Group Project
Framingham Heart Study

Arvind Ramakrishnan
Eric Reed
Sangsoo Park

April 7, 2014

```
require(ggplot2)

## Loading required package:  ggplot2

require(RCurl)

## Loading required package:  RCurl
## Loading required package:  bitops

data <- getURL("https://raw.githubusercontent.com/arvindram12/Final-Project/master/frmgham2.
    ssl.verifypeer = 0L, followlocation = 1L)
writeLines(data, "framingham.csv")
frm <- read.csv("framingham.csv")
```

1 Characteristics of the Data Set

This data set is the result of the Framingham Heart Study, which was performed to identify the main risk factors for Cardiovascular Disease (CVD), more information can be found on their website at http://www.framinghamheartstudy.org/about-fhs/history.php.

1.1 Observations

In total there were 11,627 observations made, however, there were multiple observations made for each individual in the study, as was evident that multiple observations had the same ID number. Therefore, we filtered the data to include just the first observation for each study participant.

```
frm$RANDID <- as.factor(frm$RANDID)
frm1 <- frm[which(!duplicated(frm$RANDID)), ]
nrow(frm1)

## [1] 4434
```

Here, we are left with one observation for each of the 4,434 study participants, which leaves us with a fairly large dataset.

1.2 Variables

In total, the dataset is comprised of 39 variables, of which our analysis will focus on 13. Of these 13 variables 8 are continuous and 5 are categorical.

Continous Variables

1. Total Cholestrol (mg/dL) — "TOTCHOL"
2. Age $(years)$ — "AGE"
3. Systolic blood pressure $(mmHg)$ — "SYSBP"
4. Diastolic blood pressure$(mmHg)$ — "DIABP"
5. Cigarettes Per Day — "CIGPDAY"
6. BMI : Body Mass Index (kg/m^2) — "BMI"
7. Heart Rate $(beats/min)$ — "HEARTRTE"
8. Glucose (mg/dL) — "GLUCOSE"

Categorical Variables

1. Sex $(1 = Male, 2 = Female)$ — "SEX"
2. Current Smoker? $(0 = No, 1 = Yes)$ — "CURSMOKE"
3. Diabetic? $(0 = No, 1 = Yes)$ — "DIABETES"
4. Currently on Blood Pressure Medication? $(0 = No, 1 = Yes)$ — "BPMEDS"
5. Education $(1 = Grades\ 1-11, 2 = High\ School\ Diploma\ or\ GED, 3 = Some\ College, 4 = College\ Degree)$ — "educ"

```
frm2 <- frm1[, 1:14]
```

1.3 Missing Data

```r
# Check the number of missing values of the continuous varaibles
sum(complete.cases(frm2))
```

```
## [1] 3826
```

```r
sum(is.na(frm2$SEX))
```

```
## [1] 0
```

```r
sum(is.na(frm2$TOTCHOL))
```

```
## [1] 52
```

```r
sum(is.na(frm2$AGE))
```

```
## [1] 0
```

```r
sum(is.na(frm2$SYSBP))
```

```
## [1] 0
```

```r
sum(is.na(frm2$DIABP))
```

```
## [1] 0
```

```r
sum(is.na(frm2$CURSMOKE))
```

```
## [1] 0
```

```r
sum(is.na(frm2$CIGPDAY))
```

```
## [1] 32
```

```r
sum(is.na(frm2$BMI))
```

```
## [1] 19
```

```r
sum(is.na(frm2$DIABETES))
```

```
## [1] 0
```

```r
sum(is.na(frm2$BPMEDS))
```

```
## [1] 61
```

```r
sum(is.na(frm2$HEARTRTE))
```

```
## [1] 1
```

```r
sum(is.na(frm2$GLUCOSE))
```

```
## [1] 397
```

```r
sum(is.na(frm2$educ))
```

