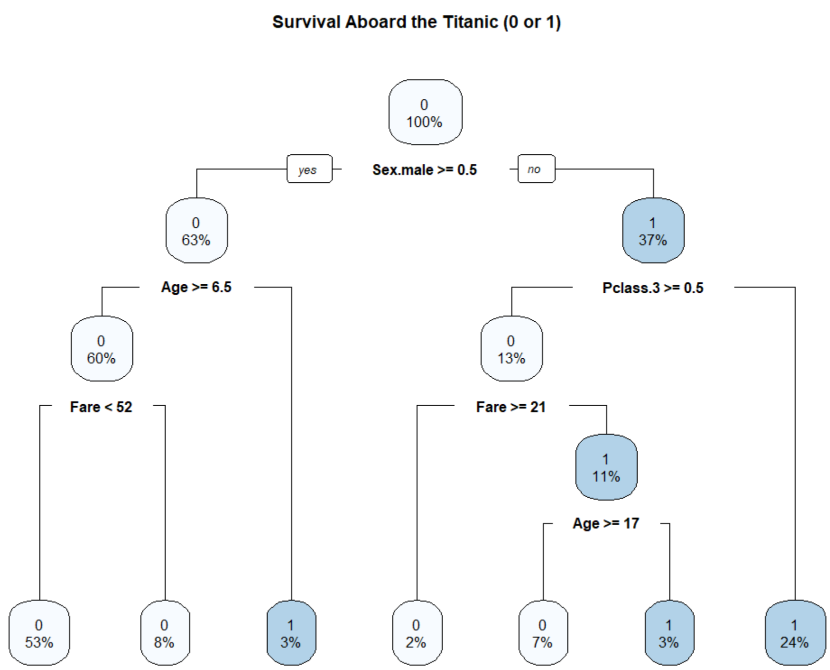
University of Essex
Department of Mathematical Sciences

MA831: Capstone Project Dissertation

Automating the Analysis Process: automodel

Robbie Mowforth, rm18379



Supervisor: **Dr. Yanchun Bao**

April 15, 2022
Colchester

Background

Whilst working on this dissertation, the goal has changed quite a few times. Firstly, this was a project on investigating methods that deal with data-sets that include missing values. After some discussion with my supervisor she mentioned the opportunity to model a larger data-set that was on the proteins found within blood samples. I decided to take on this opportunity, however, this data-set didn't have a public release when I first stated. To prepare for it's release, I did analysis on a similar data-set which I was planning to replicate.

As this analysis went on, I found myself enjoying the challenge of trying to generalize the methods I was using. I also came to learn that due to some unfortunate circumstances, the larger data-set on proteins within blood samples wouldn't be available for public release until after my dissertation. Due to these circumstances I decided to change the angle of my project and turned it into an attempt to automate the analysis process, hence the title: **Automating the Analysis Process: automodel**.

I have had an amazing time making, working and writing on this topic which naturally evolved from my given circumstances. I am hoping to continue working with the results from this dissertation beyond it's finish date and hope to have a better version in the years to come.

I would like to thank the following for their support during this dissertation:

- My supervisor Dr. Yanchun Bao for the mass amount of support and mathematical prowess
- Prof. Meena Kumari for her wealth of knowledge in the Bio-Social field
- Anna Dearman for her technical skills, know-how of Bio-Social data and proof reading
- My mother for helping me improve the words I've put on the page
- The reader for attempting to understand the ramblings I've written on the pages below

Contents

1 Introduction

This dissertation covers the learning of new analysis techniques at a technical/-mathematical level and the application of them within the statistical coding language R.

1.1 Description

The result of this dissertation is a package called **automodel**. **automodel** aims to provide an introduction to the predictive/modelling power within a given data-set. **automodel** provides a range of different models, results and statistics that can be used to inform any further in-depth analysis rather than being considered a final result.

The recommended workflow when using **automodel** is to run the main function, **autoModel**, using the desired dependant variable within a data-set. The results of this run should then be used to inform any further in-depth analysis instead of a final result. Lets say we have a data-set which has a variable we are interested in modelling/predicting. We can run **autoModel** without needing to have any context on the data-set and generate some initial models we can analyze. From here we can use the results of these models to inform the next best steps for modelling/predicting our variable of interest.

autoModel take two inputs:

- The name of the variable which the user wishes to predict (the dependant variable)

- The data-set which includes all the variables to be used when modelling, including the dependant variable

An example call to this function would be

```
autoModel("dependant variable", data)
```

The result of calling this function is a:

- cleaned, reduced & transformed version of the data-set inputted

- K-Means cluster model with predictions

- kNN cluster model with training data

- CART model with predictions

- Refined Ordinary Linear Regression model with predictions

- Best Alpha Elastic Net Regression model with predictions

1.2 Example

The data-set **titanicData** includes data on the passengers of the Titanic. Each observation of the data-set includes information on the individual passengers (**Age**, **Name**) and how they interacted with the Titanic (**Ticket**, **Fare**). In this data-set, we have the variable **Survived**, which is a binary variable where 1 means the passenger survived and 0 means the passenger didn't survive. We are going to use **automodel** to enable us to predict **Survived** using all other variables in **titanicData**.

To use **automodel** in R, we run the following command:

```
titanicResults = autoModel("Survived", titanicData)
```

This command creates all of our results and stores them in a variable called **titanicResults**. Let's evaluate some of the results within this variable.

Below are the predictions made by the kNN model. To get our results table, we execute the following in R

```
titanicResults$kNNResults$predictions
```

Table 1 is the results table from running the above shows what the model has predicted for **Survived** (if a passenger aboard the Titanic survived or not) per passenger.

	predicted	real	freq
1	0	0	86
2	1	0	27
3	0	1	23
4	1	1	41

Table 1: kNN Example Predictions

- The model predicted a passenger didn't survive (**Survived** = 0) correctly 86 times (see row 1)
- The model predicted a passenger didn't survive (**Survived** = 0) incorrectly 23 times (see row 3)
- The model predicted a passenger did survive (**Survived** = 1) correctly 41 times (see row 4)
- The model predicted a passenger did survive (**Survived** = 1) incorrectly 27 times (see row 2)
- In total: The model predicted correctly 127 times meaning the model was correct for 72% of the predictions
- In total: the model predicted in-correctly 50 times meaning the model was in-correct for 28% of the predictions

Next, let's evaluate the predictions made by the Linear Regression model. To get our results, we execute the following in R

```
summary(titanicResults$linearRegression$model)
```

The results from the model show what variables were considered significant in predicting **Survived**.

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.14503 -0.26114 -0.06787  0.23964  0.99263

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.97842    0.03987  24.539  < 2e-16 ***
Pclass.2    -0.21527    0.04990  -4.314 1.91e-05 ***
Pclass.3    -0.40917    0.04592  -8.910  < 2e-16 ***
Sex.male    -0.46835    0.03616 -12.953  < 2e-16 ***
Age         -0.08749    0.01987  -4.402 1.30e-05 ***
SibSp.4     -0.22467    0.10165  -2.210   0.0275 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3929 on 524 degrees of freedom
Multiple R-squared:  0.3745,    Adjusted R-squared:  0.3685
F-statistic: 62.73 on 5 and 524 DF,  p-value: < 2.2e-16
```

We can see that out of all the variables given to **autoModel**, it has deemed the following variables significant in predicting/modelling **Survived**:

- the binary variable of **Pclass.2**, describes if the passenger had a ticket of class 2 or not
- the binary variable of **Pclass.3**, describes if the passenger had a ticket of class 3 or not
- the binary variable **Sex.male**, describes if the passenger was male or female
- **Age**, the age of the passenger
- the binary variable **SibSp.4**, describes if the passenger had a 4 siblings/ spouses aboard the Titanic or not

Why these variables might be predictive is beyond the scope of the **automodel** package, but they certainly throw up some interesting questions for those using the **titanicData** to investigate further. This is the overall goal of the **automodel** package, to hopefully provide the user with interesting insights on their data-set which they can go on to investigate further.

1.3 Video Tutorial

A video tutorial of the **automodel** package is available at the following link: https://www.youtube.com/watch?v=uL2BBJ9gBGM

2 Data

To demonstrate the results generated from using the **automodel** package, we are going to parse multiple different data-set though the **autoModel** function. In this section we will cover the context behind each data-set used and give a brief preliminary analysis on each them.

2.1 Understanding Society & GHQ

Understanding Society (also known as The UK Household Longitudinal Study or UKHLS) is a study that measures a large amount of variables based on a participants well-being, lifestyle & life choices. It aims to have a varied pool of participants from all different backgrounds so that there's a fair representation of the population. "The sample is large enough to have around 10,000 people for each birth cohort per decade from the 1940s on-wards. They also have approximately 17,000 children who have been born into the Study since the year 2000" [1].

We are interested in using the Main Survey and Nurse Visit survey from Understanding Society. The surveys measurements are taken in "waves", where each UKHLS wave takes around two years to be fully processed across all participants. Even though the waves cross-over, each participant is interviewed roughly 1 year apart. Wave 2 was started in the year 2010 and wave 3 was started in the year 2011 with the study of both ranging over the time period 2010-2013 [2]. randomSeed To demonstrate the usage of **automodel** package, we will attempt to predict the variable **scghq1_dv**, which is a total score of a participant gathered from a survey called GHQ (more information can be found on the GHQ in Section 2.1.3).

2.1.1 Data Breakdown

These data-sets are quite large and therefore take a long time to be fully processed when using **automodel**. In Table 2 is a breakdown of each of the data-sets and their exact dimensions. All data-sets used were in *.tab* format. Table 2 describes the number of variables and observations in each data-set, if they include data on the GHQ and what wave the data is measured from.

Data-set	No. Obs	No. Vars	Contains GHQ?	Wave
b_indresp	54569	1652	TRUE	2
b_income	82111	23	FALSE	2
c_indresp	49692	3058	TRUE	3
c_income	76099	23	FALSE	3
xindresp_ns	20699	355	TRUE	2 & 3
xlabblood_ns	13258	35	FALSE	2 & 3

Table 2: Initial Understanding Society data-sets

- **b_indresp**: Main survey data from wave 2 of the UKHLS
- **b_income**: Survey data based around a participants income in wave 2. Participants could choose not to answer
- **c_indresp**: Main survey data from wave 3 of the UKHLS
- **c_income**: Survey data based around a participants income in wave 3. Participants could choose not to answer these questions
- **xindresp_ns**: Data from a nurse visiting the participants in waves 2 & 3. The nurses were to visit each participant once in both waves 2 & 3. Not all participants in the Nurse Visit were included in the Main Survey.
- **xlabblood_ns**: This is the results from the blood samples taken from the nurses during the nurse visits. Participants could choose to not have their blood sampled meaning there's significantly less observations compared to **xindresp_ns**

2.1.2 Initial Data Investigation

Understanding Society data encodes missing data into negative integers based on the reason why it's missing:

-9: missing, data is purely missing

-8: inapplicable, answering a previous question meant they didn't need to answer some following questions

-7: proxy respondent, someone else answered the question on behalf of the participant

-2: refused, refused to answer the question

-1: don't know, no clear reason on why data is missing

When we use this data with **automodel**, we are going to treat all missing values as the same regardless of context. This means we encode -9, -8, -7, -2, -1 as NA within R (See Section 3.7.1)

To preform merges/joins between these 6 data-sets, we use two variables:

- **pidp**: The unique identifier used to recognize each individual participant. Takes the form of a random integer value.
- **wave**: a column initially found in **xindresp_ns** & **xlabblood_ns**, it states the wave which the data is from, example: 2. For a join between **xindresp_ns** & **xlabblood_ns** to be possible with the other data-sets, we need to state the wave the data is from, therefore we manually add the variable wave to **b_indresp**, **b_income**, **c_indresp**, **c_income**. This process is explained further in Section 4.2.

As expected, the survey per wave have different sets of variables/measurements. In total there are:

- 1186 variables both measured in **b_indresp** & **c_indresp**

- 466 variables that are in **b_indresp** and not in **c_indresp**

- 1872 variables that are in **c_indresp** and not in **b_indresp**

Due to these differences, when analyzing **b_indresp** & **c_indresp** in the same data-set, we will only use the 1186 shared variables. More about this shared data-set will be explained in Section 4.1.

2.1.3 GHQ Preliminary Analysis

The General Health Questionnaire (GHQ) was authored by David Goldberg in 1978. "The GHQ focuses on the client's ability to carry out 'normal' functions and the appearance of any new disturbing phenomena." [3]. In more lay-mans terms, the GHQ is designed to measure the psychological distress of a participant. The version used within the Understanding Society survey is the GHQ-12, which consists of 12 questions. The GHQ-12 is descried as a "A quick screener for survey use" [3] and therefore the results should not be used to diagnose mental health. For each question, 1 of 4 responses can be chosen (with each response garnering a score between 1 to 4). The scores from each question are them summed up to create the GHQ score (**scghq1_dv**) of a participant (in the GHQ-12, the minimum GHQ score is 0 and the maximum is 36) The name of each of the questions in the GHQ-12 have the following format:

- sc: This stand for *Self Completion* and means that the questions were filled out my the participant themselves

- ghq: Represents that these questions are from the GHQ-12 used in the Understanding Society questionnaire

- a-l: The letter at the end of the question denotes each individual question used within the GHQ-12.

Knowing the above, the questions used in the GHQ-12 and their responses are:

scghqa: *Concentration*, Have you recently been able to concentrate on whatever you're doing?

- better than usual, score: 0
- same as usual, score: 1
- less than usual, score: 2
- much less than usual, score: 3

scghqb: *Loss of sleep*, Have you recently lost much sleep over worry?

- not at all, score: 0
- no more than usual, score: 1
- rather more than usual, score: 2
- much more than usual, score: 3

scghqc: *Playing a useful role*, Have you recently felt that you were playing a useful part in things?

- more than usual, score: 0
- same as usual, score: 1
- less so than usual, score: 2
- much less than usual, score: 3

scghqd: *Capable of making decisions*, Have you recently felt capable of making decisions about things?

- more so than usual, score: 0
- same as usual, score: 1
- less so than usual, score: 2
- much less capable, score: 3

scghqe: *Constantly under strain*, Have you recently felt constantly under strain?

- not at all, score: 0
- no more than usual, score: 1
- rather more than usual, score: 2
- much more than usual, score: 3

scghqf: *Problem overcoming difficulties*, Have you recently felt you couldn't overcome your difficulties?

- not at all, score: 0
- no more than usual, score: 1
- rather more than usual, score: 2
- much more than usual, score: 3

scghqg: *Enjoy day-to-day activities*, Have you recently been able to enjoy your normal day-to-day activities?

- more than usual, score: 0
- same as usual, score: 1
- less so than usual, score: 2
- much less than usual, score: 3

scghqh: *Ability to face problems*, Have you recently been able to face up to problems?

- more so than usual, score: 0
- same as usual, score: 1
- less so than usual, score: 2
- much less able, score: 3

scghqi: *Unhappy or depressed*, Have you recently been feeling unhappy or depressed?

- not at all, score: 0
- no more than usual, score: 1
- rather more than usual, score: 2
- much more than usual, score: 3

scghqj: *Losing confidence*, Have you recently been losing confidence in yourself?

- not at all, score: 0
- no more than usual, score: 1
- rather more than usual, score: 2
- much more than usual, score: 3

scghqk: *Believe in self-worth*, Have you recently been thinking of yourself as a worthless person?

- not at all, score: 0
- no more than usual, score: 1
- rather more than usual, score: 2
- much more than usual, score: 3

scghql: *General happiness*, Have you recently been feeling reasonably happy, all things considered?

- more so than usual, score: 0
- same as usual, score: 1
- less so than usual, score: 2
- much less than usual, score: 3

Since GHQ is present in multiple data-sets, we should preform some checks on the differences in distribution across the data-sets we are loading in. Please see Figure 1 for the density plots of each data-set that includes GHQ data. We can see that even though there is noticeable differences between the distributions, the GHQ score (*scghq1_dv*) falls within a similar distribution for all data-sets.

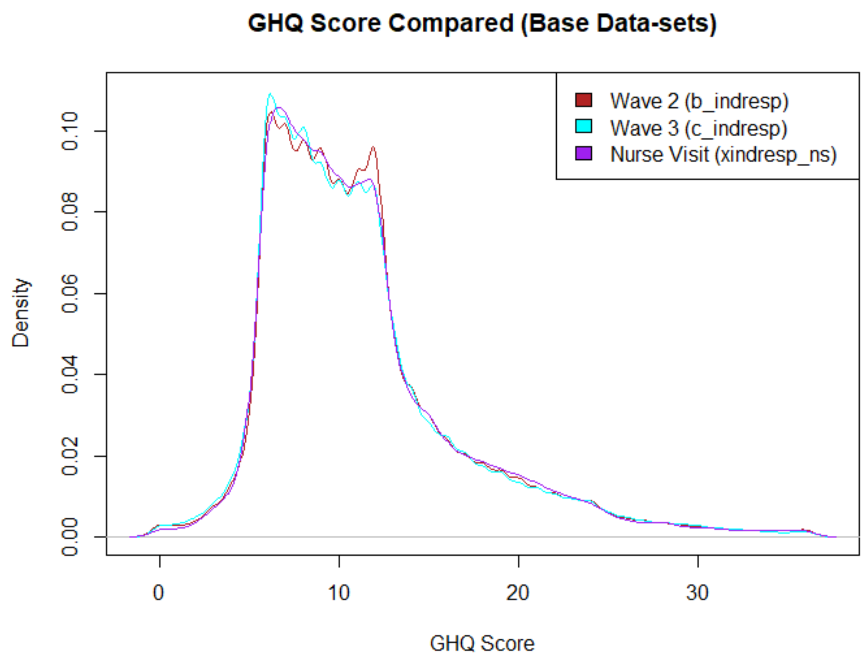


Figure 1: Density of base data-sets

Now lets do some two sided t.tests and two sided f.tests on the GHQ score from each data-set. The below test were conducted in the coding language R. The variable names used are described in Section 4.1, for now, use the titles given before each test result as a guide.

Firstly, tests for Wave 2 (**b_indresp**) and Wave 3 (**c_indresp**)

```
        Welch Two Sample t-test

data:  w2indresp$scghq1_dv and w3indresp$scghq1_dv
t = 3.3084, df = 83601, p-value = 0.0009388
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.05135789 0.20066465
sample estimates:
mean of x mean of y
 11.20125  11.07524

        F test to compare two variances

data:  w2indresp$scghq1_dv and w3indresp$scghq1_dv
F = 0.99797, num df = 43422, denom df = 40575, p-value = 0.8349
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.979047 1.017249
sample estimates:
ratio of variances
         0.9979685
```

Secondly, tests for Wave 2 (**b_indresp**) and Nurse Visit (**xindresp_ns**)

```
        Welch Two Sample t-test

data:  w2indresp$scghq1_dv and mixNurse$scghq1_dv
t = 0.30387, df = 36019, p-value = 0.7612
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.07931319  0.10841751
sample estimates:
mean of x mean of y
 11.20125  11.18670

        F test to compare two variances

data:  w2indresp$scghq1_dv and mixNurse$scghq1_dv
F = 1.0125, num df = 43422, denom df = 18842, p-value = 0.3166
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9882163 1.0371829
sample estimates:
ratio of variances
          1.012463
```

Lastly, tests for Wave 3 (**c_indresp**) and Nurse Visit (**xindresp_ns**)

```
        Welch Two Sample t-test

data:  w3indresp$scghq1_dv and mixNurse$scghq1_dv
t = -2.3021, df = 36968, p-value = 0.02133
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.20635509 -0.01656312
sample estimates:
mean of x mean of y
 11.07524  11.18670

        F test to compare two variances

data:  w3indresp$scghq1_dv and mixNurse$scghq1_dv
F = 1.0145, num df = 40575, denom df = 18842, p-value = 0.2485
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9899783 1.0395630
sample estimates:
ratio of variances
          1.014524
```

Due to the large number of observations within each data-set (degrees of freedom), the tests are rather sensitive to differences between values. This can be seen above as the t.test between Wave 2 and Wave 3 (**b_indresp** & **c_indresp**) rejects the null hypothesis at the 95% confidence interval (p-value of 0.0009388) when the mean of Wave 2 is 11.20125 and the mean of Wave 3 is 11.07524 (a difference of 0.126). The t.test between Wave 3 and Nurse Visit (**c_indresp** & **xindresp_ns**) also rejected the null hypothesis at the 95% confidence interval (p-value 0.02133) when the mean of Wave 3 is 11.07524 and the mean of Nurse Visit is 11.18670 (a difference of 0.111). Even though the true differences between these means are statistically significant, the magnitude of the difference is rather small given the context of the variable.

All of tests preformed failed the reject the null hypothesis. Considering the conditions of the rejections found in the previous tests mentioned, it's safe to assume that the differences between the means and variances of the GHQ score per data-set is negligible.

Now let's have a look at the summaries of the GHQ score (*scghq1_dv*) and the individual questions (**scghqa - scghql**). These results can be seen in Table 3, Table 4 and Table 5. Given the above investigation into the distributions, we are only going to look at the Wave 2 (**b_indresp**) summaries of the variables. The same summaries on all data-sets can be found Section 10.

scghq1_dv	
Min	0
1st Q	7
Median	10
Mode	6
Mean	11.2
3rd Q	13
Max	36
NA's	11146
Non-NA's	43423

Table 3: GHQ score Summary

Names	scghqa	scghqb	scghqc	scghqd	scghqe	scghqf
Min.	1	1	1	1	1	1
1st Q	2	1	2	2	1	1
Median	2	2	2	2	2	2
Mean	2.147	1.876	2.082	2.03	1.997	1.789
3rd Q	2	2	2	2	2	2
Max.	4	4	4	4	4	4

Table 4: GHQ Questions a - f Summary

Names	scghqg	scghqh	scghqi	scghqj	scghqk	scghql
Min.	1	1	1	1	1	1
1st Q	2	2	1	1	1	2
Median	2	2	2	2	1	2
Mean	2.159	2.063	1.844	1.728	1.445	2.056
3rd Q	2	2	2	2	2	2
Max.	4	4	4	4	4	4

Table 5: GHQ Questions g - l Summary

Secondly, we can have a look at Figure 2 for the histogram of the scghq1_dv variable to understand the distribution. We can see that we have a rather positive skew distribution in the GHQ score. The most frequent GHQ scores are within the 6 - 12 range.

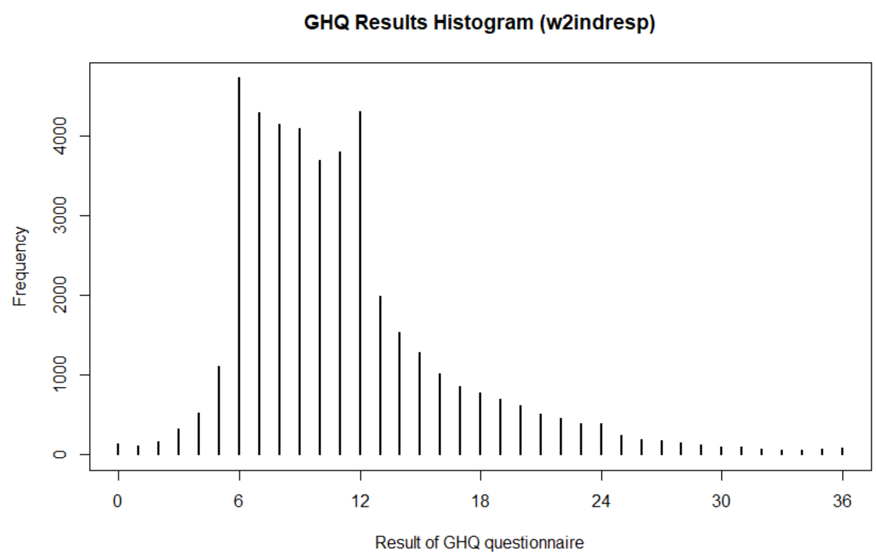


Figure 2: Distribution Plot of Wave 2

Thirdly, we will have a look at Figure 3 for the qqPlot of the GHQ score. We can see that the GHQ score doesn't fit the normal distribution and has a rather large right tail [4].

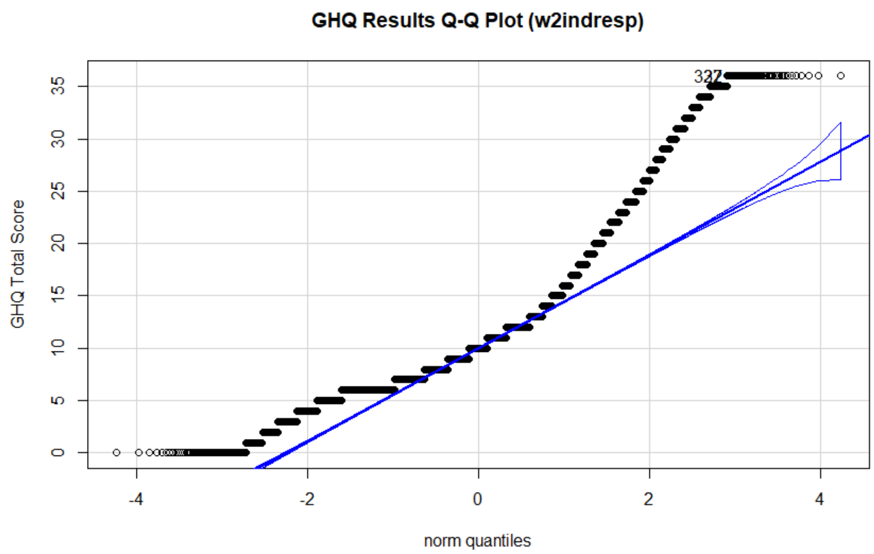


Figure 3: qqPlot of Wave 2 (Normality)

Lastly, we will have a look at Figure 4 for the correlation matrix between all of individual GHQ-12 questions. From the results we can see all that all the correlation coefficients calculated are less than 0.75 meaning there's good reason to assume the questions we have in the GHQ garner varied responses from different participants.

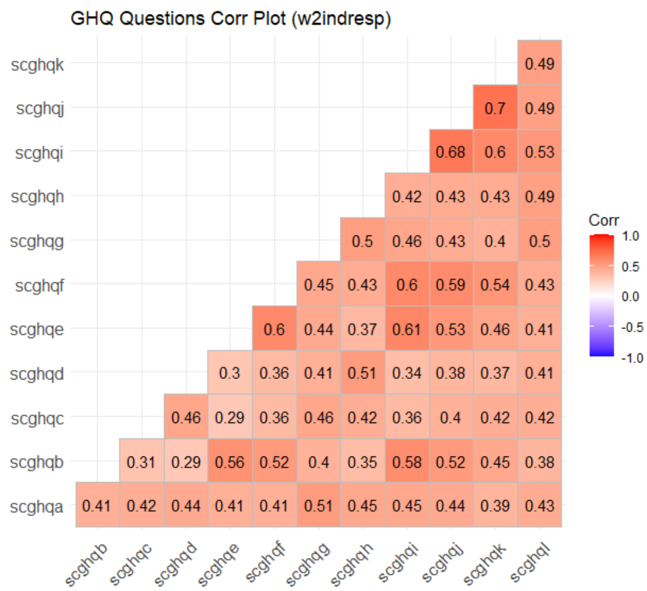


Figure 4: correlations between questions in Wave 2

Let's have a look at some of the most correlated questions:

- ghqe (Constantly under strain) & ghqf (Problem overcoming difficulties): coef of 0.6

- ghqe (Constantly under strain) & ghqi (Unhappy or depressed): coef of 0.61

- ghqf (Problem overcoming difficulties) & ghqi (Unhappy or depressed): coef of 0.6

- ghqi (Unhappy or depressed) & ghqj (Losing confidence): coef of 0.68

- ghqj (Losing confidence)& ghqk (Believe in self-worth): coef of 0.7

This finishes the GHQ preliminary analysis, we will continue the analysis of the Understanding Society Data in Section 4.

2.2 Kaggle Titanic

The Titanic data comes from a Kaggle Machine Learning Competition where participants should create a model which predicts weather a passenger survived or not [5]. This competition is considered as an entry point for those new to machine learning and Kaggle. The data-set provided by Kaggle has already been split into a train and test set. In this dissertation we will be running this data though our function and submitting our results onto the Kaggle leader-board to compare the results.

The Titanic **train.csv** data-set contains 891 observations of 12 variables. Here's a breakdown of what's included:

- **PassengerId**: A unique identifier for each passenger
- **Survived**: Boolean variable, 1 = Survived, 0 = Not Survived. This is our dependant variable, the variable we are looking to predict
- **Pclass**: A passengers ticket class on the titanic, is either 1,2 or 3 (1st, 2nd or 3rd class respectively).
- **Name**: A passengers full name
- **Sex**: Boolean variable, represents the sex of a passenger (either 'male' or 'female').
- **Age**: represents a passengers age in years
- **SibSp**: number of siblings / spouses aboard the titanic
- **Parch**: number of parents / children aboard the titanic
- **Ticket**: The unique Ticket number
- **Fare**: How much was paid by the passenger to board
- **Cabin**: the cabin number the passenger stayed in
- **Embarked**: The port of embarkation (where the passenger boarded the Titanic). C = Cherbourg, Q = Queenstown, S = Southampton.

In the data-set, we are going to be attempting to model the variable **Survived**. In Table 6, we can see a breakdown of the variable **Survived**. We can see that the majority of people didn't survive the titanic, 61% didn't survive and 39% didn't.

Survived	
0 (Didn't Survived)	1 (Survived)
549	342

Table 6: Breakdown Of **Survived** Variable

Overall this data-set provides a good amount of observations and variables to generate initial models with **automodel**. The run of this data-set using **automodel** will be documented in Section 5.

2.3 Edgar Anderson's Iris Data

The Edgar Anderson's Iris data-set is in-built into R as a data-set to be used for initial learning/testing. The data-set "gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris"[6]. The data-set is commonly used to demonstrate clustering and classification algorithms due to the nature of the 3 iris species. In this dissertation, we will be modelling the variable **Species**/ treating **Species** as our dependant variable.

The Iris data-set contains 150 observations of 5 variables. Here's a breakdown of whats included:

- **Sepal.Length**: The length of the flowers sepal (the green leaf that covers the flower before bloom) in centimeters
- **Sepal.Width**: The width of the flowers sepal in centimeters
- **Petal.Length**: The length of the petals from the flower
- **Petal.Width**: The width of the petals from the flower
- **Species**: The species of Iris flower

Next, the summary statistics for the Iris data variables are in Table 7 & Table 8. The below shows that there's a decent distribution across the continuous variables and the categorical variable, **Species**, is equally distributed.

Name	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Min	4.3	2	1	0.1
1st Q	5.1	2.8	1.6	0.3
Median	5.8	3	4.35	1.3
Mean	5.843	3.057	3.758	1.199
3rd Q	6.4	3.3	5.1	1.8
Max	7.9	4.4	6.9	2.5

Table 7: Continuous Variable Summaries

Species	Count
Setosa	50
Versicolor	50
Virginica	50

Table 8: Categorical Variable Summaries

A very neat graphical summary of the data can be seen in Figure 5 [7]. In Figure 5 we can see the scatter-plots of all the continuous variables included in Iris coloured by **Species** (black is *setosa*, red is *versicolor*, green is *virginica*). The plot shows that the Iris data-set seems to have some nice clustering within it's data and therefore would benefit from a clustering model.

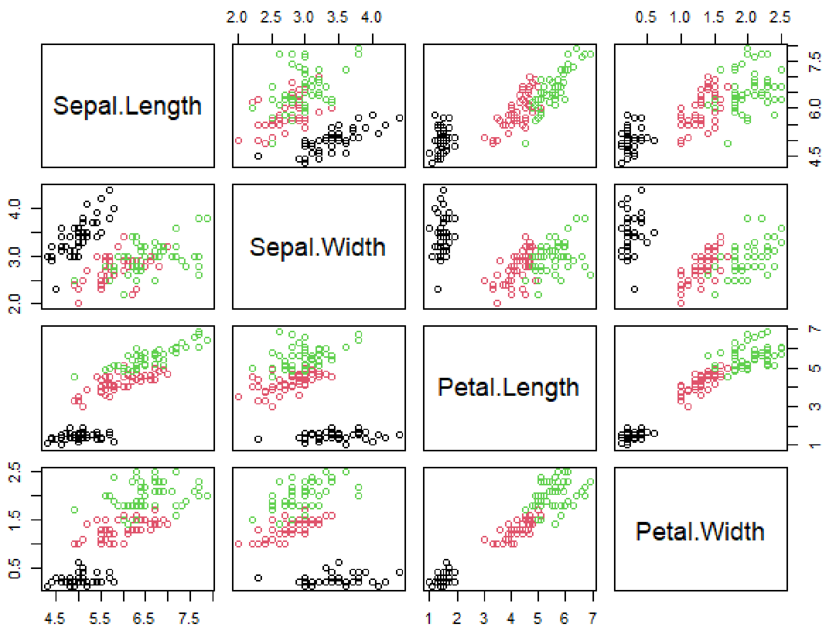


Figure 5: Iris Data-set Plotted

Since we have a small amount of variables, it's worth us doing a brief correlation check on the variables. Figure 6 shows the correlation matrix between all the continuous variables within the Iris data-set. We can see that there is a high amount of correlation between the variables, suggesting a regression model will be a poor fit due to high multicollinearity.

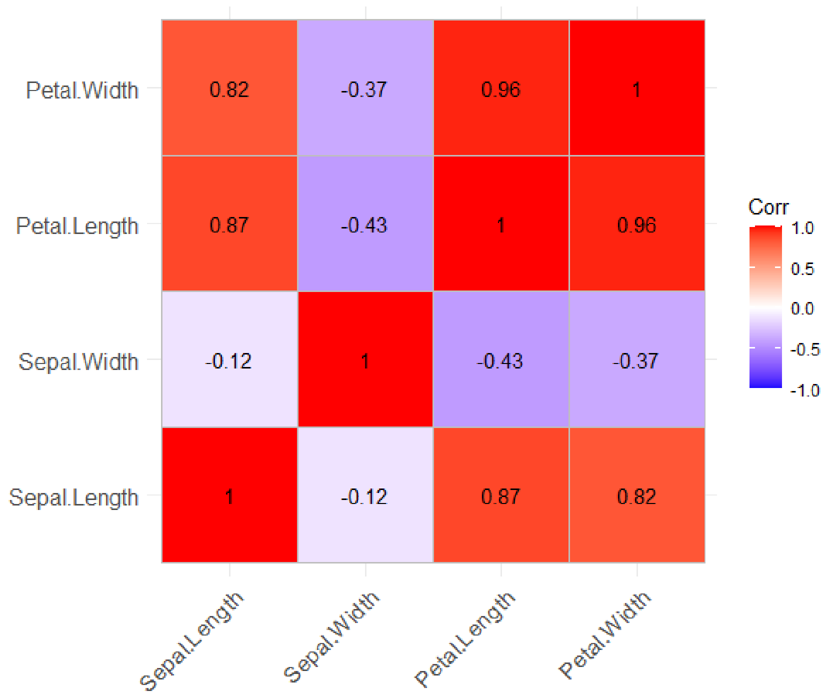


Figure 6: Iris Correlation Matrix

Overall, though our initial analysis, the Iris data looks to be a data-set where some clusters can be easily identifiable, therefore it's a good test for **automodel** to see if it picks up these clusters efficiently. We will continue the usage of the Iris data in Section 6.

3 Methodology

In the **automodel** package, we have various different methods & functions. In this Section we will be covering the details behind each R Package, method and function used within the **automodel** package.

3.1 R Packages

To produce our results, a fair amount of packages/libraries have been used. In this Section we will credit the creators of the packages/libraries and explain how they have been used in this dissertation.

- **broom**

Made/Maintained by: Simon Couch & Team

Reference: https://cran.r-project.org/web/packages/broom/broom.pdf

Usage: Data cleaning tools. The function *tidy()* is used within **automodel**.

- **car**

Made/Maintained by: John Fox & Team

Reference: https://cran.r-project.org/web/packages/car/car.pdf

Usage: Variance Inflation Factor (VIF) and qqPlots. The function *vif()* is used within **automodel** and the function *qqPlot()* is used within the custom assisting function *ghq_analyze()*

- **caret**

Made/Maintained by: Max Kuhn & Team

Reference: https://cran.r-project.org/web/packages/caret/caret.pdf

Usage: Dummy-tization of variables (creating dummy variables). The function *dummyVars()* is used within **automodel**

- **class**

Made/Maintained by: Brian Ripley, William Venables

Reference: https://cran.r-project.org/web/packages/class/class.pdf

Usage: k Nearest Neighbours (kNN). The function *knn()* is used within *automodel*

- **devtools**

Made/Maintained by: Jennifer Bryan & Team

Reference: https://cran.r-project.org/web/packages/devtools/devtools.pdf

Usage: Local R package management. Used to document, format and install **automodel** as a R package.

- dplyr

Made/Maintained by: Hadley Wickham & Team

Reference: https://cran.r-project.org/web/packages/dplyr/dplyr.pdf

Usage: tools for working with data. The function *select()* and *all_of()* is used within **automodel**

- ggcorrplot

Made/Maintained by: Alboukadel Kassambara

Reference: https://cran.r-project.org/web/packages/ggcorrplot/ggcorrplot.pdf

Usage: Correlation Matrix plots. The function *ggcorrplot()* is used within the custom assisting function *ghq_analyze()*

- glmnet

Made/Maintained by: Trevor Hastie & Team

Reference: https://cran.r-project.org/web/packages/glmnet/glmnet.pdf

Usage: Elastic Net Regression Model. The function *cv.glmnet()* is used within **automodel**

- Matrix

Made/Maintained by: Martin Maechler & Team

Reference: https://cran.r-project.org/web/packages/Matrix/Matrix.pdf

Usage: Matrix manipulation. Needed so that we can work with sparse matrices within **automodel**

- mice

Made/Maintained by: Stef van Buuren & Team

Reference: https://cran.r-project.org/web/packages/mice/mice.pdf

Usage: Missing value imputation. The main function for imputing is *mice()*, used within **automodel**.

- readr

Made/Maintained by: Jennifer Bryan & Team

Reference: https://cran.r-project.org/web/packages/readr/readr.pdf

Usage: Reading .tab files. The function *read_table()* is used to load in the Understanding Society data.

- rlang

Made/Maintained by: Lionel Henry & Team

Reference: https://cran.r-project.org/web/packages/rlang/rlang.pdf

Usage: tools for working with data. The function *is_empty()* is used within **automodel**.

- **rpart**

Made/Maintained by: Terry Therneau, Beth Atkinson, Brian Ripley

Reference: https://cran.r-project.org/web/packages/rpart/rpart.pdf

Usage: Classification and Regression Tree Models (CART). The function *rpart()* is used within *automodel*.

- **sqldf**

Made/Maintained by: G. Grothendieck

Reference: https://cran.r-project.org/web/packages/sqldf/sqldf.pdf

Usage: Running SQL queries in R. the function *sqldf()* is used to transform the income data from Understanding Society (See Section 4.3 Data Processing)

- **stats**

Made/Maintained by: R Core Team

Reference: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html

Usage: General statistical tools. The following functions are from stats: *AIC*, *coef*, *cor*, *kmeans*, *lm*, *na.omit*, *predict*, *princomp*, *runif*, *sd*.

- **stringr**

Made/Maintained by: Hadley Wickham & Team

Reference: https://cran.r-project.org/web/packages/stringr/stringr.pdf

Usage: Simple string manipulations. The functions *str_remove()* and *str_replace()* are used within **automodel**.

3.2 Cleaning Functions

The first process we need to preform to our data-sets is cleaning. This process aims to put the data-set into a readable format for the transformation process (See Section 3.3 Transforming).

3.2.1 Data-frame Checks

Sometimes the data parsed to the function might not be in the right format for analysis. In this method we make sure that the data is interpreted correctly by casting the entire data set to a data.frame as setting blank data cells as 'NA'. This is achieved though the following two lines.

```
data = as.data.frame(data)
data[data == ""] = NA
```

3.2.2 Data Type Checks

When a variable contains any other data type that numeric, we can't analyze the variable therefore we have this check to remove/change non-numeric data. This method works by detecting which variables are non-numeric and re-codes them into factor variables. Values are re-coded in numerical order then in alphabetical order, please see Table 9 for an example run. (if 1 = "male", 2 = "female", make sure variables is all numeric or text)

	Example Transform
Original	1345, 32233, "sheep", "kangaroo", "kangaroo", NA
After	1, 2, 4, 3, 3, NA

Table 9: Data Type Check Example

3.2.3 High Missing Data Variables

If a variable has a large proportion of their data missing, this may lead to all of the data being removed if **automodel** preforms List-wise deletion. To avoid this, we remove variables which have a certain proportion or more of their data as NA/missing. This proportion is called **naPercent** and can be custom set on each run (default is 20% or 0.2). This means, by default, this function removes variables which have 20% or more of their data as NA/missing.

Table 10 is an example on a arbitrary data-set with 10 observations, we will be assuming that **naPercent** is it's default value (20%) meaning we will remove variables with 2 or more observations as NA/missing

3.2.4 Missing Value Imputation

In our data-sets, it's most likely that there will be values that are missing. These missing values can make it quite hard to generate model and therefore introduce a wealth of issues as data-sets gets large (more variables and observations).

Original Data					Transformed Data		
VarA	VarB	VarC	VarD	VarE	VarA	VarC	VarE
0	NA	100	NA	"yes"	0	100	"yes"
1	28	101	NA	"yes"	1	101	"yes"
1	NA	102	NA	"yes"	1	102	"yes"
0	27	103	NA	"no"	0	103	"no"
0	NA	104	NA	"no"	0	104	"no"
1	103.2	105	NA	NA	1	105	NA
0	NA	106	NA	"no"	0	106	"no"
1	NA	107	NA	"yes"	1	107	"yes"
1	107.4	108	NA	"yes"	1	108	"yes"
1	NA	109	NA	"yes"	1	109	"yes"

Table 10: High Missing Data Variables

In this paper, we will use Single Imputation to handle missing values within a data-set. In this Dissertation, we use Classification and Regression Trees (CART) as our imputation method. This is because at the stage where we want to impute our missing values, it's possible for a data-set to have a highly correlated variables which would cause issues with most regression imputation methods. With CART, correlated variables don't effect the model as much and therefore can be used. We achieve this using the mice package (See Section 3.1) in R. Imputation works by initially assigning a temporary value for each missing observation then refines these values by generating CART models to predict these values [9].

To run an imputation, we use the following code:

```
imp = mice(data,
           seed = randomSeed,
          m=1,
           maxit = 5,
           method = "cart",
           threshold = 2)
```

Here's an in-depth explanation of the above code:

- ```
  imp
  ```

  This is the object that holds the imputed values, the object type is a MICE defined *mids* object.

- ```
  mice()
  ```

  The function which runs our missing value imputation on our data

- ```
  data
  ```

  The data-set which we wish to impute missing data for

- ```
  seed = \textbf{randomSeed}
  ```

This sets the randomness in the function to a fixed random space, makes the results reproducible if wanted. **randomSeed** is a function variable (See Section 3.6).

- ```
  m = 1
  ```

  This is what tells us that we are doing single imputation. We only create one iteration of imputed values for the missing values.

- ```
  maxit = 5
  ```

  This is the amount of iterations the single imputation goes though. This basically means how many times we create a CART model for each variable. This allows for the missing values to converge on a best fit from the initially assigned temporary values.

- ```
  method = "cart"
  ```

  This specifies that we are imputing the missing values using classification and regression trees. See Section 3.4.3 for a further explanation of this imputation method.

- ```
  threshold = 2
  ```

  MICE will automatically remove predictor variables based on co-linearity. Since we have checks for co-linearity later on, we set the variable 'threshold' to 2 so that mice doesn't automatically ignore/remove highly correlated variables (since we are using CART as our imputation method, we don't have to worry about co-linearity when imputing)

This function can ran by the user by setting the function variable **impFlag** to TRUE (**impFlag** is FALSE by default since imputation can be a timely process). The process can also be speed up by increasing the **automodel** variable **cartSplit** however this can results in the CART models made being a worse fit for the data-set.

There are some assumptions associated with missing data that are important to know for missing value imputation

- **MCAR**: Missing Completely at Random (MCAR) is when the reason for the missing data is unrelated to the data itself. Example: A sever glitch causes random data entries to go missing or corrupt. As Stef Van Buuren puts it [10]

  > *If the probability of being missing is the same for all cases, then the data are said to be missing completely at random (MCAR). This effectively implies that causes of the missing data are unrelated to the data*

- **MAR**: Missing at Random (MAR) is a more realistic assumption of randomness compared to MCAR and takes on a broad range of possible cases. MAR is when randomness is caused by a feature of the data. Example: when using a pedometer in summer, we may produce more missing values than if used in winter. If we can assume that for both in summer and winter, the data is MCAR, then we can assume the overall data is MAR. As Stef Van Buuren puts it [10]

  > *If the probability of being missing is the same only within groups defined by the observed data, then the data are missing at random (MAR). MAR is a much broader class than MCAR. For example, when placed on a soft surface, a weighing scale may produce more missing values than when placed on a hard surface. Such data are thus not MCAR.*

- **MNAR**: Missing Not at Random (MNAR) is applied if MCAR and MAR don't hold. It means that our reason for why there are missing values is unknown to us. Example: Without our knowledge, the longer the walk with a pedometer, the higher the likelihood of their being missing values. As Stef Van Buuren puts it [10]

  > *If neither MCAR nor MAR holds, then we speak of missing not at random (MNAR). In the literature one can also find the term NMAR (not missing at random) for the same concept. MNAR means that the probability of being missing varies for reasons that are unknown to us. For example, the weighing scale mechanism may wear out over time, producing more missing data as time progresses, but we may fail to note this*

If missing values are handled appropriately, unwanted bias can be introduced in the models created. Example: a group for people refused to answer a survey for a collective/similar reason. If the missing values caused by this aren't handled appropriately, the models created with this data-set could lead to Information Bias, Selection Bias and incorrect interpretations of the models generated [11] [12].

A criticism of using Single Imputation is that it doesn't account for the variability that comes with missing value imputation. The more accepted method for missing values is Multiple Imputation. In this method you impute the missing values multiple times (in terms of the *mice()* function, this means imputing the data m times), build models on each of the imputed data-sets then average/pool the results across all models made. Using Multiple Imputation accounts for the known variability that comes with imputing missing values (results can be quite different per imputation) and therefore allows for better modelling of data. We use Single Imputation in this package instead of Multiple Imputation for 2 reasons:

1. Having to generate & model m imputations can drastically increase the overall run-time in our function (which is already quite high for large data sets).

2. Having Single Imputation is better than not having any Imputation option at all (just List-wise deletion as the only option may leave some analysis a little short of options).

Multiple Imputation is included in Section 7.3 of this dissertation. Even though single imputation is a valid method for missing value imputation, there is good reason to account for the variability in imputations correctly by using multiple imputations.

3.2.5 List-wise Deletion

To be able to model our data, we must make sure that there is no NA values within the analyzed data-set. The simplest way to deal with this is by preforming List-wise Deletion. In simple terms, List-wise Deletion is if a observation has a NA value within any variable, the observation is removed from the data-set. This can be achieved with this very simple function in R:

```
na.omit()
```

One criticism of this method is that it can introduce selection bias. Selection Bias is when the participants in a study differ from the population of interest [12]. In List-wise deletion, we remove a large amount of data based on missing values, if these missing values are assumed to be MAR or MNAR (see Section 3.2.4) we may cause selection bias in our results. Using the Understanding Society data as an example, preforming List-wise deletion on the data will mean we don't consider the participants which refuse to answer certain questions (such as household income). This would cause a bias against a group of participants that don't answer survey questions about their household income.

When used in **automodel**, if variables with a large amount of missing values aren't removed (the method to handle this is detailed in Section 3.2.3), List-wise deletion could remove almost all observations in a data-set.

To show how List-wise deletion works, Table 11 is an example arbitrary data set with 10 observations which we will preform List-wise Deletion on

Original Data		
VarA	VarB	VarC
0	100	"yes"
NA	101	"yes"
1	102	"yes"
0	103	"no"
0	104	"no"
1	NA	NA
0	106	"no"
1	107	"yes"
1	108	"yes"
1	109	"yes"

Transformed Data		
VarA	VarB	VarC
0	100	"yes"
1	102	"yes"
0	103	"no"
0	104	"no"
0	106	"no"
1	107	"yes"
1	108	"yes"
1	109	"yes"

Table 11: List-wise Deletion example

3.2.6 Pair-wise Deletion

When accounting for missing values from data-set, we may have conditions meaning we would only want to remove observations with missing values in a select group of variables. Pair-wise deletion works by deleting observations based on the missing values within a select group of variables within a data-set. The resulting method will result in no missing values in the variables that were within the group and potentially some missing values in the variables not in the group (depends on the observations that are removed). It's worth noting that if the group of variables selected is the entire data-set, the deletion preformed will be the same as the List-wise deletion (See Section 3.2.5).

Table 12 is an example in a arbitrary data set with 10 observations. For our example Pair-Wise Deletion, we only need to remove observations which has NA values for **VarA**.

Original Data		
VarA	VarB	VarC
0	100	"yes"
NA	101	"yes"
NA	NA	"yes"
0	103	"no"
0	NA	"no"
1	NA	NA
0	106	"no"
1	107	"yes"
1	108	NA
1	109	"yes"

Transformed Data		
VarA	VarB	VarC
0	100	"yes"
0	103	"no"
0	NA	"no"
0	NA	NA
0	106	"no"
1	107	"yes"
1	108	NA
1	109	"yes"

Table 12: Pair-wise Deletion Example 1

in **automodel**, we use Pair-wise Deletion on the dependant variable within the data-set. We run this before removing variables which have a large proportion of their data as missing (see Section 3.2.3). Doing so means we can potentially end up with more variables in the overall analysis.

An example of this effect can be seen in Table 13. We do is a Pair-Wise deletion on **VarA**, then we remove variables from the analysis if they have 2 or more missing values. As seen below, if we were to remove variables based on their missing values before preforming Pair-Wise Deletion (removing based on the Original Data), we would of removed the entire data-set.

Original Data		
VarA	VarB	VarC
0	100	"yes"
NA	NA	NA
NA	NA	"yes"
0	103	"no"
0	104	"no"
NA	105	NA
0	106	"no"
1	NA	"yes"
1	108	NA
1	109	NA

Pair-Wise Data		
VarA	VarB	VarC
0	100	"yes"
0	103	"no"
0	104	"no"
0	106	"no"
1	NA	"yes"
1	108	NA
1	109	NA

Variable Selected Data	
VarA	VarB
0	100
0	103
0	104
0	106
1	NA
1	108
1	109

Table 13: Pair-wise Deletion Example 2

3.2.7 Low Level Removal

Some factor variables may have levels (unique values within the variable) which have a low amount of observations. This data wouldn't be reliable to model with so we remove those levels from the data-set. The amount of observations needed per level is determined by the function variable **obsPerLevel**, default value is 10. Like List-wise Deletion (See Section 3.2.5), if the missing data is assumed MAR or MNAR, the deletion process can cause Selection Bias. For example, say you have a dummy variable (variable that only takes the value 1 or 0) that asks if a participant has a rare condition. Since the condition is rare, you may expect there to only be 4-5 observations of it out of 10,000, if **obsPerLevel** is set to 10, this will remove all observations of the rare condition, meaning we will have bias against those who have this condition.

As an example, Table 14 is an arbitrary data-set of 10 observations. We are removing levels where there is only 1 observation. In our example:

- **VarA**: Factor variable with the levels 0,1 & 2
- **VarB**: Continuous variable, no defined levels
- **VarC**: Factor variable with the levels "yes", "no" & "maybe"

Original Data		
VarA	VarB	VarC
0	100	"yes"
1	101	"yes"
1	102	"yes"
0	103	"no"
0	104	"no"
1	105	"maybe"
0	106	"no"
1	107	"yes"
1	108	"yes"
2	109	"yes"

Transformed Data		
VarA	VarB	VarC
0	100	"yes"
1	102	"yes"
0	103	"no"
0	104	"no"
0	106	"no"
1	107	"yes"
1	108	"yes"

Table 14: Low Level Removal Example

3.2.8 Observation to Variable Ratio

During the analysis process, the data-set used may have a lack of observations to create meaningful models. Throughout the *automodel* process, there are checks on the observation to variable ratio where if the ratio is less than 5, *automodel* throws an error message [13].

Here are the following checks done and the error messages they throw:

- After List-wise Deletion (see Section 3.2.5) or Missing Value Imputation (see Section 3.2.4) is done the following error message is thrown if the observation to variable ratio is less than 5 (in the error message below, *¡data-set ratio¿* is the observation to variable ratio of the current data-set as calculated in R):

Observation to Variable ratio of ¡data-set ratio¿ is less than 5 after Listwise Deletion (or has always been if using Multiple Imputation). Consider the following:

  – *Lower the value of function variable: 'naPercent'*
  – *Run missing value imputation by setting impFlag = TRUE*
  – *Use less variables*
  – *Gather more observations*

- After Low Level Removal (see Section 3.2.7) is done the following error message is thrown if the observation to variable ratio is less than 5 (in the error message below, *¡data-set ratio¿* is the observation to variable ratio of the current data-set as calculated in R):

  Observation to Variable ratio of ¡data-set ratio¿ is less than 5 after Low Level Removal. Consider the following:

  – *Lower the value of function variable: 'obsPerLevel'*
  – *Use less variables*
  – *Gather more observations*

3.2.9 Variance Checks

If a variable has 0 variance, the variable provides no predictive power since the variables value is the same for every observation. Since these variables can cause issues in other methods, we remove them during this check. This is achieved simply by calculating the variance of all variables, then removing those which have 0 variance.

Table 15 is an example using an arbitrary data-set with 10 observations.

Original Data			Transformed Data	
VarA	VarB	VarC	VarB	VarC
0	100	"yes"	100	"yes"
0	101	"yes"	101	"yes"
0	102	"no"	102	"no"
0	103	"yes"	103	"yes"
0	104	"yes"	104	"yes"
0	105	"no"	105	"no"
0	106	"no"	106	"no"
0	107	"yes"	107	"yes"
0	108	"yes"	108	"yes"
0	109	"yes"	109	"yes"

Table 15: Variance Check Example

3.2.10 Correlation Checks

If two or more variables are correlated, this would introduce multicollinearity to our regression models, specifically our Ordinary Linear Regression Model. If a model has multicollinearity [14]:

- It becomes hard to choose a list of significant variables since the model will give heavily varied results
- Coefficient estimates can vary by a large degree
- The model could be over-fit to the data meaning when applying the model to another sample of the same data, the accuracy of the model will be poor.

For those reasons, we preform checks to remove variables which are highly correlated. In these checks, we generate a correlation matrix, then remove variables according to a limit that is set by the function variable **corrConfLevel**, default is 0.8. In the automatic approach/check, the method of removal is finding the variable which has the most correlation coefficients above the **corrConfLevel** level. The count of the correlation coefficients are done on the full correlation matrix and the lower triangle of the correlation matrix (both with their diagonals set to 0). These counts are done independent of each other and then summed to generate our criterion on which we remove our variables. This means, by default, if VarA correlated with 4 variables with a coefficient of 0.8 and VarB correlated with 3 variables with a coefficient of 0.8, VarA will be removed first.

This design was chosen because correlation is non-transitive and removing the variable with the most correlations can lead to removing less variables overall. Also, If there is a tie between variables, the leftmost variable in the data-set is removed in the correlation pair. This means the user can have some manual input on the correlation checks by moving variables they want to keep in these checks to the end of the data-set.

the correlation matrix calculates the Pearson's correlation coefficient. The equation is the following:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \tag{1}$$

- ρ_{XY}: The Pearson's correlation coefficient at the population level
- $Cov(X,Y)$: The Co-variance between X and Y
- $Var(X)$: The Variance of X
- $Var(Y)$: The Variance of Y
- X: The X variable
- Y: The Y variable

$$\rho_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

- ρ_{xy}: The Pearson's correlation coefficient at the sample level
- x_i: The ith observations of the X variable
- $\bar{x}$: The mean of the observations of X
- y_i: the ith observations of the Y variable
- $\bar{y}$: the mean of the observations of Y

To demonstrate the selection process, below is are 3 correlated variables. These values correlate in such a way so that X_1 correlates with X_2, X_2 correlates with X_3 but X_1 doesn't correlate with X_3 (demonstrating the non-transitive nature of correlations [15]).

$$X_1 = \{2, 0, 0, 1\}, X_2 = \{1, 0, 0, 1\}, X_3 = \{1, 0, 0, 2\}$$

Here's the full correlation matrix for the variables X_1, X_2, X_3

$$\begin{array}{c c} & \begin{array}{ccc} X_1 & X_2 & X_3 \end{array} \\ \begin{array}{c} X_1 \\ X_2 \\ X_3 \end{array} & \begin{pmatrix} 0 & 0.9 & 0.64 \\ 0.9 & 0 & 0.9 \\ 0.64 & 0.9 & 0 \end{pmatrix} \end{array}$$

and here's the lower triangle of the correlation matrix for the variables X_1, X_2, X_3

$$\begin{array}{c c} & \begin{array}{ccc} X_1 & X_2 & X_3 \end{array} \\ \begin{array}{c} X_1 \\ X_2 \\ X_3 \end{array} & \begin{pmatrix} 0 & 0 & 0 \\ 0.9 & 0 & 0 \\ 0.64 & 0.9 & 0 \end{pmatrix} \end{array}$$

We are looking to remove variables based on if their correlation coefficient is above 0.8 or not. For our first check, here are the counts of the full correlation matrix and lower correlation matrix summed.

- X_1: 2 coefficients
- X_2: 3 coefficients
- X_3: 1 coefficients

The above shows that this iteration of the check identified X_2 as the most correlated variable and has discriminated against X_1 over X_3 since X_1 is the leftmost variable (showing the effect a user can have on the selection process, if wanted). Given this, X_2 is removed from the data-set, the remaining correlation matrices are:

$$\begin{array}{c c} & \begin{array}{cc} X_1 & X_3 \end{array} \\ \begin{array}{c} X_1 \\ X_3 \end{array} & \begin{pmatrix} 0 & 0.64 \\ 0.64 & 0 \end{pmatrix} \end{array} \begin{array}{c c} & \begin{array}{cc} X_1 & X_3 \end{array} \\ \begin{array}{c} X_1 \\ X_3 \end{array} & \begin{pmatrix} 0 & 0 \\ 0.64 & 0 \end{pmatrix} \end{array}$$

No more correlation coefficients are above out limit of 0.8 therefore we end our correlation checking process.

If we were to simply remove the variables in order of their position in the data-set (left-most variables to be removed first), we would of removed X_1 then X_2, resulting in two variables being removed. This isn't ideal and by preforming our methods, we will end up keeping more variables overall whilst giving the user the functionality to control the results.

3.3 Transforming

After cleaning our data-set, we preform transformation methods that aim to better represent our data. This data will then be used to fit our models (See Section 3.4).

3.3.1 Factorization

For our predictions to be accurate, we need to model our variables appropriately. We do this by detecting if a variable should be considered as a factor based on the number of unique values/levels the variable has. This number of unique levels is controlled by the **automodel** variable **catLevels**, which by default is 15. If a variable has **catLevels** or less unique values/levels, the variable is encoded as a factor (categorical), otherwise is considered numeric (continuous). Factor variables will be later encoded as dummy variables in Section 3.3.3. Numerical/Continuous variables will be later scaled in Section 3.3.2.

If there's a categorical variable with a more unique levels than **catLevels** (example: a variable holding the name of a person observed), the variable will be treated as continuous. This may not be the right way to model the variable however since our goal is prediction there are some things to consider:

- If this variable was made a factor, we could end up removing a lot of levels which have less than **obsPerLevel** observations.

- Using name as a scalar may find a strange hidden relationship so therefore, considering the above, should still be considered for modelling (VIF, backwards selection etc can still remove this variable)

- If the variable was a abstract string column that did measured in a scalar fashion, we would capture this by converting the column to a factor (see Section 3.2.2) then into a continuous variable. We would loose the string values associated with this column (however they can be inferred since they are en-coded in a systematic manor)

3.3.2 Standardization

When working with multiple continuous variables, they will most likely all be measured on different scales. If we model with these variables without standardization, we will introduce bias into our predictions due to the differences in scale. Example: **VarA** is measured in the range 0-10, **VarB** is measured in the range 0-10000. A 1 unit increase in **VarA** would be considered more impact than a 1 unit increase in **VarB** however our model won't be able to correctly model these impacts due to their difference in scale.

Another issue is that if our continuous variable is measured on a large scale (0-1000000), calculations on the variable may results in incredibly large values. This isn't desired since large numbers are computationally hard to solve and the memory within the coding language R can only handle integer values up to 2147483647, formally known as the *integer.max* within R.

To solve this issue, we standardize (create the Z-scores) our continuous variables so that it they have a mean of 0 and standard deviation of 1. This puts all of our continuous variables on the same scale ready for modelling.

$$X = \frac{X - E(X)}{\sqrt{Var(X)}} \tag{2}$$

This is Standardization equation at the population level

- X: The X variable being standardized

- $E(X)$: The mean of the X variable (The expected value)

- $Var(X)$: The variance of the X variable

$$x = \frac{x - \bar{x}}{\sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2}}$$

This is Standardization equation at the sample level

- x: The x observation being standardized

- $\bar{x}$: The mean of the X variable

- n: The number of observations

- x_j: The jth observation of x (used to calculate the standard deviation per ith observation)

3.3.3 Dummy Variables

To correctly model our factor/categorical variables, we need to create dummy variables from them. This is achieved by taking a factor of n levels and producing n-1 dummy variables to represent it. Given this transformation, our resulting data-set for analysis will have more variables than the data-set inputted.

Table 16 is an arbitrary data set of 10 observations which we will turn into dummy variables. These variables are:

- **VarA**: a factor variable, its levels are yes & no
- **VarB**: a continuous variable
- **VarC**: a factor variable, its levels are 1,2 & 3

Original Data			Transformed Data			
VarA	VarB	VarC	VarA.yes	VarB	VarC.2	VarC.3
yes	100	1	1	100	0	0
no	101	2	0	101	1	0
no	102	1	0	102	0	0
yes	103	3	1	103	0	1
no	104	3	0	104	0	1
yes	105	2	1	105	1	0
no	106	2	0	106	1	0
no	107	1	0	107	0	0
no	108	3	0	108	0	1
yes	109	2	1	109	1	0

Table 16: Dummy Variables Example

3.3.4 Principal Component Analysis (PCA)

Optional to the user, a PCA transformation on the data-set can be preformed. This method takes our data-set and rotates it into new dimensions/axis which describes the variance the data-set in parts. This is done by calculating the co-variance matrix of the data-set, finding the eigenvalues & eigenvectors of said matrix and then rotating/transforming our data using the eigenvectors as the directions for our new axes. This then creates a data-set of Principal Components which are ordered left to right by the amount of variance explained [16]. Below we will go over the mathematics behind this transformation [17].

$$\begin{pmatrix} Var(X) & Cov(Y,X) & Cov(Z,X) \\ Cov(X,Y) & Var(Y) & Cov(Z,Y) \\ Cov(X,Z) & Cov(Y,Z) & Var(Z) \end{pmatrix} = Q\Lambda Q^{-1} \tag{3}$$

- X, Y, Z: Arbitrary predictor variables
- Q: Matrix of eigenvectors, follows the design:

$$\begin{pmatrix} q_{11} & q_{21} & \cdots & q_{n1} \\ q_{12} & q_{22} & \cdots & q_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ q_{1n} & q_{2n} & \cdots & q_{nn} \end{pmatrix} \tag{4}$$

The matrix is square with dimensions $n \times n$ since our original matrix, the co-variance matrix, is also square with dimensions $n \times n$. Every column of Q represents one eigenvector which describes the direction of the corresponding principal axis. These eigenvectors are ordered left to right in order of amount of variance explained

- Λ: Matrix of eigenvalues, follows the design:

$$\begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix} \tag{5}$$

  The matrix is square with dimensions $n \times n$ since our original matrix, the co-variance matrix, is also square with dimensions $n \times n$. Every diagonal value is the amount of variance explained by each principal component (PC). These eigenvalues are ordered left to right in order of amount of variance explained.

Dimension reduction is done using the eigenvalues/cumulative variance to remove the PC's. we keep all PC's before & including the PC where the cumulative sum of eigenvalues (cumulative variance) goes over **pcPercent** (default is 95% or 0.95) for the first time.

As an example, below is an arbitrary data set where we want to reduce our data set to only describe 95% of the variance in the data. Below is the matrix of eigenvalues for this example

$$\begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}$$

We know that the sum of the variance described by the first 2 PCs (the first two eigenvalues λ_1, λ_2) is ¿= 95%, this can be represented like so

$$let : \frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} >= 95\%$$

Therefore, to reduce the amount of dimensions we have in our data whilst still describing 95% of the variance in the data, we remove the third PC (the third eigenvalue λ_3). This creates the following matrix as our new reduced matrix of eigenvalues

$$\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

3.3.5 Train Test Split

To allow for predictions, we need to split our data into training and testing sets. The size of the testing set is defined by the function variable **testPercent**, default value of 25% or 0.25. Since it's possible for a Train Test Split to make a variable have 0 variance in the training set, this processes is preformed in a loop until it's confirmed that the training set has non 0 variance for all variables. If a split cannot be found, the following error message is thrown:

> *No good Train Test split was found, please try increasing 'obsPerLevel' or changing 'randomSeed' if it was set and try again.*

If the user so desires, they can set **testPercent** to -1 so that the split only chooses one observation for the testing set (useful if you already have a testing set, seen used for the Titanic data in Section 5.4).

Table 17 is an example using an arbitrary data set of 12 observations using the default value of **testPercent** (which is 0.25). For the example given, notice that all the variables in the Training Set (**VarA**, **VarB** & **VarC**) maintain a non-0 variance, this feature is required for a Train Test Split to be considered.

Original Data		
VarA	VarB	VarC
40	100	1
50	101	2
50	102	1
40	103	3
50	104	3
40	105	2
50	106	2
50	107	1
50	108	3
40	109	2
40	110	1
50	111	3

Training Set		
VarA	VarB	VarC
40	100	1
50	102	1
50	104	3
40	105	2
50	106	2
50	108	3
40	109	2
40	110	1
50	111	3

Testing Set		
VarA	VarB	VarC
50	101	2
40	103	3
50	107	1

Table 17: Train Test Split Example

3.4 Modelling

As our final output of the **automodel**, we generate various different models of our data-set. These models are a mixture of classification and regression techniques to provide a varied look at what models best describe the data-set.

3.4.1 K-Means

This is a clustering algorithm which aims to cluster similar observations into the same clusters [7]. In K-Means, we are aiming to minimize the sum of the sum of the squared difference between observations and the centroid of a cluster. To break this down, the sum of the squared difference between observations and the centroid of a cluster is mathematically similar to the sum of the squared residuals (SSR) from Ordinary Linear Regression (see Section 3.4.4 equation 10). We are aiming to minimise the sum of the SSR per cluster to produce our K clusters. To evaluate this model, we use within and between MSE and exact prediction accuracy (more on this in Section 3.4.6). Given this is a classification model, the MSE produced should not be compared to other models within **automodel** expect other K-Means models.

$$\underset{s}{argmin} \sum_{i=1}^{k} \sum_{x \in S_i} |x - \mu_i|^2 \tag{6}$$

- $\underset{s}{argmin}$: This is telling us that we are aiming to minimize $||x - \mu_i||^2$ for each cluster, S
- k: The amount of clusters
- x: an observation in the current cluster, S_i
- S_i: The respective cluster (the ith cluster)
- μ_i: the mean of the points in the ith cluster (the centroid of cluster S_i)

In this dissertation, we use this classification method in an unorthodox way were we use the clusters to predict a variable. This is generally not the desired use for K-Means and the model should be instead evaluated using inside and between MSE. We would expect this model to produce a lower predictive power than the regression models generated using the **automodel** function.

3.4.2 kNN

kNN, or K Nearest neighbors, is a clustering algorithm which attempts to predict a data-point based on the K closest observations. To evaluate this model, we use MSE and exact prediction accuracy (more on this in Section 3.4.6). Given this is a classification model, the MSE produced should not be compared to other models within **automodel** expect other kNN models. The algorithm step by step is as follows:

1. a Train Test split is preformed on a data-set so that we have two sets of observations of the same data-set

2. We then decide how many neighbors should be considered (the value of k) in our kNN model

3. We then take the first observation from the testing data and calculate the Euclidean distance between this observation and all observation in the training data

4. We then order the resulting Euclidean distances from smallest to largest and take the top k training observation as the neighbors to be considered.

5. Among these k nearest neighbors, we then take the most popular/frequent value and predict our test observation to be the same

6. we repeat the steps 3 to 5 for each observation in the testing data-set.

The general equation for Euclidean Distance in the context in kNN is as follows [18]:

$$\sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \tag{7}$$

- n: The number of dimensions each observation has
- q_i: The ith dimension within the training observation, q
- p_i: The ith dimension within the testing observation, p

The help explain further, please see Figure 7 [19]. In Figure 7 it shows how a test observation (see the yellow square) would be classified given different values for k using the kNN algorithm. When $k = 3$, our test observation would be classified as class B (a green triangle). When $k = 7$, our test observation would be classified as class A (a red star).

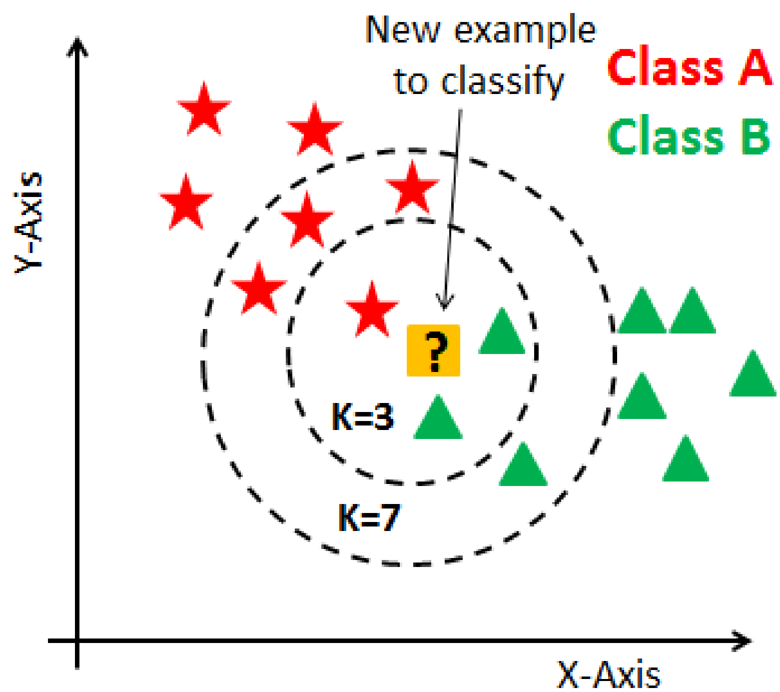


Figure 7: kNN at different values of k

3.4.3 Classification and Regression Trees (CART)

CART models work by creating thresholds that splits the data into groups per observations per variable. A certain number of observations are needed before we can decide to make a threshold at any stage in the tree which can be controlled by the function variable **cartSplit**, the default is 20. In **automodel**, CART is used to preform our missing value imputation (See Section 3.2.4) and to create a prediction model. It generates a tree which can predict the dependant variable and gives a list of variables which were the most important in predicting. We evaluate our CART model using MSE and exact prediction accuracy. The MSE created can be compared to the other regression models within **automodel** but not to the classification models.

Here's a explanation of what happens at each stage during the CART process [20] [21]:

1. a dependant variable, Y, and the predictor variables, X, are decided
2. starting from the first predictor variable, X_1, we create a threshold using the first observation, x_1. An example threshold at this stage would be if the ith observation x_i is greater than x_1.
3. We calculate the mean value of our dependant variable, Y, in the groups that are made due to the threshold.
4. Using our calculated mean values for Y in each group, we calculate the Sum of the Squared Residuals (SSR) per group using the mean value as the predicted value.
5. We then take the SSR of each group and sum them together to create a final score for the current threshold.
6. We repeat steps 2 to 5 going though each observation of each predictor variable until we have calculated a score for each possible threshold split of the data across all predictor variables.
7. We then pick the smallest score and set the associated threshold as a node on the tree. This node will then predict the mean values found in step 3.
8. Then, if the group created by the threshold has **cartSplit** or more observations, we repeat steps 2 to 7 again, treating all the data-points in one group independently from all other groups.
9. Once no more groups can be split, we have made our CART model which can be used to predict the dependant variable, Y

The general equation to calculate a groups mean in the context of CART is as follows:

$$\bar{y_i} = \frac{1}{n_i} \sum_{j=1}^{n_1} y_j \tag{8}$$

- $\bar{y_i}$: The mean value of Y for group i (one of the group created by the current threshold)
- n_i: The number of observations that have fallen into group i
- y_j: the jth observation of the dependant variable in group i.

The general equation to calculate a groups SSR in the context of CART is as follows:

$$\sum_{j=1}^{n_i} (y_j - \bar{y_i})^2 \tag{9}$$

- n_i: The number of observations that have fallen into group i due to the threshold used.
- y_j: the jth observation of the dependant variable in group i.
- $\bar{y_i}$: The mean value of Y for group i (one of the group created by the current threshold)

The help explain further, please see Figure 8 [20] .In Figure 8, we only have 1 predictor variable, *Drug Dosage*, and our dependant variable is *Drug Effectiveness*. We only split a node only if they have 7 or more observations (equivalent to setting **cartSplit** to 7). The result is a CART model with 3 nodes that predict 1 of 4 different values for **Drug Effectiveness** based on the value of **Drug Dosage**.

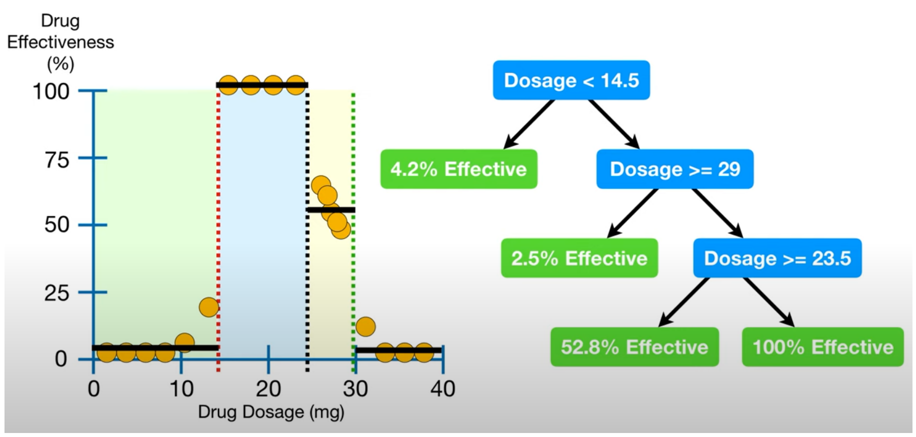


Figure 8: Simple CART model from Josh Starmer's *Regression Trees Clearly Explained*

3.4.4 Ordinary Linear Regression (OLR)

We use the base standard Ordinary Linear Regression (OLR) model as provided by R using the function **lm()**. Our model is fitted by minimizing the Sum of the Squared Residuals (SSR), which in layman's terms, is reducing the total distance between the observations and the fitted model. We evaluate our Ordinary Linear Regression model using MSE and exact prediction accuracy. The MSE created can be compared to the other regression models within **automodel** but not to the classification models.

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 ... + \beta_n x_n + \epsilon \tag{10}$$

- $\hat{y}$: The predicted value of the dependant variable
- β_i: The coefficient of the ith variable
- n: the amount of independent variables in the final model

$$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{11}$$

- SSR: This is the short hand for the Sum of Squared Residuals, what we minimize to fit the linear model
- y_i: The ith value of the dependant variable (y-variable)
- $\hat{y}$: The ith predicted value of the dependant variable (y-variable)

Starting from the full model, we use Variance Inflation Factor (VIF) to remove any leftover multicollinearity from our model. VIF is calculated per predictor variable and gives a measurement of how likely it is for the current predictor variable to be predicted by all other predictor variables. Variables are removed from a model in order largest to smallest VIF score. We stop removing predictor variables once all VIF scores are below the function variable **vifSelectionLevel**, which by default is 10.

$$VIF = \frac{1}{1 - R_n^2} \tag{12}$$

- R_n^2: This is the R^2 score from the regression of X_n on all other predictor variables. For example, given we have a Ordinary Linear Regression model fit (See Section 3.4.4 Ordinary Linear Regression, Equation 10), R_1^2 is calculated from the following equation:

  $$\hat{\beta_1} = \beta_0 + \beta_2 x_2 + \beta_3 x_3 ... + \beta_n x_n + \epsilon \tag{13}$$

  The equation used to calculated R_1^2 would then be as follows:

  $$R_1^2 = \frac{\sum(\beta_{1i} - \hat{\beta}_{1i})^2}{\beta_{1i} - \bar{\beta}_{1i})^2} \tag{14}$$

After preforming our VIF checks, we do a backwards selection on the remaining model variables based on their individual significance. The level of significance required is determined by the **automodel** variable **modelSigLevel**, which by default is set to 0.95 (meaning a 95% confidence interval). A predictor variable is kept in the final Ordinary Linear Regression model if they are considered significant at the **modelSigLevel** confidence level. To calculate a variables significance, we first divide the variables estimate coefficient by it's standard error to give us a t-value. Using this t-value, when then preform a test to determine if the absolute value of the t-distribution is greater than the absolute value of our t-value. This test then generates a p-value where the smaller the value, the more significant the predictor variable (the smaller the p-value, the more likely it is that the absolute value of our calculated t-value is greater than the absolute value of the t-distribution). We are selecting of variables based on if their p-value is less than or equal to 1 - **modelSigLevel**. [22]

3.4.5 Elastic Net Regression (ENR)

The final model within **automodel**, this creates a regression model based on minimizing the SSR of the model plus a penalty function. The penalty function used is a mixture of the penalty functions from Lasso and Ridge Regression. We evaluate our Elastic Net Regression (ENR) model using MSE and exact prediction accuracy. The MSE created can be compared to the other regression models within **automodel** but not to the classification models.

To get a proper understanding of the differences between Ordinary Linear, Lasso, Ridge and Elastic Net Regression, we will go over what each model is attempting to minimize/ The cost function of each model.

1. **Ordinary Linear Regression**

$$SSR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{15}$$

- SSR: This is the short hand for the Sum of Squared Residuals, what we minimize in Ordinary Linear Regression (see Section 3.4.4 equation 10)
- n: The number of observations
- y_i: The ith value of the dependant variable (y-variable)
- $\hat{y}$: The ith predicted value of the dependant variable (y-variable)

2. **Lasso Regression**

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{k} |\beta_j| \tag{16}$$

- $\sum_{i=1}^{n}(y_i - \hat{y}_i)$: Notice this is SSR, as seen in Ordinary Linear Regression
- n: The number of observations
- y_i: The ith value of the dependant variable (y-variable)
- $\hat{y}$: The ith predicted value of the dependant variable (y-variable)
- λ: A tuning parameter that determines the severity of the effect the slope has on fitting the model. It can take a value from 0 to $+\infty$. The value for this parameter is chosen though cross-validation in this paper.
- k: The number of predictor variables
- β_j: the coefficient of the jth predictor variable within the regression model. [23] [24]

3. **Ridge Regression**

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) + \lambda \sum_{j=1}^{k} \beta_j^2 \tag{17}$$

- $\sum_{i=1}^{n}(y_i - \hat{y}_i)$: Notice this is SSR, as seen in Ordinary Linear Regression
- n: The number of observations
- y_i: The ith value of the dependant variable (y-variable)
- $\hat{y}$: The ith predicted value of the dependant variable (y-variable)
- λ: A tuning parameter that determines the severity of the effect the slope has on fitting the model. It can take a value from 0 to $+\infty$. The value for this parameter is chosen though cross-validation in this dissertation.
- k: The number of predictor variables
- β_j: the coefficient of the jth predictor variable within the regression model. [23] [25]

4. **Elastic Net Regression**

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) + \lambda[\alpha\sum_{j=1}^{k}|\beta_j| + (1-\alpha)\sum_{j=1}^{k}\beta_j^2] \tag{18}$$

- $\sum_{i=1}^{n}(y_i - \hat{y}_i)$: Notice this is SSR, as seen in Ordinary Linear Regression
- n: The number of observations
- y_i: The ith value of the dependant variable (y-variable)
- $\hat{y}$: The ith predicted value of the dependant variable (y-variable)
- λ: A tuning parameter that determines the severity of the effect the slope has on fitting the model. It can take a value from 0 to $+\infty$. The value for this parameter is chosen though Cross Validation in this dissertation (explained further in depth within this section). We choose our λ value by picking the lowest value found during Cross Validation.
- α: a tuning parameter that determines the weight of the lasso and ridge penalties. It can take a value between 0 and 1, when $\alpha = 0$, it's identical to Ridge Regression, when $\alpha = 1$, it's identical to Lasso Regression.
- k: The number of predictor variables
- β_j: the coefficient of the jth predictor variable within the regression model. [23] [26]

When running **automodel**, we fit **elasticCount** plus 1 models each with a different value for α. The values for *alpha* are decided by incrementing from 0 in steps of size $\frac{1}{\textbf{elasticCount}}$. This would means that when **elasticCount** is 10, we fit models with the alpha value 0, 0.1, 0.2, 0.3 ... 1. The user will be able to see the MSE and R^2 value for each model but not the exact prediction accuracy. We then pick to model with the best MSE as our model of best fit and calculate an exact prediction accuracy for this model.

We use cross validation (10 fold cross validation to be precise) is used in Elastic Net Regression to choose the best λ per fit [27].

Below is an example 10 fold cross validation being preformed to choose the best value for the arbitrary parameter λ. Our error which we are looking to minimize is taken from Ridge Regression (a shortened version is used to simply the example)

$$SSR + \lambda\hat{\beta}^2 \tag{19}$$

- SSR: Represents the Sum of the Squared Residuals
- λ: Value to be determined in cross validation

- $\hat{\beta}^2$: The matrix of coefficients describing the fit of the model

In this example, as lambda increases, a larger emphasis is put on reducing the model's *slope* squared (represented as $\hat{\beta}^2$) over reducing the Sum of the Squared Residuals (SSR). This means that as λ varies, SSR and $\hat{\beta}$ changes. Since SSR and $\hat{\beta}$ change, the cost function (the equation used in this example) and λ aren't directly proportional. Then we pick our λ by a criterion of the users choosing (like picking the λ with the smallest value)

3.4.6 Model Evaluation Metrics

To evaluate our models, we calculate various statistics on the results of our predictions. Below we will list though these stats that are used.

Firstly, we use R^2 to evaluate our Ordinary Linear and Elastic Net Regression models.R^2 is a measurement of the of how much of the variance in y is explained by the model/predictor variables. In **automodel**, it's used within Ordinary Linear Regression, VIF and Elastic Net Regression. Below is the equation used to calculate R^2 [28]

$$R^2 = \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{20}$$

- y_i: The ith value of the y variable
- $\hat{y}_i$): the predicted value for y_i
- $\bar{y}$: The mean value of the y variable

Secondly, we use Mean Squared Error (MSE) to evaluate all of our prediction models. MSE is the Mean Squared Error and depending on the model, will be calculated differently. Example: the MSE used in K-Means is a calculation on the Euclidean Distance between a observation and the centroid of the cluster it's apart of. The MSE calculated in Ordinary Linear Regression is the mean of the squared error between the real and predicted values of the dependant variable. Since the method to calculate MSE is different in these two models, they shouldn't be used to compare the goodness of fit for each model. Below is the general equation used to calculate MSE for a regression model.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{21}$$

This is the general equation for the MSE in a regression model.

- n: The number of observations
- y_i: The ith value of the y variable
- $\hat{y}_i$): the predicted value for y_i

Thirdly, we use Akaike information criterion (AIC) to evaluate our Ordinary Linear Regression (OLR) model. AIC is a measurement used to compare different models of the data to determine which model was the best fit. The magnitude of AIC alone doesn't provide the user with much information, its the differences in AIC between two models trained on the same data that is useful [29].

$$AIC = 2k - 2ln(\hat{L}) \tag{22}$$

- k: The number of predictor variables in the model
- $ln()$: This is the Natural logarithm, $log_{\exp}$
- $\hat{L}$: The maximum value of the likelihood function for the model

Lastly, we calculate a prediction accuracy to evaluate all of our prediction models. We calculate the accuracy within a few different contexts/ranges:

- **Exact Accuracy**: An exact prediction accuracy is calculated by counting the number of times the model correctly predicted the dependant variables and diving it by the total number of observations. This method of evaluation is only useful when the dependant variable is categorical/ has a relatively small measurement scale.
- **Confidence Interval Accuracy**: For only the Ordinary Linear Regression Model, we predict within a 95% confidence interval. This gives us a little more information for the Ordinary Linear Regression model for us to make evaluations with. It's especially useful if our dependant variable is continuous and therefore accounts for predictions where the error in prediction may be small but not 0. The equation for this interval is:

$$\hat{y} \pm t_{\alpha/2,n-2}\sqrt{MSE(\frac{1}{n} + \frac{(x_k - \bar{x})^2}{\Sigma(x_i - \bar{x}^2)})} \tag{23}$$

  – $\hat{y}$: The given predicted value
  – α: is the confidence level, which in our case is 0.05
  – $t_{\alpha/2,n-2}$: Is the respective value from the T table distribution
  – MSE: This is the Mean Squared Error of the model
  – n: The number of observations
  – x_k: The given predictor variables
  – $\bar{x}$: The mean of the predictor variables
  – x_i: Represents the i-th predictor variable

- **+-1 Range Accuracy**: Only used for the Understanding Society data, for some variables we are predicting (like the GHQ score/scghq1_dv) we want to see if our prediction of within a 1 unit difference of the real value. This is done by if a prediction is 1 one unit change up or down (+-) from the real value, we mark the prediction as correct.

- **GHQ Range Accuracy**: Only used for the Understanding Society data, this is a calculated interval based on the results found for the GHQ score (scghq1_dv) and the grouped version of the GHQ score (scghq2_dv). The calculation for the unit change for this interval is as follows:

$$\hat{y} \pm ((\frac{unique_{ghq1}}{unique_{ghq2}}) \div 2) \tag{24}$$

  - $\hat{y}$: The given predicted value
  - $unique_{ghq1}$: The amount of unique scores in **scghq1_dv**, found in R using

```
length(unique(scghq1\_dv))
```

  - $unique_{ghq2}$: The amount of unique scores in **scghq2_dv**, found in R using

```
length(unique(scghq2\_dv))
```

3.5 Run Order

When running **automodel**, we process all of our methods in a select order. This ordered list will describes to order of which this methods are executed. As seen below, Some function may be ran multiple times, this is done to ensure a smooth run.

3.2.1 Data-frame Checks

3.2.2 Data Type Checks

3.2.6 Pair-wise Deletion

3.2.3 High Missing Data Variables

3.2.4 Missing Value Imputation, if **impFlag** = TRUE (see Section 3.6.6)

3.2.5 List-wise Deletion

3.2.8 Observation to Variable Ratio

3.3.1 Factorization

3.2.7 Low Level Removal

3.2.8 Observation to Variable Ratio

3.2.9 Variance Checks

3.3.3 Dummy Variables

3.2.9 Variance Checks

3.4.1 K-Means

3.2.10 Correlation Checks

3.3.2 Standardization

3.3.4 Principal Component Analysis (PCA), if **PCAFlag** = TRUE (see Section 3.6.11)

3.3.5 Train Test Split

3.4.2 kNN

3.4.4 Ordinary Linear Regression (OLR)

3.4.5 Elastic Net Regression (ENR)

3.6 Function Variables

To allow for user customization, we have a list of different variables which can be set by the user to change the process of the analysis. For each variable we will explain it's initial state (default value), description, provide an example case of it being used and any associated issues. The variables appear in order of usage within the **automodel** function.

3.6.1 predictVar

- initial state: Needed from user

- description: The name of the variable that is to be predicted/the name of the dependant variable

- example case: a user wants to predict the variable **age** from a data-set, **predictVar** should be set to: "age"

- associated issues: If not specified, function will not run. Value inputted needs to be a string value or else the function won't run

3.6.2 data

- initial state: Needed from user
- description: The data-set of all the predictor variables and the dependant variable
- example case: a user from to predict a variable from the loaded in data-set **footballdata**, **data** should be set to: footballdata
- associated issues: If not specified, function will not run. The data parsed needs to be able to cast to a data.frame object in R, if not, the function will not run.

3.6.3 naPercent

- initial state: 0.2 (represents 20%)
- description: This is the percentage amount of NA values allowed in a predictor variable. If the predictor variable have **naPercent** percent or more of their values as NA/missing, the predictor variable is removed
- example case: Say we are happy to keep columns with a larger amount of missing data, we can set **naPercent** to 0.4 to keep variables that have less than 40% of their data as missing/NA
- associated issues: If this value is set to high, List-wise deletion may delete almost all observations and missing value imputation may impute misleading results

3.6.4 cartSplit

- initial state: 20
- description: The amount of observations needed for a new threshold/node to be generated in a CART model. CART is used to impute missing values and to generate a prediction model of the data within the code.
- example case: When using a data-set with a low amount of observations and predictor variables (say 100 observations of 3 variables), we may want to split the nodes when there are less observations. Therefore we would set **cartSplit** to a lower value like 5
- associated issues: If this value is greater than or equal to half the amount of observations in the data-set, CART will always be creating one node which can't be split, potentially leading to very poor predictions. If **cartSplit** is too low (like 2), we risk over-fitting the model to the training data

3.6.5 impFlag

- initial state: FALSE
- description: Flag to tell the function if we are wanting to impute the missing values instead of preforming a List-wise deletion. If TRUE, a single CART imputation is done to impute the missing values.
- example case: If we have a data-set with a relatively low amount of missing data where the missing data isn't heavily bias towards any outcome (MACR or MAR assumption), we can impute the missing values to use more observations in out modelling. By changing **impFlag** to TRUE, we are telling the automodel to preform a missing value imputation on the data-set (this happens after the removal of variables with high missing values, see **naPercent**)
- associated issues: Imputation can take a long time on a large data-set and therefore should be taken into consideration. If the missing values in the data doesn't uphold the MACR or MAR assumptions, missing value imputation would badly represent the missing values in the data.

3.6.6 randomSeed

- initial state: NULL
- description: To make results reproducible, we must set a random seed within the function. **randomSeed** can be set to any integer in the range -2147483647 to 2147483647 (2147483647 is the maximum integer allowed in R).
- example case: To produce results for a report on a data-set, we would want to make sure that a observer can reproduce the results found. To ensure this, we would set **randomSeed** to any integer value, such as 1, each time we want to run the automodel function.
- associated issues: Due to the amount of models that need to have their random effect controlled, it's not possible to use **set.seed()** before calling **automodel**. Sometimes it can be quite hard to find a seed that has a Train Test split that processes correctly, to confirm a working seed exists, try running the function without setting **randomSeed** first.

3.6.7 catLevels

- initial state: 15

- description: Decides on how many unique values/levels a variable needs to be considered as a factor/categorical variable when modelling. If a variable has less than or equal to **catLevels** levels, the variable gets encoded as a factor.

- example case: Say we know that all of our categorical variables have less than 5 unique levels, we could change **catLevels** to 5 so that we ensure that no variables that were intended to be continuous get encoded as a factor (since if a continuous variable only has less than or equal to **catLevels** unique values/levels, it will be encoded as a factor).

- associated issues: If **catLevels** is set to low (for example, 2), incorrect modelling of variables may occur resulting in poor predictions. If set to high (the number of observations in the data-set), we risk over-fitting our model to the exact observations seen in continuous variables, causing poor predictions

3.6.8 obsPerLevel

- initial state: 5

- description: How many observations are needed of a level in a categorical/factor variable. If the level in the variable has less than **obsPerLevel** observations, all observations of this level are removed from the data-set

- example case: If our data-set is known to have 2 observations of a case in a categorical variable and all other levels have 3+ observations which we are happy to keep. We can set **obsPerLevel** to 3 so that this case with 2 observations is removed.

- associated issues: There are quite a few issues involved with the tweaking of this variable. Setting **obsPerLevel** too low may lead to the function not finding a valuable test, train split which maintains variance in all variables in the training data. If we have a known low amount of observations in categorical variables but wish to keep them, we can set the variable **testPercent** to -1 so that we only take one observation as our test set. This risks over-fitting and should be done at the digression of the user. Setting **testPercent** to high risks deleting large amounts of observations within the data-set, therefore making the resulting data-set unusable for modelling. Sometimes an error message from the automodel function will tell the user to modify this variable, it should be known that if this is the case, this variable should be incremented in small steps to find a workable solution.

3.6.9 clusterAmount

- initial state: determined by a logical statement

  ```
  ifelse(length(unique(data[[predictVar]])) <= catLevels,
  length(unique(data[[predictVar]])),
  catLevels)
  ```

  – `predictVar`

    The dependant variable which the user wishes to predict (function variable, **predictVar**)

  – `length(unique(data[[predictVar]]))`

    The number of unique values within the dependant variable

  – `catLevels`

    The number of unique values within the dependant variable (function variable, **catLevels**)

  The logical statement sets the number of clusters to the amount of unique values/levels in the dependant variable unless this amount is larger than **catLevels**, which if true, then sets the number of clusters to **catLevels**.

- description: Sets the amount of clusters/centroids to be used in K-Means modelling. Our K-Means model has **clusterAmount** clusters.

- example case: If we know that our dependant variable is binary (dependant variables can only take two values), we can set **clusterAmount** to two.

- associated issues: Setting **clusterAmount** to high will risk over-fitting our clustering model to the current observations (risk of making uninformative clusters). Predicting a continuous dependant variable using clustering doesn't garner good results.

3.6.10 corrConfLevel

- initial state: 0.8 (represents 80%)

- description: The cutoff point of when a correlation between two predictor variables is deemed to large/ will cause multicollinearity in a model. If the absolute value of the correlation coefficient between two predictor variables is greater than **corrConfLevel**, these variables will be considered for deletion.

- example case: If we are ok to have correlated variables in our data-set (we are more interested in the CART and clustering models than the regression models), we can set **corrConfLevel** to 0.99 so that all expect extremely similar predictor variables (basically predictor variables that are the same) are considered for deletion

- associated issues: If **corrConfLevel** is set to high (like 0.99), the results from our regression models will be bad and shouldn't be considered accurate (too much multicollinearity in the predictor variables, unstable model).

3.6.11 PCAFlag

- initial state: FALSE

- description: Flag that tells the **automodel** function to preform PCA on the data-set. If TRUE, PCA using the co-variance matrix and eigenvalue decomposition is preformed.

- example case: If we aren't worried about seeing what variables are significant towards predicting our predictor variable and instead we want to fit the most accurate model with the smallest amount of variables, we would want to preform PCA. To do this, we set **PCAFlag** to TRUE.

- associated issues: PCA will make the regression models prediction using principal components instead of predictor variables therefore making these models harder to interpret (performance/goodness statistics are still relevant though, like R^2.)

3.6.12 pcPercent

- initial state: 0.95 (represents 95%)
- description: This sets the needed amount of variance explained by the reduced principal component (PC) data-set. When preforming dimension reduction, we keep the smallest amount of PC's that describe**pcPercent** of the variance in the data-set.
- example case: We want to reduce the amount of variables in our data-set whilst still explaining 90% of it's variance. We initially set **PCAFlag** to TRUE to tell automodel that we want to preform PCA, then we set **pcPercent** to 0.9.
- associated issues: it's worth considering that setting **pcPercent** low may result in using one PC, this is fine based on the data used and the value **pcPercent** was set to.

3.6.13 testPercent

- initial state: 0.25 (represents 25%)
- description: The proportion of observations after transformation within the testing data. When preforming our Train Test split on our data, we set **testPercent** percent of our observations as our testing data and $1 -$ **testPercent** as our training data. Our training data must maintain variance in all predictor variables and therefore the splitting process iterates until a correct split can be found. If **randomSeed** is set, it only evaluates the Train Test split associated with the seed once. A user can set **testPercent** to -1 if only one observation should be tested.
- example case: We want to have a Train Test split where 10% of our observations after transformation are in the training data. To achieve this we set **testPercent** to 0.1.
- associated issues: Since we are looking for a split where the Training data maintains variance for all variables, data-sets with low observations per categorical variable levels will struggle to maintain this variance in the training data. This is unavoidable since if we don't maintain variance in the training data, those predictor variables with no variance are practically useless when modelling (since they have the same value for all observations). If a Train Test split can't be found because of this reason, it's recommended that the automodel variable **obsPerLevel** should be increased or missing value imputation is preformed on the data-set (done by setting **impFlag** to TRUE). For these reasons, it would also be hard to find a working random seed value (setting **randomSeed**) since the amount of splits a data-set of this nature could do would be limited.

3.6.14 kNNCount

- initial state: determined by the calculation

  ```
  round(sqrt(nrow(data)),0)
  ```

  This is the same as doing $\sqrt{n}$. n is the number of observations in the data-set and the answer is rounded to 0 decimal places.

- description: The number of neighbors that are considered for each test observation in the kNN model. Each test observation considers **kNNCount** neighbors to predict the dependant variable for the observation.

- example case: We want to consider the 10 closest points (the neighbors) to predict each observation in the kNN model

- associated issues: the **kNNCount** has to be an integer value between 1 and the number of observations in the transformed data-set. Setting this value to large (20% or more of the number of observations) will cause all predictions to converge to the same prediction, making the model inaccurate (will always predict the same values).

3.6.15 vifSelectionLevel

- initial state: 10

- description: We will be removing variables from the Ordinary Linear Regression model in order of their VIF score until all variables have a VIF score equal to or below **vifSelectionLevel**.

- example case: Given we have a good understanding of the co-linearity between the predictor variables and our data has many predictor variables (which naturally increases R^2), we want to increase the VIF threshold to 20. To do this, we change the value of **vifSelectionLevel** to 20

- associated issues: Setting **vifSelectionLevel** to a large value (such as 100) will create a model with high multicollinearity which generate poor/unstable predictions of our dependant variable.

3.6.16 modelSigLevel

- initial state: 0.95 (stands for 95%)
- description: To refine our Ordinary Linear Regression Model we remove variables based on their statistical significance, we use **modelSigLevel** to decide the minimum confidence level that each variable in the model needs to satisfy. Each model needs to be considered significant at the **modelSigLevel** confidence level.
- example case: The data-set we are analyzing contains a large amount of predictor variables (100+) and therefore we want to the model to only contain model that are highly significant/significant at the 99% confidence level. To do this, we set **modelSigLevel** to 0.99.
- associated issues: If this value is is set to 1, we would end up with a null model, if set to 0, we would end up with the full model after VIF checks.

3.6.17 elasticCount

- initial state: 10
- description: When fitting our elastic net models, we need to fit models for different values of α (which takes values between 0 and 1). **elasticCount** decides how many different values of alpha we try and the different values are incremented equally based on the value of **elasticCount**
- example case: Say we want to fit an Elastic Net Regression model for each 0.05 increment in α of value. To do this we would set the value of **elasticCount** to 20 (since $\frac{1}{20}$ is 0.05).
- associated issues: **elasticCount** needs to be an integer value large than 1. Setting **elasticCount** to a large integer could significantly increase the run-time of **automodel** based on the size of the data-set being processed

3.7 Assisting Custom Functions

To represent all of our data correctly, we need some custom functions to merge & analyze the data-sets to be used with **automodel**. All the functions mentioned here are for the Understanding Society (see Section 2.1) only due to the mass amount of data we are going to process from it. It's worth keeping in mind though that the idea of this dissertation was to keep most of the process in the **automodel**

3.7.1 recodeNA

This function is needed for the Understanding Society data. It takes all the different categories of missing values in the Understanding Society data and re-codes them to **NA**

```
recodeNA = function(data){
data[data == -9] = NA
data[data == -8] = NA
data[data == -7] = NA
data[data == -2] = NA
data[data == -1] = NA
return(data)
}
```

The above is finding all instances of -9, -8, -7, -2, -1 within **data** and changing them to **NA**. The returned Understanding Society data-set will have all missing entries in the data-set encoded as **NA**.

3.7.2 sqlTransform

To be able to properly analyze the income data of the participants alongside their survey data, we need to transform the income data into a form where the *pidp* is unique per row. The final SQL query is executed though the **sqldf** library.

```
sqlTransform = function(data){

#start of the SQL query
querySQL = "SELECT pidp,SUM(frmnthimp_dv) as frmnthimp_dv_total"

#for loop which turns all ficodes into binary columns
for (i in 1:max(data$ficode)){
     colName = paste("ficode", i,sep="")
     data[[colName]] = ifelse(data$ficode == i, 1, 0)
     querySQL = paste(querySQL,",SUM(",colName,") as ",colName,sep = "")
     }

#finish of the SQL query
querySQL = paste(querySQL,"FROM data GROUP BY pidp")

#SQL to turn the table into version where pidp is unique per row
return(sqldf(querySQL))
}.
```

In the above, **data** represents the respective income data passed though (**w2income** or **w3income**), **ficode** tells us what category of income the row represents & **frmnthimp_dv** is the amount of income for the respective **ficode**. In summary, what we are doing is creating a binary variable (1 or 0) for each **ficode** present therefore making our **pidp** unique per row.

To assist with explanation, please Table 18 and Table 19 below for a representation of the transformation:

Non-Transformed Income Data		
pidp	frmnthimp_dv	ficode
0001	800	1
0001	1200	2
0002	500	3
0002	700	4

Table 18: sqlTransform Example Original Data

Transformed Income Data					
pidp	frmnthimp_dv_total	ficode1	ficode2	ficode3	ficode4
0001	2000	1	1	0	0
0002	1200	0	0	1	1

Table 19: sqlTransform Example Transformed Data

The returned data-set is a transformed version of the respective income data which now has a unique *pidp* per row.

3.7.3 ghq_analyze

Before running our GHQ data though the main function, we do preliminary analysis on the GHQ data within each join of the Understanding Society data. This function includes different numerical and graphical summaries. The code for this function can be found in Section 10

To summarize what the function is doing in order of execution:

1. Numerical summaries of the *scghq1_dv* variable
2. Plots a histogram of *scghq1_dv*
3. Plots a QQ-plot of *scghq1_dv*
4. Plots the correlation matrix of all the GHQ questions

3.7.4 ghq_clean_move

Since we have some preliminary understanding of the Understanding Society and the GHQ data, we want to remove the individual GHQ questions/variables from the data. We also want to move all variables that end in *val* or *_dv* to the end of the data-set since we know these variables are derived/averages of other variables they would highly correlate with (and we prefer to keep these variables).

```
ghq_clean_move = function(data) {

#changing the name of the ghq total score and
#removing the questions from the data-set
names(data)[names(data) == names(select(data,contains("ghq1")))]
```

```
    = "total_score"
data = select(data,-contains("ghq"))

#re-ordering the columns for the correlation check so
#that we automatically keep all 'val' variables
for (i in 1:length(names(data))) {
    if (grepl("val",colnames(data)[i], fixed = T) |
        grepl("_dv",colnames(data)[i], fixed = T)) {
        data = data %>% relocate(colnames(data)[i], .after = last_col())
    }
}

#return the cleaned data-set
return(data)

}
```

To summarize what the function is doing in order of execution:

1. Changes the name of the variable *scghq1_dv* to *total_score*

2. re-orders the data parsed so that variables ending in *val* or *_dv* are at the end/rightmost within the data

3.7.5 predGHQadd

For our GHQ data, we have would like to see how the predictions for each model do within certain ranges. To achieve this we take the prediction results from all the models and calculate the accuracy of the prediction under a +-1 range and a calculated range based on the groupings found in the variable *scghq2_dv*. The code for this function can be found in Section 10.

To summarize what the function is doing in order of execution:

1. gathers the predictions for all of the models calculated by the **autoModel** function.

2. creates predictions for the K-Means and kNN model within a +-1 range

3. add +-1 and GHQ ranges to the predictions **autoModel** function calculated for CART, OLR & ENR.

4. Calculates the prediction accuracy within the given ranges for all of the models and adds returns the new prediction results.

4 Understanding Society Analysis

Here we will go though a run of the function where we create models to predict the GHQ score of a participant observed in the Understanding Society data. For this run we will be looking at different Understanding Society data-sets in different ways/joins. This will be the largest test of the automodel function in this dissertation since the Understanding Society data is rather large and we use multiple different joins between data-sets. To make this entire run reproducible, all models will have the function variable **randomSeed** set as an integer (see Section 3.6.6)

4.1 Data Breakdown

Firstly, we have to load in all the different data-sets as descried in Section 2 Understanding Society.

- **b_income**: Loaded in as **w2income**
- **b_indresp**: Loaded in as **w2indresp**
- **c_income**: Loaded in as **w3income**
- **c_indresp**: Loaded in as **w3indresp**
- **xindresp_ns**: Loaded in as **mixNurse**
- **xlabblood_ns**: Loaded in as **mixBloodData**

Then after some processing (this will be covered in Section 4.2) Table 20 contains all the different data-sets which we wish to analyze.

Data-set	No. Obs	No. Vars	GHQ Included?	Wave
w2indresp	54569	1653	TRUE	2
w2Merge	33364	1693	TRUE	2
w2MergeNurse	10921	2032	TRUE	2
w2MergeNurseBlood	6926	2065	TRUE	2
w3indresp	49692	3059	TRUE	3
w3Merge	30487	3099	TRUE	3
w3MergeNurse	3472	3438	TRUE	3
w3MergeNurseBlood	2336	3471	TRUE	3
wShared	104261	1187	TRUE	2, 3
wSMerge	63851	1227	TRUE	2, 3
wSMergeNurse	14393	1553	TRUE	2, 3
wSMergeNurseBlood	9262	1586	TRUE	2, 3
mixNurse	20699	355	TRUE	2, 3
mixNurseBlood	13247	388	TRUE	2, 3

Table 20: Analyzed Understanding Society data-sets

- **w2indresp**: The same as **b_indresp**, described in Section 2.1.1.

- **w2Merge**: This is a join between the transformed **w2income** (see Section 4.1) and **w2indresp** on the variable **pidp**.

- **w2MergeNurse**: This is a join between **w2Merge** (see previous) and **joinSurveyData** (see Section 4.1) on the variables **pidp** & **wave**.

- **w2MergeNurseBlood**: This is a join between **w2MergeNurse** (see previous) and **mixBloodData** on the variables **pidp** & **wave**.

- **w3indresp**: The same as **c_indresp**, described in Section 2.1.1.

- **w3Merge**: This is a join between the transformed **w3income** (see Section 4.1) and **w3indresp** on the variable **pidp**.

- **w3MergeNurse**: This is a join between **w3Merge** (see previous) and **joinSurveyData** on the variables **pidp** & **wave**.

- **w3MergeNurseBlood**: This is a join between **w3MergeNurse** (see previous) and **mixBloodData** on the variables **pidp** & **wave**.

- **wShared**: This is a union of the shared variables across **w2indresp** & **w3indresp**.

- **wSMerge**: This is a union of only the shared variables across **w2Merge** & **w3Merge**.

- **wSharedNurse**: This is a join between **wSMerge** (see previous) and **joinSurveyData** on the variables **pidp** & **wave**.

- **wSharedNurseBlood**: This is a join between **wSMergeNurse** (see previous) and **mixBloodData** on the variables **pidp** & **wave**.

- **mixNurse**: The same as **xindresp_ns**, described in Section 2.1.1.

- **mixNurseBlood**: This is a join between **mixNurse** and **mixNurseBlood** on the variables **pidp** & **wave**.

4.2 Data Processing

Before we can run the data though our function, there are quite a few steps we must take that are unique to the GHQ data.

Firstly, we need to deal with the encoding of missing data within the Understanding Society data (see Section 2.1.2). For this run, we are taking a blanket approach by encoding all values to NA. This is done using the function specified in Section 3.7.1.

Next, we have some small adjustments:

1. We create a version of the **mixNurse** which doesn't include any GHQ data. This is so when it comes to joining **mixNurse** to waves 2 & 3 data respectively we don't duplicate the GHQ data. This new data-set is called **joinSurveyData**

2. waves 2 & 3 data (**w2indresp** & **w3indresp**) start their variable names with **b_** and **c_** respectively. To ensure joins can be preformed correctly we must strip **b_** and **c_** from the start of all variable names in **w2indresp** & **w3indresp**.

3. To provide a base for the joins to occur, we must add a new **wave** variable to both **w2indresp** and **w3indresp** to highlight which wave the data is from. This is achieved though the following two lines of R:

```
w2indresp\$wave = 2
w3indresp\$wave = 3
```

Next, we need to transform the **w2income** & **w3income** data into a form which makes the **pidp** unique per row (**pidp** as a primary key). Using the R package **sqldf** we are able to run a SQL query which achieves this result. With a for loop, we are able to automate the process of generating the SQL query, the function used for this transformation was mentioned in Section 3.7.2.

Lastly, we need to gather all the shared variables across **w2indresp** & **w3indresp** and union them into one data-set which we will call **wShared**. This is done by running the following code in R:

```
w2Shared =
    w2indresp[,names(w2indresp)[names(w2indresp) %in% names(w3indresp)]]
w3Shared =
    w3indresp[,names(w2indresp)[names(w2indresp) %in% names(w3indresp)]]
wShared =
    rbind(w2Shared,w3Shared)
```

The same procedure above is then carried out on **w2Merge** & **w3Merge**, which replace **w2indresp** and **w3indresp** respectively, to create the data-set **wSMerge**

Our data is now ready to merge into all the different data-sets as specified in Section 4.1.

Before running all the joins though the function, we are do some initial analysis on the variables that make up the GHQ data we are interested in. To achieve this analysis per join efficiently, we will use the function ghq_analyze (see Section 3.7.3) to automate though all the joins. All of the results from this procedure can be found in Section 10

Lastly we will run the function ghq_clean_move on all the data-sets (see Section 3.7.4).

4.3 Evaluation Method

Thought-out the results gathered from Understand Society's data, we will be using some terms to explain different prediction intervals:

- **Exact**: This means the accuracy of the model when predicting the GHQ score exactly. This means that if the model predicted a score of 6, the model is correct only when the real GHQ score is 6.

- **+-1 Range**: Since a GHQ score of 6 is rather similar to a score of 7 or 8, it's worth us seeing how well our models predict within the +-1 Range. This means if the model predicts a GHQ score of 6, the model is correct when the real GHQ score is either 6,7 or 8

- **Conf. Interval**: This is a calculated 95% confidence (Mean) interval only available with the Ordinary Linear Regression model, please see Section 3.3.3.3. Model Evaluation Metrics. When the model predicts a GHQ score of 6, the model is correct when the prediction falls within a 95% confidence interval range of the value 6.

- **GHQ Range**: This is a calculated interval using the **scghq1_dv** and **scghq2_dv** variables, please see Section 3.3.3.3. Model Evaluation Metrics for more. When the model predicts a GHQ score of 6, the model is correct when the prediction values within +-1.42 of 6.

For each model in each data join, we will compare the accuracy of predictions with the following initial models all models have been calculated using the GHQ score from **w2indresp**. We only use **w2indresp** since tests preformed in Section 2 Understanding Society showed the data to be very similar for GHQ score and we are only looking for a rough benchmark to judge our models predictions accuracy by. Please see Table 21 for the benchmarks we will be using.

Model	Exact	+-1 Range
Random Choice (RC)	0.027	0.0796
Mode Choice (MC)	0.1089	0.3033

Table 21: Benchmark Predictions Accuracies

- **Random Choice**: This is the chance of getting the correct GHQ score by complete random.
  - **Exact Random Choice** is simply 1 in 37 (where 37 is the number of possible GHQ scores)
  - **+-1 Range Random Choice** is very close to 3 in 37 however we have to consider what happens when we randomly pick a score such as 0 or 36. It doesn't make sense to say *accept 0 if GHQ score is -1* since a GHQ score of -1 doesn't exist, same goes for *accept 36 if GHQ score is 37* since a GHQ score of 37 doesn't exist. Therefore when

considering all correct cases, we have to remove the bottom and top tail of the acceptable range. To explain further please see the below calculation.

$$\frac{3}{37} * \frac{37}{37} = \frac{111}{1369}$$

This calculation is taking the chance of scoring within a +-1 range and multiplying it by the number of levels there are in the GHQ score. This gives us the fraction $\frac{111}{1369}$ which tells us that there are 111 correct cases for the GHQ score as a whole. To remove our bottom and top tail from this consideration, we do the following calculation.

$$\frac{111}{1369} - \frac{2}{1369} = \frac{109}{1369}$$

This removes the bottom and top tail cases (which were *accept 0 if GHQ score is -1* and *accept 36 if GHQ score is 37*). The leftover fraction, $\frac{109}{1369}$ is the correct amount of cases we should consider for the GHQ score and therefore is used as out benchmark.

- **Mode Choice**: This is the chance of getting the correct GHQ score if you always selected the mode.
  - **Exact Mode Choice**: The mode of our GHQ score is 6 with 4728 observations. Our prediction accuracy is then calculated by taking the number of observations of our mode and dividing by the total number of observations: $\frac{4728}{43423}$
  - **+-1 Range Mode Choice**: The mode +-1 Range is (6,7,8) with total observations of 13169. Our prediction accuracy is then calculated like so $\frac{13169}{43423}$

Applying the GHQ range to the Random Choice (RC) and Mode Choice (MC) models makes little sense since the GHQ score only takes integer values however the predictions from regression models predict in a continuous space (which gives us reason to check the GHQ range).

To then gather our prediction accuracies for the *+-1 range* for all models and the *GHQ range* for regression models we run each join though the function predGHQadd (see Section 3.7.5). This then gives us all of the accuracies needed to compare our models.

4.4 Run Results

4.4.1 Introduction

The following is what we expect to see in the console for each join when running though the **autoModel** function:

- Console output telling us the changes that are made to the data for it to be cleaned
- K-Means model with prediction results
- kNN model with prediction results and training data used
- CART model with prediction results
- Ordinary Linear Regression model with prediction results
- Elastic Net Regression model with a list of fits attempted & prediction results (based on the best model)
- Console output of the model evaluation metrics calculated for each of the models above

Since we are working with lots of Understand Society data which are derived from the same data sources (See Section 4.1) we will only go over the results gathered for each prediction model together per data-set. Full runs of each data-set can be found in Section 10.

4.4.2 w2indresp

Our run of the data-set w2Merge run successfully in 2432 seconds (41 minutes) and produced all of the models correctly.

In Table 22 we have all of the prediction accuracies calculated per model. We can see that all the models had a better Exact prediction accuracy when compared to the MC and RC models, with Elastic Net Regression being the best with a 16% accuracy. When predicting in a +-1 range, the ENR model is best with a 44% accuracy and when predicting in a GHQ range the CART model is best with a 57% accuracy.

Model	Exact	Conf. Interval	+-1 range	GHQ range
ENR	0.1624	NA	0.4438	0.5541
OLR	0.1607	0.3452	0.4251	0.5326
kNN	0.1527	NA	0.3681	NA
CART	0.1464	NA	0.4136	0.5746
K-Means	0.1227	NA	0.3349	NA
Mode Choice (MC)	0.1089	NA	0.3033	NA
Random Choice (RC)	0.027	NA	0.0796	NA

Table 22: w2indresp Predictions

4.4.3 w2Merge

Our run of the data-set w2Merge run successfully 2385 seconds (40 minutes) in and produced all of the models correctly.

In Table 23 we have all of the prediction accuracies calculated per model. We can see that all the models had a better Exact prediction accuracy when compared to the MC and RC models, with Elastic Net Regression being the best with a 18% accuracy. When predicting in a +-1 range, the ENR model is best with a 47% accuracy and when predicting in a GHQ range the CART model is best with a 58% accuracy.

Model	Exact	Conf. Interval	+-1 range	GHQ range
ENR	0.1764	NA	0.4691	0.5706
kNN	0.1714	NA	0.3655	NA
OLR	0.1659	0.3809	0.4515	0.5441
K-Means	0.1298	NA	0.3422	NA
CART	0.1279	NA	0.4168	0.5843
Mode Choice (MC)	0.1089	NA	0.3033	NA
Random Choice (RC)	0.027	NA	0.0796	NA

Table 23: w2Merge Predictions

4.4.4 w2MergeNurse

Our first run of **automodel** using **w2MergeNurse** threw the us a Observation to Variable ratio error (See Section 3.2.8) since the calculated ratio was 2.31. To resolve this issue, we re-ran **automodel** with a **naPercent** value of 0.1. This run then gave use a different error related stating that there was too much multicollinearity in the model (see Section 3.4.4). To resolve this issue, we re-ran **automodel** with a **naPercent** value of 0.1 and a **corrConfLevel** = 0.5. This run managed to compile successfully in 544 seconds (9 minutes) and outputted all the desired models.

In Table 24 we have all of the prediction accuracies calculated per model for the run which managed to compile all of our models. We can see that all the models had a better exact prediction accuracy when compared to the MC and RC models, with Elastic Net Regression being the best with a 18% accuracy. When predicting in a +-1 range, the ENR model is best with a 48% accuracy and when predicting in a GHQ range the ENR model is best with a 59% accuracy.

Model	Exact	Conf. Interval	+-1 range	GHQ range
ENR	0.1808	NA	0.4777	0.5857
OLR	0.1714	0.4049	0.4601	0.5575
CART	0.1667	NA	0.4695	0.5305
kNN	0.1514	NA	0.3345	NA
K-Means	0.1385	NA	0.3563	NA
Mode Choice (MC)	0.1089	NA	0.3033	NA
Random Choice (RC)	0.027	NA	0.0796	NA

Table 24: w2MergeNurse Predictions

4.4.5 w2MergeNurseBlood

Our first run of **automodel** using **w2MergeNurseBlood** threw the us a Observation to Variable ratio error (See Section 3.2.8) since the calculated ratio was 1.36. To resolve this issue, we re-ran **automodel** with a **naPercent** value of 0.05. This run managed to compile successfully in 3217 seconds (54 minutes) and outputted all the desired models.

In Table 25 we have all of the prediction accuracies calculated per model for the run which managed to compile all of our models. We can see that all the models had a better exact prediction accuracy when compared to the MC and RC models, with Elastic Net Regression being the best with a 17% accuracy. When predicting in a +-1 range, the ENR model is best with a 46% accuracy and when predicting in a GHQ range the ENR model is best with a 56% accuracy (CART was very close to ENR).

Model	Exact	Conf. Interval	+-1 range	GHQ range
ENR	0.1697	NA	0.4627	0.5577
OLR	0.1652	0.4197	0.431	0.5396
kNN	0.1538	NA	0.3507	NA
K-Means	0.1349	NA	0.3551	NA
CART	0.1278	NA	0.3812	0.5554
Mode Choice (MC)	0.1089	NA	0.3033	NA
Random Choice (RC)	0.027	NA	0.0796	NA

Table 25: w2MergeNurseBlood Predictions

4.4.6 w3indresp

Our first run of **automodel** using **w3indresp** threw the us a VIF error (See Section 3.4.4) since their was to much multicollinearity in the data. To resolve this issue, we re-ran **automodel** with a **corrConfLevel** value of 0.5. This run managed to compile successfully in 1542 seconds (26 minutes) and outputted all the desired models.

In Table 26 we have all of the prediction accuracies calculated per model for the run which managed to compile all of our models. We can see that all the models had a better exact prediction accuracy when compared to the MC and RC models, with Elastic Net Regression being the best with a 18% accuracy. When predicting in a +-1 range, the ENR model is best with a 45% accuracy and when predicting in a GHQ range the CART model is best with a 55% accuracy.

Model	Exact	Conf. Interval	+-1 range	GHQ range
ENR	0.178	NA	0.4482	0.5397
OLR	0.1716	0.361	0.4317	0.5261
kNN	0.1344	NA	0.3117	NA
CART	0.1265	NA	0.3796	0.5482
K-Means	0.1249	NA	0.3315	NA
Mode Choice (MC)	0.1089	NA	0.3033	NA
Random Choice (RC)	0.027	NA	0.0796	NA

Table 26: w3indresp Predictions

4.4.7 w3Merge

Our first run of **automodel** using **w3Merge** threw the us a VIF error (See Section 3.4.4) since their was to much multicollinearity in the data. To resolve this issue, we re-ran **automodel** with a **corrConfLevel** value of 0.5. This run managed to compile successfully in 1381 seconds (23 minutes) and outputted all the desired models.

In Table 27 we have all of the prediction accuracies calculated per model for the run which managed to compile all of our models. We can see that all the models had a better exact prediction accuracy when compared to the MC and RC models, with Elastic Net Regression being the best with a 16% accuracy. When predicting in a +-1 range, the ENR model is best with a 46% accuracy and when predicting in a GHQ range the ENR model is best with a 57% accuracy (The CART model was close to ENR).

Model	Exact	Conf. Interval	+-1 range	GHQ range
ENR	0.1587	NA	0.46	0.5747
CART	0.1564	NA	0.4079	0.5736
OLR	0.1518	0.4508	0.4206	0.5226
K-Means	0.133	NA	0.3413	NA
kNN	0.1228	NA	0.314	NA
Mode Choice (MC)	0.1089	NA		NA
Random Choice (RC)	0.027	NA		NA

Table 27: w3Merge Predictions

4.4.8 w3MergeNurse

Our first run of **automodel** using **w3MergeNurse** threw the us a Observation to Variable ratio error (See Section 3.2.8) since the calculated ratio was 0.54. To resolve this issue, we re-ran **automodel** with a **naPercent** value of 0.009. This run then threw us a Observation to Variable ratio error again (this one however, occurred after the Lower Level Removal step) since the calculated ratio was 4.97. To resolve this issue, we re-ran **automodel** with a **naPercent** value of 0.009 and a **obsPerLevel** value of 4. This run threw us a error stating that VIF (see Section 3.4.4) couldn't run since there was too much multicollinearity in our data. To resolve this issue, we re-ran **automodel** with a **naPercent** value of 0.009, a **obsPerLevel** value of 4 and a **corrConfLevel** of 0.5. This run managed to compile successfully in 919 seconds (15 minutes) and generated all of the models.

In Table 28 we have all of the prediction accuracies calculated per model for the run which managed to compile all of our models. We can see that all the models had a better exact prediction accuracy when compared to the MC and RC models, with Elastic Net Regression being the best with a 16% accuracy. When predicting in a +-1 range, the ENR model is best with a 45% accuracy and when predicting in a GHQ range the ENR model is best with a 52% accuracy.

Model	Exact	Conf. Interval	+-1 range	GHQ range
ENR	0.1705	NA	0.4474	0.5156
CART	0.1605	NA	0.4148	0.5114
K-Means	0.1398	NA	0.3339	NA
OLR	0.1335	0.4347	0.3636	0.4574
kNN	0.1307	NA	0.3026	NA
Mode Choice (MC)	0.1089	NA	0.3033	NA
Random Choice (RC)	0.027	NA	0.0796	NA

Table 28: w3MergeNurse Predictions

4.4.9 w3MergeNurseBlood

Our first run of **automodel** using **w3MergeNurse** threw the us a Observation to Variable ratio error (See Section 3.2.8) since the calculated ratio was 0.4. To resolve this issue, we re-ran **automodel** with a **naPercent** value of 0.0001. This run then threw us a Observation to Variable ratio error again (this one however, occurred after the Lower Level Removal step) since the calculated ratio was 4.75. To resolve this issue, we re-ran **automodel** with a **naPercent** value of 0.0001 and a **obsPerLevel** value of 1. This run threw us a error that a good Train Test Split couldn't be found (most likely due to the fact **obsPerLevel** was set to 1). Given that the way to resolve a Train Test Split (see Section 3.3.5) not being found is to increase **obsPerLevel** we had come to a standstill. Since we can't increase **obsPerLevel** anymore (1 was the only value that resolved our issue seen in the 2nd run) and decreasing **naPercent** any lower doesn't make any changes, we must conclude that **automodel** can't run this data-set. The conclusion of these turn of events is that the user should go into the data-set and begin a manual variable selection process to lower the amount of variables included in their data-set.

4.4.10 wShared

Our first run compiled in a time of 2180 seconds (36 minutes) and generated all of the models desired.

In Table 29 we have all of the prediction accuracies calculated per model for the run which managed to compile all of our models. We can see that all the models had a better exact prediction accuracy when compared to the MC and RC models, with Elastic Net Regression being the best with a 16% accuracy. When predicting in a +-1 range, the ENR model is best with a 45% accuracy and when predicting in a GHQ range the CART model is best with a 54% accuracy (The ENR model is very close to the CART).

Model	Exact	Conf. Interval	+-1 range	GHQ range
ENR	0.1607	NA	0.4457	0.5338
OLR	0.1546	0.2773	0.4372	0.5301
kNN	0.1491	NA	0.3521	NA
CART	0.1403	NA	0.3849	0.537
K-Means	0.1112	NA	0.3137	NA
Mode Choice (MC)	0.1089	NA	0.3033	NA
Random Choice (RC)	0.027	NA	0.0796	NA

Table 29: wShared Predictions

4.4.11 wSMerge

Our first run threw use a VIF error (See Section 3.4.4) since there was too much multicollinearity in the OLR model. To resolve this issue, we re-ran **automodel** with a **corrConfLevel** value of 0.5. This run compile successfully in a time of 581 seconds (10 minutes) and generated all of the desired models.

In Table 30 we have all of the prediction accuracies calculated per model for the run which managed to compile all of our models. We can see that all the models had a better exact prediction accuracy when compared to the MC and RC models, with Elastic Net Regression being the best with a 16% accuracy. When predicting in a +-1 range, the ENR model is best with a 43% accuracy and when predicting in a GHQ range the OLR model is best with a 52% accuracy (The ENR model was close to the OLR model).

Model	Exact	Conf. Interval	+-1 range	GHQ range
ENR	0.1599	NA	0.4272	0.5203
OLR	0.1561	0.2701	0.43	0.5229
CART	0.1379	NA	0.3751	0.5022
kNN	0.1363	NA	0.3213	NA
K-Means	0.1114	NA	0.3108	NA
Mode Choice (MC)	0.1089	NA	0.3033	NA
Random Choice (RC)	0.027	NA	0.0796	NA

Table 30: wSMerge Predictions

4.4.12 wSMergeNurse

Our first run compiled successfully in a time of 525 seconds (9 minutes) and generated all of the desired models.

In Table 31 we have all of the prediction accuracies calculated per model for the run which managed to compile all of our models. We can see that all the models had a better exact prediction accuracy when compared to the MC and RC models, with Elastic Net Regression being the best with a 17% accuracy. When predicting in a +-1 range, the ENR model is best with a 45% accuracy and when predicting in a GHQ range the CART model is best with a 54% accuracy.

Model	Exact	Conf. Interval	+-1 range	GHQ range
ENR	0.1658	NA	0.446	0.5383
OLR	0.1532	0.3935	0.4113	0.5194
kNN	0.1522	NA	0.3326	NA
CART	0.1427	NA	0.3799	0.5435
K-Means	0.1333	NA	0.3495	NA
Mode Choice (MC)	0.1089	NA	0.3033	NA
Random Choice (RC)	0.027	NA	0.0796	NA

Table 31: wSMergeNurse Predictions

4.4.13 wSMergeNurseBlood

Our first run threw a Observation to Variable ratio error (See Section 3.2.8) with a calculated ratio of 4.5. To resolve this issue, we re-ran **automodel** with a **naPercent** value of 0.15. This run compile successfully in a time of 724 seconds (12 minutes) and generated all of the desired models.

In Table 32 we have all of the prediction accuracies calculated per model for the run which managed to compile all of our models. We can see that all the models had a better exact prediction accuracy when compared to the MC and RC models, with Elastic Net Regression being the best with a 17% accuracy. When predicting in a +-1 range, the ENR model is best with a 45% accuracy and when predicting in a GHQ range the CART model is best with a 57% accuracy.

Model	Exact	Conf. Interval	+-1 range	GHQ range
ENR	0.168	NA	0.4487	0.5523
OLR	0.1549	0.3783	0.4054	0.5091
K-Means	0.1336	NA	0.3464	NA
kNN	0.1308	NA	0.331	NA
CART	0.1217	NA	0.3994	0.5744
Mode Choice (MC)	0.1089	NA	0.3033	NA
Random Choice (RC)	0.027	NA	0.0796	NA

Table 32: wSMergeNurseBlood Predictions

4.4.14 mixNurse

Our first run compiled successfully in a time of 29 seconds and generated all of the desired models.

In Table 33 we have all of the prediction accuracies calculated per model for the run which managed to compile all of our models. We can see that only our classification models (K-Means and kNN) did better than MC with K-Means having the highest accuracy of 12%. All models still preformed better than RC. When predicting in the +-1 range, Only the kNN model preforms better than the MC model with a 33%, all preformed better than the RC model still. When predicting in the GHQ range, our regression models more preform better than the MC model with ENR being the best with a 37% accuracy.

Model	Exact	Conf. Interval	+-1 range	GHQ range
K-Means	0.1236	NA	0.3327	NA
kNN	0.1148	NA	0.2967	NA
Mode Choice (MC)	0.1089	NA	0.3033	NA
ENR	0.0949	NA	0.2839	0.3724
OLR	0.0921	0.195	0.2855	0.3652
CART	0.0853	NA	0.2883	0.3612
Random Choice (RC)	0.027	NA	0.796	NA

Table 33: mixNurse Predictions

4.4.15 mixNurseBlood

Our first run compiled successfully in a time of 31 seconds and generated all of the desired models.

In Table 34 we have all of the prediction accuracies calculated per model for the run which managed to compile all of our models. We can see that only our classification models (K-Means and kNN) did better than MC with K-Means having the highest accuracy of 13%. All models still preformed better than RC. When predicting in the +-1 range, Only the K-Means model preforms better than the MC model, getting prediction accuracy of 34% (kNN was very close to MC), all preformed better than the RC model still. When predicting in the GHQ range, our regression models more preform better than the MC model with the CART model being the best with a 36% accuracy.

Model	Exact	Conf. Interval	+-1 range	GHQ range
K-Means	0.1288	NA	0.3424	NA
kNN	0.1107	NA	0.3031	NA
Mode Choice (MC)	0.1089	NA	0.3033	NA
ENR	0.0968	NA	0.2763	0.3546
OLR	0.0951	0.226	0.2802	0.3574
CART	0.09	NA	0.269	0.363
Random Choice (RC)	0.027	NA	0.796	NA

Table 34: mixNurseBlood Predictions

4.5 Conclusions

Overall, we managed to run automodel on all joins of the Understanding Society data expect for **w3MergeNurseBlood**. Some of the joins (for example **w2MergeNurse** and **w3MergeNurse**) required a little tweaking of the **automodel** parameters for a full run to be achieved. The results generally showed that the regression models fitted the data better for predicting **Survived**. The best model for predicting the GHQ score exactly was the ENR model. When predicting within a +-1 range, the ENR model preformed the best. When predicting within the GHQ range, both CART and ENR models seemed to be the best.

When looking at the results and applying our known context of the Understanding Society data (see Section 2.1), we can see some variables that were considered the most useful for predicting the GHQ score. Below we will go over some variables in a general case and not on a variable by variable case (due to the vast amount present).

- **SF-12 Physical and Mental Component Summary**: These are similar questionnaires to the GHQ that asks questions based on the participants physical and mental health. Many different answers and the total result of all of them were deemed significant.

- **Subjective financial situation - Current**: Known as **finnow**, this is a measurement of the participants current financial situation. The levels associated with the answers *Finding it quite difficult* or *Finding it very difficult* were deemed significant.

- **Current economic activity**: Known as **jbstat**, this is a question asked on the participants current financial situation. The level associated with the answer *LT sick of disabled* was deemed significant.

- **usual type of dairy consumption**: Known as **usdairy**, this is a question asked on the participants usual type of dairy consumption. The level associated with the answer *skimmed milk* was deemed significant.

- **Health Condition**: These are question asked about any certain health conditions the participant had. A participant having mentioning health condition not asked by the other health condition questions was deemed significant **hcondn96**

- **1st mentioned important event of year**: This is an event that the participant deemed most significant in the current year. The variable is categorical (97 levels) however was modelled as continuous variable (since **catLevels** was 15) therefore further investigation into the individual levels should be conducted (since when modelled as continuous, was still deemed significant).

- **1st mentioned important event of year**: This is an event that the participant deemed most significant in the current year. The variable is categorical (97 levels) however was modelled as continuous variable (since **catLevels** was 15) therefore further investigation into the individual levels should be conducted (since when modelled as continuous, was still deemed significant).

- **Satisfaction with health, Income and Life**: These are questions asked to a participants about their current satisfaction with health, income and life. Different **sclfsat1**, **sclfsat2** & **sclfsato** levels were deemed significant.

- **Control Over Things at Home**: This is a question completed by the participant that asks how their control over things at home is (known as **schmcont**). The responses *slightly agree* and *moderately disagree*.

- **What Happens in Life is Beyond my Control**: This is a question completed by the participant that asks how if what happens in life is beyond their control. The response *strongly disagree* was considered significant.

- **Different Demands on me Hard to Combine**: This is a question completed by the participant that asks if different demands on them are hard to combine. The responses *moderately agree*, *slightly disagree* and *strongly disagree* were considered significant.

- **In General, I Have Enough Time to do Everything**: This is a question completed by the participant that asks if, in general, they have enough time to do everything. The responses *moderately agree* and *strongly disagree* were considered significant.

- **Sex**: A participants sex, was considered significant.

Overall our **automodel** package has proved to be able to work with large quantities of data. The results can provide a user who's interested in the Understanding Society's data some interesting points to continue further analysis with.

5 Titanic Analysis

This is the results seen when running to following instance of **automodel**

```
titanicData.r = autoModel("Survived", titanicData, randomSeed = 3)
```

We will be covering the initial processing, evaluation methods and the results seen from this run.

5.1 Data Processing

The Titanic data-set didn't require any pre-processing to run. All we have done is encode the **Survived** variable as Didn't Survive if **Survived** was 0 and **Survived** if **Survived** was 1. This was done so that the feature of the dependant variable legend can be seen.

5.2 Evaluation Method

Our dependant variable, **Survived**, is a categorical variable that only takes 2 unique values. For this reason we will evaluate our results by only using the exact prediction accuracy of each model (MSE can be misleading given the regression models predict on a continuous scale). MSE and more is covered in Section 3.4.6.

We will compare the accuracy of predictions with the initial models/statistics in Table 35

Model	Exact
Random Choice (RC)	0.5
Mode Choice (MC)	0.6162

Table 35: Benchmark Accuracies for Titanic

- **Random Choice**: This is the accuracy of predicting if a passenger survived the Titanic (**Survived**) by complete random. This can be viewed as predicting a 1 with a 50% chance and a 0 with a 50% chance.

- **Mode Choice**: This is the accuracy of predicting if a passenger survived the Titanic (**Survived**) by always choosing the mode. This can be viewed as always predicting 0 since the mode passenger didn't survive the Titanic (The mode of **Survived** is 0).

5.3 Run Results

The following is what we expect to see in the console when running the Titanic data-set though the **autoModel** function:

- Console outputted legend telling us how the dependant variable (**Survived** in this case) has been encoded
- Console output telling us the changes that are made to the data for it to be cleaned
- K-Means model with prediction results
- kNN model with prediction results and training data used
- CART model with prediction results
- Ordinary Linear Regression model with prediction results
- Elastic Net Regression model with a list of fits attempted & prediction results (based on the best model)
- Console output of the prediction tables, MSE and prediction accuracies of each model

The data-set was processed and we received a full output from **automodel**. It took a total of 0.69 seconds to process and the results were output into the variable **titanicData.r**. We will go over each model made by **automodel** model and their respective evaluation metrics. The full processing of the data can be found in Section 10.

5.3.1 K-Means

Our K-Means model was created successfully with a total amount of clusters 2. The Within MSE of each cluster is quite similar with Cluster 1 having a smaller MSE than Cluster 2. From having a look at the results table, we can see that both clusters mainly contained observations of passengers that didn't survive (0). This means that our K-Means model will always predict that a passenger didn't survive (**Survived** = 0).

From Table 36, we can see that our K-Means model preforms 9% better than random choice (RC) and 2% worse than mode choice (MC). Overall, not a good model to use for predictions.

Model	Exact
K-Means	0.5926
Compared to RC	0.0926
Compared to MC	-0.236

Table 36: K-Mean Predictions for Titanic

5.3.2 kNN

Our kNN model successfully ran and considered 30 neighbors per testing observation. The MSE of of the model was 0.0113. From having a look at results table/confusion matrix, we can see that:

- The model predicted a passenger didn't survive correctly 104 times
- The model predicted a passenger didn't survive incorrectly 36 times
- The model predicted a passenger did survive correctly 31 times
- The model predicted a passenger did survive incorrectly 6 times

From Table 37, we can see that out kNN model preforms 26% better than random choice (RC) and 15% better than mode choice (MC). Overall a decent prediction accuracy that could be improved by tweaked the number of neighbours appropriately.

Model	Exact
kNN	0.7627
Compared to RC	0.2627
Compared to MC	0.1465

Table 37: kNN Predictions for Titanic

5.3.3 CART prediction model

Our CART model successfully ran and produced the following tree

```
n= 530

node), split, n, deviance, yval
      * denotes terminal node

 1) root 530 128.8472000 1.416981
   2) Sex.male>=0.5 332  56.3855400 1.216867
     4) Age>=-1.143733 305  45.0819700 1.180328
       8) Fare< 0.3436248 268  31.8917900 1.138060
        16) Ticket< 1.555194 255  27.2313700 1.121569
          32) Fare< -0.1525384 210  18.0952400 1.095238 *
          33) Fare>=-0.1525384 45   8.3111110 1.244444
            66) Fare>=-0.06442615 23   1.8260870 1.086957 *
            67) Fare< -0.06442615 22   5.3181820 1.409091
             134) PassengerId< 0.4470703 15   2.4000000 1.200000 *
             135) PassengerId>=0.4470703 7   0.8571429 1.857143 *
        17) Ticket>=1.555194 13   3.2307690 1.461538 *
       9) Fare>=0.3436248 37   9.2432430 1.486486
        18) Name>=0.2926589 11   0.0000000 1.000000 *
        19) Name< 0.2926589 26   5.5384620 1.692308
          38) PassengerId< -1.026823 7   0.8571429 1.142857 *
          39) PassengerId>=-1.026823 19   1.7894740 1.894737 *
     5) Age< -1.143733 27   6.2962960 1.629630
      10) SibSp.1< 0.5 15   3.3333330 1.333333 *
      11) SibSp.1>=0.5 12   0.0000000 2.000000 *
   3) Sex.male< 0.5 198  36.8737400 1.752525
     6) Pclass.3>=0.5 78  19.4871800 1.487179
      12) Fare>=-0.2132273 13   0.9230769 1.076923 *
      13) Fare< -0.2132273 65  15.9384600 1.569231 *
     7) Pclass.3< 0.5 120   8.3250000 1.925000 *
```

The CART model deemed the the top 3 most important variables for modelling/predicting **Survived** are:

- *Sex.male*, measures if the passenger was male or not, importance of 35.5879.
- *Fare*, measures the amount the passenger paid for their ticket, importance of 19.545
- *Age*, measures the age of the passenger, importance of 11.9058

The MSE of the model when predicting the testing data is 0.1396. Looking at Table 38, we can see that the CART model had an exact prediction accuracy of 0.8136 (81%). We can see that out CART model preforms 31% better than random choice (RC) and 20% better than mode choice (MC). Overall a better prediction accuracy that K-Means and kNN and could be improved by further tweaking.

Model	Exact
CART	0.8136
Compared to RC	0.3136
Compared to MC	0.1974

Table 38: CART Predictions for Titanic

5.3.4 Ordinary Linear Regression

The process of generating an Ordinary Linear Regression from the full model to the improved model ran successfully. Our initial full model had an AIC of 531.4729. In Variance Inflation Factor (VIF) Removal, 0 variables were removed since all variables had a VIF score lower than 10. In backwards selection, 14 out of 19 variables were removed since they weren't significant at the 95% confidence level.

Please see Table 39 for the final outputted Ordinary Linear Regression Model. Our final improved model contains 5 variables. Pclass describes the class the passenger was in, SibSp describes how many spouses/ siblings that passenger had aboard the Titanic, other variables are self explanatory.

Variable	Estimate Coef	Std. Error	Significance
(Intercept)	1.9541	0.0384	¡ 2e-16
Pclass.2	-0.2196	0.0496	1.16e-05
Pclass.3	-0.3805	0.0446	¡ 2e-16
Sex.male	-0.4636	0.03652	¡ 2e-16
Age	-0.0869	0.0186	3.76e-06
SibSp.3	-0.334	0.1341	0.0131

Table 39: Ordinary Linear Regression Model Variables

The R^2 value for the model above is 0.3724, the MSE of the model is 0.1526 and the AIC is 521.6136. When predicting, the MSE is 0.128 and, as seen in Table 40, our Exact prediction accuracy is 0.8249 (we ignore the prediction within the confidence interval due to the context of our dependant variable **survived**). Looking at Table 40, we can see that out Ordinary Linear Regression model preforms 32% better than random choice (RC) and 21% better than mode choice (MC). Overall a good model for **Survived** and we have a fair amount of room to tweak this model to achieve a higher prediction accuracy.

Model	Exact
Ordinary Linear	0.8249
Compared to RC	0.3249
Compared to MC	0.2087

Table 40: Ordinary Linear Regression Predictions for Titanic

5.3.5 Elastic Net Regression

Our final model, Elastic Net Regression, ran successfully. In Table 41 we can see the list of variables which our best Elastic Net Regression model considered for prediction/modelling. Similar to the Ordinary Linear Regression model, Pclass (A passengers class), Sex & Age were considered. Unlike Ordinary Linear Regression, Fare was considered and SibSp (passengers amount of spouses/siblings on-board) wasn't.

Variable	Estimate Coef
(Intercept)	1.7498
Pclass.3	-0.1451
Sex.male	-0.4166
Age	-0.0157
Fare	0.0367

Table 41: Elastic Net Regression Model Variables

The best Elastic Net Model chosen by MSE had an α of 1. This model's predictions had an MSE of 0.13. As seen in Table 42, the model had an exact prediction accuracy of 0.8192, slightly worse than Ordinary Linear Regression. Overall a good model of **Survived** however not our best. The model could be improved by tweaking however these is very little that can be done compared to other models.

Model	Exact
Elastic Net	0.8192
Compared to RC	0.3192
Compared to MC	0.203

Table 42: Elastic Net Regression Predictions for Titanic

5.4 Conclusions

All our models were created successfully and preformed better than the RC and MC models, expect for K-Means. In Table 43 we are comparing all of the models made during the **automodel** run of the Titanic data-set. We can see that the Regression models seem to be best for modelling our dependant variable, **Survived**, with Ordinary Linear Regression being the best fit by Exact Prediction accuracy. If we were to carry on the analysis from here, it would be best to attempt fitting a better Ordinary Linear Regression model.

Model	Exact
OLR	0.8249
ENR	0.8192
CART	0.8136
kNN	0.7627
Mode Choice (MC)	0.6162
K-Means	0.5926
Random Choice (RC)	0.5

Table 43: Elastic Net Regression Predictions for Titanic

Since this data came from Kaggle, it's possible for us to use our best model made (which was our OLR model) on a testing set of data provided by Kaggle. To achieve this, we ran automodel with the same randomSeed but only used 1 test observation (which is done by setting **testPercent** to -1). Then based on the resulting model, we formatted the testing data provided by Kaggle into the formatted which our OLR model accepted. The code that generated this output can be found in Section 10.

As seen in Figure 9, at the time of submission, the OLR model generated by **automodel** scored a 0.75598 (76% accuracy in prediction) and was ranked 12289 out of roughly 14500 entries. The model's prediction accuracy is rather good but compared to other models made with the data, there still a fair amount of improvement that could be done. This represents that even though *automodel* generates useful models, the models shouldn't be used as a final consideration for the data-set.

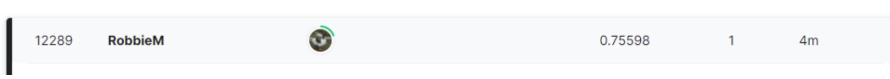


Figure 9: Kaggle leader-board Snippet

6 Iris Analysis

This is the results seen when running to following instance of **automodel**.

```
iris.r = autoModel("Species", iris, randomSeed = 1, corrConfLevel = 0.99)
```

We have tweaked the **corrConfLevel** since our brief correlation analysis on the iris data showed there was high correlation between variables which we wish to keep. This means for this run, we are expecting better results from our clustering/ classification models.

6.1 Data Processing

The Iris data-set didn't require any pre-processing to run. Since Iris is in-built into R, we don't need to load in any data and can call it within **automodel** without any hassle.

6.2 Evaluation Method

Our dependant variable, **Species**, is a categorical variable that takes 3 unique values. For this reason we will evaluate our results by only using the exact prediction accuracy of each model (MSE can be misleading given the regression models predict on a continuous scale). To understand these, please see Section 3.4.6.

We will compare the accuracy of predictions with the initial models/statistics seen in Table 44.

Model	Exact
Random Choice (RC)	0.3333

Table 44: benchmark Accuracies Iris

- **Random Choice**: This is the accuracy of predicting the species of iris (**Species**) by complete random. This can be viewed as predicting *Setosa* with a 33% chance, *Versicolor* with a 33% chance and *Virginica* with a 33% chance.

We do not use Mode Choice to predict like in the other runs (See Section 4.3 & 5.2) since Mode Choice (MC) is the same as Random Choice (RC). This is because there is a equal number of each species of iris within the Iris data (50 observations of each).

6.3 Run Results

The following is what we expect to see in the console when running the Iris data-set though the **autoModel** function:

- Console outputted legend telling us how the dependant variable (**Species** in this case) has been encoded
- A legend stating how our dependant variable has been encoded
- Console output telling us the changes that are made to the data for it to be cleaned
- K-Means model with prediction results
- kNN model with prediction results and training data used
- CART model with prediction results
- Ordinary Linear Regression model with prediction results
- Elastic Net Regression model with a list of fits attempted & prediction results (based on the best model)
- Console output of the prediction tables, MSE and prediction accuracies of each model

The data-set was processed and we received a full output from **automodel**. It took a total of 0.56 seconds to process and the results were output into the variable **iris.r**. We will go over each model made by **automodel** model and their respective evaluation metrics. The full processing of the data can be found in Section 10

6.3.1 K-Means

Our K-Means model was created successfully with a total amount of clusters being 3. From having a look at the results table in the console output, we can see that K-means has identified a cluster that only contains observations of **Species** *Setosa*, this being Cluster 3. Cluster 1 and Cluster 2 are also pretty well defined with Cluster 1 mainly being **Species** *Versicolor* and Cluster 2 mainly being **Species** *Virginica*.

From Table 45, we can see that our K-Means model preforms 56% better than random choice (RC). This is a good result and suggest that the data does indeed naturally cluster per iris species.

Model	Exact
K-Means	0.8933
Compared to RC	0.56

Table 45: K-Mean Predictions for Iris

6.3.2 kNN

Our kNN model successfully ran and considered 3 neighbors per testing observation. The MSE of of the model was 0.1053. From having a look at results table/confusion matrix, we can see that:

- The model predicted *setosa* correctly 15 times
- The model predicted *versicolor* correctly 8 times
- The model predicted *virginica* correctly 13 times
- The model predicted *virginica* in-correctly 2 times

From Table 46, we can see that out kNN model preforms 61% better than random choice (RC). Overall a very good model for *Species* and little needs to be done to improve the model.

Model	Exact
kNN	0.9474
Compared to RC	0.6141

Table 46: kNN Predictions for Iris

6.3.3 CART prediction model

Our CART model successfully ran and produced the following tree

```
n= 112

node), split, n, deviance, yval
      * denotes terminal node

 1) root 112 71.964290 2.017857
   2) Petal.Length< -0.7409513 35  0.000000 1.000000 *
   3) Petal.Length>=-0.7409513 77 19.220780 2.480519
     6) Petal.Length< 0.5619447 37  0.972973 2.027027 *
     7) Petal.Length>=0.5619447 40  3.600000 2.900000
      14) Petal.Length< 0.7318877 11  2.545455 2.636364 *
      15) Petal.Length>=0.7318877 29  0.000000 3.000000 *
```

The CART model gave the importance in each variable for predicting **Species** as follows:

- **Petal.Length**, the length of the petal in centimeters, importance of 68.44586
- **Petal.Width**, the width of the petal in centimeters, importance of 64.71597
- **Sepal.Length**, the length of the sepal in centimeters, importance of 46.59445
- **Sepal.Width**, the width of the sepal in centimeters, importance of 28.65781

The MSE of the model when predicting the testing data is 0.0406. Looking at Table 47, we can see that the CART model had an exact prediction accuracy of 0.9474 (94%). We can see that out CART model preforms 61% better than random choice (RC). Overall a very good model for *Species* and little needs to be done to improve the model.

Model	Exact
CART	0.9474
Compared to RC	0.6141

Table 47: CART Predictions for Iris

6.3.4 Ordinary Linear Regression

The process of generating an Ordinary Linear Regression from the full model to the improved model ran successfully. Our initial full model had an AIC of -12.4783. In Variance Inflation Factor (VIF) Removal, **Petal.Length** was removed since it had a VIF score of 33.0637. The resulting model after VIF had an AIC of 2.4362. In backwards selection, **Sepal.Width** and **Sepal.Length** were removed since they weren't significant at the 95% confidence level.

Please see Table 48 for the final outputted Ordinary Linear Regression Model. Our final improved model contains only Petal.Width. Since this model only has on variable in it, we can say that there is no multicollinearity between predictor variables.

Variable	Estimate Coef	Std. Error	Significance
(Intercept)	2.01155	0.02379	¡ 2e-16
Petal.Width	0.79787	0.02492	¡ 2e-16

Table 48: OLR Model Variables

The R^2 value for the model is 0.9022, the MSE of the model is 0.0623 (based on the training set) and the AIC is 12.878. When predicting, the MSE is 0.0416 and, as seen in Table 49, our exact prediction accuracy is 1 (we ignore the prediction within the confidence interval due to the context of our dependant variable **Species**). Looking at Table 49, we can see that out Ordinary Linear Regression model preforms 67% better than random choice (RC). Overall the model predicts perfectly, however, this may be because of the Train Test Split or the number of observations so the accuracy should be taken with caution.

Model	Exact
Ordinary Linear	1
Compared to RC	0.6667

Table 49: Ordinary Linear Regression Predictions for Titanic

6.3.5 Elastic Net Regression

Our final model, Elastic Net Regression, ran successfully. In Table 50 we can see the list of variables which our best Elastic Net Regression model considered for prediction/modelling. In general, ENR considered all of the variables when predicting **Species**

Variable	Estimate Coef
(Intercept)	2.0068
Sepal.Length	0.0267
Sepal.Width	-0.0662
Petal.Length	0.3283
Petal.Width	0.3814

Table 50: Elastic Net Regression Model Variables

The best Elastic Net Model chosen by MSE had an α of 0.1. This models predictions had an MSE of 0.0455. As seen in Table 51, the model had an exact prediction accuracy of 0.9737 (97%), slightly worse than Ordinary Linear Regression. Overall seemly good model for *Species* however since the model has considered all the predictor variables, we would assume the model has high multicollinearity and therefore varies drastically with small changes.

Model	Exact
Elastic Net	0.9737
Compared to RC	0.6404

Table 51: Elastic Net Regression Predictions for Iris

6.4 Conclusions

All our models were created successfully and preformed better than the RC model. In Table 52 we are comparing all of the models made during the **automodel** run of the Iris data-set. Surprisingly, our best model for prediction was Ordinary Linear Regression (OLR) with a model only using **Petal.Width**. Our classification models, K-Means and kNN still did well and showed that the data does seem to cluster based on **Species**. Generally, to better improve our calculated accuracies, more observations of iris flowers would be needed.

Model	Exact
OLR	1
Elastic Net	0.9737
CART	0.9474
kNN	0.9474
K-Means	0.8933

Table 52: Elastic Net Regression Predictions for Iris

7 Discussion

Now that we have seen what the **automodel** package is and how it works with different data-sets, we will discuss on why this data and methodology was used. We will also make mention to further improvements that could be done to **automodel** and give reasons why.

7.1 Data

To demonstrate **automodel**, 3 different data-sets were used.

Firstly, the Understanding Society data-set (see Section 2.1) was the main data source used within this dissertation. This is because of various reasons:

- When initially working on this dissertation, the goal was to fit a model that best described the GHQ score within Understanding Society. Naturally this meant **automodel** was first built using this data as a training ground.
- The Understanding Society data is incredibly large with data sources reaching up to 400MB+ in size. This gives us very large data-sets for us to work with meaning that when building **automodel**, we can account for issues that occur when using large amounts of data.
- The people that assisted in providing context to the Understanding Society data had academic links to the data-set. This meant that when building **automodel**, their wealth of knowledge could be used to understand the context around the large amount of data.

Secondly, the Titanic data-set (see Section 2.2) was used to show the how data assumed to have a linear relationship looks when using the **automodel** function. This is because of various reasons:

- Since the data-set was provided by Kaggle, the results could be uploaded to Kaggle leader-board to compare how the models generated compared with other built models submitted to Kaggle.
- The data-set included data which would test the cleaning, transforming and modelling functions to a good degree. Example: The variable **Name** is a string variable were each observation is a passengers name. This **Name** would test our cleaning methods and then our modelling methods.

- Due to it's smaller size compared to Understanding Society, this data-set became very useful to quickly preform tests on the **automodel** package after an update. A average run of the Titanic data-set took about 1 second (compared to Understanding Societies 20+ minutes)

Lastly, the Iris data-set (see Section 2.3) was used to demonstrate how well the **automodel** package represents data that is known to already have natural clustering.

- The data-set is easily recognized which gives readers a good benchmark to compare the results of their individual analysis to the **automodel** package.
- Since the variables were highly correlated with each other, this run was designed to favour our classification models and see if they were able to out preform the regression models in the model evaluation metrics.
- Due to it's smaller size compared to Understanding Society, runs with this data can be preformed very quickly, giving the reader another data-set like the Titanic data-set (see Section 2.2) to play around with.

7.2 Methodology

In this dissertation lots of different cleaning, transforming and modelling methods have been used within **automodel**. The reason for why these methods were chosen is all relatively the same; during university or outside study, these were the methods that were used the most/ were already known at a very basic level. Further investigation has been done on these methods so that an in depth understanding of them was learnt. Therefore instead of discussing why these certain methods were used, we will go over some criticisms of their usage in this dissertation.

7.2.1 Data Type Checks

Our method of casting string variables which have more than **catLevels** levels to a numerical variable lead to some strange results. Example: In our Titanic run, **Name** is considered a significant variable when considered as a continuous variable. This result isn't the correct way to model the variable **Name** and given our prior context on it, could cause and over-fit model.

7.2.2 Missing Value Imputation

Our method of Single Imputation using CART keeps time complexity low and ensures the function runs smoothly, however, this method doesn't account for the known variation that comes with imputing missing values. To account for this Multiple Imputation is generally used (more on this in Section 7.3).

7.2.3 Low Level Removal

Even though a user can adjust the value for this variable, the default fun of automodel will discriminate against factor levels which have low observations. Depending on the data-set parsed, this may be an undesired bias.

7.2.4 Correlation Checks

When checking for correlation, we discriminate against variables that come earlier in the data-set. This gives the user the functionality to have some manual input on what variables get removed at this stage however it comes at a trade-off. To achieve our result we now have a bias in our correlation selection method which could lead to more variables being removed than necessary. Another method that could of been introduced in for Section 3.2.10 is a check on the correlation coefficients between the predictor variables and the dependant variable. Then, another selection method could be done of these correlation coefficients to refine the predictor variables chosen for analysis. Example: We keep predictor variables if they have a correlation coefficient with the dependant variable of 0.5 or higher.

7.2.5 Dummy Variables

If we have a categorical variable, we won't create a dummy variable to represent the first ordered level since it's represented using the other dummies made (if all other dummies are 0, this represents the first ordered level). This reduces the amount of variables within the data-set analyzed. This may be an issue though when preforming variable selection in regression models (like OLR). Example: the first ordered level in a categorical with more than 2 levels is very significant in predicting our dependant variables whilst the other levels aren't. When encoding, our first level won't be represented as a single dummy variable but instead as a combination of all other dummies made from the same variable. This then makes the significance/importance of that first ordered level spread across the combination of dummy variables that describe it which can lead to this highly significant level to be considered insignificant due to it's other levels.

7.2.6 Factorization

The **autoModel** function decides on if a variable is a factor based on the function variable **catLevels**, which by default is 15. If there is a categorical variable within the data that has more levels than **catLevels**, it will be treated as continuous.

7.2.7 Train Test Split

To be able to get a prediction accuracy for our models, we split our data-set into a training and testing set. Since there's a random element to this process the resulting training and testing sets can have a large impact on the model created, causing variation in results each run. To minimise this, k-fold cross validation could be done to see how the models preform for each fold (more on this in Section 7.3)

7.2.8 K-Means

In our K-Means model, we decide clusters based on the number of unique levels within the dependant variable (if the dependant variable is continuous, **catLevels** clusters are made). This can lead to a bad representation of the data as clusters. Instead, the number of clusters should be decided though an evaluation method (More on this in Section 7.3). Our usage of K-Means as a prediction model means we can compare the models to the results of other models used by **automodel** (though exact prediction accuracy). Even though useful, this is an unorthodox way of using K-Means and the primary focus should be on the natural groups/clusters that form within the data-set. When modelling a continuous variable, our results tables and our exact prediction accuracies provide misleading results.

7.2.9 kNN

For our kNN model, we choose the number of neighbours to consider by taking the square root of the number of observations present ($\sqrt{n}$). This is a good estimate for the number of neighbours needed however it may not be the best. Instead, a selection method for k should be introduced where k is decided by calculating and error rate per k value and picking the k with the lowest error rate. When modelling a continuous variable, our results tables and our exact prediction accuracies provide misleading results.

7.2.10 Ordinary Linear Regression

For our backwards selection process, we choose our variables based on their statistical significance when modelling our dependant variable. Doing this can lead to having variables which are all considered statistically significant but the overall model not considered statically significant (more on this in Section 7.3). Also, our selection process could be more efficient if the selection process used was step-wise selection. This could increase accuracy but due to the time complexity this method brings, we chose to use backwards selection instead.

7.2.11 Elastic Net Regression

When choosing out λ to use in our penalty function, we preform cross validation then pick the λ which the smallest value. Instead, we could of potentially picked the λ of the smallest model that is within 1 standard error of the smallest value of λ. We chose not to since picking the smallest λ after cross validation is simpler to understand and produces the most accurate model (that comes with a trade off of more dimensions).

7.3 Future Improvements

Since there's been a time constraint on the completion of this dissertation, there are still many ways which the **automodel** package could be improved. Below I will talk about these in brief detail:

- **automodel** could also find the best value for k within K-Means clustering. By using method such as *elbow* or *average silhouette* we could instead find the best value for k in K-Means instead of it being an arbitrary amount that can be tuned by the user.

- In Ordinary Linear Regression, we improve our models using backwards selection on the statistical significance of each predictor variable at the **modelSigLevel** confidence level. If our final model includes lots of predictor variables, the entire model will be considered statistically significant at the product of all the confidence levels in each predictor variable. For example, if a model contains two variables that are statistically significant at exactly the 95% confidence level, the model will be considered statistically significant at the $95\% \times 95\% = 90.25\%$ confidence level. For this reason, using statistical significance as our metric for selection can lead to some models with overall, aren't that statistically significant.

  To get around this issue, we should use another metrics for our selection process. One example would be using the Akaike information criterion (AIC) to select our models instead. Instead of selecting a model based on the statistical significance of the variables included, the model is picked on the best overall performance at predicting the dependant variable (based on the AIC).

- In Ordinary Linear Regression, there is a lack of investigation into interactions between variables. To improve this, we could make the Ordinary Linear Regression model consider the interactions between predictor variables.

- Further investigation could be done on if the relationships between the dependant variables and predictor variables aren't linear. modelling terms as their squares, x^2, or as exponential, $\exp^x$, could lead to us finding further interesting relationships within the users data.

- For large data-sets, running the **autoModel** function can take quite a long time. We could improve this run-time by preforming some general quality of life improvements on the code which means that large calculations don't have to occur as often (such as in VIF or backwards selection, we don't recalculate the model after each variable removal, instead in jumps of 2 or 3).

- When processing the full Ordinary Linear Regression model, we have a step where we have to remove variables that the function deemed as unusable/ as NA. It's unknown the exact reasons why some variable appear like this since it's quite contextual to the data inputted, thare are some theories behind why though:

  – After correlation checks, some variables are considered so highly correlated, the lm() function doesn't model them. It's hard to know if this is the case though since there doesn't seem to be a defined limit to the correlation lm() checks for.

  – Some variables in a data-set can be composites of other variables in the data-set. For example if you have a variable called *score* that was the sum of 10+ other variables within the data-set, you may not get high correlation between *score* and these other variables however score can still turn up NA after calling lm() (creating a OLR model). It was very hard to think of a general function to tackle this issue and will be considered in the future

- There is no formal checks for outliers within the data. Besides the removal of categorical variable levels with low observations (see Section 3.2.7), we do not remove any observations if they are considered to be an outlier. This can results in very poor predictions when modelling and therefore should be accounted for correctly (using leverage or the interquartile range).

- Out function doesn't address the normally of the data/ how normally distributed each variable is. An option for a user to address this situation should be included as some model assumptions may want their variables to be considered as normally distributed.

- In testing, it's possible for a user to be shown an error message telling then to increase the **automodel** variable **obsPerLevel** (if a user couldn't find a good Train Test split, this would occur). After doing so the user could then be hit with another error message telling them to lower **obsPerLevel** (After low level observation removal, there may not be enough observations to continue). This can make the journey confusing and also seemingly impossible to run and therefore we should consider manually tuning parameters in those cases where an error occurs.

- Instead of having abstract numerical labelling in the clustering/classification models for categorical dependant variables, we should give them understandable labels. Say we were trying to predict a variable that as originally "Survived" and "Not Survived", we should aim to label our prediction tables as such (our current solution is outputting a legend for the user to follow at the start of the function).

- When predicting a binary variable using linear regression can lead to some misleading results (predicting 0.33 instead of 0 or 1). Instead we should be fitting a logistic regression model to predict variables of this nature.

- Instead of removing levels from categorical variables that have low observations, we could instead consider other options. One idea would be merging similar levels together if they have low observations, this is quite a hard task to achieve without any context of the data-set.

- To decrease the variance in the resulting models, k-folds cross validation could be done on the Train Test split. This would result in **automodel** modelling the data k times on k equally distributed folds of the data-set, then, the models results could be collated to get an average and variance across the results.

- K-Means and kNN aren't very good at representing the predictions of continuous variables. The predictions tables aren't going to format correctly since it's difficult to predict a continuous variable exactly. The SSR results for both algorithms are still valuable given the context of a continuous variable prediction.

- There is a lack of graphical output from the function. It would be hard to decide what graphs to include since there are a-lot of features which are worthy of plotting. We don't want to plot to much since it could end up breaking up the flow of the function's run-time. It's a hard one to decide on what's best to do

- The make future updates and collaboration easier, the **automodel** package should be re-designed in a modular style. This would mean that each method used would be their own function within the **automodel** package instead of being all bundled in the **autoModel** function. The **autoModel** would then call to these functions in order of best execution.

8 Conclusion

In this dissertation, we have investigated multiple different mathematical methods, created a function to use them all automatically and then analyze the results when this function was applied to different data-sets. This process has demonstrated that it is indeed possible to automate the modelling process. It has also demonstrated a technical understanding of the mathematical principles and concepts behind the techniques used to preform data science and analytics. This knowledge has then been successfully translated to the statistical coding language R where the working package, **automodel**, has been made.

The results from the **autoModel** function showed that even though the model performances were deemed useful, they aren't to be considered the best models for the specific data it's model from. Therefore the best usage of **automodel** is to use it in preliminary analysis to help advise in-depth analysis that utilizes the context of the data-set being analyzed. The audience that could get the most from this package would be those already with analytical knowledge so that the results from **automodel** can be interpreted correctly. If a person was to use **automodel** without this knowledge, there is the risk that the large quantity of output could lead to some poor/ incorrect conclusions. The package also doesn't provide a good learning experience and therefore it's recommended that those with less knowledge in data science and analytics initially preform analysis without **automodel**.

The **automodel** package in it's current stage works, but should undergo a fair amount of refinement. To improve, the features listed in Section 7.3 Future Improvements could be implemented. Hopefully this package will be continually worked on beyond this dissertation to further improve the usefulness for those looking to preform data science and analytics.

Overall this dissertation has been an in-depth learning experience and has generated a function, **automodel**, which can be used as a useful tool for analysis.

9 Bibliography

References

[1] Understanding Society (1978). **"Understanding Society: About the Study"**. Website: https://www.understandingsociety.ac.uk/about/about-the-study

[2] Understanding Society (1978). **"Understanding Society: Survey Timeline"**. Website: https://www.understandingsociety.ac.uk/documentation/mainstage/survey-timeline

[3] Goldberg D (1978). **"GHQ-12: A quick screener for survey use"**. Description: https://eprovide.mapi-trust.org/instruments/general-health-questionnaire#basic_description

[4] Glen_b (2014). **"How to interpret a QQ plot"** Post: https://stats.stackexchange.com/questions/101274/how-to-interpret-a-qq-plot

[5] Kaggle (2019). **"Titanic - Machine Learning from Disaster"** Article: https://www.kaggle.com/c/titanic/data

[6] Edgar Anderson (1988) **"iris: Edgar Anderson's Iris Data"** Documentation: https://www.rdocumentation.org/packages/datasets/versions/3.6.2/topics/iris

[7] Dr Mercedes Torres-Torres; Dr Michael Pound (2019). **"Data Analysis - Computerphile"**. Series start: https://www.youtube.com/watch?v=NxYEzbbpk-4&list=PLJ1bsfrnyoq63ncJB19w3nFm4i4K04mNq

[8]

[9] Isabella Ghement (2019). **"Multiple Imputation by Chained Equations (MICE) Explained"**. Post: https://stats.stackexchange.com/questions/421545/multiple-imputation-by-chained-equations-mice-explained

[10] Stef van Buuren (2012). **"Flexible Imputation of Missing Data, Second Edition"**. Book: https://stefvanbuuren.name/fimd/sec-MCAR.html

[11] Nunan D, Bankhead C, Aronson JK (2019). **"Catalogue of Bias Collaboration: Information bias"**. Catalogue Of Bias 2017: https://catalogofbias.org/biases/information-bias/

[12] Nunan D, Bankhead C, Aronson JK (2017). **"Catalogue of Bias Collaboration: Selection bias"**. Catalogue Of Bias 2017: http://www.catalogofbias.org/biases/selection-bias/

[13] Paul T. Barrett, Paul Kline (1981). **"Observation to Variable Ratio"** Paper: https://www.pbarrett.net/publications/Observation-to-Variable-Ratio-Barrett-and-Kline-1981.pdf

[14] Songhao Wu (2020). **"Multicollinearity in Regression"** Post: https://towardsdatascience.com/multi-collinearity-in-regression-fe7a2c1467ea

[15] Cuemath (2022) **"Transitive Relations"** Page: https://www.cuemath.com/algebra/transitive-relations/

[16] amoeba (2015). **"Making sense of principal component analysis, eigenvectors & eigenvalues"** Post: https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues

[17] Wiki, Unknown (2022). **"Eigendecomposition of a matrix"** Page: https://en.wikipedia.org/wiki/Eigendecomposition_of_a_matrix

[18] Nour Al-Rahman Al-Serw (2021). **"K-nearest Neighbor: The maths behind it, how it works and an example"**. Article: https://medium.com/analytics-vidhya/k-nearest-neighbor-the-maths-behind-it-how-it-works-and-an-example-f1de1208546c

[19] Avinash Navlani (2018). **"KNN Classification Tutorial using Scikit-learn"**. Tutorial: https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn

[20] Josh Starmer (2019). **"Regression Trees, Clearly Explained!!!"** Video: https://www.youtube.com/watch?v=g9c66TUylZ4

[21] Dr. GP Pulipaka (2016). **"An essential guide to classification and regression trees in R Language"** Article: https://medium.com/@gp_pulipaka/an-essential-guide-to-classification-and-regression-trees-in-r-language-4ced657d176b

[22] Cettt (2019). **"How is Pr(¿—t—) in a linear regression in R calculated?"** Post: https://stackoverflow.com/questions/55969011/how-is-prt-in-a-linear-regression-in-r-calculated

[23] Josh Starmer (2018). **"Machine Learning Fundamentals: Bias and Variance"**. Video: https://www.youtube.com/watch?v=EuBBz3bI-aA

[24] Josh Starmer (2018). **"Regularization Part 2: Lasso (L1) Regression"**. Video: https://www.youtube.com/watch?v=NGf0voTMlcs

[25] Josh Starmer (2018). **"Regularization Part 1: Ridge (L2) Regression"**. Video: https://www.youtube.com/watch?v=Q81RR3yKn30

[26] Josh Starmer (2018). **"Regularization Part 3: Elastic Net Regression"**. Video: https://www.youtube.com/watch?v=1dKRdX9bfIo

[27] Josh Starmer (2018). **"Machine Learning Fundamentals: Cross Validation"**. Video: https://www.youtube.com/watch?v=fSytzGwwBVw

[28] Newcastle University (2022). **"Coefficient of Determination, R-squared"** Site: https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.html

[29] Alexandre Zajic (2019). **"Introduction to AIC — Akaike Information Criterion"** Article: https://towardsdatascience.com/introduction-to-aic-akaike-information-criterion-9c9ba1c96ced

10 Appendix

10.1 Code

10.1.1 automodel

```
#' autoModel
#'
#' autoModel will attempt to model a variable from a given data-set using various different analytica
#'     It includes various cleaning & transforming procedures, K-Means model, kNN model, CART model,
#'         Elastic Net Regression Model.
#'
#' @importFrom dplyr select all_of
#' @importFrom stringr str_remove str_replace
#' @importFrom rlang is_empty
#' @import mice
#' @importFrom caret dummyVars
#' @importFrom glmnet cv.glmnet
#' @import Matrix
#' @importFrom broom tidy
#' @importFrom car vif
#' @importFrom class knn
#' @importFrom rpart rpart rpart.control
#' @importFrom stats AIC coef cor kmeans lm na.omit predict princomp runif sd
#'
#' @param predictVar The name of the variable that is to be predicted/the name of the dependent varia
#' @param data The data-set of all the predictor variables and the dependent variable
#' @param naPercent This is the percentage amount of NA values allowed in a predictor variable.
#'     If the predictor variable have naPercent percent or more of their values as NA/missing, the p
#' @param cartSplit The amount of observations needed for a new threshold/node to be generated in a C
#'     CART is used to impute missing values and to generate a prediction model of the data within th
#' @param impFlag Flag to tell the function if we are wanting to impute the missing values instead o
#'     If TRUE, a single CART imputation is done to impute the missing values.
#' @param randomSeed To make results reproducible, we must set a random seed within the function.
#'     randomSeed can be set to any integer in the range -2147483647 to 2147483647 (2147483647 is the
#' @param catLevels Decides on how many unique values/levels  a variable needs to be considered as a
#'     If a variable has less than or equal to catLevels levels, the variable gets encoded as a facto
#' @param obsPerLevel How many observations are needed of a level in a categorical/factor variable.
#'     If the level in the variable has less than obsPerLevel observations, all observations of this
#' @param clusterAmount Sets the amount of clusters/centroids to be used in K-Means modelling.
#'     Our K-Means model has clusterAmount clusters.
#' @param corrConfLevel The cuttoff point of when a correlation between two predictor variables is de
#'     If the absolute value of the correlation coefficient between two predictor variables is greate
#' @param PCAFlag Flag that tells the automodel function to preform PCA on the data-set.
#'  If TRUE, PCA using the co-variance matrix and eigenvalue decomposition is preformed.
#' @param pcPercent This sets the needed amount of variance explained by the reduced principal compo
#'     When preforming dimension reduction, we keep the smallest amount of PC's that describe pcPerce
```

```
#' @param testPercent The proportion of observations after transformation within the testing data.
#'     When preforming our Train Test split on our data, we set testPercent percent of our observatio
#'     Our training data must maintain variance in all predictor variables and therefore the splittin
#'     If randomSeed is set, it only evaluates the Train Test split associated with the seed once.
#'     A user can set testPercent to -1 if only one observation should be tested.
#' @param kNNCount The number of neighbors that are considered for each test observation in the kNN r
#'     Each test observation considers kNNCount neighbors to predict the dependent variable for the c
#' @param vifSelectionLevel We will be removing variables from the Ordinary Linear Regression model i
#' @param modelSigLevel To refine our Ordinary Linear Regression Model we remove variables based on t
#'     Each model needs to be considered significant at the modelSigLevel confidence level.
#' @param elasticCount When fitting our elastic net models, we need to fit models for different valu
#'     elasticCount decides how many different values of alpha we try and the different values are in
#'
#' @export
#'
#' @examples
#' \dontrun{
#' #reading in an example csv as our data
#' data = read.csv("data.csv")
#'
#' #running the autoModel function and setting the results to the variable 'data.results'
#' data.results = autoModel(predictVar = "Variable", data = data)
#' }
autoModel = function(
  predictVar,
  data,
  naPercent = 0.2,
  cartSplit = 20,
  impFlag = FALSE,
  randomSeed = NULL,
  catLevels = 15,
  obsPerLevel = 5,
  clusterAmount = ifelse(length(unique(data[[predictVar]])) <= catLevels, length(unique(data[[predict
  corrConfLevel = 0.8,
  PCAFlag = FALSE,
  pcPercent = 0.95,
  testPercent = 0.25,
  kNNCount = round(sqrt(nrow(data)),0),
  vifSelectionLevel = 10,
  modelSigLevel = 0.95,
  elasticCount = 10
) {
  #-----Functions needed for initial analysis-----

  #After some deletions, we have variables which now have now have 0 variance (all values in column a
  #This will be wrapped in a function as we will need to repeat this step later
  #This function is used within the autoModel function multiple times to remove columns with no varia
  #A column with no variance for example is [1,1,1,1,1] (all values the same)
  var0Remove = function(dataset,print = T) {

    #get all variances and find names of all attributes with NA variance (as.numeric() is included in
    naSdRM = names(which(sapply(na.omit(dataset),function(x) sd(as.numeric(x))==0 )))

    #This checks if we want to print the results of this test, prints results by default
    if (print){
      print("-------------Variance 0 Check-------------")
      print(paste(length(naSdRM)," variables removed since their new variance was 0"))
      print(naSdRM)

    }

    #Remove the values
    dataset = dataset[,!(names(dataset) %in% naSdRM)]

    return(dataset)
  }
  #-----
```

```
###START###
print("------------Initial Checks------------")
#as a added feature, we can add how long the function ran for, we start the timer here
timer = proc.time()

#First to fool proof the data, we need to treat as a data-frame
data = as.data.frame(data)

#next, we need to do a quick NA transform where blank string is coded to 'NA'
data[data == ""] = NA

#return the count of na values
print(paste(sum(is.na(data)), " NA cells were found across the entire dataset (",
            round((sum(is.na(data))/(nrow(data) * ncol(data)))*100,2), "% of data as NA)", sep=""))

#If the dependent variable is a factor, this prints out a legend to follow for the results
if (class(select(data, all_of(predictVar))[,1]) == "factor"){
  print("------------Dependant Variable Legend (Use to understand cluster models)------------")
  factorNames = levels(select(data, all_of(predictVar))[,1])
  facLegend = as.data.frame(factorNames)
  facLegend$values = as.numeric(unique(select(data, all_of(predictVar))[,1]))
  print(facLegend)
}

#This method finds all the columns with non-numeric data and re-codes them into dummy columns
print("------------Data Type Checks------------")
badDataType = sapply(data, function(x) typeof(x) %in% c("double", "integer")
)
badTRM = names(which(badDataType == F))
print(paste(length(badTRM), "variables recoded since all their entries aren't numeric or NA"), sep=
print("NOTE: algorithim recodes categorical data alphabetically e.g (female = 1, male = 2)")
print(badTRM)

#The below loop re-codes a column which has at least one string entry
#This works by identifying each unique level of the column, assigning them an integer based on
#alphabetical order then assign these values as the integers.
#e.g a column containing: "male, "female", "female", NA
#will be re-coded into: "2", "1", "1", NA
for (i in badTRM) { data[[i]] = as.factor(data[[i]]) }

#We remove all rows where there's a NA in the predictor variable, this is a pair-wise deletion
#we do this here so that we potentially keep more variables in the 'Low Data Removal' step
#A user can be recommended to manually remove these rows to keep them for potential test cases
data.rm = data[!is.na(data[[predictVar]]),]

#get the count of NA's per variable
print("------------Low Data Removal------------")
naCount = sapply(data.rm, function(x) sum(is.na(x)))

#if we don't do this step, when we do a listwise deletion we will most likely remove all rows
naRM = names(which(naCount >= nrow(data.rm)*naPercent)) #usage of user input here
print(paste(length(naRM), "variables removed since they had >= 'naPercent' (default 20%) NA values
print(naRM)

#remove the said attributes from data
data.rm = data.rm[,!(names(data.rm) %in% naRM)]

#before potentially running imputation, since we impute with CART,
#we want to add in the user input for how many observations are needed for a node split
rpart.control(minsplit = cartSplit)

if (impFlag){

  #creating the imputation
  print("------------CART Single Imputation------------")

  #a quirk of the function means that we can't have a variable name as 'next', which w2indresp has
  #(it thinks its a function call). We need to rename all variables called 'next' as 'nextVar'
```

```
  names(data.rm)[names(data.rm) == "next"] = "nextVar"

  #Since we impute before correlation checks, the data used may be a computationally singular matr
  #of MICE won't work, therefore we need to use the 'CART' method which is classification and regre
  #Since using multiple imputations can get quite hard, we use a single imputation created over mu
  data.imp = mice(data.rm,
                  #quirk of the MICE function, doesn't accept a seed of NULL, therefore we generate
                  seed = ifelse(is.null(randomSeed),round(runif(1, min=1, max=100000),0),randomSeed
                  m = 1,
                  maxit = 5,
                  method = "cart",
                  threshold = 2) #This threshold means we keep in all the highly correlated variab

  #This makes a data-set with the respective NA values computed
  data.imp.data = complete(data.imp,1)

  #rename the data-set so that we can use it going forward in the code
  data.recode = data.imp.data
} else{
  #list-wise deletion of all dummy NA's
  data.recode = na.omit(data.rm)
}

#now we need to do a check for the observation to variable ratio since we may have
#low observations after list-wise deletion
if (nrow(data.recode)/ncol(data.recode) < 5){
  stop(paste("Observation to Variable ratio of (",round(nrow(data.recode)/ncol(data.recode),2),") 
              "\n-Increase the function variable: 'naPercent'\n-Run missing value imputation by 

}

#now we need to encode our continuous and categorical variables
#first we need to encode categorical as factors and normalize the continuous variables
#This loop will change variables with less than 'catLevels' levels into a factor otherwise normaliz
#We also don't factorize the predictor variable since this will cause issues when making dummies
#we also don't scale the predictor variable IF it's categorical (would normally be a factor)
for (i in 1:length(data.recode)) {

  #This check is needed if the DV is a factor or the variable has more levels than catLevels, turns
  if (names(data.recode)[i] == predictVar | length(unique(data.recode[,i])) > catLevels) {
    data.recode[,i] = as.numeric(data.recode[,i])
  }

  #This recodes a column into a factor
  if (length(unique(data.recode[,i])) <= catLevels & names(data.recode)[i] != predictVar) {
    data.recode[,i] = as.factor(data.recode[,i])
  }

}

#loop goes though each columns, makes a freq table of the levels
print("----------Low Level Removal----------")
print("If a level is removed from a variable you wish to keep, reccomended to manually merge levels

#needed for console outputs
varCounter = 0
levelCounter = 0
obsCounter = 0

#start of loop
for (i in names(data.recode)) {

  #checks if variable is a factor that is does have at least 1 level with less than obsPerLevel ob
  if (is.factor(data.recode[[i]]) & !is_empty(names(which(table(data.recode[[i]])<obsPerLevel)))) 

    varCounter = varCounter + 1

    #looping though the freq table to find levels with less than obsPerLevel observations
```

```
    for (j in names(which(table(data.recode[[i]])<obsPerLevel)) ) {

      #console output
      print(paste("level",j,"in",i,"removed,",table(data.recode[[i]])[j],"observations found"))

      #before removing, we need to update the total obs removed count
      obsCounter = obsCounter + ifelse(!is.na(table(data.recode[[i]])[j]),table(data.recode[[i]])[j

      #removing the level, list-wise deletion
      data.recode = data.recode[data.recode[[i]] != j,]

      #counting the number of levels
      levelCounter = levelCounter + 1

    }

  }

}

#overall results from procedure
print(paste(levelCounter,"total levels removed from",varCounter,"different variables. In total",obs

#Check here since we may or removed enough observations to deem the ratio poor
if (nrow(data.recode)/ncol(data.recode) < 5){
  stop(paste("Observation to Variable ratio of ",round(nrow(data.recode)/ncol(data.recode),2)," is
             "\n-Increase the  function variable: 'naPercent'\n-Lower the value of function variab

}

#use the function to clean 0 var variables
data.recode.2 = var0Remove(data.recode)

#Here we need to check if the dependent variable is still apart of the data (if it's not it's becau
if (sum(names(data.recode.2) == predictVar) == 0){
  paste("Dependant Variable no longer has variance (all values the same). Consider the following:"
        "\n-Lower the value of function variable: 'obsPerLevel'\n-Gather more observations for the
}

#There are a few more steps before modelling, however we shouldn't apply these next steps to the D
#set the Dependent Variable
y = select(data.recode.2,all_of(predictVar))

#all the rest/all the predictor variables
data.recode.2 = select(data.recode.2,-all_of(predictVar))

#next we need to encode the categorical/factor variables as multiple columns of 0's and 1's to work
#the above step makes k-1 columns if a categorical variable has k levels (thanks to fullRank = T)
print("------------Dummy Variables------------")
dmy = dummyVars(" ~ .", data = data.recode.2, fullRank = T)
data.recode.dmy = data.frame(predict(dmy, newdata = data.recode.2))
print(paste("predictor variable count went from",ncol(data.recode.2),"to",ncol(data.recode.dmy)))

#After encoding the dummies, we need to remove 0 var variables again since some of the levels descr
#may no longer exist (which is what the dummy function uses to dumm-ize the data) therefore we end
data.clus = var0Remove(data.recode.dmy)

###K-Means###
#we next do K-Means since removing correlated variables and scaling negatively effect this model
#we use all the data (both train and test) excluding the dependant variable
print("------------K-Means------------")
set.seed(randomSeed) #need to set random seed here to ensure same results if set
kmClus = kmeans(data.clus,clusterAmount,1000)

#Tell user how many clusters have been made
print(paste(clusterAmount,"clusters have been made for K-Means"))

#generate the results table as a data.frame
```

```
resultsTable = as.data.frame(table(kmClus$cluster,as.matrix(y)))
names(resultsTable) = c("cluster","level","freq")
resultsTable$cluster = as.numeric(levels(resultsTable$cluster))[resultsTable$cluster] #needed to m
resultsTable$level = as.numeric(levels(resultsTable$level))[resultsTable$level]
resultsTable = resultsTable[order(resultsTable$cluster,resultsTable$level),]
resultsTable = resultsTable[order(resultsTable$cluster,resultsTable$level),]

#print table
print("K-Means results as a table, the max value in each row is a simple way to define which cluste
print(table(kmClus$cluster,as.matrix(y)))

#Here we are calculating the accuracy of the clusters, how a cluster is identified is:
#Whatever the current cluster has the most values of included within, it's defined as the cluster
#example:          y.var1    y.var2
#example: clus1 |    100   |    50    |
#example: clus1 is defined as the cluster for y.var1
clusterIden = 0
for (i in unique(resultsTable$cluster)){
  temp = subset(resultsTable,resultsTable$cluster == i)
  clusterIden = clusterIden + max(temp$freq)
}

#report back the in-cluster and between cluster SSR, along with each clusters size
print("CAUTION: Be careful comparing the MSE of this classification model to the regression models'
print(paste("Cluster ",1:length(kmClus$withinss),": Within MSE ",round(kmClus$withinss/kmClus$size
print(paste("Total between cluster MSE: ",round(kmClus$betweenss/length(kmClus$cluster),0),", Tota

#calculate the accuracy of the k-means using the sum of those maxes divided by total sum
print(paste("The K-Means model predicts exactly with an accuracy of",round(clusterIden/sum(resultsT

#Correlation Checks
print("-----------Correlation Checks-----------")

#creating a version of the data to remove the vars from
data.corr.rm = data.clus

#first, we make our initial correlation matrix
corr = round(cor(data.corr.rm),4)
diag(corr) = 0

#we then make a version which is just the lower tri correlation coefs
corrLower = corr
corrLower[upper.tri(corrLower)] = 0

#to preform our correlation checks, we do a loop that checks for the variable which has the most c
#to start it off we check if there is at all any correlation between the variables
corrCountUp = as.data.frame(corrLower[,apply(corrLower, 2, function(x) any(abs(x) > corrConfLevel)

#then append the answer in a list like so, this makes our while loop initialize
#if length(corrCountUp) is initially 0, then after the below command, length(corrCountUp) is 1
#if length(corrCountUp) is 1+, then after the below command, length(corrCountUp) is 2
corrCountUp = c(corrCountUp,1)

#beginning of the selection loop
iterC = 0
while (length(corrCountUp)-1 > 0) {
  #0.80-1.00: very strong correlation, by default we remove correlations above the 0.8 level (this
  #to deal with transitive co-linearity, we remove variables in order of the amount of coefficients
  varsRemoval = as.data.frame(corr[,apply(corr, 2, function(x) any(abs(x) > corrConfLevel))])

  #To deal with the non-transitive correlations, we sum the results from counting:
    #correlation coefs on the full correlation matrix
    #correlation coefs on the lower tri correlation matrix
  #This gives us a results which finds variables with the most correlations and discriminates agai
  #earlier in the dataset (which allows the user to have some manual control over what variables a

  #next, we get varsRemoval but with the upper tri removed
  varsRemovalUp = as.data.frame(corrLower[,apply(corrLower, 2, function(x) any(abs(x) > corrConfLe
```

```
  #we need this check for removing the last coefficient, quirk of the functions
  if (ncol(varsRemovalUp) == 1){
    #what we are doing here is getting the name of the column which has the correlation coef above
    names(varsRemovalUp) = names(apply(corrLower, 2, function(x) any(abs(x) > corrConfLevel))[apply
  }

  #We then get the amount of coefficients that are above the 'corrConfLevel' level per variable (K
  #variables that are at the start of the data-set to provide users a manual way to choose variable
  #which is moving them to the end of the data-set)
  corrCount = apply(varsRemoval, 2, function(x) sum(abs(x) > corrConfLevel))

  #do the same for varsRemovalUp
  corrCountUp = apply(varsRemovalUp, 2, function(x) sum(abs(x) > corrConfLevel))

  #now we sum corrCount and corrCountUp to get our final counts/scores for correlated variables
  for (i in names(corrCountUp)){
    if (i %in% names(corrCount)){
      corrCountUp[[i]] = corrCountUp[[i]] + corrCount[[i]]
    }
  }

  #this is the variable to be removed (if there are multiple at the max value, we select the varial
  colRemoved = corrCountUp[corrCountUp == max(corrCountUp)][1]

  #remove this column from the data.clus.rm, full correlation matrix and lower tri correlation mat
  data.corr.rm = data.corr.rm[,colnames(data.corr.rm) != names(colRemoved)]
  corr = corr[,colnames(corr) != names(colRemoved)]
  corrLower = corrLower[,colnames(corrLower) != names(colRemoved)]

  #progress marker
  iterC = iterC + 1
  print(paste(names(colRemoved),"removed, correlated with",corrCount[[names(colRemoved)]], "other 

  #to cut run-time, we only process a new correlation matrix when all variables with the same amou
  #highly correlated coefficients in the full+lower (example: all variables with 4 highly correlate
  #removed, we re-process the matrix to calculate all those with 3 and below)
  #We also don't process the matrix after we have removed the last variable
  if (length(corrCountUp[corrCountUp == max(corrCountUp)]) == 1 & ncol(varsRemovalUp) != 1) {
    corr = round(cor(data.corr.rm),4)
    diag(corr) = 0
    corrLower = corr
    corrLower[upper.tri(corrLower)] = 0
  }

}

#printing the results
print(paste(iterC,"variables removed since they had high correlation coefs"))

#This re-codes a column into a scalar (scales the continuous variables)
for (i in names(data.corr.rm)) {

  #this scales continuous variables, needed mainly for regression models
  #we can use 2 here since all factor variables will of been encoded as dummy columns by this stage
  if (length(unique(data.corr.rm[[i]])) > 2 ) {
    data.corr.rm[[i]] = as.vector(scale(data.corr.rm[[i]]))
  }

}

#since this is our last data transform, we should give the user a message stating that the clean d
print("The final cleaned dataset has been completed at this stage and is stored under the name 'cle

#finally, we now have a set of predictor variables ready to go into modelling
x = data.corr.rm

###PCA transformation###
```

```
if (PCAFlag){
  print("--------------PCA Transformation--------------")

  #preform PCA on remaining data set
  x.pca = princomp(x)

  #summary stats
  print(summary(x.pca))

  #next, we calculate the cumulative variance seen in the summary
  cutoff = cumsum(x.pca$sdev^2 / sum(x.pca$sdev^2))

  #This grabs all Comp's/PC's that are less than 'pcPercent' in cumulative variance explained.
  pcaToKeep = names(cutoff[cutoff<pcPercent])
  length(pcaToKeep)

  #ouput the number of Comps/PC's we are keeping
  print(paste("We are keeping up to Comp.",length(pcaToKeep)+1,"/PC",length(pcaToKeep)+1," since tl

  #now lets get all the PC's as a data frame (each column is a PC with the rotations as the rows)
  #we also clean the original pca to remove those PC's which are of little importance (cumlative p
  #we then set the original x as the PCA transformed data
  x = as.data.frame(x.pca$scores[,1:length(pcaToKeep)+1])
}

#split the remaining data in train and test sets
#This loop is need as sometimes the split can make a column have 0 variance therefore we keep tryin
#until we get on where the variance of each column can still be observed
print("--------------Attempting a Train Test Split--------------")
splitLoop = 0
splitLimit = 1000
randomSeedIter = randomSeed
while (splitLoop<splitLimit) {

  #we can set the user assigned seed if they want to keep the split consistent
  #we have the if statement so that if the seed provided is a bad split, it only runs the loop once
  if (!is.null(randomSeedIter)){
    set.seed(randomSeedIter)
    randomSeedIter = randomSeedIter + 1
  }

  #sample the data into a train and test split, if testPercent is -1, we only get one test value
  if (testPercent == -1) {s = sample(nrow(data.corr.rm),1)} else {s = sample(nrow(data.corr.rm),rou

  #check that variance is still seen within the training variables
  tempTrain = var0Remove(x[-s,], print = F)

  #If our training data still has variance, end the loop
  if (ncol(x) == ncol(tempTrain)){
    print("Good train, test split found")
    splitLoop = splitLimit*2
  } else {splitLoop = splitLoop + 1}

}

#In the case where a good train test split can't be found, we must throw an error
if (splitLoop >= splitLimit & splitLoop < 1500) {
  stop("No good Train Test split was found, please try increasing 'obsPerLevel', 'naPercent' or cha
}

#report back the seed if the given seed didn't work
if (!is.null(randomSeedIter)) {print(paste("The working seed found was",randomSeedIter - 1))}

#now we have a stable split, we can split the y and x data
y.test = as.matrix(y[s,])
y.train = as.matrix(y[-s,])

x.test = as.matrix(x[s,])
```

```
x.train = as.matrix(x[-s,])

###kNN###
print("--------------kNN--------------")

#Here we make the kNN model, takes train and test data. Function uses the train data so that for ea
#kNNCount train neighbors are considered for classification
set.seed(randomSeed) #need to set seed here so that the kNN choose the same neighbours if set
prT = knn(x.train, x.test, cl=y.train, k=kNNCount)

#this is making the data which can be returned to the user to use for further kNN study
kNNTrainData = as.data.frame(cbind(y.train, x.train))
names(kNNTrainData)[1] = "y"

#now lets make a table between the kNN results and the y.test data to see how well it has predictec
#we rename the data here just for the data output, not the best way since it creates new variables
predicted = prT
real = y.test
kNNTablePre = table(predicted,real)

#storing as data.frame for output
resultsDataF = as.data.frame(kNNTablePre)
names(resultsDataF) = c("predicted","real","freq")
resultsDataF$predicted = as.numeric(levels(resultsDataF$predicted))[resultsDataF$predicted] #neede
resultsDataF$real = as.numeric(levels(resultsDataF$real))[resultsDataF$real]

#getting the error per prediction
resultsDataF$error = (resultsDataF$predicted - resultsDataF$real)*resultsDataF$freq

#output the k in kNN (the number of neighbours)
print(paste(kNNCount,"neighbours considered for each test data point"))

#output results table
print("kNN results as a table, follow the diagonal for the correctly mapped clusters")
print(kNNTablePre)

#get the MSE of the predictions
print("CAUTION: Be careful comparing the MSE of this classification model to the regression models'
print(paste("The MSE of the predicted values are of",round(sum(resultsDataF$error^2)/sum(resultsDat

#generate the accuracy, done this way because the diag of the predictiosn table sometimes doesn't s
#predictions
accuracyS = 0
for (i in 1:nrow(resultsDataF)){
  if (resultsDataF$error[i] == 0){ accuracyS = accuracyS + resultsDataF$freq[i]}
}

#print results
print(paste("The kNN model predicts exactly with an accuracy of",round(accuracyS/sum(resultsDataF$

###CART###
print("--------------CART prediction model--------------")
#Here we fit our CART model and print the summary
x.data.cart = x.train
cartModel = rpart(y.train ~ x.data.cart)

#here we remove the data-frame name from the variables
cartModel$frame$var = str_remove(cartModel$frame$var,"x.data.cart")
names(cartModel$variable.importance) = str_remove(names(cartModel$variable.importance),"x.data.cart

#output the CART model
print(cartModel)

#output the variance importance
print("Variable Importance")
print(cartModel$variable.importance)

#quirk of the prediction function, need to rename the data-frame to the name used to build the mod
```

```
x.data.cart = x.test

#here we get the predictions using the model
predictionsCART = as.data.frame(predict(cartModel, as.data.frame(x.data.cart), interval = 'confiden
names(predictionsCART) = c("fit")

#add in the real y
predictionsCART$real = y.test

#Add in the error between the predicted and real
predictionsCART$error = predictionsCART$fit - predictionsCART$real

#calculations on the MSE of the model, ask about which one is best
print(paste("The MSE of the predicted values are of",round(mean(predictionsCART$error^2),4),sep=" "

#accuracy based on exact prediction rounded
exactAcc = sum(round(predictionsCART$fit,0) == predictionsCART$real)/length(predictionsCART$fit)
print(paste("The CART model predicts exactly with accuracy of",round(exactAcc,4),sep=" "))

###Ordinary Linear Regression###
print("-------------Ordinary Linear Regression (Initial)-------------")

#First lets do a basic multivariate Ordinary Linear Regression
#making our initial regression model
multiModel = lm(y.train ~ x.train)

#removing the matrix name from the coef names
names(multiModel$coefficients) = str_replace(names(multiModel$coefficients),"x.train","")

#For now we will remove any leftover NA vars from the model since we have done checks for all the a
multiModel.data = tidy(multiModel)
multiModel.data = na.omit(multiModel.data[-1,]) #removes the intercept and NA predictors from the

#removes the NA vars from x.data
x.data.linear = subset(x.train, select = multiModel.data$term)
x.test.clean = subset(x.test, select = multiModel.data$term)

#for the VIF functions to work, we need our dependent variable within our x.data dataframe
x.data.linear = cbind(x.data.linear,y.train)
dimnames(x.data.linear)[2][[1]][ncol(x.data.linear)] = "y"

#repeat for the test data
x.test.clean = cbind(x.test.clean,y.test)
dimnames(x.test.clean)[2][[1]][ncol(x.test.clean)] = "y"

#get the AIC of the full model
print(paste("The full model AIC is:",round(AIC(multiModel),4)))

#need this to store the length of the data-set, needed in backwards selection
lengthT = nrow(multiModel.data)

#Now to do a VIF check on the variables in the OLS model
print("-------------Variance Inflation Factor Removal-------------")
iterV = 0

#need to make the initial vif scores of the full model
vifModel = vif(lm(y ~ ., data = as.data.frame(x.data.linear)))

#if the VIF can't be calculated, we throw an error
if(is.na(max(vifModel))) {
  stop("VIF can't be calculated since there is too much multi-collinearity in the model, please lov
}

#here we have a loop where we remove variables based on maximum VIF score until all are removed
while (max(vifModel) > vifSelectionLevel){

  #get the name of the column that should be removed
  colRemoved = vifModel[vifModel == max(vifModel)]
```

```
  #update the length stored
  lengthT = lengthT - 1

  #remove this column from the x.train data
  x.data.linear = x.data.linear[,colnames(x.data.linear) != names(colRemoved)]

  #we may have to transpose and cast as matrix since if testPercent = -1, having 1 observation cha
  if (nrow(x.test.clean) == 1) { x.test.clean = t(as.matrix(x.test.clean[,colnames(x.test.clean) !=
    x.test.clean = x.test.clean[,colnames(x.test.clean) != colRemoved]
  }

  #progress marker
  iterV = iterV+1
  print(paste("The variable", names(colRemoved),"was removed since it had a VIF score of", round(co

  #generate the VIF scores for the full model
  vifModel = vif(lm(y ~ ., data = as.data.frame(x.data.linear)))

}

#print results from the VIF stage
print(paste(iterV,"variables removed from the Ordinary Linear Model since they have a VIF score hi

#have a look at the model AIC after VIF checks
print(paste("The full model AIC after VIF checks is:",round(AIC(lm(y ~ ., data = as.data.frame(x.d

#setting up progress marker
print("----------Backwards Selection----------")
iterA = 0

#backwards selection
#our method: remove least significant variable, re-make model, repeat. Stopping criteria: all vari

#storing the columns removed in data-frame for output later
rmNames = c()

while (max(multiModel.data$p.value) > (1 - modelSigLevel) ) {
  #get the column name of the variable with the least correlation/ highest p-value
  colRemoved = str_remove(toString(multiModel.data[which.max(multiModel.data$p.value),1]),"x.data.

