INTRODUCTION

Diabetes is one of the most common diseases in the United States and worldwide. According to the CDC (Center for Disease Control and Prevention), the number of adults in the United States who have diabetes is approximately 11%, which totals to 37.3 million people, which is double what it was twenty years ago (By the Numbers: Diabetes in America, CDC). Of these people, one in five people go undiagnosed (What is diabetes? | CDC, CDC). This may be due to multiple reasons, among which would be not realizing the health risks that contribute to getting diabetes, or perhaps their economic situation might not allow them to get treatment, let alone a diagnosis. Furthermore, approximately 96 million adults have prediabetes, a condition which emulates diabetes, just with a blood sugar level slightly lower than that of diagnosed diabetics due to the pancreas not being able to produce enough insulin to pull enough sugar out from the blood (Prediabetes - Your Chance to Prevent Type 2 Diabetes | CDC, CDC).

This project utilizes data regarding aspects of potential diabetic factors collected by the CDC. The aim of the study is to comprehend the effectiveness of specific health-related and socioeconomic factors in their ability to predict the percentage of diabetic people within the United States. Within the study, we create and evaluate regression models using a curated set of variables from the CDC's "Social Determinants of Health" dataset, and we perform exploratory factor analysis to evaluate the performance of similar variables (Social Determinants of Health - United States Diabetes Surveillance System, CDC). With this research, we aim to find the most useful ideas to aid in the diagnosis of diabetes in undiagnosed populations.

RESEARCH QUESTION

The main research question is how effective are certain factors in the prediction of the percentage of the diabetic population within the United States, and of these factors, which are the primary factors that seem to contribute to higher diabetes rates?

METHODOLOGY

We plan to approach this question by performing a cursory analysis of the different variables we collect from the CDC SDoH website, then diving deeper into the analysis by performing regressions using different methods - specifically Ordinary Least Squares, one with all the variables and one with an automatic variable subset generator, selecting the most parsimonious model with the least loss in adjusted R^2. We also use Ridge and LASSO regression models to mitigate any multicollinear effects with penalties. Then after that, we will use Principal Component Analysis to test the dimension reduction effectiveness, then use Pricipal Factor Analysis to try to find underlying/latent factors hidden within the variables we select, which we then use in regression to see how effective these compressions of data retain information.

DATA

We collected the original dataset from CDC (Center for Disease Control and Prevention). The variables, with names are on the left and their description on the right, are as follows:

Variable Name	Description
County	County Name
State	State
CountyFIPS	County ID
A.Percent.DiagDia	All, Percent Diagnosed Diabetes
M.Percent.DiagDia	Male, Percent Diagnosed Diabetes
F.Percent.DiagDia	Female, Percent Diagnosed Diabetes
A.Percent.Obes	All, Percent Obese
M.Percent.Obes	Male, Percent Obese
F.Percent.Obes	Female, Percent Obese
A.Percent.Inac	All, Percent Physically Inactive
M.Percent.Inac	Male, Percent Physically Inactive
F.Percent.Inac	Female, Percent Physically Inactive
M.Number.DiagDia	Male, Raw Number Diagnosed Diabetes
F.Number.DiagDia	Female, Raw Number Diagnosed Diabetes
M.Number.Obes	Male, Raw Number Obese
F.Number.Obes	Female, Raw Number Obese
M.Number.Inac	Male, Raw Number Physically Inactive
F.Number.Inac	Female, Raw Number Physically Inactive
A.Percent.N.Inet	All, Percent Without Internet
A.Percent.ChilPov	All, Percent Children in Poverty
A.Percent.FRLunch	All, Percent Enrolled in Free and Reduced Lunch Programs
A.Percent.FInsec	All, Percent Financially Insecure
A.Percent.NHInsur	All, Percent Without Health Insurance
A.Percent.SHCB	All, Percent With Severe Housing Cost Burden
A.Percent.LCommute	All, Percent with Commute Longer than 60 minutes

For the original dataset, we have a total of 22 variables. To ensure all variables share the same unit, we chose to utilize the percentage data as opposed to using the raw numbers, as this would standardize the entire dataset to the use of percentages. The non numeric variables were also removed for building the regression models. Two final frames of data were created each containing 1 dependent variable and 9 independent variables.

During the initial preprocessing step, some counties were completely omitted from some of the data subsets, while other counties did not collect data with respect to some variables. The counties which were completely absent from the socioeconomic factor data compilation were then completely removed from the dataset, as these counties were largely missing data for most other variables. The counties which had only missing data for some socioeconomic factors had a replacement value of the average of the column. All other null values within the dataset would remove the entire row within R. The data was then grouped into different subsets with respect to gender. The data frame “dap” contains five socioeconomic variables, while “dmp” contains male health factors and “dfp” contains female health factors. The final step bound the gender specific factors with the socioeconomic factors into two separate data frames by gender, "dm" and "df", as seen in the table below. These steps are demonstrated in Figure 1, located in the appendix.

Data Frame	Independent Variables Count	Dependent Variable Count	Independent Variable	Dependent Variable
dm (male)	9	1	M.Percent.Obes M.Percent.Inac A.Percent.N.Inet A.Percent.ChilPov A.Percent.FRlunch A.Percent.FInsec A.Percent.NHInsur A.Percent.SHCB A.Percent.LCommute	M.Percent.DiagDia
df (female)	9	1	F.Percent.Obes F.Percent.Inac A.Percent.N.Inet A.Percent.ChilPov A.Percent.FRlunch A.Percent.FInsec A.Percent.NHInsur A.Percent.SHCB A.Percent.LCommute	F.Percent.DiagDia

The final count of rows in the dataset was 3134, and from those, a training to testing split of 80% to 20% was generated.

ANALYSIS, RESULTS & FINDINGS

Variable exploration

The summary of the “dm” and “df” data sets are shown in the table shown below, where the "dm" set contains variables that start with M and A, and the "df" set contains variables that start with F and A. This is found in Figure 2 in the appendix, but it is condensed in the table below.

	Min	1Q	Median	Mean	3Q	Max
M.Percent.DiagDia	4.500	7.900	8.900	9.169	10.10	16.30
M.Percent.Obes	11.40	23.10	28.20	27.91	32.40	46.40
M.Percent.Inac	9.100	15.70	18.50	18.90	21.60	38.20
F.Percent.DiagDia	3.800	6.800	7.800	8.135	9.100	16.40
F.Percent.Obes	10.60	22.90	27.90	27.83	32.30	52.50
F.Percent.Inac	9.400	17.50	20.60	21.08	24.00	39.90
A.Percent.N.Inet	2.500	14.80	19.50	20.54	25.10	59.60
A.Percent.ChilPov	2.600	12.60	17.60	18.69	23.40	59.70
A.Percent.FRLunch	1.200	42.40	53.00	54.38	64.10	100.0
A.Percent.FInsec	2.900	10.50	12.90	13.09	15.57	29.40
A.Percent.NHInsur	0.700	5.800	8.700	9.618	12.10	40.90
A.Percent.SHCB	0.800	8.400	10.20	10.67	12.60	30.30
A.Percent.LCommute	0.200	4.500	6.800	8.079	10.30	38.70

The boxplots in Figure 3 (a-c) visualize the spread of the gender-split variables. In Figure 3 (a), the spread of the set of males who have diagnosed diabetes trends higher than the set of females with diagnosed diabetes, with males having a higher first quartile value than the median value of females. Figure 3 (b) shows male and female obesity rates and ranges as very similar. The rate of inactivity in the male population trends slightly lower than the rate of inactivity in the female population, as shown in Figure 3 (c).

Figure 3 (d) shows the pairwise scatterplots between the male and female sets of variables, where each health factor's strongest correlation was between the gendered counterparts, e.g. male and female rates of diagnosed diabetes were strongest with each other.

Pairwise correlation plots were generated based on which set each variable was in, as seen in Figure 4. Health factors tended to correlate best with each other, and the socioeconomic factors not regarding houses tended to correlate well with each other. The housing related factors tended not to correlate well with much at all, let alone between themselves. With regards to the correlation between diagnosed diabetes and the other factors, the strongest correlations were found in obesity and physical inactivity, and moderate correlations were found in the aforementioned non-housing related socioeconomic factors.

The correlation plots indicate that multicollinearity could be an issue for model building, thus techniques that reduce multicollinearity, such as Ridge regression and VIF analysis, should be considered.

