The Framingham Heart Study

Descriptive Statistics & Exploratory Analysis with Visualization

Muhammad Affan Mansoor Malik

Marquette University

COSC 5610: Data Mining

11/05/2020

Introduction

In our literature review I discussed about the Cardiovascular Disease and a study done on residents of Framingham, MA by researchers to understand the common factors that contribute to the development of CVD in the United States.

10 Step Process

Before I delve into the statistics and visualization, I believe it is important to discuss the steps that are taken for a successful data mining project. This is something I learned in one of my previous courses (Business Intelligence) and is something I will follow in our project. Out of which step 1 & 2 have been discussed in the literature review.

Here is what those 10 steps look like:
1. Develop an understanding of the purpose of the data mining project.
2. Obtain the dataset to be used in the analysis.
3. Explore, clean, and preprocess the data.
4. Reduce the data dimension, if necessary.
5. Determine the data mining task.
6. Partition the data (for supervised tasks).
7. Choose the data mining techniques to be used.
8. Use algorithms to perform the task.
9. Interpret the results of the algorithms.
10. Deploy the model.

You can find the full explanation of these bullet points at here[1].

Descriptive Statistics & Exploratory Analysis with Visualization

The dataset was downloaded from Kaggle and contained demographic, behavioral, medical and predict variables of patients which had over 4000 records and 16 attributes.

In this assignment we will be using R/RStudio to find descriptive statistical data from the dataset and use visual techniques for further data exploration.

Reading & importing the data:

```
FHR <- read.csv("D:/affanmansoor/Downloads/Marquette/DataMining/framingham.csv")
```

Let's now check the internal structure of a R object we just created in a compact display with the function `str()`.

```
> str(FHR)
'data.frame':   4240 obs. of  16 variables:
 $ male           : int  1 0 1 0 0 0 0 0 1 1 ...
 $ age            : int  39 46 48 61 46 43 63 45 52 43 ...
 $ education      : int  4 2 1 3 3 2 1 2 1 1 ...
 $ currentSmoker  : int  0 0 1 1 1 0 0 1 0 1 ...
 $ cigsPerDay     : int  0 0 20 30 23 0 0 20 0 30 ...
 $ BPMeds         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ prevalentStroke: int  0 0 0 0 0 0 0 0 0 0 ...
 $ prevalentHyp   : int  0 0 0 1 0 1 0 0 1 1 ...
 $ diabetes       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ totChol        : int  195 250 245 225 285 228 205 313 260 225 ...
 $ sysBP          : num  106 121 128 150 130 ...
 $ diaBP          : num  70 81 80 95 84 110 71 71 89 107 ...
 $ BMI            : num  27 28.7 25.3 28.6 23.1 ...
 $ heartRate      : int  80 95 75 65 85 77 60 79 76 93 ...
 $ glucose        : int  77 76 70 103 85 99 85 78 79 88 ...
 $ TenYearCHD     : int  0 0 0 1 0 0 1 0 0 0 ...
```

We will now check the head and tail of the dataset just to have a quick peak at the dataset instead of viewing it completely.

```
> head(FHR)
  male age education currentSmoker cigsPerDay BPMeds prevalentStroke prevalentHyp diabetes totChol sysBP diaBP   BMI heartRate glucose TenYearCHD
1    1  39         4             0          0      0                0            0        0     195 106.0    70 26.97        80      77          0
2    0  46         2             0          0      0                0            0        0     250 121.0    81 28.73        95      76          0
3    1  48         1             1         20      0                0            0        0     245 127.5    80 25.34        75      70          0
4    0  61         3             1         30      0                0            1        0     225 150.0    95 28.58        65     103          1
5    0  46         3             1         23      0                0            0        0     285 130.0    84 23.10        85      85          0
6    0  43         2             0          0      0                0            1        0     228 180.0   110 30.30        77      99          0
> tail(FHR)
     male age education currentSmoker cigsPerDay BPMeds prevalentStroke prevalentHyp diabetes totChol sysBP diaBP   BMI heartRate glucose TenYearCHD
4235    1  51         3             1         43      0                0            0        0     207 126.5    80 19.71        65      68          0
4236    0  48         2             1         20     NA                0            0        0     248 131.0    72 22.00        84      86          0
4237    0  44         1             1         15      0                0            0        0     210 126.5    87 19.16        86      NA          0
4238    0  52         2             0          0      0                0            0        0     269 133.5    83 21.47        80     107          0
4239    1  40         3             0          0      0                0            1        0     185 141.0    98 25.60        67      72          0
4240    0  39         3             1         30      0                0            0        0     196 133.0    86 20.91        85      80          0
>
```

