Code For Project 1 - Applied Statistics

Created by: Nachi Lieder 314399114

Introduction

This is a report which its main purpose is to examine the variables affecting the BMI of the respondents. We will go through several stages in this report

- Data formatting - appending the data into one set that is manageable.
- Feature Exploration
- Dealing With Missing Data
- Response Variable Exploration
- Univariate regressions - initial screening
- Multivariate regressions - in depth screening & attempt to describe the BMI target using the given parameters.

We will explore the data , attempt to understand the behavior of different predictors, and try to use them to explain the behavior of the BMI target.

Part 1

In this part I recieve as an input the 4 datsets , merge them , and combine 1 whole dataset. You can see in the result the dimensions of the resulted dataframe which we will explore. I merged each genders two datasets together (the demographic data and survey data) and appended it together. The final result here is a single dataframe with the dimension of 4036 x 57.

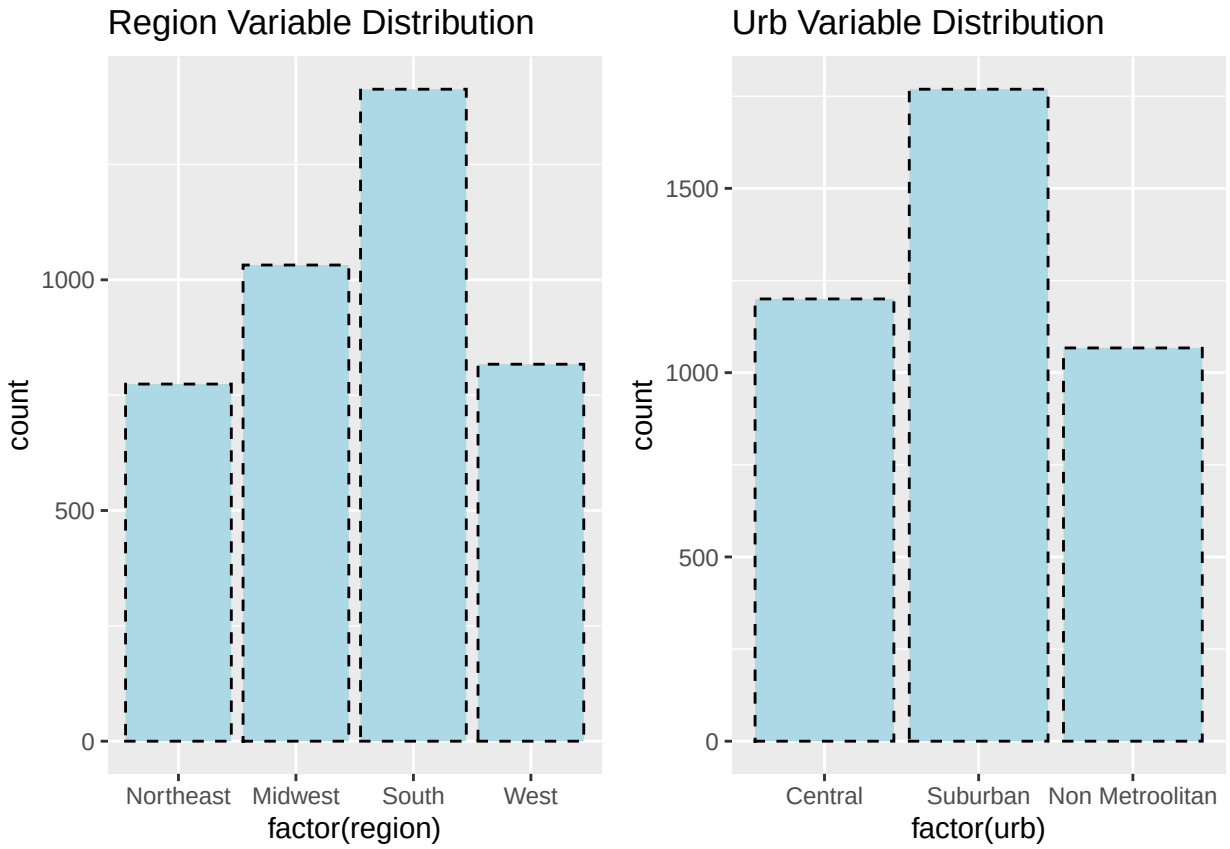
Part 2

In this part I will perform an initial exploration on the datasets features (predictors) , and attempt to understand the charecteristics of the predictors. Lets start off with looking at the key features:

1. Region
2. Urb
3. Income
4. Age
5. Gender
6. Grade
7. Exercise

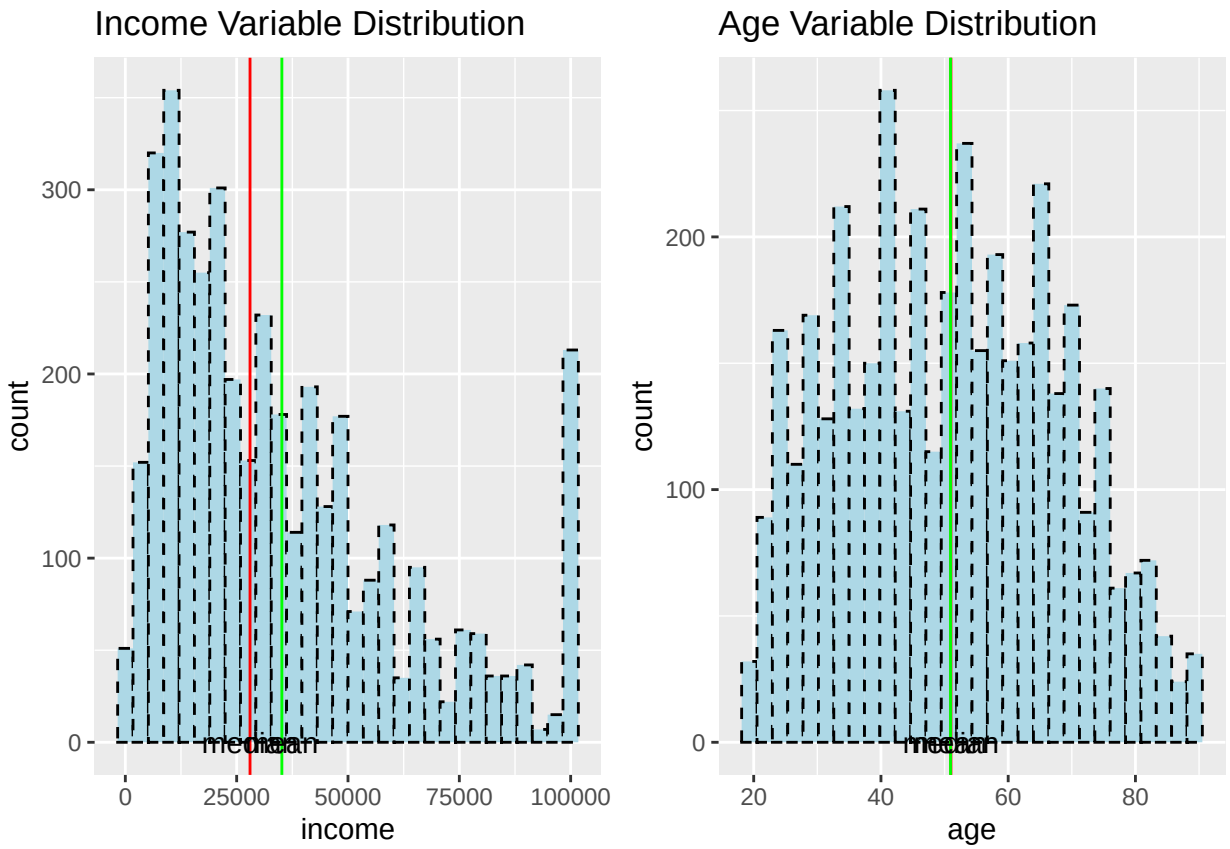
Lets observe the Region feature: We can see that overall the there is a slight advantage to observations of patients coming from the South.

Next, lets observe the Urb feature: Here what we are observing is a slight advantage in the count for the Suburban originated patients.



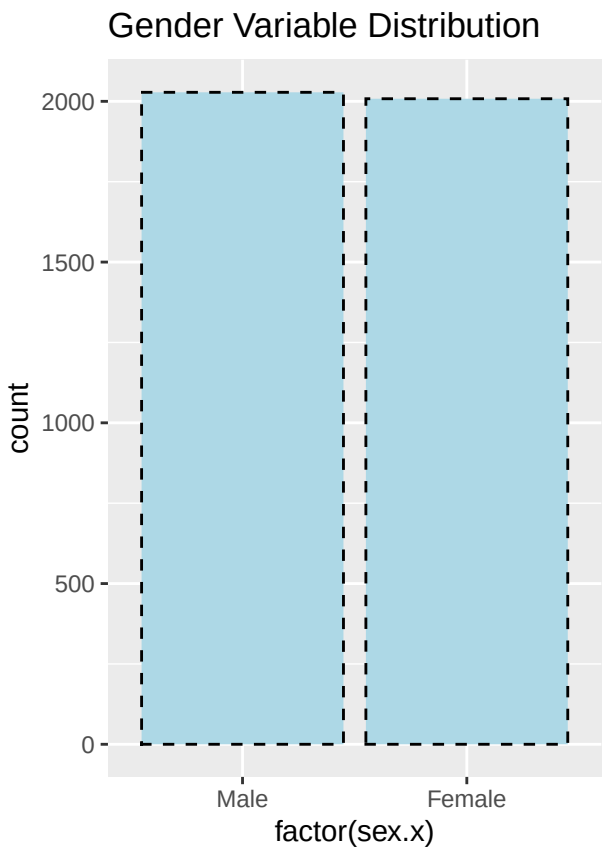
Next , lets observe the income variable. (red line - Median , Green line - mean) We can see that here there is a long right tail and what seems to be abnormal value around 100000. Due to the scaling this right tail spike can argueably make sense , therefor I have decided to leave it in the anlysis and not treat it as an outlier.

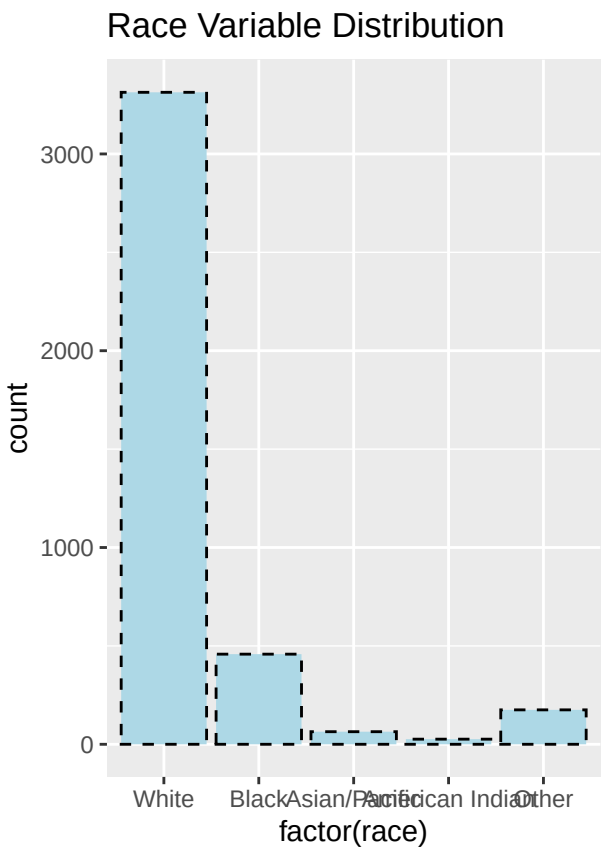
Lets review the next parameter - Age: Here the distribution surrounds the median and mean which is around 50.



Lets observe the Gender distribution, where we can see an even distribution.

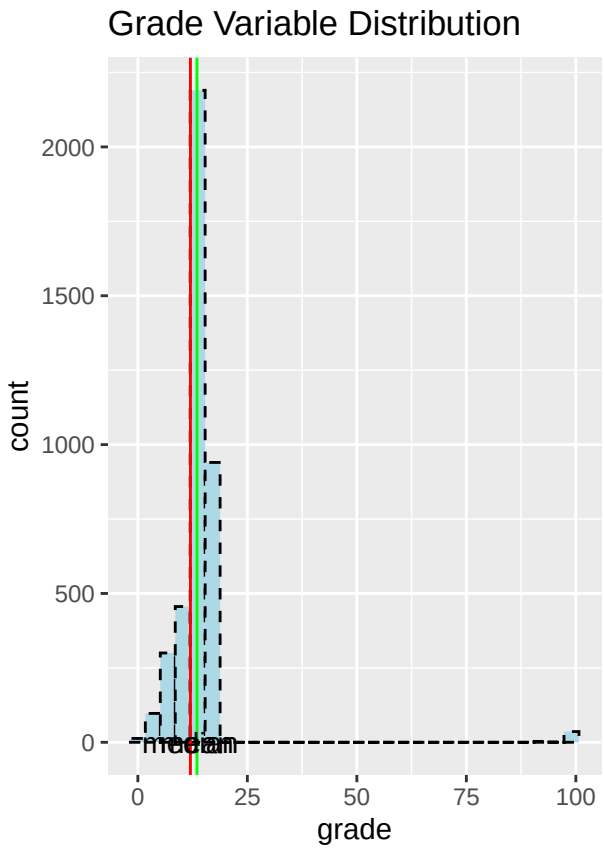
Lets look at the race feature: Here we can see an unbalanced set where there is an oversample of patients categorized under White. We will explore later on whether this feature can assist us in describing our target (BMI).

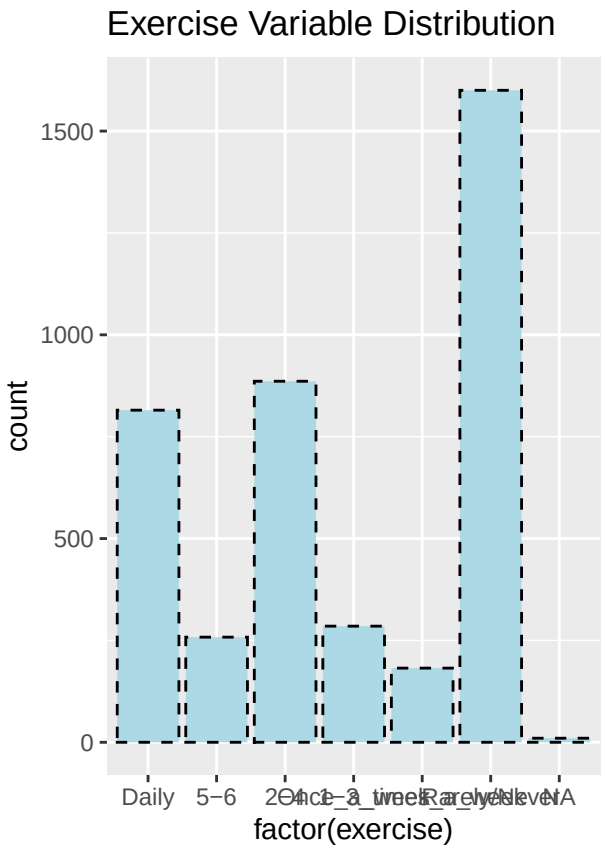




Lets view the Grade feature: Here we are able to catch the outliers where their grade is above 20. We will treat these outliers in the following parts. In addition we can see that as expected , the median and mean are near 12. which intuitively makes sense since this represents people who finished high school.