  #adding the name to the vector for output
  rmNames = c(rmNames,colRemoved)

  #remove this column from the x.train data
  x.data.linear = x.data.linear[,colnames(x.data.linear) != colRemoved]
  if (nrow(x.test.clean) == 1) { x.test.clean = t(as.matrix(x.test.clean[,colnames(x.test.clean) !=
    x.test.clean = x.test.clean[,colnames(x.test.clean) != colRemoved]
  }

  #re-fit the model
  multiModel = lm(y ~ ., data = as.data.frame(x.data.linear))

  #removing the matrix name from the coef names
  names(multiModel$coefficients) = str_replace(names(multiModel$coefficients),"x.data","")
  multiModel.data = tidy(multiModel)

  #We don't want to check the intercept for this backward selection so we remove from the list of 
  multiModel.data = multiModel.data[-1,]

  #progress marker
  iterA = iterA+1

  #print progress every 50 variables
  if ((iterA %% 50) == 0){
    print(paste(toString(iterA)," out of ",toString(lengthT)," variables removed so far.",sep=""))
  }
```

```
}

#report back the results of the backwards selection
print(paste(toString(iterA)," out of ",toString(lengthT)," variables removed in backwards selection

#output names of the regression
print(rmNames)

print("-----------Ordinary Linear Regression (Improved)-----------")
#set and have a look at the final linear model
multiModelFinal = multiModel
print(summary(multiModelFinal))
cat(paste("AIC:",round(AIC(multiModelFinal),4)))
cat(paste("\nMSE:",round(mean(multiModelFinal$residuals^2),4),"\n"))

#now we can make some predictions using the test data
#the below step is needed just so we can use the predict function (quirk of the func)
x.data.linear = x.test.clean
colnames(x.test.clean) = colnames(x.data.linear)

#Now we can put the predictions into a data-frame
predictionsLinear = as.data.frame(predict(multiModelFinal, as.data.frame(x.data.linear), interval =

#add in the real y
predictionsLinear$real = y.test

#Add in the error between the predicted and real
predictionsLinear$error = predictionsLinear$fit - predictionsLinear$real

#calculations on the MSE of the model, ask about which one is best
print(paste("The MSE of the predicted values are of",round(mean(predictionsLinear$error^2),4),sep=

#accuracy based on exact prediction rounded
exactAcc = sum(round(predictionsLinear$fit,0) == predictionsLinear$real)/length(predictionsLinear$
print(paste("The Linear Model predicts exactly with accuracy of",round(exactAcc,4),sep=" "))

#accuracy based on being within the confidence interval rounded
intervalAcc = sum(predictionsLinear$real >= round(predictionsLinear$lwr,0) & predictionsLinear$rea
print(paste("The Linear Model predicts within a confidence interval with accuracy of",round(interva

###Elastic Net Regression (includes Ridge and Lasso in theory)###
print("-----------Elastic Net Regression-----------")
#Now we are going to loop though different values of alpha to find our best fit
#alpha = 0 is ridge regression
#alpha = 1 is lasso regression
#0 < alpha < 1 is elastic net

#the below loop creates the fits for each alpha (decided by the elasticCount)
list.of.fits = list()

#if wanted, we can add a progress bar to the fitting procedure like so
#glmnet.control(itrace = 0)

#loop for fitting all models for each alpha
for (i in 0:elasticCount) {

  fit.name = paste0("alpha", i/elasticCount)

  set.seed(randomSeed) #This is set here since the lambda value in each fold is choosen randomly

  #by default, we do 10-fold cross validation to get our lambda value for each fit
  list.of.fits[[fit.name]] = cv.glmnet(x.train, y.train, type.measure="mse", alpha=i/elasticCount,
}

#now we calculate the models per alpha value generated
#predicting the values in the Testing data-set.
results = data.frame()
for (i in 0:elasticCount) {
```

```r
  fit.name = paste("alpha", i/elasticCount,sep="")

  #use each model to predict 'y' given the test data set
  predicted = predict(list.of.fits[[fit.name]], s=list.of.fits[[fit.name]]$lambda.min, newx=x.test

  #calculate the Mean Squared Error of the predicted values
  mse = mean((y.test - predicted)^2)

  #calculate the R^2 value
  r2 = cor(y.test, predicted)^2

  #store the results
  temp = data.frame(alpha=i/elasticCount, mse=mse, r2=r2, fit.name=fit.name)
  results = rbind(results, temp)
}

#naming issue with results data-frame, have to have this one line to clear up
colnames(results)[3] = "R^2"

#view the results
results

#we can then pick the best model from the tests
#get the best fit name by lowest mse value (fits in line with how our models were fitted for each a
bestFitName = toString(results[which.min(results$mse),4])
bestFitMSE = toString(results[which.min(results$mse),2])
bestFitAlpha = toString(results[which.min(results$mse),1])

#coefficients of the last alpha value (so lasso since alpha = 1)
#To show only variables which were considered significant in the model, we have to do some transfor
#first getting the significant coefficients
elasSumm = summary(coef(list.of.fits[[bestFitName]]))

#then here we are getting the names of all the predictor variables considered for the model
namesHold = as.data.frame(dimnames(coef(list.of.fits[[bestFitName]]))[1])

#now a setup for a for loop that goes though and matches the index of names in 'namesHold' to the
names(namesHold)[1] = "names"
elasSumm$names = NA
names(elasSumm) = c("iter1","iter2","val","names")
iterB = 1
for (i in 1:nrow(namesHold)) {
  if (i == elasSumm$iter1[iterB]) {
    elasSumm$names[iterB] = namesHold$names[i]
    iterB = iterB + 1
    if (iterB > nrow(elasSumm)) #This if is needed since the iterator counts above the amount if la
      break
  }
}

#lastly, cleaning up the results then printing
elasSumm$Estimate_Coefs = elasSumm$val
elasSumm$iter1 = NULL
elasSumm$iter2 = NULL
elasSumm$val = NULL
print(elasSumm)

#MSE of the best fit
print(paste("The MSE of the predicted values of the best fit model is",round(as.numeric(bestFitMSE

#Alpha of the best fit
print(paste("The Alpha of the best fit model is",bestFitAlpha,sep=" "))

#get the best fit
bestFit = list.of.fits[[bestFitName]]

#generate a predict return sheet
predictElastic = as.data.frame(predict(bestFit, s=bestFit$lambda.min, newx = x.test))
```

```
  #changing name to fall in line with Ordinary Linear Regression predictions sheet
  names(predictElastic)[1] = "fit"

  #add in real values
  predictElastic$real = y.test

  #accuracy based on exact prediction rounded
  exactAccElastic = sum(round(predictElastic$fit,0) == predictElastic$real)/length(predictElastic$fit
  print(paste("The Elastic Net Model predicts exactly with accuracy of",round(exactAccElastic,4),sep=

  #now to make the return object, we have multiple things we want to return in a list
  returnList = list(data.corr.rm,
                    list(kmClus,resultsTable),
                    list(kNNTrainData, resultsDataF),
                    list(cartModel,predictionsCART),
                    list(multiModel, predictionsLinear),
                    list(results, bestFit, predictElastic))

  #setting the names of the list for better user reabale results
  names(returnList) = c("cleanData",
                        "kMeansResults",
                        "kNNResults",
                        "cartResults",
                        "linearRegression",
                        "elasticNetRegression")

  #setting inner names
  names(returnList$kMeansResults) = c("model","predictions")
  names(returnList$kNNResults) = c("trainingData","predictions")
  names(returnList$cartResults) = c("model","predictions")
  names(returnList$linearRegression) = c("model","predictions")
  names(returnList$elasticNetRegression) = c("listOfFits","model","predictions")

  #lastly, we stop the timer and report back the run times
  print("----------Timer Results----------")
  print(proc.time() - timer)

  return(returnList)

}
```

10.1.2 mainRun

```
#Here we begin to to a run though the methods made using the UKHLS data

#TODO:
  #1. Figure out what happened to the job data (Was BPHS data, can't use)
  #2. Finish off transforming all the data for UKHLS
  #2.2 generate some tests cases from outside UKHLS
  #2.3 investigate pairwise deletion and formalize for 'imp'
  #3. Consider automatic way to find encoded NA variables (such as -9,-8,-7,-2,-1)
  #4. VIR inflation factor needs to be included
  #5. Fix up the classification section to make more sense and good results
  #6. Question weather it's ok to remove extremely significant vars from regression model
  #7. work on representing the predictions made from the Ordinary Linear Regression model
  #10. document processes and procedures; MAKE CLEAR

##################################################
##################################################
####################START HERE####################
##################################################
##################################################

graphics.off()  # clear all graphs
rm(list = ls()) # remove all files from your workspace
```

```
library(readr) #used to read the tab files
library(dplyr) #the 'select()' function is used to extract some data outside of automodel
library(stringr) #package is used to clean variable names (str_replace)

#given our package made isn't on CRAN and we have it locally, we need to install from the zip file u
library(devtools)
install_local("automodel.zip")
library(automodel) #the function made during this dissertation, all code is within 'automodel.R'

#We have some functions which are used to process the ghq data separate from everything else
#This code can be found in 'ghq_functions.R'
source(paste(getwd(),"/rCode/ghq_functions.R",sep=""))

#getting the wave 2 data
w2income = read_table("rData/selected/b_income.tab") #Income and payment
w2indresp = read_table("rData/selected/b_indresp.tab") #All information from the individual question

#getting the wave 3 data
w3income = read_table("rData/selected/c_income.tab") #Income and payment
w3indresp = read_table("rData/selected/c_indresp.tab") #All information from the individual question

#getting Nurse visit mixed data
#'xindresp_ns.tab' is Wave 2 and 3 Nurse - individual respondents data
mixNurse = read_table(file = "rData/selected/xindresp_ns.tab")

#Blood sample from nurse visit, 2010-2011 collected, 2013 analyzed
mixBloodData = read_table(file = "rData/selected/xlabblood_ns.tab")

#re-code the NA values
w2income = recodeNA(w2income)
w2indresp = recodeNA(w2indresp)
w3income = recodeNA(w3income)
w3indresp = recodeNA(w3indresp)
mixNurse = recodeNA(mixNurse)
mixBloodData = recodeNA(mixBloodData)

#we have to remove the GHQ related variables from the nurse data to avoid duplicate predictor column
#making sure we are only getting relevant columns with the function
names(select(mixNurse,contains("ghq")))

#removing the questions and making a new dataset for it
joinSurveyData = select(mixNurse,-contains("ghq"))

#next, lets create all the joins between the data we want to view
#per wave 2, wave 3 and mixed we want:
  #-all participants -> all nurse visits -> all nurse visit blood samples

#first, we need to remove some wave 2 & 3 naming conventions to assist when merging
names(w2income) = str_replace(names(w2income),"b_","")
names(w2indresp) = str_replace(names(w2indresp),"b_","")

names(w3income) = str_replace(names(w3income),"c_","")
names(w3indresp) = str_replace(names(w3indresp),"c_","")

#Having a quick investigation into the differences in variables cross wave
names(w2indresp)[names(w2indresp) %in% names(w3indresp)]
length(names(w2indresp)[names(w2indresp) %in% names(w3indresp)]) #number of shared variables
length(names(w2indresp)[!names(w2indresp) %in% names(w3indresp)]) #number of variables in w2indresp
length(names(w3indresp)[!names(w3indresp) %in% names(w2indresp)]) #number of variables in w3indresp

#adding a wave column to the survey data for joining
w2indresp$wave = 2
w3indresp$wave = 3

#we can then make a dataset that is all the variables in both wave 2 and 3 in one dataset
w2Shared = w2indresp[,names(w2indresp)[names(w2indresp) %in% names(w3indresp)]]
w3Shared = w3indresp[,names(w2indresp)[names(w2indresp) %in% names(w3indresp)]]
```

```
#creation of stacked data-set
wShared = rbind(w2Shared,w3Shared)

#next we have to join the survey data and income data together per wave 2 and 3, this requires some S
#merge income with nurse visit data (Note this use of sqlTransform here)
w2Merge = merge(w2indresp,sqlTransform(w2income), by = "pidp")

#now doing the same for wave 3
w3Merge = merge(w3indresp,sqlTransform(w3income), by = "pidp")

#Preforming the same investigation but for our transformed data w2Merge, w3Merge
length(names(w2Merge)[names(w2Merge) %in% names(w3Merge)]) #number of shared variables (tells us that
length(names(w2Merge)[!names(w2Merge) %in% names(w3Merge)]) #number of variables in w2indresp not in
length(names(w3Merge)[!names(w3Merge) %in% names(w2Merge)]) #number of variables in w3indresp not in

#we can make a versions of the data which only includes the shared variables which we can analyze
w2SMerge = w2Merge[,names(w2Merge)[names(w2Merge) %in% names(w3Merge)]]
w3SMerge = w3Merge[,names(w2Merge)[names(w2Merge) %in% names(w3Merge)]]

#now to stack the two datasets/union them to get the income version
wSMerge = rbind(w2SMerge,w3SMerge)

#Lastly, we can now make all the joins between the relevant wave data. We are only going to use INNER
#(reason is because analyze as clusters, LEFT JOIN would leave NA's for alot, doesn't share fair com
w2MergeNurse = merge(x = w2Merge, y = joinSurveyData, by = c("pidp","wave")) #w2Merge with nurse vis
w2MergeNurseBlood = merge(x = w2MergeNurse, y = mixBloodData, by = c("pidp","wave")) #w2Merge with bl

w3MergeNurse = merge(x = w3Merge, y = joinSurveyData, by = c("pidp","wave")) #w3Merge with nurse vis
w3MergeNurseBlood = merge(x = w3MergeNurse, y = mixBloodData, by = c("pidp","wave")) #w3Merge with bl

wSMergeNurse = merge(x = wSMerge, y = joinSurveyData, by = c("pidp","wave")) #wSMerge with nurse vis
wSMergeNurseBlood = merge(x = wSMergeNurse, y = mixBloodData, by = c("pidp","wave")) #wSMergeNurse w

mixNurseBlood = merge(x = mixNurse, y = mixBloodData, by = c("pidp","wave")) #nurse and blood data
#Note: the difference in mixNurseBlood and wSMergeNurseBlood observations is the number of participan
#but didn't answer the income questionnaire

#Since in this run we are wanting to look at predicting ghq outcome, we need to do some initial anal
#first, lets compare all the base datasets the comprise all of our joins
plot(density(na.omit(w2indresp$scghq1_dv)), col = "firebrick",
     xlab = "GHQ Score",
     main = "GHQ Score Compared (Base Data-sets)",
     ylim = c(0.00,0.11))
lines(density(na.omit(w3indresp$scghq1_dv)), col = "cyan")
lines(density(na.omit(mixNurse$scghq1_dv)), col = "purple")
legend("topright",legend = c("Wave 2 (b_indresp)","Wave 3 (c_indresp)","Nurse Visit (xindresp_ns)"),
       fill = c("firebrick","cyan","purple"),
       cex = 0.99)

#lets do some t.tests and f.tests on the data
#first, wave 2 and wave 3
t.test(w2indresp$scghq1_dv, w3indresp$scghq1_dv)
var.test(w2indresp$scghq1_dv, w3indresp$scghq1_dv)

#next, wave 2 and nurse visit
t.test(w2indresp$scghq1_dv, mixNurse$scghq1_dv)
var.test(w2indresp$scghq1_dv, mixNurse$scghq1_dv)

#lastly, wave 3 and nurse visit
t.test(w3indresp$scghq1_dv, mixNurse$scghq1_dv)
var.test(w3indresp$scghq1_dv, mixNurse$scghq1_dv)

#now we are going to compare the distributions of the GHQ score per merge
par(mfrow=c(2,2))

#plotting the wave 2 dataset joins
plot(density(na.omit(w2indresp$scghq1_dv)), col = "firebrick",
```

```
      xlab = "GHQ Score",
      main = "GHQ Score Compared (Wave 2)",
      ylim = c(0.00,0.11))
lines(density(na.omit(w2Merge$scghq1_dv)), col = "cyan")
lines(density(na.omit(w2MergeNurse$scghq1_dv)), col = "palegreen")
lines(density(na.omit(w2MergeNurseBlood$scghq1_dv)), col = "purple")
legend("topright",legend = c("w2indresp","w2Merge","w2MergeNurse","w2MergeNurseBlood"),
        fill = c("firebrick","cyan","palegreen","purple"),
        cex = 0.7)

#plotting the wave 3 dataset joins
plot(density(na.omit(w3indresp$scghq1_dv)), col = "firebrick",
      xlab = "GHQ Score",
      main = "GHQ Score Compared (Wave 3)",
      ylim = c(0.00,0.11))
lines(density(na.omit(w3Merge$scghq1_dv)), col = "cyan")
lines(density(na.omit(w3MergeNurse$scghq1_dv)), col = "palegreen")
lines(density(na.omit(w3MergeNurseBlood$scghq1_dv)), col = "purple")
legend("topright",legend = c("w3indresp","w3Merge","w3MergeNurse","w3MergeNurseBlood"),
        fill = c("firebrick","cyan","palegreen","purple"),
        cex = 0.7)

#plotting the shared dataset joins
plot(density(na.omit(wShared$scghq1_dv)), col = "firebrick",
      xlab = "GHQ Score",
      main = "GHQ Score Compared (Wave 2 & 3)",
      ylim = c(0.00,0.11))
lines(density(na.omit(wSMerge$scghq1_dv)), col = "cyan")
lines(density(na.omit(wSMergeNurse$scghq1_dv)), col = "palegreen")
lines(density(na.omit(wSMergeNurseBlood$scghq1_dv)), col = "purple")
legend("topright",legend = c("wShared","wSMerge","wSMergeNurse","wSMergeNurseBlood"),
        fill = c("firebrick","cyan","palegreen","purple"),
        cex = 0.7)

#plotting the mixed nurse and mix nurse blood data
plot(density(na.omit(mixNurse$scghq1_dv)), col = "firebrick",
      xlab = "GHQ Score",
      main = "GHQ Score Compared (Nurse Visits Only)",
      ylim = c(0.00,0.11))
lines(density(na.omit(mixNurseBlood$scghq1_dv)), col = "purple")
legend("topright",legend = c("mixNurse","mixNurseBlood"),
        fill = c("firebrick","purple"),
        cex = 0.7)

#reset graphing options
par(mfrow = c(1,1))

#from the above plot we can see that even those there is noticeable differences in distribution, we c
#that the distribution of sc_ghq1_dv is rather similar across all joins

#Now, we are investigating the allowable prediction range for GHQ
#Since a total score of GHQ 7 can be considered quite similar to a score of 8 or 6, when it comes to
#it would be worth seeing if the predicted values fall into a range instead of an exact value.

#to assist this calculation, we need some variables to hold to two GHQ columns
ghq1_dv = na.omit(w2indresp$scghq1_dv)
ghq2_dv = na.omit(w2indresp$scghq2_dv)

#amount of groups in the ghq score
grGHQ = length(unique(ghq2_dv))
grGHQ

#amount of total scores possible in ghq score
scGHQ = length(unique(ghq1_dv))
scGHQ

#estimate size of each group in ghq2
grSize = scGHQ / grGHQ
```

```
grSize

#how far we should look each side of the predicted values
grSide = grSize / 2
grSide

#to compare our prediction accuracy on this data, I want to compare to:
#randomly picking a GHQ score if all scores assumed equal
randomPickCh.g = 1/length(unique(ghq1_dv))
randomPickCh.g

#then the chance of always picking the most common score in scghq
mstComm.g = as.numeric(names(which(table(ghq1_dv) == max(table(ghq1_dv)))))
paste(”Mode/most frequent score is:”,mstComm.g)
modePickCh.g = max(table(ghq1_dv)) / length(ghq1_dv)
modePickCh.g

#The chance when randomly picking within a +- 1 range (so randomly picking 7 is correct if real is in
randomRangeCh.g = ((3*37)-2)/(length(unique(ghq1_dv))*37)
randomRangeCh.g

#The chance of picking within the largest +- 1 range (which is 6,7,8)
rangePickCh.g = sum(table(ghq1_dv)[c(mstComm.g + 1, mstComm.g + 2, mstComm.g + 3)]) / length(ghq1_dv
rangePickCh.g

#now let's investigate the distribution of the GHQ score and the correlations between the questions,
#after we've analyzed, we then remove GHQ questions and move some variables so that we get the best
#Then, we can run each version of the joined data though the function and review the results
####################################
#wave 2 all participants
####################################
#analyzing this merge
ghq_analyze(w2indresp)

#cleaning merge ready for analysis
w2indresp.rData = ghq_clean_move(w2indresp)

#set console output location (needed for write up, you can un-comment to store console output in a te
sink(”consoleOutput/w2indresp.txt”, type=c(”output”,”message”))

#generate models
w2indresp.r = autoModel(”total_score”, w2indresp.rData, randomSeed = 3)

#create the GHQ unique prediction ranges
predHold_w2indresp = predGHQadd(w2indresp.r,grSide)

####################################
#wave 2 all participants which provided income data
####################################
#analyzing this merge
ghq_analyze(w2Merge)

#cleaning merge ready for analysis
w2Merge.rData = ghq_clean_move(w2Merge)

#set console output location
sink(”consoleOutput/w2Merge.txt”, type=c(”output”,”message”))

w2Merge.r = autoModel(”total_score”, w2Merge.rData, randomSeed = 3)

#create the GHQ unique prediction ranges
predHold_w2Merge = predGHQadd(w2Merge.r,grSide)

####################################
#wave 2 all participants which provided income data and had a nurse visit
####################################
#analyzing this merge
ghq_analyze(w2MergeNurse)
```

```
#cleaning merge ready for analysis
w2MergeNurse.rData = ghq_clean_move(w2MergeNurse)

#set console output location
sink("consoleOutput/w2MergeNurse_1.txt", type=c("output","message"))

w2MergeNurse.r = autoModel("total_score", w2MergeNurse.rData, randomSeed = 3)

#running again due to error message of observation to variable ratio
sink("consoleOutput/w2MergeNurse_2.txt", type=c("output","message"))

w2MergeNurse.r = autoModel("total_score", w2MergeNurse.rData, randomSeed = 3, naPercent = 0.1)

#running again due to error message of VIF calculation
sink("consoleOutput/w2MergeNurse_3.txt", type=c("output","message"))

w2MergeNurse.r = autoModel("total_score", w2MergeNurse.rData, randomSeed = 3, naPercent = 0.1, corrC

#create the GHQ unique prediction ranges
predHold_w2MergeNurse = predGHQadd(w2MergeNurse.r,grSide)

####################################
#wave 2 all participants which provided income data, had a nurse visit and provided blood data
####################################
#analyzing this merge
ghq_analyze(w2MergeNurseBlood)

#cleaning merge ready for analysis
w2MergeNurseBlood.rData = ghq_clean_move(w2MergeNurseBlood)

#set console output location
sink("consoleOutput/w2MergeNurseBlood_1.txt", type=c("output","message"))

w2MergeNurseBlood.r = autoModel("total_score", w2MergeNurseBlood.rData, randomSeed = 3)

#re-run since original run threw Observation to variable Ratio error
sink("consoleOutput/w2MergeNurseBlood_2.txt", type=c("output","message"))

w2MergeNurseBlood.r = autoModel("total_score", w2MergeNurseBlood.rData, randomSeed = 3, naPercent = 

#create the GHQ unique prediction ranges
predHold_w2MergeNurseBlood = predGHQadd(w2MergeNurseBlood.r,grSide)

####################################
#wave 3 all participants
####################################
#analyzing this merge
ghq_analyze(w3indresp)

#cleaning merge ready for analysis
w3indresp.rData = ghq_clean_move(w3indresp)

#set console output location
sink("consoleOutput/w3indresp_1.txt", type=c("output","message"))

#first attempt, throws an error saying the VIF can't be calculated, we adjust the corrConfLevel acco
w3indresp.r = autoModel("total_score", w3indresp.rData, randomSeed = 3)

#second run
sink("consoleOutput/w3indresp_2.txt", type=c("output","message"))

w3indresp.r = autoModel("total_score", w3indresp.rData, randomSeed = 3, corrConfLevel = 0.5)

#create the GHQ unique prediction ranges
predHold_w3indresp = predGHQadd(w3indresp.r,grSide)

####################################
```

```
#wave 2 all participants which provided income data
####################################
#analyzing this merge
ghq_analyze(w3Merge)

#cleaning merge ready for analysis
w3Merge.rData = ghq_clean_move(w3Merge)

#set console output location
sink("consoleOutput/w3Merge_1.txt", type=c("output","message"))

w3Merge.r = autoModel("total_score", w3Merge.rData, randomSeed = 3)

sink("consoleOutput/w3Merge_2.txt", type=c("output","message"))

w3Merge.r = autoModel("total_score", w3Merge.rData, randomSeed = 3, corrConfLevel = 0.5)

#create the GHQ unique prediction ranges
predHold_w3Merge = predGHQadd(w3Merge.r, grSide)

####################################
#wave 3 all participants which provided income data and had a nurse visit
####################################
#analyzing this merge
ghq_analyze(w3MergeNurse)

#cleaning merge ready for analysis
w3MergeNurse.rData = ghq_clean_move(w3MergeNurse)

#set console output location
sink("consoleOutput/w3MergeNurse_1.txt", type=c("output","message"))

w3MergeNurse.r = autoModel("total_score", w3MergeNurse.rData, randomSeed = 3)

#re run since 1st run threw observation to variable error
sink("consoleOutput/w3MergeNurse_2.txt", type=c("output","message"))

w3MergeNurse.r = autoModel("total_score", w3MergeNurse.rData, randomSeed = 3, naPercent = 0.009)

#re run since 2nd run threw observation to variable error
sink("consoleOutput/w3MergeNurse_3.txt", type=c("output","message"))

w3MergeNurse.r = autoModel("total_score", w3MergeNurse.rData, randomSeed = 3, naPercent = 0.009, obsl

#re run since 3rd run threw VIF error
sink("consoleOutput/w3MergeNurse_4.txt", type=c("output","message"))

w3MergeNurse.r = autoModel("total_score", w3MergeNurse.rData, randomSeed = 3, naPercent = 0.009, obsl

#create the GHQ unique prediction ranges
predHold_w3MergeNurse = predGHQadd(w3MergeNurse.r, grSide)

####################################
#wave 3 all participants which provided income data, had a nurse visit and provided blood data
####################################
#analyzing this merge
ghq_analyze(w3MergeNurseBlood)

#cleaning merge ready for analysis
w3MergeNurseBlood.rData = ghq_clean_move(w3MergeNurseBlood)

#set console output location
sink("consoleOutput/w3MergeNurseBlood_1.txt", type=c("output","message"))

w3MergeNurseBlood.r = autoModel("total_score", w3MergeNurseBlood.rData, randomSeed = 3)

#given 2nd threw observation to variable ratio error
sink("consoleOutput/w3MergeNurseBlood_2.txt", type=c("output","message"))
```

```
w3MergeNurseBlood.r = autoModel("total_score", w3MergeNurseBlood.rData, randomSeed = 3, naPercent = 0

#given 3rd threw observation to variable ratio error
sink("consoleOutput/w3MergeNurseBlood_3.txt", type=c("output","message"))

#NOTE: obsPerLevel = 2 wouldn't run, threw observation to ratio error
w3MergeNurseBlood.r = autoModel("total_score", w3MergeNurseBlood.rData, randomSeed = 3, naPercent = 0

#Given the above result, we are going to say that w3MergeNurseBlood has to little observations per v
#in it's current state and a manual selection process should be done on what variables are the most i
#in analyzing.

####################################
#the shared wave participants
####################################
#analyzing this merge
ghq_analyze(wShared)

#cleaning merge ready for analysis
wShared.rData = ghq_clean_move(wShared)

#set console output location
sink("consoleOutput/wShared.txt", type=c("output","message"))

wShared.r = autoModel("total_score", wShared.rData, randomSeed = 3)

#create the GHQ unique prediction ranges
predHold_wShared = predGHQadd(wShared.r, grSide)

####################################
#the shared wave participants which provided income data
####################################
#analyzing this merge
ghq_analyze(wSMerge)

#cleaning merge ready for analysis
wSMerge.rData = ghq_clean_move(wSMerge)

#set console output location
sink("consoleOutput/wSMerge_1.txt", type=c("output","message"))

wSMerge.r = autoModel("total_score", wSMerge.rData, randomSeed = 3)

#set console output location
sink("consoleOutput/wSMerge_2.txt", type=c("output","message"))

wSMerge.r = autoModel("total_score", wSMerge.rData, randomSeed = 3, corrConfLevel = 0.5)

#create the GHQ unique prediction ranges
predHold_wSMerge = predGHQadd(wSMerge.r, grSide)

####################################
#the shared wave participants which provided income data and had a nurse visit
####################################
#analyzing this merge
ghq_analyze(wSMergeNurse)

#cleaning merge ready for analysis
wSMergeNurse.rData = ghq_clean_move(wSMergeNurse)

#set console output location
sink("consoleOutput/wSMergeNurse.txt", type=c("output","message"))

wSMergeNurse.r = autoModel("total_score", wSMergeNurse.rData, randomSeed = 3)

#create the GHQ unique prediction ranges
predHold_wSMergeNurse = predGHQadd(wSMergeNurse.r, grSide)
```

```
####################################
#the shared wave participants which provided income data, had a nurse visit and provided blood data
####################################
#analyzing this merge
ghq_analyze(wSMergeNurseBlood)

#cleaning merge ready for analysis
wSMergeNurseBlood.rData = ghq_clean_move(wSMergeNurseBlood)

#set console output location
sink("consoleOutput/wSMergeNurseBlood_1.txt", type=c("output","message"))

wSMergeNurseBlood.r = autoModel("total_score", wSMergeNurseBlood.rData, randomSeed = 3)

#given 1st run the observation to variable ratio error
sink("consoleOutput/wSMergeNurseBlood_2.txt", type=c("output","message"))

wSMergeNurseBlood.r = autoModel("total_score", wSMergeNurseBlood.rData, randomSeed = 3, naPercent = 

#create the GHQ unique prediction ranges
predHold_wSMergeNurseBlood = predGHQadd(wSMergeNurseBlood.r,grSide)

####################################
#wave 2,3 all participants which had a nurse visit  (only nurse data)
####################################
#analyzing this merge
ghq_analyze(mixNurse)

#cleaning merge ready for analysis
mixNurse.rData = ghq_clean_move(mixNurse)

#set console output location
sink("consoleOutput/mixNurse.txt", type=c("output","message"))

mixNurse.r = autoModel("total_score", mixNurse.rData, randomSeed = 3)

#create the GHQ unique prediction ranges
predHold_mixNurse = predGHQadd(mixNurse.r,grSide)

####################################
#wave 2,3 all participants which had a nurse visit and provided blood data (only nurse & blood data)
####################################
#analyzing this merge
ghq_analyze(mixNurseBlood)

#cleaning merge ready for analysis
mixNurseBlood.rData = ghq_clean_move(mixNurseBlood)

#set console output location
sink("consoleOutput/mixNurseBlood.txt", type=c("output","message"))

mixNurseBlood.r = autoModel("total_score", mixNurseBlood.rData, randomSeed = 3)

#create the GHQ unique prediction ranges
predHold_mixNurseBlood = predGHQadd(mixNurseBlood.r,grSide)

#This needs to be ran to close all the sink connections
closeAllConnections()

########Titanic Run########
#Load in the Titanic data (notice we load in 'train.csv', that because the test data Kaggle provides
titanicData = read.csv("rData/TitanicData/train.csv")

#get the random chances of picking Survival we are wanting to beat with our models
#complete random
randomCh.t = 1/length(unique(titanicData$Survived))
randomCh.t
```

```
#mode selection
mstComm.t = as.numeric(names(which(table(titanicData$Survived) == max(table(titanicData$Survived)))))
paste("Mode/most frequent score is:",mstComm.t)
modePickCh.t = max(table(titanicData$Survived)) / length(titanicData$Survived)
modePickCh.t

#we can also make the predictor variable a factor, this generates a legend we can follow in the mode
titanicData$Survived = as.factor(ifelse(titanicData$Survived == 0,"Didn't Survive","Survived"))

sink("consoleOutput/Titanic.txt", type=c("output","message"))

#run the model (random seed 3 works as-well)
titanicData.r = autoModel("Survived", titanicData, randomSeed = 3)

#This needs to be ran to close all the sink connections
closeAllConnections()

#having a look at the results
titanicData.r$elasticNetRegression$listOfFits

#now lets do a run where we only use 1 test observation, used to submit best model to Kaggle
#To make sure we get the compatible results for Kaggle, we need to change back our Survived variable
#we can also make the predictor variable a factor, this generates a legend we can follow in the mode
titanicData$Survived = as.numeric(ifelse(titanicData$Survived == "Didn't Survive",0,1))

#running the kaggle submit model
titanicData.r.k = autoModel("Survived", titanicData, randomSeed = 3, testPercent = -1)

#The linear regression model provides the best predictions, therefore we will use this model to pred
#first, we need to lead in the test.csv data
TitanicTest = read.csv("rData/TitanicData/test.csv")

#now we get the linear model
TitanicLinear.k = titanicData.r.k$linearRegression$model

#now we generate predictions, do do this, we need to format the test data like the clean data from th
#Here's the variables we need in our test data (excluding y)
print(names(TitanicLinear.k$model))

#lets clean our test data accordingly
TitanicTest.clean = subset(TitanicTest, select = c("Pclass","Sex","Age","SibSp"))
TitanicTest.clean$Pclass = as.factor(TitanicTest.clean$Pclass)
TitanicTest.clean$Sex = as.factor(TitanicTest.clean$Sex)
TitanicTest.clean$SibSp = as.factor(TitanicTest.clean$SibSp)

#Since Age has NA values, we need to impute them to get predictions for that PassengerId
library(mice)
TitanicTest.clean.imp = mice(TitanicTest.clean)

#we can check which imputation was the closest to the mean value of original data
mean(complete(TitanicTest.clean.imp,1)$Age) - mean(na.omit(TitanicTest.clean$Age))
mean(complete(TitanicTest.clean.imp,2)$Age) - mean(na.omit(TitanicTest.clean$Age))
mean(complete(TitanicTest.clean.imp,3)$Age) - mean(na.omit(TitanicTest.clean$Age))
mean(complete(TitanicTest.clean.imp,4)$Age) - mean(na.omit(TitanicTest.clean$Age))
mean(complete(TitanicTest.clean.imp,5)$Age) - mean(na.omit(TitanicTest.clean$Age))

#from the above, it seems that the first iteration was our best imputation, therefore we will use it
TitanicTest.clean = complete(TitanicTest.clean.imp,1)

#We then scale the Age column
TitanicTest.clean$Age = scale(TitanicTest.clean$Age)

#creating the dummy vars
library(caret)
dmy = dummyVars(" ~ .", data = TitanicTest.clean, fullRank = T)
TitanicTest.clean = data.frame(predict(dmy, newdata = TitanicTest.clean))
```

```
#keeping only the levels needed
TitanicTest.clean = subset(TitanicTest.clean, select = c("Pclass.2","Pclass.3","Sex.male","Age","SibS

#now we have cleaned data for the model, we can predict Survived for it
TitanicTest.clean.pred = predict(TitanicLinear.k, newdata = TitanicTest.clean)

#adding our predictions to the data-set, we can see how the results look
TitanicTest.clean$PassengerId = TitanicTest$PassengerId
TitanicTest.clean$Survived = round(TitanicTest.clean.pred,0) #we round since model predicts as if ca

#generating our csv for submission
Titanic_submission = subset(TitanicTest.clean, select = c("PassengerId","Survived"))

#writing the submission file
write.csv(Titanic_submission, "rData/TitanicData/titanic_sub.csv", row.names = FALSE)

#our submission scored us 0.75598 which put us at a position of 12288, not that high (14463 is the lo
#This shows that even though we generate models with a fair accuracy, the results should be used as

#we are going to do a separate run for plotting the CART tree
#removing some hard to interpret variables for the plot
titanicData$Name = NULL
titanicData$Ticket = NULL
titanicData$PassengerId = NULL

#run the model
titanicData.r = autoModel("Survived", titanicData, randomSeed = 3)

#plotting the rpart model using rpart.plot
#install.packages("rpart.plot")
library(rpart.plot)

#round the values since Survived in categorical
titanicData.r$cartResults$model$frame$yval = round(titanicData.r$cartResults$model$frame$yval,0)

#plotting
rpart.plot(titanicData.r$cartResults$model,roundint=FALSE, main = "Survival Aboard the Titanic (0 or


########Iris Run########
#Iris is a in-built data-set for R which we can test our function on
summary(iris)

#a nice, simple graphical summary of the data-set
pairs(iris[,1:4], col = iris$Species)

#correlation plot of the data
cor_iris = cor(iris[,1:4])
ggcorrplot(cor_iris, lab = T)

#set console output location
sink("consoleOutput/iris.txt", type=c("output","message"))

#run the model with our context on correlation
iris.r = autoModel("Species", iris, randomSeed = 1, corrConfLevel = 0.99)

#close the connection
closeAllConnections()

########Life_Expectancy_Data1########
#The life expectancy data is a dataset used within the module MA317 at UoE
life_exp = read.csv("rData/MA317_Data/Life_Expectancy_Data1.csv")
summary(life_exp)

#adding a variable which we can color the below plot with
life_exp$split_H2O = ifelse(life_exp$SH.H2O.SMDW.ZS >= 90,2,1)

#a nice linear relationship in the data
```

```
plot(life_exp$SP.DYN.CBRT.IN, life_exp$SP.DYN.LE00.IN,
     xlab = "Birth Rate",
     ylab = "Life Expectancy",
     main = "Life Expectancy vs Birth Rate per Clean Water Access",
     col = life_exp$split_H2O)
abline(lm(SP.DYN.LE00.IN~SP.DYN.CBRT.IN,data = life_exp))
legend("topright",legend = c("Clean Water Access >= 90%", "Clean Water Access < 90%"), fill = c("red

#remove the variable used for plotting
life_exp$split_H2O = NULL

#run the model
life_exp.r = autoModel("SP.DYN.LE00.IN", life_exp)

########House Data########
#The life expectnacy data is a dataset used within the module MA317 at UoE
house_sales = read.csv("rData/MA321_Data/house-data.csv")
summary(house_sales)

#run the model for two different variables
house_sales.r.oc = autoModel("OverallCond", house_sales)
house_sales.r.sp = autoModel("SalePrice", house_sales)
```

10.1.3 ghq_functions

```
#This is all the functions needed to analyze all the GHQ data

#----Libraries needed for the code----
#packages that are needed are install below
#install.packages("ggcorrplot");install.packages("car");install.packages("sqldf")

library(sqldf) #used to execute sql queries
library(ggcorrplot) #used to plot corr matrix in the preliminary analysis
library(car) #need for the qqPlot done in the preliminary analysis
#----

#----Functions needed for initial analysis----

#In our data we have missing values re-coded as various negative values, which are:
#-9, missing
#-8, inapplicable (We can try a run where we recode this to 0, everything else as NA)
#-7, proxy respondent
#-2, refused
#-1, don't know
#We should re-code this all to 'NA' to allow for the model to compile, here a function to achieve th
recodeNA = function(data){
  data[data == -9] = NA
  data[data == -8] = NA
  data[data == -7] = NA
  data[data == -2] = NA
  data[data == -1] = NA
  return(data)
}

#This function below is used so that we can join the income data into our model in a way were it kee
#here what we are doing is that for each seen finance category, we are creating a binary column stat
#pidp is associated with a ficode, (1 means yes, 0 means no)
#here we are also building the select query used for the SQL manipulation later on
sqlTransform = function(data){

  #start of the SQL query
  querySQL = "SELECT pidp,SUM(frmnthimp_dv) as frmnthimp_dv_total"

  #for loop which turns all ficodes into binary columns
  for (i in 1:max(data$ficode)){
    colName = paste("ficode", i,sep="")
```

```
    data[[colName]] = ifelse(data$ficode == i, 1, 0)
    querySQL = paste(querySQL,",SUM(",colName,") as ",colName,sep = "")
  }

  #finish of the SQL query
  querySQL = paste(querySQL,"FROM data GROUP BY pidp")

  #SQL to turn the table into version where pidp is unique per row
  return(sqldf(querySQL))
}

#This function is used to analyse the distribution of GHQ scores and correlations of questions in a d
ghq_analyze = function(data) {
  #We can have a look at the variable scghq1_dv (total GHQ score).
  mainScore = select(data,contains("ghq1"))
  print(summary(mainScore))

  #get number of non-NA Observations
  print(paste(nrow(mainScore) - sum(is.na(mainScore)),"non NA observations"))

  #Get the mode of the total score
  mstComm.an = as.numeric(names(which(table(mainScore) == max(table(mainScore)))))
  print(paste("Mode/most frequent score is:",mstComm.an))

  #getting the total score column and transforming into clean data
  totalScores = as.numeric(unlist(mainScore))

  #plotting histogram
  par(ask = TRUE)
  plot(table(totalScores),
       xlab = "Result of GHQ questionnaire",
       xaxt = 'n',
       ylab = "Frequency",
       main = paste("GHQ Results Histogram (",deparse(substitute(data)),")",sep=""))
  axis(1,at = seq(0,36,6))

  #plotting a Q-Q Plot to investigate normality of distribution
  qqPlot(totalScores,
         ylab = "GHQ Total Score",
         main = paste("GHQ Results Q-Q Plot (",deparse(substitute(data)),")",sep=""))

  #getting the ghq question ONLY (no total scores)
  x_ghq = select(data,contains("ghq"))
  x_ghq = select(x_ghq,-contains("ghq1"))
  x_ghq = select(x_ghq,-contains("ghq2"))

  #summary of all questions
  print(summary(x_ghq))

  #NA removal
  x_ghq = na.omit(x_ghq)

  #creating the correlation matrix and plotting
  corr_ghq = cor(x_ghq)
  ggcorrplot(corr_ghq,
             type = "lower",
             lab = T,
             title = paste("GHQ Questions Corr Plot (",deparse(substitute(data)),")",sep=""))
}

#this function is used to clean out the ghq question from the data so that we are left with just the
#We also then move all common known columns which are totals of other columns to the end of the datas
ghq_clean_move = function(data) {

  #chaging the name of the ghq total score and removing the questions from the dataset
  names(data)[names(data) == names(select(data,contains("ghq1")))] = "total_score"
  data = select(data,-contains("ghq"))
```

```
  #re-ordering the columns for the correlation check so that we automatically keep all 'val' variable
  for (i in 1:length(names(data))) {
    if (grepl("val",colnames(data)[i], fixed = T) | grepl("_dv",colnames(data)[i], fixed = T)) {
      data = data %>% relocate(colnames(data)[i], .after = last_col())
    }
  }

  #return the cleaned dataset
  return(data)
}

#We use this function to add prediction intervals which are unique to GHQ
predGHQadd = function(modelResults,ghq_side){

  #get prediction sheets
  modelKMeans = modelResults$kMeansResults$predictions

  modelkNN = as.data.frame(modelResults$kNNResults$predictions)

  modelCART = modelResults$cartResults$predictions

  modelLinear = modelResults$linearRegression$predictions

  modelElastic = modelResults$elasticNetRegression$predictions

  #creating 3-range score tables for K-Means using function
  addGroupScore = function(modelPred) {
    modelPred$groupScore = 0
    for (i in 1:nrow(modelPred)) {
      if (modelPred[i,2] != 0 & modelPred[i,2] != 36) {
        modelPred[i,4] = modelPred[i-1,3] + modelPred[i,3] + modelPred[i+1,3]
      }
    }

    return(modelPred)
  }

  #adding the new columns
  modelKMeans = addGroupScore(modelKMeans)

  #creating the 1 unit bounds
  modelCART$unitlwr = modelCART$fit - 1
  modelCART$unitupr = modelCART$fit + 1

  modelLinear$unitlwr = modelLinear$fit - 1
  modelLinear$unitupr = modelLinear$fit + 1

  modelElastic$unitlwr = modelElastic$fit - 1
  modelElastic$unitupr = modelElastic$fit + 1

  #creating the GHQ bounds
  modelCART$GHQlwr = modelCART$fit - ghq_side
  modelCART$GHQupr = modelCART$fit + ghq_side

  modelLinear$GHQlwr = modelLinear$fit - ghq_side
  modelLinear$GHQupr = modelLinear$fit + ghq_side

  modelElastic$GHQlwr = modelElastic$fit - ghq_side
  modelElastic$GHQupr = modelElastic$fit + ghq_side

  #getting prediction accuracy in group for K-Means
  clusterKMeans = 0
  for (i in unique(modelKMeans$cluster)){
    temp = subset(modelKMeans, cluster == i)
    clusterKMeans = clusterKMeans + max(temp$groupScore)
  }

  predkNN = 0
```

```
  for (i in unique(modelkNN$predicted)){
    temp = subset(modelkNN, predicted == i)

    for (j in 1:nrow(temp)) {
      if (temp$predicted[j] == temp$real[j]) {
        predkNN = predkNN + temp$groupScore[j]
      }
    }

  }

  #calculating the +-1 range accuracy for kMeans
  print(paste("The K-Means Model predicts within the 1 unit interval range with accuracy of",
              round(clusterKMeans/sum(modelKMeans$freq),4)))

  #calculating the +-1 range accuracy for kNN
  correct = sum(ifelse(modelkNN$predicted - modelkNN$real == 0 |
                            abs(modelkNN$predicted - modelkNN$real) == 1,1,0)*modelkNN$freq)
  print(paste("The kNN Model predicts within the 1 unit interval range with accuracy of",
              round(correct/sum(modelkNN$freq),4)))

  #creating the prediction for the (CART): 1 unit
  print(paste("The CART Model predicts within the 1 unit interval with accuracy of",
              round(sum(modelCART$real >= round(modelCART$unitlwr,0) & modelCART$real <= round(modelC

  #creating the prediction for the (CART): GHQ limits
  print(paste("The CART Model predicts within the GHQ calculated interval with accuracy of",
              round(sum(modelCART$real >= round(modelCART$GHQlwr,0) & modelCART$real <= round(modelCA

  #creating the prediction for the (Linear): 1 unit
  print(paste("The OLR Model predicts within the 1 unit interval with accuracy of",
              round(sum(modelLinear$real >= round(modelLinear$unitlwr,0) & modelLinear$real <= round(

  #creating the prediction for the (Linear): GHQ limits
  print(paste("The OLR Model predicts within the GHQ calculated interval with accuracy of",
              round(sum(modelLinear$real >= round(modelLinear$GHQlwr,0) & modelLinear$real <= round(

  #creating the prediction for the (Elastic): 1 unit
  print(paste("The ENR Model predicts within the 1 unit interval with accuracy of",
              round(sum(modelElastic$real >= round(modelElastic$unitlwr,0) & modelElastic$real <= rou

  #creating the prediction for the (Elastic): GHQ limits
  print(paste("The ENR Model predicts within the GHQ calculated interval with accuracy of",
              round(sum(modelElastic$real >= round(modelElastic$GHQlwr,0) & modelElastic$real <= rou

  #create a list which holds both the sheets
  predictionSheets = list(modelKMeans,modelkNN,modelLinear,modelElastic)
  names(predictionSheets) = c("K-Means","kNN","Linear","Elastic")

  #returning the sheets separately
  return(predictionSheets)
}
```

10.2 Run Results

10.2.1 shared results across all data-sets for summaries in GHQ

Names	scghq1_dv	scghqa - scghql (excl. scghqk)	scghqk
Min.	0	1	1
1st Q	7 - 8	**varied**	1
Median	10	2	1
Mode	6	**varied**	**varied**
3rd Qu	13 - 14	2	2
Max.	36	4	4

10.2.2 w2indresp stats (graphs)

scghq1_dv	
Mean	11.2
NA's	11146
Non-NA's	43423

scghqa	
1st Q	2
Mean	2.147
NA's	10685

scghqb	
1st Q	1
Mean	1.876
NA's	10646

scghqc	
1st Q	2
Mean	2.082
NA's	10729

scghqd	
1st Q	2
Mean	2.03
NA's	10651

scghqe	
1st Q	1
Mean	1.997
NA's	10667

scghqf	
1st Q	1
Mean	1.789
NA's	10679

scghqg	
1st Q	2
Mean	2.159
NA's	10645

scghqh	
1st Q	2
Mean	2.063
NA's	10669

scghqi	
1st Q	1
Mean	1.844
NA's	10649

scghqj	
1st Q	1
Mean	1.728
NA's	10654

scghqk	
1st Q	1
Mean	1.445
NA's	10672

scghql	
1st Q	2
Mean	2.056
NA's	10661

10.2.3 w2Merge stats (graphs)

scghq1_dv	
Mean	11.56
NA's	4824
Non-NA's	28540

scghqa	
1st Q	2
Mean	2.175
NA's	4477

scghqb	
1st Q	1
Mean	1.908
NA's	4442

scghqc	
1st Q	2
Mean	2.125
NA's	4509

scghqd	
1st Q	2
Mean	2.064
NA's	4445

scghqe	
1st Q	1
Mean	2
NA's	4464

scghqf	
1st Q	1
Mean	1.815
NA's	4474

scghqg	
1st Q	2
Mean	2.199
NA's	4446

scghqh	
1st Q	2
Mean	2.092
NA's	4467

scghqi	
1st Q	1
Mean	1.866
NA's	4449

scghqj	
1st Q	1
Mean	1.765
NA's	4458

scghqk	
1st Q	1
Mean	1.483
NA's	4461

scghql	
1st Q	2
Mean	2.081
NA's	4458

10.2.4 w2MergeNurse stats (graphs)

scghq1_dv	
Mean	11.42
NA's	1110
Non-NA's	9811

scghqa	
1st Q	2
Mean	2.17
NA's	1020

scghqb	
1st Q	1
Mean	1.885
NA's	1008

scghqc	
1st Q	2
Mean	2.116
NA's	1031

scghqd	
1st Q	2
Mean	2.062
NA's	1010

scghqe	
1st Q	1
Mean	1.976
NA's	1013

scghqf	
1st Q	1
Mean	1.793
NA's	1014

scghqg	
1st Q	2
Mean	2.192
NA's	1007

scghqh	
1st Q	2
Mean	2.093
NA's	1016

scghqi	
1st Q	1
Mean	1.851
NA's	999

scghqj	
1st Q	1
Mean	1.751
NA's	1003

scghqk	
1st Q	1
Mean	1.463
NA's	1005

scghql	
1st Q	2
Mean	2.08
NA's	999

10.2.5 w2MergeNurseBlood stats (graphs)

scghq1_dv	
Mean	11.21
Mode	6.00
NA's	632
Non-NA's	6294

scghqa	
1st Q	2
Mean	2.157
NA's	577

scghqb	
1st Q	1
Mean	1.865
NA's	573

scghqc	
1st Q	2
Mean	2.092
NA's	585

scghqd	
1st Q	2
Mean	2.056
NA's	573

scghqe	
1st Q	1
Mean	1.951
NA's	572

scghqf	
1st Q	1
Mean	1.77
NA's	577

scghqg	
1st Q	2
Mean	2.18
NA's	571

scghqh	
1st Q	2
Mean	2.09
NA's	576

scghqi	
1st Q	1
Mean	1.817
NA's	568

scghqj	
1st Q	1
Mean	1.73
NA's	571

scghqk	
1st Q	1
Mean	1.446
NA's	567

scghql	
1st Q	2
Mean	2.075
NA's	568

10.2.6 w3indresp stats (graphs)

scghq1_dv	
Mean	11.07
NA's	9116
Non-NA's	40576

scghqa	
1st Q	2
Mean	2.169
NA's	9054

scghqb	
1st Q	1
Mean	1.843
NA's	9049

scghqc	
1st Q	2
Mean	2.092
NA's	9073

scghqd	
1st Q	2
Mean	2.027
NA's	9057

scghqe	
1st Q	1
Mean	2.006
NA's	9054

scghqf	
1st Q	1
Mean	1.763
NA's	9067

scghqg	
1st Q	2
Mean	2.134
NA's	9057

scghqh	
1st Q	2
Mean	2.053
NA's	9060

scghqi	
1st Q	1
Mean	1.816
NA's	9053

scghqj	
1st Q	1
Mean	1.713
NA's	9058

scghqk	
1st Q	1
Mean	1.415
NA's	9070

scghql	
1st Q	2
Mean	2.049
NA's	9062

10.2.7 w3Merge stats (graphs)

scghq1_dv	
Mean	11.34
NA's	3939
Non-NA's	26548

scghqa	
1st Q	2
Mean	2.193
NA's	3884

scghqb	
1st Q	1
Mean	1.859
NA's	3879

scghqc	
1st Q	2
Mean	2.123
NA's	3903

scghqd	
1st Q	2
Mean	2.053
NA's	3887

scghqe	
1st Q	1
Mean	1.998
NA's	3884

scghqf	
1st Q	1
Mean	1.779
NA's	3896

scghqg	
1st Q	2
Mean	2.166
NA's	3886

scghqh	
1st Q	2
Mean	2.08
NA's	3888

scghqi	
1st Q	1
Mean	1.833
NA's	3881

scghqj	
1st Q	1
Mean	1.742
NA's	3885

scghqk	
1st Q	1
Mean	1.441
NA's	3895

scghql	
1st Q	2
Mean	2.075
NA's	3887

10.2.8 w3MergeNurse stats (graphs)

scghq1_dv	
Mean	11.36
NA's	255
Non-NA's	3217

scghqa	
1st Q	2
Mean	2.208
NA's	247

scghqb	
1st Q	1
Mean	1.86
NA's	249

scghqc	
1st Q	2
Mean	2.126
NA's	249

scghqd	
1st Q	2
Mean	2.06
NA's	247

scghqe	
1st Q	1
Mean	2.019
NA's	247

scghqf	
1st Q	1
Mean	1.779
NA's	248

scghqg	
1st Q	2
Mean	2.172
NA's	248

scghqh	
1st Q	2
Mean	2.083
NA's	249

scghqi	
1st Q	1
Mean	1.837
NA's	247

scghqj	
1st Q	1
Mean	1.723
NA's	247

scghqk	
1st Q	1
Mean	1.44
NA's	248

scghql	
1st Q	2
Mean	2.065
NA's	247

10.2.9 w3MergeNurseBlood stats (graphs)

scghq1_dv	
Mean	11.21
NA's	154
Non-NA's	2182

scghqa	
1st Q	2
Mean	2.198
NA's	149

scghqb	
1st Q	1
Mean	1.844
NA's	150

scghqc	
1st Q	2
Mean	2.11
NA's	151

scghqd	
1st Q	2
Mean	2.054
NA's	149

scghqe	
1st Q	1
Mean	2.009
NA's	149

scghqf	
1st Q	1
Mean	1.769
NA's	150

scghqg	
1st Q	2
Mean	2.151
NA's	150

scghqh	
1st Q	2
Mean	2.08
NA's	150

scghqi	
1st Q	1
Mean	1.816
NA's	149

scghqj	
1st Q	1
Mean	1.7
NA's	149

scghqk	
1st Q	1
Mean	1.423
NA's	149

scghql	
1st Q	2
Mean	2.052
NA's	149

10.2.10 wShared stats (graphs)

scghq1_dv	
Mean	11.14
NA's	20262
Non-NA's	83999

scghqa	
1st Q	2
Mean	2.158
NA's	19739

scghqb	
1st Q	1
Mean	1.86
NA's	19695

scghqc	
1st Q	2
Mean	2.087
NA's	19802

scghqd	
1st Q	2
Mean	2.029
NA's	19708

scghqe	
1st Q	1
Mean	2.001
NA's	19721

scghqf	
1st Q	1
Mean	1.777
NA's	19746

scghqg	
1st Q	2
Mean	2.147
NA's	19702

scghqh	
1st Q	2
Mean	2.058
NA's	19729

scghqi	
1st Q	1
Mean	1.83
NA's	19702

scghqj	
1st Q	1
Mean	1.721
NA's	19712

scghqk	
1st Q	1
Mean	1.43
NA's	19742

scghql	
1st Q	2
Mean	2.052
NA's	19723

10.2.11 wSMerge stats (graphs)

scghq1_dv	
Mean	11.45
NA's	8763
Non-NA's	55088

scghqa	
1st Q	2
Mean	2.184
NA's	8361

scghqb	
1st Q	1
Mean	1.885
NA's	8321

scghqc	
1st Q	2
Mean	2.124
NA's	8412

scghqd	
1st Q	2
Mean	2.059
NA's	8332

scghqe	
1st Q	1
Mean	1.999
NA's	8348

scghqf	
1st Q	1
Mean	1.798
NA's	8370

scghqg	
1st Q	2
Mean	2.183
NA's	8332

scghqh	
1st Q	2
Mean	2.086
NA's	8355

scghqi	
1st Q	1
Mean	1.850
NA's	8330

scghqj	
1st Q	1
Mean	1.754
NA's	8343

scghqk	
1st Q	1
Mean	1.463
NA's	8356

scghql	
1st Q	2
Mean	2.078
NA's	8345

10.2.12 wSMergeNurse stats (graphs)

scghq1_dv	
Mean	11.4
NA's	1365
Non-NA's	13028

scghqa	
1st Q	2
Mean	2.179
NA's	1267

scghqb	
1st Q	1
Mean	1.879
NA's	1257

scghqc	
1st Q	2
Mean	2.118
NA's	1280

scghqd	
1st Q	2
Mean	2.062
NA's	1257

scghqe	
1st Q	1
Mean	1.987
NA's	1260

scghqf	
1st Q	1
Mean	1.789
NA's	1262

scghqg	
1st Q	2
Mean	2.187
NA's	1255

scghqh	
1st Q	2
Mean	2.090
NA's	1265

scghqi	
1st Q	1
Mean	1.848
NA's	1246

scghqj	
1st Q	1
Mean	1.744
NA's	1250

scghqk	
1st Q	1
Mean	1.457
NA's	1253

scghql	
1st Q	2
Mean	2.077
NA's	1246

10.2.13 wSMergeNurseBlood stats (graphs)

scghq1_dv	
Mean	11.21
NA's	786
Non-NA's	8476

scghqa	
1st Q	2
Mean	2.168
NA's	726

scghqb	
1st Q	1
Mean	1.859
NA's	723

scghqc	
1st Q	2
Mean	2.097
NA's	736

scghqd	
1st Q	2
Mean	2.055
NA's	722

scghqe	
1st Q	1
Mean	1.966
NA's	721

scghqf	
1st Q	1
Mean	1.77
NA's	727

scghqg	
1st Q	2
Mean	2.173
NA's	721

scghqh	
1st Q	2
Mean	2.087
NA's	726

scghqi	
1st Q	1
Mean	1.817
NA's	717

scghqj	
1st Q	1
Mean	1.722
NA's	720

scghqk	
1st Q	1
Mean	1.44
NA's	716

scghql	
1st Q	2
Mean	2.07
NA's	717

10.2.14 mixNurse stats (graphs)

scghq1_dv	
Mean	11.19
NA's	1856
Non-NA's	18843

scghqa	
1st Q	2
Mean	2.161
NA's	1734

scghqb	
1st Q	1
Mean	1.863
NA's	1725

scghqc	
1st Q	2
Mean	2.085
NA's	1756

scghqd	
1st Q	2
Mean	2.039
NA's	1724

scghqe	
1st Q	1
Mean	1.998
NA's	1727

scghqf	
1st Q	1
Mean	1.776
NA's	1732

scghqg	
1st Q	2
Mean	2.156
NA's	1722

scghqh	
1st Q	2
Mean	2.067
NA's	1735

scghqi	
1st Q	1
Mean	1.841
NA's	1706

scghqj	
1st Q	1
Mean	1.722
NA's	1710

scghqk	
1st Q	1
Mean	1.431
NA's	1718

scghql	
1st Q	2
Mean	2.058
NA's	1710

10.2.15 mixNurseBlood stats (graphs)

scghq1_dv	
Mean	11.06
NA's	1091
Non-NA's	12156

scghqa	
1st Q	2
Mean	2.153
NA's	1016

scghqb	
1st Q	1
Mean	1.849
NA's	1014

scghqc	
1st Q	2
Mean	2.07
NA's	1029

scghqd	
1st Q	2
Mean	2.037
NA's	1012

scghqe	
1st Q	1
Mean	1.982
NA's	1011

scghqf	
1st Q	1
Mean	1.758
NA's	1019

scghqg	
1st Q	2
Mean	2.148
NA's	1011

scghqh	
1st Q	2
Mean	2.067
NA's	1018

scghqi	
1st Q	1
Mean	1.819
NA's	1004

scghqj	
1st Q	1
Mean	1.707
NA's	1007

scghqk	
1st Q	1
Mean	1.422
NA's	1007

scghql	
1st Q	2
Mean	2.056
NA's	1006

10.2.16 w2indresp console

```
[1] "————————Initial Checks————————"
[1] "57987325 NA cells were found across the entire dataset (64.8% of data as NA)"
[1] "————————Data Type Checks————————"
[1] "0 variables recoded since all their entries aren't numeric or NA"
[1] "NOTE: algorithim recodes categorical data alphabetically e.g (female = 1, male = 2)"
character(0)
[1] "————————Low Data Removal————————"
[1] "1163 variables removed since they had >= 'naPercent' (default 20%) NA values"
   [1] "pid"           "lvwhy"         "lvmthp"        "lvyrp"         "mvever"
"mvmnth"        "mvyr"          "mlstatchk"
   [9] "mlstat"        "drive"         "caruse"        "ukborn"        "plbornc"
"yr2uk4"        "citzn1"        "citzn2"
  [17] "citzn3"        "qfhigh"        "qualoc"        "qfvoc1"        "qfvoc2"
"qfvoc3"        "qfvoc4"        "qfvoc5"
  [25] "qfvoc6"        "qfvoc7"        "qfvoc8"        "qfvoc9"        "qfvoc10"
"qfvoc11"       "qfvoc12"       "qfvoc13"
  [33] "qfvoc14"       "qfvoc15"       "qfvoc96"       "school"        "scend"
"schlloc"       "schok"         "fenow"
  [41] "feend"         "j1none"        "j1semp"        "j1boss"        "j1mngr"
"paedqf"        "maedqf"        "edtype"
  [49] "edasp"         "fedlik"        "fednt"         "ocimpa"        "ocimpb"
"ocimpe"        "ocimpf"        "ocimpi"
  [57] "ocimpk"        "ocimpl"        "futra"         "futrb"         "futrc"
"futrd"         "futre"         "futrf"
  [65] "futrg"         "futrh"         "futri"         "futrj"         "futrk"
"futrl"         "paju"          "maju"
  [73] "pacob"         "payruk"        "macob"         "mayruk"        "natid1"
"natid2"        "natid3"        "natid4"
  [81] "natid5"        "natid6"        "natid97"       "racel"         "racelo_code"
"ethid2"        "ethid3"        "ethid4a"
  [89] "ethid4b"       "ethid5"        "ethid6"        "ethid7"        "ethid8"
"ethid9"        "ethid10"       "ethid11"
  [97] "ethid12"       "ethid13"       "ethid14"       "ethclose1"     "ethclose2a"
"ethclose2b"    "ethclose3"     "ethclose4a"
 [105] "ethclose4b"    "ethclose5"     "ethclose6"     "ethclose7"     "ethclose8"
"ethclose9"     "ethclose10"    "ethclose11"
 [113] "pride2"        "pride4a"       "pride4b"       "pride6"        "pride7"
"pride5"        "pride8"        "pride9"
 [121] "pride10"       "pride11"       "pride12"       "pride13"       "pride14"
"food1"         "food2"         "food3"
 [129] "food4"         "food5"         "food6"         "food7"         "oprlg"
"oprlg0ni"      "nirel"         "niact"
 [137] "oprlg0"        "oprlg1"        "smnow"         "ncigs"         "smcigs"
"smncigs"       "aglquit"       "smagbg"
 [145] "hospc1"        "hospdc1"       "hospc2"        "hospdc2"       "hospc3"
"hospdc3"       "hospc4"        "hospdc4"
 [153] "hospc5"        "hospdc5"       "hospc6"        "hospdc6"       "hospc7"
"hospdc7"       "hospc8"        "hospdc8"
 [161] "disdif1"       "disdif2"       "disdif3"       "disdif4"       "disdif5"
"disdif6"       "disdif7"       "disdif8"
 [169] "disdif9"       "disdif10"      "disdif11"      "disdif12"      "disdif96"
"aidhua1"       "aidhua2"       "aidhua3"
 [177] "aidhua4"       "aidhua5"       "aidhua6"       "aidhua7"       "aidhua8"
"aidhua9"       "aidhua10"      "aidhua11"
 [185] "aidhua12"      "aidhua13"      "aidhua14"      "aidhua15"      "aidhua16"
"naidxhh"       "aidhu1"        "aidhu2"
 [193] "aidhrs"        "aideft"        "lcohnpi"       "coh1bm"        "coh1by"
"coh1mr"        "coh1em"        "coh1ey"
 [201] "nmar"          "lmar1m"        "lmar1y"        "ladopt"        "lnadopt"
"lprnt"         "lnprnt"        "ch1by4"
 [209] "movy11"        "movy12"        "movy13"        "movy14"        "movy15"
"movy16"        "family"        "education"
 [217] "memploy"       "housing"       "area"          "moveoth_code"  "movdir"
"plnowm"        "plnowy4"       "mstatsam"
 [225] "lwwrong"       "mstatsamn"     "lastmstatch1"  "mstatch1"      "statcm1"
"statcy41"      "divchk1"       "divfin1"
```

```
 [233] "dvm1"          "dvy41"         "cmlstat1"      "lastmstatch2" "mstatch2"
"statcm2"       "statcy42"      "divchk2"
 [241] "divfin2"       "dvm2"          "dvy42"         "cmlstat2"      "lastmstatch3" "mstatch3"
"statcm3"       "statcy43"
 [249] "divchk3"       "divfin3"       "dvm3"          "dvy43"         "cmlstat3"
"cohab"         "cohabn"        "lmcbm1"
 [257] "lmcby41"       "currpart1"     "lmspm1"        "lmspy41"       "lmcbm2"
"lmcby42"       "currpart2"     "lmspm2"
 [265] "lmspy42"       "lmcbm3"        "lmcby43"       "currpart3"     "lmspm3"
"lmspy43"       "lmcbm4"        "lmcby44"
 [273] "currpart4"     "lmspm4"        "lmspy44"       "father"        "nchild"
"preg"          "pregm1"        "pregy41"
 [281] "pregfert1"     "invitro1"      "pregout1"      "pregend1"      "endmnth1"
"pregsmoke1"    "smkmnth11"     "smkmnth21"
 [289] "smkmnth31"     "pregsmk11"     "pregsmk21"     "pregsmk31"     "pregdrink1"
"lchmulti1"     "pregm2"        "pregy42"
 [297] "pregfert2"     "invitro2"      "pregout2"      "pregend2"      "endmnth2"
"pregsmoke2"    "smkmnth12"     "smkmnth22"
 [305] "smkmnth32"     "pregsmk12"     "pregsmk22"     "pregsmk32"     "pregdrink2"
"lchmulti2"     "pregm3"        "pregy43"
 [313] "pregfert3"     "invitro3"      "pregout3"      "pregend3"      "endmnth3"
"pregsmoke3"    "smkmnth13"     "smkmnth23"
 [321] "smkmnth33"     "pregsmk13"     "pregsmk23"     "pregsmk33"     "pregdrink3"
"lchmulti3"     "pregm4"        "pregy44"
 [329] "pregfert4"     "invitro4"      "pregout4"      "pregend4"      "endmnth4"
"pregsmoke4"    "smkmnth14"     "smkmnth24"
 [337] "smkmnth34"     "pregsmk14"     "pregsmk24"     "pregsmk34"     "pregdrink4"
"lchmulti4"     "pregm5"        "pregy45"
 [345] "pregfert5"     "invitro5"      "pregout5"      "pregend5"      "endmnth5"
"pregsmoke5"    "smkmnth15"     "smkmnth25"
 [353] "smkmnth35"     "pregsmk15"     "pregsmk25"     "pregsmk35"     "pregdrink5"
"lchmulti5"     "hcondno1"      "hcondns1"
 [361] "hcondno2"      "hcondns2"      "hcondno3"      "hcondns3"      "hcondno4"
"hcondns4"      "hcondno5"      "hcondns5"
 [369] "hcondno6"      "hcondns6"      "hcondno7"      "hcondns7"      "hcondno8"
"hcondns8"      "contft"        "ftendm"
 [377] "ftendy4"       "ftquals"       "ftedstartm1"   "ftedstarty41" "ftedend1"
"ft2endm1"      "ft2endy41"     "ftedmor1"
 [385] "ftedstartm2"   "ftedstarty42" "ftedend2"      "ft2endm2"      "ft2endy42"
"ftedmor2"      "qualnew1"      "qualnew2"
 [393] "qualnew3"      "qualnew4"      "qualnew5"      "qualnew6"      "qualnew7"
"qualnew8"      "qualnew9"      "qualnew10"
 [401] "qualnew11"     "qualnew13"     "qualnew15"     "qualnew16"     "qualnew17"
"qualnew18"     "qualnew19"     "qualnew20"
 [409] "qualnew21"     "qualnew22"     "qualnew23"     "qualnew24"     "qualnew25"
"qualnew26"     "qualnew27"     "qualnew28"
 [417] "qualnew29"     "qualnew30"     "qualnew31"     "trwho1"        "traindays1"
"trainhrs1"     "trainend1"     "trainpurp11"
 [425] "trainpurp21"   "trainpurp31"   "trainpurp41"   "trainpurp51"   "trainpurp61"
"trainpurp71"   "trainqual1"    "trwho2"
 [433] "traindays2"    "trainhrs2"     "trainend2"     "trainpurp12"   "trainpurp22"
"trainpurp32"   "trainpurp42"   "trainpurp52"
 [441] "trainpurp62"   "trainpurp72"   "trainqual2"    "trwho3"        "traindays3"
"trainhrs3"     "trainend3"     "trainpurp13"
 [449] "trainpurp23"   "trainpurp33"   "trainpurp43"   "trainpurp53"   "trainpurp63"
"trainpurp73"   "trainqual3"    "trainn"
 [457] "trqual1"       "trqual2"       "trqual3"       "trqual4"       "trqual5"
"trqual6"       "trqual7"       "trqual8"
 [465] "trqual9"       "trqual10"      "trqual11"      "trqual13"      "trqual15"
"trqual16"      "trqual17"      "trqual18"
 [473] "trqual19"      "trqual20"      "trqual21"      "trqual22"      "trqual23"
"trqual24"      "trqual25"      "trqual26"
 [481] "trqual27"      "trqual28"      "trqual29"      "trqual30"      "trqual31"
"notempchk"     "empchk"        "empstendd"
 [489] "empstendm"     "empstendy4"    "stendreas"     "nxtst"         "nxtstelse"
"cstat"         "nxtstendd"     "nxtstendm"
 [497] "nxtstendy4"    "jbsamr"        "wkplsam"       "samejob"       "matlv"
"matlvstd"      "matlvstm"      "matlvsty4"
```

```
[505] "matlvendd"    "matlvendm"    "matlvendy4"   "jbendd"       "jbendm"
"jbendy4"      "jbendreas"    "cjob"
[513] "nxtjbhrs"     "nxtjbes"      "nxtjbendd"    "nxtjbendm"    "nxtjbendy4"
"cjbatt"       "nextstat1"    "nextelse1"
[521] "currstat1"    "nextjob1"     "currjob1"     "jobhours1"    "reasend1"
"jbatt1"       "statendd1"    "statendm1"
[529] "statendy41"   "nextstat2"    "nextelse2"    "currstat2"    "nextjob2"
"currjob2"     "jobhours2"    "reasend2"
[537] "jbatt2"       "statendd2"    "statendm2"    "statendy42"   "nextstat3"
"nextelse3"    "currstat3"    "nextjob3"
[545] "currjob3"     "jobhours3"    "reasend3"     "jbatt3"       "statendd3"
"statendm3"    "statendy43"   "nextstat4"
[553] "nextelse4"    "currstat4"    "nextjob4"     "currjob4"     "jobhours4"
"reasend4"     "jbatt4"       "statendd4"
[561] "statendm4"    "statendy44"   "nextstat5"    "nextelse5"    "currstat5"
"nextjob5"     "currjob5"     "jobhours5"
[569] "reasend5"     "jbatt5"       "statendd5"    "statendm5"    "statendy45"
"nextstat6"    "nextelse6"    "currstat6"
[577] "nextjob6"     "currjob6"     "jobhours6"    "reasend6"     "jbatt6"
"statendd6"    "statendm6"    "statendy46"
[585] "nextstat7"    "nextelse7"    "currstat7"    "nextjob7"     "currjob7"
"jobhours7"    "reasend7"     "jbatt7"
[593] "statendd7"    "statendm7"    "statendy47"   "nextstat8"    "nextelse8"
"currstat8"    "nextjob8"     "currjob8"
[601] "jobhours8"    "reasend8"     "jbatt8"       "statendd8"    "statendm8"
"statendy48"   "nextstat9"    "nextelse9"
[609] "currstat9"    "nextjob9"     "currjob9"     "jobhours9"    "reasend9"
"jbatt9"       "statendd9"    "statendm9"
[617] "statendy49"   "nextstat10"   "nextelse10"   "currstat10"   "nextjob10"
"currjob10"    "jobhours10"   "reasend10"
[625] "jbatt10"      "statendd10"   "statendm10"   "statendy410"  "jboff"
"jboffy"       "jbterm1"      "jbterm2"
[633] "jbsic07chk"   "jbsoc00chk"   "jbsempchk"    "jbsemp"       "jbbgd"
"jbbgm"        "jbbgy"        "jbmngrchk"
[641] "jbmngr"       "jbsizechk"    "jbsize"       "jbsect"       "jbsectpub"
"jbhrs"        "jbot"         "jbotpd"
[649] "paygwc"       "paynwc"       "payusl"       "payu"         "payuwc"
"payug"        "paytyp"       "ovtpay"
[657] "extnsa"       "extrate"      "extrest"      "basnsa"       "basrate"
"basrest"      "ovtnsa"       "ovtrate"
[665] "ovtrest"      "jbpl"         "jbttwt"       "worktrav"     "jsboss"
"jssize"       "jshrs"        "jstypeb"
[673] "jsaccs"       "jspart"       "jsprbm"       "jsprby4"      "jsprem"
"jsprey4"      "jsprls"       "jsprtx"
[681] "jsprni"       "jspayu"       "jspayw"       "jspytx"       "jspyni"
"jspl"         "jsttwt"       "jsttwtb"
[689] "jsworktrav"   "workdis"      "twkdiff1"     "twkdiff2"     "twkdiff3"
"twkdiff4"     "twkdiff5"     "twkdiff6"
[697] "twkdiff7"     "twkdiff8"     "twkdiff97"    "twkdiffm"     "twkcar"
"twkcary1"     "twkcary2"     "twkcary3"
[705] "twkcary4"     "twkcary5"     "twkcary6"     "twkcary7"     "twkcary8"
"twkcary9"     "twkcary10"    "twkcary11"
[713] "twkcary12"    "twkcary13"    "twkcary14"    "twkcary97"    "twkcarym"
"altcar1"      "altcar2"      "altcar3"
[721] "altcar4"      "altcar5"      "altcar6"      "altcar7"      "altcar8"
"altcar9"      "altcar10"     "altcar11"
[729] "altcar12"     "altcar96"     "altcar97"     "carclub"      "carshare"
"wkhome"       "lifthh"       "liftxhh"
[737] "motcyc"       "comtaxi"      "combus"       "comtrain"     "commetro"
"combike"      "comwalk"      "comother"
[745] "jbsat"        "wkphys"       "jbperfp"      "jbonus"       "jbrise"
"tujbpl"       "tuin1"        "jbpen"
[753] "jbpenm"       "jbpeny4"      "penmcn"       "penmpy"       "penmtp"
"penspb"       "wktime"       "wkends"
[761] "jbflex1"      "jbflex2"      "jbflex3"      "jbflex4"      "jbflex5"
"jbflex6"      "jbflex7"      "jbflex8"
[769] "jbflex96"     "jbfxuse1"     "jbfxuse2"     "jbfxuse3"     "jbfxuse4"
"jbfxuse5"     "jbfxuse6"     "jbfxuse7"
```

```
 [777] "jbfxuse8"      "jbfxuse96"     "jbfxinf"       "wkaut1"        "wkaut2"
"wkaut3"        "wkaut4"        "wkaut5"
 [785] "depenth1"      "depenth2"      "depenth3"      "depenth4"      "depenth5"
"depenth6"      "jblkcha"       "jbxpcha"
 [793] "jblkchb"       "jbxpchb"       "jblkchc"       "jbxpchc"       "jblkchd"
"jbxpchd"       "jblkche"       "jbxpche"
 [801] "jbsec"         "julk4wk"       "julkjb"        "jubgn"         "julk4x1"
"julk4x2"       "julk4x3"       "julk4x4"
 [809] "julk4x5"       "julk4x6"       "julk4x96"      "jbhad"         "jlendm"
"jlendy"        "jlsemp"        "jlboss"
 [817] "jlmngr"        "jlsize"        "eprosh"        "j2semp"        "j2hrs"
"j2pay"         "retchk"        "ageret"
 [825] "rtpro1"        "rtpro2"        "rtpro3"        "rtpro4"        "rtpro5"
"rtpro6"        "rtcon1"        "rtcon2"
 [833] "rtcon3"        "rtcon4"        "penmex"        "pppex"         "pppexm"
"sppen"         "rtexpjb"       "rtfnd1"
 [841] "rtfnd2"        "rtfnd3"        "rtfnd4"        "rtfnd5"        "rtfnd6"
"rtfnd7"        "rtfnd8"        "rtfnd9"
 [849] "rtfnd10"       "rtfnd96"       "retamt"        "retsuf"        "volfreq"
"volhrs"        "charfreq"      "charam"
 [857] "ccare"         "ccwork"        "benunemp1"     "benunemp2"     "benunemp96"
"bendis1"       "bendis2"       "bendis3"
 [865] "bendis4"       "bendis5"       "bendis6"       "bendis7"       "bendis8"
"bendis9"       "bendis10"      "bendis11"
 [873] "bendis96"      "benpen1"       "benpen2"       "benpen3"       "benpen4"
"benpen5"       "benpen6"       "benpen7"
 [881] "benpen8"       "benpen96"      "niserps"       "bencb"         "benctc"
"benfam1"       "benfam2"       "benfam3"
 [889] "benfam4"       "benfam5"       "benfam96"      "bentax1"       "bentax2"
"bentax3"       "bentax4"       "bentax5"
 [897] "bentax96"      "benhou1"       "benhou2"       "benhou3"       "benhou4"
"benhou96"      "nfh01"         "nfh02"
 [905] "nfh03"         "nfh04"         "nfh05"         "nfh06"         "nfh07"
"nfh08"         "nfh09"         "nfh10"
 [913] "nfh11"         "nfh12"         "nfh13"         "nfh14"         "nfh15"
"nfh16"         "nfh17"         "nfh18"
 [921] "nfh19"         "nfh20"         "nfh21"         "nfh22"         "nfh23"
"nfh24"         "nfh25"         "nfh26"
 [929] "nfh27"         "nfh28"         "nfh29"         "nfh30"         "nfh31"
"nfh32"         "nfh33"         "nfh34"
 [937] "nfh35"         "nfh36"         "nfh37"         "nfh38"         "fiyrdb1"
"fiyrdb2"       "fiyrdb3"       "fiyrdb4"
 [945] "fiyrdb5"       "fiyrdb6"       "ppent"         "ppyrs"         "ppreg"
"ppram"         "pprampc"       "saved"
 [953] "savreg"        "savlt"         "hubuys"        "hufrys"        "humops"
"huiron"        "hupots"        "hudiy"
 [961] "husits"        "huboss"        "vote2"         "vote3"         "vote4"
"vote5"         "perpolinf"     "colbens1"
 [969] "colbens2"      "colbens3"      "civicduty"     "polcost"       "votenorm"
"perbfts"       "grpbfts"       "voteintent"
 [977] "demorient"     "vote7"         "vote8"         "arts1freq"     "arts2freq"
"libfreq"       "arcfreq"       "musfreq"
 [985] "herfreq"       "sportsfreq"    "sports3freq"   "club"          "acc1to161"
"acc1to162"     "acc1to163"     "acc1to164"
 [993] "acc1to165"     "acc1to166"     "acc1to167"     "acc1to168"     "acc1to169"
"acc1to1610"    "acc1to1611"    "acc1to1612"
 [ reached getOption("max.print") -- omitted 163 entries ]
[1] "------------Low Level Removal------------"
[1] "If a level is removed from a variable you wish to keep, reccomended to manually merge levels to
[1] "level 6 in pno removed, 4 observations found"
[1] "level 7 in pno removed, 1 observations found"
[1] "level 5 in nch415resp removed, 1 observations found"
[1] "level 6 in nchund18resp removed, 0 observations found"
[1] "level 7 in natch01 removed, 3 observations found"
[1] "level 8 in natch01 removed, 0 observations found"
[1] "level 2 in natch02 removed, 3 observations found"
[1] "level 7 in natch02 removed, 3 observations found"
[1] "level 8 in natch02 removed, 0 observations found"
```

```
[1] "level 9 in natch02 removed, 0 observations found"
[1] "level 3 in natch03 removed, 0 observations found"
[1] "level 7 in natch03 removed, 2 observations found"
[1] "level 9 in natch03 removed, 0 observations found"
[1] "level 10 in natch03 removed, 0 observations found"
[1] "level 7 in natch06 removed, 1 observations found"
[1] "level 8 in natch06 removed, 1 observations found"
[1] "level 9 in natch06 removed, 1 observations found"
[1] "level 9 in natch07 removed, 0 observations found"
[1] "level 6 in nnatch removed, 0 observations found"
[1] "level 7 in nnatch removed, 0 observations found"
[1] "level 4 in nadoptch removed, 1 observations found"
[1] "level 6 in adoptch02 removed, 2 observations found"
[1] "level 4 in adoptch03 removed, 2 observations found"
[1] "level 6 in adoptch03 removed, 0 observations found"
[1] "level 7 in adoptch03 removed, 0 observations found"
[1] "level 5 in adoptch04 removed, 0 observations found"
[1] "level 6 in adoptch04 removed, 0 observations found"
[1] "level 7 in adoptch04 removed, 0 observations found"
[1] "level 6 in nchunder16 removed, 0 observations found"
[1] "level 5 in nch5to15 removed, 0 observations found"
[1] "level 6 in nch5to15 removed, 0 observations found"
[1] "level 4 in nch10to15 removed, 4 observations found"
[1] "level 1 in allch01 removed, 3 observations found"
[1] "level 7 in allch01 removed, 0 observations found"
[1] "level 8 in allch01 removed, 0 observations found"
[1] "level 7 in allch02 removed, 3 observations found"
[1] "level 8 in allch02 removed, 1 observations found"
[1] "level 9 in allch02 removed, 0 observations found"
[1] "level 9 in allch03 removed, 0 observations found"
[1] "level 10 in allch03 removed, 0 observations found"
[1] "level 7 in allch04 removed, 1 observations found"
[1] "level 8 in allch04 removed, 2 observations found"
[1] "level 9 in allch04 removed, 0 observations found"
[1] "level 8 in allch05 removed, 0 observations found"
[1] "level 9 in allch06 removed, 0 observations found"
[1] "level 9 in jbstat removed, 4 observations found"
[1] "level 11 in fruvege removed, 2 observations found"
[1] "level 13 in fruvege removed, 1 observations found"
[1] "level 15 in fruvege removed, 1 observations found"
[1] "level 20 in fruvege removed, 1 observations found"
[1] "level 5 in relup removed, 3 observations found"
[1] "level 1 in bensta3 removed, 2 observations found"
[1] "level 7 in marstat removed, 2 observations found"
[1] "level 9 in marstat removed, 1 observations found"
[1] "level 4 in ivcoop removed, 2 observations found"
[1] "level 5 in undqus removed, 1 observations found"
[1] "level 7 in hgbiom removed, 2 observations found"
[1] "level 5 in hgbiof removed, 2 observations found"
[1] "level 6 in hgbiof removed, 0 observations found"
[1] "level 6 in pn1pno removed, 0 observations found"
[1] "level 7 in pn1pno removed, 0 observations found"
[1] "level 7 in pn2pno removed, 0 observations found"
[1] "level 6 in pns1pno removed, 0 observations found"
[1] "level 7 in pns1pno removed, 0 observations found"
[1] "level 6 in pns2pno removed, 1 observations found"
[1] "level 7 in pns2pno removed, 0 observations found"
[1] "level 1 in fiyrinvinc_tc removed, 2 observations found"
[1] "level 10 in ff_jbstat removed, 3 observations found"
[1] "level 1 in ff_bentype25 removed, 4 observations found"
[1] "level 1 in ff_bentype32 removed, 3 observations found"
[1] "level 1 in ff_bentype36 removed, 2 observations found"
[1] "level 2 in ngrp_dv removed, 1 observations found"
[1] "level 4 in nnssib_dv removed, 2 observations found"
[1] "level 5 in nnssib_dv removed, 3 observations found"
[1] "level 6 in nnssib_dv removed, 1 observations found"
[1] "level 7 in mastat_dv removed, 0 observations found"
[1] "level 9 in mastat_dv removed, 0 observations found"
```

```
[1] "level 6 in buno_dv removed, 0 observations found"
[1] "level 6 in nchild_dv removed, 1 observations found"
[1] "level 6 in hrpno removed, 1 observations found"
[1] "level 6 in ppno removed, 2 observations found"
[1] "level 5 in sppno removed, 4 observations found"
[1] "level 5 in fnpno removed, 0 observations found"
[1] "level 6 in fnpno removed, 0 observations found"
[1] "level 5 in fnspno removed, 1 observations found"
[1] "level 6 in fnspno removed, 0 observations found"
[1] "level 7 in mnpno removed, 0 observations found"
[1] "level 7 in mnspno removed, 0 observations found"
[1] "level 1 in grfpno removed, 2 observations found"
[1] "level 2 in grfpno removed, 1 observations found"
[1] "level 5 in grfpno removed, 1 observations found"
[1] "level 2 in grmpno removed, 2 observations found"
[1] "level 4 in grmpno removed, 4 observations found"
[1] "level 7 in grmpno removed, 1 observations found"
[1] "level 3 in nunmpsp_dv removed, 2 observations found"
[1] "level 4 in nunmpsp_dv removed, 1 observations found"
[1] "level 4 in nnsib_dv removed, 0 observations found"
[1] "level 5 in nnsib_dv removed, 0 observations found"
[1] "level 21 in hhtype_dv removed, 0 observations found"
[1] "level 4 in nmpsp_dv removed, 2 observations found"
[1] "level 5 in nmpsp_dv removed, 2 observations found"
[1] "level 6 in nmpsp_dv removed, 1 observations found"
[1] "level 6 in ndepchl_dv removed, 0 observations found"
[1] "103 total levels removed from 59 different variables. In total 113 observations deleted"
[1] "----------Variance 0 Check----------"
[1] "63 variables removed since their new variance was 0"
 [1] "ivfio"          "ioutcome"       "newper"         "newentrant"     "adstatus"
"natch06"        "natch07"
 [8] "natch08"        "natch09"        "natch10"        "natch11"        "natch12"
"natch13"        "natch14"
[15] "natch15"        "natch16"        "adoptch04"      "adoptch05"      "adoptch06"
"adoptch07"      "adoptch08"
[22] "adoptch09"      "adoptch10"      "adoptch11"      "adoptch12"      "adoptch13"
"adoptch14"      "adoptch15"
[29] "adoptch16"      "allch06"        "allch07"        "allch08"        "allch09"
"allch10"        "allch11"
[36] "allch12"        "allch13"        "allch14"        "allch15"        "allch16"
"chkdob"         "wlk10m"
[43] "bensta3"        "indmode"        "sceverdrnk"     "screlany"       "scfrendany"
"intdatd_if"     "intdatm_if"
[50] "intdaty_if"     "doby_if"        "age_if"         "fiyrinvinc_tc" "ff_ivlolw"
"ff_everint"     "ff_bentype25"
[57] "ff_bentype32"   "ff_bentype36"   "ff_bentype37"   "grfpno"         "fiyrinvinc_if" "wave"
"scflag_dv"
[1] "----------Dummy Variables----------"
[1] "predictor variable count went from 413 to 944"
[1] "----------Variance 0 Check----------"
[1] "87 variables removed since their new variance was 0"
 [1] "pno.6"           "pno.7"           "nch415resp.5"    "nchund18resp.6" "natch01.7"
"natch01.8"       "natch02.2"
 [8] "natch02.7"       "natch02.8"       "natch02.9"       "natch03.3"
"natch03.7"       "natch03.9"       "natch03.10"
[15] "nnatch.6"        "nnatch.7"        "nadoptch.4"      "adoptch02.6"
"adoptch03.4"     "adoptch03.6"     "adoptch03.7"
[22] "nchunder16.6"    "nch5to15.5"      "nch5to15.6"      "nch10to15.4"
"allch01.1"       "allch01.7"       "allch01.8"
[29] "allch02.7"       "allch02.8"       "allch02.9"       "allch03.9"
"allch03.10"      "allch04.7"       "allch04.8"
[36] "allch04.9"       "allch05.8"       "jbstat.9"        "fruvege.11"
"fruvege.13"      "fruvege.15"      "fruvege.20"
[43] "relup.5"         "marstat.7"       "marstat.9"       "ivcoop.4"
"undqus.5"        "hgbiom.7"        "hgbiof.5"
[50] "hgbiof.6"        "pn1pno.6"        "pn1pno.7"        "pn2pno.7"
"pns1pno.6"       "pns1pno.7"       "pns2pno.6"
[57] "pns2pno.7"       "ff_jbstat.10"    "ngrp_dv.2"       "nnssib_dv.4"
```

```
"nnssib_dv.5"     "nnssib_dv.6"     "mastat_dv.7"
[64] "mastat_dv.9"     "buno_dv.6"       "nchild_dv.6"     "hrpno.6"
"ppno.6"          "sppno.5"         "fnpno.5"
[71] "fnpno.6"         "fnspno.5"        "fnspno.6"        "mnpno.7"
"mnspno.7"        "grmpno.2"        "grmpno.4"
[78] "grmpno.7"        "nunmpsp_dv.3"    "nunmpsp_dv.4"    "nnsib_dv.4"
"nnsib_dv.5"      "hhtype_dv.21"    "nmpsp_dv.4"
[85] "nmpsp_dv.5"      "nmpsp_dv.6"      "ndepchl_dv.6"
[1] "————————K–Means————————"
[1] "15 clusters have been made for K–Means"
[1] "K–Means results as a table, the max value in each row is a simple way to define which cluster re

      0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16
17  18  19  20  21  22  23  24  25  26  27  28  29
  1   1   1   9  11  16  29 168 164 144 125 149 146 143  63  40  29  30
21  19  12  10  16  13  10   7   4   2   4   2   2
  2   2   2   1   2  10  19  83  63  70  62  45  52  55  29  17  13  13
16  12   5   9   9   7   4   4   3   2   3   3   1
  3   2   3   1  14   9  29 115 120 122 120  93  97  86  44  33  32  21
19  18  18  12   9   9   1  12   5   6   5   2   2
  4   2   4   5   7  11  27 103  92  89  98  88  84  93  48  34  28  16
16  14   7   9  10   8   5   3   5   2   0   2   1
  5   1   0   2   5   9  18  73  99  77  79  61  79  96  36  29  23  23
10  12  11  12   7  13   9   3   5   4   0   0   0
  6   3   0   1   1   5  15  66  69  57  67  53  62  60  25  26  21   6
13   5   9   8   8   9   1   1   2   2   1   0   1
  7   3   0   4   3   1   7  30  46  43  34  28  38  38  19  10  16   6
6   7   5   2   1   4   3   1   2   0   0   3   2
  8   1   1   3   2   4  17  82  73  72  79  54  66  66  27  21  13  10
6  10  13   5   5   4   8   4   5   2   3   1   1
  9   1   1   0   5  10  14  74  82  52  65  59  50  68  30  20   9  11
11  14  15  11   7   4   4   5   5   4   1   2   0
  10   2   1   1   2   4  14  65  82  72  70  60  69  59  28  20  18   7
11  13   9   8   2   6   5   4   5   1   1   1   0
  11   3   6   3   2  12  19 124  94 105 113 104  94 120  48  33  27  30
18  17  11  15   8  13  11   5   6   1   4   2   1
  12   1   0   1   3   9  19  70  68  66  64  57  58  61  22  17  16  15
9   5   9   7   6   2  11   4   7   4   0   0   2
  13   2   4   2   6  11  17 118  76 103  91  86  66  96  35  33  19  20
11   7   9  10   4  11   5   6   6   0   3   2   2
  14   0   0   3   4   5   9  73  79  69  78  53  46  66  24  21  16  15
9  11   5   8   3   4   0   9   1   3   2   2   0
  15   0   1   2   0   9   9  75  80  87  60  66  62  73  31  12  14  15
10  11   7   8  13   8   2   1   0   1   2   2   2

     30  31  32  33  34  35  36
  1   1   0   1   1   0   0   1
  2   1   1   2   2   0   0   0
  3   2   1   2   2   0   1   0
  4   1   2   0   1   1   0   0
  5   1   0   1   0   0   1   0
  6   1   0   1   0   0   0   1
  7   1   1   0   1   1   1   0
  8   1   2   1   0   0   0   0
  9   0   0   0   0   0   2   0
  10  0   0   0   0   0   0   0
  11  2   1   1   1   0   0   0
  12  1   0   0   2   0   0   1
  13  2   0   1   1   0   0   1
  14  0   0   0   0   0   1   0
  15  0   1   1   0   1   0   1
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
 [1] "Cluster 1: Within MSE 17297070334945836, Size 1394" "Cluster 2: Within MSE 358122483885442, Siz
 [3] "Cluster 3: Within MSE 9425448519842926, Size 1067"  "Cluster 4: Within MSE 2853038387595288, S
 [5] "Cluster 5: Within MSE 2451215983932508, Size 799"   "Cluster 6: Within MSE 340166019576860, Siz
 [7] "Cluster 7: Within MSE 5046890580767, Size 367"      "Cluster 8: Within MSE 394667294467190, Siz
 [9] "Cluster 9: Within MSE 351624291257316, Size 636"    "Cluster 10: Within MSE 3473914728735625, S
[11] "Cluster 11: Within MSE 293062468494208, Size 1054" "Cluster 12: Within MSE 318805103876834, S
```

```
[13] "Cluster 13: Within MSE 2472032355437551, Size 866"  "Cluster 14: Within MSE 352823784327687, S
[15] "Cluster 15: Within MSE 3473434233665862, Size 667"
[1] "Total between cluster MSE: 626153876575018880, Total within cluster MSE: 4324086962094729"
[1] "The K-Means model predicts exactly with an accuracy of 0.1227"
[1] "----------Correlation Checks----------"
[1] "indpxus_xw removed, correlated with 5 other variable(s)"
[1] "month removed, correlated with 5 other variable(s)"
[1] "dvage removed, correlated with 4 other variable(s)"
[1] "indinus_xw removed, correlated with 4 other variable(s)"
[1] "indpxus_lw removed, correlated with 5 other variable(s)"
[1] "nchunder16.5 removed, correlated with 4 other variable(s)"
[1] "pensioner_dv.2 removed, correlated with 4 other variable(s)"
[1] "sex.2 removed, correlated with 3 other variable(s)"
[1] "birthy removed, correlated with 3 other variable(s)"
[1] "nchunder16.3 removed, correlated with 3 other variable(s)"
[1] "relup.2 removed, correlated with 3 other variable(s)"
[1] "pn2pno.2 removed, correlated with 3 other variable(s)"
[1] "allch05.7 removed, correlated with 3 other variable(s)"
[1] "pidp removed, correlated with 2 other variable(s)"
[1] "pno.2 removed, correlated with 2 other variable(s)"
[1] "hhorig.7 removed, correlated with 2 other variable(s)"
[1] "nch14resp.2 removed, correlated with 2 other variable(s)"
[1] "nch14resp.3 removed, correlated with 2 other variable(s)"
[1] "nchunder16.2 removed, correlated with 2 other variable(s)"
[1] "nchunder16.4 removed, correlated with 2 other variable(s)"
[1] "allch03.5 removed, correlated with 2 other variable(s)"
[1] "chksex.2 removed, correlated with 2 other variable(s)"
[1] "relup.6 removed, correlated with 2 other variable(s)"
[1] "marstat.2 removed, correlated with 2 other variable(s)"
[1] "marstat.4 removed, correlated with 2 other variable(s)"
[1] "marstat.6 removed, correlated with 2 other variable(s)"
[1] "hgbiom.1 removed, correlated with 2 other variable(s)"
[1] "hgbiom.2 removed, correlated with 2 other variable(s)"
[1] "hgbiom.3 removed, correlated with 2 other variable(s)"
[1] "hgbiom.4 removed, correlated with 2 other variable(s)"
[1] "hgbiom.5 removed, correlated with 2 other variable(s)"
[1] "hgbiof.1 removed, correlated with 2 other variable(s)"
[1] "hgbiof.2 removed, correlated with 2 other variable(s)"
[1] "hgbiof.4 removed, correlated with 2 other variable(s)"
[1] "respm16.2 removed, correlated with 2 other variable(s)"
[1] "scdoby4 removed, correlated with 2 other variable(s)"
[1] "pns2pno.2 removed, correlated with 2 other variable(s)"
[1] "fimngrs_tc.1 removed, correlated with 2 other variable(s)"
[1] "ff_bentype01.1 removed, correlated with 2 other variable(s)"
[1] "cohab_dv.1 removed, correlated with 2 other variable(s)"
[1] "nchild_dv.4 removed, correlated with 3 other variable(s)"
[1] "fnspno.1 removed, correlated with 3 other variable(s)"
[1] "indpxbh_xw removed, correlated with 2 other variable(s)"
[1] "indinub_xw removed, correlated with 2 other variable(s)"
[1] "fimnlabgrs_dv removed, correlated with 2 other variable(s)"
[1] "fimngrs_dv removed, correlated with 3 other variable(s)"
[1] "nchresp.1 removed, correlated with 2 other variable(s)"
[1] "nchresp.4 removed, correlated with 2 other variable(s)"
[1] "indinus_lw removed, correlated with 2 other variable(s)"
[1] "hidp removed, correlated with 1 other variable(s)"
[1] "pno.3 removed, correlated with 1 other variable(s)"
[1] "pno.4 removed, correlated with 1 other variable(s)"
[1] "hhorig.2 removed, correlated with 1 other variable(s)"
[1] "hhorig.3 removed, correlated with 1 other variable(s)"
[1] "hhorig.4 removed, correlated with 1 other variable(s)"
[1] "hhorig.5 removed, correlated with 1 other variable(s)"
[1] "hhorig.6 removed, correlated with 1 other variable(s)"
[1] "memorig.7 removed, correlated with 1 other variable(s)"
[1] "strata removed, correlated with 1 other variable(s)"
[1] "nch14resp.5 removed, correlated with 1 other variable(s)"
[1] "nchresp.2 removed, correlated with 1 other variable(s)"
[1] "nchresp.3 removed, correlated with 1 other variable(s)"
[1] "natch03.5 removed, correlated with 1 other variable(s)"
```

```
[1] "natch04.6 removed, correlated with 1 other variable(s)"
[1] "natch05.7 removed, correlated with 1 other variable(s)"
[1] "nadoptch.3 removed, correlated with 1 other variable(s)"
[1] "nchunder16.1 removed, correlated with 1 other variable(s)"
[1] "istrtdaty.2011 removed, correlated with 1 other variable(s)"
[1] "istrtdaty.2012 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.2 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.3 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.4 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.5 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.6 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.7 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.8 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.9 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.10 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.11 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.12 removed, correlated with 1 other variable(s)"
[1] "istrtdatd removed, correlated with 1 other variable(s)"
[1] "jbstat.4 removed, correlated with 1 other variable(s)"
[1] "jbhas.2 removed, correlated with 1 other variable(s)"
[1] "btype5.1 removed, correlated with 1 other variable(s)"
[1] "fiyrdia removed, correlated with 1 other variable(s)"
[1] "arts1b13.1 removed, correlated with 1 other variable(s)"
[1] "marstat.3 removed, correlated with 1 other variable(s)"
[1] "hgbiof.3 removed, correlated with 1 other variable(s)"
[1] "respf16.2 removed, correlated with 1 other variable(s)"
[1] "scsex.2 removed, correlated with 1 other variable(s)"
[1] "scsf2a.2 removed, correlated with 1 other variable(s)"
[1] "scsf2b.2 removed, correlated with 1 other variable(s)"
[1] "scsf3a.5 removed, correlated with 1 other variable(s)"
[1] "scmolwp.2 removed, correlated with 1 other variable(s)"
[1] "scrletdwn.3 removed, correlated with 1 other variable(s)"
[1] "scfcritic.3 removed, correlated with 1 other variable(s)"
[1] "scfletdwn.3 removed, correlated with 1 other variable(s)"
[1] "scfannoy.3 removed, correlated with 1 other variable(s)"
[1] "istrtdathh removed, correlated with 1 other variable(s)"
[1] "pn1pno.1 removed, correlated with 1 other variable(s)"
[1] "pn1pno.2 removed, correlated with 1 other variable(s)"
[1] "pn1pno.3 removed, correlated with 1 other variable(s)"
[1] "pn1pno.4 removed, correlated with 1 other variable(s)"
[1] "pn1pno.5 removed, correlated with 1 other variable(s)"
[1] "pn2pno.3 removed, correlated with 1 other variable(s)"
[1] "pn2pno.5 removed, correlated with 1 other variable(s)"
[1] "fimnlabgrs_tc.1 removed, correlated with 1 other variable(s)"
[1] "j2paynet_dv removed, correlated with 1 other variable(s)"
[1] "ff_emplw.2 removed, correlated with 1 other variable(s)"
[1] "age_dv removed, correlated with 1 other variable(s)"
[1] "npn_dv.1 removed, correlated with 1 other variable(s)"
[1] "npn_dv.2 removed, correlated with 1 other variable(s)"
[1] "ngrp_dv.1 removed, correlated with 1 other variable(s)"
[1] "nnssib_dv.1 removed, correlated with 1 other variable(s)"
[1] "nnssib_dv.2 removed, correlated with 1 other variable(s)"
[1] "nnssib_dv.3 removed, correlated with 1 other variable(s)"
[1] "country.2 removed, correlated with 1 other variable(s)"
[1] "country.3 removed, correlated with 1 other variable(s)"
[1] "country.4 removed, correlated with 1 other variable(s)"
[1] "xtra5min_dv.1 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.5 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.6 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.7 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.8 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.9 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.10 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.11 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.12 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.13 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.15 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.13 removed, correlated with 1 other variable(s)"
```

```
[1] "mastat_dv.2 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.4 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.5 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.6 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.10 removed, correlated with 1 other variable(s)"
[1] "nchild_dv.2 removed, correlated with 1 other variable(s)"
[1] "nchild_dv.3 removed, correlated with 1 other variable(s)"
[1] "nchild_dv.5 removed, correlated with 1 other variable(s)"
[1] "ppno.1 removed, correlated with 1 other variable(s)"
[1] "ppno.2 removed, correlated with 1 other variable(s)"
[1] "fnpno.2 removed, correlated with 1 other variable(s)"
[1] "fnpno.4 removed, correlated with 1 other variable(s)"
[1] "mnpno.1 removed, correlated with 1 other variable(s)"
[1] "mnpno.2 removed, correlated with 1 other variable(s)"
[1] "mnpno.3 removed, correlated with 1 other variable(s)"
[1] "mnpno.4 removed, correlated with 1 other variable(s)"
[1] "mnpno.5 removed, correlated with 1 other variable(s)"
[1] "paygu_if.1 removed, correlated with 1 other variable(s)"
[1] "fimnlabgrs_if removed, correlated with 1 other variable(s)"
[1] "indinbh_xw removed, correlated with 1 other variable(s)"
[1] "indpxub_xw removed, correlated with 1 other variable(s)"
[1] "respm16_dv.2 removed, correlated with 1 other variable(s)"
[1] "hhtype_dv.11 removed, correlated with 1 other variable(s)"
[1] "hhtype_dv.12 removed, correlated with 1 other variable(s)"
[1] "ethn_dv removed, correlated with 1 other variable(s)"
[1] "156 variables removed since they had high correlation coefs"
[1] "The final cleaned dataset has been completed at this stage and is stored under the name 'cleanDa
[1] "——————Attempting a Train Test Split——————"
[1] "Good train, test split found"
[1] "The working seed found was 3"
[1] "——————kNN——————"
[1] "234 neighbours considered for each test data point"
[1] "kNN results as a table, follow the diagonal for the correctly mapped clusters"
         real
predicted   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15
16  17  18  19  20  21  22  23  24  25  26  27  28
        0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        2   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        3   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        4   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        5   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        6   4   3   6   9  17  40 211 194 156 115  74  83  72  27  19  13
11   3   5   1   3   3   2   1   0   0   0   0   0
        7   0   2   2   4   3  10  91  75  78  61  53  61  46  18  17  11
3   3   6   3   2   3   1   0   0   0   0   0   0
        8   2   1   0   4   4   5  29  41  34  38  32  36  34  10   8   8
7   2   7   1   3   1   2   1   0   0   1   0   0
        9   1   0   1   2   4   4  13  20  24  37  26  35  44  19  16   8
11  10  10   2   2   1   4   2   1   1   0   0   0
        10  0   0   0   0   2   1   3   5  11  12  19  22  23  11  11   9
5   2   2   4   1   1   2   5   1   1   0   0   0
        11  0   0   0   0   0   0   0   1   3   2   4  11  14   7   8   3
7   5   1   1   3   3   2   0   3   3   3   1   1
        12  0   0   1   0   1   1   2   4   6   9  24  31  53  29  25  22
19  16  22  24  17  16  15  11  16  14   5   5   3
        13  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        14  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        15  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
```

```
          16   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0    0   0   0   0   0   0   0   0   0   0   0
          17   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0    0   0   0   0   0   0   0   0   0   0   0
          18   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0    0   0   0   0   0   0   0   0   0   0   0
          19   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0    0   0   0   0   0   0   0   0   0   0   0
          20   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0    0   0   0   0   0   0   0   0   0   0   0
          21   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0    0   0   0   0   0   0   0   0   0   0   0
          22   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0    0   0   0   0   0   0   0   0   0   0   0
          23   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0    0   0   0   0   0   0   0   0   0   0   0
          24   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0    0   0   0   0   0   0   0   0   0   0   0
          25   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0    0   0   0   0   0   0   0   0   0   0   0
          26   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0    0   0   0   0   0   0   0   0   0   0   0
         real
predicted  29  30  31  32  33  35  36
        0   0   0   0   0   0   0   0
        1   0   0   0   0   0   0   0
        2   0   0   0   0   0   0   0
        3   0   0   0   0   0   0   0
        4   0   0   0   0   0   0   0
        5   0   0   0   0   0   0   0
        6   0   0   0   0   0   0   0
        7   0   0   0   0   0   0   0
        8   0   0   0   1   0   1   0
        9   0   0   0   0   0   0   0
        10  0   0   0   2   0   0   0
        11  0   1   0   0   0   0   0
        12  5   3   1   0   3   0   1
        13  0   0   0   0   0   0   0
        14  0   0   0   0   0   0   0
        15  0   0   0   0   0   0   0
        16  0   0   0   0   0   0   0
        17  0   0   0   0   0   0   0
        18  0   0   0   0   0   0   0
        19  0   0   0   0   0   0   0
        20  0   0   0   0   0   0   0
        21  0   0   0   0   0   0   0
        22  0   0   0   0   0   0   0
        23  0   0   0   0   0   0   0
        24  0   0   0   0   0   0   0
        25  0   0   0   0   0   0   0
        26  0   0   0   0   0   0   0
 [ reached getOption("max.print") -- omitted 10 rows ]
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
[1] "The MSE of the predicted values are of 103.5538"
[1] "The kNN model predicts exactly with an accuracy of 0.1527"
[1] "------------CART prediction model------------"
n= 8644

node), split, n, deviance, yval
      * denotes terminal node

 1) root 8644 196955.10 10.624480
   2) sf12mcs_dv>=-0.8633253 7175  75632.03  9.318746
     4) scsf6c.5>=0.5 3642  18824.95  7.736409 *
     5) scsf6c.5< 0.5 3533  38288.13 10.949900
      10) sf12mcs_dv>=-0.2421288 2465  20420.01 10.324140 *
      11) sf12mcs_dv< -0.2421288 1068  14675.04 12.394190 *
   3) sf12mcs_dv< -0.8633253 1469  49340.99 17.002040
```

```
     6) sf12mcs_dv>=-2.159634 1147  25732.18 15.477770 *
     7) sf12mcs_dv< -2.159634 322  11451.00 22.431680 *
[1] "Variable Importance"
sf12mcs_dv   scsf4a.3 sf12pcs_dv   scsf4b.3   scsf6c.5   scsf6c.4   scsf6c.2
scsf6a.4   scsf4a.5   scsf6a.3   scsf4a.2
99698.1084 22589.3256 18767.3168 18669.2691 18518.9440 15125.2782 14874.7192 14457.1914
4675.5702  3758.3025  2831.7890
   scsf7.2   scsf4b.2   scsf6c.3
 1397.0159  1245.9871   406.6088
[1] "The MSE of the predicted values are of 10.7688"
[1] "The CART model predicts exactly with accuracy of 0.1464"
[1] "------------Ordinary Linear Regression (Initial)------------"
[1] "The full model AIC is: 43451.4718"
[1] "------------Variance Inflation Factor Removal------------"
[1] "The variable sf12mcs_dv was removed since it had a VIF score of 15623.3772"
[1] "The variable natch02.4 was removed since it had a VIF score of 1144.9765"
[1] "The variable sppno.2 was removed since it had a VIF score of 174.1914"
[1] "The variable scsf4b.5 was removed since it had a VIF score of 165.6443"
[1] "The variable hrpid was removed since it had a VIF score of 160.8802"
[1] "The variable agegr13_dv.9 was removed since it had a VIF score of 99.5937"
[1] "The variable sf12pcs_dv was removed since it had a VIF score of 80.9875"
[1] "The variable adoptch01.3 was removed since it had a VIF score of 63.3403"
[1] "The variable fibenothr_dv was removed since it had a VIF score of 62.8912"
[1] "The variable scsf4a.5 was removed since it had a VIF score of 62.425"
[1] "The variable pns1pno.1 was removed since it had a VIF score of 56.4706"
[1] "The variable scsf7.5 was removed since it had a VIF score of 53.4345"
[1] "The variable doby_dv was removed since it had a VIF score of 52.3631"
[1] "The variable allch01.3 was removed since it had a VIF score of 51.8324"
[1] "The variable wkvege.4 was removed since it had a VIF score of 44.9598"
[1] "The variable rach16_dv.2 was removed since it had a VIF score of 41.3013"
[1] "The variable scsf6c.5 was removed since it had a VIF score of 39.8368"
[1] "The variable ndepchl_dv.2 was removed since it had a VIF score of 36.6416"
[1] "The variable sclfsato.6 was removed since it had a VIF score of 35.4007"
[1] "The variable bensta96.1 was removed since it had a VIF score of 32.0386"
[1] "The variable scsf3b.5 was removed since it had a VIF score of 28.4894"
[1] "The variable access.5 was removed since it had a VIF score of 26.8993"
[1] "The variable sclfsat1.6 was removed since it had a VIF score of 21.9783"
[1] "The variable hhtype_dv.8 was removed since it had a VIF score of 19.9613"
[1] "The variable nnatch.4 was removed since it had a VIF score of 18.5076"
[1] "The variable allch02.4 was removed since it had a VIF score of 17.8964"
[1] "The variable nnmpsp_dv.1 was removed since it had a VIF score of 15.0857"
[1] "The variable natch01.3 was removed since it had a VIF score of 14.0398"
[1] "The variable scopfamf.4 was removed since it had a VIF score of 13.2954"
[1] "The variable marstat_dv.6 was removed since it had a VIF score of 13.0452"
[1] "The variable scrannoy.3 was removed since it had a VIF score of 12.388"
[1] "The variable netuse.7 was removed since it had a VIF score of 12.2953"
[1] "The variable sclfsat7.6 was removed since it had a VIF score of 11.8198"
[1] "The variable scrcritic.3 was removed since it had a VIF score of 11.0389"
[1] "The variable sclfsat2.6 was removed since it had a VIF score of 10.5328"
[1] "The variable nunmpsp_dv.1 was removed since it had a VIF score of 10.4218"
[1] "36 variables removed from the Ordinary Linear Model since they have a VIF score higher than 10"
[1] "The full model AIC after VIF checks is: 44044.1549"
[1] "------------Backwards Selection------------"
[1] "50 out of 682 variables removed so far."
[1] "100 out of 682 variables removed so far."
[1] "150 out of 682 variables removed so far."
[1] "200 out of 682 variables removed so far."
[1] "250 out of 682 variables removed so far."
[1] "300 out of 682 variables removed so far."
[1] "350 out of 682 variables removed so far."
[1] "400 out of 682 variables removed so far."
[1] "450 out of 682 variables removed so far."
[1] "500 out of 682 variables removed so far."
[1] "538 out of 682 variables removed in backwards selection since they weren't significant at the 95
  [1] "arts1b9.1"        "ppno.5"          "usbread.6"        "scopfamb.3"
"allch01.2"        "agegr13_dv.11"
  [7] "sports17.1"       "scopfamd.3"      "jbstat.5"         "vote6.4"
"access.2"         "intdatm_dv.8"
```

```
 [13] "scalcl7d.2"        "hcondn16.1"        "sctimemnuf.4"      "arts2b10.1"
"hiqual_dv.9"       "sctimemnuf.3"
 [19] "nchund18resp.1"    "sports230.1"       "adoptch02.4"       "sports32.1"
"hhtype_dv.17"      "gor_dv.4"
 [25] "sclfsat7.7"        "nnewborn.1"        "indscus_xw"        "allch04.6"
"arts1a3.1"         "mla2.1"
 [31] "scopfamb.4"        "scwhorurac.2"      "btype9.1"          "allch02.6"
"scfannoy.4"        "ppen.2"
 [37] "pns1pno.3"         "sports219.1"       "xpmove.2"          "scwhorufam.4"
"ff_jbstat.3"       "hiqual_dv.2"
 [43] "adoptch01.2"       "hcondn8.1"         "nmpsp_dv.2"        "nch5to15.1"
"nch5to15.2"        "fimnlabnet_dv"
 [49] "rhland_code.1"     "nch415resp.3"      "fruvege.7"         "event4s"
"allch02.3"         "natch02.3"
 [55] "allch03.7"         "sports223.1"       "undqus.2"          "ivcoop.2"
"gor_dv.9"          "sports33.1"
 [61] "scopfamf.3"        "undqus.4"          "sports11.1"        "scsf1.2"
"scssupr2r.9"       "hiqual_dv.3"
 [67] "hrpno.5"           "pns1pno.2"         "scrannoy.4"        "fibenothr_tc.1"
"arts2a96.1"        "heritage3.1"
 [73] "scwhorusex.4"      "scrrely.4"         "adoptch01.4"       "sctimemnuf.5"
"scrrely.2"         "sports14.1"
 [79] "usdairy.4"         "fruvege.4"         "fimnprben_dv"      "agegr5_dv.14"
"agegr10_dv.8"      "jbiindb_dv"
 [85] "seearngrs_if.1"    "fimngrs_if"        "paynu_if.1"        "arts2a5.1"
"heritage4.1"       "event3s"
 [91] "scopfamd.4"        "netuse.3"          "j2has.2"           "usbread.3"
"npensioner_dv.3"   "pns1pno.5"
 [97] "nnatch.3"          "scloutcont.3"      "fimnsben_dv"       "fimnnet_dv"
"sports112.1"       "memorig.4"
[103] "ff_bentype07.1"    "ff_bentype23.1"    "agegr13_dv.12"     "arts2b11.1"
"scopfamf.2"        "hcondn7.1"
[109] "fruvege.12"        "hhtype_dv.10"      "hcondn4.1"         "jbstat.10"
"employ.2"          "lenindintv"
[115] "marstat_dv.4"      "ff_bentype09.1"    "usdairy.3"         "usdairy.5"
"heritage1.1"       "wkvege.2"
[121] "nunmpsp_dv.2"      "scrundstnd.3"      "scropenup.4"       "wlk30min"
"scwhorusex.2"      "arts1a4.1"
[127] "walkpace.4"        "scage1drink"       "ff_bentype13.1"    "sports396.1"
"hrpno.3"           "scfannoy.2"
[133] "save.2"            "nch10to15.2"       "scfalcdrnk.2"      "agegr13_dv.4"
"heritage6.1"       "sports12.1"
[139] "fnspno.2"          "scwhoruage.4"      "npensioner_dv.1"   "npensioner_dv.2"
"allch02.5"         "access.4"
[145] "netuse.4"          "scopfamh.4"        "hcondn14.1"        "hcondn96.1"
"sf1.3"             "scsf1.3"
[151] "memorig.2"         "btype8.1"          "event1s"           "scssupr2r.3"
"fibenothr_if"      "fruvege.10"
[157] "origadd.2"         "istrtdatss"        "fimnmisc_dv"       "hrpno.4"
"sports296.1"       "natch01.1"
[163] "mnspno.1"          "scrrely.3"         "sampst.3"          "memorig.3"
"memorig.5"         "ff_bentype38.1"
[169] "sports15.1"        "mla96.1"           "netuse.2"          "scrannoy.2"
"ff_bentype08.1"    "xtra5minosm_dv.1"
[175] "ff_bentype11.1"    "btype3.1"          "ndepchl_dv.3"      "nchund18resp.3"
"heritage96.1"      "wkfruit.3"
[181] "wkfruit.4"         "sppno.3"           "btype1.1"          "nnmpsp_dv.2"
"sclfsat1.3"        "arts1a7.1"
[187] "intdatm_dv.11"     "scfundstnd.4"      "sports35.1"        "ff_bentype21.1"
"sports13.1"        "jbstat.2"
[193] "nmpsp_dv.1"        "arts2a4.1"         "wkvege.3"          "usbread.7"
"mla3.1"            "chargv.2"
[199] "sports19.1"        "nnsib_dv.2"        "scsf3a.2"          "scopfama.5"
"ff_bentype18.1"    "btype96.1"
[205] "scopfamh.5"        "mobuse.2"          "heritage7.1"       "hcondn11.1"
"nnewborn.2"        "gor_dv.6"
[211] "gor_dv.5"          "gor_dv.2"          "hhtype_dv.16"      "bensta6.1"
"relup.4"           "fnpno.3"
```

```
[217] "respf16_dv.2"       "scwhoruedu.4"       "hiqual_dv.4"        "nch415resp.4"
"nchund18resp.5"    "scfalcdrnk.6"
[223] "scfalcdrnk.7"       "ff_bentype17.1"     "sports221.1"        "ppno.4"
"sppno.4"           "fruvege.9"
[229] "pns2pno.4"          "pns2pno.5"          "arts2b13.1"         "scopfamb.5"
"undqus.3"          "ivcoop.3"
[235] "intdatm_dv.3"       "event2s"            "sampst.2"           "nadoptch.1"
"daywlk"            "bensta5.1"
[241] "marstat_dv.3"       "ppno.3"             "scopfamb.2"         "scrletdwn.2"
"scrcritic.2"       "arts2b14.1"
[247] "sclfsat1.7"         "access.6"           "fruvege.8"          "scsf3a.4"
"scsf3a.3"          "scsf3b.4"
[253] "sclfsat1.4"         "sf1.2"              "scfopenup.2"        "scfopenup.3"
"scfrely.2"         "scfrely.3"
[259] "sports36.1"         "hiqual_dv.5"        "susp.3"             "nchresp.5"
"sclfsat1.5"        "indscbh_xw"
[265] "ienddatss"          "sclfsat2.2"         "ff_bentype06.1"     "j2pay_dv"
"fimnlabnet_tc.1"   "btype7.1"
[271] "ff_bentype26.1"     "sports196.1"        "heritage2.1"        "jbstat.97"
"ff_bentype22.1"    "indscus_lw"
[277] "hcondn15.1"         "scssup1.4"          "intdaty_dv.2011"    "ivprsnt.2"
"hcondn1.1"         "scwhorurac.3"
[283] "scwhorusex.3"       "health.2"           "scssupr2r.8"        "arts2a1.1"
"pn2pno.4"          "natch02.6"
[289] "ff_jbstat.97"       "hcondn13.1"         "scssupr2r.5"        "scopfamh.2"
"scopfamh.3"        "allch03.6"
[295] "nchild_dv.1"        "ndepchl_dv.1"       "hhtype_dv.20"       "nnatch.2"
"btype6.1"          "gor_dv.8"
[301] "scfalcdrnk.3"       "sports217.1"        "arts1b12.1"         "arts1b96.1"
"trainany.2"        "sports38.1"
[307] "scssup1.5"          "ivtrans.2"          "vote1.2"            "natch01.4"
"hcondn3.1"         "ndepchl_dv.5"
[313] "natch05.6"          "nch14resp.1"        "nch415resp.1"       "ff_bentype34.1"
"ff_bentype03.1"    "scfletdwn.2"
[319] "fruvege.6"          "fruvege.5"          "sportact.10"        "sportact.1"
"sportact.4"        "sportact.7"
[325] "sportact.9"         "sportact.8"         "aidhh.2"            "psu"
"ff_bentype12.1"    "marstat_dv.2"
[331] "fnpno.1"            "mnspno.2"           "sports39.1"         "arts2b15.1"
"sppno.1"           "agegr13_dv.10"
[337] "sports229.1"        "natch01.2"          "scsf2a.3"           "intdaty_dv.2012"
"sportact.3"        "intdatd_dv"
[343] "nnsib_dv.1"         "howlng"             "sports226.1"        "finnow.2"
"sports227.1"       "scfalcdrnk.5"
[349] "natch02.5"          "mla1.1"             "heritage5.1"        "scopfamd.5"
"nch5to15.3"        "sports225.1"
[355] "intdatm_dv.10"      "intdatm_dv.9"       "arts2a3.1"          "scwkvfast.2"
"arts1b15.1"        "scopfama.2"
[361] "walkpace.3"         "ff_jbstat.4"        "ff_jbstat.7"        "ff_jbstat.8"
"ff_bentype14.1"    "ff_jbstat.2"
[367] "ff_jbstat.5"        "scfletdwn.4"        "scrcritic.4"        "scrundstnd.2"
"scssupr2r.2"       "pno.5"
[373] "buno_dv.5"          "gor_dv.7"           "lkmove.2"           "hcondn5.1"
"aidxhh.2"          "intdatm_dv.5"
[379] "adoptch01.1"        "jbstat.7"           "depchl_dv.2"        "ff_bentype05.1"
"sclfsat2.7"        "sclfsat2.5"
[385] "cindtime"           "istrtdatmm"         "sclfsat2.4"         "hhresp_dv.2"
"allch01.5"         "susp.2"
[391] "agegr13_dv.5"       "agegr13_dv.7"       "kidlang"            "scfalcdrnk.4"
"scopfamd.2"        "sportact.2"
[397] "bensta8.1"          "hcondn12.1"         "hhtype_dv.18"       "marstat_dv.5"
"hhtype_dv.22"      "grmpno.1"
[403] "hhtype_dv.6"        "usbread.2"          "usbread.5"          "usbread.4"
"j2pay_if.1"        "smever.2"
[409] "hcondn10.1"         "natch04.7"          "indin91_lw"         "indin01_lw"
"pns2pno.3"         "sports16.1"
[415] "agegr13_dv.6"       "ind5mus_lw"         "scanyelsetxt.2"     "sports111.1"
"netuse.6"          "arts1b11.1"
```

```
[421] "ff_bentype02.1"    "ff_bentype04.1"    "scopfama.3"        "arts2b9.1"
"ff_bentype27.1"    "ff_bentype24.1"
[427] "ff_bentype31.1"    "nch14resp.4"       "nchund18resp.4"    "natch04.5"
"allch04.5"         "allch03.4"
[433] "arts1b10.1"        "arts2a2.1"         "sports113.1"       "sports231.1"
"ff_bentype30.1"    "ff_bentype15.1"
[439] "bensta1.1"         "fiyrinvinc_dv"     "hhtype_dv.19"      "arts1a6.1"
"scwhorurac.4"      "agegr13_dv.8"
[445] "sclfsat7.2"        "sclfsat7.5"        "scssupr2r.4"       "scwhoruage.3"
"memorig.6"         "nch10to15.3"
[451] "heritage8.1"       "sctimemnuf.6"      "nmpsp_dv.3"        "sports216.1"
"sports220.1"       "natch01.6"
[457] "allch01.6"         "arts1a1.1"         "sports228.1"       "intdatm_dv.6"
"intdatm_dv.4"      "sports37.1"
[463] "scfrely.4"         "scfopenup.4"       "bensta7.1"         "ff_bentype29.1"
"scrletdwn.4"       "scfcritic.4"
[469] "gor_dv.3"          "ff_bentype28.1"    "scsf1.4"           "sf1.4"
"marstat.5"         "hcondn9.1"
[475] "hhresp_dv.3"       "volun.2"           "urban_dv.2"        "fimninvnet_dv"
"gor_dv.10"         "arts1a96.1"
[481] "scwhorupol.2"      "scwhorupol.3"      "vote6.2"           "vote6.3"
"sports34.1"        "scssup1.2"
[487] "mastat_dv.3"       "fruvege.3"         "fruvege.2"         "arts2a6.1"
"sclfsat2.3"        "scopfamf.5"
[493] "nadoptch.2"        "adoptch02.3"       "hhtype_dv.23"      "natch03.6"
"nnsib_dv.3"        "agegr13_dv.3"
[499] "usdairy.2"         "usdairy.6"         "gor_dv.11"         "scwhoruage.2"
"relup.3"           "arts2b12.1"
[505] "ff_bentype20.1"    "nch10to15.1"       "allch01.4"         "nch5to15.4"
"sportact.6"        "bensta2.1"
[511] "advvoucher.2"      "pns1pno.4"         "arts1b14.1"        "natch01.5"
"finfut.3"          "sports218.1"
[517] "intdatm_dv.7"      "scsf3b.2"          "fimnpen_dv"        "btype4.1"
"sports18.1"        "scloutcont.2"
[523] "chkcoa.2"          "bensta4.1"         "ff_bentype19.1"    "ftedany.2"
"ff_jbstat.9"       "scsf5.3"
[529] "scsf5.2"           "scsf5.4"           "hrpno.2"           "sportact.5"
"scwkvfast.4"       "scwkvfast.3"
[535] "nchund18resp.2"    "nch415resp.2"      "sports224.1"       "ff_bentype33.1"
[1] "————————Ordinary Linear Regression (Improved)————————"

Call:
lm(formula = y ~ ., data = as.data.frame(x.data.linear))

Residuals:
     Min       1Q   Median       3Q      Max
-14.7725  -1.7039  -0.1596   1.5178  18.7197

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.47886    0.27389  23.655  < 2e-16 ***
natch03.4      0.72978    0.33858   2.155 0.031159 *
natch05.8     -5.95319    2.98413  -1.995 0.046080 *
adoptch01.5   -2.19938    0.94384  -2.330 0.019816 *
jbstat.3       0.69030    0.18978   3.637 0.000277 ***
jbstat.6       0.68424    0.18659   3.667 0.000247 ***
jbstat.8       1.30535    0.33525   3.894 9.95e-05 ***
netuse.5       0.27019    0.12598   2.145 0.032005 *
sf1.5          0.83916    0.34873   2.406 0.016136 *
wkfruit.2     -0.22442    0.07422  -3.024 0.002505 **
walkpace.2    -0.16041    0.06520  -2.460 0.013910 *
hcondn2.1     -0.40192    0.20165  -1.993 0.046278 *
hcondn6.1      1.62524    0.68946   2.357 0.018432 *
hcondn17.1     0.95099    0.27616   3.444 0.000577 ***
X.next.1      -0.37738    0.17812  -2.119 0.034149 *
btype2.1      -0.81909    0.24595  -3.330 0.000871 ***
finnow.3       0.33878    0.07907   4.284 1.85e-05 ***
finnow.4       0.77283    0.13581   5.691 1.31e-08 ***
```

finnow.5	1.68204	0.24757	6.794	1.16e−11	***
finfut.2	0.21088	0.08331	2.531	0.011377	*
arts1a2.1	−0.37245	0.14976	−2.487	0.012904	*
arts1a5.1	0.50497	0.20313	2.486	0.012940	*
arts2a7.1	−0.31626	0.15585	−2.029	0.042460	*
arts2b96.1	−0.27238	0.06986	−3.899	9.73e−05	***
sports110.1	0.29621	0.10933	2.709	0.006755	**
sports114.1	−0.60064	0.18478	−3.251	0.001156	**
sports222.1	0.26841	0.10488	2.559	0.010511	*
sports31.1	−0.18389	0.08524	−2.157	0.031016	*
access.3	−0.25604	0.12035	−2.128	0.033401	*
event1	−0.08702	0.03244	−2.682	0.007331	**
scsf1.5	1.09615	0.40281	2.721	0.006516	**
scsf2b.3	0.18831	0.09245	2.037	0.041681	*
scsf3b.3	−0.38167	0.12440	−3.068	0.002161	**
scsf4a.2	2.98397	0.33994	8.778	< 2e−16	***
scsf4a.3	1.21011	0.17743	6.820	9.72e−12	***
scsf4a.4	0.69669	0.11433	6.094	1.15e−09	***
scsf4b.2	1.10223	0.36769	2.998	0.002728	**
scsf4b.3	0.94789	0.19048	4.976	6.61e−07	***
scsf4b.4	0.40022	0.11487	3.484	0.000496	***
scsf5.5	0.75923	0.33984	2.234	0.025503	*
scsf6a.2	0.62171	0.16236	3.829	0.000129	***
scsf6a.3	1.42665	0.17829	8.002	1.39e−15	***
scsf6a.4	2.98460	0.20977	14.228	< 2e−16	***
scsf6a.5	4.76696	0.32854	14.509	< 2e−16	***
scsf6b.2	0.63198	0.17280	3.657	0.000256	***
scsf6b.3	1.03664	0.18200	5.696	1.27e−08	***
scsf6b.4	1.30697	0.20730	6.305	3.03e−10	***
scsf6b.5	2.00900	0.29097	6.905	5.40e−12	***
scsf6c.2	5.08327	0.21887	23.225	< 2e−16	***
scsf6c.3	2.91122	0.11511	25.291	< 2e−16	***
scsf6c.4	1.47599	0.08068	18.295	< 2e−16	***
scsf7.2	3.71653	0.25509	14.569	< 2e−16	***
scsf7.3	1.28679	0.13515	9.521	< 2e−16	***
scsf7.4	0.85556	0.09638	8.877	< 2e−16	***
scwhoruedu.2	0.19306	0.07841	2.462	0.013830	*
scwhoruedu.3	0.22484	0.08977	2.505	0.012275	*
scwhorupol.4	0.13903	0.06942	2.003	0.045251	*
scwhorufam.2	−0.25000	0.08673	−2.883	0.003955	**
scwhorufam.3	−0.60322	0.21651	−2.786	0.005346	**
scfalcdrnk.9	−0.61257	0.22973	−2.666	0.007680	**
sclfsat1.2	−0.40195	0.16282	−2.469	0.013583	*
sclfsat7.3	−0.20942	0.09145	−2.290	0.022053	*
sclfsat7.4	−0.22401	0.10502	−2.133	0.032958	*
sclfsato.2	1.29510	0.19541	6.628	3.62e−11	***
sclfsato.3	1.85872	0.14484	12.833	< 2e−16	***
sclfsato.4	0.94061	0.13940	6.748	1.60e−11	***
sclfsato.5	0.39529	0.09421	4.196	2.75e−05	***
sclfsato.7	−0.40762	0.11320	−3.601	0.000319	***
schmcont.2	0.20227	0.07468	2.709	0.006772	**
schmcont.3	0.67935	0.12153	5.590	2.34e−08	***
schmcont.4	0.69004	0.18535	3.723	0.000198	***
schmcont.5	1.04185	0.26552	3.924	8.79e−05	***
schmcont.6	0.81677	0.37170	2.197	0.028019	*
scloutcont.4	−0.32950	0.10200	−3.230	0.001241	**
scloutcont.5	−0.33502	0.10050	−3.333	0.000862	***
scloutcont.6	−0.40781	0.14804	−2.755	0.005886	**
scdem2many.2	−0.46994	0.15324	−3.067	0.002171	**
scdem2many.3	−0.82721	0.14805	−5.587	2.38e−08	***
scdem2many.4	−1.03726	0.16221	−6.395	1.69e−10	***
scdem2many.5	−1.21287	0.16979	−7.143	9.87e−13	***
scdem2many.6	−1.41007	0.19744	−7.142	9.96e−13	***
sctimemnuf.2	−0.19289	0.07323	−2.634	0.008451	**
scwkvfast.5	0.22045	0.09843	2.240	0.025140	*
scwkvfast.6	0.25746	0.13018	1.978	0.047986	*
scrundstnd.4	0.61180	0.15308	3.997	6.48e−05	***
scropenup.2	0.27084	0.07484	3.619	0.000298	***

```
scropenup.3       0.35685     0.08745    4.081  4.53e-05 ***
scfundstnd.2      0.16433     0.07672    2.142  0.032235 *
scfundstnd.3      0.21797     0.09724    2.242  0.025012 *
scfcritic.2      -0.37314     0.14041   -2.658  0.007887 **
scssupr2r.7       1.34360     0.48668    2.761  0.005780 **
scopfama.4       -0.19871     0.07138   -2.784  0.005381 **
ienddathh         0.08607     0.03286    2.619  0.008824 **
ienddatmm         0.07899     0.03209    2.462  0.013854 *
ff_jbstat.6      -0.50005     0.17786   -2.812  0.004942 **
ff_bentype10.1   -0.53190     0.24100   -2.207  0.027337 *
ff_bentype16.1   -0.66041     0.26649   -2.478  0.013225 *
ff_bentype35.1    2.25137     1.05237    2.139  0.032437 *
buno_dv.2        -0.56918     0.21253   -2.678  0.007417 **
buno_dv.3        -0.55024     0.19591   -2.809  0.004986 **
buno_dv.4        -0.94661     0.43358   -2.183  0.029044 *
fnspno.3         -2.08140     0.69728   -2.985  0.002844 **
fnspno.4         -2.31721     0.86789   -2.670  0.007601 **
sex_dv.2          0.52082     0.07623    6.832  8.92e-12 ***
intdatm_dv.2      0.34666     0.11542    3.003  0.002677 **
intdatm_dv.12     0.31980     0.14012    2.282  0.022497 *
hhtype_dv.5      -0.71167     0.22566   -3.154  0.001618 **
racel_dv         -0.08736     0.03208   -2.723  0.006477 **
ndepchl_dv.4     -0.78335     0.31712   -2.470  0.013522 *
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1

Residual standard error: 2.96 on 8535 degrees of freedom
Multiple R-squared:  0.6204,    Adjusted R-squared:  0.6156
F-statistic: 129.1 on 108 and 8535 DF,  p-value: < 2.2e-16

AIC: 43400.7249
MSE: 8.6502
[1] "The MSE of the predicted values are of 9.6928"
[1] "The Linear Model predicts exactly with accuracy of 0.1607"
[1] "The Linear Model predicts within a confidence interval with accuracy of 0.3452"
[1] "-----------Elastic Net Regression-----------"
702 x 1 sparse Matrix of class "dgCMatrix", with 50 entries
          names Estimate_Coefs
1   (Intercept)  11.4156358308
2         sf1.5   0.5267910771
3      health.2  -0.0398983292
4      finnow.3   0.0340696634
5      finnow.4   0.3068957260
6      finnow.5   0.8331880973
7      finfut.2   0.0848981857
8     arts2b9.1   0.0549366221
9    arts2b96.1  -0.0295924644
10       event1  -0.0094198832
11      scsf1.5   0.5199628899
12     scsf2a.3  -0.4879337032
13     scsf2b.3  -0.0246434269
14     scsf3b.5  -0.1681658585
15     scsf4a.2   0.4281282914
16      scsf5.2   0.0068550932
17      scsf5.4   0.1738438726
18      scsf5.5   0.1438430320
19     scsf6a.4   0.3523615123
20     scsf6a.5   0.3198647792
21     scsf6c.2   1.9096431223
22     scsf6c.3   0.4719735246
23     scsf6c.5  -0.8494586880
24      scsf7.2   1.1554873952
25      scsf7.5  -0.0398290126
26    sclfsat2.2   0.1368996958
27    sclfsato.2   0.1208265403
28    sclfsato.3   1.2095349616
29    sclfsato.4   0.1894372332
30    sclfsato.6  -0.4380431816
```

```
31    sclfsato.7  -0.6607287177
32    schmcont.3   0.1402838224
33    schmcont.5   0.0025333938
34 scloutcont.4  -0.0028789927
35 scloutcont.5  -0.0720542713
36 scloutcont.6  -0.1097880618
37 scdem2many.2   0.1386173041
38 scdem2many.5  -0.1121357756
39 scdem2many.6  -0.2584425647
40 sctimemnuf.2  -0.0381684474
41 sctimemnuf.6   0.0832948538
42 scrundstnd.4   0.1186441786
43   scropenup.3   0.0016103102
44    sf12mcs_dv  -2.4933101991
45      sex_dv.2   0.1448335206
46 marstat_dv.6  -0.2319720974
47 agegr10_dv.3  -0.0937115834
48  livesp_dv.1   0.0002540542
49      racel_dv  -0.0064602658
50    sf12pcs_dv  -0.3258411743
[1] "The MSE of the predicted values of the best fit model is 8.7763"
[1] "The Alpha of the best fit model is 0.8"
[1] "The Elastic Net Model predicts exactly with accuracy of 0.1624"
[1] "------------Timer Results------------"
   user  system elapsed
2417.25   13.58 2432.21
```

10.2.17 w2Merge console

```
[1] "------------Initial Checks------------"
[1] "34951467 NA cells were found across the entire dataset (62.36% of data as NA)"
[1] "------------Data Type Checks------------"
[1] "0 variables recoded since all their entries aren't numeric or NA"
[1] "NOTE: algorithim recodes categorical data alphabetically e.g (female = 1, male = 2)"
character(0)
[1] "------------Low Data Removal------------"
[1] "1163 variables removed since they had >= 'naPercent' (default 20%) NA values"
   [1] "pid"          "lvwhy"        "lvmthp"       "lvyrp"        "mvever"
"mvmnth"          "mvyr"         "mlstatchk"
   [9] "mlstat"       "drive"        "caruse"       "ukborn"       "plbornc"
"yr2uk4"          "citzn1"       "citzn2"
  [17] "citzn3"       "qfhigh"       "qualoc"       "qfvoc1"       "qfvoc2"
"qfvoc3"          "qfvoc4"       "qfvoc5"
  [25] "qfvoc6"       "qfvoc7"       "qfvoc8"       "qfvoc9"       "qfvoc10"
"qfvoc11"         "qfvoc12"      "qfvoc13"
  [33] "qfvoc14"      "qfvoc15"      "qfvoc96"      "school"       "scend"
"schlloc"         "schok"        "fenow"
  [41] "feend"        "j1none"       "j1semp"       "j1boss"       "j1mngr"
"paedqf"          "maedqf"       "edtype"
  [49] "edasp"        "fedlik"       "fednt"        "ocimpa"       "ocimpb"
"ocimpe"          "ocimpf"       "ocimpi"
  [57] "ocimpk"       "ocimpl"       "futra"        "futrb"        "futrc"
"futrd"           "futre"        "futrf"
  [65] "futrg"        "futrh"        "futri"        "futrj"        "futrk"
"futrl"           "paju"         "maju"
  [73] "pacob"        "payruk"       "macob"        "mayruk"       "natid1"
"natid2"          "natid3"       "natid4"
  [81] "natid5"       "natid6"       "natid97"      "racel"        "racelo_code"
"ethid2"          "ethid3"       "ethid4a"
  [89] "ethid4b"      "ethid5"       "ethid6"       "ethid7"       "ethid8"
"ethid9"          "ethid10"      "ethid11"
  [97] "ethid12"      "ethid13"      "ethid14"      "ethclose1"    "ethclose2a"
"ethclose2b"      "ethclose3"    "ethclose4a"
 [105] "ethclose4b"   "ethclose5"    "ethclose6"    "ethclose7"    "ethclose8"
"ethclose9"       "ethclose10"   "ethclose11"
```

```
 [113] "pride2"          "pride4a"         "pride4b"         "pride6"          "pride7"
"pride5"          "pride8"          "pride9"
 [121] "pride10"         "pride11"         "pride12"         "pride13"         "pride14"
"food1"           "food2"           "food3"
 [129] "food4"           "food5"           "food6"           "food7"           "oprlg"
"oprlg0ni"        "nirel"           "niact"
 [137] "oprlg0"          "oprlg1"          "smnow"           "ncigs"           "smcigs"
"smncigs"         "aglquit"         "smagbg"
 [145] "hospc1"          "hospdc1"         "hospc2"          "hospdc2"         "hospc3"
"hospdc3"         "hospc4"          "hospdc4"
 [153] "hospc5"          "hospdc5"         "hospc6"          "hospdc6"         "hospc7"
"hospdc7"         "hospc8"          "hospdc8"
 [161] "disdif1"         "disdif2"         "disdif3"         "disdif4"         "disdif5"
"disdif6"         "disdif7"         "disdif8"
 [169] "disdif9"         "disdif10"        "disdif11"        "disdif12"        "disdif96"
"aidhua1"         "aidhua2"         "aidhua3"
 [177] "aidhua4"         "aidhua5"         "aidhua6"         "aidhua7"         "aidhua8"
"aidhua9"         "aidhua10"        "aidhua11"
 [185] "aidhua12"        "aidhua13"        "aidhua14"        "aidhua15"        "aidhua16"
"naidxhh"         "aidhu1"          "aidhu2"
 [193] "aidhrs"          "aideft"          "lcohnpi"         "coh1bm"          "coh1by"
"coh1mr"          "coh1em"          "coh1ey"
 [201] "nmar"            "lmar1m"          "lmar1y"          "ladopt"          "lnadopt"
"lprnt"           "lnprnt"          "ch1by4"
 [209] "movy11"          "movy12"          "movy13"          "movy14"          "movy15"
"movy16"          "family"          "education"
 [217] "memploy"         "housing"         "area"            "moveoth_code" "movdir"
"plnowm"          "plnowy4"         "mstatsam"
 [225] "lwwrong"         "mstatsamn"       "lastmstatch1" "mstatch1"        "statcm1"
"statcy41"        "divchk1"         "divfin1"
 [233] "dvm1"            "dvy41"           "cmlstat1"        "lastmstatch2" "mstatch2"
"statcm2"         "statcy42"        "divchk2"
 [241] "divfin2"         "dvm2"            "dvy42"           "cmlstat2"        "lastmstatch3" "mstatch3"
"statcm3"         "statcy43"
 [249] "divchk3"         "divfin3"         "dvm3"            "dvy43"           "cmlstat3"
"cohab"           "cohabn"          "lmcbm1"
 [257] "lmcby41"         "currpart1"       "lmspm1"          "lmspy41"         "lmcbm2"
"lmcby42"         "currpart2"       "lmspm2"
 [265] "lmspy42"         "lmcbm3"          "lmcby43"         "currpart3"       "lmspm3"
"lmspy43"         "lmcbm4"          "lmcby44"
 [273] "currpart4"       "lmspm4"          "lmspy44"         "father"          "nchild"
"preg"            "pregm1"          "pregy41"
 [281] "pregfert1"       "invitro1"        "pregout1"        "pregend1"        "endmnth1"
"pregsmoke1"      "smkmnth11"       "smkmnth21"
 [289] "smkmnth31"       "pregsmk11"       "pregsmk21"       "pregsmk31"       "pregdrink1"
"lchmulti1"       "pregm2"          "pregy42"
 [297] "pregfert2"       "invitro2"        "pregout2"        "pregend2"        "endmnth2"
"pregsmoke2"      "smkmnth12"       "smkmnth22"
 [305] "smkmnth32"       "pregsmk12"       "pregsmk22"       "pregsmk32"       "pregdrink2"
"lchmulti2"       "pregm3"          "pregy43"
 [313] "pregfert3"       "invitro3"        "pregout3"        "pregend3"        "endmnth3"
"pregsmoke3"      "smkmnth13"       "smkmnth23"
 [321] "smkmnth33"       "pregsmk13"       "pregsmk23"       "pregsmk33"       "pregdrink3"
"lchmulti3"       "pregm4"          "pregy44"
 [329] "pregfert4"       "invitro4"        "pregout4"        "pregend4"        "endmnth4"
"pregsmoke4"      "smkmnth14"       "smkmnth24"
 [337] "smkmnth34"       "pregsmk14"       "pregsmk24"       "pregsmk34"       "pregdrink4"
"lchmulti4"       "pregm5"          "pregy45"
 [345] "pregfert5"       "invitro5"        "pregout5"        "pregend5"        "endmnth5"
"pregsmoke5"      "smkmnth15"       "smkmnth25"
 [353] "smkmnth35"       "pregsmk15"       "pregsmk25"       "pregsmk35"       "pregdrink5"
"lchmulti5"       "hcondno1"        "hcondns1"
 [361] "hcondno2"        "hcondns2"        "hcondno3"        "hcondns3"        "hcondno4"
"hcondns4"        "hcondno5"        "hcondns5"
 [369] "hcondno6"        "hcondns6"        "hcondno7"        "hcondns7"        "hcondno8"
"hcondns8"        "contft"          "ftendm"
 [377] "ftendy4"         "ftquals"         "ftedstartm1"     "ftedstarty41" "ftedend1"
"ft2endm1"        "ft2endy41"       "ftedmor1"
```

