CS 5610 Project

Aaron Carr & Daniel Reardon

4/8/2022

Cardiovascular disease is the number 1 leading cause of death in the world, accounting for 31% of all deaths (17.9 million people every year.) Four fifths of these deaths are due to sudden episodes of heart attack and stroke, and one third of these occur in patients under 70 years old. As such, constructing a model which can predict these events prior to their occurrence has wide ranging potential benefits in the medical field.

```
library(boot) library(MASS)
library(e1071) library(caTools)
```

```
## Warning: package 'caTools' was built under R version 4.1.3
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
## The following object is masked from 'package:MASS':
##
##     select
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tree)
```

```
## Warning: package 'tree' was built under R version 4.1.3
```

```
library(gbm)
```

```
## Warning: package 'gbm' was built under R version 4.1.3 ## Loaded gbm
2.1.8
```

```
library(ggplot2) library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
## The following object is masked from 'package:dplyr': ##
##     combine
```

```
library(randomForest)
```

Warning: package 'randomForest' was built under R version 4.1.3 ## randomForest

4.7-1

Type rfNews() to see new features/changes/bug fixes.

##

Attaching package: 'randomForest'

The following object is masked from 'package:gridExtra':

##

combine

The following object is masked from 'package:ggplot2':

##

margin

The following object is masked from 'package:dplyr':

##

combine

The dataset which we use was constructed using 5 different datasets from Cleveland, Hungary, Switzerland, Long Beach California, and the Stalog Heart Data from the UCI Machine Learning Repository. It contains 918 different observations of which 11 common features are abstracted. These features are Age, Sex, Type of Chest Pain (Typical Angina, Atypical Angina, Non-Anginal Pain, and Asymptomatic), Resting Blood Pressure, Cholesterol, Fasting Blood Sugar (a binary measurement (1 if > 120 ml.dl, 0 otherwise), Resting Electrocardiogram Results (Normal, Having St-T wave abnormality, or showing at least probable left ventricular hypertrophy), Maximum Heart Rate achieved, Exercise Induced angina (yes or no), Oldpeak (as a numeric value), and ST-Slope (Up, down, or flat), and finally, the absence or presence of Heart Disease. The goal of our models is to predict the value of the Heart Disease variable using the values of the other variables as consistently and accurately as possible.

```
##        Age          RestingBP       Cholesterol      FastingBS
##  Min.   :28.00   Min.   :  0.0   Min.   :  0.0   Min.   :0.0000
```

```
#Load Dataset
heart<-read.csv("heart.csv")
heart_copy=read.csv("heart.csv") #View(heart)
heart<-mutate(heart,typical_angina =as.integer(ChestPainType=="TA")) %>% mutate(atypical_angina
  =as.integer(ChestPainType=="ATA")) %>% mutate(non_angina_pain
  =as.integer(ChestPainType=="NAP")) %>% mutate(st_abnorm =as.integer(RestingECG=="ST")) %>%
  mutate(left_vent_hypertroph =as.integer(RestingECG=="LVH")) %>% mutate(ExerciseAngina
  =as.integer(ExerciseAngina=="Y")) %>% mutate(stslope_up =as.integer(ST_Slope=="Up")) %>%
  mutate(stslope_down =as.integer(ST_Slope=="Down")) %>% mutate(male
  =as.integer(Sex=="M")) %>%
  select(-c(ChestPainType, RestingECG, ST_Slope, Sex))

summary(heart)
```

```
##  1st Qu.:47.00   1st Qu.:120.0   1st Qu.:173.2   1st Qu.:0.0000
##  Median :54.00   Median :130.0   Median :223.0   Median :0.0000
##  Mean   :53.51   Mean   :132.4   Mean   :198.8   Mean   :0.2331
##  3rd Qu.:60.00   3rd Qu.:140.0   3rd Qu.:267.0   3rd Qu.:0.0000
##  Max.   :77.00   Max.   :200.0   Max.   :603.0   Max.   :1.0000
```

```
##          MaxHR          ExerciseAngina         Oldpeak          HeartDisease ##
      Min. : 60.0          Min.   :0.0000   Min.   :-2.6000   Min.   :0.0000
##    1st Qu.:120.0          1st Qu.:0.0000          1st Qu.: 0.0000          1st Qu.:0.0000
##    Median :138.0          Median :0.0000          Median : 0.6000          Median :1.0000
## Mean   :136.8  Mean   :0.4041  Mean   : 0.8874  Mean   :0.5534     ## 3rd Qu.:156.0  3rd
Qu.:1.0000  3rd Qu.: 1.5000  3rd Qu.:1.0000  ##  Max.   :202.0   Max.   :1.0000   Max.   :
6.2000   Max.   :1.0000
## typical_angina atypical_angina non_angina_pain   st_abnorm
## Min.   :0.00000  Min.   :0.0000  Min.   :0.0000  Min.   :0.0000     ##  1st Qu.:0.00000   1st
Qu.:0.0000   1st Qu.:0.0000   1stQu.:0.0000
##  Median :0.00000   Median :0.0000   Median :0.0000   Median :0.0000
## Mean   :0.05011  Mean   :0.1885  Mean   :0.2211  Mean   :0.1939     ## 3rd Qu.:0.00000
3rd Qu.:0.0000  3rd Qu.:0.0000  3rd Qu.:0.0000  ##  Max.   :1.00000   Max.   :1.0000   Max.
:1.0000   Max.   :1.0000
##  left_vent_hypertroph   stslope_up     stslope_down         male
##  Min.   :0.0000    Min.   :0.0000   Min.   :0.00000   Min.   :0.0000 ##  1st Qu.:0.0000       1st
Qu.:0.0000   1st Qu.:0.00000   1stQu.:1.0000
##  Median :0.0000      Median :0.0000   Median :0.00000   Median :1.0000
##  Mean   :0.2048    Mean   :0.4303   Mean   :0.06863   Mean   :0.7898 ##  3rd Qu.:0.0000    3rd
Qu.:1.0000    3rd Qu.:0.00000    3rd Qu.:1.0000 ##  Max.   :1.0000        Max.   :1.0000    Max.
:1.00000   Max.   :1.0000
```

```
#There appear to be some gaps in the data, where Cholesterol and RestingBP are labeled aszer #We decided to        o, which i
construct small linear models to predict their values and replace them

lm.chol_fit<-lm(Cholesterol~.-HeartDisease,data =dplyr::filter(heart, Cholesterol!=0)) lm.bp_fit<-
lm(RestingBP~.-HeartDisease,data =dplyr::filter(heart, RestingBP!=0))

PredictedCholesterol<-predict(lm.chol_fit, dplyr::filter(heart, Cholesterol==0)) PredictedBP<-
predict(lm.bp_fit, dplyr::filter(heart, RestingBP==0))

heart<-mutate(heart,Cholesterol =replace(Cholesterol, Cholesterol==0,PredictedCholesterol)) heart<-
mutate(heart,RestingBP =replace(RestingBP, RestingBP==0, PredictedBP))

summary(heart)
```

```
##          Age             RestingBP       Cholesterol      FastingBS
##    Min.   :28.00    Min.   : 80.0    Min.   : 85.0    Min.   :0.0000
##    1st Qu.:47.00    1st Qu.:120.0    1st Qu.:214.0    1st Qu.:0.0000
##    Median :54.00    Median :130.0    Median :240.1    Median :0.0000
##    Mean   :53.51    Mean   :132.5    Mean   :244.2    Mean   :0.2331
##    3rd Qu.:60.00    3rd Qu.:140.0    3rd Qu.:268.0    3rd Qu.:0.0000
##    Max.   :77.00    Max.   :200.0    Max.   :603.0    Max.   :1.0000
##          MaxHR          ExerciseAngina         Oldpeak          HeartDisease
##    Min.   : 60.0    Min.   :0.0000    Min.   :-2.6000    Min.   :0.0000
##    1st Qu.:120.0    1st Qu.:0.0000    1st Qu.: 0.0000    1st Qu.:0.0000
##    Median :138.0    Median :0.0000    Median :  0.6000   Median :1.0000
##    Mean   :136.8    Mean   :0.4041    Mean   :  0.8874   Mean   :0.5534
##    3rd Qu.:156.0    3rd Qu.:1.0000    3rd Qu.:  1.5000   3rd Qu.:1.0000
##    Max.   :202.0    Max.   :1.0000    Max.   :  6.2000   Max.   :1.0000
##    typical_angina     atypical_angina    non_angina_pain     st_abnorm
##    Min.   :0.00000    Min.   :0.0000    Min.   :0.0000    Min.   :0.0000
##    1st Qu.:0.00000    1st Qu.:0.0000    1st Qu.:0.0000    1st Qu.:0.0000
##    Median :0.00000    Median :0.0000    Median :0.0000    Median :0.0000
##    Mean   :0.05011    Mean   :0.1885    Mean   :0.2211    Mean   :0.1939
```

```
## 3rd Qu.:0.00000  3rd Qu.:0.0000  3rd Qu.:0.0000  3rd Qu.:0.0000  ## Max. :1.00000
Max. :1.0000 Max. :1.0000 Max.  :1.0000
##    left_vent_hypertroph      stslope_up       stslope_down          male
##    Min.    :0.0000      Min.    :0.0000    Min.    :0.00000    Min.    :0.0000
##    1st Qu.:0.0000      1st Qu.:0.0000    1st Qu.:0.00000    1st Qu.:1.0000
##    Median :0.0000      Median :0.0000    Median :0.00000    Median :1.0000
##    Mean    :0.2048      Mean    :0.4303    Mean    :0.06863    Mean    :0.7898
##    3rd Qu.:0.0000      3rd Qu.:1.0000    3rd Qu.:0.00000    3rd Qu.:1.0000
##    Max.    :1.0000      Max.    :1.0000    Max.    :1.00000    Max.    :1.0000
#Give it a look over
#View(heart) names(heart)


##  [1] "Age"                "RestingBP"          "Cholesterol"
##  [4] "FastingBS"          "MaxHR"              "ExerciseAngina"  ##
[7] "Oldpeak"                "HeartDisease"       "typical_angina"  ##  [10]
"atypical_angina" "non_angina_pain"  "st_abnorm"
## [13] "left_vent_hypertroph" "stslope_up"                "stslope_down"
## [16] "male"
dim(heart)

## [1] 918      16
summary(heart)

##          Age            RestingBP       Cholesterol      FastingBS
##    Min.    :28.00    Min.    : 80.0    Min.    : 85.0    Min.    :0.0000
##    1st Qu.:47.00    1st Qu.:120.0    1st Qu.:214.0    1st Qu.:0.0000
##    Median :54.00    Median :130.0    Median :240.1    Median :0.0000
##    Mean    :53.51    Mean    :132.5    Mean    :244.2    Mean    :0.2331
##    3rd Qu.:60.00    3rd Qu.:140.0    3rd Qu.:268.0    3rd Qu.:0.0000
##    Max.    :77.00    Max.    :200.0    Max.    :603.0    Max.    :1.0000
##        MaxHR        ExerciseAngina      Oldpeak        HeartDisease
##    Min.    : 60.0    Min.    :0.0000    Min.    :-2.6000    Min.    :0.0000
##    1st Qu.:120.0    1st Qu.:0.0000    1st Qu.: 0.0000    1st Qu.:0.0000
##    Median :138.0    Median :0.0000    Median :  0.6000    Median :1.0000
##    Mean    :136.8    Mean    :0.4041    Mean    :  0.8874    Mean    :0.5534
##    3rd Qu.:156.0    3rd Qu.:1.0000    3rd Qu.:  1.5000    3rd Qu.:1.0000
##    Max.    :202.0    Max.    :1.0000    Max.    :  6.2000    Max.    :1.0000
## typical_angina atypical_angina non_angina_pain   st_abnorm
## Min.  :0.00000  Min.  :0.0000  Min.  :0.0000  Min.  :0.0000     ##  1st Qu.:0.00000   1st
Qu.:0.0000   1st Qu.:0.0000   1stQu.:0.0000
##  Median :0.00000   Median :0.0000   Median :0.0000   Median :0.0000
## Mean  :0.05011  Mean  :0.1885  Mean  :0.2211  Mean  :0.1939     ## 3rd Qu.:0.00000
3rd Qu.:0.0000  3rd Qu.:0.0000  3rd Qu.:0.0000  ##  Max.  :1.00000   Max.  :1.0000   Max.
:1.0000   Max.  :1.0000
##  left_vent_hypertroph   stslope_up     stslope_down         male
##  Min.  :0.0000    Min.  :0.0000    Min.  :0.00000    Min.  :0.0000 ##  1st Qu.:0.0000       1st
Qu.:0.0000   1st Qu.:0.00000   1stQu.:1.0000
##  Median :0.0000      Median :0.0000   Median :0.00000   Median :1.0000
##  Mean  :0.2048    Mean  :0.4303   Mean  :0.06863   Mean  :0.7898 ##  3rd Qu.:0.0000    3rd
Qu.:1.0000    3rd Qu.:0.00000    3rd Qu.:1.0000 ##  Max.  :1.0000        Max.  :1.0000    Max.
:1.00000   Max.  :1.0000
```

```
#Check to see if there's any missing values
any(is.na(heart))
```