Lets explore the "Exercise" variable: here too we can see some NaN obsevations that will be treated in the following parts of the report. In terms of equally distributed data , the data here is not balanced.

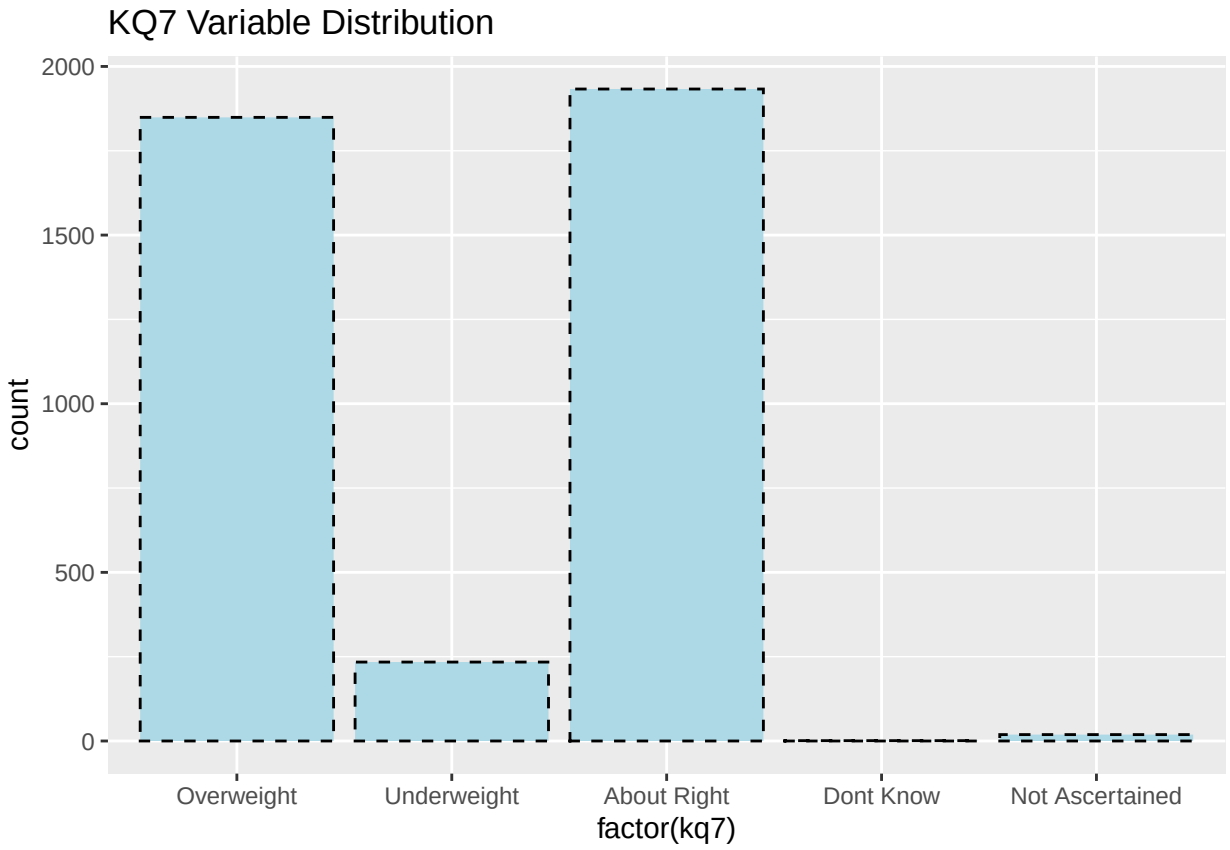




Lets view the kq7 feature: Here we can see some null or insignificant values which will be treated later on.

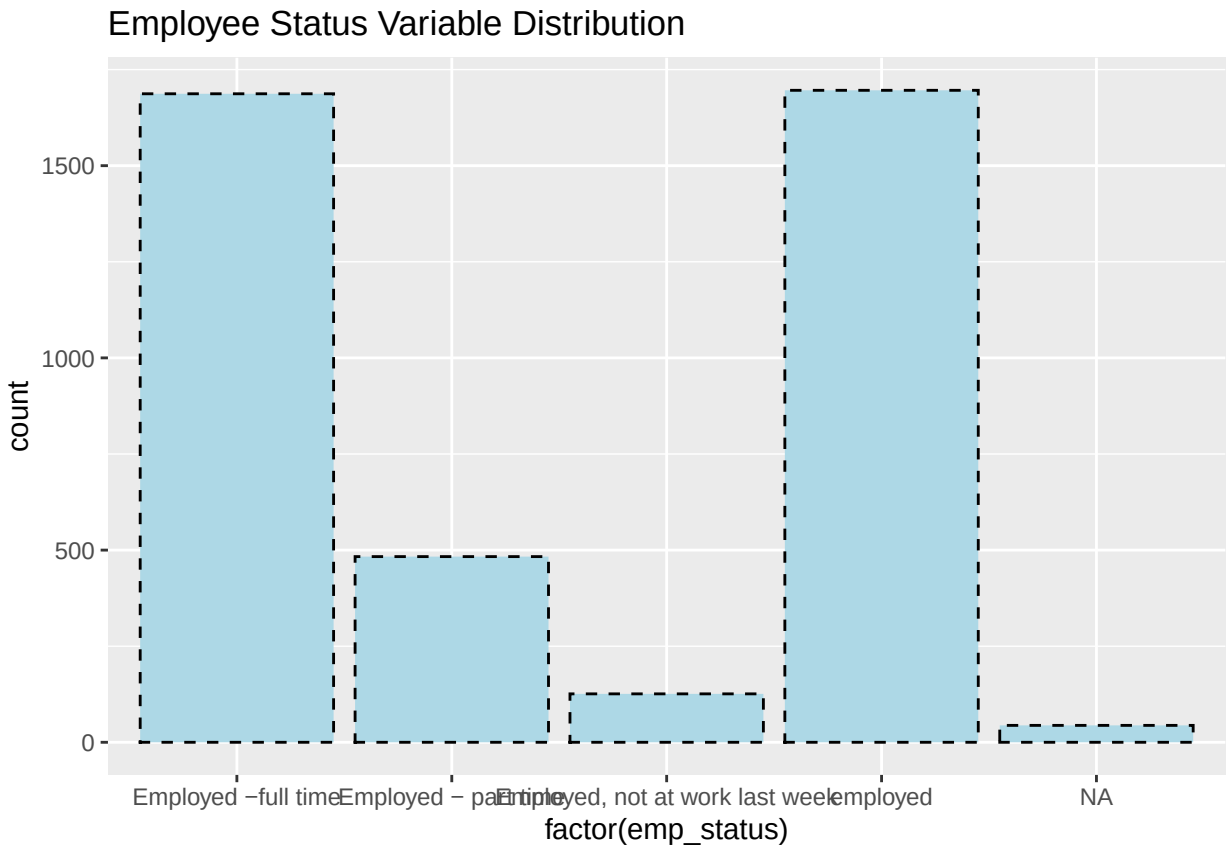
Table 1: Frequency Matrix - Doctor Questions

	dt01	dt02	dt03	dt06	dt07
	1: 245	1: 361	1: 197	1: 70	1: 138
	2:3791	2:3675	2:3839	2:3966	2:3898



Lets view the binary questions known as dt01,02,03,06,07 In the following diagram we have the frequency matrix per question x category answer. These too are not balanced and mostly are answered with a one-sided answer, we will try to take this under consideration in the following steps.

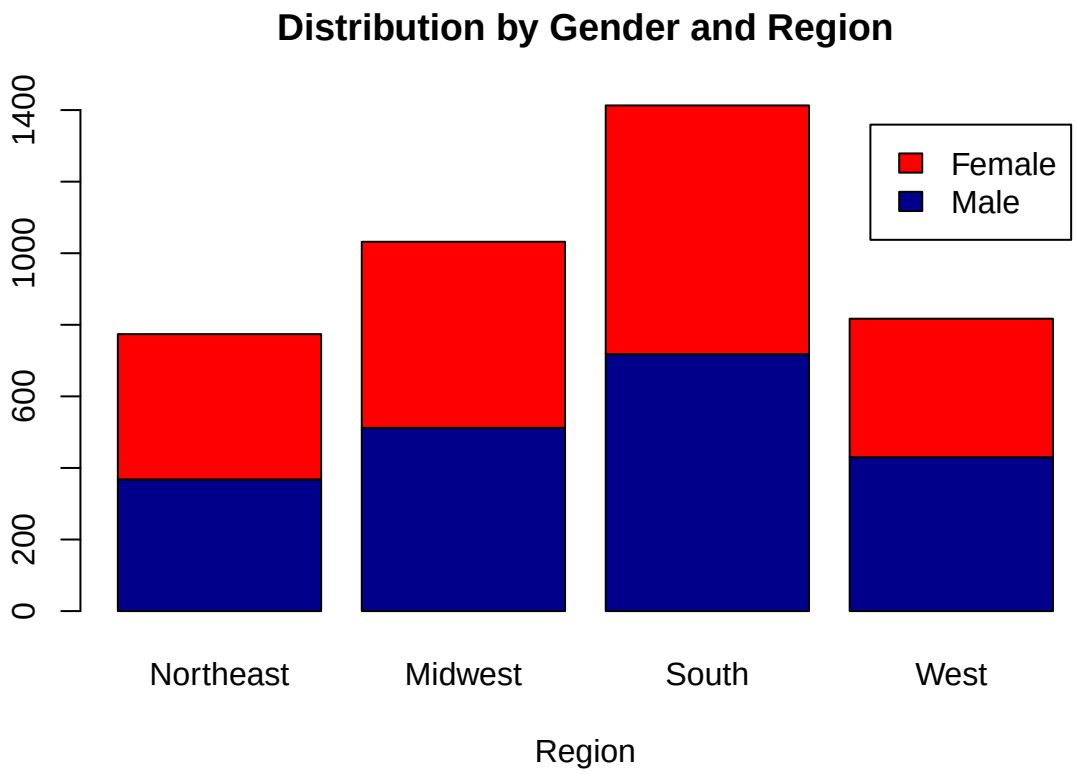
Lets look at the Employee Status feature. There are some values to be treated , and it seems that the majority comes from two classes from of the four.

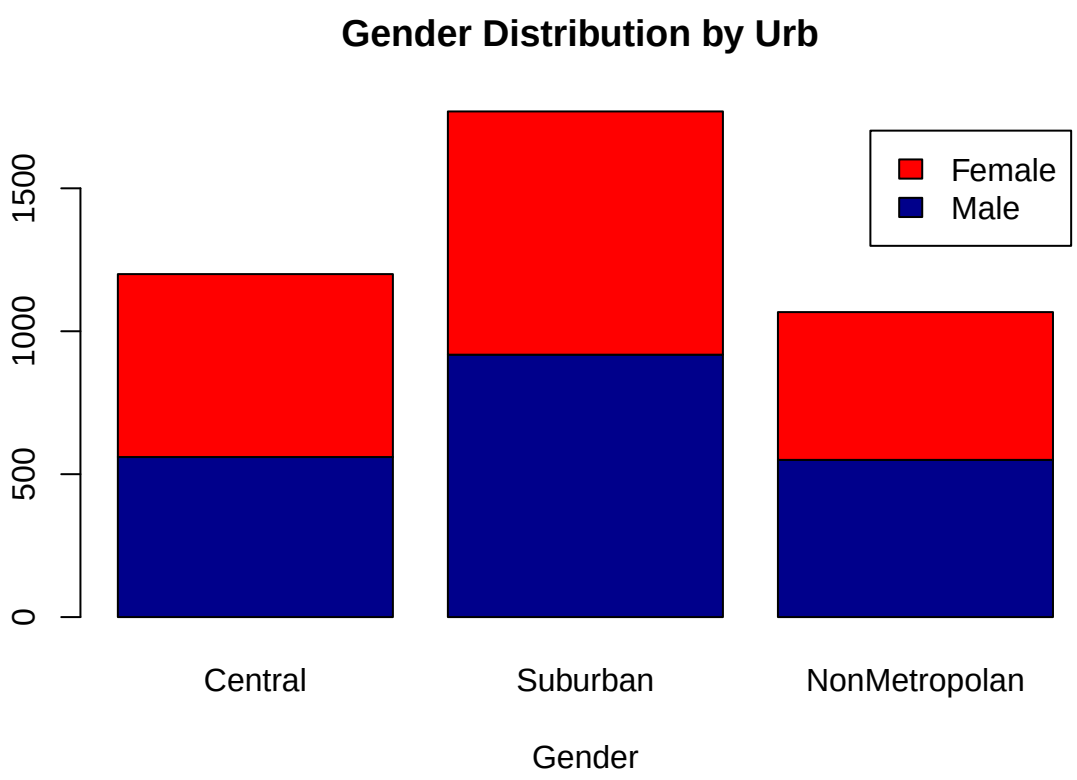


Lets explore some additional characterisitcs and relationships that may trigger some additional analysis.

In this following plot we describe the distribution of gender & Region It seems that the porportions of the gender are equal across all regions .