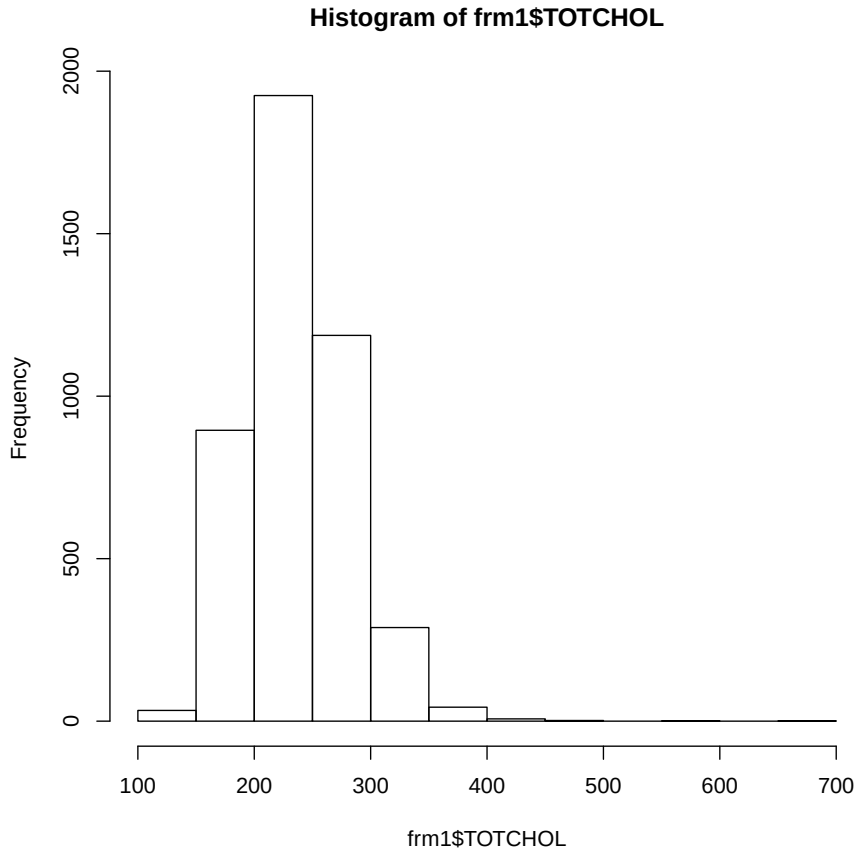
```
## [1] 113
```

Of the 4434 observations in our data, 3,826 of them have complete data. We are missing 52 observations for total cholesterol, 32 obervations for cigarettes per day, 19 observations for BMI, 61 observations for blood pressure medication, 1 observation for heart rate, 397 observations for glucose, and 113 observations for education.