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

```
set.seed(97) spl=sample.split(heart$HeartDisease,SplitRatio =0.75)

heartTrain=subset(heart, spl==TRUE)
heartTest=subset(heart, spl==FALSE)

dim(heartTrain)
```

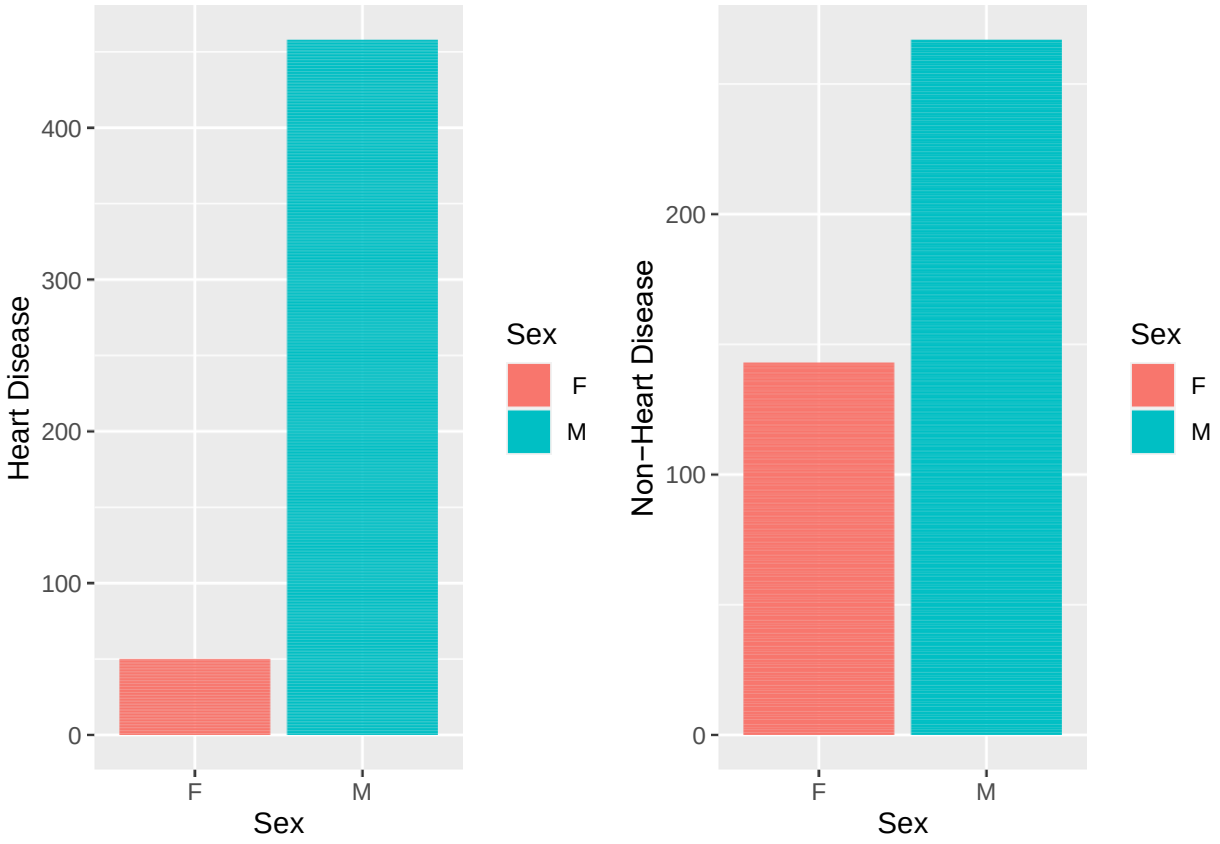
```
## [1] 689  16
```

```
dim(heartTest)
```

```
## [1] 229  16
```