```
 [385] "ftedstartm2"  "ftedstarty42" "ftedend2"     "ft2endm2"     "ft2endy42"
"ftedmor2"      "qualnew1"     "qualnew2"
 [393] "qualnew3"     "qualnew4"     "qualnew5"     "qualnew6"     "qualnew7"
"qualnew8"      "qualnew9"     "qualnew10"
 [401] "qualnew11"    "qualnew13"    "qualnew15"    "qualnew16"    "qualnew17"
"qualnew18"     "qualnew19"    "qualnew20"
 [409] "qualnew21"    "qualnew22"    "qualnew23"    "qualnew24"    "qualnew25"
"qualnew26"     "qualnew27"    "qualnew28"
 [417] "qualnew29"    "qualnew30"    "qualnew31"    "trwho1"       "traindays1"
"trainhrs1"     "trainend1"    "trainpurp11"
 [425] "trainpurp21"  "trainpurp31"  "trainpurp41"  "trainpurp51"  "trainpurp61"
"trainpurp71"   "trainqual1"   "trwho2"
 [433] "traindays2"   "trainhrs2"    "trainend2"    "trainpurp12"  "trainpurp22"
"trainpurp32"   "trainpurp42"  "trainpurp52"
 [441] "trainpurp62"  "trainpurp72"  "trainqual2"   "trwho3"       "traindays3"
"trainhrs3"     "trainend3"    "trainpurp13"
 [449] "trainpurp23"  "trainpurp33"  "trainpurp43"  "trainpurp53"  "trainpurp63"
"trainpurp73"   "trainqual3"   "trainn"
 [457] "trqual1"      "trqual2"      "trqual3"      "trqual4"      "trqual5"
"trqual6"       "trqual7"      "trqual8"
 [465] "trqual9"      "trqual10"     "trqual11"     "trqual13"     "trqual15"
"trqual16"      "trqual17"     "trqual18"
 [473] "trqual19"     "trqual20"     "trqual21"     "trqual22"     "trqual23"
"trqual24"      "trqual25"     "trqual26"
 [481] "trqual27"     "trqual28"     "trqual29"     "trqual30"     "trqual31"
"notempchk"     "empchk"       "empstendd"
 [489] "empstendm"    "empstendy4"   "stendreas"    "nxtst"        "nxtstelse"
"cstat"         "nxtstendd"    "nxtstendm"
 [497] "nxtstendy4"   "jbsamr"       "wkplsam"      "samejob"      "matlv"
"matlvstd"      "matlvstm"     "matlvsty4"
 [505] "matlvendd"    "matlvendm"    "matlvendy4"   "jbendd"       "jbendm"
"jbendy4"       "jbendreas"    "cjob"
 [513] "nxtjbhrs"     "nxtjbes"      "nxtjbendd"    "nxtjbendm"    "nxtjbendy4"
"cjbatt"        "nextstat1"    "nextelse1"
 [521] "currstat1"    "nextjob1"     "currjob1"     "jobhours1"    "reasend1"
"jbatt1"        "statendd1"    "statendm1"
 [529] "statendy41"   "nextstat2"    "nextelse2"    "currstat2"    "nextjob2"
"currjob2"      "jobhours2"    "reasend2"
 [537] "jbatt2"       "statendd2"    "statendm2"    "statendy42"   "nextstat3"
"nextelse3"     "currstat3"    "nextjob3"
 [545] "currjob3"     "jobhours3"    "reasend3"     "jbatt3"       "statendd3"
"statendm3"     "statendy43"   "nextstat4"
 [553] "nextelse4"    "currstat4"    "nextjob4"     "currjob4"     "jobhours4"
"reasend4"      "jbatt4"       "statendd4"
 [561] "statendm4"    "statendy44"   "nextstat5"    "nextelse5"    "currstat5"
"nextjob5"      "currjob5"     "jobhours5"
 [569] "reasend5"     "jbatt5"       "statendd5"    "statendm5"    "statendy45"
"nextstat6"     "nextelse6"    "currstat6"
 [577] "nextjob6"     "currjob6"     "jobhours6"    "reasend6"     "jbatt6"
"statendd6"     "statendm6"    "statendy46"
 [585] "nextstat7"    "nextelse7"    "currstat7"    "nextjob7"     "currjob7"
"jobhours7"     "reasend7"     "jbatt7"
 [593] "statendd7"    "statendm7"    "statendy47"   "nextstat8"    "nextelse8"
"currstat8"     "nextjob8"     "currjob8"
 [601] "jobhours8"    "reasend8"     "jbatt8"       "statendd8"    "statendm8"
"statendy48"    "nextstat9"    "nextelse9"
 [609] "currstat9"    "nextjob9"     "currjob9"     "jobhours9"    "reasend9"
"jbatt9"        "statendd9"    "statendm9"
 [617] "statendy49"   "nextstat10"   "nextelse10"   "currstat10"   "nextjob10"
"currjob10"     "jobhours10"   "reasend10"
 [625] "jbatt10"      "statendd10"   "statendm10"   "statendy410"  "jboff"
"jboffy"        "jbterm1"      "jbterm2"
 [633] "jbsic07chk"   "jbsoc00chk"   "jbsempchk"    "jbsemp"       "jbbgd"
"jbbgm"         "jbbgy"        "jbmngrchk"
 [641] "jbmngr"       "jbsizechk"    "jbsize"       "jbsect"       "jbsectpub"
"jbhrs"         "jbot"         "jbotpd"
 [649] "paygwc"       "paynwc"       "payusl"       "payu"         "payuwc"
"payug"         "paytyp"       "ovtpay"
```

```
 [657] "extnsa"      "extrate"     "extrest"     "basnsa"      "basrate"
"basrest"     "ovtnsa"      "ovtrate"
 [665] "ovtrest"     "jbpl"        "jbttwt"      "worktrav"    "jsboss"
"jssize"      "jshrs"       "jstypeb"
 [673] "jsaccs"      "jspart"      "jsprbm"      "jsprby4"     "jsprem"
"jsprey4"     "jsprls"      "jsprtx"
 [681] "jsprni"      "jspayu"      "jspayw"      "jspytx"      "jspyni"
"jspl"        "jsttwt"      "jsttwtb"
 [689] "jsworktrav"  "workdis"     "twkdiff1"    "twkdiff2"    "twkdiff3"
"twkdiff4"    "twkdiff5"    "twkdiff6"
 [697] "twkdiff7"    "twkdiff8"    "twkdiff97"   "twkdiffm"    "twkcar"
"twkcary1"    "twkcary2"    "twkcary3"
 [705] "twkcary4"    "twkcary5"    "twkcary6"    "twkcary7"    "twkcary8"
"twkcary9"    "twkcary10"   "twkcary11"
 [713] "twkcary12"   "twkcary13"   "twkcary14"   "twkcary97"   "twkcarym"
"altcar1"     "altcar2"     "altcar3"
 [721] "altcar4"     "altcar5"     "altcar6"     "altcar7"     "altcar8"
"altcar9"     "altcar10"    "altcar11"
 [729] "altcar12"    "altcar96"    "altcar97"    "carclub"     "carshare"
"wkhome"      "lifthh"      "liftxhh"
 [737] "motcyc"      "comtaxi"     "combus"      "comtrain"    "commetro"
"combike"     "comwalk"     "comother"
 [745] "jbsat"       "wkphys"      "jbperfp"     "jbonus"      "jbrise"
"tujbpl"      "tuin1"       "jbpen"
 [753] "jbpenm"      "jbpeny4"     "penmcn"      "penmpy"      "penmtp"
"penspb"      "wktime"      "wkends"
 [761] "jbflex1"     "jbflex2"     "jbflex3"     "jbflex4"     "jbflex5"
"jbflex6"     "jbflex7"     "jbflex8"
 [769] "jbflex96"    "jbfxuse1"    "jbfxuse2"    "jbfxuse3"    "jbfxuse4"
"jbfxuse5"    "jbfxuse6"    "jbfxuse7"
 [777] "jbfxuse8"    "jbfxuse96"   "jbfxinf"     "wkaut1"      "wkaut2"
"wkaut3"      "wkaut4"      "wkaut5"
 [785] "depenth1"    "depenth2"    "depenth3"    "depenth4"    "depenth5"
"depenth6"    "jblkcha"     "jbxpcha"
 [793] "jblkchb"     "jbxpchb"     "jblkchc"     "jbxpchc"     "jblkchd"
"jbxpchd"     "jblkche"     "jbxpche"
 [801] "jbsec"       "julk4wk"     "julkjb"      "jubgn"       "julk4x1"
"julk4x2"     "julk4x3"     "julk4x4"
 [809] "julk4x5"     "julk4x6"     "julk4x96"    "jbhad"       "jlendm"
"jlendy"      "jlsemp"      "jlboss"
 [817] "jlmngr"      "retchk"      "jlsize"      "eprosh"      "j2semp"
"j2hrs"
"j2pay"       "retchk"      "ageret"
 [825] "rtpro1"      "rtpro2"      "rtpro3"      "rtpro4"      "rtpro5"
"rtpro6"      "rtcon1"      "rtcon2"
 [833] "rtcon3"      "rtcon4"      "penmex"      "pppex"       "pppexm"
"sppen"       "rtexpjb"     "rtfnd1"
 [841] "rtfnd2"      "rtfnd3"      "rtfnd4"      "rtfnd5"      "rtfnd6"
"rtfnd7"      "rtfnd8"      "rtfnd9"
 [849] "rtfnd10"     "rtfnd96"     "retamt"      "retsuf"      "volfreq"
"volhrs"      "charfreq"    "charam"
 [857] "ccare"       "ccwork"      "benunemp1"   "benunemp2"   "benunemp96"
"bendis1"     "bendis2"     "bendis3"
 [865] "bendis4"     "bendis5"     "bendis6"     "bendis7"     "bendis8"
"bendis9"     "bendis10"    "bendis11"
 [873] "bendis96"    "benpen1"     "benpen2"     "benpen3"     "benpen4"
"benpen5"     "benpen6"     "benpen7"
 [881] "benpen8"     "benpen96"    "niserps"     "bencb"       "benctc"
"benfam1"     "benfam2"     "benfam3"
 [889] "benfam4"     "benfam5"     "benfam96"    "bentax1"     "bentax2"
"bentax3"     "bentax4"     "bentax5"
 [897] "bentax96"    "benhou1"     "benhou2"     "benhou3"     "benhou4"
"benhou96"    "nfh01"       "nfh02"
 [905] "nfh03"       "nfh04"       "nfh05"       "nfh06"       "nfh07"
"nfh08"       "nfh09"       "nfh10"
 [913] "nfh11"       "nfh12"       "nfh13"       "nfh14"       "nfh15"
"nfh16"       "nfh17"       "nfh18"
 [921] "nfh19"       "nfh20"       "nfh21"       "nfh22"       "nfh23"
"nfh24"       "nfh25"       "nfh26"
```

```
 [929] "nfh27"        "nfh28"        "nfh29"        "nfh30"        "nfh31"
"nfh32"        "nfh33"        "nfh34"
 [937] "nfh35"        "nfh36"        "nfh37"        "nfh38"        "fiyrdb1"
"fiyrdb2"      "fiyrdb3"      "fiyrdb4"
 [945] "fiyrdb5"      "fiyrdb6"      "ppent"        "ppyrs"        "ppreg"
"ppram"        "pprampc"      "saved"
 [953] "savreg"       "savlt"        "hubuys"       "hufrys"       "humops"
"huiron"       "hupots"       "hudiy"
 [961] "husits"       "huboss"       "vote2"        "vote3"        "vote4"
"vote5"        "perpolinf"    "colbens1"
 [969] "colbens2"     "colbens3"     "civicduty"    "polcost"      "votenorm"
"perbfts"      "grpbfts"      "voteintent"
 [977] "demorient"    "vote7"        "vote8"        "arts1freq"    "arts2freq"
"libfreq"      "arcfreq"      "musfreq"
 [985] "herfreq"      "sportsfreq"   "sports3freq"  "club"         "acc1to161"
"acc1to162"    "acc1to163"    "acc1to164"
 [993] "acc1to165"    "acc1to166"    "acc1to167"    "acc1to168"    "acc1to169"
"acc1to1610"   "acc1to1611"   "acc1to1612"
 [ reached getOption("max.print") -- omitted 163 entries ]
[1] "------------Low Level Removal------------"
[1] "If a level is removed from a variable you wish to keep, reccomended to manually merge levels to
[1] "level 6 in pno removed, 2 observations found"
[1] "level 7 in pno removed, 1 observations found"
[1] "level 7 in natch01 removed, 3 observations found"
[1] "level 8 in natch01 removed, 0 observations found"
[1] "level 2 in natch02 removed, 2 observations found"
[1] "level 7 in natch02 removed, 3 observations found"
[1] "level 8 in natch02 removed, 0 observations found"
[1] "level 9 in natch02 removed, 0 observations found"
[1] "level 3 in natch03 removed, 0 observations found"
[1] "level 7 in natch03 removed, 2 observations found"
[1] "level 9 in natch03 removed, 0 observations found"
[1] "level 10 in natch03 removed, 0 observations found"
[1] "level 7 in natch06 removed, 1 observations found"
[1] "level 8 in natch06 removed, 1 observations found"
[1] "level 9 in natch06 removed, 1 observations found"
[1] "level 9 in natch07 removed, 0 observations found"
[1] "level 6 in nnatch removed, 0 observations found"
[1] "level 7 in nnatch removed, 0 observations found"
[1] "level 4 in nadoptch removed, 1 observations found"
[1] "level 1 in adoptch01 removed, 4 observations found"
[1] "level 6 in adoptch02 removed, 2 observations found"
[1] "level 4 in adoptch03 removed, 0 observations found"
[1] "level 6 in adoptch03 removed, 0 observations found"
[1] "level 7 in adoptch03 removed, 0 observations found"
[1] "level 5 in adoptch04 removed, 0 observations found"
[1] "level 6 in adoptch04 removed, 0 observations found"
[1] "level 7 in adoptch04 removed, 0 observations found"
[1] "level 6 in nchunder16 removed, 0 observations found"
[1] "level 5 in nch5to15 removed, 0 observations found"
[1] "level 6 in nch5to15 removed, 0 observations found"
[1] "level 4 in nch10to15 removed, 4 observations found"
[1] "level 1 in allch01 removed, 3 observations found"
[1] "level 7 in allch01 removed, 0 observations found"
[1] "level 8 in allch01 removed, 0 observations found"
[1] "level 7 in allch02 removed, 2 observations found"
[1] "level 8 in allch02 removed, 1 observations found"
[1] "level 9 in allch02 removed, 0 observations found"
[1] "level 9 in allch03 removed, 0 observations found"
[1] "level 10 in allch03 removed, 0 observations found"
[1] "level 7 in allch04 removed, 1 observations found"
[1] "level 8 in allch04 removed, 2 observations found"
[1] "level 9 in allch04 removed, 0 observations found"
[1] "level 8 in allch05 removed, 0 observations found"
[1] "level 9 in allch06 removed, 0 observations found"
[1] "level 9 in jbstat removed, 4 observations found"
[1] "level 10 in jbstat removed, 4 observations found"
[1] "level 11 in fruvege removed, 2 observations found"
```

```
[1] "level 12 in fruvege removed, 2 observations found"
[1] "level 13 in fruvege removed, 1 observations found"
[1] "level 15 in fruvege removed, 1 observations found"
[1] "level 20 in fruvege removed, 1 observations found"
[1] "level 5 in relup removed, 3 observations found"
[1] "level 1 in bensta3 removed, 2 observations found"
[1] "level 7 in marstat removed, 2 observations found"
[1] "level 4 in ivcoop removed, 2 observations found"
[1] "level 5 in undqus removed, 1 observations found"
[1] "level 7 in hgbiom removed, 2 observations found"
[1] "level 4 in hgbiof removed, 3 observations found"
[1] "level 5 in hgbiof removed, 2 observations found"
[1] "level 6 in hgbiof removed, 0 observations found"
[1] "level 5 in scssupr2r removed, 2 observations found"
[1] "level 6 in pn1pno removed, 0 observations found"
[1] "level 7 in pn1pno removed, 0 observations found"
[1] "level 4 in pn2pno removed, 0 observations found"
[1] "level 5 in pn2pno removed, 0 observations found"
[1] "level 7 in pn2pno removed, 0 observations found"
[1] "level 6 in pns1pno removed, 0 observations found"
[1] "level 7 in pns1pno removed, 0 observations found"
[1] "level 4 in pns2pno removed, 1 observations found"
[1] "level 5 in pns2pno removed, 0 observations found"
[1] "level 7 in pns2pno removed, 0 observations found"
[1] "level 1 in fiyrinvinc_tc removed, 2 observations found"
[1] "level 10 in ff_jbstat removed, 1 observations found"
[1] "level 1 in ff_bentype25 removed, 4 observations found"
[1] "level 1 in ff_bentype32 removed, 2 observations found"
[1] "level 1 in ff_bentype36 removed, 2 observations found"
[1] "level 1 in ngrp_dv removed, 4 observations found"
[1] "level 4 in nnssib_dv removed, 2 observations found"
[1] "level 5 in nnssib_dv removed, 1 observations found"
[1] "level 7 in mastat_dv removed, 0 observations found"
[1] "level 5 in buno_dv removed, 2 observations found"
[1] "level 6 in nchild_dv removed, 1 observations found"
[1] "level 6 in hrpno removed, 0 observations found"
[1] "level 6 in ppno removed, 1 observations found"
[1] "level 5 in sppno removed, 4 observations found"
[1] "level 4 in fnpno removed, 0 observations found"
[1] "level 5 in fnpno removed, 0 observations found"
[1] "level 6 in fnpno removed, 0 observations found"
[1] "level 4 in fnspno removed, 0 observations found"
[1] "level 5 in fnspno removed, 0 observations found"
[1] "level 6 in fnspno removed, 0 observations found"
[1] "level 7 in mnpno removed, 0 observations found"
[1] "level 7 in mnspno removed, 0 observations found"
[1] "level 2 in grfpno removed, 0 observations found"
[1] "level 1 in grmpno removed, 0 observations found"
[1] "level 2 in grmpno removed, 0 observations found"
[1] "level 4 in grmpno removed, 0 observations found"
[1] "level 3 in nunmpsp_dv removed, 1 observations found"
[1] "level 4 in nunmpsp_dv removed, 1 observations found"
[1] "level 2 in ficode3 removed, 3 observations found"
[1] "level 3 in ficode3 removed, 1 observations found"
[1] "level 1 in ficode21 removed, 4 observations found"
[1] "level 1 in ficode25 removed, 0 observations found"
[1] "level 3 in ficode26 removed, 1 observations found"
[1] "level 2 in ficode27 removed, 3 observations found"
[1] "level 2 in ficode28 removed, 4 observations found"
[1] "level 1 in ficode35 removed, 2 observations found"
[1] "level 2 in ficode38 removed, 3 observations found"
[1] "level 2 in ficode39 removed, 3 observations found"
[1] "level 4 in nnsib_dv removed, 0 observations found"
[1] "level 6 in ndepchl_dv removed, 0 observations found"
[1] "level 21 in hhtype_dv removed, 0 observations found"
[1] "level 3 in nmpsp_dv removed, 4 observations found"
[1] "level 4 in nmpsp_dv removed, 1 observations found"
[1] "level 5 in nmpsp_dv removed, 1 observations found"
```

```
[1] "level 6 in nmpsp_dv removed, 1 observations found"
[1] "116 total levels removed from 68 different variables. In total 130 observations deleted"
[1] "——————Variance 0 Check——————"
[1] "69 variables removed since their new variance was 0"
 [1] "ivfio"         "ioutcome"      "newper"        "newentrant"    "adstatus"
"natch06"       "natch07"
 [8] "natch08"       "natch09"       "natch10"       "natch11"       "natch12"
"natch13"       "natch14"
[15] "natch15"       "natch16"       "adoptch04"     "adoptch05"     "adoptch06"
"adoptch07"     "adoptch08"
[22] "adoptch09"     "adoptch10"     "adoptch11"     "adoptch12"     "adoptch13"
"adoptch14"     "adoptch15"
[29] "adoptch16"     "allch06"       "allch07"       "allch08"       "allch09"
"allch10"       "allch11"
[36] "allch12"       "allch13"       "allch14"       "allch15"       "allch16"
"chkdob"        "wlk10m"
[43] "bensta3"       "indmode"       "sceverdrnk"    "screlany"      "scfrendany"
"intdatd_if"    "intdatm_if"
[50] "intdaty_if"    "doby_if"       "age_if"        "fiyrinvinc_tc" "ff_ivlolw"
"ff_everint"    "ff_bentype25"
[57] "ff_bentype32"  "ff_bentype36"  "ff_bentype37"  "ngrp_dv"       "grfpno"
"grmpno"        "fiyrinvinc_if"
[64] "wave"          "ficode21"      "ficode25"      "ficode35"      "ficode36"
"scflag_dv"
[1] "——————Dummy Variables——————"
[1] "predictor variable count went from 447 to 981"
[1] "——————Variance 0 Check——————"
[1] "95 variables removed since their new variance was 0"
 [1] "pno.6"         "pno.7"         "natch01.7"     "natch01.8"     "natch02.2"
"natch02.7"     "natch02.8"     "natch02.9"
 [9] "natch03.3"     "natch03.7"     "natch03.9"     "natch03.10"    "nnatch.6"
"nnatch.7"      "nadoptch.4"    "adoptch01.1"
[17] "adoptch02.6"   "adoptch03.4"   "adoptch03.6"   "adoptch03.7"   "nchunder16.6" "nch5to15.5"
"nch5to15.6"    "nch10to15.4"
[25] "allch01.1"     "allch01.7"     "allch01.8"     "allch02.7"     "allch02.8"
"allch02.9"     "allch03.9"     "allch03.10"
[33] "allch04.7"     "allch04.8"     "allch04.9"     "allch05.8"     "jbstat.9"
"jbstat.10"     "fruvege.11"    "fruvege.12"
[41] "fruvege.13"    "fruvege.15"    "fruvege.20"    "relup.5"       "marstat.7"
"ivcoop.4"      "undqus.5"      "hgbiom.7"
[49] "hgbiof.4"      "hgbiof.5"      "hgbiof.6"      "scssupr2r.5"   "pn1pno.6"
"pn1pno.7"      "pn2pno.4"      "pn2pno.5"
[57] "pn2pno.7"      "pns1pno.6"     "pns1pno.7"     "pns2pno.4"     "pns2pno.5"
"pns2pno.7"     "ff_jbstat.10"  "nnssib_dv.4"
[65] "nnssib_dv.5"   "mastat_dv.7"   "buno_dv.5"     "nchild_dv.6"   "hrpno.6"
"ppno.6"        "sppno.5"       "fnpno.4"
[73] "fnpno.5"       "fnpno.6"       "fnspno.4"      "fnspno.5"      "fnspno.6"
"mnpno.7"       "mnspno.7"      "nunmpsp_dv.3"
[81] "nunmpsp_dv.4"  "ficode3.2"     "ficode3.3"     "ficode26.3"    "ficode27.2"
"ficode28.2"    "ficode38.2"    "ficode39.2"
[89] "nnsib_dv.4"    "ndepchl_dv.6"  "hhtype_dv.21"  "nmpsp_dv.3"    "nmpsp_dv.4"
"nmpsp_dv.5"    "nmpsp_dv.6"
[1] "——————K—Means——————"
[1] "15 clusters have been made for K—Means"
[1] "K—Means results as a table, the max value in each row is a simple way to define which cluster r

       0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16
17  18  19  20  21  22  23  24  25  26  27  28  29
  1    1   2   0   2   5  11  55  41  46  44  33  34  39  16  13   9  14
10   5   3   6   7   3   3   3   4   2   3   4   1
  2    2   0   2   2   8  20  77 105  79  83  60  72  83  39  28  26  16
12  10   9  11  11  14   3   2   4   4   1   0   0
  3    1   1   0   0   1   3  18   9   9  18  14  17  18   9   3   1   5
0   2   2   5   2   0   1   0   1   0   0   1   0
  4    3   0   2   1   0   4  18  24  31  21  24  21  24  13   4  13   4
5   4   3   1   1   2   3   1   1   0   0   2   2
  5    0   0   2   3   6  13  73  66  83  58  61  66  64  25  15  19  13
8  16   6   6   4   8   4   6   2   2   4   2   4
```

```
   6    0    1    3    3    4    6   47   50   43   39   49   51   45   21   13    8    9
7    6    6    3   10    5    2    2    1    1    2    1    0
   7    1    2    2    8   11   13   77   96   72   84   66   55   78   39   21   10   15
12   14   18   10    7    4    8    6    6    3    0    3    0
   8    0    0    0    0    6    5   50   51   55   34   45   39   51   23    7    7    8
5    6    5    5   10    7    2    1    0    0    2    2    0
   9    0    0    0    0    2    4   24   27   33   25   23   22   19   10    8    5    8
6    6    2    2    1    2    3    3    1    1    2    0    0
  10    1    6    2    1    6   16  100   64   83   81   70   68   95   31   30   19   15
12   10    9   11    5    9    4    4    4    0    3    2    2
  11    1    0    1    2    5   10   51   42   38   40   32   37   34   17    9   11    8
5    5    5    4    5    1    6    4    6    1    0    0    2
  12    0    0    4    3    4   14   95   98   74   82   57   63   83   30   21   16    9
11   17   14   10    6    4    6    6    4    4    4    2    1
  13    1    1    0    2    5   12   48   46   27   39   32   29   47   20   12   18    8
7   11    4    5    6    2    0    3    1    1    0    1    1
  14    1    0    1    3    5   13   58   60   47   64   35   41   32   22   15   14   11
8    9    7    4    2    5    0    6    2    1    1    0    0
  15    1    1    1    0    3   12   42   52   46   54   36   45   35   18   13   11    6
5    9    7    6    1    3    2    2    5    1    1    1    0

      30   31   32   33   34   35   36
  1    1    0    1    1    0    0    0
  2    2    0    1    0    0    0    1
  3    1    0    1    0    0    0    0
  4    0    0    0    1    1    0    0
  5    1    0    0    1    0    1    1
  6    0    0    0    1    0    0    0
  7    0    2    0    0    0    2    0
  8    0    0    1    0    1    0    1
  9    1    0    0    1    0    0    0
  10   1    0    0    1    0    0    1
  11   0    0    0    2    0    0    1
  12   0    1    0    0    0    0    0
  13   1    0    0    0    0    0    0
  14   0    0    1    1    0    0    0
  15   0    0    0    0    0    0    0
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
 [1] "Cluster 1: Within MSE 804551211778054, Size 422"   "Cluster 2: Within MSE 3693587232729181, Siz
 [3] "Cluster 3: Within MSE 155463966123151, Size 143"   "Cluster 4: Within MSE 5075745083535, Size 2
 [5] "Cluster 5: Within MSE 9324060776371804, Size 643"  "Cluster 6: Within MSE 3452050729842465, Siz
 [7] "Cluster 7: Within MSE 3398382141983318, Size 745"  "Cluster 8: Within MSE 3465997619251857, Siz
 [9] "Cluster 9: Within MSE 6487899342394, Size 241"     "Cluster 10: Within MSE 3924211248383204, S
[11] "Cluster 11: Within MSE 317105354420029, Size 385"  "Cluster 12: Within MSE 3463884025430642, S
[13] "Cluster 13: Within MSE 1060357665265256, Size 390" "Cluster 14: Within MSE 3465050958987151, S
[15] "Cluster 15: Within MSE 3469377899275599, Size 419"
[1] "Total between cluster MSE: 614600548558713728, Total within cluster MSE: 3307360493513786"
[1] "The K-Means model predicts exactly with an accuracy of 0.1298"
[1] "--------------Correlation Checks--------------"
[1] "indpxus_xw removed, correlated with 5 other variable(s)"
[1] "pensioner_dv.2 removed, correlated with 5 other variable(s)"
[1] "dvage removed, correlated with 4 other variable(s)"
[1] "hgbiom.1 removed, correlated with 4 other variable(s)"
[1] "hgbiom.4 removed, correlated with 4 other variable(s)"
[1] "hgbiom.5 removed, correlated with 4 other variable(s)"
[1] "indinus_xw removed, correlated with 4 other variable(s)"
[1] "indpxus_lw removed, correlated with 5 other variable(s)"
[1] "nchresp.2 removed, correlated with 4 other variable(s)"
[1] "nchresp.3 removed, correlated with 4 other variable(s)"
[1] "nchunder16.3 removed, correlated with 4 other variable(s)"
[1] "pidp removed, correlated with 3 other variable(s)"
[1] "sex.2 removed, correlated with 3 other variable(s)"
[1] "birthy removed, correlated with 3 other variable(s)"
[1] "nchunder16.5 removed, correlated with 3 other variable(s)"
[1] "relup.2 removed, correlated with 3 other variable(s)"
[1] "pn1pno.1 removed, correlated with 3 other variable(s)"
[1] "pn1pno.4 removed, correlated with 3 other variable(s)"
[1] "pn1pno.5 removed, correlated with 3 other variable(s)"
```

```
[1] "pn2pno.2 removed, correlated with 3 other variable(s)"
[1] "ficode1.1 removed, correlated with 4 other variable(s)"
[1] "nchresp.1 removed, correlated with 3 other variable(s)"
[1] "employ.2 removed, correlated with 3 other variable(s)"
[1] "hidp removed, correlated with 2 other variable(s)"
[1] "pno.2 removed, correlated with 2 other variable(s)"
[1] "hhorig.7 removed, correlated with 2 other variable(s)"
[1] "nchunder16.2 removed, correlated with 2 other variable(s)"
[1] "nchunder16.4 removed, correlated with 2 other variable(s)"
[1] "allch05.7 removed, correlated with 2 other variable(s)"
[1] "chksex.2 removed, correlated with 2 other variable(s)"
[1] "relup.6 removed, correlated with 2 other variable(s)"
[1] "btype8.1 removed, correlated with 2 other variable(s)"
[1] "marstat.2 removed, correlated with 2 other variable(s)"
[1] "marstat.4 removed, correlated with 2 other variable(s)"
[1] "marstat.6 removed, correlated with 2 other variable(s)"
[1] "hgbiom.2 removed, correlated with 2 other variable(s)"
[1] "hgbiom.3 removed, correlated with 2 other variable(s)"
[1] "hgbiof.1 removed, correlated with 2 other variable(s)"
[1] "hgbiof.2 removed, correlated with 2 other variable(s)"
[1] "respm16.2 removed, correlated with 2 other variable(s)"
[1] "scdoby4 removed, correlated with 2 other variable(s)"
[1] "pns1pno.1 removed, correlated with 2 other variable(s)"
[1] "pns1pno.4 removed, correlated with 2 other variable(s)"
[1] "pns1pno.5 removed, correlated with 2 other variable(s)"
[1] "pns2pno.2 removed, correlated with 2 other variable(s)"
[1] "fimngrs_tc.1 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.6 removed, correlated with 2 other variable(s)"
[1] "cohab_dv.1 removed, correlated with 2 other variable(s)"
[1] "indpxbh_xw removed, correlated with 2 other variable(s)"
[1] "indscus_xw removed, correlated with 2 other variable(s)"
[1] "indinub_xw removed, correlated with 2 other variable(s)"
[1] "fimnlabgrs_dv removed, correlated with 2 other variable(s)"
[1] "ndepchl_dv.2 removed, correlated with 3 other variable(s)"
[1] "nchresp.4 removed, correlated with 2 other variable(s)"
[1] "ndepchl_dv.3 removed, correlated with 2 other variable(s)"
[1] "hhorig.2 removed, correlated with 1 other variable(s)"
[1] "hhorig.3 removed, correlated with 1 other variable(s)"
[1] "hhorig.4 removed, correlated with 1 other variable(s)"
[1] "hhorig.5 removed, correlated with 1 other variable(s)"
[1] "hhorig.6 removed, correlated with 1 other variable(s)"
[1] "memorig.7 removed, correlated with 1 other variable(s)"
[1] "strata removed, correlated with 1 other variable(s)"
[1] "month removed, correlated with 1 other variable(s)"
[1] "nch14resp.3 removed, correlated with 1 other variable(s)"
[1] "nch14resp.5 removed, correlated with 1 other variable(s)"
[1] "natch03.5 removed, correlated with 1 other variable(s)"
[1] "natch04.6 removed, correlated with 1 other variable(s)"
[1] "natch05.7 removed, correlated with 1 other variable(s)"
[1] "nadoptch.3 removed, correlated with 1 other variable(s)"
[1] "nchunder16.1 removed, correlated with 1 other variable(s)"
[1] "allch03.5 removed, correlated with 1 other variable(s)"
[1] "allch04.6 removed, correlated with 1 other variable(s)"
[1] "istrtdaty.2011 removed, correlated with 1 other variable(s)"
[1] "istrtdaty.2012 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.2 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.3 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.4 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.5 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.6 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.7 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.8 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.9 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.10 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.11 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.12 removed, correlated with 1 other variable(s)"
[1] "istrtdatd removed, correlated with 1 other variable(s)"
[1] "jbstat.4 removed, correlated with 1 other variable(s)"
```

```
[1] "btype1.1 removed, correlated with 1 other variable(s)"
[1] "btype2.1 removed, correlated with 1 other variable(s)"
[1] "btype5.1 removed, correlated with 1 other variable(s)"
[1] "btype6.1 removed, correlated with 1 other variable(s)"
[1] "btype9.1 removed, correlated with 1 other variable(s)"
[1] "bensta2.1 removed, correlated with 1 other variable(s)"
[1] "bensta4.1 removed, correlated with 1 other variable(s)"
[1] "bensta5.1 removed, correlated with 1 other variable(s)"
[1] "bensta6.1 removed, correlated with 1 other variable(s)"
[1] "bensta7.1 removed, correlated with 1 other variable(s)"
[1] "bensta8.1 removed, correlated with 1 other variable(s)"
[1] "fiyrdia removed, correlated with 1 other variable(s)"
[1] "arts1b13.1 removed, correlated with 1 other variable(s)"
[1] "marstat.3 removed, correlated with 1 other variable(s)"
[1] "hgbiof.3 removed, correlated with 1 other variable(s)"
[1] "respf16.2 removed, correlated with 1 other variable(s)"
[1] "scsex.2 removed, correlated with 1 other variable(s)"
[1] "scsf2a.2 removed, correlated with 1 other variable(s)"
[1] "scsf2b.2 removed, correlated with 1 other variable(s)"
[1] "scsf3a.5 removed, correlated with 1 other variable(s)"
[1] "scmolwp.2 removed, correlated with 1 other variable(s)"
[1] "scrletdwn.3 removed, correlated with 1 other variable(s)"
[1] "scfcritic.3 removed, correlated with 1 other variable(s)"
[1] "scfletdwn.3 removed, correlated with 1 other variable(s)"
[1] "scfannoy.3 removed, correlated with 1 other variable(s)"
[1] "istrtdathh removed, correlated with 1 other variable(s)"
[1] "pn1pno.2 removed, correlated with 1 other variable(s)"
[1] "pn1pno.3 removed, correlated with 1 other variable(s)"
[1] "pn2pno.3 removed, correlated with 1 other variable(s)"
[1] "fimnlabgrs_tc.1 removed, correlated with 1 other variable(s)"
[1] "j2paynet_dv removed, correlated with 1 other variable(s)"
[1] "ff_emplw.2 removed, correlated with 1 other variable(s)"
[1] "ff_bentype01.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype02.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype09.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype10.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype13.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype14.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype18.1 removed, correlated with 1 other variable(s)"
[1] "age_dv removed, correlated with 1 other variable(s)"
[1] "npn_dv.1 removed, correlated with 1 other variable(s)"
[1] "npn_dv.2 removed, correlated with 1 other variable(s)"
[1] "nnssib_dv.1 removed, correlated with 1 other variable(s)"
[1] "nnssib_dv.2 removed, correlated with 1 other variable(s)"
[1] "nnssib_dv.3 removed, correlated with 1 other variable(s)"
[1] "country.2 removed, correlated with 1 other variable(s)"
[1] "country.3 removed, correlated with 1 other variable(s)"
[1] "country.4 removed, correlated with 1 other variable(s)"
[1] "xtra5min_dv.1 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.5 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.7 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.8 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.9 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.10 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.11 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.12 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.13 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.15 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.5 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.13 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.2 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.4 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.5 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.6 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.10 removed, correlated with 1 other variable(s)"
[1] "nchild_dv.5 removed, correlated with 1 other variable(s)"
[1] "ppno.1 removed, correlated with 1 other variable(s)"
[1] "ppno.2 removed, correlated with 1 other variable(s)"
```

```
[1] "fnpno.1 removed, correlated with 1 other variable(s)"
[1] "fnpno.2 removed, correlated with 1 other variable(s)"
[1] "mnpno.1 removed, correlated with 1 other variable(s)"
[1] "mnpno.2 removed, correlated with 1 other variable(s)"
[1] "mnpno.3 removed, correlated with 1 other variable(s)"
[1] "mnpno.4 removed, correlated with 1 other variable(s)"
[1] "mnpno.5 removed, correlated with 1 other variable(s)"
[1] "paygu_if.1 removed, correlated with 1 other variable(s)"
[1] "indinbh_xw removed, correlated with 1 other variable(s)"
[1] "indinus_lw removed, correlated with 1 other variable(s)"
[1] "indpxub_xw removed, correlated with 1 other variable(s)"
[1] "fimnnet_dv removed, correlated with 1 other variable(s)"
[1] "respm16_dv.2 removed, correlated with 1 other variable(s)"
[1] "frmnthimp_dv_total removed, correlated with 1 other variable(s)"
[1] "fimnlabnet_dv removed, correlated with 1 other variable(s)"
[1] "ethn_dv removed, correlated with 1 other variable(s)"
[1] "171 variables removed since they had high correlation coefs"
[1] "The final cleaned dataset has been completed at this stage and is stored under the name 'cleanDa
[1] "----------Attempting a Train Test Split----------"
[1] "Good train, test split found"
[1] "The working seed found was 3"
[1] "----------kNN----------"
[1] "183 neighbours considered for each test data point"
[1] "kNN results as a table, follow the diagonal for the correctly mapped clusters"
         real
predicted   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15
16  17  18  19  20  21  22  23  24  25  26  27  28
       0    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
       1    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
       2    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
       3    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
       4    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
       5    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
       6    2   0   3   5  11  25 148 110  67  85  56  47  34  14   5   6
4   0   5   1   1   2   2   0   1   0   0   0   0
       7    0   1   0   1   5   3  47  73  45  58  43  46  36  23   6   5
6   5   3   1   2   1   2   1   1   0   0   0   0
       8    1   0   2   1   6   4  23  33  33  27  30  34  28  16   9   4
2   4   6   1   4   0   1   0   2   2   1   0   0
       9    0   0   1   1   0   1   6   7  10  14   6  18  15   6   4   5
1   2   3   1   2   3   1   0   2   2   0   0   0
       10   0   0   0   0   0   0   0   0   1   2   1   4   3   6   0   1
0   1   0   1   0   0   0   0   1   0   1   0   0
       11   0   0   0   0   0   1   0   3   3  11  10  13  19  11   9   4
6   5   4   7   2   0   1   0   2   1   0   0   1
       12   0   0   0   0   0   1   5   5   6   9  17  18  29  16   8  15
7   9  12  16   6  15   6   9   3   7   6   6   4
       13   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   1   0   0   0   0   0   0   1   0   0   0   0
       14   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
       15   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
       16   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
       17   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
       18   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
       19   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
       20   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
```

```
0   0   0   0   0   0   0   0   0   0   0   0   0
        21  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        22  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        23  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        24  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        25  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        26  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        27  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        28  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
          real
predicted  29  30  31  35  36
        0   0   0   0   0   0
        1   0   0   0   0   0
        2   0   0   0   0   0
        3   0   0   0   0   0
        4   0   0   0   0   0
        5   0   0   0   0   0
        6   0   0   0   0   0
        7   0   0   0   0   1
        8   0   0   0   0   0
        9   0   1   0   0   0
        10  0   0   0   0   0
        11  0   0   0   0   0
        12  3   0   1   1   2
        13  0   0   0   0   0
        14  0   0   0   0   0
        15  0   0   0   0   0
        16  0   0   0   0   0
        17  0   0   0   0   0
        18  0   0   0   0   0
        19  0   0   0   0   0
        20  0   0   0   0   0
        21  0   0   0   0   0
        22  0   0   0   0   0
        23  0   0   0   0   0
        24  0   0   0   0   0
        25  0   0   0   0   0
        26  0   0   0   0   0
        27  0   0   0   0   0
        28  0   0   0   0   0
 [ reached getOption("max.print") -- omitted 8 rows ]
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
[1] "The MSE of the predicted values are of 151.2839"
[1] "The kNN model predicts exactly with an accuracy of 0.1714"
[1] "----------CART prediction model----------"
n= 5441

node), split, n, deviance, yval
      * denotes terminal node

 1) root 5441 129439.000 10.784780
   2) sf12mcs_dv>=-0.5279083 4131  39272.500  9.090051
     4) scsf6c.5>=0.5 2271  11607.890  7.719947 *
     5) scsf6c.5< 0.5 1860  18196.440 10.762900 *
   3) sf12mcs_dv< -0.5279083 1310  40887.200 16.129010
     6) sf12mcs_dv>=-1.850473 994  20052.060 14.580480
      12) sf12pcs_dv>=-1.401703 864  14947.170 14.082180 *
      13) sf12pcs_dv< -1.401703 130   3464.492 17.892310 *
     7) sf12mcs_dv< -1.850473 316  10954.000 21.000000
      14) sf12mcs_dv>=-2.725523 214   5559.495 19.495330 *
```

```
     15) sf12mcs_dv< -2.725523 102   3893.490 24.156860 *
[1] "Variable Importance"
sf12mcs_dv   scsf4a.5   scsf4b.5   scsf4a.3   scsf6c.3    scsf7.5   scsf6c.5
scsf6c.4 sf12pcs_dv   scsf6c.2   scsf4a.2
66505.2299 20295.5441 17717.9706 14708.5488 14294.7533 12413.8647  9468.1744
7951.2303  4528.1156  3126.9419  2376.4896
  scsf4a.4   scsf4b.2    scsf7.2    scsf5.4   scsf3b.2      sf1.5   scsf3a.2
scsf1.5   scsf6a.5   scsf6b.5
 1537.3057  1083.4031   974.8790   593.0690   580.4505   492.1211   429.0286
340.6992   235.4533   206.0216
[1] "The MSE of the predicted values are of 10.1101"
[1] "The CART model predicts exactly with accuracy of 0.1279"
[1] "———————Ordinary Linear Regression (Initial)———————"
[1] "The full model AIC is: 27668.3085"
[1] "———————Variance Inflation Factor Removal———————"
[1] "The variable sf12mcs_dv was removed since it had a VIF score of 13645.2993"
[1] "The variable natch02.4 was removed since it had a VIF score of 912.6425"
[1] "The variable sppno.2 was removed since it had a VIF score of 273.1065"
[1] "The variable agegr13_dv.8 was removed since it had a VIF score of 234.2161"
[1] "The variable hrpid was removed since it had a VIF score of 162.6548"
[1] "The variable scsf4a.5 was removed since it had a VIF score of 139.5435"
[1] "The variable nchild_dv.2 was removed since it had a VIF score of 123.7172"
[1] "The variable hhtype_dv.8 was removed since it had a VIF score of 90.9036"
[1] "The variable sf12pcs_dv was removed since it had a VIF score of 84.0497"
[1] "The variable doby_dv was removed since it had a VIF score of 71.2651"
[1] "The variable scsf4b.5 was removed since it had a VIF score of 64.353"
[1] "The variable allch01.3 was removed since it had a VIF score of 60.9808"
[1] "The variable wkvege.4 was removed since it had a VIF score of 60.3997"
[1] "The variable adoptch01.3 was removed since it had a VIF score of 54.3575"
[1] "The variable rach16_dv.2 was removed since it had a VIF score of 50.8789"
[1] "The variable scsf7.5 was removed since it had a VIF score of 46.7712"
[1] "The variable scsf6c.5 was removed since it had a VIF score of 42.3825"
[1] "The variable sclfsato.6 was removed since it had a VIF score of 34.7969"
[1] "The variable hhtype_dv.11 was removed since it had a VIF score of 33.2222"
[1] "The variable scsf3b.5 was removed since it had a VIF score of 26.9019"
[1] "The variable allch02.4 was removed since it had a VIF score of 23.8488"
[1] "The variable nnatch.4 was removed since it had a VIF score of 21.5916"
[1] "The variable access.5 was removed since it had a VIF score of 21.3729"
[1] "The variable natch01.2 was removed since it had a VIF score of 20.6966"
[1] "The variable fimnlabgrs_if was removed since it had a VIF score of 19.8518"
[1] "The variable sclfsat1.6 was removed since it had a VIF score of 19.5143"
[1] "The variable natch01.3 was removed since it had a VIF score of 19.0922"
[1] "The variable fibenothr_dv was removed since it had a VIF score of 17.7918"
[1] "The variable ff_jbstat.4 was removed since it had a VIF score of 17.6642"
[1] "The variable scrannoy.3 was removed since it had a VIF score of 14.8249"
[1] "The variable nchund18resp.2 was removed since it had a VIF score of 14.7814"
[1] "The variable natch02.3 was removed since it had a VIF score of 13.8242"
[1] "The variable ndepchl_dv.1 was removed since it had a VIF score of 13.308"
[1] "The variable nunmpsp_dv.1 was removed since it had a VIF score of 12.4288"
[1] "The variable indscbh_xw was removed since it had a VIF score of 12.2643"
[1] "The variable sclfsat7.6 was removed since it had a VIF score of 12.106"
[1] "The variable npensioner_dv.2 was removed since it had a VIF score of 12.056"
[1] "The variable scrcritic.3 was removed since it had a VIF score of 11.9431"
[1] "The variable nchild_dv.3 was removed since it had a VIF score of 11.727"
[1] "The variable scopfamf.4 was removed since it had a VIF score of 11.6361"
[1] "The variable netuse.7 was removed since it had a VIF score of 10.6798"
[1] "The variable sclfsat2.6 was removed since it had a VIF score of 10.5211"
[1] "42 variables removed from the Ordinary Linear Model since they have a VIF score higher than 10"
[1] "The full model AIC after VIF checks is: 28067.7844"
[1] "———————Backwards Selection———————"
[1] "50 out of 700 variables removed so far."
[1] "100 out of 700 variables removed so far."
[1] "150 out of 700 variables removed so far."
[1] "200 out of 700 variables removed so far."
[1] "250 out of 700 variables removed so far."
[1] "300 out of 700 variables removed so far."
[1] "350 out of 700 variables removed so far."
[1] "400 out of 700 variables removed so far."
```

```
[1] "450 out of 700 variables removed so far."
[1] "500 out of 700 variables removed so far."
[1] "550 out of 700 variables removed so far."
[1] "554 out of 700 variables removed in backwards selection since they weren't significant at the 95
  [1] "fruvege.8"         "scssupr2r.9"       "btype96.1"         "sports113.1"
"intdatm_dv.8"      "netuse.2"
  [7] "sampst.2"          "nch415resp.2"      "istrtdatss"        "sports222.1"
"fruvege.6"         "fruvege.10"
 [13] "ficode17.1"        "arts2b11.1"        "pno.3"             "fnpno.3"
"scwhorusex.2"      "buno_dv.4"
 [19] "memorig.6"         "scssupr2r.8"       "hcondn16.1"        "sclfsat2.3"
"adoptch01.4"       "scloutcont.2"
 [25] "arts1b9.1"         "volun.2"           "intdatm_dv.10"     "scopfamb.4"
"mla96.1"           "nch14resp.2"
 [31] "nch5to15.1"        "allch03.4"         "seearngrs_if.1"    "memorig.4"
"sports230.1"       "hrpno.5"
 [37] "sports38.1"        "psu"               "istrtdatmm"        "ff_bentype17.1"
"hhtype_dv.5"       "aidhh.2"
 [43] "howlng"            "scwhorurac.4"      "scwhoruedu.3"      "ficode2.1"
"netuse.6"          "ficode32.1"
 [49] "event2s"           "nadoptch.1"        "pns1pno.3"         "arts2b10.1"
"nunmpsp_dv.2"      "scfundstnd.3"
 [55] "agegr13_dv.10"     "nch5to15.2"        "ficode13.1"        "agegr13_dv.4"
"gor_dv.5"          "scwhorusex.4"
 [61] "ff_bentype04.1"    "marstat_dv.4"      "arts2a5.1"         "j2has.2"
"sportact.2"        "scrletdwn.2"
 [67] "nchresp.5"         "usbread.5"         "aidxhh.2"          "ficode11.1"
"fibenothr_if"      "scfalcdrnk.2"
 [73] "ff_bentype31.1"    "scsf1.3"           "scsf1.2"           "sampst.3"
"sports16.1"        "sports219.1"
 [79] "arts1a7.1"         "scwhorupol.2"      "ff_bentype06.1"    "scssupr2r.3"
"hcondn5.1"         "scalcl7d.2"
 [85] "ff_bentype05.1"    "ficode5.1"         "hhtype_dv.12"      "natch03.4"
"arts1b12.1"        "arts1b96.1"
 [91] "mla3.1"            "arts2b15.1"        "sclfsat7.5"        "sf1.2"
"mnspno.1"          "natch01.1"
 [97] "hcondn3.1"         "fruvege.9"         "usbread.7"         "scropenup.4"
"hhtype_dv.20"      "sclfsat1.4"
[103] "jbstat.7"          "wlk30min"          "cindtime"          "allch01.6"
"scopfamd.4"        "hcondn8.1"
[109] "sports196.1"       "sclfsat1.5"        "hrpno.2"           "heritage6.1"
"hcondn7.1"         "hiqual_dv.9"
[115] "relup.4"           "marstat_dv.6"      "allch01.2"         "hhtype_dv.18"
"wkfruit.4"         "agegr10_dv.3"
[121] "sports34.1"        "sports37.1"        "sports19.1"        "ficode38.1"
"ficode27.1"        "lkmove.2"
[127] "ficode3.1"         "usbread.6"         "event1s"           "arts2a7.1"
"scfannoy.4"        "scanyelsetxt.2"
[133] "nchund18resp.4"    "nch5to15.4"        "intdatm_dv.3"      "intdaty_dv.2012"
"undqus.4"          "scrrely.3"
[139] "scfrely.3"         "scsf5.3"           "nch415resp.4"      "nch10to15.3"
"arts2a96.1"        "sports32.1"
[145] "health.2"          "fruvege.4"         "scopfamh.4"        "scopfamh.2"
"fruvege.5"         "fruvege.7"
[151] "usbread.2"         "usbread.4"         "buno_dv.3"         "natch02.6"
"arts1a6.1"         "arts1a2.1"
[157] "ff_bentype34.1"    "heritage96.1"      "heritage3.1"       "scopfamd.5"
"hcondn14.1"        "sports18.1"
[163] "ienddathh"         "sportact.10"       "sportact.7"        "sports14.1"
"sportact.3"        "sportact.6"
[169] "sportact.5"        "sportact.9"        "ficode39.1"        "ficode2.2"
"scwhorusex.3"      "finfut.3"
[175] "memorig.2"         "usdairy.2"         "access.6"          "gor_dv.6"
"finnow.2"          "sctimemnuf.3"
[181] "heritage1.1"       "scwhorurac.2"      "scwhorurac.3"      "sclfsat1.3"
"arts1b14.1"        "npensioner_dv.3"
[187] "scsf3a.3"          "daywlk"            "sports33.1"        "event4s"
"wkvege.2"          "heritage7.1"
```

```
[193] "arts2a3.1"        "marstat.5"         "sports17.1"        "sclfsat2.5"
"heritage8.1"       "ficode30.1"
[199] "scrrely.4"        "ficode4.1"         "btype3.1"          "sports229.1"
"gor_dv.2"          "gor_dv.4"
[205] "hcondn10.1"       "fimnmisc_dv"       "ficode28.1"        "btype7.1"
"ficode19.1"        "ff_bentype30.1"
[211] "ff_bentype03.1"   "arts1b10.1"        "sclfsat7.3"        "fimngrs_dv"
"fimnsben_dv"       "ficode22.1"
[217] "agegr13_dv.3"     "ff_bentype29.1"    "agegr5_dv.14"      "agegr10_dv.8"
"scrannoy.4"        "allch04.5"
[223] "natch04.5"        "natch05.6"         "scfletdwn.4"       "arts2b13.1"
"intdatm_dv.9"      "sports225.1"
[229] "hcondn12.1"       "buno_dv.2"         "ficode4.2"         "urban_dv.2"
"arts2b14.1"        "hhresp_dv.2"
[235] "scloutcont.3"     "scopfama.3"        "scage1drink"       "sctimemnuf.4"
"mobuse.2"          "arts2b9.1"
[241] "sports35.1"       "nnewborn.1"        "save.2"            "jbiindb_dv"
"ficode2.3"         "sctimemnuf.5"
[247] "fimngrs_if"       "ff_jbstat.2"       "nmpsp_dv.1"        "fnspno.1"
"walkpace.4"        "walkpace.2"
[253] "fimnprben_dv"     "ff_bentype26.1"    "fimnlabnet_tc.1"   "scsf3b.4"
"ficode34.1"        "memorig.5"
[259] "ff_bentype21.1"   "ff_jbstat.5"       "scwkvfast.2"       "intdaty_dv.2011"
"vote1.2"           "ficode16.1"
[265] "netuse.3"         "ficode24.2"        "bensta1.1"         "sclfsat7.7"
"scopfamb.5"        "scopfamb.2"
[271] "scopfamb.3"       "sports217.1"       "ppno.5"            "scrundstnd.2"
"ff_bentype11.1"    "nchund18resp.1"
[277] "scsf5.2"          "sports227.1"       "mnspno.5"          "heritage2.1"
"hhtype_dv.17"      "scssupr2r.4"
[283] "hiqual_dv.4"      "sclfsat7.2"        "nch14resp.1"       "intdatd_dv"
"scopfamh.5"        "sports36.1"
[289] "mastat_dv.3"      "arts1b11.1"        "sports13.1"        "access.2"
"access.4"          "j2pay_if.1"
[295] "ivcoop.2"         "undqus.2"          "scfrely.2"         "nnmpsp_dv.2"
"ff_jbstat.7"       "nnmpsp_dv.1"
[301] "relup.3"          "scropenup.2"       "scropenup.3"       "scrundstnd.3"
"ff_bentype16.1"    "ivprsnt.2"
[307] "nnatch.2"         "event3s"           "memorig.3"         "origadd.2"
"scwhoruage.4"      "scwhoruage.2"
[313] "arts2a6.1"        "indin91_lw"        "sportact.4"        "pno.4"
"fnspno.2"          "ff_bentype22.1"
[319] "jbstat.97"        "sports223.1"       "fruvege.2"         "ficode29.1"
"ficode4.3"         "gor_dv.10"
[325] "ficode26.2"       "agegr13_dv.9"      "fruvege.3"         "scrcritic.2"
"hiqual_dv.5"       "hiqual_dv.3"
[331] "hiqual_dv.2"      "scsf3a.4"          "susp.2"            "hhtype_dv.10"
"nchild_dv.1"       "ftedany.2"
[337] "scrannoy.2"       "scrletdwn.4"       "scfletdwn.2"       "scfannoy.2"
"hcondn1.1"         "ivcoop.3"
[343] "nnewborn.2"       "gor_dv.3"          "gor_dv.9"          "ficode24.1"
"arts1a3.1"         "arts1a1.1"
[349] "ff_bentype35.1"   "allch01.5"         "adoptch02.4"       "natch01.6"
"wkvege.3"          "heritage4.1"
[355] "mla2.1"           "intdatm_dv.5"      "sports396.1"       "ficode7.1"
"ff_bentype07.1"    "sclfsat2.7"
[361] "ficode20.1"       "respf16_dv.2"      "nch10to15.1"       "nch415resp.1"
"ff_bentype08.1"    "marstat_dv.2"
[367] "hhresp_dv.3"      "hhtype_dv.22"      "scwkvfast.4"       "scwkvfast.3"
"undqus.3"          "scwkvfast.6"
[373] "scwkvfast.5"      "lenindintv"        "gor_dv.8"          "nch5to15.3"
"agegr13_dv.12"     "marstat_dv.3"
[379] "npensioner_dv.1"  "hhtype_dv.6"       "ficode9.1"         "sf1.4"
"scsf1.4"           "scopfama.2"
[385] "sclfsat2.4"       "ienddatss"         "nnsib_dv.2"        "susp.3"
"agegr13_dv.7"      "scwhorufam.4"
[391] "scwhoruedu.4"     "heritage5.1"       "usbread.3"         "fiyrinvinc_dv"
"hcondn15.1"        "usdairy.5"
```

```
[397] "scopfamh.3"         "allch03.7"          "nchund18resp.5"    "ind5mus_lw"
"xtra5minosm_dv.1"  "sports221.1"
[403] "usdairy.4"          "arts2a2.1"          "sclfsat1.7"        "allch02.5"
"ivtrans.2"         "sports12.1"
[409] "ff_bentype38.1"     "finfut.2"           "hhtype_dv.16"      "agegr13_dv.11"
"ff_bentype24.1"    "allch03.6"
[415] "scrcritic.4"        "scssupr2r.7"        "nnatch.3"          "allch02.6"
"indin01_lw"        "indscus_lw"
[421] "adoptch01.2"        "adoptch02.3"        "hrpno.4"           "fibenothr_tc.1"
"fimninvnet_dv"     "ficode6.1"
[427] "xpmove.2"           "usdairy.6"          "ienddatmm"         "ff_bentype27.1"
"intdatm_dv.7"      "sports231.1"
[433] "sports226.1"        "sports228.1"        "scssup1.5"         "hrpno.3"
"ppno.3"            "sppno.3"
[439] "sportact.8"         "sports112.1"        "arts1b15.1"        "arts2a4.1"
"netuse.4"          "ndepchl_dv.5"
[445] "paynu_if.1"         "X.next.1"           "nmpsp_dv.2"        "mnspno.4"
"natch03.6"         "ficode31.1"
[451] "scwhoruage.3"       "scfrely.4"          "scfopenup.4"       "gor_dv.7"
"scopfamf.3"        "scopfamf.5"
[457] "arts1a5.1"          "scopfamf.2"         "ppen.2"            "scsf5.4"
"scsf3b.2"          "ff_jbstat.97"
[463] "natch01.4"          "ff_bentype19.1"     "ff_bentype20.1"    "finnow.3"
"sports15.1"        "sports296.1"
[469] "sports224.1"        "sports218.1"        "scssupr2r.2"       "nadoptch.2"
"natch01.5"         "nch14resp.4"
[475] "nchild_dv.4"        "ndepchl_dv.4"       "nch415resp.3"      "nchund18resp.3"
"scsf6a.2"          "jbstat.8"
[481] "ficode12.1"         "ff_bentype12.1"     "ficode10.1"        "jbstat.5"
"ficode29.3"        "wkfruit.3"
[487] "schmcont.4"         "ficode29.2"         "advvoucher.2"      "indscub_xw"
"racel_dv"          "sportact.1"
[493] "scssup1.4"          "hcondn11.1"         "trainany.2"        "vote6.4"
"schmcont.2"        "nnsib_dv.3"
[499] "sports111.1"        "sports11.1"         "scwhorufam.3"      "sppno.1"
"scfundstnd.4"      "scrundstnd.4"
[505] "j2pay_dv"           "sports110.1"        "sports31.1"        "natch04.7"
"bensta96.1"        "chargv.2"
[511] "vote6.2"            "vote6.3"            "hcondn9.1"         "hcondn17.1"
"sclfsat7.4"        "scopfamd.3"
[517] "scopfama.5"         "sports114.1"        "agegr13_dv.6"      "ficode18.1"
"arts1a4.1"         "natch02.5"
[523] "nch10to15.2"        "ficode26.1"         "allch02.3"         "ff_bentype15.1"
"ficode23.1"        "ff_bentype23.1"
[529] "intdatm_dv.11"      "arts2b12.1"         "jbstat.6"          "ficode15.1"
"scfcritic.4"       "sports39.1"
[535] "sppno.4"            "ppno.4"             "smever.2"          "sclfsato.5"
"intdatm_dv.6"      "scfalcdrnk.7"
[541] "scfalcdrnk.4"       "scfalcdrnk.5"       "scfalcdrnk.3"      "scfalcdrnk.6"
"scloutcont.4"      "pns1pno.2"
[547] "pns2pno.3"          "intdatm_dv.4"       "jbstat.2"          "mla1.1"
"scfcritic.2"       "arts2b96.1"
[553] "jbstat.3"           "ff_jbstat.3"
[1] "-----------Ordinary Linear Regression (Improved)-----------"

Call:
lm(formula = y ~ ., data = as.data.frame(x.data.linear))

Residuals:
     Min       1Q   Median       3Q      Max
-12.8217  -1.7444  -0.1915   1.5088  17.8512

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     8.44089    0.91199   9.256  < 2e-16 ***
pno.5          -4.19444    1.36613  -3.070 0.002149 **
natch05.8      -7.63071    3.05748  -2.496 0.012599 *
adoptch01.5    -2.76966    1.35031  -2.051 0.040303 *
```

allch01.4	0.34572	0.15901	2.174	0.029729	*
netuse.5	0.44492	0.16104	2.763	0.005750	**
kidlang	−0.11816	0.04168	−2.835	0.004602	**
sf1.3	0.28612	0.09326	3.068	0.002165	**
sf1.5	0.85292	0.40270	2.118	0.034220	*
usdairy.3	0.30149	0.11306	2.667	0.007685	**
wkfruit.2	−0.23066	0.09626	−2.396	0.016595	*
walkpace.3	0.33292	0.08825	3.772	0.000163	***
hcondn2.1	−0.78400	0.27628	−2.838	0.004560	**
hcondn4.1	−2.11149	0.80506	−2.623	0.008746	**
hcondn6.1	1.89986	0.83714	2.269	0.023279	*
hcondn13.1	−1.06572	0.42336	−2.517	0.011855	*
hcondn96.1	−0.70078	0.15699	−4.464	8.21e−06	***
jbhas.2	0.36002	0.10461	3.441	0.000583	***
btype4.1	−0.33263	0.12258	−2.714	0.006678	**
finnow.4	0.62472	0.16206	3.855	0.000117	***
finnow.5	1.28606	0.28010	4.591	4.50e−06	***
arts1a96.1	0.25640	0.09393	2.730	0.006361	**
arts2a1.1	0.24117	0.09160	2.633	0.008492	**
sports216.1	0.88456	0.40546	2.182	0.029179	*
sports220.1	−1.48863	0.61695	−2.413	0.015861	*
access.3	−0.38208	0.14539	−2.628	0.008615	**
chkcoa.2	−0.41326	0.15841	−2.609	0.009111	**
event1	−0.08762	0.04141	−2.116	0.034396	*
rhland_code.1	0.46184	0.19293	2.394	0.016709	*
scsf1.5	1.18447	0.45628	2.596	0.009459	**
scsf2a.3	−0.32603	0.14484	−2.251	0.024425	*
scsf2b.3	0.28829	0.13307	2.166	0.030319	*
scsf3a.2	−0.80264	0.24313	−3.301	0.000969	***
scsf3b.3	−0.31117	0.15325	−2.030	0.042356	*
scsf4a.2	2.03259	0.39765	5.111	3.31e−07	***
scsf4a.3	1.47380	0.21606	6.821	1.00e−11	***
scsf4a.4	0.75954	0.14785	5.137	2.88e−07	***
scsf4b.2	1.90858	0.45358	4.208	2.62e−05	***
scsf4b.3	1.21818	0.22914	5.316	1.10e−07	***
scsf4b.4	0.55792	0.14901	3.744	0.000183	***
scsf5.5	1.59828	0.39654	4.031	5.64e−05	***
scsf6a.3	0.94410	0.10782	8.757	< 2e−16	***
scsf6a.4	2.43050	0.17118	14.198	< 2e−16	***
scsf6a.5	3.59932	0.35721	10.076	< 2e−16	***
scsf6b.2	1.12758	0.20188	5.585	2.45e−08	***
scsf6b.3	1.39726	0.21384	6.534	6.99e−11	***
scsf6b.4	1.93447	0.24880	7.775	8.98e−15	***
scsf6b.5	2.36792	0.34405	6.882	6.56e−12	***
scsf6c.2	5.08825	0.27272	18.657	< 2e−16	***
scsf6c.3	2.91048	0.14539	20.018	< 2e−16	***
scsf6c.4	1.61289	0.10247	15.740	< 2e−16	***
scsf7.2	3.09043	0.31340	9.861	< 2e−16	***
scsf7.3	1.23953	0.16740	7.405	1.52e−13	***
scsf7.4	0.91402	0.12379	7.384	1.78e−13	***
scwhoruedu.2	−0.18072	0.08262	−2.187	0.028758	*
scwhorupol.3	0.24608	0.10224	2.407	0.016118	*
scwhorupol.4	0.21530	0.10632	2.025	0.042917	*
scwhorufam.2	−0.25374	0.11641	−2.180	0.029326	*
scfalcdrnk.9	−0.74691	0.27226	−2.743	0.006102	**
sclfsat1.2	−0.41891	0.20197	−2.074	0.038119	*
sclfsat2.2	0.33925	0.16678	2.034	0.041987	*
sclfsato.2	0.95166	0.25052	3.799	0.000147	***
sclfsato.3	1.52344	0.17258	8.827	< 2e−16	***
sclfsato.4	0.56599	0.17013	3.327	0.000885	***
sclfsato.7	−0.43600	0.13607	−3.204	0.001362	**
schmcont.3	0.52113	0.14430	3.611	0.000307	***
schmcont.5	0.92185	0.33670	2.738	0.006205	**
schmcont.6	1.22026	0.44258	2.757	0.005850	**
scloutcont.5	−0.25633	0.12694	−2.019	0.043516	*
scloutcont.6	−0.45969	0.18762	−2.450	0.014315	*
scdem2many.2	−0.53795	0.19251	−2.794	0.005218	**
scdem2many.3	−0.87587	0.18750	−4.671	3.07e−06	***

```
scdem2many.4     -1.10836     0.20603   -5.380 7.78e-08 ***
scdem2many.5     -1.20530     0.21300   -5.659 1.61e-08 ***
scdem2many.6     -1.40376     0.23772   -5.905 3.74e-09 ***
sctimemnuf.2     -0.25728     0.09184   -2.801 0.005107 **
sctimemnuf.6      0.42212     0.18883    2.235 0.025428 *
scrrely.2         0.22084     0.09703    2.276 0.022888 *
scfundstnd.2      0.17360     0.08742    1.986 0.047093 *
scfopenup.2       0.21914     0.09978    2.196 0.028115 *
scfopenup.3       0.38858     0.11461    3.391 0.000703 ***
scssup1.2         0.18034     0.08979    2.008 0.044657 *
scopfama.4       -0.26652     0.09293   -2.868 0.004145 **
scopfamd.2        0.18649     0.08541    2.184 0.029040 *
ff_jbstat.6      -0.44444     0.16158   -2.751 0.005969 **
ff_jbstat.8       1.29759     0.36790    3.527 0.000424 ***
ff_jbstat.9      -4.26423     1.61885   -2.634 0.008460 **
ff_bentype28.1   -1.62720     0.61589   -2.642 0.008266 **
ff_bentype33.1   -3.66787     0.78839   -4.652 3.36e-06 ***
fnspno.3         -5.87872     1.17145   -5.018 5.38e-07 ***
mnspno.2          1.22419     0.55018    2.225 0.026120 *
ficode8.1        -1.63569     0.67717   -2.415 0.015749 *
ficode14.1       -0.81837     0.35844   -2.283 0.022460 *
ficode33.1        2.28676     0.57018    4.011 6.14e-05 ***
ficode37.1        2.25398     1.10717    2.036 0.041820 *
sex_dv.2          0.58567     0.10268    5.704 1.23e-08 ***
intdatm_dv.2      0.46234     0.15059    3.070 0.002149 **
intdatm_dv.12     0.60001     0.18374    3.266 0.001100 **
marstat_dv.5     -0.70938     0.29312   -2.420 0.015550 *
nnsib_dv.1       -1.25724     0.41639   -3.019 0.002545 **
depchl_dv.2      -2.17270     0.81037   -2.681 0.007360 **
gor_dv.11         0.35694     0.13725    2.601 0.009329 **
fimnpen_dv        0.11393     0.04604    2.474 0.013375 *
hhtype_dv.19      0.40185     0.17076    2.353 0.018645 *
hhtype_dv.23      1.18801     0.50290    2.362 0.018196 *
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1      1

Residual standard error: 2.983 on 5336 degrees of freedom
Multiple R-squared:  0.6331,    Adjusted R-squared:  0.6259
F-statistic: 88.53 on 104 and 5336 DF,  p-value: < 2.2e-16

AIC: 27441.2675
MSE: 8.7284
[1] "The MSE of the predicted values are of 9.5976"
[1] "The Linear Model predicts exactly with accuracy of 0.1659"
[1] "The Linear Model predicts within a confidence interval with accuracy of 0.3809"
[1] "----------Elastic Net Regression----------"
716 x 1 sparse Matrix of class "dgCMatrix", with 42 entries
          names Estimate_Coefs
1   (Intercept)   12.139692835
2      jbstat.8    0.192253837
3         sf1.5    0.822665502
4     usdairy.3    0.025726878
5    hcondn96.1   -0.079707154
6      finnow.4    0.282228721
7      finnow.5    0.648871748
8        event1   -0.001617382
9       scsf1.5    0.331608681
10     scsf2a.3   -0.547916227
11     scsf3b.5   -0.076456720
12     scsf4a.3    0.068965784
13     scsf4a.5   -0.296949389
14     scsf4b.3    0.002256972
15     scsf4b.5   -0.058271257
16      scsf5.4    0.136170319
17      scsf5.5    0.551470547
18     scsf6a.2   -0.154381775
19     scsf6a.4    0.532857397
20     scsf6a.5    0.259477537
```

```
21      scsf6c.2    1.760712238
22      scsf6c.3    0.378880384
23      scsf6c.5   -1.115542689
24       scsf7.2    0.820725904
25       scsf7.5   -0.345218990
26   sclfsat1.3    0.029160954
27   sclfsat2.2    0.167701232
28   sclfsato.2    0.057572241
29   sclfsato.3    1.247813793
30   sclfsato.6   -0.320965020
31   sclfsato.7   -0.482253444
32   schmcont.3    0.235843162
33   schmcont.5    0.376610778
34 scloutcont.6   -0.098221256
35 scdem2many.2    0.075675028
36 scdem2many.5   -0.117300663
37 scdem2many.6   -0.213877121
38 sctimemnuf.2   -0.078443635
39 sctimemnuf.6    0.264320649
40   sf12mcs_dv   -2.186426369
41     sex_dv.2    0.079110217
42   sf12pcs_dv   -0.273823694
[1] "The MSE of the predicted values of the best fit model is 8.2415"
[1] "The Alpha of the best fit model is 0.4"
[1] "The Elastic Net Model predicts exactly with accuracy of 0.1764"
[1] "-----------Timer Results-----------"
   user  system elapsed
2375.16    9.03 2385.39
```

10.2.18 w2MergeNurse console

```
[1] "-----------Initial Checks-----------"
[1] "13012362 NA cells were found across the entire dataset (59.01% of data as NA)"
[1] "-----------Data Type Checks-----------"
[1] "0 variables recoded since all their entries aren't numeric or NA"
[1] "NOTE: algorithim recodes categorical data alphabetically e.g (female = 1, male = 2)"
character(0)
[1] "-----------Low Data Removal-----------"
[1] "1364 variables removed since they had >= 'naPercent' (default 20%) NA values"
   [1] "pid.x"          "lvwhy"          "lvmthp"         "lvyrp"          "mvever"
"mvmnth"         "mvyr"           "mlstatchk"
   [9] "mlstat.x"       "drive"          "caruse"         "ukborn"         "plbornc"
"yr2uk4"         "citzn1"         "citzn2"
  [17] "citzn3"         "qfhigh"         "qualoc"         "qfvoc1"         "qfvoc2"
"qfvoc3"         "qfvoc4"         "qfvoc5"
  [25] "qfvoc6"         "qfvoc7"         "qfvoc8"         "qfvoc9"         "qfvoc10"
"qfvoc11"        "qfvoc12"        "qfvoc13"
  [33] "qfvoc14"        "qfvoc15"        "qfvoc96"        "school"         "scend"
"schlloc"        "schok"          "fenow"
  [41] "feend"          "j1none"         "j1semp"         "j1boss"         "j1mngr"
"paedqf"         "maedqf"         "edtype"
  [49] "edasp"          "fedlik"         "fednt"          "ocimpa"         "ocimpb"
"ocimpe"         "ocimpf"         "ocimpi"
  [57] "ocimpk"         "ocimpl"         "futra"          "futrb"          "futrc"
"futrd"          "futre"          "futrf"
  [65] "futrg"          "futrh"          "futri"          "futrj"          "futrk"
"futrl"          "paju"           "maju"
  [73] "pacob"          "payruk"         "macob"          "mayruk"         "natid1"
"natid2"         "natid3"         "natid4"
  [81] "natid5"         "natid6"         "natid97"        "racel"          "racelo_code"
"ethid2"         "ethid3"         "ethid4a"
  [89] "ethid4b"        "ethid5"         "ethid6"         "ethid7"         "ethid8"
"ethid9"         "ethid10"        "ethid11"
  [97] "ethid12"        "ethid13"        "ethid14"        "ethclose1"      "ethclose2a"
"ethclose2b"     "ethclose3"      "ethclose4a"
```

```
 [105] "ethclose4b"    "ethclose5"     "ethclose6"     "ethclose7"     "ethclose8"
"ethclose9"     "ethclose10"    "ethclose11"
 [113] "pride2"        "pride4a"       "pride4b"       "pride6"        "pride7"
"pride5"        "pride8"        "pride9"
 [121] "pride10"       "pride11"       "pride12"       "pride13"       "pride14"
"food1"         "food2"         "food3"
 [129] "food4"         "food5"         "food6"         "food7"         "oprlg"
"oprlg0ni"      "nirel"         "niact"
 [137] "oprlg0"        "oprlg1"        "daywlk"        "wlk30min"      "walkpace"
"smnow"         "ncigs"         "smcigs"
 [145] "smncigs"       "aglquit"       "smagbg"        "hospc1"        "hospdc1"
"hospc2"        "hospdc2"       "hospc3"
 [153] "hospdc3"       "hospc4"        "hospdc4"       "hospc5"        "hospdc5"
"hospc6"        "hospdc6"       "hospc7"
 [161] "hospdc7"       "hospc8"        "hospdc8"       "disdif1"       "disdif2"
"disdif3"       "disdif4"       "disdif5"
 [169] "disdif6"       "disdif7"       "disdif8"       "disdif9"       "disdif10"
"disdif11"      "disdif12"      "disdif96"
 [177] "aidhh"         "aidhua1"       "aidhua2"       "aidhua3"       "aidhua4"
"aidhua5"       "aidhua6"       "aidhua7"
 [185] "aidhua8"       "aidhua9"       "aidhua10"      "aidhua11"      "aidhua12"
"aidhua13"      "aidhua14"      "aidhua15"
 [193] "aidhua16"      "naidxhh"       "aidhu1"        "aidhu2"        "aidhrs"
"aideft"        "lcohnpi"       "coh1bm"
 [201] "coh1by"        "coh1mr"        "coh1em"        "coh1ey"        "nmar"
"lmar1m"        "lmar1y"        "ladopt"
 [209] "lnadopt"       "lprnt"         "lnprnt"        "ch1by4"        "movy11"
"movy12"        "movy13"        "movy14"
 [217] "movy15"        "movy16"        "family"        "education"     "memploy"
"housing"       "area"          "moveoth_code"
 [225] "movdir"        "plnowm"        "plnowy4"       "mstatsam"      "lwwrong"
"mstatsamn"     "lastmstatch1" "mstatch1"
 [233] "statcm1"       "statcy41"      "divchk1"       "divfin1"       "dvm1"
"dvy41"         "cmlstat1"      "lastmstatch2"
 [241] "mstatch2"      "statcm2"       "statcy42"      "divchk2"       "divfin2"
"dvm2"          "dvy42"         "cmlstat2"
 [249] "lastmstatch3" "mstatch3"       "statcm3"       "statcy43"      "divchk3"
"divfin3"       "dvm3"          "dvy43"
 [257] "cmlstat3"      "cohab"         "cohabn"        "lmcbm1"        "lmcby41"
"currpart1"     "lmspm1"        "lmspy41"
 [265] "lmcbm2"        "lmcby42"       "currpart2"     "lmspm2"        "lmspy42"
"lmcbm3"        "lmcby43"       "currpart3"
 [273] "lmspm3"        "lmspy43"       "lmcbm4"        "lmcby44"       "currpart4"
"lmspm4"        "lmspy44"       "father"
 [281] "nchild"        "preg"          "pregm1"        "pregy41"       "pregfert1"
"invitro1"      "pregout1"      "pregend1"
 [289] "endmnth1"      "pregsmoke1"    "smkmnth11"     "smkmnth21"     "smkmnth31"
"pregsmk11"     "pregsmk21"     "pregsmk31"
 [297] "pregdrink1"    "lchmulti1"     "pregm2"        "pregy42"       "pregfert2"
"invitro2"      "pregout2"      "pregend2"
 [305] "endmnth2"      "pregsmoke2"    "smkmnth12"     "smkmnth22"     "smkmnth32"
"pregsmk12"     "pregsmk22"     "pregsmk32"
 [313] "pregdrink2"    "lchmulti2"     "pregm3"        "pregy43"       "pregfert3"
"invitro3"      "pregout3"      "pregend3"
 [321] "endmnth3"      "pregsmoke3"    "smkmnth13"     "smkmnth23"     "smkmnth33"
"pregsmk13"     "pregsmk23"     "pregsmk33"
 [329] "pregdrink3"    "lchmulti3"     "pregm4"        "pregy44"       "pregfert4"
"invitro4"      "pregout4"      "pregend4"
 [337] "endmnth4"      "pregsmoke4"    "smkmnth14"     "smkmnth24"     "smkmnth34"
"pregsmk14"     "pregsmk24"     "pregsmk34"
 [345] "pregdrink4"    "lchmulti4"     "pregm5"        "pregy45"       "pregfert5"
"invitro5"      "pregout5"      "pregend5"
 [353] "endmnth5"      "pregsmoke5"    "smkmnth15"     "smkmnth25"     "smkmnth35"
"pregsmk15"     "pregsmk25"     "pregsmk35"
 [361] "pregdrink5"    "lchmulti5"     "hcondno1"      "hcondns1"      "hcondno2"
"hcondns2"      "hcondno3"      "hcondns3"
 [369] "hcondno4"      "hcondns4"      "hcondno5"      "hcondns5"      "hcondno6"
"hcondns6"      "hcondno7"      "hcondns7"
```

```
 [377] "hcondno8"      "hcondns8"      "contft"        "ftendm"        "ftendy4"
"ftquals"      "ftedstartm1"   "ftedstarty41"
 [385] "ftedend1"      "ft2endm1"      "ft2endy41"     "ftedmor1"      "ftedstartm2"
"ftedstarty42" "ftedend2"      "ft2endm2"
 [393] "ft2endy42"     "ftedmor2"      "qualnew1"      "qualnew2"      "qualnew3"
"qualnew4"     "qualnew5"      "qualnew6"
 [401] "qualnew7"      "qualnew8"      "qualnew9"      "qualnew10"     "qualnew11"
"qualnew13"    "qualnew15"     "qualnew16"
 [409] "qualnew17"     "qualnew18"     "qualnew19"     "qualnew20"     "qualnew21"
"qualnew22"    "qualnew23"     "qualnew24"
 [417] "qualnew25"     "qualnew26"     "qualnew27"     "qualnew28"     "qualnew29"
"qualnew30"    "qualnew31"     "trwho1"
 [425] "traindays1"    "trainhrs1"     "trainend1"     "trainpurp11"   "trainpurp21"
"trainpurp31"  "trainpurp41"   "trainpurp51"
 [433] "trainpurp61"   "trainpurp71"   "trainqual1"    "trwho2"        "traindays2"
"trainhrs2"    "trainend2"     "trainpurp12"
 [441] "trainpurp22"   "trainpurp32"   "trainpurp42"   "trainpurp52"   "trainpurp62"
"trainpurp72"  "trainqual2"    "trwho3"
 [449] "traindays3"    "trainhrs3"     "trainend3"     "trainpurp13"   "trainpurp23"
"trainpurp33"  "trainpurp43"   "trainpurp53"
 [457] "trainpurp63"   "trainpurp73"   "trainqual3"    "trainn"        "trqual1"
"trqual2"      "trqual3"       "trqual4"
 [465] "trqual5"       "trqual6"       "trqual7"       "trqual8"       "trqual9"
"trqual10"     "trqual11"      "trqual13"
 [473] "trqual15"      "trqual16"      "trqual17"      "trqual18"      "trqual19"
"trqual20"     "trqual21"      "trqual22"
 [481] "trqual23"      "trqual24"      "trqual25"      "trqual26"      "trqual27"
"trqual28"     "trqual29"      "trqual30"
 [489] "trqual31"      "notempchk"     "empchk"        "empstendd"     "empstendm"
"empstendy4"   "stendreas"     "nxtst"
 [497] "nxtstelse"     "cstat"         "nxtstendd"     "nxtstendm"     "nxtstendy4"
"jbsamr"       "wkplsam"       "samejob"
 [505] "matlv"         "matlvstd"      "matlvstm"      "matlvsty4"     "matlvendd"
"matlvendm"    "matlvendy4"    "jbendd"
 [513] "jbendm"        "jbendy4"       "jbendreas"     "cjob"          "nxtjbhrs"
"nxtjbes"      "nxtjbendd"     "nxtjbendm"
 [521] "nxtjbendy4"    "cjbatt"        "nextstat1"     "nextelse1"     "currstat1"
"nextjob1"     "currjob1"      "jobhours1"
 [529] "reasend1"      "jbatt1"        "statendd1"     "statendm1"     "statendy41"
"nextstat2"    "nextelse2"     "currstat2"
 [537] "nextjob2"      "currjob2"      "jobhours2"     "reasend2"      "jbatt2"
"statendd2"    "statendm2"     "statendy42"
 [545] "nextstat3"     "nextelse3"     "currstat3"     "nextjob3"      "currjob3"
"jobhours3"    "reasend3"      "jbatt3"
 [553] "statendd3"     "statendm3"     "statendy43"    "nextstat4"     "nextelse4"
"currstat4"    "nextjob4"      "currjob4"
 [561] "jobhours4"     "reasend4"      "jbatt4"        "statendd4"     "statendm4"
"statendy44"   "nextstat5"     "nextelse5"
 [569] "currstat5"     "nextjob5"      "currjob5"      "jobhours5"     "reasend5"
"jbatt5"       "statendd5"     "statendm5"
 [577] "statendy45"    "nextstat6"     "nextelse6"     "currstat6"     "nextjob6"
"currjob6"     "jobhours6"     "reasend6"
 [585] "jbatt6"        "statendd6"     "statendm6"     "statendy46"    "nextstat7"
"nextelse7"    "currstat7"     "nextjob7"
 [593] "currjob7"      "jobhours7"     "reasend7"      "jbatt7"        "statendd7"
"statendm7"    "statendy47"    "nextstat8"
 [601] "nextelse8"     "currstat8"     "nextjob8"      "currjob8"      "jobhours8"
"reasend8"     "jbatt8"        "statendd8"
 [609] "statendm8"     "statendy48"    "nextstat9"     "nextelse9"     "currstat9"
"nextjob9"     "currjob9"      "jobhours9"
 [617] "reasend9"      "jbatt9"        "statendd9"     "statendm9"     "statendy49"
"nextstat10"   "nextelse10"    "currstat10"
 [625] "nextjob10"     "currjob10"     "jobhours10"    "reasend10"     "jbatt10"
"statendd10"   "statendm10"    "statendy410"
 [633] "jboff"         "jboffy"        "jbterm1"       "jbterm2"       "jbsic07chk"
"jbsoc00chk"   "jbsempchk"     "jbsemp"
 [641] "jbbgd"         "jbbgm"         "jbbgy"         "jbmngrchk"     "jbmngr"
"jbsizechk"    "jbsize"        "jbsect"
```

```
[649] "jbsectpub"   "jbhrs"       "jbot"        "jbotpd"      "paygwc"
"paynwc"      "payusl"      "payu"
[657] "payuwc"      "payug"       "paytyp"      "ovtpay"      "extnsa"
"extrate"     "extrest"     "basnsa"
[665] "basrate"     "basrest"     "ovtnsa"      "ovtrate"     "ovtrest"
"jbpl"        "jbttwt"      "worktrav"
[673] "jsboss"      "jssize"      "jshrs"       "jstypeb"     "jsaccs"
"jspart"      "jsprbm"      "jsprby4"
[681] "jsprem"      "jsprey4"     "jsprls"      "jsprtx"      "jsprni"
"jspayu"      "jspayw"      "jspytx"
[689] "jspyni"      "jspl"        "jsttwt"      "jsttwtb"     "jsworktrav"
"workdis"     "twkdiff1"    "twkdiff2"
[697] "twkdiff3"    "twkdiff4"    "twkdiff5"    "twkdiff6"    "twkdiff7"
"twkdiff8"    "twkdiff97"   "twkdiffm"
[705] "twkcar"      "twkcary1"    "twkcary2"    "twkcary3"    "twkcary4"
"twkcary5"    "twkcary6"    "twkcary7"
[713] "twkcary8"    "twkcary9"    "twkcary10"   "twkcary11"   "twkcary12"
"twkcary13"   "twkcary14"   "twkcary97"
[721] "twkcarym"    "altcar1"     "altcar2"     "altcar3"     "altcar4"
"altcar5"     "altcar6"     "altcar7"
[729] "altcar8"     "altcar9"     "altcar10"    "altcar11"    "altcar12"
"altcar96"    "altcar97"    "carclub"
[737] "carshare"    "wkhome"      "lifthh"      "liftxhh"     "motcyc"
"comtaxi"     "combus"      "comtrain"
[745] "commetro"    "combike"     "comwalk"     "comother"    "jbsat"
"wkphys"      "jbperfp"     "jbonus"
[753] "jbrise"      "tujbpl"      "tuin1"       "jbpen"       "jbpenm"
"jbpeny4"     "penmcn"      "penmpy"
[761] "penmtp"      "penspb"      "wktime"      "wkends"      "jbflex1"
"jbflex2"     "jbflex3"     "jbflex4"
[769] "jbflex5"     "jbflex6"     "jbflex7"     "jbflex8"     "jbflex96"
"jbfxuse1"    "jbfxuse2"    "jbfxuse3"
[777] "jbfxuse4"    "jbfxuse5"    "jbfxuse6"    "jbfxuse7"    "jbfxuse8"
"jbfxuse96"   "jbfxinf"     "wkaut1"
[785] "wkaut2"      "wkaut3"      "wkaut4"      "wkaut5"      "depenth1"
"depenth2"    "depenth3"    "depenth4"
[793] "depenth5"    "depenth6"    "jblkcha"     "jbxpcha"     "jblkchb"
"jbxpchb"     "jblkchc"     "jbxpchc"
[801] "jblkchd"     "jbxpchd"     "jblkche"     "jbxpche"     "jbsec"
"julk4wk"     "julkjb"      "jubgn"
[809] "julk4x1"     "julk4x2"     "julk4x3"     "julk4x4"     "julk4x5"
"julk4x6"     "julk4x96"    "jbhad"
[817] "jlendm"      "jlendy"      "jlsemp"      "jlboss"      "jlmngr"
"jlsize"      "eprosh"      "j2semp"
[825] "j2hrs"       "j2pay"       "retchk"      "ageret"      "rtpro1"
"rtpro2"      "rtpro3"      "rtpro4"
[833] "rtpro5"      "rtpro6"      "rtcon1"      "rtcon2"      "rtcon3"
"rtcon4"      "penmex"      "pppex"
[841] "pppexm"      "sppen"       "rtexpjb"     "rtfnd1"      "rtfnd2"
"rtfnd3"      "rtfnd4"      "rtfnd5"
[849] "rtfnd6"      "rtfnd7"      "rtfnd8"      "rtfnd9"      "rtfnd10"
"rtfnd96"     "retamt"      "retsuf"
[857] "volfreq"     "volhrs"      "charfreq"    "charam"      "ccare"
"ccwork"      "benunemp1"   "benunemp2"
[865] "benunemp96"  "bendis1"     "bendis2"     "bendis3"     "bendis4"
"bendis5"     "bendis6"     "bendis7"
[873] "bendis8"     "bendis9"     "bendis10"    "bendis11"    "bendis96"
"benpen1"     "benpen2"     "benpen3"
[881] "benpen4"     "benpen5"     "benpen6"     "benpen7"     "benpen8"
"benpen96"    "niserps"     "bencb"
[889] "benctc"      "benfam1"     "benfam2"     "benfam3"     "benfam4"
"benfam5"     "benfam96"    "bentax1"
[897] "bentax2"     "bentax3"     "bentax4"     "bentax5"     "bentax96"
"benhou1"     "benhou2"     "benhou3"
[905] "benhou4"     "benhou96"    "nfh01"       "nfh02"       "nfh03"
"nfh04"       "nfh05"       "nfh06"
[913] "nfh07"       "nfh08"       "nfh09"       "nfh10"       "nfh11"
"nfh12"       "nfh13"       "nfh14"
```

```
 [921] "nfh15"        "nfh16"        "nfh17"        "nfh18"        "nfh19"
"nfh20"        "nfh21"        "nfh22"
 [929] "nfh23"        "nfh24"        "nfh25"        "nfh26"        "nfh27"
"nfh28"        "nfh29"        "nfh30"
 [937] "nfh31"        "nfh32"        "nfh33"        "nfh34"        "nfh35"
"nfh36"        "nfh37"        "nfh38"
 [945] "fiyrdia"      "fiyrdb1"      "fiyrdb2"      "fiyrdb3"      "fiyrdb4"
"fiyrdb5"      "fiyrdb6"      "ppent"
 [953] "ppyrs"        "ppreg"        "ppram"        "pprampc"      "saved"
"savreg"       "savlt"        "hubuys"
 [961] "hufrys"       "humops"       "huiron"       "hupots"       "hudiy"
"husits"       "huboss"       "vote2"
 [969] "vote3"        "vote4"        "vote5"        "perpolinf"    "colbens1"
"colbens2"     "colbens3"     "civicduty"
 [977] "polcost"      "votenorm"     "perbfts"      "grpbfts"      "voteintent"
"demorient"    "vote7"        "vote8"
 [985] "arts1freq"    "arts2freq"    "libfreq"      "arcfreq"      "musfreq"
"herfreq"      "sportsfreq"   "sports3freq"
 [993] "club"         "acc1to161"    "acc1to162"    "acc1to163"    "acc1to164"
"acc1to165"    "acc1to166"    "acc1to167"
 [ reached getOption("max.print") -- omitted 364 entries ]
[1] "----------Low Level Removal----------"
[1] "If a level is removed from a variable you wish to keep, reccomended to manually merge levels too
[1] "level 5 in pno removed, 1 observations found"
[1] "level 3 in hhorig.x removed, 1 observations found"
[1] "level 5 in nch14resp removed, 1 observations found"
[1] "level 5 in nchresp removed, 1 observations found"
[1] "level 5 in nchund18resp removed, 0 observations found"
[1] "level 6 in natch01 removed, 2 observations found"
[1] "level 7 in natch03 removed, 1 observations found"
[1] "level 5 in natch04 removed, 3 observations found"
[1] "level 7 in natch04 removed, 3 observations found"
[1] "level 7 in natch05 removed, 2 observations found"
[1] "level 8 in natch05 removed, 1 observations found"
[1] "level 9 in natch06 removed, 0 observations found"
[1] "level 5 in nnatch removed, 0 observations found"
[1] "level 6 in nnatch removed, 0 observations found"
[1] "level 3 in nadoptch removed, 2 observations found"
[1] "level 4 in nadoptch removed, 2 observations found"
[1] "level 1 in adoptch01 removed, 3 observations found"
[1] "level 2 in adoptch01 removed, 4 observations found"
[1] "level 5 in adoptch01 removed, 2 observations found"
[1] "level 6 in adoptch01 removed, 3 observations found"
[1] "level 3 in adoptch02 removed, 0 observations found"
[1] "level 5 in adoptch02 removed, 3 observations found"
[1] "level 7 in adoptch02 removed, 0 observations found"
[1] "level 5 in adoptch03 removed, 0 observations found"
[1] "level 6 in adoptch03 removed, 0 observations found"
[1] "level 6 in adoptch04 removed, 0 observations found"
[1] "level 7 in adoptch04 removed, 0 observations found"
[1] "level 5 in nchunder16 removed, 0 observations found"
[1] "level 6 in nchunder16 removed, 0 observations found"
[1] "level 5 in nch5to15 removed, 0 observations found"
[1] "level 6 in nch5to15 removed, 0 observations found"
[1] "level 4 in nch10to15 removed, 2 observations found"
[1] "level 6 in allch01 removed, 3 observations found"
[1] "level 7 in allch02 removed, 0 observations found"
[1] "level 8 in allch02 removed, 0 observations found"
[1] "level 7 in allch03 removed, 0 observations found"
[1] "level 7 in allch04 removed, 0 observations found"
[1] "level 8 in allch04 removed, 0 observations found"
[1] "level 7 in allch05 removed, 0 observations found"
[1] "level 8 in allch05 removed, 0 observations found"
[1] "level 9 in allch06 removed, 0 observations found"
[1] "level 10 in jbstat.x removed, 4 observations found"
[1] "level 9 in kidlang removed, 1 observations found"
[1] "level 15 in kidlang removed, 2 observations found"
[1] "level 16 in kidlang removed, 1 observations found"
```

```
[1] "level 18 in kidlang removed, 1 observations found"
[1] "level 11 in fruvege removed, 1 observations found"
[1] "level 12 in fruvege removed, 2 observations found"
[1] "level 13 in fruvege removed, 1 observations found"
[1] "level 20 in fruvege removed, 0 observations found"
[1] "level 26 in fruvege removed, 1 observations found"
[1] "level 2 in nnewborn removed, 2 observations found"
[1] "level 1 in hcondn3 removed, 1 observations found"
[1] "level 1 in hcondn9 removed, 4 observations found"
[1] "level 1 in bensta3 removed, 2 observations found"
[1] "level 4 in ivcoop removed, 4 observations found"
[1] "level 3 in susp removed, 4 observations found"
[1] "level 4 in undqus removed, 4 observations found"
[1] "level 1 in ivtrans removed, 2 observations found"
[1] "level 4 in hgbiom removed, 3 observations found"
[1] "level 5 in hgbiom removed, 1 observations found"
[1] "level 3 in hgbiof removed, 3 observations found"
[1] "level 5 in hgbiof removed, 1 observations found"
[1] "level 5 in scssupr2r removed, 1 observations found"
[1] "level 22 in istrtdathh removed, 1 observations found"
[1] "level 4 in pn1pno removed, 0 observations found"
[1] "level 5 in pn1pno removed, 0 observations found"
[1] "level 3 in pn2pno removed, 0 observations found"
[1] "level 4 in pn2pno removed, 0 observations found"
[1] "level 4 in pns1pno removed, 0 observations found"
[1] "level 5 in pns1pno removed, 0 observations found"
[1] "level 3 in pns2pno removed, 0 observations found"
[1] "level 4 in pns2pno removed, 0 observations found"
[1] "level 1 in fiyrinvinc_tc removed, 1 observations found"
[1] "level 1 in fibenothr_tc removed, 3 observations found"
[1] "level 9 in ff_jbstat removed, 2 observations found"
[1] "level 10 in ff_jbstat removed, 2 observations found"
[1] "level 1 in ff_bentype06 removed, 1 observations found"
[1] "level 1 in ff_bentype21 removed, 3 observations found"
[1] "level 1 in ff_bentype25 removed, 1 observations found"
[1] "level 1 in ff_bentype30 removed, 1 observations found"
[1] "level 1 in ff_bentype35 removed, 3 observations found"
[1] "level 1 in ff_bentype36 removed, 1 observations found"
[1] "level 1 in ngrp_dv removed, 2 observations found"
[1] "level 2 in nnssib_dv removed, 3 observations found"
[1] "level 5 in nnssib_dv removed, 1 observations found"
[1] "level 3 in country removed, 1 observations found"
[1] "level 2 in agegr13_dv removed, 2 observations found"
[1] "level 4 in buno_dv removed, 2 observations found"
[1] "level 5 in buno_dv removed, 0 observations found"
[1] "level 5 in nchild_dv removed, 0 observations found"
[1] "level 6 in nchild_dv removed, 0 observations found"
[1] "level 4 in hrpno removed, 4 observations found"
[1] "level 5 in hrpno removed, 2 observations found"
[1] "level 4 in ppno removed, 4 observations found"
[1] "level 5 in ppno removed, 3 observations found"
[1] "level 4 in sppno removed, 0 observations found"
[1] "level 5 in sppno removed, 0 observations found"
[1] "level 3 in fnpno removed, 0 observations found"
[1] "level 5 in fnpno removed, 0 observations found"
[1] "level 3 in fnspno removed, 0 observations found"
[1] "level 5 in fnspno removed, 0 observations found"
[1] "level 4 in mnpno removed, 0 observations found"
[1] "level 5 in mnpno removed, 0 observations found"
[1] "level 4 in mnspno removed, 0 observations found"
[1] "level 5 in mnspno removed, 0 observations found"
[1] "level 1 in grmpno removed, 0 observations found"
[1] "level 2 in grmpno removed, 0 observations found"
[1] "level 2 in nnmpsp_dv removed, 2 observations found"
[1] "level 3 in ficode3 removed, 1 observations found"
[1] "level 1 in ficode6 removed, 4 observations found"
[1] "level 2 in ficode24 removed, 1 observations found"
[1] "level 1 in ficode25 removed, 0 observations found"
```

```
[1] "level 2 in ficode26 removed, 3 observations found"
[1] "level 2 in ficode28 removed, 1 observations found"
[1] "level 1 in ficode30 removed, 1 observations found"
[1] "level 1 in ficode35 removed, 0 observations found"
[1] "level 4 in b_pno removed, 3 observations found"
[1] "level 5 in b_pno removed, 0 observations found"
[1] "level 3 in medcnjd removed, 3 observations found"
[1] "level 1 in difbpc4 removed, 3 observations found"
[1] "level 4 in nseqno removed, 0 observations found"
[1] "level 5 in nseqno removed, 0 observations found"
[1] "level 9 in elig removed, 1 observations found"
[1] "level 96 in ag16g10 removed, 1 observations found"
[1] "level 96 in ag16g20 removed, 0 observations found"
[1] "level 4 in wstokb removed, 1 observations found"
[1] "level 7 in hhsize removed, 0 observations found"
[1] "level 8 in hhsize removed, 0 observations found"
[1] "level 9 in hhsize removed, 0 observations found"
[1] "level 10 in jbstat.y removed, 0 observations found"
[1] "level 2 in nnsib_dv removed, 0 observations found"
[1] "level 2 in agegr10_dv removed, 4 observations found"
[1] "level 1 in depchl_dv removed, 0 observations found"
[1] "level 9 in qfhigh_dv removed, 1 observations found"
[1] "level 11 in gor_dv removed, 0 observations found"
[1] "136 total levels removed from 91 different variables. In total 159 observations deleted"
[1] "————Variance 0 Check————"
[1] "122 variables removed since their new variance was 0"
  [1] "wave"          "hhorig.x"      "memorig"       "sampst"        "ivfio"
"ioutcome"      "newper"
  [8] "newentrant"    "adstatus"      "natch05"       "natch06"       "natch07"
"natch08"       "natch09"
 [15] "natch10"       "natch11"       "natch12"       "natch13"       "natch14"
"natch15"       "natch16"
 [22] "adoptch03"     "adoptch04"     "adoptch05"     "adoptch06"     "adoptch07"
"adoptch08"     "adoptch09"
 [29] "adoptch10"     "adoptch11"     "adoptch12"     "adoptch13"     "adoptch14"
"adoptch15"     "adoptch16"
 [36] "allch05"       "allch06"       "allch07"       "allch08"       "allch09"
"allch10"       "allch11"
 [43] "allch12"       "allch13"       "allch14"       "allch15"       "allch16"
"chkdob"        "hcondn3"
 [50] "hcondn9"       "hcondn15"      "bensta3"       "indmode"       "ivtrans"
"sceverdrnk"    "screlany"
 [57] "scfrendany"    "intdatd_if"    "intdatm_if"    "intdaty_if"    "doby_if"
"age_if"        "fiyrinvinc_tc"
 [64] "fibenothr_tc"  "ff_ivlolw"     "ff_everint"    "ff_bentype06"  "ff_bentype21"
"ff_bentype25"  "ff_bentype30"
 [71] "ff_bentype31"  "ff_bentype32"  "ff_bentype35"  "ff_bentype36"  "ff_bentype37"
"ff_bentype38"  "ngrp_dv"
 [78] "grfpno"        "grmpno"        "indpxbh_xw"    "indinbh_xw"    "indscbh_xw"
"indin91_lw"    "indin01_lw"
 [85] "ficode6"       "ficode21"      "ficode25"      "ficode30"      "ficode31"
"ficode32"      "ficode35"
 [92] "ficode36"      "hhorig.y"      "b_splitnum"    "tbmed"         "medtyp13"
"resphts"       "respwts"
 [99] "bfpcok"        "whintro"       "bpconst"       "respbps"       "difbpc4"
"mmgswil"       "mmgsok"
[106] "mmgssta"       "lungsurg"      "clotb"         "fit"           "dateok"
"bfck2"         "htok"
[113] "wtok"          "bmiok"         "elig"          "full1"         "full2"
"full3"         "wstokb"
[120] "depchl_dv"     "scflag_dv"     "qfhighfl_dv"
[1] "————Dummy Variables————"
[1] "predictor variable count went from 532 to 1083"
[1] "————Variance 0 Check————"
[1] "102 variables removed since their new variance was 0"
  [1] "pno.4"          "pno.5"          "nch14resp.5"    "nchresp.5"
"nchund18resp.5" "natch01.6"      "natch03.7"
  [8] "natch04.5"      "natch04.7"      "nnatch.5"       "nnatch.6"
```

```
"nadoptch.3"      "nadoptch.4"      "adoptch01.1"
 [15] "adoptch01.2"     "adoptch01.5"     "adoptch01.6"     "adoptch02.3"
"adoptch02.5"     "adoptch02.7"     "nchunder16.5"
 [22] "nchunder16.6"    "nch5to15.5"      "nch5to15.6"      "nch10to15.4"
"allch01.6"       "allch02.7"       "allch02.8"
 [29] "allch03.7"       "allch04.7"       "allch04.8"       "jbstat.x.10"
"kidlang.9"       "kidlang.15"      "kidlang.16"
 [36] "kidlang.18"      "fruvege.11"      "fruvege.12"      "fruvege.13"
"fruvege.20"      "fruvege.26"      "nnewborn.2"
 [43] "ivcoop.4"        "susp.3"          "undqus.4"        "hgbiom.4"
"hgbiom.5"        "hgbiof.3"        "hgbiof.5"
 [50] "scssupr2r.5"     "istrtdathh.22"   "pn1pno.4"        "pn1pno.5"
"pn2pno.3"        "pn2pno.4"        "pns1pno.4"
 [57] "pns1pno.5"       "pns2pno.3"       "pns2pno.4"       "ff_jbstat.9"
"ff_jbstat.10"    "nnssib_dv.2"     "nnssib_dv.5"
 [64] "country.3"       "agegr13_dv.3"    "buno_dv.4"       "buno_dv.5"
"nchild_dv.5"     "nchild_dv.6"     "hrpno.4"
 [71] "hrpno.5"         "ppno.4"          "ppno.5"          "sppno.4"
"sppno.5"         "fnpno.3"         "fnpno.5"
 [78] "fnspno.3"        "fnspno.5"        "mnpno.4"         "mnpno.5"
"mnspno.4"        "mnspno.5"        "nnmpsp_dv.2"
 [85] "ficode3.3"       "ficode24.2"      "ficode26.2"      "ficode28.2"
"b_pno.4"         "b_pno.5"         "medcnjd.3"
 [92] "nseqno.4"        "nseqno.5"        "ag16g10.96"      "ag16g20.96"
"hhsize.7"        "hhsize.8"        "hhsize.9"
 [99] "jbstat.y.10"     "nnsib_dv.2"      "qfhigh_dv.9"     "gor_dv.11"
[1] "——————K—Means——————"
[1] "15 clusters have been made for K—Means"
[1] "K—Means results as a table, the max value in each row is a simple way to define which cluster re

      0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
  1   0  0  0  2  1  6 17 18 17 16 10 17 13 11  5  6  3  4  1  1  1  0  3
3  0  2  1  0  0  0  1  1  0  0  1  0
  2   0  0  0  1  4  4 27 23 25 23 28 26 11  4  6  4  3  2  5  4  5  3  5
0  3  0  1  1  0  0  0  0  0  0  3  0
  3   0  0  0  0  1  3 10 10 11  4  3 10  6  3  0  2  2  1  0  0  1  0  0
0  0  0  0  0  0  1  0  0  0  0  0  0
  4   1  2  3  3  6  5 37 44 38 46 44 26 35 17 10  8 12  5  6  7  6  3  2
4  2  2  1  1  2  0  0  0  0  0  1  0
  5   1  0  0  1  1  4 23 24 14 22 14 12 15 17  4  4  4  1  6  1  3  4  3
1  0  0  1  0  0  2  0  0  0  0  0  0
  6   0  0  0  1  3  2 26 23 15 24 24 13 18  8  7  3  0  4  2  5  1  2  5
0  1  1  1  0  0  0  0  0  1  0  0  0
  7   0  0  0  0  1  1  6  8  1  8  3  3  3  3  1  0  1  1  0  0  1  0  1
0  0  0  1  1  0  0  0  0  0  0  0  0
  8   0  0  1  1  1 11 38 45 20 42 16 24 14 11  4  7  4  3  2  3  2  1  1
1  1  3  0  0  1  0  0  0  0  0  2  0
  9   0  0  0  0  2  6 52 34 33 36 35 28 39 16 13 12  8  4  3  1  4  3  2
4  4  2  1  2  1  0  0  0  1  0  0  0
  10  3  1  2  2  2 10 29 27 30 25 24 27 28 11  5 12  5  3  6  7  4  0  4
3  1  1  1  2  1  1  0  1  0  0  0  0
  11  0  1  2  1  3  5 43 50 34 27 31 25 26 14 10  7  5 10  4  4  3  4  2
0  6  1  2  2  0  0  1  0  0  0  0  0
  12  0  0  0  0  3  5 21 32 20 18 26 21 28  9  3  8  3  2  3  3  0  6  1
3  2  0  0  0  1  0  0  0  2  0  0  0
  13  0  0  0  0  2  2 30 38 19 30 30 21 30 11  3  5  2  4  3  4  3  6  6
1  1  0  0  3  1  0  0  0  0  2  1  1
  14  0  0  1  0  1  3 17 16 12 15 10 14 13  5  1  4  4  0  1  1  3  0  2
2  0  0  0  1  2  0  0  0  0  0  0  0
  15  0  1  0  4  0  8 47 47 44 37 30 22 31 22 12  9  5  5  2  9  5  2  1
3  0  1  1  0  0  0  1  0  0  0  0  0
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
 [1] "Cluster 1: Within MSE 6298077044284, Size 161"     "Cluster 2: Within MSE 4626688191955561, Siz
 [3] "Cluster 3: Within MSE 1869637869235, Size 68"      "Cluster 4: Within MSE 4643595469908325, Siz
 [5] "Cluster 5: Within MSE 7712486604103, Size 182"     "Cluster 6: Within MSE 7115752554144, Size 
 [7] "Cluster 7: Within MSE 1203006698712, Size 44"      "Cluster 8: Within MSE 4631190560973769, Siz
 [9] "Cluster 9: Within MSE 4619666387278534, Size 346"  "Cluster 10: Within MSE 4460938302148068, S
[11] "Cluster 11: Within MSE 4594040214760925, Size 323" "Cluster 12: Within MSE 4567981210372712, S
```

```
[13] "Cluster 13: Within MSE 4547992746337563, Size 259" "Cluster 14: Within MSE 6523634416458, Size
[15] "Cluster 15: Within MSE 4622864321597182, Size 349"
[1] "Total between cluster MSE: 768416298735565696, Total within cluster MSE: 3553087737252808"
[1] "The K-Means model predicts exactly with an accuracy of 0.1385"
[1] "----------Correlation Checks----------"
[1] "dvage removed, correlated with 41 other variable(s)"
[1] "birthy removed, correlated with 40 other variable(s)"
[1] "scdoby4 removed, correlated with 39 other variable(s)"
[1] "jbstat.x.4 removed, correlated with 33 other variable(s)"
[1] "age_dv removed, correlated with 36 other variable(s)"
[1] "btype4.1 removed, correlated with 27 other variable(s)"
[1] "ff_jbstat.4 removed, correlated with 29 other variable(s)"
[1] "age removed, correlated with 34 other variable(s)"
[1] "confage removed, correlated with 32 other variable(s)"
[1] "btype5.1 removed, correlated with 22 other variable(s)"
[1] "pensioner_dv.2 removed, correlated with 23 other variable(s)"
[1] "sex.2 removed, correlated with 17 other variable(s)"
[1] "natch01.3 removed, correlated with 17 other variable(s)"
[1] "chksex.2 removed, correlated with 16 other variable(s)"
[1] "doby_dv removed, correlated with 29 other variable(s)"
[1] "scsex.2 removed, correlated with 15 other variable(s)"
[1] "ff_bentype18.1 removed, correlated with 18 other variable(s)"
[1] "ficode1.1 removed, correlated with 19 other variable(s)"
[1] "nsex.2 removed, correlated with 14 other variable(s)"
[1] "pn2pno.2 removed, correlated with 15 other variable(s)"
[1] "jbstat.y.4 removed, correlated with 21 other variable(s)"
[1] "pns2pno.2 removed, correlated with 14 other variable(s)"
[1] "nchresp.2 removed, correlated with 12 other variable(s)"
[1] "employ.2 removed, correlated with 12 other variable(s)"
[1] "ff_bentype01.1 removed, correlated with 12 other variable(s)"
[1] "ficode18.1 removed, correlated with 15 other variable(s)"
[1] "height removed, correlated with 11 other variable(s)"
[1] "npn_dv.2 removed, correlated with 13 other variable(s)"
[1] "nchund18resp.2 removed, correlated with 11 other variable(s)"
[1] "estht removed, correlated with 10 other variable(s)"
[1] "sys1 removed, correlated with 10 other variable(s)"
[1] "nch14resp.3 removed, correlated with 9 other variable(s)"
[1] "jbstat.x.2 removed, correlated with 9 other variable(s)"
[1] "relup.6 removed, correlated with 9 other variable(s)"
[1] "weight removed, correlated with 9 other variable(s)"
[1] "mmgsnval removed, correlated with 9 other variable(s)"
[1] "dias1 removed, correlated with 9 other variable(s)"
[1] "nch14resp.1 removed, correlated with 8 other variable(s)"
[1] "nch14resp.2 removed, correlated with 8 other variable(s)"
[1] "nchresp.1 removed, correlated with 9 other variable(s)"
[1] "nchresp.4 removed, correlated with 9 other variable(s)"
[1] "jbhas.2 removed, correlated with 8 other variable(s)"
[1] "marstat.x.2 removed, correlated with 8 other variable(s)"
[1] "respm16.2 removed, correlated with 11 other variable(s)"
[1] "ff_emplw.2 removed, correlated with 9 other variable(s)"
[1] "indscus_xw removed, correlated with 9 other variable(s)"
[1] "estwt removed, correlated with 8 other variable(s)"
[1] "mmgsd1 removed, correlated with 8 other variable(s)"
[1] "map1 removed, correlated with 8 other variable(s)"
[1] "nchresp.3 removed, correlated with 8 other variable(s)"
[1] "pn1pno.1 removed, correlated with 8 other variable(s)"
[1] "pidp removed, correlated with 7 other variable(s)"
[1] "nchund18resp.4 removed, correlated with 8 other variable(s)"
[1] "natch02.4 removed, correlated with 7 other variable(s)"
[1] "hgbiom.2 removed, correlated with 7 other variable(s)"
[1] "scmolwp.2 removed, correlated with 7 other variable(s)"
[1] "agegr13_dv.13 removed, correlated with 8 other variable(s)"
[1] "mmgsn1 removed, correlated with 7 other variable(s)"
[1] "wtval removed, correlated with 7 other variable(s)"
[1] "sys2 removed, correlated with 7 other variable(s)"
[1] "nchund18resp.3 removed, correlated with 7 other variable(s)"
[1] "pns1pno.1 removed, correlated with 7 other variable(s)"
[1] "sppno.2 removed, correlated with 9 other variable(s)"
```

```
[1] "indinus_xw removed, correlated with 7 other variable(s)"
[1] "fimnlabgrs_dv removed, correlated with 8 other variable(s)"
[1] "hidp removed, correlated with 6 other variable(s)"
[1] "nchunder16.2 removed, correlated with 7 other variable(s)"
[1] "nchunder16.4 removed, correlated with 8 other variable(s)"
[1] "marstat.x.5 removed, correlated with 6 other variable(s)"
[1] "mmgsd2 removed, correlated with 6 other variable(s)"
[1] "bmivg5.18 removed, correlated with 6 other variable(s)"
[1] "dias2 removed, correlated with 6 other variable(s)"
[1] "ag16g20.56 removed, correlated with 8 other variable(s)"
[1] "nchunder16.3 removed, correlated with 7 other variable(s)"
[1] "indpxus_lw removed, correlated with 6 other variable(s)"
[1] "bmivg5.30 removed, correlated with 6 other variable(s)"
[1] "pno.2 removed, correlated with 5 other variable(s)"
[1] "psu.x removed, correlated with 5 other variable(s)"
[1] "month removed, correlated with 5 other variable(s)"
[1] "nch415resp.1 removed, correlated with 5 other variable(s)"
[1] "hgbiom.1 removed, correlated with 5 other variable(s)"
[1] "hgbiof.1 removed, correlated with 5 other variable(s)"
[1] "pn1pno.2 removed, correlated with 5 other variable(s)"
[1] "ff_jbstat.2 removed, correlated with 5 other variable(s)"
[1] "ppno.2 removed, correlated with 6 other variable(s)"
[1] "bpmedc.1 removed, correlated with 7 other variable(s)"
[1] "mmgsn2 removed, correlated with 5 other variable(s)"
[1] "map2 removed, correlated with 5 other variable(s)"
[1] "sex_dv.2 removed, correlated with 7 other variable(s)"
[1] "single_dv.1 removed, correlated with 8 other variable(s)"
[1] "natch04.6 removed, correlated with 5 other variable(s)"
[1] "allch04.6 removed, correlated with 7 other variable(s)"
[1] "scsf2a.3 removed, correlated with 5 other variable(s)"
[1] "scsf2b.3 removed, correlated with 6 other variable(s)"
[1] "indinus_lw removed, correlated with 5 other variable(s)"
[1] "bmi removed, correlated with 6 other variable(s)"
[1] "pno.3 removed, correlated with 4 other variable(s)"
[1] "strata.x removed, correlated with 4 other variable(s)"
[1] "nch415resp.2 removed, correlated with 4 other variable(s)"
[1] "natch03.5 removed, correlated with 4 other variable(s)"
[1] "allch01.3 removed, correlated with 4 other variable(s)"
[1] "allch03.5 removed, correlated with 6 other variable(s)"
[1] "jbstat.x.3 removed, correlated with 4 other variable(s)"
[1] "btype3.1 removed, correlated with 4 other variable(s)"
[1] "btype8.1 removed, correlated with 4 other variable(s)"
[1] "marstat.x.6 removed, correlated with 4 other variable(s)"
[1] "hgbiom.3 removed, correlated with 4 other variable(s)"
[1] "pns1pno.2 removed, correlated with 4 other variable(s)"
[1] "fnpno.1 removed, correlated with 4 other variable(s)"
[1] "nnmpsp_dv.1 removed, correlated with 4 other variable(s)"
[1] "medtyp1.1 removed, correlated with 6 other variable(s)"
[1] "sys3 removed, correlated with 4 other variable(s)"
[1] "nnatch.1 removed, correlated with 4 other variable(s)"
[1] "ff_jbstat.8 removed, correlated with 4 other variable(s)"
[1] "ff_bentype19.1 removed, correlated with 4 other variable(s)"
[1] "indinub_xw removed, correlated with 4 other variable(s)"
[1] "indscub_xw removed, correlated with 5 other variable(s)"
[1] "marstat.y.5 removed, correlated with 5 other variable(s)"
[1] "natch01.2 removed, correlated with 3 other variable(s)"
[1] "nnatch.2 removed, correlated with 3 other variable(s)"
[1] "nch5to15.2 removed, correlated with 3 other variable(s)"
[1] "allch02.4 removed, correlated with 4 other variable(s)"
[1] "istrtdaty.2011 removed, correlated with 3 other variable(s)"
[1] "relup.2 removed, correlated with 3 other variable(s)"
[1] "bensta4.1 removed, correlated with 3 other variable(s)"
[1] "bensta7.1 removed, correlated with 3 other variable(s)"
[1] "marstat.x.4 removed, correlated with 3 other variable(s)"
[1] "hgbiof.2 removed, correlated with 3 other variable(s)"
[1] "scsf3a.5 removed, correlated with 4 other variable(s)"
[1] "pn1pno.3 removed, correlated with 3 other variable(s)"
[1] "ff_bentype22.1 removed, correlated with 3 other variable(s)"
```

```
[1] "ff_ukborn.3 removed, correlated with 3 other variable(s)"
[1] "npn_dv.1 removed, correlated with 3 other variable(s)"
[1] "agegr5_dv.6 removed, correlated with 3 other variable(s)"
[1] "agegr5_dv.7 removed, correlated with 3 other variable(s)"
[1] "agegr5_dv.8 removed, correlated with 3 other variable(s)"
[1] "agegr5_dv.9 removed, correlated with 3 other variable(s)"
[1] "agegr5_dv.10 removed, correlated with 3 other variable(s)"
[1] "agegr5_dv.11 removed, correlated with 3 other variable(s)"
[1] "agegr5_dv.13 removed, correlated with 3 other variable(s)"
[1] "agegr5_dv.15 removed, correlated with 3 other variable(s)"
[1] "mastat_dv.6 removed, correlated with 3 other variable(s)"
[1] "buno_dv.3 removed, correlated with 3 other variable(s)"
[1] "nchild_dv.4 removed, correlated with 5 other variable(s)"
[1] "hrpid removed, correlated with 4 other variable(s)"
[1] "ppno.1 removed, correlated with 3 other variable(s)"
[1] "fnspno.1 removed, correlated with 3 other variable(s)"
[1] "mnpno.1 removed, correlated with 4 other variable(s)"
[1] "ficode3.1 removed, correlated with 4 other variable(s)"
[1] "mmgsd3 removed, correlated with 3 other variable(s)"
[1] "dias3 removed, correlated with 3 other variable(s)"
[1] "ag16g10.26 removed, correlated with 5 other variable(s)"
[1] "ag16g10.36 removed, correlated with 5 other variable(s)"
[1] "ag16g10.46 removed, correlated with 5 other variable(s)"
[1] "waist1 removed, correlated with 3 other variable(s)"
[1] "jbstat.y.2 removed, correlated with 3 other variable(s)"
[1] "psu.y removed, correlated with 3 other variable(s)"
[1] "fimnlabnet_dv removed, correlated with 5 other variable(s)"
[1] "nadoptch.2 removed, correlated with 3 other variable(s)"
[1] "nchunder16.1 removed, correlated with 3 other variable(s)"
[1] "heritage2.1 removed, correlated with 3 other variable(s)"
[1] "scsf1.2 removed, correlated with 3 other variable(s)"
[1] "scsf4a.5 removed, correlated with 3 other variable(s)"
[1] "ff_bentype14.1 removed, correlated with 3 other variable(s)"
[1] "nchild_dv.3 removed, correlated with 4 other variable(s)"
[1] "fimnlabgrs_if removed, correlated with 4 other variable(s)"
[1] "ficode19.1 removed, correlated with 3 other variable(s)"
[1] "numed2 removed, correlated with 4 other variable(s)"
[1] "natch02.3 removed, correlated with 2 other variable(s)"
[1] "jbstat.x.6 removed, correlated with 2 other variable(s)"
[1] "jbstat.x.7 removed, correlated with 2 other variable(s)"
[1] "sf1.x.3 removed, correlated with 2 other variable(s)"
[1] "sf1.x.4 removed, correlated with 2 other variable(s)"
[1] "sf1.x.5 removed, correlated with 2 other variable(s)"
[1] "btype1.1 removed, correlated with 2 other variable(s)"
[1] "btype2.1 removed, correlated with 2 other variable(s)"
[1] "marstat.x.3 removed, correlated with 2 other variable(s)"
[1] "scsf2a.2 removed, correlated with 2 other variable(s)"
[1] "scsf3b.5 removed, correlated with 3 other variable(s)"
[1] "scsf7.5 removed, correlated with 3 other variable(s)"
[1] "pns1pno.3 removed, correlated with 2 other variable(s)"
[1] "fimngrs_tc.1 removed, correlated with 2 other variable(s)"
[1] "ff_jbstat.3 removed, correlated with 2 other variable(s)"
[1] "ff_bentype02.1 removed, correlated with 2 other variable(s)"
[1] "ff_bentype23.1 removed, correlated with 2 other variable(s)"
[1] "country.2 removed, correlated with 2 other variable(s)"
[1] "xtra5min_dv.1 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.5 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.12 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.14 removed, correlated with 2 other variable(s)"
[1] "agegr13_dv.12 removed, correlated with 2 other variable(s)"
[1] "cohab_dv.1 removed, correlated with 2 other variable(s)"
[1] "mastat_dv.2 removed, correlated with 2 other variable(s)"
[1] "mastat_dv.4 removed, correlated with 2 other variable(s)"
[1] "sppno.1 removed, correlated with 2 other variable(s)"
[1] "fnpno.2 removed, correlated with 2 other variable(s)"
[1] "hiqual_dv.x.2 removed, correlated with 2 other variable(s)"
[1] "hiqual_dv.x.3 removed, correlated with 2 other variable(s)"
[1] "hiqual_dv.x.4 removed, correlated with 2 other variable(s)"
```

```
[1] "hiqual_dv.x.9 removed, correlated with 2 other variable(s)"
[1] "frmnthimp_dv_total removed, correlated with 2 other variable(s)"
[1] "ficode22.1 removed, correlated with 3 other variable(s)"
[1] "b_pno.3 removed, correlated with 2 other variable(s)"
[1] "bfpc removed, correlated with 2 other variable(s)"
[1] "consubx1.1 removed, correlated with 2 other variable(s)"
[1] "consubx2.1 removed, correlated with 2 other variable(s)"
[1] "consubx5.1 removed, correlated with 3 other variable(s)"
[1] "difbpc1.1 removed, correlated with 2 other variable(s)"
[1] "mmgsn3 removed, correlated with 2 other variable(s)"
[1] "lungeye.2 removed, correlated with 2 other variable(s)"
[1] "bswill.2 removed, correlated with 2 other variable(s)"
[1] "ethnic removed, correlated with 2 other variable(s)"
[1] "pulse1 removed, correlated with 2 other variable(s)"
[1] "map3 removed, correlated with 2 other variable(s)"
[1] "ag16g10.76 removed, correlated with 2 other variable(s)"
[1] "waist2 removed, correlated with 2 other variable(s)"
[1] "hhtype_dv.y removed, correlated with 2 other variable(s)"
[1] "strata.y removed, correlated with 2 other variable(s)"
[1] "ieqmoecd_dv removed, correlated with 3 other variable(s)"
[1] "nnatch.4 removed, correlated with 2 other variable(s)"
[1] "adoptch01.3 removed, correlated with 2 other variable(s)"
[1] "mla3.1 removed, correlated with 2 other variable(s)"
[1] "scsf4b.4 removed, correlated with 2 other variable(s)"
[1] "scsf4b.5 removed, correlated with 2 other variable(s)"
[1] "scsf6c.5 removed, correlated with 2 other variable(s)"
[1] "buno_dv.2 removed, correlated with 2 other variable(s)"
[1] "mnspno.1 removed, correlated with 2 other variable(s)"
[1] "ficode26.1 removed, correlated with 2 other variable(s)"
[1] "b_hidp removed, correlated with 2 other variable(s)"
[1] "nch415resp.3 removed, correlated with 1 other variable(s)"
[1] "nch415resp.4 removed, correlated with 1 other variable(s)"
[1] "nchund18resp.1 removed, correlated with 1 other variable(s)"
[1] "natch01.4 removed, correlated with 1 other variable(s)"
[1] "natch01.5 removed, correlated with 1 other variable(s)"
[1] "natch03.4 removed, correlated with 1 other variable(s)"
[1] "nnatch.3 removed, correlated with 1 other variable(s)"
[1] "nch5to15.1 removed, correlated with 1 other variable(s)"
[1] "allch01.2 removed, correlated with 1 other variable(s)"
[1] "allch01.4 removed, correlated with 1 other variable(s)"
[1] "istrtdaty.2012 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.2 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.3 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.4 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.5 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.6 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.7 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.8 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.9 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.10 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.11 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.12 removed, correlated with 1 other variable(s)"
[1] "istrtdatd removed, correlated with 1 other variable(s)"
[1] "jbstat.x.5 removed, correlated with 1 other variable(s)"
[1] "jbstat.x.8 removed, correlated with 1 other variable(s)"
[1] "jbstat.x.97 removed, correlated with 1 other variable(s)"
[1] "netuse.6 removed, correlated with 1 other variable(s)"
[1] "sf1.x.2 removed, correlated with 1 other variable(s)"
[1] "usdairy.2 removed, correlated with 1 other variable(s)"
[1] "wkfruit.2 removed, correlated with 1 other variable(s)"
[1] "wkfruit.3 removed, correlated with 1 other variable(s)"
[1] "wkvege.3 removed, correlated with 1 other variable(s)"
[1] "health.x.2 removed, correlated with 1 other variable(s)"
[1] "hcondn2.1 removed, correlated with 1 other variable(s)"
[1] "hcondn16.1 removed, correlated with 1 other variable(s)"
[1] "ftedany.2 removed, correlated with 1 other variable(s)"
[1] "btype6.1 removed, correlated with 1 other variable(s)"
[1] "btype9.1 removed, correlated with 1 other variable(s)"
```

```
[1] "bensta2.1 removed, correlated with 1 other variable(s)"
[1] "bensta5.1 removed, correlated with 1 other variable(s)"
[1] "bensta6.1 removed, correlated with 1 other variable(s)"
[1] "bensta8.1 removed, correlated with 1 other variable(s)"
[1] "bensta96.1 removed, correlated with 1 other variable(s)"
[1] "finfut.2 removed, correlated with 1 other variable(s)"
[1] "vote6.2 removed, correlated with 1 other variable(s)"
[1] "arts1a1.1 removed, correlated with 1 other variable(s)"
[1] "arts1a3.1 removed, correlated with 1 other variable(s)"
[1] "arts1b13.1 removed, correlated with 1 other variable(s)"
[1] "arts2a1.1 removed, correlated with 1 other variable(s)"
[1] "arts2b9.1 removed, correlated with 1 other variable(s)"
[1] "mla1.1 removed, correlated with 1 other variable(s)"
[1] "heritage1.1 removed, correlated with 1 other variable(s)"
[1] "heritage3.1 removed, correlated with 1 other variable(s)"
[1] "sports13.1 removed, correlated with 1 other variable(s)"
[1] "sports222.1 removed, correlated with 1 other variable(s)"
[1] "sports34.1 removed, correlated with 1 other variable(s)"
[1] "access.4 removed, correlated with 1 other variable(s)"
[1] "ivcoop.2 removed, correlated with 1 other variable(s)"
[1] "respf16.2 removed, correlated with 1 other variable(s)"
[1] "scsf1.3 removed, correlated with 1 other variable(s)"
[1] "scsf1.4 removed, correlated with 1 other variable(s)"
[1] "scsf1.5 removed, correlated with 1 other variable(s)"
[1] "scsf3a.2 removed, correlated with 1 other variable(s)"
[1] "scsf3a.3 removed, correlated with 1 other variable(s)"
[1] "scsf3a.4 removed, correlated with 1 other variable(s)"
[1] "scsf4a.2 removed, correlated with 1 other variable(s)"
[1] "scsf4a.3 removed, correlated with 1 other variable(s)"
[1] "scsf5.4 removed, correlated with 1 other variable(s)"
[1] "scsf6a.2 removed, correlated with 1 other variable(s)"
[1] "scsf6b.2 removed, correlated with 1 other variable(s)"
[1] "scwhorusex.4 removed, correlated with 1 other variable(s)"
[1] "scrundstnd.2 removed, correlated with 1 other variable(s)"
[1] "scrrely.4 removed, correlated with 1 other variable(s)"
[1] "scrcritic.3 removed, correlated with 1 other variable(s)"
[1] "scrletdwn.3 removed, correlated with 1 other variable(s)"
[1] "scrannoy.3 removed, correlated with 1 other variable(s)"
[1] "scfcritic.3 removed, correlated with 1 other variable(s)"
[1] "scfletdwn.3 removed, correlated with 1 other variable(s)"
[1] "scfannoy.3 removed, correlated with 1 other variable(s)"
[1] "scopfama.5 removed, correlated with 1 other variable(s)"
[1] "scopfamd.2 removed, correlated with 1 other variable(s)"
[1] "fimnlabgrs_tc.1 removed, correlated with 1 other variable(s)"
[1] "j2paynet_dv removed, correlated with 1 other variable(s)"
[1] "ff_jbstat.6 removed, correlated with 1 other variable(s)"
[1] "ff_bentype04.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype05.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype07.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype08.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype09.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype10.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype11.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype12.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype13.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype15.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype20.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype27.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype29.1 removed, correlated with 1 other variable(s)"
[1] "nnssib_dv.1 removed, correlated with 1 other variable(s)"
[1] "fimninvnet_dv removed, correlated with 1 other variable(s)"
[1] "urban_dv.x.2 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.4 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.5 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.6 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.7 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.8 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.9 removed, correlated with 1 other variable(s)"
```

```
[1] "agegr13_dv.10 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.11 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.3 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.5 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.10 removed, correlated with 1 other variable(s)"
[1] "nchild_dv.2 removed, correlated with 1 other variable(s)"
[1] "ppno.3 removed, correlated with 1 other variable(s)"
[1] "fnspno.2 removed, correlated with 1 other variable(s)"
[1] "mnpno.2 removed, correlated with 1 other variable(s)"
[1] "mnpno.3 removed, correlated with 1 other variable(s)"
[1] "hiqual_dv.x.5 removed, correlated with 1 other variable(s)"
[1] "nunmpsp_dv.1 removed, correlated with 1 other variable(s)"
[1] "paygu_if.1 removed, correlated with 1 other variable(s)"
[1] "fibenothr_if removed, correlated with 1 other variable(s)"
[1] "indpxus_xw removed, correlated with 1 other variable(s)"
[1] "ind5mus_lw removed, correlated with 1 other variable(s)"
[1] "ficode2.1 removed, correlated with 1 other variable(s)"
[1] "ficode16.1 removed, correlated with 1 other variable(s)"
[1] "b_pno.2 removed, correlated with 1 other variable(s)"
[1] "region.2 removed, correlated with 1 other variable(s)"
[1] "medcnjd.2 removed, correlated with 1 other variable(s)"
[1] "floorc.2 removed, correlated with 1 other variable(s)"
[1] "cufsize.2 removed, correlated with 1 other variable(s)"
[1] "lungex.2 removed, correlated with 1 other variable(s)"
[1] "nseqno.3 removed, correlated with 1 other variable(s)"
[1] "nuroutc.12 removed, correlated with 1 other variable(s)"
[1] "nurdayy.2011 removed, correlated with 1 other variable(s)"
[1] "pulse2 removed, correlated with 1 other variable(s)"
[1] "omsyst removed, correlated with 1 other variable(s)"
[1] "ag16g10.56 removed, correlated with 1 other variable(s)"
[1] "ag16g10.66 removed, correlated with 1 other variable(s)"
[1] "ag16g20.36 removed, correlated with 1 other variable(s)"
[1] "ag16g20.76 removed, correlated with 1 other variable(s)"
[1] "hhsize.2 removed, correlated with 1 other variable(s)"
[1] "marstat.y.2 removed, correlated with 1 other variable(s)"
[1] "marstat.y.4 removed, correlated with 1 other variable(s)"
[1] "marstat.y.6 removed, correlated with 1 other variable(s)"
[1] "indnsub_lw removed, correlated with 1 other variable(s)"
[1] "fimnnet_dv removed, correlated with 1 other variable(s)"
[1] "respm16_dv.2 removed, correlated with 1 other variable(s)"
[1] "qfhigh_dv.3 removed, correlated with 1 other variable(s)"
[1] "qfhigh_dv.7 removed, correlated with 1 other variable(s)"
[1] "qfhigh_dv.13 removed, correlated with 1 other variable(s)"
[1] "qfhigh_dv.96 removed, correlated with 1 other variable(s)"
[1] "bfpcval removed, correlated with 1 other variable(s)"
[1] "htval removed, correlated with 1 other variable(s)"
[1] "fibenothr_dv removed, correlated with 1 other variable(s)"
[1] "nmpsp_dv.1 removed, correlated with 1 other variable(s)"
[1] "wstval removed, correlated with 1 other variable(s)"
[1] "ethn_dv removed, correlated with 1 other variable(s)"
[1] "385 variables removed since they had high correlation coefs"
[1] "The final cleaned dataset has been completed at this stage and is stored under the name 'cleanDa
[1] "——————Attempting a Train Test Split——————"
[1] "Good train, test split found"
[1] "The working seed found was 3"
[1] "——————kNN——————"
[1] "105 neighbours considered for each test data point"
[1] "kNN results as a table, follow the diagonal for the correctly mapped clusters"
         real
predicted  0  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 33 3
        0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0
        1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0
        2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0
        3  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0
```

```
        4   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0
        5   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0
        6   2  2  1  6 13 56 49 30 44 30 26 20 15  5  5  3  1  2  2  2  2
1  0  0  0  0  0  1  0  0  0  0
        7   1  0  2  6  5 38 57 40 41 42 43 35 13 11  8  5  5  2  5  6  3
3  4  1  1  1  0  0  0  0  1  0
        8   0  0  0  1  0  1  0  1  1  2  2  2  2  1  1  0  0  0  0  0  1
0  0  0  0  0  0  0  0  0  0  0
        9   0  1  0  0  0  5  3  3  5  7  6  6  3  6  1  1  2  2  2  2  2
1  2  2  1  0  0  0  1  0  0  0
        10  0  0  0  0  0  1  1  1  1  2  1  4  2  1  0  0  1  0  0  0  0
0  0  0  1  0  0  0  0  0  0  0
        11  0  0  0  0  0  0  0  2  0  0  1  0  1  1  1  1  2  0  0  2  1
3  1  0  1  0  2  0  0  1  0  1
        12  0  0  0  0  0  0  0  1  1  0  3  6  0  1  3  1  2  1  1  2  1
2  1  2  1  1  1  1  0  0  1  0
        13  0  0  0  0  0  0  0  0  0  1  0  0  1  0  0  0  0  0  1  1  1
0  0  0  0  0  0  0  0  0  0  0
        14  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  1  0
        15  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0
        16  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0
        17  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0
        18  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0
        19  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0
        20  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0
        21  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0
        22  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0
        23  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0
        24  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0
        25  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0
        26  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0
        27  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0
        28  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0
        29  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0
        30  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0
 [ reached getOption("max.print") -- omitted 4 rows ]
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
[1] "The MSE of the predicted values are of 266.0141"
[1] "The kNN model predicts exactly with an accuracy of 0.1514"
[1] "------------CART prediction model------------"
n= 2555

node), split, n, deviance, yval
      * denotes terminal node

 1) root 2555 60203.100 10.567510
   2) sf12mcs_dv>=-0.4090849 1862 13718.890  8.668099
     4) sf12mcs_dv>=0.3460245 1128  6057.531  7.770390
       8) scsf6c.4< 0.5 973  3857.848  7.425488 *
```

```
        9) scsf6c.4>=0.5 155  1357.355  9.935484 *
      5) sf12mcs_dv< 0.3460245 734  5355.331 10.047680 *
    3) sf12mcs_dv< -0.4090849 693 21716.990 15.671000
      6) sf12mcs_dv>=-1.478103 458  7811.967 13.731440
       12) sf12pcs_dv>=-1.201603 382  5389.277 13.167540 *
       13) sf12pcs_dv< -1.201603 76  1690.671 16.565790 *
      7) sf12mcs_dv< -1.478103 235  8824.187 19.451060
       14) sf12mcs_dv>=-2.64619 185  4723.784 17.918920 *
       15) sf12mcs_dv< -2.64619 50  2059.280 25.120000 *
[1] "Variable Importance"
  sf12mcs_dv      scsf6c.3      scsf6a.4      scsf4b.3       scsf7.3      scsf6b.4
sf12pcs_dv      scsf6c.4      scsf6c.2
34195.210594  8940.273289  5851.529009  5465.014958  4967.741987  4217.219817
2284.831767  1929.363652  1537.331334
    scsf6a.3      scsf4b.2      scsf4a.4      scsf3b.2       sf1.y.5    schmcont.6
scsf6a.5       scsf5.5  sports221.1
  760.296404   605.375791   317.313789   231.163803   173.372852   163.289876
163.289876   130.648103   122.467407
  jbstat.y.8      scsf6b.5    hcondn11.1    lenindintv    sclfsato.3
  115.581901     96.318251     16.303128     16.303128      5.434376
[1] "The MSE of the predicted values are of 10.3192"
[1] "The CART model predicts exactly with accuracy of 0.1667"
[1] "--------------Ordinary Linear Regression (Initial)--------------"
[1] "The full model AIC is: 12953.6644"
[1] "--------------Variance Inflation Factor Removal--------------"
[1] "The variable ienddathh was removed since it had a VIF score of 866922695.8729"
[1] "The variable livesp_dv.1 was removed since it had a VIF score of 192.6294"
[1] "The variable istrtdathh.10 was removed since it had a VIF score of 86.6685"
[1] "The variable sclfsato.6 was removed since it had a VIF score of 46.8376"
[1] "The variable sclfsat1.6 was removed since it had a VIF score of 23.5633"
[1] "The variable sf12mcs_dv was removed since it had a VIF score of 21.7407"
[1] "The variable sclfsat7.6 was removed since it had a VIF score of 16.1199"
[1] "The variable agegr10_dv.5 was removed since it had a VIF score of 13.4889"
[1] "The variable scopfamf.4 was removed since it had a VIF score of 12.1611"
[1] "The variable sclfsat2.6 was removed since it had a VIF score of 12.0474"
[1] "The variable rach16_dv.2 was removed since it had a VIF score of 11.3099"
[1] "11 variables removed from the Ordinary Linear Model since they have a VIF score higher than 10"
[1] "The full model AIC after VIF checks is: 13147.5186"
[1] "--------------Backwards Selection--------------"
[1] "50 out of 595 variables removed so far."
[1] "100 out of 595 variables removed so far."
[1] "150 out of 595 variables removed so far."
[1] "200 out of 595 variables removed so far."
[1] "250 out of 595 variables removed so far."
[1] "300 out of 595 variables removed so far."
[1] "350 out of 595 variables removed so far."
[1] "400 out of 595 variables removed so far."
[1] "450 out of 595 variables removed so far."
[1] "500 out of 595 variables removed so far."
[1] "512 out of 595 variables removed in backwards selection since they weren't significant at the 95
  [1] "fimngrs_dv"        "ficode28.1"        "scanyelsetxt.2"    "arts1b14.1"
"event1s"         "btype7.1"
  [7] "arts1a4.1"         "mla96.1"           "scopfama.2"        "hhsize.3"
"intdaty_dv.2011"  "npensioner_dv.3"
 [13] "sports224.1"       "scrannoy.4"        "ag16g10.86"        "kidlang.8"
"scopfamd.5"      "bmivg5.25"
 [19] "susp.2"            "arts2a5.1"         "qfhigh_dv.11"      "heritage5.1"
"lenindintv"      "sports231.1"
 [25] "nurdaym.4"         "difbpc3.1"         "strtnurhh"         "arts1b12.1"
"scopfamf.3"      "fimngrs_if"
 [31] "fiyrinvinc_if.1"  "scrletdwn.2"       "X.next.1"          "wjrel.2"
"ficode39.1"      "scrundstnd.4"
 [37] "istrtdathh.15"     "istrtdathh.14"     "arts2a4.1"         "sports37.1"
"gor_dv.3"        "intdaty_dv.2012"
 [43] "heritage8.1"       "sclfsat7.7"        "sclfsat7.2"        "sports196.1"
"sports11.1"      "mmgsdval"
 [49] "hcondn11.1"        "sports112.1"       "ficode14.1"        "ff_jbstat.97"
"qfhigh_dv.5"     "btype96.1"
```

```
 [55] "ficode4.2"          "indscus_lw"         "medtyp6.1"          "event3s"
"nmpsp_dv.2"          "omdiast"
 [61] "scsf2b.2"           "scopfamf.5"         "hhsize.5"           "natch02.5"
"scwhorufam.4"        "arts1a7.1"
 [67] "mmgstp.4"           "nurdaym.6"          "nurdaym.10"         "hhresp_dv.3"
"arts2b10.1"          "scwhorufam.3"
 [73] "ff_bentype33.1"     "lfout.9"            "strtnurmm"          "kidlang.97"
"indnsub_xw"          "fimnmisc_dv"
 [79] "ficode13.1"         "ff_ukborn.2"        "vote6.3"            "sclfsat2.7"
"npns_dv.2"           "urban_dv.y.2"
 [85] "scfalcdrnk.6"       "scssupr2r.3"        "j2pay_if.1"         "hhtype_dv.x"
"sports223.1"         "medtyp9.1"
 [91] "lfout.4"            "qfhigh_dv.2"        "scloutcont.6"       "scloutcont.2"
"npns_dv.1"           "ff_ukborn.4"
 [97] "sclfsat7.5"         "hiqual_dv.y.5"      "scfundstnd.4"       "fruvege.2"
"consubx3.1"          "natch02.6"
[103] "bprespc.2"          "fruvege.9"          "istrtdatss"         "hcondn1.1"
"arts1b10.1"          "sportact.7"
[109] "sportact.8"         "sclfsat2.4"         "sclfsat1.4"         "sclfsat1.5"
"sclfsat1.7"          "scloutcont.3"
[115] "arts1b96.1"         "qfhigh_dv.4"        "scsf3b.4"           "sports36.1"
"arts2a6.1"           "scssupr2r.9"
[121] "scssupr2r.2"        "scssup1.2"          "sports31.1"         "sclfsat1.2"
"allch02.5"           "ficode4.3"
[127] "arts1a5.1"          "ficode3.2"          "ficode4.1"          "jbstat.y.97"
"hcondn4.1"           "omronno"
[133] "netuse.5"           "scopfamh.4"         "sports217.1"        "fruvege.3"
"fruvege.8"           "sctimemnuf.3"
[139] "sctimemnuf.4"       "calciumb.1"         "arts2b13.1"         "scage1drink"
"usdairy.3"           "j2pay_dv"
[145] "qfhigh_dv.15"       "ff_bentype34.1"     "lfout.3"            "sports221.1"
"scrletdwn.4"         "ficode5.1"
[151] "allch01.5"          "scsf3b.2"           "scsf3b.3"           "sf1.y.4"
"intdatm_dv.3"        "nurdaym.8"
[157] "scfopenup.2"        "marstat.y.3"        "ficode9.1"          "qfhigh_dv.10"
"wlk10m.2"            "sportact.3"
[163] "arts1a2.1"          "nurdayw.6"          "nurdayw.1"          "nurdayw.2"
"nurdayw.5"           "lfout.5"
[169] "gor_dv.5"           "vote1.2"            "access.5"           "mobuse.2"
"hcondn96.1"          "airtemp"
[175] "access.2"           "ficode12.1"         "ficode29.1"         "fruvege.6"
"relup.5"             "ivcoop.3"
[181] "sportact.9"         "ficode10.1"         "wkvege.4"           "access.6"
"scfcritic.4"         "sports216.1"
[187] "chkcoa.2"           "jbstat.y.6"         "sppno.3"            "bswill.3"
"nnsib_dv.1"          "sports14.1"
[193] "relup.4"            "sf1.y.3"            "scrundstnd.3"       "volun.2"
"intdatm_dv.11"       "agegr10_dv.3"
[199] "mmgstp.3"           "medtyp5.1"          "scfalcdrnk.3"       "scalcl7d.2"
"sports218.1"         "hcondn12.1"
[205] "mmgsres.2"          "beta.1"             "lipid.1"            "arts2a2.1"
"qfhigh_dv.16"        "bensta1.1"
[211] "arts2b11.1"         "medtyp8.1"          "j2has.2"            "ficode29.3"
"chargv.2"            "sctimemnuf.2"
[217] "allch02.6"          "natch03.6"          "sclfsat7.3"         "nurdayd"
"mnspno.3"            "scssup1.4"
[223] "fimnsben_dv"        "ficode20.1"         "ff_jbstat.7"        "bmival"
"marstat_dv.4"        "mnspno.2"
[229] "xtra5minosm_dv.1"   "relwaitb.2"         "save.2"             "pulse3"
"sports33.1"          "hcondn14.1"
[235] "gor_dv.10"          "gor_dv.8"           "gor_dv.6"           "gor_dv.2"
"adoptch02.4"         "nchild_dv.1"
[241] "kidlang.11"         "arts2a96.1"         "intdatm_dv.2"       "paynu_if.1"
"sports16.1"          "sports229.1"
[247] "usdairy.5"          "hcondn6.1"          "mmgstp.2"           "ff_bentype17.1"
"marstat_dv.2"        "scopfama.4"
[253] "scwhoruage.4"       "arts1b9.1"          "sports230.1"        "ienddatss"
"aceinh.1"            "usbread.7"
```

```
[259] "sclfsat2.5"        "finnow.3"           "fimnpen_dv"         "finnow.2"
"scwhorusex.2"      "scwhorusex.3"
[265] "arts2b14.1"        "arts2a7.1"          "nch10to15.2"        "istrtdathh.16"
"netuse.7"          "netuse.2"
[271] "netuse.3"          "relhite.2"          "hcondn5.1"          "vote6.4"
"fimnprben_dv"      "ienddatmm"
[277] "istrtdathh.11"     "ficode29.2"         "netuse.4"           "event1"
"sports228.1"       "arts1b15.1"
[283] "fruvege.10"        "ff_bentype03.1"     "istrtdathh.13"      "marstat_dv.3"
"ff_oprlg.2"        "sports227.1"
[289] "hiqual_dv.y.2"     "aidxhh.2"           "nch10to15.3"        "hhsize.6"
"allch02.3"         "nch10to15.1"
[295] "nurdayw.3"         "istrtdathh.19"      "sports111.1"        "sclfsat7.4"
"wkfruit.4"         "scopfamh.5"
[301] "difbpc5.1"         "arts2b12.1"         "sports12.1"         "event4s"
"istrtdathh.21"     "ppen.2"
[307] "difbpc2.1"         "hcondn10.1"         "istrtdathh.18"      "istrtdathh.12"
"scfalcdrnk.9"      "intdatd_dv"
[313] "medtyp4.1"         "sports219.1"        "sports110.1"        "consubx4.1"
"jbstat.y.5"        "sports19.1"
[319] "access.3"          "hcondn13.1"         "gor_dv.4"           "arts2a3.1"
"sports18.1"        "mmgsdom.2"
[325] "sclfsat2.2"        "medtyp11.1"         "diur.1"             "scfalcdrnk.7"
"nurdayy.2012"      "nurdaym.11"
[331] "intdatm_dv.7"      "heritage6.1"        "qfhigh_dv.14"       "scopfamd.4"
"medtyp3.1"         "mla2.1"
[337] "scfundstnd.2"      "scfrely.2"          "sports39.1"         "nch5to15.3"
"ff_bentype16.1"    "scfannoy.4"
[343] "finfut.3"          "arts2b15.1"         "advvoucher.2"       "nurdayw.4"
"scfopenup.4"       "sportact.1"
[349] "sportact.10"       "ficode24.1"         "sportact.4"         "istrtdatmm"
"cindtime"          "heritage4.1"
[355] "ff_bentype28.1"    "wlk10m.3"           "istrtdathh.17"      "ficode34.1"
"sports114.1"       "nurdaym.3"
[361] "nurdaym.7"         "kidlang.12"         "origadd.2"          "medtyp2.1"
"medcnj.2"          "heritage96.1"
[367] "agegr10_dv.4"      "scfletdwn.4"        "scfletdwn.2"        "wjrel.3"
"intdatm_dv.4"      "intdatm_dv.5"
[373] "fruvege.7"         "bsoute.3"           "sclfsat2.3"         "sports220.1"
"usdairy.4"         "fruvege.4"
[379] "wkvege.2"          "scopfamd.3"         "event2s"            "scwkvfast.4"
"scwkvfast.2"       "smever.2"
[385] "agegr10_dv.7"      "scopfamb.4"         "scopfamb.5"         "sports225.1"
"undqus.2"          "cufsize.3"
[391] "difbpc6.1"         "ficode17.1"         "arts1b11.1"         "scfannoy.2"
"scwhoruage.2"      "kidlang.2"
[397] "natch01.1"         "ficode7.1"          "usbread.6"          "sports15.1"
"scssupr2r.8"       "sports226.1"
[403] "sports296.1"       "scopfamh.3"         "heritage7.1"        "hrpno.2"
"marstat_dv.6"      "scdem2many.2"
[409] "seearngrs_if.1"    "undqus.3"           "sports113.1"        "sports35.1"
"scopfamb.3"        "scopfama.3"
[415] "fimnlabnet_tc.1"   "nadoptch.1"         "lfout.2"            "ficode2.2"
"npensioner_dv.2"   "npensioner_dv.1"
[421] "medtyp12.1"        "sports396.1"        "arts1a96.1"         "jbiindb_dv"
"scwkvfast.5"       "scwkvfast.6"
[427] "scwkvfast.3"       "kidlang.6"          "ivprsnt.2"          "floorc.3"
"trainany.2"        "sf1.y.5"
[433] "allch03.4"         "relup.3"            "usdairy.6"          "rhland_code.1"
"scfcritic.2"       "medtyp10.1"
[439] "sclfsat1.3"        "sf1.y.2"            "nch5to15.4"         "nch14resp.4"
"jbstat.y.7"        "indpxub_xw"
[445] "marstat_dv.5"      "sclfsato.5"         "scfalcdrnk.4"       "scssupr2r.4"
"scwhorufam.2"      "scssup1.5"
[451] "ficode2.3"         "adoptch01.4"        "usbread.4"          "usbread.5"
"usbread.3"         "ff_bentype24.1"
[457] "scrannoy.2"        "scopfamh.2"         "health.y.2"         "bsoute.2"
"howlng"            "lfout.8"
```

```
[463] "lfout.11"          "intdatm_dv.8"      "intdatm_dv.10"     "nurdaym.2"
"bmivg5.40"         "hcondn7.1"
[469] "sportact.5"        "qfhigh_dv.12"      "scdem2many.3"      "hcondn8.1"
"nnewborn.1"        "intdatm_dv.9"
[475] "scwhoruage.3"      "sports17.1"        "scfopenup.3"       "bsoute.5"
"scopfamb.2"        "nurdaym.5"
[481] "istrtdathh.20"     "sctimemnuf.5"      "nseqno.2"          "arts1a6.1"
"agegr10_dv.6"      "xpmove.2"
[487] "lkmove.2"          "nurdaym.12"        "nurdaym.9"         "hrpno.3"
"istrtdathh.9"      "iron.1"
[493] "scfrely.3"         "difbpc95.1"        "ficode38.1"        "sportact.6"
"scfalcdrnk.5"      "ficode37.1"
[499] "statins.2"         "ff_ukborn.5"       "kidlang.4"         "hiqual_dv.y.4"
"hiqual_dv.y.3"     "hiqual_dv.y.9"
[505] "ficode23.1"        "ficode15.1"        "ficode33.1"        "allch03.6"
"scfundstnd.3"      "sportact.2"
[511] "scfrely.4"         "fruvege.5"
[1] "————————Ordinary Linear Regression (Improved)————————"

Call:
lm(formula = y ~ ., data = as.data.frame(x.data.linear))

Residuals:
     Min       1Q   Median       3Q      Max
-10.1248  -1.5998  -0.1075   1.4254  13.4173

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.33427    0.27805  22.781  < 2e-16 ***
kidlang.5      2.32327    0.97594   2.381 0.017362 *
kidlang.7      1.88537    0.95289   1.979 0.047973 *
usbread.2      0.29340    0.12653   2.319 0.020488 *
hcondn17.1     1.75130    0.45893   3.816 0.000139 ***
finnow.4       0.58579    0.23892   2.452 0.014280 *
finnow.5       1.74541    0.39488   4.420 1.03e-05 ***
arts2b96.1    -0.32941    0.12118  -2.718 0.006605 **
sports32.1    -0.92751    0.21833  -4.248 2.24e-05 ***
sports38.1     1.17506    0.41727   2.816 0.004900 **
scsf4a.4       0.46142    0.16160   2.855 0.004335 **
scsf4b.2       3.02086    0.47395   6.374 2.19e-10 ***
scsf4b.3       1.75571    0.23384   7.508 8.32e-14 ***
scsf5.2        0.44624    0.13825   3.228 0.001264 **
scsf5.3        0.91956    0.24815   3.706 0.000215 ***
scsf5.5        1.49470    0.51429   2.906 0.003690 **
scsf6a.3       0.96184    0.15496   6.207 6.30e-10 ***
scsf6a.4       2.02687    0.25466   7.959 2.61e-15 ***
scsf6a.5       4.23850    0.52616   8.056 1.22e-15 ***
scsf6b.3       0.36666    0.13984   2.622 0.008796 **
scsf6b.4       1.25047    0.22868   5.468 5.00e-08 ***
scsf6b.5       2.66648    0.39349   6.777 1.53e-11 ***
scsf6c.2       4.98225    0.39226  12.701  < 2e-16 ***
scsf6c.3       3.06474    0.19723  15.539  < 2e-16 ***
scsf6c.4       1.62963    0.14179  11.494  < 2e-16 ***
scsf7.2        3.69377    0.41743   8.849  < 2e-16 ***
scsf7.3        1.45103    0.21953   6.610 4.70e-11 ***
scsf7.4        0.96989    0.17099   5.672 1.57e-08 ***
scfalcdrnk.2   0.56620    0.21428   2.642 0.008285 **
sclfsato.2     1.87807    0.33648   5.582 2.64e-08 ***
sclfsato.3     1.89406    0.24203   7.826 7.42e-15 ***
sclfsato.4     0.98849    0.22865   4.323 1.60e-05 ***
sclfsato.7    -0.64342    0.17814  -3.612 0.000310 ***
schmcont.2     0.58132    0.12802   4.541 5.87e-06 ***
schmcont.3     1.49825    0.23845   6.283 3.90e-10 ***
schmcont.4     1.44390    0.32547   4.436 9.55e-06 ***
schmcont.5     1.82815    0.50995   3.585 0.000344 ***
schmcont.6     3.50821    0.63000   5.569 2.84e-08 ***
scloutcont.4  -0.44169    0.19619  -2.251 0.024452 *
scloutcont.5  -0.49911    0.17548  -2.844 0.004489 **
```

```
scdem2many.4     -0.43423    0.16655   -2.607 0.009185 **
scdem2many.5     -0.37012    0.17308   -2.138 0.032575 *
scdem2many.6     -0.55573    0.20197   -2.752 0.005974 **
sctimemnuf.6      0.82051    0.26754    3.067 0.002187 **
scrrely.2        -0.41673    0.15060   -2.767 0.005697 **
scrrely.3        -0.55955    0.21264   -2.631 0.008556 **
scropenup.2       0.44225    0.13902    3.181 0.001485 **
scropenup.3       0.54480    0.18370    2.966 0.003049 **
scropenup.4       1.05133    0.26022    4.040 5.50e-05 ***
scrcritic.2      -0.50654    0.19277   -2.628 0.008650 **
scrcritic.4      -0.25707    0.12161   -2.114 0.034617 *
scssupr2r.7       1.98772    0.83888    2.369 0.017889 *
scopfamf.2        0.37220    0.18737    1.986 0.047098 *
ff_jbstat.5      -1.30718    0.54523   -2.397 0.016582 *
ff_bentype26.1    0.99175    0.33944    2.922 0.003512 **
respf16_dv.2      0.62052    0.20327    3.053 0.002292 **
nunmpsp_dv.2     -4.39870    1.22399   -3.594 0.000332 ***
ficode8.1        -3.08851    1.10881   -2.785 0.005386 **
ficode11.1       -1.95461    0.77482   -2.523 0.011710 *
ficode27.1       -1.26075    0.51954   -2.427 0.015309 *
obpdrug.1        -1.23293    0.39352   -3.133 0.001750 **
medtyp7.1        -0.74072    0.21162   -3.500 0.000473 ***
hhsize.4          0.32472    0.14858    2.186 0.028943 *
jbstat.y.3        0.78865    0.33385    2.362 0.018239 *
jbstat.y.8        2.19623    0.44395    4.947 8.04e-07 ***
fiyrinvinc_dv    -0.11397    0.05261   -2.166 0.030374 *
intdatm_dv.6     -0.43921    0.19554   -2.246 0.024784 *
intdatm_dv.12     0.78189    0.23461    3.333 0.000872 ***
hhresp_dv.2      -0.44509    0.20824   -2.137 0.032662 *
racel_dv         -0.16264    0.05117   -3.178 0.001499 **
gor_dv.7          1.03627    0.23218    4.463 8.43e-06 ***
gor_dv.9          0.36533    0.16853    2.168 0.030278 *
sf12pcs_dv        0.41689    0.08276    5.037 5.06e-07 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1    1

Residual standard error: 2.819 on 2482 degrees of freedom
Multiple R-squared:  0.6724,    Adjusted R-squared:  0.6629
F-statistic: 70.75 on 72 and 2482 DF,  p-value: < 2.2e-16

AIC: 12620.4516
MSE: 7.7192
[1] "The MSE of the predicted values are of 10.1981"
[1] "The Linear Model predicts exactly with accuracy of 0.1714"
[1] "The Linear Model predicts within a confidence interval with accuracy of 0.4049"
[1] "----------Elastic Net Regression----------"
597 x 1 sparse Matrix of class "dgCMatrix", with 23 entries
           names Estimate_Coefs
1    (Intercept)   10.424436813
2       finnow.4    0.010021598
3       finnow.5    0.111482306
4        scsf5.3    0.084779191
5       scsf6c.2    0.452574688
6       scsf6c.3    0.046343931
7        scsf7.2    0.466253773
8     sclfsat1.3    0.123493530
9     sclfsato.2    1.282333043
10    sclfsato.3    1.518937012
11    sclfsato.4    0.150233845
12    sclfsato.6   -0.151810797
13    sclfsato.7   -0.484736501
14     schmcont.3   0.356700024
15     schmcont.4   0.125984355
16     schmcont.6   0.067485500
17  scloutcont.5   -0.005755865
18  sctimemnuf.6    0.332769046
19    sf12mcs_dv   -3.193887636
20    jbstat.y.8    0.012169014
```

```
21       sf1.y.5     0.186353448
22    health.y.2    -0.021979541
23    sf12pcs_dv    -0.629906160
[1] "The MSE of the predicted values of the best fit model is 8.6719"
[1] "The Alpha of the best fit model is 1"
[1] "The Elastic Net Model predicts exactly with accuracy of 0.1808"
[1] "------------Timer Results------------"
   user  system elapsed
 538.69    4.38  543.56
```

10.2.19 w2MergeNurseBlood console

```
[1] "------------Initial Checks------------"
[1] "8222559 NA cells were found across the entire dataset (57.86% of data as NA)"
[1] "------------Data Type Checks------------"
[1] "0 variables recoded since all their entries aren't numeric or NA"
[1] "NOTE: algorithim recodes categorical data alphabetically e.g (female = 1, male = 2)"
character(0)
[1] "------------Low Data Removal------------"
[1] "1361 variables removed since they had >= 'naPercent' (default 20%) NA values"
   [1] "pid.x"          "lvwhy"          "lvmthp"         "lvyrp"          "mvever"
"mvmnth"         "mvyr"           "mlstatchk"
   [9] "mlstat.x"       "drive"          "caruse"         "ukborn"         "plbornc"
"yr2uk4"         "citzn1"         "citzn2"
  [17] "citzn3"         "qfhigh"         "qualoc"         "qfvoc1"         "qfvoc2"
"qfvoc3"         "qfvoc4"         "qfvoc5"
  [25] "qfvoc6"         "qfvoc7"         "qfvoc8"         "qfvoc9"         "qfvoc10"
"qfvoc11"        "qfvoc12"        "qfvoc13"
  [33] "qfvoc14"        "qfvoc15"        "qfvoc96"        "school"         "scend"
"schlloc"        "schok"          "fenow"
  [41] "feend"          "j1none"         "j1semp"         "j1boss"         "j1mngr"
"paedqf"         "maedqf"         "edtype"
  [49] "edasp"          "fedlik"         "fednt"          "ocimpa"         "ocimpb"
"ocimpe"         "ocimpf"         "ocimpi"
  [57] "ocimpk"         "ocimpl"         "futra"          "futrb"          "futrc"
"futrd"          "futre"          "futrf"
  [65] "futrg"          "futrh"          "futri"          "futrj"          "futrk"
"futrl"          "paju"           "maju"
  [73] "pacob"          "payruk"         "macob"          "mayruk"         "natid1"
"natid2"         "natid3"         "natid4"
  [81] "natid5"         "natid6"         "natid97"        "racel"          "racelo_code"
"ethid2"         "ethid3"         "ethid4a"
  [89] "ethid4b"        "ethid5"         "ethid6"         "ethid7"         "ethid8"
"ethid9"         "ethid10"        "ethid11"
  [97] "ethid12"        "ethid13"        "ethid14"        "ethclose1"      "ethclose2a"
"ethclose2b"     "ethclose3"      "ethclose4a"
 [105] "ethclose4b"     "ethclose5"      "ethclose6"      "ethclose7"      "ethclose8"
"ethclose9"      "ethclose10"     "ethclose11"
 [113] "pride2"         "pride4a"        "pride4b"        "pride6"         "pride7"
"pride5"         "pride8"         "pride9"
 [121] "pride10"        "pride11"        "pride12"        "pride13"        "pride14"
"food1"          "food2"          "food3"
 [129] "food4"          "food5"          "food6"          "food7"          "oprlg"
"oprlg0ni"       "nirel"          "niact"
 [137] "oprlg0"         "oprlg1"         "daywlk"         "wlk30min"       "walkpace"
"smnow"          "ncigs"          "smcigs"
 [145] "smncigs"        "aglquit"        "smagbg"         "hospc1"         "hospdc1"
"hospc2"         "hospdc2"        "hospc3"
 [153] "hospdc3"        "hospc4"         "hospdc4"        "hospc5"         "hospdc5"
"hospc6"         "hospdc6"        "hospc7"
 [161] "hospdc7"        "hospc8"         "hospdc8"        "disdif1"        "disdif2"
"disdif3"        "disdif4"        "disdif5"
 [169] "disdif6"        "disdif7"        "disdif8"        "disdif9"        "disdif10"
"disdif11"       "disdif12"       "disdif96"
 [177] "aidhh"          "aidhua1"        "aidhua2"        "aidhua3"        "aidhua4"
"aidhua5"        "aidhua6"        "aidhua7"
```

```
 [185] "aidhua8"       "aidhua9"       "aidhua10"      "aidhua11"      "aidhua12"
"aidhua13"      "aidhua14"      "aidhua15"
 [193] "aidhua16"      "naidxhh"       "aidhu1"        "aidhu2"        "aidhrs"
"aideft"        "lcohnpi"       "coh1bm"
 [201] "coh1by"        "coh1mr"        "coh1em"        "coh1ey"        "nmar"
"lmar1m"        "lmar1y"        "ladopt"
 [209] "lnadopt"       "lprnt"         "lnprnt"        "ch1by4"        "movy11"
"movy12"        "movy13"        "movy14"
 [217] "movy15"        "movy16"        "family"        "education"     "memploy"
"housing"       "area"          "moveoth_code"
 [225] "movdir"        "plnowm"        "plnowy4"       "mstatsam"      "lwwrong"
"mstatsamn"     "lastmstatch1"  "mstatch1"
 [233] "statcm1"       "statcy41"      "divchk1"       "divfin1"       "dvm1"
"dvy41"         "cmlstat1"      "lastmstatch2"
 [241] "mstatch2"      "statcm2"       "statcy42"      "divchk2"       "divfin2"
"dvm2"          "dvy42"         "cmlstat2"
 [249] "lastmstatch3"  "mstatch3"      "statcm3"       "statcy43"      "divchk3"
"divfin3"       "dvm3"          "dvy43"
 [257] "cmlstat3"      "cohab"         "cohabn"        "lmcbm1"        "lmcby41"
"currpart1"     "lmspm1"        "lmspy41"
 [265] "lmcbm2"        "lmcby42"       "currpart2"     "lmspm2"        "lmspy42"
"lmcbm3"        "lmcby43"       "currpart3"
 [273] "lmspm3"        "lmspy43"       "lmcbm4"        "lmcby44"       "currpart4"
"lmspm4"        "lmspy44"       "father"
 [281] "nchild"        "preg"          "pregm1"        "pregy41"       "pregfert1"
"invitro1"      "pregout1"      "pregend1"
 [289] "endmnth1"      "pregsmoke1"    "smkmnth11"     "smkmnth21"     "smkmnth31"
"pregsmk11"     "pregsmk21"     "pregsmk31"
 [297] "pregdrink1"    "lchmulti1"     "pregm2"        "pregy42"       "pregfert2"
"invitro2"      "pregout2"      "pregend2"
 [305] "endmnth2"      "pregsmoke2"    "smkmnth12"     "smkmnth22"     "smkmnth32"
"pregsmk12"     "pregsmk22"     "pregsmk32"
 [313] "pregdrink2"    "lchmulti2"     "pregm3"        "pregy43"       "pregfert3"
"invitro3"      "pregout3"      "pregend3"
 [321] "endmnth3"      "pregsmoke3"    "smkmnth13"     "smkmnth23"     "smkmnth33"
"pregsmk13"     "pregsmk23"     "pregsmk33"
 [329] "pregdrink3"    "lchmulti3"     "pregm4"        "pregy44"       "pregfert4"
"invitro4"      "pregout4"      "pregend4"
 [337] "endmnth4"      "pregsmoke4"    "smkmnth14"     "smkmnth24"     "smkmnth34"
"pregsmk14"     "pregsmk24"     "pregsmk34"
 [345] "pregdrink4"    "lchmulti4"     "pregm5"        "pregy45"       "pregfert5"
"invitro5"      "pregout5"      "pregend5"
 [353] "endmnth5"      "pregsmoke5"    "smkmnth15"     "smkmnth25"     "smkmnth35"
"pregsmk15"     "pregsmk25"     "pregsmk35"
 [361] "pregdrink5"    "lchmulti5"     "hcondno1"      "hcondns1"      "hcondno2"
"hcondns2"      "hcondno3"      "hcondns3"
 [369] "hcondno4"      "hcondns4"      "hcondno5"      "hcondns5"      "hcondno6"
"hcondns6"      "hcondno7"      "hcondns7"
 [377] "hcondno8"      "hcondns8"      "contft"        "ftendm"        "ftendy4"
"ftquals"       "ftedstartm1"   "ftedstarty41"
 [385] "ftedend1"      "ft2endm1"      "ft2endy41"     "ftedmor1"      "ftedstartm2"
"ftedstarty42"  "ftedend2"      "ft2endm2"
 [393] "ft2endy42"     "ftedmor2"      "qualnew1"      "qualnew2"      "qualnew3"
"qualnew4"      "qualnew5"      "qualnew6"
 [401] "qualnew7"      "qualnew8"      "qualnew9"      "qualnew10"     "qualnew11"
"qualnew13"     "qualnew15"     "qualnew16"
 [409] "qualnew17"     "qualnew18"     "qualnew19"     "qualnew20"     "qualnew21"
"qualnew22"     "qualnew23"     "qualnew24"
 [417] "qualnew25"     "qualnew26"     "qualnew27"     "qualnew28"     "qualnew29"
"qualnew30"     "qualnew31"     "trwho1"
 [425] "traindays1"    "trainhrs1"     "trainend1"     "trainpurp11"   "trainpurp21"
"trainpurp31"   "trainpurp41"   "trainpurp51"
 [433] "trainpurp61"   "trainpurp71"   "trainqual1"    "trwho2"        "traindays2"
"trainhrs2"     "trainend2"     "trainpurp12"
 [441] "trainpurp22"   "trainpurp32"   "trainpurp42"   "trainpurp52"   "trainpurp62"
"trainpurp72"   "trainqual2"    "trwho3"
 [449] "traindays3"    "trainhrs3"     "trainend3"     "trainpurp13"   "trainpurp23"
"trainpurp33"   "trainpurp43"   "trainpurp53"
```

```
[457] "trainpurp63"  "trainpurp73"  "trainqual3"   "trainn"       "trqual1"
"trqual2"       "trqual3"       "trqual4"
[465] "trqual5"      "trqual6"      "trqual7"      "trqual8"      "trqual9"
"trqual10"      "trqual11"      "trqual13"
[473] "trqual15"     "trqual16"     "trqual17"     "trqual18"     "trqual19"
"trqual20"      "trqual21"      "trqual22"
[481] "trqual23"     "trqual24"     "trqual25"     "trqual26"     "trqual27"
"trqual28"      "trqual29"      "trqual30"
[489] "trqual31"     "notempchk"    "empchk"       "empstendd"    "empstendm"
"empstendy4"    "stendreas"     "nxtst"
[497] "nxtstelse"    "cstat"        "nxtstendd"    "nxtstendm"    "nxtstendy4"
"jbsamr"        "wkplsam"       "samejob"
[505] "matlv"        "matlvstd"     "matlvstm"     "matlvsty4"    "matlvendd"
"matlvendm"     "matlvendy4"    "jbendd"
[513] "jbendm"       "jbendy4"      "jbendreas"    "cjob"         "nxtjbhrs"
"nxtjbes"       "nxtjbendd"     "nxtjbendm"
[521] "nxtjbendy4"   "cjbatt"       "nextstat1"    "nextelse1"    "currstat1"
"nextjob1"      "currjob1"      "jobhours1"
[529] "reasend1"     "jbatt1"       "statendd1"    "statendm1"    "statendy41"
"nextstat2"     "nextelse2"     "currstat2"
[537] "nextjob2"     "currjob2"     "jobhours2"    "reasend2"     "jbatt2"
"statendd2"     "statendm2"     "statendy42"
[545] "nextstat3"    "nextelse3"    "currstat3"    "nextjob3"     "currjob3"
"jobhours3"     "reasend3"      "jbatt3"
[553] "statendd3"    "statendm3"    "statendy43"   "nextstat4"    "nextelse4"
"currstat4"     "nextjob4"      "currjob4"
[561] "jobhours4"    "reasend4"     "jbatt4"       "statendd4"    "statendm4"
"statendy44"    "nextstat5"     "nextelse5"
[569] "currstat5"    "nextjob5"     "currjob5"     "jobhours5"    "reasend5"
"jbatt5"        "statendd5"     "statendm5"
[577] "statendy45"   "nextstat6"    "nextelse6"    "currstat6"    "nextjob6"
"currjob6"      "jobhours6"     "reasend6"
[585] "jbatt6"       "statendd6"    "statendm6"    "statendy46"   "nextstat7"
"nextelse7"     "currstat7"     "nextjob7"
[593] "currjob7"     "jobhours7"    "reasend7"     "jbatt7"       "statendd7"
"statendm7"     "statendy47"    "nextstat8"
[601] "nextelse8"    "currstat8"    "nextjob8"     "currjob8"     "jobhours8"
"reasend8"      "jbatt8"        "statendd8"
[609] "statendm8"    "statendy48"   "nextstat9"    "nextelse9"    "currstat9"
"nextjob9"      "currjob9"      "jobhours9"
[617] "reasend9"     "jbatt9"       "statendd9"    "statendm9"    "statendy49"
"nextstat10"    "nextelse10"    "currstat10"
[625] "nextjob10"    "currjob10"    "jobhours10"   "reasend10"    "jbatt10"
"statendd10"    "statendm10"    "statendy410"
[633] "jboff"        "jboffy"       "jbterm1"      "jbterm2"      "jbsic07chk"
"jbsoc00chk"    "jbsempchk"     "jbsemp"
[641] "jbbgd"        "jbbgm"        "jbbgy"        "jbmngrchk"    "jbmngr"
"jbsizechk"     "jbsize"        "jbsect"
[649] "jbsectpub"    "jbhrs"        "jbot"         "jbotpd"       "paygwc"
"paynwc"        "payusl"        "payu"
[657] "payuwc"       "payug"        "paytyp"       "ovtpay"       "extnsa"
"extrate"       "extrest"       "basnsa"
[665] "basrate"      "basrest"      "ovtnsa"       "ovtrate"      "ovtrest"
"jbpl"          "jbttwt"        "worktrav"
[673] "jsboss"       "jssize"       "jshrs"        "jstypeb"      "jsaccs"
"jspart"        "jsprbm"        "jsprby4"
[681] "jsprem"       "jsprey4"      "jsprls"       "jsprtx"       "jsprni"
"jspayu"        "jspayw"        "jspytx"
[689] "jspyni"       "jspl"         "jsttwt"       "jsttwtb"      "jsworktrav"
"workdis"       "twkdiff1"      "twkdiff2"
[697] "twkdiff3"     "twkdiff4"     "twkdiff5"     "twkdiff6"     "twkdiff7"
"twkdiff8"      "twkdiff97"     "twkdiffm"
[705] "twkcar"       "twkcary1"     "twkcary2"     "twkcary3"     "twkcary4"
"twkcary5"      "twkcary6"      "twkcary7"
[713] "twkcary8"     "twkcary9"     "twkcary10"    "twkcary11"    "twkcary12"
"twkcary13"     "twkcary14"     "twkcary97"
[721] "twkcarym"     "altcar1"      "altcar2"      "altcar3"      "altcar4"
"altcar5"       "altcar6"       "altcar7"
```