Regression Analysis

The regression techniques applied here were the Ordinary Least Squares model via the lm() function in Base R, the Ordinary Least Squares model with an exhaustive subset of variables using the regsubsets() function from the leaps library to generate the subset plots for choosing a parsimonious model, the Ridge model found in the glmnet library, using cv.glmnet() to generate the appropriate lambda for use in the glmnet() function, with alpha = 0 to indicate the use of Ridge regression, and the LASSO model found in the glmnet library, following the same techniques as LASSO, but with alpha = 1 instead. The lambdas selected were the "one standard error" (1SE) lambdas. The models had the following performances:

	Male RMSE (Train)	Female RMSE (Train)
OLS	1.077109	1.053210
OLS Subset	1.081232	1.055452
Ridge	1.098473	1.077201
LASSO	1.093123	1.069711

	Male RMSE (Test)	Female RMSE (Test)
OLS	1.128351	1.080419
OLS Subset	1.127030	1.083424
Ridge	1.154599	1.114222
LASSO	1.151534	1.107140

As the training and testing models performed similarly, the ratios between all the test and train errors returned near 1, showing that the models are not overfitting. With this in mind, we find that the model with the lowest RMSE is the OLS model fully inclusive of all the variables. The model summary is as shown in Figure 5.

Principal Component Analysis (PCA)

Principal Component Analysis was performed to measure the effectiveness of dimension reduction within the data. The method used to generate the PCs was to pass the correlation matrix into the prcomp() function in the stats library, as opposed to passing the entire dataset into the function. This performs the same effect of passing a scaled version of the data into prcomp(). The table below describes the different PCs with respect to their proportions of variance.

	PC1	PC2	PC3	PC4
Male	0.5661	0.2289	0.1057	0.0624
Female	0.5303	0.2529	0.1105	0.0674

With 3 PCs, we achieve around 85-90% variance explained, which would be a 3:1 compression ratio with 10-15% loss. This is a fair tradeoff for the amount of compression offered. The scree plots of both genders' PCA models show that the knee is at 3 components, as shown in Figure 7. When performing selection by cumulative proportion of variance, we set our cutoff value around .90 ± .005, which generated the same result: selecting 3 components/factors, as shown in Figure 8.

Figure 6 (a) and Figure 6 (b) show the loadings of variables onto the top 3 PCs. The first PC is positively correlated with five socioeconomic variables and negatively correlated with health variables. The second PC is mostly highly correlated with the health variables. The third PC is mostly positively correlated with the percentage of severe housing cost burden.

Principal Factor Analysis (PFA)

Principal Factor Analysis was performed to find latent or hidden factors composed of the variables selected for the dataset. Figure 9 shows the fa.parallel function was used to provide the optimal number of factors. According to the screen plots (Figure 10) generated by analyzing the variance loadings of each potential factor, three factors were selected as optimal for maximizing variance capture while minimizing the number of factors.

Figure 11 shows the loadings of three factors with a cut off value of 0.4. We can tell from the output that the variables are distinctly sepatreated into three factors. RC1 captures around 0.38 variance, RC 2 captures about 0.2 variance and RC3 captures about 0.13 variance. RC1 is highly correlated with five socioeconomic variables. RC2 is highly correlated with health related variables and RC3 is highly correlated with housing related variables.

PFA suggested that there are three main contributing factors to the percentage of diabetes rate in male and female population. The first factor is related to the socioeconomic status of the region, the lower the socioeconomic status is, the diabetes rate is likley to be higher for the population. Similarly, the second factor is related to the health of the demographic. The higher

obese rate and higher inactive rate contributes to the higher diabetes rate. The third factor is related to the housing status of the demographic. This reveals a hidden factor that contributes to diabetes rate. There was no direct connection between the housing burden or long commute to the diabetes rate. But it indeed contributes to the likelihood of diabetes.

Principal Component Regression

In previous principal factor analysis, the fa.parallel screen plots generated optimal number of principal components, which was found to be three components. This became the basis for the number of principal components to use for testing the effectiveness of regression through pure dimension reduction.

The below table consists of the beta values for PC1, PC2, PC3, with the adjusted R^2 score and F-test of the principal component regression. Further information can be found in Figure 12.

	β_{PC1}	β_{PC2}	β_{PC3}*	Adj-R^2	p(F)
Male	-0.027683	+0.161586	-0.006763	0.4646	< 2.2e-16
Female	-0.013618	+0.162722	+0.007013	0.5432	< 2.2e-16

β_{PC3} is marked to indicate that it failed its t-test for significance, which is true for both the male and female variants.

The dimension reduction regression captures 46-54% of the variance in diagnosed diabetes, which runs very close (between 3-12%) to the amount of variance captured by the ordinary least squares models without dimension reduction.

Principal Factor Regression

To analyze the effectiveness of the possible composite factors in the principal factor analysis, an ordinary least squares regression was performed with the factors RC1, RC2, and RC3 against their respective gender specific rate of diagnosed diabetes.

Furthermore, each factor was decomposed into its primary variables, as determined by the loading charts, and regression using each set of primary variables was used to determine the importance of each factor towards capturing the most variance in diagnosed diabetes.

Shown below are the values for the betas for RC1, RC2, RC3, with the adjusted R^2 score and F-test of the principal factor regression. Further information can be found in Figure 13.

	β_{RC1}	β_{RC2}	β_{RC3}	Adj-R^2	p(F)
Male	+0.48447	+1.16703	+0.11464	0.5753	< 2.2e-16

Female	+0.60803	+1.23576	+0.17058	0.6272	< 2.2e-16

All betas within this set of regressions did not fail the t-test for significance.

The amount of variance captured by this regression performed on par or better than the ordinary least squares regression without the factor analysis, capturing 58-62% of the variance.

Per Principal Factor Regression

Shown below are the values for the betas for variables within RC1 with the adjusted R^2 score and F-test of the RC1 regression. Further information can be found in Figure 14.

RC1	$\beta_{N.Inet}$	$\beta_{ChilPov}$	$\beta_{FRLunch}$	β_{FInsec}	$\beta_{NHInsur}$	Adj-R^2	p(F)
Male	-0.028927	+0.027034	+0.018678	+0.110954	-0.035351	0.1642	< 2.2e-16
Female	-0.031652	+0.047197	+0.025053	+0.080976	-0.022668	0.2285	< 2.2e-16

Shown below are the values for the betas for variables within RC2 with the adjusted R^2 score and F-test of the RC2 regression. Further information can be found in Figure 15.

RC2	β_{Obes}	β_{Inac}	Adj-R^2	p(F)
Male	+0.076797	+0.203136	0.5101	< 2.2e-16
Female	+0.073641	+0.199550	0.5522	< 2.2e-16

Shown below are the values for the betas for variables within RC3 with the adjusted R^2 score and F-test of the RC3 regression. Further information can be found in Figure 16.

RC3	β_{SHCB}	$\beta_{LCommute}$**	Adj-R^2	p(F)
Male	+0.0564034	-0.0008949	0.01325	2.021e-08
Female	+0.0927617	-0.0000665	0.03394	< 2.2e-16

$\beta_{LCommute}$ is double marked to indicate how extremely it failed its t-test in both sets of data. All other betas succeeded the t-test with an alpha of .05.

The variance captured by the variables in RC2, Obesity and Physical Inactivity, is 51-55%, which is most of the variance from the PFA regression, and the variables in RC1 captures 16-22% of the variance of diagnosed diabetes. RC3 captures only 1-3% of the variance.

FINAL MODEL

The regression model chosen to be the final model is the subset model due to the parsimonious quality while minimizing losses in adjusted R^2. The independent variables selected are percentage of obese, percentage of inactive, percentage of child poverty, percentage of free lunch and percentage of severe housing cost burden for both female and male models.

CONCLUSION

With our findings in the Per Factor Regression section, we find that the most impactful factor is the health factors in RC2, as they encompass the majority of the variation in the regression, despite the fact that the exploratory factor analysis shows that they share a lesser proportion of variance than the socioeconomic factors during their rotation into factors. RC1 variables contributed to 16-20% of the variance of diagnosed diabetes, where RC2 variables contributed to 51-55% of the variance. A model which would select these variables would then produce a high quality regression predicting the rate of diabetes within the United States. As the variables selected in the parsimonious model fall into these categories as their top contributors, we can confirm that this model captures the right variables. Due to the numerous similarities between the male and female regression models with regards to variables selected all the way up to the loadings of each variable into each factor, there is not much ability for the models to be able to discriminate between males and females with respect to diagnosed diabetes. This implies that gender has no effect in terms of which factors are effective predictors in predicting the rate of diagnosed diabetes.

DISCUSSION

Other data, such as data based on race and age groups, might be helpful in analysis towards predicting more specific rates of diabetes in populations. Furthermore, the socioeconomic data did not have gender splits when retrieved from the CDC website. Perhaps the regressions by gender would capture higher variance if the splits were available.

Another idea to test the difference between the two sets of regressions might be to perform a goodness of fit test to see if there was a significant difference. This was not performed due to lack of technical ability, but it might be an interesting result to find. We could also perform this test between the gender-specific health factors.

The data provided on the website makes it possible to perform time-series analysis, and it would be interesting to see if past observations might lend itself to performing differently than current observations in the regression models, or discovering trends that might lead to a shift in how each variable contributed to the prediction of diagnosed diabetes.