1.4 Distributions of Continuous Variables

Total Choesterol

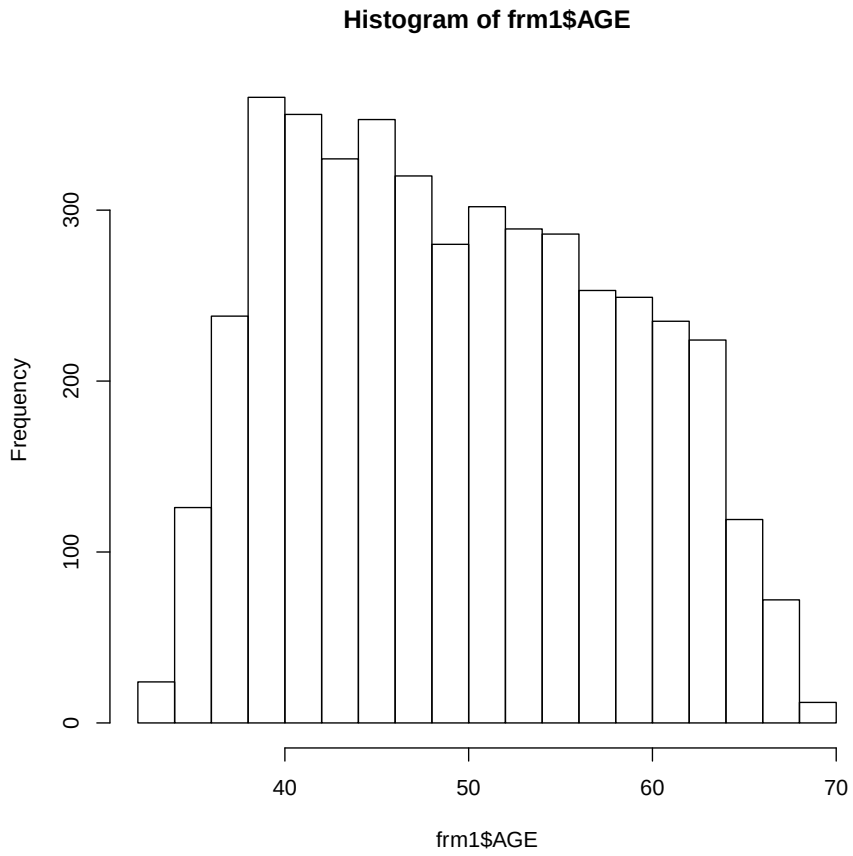
```
hist(frm1$TOTCHOL)
```



The data for total choesterol appears to be normally distributed

Age

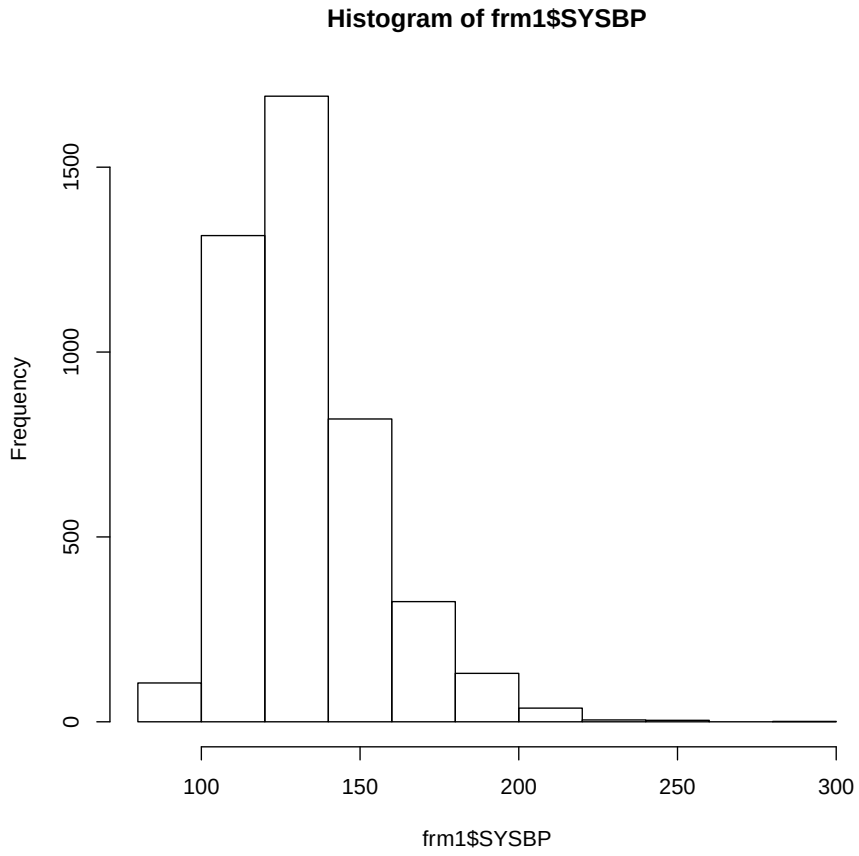
```
hist(frm1$AGE)
```



The data for age, has no obvous distribution, though it does taper of at either tail.

Systolic Blood Pressure

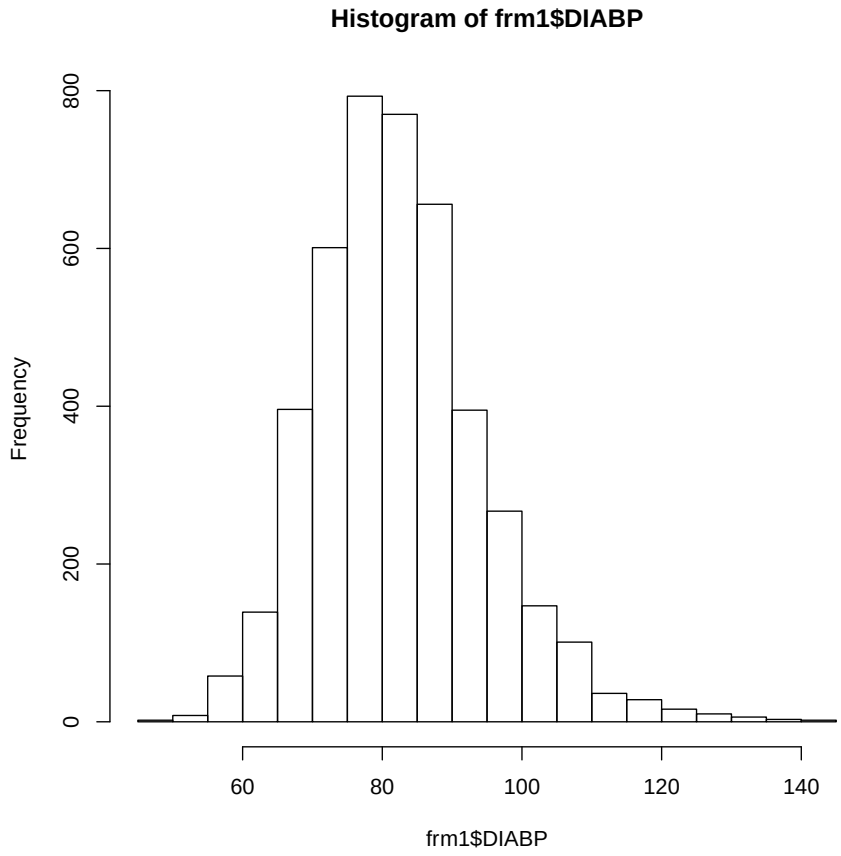
```
hist(frm1$SYSBP)
```



The data for systolic blood pressure appears positively skewed.