Upon quick inspection, I have noticed that there are NA values present in the dataset which can end up in hindering our results & end goal so it will be better if we left out those rows that contain these incomplete values.

```
> sum(is.na(FHR))
[1] 645
> sum(apply(FHR, 1, anyNA))
[1] 582
```

```
> sum(complete.cases(FHR))
[1] 3658
>
```

In this data set, 3658 rows are complete whereas there are 582 rows and 645 cells with NA values. Out of which cigsPerDay has 29 NAs, BPMeds has 53 NAs, totChol has 50NAs, glucose has 388 NAs, heartRate has 1 NAs and BMI has 19 NAs. When taken into consideration this ends up being only 7.28% of our data and would not affect our analysis upon being excluded/discarded.

To remove these rows we will use `FHS <- na.omit(FHR)`.

```
> sum(is.na(FHS))
[1] 0
> sum(complete.cases(FHS))
[1] 3658
```

After exploring, cleaning and preprocessing the dataset, let's begin with our analysis.

We want to know the age summary of study participants:

summary(age), sd(age)

```
summary(age)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
32.00   42.00   49.00   49.55   56.00   70.00
```

```
> sd(age)
[1] 8.562029
```

Mean age of participants is 49.55 years with a standard deviation of 8.56 years. The minimum age is 32 years, and the maximum is 70 years.

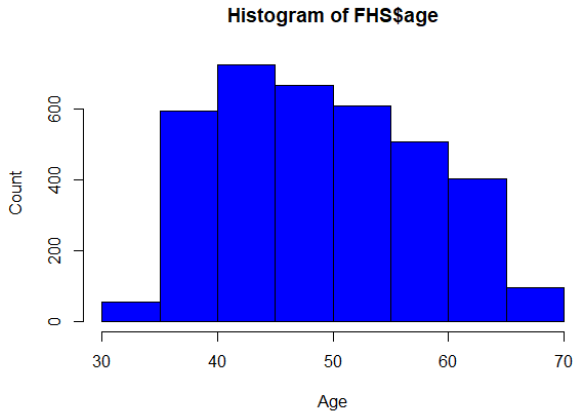
Let's look at the number of Male and Female participants in the study.

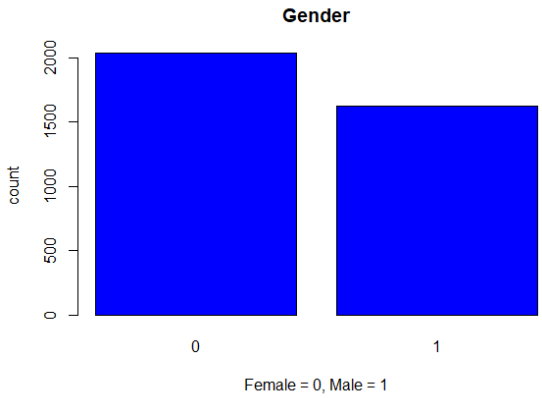
table(FHS$male)

There are 1623 males and 2035 females in our updated dataset.