```
 [729] "altcar8"      "altcar9"      "altcar10"     "altcar11"     "altcar12"
"altcar96"     "altcar97"     "carclub"
 [737] "carshare"     "wkhome"       "lifthh"       "liftxhh"      "motcyc"
"comtaxi"      "combus"       "comtrain"
 [745] "commetro"     "combike"      "comwalk"      "comother"     "jbsat"
"wkphys"       "jbperfp"      "jbonus"
 [753] "jbrise"       "tujbpl"       "tuin1"        "jbpen"        "jbpenm"
"jbpeny4"      "penmcn"       "penmpy"
 [761] "penmtp"       "penspb"       "wktime"       "wkends"       "jbflex1"
"jbflex2"      "jbflex3"      "jbflex4"
 [769] "jbflex5"      "jbflex6"      "jbflex7"      "jbflex8"      "jbflex96"
"jbfxuse1"     "jbfxuse2"     "jbfxuse3"
 [777] "jbfxuse4"     "jbfxuse5"     "jbfxuse6"     "jbfxuse7"     "jbfxuse8"
"jbfxuse96"    "jbfxinf"      "wkaut1"
 [785] "wkaut2"       "wkaut3"       "wkaut4"       "wkaut5"       "depenth1"
"depenth2"     "depenth3"     "depenth4"
 [793] "depenth5"     "depenth6"     "jblkcha"      "jbxpcha"      "jblkchb"
"jbxpchb"      "jblkchc"      "jbxpchc"
 [801] "jblkchd"      "jbxpchd"      "jblkche"      "jbxpche"      "jbsec"
"julk4wk"      "julkjb"       "jubgn"
 [809] "julk4x1"      "julk4x2"      "julk4x3"      "julk4x4"      "julk4x5"
"julk4x6"      "julk4x96"     "jbhad"
 [817] "jlendm"       "jlendy"       "jlsemp"       "jlboss"       "jlmngr"
"jlsize"       "eprosh"       "j2semp"
 [825] "j2hrs"        "j2pay"        "retchk"       "ageret"       "rtpro1"
"rtpro2"       "rtpro3"       "rtpro4"
 [833] "rtpro5"       "rtpro6"       "rtcon1"       "rtcon2"       "rtcon3"
"rtcon4"       "penmex"       "pppex"
 [841] "pppexm"       "sppen"        "rtexpjb"      "rtfnd1"       "rtfnd2"
"rtfnd3"       "rtfnd4"       "rtfnd5"
 [849] "rtfnd6"       "rtfnd7"       "rtfnd8"       "rtfnd9"       "rtfnd10"
"rtfnd96"      "retamt"       "retsuf"
 [857] "volfreq"      "volhrs"       "charfreq"     "charam"       "ccare"
"ccwork"       "benunemp1"    "benunemp2"
 [865] "benunemp96"   "bendis1"      "bendis2"      "bendis3"      "bendis4"
"bendis5"      "bendis6"      "bendis7"
 [873] "bendis8"      "bendis9"      "bendis10"     "bendis11"     "bendis96"
"benpen1"      "benpen2"      "benpen3"
 [881] "benpen4"      "benpen5"      "benpen6"      "benpen7"      "benpen8"
"benpen96"     "niserps"      "bencb"
 [889] "benctc"       "benfam1"      "benfam2"      "benfam3"      "benfam4"
"benfam5"      "benfam96"     "bentax1"
 [897] "bentax2"      "bentax3"      "bentax4"      "bentax5"      "bentax96"
"benhou1"      "benhou2"      "benhou3"
 [905] "benhou4"      "benhou96"     "nfh01"        "nfh02"        "nfh03"
"nfh04"        "nfh05"        "nfh06"
 [913] "nfh07"        "nfh08"        "nfh09"        "nfh10"        "nfh11"
"nfh12"        "nfh13"        "nfh14"
 [921] "nfh15"        "nfh16"        "nfh17"        "nfh18"        "nfh19"
"nfh20"        "nfh21"        "nfh22"
 [929] "nfh23"        "nfh24"        "nfh25"        "nfh26"        "nfh27"
"nfh28"        "nfh29"        "nfh30"
 [937] "nfh31"        "nfh32"        "nfh33"        "nfh34"        "nfh35"
"nfh36"        "nfh37"        "nfh38"
 [945] "fiyrdia"      "fiyrdb1"      "fiyrdb2"      "fiyrdb3"      "fiyrdb4"
"fiyrdb5"      "fiyrdb6"      "ppent"
 [953] "ppyrs"        "ppreg"        "ppram"        "pprampc"      "saved"
"savreg"       "savlt"        "hubuys"
 [961] "hufrys"       "humops"       "huiron"       "hupots"       "hudiy"
"husits"       "huboss"       "vote2"
 [969] "vote3"        "vote4"        "vote5"        "perpolinf"    "colbens1"
"colbens2"     "colbens3"     "civicduty"
 [977] "polcost"      "votenorm"     "perbfts"      "grpbfts"      "voteintent"
"demorient"    "vote7"        "vote8"
 [985] "arts1freq"    "arts2freq"    "libfreq"      "arcfreq"      "musfreq"
"herfreq"      "sportsfreq"   "sports3freq"
 [993] "club"         "acc1to161"    "acc1to162"    "acc1to163"    "acc1to164"
"acc1to165"    "acc1to166"    "acc1to167"
```

```
 [ reached getOption("max.print") -- omitted 361 entries ]
[1] "-------------Low Level Removal-------------"
[1] "If a level is removed from a variable you wish to keep, reccomended to manually merge levels to
[1] "level 5 in pno removed, 1 observations found"
[1] "level 3 in hhorig.x removed, 1 observations found"
[1] "level 5 in nch14resp removed, 2 observations found"
[1] "level 5 in nch415resp removed, 0 observations found"
[1] "level 5 in nchresp removed, 0 observations found"
[1] "level 5 in nchund18resp removed, 1 observations found"
[1] "level 6 in natch01 removed, 2 observations found"
[1] "level 7 in natch02 removed, 0 observations found"
[1] "level 7 in natch03 removed, 1 observations found"
[1] "level 5 in natch04 removed, 4 observations found"
[1] "level 7 in natch04 removed, 2 observations found"
[1] "level 6 in natch05 removed, 0 observations found"
[1] "level 7 in natch05 removed, 2 observations found"
[1] "level 8 in natch05 removed, 1 observations found"
[1] "level 9 in natch06 removed, 0 observations found"
[1] "level 5 in nnatch removed, 0 observations found"
[1] "level 6 in nnatch removed, 0 observations found"
[1] "level 3 in nadoptch removed, 2 observations found"
[1] "level 4 in nadoptch removed, 1 observations found"
[1] "level 1 in adoptch01 removed, 2 observations found"
[1] "level 2 in adoptch01 removed, 4 observations found"
[1] "level 5 in adoptch01 removed, 4 observations found"
[1] "level 6 in adoptch01 removed, 1 observations found"
[1] "level 3 in adoptch02 removed, 0 observations found"
[1] "level 5 in adoptch02 removed, 2 observations found"
[1] "level 5 in adoptch03 removed, 0 observations found"
[1] "level 6 in adoptch03 removed, 0 observations found"
[1] "level 7 in adoptch04 removed, 0 observations found"
[1] "level 5 in nchunder16 removed, 0 observations found"
[1] "level 6 in nchunder16 removed, 0 observations found"
[1] "level 6 in nch5to15 removed, 0 observations found"
[1] "level 4 in nch10to15 removed, 1 observations found"
[1] "level 6 in allch01 removed, 2 observations found"
[1] "level 7 in allch02 removed, 0 observations found"
[1] "level 8 in allch02 removed, 0 observations found"
[1] "level 7 in allch03 removed, 0 observations found"
[1] "level 5 in allch04 removed, 0 observations found"
[1] "level 7 in allch04 removed, 0 observations found"
[1] "level 8 in allch04 removed, 0 observations found"
[1] "level 6 in allch05 removed, 0 observations found"
[1] "level 7 in allch05 removed, 0 observations found"
[1] "level 8 in allch05 removed, 0 observations found"
[1] "level 9 in allch06 removed, 0 observations found"
[1] "level 1 in hcondn3 removed, 1 observations found"
[1] "level 1 in hcondn9 removed, 4 observations found"
[1] "level 1 in bensta3 removed, 3 observations found"
[1] "level 7 in marstat.x removed, 1 observations found"
[1] "level 4 in ivcoop removed, 2 observations found"
[1] "level 4 in undqus removed, 3 observations found"
[1] "level 1 in ivtrans removed, 3 observations found"
[1] "level 4 in hgbiom removed, 2 observations found"
[1] "level 5 in hgbiom removed, 2 observations found"
[1] "level 3 in hgbiof removed, 1 observations found"
[1] "level 4 in pn1pno removed, 0 observations found"
[1] "level 5 in pn1pno removed, 0 observations found"
[1] "level 3 in pn2pno removed, 1 observations found"
[1] "level 4 in pn2pno removed, 0 observations found"
[1] "level 4 in pns1pno removed, 0 observations found"
[1] "level 5 in pns1pno removed, 0 observations found"
[1] "level 3 in pns2pno removed, 0 observations found"
[1] "level 4 in pns2pno removed, 0 observations found"
[1] "level 1 in fiyrinvinc_tc removed, 1 observations found"
[1] "level 1 in fibenothr_tc removed, 2 observations found"
[1] "level 9 in ff_jbstat removed, 3 observations found"
[1] "level 10 in ff_jbstat removed, 2 observations found"
```

```
[1] "level 1 in ff_bentype06 removed, 2 observations found"
[1] "level 1 in ff_bentype21 removed, 2 observations found"
[1] "level 1 in ff_bentype25 removed, 2 observations found"
[1] "level 1 in ff_bentype30 removed, 1 observations found"
[1] "level 1 in ff_bentype35 removed, 2 observations found"
[1] "level 1 in ngrp_dv removed, 1 observations found"
[1] "level 2 in nnssib_dv removed, 2 observations found"
[1] "level 5 in nnssib_dv removed, 1 observations found"
[1] "level 2 in agegr13_dv removed, 1 observations found"
[1] "level 7 in mastat_dv removed, 0 observations found"
[1] "level 4 in buno_dv removed, 3 observations found"
[1] "level 5 in buno_dv removed, 0 observations found"
[1] "level 5 in nchild_dv removed, 0 observations found"
[1] "level 6 in nchild_dv removed, 0 observations found"
[1] "level 5 in hrpno removed, 2 observations found"
[1] "level 5 in ppno removed, 3 observations found"
[1] "level 5 in sppno removed, 0 observations found"
[1] "level 3 in fnpno removed, 0 observations found"
[1] "level 3 in fnspno removed, 0 observations found"
[1] "level 4 in mnpno removed, 0 observations found"
[1] "level 5 in mnpno removed, 0 observations found"
[1] "level 4 in mnspno removed, 0 observations found"
[1] "level 5 in mnspno removed, 0 observations found"
[1] "level 1 in grmpno removed, 0 observations found"
[1] "level 2 in nnmpsp_dv removed, 3 observations found"
[1] "level 3 in nunmpsp_dv removed, 1 observations found"
[1] "level 3 in ficode3 removed, 2 observations found"
[1] "level 1 in ficode6 removed, 3 observations found"
[1] "level 1 in ficode21 removed, 1 observations found"
[1] "level 2 in ficode24 removed, 1 observations found"
[1] "level 1 in ficode25 removed, 0 observations found"
[1] "level 2 in ficode26 removed, 2 observations found"
[1] "level 2 in ficode27 removed, 1 observations found"
[1] "level 2 in ficode28 removed, 1 observations found"
[1] "level 1 in ficode30 removed, 1 observations found"
[1] "level 1 in ficode37 removed, 3 observations found"
[1] "level 2 in ficode39 removed, 1 observations found"
[1] "level 4 in b_pno.x removed, 3 observations found"
[1] "level 5 in b_pno.x removed, 0 observations found"
[1] "level 1 in b_splitnum.x removed, 1 observations found"
[1] "level 3 in medcnjd removed, 1 observations found"
[1] "level 1 in difbpc4 removed, 4 observations found"
[1] "level 1 in difbpc5 removed, 4 observations found"
[1] "level 1 in vppress3 removed, 2 observations found"
[1] "level 4 in nseqno removed, 0 observations found"
[1] "level 5 in nseqno removed, 0 observations found"
[1] "level 7 in lfout removed, 4 observations found"
[1] "level 9 in elig removed, 4 observations found"
[1] "level 6 in ethnic removed, 3 observations found"
[1] "level 10 in ethnic removed, 4 observations found"
[1] "level 12 in ethnic removed, 4 observations found"
[1] "level 16 in ethnic removed, 1 observations found"
[1] "level 4 in wstokb removed, 1 observations found"
[1] "level 7 in hhsize removed, 0 observations found"
[1] "level 8 in hhsize removed, 0 observations found"
[1] "level 9 in hhsize removed, 0 observations found"
[1] "level 10 in jbstat.y removed, 3 observations found"
[1] "level 7 in marstat.y removed, 0 observations found"
[1] "level 4 in b_pno.y removed, 0 observations found"
[1] "level 5 in b_pno.y removed, 0 observations found"
[1] "level 1 in b_splitnum.y removed, 0 observations found"
[1] "level 2 in nnsib_dv removed, 0 observations found"
[1] "level 1 in depchl_dv removed, 1 observations found"
[1] "level 9 in qfhigh_dv removed, 1 observations found"
[1] "level 3 in nmpsp_dv removed, 1 observations found"
[1] "130 total levels removed from 95 different variables. In total 151 observations deleted"
[1] "——————Variance 0 Check——————"
[1] "126 variables removed since their new variance was 0"
```

```
  [1] "wave"            "hhorig.x"        "memorig"         "sampst"          "ivfio"
"ioutcome"        "newper"
  [8] "newentrant"      "adstatus"        "natch05"         "natch06"         "natch07"
"natch08"         "natch09"
 [15] "natch10"         "natch11"         "natch12"         "natch13"         "natch14"
"natch15"         "natch16"
 [22] "adoptch03"       "adoptch04"       "adoptch05"       "adoptch06"       "adoptch07"
"adoptch08"       "adoptch09"
 [29] "adoptch10"       "adoptch11"       "adoptch12"       "adoptch13"       "adoptch14"
"adoptch15"       "adoptch16"
 [36] "allch05"         "allch06"         "allch07"         "allch08"         "allch09"
"allch10"         "allch11"
 [43] "allch12"         "allch13"         "allch14"         "allch15"         "allch16"
"chkdob"          "hcondn3"
 [50] "hcondn9"         "hcondn15"        "bensta3"         "indmode"         "ivtrans"
"sceverdrnk"      "screlany"
 [57] "intdatd_if"      "intdatm_if"      "intdaty_if"      "doby_if"         "age_if"
"fiyrinvinc_tc" "fibenothr_tc"
 [64] "ff_ivlolw"       "ff_everint"      "ff_bentype06"    "ff_bentype21"    "ff_bentype25"
"ff_bentype30"    "ff_bentype31"
 [71] "ff_bentype32"    "ff_bentype35"    "ff_bentype36"    "ff_bentype37"    "ff_bentype38"
"ngrp_dv"         "grfpno"
 [78] "grmpno"          "indpxbh_xw"      "indinbh_xw"      "indscbh_xw"      "indin91_lw"
"indin01_lw"      "ficode6"
 [85] "ficode21"        "ficode25"        "ficode30"        "ficode31"        "ficode32"
"ficode35"        "ficode36"
 [92] "ficode37"        "hhorig.y"        "b_splitnum.x"    "medtyp13"        "resphts"
"whintro"         "bpconst"
 [99] "respbps"         "difbpc4"         "difbpc5"         "mmgswil"         "mmgssta"
"clotb"           "fit"
[106] "bswill"          "constorb"        "samdifc6"        "vppress3"        "dateok"
"bfck2"           "nuroutc"
[113] "bsoute"          "htok"            "wtok"            "bmiok"           "elig"
"full1"           "full2"
[120] "full3"           "wstokb"          "b_splitnum.y"    "hhorig"          "depchl_dv"
"scflag_dv"       "qfhighfl_dv"
[1] "———————Dummy Variables———————"
[1] "predictor variable count went from 564 to 1097"
[1] "———————Variance 0 Check———————"
[1] "93 variables removed since their new variance was 0"
 [1] "pno.4"            "pno.5"            "nch14resp.5"      "nch415resp.5"
"nchresp.5"        "nchund18resp.5" "natch01.6"
 [8] "natch02.7"        "natch03.7"        "natch04.5"        "natch04.7"
"nnatch.5"         "nnatch.6"         "nadoptch.3"
[15] "nadoptch.4"       "adoptch01.1"      "adoptch01.2"      "adoptch01.5"
"adoptch01.6"      "adoptch02.3"      "adoptch02.5"
[22] "nchunder16.5"     "nchunder16.6"     "nch5to15.6"       "nch10to15.4"
"allch01.6"        "allch02.7"        "allch02.8"
[29] "allch03.7"        "allch04.5"        "allch04.7"        "allch04.8"
"jbstat.x.10"      "marstat.x.7"      "ivcoop.4"
[36] "undqus.4"         "hgbiom.4"         "hgbiom.5"         "hgbiof.3"
"pn1pno.4"         "pn1pno.5"         "pn2pno.3"
[43] "pn2pno.4"         "pns1pno.4"        "pns1pno.5"        "pns2pno.3"
"pns2pno.4"        "ff_jbstat.9"      "ff_jbstat.10"
[50] "nnssib_dv.2"      "nnssib_dv.5"      "mastat_dv.7"      "buno_dv.4"
"buno_dv.5"        "nchild_dv.5"      "nchild_dv.6"
[57] "hrpno.5"          "ppno.5"           "sppno.5"          "fnpno.3"
"fnspno.3"         "mnpno.4"          "mnpno.5"
[64] "mnspno.4"         "mnspno.5"         "nnmpsp_dv.2"      "nunmpsp_dv.3"
"ficode3.3"        "ficode24.2"       "ficode26.2"
[71] "ficode27.2"       "ficode28.2"       "ficode39.2"       "b_pno.x.4"
"b_pno.x.5"        "medcnjd.3"        "nseqno.4"
[78] "nseqno.5"         "lfout.7"          "ethnic.6"         "ethnic.10"
"ethnic.12"        "ethnic.16"        "hhsize.7"
[85] "hhsize.8"         "hhsize.9"         "jbstat.y.10"      "marstat.y.7"
"b_pno.y.4"        "b_pno.y.5"        "nnsib_dv.2"
[92] "qfhigh_dv.9"      "nmpsp_dv.3"
[1] "———————K-Means———————"
```

```
[1] "15 clusters have been made for K—Means"
[1] "K—Means results as a table, the max value in each row is a simple way to define which cluster re

      0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
  1   0  0  0  1  1  2 10 15  3  9  3  8 10  7  2  1  2  1  2  0  1  3  1
1  0  0  0  0  0  0  0  0  0  0  0  0  0
  2   1  0  1  1  1  6 25 28 29 33 27 21 18  6  6  5  5  6  5  4  4  2  3
0  3  0  0  0  1  0  0  0  0  0  0  1  0
  3   0  1  0  3  1  7 39 41 33 33 29 36 35 20 16 11  8  6  5  3  2  5  1
3  3  4  0  0  1  0  1  1  1  1  0  0  0
  4   0  2  3  2  5  5 34 43 38 42 41 32 37 18 13 11  9  5  9  7  6  2  3
6  3  2  1  1  2  1  0  0  0  1  0  1  1
  5   0  0  0  0  0  3 10  8  8  5 10  7  5  6  2  1  2  1  3  2  1  3  1
1  2  0  1  0  0  1  0  0  0  0  0  0  0
  6   0  0  0  3  1  7 22 16 16 17 12 14 12  9  6  7  3  1  0  2  1  0  5
2  0  2  1  0  3  0  1  1  1  0  0  1  1
  7   0  0  1  1  1  4 27 26 26 20 20 10 12  6  5  6  3  2  1  1  2  3  2
2  1  2  2  2  1  0  0  0  0  1  0  0  0
  8   0  0  0  0  1  4 18 24 16 24 13 17  7  8  2  3  4  2  3  1  3  2  2
1  0  1  0  0  0  1  1  0  0  0  0  1  0
  9   0  0  0  1  3  3 21 14 18 13 11 12  6  4  2  3  3  1  4  2  3  1  2
0  1  0  1  1  0  0  0  0  0  0  0  1  0
  10  3  1  2  2  1  6 34 27 28 23 28 34 27 11  5 15  1  4  7  7  6  2  3
2  2  1  1  2  1  1  0  2  0  0  0  0  1
  11  0  1  1  2  2  8 82 72 65 42 40 40 37 25 18 11 10 12  7  8  6  3  2
4  8  3  3  2  0  0  0  1  0  0  1  0  0
  12  0  0  1  1  3  6 32 33 20 25 15 22 23 10  6  9  5  1  3  1  5  3  3
4  2  0  0  1  2  1  0  1  0  1  0  0  1
  13  0  0  0  1  4  7 54 58 39 49 51 39 54 24 10 10  4  9  7 11  4  8 10
2  4  0  2  2  1  0  0  0  1  1  2  2  1
  14  0  0  0  0  2  3 16 15 12  9  7 15  8  6  1  2  3  2  0  1  1  1  4
0  0  0  1  0  0  1  0  1  0  0  0  0  0
  15  0  0  0  1  2  0 11 18 10  8 14  7 16  5  4  3  1  3  0  3  0  5  0
1  2  0  0  0  0  0  0  1  1  2  0  0  0
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
 [1] "Cluster 1: Within MSE 3730828753998, Size 83"       "Cluster 2: Within MSE 5744802398308633, S
 [3] "Cluster 3: Within MSE 5741824631339592, Size 350"   "Cluster 4: Within MSE 5803497824055890, S
 [5] "Cluster 5: Within MSE 3047346675020, Size 83"       "Cluster 6: Within MSE 11939825021248, Size
 [7] "Cluster 7: Within MSE 12577155829388, Size 190"     "Cluster 8: Within MSE 8159082429605, Size
 [9] "Cluster 9: Within MSE 8827974911371, Size 131"      "Cluster 10: Within MSE 5706569156089935, S
[11] "Cluster 11: Within MSE 15090016999551488, Size 516" "Cluster 12: Within MSE 5679924775278238, S
[13] "Cluster 13: Within MSE 14666377656828436, Size 471" "Cluster 14: Within MSE 8853056556388, Size
[15] "Cluster 15: Within MSE 7354231489496, Size 118"
[1] "Total between cluster MSE: 958612412723509376, Total within cluster MSE: 6604942340195843"
[1] "The K—Means model predicts exactly with an accuracy of 0.1349"
[1] "—————Correlation Checks—————"
[1] "map1 removed, correlated with 10 other variable(s)"
[1] "pidp removed, correlated with 7 other variable(s)"
[1] "weight removed, correlated with 7 other variable(s)"
[1] "map2 removed, correlated with 9 other variable(s)"
[1] "hidp removed, correlated with 6 other variable(s)"
[1] "dvage removed, correlated with 6 other variable(s)"
[1] "employ.2 removed, correlated with 7 other variable(s)"
[1] "estwt removed, correlated with 6 other variable(s)"
[1] "pno.2 removed, correlated with 5 other variable(s)"
[1] "month removed, correlated with 5 other variable(s)"
[1] "sex.2 removed, correlated with 5 other variable(s)"
[1] "birthy removed, correlated with 5 other variable(s)"
[1] "indpxus_xw removed, correlated with 5 other variable(s)"
[1] "mmgsd1 removed, correlated with 5 other variable(s)"
[1] "wtval removed, correlated with 5 other variable(s)"
[1] "alb removed, correlated with 5 other variable(s)"
[1] "map3 removed, correlated with 7 other variable(s)"
[1] "pno.3 removed, correlated with 4 other variable(s)"
[1] "nchresp.4 removed, correlated with 4 other variable(s)"
[1] "istrtdaty.2011 removed, correlated with 4 other variable(s)"
[1] "chksex.2 removed, correlated with 4 other variable(s)"
[1] "hgbiom.1 removed, correlated with 4 other variable(s)"
```

```
[1] "hgbiom.3 removed, correlated with 4 other variable(s)"
[1] "scdoby4 removed, correlated with 4 other variable(s)"
[1] "ppno.1 removed, correlated with 4 other variable(s)"
[1] "indinus_xw removed, correlated with 4 other variable(s)"
[1] "mmgsn1 removed, correlated with 4 other variable(s)"
[1] "waist1 removed, correlated with 4 other variable(s)"
[1] "chol removed, correlated with 4 other variable(s)"
[1] "nchresp.2 removed, correlated with 4 other variable(s)"
[1] "nchresp.3 removed, correlated with 4 other variable(s)"
[1] "psu.x removed, correlated with 3 other variable(s)"
[1] "relup.2 removed, correlated with 3 other variable(s)"
[1] "marstat.x.2 removed, correlated with 3 other variable(s)"
[1] "marstat.x.4 removed, correlated with 3 other variable(s)"
[1] "marstat.x.5 removed, correlated with 3 other variable(s)"
[1] "marstat.x.6 removed, correlated with 3 other variable(s)"
[1] "scsex.2 removed, correlated with 3 other variable(s)"
[1] "pn1pno.1 removed, correlated with 3 other variable(s)"
[1] "pn1pno.3 removed, correlated with 3 other variable(s)"
[1] "pn2pno.2 removed, correlated with 3 other variable(s)"
[1] "age_dv removed, correlated with 3 other variable(s)"
[1] "intdaty_dv.2011 removed, correlated with 3 other variable(s)"
[1] "country.3 removed, correlated with 3 other variable(s)"
[1] "buno_dv.3 removed, correlated with 3 other variable(s)"
[1] "sppno.1 removed, correlated with 3 other variable(s)"
[1] "indscus_xw removed, correlated with 3 other variable(s)"
[1] "mmgsd2 removed, correlated with 3 other variable(s)"
[1] "sys1 removed, correlated with 3 other variable(s)"
[1] "dias1 removed, correlated with 3 other variable(s)"
[1] "waist2 removed, correlated with 3 other variable(s)"
[1] "nchresp.1 removed, correlated with 3 other variable(s)"
[1] "ff_jbstat.2 removed, correlated with 3 other variable(s)"
[1] "ff_emplw.2 removed, correlated with 3 other variable(s)"
[1] "pensioner_dv.2 removed, correlated with 3 other variable(s)"
[1] "hdl removed, correlated with 3 other variable(s)"
[1] "strata.x removed, correlated with 2 other variable(s)"
[1] "natch04.6 removed, correlated with 2 other variable(s)"
[1] "nchunder16.3 removed, correlated with 2 other variable(s)"
[1] "nchunder16.4 removed, correlated with 2 other variable(s)"
[1] "jbstat.x.4 removed, correlated with 2 other variable(s)"
[1] "relup.6 removed, correlated with 2 other variable(s)"
[1] "jbhas.2 removed, correlated with 2 other variable(s)"
[1] "btype5.1 removed, correlated with 2 other variable(s)"
[1] "btype8.1 removed, correlated with 2 other variable(s)"
[1] "marstat.x.3 removed, correlated with 2 other variable(s)"
[1] "hgbiom.2 removed, correlated with 2 other variable(s)"
[1] "hgbiof.1 removed, correlated with 2 other variable(s)"
[1] "hgbiof.2 removed, correlated with 2 other variable(s)"
[1] "respm16.2 removed, correlated with 2 other variable(s)"
[1] "pns1pno.1 removed, correlated with 2 other variable(s)"
[1] "pns1pno.3 removed, correlated with 2 other variable(s)"
[1] "pns2pno.2 removed, correlated with 2 other variable(s)"
[1] "fimngrs_tc.1 removed, correlated with 2 other variable(s)"
[1] "country.2 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.6 removed, correlated with 2 other variable(s)"
[1] "cohab_dv.1 removed, correlated with 2 other variable(s)"
[1] "mastat_dv.2 removed, correlated with 2 other variable(s)"
[1] "mastat_dv.4 removed, correlated with 2 other variable(s)"
[1] "mastat_dv.5 removed, correlated with 2 other variable(s)"
[1] "mastat_dv.6 removed, correlated with 2 other variable(s)"
[1] "hrpid removed, correlated with 2 other variable(s)"
[1] "indinub_xw removed, correlated with 2 other variable(s)"
[1] "b_pno.x.2 removed, correlated with 2 other variable(s)"
[1] "b_pno.x.3 removed, correlated with 2 other variable(s)"
[1] "nsex.2 removed, correlated with 2 other variable(s)"
[1] "confage removed, correlated with 2 other variable(s)"
[1] "region.3 removed, correlated with 2 other variable(s)"
[1] "height removed, correlated with 2 other variable(s)"
[1] "mmgsn2 removed, correlated with 2 other variable(s)"
```

```
[1] "pulse1 removed, correlated with 2 other variable(s)"
[1] "sys2 removed, correlated with 2 other variable(s)"
[1] "dias2 removed, correlated with 2 other variable(s)"
[1] "bmi removed, correlated with 2 other variable(s)"
[1] "indnsub_lw removed, correlated with 2 other variable(s)"
[1] "fimnlabgrs_dv removed, correlated with 2 other variable(s)"
[1] "single_dv.1 removed, correlated with 3 other variable(s)"
[1] "trig removed, correlated with 2 other variable(s)"
[1] "natch03.5 removed, correlated with 1 other variable(s)"
[1] "nnatch.4 removed, correlated with 1 other variable(s)"
[1] "nadoptch.2 removed, correlated with 1 other variable(s)"
[1] "nchunder16.1 removed, correlated with 1 other variable(s)"
[1] "nchunder16.2 removed, correlated with 1 other variable(s)"
[1] "allch03.5 removed, correlated with 1 other variable(s)"
[1] "allch04.6 removed, correlated with 1 other variable(s)"
[1] "istrtdaty.2012 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.2 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.3 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.4 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.5 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.6 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.7 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.8 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.9 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.10 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.11 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.12 removed, correlated with 1 other variable(s)"
[1] "istrtdatd removed, correlated with 1 other variable(s)"
[1] "jbstat.x.2 removed, correlated with 1 other variable(s)"
[1] "jbstat.x.3 removed, correlated with 1 other variable(s)"
[1] "jbstat.x.5 removed, correlated with 1 other variable(s)"
[1] "jbstat.x.6 removed, correlated with 1 other variable(s)"
[1] "jbstat.x.7 removed, correlated with 1 other variable(s)"
[1] "jbstat.x.8 removed, correlated with 1 other variable(s)"
[1] "jbstat.x.97 removed, correlated with 1 other variable(s)"
[1] "sf1.x.2 removed, correlated with 1 other variable(s)"
[1] "sf1.x.3 removed, correlated with 1 other variable(s)"
[1] "sf1.x.4 removed, correlated with 1 other variable(s)"
[1] "sf1.x.5 removed, correlated with 1 other variable(s)"
[1] "health.x.2 removed, correlated with 1 other variable(s)"
[1] "btype2.1 removed, correlated with 1 other variable(s)"
[1] "btype6.1 removed, correlated with 1 other variable(s)"
[1] "btype9.1 removed, correlated with 1 other variable(s)"
[1] "bensta2.1 removed, correlated with 1 other variable(s)"
[1] "bensta4.1 removed, correlated with 1 other variable(s)"
[1] "bensta5.1 removed, correlated with 1 other variable(s)"
[1] "bensta6.1 removed, correlated with 1 other variable(s)"
[1] "bensta7.1 removed, correlated with 1 other variable(s)"
[1] "bensta8.1 removed, correlated with 1 other variable(s)"
[1] "arts1b13.1 removed, correlated with 1 other variable(s)"
[1] "respf16.2 removed, correlated with 1 other variable(s)"
[1] "scsf2a.2 removed, correlated with 1 other variable(s)"
[1] "scsf3a.5 removed, correlated with 1 other variable(s)"
[1] "scrletdwn.3 removed, correlated with 1 other variable(s)"
[1] "scrannoy.3 removed, correlated with 1 other variable(s)"
[1] "istrtdathh removed, correlated with 1 other variable(s)"
[1] "pn1pno.2 removed, correlated with 1 other variable(s)"
[1] "fimnlabgrs_tc.1 removed, correlated with 1 other variable(s)"
[1] "j2paynet_dv removed, correlated with 1 other variable(s)"
[1] "ff_jbstat.3 removed, correlated with 1 other variable(s)"
[1] "ff_jbstat.4 removed, correlated with 1 other variable(s)"
[1] "ff_bentype01.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype09.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype10.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype13.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype14.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype18.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype22.1 removed, correlated with 1 other variable(s)"
```

```
[1] "npn_dv.1 removed, correlated with 1 other variable(s)"
[1] "npn_dv.2 removed, correlated with 1 other variable(s)"
[1] "nnssib_dv.1 removed, correlated with 1 other variable(s)"
[1] "urban_dv.x.2 removed, correlated with 1 other variable(s)"
[1] "xtra5min_dv.1 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.5 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.7 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.8 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.9 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.10 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.11 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.12 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.13 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.15 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.5 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.13 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.3 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.10 removed, correlated with 1 other variable(s)"
[1] "ppno.2 removed, correlated with 1 other variable(s)"
[1] "fnpno.1 removed, correlated with 1 other variable(s)"
[1] "fnpno.2 removed, correlated with 1 other variable(s)"
[1] "mnpno.1 removed, correlated with 1 other variable(s)"
[1] "mnpno.2 removed, correlated with 1 other variable(s)"
[1] "mnpno.3 removed, correlated with 1 other variable(s)"
[1] "hiqual_dv.x.2 removed, correlated with 1 other variable(s)"
[1] "hiqual_dv.x.3 removed, correlated with 1 other variable(s)"
[1] "hiqual_dv.x.4 removed, correlated with 1 other variable(s)"
[1] "hiqual_dv.x.5 removed, correlated with 1 other variable(s)"
[1] "hiqual_dv.x.9 removed, correlated with 1 other variable(s)"
[1] "paygu_if.1 removed, correlated with 1 other variable(s)"
[1] "indpxus_lw removed, correlated with 1 other variable(s)"
[1] "ind5mus_lw removed, correlated with 1 other variable(s)"
[1] "indpxub_xw removed, correlated with 1 other variable(s)"
[1] "frmnthimp_dv_total removed, correlated with 1 other variable(s)"
[1] "b_hidp.x removed, correlated with 1 other variable(s)"
[1] "age removed, correlated with 1 other variable(s)"
[1] "region.2 removed, correlated with 1 other variable(s)"
[1] "medcnjd.2 removed, correlated with 1 other variable(s)"
[1] "bpmedc.1 removed, correlated with 1 other variable(s)"
[1] "estht removed, correlated with 1 other variable(s)"
[1] "floorc.2 removed, correlated with 1 other variable(s)"
[1] "consubx5.1 removed, correlated with 1 other variable(s)"
[1] "cufsize.2 removed, correlated with 1 other variable(s)"
[1] "mmgsdom.2 removed, correlated with 1 other variable(s)"
[1] "mmgsd3 removed, correlated with 1 other variable(s)"
[1] "samparm.2 removed, correlated with 1 other variable(s)"
[1] "samparm.3 removed, correlated with 1 other variable(s)"
[1] "nseqno.2 removed, correlated with 1 other variable(s)"
[1] "nseqno.3 removed, correlated with 1 other variable(s)"
[1] "lfout.6 removed, correlated with 1 other variable(s)"
[1] "nurdaym.2 removed, correlated with 1 other variable(s)"
[1] "nurdaym.3 removed, correlated with 1 other variable(s)"
[1] "nurdaym.4 removed, correlated with 1 other variable(s)"
[1] "nurdaym.5 removed, correlated with 1 other variable(s)"
[1] "nurdaym.6 removed, correlated with 1 other variable(s)"
[1] "nurdaym.7 removed, correlated with 1 other variable(s)"
[1] "nurdaym.8 removed, correlated with 1 other variable(s)"
[1] "nurdaym.9 removed, correlated with 1 other variable(s)"
[1] "nurdaym.10 removed, correlated with 1 other variable(s)"
[1] "nurdaym.11 removed, correlated with 1 other variable(s)"
[1] "nurdaym.12 removed, correlated with 1 other variable(s)"
[1] "nurdayy.2011 removed, correlated with 1 other variable(s)"
[1] "nurdayy.2012 removed, correlated with 1 other variable(s)"
[1] "pulse2 removed, correlated with 1 other variable(s)"
[1] "sys3 removed, correlated with 1 other variable(s)"
[1] "dias3 removed, correlated with 1 other variable(s)"
[1] "ag16g10.76 removed, correlated with 1 other variable(s)"
[1] "hhtype_dv.y removed, correlated with 1 other variable(s)"
```

```
[1] "marstat.y.2 removed, correlated with 1 other variable(s)"
[1] "marstat.y.4 removed, correlated with 1 other variable(s)"
[1] "marstat.y.5 removed, correlated with 1 other variable(s)"
[1] "marstat.y.6 removed, correlated with 1 other variable(s)"
[1] "vpstimehh removed, correlated with 1 other variable(s)"
[1] "psu.y removed, correlated with 1 other variable(s)"
[1] "indnsub_xw removed, correlated with 1 other variable(s)"
[1] "alt removed, correlated with 1 other variable(s)"
[1] "hba1c removed, correlated with 1 other variable(s)"
[1] "testo removed, correlated with 1 other variable(s)"
[1] "uscmg..22 removed, correlated with 1 other variable(s)"
[1] "uscmg.1 removed, correlated with 1 other variable(s)"
[1] "fimnnet_dv removed, correlated with 1 other variable(s)"
[1] "respm16_dv.2 removed, correlated with 1 other variable(s)"
[1] "wstval removed, correlated with 1 other variable(s)"
[1] "ethn_dv removed, correlated with 1 other variable(s)"
[1] "242 variables removed since they had high correlation coefs"
[1] "The final cleaned dataset has been completed at this stage and is stored under the name 'cleanDa
[1] "------------Attempting a Train Test Split------------"
[1] "Good train, test split found"
[1] "The working seed found was 3"
[1] "------------kNN------------"
[1] "83 neighbours considered for each test data point"
[1] "kNN results as a table, follow the diagonal for the correctly mapped clusters"
         real
predicted  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 3
        0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
        1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
        2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
        3  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
        4  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
        5  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
        6  1  1  3  1  1  9 59 40 40 26 33 22 19  7  7  4  4  5  3  4  2
1  2  1  2  1  1  0  1  1  0  0  0  0
        7  0  0  1  4  4 11 39 48 25 35 28 19 20 12  7  6  6  3  4  0  3
4  2  0  1  1  0  1  2  1  1  1  1  0
        8  0  0  1  0  0  0  2  4  2  3  5  3  3  2  0  0  0  0  1  2  0
1  0  0  0  0  1  0  0  0  0  0  1  0
        9  0  0  0  0  0  1  8 14 12 11 15 11  5 12  2  3  1  1  1  2  1
3  3  1  0  1  1  0  0  1  0  0  0  0
       10  0  0  0  0  0  0  0  2  2  1  2  5  6  1  0  0  0  1  0  1  0
2  1  0  1  1  0  0  0  0  0  0  0  0
       11  0  0  0  0  0  0  0  0  0  2  4  7  5  6  7  7  1  4  0  3  0
3  2  4  1  1  1  0  0  0  0  0  0  1
       12  0  0  0  0  0  0  0  6  5  4  6  6  7  6  2  5  5  0  1  1  2
2  2  1  2  0  1  0  0  1  0  0  0  0
       13  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1
1  0  0  0  0  0  0  0  0  0  0  0  0
       14  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
       15  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
       16  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
       17  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
       18  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
       19  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
       20  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
```

```
       21  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
       22  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
       23  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
       24  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
       25  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
       26  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
       27  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
       28  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
 [ reached getOption("max.print") -- omitted 8 rows ]
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
[1] "The MSE of the predicted values are of 313.8303"
[1] "The kNN model predicts exactly with an accuracy of 0.1538"
[1] "----------CART prediction model----------"
n= 2653

node), split, n, deviance, yval
      * denotes terminal node

 1) root 2653 73496.3000 10.890690
   2) scsf4a.5>=0.5 1828 17091.8800  8.885120
     4) scsf6c.5>=0.5 1075  4732.7910  7.613953 *
     5) scsf6c.5< 0.5 753  8142.1700 10.699870
      10) scsf6a.2>=0.5 424  2627.0170  9.818396 *
      11) scsf6a.2< 0.5 329  4761.1370 11.835870 *
   3) scsf4a.5< 0.5 825 32759.6700 15.334550
     6) scsf4a.4>=0.5 455  7743.9690 12.830770
      12) scsf7.5>=0.5 221  2588.1720 11.348420 *
      13) scsf7.5< 0.5 234  4211.5380 14.230770 *
     7) scsf4a.4< 0.5 370 18655.7300 18.413510
      14) scsf4a.3>=0.5 261  8823.2260 16.509580
        28) scsf6a.2>=0.5 55  1137.9270 12.236360 *
        29) scsf6a.2< 0.5 206  6412.8350 17.650490 *
      15) scsf4a.3< 0.5 109  6620.9170 22.972480
        30) scsf6a.5< 0.5 90  4919.7890 21.588890
          60) scsf7.2< 0.5 57  2411.3680 18.894740 *
          61) scsf7.2>=0.5 33  1380.0610 26.242420 *
        31) scsf6a.5>=0.5 19   712.7368 29.526320 *
[1] "Variable Importance"
    scsf4a.5     scsf4a.4     scsf4b.5     scsf4a.3     scsf4b.4      scsf7.5
scsf6c.5     scsf4a.2     scsf6c.4
 23644.75931  19400.40680  15390.58879  15178.28382   9795.22247   8450.94668
4332.59312   4012.17607   3225.68765
    scsf4b.3     scsf6a.2      scsf7.2     scsf6a.3     scsf4b.2     scsf2a.3
scsf6c.3     scsf6a.5      scsf7.4
  2543.98552   2026.48053   1540.85752   1394.40938   1384.81358   1357.93821
1113.64003    988.39170    811.43568
     scsf7.3     scsf6b.5     scsf3b.2     finnow.5   indscub_xw qfhigh_dv.13
sclfsat1.6        htval       sf1.y.2
   585.45769    441.96178    383.03354    341.92723    271.74038    239.34906
234.99654    205.15634    200.81522
    finfut.2       ppen.2     scsf6b.2     scsf6a.4   arts2b10.1    istrtdatss
b_hidp.y      doby_dv   ficode18.1
   170.96362    156.06185    155.84540    112.30036    104.04123    104.04123
92.54282     77.92270     77.92270
  fimnpen_dv    usbread.4    usdairy.4
    69.40712     52.02062     46.27141
[1] "The MSE of the predicted values are of 13.4942"
[1] "The CART model predicts exactly with accuracy of 0.1278"
[1] "----------Ordinary Linear Regression (Initial)----------"
[1] "The full model AIC is: 13731.9241"
```

```
[1] "------------Variance Inflation Factor Removal------------"
[1] "The variable natch02.4 was removed since it had a VIF score of 1505.3133"
[1] "The variable dory.2012 was removed since it had a VIF score of 656.6149"
[1] "The variable hhsize.4 was removed since it had a VIF score of 514.2299"
[1] "The variable livesp_dv.1 was removed since it had a VIF score of 457.1411"
[1] "The variable fibenothr_dv was removed since it had a VIF score of 438.3819"
[1] "The variable nchild_dv.2 was removed since it had a VIF score of 267.0221"
[1] "The variable ag16g10.66 was removed since it had a VIF score of 206.0762"
[1] "The variable doby_dv was removed since it had a VIF score of 141.4947"
[1] "The variable gor_dv.8 was removed since it had a VIF score of 113.0756"
[1] "The variable scsf4b.5 was removed since it had a VIF score of 107.9911"
[1] "The variable allch01.3 was removed since it had a VIF score of 91.3055"
[1] "The variable uscmm.2 was removed since it had a VIF score of 81.3871"
[1] "The variable vpprob96.1 was removed since it had a VIF score of 80.6047"
[1] "The variable rach16_dv.2 was removed since it had a VIF score of 78.1855"
[1] "The variable fimnlabnet_dv was removed since it had a VIF score of 71.5841"
[1] "The variable ieqmoecd_dv was removed since it had a VIF score of 67.5423"
[1] "The variable bmivg5.30 was removed since it had a VIF score of 63.1864"
[1] "The variable wkvege.4 was removed since it had a VIF score of 61.2917"
[1] "The variable scsf4a.5 was removed since it had a VIF score of 59.6314"
[1] "The variable scsf7.5 was removed since it had a VIF score of 49.9417"
[1] "The variable scsf6c.5 was removed since it had a VIF score of 43.8175"
[1] "The variable ag16g10.36 was removed since it had a VIF score of 42.143"
[1] "The variable sclfsato.6 was removed since it had a VIF score of 38.2638"
[1] "The variable natch01.3 was removed since it had a VIF score of 36.7014"
[1] "The variable allch02.4 was removed since it had a VIF score of 28.9813"
[1] "The variable jbstat.y.4 was removed since it had a VIF score of 28.8451"
[1] "The variable scmolwp.2 was removed since it had a VIF score of 26.3096"
[1] "The variable nnatch.2 was removed since it had a VIF score of 24.1225"
[1] "The variable scsf3b.5 was removed since it had a VIF score of 23.9725"
[1] "The variable sclfsat1.6 was removed since it had a VIF score of 23.7693"
[1] "The variable access.5 was removed since it had a VIF score of 22.7984"
[1] "The variable fimnlabgrs_if was removed since it had a VIF score of 19.1874"
[1] "The variable agegr13_dv.6 was removed since it had a VIF score of 19.0681"
[1] "The variable bprespc.2 was removed since it had a VIF score of 18.9972"
[1] "The variable sppno.2 was removed since it had a VIF score of 18.8697"
[1] "The variable nchund18resp.2 was removed since it had a VIF score of 18.3218"
[1] "The variable numed2 was removed since it had a VIF score of 16.3964"
[1] "The variable npensioner_dv.2 was removed since it had a VIF score of 15.4048"
[1] "The variable sclfsat7.7 was removed since it had a VIF score of 15.0819"
[1] "The variable nnatch.3 was removed since it had a VIF score of 14.7541"
[1] "The variable samdifc1.1 was removed since it had a VIF score of 14.7347"
[1] "The variable scrcritic.3 was removed since it had a VIF score of 14.1016"
[1] "The variable natch01.2 was removed since it had a VIF score of 13.8577"
[1] "The variable bensta96.1 was removed since it had a VIF score of 13.796"
[1] "The variable nchild_dv.4 was removed since it had a VIF score of 13.3335"
[1] "The variable agegr13_dv.9 was removed since it had a VIF score of 13.1952"
[1] "The variable nch415resp.2 was removed since it had a VIF score of 12.0694"
[1] "The variable mnspno.2 was removed since it had a VIF score of 11.849"
[1] "The variable qfhigh_dv.13 was removed since it had a VIF score of 11.7236"
[1] "The variable sclfsat2.6 was removed since it had a VIF score of 11.2772"
[1] "The variable nch14resp.3 was removed since it had a VIF score of 11.0041"
[1] "The variable scsf6b.3 was removed since it had a VIF score of 10.618"
[1] "The variable scopfamf.4 was removed since it had a VIF score of 10.3804"
[1] "The variable nnmpsp_dv.1 was removed since it had a VIF score of 10.3548"
[1] "The variable btype4.1 was removed since it had a VIF score of 10.1054"
[1] "The variable wkfruit.4 was removed since it had a VIF score of 10.0937"
[1] "56 variables removed from the Ordinary Linear Model since they have a VIF score higher than 10"
[1] "The full model AIC after VIF checks is: 13985.8788"
[1] "----------Backwards Selection----------"
[1] "50 out of 749 variables removed so far."
[1] "100 out of 749 variables removed so far."
[1] "150 out of 749 variables removed so far."
[1] "200 out of 749 variables removed so far."
[1] "250 out of 749 variables removed so far."
[1] "300 out of 749 variables removed so far."
[1] "350 out of 749 variables removed so far."
[1] "400 out of 749 variables removed so far."
```

```
[1] "450 out of 749 variables removed so far."
[1] "500 out of 749 variables removed so far."
[1] "550 out of 749 variables removed so far."
[1] "600 out of 749 variables removed so far."
[1] "606 out of 749 variables removed in backwards selection since they weren't significant at the 95
  [1] "numed2"            "netuse.2"          "dorm.7"            "rtin"
"relhite.2"         "cufsize.3"
  [7] "sf1.y.3"           "sports16.1"        "dord"              "hrpno.3"
"nurdayd"           "btype7.1"
 [13] "sports110.1"       "ficode29.2"        "scssup1.2"         "howlng"
"netuse.6"          "dorm.11"
 [19] "mla3.1"            "vpprob95.1"        "lfout.4"           "calciumb.1"
"sports34.1"        "scfalcdrnk.9"
 [25] "hiqual_dv.y.2"     "sports227.1"       "ficode29.1"        "origadd.2"
"ftedany.2"         "scsf5.3"
 [31] "j2has.2"           "scsf2b.3"          "bmivg5.18"         "beta.1"
"ff_oprlg.2"        "sf1.y.4"
 [37] "adoptch01.3"       "gor_dv.4"          "gor_dv.3"          "difbpc1.1"
"nadoptch.1"        "nch14resp.2"
 [43] "netuse.5"          "ppno.4"            "medtyp9.1"         "sports224.1"
"scfrendany.2"      "arts1a3.1"
 [49] "hcondn10.1"        "hhresp_dv.3"       "ficode7.1"         "scopfamf.3"
"hhsize.6"          "qfhigh_dv.7"
 [55] "ficode2.1"         "vpprob2.1"         "scrannoy.4"        "medtyp1.1"
"paynu_if.1"        "consubx1.1"
 [61] "ff_bentype28.1"    "hscrp"             "sports33.1"        "arts1b12.1"
"gor_dv.6"          "allch03.4"
 [67] "intdatd_dv"        "istrtdatss"        "hcondn16.1"        "sppno.3"
"cindtime"          "finnow.2"
 [73] "medtyp10.1"        "susp.2"            "ff_bentype15.1"    "allch01.5"
"ff_bentype16.1"    "scrletdwn.4"
 [79] "hhtype_dv.x"       "allch02.3"         "natch02.3"         "ggt"
"advvoucher.2"      "undqus.2"
 [85] "qfhigh_dv.4"       "intdatm_dv.5"      "intdatm_dv.3"      "j2pay_if.1"
"scfalcdrnk.2"      "ff_bentype08.1"
 [91] "arts1b11.1"        "obpdrug.1"         "hrpno.2"           "lfout.3"
"nchild_dv.1"       "hgb"
 [97] "samdifc95.1"       "vppress1.1"        "gor_dv.11"         "netuse.4"
"indinus_lw"        "istrtdatmm"
[103] "ienddatmm"         "nmpsp_dv.1"        "ff_bentype23.1"    "lfout.9"
"susp.3"            "arts1b10.1"
[109] "nch14resp.4"       "fnspno.1"          "difbpc3.1"         "wjrel.2"
"scsf5.5"           "sclfsat7.5"
[115] "jbstat.y.97"       "fibenothr_if"      "fimngrs_if"        "dorm.8"
"scopfamb.2"        "scopfamb.3"
[121] "scopfamb.4"        "respwts.2"         "usdairy.5"         "sports19.1"
"access.4"          "arts1b9.1"
[127] "difbpc2.1"         "sclfsat7.3"        "qfhigh_dv.2"       "sports223.1"
"medtyp8.1"         "sports113.1"
[133] "arts2b13.1"        "fimnprben_dv"      "aceinh.1"          "ficode4.2"
"hrpno.4"           "gor_dv.5"
[139] "airtemp"           "gor_dv.10"         "gor_dv.9"          "wlk10m.2"
"iron.1"            "natch01.5"
[145] "medcnj.2"          "ff_bentype04.1"    "scopfamb.5"        "sports296.1"
"sports225.1"       "jbstat.y.3"
[151] "wkvege.2"          "sports18.1"        "wkvege.3"          "arts1b14.1"
"nurdayw.1"         "heritage6.1"
[157] "mobuse.2"          "sports15.1"        "marstat_dv.6"      "allch01.4"
"scssup1.5"         "ff_bentype20.1"
[163] "dorm.2"            "medtyp2.1"         "omronno"           "natch01.4"
"usbread.7"         "agegr13_dv.7"
[169] "undqus.3"          "vote6.3"           "scropenup.3"       "arts2a2.1"
"arts2a96.1"        "dorm.9"
[175] "dheas"             "sclfsat1.5"        "hcondn96.1"        "alkp"
"sportact.10"       "sports226.1"
[181] "ficode10.1"        "btype3.1"          "jbstat.y.6"        "sclfsat2.3"
"nch5to15.4"        "sports220.1"
[187] "sports218.1"       "strtnurhh"         "qfhigh_dv.15"      "medtyp6.1"
```

```
"qfhigh_dv.3"        "ethnic.4"
[193] "medtyp3.1"          "scfalcdrnk.7"       "scropenup.2"        "save.2"
"event1s"            "ienddatss"
[199] "scsf2b.2"           "hhresp_dv.2"        "nch415resp.3"       "sportact.9"
"lfout.8"            "arts1a1.1"
[205] "vpalco.2"           "heritage5.1"        "sports39.1"         "ivprsnt.2"
"bmivg5.25"          "jbstat.y.5"
[211] "allch01.2"          "scopfamd.5"         "ficode27.1"         "sportact.2"
"lfout.5"            "vppress2.1"
[217] "vpsys.2"            "fimninvnet_dv"      "ag16g10.86"         "hcondn12.1"
"ff_bentype07.1"     "ff_bentype17.1"
[223] "sports112.1"        "samdifc4.1"         "vpsam.2"            "ff_jbstat.5"
"samdifc3.1"         "sports12.1"
[229] "samdifc5.1"         "scsf1.3"            "scsf1.2"            "lfout.2"
"mla2.1"             "allch02.6"
[235] "natch03.6"          "access.2"           "sports217.1"        "xpmove.2"
"hcondn8.1"          "ethnic.15"
[241] "smever.2"           "consubx2.1"         "hiqual_dv.y.3"      "agegr10_dv.3"
"ficode5.1"          "scsf3a.2"
[247] "ff_bentype02.1"     "ppen.2"             "heritage7.1"        "mla96.1"
"samdifc2.1"         "cfib"
[253] "ppno.3"             "arts1a4.1"          "ficode19.1"         "hcondn13.1"
"sports216.1"        "strtnurmm"
[259] "fimnlabnet_tc.1"    "scsf3b.3"           "scsf3b.4"           "scsf3a.3"
"agegr13_dv.4"       "relwaitb.2"
[265] "arts2b10.1"         "hcondn1.1"          "sclfsat7.2"         "ivcoop.2"
"indscub_xw"         "indscus_lw"
[271] "seearngrs_if.1"     "ag16g10.56"         "arts1b96.1"         "arts2a5.1"
"adoptch01.4"        "X.next.1"
[277] "natch02.5"          "heritage96.1"       "heritage1.1"        "bmival"
"bmivg5.40"          "ff_bentype34.1"
[283] "scrundstnd.2"       "netuse.3"           "agegr13_dv.11"      "arts2a1.1"
"nch5to15.3"         "event2s"
[289] "ficode17.1"         "medtyp11.1"         "jbstat.y.2"         "jbiindb_dv"
"finnow.3"           "scopfama.2"
[295] "scopfama.3"         "sclfsat1.4"         "arts2b12.1"         "hcondn11.1"
"wlk10m.3"           "usdairy.6"
[301] "mnspno.1"           "pns1pno.2"          "agegr13_dv.12"      "npensioner_dv.1"
"intdatm_dv.10"      "ff_bentype03.1"
[307] "usbread.2"          "uscmg.3"            "agegr5_dv.14"       "ethnic.2"
"natch02.6"          "arts2b96.1"
[313] "floorc.3"           "dory.2011"          "consubx3.1"         "sports14.1"
"wkfruit.2"          "wkfruit.3"
[319] "sports230.1"        "urban_dv.y.2"       "uscmm..22"          "scwkvfast.6"
"sctimemnuf.6"       "ficode11.1"
[325] "fimnsben_dv"        "consubx4.1"         "scopfamf.5"         "hcondn7.1"
"scopfamf.2"         "mmgstp.2"
[331] "scopfamh.3"         "trainany.2"         "b_hidp.y"           "vparm.3"
"sports114.1"        "access.3"
[337] "ficode12.1"         "vphand.2"           "nnewborn.1"         "intdatm_dv.7"
"xtra5minosm_dv.1"   "ethnic.5"
[343] "vpcheck.2"          "fimnpen_dv"         "ff_bentype29.1"     "fimngrs_dv"
"j2pay_dv"           "scrrely.2"
[349] "event3s"            "sclfsat1.3"         "sclfsat2.5"         "sclfsat2.4"
"scopfama.5"         "sports231.1"
[355] "scssup1.3"          "ficode8.1"          "arts2a3.1"          "ficode26.1"
"sportact.7"         "sportact.8"
[361] "sportact.4"         "sportact.3"         "sports229.1"        "scopfama.4"
"sports222.1"        "scrrely.4"
[367] "volun.2"            "sportact.1"         "relup.4"            "ficode23.1"
"ficode29.3"         "usbread.5"
[373] "usbread.3"          "ficode39.1"         "scwkvfast.5"        "scwkvfast.2"
"scwkvfast.3"        "scwkvfast.4"
[379] "scwhoruage.2"       "scwhoruage.3"       "qfhigh_dv.11"       "mnspno.3"
"arts2a7.1"          "usdairy.4"
[385] "sportact.6"         "fruvege"            "usbread.4"          "ff_ukborn.2"
"hcondn14.1"         "scrundstnd.4"
[391] "scrrely.3"          "scrundstnd.3"       "sf1.y.2"            "scwhorufam.2"
```

```
"kidlang"            "ff_ukborn.5"
[397] "bensta1.1"          "ficode4.1"          "scopfamh.5"          "sports228.1"
"scloutcont.3"       "sclfsato.7"
[403] "nch10to15.2"        "scopfamh.2"         "ff_jbstat.97"        "ivcoop.3"
"natch01.1"          "buno_dv.2"
[409] "lkmove.2"           "ethnic.9"           "ethnic.97"           "nch10to15.1"
"nch5to15.2"         "vote6.4"
[415] "b_pno.y.2"          "sports221.1"        "heritage3.1"         "heritage2.1"
"tbmed.2"            "statins.2"
[421] "ff_bentype19.1"     "jbstat.y.7"         "agegr13_dv.3"        "hhsize.3"
"scopfamd.4"         "ff_jbstat.6"
[427] "scopfamd.3"         "vote1.2"            "sports31.1"          "bfpcok.1"
"ethnic.14"          "sports396.1"
[433] "sports35.1"         "sclfsat2.7"         "sports36.1"          "arts1a5.1"
"scsf6b.2"           "ast"
[439] "ure"                "ecre"               "heritage4.1"         "heritage8.1"
"ficode4.3"          "medtyp4.1"
[445] "ff_jbstat.7"        "uscmm.3"            "ficode3.1"           "ficode3.2"
"lipid.1"            "scfalcdrnk.6"
[451] "sports37.1"         "ficode33.1"         "ficode16.1"          "btype1.1"
"difbpc6.1"          "nchund18resp.3"
[457] "qfhigh_dv.12"       "lenindintv"         "event4s"             "arts2b11.1"
"ethnic.8"           "ficode2.2"
[463] "aidxhh.2"           "fnspno.2"           "b_pno.y.3"           "scsf2a.3"
"ff_bentype11.1"     "relup.3"
[469] "ag16g20.76"         "finfut.2"           "finfut.3"            "usbread.6"
"ethnic.7"           "scsf6b.4"
[475] "agegr13_dv.10"      "scopfamh.4"         "hcondn5.1"           "dorm.6"
"intdaty_dv.2012"    "ff_bentype27.1"
[481] "nch10to15.3"        "nch415resp.4"       "nchund18resp.4"      "allch03.6"
"nch415resp.1"       "ficode9.1"
[487] "sppno.4"            "fiyrinvinc_if.1"    "sclfsat7.6"          "dorm.12"
"qfhigh_dv.5"        "arts1a96.1"
[493] "ficode34.1"         "sports11.1"         "arts2b9.1"           "vpskin.2"
"scfalcdrnk.5"       "htval"
[499] "sex_dv.2"           "mmgsn3"             "mmgsres.2"           "mla1.1"
"ficode28.1"         "hcondn4.1"
[505] "hcondn6.1"          "intdatm_dv.8"       "ff_bentype12.1"      "medtyp5.1"
"scropenup.4"        "nmpsp_dv.2"
[511] "qfhigh_dv.10"       "event1"             "intdatm_dv.4"        "intdatm_dv.9"
"marstat_dv.5"       "marstat_dv.4"
[517] "marstat_dv.3"       "hhsize.2"           "ficode18.1"          "indbdub_xw"
"ficode13.1"         "gor_dv.7"
[523] "dorm.4"             "dorm.5"             "dorm.3"              "gor_dv.2"
"strata.y"           "ff_ukborn.3"
[529] "mmgsok.1"           "wjrel.3"            "netuse.7"            "qfhigh_dv.14"
"intdatm_dv.12"      "ff_jbstat.8"
[535] "scwhorufam.4"       "scwhoruage.4"       "scanyelsetxt.2"      "marstat_dv.2"
"arts2b14.1"         "qfhigh_dv.16"
[541] "scopfamd.2"         "arts2b15.1"         "arts1a7.1"           "scfalcdrnk.3"
"fimnmisc_dv"        "ficode38.1"
[547] "sctimemnuf.2"       "sctimemnuf.3"       "sctimemnuf.4"        "sctimemnuf.5"
"ff_bentype05.1"     "scsf5.2"
[553] "scsf3a.4"           "scsf3b.2"           "scsf5.4"             "chkcoa.2"
"hcondn2.1"          "arts1a2.1"
[559] "nch14resp.1"        "sports38.1"         "sports196.1"         "sports13.1"
"nchund18resp.1"     "scloutcont.2"
[565] "scloutcont.4"       "sclfsato.5"         "vpstimemm"           "arts1a6.1"
"ficode24.1"         "scloutcont.5"
[571] "scloutcont.6"       "dorm.10"            "btype96.1"           "sports219.1"
"mmgstp.3"           "uscmg.2"
[577] "ethnic.13"          "agegr13_dv.8"       "npensioner_dv.3"     "igfi"
"health.y.2"         "ff_ukborn.4"
[583] "vote6.2"            "condna.2"           "access.6"            "sclfsat7.4"
"sclfsato.4"         "sclfsat2.2"
[589] "intdatm_dv.11"      "nunmpsp_dv.1"       "ienddathh"           "nchild_dv.3"
"pulse3"             "hhsize.5"
[595] "scrletdwn.2"        "sports17.1"         "sports111.1"         "ag16g10.46"
```

```
"ag16g10.26"        "ficode1.1"
[601] "omdiast"          "omsyst"            "arts1b15.1"        "arts2a4.1"
"ff_bentype24.1"   "marstat.y.3"
[1] "————————Ordinary Linear Regression (Improved)————————"

Call:
lm(formula = y ~ ., data = as.data.frame(x.data.linear))

Residuals:
     Min       1Q   Median       3Q      Max
-11.4843  -1.6614  -0.0728   1.3920  15.6097

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.18859    0.56066   9.254  < 2e-16 ***
natch03.4       -2.06469    0.75770  -2.725 0.006475 **
nch5to15.1       0.46314    0.17589   2.633 0.008510 **
allch02.5        1.26645    0.45425   2.788 0.005343 **
usdairy.2        0.35523    0.16333   2.175 0.029732 *
usdairy.3        0.46559    0.20115   2.315 0.020711 *
relup.5          2.55120    0.87656   2.910 0.003640 **
hcondn17.1       2.51815    0.46084   5.464 5.10e-08 ***
chargv.2        -0.32268    0.14708  -2.194 0.028327 *
finnow.4         0.98422    0.24462   4.023 5.90e-05 ***
finnow.5         2.16185    0.44173   4.894 1.05e-06 ***
arts2a6.1        0.38652    0.17876   2.162 0.030692 *
sports32.1      -0.61984    0.23729  -2.612 0.009049 **
sportact.5       0.38515    0.17749   2.170 0.030096 *
scsf1.4          0.43316    0.19999   2.166 0.030409 *
scsf1.5          3.06044    0.57500   5.322 1.11e-07 ***
scsf4a.2         2.01130    0.50859   3.955 7.87e-05 ***
scsf4a.3         1.37637    0.29481   4.669 3.19e-06 ***
scsf4a.4         0.61780    0.20742   2.979 0.002924 **
scsf4b.2         3.18090    0.55393   5.742 1.04e-08 ***
scsf4b.3         1.04678    0.30854   3.393 0.000703 ***
scsf4b.4         0.57403    0.20500   2.800 0.005147 **
scsf6a.2         0.84145    0.24814   3.391 0.000707 ***
scsf6a.3         1.73551    0.28252   6.143 9.36e-10 ***
scsf6a.4         3.62675    0.33412  10.855  < 2e-16 ***
scsf6a.5         7.55134    0.53626  14.081  < 2e-16 ***
scsf6b.5         1.60336    0.35643   4.498 7.15e-06 ***
scsf6c.2         4.73450    0.37933  12.481  < 2e-16 ***
scsf6c.3         2.66915    0.20259  13.175  < 2e-16 ***
scsf6c.4         1.42033    0.14642   9.700  < 2e-16 ***
scsf7.2          3.10210    0.37484   8.276  < 2e-16 ***
scsf7.3          1.52607    0.22425   6.805 1.25e-11 ***
scsf7.4          1.27147    0.17883   7.110 1.50e-12 ***
scwhorufam.3     0.79713    0.39910   1.997 0.045894 *
scfalcdrnk.4    -0.26861    0.13062  -2.056 0.039844 *
scfalcdrnk.8    -1.27820    0.34270  -3.730 0.000196 ***
sclfsat1.2      -0.74917    0.28711  -2.609 0.009125 **
sclfsat1.7      -0.47121    0.19911  -2.367 0.018028 *
sclfsato.2       1.83659    0.36733   5.000 6.12e-07 ***
sclfsato.3       1.35722    0.24314   5.582 2.63e-08 ***
schmcont.2       0.69263    0.13212   5.242 1.71e-07 ***
schmcont.3       1.44129    0.23252   6.199 6.62e-10 ***
schmcont.4       1.31781    0.37248   3.538 0.000410 ***
schmcont.5       1.41053    0.47278   2.983 0.002877 **
schmcont.6       2.53909    0.61751   4.112 4.05e-05 ***
scdem2many.2    -0.74309    0.27319  -2.720 0.006571 **
scdem2many.3    -0.78234    0.26260  -2.979 0.002917 **
scdem2many.4    -1.11096    0.29176  -3.808 0.000144 ***
scdem2many.5    -1.39903    0.29098  -4.808 1.61e-06 ***
scdem2many.6    -1.58597    0.30760  -5.156 2.72e-07 ***
scrcritic.2     -0.40271    0.19239  -2.093 0.036430 *
scrcritic.4     -0.36101    0.12632  -2.858 0.004297 **
scrannoy.2       0.52510    0.21891   2.399 0.016522 *
scssup1.4        3.89462    1.74684   2.230 0.025865 *
```

```
ff_bentype26.1    0.89149     0.41206    2.163 0.030595 *
ff_bentype33.1   -1.83436     0.81647   -2.247 0.024744 *
respf16_dv.2      0.72592     0.21612    3.359 0.000794 ***
nunmpsp_dv.2     -5.53216     1.53839   -3.596 0.000329 ***
ficode2.3         2.06090     0.80243    2.568 0.010275 *
ficode14.1       -1.69044     0.45202   -3.740 0.000188 ***
ficode15.1       -1.16008     0.36954   -3.139 0.001713 **
ficode20.1       -0.59947     0.21109   -2.840 0.004549 **
ficode22.1        0.83237     0.23381    3.560 0.000378 ***
diur.1           -0.62596     0.21442   -2.919 0.003539 **
medtyp7.1        -0.50800     0.22352   -2.273 0.023124 *
medtyp12.1        0.94444     0.32780    2.881 0.003996 **
mmgstp.4          1.09679     0.39027    2.810 0.004987 **
vparm.2           0.28141     0.11719    2.401 0.016407 *
vpsens.2          0.82124     0.34599    2.374 0.017691 *
vpsens.3          1.54300     0.73638    2.095 0.036235 *
vpprob1.1       -10.91562     2.20783   -4.944 8.15e-07 ***
vpprob3.1         7.40800     1.37366    5.393 7.57e-08 ***
lfout.11          0.70114     0.32613    2.150 0.031655 *
nurdayw.2         0.43262     0.17939    2.412 0.015951 *
nurdayw.3         0.62238     0.18044    3.449 0.000571 ***
nurdayw.4         0.48291     0.18742    2.577 0.010030 *
nurdayw.5         0.59484     0.19408    3.065 0.002200 **
nurdayw.6         0.75187     0.25637    2.933 0.003389 **
jbstat.y.8        1.67507     0.49041    3.416 0.000646 ***
sf1.y.5          -1.28474     0.48154   -2.668 0.007678 **
uscmm.1          -0.95355     0.42538   -2.242 0.025071 *
fiyrinvinc_dv    -0.20043     0.06673   -3.004 0.002693 **
intdatm_dv.2      0.46164     0.23252    1.985 0.047207 *
intdatm_dv.6     -0.51020     0.20254   -2.519 0.011829 *
nnsib_dv.1       -2.42640     1.14449   -2.120 0.034096 *
qfhigh_dv.96     -0.36303     0.15081   -2.407 0.016145 *
racel_dv         -0.12793     0.05171   -2.474 0.013433 *
hiqual_dv.y.4     0.34820     0.15066    2.311 0.020897 *
hiqual_dv.y.5     0.40903     0.19343    2.115 0.034564 *
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1       1

Residual standard error: 2.957 on 2564 degrees of freedom
Multiple R-squared:  0.695,     Adjusted R-squared:  0.6845
F-statistic: 66.39 on 88 and 2564 DF,  p-value: < 2.2e-16

AIC: 13370.6971
MSE: 8.4496
[1] "The MSE of the predicted values are of 11.0779"
[1] "The Linear Model predicts exactly with accuracy of 0.1652"
[1] "The Linear Model predicts within a confidence interval with accuracy of 0.4197"
[1] "------------Elastic Net Regression------------"
763 x 1 sparse Matrix of class "dgCMatrix", with 63 entries
           names Estimate_Coefs
1    (Intercept)  13.0429379133
2  nchund18resp.2   0.1050112470
3      nch5to15.1   0.0094227740
4         relup.5   0.1927910771
5      hcondn17.1   1.4911004853
6        finnow.4   0.6356914104
7        finnow.5   1.3528168921
8      sports32.1  -0.1123650005
9         scsf1.5   0.6778287547
10       scsf4a.2   1.2224833803
11       scsf4a.3   0.4138647663
12       scsf4a.4  -0.0673643900
13       scsf4a.5  -1.0579078897
14       scsf4b.2   1.7832048860
15       scsf4b.3   0.1296521921
16       scsf4b.5  -0.8177267420
17       scsf6a.3   0.6720444089
18       scsf6a.4   2.4652166646
```

```
19          scsf6a.5    5.6387410582
20          scsf6b.4    0.2869480842
21          scsf6b.5    1.2696024177
22          scsf6c.2    2.9986238344
23          scsf6c.3    0.9569024712
24          scsf6c.5   -1.8160797439
25           scsf7.2    1.6243604713
26           scsf7.5   -1.5628186638
27     scfalcdrnk.2    0.0288047130
28     scfalcdrnk.8   -0.1199763838
29        sclfsat7.3    0.0419511975
30        sclfsat7.7   -0.0700506434
31        sclfsato.2    1.0036824607
32        sclfsato.3    1.0523841815
33        sclfsato.6   -0.3241134655
34        sclfsato.7   -0.6644138995
35        schmcont.2    0.0879498125
36        schmcont.3    0.5289183940
37        schmcont.4    0.2187476235
38        schmcont.5    0.1745775253
39        schmcont.6    0.5248072053
40      scloutcont.6   -0.2558763284
41      scdem2many.5   -0.0667307921
42      scdem2many.6   -0.1730848528
43       scwkvfast.6   -0.0434610289
44      scrundstnd.3    0.0659876648
45       scropenup.4    0.0409904597
46       scrcritic.4   -0.0311360012
47       scrletdwn.2    0.1313609060
48       scrletdwn.4   -0.0289020127
49        scrannoy.2    0.0273222973
50         scssup1.4    0.3522815661
51     nunmpsp_dv.2   -0.0463337783
52        ficode33.1    0.2624450640
53            diur.1   -0.1025739043
54         vpprob1.1   -2.2261418543
55         vpprob3.1    1.2732456933
56         nurdayw.1   -0.1463398029
57            omsyst   -0.0396254398
58        ag16g10.46    0.0009486183
59        jbstat.y.8    0.5983814600
60      intdatm_dv.6   -0.1072784003
61        nnsib_dv.1   -0.3655455360
62       rach16_dv.2   -0.0204310019
63    hiqual_dv.y.9   -0.1243359027
[1] "The MSE of the predicted values of the best fit model is 9.4568"
[1] "The Alpha of the best fit model is 0.9"
[1] "The Elastic Net Model predicts exactly with accuracy of 0.1697"
[1] "----------Timer Results----------"
   user  system elapsed
3189.06   26.47 3217.34
```

10.2.20 w3indresp console

```
[1] "----------Initial Checks----------"
[1] "116063193 NA cells were found across the entire dataset (76.68% of data as NA)"
[1] "----------Data Type Checks----------"
[1] "0 variables recoded since all their entries aren't numeric or NA"
[1] "NOTE: algorithim recodes categorical data alphabetically e.g (female = 1, male = 2)"
character(0)
[1] "----------Low Data Removal----------"
[1] "2495 variables removed since they had >= 'naPercent' (default 20%) NA values"
   [1] "pid"                "childpno"           "lvwhy"              "lvmthp"
"lvyrp"              "liwho"
   [7] "lihow"              "mvever"             "mvmnth"             "mvyr"
"mlstatchk"          "mlstat"
```

```
  [13] "drive"             "caruse"            "britid"            "ukborn"
"plbornc"          "yr2uk4"
  [19] "citzn1"            "citzn2"            "citzn3"            "qfhigh"
"qualoc"           "qfvoc1"
  [25] "qfvoc2"            "qfvoc3"            "qfvoc4"            "qfvoc5"
"qfvoc6"           "qfvoc7"
  [31] "qfvoc8"            "qfvoc9"            "qfvoc10"           "qfvoc11"
"qfvoc12"          "qfvoc13"
  [37] "qfvoc14"           "qfvoc15"           "qfvoc96"           "school"
"scend"            "schlloc"
  [43] "schok"             "fenow"             "feend"             "j1none"
"j1semp"           "j1boss"
  [49] "j1mngr"            "edtype"            "edasp"             "lvscdo"
"ahvwell"          "fedlik"
  [55] "fednt"             "futureint"         "lvhm"              "ocimpa"
"ocimpb"           "ocimpe"
  [61] "ocimpf"            "ocimpi"            "ocimpk"            "ocimpl"
"futra"            "futrb"
  [67] "futrc"             "futrd"             "futre"             "futrf"
"futrg"            "futrh"
  [73] "futri"             "futrj"             "futrk"             "futrl"
"paju"             "maju"
  [79] "pacob"             "payruk"            "macob"             "mayruk"
"natid1"           "natid2"
  [85] "natid3"            "natid4"            "natid5"            "natid6"
"natid97"          "racel"
  [91] "racelo_code"       "racelt"            "racelwt"           "racelmt"
"racelat"          "racelbt"
  [97] "racelot_code"      "oprlg"             "oprlg0ni"          "nirel"
"niact"            "oprlg0"
 [103] "oprlg1"            "unsafe1"           "unsafe2"           "unsafe3"
"unsafe4"          "unsafe5"
 [109] "unsafe6"           "unsafe7"           "unsafe8"           "unsafe9"
"unsafe10"         "unsafe11"
 [115] "unsafe96"          "unsafe97"          "unsafeo_code"      "resunsafe1_1"
"resunsafe2_1"     "resunsafe3_1"
 [121] "resunsafe4_1"      "resunsafe5_1"      "resunsafe6_1"      "resunsafe7_1"
"resunsafe8_1"     "resunsafe9_1"
 [127] "resunsafe96_1"     "resunsafe97_1"     "resunsafe1_2"      "resunsafe2_2"
"resunsafe3_2"     "resunsafe4_2"
 [133] "resunsafe5_2"      "resunsafe6_2"      "resunsafe7_2"      "resunsafe8_2"
"resunsafe9_2"     "resunsafe96_2"
 [139] "resunsafe97_2"     "resunsafe1_3"      "resunsafe2_3"      "resunsafe3_3"
"resunsafe4_3"     "resunsafe5_3"
 [145] "resunsafe6_3"      "resunsafe7_3"      "resunsafe8_3"      "resunsafe9_3"
"resunsafe96_3"    "resunsafe97_3"
 [151] "resunsafe1_4"      "resunsafe2_4"      "resunsafe3_4"      "resunsafe4_4"
"resunsafe5_4"     "resunsafe6_4"
 [157] "resunsafe7_4"      "resunsafe8_4"      "resunsafe9_4"      "resunsafe96_4"
"resunsafe97_4"    "resunsafe1_5"
 [163] "resunsafe2_5"      "resunsafe3_5"      "resunsafe4_5"      "resunsafe5_5"
"resunsafe6_5"     "resunsafe7_5"
 [169] "resunsafe8_5"      "resunsafe9_5"      "resunsafe96_5"     "resunsafe97_5"
"resunsafe1_6"     "resunsafe2_6"
 [175] "resunsafe3_6"      "resunsafe4_6"      "resunsafe5_6"      "resunsafe6_6"
"resunsafe7_6"     "resunsafe8_6"
 [181] "resunsafe9_6"      "resunsafe96_6"     "resunsafe97_6"     "resunsafe1_7"
"resunsafe2_7"     "resunsafe3_7"
 [187] "resunsafe4_7"      "resunsafe5_7"      "resunsafe6_7"      "resunsafe7_7"
"resunsafe8_7"     "resunsafe9_7"
 [193] "resunsafe96_7"     "resunsafe97_7"     "resunsafe1_8"      "resunsafe2_8"
"resunsafe3_8"     "resunsafe4_8"
 [199] "resunsafe5_8"      "resunsafe6_8"      "resunsafe7_8"      "resunsafe8_8"
"resunsafe9_8"     "resunsafe96_8"
 [205] "resunsafe97_8"     "resunsafe1_9"      "resunsafe2_9"      "resunsafe3_9"
"resunsafe4_9"     "resunsafe5_9"
 [211] "resunsafe6_9"      "resunsafe7_9"      "resunsafe8_9"      "resunsafe9_9"
"resunsafe96_9"    "resunsafe97_9"
```

```
 [217] "resunsafe1_10"     "resunsafe2_10"     "resunsafe3_10"     "resunsafe4_10"    
"resunsafe5_10"     "resunsafe6_10"    
 [223] "resunsafe7_10"     "resunsafe8_10"     "resunsafe9_10"     "resunsafe96_10"   
"resunsafe97_10"    "resunsafe1_11"    
 [229] "resunsafe2_11"     "resunsafe3_11"     "resunsafe4_11"     "resunsafe5_11"    
"resunsafe6_11"     "resunsafe7_11"    
 [235] "resunsafe8_11"     "resunsafe9_11"     "resunsafe96_11"    "resunsafe97_11"   
"resunsafe1_12"     "resunsafe2_12"    
 [241] "resunsafe3_12"     "resunsafe4_12"     "resunsafe5_12"     "resunsafe6_12"    
"resunsafe7_12"     "resunsafe8_12"    
 [247] "resunsafe9_12"     "resunsafe96_12"    "resunsafe97_12"    "avoidance1"       
"avoidance2"        "avoidance3"       
 [253] "avoidance4"        "avoidance5"        "avoidance6"        "avoidance7"       
"avoidance8"        "avoidance9"       
 [259] "avoidance10"       "avoidance11"       "avoidance96"       "avoidance97"      
"avoidanceo_code"   "resavoid1_1"      
 [265] "resavoid2_1"       "resavoid3_1"       "resavoid4_1"       "resavoid5_1"      
"resavoid6_1"       "resavoid7_1"      
 [271] "resavoid8_1"       "resavoid9_1"       "resavoid96_1"      "resavoid97_1"     
"resavoid1_2"       "resavoid2_2"      
 [277] "resavoid3_2"       "resavoid4_2"       "resavoid5_2"       "resavoid6_2"      
"resavoid7_2"       "resavoid8_2"      
 [283] "resavoid9_2"       "resavoid96_2"      "resavoid97_2"      "resavoid1_3"      
"resavoid2_3"       "resavoid3_3"      
 [289] "resavoid4_3"       "resavoid5_3"       "resavoid6_3"       "resavoid7_3"      
"resavoid8_3"       "resavoid9_3"      
 [295] "resavoid96_3"      "resavoid97_3"      "resavoid1_4"       "resavoid2_4"      
"resavoid3_4"       "resavoid4_4"      
 [301] "resavoid5_4"       "resavoid6_4"       "resavoid7_4"       "resavoid8_4"      
"resavoid9_4"       "resavoid96_4"     
 [307] "resavoid97_4"      "resavoid1_5"       "resavoid2_5"       "resavoid3_5"      
"resavoid4_5"       "resavoid5_5"      
 [313] "resavoid6_5"       "resavoid7_5"       "resavoid8_5"       "resavoid9_5"      
"resavoid96_5"      "resavoid97_5"     
 [319] "resavoid1_6"       "resavoid2_6"       "resavoid3_6"       "resavoid4_6"      
"resavoid5_6"       "resavoid6_6"      
 [325] "resavoid7_6"       "resavoid8_6"       "resavoid9_6"       "resavoid96_6"     
"resavoid97_6"      "resavoid1_7"      
 [331] "resavoid2_7"       "resavoid3_7"       "resavoid4_7"       "resavoid5_7"      
"resavoid6_7"       "resavoid7_7"      
 [337] "resavoid8_7"       "resavoid9_7"       "resavoid96_7"      "resavoid97_7"     
"resavoid1_8"       "resavoid2_8"      
 [343] "resavoid3_8"       "resavoid4_8"       "resavoid5_8"       "resavoid6_8"      
"resavoid7_8"       "resavoid8_8"      
 [349] "resavoid9_8"       "resavoid96_8"      "resavoid97_8"      "resavoid1_9"      
"resavoid2_9"       "resavoid3_9"      
 [355] "resavoid4_9"       "resavoid5_9"       "resavoid6_9"       "resavoid7_9"      
"resavoid8_9"       "resavoid9_9"      
 [361] "resavoid96_9"      "resavoid97_9"      "resavoid1_10"      "resavoid2_10"     
"resavoid3_10"      "resavoid4_10"     
 [367] "resavoid5_10"      "resavoid6_10"      "resavoid7_10"      "resavoid8_10"     
"resavoid9_10"      "resavoid96_10"    
 [373] "resavoid97_10"     "resavoid1_11"      "resavoid2_11"      "resavoid3_11"     
"resavoid4_11"      "resavoid5_11"     
 [379] "resavoid6_11"      "resavoid7_11"      "resavoid8_11"      "resavoid9_11"     
"resavoid96_11"     "resavoid97_11"    
 [385] "resavoid1_12"      "resavoid2_12"      "resavoid3_12"      "resavoid4_12"     
"resavoid5_12"      "resavoid6_12"     
 [391] "resavoid7_12"      "resavoid8_12"      "resavoid9_12"      "resavoid96_12"    
"resavoid97_12"     "insulted1"        
 [397] "insulted2"         "insulted3"         "insulted4"         "insulted5"        
"insulted6"         "insulted7"        
 [403] "insulted8"         "insulted9"         "insulted10"        "insulted11"       
"insulted96"        "insulted97"       
 [409] "insultedo_code"    "resinsulted1_1"    "resinsulted2_1"    "resinsulted3_1"   
"resinsulted4_1"    "resinsulted5_1"   
 [415] "resinsulted6_1"    "resinsulted7_1"    "resinsulted8_1"    "resinsulted9_1"   
"resinsulted96_1"   "resinsulted97_1"  
```

```
 [421] "resinsulted1_2"  "resinsulted2_2"  "resinsulted3_2"  "resinsulted4_2"
"resinsulted5_2"  "resinsulted6_2"
 [427] "resinsulted7_2"  "resinsulted8_2"  "resinsulted9_2"  "resinsulted96_2"
"resinsulted97_2"  "resinsulted1_3"
 [433] "resinsulted2_3"  "resinsulted3_3"  "resinsulted4_3"  "resinsulted5_3"
"resinsulted6_3"  "resinsulted7_3"
 [439] "resinsulted8_3"  "resinsulted9_3"  "resinsulted96_3"  "resinsulted97_3"
"resinsulted1_4"  "resinsulted2_4"
 [445] "resinsulted3_4"  "resinsulted4_4"  "resinsulted5_4"  "resinsulted6_4"
"resinsulted7_4"  "resinsulted8_4"
 [451] "resinsulted9_4"  "resinsulted96_4"  "resinsulted97_4"  "resinsulted1_5"
"resinsulted2_5"  "resinsulted3_5"
 [457] "resinsulted4_5"  "resinsulted5_5"  "resinsulted6_5"  "resinsulted7_5"
"resinsulted8_5"  "resinsulted9_5"
 [463] "resinsulted96_5"  "resinsulted97_5"  "resinsulted1_6"  "resinsulted2_6"
"resinsulted3_6"  "resinsulted4_6"
 [469] "resinsulted5_6"  "resinsulted6_6"  "resinsulted7_6"  "resinsulted8_6"
"resinsulted9_6"  "resinsulted96_6"
 [475] "resinsulted97_6"  "resinsulted1_7"  "resinsulted2_7"  "resinsulted3_7"
"resinsulted4_7"  "resinsulted5_7"
 [481] "resinsulted6_7"  "resinsulted7_7"  "resinsulted8_7"  "resinsulted9_7"
"resinsulted96_7"  "resinsulted97_7"
 [487] "resinsulted1_8"  "resinsulted2_8"  "resinsulted3_8"  "resinsulted4_8"
"resinsulted5_8"  "resinsulted6_8"
 [493] "resinsulted7_8"  "resinsulted8_8"  "resinsulted9_8"  "resinsulted96_8"
"resinsulted97_8"  "resinsulted1_9"
 [499] "resinsulted2_9"  "resinsulted3_9"  "resinsulted4_9"  "resinsulted5_9"
"resinsulted6_9"  "resinsulted7_9"
 [505] "resinsulted8_9"  "resinsulted9_9"  "resinsulted96_9"  "resinsulted97_9"
"resinsulted1_10"  "resinsulted2_10"
 [511] "resinsulted3_10"  "resinsulted4_10"  "resinsulted5_10"  "resinsulted6_10"
"resinsulted7_10"  "resinsulted8_10"
 [517] "resinsulted9_10"  "resinsulted96_10" "resinsulted97_10" "resinsulted1_11"
"resinsulted2_11"  "resinsulted3_11"
 [523] "resinsulted4_11"  "resinsulted5_11"  "resinsulted6_11"  "resinsulted7_11"
"resinsulted8_11"  "resinsulted9_11"
 [529] "resinsulted96_11" "resinsulted97_11" "resinsulted1_12"  "resinsulted2_12"
"resinsulted3_12"  "resinsulted4_12"
 [535] "resinsulted5_12"  "resinsulted6_12"  "resinsulted7_12"  "resinsulted8_12"
"resinsulted9_12"  "resinsulted96_12"
 [541] "resinsulted97_12" "attacked1"  "attacked2"  "attacked3"
"attacked4"  "attacked5"
 [547] "attacked6"  "attacked7"  "attacked8"  "attacked9"
"attacked10"  "attacked11"
 [553] "attacked96"  "attacked97"  "attackedo_code"  "resattacked1_1"
"resattacked2_1"  "resattacked3_1"
 [559] "resattacked4_1"  "resattacked5_1"  "resattacked6_1"  "resattacked7_1"
"resattacked8_1"  "resattacked9_1"
 [565] "resattacked96_1"  "resattacked97_1"  "resattacked1_2"  "resattacked2_2"
"resattacked3_2"  "resattacked4_2"
 [571] "resattacked5_2"  "resattacked6_2"  "resattacked7_2"  "resattacked8_2"
"resattacked9_2"  "resattacked96_2"
 [577] "resattacked97_2"  "resattacked1_3"  "resattacked2_3"  "resattacked3_3"
"resattacked4_3"  "resattacked5_3"
 [583] "resattacked6_3"  "resattacked7_3"  "resattacked8_3"  "resattacked9_3"
"resattacked96_3"  "resattacked97_3"
 [589] "resattacked1_4"  "resattacked2_4"  "resattacked3_4"  "resattacked4_4"
"resattacked5_4"  "resattacked6_4"
 [595] "resattacked7_4"  "resattacked8_4"  "resattacked9_4"  "resattacked96_4"
"resattacked97_4"  "resattacked1_5"
 [601] "resattacked2_5"  "resattacked3_5"  "resattacked4_5"  "resattacked5_5"
"resattacked6_5"  "resattacked7_5"
 [607] "resattacked8_5"  "resattacked9_5"  "resattacked96_5"  "resattacked97_5"
"resattacked1_6"  "resattacked2_6"
 [613] "resattacked3_6"  "resattacked4_6"  "resattacked5_6"  "resattacked6_6"
"resattacked7_6"  "resattacked8_6"
 [619] "resattacked9_6"  "resattacked96_6"  "resattacked97_6"  "resattacked1_7"
"resattacked2_7"  "resattacked3_7"
```

```
 [625] "resattacked4_7"   "resattacked5_7"   "resattacked6_7"   "resattacked7_7"
"resattacked8_7"   "resattacked9_7"
 [631] "resattacked96_7"  "resattacked97_7"  "resattacked1_8"   "resattacked2_8"
"resattacked3_8"   "resattacked4_8"
 [637] "resattacked5_8"   "resattacked6_8"   "resattacked7_8"   "resattacked8_8"
"resattacked9_8"   "resattacked96_8"
 [643] "resattacked97_8"  "resattacked1_9"   "resattacked2_9"   "resattacked3_9"
"resattacked4_9"   "resattacked5_9"
 [649] "resattacked6_9"   "resattacked7_9"   "resattacked8_9"   "resattacked9_9"
"resattacked96_9"  "resattacked97_9"
 [655] "resattacked1_10"  "resattacked2_10"  "resattacked3_10"  "resattacked4_10"
"resattacked5_10"  "resattacked6_10"
 [661] "resattacked7_10"  "resattacked8_10"  "resattacked9_10"  "resattacked96_10" "resattacked97_10"
 [667] "resattacked2_11"  "resattacked3_11"  "resattacked4_11"  "resattacked5_11"
"resattacked6_11"  "resattacked7_11"
 [673] "resattacked8_11"  "resattacked9_11"  "resattacked96_11" "resattacked97_11" "resattacked1_12"
"resattacked2_12"
 [679] "resattacked3_12"  "resattacked4_12"  "resattacked5_12"  "resattacked6_12"
"resattacked7_12"  "resattacked8_12"
 [685] "resattacked9_12"  "resattacked96_12" "resattacked97_12" "servaccy1"
"servaccy2"        "servaccy3"
 [691] "servaccy4"        "servaccy5"        "servaccy6"        "servaccy7"
"servaccy8"        "servaccy9"
 [697] "servaccy10"       "servaccy11"       "servaccy12"       "servaccy13"
"servaccy14"       "servaccy15"
 [703] "servaccy97"       "sattransy1"       "sattransy2"       "sattransy3"
"sattransy4"       "sattransy5"
 [709] "sattransy6"       "sattransy7"       "sattransy8"       "sattransy9"
"sattransy10"      "sattransy97"
 [715] "crworb"           "netcht"           "visfrndsy1"       "visfrndsy2"
"visfrndsy3"       "visfrndsy4"
 [721] "visfrndsy5"       "visfrndsy6"       "visfrndsy7"       "visfrndsy8"
"visfrndsy9"       "visfrndsy10"
 [727] "visfrndsy11"      "visfrndsy12"      "visfrndsy13"      "visfrndsy14"
"visfrndsy97"      "netrl_1"
 [733] "netmet_1"         "netweb_1"         "nettalk1_1"       "nettalk2_1"
"nettalk3_1"       "nettalk4_1"
 [739] "nettalk5_1"       "nettalk6_1"       "nettalk7_1"       "nettalk8_1"
"nettalk9_1"       "nettalk10_1"
 [745] "nettalk11_1"      "nettalk12_1"      "nettalk13_1"      "nettalk97_1"
"nettalk96_1"      "netdo1_1"
 [751] "netdo2_1"         "netdo3_1"         "netdo4_1"         "netdo5_1"
"netdo6_1"         "netdo7_1"
 [757] "netdo8_1"         "netdo9_1"         "netdo10_1"        "netdo97_1"
"netdo96_1"        "netrl_2"
 [763] "netmet_2"         "netweb_2"         "nettalk1_2"       "nettalk2_2"
"nettalk3_2"       "nettalk4_2"
 [769] "nettalk5_2"       "nettalk6_2"       "nettalk7_2"       "nettalk8_2"
"nettalk9_2"       "nettalk10_2"
 [775] "nettalk11_2"      "nettalk12_2"      "nettalk13_2"      "nettalk97_2"
"nettalk96_2"      "netdo1_2"
 [781] "netdo2_2"         "netdo3_2"         "netdo4_2"         "netdo5_2"
"netdo6_2"         "netdo7_2"
 [787] "netdo8_2"         "netdo9_2"         "netdo10_2"        "netdo97_2"
"netdo96_2"        "netsx_3"
 [793] "netwr_3"          "netrl_3"          "netag_3"          "netkn_3"
"netph_3"          "netlv_3"
 [799] "netjb_3"          "netet_3"          "netmet_3"         "netweb_3"
"nettalk1_3"       "nettalk2_3"
 [805] "nettalk3_3"       "nettalk4_3"       "nettalk5_3"       "nettalk6_3"
"nettalk7_3"       "nettalk8_3"
 [811] "nettalk9_3"       "nettalk10_3"      "nettalk11_3"      "nettalk12_3"
"nettalk13_3"      "nettalk97_3"
 [817] "nettalk96_3"      "netdo1_3"         "netdo2_3"         "netdo3_3"
"netdo4_3"         "netdo5_3"
 [823] "netdo6_3"         "netdo7_3"         "netdo8_3"         "netdo9_3"
"netdo10_3"        "netdo97_3"
 [829] "netdo96_3"        "orgm1"            "orgm2"            "orgm3"
```

```
"orgm4"            "orgm5"
 [835] "orgm6"            "orgm7"            "orgm8"            "orgm9"
"orgm10"           "orgm11"
 [841] "orgm12"           "orgm13"           "orgm14"           "orgm15"
"orgm16"           "orgm96"
 [847] "orgmt1"           "orgmt2"           "orgmt3"           "orgmt4"
"orgmt5"           "orgmt6"
 [853] "orgmt7"           "orgmt8"           "orgmt9"           "orgmt10"
"orgmt11"          "orgmt12"
 [859] "orgmt13"          "orgmt14"          "orgmt15"          "orgmt16"
"orgmt96"          "orgat1"
 [865] "orgat2"           "orgat3"           "orgat4"           "orgat5"
"orgat6"           "orgat7"
 [871] "orgat8"           "orgat9"           "orgat10"          "orgat11"
"orgat12"          "orgat13"
 [877] "orgat14"          "orgat15"          "orgat16"          "orgat96"
"hospc1"           "hospdc1"
 [883] "hospc2"           "hospdc2"          "hospc3"           "hospdc3"
"hospc4"           "hospdc4"
 [889] "hospc5"           "hospdc5"          "hospc6"           "hospdc6"
"hospc7"           "hospdc7"
 [895] "hospc8"           "hospdc8"          "disdif1"          "disdif2"
"disdif3"          "disdif4"
 [901] "disdif5"          "disdif6"          "disdif7"          "disdif8"
"disdif9"          "disdif10"
 [907] "disdif11"         "disdif12"         "disdif96"         "hcond1"
"hcond2"           "hcond3"
 [913] "hcond4"           "hcond5"           "hcond6"           "hcond7"
"hcond8"           "hcond9"
 [919] "hcond10"          "hcond11"          "hcond12"          "hcond13"
"hcond14"          "hcond15"
 [925] "hcond16"          "hcond17"          "hcond96"          "hconds01"
"hconds02"         "hconds03"
 [931] "hconds04"         "hconds05"         "hconds08"         "hconds09"
"hconds10"         "hconds11"
 [937] "hconds12"         "hconds13"         "hconds14"         "hconds15"
"hconds16"         "hconds17"
 [943] "hconda01"         "hconda02"         "hconda03"         "hconda04"
"hconda05"         "hconda06"
 [949] "hconda07"         "hconda08"         "hconda09"         "hconda10"
"hconda11"         "hconda12"
 [955] "hconda13"         "hconda14"         "hconda15"         "hconda16"
"hconda17"         "aidhua1"
 [961] "aidhua2"          "aidhua3"          "aidhua4"          "aidhua5"
"aidhua6"          "aidhua7"
 [967] "aidhua8"          "aidhua9"          "aidhua10"         "aidhua11"
"aidhua12"         "aidhua13"
 [973] "aidhua14"         "aidhua15"         "aidhua16"         "naidxhh"
"aidhu1"           "aidhu2"
 [979] "aidhrs"           "aideft"           "lcohnpi"          "coh1bm"
"coh1by"           "coh1mr"
 [985] "coh1em"           "coh1ey"           "nmar"             "lmar1m"
"lmar1y"           "ladopt"
 [991] "lnadopt"          "lprnt"            "lnprnt"           "ch1by4"
"movy11"           "movy12"
 [997] "movy13"           "movy14"           "movy15"           "movy16"
 [ reached getOption("max.print") -- omitted 1495 entries ]
[1] "------------Low Level Removal------------"
[1] "If a level is removed from a variable you wish to keep, reccomended to manually merge levels toe
[1] "level 7 in hhorig removed, 1 observations found"
[1] "level 7 in memorig removed, 0 observations found"
[1] "level 5 in nch14resp removed, 4 observations found"
[1] "level 6 in nch14resp removed, 2 observations found"
[1] "level 2 in nch3resp removed, 1 observations found"
[1] "level 2 in nch5resp removed, 2 observations found"
[1] "level 5 in nch415resp removed, 1 observations found"
[1] "level 6 in nch415resp removed, 0 observations found"
[1] "level 5 in nchresp removed, 2 observations found"
```

```
[1] "level 6 in nchresp removed, 0 observations found"
[1] "level 5 in nchund18resp removed, 3 observations found"
[1] "level 6 in nchund18resp removed, 0 observations found"
[1] "level 7 in natch01 removed, 2 observations found"
[1] "level 9 in natch01 removed, 1 observations found"
[1] "level 2 in natch02 removed, 3 observations found"
[1] "level 7 in natch02 removed, 3 observations found"
[1] "level 8 in natch02 removed, 0 observations found"
[1] "level 3 in natch03 removed, 0 observations found"
[1] "level 9 in natch03 removed, 0 observations found"
[1] "level 7 in natch04 removed, 2 observations found"
[1] "level 10 in natch04 removed, 0 observations found"
[1] "level 6 in natch05 removed, 0 observations found"
[1] "level 8 in natch05 removed, 0 observations found"
[1] "level 8 in natch06 removed, 4 observations found"
[1] "level 6 in nnatch removed, 0 observations found"
[1] "level 4 in nadoptch removed, 0 observations found"
[1] "level 5 in nadoptch removed, 1 observations found"
[1] "level 1 in adoptch01 removed, 4 observations found"
[1] "level 6 in adoptch01 removed, 1 observations found"
[1] "level 3 in adoptch02 removed, 2 observations found"
[1] "level 5 in adoptch02 removed, 3 observations found"
[1] "level 6 in adoptch02 removed, 1 observations found"
[1] "level 4 in adoptch03 removed, 0 observations found"
[1] "level 6 in adoptch04 removed, 0 observations found"
[1] "level 7 in adoptch05 removed, 0 observations found"
[1] "level 5 in nchunder16 removed, 2 observations found"
[1] "level 6 in nchunder16 removed, 0 observations found"
[1] "level 5 in nch5to15 removed, 0 observations found"
[1] "level 4 in nch10to15 removed, 1 observations found"
[1] "level 2 in nch10 removed, 4 observations found"
[1] "level 1 in allch01 removed, 3 observations found"
[1] "level 7 in allch01 removed, 0 observations found"
[1] "level 9 in allch01 removed, 0 observations found"
[1] "level 7 in allch02 removed, 2 observations found"
[1] "level 8 in allch02 removed, 0 observations found"
[1] "level 8 in allch03 removed, 0 observations found"
[1] "level 9 in allch03 removed, 0 observations found"
[1] "level 5 in allch04 removed, 3 observations found"
[1] "level 7 in allch04 removed, 1 observations found"
[1] "level 8 in allch04 removed, 0 observations found"
[1] "level 10 in allch04 removed, 0 observations found"
[1] "level 6 in allch05 removed, 0 observations found"
[1] "level 7 in allch05 removed, 0 observations found"
[1] "level 8 in allch05 removed, 0 observations found"
[1] "level 8 in allch06 removed, 0 observations found"
[1] "level 9 in jbstat removed, 1 observations found"
[1] "level 10 in jbstat removed, 3 observations found"
[1] "level 11 in jbstat removed, 1 observations found"
[1] "level 5 in relup removed, 3 observations found"
[1] "level 2 in nnewborn removed, 0 observations found"
[1] "level 1 in hcondn3 removed, 2 observations found"
[1] "level 1 in hcondn9 removed, 4 observations found"
[1] "level 1 in bensta3 removed, 2 observations found"
[1] "level 10 in newsmain removed, 3 observations found"
[1] "level 12 in newsmain removed, 3 observations found"
[1] "level 9 in tvm2 removed, 4 observations found"
[1] "level 1 in sub7_1 removed, 1 observations found"
[1] "level 3 in sub7_1 removed, 2 observations found"
[1] "level 73 in sub7_1 removed, 2 observations found"
[1] "level 86 in sub7_1 removed, 2 observations found"
[1] "level 89 in sub7_1 removed, 2 observations found"
[1] "level 91 in sub7_1 removed, 1 observations found"
[1] "level 96 in sub7_1 removed, 2 observations found"
[1] "level 97 in sub7_1 removed, 2 observations found"
[1] "level 4 in clangab removed, 4 observations found"
[1] "level 4 in ivcoop removed, 3 observations found"
[1] "level 4 in undqus removed, 0 observations found"
```

```
[1] "level 2 in ivlieng removed, 3 observations found"
[1] "level 97 in ivlitrans removed, 0 observations found"
[1] "level 5 in hgbiof removed, 2 observations found"
[1] "level 2 in origadd removed, 2 observations found"
[1] "level 1 in fiyrinvinc_tc removed, 1 observations found"
[1] "level 5 in pn1pno removed, 4 observations found"
[1] "level 5 in pn2pno removed, 1 observations found"
[1] "level 5 in pns1pno removed, 0 observations found"
[1] "level 5 in pns2pno removed, 0 observations found"
[1] "level 6 in pns2pno removed, 1 observations found"
[1] "level 10 in ff_jbstat removed, 4 observations found"
[1] "level 1 in ff_bentype06 removed, 3 observations found"
[1] "level 1 in ff_bentype21 removed, 1 observations found"
[1] "level 1 in ff_bentype25 removed, 1 observations found"
[1] "level 1 in ff_bentype30 removed, 2 observations found"
[1] "level 1 in ff_bentype31 removed, 3 observations found"
[1] "level 1 in ff_bentype32 removed, 2 observations found"
[1] "level 1 in ff_bentype34 removed, 4 observations found"
[1] "level 1 in ff_bentype35 removed, 1 observations found"
[1] "level 2 in ngrp_dv removed, 2 observations found"
[1] "level 4 in nnssib_dv removed, 2 observations found"
[1] "level 5 in nnssib_dv removed, 2 observations found"
[1] "level 1 in xtra5min_dv removed, 2 observations found"
[1] "level 6 in buno_dv removed, 1 observations found"
[1] "level 5 in nchild_dv removed, 0 observations found"
[1] "level 6 in nchild_dv removed, 0 observations found"
[1] "level 5 in hrpno removed, 2 observations found"
[1] "level 6 in hrpno removed, 4 observations found"
[1] "level 6 in ppno removed, 1 observations found"
[1] "level 8 in ppno removed, 0 observations found"
[1] "level 5 in sppno removed, 1 observations found"
[1] "level 6 in sppno removed, 0 observations found"
[1] "level 8 in sppno removed, 0 observations found"
[1] "level 5 in fnpno removed, 0 observations found"
[1] "level 5 in fnspno removed, 0 observations found"
[1] "level 6 in fnspno removed, 0 observations found"
[1] "level 5 in mnpno removed, 0 observations found"
[1] "level 5 in mnspno removed, 0 observations found"
[1] "level 1 in grfpno removed, 1 observations found"
[1] "level 2 in grfpno removed, 1 observations found"
[1] "level 2 in grmpno removed, 1 observations found"
[1] "level 3 in grmpno removed, 1 observations found"
[1] "level 4 in grmpno removed, 1 observations found"
[1] "level 4 in nnmpsp_dv removed, 1 observations found"
[1] "level 1 in big5c_dv removed, 3 observations found"
[1] "level 0 in cgwri_dv removed, 2 observations found"
[1] "level 0 in cgs7cs_dv removed, 3 observations found"
[1] "level 0.177194967865944 in ind5mus_xw removed, 0 observations found"
[1] "level 5 in ndepchl_dv removed, 0 observations found"
[1] "level 6 in ndepchl_dv removed, 0 observations found"
[1] "level 3 in nmpsp_dv removed, 4 observations found"
[1] "level 1 in big5a_dv removed, 3 observations found"
[1] "level 4 in nnsib_dv removed, 0 observations found"
[1] "level 5 in nnsib_dv removed, 0 observations found"
[1] "131 total levels removed from 84 different variables. In total 179 observations deleted"
[1] "———————Variance 0 Check——————"
[1] "101 variables removed since their new variance was 0"
  [1] "ivfio"          "ioutcome"       "adstatus"       "natch06"
"natch07"        "natch08"        "natch09"
  [8] "natch10"        "natch11"        "natch12"        "natch13"
"natch14"        "natch15"        "natch16"
 [15] "adoptch04"      "adoptch05"      "adoptch06"      "adoptch07"
"adoptch08"      "adoptch09"      "adoptch10"
 [22] "adoptch11"      "adoptch12"      "adoptch13"      "adoptch14"
"adoptch15"      "adoptch16"      "allch05"
 [29] "allch06"        "allch07"        "allch08"        "allch09"
"allch10"        "allch11"        "allch12"
 [36] "allch13"        "allch14"        "allch15"        "allch16"
```

```
"lieng"          "litrans"        "chkresp"
 [43] "hcondn3"        "hcondn9"        "btype10"        "btype11"
"btype12"        "btype13"        "btype14"
 [50] "bensta3"        "bensta9"        "bensta10"       "bensta11"
"bensta12"       "bensta13"       "newsource6"
 [57] "newsource96"    "casiend"        "precog"         "wrdrecpre"
"wrcomppre"      "ns242"          "vfpre"
 [64] "nuabpre"        "cogend"         "hearcomputer"   "readtest"
"liceng"         "lictrans"       "ivlieng"
 [71] "ivlitrans"      "origadd"        "indmode"        "fiyrinvinc_tc"
"intdatd_if"     "intdatm_if"     "intdaty_if"
 [78] "doby_if"        "age_if"         "ff_ivlolw"      "ff_everint"
"ff_bentype06"   "ff_bentype21"   "ff_bentype25"
 [85] "ff_bentype30"   "ff_bentype31"   "ff_bentype32"   "ff_bentype34"
"ff_bentype35"   "ff_bentype36"   "ff_ivintlang"
 [92] "xtra5min_dv"    "grfpno"         "fiyrinvinc_if"  "ind5mus_lw"
"ind5mus_xw"     "wave"           "nbrcohdk_dv"
 [99] "xtra5minosm_dv" "scflag_dv"      "cgs7n_dv"
[1] "----------Dummy Variables----------"
[1] "predictor variable count went from 449 to 1181"
[1] "----------Variance 0 Check----------"
[1] "101 variables removed since their new variance was 0"
  [1] "hhorig.7"       "memorig.7"      "nch14resp.5"    "nch14resp.6"
"nch3resp.2"     "nch5resp.2"     "nch415resp.5"
  [8] "nch415resp.6"   "nchresp.5"      "nchresp.6"      "nchund18resp.5" "nchund18resp.6" "natch01.7
"natch01.9"
 [15] "natch02.2"      "natch02.7"      "natch02.8"      "natch03.3"
"natch03.9"      "natch04.7"      "natch04.10"
 [22] "natch05.6"      "natch05.8"      "nnatch.6"       "nadoptch.4"
"nadoptch.5"     "adoptch01.1"    "adoptch01.6"
 [29] "adoptch02.3"    "adoptch02.5"    "adoptch02.6"    "adoptch03.4"
"nchunder16.5"   "nchunder16.6"   "nch5to15.5"
 [36] "nch10to15.4"    "nch10.2"        "allch01.1"      "allch01.7"
"allch01.9"      "allch02.7"      "allch02.8"
 [43] "allch03.8"      "allch03.9"      "allch04.5"      "allch04.7"
"allch04.8"      "allch04.10"     "jbstat.9"
 [50] "jbstat.10"      "jbstat.11"      "relup.5"        "nnewborn.2"
"newsmain.10"    "newsmain.12"    "tvm2.9"
 [57] "sub7_1.3"       "sub7_1.73"      "sub7_1.86"      "sub7_1.89"
"sub7_1.91"      "sub7_1.96"      "sub7_1.97"
 [64] "clangab.4"      "ivcoop.4"       "undqus.4"       "hgbiom.5"
"hgbiof.5"       "pn1pno.5"       "pn2pno.5"
 [71] "pns1pno.5"      "pns2pno.5"      "pns2pno.6"      "ff_jbstat.10"
"ngrp_dv.2"      "nnssib_dv.4"    "nnssib_dv.5"
 [78] "buno_dv.6"      "nchild_dv.5"    "nchild_dv.6"    "hrpno.5"
"hrpno.6"        "ppno.6"         "ppno.8"
 [85] "sppno.5"        "sppno.6"        "sppno.8"        "fnpno.5"
"fnspno.5"       "fnspno.6"       "mnpno.5"
 [92] "mnspno.5"       "grmpno.2"       "grmpno.3"       "grmpno.4"
"nnmpsp_dv.4"    "ndepchl_dv.5"   "ndepchl_dv.6"
 [99] "nmpsp_dv.3"     "nnsib_dv.4"     "nnsib_dv.5"
[1] "----------K-Means----------"
[1] "15 clusters have been made for K-Means"
[1] "K-Means results as a table, the max value in each row is a simple way to define which cluster re

      0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
  1   0  0  1  3  5  7 32 22 22 25 16 28 12  9  6  5  7  6  3  3  1  1  1
2  0  2  1  1  0  0  0  0  0  1  0  0  0
  2   1  0  4  2  8  9 50 63 73 60 50 43 45 29 15 10 17 16 14  9  8  7  6
3  5  5  1  1  0  3  4  4  1  1  1  0  0
  3   2  1  2  4  6 13 31 30 25 22 30 32 28 19  9  5  5  9  7  9  4  2  3
3  0  2  1  0  2  0  1  1  1  0  0  1  0
  4   1  0  0  2  4  2 25 25 26 22 22 18 26 12  4  5  4  2  2  4  2  1  3
2  0  1  0  0  0  1  0  0  2  0  0  0  0
  5   0  1  0  6  8 13 32 52 54 49 47 38 52 23 12  6 16  4  8  6  5  7  4
5  2  3  2  2  1  2  0  0  2  2  0  0  0
  6   1  2  2  2  7 20 49 46 39 47 39 35 36 12 18  7  6  1  2  4  8  3  2
2  5  2  4  3  1  0  1  1  0  0  0  0  0
```

```
   7    0   1   1   2   1  10  28  26  18  31  21  15  15  10  15   4   5   6   8   3   3   2   4
1   3   2   1   0   0   0   0   0   0   0   0   0   0
   8    3   1   1   2   2  10  35  27  33  29  19  35  16  16   5   7   5   6   3   2   3   3   3
1   3   1   0   1   0   0   2   3   0   1   0   1   0
   9    0   1   2   1   2   5  32  24  28  25  28  24  28  13   8   7   9  12   2   6   3   2   4
2   2   1   0   0   1   1   0   0   0   0   0   1   0
   10   1   3   1   2   9  17  68  65  48  50  48  51  48  27  24  10  11  10  10  10   7   5   6
6   4   1   0   1   0   1   4   0   0   1   0   0   0
   11   0   1   2   4   4   7  43  40  44  36  28  52  40  24  13   9   9  10   4   8   6   6   4
0   3   4   4   1   1   1   1   0   0   0   0   1   0
   12   0   1   2   2   7  10  56  44  48  38  30  41  34  18  10  16  11   8   8   4   3   5   2
6   5   2   1   1   4   1   0   0   2   0   0   1   0
   13   0   0   0   0   3   6  34  15  22  10  11  17  23  16  11   2   3   4   7   3   3   1   3
2   2   2   2   0   0   0   1   0   0   1   0   0   2
   14   2   2   0   5   5  19  49  52  47  60  47  60  33  25  28  14  12   7  11   9   6   8   6
4   4   6   1   1   1   1   0   1   1   0   1   0   1
   15   1   1   1   1   8  10  65  55  60  57  39  46  40  21  19  11   8   7   8  12   5   7   4
3   3   0   2   0   1   2   0   1   1   1   1   0   1
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
 [1] "Cluster 1: Within MSE 154789078038118, Size 222"   "Cluster 2: Within MSE 9115237544271338, Siz
 [3] "Cluster 3: Within MSE 68034076318965, Size 310"    "Cluster 4: Within MSE 4657205479062, Size
 [5] "Cluster 5: Within MSE 3339619098082804, Size 464"  "Cluster 6: Within MSE 3387094364059264, Siz
 [7] "Cluster 7: Within MSE 48815349853650, Size 236"    "Cluster 8: Within MSE 55718772703504, Size
 [9] "Cluster 9: Within MSE 38859214686774, Size 274"    "Cluster 10: Within MSE 3419690212061242, S
[11] "Cluster 11: Within MSE 3473919215490826, Size 410" "Cluster 12: Within MSE 3446761046466573, S
[13] "Cluster 13: Within MSE 5614869228997, Size 206"    "Cluster 14: Within MSE 3501963590741360, S
[15] "Cluster 15: Within MSE 3537771558300898, Size 502"
[1] "Total between cluster MSE: 665335567113566336, Total within cluster MSE: 2963758340973377"
[1] "The K-Means model predicts exactly with an accuracy of 0.1249"
[1] "----------Correlation Checks----------"
[1] "plivpar.2 removed, correlated with 28 other variable(s)"
[1] "jbstat.4 removed, correlated with 28 other variable(s)"
[1] "dvage removed, correlated with 21 other variable(s)"
[1] "pn1pno.1 removed, correlated with 23 other variable(s)"
[1] "birthy removed, correlated with 20 other variable(s)"
[1] "pns1pno.1 removed, correlated with 22 other variable(s)"
[1] "ff_jbstat.4 removed, correlated with 24 other variable(s)"
[1] "memcont11.1 removed, correlated with 18 other variable(s)"
[1] "memcont13.1 removed, correlated with 17 other variable(s)"
[1] "ff_bentype01.1 removed, correlated with 21 other variable(s)"
[1] "subcont11.1 removed, correlated with 16 other variable(s)"
[1] "pensioner_dv.2 removed, correlated with 21 other variable(s)"
[1] "subcont13.1 removed, correlated with 15 other variable(s)"
[1] "nscont11.1 removed, correlated with 14 other variable(s)"
[1] "nscont13.1 removed, correlated with 13 other variable(s)"
[1] "pn2pno.2 removed, correlated with 15 other variable(s)"
[1] "btype4.1 removed, correlated with 15 other variable(s)"
[1] "dmemcont11.1 removed, correlated with 12 other variable(s)"
[1] "pns2pno.2 removed, correlated with 14 other variable(s)"
[1] "memcont12.1 removed, correlated with 11 other variable(s)"
[1] "dmemcont13.1 removed, correlated with 11 other variable(s)"
[1] "age_dv removed, correlated with 15 other variable(s)"
[1] "npn_dv.2 removed, correlated with 13 other variable(s)"
[1] "agegr13_dv.13 removed, correlated with 16 other variable(s)"
[1] "nch14resp.2 removed, correlated with 10 other variable(s)"
[1] "nchresp.2 removed, correlated with 11 other variable(s)"
[1] "subcont12.1 removed, correlated with 10 other variable(s)"
[1] "vfcont11.1 removed, correlated with 10 other variable(s)"
[1] "nch14resp.3 removed, correlated with 9 other variable(s)"
[1] "nchresp.1 removed, correlated with 10 other variable(s)"
[1] "nchresp.3 removed, correlated with 10 other variable(s)"
[1] "marstat.2 removed, correlated with 10 other variable(s)"
[1] "employ.2 removed, correlated with 10 other variable(s)"
[1] "nscont12.1 removed, correlated with 9 other variable(s)"
[1] "vfcont13.1 removed, correlated with 9 other variable(s)"
[1] "nchund18resp.2 removed, correlated with 9 other variable(s)"
[1] "doby_dv removed, correlated with 13 other variable(s)"
[1] "npns_dv.2 removed, correlated with 16 other variable(s)"
```

```
[1] "dmemcont12.1 removed, correlated with 8 other variable(s)"
[1] "nacont11.1 removed, correlated with 8 other variable(s)"
[1] "nch415resp.1 removed, correlated with 8 other variable(s)"
[1] "nchund18resp.3 removed, correlated with 8 other variable(s)"
[1] "nchunder16.2 removed, correlated with 9 other variable(s)"
[1] "jbhas.2 removed, correlated with 8 other variable(s)"
[1] "btype5.1 removed, correlated with 8 other variable(s)"
[1] "mastat_dv.2 removed, correlated with 9 other variable(s)"
[1] "hhorig.3 removed, correlated with 7 other variable(s)"
[1] "natch01.2 removed, correlated with 7 other variable(s)"
[1] "vfcont12.1 removed, correlated with 7 other variable(s)"
[1] "nacont13.1 removed, correlated with 7 other variable(s)"
[1] "hgbiom.1 removed, correlated with 7 other variable(s)"
[1] "indinub_lw removed, correlated with 7 other variable(s)"
[1] "hhtype_dv.8 removed, correlated with 11 other variable(s)"
[1] "nchunder16.3 removed, correlated with 8 other variable(s)"
[1] "pidp removed, correlated with 6 other variable(s)"
[1] "memorig.3 removed, correlated with 6 other variable(s)"
[1] "nch14resp.1 removed, correlated with 6 other variable(s)"
[1] "jbstat.2 removed, correlated with 6 other variable(s)"
[1] "nacont12.1 removed, correlated with 6 other variable(s)"
[1] "ivprsnt.2 removed, correlated with 6 other variable(s)"
[1] "hgbiom.2 removed, correlated with 6 other variable(s)"
[1] "ff_emplw.2 removed, correlated with 7 other variable(s)"
[1] "buno_dv.3 removed, correlated with 7 other variable(s)"
[1] "indpxus_lw removed, correlated with 6 other variable(s)"
[1] "natch02.4 removed, correlated with 6 other variable(s)"
[1] "natch04.6 removed, correlated with 6 other variable(s)"
[1] "livesp_dv.1 removed, correlated with 8 other variable(s)"
[1] "hidp removed, correlated with 5 other variable(s)"
[1] "nch14resp.4 removed, correlated with 5 other variable(s)"
[1] "nch415resp.2 removed, correlated with 5 other variable(s)"
[1] "nchresp.4 removed, correlated with 6 other variable(s)"
[1] "natch03.5 removed, correlated with 5 other variable(s)"
[1] "cgivns1_dv.2 removed, correlated with 5 other variable(s)"
[1] "cgivns1_dv.3 removed, correlated with 5 other variable(s)"
[1] "cgivns1_dv.4 removed, correlated with 5 other variable(s)"
[1] "indpxub_lw removed, correlated with 5 other variable(s)"
[1] "nnatch.4 removed, correlated with 5 other variable(s)"
[1] "nchunder16.1 removed, correlated with 5 other variable(s)"
[1] "scsf2a.3 removed, correlated with 5 other variable(s)"
[1] "nnmpsp_dv.1 removed, correlated with 5 other variable(s)"
[1] "npensioner_dv.2 removed, correlated with 7 other variable(s)"
[1] "pno.2 removed, correlated with 4 other variable(s)"
[1] "month removed, correlated with 4 other variable(s)"
[1] "sex.2 removed, correlated with 4 other variable(s)"
[1] "nnatch.3 removed, correlated with 4 other variable(s)"
[1] "allch02.4 removed, correlated with 5 other variable(s)"
[1] "allch03.5 removed, correlated with 5 other variable(s)"
[1] "netag_1 removed, correlated with 4 other variable(s)"
[1] "hgbiom.3 removed, correlated with 4 other variable(s)"
[1] "hgbiom.4 removed, correlated with 4 other variable(s)"
[1] "hgbiof.1 removed, correlated with 4 other variable(s)"
[1] "hgbiof.2 removed, correlated with 4 other variable(s)"
[1] "hgbiof.3 removed, correlated with 4 other variable(s)"
[1] "hgbiof.4 removed, correlated with 4 other variable(s)"
[1] "respm16.2 removed, correlated with 4 other variable(s)"
[1] "fimngrs_tc.1 removed, correlated with 4 other variable(s)"
[1] "npn_dv.1 removed, correlated with 4 other variable(s)"
[1] "nchild_dv.3 removed, correlated with 6 other variable(s)"
[1] "mnpno.1 removed, correlated with 6 other variable(s)"
[1] "indin91_lw removed, correlated with 4 other variable(s)"
[1] "fimnnet_dv removed, correlated with 6 other variable(s)"
[1] "cgivwrd1_dv.2 removed, correlated with 4 other variable(s)"
[1] "cgivwrd1_dv.3 removed, correlated with 4 other variable(s)"
[1] "cgivwrd1_dv.4 removed, correlated with 4 other variable(s)"
[1] "marstat_dv.6 removed, correlated with 7 other variable(s)"
[1] "natch02.6 removed, correlated with 4 other variable(s)"
```

```
[1] "scsf2b.3 removed, correlated with 4 other variable(s)"
[1] "indinub_xw removed, correlated with 4 other variable(s)"
[1] "hhorig.2 removed, correlated with 3 other variable(s)"
[1] "isyear.2012 removed, correlated with 3 other variable(s)"
[1] "natch02.3 removed, correlated with 3 other variable(s)"
[1] "nnatch.2 removed, correlated with 3 other variable(s)"
[1] "nchunder16.4 removed, correlated with 3 other variable(s)"
[1] "relup.2 removed, correlated with 3 other variable(s)"
[1] "btype8.1 removed, correlated with 3 other variable(s)"
[1] "pn1pno.3 removed, correlated with 3 other variable(s)"
[1] "pn1pno.4 removed, correlated with 3 other variable(s)"
[1] "pn2pno.3 removed, correlated with 3 other variable(s)"
[1] "pn2pno.4 removed, correlated with 3 other variable(s)"
[1] "ff_bentype18.1 removed, correlated with 3 other variable(s)"
[1] "ff_bentype19.1 removed, correlated with 4 other variable(s)"
[1] "nchild_dv.1 removed, correlated with 4 other variable(s)"
[1] "nchild_dv.2 removed, correlated with 4 other variable(s)"
[1] "ppno.1 removed, correlated with 3 other variable(s)"
[1] "fnpno.1 removed, correlated with 3 other variable(s)"
[1] "cgs7cs_dv.5 removed, correlated with 4 other variable(s)"
[1] "paygu_if.1 removed, correlated with 3 other variable(s)"
[1] "cgivs71_dv.2 removed, correlated with 3 other variable(s)"
[1] "cgivs71_dv.3 removed, correlated with 3 other variable(s)"
[1] "strata removed, correlated with 3 other variable(s)"
[1] "nadoptch.2 removed, correlated with 3 other variable(s)"
[1] "scsf3a.5 removed, correlated with 3 other variable(s)"
[1] "scsf4a.5 removed, correlated with 3 other variable(s)"
[1] "sub7_3 removed, correlated with 3 other variable(s)"
[1] "indin01_lw removed, correlated with 3 other variable(s)"
[1] "indscub_lw removed, correlated with 3 other variable(s)"
[1] "nmpsp_dv.1 removed, correlated with 3 other variable(s)"
[1] "memorig.2 removed, correlated with 2 other variable(s)"
[1] "isyear.2013 removed, correlated with 2 other variable(s)"
[1] "psiblings.2 removed, correlated with 2 other variable(s)"
[1] "natch03.7 removed, correlated with 2 other variable(s)"
[1] "nnatch.1 removed, correlated with 2 other variable(s)"
[1] "allch01.2 removed, correlated with 2 other variable(s)"
[1] "allch04.6 removed, correlated with 2 other variable(s)"
[1] "istrtdaty.2012 removed, correlated with 2 other variable(s)"
[1] "jbstat.7 removed, correlated with 2 other variable(s)"
[1] "netsx_1.2 removed, correlated with 2 other variable(s)"
[1] "netag_2 removed, correlated with 2 other variable(s)"
[1] "bensta4.1 removed, correlated with 2 other variable(s)"
[1] "bensta7.1 removed, correlated with 2 other variable(s)"
[1] "fiyrdia removed, correlated with 2 other variable(s)"
[1] "cmroute.1 removed, correlated with 3 other variable(s)"
[1] "civicduty.6 removed, correlated with 2 other variable(s)"
[1] "scsf6a.2 removed, correlated with 2 other variable(s)"
[1] "scsf7.5 removed, correlated with 3 other variable(s)"
[1] "marstat.4 removed, correlated with 2 other variable(s)"
[1] "marstat.5 removed, correlated with 2 other variable(s)"
[1] "marstat.6 removed, correlated with 2 other variable(s)"
[1] "memper.3 removed, correlated with 2 other variable(s)"
[1] "memper.4 removed, correlated with 2 other variable(s)"
[1] "pns1pno.3 removed, correlated with 2 other variable(s)"
[1] "pns1pno.4 removed, correlated with 2 other variable(s)"
[1] "pns2pno.3 removed, correlated with 2 other variable(s)"
[1] "pns2pno.4 removed, correlated with 2 other variable(s)"
[1] "ff_jbstat.3 removed, correlated with 2 other variable(s)"
[1] "ff_bentype22.1 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.5 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.6 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.7 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.8 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.9 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.10 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.11 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.12 removed, correlated with 2 other variable(s)"
```

```
[1] "agegr5_dv.13 removed, correlated with 2 other variable(s)"
[1] "cohab_dv.1 removed, correlated with 2 other variable(s)"
[1] "ppno.2 removed, correlated with 2 other variable(s)"
[1] "fnpno.2 removed, correlated with 2 other variable(s)"
[1] "fnspno.1 removed, correlated with 2 other variable(s)"
[1] "fnspno.2 removed, correlated with 3 other variable(s)"
[1] "paynu_if.1 removed, correlated with 2 other variable(s)"
[1] "seearngrs_if.1 removed, correlated with 2 other variable(s)"
[1] "fimnlabgrs_if removed, correlated with 3 other variable(s)"
[1] "indinus_lw removed, correlated with 2 other variable(s)"
[1] "fimnlabgrs_dv removed, correlated with 2 other variable(s)"
[1] "cgivvf1_dv.2 removed, correlated with 2 other variable(s)"
[1] "cgivvf1_dv.3 removed, correlated with 2 other variable(s)"
[1] "cgivvf1_dv.4 removed, correlated with 2 other variable(s)"
[1] "cgivwri1_dv.4 removed, correlated with 3 other variable(s)"
[1] "adoptch01.3 removed, correlated with 2 other variable(s)"
[1] "nch5to15.3 removed, correlated with 2 other variable(s)"
[1] "nbrcoh4.4 removed, correlated with 2 other variable(s)"
[1] "scsf3b.4 removed, correlated with 2 other variable(s)"
[1] "scsf3b.5 removed, correlated with 2 other variable(s)"
[1] "scsf4b.5 removed, correlated with 2 other variable(s)"
[1] "scsf6c.5 removed, correlated with 2 other variable(s)"
[1] "scptrt5c1.7 removed, correlated with 2 other variable(s)"
[1] "sub7_1.93 removed, correlated with 2 other variable(s)"
[1] "ff_jbstat.8 removed, correlated with 2 other variable(s)"
[1] "nbrcoh_dv removed, correlated with 2 other variable(s)"
[1] "cgna_dv.3 removed, correlated with 2 other variable(s)"
[1] "hhtype_dv.12 removed, correlated with 2 other variable(s)"
[1] "jbiindb_dv removed, correlated with 2 other variable(s)"
[1] "pno.4 removed, correlated with 1 other variable(s)"
[1] "pno.5 removed, correlated with 1 other variable(s)"
[1] "hhorig.4 removed, correlated with 1 other variable(s)"
[1] "hhorig.5 removed, correlated with 1 other variable(s)"
[1] "hhorig.6 removed, correlated with 1 other variable(s)"
[1] "psu removed, correlated with 1 other variable(s)"
[1] "nchund18resp.1 removed, correlated with 1 other variable(s)"
[1] "nchund18resp.4 removed, correlated with 1 other variable(s)"
[1] "natch01.3 removed, correlated with 1 other variable(s)"
[1] "natch01.6 removed, correlated with 1 other variable(s)"
[1] "natch03.4 removed, correlated with 1 other variable(s)"
[1] "natch05.7 removed, correlated with 1 other variable(s)"
[1] "nadoptch.3 removed, correlated with 1 other variable(s)"
[1] "nch5to15.1 removed, correlated with 1 other variable(s)"
[1] "nch5to15.2 removed, correlated with 1 other variable(s)"
[1] "allch02.3 removed, correlated with 1 other variable(s)"
[1] "allch02.6 removed, correlated with 1 other variable(s)"
[1] "istrtdaty.2013 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.2 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.3 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.4 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.5 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.6 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.7 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.8 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.9 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.10 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.11 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.12 removed, correlated with 1 other variable(s)"
[1] "istrtdatd removed, correlated with 1 other variable(s)"
[1] "jbstat.3 removed, correlated with 1 other variable(s)"
[1] "jbstat.6 removed, correlated with 1 other variable(s)"
[1] "locseras.2 removed, correlated with 1 other variable(s)"
[1] "locserb.2 removed, correlated with 1 other variable(s)"
[1] "locserd.2 removed, correlated with 1 other variable(s)"
[1] "locsere.2 removed, correlated with 1 other variable(s)"
[1] "locserc.2 removed, correlated with 1 other variable(s)"
[1] "nbrcoh1.2 removed, correlated with 1 other variable(s)"
[1] "nbrcoh1.5 removed, correlated with 1 other variable(s)"
```

```
[1] "nbrcoh2.2 removed, correlated with 1 other variable(s)"
[1] "nbrcoh3.2 removed, correlated with 1 other variable(s)"
[1] "simfam.3 removed, correlated with 1 other variable(s)"
[1] "netkn_1.3 removed, correlated with 1 other variable(s)"
[1] "netet_1 removed, correlated with 1 other variable(s)"
[1] "netsx_2.2 removed, correlated with 1 other variable(s)"
[1] "netkn_2.3 removed, correlated with 1 other variable(s)"
[1] "netph_2.2 removed, correlated with 1 other variable(s)"
[1] "org.2 removed, correlated with 1 other variable(s)"
[1] "sf1.2 removed, correlated with 1 other variable(s)"
[1] "sf1.3 removed, correlated with 1 other variable(s)"
[1] "sf1.4 removed, correlated with 1 other variable(s)"
[1] "sf1.5 removed, correlated with 1 other variable(s)"
[1] "relup.6 removed, correlated with 1 other variable(s)"
[1] "hcondn16.1 removed, correlated with 1 other variable(s)"
[1] "hconde6.1 removed, correlated with 1 other variable(s)"
[1] "hconde7.1 removed, correlated with 1 other variable(s)"
[1] "lvrel3.1 removed, correlated with 1 other variable(s)"
[1] "lvrel96.1 removed, correlated with 1 other variable(s)"
[1] "visfam.4 removed, correlated with 1 other variable(s)"
[1] "btype2.1 removed, correlated with 1 other variable(s)"
[1] "btype3.1 removed, correlated with 1 other variable(s)"
[1] "finfut.2 removed, correlated with 1 other variable(s)"
[1] "vote6.2 removed, correlated with 1 other variable(s)"
[1] "perpolinf.11 removed, correlated with 1 other variable(s)"
[1] "perbfts.2 removed, correlated with 1 other variable(s)"
[1] "grpbfts.2 removed, correlated with 1 other variable(s)"
[1] "demorient.2 removed, correlated with 1 other variable(s)"
[1] "poleff1.5 removed, correlated with 1 other variable(s)"
[1] "poleff2.3 removed, correlated with 1 other variable(s)"
[1] "scsf3a.2 removed, correlated with 1 other variable(s)"
[1] "scsf3a.3 removed, correlated with 1 other variable(s)"
[1] "scsf6b.2 removed, correlated with 1 other variable(s)"
[1] "scopngbha.2 removed, correlated with 1 other variable(s)"
[1] "scopngbhb.2 removed, correlated with 1 other variable(s)"
[1] "scopngbhe.2 removed, correlated with 1 other variable(s)"
[1] "scopngbhg.2 removed, correlated with 1 other variable(s)"
[1] "scptrt5n1.2 removed, correlated with 1 other variable(s)"
[1] "scptrt5o1.7 removed, correlated with 1 other variable(s)"
[1] "scptrt5a2.7 removed, correlated with 1 other variable(s)"
[1] "scptrt5e2.7 removed, correlated with 1 other variable(s)"
[1] "scptrt5n2.7 removed, correlated with 1 other variable(s)"
[1] "scptrt5o2.7 removed, correlated with 1 other variable(s)"
[1] "scptrt5a3.7 removed, correlated with 1 other variable(s)"
[1] "scptrt5c3.7 removed, correlated with 1 other variable(s)"
[1] "scptrt5o3.7 removed, correlated with 1 other variable(s)"
[1] "marstat.3 removed, correlated with 1 other variable(s)"
[1] "memper.2 removed, correlated with 1 other variable(s)"
[1] "memper.5 removed, correlated with 1 other variable(s)"
[1] "sub7_4 removed, correlated with 1 other variable(s)"
[1] "vftctw removed, correlated with 1 other variable(s)"
[1] "nadisease.2 removed, correlated with 1 other variable(s)"
[1] "vfct removed, correlated with 1 other variable(s)"
[1] "memprob1.1 removed, correlated with 1 other variable(s)"
[1] "memprob2.1 removed, correlated with 1 other variable(s)"
[1] "memprob3.1 removed, correlated with 1 other variable(s)"
[1] "memaid.2 removed, correlated with 1 other variable(s)"
[1] "clangab.2 removed, correlated with 1 other variable(s)"
[1] "ivcoop.2 removed, correlated with 1 other variable(s)"
[1] "respf16.2 removed, correlated with 1 other variable(s)"
[1] "fimnlabgrs_tc.1 removed, correlated with 1 other variable(s)"
[1] "j2paynet_dv removed, correlated with 1 other variable(s)"
[1] "pn1pno.2 removed, correlated with 1 other variable(s)"
[1] "ff_bentype02.1 removed, correlated with 1 other variable(s)"
[1] "fimninvnet_dv removed, correlated with 1 other variable(s)"
[1] "intdaty_dv.2012 removed, correlated with 1 other variable(s)"
[1] "ngrp_dv.1 removed, correlated with 1 other variable(s)"
[1] "nnssib_dv.1 removed, correlated with 1 other variable(s)"
```

```
[1] "nnssib_dv.2 removed, correlated with 1 other variable(s)"
[1] "nnssib_dv.3 removed, correlated with 1 other variable(s)"
[1] "country.2 removed, correlated with 1 other variable(s)"
[1] "country.3 removed, correlated with 1 other variable(s)"
[1] "country.4 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.14 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.15 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.4 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.5 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.6 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.7 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.8 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.9 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.10 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.11 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.12 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.4 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.5 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.6 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.10 removed, correlated with 1 other variable(s)"
[1] "buno_dv.2 removed, correlated with 1 other variable(s)"
[1] "nchild_dv.4 removed, correlated with 1 other variable(s)"
[1] "ppno.3 removed, correlated with 1 other variable(s)"
[1] "ppno.4 removed, correlated with 1 other variable(s)"
[1] "sppno.1 removed, correlated with 1 other variable(s)"
[1] "fnpno.3 removed, correlated with 1 other variable(s)"
[1] "fnpno.4 removed, correlated with 1 other variable(s)"
[1] "mnpno.2 removed, correlated with 1 other variable(s)"
[1] "mnpno.3 removed, correlated with 1 other variable(s)"
[1] "mnpno.4 removed, correlated with 1 other variable(s)"
[1] "hiqual_dv.9 removed, correlated with 1 other variable(s)"
[1] "big5c_dv.5 removed, correlated with 1 other variable(s)"
[1] "cgsrmem_dv.3 removed, correlated with 1 other variable(s)"
[1] "cgsrmem_dv.4 removed, correlated with 1 other variable(s)"
[1] "cgs7cs_dv.3 removed, correlated with 1 other variable(s)"
[1] "cgs7cs_dv.4 removed, correlated with 1 other variable(s)"
[1] "cgs7ca_dv.1 removed, correlated with 1 other variable(s)"
[1] "indns91_lw removed, correlated with 1 other variable(s)"
[1] "indnsub_lw removed, correlated with 1 other variable(s)"
[1] "indpxub_xw removed, correlated with 1 other variable(s)"
[1] "hhtype_dv.10 removed, correlated with 1 other variable(s)"
[1] "hhtype_dv.11 removed, correlated with 1 other variable(s)"
[1] "rach16_dv.2 removed, correlated with 1 other variable(s)"
[1] "fimngrs_dv removed, correlated with 1 other variable(s)"
[1] "fibenothr_dv removed, correlated with 1 other variable(s)"
[1] "cgivwri1_dv.2 removed, correlated with 1 other variable(s)"
[1] "cgivwri1_dv.3 removed, correlated with 1 other variable(s)"
[1] "racel_dv removed, correlated with 1 other variable(s)"
[1] "big5a_dv.5 removed, correlated with 1 other variable(s)"
[1] "359 variables removed since they had high correlation coefs"
[1] "The final cleaned dataset has been completed at this stage and is stored under the name 'cleanDa
[1] "——————Attempting a Train Test Split——————"
[1] "Good train, test split found"
[1] "The working seed found was 3"
[1] "——————kNN——————"
[1] "223 neighbours considered for each test data point"
[1] "kNN results as a table, follow the diagonal for the correctly mapped clusters"
         real
predicted  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
16 17 18 19 20 21 22 23 24 25 26 27 28
        0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
        1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
        2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
        3  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
```

4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	1	0	2	3	9	21	106	76	72	69	51	51	49	19	21	14	9	10	11	10	4	5	5	4	1	2	0	0	1
7	1	2	1	2	4	7	29	32	32	38	35	38	21	12	5	2	7	1	3	2	4	1	1	0	2	0	0	1	0
8	1	0	1	3	1	4	18	31	33	27	24	28	24	9	9	5	3	3	6	5	2	3	2	3	2	0	3	1	0
9	0	0	0	0	1	0	3	7	7	10	7	17	13	11	9	3	4	4	2	3	3	2	2	1	2	2	0	0	0
10	0	0	0	0	0	0	0	1	0	1	0	4	3	2	0	0	0	2	1	1	0	1	1	0	1	0	0	0	0
11	0	0	0	0	0	0	1	0	0	0	2	3	6	4	3	3	1	3	1	5	2	5	2	3	1	0	2	0	0
12	0	0	0	0	0	0	0	0	0	1	1	1	4	4	0	1	4	2	4	1	6	2	2	0	2	1	1	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

	real						
predicted	29	30	31	32	33	35	36
0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0
6	2	0	0	1	2	0	0
7	1	0	0	0	0	0	0
8	2	0	0	0	0	0	0
9	0	0	0	1	0	1	1
10	0	1	1	0	0	1	0
11	0	0	1	0	1	0	0
12	1	1	2	1	0	0	1
13	0	1	0	0	0	0	0
14	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0

```
        20    0    0    0    0    0    0    0
        21    0    0    0    0    0    0    0
        22    0    0    0    0    0    0    0
        23    0    0    0    0    0    0    0
        24    0    0    0    0    0    0    0
        25    0    0    0    0    0    0    0
        26    0    0    0    0    0    0    0
 [ reached getOption("max.print") -- omitted 10 rows ]
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
[1] "The MSE of the predicted values are of 213.3252"
[1] "The kNN model predicts exactly with an accuracy of 0.1344"
[1] "----------CART prediction model----------"
n= 4196

node), split, n, deviance, yval
      * denotes terminal node

 1) root 4196 110576.200 10.774790
   2) sf12mcs_dv>=-0.978582 3550  45293.300  9.414085
     4) sf12mcs_dv>=0.1249772 2338  17827.060  8.195038
       8) scsf6c.4< 0.5 1652   9792.506  7.611380 *
       9) scsf6c.4>=0.5 686   6116.560  9.600583 *
     5) sf12mcs_dv< 0.1249772 1212  17289.450 11.765680
      10) sf12mcs_dv>=-0.4229895 698   8040.345 10.918340 *
      11) sf12mcs_dv< -0.4229895 514   8067.403 12.916340 *
   3) sf12mcs_dv< -0.978582 646  22589.870 18.252320
     6) sf12mcs_dv>=-2.359938 526  12879.940 16.836500 *
     7) sf12mcs_dv< -2.359938 120   4033.792 24.458330 *
[1] "Variable Importance"
  sf12mcs_dv      scsf6c.2      scsf4a.3    sf12pcs_dv      scsf4a.2       scsf7.3
scsf4a.4      scsf6c.3      scsf6a.3
 60135.83440    9714.97173    9494.88421    9305.64218    6080.78535    5508.89831
2754.11240    2720.62023    2183.13788
    scsf6c.4      scsf4b.4      scsf4b.2      scsf6a.4       scsf7.2         hrpid fimnprben_dv
finnow.5    indscub_xw
  1917.99663    1721.32025     425.71055     179.32475     141.90352      94.60234
8.38774       8.38774       8.38774
[1] "The MSE of the predicted values are of 13.3525"
[1] "The CART model predicts exactly with accuracy of 0.1265"
[1] "----------Ordinary Linear Regression (Initial)----------"
[1] "The full model AIC is: 21794.4735"
[1] "----------Variance Inflation Factor Removal----------"
[1] "The variable ienddathh.20 was removed since it had a VIF score of 864337892.7596"
[1] "The variable agegr10_dv.5 was removed since it had a VIF score of 201.2939"
[1] "The variable nbrsnci_dv was removed since it had a VIF score of 71.6345"
[1] "The variable big5e_dv.5 was removed since it had a VIF score of 65.7453"
[1] "The variable newsmain.6 was removed since it had a VIF score of 58.5249"
[1] "The variable big5o_dv.5 was removed since it had a VIF score of 43.4941"
[1] "The variable istrtdathh was removed since it had a VIF score of 38.7722"
[1] "The variable sclfsato.6 was removed since it had a VIF score of 25.2837"
[1] "The variable sf12mcs_dv was removed since it had a VIF score of 24.052"
[1] "The variable cgwrd_dv.6 was removed since it had a VIF score of 19.4726"
[1] "The variable big5n_dv.4 was removed since it had a VIF score of 18.2419"
[1] "The variable respm16_dv.2 was removed since it had a VIF score of 17.8568"
[1] "The variable cgwri_dv.7 was removed since it had a VIF score of 17.7719"
[1] "The variable scptrt5n3.5 was removed since it had a VIF score of 16.2238"
[1] "The variable scptrt5e1.7 was removed since it had a VIF score of 13.3127"
[1] "The variable sf12pcs_dv was removed since it had a VIF score of 13.0973"
[1] "The variable single_dv.1 was removed since it had a VIF score of 10.4516"
[1] "17 variables removed from the Ordinary Linear Model since they have a VIF score higher than 10"
[1] "The full model AIC after VIF checks is: 22025.9175"
[1] "----------Backwards Selection----------"
[1] "50 out of 720 variables removed so far."
[1] "100 out of 720 variables removed so far."
[1] "150 out of 720 variables removed so far."
[1] "200 out of 720 variables removed so far."
[1] "250 out of 720 variables removed so far."
[1] "300 out of 720 variables removed so far."
```

```
[1] "350 out of 720 variables removed so far."
[1] "400 out of 720 variables removed so far."
[1] "450 out of 720 variables removed so far."
[1] "500 out of 720 variables removed so far."
[1] "550 out of 720 variables removed so far."
[1] "600 out of 720 variables removed so far."
[1] "614 out of 720 variables removed in backwards selection since they weren't significant at the 95
  [1] "simrace.4"        "scsf1.3"          "nnatch.5"         "vote6.3"
"hhtype_dv.19"     "simeduc.3"
  [7] "gor_dv.12"        "adoptch01.5"      "orga2.1"          "ns200pre.2"
"voteintent.11"    "scptrt5n3.3"
 [13] "newsmain.3"       "netpuse.4"        "big5o_dv.3"       "nch10to15.2"
"intdatm_dv.3"     "relup.3"
 [19] "wrdrcl.3"         "gor_dv.2"         "perpolinf.2"      "netpuse.7"
"crdark.5"         "tvm2.7"
 [25] "nch8resp.1"       "fimngrs_if"       "locserb.3"        "ff_jbstat.97"
"marstat_dv.2"     "orga10.1"
 [31] "ff_bentype17.1"   "hhtype_dv.23"     "ndepchl_dv.3"     "gor_dv.4"
"servacc.2"        "big5a_dv.2"
 [37] "scptrt5a1.3"      "hhtype_dv.17"     "scptrt5o3.3"      "newsource3.1"
"adoptch02.4"      "poleff4.4"
 [43] "cgna_dv.5"        "big5e_dv.3"       "cgsrmem_dv.2"     "nch415resp.4"
"cgvfw_dv"         "rhland_code.1"
 [49] "civicduty.2"      "intdaty_dv.2013"  "btype1.1"         "nunmpsp_dv.1"
"mnspno.3"         "orga4.1"
 [55] "ff_bentype38.1"   "netlv_2.2"        "scptrt5o2.2"      "scopngbhb.5"
"j2pay_if.1"       "orga12.1"
 [61] "natch02.5"        "scptrt5c2.2"      "simjob.4"         "scptrt5e1.2"
"buno_dv.5"        "pns1pno.2"
 [67] "lvrel8.1"         "poleff4.2"        "big5n_dv.6"       "scptrt5o1.6"
"voteintent.8"     "simage.4"
 [73] "netjb_1.6"        "crwora.2"         "respf16_dv.2"     "perpolinf.5"
"ienddathh.23"     "ns200pre.3"
 [79] "scptrt5c1.2"      "big5n_dv.5"       "hcondn15.1"       "lenindintv"
"cgwrd_dv.3"       "cgwrd_dv.1"
 [85] "gor_dv.5"         "gor_dv.6"         "gor_dv.9"         "agegr10_dv.4"
"hcondn11.1"       "newsmain.9"
 [91] "poleff1.2"        "voteintent.1"     "perpolinf.1"      "cgwrd_dv.9"
"intdatm_dv.12"    "ffbrfedlw.1"
 [97] "newsmain.5"       "memaid.3"         "voteintent.2"     "scptrt5e2.6"
"scptrt5e2.4"      "ienddathh.14"
[103] "netjb_1.4"        "scptrt5o2.5"      "sclfsat7.7"       "netkn_1.4"
"big5o_dv.6"       "hcondn17.1"
[109] "simeduc.4"        "siminc.4"         "nch10to15.3"      "ff_bentype37.1"
"mnspno.2"         "visfam.6"
[115] "demorient.3"      "demorient.4"      "grpbfts.4"        "scsf1.2"
"cgwri_dv.1"       "closenum"
[121] "scptrt5c2.7"      "tvm2.6"           "grmpno.1"         "nnsib_dv.1"
"hiqual_dv.5"      "newsource12.1"
[127] "sppno.2"          "netlv_2.5"        "tvm2.97"          "simage.2"
"memorig.4"        "orga14.1"
[133] "ienddatmm"        "sppno.4"          "nch10.1"          "lvrel7.1"
"netjb_1.2"        "sppno.3"
[139] "scptrt5o3.4"      "cgwrd_dv.7"       "tvm2.8"           "scptrt5c3.3"
"scptrt5c2.4"      "mastat_dv.3"
[145] "cgwri_dv.9"       "cgwri_dv.4"       "simarea.5"        "hood15.4"
"agegr10_dv.8"     "vote6.4"
[151] "ff_bentype11.1"   "ff_bentype23.1"   "indbdub_lw"       "scptrt5e2.5"
"netwr_2.2"        "intdatd_dv"
[157] "allch01.6"        "allch01.4"        "ienddathh.21"     "poleff4.5"
"scptrt5n3.4"      "tvm2.4"
[163] "mobuse.2"         "locseras.4"       "netjb_2.3"        "netkn_2.4"
"scopngbhf.4"      "newsource1.1"
[169] "ienddathh.10"     "scptrt5a3.6"      "nbrcoh2.4"        "scptrt5e1.5"
"tvm2.3"           "locsere.4"
[175] "adoptch03.5"      "scsf2b.2"         "scopngbhb.3"      "scopngbhb.4"
"scptrt5o2.3"      "polcost.5"
[181] "istrtdatmm"       "cindtime"         "scptrt5a2.4"      "big5a_dv.6"
```

```
"big5a_dv.7"         "sclfsat1.7"
[187] "pno.6"            "voteintent.6"       "gor_dv.3"           "scptrt5e3.4"
"nbrcoh4.3"          "ff_bentype28.1"
[193] "perpolinf.9"      "hhtype_dv.5"        "netjb_2.4"          "intdatm_dv.5"
"scptrt5e2.3"        "adoptch01.2"
[199] "ff_bentype03.1"   "orga6.1"            "vote1.2"            "visfam.5"
"lvrel1.1"           "scsf5.2"
[205] "simfam.5"         "simfam.4"           "ndepchl_dv.2"       "scsf3b.3"
"crdark.3"           "allch01.3"
[211] "relup.4"          "nbrcoh1.3"          "scptrt5a1.2"        "scptrt5a1.4"
"lvrel9.1"           "locserap.4"
[217] "netlv_2.4"        "netlv_2.3"          "scptrt5o3.2"        "netph_1.3"
"cgs7cs_dv.2"        "newsource11.1"
[223] "newsource9.1"     "netpuse.3"          "voteintent.5"       "susp.2"
"ff_bentype20.1"     "ienddathh.13"
[229] "allch02.5"        "fimnmisc_dv"        "hcondn96.1"         "hcondn14.1"
"scptrt5o1.5"        "big5o_dv.4"
[235] "scptrt5c3.4"      "big5c_dv.6"         "scopngbhd.5"        "big5e_dv.4"
"natch03.6"          "locsere.3"
[241] "cgs7ca_dv.2"      "fimnlabnet_tc.1"    "perbfts.5"          "poleff1.4"
"bensta1.1"          "simarea.4"
[247] "undqus.3"         "llknbrd.2"          "scopngbhg.4"        "fimnpen_dv"
"newsource4.1"       "newsmain.4"
[253] "netkn_2.2"        "cgwri_dv.8"         "cgwrd_dv.8"         "netph_1.4"
"scopngbhe.3"        "finfut.3"
[259] "netph_2.3"        "netjb_2.5"          "j2has.2"            "sevenspap.3"
"ienddathh.17"       "intdatm_dv.10"
[265] "intdatm_dv.7"     "orga96.1"           "cgwrd_dv.2"         "newsmain.7"
"j2pay_dv"           "orga7.1"
[271] "hcondn8.1"        "nbrcoh3.3"          "cgivs71_dv.4"       "nbrcoh4.5"
"ndepchl_dv.1"       "cgwrd_dv.5"
[277] "cgwri_dv.6"       "lvrel10.1"          "scopngbhh.4"        "hhtype_dv.20"
"cgsrmem2_dv.1"      "poleff2.5"
[283] "scopngbhh.3"      "fimnprben_dv"       "hrpno.4"            "cgivna1_dv.3"
"sampst.3"           "sub7_1.94"
[289] "indscus_lw"       "gor_dv.7"           "hood15.3"           "scptrt5o2.6"
"scptrt5o2.4"        "voteintent.9"
[295] "voteintent.10"    "newsource10.1"      "newsource2.1"       "scsf3a.4"
"btype7.1"           "simeduc.2"
[301] "ff_bentype26.1"   "ff_bentype29.1"     "bensta97.1"         "bensta96.1"
"ff_bentype04.1"     "hrpid"
[307] "memprob4.1"       "hhresp_dv.2"        "simage.3"           "nunmpsp_dv.2"
"sclfsat7.6"         "sclfsat7.3"
[313] "sclfsat7.4"       "scptrt5e2.2"        "mnspno.4"           "nnsib_dv.2"
"nnewborn.1"         "big5a_dv.4"
[319] "simjob.2"         "scopngbhf.3"        "crdark.2"           "hiqual_dv.2"
"orga9.1"            "susp.3"
[325] "orga8.1"          "cgs7cs_dv.1"        "addrmov_dv.2"       "ff_jbstat.5"
"ff_jbstat.2"        "scopngbhf.5"
[331] "btype9.1"         "ff_bentype39.1"     "perbfts.6"          "cgs7ca_dv.3"
"scptrt5o3.6"        "scptrt5o3.5"
[337] "nch3resp.1"       "scptrt5e1.4"        "visfrnds.2"         "scptrt5a1.7"
"cgvfc_dv"           "nsran.2"
[343] "big5e_dv.6"       "siminc.2"           "siminc.3"           "netpuse.2"
"socweb.2"           "scac.3"
[349] "agegr10_dv.7"     "tvm2.5"             "big5n_dv.2"         "scopngbhd.2"
"intdatm_dv.11"      "intdatm_dv.9"
[355] "intdatm_dv.8"     "btype96.1"          "scopngbhe.4"        "jbstat.8"
"scsf1.4"            "scsf2a.2"
[361] "scopngbhc.5"      "lvrel5.1"           "jbstat.97"          "nnmpsp_dv.2"
"buno_dv.4"          "pno.3"
[367] "hhresp_dv.3"      "npensioner_dv.1"    "sub7_2"             "nbrcoh3.5"
"locserb.4"          "locserd.4"
[373] "locserc.4"        "locserc.3"          "locserc.5"          "voteintent.3"
"wrdrcl.4"           "nbrcoh1.4"
[379] "ff_bentype12.1"   "scopngbhd.3"        "civicduty.5"        "ienddathh.15"
"ienddathh.16"       "ienddathh.22"
[385] "perpolinf.6"      "perpolinf.8"        "perpolinf.10"       "nch5to15.4"
```

```
"hhtype_dv.6"       "adcts.2"
[391] "newsmain.8"        "netkn_1.2"         "natch01.4"         "adoptch01.4"
"sampst.2"          "cgna_dv.1"
[397] "natch01.5"         "ff_bentype07.1"    "btype6.1"          "fimnsben_dv"
"scopngbha.3"       "scptrt5e3.3"
[403] "scptrt5a2.3"       "simjob.3"          "scptrt5n2.6"       "scptrt5n2.4"
"scptrt5n2.5"       "scptrt5n2.3"
[409] "big5o_dv.7"        "istrtdatss"        "marstat_dv.3"      "memorig.5"
"indbd91_lw"        "memorig.6"
[415] "depchl_dv.2"       "newsmain.11"       "hhtype_dv.22"      "hhtype_dv.16"
"hcondn6.1"         "mnspno.1"
[421] "npns_dv.1"         "hhtype_dv.18"      "lvrel6.1"          "netwr_1.2"
"sub7_1.92"         "cgs7ca_dv.4"
[427] "orga1.1"           "simrace.2"         "visfam.3"          "newsource8.1"
"scptrt5a1.5"       "nasofa.2"
[433] "netlv_1.2"         "voteintent.7"      "indscub_xw"        "ethn_dv"
"big5c_dv.2"        "urban_dv.2"
[439] "ff_bentype08.1"    "fnspno.3"          "scptrt5n3.2"       "orga11.1"
"poleff2.4"         "sex_dv.2"
[445] "ff_bentype27.1"    "bensta5.1"         "scptrt5c3.5"       "scptrt5c3.6"
"clangab.3"         "health.2"
[451] "ff_bentype09.1"    "ff_jbstat.7"       "ff_bentype24.1"    "npensioner_dv.3"  "hrpno.3"
"orga16.1"
[457] "orga3.1"           "poleff2.2"         "cgwri_dv.5"        "gor_dv.10"
"scsf3b.2"          "ienddathh.12"
[463] "locserd.3"         "fibenothr_tc.1"    "finnow.2"          "nbrcoh3.4"
"bensta2.1"         "nnsib_dv.3"
[469] "big5o_dv.2"        "hiqual_dv.4"       "hiqual_dv.3"       "netpuse.6"
"tvhours"           "ff_bentype05.1"
[475] "poleff3.5"         "ff_bentype10.1"    "big5e_dv.2"        "natch01.1"
"cgwrd_dv.4"        "scopngbhh.5"
[481] "hood15.6"          "poleff4.3"         "nch415resp.3"      "fimnlabnet_dv"
"scopngbhg.3"       "scptrt5a1.6"
[487] "big5a_dv.3"        "perpolinf.3"       "scsf5.5"           "voteintent.4"
"netph_2.4"         "cgsrmem_dv.5"
[493] "big5e_dv.7"        "netpuse.5"         "scptrt5e1.6"       "scptrt5e3.7"
"cogdist.2"         "cgwri_dv.2"
[499] "simfam.2"          "scptrt5n2.2"       "perbfts.3"         "perbfts.4"
"sclfsat1.4"        "aidxhh.2"
[505] "cgivna1_dv.4"      "orga5.1"           "scopngbhg.5"       "nch5resp.1"
"jbstat.5"          "ff_jbstat.9"
[511] "scopngbhf.2"       "scptrt5a3.4"       "scptrt5a3.3"       "scptrt5a3.5"
"scptrt5a3.2"       "scptrt5c3.2"
[517] "nbrcoh4.2"         "crdark.4"          "scptrt5o1.2"       "scptrt5o1.4"
"hconde96.1"        "hcondn5.1"
[523] "hcondn10.1"        "gor_dv.8"          "nadoptch.1"        "grpbfts.5"
"netet_2"           "nmpsp_dv.2"
[529] "allch03.4"         "sub7_5"            "orga13.1"          "marstat_dv.4"
"nch10to15.1"       "hood15.2"
[535] "xpmove.2"          "perpolinf.4"       "locserap.2"        "locserap.3"
"locseras.3"        "ienddatss"
[541] "allch01.5"         "allch03.7"         "grpbfts.3"         "ienddathh.11"
"hcondn4.1"         "scac.2"
[547] "perpolinf.7"       "sclfsat1.5"        "sclfsat1.3"        "sclfsat1.6"
"sclfsat1.2"        "natch04.5"
[553] "intdatm_dv.4"      "sclfsat2.6"        "sclfsat2.3"        "sclfsat2.7"
"sclfsat2.5"        "cgwrd_dv.10"
[559] "fiyrinvinc_dv"     "scptrt5e3.6"       "scptrt5e3.2"       "poleff1.3"
"netlv_1.4"         "netph_1.2"
[565] "big5n_dv.3"        "tvm2.2"            "scptrt5c1.5"       "scptrt5c1.6"
"big5c_dv.7"        "scptrt5c1.4"
[571] "scptrt5c2.3"       "scptrt5e1.3"       "scptrt5e3.5"       "aidhh.2"
"netlv_1.5"         "hcondn13.1"
[577] "hcondn1.1"         "newsource7.1"      "cgna_dv.4"         "ndepchl_dv.4"
"allch03.6"         "nacar.2"
[583] "cgna_dv.2"         "civicduty.3"       "civicduty.4"       "polcost.3"
"polcost.2"         "polcost.4"
[589] "hcondn2.1"         "lvrel4.1"          "hcondn12.1"        "simrace.3"
```

```
"scsf6b.5"          "hcondn7.1"
[595] "sclfsat7.2"        "nbrcoh2.3"         "intdatm_dv.2"     "undqus.2"
"scsf5.4"           "scsf5.3"
[601] "scsf6b.3"          "hrpno.2"           "netjb_1.5"        "netjb_1.3"
"scptrt5c1.3"       "scptrt5a2.5"
[607] "scopngbhh.2"       "scopngbhc.2"       "scopngbhc.4"      "scopngbhc.3"
"scopngbhd.4"       "hood15.5"
[613] "ivcoop.3"          "ff_bentype33.1"
[1] "----------Ordinary Linear Regression (Improved)----------"

Call:
lm(formula = y ~ ., data = as.data.frame(x.data.linear))

Residuals:
     Min       1Q   Median       3Q      Max
-11.2032  -1.7256  -0.1398   1.5329  15.6311

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.32509    0.50381  12.555  < 2e-16 ***
wrdrcl.2      -0.22191    0.10749  -2.065 0.039032 *
lkmove.2      -0.24864    0.10331  -2.407 0.016137 *
nbrcoh2.5      1.25181    0.51439   2.434 0.014993 *
simarea.2      0.28441    0.11000   2.585 0.009760 **
simarea.3      0.28024    0.12381   2.263 0.023658 *
netlv_1.3      0.25905    0.10556   2.454 0.014164 *
netjb_2.2     -0.27679    0.12355  -2.240 0.025122 *
netjb_2.6     -0.32233    0.14817  -2.175 0.029658 *
orga15.1       1.13883    0.41422   2.749 0.005998 **
ftedany.2      1.11094    0.39329   2.825 0.004755 **
trainany.2     0.21637    0.10578   2.046 0.040862 *
lvrel2.1      -0.29283    0.10472  -2.796 0.005194 **
visfam.2       0.39755    0.18107   2.196 0.028178 *
bensta6.1     -1.29951    0.53778  -2.416 0.015717 *
finnow.3       0.45511    0.11780   3.864 0.000113 ***
finnow.4       0.89903    0.19196   4.683 2.91e-06 ***
finnow.5       2.47257    0.34025   7.267 4.38e-13 ***
poleff3.2     -0.60124    0.16796  -3.580 0.000348 ***
poleff3.3     -0.58103    0.17521  -3.316 0.000920 ***
poleff3.4     -0.87915    0.18106  -4.856 1.24e-06 ***
newsource5.1   0.19858    0.09637   2.061 0.039400 *
newsmain.2     1.17453    0.47565   2.469 0.013578 *
scsf1.5        1.16144    0.30879   3.761 0.000171 ***
scsf4a.2       2.38637    0.36820   6.481 1.02e-10 ***
scsf4a.3       1.11031    0.20919   5.308 1.17e-07 ***
scsf4a.4       0.49847    0.14295   3.487 0.000493 ***
scsf4b.2       1.92593    0.40795   4.721 2.43e-06 ***
scsf4b.3       1.16145    0.20352   5.707 1.23e-08 ***
scsf4b.4       0.62238    0.13293   4.682 2.93e-06 ***
scsf6a.3       0.86901    0.11827   7.348 2.42e-13 ***
scsf6a.4       2.27025    0.18184  12.485  < 2e-16 ***
scsf6a.5       4.66135    0.33612  13.868  < 2e-16 ***
scsf6b.4       0.67171    0.15224   4.412 1.05e-05 ***
scsf6c.2       5.90540    0.30314  19.481  < 2e-16 ***
scsf6c.3       2.59421    0.17052  15.214  < 2e-16 ***
scsf6c.4       1.33075    0.12172  10.933  < 2e-16 ***
scsf7.2        2.58438    0.30629   8.438  < 2e-16 ***
scsf7.3        1.01494    0.18506   5.484 4.40e-08 ***
scsf7.4        0.74615    0.14361   5.195 2.14e-07 ***
scopngbha.4    0.51083    0.23131   2.208 0.027267 *
scopngbha.5    1.84077    0.48884   3.766 0.000168 ***
scopngbhe.5   -1.49150    0.51548  -2.893 0.003831 **
sclfsat2.2    -0.42449    0.15862  -2.676 0.007479 **
sclfsat2.4    -0.43259    0.15823  -2.734 0.006287 **
sclfsat7.5    -0.24744    0.12015  -2.059 0.039510 *
sclfsato.2     1.22103    0.21847   5.589 2.43e-08 ***
sclfsato.3     1.94409    0.20271   9.591  < 2e-16 ***
sclfsato.4     0.90535    0.20176   4.487 7.41e-06 ***
```

```
sclfsato.5        0.54451      0.14172    3.842 0.000124 ***
sclfsato.7       -0.77467      0.16750   -4.625 3.86e-06 ***
scptrt5n1.3       0.41902      0.15462    2.710 0.006758 **
scptrt5n1.4       0.63644      0.15278    4.166 3.17e-05 ***
scptrt5n1.5       1.06894      0.15917    6.716 2.13e-11 ***
scptrt5n1.6       1.42815      0.18761    7.612 3.32e-14 ***
scptrt5n1.7       2.22834      0.20693   10.769  < 2e-16 ***
scptrt5o1.3       0.32263      0.13998    2.305 0.021224 *
scptrt5a2.2       0.62981      0.25546    2.465 0.013727 *
scptrt5a2.6       0.33186      0.10353    3.205 0.001359 **
scptrt5c2.5      -0.48582      0.17229   -2.820 0.004829 **
scptrt5c2.6      -0.97209      0.25967   -3.744 0.000184 ***
scptrt5n3.6      -0.29348      0.12663   -2.318 0.020522 *
scptrt5n3.7      -0.62388      0.19656   -3.174 0.001515 **
sub7_1.83         3.87032      1.27647    3.032 0.002444 **
sub7_1.95         6.30398      2.20808    2.855 0.004326 **
sevenspap.2       5.10498      1.38174    3.695 0.000223 ***
ienddathh.18      0.52205      0.17105    3.052 0.002288 **
ienddathh.19      0.30351      0.13946    2.176 0.029594 *
ff_jbstat.6      -0.44147      0.21149   -2.087 0.036912 *
ff_bentype13.1    2.69919      1.25164    2.157 0.031100 *
ff_bentype14.1   -1.72510      0.42770   -4.033 5.60e-05 ***
ff_bentype15.1    0.86417      0.36163    2.390 0.016911 *
ff_bentype16.1   -1.22957      0.39655   -3.101 0.001944 **
agegr13_dv.3     -1.28396      0.38395   -3.344 0.000833 ***
ppno.5           -4.33309      1.77778   -2.437 0.014837 *
fnspno.4          3.09070      1.39119    2.222 0.026363 *
big5c_dv.3        1.07864      0.32115    3.359 0.000790 ***
big5c_dv.4        0.46613      0.16647    2.800 0.005132 **
big5n_dv.7        1.99953      0.36754    5.440 5.63e-08 ***
cgwri_dv.3       -1.19340      0.47889   -2.492 0.012741 *
cgs7ca_dv.5       0.29610      0.11531    2.568 0.010270 *
fibenothr_if      0.10644      0.04920    2.163 0.030568 *
gor_dv.11         0.39588      0.17913    2.210 0.027160 *
qfhighfl_dv.1    -0.60566      0.21242   -2.851 0.004377 **
marstat_dv.5      1.01691      0.49605    2.050 0.040428 *
agegr10_dv.3     -0.52140      0.17078   -3.053 0.002280 **
agegr10_dv.6      0.24263      0.12365    1.962 0.049801 *
qfhigh_dv        -0.14013      0.05559   -2.521 0.011749 *
intdatm_dv.6     -0.35425      0.16951   -2.090 0.036691 *
cgivna1_dv.2      1.00659      0.45553    2.210 0.027180 *
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1       1

Residual standard error: 3.04 on 4106 degrees of freedom
Multiple R-squared:  0.6568,    Adjusted R-squared:  0.6494
F-statistic:  88.3 on 89 and 4106 DF,  p-value: < 2.2e-16

AIC: 21329.6628
MSE: 9.0438
[1] "The MSE of the predicted values are of 12.6348"
[1] "The Linear Model predicts exactly with accuracy of 0.1716"
[1] "The Linear Model predicts within a confidence interval with accuracy of 0.361"
[1] "-----------Elastic Net Regression-----------"
722 x 1 sparse Matrix of class "dgCMatrix", with 29 entries
          names Estimate_Coefs
1   (Intercept)   10.772054594
2       lvrel6.1  -0.152098817
3       finnow.3   0.021318487
4       finnow.4   0.403978812
5       finnow.5   1.831646368
6      poleff3.4  -0.101001374
7       scsf6a.4   0.170637112
8       scsf6a.5   0.198148380
9       scsf6c.2   2.430406712
10      scsf6c.3   0.289260947
11       scsf7.2   0.160804397
12    sclfsat1.3   0.012429074
```

```
13     sclfsat1.7    -0.076729959
14     sclfsat2.3     0.020980244
15     sclfsato.2     0.025315221
16     sclfsato.3     0.877483089
17     sclfsato.6    -0.634674395
18     sclfsato.7    -1.074582092
19   scptrt5n1.5     0.162193831
20   scptrt5n1.6     0.469914707
21   scptrt5n1.7     1.344284057
22   scptrt5n3.6    -0.108151802
23   scptrt5n3.7    -0.233878768
24    sf12mcs_dv    -2.915019293
25    big5n_dv.2    -0.090316823
26    big5n_dv.7     0.655606088
27    sf12pcs_dv    -0.595613175
28 agegr10_dv.6     0.003087375
29    nnsib_dv.1    -0.161030480
[1] "The MSE of the predicted values of the best fit model is 10.8909"
[1] "The Alpha of the best fit model is 1"
[1] "The Elastic Net Model predicts exactly with accuracy of 0.178"
[1] "—————Timer Results—————"
   user  system elapsed
1529.39    12.11 1542.41
```

10.2.21 w3Merge console

```
[1] "—————Initial Checks—————"
[1] "69981598 NA cells were found across the entire dataset (74.38% of data as NA)"
[1] "—————Data Type Checks—————"
[1] "0 variables recoded since all their entries aren't numeric or NA"
[1] "NOTE: algorithim recodes categorical data alphabetically e.g (female = 1, male = 2)"
character(0)
[1] "—————Low Data Removal—————"
[1] "2493 variables removed since they had >= 'naPercent' (default 20%) NA values"
   [1] "pid"            "childpno"       "lvwhy"          "lvmthp"
"lvyrp"          "liwho"
   [7] "lihow"          "mvever"         "mvmnth"         "mvyr"
"mlstatchk"      "mlstat"
  [13] "drive"          "caruse"         "britid"         "ukborn"
"plbornc"        "yr2uk4"
  [19] "citzn1"         "citzn2"         "citzn3"         "qfhigh"
"qualoc"         "qfvoc1"
  [25] "qfvoc2"         "qfvoc3"         "qfvoc4"         "qfvoc5"
"qfvoc6"         "qfvoc7"
  [31] "qfvoc8"         "qfvoc9"         "qfvoc10"        "qfvoc11"
"qfvoc12"        "qfvoc13"
  [37] "qfvoc14"        "qfvoc15"        "qfvoc96"        "school"
"scend"          "schlloc"
  [43] "schok"          "fenow"          "feend"          "j1none"
"j1semp"         "j1boss"
  [49] "j1mngr"         "edtype"         "edasp"          "lvscdo"
"ahvwell"        "fedlik"
  [55] "fednt"          "futureint"      "lvhm"           "ocimpa"
"ocimpb"         "ocimpe"
  [61] "ocimpf"         "ocimpi"         "ocimpk"         "ocimpl"
"futra"          "futrb"
  [67] "futrc"          "futrd"          "futre"          "futrf"
"futrg"          "futrh"
  [73] "futri"          "futrj"          "futrk"          "futrl"
"paju"           "maju"
  [79] "pacob"          "payruk"         "macob"          "mayruk"
"natid1"         "natid2"
  [85] "natid3"         "natid4"         "natid5"         "natid6"
"natid97"        "racel"
  [91] "racelo_code"    "racelt"         "racelwt"        "racelmt"
"racelat"        "racelbt"
```