Diabolic Blood Pressure

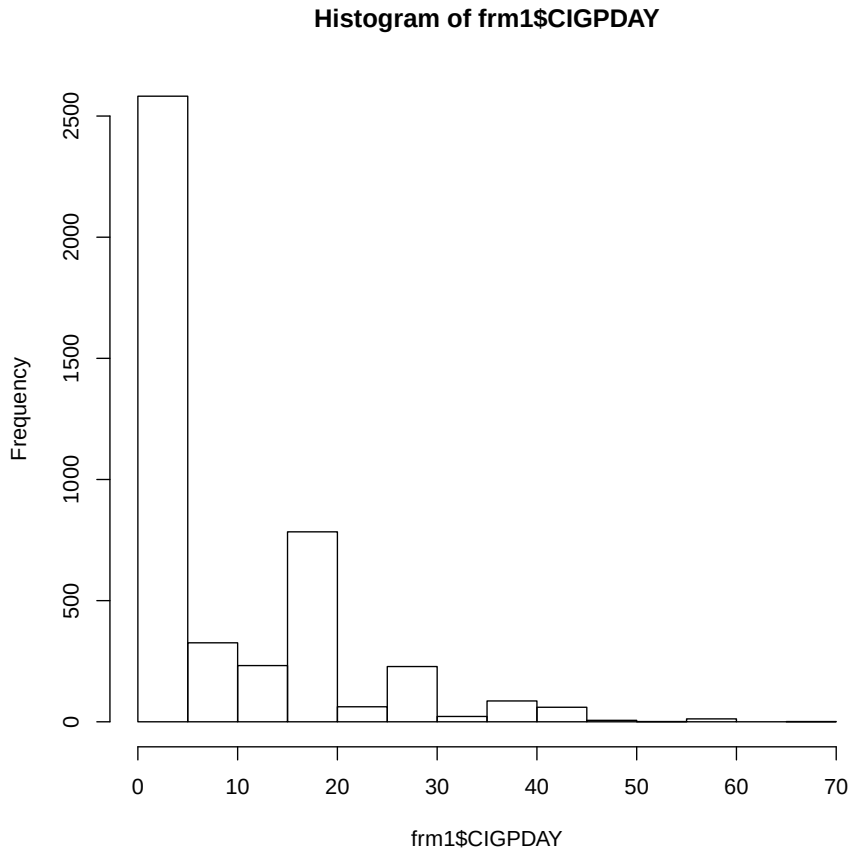
```
hist(frm1$DIABP)
```



The data for diaboloic blood pressure appears normally distributed, though it seems a little positively skewed.

Cigarettes per Day

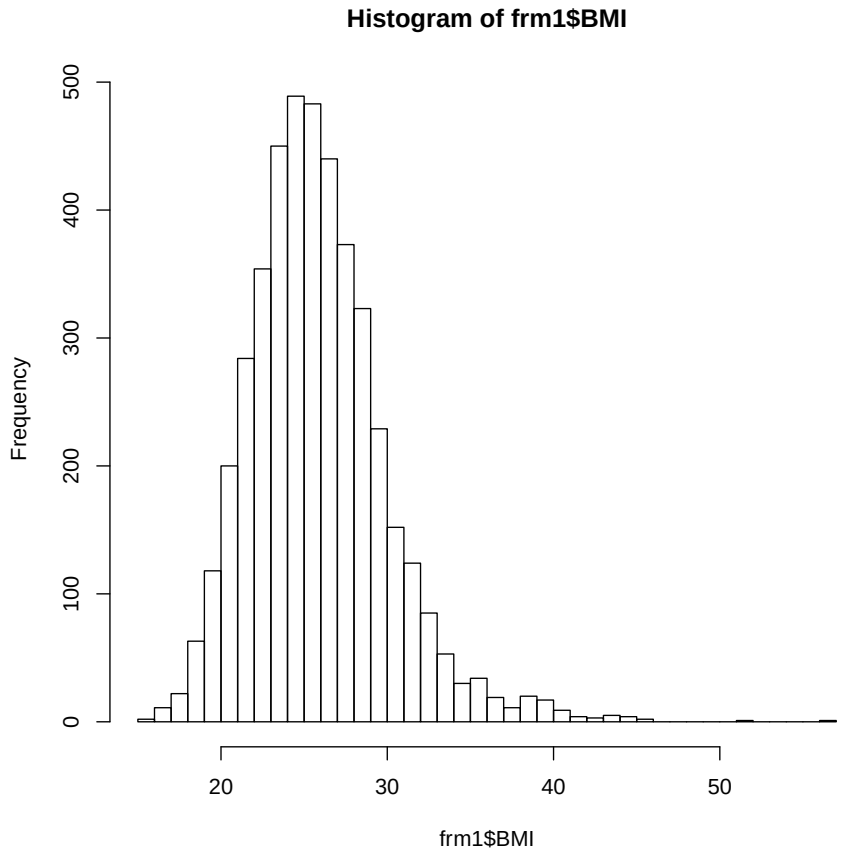
```
hist(frm1$CIGPDAY)
```



The frequency of cigarettes per day, appears to decrease as the number of cigarettes increases.

Body Mass Index

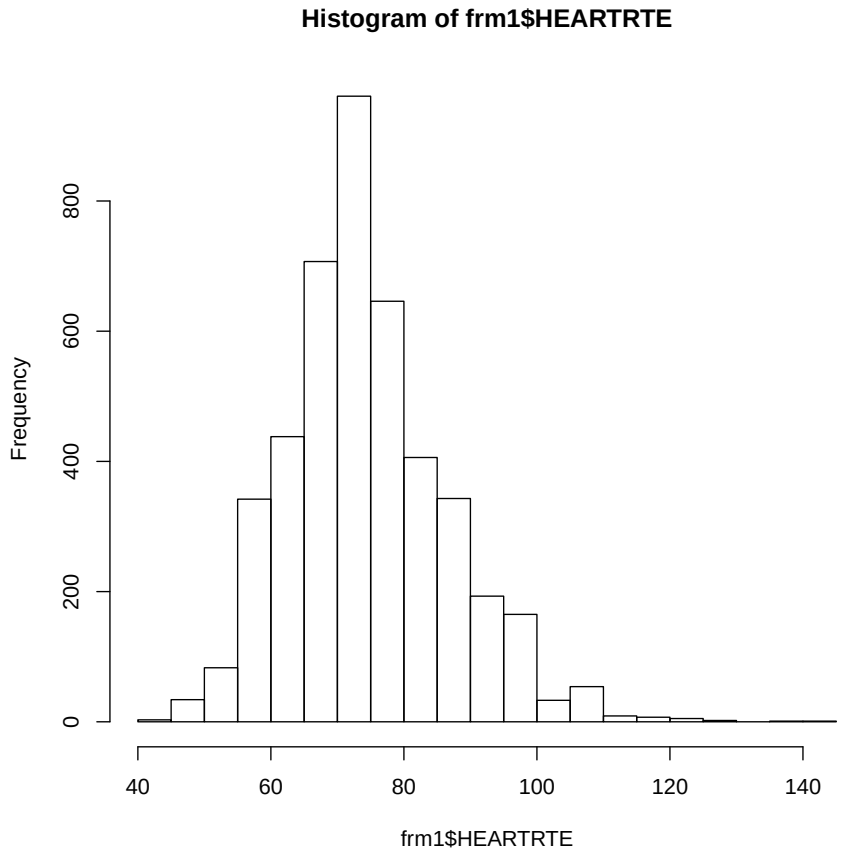
```
hist(frm1$BMI, breaks = 50)
```



The data for body mass index appears normally distributed, though it may be postively skewed.