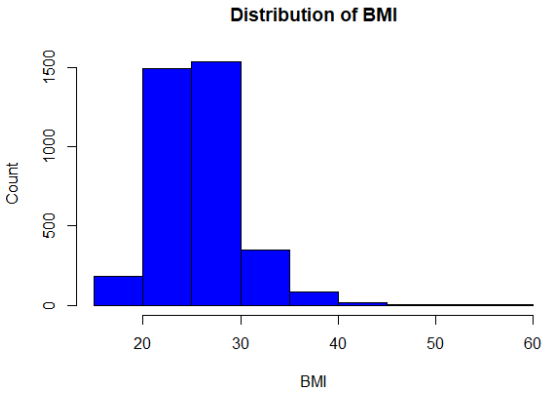
Visualization

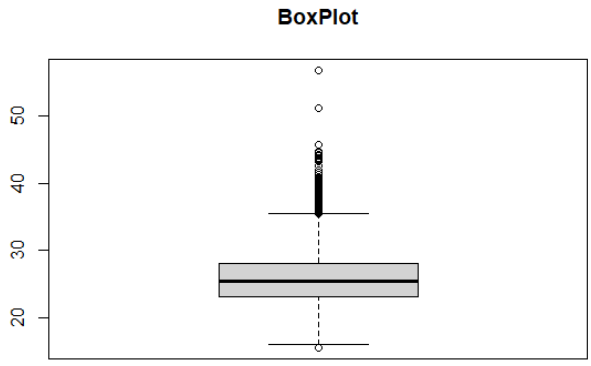
Male:Female ratio, there are more females than male in this study.

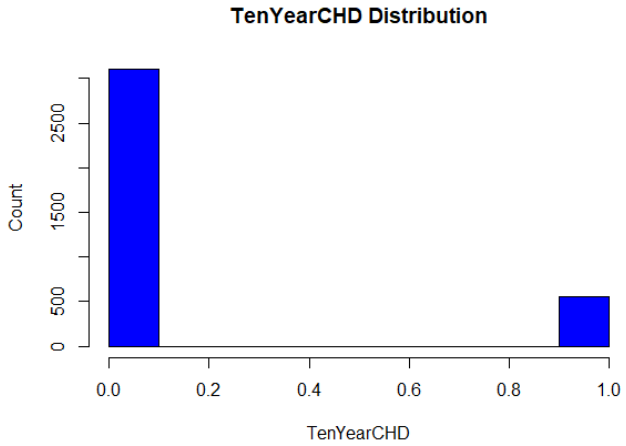


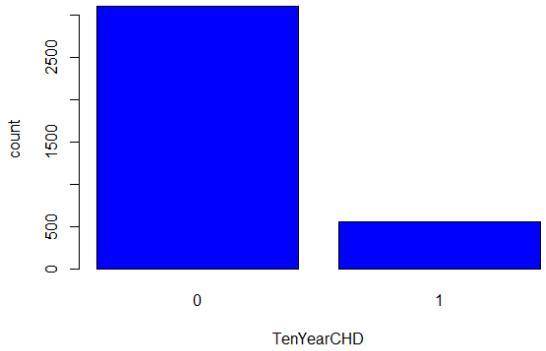


BoxPlot shows that BMI's normal distribution with mean BMI equals to 25.78 and median BMI equals to 25.38.