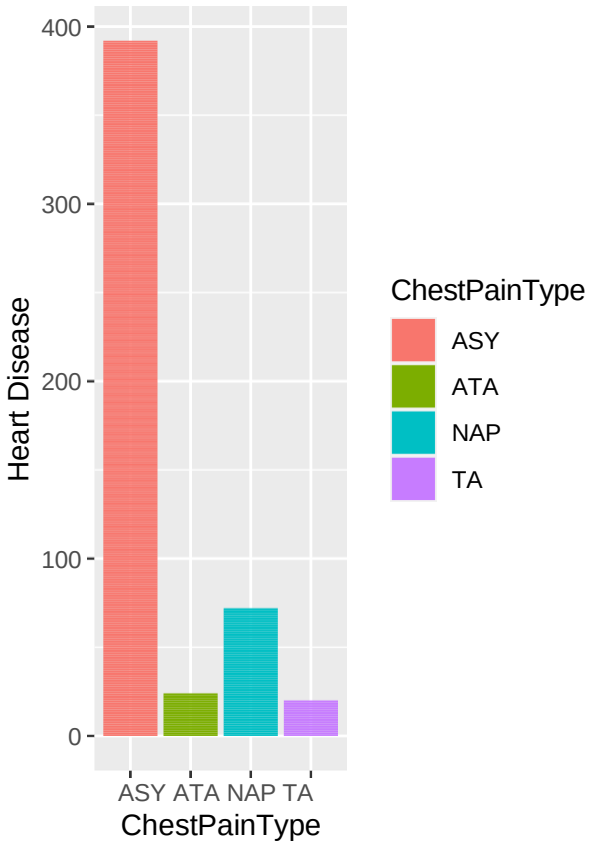
Our first task involves wrangling the dataset so that it optimal for use by the models. Glancing over the dataset, it is clear that the some of the variables are categorical and consist of three or more categories; in situations such as these, it is vital to split these categories into binary categories. For the sake of clarity, this is explicitly performed by the dplyr pipeline above and not left to the models themselves. The chest pain type, resting ECG type, exercised induced angina and patient sex have all been reformatted such that the baseline patient is a female with no resting chest pain, a standard ECG, and no exercised induced chest pain. Next, when looking over a summary of the original dataset, it becomes apparent that while there are technically no missing values, some Cholesterol values (and a single Resting Blood Pressure value) are listed as 0, which is clearly erroneous. In order to best address this without skewing the data or removing a large chunk of training data, toy linear regression models were constructed to predict the appropriate values for these variables. In situations where the data appeared to be erroneous, the erroneous value was replaced with the predicted value.

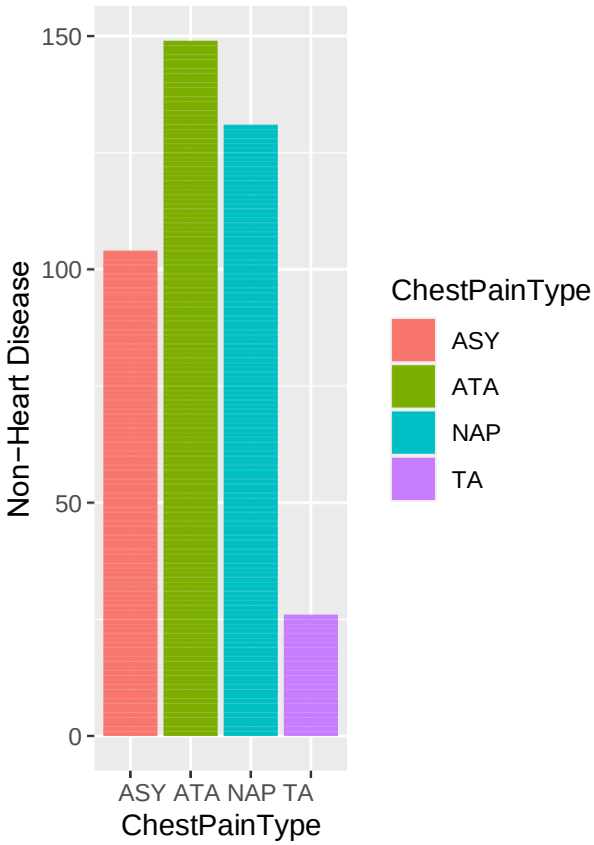
```
#Heart Disease by sex
one=ggplot(data =heart_copy)+
  geom_col(mapping =aes(x =Sex,y =HeartDisease,fill =Sex))+labs(y ="Heart Disease") zero=ggplot(data
=heart_copy)+
  geom_col(mapping =aes(x =Sex,y =1-HeartDisease,fill =Sex))+labs(y ="Non-Heart Disease")

grid.arrange(one,zero,ncol=2)
```



This dataset seems to suggest that Heart Disease is more common among men than women.

