

# Advanced Data Analysis (DSC424-710)

## Final Project

*Louis Huang, Mark Wismer, Michal Chowaniak, Ryan Patrick, Sidney Fox*

*November 14, 2018*

### **1 Executive Summary**

The cost of healthcare within the United States has become a hot button issue and major burden for many individuals and families. In 2016, the most recent year in which statistics were published, the average person spent over \$10,000 a year on health expenditures (CMS 2018). At the national level, that equates to \$3.3 trillion, or 17.9% of the Gross Domestic Product (GDP). The Center for Medicare and Medicaid Services (CMS) projects healthcare spending to increase by an average rate of 5.5 percent – 2.5 percentage points higher than the average rate of inflation.

With healthcare costs rising at a steady rate, understanding the contributing factors that lead to high and low medical costs are extremely important. This project sought to identify contributing factors, as well as determine if a medical cost range could be predicted based on National Health Survey data collected and made publicly available by The National Center for Health Statistics (NCHS).

Beginning with the raw “Family” data set from NCHS, a number of preprocessing actions were applied to the survey data in an effort to identify variables that were missing values. Once the data was determined to be in a structure conducive to analysis, a battery of statistical models were applied to the modified survey data. Methods were applied to ensure

the data contained variables that displayed semblance of predictive power towards a survey respondent's self-reported health care costs.

The original data set contained over 33,000 observations and 127 variables. Within the preprocessing phase of the project, 46 variables were determined to be sufficient and included enough data to be valuable during the modeling phase. Correspondence analysis (CA) helped identify what variables had weak to no correlation, and principal component analysis (PCA) produced components that explained a large amount of the data's variance in eight components. Once the preprocessing steps were complete and the data was determined to be conducive to regression analysis, a number of statistical models were applied in an attempt to predict the health care cost range for survey respondents.

As an exploratory approach, the principal component analysis was conducted using the data set. Forty-six continuous variables in the dataset were analyzed using principal component analysis with the goal of reducing the variables to a smaller set of components representing the variance of the underlying variables. Exploratory analysis revealed that five variables should be removed due to severe imbalance of missing data, and 12 variables were removed due to low correlation with other variables. The PCA analyzed 29 variables and produced eight components, representing .64 of the cumulative proportion for the data set.

The components generated were then used as initial inputs into a linear discriminant model and yielded results of approximately 37% accuracy. Removing the components and using the subset of variables that were used as inputs into the PCA model only produced a model that was one percentage point more accurate than the previous model. Using a subset of variables found in the original data set, alongside generated "dummy" variables, yielded a model with 72% accuracy.

Factor analysis was also conducted to determine if there were latent constructs to represent the variables. This analyses reduced the data set down to 59 variables including the dependent *FHICOST* variable. The 59 were further reduced to 29 and have five major

factors were generated from these 29 variables, providing around 60% cumulative variance. A loading coefficients threshold of was set at a cutoff of at least 0.4. Various subsets of the original data set were used as inputs and the most favorable outcome, and thereby the most accurate model, was achieved by leveraging a modified data set that also contained dummy variables. Said model achieved 72% accuracy.

The five factors generated from the variables are as follows:

1. Family Demographics
2. Family Health Needs
3. Family Coverage
4. Family with Elderly
5. Family Work Status

Correspondence analysis provided assurance that at least a subset of the variables were correlated and could thus be used as inputs for other phases of the project without losing a lot variance. The initial data set contained 127 variables through iterations of filtering by employing analysis like correspondence analysis, the project group was able to reduce the number of variables within the data set to approximately 40.

Following correspondence analysis, we analyzed the data using logistic regression in an attempt to predict whether or not Medical Dental costs were high or low. The input for the model included approximately 20 variables and achieved 70% accuracy on training data. However, the generated logistic regression model executed on the test data set was not as performant, as some variables were insignificant.

Along with logistic regression, linear discriminant analysis produced extremely favorable results that definitely provides a solid foundation for further analysis. After three iterations of LDA with different train and test splits and different subsets of input variables or components, the final model exhibited 72% accuracy. The model showed that there are many variables, which predict if a family spends \$500 or less or more than \$500 per year in

medical/dental costs. Income, housing, family structure, type of insurance are among those variables which influence a family medical cost in positive or negative way.

Model results were strong and favorable as the outcomes of Logistic Regression. It was determined that one could accurately predict a respondent's health care cost range with a subset of the variables included in the survey and a year's worth of observations. The executive summary visualizations in the appendix underscore the in-depth analysis completed and the process developed as a product of this analysis. Figure one displays a histogram of the proportion of columns that are missing data, figure two displays a correlation plot between a subset of the variables, and figure three displays a scree plot of the components created as a product of PCA.

---

## 2 Abstract

Determining the contributing factors of medical costs is necessary to address the consistent rise in costs over time as well as for informing healthcare economics policy choices. This project attempted to identify contributing factors for predicting a medical cost range using National Health Survey data made publicly available by The National Center for Health Statistics (NCHS). The analyses focuses on investigating the significant predictors and interesting dependencies of Cost of family medical/dental care ( *FHICOST* ) in the past 12 months. Among the many methods of analysis available to the group, the group settled on what was perceived to be the best and the most appropriate models for our survey data set.

After conducting exploratory cluster analysis and reducing the number of variables using principal components and factor analyses, linear discriminant analysis as well as logistic regression and correspondence analysis were used to determine which variables would be significant for predicting or expressing relationships between the dependent variable, *FHICOST*,

and the remaining independent variables.

Exploratory factor and cluster analyses revealed interesting variable groupings that could be used to inform future studies, but were not useful when included in the LDA, logistic regression, and correspondence analysis for this study. Along with logistic regression, linear discriminant analysis produced extremely favorable results that provides a solid basis for further studies. The final LDA model exhibited 72% accuracy, while correspondence analysis yielded a small subset of variables that exhibited strong correlation and contained insightful information when used to complete downstream phases of the analysis.

---

### 3 Introduction

The goal of the National Health Survey provides estimates on health status of the United States population and was authorized by the National Health Survey Act in 1956. This study was conducted using the “Family” data set from the 2017 National Health Interview Survey. Most variables in the data set relate to family structure, education, health condition, health support from the government and income. The numeric data measures health and medical services as well as various economic measurements to gauge the socioeconomic status of the household. As a result, many of the metric variables concern the number of family members within the household and where they lie on healthcare needs and incomes scales. Categorical variables in the data set are most interesting because they may have the most significant effect on a dependent variable. Extensive data cleaning was performed during the preprocessing phase due to variables containing an imbalance of missing values.

---

## 4 Literature Review

In any given year, a number of research projects are completed based on the previously mentioned health survey - among them the following:

There have been numerous studies that used the previously detailed health survey data to analyze various aspects of healthcare and its effects on the population. For example, one study analyzed the health survey data to review the effects of complementary health on children (Black et al. 2015). The data found that the highest predictor of whether a child used complementary healthcare (i.e. alternative medicine like yoga or herbalists) was the parent. Therefore, while reviewing this data it was important to understand the effects of the household and how healthcare costs are related to the household. If a parent was having health issues, the child would also be affected.

Another such study reviewed the health survey data in an attempt to find an association between obesity and sleep (Jean-Louis et al. 2014). Again providing evidence that the cost of healthcare has various factors that can affect it. The research also presented evidence that the cost of healthcare has other factors, such as sleep apnea or a small child in the household to feed it or change it that can cause healthcare costs to increase.

A third study provided evidence that injuries and the need for rehabilitation from those injuries increased one's cost (Ma, Chan, and Carruthers 2014). The paper found that reducing back pain and spinal injuries would reduce the cost of healthcare in totality by upwards of \$200 billion. This paper provided one area of focus within the data for family members that needed help with routine care or personal care needs.

## 5 Methods

After coding missing values (NA) as negative one (-1), correspondence analysis was conducted on a subset of the variables in an attempt to identify correlation between different variables. The model summaries that contained the results of two variables were reviewed to determine if correlation existed between the two variables that were used as inputs. Once correspondence analysis was applied to each variable, the variables that did not exhibit 0.3 or greater correlation were removed from the data set.

Input variables changed and various analyses were run, as well as the training and testing splits. The project group employed LDA as a model and also explored the results produced from enabling the cross validation parameter in order to gain insight into the variables the model identified to be the most impactful.

Parsimonious variable selection is considered a best practice for building efficient and interpretable statistical models. Principal component analysis (PCA) is a common method for reducing the number of variables in a data set down to a few representative components that maximize the variance between the underlying variables. PCA was used as a preprocessing step to reduce the number of variables used in some of the statistical models for this study, while also preserving the variance of the original underlying variables in the data set.

After preprocessing variables, there are 29 numeric variables with the same scale (count of family members), which indicates that the data set is conducive to factor analysis. After processing numerical variables, factor analyses were executed to pair down the original data set continuous variables to 29, providing above a 0.40 loading within a factor while also providing around 60% variance explanation for those 29 variables.

Logistic regression was performed on the modified family health survey data set. The Medical Expense Cost variable was transformed into a binary variable with 0 representing

cost of family medical/dental care in past 12 months less than \$500, and 1 representing the same but over \$500. A substantial amount of time was spent on preprocessing the data set. The logistic regression model was executed using dummy categorical variables and numerical variables on a data set containing 26,000 observations and 128 variables. Logistic regression identified 23 significant variables as key influencers of the dependent variable.

---

## 6 Discussion and Results

In order to make our analysis simpler, the group refactored *FHICOST* from six levels to two levels. So the project team could only analyze and predict whether the family spent more than \$500 in medical/dental care last year, which is a main limitation of the research. For further study in this area, one could try to predict the original six levels of *FHICOST*, but additional changes may arise, like imbalanced data since there are not many families spending a lot of money (more than \$5,000) in our data set. In addition to these challenges, LDA and logistic regression models do not have a very high accuracy (around 70%) in predicting the levels. On the one hand, one can add some extra data set which may have strong correlation with *FHICOST*. Demographic data, geographic data and insurance data are examples of data sets that may allow future analysis to represent families more accurately. In addition to data set augmentation, the project group could improve implemented models by making full use of feature engineering work. One major shortcoming of the research conducted was not full implementation of PCA and FA results to represent all of the continuous variables as a part of input for logistic regression. Ideally, this could be completed in the future. Machine learning techniques like decision tree, random forest and gradient boost may aid in the development of classification models as well.

With an original dataset of well over 100 variables, exploratory analysis was con-



ducted to determine which variables could be easily removed as either insignificant for the study, or containing a large imbalance of missing values that would make the variable unusable for analysis purposes. While categorical variables could be used for regression and other modeling approaches, it was determined that the continuous variables would be analyzed using principal component analysis to reduce the original 46 continuous variables to a few components which would be representative of the variance within the original variables. The relationships between the variables were examined using a correlation matrix and plot (figure 27 in the *Visualization Appendix*). The correlation plot in the *Visualization Appendix* clearly identifies variables with strong and weak correlations between the variables. As a result, 12 variables were removed from the PCA that had weak correlations with other variables of less than .30. After removing five more variables that had greater than 30% of observations missing, the final data set used for the PCA contained 29 variables. The PCA was run with a loading value cutoff of .3, which produced 29 total components. Of the initial 29 components, 12 components contained .792 of the cumulative variance for the variables (figure 29). The scree plot shown in figure 28 identifies eight components with an eigenvalue of greater than or equal to 1. These eight components compose .639 of the cumulative variance for the variables. A few variables were found to be cross loaded on multiple components at the .3 loading score level. It was observed that most of the cross loadings had loading scores below .5, which were then addressed by rerunning the model with a loading score cutoff of .5. The final model only contained two cross loaded variables. These variables were assigned to components according to their highest load scores.

The factor analysis and the principal component analysis that were done did result in similar factors/components for what the factor analysis described as Family Health Needs, therefore showing that these two analyses did result in similar components/factors. It would then be important to run potentially a future regression with the factors and components and find how these factors/components do affect healthcare cost. The project team did find that factor analysis could explain around 60% of the data using five factors, while the

principal component analysis provided 64% with eight factors so it may be important to keep reviewing and re-writing the analysis to find a happy medium to provide that needed variance percentage while also providing good factors/components to research.

For cluster analysis, only 29 continuous variables were used as input, thus the model ignored many useful categorical and ordinal variables. The results of the cluster analysis are not very good at differentiating two levels of *FHICOST* clearly. If one were to try cluster analysis on categorical data using Jaccard dissimilarity, interesting and insightful patterns may be discovered. In addition to the previously mentioned shortcomings, the project group did not check other categorical variables to see if the clusters might differ, which could potentially provide the group with useful insights for other research topics. The variables vary a lot in different clusters, which could also be used for future analysis.

Executing correspondence analysis did not allow the project group to explicitly determine if a health spending range could be determined from the other variables in the data set. That was not the intent of this particular exercise. Correspondence analysis did however provide evidence that a subset of the variables were correlated to some extent. This analysis was insightful and allowed the group to gain greater understanding of the relationship between different variables and the potential trends that may be present within the data set. Further analysis should include correspondence analysis if a new data set were introduced, or if additional variables were added to the existing data set.

Linear discriminant analysis produced favorable results that answered the primary research directly - based on the results of the national health survey, one can predict the health spending range of a family or individual. Cross validation was applied to the subset of variables used as inputs for the lda model to identify what variables the model found to be the most impactful in determining the target variable's class. The model did not exhibit extremely strong accuracy, and that would most likely be the main focus of future analysis that included this data set.

The factor analysis conducted on the data set did not directly answer or explain the research question, as it was not related to determining significant predictors of healthcare costs. Factor analysis did however provide factors that could be used in a regression to find the factors' effects on healthcare costs. The factor analysis showed that 29 variables within the study can be used to predict approximately 60% of the variable contribution to healthcare costs and can be broken out by the five factors and their effects. It would be interesting and insightful to augment the current data set with detailed family demographic information.

As discussed in the methods, the results for the factor analysis provided 19 variables broken into three factors that explained around 50% of the variance. The three factors can be described as follows for each family: 1) Likelihood of Insured 2) Likelihood of health issues 3) Likelihood of Being Elderly. See the appendix for the loadings for each variable and the total variance for each factor. Furthermore, each factor did have specific variables that influenced the factor more. For the first factor, a family with insurance and private insurance affected the value of the factor by around 80% each. The second factor was influenced at around 85% by family with work limits or family that is limited in any way. Finally the third factor was heavily influenced by family on medicare and family members over the age of 65. The benefit of reducing these 19 variables from the survey into 3 main factors will provide future opportunity of regression analysis and how healthcare cost potentially is affected by these three factors within the survey.

The first cluster analysis tested  $k$  with a range from two to ten to find the best  $k$  to do the analysis. From the cluster figure1 (screeplot of Within-groups sum-of-squares), one can determine that setting  $k$  equal to three and five might be good candidates. Since the response variable *FHICOST* has six levels, the final analysis was run on three different  $k$  values - 3, 5, and 6. The visualizations of the cluster analysis were plotted using three significant principal components. The clusters are split clearly, as shown in figures 16 through 18 in the *Visualization Appendix*. If one plotted assigned color to the points by the levels

of *FHICOST*, the points are not easily separated, according to cluster-figure5. Refactoring *FHICOST* was a necessary step and after refactoring, one can finally start to identify patterns within the data.

The kmeans cluster analysis using  $k$  equals to 3 is the final choice. The total withness is 710876.3 and the betweenness is 225214.7.

Cluster 1 can be named as small size family with normal health condition and living condition. They have the most proportion of low cost in health and medical care. This kind of family might have fewer people and they are in normal health condition and without too much disabilities. Cluster 2 can be named as big size family with stable source of income and good health condition. This kind of big family need to cover all family member's health cost so it's much more than other clusters. They also have enough health insurance to cover their cost. Cluster 3 can be named as medium size family with some family members in poor health condition or disability. This family has much more problems with living. They might have more old people who cannot work but need to be taken care of. So this kind of family will spend more money in health care than cluster 1. According to the analysis, one can identify important features in classifying different levels of *FHICOST*, which can help continue feature engineering for other models.