Heart Rate

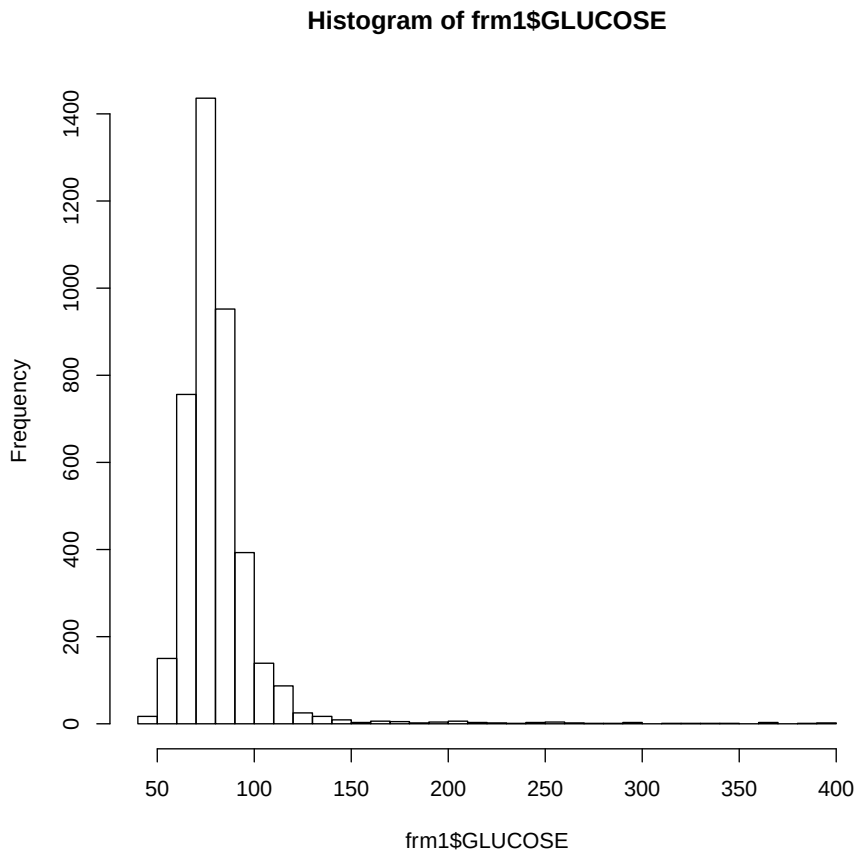
```
hist(frm1$HEARTRTE, breaks = 20)
```



The data for heart rate appears normally distributed, though it may be positively skewed.

Glucose

```
hist(frm1$GLUCOSE, breaks = 40)
```



The data for glucose appears slightly positively skewed, with positive outliers on the postive end.

2 Candidate Continuous Variables for Linear Regression

In the following example I utilized the GGally package to create a matrix of plots and correlational ceofficients for candidate continuous predictor and outcome variables, after removing observations with missing data.