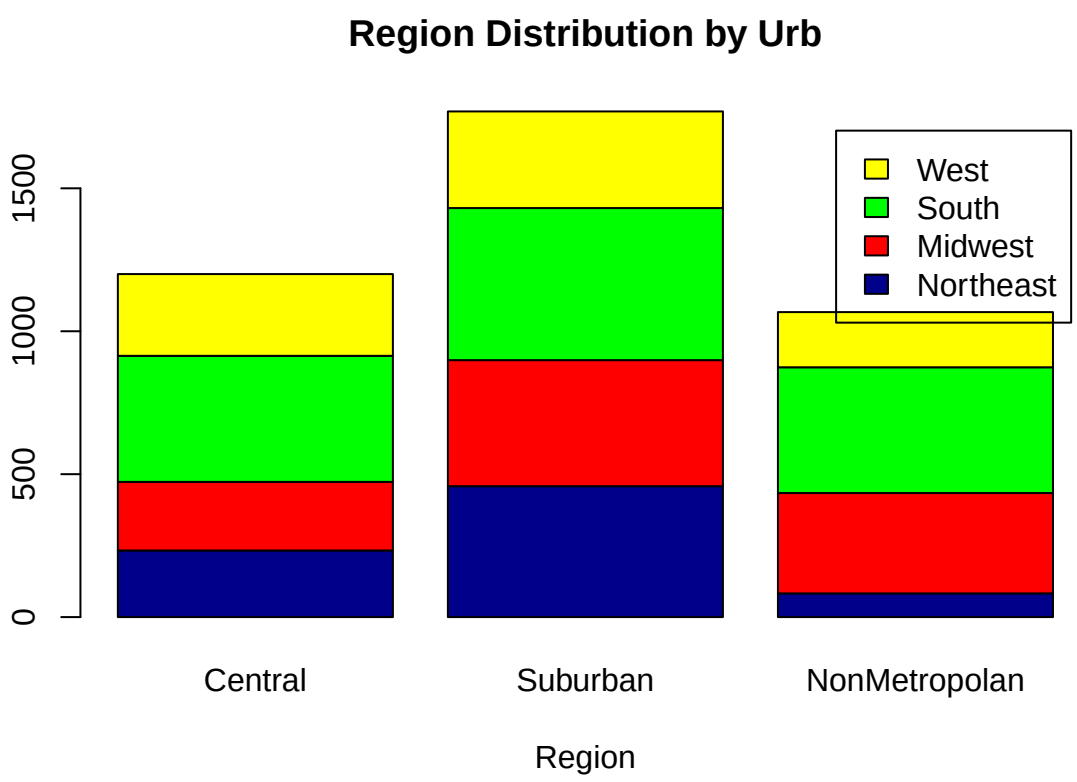
Next lets view the split between the gender and the Urb feature. Here too there seems to be an equal split.

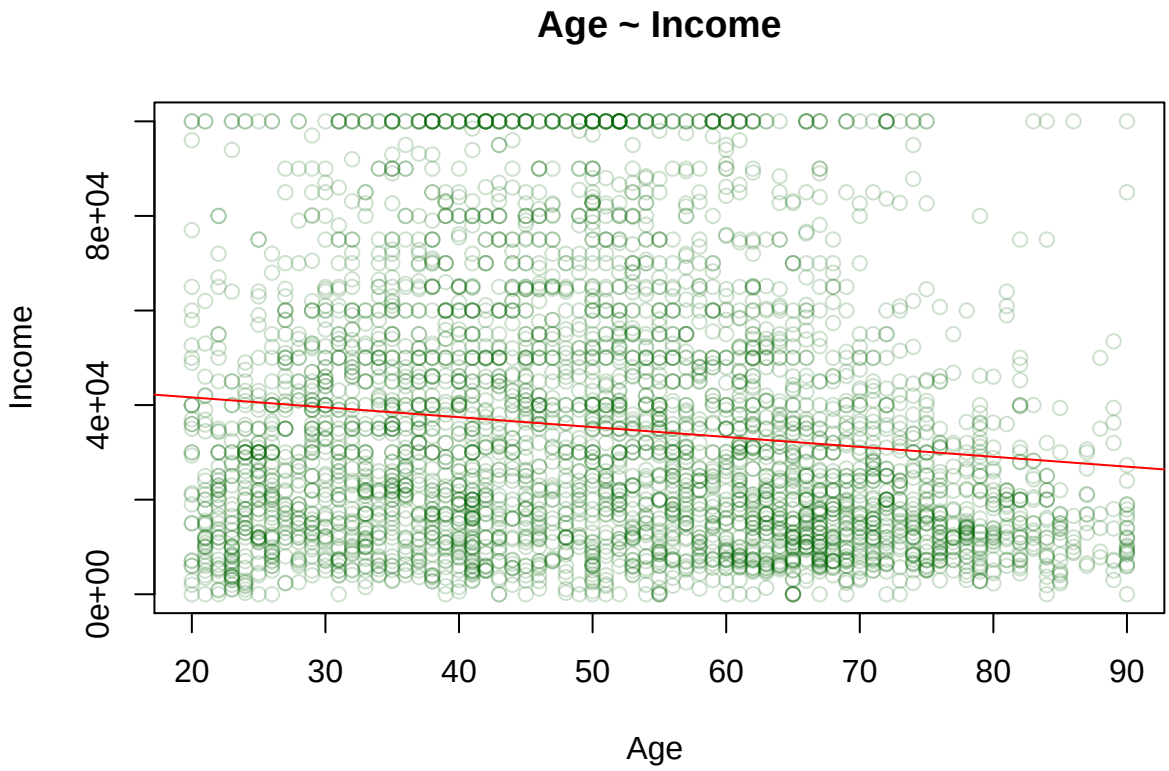

Distribution by Gender and Region
0
200
600
1000
1400
Female
Male
Northeast
Midwest
South
West
Region



Now lets observe the split between Urb and Region. This is to understand the different distributions between inner classes. In this example we can see that the Northeast under the Non Metropolitan has a lower frequency , than the other two Urb sections, and is porportionally less.

In this following plot we observe the relationship and linear trend between the Age and the Income. The scatter is pretty well spread , though there is a slight negative trend/correlation overall between the two. This might be due to the fact that the elderly dont work and possibly have a lower income.(Thoughts to consider)


Region Distribution by Urb
0
500
1000
1500
West
South
Midwest
Northeast
Central
Suburban
NonMetropolan
Region



Part 3 - Missing and Problematic Data

Given the insights from the prior section , we decide now what data to be changed to NA and follow by removing for regressional tests etc. In this part we deal with NaNs for exercise.

Deal with NaNs and problematic answers that are uninformative for the kq7 variable.

We deal with grades that are above 18 which will be considered outliers.

We deal with Null values for the 5 questions from the doctors.

We deal with Null and missing values for the employee status.

In total we remove about 100 observations from the initial dataset, given that this is a very small amount , we can afford to give up these observations and continue the analysis without them.

We can see that the dimension after the data cleaning is slightly reduced to 3935 x 57

Part 4 - The Response Variable - Need For Transformation?

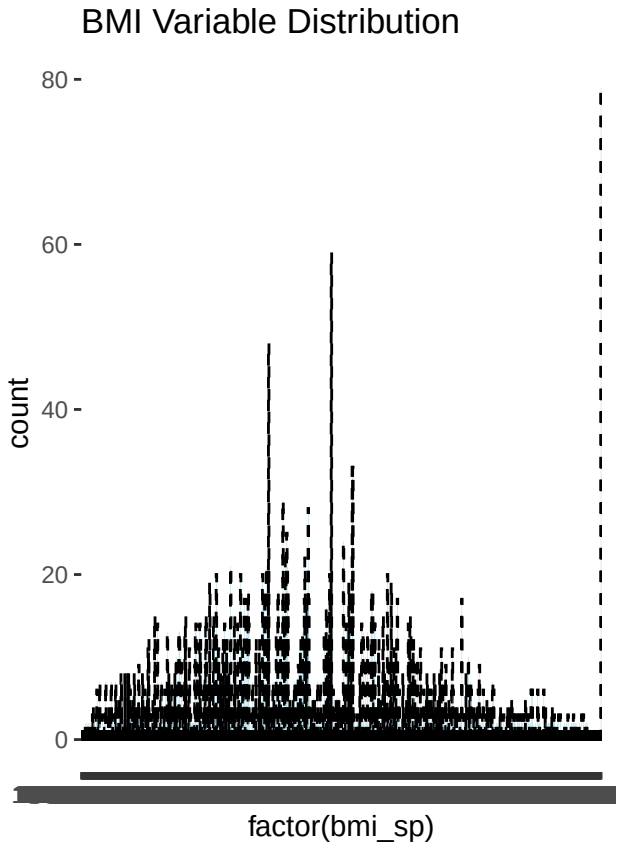
Lets look at the target variable. In this step , we will observe the BMI Measurement and evaluate the metrics , understand whether we will need to transform the value for further perdictive analysis and deal with outlier data.

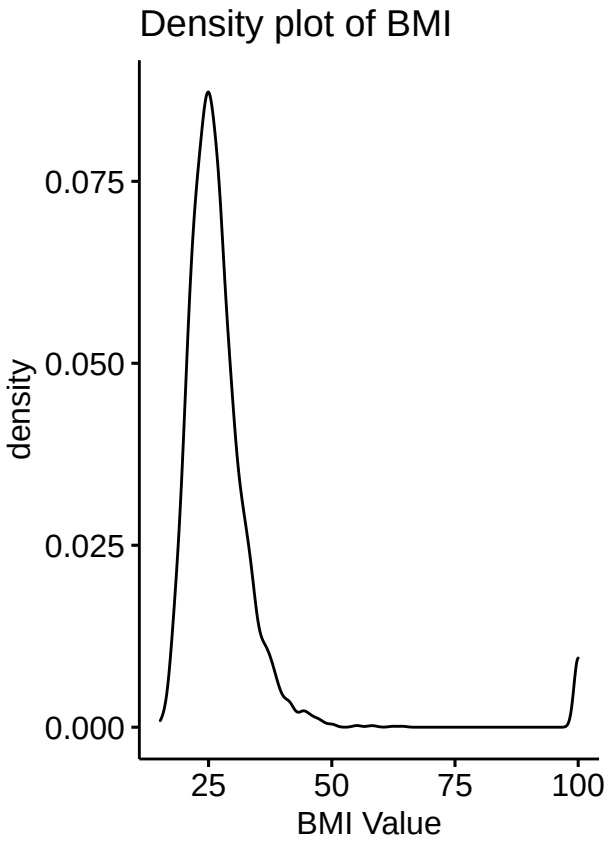
First off we know that values of BMI *tend* to range between 0-30 , where over 30 is considered obese. We can see here that there are certain outliers , where an entire quartile is above 30 (and the top 2 percentile are above 75). We will consider the top two percentile as outliers and deal with them in the following sections.

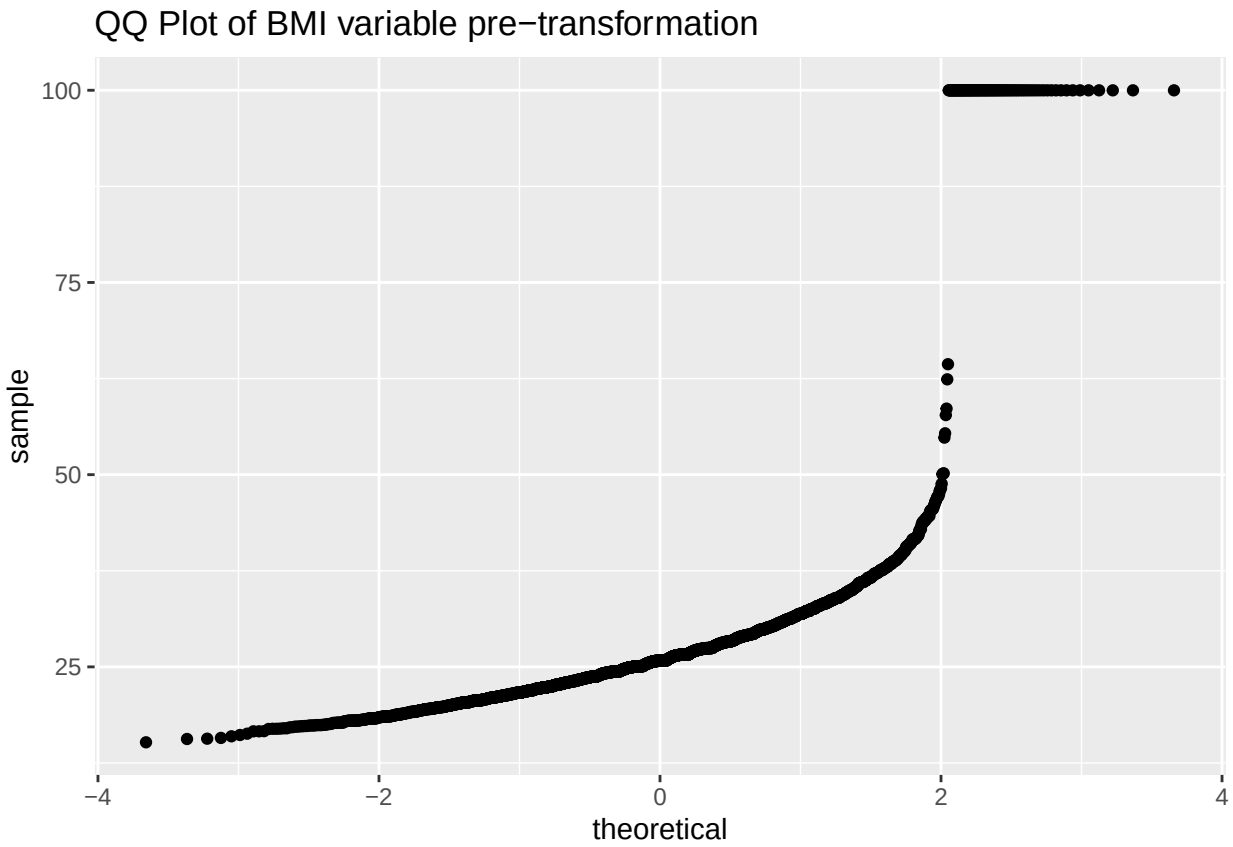
Table 2: Quantile Matrix - BMI results

	x
0%	15.1900
25%	22.9200
50%	25.8200
75%	29.4150
90%	34.1340
98%	75.7684
100%	99.9900

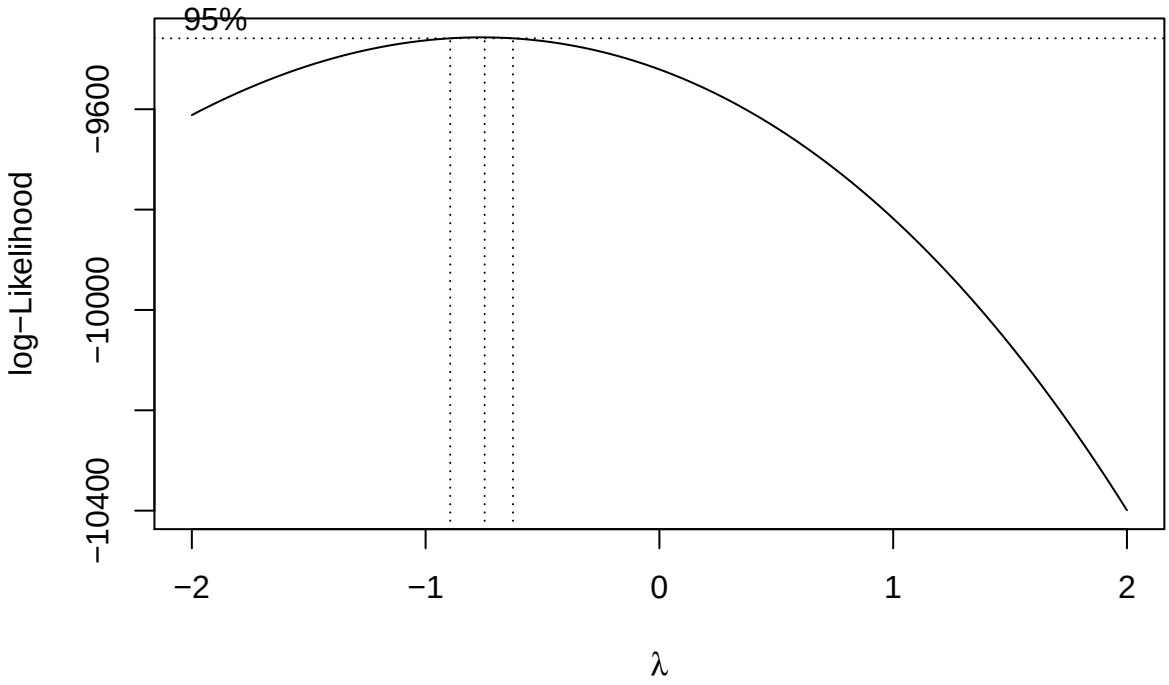
In addition ,from the following figure, we are dealing with data that has a gaussian-like curve , though not specifically a normal distribution. We would like to verify that it is actually normal as a pre condition for future analysis such as log likelihood etc. For this we performed the following QQ plot where we can see the huge right tail of our outliers affecting this analysis. For the sake of the checkup , I removed these outliers and re-checked the QQ plot finding a somewhat smoother plot , though not a normal disrtibution.