The binary logistic regression model was first executed on the test data set, which included 80% of the observations. Observations with missing values and answers including "Refused," "Not Ascertained," or "Don't Know" were removed. Categorical variables were recoded to binary dummy variables k-1 using the Python pandas library. Automatic forward selection method based on Chi-squared values produced a model with over 50 variables, so a different approach was needed. Over several model iterations, non-significant variables with p-values over 0.05 and VIF score over 5 were dropped. At the end 23 of the most significant variables were chosen. The confusion matrix, figure 22 in the *Visualization Appendix*, showed the the model correctly predicted 7,194 observations as false negative, and 8,304 observations

as true positives (figure 23). The final model was produced using the following equation:

$$\begin{aligned} \text{Log}(FHICOST = 1 / FHICOST = 0) = & -0.15 * FHIMILCT + 0.19 * FHIPRVCT - \\ & 0.57 * FHIIHSCT - 0.16 * FHICADCT - 0.34 * FM\_STRCP\_12 + 0.40 * \\ & FM\_STRCP\_21 + 0.17 * FM\_STRP\_41 - 0.22 * CURWRKN\_2 - 0.16 * TELCELN\_2 - \\ & 0.45 * FHCPHRYN\_2 - 0.57 * FHOSP2YN\_2 + 0.18 * FSBALANC\_3 - 0.50 * \\ & HOUSEOWN\_2 - 0.30 * HOUSEOWN\_3 + 0.37 * FSNAP\_2 + 0.42 * INCGRP4\_2 + \\ & 0.63 * INCGRP4\_3 + 0.77 * INCGRP4\_4 + 1.1 * INCGRP4\_5 - 0.42 * FINTR1YN\_2 - \\ & 1.08 * FMEDBILL\_2 - 1.14 * FMEDBPAY\_2 - 0.30 * FSAF\_2 \end{aligned}$$

There were not significant correlations between variables (figure 24) and the variance inflation factors (VIF) were under 5.

### 6.0.1 Training Data Set Statistics (figure 25)

Confusion matrix summary:  $TP = 8,232$ ,  $TN = 7,021$ ,  $FP = 3,205$ , and  $FN = 2,726$

$$Accuracy = (8232 + 7021) / (8232 + 7021 + 3205 + 2726) = 15253 / 21184 = 0.72$$

$$Specificity = 7021 / (7021 + 3205) = 7021 / 10226 = 0.68$$

### 6.0.2 Test Data Set Statistics (figure 26)

$$Accuracy = 0.71$$

$$Specificity = 0.67$$

$$AUC = 0.78$$

The above statistics show that the logistic regression model performed fairly well on the test data set.

A family is more likely to spend \$500 or more in medical/dental cost per year when:

- When total combined family income is more than \$35,000 per year
- When a family member received income from interest bearing account
- When a family always was able to afford to eat balanced meal
- When a family member did not receive food stamps
- When the number of family members with private insurance increased
- When family was a married couple
- When family includes parent, step parent and children

A family is less likely to spend \$500 or more in medical/dental cost per year?

- When there is no working phone or cell phone in household
- When of family members in Military or Medicaid decreased
- When a family rents a house, an apartment or is living with roommates
- When a family member did not get medical advice or test in 2 weeks or stayed in hospital in 12 months,
- When a family does not pay medical bills over the time
- When family does not have flexible spending account

Overall, the dataset is quite wide with 127 variables, but the project group only focused on *FHICOST* (Cost of family medical/dental care in the past 12 months) since this variable can be used for many medical related businesses. Medical services companies might want to know what kind of families spend a lot of money in medical care. The developed LDA and logistic regression models could help said companies target these kind of families more efficiently so they could develop business arrangements with them to increase their profits. For insurance companies, they could also utilize the developed cluster analysis and LDA to design corresponding insurance plans for different kinds of families to increase their profits. When they collect information from customers, they can easily see whether the customers will spend lots of money in medical care and ask the insurance company to reimburse these payments.

---

## 7 Conclusion

In this paper, the project group analyzed the family data set for the National Health Survey Data with the goals of identifying factors that contribute to healthcare costs for families and determining significant predictors of medical cost ranges for families. Using various statistical modeling techniques, we determined that logistic regression yielded the strongest model for predicting the health care cost range for survey respondents. The logistic regression model showed that less fortunate individuals who earn less than \$35,000 per year, have Medicaid or food stamps, have private insurance, who do not have a landline or cell phone, who rent, or live with roommates who have substandard nutrition habits, who are not married, have problems with paying medical bill make up the population who spend less than \$500 for Medical/Dental cost in a year. It was also determined that one could accurately predict a respondent's health care cost range with a subset of the variables included in the survey with a large enough sample size. Additional statistical models were used for examining the data set. After executing correspondence analysis and linear discriminant analysis, the project group determined that some of the variables exhibited significant correlation with other variables. While this was an insightful and helpful step in our project, it did not provide a model with strong results. Linear discriminant analysis yielded moderate accuracy, but the project group recommends that future studies include additional data sets in order to bring other variables to enrich the model and achieve higher accuracy. Cluster analysis was used to define families having different levels of *FHICOST*. Families with more members having health insurance coverage, more members in good health and more members working tend to spend more money on medical care, while families with more members having disabilities and working limitations tend to spend less. Future studies could improve on the analyses discussed in this paper by combining new data sets from other sources which could be compared with

the National Health Survey Data in order to help determine significant factors and significant predictors of annual health care costs for families.

---



## 8 Appendix

Please find the code used to complete the analysis discussed throughout this paper below:

### 8.1 Cluster Analysis

```
## read data

setwd("~/Desktop/MSPA/advanced data analysis/final/data")

library(tidyverse)
library(corrplot)
library(plyr)
library(plotly)
library(ggplot2)
library("RColorBrewer")

data <- read.csv("family_modified_001.csv")
data <- data[-which(data$FHICOST > 5), ]

## Select numeric variables
NewSub <- data[, c("FM_SIZE", "FCHLMCT", "FSPEDCT",
  "FLAADLCT", "FLIADLCT", "FWKLIMCT", "FWALKCT",
  "FREMECT", "FANYLCT", "FHSTATEX", "FHSTATVG",
  "FHSTATG", "FHSTATFR", "FHSTATPR", "FHICOVCT",
  "FHIPRVCT", "FHIEXCT", "FHISINCT", "FHICARCT",
  "FHICADCT", "FHICHPCT", "FHIMILCT", "FHIIHSCT",
  "FHIPUBCT", "FHIOGVCT", "FHIEBCCT", "FHDSTCT",
  "FDGLWCT1", "FDGLWCT2", "FWRKLWCT", "FSALCT", "FSEINCCT",
```

```

"FSSRRCT", "FPENSCT", "FOPENSCT", "FSSICT", "FTANFCT",
"FOWBENCT", "FINTR1CT", "FDIVDCT", "FCHSPCT", "FINCOTCT",
"FSSAPLCT", "FSDAPLCT", "FWICCT", "FM_ELDR"]

## convert to Matrix
DataMatrix <- as.matrix(as.data.frame(NewSub))

## Remove Variables with missing values coded as -1
NewSub1 <- NewSub[-c(2, 27, 3, 45, 30)]
DataMatrix <- as.matrix(as.data.frame(NewSub1))

## correlation plot
CorData <- cor(DataMatrix)
corrplot(CorData, method = "circle", tl.cex = 0.5)

## Drop Variables with correlation <.3 Final PCA Set
PCASet <- NewSub1[-c(match(c("FHIEXCT", "FHIMILCT",
    "FHIHSCT", "FHIPUBCT", "FHIOGVCT", "FDGLWCT2",
    "FSEINCCT", "FTANFCT", "FOWBENCT", "FINTR1CT",
    "FCHSPCT", "FINCOTCT"), names(NewSub1)))]
DataMatrix <- as.matrix(as.data.frame(PCASet))

## corrplot again
CorData <- cor(DataMatrix)
corrplot(CorData, method = "circle", tl.cex = 0.5)

## Add FHICOST;Normalize data

```

```

std <- apply(PCASet[, -30], 2, sd) # finding standard deviations of variables
PCASet.std <- sweep(PCASet[, -30], 2, std, FUN = "/")

my.k.choices <- 3:10
n <- length(PCASet.std[, 1])
wss1 <- (n - 1) * sum(apply(PCASet.std, 2, var))
wss <- numeric(0)
for (i in my.k.choices)
{
  W <- sum(kmeans(PCASet.std, i)$withinss)
  wss <- c(wss, W)
}
wss <- c(wss1, wss)
plot(c(1, my.k.choices), wss, type = "l", xlab = "Number of clusters",
      ylab = "Within-groups sum-of-squares", lwd = 2)

## FHICOST has 6 levels so try k = 6 first
PCASet.k6 <- kmeans(PCASet.std, centers = 6, iter.max = 500,
  nstart = 25)

## According to the plot, we can choose k = 3,5
PCASet.k3 <- kmeans(PCASet.std, centers = 3, iter.max = 500,
  nstart = 25)

## Also want to try k = 5
PCASet.k5 <- kmeans(PCASet.std, centers = 5, iter.max = 500,
  nstart = 25)

```

```

## Run pca
health.pc <- princomp(PCASet.std, cor = T)

## 3d plot of k = 3,5,6
colors <- brewer.pal(n = 6, name = "Set2")
pc <- data_frame(health.pc$scores[, 1], health.pc$scores[,
  2], health.pc$scores[, 3], PCASet.k6$cluster)
colnames(pc) <- c("pc1", "pc2", "pc3", "cluster")
p <- plot_ly(pc,
  x = ~pc1,
  y = ~pc2,
  z = ~pc3,
  color = ~cluster,
  colors = colors) %>%
  add_markers() %>%
  layout(scene = list(xaxis = list(title = "pc1"),
    yaxis = list(title = "pc2"),
    zaxis = list(title = "pc3")))
p

colors <- brewer.pal(n = 5, name = "Set2")
pc <- data_frame(health.pc$scores[, 1], health.pc$scores[,
  2], health.pc$scores[, 3], PCASet.k5$cluster)
colnames(pc) <- c("pc1", "pc2", "pc3", "cluster")
p <- plot_ly(pc, x = ~pc1, y = ~pc2, z = ~pc3, color = ~cluster,
  colors = colors) %>% add_markers() %>%

```

```

    layout(scene = list(xaxis = list(title = "pc1"),
      yaxis = list(title = "pc2"), zaxis = list(title = "pc3")))
p

colors <- brewer.pal(n = 3, name = "Set2")
pc <- data_frame(health.pc$scores[, 1], health.pc$scores[,
  2], health.pc$scores[, 3], PCASet.k3$cluster)
colnames(pc) <- c("pc1", "pc2", "pc3", "cluster")
p <- plot_ly(pc, x = ~pc1, y = ~pc2, z = ~pc3, color = ~cluster,
  colors = colors) %>% add_markers() %>%
  layout(scene = list(xaxis = list(title = "pc1"),
    yaxis = list(title = "pc2"), zaxis = list(title = "pc3")))
p

## Prepare data for checking the percentage of
## differenrent levels of FHICOST
dt6 <- data_frame(as.numeric(data$FHICOST), PCASet.k6$cluster)
colnames(dt6) <- c("FHICOST", "cluster")
dt5 <- data_frame(as.numeric(data$FHICOST), PCASet.k5$cluster)
colnames(dt5) <- c("FHICOST", "cluster")
dt3 <- data_frame(as.numeric(data$FHICOST), PCASet.k3$cluster)
colnames(dt3) <- c("FHICOST", "cluster")

## FHICOST percentage in each cluster
## tapply(dt6$FHICOST, dt6$cluster, table)
dt6 %>% group_by(cluster) %>% table() %>% prop.table(.,
  1)

```

```

dt6 %>% group_by(cluster) %>% table() %>% prop.table(.,
  2)

# tapply(dt5$FHICOST, dt5$cluster, table)
dt5 %>% group_by(cluster) %>% table() %>% prop.table(.,
  1)
dt5 %>% group_by(cluster) %>% table() %>% prop.table(.,
  2)

dt3 %>% group_by(cluster) %>% table() %>% prop.table(.,
  1)
dt3 %>% group_by(cluster) %>% table() %>% prop.table(.,
  2)

## 3d plot of FHICOST using 3 principal components
colors <- brewer.pal(n = 6, name = "Set2")
pc <- data_frame(health.pc$scores[, 1], health.pc$scores[,
  2], health.pc$scores[, 3], data$FHICOST)
colnames(pc) <- c("pc1", "pc2", "pc3", "FHICOST")
p <- plot_ly(pc, x = ~pc1, y = ~pc2, z = ~pc3, color = ~FHICOST,
  colors = colors) %>% add_markers() %>%
  layout(scene = list(xaxis = list(title = "pc1"),
    yaxis = list(title = "pc2"), zaxis = list(title = "pc3")))
p

## rescale by the size of FHICOST level
colors <- brewer.pal(n = 6, name = "Set2")

```

```

dt <- as.numeric(data$FHICOST) + 1
pc <- data_frame(health.pc$scores[, 1], health.pc$scores[,
  2], health.pc$scores[, 3], dt)
colnames(pc) <- c("pc1", "pc2", "pc3", "FHICOST")
p <- plot_ly(pc, x = ~pc1, y = ~pc2, z = ~pc3, color = ~FHICOST,
  size = ~FHICOST, colors = colors) %>% add_markers() %>%
  layout(scene = list(xaxis = list(title = "pc1"),
    yaxis = list(title = "pc2"), zaxis = list(title = "pc3")))
p

## refactor FHICOST from six levels to three levels
data <- within(data, {
  y3 <- NA
  y3[FHICOST <= 1] <- "low"
  y3[FHICOST <= 3 & FHICOST > 1] <- "Middle"
  y3[FHICOST > 3] <- "high"
})

colors <- brewer.pal(n = 3, name = "Set2")
pc <- data_frame(health.pc$scores[, 1], health.pc$scores[,
  2], health.pc$scores[, 3], data$y3)
colnames(pc) <- c("pc1", "pc2", "pc3", "FHICOST")
p <- plot_ly(pc, x = ~pc1, y = ~pc2, z = ~pc3, color = ~FHICOST,
  colors = colors) %>% add_markers() %>%
  layout(scene = list(xaxis = list(title = "pc1"),
    yaxis = list(title = "pc2"), zaxis = list(title = "pc3")))
p

```

```

dt6 <- data_frame(data$y3, PCASet.k6$cluster)
colnames(dt6) <- c("FHICOST", "cluster")
dt5 <- data_frame(data$y3, PCASet.k5$cluster)
colnames(dt5) <- c("FHICOST", "cluster")
dt3 <- data_frame(data$y3, PCASet.k3$cluster)
colnames(dt3) <- c("FHICOST", "cluster")

## new FHICOST percentage in each cluster
dt6 %>% group_by(FHICOST) %>% table() %>% prop.table(.,
  1)
dt6 %>% group_by(FHICOST) %>% table() %>% prop.table(.,
  2)

dt5 %>% group_by(FHICOST) %>% table() %>% prop.table(.,
  1)
dt5 %>% group_by(FHICOST) %>% table() %>% prop.table(.,
  2)

dt3 %>% group_by(FHICOST) %>% table() %>% prop.table(.,
  1)
dt3 %>% group_by(FHICOST) %>% table() %>% prop.table(.,
  2)

# According to cluster analysis above, it is quite
# difficult to cluster different FHICOST very
# clearly. After refactoring FHICOST, one cluster

```



```

# has higher FHICOST which means the people in this
# cluster spend more money in healthcare. The rest
# clusters are more mixed with low FHICOST.

## refactor based on Michal's code
data <- within(data, {
  y4 <- NA
  y4[FHICOST <= 1] <- "low"
  y4[FHICOST > 1] <- "high"
})

colors <- brewer.pal(n = 2, name = "Set2")
pc <- data_frame(health.pc$scores[, 1], health.pc$scores[,
  2], health.pc$scores[, 3], data$y4)
colnames(pc) <- c("pc1", "pc2", "pc3", "FHICOST")
p <- plot_ly(pc, x = ~pc1, y = ~pc2, z = ~pc3, color = ~FHICOST,
  colors = colors) %>% add_markers() %>%
  layout(scene = list(xaxis = list(title = "pc1"),
    yaxis = list(title = "pc2"), zaxis = list(title = "pc3")))
p

## Prepare data
dt6 <- data_frame(data$y4, PCASet.k6$cluster)
colnames(dt6) <- c("FHICOST", "cluster")
dt5 <- data_frame(data$y4, PCASet.k5$cluster)
colnames(dt5) <- c("FHICOST", "cluster")
dt3 <- data_frame(data$y4, PCASet.k3$cluster)

```

```

colnames(dt3) <- c("FHICOST", "cluster")

dt6 %>% group_by(FHICOST) %>% table() %>% prop.table(.,
  1)
dt6 %>% group_by(FHICOST) %>% table() %>% prop.table(.,
  2)

dt5 %>% group_by(FHICOST) %>% table() %>% prop.table(.,
  1)
dt5 %>% group_by(FHICOST) %>% table() %>% prop.table(.,
  2)

table(data$y4)
dt3 %>% group_by(FHICOST) %>% table() %>% prop.table(.,
  1)
dt3 %>% group_by(FHICOST) %>% table() %>% prop.table(.,
  2)

# If we refacor FHICOST again to just tow category,
# then it is clear that one cluster has high FHICOST
# or low FHICOST dominating inside them. So this
# can help some companies to target those people
# who pay more in healthcare.

PCASet.k3$centers

# Withiness
PCASet.k3["tot.withinss"]