It would be insightful to see if reshaping the data to include labels of gender, age, race, or even average salary (binned into categories) would allow for the creation of classification models such as k-means, decision trees, SVM, etc. Furthermore, it would be interesting to see how it would cluster together, to see if a clustering algorithm could pick out these factors. In our analysis, we did plot the PCA and PFA rotations, but included in those rotations were the non-specific socioeconomic factors, and as the socioeconomic factors were selected as the highest proportion of variance, this might have muted the effects of gender in the plot, which might have obscured the ability to cluster.

R CODE

```
## install libraries on first time run only
install.packages(c('tidyverse', 'corrplot',
                   'psych', 'stats', 'caret', 'MASS',
                   'glmnet', 'leaps', 'factoextra',
                   'car'))
library(tidyverse)
library(corrplot)
library(psych)
library(stats)
library(caret)
library(MASS)
library(glmnet)
library(leaps)
library(factoextra)
library(car)

## Initial preprocessing
diabetes <- read.csv("~/Downloads/diabetes.csv", stringsAsFactors=T)
diabetes <- na.omit(diabetes)
lend <- nrow(diabetes)
dap <- diabetes %>% dplyr::select(c(A.Percent.N.Inet, A.Percent.ChilPov,
                                    A.Percent.FRLunch, A.Percent.FInsec,
                                    A.Percent.NHInsur, A.Percent.SHCB,
                                    A.Percent.LCommute))
dmp <- diabetes %>% dplyr::select(c(M.Percent.DiagDia, M.Percent.Obes,
M.Percent.Inac))
dfp <- diabetes %>% dplyr::select(c(F.Percent.DiagDia, F.Percent.Obes,
F.Percent.Inac))

## Binding gender based groups by column
dm <- cbind(dmp, dap)
df <- cbind(dfp, dap)

#visualization

summary(dm)
summary(df)


boxplot(dm$M.Percent.DiagDia, df$F.Percent.DiagDia,
        at = c(1,2), main = "DiagDia",
        names = c("male", "female"),horizontal = TRUE,
        notch = TRUE
)

boxplot(dm$M.Percent.Obes,df$F.Percent.Obes,main = "Obes",at =
c(1,2),names=c("male","female"),
```

```
horizontal=TRUE,notch=TRUE)



boxplot(dm$M.Percent.Inac,df$F.Percent.Inac, main= "Inactive",at =
c(1,2),names=c("male","female"),
        horizontal=TRUE,notch=TRUE)

## Evaluating scatters and multicollinear data
plot(cbind(dmp, dfp))
plot(diabetes$M.Percent.DiagDia, diabetes$F.Percent.DiagDia)
plot(diabetes$M.Percent.Inac, diabetes$F.Percent.Inac)
plot(diabetes$M.Percent.Obes, diabetes$F.Percent.Obes)
m_cor_pear <- cor(dm[,-1], method = 'pearson')
f_cor_pear <- cor(df[,-1], method = 'pearson')
m_cor_spear <- cor(dm[,-1], method = 'spearman')
f_cor_spear <- cor(df[,-1], method = 'spearman')
m_cor_kend <- cor(dm[,-1], method = 'kendall')
f_cor_kend <- cor(df[,-1], method = 'kendall')

## Evaluating correlation method to find factors
i_kmo <- KMO(m_cor_spear)
m_kmo <- data.frame(pear = KMO(m_cor_pear)$MSAi,
                    spear = KMO(m_cor_spear)$MSAi,
                    kend = KMO(m_cor_kend)$MSAi)
m_kmo["MSA Overall",] <- c(KMO(m_cor_pear)$MSA,
                           KMO(m_cor_spear)$MSA,
                           KMO(m_cor_kend)$MSA)
f_kmo <- data.frame(pear = KMO(f_cor_pear)$MSAi,
                    spear = KMO(f_cor_spear)$MSAi,
                    kend = KMO(f_cor_kend)$MSAi)
f_kmo["MSA Overall",] <- c(KMO(f_cor_pear)$MSA,
                           KMO(f_cor_spear)$MSA,
                           KMO(f_cor_kend)$MSA)
m_kmo
f_kmo
corrplot(m_cor_pear, method = 'ellipse')
corrplot(f_cor_pear, method = 'ellipse')
corrplot(m_cor_spear, method = 'ellipse')
corrplot(f_cor_spear, method = 'ellipse')
corrplot(m_cor_kend, method = 'ellipse')
corrplot(f_cor_kend, method = 'ellipse')

## PCA/Evaluating Factor Count
m_pca <- prcomp(m_cor_pear)
f_pca <- prcomp(f_cor_pear)
summary(m_pca)
```

```
summary(f_pca)
plot(m_pca)
plot(f_pca)
fa.parallel(m_cor_pear, n.obs = lend)
fa.parallel(f_cor_pear, n.obs = lend)
# results: 3 factors optimal.

## PFA/Evaluating Factors
m_pfa <- principal(m_cor_pear, nfactors = 3, n.obs = lend)
f_pfa <- principal(f_cor_pear, nfactors = 3, n.obs = lend)
m_pfa
f_pfa
print(m_pfa$loadings, cutoff = .4)
print(f_pfa$loadings, cutoff = .4)
# results:
# RC1: Lack of common human needs (Internet, Food, Money, Insurance)
# RC2: Health factors (obesity, inactivity, diabetes)
# RC3: Housing convenience (Housing costs and commute to work)

## Maximum likelihood factor analysis
m_mlfa <- fa(m_cor_pear, nfactors = 3, n.obs = lend, fm = 'ml')
f_mlfa <- fa(f_cor_pear, nfactors = 3, n.obs = lend, fm = 'ml')

print(m_mlfa$loadings, cutoff = .4)
print(f_mlfa$loadings, cutoff = .4)

## Evaluating factors' goodness of fit
m_pfa_fit <- factor.stats(m_cor_pear, m_pfa, n.obs = lend, alpha = .05)
f_pfa_fit <- factor.stats(f_cor_pear, f_pfa, n.obs = lend, alpha = .05)
m_pfa_fit$RMSEA
m_pfa_fit
f_pfa_fit$RMSEA
f_pfa_fit
# results:
# Chi-Sq test returns sufficient, likely from the volume of observations.
# RMSEA returns .17-.20 ish (low fit) meaning the model doesn't fit very well.
# Let's look at the plots.

## Plotting PF/PCs
m_pfp <- as.data.frame(predict(m_pfa, dm))
f_pfp <- as.data.frame(predict(f_pfa, df))
pfplot <- ggplot() + geom_point(data = m_pfp, aes(x = RC1, y = RC2), alpha = 0.3, col
= 'green') +
  geom_point(data = f_pfp, aes(x = RC1, y = RC2), alpha = 0.3, col = 'red')
pfplot
m_pcp <- as.data.frame(predict(m_pca, dm))
f_pcp <- as.data.frame(predict(f_pca, df))
```

```
pcplot <- ggplot() + geom_point(data = m_pcp, aes(x = PC1, y = PC2), alpha = 0.3, col
= 'green') +
  geom_point(data = f_pcp, aes(x = PC1, y = PC2), alpha = 0.3, col = 'red')
pcplot
m_mlp <- as.data.frame(predict(m_mlfa, dm))
f_mlp <- as.data.frame(predict(f_mlfa, df))
mlplot <- ggplot() + geom_point(data = m_mlp, aes(x = ML1, y = ML2), alpha = 0.3, col
= 'green') +
  geom_point(data = f_mlp, aes(x = ML1, y = ML2), alpha = 0.3, col = 'red')
mlplot

## Regression Preprocessing
dm_i <- cbind(CountyFIPS = diabetes$CountyFIPS, dm)
df_i <- cbind(CountyFIPS = diabetes$CountyFIPS, df)
set.seed(123)
ind <- createDataPartition(dm_i$CountyFIPS, p = 0.8, list = F)
dm_train <- dm_i[ind,] %>% dplyr::select(-c(CountyFIPS))
dm_test <- dm_i[-ind,] %>% dplyr::select(-c(CountyFIPS))
dm_XTr <- as.matrix(dm_train[,-1])
dm_YTr <- as.matrix(dm_train[,1])
dm_XTs <- as.matrix(dm_test[,-1])
df_train <- df_i[ind,] %>% dplyr::select(-c(CountyFIPS))
df_test <- df_i[-ind,] %>% dplyr::select(-c(CountyFIPS))
df_XTr <- as.matrix(df_train[,-1])
df_YTr <- as.matrix(df_train[,1])
df_XTs <- as.matrix(df_test[,-1])
lrange <- seq(0, 10, .01)

## Regression analysis
m_subset <- regsubsets(M.Percent.DiagDia ~ ., data = dm_train, nvmax = 9,
                       method = c('exhaustive', 'forward', 'backward'))
f_subset <- regsubsets(F.Percent.DiagDia ~ ., data = df_train, nvmax = 9,
                       method = c('exhaustive', 'forward', 'backward'))
plot(m_subset, scale = 'adjr2')
plot(f_subset, scale = 'adjr2')

ols_m_dd <- lm(M.Percent.DiagDia  ~ ., data = dm_train)
summary(ols_m_dd)
vif(ols_m_dd)
#plot(ols_m_dd)
ols_f_dd <- lm(F.Percent.DiagDia  ~ ., data = df_train)
summary(ols_f_dd)
vif(ols_f_dd)
#plot(ols_f_dd)
ols_ss_m <- lm(M.Percent.DiagDia ~ M.Percent.Obes + M.Percent.Inac +
               A.Percent.ChilPov + A.Percent.FRLunch + A.Percent.SHCB,
               data = dm_train)
summary(ols_ss_m)
```

```
vif(ols_ss_m)
#plot(ols_ss_m)
ols_ss_f <- lm(F.Percent.DiagDia ~ F.Percent.Obes + F.Percent.Inac +
               A.Percent.ChilPov + A.Percent.FRLunch + A.Percent.SHCB,
               data = df_train)
summary(ols_ss_f)
vif(ols_ss_f)
#plot(ols_ss_f)
ridge_m_dd <- cv.glmnet(dm_XTr, dm_YTr, alpha = 0, lambda = lrange)
ridge_f_dd <- cv.glmnet(df_XTr, df_YTr, alpha = 0, lambda = lrange)
lasso_m_dd <- cv.glmnet(dm_XTr, dm_YTr, alpha = 1, lambda = lrange)
lasso_f_dd <- cv.glmnet(df_XTr, df_YTr, alpha = 1, lambda = lrange)

