# Data Science Capstone : Heart Attack Prediction

Omid Airom

10/27/2021

## Introduction

This project report is provided for the final project of the Data Science : Capstone course provided by Harvard University & Edx. The data has been downloaded from the Kaggle website and is the data of 1025 patients which shows either each patient with its own data has experienced heart attack or not. Worldwide, about 15.9 million myocardial infarctions occurred in 2015. Heart disease is one in all the foremost vital human diseases. The curing of all the diseases have a greater scope but when it comes to heart , the accuracy and the risk factors are in down line. People who suffer from the symptoms are aware only during the last stages of occurrences and it becomes difficult for the doctors to treat them as well. To bring out a solution that is more efficient, machine learning algorithms are used for testing the attributes of the patients and bring out their issues earlier. Prediction is used to find the future results based upon the current trends. Machine learning algorithms are used for the prediction mostly when there is a need for higher accuracy rates than the existing system and to provide best results. The traditional methods have less scope than the computer based test results. It will be useful for the medical practitioners to treat the patient with higher accuracy rate of the computer aided diagnosis produced by the machine learning algorithms and it also helps them to treat the patients at the earliest based on their severity rates. In this project i first loaded the data, then explored it and took a look at each feature of the patients data and at last built 3 models with the original data and the normalized data and reached the accuracy of 0.883.

## Exploring the Data

### Data Analysis

As mentioned before the data includes 1025 rows each with 14 feature for each of the 1025 patients. the features are as following :

Age
Gender (sex) Chest Pain Type (cp)
Resting Blood Pressure (trestbps)
Cholesterol (chol)
Fasting Blood Sugar (fbs) Resting Electrocardiographic Results (restecg)
Maximum Heart Rate Achieved (thalach)
Exercise Induced Angina (exang)
ST Depression Induced by Exercise Relative to Rest (oldpeak)
Slope of the Peak Exercise ST Segment
Number of Major Vessels (ca)
Thalassemia Blood Disorder (thal)
Presence of Heart Disease (target)

## Exploring the features

Taking a look at the first rows of the data :

| age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 52 | 1 | 0 | 125 | 212 | 0 | 1 | 168 | 0 | 1.0 | 2 | 2 | 3 | 0 |
| 53 | 1 | 0 | 140 | 203 | 1 | 0 | 155 | 1 | 3.1 | 0 | 0 | 3 | 0 |
| 70 | 1 | 0 | 145 | 174 | 0 | 1 | 125 | 1 | 2.6 | 0 | 0 | 3 | 0 |
| 61 | 1 | 0 | 148 | 203 | 0 | 1 | 161 | 0 | 0.0 | 2 | 1 | 3 | 0 |
| 62 | 0 | 0 | 138 | 294 | 1 | 1 | 106 | 0 | 1.9 | 1 | 3 | 2 | 0 |

The dimension of the data :

```
## [1] 1025   14
```

The number of positive and negative cases and the summary of the whole data :

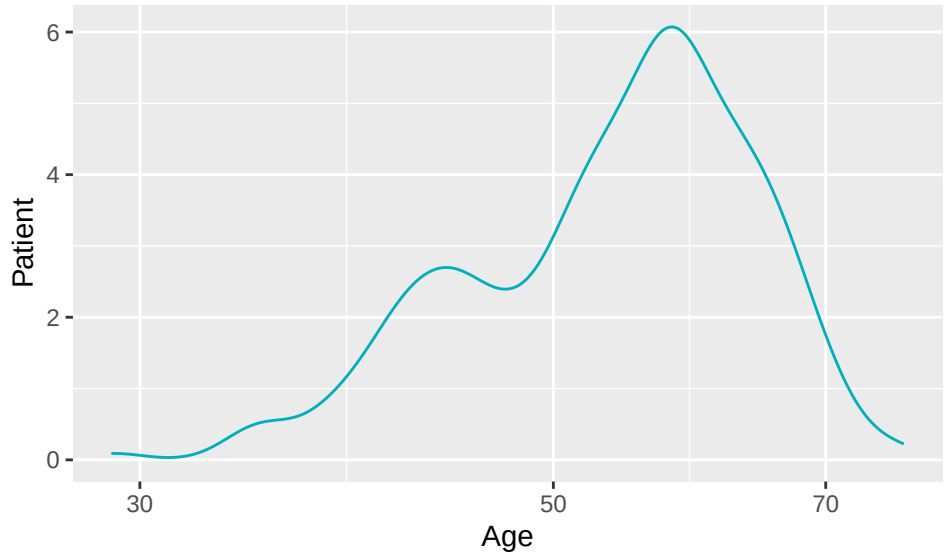
| targets | Freq |
|---|---|
| negative | 499 |
| positive | 526 |

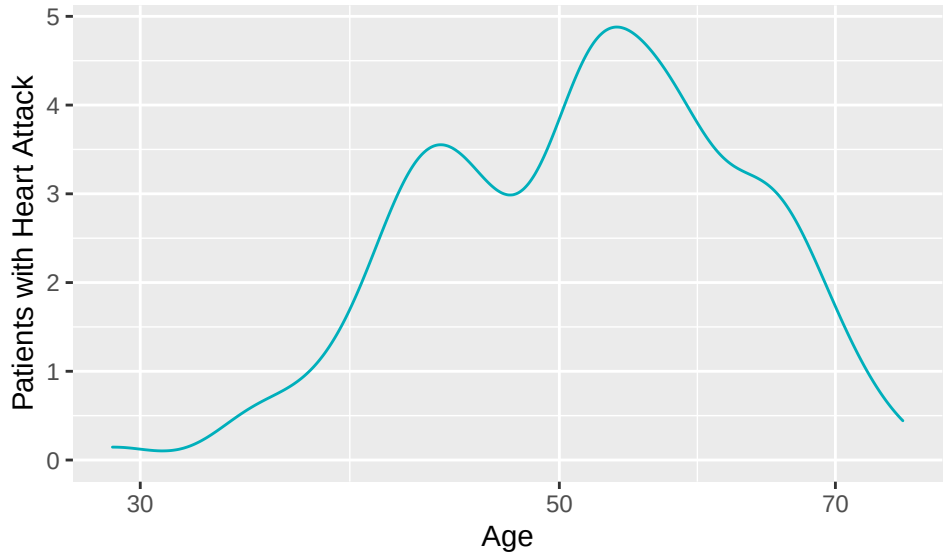
```
##       age             sex               cp            trestbps    
##  Min.   :29.00   Min.   :0.0000   Min.   :0.0000   Min.   : 94.0  
##  1st Qu.:48.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:120.0  
##  Median :56.00   Median :1.0000   Median :1.0000   Median :130.0  
##  Mean   :54.43   Mean   :0.6956   Mean   :0.9424   Mean   :131.6  
##  3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.0000   3rd Qu.:140.0  
##  Max.   :77.00   Max.   :1.0000   Max.   :3.0000   Max.   :200.0  
##       chol            fbs             restecg          thalach     
##  Min.   :126   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0  
##  1st Qu.:211   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:132.0  
##  Median :240   Median :0.0000   Median :1.0000   Median :152.0  
##  Mean   :246   Mean   :0.1493   Mean   :0.5298   Mean   :149.1  
##  3rd Qu.:275   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:166.0  
##  Max.   :564   Max.   :1.0000   Max.   :2.0000   Max.   :202.0  
##      exang           oldpeak          slope             ca        
##  Min.   :0.0000   Min.   :0.000   Min.   :0.000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:1.000   1st Qu.:0.0000  
##  Median :0.0000   Median :0.800   Median :1.000   Median :0.0000  
##  Mean   :0.3366   Mean   :1.072   Mean   :1.385   Mean   :0.7541  
##  3rd Qu.:1.0000   3rd Qu.:1.800   3rd Qu.:2.000   3rd Qu.:1.0000  
##  Max.   :1.0000   Max.   :6.200   Max.   :2.000   Max.   :4.0000  
##       thal           target      
##  Min.   :0.000   Min.   :0.0000  
##  1st Qu.:2.000   1st Qu.:0.0000  
##  Median :2.000   Median :1.0000  
##  Mean   :2.324   Mean   :0.5132  
##  3rd Qu.:3.000   3rd Qu.:1.0000  
##  Max.   :3.000   Max.   :1.0000
```

