Logistic Regression on 10 Year Heart Disease Risk

By: Evan Parker, Edward Coleman, Zack Shin, and Johnny Zhang

December 11, 2022

Contents

1 Introduction

Our analysis will be a logistic analysis on the Framingham Heart Study dataset, a dataset from a long-term ongoing cardiovascular cohort study of residents on Framingham, Massachusetts.

The Framingham study began in 1948 with over 5,000 research participates. These research participants were 30 to 60 years of age and were medically examined every 2 years. In 1971, the study organizers began obtaining new research participants to replace some of the original research participants from 1948 who had passed away. Our starting dataset contains 4240 observations of patients from both the original 1948 study and the 1971 re-roster.

This analysis will hope to uncover key characteristics of patients who have been diagnosed with Coronary Heart Disease in an attempt to help diagnose patients sooner and more effectively, potentially saving the lives of the diagnosed patients.

2 Data Management and Analytic Data Set Creation

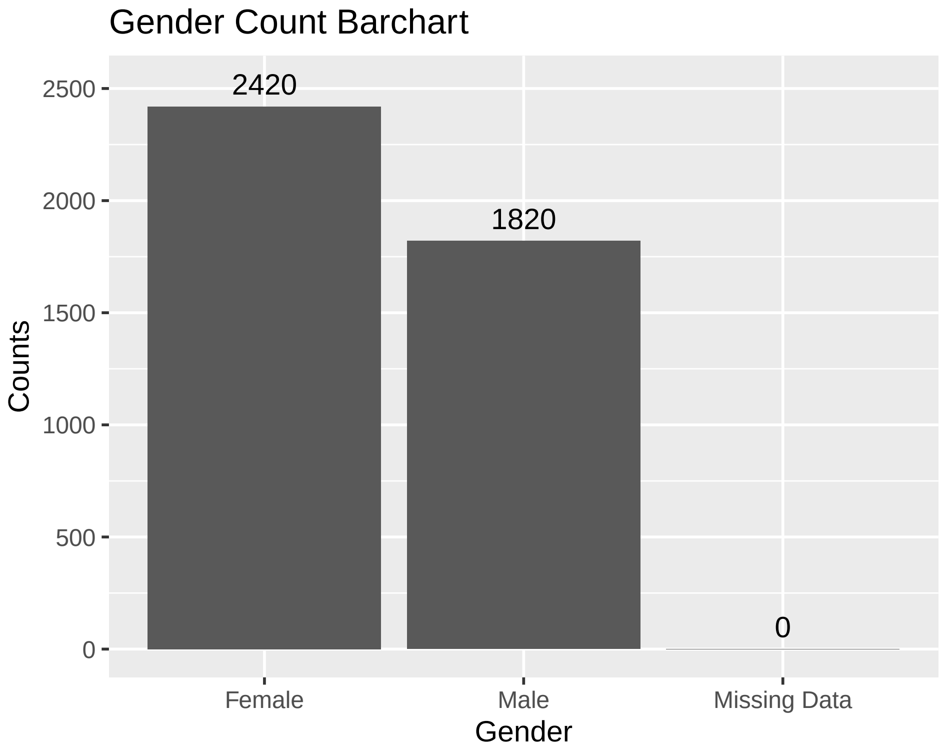
First, we obtained the dataset and uploaded it to Github, making it publicly available for our team. Next, we converted the data file into a .csv file, allowing us to load the dataset into R for further analysis.

2.1 Variable Breakdown

Here, we will provide a breakdown of all the variables within the dataset along with some analyses of the counts and potential histograms for numerical variables. We will provide insight into missing data values later in this report.

2.1.1 Male

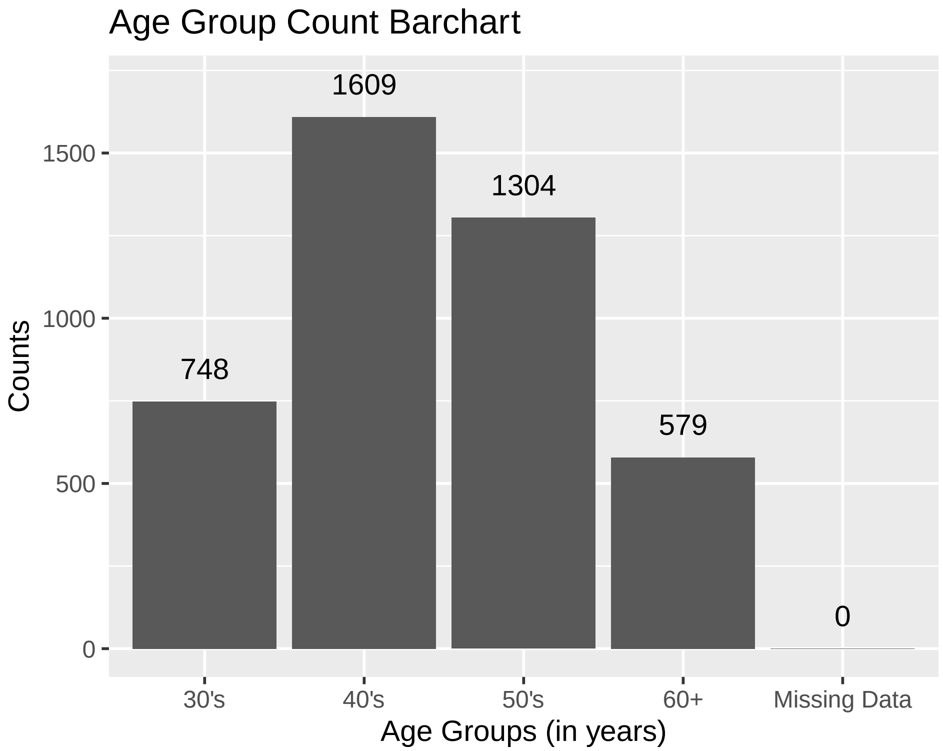
The first variable our data set is **male**, a dummy variable where a *0* represents that the patient is female and a *1* represents that the patient is male. Below is a breakdown of the amount of males and females within our dataset.



As shown in the barchart above, our data set has *2420* females and *1820* males, with no patients missing their gender.

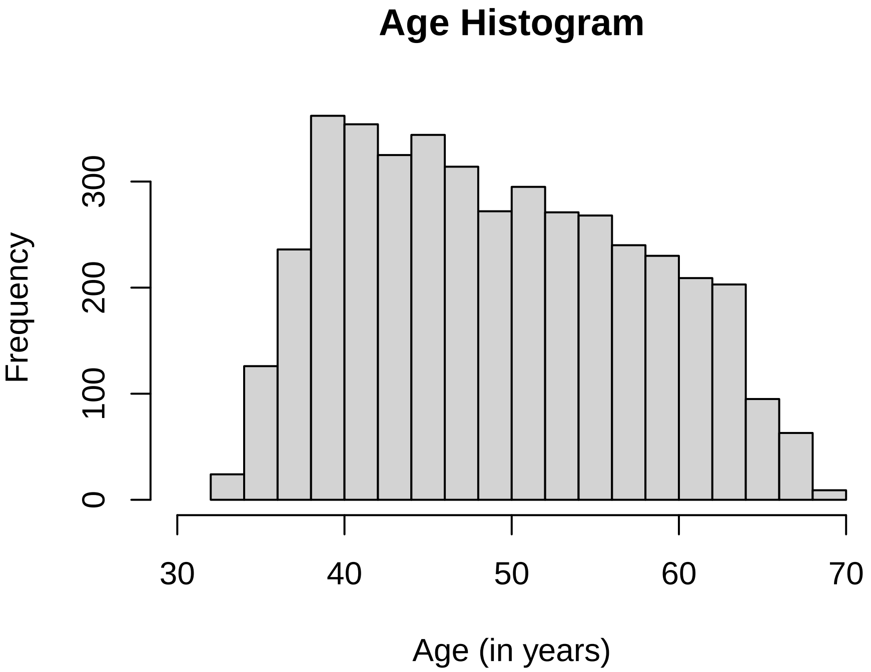
2.1.2 Age

The second variable in our data set is **age**, a numerical variable measuring the patient's age in years for that given medical examination. Below is a histogram of the patient's ages.



As shown in the barchart above, our data set has *748* patients in their 30's, *1609* patients in their 40's, *1304* patients in their 50's, *579* patients in their 60's and higher, and *0* patients missing their age.

Since **age** is a continuous quantitative variable, below is a histogram.