```
 [97] "racelot_code"      "oprlg"             "oprlg0ni"          "nirel"
"niact"             "oprlg0"
 [103] "oprlg1"           "unsafe1"           "unsafe2"           "unsafe3"
"unsafe4"           "unsafe5"
 [109] "unsafe6"          "unsafe7"           "unsafe8"           "unsafe9"
"unsafe10"          "unsafe11"
 [115] "unsafe96"         "unsafe97"          "unsafeo_code"      "resunsafe1_1"
"resunsafe2_1"      "resunsafe3_1"
 [121] "resunsafe4_1"     "resunsafe5_1"      "resunsafe6_1"      "resunsafe7_1"
"resunsafe8_1"      "resunsafe9_1"
 [127] "resunsafe96_1"    "resunsafe97_1"     "resunsafe1_2"      "resunsafe2_2"
"resunsafe3_2"      "resunsafe4_2"
 [133] "resunsafe5_2"     "resunsafe6_2"      "resunsafe7_2"      "resunsafe8_2"
"resunsafe9_2"      "resunsafe96_2"
 [139] "resunsafe97_2"    "resunsafe1_3"      "resunsafe2_3"      "resunsafe3_3"
"resunsafe4_3"      "resunsafe5_3"
 [145] "resunsafe6_3"     "resunsafe7_3"      "resunsafe8_3"      "resunsafe9_3"
"resunsafe96_3"     "resunsafe97_3"
 [151] "resunsafe1_4"     "resunsafe2_4"      "resunsafe3_4"      "resunsafe4_4"
"resunsafe5_4"      "resunsafe6_4"
 [157] "resunsafe7_4"     "resunsafe8_4"      "resunsafe9_4"      "resunsafe96_4"
"resunsafe97_4"     "resunsafe1_5"
 [163] "resunsafe2_5"     "resunsafe3_5"      "resunsafe4_5"      "resunsafe5_5"
"resunsafe6_5"      "resunsafe7_5"
 [169] "resunsafe8_5"     "resunsafe9_5"      "resunsafe96_5"     "resunsafe97_5"
"resunsafe1_6"      "resunsafe2_6"
 [175] "resunsafe3_6"     "resunsafe4_6"      "resunsafe5_6"      "resunsafe6_6"
"resunsafe7_6"      "resunsafe8_6"
 [181] "resunsafe9_6"     "resunsafe96_6"     "resunsafe97_6"     "resunsafe1_7"
"resunsafe2_7"      "resunsafe3_7"
 [187] "resunsafe4_7"     "resunsafe5_7"      "resunsafe6_7"      "resunsafe7_7"
"resunsafe8_7"      "resunsafe9_7"
 [193] "resunsafe96_7"    "resunsafe97_7"     "resunsafe1_8"      "resunsafe2_8"
"resunsafe3_8"      "resunsafe4_8"
 [199] "resunsafe5_8"     "resunsafe6_8"      "resunsafe7_8"      "resunsafe8_8"
"resunsafe9_8"      "resunsafe96_8"
 [205] "resunsafe97_8"    "resunsafe1_9"      "resunsafe2_9"      "resunsafe3_9"
"resunsafe4_9"      "resunsafe5_9"
 [211] "resunsafe6_9"     "resunsafe7_9"      "resunsafe8_9"      "resunsafe9_9"
"resunsafe96_9"     "resunsafe97_9"
 [217] "resunsafe1_10"    "resunsafe2_10"     "resunsafe3_10"     "resunsafe4_10"
"resunsafe5_10"     "resunsafe6_10"
 [223] "resunsafe7_10"    "resunsafe8_10"     "resunsafe9_10"     "resunsafe96_10"
"resunsafe97_10"    "resunsafe1_11"
 [229] "resunsafe2_11"    "resunsafe3_11"     "resunsafe4_11"     "resunsafe5_11"
"resunsafe6_11"     "resunsafe7_11"
 [235] "resunsafe8_11"    "resunsafe9_11"     "resunsafe96_11"    "resunsafe97_11"
"resunsafe1_12"     "resunsafe2_12"
 [241] "resunsafe3_12"    "resunsafe4_12"     "resunsafe5_12"     "resunsafe6_12"
"resunsafe7_12"     "resunsafe8_12"
 [247] "resunsafe9_12"    "resunsafe96_12"    "resunsafe97_12"    "avoidance1"
"avoidance2"        "avoidance3"
 [253] "avoidance4"       "avoidance5"        "avoidance6"        "avoidance7"
"avoidance8"        "avoidance9"
 [259] "avoidance10"      "avoidance11"       "avoidance96"       "avoidance97"
"avoidanceo_code"   "resavoid1_1"
 [265] "resavoid2_1"      "resavoid3_1"       "resavoid4_1"       "resavoid5_1"
"resavoid6_1"       "resavoid7_1"
 [271] "resavoid8_1"      "resavoid9_1"       "resavoid96_1"      "resavoid97_1"
"resavoid1_2"       "resavoid2_2"
 [277] "resavoid3_2"      "resavoid4_2"       "resavoid5_2"       "resavoid6_2"
"resavoid7_2"       "resavoid8_2"
 [283] "resavoid9_2"      "resavoid96_2"      "resavoid97_2"      "resavoid1_3"
"resavoid2_3"       "resavoid3_3"
 [289] "resavoid4_3"      "resavoid5_3"       "resavoid6_3"       "resavoid7_3"
"resavoid8_3"       "resavoid9_3"
 [295] "resavoid96_3"     "resavoid97_3"      "resavoid1_4"       "resavoid2_4"
"resavoid3_4"       "resavoid4_4"
```

```
 [301] "resavoid5_4"      "resavoid6_4"      "resavoid7_4"      "resavoid8_4"
"resavoid9_4"      "resavoid96_4"
 [307] "resavoid97_4"     "resavoid1_5"      "resavoid2_5"      "resavoid3_5"
"resavoid4_5"      "resavoid5_5"
 [313] "resavoid6_5"      "resavoid7_5"      "resavoid8_5"      "resavoid9_5"
"resavoid96_5"     "resavoid97_5"
 [319] "resavoid1_6"      "resavoid2_6"      "resavoid3_6"      "resavoid4_6"
"resavoid5_6"      "resavoid6_6"
 [325] "resavoid7_6"      "resavoid8_6"      "resavoid9_6"      "resavoid96_6"
"resavoid97_6"     "resavoid1_7"
 [331] "resavoid2_7"      "resavoid3_7"      "resavoid4_7"      "resavoid5_7"
"resavoid6_7"      "resavoid7_7"
 [337] "resavoid8_7"      "resavoid9_7"      "resavoid96_7"     "resavoid97_7"
"resavoid1_8"      "resavoid2_8"
 [343] "resavoid3_8"      "resavoid4_8"      "resavoid5_8"      "resavoid6_8"
"resavoid7_8"      "resavoid8_8"
 [349] "resavoid9_8"      "resavoid96_8"     "resavoid97_8"     "resavoid1_9"
"resavoid2_9"      "resavoid3_9"
 [355] "resavoid4_9"      "resavoid5_9"      "resavoid6_9"      "resavoid7_9"
"resavoid8_9"      "resavoid9_9"
 [361] "resavoid96_9"     "resavoid97_9"     "resavoid1_10"     "resavoid2_10"
"resavoid3_10"     "resavoid4_10"
 [367] "resavoid5_10"     "resavoid6_10"     "resavoid7_10"     "resavoid8_10"
"resavoid9_10"     "resavoid96_10"
 [373] "resavoid97_10"    "resavoid1_11"     "resavoid2_11"     "resavoid3_11"
"resavoid4_11"     "resavoid5_11"
 [379] "resavoid6_11"     "resavoid7_11"     "resavoid8_11"     "resavoid9_11"
"resavoid96_11"    "resavoid97_11"
 [385] "resavoid1_12"     "resavoid2_12"     "resavoid3_12"     "resavoid4_12"
"resavoid5_12"     "resavoid6_12"
 [391] "resavoid7_12"     "resavoid8_12"     "resavoid9_12"     "resavoid96_12"
"resavoid97_12"    "insulted1"
 [397] "insulted2"        "insulted3"        "insulted4"        "insulted5"
"insulted6"        "insulted7"
 [403] "insulted8"        "insulted9"        "insulted10"       "insulted11"
"insulted96"       "insulted97"
 [409] "insultedo_code"   "resinsulted1_1"   "resinsulted2_1"   "resinsulted3_1"
"resinsulted4_1"   "resinsulted5_1"
 [415] "resinsulted6_1"   "resinsulted7_1"   "resinsulted8_1"   "resinsulted9_1"
"resinsulted96_1"  "resinsulted97_1"
 [421] "resinsulted1_2"   "resinsulted2_2"   "resinsulted3_2"   "resinsulted4_2"
"resinsulted5_2"   "resinsulted6_2"
 [427] "resinsulted7_2"   "resinsulted8_2"   "resinsulted9_2"   "resinsulted96_2"
"resinsulted97_2"  "resinsulted1_3"
 [433] "resinsulted2_3"   "resinsulted3_3"   "resinsulted4_3"   "resinsulted5_3"
"resinsulted6_3"   "resinsulted7_3"
 [439] "resinsulted8_3"   "resinsulted9_3"   "resinsulted96_3"  "resinsulted97_3"
"resinsulted1_4"   "resinsulted2_4"
 [445] "resinsulted3_4"   "resinsulted4_4"   "resinsulted5_4"   "resinsulted6_4"
"resinsulted7_4"   "resinsulted8_4"
 [451] "resinsulted9_4"   "resinsulted96_4"  "resinsulted97_4"  "resinsulted1_5"
"resinsulted2_5"   "resinsulted3_5"
 [457] "resinsulted4_5"   "resinsulted5_5"   "resinsulted6_5"   "resinsulted7_5"
"resinsulted8_5"   "resinsulted9_5"
 [463] "resinsulted96_5"  "resinsulted97_5"  "resinsulted1_6"   "resinsulted2_6"
"resinsulted3_6"   "resinsulted4_6"
 [469] "resinsulted5_6"   "resinsulted6_6"   "resinsulted7_6"   "resinsulted8_6"
"resinsulted9_6"   "resinsulted96_6"
 [475] "resinsulted97_6"  "resinsulted1_7"   "resinsulted2_7"   "resinsulted3_7"
"resinsulted4_7"   "resinsulted5_7"
 [481] "resinsulted6_7"   "resinsulted7_7"   "resinsulted8_7"   "resinsulted9_7"
"resinsulted96_7"  "resinsulted97_7"
 [487] "resinsulted1_8"   "resinsulted2_8"   "resinsulted3_8"   "resinsulted4_8"
"resinsulted5_8"   "resinsulted6_8"
 [493] "resinsulted7_8"   "resinsulted8_8"   "resinsulted9_8"   "resinsulted96_8"
"resinsulted97_8"  "resinsulted1_9"
 [499] "resinsulted2_9"   "resinsulted3_9"   "resinsulted4_9"   "resinsulted5_9"
"resinsulted6_9"   "resinsulted7_9"
```

```
 [505] "resinsulted8_9"   "resinsulted9_9"   "resinsulted96_9"  "resinsulted97_9"
"resinsulted1_10"  "resinsulted2_10"
 [511] "resinsulted3_10"  "resinsulted4_10"  "resinsulted5_10"  "resinsulted6_10"
"resinsulted7_10"  "resinsulted8_10"
 [517] "resinsulted9_10"  "resinsulted96_10" "resinsulted97_10" "resinsulted1_11"
"resinsulted2_11"  "resinsulted3_11"
 [523] "resinsulted4_11"  "resinsulted5_11"  "resinsulted6_11"  "resinsulted7_11"
"resinsulted8_11"  "resinsulted9_11"
 [529] "resinsulted96_11" "resinsulted97_11" "resinsulted1_12"  "resinsulted2_12"
"resinsulted3_12"  "resinsulted4_12"
 [535] "resinsulted5_12"  "resinsulted6_12"  "resinsulted7_12"  "resinsulted8_12"
"resinsulted9_12"  "resinsulted96_12"
 [541] "resinsulted97_12" "attacked1"        "attacked2"        "attacked3"
"attacked4"        "attacked5"
 [547] "attacked6"        "attacked7"        "attacked8"        "attacked9"
"attacked10"       "attacked11"
 [553] "attacked96"       "attacked97"       "attackedo_code"   "resattacked1_1"
"resattacked2_1"   "resattacked3_1"
 [559] "resattacked4_1"   "resattacked5_1"   "resattacked6_1"   "resattacked7_1"
"resattacked8_1"   "resattacked9_1"
 [565] "resattacked96_1"  "resattacked97_1"  "resattacked1_2"   "resattacked2_2"
"resattacked3_2"   "resattacked4_2"
 [571] "resattacked5_2"   "resattacked6_2"   "resattacked7_2"   "resattacked8_2"
"resattacked9_2"   "resattacked96_2"
 [577] "resattacked97_2"  "resattacked1_3"   "resattacked2_3"   "resattacked3_3"
"resattacked4_3"   "resattacked5_3"
 [583] "resattacked6_3"   "resattacked7_3"   "resattacked8_3"   "resattacked9_3"
"resattacked96_3"  "resattacked97_3"
 [589] "resattacked1_4"   "resattacked2_4"   "resattacked3_4"   "resattacked4_4"
"resattacked5_4"   "resattacked6_4"
 [595] "resattacked7_4"   "resattacked8_4"   "resattacked9_4"   "resattacked96_4"
"resattacked97_4"  "resattacked1_5"
 [601] "resattacked2_5"   "resattacked3_5"   "resattacked4_5"   "resattacked5_5"
"resattacked6_5"   "resattacked7_5"
 [607] "resattacked8_5"   "resattacked9_5"   "resattacked96_5"  "resattacked97_5"
"resattacked1_6"   "resattacked2_6"
 [613] "resattacked3_6"   "resattacked4_6"   "resattacked5_6"   "resattacked6_6"
"resattacked7_6"   "resattacked8_6"
 [619] "resattacked9_6"   "resattacked96_6"  "resattacked97_6"  "resattacked1_7"
"resattacked2_7"   "resattacked3_7"
 [625] "resattacked4_7"   "resattacked5_7"   "resattacked6_7"   "resattacked7_7"
"resattacked8_7"   "resattacked9_7"
 [631] "resattacked96_7"  "resattacked97_7"  "resattacked1_8"   "resattacked2_8"
"resattacked3_8"   "resattacked4_8"
 [637] "resattacked5_8"   "resattacked6_8"   "resattacked7_8"   "resattacked8_8"
"resattacked9_8"   "resattacked96_8"
 [643] "resattacked97_8"  "resattacked1_9"   "resattacked2_9"   "resattacked3_9"
"resattacked4_9"   "resattacked5_9"
 [649] "resattacked6_9"   "resattacked7_9"   "resattacked8_9"   "resattacked9_9"
"resattacked96_9"  "resattacked97_9"
 [655] "resattacked1_10"  "resattacked2_10"  "resattacked3_10"  "resattacked4_10"
"resattacked5_10"  "resattacked6_10"
 [661] "resattacked7_10"  "resattacked8_10"  "resattacked9_10"  "resattacked96_10" "resattacked97_10"
 [667] "resattacked2_11"  "resattacked3_11"  "resattacked4_11"  "resattacked5_11"
"resattacked6_11"  "resattacked7_11"
 [673] "resattacked8_11"  "resattacked9_11"  "resattacked96_11" "resattacked97_11" "resattacked1_12"
"resattacked2_12"
 [679] "resattacked3_12"  "resattacked4_12"  "resattacked5_12"  "resattacked6_12"
"resattacked7_12"  "resattacked8_12"
 [685] "resattacked9_12"  "resattacked96_12" "resattacked97_12" "servaccy1"
"servaccy2"        "servaccy3"
 [691] "servaccy4"        "servaccy5"        "servaccy6"        "servaccy7"
"servaccy8"        "servaccy9"
 [697] "servaccy10"       "servaccy11"       "servaccy12"       "servaccy13"
"servaccy14"       "servaccy15"
 [703] "servaccy97"       "sattransy1"       "sattransy2"       "sattransy3"
"sattransy4"       "sattransy5"
 [709] "sattransy6"       "sattransy7"       "sattransy8"       "sattransy9"
```

```
"sattransy10"       "sattransy97"
 [715] "crworb"            "netcht"            "visfrndsy1"        "visfrndsy2"
"visfrndsy3"        "visfrndsy4"
 [721] "visfrndsy5"        "visfrndsy6"        "visfrndsy7"        "visfrndsy8"
"visfrndsy9"        "visfrndsy10"
 [727] "visfrndsy11"       "visfrndsy12"       "visfrndsy13"       "visfrndsy14"
"visfrndsy97"       "netrl_1"
 [733] "netmet_1"          "netweb_1"          "nettalk1_1"        "nettalk2_1"
"nettalk3_1"        "nettalk4_1"
 [739] "nettalk5_1"        "nettalk6_1"        "nettalk7_1"        "nettalk8_1"
"nettalk9_1"        "nettalk10_1"
 [745] "nettalk11_1"       "nettalk12_1"       "nettalk13_1"       "nettalk97_1"
"nettalk96_1"       "netdo1_1"
 [751] "netdo2_1"          "netdo3_1"          "netdo4_1"          "netdo5_1"
"netdo6_1"          "netdo7_1"
 [757] "netdo8_1"          "netdo9_1"          "netdo10_1"         "netdo97_1"
"netdo96_1"         "netrl_2"
 [763] "netmet_2"          "netweb_2"          "nettalk1_2"        "nettalk2_2"
"nettalk3_2"        "nettalk4_2"
 [769] "nettalk5_2"        "nettalk6_2"        "nettalk7_2"        "nettalk8_2"
"nettalk9_2"        "nettalk10_2"
 [775] "nettalk11_2"       "nettalk12_2"       "nettalk13_2"       "nettalk97_2"
"nettalk96_2"       "netdo1_2"
 [781] "netdo2_2"          "netdo3_2"          "netdo4_2"          "netdo5_2"
"netdo6_2"          "netdo7_2"
 [787] "netdo8_2"          "netdo9_2"          "netdo10_2"         "netdo97_2"
"netdo96_2"         "netsx_3"
 [793] "netwr_3"           "netrl_3"           "netag_3"           "netkn_3"
"netph_3"           "netlv_3"
 [799] "netjb_3"           "netet_3"           "netmet_3"          "netweb_3"
"nettalk1_3"        "nettalk2_3"
 [805] "nettalk3_3"        "nettalk4_3"        "nettalk5_3"        "nettalk6_3"
"nettalk7_3"        "nettalk8_3"
 [811] "nettalk9_3"        "nettalk10_3"       "nettalk11_3"       "nettalk12_3"
"nettalk13_3"       "nettalk97_3"
 [817] "nettalk96_3"       "netdo1_3"          "netdo2_3"          "netdo3_3"
"netdo4_3"          "netdo5_3"
 [823] "netdo6_3"          "netdo7_3"          "netdo8_3"          "netdo9_3"
"netdo10_3"         "netdo97_3"
 [829] "netdo96_3"         "orgm1"             "orgm2"             "orgm3"
"orgm4"             "orgm5"
 [835] "orgm6"             "orgm7"             "orgm8"             "orgm9"
"orgm10"            "orgm11"
 [841] "orgm12"            "orgm13"            "orgm14"            "orgm15"
"orgm16"            "orgm96"
 [847] "orgmt1"            "orgmt2"            "orgmt3"            "orgmt4"
"orgmt5"            "orgmt6"
 [853] "orgmt7"            "orgmt8"            "orgmt9"            "orgmt10"
"orgmt11"           "orgmt12"
 [859] "orgmt13"           "orgmt14"           "orgmt15"           "orgmt16"
"orgmt96"           "orgat1"
 [865] "orgat2"            "orgat3"            "orgat4"            "orgat5"
"orgat6"            "orgat7"
 [871] "orgat8"            "orgat9"            "orgat10"           "orgat11"
"orgat12"           "orgat13"
 [877] "orgat14"           "orgat15"           "orgat16"           "orgat96"
"hospc1"            "hospdc1"
 [883] "hospc2"            "hospdc2"           "hospc3"            "hospdc3"
"hospc4"            "hospdc4"
 [889] "hospc5"            "hospdc5"           "hospc6"            "hospdc6"
"hospc7"            "hospdc7"
 [895] "hospc8"            "hospdc8"           "disdif1"           "disdif2"
"disdif3"           "disdif4"
 [901] "disdif5"           "disdif6"           "disdif7"           "disdif8"
"disdif9"           "disdif10"
 [907] "disdif11"          "disdif12"          "disdif96"          "hcond1"
"hcond2"            "hcond3"
 [913] "hcond4"            "hcond5"            "hcond6"            "hcond7"
```

```
"hcond8"           "hcond9"
 [919] "hcond10"          "hcond11"          "hcond12"          "hcond13"
"hcond14"          "hcond15"
 [925] "hcond16"          "hcond17"          "hcond96"          "hconds01"
"hconds02"         "hconds03"
 [931] "hconds04"         "hconds05"         "hconds08"         "hconds09"
"hconds10"         "hconds11"
 [937] "hconds12"         "hconds13"         "hconds14"         "hconds15"
"hconds16"         "hconds17"
 [943] "hconda01"         "hconda02"         "hconda03"         "hconda04"
"hconda05"         "hconda06"
 [949] "hconda07"         "hconda08"         "hconda09"         "hconda10"
"hconda11"         "hconda12"
 [955] "hconda13"         "hconda14"         "hconda15"         "hconda16"
"hconda17"         "aidhua1"
 [961] "aidhua2"          "aidhua3"          "aidhua4"          "aidhua5"
"aidhua6"          "aidhua7"
 [967] "aidhua8"          "aidhua9"          "aidhua10"         "aidhua11"
"aidhua12"         "aidhua13"
 [973] "aidhua14"         "aidhua15"         "aidhua16"         "naidxhh"
"aidhu1"           "aidhu2"
 [979] "aidhrs"           "aideft"           "lcohnpi"          "coh1bm"
"coh1by"           "coh1mr"
 [985] "coh1em"           "coh1ey"           "nmar"             "lmar1m"
"lmar1y"           "ladopt"
 [991] "lnadopt"          "lprnt"            "lnprnt"           "ch1by4"
"movy11"           "movy12"
 [997] "movy13"           "movy14"           "movy15"           "movy16"
 [ reached getOption("max.print") -- omitted 1493 entries ]
[1] "----------Low Level Removal----------"
[1] "If a level is removed from a variable you wish to keep, reccomended to manually merge levels too
[1] "level 5 in pno removed, 4 observations found"
[1] "level 6 in pno removed, 3 observations found"
[1] "level 7 in hhorig removed, 1 observations found"
[1] "level 7 in memorig removed, 0 observations found"
[1] "level 5 in nch14resp removed, 4 observations found"
[1] "level 6 in nch14resp removed, 1 observations found"
[1] "level 2 in nch3resp removed, 1 observations found"
[1] "level 2 in nch5resp removed, 2 observations found"
[1] "level 2 in nch8resp removed, 1 observations found"
[1] "level 5 in nch415resp removed, 2 observations found"
[1] "level 6 in nch415resp removed, 0 observations found"
[1] "level 5 in nchresp removed, 3 observations found"
[1] "level 6 in nchresp removed, 0 observations found"
[1] "level 5 in nchund18resp removed, 1 observations found"
[1] "level 6 in nchund18resp removed, 0 observations found"
[1] "level 6 in natch01 removed, 3 observations found"
[1] "level 7 in natch01 removed, 2 observations found"
[1] "level 2 in natch02 removed, 2 observations found"
[1] "level 7 in natch02 removed, 0 observations found"
[1] "level 8 in natch02 removed, 0 observations found"
[1] "level 3 in natch03 removed, 0 observations found"
[1] "level 7 in natch03 removed, 4 observations found"
[1] "level 9 in natch03 removed, 0 observations found"
[1] "level 7 in natch04 removed, 2 observations found"
[1] "level 10 in natch04 removed, 0 observations found"
[1] "level 6 in natch05 removed, 0 observations found"
[1] "level 8 in natch05 removed, 0 observations found"
[1] "level 8 in natch06 removed, 3 observations found"
[1] "level 6 in nnatch removed, 0 observations found"
[1] "level 4 in nadoptch removed, 0 observations found"
[1] "level 1 in adoptch01 removed, 2 observations found"
[1] "level 6 in adoptch01 removed, 1 observations found"
[1] "level 3 in adoptch02 removed, 1 observations found"
[1] "level 5 in adoptch02 removed, 3 observations found"
[1] "level 6 in adoptch02 removed, 2 observations found"
[1] "level 4 in adoptch03 removed, 0 observations found"
[1] "level 6 in adoptch04 removed, 0 observations found"
```

```
[1] "level 5 in nchunder16 removed, 3 observations found"
[1] "level 6 in nchunder16 removed, 0 observations found"
[1] "level 5 in nch5to15 removed, 0 observations found"
[1] "level 4 in nch10to15 removed, 1 observations found"
[1] "level 2 in nch10 removed, 3 observations found"
[1] "level 1 in allch01 removed, 3 observations found"
[1] "level 7 in allch01 removed, 0 observations found"
[1] "level 7 in allch02 removed, 1 observations found"
[1] "level 8 in allch02 removed, 0 observations found"
[1] "level 7 in allch03 removed, 1 observations found"
[1] "level 8 in allch03 removed, 0 observations found"
[1] "level 9 in allch03 removed, 0 observations found"
[1] "level 5 in allch04 removed, 3 observations found"
[1] "level 7 in allch04 removed, 1 observations found"
[1] "level 8 in allch04 removed, 0 observations found"
[1] "level 10 in allch04 removed, 0 observations found"
[1] "level 6 in allch05 removed, 0 observations found"
[1] "level 7 in allch05 removed, 0 observations found"
[1] "level 8 in allch05 removed, 0 observations found"
[1] "level 8 in allch06 removed, 0 observations found"
[1] "level 9 in jbstat removed, 1 observations found"
[1] "level 10 in jbstat removed, 2 observations found"
[1] "level 2 in nnewborn removed, 0 observations found"
[1] "level 1 in hcondn3 removed, 4 observations found"
[1] "level 1 in hcondn7 removed, 3 observations found"
[1] "level 1 in hcondn9 removed, 3 observations found"
[1] "level 1 in hcondn15 removed, 1 observations found"
[1] "level 11 in nrels2 removed, 1 observations found"
[1] "level 13 in nrels2 removed, 0 observations found"
[1] "level 17 in nrels2 removed, 1 observations found"
[1] "level 6 in visfam removed, 2 observations found"
[1] "level 1 in bensta3 removed, 2 observations found"
[1] "level 10 in newsmain removed, 2 observations found"
[1] "level 11 in newsmain removed, 4 observations found"
[1] "level 12 in newsmain removed, 4 observations found"
[1] "level 9 in tvm2 removed, 1 observations found"
[1] "level 9 in marstat removed, 1 observations found"
[1] "level 1 in sub7_1 removed, 1 observations found"
[1] "level 3 in sub7_1 removed, 2 observations found"
[1] "level 73 in sub7_1 removed, 1 observations found"
[1] "level 83 in sub7_1 removed, 2 observations found"
[1] "level 86 in sub7_1 removed, 1 observations found"
[1] "level 89 in sub7_1 removed, 1 observations found"
[1] "level 95 in sub7_1 removed, 3 observations found"
[1] "level 96 in sub7_1 removed, 1 observations found"
[1] "level 97 in sub7_1 removed, 1 observations found"
[1] "level 6 in vftctw removed, 1 observations found"
[1] "level 8 in vftctw removed, 1 observations found"
[1] "level 9 in vftctw removed, 2 observations found"
[1] "level 11 in vftctw removed, 4 observations found"
[1] "level 12 in vftctw removed, 1 observations found"
[1] "level 14 in vftctw removed, 1 observations found"
[1] "level 17 in vftctw removed, 1 observations found"
[1] "level 18 in vftctw removed, 1 observations found"
[1] "level 4 in clangab removed, 3 observations found"
[1] "level 4 in ivcoop removed, 4 observations found"
[1] "level 4 in undqus removed, 0 observations found"
[1] "level 2 in ivlieng removed, 4 observations found"
[1] "level 9 in ivlitrans removed, 0 observations found"
[1] "level 97 in ivlitrans removed, 0 observations found"
[1] "level 4 in hgbiom removed, 2 observations found"
[1] "level 5 in hgbiom removed, 3 observations found"
[1] "level 6 in hgbiom removed, 1 observations found"
[1] "level 3 in hgbiof removed, 3 observations found"
[1] "level 4 in hgbiof removed, 2 observations found"
[1] "level 2 in origadd removed, 0 observations found"
[1] "level 22 in istrtdathh removed, 3 observations found"
[1] "level 1 in fiyrinvinc_tc removed, 1 observations found"
```

```
[1] "level 4 in pn1pno removed, 0 observations found"
[1] "level 5 in pn1pno removed, 0 observations found"
[1] "level 6 in pn1pno removed, 0 observations found"
[1] "level 3 in pn2pno removed, 2 observations found"
[1] "level 4 in pn2pno removed, 0 observations found"
[1] "level 4 in pns1pno removed, 0 observations found"
[1] "level 5 in pns1pno removed, 0 observations found"
[1] "level 6 in pns1pno removed, 0 observations found"
[1] "level 3 in pns2pno removed, 3 observations found"
[1] "level 4 in pns2pno removed, 0 observations found"
[1] "level 9 in ff_jbstat removed, 3 observations found"
[1] "level 10 in ff_jbstat removed, 4 observations found"
[1] "level 1 in ff_bentype06 removed, 3 observations found"
[1] "level 1 in ff_bentype21 removed, 1 observations found"
[1] "level 1 in ff_bentype25 removed, 1 observations found"
[1] "level 1 in ff_bentype30 removed, 1 observations found"
[1] "level 1 in ff_bentype34 removed, 2 observations found"
[1] "level 1 in ff_bentype35 removed, 1 observations found"
[1] "level 1 in ff_ivintlang removed, 1 observations found"
[1] "level 9 in ff_ivintlang removed, 0 observations found"
[1] "level 1 in ngrp_dv removed, 1 observations found"
[1] "level 2 in ngrp_dv removed, 1 observations found"
[1] "level 2 in nnssib_dv removed, 1 observations found"
[1] "level 3 in nnssib_dv removed, 1 observations found"
[1] "level 4 in nnssib_dv removed, 0 observations found"
[1] "level 1 in xtra5min_dv removed, 4 observations found"
[1] "level 2 in agegr13_dv removed, 2 observations found"
[1] "level 9 in mastat_dv removed, 0 observations found"
[1] "level 4 in buno_dv removed, 2 observations found"
[1] "level 5 in buno_dv removed, 0 observations found"
[1] "level 5 in nchild_dv removed, 0 observations found"
[1] "level 6 in nchild_dv removed, 0 observations found"
[1] "level 5 in hrpno removed, 1 observations found"
[1] "level 6 in hrpno removed, 1 observations found"
[1] "level 5 in ppno removed, 4 observations found"
[1] "level 6 in ppno removed, 0 observations found"
[1] "level 5 in sppno removed, 0 observations found"
[1] "level 3 in fnpno removed, 0 observations found"
[1] "level 4 in fnpno removed, 0 observations found"
[1] "level 3 in fnspno removed, 0 observations found"
[1] "level 4 in fnspno removed, 0 observations found"
[1] "level 4 in mnpno removed, 0 observations found"
[1] "level 5 in mnpno removed, 0 observations found"
[1] "level 6 in mnpno removed, 0 observations found"
[1] "level 4 in mnspno removed, 0 observations found"
[1] "level 5 in mnspno removed, 0 observations found"
[1] "level 6 in mnspno removed, 0 observations found"
[1] "level 2 in grfpno removed, 0 observations found"
[1] "level 1 in grmpno removed, 0 observations found"
[1] "level 2 in nnmpsp_dv removed, 3 observations found"
[1] "level 4 in nnmpsp_dv removed, 0 observations found"
[1] "level 2 in nunmpsp_dv removed, 3 observations found"
[1] "level 1 in big5c_dv removed, 1 observations found"
[1] "level 2 in big5c_dv removed, 2 observations found"
[1] "level 0 in cgwri_dv removed, 1 observations found"
[1] "level 1 in cgwri_dv removed, 4 observations found"
[1] "level 0 in cgs7cs_dv removed, 2 observations found"
[1] "level 5 in cgvfw_dv removed, 4 observations found"
[1] "level 6 in cgvfw_dv removed, 0 observations found"
[1] "level 8 in cgvfw_dv removed, 0 observations found"
[1] "level 9 in cgvfw_dv removed, 0 observations found"
[1] "level 11 in cgvfw_dv removed, 0 observations found"
[1] "level 12 in cgvfw_dv removed, 0 observations found"
[1] "level 14 in cgvfw_dv removed, 0 observations found"
[1] "level 17 in cgvfw_dv removed, 0 observations found"
[1] "level 18 in cgvfw_dv removed, 0 observations found"
[1] "level 1 in cgna_dv removed, 3 observations found"
[1] "level 0.159655138850212 in ind5mus_lw removed, 0 observations found"
```

```
[1] "level 0.164244398474693 in ind5mus_lw removed, 0 observations found"
[1] "level 0.177194967865944 in ind5mus_xw removed, 0 observations found"
[1] "level 0.186433836817741 in ind5mus_xw removed, 0 observations found"
[1] "level 0.191792845726013 in ind5mus_xw removed, 0 observations found"
[1] "level 2 in ficode3 removed, 0 observations found"
[1] "level 3 in ficode4 removed, 1 observations found"
[1] "level 1 in ficode6 removed, 2 observations found"
[1] "level 1 in ficode21 removed, 2 observations found"
[1] "level 2 in ficode24 removed, 4 observations found"
[1] "level 1 in ficode25 removed, 0 observations found"
[1] "level 2 in ficode26 removed, 4 observations found"
[1] "level 2 in ficode28 removed, 2 observations found"
[1] "level 1 in ficode30 removed, 0 observations found"
[1] "level 1 in ficode35 removed, 2 observations found"
[1] "level 1 in xtra5minosm_dv removed, 0 observations found"
[1] "level 5 in ndepchl_dv removed, 0 observations found"
[1] "level 6 in ndepchl_dv removed, 0 observations found"
[1] "level 3 in nmpsp_dv removed, 1 observations found"
[1] "level 2 in nnsib_dv removed, 0 observations found"
[1] "level 3 in nnsib_dv removed, 0 observations found"
[1] "level 4 in nnsib_dv removed, 0 observations found"
[1] "level 1 in big5a_dv removed, 3 observations found"
[1] "195 total levels removed from 108 different variables. In total 231 observations deleted"
[1] "———Variance 0 Check———"
[1] "114 variables removed since their new variance was 0"
  [1] "ivfio"          "ioutcome"       "adstatus"       "natch05"
"natch06"        "natch07"        "natch08"
  [8] "natch09"        "natch10"        "natch11"        "natch12"
"natch13"        "natch14"        "natch15"
 [15] "natch16"        "adoptch04"      "adoptch05"      "adoptch06"
"adoptch07"      "adoptch08"      "adoptch09"
 [22] "adoptch10"      "adoptch11"      "adoptch12"      "adoptch13"
"adoptch14"      "adoptch15"      "adoptch16"
 [29] "allch05"        "allch06"        "allch07"        "allch08"
"allch09"        "allch10"        "allch11"
 [36] "allch12"        "allch13"        "allch14"        "allch15"
"allch16"        "lieng"          "litrans"
 [43] "chkresp"        "hcondn3"        "hcondn7"        "hcondn9"
"hcondn15"       "lvrel4"         "lvrel96"
 [50] "btype10"        "btype11"        "btype12"        "btype13"
"btype14"        "bensta3"        "bensta9"
 [57] "bensta10"       "bensta11"       "bensta12"       "bensta13"
"newsource6"     "newsource96"    "casiend"
 [64] "precog"         "wrdrecpre"      "ns242"          "vfpre"
"nuabpre"        "cogend"         "hearcomputer"
 [71] "liceng"         "lictrans"       "ivlieng"        "ivlitrans"
"origadd"        "indmode"        "fiyrinvinc_tc"
 [78] "intdatd_if"     "intdatm_if"     "intdaty_if"     "doby_if"
"age_if"         "ff_ivlolw"      "ff_everint"
 [85] "ff_bentype06"   "ff_bentype21"   "ff_bentype25"   "ff_bentype30"
"ff_bentype31"   "ff_bentype32"   "ff_bentype34"
 [92] "ff_bentype35"   "ff_bentype36"   "ff_ivintlang"   "ngrp_dv"
"xtra5min_dv"    "grfpno"         "grmpno"
 [99] "fiyrinvinc_if"  "ind5mus_lw"     "ind5mus_xw"     "wave"
"ficode6"        "ficode21"       "ficode25"
[106] "ficode30"       "ficode31"       "ficode32"       "ficode35"
"ficode36"       "nbrcohdk_dv"    "xtra5minosm_dv"
[113] "scflag_dv"      "cgs7n_dv"
[1] "———Dummy Variables———"
[1] "predictor variable count went from 478 to 1266"
[1] "———Variance 0 Check———"
[1] "149 variables removed since their new variance was 0"
  [1] "pno.5"          "pno.6"          "hhorig.7"       "memorig.7"
"nch14resp.5"    "nch14resp.6"    "nch3resp.2"
  [8] "nch5resp.2"     "nch8resp.2"     "nch415resp.5"   "nch415resp.6"
"nchresp.5"      "nchresp.6"      "nchund18resp.5"
 [15] "nchund18resp.6" "natch01.6"      "natch01.7"      "natch02.2"
"natch02.7"      "natch02.8"      "natch03.3"
```

```
 [22] "natch03.7"         "natch03.9"         "natch04.7"         "natch04.10"
"nnatch.5"          "nnatch.6"          "nadoptch.4"
 [29] "adoptch01.1"       "adoptch01.6"       "adoptch02.3"       "adoptch02.5"
"adoptch02.6"       "adoptch03.4"       "nchunder16.5"
 [36] "nchunder16.6"      "nch5to15.5"        "nch10to15.4"       "nch10.2"
"allch01.1"         "allch01.7"         "allch02.7"
 [43] "allch02.8"         "allch03.7"         "allch03.8"         "allch03.9"
"allch04.5"         "allch04.7"         "allch04.8"
 [50] "allch04.10"        "jbstat.9"          "jbstat.10"         "nnewborn.2"
"nrels2.11"         "nrels2.13"         "nrels2.17"
 [57] "visfam.6"          "newsmain.10"       "newsmain.11"       "newsmain.12"
"tvm2.9"            "marstat.9"         "sub7_1.3"
 [64] "sub7_1.73"         "sub7_1.83"         "sub7_1.86"         "sub7_1.89"
"sub7_1.95"         "sub7_1.96"         "sub7_1.97"
 [71] "vftctw.5"          "vftctw.6"          "vftctw.8"          "vftctw.9"
"vftctw.11"         "vftctw.12"         "vftctw.14"
 [78] "vftctw.17"         "vftctw.18"         "clangab.4"         "ivcoop.4"
"undqus.4"          "hgbiom.4"          "hgbiom.5"
 [85] "hgbiom.6"          "hgbiof.3"          "hgbiof.4"          "istrtdathh.22"
"pn1pno.4"          "pn1pno.5"          "pn1pno.6"
 [92] "pn2pno.3"          "pn2pno.4"          "pns1pno.4"         "pns1pno.5"
"pns1pno.6"         "pns2pno.3"         "pns2pno.4"
 [99] "ff_jbstat.9"       "ff_jbstat.10"      "nnssib_dv.2"       "nnssib_dv.3"
"nnssib_dv.4"       "mastat_dv.9"       "buno_dv.4"
[106] "buno_dv.5"         "nchild_dv.5"       "nchild_dv.6"       "hrpno.5"
"hrpno.6"           "ppno.5"            "ppno.6"
[113] "sppno.5"           "fnpno.3"           "fnpno.4"           "fnspno.3"
"fnspno.4"          "mnpno.4"           "mnpno.5"
[120] "mnpno.6"           "mnspno.4"          "mnspno.5"          "mnspno.6"
"nnmpsp_dv.2"       "nnmpsp_dv.4"       "nunmpsp_dv.2"
[127] "big5c_dv.2"        "cgwri_dv.1"        "cgvfw_dv.5"        "cgvfw_dv.6"
"cgvfw_dv.8"        "cgvfw_dv.9"        "cgvfw_dv.11"
[134] "cgvfw_dv.12"       "cgvfw_dv.14"       "cgvfw_dv.17"       "cgvfw_dv.18"
"cgna_dv.1"         "ficode3.2"         "ficode4.3"
[141] "ficode24.2"        "ficode26.2"        "ficode28.2"        "ndepchl_dv.5"
"ndepchl_dv.6"      "nmpsp_dv.3"        "nnsib_dv.2"
[148] "nnsib_dv.3"        "nnsib_dv.4"
[1] "——————K—Means——————"
[1] "15 clusters have been made for K—Means"
[1] "K—Means results as a table, the max value in each row is a simple way to define which cluster re

      0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
  1   0  1  0  1  1  7 22 14 15 20 14 13 16  4  6  2  1  0  2  3  1  2  0
2  3  0  0  1  0  0  1  0  0  1  0  0  0
  2   2  2  1  2  2  5 37 35 38 32 22 34 18 16 11  7  8  6  5  4  2  4  3
2  0  2  1  0  0  0  1  1  1  1  0  1  0
  3   0  0  0  0  1  0  4  8  4  4  3  4  5  2  2  2  1  0  2  0  2  0  1
0  0  0  0  0  0  0  0  0  0  0  0  0  1
  4   0  0  5  1  2  7 40 46 60 38 40 30 30 19 11  5 11  9 10  4  1  4  4
3  3  3  2  0  0  0  2  3  0  1  1  0  1
  5   0  0  0  0  2  3 15 16 11  8 12  9  8  5  6  2  2  1  1  1  2  2  1
0  2  2  0  1  0  0  0  1  0  0  0  0  0
  6   1  1  0  1  1  6 43 20 28 16 24 22 33 15  6  7  4  3  2  7  4  2  6
3  2  0  1  0  0  0  1  0  2  1  0  0  1
  7   0  1  0  0  1  3 10 13 10 11  6  8  9  2  3  2  1  3  2  0  0  2  1
1  1  0  2  0  0  0  0  1  0  0  1  0  0
  8   2  1  1  1  1  7 19 19 12 12 12 17 19 13  7  5  6  6  1  1  3  0  2
2  0  0  1  0  1  0  1  1  1  0  0  0  0
  9   0  0  0  2  1  5 10 18 14 19 17 17 13 10 11  2  3  4  3  1  3  2  2
1  2  2  0  0  1  1  0  0  0  0  1  0  1
  10  0  0  1  1  0  5 34 31 40 30 28 21 28 13 13  9  9  9  4  3  3  0  6
3  1  4  1  0  0  1  0  0  0  0  0  1  1
  11  0  1  1  2  5 12 35 43 33 30 37 42 38 19 11  3 14  6  6  8  7  6  3
6  3  3  5  1  1  2  0  0  0  1  0  1  0
  12  0  0  2  0  4  8 42 29 39 36 24 26 26 16 13  6  5  7  5 10  2  7  4
3  4  0  0  0  1  0  0  0  0  0  0  0  0
  13  0  1  0  0  3  8 31 34 40 27 24 27 21  8 10 11  7  1  7  3  5  5  1
3  2  1  0  1  2  1  0  0  0  0  0  1  0
```

```
  14  0  0  0  0  0  0  2  0  0  1  0  2  0  0  1  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0
  15  0  2  1  4  3 12 44 50 51 42 35 54 43 19 16 11 11 10  7  5  3  3  6
2  3  2  0  2  0  0  2  0  1  2  0  0  0
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
 [1] "Cluster 1: Within MSE 4474753972529, Size 153"     "Cluster 2: Within MSE 3516814985944699, Siz
 [3] "Cluster 3: Within MSE 7007148303718, Size 46"      "Cluster 4: Within MSE 9251324534650154, Siz
 [5] "Cluster 5: Within MSE 4480605996068, Size 113"     "Cluster 6: Within MSE 3435002413407388, Siz
 [7] "Cluster 7: Within MSE 2513571980872, Size 94"      "Cluster 8: Within MSE 87330277256724, Size
 [9] "Cluster 9: Within MSE 22536657459220, Size 166"    "Cluster 10: Within MSE 3480539750972536, S
[11] "Cluster 11: Within MSE 9373485885903954, Size 385" "Cluster 12: Within MSE 3569884227357624, S
[13] "Cluster 13: Within MSE 3438691020388832, Size 285" "Cluster 14: Within MSE 99086532899296, Size
[15] "Cluster 15: Within MSE 9339191092850974, Size 446"
[1] "Total between cluster MSE: 668608915740519552, Total within cluster MSE: 4809214770354206"
[1] "The K-Means model predicts exactly with an accuracy of 0.133"
[1] "-----------Correlation Checks-----------"
[1] "dvage removed, correlated with 35 other variable(s)"
[1] "birthy removed, correlated with 34 other variable(s)"
[1] "jbstat.4 removed, correlated with 35 other variable(s)"
[1] "ff_jbstat.4 removed, correlated with 31 other variable(s)"
[1] "btype4.1 removed, correlated with 28 other variable(s)"
[1] "age_dv removed, correlated with 30 other variable(s)"
[1] "pensioner_dv.2 removed, correlated with 26 other variable(s)"
[1] "plivpar.2 removed, correlated with 19 other variable(s)"
[1] "memcont11.1 removed, correlated with 18 other variable(s)"
[1] "doby_dv removed, correlated with 28 other variable(s)"
[1] "memcont13.1 removed, correlated with 17 other variable(s)"
[1] "subcont11.1 removed, correlated with 16 other variable(s)"
[1] "subcont13.1 removed, correlated with 15 other variable(s)"
[1] "ff_bentype01.1 removed, correlated with 20 other variable(s)"
[1] "ficode1.1 removed, correlated with 24 other variable(s)"
[1] "nscont11.1 removed, correlated with 14 other variable(s)"
[1] "pn1pno.1 removed, correlated with 15 other variable(s)"
[1] "nscont13.1 removed, correlated with 13 other variable(s)"
[1] "ficode18.1 removed, correlated with 20 other variable(s)"
[1] "nch14resp.2 removed, correlated with 12 other variable(s)"
[1] "nchresp.2 removed, correlated with 13 other variable(s)"
[1] "dmemcont11.1 removed, correlated with 12 other variable(s)"
[1] "pns1pno.1 removed, correlated with 14 other variable(s)"
[1] "natch01.2 removed, correlated with 11 other variable(s)"
[1] "netag_1 removed, correlated with 11 other variable(s)"
[1] "memcont12.1 removed, correlated with 11 other variable(s)"
[1] "dmemcont13.1 removed, correlated with 11 other variable(s)"
[1] "hhtype_dv.8 removed, correlated with 19 other variable(s)"
[1] "nch14resp.3 removed, correlated with 10 other variable(s)"
[1] "nchresp.3 removed, correlated with 11 other variable(s)"
[1] "subcont12.1 removed, correlated with 10 other variable(s)"
[1] "vfcont11.1 removed, correlated with 10 other variable(s)"
[1] "hgbiom.2 removed, correlated with 10 other variable(s)"
[1] "nch14resp.1 removed, correlated with 9 other variable(s)"
[1] "nchresp.1 removed, correlated with 10 other variable(s)"
[1] "marstat.2 removed, correlated with 10 other variable(s)"
[1] "nscont12.1 removed, correlated with 9 other variable(s)"
[1] "vfcont13.1 removed, correlated with 9 other variable(s)"
[1] "hgbiof.1 removed, correlated with 9 other variable(s)"
[1] "pn2pno.2 removed, correlated with 10 other variable(s)"
[1] "nchund18resp.2 removed, correlated with 9 other variable(s)"
[1] "jbhas.2 removed, correlated with 9 other variable(s)"
[1] "btype5.1 removed, correlated with 11 other variable(s)"
[1] "sex.2 removed, correlated with 8 other variable(s)"
[1] "nchresp.4 removed, correlated with 9 other variable(s)"
[1] "dmemcont12.1 removed, correlated with 8 other variable(s)"
[1] "nacont11.1 removed, correlated with 8 other variable(s)"
[1] "indpxus_lw removed, correlated with 9 other variable(s)"
[1] "nchund18resp.3 removed, correlated with 8 other variable(s)"
[1] "relup.6 removed, correlated with 8 other variable(s)"
[1] "employ.2 removed, correlated with 8 other variable(s)"
[1] "pns2pno.2 removed, correlated with 8 other variable(s)"
```

```
[1] "hhorig.3 removed, correlated with 7 other variable(s)"
[1] "nch415resp.1 removed, correlated with 7 other variable(s)"
[1] "vfcont12.1 removed, correlated with 7 other variable(s)"
[1] "nacont13.1 removed, correlated with 7 other variable(s)"
[1] "nchund18resp.4 removed, correlated with 7 other variable(s)"
[1] "natch02.4 removed, correlated with 7 other variable(s)"
[1] "nchunder16.2 removed, correlated with 8 other variable(s)"
[1] "nchunder16.3 removed, correlated with 8 other variable(s)"
[1] "npn_dv.2 removed, correlated with 7 other variable(s)"
[1] "agegr13_dv.13 removed, correlated with 10 other variable(s)"
[1] "mastat_dv.2 removed, correlated with 8 other variable(s)"
[1] "pidp removed, correlated with 6 other variable(s)"
[1] "memorig.3 removed, correlated with 6 other variable(s)"
[1] "jbstat.2 removed, correlated with 6 other variable(s)"
[1] "nacont12.1 removed, correlated with 6 other variable(s)"
[1] "ivprsnt.2 removed, correlated with 6 other variable(s)"
[1] "ff_emplw.2 removed, correlated with 7 other variable(s)"
[1] "indinub_lw removed, correlated with 6 other variable(s)"
[1] "natch04.6 removed, correlated with 6 other variable(s)"
[1] "allch01.3 removed, correlated with 6 other variable(s)"
[1] "cmroute.1 removed, correlated with 7 other variable(s)"
[1] "buno_dv.3 removed, correlated with 6 other variable(s)"
[1] "hidp removed, correlated with 5 other variable(s)"
[1] "natch03.5 removed, correlated with 5 other variable(s)"
[1] "nchunder16.1 removed, correlated with 6 other variable(s)"
[1] "allch03.5 removed, correlated with 7 other variable(s)"
[1] "ff_bentype18.1 removed, correlated with 7 other variable(s)"
[1] "cgivns1_dv.2 removed, correlated with 5 other variable(s)"
[1] "cgivns1_dv.3 removed, correlated with 5 other variable(s)"
[1] "cgivns1_dv.4 removed, correlated with 5 other variable(s)"
[1] "indpxub_lw removed, correlated with 5 other variable(s)"
[1] "scsf2a.3 removed, correlated with 5 other variable(s)"
[1] "npn_dv.1 removed, correlated with 6 other variable(s)"
[1] "nnmpsp_dv.1 removed, correlated with 5 other variable(s)"
[1] "fimnlabgrs_dv removed, correlated with 5 other variable(s)"
[1] "sex_dv.2 removed, correlated with 7 other variable(s)"
[1] "npensioner_dv.2 removed, correlated with 8 other variable(s)"
[1] "livesp_dv.1 removed, correlated with 7 other variable(s)"
[1] "pno.2 removed, correlated with 4 other variable(s)"
[1] "month removed, correlated with 4 other variable(s)"
[1] "nch14resp.4 removed, correlated with 4 other variable(s)"
[1] "nnatch.2 removed, correlated with 4 other variable(s)"
[1] "nnatch.4 removed, correlated with 4 other variable(s)"
[1] "nchunder16.4 removed, correlated with 5 other variable(s)"
[1] "allch02.4 removed, correlated with 5 other variable(s)"
[1] "btype8.1 removed, correlated with 4 other variable(s)"
[1] "hgbiom.3 removed, correlated with 4 other variable(s)"
[1] "fimngrs_tc.1 removed, correlated with 4 other variable(s)"
[1] "ff_bentype22.1 removed, correlated with 5 other variable(s)"
[1] "nchild_dv.2 removed, correlated with 6 other variable(s)"
[1] "fnpno.1 removed, correlated with 4 other variable(s)"
[1] "fnspno.1 removed, correlated with 5 other variable(s)"
[1] "indin91_lw removed, correlated with 4 other variable(s)"
[1] "indin01_lw removed, correlated with 5 other variable(s)"
[1] "indinus_lw removed, correlated with 4 other variable(s)"
[1] "cgivwrd1_dv.2 removed, correlated with 4 other variable(s)"
[1] "cgivwrd1_dv.3 removed, correlated with 4 other variable(s)"
[1] "cgivwrd1_dv.4 removed, correlated with 4 other variable(s)"
[1] "nnatch.1 removed, correlated with 4 other variable(s)"
[1] "scsf2b.3 removed, correlated with 4 other variable(s)"
[1] "nchild_dv.1 removed, correlated with 5 other variable(s)"
[1] "nchild_dv.3 removed, correlated with 5 other variable(s)"
[1] "single_dv.1 removed, correlated with 5 other variable(s)"
[1] "isyear.2012 removed, correlated with 3 other variable(s)"
[1] "natch01.5 removed, correlated with 3 other variable(s)"
[1] "jbstat.3 removed, correlated with 3 other variable(s)"
[1] "jbstat.7 removed, correlated with 3 other variable(s)"
[1] "jbstat.8 removed, correlated with 3 other variable(s)"
```

```
[1] "netag_2 removed, correlated with 3 other variable(s)"
[1] "relup.2 removed, correlated with 3 other variable(s)"
[1] "btype6.1 removed, correlated with 3 other variable(s)"
[1] "bensta7.1 removed, correlated with 3 other variable(s)"
[1] "perpolinf.11 removed, correlated with 3 other variable(s)"
[1] "hgbiom.1 removed, correlated with 3 other variable(s)"
[1] "pn1pno.3 removed, correlated with 3 other variable(s)"
[1] "ppno.1 removed, correlated with 3 other variable(s)"
[1] "paygu_if.1 removed, correlated with 3 other variable(s)"
[1] "indinub_xw removed, correlated with 3 other variable(s)"
[1] "indscub_lw removed, correlated with 4 other variable(s)"
[1] "fimnnet_dv removed, correlated with 4 other variable(s)"
[1] "cgivs71_dv.2 removed, correlated with 3 other variable(s)"
[1] "cgivs71_dv.3 removed, correlated with 3 other variable(s)"
[1] "cgivs71_dv.4 removed, correlated with 3 other variable(s)"
[1] "scsf3a.5 removed, correlated with 3 other variable(s)"
[1] "scsf4a.5 removed, correlated with 3 other variable(s)"
[1] "sub7_3 removed, correlated with 3 other variable(s)"
[1] "cgs7cs_dv.5 removed, correlated with 3 other variable(s)"
[1] "hhtype_dv.12 removed, correlated with 4 other variable(s)"
[1] "nmpsp_dv.1 removed, correlated with 3 other variable(s)"
[1] "fimnpen_dv removed, correlated with 4 other variable(s)"
[1] "psu removed, correlated with 2 other variable(s)"
[1] "isyear.2013 removed, correlated with 2 other variable(s)"
[1] "psiblings.2 removed, correlated with 2 other variable(s)"
[1] "natch02.3 removed, correlated with 2 other variable(s)"
[1] "nadoptch.2 removed, correlated with 2 other variable(s)"
[1] "adoptch01.3 removed, correlated with 3 other variable(s)"
[1] "allch01.2 removed, correlated with 2 other variable(s)"
[1] "allch02.3 removed, correlated with 3 other variable(s)"
[1] "allch04.6 removed, correlated with 2 other variable(s)"
[1] "istrtdaty.2012 removed, correlated with 2 other variable(s)"
[1] "btype2.1 removed, correlated with 2 other variable(s)"
[1] "btype3.1 removed, correlated with 2 other variable(s)"
[1] "bensta4.1 removed, correlated with 2 other variable(s)"
[1] "fiyrdia removed, correlated with 2 other variable(s)"
[1] "civicduty.6 removed, correlated with 2 other variable(s)"
[1] "scsf6a.2 removed, correlated with 2 other variable(s)"
[1] "scsf7.5 removed, correlated with 3 other variable(s)"
[1] "marstat.4 removed, correlated with 2 other variable(s)"
[1] "marstat.5 removed, correlated with 2 other variable(s)"
[1] "marstat.6 removed, correlated with 2 other variable(s)"
[1] "memper.3 removed, correlated with 2 other variable(s)"
[1] "memper.4 removed, correlated with 2 other variable(s)"
[1] "hgbiof.2 removed, correlated with 2 other variable(s)"
[1] "respm16.2 removed, correlated with 2 other variable(s)"
[1] "pn1pno.2 removed, correlated with 2 other variable(s)"
[1] "pns1pno.3 removed, correlated with 2 other variable(s)"
[1] "ff_jbstat.3 removed, correlated with 2 other variable(s)"
[1] "ff_jbstat.8 removed, correlated with 2 other variable(s)"
[1] "ff_bentype23.1 removed, correlated with 2 other variable(s)"
[1] "fimnsben_dv removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.6 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.7 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.8 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.9 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.10 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.11 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.12 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.13 removed, correlated with 2 other variable(s)"
[1] "cohab_dv.1 removed, correlated with 2 other variable(s)"
[1] "ppno.2 removed, correlated with 2 other variable(s)"
[1] "mnpno.1 removed, correlated with 2 other variable(s)"
[1] "mnpno.2 removed, correlated with 2 other variable(s)"
[1] "paynu_if.1 removed, correlated with 2 other variable(s)"
[1] "fimnlabgrs_if removed, correlated with 3 other variable(s)"
[1] "indns91_lw removed, correlated with 2 other variable(s)"
[1] "ficode22.1 removed, correlated with 3 other variable(s)"
```

```
[1] "cgivvf1_dv.2 removed, correlated with 2 other variable(s)"
[1] "cgivvf1_dv.3 removed, correlated with 2 other variable(s)"
[1] "cgivvf1_dv.4 removed, correlated with 2 other variable(s)"
[1] "nbrcoh4.4 removed, correlated with 2 other variable(s)"
[1] "netjb_1.6 removed, correlated with 2 other variable(s)"
[1] "scsf3b.4 removed, correlated with 2 other variable(s)"
[1] "scsf3b.5 removed, correlated with 2 other variable(s)"
[1] "scsf4b.5 removed, correlated with 2 other variable(s)"
[1] "scsf6c.5 removed, correlated with 2 other variable(s)"
[1] "scptrt5c1.7 removed, correlated with 2 other variable(s)"
[1] "sub7_1.93 removed, correlated with 2 other variable(s)"
[1] "sub7_4 removed, correlated with 2 other variable(s)"
[1] "nbrcoh_dv removed, correlated with 2 other variable(s)"
[1] "cgna_dv.3 removed, correlated with 2 other variable(s)"
[1] "jbiindb_dv removed, correlated with 2 other variable(s)"
[1] "fimnlabnet_dv removed, correlated with 2 other variable(s)"
[1] "pno.3 removed, correlated with 1 other variable(s)"
[1] "hhorig.4 removed, correlated with 1 other variable(s)"
[1] "hhorig.5 removed, correlated with 1 other variable(s)"
[1] "strata removed, correlated with 1 other variable(s)"
[1] "nch415resp.2 removed, correlated with 1 other variable(s)"
[1] "nch415resp.3 removed, correlated with 1 other variable(s)"
[1] "nchund18resp.1 removed, correlated with 1 other variable(s)"
[1] "natch01.1 removed, correlated with 1 other variable(s)"
[1] "natch03.4 removed, correlated with 1 other variable(s)"
[1] "nnatch.3 removed, correlated with 1 other variable(s)"
[1] "nadoptch.3 removed, correlated with 1 other variable(s)"
[1] "allch01.4 removed, correlated with 1 other variable(s)"
[1] "istrtdaty.2013 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.2 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.3 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.4 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.5 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.6 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.7 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.8 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.9 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.10 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.11 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.12 removed, correlated with 1 other variable(s)"
[1] "istrtdatd removed, correlated with 1 other variable(s)"
[1] "jbstat.6 removed, correlated with 1 other variable(s)"
[1] "locseras.2 removed, correlated with 1 other variable(s)"
[1] "locserb.2 removed, correlated with 1 other variable(s)"
[1] "locserd.2 removed, correlated with 1 other variable(s)"
[1] "locsere.2 removed, correlated with 1 other variable(s)"
[1] "locserc.2 removed, correlated with 1 other variable(s)"
[1] "nbrcoh1.2 removed, correlated with 1 other variable(s)"
[1] "nbrcoh1.5 removed, correlated with 1 other variable(s)"
[1] "nbrcoh2.2 removed, correlated with 1 other variable(s)"
[1] "nbrcoh3.2 removed, correlated with 1 other variable(s)"
[1] "simfam.3 removed, correlated with 1 other variable(s)"
[1] "netsx_1.2 removed, correlated with 1 other variable(s)"
[1] "netkn_1.3 removed, correlated with 1 other variable(s)"
[1] "netkn_2.3 removed, correlated with 1 other variable(s)"
[1] "netph_2.2 removed, correlated with 1 other variable(s)"
[1] "sf1.2 removed, correlated with 1 other variable(s)"
[1] "sf1.3 removed, correlated with 1 other variable(s)"
[1] "sf1.4 removed, correlated with 1 other variable(s)"
[1] "sf1.5 removed, correlated with 1 other variable(s)"
[1] "hcondn16.1 removed, correlated with 1 other variable(s)"
[1] "hconde6.1 removed, correlated with 1 other variable(s)"
[1] "hconde7.1 removed, correlated with 1 other variable(s)"
[1] "lvrel1.1 removed, correlated with 1 other variable(s)"
[1] "lvrel3.1 removed, correlated with 1 other variable(s)"
[1] "visfam.4 removed, correlated with 1 other variable(s)"
[1] "btype1.1 removed, correlated with 1 other variable(s)"
[1] "btype9.1 removed, correlated with 1 other variable(s)"
```

```
[1] "bensta2.1 removed, correlated with 1 other variable(s)"
[1] "bensta5.1 removed, correlated with 1 other variable(s)"
[1] "bensta6.1 removed, correlated with 1 other variable(s)"
[1] "bensta96.1 removed, correlated with 1 other variable(s)"
[1] "bensta97.1 removed, correlated with 1 other variable(s)"
[1] "finfut.2 removed, correlated with 1 other variable(s)"
[1] "vote6.2 removed, correlated with 1 other variable(s)"
[1] "polcost.2 removed, correlated with 1 other variable(s)"
[1] "perbfts.2 removed, correlated with 1 other variable(s)"
[1] "perbfts.6 removed, correlated with 1 other variable(s)"
[1] "grpbfts.2 removed, correlated with 1 other variable(s)"
[1] "demorient.2 removed, correlated with 1 other variable(s)"
[1] "poleff1.5 removed, correlated with 1 other variable(s)"
[1] "poleff2.3 removed, correlated with 1 other variable(s)"
[1] "scsf3a.2 removed, correlated with 1 other variable(s)"
[1] "scsf3a.3 removed, correlated with 1 other variable(s)"
[1] "scsf6b.2 removed, correlated with 1 other variable(s)"
[1] "scopngbha.2 removed, correlated with 1 other variable(s)"
[1] "scopngbha.5 removed, correlated with 1 other variable(s)"
[1] "scopngbhb.2 removed, correlated with 1 other variable(s)"
[1] "scopngbhe.2 removed, correlated with 1 other variable(s)"
[1] "scopngbhg.2 removed, correlated with 1 other variable(s)"
[1] "scptrt5n1.2 removed, correlated with 1 other variable(s)"
[1] "scptrt5o1.7 removed, correlated with 1 other variable(s)"
[1] "scptrt5a2.7 removed, correlated with 1 other variable(s)"
[1] "scptrt5e2.7 removed, correlated with 1 other variable(s)"
[1] "scptrt5n2.7 removed, correlated with 1 other variable(s)"
[1] "scptrt5o2.7 removed, correlated with 1 other variable(s)"
[1] "scptrt5a3.7 removed, correlated with 1 other variable(s)"
[1] "scptrt5c3.7 removed, correlated with 1 other variable(s)"
[1] "scptrt5o3.7 removed, correlated with 1 other variable(s)"
[1] "marstat.3 removed, correlated with 1 other variable(s)"
[1] "memper.2 removed, correlated with 1 other variable(s)"
[1] "memper.5 removed, correlated with 1 other variable(s)"
[1] "vftctw.1 removed, correlated with 1 other variable(s)"
[1] "vftctw.2 removed, correlated with 1 other variable(s)"
[1] "vftctw.3 removed, correlated with 1 other variable(s)"
[1] "vftctw.4 removed, correlated with 1 other variable(s)"
[1] "nadisease.2 removed, correlated with 1 other variable(s)"
[1] "vfct removed, correlated with 1 other variable(s)"
[1] "memprob1.1 removed, correlated with 1 other variable(s)"
[1] "memprob2.1 removed, correlated with 1 other variable(s)"
[1] "memprob3.1 removed, correlated with 1 other variable(s)"
[1] "memaid.2 removed, correlated with 1 other variable(s)"
[1] "clangab.2 removed, correlated with 1 other variable(s)"
[1] "respf16.2 removed, correlated with 1 other variable(s)"
[1] "istrtdathh.9 removed, correlated with 1 other variable(s)"
[1] "istrtdathh.10 removed, correlated with 1 other variable(s)"
[1] "istrtdathh.11 removed, correlated with 1 other variable(s)"
[1] "istrtdathh.12 removed, correlated with 1 other variable(s)"
[1] "istrtdathh.13 removed, correlated with 1 other variable(s)"
[1] "istrtdathh.14 removed, correlated with 1 other variable(s)"
[1] "istrtdathh.15 removed, correlated with 1 other variable(s)"
[1] "istrtdathh.16 removed, correlated with 1 other variable(s)"
[1] "istrtdathh.17 removed, correlated with 1 other variable(s)"
[1] "istrtdathh.18 removed, correlated with 1 other variable(s)"
[1] "istrtdathh.19 removed, correlated with 1 other variable(s)"
[1] "istrtdathh.20 removed, correlated with 1 other variable(s)"
[1] "istrtdathh.21 removed, correlated with 1 other variable(s)"
[1] "fimnlabgrs_tc.1 removed, correlated with 1 other variable(s)"
[1] "j2paynet_dv removed, correlated with 1 other variable(s)"
[1] "pns1pno.2 removed, correlated with 1 other variable(s)"
[1] "ff_jbstat.7 removed, correlated with 1 other variable(s)"
[1] "ff_bentype02.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype03.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype04.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype05.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype07.1 removed, correlated with 1 other variable(s)"
```

```
[1] "ff_bentype08.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype09.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype10.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype11.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype12.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype13.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype14.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype15.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype19.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype20.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype26.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype29.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype33.1 removed, correlated with 1 other variable(s)"
[1] "fimninvnet_dv removed, correlated with 1 other variable(s)"
[1] "intdaty_dv.2012 removed, correlated with 1 other variable(s)"
[1] "nnssib_dv.1 removed, correlated with 1 other variable(s)"
[1] "country.2 removed, correlated with 1 other variable(s)"
[1] "country.3 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.5 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.14 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.15 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.3 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.5 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.6 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.7 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.8 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.9 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.10 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.11 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.12 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.4 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.5 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.6 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.10 removed, correlated with 1 other variable(s)"
[1] "nchild_dv.4 removed, correlated with 1 other variable(s)"
[1] "ppno.3 removed, correlated with 1 other variable(s)"
[1] "ppno.4 removed, correlated with 1 other variable(s)"
[1] "sppno.1 removed, correlated with 1 other variable(s)"
[1] "fnpno.2 removed, correlated with 1 other variable(s)"
[1] "mnpno.3 removed, correlated with 1 other variable(s)"
[1] "mnspno.1 removed, correlated with 1 other variable(s)"
[1] "mnspno.2 removed, correlated with 1 other variable(s)"
[1] "hiqual_dv.9 removed, correlated with 1 other variable(s)"
[1] "cgsrmem_dv.3 removed, correlated with 1 other variable(s)"
[1] "cgsrmem_dv.4 removed, correlated with 1 other variable(s)"
[1] "cgs7cs_dv.4 removed, correlated with 1 other variable(s)"
[1] "cgs7ca_dv.1 removed, correlated with 1 other variable(s)"
[1] "fibenothr_if removed, correlated with 1 other variable(s)"
[1] "indnsub_lw removed, correlated with 1 other variable(s)"
[1] "indpxub_xw removed, correlated with 1 other variable(s)"
[1] "hhtype_dv.10 removed, correlated with 1 other variable(s)"
[1] "hhtype_dv.11 removed, correlated with 1 other variable(s)"
[1] "rach16_dv.2 removed, correlated with 1 other variable(s)"
[1] "frmnthimp_dv_total removed, correlated with 1 other variable(s)"
[1] "ethn_dv removed, correlated with 1 other variable(s)"
[1] "big5a_dv.5 removed, correlated with 1 other variable(s)"
[1] "cgivna1_dv.2 removed, correlated with 1 other variable(s)"
[1] "cgivna1_dv.3 removed, correlated with 1 other variable(s)"
[1] "cgivna1_dv.4 removed, correlated with 1 other variable(s)"
[1] "383 variables removed since they had high correlation coefs"
[1] "The final cleaned dataset has been completed at this stage and is stored under the name 'cleanDa
[1] "—————Attempting a Train Test Split—————"
[1] "Good train, test split found"
[1] "The working seed found was 3"
[1] "—————kNN—————"
[1] "175 neighbours considered for each test data point"
[1] "kNN results as a table, follow the diagonal for the correctly mapped clusters"
     real
```

```
predicted  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 3
        0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
        1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
        2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
        3  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
        4  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
        5  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
        6  1  1  1  1  7 10 70 70 63 43 34 32 24 18 13  4  8  3  4  4  1
3  2  2  1  1  0  1  1  0  0  0  1
        7  0  0  1  0  0  4 19 15 23 21 25 22 12  6  7  3  2  3  2  1  0
1  1  1  1  1  0  0  1  0  0  0  0
        8  1  1  0  0  0  5 11 12 15 20 18 27 24 10 11  6  8  5  4  2  3
2  3  2  1  1  0  0  1  0  0  0  0
        9  0  0  0  0  0  0  0  0  0  0  2  0  2  1  0  0  1  0  1  1  1
2  2  1  0  0  0  0  0  1  1  0  0
       10  0  0  0  0  0  0  0  0  0  1  1  2  2  0  0  1  1  0  0  1  0
0  1  0  0  0  0  0  0  0  0  0  0
       11  0  0  0  0  0  0  0  0  0  1  1  2  3  1  3  2  3  1  1  1  1
0  0  1  0  1  0  0  0  0  0  0  0
       12  0  0  0  0  0  0  0  0  0  1  1  0  3  2  0  1  2  0  2  1  0
3  1  3  1  2  1  0  0  0  0  0  0
       13  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1
0  0  0  0  0  0  0  0  0  0  1  0
       14  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
       15  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
       16  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
       17  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
       18  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
       19  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
       20  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
       21  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
       22  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
       23  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
       24  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
       25  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
       26  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
       27  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
       28  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
       29  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
 [ reached getOption("max.print") -- omitted 7 rows ]
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
[1] "The MSE of the predicted values are of 298.701"
[1] "The kNN model predicts exactly with an accuracy of 0.1228"
[1] "------------CART prediction model------------"
n= 2589
```

```
node), split, n, deviance, yval
      * denotes terminal node

 1) root 2589 73599.850 10.992280
   2) sf12mcs_dv>=-0.7669562 2085 24971.360  9.354916
     4) sf12mcs_dv>=0.1613441 1398 10027.130  8.263233
       8) scsf6c.4< 0.5 1024  5548.687  7.657227 *
       9) scsf6c.4>=0.5 374  3072.751  9.922460 *
     5) sf12mcs_dv< 0.1613441 687  9887.738 11.576420 *
   3) sf12mcs_dv< -0.7669562 504 19914.370 17.765870
     6) sf12mcs_dv>=-2.159407 416 10384.530 16.216350 *
     7) sf12mcs_dv< -2.159407 88  3809.273 25.090910
      14) scsf6c.3>=0.5 20   380.550 18.650000 *
      15) scsf6c.3< 0.5 68  2354.985 26.985290 *
[1] "Variable Importance"
 sf12mcs_dv     scsf4a.3      scsf7.3     scsf6a.4     scsf4b.3     scsf6c.2
scsf6c.3  sf12pcs_dv     scsf6c.4     scsf6a.3
39847.68022  8204.03154  6608.80318  5127.51971  5070.54727  4744.78073
2140.97414  1583.35964  1405.69233  1104.03798
   scsf4a.4     scsf4b.4     scsf6a.5  indbd91_lw  istrtdatss     memaid.3
netet_2   nbrcoh4.5      orga7.1  sclfsato.3
 1060.19009   721.30481   715.07143   214.74749   161.06061   130.01299
130.01299   107.37374   107.37374    26.30975
 fimngrs_dv
   18.79268
[1] "The MSE of the predicted values are of 10.7812"
[1] "The CART model predicts exactly with accuracy of 0.1564"
[1] "—————Ordinary Linear Regression (Initial)—————"
[1] "The full model AIC is: 13820.5524"
[1] "—————Variance Inflation Factor Removal—————"
[1] "The variable agegr10_dv.5 was removed since it had a VIF score of 485.394"
[1] "The variable ienddathh.19 was removed since it had a VIF score of 89.4811"
[1] "The variable newsmain.6 was removed since it had a VIF score of 80.8195"
[1] "The variable big5e_dv.5 was removed since it had a VIF score of 80.7827"
[1] "The variable nbrsnci_dv was removed since it had a VIF score of 69.8847"
[1] "The variable big5o_dv.4 was removed since it had a VIF score of 52.6054"
[1] "The variable sclfsato.6 was removed since it had a VIF score of 24.4873"
[1] "The variable sf12pcs_dv was removed since it had a VIF score of 24.2057"
[1] "The variable ndepchl_dv.2 was removed since it had a VIF score of 23.8388"
[1] "The variable big5n_dv.4 was removed since it had a VIF score of 21.7034"
[1] "The variable cgwrd_dv.6 was removed since it had a VIF score of 17.648"
[1] "The variable cgwri_dv.7 was removed since it had a VIF score of 17.3625"
[1] "The variable scptrt5n3.5 was removed since it had a VIF score of 16.8755"
[1] "The variable scptrt5e1.7 was removed since it had a VIF score of 16.7791"
[1] "The variable respm16_dv.2 was removed since it had a VIF score of 14.8577"
[1] "The variable sf12mcs_dv was removed since it had a VIF score of 14.621"
[1] "The variable natch01.3 was removed since it had a VIF score of 10.0156"
[1] "17 variables removed from the Ordinary Linear Model since they have a VIF score higher than 10"
[1] "The full model AIC after VIF checks is: 14060.8218"
[1] "—————Backwards Selection—————"
[1] "50 out of 732 variables removed so far."
[1] "100 out of 732 variables removed so far."
[1] "150 out of 732 variables removed so far."
[1] "200 out of 732 variables removed so far."
[1] "250 out of 732 variables removed so far."
[1] "300 out of 732 variables removed so far."
[1] "350 out of 732 variables removed so far."
[1] "400 out of 732 variables removed so far."
[1] "450 out of 732 variables removed so far."
[1] "500 out of 732 variables removed so far."
[1] "550 out of 732 variables removed so far."
[1] "600 out of 732 variables removed so far."
[1] "611 out of 732 variables removed in backwards selection since they weren't significant at the 95
  [1] "colbens2.2"       "hhresp_dv.3"      "gor_dv.2"         "ficode16.1"
"netpuse.4"        "relup.3"
  [7] "simarea.4"        "gor_dv.8"         "scopngbhd.3"      "orga10.1"
"locseras.4"       "sclfsat1.3"
 [13] "poleff4.4"        "sclfsat7.5"       "nnewborn.1"       "nrels2.6"
```

```
"simage.2"          "scptrt5e1.2"
 [19] "colbens1.4"        "hiqual_dv.5"       "scopngbhb.3"       "perpolinf.9"
"netjb_2.5"         "hhtype_dv.18"
 [25] "nunmpsp_dv.1"      "trainany.2"        "fnspno.2"          "netlv_2.5"
"orga8.1"           "locserap.4"
 [31] "scptrt5o1.3"       "big5o_dv.7"        "ficode11.1"        "orga7.1"
"scptrt5c1.4"       "gor_dv.3"
 [37] "voteintent.6"      "cgsrmem_dv.2"      "orga2.1"           "colbens1.5"
"ficode24.1"        "nsran.2"
 [43] "sclfsat1.5"        "ndepchl_dv.4"      "cgwrd_dv.7"        "intdatm_dv.2"
"siminc.2"          "scsf5.4"
 [49] "marstat_dv.4"      "cgwrd_dv.5"        "hhresp_dv.2"       "poleff2.5"
"simeduc.3"         "fibenothr_tc.1"
 [55] "simfam.5"          "scptrt5o3.4"       "ff_bentype24.1"    "netlv_2.2"
"demorient.3"       "grpbfts.4"
 [61] "orga14.1"          "finnow.2"          "scsf1.3"           "newsmain.9"
"nch10to15.1"       "nch10.1"
 [67] "netjb_1.3"         "newsource5.1"      "nbrcoh1.3"         "crwora.2"
"ienddathh.22"      "urban_dv.2"
 [73] "tvm2.6"            "sampst.2"          "ienddathh.16"      "locserd.4"
"marstat_dv.5"      "sub7_2"
 [79] "cgs7cs_dv.3"       "scptrt5o1.2"       "ficode2.1"         "ficode20.1"
"newsource12.1"     "scptrt5e3.4"
 [85] "hcondn12.1"        "respf16_dv.2"      "perpolinf.4"       "colbens2.3"
"scsf2a.2"          "scsf3b.3"
 [91] "lvrel10.1"         "scptrt5a3.2"       "orga3.1"           "jbstat.97"
"memorig.4"         "civicduty.3"
 [97] "civicduty.2"       "scsf2b.2"          "perbfts.4"         "scopngbhc.2"
"intdatm_dv.9"      "big5n_dv.6"
[103] "ff_jbstat.97"      "natch02.6"         "big5n_dv.5"        "hood15.4"
"marstat_dv.6"      "netjb_1.2"
[109] "cgwrd_dv.2"        "ficode12.1"        "hhtype_dv.5"       "mnspno.3"
"npns_dv.1"         "scptrt5e1.3"
[115] "ficode5.1"         "scptrt5c3.4"       "scptrt5c3.5"       "nch5to15.1"
"nch5to15.2"        "hhtype_dv.19"
[121] "nrels2.7"          "ficode9.1"         "scptrt5a2.5"       "scptrt5a2.4"
"scptrt5a1.4"       "scptrt5a2.2"
[127] "cgna_dv.5"         "indscus_lw"        "hhtype_dv.16"      "npensioner_dv.1" "depchl_dv.2"
"colbens2.8"
[133] "lvrel7.1"          "allch03.4"         "hcondn11.1"        "btype7.1"
"scsf3a.4"          "pno.4"
[139] "netjb_2.2"         "netsx_2.2"         "scptrt5e2.6"       "scptrt5e2.5"
"lvrel2.1"          "susp.2"
[145] "intdatm_dv.7"      "ienddathh.11"      "ienddathh.10"      "scopngbhd.2"
"cgs7cs_dv.1"       "ff_bentype38.1"
[151] "sub7_1.92"         "scptrt5a1.5"       "scptrt5e2.4"       "ff_jbstat.5"
"nch415resp.4"      "intdatm_dv.12"
[157] "hrpid"             "health.2"          "scopngbhh.3"       "ns200pre.2"
"cgs7ca_dv.4"       "cgs7ca_dv.3"
[163] "netpuse.2"         "ff_bentype17.1"    "perbfts.5"         "lenindintv"
"orga11.1"          "istrtdatmm"
[169] "ficode34.1"        "nch5to15.3"        "big5n_dv.2"        "perpolinf.10"
"ficode4.1"         "xpmove.2"
[175] "tvm2.97"           "clangab.3"         "vote6.3"           "orga12.1"
"cgwri_dv.4"        "lvrel6.1"
[181] "nch10to15.3"       "fibenothr_dv"      "fimnlabnet_tc.1"   "nch5resp.1"
"colbens1.9"        "netph_2.3"
[187] "newsmain.3"        "ficode39.1"        "wrdrcl.3"          "scsf1.2"
"ficode29.1"        "tvm2.7"
[193] "adoptch01.2"       "scptrt5c2.3"       "simfam.4"          "natch04.5"
"newsource10.1"     "scptrt5c3.2"
[199] "scptrt5a3.3"       "scptrt5a3.6"       "poleff4.3"         "poleff4.2"
"poleff4.5"         "npns_dv.2"
[205] "buno_dv.2"         "nnsib_dv.1"        "colbens1.10"       "intdatm_dv.10"
"intdatm_dv.6"      "intdatm_dv.4"
[211] "voteintent.8"      "nrels2.8"          "natch02.5"         "cgs7ca_dv.5"
"poleff1.4"         "cogdist.2"
[217] "cgs7ca_dv.2"       "locserc.5"         "newsmain.7"        "cgwrd_dv.10"
```

```
"locserd.3"       "cgvfc_dv"
[223] "ficode23.1"      "bensta1.1"       "btype96.1"       "gor_dv.11"
"netph_2.4"       "allch02.5"
[229] "locserb.4"       "voteintent.5"    "voteintent.9"    "voteintent.11"
"netet_1"         "simjob.2"
[235] "scptrt5n2.4"     "scptrt5n2.3"     "big5c_dv.6"      "hiqual_dv.3"
"nrels2.3"        "scptrt5a1.2"
[241] "sevenspap.3"     "ienddathh.20"    "ienddathh.21"    "ienddathh.23"
"socweb.2"        "perpolinf.3"
[247] "ficode2.3"       "newsource3.1"    "ienddathh.17"    "netlv_1.2"
"hhtype_dv.6"     "cgivwri1_dv.3"
[253] "susp.3"          "llknbrd.2"       "scptrt5e1.4"     "scac.3"
"scopngbhe.4"     "scptrt5a1.3"
[259] "nrels2.9"        "marstat_dv.3"    "relup.5"         "colbens2.1"
"scptrt5a1.7"     "scopngbhh.5"
[265] "agegr10_dv.6"    "nch10to15.2"     "allch01.5"       "ivcoop.3"
"indbdub_lw"      "civicduty.5"
[271] "scptrt5n2.5"     "colbens1.6"      "polcost.4"       "ficode13.1"
"org.2"           "ficode29.3"
[277] "ff_bentype28.1"  "ficode2.2"       "ficode8.1"       "simeduc.2"
"netlv_1.3"       "netlv_1.4"
[283] "netlv_1.5"       "netph_1.2"       "netph_1.3"       "scopngbhc.5"
"netwr_1.2"       "istrtdatss"
[289] "netkn_2.4"       "netwr_2.2"       "nch8resp.1"      "cgwri_dv.6"
"cgwri_dv.9"      "orga5.1"
[295] "sppno.2"         "mastat_dv.3"     "ficode38.1"      "big5a_dv.3"
"scopngbhf.4"     "cgs7cs_dv.2"
[301] "adoptch01.5"     "tvm2.5"          "nbrcoh2.4"       "newsmain.5"
"nch3resp.1"      "newsource8.1"
[307] "netpuse.6"       "ficode15.1"      "netjb_1.4"       "netjb_1.5"
"simjob.4"        "hrpno.3"
[313] "sppno.3"         "poleff1.2"       "intdatm_dv.5"    "scopngbhd.4"
"scsf1.4"         "scsf5.2"
[319] "locserc.4"       "locserc.3"       "adoptch03.5"     "adoptch02.4"
"ienddathh.15"    "scptrt5o3.5"
[325] "scptrt5o3.6"     "newsource1.1"    "big5o_dv.5"      "big5o_dv.6"
"scptrt5o2.5"     "scopngbhh.4"
[331] "colbens2.7"      "colbens2.5"      "ficode3.1"       "simrace.2"
"qfhigh_dv"       "fimnmisc_dv"
[337] "ficode27.1"      "poleff3.5"       "sclfsat1.2"      "hiqual_dv.4"
"hiqual_dv.2"     "ficode4.2"
[343] "voteintent.3"    "scptrt5n3.2"     "nmpsp_dv.2"      "intdaty_dv.2013" "colbens2.9"
"orga1.1"
[349] "sclfsat7.3"      "sclfsat7.6"      "rhland_code.1"   "demorient.4"
"scptrt5n2.2"     "gor_dv.9"
[355] "grpbfts.3"       "aidxhh.2"        "cgna_dv.4"       "cgwrd_dv.3"
"ficode17.1"      "scptrt5e2.2"
[361] "big5e_dv.2"      "colbens1.2"      "newsmain.4"      "simrace.4"
"racel_dv"        "gor_dv.10"
[367] "scptrt5e1.6"     "scptrt5o2.4"     "scptrt5o2.3"     "sclfsat7.2"
"sclfsat7.4"      "scptrt5e1.5"
[373] "sclfsat1.6"      "gor_dv.4"        "scptrt5c2.5"     "cgivwri1_dv.4"
"servacc.2"       "scopngbhb.5"
[379] "scptrt5c1.2"     "big5c_dv.3"      "scptrt5c2.6"     "orga4.1"
"orga16.1"        "jbstat.5"
[385] "hhtype_dv.20"    "netjb_2.4"       "scopngbha.3"     "ficode19.1"
"vote6.4"         "visfrnds.2"
[391] "hhtype_dv.23"    "nbrcoh1.4"       "netkn_1.2"       "scsf4b.4"
"gor_dv.5"        "ienddathh.18"
[397] "hcondn13.1"      "gor_dv.6"        "ficode38.2"      "sclfsat1.4"
"nrels2.2"        "voteintent.7"
[403] "polcost.3"       "ff_bentype37.1"  "polcost.5"       "colbens2.10"
"cgwri_dv.8"      "simfam.2"
[409] "ienddatmm"       "adoptch01.4"     "grpbfts.5"       "netpuse.5"
"hood15.6"        "nrels2.4"
[415] "cgna_dv.2"       "perpolinf.1"     "netkn_1.4"       "nbrcoh3.4"
"allch03.6"       "j2pay_if.1"
[421] "simarea.3"       "lvrel5.1"        "visfam.5"        "agegr10_dv.4"
```

```
"orga9.1"           "colbens1.7"
[427] "cgsrmem2_dv.1"    "big5o_dv.2"        "sub7_5"            "ficode7.1"
"closenum"          "perpolinf.6"
[433] "netlv_2.4"        "colbens1.1"        "intdatm_dv.8"      "big5e_dv.7"
"netjb_2.3"         "simeduc.4"
[439] "wrdrcl.4"         "cgwri_dv.5"        "big5e_dv.3"        "scptrt5e3.2"
"scptrt5e3.3"       "hcondn5.1"
[445] "scptrt5o3.3"      "hconde96.1"        "intdatd_dv"        "memorig.5"
"big5a_dv.4"        "netkn_2.2"
[451] "nbrcoh3.5"        "relup.4"           "scopngbhf.3"       "lkmove.2"
"allch01.6"         "cgvfw_dv.4"
[457] "hcondn4.1"        "ienddathh.13"      "lvrel8.1"          "ff_bentype39.1"
"scptrt5c2.2"       "scptrt5c3.6"
[463] "simage.3"         "crdark.5"          "poleff2.2"         "poleff2.4"
"big5o_dv.3"        "qfhighfl_dv.1"
[469] "orga6.1"          "hcondn14.1"        "natch03.6"         "ficode33.1"
"tvhours"           "indbd91_lw"
[475] "colbens2.6"       "nbrcoh4.5"         "cgwrd_dv.4"        "scopngbhe.3"
"ns200pre.3"        "ficode29.2"
[481] "perpolinf.7"      "perpolinf.2"       "perpolinf.5"       "scopngbhh.2"
"scopngbhf.5"       "locserb.3"
[487] "locsere.4"        "scsf5.5"           "agegr10_dv.8"      "nch5to15.4"
"cgivwri1_dv.2"     "newsource2.1"
[493] "colbens1.3"       "scptrt5o3.2"       "finfut.3"          "cgwrd_dv.1"
"scopngbhf.2"       "j2pay_dv"
[499] "j2has.2"          "wrdrcl.2"          "cgvfw_dv.3"        "mobuse.2"
"hcondn8.1"         "scopngbhg.4"
[505] "memaid.3"         "big5n_dv.3"        "undqus.3"          "ivcoop.2"
"undqus.2"          "poleff1.3"
[511] "ffbrfedlw.1"      "newsmain.8"        "ndepchl_dv.3"      "agegr10_dv.7"
"nbrcoh2.3"         "cindtime"
[517] "sclfsat2.3"       "sclfsat2.2"        "sclfsat2.7"        "sclfsat2.6"
"sclfsat2.5"        "perpolinf.8"
[523] "siminc.4"         "simage.4"          "simjob.3"          "newsource11.1"
"siminc.3"          "scptrt5e3.5"
[529] "intdatm_dv.3"     "hcondn1.1"         "ienddathh.12"      "ftedany.2"
"adcts.2"           "addrmov_dv.2"
[535] "scptrt5o2.6"      "ficode26.1"        "gor_dv.7"          "locserap.2"
"locserap.3"        "fiyrinvinc_dv"
[541] "ienddathh.14"     "ienddatss"         "agegr13_dv.4"      "agegr10_dv.3"
"memprob4.1"        "hhtype_dv.22"
[547] "hood15.3"         "scptrt5n2.6"       "hood15.2"          "scptrt5e2.3"
"newsource7.1"      "netet_2"
[553] "indscub_xw"       "ff_bentype27.1"    "crdark.4"          "scopngbhd.5"
"cgwrd_dv.8"        "locsere.3"
[559] "intdatm_dv.11"    "hrpno.4"           "seearngrs_if.1"    "fimngrs_if"
"hrpno.2"           "scptrt5c1.3"
[565] "voteintent.2"     "big5e_dv.6"        "scptrt5a1.6"       "ficode14.1"
"sub7_1.94"         "cgwrd_dv.9"
[571] "ff_bentype16.1"   "netjb_2.6"         "ff_jbstat.2"       "scopngbhc.3"
"scopngbhc.4"       "hood15.5"
[577] "scptrt5n3.6"      "simarea.5"         "nrels2.5"          "perbfts.3"
"marstat_dv.2"      "newsmain.2"
[583] "hhtype_dv.17"     "nasofa.2"          "big5a_dv.6"        "scptrt5a3.5"
"cgvfw_dv.1"        "ndepchl_dv.1"
[589] "fimnprben_dv"     "fimngrs_dv"        "ficode28.1"        "lvrel9.1"
"big5c_dv.4"        "big5e_dv.4"
[595] "scopngbhg.3"      "scptrt5c3.3"       "hcondn2.1"         "natch01.4"
"nadoptch.1"        "scptrt5a3.4"
[601] "scac.2"           "nbrcoh4.2"         "nbrcoh4.3"         "voteintent.10"
"npensioner_dv.3"   "cgwri_dv.2"
[607] "scptrt5c1.6"      "colbens1.8"        "scptrt5c2.4"       "aidhh.2"
"cgvfw_dv.2"
[1] "----------Ordinary Linear Regression (Improved)----------"

Call:
lm(formula = y ~ ., data = as.data.frame(x.data.linear))
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-13.0782  -1.8058  -0.2139   1.5586  16.4628

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.6670     0.4187  15.922  < 2e-16 ***
sampst.3         1.2076     0.5788   2.086 0.037037 *
allch02.6       -3.8911     1.3332  -2.919 0.003547 **
netpuse.3        0.5931     0.2822   2.102 0.035687 *
netpuse.7        1.0304     0.3971   2.595 0.009515 **
locseras.3      -0.3375     0.1487  -2.270 0.023263 *
nbrcoh2.5        2.0541     0.6919   2.969 0.003016 **
nbrcoh3.3       -0.4011     0.1558  -2.574 0.010115 *
crdark.2        -0.4464     0.1367  -3.265 0.001108 **
crdark.3        -0.4113     0.2069  -1.988 0.046895 *
simrace.3        0.7886     0.2935   2.687 0.007248 **
simarea.2        0.3555     0.1374   2.587 0.009744 **
netph_1.4       -0.6116     0.2307  -2.651 0.008088 **
netlv_2.3        0.3534     0.1376   2.568 0.010274 *
orga13.1         0.4488     0.1793   2.503 0.012390 *
orga15.1         1.2485     0.5559   2.246 0.024814 *
orga96.1         0.3748     0.1451   2.583 0.009849 **
hcondn6.1        2.4393     1.0850   2.248 0.024646 *
hcondn10.1       1.9126     0.8229   2.324 0.020198 *
hcondn17.1       2.4473     0.5881   4.161 3.27e-05 ***
hcondn96.1       0.6263     0.2290   2.735 0.006279 **
nrels2.10        3.3226     1.3368   2.485 0.013003 *
nrels2.12       -4.2024     1.8559  -2.264 0.023639 *
visfam.2         0.6335     0.2326   2.723 0.006512 **
visfam.3         0.5194     0.1871   2.776 0.005548 **
finnow.3         0.5438     0.1534   3.544 0.000401 ***
finnow.4         0.7265     0.2374   3.061 0.002231 **
finnow.5         1.0328     0.4187   2.467 0.013698 *
vote1.2          0.2689     0.1359   1.979 0.047912 *
colbens2.4      -0.4585     0.1984  -2.311 0.020906 *
civicduty.4     -0.5511     0.2304  -2.392 0.016836 *
voteintent.1     1.1126     0.4964   2.241 0.025090 *
voteintent.4     1.2135     0.5595   2.169 0.030182 *
poleff3.2       -0.7111     0.2325  -3.058 0.002253 **
poleff3.3       -0.7689     0.2420  -3.178 0.001503 **
poleff3.4       -1.0276     0.2504  -4.103 4.20e-05 ***
newsource4.1     0.3352     0.1350   2.483 0.013108 *
newsource9.1     0.5439     0.2116   2.571 0.010203 *
tvm2.2          -1.1071     0.4946  -2.238 0.025280 *
tvm2.3          -0.7868     0.2702  -2.912 0.003619 **
tvm2.4          -0.3986     0.1713  -2.327 0.020064 *
tvm2.8          -5.4369     2.5934  -2.096 0.036147 *
scsf1.5          1.3073     0.3941   3.317 0.000923 ***
scsf3b.2        -0.7151     0.3241  -2.206 0.027465 *
scsf4a.2         2.1966     0.4666   4.708 2.64e-06 ***
scsf4a.3         1.0819     0.2722   3.974 7.27e-05 ***
scsf4a.4         0.8279     0.1798   4.605 4.32e-06 ***
scsf4b.2         1.8962     0.5087   3.727 0.000198 ***
scsf4b.3         0.6247     0.2499   2.500 0.012500 *
scsf5.3         -0.4836     0.2269  -2.132 0.033135 *
scsf6a.3         0.7011     0.1608   4.360 1.35e-05 ***
scsf6a.4         1.9881     0.2365   8.407  < 2e-16 ***
scsf6a.5         6.2315     0.4474  13.928  < 2e-16 ***
scsf6b.3         0.3521     0.1559   2.259 0.023976 *
scsf6b.4         0.9753     0.2274   4.288 1.87e-05 ***
scsf6b.5         1.8063     0.3943   4.581 4.86e-06 ***
scsf6c.2         5.8652     0.4191  13.996  < 2e-16 ***
scsf6c.3         2.5439     0.2249  11.310  < 2e-16 ***
scsf6c.4         1.1825     0.1622   7.292 4.10e-13 ***
scsf7.2          3.5024     0.3951   8.864  < 2e-16 ***
scsf7.3          1.2432     0.2399   5.183 2.36e-07 ***
scsf7.4          0.9698     0.1918   5.057 4.57e-07 ***
```

```
scopngbha.4     0.9067     0.3327   2.725 0.006468 **
scopngbhb.4    -0.8946     0.2884  -3.102 0.001945 **
scopngbhe.5    -2.3010     1.1158  -2.062 0.039287 *
scopngbhg.5     1.5018     0.4939   3.041 0.002384 **
sclfsat1.7     -0.6683     0.2381  -2.807 0.005038 **
sclfsat2.4     -0.4641     0.2041  -2.274 0.023029 *
sclfsat7.7     -0.4042     0.2050  -1.971 0.048797 *
sclfsato.2      1.2663     0.2846   4.449 9.01e-06 ***
sclfsato.3      1.6747     0.2697   6.210 6.19e-10 ***
sclfsato.4      0.6474     0.2749   2.355 0.018581 *
sclfsato.5      0.4339     0.1857   2.337 0.019533 *
sclfsato.7     -0.5426     0.2361  -2.299 0.021611 *
scptrt5c1.5     0.4471     0.1828   2.446 0.014526 *
scptrt5n1.3     0.4592     0.2027   2.265 0.023574 *
scptrt5n1.4     0.6575     0.2006   3.278 0.001059 **
scptrt5n1.5     1.0485     0.2096   5.003 6.03e-07 ***
scptrt5n1.6     1.2100     0.2496   4.848 1.32e-06 ***
scptrt5n1.7     1.4954     0.2700   5.539 3.35e-08 ***
scptrt5o1.4    -0.4611     0.1740  -2.650 0.008109 **
scptrt5o1.5    -0.4976     0.1642  -3.031 0.002463 **
scptrt5o1.6    -0.4779     0.1942  -2.460 0.013957 *
scptrt5a2.3     0.5873     0.2453   2.394 0.016743 *
scptrt5a2.6     0.4056     0.1405   2.888 0.003912 **
scptrt5c2.7     1.4173     0.6672   2.124 0.033750 *
scptrt5o2.2    -0.5295     0.2124  -2.493 0.012724 *
scptrt5e3.6     0.5847     0.2044   2.861 0.004263 **
scptrt5e3.7     0.8576     0.2833   3.027 0.002495 **
scptrt5n3.3     0.6377     0.1995   3.197 0.001408 **
scptrt5n3.4     0.3801     0.1691   2.248 0.024679 *
scptrt5n3.7    -0.8120     0.2590  -3.135 0.001741 **
nacar.2         0.4752     0.1477   3.217 0.001312 **
readtest.2     -0.4519     0.1747  -2.586 0.009763 **
sevenspap.2     3.0476     1.4378   2.120 0.034135 *
ff_jbstat.6    -0.5067     0.2340  -2.166 0.030432 *
sppno.4         1.8267     0.8993   2.031 0.042332 *
big5c_dv.5     -0.3104     0.1476  -2.103 0.035602 *
big5n_dv.7      1.7424     0.4816   3.618 0.000303 ***
cgsrmem_dv.5    0.6218     0.2427   2.562 0.010454 *
cgwri_dv.3     -1.5997     0.5517  -2.899 0.003770 **
ficode10.1     -0.8782     0.3011  -2.916 0.003572 **
ficode37.1      2.0179     0.6860   2.942 0.003295 **
big5a_dv.2      2.8981     1.0797   2.684 0.007320 **
big5a_dv.7     -0.4032     0.1686  -2.391 0.016862 *
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1    1

Residual standard error: 3.132 on 2484 degrees of freedom
Multiple R-squared:  0.669,    Adjusted R-squared:  0.6551
F-statistic: 48.27 on 104 and 2484 DF,  p-value: < 2.2e-16

AIC: 13363.2143
MSE: 9.41
[1] "The MSE of the predicted values are of 11.1433"
[1] "The Linear Model predicts exactly with accuracy of 0.1518"
[1] "The Linear Model predicts within a confidence interval with accuracy of 0.4508"
[1] "-----------Elastic Net Regression-----------"
735 x 1 sparse Matrix of class "dgCMatrix", with 26 entries
         names Estimate_Coefs
1  (Intercept)   10.893652394
2     health.2   -0.095325666
3     finnow.4    0.107054147
4     finnow.5    0.517153882
5      scsf1.5    0.394014360
6     scsf4a.2    0.026392935
7     scsf6a.4    0.267169821
8     scsf6a.5    1.558632701
9     scsf6c.2    2.686415112
10    scsf6c.3    0.491407978
```

```
11      scsf7.2     1.388712329
12      scsf7.4     0.099420150
13   sclfsat1.7    -0.105193207
14   sclfsato.2     0.222075748
15   sclfsato.3     0.790901401
16   sclfsato.6    -0.576691657
17   sclfsato.7    -0.763174933
18 scptrt5n1.6     0.295475747
19 scptrt5n1.7     0.741945839
20 scptrt5n3.6    -0.006087259
21      nacar.2     0.040715591
22   sf12mcs_dv    -2.526571494
23   big5n_dv.2    -0.224583380
24   big5n_dv.7     0.461350069
25   sf12pcs_dv    -0.397836769
26   nbrsnci_dv    -0.047772278
[1] "The MSE of the predicted values of the best fit model is 8.882"
[1] "The Alpha of the best fit model is 0.3"
[1] "The Elastic Net Model predicts exactly with accuracy of 0.1587"
[1] "-----------Timer Results-----------"
   user  system elapsed
1367.63   12.52 1380.93
```

10.2.22 w3MergeNurse console

```
[1] "-----------Initial Checks-----------"
[1] "8492259 NA cells were found across the entire dataset (71.41% of data as NA)"
[1] "-----------Data Type Checks-----------"
[1] "0 variables recoded since all their entries aren't numeric or NA"
[1] "NOTE: algorithim recodes categorical data alphabetically e.g (female = 1, male = 2)"
character(0)
[1] "-----------Low Data Removal-----------"
[1] "2867 variables removed since they had >= 'naPercent' (default 20%) NA values"
   [1] "pid.x"           "childpno"        "cindtime"        "lvwhy"
"lvmthp"          "lvyrp"
   [7] "liwho"           "lihow"           "mvever"          "mvmnth"
"mvyr"            "mlstatchk"
  [13] "mlstat.x"        "drive"           "caruse"          "britid"
"ukborn"          "plbornc"
  [19] "yr2uk4"          "citzn1"          "citzn2"          "citzn3"
"qfhigh"          "qualoc"
  [25] "qfvoc1"          "qfvoc2"          "qfvoc3"          "qfvoc4"
"qfvoc5"          "qfvoc6"
  [31] "qfvoc7"          "qfvoc8"          "qfvoc9"          "qfvoc10"
"qfvoc11"         "qfvoc12"
  [37] "qfvoc13"         "qfvoc14"         "qfvoc15"         "qfvoc96"
"school"          "scend"
  [43] "schlloc"         "schok"           "fenow"           "feend"
"j1none"          "j1semp"
  [49] "j1boss"          "j1mngr"          "edtype"          "edasp"
"lvscdo"          "ahvwell"
  [55] "fedlik"          "fednt"           "futureint"       "lvhm"
"ocimpa"          "ocimpb"
  [61] "ocimpe"          "ocimpf"          "ocimpi"          "ocimpk"
"ocimpl"          "futra"
  [67] "futrb"           "futrc"           "futrd"           "futre"
"futrf"           "futrg"
  [73] "futrh"           "futri"           "futrj"           "futrk"
"futrl"           "paju"
  [79] "maju"            "pacob"           "payruk"          "macob"
"mayruk"          "natid1"
  [85] "natid2"          "natid3"          "natid4"          "natid5"
"natid6"          "natid97"
  [91] "racel"           "racelo_code"     "racelt"          "racelwt"
"racelmt"         "racelat"
```

```
 [97] "racelbt"          "racelot_code"     "oprlg"            "oprlg0ni"
"nirel"           "niact"
 [103] "oprlg0"          "oprlg1"           "unsafe1"          "unsafe2"
"unsafe3"         "unsafe4"
 [109] "unsafe5"         "unsafe6"          "unsafe7"          "unsafe8"
"unsafe9"         "unsafe10"
 [115] "unsafe11"        "unsafe96"         "unsafe97"         "unsafeo_code"
"resunsafe1_1"    "resunsafe2_1"
 [121] "resunsafe3_1"    "resunsafe4_1"     "resunsafe5_1"     "resunsafe6_1"
"resunsafe7_1"    "resunsafe8_1"
 [127] "resunsafe9_1"    "resunsafe96_1"    "resunsafe97_1"    "resunsafe1_2"
"resunsafe2_2"    "resunsafe3_2"
 [133] "resunsafe4_2"    "resunsafe5_2"     "resunsafe6_2"     "resunsafe7_2"
"resunsafe8_2"    "resunsafe9_2"
 [139] "resunsafe96_2"   "resunsafe97_2"    "resunsafe1_3"     "resunsafe2_3"
"resunsafe3_3"    "resunsafe4_3"
 [145] "resunsafe5_3"    "resunsafe6_3"     "resunsafe7_3"     "resunsafe8_3"
"resunsafe9_3"    "resunsafe96_3"
 [151] "resunsafe97_3"   "resunsafe1_4"     "resunsafe2_4"     "resunsafe3_4"
"resunsafe4_4"    "resunsafe5_4"
 [157] "resunsafe6_4"    "resunsafe7_4"     "resunsafe8_4"     "resunsafe9_4"
"resunsafe96_4"   "resunsafe97_4"
 [163] "resunsafe1_5"    "resunsafe2_5"     "resunsafe3_5"     "resunsafe4_5"
"resunsafe5_5"    "resunsafe6_5"
 [169] "resunsafe7_5"    "resunsafe8_5"     "resunsafe9_5"     "resunsafe96_5"
"resunsafe97_5"   "resunsafe1_6"
 [175] "resunsafe2_6"    "resunsafe3_6"     "resunsafe4_6"     "resunsafe5_6"
"resunsafe6_6"    "resunsafe7_6"
 [181] "resunsafe8_6"    "resunsafe9_6"     "resunsafe96_6"    "resunsafe97_6"
"resunsafe1_7"    "resunsafe2_7"
 [187] "resunsafe3_7"    "resunsafe4_7"     "resunsafe5_7"     "resunsafe6_7"
"resunsafe7_7"    "resunsafe8_7"
 [193] "resunsafe9_7"    "resunsafe96_7"    "resunsafe97_7"    "resunsafe1_8"
"resunsafe2_8"    "resunsafe3_8"
 [199] "resunsafe4_8"    "resunsafe5_8"     "resunsafe6_8"     "resunsafe7_8"
"resunsafe8_8"    "resunsafe9_8"
 [205] "resunsafe96_8"   "resunsafe97_8"    "resunsafe1_9"     "resunsafe2_9"
"resunsafe3_9"    "resunsafe4_9"
 [211] "resunsafe5_9"    "resunsafe6_9"     "resunsafe7_9"     "resunsafe8_9"
"resunsafe9_9"    "resunsafe96_9"
 [217] "resunsafe97_9"   "resunsafe1_10"    "resunsafe2_10"    "resunsafe3_10"
"resunsafe4_10"   "resunsafe5_10"
 [223] "resunsafe6_10"   "resunsafe7_10"    "resunsafe8_10"    "resunsafe9_10"
"resunsafe96_10"  "resunsafe97_10"
 [229] "resunsafe1_11"   "resunsafe2_11"    "resunsafe3_11"    "resunsafe4_11"
"resunsafe5_11"   "resunsafe6_11"
 [235] "resunsafe7_11"   "resunsafe8_11"    "resunsafe9_11"    "resunsafe96_11"
"resunsafe97_11"  "resunsafe1_12"
 [241] "resunsafe2_12"   "resunsafe3_12"    "resunsafe4_12"    "resunsafe5_12"
"resunsafe6_12"   "resunsafe7_12"
 [247] "resunsafe8_12"   "resunsafe9_12"    "resunsafe96_12"   "resunsafe97_12"
"avoidance1"      "avoidance2"
 [253] "avoidance3"      "avoidance4"       "avoidance5"       "avoidance6"
"avoidance7"      "avoidance8"
 [259] "avoidance9"      "avoidance10"      "avoidance11"      "avoidance96"
"avoidance97"     "avoidanceo_code"
 [265] "resavoid1_1"     "resavoid2_1"      "resavoid3_1"      "resavoid4_1"
"resavoid5_1"     "resavoid6_1"
 [271] "resavoid7_1"     "resavoid8_1"      "resavoid9_1"      "resavoid96_1"
"resavoid97_1"    "resavoid1_2"
 [277] "resavoid2_2"     "resavoid3_2"      "resavoid4_2"      "resavoid5_2"
"resavoid6_2"     "resavoid7_2"
 [283] "resavoid8_2"     "resavoid9_2"      "resavoid96_2"     "resavoid97_2"
"resavoid1_3"     "resavoid2_3"
 [289] "resavoid3_3"     "resavoid4_3"      "resavoid5_3"      "resavoid6_3"
"resavoid7_3"     "resavoid8_3"
 [295] "resavoid9_3"     "resavoid96_3"     "resavoid97_3"     "resavoid1_4"
"resavoid2_4"     "resavoid3_4"
```

```
 [301] "resavoid4_4"      "resavoid5_4"      "resavoid6_4"      "resavoid7_4"
"resavoid8_4"      "resavoid9_4"
 [307] "resavoid96_4"     "resavoid97_4"     "resavoid1_5"      "resavoid2_5"
"resavoid3_5"      "resavoid4_5"
 [313] "resavoid5_5"      "resavoid6_5"      "resavoid7_5"      "resavoid8_5"
"resavoid9_5"      "resavoid96_5"
 [319] "resavoid97_5"     "resavoid1_6"      "resavoid2_6"      "resavoid3_6"
"resavoid4_6"      "resavoid5_6"
 [325] "resavoid6_6"      "resavoid7_6"      "resavoid8_6"      "resavoid9_6"
"resavoid96_6"     "resavoid97_6"
 [331] "resavoid1_7"      "resavoid2_7"      "resavoid3_7"      "resavoid4_7"
"resavoid5_7"      "resavoid6_7"
 [337] "resavoid7_7"      "resavoid8_7"      "resavoid9_7"      "resavoid96_7"
"resavoid97_7"     "resavoid1_8"
 [343] "resavoid2_8"      "resavoid3_8"      "resavoid4_8"      "resavoid5_8"
"resavoid6_8"      "resavoid7_8"
 [349] "resavoid8_8"      "resavoid9_8"      "resavoid96_8"     "resavoid97_8"
"resavoid1_9"      "resavoid2_9"
 [355] "resavoid3_9"      "resavoid4_9"      "resavoid5_9"      "resavoid6_9"
"resavoid7_9"      "resavoid8_9"
 [361] "resavoid9_9"      "resavoid96_9"     "resavoid97_9"     "resavoid1_10"
"resavoid2_10"     "resavoid3_10"
 [367] "resavoid4_10"     "resavoid5_10"     "resavoid6_10"     "resavoid7_10"
"resavoid8_10"     "resavoid9_10"
 [373] "resavoid96_10"    "resavoid97_10"    "resavoid1_11"     "resavoid2_11"
"resavoid3_11"     "resavoid4_11"
 [379] "resavoid5_11"     "resavoid6_11"     "resavoid7_11"     "resavoid8_11"
"resavoid9_11"     "resavoid96_11"
 [385] "resavoid97_11"    "resavoid1_12"     "resavoid2_12"     "resavoid3_12"
"resavoid4_12"     "resavoid5_12"
 [391] "resavoid6_12"     "resavoid7_12"     "resavoid8_12"     "resavoid9_12"
"resavoid96_12"    "resavoid97_12"
 [397] "insulted1"        "insulted2"        "insulted3"        "insulted4"
"insulted5"        "insulted6"
 [403] "insulted7"        "insulted8"        "insulted9"        "insulted10"
"insulted11"       "insulted96"
 [409] "insulted97"       "insultedo_code"   "resinsulted1_1"   "resinsulted2_1"
"resinsulted3_1"   "resinsulted4_1"
 [415] "resinsulted5_1"   "resinsulted6_1"   "resinsulted7_1"   "resinsulted8_1"
"resinsulted9_1"   "resinsulted96_1"
 [421] "resinsulted97_1"  "resinsulted1_2"   "resinsulted2_2"   "resinsulted3_2"
"resinsulted4_2"   "resinsulted5_2"
 [427] "resinsulted6_2"   "resinsulted7_2"   "resinsulted8_2"   "resinsulted9_2"
"resinsulted96_2"  "resinsulted97_2"
 [433] "resinsulted1_3"   "resinsulted2_3"   "resinsulted3_3"   "resinsulted4_3"
"resinsulted5_3"   "resinsulted6_3"
 [439] "resinsulted7_3"   "resinsulted8_3"   "resinsulted9_3"   "resinsulted96_3"
"resinsulted97_3"  "resinsulted1_4"
 [445] "resinsulted2_4"   "resinsulted3_4"   "resinsulted4_4"   "resinsulted5_4"
"resinsulted6_4"   "resinsulted7_4"
 [451] "resinsulted8_4"   "resinsulted9_4"   "resinsulted96_4"  "resinsulted97_4"
"resinsulted1_5"   "resinsulted2_5"
 [457] "resinsulted3_5"   "resinsulted4_5"   "resinsulted5_5"   "resinsulted6_5"
"resinsulted7_5"   "resinsulted8_5"
 [463] "resinsulted9_5"   "resinsulted96_5"  "resinsulted97_5"  "resinsulted1_6"
"resinsulted2_6"   "resinsulted3_6"
 [469] "resinsulted4_6"   "resinsulted5_6"   "resinsulted6_6"   "resinsulted7_6"
"resinsulted8_6"   "resinsulted9_6"
 [475] "resinsulted96_6"  "resinsulted97_6"  "resinsulted1_7"   "resinsulted2_7"
"resinsulted3_7"   "resinsulted4_7"
 [481] "resinsulted5_7"   "resinsulted6_7"   "resinsulted7_7"   "resinsulted8_7"
"resinsulted9_7"   "resinsulted96_7"
 [487] "resinsulted97_7"  "resinsulted1_8"   "resinsulted2_8"   "resinsulted3_8"
"resinsulted4_8"   "resinsulted5_8"
 [493] "resinsulted6_8"   "resinsulted7_8"   "resinsulted8_8"   "resinsulted9_8"
"resinsulted96_8"  "resinsulted97_8"
 [499] "resinsulted1_9"   "resinsulted2_9"   "resinsulted3_9"   "resinsulted4_9"
"resinsulted5_9"   "resinsulted6_9"
```

```
 [505] "resinsulted7_9"   "resinsulted8_9"   "resinsulted9_9"   "resinsulted96_9"
"resinsulted97_9"  "resinsulted1_10"
 [511] "resinsulted2_10"  "resinsulted3_10"  "resinsulted4_10"  "resinsulted5_10"
"resinsulted6_10"  "resinsulted7_10"
 [517] "resinsulted8_10"  "resinsulted9_10"  "resinsulted96_10" "resinsulted97_10" "resinsulted1_11"
"resinsulted2_11"
 [523] "resinsulted3_11"  "resinsulted4_11"  "resinsulted5_11"  "resinsulted6_11"
"resinsulted7_11"  "resinsulted8_11"
 [529] "resinsulted9_11"  "resinsulted96_11" "resinsulted97_11" "resinsulted1_12"
"resinsulted2_12"  "resinsulted3_12"
 [535] "resinsulted4_12"  "resinsulted5_12"  "resinsulted6_12"  "resinsulted7_12"
"resinsulted8_12"  "resinsulted9_12"
 [541] "resinsulted96_12" "resinsulted97_12" "attacked1"        "attacked2"
"attacked3"        "attacked4"
 [547] "attacked5"        "attacked6"        "attacked7"        "attacked8"
"attacked9"        "attacked10"
 [553] "attacked11"       "attacked96"       "attacked97"       "attackedo_code"
"resattacked1_1"   "resattacked2_1"
 [559] "resattacked3_1"   "resattacked4_1"   "resattacked5_1"   "resattacked6_1"
"resattacked7_1"   "resattacked8_1"
 [565] "resattacked9_1"   "resattacked96_1"  "resattacked97_1"  "resattacked1_2"
"resattacked2_2"   "resattacked3_2"
 [571] "resattacked4_2"   "resattacked5_2"   "resattacked6_2"   "resattacked7_2"
"resattacked8_2"   "resattacked9_2"
 [577] "resattacked96_2"  "resattacked97_2"  "resattacked1_3"   "resattacked2_3"
"resattacked3_3"   "resattacked4_3"
 [583] "resattacked5_3"   "resattacked6_3"   "resattacked7_3"   "resattacked8_3"
"resattacked9_3"   "resattacked96_3"
 [589] "resattacked97_3"  "resattacked1_4"   "resattacked2_4"   "resattacked3_4"
"resattacked4_4"   "resattacked5_4"
 [595] "resattacked6_4"   "resattacked7_4"   "resattacked8_4"   "resattacked9_4"
"resattacked96_4"  "resattacked97_4"
 [601] "resattacked1_5"   "resattacked2_5"   "resattacked3_5"   "resattacked4_5"
"resattacked5_5"   "resattacked6_5"
 [607] "resattacked7_5"   "resattacked8_5"   "resattacked9_5"   "resattacked96_5"
"resattacked97_5"  "resattacked1_6"
 [613] "resattacked2_6"   "resattacked3_6"   "resattacked4_6"   "resattacked5_6"
"resattacked6_6"   "resattacked7_6"
 [619] "resattacked8_6"   "resattacked9_6"   "resattacked96_6"  "resattacked97_6"
"resattacked1_7"   "resattacked2_7"
 [625] "resattacked3_7"   "resattacked4_7"   "resattacked5_7"   "resattacked6_7"
"resattacked7_7"   "resattacked8_7"
 [631] "resattacked9_7"   "resattacked96_7"  "resattacked97_7"  "resattacked1_8"
"resattacked2_8"   "resattacked3_8"
 [637] "resattacked4_8"   "resattacked5_8"   "resattacked6_8"   "resattacked7_8"
"resattacked8_8"   "resattacked9_8"
 [643] "resattacked96_8"  "resattacked97_8"  "resattacked1_9"   "resattacked2_9"
"resattacked3_9"   "resattacked4_9"
 [649] "resattacked5_9"   "resattacked6_9"   "resattacked7_9"   "resattacked8_9"
"resattacked9_9"   "resattacked96_9"
 [655] "resattacked97_9"  "resattacked1_10"  "resattacked2_10"  "resattacked3_10"
"resattacked4_10"  "resattacked5_10"
 [661] "resattacked6_10"  "resattacked7_10"  "resattacked8_10"  "resattacked9_10"
"resattacked96_10" "resattacked97_10"
 [667] "resattacked1_11"  "resattacked2_11"  "resattacked3_11"  "resattacked4_11"
"resattacked5_11"  "resattacked6_11"
 [673] "resattacked7_11"  "resattacked8_11"  "resattacked9_11"  "resattacked96_11" "resattacked97_11"
 [679] "resattacked2_12"  "resattacked3_12"  "resattacked4_12"  "resattacked5_12"
"resattacked6_12"  "resattacked7_12"
 [685] "resattacked8_12"  "resattacked9_12"  "resattacked96_12" "resattacked97_12" "locserap"
"locseras"
 [691] "locserb"          "locsere"          "servaccy1"        "servaccy2"
"servaccy3"        "servaccy4"
 [697] "servaccy5"        "servaccy6"        "servaccy7"        "servaccy8"
"servaccy9"        "servaccy10"
 [703] "servaccy11"       "servaccy12"       "servaccy13"       "servaccy14"
"servaccy15"       "servaccy97"
 [709] "locserc"          "sattransy1"       "sattransy2"       "sattransy3"
```

```
"sattransy4"        "sattransy5"
 [715] "sattransy6"        "sattransy7"        "sattransy8"        "sattransy9"
"sattransy10"       "sattransy97"
 [721] "crworb"            "simrace"           "simeduc"           "simjob"
"siminc"            "simarea"
 [727] "netcht"            "visfrndsy1"        "visfrndsy2"        "visfrndsy3"
"visfrndsy4"        "visfrndsy5"
 [733] "visfrndsy6"        "visfrndsy7"        "visfrndsy8"        "visfrndsy9"
"visfrndsy10"       "visfrndsy11"
 [739] "visfrndsy12"       "visfrndsy13"       "visfrndsy14"       "visfrndsy97"
"netsx_1"           "netwr_1"
 [745] "netrl_1"           "netag_1"           "netkn_1"           "netph_1"
"netlv_1"           "netjb_1"
 [751] "netet_1"           "netmet_1"          "netweb_1"          "nettalk1_1"
"nettalk2_1"        "nettalk3_1"
 [757] "nettalk4_1"        "nettalk5_1"        "nettalk6_1"        "nettalk7_1"
"nettalk8_1"        "nettalk9_1"
 [763] "nettalk10_1"       "nettalk11_1"       "nettalk12_1"       "nettalk13_1"
"nettalk97_1"       "nettalk96_1"
 [769] "netdo1_1"          "netdo2_1"          "netdo3_1"          "netdo4_1"
"netdo5_1"          "netdo6_1"
 [775] "netdo7_1"          "netdo8_1"          "netdo9_1"          "netdo10_1"
"netdo97_1"         "netdo96_1"
 [781] "netsx_2"           "netwr_2"           "netrl_2"           "netag_2"
"netkn_2"           "netph_2"
 [787] "netlv_2"           "netjb_2"           "netet_2"           "netmet_2"
"netweb_2"          "nettalk1_2"
 [793] "nettalk2_2"        "nettalk3_2"        "nettalk4_2"        "nettalk5_2"
"nettalk6_2"        "nettalk7_2"
 [799] "nettalk8_2"        "nettalk9_2"        "nettalk10_2"       "nettalk11_2"
"nettalk12_2"       "nettalk13_2"
 [805] "nettalk97_2"       "nettalk96_2"       "netdo1_2"          "netdo2_2"
"netdo3_2"          "netdo4_2"
 [811] "netdo5_2"          "netdo6_2"          "netdo7_2"          "netdo8_2"
"netdo9_2"          "netdo10_2"
 [817] "netdo97_2"         "netdo96_2"         "netsx_3"           "netwr_3"
"netrl_3"           "netag_3"
 [823] "netkn_3"           "netph_3"           "netlv_3"           "netjb_3"
"netet_3"           "netmet_3"
 [829] "netweb_3"          "nettalk1_3"        "nettalk2_3"        "nettalk3_3"
"nettalk4_3"        "nettalk5_3"
 [835] "nettalk6_3"        "nettalk7_3"        "nettalk8_3"        "nettalk9_3"
"nettalk10_3"       "nettalk11_3"
 [841] "nettalk12_3"       "nettalk13_3"       "nettalk97_3"       "nettalk96_3"
"netdo1_3"          "netdo2_3"
 [847] "netdo3_3"          "netdo4_3"          "netdo5_3"          "netdo6_3"
"netdo7_3"          "netdo8_3"
 [853] "netdo9_3"          "netdo10_3"         "netdo97_3"         "netdo96_3"
"orgm1"             "orgm2"
 [859] "orgm3"             "orgm4"             "orgm5"             "orgm6"
"orgm7"             "orgm8"
 [865] "orgm9"             "orgm10"            "orgm11"            "orgm12"
"orgm13"            "orgm14"
 [871] "orgm15"            "orgm16"            "orgm96"            "orgmt1"
"orgmt2"            "orgmt3"
 [877] "orgmt4"            "orgmt5"            "orgmt6"            "orgmt7"
"orgmt8"            "orgmt9"
 [883] "orgmt10"           "orgmt11"           "orgmt12"           "orgmt13"
"orgmt14"           "orgmt15"
 [889] "orgmt16"           "orgmt96"           "orgat1"            "orgat2"
"orgat3"            "orgat4"
 [895] "orgat5"            "orgat6"            "orgat7"            "orgat8"
"orgat9"            "orgat10"
 [901] "orgat11"           "orgat12"           "orgat13"           "orgat14"
"orgat15"           "orgat16"
 [907] "orgat96"           "hospc1"            "hospdc1"           "hospc2"
"hospdc2"           "hospc3"
 [913] "hospdc3"           "hospc4"            "hospdc4"           "hospc5"
```

```
"hospdc5"          "hospc6"
 [919] "hospdc6"          "hospc7"           "hospdc7"          "hospc8"
"hospdc8"          "disdif1"
 [925] "disdif2"          "disdif3"          "disdif4"          "disdif5"
"disdif6"          "disdif7"
 [931] "disdif8"          "disdif9"          "disdif10"         "disdif11"
"disdif12"         "disdif96"
 [937] "hcond1"           "hcond2"           "hcond3"           "hcond4"
"hcond5"           "hcond6"
 [943] "hcond7"           "hcond8"           "hcond9"           "hcond10"
"hcond11"          "hcond12"
 [949] "hcond13"          "hcond14"          "hcond15"          "hcond16"
"hcond17"          "hcond96"
 [955] "hconds01"         "hconds02"         "hconds03"         "hconds04"
"hconds05"         "hconds08"
 [961] "hconds09"         "hconds10"         "hconds11"         "hconds12"
"hconds13"         "hconds14"
 [967] "hconds15"         "hconds16"         "hconds17"         "hconda01"
"hconda02"         "hconda03"
 [973] "hconda04"         "hconda05"         "hconda06"         "hconda07"
"hconda08"         "hconda09"
 [979] "hconda10"         "hconda11"         "hconda12"         "hconda13"
"hconda14"         "hconda15"
 [985] "hconda16"         "hconda17"         "aidhh"            "aidhua1"
"aidhua2"          "aidhua3"
 [991] "aidhua4"          "aidhua5"          "aidhua6"          "aidhua7"
"aidhua8"          "aidhua9"
 [997] "aidhua10"         "aidhua11"         "aidhua12"         "aidhua13"
 [ reached getOption("max.print") -- omitted 1867 entries ]
[1] "----------Low Level Removal----------"
[1] "If a level is removed from a variable you wish to keep, reccomended to manually merge levels to
[1] "level 6 in pno removed, 1 observations found"
[1] "level 7 in pno removed, 1 observations found"
[1] "level 9 in pno removed, 1 observations found"
[1] "level 10 in pno removed, 1 observations found"
[1] "level 5 in nch14resp removed, 1 observations found"
[1] "level 2 in nch3resp removed, 1 observations found"
[1] "level 2 in nch5resp removed, 1 observations found"
[1] "level 2 in nch8resp removed, 1 observations found"
[1] "level 4 in nch415resp removed, 3 observations found"
[1] "level 5 in nch415resp removed, 1 observations found"
[1] "level 5 in nchresp removed, 0 observations found"
[1] "level 5 in nchund18resp removed, 0 observations found"
[1] "level 6 in natch01 removed, 2 observations found"
[1] "level 2 in natch02 removed, 1 observations found"
[1] "level 8 in natch02 removed, 0 observations found"
[1] "level 7 in natch03 removed, 2 observations found"
[1] "level 7 in natch04 removed, 2 observations found"
[1] "level 6 in natch05 removed, 1 observations found"
[1] "level 8 in natch06 removed, 2 observations found"
[1] "level 9 in natch07 removed, 0 observations found"
[1] "level 5 in nnatch removed, 3 observations found"
[1] "level 6 in nnatch removed, 0 observations found"
[1] "level 7 in nnatch removed, 0 observations found"
[1] "level 5 in adoptch01 removed, 3 observations found"
[1] "level 6 in adoptch01 removed, 1 observations found"
[1] "level 5 in adoptch02 removed, 1 observations found"
[1] "level 6 in adoptch02 removed, 0 observations found"
[1] "level 5 in adoptch03 removed, 2 observations found"
[1] "level 7 in adoptch03 removed, 0 observations found"
[1] "level 5 in nchunder16 removed, 0 observations found"
[1] "level 4 in nch5to15 removed, 1 observations found"
[1] "level 2 in nch10 removed, 0 observations found"
[1] "level 7 in allch02 removed, 0 observations found"
[1] "level 8 in allch02 removed, 0 observations found"
[1] "level 8 in allch03 removed, 0 observations found"
[1] "level 5 in allch04 removed, 3 observations found"
[1] "level 7 in allch04 removed, 0 observations found"
```

```
[1] "level 9 in allch04 removed, 0 observations found"
[1] "level 7 in allch05 removed, 0 observations found"
[1] "level 10 in jbstat.x removed, 1 observations found"
[1] "level 1 in bensta3 removed, 1 observations found"
[1] "level 3 in marstat.x removed, 2 observations found"
[1] "level 5 in vftctw removed, 2 observations found"
[1] "level 6 in vftctw removed, 2 observations found"
[1] "level 7 in vftctw removed, 2 observations found"
[1] "level 9 in vftctw removed, 1 observations found"
[1] "level 14 in vftctw removed, 3 observations found"
[1] "level 16 in vftctw removed, 2 observations found"
[1] "level 17 in vftctw removed, 1 observations found"
[1] "level 18 in vftctw removed, 3 observations found"
[1] "level 19 in vftctw removed, 1 observations found"
[1] "level 5 in clangab removed, 2 observations found"
[1] "level 4 in ivcoop removed, 1 observations found"
[1] "level 5 in undqus removed, 0 observations found"
[1] "level 4 in hgbiom removed, 1 observations found"
[1] "level 5 in hgbiom removed, 1 observations found"
[1] "level 4 in hgbiof removed, 1 observations found"
[1] "level 7 in hgbiof removed, 1 observations found"
[1] "level 1 in fiyrinvinc_tc removed, 2 observations found"
[1] "level 1 in fibenothr_tc removed, 2 observations found"
[1] "level 4 in pn1pno removed, 0 observations found"
[1] "level 5 in pn1pno removed, 0 observations found"
[1] "level 6 in pn1pno removed, 0 observations found"
[1] "level 4 in pn2pno removed, 1 observations found"
[1] "level 7 in pn2pno removed, 0 observations found"
[1] "level 4 in pns1pno removed, 0 observations found"
[1] "level 5 in pns1pno removed, 0 observations found"
[1] "level 6 in pns1pno removed, 0 observations found"
[1] "level 4 in pns2pno removed, 1 observations found"
[1] "level 7 in pns2pno removed, 0 observations found"
[1] "level 1 in ngrp_dv removed, 3 observations found"
[1] "level 4 in nnssib_dv removed, 1 observations found"
[1] "level 5 in nnssib_dv removed, 1 observations found"
[1] "level 6 in nnssib_dv removed, 3 observations found"
[1] "level 7 in nnssib_dv removed, 0 observations found"
[1] "level 3 in mastat_dv removed, 0 observations found"
[1] "level 5 in buno_dv removed, 3 observations found"
[1] "level 6 in buno_dv removed, 0 observations found"
[1] "level 10 in buno_dv removed, 0 observations found"
[1] "level 5 in nchild_dv removed, 0 observations found"
[1] "level 6 in hrpno removed, 0 observations found"
[1] "level 10 in hrpno removed, 1 observations found"
[1] "level 6 in ppno removed, 1 observations found"
[1] "level 7 in ppno removed, 1 observations found"
[1] "level 5 in sppno removed, 3 observations found"
[1] "level 6 in sppno removed, 0 observations found"
[1] "level 7 in sppno removed, 0 observations found"
[1] "level 4 in fnpno removed, 0 observations found"
[1] "level 6 in fnpno removed, 0 observations found"
[1] "level 4 in fnspno removed, 0 observations found"
[1] "level 6 in fnspno removed, 0 observations found"
[1] "level 4 in mnpno removed, 0 observations found"
[1] "level 5 in mnpno removed, 0 observations found"
[1] "level 4 in mnspno removed, 0 observations found"
[1] "level 5 in mnspno removed, 0 observations found"
[1] "level 1 in grfpno removed, 3 observations found"
[1] "level 4 in grfpno removed, 1 observations found"
[1] "level 1 in grmpno removed, 0 observations found"
[1] "level 2 in grmpno removed, 0 observations found"
[1] "level 3 in grmpno removed, 0 observations found"
[1] "level 4 in grmpno removed, 0 observations found"
[1] "level 5 in cgvfw_dv removed, 0 observations found"
[1] "level 6 in cgvfw_dv removed, 0 observations found"
[1] "level 7 in cgvfw_dv removed, 0 observations found"
[1] "level 9 in cgvfw_dv removed, 0 observations found"
```

```
[1] "level 14 in cgvfw_dv removed, 0 observations found"
[1] "level 16 in cgvfw_dv removed, 0 observations found"
[1] "level 17 in cgvfw_dv removed, 0 observations found"
[1] "level 18 in cgvfw_dv removed, 0 observations found"
[1] "level 19 in cgvfw_dv removed, 0 observations found"
[1] "level 0.0299999993294477 in fimnlabgrs_if removed, 2 observations found"
[1] "level 0.0399999991059303 in fimnlabgrs_if removed, 1 observations found"
[1] "level 0.140000000596046 in fimnlabgrs_if removed, 1 observations found"
[1] "level 0.620000004768372 in fimnlabgrs_if removed, 1 observations found"
[1] "level 0.689999997615814 in fimnlabgrs_if removed, 1 observations found"
[1] "level 0.819999992847443 in fimnlabgrs_if removed, 1 observations found"
[1] "level 0.850000023841858 in fimnlabgrs_if removed, 1 observations found"
[1] "level 0.939999997615814 in fimnlabgrs_if removed, 1 observations found"
[1] "level 3 in ficode3 removed, 3 observations found"
[1] "level 3 in ficode4 removed, 3 observations found"
[1] "level 1 in ficode21 removed, 1 observations found"
[1] "level 2 in ficode24 removed, 3 observations found"
[1] "level 1 in ficode25 removed, 1 observations found"
[1] "level 2 in ficode27 removed, 2 observations found"
[1] "level 2 in ficode28 removed, 1 observations found"
[1] "level 3 in ficode29 removed, 3 observations found"
[1] "level 1 in ficode30 removed, 0 observations found"
[1] "level 1 in ficode34 removed, 3 observations found"
[1] "level 2 in ficode38 removed, 3 observations found"
[1] "level 6 in c_pno removed, 0 observations found"
[1] "level 7 in c_pno removed, 0 observations found"
[1] "level 9 in c_pno removed, 0 observations found"
[1] "level 10 in c_pno removed, 0 observations found"
[1] "level 1 in c_splitnum removed, 2 observations found"
[1] "level 2 in c_splitnum removed, 1 observations found"
[1] "level 3 in medcnjd removed, 1 observations found"
[1] "level 1 in medtyp13 removed, 1 observations found"
[1] "level 2 in resphts removed, 2 observations found"
[1] "level 5 in respbps removed, 2 observations found"
[1] "level 6 in nseqno removed, 0 observations found"
[1] "level 7 in nseqno removed, 0 observations found"
[1] "level 9 in nseqno removed, 0 observations found"
[1] "level 10 in nseqno removed, 0 observations found"
[1] "level 2 in dateok removed, 2 observations found"
[1] "level 3 in htok removed, 0 observations found"
[1] "level 2 in bmiok removed, 2 observations found"
[1] "level 3 in bmiok removed, 0 observations found"
[1] "level 4 in bmiok removed, 1 observations found"
[1] "level 96 in ag16g10 removed, 2 observations found"
[1] "level 96 in ag16g20 removed, 0 observations found"
[1] "level 5 in wstokb removed, 1 observations found"
[1] "level 8 in hhsize removed, 1 observations found"
[1] "level 9 in hhsize removed, 0 observations found"
[1] "level 10 in hhsize removed, 0 observations found"
[1] "level 14 in hhsize removed, 0 observations found"
[1] "level 10 in jbstat.y removed, 0 observations found"
[1] "level 3 in marstat.y removed, 0 observations found"
[1] "level 3 in npensioner_dv removed, 3 observations found"
[1] "level 4 in nnsib_dv removed, 0 observations found"
[1] "level 5 in nnsib_dv removed, 0 observations found"
[1] "level 6 in nnsib_dv removed, 0 observations found"
[1] "level 7 in nnsib_dv removed, 0 observations found"
[1] "level 1 in big5a_dv removed, 1 observations found"
[1] "163 total levels removed from 87 different variables. In total 145 observations deleted"
[1] "————Variance 0 Check————"
[1] "100 variables removed since their new variance was 0"
  [1] "wave"          "ivfio"         "ioutcome"      "adstatus"
"natch05"       "natch06"       "natch07"
  [8] "natch08"       "natch09"       "natch10"       "natch11"
"natch12"       "natch13"       "natch14"
 [15] "natch15"       "natch16"       "adoptch04"     "adoptch05"
"adoptch06"     "adoptch07"     "adoptch08"
 [22] "adoptch09"     "adoptch10"     "adoptch11"     "adoptch12"
```

```
"adoptch13"        "adoptch14"        "adoptch15"
 [29] "adoptch16"        "allch05"          "allch06"          "allch07"
"allch08"          "allch09"          "allch10"
 [36] "allch11"          "allch12"          "allch13"          "allch14"
"allch15"          "allch16"          "lieng"
 [43] "litrans"          "chkresp"          "btype10"          "btype11"
"btype12"          "btype13"          "btype14"
 [50] "bensta3"          "bensta9"          "bensta10"         "bensta11"
"bensta12"         "bensta13"         "casiend"
 [57] "precog"           "wrdrecpre"        "vfpre"            "nuabpre"
"cogend"           "hearcomputer"     "liceng"
 [64] "lictrans"         "ivlieng"          "ivlitrans"        "indmode"
"fiyrinvinc_tc"    "fibenothr_tc"     "intdatd_if"
 [71] "intdatm_if"       "intdaty_if"       "doby_if"          "age_if"
"ff_tel"           "ngrp_dv"          "xtra5min_dv"
 [78] "grfpno"           "grmpno"           "indpxus_lw"       "indinus_lw"
"indscus_lw"       "ind5mus_lw"       "ind5mus_xw"
 [85] "ficode21"         "ficode25"         "ficode30"         "ficode31"
"ficode32"         "ficode34"         "ficode35"
 [92] "ficode36"         "c_splitnum"       "medtyp13"         "bpconst"
"dateok"           "elig"             "nbrcohdk_dv"
 [99] "xtra5minosm_dv" "scflag_dv"
[1] "————————Dummy Variables————————"
[1] "predictor variable count went from 457 to 1168"
[1] "————————Variance 0 Check————————"
[1] "140 variables removed since their new variance was 0"
  [1] "pno.6"                              "pno.7"
"pno.9"
  [4] "pno.10"                             "nch14resp.5"
"nch3resp.2"
  [7] "nch5resp.2"                         "nch8resp.2"
"nch415resp.4"
 [10] "nch415resp.5"                       "nchresp.5"
"nchund18resp.5"
 [13] "natch01.6"                          "natch02.2"
"natch02.8"
 [16] "natch03.7"                          "natch04.7"
"nnatch.5"
 [19] "nnatch.6"                           "nnatch.7"
"adoptch01.5"
 [22] "adoptch01.6"                        "adoptch02.5"
"adoptch02.6"
 [25] "adoptch03.5"                        "adoptch03.7"
"nchunder16.5"
 [28] "nch5to15.4"                         "nch10.2"
"allch02.7"
 [31] "allch02.8"                          "allch03.8"
"allch04.5"
 [34] "allch04.7"                          "allch04.9"
"jbstat.x.10"
 [37] "marstat.x.3"                        "vftctw.5"
"vftctw.6"
 [40] "vftctw.7"                           "vftctw.9"
"vftctw.14"
 [43] "vftctw.16"                          "vftctw.17"
"vftctw.18"
 [46] "vftctw.19"                          "clangab.5"
"ivcoop.4"
 [49] "undqus.5"                           "hgbiom.4"
"hgbiom.5"
 [52] "hgbiof.4"                           "hgbiof.7"
"pn1pno.4"
 [55] "pn1pno.5"                           "pn1pno.6"
"pn2pno.4"
 [58] "pn2pno.7"                           "pns1pno.4"
"pns1pno.5"
 [61] "pns1pno.6"                          "pns2pno.4"
"pns2pno.7"
```

```
 [64] "nnssib_dv.4"                      "nnssib_dv.5"
"nnssib_dv.6"
 [67] "nnssib_dv.7"                      "mastat_dv.3"
"buno_dv.5"
 [70] "buno_dv.6"                        "buno_dv.10"
"nchild_dv.5"
 [73] "hrpno.6"                          "hrpno.10"
"ppno.6"
 [76] "ppno.7"                           "sppno.5"
"sppno.6"
 [79] "sppno.7"                          "fnpno.4"
"fnpno.6"
 [82] "fnspno.4"                         "fnspno.6"
"mnpno.4"
 [85] "mnpno.5"                          "mnspno.4"
"mnspno.5"
 [88] "cgvfw_dv.5"                       "cgvfw_dv.6"
"cgvfw_dv.7"
 [91] "cgvfw_dv.9"                       "cgvfw_dv.14"
"cgvfw_dv.16"
 [94] "cgvfw_dv.17"                      "cgvfw_dv.18"
"cgvfw_dv.19"
 [97] "fimnlabgrs_if.0.0299999993294477" "fimnlabgrs_if.0.0399999991059303" "fimnlabgrs_if.0.14000000
[100] "fimnlabgrs_if.0.620000004768372"  "fimnlabgrs_if.0.689999997615814"
"fimnlabgrs_if.0.819999992847443"
[103] "fimnlabgrs_if.0.850000023841858"  "fimnlabgrs_if.0.939999997615814"
"ficode3.3"
[106] "ficode4.3"                        "ficode24.2"
"ficode27.2"
[109] "ficode28.2"                       "ficode29.3"
"ficode38.2"
[112] "c_pno.6"                          "c_pno.7"
"c_pno.9"
[115] "c_pno.10"                         "medcnjd.3"
"resphts.2"
[118] "respbps.5"                        "nseqno.6"
"nseqno.7"
[121] "nseqno.9"                         "nseqno.10"
"htok.3"
[124] "bmiok.2"                          "bmiok.3"
"bmiok.4"
[127] "ag16g10.96"                       "ag16g20.96"
"wstokb.5"
[130] "hhsize.8"                         "hhsize.9"
"hhsize.10"
[133] "hhsize.14"                        "jbstat.y.10"
"marstat.y.3"
[136] "npensioner_dv.3"                  "nnsib_dv.4"
"nnsib_dv.5"
[139] "nnsib_dv.6"                       "nnsib_dv.7"
[1] "————————K–Means————————"
[1] "15 clusters have been made for K–Means"
[1] "K–Means results as a table, the max value in each row is a simple way to define which cluster re

       0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31
   1   2   0   1   1   2   7  34  29  21  20  19  23  17  11   7   7   2   5   4   3   4   0   1
5   5   1   0   1   4   0   2   1   0   1   1   0   0
   2   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0   0
   3   0   1   1   3   3  12  68  43  40  47  50  31  47  14  15  18  13  11  10   5   9   7  10
3   5   3   3   3   3   1   3   1   0   0   2   1   0
   4   0   0   0   1   0   1   5   7  11   7   3   7   2   3   3   2   0   2   1   1   0   0   0
1   2   1   0   0   1   1   0   0   1   0   0   0   0
   5   0   0   0   0   0   0   5   4   4   2   8   3   5   2   0   4   1   0   0   0   0   2   0
0   0   0   0   0   0   0   0   0   0   0   0   0   0
   6   3   0   0   2   3   7  21  25  25  16  19  18  29  13   7   5   4   6   3   0   6   5   7
1   5   1   2   0   1   0   0   1   1   1   0   0   1
   7   1   3   0   0   1   5  23  29  18  11  19  12  16   3   3   3   3   2   3   3   3   3   2
```

```
3  1  1  0  0  0  0  2  0  0  1  0  0  0
   8   1  0  3  2  3  9 40 31 24 25 17 22 20 13 11 11 12  7  6  7  1  6  0
3  4  1  1  2  0  1  0  1  1  1  0  1  1
   9   0  0  0  0  0  3  8 13 14  6  5  8  6  7  2  2  0  1  1  0  1  1  0
2  0  1  1  0  1  0  0  0  0  0  0  0  2
   10  0  0  0  1  1 10 33 23 19 33 18 13 15 13  8  4  8  4  3  6  5  2  1
2  1  1  1  1  0  0  0  0  0  0  1  0  0
   11  0  0  0  3  0  1 34 24 25 22 15 13 28  9 12  3  4  4  9  1  3  4  1
1  3  3  3  0  0  0  0  1  1  0  1  0  0
   12  0  0  0  2  2 11 49 43 32 31 26 38 40 23 10 13 13  9  5  7 10  6  4
5  5  2  1  1  1  0  1  3  1  0  1  3  1
   13  0  1  0  1  0  5 18 17 14 15 14 12 12  5  8  2  6  4  3  2  0  3  2
0  1  0  0  0  0  1  1  0  0  1  1  0  0
   14  0  1  0  1  3  6 14 14 14 15 17 13 20  8  3  2  4  2  4  4  4  2  2
2  3  0  1  1  0  0  1  0  0  1  0  0  0
   15  1  0  0  0  0  0  2  3  4  2  4  6  2  0  1  2  1  1  0  2  0  1  2
0  0  0  1  0  0  1  0  0  0  1  0  0  0
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
 [1] "Cluster 1: Within MSE 129039958660147, Size 241"   "Cluster 2: Within MSE 0, Size 1"
 [3] "Cluster 3: Within MSE 4801524085498109, Size 486"  "Cluster 4: Within MSE 65531106972525, Size
 [5] "Cluster 5: Within MSE 318105096331, Size 40"       "Cluster 6: Within MSE 125586026905488, Size
 [7] "Cluster 7: Within MSE 737346617508519, Size 174"   "Cluster 8: Within MSE 118001190470054, Size
 [9] "Cluster 9: Within MSE 53355138877964, Size 85"     "Cluster 10: Within MSE 106404596284972, Siz
[11] "Cluster 11: Within MSE 137846883507380, Size 228"  "Cluster 12: Within MSE 4815943569162807, S
[13] "Cluster 13: Within MSE 5898365508968, Size 149"    "Cluster 14: Within MSE 5919194388446, Size
[15] "Cluster 15: Within MSE 49836345053257, Size 37"
[1] "Total between cluster MSE: 224333987242009920, Total within cluster MSE: 1613312264958607"
[1] "The K-Means model predicts exactly with an accuracy of 0.1398"
[1] "——————Correlation Checks——————"
[1] "jbstat.x.4 removed, correlated with 29 other variable(s)"
[1] "dvage removed, correlated with 26 other variable(s)"
[1] "birthy removed, correlated with 25 other variable(s)"
[1] "plivpar.2 removed, correlated with 22 other variable(s)"
[1] "pensioner_dv.2 removed, correlated with 25 other variable(s)"
[1] "pns1pno.1 removed, correlated with 23 other variable(s)"
[1] "btype4.1 removed, correlated with 21 other variable(s)"
[1] "age_dv removed, correlated with 23 other variable(s)"
[1] "memcont11.1 removed, correlated with 18 other variable(s)"
[1] "ficode1.1 removed, correlated with 22 other variable(s)"
[1] "memcont13.1 removed, correlated with 17 other variable(s)"
[1] "pn1pno.1 removed, correlated with 19 other variable(s)"
[1] "subcont11.1 removed, correlated with 16 other variable(s)"
[1] "ficode18.1 removed, correlated with 23 other variable(s)"
[1] "subcont13.1 removed, correlated with 15 other variable(s)"
[1] "confage removed, correlated with 19 other variable(s)"
[1] "nscont11.1 removed, correlated with 14 other variable(s)"
[1] "age removed, correlated with 18 other variable(s)"
[1] "nscont13.1 removed, correlated with 13 other variable(s)"
[1] "indinub_lw removed, correlated with 13 other variable(s)"
[1] "jbstat.y.4 removed, correlated with 19 other variable(s)"
[1] "dmemcont11.1 removed, correlated with 12 other variable(s)"
[1] "pns2pno.2 removed, correlated with 14 other variable(s)"
[1] "indpxub_lw removed, correlated with 12 other variable(s)"
[1] "psu.x removed, correlated with 13 other variable(s)"
[1] "nch14resp.2 removed, correlated with 11 other variable(s)"
[1] "nchresp.2 removed, correlated with 12 other variable(s)"
[1] "btype5.1 removed, correlated with 15 other variable(s)"
[1] "memcont12.1 removed, correlated with 11 other variable(s)"
[1] "dmemcont13.1 removed, correlated with 11 other variable(s)"
[1] "pn2pno.2 removed, correlated with 12 other variable(s)"
[1] "subcont12.1 removed, correlated with 10 other variable(s)"
[1] "vfcont11.1 removed, correlated with 10 other variable(s)"
[1] "agegr13_dv.13 removed, correlated with 13 other variable(s)"
[1] "nchund18resp.2 removed, correlated with 10 other variable(s)"
[1] "npn_dv.2 removed, correlated with 11 other variable(s)"
[1] "hhorig.x.4 removed, correlated with 9 other variable(s)"
[1] "nscont12.1 removed, correlated with 9 other variable(s)"
[1] "vfcont13.1 removed, correlated with 9 other variable(s)"
```

```
[1] "indin91_lw removed, correlated with 9 other variable(s)"
[1] "indscub_lw removed, correlated with 10 other variable(s)"
[1] "indnsub_lw.x removed, correlated with 11 other variable(s)"
[1] "natch02.4 removed, correlated with 9 other variable(s)"
[1] "doby_dv removed, correlated with 14 other variable(s)"
[1] "memorig.4 removed, correlated with 8 other variable(s)"
[1] "nch14resp.1 removed, correlated with 8 other variable(s)"
[1] "nch14resp.3 removed, correlated with 8 other variable(s)"
[1] "nchresp.1 removed, correlated with 9 other variable(s)"
[1] "dmemcont12.1 removed, correlated with 8 other variable(s)"
[1] "nacont11.1 removed, correlated with 8 other variable(s)"
[1] "country.3 removed, correlated with 9 other variable(s)"
[1] "nch14resp.4 removed, correlated with 7 other variable(s)"
[1] "nchund18resp.3 removed, correlated with 8 other variable(s)"
[1] "vfcont12.1 removed, correlated with 7 other variable(s)"
[1] "nacont13.1 removed, correlated with 7 other variable(s)"
[1] "hhorig.x.5 removed, correlated with 6 other variable(s)"
[1] "strata.x removed, correlated with 6 other variable(s)"
[1] "month.12 removed, correlated with 6 other variable(s)"
[1] "nch415resp.1 removed, correlated with 6 other variable(s)"
[1] "nch415resp.2 removed, correlated with 6 other variable(s)"
[1] "nchresp.4 removed, correlated with 6 other variable(s)"
[1] "natch04.6 removed, correlated with 7 other variable(s)"
[1] "nchunder16.2 removed, correlated with 7 other variable(s)"
[1] "nchunder16.3 removed, correlated with 8 other variable(s)"
[1] "allch01.3 removed, correlated with 7 other variable(s)"
[1] "jbstat.x.2 removed, correlated with 6 other variable(s)"
[1] "marstat.x.2 removed, correlated with 6 other variable(s)"
[1] "nacont12.1 removed, correlated with 6 other variable(s)"
[1] "ivprsnt.2 removed, correlated with 6 other variable(s)"
[1] "hgbiom.2 removed, correlated with 6 other variable(s)"
[1] "sppno.1 removed, correlated with 8 other variable(s)"
[1] "region.3 removed, correlated with 7 other variable(s)"
[1] "fimnlabgrs_dv removed, correlated with 8 other variable(s)"
[1] "mastat_dv.6 removed, correlated with 6 other variable(s)"
[1] "indns91_lw.x removed, correlated with 6 other variable(s)"
[1] "respwts.5 removed, correlated with 6 other variable(s)"
[1] "pno.5 removed, correlated with 5 other variable(s)"
[1] "memorig.5 removed, correlated with 5 other variable(s)"
[1] "adoptch02.3 removed, correlated with 7 other variable(s)"
[1] "hgbiof.1 removed, correlated with 5 other variable(s)"
[1] "hgbiof.3 removed, correlated with 5 other variable(s)"
[1] "pn1pno.3 removed, correlated with 6 other variable(s)"
[1] "cgivns1_dv.2 removed, correlated with 5 other variable(s)"
[1] "cgivns1_dv.3 removed, correlated with 5 other variable(s)"
[1] "cgivns1_dv.4 removed, correlated with 5 other variable(s)"
[1] "indinub_xw removed, correlated with 5 other variable(s)"
[1] "bpmedc.1 removed, correlated with 7 other variable(s)"
[1] "nchresp.3 removed, correlated with 5 other variable(s)"
[1] "natch03.5 removed, correlated with 5 other variable(s)"
[1] "nnatch.4 removed, correlated with 5 other variable(s)"
[1] "scsf2a.3 removed, correlated with 5 other variable(s)"
[1] "scsf4a.5 removed, correlated with 5 other variable(s)"
[1] "full3.2 removed, correlated with 7 other variable(s)"
[1] "pidp removed, correlated with 4 other variable(s)"
[1] "pno.2 removed, correlated with 4 other variable(s)"
[1] "pno.3 removed, correlated with 4 other variable(s)"
[1] "jbstat.x.7 removed, correlated with 4 other variable(s)"
[1] "jbhas.2 removed, correlated with 4 other variable(s)"
[1] "marstat.x.6 removed, correlated with 4 other variable(s)"
[1] "memper.3 removed, correlated with 4 other variable(s)"
[1] "hgbiom.1 removed, correlated with 4 other variable(s)"
[1] "fimnsben_dv removed, correlated with 4 other variable(s)"
[1] "mastat_dv.2 removed, correlated with 4 other variable(s)"
[1] "ppno.2 removed, correlated with 5 other variable(s)"
[1] "fnpno.1 removed, correlated with 4 other variable(s)"
[1] "indpxub_xw removed, correlated with 4 other variable(s)"
[1] "hhorig.y.4 removed, correlated with 4 other variable(s)"
```

```
[1] "lfout.6 removed, correlated with 5 other variable(s)"
[1] "numed2 removed, correlated with 7 other variable(s)"
[1] "indns91_lw.y removed, correlated with 5 other variable(s)"
[1] "indnsub_lw.y removed, correlated with 6 other variable(s)"
[1] "indnsbh_xw removed, correlated with 7 other variable(s)"
[1] "cgivwrd1_dv.2 removed, correlated with 4 other variable(s)"
[1] "cgivwrd1_dv.3 removed, correlated with 4 other variable(s)"
[1] "cgivwrd1_dv.4 removed, correlated with 4 other variable(s)"
[1] "natch02.3 removed, correlated with 4 other variable(s)"
[1] "nnatch.1 removed, correlated with 4 other variable(s)"
[1] "nchunder16.4 removed, correlated with 4 other variable(s)"
[1] "allch02.4 removed, correlated with 5 other variable(s)"
[1] "btype3.1 removed, correlated with 4 other variable(s)"
[1] "scsf2b.3 removed, correlated with 4 other variable(s)"
[1] "pns1pno.3 removed, correlated with 4 other variable(s)"
[1] "hhorig.y.5 removed, correlated with 4 other variable(s)"
[1] "wtok.5 removed, correlated with 5 other variable(s)"
[1] "hidp removed, correlated with 3 other variable(s)"
[1] "pno.4 removed, correlated with 3 other variable(s)"
[1] "month.2 removed, correlated with 3 other variable(s)"
[1] "month.3 removed, correlated with 3 other variable(s)"
[1] "month.4 removed, correlated with 3 other variable(s)"
[1] "month.8 removed, correlated with 3 other variable(s)"
[1] "month.9 removed, correlated with 3 other variable(s)"
[1] "month.10 removed, correlated with 3 other variable(s)"
[1] "sex.2 removed, correlated with 3 other variable(s)"
[1] "psiblings.2 removed, correlated with 3 other variable(s)"
[1] "nnatch.3 removed, correlated with 3 other variable(s)"
[1] "nadoptch.3 removed, correlated with 3 other variable(s)"
[1] "jbstat.x.3 removed, correlated with 3 other variable(s)"
[1] "sf1.x.5 removed, correlated with 3 other variable(s)"
[1] "bensta4.1 removed, correlated with 3 other variable(s)"
[1] "marstat.x.4 removed, correlated with 3 other variable(s)"
[1] "marstat.x.5 removed, correlated with 3 other variable(s)"
[1] "employ.2 removed, correlated with 3 other variable(s)"
[1] "memper.4 removed, correlated with 3 other variable(s)"
[1] "npn_dv.1 removed, correlated with 3 other variable(s)"
[1] "agegr5_dv.6 removed, correlated with 3 other variable(s)"
[1] "agegr5_dv.7 removed, correlated with 3 other variable(s)"
[1] "agegr5_dv.8 removed, correlated with 3 other variable(s)"
[1] "agegr5_dv.9 removed, correlated with 3 other variable(s)"
[1] "agegr5_dv.10 removed, correlated with 3 other variable(s)"
[1] "agegr5_dv.11 removed, correlated with 3 other variable(s)"
[1] "agegr5_dv.13 removed, correlated with 3 other variable(s)"
[1] "agegr5_dv.15 removed, correlated with 3 other variable(s)"
[1] "buno_dv.3 removed, correlated with 3 other variable(s)"
[1] "nchild_dv.3 removed, correlated with 5 other variable(s)"
[1] "fnspno.1 removed, correlated with 3 other variable(s)"
[1] "respbps.4 removed, correlated with 3 other variable(s)"
[1] "ag16g10.26 removed, correlated with 5 other variable(s)"
[1] "ag16g10.36 removed, correlated with 5 other variable(s)"
[1] "ag16g10.46 removed, correlated with 5 other variable(s)"
[1] "ag16g10.56 removed, correlated with 4 other variable(s)"
[1] "ag16g20.36 removed, correlated with 4 other variable(s)"
[1] "ag16g20.56 removed, correlated with 5 other variable(s)"
[1] "cgivs71_dv.2 removed, correlated with 3 other variable(s)"
[1] "cgivs71_dv.3 removed, correlated with 3 other variable(s)"
[1] "cgivs71_dv.4 removed, correlated with 3 other variable(s)"
[1] "nchunder16.1 removed, correlated with 3 other variable(s)"
[1] "nch5to15.1 removed, correlated with 3 other variable(s)"
[1] "allch01.2 removed, correlated with 3 other variable(s)"
[1] "allch04.6 removed, correlated with 3 other variable(s)"
[1] "lvrel2.1 removed, correlated with 3 other variable(s)"
[1] "scsf3a.5 removed, correlated with 3 other variable(s)"
[1] "fimnlabgrs_if.1 removed, correlated with 4 other variable(s)"
[1] "indbdub_lw removed, correlated with 3 other variable(s)"
[1] "whintro.2 removed, correlated with 3 other variable(s)"
[1] "marstat.y.2 removed, correlated with 3 other variable(s)"
```

```
[1] "marstat.y.6 removed, correlated with 3 other variable(s)"
[1] "fimnnet_dv removed, correlated with 3 other variable(s)"
[1] "marstat_dv.3 removed, correlated with 4 other variable(s)"
[1] "month.7 removed, correlated with 2 other variable(s)"
[1] "month.11 removed, correlated with 2 other variable(s)"
[1] "isyear.2012 removed, correlated with 2 other variable(s)"
[1] "natch01.3 removed, correlated with 2 other variable(s)"
[1] "nnatch.2 removed, correlated with 2 other variable(s)"
[1] "adoptch03.4 removed, correlated with 2 other variable(s)"
[1] "jbstat.x.8 removed, correlated with 2 other variable(s)"
[1] "sf1.x.2 removed, correlated with 2 other variable(s)"
[1] "sf1.x.3 removed, correlated with 2 other variable(s)"
[1] "sf1.x.4 removed, correlated with 2 other variable(s)"
[1] "health.x.2 removed, correlated with 2 other variable(s)"
[1] "j2has.2 removed, correlated with 2 other variable(s)"
[1] "btype1.1 removed, correlated with 2 other variable(s)"
[1] "btype6.1 removed, correlated with 2 other variable(s)"
[1] "btype8.1 removed, correlated with 2 other variable(s)"
[1] "bensta2.1 removed, correlated with 2 other variable(s)"
[1] "bensta7.1 removed, correlated with 2 other variable(s)"
[1] "hgbiom.3 removed, correlated with 2 other variable(s)"
[1] "hgbiof.2 removed, correlated with 2 other variable(s)"
[1] "respm16.2 removed, correlated with 2 other variable(s)"
[1] "fimngrs_tc.1 removed, correlated with 2 other variable(s)"
[1] "country.2 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.5 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.12 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.14 removed, correlated with 2 other variable(s)"
[1] "cohab_dv.1 removed, correlated with 2 other variable(s)"
[1] "mastat_dv.4 removed, correlated with 2 other variable(s)"
[1] "mastat_dv.5 removed, correlated with 2 other variable(s)"
[1] "buno_dv.4 removed, correlated with 2 other variable(s)"
[1] "nchild_dv.2 removed, correlated with 3 other variable(s)"
[1] "hrpid removed, correlated with 2 other variable(s)"
[1] "ppno.1 removed, correlated with 2 other variable(s)"
[1] "fnspno.2 removed, correlated with 3 other variable(s)"
[1] "mnpno.1 removed, correlated with 2 other variable(s)"
[1] "mnpno.2 removed, correlated with 2 other variable(s)"
[1] "cgsrmem_dv.3 removed, correlated with 2 other variable(s)"
[1] "frmnthimp_dv_total removed, correlated with 2 other variable(s)"
[1] "c_pno.3 removed, correlated with 2 other variable(s)"
[1] "nsex.2 removed, correlated with 2 other variable(s)"
[1] "medcnjd.2 removed, correlated with 2 other variable(s)"
[1] "medtyp1.1 removed, correlated with 3 other variable(s)"
[1] "resphts.4 removed, correlated with 2 other variable(s)"
[1] "whintro.3 removed, correlated with 2 other variable(s)"
[1] "consubx1.1 removed, correlated with 2 other variable(s)"
[1] "consubx2.1 removed, correlated with 2 other variable(s)"
[1] "consubx5.1 removed, correlated with 3 other variable(s)"
[1] "respbps.3 removed, correlated with 2 other variable(s)"
[1] "ethnic removed, correlated with 2 other variable(s)"
[1] "full1.2 removed, correlated with 2 other variable(s)"
[1] "full2.2 removed, correlated with 3 other variable(s)"
[1] "ag16g10.76 removed, correlated with 2 other variable(s)"
[1] "hhtype_dv.y removed, correlated with 2 other variable(s)"
[1] "jbstat.y.2 removed, correlated with 2 other variable(s)"
[1] "jbstat.y.7 removed, correlated with 3 other variable(s)"
[1] "psu.y removed, correlated with 2 other variable(s)"
[1] "rach16_dv.2 removed, correlated with 3 other variable(s)"
[1] "cgivvf1_dv.2 removed, correlated with 2 other variable(s)"
[1] "cgivvf1_dv.3 removed, correlated with 2 other variable(s)"
[1] "cgivvf1_dv.4 removed, correlated with 2 other variable(s)"
[1] "cmroute.1 removed, correlated with 2 other variable(s)"
[1] "scsf3b.4 removed, correlated with 2 other variable(s)"
[1] "scsf3b.5 removed, correlated with 2 other variable(s)"
[1] "scsf4b.5 removed, correlated with 2 other variable(s)"
[1] "scsf6c.5 removed, correlated with 2 other variable(s)"
[1] "scsf7.5 removed, correlated with 2 other variable(s)"
```

```
[1] "nbrcoh_dv removed, correlated with 2 other variable(s)"
[1] "nchild_dv.4 removed, correlated with 2 other variable(s)"
[1] "fnpno.3 removed, correlated with 2 other variable(s)"
[1] "ficode26.1 removed, correlated with 2 other variable(s)"
[1] "sf1.y.5 removed, correlated with 2 other variable(s)"
[1] "livesp_dv.1 removed, correlated with 2 other variable(s)"
[1] "month.5 removed, correlated with 1 other variable(s)"
[1] "month.6 removed, correlated with 1 other variable(s)"
[1] "nch415resp.3 removed, correlated with 1 other variable(s)"
[1] "nchund18resp.1 removed, correlated with 1 other variable(s)"
[1] "natch01.5 removed, correlated with 1 other variable(s)"
[1] "natch03.4 removed, correlated with 1 other variable(s)"
[1] "nadoptch.1 removed, correlated with 1 other variable(s)"
[1] "nadoptch.2 removed, correlated with 1 other variable(s)"
[1] "allch02.3 removed, correlated with 1 other variable(s)"
[1] "allch02.5 removed, correlated with 1 other variable(s)"
[1] "allch03.5 removed, correlated with 1 other variable(s)"
[1] "istrtdaty.2012 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.2 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.3 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.4 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.5 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.6 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.7 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.8 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.9 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.10 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.11 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.12 removed, correlated with 1 other variable(s)"
[1] "istrtdatd removed, correlated with 1 other variable(s)"
[1] "jbstat.x.5 removed, correlated with 1 other variable(s)"
[1] "jbstat.x.6 removed, correlated with 1 other variable(s)"
[1] "jbstat.x.97 removed, correlated with 1 other variable(s)"
[1] "locserd.2 removed, correlated with 1 other variable(s)"
[1] "nbrcoh1.2 removed, correlated with 1 other variable(s)"
[1] "nbrcoh2.2 removed, correlated with 1 other variable(s)"
[1] "nbrcoh3.2 removed, correlated with 1 other variable(s)"
[1] "nbrcoh4.3 removed, correlated with 1 other variable(s)"
[1] "simfam.3 removed, correlated with 1 other variable(s)"
[1] "org.2 removed, correlated with 1 other variable(s)"
[1] "lvrel1.1 removed, correlated with 1 other variable(s)"
[1] "lvrel3.1 removed, correlated with 1 other variable(s)"
[1] "lvrel96.1 removed, correlated with 1 other variable(s)"
[1] "visfam.4 removed, correlated with 1 other variable(s)"
[1] "btype2.1 removed, correlated with 1 other variable(s)"
[1] "btype9.1 removed, correlated with 1 other variable(s)"
[1] "bensta5.1 removed, correlated with 1 other variable(s)"
[1] "bensta6.1 removed, correlated with 1 other variable(s)"
[1] "bensta96.1 removed, correlated with 1 other variable(s)"
[1] "bensta97.1 removed, correlated with 1 other variable(s)"
[1] "civicduty.5 removed, correlated with 1 other variable(s)"
[1] "civicduty.6 removed, correlated with 1 other variable(s)"
[1] "perbfts.2 removed, correlated with 1 other variable(s)"
[1] "poleff1.5 removed, correlated with 1 other variable(s)"
[1] "poleff2.3 removed, correlated with 1 other variable(s)"
[1] "poleff3.2 removed, correlated with 1 other variable(s)"
[1] "scsf1.2 removed, correlated with 1 other variable(s)"
[1] "scsf1.3 removed, correlated with 1 other variable(s)"
[1] "scsf1.4 removed, correlated with 1 other variable(s)"
[1] "scsf3a.2 removed, correlated with 1 other variable(s)"
[1] "scsf3a.3 removed, correlated with 1 other variable(s)"
[1] "scsf6a.2 removed, correlated with 1 other variable(s)"
[1] "scsf6b.2 removed, correlated with 1 other variable(s)"
[1] "scopngbha.2 removed, correlated with 1 other variable(s)"
[1] "scopngbhb.2 removed, correlated with 1 other variable(s)"
[1] "scopngbhe.2 removed, correlated with 1 other variable(s)"
[1] "scopngbhg.2 removed, correlated with 1 other variable(s)"
[1] "scptrt5c1.7 removed, correlated with 1 other variable(s)"
```

```
[1] "scptrt5o1.7 removed, correlated with 1 other variable(s)"
[1] "scptrt5a2.7 removed, correlated with 1 other variable(s)"
[1] "scptrt5e2.7 removed, correlated with 1 other variable(s)"
[1] "scptrt5n2.7 removed, correlated with 1 other variable(s)"
[1] "scptrt5o2.7 removed, correlated with 1 other variable(s)"
[1] "scptrt5a3.7 removed, correlated with 1 other variable(s)"
[1] "scptrt5c3.7 removed, correlated with 1 other variable(s)"
[1] "memper.2 removed, correlated with 1 other variable(s)"
[1] "memper.5 removed, correlated with 1 other variable(s)"
[1] "vftctw.1 removed, correlated with 1 other variable(s)"
[1] "vftctw.2 removed, correlated with 1 other variable(s)"
[1] "vftctw.3 removed, correlated with 1 other variable(s)"
[1] "vftctw.4 removed, correlated with 1 other variable(s)"
[1] "nadisease.2 removed, correlated with 1 other variable(s)"
[1] "vfct removed, correlated with 1 other variable(s)"
[1] "memprob1.1 removed, correlated with 1 other variable(s)"
[1] "memprob3.1 removed, correlated with 1 other variable(s)"
[1] "memaid.2 removed, correlated with 1 other variable(s)"
[1] "respf16.2 removed, correlated with 1 other variable(s)"
[1] "istrtdathh removed, correlated with 1 other variable(s)"
[1] "fimnlabgrs_tc.1 removed, correlated with 1 other variable(s)"
[1] "j2paynet_dv removed, correlated with 1 other variable(s)"
[1] "pn1pno.2 removed, correlated with 1 other variable(s)"
[1] "pns2pno.3 removed, correlated with 1 other variable(s)"
[1] "fimninvnet_dv removed, correlated with 1 other variable(s)"
[1] "nnssib_dv.1 removed, correlated with 1 other variable(s)"
[1] "nnssib_dv.2 removed, correlated with 1 other variable(s)"
[1] "nnssib_dv.3 removed, correlated with 1 other variable(s)"
[1] "urban_dv.x.2 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.4 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.5 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.6 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.7 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.8 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.9 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.10 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.11 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.12 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.10 removed, correlated with 1 other variable(s)"
[1] "ppno.3 removed, correlated with 1 other variable(s)"
[1] "ppno.4 removed, correlated with 1 other variable(s)"
[1] "mnpno.3 removed, correlated with 1 other variable(s)"
[1] "mnspno.1 removed, correlated with 1 other variable(s)"
[1] "mnspno.2 removed, correlated with 1 other variable(s)"
[1] "cgsrmem_dv.4 removed, correlated with 1 other variable(s)"
[1] "paygu_if.1 removed, correlated with 1 other variable(s)"
[1] "fibenothr_if removed, correlated with 1 other variable(s)"
[1] "indin01_lw removed, correlated with 1 other variable(s)"
[1] "indscub_xw removed, correlated with 1 other variable(s)"
[1] "ficode2.1 removed, correlated with 1 other variable(s)"
[1] "ficode14.1 removed, correlated with 1 other variable(s)"
[1] "ficode16.1 removed, correlated with 1 other variable(s)"
[1] "ficode22.1 removed, correlated with 1 other variable(s)"
[1] "c_hidp removed, correlated with 1 other variable(s)"
[1] "c_pno.2 removed, correlated with 1 other variable(s)"
[1] "c_pno.4 removed, correlated with 1 other variable(s)"
[1] "c_pno.5 removed, correlated with 1 other variable(s)"
[1] "region.2 removed, correlated with 1 other variable(s)"
[1] "resphts.3 removed, correlated with 1 other variable(s)"
[1] "estht removed, correlated with 1 other variable(s)"
[1] "respwts.2 removed, correlated with 1 other variable(s)"
[1] "respwts.3 removed, correlated with 1 other variable(s)"
[1] "respwts.4 removed, correlated with 1 other variable(s)"
[1] "omronno removed, correlated with 1 other variable(s)"
[1] "cufsize.2 removed, correlated with 1 other variable(s)"
[1] "respbps.2 removed, correlated with 1 other variable(s)"
[1] "clotb.2 removed, correlated with 1 other variable(s)"
[1] "nseqno.3 removed, correlated with 1 other variable(s)"
```

```
[1] "nuroutc.12 removed, correlated with 1 other variable(s)"
[1] "htok.5 removed, correlated with 1 other variable(s)"
[1] "ag16g10.66 removed, correlated with 1 other variable(s)"
[1] "ag16g20.76 removed, correlated with 1 other variable(s)"
[1] "hhsize.2 removed, correlated with 1 other variable(s)"
[1] "health.y.2 removed, correlated with 1 other variable(s)"
[1] "marstat.y.4 removed, correlated with 1 other variable(s)"
[1] "marstat.y.5 removed, correlated with 1 other variable(s)"
[1] "strata.y removed, correlated with 1 other variable(s)"
[1] "hhtype_dv.x removed, correlated with 1 other variable(s)"
[1] "jbiindb_dv removed, correlated with 1 other variable(s)"
[1] "fimngrs_dv removed, correlated with 1 other variable(s)"
[1] "fibenothr_dv removed, correlated with 1 other variable(s)"
[1] "marstat_dv.6 removed, correlated with 1 other variable(s)"
[1] "cgivwri1_dv.2 removed, correlated with 1 other variable(s)"
[1] "cgivwri1_dv.3 removed, correlated with 1 other variable(s)"
[1] "cgivwri1_dv.4 removed, correlated with 1 other variable(s)"
[1] "racel_dv removed, correlated with 1 other variable(s)"
[1] "cgs7n_dv.1 removed, correlated with 1 other variable(s)"
[1] "big5a_dv.5 removed, correlated with 1 other variable(s)"
[1] "399 variables removed since they had high correlation coefs"
[1] "The final cleaned dataset has been completed at this stage and is stored under the name 'cleanDa
[1] "—————Attempting a Train Test Split—————"
[1] "Good train, test split found"
[1] "The working seed found was 3"
[1] "—————kNN—————"
[1] "59 neighbours considered for each test data point"
[1] "kNN results as a table, follow the diagonal for the correctly mapped clusters"
         real
predicted  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 3
       0   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
       1   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
       2   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
       3   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
       4   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
       5   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
       6   0  0  1  1  1 14 61 54 39 40 37 21 27 13 12  8  5  4  3  5  6
3  4  4  3  1  4  0  2  2  0  0  0  0
       7   1  1  0  0  2  4 16 14 15 12 14 14  8  9  4  1  3  5  1  2  1
2  0  1  5  0  2  1  0  0  0  0  0  0
       8   1  0  0  1  0  1  7  3  6  1  2  9  8  4  3  2  0  2  0  1  1
2  2  1  1  3  0  0  0  0  0  1  0  0
       9   0  0  0  0  0  0  2  5  2  4  6  3  5  4  1  2  1  2  3  1  3
1  1  0  1  1  0  1  1  0  0  0  1  0
       10  0  0  0  0  0  0  1  0  0  1  3  0  3  2  1  1  3  0  0  0  0
0  0  0  1  0  0  0  1  0  0  0  0  0
       11  0  0  0  0  0  0  0  0  0  0  2  0  5  1  2  0  2  0  1  1  3
1  1  0  0  0  0  0  0  0  1  0  0  0
       12  0  0  0  0  0  0  0  2  1  4  0  0  4  2  3  0  2  3  0  0  1
2  1  0  2  2  1  0  0  0  1  0  2  1
       13  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
       14  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
       15  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
       16  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  1
1  0  0  0  0  0  0  0  0  0  0  0  0
       17  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
       18  0  0  0  0  0  0  0  0  0  0  0  0  0  0  2  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  1  0  1
```