## Age

Age is the first feature in the data we have. In this data we can see that the range of the age feature is between 29 and 77 years. the median is 56 and we have a mean equal to 54.43. According to the histogram of the age feature we can see that most of the patients are between 40 and 70 and most of the patients are near to 60 years old.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   29.00   48.00   56.00   54.43   61.00   77.00
```



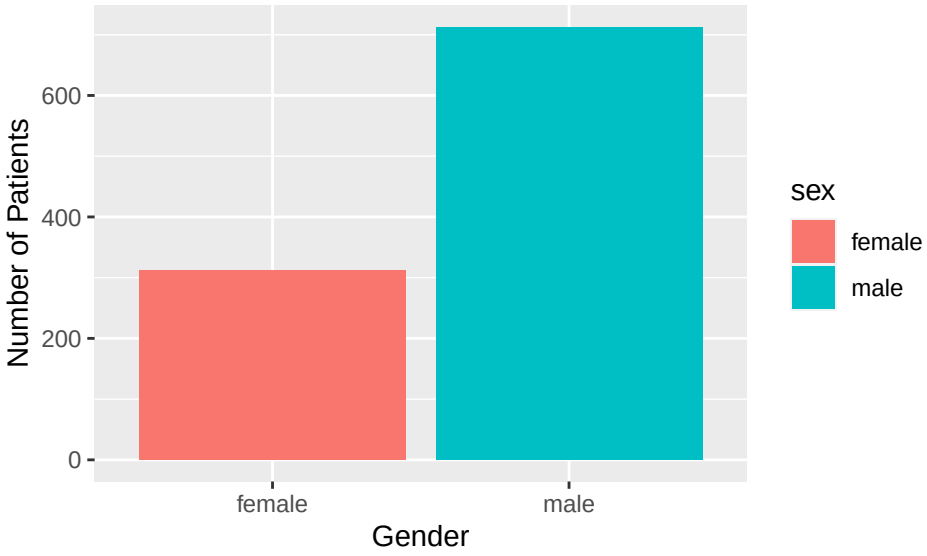




So we can see that the age of the patient effects heart attack and most of the positive cases were for the patients between 40 and 70.

## Sex

The next feature in the data is the sex column. In the original data women are shown with "0" and men are shown with "1". I changed it to female and male. We can see that there are 312 women which here are 226 cases with heart attack in them.

| Var1 | Freq |
|---|---|
| female | 312 |
| male | 713 |



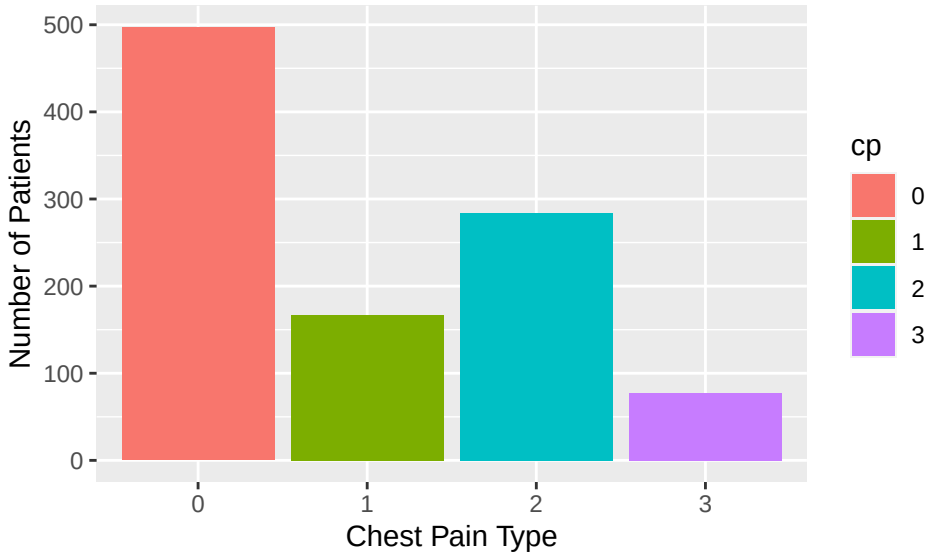

| gender | number |
|---|---|
| total women | 312.0000000 |
| positive women | 226.0000000 |
| positive women to total women ratio | 0.7243590 |
| total men | 713.0000000 |
| positive men | 300.0000000 |
| positive men to total men ratio | 0.4207574 |

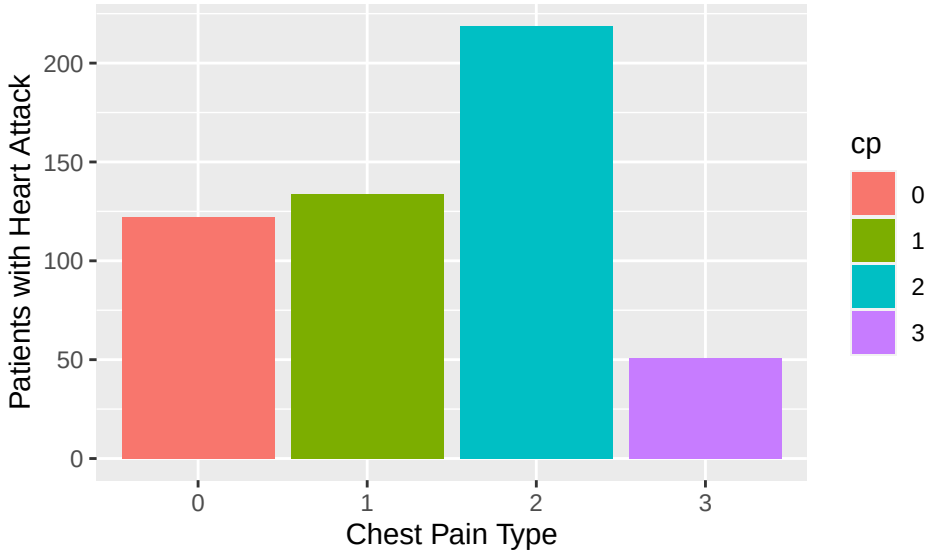
So we can see that near 72.5% of the total women in this data are positive cases while this ration is only 42% in men. This shows that the gender of the patients has a great effect in heart attack and women suffer more than men.

## Chest Pain

The next feature in the patients data is the chest pain. In the data the chest pain is defined with four values including ("0", "1", "2", "3").

| Var1 | Freq |
|---|---|
| 0 | 497 |
| 1 | 167 |
| 2 | 284 |
| 3 | 77 |



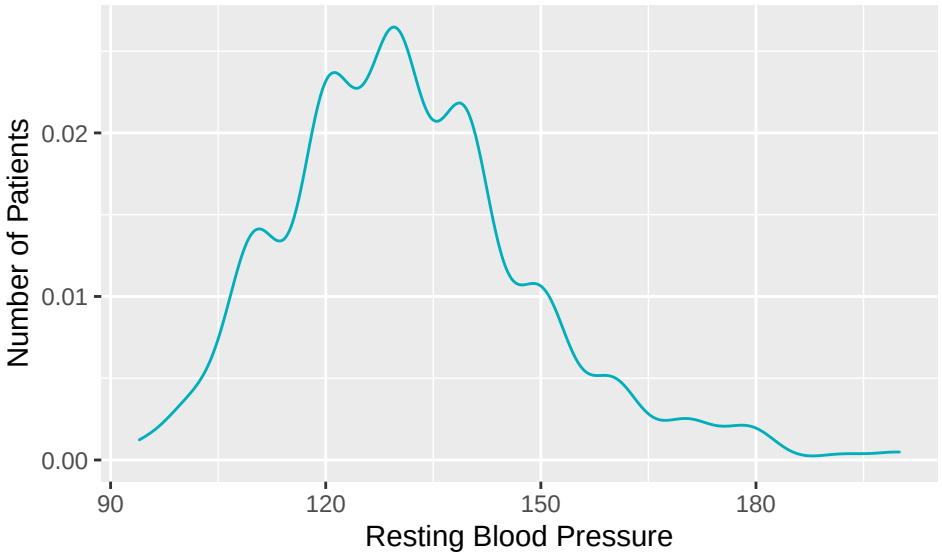
| chest_pain | number |
|---|---|
| type 0 total | 497.0000000 |
| type 0 positives | 122.0000000 |
| positive type 0 to total ratio | 0.2454728 |
| type 1 total | 167.0000000 |
| type 1 positive | 134.0000000 |
| positive type 1 to total ratio | 0.8023952 |
| type 2 total | 284.0000000 |
| type 2 positive | 219.0000000 |
| positive type 2 to total ratio | 0.7711268 |
| type 3 total | 77.0000000 |
| type 3 positive | 51.0000000 |
| positive type 3 to total ratio | 0.6623377 |

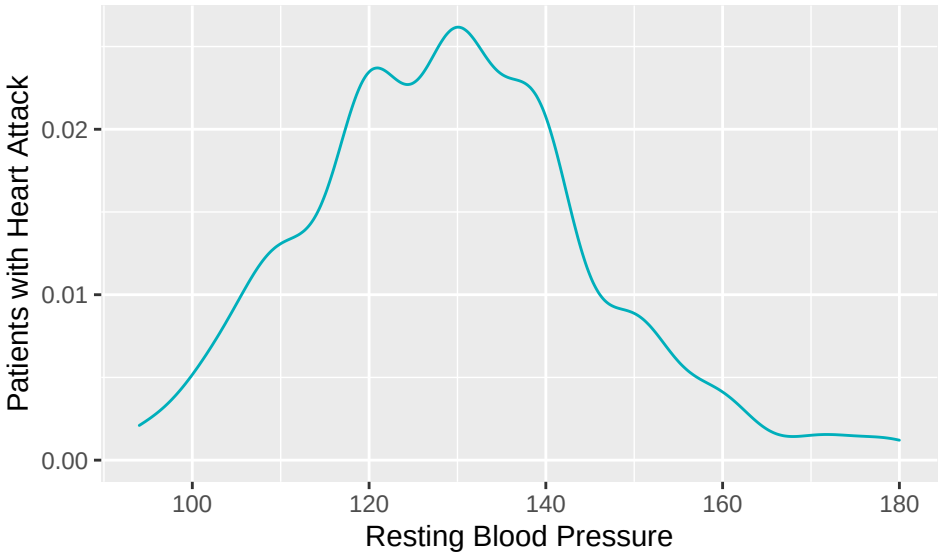
We can see that most of the patients who had heart attacks were those who had the type 2 chest pain and after that type 2 and 1 were in the second and third position with a little difference. But the import point is that 80% of the patients with chest pain type 1 had heart attacks which means type 1 is the one which most causes heart attack, then there was type 2 with 77%, type 3 with 66% and at last type 0 with 24.5%.

### Resting Blood Pressure

The next feature of the patients of the data is their resting blood pressure. We can see that most of the patients in the data have a resting blood pressure between 105 and 150. The median of the resting blood pressure is 130, the mean is 131.6 and the range of it is from 94 to 200.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    94.0   120.0   130.0   131.6   140.0   200.0
```