```

```

# Betweenness
PCASet.k3["betweenss"]

# If we look at the centers of three clusters by
# kmeans using k = 3.

# In the cluster that has highest percentage of
# high FHICOST, variable FHSTATEX(# fam mem in
# excellent health) and FHSTATVG(# fam mem in very
# good health) is higher then other clusters, which
# means families in this cluster are willing to pay
# more money to keep the family healthy.

# In cluster that has highest percetage of low
# FHICOST, it has much more family in pooe health
# but there are more people not pay much in health
# care. In this cluster, there are more people
# receiving money from sources other than wages and
# salary. It has much higher FSSICT(# fam mem
# receive income from SSI (Supplemental Security
# Income)) and FSSRRCT(# fam mem receive income
# from Social Security or Railroad retirement
# inc),FSSAPLCT(# fam mem ever apply for SSI
# (Supplemental Security Income)), FSDAPLCT(# fam
# mem ever apply for SSDI (Social Security
# Disability)) which indicates they try to find
# other sources to help cover their healthcare

```

```
# cost. These families might have more difficulty  
# living than other families which can also be  
# found in higher FPENSCT(# fam mem receive  
# disability pensions), lower FDGLWCT1(# of fam mem  
# working last week) and higher FM_ELDR(# of family  
# members over 65).
```

## 8.2 Factor Analysis

```
# libraries  
  
library(foreign)  
library(corrplot)  
library(car)  
library(QuantPsyc)  
library(leaps)  
library(RColorBrewer)  
library(Hmisc)  
library(psych)  
library(dplyr)  
  
# Set the working directory  
setwd("C:/Users/User/Documents/DSC 424/Final Dataset")  
  
# import final file#  
a1 <- read.csv("family_modified_001.csv")
```

```

# read first files
head(a1, 10)

# create obs based upon values from 18 to 127
b1 <- a1[, c(7, 18:127)]

# reduce variables
z1 <- a1[, c(7, 18, 19, 21, 23, 25, 27, 29, 31, 33,
           39, 44:48, 57:69, 73, 93:119, 124, 127)]

pc.z1 <- prcomp(z1, scale = T)
pc.z1

screeplot(pc.z1, npcs = 58, type = "barplot", main = "Scree Plot")
title(ylab = "Variances", xlab = "PCA Number")
abline(h = 1, lwd = 4, col = "black")
summary(pc.z1)

xy1 <- psych::principal(z1, rotate = "varimax", nfactors = 8,
                        score = TRUE)
print(xy1$loadings, sort = T)
print(xy1$loadings, cutoff = 0.559, sort = T)
print(xy1$loadings, cutoff = 0.3, sort = T)
print(xy1$loadings, cutoff = 0.5, sort = T)
print(xy1$loadings, cutoff = 0.54, sort = T)

xy2 <- psych::principal(z1, rotate = "varimax", nfactors = 16,

```

```

    score = TRUE)
print(xy2$loadings, sort = T)
print(xy2$loadings, cutoff = 0.559, sort = T)
print(xy2$loadings, cutoff = 0.3, sort = T)
print(xy2$loadings, cutoff = 0.5, sort = T)

# Remove non good variables
z1$FHIPUBCT <- NULL
z1$FHIOGVCT <- NULL
z1$FHIIHSCT <- NULL
z1$FINCOTCT <- NULL
z1$FHICHPCT <- NULL
z1$FPENSCT <- NULL
z1$FTANFCT <- NULL
z1$FOWBENCT <- NULL
z1$FDGLWCT2 <- NULL
z1$FHSTATVG <- NULL
z1$FHCDVCT <- NULL
z1$FHCPHRCT <- NULL

# run some more
xy1 <- psych::principal(z1, rotate = "varimax", nfactors = 8,
    score = TRUE)
print(xy1$loadings, sort = T)
print(xy1$loadings, cutoff = 0.559, sort = T)
print(xy1$loadings, cutoff = 0.3, sort = T)
print(xy1$loadings, cutoff = 0.5, sort = T)

```

```

print(xy1$loadings, cutoff = 0.54, sort = T)

# Remove additional variables
z1$FINTR1CT <- NULL
z1$F10DVCT <- NULL
z1$FHOSP2CT <- NULL
z1$FHSTATG <- NULL
z1$FHSTATFR <- NULL
z1$FSNAPMYR <- NULL
z1$FCHSPCT <- NULL
z1$FHIMILCT <- NULL
z1$FHIEXCT <- NULL

# run some more
xy1 <- psych::principal(z1, rotate = "varimax", nfactors = 8,
  score = TRUE)
print(xy1$loadings, sort = T)
print(xy1$loadings, cutoff = 0.559, sort = T)
print(xy1$loadings, cutoff = 0.3, sort = T)
print(xy1$loadings, cutoff = 0.5, sort = T)
print(xy1$loadings, cutoff = 0.54, sort = T)
print(xy1$loadings, cutoff = 0.351, sort = T)

# run some more
xy3 <- psych::principal(z1, rotate = "varimax", nfactors = 5,
  score = TRUE)
print(xy3$loadings, sort = T)

```

```

print(xy3$loadings, cutoff = 0.559, sort = T)
print(xy3$loadings, cutoff = 0.3, sort = T)

# drop some more variables
z1$FSSKDAY5 <- NULL
z1$FSSKDAY6 <- NULL
z1$FSSKDAY7 <- NULL
z1$FSSKDAY8 <- NULL

# run some more
xy1 <- psych::principal(z1, rotate = "varimax", nfactors = 8,
  score = TRUE)
print(xy1$loadings, sort = T)
print(xy1$loadings, cutoff = 0.559, sort = T)
print(xy1$loadings, cutoff = 0.3, sort = T)
print(xy1$loadings, cutoff = 0.5, sort = T)

xy3 <- psych::principal(z1, rotate = "varimax", nfactors = 5,
  score = TRUE)
print(xy3$loadings, sort = T)
print(xy3$loadings, cutoff = 0.559, sort = T)
print(xy3$loadings, cutoff = 0.3, sort = T)

# Factoral Analysis

fz1 <- factanal(z1, 8)
print(fz1$loadings, cutoff = 0.4, sort = T)

```



```

fz2 <- factanal(z1, 5)
print(fz2$loadings, cutoff = 0.4, sort = T)

# drop additional variables
z1$FSSICT <- NULL
z1$FDIVDCT <- NULL
z1$FHCHMCT <- NULL
z1$FSEINCCT <- NULL
z1$FHCHMCT <- NULL

# factor analysis
fz3 <- factanal(z1, 5)
print(fz3$loadings, cutoff = 0.44, sort = T)

# drop cost variables
z1$FHICOST <- NULL

# run factor analysis
fz4 <- factanal(z1, 5)
print(fz4$loadings, cutoff = 0.44, sort = T)

```

## 8.3 Logistic Regression

### 8.3.1 Preprocessing

```
# import libraries

library(foreign)
library(corrplot)
library(car)
library(QuantPsyc)
library(leaps)
library(RColorBrewer)
library(Hmisc)
library(psych)
library(ggplot2)
library(MASS)
library(reshape2)
library(RCurl)
library(tidyverse)
library(plyr)
library(caTools)

# family_data <-
# getURL('https://raw.githubusercontent.com/stfox13/
#       DSC424FinalProject/master/family_modified_001.csv')
# family_df <- read.csv(text = family_data) View(family_df)
# Set the working directory
setwd("C:/Depaul_Win7/DSC 424 Advanced Data Analysis/Project/family")

# Read data

fam <- read.csv("familyxx.csv", sep = ",", header = T)
```

```

dim(fam)
str(fam)
head(fam)

# FHICOST Cost of family medical/dental care in the past 12 months
# put FHICOST in first column
fam <- fam[, c(119, 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, 32,
              33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49,
              50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66,
              67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83,
              84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100,
              101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114,
              115, 116, 117, 118, 120, 121, 122, 123, 124, 125, 126, 127)]

str(fam)

# clean NAs check na after cleaning
sum(is.na(fam))

# #recode missing values -1 to NA fam[fam==-1] <-NA
# recode missing values 99 to NA
fam[fam == 99] <- NA

# recode missing values 98 to NA
fam[fam == 98] <- NA

# recode missing values 97 to NA
fam[fam == 97] <- NA

# recode missing values 96 to NA
fam[fam == 96] <- NA

# recode missing values 7 to NA
fam[fam == 7] <- NA

```

```

# recode missing values 8 to NA
fam[fam == 8] <- NA

# recode missing values 9 to NA
fam[fam == 9] <- NA

# check for missing values
sum(is.na(fam))

# check which columns have the most NAs
map(fam, ~sum(is.na(.)))

# drop variables which have too many NAs sum(is.na(fam$WRKCELN))

fam$WRKCELN <- NULL #
fam$PHONEUSE <- NULL #
fam$FM_EDUC1 <- NULL #
fam$FSSKIP <- NULL #
fam$FSSKDAY <- NULL #
fam$FSLESS <- NULL #
fam$FSHUNGRY <- NULL #
fam$FSWEIGHT <- NULL #
fam$FSNOTEAT <- NULL #
fam$FSNEDAYS <- NULL #
fam$FHDSTCT <- NULL #
fam$FWRKLWCT <- NULL #
fam$FCHLMYN <- NULL #
fam$FSPEDYN <- NULL #
fam$FCHLMCT <- NULL #
fam$FSPEDCT <- NULL #
fam$FGAH <- NULL #
fam$FSNAPMYR <- NULL #

```

```

fam$RAT_CAT4 <- NULL #
fam$RAT_CAT5 <- NULL #
fam$FWICYN <- NULL #
fam$FWICCT <- NULL #
fam$COVCONF <- NULL #
fam$FMEDBNOP <- NULL #
fam$FPRCOOH <- NULL #
fam$FHIEBCCT <- NULL #
# drop not important variables
fam$RECTYPE <- NULL #
fam$WTFA_FAM <- NULL #
fam$FINT_Y_P <- NULL #
fam$SRVY_YR <- NULL #
fam$FMX <- NULL #
fam$HHX <- NULL #
fam$FINT_M_P <- NULL #
fam$FINT_Y_P <- NULL #
map(fam, ~sum(is.na(.)))
# check NA by rows
fam[!complete.cases(fam), ]
# remove rows which include NAs
famcleaned <- na.omit(fam)
head(famcleaned)
dim(fam)
dim(famcleaned)
# check na after cleaning
sum(is.na(famcleaned))

```

```

# no NAs, 19800 rows, 101 columns
# not recognized as categorical variable
is.factor(famcleaned$FHICOST)
# no dummy variables contrasts(famcleaned$FHICOST)
# bin FHICOST to low and high to use it in logistic regression low=0 -
# less then $500 per year, high=1 - more than $500 per year 0 and 1
# will be 0 2, 3, 4, 5, will be 1
famcleaned$FHICOST[famcleaned$FHICOST == 1] <- 0
famcleaned$FHICOST[famcleaned$FHICOST == 0] <- 0
famcleaned$FHICOST[famcleaned$FHICOST == 2] <- 1
famcleaned$FHICOST[famcleaned$FHICOST == 3] <- 1
famcleaned$FHICOST[famcleaned$FHICOST == 4] <- 1
famcleaned$FHICOST[famcleaned$FHICOST == 5] <- 1
head(famcleaned)
plot(famcleaned$FHICOST)
hist(famcleaned$FHICOST, col = "green")
sum(is.na(famcleaned))

# #create dataset with categorical variables only drop
# <-c('FM_SIZE', 'FM_KIDS', 'FM_ELDR', 'F10DVCT',
# 'FDMEDCT', 'FHCDVCT', 'FHCHMCT', 'FHCPHRCT',
# 'FHOSP2CT', 'FNMEDCT', 'FDGLWCT1', 'FDGLWCT2', 'FLIADLCT', 'FWKLIMCT',
# 'FWALKCT', 'FREMEMCT', 'FANYLCT', 'FHSTATEX',
# 'FHSTATVG', 'FHSTATG', 'FHSTATFR', 'FHSTATPR',
# 'FLAADLCT', 'FSALCT', 'FSEINCCT',
# 'FSSRRCT', 'FPENSCT', 'FOPENSCT', 'FSSICT',
# 'FTANFCT', 'FOWBENCT', 'FINTR1CT', 'FDIVDCT',
# 'FCHSPCT', 'FINCOTCT', 'FSSAPLCT', 'FSDAPLCT',

```

```

# 'FHIPRVCT', 'FHISINCT', 'FHICARCT',
# 'FHICADCT', 'FHICHPCT', 'FHIMILCT', 'FHIPUBCT',
# 'FHIOGVCT', 'FHIIHSCT', 'FHIEXCT', 'FHICOVCT')
# fam_cat=famcleaned[,!(names(famcleaned)%in%drop)]
head(fam_cat)
dim(fam_cat)
# 46 variables left, 19800 rows
famnum <- famcleaned[c("FHICOST", "F10DVCT", "FHCPHRCT", "FDGLWCT2", "FWKLIMCT",
  "FHSTATEX", "FHSTATPR", "FSEINCCT", "FSSICT", "FDIVDCT", "FSDAPLCT",
  "FHICARCT", "FHIPUBCT", "FHICOVCT", "FM_ELDR", "FHCHMCT", "FDGLWCT1",
  "FLIADLCT", "FANYLCT", "FHSTATFR", "FSALCT", "FOPENSCT", "FINTR1CT",
  "FSSAPLCT", "FHISINCT", "FHIMILCT", "FHIEXCT", "FM_KIDS", "FHCDVCT",
  "FNMEDCT", "FREMECT", "FHSTATG", "FLAADLCT", "FPENSCT", "FOWBENCT",
  "FINCOTCT", "FHIPRVCT", "FHICHPCT", "FHIIHSCT", "FM_SIZE", "FDMEDCT",
  "FHOSP2CT", "FWALKCT", "FHSTATVG", "FSSRRCT", "FTANFCT", "FCHSPCT",
  "FHICADCT", "FHIOGVCT"))]
head(famnum)
# save famnum to csv
write.csv(famnum, "famnum.csv")
# categorical
famcat <- famcleaned[c("FM_STRCP", "FM_TYPE", "FM_STRP", "TELN_FLG", "CURWRKN",
  "TELCELN", "FLNGINTV", "F10DVYN", "FDMEDYN", "FHCDVYN", "FHCHMYN",
  "FHCPHRYN", "FHOSP2YN", "FNMEDYN", "FSRUNOUT", "FSLAST", "FSBALANC",
  "FLAADLYN", "FLIADLYN", "FWKLIMYN", "FWALKYN", "FREMEMYN", "FANYLYN",
  "HOUSEOWN", "FSNAP", "INCGRP4", "INCGRP5", "FSALYN", "FSEINCYN", "FSSRRYN",
  "FPENSYN", "FOPENSYN", "FSSIYN", "FTANFYN", "FOWBENYN", "FINTR1YN",
  "FDIVDYN", "FCHSPYN", "FINCOTYN", "FSSAPLYN", "FSDAPLYN", "FMEDBILL",

```

```

    "FMEDBPAY", "FSAF", "FHICOVYN")]
```

*# save famcleaned to csv to create dummy variables using Python*

```

write.csv(famcat, "famcat.csv")
```

*#DUMMY VARIABLES - PYTHON, PANDAS*

*# coding: utf-8*

*#CSC424 create dummies*

*## Import Modules*

```

import numpy as np
import pandas as pd
import os
import matplotlib.pyplot as plt
get_ipython().magic(u'matplotlib inline')
```

*##get working directory*

```

os.getcwd()
```

*##load famcleaned2.csv via read\_csv*

```

fam_df = pd.read_csv('C:/Depaul_Win7/DSC 424 Advanced Data Analysis/'
                    'Project/family/famcat.csv', sep=",")
fam_df = fam_df.drop(fam_df.columns[0],axis=1)
fam_df.head()
```

```

fam_cat = fam_df.astype('category')
```



```

fam_cat.dtypes

# create dummy variables k-1

fam_dummy = pd.get_dummies(fam_cat, drop_first=True)
fam_dummy.head()

##load famnum via read_csv

fam_num = pd.read_csv('C:/Depaul_Win7/DSC 424 Advanced Data Analysis/Project/'
                      'family/famnum.csv', sep=",")
fam_num = fam_num.drop(fam_num.columns[0],axis=1)
fam_num.head()

#merge data frames

famreg = pd.concat([fam_num, fam_dummy], axis = 1, join_axes=[fam_num.index])
famreg.head()

#save dummy variables to csv file

famreg.to_csv('fam_all_dummies.csv', sep=",")

# END OF PYTHON CODE