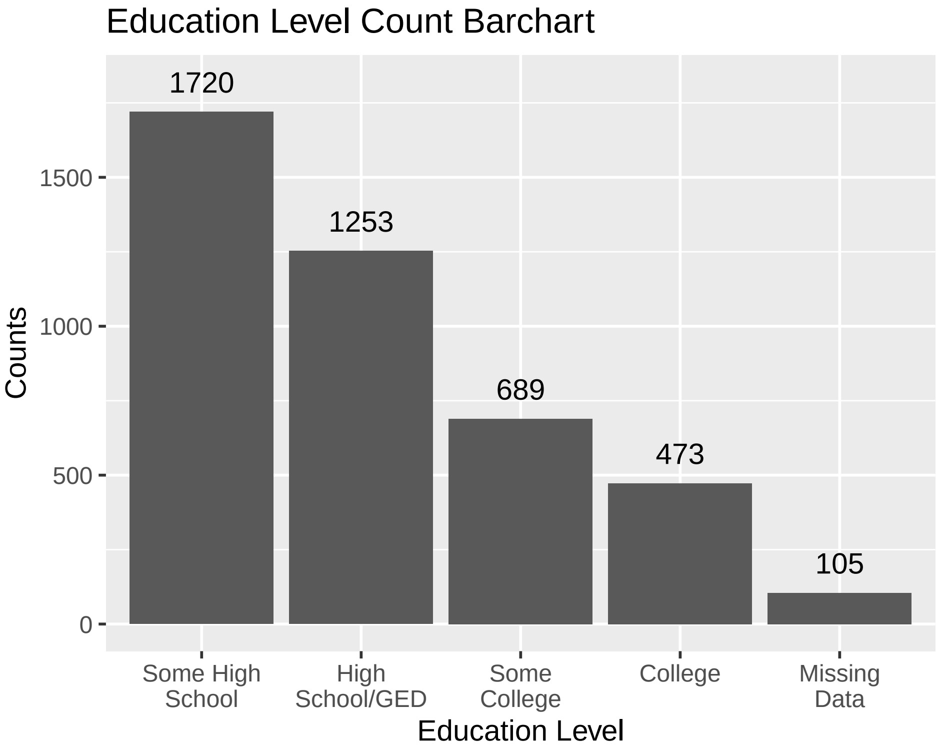


As shown above, the variable **age** appears to be approximately normal.

2.1.3 Education

The third variable in our data set is **education**, a variable ranging from 1 through 4 where each number represents a different level in the patients most completed educational level. The breakdown of each educational level is below.

- 1: Some High School
- 2: High School/GED
- 3: Some College/Vocational School
- 4: College

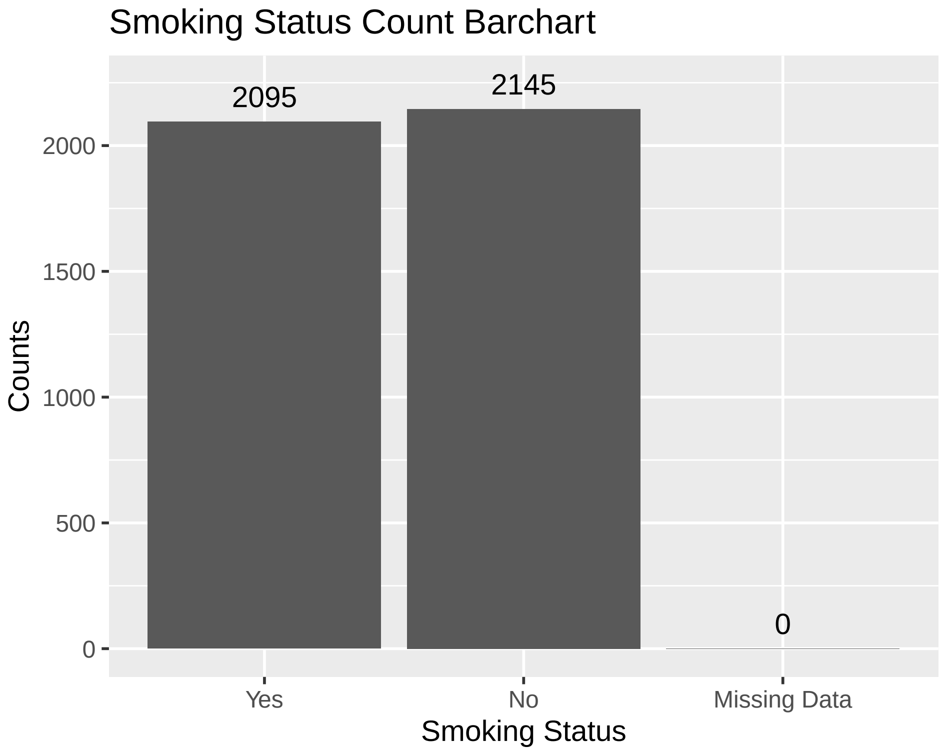


As shown in the barchart above, our data set has *1720* patients have some high school education, *1253* patients have a high school/GED education, *689* patients have some college/vocational school education, *473* patients have college education, and *105* patients are missing their education data.

We will address our attempts of recovering this missing data in a later section of this report.

2.1.4 Smoking Status

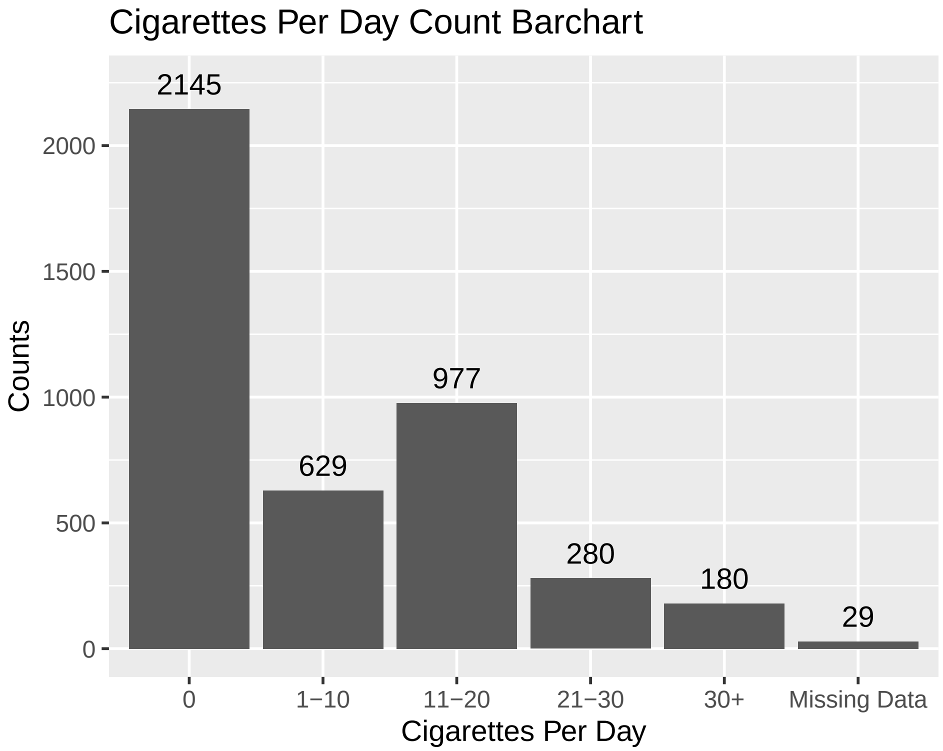
The fourth variable in our data set is **SmokingStatus**, a dummy variable where a **0** represents that the patient is not a smoker and a **1** represents that the patient is a smoker Below is a breakdown of the amount of smokers and non-smokers within our dataset.



As shown in the barchart above, our data set has *2095* smokers and *2145* non-smokers, with no patients missing their smoking status.