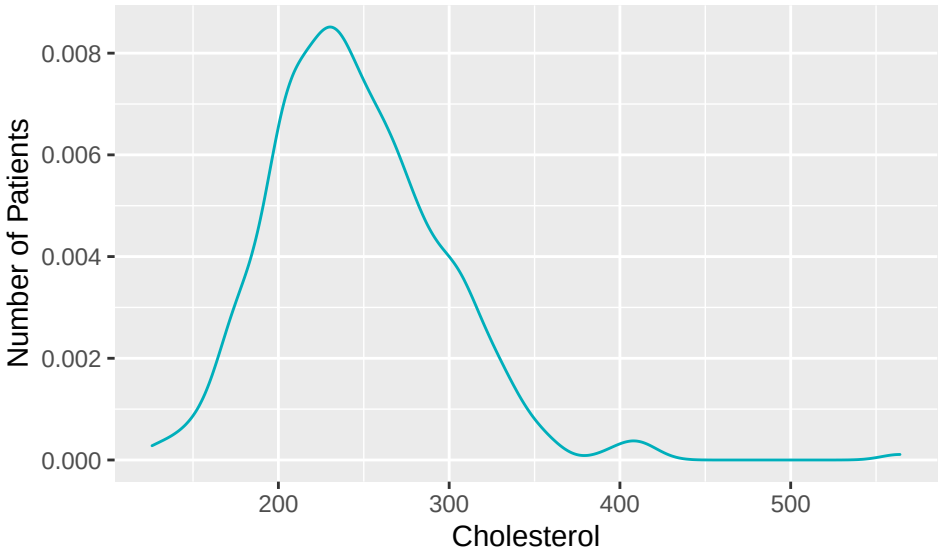



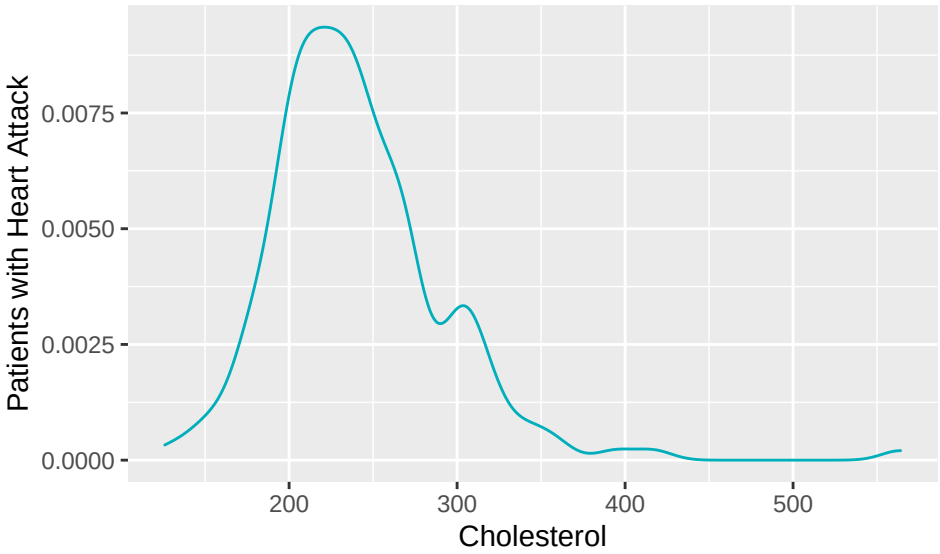

We can see that most of the patients who had heart attack had a resting blood pressure from 110 to 145.