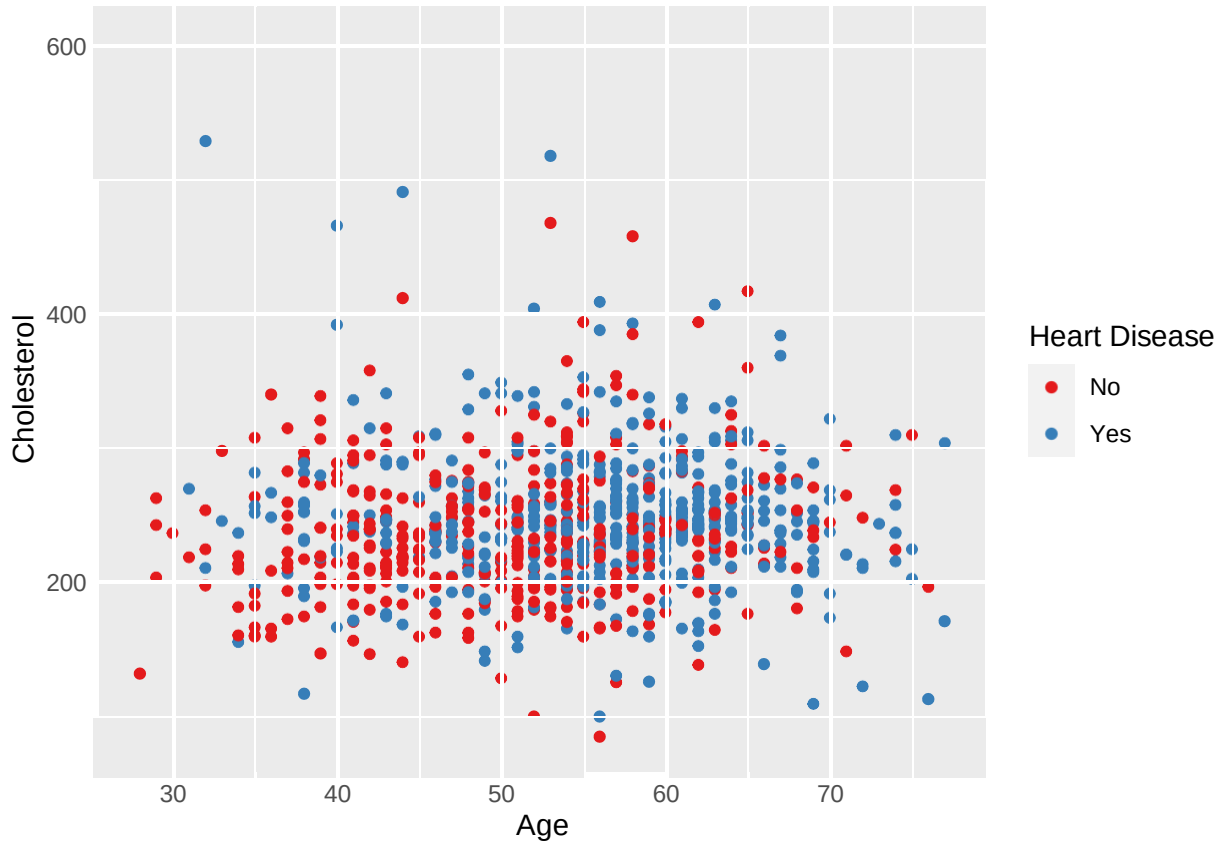
```
#Heart disease by chest pain type
two=ggplot(data =heart_copy)+
  geom_col(mapping =aes(x =ChestPainType,y =HeartDisease,fill =ChestPainType))+labs(y ="HeartDisease") three=ggplot(data
=heart_copy)+
  geom_col(mapping =aes(x =ChestPainType,y =1-HeartDisease,fill =ChestPainType))+labs(y ="Non-HeartDise

grid.arrange(two,three,ncol=2)
```





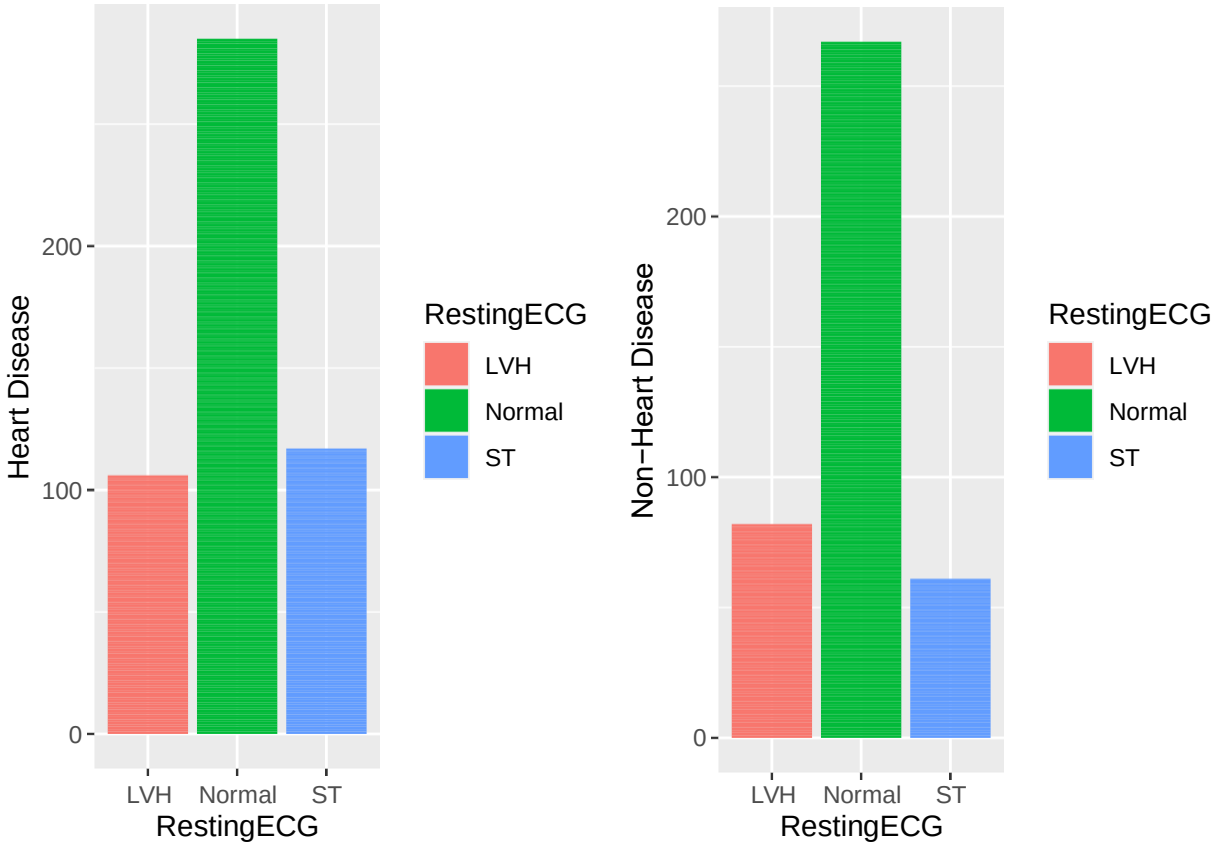
Asymptomatic individuals seem to dominate the heart disease group. There is a condition known as silent ischemia which restricts blood flow to the heart while the person feels no pain.

```
#Heart disease by age and cholesterol
ggplot(data =heart)+
  geom_point(mapping =aes(y =Cholesterol,x =Age,color =factor(HeartDisease,labels=c("No","Yes"))))+
        scale_color_brewer(palette ="Set1")+
        labs(color ="Heart Disease")
```