```

### 8.3.2 Model Development

```

#DSC 424 project
#logistic regression
#Michal Chowaniak

```

```
#import libraries
library(foreign)
library(corrplot)
library(car)
library(QuantPsyc)
library(leaps)
library(RColorBrewer)
library(Hmisc)
library(psych)
library(ggplot2)
library(MASS)
library(reshape2)
library(RCurl)
library(tidyverse)
library(plyr)
library(caTools)
library(ROCR)
library(caret)
library(glmnet)
install.packages('e1071', dependencies=TRUE)
library(Amelia)
library(pscl)

# Set the working directory
```

```

setwd("C:/Depaul_Win7/DSC 424 Advanced Data Analysis/Project/family")

#Read data
famreg <- read.csv("fam_all_dummies.csv", sep=",", header=T)

famreg$FHICOST <- as.factor(famreg$FHICOST)
famreg$X = NULL
dim(famreg)
str(famreg)
head(famreg)

# Bar Plot
counts <- table(famreg$FHICOST)
barplot(counts, main="Cost of family medical/
          dental care in the past 12 months low (0) or high (1)",
        xlab="0-less than $500, 1- more than $500")

# per above bar plot Y is balanced

#create a list of names for barplots
names <- c("FHICOST", "F10DVCT", "FHCPHRCT",
           "FDGLWCT2", "FWKLIMCT", "FHSTATEX",
           "FHSTATPR", "FSEINCCT", "FSSICT",
           "FDIVDCT", "FSDAPLCT", "FHICARCT",
           "FHIPUBCT", "FHICOVCT", "FM_ELDR",

```

"FHCHMCT", "FDGLWCT1", "FLIADLCT",  
 "FANYLCT", "FHSTATFR", "FSALCT",  
 "FOPENSCT", "FINTR1CT", "FSSAPLCT",  
 "FHISINCT", "FHIMILCT", "FHIEXCT",  
 "FM\_KIDS", "FHCDVCT", "FNMEDCT",  
 "FREMECT", "FHSTATG", "FLAADLCT",  
 "FPENSCT", "FOWBENCT", "FINCOTCT",  
 "FHIPRVCT", "FHICHPCT", "FHIIHSCT",  
 "FM\_SIZE", "FDMEDCT", "FHOSP2CT",  
 "FWALKCT", "FHSTATVG", "FSSRRCT",  
 "FTANFCT", "FCHSPCT", "FHICADCT",  
 "FHIOGVCT", "FM\_STRCP\_12", "FM\_STRCP\_21",  
 "FM\_STRCP\_22", "FM\_STRCP\_23", "FM\_STRCP\_31",  
 "FM\_STRCP\_32", "FM\_STRCP\_33", "FM\_STRCP\_41",  
 "FM\_STRCP\_42", "FM\_STRCP\_43", "FM\_STRCP\_44",  
 "FM\_STRCP\_45", "FM\_TYPE\_2", "FM\_TYPE\_3",  
 "FM\_TYPE\_4", "FM\_STRP\_12", "FM\_STRP\_21",  
 "FM\_STRP\_22", "FM\_STRP\_23", "FM\_STRP\_31",  
 "FM\_STRP\_32", "FM\_STRP\_33", "FM\_STRP\_41",  
 "FM\_STRP\_42", "FM\_STRP\_43", "FM\_STRP\_44",  
 "FM\_STRP\_45", "CURWRKN\_2", "TELCELN\_2",  
 "FLNGINTV\_2", "FLNGINTV\_3", "FLNGINTV\_4",  
 "F10DVYN\_2", "FDMEDYN\_2", "FHCDVYN\_2",  
 "FHCHMYN\_2", "FHCPHRYN\_2", "FHOSP2YN\_2",  
 "FNMEDYN\_2", "FSRUNOUT\_2", "FSRUNOUT\_3",  
 "FSLAST\_2", "FSLAST\_3", "FSBALANC\_2",  
 "FSBALANC\_3", "FLAADLYN\_2", "FLIADLYN\_2",

```

"FWKLIMYN_2", "FWALKYN_2", "FREMEMYN_2",
"FANYLYN_2", "HOUSEOWN_2", "HOUSEOWN_3",
"FSNAP_2", "INCGRP4_2", "INCGRP4_3",
"INCGRP4_4", "INCGRP4_5", "INCGRP5_2",
"INCGRP5_3", "INCGRP5_4", "FSALYN_2",
"FSEINCYN_2", "FSSRRYN_2", "FPENSYN_2",
"FOPENSYN_2", "FSSIYN_2", "FTANFYN_2",
"FOWBENYN_2", "FINTR1YN_2", "FDIVDYN_2",
"FCHSPYN_2", "FINCOTYN_2", "FSSAPLYN_2",
"FSDAPLYN_2", "FMEDBILL_2", "FMEDBPAY_2",
"FSAF_2", "FHICOVYN_2")

```

```
names[1]
```

```
#Create barplots for all variables
```

```

for ( i in 1:length(famreg)){
  count <- table(famreg[[i]])
  name <- names[i]
  barplot(count, main = name)
}

```

```
#barplots show x variables are imbalanced
```

```
#check for correlation on numerical variables
```

```
famnum <- famreg[c('F10DVCT', 'FHCPHRCT', 'FDGLWCT2', 'FWKLIMCT', 'FHSTATEX',  
                  'FHSTATPR', 'FSEINCCT', 'FSSICT', 'FDIVDCT', 'FSDAPLCT',  
                  'FHICARCT', 'FHIPUBCT', 'FHICOVCT', 'FM_ELDR', 'FHCHMCT',  
                  'FDGLWCT1', 'FLIADLCT', 'FANYLCT', 'FHSTATFR', 'FSALCT',  
                  'FOPENSCT', 'FINTR1CT', 'FSSAPLCT', 'FHISINCT', 'FHIMILCT',  
                  'FHIEXCT', 'FM_KIDS', 'FHCDVCT', 'FNMEDCT', 'FREMEMCT',  
                  'FHSTATG', 'FLAADLCT', 'FPENSCT', 'FOWBENCT', 'FINCOTCT',  
                  'FHIPRVCT', 'FHICHPCT', 'FHIIHSCT', 'FM_SIZE', 'FDMEDCT',  
                  'FHOSP2CT', 'FWALKCT', 'FHSTATVG', 'FSSRRCT', 'FTANFCT',  
                  'FCHSPCT', 'FHICADCT', 'FHIOGVCT')]
```

```
#check for correlation
```

```
cor.famnum = cor(famnum)  
cor.famnum  
corrplot(cor.famnum, method = "number")  
corrplot(cor.famnum, method = "ellipse")
```

```
#drop correlated variables
```

```
famreg$FM_ELDR = NULL #  
famreg$FANYLCT = NULL #  
famreg$FSALCT = NULL #  
famreg$FHICOVCT = NULL #  
famreg$FSSRRCT = NULL #
```

```

famreg$FM_SIZE = NULL #
famreg$FSSAPLCT = NULL #
famreg$FDMEDCT = NULL #
famreg$FDGLWCT1 = NULL #

#check for correclation on numerical
famnum <- famreg[c('F10DVCT', 'FHCPHRCT', 'FDGLWCT2', 'FWKLIMCT', 'FHSTATEX',
                  'FHSTATPR', 'FSEINCCT', 'FSSICT', 'FDIVDCT', 'FSDAPLCT',
                  'FHICARCT', 'FHIPUBCT', 'FHCHMCT',
                  'FLIADLCT', 'FHSTATFR',
                  'FOPENSCT', 'FINTR1CT', 'FHISINCT', 'FHIMILCT',
                  'FHIEXCT', 'FM_KIDS', 'FHCDVCT', 'FNMEDCT', 'FREMEMCT',
                  'FHSTATG', 'FLAADLCT', 'FPENSCT', 'FOWBENCT', 'FINCOTCT',
                  'FHIPRVCT', 'FHICHPCT', 'FHIIHSCT',
                  'FHOSP2CT', 'FWALKCT', 'FHSTATVG', 'FTANFCT',
                  'FCHSPCT', 'FHICADCT', 'FHIOGVCT'))]

#check for correlation
cor.famnum = cor(famnum)
cor.famnum
corrplot(cor.famnum, method = "ellipse")
corrplot(cor.famnum, method = "number")

```

```

#split data set to training and test
set.seed(314)
split <- sample.split(famreg$FHICOST, SplitRatio = 0.80)
fam_train = subset(famreg, split == TRUE)
fam_test = subset(famreg, split == FALSE)
head(fam_train)
dim(fam_train)
dim(fam_test)

#check for missing
#missmap(fam_train, main = "Missing values vs observed")

#check for categorical status
is.factor(fam_train$FHICOST)

#####

#automatic logistic regression model, forward selection on test

```



```

model.null = glm(FHICOST ~ 1, data=fam_train, family = binomial(link="logit"))

model.full = glm(FHICOST ~ ., data=fam_train, family = binomial(link="logit"))

#step(model.null, scope = list(upper=model.full),
#direction="forward", test="Chisq", data=fam_train)

#result of above
#56 variables
model6 <- glm(formula = FHICOST ~ FHIPRVCT + FMEDBPAY_2 + HOUSEOWN_2 +
              F10DVYN_2 + FINTR1CT + INCGRP4_5 + FMEDBILL_2 + FSNAP_2 +
              FM_TYPE_2 + FHCDVCT + FHICADCT + INCGRP4_4 + INCGRP5_2 +
              FDMEDYN_2 + FHICARCT + FSEINCYN_2 + FSAF_2 + FHOSP2YN_2 +
              FSRUNOUT_3 + FHIIHSCT + FDIVDYN_2 + FSSICT + FANYLYN_2 +
              FHIMILCT + FM_TYPE_4 + INCGRP4_2 + CURWRKN_2 + FHSTATEX +
              FM_KIDS + FHCPHRYN_2 + FM_STRP_42 + HOUSEOWN_3 + FHSTATFR +
              FSBALANC_3 + FM_STRCP_21 + FINTR1YN_2 + F10DVCT + FTANFCT +
              TELCELN_2 + FM_STRCP_12 + FLNGINTV_4 + FHSTATPR + FLIADLCT +
              FM_TYPE_3 + FREMEMCT + FHCHMYN_2 + FM_STRCP_23 + FHIEXCT +
              FHIOGVCT + FDIVDCT + FWKLIMYN_2 + FWKLIMCT + FLIADLYN_2,
              family = binomial(link = "logit"), data = fam_train)

summary(model6) #not all variable under 0.05
vif(model6) #vif under 5

```

```

anova(model6, test="Chisq")

famfit <- predict(model6, type = 'response')

#confusion matrix
table(fam_train$FHICOST, famfit > 0.5)

prop.table(table(fam_train$FHICOST, famfit > 0.5))

tp = 8304
tn = 7193
fp = 3033
fn = 2654

sensitivity = tp/(tp+fn)
sensitivity
accuracy = (tp+tn)/(tp+tn+fp+fn)
accuracy
precision = tp/(tp+fp)
precision
specificity = tn/(tn+fp)
specificity

```

```

# model is too complicated, too many variables

#End of automatic model selection forward on test
#####

#####

#logistic regression model - manual variable selection

model <- glm (FHICOST ~ ., data = fam_train, family = binomial)
summary(model)

predict <- predict(model, type = 'response')

#confusion matrix
table(fam_train$FHICOST, predict > 0.5)

ROCRpred <- prediction(predict, fam_train$FHICOST)

```

```

ROCPerf <- performance(ROCpred, 'tpr','fpr')
plot(ROCPerf, colorize = TRUE, text.adj = c(-0.2,1.7))

#drop non significant variables
fam_train_sig <- fam_train[c('FHICOST','F10DVCT','FHIMILCT','FHIPRVCT','FHIIHSCT',
                             'FHICADCT','FM_STRCP_12','FM_STRCP_21','FM_STRP_41',
                             'CURWRKN_2','TELCELN_2','F10DVYN_2','FHCPHRYN_2',
                             'FHOSP2YN_2','FSRUNOUT_2','FSBALANC_3','HOUSEOWN_2',
                             'HOUSEOWN_3','FSNAP_2','INCGRP4_2','INCGRP4_3',
                             'INCGRP4_4','INCGRP4_5','FSALYN_2','FINTR1YN_2',
                             'FMEDBILL_2','FMEDBPAY_2','FSAF_2')]

model2 <- glm (FHICOST ~ ., data = fam_train_sig, family = binomial)
summary(model2)

#drop more non significant variables
fam_train_sig2 <- fam_train[c('FHICOST','F10DVCT','FHIMILCT','FHIPRVCT','FHIIHSCT',
                              'FHICADCT','FM_STRCP_12','FM_STRCP_21','FM_STRP_41',
                              'CURWRKN_2','TELCELN_2','F10DVYN_2','FHCPHRYN_2',
                              'FHOSP2YN_2','FSBALANC_3','HOUSEOWN_2',
                              'HOUSEOWN_3','FSNAP_2','INCGRP4_2','INCGRP4_3',

```

```

      'INCGRP4_4', 'INCGRP4_5', 'FINTR1YN_2',
      'FMEDBILL_2', 'FMEDBPAY_2', 'FSAF_2'])

model3 <- glm (FHICOST ~ ., data = fam_train_sig2, family = binomial)
summary(model3)

vif(model3) #vif normal, under 5

#drop variables vif over 5, F10DVCT, F10DVYN_2
fam_train_sig3 <- fam_train[c('FHICOST', 'FHIMILCT', 'FHIPRVCT', 'FHIHSCT',
                              'FHICADCT', 'FM_STRCP_12', 'FM_STRCP_21', 'FM_STRP_41',
                              'CURWRKN_2', 'TELCELN_2', 'FHCPHRYN_2',
                              'FHOSP2YN_2', 'FSBALANC_3', 'HOUSEOWN_2',
                              'HOUSEOWN_3', 'FSNAP_2', 'INCGRP4_2', 'INCGRP4_3',
                              'INCGRP4_4', 'INCGRP4_5', 'FINTR1YN_2',
                              'FMEDBILL_2', 'FMEDBPAY_2', 'FSAF_2'])

model4 <- glm (FHICOST ~ ., data = fam_train_sig3, family = binomial)
summary(model4) #all variable under 0.05
vif(model4) #vif under 5

predict2 <- predict(model4, type = 'response')

```

```

#confusion matrix
table(fam_train_sig3$FHICOST, predict2 > 0.5)

ROCRpred <- prediction(predict2, fam_train_sig3$FHICOST)
ROCRperf <- performance(ROCRpred, 'tpr','fpr')
plot(ROCRperf, colorize = TRUE, text.adj = c(-0.2,1.7))

#####

#run model on test data

fam_test2 <- fam_test[c('FHICOST','FHIMILCT','FHIPRVCT','FHIHSCT',
                        'FHICADCT','FM_STRCP_12','FM_STRCP_21','FM_STRP_41',
                        'CURWRKN_2','TELCELN_2','FHCPHRYN_2',
                        'FHOSP2YN_2','FSBALANC_3','HOUSEOWN_2',
                        'HOUSEOWN_3','FSNAP_2','INCGRP4_2','INCGRP4_3',
                        'INCGRP4_4','INCGRP4_5','FINTR1YN_2',
                        'FMEDBILL_2','FMEDBPAY_2','FSAF_2'))]

model5 <- glm (FHICOST ~ ., data = fam_test2, family = binomial)

```

```

summary(model5) #not all variable under 0.05
vif(model5) #vif under 5

anova(model5, test="Chisq")

pR2(model5)

famfit <- predict(model5, type = 'response')

#confusion matrix
table(fam_test2$FHICOST, famfit > 0.5)

prop.table(table(fam_test2$FHICOST, famfit > 0.5))

tp = 2051
tn = 1729
fp = 828
fn = 689

sensitivity = tp/(tp+fn)
sensitivity
accuracy = (tp+tn)/(tp+tn+fp+fn)

```

```

accuracy
precision = tp/(tp+fp)
precision
specificity = tn/(tn+fp)
specificity

ROCRpred <- prediction(famfit, fam_test2$FHICOST)
ROCRperf <- performance(ROCRpred, 'tpr','fpr')
plot(ROCRperf, colorize = TRUE, text.adj = c(-0.2,1.7))

# AUC- the probability that the model will rank a randomly chosen positive example
#higher than a randomly chosen negative example
auc <- performance(ROCRpred, measure = "auc")
auc <- auc@y.values[[1]]
auc

#the end
#####

```

## 8.4 Principle Component Analysis

```

setwd("C:/Users/izl7729/Desktop/DePaul/Project")
getwd()
data <- read.csv("family_modified_001.csv")
library(tidyverse)