## Cholesterol

Another feature from the patients data is their Cholesterol. The range of the patients cholesterol is between 126 and 564, the mean is 246 which means the peak in the histogram would be in the left side and the median is equal to 240. Most of the patients cholesterol is 175 and 300.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     126     211     240     246     275     564
```



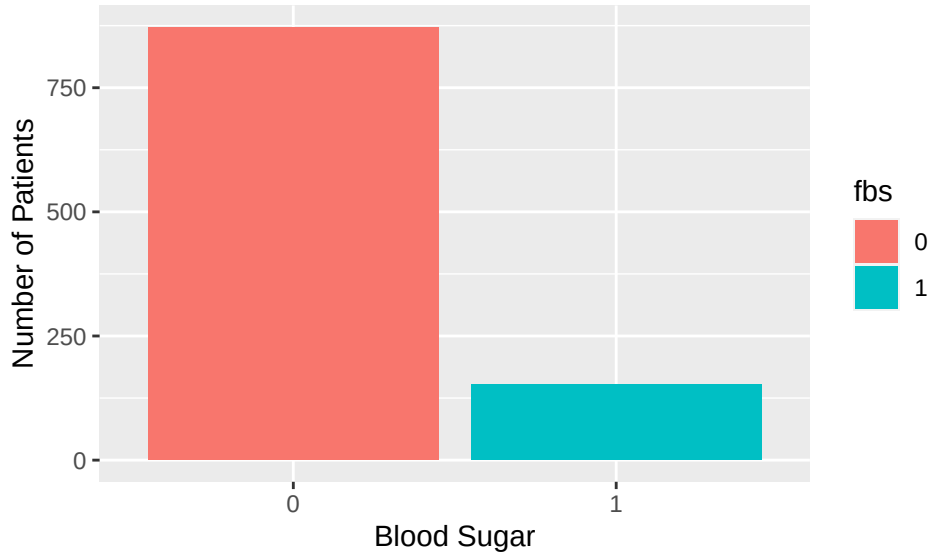



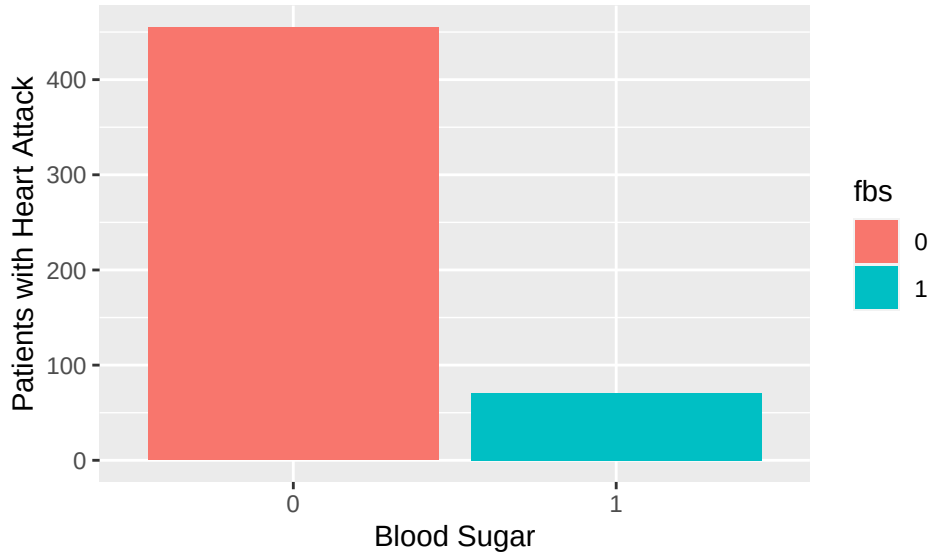

It is clear that most of the patients who are in the positive cases have a cholesterol between 180 and 280.

## Blood Sugar (fbs)

One of the 13 feature of the patients data is their blood sugar, it is defined with 0 and 1 types and 872 patients had type 0 and 153 had type 1.

| Var1 | Freq |
|---|---|
| 0 | 872 |
| 1 | 153 |



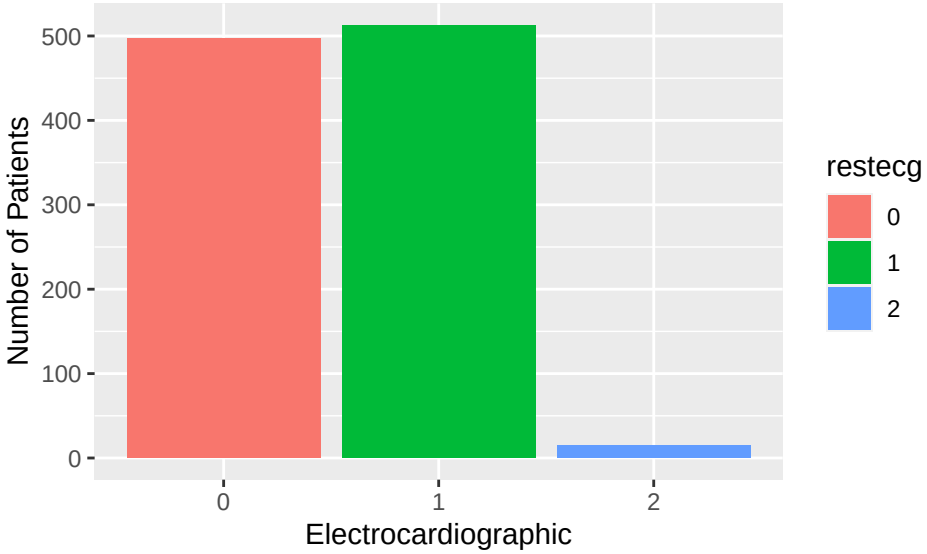
| fbs | number |
|---|---|
| total type 0 | 872.0000000 |
| positive type 0 | 455.0000000 |
| positive type 0 to total ratio | 0.5217890 |
| total type 1 | 153.0000000 |
| positive type 1 | 71.0000000 |
| positive type 1 to total ratio | 0.4640523 |

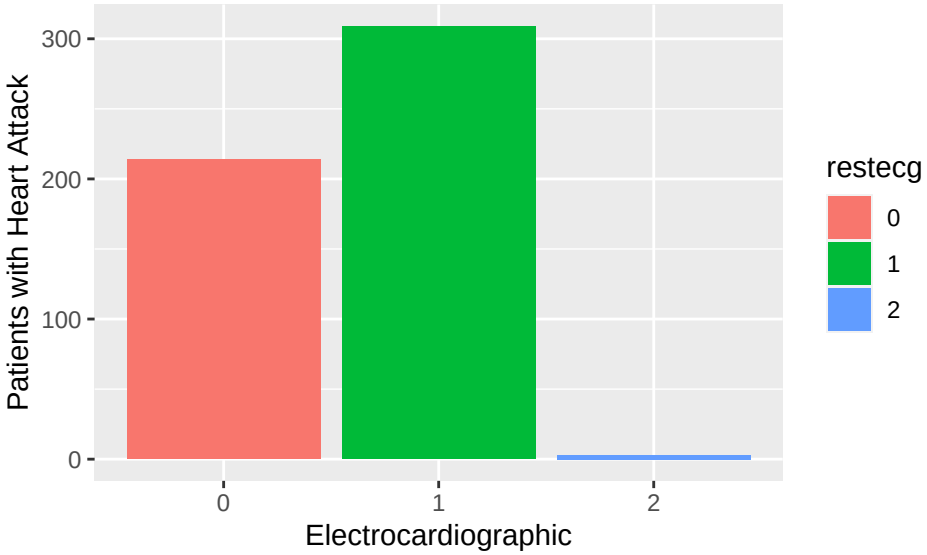
We can see that most of the patients had type 0 blood sugar, also most of the positive cases had type 0 blood pressure but, this does not mean when a patient has type 0 blood pressure we can say that this patient has had heart attack before. Because due to the ratios (proportions) we can see that 46% of the patients with type 0 blood sugar had heart attacks but this number is 52% for those with type 1 blood sugar which means patients with type 1 blood sugar are more likely to have heart attacks but in total we can also build are models without the blood sugar effect to see how it will be because 46 and 52 are close numbers.

### Electrocardiographic

The next feature is the Electrocardiograph which has three values : 0, 1, 2. There are 497 patients with the value 0, 513 with 1 and just 15 with 2 in the 1025v patients in the data.

| Var1 | Freq |
|---|---|
| 0 | 497 |
| 1 | 513 |
| 2 | 15 |



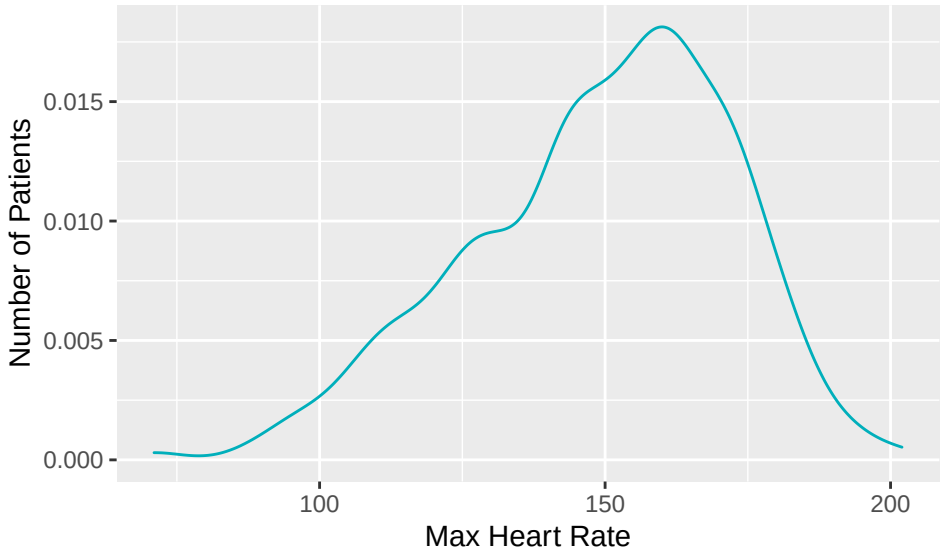
| restecg | number |
|---|---|
| total type 0 | 497.0000000 |
| positive type 0 | 214.0000000 |
| positive type 0 to total ratio | 0.4305835 |
| total type 1 | 513.0000000 |
| positive type 1 | 309.0000000 |
| positive type 1 to total ratio | 0.6023392 |
| total type 1 | 15.0000000 |
| positive type 2 | 3.0000000 |
| positive type 2 to total ratio | 0.2000000 |

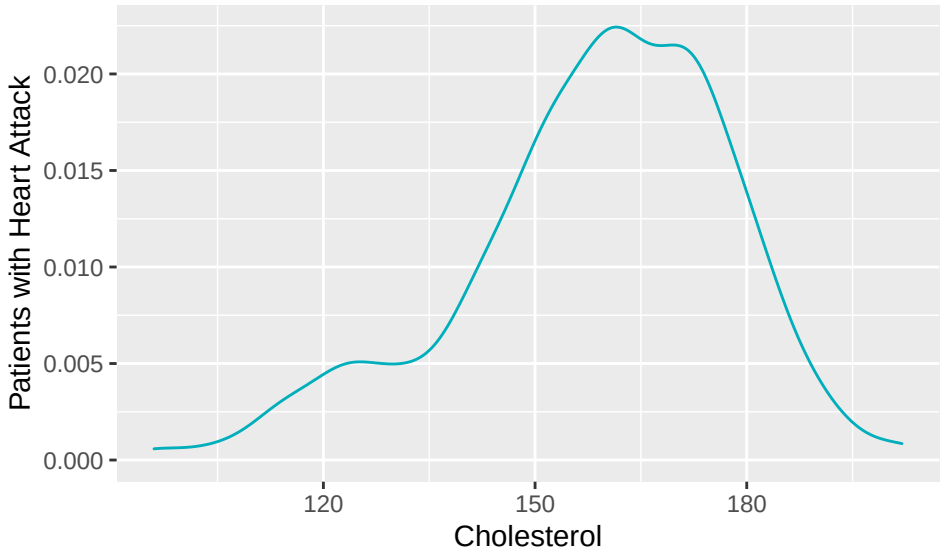
We can see that most of the patients had a type 0 or 1 Electrocardiographic in their data but only 15 had a type 2 and also the ratio of positive cases to total in each type is equal to 43%, 60% and 20%.

### Max Heart Rate

The maximum heart rate of the patients is another factor which can effect the occurrence of heart attack in them. In the data we have we can see that most of the patients have a max heart rate near 162, the mean of the data is 149.1 and the median is 152. Also its noteworthy to mention that the range of the data is between 71 and 202.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    71.0   132.0   152.0   149.1   166.0   202.0
```