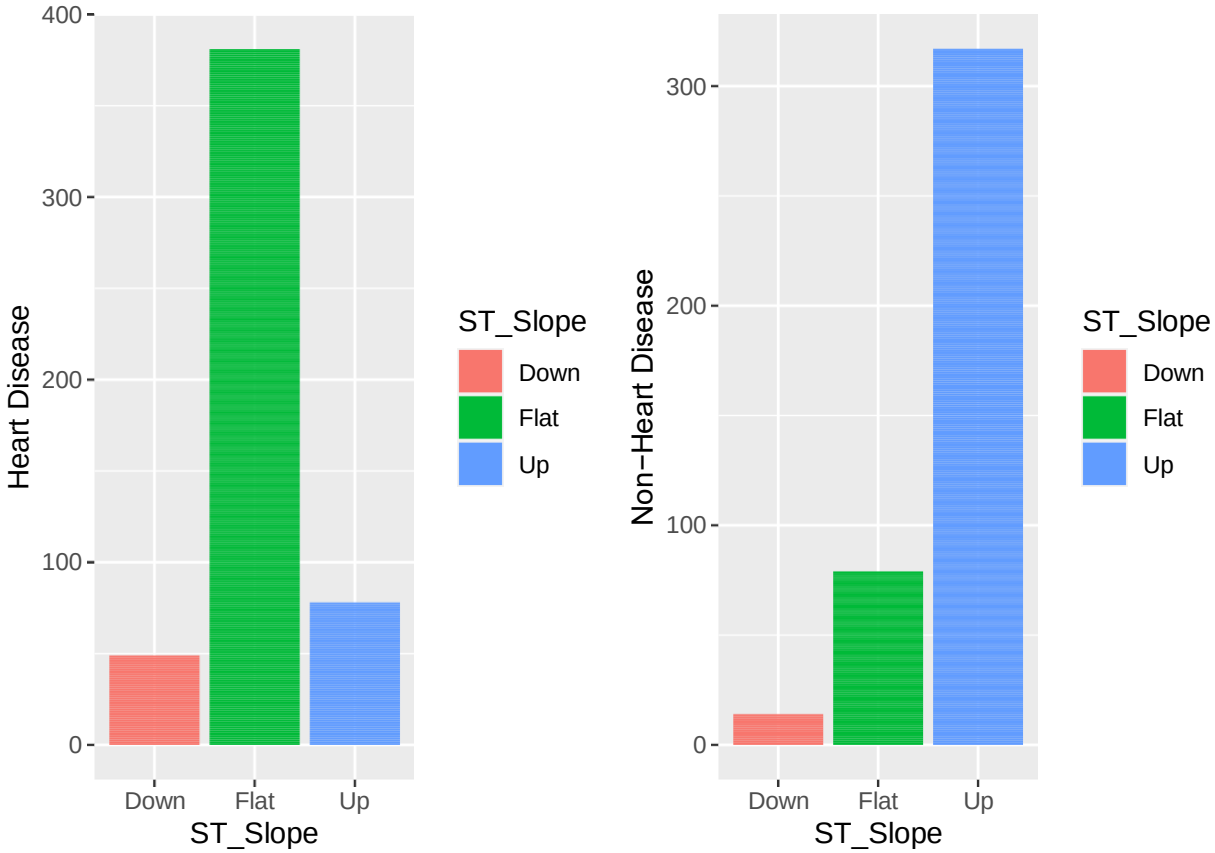
It is difficult to establish a relationship of age, cholesterol, and heart disease based on this scatterplot. It does appear that there that heart disease risk increases with age.

```
#Heart disease by Resting ECG
four=ggplot(data =heart_copy)+
  geom_col(mapping =aes(x =RestingECG,y =HeartDisease,fill =RestingECG))+labs(y ="HeartDisease") five=ggplot(data
=heart_copy)+
  geom_col(mapping =aes(x =RestingECG,y =1-HeartDisease,fill =RestingECG))+labs(y ="Non-Heart Disease")

grid.arrange(four,five,ncol=2)
```



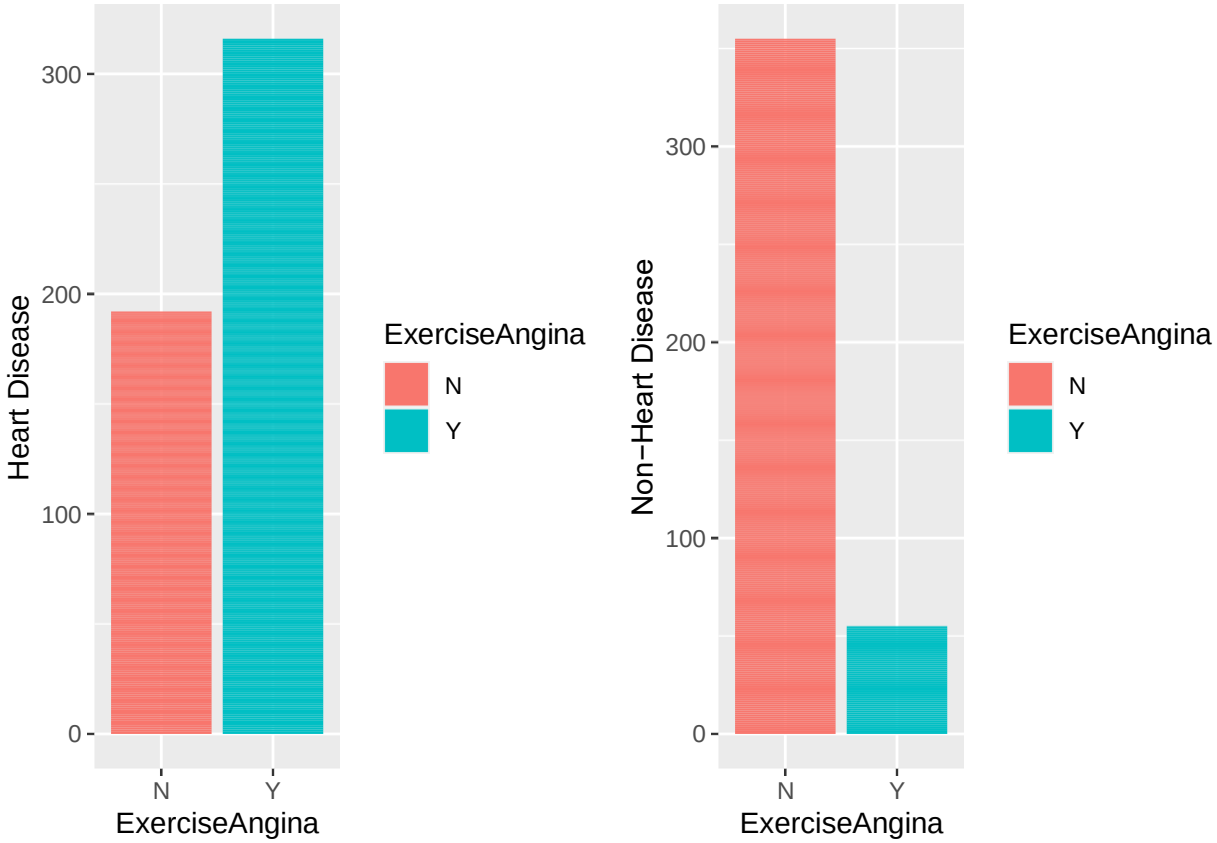
RestingECG seems to have no affect on heart disease risk.

```
#Heart Disease by ST_slope
six=ggplot(data =heart_copy)+
  geom_col(mapping =aes(x =ST_Slope,y =HeartDisease,fill =ST_Slope))+labs(y ="HeartDisease") seven=ggplot(data
=heart_copy)+
  geom_col(mapping =aes(x =ST_Slope,y =1-HeartDisease,fill =ST_Slope))+labs(y ="Non-Heart Disease")

grid.arrange(six,seven,ncol=2)
```



A flat ST_slope seems to be a risk factor for heart disease.

```
eight=ggplot(data =heart_copy)+
  geom_col(mapping =aes(x =ExerciseAngina,y =HeartDisease,fill =ExerciseAngina))+labs(y ="HeartDisease" nine=ggplot(data
=heart_copy)+
  geom_col(mapping =aes(x =ExerciseAngina,y =1-HeartDisease,fill =ExerciseAngina))+labs(y ="Non-HeartDi

grid.arrange(eight,nine,ncol=2)
```