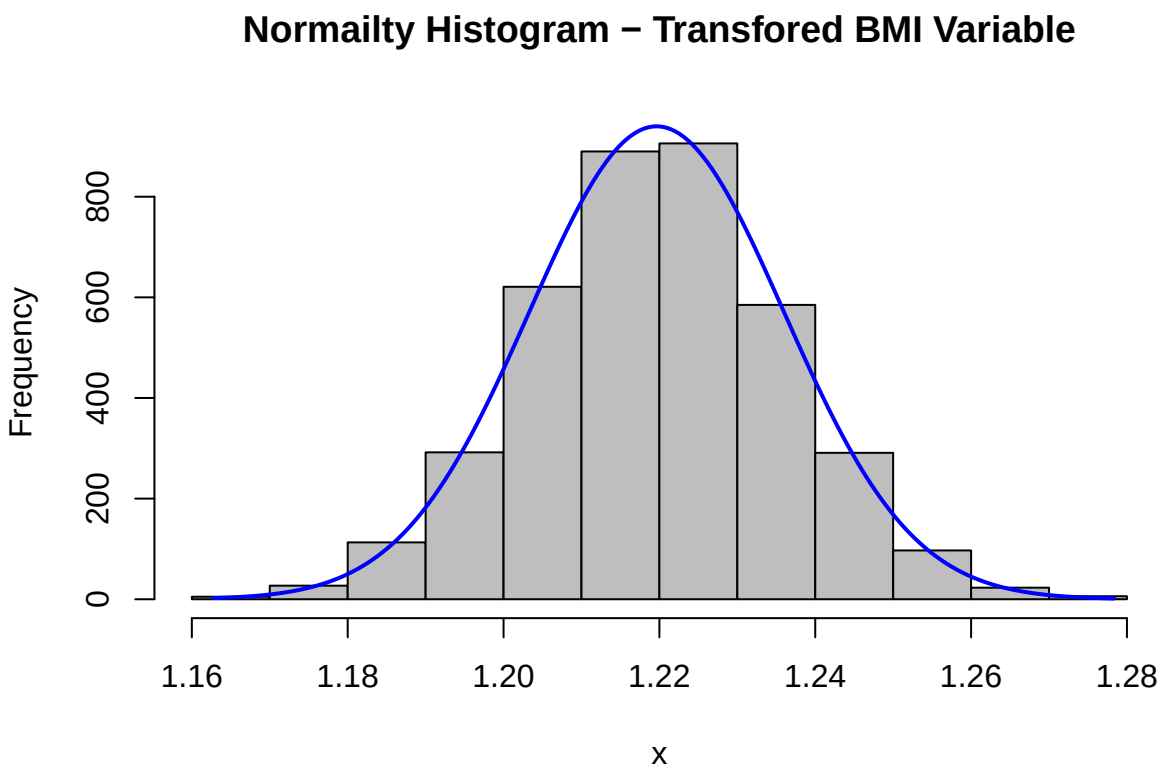


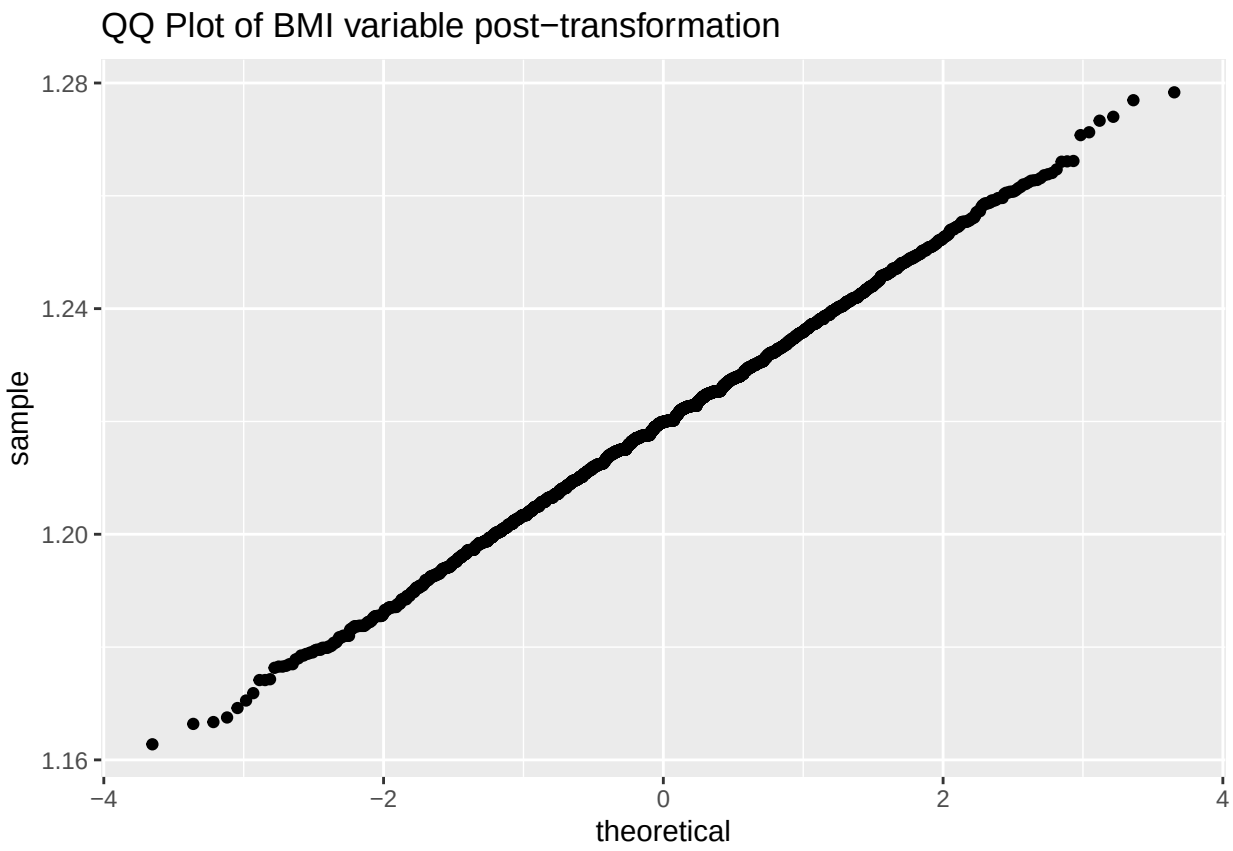




Given the results we saw regarding the BMI distribution , we will perform a Box-Cox tranformation , and re-evaluate the distribution. In the following plot we perform the box-cox , and plot the updated histogram. We can see now that the distribution plot look a lot better and are able to take under normality assumptions.


95%
log−Likelihood
−9600
−10000
−10400
−2
−1
0
1
2
λ


Normailty Histogram − Transfored BMI Variable
Frequency
0
200
400
600
800
1.16
1.18
1.20
1.22
1.24
1.26
1.28
x



Part 5 - Relation of BMI with other values

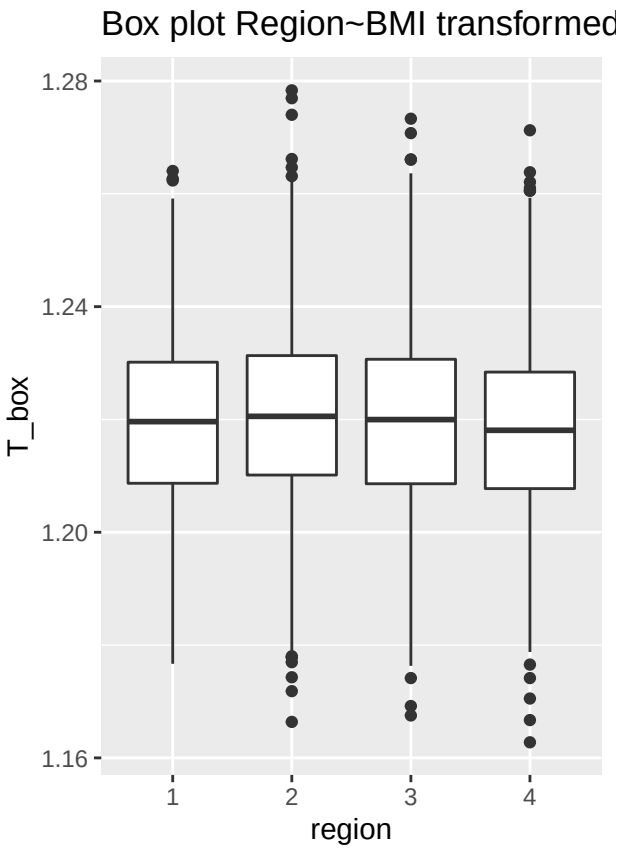
Lets review the relationship of the BMI with other features , identify some correlations and potential descriptive analysis.

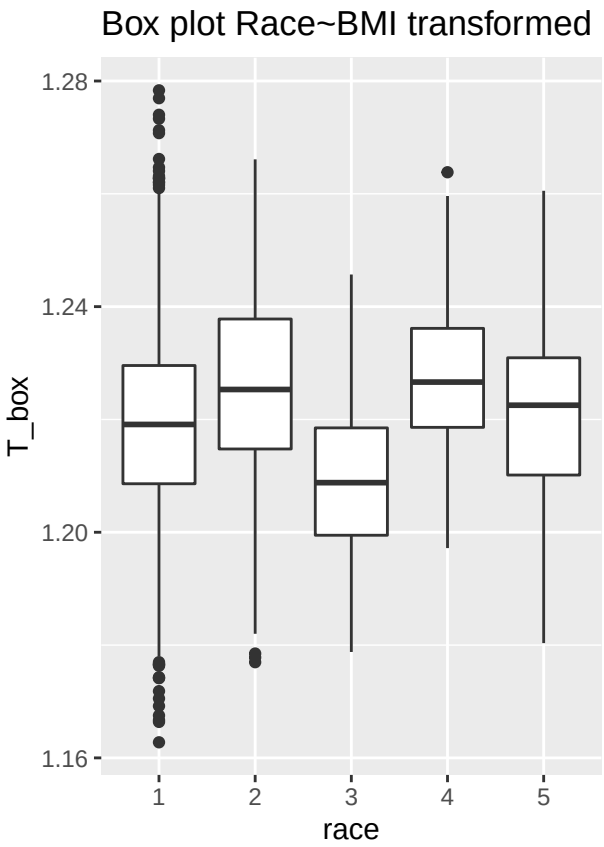
Looking at the BMI grouped by regions , we can see that Northeast is with the smallest range , and least amount of outliers. We can also see that the medians are similar to all groups.

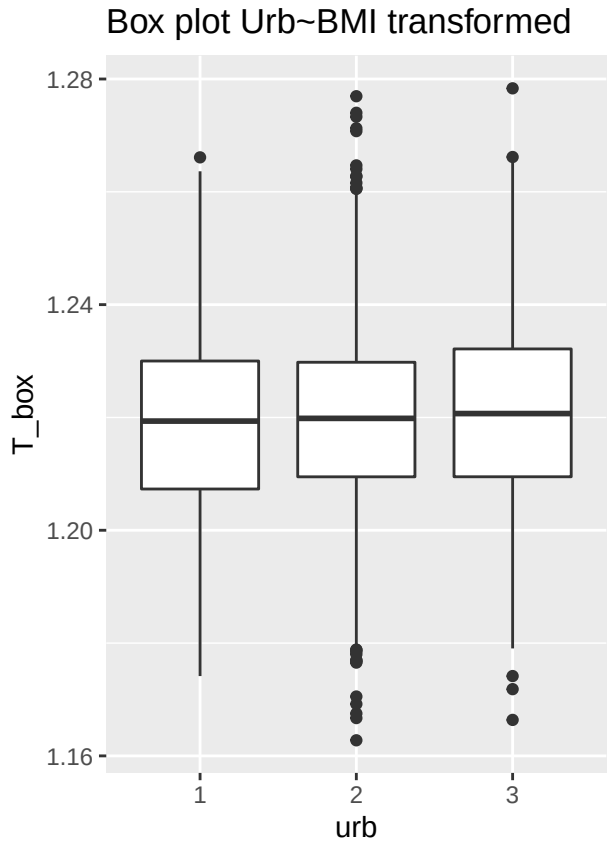
While plotting the box plots of the BMI , vs the Race feature ,we see different ranges through the grouop , having “white” as expected with the largest range . This is expected since it holds the majority of the observations. What is interesting is that most (if not all) the outliers that are potential from this set , are under the “white” category.