2.1.5 Cigarettes Per Day

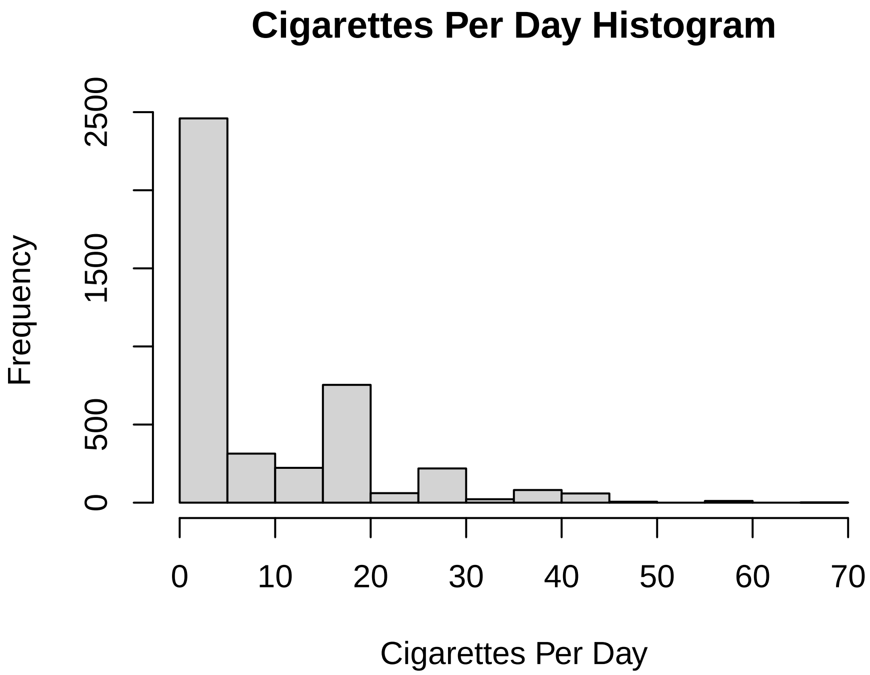
The fifth variable in our data set is **CigarettsPerDay**, a variable where patients estimate how many cigarettes they smoke per day at the time of their examinations. Below is a breakdown of **CigarettsPerDay**



As shown in the barchart above, our data set has *2145* patients who do not smoke, *629* patients who smoke 1 to 10 cigarettes per day, *977* patients who smokes 11 to 20 cigarettes per day, *280* patients who smokes 21 to 30 cigarettes per day, *180* patients who smoke over 30 cigarettes per day, and *29* patients who are missing their cigarettes per day data.

We will address our attempts of recovering this missing data in a later section of this report.

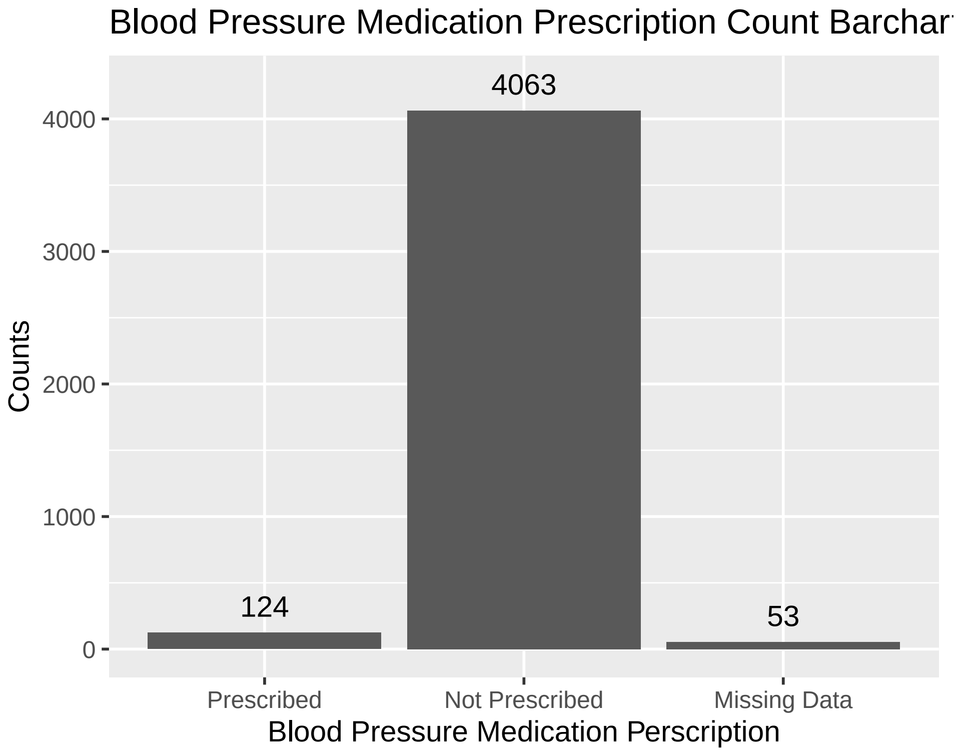
Since **CigarettesPerDay** is a continuous quantitative variable, below is a histogram.



As shown above, the variable **CigarettesPerDay** appears to be skewed right, which would make sense given the large proportion of our sample that does not smoke.

2.1.6 Blood Pressure Meds

The sixth variable in our data set is **BloodPressureMeds**, a dummy variable where a *0* represents that the patient does not take blood pressure medication and a *1* represents that the patient does take blood pressure medication. Below is a breakdown of **BloodPressureMeds**:

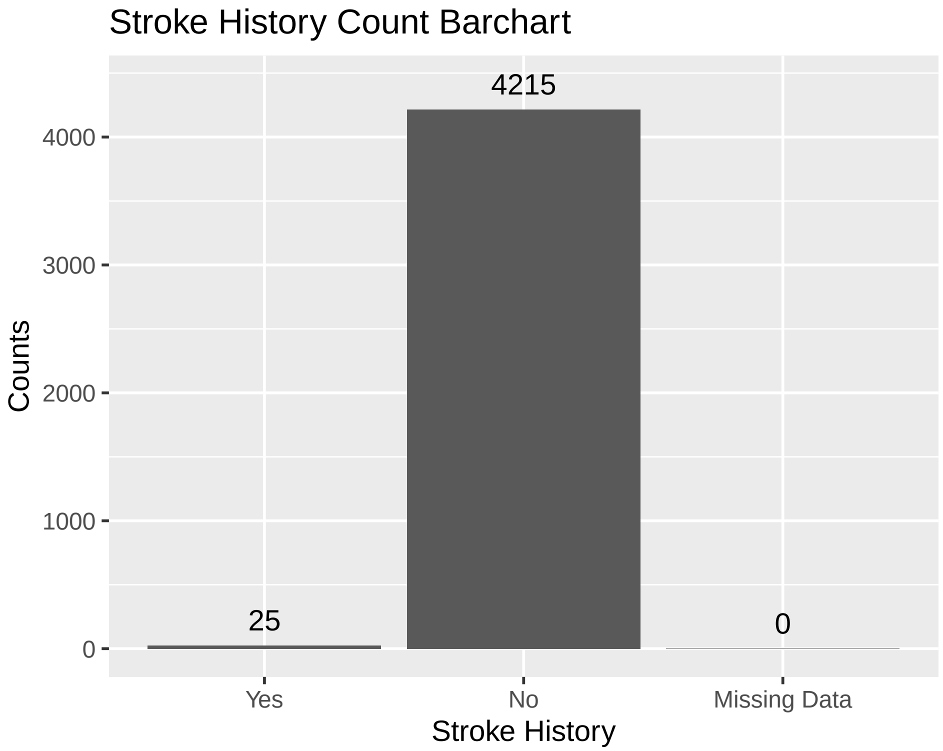


As shown in the barchart above, our data set has *124* patients who have been prescribed blood pressure medication, *4063* patients who are not prescribed blood pressure medication, and *53* patients who are missing their blood pressure medication status data.

We will address our attempts of recovering this missing data in a later section of this report.

2.1.7 Stroke History

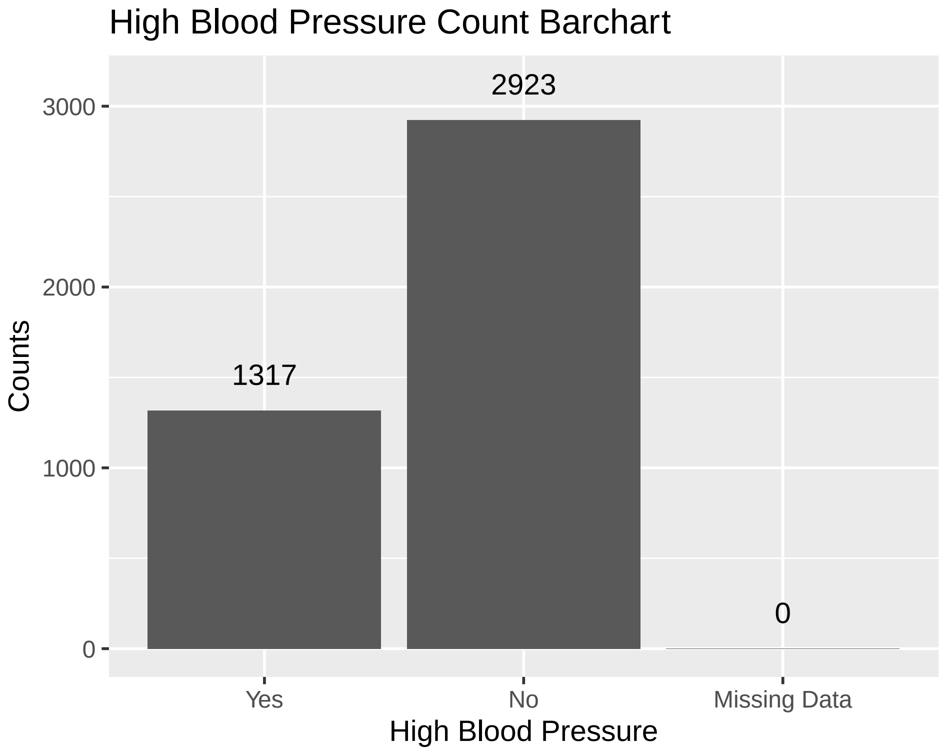
The seventh variable in our data set is **StrokeHistory**, a dummy variable where a *0* represents that the patient has not had a stroke and a *1* represents that the patient has had a stroke. Below is a breakdown of **StrokeHistory**:



As shown in the barchart above, our data set has *25* patients who have had a stroke, *4215* patients who have not had a stroke, and *0* patients who are missing their stroke data.