```



```

library(corrplot)
library(plyr)
library(ggplot2)
library(psych)

# Subset continuous variables
NewSub <- data[, c("FM_SIZE", "FCHLMCT", "FSPEDCT", "FLAADLCT", "FLIADLCT",
  "FWKLIMCT", "FWALKCT", "FREMEMCT", "FANYLCT", "FHSTATEX", "FHSTATVG",
  "FHSTATG", "FHSTATFR", "FHSTATPR", "FHICOVCT", "FHIPRVCT", "FHIEXCT",
  "FHISINCT", "FHICARCT", "FHICADCT", "FHICHPCT", "FHIMILCT", "FHIIHSCT",
  "FHIPUBCT", "FHIOGVCT", "FHIEBCCT", "FHDSTCT", "FDGLWCT1", "FDGLWCT2",
  "FWRKLWCT", "FSALCT", "FSEINCCT", "FSSRRCT", "FPENSCT", "FOPENSCT",
  "FSSICT", "FTANFCT", "FOWBENCT", "FINTR1CT", "FDIVDCT", "FCHSPCT",
  "FINCOTCT", "FSSAPLCT", "FSDAPLCT", "FWICCT", "FM_ELDLCT")]

# convert to Matrix
DataMatrix <- as.matrix(as.data.frame(NewSub))

# Reveiw Missing data
colSums(is.na(DataMatrix))

# Histograms
ggplot(gather(NewSub), aes(value)) + geom_histogram(bins = 20) + facet_wrap(~key,
  scales = "free_x")

# Correlation Matrices
CorData <- cor(DataMatrix)
corrplot(CorData, method = "circle")

# Remove Variables with missing values coded as -1
NewSub1 <- NewSub[-c(2, 27, 3, 45, 30)]

DataMatrix <- as.matrix(as.data.frame(NewSub1))

# Drop Variables with correlation <.3 Final PCA Set

```

```

PCASet <- NewSub1[-c(match(c("FHIEXCT", "FHIMILCT", "FHIHSCT", "FHIPUBCT",
    "FHIQVCT", "FDGLWCT2", "FSEINCCT", "FTANFCT", "FOWBENCT", "FINTR1CT",
    "FCHSPCT", "FINCOTCT"), names(NewSub1)))]
DataMatrix <- as.matrix(as.data.frame(PCASet))
describe(DataMatrix)
options(scipen = 100, digits = 5)
round(cor(DataMatrix), 2)
MCorrTest <- corr.test(DataMatrix, adjust = "none")
MCorrTest
M <- MCorrTest$p
M
MTest <- ifelse(M < 0.01, T, F)
MTest
colSums(MTest) - 1
# PCA Analysis
p2 <- psych::principal(DataMatrix, rotate = "varimax", nfactors = 12, scores = TRUE,
    oblique.scores = TRUE)
p2
plot(p2$values)
abline(1, 0, col = "red")
# Validation Statistics
print(p2$loadings, cutoff = 0.5, sort = T)
p2$loadings
p2$values
p2$communality
p2$rot.mat

```

## 8.5 Correspondence Analysis

```
setwd("/Users/sidneyfox/Documents/DePaul/Fall2018/DSC424/FinalProject/")
family_csv2 <- read.csv("familyxx.csv", header = T, sep = ",")
family_csv <- read.csv("familyxx.csv", header = T, sep = ",")
family_csv[is.na(family_csv)] <- -1
write.csv(family_csv, "family_modified_001.csv", row.names = F)

# read in data from GitHub
library(RCurl)
family_data <- getURL("https://raw.githubusercontent.com/stfox13/
                      DSC424FinalProject/master/family_modified_001.csv")
family_df <- read.csv(text = family_data)
View(family_df)

#### Correspondence Analysis ####
library(ca)

# compare variables for family type and highest education of family
# attained:
family_type_education <- with(family_df, table(FM_TYPE, FM_EDUC1))
prop.table(family_type_education, 1) # row percentages
prop.table(family_type_education, 2) # column percentages
fit <- ca(family_type_education)
print(fit)
print(summary(fit))
plot(fit, main = "Correspondence Analysis (symetric map)\n
      1. family type 2. highest education level")
plot(fit, mass = TRUE, contrib = "absolute", map = "rowgreen", arrows = c(F,
T), main = "Correspondence Analysis (assymetric map)\n
```

```

    1. family type 2. highest education level")
# compare variables for language of interview and family structure:
int_lang_fam_struct <- with(family_df, table(FLNGINTV, FM_STRCP))
prop.table(int_lang_fam_struct, 1) # row percentages
prop.table(int_lang_fam_struct, 2) # column percentages
fit <- ca(int_lang_fam_struct)
print(fit)
print(summary(fit))
plot(fit, main = "Correspondence Analysis (symetric map)\n
    1. language of interview 2. family structure")
plot(fit, mass = TRUE, contrib = "absolute", map = "rowgreen", arrows = c(F,
    T), main = "Correspondence Analysis (assymetric map)\n
    1. language of interview 2. family structure")
# compare variables for language of interview and family structure:
cov_inc_grp <- with(family_df, table(COVCONF, INCGRP4))
prop.table(cov_inc_grp, 1) # row percentages
prop.table(cov_inc_grp, 2) # column percentages
fit <- ca(cov_inc_grp)
print(fit)
print(summary(fit))
plot(fit, main = "Correspondence Analysis (symetric map)\n
    1. coverage confidence 2. income group")
plot(fit, mass = TRUE, contrib = "absolute", map = "rowgreen", arrows = c(F,
    T), main = "Correspondence Analysis (assymetric map)\n
    1. coverage confidence 2. income group")
# compare variables for language of interview and family structure:
phone_house <- with(family_df, table(PHONEUSE, HOUSEOWN))

```

```

prop.table(phone_house, 1) # row percentages
prop.table(phone_house, 2) # column percentages
fit <- ca(phone_house)
print(fit)
print(summary(fit))
plot(fit, main = "Correspondence Analysis (symetric map)\n
      1. working cell phone / landline 2. home ownership status")
plot(fit, mass = TRUE, contrib = "absolute", map = "rowgreen", arrows = c(F,
      T), main = "Correspondence Analysis (assymetric map)\n
      1. working cell phone / landline 2. home ownership status")
# count the number of NAs
family_na_summary <- data.frame(colSums_vals = colSums(is.na(family_csv)))
family_na_per <- data.frame(per_na = colMeans(is.na(family_csv)))
# visualize: histogram:
hist(family_na_per$per_na,
      main = "Percentage of missing values within data set \n
      total number of variables = 127",
      xlab = "Percentage of missing values")

```

## 8.6 Linear Discriminant Analysis

```

library(tidyverse)
library(corrplot)
library(plyr)
library(ggplot2)
library(RCurl)

```

```

library(psych)
require(MASS)

family_data <- getURL("https://raw.githubusercontent.com/stfox13/
                        DSC424FinalProject/master/family_modified_001.csv")

data <- read.csv(text = family_data)

setwd("/Users/sidneyfox/Documents/DePaul/Fall2018/DSC424/FinalProject/")

data_w_dummy_vars <- read.csv("fam_all_dummies.csv")[,
  -1] # drop row_names

# data = read.csv('family_modified_001.csv') data2
# = read.csv('familyxx.csv')

data <- family_df
data2

head(SubData, 2)

NewSub <- data[, c("FM_SIZE", "FCHLMCT", "FSPEDCT",
  "FLAADLCT", "FLIADLCT", "FWKLIMCT", "FWALKCT",
  "FREMECT", "FANYLCT", "FHSTATEX", "FHSTATVG",
  "FHSTATG", "FHSTATFR", "FHSTATPR", "FHICOVCT",
  "FHIPRVCT", "FHIEXCT", "FHISINCT", "FHICARCT",
  "FHICADCT", "FHICHPCT", "FHIMILCT", "FHIIHSCT",
  "FHIPUBCT", "FHIOGVCT", "FHIEBCCT", "FHDSTCT",
  "FDGLWCT1", "FDGLWCT2", "FWRKLWCT", "FSALCT", "FSEINCCT",
  "FSSRRCT", "FPENSCT", "FOPENSCT", "FSSICT", "FTANFCT",
  "FOWBENCT", "FINTR1CT", "FDIVDCT", "FCHSPCT", "FINCOTCT",
  "FSSAPLCT", "FSDAPLCT", "FWICCT", "FM_ELDR")]

# convert to Matrix

DataMatrix <- as.matrix(as.data.frame(NewSub))

# Missing data

```

```

colSums(is.na(DataMatrix))

# Histograms
ggplot(gather(NewSub), aes(value)) + geom_histogram(bins = 20) +
  facet_wrap(~key, scales = "free_x")

# Correlation Matrices
CorData <- cor(DataMatrix)
corrplot(CorData, method = "circle")

# Remove Variables with missing values coded as -1
NewSub1 <- NewSub[-c(2, 27, 3, 45, 30)]
DataMatrix <- as.matrix(as.data.frame(NewSub1))

# Drop Variables with correlation <.3 Final PCA Set
PCASet <- NewSub1[-c(match(c("FHIEXCT", "FHIMILCT",
  "FHIHSCT", "FHIPUBCT", "FHIOGVCT", "FDGLWCT2",
  "FSEINCCT", "FTANFCT", "FOWBENCT", "FINTR1CT",
  "FCHSPCT", "FINCOTCT"), names(NewSub1)))]
DataMatrix <- as.matrix(as.data.frame(PCASet))
CorData <- cor(DataMatrix)
corrplot(CorData, method = "circle")
library(psych)
describe(DataMatrix)
options(scipen = 100, digits = 5)
round(cor(DataMatrix), 2)
MCorrTest <- corr.test(DataMatrix, adjust = "none")
MCorrTest
M <- MCorrTest$p
# M
MTest <- ifelse(M < 0.01, T, F)

```

```

MT
colSums(MTest) - 1

p2 <- psych::principal(DataMatrix, rotate = "varimax",
  nfactors = 12, scores = TRUE, oblique.scores = TRUE)

p2
# plot(p2$values) abline(1, 0, col = 'red')
# combine principal component scores with target
# variable

pc_data <- merge(p2$scores[, 0:8], data$FHICOST, by = "row.names",
  all = T)[, -1]

cleaned_data <- merge(PCASet, data$FHICOST, by = "row.names",
  all = T)[, -1]

#### linear discriminant analysis ####

# break data into test and train (80 / 20 split):
set.seed(101)

sample_pdata <- sample.int(n = nrow(pc_data), size = floor(0.8 *
  nrow(pc_data)), replace = F)

pc_data_train <- as.data.frame(pc_data[sample_pdata,
  ])

pc_data_test <- as.data.frame(pc_data[-sample_pdata,
  ])

familyLDA <- lda(y ~ ., data = pc_data_train)

# plot(familyLDA)
# predict

familyLDA.values <- predict(familyLDA, pc_data_test)

p <- predict(familyLDA, newdata = pc_data_test)

mean(p$class == pc_data_test$y)

```



```

# 37% - not great :(
# cleaned data # break data into test and train (80
# / 20 split):
set.seed(101)
sample_cleaneddata <- sample.int(n = nrow(cleaned_data),
  size = floor(0.8 * nrow(cleaned_data)), replace = F)
cleaned_data_train <- as.data.frame(cleaned_data[sample_cleaneddata,
  ])
cleaned_data_test <- as.data.frame(cleaned_data[-sample_cleaneddata,
  ])
familyLDA2 <- lda(y ~ ., data = cleaned_data_train)
# plot(familyLDA)
# predict
familyLDA2.values <- predict(familyLDA2, cleaned_data_test)
p2 <- predict(familyLDA2, newdata = cleaned_data_test)
mean(p2$class == cleaned_data_test$y)
# 38% - not great :(
# data with dummy vars # break data into test and
# train (80 / 20 split):
set.seed(101)
sample_dummyvardata <- sample.int(n = nrow(data_w_dummy_vars2),
  size = floor(0.8 * nrow(data_w_dummy_vars2)), replace = F)
dummy_var_data_train <- as.data.frame(data_w_dummy_vars2[sample_dummyvardata,
  ])
dummy_var_data_test <- as.data.frame(data_w_dummy_vars2[-sample_dummyvardata,
  ])
familyLDA3 <- lda(FHICOST ~ ., data = dummy_var_data_train)

```

```

# plot(familyLDA)

# predict

familyLDA3.values <- predict(familyLDA3, dummy_var_data_test)

ldahist(data = familyLDA3.values$x[, 2], g = FHICOST)

plot(familyLDA3.values$x[, 1], familyLDA3.values$x[,
      2])

p3 <- predict(familyLDA3, newdata = dummy_var_data_test)

# accuracy: 72% - pretty good! :)

mean(p3$class == dummy_var_data_test$FHICOST)

p3_df <- data.frame(LD1 = p3$x, class = p3$class)

# display density plot of linear discriminants and
# predicted values

ggplot(p3_df) + geom_density(aes(LD1, fill = class),
                              alpha = 0.2) + ggtitle("Density plot displaying the density of LD1\n
                                                    color coded by the class variable.")

reg_formula <- paste("FHICOST ~", paste(names(data_w_dummy_vars2)[-1],
                                         collapse = " + "))

# observation: aliased coefficients:

alias(lm(reg_formula, data = data_w_dummy_vars2))

fit <- lm(reg_formula, data = data_w_dummy_vars2)

vif_results <- car::vif(fit)

# observation: rampant multicollinearity (vif > 10)

vif_results

remove_factors <- c(FM_STRCP_12, FM_STRCP_21, FM_STRCP_22,
                   FM_STRCP_23, FM_STRCP_31, FM_STRCP_32, FM_STRCP_33,
                   FM_STRCP_41, FM_STRCP_42, FM_STRCP_43, FM_STRCP_44,
                   FM_STRCP_45)

```

```
data_w_dummy_vars2 <- subset(data_w_dummy_vars, select = -c(FM_STRCP_12,
  FM_STRCP_21, FM_STRCP_22, FM_STRCP_23, FM_STRCP_31,
  FM_STRCP_32, FM_STRCP_33, FM_STRCP_41, FM_STRCP_42,
  FM_STRCP_43, FM_STRCP_44, FM_STRCP_45, FM_TYPE_2,
  FM_TYPE_3, FM_TYPE_4, INCGRP4_2, INCGRP4_3, INCGRP4_4,
  INCGRP4_5, FHCPHRCT, FWKLIMCT, FHSTATEX, FHSTATPR,
  FSSICT, FSDAPLCT, FHICOVCT, FLIADLCT, FANYLCT,
  FHSTATFR, FOPENSCT, FSSAPLCT, FM_KIDS, FREMEMCT,
  FHSTATG, FLAADLCT, FPENSCT, FHIPRVCT, FM_SIZE,
  FWALKCT, FHSTATVG, FSSRRCT, FHCHMYN_2, FHCPHRYN_2,
  FLAADLYN_2, FLIADLYN_2, FWKLIMYN_2, FLIADLYN_2,
  FWKLIMYN_2, FWALKYN_2, FREMEMYN_2, FANYLYN_2, FPENSYN_2,
  FOPENSYN_2, FSSIYN_2, FSSAPLYN_2, FSDAPLYN_2))
```

## 9 Visualization Appendix

Please find the figures referenced throughout this paper below:

### 9.1 Executive Summary

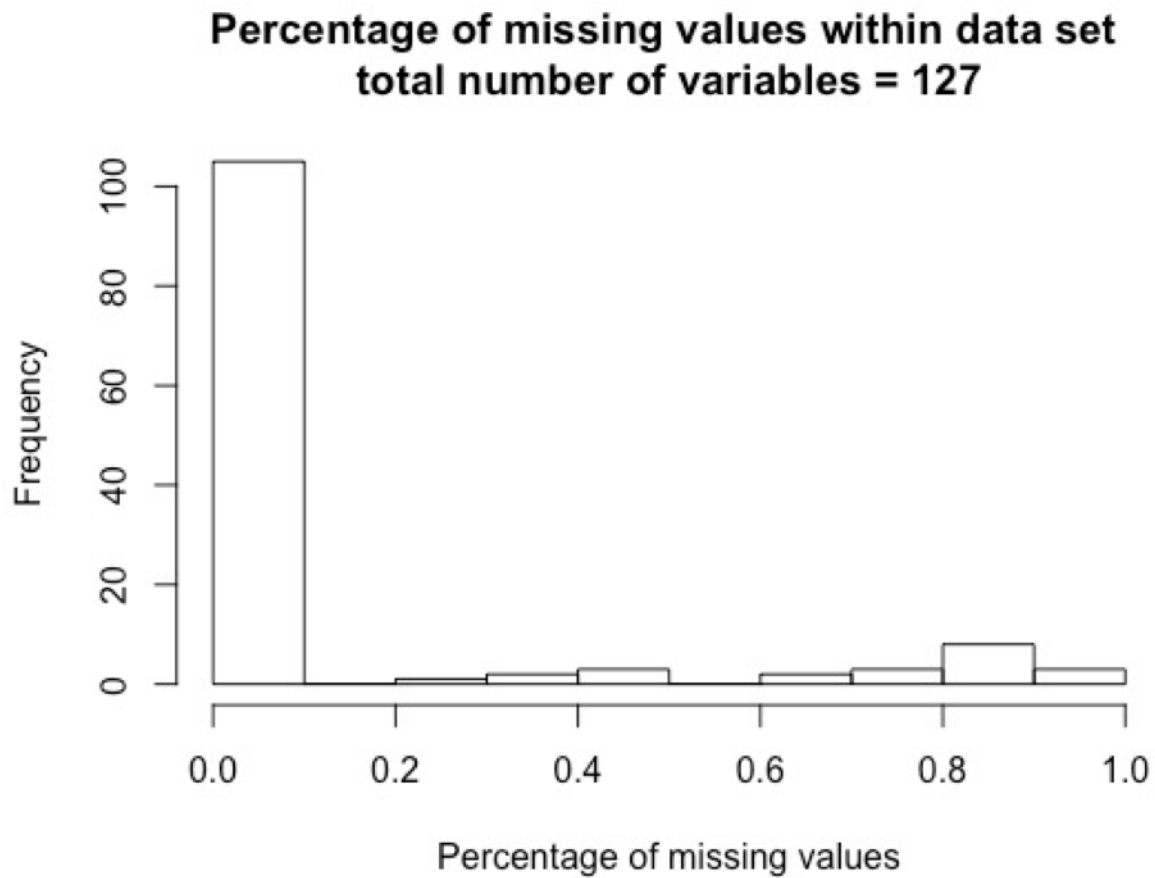


Figure 1: Histogram of Missing Values

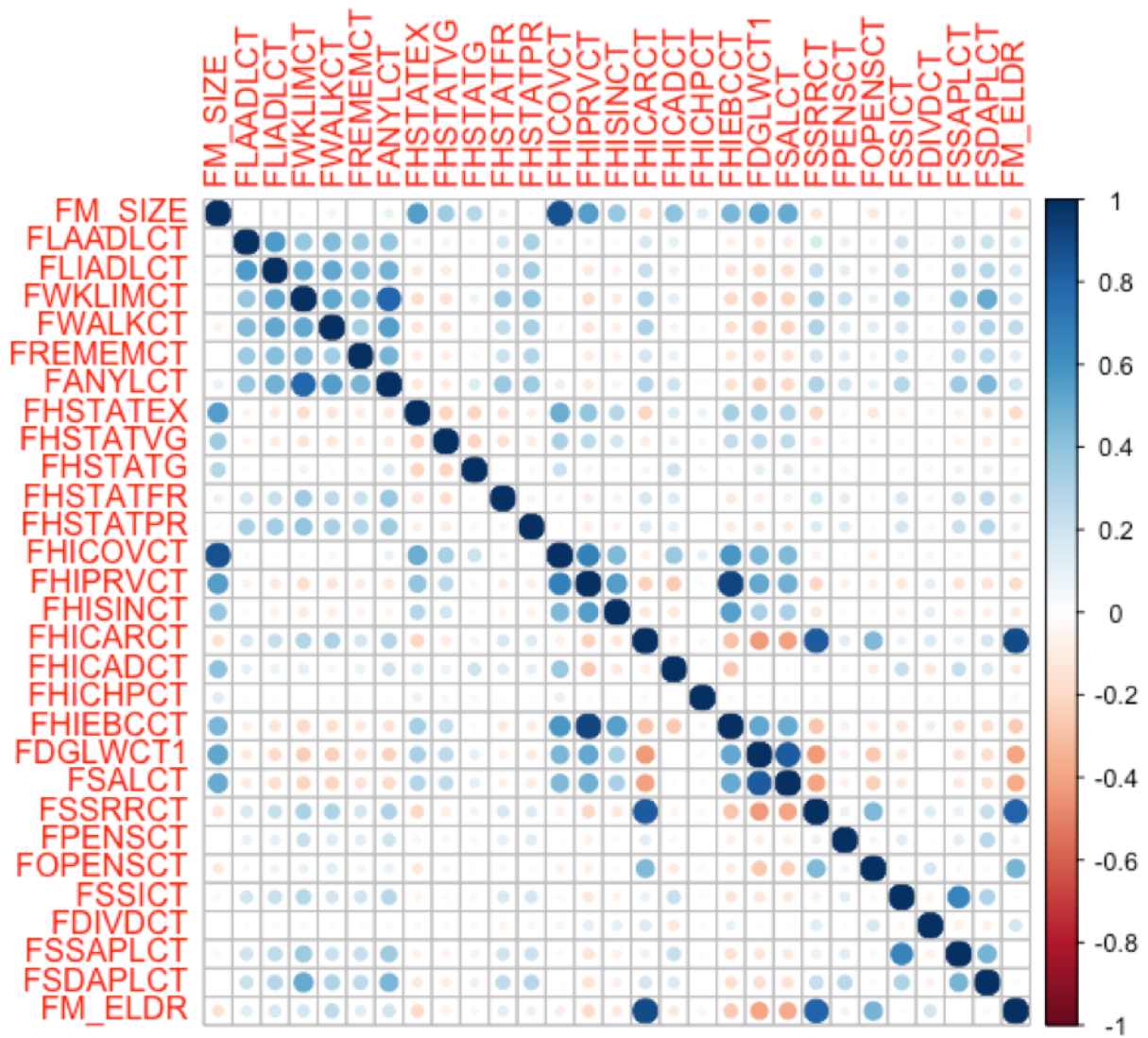


Figure 2: Correlation Plot Displaying a Subset of Variables

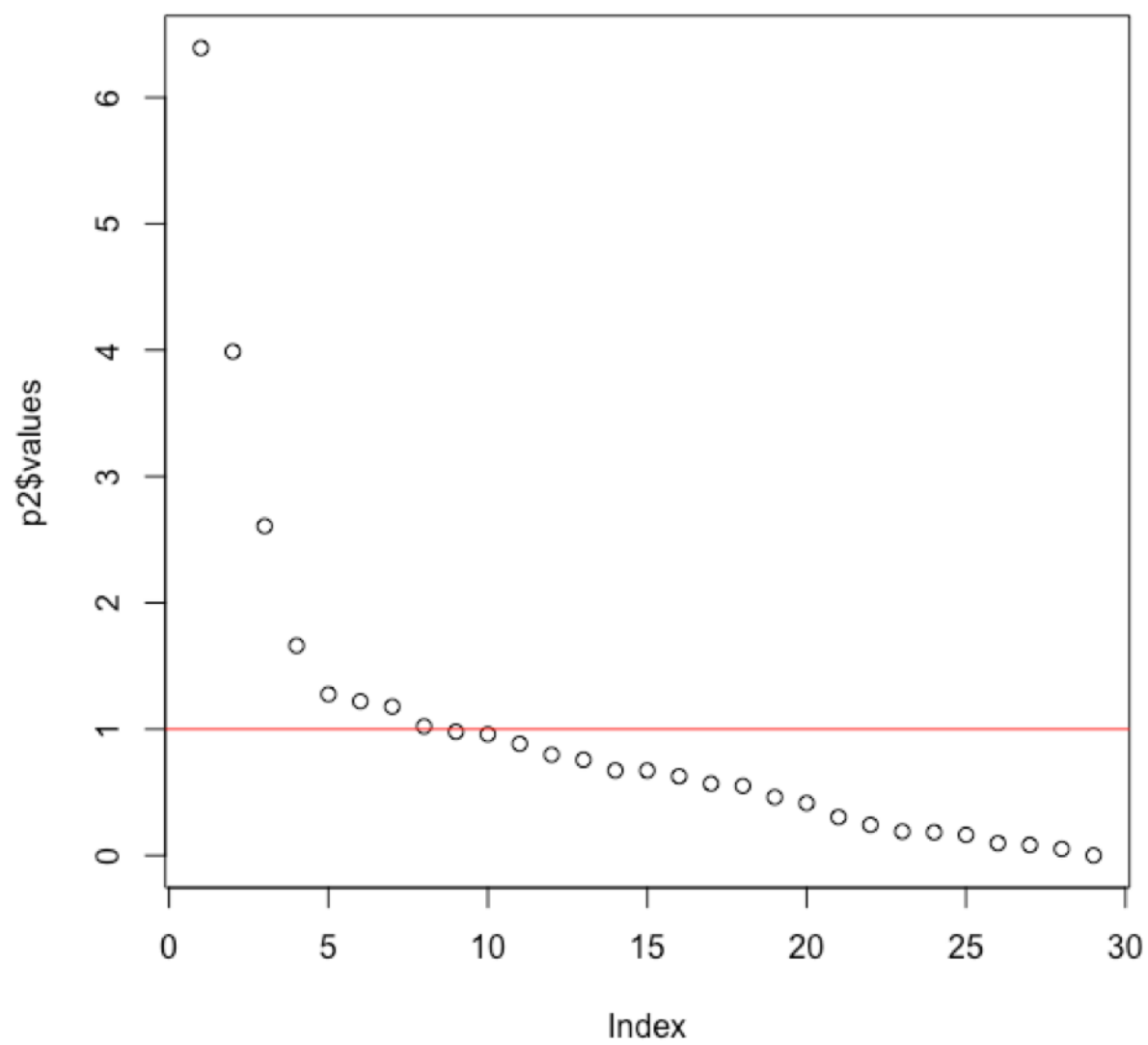


Figure 3: PCA Scree Plot

## 9.2 Correspondence Analysis

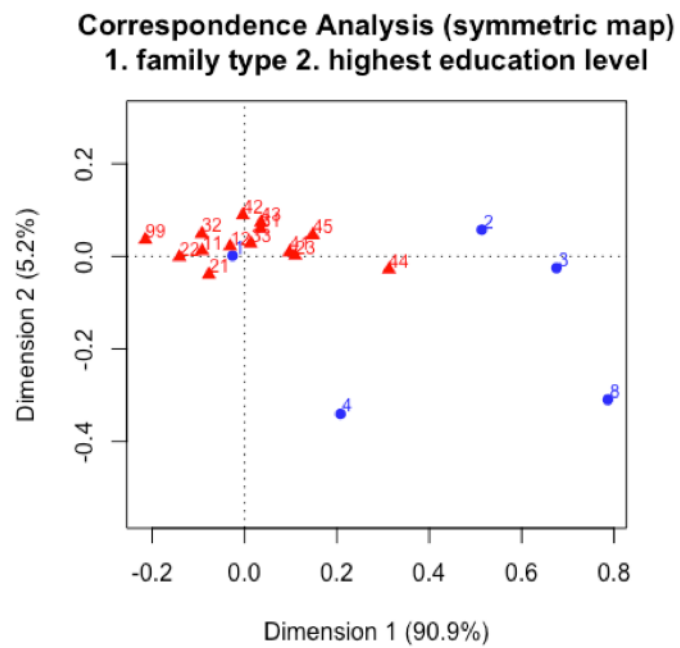


Figure 4: Correspondence Analysis Plot 1

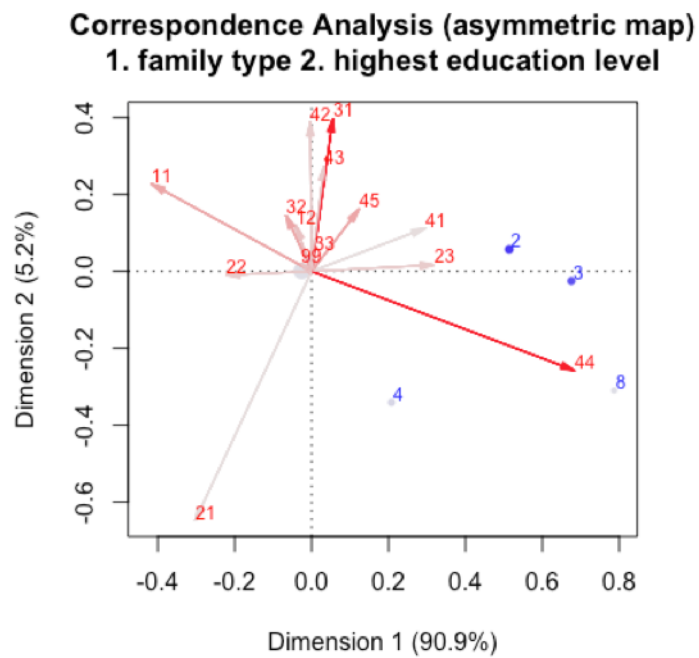


Figure 5: Correspondence Analysis Plot 2

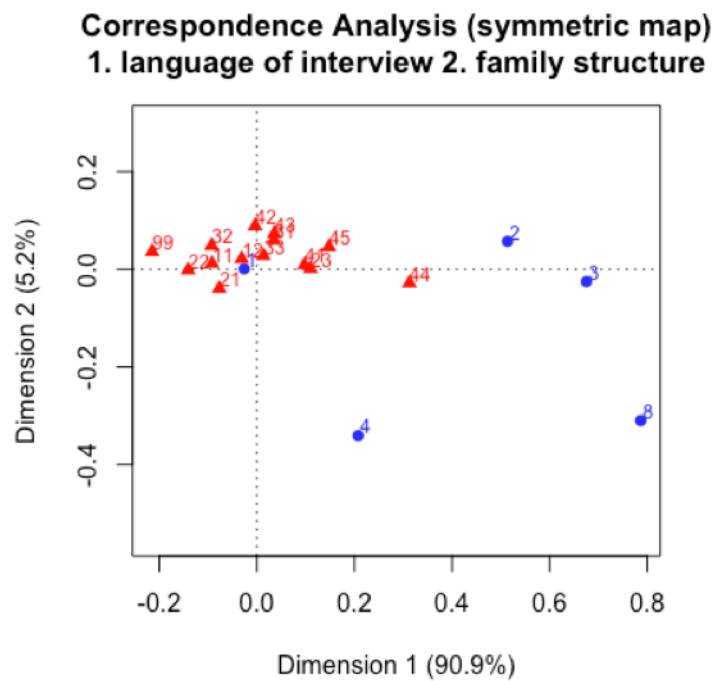


Figure 6: Correspondence Analysis Plot 3

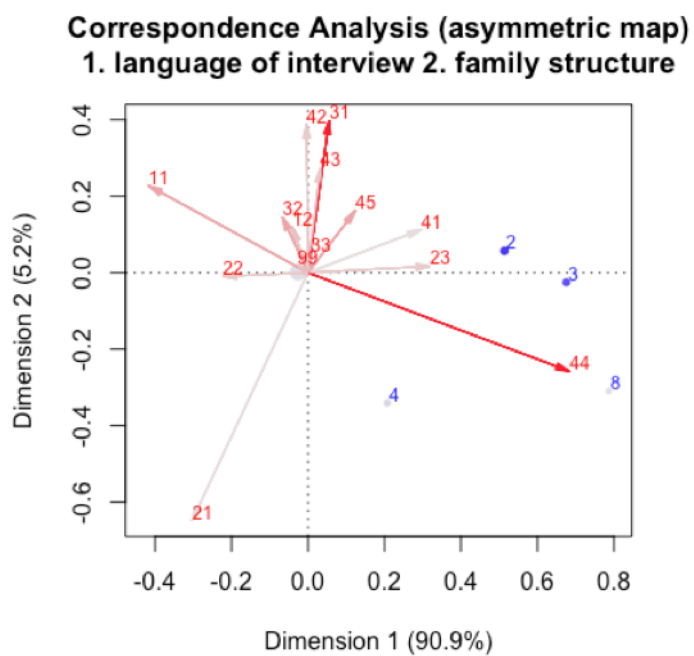


Figure 7: Correspondence Analysis Plot 4



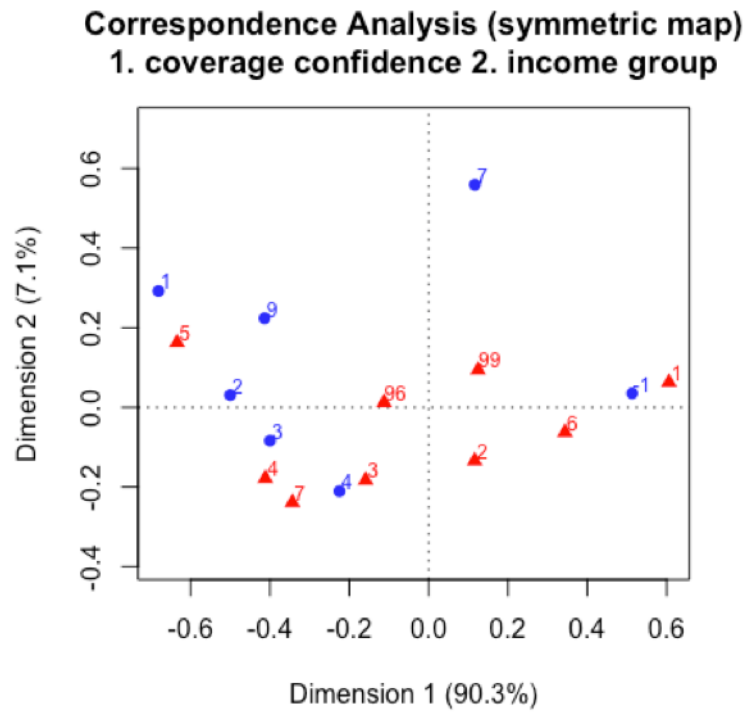


Figure 8: Correspondence Analysis Plot 5

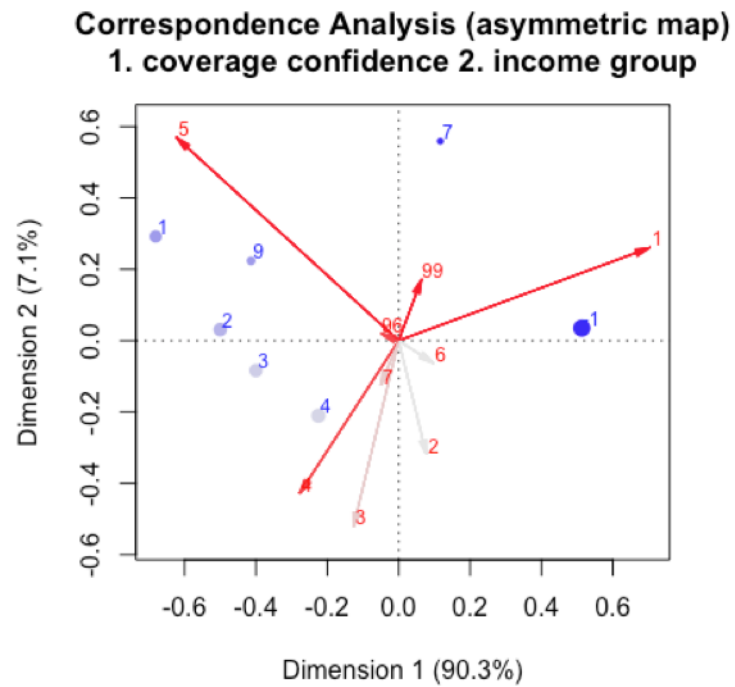


Figure 9: Correspondence Analysis Plot 6

**Correspondence Analysis (symmetric map)**  
**working cell phone / landline 2. home ownership**

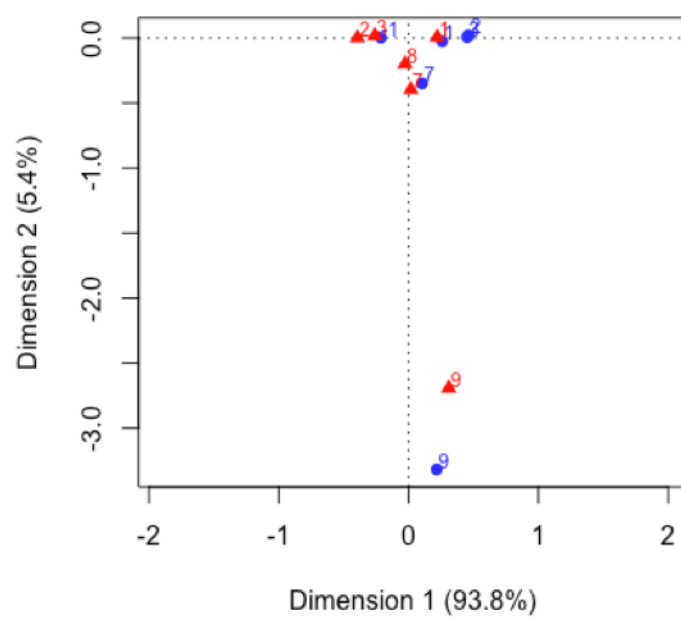


Figure 10: Correspondence Analysis Plot 7

**Correspondence Analysis (asymmetric map)**  
**working cell phone / landline 2. home ownership**

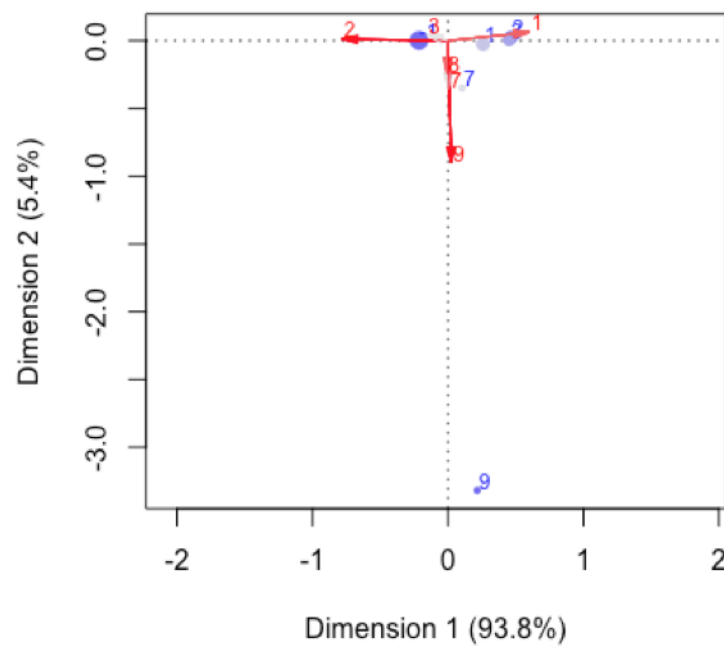


Figure 11: Correspondence Analysis Plot 8

### 9.3 Linear Discriminant Analysis

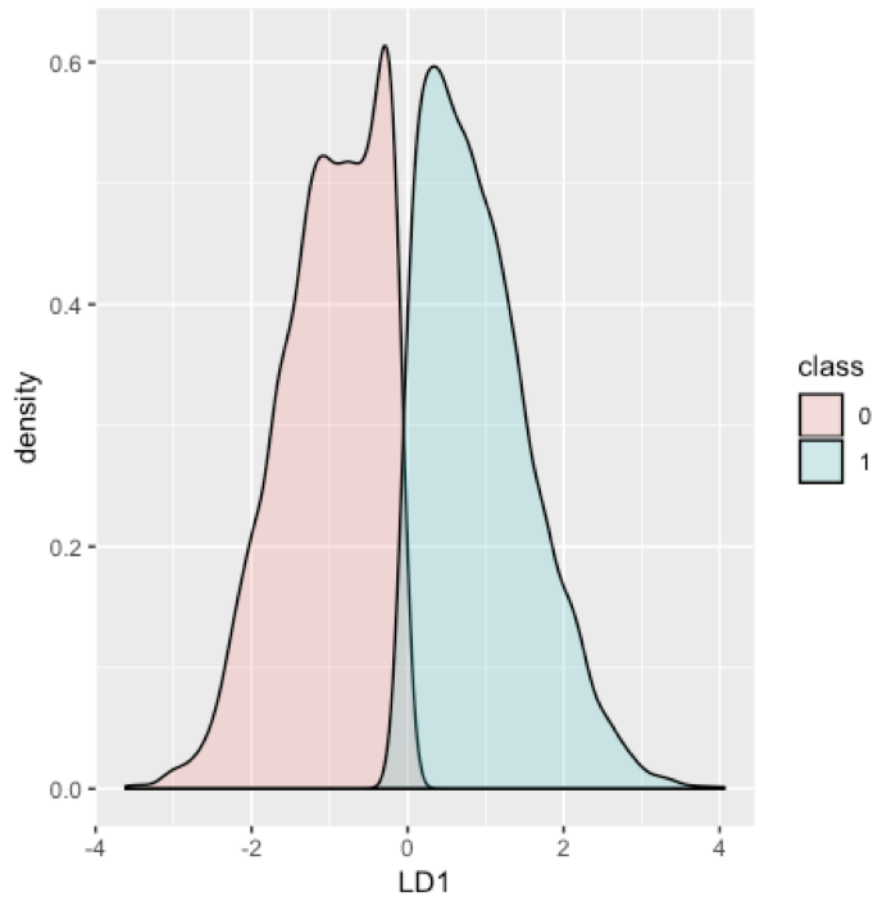