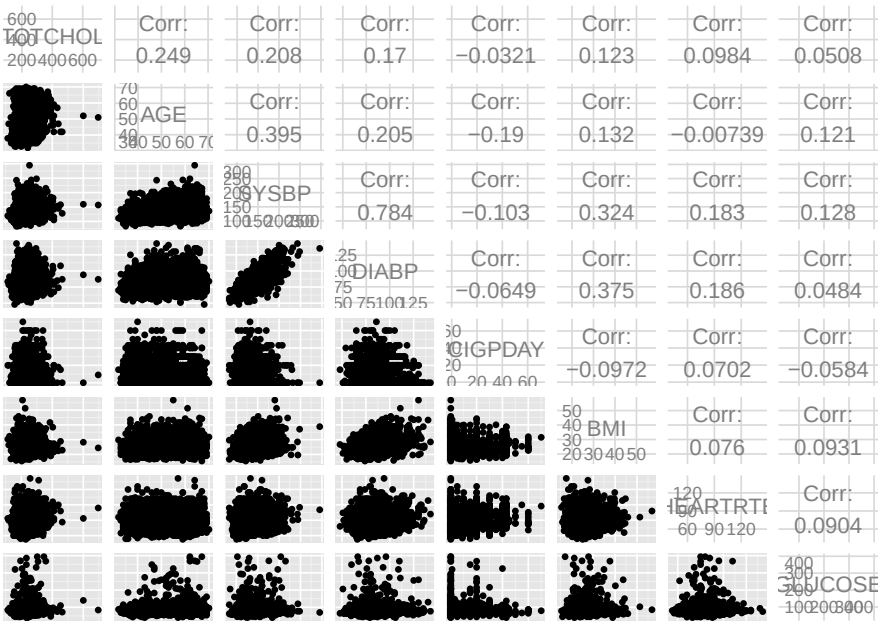


Figure 1: Plot and Correlational Matrix of Continuous Variables

From the figure we can see that the most consistent correlation between one variable and the others seems to be Systolic Blood pressure. This is a good outcome variable because it intuitively makes sense as an outcome, whereas other variables such as BMI occur in a large part due to environmental factors that aren't factored into this study such as diet. Originally, Total Cholestrol was considered as the outcome variable, however it has relatively weak correlation with many of the factors of interest. Systolic Blood Pressure has the strongest correlation with Diabolic Blood Pressure, however this is a relatively uninteresting relationship to study as it they can be considered more or less colinear. We can move on instead to the next highest correlation which is BMI.

2.1 Simple Linear Regression Analysis of Systolic Blood Pressure vs. BMI

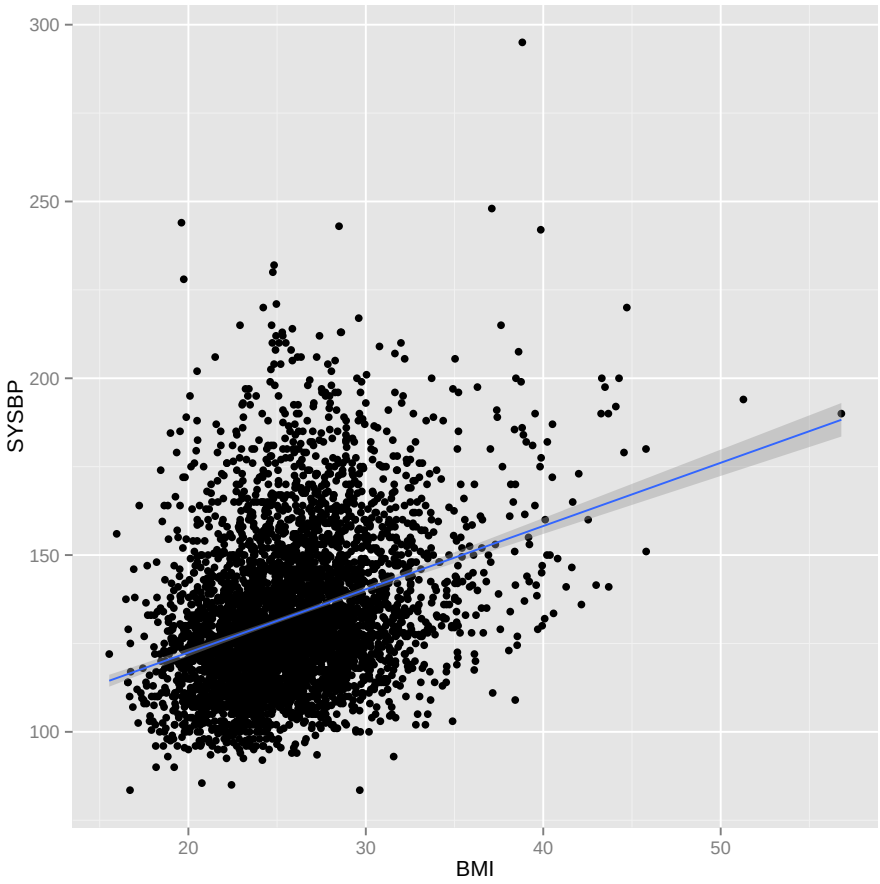
```
m1 <- lm(SYSBP ~ BMI, data = frm2)
summary(m1)

##
## Call:
```

```
## lm(formula = SYSBP ~ BMI, data = frm2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -56.21 -14.78  -3.83  10.42 138.90
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  86.6668     2.0286    42.7   <2e-16 ***
## BMI           1.7885     0.0775    23.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.1 on 4413 degrees of freedom
##   (19 observations deleted due to missingness)
## Multiple R-squared:  0.108,Adjusted R-squared:  0.107
## F-statistic:  532 on 1 and 4413 DF,  p-value: <2e-16

qplot(BMI, SYSBP, data = frm2) + geom_smooth(method = "lm", se = TRUE)

## Warning:  Removed 19 rows containing missing values (stat_smooth).
## Warning:  Removed 19 rows containing missing values (geom_point).
```



From our simple linear modeling summary we can see a conclude that there is an association between systolic blood pressure and BMI. The adjusted R^2 is relatively low. It would be interesting to see then the effect of adding new variables to our model.