2.1.8 High Blood Pressure

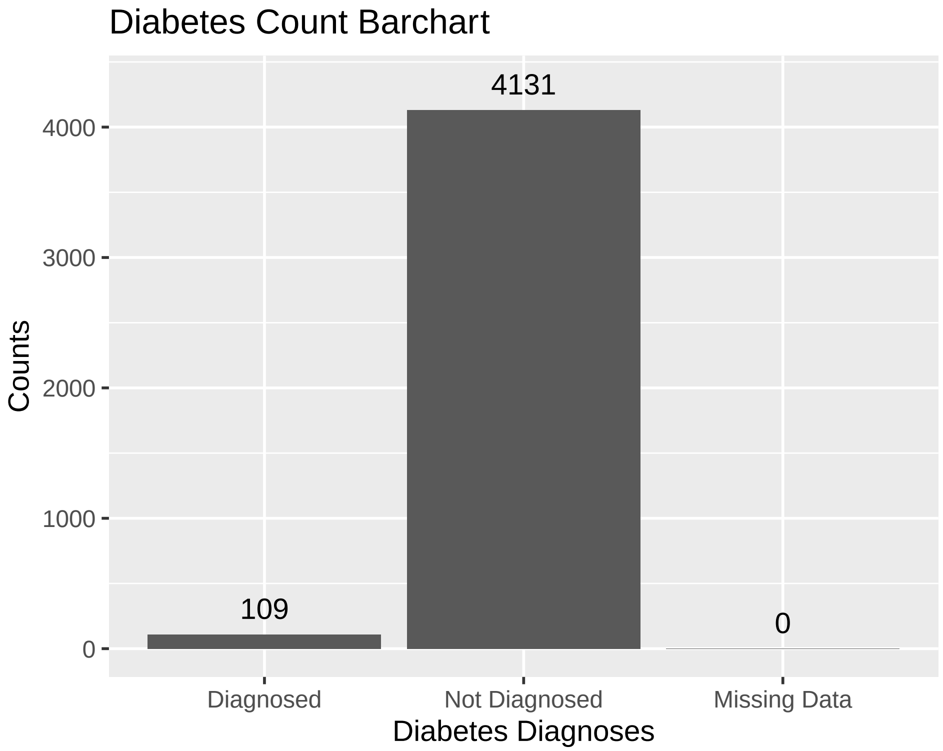
The eighth variable in our data set is **HighBloodPressure**, a dummy variable where a *0* represents that the patient did not have high blood pressure at the time of examination and a *1* represents that the patient did have high blood pressure at the time of examination. Below is a breakdown of **HighBloodPressure**:



As shown in the barchart above, our data set has *1317* patients who had high blood pressure at the time of examination, *4215* patients who did not have high blood pressure at the time of examination, and *0* patients who are missing high blood pressure data.

2.1.9 Diabetes

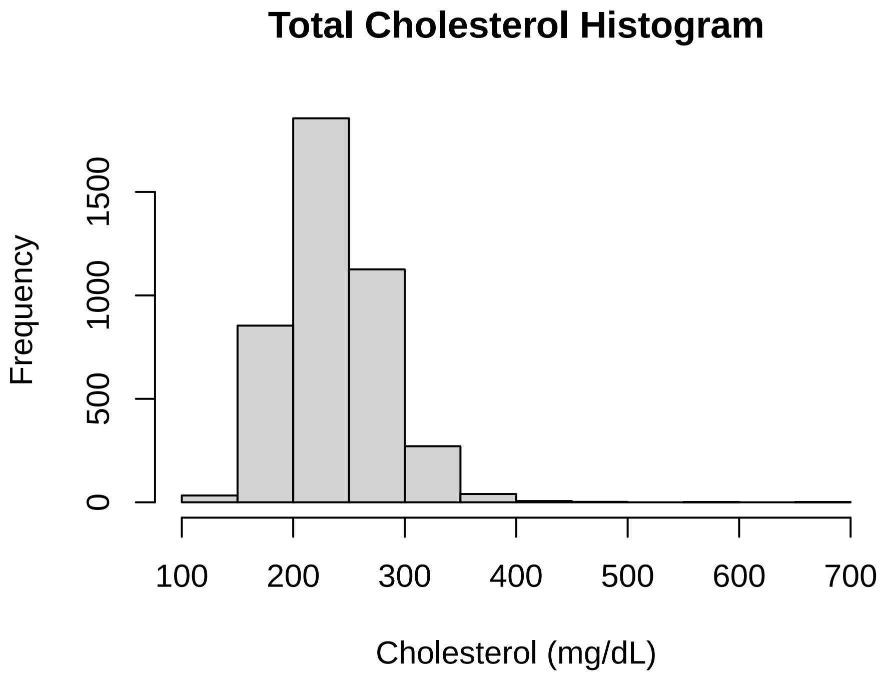
The ninth variable in our data set is **Diabetes**, a dummy variable where a *0* represents that the patient has not been diagnosed with diabetes and a *1* represents that the patient has been diagnosed diabetes. Below is a breakdown of **Diabetes**:



As shown in the barchart above, our data set has *109* patients who have been diagnosed with diabetes, *4131* patients who have not been diagnosed with diabetes, and *0* patients who are missing diabetes data.

2.1.10 Total Cholesterol

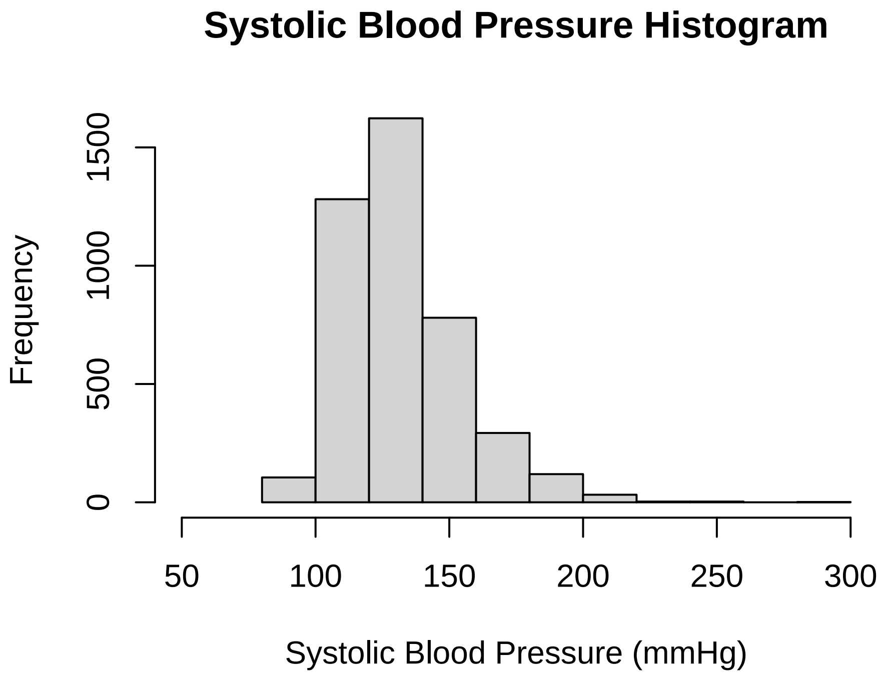
The tenth variable in our data set is **TotalCholesterol**, a continuous quantitative variable of the patients total cholesterol in milligrams per deciliter (mg/dL). Below is a breakdown of **TotalCholesterol**:



As shown above, the variable **TotalCholesterol** appears to be Normal, with some high outliers in the 600's that are hard to visualize due to the small amount of patients in the 600 range.