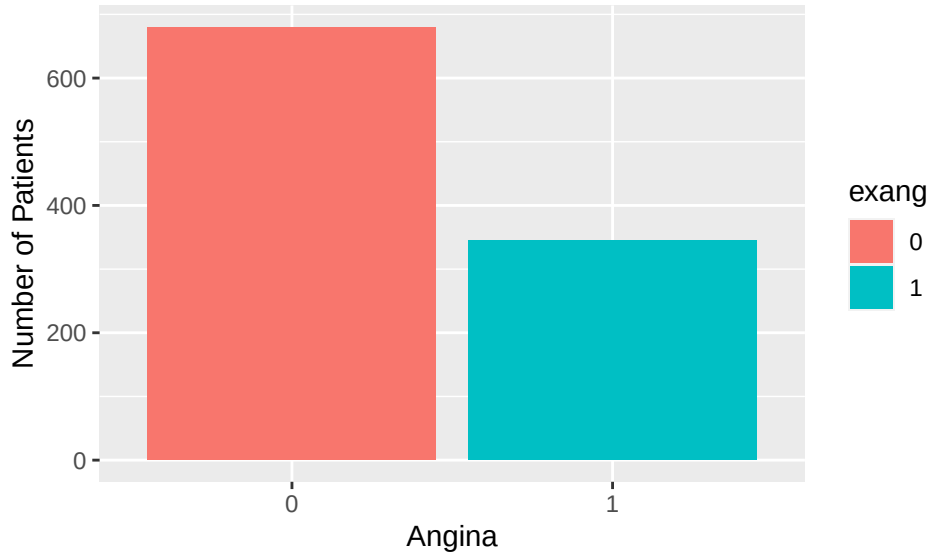



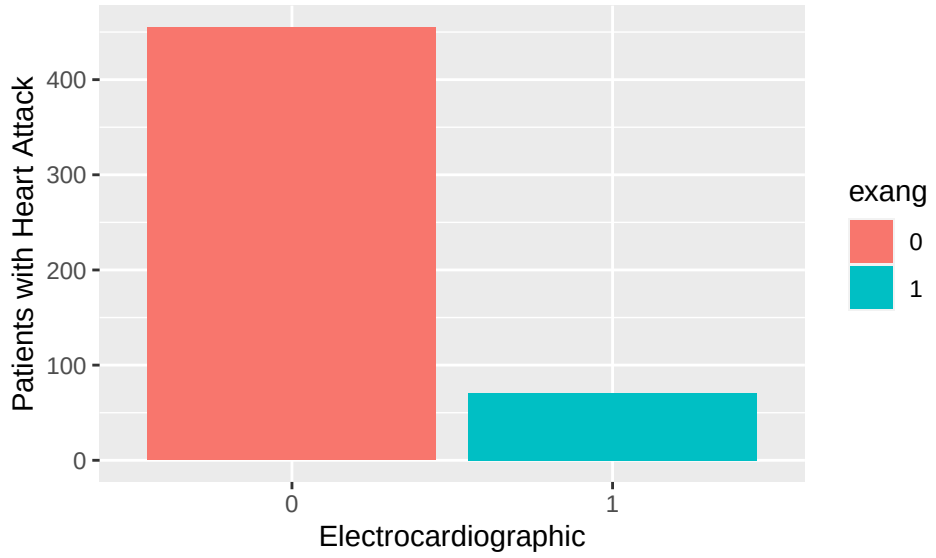

According to the histogram that shows the patients with heart attack we can understand hat most of them were in the range of 145 to 185.

## Angina

Another one of the columns of the patients data is the Angina column which has values equal to 0 or 1. From the 1025 patients in the data we are exploring 680 have type 0 angina and 345 have type 1.

| Var1 | Freq |
|---|---|
| 0 | 680 |
| 1 | 345 |



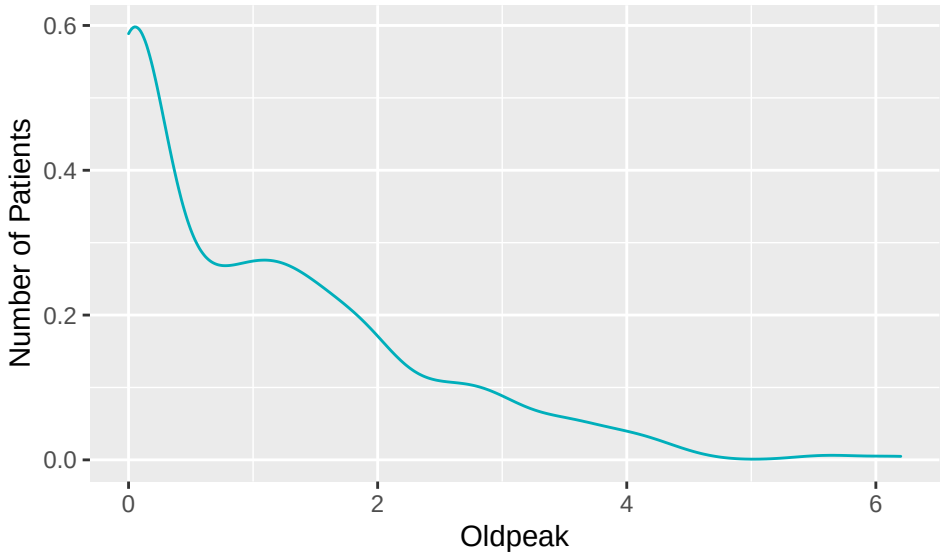
| angina | number |
|---|---|
| total type 0 | 680.0000000 |
| positive type 0 | 455.0000000 |
| positive type 0 to total ratio | 0.6691176 |
| total type 1 | 345.0000000 |
| positive type 1 | 71.0000000 |
| positive type 1 to total ratio | 0.2057971 |