The presence of exercise angina appears to be an indicator of heart disease.

```
#Logistic Regression (87 )        %
glm.fits<-glm(HeartDisease~., family
   =binomial,data =heartTrain
)
summary(glm.fits)
```

```
##
## Call:
## glm(formula = HeartDisease ~ ., family = binomial, data = heartTrain) ##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max ## -2.8350  -0.4045
0.1933  0.5045  2.5901  ##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
##  (Intercept)                 -0.788478   1.611022  -0.489  0.624539
##  Age                          0.011393   0.014997   0.760  0.447443
##  RestingBP                    0.004111   0.007047   0.583  0.559639
##  Cholesterol                  0.002555   0.002384   1.071  0.283995
##  FastingBS                    1.065646   0.302524   3.523  0.000427 ***
##  MaxHR                       -0.008234   0.005573  -1.477  0.139546
##  ExerciseAngina               0.938880   0.274380   3.422  0.000622 ***
##  Oldpeak                      0.357433   0.130584   2.737  0.006197 **
##  typical_angina              -1.192704   0.461671  -2.583  0.009782 **
```

```
## atypical_angina          -1.824644   0.365610  -4.991 6.02e-07 ***
## non_angina_pain          -1.611619   0.297733  -5.413 6.20e-08 ***
## st_abnorm                 0.267192   0.334449   0.799 0.424346
## left_vent_hypertroph      0.221632   0.305032   0.727 0.467479
## stslope_up               -2.302289   0.273953  -8.404  < 2e-16 ***
## stslope_down             -1.346388   0.498025  -2.703 0.006862 **
## male                      1.654871   0.340736   4.857 1.19e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 ##
## (Dispersion parameter for binomial family taken to be 1) ##
##     Null deviance: 947.41  on 688  degrees of freedom ## Residual
deviance: 463.91 on 673 degrees of freedom ## AIC: 495.91
##
## Number of Fisher Scoring iterations: 5
```

```
glm.probs<-predict(glm.fits,type ="response") glm.pred<-
rep(0,689)
glm.pred[glm.probs>.5]=1

table(glm.pred, heartTrain$HeartDisease)
```

```
##
##  glm.pred    0   1
##         0 253  42
##         1  55 339
```

```
mean(glm.pred==heartTrain$HeartDisease)
```

```
## [1] 0.8592163
```

```
glm.probs<-predict(glm.fits,type ="response",newdata =heartTest) glm.predTest<-
rep(0,229)
glm.predTest[glm.probs>.5]=1

table(glm.predTest, heartTest$HeartDisease)
```

```
##
##  glm.predTest   0   1
##             0  86  13
##             1  16 114
```

```
mean(glm.predTest==heartTest$HeartDisease)
```

```
## [1] 0.8733624
```

```
##Feature Selection algorithm
i<-glm(HeartDisease~1,
                    family =binomial,data =heartTrain)

glm.new<-step(i,direction=          'both',scope=formula(glm.fits),trace=0)

summary(glm.new)
##
## Call:
```

```
## glm(formula = HeartDisease ~ stslope_up + ExerciseAngina + male +
##     non_angina_pain + atypical_angina + FastingBS + typical_angina +
##     Oldpeak + stslope_down + MaxHR, family = binomial, data =heartTrain) ##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8268  -0.4138   0.1946   0.5056   2.7572
##
## Coefficients:
##                   Estimate   Std. Error  z value  Pr(>|z|)
## (Intercept)       1.376374   0.797648    1.726    0.084430 .
## stslope_up        -2.316734  0.270974    -8.550   < 2e-16  ***
## ExerciseAngina    0.988086   0.270632    3.651    0.000261 ***
## male              1.557620   0.332344    4.687    2.78e-06 ***
## non_angina_pain   -1.620591  0.293086    -5.529   3.21e-08 ***
## atypical_angina   -1.824558  0.362917    -5.027   4.97e-07 ***
## FastingBS         1.130285   0.294911    3.833    0.000127 ***
## typical_angina    -1.152981  0.450237    -2.561   0.010442 *
## Oldpeak           0.393905   0.127162    3.098    0.001950 **
## stslope_down      -1.406604  0.493726    -2.849   0.004386 **
## MaxHR             -0.010062  0.005077    -1.982   0.047503 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 ##
## (Dispersion parameter for binomial family taken to be 1) ##
##    Null deviance: 947.41  on 688  degrees of freedom ## Residual
deviance: 467.97 on 678 degrees of freedom ## AIC: 489.97
##
## Number of Fisher Scoring iterations: 5
```

```
glm.probs<-predict(glm.new,type ="response",newdata =heartTest)
glm.predTest<-rep(0,229)
glm.predTest[glm.probs>.5]=1

table(glm.predTest, heartTest$HeartDisease)
```

```
##
##  glm.predTest   0   1
##             0  86  11
##             1  16 116
```

```
mean(glm.predTest==heartTest$HeartDisease)
```

```
## [1] 0.8820961
```

The Logistic function models the probability of the outcome being a “success” (in this case success means that a patient has heart disease). The estimates return the log odds of a success given a value. For example, the estimate for the sex of the patient is 1.642. Meaning that a male has eˆ1.642 higher odds (or about 5 times higher odds) of having heart disease than a female.

For a continuous variable like cholesterol, the interpretation is similar except it relates to an increase of one unit. For example if a patient has one cholesterol unit higher than another patient, that patient with have an increase in odds of eˆ0.003597 (or about 1.0036 times higher odds).

The algorithm used to select the features used in the model was the stepwise regression algorithm. The

algorithm begins with nothing in the model but the intercept, and adds the most significant variable. It continues this step until there are no more significant variables. However, at each step it also looks to see that no variables have been made insignificant by the addition of new variables. While this method is not foolproof, it does generally give a good idea of what features to include.

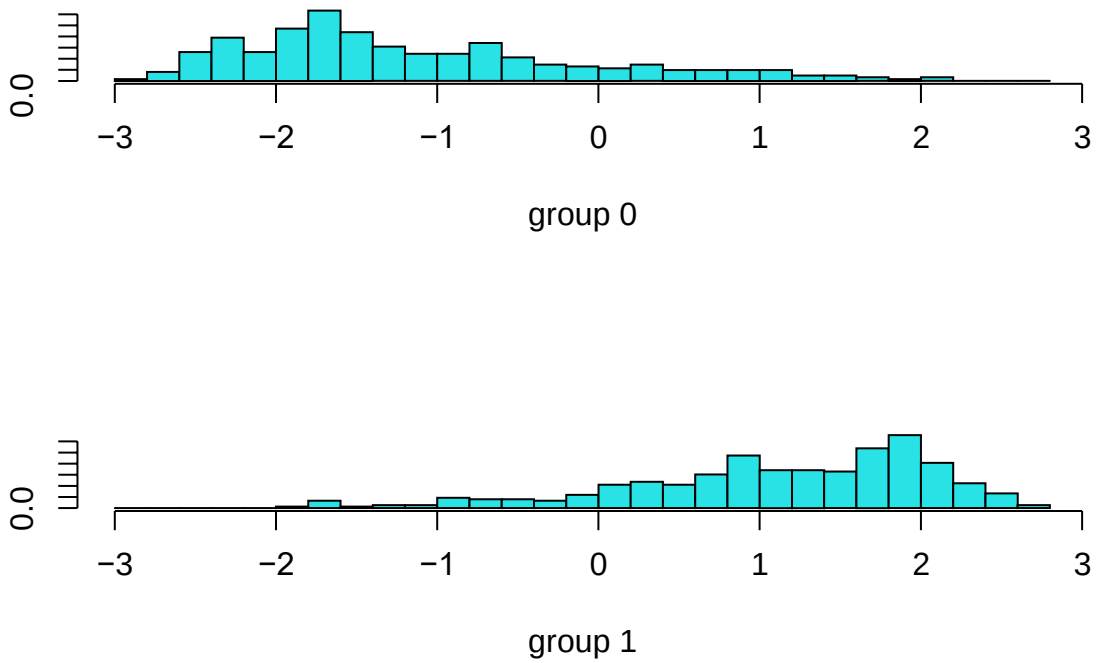
The model seems to agree that the variables we identified as significant in the visualization stage should be included in the model. It also agrees with not including RestingECG in the model as that seemed to have no effect when looking at the bar chart.