plot(ridge_m_dd)
plot(lasso_m_dd)
plot(ridge_f_dd)
plot(lasso_f_dd)
rm1se <- ridge_m_dd$lambda.1se
lm1se <- lasso_m_dd$lambda.1se
rf1se <- ridge_f_dd$lambda.1se
lf1se <- lasso_f_dd$lambda.1se

ridge_m_1se <- glmnet(dm_XTr, dm_YTr, alpha = 0, lambda = rm1se)
ridge_m_1se$beta
lasso_m_1se <- glmnet(dm_XTr, dm_YTr, alpha = 1, lambda = lm1se)
lasso_m_1se$beta
ridge_f_1se <- glmnet(df_XTr, df_YTr, alpha = 0, lambda = rf1se)
ridge_f_1se$beta
lasso_f_1se <- glmnet(df_XTr, df_YTr, alpha = 1, lambda = lf1se)
lasso_f_1se$beta

## Regression Evaluation
train_avp <- data.frame(m_actual = dm_YTr, f_actual = df_YTr)
train_avp$m_ols <- predict(ols_m_dd, as.data.frame(dm_XTr))
train_avp$f_ols <- predict(ols_f_dd, as.data.frame(df_XTr))
train_avp$m_ols_ss <- predict(ols_ss_m, as.data.frame(dm_XTr))
train_avp$f_ols_ss <- predict(ols_ss_f, as.data.frame(df_XTr))
train_avp$m_ridge <- predict(ridge_m_1se, dm_XTr)
train_avp$f_ridge <- predict(ridge_f_1se, df_XTr)
train_avp$m_lasso <- predict(lasso_m_1se, dm_XTr)
train_avp$f_lasso <- predict(lasso_f_1se, df_XTr)
plot(train_avp$m_actual, train_avp$m_ols)
plot(train_avp$f_actual, train_avp$f_ols)
plot(train_avp$m_actual, train_avp$m_ols_ss)
plot(train_avp$f_actual, train_avp$f_ols_ss)
plot(train_avp$m_actual, train_avp$m_ridge)
plot(train_avp$f_actual, train_avp$f_ridge)
plot(train_avp$m_actual, train_avp$m_lasso)
```

```
plot(train_avp$f_actual, train_avp$f_lasso)
test_avp <- data.frame(m_actual = dm_test$M.Percent.DiagDia,
                       f_actual = df_test$F.Percent.DiagDia)
test_avp$m_ols <- predict(ols_m_dd, as.data.frame(dm_XTs))
test_avp$f_ols <- predict(ols_f_dd, as.data.frame(df_XTs))
test_avp$m_ols_ss <- predict(ols_ss_m, as.data.frame(dm_XTs))
test_avp$f_ols_ss <- predict(ols_ss_f, as.data.frame(df_XTs))
test_avp$m_ridge <- predict(ridge_m_1se, dm_XTs)
test_avp$f_ridge <- predict(ridge_f_1se, df_XTs)
test_avp$m_lasso <- predict(lasso_m_1se, dm_XTs)
test_avp$f_lasso <- predict(lasso_f_1se, df_XTs)
plot(test_avp$m_actual, test_avp$m_ols)
plot(test_avp$f_actual, test_avp$f_ols)
plot(test_avp$m_actual, test_avp$m_ols_ss)
plot(test_avp$f_actual, test_avp$f_ols_ss)
plot(test_avp$m_actual, test_avp$m_ridge)
plot(test_avp$f_actual, test_avp$f_ridge)
plot(test_avp$m_actual, test_avp$m_lasso)
plot(test_avp$f_actual, test_avp$f_lasso)

train_acc <- c(
sqrt(mean((train_avp$m_actual - train_avp$m_ols)^2)),
sqrt(mean((train_avp$f_actual - train_avp$f_ols)^2)),
sqrt(mean((train_avp$m_actual - train_avp$m_ols_ss)^2)),
sqrt(mean((train_avp$f_actual - train_avp$f_ols_ss)^2)),
sqrt(mean((train_avp$m_actual - train_avp$m_ridge)^2)),
sqrt(mean((train_avp$f_actual - train_avp$f_ridge)^2)),
sqrt(mean((train_avp$m_actual - train_avp$m_lasso)^2)),
sqrt(mean((train_avp$f_actual - train_avp$f_lasso)^2))
)

test_acc <- c(
sqrt(mean((test_avp$m_actual - test_avp$m_ols)^2)),
sqrt(mean((test_avp$f_actual - test_avp$f_ols)^2)),
sqrt(mean((test_avp$m_actual - test_avp$m_ols_ss)^2)),
sqrt(mean((test_avp$f_actual - test_avp$f_ols_ss)^2)),
sqrt(mean((test_avp$m_actual - test_avp$m_ridge)^2)),
sqrt(mean((test_avp$f_actual - test_avp$f_ridge)^2)),
sqrt(mean((test_avp$m_actual - test_avp$m_lasso)^2)),
sqrt(mean((test_avp$f_actual - test_avp$f_lasso)^2))
)

trvts_stats <- data.frame(reg_method = colnames(train_avp)[3:10], train_acc =
train_acc, test_acc = test_acc)
## If Test < Train -> negative, else if Test > Train, Positive.
## Want Ratio <= 0 to prevent overfit.
trvts_stats$ratio <- (trvts_stats$test_acc/trvts_stats$train_acc)
trvts_stats
```

```
# results: (lowest to highest ratio, based on proximity to 1)
# LASSO < Ridge < subset OLS < full set OLS

## PCA/PFA Regression: (probably PCA best)
ind_dm <- dm_train[,-1]
ind_df <- df_train[,-1]
res_dm <- dm_train[,1]
res_df <- df_train[,1]
ind_mpca <- prcomp(cor(ind_dm))
ind_fpca <- prcomp(cor(ind_df))
ind_mpca
ind_fpca
fa.parallel(cor(ind_dm), n.obs = lend)
fa.parallel(cor(ind_df), n.obs = lend)
ind_mpfa <- principal(cor(ind_dm), nfactors = 3, n.obs = lend)
ind_fpfa <- principal(cor(ind_df), nfactors = 3, n.obs = lend)
print(ind_mpfa$loadings, cutoff = .4)
print(ind_fpfa$loadings, cutoff = .4)
ind_rot_m <- as.data.frame(predict(ind_mpca, ind_dm))
ind_rot_f <- as.data.frame(predict(ind_fpca, ind_df))
ind_var_m <- as.data.frame(predict(ind_mpfa, ind_dm))
ind_var_f <- as.data.frame(predict(ind_fpfa, ind_df))
ind_rot_m
ind_rot_f
reg_pca_dm <- cbind(res_dm, ind_rot_m)
reg_pca_df <- cbind(res_df, ind_rot_f)
reg_pfa_dm <- cbind(res_dm, ind_var_m)
reg_pfa_df <- cbind(res_df, ind_var_f)

pc_ols_dm <- lm(res_dm ~ PC1 + PC2 + PC3, data = reg_pca_dm)
pc_ols_df <- lm(res_df ~ PC1 + PC2 + PC3, data = reg_pca_df)
pf_ols_dm <- lm(res_dm ~ RC1 + RC2 + RC3, data = reg_pfa_dm)
pf_ols_df <- lm(res_df ~ RC1 + RC2 + RC3, data = reg_pfa_df)
rc1_ols_dm <- lm(M.Percent.DiagDia ~ A.Percent.N.Inet + A.Percent.ChilPov +
                 A.Percent.FRLunch + A.Percent.FInsec + A.Percent.NHInsur,
                 data = dm_train)
rc1_ols_df <- lm(F.Percent.DiagDia ~ A.Percent.N.Inet + A.Percent.ChilPov +
                 A.Percent.FRLunch + A.Percent.FInsec + A.Percent.NHInsur,
                 data = df_train)
rc2_ols_dm <- lm(M.Percent.DiagDia ~ M.Percent.Obes + M.Percent.Inac,
                 data = dm_train)
rc2_ols_df <- lm(F.Percent.DiagDia ~ F.Percent.Obes + F.Percent.Inac,
                 data = df_train)
rc3_ols_dm <- lm(M.Percent.DiagDia ~ A.Percent.SHCB + A.Percent.LCommute,
                 data = dm_train)
rc3_ols_df <- lm(F.Percent.DiagDia ~ A.Percent.SHCB + A.Percent.LCommute,
                 data = df_train)
summary(pc_ols_dm)
```

```
summary(pc_ols_df)
summary(pf_ols_dm)
summary(pf_ols_df)

summary(rc1_ols_dm)
#plot(rc1_ols_dm)
summary(rc1_ols_df)
#plot(rc1_ols_df)
summary(rc2_ols_dm)
#plot(rc2_ols_dm)
summary(rc2_ols_df)
#plot(rc2_ols_df)
summary(rc3_ols_dm)
#plot(rc3_ols_dm)
summary(rc3_ols_df)
#plot(rc3_ols_df)

## Sure, let's do clusters.
## Pre-processing (scaling/centering)
dm_preproc <- preProcess(dm, method = c('center','scale'))
dm_p <- predict(dm_preproc, dm)
dm_p

## Determining cluster count
fviz_nbclust(dm_p, FUN = kmeans, method = 'wss')
fviz_nbclust(dm_p, FUN = kmeans, method = 'silhouette')
# results:
# silhouette returns 2 clusters
# knee returns 4 clusters
# might as well try both while we're here.

## KMeans clustering:
dm_kmeans_2 <- kmeans(dm_p, centers = 2, nstart = 25)
dm_kmeans_4 <- kmeans(dm_p, centers = 4, nstart = 25)
m_pfp$cluster2 <- as.factor(dm_kmeans_2$cluster)
m_pfp$cluster4 <- as.factor(dm_kmeans_4$cluster)
ggplot() + geom_point(data = m_pfp, aes(x = RC1, y = RC2, col = cluster2), alpha =
0.3)
ggplot() + geom_point(data = m_pfp, aes(x = RC1, y = RC2, col = cluster4), alpha =
0.3)

# not sure what each cluster could mean.
# That said, each cluster is determined by distance with 9 variables. Maybe HAC?

fviz_nbclust(dm_p, FUN = hcut, method = 'wss')
fviz_nbclust(dm_p, FUN = hcut, method = 'silhouette')

dm_p.e.dist <- dist(dm_p, method = 'euclidean')
```

```
dm_p.m.dist <- dist(dm_p, method = 'manhattan')
dm_p.e1 <- hclust(dm_p.e.dist, method = 'ward.D')
dm_p.e2 <- hclust(dm_p.e.dist, method = 'complete')
dm_p.m1 <- hclust(dm_p.m.dist, method = 'ward.D')
dm_p.m2 <- hclust(dm_p.m.dist, method = 'complete')
plot(dm_p.e1)
plot(dm_p.e2)
plot(dm_p.m1)
plot(dm_p.m2)

dm.e1 <- cutree(dm_p.e1, k = 4)
dm.e2 <- cutree(dm_p.e2, k = 4)
dm.m1 <- cutree(dm_p.m1, k = 4)
dm.m2 <- cutree(dm_p.m2, k = 4)

m_pfp$e1 <- as.factor(dm.e1)
m_pfp$e2 <- as.factor(dm.e2)
m_pfp$m1 <- as.factor(dm.m1)
m_pfp$m2 <- as.factor(dm.m2)

ggplot() + geom_point(data = m_pfp, aes(x = RC1, y = RC2, col = e1), alpha = 0.3)
ggplot() + geom_point(data = m_pfp, aes(x = RC1, y = RC2, col = e2), alpha = 0.3)
ggplot() + geom_point(data = m_pfp, aes(x = RC1, y = RC2, col = m1), alpha = 0.3)
ggplot() + geom_point(data = m_pfp, aes(x = RC1, y = RC2, col = m2), alpha = 0.3)
```