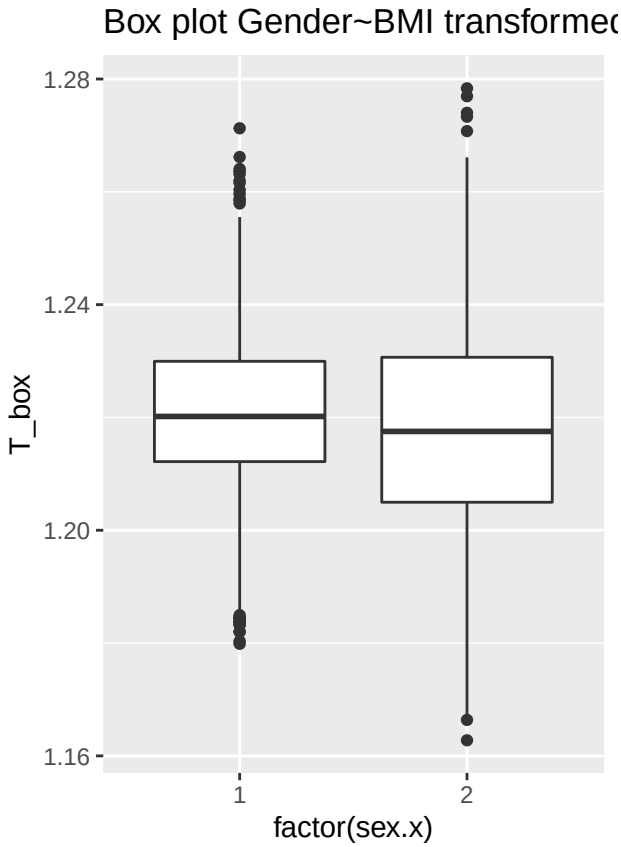
Looking at the relationship between BMI and Urb , we find equal medians with Class 2 (Suburban) containing some additional long tails , though we are still under standard acceptable ranges , so this is considered fine.

Next we can see the box plots per gender. The body of Group 2 is larger under almost equaly sized groups , which is interesting to understand full relation and correlation between the two .


Box plot Region~BMI transformed
1.28
1.24
1.20
1.16
T_box
1
2
3
4
region


Box plot Race~BMI transformed
1.28
1.24
1.20
1.16
T_box
1
2
3
4
5
race

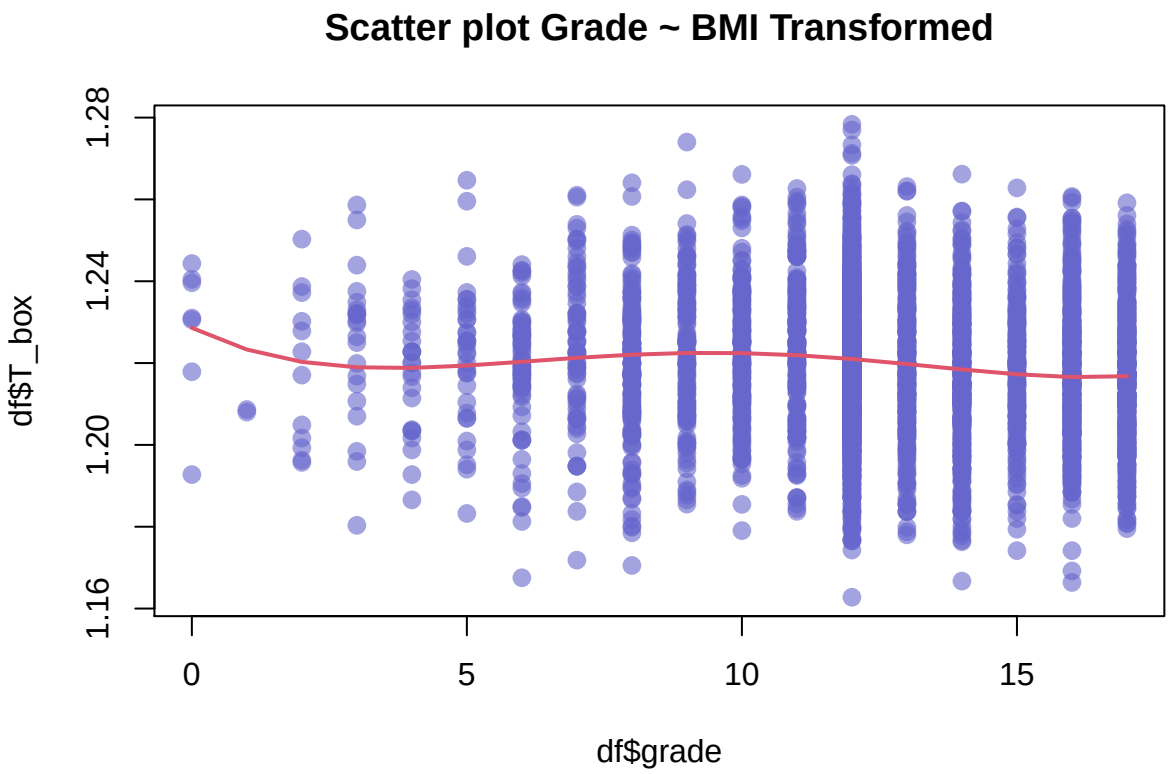


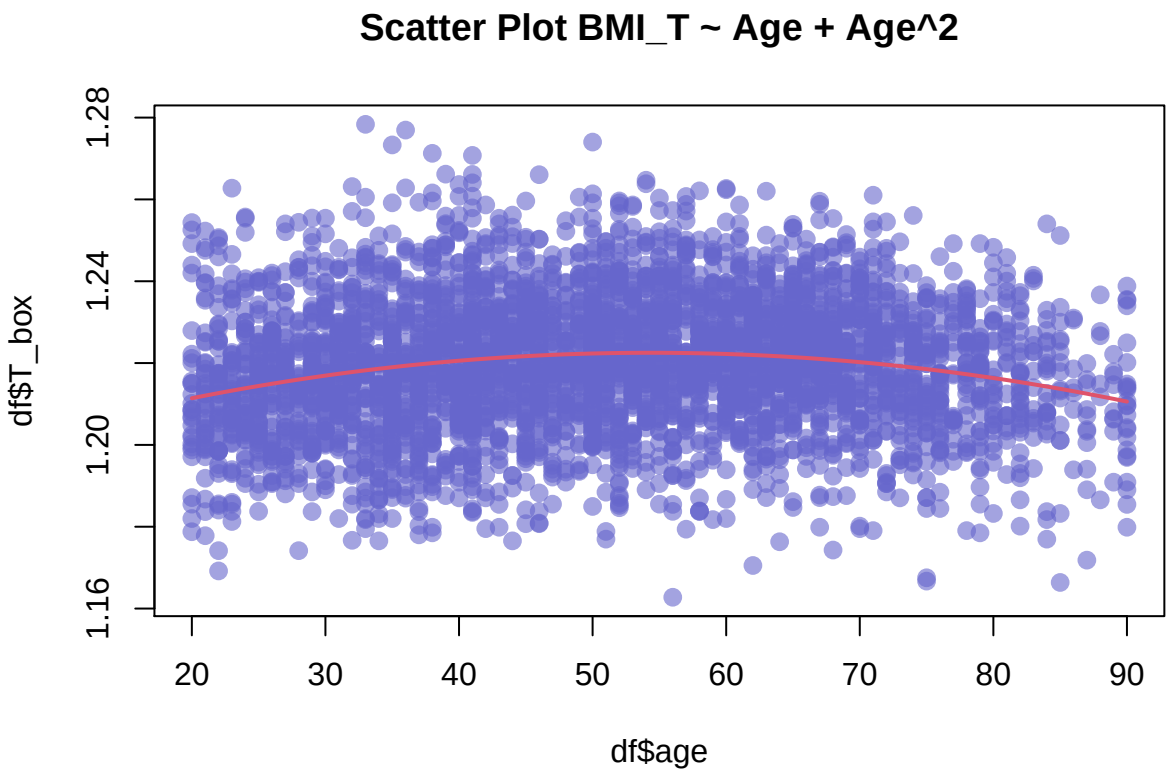


Next we will look at the negative correlation between the BMI transformed and the grade. The ranges are much larger for grades 12 + , this makes sense , since the age range is probably larger and more diverse , vs lower grades which probably contains a more unified set of patients. What is interesting here is using a polinomial regression line we can see a small peak around 11-12 , and minimums around 3-6 and around 15.

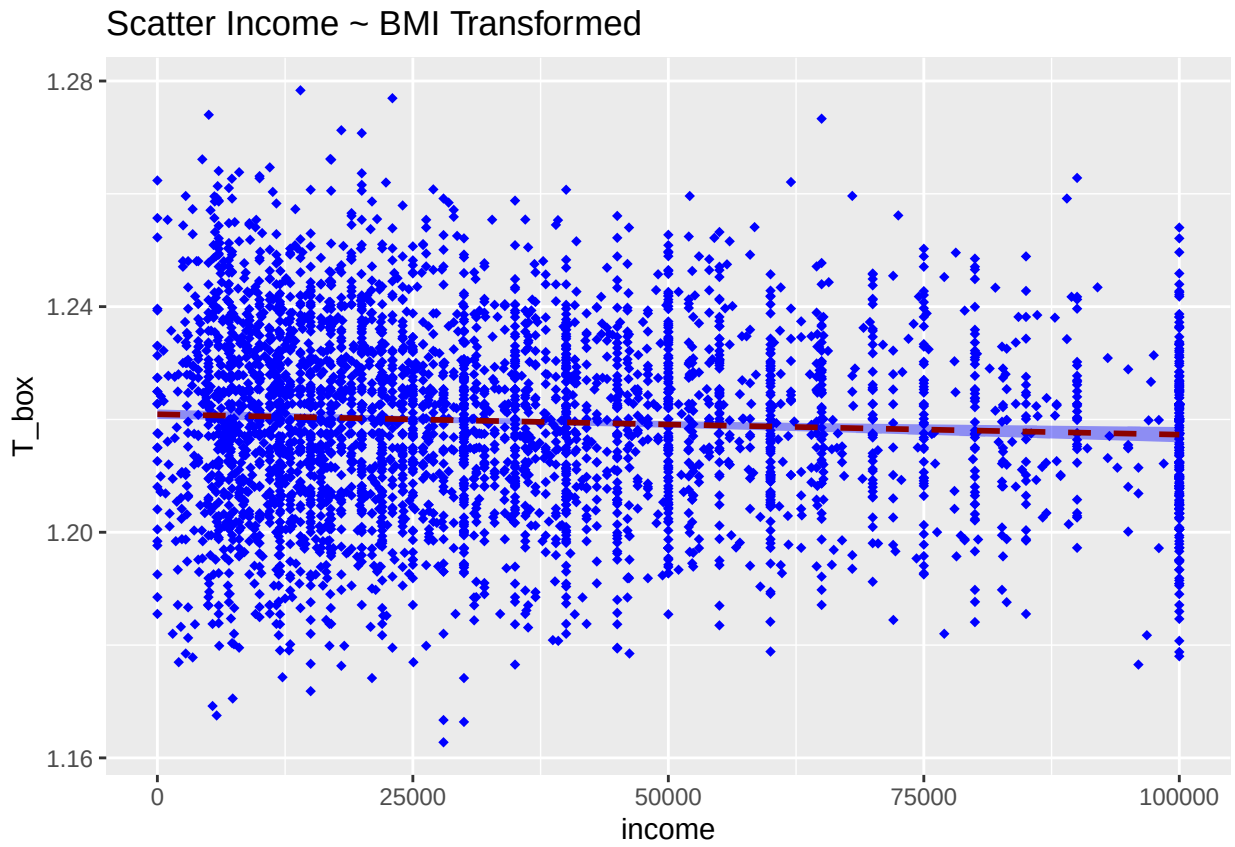
In the following figure we demonstrate a similar analysis to the previous though with age vs BMI .

Here we can see that the highest BMI range around the ages of 50-70. Leveraging this , we can consider using a Transformation of this variable for the prediction of the BMI.

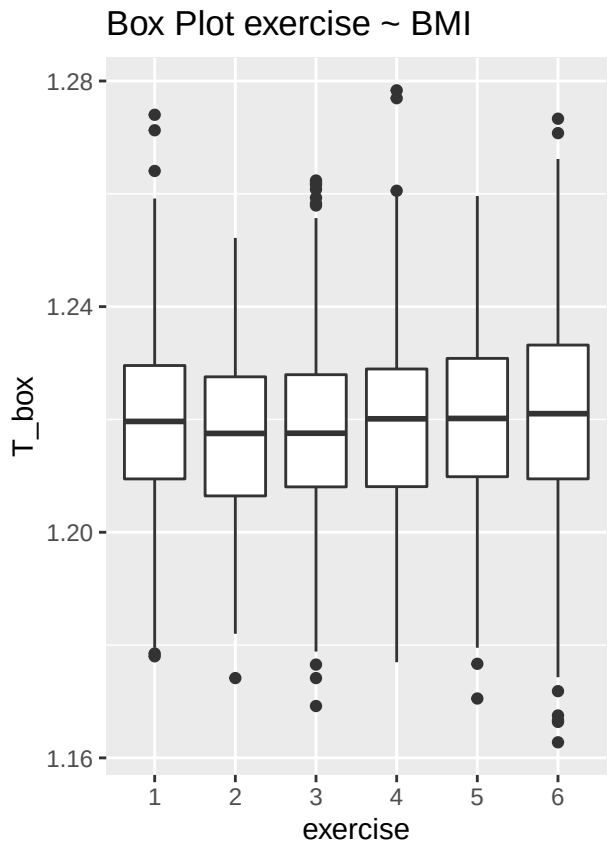

Scatter plot Grade ~ BMI Transformed
df$T_box
1.16
1.20
1.24
1.28
0
5
10
15
df$grade

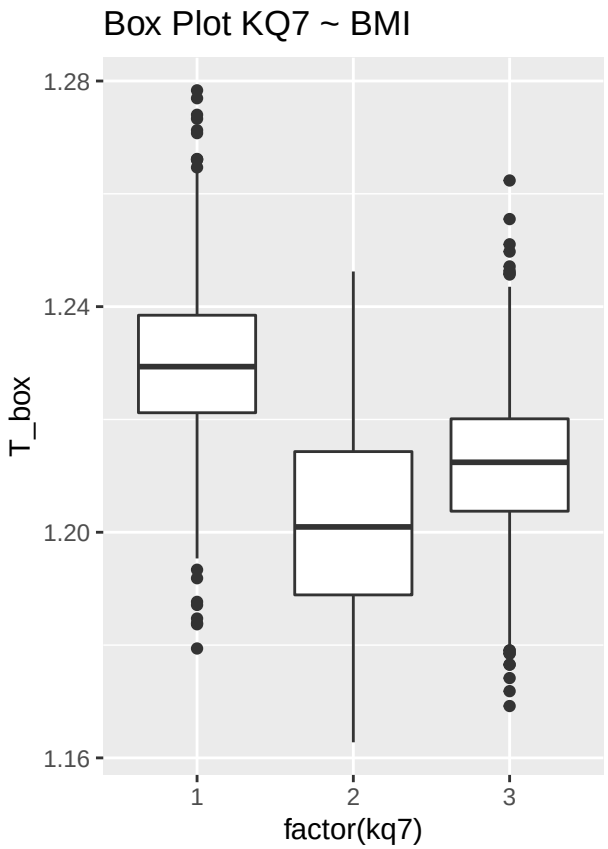


Here we can see a scatter plot of the relationship between the income and the income. The trend is pretty linear and subtle. We will suspect that there is not significant relationship between the two vatiables and will verify this is the following steps.