The feature selected model performs almost identically to the model which includes all features, each having an accuracy score of around 88%. (In fact, on the test data, the feature selection model performs slightly better).

```
#Linear Discriminant Analysis (87 )          %
lda.fit<-lda(HeartDisease~.,data =heartTrain) lda.fit
```

```
##  Call:
##  lda(HeartDisease ~ ., data = heartTrain)
##
##  Prior probabilities of groups:
##                 0                 1
##  0.4470247 0.5529753
##
##  Group means:
##                Age RestingBP Cholesterol FastingBS                 MaxHR ExerciseAngina          Oldpeak
##  0 50.72727       130.3442        240.4317 0.1201299 148.0227                0.1331169 0.4123377
##  1 55.92388       135.3390        251.4317 0.3307087 128.2756                0.6246719 1.2958005
##      typical_angina atypical_angina non_angina_pain st_abnorm left_vent_hypertroph
##  0         0.06493506          0.36038961              0.3344156  0.1493506                    0.1948052
##  1         0.05249344          0.04986877              0.1548556  0.2309711                    0.2125984
##      stslope_up stslope_down                male
##  0     0.7889610      0.03246753 0.6623377
##  1     0.1653543      0.09711286 0.9160105
##
## Coefficients of linear discriminants: ##   LD1
## Age                                  0.007225944
## RestingBP                            0.001151866
## Cholesterol                          0.001079759
## FastingBS                            0.515764906
## MaxHR                               -0.004872809
## ExerciseAngina                       0.554412550
## Oldpeak                              0.164861467
## typical_angina                      -0.584059677
## atypical_angina                     -1.020089731
## non_angina_pain                     -0.903042790
## st_abnorm                            0.150132055
## left_vent_hypertroph                 0.101949823
## stslope_up                          -1.510016605
## stslope_down                        -0.595821588
## male                                 0.718177361
```

```
plot(lda.fit)
```




```
lda.pred<-predict(lda.fit,newdata =heartTest) lda.class<-
lda.pred$class
table(lda.class, heartTest$HeartDisease)
```

```
##
## lda.class          0     1
##                0   85    12
##                1   17 115
```

```
mean(lda.class==heartTest$HeartDisease)
```

```
## [1] 0.8733624
```

The linear discriminant analysis model also displays a relatively high level of accuracy, at 87%.

```
#Quadratic Discriminant Analysis (86 )          %

qda.fit<-qda(HeartDisease~.,data =heartTrain) qda.fit
```

```
## Call:
## qda(HeartDisease ~ ., data = heartTrain) ##
## Prior probabilities of groups: ##
                   0                 1
## 0.4470247 0.5529753
```

```
##
## Group means:
##              Age RestingBP Cholesterol FastingBS               MaxHR ExerciseAngina         Oldpeak
## 0 50.72727       130.3442       240.4317 0.1201299 148.0227              0.1331169 0.4123377
## 1 55.92388       135.3390       251.4317 0.3307087 128.2756              0.6246719 1.2958005
##      typical_angina atypical_angina non_angina_pain st_abnorm left_vent_hypertroph
## 0         0.06493506       0.36038961      0.3344156 0.1493506                0.1948052
## 1         0.05249344       0.04986877      0.1548556 0.2309711                0.2125984
##      stslope_up stslope_down      male
## 0    0.7889610    0.03246753 0.6623377
## 1    0.1653543    0.09711286 0.9160105
```

```
qda.class<-predict(qda.fit,newdata =heartTest)$class table(qda.class,
heartTest$HeartDisease)
```

```
##
## qda.class        0     1
##             0    83    16
##             1    19   111
```

```
mean(qda.class==heartTest$HeartDisease)
```

```
## [1] 0.8471616
```

The quandradic discriminant model has an accuracy of only around 85%, close to the other models but slightly weaker. This might be an indication that the boundary between categories can be representedlinearly.

```
#Naive Bayes (87 )   %
nb.fit<-naiveBayes(HeartDisease~.,data =heartTrain) nb.fit
```

```
##
## Naive Bayes Classifier for DiscretePredictors ##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace) ##
## A-priori probabilities: ## Y
##         0         1
## 0.4470247 0.5529753 ##
## Conditional probabilities: ##     Age

## Y          [,1]     [,2]
##      0 50.72727 9.347628
##      1 55.92388 8.595102
##
##      RestingBP
## Y          [,1]     [,2]
##      0 130.3442 16.68305
##      1 135.3390 19.15541
##
##      Cholesterol
## Y          [,1]     [,2]
##      0 240.4317 53.19314
##      1 251.4317 54.68725
```

```
##
##       FastingBS
##  Y             [,1]          [,2]
##       0 0.1201299  0.3256424
##       1 0.3307087  0.4710870
##
##       MaxHR
##  Y          [,1]          [,2]
##     0 148.0227     22.91983
##     1 128.2756 23.67989
##
##       ExerciseAngina
##  Y             [,1]          [,2]
##       0 0.1331169  0.3402538
##       1 0.6246719  0.4848442
##
##       Oldpeak
##  Y           [,1]          [,2]
##       0 0.4123377  0.720055
##       1 1.2958005  1.175779
##
##       typical_angina
##  Y             [,1]              [,2]
##     0 0.06493506       0.2468122
##     1 0.05249344       0.2233132
##
##       atypical_angina
##  Y             [,1]             [,2]
##     0 0.36038961 0.4808948
##     1 0.04986877 0.2179600
##
##       non_angina_pain
##  Y             [,1]          [,2]
##       0 0.3344156  0.4725535
##       1 0.1548556  0.3622427
##
##       st_abnorm
##  Y          [,1]          [,2]
##       0 0.1493506  0.3570138
##       1 0.2309711  0.4220082
##
##       left_vent_hypertroph
##  Y             [,1]          [,2]
##       0 0.1948052  0.3966952
##       1 0.2125984  0.4096839
##
##          stslope_up
##  Y             [,1]          [,2]
##       0 0.7889610  0.4087100
##       1 0.1653543  0.3719885
##
##       stslope_down
##  Y             [,1]              [,2]
##     0 0.03246753       0.1775267
```

```
##      1 0.09711286 0.2965008 ##
##        male
## Y                [,1]              [,2]
##      0 0.6623377 0.4736824
##      1 0.9160105 0.2777368
```

```
nb.class<-predict(nb.fit,newdata =heartTest) table(nb.class,
heartTest$HeartDisease)
```

```
##
## nb.class          0     1
##              0    84    18
##              1    18 109
```

```
mean(nb.class==heartTest$HeartDisease)
```

```
## [1] 0.8427948
```

