Final Project R Script

Alexander Hernandez, Tony Ortiz, Diego Ramirez

12/7/2022

0.1) Source

"The dataset is publically available on the Kaggle website, and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD).The dataset provides the patients' information. It includes over 4,000 records and 15 attributes."

https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression?resource=download

0.2) Importing data:

```
heart = read.csv("C:\\repos\\STAT 50001\\Final Project\\framingham.csv")

head(heart)
```

```
##   male age education currentSmoker cigsPerDay BPMeds prevalentStroke
## 1    1  39         4             0          0      0               0
## 2    0  46         2             0          0      0               0
## 3    1  48         1             1         20      0               0
## 4    0  61         3             1         30      0               0
## 5    0  46         3             1         23      0               0
## 6    0  43         2             0          0      0               0
##   prevalentHyp diabetes totChol sysBP diaBP   BMI heartRate glucose TenYearCHD
## 1            0        0     195 106.0    70 26.97        80      77          0
## 2            0        0     250 121.0    81 28.73        95      76          0
## 3            0        0     245 127.5    80 25.34        75      70          0
## 4            1        0     225 150.0    95 28.58        65     103          1
## 5            0        0     285 130.0    84 23.10        85      85          0
## 6            1        0     228 180.0   110 30.30        77      99          0
```

```
attach(heart)
```

1) Introduction

Relevancy and Data/Variable Information ## Relevancy ### Demographics and Risk Factors sex, age, smoking status, cholesterol level, blood pressure, etc... ### Research Topic Trends and Changes in patient demographics, risk factors, and medical statistics may be used to predict heart disease. ## Data Information ### Demographic: Sex: male or female(Nominal) Age: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous) ### Behavioral Current Smoker: whether or not the patient is a current smoker (Nominal) Cigs Per Day: the number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarettes, even half a cigarette.) ### Medical(history) BP Meds: whether or not the patient was on blood pressure medication (Nominal) Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal) Prevalent Hyp: whether or not the patient was hypertensive (Nominal) Diabetes: whether or not the patient had diabetes (Nominal) ### Medical(current) Tot Chol: total cholesterol level (Continuous) Sys BP: systolic blood pressure (Continuous) Dia BP: diastolic blood pressure (Continuous) BMI: Body Mass Index (Continuous) Heart Rate: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.) Glucose: glucose level (Continuous) ### Predict variable (desired target) 10 year risk of coronary heart disease CHD (binary: “1”, means “Yes”, “0” means “No”)

2) Model and Logistic Equation Creation ## Model Creation

```
model = glm(TenYearCHD ~ .,
            family = binomial(logit),
            data = heart)
```

Significant Variables:

```
summary(model)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ ., family = binomial(logit), data = heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9581  -0.5944  -0.4267  -0.2829   2.8402
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -8.322206   0.715476 -11.632  < 2e-16 ***
## male             0.555098   0.109046   5.091 3.57e-07 ***
## age              0.063453   0.006680   9.499  < 2e-16 ***
## education       -0.047497   0.049390  -0.962  0.33621
## currentSmoker    0.070875   0.156749   0.452  0.65115
## cigsPerDay       0.017929   0.006238   2.874  0.00405 **
## BPMeds           0.162255   0.234309   0.692  0.48863
## prevalentStroke  0.693502   0.489532   1.417  0.15658
## prevalentHyp     0.234638   0.138037   1.700  0.08917 .
## diabetes         0.039461   0.315483   0.125  0.90046
## totChol          0.002324   0.001127   2.062  0.03920 *
## sysBP            0.015398   0.003808   4.043 5.27e-05 ***
## diaBP           -0.004132   0.006438  -0.642  0.52096
## BMI              0.006603   0.012758   0.518  0.60476
## heartRate       -0.003250   0.004211  -0.772  0.44030
## glucose          0.007124   0.002234   3.189  0.00143 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3120.5  on 3655  degrees of freedom
## Residual deviance: 2754.2  on 3640  degrees of freedom
##   (582 observations deleted due to missingness)
## AIC: 2786.2
##
## Number of Fisher Scoring iterations: 5
```