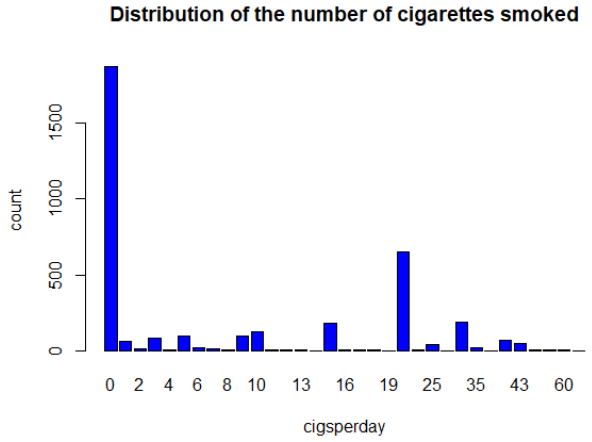


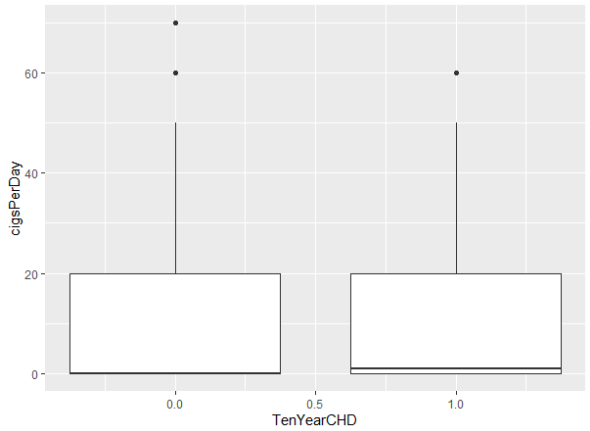


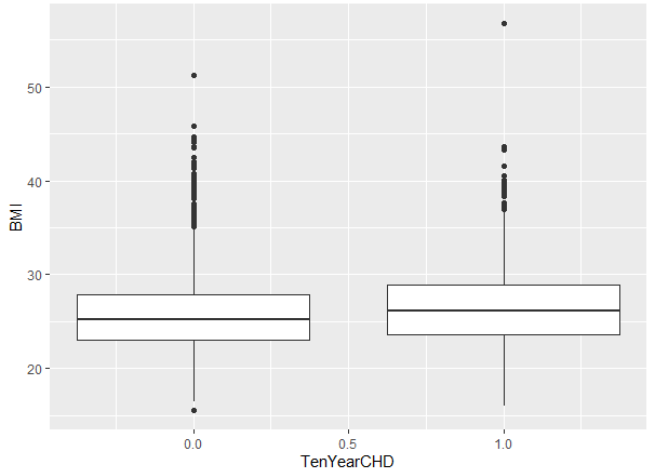


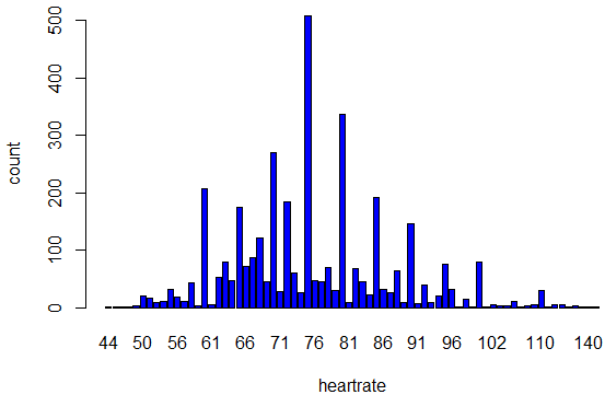


Only 15% of patients have been diagnosed with a CHD.









We will pick two variables, namely the male and female gender and come up with the following conclusions.

Studying the Male Gender

```
> mod4 <- lm( TenYearCHD ~ diabetes + prevalentHyp+currentSmoker, data = FHSmale, family=binomial)
Warning message:
In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :
 extra argument 'family' will be disregarded
> mod4

Call:
lm(formula = TenYearCHD ~ diabetes + prevalentHyp + currentSmoker,
    data = FHSmale, family = binomial)

Coefficients:
  (Intercept)       diabetes   prevalentHyp  currentSmoker
      0.11771        0.22947        0.12799        0.04082

> summary(mod4)

Call:
lm(formula = TenYearCHD ~ diabetes + prevalentHyp + currentSmoker,
    data = FHSmale, family = binomial)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5160 -0.1585 -0.1585 -0.1177  0.8823

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.11771    0.01689   6.971 4.56e-12 ***
diabetes       0.22947    0.05663   4.052 5.32e-05 ***
prevalentHyp   0.12799    0.02066   6.194 7.41e-10 ***
currentSmoker  0.04082    0.01962   2.080   0.0377 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3852 on 1619 degrees of freedom
Multiple R-squared:  0.03513,   Adjusted R-squared:  0.03334
F-statistic: 19.65 on 3 and 1619 DF,  p-value: 1.637e-12
```