APPENDIX OF FIGURES

```
## Initial preprocessing
diabetes <- read.csv("~/Downloads/diabetes.csv", stringsAsFactors=T)
diabetes <- na.omit(diabetes)
lend <- nrow(diabetes)
dap <- diabetes %>% dplyr::select(c(A.Percent.N.Inet, A.Percent.ChilPov,
                                    A.Percent.FRLunch, A.Percent.FInsec,
                                    A.Percent.NHInsur, A.Percent.SHCB,
                                    A.Percent.LCommute))
dmp <- diabetes %>% dplyr::select(c(M.Percent.DiagDia, M.Percent.Obes, M.Percent.Inac))
dfp <- diabetes %>% dplyr::select(c(F.Percent.DiagDia, F.Percent.Obes, F.Percent.Inac))

## Binding gender based groups by column
dm <- cbind(dmp, dap)
df <- cbind(dfp, dap)
```

Figure 1 - R Pre-Processing Steps

```
> summary(dm)
 M.Percent.DiagDia M.Percent.Obes  M.Percent.Inac A.Percent.N.Inet A.Percent.ChilPov A.Percent.FRLunch
 Min.   : 4.500    Min.   :11.40   Min.   : 9.1   Min.   : 2.50    Min.   : 2.60     Min.   :  1.20
 1st Qu.: 7.900    1st Qu.:23.10   1st Qu.:15.7   1st Qu.:14.80    1st Qu.:12.60     1st Qu.: 42.40
 Median : 8.900    Median :28.20   Median :18.5   Median :19.50    Median :17.60     Median : 53.00
 Mean   : 9.169    Mean   :27.91   Mean   :18.9   Mean   :20.54    Mean   :18.69     Mean   : 54.38
 3rd Qu.:10.100    3rd Qu.:32.40   3rd Qu.:21.6   3rd Qu.:25.10    3rd Qu.:23.40     3rd Qu.: 64.10
 Max.   :16.300    Max.   :46.40   Max.   :38.2   Max.   :59.60    Max.   :59.70     Max.   :100.00
 A.Percent.FInsec A.Percent.NHInsur A.Percent.SHCB  A.Percent.LCommute
 Min.   : 2.90    Min.   : 0.700    Min.   : 0.80   Min.   : 0.200
 1st Qu.:10.50    1st Qu.: 5.800    1st Qu.: 8.40   1st Qu.: 4.500
 Median :12.90    Median : 8.700    Median :10.20   Median : 6.800
 Mean   :13.09    Mean   : 9.618    Mean   :10.67   Mean   : 8.079
 3rd Qu.:15.57    3rd Qu.:12.100    3rd Qu.:12.60   3rd Qu.:10.300
 Max.   :29.40    Max.   :40.900    Max.   :30.30   Max.   :38.700
> summary(df)
 F.Percent.DiagDia F.Percent.Obes  F.Percent.Inac  A.Percent.N.Inet A.Percent.ChilPov A.Percent.FRLunch
 Min.   : 3.800    Min.   :10.60   Min.   : 9.40   Min.   : 2.50    Min.   : 2.60     Min.   :  1.20
 1st Qu.: 6.800    1st Qu.:22.90   1st Qu.:17.50   1st Qu.:14.80    1st Qu.:12.60     1st Qu.: 42.40
 Median : 7.800    Median :27.90   Median :20.60   Median :19.50    Median :17.60     Median : 53.00
 Mean   : 8.135    Mean   :27.83   Mean   :21.08   Mean   :20.54    Mean   :18.69     Mean   : 54.38
 3rd Qu.: 9.100    3rd Qu.:32.30   3rd Qu.:24.00   3rd Qu.:25.10    3rd Qu.:23.40     3rd Qu.: 64.10
 Max.   :16.400    Max.   :52.50   Max.   :39.90   Max.   :59.60    Max.   :59.70     Max.   :100.00
 A.Percent.FInsec A.Percent.NHInsur A.Percent.SHCB  A.Percent.LCommute
 Min.   : 2.90    Min.   : 0.700    Min.   : 0.80   Min.   : 0.200
 1st Qu.:10.50    1st Qu.: 5.800    1st Qu.: 8.40   1st Qu.: 4.500
 Median :12.90    Median : 8.700    Median :10.20   Median : 6.800
 Mean   :13.09    Mean   : 9.618    Mean   :10.67   Mean   : 8.079
 3rd Qu.:15.57    3rd Qu.:12.100    3rd Qu.:12.60   3rd Qu.:10.300
 Max.   :29.40    Max.   :40.900    Max.   :30.30   Max.   :38.700
```