```
        19  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
        20  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
        21  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
        22  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
        23  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
        24  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
        25  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
        26  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
        27  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
        28  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0
 [ reached getOption("max.print") -- omitted 8 rows ]
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
[1] "The MSE of the predicted values are of 387.8693"
[1] "The kNN model predicts exactly with an accuracy of 0.1307"
[1] "------------CART prediction model------------"
n= 2114

node), split, n, deviance, yval
      * denotes terminal node

 1) root 2114 68208.4700 11.218070
   2) sf12mcs_dv>=-0.5503648 1591 19095.6700  9.184789
     4) sf12mcs_dv>=0.2706947 1003  6681.1570  7.957129
       8) scsf6c.4< 0.5 754  3986.2710  7.445623 *
       9) scsf6c.4>=0.5 249  1900.2410  9.506024 *
     5) sf12mcs_dv< 0.2706947 588  8324.2590 11.278910
      10) sf12pcs_dv>=-0.1943717 397  4069.1080 10.481110 *
      11) sf12pcs_dv< -0.1943717 191  3477.2460 12.937170 *
   3) sf12mcs_dv< -0.5503648 523 22525.8700 17.403440
     6) sf12mcs_dv>=-1.94745 430 11404.1000 15.662790
      12) sf12mcs_dv>=-1.226677 281  6407.5590 14.597860 *
      13) sf12mcs_dv< -1.226677 149  4076.8860 17.671140 *
     7) sf12mcs_dv< -1.94745 93  3795.0320 25.451610
      14) scsf6c.3>=0.5 18   294.2778 19.611110 *
      15) scsf6c.3< 0.5 75  2739.3870 26.853330 *
[1] "Variable Importance"
   sf12mcs_dv      scsf6c.3      scsf6a.4      scsf4a.3      scsf4a.2
scsf6c.2    sf12pcs_dv      scsf6c.4      scsf6a.3
 39035.274899   7661.583270   6151.085671   5083.541877   4827.907084   4546.125732
1975.976572   1504.179365   1266.031889
     scsf6a.5      scsf4a.4    big5n_dv.7      scsf2a.2   scopngbha.5
scsf5.4       scsf3b.3      scsf2b.2       sf1.y.4
  1102.949640    980.826903    945.385406    317.678121    315.128469
281.022954    268.804564    264.731768    256.586175
   nbrsnci_dv    cgwrd_dv.9   cgwri_dv.10    istrtdatmm     netpuse.4
sclfsat1.7    ficode15.1     nbrcoh3.5  cgivna1_dv.2
   236.346351    126.894636     84.596424     84.596424     84.596424
84.596424     30.861075     24.688860     15.956727
    perbfts.6 newsource11.1
    12.765382      9.574036
[1] "The MSE of the predicted values are of 13.6203"
[1] "The CART model predicts exactly with accuracy of 0.1605"
[1] "------------Ordinary Linear Regression (Initial)------------"
[1] "The full model AIC is: 11293.1656"
[1] "------------Variance Inflation Factor Removal------------"
[1] "The variable big5c_dv.6 was removed since it had a VIF score of 244.1626"
[1] "The variable nurdayy.2012 was removed since it had a VIF score of 204.2449"
[1] "The variable agegr10_dv.8 was removed since it had a VIF score of 174.9713"
```

```
[1] "The variable cgna_dv.3 was removed since it had a VIF score of 122.0473"
[1] "The variable cgwri_dv.7 was removed since it had a VIF score of 94.2737"
[1] "The variable nbrsnci_dv was removed since it had a VIF score of 75.8486"
[1] "The variable big5e_dv.5 was removed since it had a VIF score of 54.3504"
[1] "The variable ieqmoecd_dv was removed since it had a VIF score of 53.1611"
[1] "The variable big5o_dv.5 was removed since it had a VIF score of 37.1838"
[1] "The variable big5n_dv.4 was removed since it had a VIF score of 25.0977"
[1] "The variable sf12pcs_dv was removed since it had a VIF score of 17.7337"
[1] "The variable intdatm_dv.11 was removed since it had a VIF score of 16.4275"
[1] "The variable single_dv.1 was removed since it had a VIF score of 16.0616"
[1] "The variable sclfsato.6 was removed since it had a VIF score of 14.9702"
[1] "The variable scptrt5n3.5 was removed since it had a VIF score of 13.9906"
[1] "The variable respm16_dv.2 was removed since it had a VIF score of 13.3099"
[1] "The variable agegr10_dv.4 was removed since it had a VIF score of 12.8163"
[1] "The variable scptrt5o3.5 was removed since it had a VIF score of 11.8657"
[1] "The variable scptrt5e1.7 was removed since it had a VIF score of 11.6643"
[1] "The variable sf12mcs_dv was removed since it had a VIF score of 11.4437"
[1] "The variable polcost.3 was removed since it had a VIF score of 10.8778"
[1] "The variable cgwrd_dv.6 was removed since it had a VIF score of 10.8563"
[1] "The variable gor_dv.10 was removed since it had a VIF score of 10.6787"
[1] "23 variables removed from the Ordinary Linear Model since they have a VIF score higher than 10"
[1] "The full model AIC after VIF checks is: 11579.5823"
[1] "————Backwards Selection————"
[1] "50 out of 629 variables removed so far."
[1] "100 out of 629 variables removed so far."
[1] "150 out of 629 variables removed so far."
[1] "200 out of 629 variables removed so far."
[1] "250 out of 629 variables removed so far."
[1] "300 out of 629 variables removed so far."
[1] "350 out of 629 variables removed so far."
[1] "400 out of 629 variables removed so far."
[1] "450 out of 629 variables removed so far."
[1] "499 out of 629 variables removed in backwards selection since they weren't significant at the 95
  [1] "finnow.2"        "nasofa.2"        "fimnpen_dv"      "gor_dv.3"
"lfout.5"         "scptrt5e1.6"
  [7] "ienddathh"       "big5e_dv.3"      "adoptch02.4"     "newsource4.1"
"adoptch01.3"     "scptrt5a1.5"
 [13] "orga13.1"        "nurdaym.7"       "buno_dv.2"       "simage.3"
"closenum"        "scptrt5e3.2"
 [19] "strtnurhh"       "paynu_if.1"      "sclfsat7.3"      "poleff4.2"
"cgwrd_dv.1"      "btype96.1"
 [25] "scptrt5o1.3"     "scptrt5o1.4"     "fiyrinvinc_dv"   "scsf3b.3"
"bsoute.4"        "cgwri_dv.1"
 [31] "big5a_dv.4"      "medtyp5.1"       "intdatm_dv.9"    "sf1.y.2"
"poleff2.4"       "cgwri_dv.9"
 [37] "hood15.6"        "polcost.4"       "scopngbhc.4"     "voteintent.7"
"cgsrmem_dv.2"    "visfrnds.2"
 [43] "netpuse.6"       "ficode5.1"       "ag16g10.86"      "marstat_dv.4"
"indns01_lw"      "scptrt5a2.5"
 [49] "ficode6.1"       "scptrt5n2.5"     "orga16.1"        "scsf4b.4"
"sclfsat1.3"      "memprob4.1"
 [55] "agegr10_dv.3"    "medcnj.2"        "hhsize.4"        "ienddatmm"
"lvrel5.1"        "scopngbha.3"
 [61] "vote6.4"         "nurdaym.3"       "crdark.3"        "medtyp12.1"
"scopngbhg.3"     "netpuse.5"
 [67] "beta.1"          "medtyp2.1"       "poleff4.4"       "poleff2.5"
"scopngbhb.4"     "scopngbhc.3"
 [73] "scopngbhh.4"     "ficode19.1"      "agegr10_dv.7"    "scptrt5c3.2"
"mmgssta.4"       "ficode2.3"
 [79] "lvrel9.1"        "scac.2"          "bsoute.2"        "lfout.2"
"orga4.1"         "sampst.2"
 [85] "marstat_dv.2"    "allch01.5"       "big5a_dv.2"      "scptrt5n3.3"
"scptrt5n3.2"     "jbstat.y.8"
 [91] "orga10.1"        "tbmed.2"         "scptrt5e3.6"     "ficode3.2"
"gor_dv.7"        "newsource8.1"
 [97] "newsource3.1"    "newsource2.1"    "lvrel4.1"        "intdatm_dv.4"
"nch5resp.1"      "scopngbhc.2"
[103] "wrdrcl.4"        "wrdrcl.3"        "orga7.1"         "poleff3.5"
```

```
"nch10.1"           "nseqno.4"
[109] "consubx4.1"        "htok.4"             "cgwrd_dv.8"         "cgwrd_dv.5"
"sclfsat7.6"        "big5a_dv.3"
[115] "mnspno.3"          "scopngbhd.2"        "cgvfw_dv.2"         "nbrcoh4.2"
"big5n_dv.3"        "bfpcok.1"
[121] "newsource12.1"     "ficode3.1"          "ficode9.1"          "ficode13.1"
"lfout.8"           "sf1.y.3"
[127] "lvrel6.1"          "scptrt5a1.4"        "locserd.3"          "medtyp9.1"
"nbrcoh2.4"         "bprespc.6"
[133] "ficode39.1"        "voteintent.11"      "perbfts.4"          "lfout.7"
"orga1.1"           "clangab.2"
[139] "statins.2"         "obpdrug.1"          "orga3.1"            "lfout.4"
"ficode12.1"        "simfam.4"
[145] "simfam.2"          "undqus.2"           "orga6.1"            "jbstat.y.5"
"indnsub_xw"        "bmiok.6"
[151] "scptrt5o2.5"       "hhresp_dv.3"        "ficode38.1"         "scptrt5o1.2"
"ivcoop.2"          "big5o_dv.7"
[157] "hood15.4"          "mmgsok.1"           "npns_dv.2"          "nurdayw.5"
"nurdayw.1"         "nurdayw.3"
[163] "nurdayw.2"         "socweb.2"           "lvrel8.1"           "scptrt5e1.4"
"scsf3a.4"          "visfam.5"
[169] "netpuse.3"         "scopngbhc.5"        "big5o_dv.2"         "poleff4.5"
"poleff3.4"         "wtok.4"
[175] "cgvfw_dv.4"        "scptrt5c2.2"        "big5c_dv.7"         "scopngbhb.3"
"newsource5.1"      "aceinh.1"
[181] "calciumb.1"        "mobuse.2"           "vote6.3"            "vote6.2"
"vote1.2"           "intdaty_dv.2012"
[187] "intdatd_dv"        "intdatm_dv.12"      "nurdaym.6"          "scptrt5e3.5"
"ficode4.1"         "cgna_dv.2"
[193] "sppno.4"           "nurdaym.8"          "cgwrd_dv.10"        "scptrt5n3.6"
"scopngbhe.3"       "jbstat.y.3"
[199] "medtyp4.1"         "memaid.3"           "j2pay_if.1"         "orga12.1"
"orga11.1"          "voteintent.2"
[205] "intdatm_dv.6"      "intdatm_dv.5"       "nurdaym.9"          "nurdaym.12"
"scopngbhd.4"       "scptrt5c2.7"
[211] "scptrt5c2.3"       "visfam.3"           "allch03.4"          "lvrel7.1"
"fimnprben_dv"      "marstat_dv.5"
[217] "civicduty.3"       "hhresp_dv.2"        "natch02.6"          "nurdaym.2"
"nurdaym.4"         "bsoute.3"
[223] "lfout.3"           "iron.1"             "ficode37.1"         "newsource96.1"
"scptrt5a1.7"       "civicduty.2"
[229] "scptrt5n2.6"       "cgwri_dv.6"         "cgwri_dv.8"         "cgwri_dv.2"
"scptrt5o2.2"       "scsf5.3"
[235] "scsf5.4"           "newsource1.1"       "scptrt5o3.3"        "sclfsat7.2"
"newsource11.1"     "netpuse.4"
[241] "urban_dv.y.2"      "seearngrs_if.1"     "ficode29.2"         "cgwrd_dv.2"
"hrpno.5"           "nseqno.5"
[247] "nnsib_dv.3"        "allch01.4"          "nch10to15.1"        "nch5to15.2"
"intdatm_dv.10"     "big5a_dv.7"
[253] "hood15.2"          "polcost.2"          "nseqno.2"           "scptrt5a3.6"
"cogdist.2"         "fiyrinvinc_if.1"
[259] "fnspno.3"          "poleff1.3"          "poleff1.4"          "cgsrmem_dv.5"
"depchl_dv.2"       "poleff3.3"
[265] "scptrt5o3.7"       "big5o_dv.4"         "memprob2.1"         "scptrt5c1.2"
"strtnurmm"         "crdark.2"
[271] "sppno.2"           "intdatm_dv.2"       "consubx3.1"         "gor_dv.8"
"scopngbhh.3"       "scopngbhh.2"
[277] "cgwri_dv.10"       "ficode28.1"         "diur.1"             "sppno.3"
"natch01.1"         "visfam.6"
[283] "scsf5.5"           "tvhours"            "j2pay_dv"           "ficode24.1"
"clangab.3"         "newsource6.1"
[289] "cgwri_dv.4"        "nurdayw.6"          "istrtdatmm"         "nurdayw.4"
"ivcoop.3"          "susp.3"
[295] "gor_dv.11"         "cgna_dv.4"          "indbd91_lw"         "netpuse.2"
"scsf5.2"           "mmgssta.2"
[301] "nacar.2"           "visfam.2"           "nch10to15.3"        "nch5to15.3"
"poleff4.3"         "voteintent.8"
[307] "voteintent.10"     "perbfts.5"          "cgvfc_dv"           "cgwri_dv.3"
```

```
"cgwri_dv.5"        "mmgsdom.2"
[313] "newsource9.1"      "hrpno.3"           "scptrt5n3.7"       "scptrt5a2.6"
"nchund18resp.4"    "allch02.6"
[319] "voteintent.3"      "natch04.5"         "crdark.4"          "hrpno.2"
"medtyp11.1"        "bsoute.5"
[325] "orga15.1"          "sex_dv.2"          "voteintent.6"      "clangab.4"
"undqus.4"          "readtest.2"
[331] "medtyp3.1"         "lfout.11"          "nsran.2"           "lkmove.2"
"nbrcoh2.3"         "scptrt5a1.6"
[337] "cgwrd_dv.4"        "cgs7n_dv.4"        "btype7.1"          "jbstat.y.6"
"cgwrd_dv.9"        "scptrt5e3.3"
[343] "scptrt5c3.3"       "fimnmisc_dv"       "cgivna1_dv.4"      "simage.2"
"simage.4"          "bprespc.4"
[349] "ficode17.1"        "scopngbhe.4"       "ffbrfedlw.1"       "ethn_dv"
"ficode8.1"         "scsf2b.2"
[355] "gor_dv.4"          "big5c_dv.3"        "crdark.5"          "poleff1.2"
"locserd.4"         "scsf6b.3"
[361] "scptrt5e1.3"       "poleff2.2"         "scopngbhg.5"       "sclfsat7.7"
"scptrt5o3.2"       "nurdaym.5"
[367] "newsource7.1"      "newsource10.1"     "istrtdatss"        "scptrt5a1.3"
"scptrt5a1.2"       "agegr10_dv.6"
[373] "big5a_dv.6"        "scptrt5a3.5"       "scptrt5a3.2"       "scptrt5a3.3"
"qfhighfl_dv.1"     "hhsize.5"
[379] "npensioner_dv.1" "npensioner_dv.2" "ficode20.1"        "nbrcoh3.5"
"nbrcoh1.5"         "scopngbhh.5"
[385] "voteintent.9"      "scptrt5o2.3"       "nch10to15.2"       "cgs7n_dv.2"
"cgs7n_dv.5"        "nurdaym.11"
[391] "scptrt5e1.5"       "intdatm_dv.3"      "ficode27.1"        "wtok.2"
"ficode7.1"         "nbrcoh1.4"
[397] "medtyp10.1"        "ficode10.1"        "scsf2a.2"          "scptrt5e3.4"
"scptrt5e1.2"       "nurdaym.10"
[403] "scptrt5o2.4"       "big5o_dv.3"        "nchild_dv.1"       "bensta1.1"
"polcost.5"         "scptrt5c1.3"
[409] "nbrcoh1.3"         "gor_dv.9"          "servacc.2"         "allch01.6"
"hhsize.6"          "allch03.6"
[415] "nurdayd"           "hrpno.4"           "natch03.6"         "scptrt5e3.7"
"nbrcoh3.4"         "nbrcoh4.5"
[421] "nbrcoh4.4"         "lipid.1"           "undqus.3"          "hood15.3"
"scptrt5o3.4"       "scptrt5o3.6"
[427] "orga8.1"           "bmiok.5"           "cgna_dv.1"         "fnpno.2"
"nbrcoh3.3"         "big5o_dv.6"
[433] "orga96.1"          "scopngbhd.5"       "sclfsat2.7"        "sclfsat2.4"
"sclfsat2.5"        "sclfsat2.3"
[439] "scopngbhf.4"       "scopngbhf.3"       "scopngbhf.2"       "scopngbhf.5"
"fimnlabnet_tc.1" "scopngbhd.3"
[445] "cgsrmem2_dv.1"     "scptrt5a3.4"       "orga5.1"           "orga14.1"
"llknbrd.2"         "sclfsato.4"
[451] "lvrel10.1"         "scptrt5n1.2"       "scptrt5n1.3"       "scptrt5n1.4"
"sclfsat7.5"        "scsf4b.3"
[457] "cgna_dv.5"         "scptrt5o2.6"       "sclfsat1.4"        "hhsize.7"
"cgs7n_dv.3"        "cgivna1_dv.3"
[463] "fimngrs_if"        "medtyp7.1"         "cgwrd_dv.3"        "civicduty.4"
"bprespc.2"         "npns_dv.1"
[469] "pns1pno.2"         "pn2pno.3"          "sclfsat1.6"        "sclfsat1.2"
"gor_dv.2"          "big5c_dv.5"
[475] "big5c_dv.4"        "scptrt5c2.5"       "big5c_dv.2"        "scopngbhe.5"
"scopngbhb.5"       "scptrt5a2.4"
[481] "adoptch01.2"       "scptrt5a2.2"       "scptrt5o1.5"       "scptrt5n1.5"
"sclfsat1.5"        "orga9.1"
[487] "netpuse.7"         "ficode29.1"        "respf16_dv.2"      "fimnlabnet_dv"
"big5e_dv.4"        "medtyp6.1"
[493] "natch01.2"         "nnsib_dv.2"        "ficode33.1"        "simage.5"
"scptrt5c1.4"       "ficode2.2"
[499] "sclfsat1.7"
[1] "——————Ordinary Linear Regression (Improved)——————"

Call:
lm(formula = y ~ ., data = as.data.frame(x.data.linear))
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-10.274  -1.843  -0.087   1.639  15.133

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     9.28861    0.45557  20.389  < 2e-16 ***
sampst.3        1.02658    0.27239   3.769 0.000169 ***
wrdrcl.2       -0.55712    0.16505  -3.375 0.000751 ***
nch3resp.1     -1.31803    0.46047  -2.862 0.004249 **
nch8resp.1     -1.54074    0.41713  -3.694 0.000227 ***
natch01.4      -1.84529    0.70787  -2.607 0.009206 **
natch02.5       4.65850    1.16588   3.996 6.68e-05 ***
adoptch01.4    12.21450    3.39957   3.593 0.000335 ***
xpmove.2       -0.68289    0.25504  -2.678 0.007477 **
nbrcoh2.5      -3.32327    0.86569  -3.839 0.000127 ***
crwora.2       -0.41694    0.15523  -2.686 0.007292 **
hood15.5        0.52080    0.24049   2.166 0.030464 *
simfam.5       -0.80957    0.40550  -1.996 0.046018 *
orga2.1         1.05741    0.50425   2.097 0.036117 *
aidxhh.2       -0.46598    0.20615  -2.260 0.023906 *
finnow.3        0.47762    0.18247   2.618 0.008924 **
finnow.4        1.70388    0.30928   5.509 4.07e-08 ***
finnow.5        2.83239    0.47084   6.016 2.12e-09 ***
perbfts.3       0.43747    0.16884   2.591 0.009637 **
perbfts.6       2.41110    0.78824   3.059 0.002251 **
voteintent.1    1.04974    0.53023   1.980 0.047863 *
voteintent.4    1.53827    0.65362   2.353 0.018695 *
voteintent.5    0.81599    0.31112   2.623 0.008789 **
scac.3         -4.89557    1.37507  -3.560 0.000379 ***
scsf1.5         1.21432    0.37227   3.262 0.001125 **
scsf3b.2       -1.15404    0.31190  -3.700 0.000221 ***
scsf4a.2        2.49281    0.42130   5.917 3.85e-09 ***
scsf4a.3        1.79666    0.27094   6.631 4.27e-11 ***
scsf4a.4        1.04508    0.21057   4.963 7.53e-07 ***
scsf4b.2        0.92744    0.39956   2.321 0.020379 *
scsf6a.3        0.96942    0.19158   5.060 4.57e-07 ***
scsf6a.4        2.86873    0.26892  10.667  < 2e-16 ***
scsf6a.5        4.96666    0.43800  11.339  < 2e-16 ***
scsf6b.4        0.96162    0.22533   4.268 2.07e-05 ***
scsf6b.5        2.17220    0.35264   6.160 8.78e-10 ***
scsf6c.2        4.14856    0.44988   9.222  < 2e-16 ***
scsf6c.3        2.23408    0.25191   8.868  < 2e-16 ***
scsf6c.4        0.75526    0.19145   3.945 8.26e-05 ***
scsf7.2         2.40233    0.40662   5.908 4.06e-09 ***
scsf7.3         1.12649    0.25821   4.363 1.35e-05 ***
scsf7.4         0.55510    0.21287   2.608 0.009183 **
scopngbha.4     1.05220    0.36920   2.850 0.004417 **
scopngbha.5     2.55955    0.77930   3.284 0.001039 **
scopngbhg.4    -0.81471    0.27819  -2.929 0.003443 **
sclfsat2.2     -0.69531    0.23374  -2.975 0.002967 **
sclfsat2.6     -0.40157    0.18237  -2.202 0.027784 *
sclfsat7.4      0.53492    0.24254   2.205 0.027533 *
sclfsato.2      1.04015    0.30182   3.446 0.000580 ***
sclfsato.3      1.21892    0.31271   3.898 0.000100 ***
sclfsato.5      0.52752    0.20684   2.550 0.010833 *
sclfsato.7     -0.91429    0.25746  -3.551 0.000392 ***
scptrt5c1.5    -0.48562    0.22606  -2.148 0.031815 *
scptrt5c1.6    -0.41310    0.17989  -2.296 0.021754 *
scptrt5n1.6     0.85407    0.29306   2.914 0.003605 **
scptrt5n1.7     1.14144    0.34884   3.272 0.001085 **
scptrt5o1.6    -0.44203    0.21865  -2.022 0.043348 *
scptrt5a2.3     0.72317    0.28252   2.560 0.010549 *
scptrt5c2.4    -0.46136    0.23222  -1.987 0.047093 *
scptrt5c2.6    -0.87693    0.37165  -2.360 0.018392 *
scptrt5e2.2    -1.04911    0.38712  -2.710 0.006785 **
scptrt5e2.3    -0.72606    0.31407  -2.312 0.020891 *
```

```
scptrt5e2.4   -1.02143    0.28611  -3.570 0.000365 ***
scptrt5e2.5   -0.70827    0.26472  -2.676 0.007522 **
scptrt5e2.6   -0.64445    0.24503  -2.630 0.008602 **
scptrt5n2.2    0.61519    0.20742   2.966 0.003054 **
scptrt5n2.3    0.96357    0.22834   4.220 2.55e-05 ***
scptrt5n2.4    0.85205    0.23212   3.671 0.000248 ***
scptrt5c3.4    0.88683    0.27482   3.227 0.001271 **
scptrt5c3.5    0.77032    0.22960   3.355 0.000808 ***
scptrt5c3.6    0.52655    0.20979   2.510 0.012157 *
scptrt5n3.4   -0.62904    0.18912  -3.326 0.000896 ***
susp.2        -1.45665    0.68744  -2.119 0.034218 *
origadd.2     -0.76069    0.28273  -2.691 0.007192 **
ienddatss     -0.20187    0.07262  -2.780 0.005487 **
agegr13_dv.3   1.69114    0.72938   2.319 0.020517 *
ppno.5         3.66425    1.66984   2.194 0.028323 *
big5n_dv.2    -0.60596    0.21216  -2.856 0.004333 **
big5n_dv.5     1.14605    0.25371   4.517 6.63e-06 ***
big5n_dv.6     2.08117    0.41719   4.989 6.61e-07 ***
big5n_dv.7     3.22042    0.56024   5.748 1.04e-08 ***
cgwrd_dv.7    -0.74686    0.21571  -3.462 0.000547 ***
cgvfw_dv.1    -0.46003    0.22232  -2.069 0.038658 *
cgvfw_dv.3     1.82678    0.84245   2.168 0.030244 *
ficode4.2      1.92982    0.81323   2.373 0.017737 *
ficode11.1     1.36665    0.48532   2.816 0.004910 **
ficode15.1     1.16893    0.41494   2.817 0.004894 **
ficode23.1    -0.93635    0.22933  -4.083 4.62e-05 ***
medtyp8.1      1.42208    0.65567   2.169 0.030208 *
cufsize.3      0.89379    0.32678   2.735 0.006291 **
airtemp       -0.15798    0.07607  -2.077 0.037941 *
mmgssta.3      1.70984    0.74010   2.310 0.020973 *
htok.2        -2.16679    0.90238  -2.401 0.016432 *
wtok.3        -3.36093    1.27383  -2.638 0.008393 **
lfout.9        1.55628    0.51650   3.013 0.002618 **
wstokb.9       1.43575    0.60419   2.376 0.017580 *
hhsize.3       0.40236    0.19709   2.042 0.041326 *
jbstat.y.97   -2.65980    0.93775  -2.836 0.004609 **
sf1.y.4       -0.52089    0.20510  -2.540 0.011170 *
gor_dv.5      -0.72965    0.32891  -2.218 0.026641 *
gor_dv.6      -0.66278    0.27727  -2.390 0.016925 *
big5e_dv.2     1.12979    0.40512   2.789 0.005341 **
big5e_dv.6    -0.60850    0.21904  -2.778 0.005520 **
big5e_dv.7    -0.73204    0.31042  -2.358 0.018456 *
agegr10_dv.5   0.57006    0.19838   2.874 0.004101 **
intdatm_dv.7  -0.93915    0.25453  -3.690 0.000230 ***
intdatm_dv.8  -1.03386    0.27939  -3.700 0.000221 ***
cgivna1_dv.2  -1.56082    0.59651  -2.617 0.008948 **
nnsib_dv.1    -2.28601    0.58146  -3.932 8.73e-05 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1    1

Residual standard error: 3.265 on 2006 degrees of freedom
Multiple R-squared:  0.6865,    Adjusted R-squared:  0.6698
F-statistic: 41.05 on 107 and 2006 DF,  p-value: < 2.2e-16

AIC: 11109.2295
MSE: 10.1156
[1] "The MSE of the predicted values are of 15.8178"
[1] "The Linear Model predicts exactly with accuracy of 0.1335"
[1] "The Linear Model predicts within a confidence interval with accuracy of 0.4347"
[1] "----------Elastic Net Regression----------"
630 x 1 sparse Matrix of class "dgCMatrix", with 17 entries
        names Estimate_Coefs
1 (Intercept)    11.28002446
2    finnow.4     0.73553433
3    finnow.5     0.92617519
4     scsf1.5     0.58452450
5    scsf6a.4     0.18971907
6    scsf6c.2     0.73423146
```

```
7    sclfsato.3      0.28129089
8    sclfsato.6     −0.70826597
9    sclfsato.7     −0.76404977
10  scptrt5n1.6      0.51871227
11  scptrt5n1.7      1.57517539
12   sf12mcs_dv     −3.48322620
13   big5n_dv.2     −0.10739561
14   big5n_dv.5      0.05320476
15   big5n_dv.7      0.19733581
16   sf12pcs_dv     −0.63709686
17   big5e_dv.2      0.04171344
[1] "The MSE of the predicted values of the best fit model is 11.0942"
[1] "The Alpha of the best fit model is 1"
[1] "The Elastic Net Model predicts exactly with accuracy of 0.1705"
[1] "————Timer Results————"
   user  system elapsed
 907.82   10.61  918.73
```

10.2.23 wShared console

```
[1] "————Initial Checks————"
[1] "83670404 NA cells were found across the entire dataset (68.36% of data as NA)"
[1] "————Data Type Checks————"
[1] "0 variables recoded since all their entries aren't numeric or NA"
[1] "NOTE: algorithim recodes categorical data alphabetically e.g (female = 1, male = 2)"
character(0)
[1] "————Low Data Removal————"
[1] "862 variables removed since they had >= 'naPercent' (default 20%) NA values"
  [1] "pid"          "lvwhy"        "lvmthp"       "lvyrp"        "mvever"
"mvmnth"       "mvyr"         "mlstatchk"
  [9] "mlstat"       "drive"        "caruse"       "netuse"       "ukborn"
"plbornc"      "yr2uk4"       "citzn1"
 [17] "citzn2"       "citzn3"       "qfhigh"       "qualoc"       "qfvoc1"
"qfvoc2"       "qfvoc3"       "qfvoc4"
 [25] "qfvoc5"       "qfvoc6"       "qfvoc7"       "qfvoc8"       "qfvoc9"
"qfvoc10"      "qfvoc11"      "qfvoc12"
 [33] "qfvoc13"      "qfvoc14"      "qfvoc15"      "qfvoc96"      "school"
"scend"        "schlloc"      "schok"
 [41] "fenow"        "feend"        "j1none"       "j1semp"       "j1boss"
"j1mngr"       "edtype"       "edasp"
 [49] "fedlik"       "fednt"        "ocimpa"       "ocimpb"       "ocimpe"
"ocimpf"       "ocimpi"       "ocimpk"
 [57] "ocimpl"       "futra"        "futrb"        "futrc"        "futrd"
"futre"        "futrf"        "futrg"
 [65] "futrh"        "futri"        "futrj"        "futrk"        "futrl"
"paju"         "maju"         "pacob"
 [73] "payruk"       "macob"        "mayruk"       "natid1"       "natid2"
"natid3"       "natid4"       "natid5"
 [81] "natid6"       "natid97"      "racel"        "racelo_code"  "oprlg"
"oprlg0ni"     "nirel"        "niact"
 [89] "oprlg0"       "oprlg1"       "hospc1"       "hospdc1"      "hospc2"
"hospdc2"      "hospc3"       "hospdc3"
 [97] "hospc4"       "hospdc4"      "hospc5"       "hospdc5"      "hospc6"
"hospdc6"      "hospc7"       "hospdc7"
[105] "hospc8"       "hospdc8"      "disdif1"      "disdif2"      "disdif3"
"disdif4"      "disdif5"      "disdif6"
[113] "disdif7"      "disdif8"      "disdif9"      "disdif10"     "disdif11"
"disdif12"     "disdif96"     "aidhua1"
[121] "aidhua2"      "aidhua3"      "aidhua4"      "aidhua5"      "aidhua6"
"aidhua7"      "aidhua8"      "aidhua9"
[129] "aidhua10"     "aidhua11"     "aidhua12"     "aidhua13"     "aidhua14"
"aidhua15"     "aidhua16"     "naidxhh"
[137] "aidhu1"       "aidhu2"       "aidhrs"       "aideft"       "lcohnpi"
"coh1bm"       "coh1by"       "coh1mr"
[145] "coh1em"       "coh1ey"       "nmar"         "lmar1m"       "lmar1y"
"ladopt"       "lnadopt"      "lprnt"
```

```
[153] "lnprnt"        "ch1by4"        "movy11"        "movy12"        "movy13"
"movy14"        "movy15"        "movy16"
[161] "family"        "education"     "memploy"       "housing"       "area"
"moveoth_code"  "movdir"        "plnowm"
[169] "plnowy4"       "mstatsam"      "lwwrong"       "mstatsamn"     "mstatch1"
"statcm1"       "statcy41"      "divchk1"
[177] "divfin1"       "dvm1"          "dvy41"         "cmlstat1"      "mstatch2"
"statcm2"       "statcy42"      "divchk2"
[185] "divfin2"       "dvm2"          "dvy42"         "cmlstat2"      "cohab"
"cohabn"        "lmcbm1"        "lmcby41"
[193] "currpart1"     "lmspm1"        "lmspy41"       "lmcbm2"        "lmcby42"
"currpart2"     "lmspm2"        "lmspy42"
[201] "lmcbm3"        "lmcby43"       "currpart3"     "lmspm3"        "lmspy43"
"father"        "nchild"        "preg"
[209] "pregm1"        "pregy41"       "pregfert1"     "invitro1"      "pregout1"
"pregend1"      "endmnth1"      "pregsmoke1"
[217] "smkmnth11"     "smkmnth21"     "smkmnth31"     "pregsmk11"     "pregsmk21"
"pregsmk31"     "lchmulti1"     "pregm2"
[225] "pregy42"       "pregfert2"     "invitro2"      "pregout2"      "pregend2"
"endmnth2"      "pregsmoke2"    "smkmnth12"
[233] "smkmnth22"     "smkmnth32"     "pregsmk12"     "pregsmk22"     "pregsmk32"
"lchmulti2"     "pregm3"        "pregy43"
[241] "pregfert3"     "invitro3"      "pregout3"      "pregend3"      "endmnth3"
"pregsmoke3"    "smkmnth13"     "smkmnth23"
[249] "smkmnth33"     "pregsmk13"     "pregsmk23"     "pregsmk33"     "lchmulti3"
"hcondno1"      "hcondns1"      "hcondno2"
[257] "hcondns2"      "hcondno3"      "hcondns3"      "hcondno4"      "hcondns4"
"hcondno5"      "hcondns5"      "hcondno6"
[265] "hcondns6"      "hcondno7"      "hcondns7"      "hcondno8"      "hcondns8"
"contft"        "ftendm"        "ftendy4"
[273] "ftquals"       "ftedstartm1"   "ftedstarty41"  "ftedend1"      "ft2endm1"
"ft2endy41"     "ftedmor1"      "ftedstartm2"
[281] "ftedstarty42"  "ftedend2"      "ft2endm2"      "ft2endy42"     "ftedmor2"
"qualnew1"      "qualnew2"      "qualnew3"
[289] "qualnew4"      "qualnew5"      "qualnew6"      "qualnew7"      "qualnew8"
"qualnew9"      "qualnew10"     "qualnew11"
[297] "qualnew13"     "qualnew15"     "qualnew16"     "qualnew17"     "qualnew18"
"qualnew19"     "qualnew20"     "qualnew21"
[305] "qualnew22"     "qualnew23"     "qualnew24"     "qualnew25"     "qualnew26"
"qualnew27"     "qualnew28"     "qualnew29"
[313] "qualnew30"     "qualnew31"     "trwho1"        "traindays1"    "trainhrs1"
"trainend1"     "trainpurp11"   "trainpurp21"
[321] "trainpurp31"   "trainpurp41"   "trainpurp51"   "trainpurp61"   "trainpurp71"
"trainqual1"    "trwho2"        "traindays2"
[329] "trainhrs2"     "trainend2"     "trainpurp12"   "trainpurp22"   "trainpurp32"
"trainpurp42"   "trainpurp52"   "trainpurp62"
[337] "trainpurp72"   "trainqual2"    "trwho3"        "traindays3"    "trainhrs3"
"trainend3"     "trainpurp13"   "trainpurp23"
[345] "trainpurp33"   "trainpurp43"   "trainpurp53"   "trainpurp63"   "trainpurp73"
"trainqual3"    "trainn"        "trqual1"
[353] "trqual2"       "trqual3"       "trqual4"       "trqual5"       "trqual6"
"trqual7"       "trqual8"       "trqual9"
[361] "trqual10"      "trqual11"      "trqual13"      "trqual15"      "trqual16"
"trqual17"      "trqual18"      "trqual19"
[369] "trqual20"      "trqual21"      "trqual22"      "trqual23"      "trqual24"
"trqual25"      "trqual26"      "trqual27"
[377] "trqual28"      "trqual29"      "trqual30"      "trqual31"      "notempchk"
"empchk"        "empstendd"     "empstendm"
[385] "empstendy4"    "stendreas"     "nxtst"         "nxtstelse"     "cstat"
"nxtstendd"     "nxtstendm"     "nxtstendy4"
[393] "jbsamr"        "wkplsam"       "samejob"       "matlv"         "matlvstd"
"matlvstm"      "matlvsty4"     "matlvendd"
[401] "matlvendm"     "matlvendy4"    "jbendd"        "jbendm"        "jbendy4"
"jbendreas"     "cjob"          "nxtjbhrs"
[409] "nxtjbes"       "nxtjbendd"     "nxtjbendm"     "nxtjbendy4"    "cjbatt"
"nextstat1"     "nextelse1"     "currstat1"
[417] "nextjob1"      "currjob1"      "jobhours1"     "reasend1"      "jbatt1"
"statendd1"     "statendm1"     "statendy41"
```

```
[425] "nextstat2"    "nextelse2"    "currstat2"    "nextjob2"     "currjob2"
"jobhours2"    "reasend2"     "jbatt2"
[433] "statendd2"    "statendm2"    "statendy42"   "nextstat3"    "nextelse3"
"currstat3"    "nextjob3"     "currjob3"
[441] "jobhours3"    "reasend3"     "jbatt3"       "statendd3"    "statendm3"
"statendy43"   "nextstat4"    "nextelse4"
[449] "currstat4"    "nextjob4"     "currjob4"     "jobhours4"    "reasend4"
"jbatt4"       "statendd4"    "statendm4"
[457] "statendy44"   "nextstat5"    "nextelse5"    "currstat5"    "nextjob5"
"currjob5"     "jobhours5"    "reasend5"
[465] "jbatt5"       "statendd5"    "statendm5"    "statendy45"   "nextstat6"
"nextelse6"    "currstat6"    "nextjob6"
[473] "currjob6"     "jobhours6"    "reasend6"     "jbatt6"       "statendd6"
"statendm6"    "statendy46"   "nextstat7"
[481] "nextelse7"    "currstat7"    "nextjob7"     "currjob7"     "jobhours7"
"reasend7"     "jbatt7"       "statendd7"
[489] "statendm7"    "statendy47"   "nextstat8"    "nextelse8"    "currstat8"
"nextjob8"     "currjob8"     "jobhours8"
[497] "reasend8"     "jbatt8"       "statendd8"    "statendm8"    "statendy48"
"nextstat9"    "nextelse9"    "currstat9"
[505] "nextjob9"     "currjob9"     "jobhours9"    "reasend9"     "jbatt9"
"statendd9"    "statendm9"    "statendy49"
[513] "jboff"        "jboffy"       "jbterm1"      "jbterm2"      "jbsic07chk"
"jbsoc00chk"   "jbsempchk"    "jbsemp"
[521] "jbbgd"        "jbbgm"        "jbbgy"        "jbmngrchk"    "jbmngr"
"jbsizechk"    "jbsize"       "jbsect"
[529] "jbsectpub"    "jbhrs"        "jbot"         "jbotpd"       "paygwc"
"paynwc"       "payusl"       "payu"
[537] "payuwc"       "payug"        "paytyp"       "ovtpay"       "extnsa"
"extrate"      "extrest"      "basnsa"
[545] "basrate"      "basrest"      "ovtnsa"       "ovtrate"      "ovtrest"
"jbpl"         "jbttwt"       "worktrav"
[553] "jsboss"       "jssize"       "jshrs"        "jstypeb"      "jsaccs"
"jspart"       "jsprbm"       "jsprby4"
[561] "jsprem"       "jsprey4"      "jsprls"       "jsprtx"       "jsprni"
"jspayu"       "jspayw"       "jspytx"
[569] "jspyni"       "jspl"         "jsttwt"       "jsttwtb"      "jsworktrav"
"jbsat"        "julk4wk"      "julkjb"
[577] "jubgn"        "julk4x1"      "julk4x2"      "julk4x3"      "julk4x4"
"julk4x5"      "julk4x6"      "julk4x96"
[585] "jbhad"        "jlendm"       "jlendy"       "jlsemp"       "jlboss"
"jlmngr"       "jlsize"       "eprosh"
[593] "j2semp"       "j2hrs"        "j2pay"        "retchk"       "ageret"
"rtpro1"       "rtpro2"       "rtpro3"
[601] "rtpro4"       "rtpro5"       "rtpro6"       "rtcon1"       "rtcon2"
"rtcon3"       "rtcon4"       "penmex"
[609] "pppex"        "pppexm"       "sppen"        "rtexpjb"      "rtfnd1"
"rtfnd2"       "rtfnd3"       "rtfnd4"
[617] "rtfnd5"       "rtfnd6"       "rtfnd7"       "rtfnd8"       "rtfnd9"
"rtfnd10"      "rtfnd96"      "retamt"
[625] "retsuf"       "ccare"        "ccwork"       "benunemp1"    "benunemp2"
"benunemp96"   "bendis1"      "bendis2"
[633] "bendis3"      "bendis4"      "bendis5"      "bendis6"      "bendis7"
"bendis8"      "bendis9"      "bendis10"
[641] "bendis96"     "benpen1"      "benpen2"      "benpen3"      "benpen4"
"benpen5"      "benpen6"      "benpen7"
[649] "benpen8"      "benpen96"     "niserps"      "bencb"        "benctc"
"benfam1"      "benfam2"      "benfam3"
[657] "benfam4"      "benfam5"      "benfam96"     "bentax1"      "bentax2"
"bentax3"      "bentax4"      "bentax5"
[665] "bentax96"     "benhou1"      "benhou2"      "benhou3"      "benhou4"
"benhou96"     "nfh01"        "nfh02"
[673] "nfh03"        "nfh04"        "nfh05"        "nfh06"        "nfh07"
"nfh08"        "nfh09"        "nfh10"
[681] "nfh11"        "nfh12"        "nfh13"        "nfh14"        "nfh15"
"nfh16"        "nfh17"        "nfh18"
[689] "nfh19"        "nfh20"        "nfh21"        "nfh22"        "nfh23"
"nfh24"        "nfh25"        "nfh26"
```

```
[697] "nfh27"          "nfh28"          "nfh29"          "nfh30"          "nfh31"
"nfh32"          "nfh33"          "nfh34"
[705] "nfh35"          "nfh36"          "nfh37"          "nfh38"          "fiyrdb1"
"fiyrdb2"        "fiyrdb3"        "fiyrdb4"
[713] "fiyrdb5"        "fiyrdb6"        "vote2"          "vote3"          "vote4"
"vote5"          "perpolinf"      "colbens1"
[721] "colbens2"       "colbens3"       "civicduty"      "polcost"        "votenorm"
"perbfts"        "grpbfts"        "voteintent"
[729] "demorient"      "rphmob_code"    "rphwrk_code"    "remail_code"    "ctadd1_code"
"ctadd2_code"    "cttown_code"    "ctcnty_code"
[737] "ctpcode_code"   "cttel1_code"    "cttel2_code"    "ctemail_code"   "livesp"
"livewith"       "lingua"         "jbbgdatd"
[745] "jbbgdatm"       "jbbgdaty"       "prel"           "preason"        "pripn"
"pjulk4wk"       "pjbptft"        "pjsptft"
[753] "prearn"         "pbnft1"         "pbnft2"         "pbnft3"         "pbnft4"
"pbnft5"         "pbnft6"         "pbnft7"
[761] "pbnft8"         "pbnft9"         "pbnft10"        "pbnft11"        "pbnft12"
"pbnft96"        "prfitb"         "ivinfnce"
[769] "ivaffct11"      "ivaffct12"      "ivaffct13"      "ivaffct14"      "ivaffct15"
"ivaffct17"      "ivaffct21"      "ivaffct22"
[777] "ivaffct23"      "ivaffct27"      "ivaffct30"      "ivaffct97"      "hgadoptm"
"hgadoptf"       "hgpart"         "ppsex"
[785] "fnpid"          "mnpid"          "pn1pid"         "pn1sex"         "pn2pid"
"pn2sex"         "pns1pid"        "pns1sex"
[793] "pns2pid"        "pns2sex"        "grfpid"         "grmpid"         "paygl"
"paynl"          "jsprf"          "payg_dv"
[801] "payn_dv"        "seearnnet_dv"   "ff_tel"         "ff_jbsemp"      "ff_jbmngr"
"ff_jbsize"      "ff_oprlg"       "ff_oprlg0"
[809] "ff_oprlg0ni"    "ff_ukborn"      "ff_yr2uk4"      "j1nssec8_dv"    "manssec8_dv"
"adresp15_dv"    "ppid"           "sppid"
[817] "fnspid"         "mnspid"         "nqfhigh_dv"     "jbsoc00_cc"     "jbsoc10_cc"
"jbsic07_cc"     "jbes2000"       "jbrgsc_dv"
[825] "jbnssec8_dv"    "jbisco88_cc"    "jlsoc00_cc"     "jlsoc10_cc"     "jlsic07_cc"
"jles2000"       "jlrgsc_dv"      "jlnssec8_dv"
[833] "jlisco88_cc"    "pasoc90_cc"     "pasoc00_cc"     "pasoc10_cc"     "masoc90_cc"
"masoc00_cc"     "masoc10_cc"     "j1soc90_cc"
[841] "j1soc00_cc"     "j1soc10_cc"     "j2soc90_cc"     "j2soc00_cc"     "j2soc10_cc"
"payu_dv"        "jlnssec3_dv"    "paygu_dv"
[849] "seearngrs_dv"   "panssec8_dv"    "nhiqual_dv"     "jbft_dv"        "jbseg_dv"
"jbnssec5_dv"    "jlseg_dv"       "jlnssec5_dv"
[857] "paynu_dv"       "j2nssec8_dv"    "jliindb_dv"     "jbnssec_dv"     "jbnssec3_dv"
"jlnssec_dv"
[1] "————————Low Level Removal————————"
[1] "If a level is removed from a variable you wish to keep, reccomended to manually merge levels to
[1] "level 8 in pno removed, 2 observations found"
[1] "level 1 in adstatus removed, 1 observations found"
[1] "level 7 in nch14resp removed, 2 observations found"
[1] "level 6 in nch415resp removed, 3 observations found"
[1] "level 7 in nch415resp removed, 0 observations found"
[1] "level 7 in nchresp removed, 3 observations found"
[1] "level 8 in nchresp removed, 0 observations found"
[1] "level 7 in nchund18resp removed, 0 observations found"
[1] "level 8 in nchund18resp removed, 0 observations found"
[1] "level 10 in natch01 removed, 2 observations found"
[1] "level 11 in natch01 removed, 2 observations found"
[1] "level 12 in natch01 removed, 1 observations found"
[1] "level 10 in natch02 removed, 3 observations found"
[1] "level 11 in natch02 removed, 0 observations found"
[1] "level 13 in natch02 removed, 0 observations found"
[1] "level 14 in natch02 removed, 0 observations found"
[1] "level 9 in natch03 removed, 2 observations found"
[1] "level 10 in natch03 removed, 4 observations found"
[1] "level 12 in natch03 removed, 0 observations found"
[1] "level 15 in natch03 removed, 0 observations found"
[1] "level 8 in natch04 removed, 1 observations found"
[1] "level 9 in natch04 removed, 1 observations found"
[1] "level 10 in natch04 removed, 0 observations found"
[1] "level 5 in natch05 removed, 3 observations found"
```

```
[1] "level 9 in natch05 removed, 1 observations found"
[1] "level 7 in natch06 removed, 2 observations found"
[1] "level 9 in natch06 removed, 1 observations found"
[1] "level 8 in natch07 removed, 0 observations found"
[1] "level 10 in natch08 removed, 2 observations found"
[1] "level 11 in natch09 removed, 0 observations found"
[1] "level 12 in natch10 removed, 0 observations found"
[1] "level 8 in nnatch removed, 0 observations found"
[1] "level 10 in nnatch removed, 0 observations found"
[1] "level 5 in nadoptch removed, 3 observations found"
[1] "level 7 in adoptch02 removed, 4 observations found"
[1] "level 4 in adoptch03 removed, 4 observations found"
[1] "level 7 in adoptch03 removed, 1 observations found"
[1] "level 8 in adoptch03 removed, 0 observations found"
[1] "level 7 in adoptch04 removed, 2 observations found"
[1] "level 7 in adoptch05 removed, 0 observations found"
[1] "level 7 in nchunder16 removed, 1 observations found"
[1] "level 6 in nch5to15 removed, 0 observations found"
[1] "level 5 in nch10to15 removed, 0 observations found"
[1] "level 8 in allch01 removed, 4 observations found"
[1] "level 9 in allch01 removed, 3 observations found"
[1] "level 10 in allch01 removed, 0 observations found"
[1] "level 11 in allch01 removed, 0 observations found"
[1] "level 12 in allch01 removed, 0 observations found"
[1] "level 2 in allch02 removed, 2 observations found"
[1] "level 9 in allch02 removed, 0 observations found"
[1] "level 10 in allch02 removed, 0 observations found"
[1] "level 11 in allch02 removed, 0 observations found"
[1] "level 13 in allch02 removed, 0 observations found"
[1] "level 14 in allch02 removed, 0 observations found"
[1] "level 9 in allch03 removed, 3 observations found"
[1] "level 10 in allch03 removed, 0 observations found"
[1] "level 12 in allch03 removed, 0 observations found"
[1] "level 15 in allch03 removed, 0 observations found"
[1] "level 9 in allch04 removed, 2 observations found"
[1] "level 10 in allch04 removed, 0 observations found"
[1] "level 9 in allch05 removed, 2 observations found"
[1] "level 11 in allch05 removed, 0 observations found"
[1] "level 7 in allch06 removed, 0 observations found"
[1] "level 9 in allch06 removed, 0 observations found"
[1] "level 12 in allch06 removed, 0 observations found"
[1] "level 9 in allch07 removed, 0 observations found"
[1] "level 10 in allch07 removed, 0 observations found"
[1] "level 8 in marstat removed, 1 observations found"
[1] "level 9 in marstat removed, 3 observations found"
[1] "level 8 in hgbiom removed, 1 observations found"
[1] "level 13 in hgbiom removed, 0 observations found"
[1] "level 8 in hgbiof removed, 0 observations found"
[1] "level 10 in hgbiof removed, 0 observations found"
[1] "level 7 in pn1pno removed, 2 observations found"
[1] "level 8 in pn1pno removed, 0 observations found"
[1] "level 10 in pn1pno removed, 0 observations found"
[1] "level 13 in pn1pno removed, 0 observations found"
[1] "level 8 in pn2pno removed, 0 observations found"
[1] "level 7 in pns1pno removed, 0 observations found"
[1] "level 8 in pns1pno removed, 0 observations found"
[1] "level 10 in pns1pno removed, 0 observations found"
[1] "level 13 in pns1pno removed, 0 observations found"
[1] "level 8 in pns2pno removed, 0 observations found"
[1] "level 1 in ff_bentype36 removed, 2 observations found"
[1] "level 7 in nnssib_dv removed, 2 observations found"
[1] "level 10 in nnssib_dv removed, 2 observations found"
[1] "level 8 in mastat_dv removed, 0 observations found"
[1] "level 9 in mastat_dv removed, 0 observations found"
[1] "level 8 in buno_dv removed, 0 observations found"
[1] "level 7 in nchild_dv removed, 0 observations found"
[1] "level 8 in nchild_dv removed, 0 observations found"
[1] "level 8 in hrpno removed, 1 observations found"
```

```
[1] "level 9 in hrpno removed, 1 observations found"
[1] "level 10 in hrpno removed, 0 observations found"
[1] "level 8 in ppno removed, 4 observations found"
[1] "level 9 in ppno removed, 0 observations found"
[1] "level 8 in sppno removed, 0 observations found"
[1] "level 9 in sppno removed, 0 observations found"
[1] "level 8 in fnpno removed, 0 observations found"
[1] "level 10 in fnpno removed, 0 observations found"
[1] "level 8 in fnspno removed, 0 observations found"
[1] "level 10 in fnspno removed, 0 observations found"
[1] "level 8 in mnpno removed, 0 observations found"
[1] "level 13 in mnpno removed, 0 observations found"
[1] "level 8 in mnspno removed, 0 observations found"
[1] "level 13 in mnspno removed, 0 observations found"
[1] "level 3 in grfpno removed, 2 observations found"
[1] "level 5 in grmpno removed, 4 observations found"
[1] "level 6 in grmpno removed, 2 observations found"
[1] "level 7 in grmpno removed, 1 observations found"
[1] "level 4 in nnmpsp_dv removed, 2 observations found"
[1] "level 3 in nunmpsp_dv removed, 4 observations found"
[1] "level 5 in nunmpsp_dv removed, 1 observations found"
[1] "level 0 in sex_dv removed, 2 observations found"
[1] "level 7 in nnsib_dv removed, 0 observations found"
[1] "level 10 in nnsib_dv removed, 0 observations found"
[1] "level 0 in scflag_dv removed, 1 observations found"
[1] "level 7 in ndepchl_dv removed, 0 observations found"
[1] "level 8 in ndepchl_dv removed, 0 observations found"
[1] "level 4 in npensioner_dv removed, 1 observations found"
[1] "level 4 in nmpsp_dv removed, 1 observations found"
[1] "level 5 in nmpsp_dv removed, 4 observations found"
[1] "level 6 in nmpsp_dv removed, 1 observations found"
[1] "123 total levels removed from 61 different variables. In total 112 observations deleted"
[1] "---------Variance 0 Check---------"
[1] "45 variables removed since their new variance was 0"
 [1] "ivfio"         "ioutcome"      "adstatus"      "natch08"       "natch09"
"natch10"       "natch11"
 [8] "natch12"       "natch13"       "natch14"       "natch15"       "natch16"
"adoptch05"     "adoptch06"
[15] "adoptch07"     "adoptch08"     "adoptch09"     "adoptch10"     "adoptch11"
"adoptch12"     "adoptch13"
[22] "adoptch14"     "adoptch15"     "adoptch16"     "allch07"       "allch08"
"allch09"       "allch10"
[29] "allch11"       "allch12"       "allch13"       "allch14"       "allch15"
"allch16"       "indmode"
[36] "intdatd_if"    "intdatm_if"    "intdaty_if"    "doby_if"       "age_if"
"ff_ivlolw"     "ff_everint"
[43] "ff_bentype36"  "fiyrinvinc_if" "scflag_dv"
[1] "---------Dummy Variables---------"
[1] "predictor variable count went from 266 to 787"
[1] "---------Variance 0 Check---------"
[1] "116 variables removed since their new variance was 0"
  [1] "pno.8"          "nch14resp.7"    "nch415resp.6"   "nch415resp.7"
"nchresp.7"      "nchresp.8"
  [7] "nchund18resp.7" "nchund18resp.8" "natch01.8"      "natch01.9"
"natch01.10"     "natch01.11"
 [13] "natch01.12"     "natch02.9"      "natch02.10"     "natch02.11"
"natch02.13"     "natch02.14"
 [19] "natch03.9"      "natch03.10"     "natch03.12"     "natch03.15"
"natch04.8"      "natch04.9"
 [25] "natch04.10"     "natch05.5"      "natch05.9"      "natch06.7"
"natch06.9"      "natch07.8"
 [31] "nnatch.8"       "nnatch.10"      "nadoptch.5"     "adoptch02.7"
"adoptch03.4"    "adoptch03.7"
 [37] "adoptch03.8"    "adoptch04.7"    "nchunder16.7"   "nch5to15.6"
"nch10to15.5"    "allch01.8"
 [43] "allch01.9"      "allch01.10"     "allch01.11"     "allch01.12"
"allch02.2"      "allch02.9"
 [49] "allch02.10"     "allch02.11"     "allch02.13"     "allch02.14"
```

```
"allch03.9"         "allch03.10"
 [55] "allch03.12"        "allch03.15"        "allch04.9"         "allch04.10"
"allch05.9"         "allch05.11"
 [61] "allch06.7"         "allch06.9"         "allch06.12"        "marstat.8"
"marstat.9"         "hgbiom.8"
 [67] "hgbiom.13"         "hgbiof.8"          "hgbiof.10"         "pn1pno.7"
"pn1pno.8"          "pn1pno.10"
 [73] "pn1pno.13"         "pn2pno.8"          "pns1pno.7"         "pns1pno.8"
"pns1pno.10"        "pns1pno.13"
 [79] "pns2pno.8"         "nnssib_dv.7"       "nnssib_dv.10"      "mastat_dv.8"
"mastat_dv.9"       "buno_dv.8"
 [85] "nchild_dv.7"       "nchild_dv.8"       "hrpno.8"           "hrpno.9"
"hrpno.10"          "ppno.8"
 [91] "ppno.9"            "sppno.8"           "sppno.9"           "fnpno.8"
"fnpno.10"          "fnspno.8"
 [97] "fnspno.10"         "mnpno.8"           "mnpno.13"          "mnspno.8"
"mnspno.13"         "grfpno.3"
[103] "grmpno.5"          "grmpno.6"          "grmpno.7"          "nnmpsp_dv.4"
"nunmpsp_dv.3"      "nunmpsp_dv.5"
[109] "nnsib_dv.7"        "nnsib_dv.10"       "ndepchl_dv.7"      "ndepchl_dv.8"
"npensioner_dv.4" "nmpsp_dv.4"
[115] "nmpsp_dv.5"        "nmpsp_dv.6"
[1] "————————K–Means————————"
[1] "15 clusters have been made for K–Means"
[1] "K–Means results as a table, the max value in each row is a simple way to define which cluster re

      0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16
17  18  19  20  21  22  23  24  25  26  27  28  29
  1  10   5   8  24  39  90 330 336 309 305 272 247 270 138 108  82  68
70  46  38  38  38  32  28  31  19  13  12   6   6
  2  10  12  12  21  36  80 350 328 313 286 236 258 258 149 122  73  72
49  42  48  34  32  26  24  22  20   6   8   7   7
  3   7   5  11  23  42  65 289 265 247 240 253 252 254 113  91  80  60
46  43  38  33  32  25  22  19   9  13  13  11   2
  4   3   7   9  12  20  44 173 149 161 142 150 128 174  84  67  39  42
41  28  25  24  13  13  15   4  10   4   4   7   4
  5   3   2   8   7  24  47 133 149 135 134 118 116 131  58  60  37  24
33  26  25  23  12  20  16  10  11   6   2   0   1
  6   5   7   9  18  40  77 332 281 271 283 252 255 268 151 109  74  60
53  54  37  38  38  35  26  29  16  13   6  13  12
  7   7   9  22  31  57  82 409 406 417 371 350 359 378 178 134 106  89
80  64  63  55  55  42  42  24  21  20  19  17  16
  8   4   6  11  11  27  56 248 274 252 227 220 213 214 123  81  59  58
44  41  33  24  31  26  15  22  10  10   6   4  10
  9   5   3   4  14  16  52 190 168 152 156 149 131 153  80  48  38  31
34  36  36  15  16  14  11  14  13   4   6   1   1
  10  8   7  10  20  41  71 265 250 249 255 208 209 227 125  89  59  54
58  56  36  38  35  20  11  27  18  16  10   6   8
  11  3   5   6  13  12  43 163 138 150 122 129 113 110  48  49  27  26
23  19  22  16  11  11  10  12   5   9   9   4   2
  12  2   2   6   6  13  37 172 159 166 152 148 110 153  75  49  50  46
40  29  15  17  13  25  15  20   6   4   8   4   2
  13  9   6   9  25  35  67 283 295 302 292 254 305 290 136 117  91  56
66  39  35  32  32  40  29  12  17  14  14   7  10
  14  9   5   4   5  18  30 140 135 136 120 113  98 129  66  46  37  37
18  17  19  19  11  17  13   8   7   6   3   6   6
  15  2   3   8  10  14  33 179 141 134 119 121 124 144  76  40  38  31
18  37  26  12  18  16  12  13   8   9   3   2   4

     30  31  32  33  34  35  36
  1   5   9   7   4   5   1   3
  2   9   6   3   3   1   1   3
  3   7   8   5   1   5   3   5
  4   5   3   3   3   2   1   6
  5   3   1   1   2   1   1   4
  6  10   2   2  10   4   8   8
  7   6   5  13   3   4   4   1
  8   4   8   2   6   3   3   2
```

```
  9    2   4   1   3   1   0   0
  10  13   4   3   5   3   6   2
  11   2   3   2   3   3   1   2
  12   3   2   7   3   1   1   0
  13   9   5   8   3   1   3   2
  14   1   2   3   2   0   1   2
  15   5   5   2   4   3   2   5
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
 [1] "Cluster 1: Within MSE 3584085606540662, Size 3052"  "Cluster 2: Within MSE 3681464631217730, S
 [3] "Cluster 3: Within MSE 3454927243779826, Size 2637"  "Cluster 4: Within MSE 71279730475333, Size
 [5] "Cluster 5: Within MSE 52574171240992, Size 1384"    "Cluster 6: Within MSE 3516275463207399, S
 [7] "Cluster 7: Within MSE 9408366841507118, Size 3959"  "Cluster 8: Within MSE 3455435029868216, S
 [9] "Cluster 9: Within MSE 173963140060023, Size 1602"   "Cluster 10: Within MSE 3465774404360946, S
[11] "Cluster 11: Within MSE 5381495840663, Size 1326"    "Cluster 12: Within MSE 72584845168120, Siz
[13] "Cluster 13: Within MSE 3513152550378520, Size 2950" "Cluster 14: Within MSE 5939365367012, Size
[15] "Cluster 15: Within MSE 6267162588301, Size 1421"
[1] "Total between cluster MSE: 663771303894459136, Total within cluster MSE: 3168213575082083"
[1] "The K-Means model predicts exactly with an accuracy of 0.1112"
[1] "-----------Correlation Checks-----------"
[1] "employ.2 removed, correlated with 6 other variable(s)"
[1] "pensioner_dv.2 removed, correlated with 5 other variable(s)"
[1] "pidp removed, correlated with 3 other variable(s)"
[1] "hhorig.2 removed, correlated with 3 other variable(s)"
[1] "hhorig.7 removed, correlated with 3 other variable(s)"
[1] "dvage removed, correlated with 3 other variable(s)"
[1] "relup.2 removed, correlated with 3 other variable(s)"
[1] "pn2pno.2 removed, correlated with 3 other variable(s)"
[1] "hidp removed, correlated with 2 other variable(s)"
[1] "pno.2 removed, correlated with 2 other variable(s)"
[1] "memorig.2 removed, correlated with 2 other variable(s)"
[1] "memorig.7 removed, correlated with 2 other variable(s)"
[1] "sex.2 removed, correlated with 2 other variable(s)"
[1] "birthy removed, correlated with 2 other variable(s)"
[1] "nch14resp.2 removed, correlated with 2 other variable(s)"
[1] "nchunder16.2 removed, correlated with 2 other variable(s)"
[1] "nchunder16.4 removed, correlated with 2 other variable(s)"
[1] "nchunder16.6 removed, correlated with 2 other variable(s)"
[1] "marstat.2 removed, correlated with 2 other variable(s)"
[1] "marstat.4 removed, correlated with 2 other variable(s)"
[1] "marstat.6 removed, correlated with 2 other variable(s)"
[1] "hgbiom.1 removed, correlated with 2 other variable(s)"
[1] "hgbiom.2 removed, correlated with 2 other variable(s)"
[1] "hgbiom.3 removed, correlated with 2 other variable(s)"
[1] "hgbiom.4 removed, correlated with 2 other variable(s)"
[1] "hgbiom.5 removed, correlated with 2 other variable(s)"
[1] "hgbiom.6 removed, correlated with 2 other variable(s)"
[1] "hgbiom.7 removed, correlated with 2 other variable(s)"
[1] "hgbiof.1 removed, correlated with 2 other variable(s)"
[1] "hgbiof.2 removed, correlated with 2 other variable(s)"
[1] "hgbiof.3 removed, correlated with 2 other variable(s)"
[1] "hgbiof.4 removed, correlated with 2 other variable(s)"
[1] "hgbiof.5 removed, correlated with 2 other variable(s)"
[1] "hgbiof.6 removed, correlated with 2 other variable(s)"
[1] "hgbiof.7 removed, correlated with 2 other variable(s)"
[1] "respm16.2 removed, correlated with 2 other variable(s)"
[1] "pn1pno.1 removed, correlated with 2 other variable(s)"
[1] "pns2pno.2 removed, correlated with 2 other variable(s)"
[1] "fimngrs_tc.1 removed, correlated with 2 other variable(s)"
[1] "ff_bentype01.1 removed, correlated with 2 other variable(s)"
[1] "cohab_dv.1 removed, correlated with 2 other variable(s)"
[1] "mnpno.1 removed, correlated with 3 other variable(s)"
[1] "indinub_xw removed, correlated with 2 other variable(s)"
[1] "fimnlabgrs_dv removed, correlated with 2 other variable(s)"
[1] "ndepchl_dv.2 removed, correlated with 3 other variable(s)"
[1] "nchresp.1 removed, correlated with 2 other variable(s)"
[1] "nchresp.3 removed, correlated with 2 other variable(s)"
[1] "nchresp.4 removed, correlated with 2 other variable(s)"
[1] "ff_emplw.2 removed, correlated with 2 other variable(s)"
```

```
[1] "nchild_dv.3 removed, correlated with 2 other variable(s)"
[1] "pno.3 removed, correlated with 1 other variable(s)"
[1] "pno.4 removed, correlated with 1 other variable(s)"
[1] "pno.7 removed, correlated with 1 other variable(s)"
[1] "hhorig.3 removed, correlated with 1 other variable(s)"
[1] "hhorig.4 removed, correlated with 1 other variable(s)"
[1] "hhorig.5 removed, correlated with 1 other variable(s)"
[1] "hhorig.6 removed, correlated with 1 other variable(s)"
[1] "month removed, correlated with 1 other variable(s)"
[1] "nch14resp.6 removed, correlated with 1 other variable(s)"
[1] "nchresp.2 removed, correlated with 1 other variable(s)"
[1] "nchresp.5 removed, correlated with 1 other variable(s)"
[1] "natch02.8 removed, correlated with 1 other variable(s)"
[1] "natch06.8 removed, correlated with 1 other variable(s)"
[1] "natch07.9 removed, correlated with 1 other variable(s)"
[1] "nadoptch.2 removed, correlated with 1 other variable(s)"
[1] "nadoptch.3 removed, correlated with 1 other variable(s)"
[1] "nadoptch.4 removed, correlated with 1 other variable(s)"
[1] "nchunder16.1 removed, correlated with 1 other variable(s)"
[1] "nchunder16.5 removed, correlated with 1 other variable(s)"
[1] "allch03.5 removed, correlated with 1 other variable(s)"
[1] "allch06.8 removed, correlated with 1 other variable(s)"
[1] "istrtdaty.2011 removed, correlated with 1 other variable(s)"
[1] "istrtdaty.2012 removed, correlated with 1 other variable(s)"
[1] "istrtdaty.2013 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.2 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.3 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.4 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.5 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.6 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.7 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.8 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.9 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.10 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.11 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.12 removed, correlated with 1 other variable(s)"
[1] "istrtdatd removed, correlated with 1 other variable(s)"
[1] "jbstat.4 removed, correlated with 1 other variable(s)"
[1] "sf1.5 removed, correlated with 1 other variable(s)"
[1] "relup.6 removed, correlated with 1 other variable(s)"
[1] "btype5.1 removed, correlated with 1 other variable(s)"
[1] "fiyrdia removed, correlated with 1 other variable(s)"
[1] "marstat.3 removed, correlated with 1 other variable(s)"
[1] "marstat.7 removed, correlated with 1 other variable(s)"
[1] "respf16.2 removed, correlated with 1 other variable(s)"
[1] "scsf2a.2 removed, correlated with 1 other variable(s)"
[1] "scsf3a.5 removed, correlated with 1 other variable(s)"
[1] "istrtdathh removed, correlated with 1 other variable(s)"
[1] "pn1pno.2 removed, correlated with 1 other variable(s)"
[1] "pn1pno.3 removed, correlated with 1 other variable(s)"
[1] "pn1pno.4 removed, correlated with 1 other variable(s)"
[1] "pn1pno.5 removed, correlated with 1 other variable(s)"
[1] "pn1pno.6 removed, correlated with 1 other variable(s)"
[1] "pn2pno.3 removed, correlated with 1 other variable(s)"
[1] "pn2pno.4 removed, correlated with 1 other variable(s)"
[1] "pn2pno.5 removed, correlated with 1 other variable(s)"
[1] "pn2pno.6 removed, correlated with 1 other variable(s)"
[1] "pn2pno.7 removed, correlated with 1 other variable(s)"
[1] "fimnlabgrs_tc.1 removed, correlated with 1 other variable(s)"
[1] "j2paynet_dv removed, correlated with 1 other variable(s)"
[1] "age_dv removed, correlated with 1 other variable(s)"
[1] "npn_dv.1 removed, correlated with 1 other variable(s)"
[1] "npn_dv.2 removed, correlated with 1 other variable(s)"
[1] "nnssib_dv.1 removed, correlated with 1 other variable(s)"
[1] "nnssib_dv.2 removed, correlated with 1 other variable(s)"
[1] "nnssib_dv.3 removed, correlated with 1 other variable(s)"
[1] "nnssib_dv.4 removed, correlated with 1 other variable(s)"
[1] "nnssib_dv.5 removed, correlated with 1 other variable(s)"
```

```
[1] "nnssib_dv.6 removed, correlated with 1 other variable(s)"
[1] "country.2 removed, correlated with 1 other variable(s)"
[1] "country.3 removed, correlated with 1 other variable(s)"
[1] "country.4 removed, correlated with 1 other variable(s)"
[1] "xtra5min_dv.1 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.5 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.6 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.7 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.8 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.9 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.10 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.11 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.12 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.13 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.15 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.13 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.2 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.4 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.5 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.6 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.10 removed, correlated with 1 other variable(s)"
[1] "ppno.1 removed, correlated with 1 other variable(s)"
[1] "ppno.2 removed, correlated with 1 other variable(s)"
[1] "ppno.7 removed, correlated with 1 other variable(s)"
[1] "fnpno.1 removed, correlated with 1 other variable(s)"
[1] "fnpno.2 removed, correlated with 1 other variable(s)"
[1] "fnpno.3 removed, correlated with 1 other variable(s)"
[1] "fnpno.4 removed, correlated with 1 other variable(s)"
[1] "fnpno.5 removed, correlated with 1 other variable(s)"
[1] "fnpno.6 removed, correlated with 1 other variable(s)"
[1] "fnpno.7 removed, correlated with 1 other variable(s)"
[1] "mnpno.2 removed, correlated with 1 other variable(s)"
[1] "mnpno.3 removed, correlated with 1 other variable(s)"
[1] "mnpno.4 removed, correlated with 1 other variable(s)"
[1] "mnpno.5 removed, correlated with 1 other variable(s)"
[1] "mnpno.6 removed, correlated with 1 other variable(s)"
[1] "mnpno.7 removed, correlated with 1 other variable(s)"
[1] "paygu_if.1 removed, correlated with 1 other variable(s)"
[1] "indpxus_lw removed, correlated with 1 other variable(s)"
[1] "indin91_lw removed, correlated with 1 other variable(s)"
[1] "indpxub_xw removed, correlated with 1 other variable(s)"
[1] "sex_dv.1 removed, correlated with 1 other variable(s)"
[1] "fimnnet_dv removed, correlated with 1 other variable(s)"
[1] "respm16_dv.2 removed, correlated with 1 other variable(s)"
[1] "racel_dv removed, correlated with 1 other variable(s)"
[1] "fimnlabnet_dv removed, correlated with 1 other variable(s)"
[1] "163 variables removed since they had high correlation coefs"
[1] "The final cleaned dataset has been completed at this stage and is stored under the name 'cleanDa
[1] "----------Attempting a Train Test Split----------"
[1] "Good train, test split found"
[1] "The working seed found was 3"
[1] "----------kNN----------"
[1] "323 neighbours considered for each test data point"
[1] "kNN results as a table, follow the diagonal for the correctly mapped clusters"
          real
predicted   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15
16  17  18  19  20  21  22  23  24  25  26  27  28
        0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        2   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        3   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        4   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        5   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
```

```
0    0    0    0    0    0    0    0    0    0    0    0    0
          6   14    7   12   25   58  102  576  410  393  318  232  178  169   55   61   31
22   11   11    6    6    7    7    4    5    1    0    1    2
          7    6    8   12   21   33   77  232  254  276  262  196  193  200   70   55   34
24   23   15    8    6    5    4    1    3    2    1    2    2
          8    4    1    3    8   12   24   82   98  126  125  103  104  100   38   21   13
17    8   12    7    9    5    2    2    1    2    2    1    1
          9    0    0    0    1    4    2   12   12   20   33   38   42   39   16   15    9
7     6   10    4    2    2    3    1    3    0    1    1    1
         10    1    0    1    2    2    4    8   18   21   30   41   53   48   28   21   23
18   22   16   10   13    9    5    5   12    4    1    2    2
         11    1    0    0    1    3    2   14   10   23   33   42   60  108   49   36   46
31   30   30   19   16   17   13    5   10    7    5    3    3
         12    0    0    3    0    2    4   10   13   23   35   44   64  157  107  100   76
72   68   60   57   64   51   56   39   38   30   21   21   21
         13    0    0    0    0    0    0    0    0    0    2    2    0    2    5    1    1
2     3    1    3    2    2    1    2    2    1    2    2    0
         14    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0     0    0    0    0    0    0    0    0    0    0    0    0
         15    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0     0    0    0    0    0    1    0    0    0    0    0    1
         16    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0     0    0    0    0    0    0    0    0    1    0    0    0
         17    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0     0    0    0    0    0    0    0    0    0    0    0    0
         18    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0     0    0    0    0    0    0    0    0    0    0    0    0
         19    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0     0    0    0    0    0    0    0    0    0    0    0    0
         20    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0     0    0    0    0    0    0    0    0    0    0    0    0
         21    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0     0    0    0    0    0    0    0    0    0    0    0    0
         22    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0     1    0    0    0    0    0    0    0    0    0    0    0
         23    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0     0    0    0    0    0    0    0    0    0    0    0    0
         24    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0     0    0    0    0    0    0    0    0    0    0    0    0
         25    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0     0    0    0    0    0    0    0    0    0    0    0    0
         26    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0     0    0    0    0    0    0    0    0    0    0    0    0
```

	real							
predicted	29	30	31	32	33	34	35	36
0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	2	0	0	1	0	0	0
7	1	0	0	0	0	0	0	0
8	1	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0
10	0	1	0	1	0	0	1	0
11	1	0	2	0	0	1	0	1
12	15	14	16	13	11	7	10	14
13	0	1	0	0	0	2	0	0
14	0	0	0	0	0	0	0	0
15	1	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0
21	0	0	0	1	0	0	0	0
22	0	0	0	0	0	0	0	0

```
       23    0    0    1    0    0    0    0    0
       24    0    0    0    0    0    0    0    0
       25    0    0    0    0    0    0    0    0
       26    0    0    0    0    0    0    0    0
 [ reached getOption("max.print") -- omitted 10 rows ]
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
[1] "The MSE of the predicted values are of 37.1763"
[1] "The kNN model predicts exactly with an accuracy of 0.1491"
[1] "-----------CART prediction model-----------"
n= 25187

node), split, n, deviance, yval
      * denotes terminal node

 1) root 25187 723174.00 11.025170
   2) sf12mcs_dv>=-0.6467152 19606 233242.40  9.273131
     4) scsf6c.5>=0.5 10020  66033.63  7.746607 *
     5) scsf6c.5< 0.5 9586 119452.90 10.868770
      10) sf12mcs_dv>=-0.1142402 6495  63465.24 10.149810 *
      11) sf12mcs_dv< -0.1142402 3091  45575.86 12.379490 *
   3) sf12mcs_dv< -0.6467152 5581 218324.00 17.180080
     6) sf12mcs_dv>=-2.05382 4476 119378.90 15.670460
      12) sf12mcs_dv>=-1.243509 2673  56406.25 14.501310 *
      13) sf12mcs_dv< -1.243509 1803  53902.05 17.403770 *
     7) sf12mcs_dv< -2.05382 1105  47425.82 23.295020 *
[1] "Variable Importance"
 sf12mcs_dv     scsf4a.3     scsf4a.5     scsf6a.4     scsf6c.2     scsf4b.3
scsf6c.5    scsf6c.4    scsf6c.3    scsf6a.3
372719.5382  69621.3187  64183.4149  59121.6028  56115.6734  54993.1237
47755.8418  38868.5223  11227.6188   9679.6996
   scsf6a.5  sf12pcs_dv     scsf4a.2      scsf7.2     scsf4b.2  hcondn17.1
  4522.5057   2983.9213   1943.6777    563.4560    467.8697    466.2377
[1] "The MSE of the predicted values are of 13.1256"
[1] "The CART model predicts exactly with accuracy of 0.1403"
[1] "-----------Ordinary Linear Regression (Initial)-----------"
[1] "The full model AIC is: 131477.5086"
[1] "-----------Variance Inflation Factor Removal-----------"
[1] "The variable sf12mcs_dv was removed since it had a VIF score of 13291.9871"
[1] "The variable natch01.3 was removed since it had a VIF score of 5701.2035"
[1] "The variable pns1pno.1 was removed since it had a VIF score of 1621.3657"
[1] "The variable intdaty_dv.2012 was removed since it had a VIF score of 340.4403"
[1] "The variable adoptch01.3 was removed since it had a VIF score of 183.3748"
[1] "The variable sf12pcs_dv was removed since it had a VIF score of 142.1695"
[1] "The variable sppno.2 was removed since it had a VIF score of 121.8488"
[1] "The variable natch05.7 was removed since it had a VIF score of 120.1338"
[1] "The variable agegr10_dv.8 was removed since it had a VIF score of 110.326"
[1] "The variable ngrp_dv.2 was removed since it had a VIF score of 60.1317"
[1] "The variable hhtype_dv.8 was removed since it had a VIF score of 52.1474"
[1] "The variable allch01.3 was removed since it had a VIF score of 51.5323"
[1] "The variable strata was removed since it had a VIF score of 41.9373"
[1] "The variable rach16_dv.2 was removed since it had a VIF score of 39.9479"
[1] "The variable allch02.4 was removed since it had a VIF score of 35.2547"
[1] "The variable scsf4b.5 was removed since it had a VIF score of 34.2301"
[1] "The variable doby_dv was removed since it had a VIF score of 26.8292"
[1] "The variable scsf6c.5 was removed since it had a VIF score of 26.5469"
[1] "The variable fimnlabgrs_if was removed since it had a VIF score of 23.9128"
[1] "The variable fibenothr_dv was removed since it had a VIF score of 22.0494"
[1] "The variable scsf4a.5 was removed since it had a VIF score of 21.9659"
[1] "The variable nchild_dv.5 was removed since it had a VIF score of 21.3944"
[1] "The variable allch04.6 was removed since it had a VIF score of 20.307"
[1] "The variable scsf7.5 was removed since it had a VIF score of 17.8782"
[1] "The variable ff_jbstat.4 was removed since it had a VIF score of 17.7017"
[1] "The variable natch02.4 was removed since it had a VIF score of 16.5516"
[1] "The variable sclfsato.6 was removed since it had a VIF score of 15.9871"
[1] "The variable hhtype_dv.12 was removed since it had a VIF score of 15.2334"
[1] "The variable scsf3b.5 was removed since it had a VIF score of 15.0031"
[1] "The variable nchild_dv.2 was removed since it had a VIF score of 13.0634"
[1] "The variable qfhighfl_dv.1 was removed since it had a VIF score of 11.911"
```

```
[1] "The variable nchunder16.3 was removed since it had a VIF score of 11.4346"
[1] "The variable mnspno.2 was removed since it had a VIF score of 10.8308"
[1] "The variable nunmpsp_dv.1 was removed since it had a VIF score of 10.3936"
[1] "The variable marstat_dv.6 was removed since it had a VIF score of 10.2124"
[1] "35 variables removed from the Ordinary Linear Model since they have a VIF score higher than 10"
[1] "The full model AIC after VIF checks is: 134280.2181"
[1] "--------------Backwards Selection--------------"
[1] "50 out of 489 variables removed so far."
[1] "100 out of 489 variables removed so far."
[1] "150 out of 489 variables removed so far."
[1] "200 out of 489 variables removed so far."
[1] "250 out of 489 variables removed so far."
[1] "300 out of 489 variables removed so far."
[1] "335 out of 489 variables removed in backwards selection since they weren't significant at the 95
  [1] "pns2pno.4"         "sppno.1"           "agegr13_dv.12"     "hcondn2.1"
"susp.3"             "bensta5.1"
  [7] "sampst.3"          "intdatm_dv.2"      "adoptch02.3"       "sampst.2"
"npensioner_dv.3"    "vote6.3"
 [13] "nchresp.6"         "jbiindb_dv"        "grmpno.3"          "fimnpen_dv"
"sppno.3"            "susp.2"
 [19] "nmpsp_dv.3"        "hcondn96.1"        "ndepchl_dv.3"      "relup.4"
"nnsib_dv.5"         "natch04.5"
 [25] "vote6.2"           "lkmove.2"          "jbstat.10"         "ff_bentype12.1"
"npensioner_dv.1"    "natch02.2"
 [31] "nnsib_dv.6"        "nchild_dv.6"       "memorig.4"         "grfpno.2"
"gor_dv.8"           "nch10to15.4"
 [37] "natch03.7"         "aidhh.2"           "agegr13_dv.8"      "natch03.8"
"j2pay_if.1"         "buno_dv.5"
 [43] "origadd.2"         "ff_bentype30.1"    "allch02.8"         "natch05.8"
"sppno.7"            "intdatm_dv.7"
 [49] "ff_bentype17.1"    "intdaty_dv.2013"   "intdatm_dv.11"     "ppno.6"
"nch415resp.5"       "fiyrinvinc_tc.1"
 [55] "jbstat.7"          "ff_jbstat.7"       "ff_bentype24.1"    "hcondn1.1"
"pns1pno.4"          "hrpno.4"
 [61] "gor_dv.6"          "nnewborn.2"        "jbstat.2"          "fnspno.5"
"nch10to15.2"        "ivcoop.5"
 [67] "hcondn3.1"         "istrtdatmm"        "hcondn7.1"         "hhtype_dv.16"
"ivprsnt.2"          "hiqual_dv.2"
 [73] "fimngrs_if"        "allch03.7"         "ndepchl_dv.5"      "natch03.6"
"adoptch01.6"        "pns1pno.5"
 [79] "natch02.5"         "nnatch.1"          "memorig.6"         "hhresp_dv.2"
"indscub_xw"         "allch02.6"
 [85] "ff_bentype34.1"    "nch415resp.4"      "ff_bentype33.1"    "ff_jbstat.10"
"scsf2a.3"           "grfpno.1"
 [91] "grmpno.1"          "ngrp_dv.1"         "ff_bentype38.1"    "adoptch01.5"
"mastat_dv.7"        "ff_bentype05.1"
 [97] "fimnmisc_dv"       "natch01.7"         "intdatm_dv.9"      "memorig.5"
"adoptch01.1"        "allch02.3"
[103] "nchund18resp.5"    "natch03.4"         "rhland_code.1"     "vote1.2"
"ff_bentype07.1"     "ndepchl_dv.1"
[109] "fnspno.3"          "marstat_dv.2"      "seearngrs_if.1"    "nnsib_dv.2"
"hhtype_dv.18"       "bensta7.1"
[115] "ff_bentype09.1"    "adoptch02.5"       "hcondn11.1"        "ftedany.2"
"bensta2.1"          "agegr5_dv.14"
[121] "ff_bentype25.1"    "intdatm_dv.5"      "btype1.1"          "btype9.1"
"btype96.1"          "agegr13_dv.11"
[127] "intdaty_dv.2011"   "allch02.5"         "allch01.7"         "allch02.7"
"nnatch.5"           "natch05.6"
[133] "hrpid"             "indinus_lw"        "indscus_lw"        "ff_bentype28.1"
"cindtime"           "hcondn8.1"
[139] "npensioner_dv.2"   "hiqual_dv.5"       "hcondn10.1"        "ff_bentype03.1"
"allch01.4"          "nchild_dv.1"
[145] "nmpsp_dv.2"        "natch04.6"         "natch04.7"         "gor_dv.5"
"ff_bentype23.1"     "btype7.1"
[151] "nch10to15.3"       "grmpno.4"          "nnewborn.1"        "ff_bentype37.1"
"undqus.5"           "buno_dv.7"
[157] "scsf3b.3"          "fimnsben_dv"       "ppno.5"            "sppno.5"
"mobuse.2"           "jbstat.97"
```

```
[163] "jbstat.8"          "allch05.8"          "lenindintv"         "nchund18resp.3"
"ff_bentype13.1"    "allch01.1"
[169] "ppno.3"            "istrtdatss"         "natch03.3"          "ff_bentype04.1"
"jbstat.9"          "bensta4.1"
[175] "hhtype_dv.6"       "pno.5"              "buno_dv.4"          "pns1pno.2"
"buno_dv.3"         "fnspno.2"
[181] "ff_bentype22.1"    "ff_bentype15.1"     "btype2.1"           "ienddathh"
"ff_bentype06.1"    "ff_bentype20.1"
[187] "depchl_dv.2"       "natch01.1"          "pns2pno.3"          "ienddatss"
"nadoptch.1"        "adoptch01.4"
[193] "qfhigh_dv"         "relup.3"            "hhtype_dv.19"       "hhtype_dv.20"
"fnspno.1"          "nnmpsp_dv.2"
[199] "fimnprben_dv"      "fimnlabnet_tc.1"    "sclfsat1.4"         "btype8.1"
"nchild_dv.4"       "pns2pno.7"
[205] "pns1pno.6"         "intdatm_dv.6"       "intdatm_dv.3"       "bensta1.1"
"gor_dv.2"          "gor_dv.12"
[211] "ff_bentype11.1"    "health.2"           "ff_bentype18.1"     "adoptch04.6"
"allch03.8"         "gor_dv.9"
[217] "gor_dv.3"          "gor_dv.11"          "j2pay_dv"           "gor_dv.4"
"j2has.2"           "hcondn6.1"
[223] "nch10to15.1"       "nchund18resp.6"     "ff_jbstat.97"       "nch14resp.1"
"marstat.5"         "marstat_dv.4"
[229] "intdatd_dv"        "adoptch03.6"        "scsf3b.4"           "fibenothr_tc.1"
"fimninvnet_dv"     "bensta6.1"
[235] "undqus.2"          "ivcoop.2"           "indin01_lw"         "relup.5"
"marstat_dv.3"      "ff_bentype14.1"
[241] "ff_jbstat.9"       "hcondn9.1"          "natch01.4"          "natch03.5"
"allch01.5"         "natch01.5"
[247] "adoptch02.4"       "adoptch03.5"        "nnmpsp_dv.1"        "ff_jbstat.5"
"gor_dv.10"         "jbstat.11"
[253] "hrpno.7"           "sppno.6"            "xtra5minosm_dv.1"   "scsf5.3"
"jbstat.6"          "ff_jbstat.8"
[259] "adoptch01.2"       "fiyrinvinc_dv"      "hcondn14.1"         "hcondn12.1"
"hcondn5.1"         "pns2pno.5"
[265] "fnspno.6"          "pns2pno.6"          "fnspno.7"           "buno_dv.6"
"pno.6"             "nnsib_dv.3"
[271] "btype6.1"          "ff_bentype19.1"     "fimngrs_dv"         "undqus.4"
"grmpno.2"          "nch5to15.5"
[277] "nnatch.6"          "ndepchl_dv.6"       "ienddatmm"          "urban_dv.2"
"ind5mus_lw"        "ff_bentype02.1"
[283] "mastat_dv.3"       "nch5to15.3"         "nch5to15.2"         "nch5to15.4"
"allch04.7"         "ndepchl_dv.4"
[289] "nch14resp.4"       "natch02.6"          "intdatm_dv.10"      "intdatm_dv.12"
"bensta96.1"        "ff_bentype16.1"
[295] "nunmpsp_dv.2"      "nch5to15.1"         "nch415resp.1"       "nch415resp.2"
"nch14resp.3"       "nch415resp.3"
[301] "respf16_dv.2"      "jbstat.5"           "hcondn15.1"         "hcondn4.1"
"sclfsat1.3"        "allch03.4"
[307] "hhtype_dv.22"      "hrpno.3"            "hhresp_dv.3"        "vote6.4"
"marstat_dv.5"      "hcondn16.1"
[313] "allch03.6"         "paynu_if.1"         "fibenothr_if"       "ff_jbstat.2"
"intdatm_dv.4"      "ff_bentype21.1"
[319] "natch01.6"         "natch02.7"          "hhtype_dv.5"        "natch02.3"
"natch01.2"         "allch01.2"
[325] "ppno.4"            "ff_bentype32.1"     "scsf5.2"            "bensta3.1"
"nch14resp.5"       "ff_bentype29.1"
[331] "trainany.2"        "aidxhh.2"           "btype4.1"           "ff_jbstat.6"
"ivcoop.4"
[1] "————Ordinary Linear Regression (Improved)————"

Call:
lm(formula = y ~ ., data = as.data.frame(x.data.linear))

Residuals:
     Min       1Q   Median       3Q      Max
-18.1936  -1.9266  -0.2299   1.6318  22.6767

Coefficients:
```

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	6.71367	0.24086	27.874	< 2e−16	***
memorig.3	0.26756	0.12186	2.196	0.028134	*
psu	−0.11740	0.02538	−4.626	3.74e−06	***
nchund18resp.1	−0.19863	0.08340	−2.382	0.017249	*
nchund18resp.2	−0.28465	0.09562	−2.977	0.002913	**
nchund18resp.4	−0.90327	0.26180	−3.450	0.000561	***
natch04.4	−3.50694	1.60409	−2.186	0.028806	*
allch01.6	1.03421	0.37657	2.746	0.006030	**
allch04.5	2.90681	0.71471	4.067	4.77e−05	***
allch04.8	4.40063	1.72758	2.547	0.010862	*
allch05.6	−3.30685	1.48797	−2.222	0.026265	*
allch05.7	−1.08372	0.50573	−2.143	0.032132	*
xpmove.2	−0.13799	0.06936	−1.990	0.046656	*
jbstat.3	0.54212	0.11877	4.565	5.03e−06	***
sf1.2	−0.20729	0.08570	−2.419	0.015582	*
sf1.3	−0.25284	0.09561	−2.644	0.008189	**
sf1.4	−0.29009	0.11426	−2.539	0.011131	*
hcondn13.1	0.94355	0.24989	3.776	0.000160	***
hcondn17.1	2.15118	0.18174	11.837	< 2e−16	***
jbhas.2	0.48181	0.07666	6.285	3.33e−10	***
btype3.1	0.29803	0.12103	2.463	0.013802	*
finnow.2	0.12432	0.06088	2.042	0.041144	*
finnow.3	0.44533	0.07049	6.318	2.70e−10	***
finnow.4	1.03103	0.10067	10.242	< 2e−16	***
finnow.5	1.88114	0.14352	13.107	< 2e−16	***
finfut.2	0.49980	0.06935	7.207	5.88e−13	***
finfut.3	0.13190	0.05627	2.344	0.019093	*
ivcoop.3	0.45195	0.21709	2.082	0.037362	*
undqus.3	−0.33223	0.12924	−2.571	0.010157	*
scsf1.2	0.31849	0.09127	3.489	0.000485	***
scsf1.3	0.37174	0.10406	3.572	0.000354	***
scsf1.4	0.57802	0.13006	4.444	8.86e−06	***
scsf1.5	1.94374	0.15401	12.621	< 2e−16	***
scsf2b.2	0.24597	0.10123	2.430	0.015115	*
scsf2b.3	0.26568	0.10171	2.612	0.008999	**
scsf3a.2	−0.45456	0.13122	−3.464	0.000533	***
scsf3a.3	−0.50942	0.07997	−6.370	1.92e−10	***
scsf3a.4	−0.18759	0.06317	−2.970	0.002984	**
scsf3b.2	−0.57237	0.13487	−4.244	2.20e−05	***
scsf4a.2	1.89182	0.15314	12.353	< 2e−16	***
scsf4a.3	1.14016	0.09614	11.859	< 2e−16	***
scsf4a.4	0.63788	0.07035	9.068	< 2e−16	***
scsf4b.2	1.82323	0.16647	10.952	< 2e−16	***
scsf4b.3	0.97562	0.09777	9.979	< 2e−16	***
scsf4b.4	0.57223	0.06827	8.382	< 2e−16	***
scsf5.4	0.21086	0.09203	2.291	0.021959	*
scsf5.5	1.32144	0.12342	10.707	< 2e−16	***
scsf6a.2	0.50961	0.09934	5.130	2.92e−07	***
scsf6a.3	1.52250	0.10921	13.941	< 2e−16	***
scsf6a.4	3.28637	0.12515	26.260	< 2e−16	***
scsf6a.5	5.87292	0.16873	34.807	< 2e−16	***
scsf6b.2	0.84858	0.11047	7.681	1.63e−14	***
scsf6b.3	1.21111	0.11732	10.323	< 2e−16	***
scsf6b.4	2.01054	0.13186	15.248	< 2e−16	***
scsf6b.5	2.83977	0.16411	17.304	< 2e−16	***
scsf6c.2	4.97476	0.12481	39.858	< 2e−16	***
scsf6c.3	2.47942	0.07454	33.263	< 2e−16	***
scsf6c.4	1.23938	0.05577	22.223	< 2e−16	***
scsf7.2	1.95470	0.13098	14.923	< 2e−16	***
scsf7.3	1.11407	0.08336	13.365	< 2e−16	***
scsf7.4	0.86095	0.06591	13.062	< 2e−16	***
sclfsat1.2	−0.29757	0.08449	−3.522	0.000429	***
sclfsat1.5	−0.21864	0.07655	−2.856	0.004292	**
sclfsat1.6	−0.17549	0.06842	−2.565	0.010322	*
sclfsat1.7	−0.33903	0.10239	−3.311	0.000930	***
sclfsat2.2	−0.91984	0.11962	−7.690	1.53e−14	***
sclfsat2.3	−0.86403	0.11504	−7.511	6.06e−14	***

```
sclfsat2.4      -0.89389    0.12231   -7.308  2.79e-13 ***
sclfsat2.5      -0.93689    0.11982   -7.819  5.52e-15 ***
sclfsat2.6      -0.70065    0.12134   -5.774  7.82e-09 ***
sclfsat2.7      -0.65049    0.14684   -4.430  9.46e-06 ***
sclfsat7.2      -0.57684    0.13474   -4.281  1.87e-05 ***
sclfsat7.3      -0.67148    0.12695   -5.289  1.24e-07 ***
sclfsat7.4      -0.69145    0.12925   -5.350  8.88e-08 ***
sclfsat7.5      -0.66919    0.12732   -5.256  1.48e-07 ***
sclfsat7.6      -0.74749    0.12665   -5.902  3.64e-09 ***
sclfsat7.7      -0.79990    0.13889   -5.759  8.55e-09 ***
sclfsato.2       1.26829    0.11280   11.244   < 2e-16 ***
sclfsato.3       1.59755    0.09507   16.803   < 2e-16 ***
sclfsato.4       0.81781    0.09069    9.018   < 2e-16 ***
sclfsato.5       0.46881    0.06654    7.045  1.90e-12 ***
sclfsato.7      -0.81120    0.08121   -9.989   < 2e-16 ***
pns1pno.3       -0.82215    0.32502   -2.530  0.011428 *
ff_jbstat.3     -0.38748    0.10892   -3.558  0.000375 ***
ff_bentype08.1  -1.26570    0.34678   -3.650  0.000263 ***
ff_bentype10.1  -0.28451    0.13733   -2.072  0.038299 *
ff_bentype26.1   0.31670    0.14627    2.165  0.030386 *
ff_bentype27.1  -0.54851    0.26900   -2.039  0.041455 *
ff_bentype31.1   1.71258    0.61396    2.789  0.005284 **
ff_bentype35.1   2.41398    1.22124    1.977  0.048091 *
agegr13_dv.3    -0.43939    0.18150   -2.421  0.015486 *
agegr13_dv.4    -0.29972    0.11729   -2.555  0.010613 *
agegr13_dv.5    -0.28791    0.09478   -3.038  0.002386 **
agegr13_dv.6    -0.21545    0.08396   -2.566  0.010292 *
agegr13_dv.7    -0.21988    0.07992   -2.751  0.005944 **
agegr13_dv.9     0.18408    0.07642    2.409  0.016022 *
agegr13_dv.10    0.20673    0.07934    2.606  0.009176 **
buno_dv.2       -0.35854    0.14678   -2.443  0.014584 *
hrpno.2          0.11164    0.04531    2.464  0.013749 *
hrpno.5          1.61488    0.54174    2.981  0.002877 **
hrpno.6          1.60735    0.63957    2.513  0.011971 *
sppno.4          0.84872    0.37479    2.265  0.023551 *
fnspno.4         1.58284    0.50881    3.111  0.001868 **
hiqual_dv.3     -0.14989    0.05978   -2.507  0.012170 *
hiqual_dv.4     -0.17295    0.05790   -2.987  0.002819 **
hiqual_dv.9     -0.29843    0.07940   -3.759  0.000171 ***
wave.3          -0.34078    0.04481   -7.604  2.96e-14 ***
sex_dv.2         0.42471    0.05183    8.194  2.64e-16 ***
intdatm_dv.8    -0.19569    0.07780   -2.515  0.011907 *
nnsib_dv.1      -0.32176    0.13452   -2.392  0.016766 *
nnsib_dv.4      -1.68954    0.56406   -2.995  0.002744 **
ethn_dv         -0.08900    0.02429   -3.664  0.000249 ***
npns_dv.1        0.36884    0.15055    2.450  0.014291 *
gor_dv.7        -0.17797    0.07836   -2.271  0.023148 *
hhtype_dv.10     0.17400    0.08276    2.102  0.035536 *
hhtype_dv.11     0.21037    0.08032    2.619  0.008820 **
hhtype_dv.17    -0.56524    0.17653   -3.202  0.001367 **
hhtype_dv.21     2.30039    1.09938    2.092  0.036409 *
hhtype_dv.23     0.42973    0.19403    2.215  0.026787 *
nmpsp_dv.1       0.39232    0.07911    4.959  7.13e-07 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1    1

Residual standard error: 3.443 on 25067 degrees of freedom
Multiple R-squared:  0.5891,    Adjusted R-squared:  0.5871
F-statistic:   302 on 119 and 25067 DF,  p-value: < 2.2e-16

AIC: 133881.4653
MSE: 11.799
[1] "The MSE of the predicted values are of 12.1347"
[1] "The Linear Model predicts exactly with accuracy of 0.1546"
[1] "The Linear Model predicts within a confidence interval with accuracy of 0.2773"
[1] "----------Elastic Net Regression----------"
509 x 1 sparse Matrix of class "dgCMatrix", with 46 entries
          names Estimate_Coefs
```

```
1     (Intercept)    12.002139202
2             psu    -0.095149217
3        jbstat.3     0.126896647
4        health.2    -0.028397220
5        aidxhh.2    -0.017451406
6      hcondn13.1     0.297566142
7      hcondn17.1     0.864670420
8        finnow.2    -0.022805022
9        finnow.3     0.045718960
10       finnow.4     0.546251989
11       finnow.5     1.334194128
12       finfut.2     0.261352948
13        scsf1.5     0.680068257
14       scsf2a.3    -0.328916542
15       scsf2b.3    -0.284846144
16       scsf3b.4     0.006714031
17        scsf5.4     0.249920293
18        scsf5.5     1.051762907
19       scsf6a.4     0.621381593
20       scsf6a.5     1.166761273
21       scsf6b.3    -0.032496140
22       scsf6c.2     1.630521319
23       scsf6c.4    -0.226710998
24       scsf6c.5    -0.991066454
25        scsf7.5    -0.072783900
26     sclfsat2.5    -0.066973552
27     sclfsato.2     0.325232859
28     sclfsato.3     0.884016402
29     sclfsato.5    -0.086673500
30     sclfsato.6    -0.723192987
31     sclfsato.7    -1.272064925
32   agegr13_dv.9     0.075816255
33  agegr13_dv.10     0.024411866
34     sf12mcs_dv    -2.787918383
35         wave.3    -0.158510357
36       sex_dv.2     0.187356016
37   marstat_dv.6    -0.120092805
38     nnsib_dv.1    -0.079852571
39   agegr10_dv.3    -0.023301859
40   agegr10_dv.5     0.049833890
41   agegr10_dv.6     0.081246030
42    single_dv.1    -0.034049814
43        ethn_dv    -0.040211196
44       gor_dv.7    -0.014216504
45     sf12pcs_dv    -0.476586151
46  hhtype_dv.17    -0.083348641
[1] "The MSE of the predicted values of the best fit model is 10.6367"
[1] "The Alpha of the best fit model is 1"
[1] "The Elastic Net Model predicts exactly with accuracy of 0.1607"
[1] "------------Timer Results------------"
   user  system elapsed
2171.99    7.11 2179.96
```

10.2.24 wSMerge console

```
[1] "------------Initial Checks------------"
[1] "50598476 NA cells were found across the entire dataset (65.28% of data as NA)"
[1] "------------Data Type Checks------------"
[1] "0 variables recoded since all their entries aren't numeric or NA"
[1] "NOTE: algorithim recodes categorical data alphabetically e.g (female = 1, male = 2)"
character(0)
[1] "------------Low Data Removal------------"
[1] "862 variables removed since they had >= 'naPercent' (default 20%) NA values"
  [1] "pid"          "lvwhy"        "lvmthp"       "lvyrp"        "mvever"
"mvmnth"       "mvyr"         "mlstatchk"
```

```
  [9] "mlstat"        "drive"         "caruse"        "netuse"        "ukborn"
"plbornc"        "yr2uk4"        "citzn1"
 [17] "citzn2"        "citzn3"        "qfhigh"        "qualoc"        "qfvoc1"
"qfvoc2"         "qfvoc3"        "qfvoc4"
 [25] "qfvoc5"        "qfvoc6"        "qfvoc7"        "qfvoc8"        "qfvoc9"
"qfvoc10"        "qfvoc11"       "qfvoc12"
 [33] "qfvoc13"       "qfvoc14"       "qfvoc15"       "qfvoc96"       "school"
"scend"          "schlloc"       "schok"
 [41] "fenow"         "feend"         "j1none"        "j1semp"        "j1boss"
"j1mngr"         "edtype"        "edasp"
 [49] "fedlik"        "fednt"         "ocimpa"        "ocimpb"        "ocimpe"
"ocimpf"         "ocimpi"        "ocimpk"
 [57] "ocimpl"        "futra"         "futrb"         "futrc"         "futrd"
"futre"          "futrf"         "futrg"
 [65] "futrh"         "futri"         "futrj"         "futrk"         "futrl"
"paju"           "maju"          "pacob"
 [73] "payruk"        "macob"         "mayruk"        "natid1"        "natid2"
"natid3"         "natid4"        "natid5"
 [81] "natid6"        "natid97"       "racel"         "racelo_code"   "oprlg"
"oprlg0ni"       "nirel"         "niact"
 [89] "oprlg0"        "oprlg1"        "hospc1"        "hospdc1"       "hospc2"
"hospdc2"        "hospc3"        "hospdc3"
 [97] "hospc4"        "hospdc4"       "hospc5"        "hospdc5"       "hospc6"
"hospdc6"        "hospc7"        "hospdc7"
[105] "hospc8"        "hospdc8"       "disdif1"       "disdif2"       "disdif3"
"disdif4"        "disdif5"       "disdif6"
[113] "disdif7"       "disdif8"       "disdif9"       "disdif10"      "disdif11"
"disdif12"       "disdif96"      "aidhua1"
[121] "aidhua2"       "aidhua3"       "aidhua4"       "aidhua5"       "aidhua6"
"aidhua7"        "aidhua8"       "aidhua9"
[129] "aidhua10"      "aidhua11"      "aidhua12"      "aidhua13"      "aidhua14"
"aidhua15"       "aidhua16"      "naidxhh"
[137] "aidhu1"        "aidhu2"        "aidhrs"        "aideft"        "lcohnpi"
"coh1bm"         "coh1by"        "coh1mr"
[145] "coh1em"        "coh1ey"        "nmar"          "lmar1m"        "lmar1y"
"ladopt"         "lnadopt"       "lprnt"
[153] "lnprnt"        "ch1by4"        "movy11"        "movy12"        "movy13"
"movy14"         "movy15"        "movy16"
[161] "family"        "education"     "memploy"       "housing"       "area"
"moveoth_code"   "movdir"        "plnowm"
[169] "plnowy4"       "mstatsam"      "lwwrong"       "mstatsamn"     "mstatch1"
"statcm1"        "statcy41"      "divchk1"
[177] "divfin1"       "dvm1"          "dvy41"         "cmlstat1"      "mstatch2"
"statcm2"        "statcy42"      "divchk2"
[185] "divfin2"       "dvm2"          "dvy42"         "cmlstat2"      "cohab"
"cohabn"         "lmcbm1"        "lmcby41"
[193] "currpart1"     "lmspm1"        "lmspy41"       "lmcbm2"        "lmcby42"
"currpart2"      "lmspm2"        "lmspy42"
[201] "lmcbm3"        "lmcby43"       "currpart3"     "lmspm3"        "lmspy43"
"father"         "nchild"        "preg"
[209] "pregm1"        "pregy41"       "pregfert1"     "invitro1"      "pregout1"
"pregend1"       "endmnth1"      "pregsmoke1"
[217] "smkmnth11"     "smkmnth21"     "smkmnth31"     "pregsmk11"     "pregsmk21"
"pregsmk31"      "lchmulti1"     "pregm2"
[225] "pregy42"       "pregfert2"     "invitro2"      "pregout2"      "pregend2"
"endmnth2"       "pregsmoke2"    "smkmnth12"
[233] "smkmnth22"     "smkmnth32"     "pregsmk12"     "pregsmk22"     "pregsmk32"
"lchmulti2"      "pregm3"        "pregy43"
[241] "pregfert3"     "invitro3"      "pregout3"      "pregend3"      "endmnth3"
"pregsmoke3"     "smkmnth13"     "smkmnth23"
[249] "smkmnth33"     "pregsmk13"     "pregsmk23"     "pregsmk33"     "lchmulti3"
"hcondno1"       "hcondns1"      "hcondno2"
[257] "hcondns2"      "hcondno3"      "hcondns3"      "hcondno4"      "hcondns4"
"hcondno5"       "hcondns5"      "hcondno6"
[265] "hcondns6"      "hcondno7"      "hcondns7"      "hcondno8"      "hcondns8"
"contft"         "ftendm"        "ftendy4"
[273] "ftquals"       "ftedstartm1"   "ftedstarty41"  "ftedend1"      "ft2endm1"
"ft2endy41"      "ftedmor1"      "ftedstartm2"
```

```
[281] "ftedstarty42" "ftedend2"      "ft2endm2"      "ft2endy42"     "ftedmor2"
"qualnew1"      "qualnew2"      "qualnew3"
[289] "qualnew4"      "qualnew5"      "qualnew6"      "qualnew7"      "qualnew8"
"qualnew9"      "qualnew10"     "qualnew11"
[297] "qualnew13"     "qualnew15"     "qualnew16"     "qualnew17"     "qualnew18"
"qualnew19"     "qualnew20"     "qualnew21"
[305] "qualnew22"     "qualnew23"     "qualnew24"     "qualnew25"     "qualnew26"
"qualnew27"     "qualnew28"     "qualnew29"
[313] "qualnew30"     "qualnew31"     "trwho1"        "traindays1"    "trainhrs1"
"trainend1"     "trainpurp11"   "trainpurp21"
[321] "trainpurp31"   "trainpurp41"   "trainpurp51"   "trainpurp61"   "trainpurp71"
"trainqual1"    "trwho2"        "traindays2"
[329] "trainhrs2"     "trainend2"     "trainpurp12"   "trainpurp22"   "trainpurp32"
"trainpurp42"   "trainpurp52"   "trainpurp62"
[337] "trainpurp72"   "trainqual2"    "trwho3"        "traindays3"    "trainhrs3"
"trainend3"     "trainpurp13"   "trainpurp23"
[345] "trainpurp33"   "trainpurp43"   "trainpurp53"   "trainpurp63"   "trainpurp73"
"trainqual3"    "trainn"        "trqual1"
[353] "trqual2"       "trqual3"       "trqual4"       "trqual5"       "trqual6"
"trqual7"       "trqual8"       "trqual9"
[361] "trqual10"      "trqual11"      "trqual13"      "trqual15"      "trqual16"
"trqual17"      "trqual18"      "trqual19"
[369] "trqual20"      "trqual21"      "trqual22"      "trqual23"      "trqual24"
"trqual25"      "trqual26"      "trqual27"
[377] "trqual28"      "trqual29"      "trqual30"      "trqual31"      "notempchk"
"empchk"        "empstendd"     "empstendm"
[385] "empstendy4"    "stendreas"     "nxtst"         "nxtstelse"     "cstat"
"nxtstendd"     "nxtstendm"     "nxtstendy4"
[393] "jbsamr"        "wkplsam"       "samejob"       "matlv"         "matlvstd"
"matlvstm"      "matlvsty4"     "matlvendd"
[401] "matlvendm"     "matlvendy4"    "jbendd"        "jbendm"        "jbendy4"
"jbendreas"     "cjob"          "nxtjbhrs"
[409] "nxtjbes"       "nxtjbendd"     "nxtjbendm"     "nxtjbendy4"    "cjbatt"
"nextstat1"     "nextelse1"     "currstat1"
[417] "nextjob1"      "currjob1"      "jobhours1"     "reasend1"      "jbatt1"
"statendd1"     "statendm1"     "statendy41"
[425] "nextstat2"     "nextelse2"     "currstat2"     "nextjob2"      "currjob2"
"jobhours2"     "reasend2"      "jbatt2"
[433] "statendd2"     "statendm2"     "statendy42"    "nextstat3"     "nextelse3"
"currstat3"     "nextjob3"      "currjob3"
[441] "jobhours3"     "reasend3"      "jbatt3"        "statendd3"     "statendm3"
"statendy43"    "nextstat4"     "nextelse4"
[449] "currstat4"     "nextjob4"      "currjob4"      "jobhours4"     "reasend4"
"jbatt4"        "statendd4"     "statendm4"
[457] "statendy44"    "nextstat5"     "nextelse5"     "currstat5"     "nextjob5"
"currjob5"      "jobhours5"     "reasend5"
[465] "jbatt5"        "statendd5"     "statendm5"     "statendy45"    "nextstat6"
"nextelse6"     "currstat6"     "nextjob6"
[473] "currjob6"      "jobhours6"     "reasend6"      "jbatt6"        "statendd6"
"statendm6"     "statendy46"    "nextstat7"
[481] "nextelse7"     "currstat7"     "nextjob7"      "currjob7"      "jobhours7"
"reasend7"      "jbatt7"        "statendd7"
[489] "statendm7"     "statendy47"    "nextstat8"     "nextelse8"     "currstat8"
"nextjob8"      "currjob8"      "jobhours8"
[497] "reasend8"      "jbatt8"        "statendd8"     "statendm8"     "statendy48"
"nextstat9"     "nextelse9"     "currstat9"
[505] "nextjob9"      "currjob9"      "jobhours9"     "reasend9"      "jbatt9"
"statendd9"     "statendm9"     "statendy49"
[513] "jboff"         "jboffy"        "jbterm1"       "jbterm2"       "jbsic07chk"
"jbsoc00chk"    "jbsempchk"     "jbsemp"
[521] "jbbgd"         "jbbgm"         "jbbgy"         "jbmngrchk"     "jbmngr"
"jbsizechk"     "jbsize"        "jbsect"
[529] "jbsectpub"     "jbhrs"         "jbot"          "jbotpd"        "paygwc"
"paynwc"        "payusl"        "payu"
[537] "payuwc"        "payug"         "paytyp"        "ovtpay"        "extnsa"
"extrate"       "extrest"       "basnsa"
[545] "basrate"       "basrest"       "ovtnsa"        "ovtrate"       "ovtrest"
"jbpl"          "jbttwt"        "worktrav"
```

```
[553] "jsboss"        "jssize"        "jshrs"         "jstypeb"       "jsaccs"
"jspart"        "jsprbm"        "jsprby4"
[561] "jsprem"        "jsprey4"       "jsprls"        "jsprtx"        "jsprni"
"jspayu"        "jspayw"        "jspytx"
[569] "jspyni"        "jspl"          "jsttwt"        "jsttwtb"       "jsworktrav"
"jbsat"         "julk4wk"       "julkjb"
[577] "jubgn"         "julk4x1"       "julk4x2"       "julk4x3"       "julk4x4"
"julk4x5"       "julk4x6"       "julk4x96"
[585] "jbhad"         "jlendm"        "jlendy"        "jlsemp"        "jlboss"
"jlmngr"        "jlsize"        "eprosh"
[593] "j2semp"        "j2hrs"         "j2pay"         "retchk"        "ageret"
"rtpro1"        "rtpro2"        "rtpro3"
[601] "rtpro4"        "rtpro5"        "rtpro6"        "rtcon1"        "rtcon2"
"rtcon3"        "rtcon4"        "penmex"
[609] "pppex"         "pppexm"        "sppen"         "rtexpjb"       "rtfnd1"
"rtfnd2"        "rtfnd3"        "rtfnd4"
[617] "rtfnd5"        "rtfnd6"        "rtfnd7"        "rtfnd8"        "rtfnd9"
"rtfnd10"       "rtfnd96"       "retamt"
[625] "retsuf"        "ccare"         "ccwork"        "benunemp1"     "benunemp2"
"benunemp96"    "bendis1"       "bendis2"
[633] "bendis3"       "bendis4"       "bendis5"       "bendis6"       "bendis7"
"bendis8"       "bendis9"       "bendis10"
[641] "bendis96"      "benpen1"       "benpen2"       "benpen3"       "benpen4"
"benpen5"       "benpen6"       "benpen7"
[649] "benpen8"       "benpen96"      "niserps"       "bencb"         "benctc"
"benfam1"       "benfam2"       "benfam3"
[657] "benfam4"       "benfam5"       "benfam96"      "bentax1"       "bentax2"
"bentax3"       "bentax4"       "bentax5"
[665] "bentax96"      "benhou1"       "benhou2"       "benhou3"       "benhou4"
"benhou96"      "nfh01"         "nfh02"
[673] "nfh03"         "nfh04"         "nfh05"         "nfh06"         "nfh07"
"nfh08"         "nfh09"         "nfh10"
[681] "nfh11"         "nfh12"         "nfh13"         "nfh14"         "nfh15"
"nfh16"         "nfh17"         "nfh18"
[689] "nfh19"         "nfh20"         "nfh21"         "nfh22"         "nfh23"
"nfh24"         "nfh25"         "nfh26"
[697] "nfh27"         "nfh28"         "nfh29"         "nfh30"         "nfh31"
"nfh32"         "nfh33"         "nfh34"
[705] "nfh35"         "nfh36"         "nfh37"         "nfh38"         "fiyrdb1"
"fiyrdb2"       "fiyrdb3"       "fiyrdb4"
[713] "fiyrdb5"       "fiyrdb6"       "vote2"         "vote3"         "vote4"
"vote5"         "perpolinf"     "colbens1"
[721] "colbens2"      "colbens3"      "civicduty"     "polcost"       "votenorm"
"perbfts"       "grpbfts"       "voteintent"
[729] "demorient"     "rphmob_code"   "rphwrk_code"   "remail_code"   "ctadd1_code"
"ctadd2_code"   "cttown_code"   "ctcnty_code"
[737] "ctpcode_code"  "cttel1_code"   "cttel2_code"   "ctemail_code"  "livesp"
"livewith"      "lingua"        "jbbgdatd"
[745] "jbbgdatm"      "jbbgdaty"      "prel"          "preason"       "pripn"
"pjulk4wk"      "pjbptft"       "pjsptft"
[753] "prearn"        "pbnft1"        "pbnft2"        "pbnft3"        "pbnft4"
"pbnft5"        "pbnft6"        "pbnft7"
[761] "pbnft8"        "pbnft9"        "pbnft10"       "pbnft11"       "pbnft12"
"pbnft96"       "prfitb"        "ivinfnce"
[769] "ivaffct11"     "ivaffct12"     "ivaffct13"     "ivaffct14"     "ivaffct15"
"ivaffct17"     "ivaffct21"     "ivaffct22"
[777] "ivaffct23"     "ivaffct27"     "ivaffct30"     "ivaffct97"     "hgadoptm"
"hgadoptf"      "hgpart"        "ppsex"
[785] "fnpid"         "mnpid"         "pn1pid"        "pn1sex"        "pn2pid"
"pn2sex"        "pns1pid"       "pns1sex"
[793] "pns2pid"       "pns2sex"       "grfpid"        "grmpid"        "paygl"
"paynl"         "jsprf"         "payg_dv"
[801] "payn_dv"       "seearnnet_dv"  "ff_tel"        "ff_jbsemp"     "ff_jbmngr"
"ff_jbsize"     "ff_oprlg"      "ff_oprlg0"
[809] "ff_oprlg0ni"   "ff_ukborn"     "ff_yr2uk4"     "j1nssec8_dv"   "manssec8_dv"
"adresp15_dv"   "ppid"          "sppid"
[817] "fnspid"        "mnspid"        "nqfhigh_dv"    "jbsoc00_cc"    "jbsoc10_cc"
"jbsic07_cc"    "jbes2000"      "jbrgsc_dv"
```

```
[825] "jbnssec8_dv"  "jbisco88_cc"  "jlsoc00_cc"   "jlsoc10_cc"   "jlsic07_cc"
"jles2000"     "jlrgsc_dv"    "jlnssec8_dv"
[833] "jlisco88_cc"  "pasoc90_cc"   "pasoc00_cc"   "pasoc10_cc"   "masoc90_cc"
"masoc00_cc"   "masoc10_cc"   "j1soc90_cc"
[841] "j1soc00_cc"   "j1soc10_cc"   "j2soc90_cc"   "j2soc00_cc"   "j2soc10_cc"
"payu_dv"      "jlnssec3_dv"  "paygu_dv"
[849] "seearngrs_dv" "panssec8_dv"  "nhiqual_dv"   "jbft_dv"      "jbseg_dv"
"jbnssec5_dv"  "jlseg_dv"     "jlnssec5_dv"
[857] "jbnssec_dv"   "jlnssec_dv"   "jbnssec3_dv"  "j2nssec8_dv"  "jliindb_dv"
"paynu_dv"
[1] "——————Low Level Removal——————"
[1] "If a level is removed from a variable you wish to keep, reccomended to manually merge levels too
[1] "level 8 in pno removed, 2 observations found"
[1] "level 7 in nch14resp removed, 2 observations found"
[1] "level 6 in nch415resp removed, 3 observations found"
[1] "level 7 in nch415resp removed, 0 observations found"
[1] "level 7 in nchresp removed, 3 observations found"
[1] "level 8 in nchresp removed, 0 observations found"
[1] "level 7 in nchund18resp removed, 0 observations found"
[1] "level 8 in nchund18resp removed, 0 observations found"
[1] "level 10 in natch01 removed, 2 observations found"
[1] "level 11 in natch01 removed, 2 observations found"
[1] "level 12 in natch01 removed, 1 observations found"
[1] "level 10 in natch02 removed, 3 observations found"
[1] "level 11 in natch02 removed, 0 observations found"
[1] "level 13 in natch02 removed, 0 observations found"
[1] "level 14 in natch02 removed, 0 observations found"
[1] "level 9 in natch03 removed, 2 observations found"
[1] "level 10 in natch03 removed, 4 observations found"
[1] "level 12 in natch03 removed, 0 observations found"
[1] "level 15 in natch03 removed, 0 observations found"
[1] "level 8 in natch04 removed, 1 observations found"
[1] "level 9 in natch04 removed, 1 observations found"
[1] "level 10 in natch04 removed, 0 observations found"
[1] "level 5 in natch05 removed, 3 observations found"
[1] "level 9 in natch05 removed, 1 observations found"
[1] "level 7 in natch06 removed, 2 observations found"
[1] "level 9 in natch06 removed, 1 observations found"
[1] "level 8 in natch07 removed, 0 observations found"
[1] "level 10 in natch08 removed, 2 observations found"
[1] "level 11 in natch09 removed, 0 observations found"
[1] "level 12 in natch10 removed, 0 observations found"
[1] "level 8 in nnatch removed, 0 observations found"
[1] "level 10 in nnatch removed, 0 observations found"
[1] "level 5 in nadoptch removed, 3 observations found"
[1] "level 7 in adoptch02 removed, 4 observations found"
[1] "level 4 in adoptch03 removed, 4 observations found"
[1] "level 7 in adoptch03 removed, 1 observations found"
[1] "level 8 in adoptch03 removed, 0 observations found"
[1] "level 7 in adoptch04 removed, 2 observations found"
[1] "level 7 in adoptch05 removed, 0 observations found"
[1] "level 7 in nchunder16 removed, 1 observations found"
[1] "level 6 in nch5to15 removed, 0 observations found"
[1] "level 5 in nch10to15 removed, 0 observations found"
[1] "level 8 in allch01 removed, 2 observations found"
[1] "level 9 in allch01 removed, 3 observations found"
[1] "level 10 in allch01 removed, 0 observations found"
[1] "level 11 in allch01 removed, 0 observations found"
[1] "level 12 in allch01 removed, 0 observations found"
[1] "level 2 in allch02 removed, 2 observations found"
[1] "level 9 in allch02 removed, 0 observations found"
[1] "level 10 in allch02 removed, 0 observations found"
[1] "level 11 in allch02 removed, 0 observations found"
[1] "level 13 in allch02 removed, 0 observations found"
[1] "level 14 in allch02 removed, 0 observations found"
[1] "level 9 in allch03 removed, 3 observations found"
[1] "level 10 in allch03 removed, 0 observations found"
[1] "level 12 in allch03 removed, 0 observations found"
```

```
[1] "level 15 in allch03 removed, 0 observations found"
[1] "level 9 in allch04 removed, 2 observations found"
[1] "level 10 in allch04 removed, 0 observations found"
[1] "level 9 in allch05 removed, 2 observations found"
[1] "level 11 in allch05 removed, 0 observations found"
[1] "level 7 in allch06 removed, 0 observations found"
[1] "level 9 in allch06 removed, 0 observations found"
[1] "level 12 in allch06 removed, 0 observations found"
[1] "level 9 in allch07 removed, 0 observations found"
[1] "level 10 in allch07 removed, 0 observations found"
[1] "level 8 in marstat removed, 1 observations found"
[1] "level 9 in marstat removed, 1 observations found"
[1] "level 5 in ivcoop removed, 3 observations found"
[1] "level 8 in hgbiom removed, 1 observations found"
[1] "level 13 in hgbiom removed, 0 observations found"
[1] "level 7 in hgbiof removed, 4 observations found"
[1] "level 8 in hgbiof removed, 0 observations found"
[1] "level 10 in hgbiof removed, 0 observations found"
[1] "level 7 in pn1pno removed, 2 observations found"
[1] "level 8 in pn1pno removed, 0 observations found"
[1] "level 10 in pn1pno removed, 0 observations found"
[1] "level 13 in pn1pno removed, 0 observations found"
[1] "level 8 in pn2pno removed, 0 observations found"
[1] "level 7 in pns1pno removed, 0 observations found"
[1] "level 8 in pns1pno removed, 0 observations found"
[1] "level 10 in pns1pno removed, 0 observations found"
[1] "level 13 in pns1pno removed, 0 observations found"
[1] "level 8 in pns2pno removed, 0 observations found"
[1] "level 1 in ff_bentype36 removed, 2 observations found"
[1] "level 6 in nnssib_dv removed, 4 observations found"
[1] "level 10 in nnssib_dv removed, 1 observations found"
[1] "level 8 in mastat_dv removed, 0 observations found"
[1] "level 9 in mastat_dv removed, 0 observations found"
[1] "level 7 in buno_dv removed, 3 observations found"
[1] "level 8 in buno_dv removed, 0 observations found"
[1] "level 7 in nchild_dv removed, 0 observations found"
[1] "level 8 in nchild_dv removed, 0 observations found"
[1] "level 7 in hrpno removed, 3 observations found"
[1] "level 8 in hrpno removed, 1 observations found"
[1] "level 9 in hrpno removed, 1 observations found"
[1] "level 10 in hrpno removed, 0 observations found"
[1] "level 8 in ppno removed, 2 observations found"
[1] "level 9 in ppno removed, 0 observations found"
[1] "level 8 in sppno removed, 0 observations found"
[1] "level 9 in sppno removed, 0 observations found"
[1] "level 7 in fnpno removed, 0 observations found"
[1] "level 8 in fnpno removed, 0 observations found"
[1] "level 10 in fnpno removed, 0 observations found"
[1] "level 7 in fnspno removed, 0 observations found"
[1] "level 8 in fnspno removed, 0 observations found"
[1] "level 10 in fnspno removed, 0 observations found"
[1] "level 8 in mnpno removed, 0 observations found"
[1] "level 13 in mnpno removed, 0 observations found"
[1] "level 8 in mnspno removed, 0 observations found"
[1] "level 13 in mnspno removed, 0 observations found"
[1] "level 1 in grfpno removed, 2 observations found"
[1] "level 3 in grfpno removed, 1 observations found"
[1] "level 2 in grmpno removed, 4 observations found"
[1] "level 3 in grmpno removed, 4 observations found"
[1] "level 4 in grmpno removed, 4 observations found"
[1] "level 5 in grmpno removed, 2 observations found"
[1] "level 6 in grmpno removed, 1 observations found"
[1] "level 4 in nnmpsp_dv removed, 2 observations found"
[1] "level 3 in nunmpsp_dv removed, 2 observations found"
[1] "level 5 in nunmpsp_dv removed, 1 observations found"
[1] "level 3 in ficode3 removed, 3 observations found"
[1] "level 3 in ficode26 removed, 2 observations found"
[1] "level 3 in ficode28 removed, 1 observations found"
```

```
[1] "level 3 in ficode38 removed, 2 observations found"
[1] "level 0 in sex_dv removed, 2 observations found"
[1] "level 6 in nnsib_dv removed, 0 observations found"
[1] "level 10 in nnsib_dv removed, 0 observations found"
[1] "level 4 in npensioner_dv removed, 1 observations found"
[1] "level 3 in nmpsp_dv removed, 4 observations found"
[1] "level 4 in nmpsp_dv removed, 1 observations found"
[1] "level 5 in nmpsp_dv removed, 1 observations found"
[1] "level 6 in nmpsp_dv removed, 1 observations found"
[1] "level 7 in ndepchl_dv removed, 0 observations found"
[1] "level 8 in ndepchl_dv removed, 0 observations found"
[1] "135 total levels removed from 64 different variables. In total 134 observations deleted"
[1] "----------Variance 0 Check----------"
[1] "46 variables removed since their new variance was 0"
 [1] "ivfio"          "ioutcome"       "adstatus"       "natch08"        "natch09"
"natch10"        "natch11"
 [8] "natch12"        "natch13"        "natch14"        "natch15"        "natch16"
"adoptch05"      "adoptch06"
[15] "adoptch07"      "adoptch08"      "adoptch09"      "adoptch10"      "adoptch11"
"adoptch12"      "adoptch13"
[22] "adoptch14"      "adoptch15"      "adoptch16"      "allch07"        "allch08"
"allch09"        "allch10"
[29] "allch11"        "allch12"        "allch13"        "allch14"        "allch15"
"allch16"        "indmode"
[36] "intdatd_if"     "intdatm_if"     "intdaty_if"     "doby_if"        "age_if"
"ff_ivlolw"      "ff_everint"
[43] "ff_bentype36"   "fiyrinvinc_if" "ficode36"       "scflag_dv"
[1] "----------Dummy Variables----------"
[1] "predictor variable count went from 305 to 839"
[1] "----------Variance 0 Check----------"
[1] "130 variables removed since their new variance was 0"
  [1] "pno.8"             "nch14resp.7"      "nch415resp.6"     "nch415resp.7"
"nchresp.7"          "nchresp.8"
  [7] "nchund18resp.7"   "nchund18resp.8"   "natch01.8"         "natch01.9"
"natch01.10"         "natch01.11"
 [13] "natch01.12"        "natch02.9"         "natch02.10"        "natch02.11"
"natch02.13"         "natch02.14"
 [19] "natch03.9"         "natch03.10"        "natch03.12"        "natch03.15"
"natch04.8"          "natch04.9"
 [25] "natch04.10"        "natch05.5"         "natch05.9"         "natch06.7"
"natch06.9"          "natch07.8"
 [31] "nnatch.8"          "nnatch.10"         "nadoptch.5"        "adoptch02.7"
"adoptch03.4"        "adoptch03.7"
 [37] "adoptch03.8"       "adoptch04.7"       "nchunder16.7"     "nch5to15.6"
"nch10to15.5"        "allch01.8"
 [43] "allch01.9"         "allch01.10"        "allch01.11"        "allch01.12"
"allch02.2"          "allch02.9"
 [49] "allch02.10"        "allch02.11"        "allch02.13"        "allch02.14"
"allch03.9"          "allch03.10"
 [55] "allch03.12"        "allch03.15"        "allch04.9"         "allch04.10"
"allch05.9"          "allch05.11"
 [61] "allch06.7"         "allch06.9"         "allch06.12"        "marstat.8"
"marstat.9"          "ivcoop.5"
 [67] "hgbiom.8"          "hgbiom.13"         "hgbiof.7"          "hgbiof.8"
"hgbiof.10"          "pn1pno.7"
 [73] "pn1pno.8"          "pn1pno.10"         "pn1pno.13"         "pn2pno.8"
"pns1pno.7"          "pns1pno.8"
 [79] "pns1pno.10"        "pns1pno.13"        "pns2pno.8"         "nnssib_dv.6"
"nnssib_dv.10"       "mastat_dv.8"
 [85] "mastat_dv.9"       "buno_dv.7"         "buno_dv.8"         "nchild_dv.7"
"nchild_dv.8"        "hrpno.7"
 [91] "hrpno.8"           "hrpno.9"           "hrpno.10"          "ppno.8"
"ppno.9"             "sppno.8"
 [97] "sppno.9"           "fnpno.7"           "fnpno.8"           "fnpno.10"
"fnspno.7"           "fnspno.8"
[103] "fnspno.10"         "mnpno.8"           "mnpno.13"          "mnspno.8"
"mnspno.13"          "grfpno.1"
[109] "grfpno.3"          "grmpno.2"          "grmpno.3"          "grmpno.4"
```

```
"grmpno.5"          "grmpno.6"
[115] "nnmpsp_dv.4"       "nunmpsp_dv.3"      "nunmpsp_dv.5"      "ficode3.3"
"ficode26.3"        "ficode28.3"
[121] "ficode38.3"        "nnsib_dv.6"        "nnsib_dv.10"       "npensioner_dv.4"  "nmpsp_dv.3"
"nmpsp_dv.4"
[127] "nmpsp_dv.5"        "nmpsp_dv.6"        "ndepchl_dv.7"      "ndepchl_dv.8"
[1] "—————K—Means—————"
[1] "15 clusters have been made for K—Means"
[1] "K—Means results as a table, the max value in each row is a simple way to define which cluster re

       0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16
17  18  19  20  21  22  23  24  25  26  27  28  29
  1    2   1   3   3  13  27  95 101  82  83  69  73  84  37  44  25  14
24  16  14  16   7  14  10   7  10   3   1   0   1
  2    1   3   2   5   7  28 122  86 108  82  84  77  81  36  36  18  16
16  18  14  13   8   8   8   9   2   6   7   3   2
  3    8   7   4   8  17  41 209 164 177 138 165 133 187  87  56  52  49
24  38  35  21  17  28  20  16  11  14   6   7   6
  4    0   4   3  13  21  31 190 176 171 154 162 164 165  86  58  38  42
34  31  24  30  33  24  24  11  15   9  12  11   7
  5    4   2   8   2  10  17  83  90 106  82  87  77  95  41  28  34  17
23  14  17  12  11  10   9   7   4   9   4   3   4
  6    0   0   5   4  13  25  88  97  85  98  74  79  83  46  26  22  20
21  25  12  11  13   7   6  10   9   4   3   2   4
  7    1   1   5   3   7  21 102  96  85  82  81  88  92  53  36  22  29
17  17  14   9  12  10   7   8   5   3   0   1   3
  8    5   5   5  27  27  69 300 293 281 268 234 279 285 146 106  92  59
64  50  31  34  37  47  29  20  21  13  10   9  12
  9    1   4   2   7  12  22  92  81  89  80  70  74  71  43  33  17  16
18  16  12  15   9   9   3  14   5   9   4   3   3
  10   3   6  10  11  14  48 220 192 209 180 188 153 223 114  78  59  53
55  34  23  32  16  25  19  19  13   7   9   7   5
  11   0   0   0   0   1   0   2   2   6   2   3   2   5   2   1   0   1
0   0   2   1   0   0   0   0   0   0   0   0   0
  12   4   9  13  19  23  65 331 326 297 284 247 253 277 146 124  78  72
65  45  49  43  33  28  25  29  23  11   9   6   9
  13  10   7   8  16  40  76 356 313 314 292 274 238 284 166 123  73  75
69  61  50  36  44  33  34  34  24  15  14  12   9
  14   6   4   7  13  22  34 187 175 168 165 166 177 164  75  61  61  41
34  33  27  24  27  18  15  13   7   8  11   9   2
  15   0   2   4   3   7  20  79  90  95  69  64  58  66  37  21  21  18
14  14  10   9  13   7   2   7   3   5   3   3   4

      30  31  32  33  34  35  36
  1    2   1   1   2   1   1   3
  2    2   2   1   3   3   1   2
  3    6   5   4   5   2   2   6
  4    1   1   6   1   3   2   1
  5    1   2   3   2   1   1   0
  6    8   3   1   4   1   3   2
  7    2   3   0   3   2   2   1
  8   15   3   6   4   2   6   4
  9    3   1   2   1   2   3   0
  10   7   3   7   6   3   1   6
  11   0   0   0   0   0   0   0
  12   8   7   8   4   3   2   5
  13   7   8   5   9   4   5   5
  14   3   7   3   1   5   3   5
  15   2   3   2   2   1   1   0
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
 [1] "Cluster 1: Within MSE 7124835120716, Size 890"      "Cluster 2: Within MSE 5142397971132, Size
 [3] "Cluster 3: Within MSE 3477802944247110, Size 1775"  "Cluster 4: Within MSE 3586225165893482, S
 [5] "Cluster 5: Within MSE 6423795704857, Size 920"      "Cluster 6: Within MSE 6047704567082, Size
 [7] "Cluster 7: Within MSE 6267644827916, Size 923"      "Cluster 8: Within MSE 9066981266431492, S
 [9] "Cluster 9: Within MSE 6316464542372, Size 846"      "Cluster 10: Within MSE 3539336378050192, 
[11] "Cluster 11: Within MSE 279905376632843, Size 30"    "Cluster 12: Within MSE 9651025422364676, 
[13] "Cluster 13: Within MSE 9203214363338766, Size 3143" "Cluster 14: Within MSE 3454114043228132, 
[15] "Cluster 15: Within MSE 5404599031262, Size 759"
```

```
[1] "Total between cluster MSE: 656108138064986880, Total within cluster MSE: 4864856086655954"
[1] "The K-Means model predicts exactly with an accuracy of 0.1114"
[1] "----------Correlation Checks----------"
[1] "jbstat.4 removed, correlated with 26 other variable(s)"
[1] "dvage removed, correlated with 23 other variable(s)"
[1] "birthy removed, correlated with 21 other variable(s)"
[1] "ff_jbstat.4 removed, correlated with 24 other variable(s)"
[1] "age_dv removed, correlated with 20 other variable(s)"
[1] "btype5.1 removed, correlated with 16 other variable(s)"
[1] "pensioner_dv.2 removed, correlated with 15 other variable(s)"
[1] "pn1pno.1 removed, correlated with 14 other variable(s)"
[1] "pn2pno.2 removed, correlated with 15 other variable(s)"
[1] "respm16.2 removed, correlated with 15 other variable(s)"
[1] "ff_bentype01.1 removed, correlated with 12 other variable(s)"
[1] "npn_dv.2 removed, correlated with 14 other variable(s)"
[1] "nch14resp.2 removed, correlated with 10 other variable(s)"
[1] "btype4.1 removed, correlated with 10 other variable(s)"
[1] "ficode1.1 removed, correlated with 13 other variable(s)"
[1] "ficode18.1 removed, correlated with 14 other variable(s)"
[1] "hgbiof.6 removed, correlated with 10 other variable(s)"
[1] "pns1pno.1 removed, correlated with 11 other variable(s)"
[1] "nch14resp.3 removed, correlated with 9 other variable(s)"
[1] "jbstat.2 removed, correlated with 9 other variable(s)"
[1] "hgbiom.2 removed, correlated with 9 other variable(s)"
[1] "nchresp.2 removed, correlated with 9 other variable(s)"
[1] "marstat.2 removed, correlated with 10 other variable(s)"
[1] "nch14resp.1 removed, correlated with 8 other variable(s)"
[1] "nchund18resp.3 removed, correlated with 9 other variable(s)"
[1] "jbhas.2 removed, correlated with 8 other variable(s)"
[1] "hgbiom.7 removed, correlated with 8 other variable(s)"
[1] "pns2pno.2 removed, correlated with 9 other variable(s)"
[1] "nchund18resp.2 removed, correlated with 8 other variable(s)"
[1] "nchresp.4 removed, correlated with 8 other variable(s)"
[1] "natch01.2 removed, correlated with 7 other variable(s)"
[1] "nchunder16.3 removed, correlated with 9 other variable(s)"
[1] "employ.2 removed, correlated with 7 other variable(s)"
[1] "pn1pno.6 removed, correlated with 7 other variable(s)"
[1] "mastat_dv.2 removed, correlated with 9 other variable(s)"
[1] "doby_dv removed, correlated with 11 other variable(s)"
[1] "nchresp.1 removed, correlated with 7 other variable(s)"
[1] "natch02.4 removed, correlated with 7 other variable(s)"
[1] "nchunder16.2 removed, correlated with 8 other variable(s)"
[1] "nchunder16.5 removed, correlated with 9 other variable(s)"
[1] "sex.2 removed, correlated with 6 other variable(s)"
[1] "istrtdaty.2012 removed, correlated with 8 other variable(s)"
[1] "pn2pno.7 removed, correlated with 6 other variable(s)"
[1] "fimnlabgrs_dv removed, correlated with 7 other variable(s)"
[1] "nchresp.3 removed, correlated with 6 other variable(s)"
[1] "nchund18resp.4 removed, correlated with 6 other variable(s)"
[1] "natch04.6 removed, correlated with 6 other variable(s)"
[1] "nchunder16.1 removed, correlated with 7 other variable(s)"
[1] "ff_bentype18.1 removed, correlated with 6 other variable(s)"
[1] "agegr13_dv.13 removed, correlated with 7 other variable(s)"
[1] "hhorig.3 removed, correlated with 5 other variable(s)"
[1] "nch14resp.6 removed, correlated with 5 other variable(s)"
[1] "natch03.5 removed, correlated with 5 other variable(s)"
[1] "nchunder16.6 removed, correlated with 7 other variable(s)"
[1] "btype8.1 removed, correlated with 5 other variable(s)"
[1] "hgbiof.1 removed, correlated with 5 other variable(s)"
[1] "pns1pno.6 removed, correlated with 5 other variable(s)"
[1] "ff_jbstat.2 removed, correlated with 5 other variable(s)"
[1] "indpxus_lw removed, correlated with 5 other variable(s)"
[1] "nchund18resp.5 removed, correlated with 6 other variable(s)"
[1] "scsf2a.3 removed, correlated with 5 other variable(s)"
[1] "intdaty_dv.2012 removed, correlated with 7 other variable(s)"
[1] "pno.2 removed, correlated with 4 other variable(s)"
[1] "hhorig.7 removed, correlated with 4 other variable(s)"
[1] "memorig.3 removed, correlated with 4 other variable(s)"
```

```
[1] "psu removed, correlated with 5 other variable(s)"
[1] "nch14resp.4 removed, correlated with 4 other variable(s)"
[1] "nch415resp.1 removed, correlated with 4 other variable(s)"
[1] "natch05.7 removed, correlated with 4 other variable(s)"
[1] "nnatch.3 removed, correlated with 4 other variable(s)"
[1] "nnatch.4 removed, correlated with 4 other variable(s)"
[1] "nchunder16.4 removed, correlated with 5 other variable(s)"
[1] "allch03.5 removed, correlated with 5 other variable(s)"
[1] "relup.2 removed, correlated with 4 other variable(s)"
[1] "hgbiom.1 removed, correlated with 4 other variable(s)"
[1] "hgbiom.3 removed, correlated with 4 other variable(s)"
[1] "hgbiom.4 removed, correlated with 4 other variable(s)"
[1] "hgbiom.5 removed, correlated with 4 other variable(s)"
[1] "hgbiof.4 removed, correlated with 4 other variable(s)"
[1] "pns2pno.7 removed, correlated with 4 other variable(s)"
[1] "ff_emplw.2 removed, correlated with 4 other variable(s)"
[1] "nchild_dv.1 removed, correlated with 6 other variable(s)"
[1] "nchild_dv.2 removed, correlated with 6 other variable(s)"
[1] "nchild_dv.3 removed, correlated with 6 other variable(s)"
[1] "fnpno.1 removed, correlated with 4 other variable(s)"
[1] "fnpno.6 removed, correlated with 5 other variable(s)"
[1] "mnpno.2 removed, correlated with 6 other variable(s)"
[1] "indinus_lw removed, correlated with 4 other variable(s)"
[1] "sex_dv.1 removed, correlated with 5 other variable(s)"
[1] "allch06.8 removed, correlated with 5 other variable(s)"
[1] "btype3.1 removed, correlated with 4 other variable(s)"
[1] "scsf2b.3 removed, correlated with 4 other variable(s)"
[1] "scsf4a.5 removed, correlated with 4 other variable(s)"
[1] "ff_bentype19.1 removed, correlated with 4 other variable(s)"
[1] "ff_bentype22.1 removed, correlated with 4 other variable(s)"
[1] "pidp removed, correlated with 3 other variable(s)"
[1] "hhorig.2 removed, correlated with 3 other variable(s)"
[1] "nchresp.5 removed, correlated with 4 other variable(s)"
[1] "natch01.7 removed, correlated with 3 other variable(s)"
[1] "natch02.3 removed, correlated with 3 other variable(s)"
[1] "jbstat.3 removed, correlated with 3 other variable(s)"
[1] "jbstat.7 removed, correlated with 3 other variable(s)"
[1] "relup.6 removed, correlated with 3 other variable(s)"
[1] "bensta7.1 removed, correlated with 3 other variable(s)"
[1] "fiyrdia removed, correlated with 3 other variable(s)"
[1] "hgbiof.2 removed, correlated with 3 other variable(s)"
[1] "hgbiof.3 removed, correlated with 3 other variable(s)"
[1] "pn1pno.3 removed, correlated with 3 other variable(s)"
[1] "pn1pno.4 removed, correlated with 3 other variable(s)"
[1] "pn1pno.5 removed, correlated with 3 other variable(s)"
[1] "pn2pno.4 removed, correlated with 3 other variable(s)"
[1] "fimngrs_tc.1 removed, correlated with 3 other variable(s)"
[1] "npn_dv.1 removed, correlated with 3 other variable(s)"
[1] "agegr5_dv.15 removed, correlated with 3 other variable(s)"
[1] "cohab_dv.1 removed, correlated with 3 other variable(s)"
[1] "nchild_dv.5 removed, correlated with 5 other variable(s)"
[1] "ppno.1 removed, correlated with 3 other variable(s)"
[1] "nnmpsp_dv.1 removed, correlated with 3 other variable(s)"
[1] "sex_dv.2 removed, correlated with 4 other variable(s)"
[1] "fimnlabnet_dv removed, correlated with 5 other variable(s)"
[1] "nnatch.1 removed, correlated with 3 other variable(s)"
[1] "allch02.4 removed, correlated with 3 other variable(s)"
[1] "scsf3a.5 removed, correlated with 3 other variable(s)"
[1] "ff_jbstat.8 removed, correlated with 3 other variable(s)"
[1] "nchild_dv.6 removed, correlated with 4 other variable(s)"
[1] "fnspno.6 removed, correlated with 3 other variable(s)"
[1] "mnspno.2 removed, correlated with 4 other variable(s)"
[1] "fimnlabgrs_if removed, correlated with 4 other variable(s)"
[1] "agegr10_dv.7 removed, correlated with 4 other variable(s)"
[1] "hidp removed, correlated with 2 other variable(s)"
[1] "pno.3 removed, correlated with 2 other variable(s)"
[1] "memorig.2 removed, correlated with 2 other variable(s)"
[1] "memorig.7 removed, correlated with 2 other variable(s)"
```

```
[1] "natch01.3 removed, correlated with 2 other variable(s)"
[1] "natch02.2 removed, correlated with 2 other variable(s)"
[1] "natch02.8 removed, correlated with 2 other variable(s)"
[1] "natch03.4 removed, correlated with 2 other variable(s)"
[1] "nnatch.2 removed, correlated with 2 other variable(s)"
[1] "nadoptch.2 removed, correlated with 2 other variable(s)"
[1] "adoptch01.3 removed, correlated with 3 other variable(s)"
[1] "allch01.2 removed, correlated with 2 other variable(s)"
[1] "allch04.6 removed, correlated with 2 other variable(s)"
[1] "sf1.5 removed, correlated with 2 other variable(s)"
[1] "btype2.1 removed, correlated with 2 other variable(s)"
[1] "btype6.1 removed, correlated with 2 other variable(s)"
[1] "bensta4.1 removed, correlated with 2 other variable(s)"
[1] "marstat.4 removed, correlated with 2 other variable(s)"
[1] "marstat.5 removed, correlated with 2 other variable(s)"
[1] "marstat.6 removed, correlated with 2 other variable(s)"
[1] "hgbiom.6 removed, correlated with 2 other variable(s)"
[1] "hgbiof.5 removed, correlated with 2 other variable(s)"
[1] "scsf7.5 removed, correlated with 3 other variable(s)"
[1] "pn2pno.3 removed, correlated with 2 other variable(s)"
[1] "pns1pno.3 removed, correlated with 2 other variable(s)"
[1] "pns1pno.4 removed, correlated with 2 other variable(s)"
[1] "pns1pno.5 removed, correlated with 2 other variable(s)"
[1] "pns2pno.4 removed, correlated with 2 other variable(s)"
[1] "ff_jbstat.3 removed, correlated with 2 other variable(s)"
[1] "ff_bentype02.1 removed, correlated with 2 other variable(s)"
[1] "ff_bentype23.1 removed, correlated with 2 other variable(s)"
[1] "ngrp_dv.2 removed, correlated with 2 other variable(s)"
[1] "fimnsben_dv removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.5 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.6 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.7 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.8 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.9 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.10 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.11 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.12 removed, correlated with 2 other variable(s)"
[1] "mastat_dv.10 removed, correlated with 2 other variable(s)"
[1] "ppno.2 removed, correlated with 2 other variable(s)"
[1] "sppno.2 removed, correlated with 3 other variable(s)"
[1] "fnpno.2 removed, correlated with 2 other variable(s)"
[1] "mnpno.1 removed, correlated with 2 other variable(s)"
[1] "indin91_lw removed, correlated with 2 other variable(s)"
[1] "indinub_xw removed, correlated with 2 other variable(s)"
[1] "ficode22.1 removed, correlated with 3 other variable(s)"
[1] "agegr10_dv.8 removed, correlated with 2 other variable(s)"
[1] "fibenothr_dv removed, correlated with 2 other variable(s)"
[1] "nchund18resp.6 removed, correlated with 2 other variable(s)"
[1] "natch02.5 removed, correlated with 2 other variable(s)"
[1] "natch02.6 removed, correlated with 2 other variable(s)"
[1] "natch03.6 removed, correlated with 2 other variable(s)"
[1] "allch02.6 removed, correlated with 2 other variable(s)"
[1] "allch04.7 removed, correlated with 2 other variable(s)"
[1] "scsf3b.4 removed, correlated with 2 other variable(s)"
[1] "scsf3b.5 removed, correlated with 2 other variable(s)"
[1] "scsf4b.5 removed, correlated with 2 other variable(s)"
[1] "scsf6c.5 removed, correlated with 2 other variable(s)"
[1] "single_dv.1 removed, correlated with 2 other variable(s)"
[1] "pno.4 removed, correlated with 1 other variable(s)"
[1] "pno.5 removed, correlated with 1 other variable(s)"
[1] "pno.6 removed, correlated with 1 other variable(s)"
[1] "hhorig.4 removed, correlated with 1 other variable(s)"
[1] "hhorig.5 removed, correlated with 1 other variable(s)"
[1] "hhorig.6 removed, correlated with 1 other variable(s)"
[1] "strata removed, correlated with 1 other variable(s)"
[1] "month removed, correlated with 1 other variable(s)"
[1] "nch415resp.2 removed, correlated with 1 other variable(s)"
[1] "nch415resp.3 removed, correlated with 1 other variable(s)"
```

```
[1] "nch415resp.4 removed, correlated with 1 other variable(s)"
[1] "nch415resp.5 removed, correlated with 1 other variable(s)"
[1] "nchund18resp.1 removed, correlated with 1 other variable(s)"
[1] "natch03.3 removed, correlated with 1 other variable(s)"
[1] "natch04.5 removed, correlated with 1 other variable(s)"
[1] "natch05.6 removed, correlated with 1 other variable(s)"
[1] "natch06.8 removed, correlated with 1 other variable(s)"
[1] "natch07.9 removed, correlated with 1 other variable(s)"
[1] "nnatch.5 removed, correlated with 1 other variable(s)"
[1] "nadoptch.3 removed, correlated with 1 other variable(s)"
[1] "nadoptch.4 removed, correlated with 1 other variable(s)"
[1] "allch01.4 removed, correlated with 1 other variable(s)"
[1] "allch01.7 removed, correlated with 1 other variable(s)"
[1] "allch02.3 removed, correlated with 1 other variable(s)"
[1] "allch03.4 removed, correlated with 1 other variable(s)"
[1] "allch05.7 removed, correlated with 1 other variable(s)"
[1] "istrtdaty.2011 removed, correlated with 1 other variable(s)"
[1] "istrtdaty.2013 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.2 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.3 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.4 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.5 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.6 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.7 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.8 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.9 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.10 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.11 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.12 removed, correlated with 1 other variable(s)"
[1] "istrtdatd removed, correlated with 1 other variable(s)"
[1] "jbstat.6 removed, correlated with 1 other variable(s)"
[1] "jbstat.8 removed, correlated with 1 other variable(s)"
[1] "sf1.2 removed, correlated with 1 other variable(s)"
[1] "sf1.3 removed, correlated with 1 other variable(s)"
[1] "sf1.4 removed, correlated with 1 other variable(s)"
[1] "health.2 removed, correlated with 1 other variable(s)"
[1] "hcondn16.1 removed, correlated with 1 other variable(s)"
[1] "btype1.1 removed, correlated with 1 other variable(s)"
[1] "btype9.1 removed, correlated with 1 other variable(s)"
[1] "bensta2.1 removed, correlated with 1 other variable(s)"
[1] "bensta3.1 removed, correlated with 1 other variable(s)"
[1] "bensta5.1 removed, correlated with 1 other variable(s)"
[1] "bensta6.1 removed, correlated with 1 other variable(s)"
[1] "bensta96.1 removed, correlated with 1 other variable(s)"
[1] "finfut.2 removed, correlated with 1 other variable(s)"
[1] "lenindintv removed, correlated with 1 other variable(s)"
[1] "marstat.3 removed, correlated with 1 other variable(s)"
[1] "marstat.7 removed, correlated with 1 other variable(s)"
[1] "respf16.2 removed, correlated with 1 other variable(s)"
[1] "scsf1.5 removed, correlated with 1 other variable(s)"
[1] "scsf3a.2 removed, correlated with 1 other variable(s)"
[1] "scsf3a.3 removed, correlated with 1 other variable(s)"
[1] "scsf6a.2 removed, correlated with 1 other variable(s)"
[1] "scsf6b.2 removed, correlated with 1 other variable(s)"
[1] "istrtdathh removed, correlated with 1 other variable(s)"
[1] "pn1pno.2 removed, correlated with 1 other variable(s)"
[1] "pn2pno.5 removed, correlated with 1 other variable(s)"
[1] "pn2pno.6 removed, correlated with 1 other variable(s)"
[1] "pns2pno.3 removed, correlated with 1 other variable(s)"
[1] "fimnlabgrs_tc.1 removed, correlated with 1 other variable(s)"
[1] "j2paynet_dv removed, correlated with 1 other variable(s)"
[1] "ff_bentype03.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype04.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype05.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype06.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype07.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype08.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype09.1 removed, correlated with 1 other variable(s)"
```

```
[1] "ff_bentype10.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype11.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype12.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype13.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype14.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype15.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype20.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype25.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype26.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype27.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype29.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype30.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype31.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype32.1 removed, correlated with 1 other variable(s)"
[1] "ngrp_dv.1 removed, correlated with 1 other variable(s)"
[1] "nnssib_dv.1 removed, correlated with 1 other variable(s)"
[1] "nnssib_dv.2 removed, correlated with 1 other variable(s)"
[1] "nnssib_dv.3 removed, correlated with 1 other variable(s)"
[1] "nnssib_dv.4 removed, correlated with 1 other variable(s)"
[1] "nnssib_dv.5 removed, correlated with 1 other variable(s)"
[1] "fimninvnet_dv removed, correlated with 1 other variable(s)"
[1] "country.2 removed, correlated with 1 other variable(s)"
[1] "country.3 removed, correlated with 1 other variable(s)"
[1] "country.4 removed, correlated with 1 other variable(s)"
[1] "xtra5min_dv.1 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.13 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.3 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.4 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.5 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.6 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.7 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.8 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.9 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.10 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.11 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.4 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.5 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.6 removed, correlated with 1 other variable(s)"
[1] "buno_dv.3 removed, correlated with 1 other variable(s)"
[1] "nchild_dv.4 removed, correlated with 1 other variable(s)"
[1] "ppno.3 removed, correlated with 1 other variable(s)"
[1] "ppno.4 removed, correlated with 1 other variable(s)"
[1] "ppno.5 removed, correlated with 1 other variable(s)"
[1] "ppno.6 removed, correlated with 1 other variable(s)"
[1] "ppno.7 removed, correlated with 1 other variable(s)"
[1] "fnpno.3 removed, correlated with 1 other variable(s)"
[1] "fnpno.4 removed, correlated with 1 other variable(s)"
[1] "fnpno.5 removed, correlated with 1 other variable(s)"
[1] "fnspno.1 removed, correlated with 1 other variable(s)"
[1] "fnspno.2 removed, correlated with 1 other variable(s)"
[1] "mnpno.3 removed, correlated with 1 other variable(s)"
[1] "mnpno.4 removed, correlated with 1 other variable(s)"
[1] "mnpno.5 removed, correlated with 1 other variable(s)"
[1] "mnpno.6 removed, correlated with 1 other variable(s)"
[1] "mnpno.7 removed, correlated with 1 other variable(s)"
[1] "mnspno.1 removed, correlated with 1 other variable(s)"
[1] "grfpno.2 removed, correlated with 1 other variable(s)"
[1] "hiqual_dv.9 removed, correlated with 1 other variable(s)"
[1] "nunmpsp_dv.1 removed, correlated with 1 other variable(s)"
[1] "paygu_if.1 removed, correlated with 1 other variable(s)"
[1] "fibenothr_if removed, correlated with 1 other variable(s)"
[1] "indin01_lw removed, correlated with 1 other variable(s)"
[1] "indpxub_xw removed, correlated with 1 other variable(s)"
[1] "ficode19.1 removed, correlated with 1 other variable(s)"
[1] "marstat_dv.2 removed, correlated with 1 other variable(s)"
[1] "fimnnet_dv removed, correlated with 1 other variable(s)"
[1] "respm16_dv.2 removed, correlated with 1 other variable(s)"
[1] "racel_dv removed, correlated with 1 other variable(s)"
```

```
[1] "npensioner_dv.2 removed, correlated with 1 other variable(s)"
[1] "fimnpen_dv removed, correlated with 1 other variable(s)"
[1] "hhtype_dv.10 removed, correlated with 1 other variable(s)"
[1] "hhtype_dv.11 removed, correlated with 1 other variable(s)"
[1] "hhtype_dv.12 removed, correlated with 1 other variable(s)"
[1] "nmpsp_dv.1 removed, correlated with 1 other variable(s)"
[1] "343 variables removed since they had high correlation coefs"
[1] "The final cleaned dataset has been completed at this stage and is stored under the name 'cleanDa
[1] "-----------Attempting a Train Test Split-----------"
[1] "Good train, test split found"
[1] "The working seed found was 6"
[1] "-----------kNN-----------"
[1] "253 neighbours considered for each test data point"
[1] "kNN results as a table, follow the diagonal for the correctly mapped clusters"
         real
predicted   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15
16  17  18  19  20  21  22  23  24  25  26  27  28
        0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        2   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        3   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        4   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        5   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        6   1  10  15  19  29  81 443 387 351 274 211 202 181  82  49  35
26  14  16  10   5   9   7   2   6   1   0   1   3
        7   1   4   9   7  16  30  85 124 141 126 100 100 104  44  32  14
19  10  11  13   6   3   1   2   3   0   0   0   1
        8   1   1   4   9   3   6  44  52  73  63  72  69  62  37  24  14
10  12   5   8   7   4   5   1   2   2   4   2   2
        9   0   1   1   1   2   4   6   9  13  17  32  32  28  16  14   7
8   5   1   0   5   1   2   1   1   0   0   3   1
       10   0   0   0   0   0   1   4   6   6   9   8   9  22  18   8   3
7   9   6   5   2   3   6   0   1   1   2   1   1
       11   0   0   0   1   1   1   2   4  12  25  26  25  38  39  23  15
18  12  12  10  10  10  11   6   3   5   2   2   3
       12   0   0   0   0   0   2   7   7  10  16  31  47  80  55  55  45
43  41  43  40  37  40  37  26  35  23  19  19  15
       13   0   0   0   0   0   0   0   0   0   1   0   0   4   0   1   2
2   0   0   1   0   0   1   0   1   0   0   0   0
       14   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
1   0   0   1   0   0   0   0   0   0   0   0   0
       15   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0
1   0   0   2   1   0   0   0   0   0   0   1   0
       16   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
       17   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0
0   0   1   0   0   0   0   0   0   0   0   0   0
       18   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0
0   0   0   0   0   1   0   0   1   0   0   0   0
       19   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
       20   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
       21   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   1   0   0   0   0   1   0   0   0   0
       22   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   1   0   0   0   0   1   1   0   0
       23   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   1   0   0   0
       24   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
       25   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
```

```
0    0    0    0    0    0    0    0    0    0    0    0    0
          26    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0    0    0    0    0    0    0    0    0    0    0    0    0
           real
predicted  29  30  31  32  33  34  35  36
        0   0   0   0   0   0   0   0   0
        1   0   0   0   0   0   0   0   0
        2   0   0   0   0   0   0   0   0
        3   0   0   0   0   0   0   0   0
        4   0   0   0   0   0   0   0   0
        5   0   0   0   0   0   0   0   0
        6   0   2   1   2   0   0   0   0
        7   0   1   0   0   0   0   0   0
        8   0   1   0   0   0   0   0   0
        9   0   0   0   0   0   0   0   0
        10  2   1   1   1   1   0   1   0
        11  0   0   1   0   1   0   0   0
        12  9  12   9  10   5  11   7   9
        13  0   1   1   0   0   0   0   0
        14  0   0   0   0   0   0   0   0
        15  0   0   1   1   0   0   0   1
        16  0   1   0   0   0   0   0   0
        17  0   1   0   1   1   0   0   0
        18  0   0   0   0   0   0   0   0
        19  0   0   0   0   0   0   0   0
        20  0   0   0   0   0   0   1   0
        21  1   0   0   0   0   0   0   0
        22  0   0   0   0   0   0   0   1
        23  0   0   1   0   0   1   0   0
        24  0   0   0   0   0   1   0   0
        25  0   0   0   0   0   0   0   0
        26  0   0   0   0   0   1   0   1
 [ reached getOption("max.print") -- omitted 10 rows ]
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
[1] "The MSE of the predicted values are of 55.2544"
[1] "The kNN model predicts exactly with an accuracy of 0.1363"
[1] "----------CART prediction model----------"
n= 16946

node), split, n, deviance, yval
      * denotes terminal node

1) root 16946 522353.10 11.305150
  2) sf12mcs_dv>=-0.6553196 13180 167095.20  9.463202
    4) sf12mcs_dv>=0.2125211 8693  68142.02  8.323364
      8) scsf6c.4< 0.5 6376  39565.39  7.731807 *
      9) scsf6c.4>=0.5 2317  20205.49  9.951230 *
    5) sf12mcs_dv< 0.2125211 4487  65777.78 11.671500 *
  3) sf12mcs_dv< -0.6553196 3766 154045.40 17.751460
    6) sf12mcs_dv>=-1.981356 2965  81334.03 16.138620 *
    7) sf12mcs_dv< -1.981356 801  36448.92 23.721600 *
[1] "Variable Importance"
  sf12mcs_dv      scsf4a.3      scsf6c.2      scsf6a.4      scsf4b.3      scsf4a.2
scsf6a.3   sf12pcs_dv      scsf6c.4
272268.92771  47978.99340  40965.14159  40552.40088  39323.54025  32431.23496
8672.76092   8453.57099   8371.14627
    scsf6c.3      scsf4a.4      scsf4b.4      scsf6a.5   hcondn17.1     allch01.1 fimnprben_dv
fnspno.5
  7881.63950   6682.35662   4635.82361   3938.61515    181.08575     90.54288
18.06462     14.45170
[1] "The MSE of the predicted values are of 14.8036"
[1] "The CART model predicts exactly with accuracy of 0.1379"
[1] "----------Ordinary Linear Regression (Initial)----------"
[1] "The full model AIC is: 89396.5361"
[1] "----------Variance Inflation Factor Removal----------"
[1] "The variable hhtype_dv.8 was removed since it had a VIF score of 19.9273"
[1] "The variable sclfsato.6 was removed since it had a VIF score of 14.3517"
[1] "2 variables removed from the Ordinary Linear Model since they have a VIF score higher than 10"
```

```
[1] "The full model AIC after VIF checks is: 89525.7878"
[1] "———————Backwards Selection———————"
[1] "50 out of 366 variables removed so far."
[1] "100 out of 366 variables removed so far."
[1] "150 out of 366 variables removed so far."
[1] "200 out of 366 variables removed so far."
[1] "250 out of 366 variables removed so far."
[1] "279 out of 366 variables removed in backwards selection since they weren't significant at the 95
  [1] "nch10to15.1"        "buno_dv.6"          "ff_bentype34.1"     "ff_bentype21.1"
"ficode5.1"
  [6] "ff_bentype37.1"     "allch03.6"          "vote6.2"            "ficode30.1"
"hcondn8.1"
 [11] "intdatm_dv.2"       "hcondn6.1"          "ndepchl_dv.3"       "hcondn12.1"
"adoptch02.5"
 [16] "hhtype_dv.22"       "gor_dv.10"          "gor_dv.11"          "sppno.1"
"hhresp_dv.3"
 [21] "ficode4.2"          "cindtime"           "scsf5.3"            "fimngrs_dv"
"ficode38.1"
 [26] "ficode27.1"         "memorig.6"          "agegr5_dv.14"       "ff_bentype28.1"
"fiyrinvinc_tc.1"
 [31] "ndepchl_dv.5"       "rhland_code.1"      "mnspno.5"           "susp.3"
"ficode35.1"
 [36] "agegr13_dv.12"      "origadd.2"          "hcondn5.1"          "ficode3.2"
"ficode24.1"
 [41] "ivcoop.3"           "pno.7"              "agegr10_dv.3"       "ff_bentype38.1"
"hhtype_dv.19"
 [46] "hhtype_dv.16"       "ff_bentype24.1"     "nchresp.6"          "hcondn11.1"
"jbstat.97"
 [51] "intdatd_dv"         "hrpno.3"            "adoptch01.6"        "hcondn7.1"
"hcondn2.1"
 [56] "hcondn1.1"          "hrpid"              "jbstat.5"           "nnsib_dv.2"
"npns_dv.2"
 [61] "ficode2.1"          "undqus.2"           "memorig.4"          "hhtype_dv.23"
"hiqual_dv.4"
 [66] "allch03.7"          "scsf4a.2"           "natch01.1"          "mnspno.4"
"ficode7.1"
 [71] "nch5to15.2"         "bensta1.1"          "ficode31.1"         "gor_dv.12"
"jbstat.10"
 [76] "hrpno.6"            "buno_dv.5"          "ff_jbstat.10"       "lkmove.2"
"intdatm_dv.10"
 [81] "nch5to15.4"         "ficode37.1"         "allch02.5"          "indscus_lw"
"adoptch01.4"
 [86] "adoptch01.5"        "nadoptch.1"         "fnspno.5"           "natch05.8"
"ficode13.1"
 [91] "undqus.3"           "ficode24.2"         "hhresp_dv.2"        "fimnlabnet_tc.1"
"btype96.1"
 [96] "hcondn4.1"          "agegr10_dv.4"       "nmpsp_dv.2"         "ficode39.1"
"ficode26.2"
[101] "natch01.5"          "allch01.1"          "hhtype_dv.21"       "ficode2.3"
"ficode2.2"
[106] "ficode16.1"         "ficode38.2"         "intdatm_dv.9"       "btype7.1"
"j2pay_if.1"
[111] "sppno.7"            "ienddathh"          "depchl_dv.2"        "npensioner_dv.1"
"natch01.4"
[116] "allch02.7"          "allch03.8"          "adoptch01.1"        "fibenothr_tc.1"
"ff_jbstat.7"
[121] "susp.2"             "fiyrinvinc_dv"      "ficode27.2"         "vote6.4"
"indscub_xw"
[126] "nnmpsp_dv.2"        "sclfsat7.3"         "ficode34.1"         "j2has.2"
"ficode28.2"
[131] "ivcoop.2"           "ficode39.2"         "ficode9.1"          "trainany.2"
"intdatm_dv.11"
[136] "mnspno.7"           "ndepchl_dv.6"       "undqus.4"           "sclfsat1.4"
"sclfsat1.3"
[141] "ficode14.1"         "nch5to15.3"         "allch02.8"          "undqus.5"
"ficode29.2"
[146] "ficode29.3"         "vote1.2"            "j2pay_dv"           "relup.3"
"buno_dv.2"
```

```
[151] "npns_dv.1"           "pns1pno.2"          "intdatm_dv.7"        "intdaty_dv.2013"
"ienddatss"
[156] "hiqual_dv.2"         "ficode12.1"         "istrtdatss"          "istrtdatmm"
"natch03.8"
[161] "intdatm_dv.4"        "nch10to15.3"        "rach16_dv.2"         "hhtype_dv.5"
"relup.4"
[166] "finnow.2"            "ficode21.1"         "scsf4a.4"            "gor_dv.9"
"gor_dv.2"
[171] "sampst.2"            "ficode6.1"          "fimnprben_dv"        "ficode17.1"
"ff_bentype17.1"
[176] "natch04.7"           "nnatch.6"           "gor_dv.6"            "gor_dv.5"
"gor_dv.4"
[181] "nch10to15.4"         "allch05.6"          "scsf5.2"             "ftedany.2"
"scsf3a.4"
[186] "gor_dv.7"            "adoptch01.2"        "marstat_dv.5"        "buno_dv.4"
"natch03.7"
[191] "allch01.5"           "nnewborn.2"         "sppno.4"             "ff_jbstat.5"
"ivprsnt.2"
[196] "relup.5"             "hcondn10.1"         "adoptch02.4"         "adoptch03.5"
"adoptch04.6"
[201] "ficode11.1"          "fimnmisc_dv"        "frmnthimp_dv_total"  "ficode29.1"
"marstat_dv.6"
[206] "nch5to15.1"          "mnspno.6"           "nch10to15.2"         "hhtype_dv.20"
"ndepchl_dv.1"
[211] "hrpno.5"             "intdatm_dv.12"      "hiqual_dv.5"         "nnatch.7"
"grmpno.1"
[216] "gor_dv.3"            "vote6.3"            "ficode4.1"           "intdaty_dv.2011"
"adoptch02.6"
[221] "sclfsat7.4"          "sclfsat7.2"         "sclfsat7.7"          "sclfsat7.5"
"sclfsat7.6"
[226] "ficode32.1"          "paynu_if.1"         "fimngrs_if"          "seearngrs_if.1"
"hcondn3.1"
[231] "marstat_dv.4"        "marstat_dv.3"       "ficode3.1"           "hhtype_dv.18"
"mnspno.3"
[236] "allch01.3"           "scsf1.2"            "ienddatmm"           "adoptch02.3"
"adoptch03.6"
[241] "natch01.6"           "nch14resp.5"        "ficode23.1"          "natch02.7"
"sppno.6"
[246] "sampst.3"            "ff_jbstat.9"        "jbstat.9"            "mobuse.2"
"ficode25.1"
[251] "ff_jbstat.97"        "intdatm_dv.5"       "gor_dv.8"            "scsf7.4"
"pns2pno.6"
[256] "sppno.3"             "mastat_dv.7"        "natch04.4"           "hhtype_dv.6"
"ficode28.1"
[261] "ficode20.1"          "intdatm_dv.6"       "pns2pno.5"           "scsf6a.3"
"ethn_dv"
[266] "nnewborn.1"          "npensioner_dv.3"    "hcondn96.1"          "ndepchl_dv.2"
"hiqual_dv.3"
[271] "nunmpsp_dv.2"        "ficode4.3"          "ficode15.1"          "intdatm_dv.3"
"memorig.5"
[276] "xpmove.2"            "scsf4a.3"           "nch5to15.5"          "allch05.8"
[1] "——————Ordinary Linear Regression (Improved)——————"

Call:
lm(formula = y ~ ., data = as.data.frame(x.data.linear))

Residuals:
     Min       1Q   Median       3Q      Max
-20.3915  -1.8721  -0.1994   1.6828  19.1625