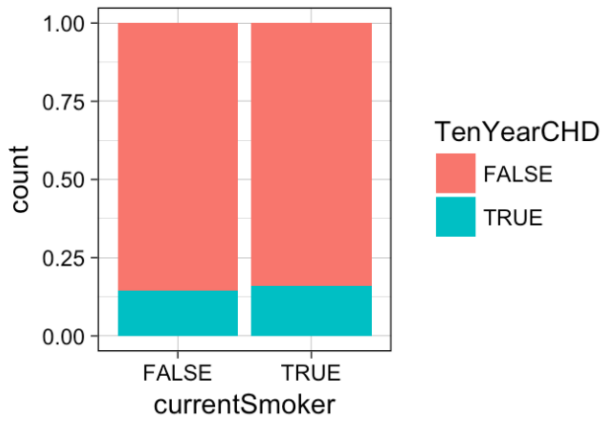
Studying the Female Gender

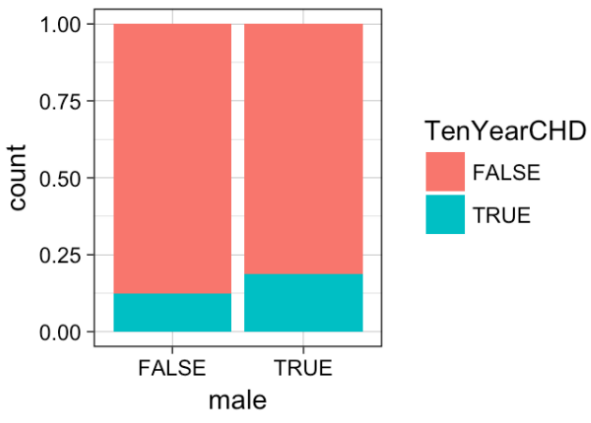
```
> mod5 <- lm( TenYearCHD ~  diabetes + prevalentHyp + currentSmoker, data = FHS ,family=binomial)
Warning message:
In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :
 extra argument 'family' will be disregarded
> mod5

Call:
lm(formula = TenYearCHD ~ diabetes + prevalentHyp + currentSmoker,
    data = FHS, family = binomial)

Coefficients:
  (Intercept)       diabetes   prevalentHyp  currentSmoker
      0.08932        0.17875        0.13917        0.03014

> summary(mod5)

Call:
lm(formula = TenYearCHD ~ diabetes + prevalentHyp + currentSmoker,
    data = FHS, family = binomial)

Residuals:
     Min       1Q   Median       3Q      Max
-0.43738 -0.11946 -0.11946 -0.08932  0.91068

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.089318   0.009372   9.530  < 2e-16 ***
diabetes      0.178750   0.036008   4.964 7.22e-07 ***
prevalentHyp  0.139174   0.012679  10.977  < 2e-16 ***
currentSmoker 0.030140   0.011720   2.572   0.0102 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3521 on 3654 degrees of freedom
Multiple R-squared:  0.04089,   Adjusted R-squared:  0.0401
F-statistic: 51.93 on 3 and 3654 DF,  p-value: < 2.2e-16
```