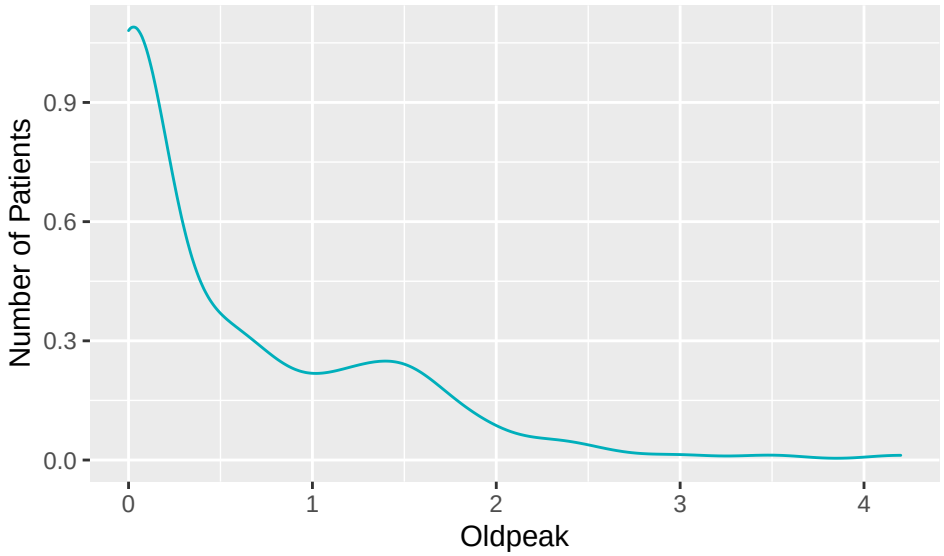
We can see that among the patients with type 0 angina about 67% of them have had heart attack before and are in the positive cases in the data but this number is only 20% among the patients with type 1 angina which means if a person has type 0 angina that person is more likely to have a heart attack in the future.