2.1.11 Systolic Blood Pressure

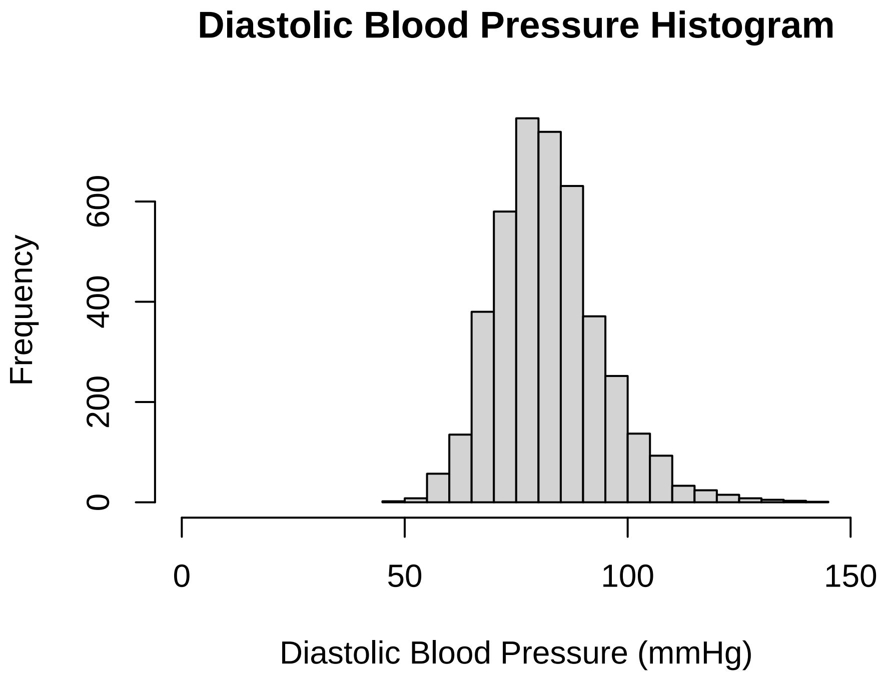
The eleventh variable in our data set is **SystolicBloodPressure**, a continuous quantitative variable of the patients pressure in their arteries when their heart beats, measured in millimeters of mercury (mmHg). Below is a breakdown of **SystolicBloodPressure**:



As shown above, the variable **SystolicBloodPressure** appears to be Normal.

2.1.12 Diastolic Blood Pressure

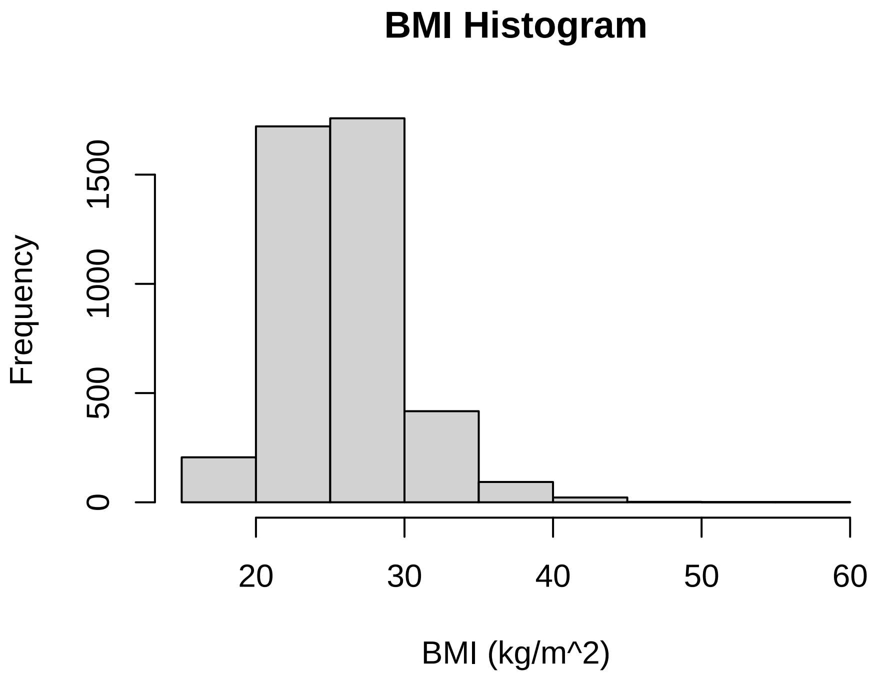
The twelfth variable in our data set is **DiastolicBloodPressure**, a continuous quantitative variable of the patients pressure in their arteries in between heart beats, measured in millimeters of mercury (mmHg). Below is a breakdown of **DiastolicBloodPressure**:



As shown above, the variable **DiastolicBloodPressure** appears to be Normal.

2.1.13 BMI

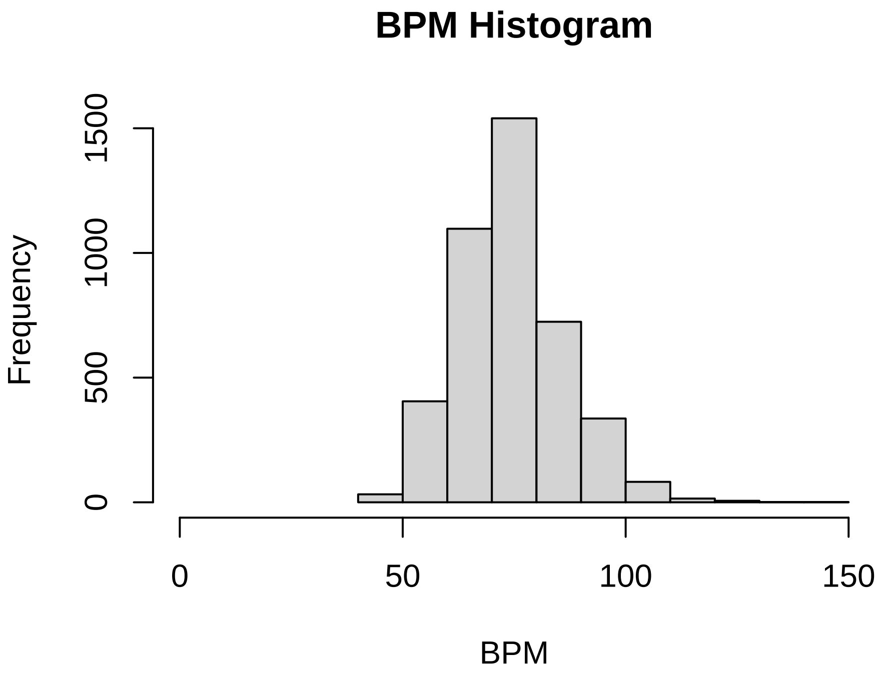
The thirteenth variable in our data set is **BMI**, a continuous quantitative variable that measures the patients height and weight, measured in kilograms over meters-squared(kg/mˆ2). Below is a breakdown of **BMI**:



As shown above, the variable **BMI** appears to be Normal, with some high outliers in the 50's and 60's that are hard to visualize due to the small amount of patients in the 50 and 60 range.

2.1.14 BPM

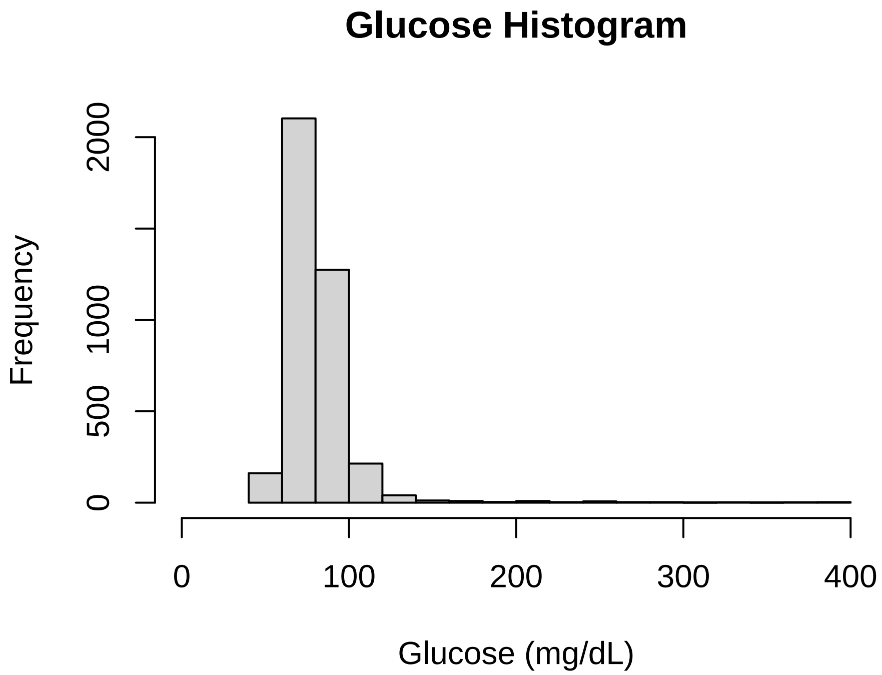
The fourteenth variable in our data set is **BPM**, a continuous quantitative variable that measures the patients heartbeats per minute. Below is a breakdown of **BPM**:



As shown above, the variable **BMI** appears to be Normal, with some high outliers in the 50's and 60's that are hard to visualize due to the small amount of patients in the 50 and 60 range.