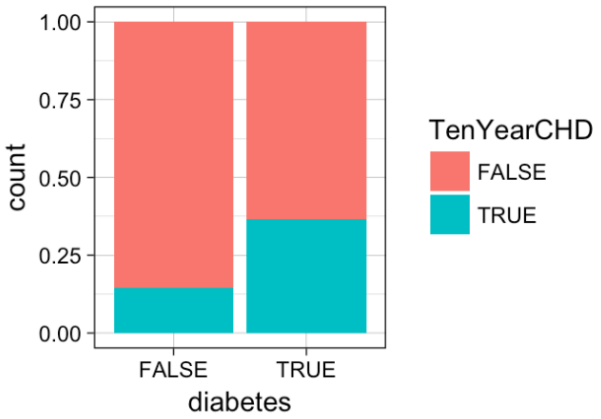
Studying the Age

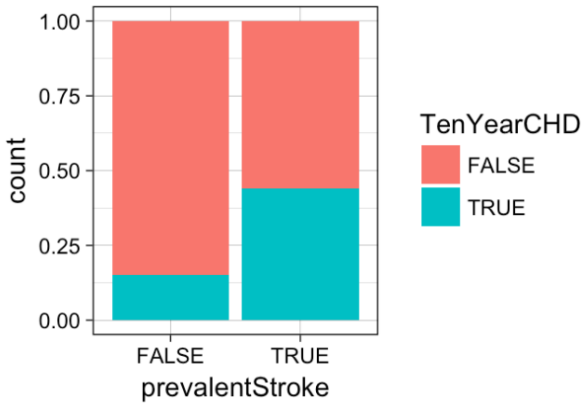
```
> mod3 <- glm( TenYearCHD ~  male+diabetes + prevalentHyp+currentSmoker, data = FHSage50, family=binomial)

> mod3

Call:  glm(formula = TenYearCHD ~ male + diabetes + prevalentHyp + currentSmoker,
    family = binomial, data = FHSage50)

Coefficients:
  (Intercept)           male       diabetes   prevalentHyp  currentSmoker
      -1.9372         0.5015         0.7275         0.7698         0.2614

Degrees of Freedom: 1615 Total (i.e. Null);  1611 Residual
Null Deviance:      1760
Residual Deviance: 1692         AIC: 1702
> summary(mod3)

Call:
glm(formula = TenYearCHD ~ male + diabetes + prevalentHyp + currentSmoker,
    family = binomial, data = FHSage50)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3173  -0.7361  -0.6534  -0.5189   2.0356

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.9372     0.1205 -16.080  < 2e-16 ***
male            0.5015     0.1262   3.973 7.09e-05 ***
diabetes        0.7275     0.2556   2.846  0.00443 **
prevalentHyp    0.7698     0.1222   6.299 2.99e-10 ***
currentSmoker   0.2614     0.1275   2.050  0.04041 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1760.4  on 1615  degrees of freedom
Residual deviance: 1691.8  on 1611  degrees of freedom
AIC: 1701.8

Number of Fisher Scoring iterations: 4
```

To illustrate the statistical dominance of males with TenYearCHD in this study, here are the following plots:









Based on what we have here, being male is directly related to TenYearCHD, thus the variable male seems a relatively good predictor. Similarly, Age seems a good predictor since the patients with TenYearCHD=TRUE, have higher median of age, with almost a similar distribution. In contrast, there seems no relation between different categories of the education and the response variable which is why no graphs were plotted. The current Smoker variable shows a slight relation with the response variable, as the current smokers have a slightly higher risk of TenYearCHD.

References

[1] 10Step Process www.shorturl.at/mxVY2

Centers for Disease Control and Prevention. Underlying Cause of Death, 1999–2018. CDC WONDER Online Database. Atlanta, GA: Centers for Disease Control and Prevention; 2018. Accessed March 12, 2020. https://www.cdc.gov/heartdisease/facts.htm

Rosano GM, Vitale C, Seferovic P. Heart Failure in Patients with Diabetes Mellitus. Card Fail Rev. 2017;3(1):52-55. doi:10.15420/cfr.2016:20:2

Barplot https://www.rdocumentation.org/packages/graphics/versions/3.6.2/topics/barplot

Histogram https://www.datamentor.io/r-programming/histogram/

Boxplot https://www.datamentor.io/r-programming/box-plot/#:~:text=In%20R%2C%20boxplot%20(and%20whisker,numeric%20vectors%20as%20its%20components.