Figure 2 - Summary of Variables

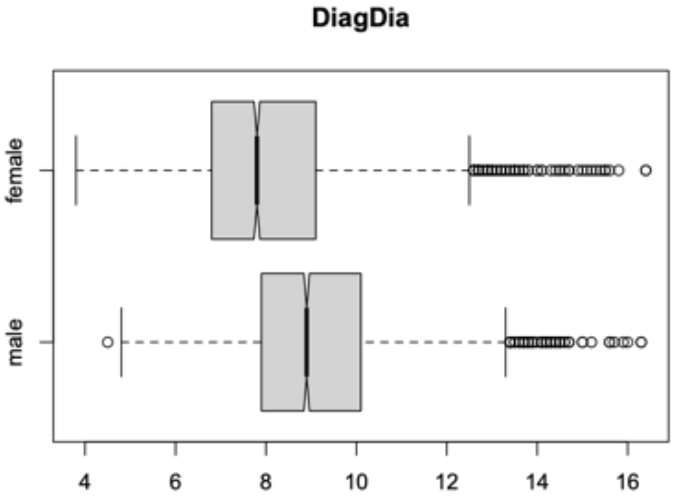

Figure 3 (a) - Diagnosed Diabetes Boxplot

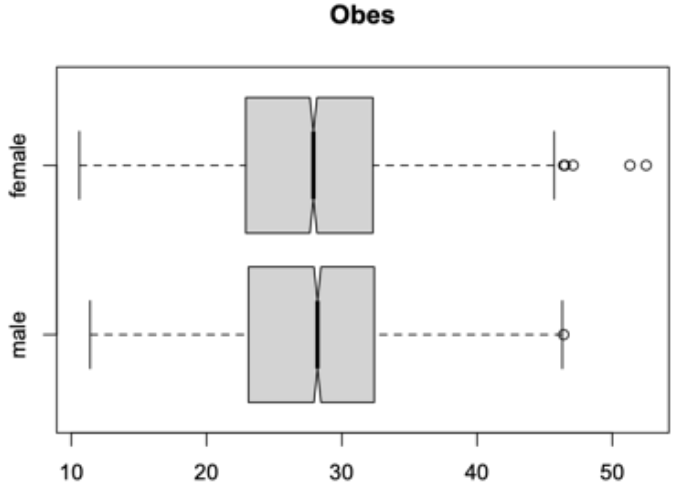

Figure 3 (b) - Obesity Boxplot

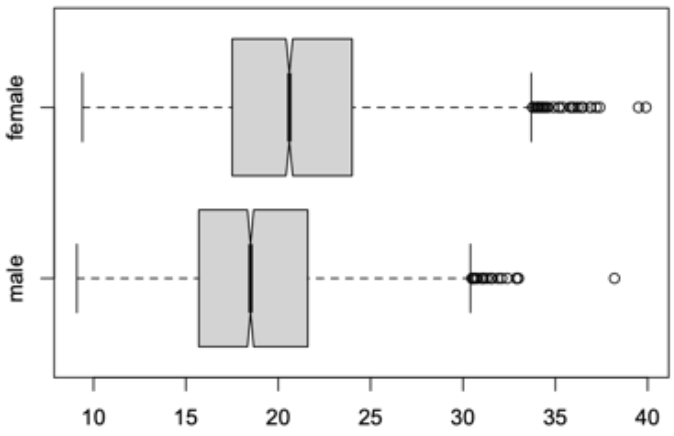

Figure 3 (c) - Physical Inactivity Boxplot

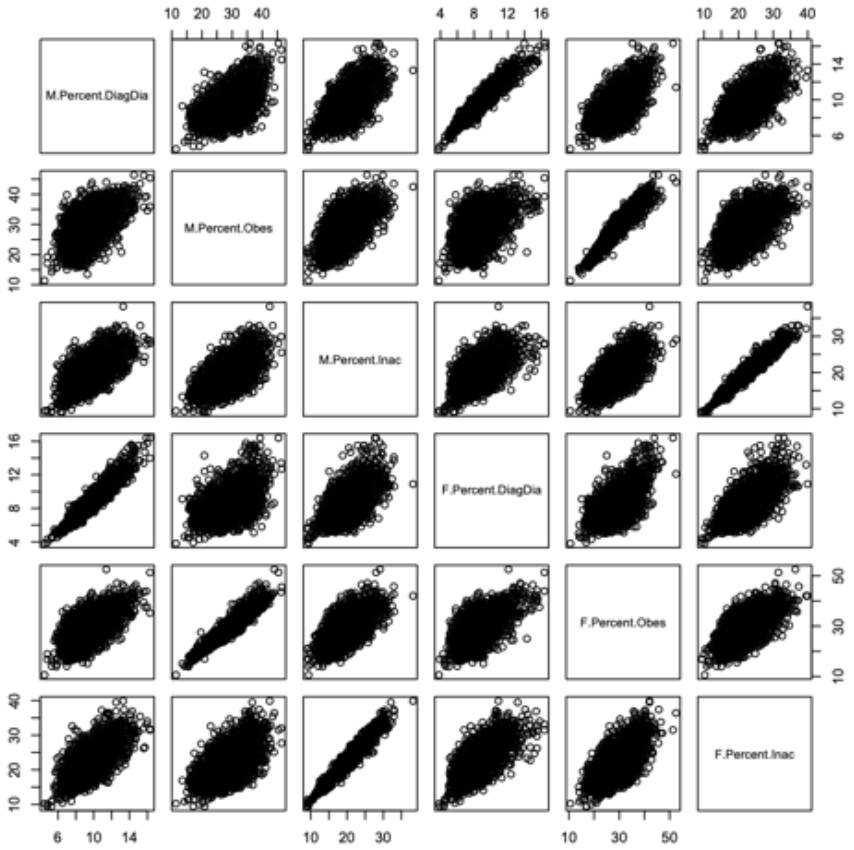

Figure 3 (d) - Gendered scatterplot

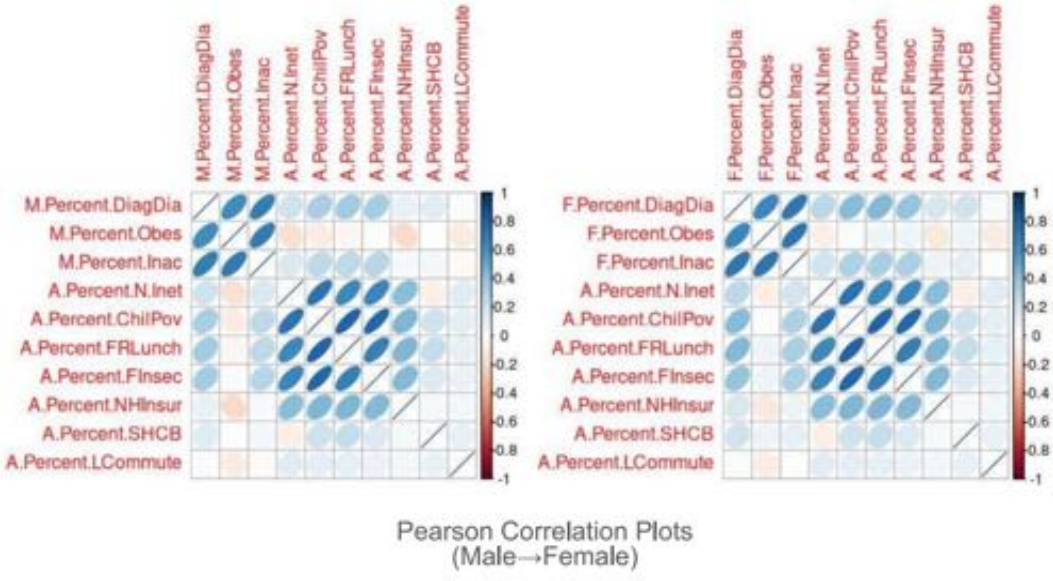

Figure 4 - Gender split correlation plots

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.744863   0.157372  11.087  < 2e-16 ***
M.Percent.Obes  0.118741   0.005559  21.361  < 2e-16 ***
M.Percent.Inac  0.125223   0.008337  15.019  < 2e-16 ***
A.Percent.N.Inet -0.004371  0.004318  -1.012  0.31148
A.Percent.ChilPov 0.030809  0.005702   5.404 7.15e-08 ***
A.Percent.FRLunch 0.011005  0.001971   5.585 2.59e-08 ***
A.Percent.FInsec  0.026407  0.009720   2.717  0.00664 **
A.Percent.NHInsur 0.003826  0.005067   0.755  0.45030
A.Percent.SHCB    0.017001  0.006884   2.470  0.01359 *
A.Percent.LCommute 0.011423  0.004452   2.566  0.01035 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.089 on 2500 degrees of freedom
Multiple R-squared:  0.5847,    Adjusted R-squared:  0.5832
F-statistic:   391 on 9 and 2500 DF,  p-value: < 2.2e-16
```

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.5501699  0.1437365   3.828 0.000133 ***
F.Percent.Obes  0.1133409  0.0053297  21.266  < 2e-16 ***
F.Percent.Inac  0.1194575  0.0074432  16.049  < 2e-16 ***
A.Percent.N.Inet 0.0014374 0.0041938   0.343 0.731810
A.Percent.ChilPov 0.0404959 0.0055217  7.334 3.00e-13 ***
A.Percent.FRLunch 0.0118260 0.0019206  6.157 8.59e-10 ***
A.Percent.FInsec  0.0003755 0.0094230  0.040 0.968220
A.Percent.NHInsur 0.0108599 0.0049026  2.215 0.026841 *
A.Percent.SHCB    0.0282008 0.0066721  4.227 2.46e-05 ***
A.Percent.LCommute 0.0088635 0.0043286  2.048 0.040697 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.057 on 2500 degrees of freedom
Multiple R-squared:  0.6425,    Adjusted R-squared:  0.6412
F-statistic: 499.2 on 9 and 2500 DF,  p-value: < 2.2e-16
```