2.2 Linear Regression of Systolic Blood Pressure vs. Interactions with Sex

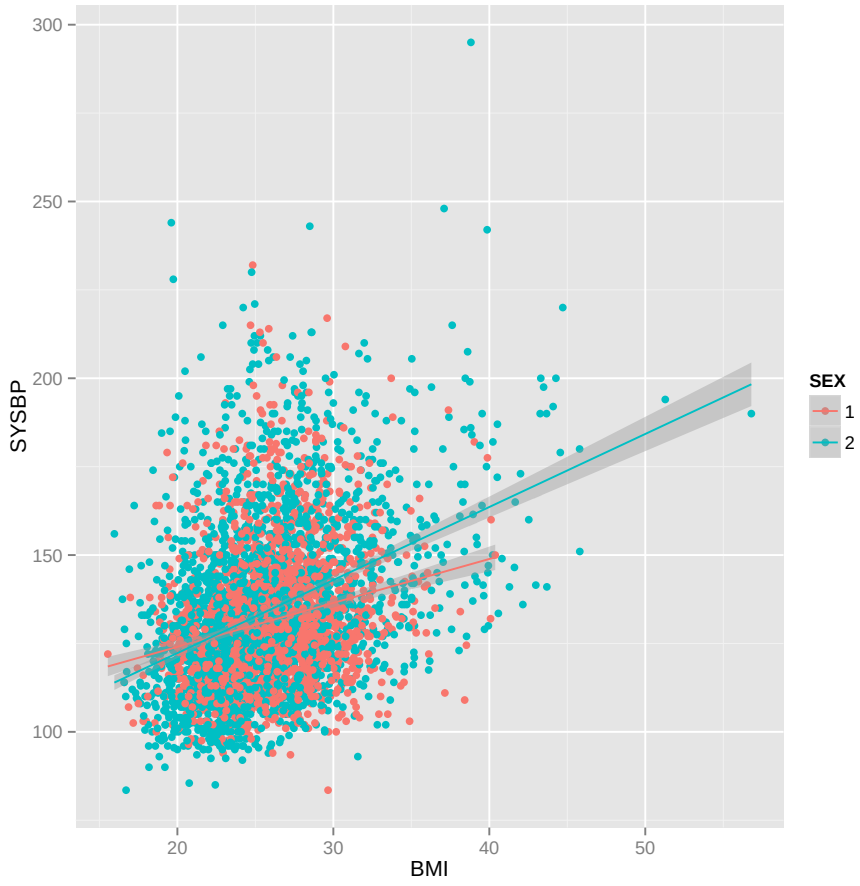
```
# Convert data type of SEX variable from integer to factor
frm1_du <- frm1
frm1_du$SEX <- as.factor(frm1_du$SEX)

# Multiple regression model, 1 continuous and 1 categorical variables
m2 <- lm(SYSBP ~ BMI * factor(SEX), data = frm1_du)
summary(m2)
```

```
##
## Call:
## lm(formula = SYSBP ~ BMI * factor(SEX), data = frm1_du)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -53.17 -14.80  -3.57  10.48 133.85
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         99.176      3.697   26.82  < 2e-16 ***
## BMI                  1.243      0.140    8.87  < 2e-16 ***
## factor(SEX)2       -18.220      4.413   -4.13  3.7e-05 ***
## BMI:factor(SEX)2     0.823      0.168    4.90  1.0e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21 on 4411 degrees of freedom
##   (19 observations deleted due to missingness)
## Multiple R-squared:  0.117,Adjusted R-squared:  0.117
## F-statistic:  196 on 3 and 4411 DF,  p-value: <2e-16
```

```
qplot(BMI, SYSBP, data = frm1_du, color = SEX) + geom_smooth(method = "lm",
    se = TRUE)
```

```
## Warning:  Removed 5 rows containing missing values (stat_smooth).
## Warning:  Removed 14 rows containing missing values (stat_smooth).
## Warning:  Removed 19 rows containing missing values (geom_point).
```

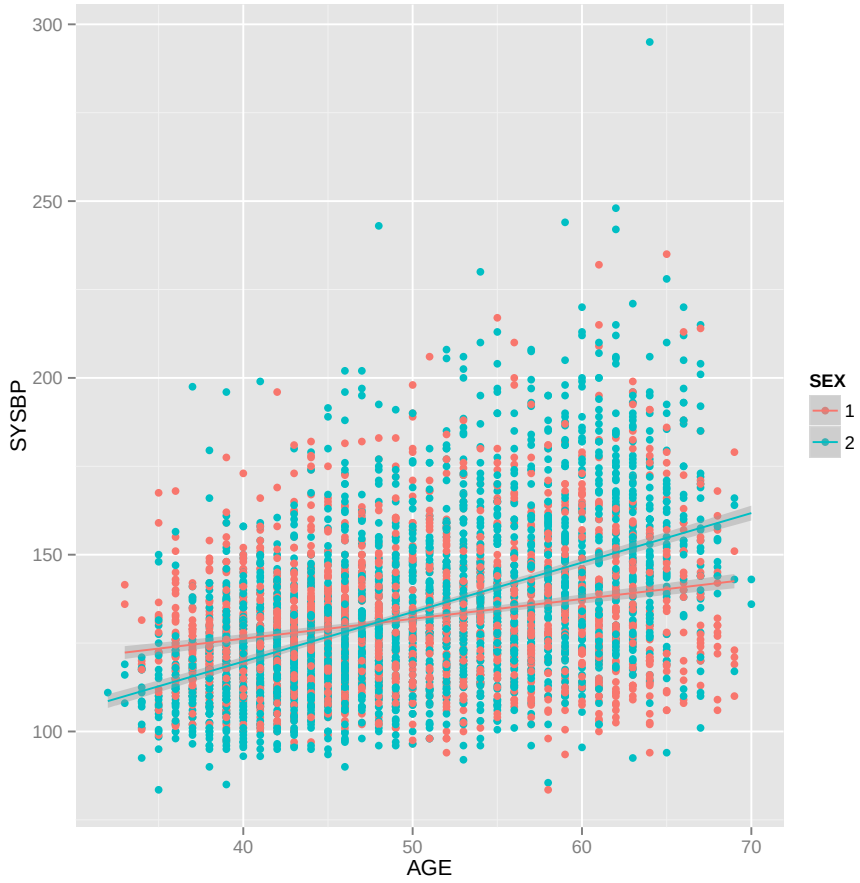



```
# Another trial with another continuous variable
m3 <- lm(SYSBP ~ AGE * factor(SEX), data = frm1_du)
summary(m3)

##
## Call:
## lm(formula = SYSBP ~ AGE * factor(SEX), data = frm1_du)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.76 -13.23  -2.47  10.17 141.64
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      103.8011     2.6604    39.0   <2e-16 ***
## AGE                0.5611     0.0526    10.7   <2e-16 ***
```

```
## factor(SEX)2     -39.9956     3.5711   -11.2   <2e-16 ***
## AGE:factor(SEX)2   0.8382     0.0705    11.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.2 on 4430 degrees of freedom
## Multiple R-squared:  0.186,Adjusted R-squared:  0.186
## F-statistic:  338 on 3 and 4430 DF,  p-value: <2e-16
```

```
qplot(AGE, SYSBP, data = frm1_du, color = SEX) + geom_smooth(method = "lm",
    se = TRUE)
```