Here there is a split according to the categorical sets of the Excercise variable. The medians are pretty equal , per all sets. Though by splitting by the kq07 , into its 3 categories , we can see a variance within the groups. The right tail observations belong to group 1 which is interesting.





Part 6 - Univariate Regressions

In this phase we will choose the set of variables ('region','urb','income','age','sex.x','race','grade','exercise','kq7','dt01', 'dt02','dt03','dt06','dt07','emp_status') and run univariate regressions of BMI ~ X , where X will be a univariate variable from the former list. Given the results , we can evaluate which of the following features are not worth consideration in the building of the model for the BMI. I decided to be relaxed with this constraint and set the threshold to 0.1 , and let the feature 'urb' enter the following steps.

In the following table we can see the results of the univariate analysis. I decided to remove the feature dt06 due to the insignificant p value, giving us an indicative that the feature will probably not contribute to the model.

Characteristic	**N**	**Beta**	**95% CI**	**p-value**
region	3,856			0.007
1				
2		0.00	0.00, 0.00	
3		0.00	0.00, 0.00	
4		0.00	0.00, 0.00	
urb	3,856			0.088
1				
2		0.00	0.00, 0.00	
3		0.00	0.00, 0.00	
income	3,856	0.00	0.00, 0.00	<0.001
age	3,856	0.00	0.00, 0.00	0.004
sex.x	3,856			<0.001
1				
2		0.00	0.00, 0.00	
race	3,856			<0.001
1				
2		0.01	0.00, 0.01	
3		-0.01	-0.01, -0.01	
4		0.01	0.00, 0.02	
5		0.00	0.00, 0.00	
grade	3,856	0.00	0.00, 0.00	<0.001
exercise	3,856			<0.001
1				
2		0.00	0.00, 0.00	
3		0.00	0.00, 0.00	
4		0.00	0.00, 0.00	
5		0.00	0.00, 0.00	
6		0.00	0.00, 0.00	
kq7	3,856			<0.001
1				
2		-0.03	-0.03, -0.03	
3		-0.02	-0.02, -0.02	
dt01	3,856			<0.001
1				
2		-0.01	-0.01, -0.01	
dt02	3,856			<0.001
1				
2		0.00	0.00, 0.00	
dt03	3,856			<0.001
1				
2		0.00	-0.01, 0.00	
dt06	3,856			0.52
1				
2		0.00	-0.01, 0.00	
dt07	3,856			<0.001
1				
2		-0.01	-0.01, 0.00	
emp_status	3,856			0.011
1				
2		0.00	0.00, 0.00	
3		0.00	0.00, 0.00	
4		0.00	0.00, 0.00	

Part 7 - Multiple Regression

Variable Screening

We screened the datasets features using the former runs results , and the pvalue outcomes.

STEP AIC

Once we have the fileterd set , we can run the Step AIC model , (using both sides) to filter out more features using a more sophisticated method. AIC is found to be pretty useful in this situation, and since the target variable is converted to a normal distribution , we are able to apply a log likelihood model.

We can see that the final run gives us an R squared of 0.07 , F statistic of 15.9. Although there are some predictors here which come out to be not as significant in their contribution , I decided to leave them in .

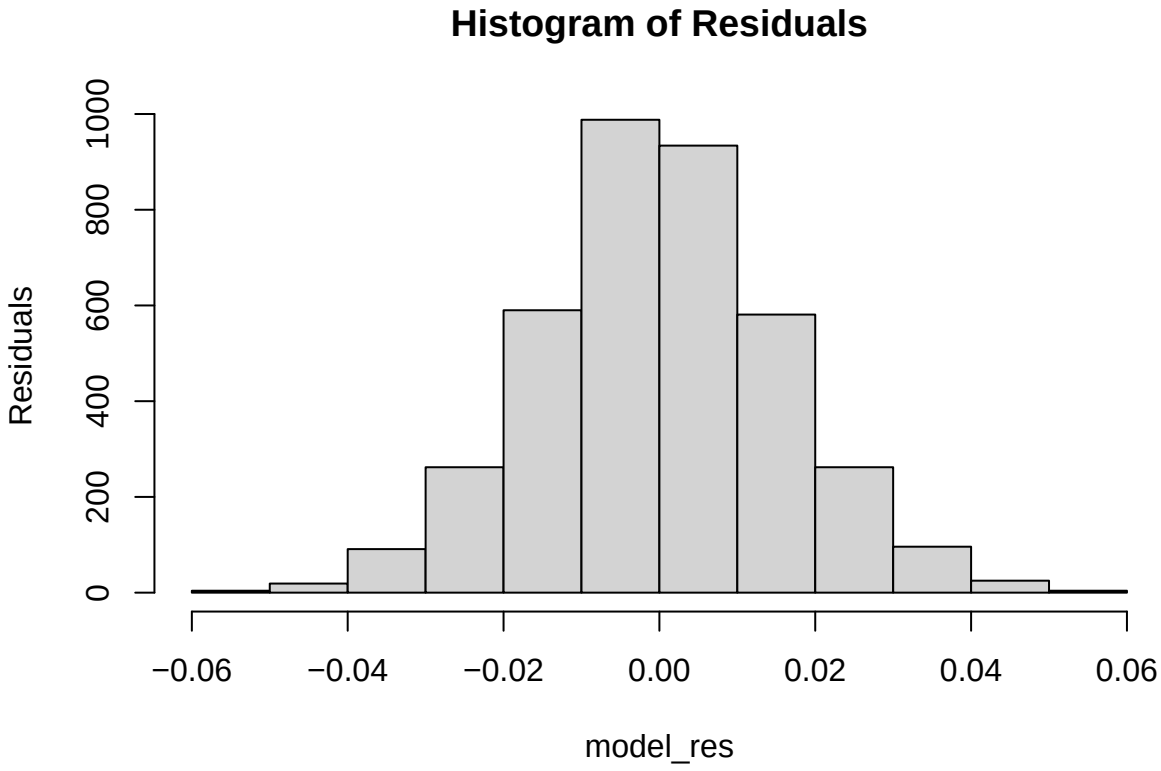
Another thought to be considered is the fact that there are some features that one sub group within the categorical feature has a significant P val , and other members of the domain dont. This could be due to a redundant split of the category, which could result in us merging some of the groups back together.

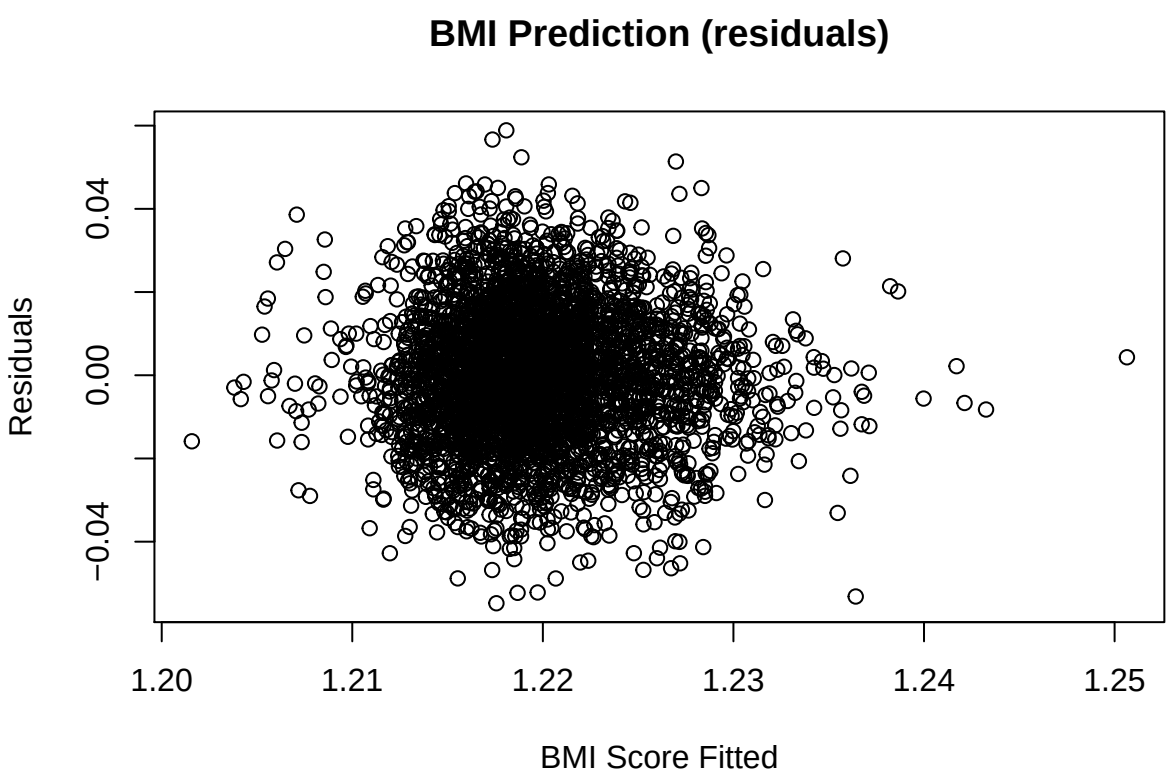
An example for this could be seen under the category Race , where some sub groups have a quite high p value and low |t-stat| , and could possibly be merged into another one. (Possibly expand Other to be all but White)

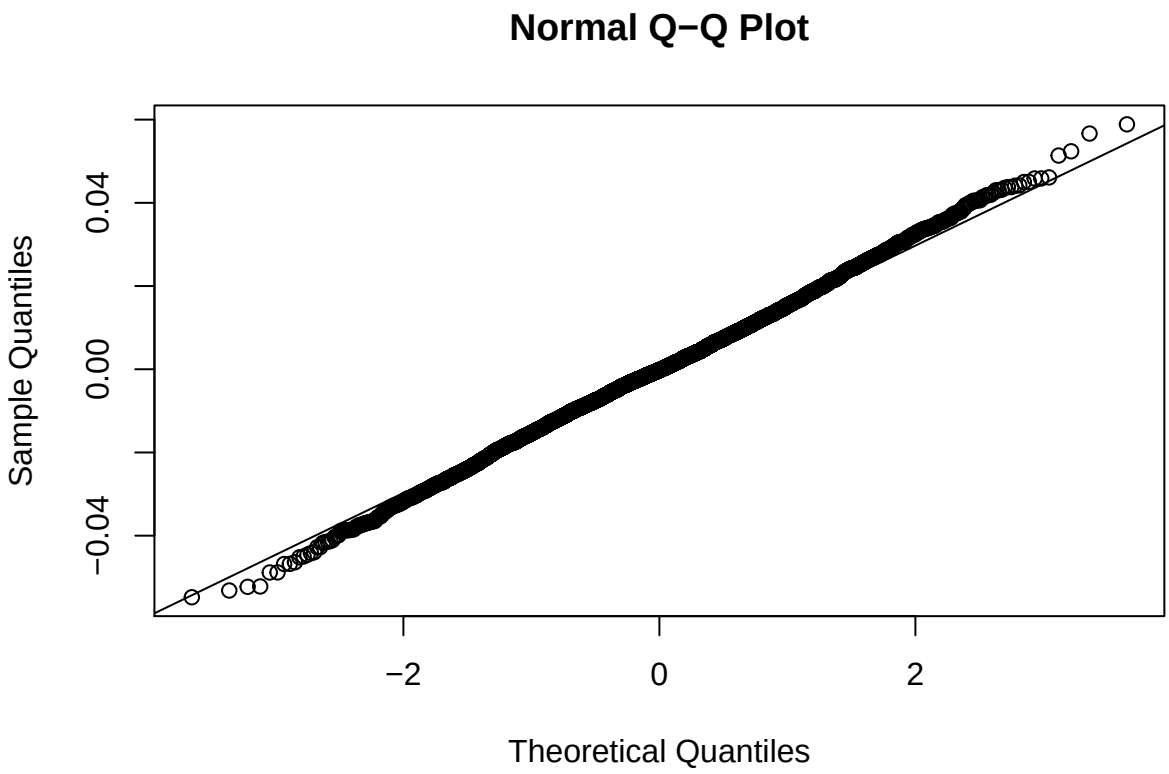
```
##
## Call:
## lm(formula = T_box ~ region + sex.x + race + grade + exercise +
##     dt01 + dt02 + dt07 + emp_status, data = uvariate_filtered_df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.054797 -0.010044 -0.000221  0.010010  0.058866
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.243e+00  2.332e-03 532.930  < 2e-16 ***
## region2      1.291e-03  7.692e-04   1.679 0.093321 .
## region3     -9.407e-05  7.291e-04  -0.129 0.897344
## region4     -7.998e-04  8.290e-04  -0.965 0.334728
## sex.x2      -3.761e-03  5.302e-04  -7.095 1.54e-12 ***
## race2        6.327e-03  8.205e-04   7.711 1.58e-14 ***
## race3       -8.190e-03  2.076e-03  -3.946 8.10e-05 ***
## race4        8.127e-03  3.258e-03   2.495 0.012655 *
## race5        1.483e-03  1.354e-03   1.095 0.273687
## grade       -4.359e-04  9.133e-05  -4.773 1.88e-06 ***
## exercise2   -9.331e-04  1.155e-03  -0.808 0.419188
## exercise3    1.270e-04  7.961e-04   0.160 0.873232
## exercise4    8.762e-04  1.119e-03   0.783 0.433709
## exercise5    1.636e-03  1.328e-03   1.232 0.217908
## exercise6    2.448e-03  7.077e-04   3.458 0.000549 ***
## dt012       -1.005e-02  1.122e-03  -8.961  < 2e-16 ***
## dt022       -1.920e-03  9.163e-04  -2.096 0.036168 *
## dt072       -6.165e-03  1.409e-03  -4.375 1.25e-05 ***
## emp_status2 -2.230e-03  8.396e-04  -2.656 0.007932 **
## emp_status3 -3.322e-04  1.487e-03  -0.223 0.823269
```

```
## emp_status4 -1.156e-03  5.974e-04  -1.934 0.053145 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01577 on 3835 degrees of freedom
## Multiple R-squared:  0.07661,    Adjusted R-squared:  0.07179
## F-statistic: 15.91 on 20 and 3835 DF,  p-value: < 2.2e-16
```

Lets review the residuals and break them down to understand their behavior. In the following plots we can see the histogram of the residuals and the scatter plot of the fitted vs the residuals. We can see that the scatter surrounds 0 which is encouraging. The QQ plot confirms the normality of the residuals. We can also see that looking at the Durbin Watson metric, we can see that the score is 2.03 which given that it is close to 2 , indicates that there is close to no auto correlation within the residuals - meanining that the residuals are independendant.