The Naive Bayes model performs about as well as the quadradic discriminant model, at around 84%.

```
#Classification Tree 85%
tree.heart<-tree(as.factor(HeartDisease)~.,data =heartTrain) summary(tree.heart)
```

```
##
## Classification tree:
## tree(formula = as.factor(HeartDisease) ~ ., data = heartTrain) ## Variables actually
used in tree construction:
##    [1] "stslope_up"                "MaxHR"                     "Oldpeak"
##    [4] "ExerciseAngina"            "male"                      "Cholesterol"
##    [7] "left_vent_hypertroph" "atypical_angina"                "non_angina_pain"
## [10] "RestingBP"
## Number of terminal nodes:           17
## Residual mean deviance:          0.6075 = 408.2 / 672
## Misclassification error rate: 0.1277 = 88 / 689
```

```
tree:::print.tree(tree.heart)
```

```
## node), split, n, deviance, yval,(yprob) ##           *
denotes terminal node
##
##    1) root 689 947.400 1 ( 0.44702 0.55298 )
##      2) stslope_up < 0.5 383 348.900 1 ( 0.16971 0.83029 )
##        4) MaxHR < 140.5 260 145.900 1 ( 0.08077 0.91923 )
##          8) Oldpeak < 0.05 59          0.000 1 ( 0.00000 1.00000 ) *
##          9) Oldpeak > 0.05 201 134.600 1 ( 0.10448 0.89552 )
##            18) Oldpeak < 1.65 101          97.660 1 ( 0.18812 0.81188 )
##              36) ExerciseAngina < 0.5 23           31.490 1 ( 0.43478 0.56522 )
##                72) male < 0.5 5          0.000 0 ( 1.00000 0.00000 ) *
##                73) male > 0.5 18         21.270 1 ( 0.27778 0.72222 ) *
##              37) ExerciseAngina > 0.5 78           55.790 1 ( 0.11538 0.88462 ) *
##            19) Oldpeak > 1.65 100          19.610 1 ( 0.02000 0.98000 ) *
##        5) MaxHR > 140.5 123 160.400 1 ( 0.35772 0.64228 )
##          10) Cholesterol < 245.153 62           85.950 0 ( 0.50000 0.50000 )
##            20) Oldpeak < 2.4 51          69.100 0 ( 0.58824 0.41176 ) *
##            21) Oldpeak > 2.4 11           6.702 1 ( 0.09091 0.90909 ) *
```

```
##           11) Cholesterol > 245.153 61          63.200 1 ( 0.21311 0.78689 ) *
##       3) stslope_up > 0.5 306 311.200 0 ( 0.79412 0.20588 )
##         6) Oldpeak < 0.45 233 163.000 0 ( 0.88841 0.11159 )
##          12) left_vent_hypertroph < 0.5 196 105.900 0 ( 0.92347 0.07653 )
##            24) atypical_angina < 0.5 112          84.400 0 ( 0.87500 0.12500 )
##              48) non_angina_pain < 0.5 54          57.210 0 ( 0.77778 0.22222 ) *
##              49) non_angina_pain > 0.5 58          17.400 0 ( 0.96552 0.03448 ) *
##            25) atypical_angina > 0.5 84          10.850 0 ( 0.98810 0.01190 ) *
##          13) left_vent_hypertroph > 0.5 37          45.030 0 ( 0.70270 0.29730 )
##            26) non_angina_pain < 0.5 26          35.430 0 ( 0.57692 0.42308 ) *
##            27) non_angina_pain > 0.5 11          0.000 0 ( 1.00000 0.00000 ) *
##         7) Oldpeak > 0.45 73 101.200 1 ( 0.49315 0.50685 )
##          14) male < 0.5 19          7.835 0 ( 0.94737 0.05263 ) *
##          15) male > 0.5 54          68.740 1 ( 0.33333 0.66667 )
##            30) MaxHR < 124.5 13          0.000 1 ( 0.00000 1.00000 ) *
##            31) MaxHR > 124.5 41          56.230 1 ( 0.43902 0.56098 )
##              62) RestingBP < 146.5 31          37.350 1 ( 0.29032 0.70968 ) *
##              63) RestingBP > 146.5 10          6.502 0 ( 0.90000 0.10000 ) *
```

```
tree.pred<-predict(tree.heart, heartTest,type ="class") table(tree.pred,
heartTest$HeartDisease)
```

```
##
##  tree.pred   0   1
##          0  78  17
##          1  24 110
```

```
(78+110)/229
```

```
## [1] 0.8209607
```

The single decision tree model was relatively inaccurate, at only 82% accuracy.

```
#Random Forest 84 %
rf.heart<-randomForest(as.factor(HeartDisease)~.,data =heartTrain,importance =TRUE) rf.pred<-
predict(rf.heart, heartTest,type ="class")
table(rf.pred, heartTest$HeartDisease)
```

```
##
##  rf.pred   0   1
##        0  85  11
##        1  17 116
```

```
(86+114)/229
```

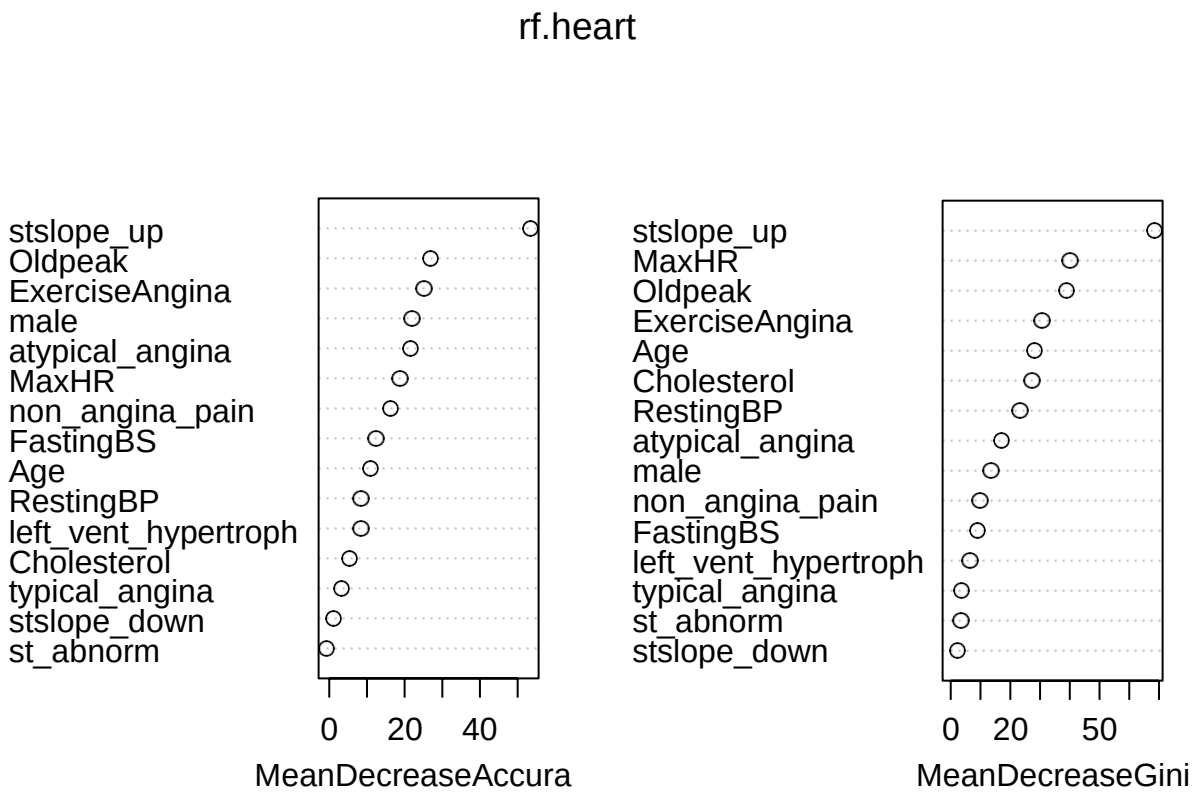
```
## [1] 0.8733624
```

```
importance(rf.heart)
```

	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
Age	13.824741	1.2336320	10.9172439	28.118585
RestingBP	4.839518	7.0998126	8.4058077	23.326737
Cholesterol	2.986400	4.6324464	5.4056095	27.306245
FastingBS	10.043454	7.9555532	12.4024993	8.934894
MaxHR	9.512849	16.8482427	18.8029124	40.103336
ExerciseAngina	21.358172	15.3484944	25.1797255	30.679202
Oldpeak	29.769385	5.3210615	26.8222659	38.824826
typical_angina	3.860771	1.2198604	3.2771607	3.704032

## atypical_angina	14.997631	15.8600948	21.6066217	17.098950
## non_angina_pain	13.883831	9.3832491	16.2090330	9.907325
## st_abnorm	2.383083	-2.9399912	-0.6736414	3.494780
## left_vent_hypertroph	6.561548	5.8332081	8.3881652	6.432005
## stslope_up	48.107553	33.9892617	53.4126063	68.461485
## stslope_down	2.685078	-0.8650802	1.1154627	2.194090
## male	14.447781	19.2995568	21.9599960	13.500637

```
varImpPlot(rf.heart)
```



But the random forest model pushed the accuracy back up to around 87%. The random forest model also seems to indicate that upward sloping ST region is overwhelmingly indicative of the absence of heart disease, in agreement with the linear model.

In summary, we have decided that the logistic model is the best model to use to solve this problem. It performs at or above the level of the other methods we tried, but it has the added benefit of being the easiest model to understand for someone who is not statistically literate.