Significant Variables:

variables: significance: male (0.001) ***age (0.001)*** cigsPerDay (0.01) **prevalentHyp (0.1) . totChol (0.05)** ***sysBP (0.001)*** glucose (0.01) ** ## Logistic Equation from Model With confidence: a = 0.05 TenYearCHD = [1 + exp(8.322206231 - 0.555097538(male) - 0.063453347(age) - 0.017929305(cigsPerDay) - 0.002324(totChol) - 0.015398(sysBP) - 0.007124(glucose))]^-1

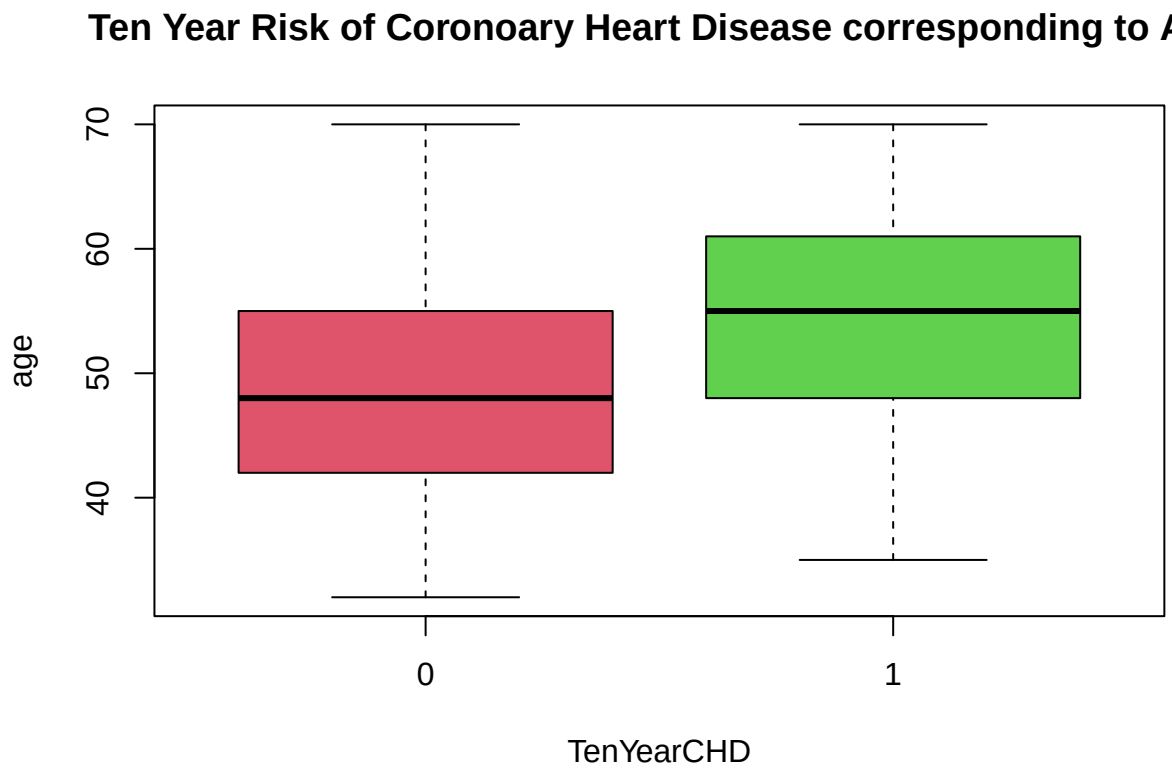
3) Data Distribution and Hypothesis Testing

```
# Binary vs Binary: boxplot(male ~ TenYearCHD)
# USE DIFFERENT PLOT

# Binary vs Binary: boxplot(prevalentStroke ~ TenYearCHD)
# USE DIFFERENT PLOT
```