2.1.15 Glucose Level

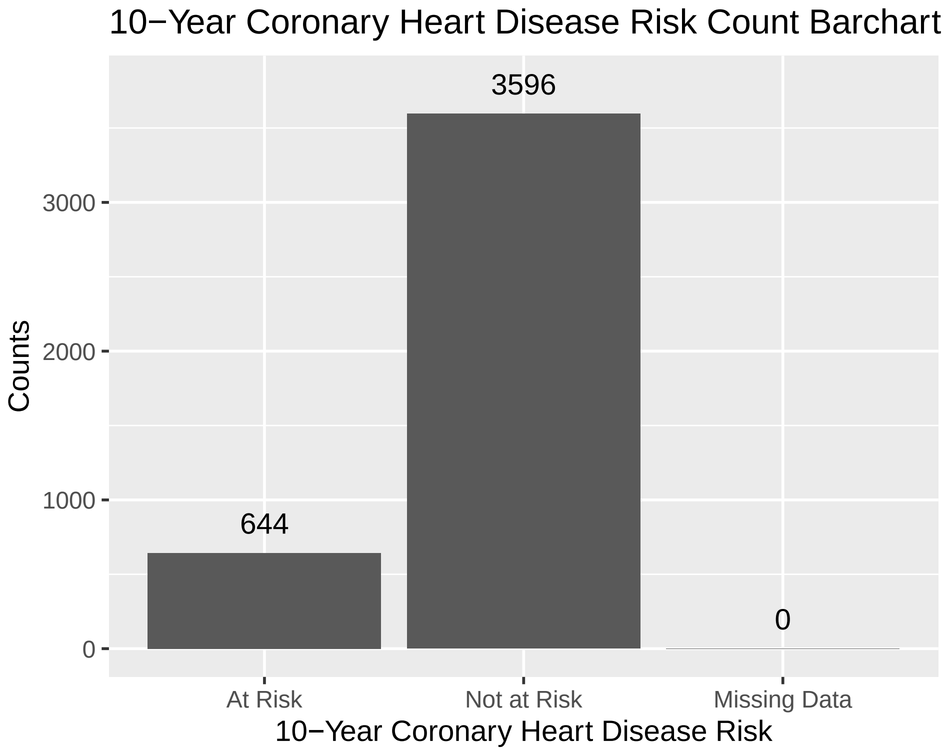
The fifteenth and final predictor variable in our data set is **GlucoseLevel**, a continuous quantitative variable that measures the patients blood glucose level in milligrams per deciliter (mg/dL). Below is a breakdown of **GlucoseLevel**:



As shown above, the variable **Glucose** appears to be Normal, with some high outliers in the 200's and 300's that are hard to visualize due to the small amount of patients in the 200 and 300 range.

2.1.16 Coronary Heart Disease

The response variable in our data set is **CoronaryHeartDisease**, a dummy variable where a **0** represents that the patient did not have a 10-year risk of contracting coronary heart disease and a **1** represents that the patient did have a 10-year risk of contracting coronary heart disease. Below is a breakdown of **CoronaryHeartDisease**:



As shown in the barchart above, our data set has *644* patients who were at risk of contracting coronary heart disease within 10 years, *3596* patients who were not at risk of contracting coronary heart disease within 10 years, and *0* patients who are missing their 10-year coronary heart disease risk data.

2.2 Missing Data Values

The variables that have missing data values are listed below:

- **Education**
- **CigarettesPerDay**
- **BloodPressureMeds**
- **TotalCholesterol**
- **BMI**
- **BPM**
- **GlucoseLevel**

To deal with the missing data, we can create a new temporary dataset that exclude these missing variables to create a logic regression model that we can then use to predict the values for the missing values. However, due many variables being highly correlated with the predictor variables with missing values, we cannot perform an imputation by simple linear regression. Thus, we can replace the missing values with the mode for the corresponding predictor variables given that there are so few missing values.

2.3 Variable Re-Coding

Next, we re-coded the predictor variables and then modified them as per the parameters below, many of which are similar if not identical to the groups from the variable breakdown portion of the report:

- **Gender**: *0* ~ Female, *1* ~ Male
- **Ages**: *30 to 39* ~ 30's, *40 to 49* ~ 40's, *50 to 59* ~ 50's, *60+* ~ 60+
- **EducationLevel**: *0* ~ No College, *1-3* ~ Some College, *4* ~ Graduated College
- **Smoking**: *0* ~ No, *1* ~ Yes,
- **CigsaDay**: *0* ~ 0, *1-10* ~ 1-10, *11-20* ~ 11-10, *21-30* ~ 21-30, *30+* ~ 30+
- **BPMedication**: *0*~ Not Prescribed, *1* ~ Prescribed
- **Stroke**: *0* ~ No, *1* ~ Yes
- **HighBP**: *0* ~ No, *1* ~ Yes
- **DiabetesStatus**: *0* ~ Not Diagnosed, *1* ~ Diagnosed
- **CHDRisk**: *0* ~ Not at Risk, *1* ~ At Risk