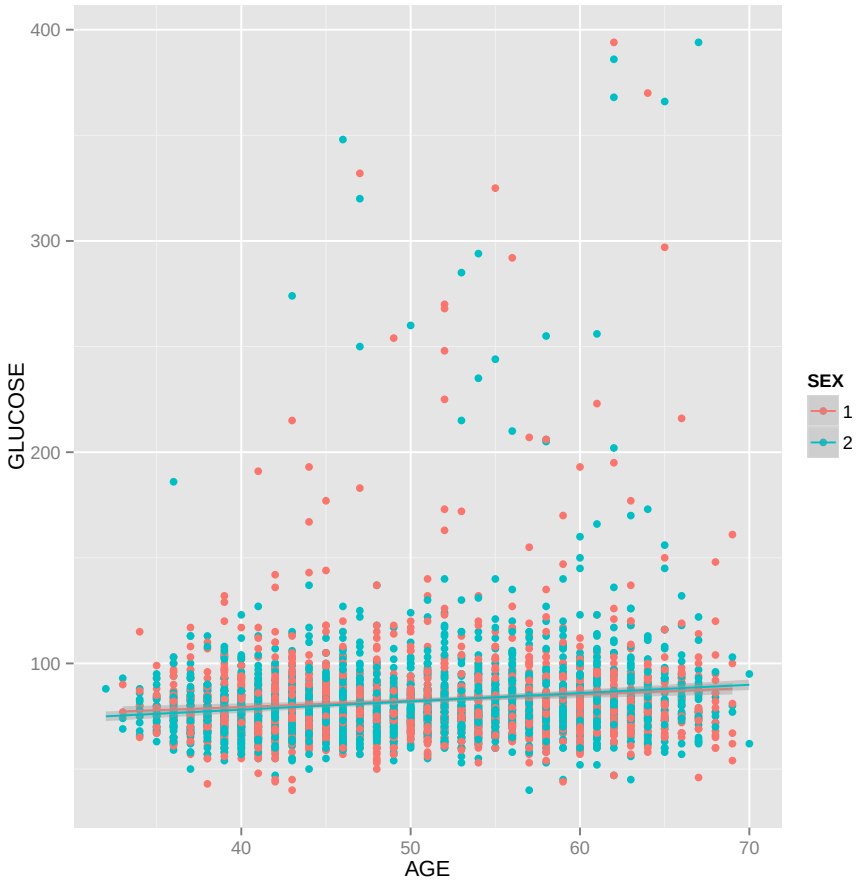



```
# Another trial with another continuous variable
m4 <- lm(GLUCOSE ~ AGE * factor(SEX), data = frm1_du)
summary(m4)
```

```
##
## Call:
## lm(formula = GLUCOSE ~ AGE * factor(SEX), data = frm1_du)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -44.78 -10.73  -3.72   4.97 308.06
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       67.4459     3.2852   20.53  < 2e-16 ***
## AGE                0.2983     0.0649    4.60  4.4e-06 ***
## factor(SEX)2      -5.0351     4.4621   -1.13     0.26
## AGE:factor(SEX)2   0.0942     0.0880    1.07     0.28
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.2 on 4033 degrees of freedom
##   (397 observations deleted due to missingness)
## Multiple R-squared:  0.0158,Adjusted R-squared:  0.0151
## F-statistic: 21.6 on 3 and 4033 DF,  p-value: 6.73e-14
```

```
qplot(AGE, GLUCOSE, data = frm1_du, color = SEX) + geom_smooth(method = "lm",
    se = TRUE)
```

```
## Warning:  Removed 120 rows containing missing values (stat_smooth).
## Warning:  Removed 277 rows containing missing values (stat_smooth).
## Warning:  Removed 397 rows containing missing values (geom_point).
```


400
300
200
100
GLUCOSE
40
50
60
70
AGE
SEX
1
2