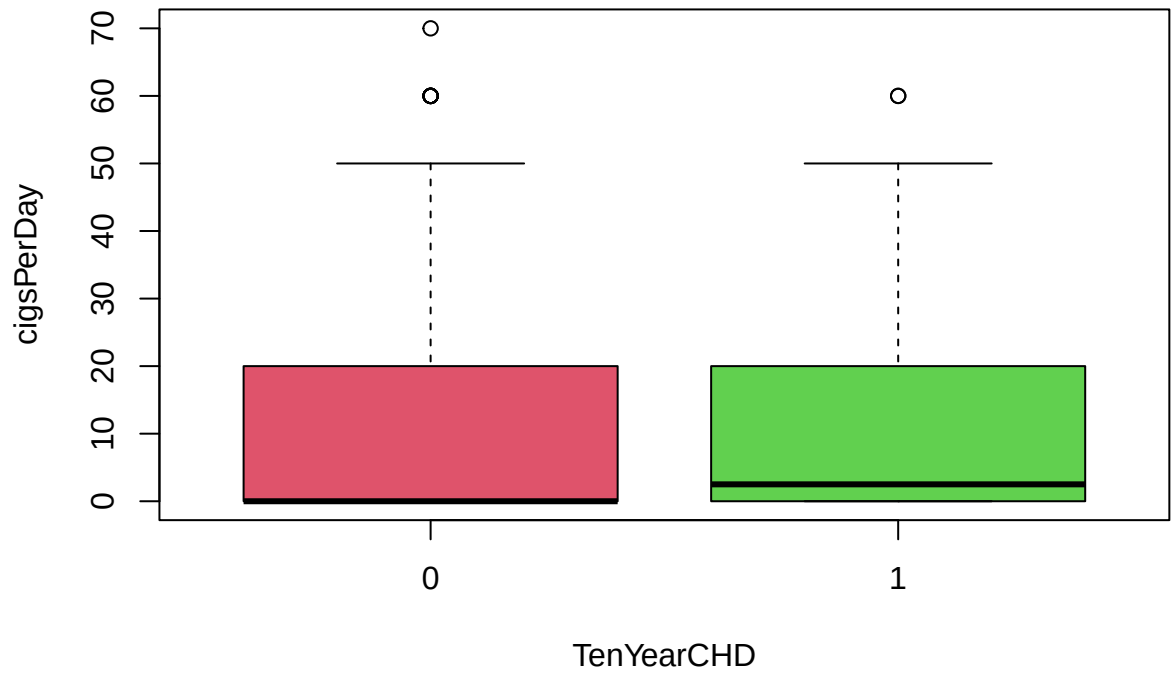
Ten Year Risk of Coronoary Heart Disease corresponding to Age

```
boxplot(age ~ TenYearCHD,
        main="Ten Year Risk of Coronoary Heart Disease corresponding to Age",
        col=c(2,3))
```



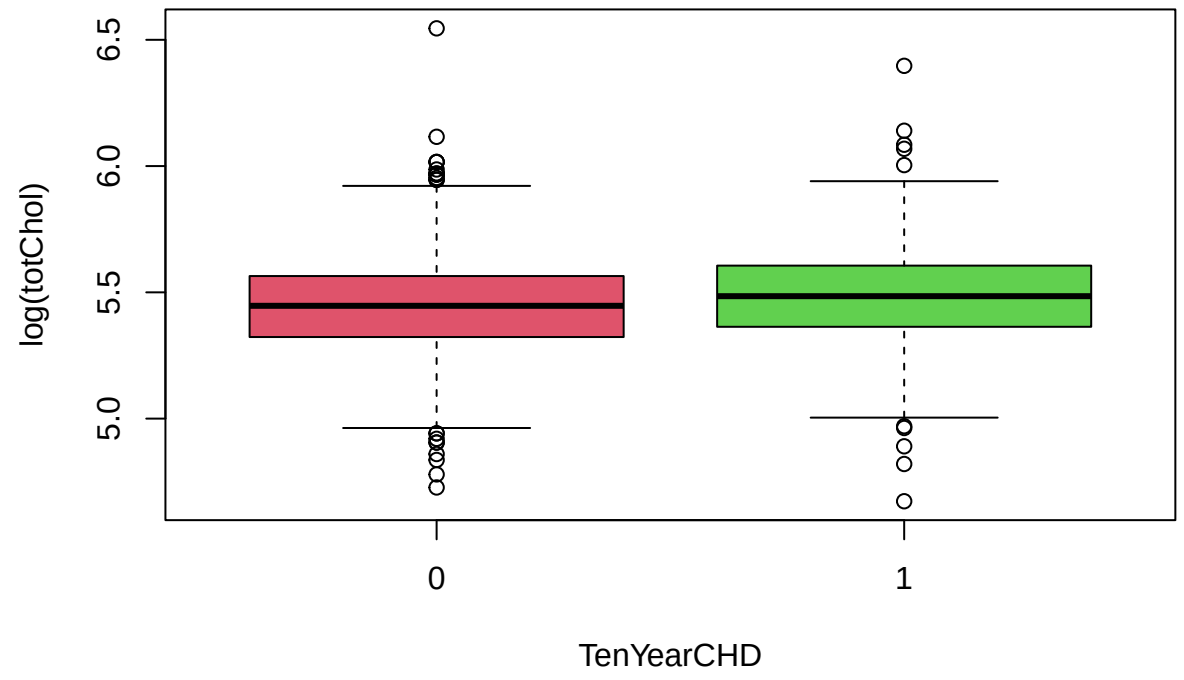

```
# PERFORM TEST

# Ten Year Risk of Coronoary Heart Disease corresponding to Cigarettes per Day
boxplot(cigsPerDay ~ TenYearCHD,
        main="Ten Year Risk of Coronoary Heart Disease corresponding to Cigarettes per Day",
        col=c(2,3))
```