Figure 12: Density Plot Comparing Outcomes of LDA

### 9.4 Common Factor Analysis

Var Char	Variable Description	Loadings	Factor Name
FDGLWCT1	Family Working Last Week	0.659	Likelihood of Insured
FHSTATEX	Family in Excellent Health	0.501	
FSALCT	Family Receiving Income from Wages	0.642	
FHIPRVCT	Family on Private Insurance	0.798	
FHICOVCT	Family with Insurance	0.814	
FHISINCT	Family on Single Plan Health Insurance	0.571	
FLIADLCT	Family with IADL	0.605	Likelihood of Health Issues
FWKLIMCT	Family with work limits due to health	0.858	
FWALKCT	Family difficult walking	0.604	
FREMEMCT	Family difficult remembering	0.531	
FANYLCT	Family limited in any way	0.858	
FSDAPLCT	Family apply for SSDI	0.56	
FHSTATPR	Family in Poor Health	0.459	
FLAADLCT	Family with ADL Help	0.501	
FSSAPLCT	Family apply for SSI	0.447	
FM_ELDR	Family Members over 65	0.932	Likelihood of Being Elderly
FSSRRCT	Family receive income from R/R or Pension	0.823	
FHICARCT	Family on Medicare	0.921	
FOPENSCT	Family receive survivor/spousal benefits	0.477	

Figure 13: Factors with Loadings

	Likelihood of Insured	Likelihood of Health Issues	Likelihood of Being Elderly
<b>SS loadings</b>	3.661	2.943	2.842
<b>Proportion Var</b>	0.193	0.155	0.15
<b>Cumulative Var</b>	0.193	0.348	0.497