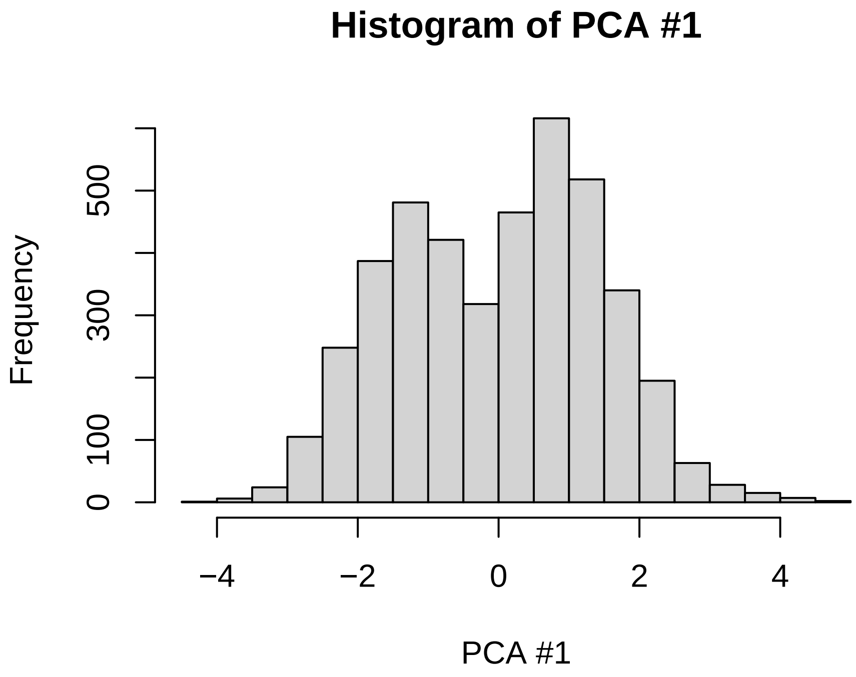
2.4 PCA's

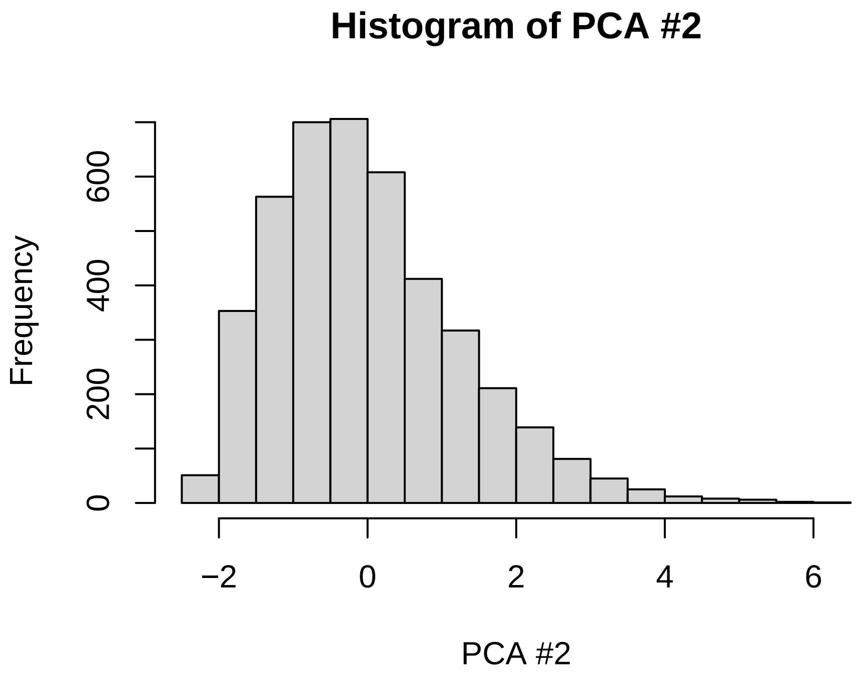
Next, we will performs both Principal Component Analysis (PCA) and Exploratory Factor Analysis (EFA). PCA's are utilized to help analyze large datasets that may be difficult to interpret and/or visualize. They can also be used to reduce the amount of dimensions of the dataset. In this instance, we will utilize the PCA's as an exploratory data analysis for the dataset above as well. Using PCA's in this way can help creative predictive models and interpret clearly through visualizations and statistics how each of the variables affects the response variable differently.

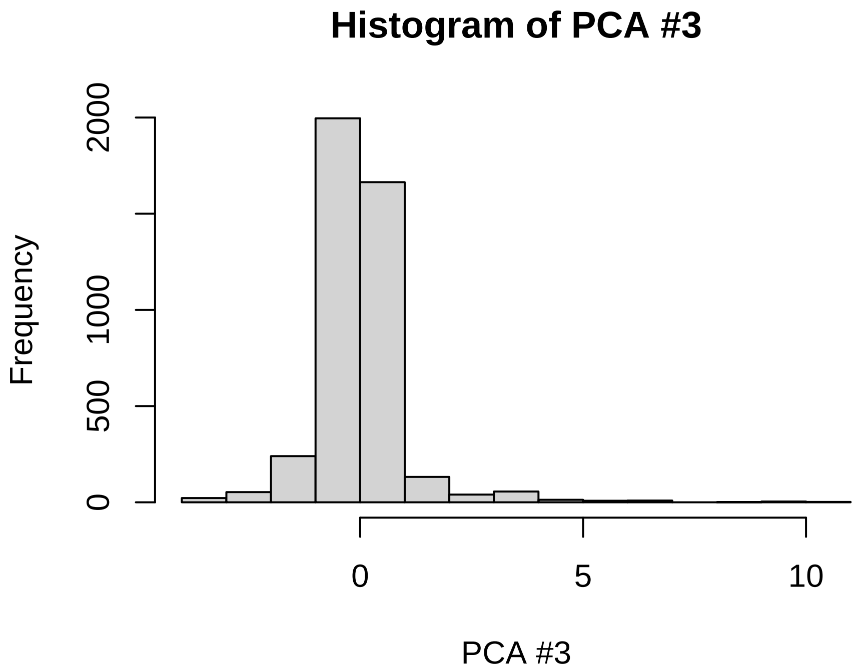
We begin by extracting the TEMP index from the dataset, and calculate the PCA's from this extracted TEMP index. Below are the results from building the PCA's

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.4444 1.2652 1.0315 0.9940 0.96822 0.96009 0.90689
## Proportion of Variance 0.2086 0.1601 0.1064 0.0988 0.09375 0.09218 0.08224
## Cumulative Proportion  0.2086 0.3687 0.4751 0.5739 0.66764 0.75981 0.84206
##                            PC8     PC9    PC10
## Standard deviation     0.84923 0.79349 0.47809
## Proportion of Variance 0.07212 0.06296 0.02286
## Cumulative Proportion  0.91418 0.97714 1.00000
```

As seen above, the first three PCA's hold almost *50%* of the variance. To ensure that these PCA's are Normal, below are histograms of the first three PCA's.


Histogram of PCA #1
Frequency
0
100
300
500
−4
−2
0
2
4
PCA #1


Histogram of PCA #2
Frequency
0
200
400
600
−2
0
2
4
6
PCA #2



As shown above, PCA's 1 and 2 appear Normal, but PCA 3 does not. Therefore, we will keep **PC1** and **PC2** and add them to our final data set.

2.5 Final Analytical Data Set

Now that we have replaced the missing values for all variables and defined the predictor variables and the response variable, along with calculating the PCA's, we have our final analytical data set and can move forward with analysis.

3 Analysis

3.1 Logistic Regression

Given that our response variable **CHDRisk** is a dummy variable, we will be running a logistic regression analysis utilizing different predictor variables to determine the best model for predicting Coronary Heart Disease.

3.1.1 Model 1: All Predictors

The very first model that we are going to consider, named *Model1*, will include all possible predictor variables including both categorical and numerical variables. It is common practice when generating possible regression models to begin with a model which includes every possible predictor variable to analyze which variables overall are important to keep in further testing. Below shows a summary of *Model1*.

```
##
```

```
## Call:
## glm(formula = response ~ Gender.cat + Ages.cat + EducationLevel.cat +
##     Smoking.cat + CigsaDay.cat + BPMedication.cat + Stroke.cat +
##     HighBP.cat + DiabetesStatus.cat + totChol + sysBP + diaBP +
##     BMI + heartRate + glucose + pca1 + pca2, data = final.data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.68690  -0.06422  -0.00146   0.06136   0.36841
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          1.271e+00  2.290e-02  55.519  < 2e-16 ***
## Gender.catMale                      -4.996e-01  8.550e-03 -58.433  < 2e-16 ***
## Ages.cat40’s                        -1.659e-01  5.199e-03 -31.916  < 2e-16 ***
## Ages.cat50’s                        -3.964e-01  8.129e-03 -48.764  < 2e-16 ***
## Ages.cat60+                         -5.961e-01  1.142e-02 -52.191  < 2e-16 ***
## EducationLevel.catSome College       1.894e-01  3.622e-03  52.286  < 2e-16 ***
## EducationLevel.catGraduated College  4.123e-01  5.901e-03  69.868  < 2e-16 ***
## Smoking.catYes                      -5.369e-01  2.280e-02 -23.546  < 2e-16 ***
## CigsaDay.cat1-10                    -1.485e-01  1.917e-02  -7.746 1.18e-14 ***
## CigsaDay.cat11-20                   -4.599e-01  2.178e-02 -21.120  < 2e-16 ***
## CigsaDay.cat21-30                   -7.120e-01  2.585e-02 -27.541  < 2e-16 ***
## CigsaDay.cat30+                     -1.051e+00  3.190e-02 -32.941  < 2e-16 ***
## BPMedication.catPrescribed          -1.230e+00  1.373e-02 -89.556  < 2e-16 ***
## Stroke.catYes                       -1.491e+00  2.337e-02 -63.817  < 2e-16 ***
## HighBP.catYes                       -6.388e-01  6.957e-03 -91.815  < 2e-16 ***
## DiabetesStatus.catDiagnosed         -8.538e-01  1.360e-02 -62.759  < 2e-16 ***
## totChol                             -1.360e-04  3.616e-05  -3.763  0.00017 ***
## sysBP                               -2.852e-04  1.339e-04  -2.129  0.03330 *
## diaBP                                3.342e-04  2.210e-04   1.512  0.13065
## BMI                                  3.147e-04  4.156e-04   0.757  0.44892
## heartRate                           -7.206e-05  1.322e-04  -0.545  0.58561
## glucose                              2.211e-04  8.432e-05   2.622  0.00877 **
## pca1                                -1.139e-01  1.075e-02 -10.600  < 2e-16 ***
## pca2                                 7.346e-01  4.148e-03 177.097  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 0.009734069)
##
##     Null deviance: 546.185  on 4239  degrees of freedom
## Residual deviance:  41.039  on 4216  degrees of freedom
## AIC: -7581.7
##
## Number of Fisher Scoring iterations: 2
```

As shown in the summary above, there are various variables that are statistically significant like **EducationLevel** and **DiabetesStatus**, while other variable such as **diaBP** and **BMI** are not statistically significant.

We will do comparative tests once more models have been developed.

3.1.2 Model 2: Categorical Predictors

Our second model that we are going to consider will include only categorical predictors, which include **Gender**, **Ages**, **EducationalLevel**, **Smoking**, **CigsaDay**, **BPMedication**, **Stroke**, **HighBP**, and **DiabetesStatus**. We will still include **PCA1** and **PCA2** in this model as the PCA's help reduce the dimentions within the dataset. This model, named *Model2*, if chosen, will help healthcare professionals predict if the patient will be diagnosed with Coronary Heart Disease without having to do any medical tests on the patient. Below shows a summary of *Model2*.