Figure 5 - OLS Full, M/F

	PC1	PC2	PC3
M.Percent.Obes	-0.416229905	0.48472284	0.02365720
M.Percent.Inac	-0.211043816	0.54034125	0.01880290
A.Percent.N.Inet	0.444558557	0.18993433	0.23001581
A.Percent.ChilPov	0.428631567	0.17956436	-0.12532677
A.Percent.FRLunch	0.378584049	0.16454563	-0.18745223
A.Percent.FInsec	0.376426446	0.22778046	-0.08742495
A.Percent.NHInsur	0.336920332	-0.02126019	0.08737099
A.Percent.SHCB	-0.048099137	-0.30172099	-0.80507912
A.Percent.LCommute	0.002062844	-0.48405571	0.48156533

Figure 6 (a) - PC Loadings, Male

	PC1	PC2	PC3
F.Percent.Obes	-0.37556706	0.53601706	0.08328498
F.Percent.Inac	-0.18620982	0.55012709	0.09339776
A.Percent.N.Inet	0.47019698	0.13479330	0.21996905
A.Percent.ChilPov	0.42920232	0.17466091	-0.12911524
A.Percent.FRLunch	0.37219244	0.17454327	-0.18739251
A.Percent.FInsec	0.38789332	0.21037109	-0.08374773
A.Percent.NHInsur	0.34020108	-0.04750013	0.04797168
A.Percent.SHCB	-0.11672612	-0.19337905	-0.83284226
A.Percent.LCommute	-0.02500489	-0.49698809	0.42570670

Figure 6 (b) - PC Loadings, Female

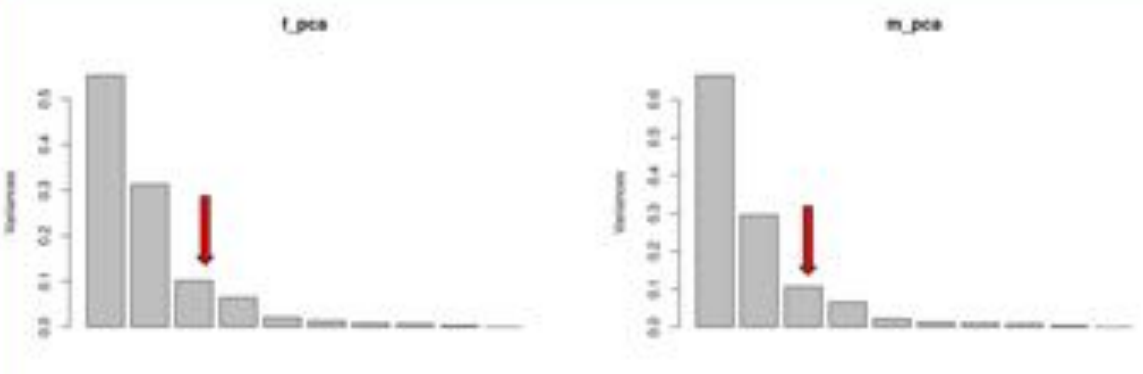

Figure 7 - PCA, scree plot knee selection

```
> summary(ind_mpca)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8       PC9
Standard deviation     0.8259 0.4986 0.3486 0.26827 0.13352 0.11366 0.09158 0.05341 3.927e-17
Proportion of Variance 0.5849 0.2131 0.1042 0.06171 0.01529 0.01108 0.00719 0.00245 0.000e+00
Cumulative Proportion  0.5849 0.7981 0.9023 0.96400 0.97929 0.99036 0.99755 1.00000 1.000e+00
> summary(ind_fpca)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8       PC9
Standard deviation     0.7679 0.5027 0.3415 0.26736 0.13242 0.11390 0.08491 0.05388 1.606e-17
Proportion of Variance 0.5505 0.2359 0.1089 0.06673 0.01637 0.01211 0.00673 0.00271 0.000e+00
Cumulative Proportion  0.5505 0.7865 0.8953 0.96208 0.97845 0.99056 0.99729 1.00000 1.000e+00
```

Figure 8 - PCA, cumulative proportion of variance selection

```
fa.parallel(m_cor_pear, n.obs = lend)
fa.parallel(f_cor_pear, n.obs = lend)
# results: 3 factors optimal.

## PFA/Evaluating Factors
m_pfa <- principal(m_cor_pear, nfactors = 3, n.obs = lend)
f_pfa <- principal(f_cor_pear, nfactors = 3, n.obs = lend)
m_pfa
f_pfa
print(m_pfa$loadings, cutoff = .4)
print(f_pfa$loadings, cutoff = .4)
```

Figure 9 - Parallel function and principal factor analysis

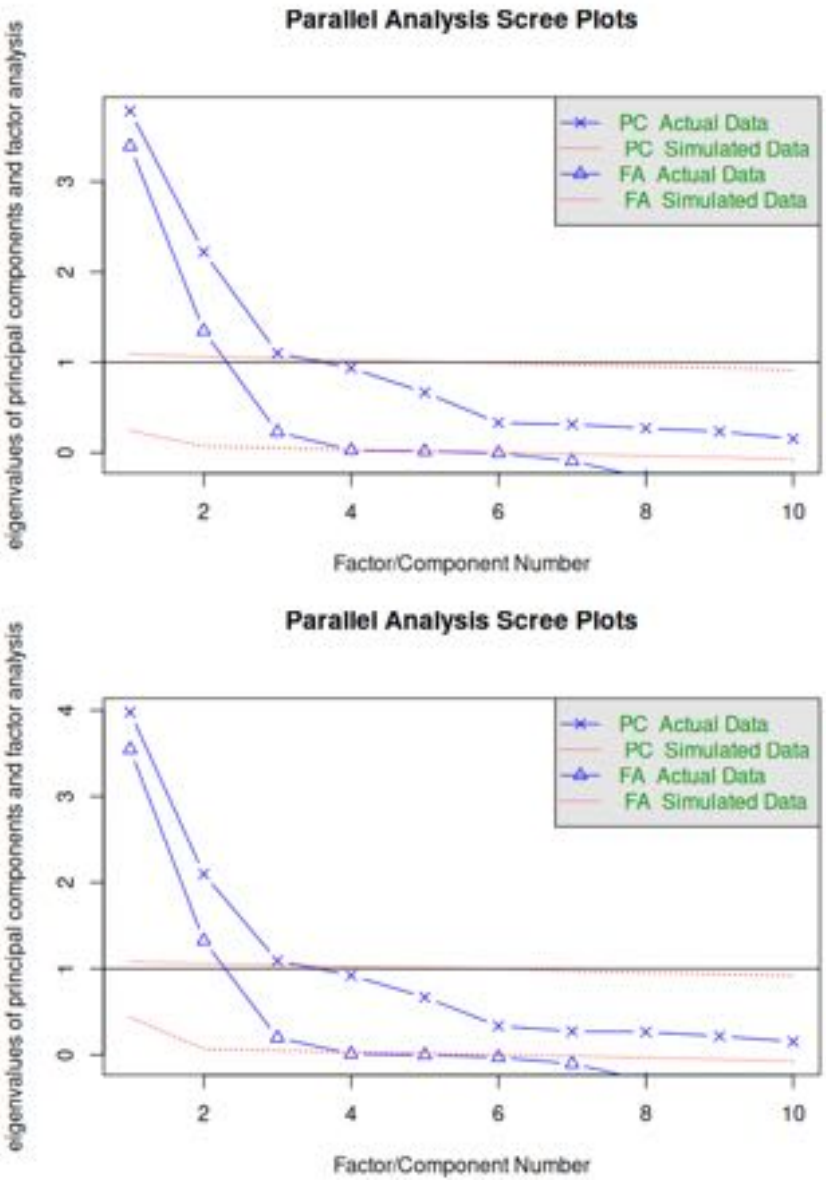

Figure 10 - fa.parallel() scree plots for factor/component count selection

```
> print(ind_mpfa$loadings, cutoff = .4)

Loadings:
                   RC1    RC2    RC3
M.Percent.Obes              0.917
M.Percent.Inac              0.897
A.Percent.N.Inet    0.868
A.Percent.ChilPov   0.903
A.Percent.FRLunch   0.836
A.Percent.FInsec    0.848
A.Percent.NHInsur   0.645
A.Percent.SHCB                     0.895
A.Percent.LCommute                 0.480

                 RC1   RC2   RC3
SS loadings    3.490 1.748 1.177
Proportion Var 0.388 0.194 0.131
Cumulative Var 0.388 0.582 0.713
```

```
> print(ind_fpfa$loadings, cutoff = .4)

Loadings:
                   RC1    RC2    RC3
F.Percent.Obes              0.931
F.Percent.Inac              0.891
A.Percent.N.Inet    0.869
A.Percent.ChilPov   0.900
A.Percent.FRLunch   0.831
A.Percent.FInsec    0.844
A.Percent.NHInsur   0.651
A.Percent.SHCB                     0.848
A.Percent.LCommute                 0.569

                 RC1   RC2   RC3
SS loadings    3.471 1.814 1.159
Proportion Var 0.386 0.202 0.129
Cumulative Var 0.386 0.587 0.716
```