Figure 14: Overview of Variance with Factor Names

## 9.5 Cluster Analysis

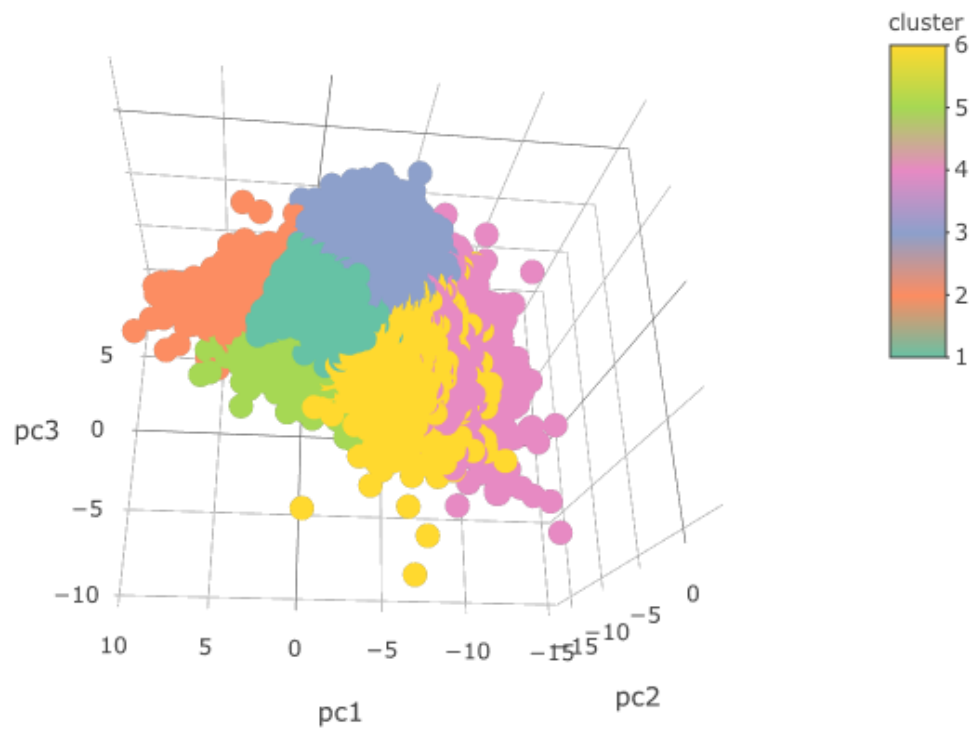


Figure 15: Cluster Representation

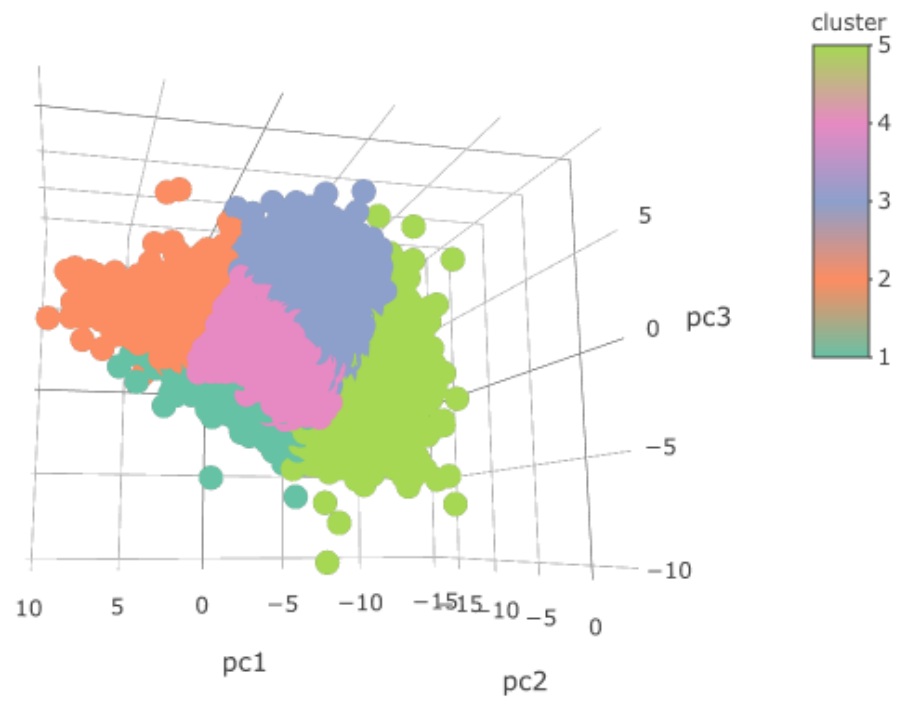


Figure 16: Cluster Representation

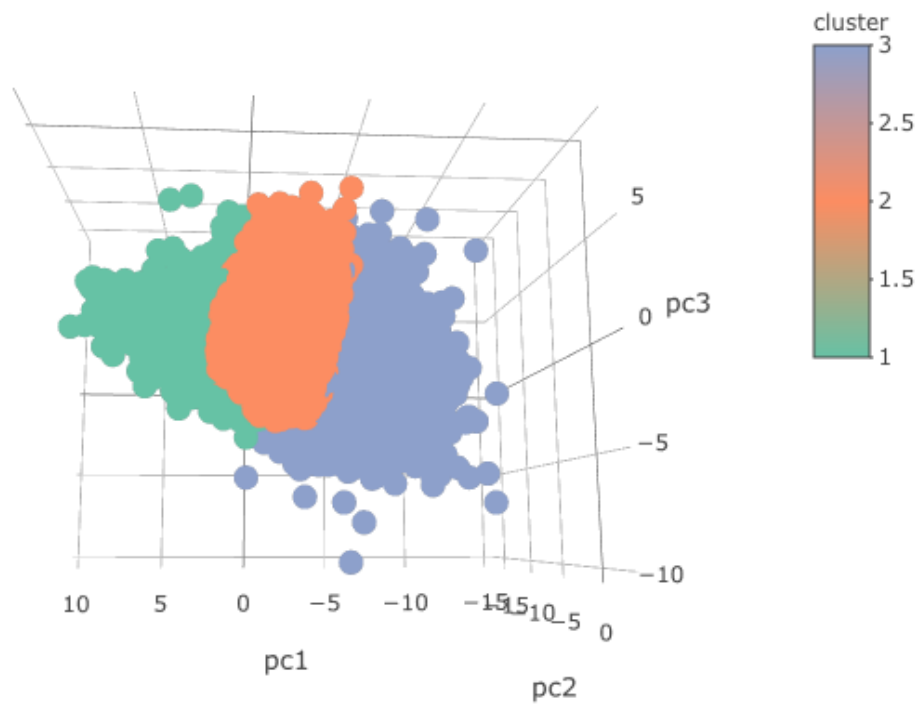


Figure 17: Cluster Representation



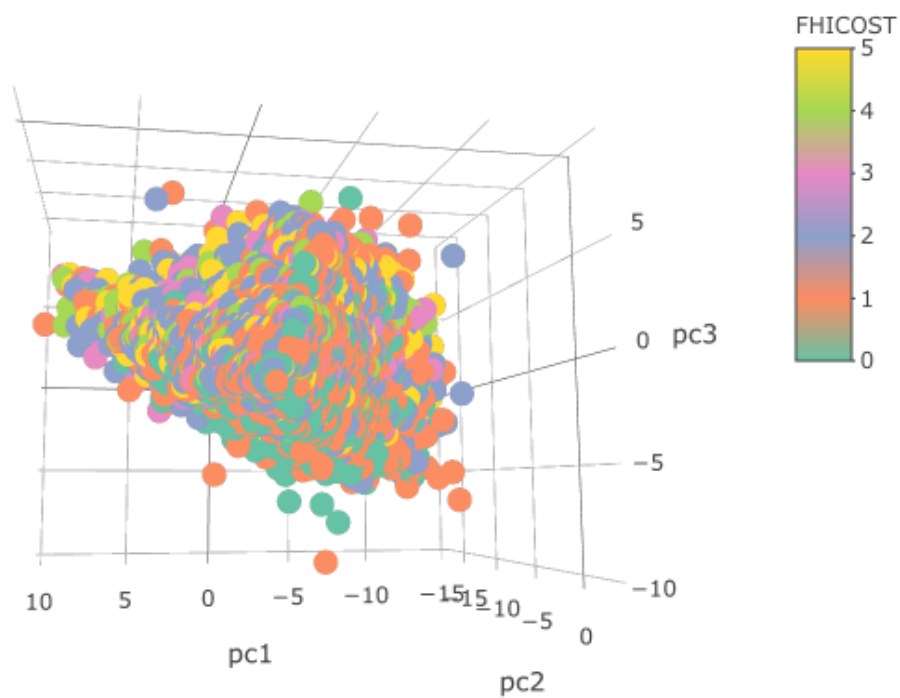


Figure 18: Cluster Representation

##	cluster			
##	FHICOST	1	2	3
##	high	0.4238197	0.6340136	0.4970060
##	low	0.5761803	0.3659864	0.5029940

Figure 19: Cluster Outputs

##	FM_SIZE	FLAADLCT	FLIADLCT	FWKLIMCT	FMALKCT	FREMEMCT	FANYLCT
## 1	1.092025	0.01065537	0.03121684	0.2077038	0.11642761	0.06340858	0.2818646
## 2	2.495778	0.03607345	0.03139931	0.1152362	0.04542679	0.04812904	0.2892039
## 3	1.567117	1.38307279	1.76483329	2.2029402	1.86077461	1.49034298	2.2022212
##	FHSTATEX	FHSTATVG	FHSTATG	FHSTATFR	FHSTATPR	FHICOVCT	FHIPRVCT
## 1	0.3142845	0.4920817	0.5014329	0.2470549	0.04535570	0.927514	0.5495737
## 2	1.2094439	1.0923727	0.6895353	0.2176955	0.03652751	2.369624	1.8628696
## 3	0.2058218	0.3020967	0.7396466	1.2912004	1.22773659	1.460852	0.5216723
##	FHISINCT	FHICARCT	FHICADCT	FHICHPCT	FHIEBCCT	FDGLWCT1	
## 1	0.3194660	0.79648478	0.1922279	0.03132827	0.07747224	0.7862517	
## 2	1.2439924	0.06941039	0.4979016	0.17503870	1.33046425	2.1076456	
## 3	0.3412092	1.42163553	0.7899674	0.07180846	-0.05442460	0.5183454	
##	FSALCT	FSSRRCT	FPENSCT	FOPENSCT	FSSICT	FDIVDCT	
## 1	0.8010731	0.77836747	0.12290732	0.57603634	0.05100196	0.4256121	
## 2	2.0732982	0.08198257	0.09199469	0.09147293	0.04176725	0.4568340	
## 3	0.5920391	1.42759461	0.80741273	0.64481988	1.17241969	0.2935624	
##	FSSAPLCT	FSDAPLCT	FM_ELDR				
## 1	0.07897665	0.1591161	0.82287901				
## 2	0.08575811	0.1050765	0.08296434				
## 3	1.42454694	1.6990005	1.11643645				

Figure 20: Cluster Variable Matrix

## 9.6 Logistic Regression

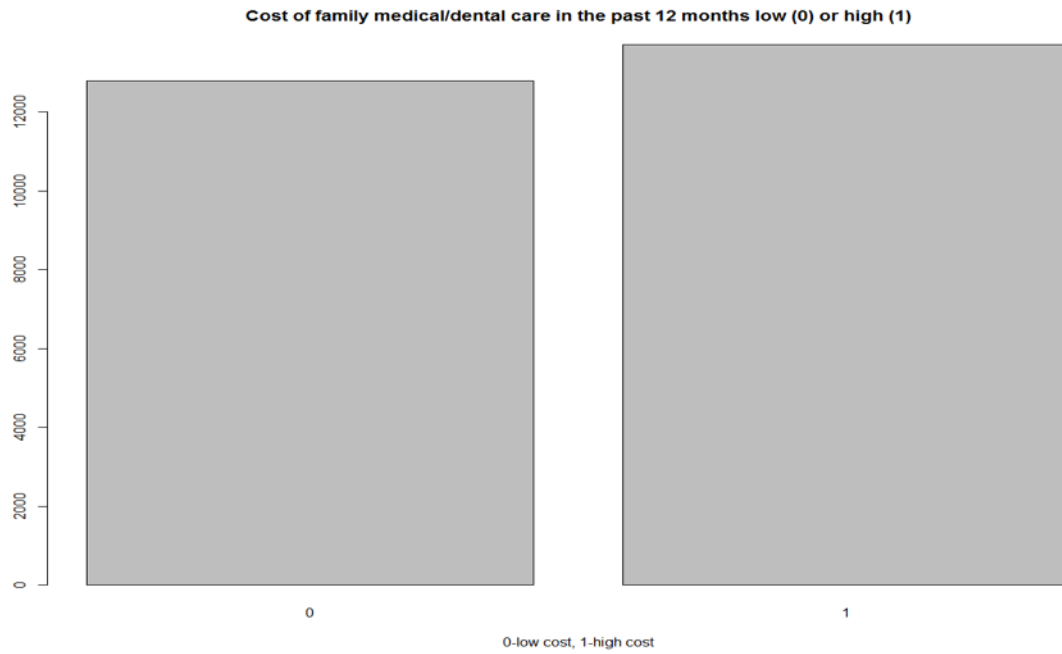


Figure 21: Logistic Regression

	FALSE	TRUE
0	7194	3032
1	2654	8304

Figure 22: Logistic Regression

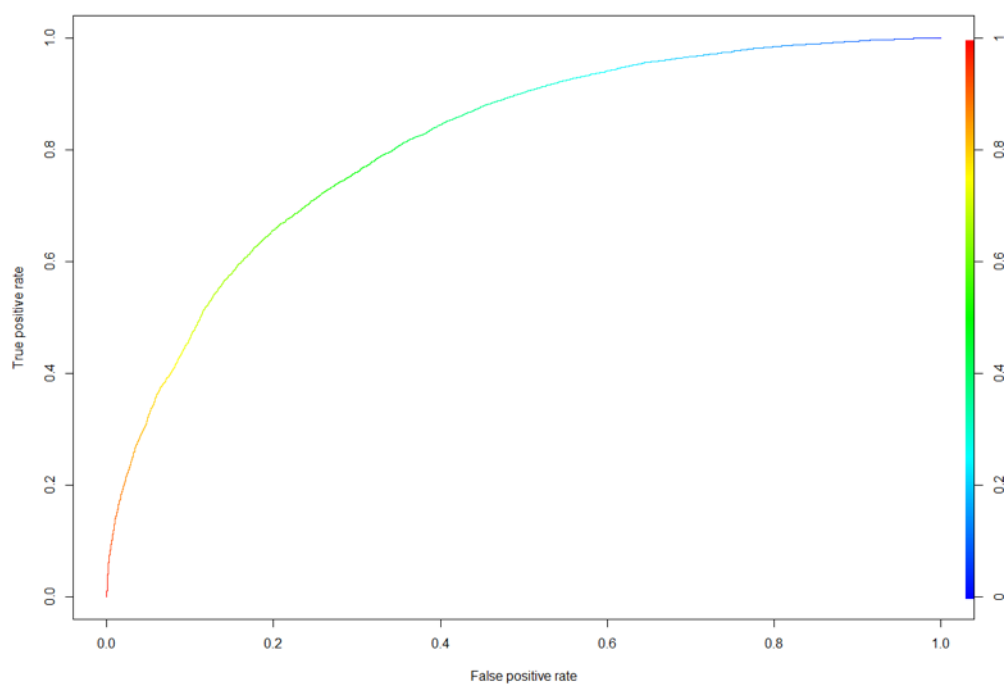


Figure 23: Logistic Regression

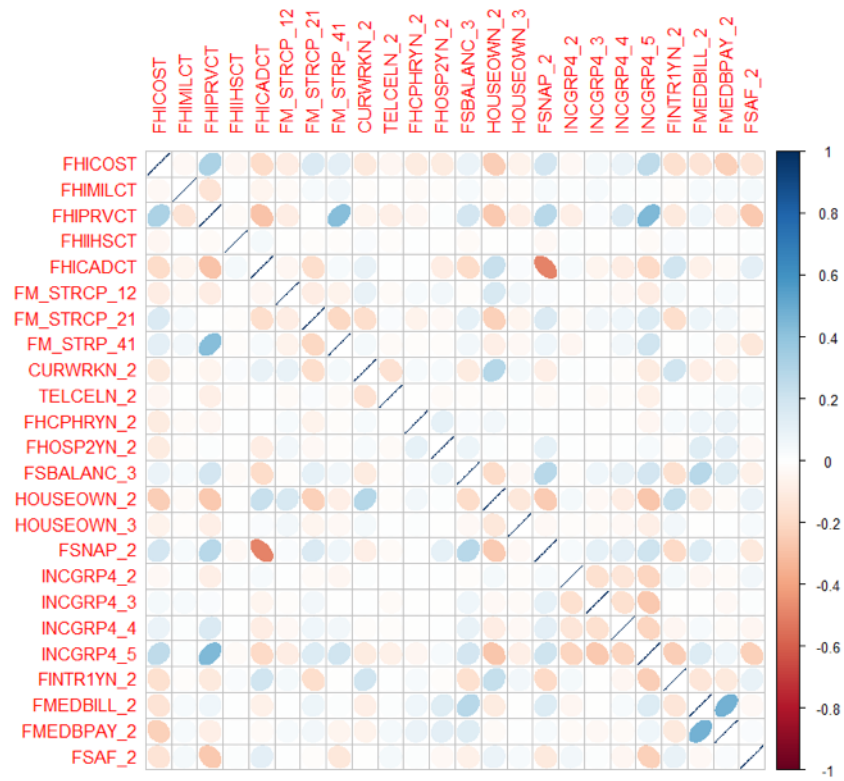


Figure 24: Logistic Regression

	FALSE	TRUE
0	7021	3205
1	2726	8232

Figure 25: Logistic Regression

	FALSE	TRUE
0	1729	828
1	689	2051

Figure 26: Logistic Regression

## 9.7 Principal Component Analysis

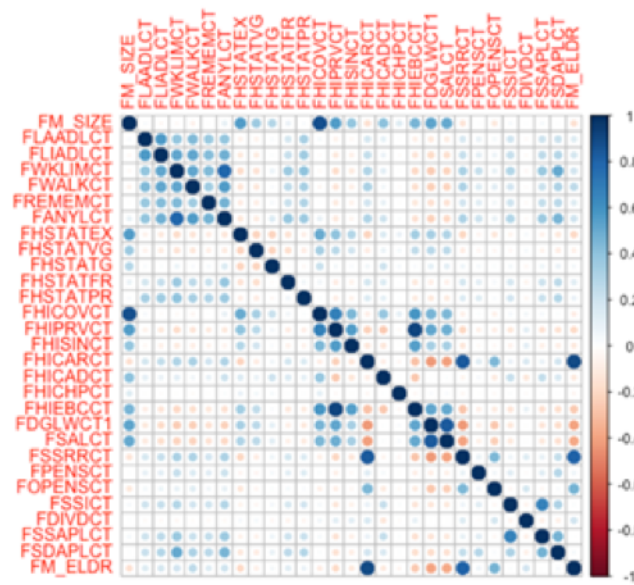


Figure 27: Principal Component Analysis

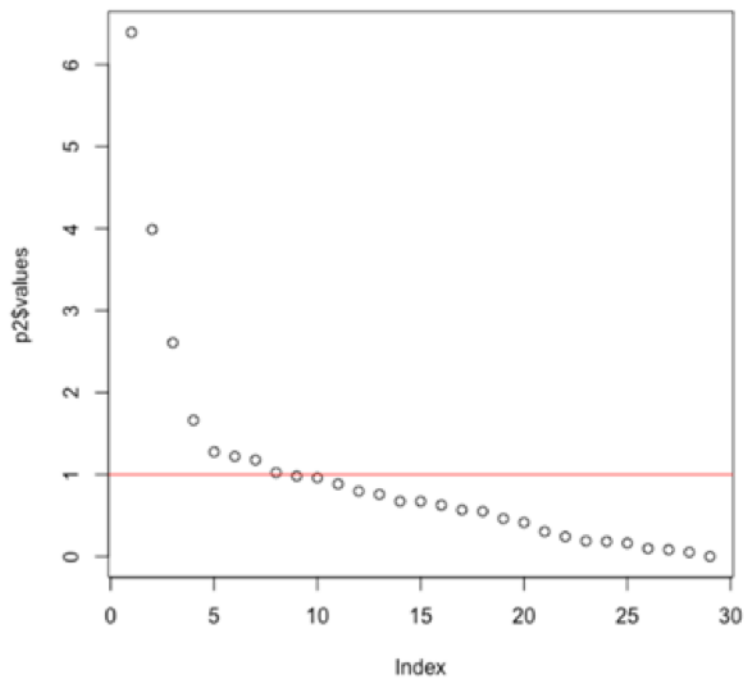


Figure 28: Principal Component Analysis

Loadings:												
	RC1	RC2	RC3	RC4	RC5	RC12	RC8	RC7	RC6	RC10	RC9	RC11
FLAADLCT	0.754											
FLIADLCT	0.793											
FWKLIMCT	0.681											
FWALKCT	0.695											
FREMEMCT	0.636											
FANYLCT	0.680											
FHSTATPR	0.592											
FHICOVCT		0.679		0.632								
FHIPRVCT		0.913										
FHISINCT		0.753										
FHIEBCT		0.879										
FHICARCT			0.912									
FSSRRCT			0.868									
FOPENSCT			0.608									
FM_ELDLDR			0.930									
FM_SIZE		0.551		0.695								
FHICADCT				0.870								
FSSICT					0.879							
FSSAPLCT					0.870							
FDGLWCTL						0.765						
FSALCT						0.779						
FPENSCT							0.893					
FHSTATG								0.963				
FHSTATEX									-0.601			
FHSTATVG									0.903			
FHSTATFR										0.857		
FHICHPCT											0.995	
FDIVDCT												0.963
FSDAPLCT												
ss loadings	3.768	3.565	3.210	1.979	1.862	1.680	1.245	1.216	1.216	1.202	1.036	0.984
Proportion Var	0.130	0.123	0.111	0.068	0.064	0.058	0.043	0.042	0.042	0.041	0.036	0.034
cumulative var	0.130	0.253	0.364	0.432	0.496	0.554	0.597	0.639	0.681	0.722	0.758	0.792

Figure 29: Principal Component Analysis

---

## References

Black, L. I., T. C. Clarke, P. M. Barnes, B. J. Stussman, and R. L. Nahin. 2015. “Use of Complementary Health Approaches Among Children Aged 4-17 Years in the United States: National Health Interview Survey, 2007-2012.” *National Health Statistics Reports* 78: 1–19.

CMS. 2018. “NHE-Fact-Sheet.” <https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/nationalhealthexpenddata/nhe-fact-sheet.html>.

Jean-Louis, G., N. J. Williams, D. Sarpong, A. Pandey, S. Youngstedt, F. Zizi, and G. Ogedegbe. 2014. “Associations Between Inadequate Sleep and Obesity in the Us Adult Population: Analysis of the National Health Interview Survey (1977-2009).” *BMC Public Health* 14: 290.

Ma, V. Y., L. Chan, and K. J. Carruthers. 2014. “Incidence, Prevalence, Costs, and Impact on Disability of Common Conditions Requiring Rehabilitation in the United States: Stroke, Spinal Cord Injury, Traumatic Brain Injury, Multiple Sclerosis, Osteoarthritis, Rheumatoid Arthritis, Limb Loss, and Back Pain.” *Archives of Physical Medicine and Rehabilitation* 95 (5): 986–95.