### Oldpeak

The Oldpeak is the next column in the data which has a mean equal to 1.072, and a range between 0 and 6.2. The trend in our histogram shows that by the increase of oldpeak we have a less patients. This trend is also the same in the histogram with positive cases but in this one the decrease of patients by the increase of oldpeak is faster.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.800   1.072   1.800   6.200
```

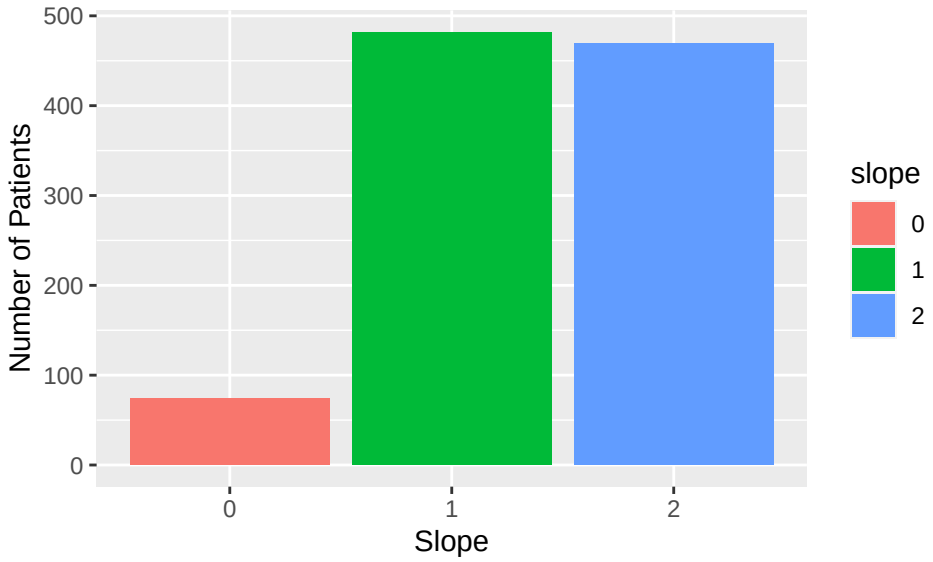


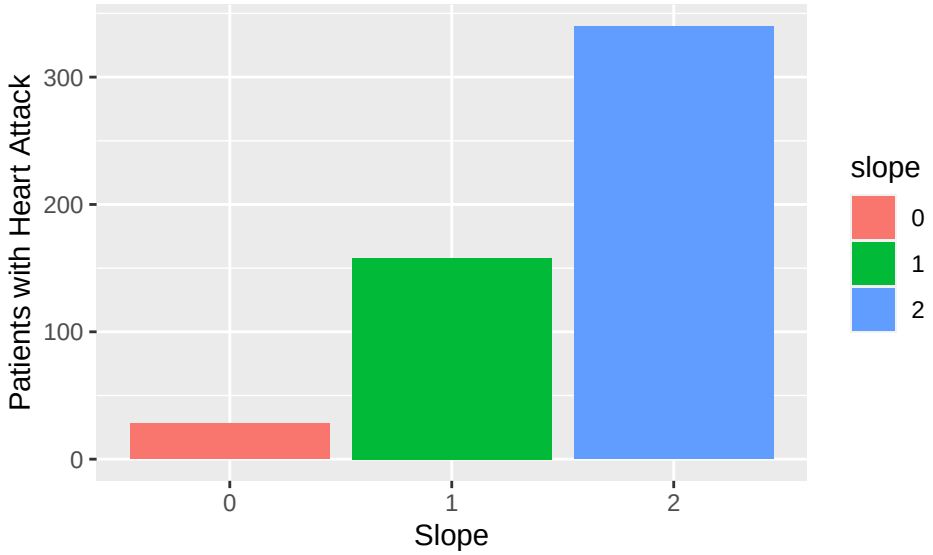



## Slope

The next feature in the data of the patients is Slope. Slope has three values including 0, 1 and 2. There are 74 patients with type 0 slope, 482 with type 1 and 469 with type 2.

| Var1 | Freq |
|---|---|
| 0 | 74 |
| 1 | 482 |
| 2 | 469 |



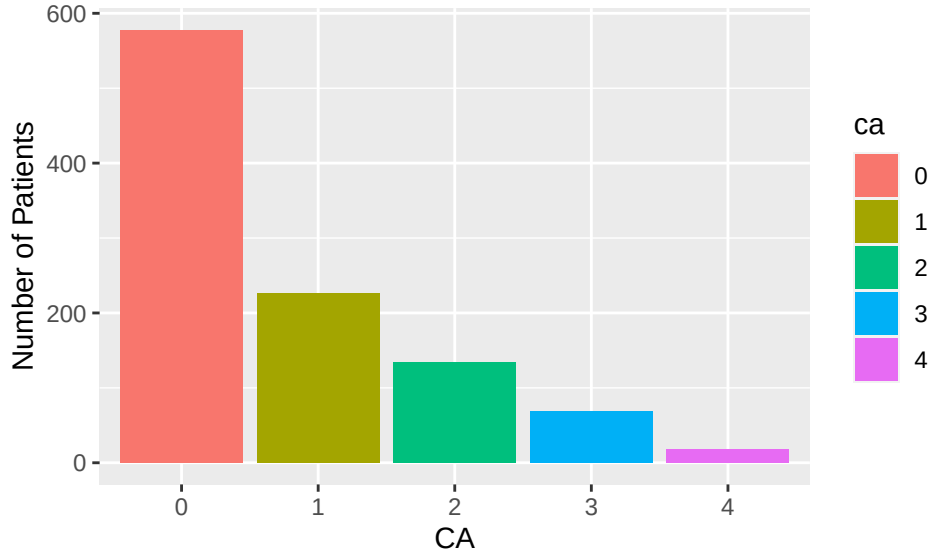
| slope | number |
|---|---|
| total type 0 | 74.0000000 |
| positive type 0 | 28.0000000 |
| positive type 0 to total ratio | 0.3783784 |
| total type 1 | 482.0000000 |
| positive type 1 | 158.0000000 |
| positive type 1 to total ratio | 0.3278008 |
| total type 2 | 469.0000000 |
| positive type 2 | 340.0000000 |
| positive type 2 to total ratio | 0.7249467 |

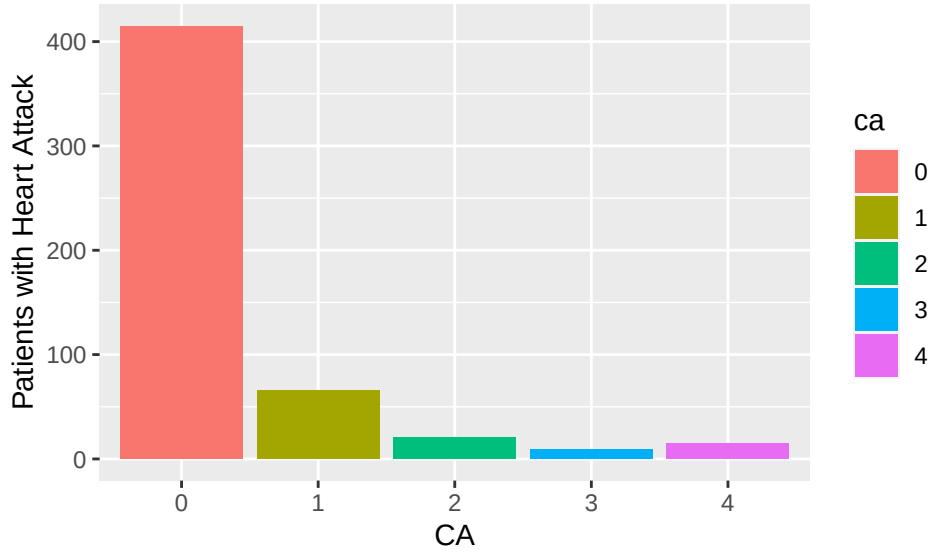
37% of the patients who had a type 0 slope have suffered hart attack and this number is for type 1 slope 32% and 3, 72% which shows patients with type 2 slope are the most in danger of heart attack and then there are type 0 and type 1.

## CA

The next feature in the data of the patients is CA. CA has five values including 0, 1, 2, 3 and 4. There are 578 patients with type 0 CA, 226 with type 1, 134 with type 2, 69 with type 3 and 18 with type 4.

| Var1 | Freq |
|---|---|
| 0 | 578 |
| 1 | 226 |
| 2 | 134 |
| 3 | 69 |
| 4 | 18 |



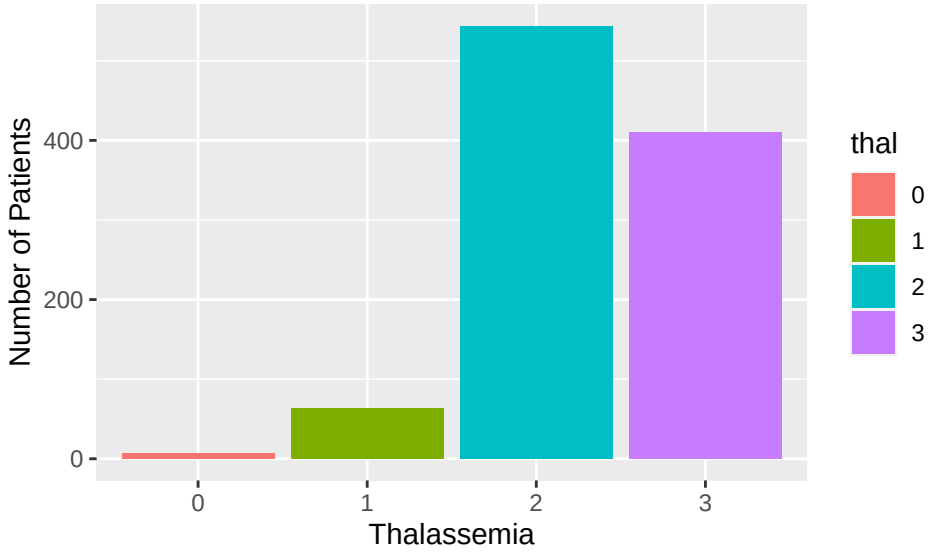
| CA | number |
|---|---|
| total type 0 | 578.0000000 |
| positive type 0 | 415.0000000 |
| positive type 0 to total ratio | 0.7179931 |
| total type 1 | 226.0000000 |
| positive type 1 | 66.0000000 |
| positive type 1 to total ratio | 0.2920354 |
| total type 2 | 134.0000000 |
| positive type 2 | 21.0000000 |
| positive type 2 to total ratio | 0.1567164 |
| total type 3 | 69.0000000 |
| positive type 3 | 9.0000000 |
| positive type 3 to total ratio | 0.1304348 |
| total type 4 | 18.0000000 |
| positive type 4 | 15.0000000 |
| positive type 4 to total ratio | 0.8333333 |

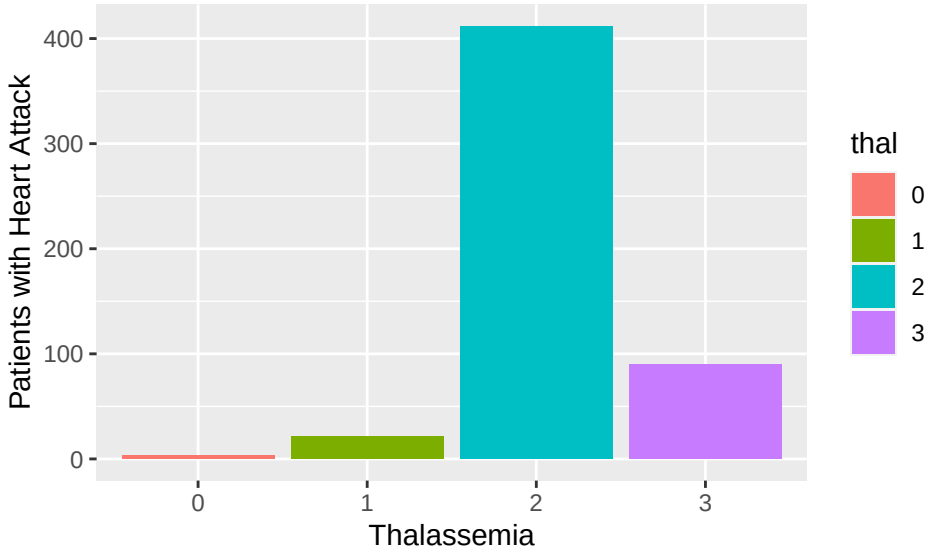
71% of the patients who had a type 0 CA have suffered hart attack and this number is for type 1 CA 29% , type 2 16%, type 3 13% and type 4 83% which shows patients with type 4 and after that type 0 are the most in danger and are more likely to have a heart attack.

### Thalassemia

The Last feature in the data of the patients is Thalassemia. Thalassemia has four values including 0, 1, 2 and 3. There are 7 patients with type 0 CA, 64 with type 1, 544 with type 2 and 410 with type 3.

| Var1 | Freq |
|---|---|
| 0 | 578 |
| 1 | 226 |
| 2 | 134 |
| 3 | 69 |
| 4 | 18 |



| Thalassemia | number |
|---|---|
| total type 0 | 7.0000000 |
| positive type 0 | 3.0000000 |
| positive type 0 to total ratio | 0.4285714 |
| total type 1 | 64.0000000 |
| positive type 1 | 21.0000000 |
| positive type 1 to total ratio | 0.3281250 |
| total type 2 | 544.0000000 |
| positive type 2 | 412.0000000 |
| positive type 2 to total ratio | 0.7573529 |
| total type 3 | 410.0000000 |
| positive type 3 | 90.0000000 |
| positive type 3 to total ratio | 0.2195122 |

43% of the patients who had a type 0 Thalassemia have suffered hart attack and this number is for type 1 Thalassemia 33% , type 2 76% and type 3 22%, which shows patients with type 2 are the most in danger and are more likely to have a heart attack and should start their treatment.

# Models

Now the features described in the previous parts are going to be used to build the three models below :

1- linear regression, lm

2- logistic regression, glm

3- k nearest neighbors

not only the three models mentioned above are going to be built based on the data we had, after that we are going to build a function which normalizes the data and changes all of them into a range between 0 and 1 and the three mentioned models are going to be built once again with the normalized data.

First of all the 1025 data from the patients is going to be split into two new sets, train and test set which they are going to be used to first train the models with the train set and then evaluate them with the test set. To so this the createDataPartition() function is going to be used.