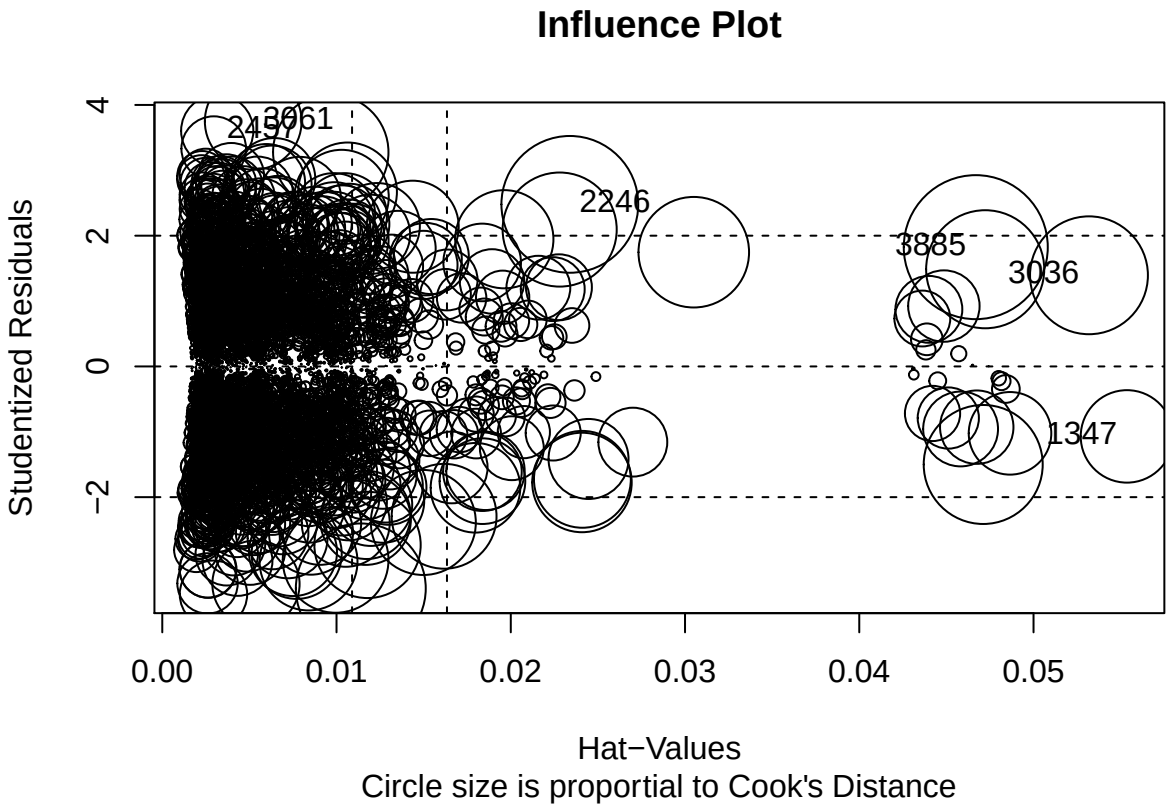



BMI Prediction (residuals)
Residuals
0.04
0.00
−0.04
1.20
1.21
1.22
1.23
1.24
1.25
BMI Score Fitted



Influence Plot

Here we will test out the influence of the different predictors , extarcting the Hat values , and the Cooks distance. In the following plot we can see that there are two clusters whcih are differed by the Hat values (above and below 0.04).



Interactions Effects Lets consider using interactions between variables. For this , we can leverage the previos architechture of the AIC model , just run it including the interactions of the different predictors. We will test this using the backward direction starting off with all the predictors. The results were quite encouraging. While adding the additional interactions we manage to filter out most of them using the AIC , and remain with:

BMI_transformed ~ region + income + age + sex.x + race + grade + exercise + dt01 + dt02 + dt07 + emp_status + region:dt07 + income:age + income:sex.x + income:dt02 + income:dt07 + income:emp_status + age:sex.x + age:grade + age:exercise + age:dt01 + age:dt02 + age:dt07 + age:emp_status + sex.x:race + sex.x:dt02 + sex.x:dt07 + dt01:emp_status

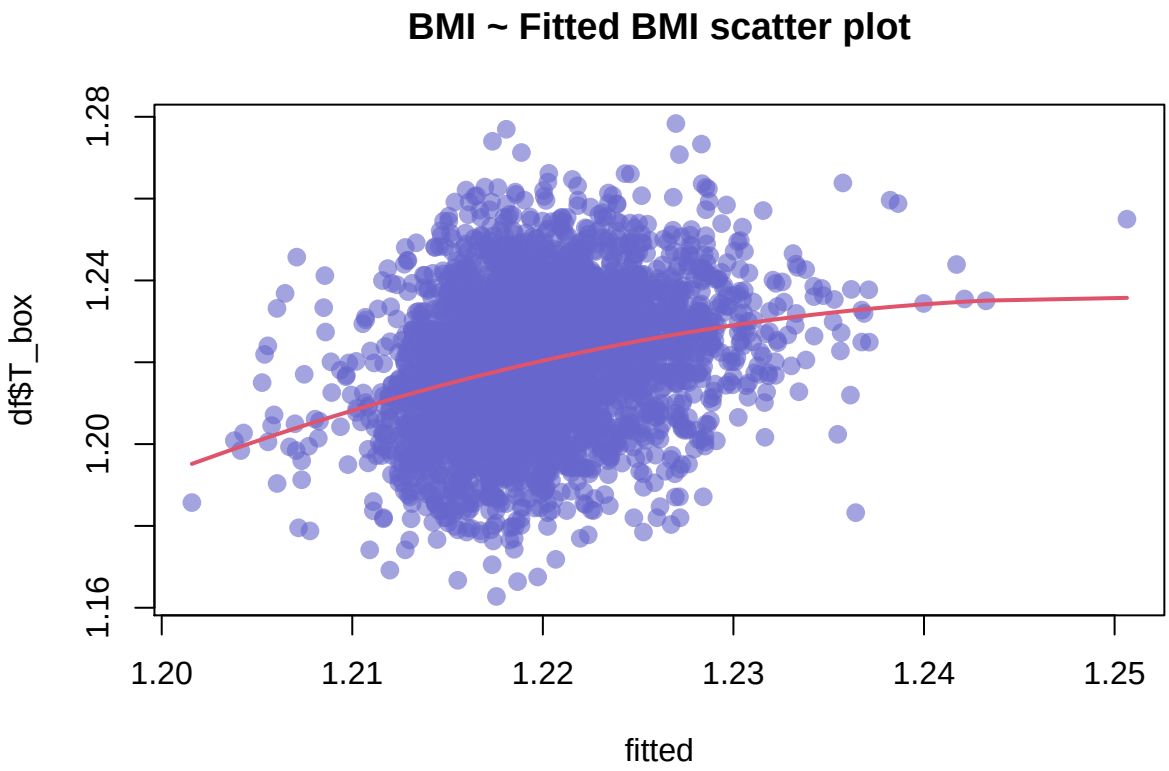
Our R squared increases from 0.07 to 0.11 , which means that there is a significant amount of variance explained from the interactions. Our F statistic is still significant , and P value as well.

For the sake of the inspection , I plotted out the fitted preditcions vs the actual ones. The intuition here was to validate whether a linear model fits the case we are trying to solve, or whether we might need to address this with a polinomial curve. Given the plot , we can assume that a linear model generalizes the problem well enough.

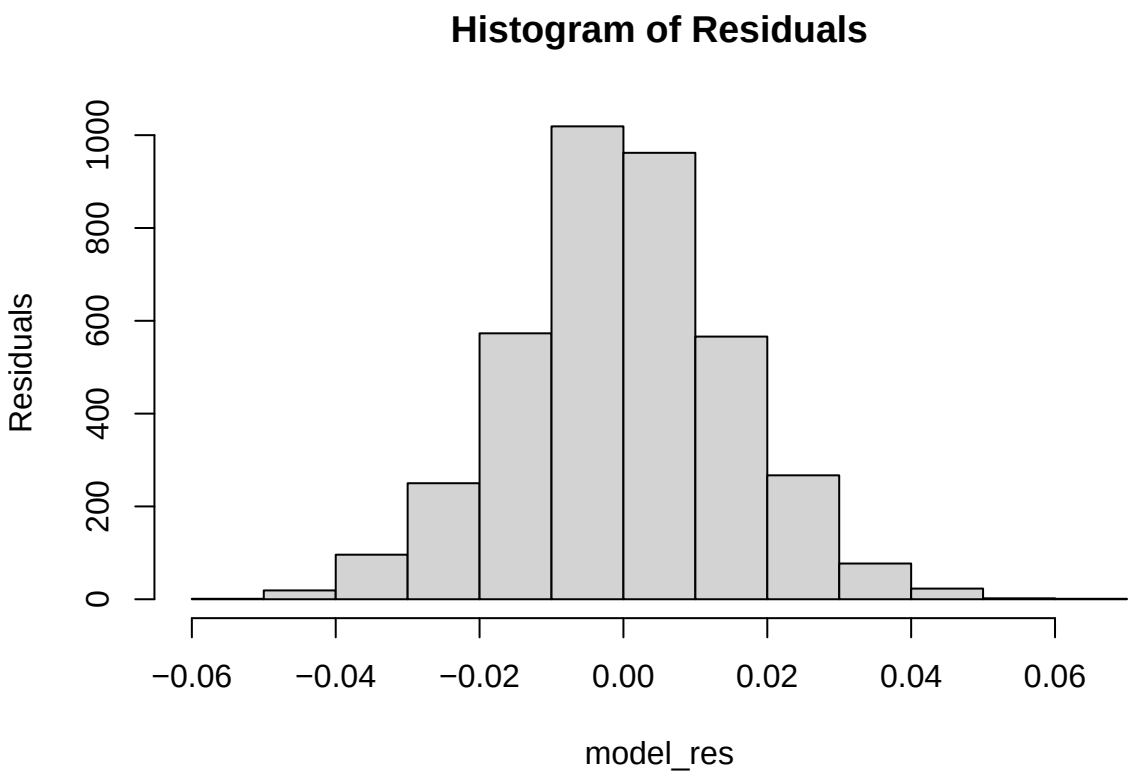
```
##
## Call:
## lm(formula = T_box ~ region + income + age + sex.x + race + grade +
##     exercise + dt01 + dt02 + dt07 + emp_status + region:dt07 +
##     income:age + income:sex.x + income:dt02 + income:dt07 + income:emp_status +
##     age:sex.x + age:grade + age:exercise + age:dt01 + age:dt02 +
##     age:dt07 + age:emp_status + sex.x:race + sex.x:dt02 + sex.x:dt07 +
##     dt01:emp_status, data = uvariate_filtered_df)
##
```

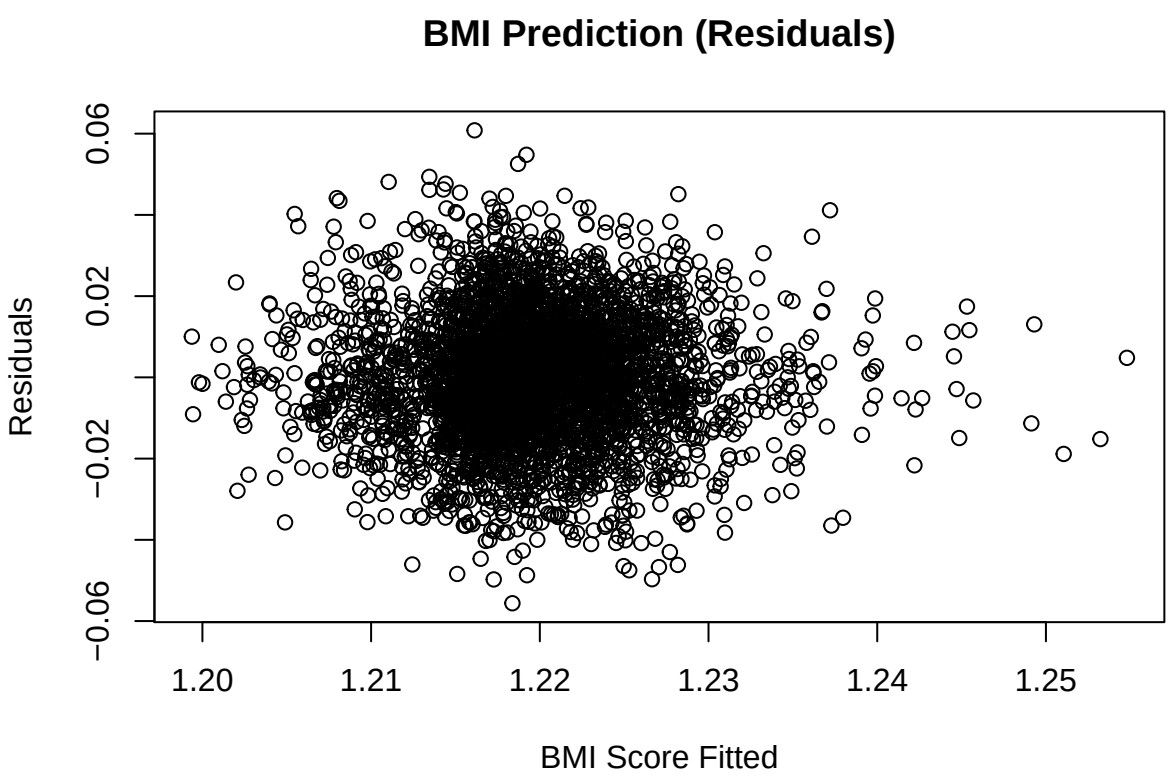
```
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.055610 -0.009614 -0.000370  0.009559  0.060815
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.305e+00  1.108e-02 117.826  < 2e-16 ***
## region2            -6.337e-03  3.720e-03  -1.704 0.088533 .
## region3            -8.727e-03  3.631e-03  -2.403 0.016292 *
## region4            -1.124e-02  4.694e-03  -2.394 0.016732 *
## income             -2.776e-07  8.150e-08  -3.406 0.000666 ***
## age                -9.775e-04  1.718e-04  -5.689 1.38e-08 ***
## sex.x2             -1.308e-03  3.923e-03  -0.333 0.738784
## race2               1.791e-03  1.242e-03   1.442 0.149383
## race3              -8.549e-03  3.013e-03  -2.838 0.004568 **
## race4               5.497e-03  4.695e-03   1.171 0.241770
## race5              -1.307e-03  1.814e-03  -0.720 0.471290
## grade              -1.223e-03  3.135e-04  -3.900 9.78e-05 ***
## exercise2          -5.745e-03  3.572e-03  -1.608 0.107860
## exercise3          -3.808e-03  2.444e-03  -1.558 0.119389
## exercise4          -2.036e-03  3.420e-03  -0.595 0.551601
## exercise5          -5.569e-04  3.909e-03  -0.142 0.886733
## exercise6           6.787e-03  2.265e-03   2.997 0.002749 **
## dt012              -2.201e-02  3.731e-03  -5.900 3.96e-09 ***
## dt022              -1.927e-02  4.371e-03  -4.409 1.07e-05 ***
## dt072              -3.056e-02  8.744e-03  -3.495 0.000480 ***
## emp_status2        -3.941e-03  4.336e-03  -0.909 0.363469
## emp_status3         5.872e-03  9.198e-03   0.638 0.523264
## emp_status4         1.371e-02  3.286e-03   4.174 3.06e-05 ***
## region2:dt072       8.149e-03  3.796e-03   2.147 0.031865 *
## region3:dt072       9.109e-03  3.699e-03   2.462 0.013846 *
## region4:dt072       1.081e-02  4.761e-03   2.270 0.023267 *
## income:age          1.837e-09  7.833e-10   2.345 0.019101 *
## income:sex.x2      -6.696e-08  2.071e-08  -3.233 0.001236 **
## income:dt022        1.016e-07  3.422e-08   2.970 0.002999 **
## income:dt072        1.382e-07  6.007e-08   2.300 0.021495 *
## income:emp_status2 -2.433e-08  3.117e-08  -0.781 0.435096
## income:emp_status3 -1.002e-07  5.469e-08  -1.832 0.067085 .
## income:emp_status4 -4.989e-08  2.548e-08  -1.959 0.050244 .
## age:sex.x2          9.365e-05  3.170e-05   2.955 0.003151 **
## age:grade           1.467e-05  5.379e-06   2.727 0.006425 **
## age:exercise2       1.122e-04  7.128e-05   1.574 0.115489
## age:exercise3       8.952e-05  4.717e-05   1.898 0.057774 .
## age:exercise4       7.174e-05  6.929e-05   1.035 0.300562
## age:exercise5       5.588e-05  8.140e-05   0.686 0.492475
## age:exercise6      -7.832e-05  4.101e-05  -1.910 0.056235 .
## age:dt012           2.905e-04  7.615e-05   3.815 0.000138 ***
## age:dt022           2.025e-04  6.306e-05   3.212 0.001330 **
## age:dt072           3.044e-04  1.221e-04   2.492 0.012749 *
## age:emp_status2     8.719e-05  5.463e-05   1.596 0.110564
## age:emp_status3    -5.758e-05  1.176e-04  -0.490 0.624331
## age:emp_status4    -1.259e-04  4.330e-05  -2.908 0.003664 **
## sex.x2:race2        7.160e-03  1.635e-03   4.378 1.23e-05 ***
## sex.x2:race3        1.883e-03  4.062e-03   0.464 0.642866
```