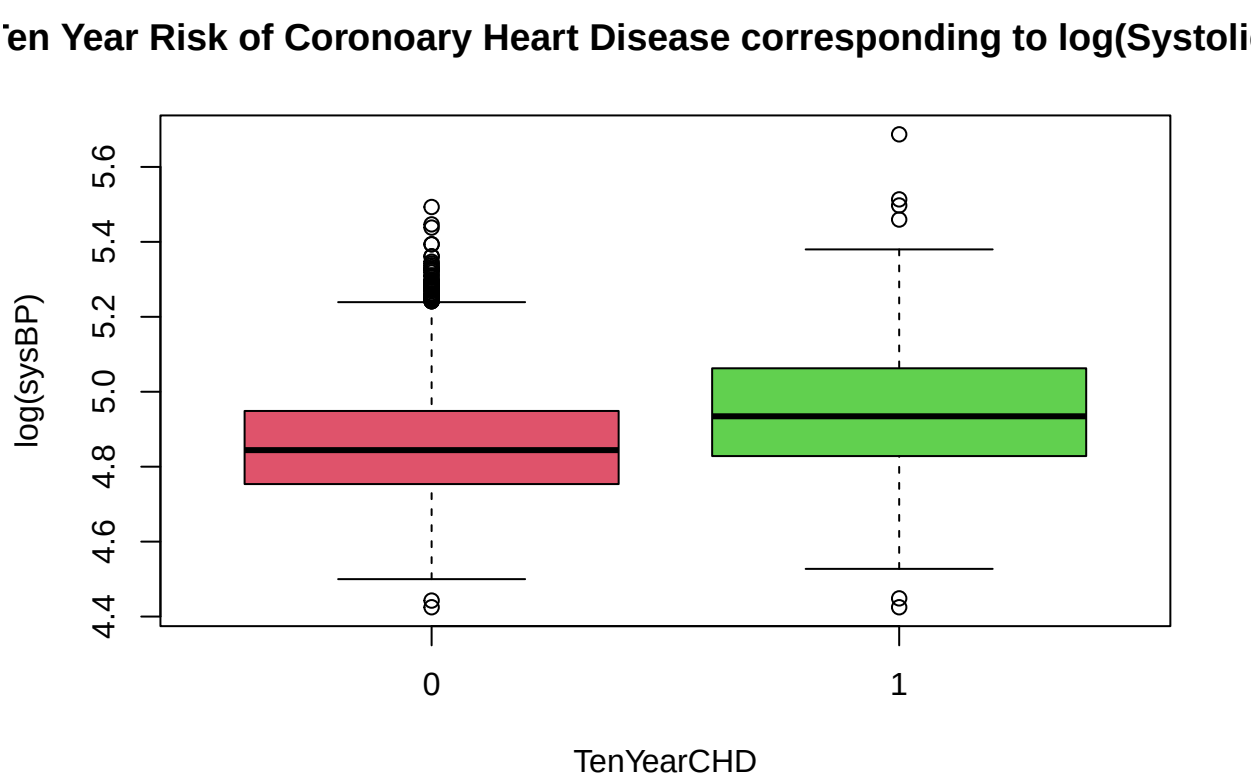

```
# PERFORM TEST

# Ten Year Risk of Coronoary Heart Disease corresponding to Total Cholesterol
boxplot(log(totChol) ~ TenYearCHD,
        main="Ten Year Risk of Coronoary Heart Disease corresponding to Total Cholesterol",
        col=c(2,3))
```