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    12.72706    0.25733  49.458  < 2e-16 ***
allch01.6       1.52377    0.41799   3.645 0.000268 ***
allch04.5       1.47913    0.58034   2.549 0.010821 *
allch04.8       3.69106    1.50305   2.456 0.014071 *
aidhh.2        -0.17431    0.08843  -1.971 0.048727 *
aidxhh.2       -0.18375    0.07491  -2.453 0.014174 *
```

hcondn9.1	1.74635	0.56956	3.066	0.002172	**
hcondn13.1	0.81390	0.26630	3.056	0.002244	**
hcondn14.1	−0.57976	0.22826	−2.540	0.011096	*
hcondn15.1	−2.04112	0.67771	−3.012	0.002601	**
hcondn17.1	1.11552	0.20455	5.454	5.00e−08	***
finnow.3	0.28165	0.06710	4.198	2.71e−05	***
finnow.4	0.75688	0.10495	7.212	5.76e−13	***
finnow.5	1.44639	0.15666	9.232	< 2e−16	***
finfut.3	−0.17016	0.05348	−3.182	0.001467	**
ivcoop.4	−1.41120	0.64915	−2.174	0.029726	*
scsf1.3	−0.31178	0.06455	−4.830	1.38e−06	***
scsf1.4	−0.45386	0.09160	−4.955	7.31e−07	***
scsf2a.2	0.17091	0.07961	2.147	0.031823	*
scsf2b.2	0.26309	0.07380	3.565	0.000365	***
scsf3b.2	−0.34076	0.13200	−2.582	0.009843	**
scsf3b.3	−0.25572	0.09271	−2.758	0.005817	**
scsf4b.2	−0.72417	0.17137	−4.226	2.39e−05	***
scsf4b.3	−0.70928	0.10693	−6.633	3.39e−11	***
scsf4b.4	−0.17186	0.07540	−2.279	0.022657	*
scsf5.4	0.25797	0.11291	2.285	0.022344	*
scsf5.5	0.64063	0.16049	3.992	6.59e−05	***
scsf6a.4	0.59027	0.09400	6.280	3.47e−10	***
scsf6a.5	1.21959	0.17581	6.937	4.15e−12	***
scsf6b.3	−0.62924	0.06947	−9.057	< 2e−16	***
scsf6b.4	−0.77022	0.10496	−7.339	2.26e−13	***
scsf6b.5	−1.48329	0.16428	−9.029	< 2e−16	***
scsf6c.2	2.13664	0.15563	13.729	< 2e−16	***
scsf6c.3	0.86502	0.09406	9.196	< 2e−16	***
scsf6c.4	0.53380	0.06761	7.896	3.07e−15	***
scsf7.2	−0.46099	0.14404	−3.201	0.001374	**
scsf7.3	−0.58251	0.09136	−6.376	1.87e−10	***
sclfsat1.2	−0.32041	0.09628	−3.328	0.000877	***
sclfsat1.5	−0.19656	0.09084	−2.164	0.030486	*
sclfsat1.6	−0.17937	0.07910	−2.268	0.023357	*
sclfsat1.7	−0.39162	0.11910	−3.288	0.001010	**
sclfsat2.2	−1.16037	0.13294	−8.728	< 2e−16	***
sclfsat2.3	−1.01822	0.12711	−8.010	1.22e−15	***
sclfsat2.4	−1.15241	0.13484	−8.546	< 2e−16	***
sclfsat2.5	−1.13136	0.13308	−8.501	< 2e−16	***
sclfsat2.6	−1.03872	0.13469	−7.712	1.31e−14	***
sclfsat2.7	−1.03359	0.16338	−6.326	2.58e−10	***
sclfsato.2	1.26176	0.12889	9.789	< 2e−16	***
sclfsato.3	1.37317	0.10948	12.543	< 2e−16	***
sclfsato.4	0.74012	0.10336	7.161	8.37e−13	***
sclfsato.5	0.39689	0.07932	5.004	5.68e−07	***
sclfsato.7	−0.67241	0.09051	−7.429	1.15e−13	***
ff_jbstat.6	−0.20804	0.09214	−2.258	0.023964	*
ff_bentype16.1	−0.44265	0.17139	−2.583	0.009812	**
ff_bentype33.1	−0.82278	0.33848	−2.431	0.015075	*
ff_bentype35.1	5.12847	1.50648	3.404	0.000665	***
urban_dv.2	0.14839	0.06069	2.445	0.014492	*
mastat_dv.3	1.92280	0.66042	2.911	0.003602	**
respf16_dv.2	0.19682	0.07928	2.482	0.013057	*
hrpno.2	0.12892	0.05430	2.374	0.017595	*
hrpno.4	0.85374	0.41275	2.068	0.038617	*
sppno.5	1.48391	0.70892	2.093	0.036345	*
fnspno.3	−1.06910	0.50609	−2.112	0.034660	*
fnspno.4	1.77163	0.75753	2.339	0.019364	*
sf12mcs_dv	−3.65980	0.06698	−54.639	< 2e−16	***
ind5mus_lw	0.09915	0.02713	3.655	0.000258	***
wave.3	−0.37750	0.05264	−7.172	7.70e−13	***
ficode8.1	−0.72360	0.30586	−2.366	0.018004	*
ficode10.1	−0.29075	0.11514	−2.525	0.011571	*
ficode26.1	0.41099	0.13923	2.952	0.003163	**
ficode33.1	1.01917	0.26396	3.861	0.000113	***
intdatm_dv.8	−0.20921	0.09231	−2.266	0.023439	*
nnsib_dv.1	−0.59289	0.20739	−2.859	0.004257	**
nnsib_dv.3	1.24850	0.62996	1.982	0.047509	*

```
nnsib_dv.4         -1.77300    0.82378   -2.152 0.031390 *
nnsib_dv.5         -3.90693    1.51133   -2.585 0.009744 **
agegr10_dv.5        0.20247    0.06753    2.998 0.002720 **
agegr10_dv.6        0.22354    0.08045    2.779 0.005466 **
qfhigh_dv          -0.10082    0.02894   -3.484 0.000496 ***
xtra5minosm_dv.1   -0.60319    0.09483   -6.361 2.05e-10 ***
hhtype_dv.17       -0.43348    0.19706   -2.200 0.027840 *
livesp_dv.1         0.20029    0.06622    3.025 0.002493 **
qfhighfl_dv.1      -0.30450    0.14472   -2.104 0.035383 *
ndepchl_dv.4       -0.61575    0.19005   -3.240 0.001198 **
sf12pcs_dv         -1.32902    0.06306  -21.075  < 2e-16 ***
jbiindb_dv         -0.08436    0.03143   -2.684 0.007285 **
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1      1

Residual standard error: 3.353 on 16860 degrees of freedom
Multiple R-squared:  0.6371,    Adjusted R-squared:  0.6352
F-statistic: 348.2 on 85 and 16860 DF,  p-value: < 2.2e-16

AIC: 89185.9031
MSE: 11.1876
[1] "The MSE of the predicted values are of 11.5242"
[1] "The Linear Model predicts exactly with accuracy of 0.1561"
[1] "The Linear Model predicts within a confidence interval with accuracy of 0.2701"
[1] "----------Elastic Net Regression----------"
367 x 1 sparse Matrix of class "dgCMatrix", with 45 entries
              names Estimate_Coefs
1       (Intercept)   11.480159755
2          allch01.6   0.210412263
3           aidxhh.2  -0.049496085
4          hcondn9.1   0.115904877
5         hcondn13.1   0.126743409
6         hcondn17.1   0.717017319
7         hcondn96.1  -0.025447674
8           finnow.2  -0.053298584
9           finnow.4   0.502190273
10          finnow.5   1.330274966
11          finfut.3  -0.070545110
12          scsf2b.2   0.111283476
13          scsf4b.3  -0.122587016
14           scsf5.4   0.099508362
15           scsf5.5   0.802599635
16          scsf6a.4   0.503134480
17          scsf6a.5   0.952297720
18          scsf6b.3  -0.089409347
19          scsf6c.2   1.805948444
20          scsf6c.3   0.556318091
21          scsf6c.4   0.265358831
22           scsf7.3  -0.061716240
23        sclfsat1.7  -0.054066780
24        sclfsat2.5  -0.016842378
25        sclfsato.2   0.479894965
26        sclfsato.3   0.680900029
27        sclfsato.5  -0.106144420
28        sclfsato.6  -0.747797784
29        sclfsato.7  -1.307824507
30   ff_bentype35.1   0.780775079
31       urban_dv.2   0.001213899
32      mastat_dv.3   0.285463546
33       sf12mcs_dv  -3.257148268
34           wave.3  -0.217548637
35       ficode26.1   0.007513737
36       ficode33.1   0.266309006
37       nnsib_dv.1  -0.083017710
38     agegr10_dv.5   0.007502568
39     agegr10_dv.6   0.049361906
40        qfhigh_dv  -0.000647016
41 xtra5minosm_dv.1  -0.214839776
```

```
42       hhtype_dv.17    -0.197662926
43            ethn_dv    -0.017819791
44      ndepchl_dv.4     -0.005071537
45        sf12pcs_dv     -0.867501275
[1] "The MSE of the predicted values of the best fit model is 11.4676"
[1] "The Alpha of the best fit model is 0.9"
[1] "The Elastic Net Model predicts exactly with accuracy of 0.1599"
[1] "———————Timer Results———————"
   user  system elapsed
 575.34    5.03  580.54
```

10.2.25 wSMergeNurse console

```
[1] "———————Initial Checks———————"
[1] "13521436 NA cells were found across the entire dataset (60.49% of data as NA)"
[1] "———————Data Type Checks———————"
[1] "0 variables recoded since all their entries aren't numeric or NA"
[1] "NOTE: algorithim recodes categorical data alphabetically e.g (female = 1, male = 2)"
character(0)
[1] "———————Low Data Removal———————"
[1] "1053 variables removed since they had >= 'naPercent' (default 20%) NA values"
   [1] "pid.x"        "lvwhy"        "lvmthp"       "lvyrp"        "mvever"
"mvmnth"       "mvyr"
   [8] "mlstatchk"    "mlstat.x"     "drive"        "caruse"       "netuse"
"ukborn"       "plbornc"
  [15] "yr2uk4"       "citzn1"       "citzn2"       "citzn3"       "qfhigh"
"qualoc"       "qfvoc1"
  [22] "qfvoc2"       "qfvoc3"       "qfvoc4"       "qfvoc5"       "qfvoc6"
"qfvoc7"       "qfvoc8"
  [29] "qfvoc9"       "qfvoc10"      "qfvoc11"      "qfvoc12"      "qfvoc13"
"qfvoc14"      "qfvoc15"
  [36] "qfvoc96"      "school"       "scend"        "schlloc"      "schok"
"fenow"        "feend"
  [43] "j1none"       "j1semp"       "j1boss"       "j1mngr"       "edtype"
"edasp"        "fedlik"
  [50] "fednt"        "ocimpa"       "ocimpb"       "ocimpe"       "ocimpf"
"ocimpi"       "ocimpk"
  [57] "ocimpl"       "futra"        "futrb"        "futrc"        "futrd"
"futre"        "futrf"
  [64] "futrg"        "futrh"        "futri"        "futrj"        "futrk"
"futrl"        "paju"
  [71] "maju"         "pacob"        "payruk"       "macob"        "mayruk"
"natid1"       "natid2"
  [78] "natid3"       "natid4"       "natid5"       "natid6"       "natid97"
"racel"        "racelo_code"
  [85] "oprlg"        "oprlg0ni"     "nirel"        "niact"        "oprlg0"
"oprlg1"       "hospc1"
  [92] "hospdc1"      "hospc2"       "hospdc2"      "hospc3"       "hospdc3"
"hospc4"       "hospdc4"
  [99] "hospc5"       "hospdc5"      "hospc6"       "hospdc6"      "hospc7"
"hospdc7"      "hospc8"
 [106] "hospdc8"      "disdif1"      "disdif2"      "disdif3"      "disdif4"
"disdif5"      "disdif6"
 [113] "disdif7"      "disdif8"      "disdif9"      "disdif10"     "disdif11"
"disdif12"     "disdif96"
 [120] "aidhua1"      "aidhua2"      "aidhua3"      "aidhua4"      "aidhua5"
"aidhua6"      "aidhua7"
 [127] "aidhua8"      "aidhua9"      "aidhua10"     "aidhua11"     "aidhua12"
"aidhua13"     "aidhua14"
 [134] "aidhua15"     "aidhua16"     "naidxhh"      "aidhu1"       "aidhu2"
"aidhrs"       "aideft"
 [141] "lcohnpi"      "coh1bm"       "coh1by"       "coh1mr"       "coh1em"
"coh1ey"       "nmar"
 [148] "lmar1m"       "lmar1y"       "ladopt"       "lnadopt"      "lprnt"
"lnprnt"       "ch1by4"
```

```
 [155] "movy11"        "movy12"        "movy13"        "movy14"        "movy15"
"movy16"        "family"
 [162] "education"     "memploy"       "housing"       "area"          "moveoth_code"
"movdir"        "plnowm"
 [169] "plnowy4"       "mstatsam"      "lwwrong"       "mstatsamn"     "mstatch1"
"statcm1"       "statcy41"
 [176] "divchk1"       "divfin1"       "dvm1"          "dvy41"         "cmlstat1"
"mstatch2"      "statcm2"
 [183] "statcy42"      "divchk2"       "divfin2"       "dvm2"          "dvy42"
"cmlstat2"      "cohab"
 [190] "cohabn"        "lmcbm1"        "lmcby41"       "currpart1"     "lmspm1"
"lmspy41"       "lmcbm2"
 [197] "lmcby42"       "currpart2"     "lmspm2"        "lmspy42"       "lmcbm3"
"lmcby43"       "currpart3"
 [204] "lmspm3"        "lmspy43"       "father"        "nchild"        "preg"
"pregm1"        "pregy41"
 [211] "pregfert1"     "invitro1"      "pregout1"      "pregend1"      "endmnth1"
"pregsmoke1"    "smkmnth11"
 [218] "smkmnth21"     "smkmnth31"     "pregsmk11"     "pregsmk21"     "pregsmk31"
"lchmulti1"     "pregm2"
 [225] "pregy42"       "pregfert2"     "invitro2"      "pregout2"      "pregend2"
"endmnth2"      "pregsmoke2"
 [232] "smkmnth12"     "smkmnth22"     "smkmnth32"     "pregsmk12"     "pregsmk22"
"pregsmk32"     "lchmulti2"
 [239] "pregm3"        "pregy43"       "pregfert3"     "invitro3"      "pregout3"
"pregend3"      "endmnth3"
 [246] "pregsmoke3"    "smkmnth13"     "smkmnth23"     "smkmnth33"     "pregsmk13"
"pregsmk23"     "pregsmk33"
 [253] "lchmulti3"     "hcondno1"      "hcondns1"      "hcondno2"      "hcondns2"
"hcondno3"      "hcondns3"
 [260] "hcondno4"      "hcondns4"      "hcondno5"      "hcondns5"      "hcondno6"
"hcondns6"      "hcondno7"
 [267] "hcondns7"      "hcondno8"      "hcondns8"      "contft"        "ftendm"
"ftendy4"       "ftquals"
 [274] "ftedstartm1"   "ftedstarty41"  "ftedend1"      "ft2endm1"      "ft2endy41"
"ftedmor1"      "ftedstartm2"
 [281] "ftedstarty42"  "ftedend2"      "ft2endm2"      "ft2endy42"     "ftedmor2"
"qualnew1"      "qualnew2"
 [288] "qualnew3"      "qualnew4"      "qualnew5"      "qualnew6"      "qualnew7"
"qualnew8"      "qualnew9"
 [295] "qualnew10"     "qualnew11"     "qualnew13"     "qualnew15"     "qualnew16"
"qualnew17"     "qualnew18"
 [302] "qualnew19"     "qualnew20"     "qualnew21"     "qualnew22"     "qualnew23"
"qualnew24"     "qualnew25"
 [309] "qualnew26"     "qualnew27"     "qualnew28"     "qualnew29"     "qualnew30"
"qualnew31"     "trwho1"
 [316] "traindays1"    "trainhrs1"     "trainend1"     "trainpurp11"   "trainpurp21"
"trainpurp31"   "trainpurp41"
 [323] "trainpurp51"   "trainpurp61"   "trainpurp71"   "trainqual1"    "trwho2"
"traindays2"    "trainhrs2"
 [330] "trainend2"     "trainpurp12"   "trainpurp22"   "trainpurp32"   "trainpurp42"
"trainpurp52"   "trainpurp62"
 [337] "trainpurp72"   "trainqual2"    "trwho3"        "traindays3"    "trainhrs3"
"trainend3"     "trainpurp13"
 [344] "trainpurp23"   "trainpurp33"   "trainpurp43"   "trainpurp53"   "trainpurp63"
"trainpurp73"   "trainqual3"
 [351] "trainn"        "trqual1"       "trqual2"       "trqual3"       "trqual4"
"trqual5"       "trqual6"
 [358] "trqual7"       "trqual8"       "trqual9"       "trqual10"      "trqual11"
"trqual13"      "trqual15"
 [365] "trqual16"      "trqual17"      "trqual18"      "trqual19"      "trqual20"
"trqual21"      "trqual22"
 [372] "trqual23"      "trqual24"      "trqual25"      "trqual26"      "trqual27"
"trqual28"      "trqual29"
 [379] "trqual30"      "trqual31"      "notempchk"     "empchk"        "empstendd"
"empstendm"     "empstendy4"
 [386] "stendreas"     "nxtst"         "nxtstelse"     "cstat"         "nxtstendd"
"nxtstendm"     "nxtstendy4"
```

```
 [393] "jbsamr"        "wkplsam"       "samejob"       "matlv"         "matlvstd"
"matlvstm"      "matlvsty4"
 [400] "matlvendd"     "matlvendm"     "matlvendy4"    "jbendd"        "jbendm"
"jbendy4"       "jbendreas"
 [407] "cjob"          "nxtjbhrs"      "nxtjbes"       "nxtjbendd"     "nxtjbendm"
"nxtjbendy4"    "cjbatt"
 [414] "nextstat1"     "nextelse1"     "currstat1"     "nextjob1"      "currjob1"
"jobhours1"     "reasend1"
 [421] "jbatt1"        "statendd1"     "statendm1"     "statendy41"    "nextstat2"
"nextelse2"     "currstat2"
 [428] "nextjob2"      "currjob2"      "jobhours2"     "reasend2"      "jbatt2"
"statendd2"     "statendm2"
 [435] "statendy42"    "nextstat3"     "nextelse3"     "currstat3"     "nextjob3"
"currjob3"      "jobhours3"
 [442] "reasend3"      "jbatt3"        "statendd3"     "statendm3"     "statendy43"
"nextstat4"     "nextelse4"
 [449] "currstat4"     "nextjob4"      "currjob4"      "jobhours4"     "reasend4"
"jbatt4"        "statendd4"
 [456] "statendm4"     "statendy44"    "nextstat5"     "nextelse5"     "currstat5"
"nextjob5"      "currjob5"
 [463] "jobhours5"     "reasend5"      "jbatt5"        "statendd5"     "statendm5"
"statendy45"    "nextstat6"
 [470] "nextelse6"     "currstat6"     "nextjob6"      "currjob6"      "jobhours6"
"reasend6"      "jbatt6"
 [477] "statendd6"     "statendm6"     "statendy46"    "nextstat7"     "nextelse7"
"currstat7"     "nextjob7"
 [484] "currjob7"      "jobhours7"     "reasend7"      "jbatt7"        "statendd7"
"statendm7"     "statendy47"
 [491] "nextstat8"     "nextelse8"     "currstat8"     "nextjob8"      "currjob8"
"jobhours8"     "reasend8"
 [498] "jbatt8"        "statendd8"     "statendm8"     "statendy48"    "nextstat9"
"nextelse9"     "currstat9"
 [505] "nextjob9"      "currjob9"      "jobhours9"     "reasend9"      "jbatt9"
"statendd9"     "statendm9"
 [512] "statendy49"    "jboff"         "jboffy"        "jbterm1"       "jbterm2"
"jbsic07chk"    "jbsoc00chk"
 [519] "jbsempchk"     "jbsemp"        "jbbgd"         "jbbgm"         "jbbgy"
"jbmngrchk"     "jbmngr"
 [526] "jbsizechk"     "jbsize"        "jbsect"        "jbsectpub"     "jbhrs"
"jbot"          "jbotpd"
 [533] "paygwc"        "paynwc"        "payusl"        "payu"          "payuwc"
"payug"         "paytyp"
 [540] "ovtpay"        "extnsa"        "extrate"       "extrest"       "basnsa"
"basrate"       "basrest"
 [547] "ovtnsa"        "ovtrate"       "ovtrest"       "jbpl"          "jbttwt"
"worktrav"      "jsboss"
 [554] "jssize"        "jshrs"         "jstypeb"       "jsaccs"        "jspart"
"jsprbm"        "jsprby4"
 [561] "jsprem"        "jsprey4"       "jsprls"        "jsprtx"        "jsprni"
"jspayu"        "jspayw"
 [568] "jspytx"        "jspyni"        "jspl"          "jsttwt"        "jsttwtb"
"jsworktrav"    "jbsat"
 [575] "julk4wk"       "julkjb"        "jubgn"         "julk4x1"       "julk4x2"
"julk4x3"       "julk4x4"
 [582] "julk4x5"       "julk4x6"       "julk4x96"      "jbhad"         "jlendm"
"jlendy"        "jlsemp"
 [589] "jlboss"        "jlmngr"        "jlsize"        "eprosh"        "j2semp"
"j2hrs"         "j2pay"
 [596] "retchk"        "ageret"        "rtpro1"        "rtpro2"        "rtpro3"
"rtpro4"        "rtpro5"
 [603] "rtpro6"        "rtcon1"        "rtcon2"        "rtcon3"        "rtcon4"
"penmex"        "pppex"
 [610] "pppexm"        "sppen"         "rtexpjb"       "rtfnd1"        "rtfnd2"
"rtfnd3"        "rtfnd4"
 [617] "rtfnd5"        "rtfnd6"        "rtfnd7"        "rtfnd8"        "rtfnd9"
"rtfnd10"       "rtfnd96"
 [624] "retamt"        "retsuf"        "ccare"         "ccwork"        "benunemp1"
"benunemp2"     "benunemp96"
```

```
[631] "bendis1"        "bendis2"        "bendis3"        "bendis4"        "bendis5"
"bendis6"        "bendis7"
[638] "bendis8"        "bendis9"        "bendis10"       "bendis96"       "benpen1"
"benpen2"        "benpen3"
[645] "benpen4"        "benpen5"        "benpen6"        "benpen7"        "benpen8"
"benpen96"       "niserps"
[652] "bencb"          "benctc"         "benfam1"        "benfam2"        "benfam3"
"benfam4"        "benfam5"
[659] "benfam96"       "bentax1"        "bentax2"        "bentax3"        "bentax4"
"bentax5"        "bentax96"
[666] "benhou1"        "benhou2"        "benhou3"        "benhou4"        "benhou96"
"nfh01"          "nfh02"
[673] "nfh03"          "nfh04"          "nfh05"          "nfh06"          "nfh07"
"nfh08"          "nfh09"
[680] "nfh10"          "nfh11"          "nfh12"          "nfh13"          "nfh14"
"nfh15"          "nfh16"
[687] "nfh17"          "nfh18"          "nfh19"          "nfh20"          "nfh21"
"nfh22"          "nfh23"
[694] "nfh24"          "nfh25"          "nfh26"          "nfh27"          "nfh28"
"nfh29"          "nfh30"
[701] "nfh31"          "nfh32"          "nfh33"          "nfh34"          "nfh35"
"nfh36"          "nfh37"
[708] "nfh38"          "fiyrdb1"        "fiyrdb2"        "fiyrdb3"        "fiyrdb4"
"fiyrdb5"        "fiyrdb6"
[715] "vote2"          "vote3"          "vote4"          "vote5"          "perpolinf"
"colbens1"       "colbens2"
[722] "colbens3"       "civicduty"      "polcost"        "votenorm"       "perbfts"
"grpbfts"        "voteintent"
[729] "demorient"      "rphmob_code"    "rphwrk_code"    "remail_code"    "ctadd1_code"
"ctadd2_code"    "cttown_code"
[736] "ctcnty_code"    "ctpcode_code"   "cttel1_code"    "cttel2_code"    "ctemail_code"
"livesp"         "livewith"
[743] "lingua"         "jbbgdatd"       "jbbgdatm"       "jbbgdaty"       "prel"
"preason"        "pripn"
[750] "pjulk4wk"       "pjbptft"        "pjsptft"        "prearn"         "pbnft1"
"pbnft2"         "pbnft3"
[757] "pbnft4"         "pbnft5"         "pbnft6"         "pbnft7"         "pbnft8"
"pbnft9"         "pbnft10"
[764] "pbnft11"        "pbnft12"        "pbnft96"        "prfitb"         "ivinfnce"
"ivaffct11"      "ivaffct12"
[771] "ivaffct13"      "ivaffct14"      "ivaffct15"      "ivaffct17"      "ivaffct21"
"ivaffct22"      "ivaffct23"
[778] "ivaffct27"      "ivaffct30"      "ivaffct97"      "hgadoptm"       "hgadoptf"
"hgpart"         "ppsex"
[785] "fnpid"          "mnpid"          "pn1pid"         "pn1sex"         "pn2pid"
"pn2sex"         "pns1pid"
[792] "pns1sex"        "pns2pid"        "pns2sex"        "grfpid"         "grmpid"
"paygl"          "paynl"
[799] "jsprf"          "payg_dv"        "payn_dv"        "seearnnet_dv"   "ff_tel"
"ff_jbsemp"      "ff_jbmngr"
[806] "ff_jbsize"      "ff_oprlg"       "ff_oprlg0"      "ff_oprlg0ni"    "ff_ukborn"
"ff_yr2uk4"      "j1nssec8_dv"
[813] "manssec8_dv"    "adresp15_dv"    "ppid"           "sppid"          "fnspid"
"mnspid"         "nqfhigh_dv"
[820] "jbsoc00_cc"     "jbsoc10_cc"     "jbsic07_cc"     "jbes2000"       "jbrgsc_dv"
"jbnssec8_dv.x"  "jbisco88_cc"
[827] "jlsoc00_cc"     "jlsoc10_cc"     "jlsic07_cc"     "jles2000"       "jlrgsc_dv"
"jlnssec8_dv.x"  "jlisco88_cc"
[834] "pasoc90_cc"     "pasoc00_cc"     "pasoc10_cc"     "masoc90_cc"     "masoc00_cc"
"masoc10_cc"     "j1soc90_cc"
[841] "j1soc00_cc"     "j1soc10_cc"     "j2soc90_cc"     "j2soc00_cc"     "j2soc10_cc"
"payu_dv"        "jlnssec3_dv.x"
[848] "b_hidp"         "b_pno"          "b_splitnum"     "c_hidp"         "c_pno"
"c_splitnum"     "statina"
[855] "folic"          "folpreghr"      "resnhi"         "ehtch"          "ehtm"
"ehtft"          "ehtin"
[862] "nohtbc1"        "nohtbc2"        "nohtbc3"        "nohtbc4"        "nohtbc5"
"nohtbc6"        "nohtbc7"
```

```
 [869] "nohtbc8"     "hinrel"      "nobf1"       "nobf2"       "nobf3"
"resnwt"      "nowtbc1"
 [876] "nowtbc2"     "nowtbc3"     "nowtbc4"     "nowtbc5"     "nowtbc6"
"nowtbc7"     "nowtbc8"
 [883] "nowtbc9"     "ewtch"       "ewtkg"       "ewtst"       "ewtl"
"ynowh"       "whpnabm1"
 [890] "whpnabm2"    "whpnabm3"    "whpnabm4"    "whpnabm5"    "whpnabm6"
"whpnabm95"   "probwj"
 [897] "ynobp"       "nattbpd0"    "nattbpd1"    "nattbpd2"    "nattbpd3"
"nattbpd4"    "nattbpd95"
 [904] "mmgsprb1"    "mmgsprb2"    "mmgsprb3"    "mmgsprb95"   "noattlf0"
"noattlf1"    "noattlf2"
 [911] "noattlf3"    "noattlf4"    "noattlf5"    "noattlf95"   "lungsmhr"
"lunginhr"    "htfvc"
 [918] "htfev"       "htpef"       "htfevfvc"    "fev1pred"    "fvcpred"
"fev1fvcp"    "htfvc_sc"
 [925] "htfev_sc"    "htpef_sc"    "htfevfvc_sc" "fev1pred_sc" "fvcpred_sc"
"fev1fvcp_sc" "qualcdf0"
 [932] "qualcdf1"    "qualcdf2"    "qualcdf3"    "qualcdf4"    "qualcdf5"
"qualcdf6"    "qualcdf7"
 [939] "qualcdf95"   "qualab"      "nulllf0"     "nulllf1"     "nulllf2"
"nulllf3"     "nulllf4"
 [946] "nulllf5"     "nulllf6"     "nulllf95"    "hasurg"      "haeysurg"
"hastro"      "chestinf"
 [953] "inhaler"     "inhalhrs"    "lfwill"      "spirno"      "lftemp"
"noread"      "nlsatlf"
 [960] "htfvc2"      "ynolf"       "lfstand"     "lfresp"      "problf1"
"problf2"     "problf3"
 [967] "problf4"     "problf5"     "noattlf"     "ncgplf"      "ncguard"
"refbsc1"     "refbsc2"
 [974] "refbsc3"     "refbsc4"     "refbsc5"     "refbsc6"     "refbsc7"
"refbsc95"    "constorb"
 [981] "condna"      "samparm"     "samdifc1"    "samdifc2"    "samdifc3"
"samdifc4"    "samdifc5"
 [988] "samdifc6"    "samdifc95"   "nobsm1"      "nobsm2"      "nobsm3"
"nobsm95"     "vpsys"
 [995] "vphand"      "vparm"       "vpskin"      "vpalco"      "vpsam"
"vppress1"
 [ reached getOption("max.print") -- omitted 53 entries ]
[1] "------------Low Level Removal------------"
[1] "If a level is removed from a variable you wish to keep, reccomended to manually merge levels to
[1] "level 4 in hhorig.x removed, 4 observations found"
[1] "level 4 in memorig removed, 0 observations found"
[1] "level 5 in nch14resp removed, 3 observations found"
[1] "level 5 in nch415resp removed, 2 observations found"
[1] "level 5 in nchresp removed, 1 observations found"
[1] "level 5 in nchund18resp removed, 1 observations found"
[1] "level 6 in natch01 removed, 4 observations found"
[1] "level 7 in natch02 removed, 0 observations found"
[1] "level 7 in natch03 removed, 3 observations found"
[1] "level 7 in natch04 removed, 4 observations found"
[1] "level 6 in natch05 removed, 0 observations found"
[1] "level 7 in natch05 removed, 3 observations found"
[1] "level 8 in natch05 removed, 1 observations found"
[1] "level 8 in natch06 removed, 0 observations found"
[1] "level 9 in natch07 removed, 0 observations found"
[1] "level 5 in nnatch removed, 0 observations found"
[1] "level 7 in nnatch removed, 0 observations found"
[1] "level 4 in nadoptch removed, 2 observations found"
[1] "level 6 in adoptch01 removed, 2 observations found"
[1] "level 3 in adoptch02 removed, 2 observations found"
[1] "level 5 in adoptch02 removed, 3 observations found"
[1] "level 6 in adoptch02 removed, 1 observations found"
[1] "level 4 in adoptch03 removed, 0 observations found"
[1] "level 6 in adoptch03 removed, 0 observations found"
[1] "level 7 in adoptch03 removed, 0 observations found"
[1] "level 6 in adoptch04 removed, 0 observations found"
[1] "level 7 in adoptch04 removed, 0 observations found"
```

```
[1] "level 5 in nchunder16 removed, 0 observations found"
[1] "level 4 in nch10to15 removed, 2 observations found"
[1] "level 1 in allch01 removed, 1 observations found"
[1] "level 6 in allch01 removed, 2 observations found"
[1] "level 7 in allch02 removed, 0 observations found"
[1] "level 8 in allch02 removed, 0 observations found"
[1] "level 7 in allch03 removed, 0 observations found"
[1] "level 8 in allch03 removed, 0 observations found"
[1] "level 5 in allch04 removed, 1 observations found"
[1] "level 7 in allch04 removed, 0 observations found"
[1] "level 8 in allch04 removed, 0 observations found"
[1] "level 9 in allch04 removed, 0 observations found"
[1] "level 6 in allch05 removed, 0 observations found"
[1] "level 7 in allch05 removed, 0 observations found"
[1] "level 9 in jbstat.x removed, 1 observations found"
[1] "level 10 in jbstat.x removed, 4 observations found"
[1] "level 5 in relup removed, 4 observations found"
[1] "level 2 in nnewborn removed, 3 observations found"
[1] "level 1 in hcondn3 removed, 2 observations found"
[1] "level 1 in hcondn8 removed, 4 observations found"
[1] "level 1 in hcondn15 removed, 1 observations found"
[1] "level 1 in bensta3 removed, 2 observations found"
[1] "level 4 in ivcoop removed, 2 observations found"
[1] "level 4 in undqus removed, 4 observations found"
[1] "level 4 in hgbiom removed, 2 observations found"
[1] "level 5 in hgbiom removed, 1 observations found"
[1] "level 6 in hgbiom removed, 1 observations found"
[1] "level 4 in hgbiof removed, 1 observations found"
[1] "level 5 in hgbiof removed, 1 observations found"
[1] "level 6 in hgbiof removed, 1 observations found"
[1] "level 4 in pn1pno removed, 0 observations found"
[1] "level 5 in pn1pno removed, 0 observations found"
[1] "level 6 in pn1pno removed, 0 observations found"
[1] "level 3 in pn2pno removed, 1 observations found"
[1] "level 4 in pn2pno removed, 0 observations found"
[1] "level 5 in pn2pno removed, 0 observations found"
[1] "level 4 in pns1pno removed, 0 observations found"
[1] "level 5 in pns1pno removed, 0 observations found"
[1] "level 6 in pns1pno removed, 0 observations found"
[1] "level 3 in pns2pno removed, 0 observations found"
[1] "level 4 in pns2pno removed, 1 observations found"
[1] "level 5 in pns2pno removed, 0 observations found"
[1] "level 1 in fiyrinvinc_tc removed, 2 observations found"
[1] "level 1 in fibenothr_tc removed, 4 observations found"
[1] "level 9 in ff_jbstat removed, 2 observations found"
[1] "level 10 in ff_jbstat removed, 3 observations found"
[1] "level 1 in ff_bentype21 removed, 4 observations found"
[1] "level 1 in ff_bentype25 removed, 1 observations found"
[1] "level 1 in ff_bentype30 removed, 2 observations found"
[1] "level 1 in ff_bentype35 removed, 1 observations found"
[1] "level 1 in ff_bentype36 removed, 1 observations found"
[1] "level 1 in ff_bentype37 removed, 3 observations found"
[1] "level 1 in ngrp_dv removed, 3 observations found"
[1] "level 2 in nnssib_dv removed, 3 observations found"
[1] "level 3 in nnssib_dv removed, 3 observations found"
[1] "level 5 in nnssib_dv removed, 2 observations found"
[1] "level 2 in agegr13_dv removed, 2 observations found"
[1] "level 4 in buno_dv removed, 4 observations found"
[1] "level 5 in buno_dv removed, 2 observations found"
[1] "level 5 in nchild_dv removed, 0 observations found"
[1] "level 5 in hrpno removed, 3 observations found"
[1] "level 7 in ppno removed, 0 observations found"
[1] "level 8 in ppno removed, 0 observations found"
[1] "level 5 in sppno removed, 2 observations found"
[1] "level 7 in sppno removed, 0 observations found"
[1] "level 8 in sppno removed, 0 observations found"
[1] "level 4 in fnpno removed, 0 observations found"
[1] "level 5 in fnpno removed, 0 observations found"
```

```
[1] "level 6 in fnpno removed, 0 observations found"
[1] "level 4 in fnspno removed, 0 observations found"
[1] "level 5 in fnspno removed, 0 observations found"
[1] "level 6 in fnspno removed, 0 observations found"
[1] "level 4 in mnpno removed, 0 observations found"
[1] "level 5 in mnpno removed, 0 observations found"
[1] "level 6 in mnpno removed, 0 observations found"
[1] "level 4 in mnspno removed, 0 observations found"
[1] "level 5 in mnspno removed, 0 observations found"
[1] "level 6 in mnspno removed, 0 observations found"
[1] "level 1 in grfpno removed, 0 observations found"
[1] "level 1 in grmpno removed, 0 observations found"
[1] "level 2 in grmpno removed, 0 observations found"
[1] "level 2 in nnmpsp_dv removed, 4 observations found"
[1] "level 2 in ficode24 removed, 1 observations found"
[1] "level 1 in ficode25 removed, 0 observations found"
[1] "level 2 in ficode26 removed, 2 observations found"
[1] "level 2 in ficode28 removed, 2 observations found"
[1] "level 1 in ficode30 removed, 1 observations found"
[1] "level 1 in ficode37 removed, 2 observations found"
[1] "level 2 in ficode39 removed, 1 observations found"
[1] "level 4 in hhorig.y removed, 0 observations found"
[1] "level 3 in medcnjd removed, 1 observations found"
[1] "level 1 in medtyp13 removed, 1 observations found"
[1] "level 1 in difbpc4 removed, 3 observations found"
[1] "level 4 in nseqno removed, 4 observations found"
[1] "level 5 in nseqno removed, 0 observations found"
[1] "level 8 in elig removed, 2 observations found"
[1] "level 9 in elig removed, 1 observations found"
[1] "level 4 in wstokb removed, 2 observations found"
[1] "level 7 in hhsize removed, 1 observations found"
[1] "level 8 in hhsize removed, 1 observations found"
[1] "level 9 in hhsize removed, 0 observations found"
[1] "level 9 in jbstat.y removed, 0 observations found"
[1] "level 10 in jbstat.y removed, 0 observations found"
[1] "level 2 in nnsib_dv removed, 0 observations found"
[1] "level 3 in nnsib_dv removed, 0 observations found"
[1] "level 5 in nnsib_dv removed, 0 observations found"
[1] "level 1 in depchl_dv removed, 4 observations found"
[1] "level 5 in ndepchl_dv removed, 0 observations found"
[1] "level 6 in ndepchl_dv removed, 0 observations found"
[1] "136 total levels removed from 84 different variables. In total 158 observations deleted"
[1] "———————Variance 0 Check———————"
[1] "111 variables removed since their new variance was 0"
  [1] "ivfio"          "ioutcome"       "adstatus"       "natch05"        "natch06"
"natch07"        "natch08"
  [8] "natch09"        "natch10"        "natch11"        "natch12"        "natch13"
"natch14"        "natch15"
 [15] "natch16"        "adoptch04"      "adoptch05"      "adoptch06"      "adoptch07"
"adoptch08"      "adoptch09"
 [22] "adoptch10"      "adoptch11"      "adoptch12"      "adoptch13"      "adoptch14"
"adoptch15"      "adoptch16"
 [29] "allch05"        "allch06"        "allch07"        "allch08"        "allch09"
"allch10"        "allch11"
 [36] "allch12"        "allch13"        "allch14"        "allch15"        "allch16"
"hcondn3"        "hcondn8"
 [43] "hcondn15"       "bensta3"        "indmode"        "intdatd_if"     "intdatm_if"
"intdaty_if"     "doby_if"
 [50] "age_if"         "fiyrinvinc_tc" "fibenothr_tc"   "ff_ivlolw"      "ff_everint"
"ff_bentype21"   "ff_bentype25"
 [57] "ff_bentype30"   "ff_bentype31"   "ff_bentype32"   "ff_bentype35"   "ff_bentype36"
"ff_bentype37"   "ngrp_dv"
 [64] "grfpno"         "grmpno"         "fiyrinvinc_if" "ficode21"       "ficode25"
"ficode30"       "ficode31"
 [71] "ficode32"       "ficode35"       "ficode36"       "ficode37"       "tbmed"
"medtyp13"       "resphts"
 [78] "respwts"        "bfpcok"         "whintro"        "bpconst"        "consubx1"
"consubx2"       "consubx3"
```

```
 [85] "consubx4"       "consubx5"       "respbps"        "difbpc4"        "mmgswil"
"mmgsok"         "mmgssta"
 [92] "lungsurg"       "lungeye"        "lunghrt"        "lunghosp"       "lungex"
"lungtest"       "clotb"
 [99] "fit"            "dateok"         "htok"           "wtok"           "bmiok"
"elig"           "full1"
[106] "full2"          "full3"          "bprespc"        "wstokb"         "depchl_dv"
"scflag_dv"
[1] "————————Dummy Variables————————"
[1] "predictor variable count went from 388 to 826"
[1] "————————Variance 0 Check————————"
[1] "103 variables removed since their new variance was 0"
  [1] "pno.4"          "pno.5"          "hhorig.x.4"     "memorig.4"
"nch14resp.5"    "nch415resp.5"   "nchresp.5"
  [8] "nchund18resp.5" "natch01.6"      "natch02.7"      "natch03.7"
"natch04.7"      "nnatch.5"       "nnatch.7"
 [15] "nadoptch.4"     "adoptch01.6"    "adoptch02.3"    "adoptch02.5"
"adoptch02.6"    "adoptch03.4"    "adoptch03.6"
 [22] "adoptch03.7"    "nchunder16.5"   "nch10to15.4"    "allch01.1"
"allch01.6"      "allch02.7"      "allch02.8"
 [29] "allch03.7"      "allch03.8"      "allch04.5"      "allch04.7"
"allch04.8"      "allch04.9"      "jbstat.x.9"
 [36] "jbstat.x.10"    "relup.5"        "nnewborn.2"     "ivcoop.4"
"undqus.4"       "hgbiom.4"       "hgbiom.5"
 [43] "hgbiom.6"       "hgbiof.4"       "hgbiof.5"       "hgbiof.6"
"pn1pno.4"       "pn1pno.5"       "pn1pno.6"
 [50] "pn2pno.3"       "pn2pno.4"       "pn2pno.5"       "pns1pno.4"
"pns1pno.5"      "pns1pno.6"      "pns2pno.3"
 [57] "pns2pno.4"      "pns2pno.5"      "ff_jbstat.9"    "ff_jbstat.10"
"nnssib_dv.2"    "nnssib_dv.3"    "nnssib_dv.5"
 [64] "buno_dv.4"      "buno_dv.5"      "nchild_dv.5"    "hrpno.5"
"ppno.7"         "ppno.8"         "sppno.5"
 [71] "sppno.7"        "sppno.8"        "fnpno.4"        "fnpno.5"
"fnpno.6"        "fnspno.4"       "fnspno.5"
 [78] "fnspno.6"       "mnpno.4"        "mnpno.5"        "mnpno.6"
"mnspno.4"       "mnspno.5"       "mnspno.6"
 [85] "nnmpsp_dv.2"    "ficode24.2"     "ficode26.2"     "ficode28.2"
"ficode39.2"     "hhorig.y.4"     "medcnjd.3"
 [92] "nseqno.4"       "nseqno.5"       "hhsize.7"       "hhsize.8"
"hhsize.9"       "jbstat.y.9"     "jbstat.y.10"
 [99] "nnsib_dv.2"     "nnsib_dv.3"     "nnsib_dv.5"     "ndepchl_dv.5"
"ndepchl_dv.6"
[1] "————————K—Means————————"
[1] "15 clusters have been made for K—Means"
[1] "K—Means results as a table, the max value in each row is a simple way to define which cluster r

      0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16
 17  18  19  20  21  22  23  24  25  26  27  28  29
  1   0   0   0   0   1   1   2   0   0   4   1   1   2   1   0   0   1
  0   0   1   0   0   2   1   1   0   0   0   0   0
  2   2   0   2   0   0   2  14  15  17  12  14  10   9   6   2  11   2
  2   2   3   1   0   1   0   1   0   1   0   1   0
  3   0   0   0   1   1   2  12  10  13  11   8  13   4   1   3   4   0
  0   1   2   4   1   2   0   0   0   1   0   0   0
  4   0   0   1   0   0   2  19  12   9  10   8  10  11   4   1   5   4
  1   2   1   3   1   1   1   0   0   1   1   0   0
  5   0   1   1   2   5  10  51  56  41  39  41  29  38  15  14   7   5
  4   5   6   3   6   2   5   2   3   3   0   0   0
  6   0   1   2   3   1   3  25  23  25  22  21  16  22  12   5   3   8
  5   3   5   6   2   0   2   1   0   0   0   1   0
  7   1   1   0   2   4  14  53  40  37  50  39  36  37  17  10  20   9
  4   7   4  11   6   7   6   3   4   2   2   3   2
  8   0   0   0   0   2   4  15  15  10   6   8  12   5   8   1   1   4
  1   0   1   3   0   1   0   0   0   0   0   0   1
  9   0   0   1   1   2  10  14  24  18  27  16  18  13   8   6   6   3
  4   4   2   3   2   2   0   3   0   0   2   1   0
 10   3   2   0   4   4  21 100  88  86  78  67  68  76  36  30  21  17
 16   8   5   6  11   4   9   9   2   0   1   1   1
```

```
  11    0    2    0    0    1    8   34   32   19   24   19   20   19   11    2    6    3
2    3    2    4    2    4    1    3    2    0    1    0    0
  12    1    1    0    5    1    8   59   64   52   43   40   31   39   17   14   11    8
12   13    7    6    8    5    3    5    1    3    2    0    0
  13    0    0    0    0    2    1   12   11    6   14    7    4   12    6    2    2    3
1    2    2    1    2    1    0    1    0    0    2    0    0
  14    0    0    0    2    4    8   37   42   40   35   37   34   39   16   11    7    3
5    5    8    2    5    5    3    3    1    2    2    1    0
  15    0    0    0    1    1    4   18   25   18   17   20   13   25    8    2    7    2
3    4    2    1    5    2    3    3    0    1    1    1    0

      30   31   32   33   34   35   36
  1    0    0    0    0    0    0    0
  2    0    0    0    0    0    0    0
  3    0    0    0    0    0    1    0
  4    0    0    0    0    0    0    1
  5    0    0    0    0    1    0    0
  6    0    0    0    0    1    2    0
  7    2    0    1    0    0    2    1
  8    0    0    0    0    0    0    0
  9    0    0    0    0    0    2    0
  10   1    1    0    3    1    1    1
  11   0    0    0    0    1    1    0
  12   0    1    1    0    0    1    0
  13   0    0    1    0    2    0    1
  14   0    0    0    0    0    0    0
  15   0    0    1    1    1    0    1
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
 [1] "Cluster 1: Within MSE 118654435467980, Size 19"     "Cluster 2: Within MSE 3674751882295, Size
 [3] "Cluster 3: Within MSE 2476784669883, Size 95"       "Cluster 4: Within MSE 3377946689797, Size
 [5] "Cluster 5: Within MSE 3421540983047066, Size 395"   "Cluster 6: Within MSE 72255459388514, Size
 [7] "Cluster 7: Within MSE 3508538847650620, Size 437"   "Cluster 8: Within MSE 2832346318480, Size
 [9] "Cluster 9: Within MSE 3424643036131900, Size 192"   "Cluster 10: Within MSE 13406561361420700,
[11] "Cluster 11: Within MSE 3425935703097446, Size 226"  "Cluster 12: Within MSE 3536240200672631, S
[13] "Cluster 13: Within MSE 3157076800213, Size 98"      "Cluster 14: Within MSE 3034514229731599, S
[15] "Cluster 15: Within MSE 3443450205659846, Size 191"
[1] "Total between cluster MSE: 555758963252273088, Total within cluster MSE: 4774349627521316"
[1] "The K-Means model predicts exactly with an accuracy of 0.1333"
[1] "————————Correlation Checks————————"
[1] "map1 removed, correlated with 13 other variable(s)"
[1] "wave.3 removed, correlated with 10 other variable(s)"
[1] "map2 removed, correlated with 12 other variable(s)"
[1] "strata.x removed, correlated with 9 other variable(s)"
[1] "map3 removed, correlated with 10 other variable(s)"
[1] "weight removed, correlated with 7 other variable(s)"
[1] "mmgsnval removed, correlated with 7 other variable(s)"
[1] "ommapval removed, correlated with 10 other variable(s)"
[1] "hhorig.x.3 removed, correlated with 6 other variable(s)"
[1] "psu.x removed, correlated with 7 other variable(s)"
[1] "employ.2 removed, correlated with 7 other variable(s)"
[1] "estwt removed, correlated with 6 other variable(s)"
[1] "mmgsd1 removed, correlated with 6 other variable(s)"
[1] "pensioner_dv.2 removed, correlated with 6 other variable(s)"
[1] "dvage removed, correlated with 5 other variable(s)"
[1] "indpxus_lw removed, correlated with 5 other variable(s)"
[1] "mmgsn1 removed, correlated with 5 other variable(s)"
[1] "wtval removed, correlated with 5 other variable(s)"
[1] "memorig.3 removed, correlated with 4 other variable(s)"
[1] "birthy removed, correlated with 4 other variable(s)"
[1] "hgbiom.3 removed, correlated with 4 other variable(s)"
[1] "ficode1.1 removed, correlated with 5 other variable(s)"
[1] "mmgsd2 removed, correlated with 4 other variable(s)"
[1] "sys1 removed, correlated with 4 other variable(s)"
[1] "dias1 removed, correlated with 4 other variable(s)"
[1] "waist1 removed, correlated with 4 other variable(s)"
[1] "nchresp.3 removed, correlated with 4 other variable(s)"
[1] "pns2pno.2 removed, correlated with 4 other variable(s)"
[1] "pidp removed, correlated with 3 other variable(s)"
```

```
[1] "pno.2 removed, correlated with 3 other variable(s)"
[1] "relup.2 removed, correlated with 3 other variable(s)"
[1] "marstat.x.2 removed, correlated with 3 other variable(s)"
[1] "marstat.x.4 removed, correlated with 3 other variable(s)"
[1] "marstat.x.6 removed, correlated with 3 other variable(s)"
[1] "pn1pno.3 removed, correlated with 3 other variable(s)"
[1] "age_dv removed, correlated with 3 other variable(s)"
[1] "indinus_lw removed, correlated with 3 other variable(s)"
[1] "hhorig.y.3 removed, correlated with 3 other variable(s)"
[1] "mmgsn2 removed, correlated with 3 other variable(s)"
[1] "pulse1 removed, correlated with 3 other variable(s)"
[1] "sys2 removed, correlated with 3 other variable(s)"
[1] "dias2 removed, correlated with 3 other variable(s)"
[1] "waist2 removed, correlated with 3 other variable(s)"
[1] "ndepchl_dv.2 removed, correlated with 5 other variable(s)"
[1] "nchresp.1 removed, correlated with 3 other variable(s)"
[1] "nchresp.2 removed, correlated with 3 other variable(s)"
[1] "nchresp.4 removed, correlated with 3 other variable(s)"
[1] "nchunder16.4 removed, correlated with 3 other variable(s)"
[1] "allch03.5 removed, correlated with 3 other variable(s)"
[1] "ff_jbstat.2 removed, correlated with 3 other variable(s)"
[1] "ff_emplw.2 removed, correlated with 3 other variable(s)"
[1] "buno_dv.3 removed, correlated with 3 other variable(s)"
[1] "hidp removed, correlated with 2 other variable(s)"
[1] "hhorig.x.5 removed, correlated with 2 other variable(s)"
[1] "sex.2 removed, correlated with 2 other variable(s)"
[1] "natch04.6 removed, correlated with 2 other variable(s)"
[1] "jbstat.x.4 removed, correlated with 2 other variable(s)"
[1] "jbhas.2 removed, correlated with 2 other variable(s)"
[1] "btype5.1 removed, correlated with 2 other variable(s)"
[1] "marstat.x.3 removed, correlated with 2 other variable(s)"
[1] "hgbiom.1 removed, correlated with 2 other variable(s)"
[1] "hgbiom.2 removed, correlated with 2 other variable(s)"
[1] "hgbiof.1 removed, correlated with 2 other variable(s)"
[1] "hgbiof.2 removed, correlated with 2 other variable(s)"
[1] "hgbiof.3 removed, correlated with 2 other variable(s)"
[1] "respm16.2 removed, correlated with 2 other variable(s)"
[1] "pn2pno.2 removed, correlated with 2 other variable(s)"
[1] "pns1pno.3 removed, correlated with 2 other variable(s)"
[1] "fimngrs_tc.1 removed, correlated with 2 other variable(s)"
[1] "country.2 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.6 removed, correlated with 2 other variable(s)"
[1] "agegr13_dv.13 removed, correlated with 2 other variable(s)"
[1] "cohab_dv.1 removed, correlated with 2 other variable(s)"
[1] "mastat_dv.2 removed, correlated with 2 other variable(s)"
[1] "mastat_dv.4 removed, correlated with 2 other variable(s)"
[1] "mastat_dv.6 removed, correlated with 2 other variable(s)"
[1] "ppno.1 removed, correlated with 2 other variable(s)"
[1] "indinub_xw removed, correlated with 2 other variable(s)"
[1] "confage removed, correlated with 2 other variable(s)"
[1] "height removed, correlated with 2 other variable(s)"
[1] "mmgsd3 removed, correlated with 2 other variable(s)"
[1] "bswill.2 removed, correlated with 2 other variable(s)"
[1] "pulse2 removed, correlated with 2 other variable(s)"
[1] "sys3 removed, correlated with 2 other variable(s)"
[1] "dias3 removed, correlated with 2 other variable(s)"
[1] "hhsize.4 removed, correlated with 2 other variable(s)"
[1] "bmi removed, correlated with 2 other variable(s)"
[1] "psu.y removed, correlated with 2 other variable(s)"
[1] "fimnlabgrs_dv removed, correlated with 2 other variable(s)"
[1] "nchild_dv.3 removed, correlated with 2 other variable(s)"
[1] "nchild_dv.4 removed, correlated with 2 other variable(s)"
[1] "ndepchl_dv.3 removed, correlated with 2 other variable(s)"
[1] "pno.3 removed, correlated with 1 other variable(s)"
[1] "memorig.5 removed, correlated with 1 other variable(s)"
[1] "month removed, correlated with 1 other variable(s)"
[1] "nch14resp.2 removed, correlated with 1 other variable(s)"
[1] "nch14resp.3 removed, correlated with 1 other variable(s)"
```

```
[1] "natch01.2 removed, correlated with 1 other variable(s)"
[1] "natch03.5 removed, correlated with 1 other variable(s)"
[1] "nnatch.4 removed, correlated with 1 other variable(s)"
[1] "nadoptch.2 removed, correlated with 1 other variable(s)"
[1] "nadoptch.3 removed, correlated with 1 other variable(s)"
[1] "nchunder16.1 removed, correlated with 1 other variable(s)"
[1] "nchunder16.2 removed, correlated with 1 other variable(s)"
[1] "istrtdaty.2011 removed, correlated with 1 other variable(s)"
[1] "istrtdaty.2012 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.2 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.3 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.4 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.5 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.6 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.7 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.8 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.9 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.10 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.11 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.12 removed, correlated with 1 other variable(s)"
[1] "istrtdatd removed, correlated with 1 other variable(s)"
[1] "jbstat.x.2 removed, correlated with 1 other variable(s)"
[1] "jbstat.x.3 removed, correlated with 1 other variable(s)"
[1] "jbstat.x.5 removed, correlated with 1 other variable(s)"
[1] "jbstat.x.6 removed, correlated with 1 other variable(s)"
[1] "jbstat.x.7 removed, correlated with 1 other variable(s)"
[1] "jbstat.x.8 removed, correlated with 1 other variable(s)"
[1] "jbstat.x.97 removed, correlated with 1 other variable(s)"
[1] "sf1.x.2 removed, correlated with 1 other variable(s)"
[1] "sf1.x.3 removed, correlated with 1 other variable(s)"
[1] "sf1.x.4 removed, correlated with 1 other variable(s)"
[1] "sf1.x.5 removed, correlated with 1 other variable(s)"
[1] "health.x.2 removed, correlated with 1 other variable(s)"
[1] "relup.6 removed, correlated with 1 other variable(s)"
[1] "btype1.1 removed, correlated with 1 other variable(s)"
[1] "btype2.1 removed, correlated with 1 other variable(s)"
[1] "btype6.1 removed, correlated with 1 other variable(s)"
[1] "btype8.1 removed, correlated with 1 other variable(s)"
[1] "btype9.1 removed, correlated with 1 other variable(s)"
[1] "bensta2.1 removed, correlated with 1 other variable(s)"
[1] "bensta4.1 removed, correlated with 1 other variable(s)"
[1] "bensta5.1 removed, correlated with 1 other variable(s)"
[1] "bensta6.1 removed, correlated with 1 other variable(s)"
[1] "bensta7.1 removed, correlated with 1 other variable(s)"
[1] "fiyrdia removed, correlated with 1 other variable(s)"
[1] "marstat.x.5 removed, correlated with 1 other variable(s)"
[1] "respf16.2 removed, correlated with 1 other variable(s)"
[1] "scsf2a.2 removed, correlated with 1 other variable(s)"
[1] "scsf2b.2 removed, correlated with 1 other variable(s)"
[1] "scsf3a.5 removed, correlated with 1 other variable(s)"
[1] "istrtdathh removed, correlated with 1 other variable(s)"
[1] "pn1pno.1 removed, correlated with 1 other variable(s)"
[1] "pn1pno.2 removed, correlated with 1 other variable(s)"
[1] "fimnlabgrs_tc.1 removed, correlated with 1 other variable(s)"
[1] "j2paynet_dv removed, correlated with 1 other variable(s)"
[1] "ff_jbstat.4 removed, correlated with 1 other variable(s)"
[1] "ff_bentype09.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype10.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype13.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype18.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype22.1 removed, correlated with 1 other variable(s)"
[1] "npn_dv.1 removed, correlated with 1 other variable(s)"
[1] "npn_dv.2 removed, correlated with 1 other variable(s)"
[1] "nnssib_dv.1 removed, correlated with 1 other variable(s)"
[1] "urban_dv.x.2 removed, correlated with 1 other variable(s)"
[1] "xtra5min_dv.1 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.5 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.7 removed, correlated with 1 other variable(s)"
```

```
[1] "agegr5_dv.8 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.9 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.10 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.11 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.12 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.13 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.15 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.5 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.3 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.5 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.10 removed, correlated with 1 other variable(s)"
[1] "ppno.2 removed, correlated with 1 other variable(s)"
[1] "ppno.3 removed, correlated with 1 other variable(s)"
[1] "sppno.1 removed, correlated with 1 other variable(s)"
[1] "fnpno.1 removed, correlated with 1 other variable(s)"
[1] "fnpno.2 removed, correlated with 1 other variable(s)"
[1] "fnpno.3 removed, correlated with 1 other variable(s)"
[1] "mnpno.1 removed, correlated with 1 other variable(s)"
[1] "mnpno.2 removed, correlated with 1 other variable(s)"
[1] "mnpno.3 removed, correlated with 1 other variable(s)"
[1] "hiqual_dv.x.2 removed, correlated with 1 other variable(s)"
[1] "hiqual_dv.x.3 removed, correlated with 1 other variable(s)"
[1] "hiqual_dv.x.4 removed, correlated with 1 other variable(s)"
[1] "hiqual_dv.x.5 removed, correlated with 1 other variable(s)"
[1] "hiqual_dv.x.9 removed, correlated with 1 other variable(s)"
[1] "paygu_if.1 removed, correlated with 1 other variable(s)"
[1] "indpxub_xw removed, correlated with 1 other variable(s)"
[1] "frmnthimp_dv_total removed, correlated with 1 other variable(s)"
[1] "nsex.2 removed, correlated with 1 other variable(s)"
[1] "age removed, correlated with 1 other variable(s)"
[1] "region.2 removed, correlated with 1 other variable(s)"
[1] "medcnjd.2 removed, correlated with 1 other variable(s)"
[1] "bpmedc.1 removed, correlated with 1 other variable(s)"
[1] "estht removed, correlated with 1 other variable(s)"
[1] "bfpc removed, correlated with 1 other variable(s)"
[1] "floorc.2 removed, correlated with 1 other variable(s)"
[1] "cufsize.2 removed, correlated with 1 other variable(s)"
[1] "mmgsn3 removed, correlated with 1 other variable(s)"
[1] "nuroutc.12 removed, correlated with 1 other variable(s)"
[1] "pulse3 removed, correlated with 1 other variable(s)"
[1] "omdiaval removed, correlated with 1 other variable(s)"
[1] "omsyst removed, correlated with 1 other variable(s)"
[1] "ag16g10.76 removed, correlated with 1 other variable(s)"
[1] "hhsize.6 removed, correlated with 1 other variable(s)"
[1] "hhtype_dv.y.5 removed, correlated with 1 other variable(s)"
[1] "hhtype_dv.y.6 removed, correlated with 1 other variable(s)"
[1] "hhtype_dv.y.8 removed, correlated with 1 other variable(s)"
[1] "hhtype_dv.y.10 removed, correlated with 1 other variable(s)"
[1] "hhtype_dv.y.11 removed, correlated with 1 other variable(s)"
[1] "hhtype_dv.y.12 removed, correlated with 1 other variable(s)"
[1] "hhtype_dv.y.16 removed, correlated with 1 other variable(s)"
[1] "hhtype_dv.y.17 removed, correlated with 1 other variable(s)"
[1] "hhtype_dv.y.18 removed, correlated with 1 other variable(s)"
[1] "hhtype_dv.y.19 removed, correlated with 1 other variable(s)"
[1] "hhtype_dv.y.20 removed, correlated with 1 other variable(s)"
[1] "hhtype_dv.y.22 removed, correlated with 1 other variable(s)"
[1] "hhtype_dv.y.23 removed, correlated with 1 other variable(s)"
[1] "marstat.y.2 removed, correlated with 1 other variable(s)"
[1] "marstat.y.4 removed, correlated with 1 other variable(s)"
[1] "marstat.y.6 removed, correlated with 1 other variable(s)"
[1] "strata.y removed, correlated with 1 other variable(s)"
[1] "indnsub_lw removed, correlated with 1 other variable(s)"
[1] "fimnnet_dv removed, correlated with 1 other variable(s)"
[1] "respm16_dv.2 removed, correlated with 1 other variable(s)"
[1] "racel_dv removed, correlated with 1 other variable(s)"
[1] "wstval removed, correlated with 1 other variable(s)"
[1] "fimnlabnet_dv removed, correlated with 1 other variable(s)"
[1] "232 variables removed since they had high correlation coefs"
```

```
[1] "The final cleaned dataset has been completed at this stage and is stored under the name 'cleanDa
[1] "—————————Attempting a Train Test Split—————————"
[1] "Good train, test split found"
[1] "The working seed found was 3"
[1] "—————————kNN—————————"
[1] "120 neighbours considered for each test data point"
[1] "kNN results as a table, follow the diagonal for the correctly mapped clusters"
         real
predicted  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 30 3
        0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0
        1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0
        2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0
        3  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0
        4  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0
        5  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0
        6  3  1  2  1  5 12 66 50 47 33 29 25 21  9  5  8  3  2  0  3  0
4  2  1  1  0  0  0  0  0  0  0  0  0  0
        7  0  1  0  3  0  5 41 56 32 35 35 30 34 15  7 10  5  7  6  5  6
2  2  0  2  1  0  0  0  0  0  0  0  0  0
        8  0  0  0  0  1  0  1  5  5  5  4  5  6  4  2  0  2  0  2  0  2
2  1  0  2  0  0  0  0  0  0  0  0  0  1
        9  0  0  2  1  1  0  6  5  7  8 10 12 15  9  6  5  3  0  1  1  4
3  0  0  0  0  0  1  0  0  0  0  0  0  0
        10 0  0  0  0  0  0  0  3  0  3  2  3  6  3  4  2  0  0  0  0  1
2  0  1  0  0  0  0  0  0  0  0  0  0  0
        11 0  0  0  0  0  0  0  0  0  0  0  1  3  3  2  2  3  0  1  1  2
1  0  1  0  0  1  0  1  0  0  0  0  0  0
        12 0  0  0  0  0  0  0  0  3  1  2  4  7  5  4  1  5  3  2  3  1
4  4  0 10  1  3  1  1  1  3  1  2  2  0
        13 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  1  0  0  0  0  0  0  0  0  0
        14 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  1  0  0  0  0  0
        15 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0
        16 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  1  0  0
        17 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  1  0  0  0  0  0  0  0  0  0  0
        18 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0
        19 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0
        20 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0
        21 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0
        22 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0
        23 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0
        24 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0
        25 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0
        26 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0
        27 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0
 [ reached getOption("max.print") -- omitted 9 rows ]
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
[1] "The MSE of the predicted values are of 307.4145"
```

```
[1] "The kNN model predicts exactly with an accuracy of 0.1522"
[1] "——————CART prediction model——————"
n= 2858

node), split, n, deviance, yval
      * denotes terminal node

 1) root 2858 75523.0500 10.744930
   2) sf12mcs_dv>=-0.7520939 2294 22856.0800  9.085876
     4) scsf6c.5>=0.5 1253  6403.8530  7.690343 *
     5) scsf6c.5< 0.5 1041 11074.8100 10.765610
      10) sf12mcs_dv>=-0.2410445 739  5757.1940 10.167790 *
      11) sf12mcs_dv< -0.2410445 302  4407.2350 12.228480 *
   3) sf12mcs_dv< -0.7520939 564 20670.9700 17.492910
     6) sf12mcs_dv>=-1.993751 432  9318.8800 15.800930 *
     7) sf12mcs_dv< -1.993751 132  6067.8790 23.030300
      14) sf12mcs_dv>=-3.308415 110  4286.3730 21.845450 *
      15) sf12mcs_dv< -3.308415 22   854.9545 28.954550 *
[1] "Variable Importance"
    sf12mcs_dv       scsf4a.3       scsf6a.4       scsf4b.3       scsf6c.2
sf12pcs_dv       scsf6c.5       scsf6c.4
   42578.11420     8345.95640     7653.67757     7034.58099     6823.91101
5972.31925      5377.42040     4369.74219
      scsf4a.5       scsf4a.4       scsf6a.3       scsf4a.2       scsf6a.5
scsf4b.2       scsf6b.5       omronno
    1430.87940     1260.41362     1151.93540      760.60645      720.57453
640.51069      480.38302      126.34793
      scsf6c.3 ff_bentype33.1   fibenothr_if        j2has.2       undqus.3
     120.58016       42.11598       42.11598       42.11598       42.11598
[1] "The MSE of the predicted values are of 12.1791"
[1] "The CART model predicts exactly with accuracy of 0.1427"
[1] "——————Ordinary Linear Regression (Initial)——————"
[1] "The full model AIC is: 14678.3502"
[1] "——————Variance Inflation Factor Removal——————"
[1] "The variable sf12mcs_dv was removed since it had a VIF score of 15269.1062"
[1] "The variable natch02.4 was removed since it had a VIF score of 575.9812"
[1] "The variable nurdayy.2012 was removed since it had a VIF score of 394.6693"
[1] "The variable hhtype_dv.x.8 was removed since it had a VIF score of 187.5185"
[1] "The variable fibenothr_dv was removed since it had a VIF score of 176.0957"
[1] "The variable ag16g10.66 was removed since it had a VIF score of 172.3822"
[1] "The variable doby_dv was removed since it had a VIF score of 131.6172"
[1] "The variable single_dv.1 was removed since it had a VIF score of 128.0351"
[1] "The variable sf12pcs_dv was removed since it had a VIF score of 99.2468"
[1] "The variable allch01.3 was removed since it had a VIF score of 79.427"
[1] "The variable bmivg5.25 was removed since it had a VIF score of 78.0142"
[1] "The variable hhtype_dv.x.11 was removed since it had a VIF score of 74.9352"
[1] "The variable scsf4b.5 was removed since it had a VIF score of 70.9156"
[1] "The variable difbpc1.1 was removed since it had a VIF score of 69.2959"
[1] "The variable hrpid was removed since it had a VIF score of 64.9236"
[1] "The variable nnatch.2 was removed since it had a VIF score of 64.5787"
[1] "The variable rach16_dv.2 was removed since it had a VIF score of 59.3979"
[1] "The variable allch02.4 was removed since it had a VIF score of 51.4805"
[1] "The variable scsf4a.5 was removed since it had a VIF score of 47.4194"
[1] "The variable ag16g10.36 was removed since it had a VIF score of 43.6483"
[1] "The variable scsf6c.5 was removed since it had a VIF score of 42.1724"
[1] "The variable scsf7.5 was removed since it had a VIF score of 39.4099"
[1] "The variable natch01.3 was removed since it had a VIF score of 35.968"
[1] "The variable sclfsato.6 was removed since it had a VIF score of 23.573"
[1] "The variable scsf3b.5 was removed since it had a VIF score of 23.3694"
[1] "The variable hhtype_dv.x.10 was removed since it had a VIF score of 23.2533"
[1] "The variable jbstat.y.4 was removed since it had a VIF score of 22.3167"
[1] "The variable nnatch.3 was removed since it had a VIF score of 18.7694"
[1] "The variable fimnlabgrs_if was removed since it had a VIF score of 18.2044"
[1] "The variable agegr13_dv.6 was removed since it had a VIF score of 17.8372"
[1] "The variable natch02.3 was removed since it had a VIF score of 16.9204"
[1] "The variable nchild_dv.2 was removed since it had a VIF score of 15.0187"
[1] "The variable npensioner_dv.2 was removed since it had a VIF score of 14.1436"
[1] "The variable mnspno.2 was removed since it had a VIF score of 13.7325"
```

```
[1] "The variable numed2 was removed since it had a VIF score of 13.3533"
[1] "The variable sclfsat1.6 was removed since it had a VIF score of 13.1853"
[1] "The variable sppno.2 was removed since it had a VIF score of 13.0057"
[1] "The variable nunmpsp_dv.1 was removed since it had a VIF score of 12.5097"
[1] "The variable ieqmoecd_dv was removed since it had a VIF score of 12.3793"
[1] "The variable hhsize.3 was removed since it had a VIF score of 11.4995"
[1] "The variable hcondn96.1 was removed since it had a VIF score of 11.4649"
[1] "The variable hhtype_dv.x.12 was removed since it had a VIF score of 11.0096"
[1] "The variable ag16g10.46 was removed since it had a VIF score of 10.932"
[1] "The variable sclfsat2.6 was removed since it had a VIF score of 10.3445"
[1] "The variable nchund18resp.2 was removed since it had a VIF score of 10.0996"
[1] "45 variables removed from the Ordinary Linear Model since they have a VIF score higher than 10"
[1] "The full model AIC after VIF checks is: 14977.3855"
[1] "————————Backwards Selection————————"
[1] "50 out of 478 variables removed so far."
[1] "100 out of 478 variables removed so far."
[1] "150 out of 478 variables removed so far."
[1] "200 out of 478 variables removed so far."
[1] "250 out of 478 variables removed so far."
[1] "300 out of 478 variables removed so far."
[1] "345 out of 478 variables removed in backwards selection since they weren't significant at the 95
  [1] "mmgsdom.2"         "adoptch01.3"       "btype7.1"          "scsf3b.2"
"nchund18resp.4"    "intdaty_dv.2011"
  [7] "diur.1"            "lfout.4"           "intdatm_dv.5"      "ompulval"
"adoptch01.2"       "adoptch01.4"
 [13] "gor_dv.3"          "marstat.y.5"       "nch415resp.3"      "hiqual_dv.y.4"
"buno_dv.2"         "relwaitb.2"
 [19] "fimninvnet_dv"     "agegr13_dv.3"      "ficode38.1"        "mmgstp.3"
"mnspno.3"          "scsf2a.3"
 [25] "xtra5minosm_dv.1" "sclfsat2.7"        "relup.4"           "ficode2.3"
"nch5to15.4"        "hhtype_dv.x.17"
 [31] "marstat_dv.3"      "indnsub_xw"        "bsoute.2"          "intdatm_dv.11"
"vote6.4"           "medtyp3.1"
 [37] "sppno.4"           "nurdayy.2011"      "bmivg5.18"         "ff_bentype20.1"
"allch03.6"         "susp.2"
 [43] "ficode10.1"        "nurdaym.2"         "ag16g10.56"        "ficode19.1"
"finnow.3"          "aceinh.1"
 [49] "allch01.5"         "intdatm_dv.6"      "ivprsnt.2"         "pns1pno.1"
"gor_dv.4"          "ff_bentype03.1"
 [55] "ficode29.1"        "npensioner_dv.3"   "ficode5.1"         "ficode7.1"
"nch10to15.2"       "istrtdatmm"
 [61] "ficode15.1"        "strtnurhh"         "ienddathh"         "allch02.5"
"hrpno.2"           "qfhighfl_dv.1"
 [67] "jbstat.y.5"        "ff_bentype11.1"    "intdaty_dv.2012"   "sclfsat2.5"
"marstat_dv.6"      "ficode28.1"
 [73] "bswill.3"          "ff_jbstat.6"       "fimngrs_if"        "btype4.1"
"sclfsat1.3"        "medtyp9.1"
 [79] "aidhh.2"           "hcondn4.1"         "relhite.2"         "nch415resp.4"
"ff_bentype16.1"    "medtyp2.1"
 [85] "fimnlabnet_tc.1"   "finnow.2"          "natch01.4"         "nadoptch.1"
"adoptch01.5"       "scsf5.5"
 [91] "sclfsat7.2"        "medtyp11.1"        "lkmove.2"          "agegr13_dv.4"
"hhresp_dv.2"       "relup.3"
 [97] "scsf3b.4"          "nchild_dv.1"       "nch10to15.1"       "nurdayw.1"
"nurdayw.6"         "difbpc2.1"
[103] "gor_dv.5"          "hcondn7.1"         "rhland_code.1"     "ind5mus_lw"
"natch01.5"         "ficode3.2"
[109] "sampst.2"          "ficode4.3"         "scsf1.4"           "mmgstp.2"
"sf1.y.4"           "nchunder16.3"
[115] "hcondn16.1"        "bensta1.1"         "nurdaym.3"         "nurdaym.6"
"nurdaym.12"        "gor_dv.10"
[121] "indscub_xw"        "jbstat.y.3"        "ff_jbstat.7"       "ficode12.1"
"susp.3"            "intdatd_dv"
[127] "sclfsat7.3"        "difbpc5.1"         "sppno.3"           "hcondn11.1"
"trainany.2"        "natch03.6"
[133] "natch02.5"         "npensioner_dv.1"   "sclfsat1.4"        "ff_bentype38.1"
"ficode2.2"         "ff_bentype01.1"
[139] "hhorig.y.5"        "nurdayw.4"         "nurdayw.5"         "nurdayw.3"
```

```
"ficode29.2"        "scsf5.4"
[145] "pns1pno.2"         "beta.1"             "allch01.2"          "allch01.4"
"ff_bentype23.1"    "urban_dv.y.2"
[151] "nmpsp_dv.2"        "hcondn5.1"          "sclfsat1.5"         "hiqual_dv.y.2"
"nurdayw.2"         "nch5to15.2"
[157] "nnewborn.1"        "scsf5.2"            "scsf3a.4"           "scsf3a.3"
"nnsib_dv.1"        "nurdayd"
[163] "strtnurmm"         "fimnmisc_dv"        "medtyp8.1"          "istrtdatss"
"jbstat.y.2"        "nnmpsp_dv.1"
[169] "gor_dv.7"          "medtyp1.1"          "obpdrug.1"          "lipid.1"
"jbstat.y.6"        "gor_dv.8"
[175] "gor_dv.9"          "ficode11.1"         "ficode39.1"         "vote6.2"
"nch10to15.3"       "lfout.8"
[181] "agegr5_dv.14"      "ag16g20.76"         "nseqno.3"           "natch01.1"
"nseqno.2"          "ficode29.3"
[187] "ficode22.1"        "nch415resp.1"       "marstat_dv.5"       "hhtype_dv.x.18"
"intdatm_dv.12"     "agegr10_dv.3"
[193] "marstat.y.3"       "jbstat.y.97"        "jbiindb_dv"         "nmpsp_dv.1"
"ff_bentype14.1"    "intdatm_dv.4"
[199] "intdatm_dv.2"      "sclfsat2.2"         "lunginhl.2"         "lfout.2"
"bmivg5.40"         "bmivg5.30"
[205] "bmival"            "hcondn14.1"         "ficode9.1"          "hiqual_dv.y.9"
"fnspno.2"          "hhtype_dv.x.19"
[211] "ndepchl_dv.1"      "j2has.2"            "scsf1.2"            "cindtime"
"hcondn9.1"         "sclfsat7.4"
[217] "ff_bentype27.1"    "sf1.y.2"            "ff_jbstat.3"        "intdatm_dv.3"
"difbpc6.1"         "hhresp_dv.3"
[223] "hhtype_dv.x.22"    "ficode3.1"          "intdatm_dv.7"       "nch415resp.2"
"nch5to15.3"        "marstat_dv.2"
[229] "ff_bentype19.1"    "bfpcval"            "omdiast"            "gor_dv.6"
"btype96.1"         "medtyp5.1"
[235] "ficode23.1"        "nurdaym.4"          "ff_bentype02.1"     "nnatch.1"
"statins.2"         "sclfsat7.5"
[241] "sclfsat7.7"        "ivcoop.3"           "undqus.2"           "ivcoop.2"
"lenindintv"        "indin01_lw"
[247] "indin91_lw"        "undqus.3"           "ficode24.1"         "ff_bentype29.1"
"vote1.2"           "hrpno.3"
[253] "ficode13.1"        "gor_dv.2"           "omsysval"           "natch04.5"
"ff_jbstat.5"       "iron.1"
[259] "mmgstp.4"          "j2pay_if.1"         "ienddatmm"          "scsf2b.3"
"ff_bentype04.1"    "ficode4.1"
[265] "ff_bentype07.1"    "origadd.2"          "bsoute.5"           "bsoute.3"
"ff_bentype34.1"    "jbstat.y.7"
[271] "sclfsato.5"        "nurdaym.5"          "htval"              "medcnj.2"
"nurdaym.10"        "nurdaym.11"
[277] "nurdaym.8"         "intdatm_dv.9"       "nurdaym.7"          "intdatm_dv.10"
"intdatm_dv.8"      "nurdaym.9"
[283] "scsf5.3"           "ff_bentype05.1"     "bensta96.1"         "ficode26.1"
"floorc.3"          "agegr13_dv.7"
[289] "agegr13_dv.12"     "ag16g10.26"         "agegr13_dv.8"       "agegr13_dv.11"
"fimnsben_dv"       "ficode14.1"
[295] "ff_bentype15.1"    "ficode20.1"         "ficode6.1"          "airtemp"
"hiqual_dv.y.5"     "hiqual_dv.y.3"
[301] "natch02.6"         "hrpno.4"            "seearngrs_if.1"     "xpmove.2"
"hcondn1.1"         "ff_bentype33.1"
[307] "ff_bentype17.1"    "fimnprben_dv"       "mobuse.2"           "difbpc95.1"
"ftedany.2"         "ethnic"
[313] "hcondn10.1"        "scsf6b.2"           "ndepchl_dv.4"       "nch14resp.4"
"respf16_dv.2"      "nch5to15.1"
[319] "omronno"           "agegr13_dv.9"       "vote6.3"            "ff_jbstat.97"
"hhtype_dv.x.16"    "fiyrinvinc_dv"
[325] "hcondn6.1"         "cufsize.3"          "hcondn12.1"         "difbpc3.1"
"lfout.5"           "sclfsat2.3"
[331] "wjrel.3"           "btype3.1"           "hhsize.5"           "hhtype_dv.x.20"
"nchund18resp.3"    "ficode18.1"
[337] "mmgsdval"          "ficode27.1"         "ff_bentype28.1"     "fimnpen_dv"
"fimngrs_dv"        "hhtype_dv.x.6"
[343] "sclfsat2.4"        "allch02.6"          "aidxhh.2"
```

```
[1] "—————Ordinary Linear Regression (Improved)—————"

Call:
lm(formula = y ~ ., data = as.data.frame(x.data.linear))

Residuals:
     Min       1Q   Median       3Q      Max
-10.7962  -1.7516  -0.2211   1.5620  16.1534

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.33020    0.35257  17.954  < 2e-16 ***
sampst.3          0.79727    0.39228   2.032 0.042206 *
nch14resp.1       0.74493    0.23924   3.114 0.001867 **
nchund18resp.1   -0.69102    0.23192  -2.980 0.002912 **
natch03.4        -4.23213    0.82970  -5.101 3.61e-07 ***
adoptch03.5      -8.84472    2.25611  -3.920 9.06e-05 ***
allch02.3        -3.40745    0.74060  -4.601 4.40e-06 ***
allch03.4         4.24674    1.25028   3.397 0.000692 ***
hcondn2.1        -0.84385    0.32695  -2.581 0.009903 **
hcondn13.1        1.30553    0.50545   2.583 0.009847 **
hcondn17.1        3.31329    0.49015   6.760 1.68e-11 ***
finnow.4          0.82899    0.24678   3.359 0.000792 ***
finnow.5          1.31933    0.38834   3.397 0.000690 ***
finfut.2          0.63651    0.18952   3.359 0.000794 ***
finfut.3          0.35359    0.15992   2.211 0.027115 *
scsf1.3          -0.56190    0.16658  -3.373 0.000753 ***
scsf1.5          -1.68774    0.54632  -3.089 0.002026 **
scsf3a.2         -0.72460    0.32168  -2.253 0.024366 *
scsf3b.3         -0.42645    0.20055  -2.126 0.033562 *
scsf4a.2          2.16921    0.49123   4.416 1.04e-05 ***
scsf4a.3          1.81259    0.28999   6.251 4.72e-10 ***
scsf4a.4          0.94220    0.19932   4.727 2.39e-06 ***
scsf4b.2          3.36538    0.54751   6.147 9.05e-10 ***
scsf4b.3          1.77039    0.28990   6.107 1.16e-09 ***
scsf4b.4          0.56960    0.19056   2.989 0.002824 **
scsf6a.2          0.73935    0.24451   3.024 0.002519 **
scsf6a.3          1.66039    0.28067   5.916 3.71e-09 ***
scsf6a.4          3.43547    0.33415  10.281  < 2e-16 ***
scsf6a.5          6.69214    0.46691  14.333  < 2e-16 ***
scsf6b.3          0.30949    0.14363   2.155 0.031261 *
scsf6b.4          0.87030    0.21788   3.994 6.65e-05 ***
scsf6b.5          2.12966    0.36462   5.841 5.80e-09 ***
scsf6c.2          3.72927    0.36468  10.226  < 2e-16 ***
scsf6c.3          2.73785    0.20511  13.348  < 2e-16 ***
scsf6c.4          1.26347    0.14950   8.451  < 2e-16 ***
scsf7.2           3.55582    0.41519   8.564  < 2e-16 ***
scsf7.3           1.37732    0.22805   6.039 1.75e-09 ***
scsf7.4           0.73480    0.17839   4.119 3.91e-05 ***
sclfsat1.2       -0.61629    0.24899  -2.475 0.013377 *
sclfsat1.7       -0.54970    0.21058  -2.610 0.009094 **
sclfsat7.6       -0.39684    0.13534  -2.932 0.003393 **
sclfsato.2        1.51476    0.32636   4.641 3.62e-06 ***
sclfsato.3        1.71556    0.25027   6.855 8.77e-12 ***
sclfsato.4        0.90907    0.23100   3.935 8.51e-05 ***
sclfsato.7       -0.70874    0.19030  -3.724 0.000200 ***
ienddatss        -0.14096    0.05812  -2.425 0.015364 *
ff_jbstat.8      -2.48925    0.58698  -4.241 2.30e-05 ***
ff_bentype06.1   -5.01451    1.82943  -2.741 0.006164 **
ff_bentype08.1   11.33031    2.01105   5.634 1.94e-08 ***
ff_bentype12.1    1.35793    0.47605   2.853 0.004370 **
ff_bentype24.1   -2.00364    0.84862  -2.361 0.018292 *
ff_bentype26.1    1.18347    0.35725   3.313 0.000936 ***
agegr13_dv.10     0.58667    0.22748   2.579 0.009958 **
ppno.4            2.12410    1.03638   2.050 0.040502 *
ppno.5           12.01716    1.98974   6.040 1.75e-09 ***
fnspno.1         -3.31382    0.97999  -3.381 0.000731 ***
fnspno.3         -4.29641    1.55673  -2.760 0.005820 **
```

```
nunmpsp_dv.2     -4.60335     1.56159   -2.948 0.003226 **
paynu_if.1        1.05472     0.46148    2.285 0.022359 *
fibenothr_if      0.19464     0.05737    3.393 0.000702 ***
indscus_lw        0.13173     0.06099    2.160 0.030869 *
ficode2.1         0.33531     0.15118    2.218 0.026637 *
ficode4.2         1.33254     0.64775    2.057 0.039763 *
ficode8.1        -4.00725     1.62170   -2.471 0.013532 *
ficode16.1        1.61856     0.53313    3.036 0.002420 **
ficode17.1       -2.74751     0.79726   -3.446 0.000577 ***
ficode33.1        1.91523     0.77200    2.481 0.013166 *
ficode34.1        3.65604     1.59091    2.298 0.021632 *
calciumb.1       -0.61144     0.22343   -2.737 0.006248 **
medtyp4.1        -0.38856     0.17999   -2.159 0.030956 *
medtyp6.1        -0.54756     0.18452   -2.967 0.003029 **
medtyp7.1        -0.46523     0.21363   -2.178 0.029511 *
medtyp10.1       -0.57406     0.24765   -2.318 0.020520 *
medtyp12.1        0.83092     0.34337    2.420 0.015589 *
wjrel.2          -1.51314     0.63946   -2.366 0.018035 *
mmgsres.2        -2.16762     0.99155   -2.186 0.028892 *
lungsmok.2        0.47512     0.17809    2.668 0.007679 **
lfout.3          -0.52997     0.24380   -2.174 0.029810 *
ag16g10.86        4.56744     1.38886    3.289 0.001019 **
jbstat.y.8        2.81096     0.64928    4.329 1.55e-05 ***
sf1.y.3           0.45436     0.16893    2.690 0.007196 **
sf1.y.5           1.68351     0.46538    3.617 0.000303 ***
health.y.2       -0.41961     0.14059   -2.985 0.002863 **
sex_dv.2          0.34539     0.13326    2.592 0.009597 **
marstat_dv.4     -0.68630     0.33377   -2.056 0.039854 *
ethn_dv          -0.17149     0.06334   -2.707 0.006824 **
j2pay_dv         -0.15375     0.06117   -2.513 0.012014 *
hhtype_dv.x.5     2.98529     0.60930    4.900 1.02e-06 ***
hhtype_dv.x.23    3.84607     0.99529    3.864 0.000114 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 3.053 on 2769 degrees of freedom
Multiple R-squared:  0.6582,    Adjusted R-squared:  0.6473
F-statistic: 60.59 on 88 and 2769 DF,  p-value: < 2.2e-16

AIC: 14580.5955
MSE: 9.0324
[1] "The MSE of the predicted values are of 13.6724"
[1] "The Linear Model predicts exactly with accuracy of 0.1532"
[1] "The Linear Model predicts within a confidence interval with accuracy of 0.3935"
[1] "----------Elastic Net Regression----------"
492 x 1 sparse Matrix of class "dgCMatrix", with 14 entries
         names Estimate_Coefs
1  (Intercept)   11.156935944
2      scsf2a.3   -0.208869851
3      scsf2b.3   -0.033897896
4      scsf4a.5   -0.227007047
5      scsf6c.5   -0.131738572
6       scsf7.2    0.077119620
7    sclfsato.3    0.777888816
8    sclfsato.6   -0.047364764
9    sclfsato.7   -0.033959843
10       ppno.5    0.457688016
11   sf12mcs_dv   -3.472037434
12      sf1.y.5    0.255087363
13   health.y.2   -0.006912849
14   sf12pcs_dv   -0.537718127
[1] "The MSE of the predicted values of the best fit model is 10.2771"
[1] "The Alpha of the best fit model is 1"
[1] "The Elastic Net Model predicts exactly with accuracy of 0.1658"
[1] "----------Timer Results----------"
   user  system elapsed
 521.12    3.11  524.92
```

10.2.26 wSMergeNurseBlood console

```
[1] "————————Initial Checks————————"
[1] "8673779 NA cells were found across the entire dataset (59.05% of data as NA)"
[1] "————————Data Type Checks————————"
[1] "0 variables recoded since all their entries aren't numeric or NA"
[1] "NOTE: algorithim recodes categorical data alphabetically e.g (female = 1, male = 2)"
character(0)
[1] "————————Low Data Removal————————"
[1] "1041 variables removed since they had >= 'naPercent' (default 20%) NA values"
   [1] "pid.x"         "lvwhy"         "lvmthp"        "lvyrp"         "mvever"
"mvmnth"       "mvyr"
   [8] "mlstatchk"     "mlstat.x"      "drive"         "caruse"        "netuse"
"ukborn"       "plbornc"
  [15] "yr2uk4"        "citzn1"        "citzn2"        "citzn3"        "qfhigh"
"qualoc"       "qfvoc1"
  [22] "qfvoc2"        "qfvoc3"        "qfvoc4"        "qfvoc5"        "qfvoc6"
"qfvoc7"       "qfvoc8"
  [29] "qfvoc9"        "qfvoc10"       "qfvoc11"       "qfvoc12"       "qfvoc13"
"qfvoc14"      "qfvoc15"
  [36] "qfvoc96"       "school"        "scend"         "schlloc"       "schok"
"fenow"        "feend"
  [43] "j1none"        "j1semp"        "j1boss"        "j1mngr"        "edtype"
"edasp"        "fedlik"
  [50] "fednt"         "ocimpa"        "ocimpb"        "ocimpe"        "ocimpf"
"ocimpi"       "ocimpk"
  [57] "ocimpl"        "futra"         "futrb"         "futrc"         "futrd"
"futre"        "futrf"
  [64] "futrg"         "futrh"         "futri"         "futrj"         "futrk"
"futrl"        "paju"
  [71] "maju"          "pacob"         "payruk"        "macob"         "mayruk"
"natid1"       "natid2"
  [78] "natid3"        "natid4"        "natid5"        "natid6"        "natid97"
"racel"        "racelo_code"
  [85] "oprlg"         "oprlg0ni"      "nirel"         "niact"         "oprlg0"
"oprlg1"       "hospc1"
  [92] "hospdc1"       "hospc2"        "hospdc2"       "hospc3"        "hospdc3"
"hospc4"       "hospdc4"
  [99] "hospc5"        "hospdc5"       "hospc6"        "hospdc6"       "hospc7"
"hospdc7"      "hospc8"
 [106] "hospdc8"       "disdif1"       "disdif2"       "disdif3"       "disdif4"
"disdif5"      "disdif6"
 [113] "disdif7"       "disdif8"       "disdif9"       "disdif10"      "disdif11"
"disdif12"     "disdif96"
 [120] "aidhh"         "aidhua1"       "aidhua2"       "aidhua3"       "aidhua4"
"aidhua5"      "aidhua6"
 [127] "aidhua7"       "aidhua8"       "aidhua9"       "aidhua10"      "aidhua11"
"aidhua12"     "aidhua13"
 [134] "aidhua14"      "aidhua15"      "aidhua16"      "naidxhh"       "aidhu1"
"aidhu2"       "aidhrs"
 [141] "aideft"        "lcohnpi"       "coh1bm"        "coh1by"        "coh1mr"
"coh1em"       "coh1ey"
 [148] "nmar"          "lmar1m"        "lmar1y"        "ladopt"        "lnadopt"
"lprnt"        "lnprnt"
 [155] "ch1by4"        "movy11"        "movy12"        "movy13"        "movy14"
"movy15"       "movy16"
 [162] "family"        "education"     "memploy"       "housing"       "area"
"moveoth_code" "movdir"
 [169] "plnowm"        "plnowy4"       "mstatsam"      "lwwrong"       "mstatsamn"
"mstatch1"     "statcm1"
 [176] "statcy41"      "divchk1"       "divfin1"       "dvm1"          "dvy41"
"cmlstat1"     "mstatch2"
 [183] "statcm2"       "statcy42"      "divchk2"       "divfin2"       "dvm2"
"dvy42"        "cmlstat2"
 [190] "cohab"         "cohabn"        "lmcbm1"        "lmcby41"       "currpart1"
"lmspm1"       "lmspy41"
 [197] "lmcbm2"        "lmcby42"       "currpart2"     "lmspm2"        "lmspy42"
"lmcbm3"       "lmcby43"
```

```
 [204] "currpart3"      "lmspm3"         "lmspy43"        "father"         "nchild"
"preg"            "pregm1"
 [211] "pregy41"        "pregfert1"      "invitro1"       "pregout1"       "pregend1"
"endmnth1"        "pregsmoke1"
 [218] "smkmnth11"      "smkmnth21"      "smkmnth31"      "pregsmk11"      "pregsmk21"
"pregsmk31"       "lchmulti1"
 [225] "pregm2"         "pregy42"        "pregfert2"      "invitro2"       "pregout2"
"pregend2"        "endmnth2"
 [232] "pregsmoke2"     "smkmnth12"      "smkmnth22"      "smkmnth32"      "pregsmk12"
"pregsmk22"       "pregsmk32"
 [239] "lchmulti2"      "pregm3"         "pregy43"        "pregfert3"      "invitro3"
"pregout3"        "pregend3"
 [246] "endmnth3"       "pregsmoke3"     "smkmnth13"      "smkmnth23"      "smkmnth33"
"pregsmk13"       "pregsmk23"
 [253] "pregsmk33"      "lchmulti3"      "hcondno1"       "hcondns1"       "hcondno2"
"hcondns2"        "hcondno3"
 [260] "hcondns3"       "hcondno4"       "hcondns4"       "hcondno5"       "hcondns5"
"hcondno6"        "hcondns6"
 [267] "hcondno7"       "hcondns7"       "hcondno8"       "hcondns8"       "contft"
"ftendm"          "ftendy4"
 [274] "ftquals"        "ftedstartm1"    "ftedstarty41"   "ftedend1"       "ft2endm1"
"ft2endy41"       "ftedmor1"
 [281] "ftedstartm2"    "ftedstarty42"   "ftedend2"       "ft2endm2"       "ft2endy42"
"ftedmor2"        "qualnew1"
 [288] "qualnew2"       "qualnew3"       "qualnew4"       "qualnew5"       "qualnew6"
"qualnew7"        "qualnew8"
 [295] "qualnew9"       "qualnew10"      "qualnew11"      "qualnew13"      "qualnew15"
"qualnew16"       "qualnew17"
 [302] "qualnew18"      "qualnew19"      "qualnew20"      "qualnew21"      "qualnew22"
"qualnew23"       "qualnew24"
 [309] "qualnew25"      "qualnew26"      "qualnew27"      "qualnew28"      "qualnew29"
"qualnew30"       "qualnew31"
 [316] "trwho1"         "traindays1"     "trainhrs1"      "trainend1"      "trainpurp11"
"trainpurp21"     "trainpurp31"
 [323] "trainpurp41"    "trainpurp51"    "trainpurp61"    "trainpurp71"    "trainqual1"
"trwho2"          "traindays2"
 [330] "trainhrs2"      "trainend2"      "trainpurp12"    "trainpurp22"    "trainpurp32"
"trainpurp42"     "trainpurp52"
 [337] "trainpurp62"    "trainpurp72"    "trainqual2"     "trwho3"         "traindays3"
"trainhrs3"       "trainend3"
 [344] "trainpurp13"    "trainpurp23"    "trainpurp33"    "trainpurp43"    "trainpurp53"
"trainpurp63"     "trainpurp73"
 [351] "trainqual3"     "trainn"         "trqual1"        "trqual2"        "trqual3"
"trqual4"         "trqual5"
 [358] "trqual6"        "trqual7"        "trqual8"        "trqual9"        "trqual10"
"trqual11"        "trqual13"
 [365] "trqual15"       "trqual16"       "trqual17"       "trqual18"       "trqual19"
"trqual20"        "trqual21"
 [372] "trqual22"       "trqual23"       "trqual24"       "trqual25"       "trqual26"
"trqual27"        "trqual28"
 [379] "trqual29"       "trqual30"       "trqual31"       "notempchk"      "empchk"
"empstendd"       "empstendm"
 [386] "empstendy4"     "stendreas"      "nxtst"          "nxtstelse"      "cstat"
"nxtstendd"       "nxtstendm"
 [393] "nxtstendy4"     "jbsamr"         "wkplsam"        "samejob"        "matlv"
"matlvstd"        "matlvstm"
 [400] "matlvsty4"      "matlvendd"      "matlvendm"      "matlvendy4"     "jbendd"
"jbendm"          "jbendy4"
 [407] "jbendreas"      "cjob"           "nxtjbhrs"       "nxtjbes"        "nxtjbendd"
"nxtjbendm"       "nxtjbendy4"
 [414] "cjbatt"         "nextstat1"      "nextelse1"      "currstat1"      "nextjob1"
"currjob1"        "jobhours1"
 [421] "reasend1"       "jbatt1"         "statendd1"      "statendm1"      "statendy41"
"nextstat2"       "nextelse2"
 [428] "currstat2"      "nextjob2"       "currjob2"       "jobhours2"      "reasend2"
"jbatt2"          "statendd2"
 [435] "statendm2"      "statendy42"     "nextstat3"      "nextelse3"      "currstat3"
"nextjob3"        "currjob3"
```

```
[442] "jobhours3"     "reasend3"      "jbatt3"        "statendd3"     "statendm3"
"statendy43"    "nextstat4"
[449] "nextelse4"     "currstat4"     "nextjob4"      "currjob4"      "jobhours4"
"reasend4"      "jbatt4"
[456] "statendd4"     "statendm4"     "statendy44"    "nextstat5"     "nextelse5"
"currstat5"     "nextjob5"
[463] "currjob5"      "jobhours5"     "reasend5"      "jbatt5"        "statendd5"
"statendm5"     "statendy45"
[470] "nextstat6"     "nextelse6"     "currstat6"     "nextjob6"      "currjob6"
"jobhours6"     "reasend6"
[477] "jbatt6"        "statendd6"     "statendm6"     "statendy46"    "nextstat7"
"nextelse7"     "currstat7"
[484] "nextjob7"      "currjob7"      "jobhours7"     "reasend7"      "jbatt7"
"statendd7"     "statendm7"
[491] "statendy47"    "nextstat8"     "nextelse8"     "currstat8"     "nextjob8"
"currjob8"      "jobhours8"
[498] "reasend8"      "jbatt8"        "statendd8"     "statendm8"     "statendy48"
"nextstat9"     "nextelse9"
[505] "currstat9"     "nextjob9"      "currjob9"      "jobhours9"     "reasend9"
"jbatt9"        "statendd9"
[512] "statendm9"     "statendy49"    "jboff"         "jboffy"        "jbterm1"
"jbterm2"       "jbsic07chk"
[519] "jbsoc00chk"    "jbsempchk"     "jbsemp"        "jbbgd"         "jbbgm"
"jbbgy"         "jbmngrchk"
[526] "jbmngr"        "jbsizechk"     "jbsize"        "jbsect"        "jbsectpub"
"jbhrs"         "jbot"
[533] "jbotpd"        "paygwc"        "paynwc"        "payusl"        "payu"
"payuwc"        "payug"
[540] "paytyp"        "ovtpay"        "extnsa"        "extrate"       "extrest"
"basnsa"        "basrate"
[547] "basrest"       "ovtnsa"        "ovtrate"       "ovtrest"       "jbpl"
"jbttwt"        "worktrav"
[554] "jsboss"        "jssize"        "jshrs"         "jstypeb"       "jsaccs"
"jspart"        "jsprbm"
[561] "jsprby4"       "jsprem"        "jsprey4"       "jsprls"        "jsprtx"
"jsprni"        "jspayu"
[568] "jspayw"        "jspytx"        "jspyni"        "jspl"          "jsttwt"
"jsttwtb"       "jsworktrav"
[575] "jbsat"         "julk4wk"       "julkjb"        "jubgn"         "julk4x1"
"julk4x2"       "julk4x3"
[582] "julk4x4"       "julk4x5"       "julk4x6"       "julk4x96"      "jbhad"
"jlendm"        "jlendy"
[589] "jlsemp"        "jlboss"        "jlmngr"        "jlsize"        "eprosh"
"j2semp"        "j2hrs"
[596] "j2pay"         "retchk"        "ageret"        "rtpro1"        "rtpro2"
"rtpro3"        "rtpro4"
[603] "rtpro5"        "rtpro6"        "rtcon1"        "rtcon2"        "rtcon3"
"rtcon4"        "penmex"
[610] "pppex"         "pppexm"        "sppen"         "rtexpjb"       "rtfnd1"
"rtfnd2"        "rtfnd3"
[617] "rtfnd4"        "rtfnd5"        "rtfnd6"        "rtfnd7"        "rtfnd8"
"rtfnd9"        "rtfnd10"
[624] "rtfnd96"       "retamt"        "retsuf"        "ccare"         "ccwork"
"benunemp1"     "benunemp2"
[631] "benunemp96"    "bendis1"       "bendis2"       "bendis3"       "bendis4"
"bendis5"       "bendis6"
[638] "bendis7"       "bendis8"       "bendis9"       "bendis10"      "bendis96"
"benpen1"       "benpen2"
[645] "benpen3"       "benpen4"       "benpen5"       "benpen6"       "benpen7"
"benpen8"       "benpen96"
[652] "niserps"       "bencb"         "benctc"        "benfam1"       "benfam2"
"benfam3"       "benfam4"
[659] "benfam5"       "benfam96"      "bentax1"       "bentax2"       "bentax3"
"bentax4"       "bentax5"
[666] "bentax96"      "benhou1"       "benhou2"       "benhou3"       "benhou4"
"benhou96"      "nfh01"
[673] "nfh02"         "nfh03"         "nfh04"         "nfh05"         "nfh06"
"nfh07"         "nfh08"
```

```
 [680] "nfh09"          "nfh10"          "nfh11"          "nfh12"          "nfh13"
"nfh14"          "nfh15"
 [687] "nfh16"          "nfh17"          "nfh18"          "nfh19"          "nfh20"
"nfh21"          "nfh22"
 [694] "nfh23"          "nfh24"          "nfh25"          "nfh26"          "nfh27"
"nfh28"          "nfh29"
 [701] "nfh30"          "nfh31"          "nfh32"          "nfh33"          "nfh34"
"nfh35"          "nfh36"
 [708] "nfh37"          "nfh38"          "fiyrdb1"        "fiyrdb2"        "fiyrdb3"
"fiyrdb4"        "fiyrdb5"
 [715] "fiyrdb6"        "vote2"          "vote3"          "vote4"          "vote5"
"perpolinf"      "colbens1"
 [722] "colbens2"       "colbens3"       "civicduty"      "polcost"        "votenorm"
"perbfts"        "grpbfts"
 [729] "voteintent"     "demorient"      "rphmob_code"    "rphwrk_code"    "remail_code"
"ctadd1_code"    "ctadd2_code"
 [736] "cttown_code"    "ctcnty_code"    "ctpcode_code"   "cttel1_code"    "cttel2_code"
"ctemail_code"   "livesp"
 [743] "livewith"       "lingua"         "jbbgdatd"       "jbbgdatm"       "jbbgdaty"
"prel"           "preason"
 [750] "pripn"          "pjulk4wk"       "pjbptft"        "pjsptft"        "prearn"
"pbnft1"         "pbnft2"
 [757] "pbnft3"         "pbnft4"         "pbnft5"         "pbnft6"         "pbnft7"
"pbnft8"         "pbnft9"
 [764] "pbnft10"        "pbnft11"        "pbnft12"        "pbnft96"        "prfitb"
"ivinfnce"       "ivaffct11"
 [771] "ivaffct12"      "ivaffct13"      "ivaffct14"      "ivaffct15"      "ivaffct17"
"ivaffct21"      "ivaffct22"
 [778] "ivaffct23"      "ivaffct27"      "ivaffct30"      "ivaffct97"      "hgadoptm"
"hgadoptf"       "hgpart"
 [785] "ppsex"          "fnpid"          "mnpid"          "pn1pid"         "pn1sex"
"pn2pid"         "pn2sex"
 [792] "pns1pid"        "pns1sex"        "pns2pid"        "pns2sex"        "grfpid"
"grmpid"         "paygl"
 [799] "paynl"          "jsprf"          "payg_dv"        "payn_dv"        "seearnnet_dv"
"ff_tel"         "ff_jbsemp"
 [806] "ff_jbmngr"      "ff_jbsize"      "ff_oprlg"       "ff_oprlg0"      "ff_oprlg0ni"
"ff_ukborn"      "ff_yr2uk4"
 [813] "j1nssec8_dv"    "manssec8_dv"    "adresp15_dv"    "ppid"           "sppid"
"fnspid"         "mnspid"
 [820] "nqfhigh_dv"     "jbsoc00_cc"     "jbsoc10_cc"     "jbsic07_cc"     "jbes2000"
"jbrgsc_dv"      "jbnssec8_dv.x"
 [827] "jbisco88_cc"    "jlsoc00_cc"     "jlsoc10_cc"     "jlsic07_cc"     "jles2000"
"jlrgsc_dv"      "jlnssec8_dv.x"
 [834] "jlisco88_cc"    "pasoc90_cc"     "pasoc00_cc"     "pasoc10_cc"     "masoc90_cc"
"masoc00_cc"     "masoc10_cc"
 [841] "j1soc90_cc"     "j1soc00_cc"     "j1soc10_cc"     "j2soc90_cc"     "j2soc00_cc"
"j2soc10_cc"     "payu_dv"
 [848] "jlnssec3_dv.x"  "b_hidp.x"       "b_pno.x"        "b_splitnum.x"   "c_hidp.x"
"c_pno.x"        "c_splitnum.x"
 [855] "statina"        "folic"          "folpreghr"      "resnhi"         "ehtch"
"ehtm"           "ehtft"
 [862] "ehtin"          "nohtbc1"        "nohtbc2"        "nohtbc3"        "nohtbc4"
"nohtbc5"        "nohtbc6"
 [869] "nohtbc7"        "nohtbc8"        "hinrel"         "nobf1"          "nobf2"
"nobf3"          "resnwt"
 [876] "nowtbc1"        "nowtbc2"        "nowtbc3"        "nowtbc4"        "nowtbc5"
"nowtbc6"        "nowtbc7"
 [883] "nowtbc8"        "nowtbc9"        "ewtch"          "ewtkg"          "ewtst"
"ewtl"           "ynowh"
 [890] "whpnabm1"       "whpnabm2"       "whpnabm3"       "whpnabm4"       "whpnabm5"
"whpnabm6"       "whpnabm95"
 [897] "probwj"         "ynobp"          "nattbpd0"       "nattbpd1"       "nattbpd2"
"nattbpd3"       "nattbpd4"
 [904] "nattbpd95"      "mmgsprb1"       "mmgsprb2"       "mmgsprb3"       "mmgsprb95"
"noattlf0"       "noattlf1"
 [911] "noattlf2"       "noattlf3"       "noattlf4"       "noattlf5"       "noattlf95"
"lungsmok"       "lungsmhr"
```

```
 [918] "lunginhl"      "lunginhr"      "htfvc"         "htfev"         "htpef"
"htfevfvc"      "fev1pred"
 [925] "fvcpred"       "fev1fvcp"      "htfvc_sc"      "htfev_sc"      "htpef_sc"
"htfevfvc_sc"   "fev1pred_sc"
 [932] "fvcpred_sc"    "fev1fvcp_sc"   "qualcdf0"      "qualcdf1"      "qualcdf2"
"qualcdf3"      "qualcdf4"
 [939] "qualcdf5"      "qualcdf6"      "qualcdf7"      "qualcdf95"     "qualab"
"nulllf0"       "nulllf1"
 [946] "nulllf2"       "nulllf3"       "nulllf4"       "nulllf5"       "nulllf6"
"nulllf95"      "hasurg"
 [953] "haeysurg"      "hastro"        "chestinf"      "inhaler"       "inhalhrs"
"lfwill"        "spirno"
 [960] "lftemp"        "noread"        "nlsatlf"       "htfvc2"        "ynolf"
"lfstand"       "lfresp"
 [967] "problf1"       "problf2"       "problf3"       "problf4"       "problf5"
"noattlf"       "ncgplf"
 [974] "ncguard"       "refbsc1"       "refbsc2"       "refbsc3"       "refbsc4"
"refbsc5"       "refbsc6"
 [981] "refbsc7"       "refbsc95"      "nobsm1"        "nobsm2"        "nobsm3"
"nobsm95"       "wtpc"
 [988] "feet"          "mmgspr"        "antic"         "bfck2"         "omdiaval"
"ommapval"      "hyper2om"
 [995] "waist3"        "mlstat.y"      "jbnssec8_dv.y" "jbnssec3_dv.y" "jlnssec5_dv.y" "hyper1"
 [ reached getOption("max.print") -- omitted 41 entries ]
[1] "------------Low Level Removal------------"
[1] "If a level is removed from a variable you wish to keep, reccomended to manually merge levels to
[1] "level 5 in pno removed, 4 observations found"
[1] "level 7 in pno removed, 1 observations found"
[1] "level 5 in nch14resp removed, 2 observations found"
[1] "level 5 in nch415resp removed, 2 observations found"
[1] "level 5 in nchresp removed, 0 observations found"
[1] "level 5 in nchund18resp removed, 0 observations found"
[1] "level 6 in natch01 removed, 2 observations found"
[1] "level 7 in natch02 removed, 0 observations found"
[1] "level 7 in natch03 removed, 1 observations found"
[1] "level 8 in natch03 removed, 0 observations found"
[1] "level 5 in natch04 removed, 4 observations found"
[1] "level 7 in natch04 removed, 2 observations found"
[1] "level 6 in natch05 removed, 0 observations found"
[1] "level 7 in natch05 removed, 4 observations found"
[1] "level 8 in natch05 removed, 1 observations found"
[1] "level 8 in natch06 removed, 0 observations found"
[1] "level 9 in natch06 removed, 0 observations found"
[1] "level 9 in natch07 removed, 0 observations found"
[1] "level 5 in nnatch removed, 0 observations found"
[1] "level 6 in nnatch removed, 0 observations found"
[1] "level 7 in nnatch removed, 0 observations found"
[1] "level 4 in nadoptch removed, 1 observations found"
[1] "level 1 in adoptch01 removed, 2 observations found"
[1] "level 6 in adoptch01 removed, 1 observations found"
[1] "level 3 in adoptch02 removed, 2 observations found"
[1] "level 5 in adoptch02 removed, 3 observations found"
[1] "level 6 in adoptch02 removed, 1 observations found"
[1] "level 4 in adoptch03 removed, 0 observations found"
[1] "level 5 in adoptch03 removed, 4 observations found"
[1] "level 6 in adoptch03 removed, 0 observations found"
[1] "level 7 in adoptch03 removed, 0 observations found"
[1] "level 7 in adoptch04 removed, 0 observations found"
[1] "level 5 in nchunder16 removed, 0 observations found"
[1] "level 6 in nchunder16 removed, 0 observations found"
[1] "level 6 in nch5to15 removed, 0 observations found"
[1] "level 4 in nch10to15 removed, 1 observations found"
[1] "level 6 in allch01 removed, 3 observations found"
[1] "level 7 in allch02 removed, 0 observations found"
[1] "level 8 in allch02 removed, 0 observations found"
[1] "level 7 in allch03 removed, 0 observations found"
[1] "level 8 in allch03 removed, 0 observations found"
[1] "level 5 in allch04 removed, 0 observations found"
```

```
[1] "level 7 in allch04 removed, 0 observations found"
[1] "level 8 in allch04 removed, 0 observations found"
[1] "level 9 in allch04 removed, 0 observations found"
[1] "level 6 in allch05 removed, 0 observations found"
[1] "level 7 in allch05 removed, 0 observations found"
[1] "level 8 in allch05 removed, 0 observations found"
[1] "level 8 in allch06 removed, 0 observations found"
[1] "level 9 in allch06 removed, 0 observations found"
[1] "level 10 in jbstat.x removed, 2 observations found"
[1] "level 2 in nnewborn removed, 1 observations found"
[1] "level 1 in hcondn3 removed, 2 observations found"
[1] "level 1 in bensta3 removed, 2 observations found"
[1] "level 4 in ivcoop removed, 1 observations found"
[1] "level 5 in ivcoop removed, 1 observations found"
[1] "level 5 in undqus removed, 0 observations found"
[1] "level 4 in hgbiom removed, 3 observations found"
[1] "level 5 in hgbiom removed, 2 observations found"
[1] "level 6 in hgbiom removed, 1 observations found"
[1] "level 4 in hgbiof removed, 1 observations found"
[1] "level 7 in hgbiof removed, 1 observations found"
[1] "level 4 in pn1pno removed, 0 observations found"
[1] "level 5 in pn1pno removed, 0 observations found"
[1] "level 6 in pn1pno removed, 0 observations found"
[1] "level 3 in pn2pno removed, 4 observations found"
[1] "level 4 in pn2pno removed, 0 observations found"
[1] "level 7 in pn2pno removed, 0 observations found"
[1] "level 4 in pns1pno removed, 0 observations found"
[1] "level 5 in pns1pno removed, 0 observations found"
[1] "level 6 in pns1pno removed, 0 observations found"
[1] "level 3 in pns2pno removed, 0 observations found"
[1] "level 4 in pns2pno removed, 1 observations found"
[1] "level 7 in pns2pno removed, 0 observations found"
[1] "level 1 in fiyrinvinc_tc removed, 2 observations found"
[1] "level 9 in ff_jbstat removed, 2 observations found"
[1] "level 10 in ff_jbstat removed, 2 observations found"
[1] "level 1 in ff_bentype21 removed, 2 observations found"
[1] "level 1 in ff_bentype25 removed, 1 observations found"
[1] "level 1 in ff_bentype30 removed, 2 observations found"
[1] "level 1 in ff_bentype35 removed, 2 observations found"
[1] "level 1 in ff_bentype37 removed, 3 observations found"
[1] "level 1 in ngrp_dv removed, 1 observations found"
[1] "level 2 in nnssib_dv removed, 3 observations found"
[1] "level 3 in nnssib_dv removed, 0 observations found"
[1] "level 5 in nnssib_dv removed, 1 observations found"
[1] "level 6 in nnssib_dv removed, 1 observations found"
[1] "level 3 in country removed, 1 observations found"
[1] "level 2 in agegr13_dv removed, 2 observations found"
[1] "level 4 in buno_dv removed, 2 observations found"
[1] "level 5 in buno_dv removed, 0 observations found"
[1] "level 6 in buno_dv removed, 0 observations found"
[1] "level 5 in nchild_dv removed, 0 observations found"
[1] "level 6 in nchild_dv removed, 0 observations found"
[1] "level 5 in hrpno removed, 2 observations found"
[1] "level 6 in hrpno removed, 0 observations found"
[1] "level 10 in hrpno removed, 1 observations found"
[1] "level 5 in ppno removed, 4 observations found"
[1] "level 6 in ppno removed, 0 observations found"
[1] "level 5 in sppno removed, 0 observations found"
[1] "level 6 in sppno removed, 0 observations found"
[1] "level 4 in fnpno removed, 0 observations found"
[1] "level 6 in fnpno removed, 0 observations found"
[1] "level 4 in fnspno removed, 0 observations found"
[1] "level 6 in fnspno removed, 0 observations found"
[1] "level 4 in mnpno removed, 0 observations found"
[1] "level 5 in mnpno removed, 0 observations found"
[1] "level 6 in mnpno removed, 0 observations found"
[1] "level 4 in mnspno removed, 0 observations found"
[1] "level 5 in mnspno removed, 0 observations found"
```

```
[1] "level 6 in mnspno removed, 0 observations found"
[1] "level 1 in grmpno removed, 0 observations found"
[1] "level 2 in nnmpsp_dv removed, 4 observations found"
[1] "level 3 in ficode3 removed, 3 observations found"
[1] "level 3 in ficode4 removed, 4 observations found"
[1] "level 1 in ficode25 removed, 0 observations found"
[1] "level 2 in ficode26 removed, 1 observations found"
[1] "level 2 in ficode28 removed, 2 observations found"
[1] "level 1 in ficode30 removed, 1 observations found"
[1] "level 2 in ficode39 removed, 1 observations found"
[1] "level 1 in medtyp13 removed, 1 observations found"
[1] "level 1 in difbpc4 removed, 4 observations found"
[1] "level 1 in vppress3 removed, 4 observations found"
[1] "level 4 in nseqno removed, 4 observations found"
[1] "level 5 in nseqno removed, 0 observations found"
[1] "level 7 in nseqno removed, 0 observations found"
[1] "level 2 in dateok removed, 2 observations found"
[1] "level 8 in elig removed, 1 observations found"
[1] "level 9 in elig removed, 2 observations found"
[1] "level 4 in wstokb removed, 1 observations found"
[1] "level 7 in hhsize removed, 1 observations found"
[1] "level 8 in hhsize removed, 1 observations found"
[1] "level 9 in hhsize removed, 0 observations found"
[1] "level 10 in hhsize removed, 0 observations found"
[1] "level 10 in jbstat.y removed, 0 observations found"
[1] "level 2 in nnsib_dv removed, 0 observations found"
[1] "level 3 in nnsib_dv removed, 0 observations found"
[1] "level 6 in nnsib_dv removed, 0 observations found"
[1] "level 1 in depchl_dv removed, 4 observations found"
[1] "level 11 in gor_dv removed, 0 observations found"
[1] "140 total levels removed from 80 different variables. In total 138 observations deleted"
[1] "———Variance 0 Check———"
[1] "108 variables removed since their new variance was 0"
  [1] "ivfio"          "ioutcome"       "adstatus"       "natch05"        "natch06"
"natch07"        "natch08"
  [8] "natch09"        "natch10"        "natch11"        "natch12"        "natch13"
"natch14"        "natch15"
 [15] "natch16"        "adoptch03"      "adoptch04"      "adoptch05"      "adoptch06"
"adoptch07"      "adoptch08"
 [22] "adoptch09"      "adoptch10"      "adoptch11"      "adoptch12"      "adoptch13"
"adoptch14"      "adoptch15"
 [29] "adoptch16"      "allch05"        "allch06"        "allch07"        "allch08"
"allch09"        "allch10"
 [36] "allch11"        "allch12"        "allch13"        "allch14"        "allch15"
"allch16"        "hcondn3"
 [43] "hcondn15"       "bensta3"        "indmode"        "intdatd_if"     "intdatm_if"
"intdaty_if"     "doby_if"
 [50] "age_if"         "fiyrinvinc_tc"  "ff_ivlolw"      "ff_everint"     "ff_bentype21"
"ff_bentype25"   "ff_bentype30"
 [57] "ff_bentype31"   "ff_bentype32"   "ff_bentype35"   "ff_bentype36"   "ff_bentype37"
"ngrp_dv"        "grfpno"
 [64] "grmpno"         "fiyrinvinc_if"  "ficode21"       "ficode25"       "ficode30"
"ficode31"       "ficode32"
 [71] "ficode35"       "ficode36"       "tbmed"          "medtyp13"       "resphts"
"respwts"        "bfpcok"
 [78] "whintro"        "bpconst"        "respbps"        "difbpc4"        "mmgswil"
"mmgsok"         "mmgssta"
 [85] "lungsurg"       "lungeye"        "lunghrt"        "lunghosp"       "lungex"
"clotb"          "fit"
 [92] "bswill"         "constorb"       "samdifc6"       "vppress3"       "dateok"
"nuroutc"        "bsoute"
 [99] "htok"           "wtok"           "bmiok"          "elig"           "full1"
"full2"          "full3"
[106] "wstokb"         "depchl_dv"      "scflag_dv"
[1] "———Dummy Variables———"
[1] "predictor variable count went from 436 to 873"
[1] "———Variance 0 Check———"
[1] "105 variables removed since their new variance was 0"
```

```
  [1] "pno.4"            "pno.5"            "pno.7"            "nch14resp.5"
"nch415resp.5"    "nchresp.5"        "nchund18resp.5"
  [8] "natch01.6"        "natch02.7"        "natch03.7"        "natch03.8"
"natch04.5"        "natch04.7"        "nnatch.5"
 [15] "nnatch.6"         "nnatch.7"         "nadoptch.3"       "nadoptch.4"
"adoptch01.1"      "adoptch01.6"      "adoptch02.3"
 [22] "adoptch02.5"      "adoptch02.6"      "nchunder16.5"     "nchunder16.6"
"nch5to15.6"       "nch10to15.4"      "allch01.6"
 [29] "allch02.7"        "allch02.8"        "allch03.7"        "allch03.8"
"allch04.5"        "allch04.7"        "allch04.8"
 [36] "allch04.9"        "jbstat.x.10"      "nnewborn.2"       "ivcoop.4"
"ivcoop.5"         "undqus.5"         "hgbiom.4"
 [43] "hgbiom.5"         "hgbiom.6"         "hgbiof.4"         "hgbiof.7"
"pn1pno.4"         "pn1pno.5"         "pn1pno.6"
 [50] "pn2pno.3"         "pn2pno.4"         "pn2pno.7"         "pns1pno.4"
"pns1pno.5"        "pns1pno.6"        "pns2pno.3"
 [57] "pns2pno.4"        "pns2pno.7"        "ff_jbstat.9"      "ff_jbstat.10"
"nnssib_dv.2"      "nnssib_dv.3"      "nnssib_dv.5"
 [64] "nnssib_dv.6"      "country.3"        "buno_dv.4"        "buno_dv.5"
"buno_dv.6"        "nchild_dv.5"      "nchild_dv.6"
 [71] "hrpno.5"          "hrpno.6"          "hrpno.10"         "ppno.5"
"ppno.6"           "sppno.5"          "sppno.6"
 [78] "fnpno.4"          "fnpno.6"          "fnspno.4"         "fnspno.6"
"mnpno.4"          "mnpno.5"          "mnpno.6"
 [85] "mnspno.4"         "mnspno.5"         "mnspno.6"         "nnmpsp_dv.2"
"ficode3.3"        "ficode4.3"        "ficode26.2"
 [92] "ficode28.2"       "ficode39.2"       "nseqno.4"         "nseqno.5"
"nseqno.7"         "hhsize.7"         "hhsize.8"
 [99] "hhsize.9"         "hhsize.10"        "jbstat.y.10"      "nnsib_dv.2"
"nnsib_dv.3"       "nnsib_dv.6"       "gor_dv.11"
[1] "————————K—Means————————"
[1] "15 clusters have been made for K—Means"
[1] "K—Means results as a table, the max value in each row is a simple way to define which cluster re

      0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
  1   0  0  0  1  1  3  9  4  5  6  4  8  4  8  1  0  3  2  0  1  5  2  1
1  1  0  0  0  0  0  0  0  0  0  1  1  0
  2   3  0  2  1  2  3 27 27 20 28 27 17 19 14  6  9  3  4  3  5  4  2  2
2  2  0  1  3  0  1  0  0  1  0  2  0  1
  3   0  0  0  1  0  2  9  9  6 10 10  5 10  4  4  1  2  1  2  1  0  1  1
1  0  0  0  0  0  1  1  0  0  0  0  0  0
  4   0  0  1  0  3  4 40 43 24 28 28 21 36 16 10 12  5  3  5  5  4  5  3
5  2  0  2  1  3  1  0  0  0  1  1  0  2
  5   0  1  1  1  7  8 51 56 51 44 32 39 39 17 17  7  6  3  7  7  4  5  5
6  4  5  3  0  1  0  0  0  1  1  1  0  0
  6   0  1  2  1  1  4 15 18 16 19 23 14 16  6  6  4  6  4  5  4  4  0  1
3  1  0  0  0  2  0  0  0  0  1  0  1  0
  7   0  1  0  1  1  6 36 25 12 24 24 13 17 10  5  5  7  5  5  3  7  7  4
3  1  1  1  0  1  1  1  1  0  1  1  0  0
  8   0  0  1  1  5  3 27 22 29 20 16 19 15  8  5  5  4  3  5  3  7  0  4
0  1  0  2  0  1  0  0  0  0  0  0  1  0
  9   0  0  1  1  1  7 18 24 20 23 20 18 11  9  5  6  3  4  4  2  2  2  2
0  3  0  0  0  1  0  1  0  0  0  0  1  0
  10  0  0  1  2  1  8 37 24 31 25 25 18 12  9  6  7  8  4  3  3  4  4  2
3  4  0  0  1  1  0  0  1  0  2  0  1  0
  11  0  1  0  1  2  7 28 35 15 28 21 23 23  9  2  8  4  2  3  1  7  3  4
1  3  2  1  1  0  0  1  2  0  0  0  1  0
  12  1  1  1  4  1  8 73 68 55 36 30 36 38 17 16 11  9 14 11  7  5  6  5
3  6  2  0  2  0  0  0  1  2  0  0  1  0
  13  0  0  0  0  1  3 10 19 17 12 16 12 15  7  2  3  0  2  2  3  2  3  4
3  2  0  0  1  1  1  0  0  0  0  0  1  1
  14  0  0  0  3  4  7 55 47 35 49 52 44 48 18 12 18  9  7  3 10  7  6  8
4  1  2  3  2  1  0  3  1  1  0  0  1  0
  15  3  1  0  1  3 15 54 53 49 42 40 38 52 26 23 14 10 11  6  5  2  7  2
3  6  2  0  0  0  0  1  0  0  2  1  0  0
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
 [1] "Cluster 1: Within MSE 78821342862905, Size 72"      "Cluster 2: Within MSE 3444543717120264, Siz
 [3] "Cluster 3: Within MSE 66503242244584, Size 82"      "Cluster 4: Within MSE 9501886638661974, Siz
```

```
 [5] "Cluster 5: Within MSE 3407462903210946, Size 430"  "Cluster 6: Within MSE 69286036395703, Size
 [7] "Cluster 7: Within MSE 68744013791371, Size 230"    "Cluster 8: Within MSE 3459407403108016, Siz
 [9] "Cluster 9: Within MSE 3410523270071764, Size 189"  "Cluster 10: Within MSE 46642625076268, Size
[11] "Cluster 11: Within MSE 3455171679136462, Size 239" "Cluster 12: Within MSE 3552888384038894, S
[13] "Cluster 13: Within MSE 41379789807503, Size 143"   "Cluster 14: Within MSE 3452894573863155, S
[15] "Cluster 15: Within MSE 4509862097671165, Size 472"
[1] "Total between cluster MSE: 55411321407497 7152, Total within cluster MSE: 3248406560562017"
[1] "The K-Means model predicts exactly with an accuracy of 0.1336"
[1] "----------Correlation Checks----------"
[1] "wave.3 removed, correlated with 11 other variable(s)"
[1] "map1 removed, correlated with 10 other variable(s)"
[1] "strata.x removed, correlated with 10 other variable(s)"
[1] "hhorig.x.3 removed, correlated with 7 other variable(s)"
[1] "psu.x removed, correlated with 8 other variable(s)"
[1] "dvage removed, correlated with 7 other variable(s)"
[1] "pensioner_dv.2 removed, correlated with 9 other variable(s)"
[1] "weight removed, correlated with 7 other variable(s)"
[1] "mmgsnval removed, correlated with 7 other variable(s)"
[1] "map2 removed, correlated with 9 other variable(s)"
[1] "employ.2 removed, correlated with 7 other variable(s)"
[1] "estwt removed, correlated with 6 other variable(s)"
[1] "mmgsd1 removed, correlated with 6 other variable(s)"
[1] "memorig.3 removed, correlated with 5 other variable(s)"
[1] "indpxus_lw removed, correlated with 5 other variable(s)"
[1] "mmgsn1 removed, correlated with 5 other variable(s)"
[1] "wtval removed, correlated with 5 other variable(s)"
[1] "alb removed, correlated with 5 other variable(s)"
[1] "jbhas.2 removed, correlated with 5 other variable(s)"
[1] "age_dv removed, correlated with 5 other variable(s)"
[1] "map3 removed, correlated with 7 other variable(s)"
[1] "hgbiom.1 removed, correlated with 4 other variable(s)"
[1] "hgbiom.3 removed, correlated with 4 other variable(s)"
[1] "hhorig.y.3 removed, correlated with 4 other variable(s)"
[1] "mmgsd2 removed, correlated with 4 other variable(s)"
[1] "waist1 removed, correlated with 4 other variable(s)"
[1] "chol removed, correlated with 4 other variable(s)"
[1] "nchresp.2 removed, correlated with 4 other variable(s)"
[1] "pidp removed, correlated with 3 other variable(s)"
[1] "pno.2 removed, correlated with 3 other variable(s)"
[1] "hhorig.x.4 removed, correlated with 3 other variable(s)"
[1] "hhorig.x.5 removed, correlated with 3 other variable(s)"
[1] "sex.2 removed, correlated with 3 other variable(s)"
[1] "birthy removed, correlated with 3 other variable(s)"
[1] "relup.2 removed, correlated with 3 other variable(s)"
[1] "marstat.x.2 removed, correlated with 3 other variable(s)"
[1] "marstat.x.4 removed, correlated with 3 other variable(s)"
[1] "marstat.x.5 removed, correlated with 3 other variable(s)"
[1] "marstat.x.6 removed, correlated with 3 other variable(s)"
[1] "pn1pno.1 removed, correlated with 3 other variable(s)"
[1] "pn1pno.3 removed, correlated with 3 other variable(s)"
[1] "pn2pno.2 removed, correlated with 3 other variable(s)"
[1] "indinus_lw removed, correlated with 3 other variable(s)"
[1] "ficode1.1 removed, correlated with 4 other variable(s)"
[1] "confage removed, correlated with 4 other variable(s)"
[1] "mmgsn2 removed, correlated with 3 other variable(s)"
[1] "sys1 removed, correlated with 3 other variable(s)"
[1] "dias1 removed, correlated with 3 other variable(s)"
[1] "waist2 removed, correlated with 3 other variable(s)"
[1] "psu.y removed, correlated with 3 other variable(s)"
[1] "strata.y removed, correlated with 4 other variable(s)"
[1] "nchresp.1 removed, correlated with 3 other variable(s)"
[1] "ff_jbstat.2 removed, correlated with 3 other variable(s)"
[1] "ff_emplw.2 removed, correlated with 3 other variable(s)"
[1] "hdl removed, correlated with 3 other variable(s)"
[1] "hidp removed, correlated with 2 other variable(s)"
[1] "memorig.4 removed, correlated with 2 other variable(s)"
[1] "memorig.5 removed, correlated with 2 other variable(s)"
[1] "nch14resp.3 removed, correlated with 2 other variable(s)"
```

```
[1] "nchunder16.3 removed, correlated with 2 other variable(s)"
[1] "nchunder16.4 removed, correlated with 2 other variable(s)"
[1] "jbstat.x.4 removed, correlated with 2 other variable(s)"
[1] "sf1.x.5 removed, correlated with 2 other variable(s)"
[1] "btype5.1 removed, correlated with 2 other variable(s)"
[1] "marstat.x.3 removed, correlated with 2 other variable(s)"
[1] "hgbiom.2 removed, correlated with 2 other variable(s)"
[1] "hgbiof.1 removed, correlated with 2 other variable(s)"
[1] "hgbiof.2 removed, correlated with 2 other variable(s)"
[1] "hgbiof.3 removed, correlated with 2 other variable(s)"
[1] "respm16.2 removed, correlated with 2 other variable(s)"
[1] "pns1pno.1 removed, correlated with 2 other variable(s)"
[1] "pns1pno.3 removed, correlated with 2 other variable(s)"
[1] "pns2pno.2 removed, correlated with 2 other variable(s)"
[1] "fimngrs_tc.1 removed, correlated with 2 other variable(s)"
[1] "country.2 removed, correlated with 2 other variable(s)"
[1] "agegr5_dv.6 removed, correlated with 2 other variable(s)"
[1] "cohab_dv.1 removed, correlated with 2 other variable(s)"
[1] "mastat_dv.2 removed, correlated with 2 other variable(s)"
[1] "mastat_dv.4 removed, correlated with 2 other variable(s)"
[1] "mastat_dv.5 removed, correlated with 2 other variable(s)"
[1] "mastat_dv.6 removed, correlated with 2 other variable(s)"
[1] "ppno.1 removed, correlated with 2 other variable(s)"
[1] "indinub_xw removed, correlated with 2 other variable(s)"
[1] "height removed, correlated with 2 other variable(s)"
[1] "mmgsd3 removed, correlated with 2 other variable(s)"
[1] "pulse1 removed, correlated with 2 other variable(s)"
[1] "sys2 removed, correlated with 2 other variable(s)"
[1] "dias2 removed, correlated with 2 other variable(s)"
[1] "bmi removed, correlated with 2 other variable(s)"
[1] "indnsub_lw removed, correlated with 2 other variable(s)"
[1] "fimnlabgrs_dv removed, correlated with 2 other variable(s)"
[1] "trig removed, correlated with 2 other variable(s)"
[1] "single_dv.1 removed, correlated with 2 other variable(s)"
[1] "pno.3 removed, correlated with 1 other variable(s)"
[1] "month removed, correlated with 1 other variable(s)"
[1] "nch14resp.4 removed, correlated with 1 other variable(s)"
[1] "nchresp.3 removed, correlated with 1 other variable(s)"
[1] "natch03.5 removed, correlated with 1 other variable(s)"
[1] "natch04.6 removed, correlated with 1 other variable(s)"
[1] "nadoptch.2 removed, correlated with 1 other variable(s)"
[1] "nchunder16.1 removed, correlated with 1 other variable(s)"
[1] "nchunder16.2 removed, correlated with 1 other variable(s)"
[1] "allch03.5 removed, correlated with 1 other variable(s)"
[1] "allch04.6 removed, correlated with 1 other variable(s)"
[1] "istrtdaty.2011 removed, correlated with 1 other variable(s)"
[1] "istrtdaty.2012 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.2 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.3 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.4 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.5 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.6 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.7 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.8 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.9 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.10 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.11 removed, correlated with 1 other variable(s)"
[1] "istrtdatm.12 removed, correlated with 1 other variable(s)"
[1] "istrtdatd removed, correlated with 1 other variable(s)"
[1] "jbstat.x.2 removed, correlated with 1 other variable(s)"
[1] "jbstat.x.3 removed, correlated with 1 other variable(s)"
[1] "jbstat.x.5 removed, correlated with 1 other variable(s)"
[1] "jbstat.x.6 removed, correlated with 1 other variable(s)"
[1] "jbstat.x.7 removed, correlated with 1 other variable(s)"
[1] "jbstat.x.8 removed, correlated with 1 other variable(s)"
[1] "jbstat.x.97 removed, correlated with 1 other variable(s)"
[1] "sf1.x.2 removed, correlated with 1 other variable(s)"
[1] "sf1.x.3 removed, correlated with 1 other variable(s)"
```

```
[1] "sf1.x.4 removed, correlated with 1 other variable(s)"
[1] "health.x.2 removed, correlated with 1 other variable(s)"
[1] "btype2.1 removed, correlated with 1 other variable(s)"
[1] "btype6.1 removed, correlated with 1 other variable(s)"
[1] "btype8.1 removed, correlated with 1 other variable(s)"
[1] "btype9.1 removed, correlated with 1 other variable(s)"
[1] "bensta2.1 removed, correlated with 1 other variable(s)"
[1] "bensta4.1 removed, correlated with 1 other variable(s)"
[1] "bensta5.1 removed, correlated with 1 other variable(s)"
[1] "bensta6.1 removed, correlated with 1 other variable(s)"
[1] "bensta7.1 removed, correlated with 1 other variable(s)"
[1] "fiyrdia removed, correlated with 1 other variable(s)"
[1] "respf16.2 removed, correlated with 1 other variable(s)"
[1] "scsf1.5 removed, correlated with 1 other variable(s)"
[1] "scsf2a.2 removed, correlated with 1 other variable(s)"
[1] "scsf3a.5 removed, correlated with 1 other variable(s)"
[1] "istrtdathh removed, correlated with 1 other variable(s)"
[1] "pn1pno.2 removed, correlated with 1 other variable(s)"
[1] "fimnlabgrs_tc.1 removed, correlated with 1 other variable(s)"
[1] "j2paynet_dv removed, correlated with 1 other variable(s)"
[1] "ff_jbstat.4 removed, correlated with 1 other variable(s)"
[1] "ff_bentype09.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype10.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype13.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype18.1 removed, correlated with 1 other variable(s)"
[1] "ff_bentype22.1 removed, correlated with 1 other variable(s)"
[1] "npn_dv.1 removed, correlated with 1 other variable(s)"
[1] "npn_dv.2 removed, correlated with 1 other variable(s)"
[1] "nnssib_dv.1 removed, correlated with 1 other variable(s)"
[1] "urban_dv.x.2 removed, correlated with 1 other variable(s)"
[1] "xtra5min_dv.1 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.5 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.7 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.8 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.9 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.10 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.11 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.12 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.13 removed, correlated with 1 other variable(s)"
[1] "agegr5_dv.15 removed, correlated with 1 other variable(s)"
[1] "agegr13_dv.5 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.3 removed, correlated with 1 other variable(s)"
[1] "mastat_dv.10 removed, correlated with 1 other variable(s)"
[1] "ppno.2 removed, correlated with 1 other variable(s)"
[1] "ppno.3 removed, correlated with 1 other variable(s)"
[1] "ppno.4 removed, correlated with 1 other variable(s)"
[1] "sppno.1 removed, correlated with 1 other variable(s)"
[1] "fnpno.1 removed, correlated with 1 other variable(s)"
[1] "fnpno.2 removed, correlated with 1 other variable(s)"
[1] "fnpno.3 removed, correlated with 1 other variable(s)"
[1] "mnpno.1 removed, correlated with 1 other variable(s)"
[1] "mnpno.2 removed, correlated with 1 other variable(s)"
[1] "mnpno.3 removed, correlated with 1 other variable(s)"
[1] "hiqual_dv.x.2 removed, correlated with 1 other variable(s)"
[1] "hiqual_dv.x.3 removed, correlated with 1 other variable(s)"
[1] "hiqual_dv.x.4 removed, correlated with 1 other variable(s)"
[1] "hiqual_dv.x.5 removed, correlated with 1 other variable(s)"
[1] "hiqual_dv.x.9 removed, correlated with 1 other variable(s)"
[1] "paygu_if.1 removed, correlated with 1 other variable(s)"
[1] "ind5mus_lw removed, correlated with 1 other variable(s)"
[1] "indpxub_xw removed, correlated with 1 other variable(s)"
[1] "frmnthimp_dv_total removed, correlated with 1 other variable(s)"
[1] "hhorig.y.4 removed, correlated with 1 other variable(s)"
[1] "hhorig.y.5 removed, correlated with 1 other variable(s)"
[1] "nsex.2 removed, correlated with 1 other variable(s)"
[1] "age removed, correlated with 1 other variable(s)"
[1] "region.2 removed, correlated with 1 other variable(s)"
[1] "medcnjd.2 removed, correlated with 1 other variable(s)"
```

```
[1] "bpmedc.1 removed, correlated with 1 other variable(s)"
[1] "estht removed, correlated with 1 other variable(s)"
[1] "bfpc removed, correlated with 1 other variable(s)"
[1] "floorc.2 removed, correlated with 1 other variable(s)"
[1] "consubx5.1 removed, correlated with 1 other variable(s)"
[1] "cufsize.2 removed, correlated with 1 other variable(s)"
[1] "mmgsdom.2 removed, correlated with 1 other variable(s)"
[1] "mmgsn3 removed, correlated with 1 other variable(s)"
[1] "lungtest.2 removed, correlated with 1 other variable(s)"
[1] "samparm.2 removed, correlated with 1 other variable(s)"
[1] "samparm.3 removed, correlated with 1 other variable(s)"
[1] "nurdaym.2 removed, correlated with 1 other variable(s)"
[1] "nurdaym.3 removed, correlated with 1 other variable(s)"
[1] "nurdaym.4 removed, correlated with 1 other variable(s)"
[1] "nurdaym.5 removed, correlated with 1 other variable(s)"
[1] "nurdaym.6 removed, correlated with 1 other variable(s)"
[1] "nurdaym.7 removed, correlated with 1 other variable(s)"
[1] "nurdaym.8 removed, correlated with 1 other variable(s)"
[1] "nurdaym.9 removed, correlated with 1 other variable(s)"
[1] "nurdaym.10 removed, correlated with 1 other variable(s)"
[1] "nurdaym.11 removed, correlated with 1 other variable(s)"
[1] "nurdaym.12 removed, correlated with 1 other variable(s)"
[1] "nurdayy.2011 removed, correlated with 1 other variable(s)"
[1] "nurdayy.2012 removed, correlated with 1 other variable(s)"
[1] "pulse2 removed, correlated with 1 other variable(s)"
[1] "sys3 removed, correlated with 1 other variable(s)"
[1] "dias3 removed, correlated with 1 other variable(s)"
[1] "ag16g10.76 removed, correlated with 1 other variable(s)"
[1] "hhtype_dv.y removed, correlated with 1 other variable(s)"
[1] "marstat.y.2 removed, correlated with 1 other variable(s)"
[1] "marstat.y.4 removed, correlated with 1 other variable(s)"
[1] "marstat.y.5 removed, correlated with 1 other variable(s)"
[1] "marstat.y.6 removed, correlated with 1 other variable(s)"
[1] "vpstimehh removed, correlated with 1 other variable(s)"
[1] "indnsub_xw removed, correlated with 1 other variable(s)"
[1] "alt removed, correlated with 1 other variable(s)"
[1] "hba1c removed, correlated with 1 other variable(s)"
[1] "testo removed, correlated with 1 other variable(s)"
[1] "uscmg..22 removed, correlated with 1 other variable(s)"
[1] "uscmg.1 removed, correlated with 1 other variable(s)"
[1] "hhorig.3 removed, correlated with 1 other variable(s)"
[1] "fimnnet_dv removed, correlated with 1 other variable(s)"
[1] "respm16_dv.2 removed, correlated with 1 other variable(s)"
[1] "racel_dv removed, correlated with 1 other variable(s)"
[1] "wstval removed, correlated with 1 other variable(s)"
[1] "240 variables removed since they had high correlation coefs"
[1] "The final cleaned dataset has been completed at this stage and is stored under the name 'cleanDa
[1] "------------Attempting a Train Test Split------------"
[1] "Good train, test split found"
[1] "The working seed found was 3"
[1] "------------kNN------------"
[1] "96 neighbours considered for each test data point"
[1] "kNN results as a table, follow the diagonal for the correctly mapped clusters"
         real
predicted  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 28 30 31 32 3
        0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
        1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
        2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
        3  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
        4  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
        5  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
        6  0  0  1  2 14 49 53 42 36 42 23 18  5  5  4  1  2  1  3  2  1
```

```
0  1  1  0  0  1  0  0  0  0  0  0
       7   1  1  2  2  8 53 54 23 41 40 31 43 16 12  5  5  2  4  4  2  3
2  2  1  0  0  0  0  0  1  0  0  0
       8   0  0  0  1  0  6  6  3  6  9  6  4  3  1  4  1  1  1  0  1  0
0  1  1  0  0  0  0  0  0  0  0  0
       9   0  2  0  1  1  5  2  5  6  6  4  6  5  5  6  0  1  3  1  0  1
1  3  0  1  0  0  0  0  0  0  0  0
       10  0  0  0  0  1  2  2  1  2  2  2  1  2  1  0  0  1  1  3  2  1
1  2  0  1  0  0  0  0  0  0  0  0
       11  0  0  0  1  0  0  4  1  2  2  7  4  3  2  3  6  2  1  3  1  4
3  1  1  0  0  1  0  0  0  0  1  0
       12  0  0  0  1  2  0  2  4  5  5 10  9 11  7  8  6  5  3  6  1  3
6  2  9  3  2  0  1  1  0  1  1  2
       13  0  0  0  0  0  0  0  1  0  0  0  0  0  1  0  0  0  0  0  2  0
0  0  0  0  0  1  1  0  0  0  0  0
       14  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  1  0  1  0
1  0  0  0  1  0  0  0  0  0  0  0
       15  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
       16  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
       17  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
       18  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
       19  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
       20  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
       21  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
       22  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
       23  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
       24  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
       25  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
       26  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
       27  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
       28  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
       29  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0
 [ reached getOption("max.print") -- omitted 7 rows ]
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
[1] "The MSE of the predicted values are of 265.4024"
[1] "The kNN model predicts exactly with an accuracy of 0.1308"
[1] "------------CART prediction model------------"
n= 2981

node), split, n, deviance, yval
      * denotes terminal node

 1) root 2981 86512.0900 10.963770
   2) sf12mcs_dv>=-0.5308194 2264 22242.8700  9.007509
     4) scsf6c.5>=0.5 1253  6276.6800  7.600160 *
     5) scsf6c.5< 0.5 1011 10408.6800 10.751730
      10) sf12pcs_dv>=-1.556036 915  7917.9630 10.444810 *
      11) sf12pcs_dv< -1.556036 96  1582.9900 13.677080 *
   3) sf12mcs_dv< -0.5308194 717 28246.7700 17.140860
     6) sf12mcs_dv>=-2.154291 590 14287.7100 15.628810
      12) sf12mcs_dv>=-1.031695 275  4908.6690 13.741820 *
      13) sf12mcs_dv< -1.031695 315  7544.9710 17.276190 *
     7) sf12mcs_dv< -2.154291 127  6343.5280 24.165350
```

```
      14) sf12pcs_dv>=0.1176167 66  2835.5300 21.287880
        28) sf12mcs_dv>=-3.000744 45  1218.3110 18.644440 *
        29) sf12mcs_dv< -3.000744 21   628.9524 26.952380 *
      15) sf12pcs_dv< 0.1176167 61  2370.2620 27.278690 *
[1] "Variable Importance"
 sf12mcs_dv    scsf4a.5    scsf4a.3    scsf4b.5    scsf6c.3     scsf7.5
scsf6c.5    scsf6c.4  sf12pcs_dv    scsf4a.4
 49879.4728  14140.6139   9959.6362   9796.8985   8892.5694   8720.0765
5557.5079   4791.5670   2700.5225   1110.4022
   scsf6a.3    scsf2a.3    scsf2b.3    scsf6a.5    scsf3b.5      numed2
health.y.2    scsf6c.2     scsf5.4    scsf4b.2
  1082.9170    913.9182    802.0099    659.6133    634.1474    615.4960
578.1932    479.7187    359.3106    329.4223
 sclfsato.6         ast    scsf3b.2    scsf4a.2     sf1.y.5    scsf3a.2
alkp    btype3.1 ff_jbstat.8     scsf7.2
   313.4592    299.8242    255.2996    239.8594    236.3886    217.4775
188.2413    188.2413    188.2413    132.3776
[1] "The MSE of the predicted values are of 11.48"
[1] "The CART model predicts exactly with accuracy of 0.1217"
[1] "————Ordinary Linear Regression (Initial)————"
[1] "The full model AIC is: 15490.32"
[1] "————Variance Inflation Factor Removal————"
[1] "The variable sf12mcs_dv was removed since it had a VIF score of 1591.0528"
[1] "The variable natch02.4 was removed since it had a VIF score of 1387.4765"
[1] "The variable dory.2012 was removed since it had a VIF score of 412.0562"
[1] "The variable hhsize.4 was removed since it had a VIF score of 387.1691"
[1] "The variable agegr13_dv.13 was removed since it had a VIF score of 262.465"
[1] "The variable livesp_dv.1 was removed since it had a VIF score of 241.4685"
[1] "The variable ag16g10.66 was removed since it had a VIF score of 182.8464"
[1] "The variable nchild_dv.2 was removed since it had a VIF score of 148.824"
[1] "The variable fibenothr_dv was removed since it had a VIF score of 115.3215"
[1] "The variable sf12pcs_dv was removed since it had a VIF score of 97.6994"
[1] "The variable hrpid was removed since it had a VIF score of 73.826"
[1] "The variable allch01.3 was removed since it had a VIF score of 73.1975"
[1] "The variable fimngrs_dv was removed since it had a VIF score of 67.0226"
[1] "The variable scsf4b.5 was removed since it had a VIF score of 65.8248"
[1] "The variable uscmm.2 was removed since it had a VIF score of 62.005"
[1] "The variable rach16_dv.2 was removed since it had a VIF score of 60.8486"
[1] "The variable bmivg5.30 was removed since it had a VIF score of 56.927"
[1] "The variable ieqmoecd_dv was removed since it had a VIF score of 53.2752"
[1] "The variable doby_dv was removed since it had a VIF score of 43.4841"
[1] "The variable scsf4a.5 was removed since it had a VIF score of 35.8456"
[1] "The variable relup.6 was removed since it had a VIF score of 34.483"
[1] "The variable scsf6c.5 was removed since it had a VIF score of 31.0963"
[1] "The variable fimnlabgrs_if was removed since it had a VIF score of 27.2994"
[1] "The variable natch01.3 was removed since it had a VIF score of 27.0685"
[1] "The variable ag16g10.36 was removed since it had a VIF score of 26.7054"
[1] "The variable scsf7.5 was removed since it had a VIF score of 26.682"
[1] "The variable jbstat.y.4 was removed since it had a VIF score of 24.3389"
[1] "The variable sclfsato.6 was removed since it had a VIF score of 24.1209"
[1] "The variable allch02.4 was removed since it had a VIF score of 23.2348"
[1] "The variable scsf3b.5 was removed since it had a VIF score of 20.0459"
[1] "The variable bprespc.2 was removed since it had a VIF score of 19.4412"
[1] "The variable nnatch.2 was removed since it had a VIF score of 17.9845"
[1] "The variable sppno.2 was removed since it had a VIF score of 15.0773"
[1] "The variable nunmpsp_dv.1 was removed since it had a VIF score of 14.644"
[1] "The variable numed2 was removed since it had a VIF score of 13.9943"
[1] "The variable nchund18resp.2 was removed since it had a VIF score of 13.5688"
[1] "The variable mnspno.2 was removed since it had a VIF score of 13.2092"
[1] "The variable sclfsat1.6 was removed since it had a VIF score of 12.2243"
[1] "The variable samdifc1.1 was removed since it had a VIF score of 11.3241"
[1] "The variable nnatch.3 was removed since it had a VIF score of 11.2546"
[1] "The variable bensta96.1 was removed since it had a VIF score of 11.1756"
[1] "The variable agegr13_dv.9 was removed since it had a VIF score of 10.9849"
[1] "The variable natch01.2 was removed since it had a VIF score of 10.4111"
[1] "The variable buno_dv.3 was removed since it had a VIF score of 10.1504"
[1] "44 variables removed from the Ordinary Linear Model since they have a VIF score higher than 10"
[1] "The full model AIC after VIF checks is: 15947.9796"
```

```
[1] "———————Backwards Selection———————"
[1] "50 out of 512 variables removed so far."
[1] "100 out of 512 variables removed so far."
[1] "150 out of 512 variables removed so far."
[1] "200 out of 512 variables removed so far."
[1] "250 out of 512 variables removed so far."
[1] "300 out of 512 variables removed so far."
[1] "350 out of 512 variables removed so far."
[1] "383 out of 512 variables removed in backwards selection since they weren't significant at the 95
  [1] "floorc.3"          "ff_bentype03.1"   "ficode2.3"        "bensta1.1"
"nch10to15.1"       "ethn_dv"
  [7] "uscmm.3"           "scsf1.2"          "hiqual_dv.y.5"    "npensioner_dv.2"
"dorm.12"           "medtyp8.1"
 [13] "ivcoop.2"          "lfout.9"          "sf1.y.4"          "hcondn96.1"
"gor_dv.3"          "ff_jbstat.97"
 [19] "fimninvnet_dv"     "hrpno.3"          "sppno.4"          "samdifc95.1"
"dory.2011"         "difbpc1.1"
 [25] "difbpc3.1"         "btype4.1"         "nch5to15.1"       "ag16g20.76"
"vpprob1.1"         "ficode4.2"
 [31] "medtyp5.1"         "agegr13_dv.11"    "scsf1.4"          "sclfsat1.3"
"sclfsat1.4"        "ficode2.1"
 [37] "ficode19.1"        "nch5to15.3"       "sf1.y.2"          "allch01.2"
"scsf5.3"           "obpdrug.1"
 [43] "istrtdatss"        "consubx3.1"       "natch01.4"        "ficode39.1"
"ficode10.1"        "scsf2b.3"
 [49] "dord"              "nurdayw.3"        "hcondn4.1"        "jbstat.y.2"
"jbstat.y.3"        "ff_jbstat.5"
 [55] "hcondn11.1"        "relup.3"          "fimngrs_if"       "fibenothr_if"
"samdifc5.1"        "vpsys.2"
 [61] "btype3.1"          "nnsib_dv.1"       "finnow.2"         "ff_bentype05.1"
"scsf2a.3"          "fimnlabnet_tc.1"
 [67] "nurdayw.4"         "indin01_lw"       "indin91_lw"       "vpsens.3"
"nurdayw.2"         "nch415resp.3"
 [73] "ficode18.1"        "sclfsat7.3"       "lfout.4"          "nchild_dv.1"
"gor_dv.9"          "gor_dv.5"
 [79] "gor_dv.2"          "agegr13_dv.3"     "ag16g10.46"       "ficode28.1"
"dheas"             "seearngrs_if.1"
 [85] "hcondn14.1"        "paynu_if.1"       "samdifc4.1"       "samdifc3.1"
"fimnpen_dv"        "jbstat.y.97"
 [91] "hhsize.6"          "allch02.3"        "intdatm_dv.11"    "intdatm_dv.12"
"btype7.1"          "iron.1"
 [97] "urban_dv.y.2"      "agegr13_dv.8"     "difbpc5.1"        "jbiindb_dv"
"ff_jbstat.6"       "lipid.1"
[103] "cindtime"          "ficode37.1"       "gor_dv.4"         "susp.3"
"sampst.2"          "ff_bentype29.1"
[109] "vppress2.1"        "hrpno.4"          "wjrel.2"          "lfout.2"
"undqus.3"          "susp.2"
[115] "hhtype_dv.x"       "hhresp_dv.3"      "hhsize.2"         "marstat.y.3"
"hiqual_dv.y.9"     "ivprsnt.2"
[121] "nadoptch.1"        "adoptch01.3"      "ficode29.2"       "consubx4.1"
"fiyrinvinc_dv"     "fimnmisc_dv"
[127] "bmivg5.40"         "ficode22.1"       "origadd.2"        "nurdayw.5"
"nurdayw.1"         "ff_bentype19.1"
[133] "respf16_dv.2"      "rtin"             "gor_dv.8"         "hcondn12.1"
"nchresp.4"         "allch03.6"
[139] "nchild_dv.3"       "alkp"             "strtnurhh"        "ienddathh"
"ficode3.1"         "vpprob2.1"
[145] "ficode27.1"        "mmgsres.2"        "sclfsat1.5"       "dorm.4"
"intdatm_dv.2"      "lfout.5"
[151] "vphand.2"          "scsf2b.2"         "ff_bentype38.1"   "ficode38.1"
"ff_bentype01.1"    "medtyp11.1"
[157] "gor_dv.7"          "indbdub_xw"       "gor_dv.6"         "adoptch01.5"
"agegr13_dv.6"      "ag16g10.26"
[163] "j2pay_if.1"        "bmivg5.25"        "intdaty_dv.2011"  "qfhighfl_dv.1"
"vote1.2"           "nch415resp.4"
[169] "intdatd_dv"        "consubx1.1"       "pulse3"           "vparm.3"
"ienddatmm"         "uscmm.1"
[175] "nurdayd"           "medtyp6.1"        "hcondn13.1"       "rhland_code.1"
```

```
"ff_bentype34.1"    "ficode15.1"
[181] "ficode23.1"          "ff_bentype23.1"    "hcondn6.1"           "uscmm..22"
"igfi"                "allch03.4"
[187] "relup.4"             "sclfsat7.5"          "ficode29.3"          "nnewborn.1"
"jbstat.y.5"          "jbstat.y.6"
[193] "health.y.2"          "ivcoop.3"            "ficode11.1"          "consubx2.1"
"allch01.4"           "hhsize.3"
[199] "vpsam.2"             "hhorig.5"            "gor_dv.10"           "hgb"
"natch01.1"           "mnspno.1"
[205] "ff_bentype12.1"    "vpcheck.2"           "lkmove.2"            "natch02.6"
"hhresp_dv.2"         "marstat_dv.6"
[211] "buno_dv.2"           "ecre"                "ure"                 "ff_bentype11.1"
"ff_bentype33.1"    "nch14resp.2"
[217] "nch415resp.2"        "agegr13_dv.4"        "mmgstp.2"            "vote6.3"
"vote6.2"             "omsyst"
[223] "omdiast"             "vote6.4"             "hscrp"               "lungtest.3"
"lfout.8"             "intdatm_dv.5"
[229] "intdatm_dv.3"        "scsf5.2"             "fimnlabnet_dv"       "scsf3a.2"
"mmgsdval"            "hhsize.5"
[235] "scsf6b.2"            "ficode2.2"           "ff_bentype17.1"      "uscmg.3"
"ficode7.1"           "ficode9.1"
[241] "bfpcval"             "ienddatss"           "ff_bentype04.1"      "ficode4.1"
"nch10to15.3"         "fnspno.3"
[247] "natch03.6"           "allch02.6"           "hiqual_dv.y.2"       "wjrel.3"
"btype96.1"           "mmgstp.3"
[253] "agegr5_dv.14"        "fimnsben_dv"         "hcondn9.1"           "marstat_dv.4"
"hcondn16.1"          "hcondn7.1"
[259] "medtyp1.1"           "beta.1"              "medtyp10.1"          "medcnj.2"
"lenindintv"          "nseqno.2"
[265] "sampst.3"            "nch10to15.2"         "mnspno.3"            "ficode29.1"
"nch5to15.2"          "dorm.2"
[271] "dorm.3"              "vpsam.4"             "ficode13.1"          "statins.2"
"natch02.5"           "fibenothr_tc.1"
[277] "ficode3.2"           "difbpc6.1"           "hhorig.4"            "ff_bentype16.1"
"natch01.5"           "ff_bentype20.1"
[283] "nchund18resp.3"      "ff_bentype07.1"      "intdatm_dv.4"        "ff_bentype14.1"
"ff_jbstat.8"         "medtyp9.1"
[289] "ff_bentype27.1"      "trainany.2"          "hcondn1.1"           "hcondn2.1"
"vparm.2"             "lfout.3"
[295] "sppno.3"             "istrtdatmm"          "hrpno.2"             "aidxhh.2"
"pns1pno.2"           "npensioner_dv.1"
[301] "natch02.3"           "ast"                 "condna.2"            "adoptch01.2"
"agegr13_dv.7"        "marstat_dv.3"
[307] "nurdayw.6"           "medtyp3.1"           "npensioner_dv.3"     "dorm.8"
"intdatm_dv.7"        "dorm.6"
[313] "intdatm_dv.9"        "intdatm_dv.8"        "intdatm_dv.10"       "dorm.5"
"dorm.7"              "intdatm_dv.6"
[319] "vppress1.1"          "omronno"             "ff_bentype24.1"      "indscub_xw"
"medtyp7.1"           "natch03.4"
[325] "undqus.2"            "intdaty_dv.2012"     "sf1.y.5"             "sclfsat1.2"
"hcondn5.1"           "fnspno.2"
[331] "fnspno.1"            "j2has.2"             "j2pay_dv"            "nchild_dv.4"
"nchund18resp.4"      "adoptch01.4"
[337] "vpskin.2"            "ag16g10.86"          "aceinh.1"            "calciumb.1"
"sclfsat7.4"          "sclfsat7.2"
[343] "sclfsat7.7"          "nmpsp_dv.2"          "nnmpsp_dv.1"         "marstat_dv.2"
"nchund18resp.1"      "nmpsp_dv.1"
[349] "hiqual_dv.y.4"       "sf1.y.3"             "scsf1.3"             "scsf6b.3"
"vpprob95.1"          "finnow.3"
[355] "ff_bentype02.1"      "relwaitb.2"          "medtyp4.1"           "nch415resp.1"
"xtra5minosm_dv.1"    "scsf3b.4"
[361] "scsf3a.3"            "scsf3a.4"            "hcondn10.1"          "nseqno.3"
"ficode24.1"          "bmivg5.18"
[367] "cufsize.3"           "bmival"              "scsf4a.2"            "dorm.10"
"medtyp2.1"           "dorm.11"
[373] "samdifc2.1"          "cfib"                "dorm.9"              "strtnurmm"
"ff_bentype06.1"      "ficode6.1"
[379] "relhite.2"           "vpstimemm"           "sclfsat7.6"          "ff_jbstat.3"
```

```
"scsf3b.3"
[1] "———————Ordinary Linear Regression (Improved)———————"

Call:
lm(formula = y ~ ., data = as.data.frame(x.data.linear))

Residuals:
     Min       1Q   Median       3Q      Max
-13.8026  -1.7404  -0.1928   1.5474  17.7423

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       9.97602    1.01295   9.848  < 2e-16 ***
nch14resp.1       0.71787    0.20929   3.430 0.000612 ***
allch01.5        -1.32566    0.60396  -2.195 0.028245 *
allch02.5         1.21072    0.48619   2.490 0.012823 *
xpmove.2         -0.62314    0.22763  -2.738 0.006228 **
mobuse.2         -0.49031    0.22557  -2.174 0.029814 *
relup.5           2.14256    0.78541   2.728 0.006411 **
hcondn8.1        -2.62359    1.10273  -2.379 0.017416 *
hcondn17.1        2.10619    0.47179   4.464 8.34e-06 ***
ftedany.2        -2.21682    0.84783  -2.615 0.008977 **
btype1.1         -2.42581    0.77520  -3.129 0.001770 **
finnow.4          1.28739    0.25556   5.037 5.01e-07 ***
finnow.5          1.37251    0.45309   3.029 0.002473 **
finfut.2          0.68868    0.19460   3.539 0.000408 ***
finfut.3          0.34493    0.16764   2.058 0.039726 *
undqus.4         -4.33886    1.74452  -2.487 0.012934 *
scsf3b.2         -0.97725    0.33395  -2.926 0.003456 **
scsf4a.3          1.13740    0.27005   4.212 2.61e-05 ***
scsf4a.4          0.60004    0.19653   3.053 0.002284 **
scsf4b.2          4.41908    0.40670  10.866  < 2e-16 ***
scsf4b.3          1.07172    0.26683   4.016 6.06e-05 ***
scsf4b.4          0.69950    0.19611   3.567 0.000367 ***
scsf5.4           0.58791    0.24960   2.355 0.018568 *
scsf5.5           2.46694    0.39054   6.317 3.08e-10 ***
scsf6a.2          0.68457    0.24811   2.759 0.005833 **
scsf6a.3          1.82861    0.28370   6.446 1.34e-10 ***
scsf6a.4          4.00346    0.33422  11.978  < 2e-16 ***
scsf6a.5          6.67558    0.45131  14.792  < 2e-16 ***
scsf6b.4          0.58342    0.19802   2.946 0.003242 **
scsf6b.5          1.78960    0.32873   5.444 5.65e-08 ***
scsf6c.2          4.13327    0.36475  11.332  < 2e-16 ***
scsf6c.3          2.58150    0.20734  12.451  < 2e-16 ***
scsf6c.4          1.22399    0.15410   7.943 2.80e-15 ***
scsf7.2           2.45616    0.37490   6.552 6.72e-11 ***
scsf7.3           1.27394    0.23008   5.537 3.35e-08 ***
scsf7.4           1.06966    0.18065   5.921 3.57e-09 ***
sclfsat1.7       -0.46214    0.21953  -2.105 0.035367 *
sclfsat2.2       -1.32388    0.34657  -3.820 0.000136 ***
sclfsat2.3       -1.27487    0.32406  -3.934 8.55e-05 ***
sclfsat2.4       -1.07921    0.34259  -3.150 0.001649 **
sclfsat2.5       -1.51107    0.32876  -4.596 4.49e-06 ***
sclfsat2.6       -1.60697    0.32228  -4.986 6.52e-07 ***
sclfsat2.7       -1.43401    0.37171  -3.858 0.000117 ***
sclfsato.2        1.72977    0.32666   5.295 1.28e-07 ***
sclfsato.3        1.72775    0.26673   6.478 1.09e-10 ***
sclfsato.4        1.12531    0.25351   4.439 9.38e-06 ***
sclfsato.5        0.42235    0.18471   2.287 0.022294 *
sclfsato.7       -0.73598    0.20020  -3.676 0.000241 ***
ff_jbstat.7      -2.83664    1.04216  -2.722 0.006530 **
ff_bentype08.1    4.43479    1.31517   3.372 0.000756 ***
ff_bentype15.1   -1.00126    0.35651  -2.808 0.005011 **
ff_bentype26.1    1.20540    0.51615   2.335 0.019592 *
ff_bentype28.1   -1.76463    0.63918  -2.761 0.005803 **
agegr13_dv.10     0.53934    0.23984   2.249 0.024601 *
agegr13_dv.12    -0.60699    0.27196  -2.232 0.025696 *
nunmpsp_dv.2     -5.17366    1.48287  -3.489 0.000492 ***
```

```
indscus_lw       0.16974     0.06053   2.804 0.005080 **
ficode5.1       -1.44440     0.72832  -1.983 0.047439 *
ficode8.1       -2.71094     1.23576  -2.194 0.028334 *
ficode12.1       0.87967     0.44113   1.994 0.046232 *
ficode14.1      -1.63326     0.48316  -3.380 0.000733 ***
ficode16.1       2.75600     0.75619   3.645 0.000273 ***
ficode17.1      -1.98911     0.81910  -2.428 0.015225 *
ficode20.1      -0.73540     0.21366  -3.442 0.000586 ***
ficode26.1      -1.29063     0.56101  -2.301 0.021488 *
ficode33.1       1.93712     0.72103   2.687 0.007260 **
ficode34.1       3.37518     1.48291   2.276 0.022916 *
diur.1          -0.54367     0.21949  -2.477 0.013308 *
medtyp12.1       0.81567     0.33942   2.403 0.016320 *
airtemp          0.14314     0.05965   2.400 0.016467 *
difbpc2.1       -2.66576     0.97890  -2.723 0.006504 **
mmgstp.4         0.83187     0.39534   2.104 0.035447 *
vpalco.2        -0.55589     0.23338  -2.382 0.017290 *
vpsens.2         0.63127     0.30936   2.041 0.041381 *
vpprob3.1        4.12552     1.89360   2.179 0.029437 *
ethnic          -0.17838     0.06224  -2.866 0.004185 **
ag16g10.56       0.61549     0.23334   2.638 0.008392 **
jbstat.y.7      -2.72952     1.23440  -2.211 0.027099 *
jbstat.y.8       2.05354     0.50650   4.054 5.16e-05 ***
ggt              0.12617     0.05971   2.113 0.034700 *
uscmg.2         -0.30476     0.12204  -2.497 0.012573 *
sex_dv.2         0.52344     0.18306   2.859 0.004276 **
marstat_dv.5    -0.84639     0.41697  -2.030 0.042463 *
fimnprben_dv     0.24773     0.07860   3.152 0.001639 **
hiqual_dv.y.3   -0.41853     0.15727  -2.661 0.007827 **
htval            0.21253     0.08577   2.478 0.013271 *
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1     1

Residual standard error: 3.209 on 2895 degrees of freedom
Multiple R-squared:  0.6553,    Adjusted R-squared:  0.6452
F-statistic: 64.76 on 85 and 2895 DF,  p-value: < 2.2e-16

AIC: 15498.5599
MSE: 10.0028
[1] "The MSE of the predicted values are of 12.6721"
[1] "The Linear Model predicts exactly with accuracy of 0.1549"
[1] "The Linear Model predicts within a confidence interval with accuracy of 0.3783"
[1] "----------Elastic Net Regression----------"
529 x 1 sparse Matrix of class "dgCMatrix", with 15 entries
         names Estimate_Coefs
1  (Intercept)    11.32007230
2     finnow.4     0.37401965
3     finnow.5     0.12557945
4     scsf2a.3    -0.10674289
5      scsf5.5     0.37649383
6     scsf6a.4     0.42315560
7     scsf6c.5    -0.42567006
8      scsf7.5    -0.06653464
9   sclfsato.2     0.49257211
10  sclfsato.3     0.52198279
11  sclfsato.6    -0.32577606
12  sclfsato.7    -0.36721460
13  sf12mcs_dv    -3.43829935
14     sf1.y.5     0.32819111
15  sf12pcs_dv    -0.66640752
[1] "The MSE of the predicted values of the best fit model is 10.2971"
[1] "The Alpha of the best fit model is 1"
[1] "The Elastic Net Model predicts exactly with accuracy of 0.168"
[1] "----------Timer Results----------"
   user  system elapsed
 709.81   13.93  724.21
```

10.2.27 mixNurse console

```
[1] "—————Initial Checks—————"
[1] "3044105 NA cells were found across the entire dataset (43% of data as NA)"
[1] "—————Data Type Checks—————"
[1] "0 variables recoded since all their entries aren't numeric or NA"
[1] "NOTE: algorithim recodes categorical data alphabetically e.g (female = 1, male = 2)"
character(0)
[1] "—————Low Data Removal—————"
[1] "186 variables removed since they had >= 'naPercent' (default 20%) NA values"
  [1] "b_hidp"        "b_pno"         "b_splitnum"    "c_hidp"
"c_pno"          "c_splitnum"
  [7] "statina"       "folic"         "folpreghr"     "resnhi"
"ehtch"          "ehtm"
 [13] "ehtft"         "ehtin"         "nohtbc1"       "nohtbc2"
"nohtbc3"        "nohtbc4"
 [19] "nohtbc5"       "nohtbc6"       "nohtbc7"       "nohtbc8"
"hinrel"         "nobf1"
 [25] "nobf2"         "nobf3"         "resnwt"        "nowtbc1"
"nowtbc2"        "nowtbc3"
 [31] "nowtbc4"       "nowtbc5"       "nowtbc6"       "nowtbc7"
"nowtbc8"        "nowtbc9"
 [37] "ewtch"         "ewtkg"         "ewtst"         "ewtl"
"ynowh"          "whpnabm1"
 [43] "whpnabm2"      "whpnabm3"      "whpnabm4"      "whpnabm5"
"whpnabm6"       "whpnabm95"
 [49] "probwj"        "ynobp"         "nattbpd0"      "nattbpd1"
"nattbpd2"       "nattbpd3"
 [55] "nattbpd4"      "nattbpd95"     "mmgsprb1"      "mmgsprb2"
"mmgsprb3"       "mmgsprb95"
 [61] "noattlf0"      "noattlf1"      "noattlf2"      "noattlf3"
"noattlf4"       "noattlf5"
 [67] "noattlf95"     "lungsmhr"      "lunginhr"      "fev1pred"
"fvcpred"        "fev1fvcp"
 [73] "htfvc_sc"      "htfev_sc"      "htpef_sc"      "htfevfvc_sc"
"fev1pred_sc"    "fvcpred_sc"
 [79] "fev1fvcp_sc"   "qualcdf0"      "qualcdf1"      "qualcdf2"
"qualcdf3"       "qualcdf4"
 [85] "qualcdf5"      "qualcdf6"      "qualcdf7"      "qualcdf95"
"qualab"         "nulllf0"
 [91] "nulllf1"       "nulllf2"       "nulllf3"       "nulllf4"
"nulllf5"        "nulllf6"
 [97] "nulllf95"      "hasurg"        "haeysurg"      "hastro"
"chestinf"       "inhaler"
[103] "inhalhrs"      "lfwill"        "spirno"        "lftemp"
"noread"         "nlsatlf"
[109] "htfvc2"        "ynolf"         "lfstand"       "lfresp"
"problf1"        "problf2"
[115] "problf3"       "problf4"       "problf5"       "noattlf"
"ncgplf"         "ncguard"
[121] "refbsc1"       "refbsc2"       "refbsc3"       "refbsc4"
"refbsc5"        "refbsc6"
[127] "refbsc7"       "refbsc95"      "constorb"      "condna"
"samparm"        "samdifc1"
[133] "samdifc2"      "samdifc3"      "samdifc4"      "samdifc5"
"samdifc6"       "samdifc95"
[139] "nobsm1"        "nobsm2"        "nobsm3"        "nobsm95"
"vpsys"          "vphand"
[145] "vparm"         "vpskin"        "vpalco"        "vpsam"
"vppress1"       "vppress2"
[151] "vppress3"      "vpsens"        "vpprob1"       "vpprob2"
"vpprob3"        "vpprob95"
[157] "vpprob96"      "vpcheck"       "wtpc"          "feet"
"mmgspr"         "antic"
[163] "bfck2"         "hyper2om"      "waist3"        "mlstat"
"jbnssec8_dv"    "jbnssec3_dv"
[169] "jlnssec5_dv"   "hyper1"        "hyper2"        "vpstimehh"
"vpstimemm"      "pid"
```

```
[175] "bnf7_conhrt"     "bnf7_antifibs"   "bnf7_aspirin"    "bnf7_statins"
"bnf7_antiinflam" "bnf7_antiep"
[181] "indns91_lw"      "indns01_lw"      "indnsbh_xw"      "jlnssec8_dv"
"jlnssec3_dv"     "jbnssec5_dv"
[1] "—————Low Level Removal—————"
[1] "If a level is removed from a variable you wish to keep, reccomended to manually merge levels to
[1] "level 1 in medtyp13 removed, 2 observations found"
[1] "level 7 in nseqno removed, 2 observations found"
[1] "level 8 in nseqno removed, 1 observations found"
[1] "level 9 in nseqno removed, 1 observations found"
[1] "level 2 in dateok removed, 3 observations found"
[1] "level 8 in elig removed, 2 observations found"
[1] "level 4 in wstokb removed, 4 observations found"
[1] "level 14 in hhsize removed, 1 observations found"
[1] "level 9 in jbstat removed, 4 observations found"
[1] "level 7 in marstat removed, 2 observations found"
[1] "level 9 in marstat removed, 2 observations found"
[1] "11 total levels removed from 8 different variables. In total 24 observations deleted"
[1] "—————Variance 0 Check—————"
[1] "33 variables removed since their new variance was 0"
 [1] "tbmed"     "medtyp13" "resphts"  "respwts"  "bfpcok"   "whintro"  "bpconst"
"consubx1" "consubx2" "consubx3" "consubx4"
[12] "consubx5" "respbps"  "mmgswil"  "mmgsok"   "mmgssta"  "lungsurg" "lungeye"
"lunghrt"  "lunghosp" "lungex"   "lungtest"
[23] "clotb"    "fit"      "dateok"   "htok"     "wtok"     "bmiok"    "full1"
"full2"    "full3"    "bprespc"  "wstokb"
[1] "—————Dummy Variables—————"
[1] "predictor variable count went from 122 to 201"
[1] "—————Variance 0 Check—————"
[1] "7 variables removed since their new variance was 0"
[1] "nseqno.7"  "nseqno.8"  "nseqno.9"  "hhsize.14" "jbstat.9"  "marstat.7" "marstat.9"
[1] "—————K—Means—————"
[1] "15 clusters have been made for K—Means"
[1] "K—Means results as a table, the max value in each row is a simple way to define which cluster r