Figure 11 - PFA Factor Loading

```
> summary(pc_ols_dm)

Call:
lm(formula = res_dm ~ PC1 + PC2 + PC3, data = reg_pca_dm)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8364 -0.8235 -0.1356  0.7329  5.2511

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.183618   0.110851  37.741   <2e-16 ***
PC1         -0.028188   0.002036 -13.845   <2e-16 ***
PC2          0.164704   0.003837  42.924   <2e-16 ***
PC3         -0.009758   0.005295  -1.843   0.0655 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.235 on 2506 degrees of freedom
Multiple R-squared:  0.468,     Adjusted R-squared:  0.4674
F-statistic: 734.9 on 3 and 2506 DF,  p-value: < 2.2e-16
```

```
> summary(pc_ols_df)

Call:
lm(formula = res_df ~ PC1 + PC2 + PC3, data = reg_pca_df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5441 -0.7837 -0.1362  0.6846  5.4369

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.455670   0.106300  23.101  < 2e-16 ***
PC1         -0.016159   0.002023  -7.986 2.11e-15 ***
PC2          0.166299   0.003342  49.757  < 2e-16 ***
PC3          0.006705   0.005288   1.268    0.205
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.199 on 2506 degrees of freedom
Multiple R-squared:  0.5485,    Adjusted R-squared:  0.5479
F-statistic:  1015 on 3 and 2506 DF,  p-value: < 2.2e-16
```

Figure 12 - PCA Regression, 3 components

```
> summary(pf_ols_dm)

Call:
lm(formula = res_dm ~ RC1 + RC2 + RC3, data = reg_pfa_dm)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5761 -0.6948 -0.0427  0.6285  5.0534

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.18183    0.02199 417.460  < 2e-16 ***
RC1          0.49411    0.02200  22.461  < 2e-16 ***
RC2          1.17917    0.02200  53.601  < 2e-16 ***
RC3          0.13074    0.02200   5.943 3.19e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.102 on 2506 degrees of freedom
Multiple R-squared:  0.5766,    Adjusted R-squared:  0.5761
F-statistic:  1138 on 3 and 2506 DF,  p-value: < 2.2e-16
```

```
> summary(pf_ols_df)

Call:
lm(formula = res_df ~ RC1 + RC2 + RC3, data = reg_pfa_df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3589 -0.6956 -0.0612  0.5922  5.4124

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.14984    0.02171 375.314   <2e-16 ***
RC1          0.62519    0.02172  28.785   <2e-16 ***
RC2          1.25383    0.02172  57.730   <2e-16 ***
RC3          0.18634    0.02172   8.579   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.088 on 2506 degrees of freedom
Multiple R-squared:  0.6282,    Adjusted R-squared:  0.6278
F-statistic:  1412 on 3 and 2506 DF,  p-value: < 2.2e-16
```

Figure 13 - PFA regression, 3 factors

```
> summary(rc1_ols_df)

Call:
lm(formula = F.Percent.DiagDia ~ A.Percent.N.Inet + A.Percent.ChilPov +
    A.Percent.FRLunch + A.Percent.FInsec + A.Percent.NHInsur,
    data = df_train)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8449 -1.1171 -0.2236  0.8739  7.2475

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.675121   0.127157  44.631  < 2e-16 ***
A.Percent.N.Inet -0.027193   0.005732  -4.744 2.21e-06 ***
A.Percent.ChilPov  0.042488   0.008239   5.157 2.71e-07 ***
A.Percent.FRLunch  0.029671   0.002831  10.479  < 2e-16 ***
A.Percent.FInsec   0.064419   0.014045   4.587 4.73e-06 ***
A.Percent.NHInsur -0.022842   0.007012  -3.258  0.00114 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.569 on 2504 degrees of freedom
Multiple R-squared:  0.2275,    Adjusted R-squared:  0.2259
F-statistic: 147.5 on 5 and 2504 DF,  p-value: < 2.2e-16
```


```
> summary(rc1_ols_df)

Call:
lm(formula = F.Percent.DiagDia ~ A.Percent.N.Inet + A.Percent.ChilPov +
    A.Percent.FRLunch + A.Percent.FInsec + A.Percent.NHInsur,
    data = df_train)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8449 -1.1171 -0.2236  0.8739  7.2475

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.675121   0.127157  44.631  < 2e-16 ***
A.Percent.N.Inet -0.027193   0.005732  -4.744 2.21e-06 ***
A.Percent.ChilPov  0.042488   0.008239   5.157 2.71e-07 ***
A.Percent.FRLunch  0.029671   0.002831  10.479  < 2e-16 ***
A.Percent.FInsec   0.064419   0.014045   4.587 4.73e-06 ***
A.Percent.NHInsur -0.022842   0.007012  -3.258  0.00114 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.569 on 2504 degrees of freedom
Multiple R-squared:  0.2275,    Adjusted R-squared:  0.2259
F-statistic: 147.5 on 5 and 2504 DF,  p-value: < 2.2e-16
```


Figure 14 - RC1 regression

```
> summary(rc2_ols_dm)

Call:
lm(formula = M.Percent.DiagDia ~ M.Percent.Obes + M.Percent.Inac,
    data = dm_train)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8825 -0.7626  0.0108  0.6610  4.5914

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.166140   0.121945   25.96   <2e-16 ***
M.Percent.Obes 0.072603   0.005439   13.35   <2e-16 ***
M.Percent.Inac 0.210325   0.007955   26.44   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.188 on 2507 degrees of freedom
Multiple R-squared:  0.5079,    Adjusted R-squared:  0.5075
F-statistic:  1294 on 2 and 2507 DF,  p-value: < 2.2e-16
```

```
> summary(rc2_ols_df)

Call:
lm(formula = F.Percent.DiagDia ~ F.Percent.Obes + F.Percent.Inac,
    data = df_train)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6629 -0.7682 -0.0371  0.6164  5.9084

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.821477   0.118302   15.40   <2e-16 ***
F.Percent.Obes 0.070896   0.005571   12.72   <2e-16 ***
F.Percent.Inac 0.205911   0.007426   27.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.2 on 2507 degrees of freedom
Multiple R-squared:  0.5479,    Adjusted R-squared:  0.5475
F-statistic:  1519 on 2 and 2507 DF,  p-value: < 2.2e-16
```

Figure 15 - RC2 regression

```
> summary(rc3_ols_dm)

Call:
lm(formula = M.Percent.DiagDia ~ A.Percent.SHCB + A.Percent.LCommute,
    data = dm_train)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6965 -1.2479 -0.3181  1.0203  7.2069

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       8.535551   0.117584  72.591  < 2e-16 ***
A.Percent.SHCB    0.056462   0.009676   5.835 6.07e-09 ***
A.Percent.LCommute 0.005465   0.006698   0.816    0.415
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.681 on 2507 degrees of freedom
Multiple R-squared:  0.01404,   Adjusted R-squared:  0.01326
F-statistic: 17.85 on 2 and 2507 DF,  p-value: 1.999e-08
```

```
> summary(rc3_ols_df)

Call:
lm(formula = F.Percent.DiagDia ~ A.Percent.SHCB + A.Percent.LCommute,
    data = df_train)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5723 -1.2387 -0.3572  0.9798  8.3793

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       7.096117   0.122590  57.885   <2e-16 ***
A.Percent.SHCB    0.094178   0.010088   9.335   <2e-16 ***
A.Percent.LCommute 0.006129   0.006983   0.878     0.38
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.753 on 2507 degrees of freedom
Multiple R-squared:  0.03459,   Adjusted R-squared:  0.03382
F-statistic: 44.91 on 2 and 2507 DF,  p-value: < 2.2e-16
```

Figure 16 - RC3 regression