```
##
## Call:
## glm(formula = response ~ Gender.cat + Ages.cat + EducationLevel.cat +
##     Smoking.cat + CigsaDay.cat + BPMedication.cat + Stroke.cat +
##     HighBP.cat + DiabetesStatus.cat + pca1 + pca2, data = final.data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.68486  -0.06489  -0.00192   0.06160   0.37330
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         1.254382   0.012383 101.296  < 2e-16 ***
## Gender.catMale                     -0.497927   0.008539 -58.310  < 2e-16 ***
## Ages.cat40’s                       -0.167604   0.005180 -32.354  < 2e-16 ***
## Ages.cat50’s                       -0.400115   0.008091 -49.452  < 2e-16 ***
## Ages.cat60+                        -0.600226   0.011408 -52.615  < 2e-16 ***
## EducationLevel.catSome College      0.188652   0.003599  52.420  < 2e-16 ***
## EducationLevel.catGraduated College  0.411558   0.005884  69.949  < 2e-16 ***
## Smoking.catYes                     -0.539599   0.022819 -23.647  < 2e-16 ***
## CigsaDay.cat1-10                   -0.148394   0.019200  -7.729 1.35e-14 ***
## CigsaDay.cat11-20                  -0.461991   0.021809 -21.184  < 2e-16 ***
## CigsaDay.cat21-30                  -0.715400   0.025889 -27.633  < 2e-16 ***
## CigsaDay.cat30+                    -1.055136   0.031941 -33.034  < 2e-16 ***
## BPMedication.catPrescribed         -1.229436   0.013749 -89.419  < 2e-16 ***
## Stroke.catYes                      -1.485129   0.023336 -63.642  < 2e-16 ***
## HighBP.catYes                      -0.641182   0.006457 -99.306  < 2e-16 ***
## DiabetesStatus.catDiagnosed        -0.834947   0.011733 -71.164  < 2e-16 ***
## pca1                               -0.115750   0.010742 -10.775  < 2e-16 ***
## pca2                                0.734269   0.004151 176.908  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 0.009780909)
##
##     Null deviance: 546.185  on 4239  degrees of freedom
## Residual deviance:  41.295  on 4222  degrees of freedom
## AIC: -7567.3
##
## Number of Fisher Scoring iterations: 2
```

As shown above, all variables are statistically significant at the $\alpha = 0.5$ level. Since *Model2* is a nested model of *Model1*, we will run a likelihood ratio test to compare the two models. Below are the results from the likelihood ratio test.

```
## Likelihood ratio test
##
## Model 1: response ~ Gender.cat + Ages.cat + EducationLevel.cat + Smoking.cat +
##     CigsaDay.cat + BPMedication.cat + Stroke.cat + HighBP.cat +
##     DiabetesStatus.cat + totChol + sysBP + diaBP + BMI + heartRate +
##     glucose + pca1 + pca2
## Model 2: response ~ Gender.cat + Ages.cat + EducationLevel.cat + Smoking.cat +
##     CigsaDay.cat + BPMedication.cat + Stroke.cat + HighBP.cat +
##     DiabetesStatus.cat + pca1 + pca2
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1  25 3815.8
## 2  19 3802.6 -6 26.383  0.0001889 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As shown above, we get a Chi-Squared test statistic of *26.383* and p-value of *0.0001889*, which is statistically significant. This means that *Model2* performs worse than *Model1*, meaning the numerical variables are important Coronary Heart Disease diagnoses

3.1.3 Model 3: Numerical Predictors

Our third model that we are going to consider will include only numerical predictors, which include **totChol**, **sysBP**, **diaBP**, **BMI**, **heartRate**, and **glucose**. Again, both PCA's will be included This model, named *Model3*, if chosen, will show that medical tests alone are the best way to predict if the patient is going to be diagnosed with Coronary Heart Disease. Below shows a summary of *Model3*.

```
##
## Call:
## glm(formula = response ~ totChol + sysBP + diaBP + BMI + heartRate +
##     glucose + pca1 + pca2, data = final.data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.14411  -0.15701  -0.04207   0.07963   0.96536
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.6866524  0.0526951  13.031  < 2e-16 ***
## totChol     -0.0001404  0.0001013  -1.385   0.1661
## sysBP       -0.0026049  0.0003633  -7.171 8.78e-13 ***
## diaBP       -0.0007251  0.0006067  -1.195   0.2321
## BMI         -0.0034422  0.0011670  -2.950   0.0032 **
## heartRate    0.0002389  0.0003717   0.643   0.5205
## glucose     -0.0003170  0.0001979  -1.601   0.1094
## pca1         0.0549693  0.0034765  15.812  < 2e-16 ***
## pca2         0.1926334  0.0041265  46.682  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.08011021)
##
##     Null deviance: 546.18  on 4239  degrees of freedom
## Residual deviance: 338.95  on 4231  degrees of freedom
## AIC: 1340.3
##
## Number of Fisher Scoring iterations: 2
```

As shown above, only some variables are statistically significant at the $\alpha = 0.5$ level. Below are the results from the likelihood ratio tests comparing *Model3* to *Model1*.

```
## Likelihood ratio test
##
## Model 1: response ~ Gender.cat + Ages.cat + EducationLevel.cat + Smoking.cat +
##     CigsaDay.cat + BPMedication.cat + Stroke.cat + HighBP.cat +
##     DiabetesStatus.cat + totChol + sysBP + diaBP + BMI + heartRate +
##     glucose + pca1 + pca2
## Model 2: response ~ totChol + sysBP + diaBP + BMI + heartRate + glucose +
##     pca1 + pca2
##   #Df LogLik  Df Chisq Pr(>Chisq)
## 1  25 3815.8
## 2  10 -660.2 -15  8952  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As shown above, we get a Chi-Squared test statistic of *8952* and p-value of practically *0*. This means that *Model3* performs worse than *Model1*, meaning the categorical variables are important Coronary Heart Disease diagnoses.

3.2 Final Model

Out of the three models, *Model1* is the best model for predicting Coronary Heart Disease diagnoses. *Model1* includes both the numerical predictor variables and categorical predictor variables. This means that medical professionals need to perform the standard medical testing along with looking at demographical information screening for Coronary Heart Disease risk. Below is the final model formula.

$CHDRisk = 1.271 - 0.4996 * (Gender.Male) - 0.1659 * (Ages.40s) - 0.3964 * (Ages.50s) - 0.5961 * (Ages.60Plus) + 0.1894 * (Education.SomeCollege) + 0.4123 * (Education.GraduatedCollege) - 0.5369 * (Smoking.Yes) - 0.1485 * (CigsaDay.1to10) - 0.4599 * (CigsaDay.11to20) - 0.712 * (CigsaDay.21to30) - 1.051 * (CigsaDay.30Plus) - 1.23 * (BPMedication.Prescribed) - 1.491 * (Stroke.Yes) - 0.6388 * (HighBP.Yes) - 0.8538 * (Diabetes.Diagnosed) - 0.000136 * (totChol) - 0.0002852 * (sysBP) + 0.0003342 * (diaBP) + 0.0003147 * (BMI) - 0.00007206 * (heartRate) + 0.0002211 * (glucose) - 0.1139 * (pca1) + 0.7346 * (pca2)$

It should be noted that in the above model, some variables such as **CigsaDay** are repeated in the model. The condition after the period represents that if the patent falls into that category, therefore the other categories of the variable should be ignored when using this model in application (i.e., if the patient smokes 15 cigarettes a day, only use the *CigsaDay.11to20* condition in the model and ignore the *CigsaDay.1to10*, *CigsaDay.21to30*, *CigsaDay.30Plus*). This is similar to if the patient does not fall into a category (i.e., if the patient is 32, ignore all categories of **Ages**).

4 Conclusion and Discussion

In conclusion, *Model1* was the best model for our sample data. Further analyses can be done if need be, as no models using interaction nor hierarchical terms were considered for this analysis. This can be attributed to the variables no having an obvious correlation with each other in an applied sense where having interaction or hierarchical terms would make sense.

When using this model in application, the user must be considerate that the sample data used to build the model was from a specific location during a specific time period. Therefore, predicting the Coronary Heart Disease diagnosis risk of patients during the current times and in different locations using our model might not be as accurate given that there have been numerous medical advances since 1948 and the 1971 re-roster.