```
set.seed(1)
index <- createDataPartition(y = data$target, times = 1, p = 0.1, list = FALSE)
train <- data[-index,]
test <- data[index,]

dim(train) # the train data set is 90% of the data
```

```
## [1] 922  14
```

```
dim(test) # the test data set is 10% of the data
```

```
## [1] 103  14
```

### first model : linear regression, lm

Now the first model which is the linear regression model can be trained and evaluated.

```
train_1 <- train(target~., method = "lm", data = train)
```

```
## Warning in train.default(x, y, weights = w, ...): You are trying to do
## regression and your outcome only has two possible values Are you trying to do
## classification? If so, use a 2 level factor as your outcome column.
```

```
prediction_1 <- round(predict(train_1, test))
confusionmatrix_1 <- confusionMatrix(as.factor(prediction_1), as.factor(test$target))
acc_1 <- confusionmatrix_1$overall[["Accuracy"]]
results <- tibble(model = "linear regression, lm", Accuracy = acc_1)
results %>% knitr::kable()
```

| model | Accuracy |
|---|---|
| linear regression, lm | 0.9029126 |

**second model : logistic regression, glm**

The second model as mentioned before is the logistic regression model.

```
train_2 <- train(target~., method="glm", train)
```

```
## Warning in train.default(x, y, weights = w, ...): You are trying to do
## regression and your outcome only has two possible values Are you trying to do
## classification? If so, use a 2 level factor as your outcome column.
```

```
prediction_2 <- round(predict(train_2, test))
confusionmatrix_2 <- confusionMatrix(as.factor(prediction_2), as.factor(test$target))
acc_2 <- confusionmatrix_2$overall[["Accuracy"]]
results <- bind_rows(results, tibble(model = "logistic regression, glm"
                                     ,Accuracy = acc_2 ))
results %>% knitr::kable()
```

| model | Accuracy |
|---|---|
| linear regression, lm | 0.9029126 |
| logistic regression, glm | 0.9029126 |

**last model : knn**

And the last model is the k nearest neighbors.

```
train_3 <- train(target~., method="knn", train)
```

```
## Warning in train.default(x, y, weights = w, ...): You are trying to do
## regression and your outcome only has two possible values Are you trying to do
## classification? If so, use a 2 level factor as your outcome column.
```

```
prediction_3 <- round(predict(train_3, test))
confusionmatrix_3 <- confusionMatrix(as.factor(prediction_3), as.factor(test$target))
acc_3 <- confusionmatrix_3$overall[["Accuracy"]]
results <- bind_rows(results, tibble(model = "k nearest neighbors"
                                     ,Accuracy = acc_3 ))
results %>% knitr::kable()
```

| model | Accuracy |
|---|---|
| linear regression, lm | 0.9029126 |
| logistic regression, glm | 0.9029126 |
| k nearest neighbors | 0.6893204 |

## Normalizing the Data

Now the same models are going to be built but just with new data that is the normalized version of the data used to build the three previous models.

Normalizing function :

```
new_data <- read.csv("/Users/omid/Desktop/harvard/heart/heart//heart.csv-edx.xls")

normalize <- function(x) {
  return ((x - min(x)) / (max(x) - min(x)))
}

normal <- as.data.frame(lapply(new_data, normalize))
```

Also to do this new train and test data sets are needed from the normalized data:

```
set.seed(1)
index_normal <- createDataPartition(y = normal$target, times = 1, p = 0.1, list = FALSE)
train_normal <- normal[-index_normal,]
test_normal <- normal[index_normal,]

dim(train_normal) # the train data set is 90% of the data
```

```
## [1] 922  14
```

```
dim(test_normal) # the test data set is 10% of the data
```

```
## [1] 103  14
```

### first model with normalized data : linear regression, lm

Now the first model which is the linear regression model can be trained and evaluated with the normalized data.

```
train_1_normal <- train(target~., method = "lm", data = normal)
```

```
## Warning in train.default(x, y, weights = w, ...): You are trying to do
## regression and your outcome only has two possible values Are you trying to do
## classification? If so, use a 2 level factor as your outcome column.
```

```
prediction_1_normal <- round(predict(train_1_normal, test_normal))
confusionmatrix_1_normal <- confusionMatrix(
  as.factor(prediction_1_normal), as.factor(test_normal$target))
acc_1_normal <- confusionmatrix_1_normal$overall[["Accuracy"]]
results <- bind_rows(results, tibble(model = "linear regression, lm, normalized data"
                                     ,Accuracy = acc_1_normal ))
```

```
results %>% knitr::kable()
```

| model | Accuracy |
|---|---|
| linear regression, lm | 0.9029126 |
| logistic regression, glm | 0.9029126 |
| k nearest neighbors | 0.6893204 |
| linear regression, lm, normalized data | 0.8737864 |

**second model with normalized data : logistic regression, glm**

The second model as mentioned before is the logistic regression model but now built with the normalized data.

```
train_2_normal <- train(target~., method = "glm", data = normal)
```

```
## Warning in train.default(x, y, weights = w, ...): You are trying to do
## regression and your outcome only has two possible values Are you trying to do
## classification? If so, use a 2 level factor as your outcome column.
```

```
prediction_2_normal <- round(predict(train_2_normal, test_normal))
confusionmatrix_2_normal <- confusionMatrix(
  as.factor(prediction_2_normal), as.factor(test_normal$target))
acc_2_normal <- confusionmatrix_2_normal$overall[["Accuracy"]]
results <- bind_rows(results, tibble(model = "linear regression, glm, normalized data"
                                     ,Accuracy = acc_2_normal ))
results %>% knitr::kable()
```

| model | Accuracy |
|---|---|
| linear regression, lm | 0.9029126 |
| logistic regression, glm | 0.9029126 |
| k nearest neighbors | 0.6893204 |
| linear regression, lm, normalized data | 0.8737864 |
| linear regression, glm, normalized data | 0.8737864 |

**last model with normalized data : knn**

And the last model is the k nearest neighbors with normalized data.

```
train_3_normal <- train(target~., method = "knn", data = normal)
```

```
## Warning in train.default(x, y, weights = w, ...): You are trying to do
## regression and your outcome only has two possible values Are you trying to do
## classification? If so, use a 2 level factor as your outcome column.
```

```
prediction_3_normal <- round(predict(train_3_normal, test_normal))
confusionmatrix_3_normal <- confusionMatrix(
  as.factor(prediction_3_normal), as.factor(test_normal$target))
acc_3_normal <- confusionmatrix_3_normal$overall[["Accuracy"]]
results <- bind_rows(results, tibble(model = "k nearest neighbors normalized data"
                                     ,Accuracy = acc_3_normal ))
results %>% knitr::kable()
```

| model | Accuracy |
|---|---|
| linear regression, lm | 0.9029126 |
| logistic regression, glm | 0.9029126 |
| k nearest neighbors | 0.6893204 |
| linear regression, lm, normalized data | 0.8737864 |
| linear regression, glm, normalized data | 0.8737864 |
| k nearest neighbors normalized data | 0.9417476 |

# Conclusion

According to the three models built with the original data and the normalized data we can see that it is possible to build a model with 94.1% accuracy to make predictions and saves the life of people in danger of having a heart attack.

In the three models built with the original and normalized data, the linear regression and logistic regression had better accuracy with the original data, both 90.2% in comparison to the knn model which had an accuracy equal to 68.9%. But when the normalized data is used to build the models and make predictions the linear and logistic regression models had an accuracy equal to 87.3% while the knn model had 94.1% of accuracy which means by normalizing the data and the using the knn model, better predictions can be made and therefore more lives can be saved.

Also further studies can be done to find more features in the patients like smoking and etc. to make better and more accurate predictions.