      0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16
17  18  19  20  21  22  23  24  25  26  27  28  29
  1   1   0   2   6   7  15  64  64  49  61  39  42  44  18  18  13   9
9   5   9   8   7   4   5   1   4   2   1   0   0
  2   0   2   4  12  10  28 124 130 100 100  89  90 104  49  33  36  21
14  21  12  20  18  12  11   6   2   3   3   3   2
  3   1   2   2   5   9  16  75  55  54  63  52  45  51  25  14  17  13
9  11   7   9  11   9   2   6   4   4   1   3   2
  4   0   1   2   5   7  16  60  53  56  69  68  47  58  28  26  17  10
13  10  16   5  11   5   4   6   1   2   2   1   1
  5   0   0   2   2   6  13  58  72  61  62  52  59  56  29  13  18  10
7  11  11   4   7   6   6   5   1   1   6   1   1
  6   0   0   0   1   1   4  40  34  19  37  25  26  28  14   5   9   6
4   6   3   4   3   5   2   2   2   0   0   1   1
  7   0   1   0   9   9  14  75  89  65  59  55  67  59  21  21  18  12
17  18   9   5   8   6   3   8   3   2   2   1   0
  8   1   4   2   4   7  18  63  74  60  58  62  56  60  42  13  16  23
8   8  15  10   8   5   7   4   3   3   2   4   2
  9   4   1   1   6  10  18  79  79  78  86  80  71  49  27  16  23  14
15  11  11  14  11   5   3   7   3   3   2   1   0
  10   2   2   0   3   9  14  62  58  60  57  62  56  73  27  15  16  18
15  13   7   5  12   7   8   5   4   0   1   1   0
  11   0   2   3   4   2  17  91  97  80  72  73  70  64  40  20  17  19
15  10  13   9  12   8   8   6   3   0   4   1   1
  12   4   3   4  15  10  30 151 139 128 119 110  75 117  58  42  29  27
22  21  12  18  15  10  12  10   4   1   2   3   2
  13   2   3   3   4  12  13  80  59  64  55  68  45  52  29  21  13  10
9   4   7   6   8   2   4   4   2   2   0   1   1
  14   4   2   4   3   4  23  75  66  74  52  55  52  54  21  16  22   9
8  10   9  12   3   6   2   4   2   3   3   2   1
  15   1   0   3   3   6  13  65  56  56  60  49  53  48  23  17  27   5
13   9   8  10   6   8   8   2   6   3   1   4   0
```

```
     30  31  32  33  34  35  36
  1   0   0   0   1   1   0   0
  2   1   1   1   3   1   1   3
  3   0   0   0   1   2   1   1
  4   1   1   0   0   0   0   1
  5   0   0   2   0   3   2   2
  6   0   1   0   0   1   1   0
  7   0   2   1   0   0   1   1
  8   0   2   0   3   2   2   0
  9   2   1   0   0   0   3   0
  10  2   1   1   1   1   1   0
  11  1   2   1   1   1   1   1
  12  2   1   1   1   2   0   0
  13  0   0   0   0   1   0   0
  14  1   1   0   2   0   1   1
  15  3   0   1   0   0   1   1
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
 [1] "Cluster 1: Within MSE 17394344425388, Size 509"     "Cluster 2: Within MSE 5628937473069624, S
 [3] "Cluster 3: Within MSE 20879636704489, Size 582"     "Cluster 4: Within MSE 18454806611295, Size
 [5] "Cluster 5: Within MSE 1152987749636850, Size 589"   "Cluster 6: Within MSE 157503308274, Size 2
 [7] "Cluster 7: Within MSE 22763082238819, Size 661"     "Cluster 8: Within MSE 18714436852944, Size
 [9] "Cluster 9: Within MSE 3053281251770580, Size 734"   "Cluster 10: Within MSE 22876625950467, Siz
[11] "Cluster 11: Within MSE 801652608599906, Size 769"   "Cluster 12: Within MSE 1151041806650814, S
[13] "Cluster 13: Within MSE 20900549854342, Size 584"    "Cluster 14: Within MSE 1144429210127181, S
[15] "Cluster 15: Within MSE 10668416979622, Size 569"
[1] "Total between cluster MSE: 184850596173291648, Total within cluster MSE: 1169000803265204"
[1] "The K-Means model predicts exactly with an accuracy of 0.1236"
[1] "----------Correlation Checks----------"
[1] "map1 removed, correlated with 13 other variable(s)"
[1] "map2 removed, correlated with 12 other variable(s)"
[1] "map3 removed, correlated with 10 other variable(s)"
[1] "weight removed, correlated with 7 other variable(s)"
[1] "mmgsnval removed, correlated with 7 other variable(s)"
[1] "ommapval removed, correlated with 10 other variable(s)"
[1] "estwt removed, correlated with 6 other variable(s)"
[1] "mmgsd1 removed, correlated with 6 other variable(s)"
[1] "mmgsn1 removed, correlated with 5 other variable(s)"
[1] "wtval removed, correlated with 5 other variable(s)"
[1] "mmgsd2 removed, correlated with 4 other variable(s)"
[1] "sys1 removed, correlated with 4 other variable(s)"
[1] "dias1 removed, correlated with 4 other variable(s)"
[1] "waist1 removed, correlated with 4 other variable(s)"
[1] "wave.3 removed, correlated with 3 other variable(s)"
[1] "mmgsn2 removed, correlated with 3 other variable(s)"
[1] "pulse1 removed, correlated with 3 other variable(s)"
[1] "sys2 removed, correlated with 3 other variable(s)"
[1] "dias2 removed, correlated with 3 other variable(s)"
[1] "waist2 removed, correlated with 3 other variable(s)"
[1] "hhorig.3 removed, correlated with 2 other variable(s)"
[1] "height removed, correlated with 2 other variable(s)"
[1] "mmgsd3 removed, correlated with 2 other variable(s)"
[1] "bswill.2 removed, correlated with 2 other variable(s)"
[1] "pulse2 removed, correlated with 2 other variable(s)"
[1] "sys3 removed, correlated with 2 other variable(s)"
[1] "dias3 removed, correlated with 2 other variable(s)"
[1] "bmi removed, correlated with 2 other variable(s)"
[1] "confage removed, correlated with 1 other variable(s)"
[1] "medcnjd.2 removed, correlated with 1 other variable(s)"
[1] "bpmedc.1 removed, correlated with 1 other variable(s)"
[1] "estht removed, correlated with 1 other variable(s)"
[1] "bfpc removed, correlated with 1 other variable(s)"
[1] "floorc.2 removed, correlated with 1 other variable(s)"
[1] "cufsize.2 removed, correlated with 1 other variable(s)"
[1] "mmgsn3 removed, correlated with 1 other variable(s)"
[1] "htfvc removed, correlated with 1 other variable(s)"
[1] "nuroutc.12 removed, correlated with 1 other variable(s)"
[1] "elig.9 removed, correlated with 1 other variable(s)"
[1] "pulse3 removed, correlated with 1 other variable(s)"
```

```
[1] "omdiaval removed, correlated with 1 other variable(s)"
[1] "omsyst removed, correlated with 1 other variable(s)"
[1] "ag16g10.76 removed, correlated with 1 other variable(s)"
[1] "psu removed, correlated with 1 other variable(s)"
[1] "indnsub_lw removed, correlated with 1 other variable(s)"
[1] "wstval removed, correlated with 1 other variable(s)"
[1] "46 variables removed since they had high correlation coefs"
[1] "The final cleaned dataset has been completed at this stage and is stored under the name 'cleanDa
[1] "——————————Attempting a Train Test Split——————————"
[1] "Good train, test split found"
[1] "The working seed found was 3"
[1] "——————————kNN——————————"
[1] "144 neighbours considered for each test data point"
[1] "kNN results as a table, follow the diagonal for the correctly mapped clusters"
         real
predicted   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15
16  17  18  19  20  21  22  23  24  25  26  27  28
        0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        2   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        3   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        4   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        5   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        6   0   4   1   8   6  33 122 102  81  97  70  58  76  24  30  20
14  11    6  11    9    5    5  10    5    4    1    3    1
        7   1   2   0   4   7  18  77  86  63  89  71  77  80  32  15  29
11  12    9  11    7  11    5    4    3    5    3    1    2
        8   2   2   0   3   2   5  37  44  28  35  30  28  30  21   7   7
6   8   7   6   5   2   2   5   0   0   3   0   0
        9   2   0   0   5   4   5  16  25  35  22  23  22  34  14   8   6
9   1   8   6   4   4   6   2   2   1   0   0   2
        10  0   0   2   0   2   2  16  13  22  15  15  15  12   7   8   6
2   1   1   4   5   2   4   1   1   0   1   0   0
        11  0   0   0   4   1   1   6   5   8   6  10   8  11   5   3   1
2   5   5   4   0   1   1   1   0   0   0   1   0
        12  0   0   0   1   0   1   8   5   9   4   9   7   7  11   1   4
2   5   4   1   3   4   2   2   2   2   0   1   1
        13  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   1   0   0   0   0   0   0   0   0   0   0   0
        14  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        15  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        16  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        17  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        18  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        19  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        20  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        21  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        22  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        23  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        24  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0
        25  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
```

```
0    0    0    0    0    0    0    0    0    0    0    0    0
         26    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0    0    0    0    0    0    0    0    0    0    0    0    0
          real
predicted  29  30  31  32  33  34  35  36
        0   0   0   0   0   0   0   0   0
        1   0   0   0   0   0   0   0   0
        2   0   0   0   0   0   0   0   0
        3   0   0   0   0   0   0   0   0
        4   0   0   0   0   0   0   0   0
        5   0   0   0   0   0   0   0   0
        6   0   0   0   1   1   2   0   0
        7   0   1   1   0   2   1   1   0
        8   1   0   0   1   0   0   0   1
        9   0   1   0   0   0   0   0   0
        10  0   1   0   0   1   1   0   0
        11  0   2   0   0   0   0   1   1
        12  1   0   0   0   0   0   0   0
        13  0   0   0   0   0   0   0   0
        14  0   0   0   0   0   0   0   0
        15  0   0   0   0   0   0   0   0
        16  0   0   0   0   0   0   0   0
        17  0   0   0   0   0   0   0   0
        18  0   0   0   0   0   0   0   0
        19  0   0   0   0   0   0   0   0
        20  0   0   0   0   0   0   0   0
        21  0   0   0   0   0   0   0   0
        22  0   0   0   0   0   0   0   0
        23  0   0   0   0   0   0   0   0
        24  0   0   0   0   0   0   0   0
        25  0   0   0   0   0   0   0   0
        26  0   0   0   0   0   0   0   0
 [ reached getOption("max.print") -- omitted 10 rows ]
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
[1] "The MSE of the predicted values are of 124.4545"
[1] "The kNN model predicts exactly with an accuracy of 0.1148"
[1] "----------CART prediction model----------"
n= 7524

node), split, n, deviance, yval
      * denotes terminal node

 1) root 7524 204018.600 10.845160
   2) medtyp4.1< 0.5 6351 135679.300 10.259800
     4) sf1.4< 0.5 5718 110080.800  9.985310
       8) sf1.5< 0.5 5620 103457.400  9.891993 *
       9) sf1.5>=0.5 98   3767.888 15.336730 *
     5) sf1.4>=0.5 633  21275.990 12.739340 *
   3) medtyp4.1>=0.5 1173  54380.750 14.014490
     6) sf1.5< 0.5 990  38335.000 13.089900
      12) age>=0.4734806 483  13504.600 11.612840 *
      13) age< 0.4734806 507  22772.750 14.497040 *
     7) sf1.5>=0.5 183  10620.950 19.016390 *
[1] "Variable Importance"
  medtyp4.1        sf1.5        sf1.4          age    jbstat.8    jbstat.4
ag16g20.56  ag16g20.36       numed2    jbstat.2
13958.53293  9291.73027  4322.57083  2057.65301  2018.12644  1571.99578
1376.02883  1333.42732  1125.20350   979.83477
ieqmoecd_dv     mmgsdval        htfev      bfpcval     htfevfvc
  937.23326   487.81977     94.98705     88.93121     87.41138
[1] "The MSE of the predicted values are of 23.2943"
[1] "The CART model predicts exactly with accuracy of 0.0853"
[1] "----------Ordinary Linear Regression (Initial)----------"
[1] "The full model AIC is: 44591.3901"
[1] "----------Variance Inflation Factor Removal----------"
[1] "The variable difbpc1.1 was removed since it had a VIF score of 109.8307"
[1] "The variable nurdayy.2012 was removed since it had a VIF score of 55.0732"
[1] "The variable age was removed since it had a VIF score of 43.1209"
```

```
[1] "The variable ieqmoecd_dv was removed since it had a VIF score of 40.9327"
[1] "The variable bmivg5.30 was removed since it had a VIF score of 34.1258"
[1] "The variable numed2 was removed since it had a VIF score of 12.2851"
[1] "6 variables removed from the Ordinary Linear Model since they have a VIF score higher than 10"
[1] "The full model AIC after VIF checks is: 44586.97"
[1] "-------------Backwards Selection-------------"
[1] "50 out of 146 variables removed so far."
[1] "100 out of 146 variables removed so far."
[1] "111 out of 146 variables removed in backwards selection since they weren't significant at the 95
  [1] "iron.1"       "bmivg5.40"    "urban_dv.2"   "mmgsres.2"    "hhsize.5"
"nurdaym.7"    "medtyp9.1"    "nurdaym.3"
  [9] "bmivg5.18"    "cufsize.3"    "indnsub_xw"   "bsoute.3"     "medtyp11.1"
"nurdayw.6"    "mmgstp.2"     "lfout.5"
 [17] "bmival"       "hhorig.4"     "hhsize.2"     "nseqno.2"     "omronno"
"elig.11"      "bsoute.5"     "ag16g10.66"
 [25] "strtnurhh"    "marstat.3"    "hhorig.5"     "airtemp"      "nurdaym.11"
"lfout.2"      "region.2"     "difbpc2.1"
 [33] "htval"        "difbpc6.1"    "marstat.5"    "nurdaym.9"    "nurdaym.10"
"bfpcval"      "floorc.3"     "hhsize.9"
 [41] "medcnjd.3"    "hiqual_dv.3"  "hiqual_dv.4"  "ethnic"       "beta.1"
"hhsize.4"     "hhsize.3"     "hiqual_dv.2"
 [49] "nurdayd"      "calciumb.1"   "marstat.4"    "medtyp1.1"    "htfevfvc"
"jbstat.10"    "ag16g20.76"   "ag16g10.56"
 [57] "nurdayw.4"    "jbstat.4"     "nseqno.4"     "bsoute.2"     "nurdaym.2"
"nurdaym.6"    "statins.2"    "medtyp12.1"
 [65] "bswill.3"     "lfout.3"      "relhite.2"    "difbpc95.1"   "medtyp2.1"
"strtnurmm"    "wjrel.3"      "hhsize.7"
 [73] "nurdaym.8"    "difbpc3.1"    "difbpc5.1"    "pidp"         "nurdayw.3"
"nurdayw.2"    "mmgsdom.2"    "relwaitb.2"
 [81] "nurdayw.5"    "nurdayw.1"    "medtyp3.1"    "medtyp5.1"    "ompulval"
"lungsmok.2"   "hiqual_dv.9"  "wjrel.2"
 [89] "hiqual_dv.5"  "bmivg5.25"    "htpef"        "htfev"        "ag16g10.86"
"hhsize.8"     "medtyp7.1"    "marstat.6"
 [97] "mmgstp.4"     "mmgstp.3"     "nseqno.5"     "hhsize.6"     "nurdayy.2011" "jbstat.97"
"jbstat.5"     "diur.1"
[105] "nurdaym.12"   "strata"       "lfout.4"      "mmgsdval"     "hhtype_dv"
"nseqno.3"     "nurdaym.4"
[1] "-------------Ordinary Linear Regression (Improved)-------------"

Call:
lm(formula = y ~ ., data = as.data.frame(x.data.linear))

Residuals:
     Min       1Q   Median       3Q      Max
-14.5900  -2.9427  -0.8057   1.9218  25.0074

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.98157    0.31056  25.701  < 2e-16 ***
nsex.2       0.77828    0.11649   6.681 2.54e-11 ***
aceinh.1    -0.59561    0.21576  -2.761 0.005785 **
obpdrug.1   -0.93587    0.42411  -2.207 0.027367 *
lipid.1     -0.49278    0.18794  -2.622 0.008757 **
medtyp4.1    1.81570    0.17454  10.403  < 2e-16 ***
medtyp6.1   -0.61441    0.18090  -3.396 0.000686 ***
medtyp8.1   -1.40267    0.57022  -2.460 0.013921 *
medtyp10.1  -0.75385    0.22697  -3.321 0.000900 ***
medcnj.2    -0.34934    0.14907  -2.343 0.019131 *
difbpc4.1    4.91924    1.54900   3.176 0.001501 **
lunginhl.2   0.47257    0.22912   2.063 0.039188 *
nseqno.6     4.75474    2.08087   2.285 0.022342 *
nurdaym.5    0.39357    0.18350   2.145 0.032000 *
omdiast      0.22340    0.07535   2.965 0.003037 **
ag16g10.26   0.79172    0.18668   4.241 2.25e-05 ***
ag16g10.36   0.87570    0.17431   5.024 5.18e-07 ***
ag16g10.46   0.99222    0.16440   6.035 1.66e-09 ***
jbstat.2     0.55749    0.14401   3.871 0.000109 ***
jbstat.3     2.21389    0.30713   7.208 6.21e-13 ***
```

```
jbstat.6      0.83169    0.26202   3.174 0.001509 **
jbstat.7      1.02627    0.29555   3.472 0.000519 ***
jbstat.8      3.16634    0.41376   7.653 2.22e-14 ***
sf1.2         1.18366    0.15639   7.569 4.22e-14 ***
sf1.3         2.11309    0.17036  12.403  < 2e-16 ***
sf1.4         3.97284    0.21827  18.202  < 2e-16 ***
sf1.5         6.87244    0.35870  19.159  < 2e-16 ***
health.2     -0.82660    0.13802  -5.989 2.21e-09 ***
marstat.2    -0.26667    0.11528  -2.313 0.020737 *
omsysval     -0.29117    0.07944  -3.666 0.000249 ***
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1       1

Residual standard error: 4.636 on 7494 degrees of freedom
Multiple R-squared:  0.2104,    Adjusted R-squared:  0.2073
F-statistic: 68.85 on 29 and 7494 DF,  p-value: < 2.2e-16

AIC:  44467.1142
MSE:  21.4113
[1] "The MSE of the predicted values are of 21.5282"
[1] "The Linear Model predicts exactly with accuracy of 0.0921"
[1] "The Linear Model predicts within a confidence interval with accuracy of 0.195"
[1] "----------Elastic Net Regression----------"
149 x 1 sparse Matrix of class "dgCMatrix", with 16 entries
          names Estimate_Coefs
1   (Intercept)   10.214736288
2        nsex.2    0.590923229
3           age   -0.053872206
4      aceinh.1   -0.050593422
5       lipid.1   -0.103000815
6     medtyp4.1    1.730722992
7    ag16g20.36    0.235441793
8    ag16g20.56   -0.206443988
9      jbstat.3    0.878555732
10     jbstat.4   -0.805322032
11     jbstat.8    2.210511671
12        sf1.3    0.541234661
13        sf1.4    2.064175295
14        sf1.5    4.520340298
15     health.2   -0.678852042
16     omsysval   -0.004364251
[1] "The MSE of the predicted values of the best fit model is 21.4601"
[1] "The Alpha of the best fit model is 0.6"
[1] "The Elastic Net Model predicts exactly with accuracy of 0.0949"
[1] "----------Timer Results----------"
   user  system elapsed
  29.06    0.14   29.22
```

10.2.28 mixNurseBlood console

```
[1] "----------Initial Checks----------"
[1] "1910999 NA cells were found across the entire dataset (38.47% of data as NA)"
[1] "----------Data Type Checks----------"
[1] "0 variables recoded since all their entries aren't numeric or NA"
[1] "NOTE: algorithim recodes categorical data alphabetically e.g (female = 1, male = 2)"
character(0)
[1] "----------Low Data Removal----------"
[1] "166 variables removed since they had >= 'naPercent' (default 20%) NA values"
  [1] "b_hidp.x"       "b_pno.x"        "b_splitnum.x"   "c_hidp.x"
"c_pno.x"        "c_splitnum.x"
  [7] "statina"        "folic"          "folpreghr"      "resnhi"
"ehtch"          "ehtm"
 [13] "ehtft"          "ehtin"          "nohtbc1"        "nohtbc2"
"nohtbc3"        "nohtbc4"
 [19] "nohtbc5"        "nohtbc6"        "nohtbc7"        "nohtbc8"
"hinrel"         "nobf1"
```

```
 [25] "nobf2"          "nobf3"          "resnwt"         "nowtbc1"
"nowtbc2"         "nowtbc3"
 [31] "nowtbc4"        "nowtbc5"        "nowtbc6"        "nowtbc7"
"nowtbc8"         "nowtbc9"
 [37] "ewtch"          "ewtkg"          "ewtst"          "ewtl"
"ynowh"           "whpnabm1"
 [43] "whpnabm2"       "whpnabm3"       "whpnabm4"       "whpnabm5"
"whpnabm6"        "whpnabm95"
 [49] "probwj"         "ynobp"          "nattbpd0"       "nattbpd1"
"nattbpd2"        "nattbpd3"
 [55] "nattbpd4"       "nattbpd95"      "mmgsprb1"       "mmgsprb2"
"mmgsprb3"        "mmgsprb95"
 [61] "noattlf0"       "noattlf1"       "noattlf2"       "noattlf3"
"noattlf4"        "noattlf5"
 [67] "noattlf95"      "lungsmhr"       "lunginhr"       "fev1pred"
"fvcpred"         "fev1fvcp"
 [73] "htfvc_sc"       "htfev_sc"       "htpef_sc"       "htfevfvc_sc"
"fev1pred_sc"     "fvcpred_sc"
 [79] "fev1fvcp_sc"    "qualcdf0"       "qualcdf1"       "qualcdf2"
"qualcdf3"        "qualcdf4"
 [85] "qualcdf5"       "qualcdf6"       "qualcdf7"       "qualcdf95"
"qualab"          "nulllf0"
 [91] "nulllf1"        "nulllf2"        "nulllf3"        "nulllf4"
"nulllf5"         "nulllf6"
 [97] "nulllf95"       "hasurg"         "haeysurg"       "hastro"
"chestinf"        "inhaler"
[103] "inhalhrs"       "lfwill"         "spirno"         "lftemp"
"noread"          "nlsatlf"
[109] "htfvc2"         "ynolf"          "lfstand"        "lfresp"
"problf1"         "problf2"
[115] "problf3"        "problf4"        "problf5"        "noattlf"
"ncgplf"          "ncguard"
[121] "refbsc1"        "refbsc2"        "refbsc3"        "refbsc4"
"refbsc5"         "refbsc6"
[127] "refbsc7"        "refbsc95"       "nobsm1"         "nobsm2"
"nobsm3"          "nobsm95"
[133] "wtpc"           "feet"           "mmgspr"         "antic"
"bfck2"           "hyper2om"
[139] "waist3"         "mlstat"         "jbnssec8_dv"    "jbnssec3_dv"
"jlnssec5_dv"     "hyper1"
[145] "hyper2"         "pid.x"          "bnf7_conhrt"    "bnf7_antifibs"
"bnf7_aspirin"    "bnf7_statins"
[151] "bnf7_antiinflam" "bnf7_antiep"   "indns91_lw"     "indns01_lw"
"indnsbh_xw"      "pid.y"
[157] "b_hidp.y"       "b_pno.y"        "b_splitnum.y"   "c_hidp.y"
"c_pno.y"         "c_splitnum.y"
[163] "cmvavc"         "jlnssec8_dv"    "jlnssec3_dv"    "jbnssec5_dv"
[1] "-------------Low Level Removal-------------"
[1] "If a level is removed from a variable you wish to keep, reccomended to manually merge levels to
[1] "level 3 in medcnjd removed, 1 observations found"
[1] "level 1 in medtyp13 removed, 2 observations found"
[1] "level 3 in vpsam removed, 1 observations found"
[1] "level 7 in nseqno removed, 1 observations found"
[1] "level 9 in nseqno removed, 1 observations found"
[1] "level 2 in dateok removed, 2 observations found"
[1] "level 8 in elig removed, 1 observations found"
[1] "level 9 in elig removed, 4 observations found"
[1] "level 4 in wstokb removed, 2 observations found"
[1] "level 9 in hhsize removed, 2 observations found"
[1] "level 14 in hhsize removed, 1 observations found"
[1] "level 10 in jbstat removed, 4 observations found"
[1] "level 7 in marstat removed, 2 observations found"
[1] "13 total levels removed from 10 different variables. In total 24 observations deleted"
[1] "-------------Variance 0 Check-------------"
[1] "39 variables removed since their new variance was 0"
 [1] "tbmed"     "medtyp13" "resphts"  "respwts"  "bfpcok"   "whintro"  "bpconst"
"consubx1" "consubx2" "consubx3" "consubx4"
[12] "consubx5" "respbps"  "mmgswil"  "mmgsok"   "mmgssta"  "lungsurg" "lungeye"
```

```
"lunghrt"  "lunghosp" "lungex"   "lungtest"
[23] "clotb"     "fit"       "bswill"    "constorb" "samdifc6" "dateok"    "nuroutc"
"bsoute"    "htok"      "wtok"      "bmiok"
[34] "elig"      "full1"     "full2"     "full3"     "bprespc"  "wstokb"
[1] "——————Dummy Variables——————"
[1] "predictor variable count went from 169 to 265"
[1] "——————Variance 0 Check——————"
[1] "8 variables removed since their new variance was 0"
[1] "medcnjd.3" "vpsam.3"   "nseqno.7"  "nseqno.9"  "hhsize.9"  "hhsize.14" "jbstat.10" "marstat.7"
[1] "——————K—Means——————"
[1] "15 clusters have been made for K—Means"
[1] "K—Means results as a table, the max value in each row is a simple way to define which cluster re

       0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16
17  18  19  20  21  22  23  24  25  26  27  28  29
  1    2   1   3   4   6  20 115 107  98  89  92  87  92  45  19  22  26
24  17  13   7  18  12  10   6   6   0   4   2   1
  2    2   3   3   8  17  15 110 109  96  79  83  85  84  31  35  21  13
14  15   9   9  11   6   6   6   5   2   2   2   0
  3    0   0   1   2   4   7  47  53  32  56  40  38  30  17   6  13  11
4   4   3   7   6   4   3   4   2   0   0   2   1
  4    0   4   2   3   5   8  42  48  43  40  50  42  40  30   6  11  16
7   7   7   7   4   2   4   3   2   1   0   4   2
  5    0   0   3   6   8  18  92  78  77  99  82  72  84  30  29  30  14
19  11  18   9  12   9   9   5   4   3   3   4   1
  6    0   0   2   0   6  12  39  55  44  44  42  46  45  21   9  13   5
7   9   7   4   5   6   5   5   1   1   4   1   0
  7    1   1   1   8   3  12  62  55  56  42  38  25  39  25  22  10   9
6   9   3   4   3   3   4   2   1   0   2   1   0
  8    1   0   2   5   5  11  43  50  35  42  31  30  29  12  14  10   7
7   4   6   6   4   3   4   1   3   2   0   0   0
  9    0   1   0   1   3   7  27  22  20  14  17  18  26   7   4   4   2
2   3   1   6   1   3   0   3   1   2   1   1   1
  10   4   1   1   4   6  10  58  52  64  57  56  52  30  20   8  14   9
10   7  10  13   7   6   0   4   2   2   0   0   0
  11   3   0   3   2   0   7  30  25  32  19  20  21  14   9   7  12   2
2   2   4   5   0   2   2   1   1   1   1   0   0
  12   0   2   1   5   6  15  75  80  57  43  44  42  61  23  19  17  13
6  10   7  10  13   4   7   3   1   2   2   2   1
  13   1   0   1   4   4   8  63  46  42  42  35  27  35  17   9  11   9
9   6   4   8   9   5   4   4   2   0   1   0   0
  14   0   1   2   4   5   9  53  43  35  39  32  29  37  19  10  11   7
5   8   3   7   8   7   2   4   2   1   0   2   1
  15   0   0   0   0   1   2   2   4   0   2   2   5   3   2   2   1   1
1   0   1   1   1   2   1   1   0   2   0   0   0

      30  31  32  33  34  35  36
  1    3   2   2   2   1   2   0
  2    0   1   1   0   1   1   1
  3    0   0   0   0   0   1   0
  4    0   2   0   3   2   2   0
  5    3   1   1   0   0   1   2
  6    0   0   1   0   3   1   2
  7    2   2   0   1   1   0   0
  8    0   0   0   1   1   0   0
  9    0   0   0   1   0   0   1
  10   0   0   0   0   0   1   0
  11   0   1   0   1   0   1   0
  12   1   1   0   3   1   1   3
  13   0   0   1   0   1   0   0
  14   0   0   0   1   2   0   0
  15   0   0   0   0   0   0   0
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
 [1] "Cluster 1: Within MSE 2022330493971919, Size 960"  "Cluster 2: Within MSE 1171468871030527, Siz
 [3] "Cluster 3: Within MSE 1157520652453464, Size 398"  "Cluster 4: Within MSE 8053373330729, Size 4
 [5] "Cluster 5: Within MSE 1163075106126793, Size 837"  "Cluster 6: Within MSE 1168100060880259, Siz
 [7] "Cluster 7: Within MSE 212295010367327, Size 453"   "Cluster 8: Within MSE 13828941946017, Size
 [9] "Cluster 9: Within MSE 108173798541, Size 200"      "Cluster 10: Within MSE 3065692599088350, S
```

```
[11] "Cluster 11: Within MSE 128858183819, Size 230"     "Cluster 12: Within MSE 3230922626436751, S
[13] "Cluster 13: Within MSE 10186481874701, Size 408"   "Cluster 14: Within MSE 22654169752371, Size
[15] "Cluster 15: Within MSE 37352999524724, Size 37"
[1] "Total between cluster MSE: 183795444183573472, Total within cluster MSE: 1187017981049227"
[1] "The K-Means model predicts exactly with an accuracy of 0.1288"
[1] "----------Correlation Checks----------"
[1] "map1 removed, correlated with 13 other variable(s)"
[1] "map2 removed, correlated with 12 other variable(s)"
[1] "map3 removed, correlated with 10 other variable(s)"
[1] "weight removed, correlated with 7 other variable(s)"
[1] "mmgsnval removed, correlated with 7 other variable(s)"
[1] "ommapval removed, correlated with 10 other variable(s)"
[1] "estwt removed, correlated with 6 other variable(s)"
[1] "mmgsd1 removed, correlated with 6 other variable(s)"
[1] "mmgsn1 removed, correlated with 5 other variable(s)"
[1] "wtval removed, correlated with 5 other variable(s)"
[1] "wave.3 removed, correlated with 4 other variable(s)"
[1] "mmgsd2 removed, correlated with 4 other variable(s)"
[1] "sys1 removed, correlated with 4 other variable(s)"
[1] "dias1 removed, correlated with 4 other variable(s)"
[1] "waist1 removed, correlated with 4 other variable(s)"
[1] "alb removed, correlated with 4 other variable(s)"
[1] "hhorig.x.3 removed, correlated with 3 other variable(s)"
[1] "mmgsn2 removed, correlated with 3 other variable(s)"
[1] "pulse1 removed, correlated with 3 other variable(s)"
[1] "sys2 removed, correlated with 3 other variable(s)"
[1] "dias2 removed, correlated with 3 other variable(s)"
[1] "waist2 removed, correlated with 3 other variable(s)"
[1] "chol removed, correlated with 3 other variable(s)"
[1] "height removed, correlated with 2 other variable(s)"
[1] "mmgsd3 removed, correlated with 2 other variable(s)"
[1] "pulse2 removed, correlated with 2 other variable(s)"
[1] "sys3 removed, correlated with 2 other variable(s)"
[1] "dias3 removed, correlated with 2 other variable(s)"
[1] "bmi removed, correlated with 2 other variable(s)"
[1] "psu removed, correlated with 2 other variable(s)"
[1] "indnsub_lw removed, correlated with 2 other variable(s)"
[1] "hdl removed, correlated with 2 other variable(s)"
[1] "hhorig.x.4 removed, correlated with 1 other variable(s)"
[1] "hhorig.x.5 removed, correlated with 1 other variable(s)"
[1] "nsex.2 removed, correlated with 1 other variable(s)"
[1] "confage removed, correlated with 1 other variable(s)"
[1] "medcnjd.2 removed, correlated with 1 other variable(s)"
[1] "bpmedc.1 removed, correlated with 1 other variable(s)"
[1] "estht removed, correlated with 1 other variable(s)"
[1] "bfpc removed, correlated with 1 other variable(s)"
[1] "floorc.2 removed, correlated with 1 other variable(s)"
[1] "cufsize.2 removed, correlated with 1 other variable(s)"
[1] "mmgsdom.2 removed, correlated with 1 other variable(s)"
[1] "mmgsn3 removed, correlated with 1 other variable(s)"
[1] "htfvc removed, correlated with 1 other variable(s)"
[1] "samparm.2 removed, correlated with 1 other variable(s)"
[1] "samparm.3 removed, correlated with 1 other variable(s)"
[1] "nurdaym.2 removed, correlated with 1 other variable(s)"
[1] "nurdaym.3 removed, correlated with 1 other variable(s)"
[1] "nurdaym.4 removed, correlated with 1 other variable(s)"
[1] "nurdaym.5 removed, correlated with 1 other variable(s)"
[1] "nurdaym.6 removed, correlated with 1 other variable(s)"
[1] "nurdaym.7 removed, correlated with 1 other variable(s)"
[1] "nurdaym.8 removed, correlated with 1 other variable(s)"
[1] "nurdaym.9 removed, correlated with 1 other variable(s)"
[1] "nurdaym.10 removed, correlated with 1 other variable(s)"
[1] "nurdaym.11 removed, correlated with 1 other variable(s)"
[1] "nurdaym.12 removed, correlated with 1 other variable(s)"
[1] "nurdayy.2011 removed, correlated with 1 other variable(s)"
[1] "nurdayy.2012 removed, correlated with 1 other variable(s)"
[1] "pulse3 removed, correlated with 1 other variable(s)"
[1] "omdiaval removed, correlated with 1 other variable(s)"
```

```
[1] "omsyst removed, correlated with 1 other variable(s)"
[1] "ag16g10.76 removed, correlated with 1 other variable(s)"
[1] "hhtype_dv removed, correlated with 1 other variable(s)"
[1] "vpstimehh removed, correlated with 1 other variable(s)"
[1] "strata removed, correlated with 1 other variable(s)"
[1] "indnsub_xw removed, correlated with 1 other variable(s)"
[1] "hba1c removed, correlated with 1 other variable(s)"
[1] "trig removed, correlated with 1 other variable(s)"
[1] "uscmg..22 removed, correlated with 1 other variable(s)"
[1] "uscmg.1 removed, correlated with 1 other variable(s)"
[1] "wstval removed, correlated with 1 other variable(s)"
[1] "73 variables removed since they had high correlation coefs"
[1] "The final cleaned dataset has been completed at this stage and is stored under the name 'cleanDa
[1] "————————Attempting a Train Test Split————————"
[1] "Good train, test split found"
[1] "The working seed found was 3"
[1] "————————kNN————————"
[1] "115 neighbours considered for each test data point"
[1] "kNN results as a table, follow the diagonal for the correctly mapped clusters"
         real
predicted  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 3
       0   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
       1   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
       2   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
       3   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
       4   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
       5   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
       6   1  1  1  4 11 13 91 71 51 61 52 36 47 21 11 14  9 10  7  7  5
5  4  4  3  3  2  1  1  1  0  0  1  1  0  0  0
       7   2  1  4  7  5 18 67 63 78 60 60 48 65 18  9 15 16  8  9  6 14
7  5  3  3  2  0  0  2  1  0  0  0  1  0  0  1
       8   0  1  0  2  1  4 19 15 12 24 17 17 15  8  4  5  3  2  4  2  4
3  4  1  1  0  0  0  0  1  0  1  0  0  0  1  0
       9   0  1  1  2  1  4 24 21 21 20 21 19 12 11  4  5 10  5  4  3  2
0  1  3  0  2  0  1  1  0  0  0  0  0  0  0  0
       10  0  1  0  1  1  2  6  9  5  7  3  6  4  5  1  1  2  1  1  2  1
2  1  1  0  0  0  0  1  0  0  0  0  0  0  1  0
       11  0  0  0  0  0  3  7  8  8  8  9  6 15  2  3  2  2  1  3  1  3
2  1  2  1  1  0  1  0  0  1  0  0  0  1  0  1
       12  0  0  1  2  1  5 11 11  9  5  9 12  3  8  4  4  3  3  1  4  2
5  1  2  1  0  1  1  2  0  0  0  1  0  0  0  0
       13  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
       14  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
       15  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
       16  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
       17  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
       18  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
       19  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
       20  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
       21  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
       22  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
       23  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
```

```
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
       24  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
       25  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
       26  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 [ reached getOption("max.print") -- omitted 10 rows ]
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
[1] "The MSE of the predicted values are of 174.5705"
[1] "The kNN model predicts exactly with an accuracy of 0.1107"
[1] "-----------CART prediction model-----------"
n= 5362

node), split, n, deviance, yval
      * denotes terminal node

 1) root 5362 148075.500 10.867210
   2) medtyp4.1< 0.5 4440  93290.360 10.223420
     4) sf1.4< 0.5 4001  75833.580  9.970757
       8) sf1.5< 0.5 3936  71417.260  9.888720 *
       9) sf1.5>=0.5 65    2785.754 14.938460 *
     5) sf1.4>=0.5 439  14873.450 12.526200 *
   3) medtyp4.1>=0.5 922  44083.020 13.967460
     6) sf1.5< 0.5 781  31399.980 13.071700
      12) age>=0.4984499 347    9829.228 11.302590 *
      13) age< 0.4984499 434  19616.420 14.486180 *
     7) sf1.5>=0.5 141    8585.291 18.929080 *
[1] "Variable Importance"
  medtyp4.1        sf1.5        sf1.4          age     jbstat.4     jbstat.8
numed2  ag16g20.36  ag16g20.56     jbstat.2
10702.06835  6610.48021  2583.33661  1954.33992  1616.41371  1479.98307
1195.91467  1115.15650   951.82549   895.50446
      htfev     mmgsdval          ast          alt       bmival     htfevfvc
ecre          ure         alkp         rtin
  827.91922   528.77833   174.37227   145.31023    81.25215    75.25687
35.30756     17.65378     11.76919     11.76919
[1] "The MSE of the predicted values are of 22.7697"
[1] "The CART model predicts exactly with accuracy of 0.09"
[1] "-----------Ordinary Linear Regression (Initial)-----------"
[1] "The full model AIC is: 31934.6338"
[1] "-----------Variance Inflation Factor Removal-----------"
[1] "The variable dory.2012 was removed since it had a VIF score of 464.7289"
[1] "The variable uscmm.2 was removed since it had a VIF score of 51.9098"
[1] "The variable ag16g10.66 was removed since it had a VIF score of 43.9516"
[1] "The variable bmivg5.30 was removed since it had a VIF score of 39.2688"
[1] "The variable ieqmoecd_dv was removed since it had a VIF score of 29.2878"
[1] "The variable numed2 was removed since it had a VIF score of 11.8883"
[1] "The variable samdifc1.1 was removed since it had a VIF score of 10.8229"
[1] "7 variables removed from the Ordinary Linear Model since they have a VIF score higher than 10"
[1] "The full model AIC after VIF checks is: 31925.6378"
[1] "-----------Backwards Selection-----------"
[1] "50 out of 180 variables removed so far."
[1] "100 out of 180 variables removed so far."
[1] "135 out of 180 variables removed in backwards selection since they weren't significant at the 95
  [1] "vpprob95.1"  "samdifc95.1" "age"          "vppress3.1"  "bmivg5.25"
"dorm.4"       "hiqual_dv.2" "relhite.2"
  [9] "vpprob3.1"   "vpsam.4"     "jbstat.7"     "ag16g10.86"  "mmgstp.2"
"difbpc2.1"   "mmgstp.3"    "vpsys.2"
 [17] "nurdayw.5"   "indbdub_xw"  "pidp"         "floorc.3"    "marstat.4"
"vppress2.1"  "hhorig.y.3"  "vparm.2"
 [25] "cfib"        "jbstat.2"    "vppress1.1"   "bfpcval"     "iron.1"
"hhsize.2"    "hhsize.5"    "hhsize.4"
 [33] "htfevfvc"    "lfout.5"     "wjrel.3"      "nseqno.2"    "hiqual_dv.3" "strtnurhh"
"dorm.8"       "samdifc5.1"
 [41] "marstat.6"   "dory.2011"   "nseqno.3"     "bmivg5.18"   "ggt"
"samdifc3.1"  "marstat.3"   "dorm.2"
 [49] "lfout.2"     "vpsens.3"    "medtyp12.1"   "strtnurmm"   "medtyp1.1"
```

```
"hscrp"        "ure"          "region.2"
 [57] "hhorig.y.5"  "omronno"      "urban_dv.2"   "nseqno.4"     "vpprob1.1"
"vpskin.2"     "lfout.3"      "lfout.4"
 [65] "hiqual_dv.4" "medtyp7.1"    "nurdayw.2"    "samdifc2.1"   "bmivg5.40"
"dorm.10"      "vparm.3"      "vpsam.2"
 [73] "uscmm.3"     "beta.1"       "obpdrug.1"    "alt"          "difbpc3.1"
"difbpc1.1"    "difbpc6.1"    "airtemp"
 [81] "ag16g10.56"  "ag16g20.76"   "igfi"         "hgb"          "hiqual_dv.5"  "marstat.5"
"vpsens.2"     "dorm.3"
 [89] "dord"        "condna.2"     "nseqno.5"     "nurdayd"      "medtyp9.1"
"mmgsres.2"    "alkp"         "lunginhl.2"
 [97] "hhsize.6"    "uscmg.3"      "statins.2"    "wjrel.2"      "medtyp3.1"
"difbpc5.1"    "medtyp11.1"   "calciumb.1"
[105] "dorm.11"     "nurdayw.1"    "nurdayw.3"    "nurdayw.4"    "hiqual_dv.9"  "bmival"
"cufsize.3"    "uscmm..22"
[113] "vpcheck.2"   "dorm.7"       "vpprob2.1"    "aceinh.1"     "uscmg.2"
"jbstat.5"     "vphand.2"     "ecre"
[121] "dorm.6"      "vpstimemm"    "nurdayw.6"    "rtin"         "ast"
"dorm.5"       "hhsize.7"     "ethnic"
[129] "uscmm.1"     "difbpc4.1"    "jbstat.97"    "mmgstp.4"     "medtyp2.1"
"lungsmok.2"   "medtyp8.1"
[1] "———————Ordinary Linear Regression (Improved)———————"

Call:
lm(formula = y ~ ., data = as.data.frame(x.data.linear))

Residuals:
     Min       1Q   Median       3Q      Max
-14.5028  -2.9706  -0.7764   1.9387  24.7134

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.06205    0.28219  35.658  < 2e-16 ***
diur.1      -0.79227    0.26295  -3.013 0.002599 **
lipid.1     -0.62156    0.20695  -3.003 0.002682 **
medtyp4.1    1.78195    0.20087   8.871  < 2e-16 ***
medtyp5.1    1.05789    0.43635   2.424 0.015367 *
medtyp6.1   -0.61764    0.20538  -3.007 0.002647 **
medtyp10.1  -1.16232    0.25492  -4.560 5.24e-06 ***
medcnj.2    -0.40833    0.17783  -2.296 0.021705 *
relwaitb.2   2.41160    1.10895   2.175 0.029699 *
htfev        0.35218    0.13336   2.641 0.008294 **
htpef       -0.28902    0.11070  -2.611 0.009058 **
samdifc4.1  -0.57377    0.27917  -2.055 0.039903 *
vpalco.2    -0.65518    0.25275  -2.592 0.009563 **
nseqno.6     4.90321    2.33880   2.096 0.036088 *
omdiast      0.27499    0.09410   2.922 0.003491 **
ag16g10.26   0.78366    0.24698   3.173 0.001518 **
ag16g10.36   0.72717    0.21597   3.367 0.000765 ***
ag16g10.46   0.87426    0.20147   4.339 1.46e-05 ***
hhsize.3    -0.35142    0.17130  -2.051 0.040274 *
hhsize.8     3.85952    1.78225   2.166 0.030391 *
jbstat.3     1.33494    0.37043   3.604 0.000316 ***
jbstat.4    -0.88094    0.20747  -4.246 2.21e-05 ***
jbstat.6     0.63912    0.30473   2.097 0.036015 *
jbstat.8     1.93333    0.48953   3.949 7.94e-05 ***
sf1.2        1.07970    0.18692   5.776 8.07e-09 ***
sf1.3        1.99495    0.20422   9.769  < 2e-16 ***
sf1.4        3.63429    0.26223  13.859  < 2e-16 ***
sf1.5        6.85592    0.42337  16.193  < 2e-16 ***
health.2    -0.91962    0.16402  -5.607 2.16e-08 ***
marstat.2   -0.33985    0.13609  -2.497 0.012544 *
dorm.9      -0.44807    0.22264  -2.013 0.044217 *
dorm.12     -0.67898    0.28770  -2.360 0.018308 *
dheas        0.18547    0.07240   2.562 0.010440 *
testo       -0.18941    0.09248  -2.048 0.040593 *
hhorig.y.4   9.61634    2.70427   3.556 0.000380 ***
mmgsdval    -0.20657    0.09933  -2.080 0.037610 *
```

```
omsysval     -0.21835     0.09557   -2.285 0.022362 *
htval        -0.23780     0.10491   -2.267 0.023447 *
ompulval     -0.13761     0.06703   -2.053 0.040130 *
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1      1

Residual standard error: 4.662 on 5323 degrees of freedom
Multiple R-squared:  0.2188,    Adjusted R-squared:  0.2132
F-statistic: 39.22 on 38 and 5323 DF,  p-value: < 2.2e-16

AIC: 31766.1546
MSE: 21.5745
[1] "The MSE of the predicted values are of 21.255"
[1] "The Linear Model predicts exactly with accuracy of 0.0951"
[1] "The Linear Model predicts within a confidence interval with accuracy of 0.226"
[1] "-----------Elastic Net Regression-----------"
185 x 1 sparse Matrix of class "dgCMatrix", with 19 entries
         names Estimate_Coefs
1  (Intercept)    10.94870637
2          age    -0.07374312
3      lipid.1    -0.15981112
4    medtyp4.1     1.77390305
5   lungsmok.2    -0.11854353
6   ag16g20.36     0.15055830
7   ag16g20.56    -0.18061145
8     jbstat.3     0.44057481
9     jbstat.4    -0.92794620
10    jbstat.8     1.24520719
11       sf1.3     0.28753193
12       sf1.4     1.66543825
13       sf1.5     4.36597553
14    health.2    -0.78749736
15       testo    -0.01841654
16  hhorig.y.4     0.02256058
17    mmgsdval    -0.06641539
18       htval    -0.09287567
19     bfpcval     0.06780254
[1] "The MSE of the predicted values of the best fit model is 21.1209"
[1] "The Alpha of the best fit model is 1"
[1] "The Elastic Net Model predicts exactly with accuracy of 0.0968"
[1] "-----------Timer Results-----------"
   user  system elapsed
  31.04    0.14   31.25
```

10.2.29 Titanic console

```
[1] "-----------Initial Checks-----------"
[1] "866 NA cells were found across the entire dataset (8.1% of data as NA)"
[1] "-----------Dependant Variable Legend (Use to understand cluster models)-----------"
     factorNames values
1 Didn't Survive      1
2       Survived      2
[1] "-----------Data Type Checks-----------"
[1] "5 variables recoded since all their entries aren't numeric or NA"
[1] "NOTE: algorithim recodes categorical data alphabetically e.g (female = 1, male = 2)"
[1] "Name"      "Sex"       "Ticket"    "Cabin"     "Embarked"
[1] "-----------Low Data Removal-----------"
[1] "1 variables removed since they had >= 'naPercent' (default 20%) NA values"
[1] "Cabin"
[1] "-----------Low Level Removal-----------"
[1] "If a level is removed from a variable you wish to keep, reccomended to manually merge levels to
[1] "level 4 in Parch removed, 4 observations found"
[1] "level 6 in Parch removed, 1 observations found"
[1] "2 total levels removed from 1 different variables. In total 5 observations deleted"
[1] "-----------Variance 0 Check-----------"
[1] "0 variables removed since their new variance was 0"
```

```
character(0)
[1] "—————Dummy Variables—————"
[1] "predictor variable count went from 10 to 21"
[1] "—————Variance 0 Check—————"
[1] "2 variables removed since their new variance was 0"
[1] "Parch.4" "Parch.6"
[1] "—————K—Means—————"
[1] "2 clusters have been made for K—Means"
[1] "K—Means results as a table, the max value in each row is a simple way to define which cluster re

      1   2
  1 195 124
  2 224 164
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
[1] "Cluster 1: Within MSE 126679, Size 319" "Cluster 2: Within MSE 128720, Size 388"
[1] "Total between cluster MSE: 52977, Total within cluster MSE: 127799"
[1] "The K—Means model predicts exactly with an accuracy of 0.5926"
[1] "—————Correlation Checks—————"
[1] "0 variables removed since they had high correlation coefs"
[1] "The final cleaned dataset has been completed at this stage and is stored under the name 'cleanDa
[1] "—————Attempting a Train Test Split—————"
[1] "Good train, test split found"
[1] "The working seed found was 3"
[1] "—————kNN—————"
[1] "30 neighbours considered for each test data point"
[1] "kNN results as a table, follow the diagonal for the correctly mapped clusters"
         real
predicted   1   2
        1 104  36
        2   6  31
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
[1] "The MSE of the predicted values are of 0.0113"
[1] "The kNN model predicts exactly with an accuracy of 0.7627"
[1] "—————CART prediction model—————"
n= 530

node), split, n, deviance, yval
      * denotes terminal node

  1) root 530 128.8472000 1.416981
    2) Sex.male>=0.5 332  56.3855400 1.216867
      4) Age>=-1.143733 305  45.0819700 1.180328
        8) Fare< 0.3436248 268  31.8917900 1.138060
         16) Ticket< 1.555194 255  27.2313700 1.121569
           32) Fare< -0.1525384 210  18.0952400 1.095238 *
           33) Fare>=-0.1525384 45   8.3111110 1.244444
             66) Fare>=-0.06442615 23   1.8260870 1.086957 *
             67) Fare< -0.06442615 22   5.3181820 1.409091
              134) PassengerId< 0.4470703 15   2.4000000 1.200000 *
              135) PassengerId>=0.4470703 7   0.8571429 1.857143 *
         17) Ticket>=1.555194 13   3.2307690 1.461538 *
        9) Fare>=0.3436248 37   9.2432430 1.486486
         18) Name>=0.2926589 11   0.0000000 1.000000 *
         19) Name< 0.2926589 26   5.5384620 1.692308
           38) PassengerId< -1.026823 7   0.8571429 1.142857 *
           39) PassengerId>=-1.026823 19   1.7894740 1.894737 *
      5) Age< -1.143733 27   6.2962960 1.629630
       10) SibSp.1< 0.5 15   3.3333330 1.333333 *
       11) SibSp.1>=0.5 12   0.0000000 2.000000 *
    3) Sex.male< 0.5 198  36.8737400 1.752525
      6) Pclass.3>=0.5 78  19.4871800 1.487179
       12) Fare>=-0.2132273 13   0.9230769 1.076923 *
       13) Fare< -0.2132273 65  15.9384600 1.569231 *
      7) Pclass.3< 0.5 120   8.3250000 1.925000 *
[1] "Variable Importance"
   Sex.male        Fare         Age    Pclass.3      Ticket PassengerId
Name      Parch.2      SibSp.1      SibSp.4
 35.5878903  19.5449523  11.9058290  10.2083271   9.3041297   7.2408534
```

```
5.9065782    3.3413589    2.9629630    1.3844039
 Embarked.Q      Pclass.2       Parch.5       SibSp.3      SibSp.5
  1.0455644    1.0455644    0.4039448    0.3709095    0.1854548
[1] "The MSE of the predicted values are of 0.1396"
[1] "The CART model predicts exactly with accuracy of 0.8136"
[1] "------------Ordinary Linear Regression (Initial)------------"
[1] "The full model AIC is: 531.4729"
[1] "------------Variance Inflation Factor Removal------------"
[1] "0 variables removed from the Ordinary Linear Model since they have a VIF score higher than 10"
[1] "The full model AIC after VIF checks is: 531.4729"
[1] "------------Backwards Selection------------"
[1] "14 out of 19 variables removed in backwards selection since they weren't significant at the 95%
 [1] "Parch.3"      "Ticket"       "SibSp.1"      "PassengerId" "Embarked.S"
"Embarked.Q"  "Parch.2"      "Parch.1"
 [9] "SibSp.2"      "Name"         "Parch.5"      "Fare"         "SibSp.5"
"SibSp.4"
[1] "------------Ordinary Linear Regression (Improved)------------"

Call:
lm(formula = y ~ ., data = as.data.frame(x.data.linear))

Residuals:
     Min       1Q    Median       3Q      Max
-1.11961 -0.24036 -0.08313  0.25484  0.98300

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.95405    0.03843  50.850  < 2e-16 ***
Pclass.2    -0.21959    0.04959  -4.428 1.16e-05 ***
Pclass.3    -0.38046    0.04463  -8.524  < 2e-16 ***
Sex.male    -0.46364    0.03652 -12.697  < 2e-16 ***
Age         -0.08693    0.01860  -4.674 3.76e-06 ***
SibSp.3     -0.33396    0.13413  -2.490   0.0131 *
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1       1

Residual standard error: 0.3928 on 524 degrees of freedom
Multiple R-squared:  0.3724,    Adjusted R-squared:  0.3664
F-statistic: 62.19 on 5 and 524 DF,  p-value: < 2.2e-16

AIC: 521.6136
MSE: 0.1526
[1] "The MSE of the predicted values are of 0.128"
[1] "The Linear Model predicts exactly with accuracy of 0.8249"
[1] "The Linear Model predicts within a confidence interval with accuracy of 0.8983"
[1] "------------Elastic Net Regression------------"
20 x 1 sparse Matrix of class "dgCMatrix", with 5 entries
        names Estimate_Coefs
1 (Intercept)     1.74982822
2    Pclass.3    -0.14507039
3    Sex.male    -0.41658008
4         Age    -0.01567631
5        Fare     0.03670566
[1] "The MSE of the predicted values of the best fit model is 0.13"
[1] "The Alpha of the best fit model is 1"
[1] "The Elastic Net Model predicts exactly with accuracy of 0.8192"
[1] "------------Timer Results------------"
   user  system elapsed
   0.69    0.00    0.69
```

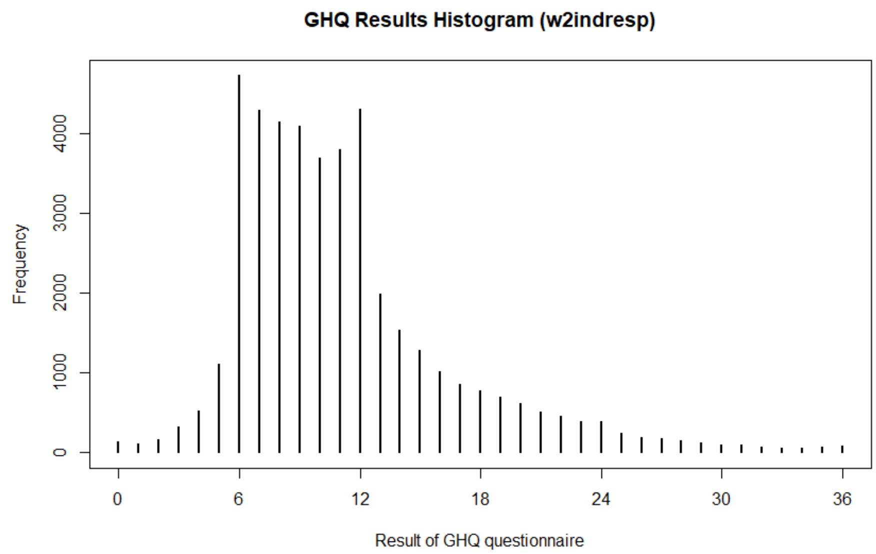
10.2.30 Iris console

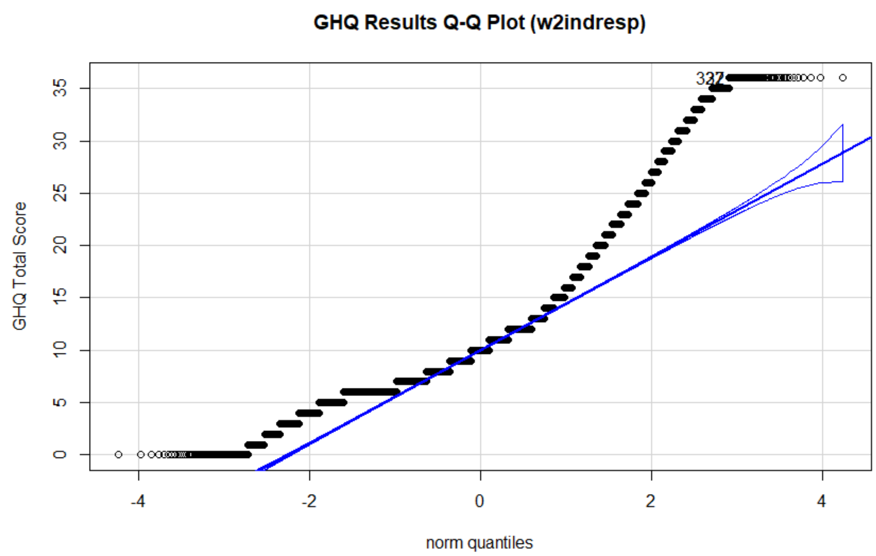
```
[1] "------------Initial Checks------------"
[1] "0 NA cells were found across the entire dataset (0% of data as NA)"
[1] "------------Dependant Variable Legend (Use to understand cluster models)------------"
  factorNames values
1      setosa      1
```

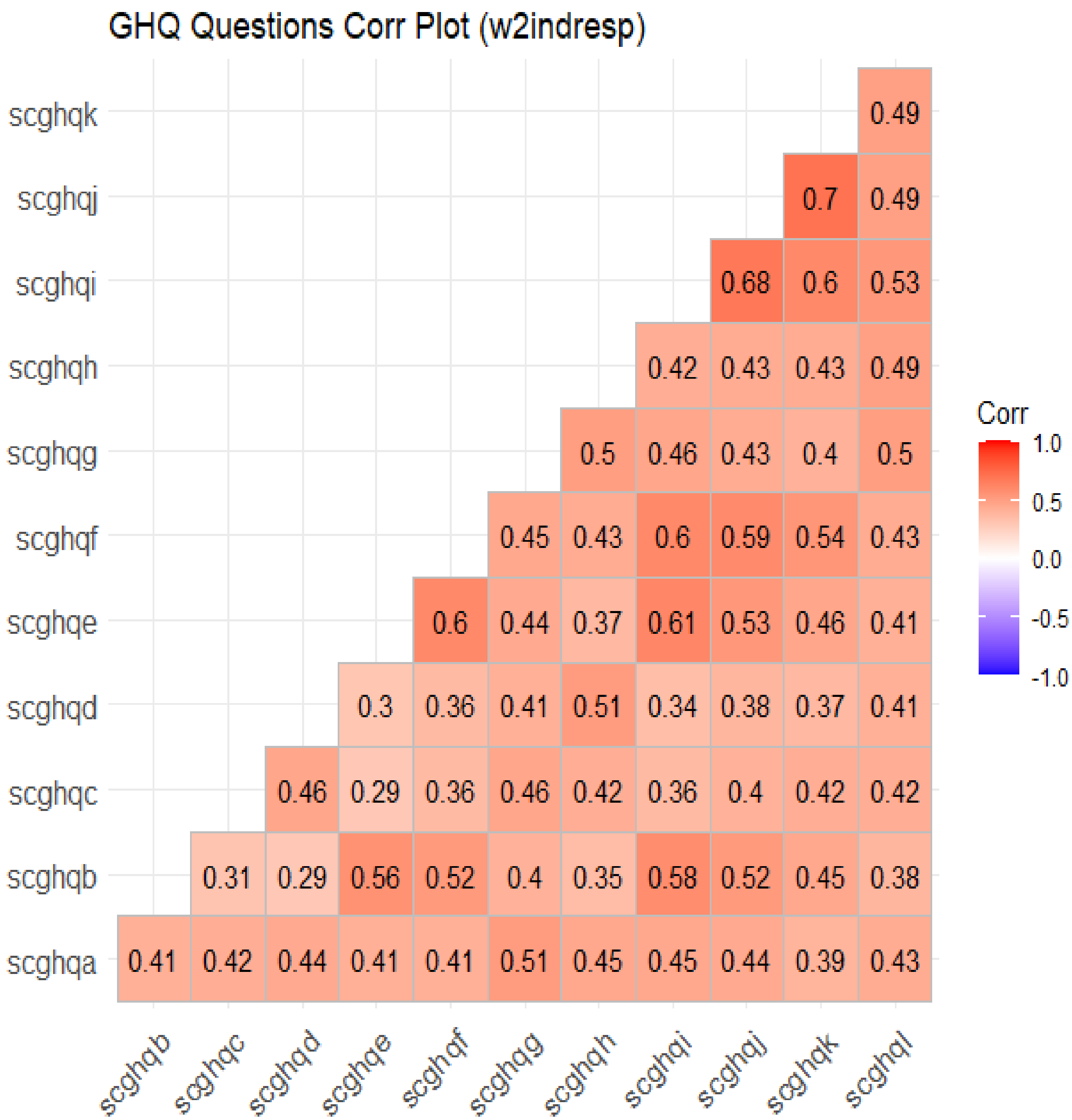
```
2   versicolor       2
3    virginica       3
[1] "————————Data Type Checks————————"
[1] "0 variables recoded since all their entries aren't numeric or NA"
[1] "NOTE: algorithim recodes categorical data alphabetically e.g (female = 1, male = 2)"
character(0)
[1] "————————Low Data Removal————————"
[1] "0 variables removed since they had >= 'naPercent' (default 20%) NA values"
character(0)
[1] "————————Low Level Removal————————"
[1] "If a level is removed from a variable you wish to keep, reccomended to manually merge levels toc
[1] "0 total levels removed from 0 different variables. In total 0 observations deleted"
[1] "————————Variance 0 Check————————"
[1] "0 variables removed since their new variance was 0"
character(0)
[1] "————————Dummy Variables————————"
[1] "predictor variable count went from 4 to 4"
[1] "————————Variance 0 Check————————"
[1] "0 variables removed since their new variance was 0"
character(0)
[1] "————————K—Means————————"
[1] "3 clusters have been made for K—Means"
[1] "K—Means results as a table, the max value in each row is a simple way to define which cluster re

     1  2  3
  1  0 48 14
  2  0  2 36
  3 50  0  0
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
[1] "Cluster 1: Within MSE 1, Size 62" "Cluster 2: Within MSE 1, Size 38" "Cluster 3: Within MSE 0, S
[1] "Total between cluster MSE: 4, Total within cluster MSE: 1"
[1] "The K—Means model predicts exactly with an accuracy of 0.8933"
[1] "————————Correlation Checks————————"
[1] "0 variables removed since they had high correlation coefs"
[1] "The final cleaned dataset has been completed at this stage and is stored under the name 'cleanDa
[1] "————————Attempting a Train Test Split————————"
[1] "Good train, test split found"
[1] "The working seed found was 1"
[1] "————————kNN————————"
[1] "12 neighbours considered for each test data point"
[1] "kNN results as a table, follow the diagonal for the correctly mapped clusters"
         real
predicted  1  2  3
        1 15  0  0
        2  0  8  0
        3  0  2 13
[1] "CAUTION: Be careful comparing the MSE of this classification model to the regression models"
[1] "The MSE of the predicted values are of 0.1053"
[1] "The kNN model predicts exactly with an accuracy of 0.9474"
[1] "————————CART prediction model————————"
n= 112

node), split, n, deviance, yval
      * denotes terminal node

 1) root 112 71.964290 2.017857
   2) Petal.Length< -0.7409513 35  0.000000 1.000000 *
   3) Petal.Length>=-0.7409513 77 19.220780 2.480519
     6) Petal.Length< 0.5619447 37  0.972973 2.027027 *
     7) Petal.Length>=0.5619447 40  3.600000 2.900000
      14) Petal.Length< 0.7318877 11  2.545455 2.636364 *
      15) Petal.Length>=0.7318877 29  0.000000 3.000000 *
[1] "Variable Importance"
Petal.Length  Petal.Width Sepal.Length  Sepal.Width
    68.44586     64.71597     46.59445     28.65781
[1] "The MSE of the predicted values are of 0.0406"
[1] "The CART model predicts exactly with accuracy of 0.9474"
[1] "————————Ordinary Linear Regression (Initial)————————"
```

```
[1] "The full model AIC is: -12.4783"
[1] "-----------Variance Inflation Factor Removal-----------"
[1] "The variable Petal.Length was removed since it had a VIF score of 33.0637"
[1] "1 variables removed from the Ordinary Linear Model since they have a VIF score higher than 10"
[1] "The full model AIC after VIF checks is: 2.4362"
[1] "-----------Backwards Selection-----------"
[1] "2 out of 3 variables removed in backwards selection since they weren't significant at the 95% co
[1] "Sepal.Width"  "Sepal.Length"
[1] "-----------Ordinary Linear Regression (Improved)-----------"

Call:
lm(formula = y ~ ., data = as.data.frame(x.data.linear))

Residuals:
     Min       1Q   Median       3Q      Max
-0.64029 -0.17486  0.03449  0.15036  0.77840

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.01155    0.02379   84.55   <2e-16 ***
Petal.Width  0.79787    0.02492   32.02   <2e-16 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1

Residual standard error: 0.2518 on 110 degrees of freedom
Multiple R-squared:  0.9031,    Adjusted R-squared:  0.9022
F-statistic:  1025 on 1 and 110 DF,  p-value: < 2.2e-16

AIC: 12.878
MSE: 0.0623
[1] "The MSE of the predicted values are of 0.0416"
[1] "The Linear Model predicts exactly with accuracy of 1"
[1] "The Linear Model predicts within a confidence interval with accuracy of 1"
[1] "-----------Elastic Net Regression-----------"
5 x 1 sparse Matrix of class "dgCMatrix", with 5 entries
        names Estimate_Coefs
1  (Intercept)     2.00682443
2 Sepal.Length     0.02668087
3  Sepal.Width    -0.06624363
4 Petal.Length     0.32826557
5  Petal.Width     0.38142479
[1] "The MSE of the predicted values of the best fit model is 0.0455"
[1] "The Alpha of the best fit model is 0.1"
[1] "The Elastic Net Model predicts exactly with accuracy of 0.9737"
[1] "-----------Timer Results-----------"
   user  system elapsed
   0.56    0.00    0.56
```

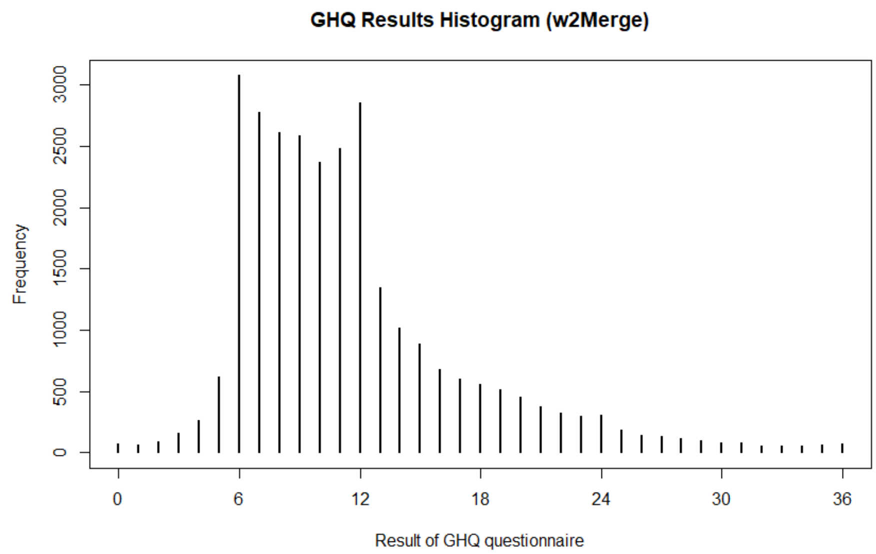
10.2.31 w2indresp graphs

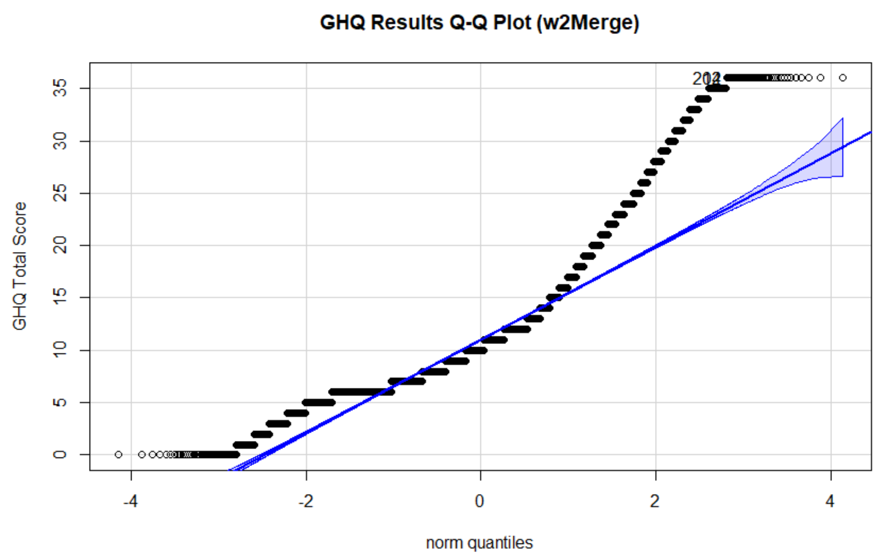


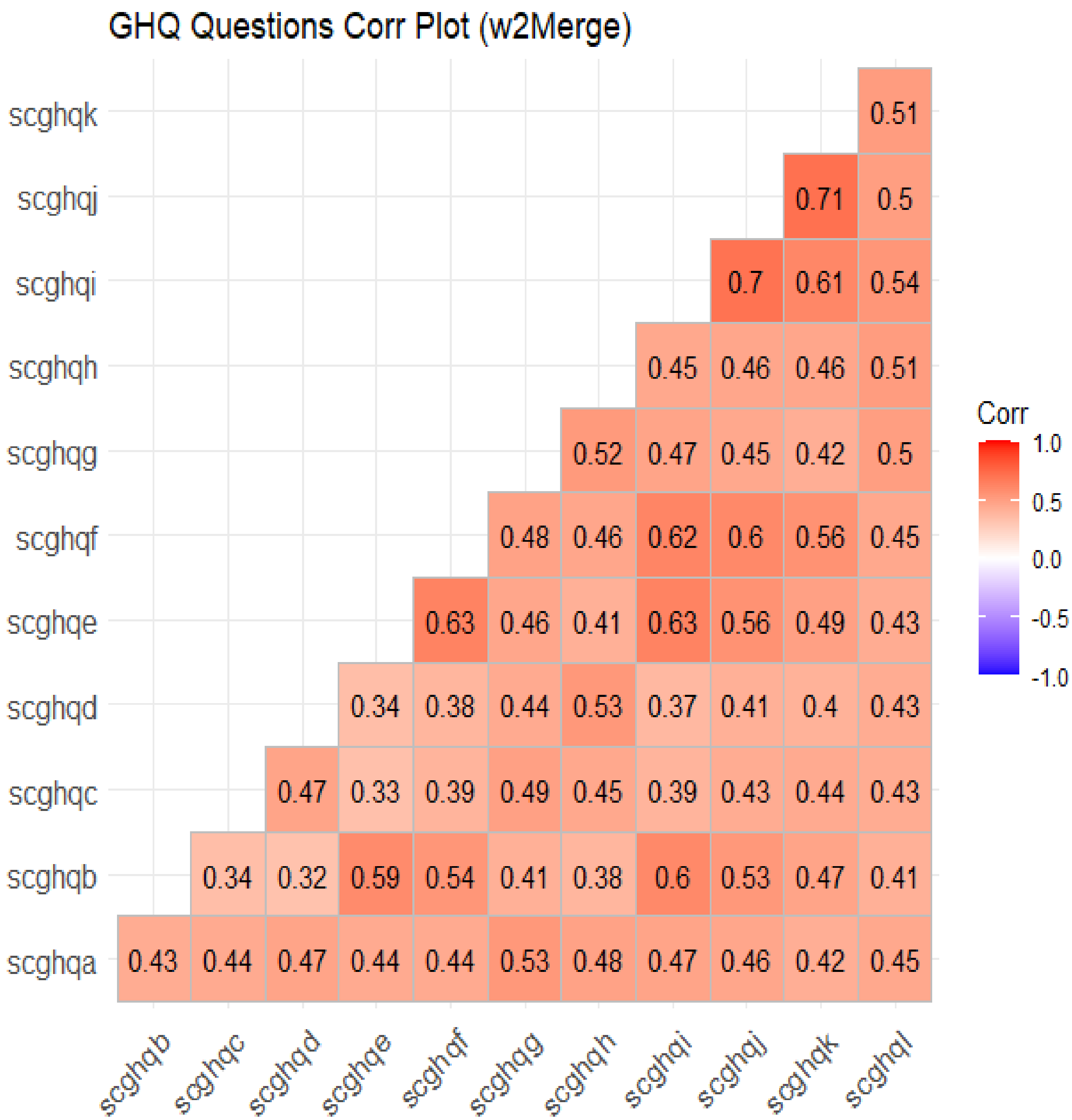



GHQ Questions Corr Plot (w2indresp)
scghqk
scghqj
scghqi
scghqh
scghqg
scghqf
scghqe
scghqd
scghqc
scghqb
scghqa
0.49
0.7 0.49
0.68 0.6 0.53
0.42 0.43 0.43 0.49
0.5 0.46 0.43 0.4 0.5
0.45 0.43 0.6 0.59 0.54 0.43
0.6 0.44 0.37 0.61 0.53 0.46 0.41
0.3 0.36 0.41 0.51 0.34 0.38 0.37 0.41
0.46 0.29 0.36 0.46 0.42 0.36 0.4 0.42 0.42
0.31 0.29 0.56 0.52 0.4 0.35 0.58 0.52 0.45 0.38
0.41 0.42 0.44 0.41 0.41 0.51 0.45 0.45 0.44 0.39 0.43
scghqb
scghqc
scghqd
scghqe
scghqf
scghqg
scghqh
scghqi
scghqj
scghqk
scghql
Corr
1.0
0.5
0.0
-0.5
-1.0

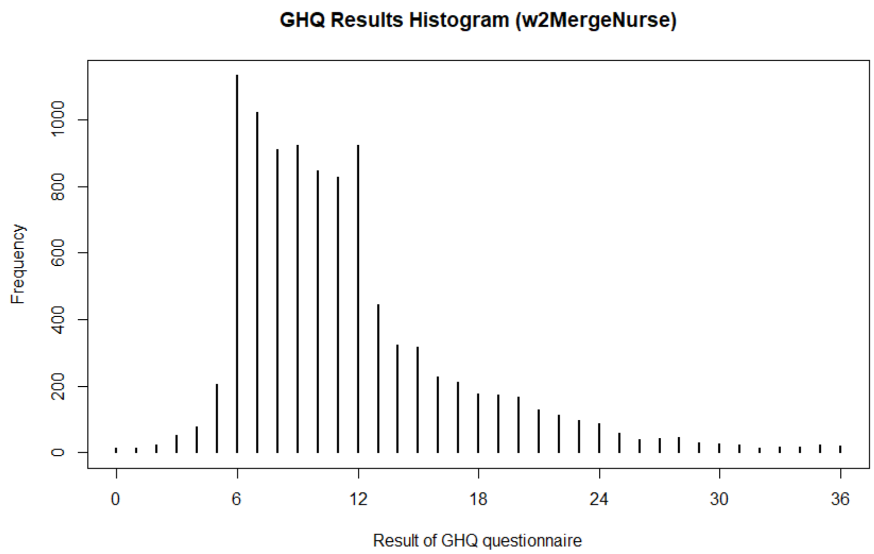
10.2.32 w2Merge graphs

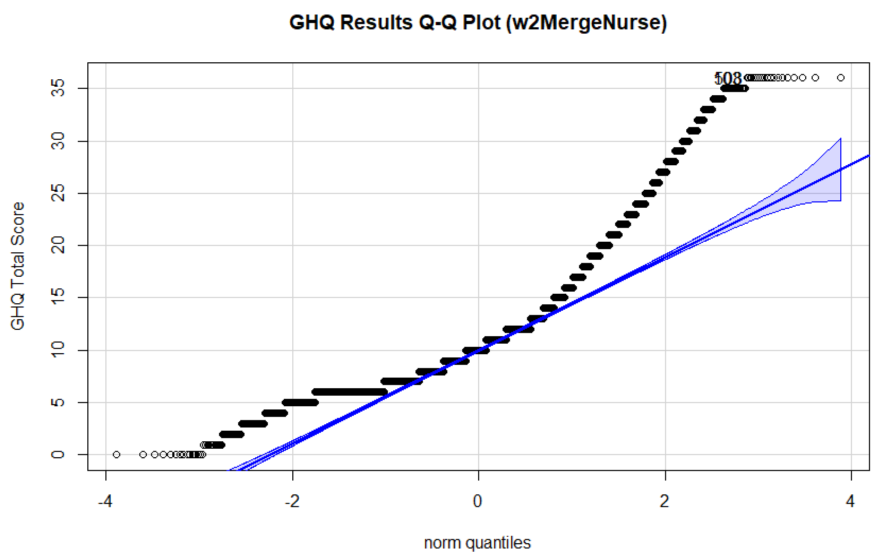


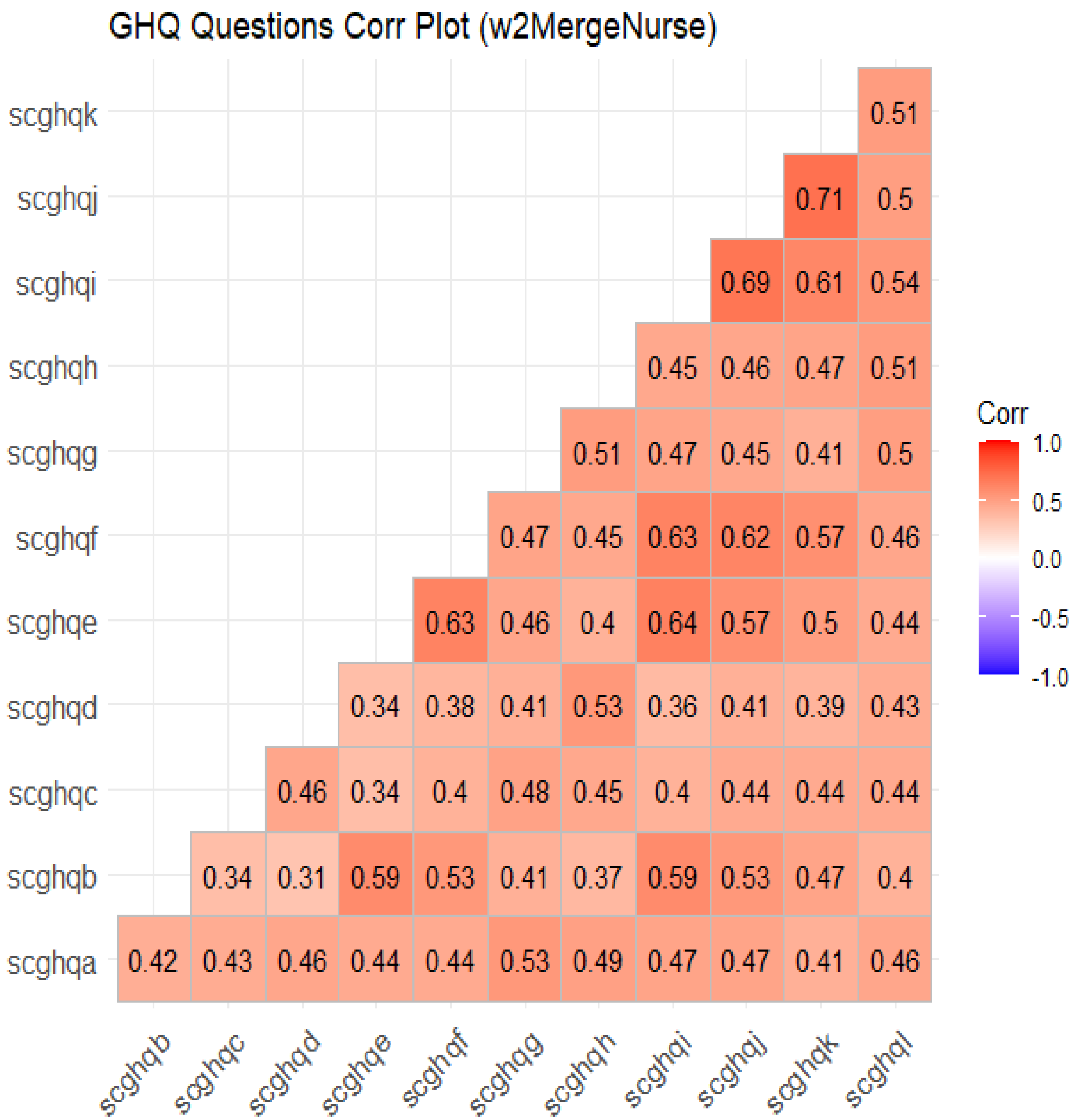



GHQ Questions Corr Plot (w2Merge)
scghqk
scghqj
scghqi
scghqh
scghqg
scghqf
scghqe
scghqd
scghqc
scghqb
scghqa
scghqb
scghqc
scghqd
scghqe
scghqf
scghqg
scghqh
scghqi
scghqj
scghqk
scghql
0.51
0.71 0.5
0.7 0.61 0.54
0.45 0.46 0.46 0.51
0.52 0.47 0.45 0.42 0.5
0.48 0.46 0.62 0.6 0.56 0.45
0.63 0.46 0.41 0.63 0.56 0.49 0.43
0.34 0.38 0.44 0.53 0.37 0.41 0.4 0.43
0.47 0.33 0.39 0.49 0.45 0.39 0.43 0.44 0.43
0.34 0.32 0.59 0.54 0.41 0.38 0.6 0.53 0.47 0.41
0.43 0.44 0.47 0.44 0.44 0.53 0.48 0.47 0.46 0.42 0.45
Corr
1.0
0.5
0.0
-0.5
-1.0

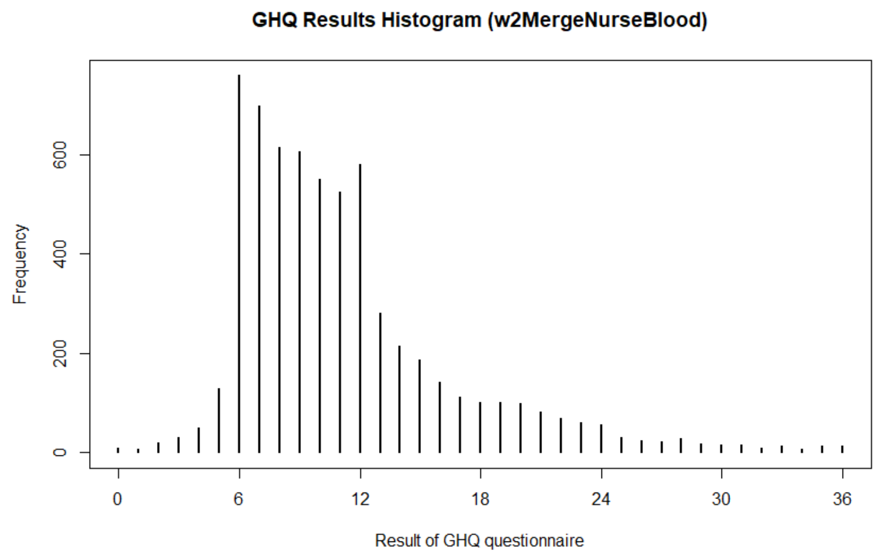
10.2.33 w2MergeNurse graphs

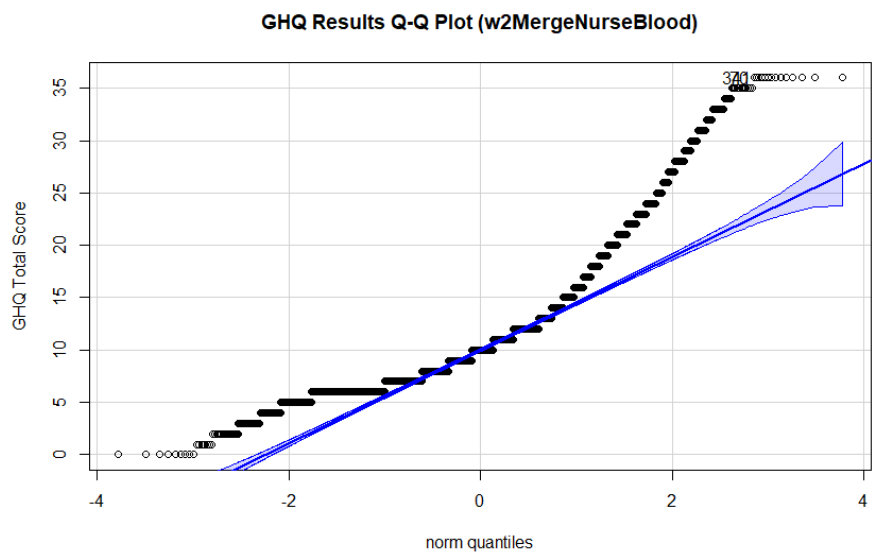


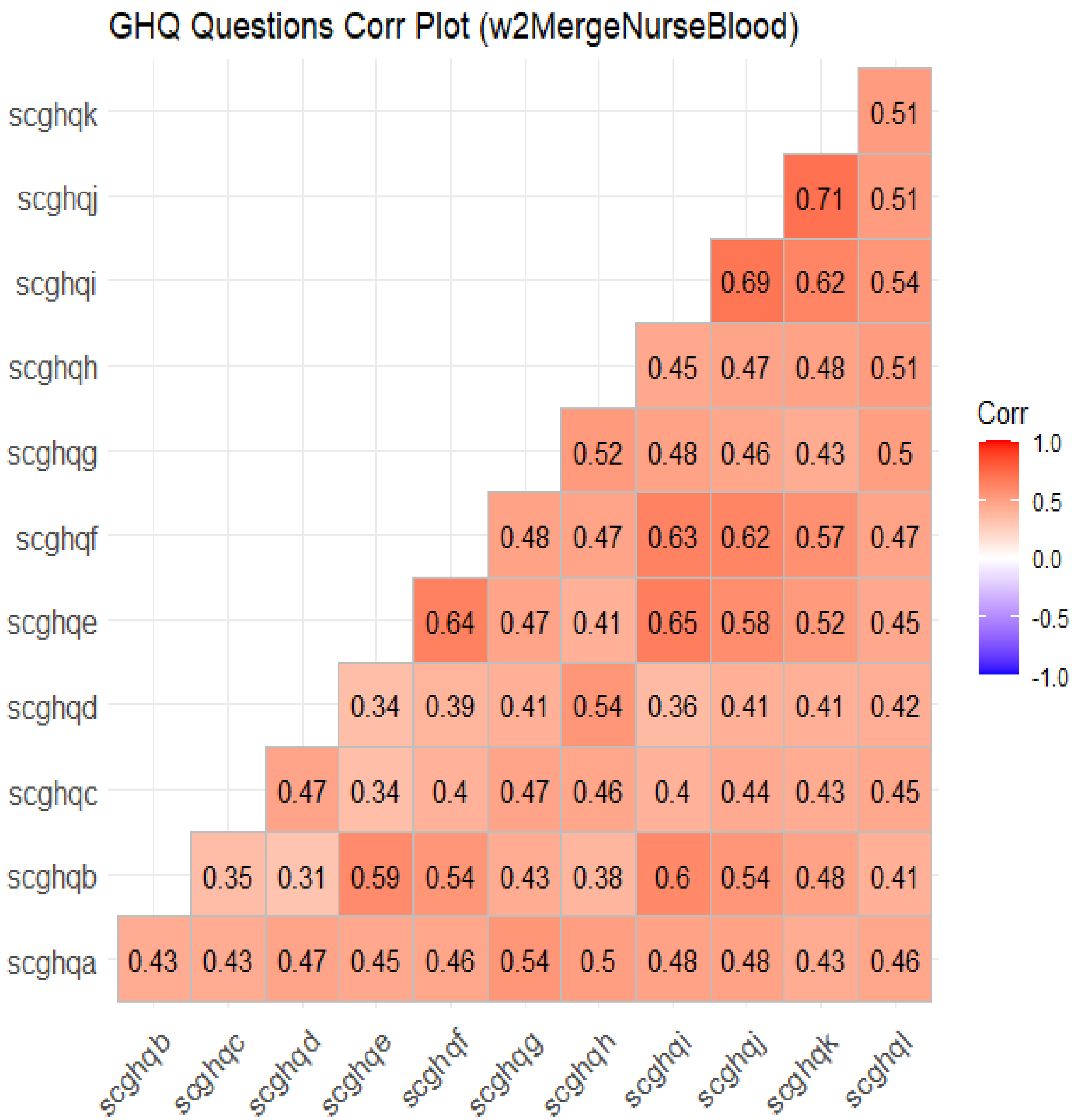



GHQ Questions Corr Plot (w2MergeNurse)
scghqk
scghqj
scghqi
scghqh
scghqg
scghqf
scghqe
scghqd
scghqc
scghqb
scghqa
0.51
0.71 0.5
0.69 0.61 0.54
0.45 0.46 0.47 0.51
0.51 0.47 0.45 0.41 0.5
0.47 0.45 0.63 0.62 0.57 0.46
0.63 0.46 0.4 0.64 0.57 0.5 0.44
0.34 0.38 0.41 0.53 0.36 0.41 0.39 0.43
0.46 0.34 0.4 0.48 0.45 0.4 0.44 0.44 0.44
0.34 0.31 0.59 0.53 0.41 0.37 0.59 0.53 0.47 0.4
0.42 0.43 0.46 0.44 0.44 0.53 0.49 0.47 0.47 0.41 0.46
scghqb
scghqc
scghqd
scghqe
scghqf
scghqg
scghqh
scghqi
scghqj
scghqk
scghql
Corr
1.0
0.5
0.0
-0.5
-1.0

10.2.34 w2MergeNurseBlood graphs

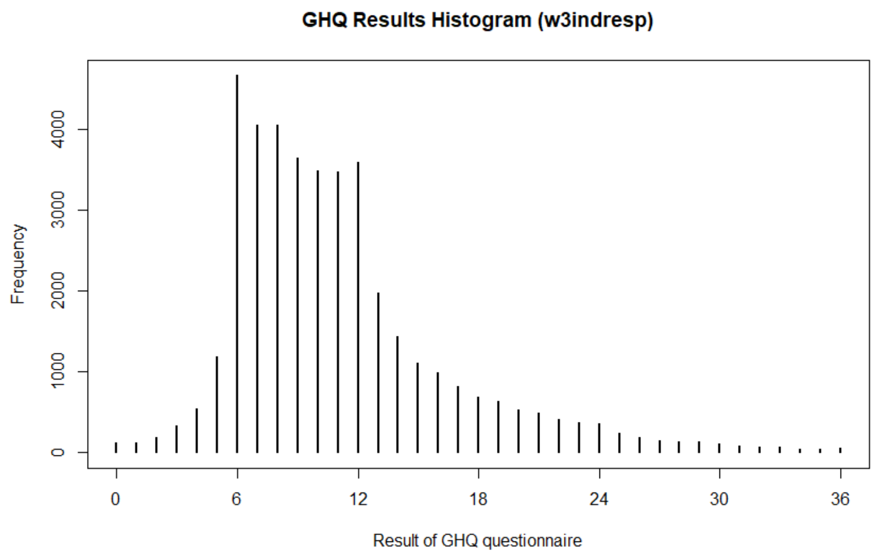


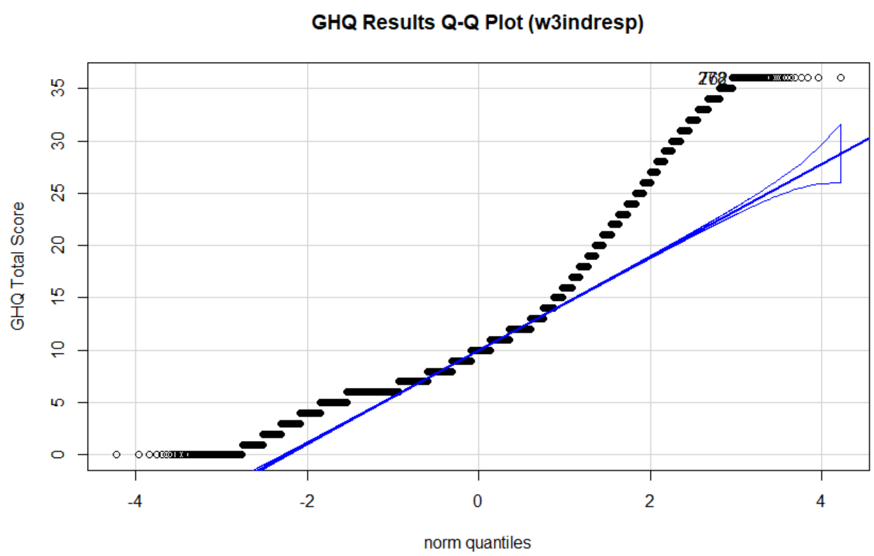


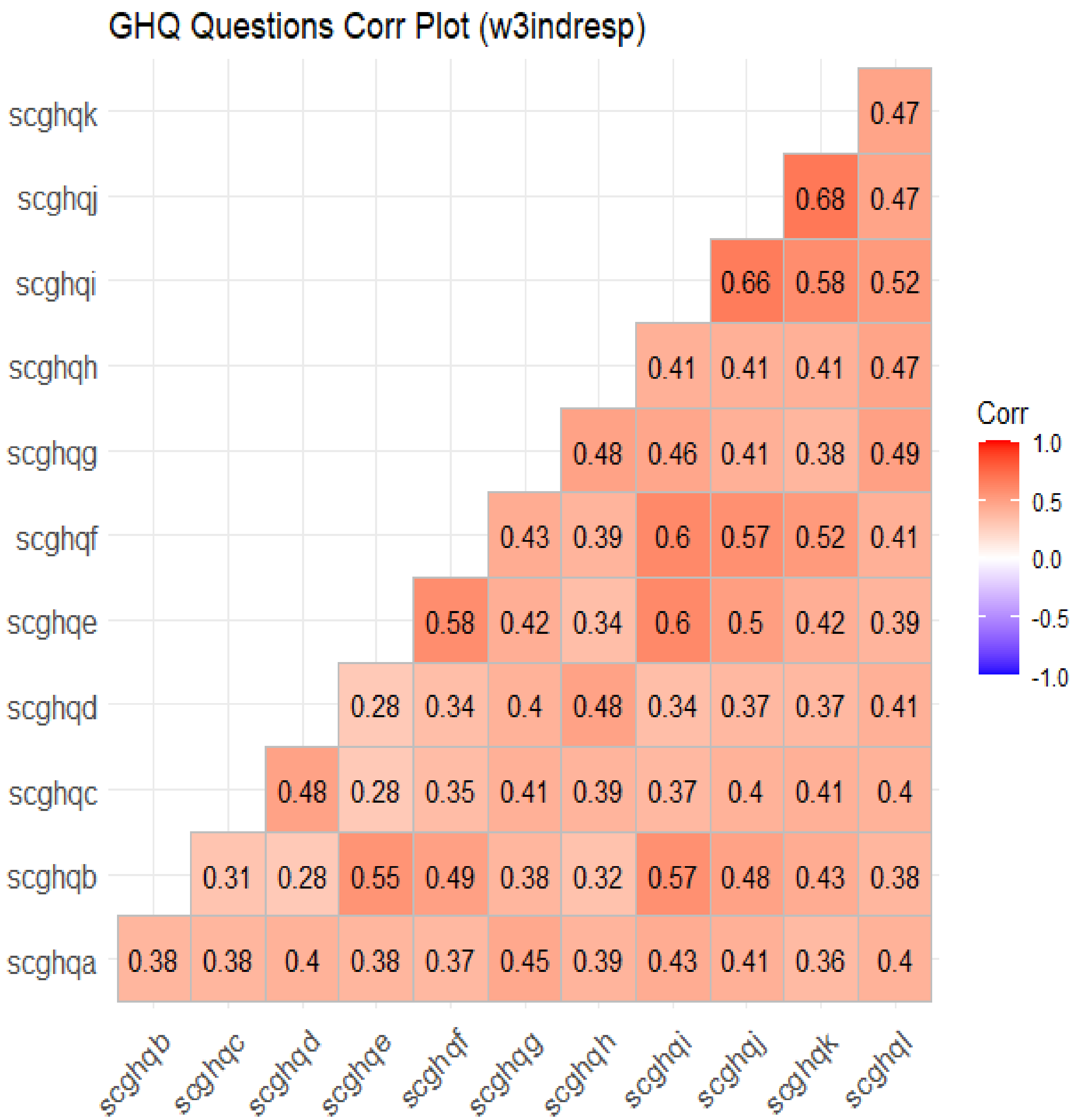


GHQ Questions Corr Plot (w2MergeNurseBlood)
scghqk
scghqj
scghqi
scghqh
scghqg
scghqf
scghqe
scghqd
scghqc
scghqb
scghqa
0.51
0.71 0.51
0.69 0.62 0.54
0.45 0.47 0.48 0.51
0.52 0.48 0.46 0.43 0.5
0.48 0.47 0.63 0.62 0.57 0.47
0.64 0.47 0.41 0.65 0.58 0.52 0.45
0.34 0.39 0.41 0.54 0.36 0.41 0.41 0.42
0.47 0.34 0.4 0.47 0.46 0.4 0.44 0.43 0.45
0.35 0.31 0.59 0.54 0.43 0.38 0.6 0.54 0.48 0.41
0.43 0.43 0.47 0.45 0.46 0.54 0.5 0.48 0.48 0.43 0.46
scghqb
scghqc
scghqd
scghqe
scghqf
scghqg
scghqh
scghqi
scghqj
scghqk
scghql
Corr
1.0
0.5
0.0
-0.5
-1.0

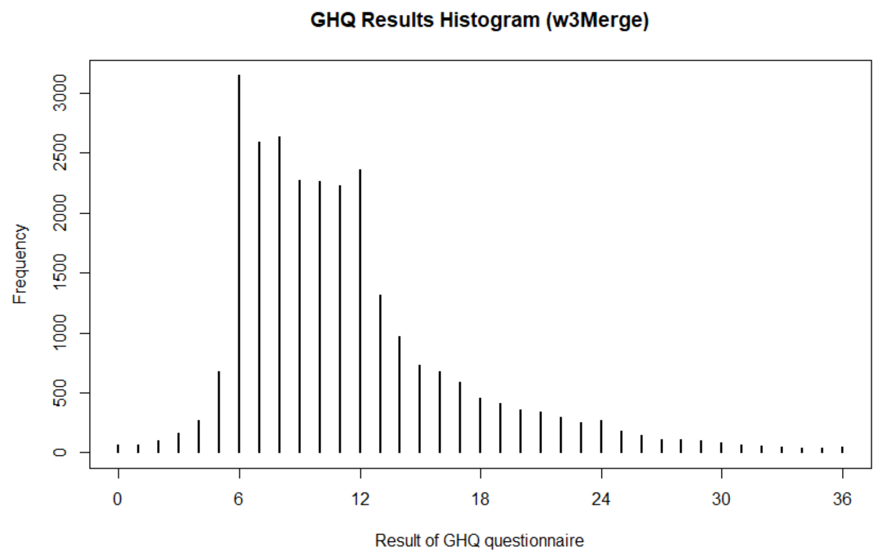
10.2.35 w3indresp graphs

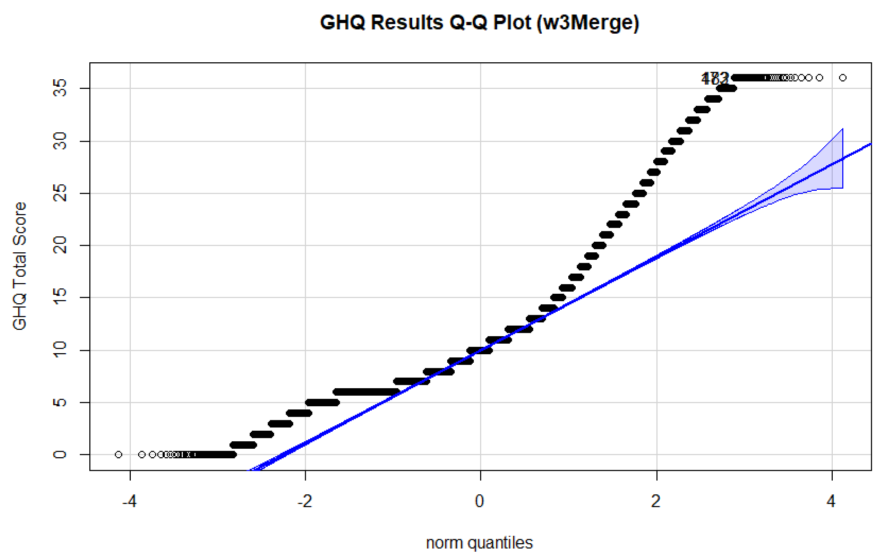


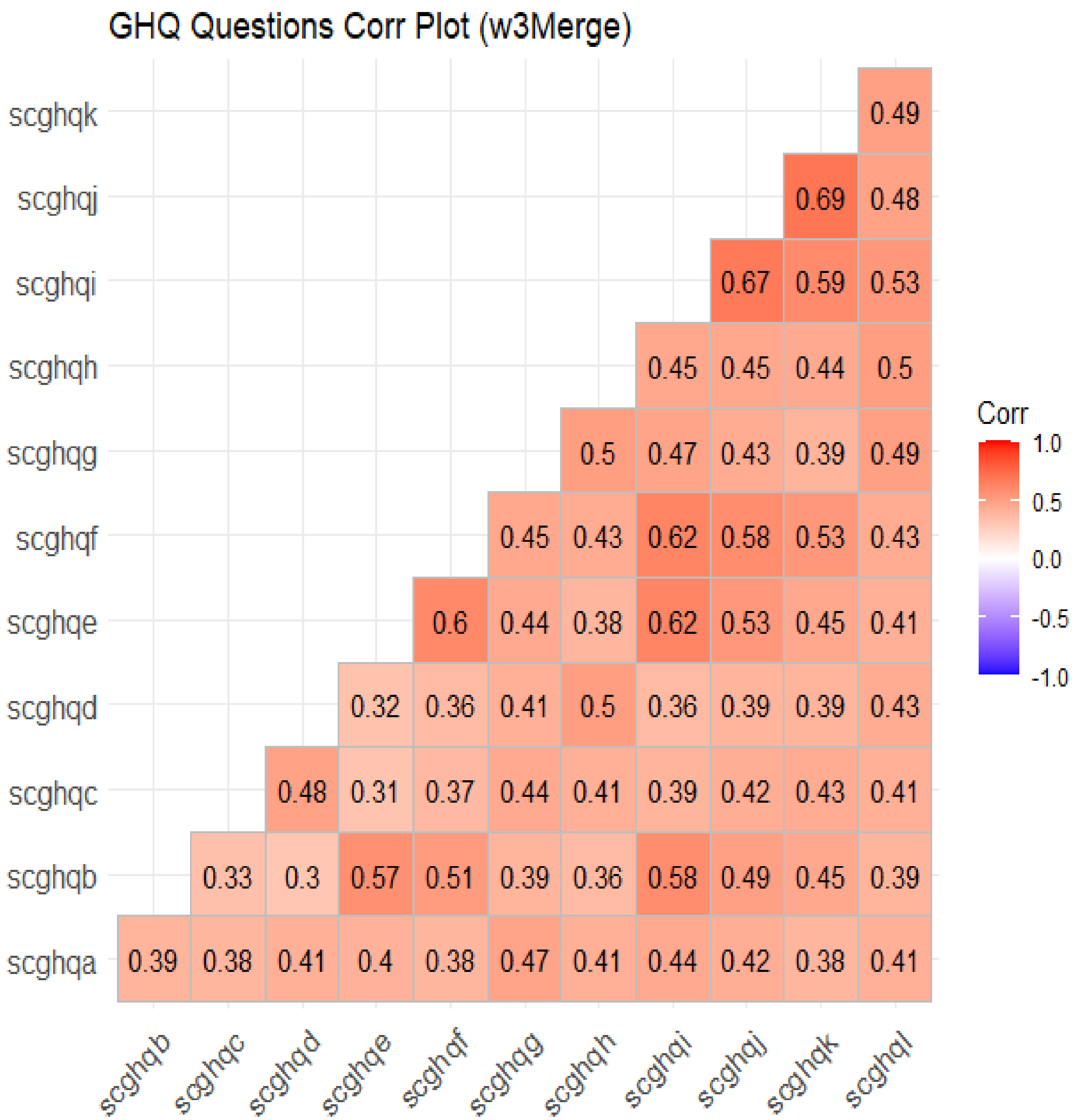



GHQ Questions Corr Plot (w3indresp)
scghqk
scghqj
scghqi
scghqh
scghqg
scghqf
scghqe
scghqd
scghqc
scghqb
scghqa
0.47
0.68 0.47
0.66 0.58 0.52
0.41 0.41 0.41 0.47
0.48 0.46 0.41 0.38 0.49
0.43 0.39 0.6 0.57 0.52 0.41
0.58 0.42 0.34 0.6 0.5 0.42 0.39
0.28 0.34 0.4 0.48 0.34 0.37 0.37 0.41
0.48 0.28 0.35 0.41 0.39 0.37 0.4 0.41 0.4
0.31 0.28 0.55 0.49 0.38 0.32 0.57 0.48 0.43 0.38
0.38 0.38 0.4 0.38 0.37 0.45 0.39 0.43 0.41 0.36 0.4
scghqb
scghqc
scghqd
scghqe
scghqf
scghqg
scghqh
scghqi
scghqj
scghqk
scghql
Corr
1.0
0.5
0.0
-0.5
-1.0

10.2.36 w3Merge graphs

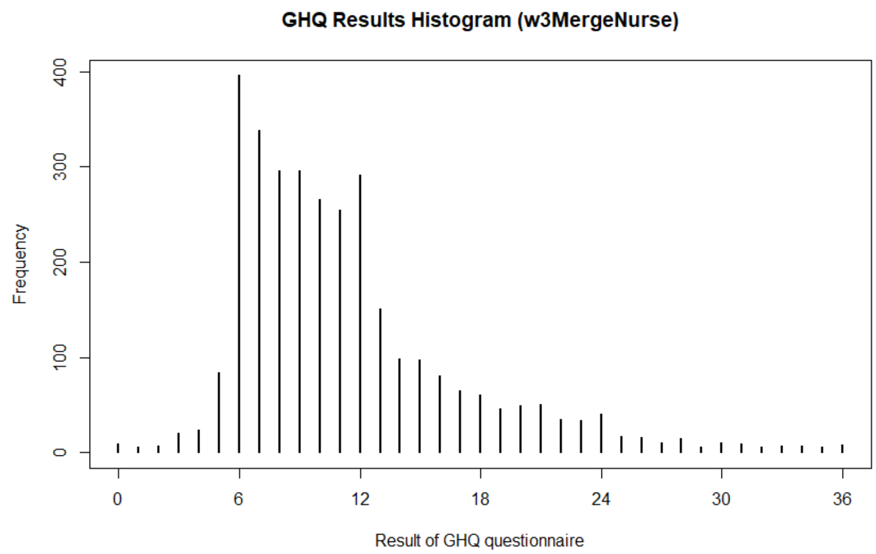


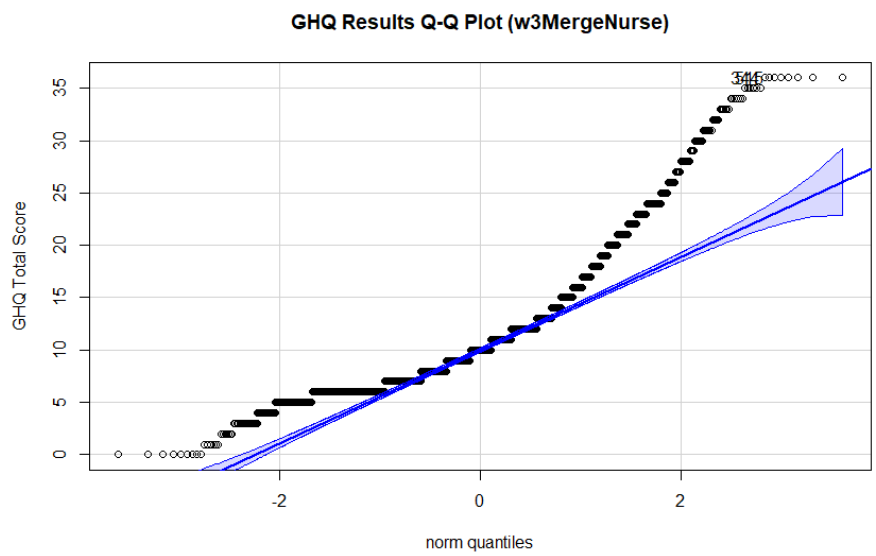


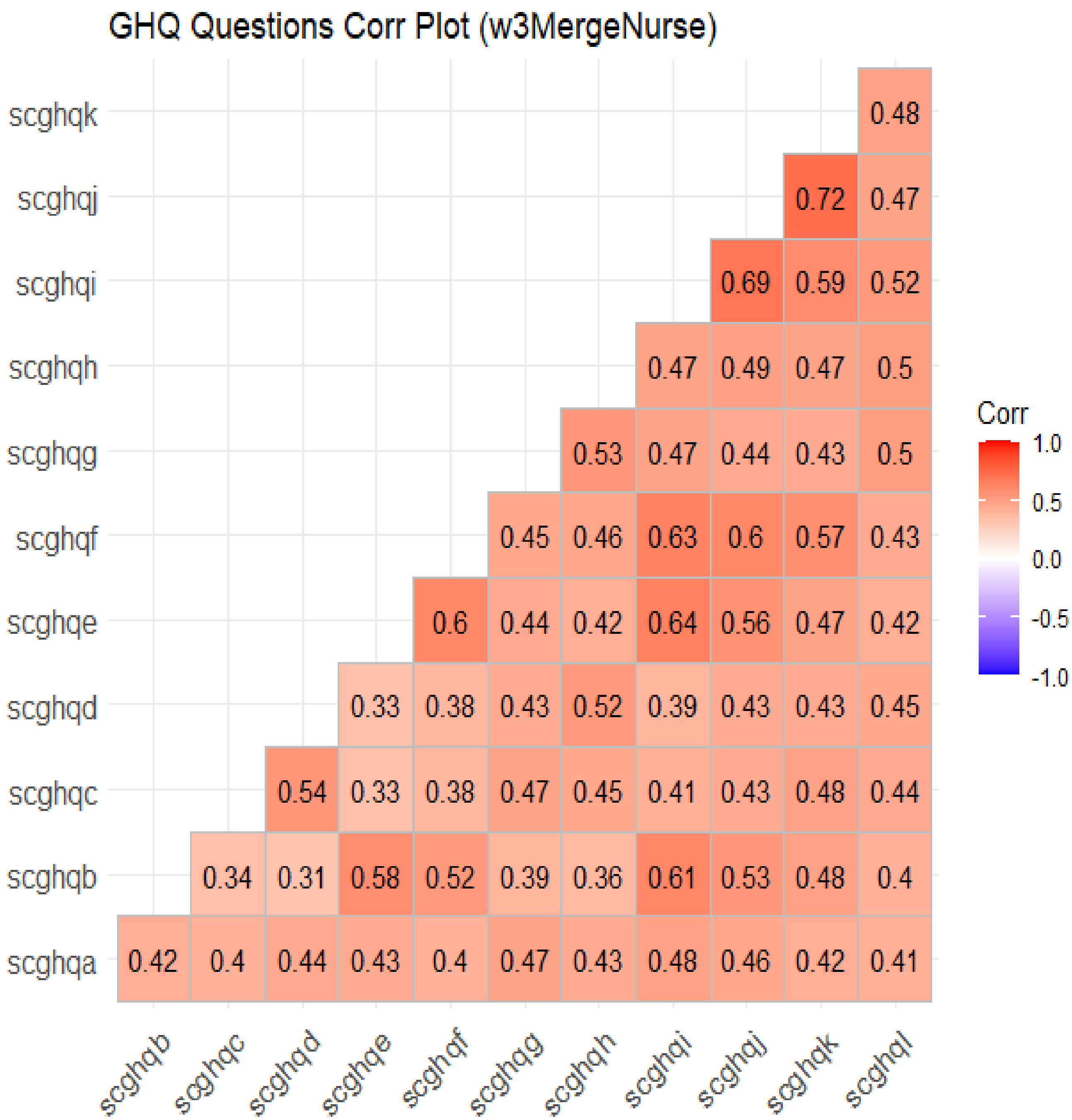


GHQ Questions Corr Plot (w3Merge)
scghqk
scghqj
scghqi
scghqh
scghqg
scghqf
scghqe
scghqd
scghqc
scghqb
scghqa
0.49
0.69 0.48
0.67 0.59 0.53
0.45 0.45 0.44 0.5
0.5 0.47 0.43 0.39 0.49
0.45 0.43 0.62 0.58 0.53 0.43
0.6 0.44 0.38 0.62 0.53 0.45 0.41
0.32 0.36 0.41 0.5 0.36 0.39 0.39 0.43
0.48 0.31 0.37 0.44 0.41 0.39 0.42 0.43 0.41
0.33 0.3 0.57 0.51 0.39 0.36 0.58 0.49 0.45 0.39
0.39 0.38 0.41 0.4 0.38 0.47 0.41 0.44 0.42 0.38 0.41
scghqb
scghqc
scghqd
scghqe
scghqf
scghqg
scghqh
scghqi
scghqj
scghqk
scghql
Corr
1.0
0.5
0.0
-0.5
-1.0

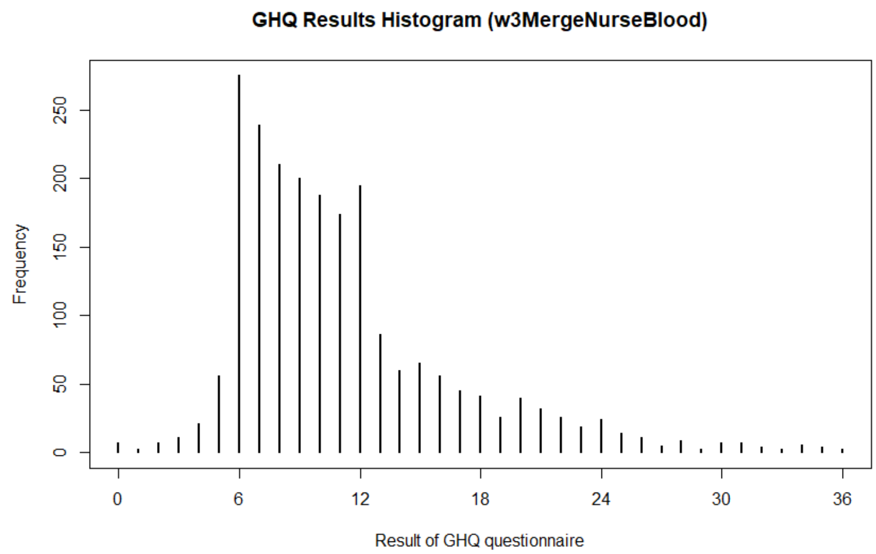
10.2.37 w3MergeNurse graphs

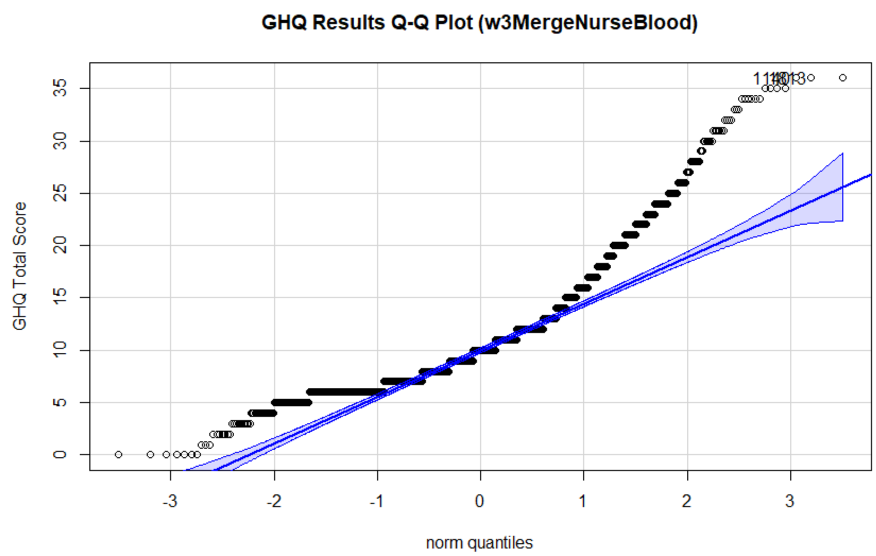


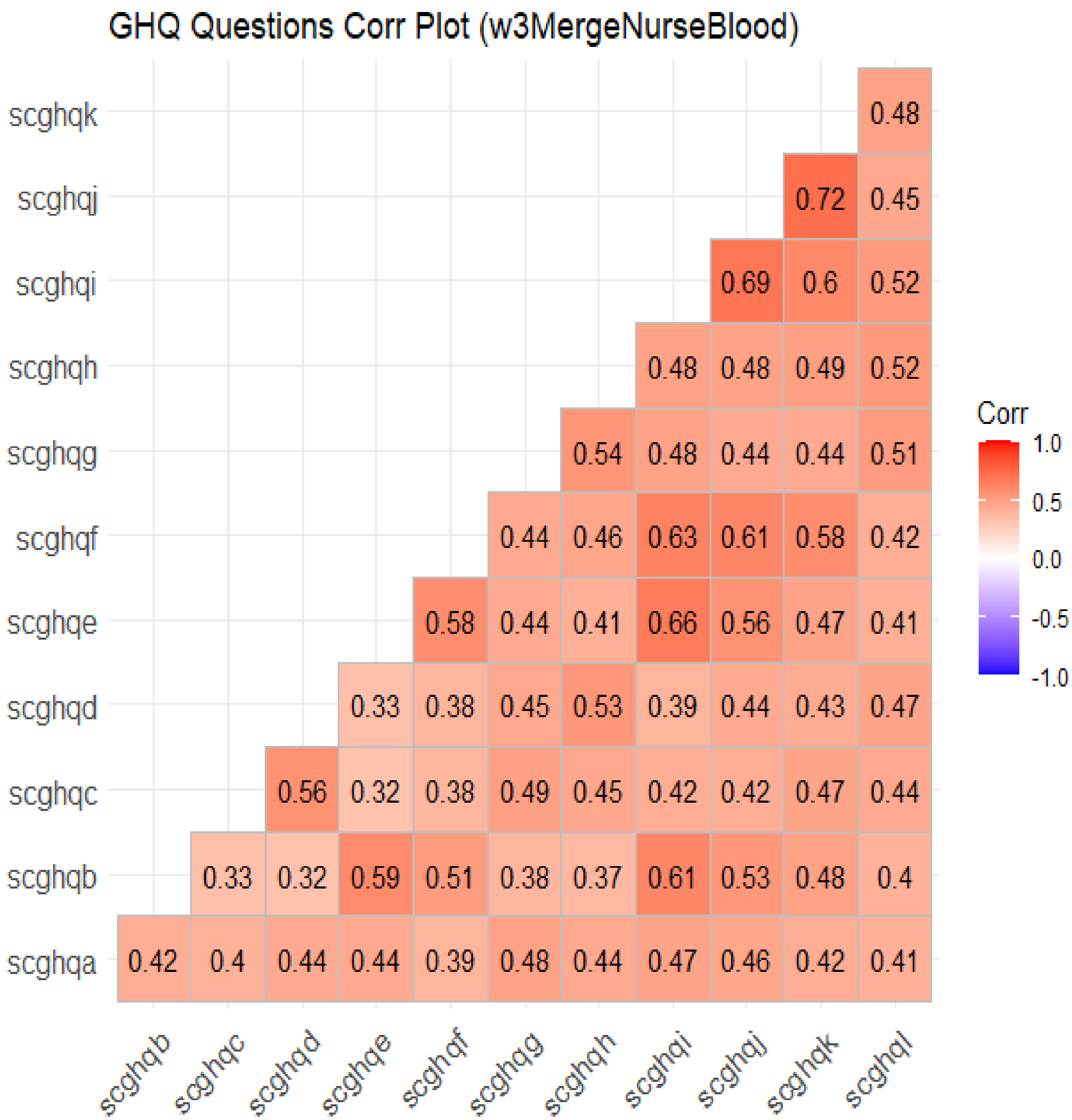



GHQ Questions Corr Plot (w3MergeNurse)
scghqk
scghqj
scghqi
scghqh
scghqg
scghqf
scghqe
scghqd
scghqc
scghqb
scghqa
scghqb
scghqc
scghqd
scghqe
scghqf
scghqg
scghqh
scghqi
scghqj
scghqk
scghql
0.48
0.72 0.47
0.69 0.59 0.52
0.47 0.49 0.47 0.5
0.53 0.47 0.44 0.43 0.5
0.45 0.46 0.63 0.6 0.57 0.43
0.6 0.44 0.42 0.64 0.56 0.47 0.42
0.33 0.38 0.43 0.52 0.39 0.43 0.43 0.45
0.54 0.33 0.38 0.47 0.45 0.41 0.43 0.48 0.44
0.34 0.31 0.58 0.52 0.39 0.36 0.61 0.53 0.48 0.4
0.42 0.4 0.44 0.43 0.4 0.47 0.43 0.48 0.46 0.42 0.41
Corr
1.0
0.5
0.0
-0.5
-1.0

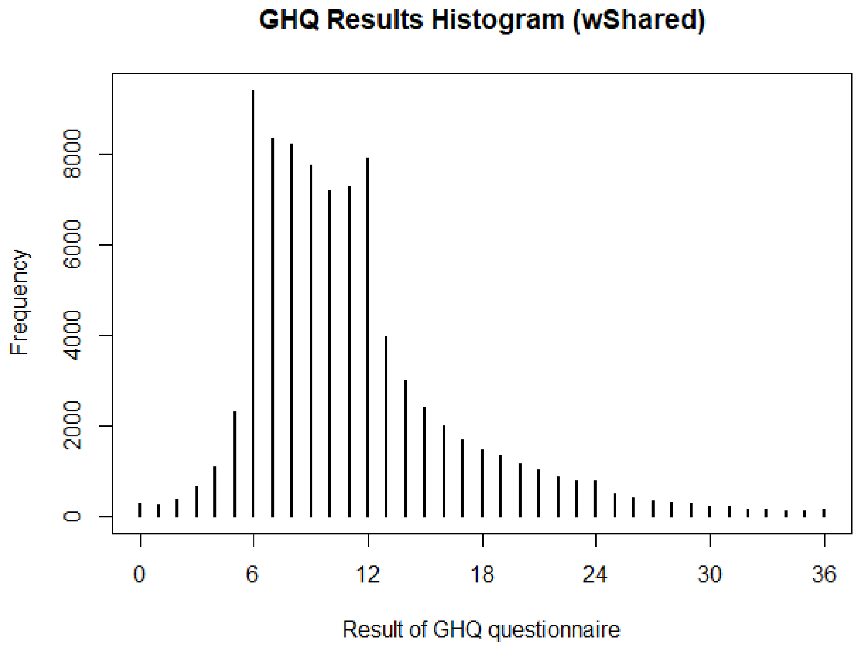
10.2.38 w3MergeNurseBlood graphs

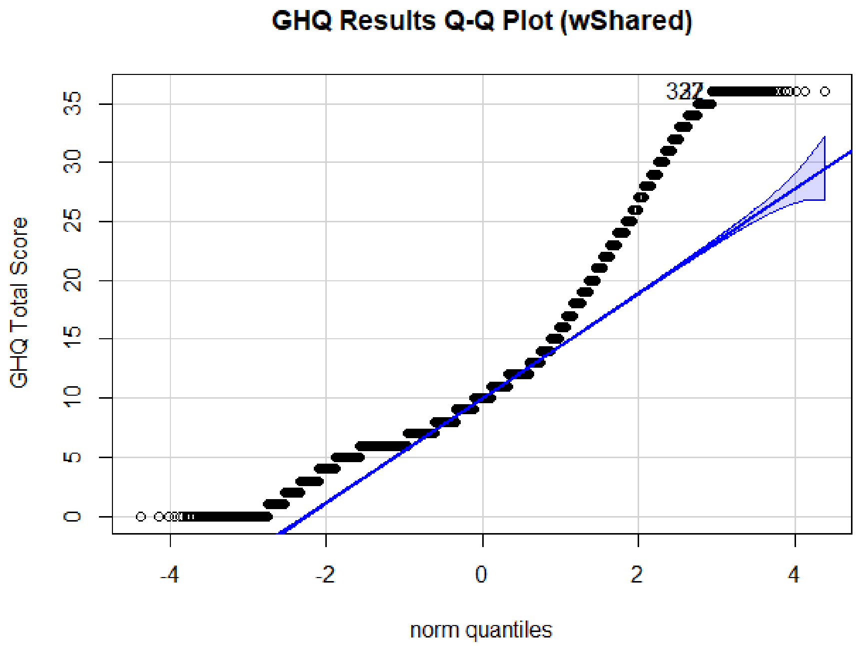


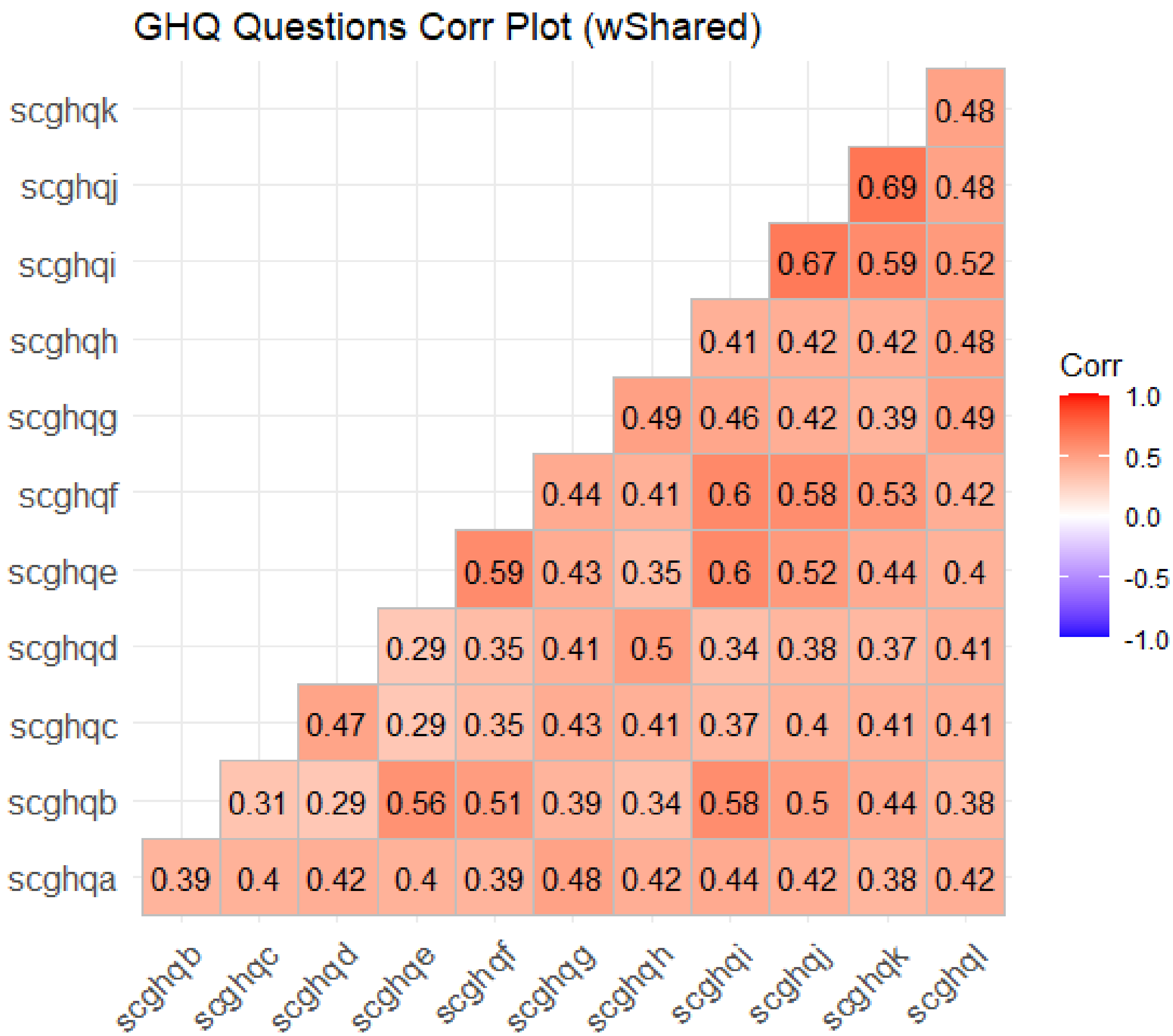



GHQ Questions Corr Plot (w3MergeNurseBlood)
scghqk
scghqj
scghqi
scghqh
scghqg
scghqf
scghqe
scghqd
scghqc
scghqb
scghqa
0.48
0.72 0.45
0.69 0.6 0.52
0.48 0.48 0.49 0.52
0.54 0.48 0.44 0.44 0.51
0.44 0.46 0.63 0.61 0.58 0.42
0.58 0.44 0.41 0.66 0.56 0.47 0.41
0.33 0.38 0.45 0.53 0.39 0.44 0.43 0.47
0.56 0.32 0.38 0.49 0.45 0.42 0.42 0.47 0.44
0.33 0.32 0.59 0.51 0.38 0.37 0.61 0.53 0.48 0.4
0.42 0.4 0.44 0.44 0.39 0.48 0.44 0.47 0.46 0.42 0.41
scghqb
scghqc
scghqd
scghqe
scghqf
scghqg
scghqh
scghqi
scghqj
scghqk
scghql
Corr
1.0
0.5
0.0
-0.5
-1.0

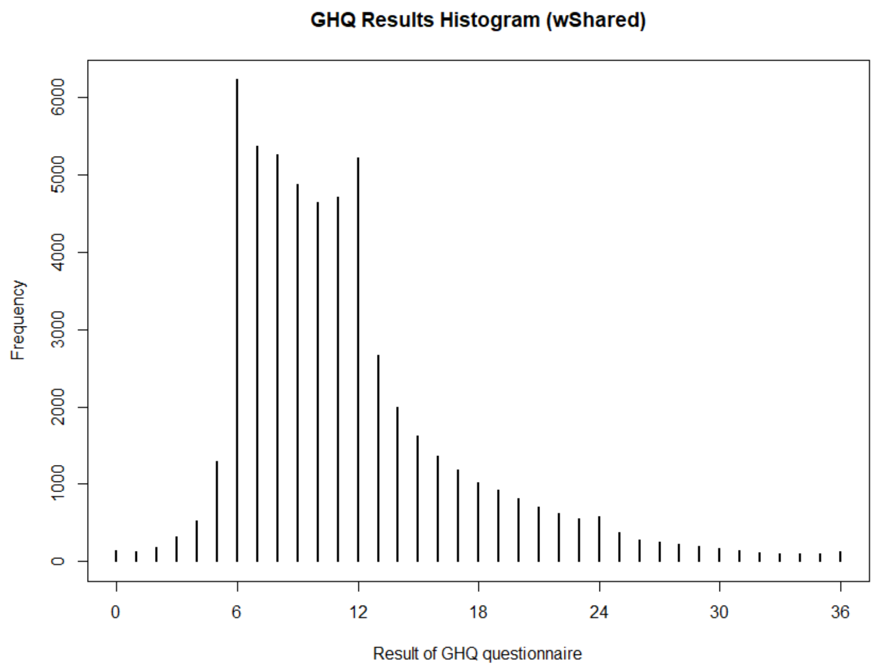
10.2.39 wShared graphs



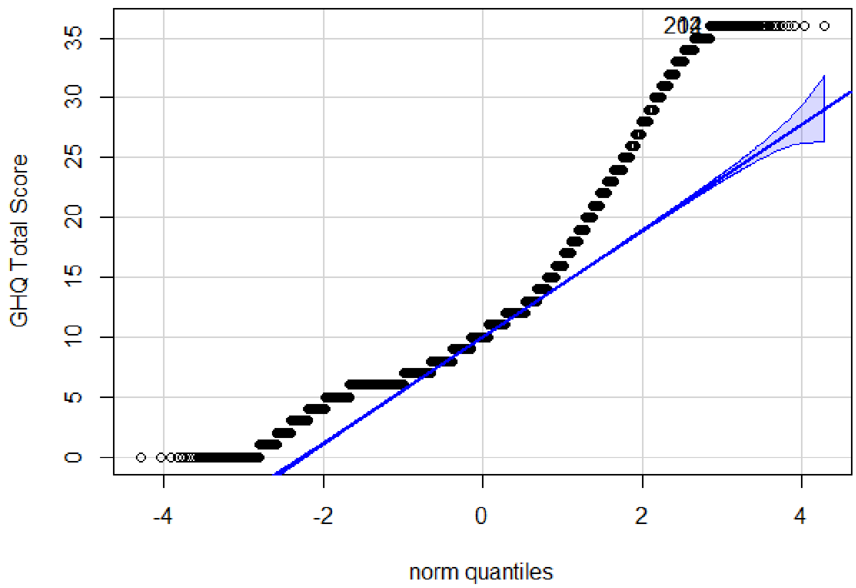

GHQ Results Q-Q Plot (wShared)
GHQ Total Score
norm quantiles
0
5
10
15
20
25
30
35
-4
-2
0
2
4

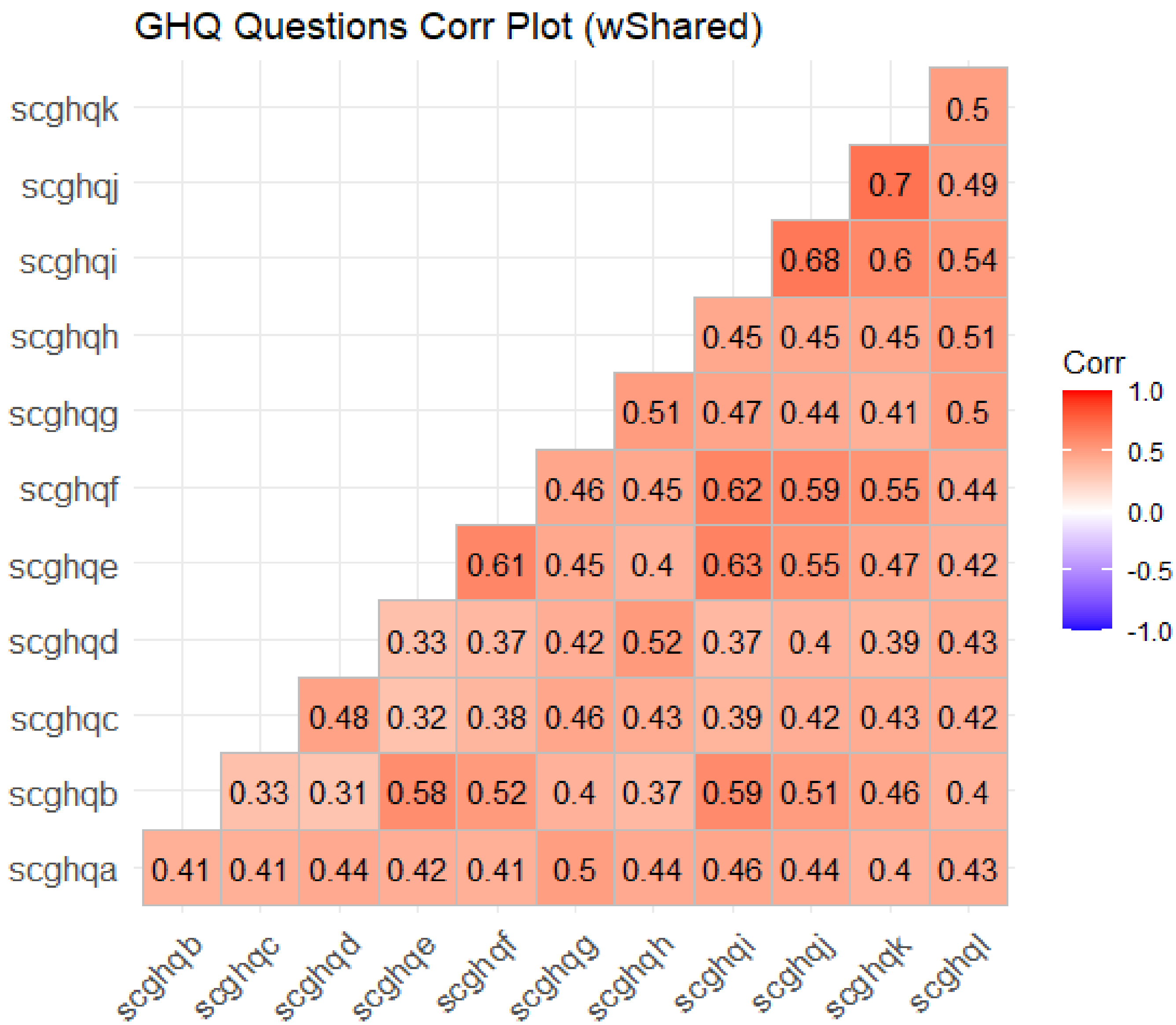

GHQ Questions Corr Plot (wShared)
scghqk
scghqj
scghqi
scghqh
scghqg
scghqf
scghqe
scghqd
scghqc
scghqb
scghqa
scghqb
scghqc
scghqd
scghqe
scghqf
scghqg
scghqh
scghqi
scghqj
scghqk
scghql
Corr
1.0
0.5
0.0
-0.5
-1.0

10.2.40 wSMerge graphs

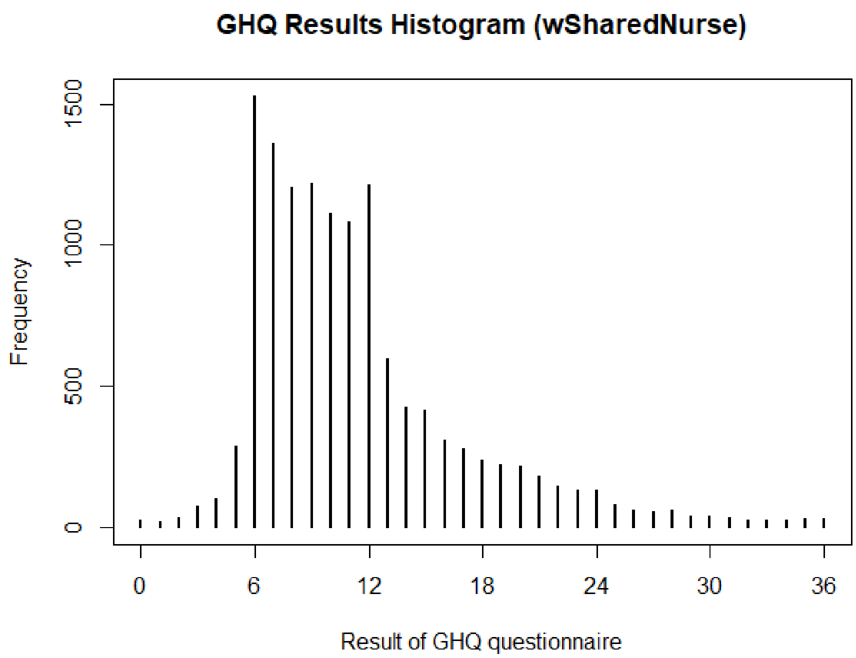


GHQ Results Q-Q Plot (wShared)

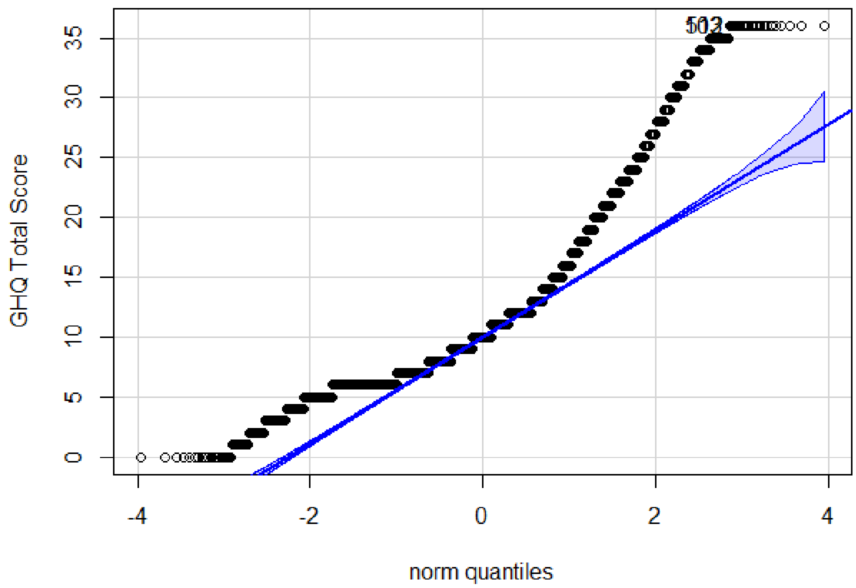


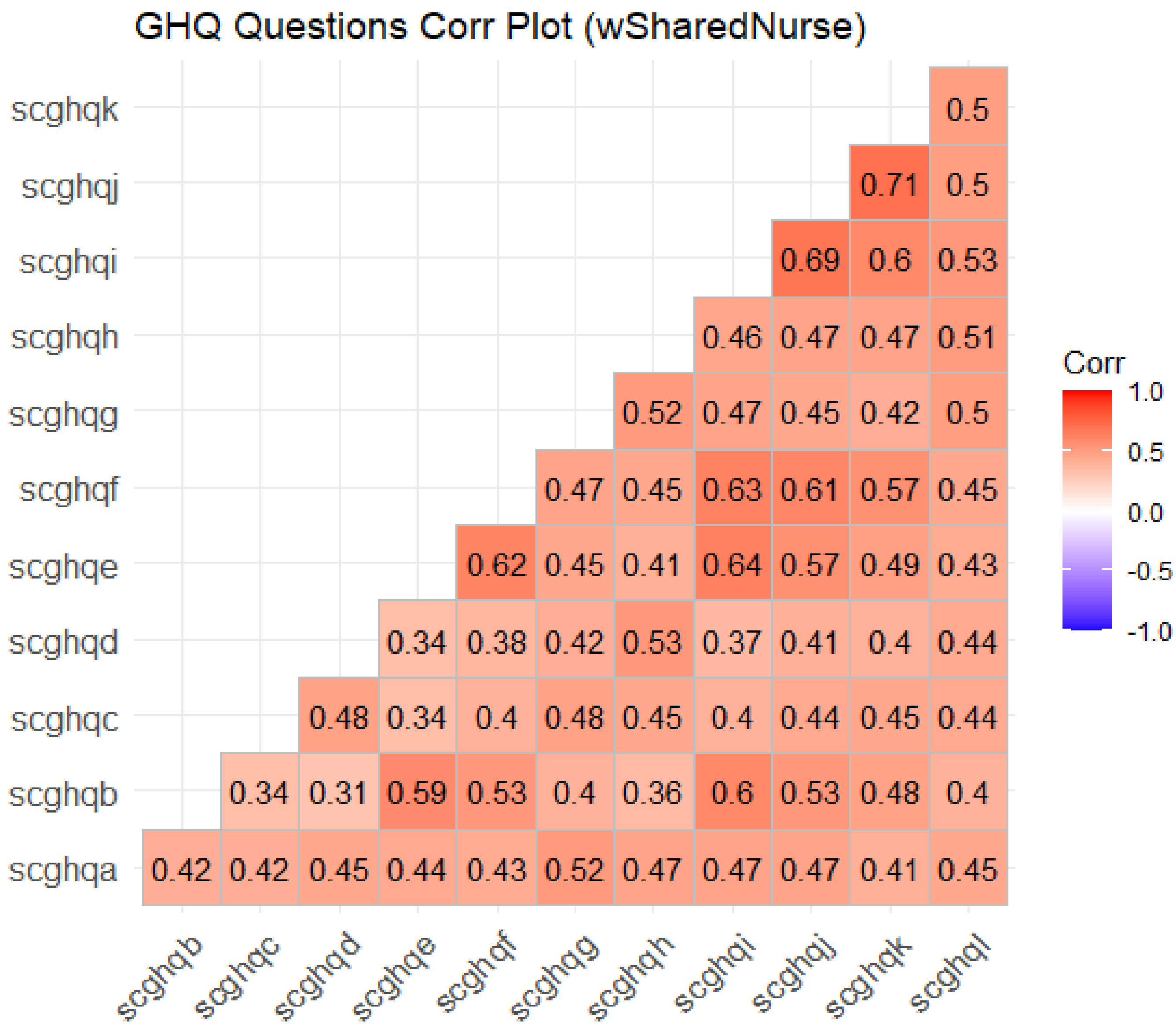

GHQ Questions Corr Plot (wShared)
scghqk
scghqj
scghqi
scghqh
scghqg
scghqf
scghqe
scghqd
scghqc
scghqb
scghqa
0.5
0.7 0.49
0.68 0.6 0.54
0.45 0.45 0.45 0.51
0.51 0.47 0.44 0.41 0.5
0.46 0.45 0.62 0.59 0.55 0.44
0.61 0.45 0.4 0.63 0.55 0.47 0.42
0.33 0.37 0.42 0.52 0.37 0.4 0.39 0.43
0.48 0.32 0.38 0.46 0.43 0.39 0.42 0.43 0.42
0.33 0.31 0.58 0.52 0.4 0.37 0.59 0.51 0.46 0.4
0.41 0.41 0.44 0.42 0.41 0.5 0.44 0.46 0.44 0.4 0.43
scghqb
scghqc
scghqd
scghqe
scghqf
scghqg
scghqh
scghqi
scghqj
scghqk
scghql
Corr
1.0
0.5
0.0
-0.5
-1.0

10.2.41 wSMergeNurse graphs



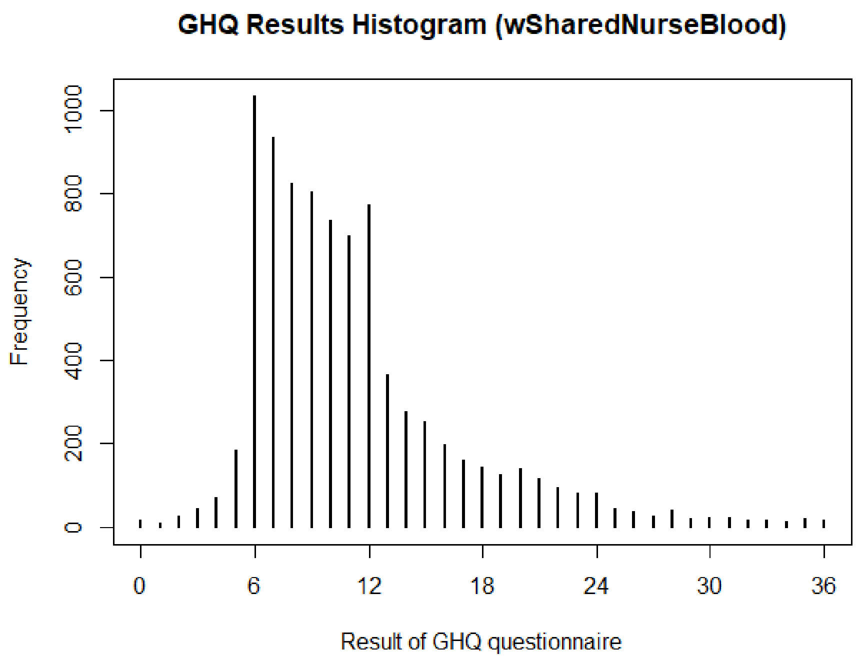
GHQ Results Q-Q Plot (wSharedNurse)

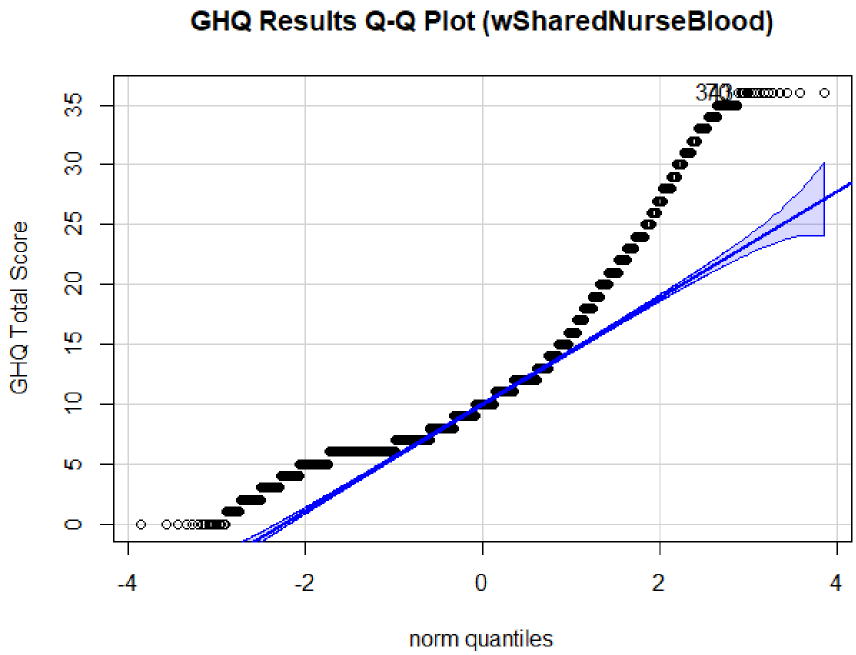


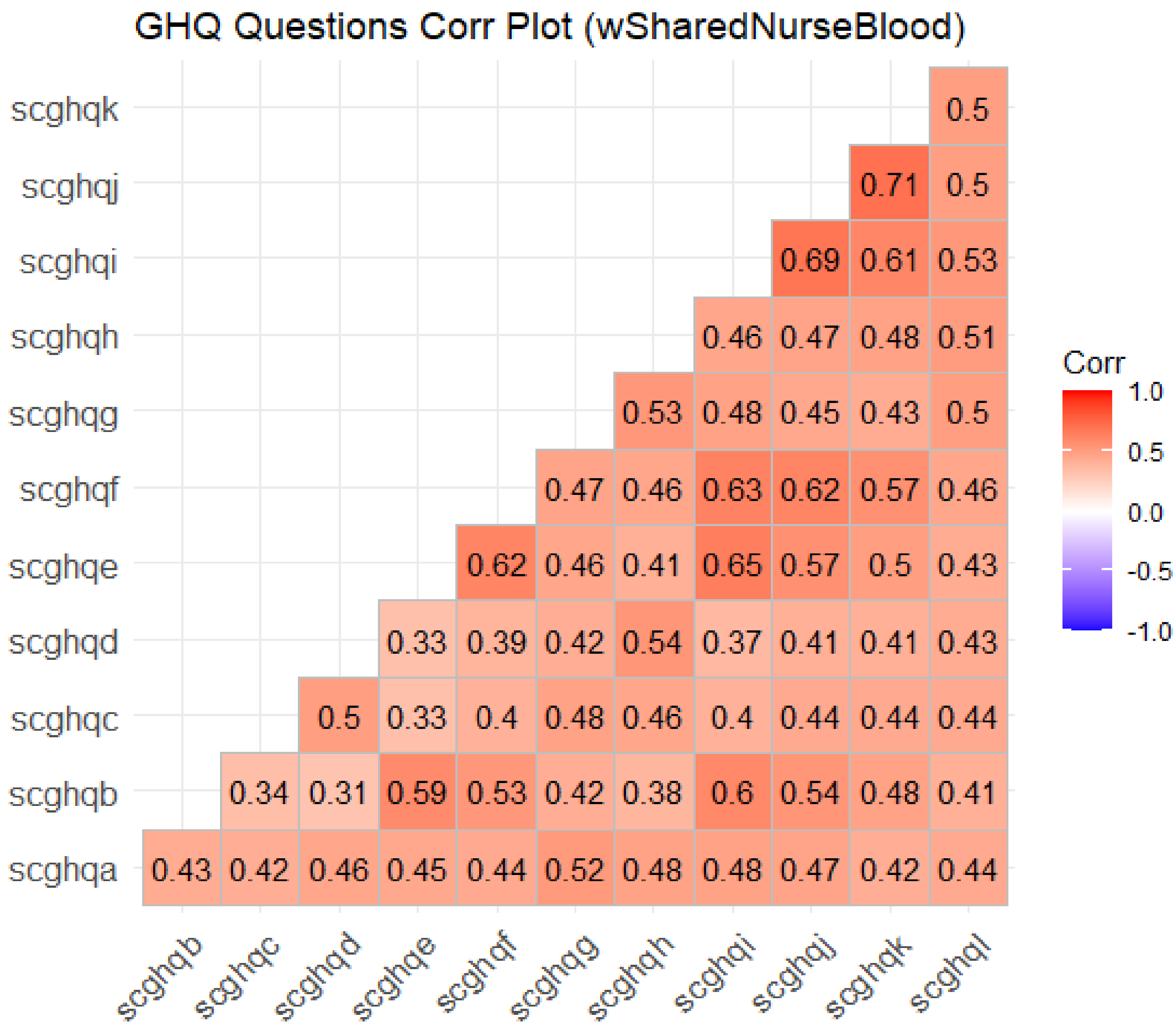


GHQ Questions Corr Plot (wSharedNurse)
scghqk
scghqj
scghqi
scghqh
scghqg
scghqf
scghqe
scghqd
scghqc
scghqb
scghqa
scghqb
scghqc
scghqd
scghqe
scghqf
scghqg
scghqh
scghqi
scghqj
scghqk
scghql
Corr
1.0
0.5
0.0
-0.5
-1.0
0.5
0.71 0.5
0.69 0.6 0.53
0.46 0.47 0.47 0.51
0.52 0.47 0.45 0.42 0.5
0.47 0.45 0.63 0.61 0.57 0.45
0.62 0.45 0.41 0.64 0.57 0.49 0.43
0.34 0.38 0.42 0.53 0.37 0.41 0.4 0.44
0.48 0.34 0.4 0.48 0.45 0.4 0.44 0.45 0.44
0.34 0.31 0.59 0.53 0.4 0.36 0.6 0.53 0.48 0.4
0.42 0.42 0.45 0.44 0.43 0.52 0.47 0.47 0.47 0.41 0.45

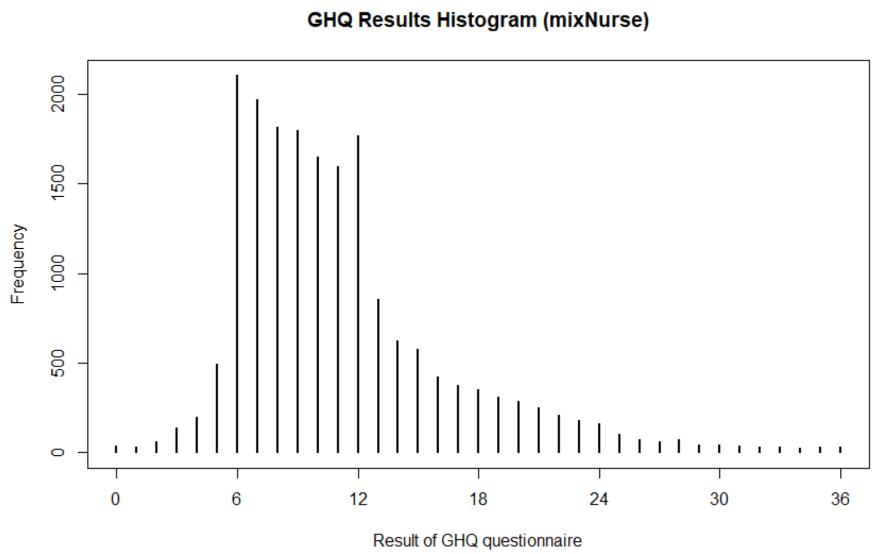
10.2.42 wSMergeNurseBlood graphs

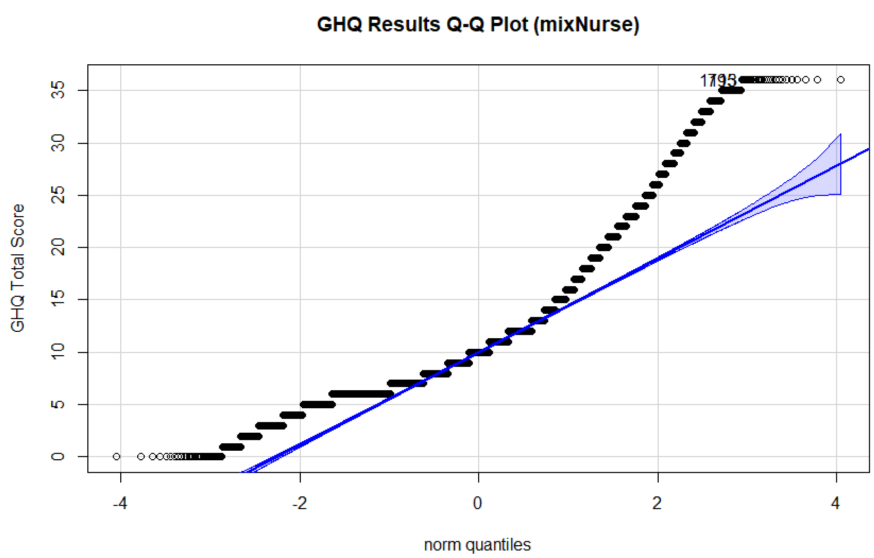


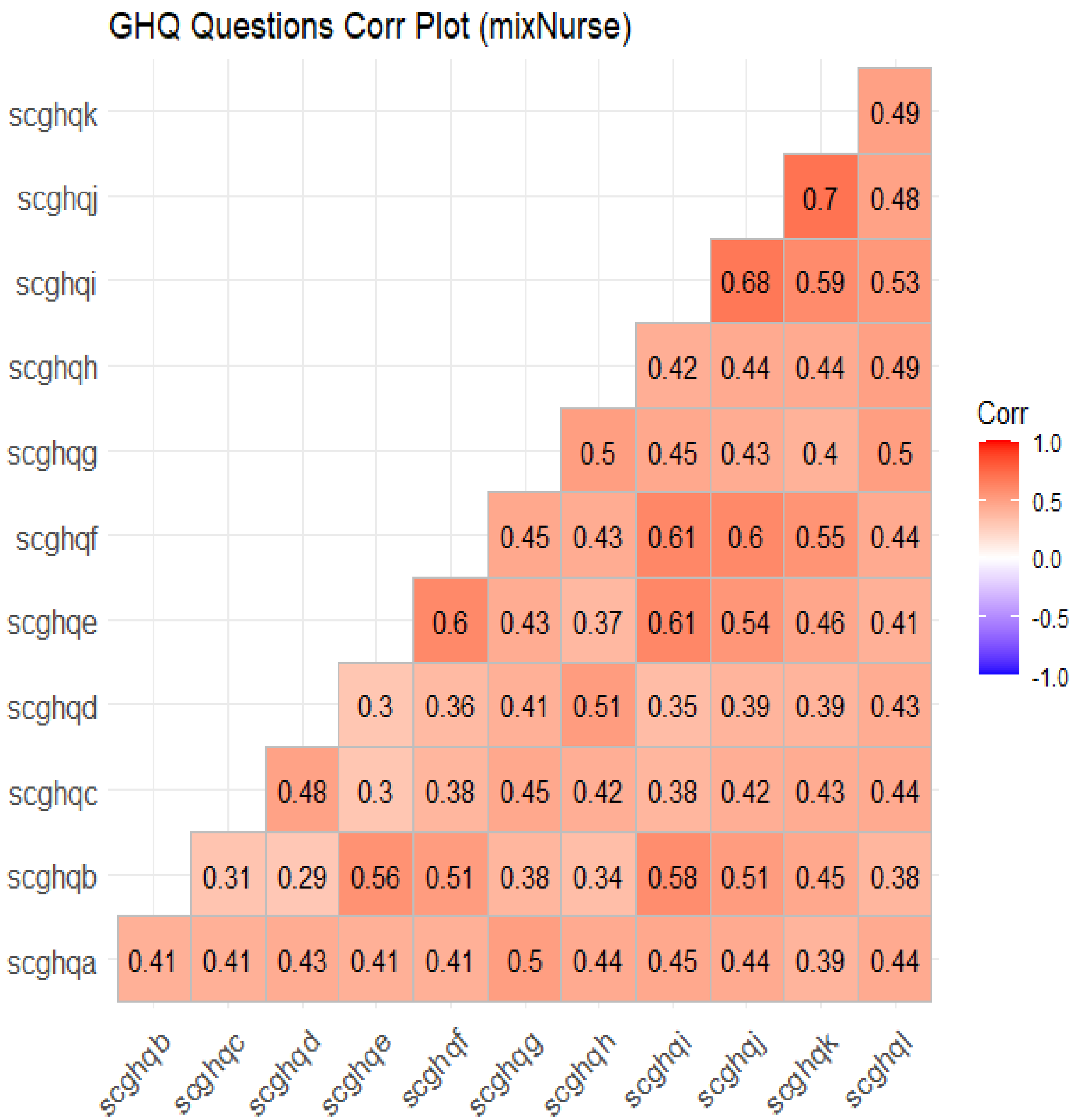

GHQ Results Q-Q Plot (wSharedNurseBlood)
GHQ Total Score
0
5
10
15
20
25
30
35
-4
-2
0
2
4
norm quantiles
370


GHQ Questions Corr Plot (wSharedNurseBlood)
scghqk
scghqj
scghqi
scghqh
scghqg
scghqf
scghqe
scghqd
scghqc
scghqb
scghqa
0.5
0.71 0.5
0.69 0.61 0.53
0.46 0.47 0.48 0.51
0.53 0.48 0.45 0.43 0.5
0.47 0.46 0.63 0.62 0.57 0.46
0.62 0.46 0.41 0.65 0.57 0.5 0.43
0.33 0.39 0.42 0.54 0.37 0.41 0.41 0.43
0.5 0.33 0.4 0.48 0.46 0.4 0.44 0.44 0.44
0.34 0.31 0.59 0.53 0.42 0.38 0.6 0.54 0.48 0.41
0.43 0.42 0.46 0.45 0.44 0.52 0.48 0.48 0.47 0.42 0.44
scghqb
scghqc
scghqd
scghqe
scghqf
scghqg
scghqh
scghqi
scghqj
scghqk
scghql
Corr
1.0
0.5
0.0
-0.5
-1.0

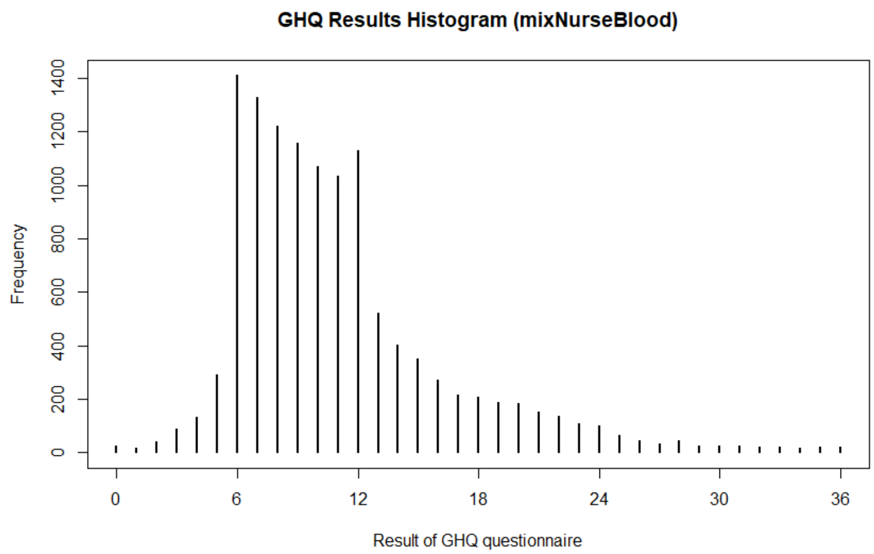
10.2.43 mixNurse graphs

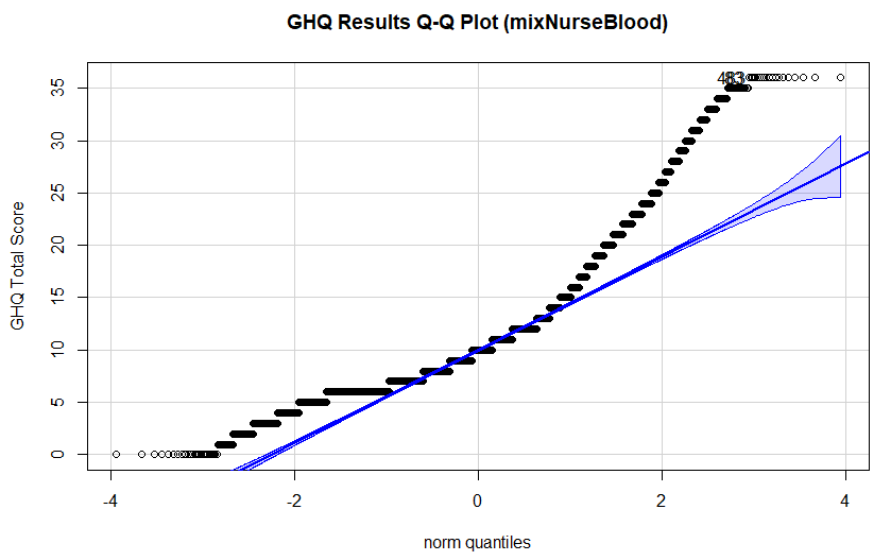


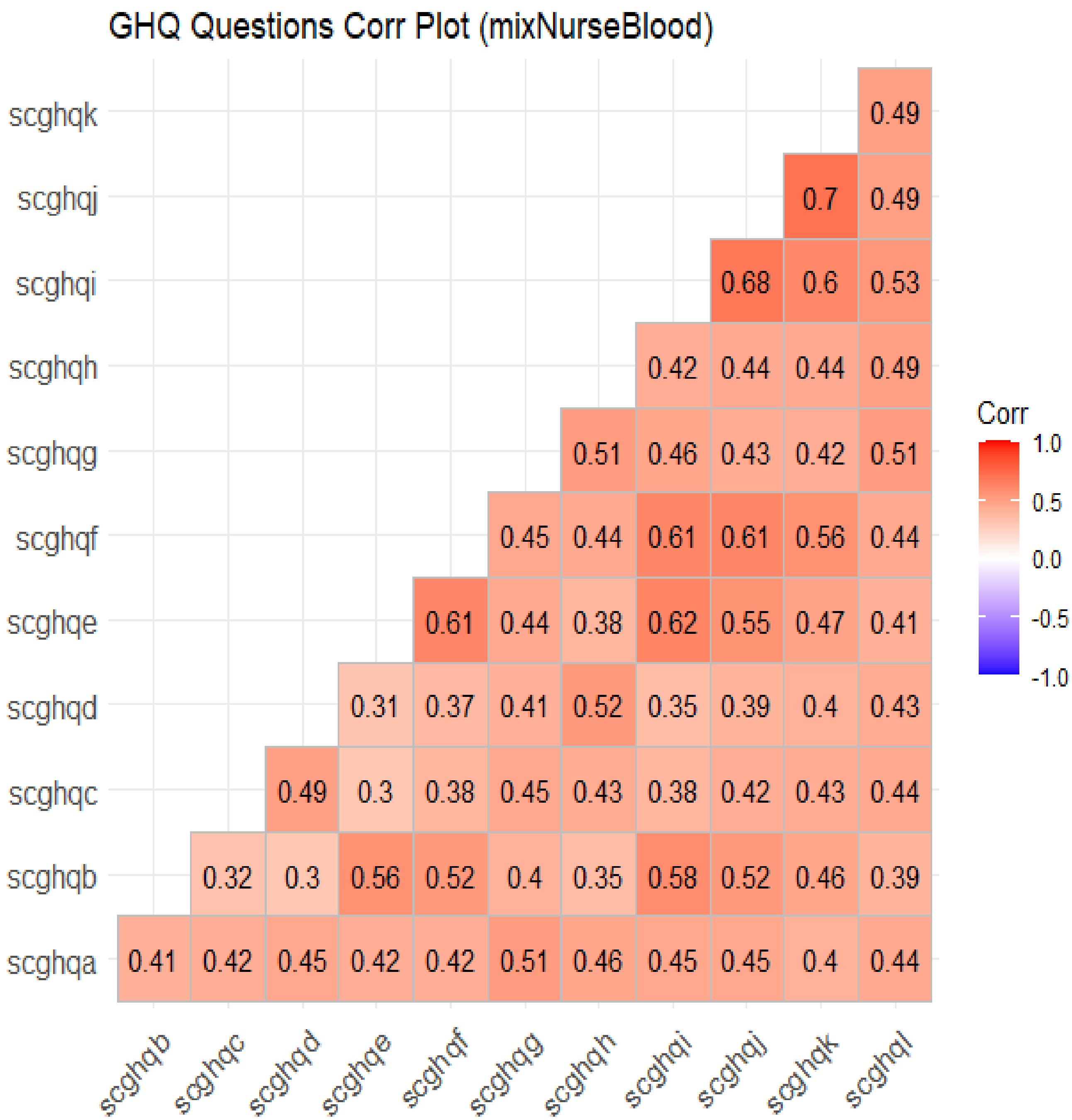



GHQ Questions Corr Plot (mixNurse)
scghqk
scghqj
scghqi
scghqh
scghqg
scghqf
scghqe
scghqd
scghqc
scghqb
scghqa
0.49
0.7 0.48
0.68 0.59 0.53
0.42 0.44 0.44 0.49
0.5 0.45 0.43 0.4 0.5
0.45 0.43 0.61 0.6 0.55 0.44
0.6 0.43 0.37 0.61 0.54 0.46 0.41
0.3 0.36 0.41 0.51 0.35 0.39 0.39 0.43
0.48 0.3 0.38 0.45 0.42 0.38 0.42 0.43 0.44
0.31 0.29 0.56 0.51 0.38 0.34 0.58 0.51 0.45 0.38
0.41 0.41 0.43 0.41 0.41 0.5 0.44 0.45 0.44 0.39 0.44
scghqb
scghqc
scghqd
scghqe
scghqf
scghqg
scghqh
scghqi
scghqj
scghqk
scghql
Corr
1.0
0.5
0.0
-0.5
-1.0

10.2.44 mixNurseBlood graphs






GHQ Questions Corr Plot (mixNurseBlood)
scghqk
scghqj
scghqi
scghqh
scghqg
scghqf
scghqe
scghqd
scghqc
scghqb
scghqa
0.49
0.7 0.49
0.68 0.6 0.53
0.42 0.44 0.44 0.49
0.51 0.46 0.43 0.42 0.51
0.45 0.44 0.61 0.61 0.56 0.44
0.61 0.44 0.38 0.62 0.55 0.47 0.41
0.31 0.37 0.41 0.52 0.35 0.39 0.4 0.43
0.49 0.3 0.38 0.45 0.43 0.38 0.42 0.43 0.44
0.32 0.3 0.56 0.52 0.4 0.35 0.58 0.52 0.46 0.39
0.41 0.42 0.45 0.42 0.42 0.51 0.46 0.45 0.45 0.4 0.44
scghqb
scghqc
scghqd
scghqe
scghqf
scghqg
scghqh
scghqi
scghqj
scghqk
scghql
Corr
1.0
0.5
0.0
-0.5
-1.0