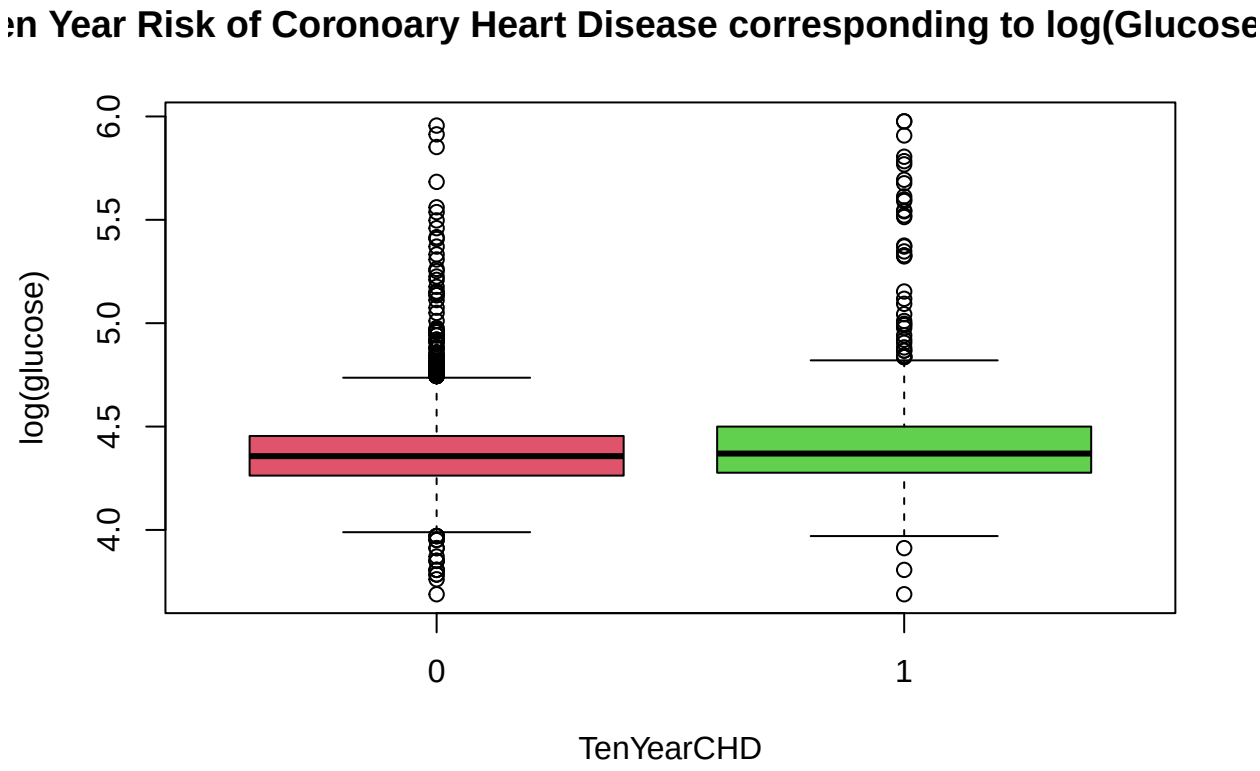

```
# PERFORM TEST

# Ten Year Risk of Coronoary Heart Disease corresponding to Systolic BP
boxplot(log(sysBP) ~ TenYearCHD,
        main="Ten Year Risk of Coronoary Heart Disease corresponding to log(Systolic BP)",
        col=c(2,3))
```




```
# PERFORM TEST

# Ten Year Risk of Coronoary Heart Disease corresponding to Glucose level
boxplot(log(glucose) ~ TenYearCHD,
        main="Ten Year Risk of Coronoary Heart Disease corresponding to log(Glucose level)",
        col=c(2,3))
```




```
# PERFORM TEST
```