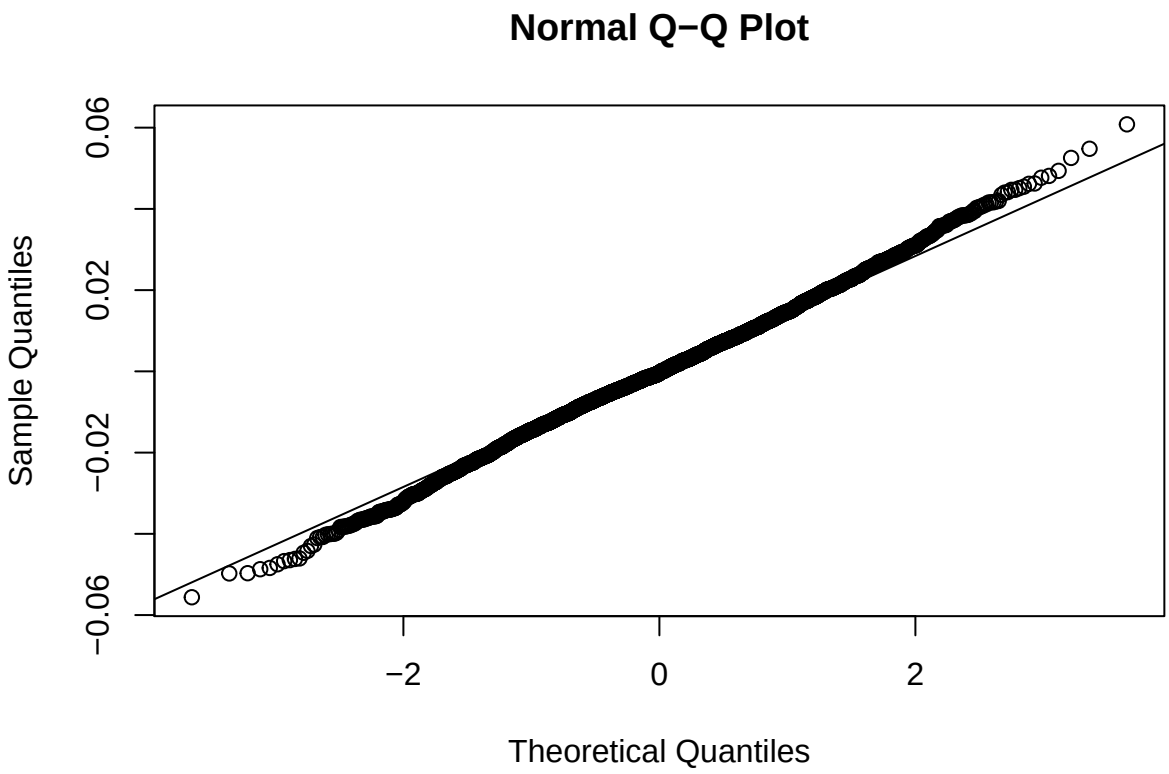
```
## sex.x2:race4        3.080e-03  6.381e-03   0.483 0.629313
## sex.x2:race5        4.028e-03  2.596e-03   1.551 0.120873
## sex.x2:dt022        3.619e-03  1.794e-03   2.017 0.043774 *
## sex.x2:dt072       -9.590e-03  2.864e-03  -3.348 0.000822 ***
## dt012:emp_status2  -9.527e-04  3.388e-03  -0.281 0.778552
## dt012:emp_status3   2.997e-04  6.385e-03   0.047 0.962567
## dt012:emp_status4  -7.092e-03  2.534e-03  -2.799 0.005156 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01538 on 3801 degrees of freedom
## Multiple R-squared:  0.1288, Adjusted R-squared:  0.1165
## F-statistic: 10.41 on 54 and 3801 DF,  p-value: < 2.2e-16
```



And by re-iterating the residuals evaluation again , we get:


Histogram of Residuals
Residuals
0
200
400
600
800
1000
−0.06
−0.04
−0.02
0.00
0.02
0.04
0.06
model_res

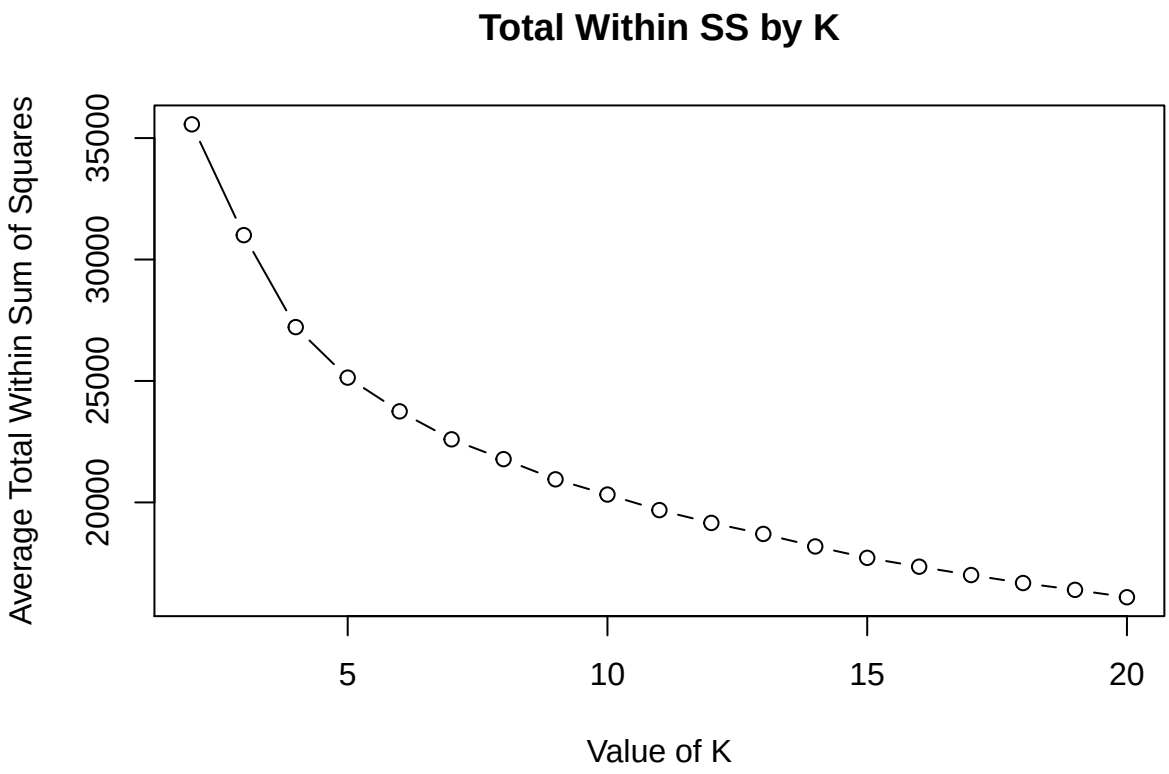

BMI Prediction (Residuals)
Residuals
0.06
0.02
−0.02
−0.06
1.20
1.21
1.22
1.23
1.24
1.25
BMI Score Fitted



Additional Analysis

This part will conduct some additional analysis on the dataset to characterize the data just a bit more and attempt to aid us with fully understanding the behavior and caveats of the dataset.

In the first part I performed a clustering method of K means to try and classify the observants within predefined clusters. We use the elbow method to determine that there should be 4 clusters.



Once we have the optimal number of clusters , we can re-run the kmeans on the set, and assign to each observation its label.

The following plot , describes performance of the Kmeans , reducing the dimension to 2 while still holding the optimal amount of variance using PCA.

We can see that the clsuters have an interesting fit , and the number of elements in each cluster is pretty significant with 997,1124,942,793 obs respectfully.

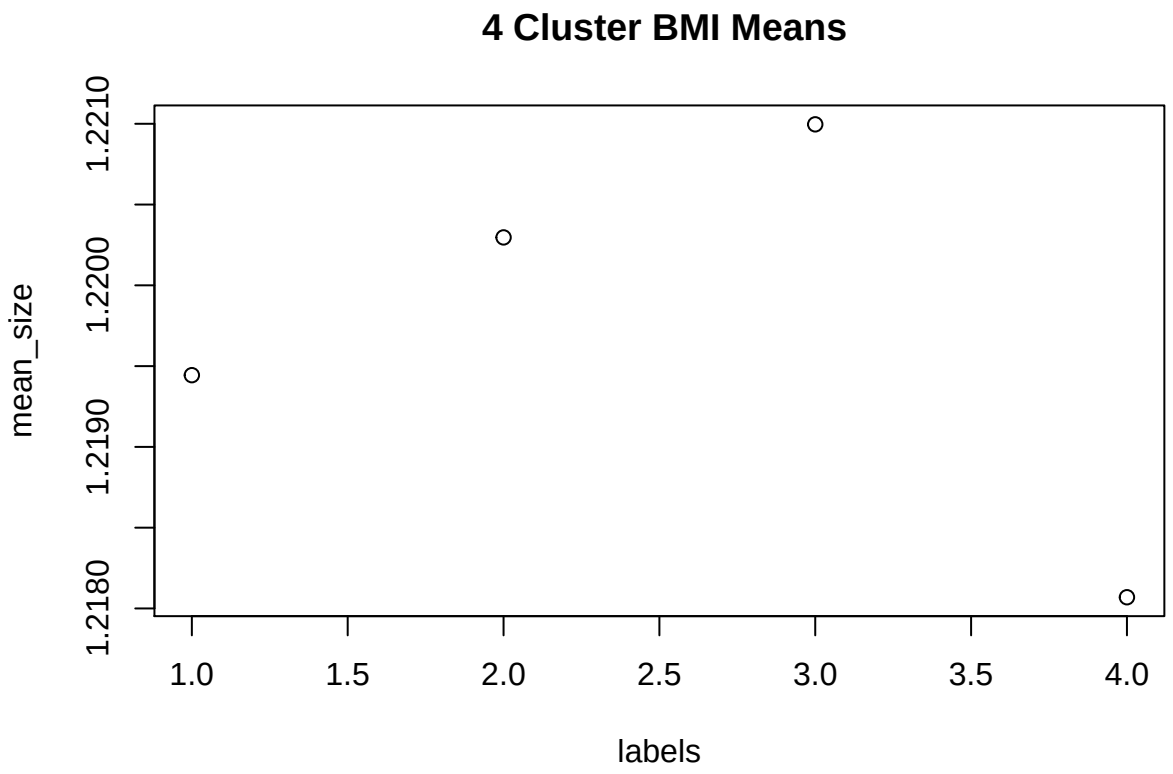
To conclude this analysis, we assign to each observation its label , and group the dataframe by the labels assigned -> and calculate the BMI mean per group.

The motivation here is to see whether the averages are different per group (reminder that BMI was not)

What we can see in the figure below the plot of the 4 averages of BMI , per group. We can see that visually there is a distinctive measurement per group that we can consider while characterizing the BMI set in the future.

Another recommendation here would be to categorize the BMI into 4 groups , and maybe analyze accordingly. We might be able to create stroger relationships while perfroming the analysis on discrete groups rather than a continuous variable.


Plot of clusters after PCA dimension reduction
0.04
0.02
0.00
−0.02
−0.04
PC2 (21.81%)
−0.02
0.00
0.02
PC1 (33.59%)
cluster
1
2
3
4



Conclussion

To conclude this analysis , we can see that the BMI parameter can be explained and described by the current set of features. There are several more options to enhance this analysis , such as additional features (new features, or transformations of current ones.) We found that there is a clear connection and relationship between the BMI and different elements of the patients such as demographic informtion , different surveys etc.

Another next step that may be interesting to check is testing additional models (consider decision tree based models) and perform a comparison of predictions between the models . Compare feature importance , and residuals variance.