Heart Disease Dataset Analysis

By Anupma Mangla
June 6th, 2019

Table of Contents

Problem Statement

To analyze and find any trends in the heart disease dataset to predict presence/absence of heart disease and to predict certain cardiovascular events.

Research Objectives

Heart disease describes a range of conditions that affect your heart. Diseases under the heart disease umbrella include blood vessel diseases, such as coronary artery disease, heart rhythm problems (arrhythmias), and heart defects you're born with (congenital heart defects), to name a few.

The term "heart disease" is often used interchangeably with the term "cardiovascular disease." Cardiovascular disease generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina) or stroke. Other heart conditions, such as those that affect your heart's muscle, valves or rhythm, also are considered forms of heart disease.

Among the many types of heart disease, here are some examples: Angina, Arrhythmia, Congenital heart disease, Coronary artery disease (CAD), Dilated cardiomyopathy, Heart attack (myocardial infarction), Heart failure, Hypertrophic cardiomyopathy, Mitral regurgitation, Mitral valve prolapse, and Pulmonary stenosis. Coronary artery disease (CAD) is the most common type of heart disease in the US. Coronary arteries supply blood to the heart muscle and coronary artery disease occurs when there is a buildup of cholesterol plaque inside the artery walls.

According to data from CDC- Centers for disease control and prevention, about 610,000 people die of heart disease in the United States every year. That’s 1 in every 4 deaths and heart disease is the leading cause of death for both men and women. More than half of the deaths due to heart disease in 2009 were in men and Coronary heart disease (CHD) is the most common type of heart disease, killing over 370,000 people annually.

So, our motivation was to research and analyze the heart disease dataset to understand the given attributes to determine which are the key indicators of presence/ absence of heart disease:

In this dataset we expect to analyze the given attributes to answer some questions we had:

1. How does gender play a role in heart-disease?
2. How does Age play a role in heart disease?
3. How does the known risk factors like cholesterol, blood pressure and fasting blood sugar play a role in predicting heart disease?

4. Does symptoms like chest pain (different types), exercise induced angina, maximum heart rate have any correlation to predict heart disease?
5. How does tests like treadmill ECG stress test help in predicting heart disease?
6. Does genetic blood disorder (thalassemia) be a cause of heart disease?
7. Can we use predictive analysis using logistic regression, and create a model which can tell us the presence or absence of a heart disease in a person based on the given attributes.

Dataset Information

Dataset is taken from kaggle. URL: https://www.kaggle.com/ronitf/heart-disease-uci
It uses the relevant health exam indicators in patients and analyzes their influences on heart disease. The dataset has 14 key attributes:

1. Age: age of the patient in years
2. Sex: gender of a patient, 1 = male; 0 = female
3. cp: chest pain type
    - Value 1: typical angina,
    - Value 2: atypical angina,
    - Value 3: non-anginal pain,
    - Value 4: asymptomatic

4. trestbps: resting blood pressure of a patient (in mm Hg on admission to the hospital)
5. Chol: serum cholesterol in mg/dl
6. Fbs: fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
7. restecg : resting electrocardiographic results
8. Thalach: maximum heart rate of a patient
9. exang: exercise induced angina (1 = yes; 0 = no)
10. oldpeak: ST depression induced by exercise relative to rest

    This is an observation in treadmill ECG stress test used to diagnose coronary artery disease(CAD). The treadmill electrocardiogram (ECG) stress test is widely used to screen for obstructive coronary artery disease. The presence of ST segment changes, either depression or elevation, on the ECG during the treadmill test often suggests presence of CAD and warrants further management. CAD is diagnosed as the plaque buildup in the arteries, which move oxygen-rich blood through the heart and lungs
11. Slope: the slope of the peak exercise ST segment

The slope is the ST segment/ Heart rate slope which is used as a indicator of coronary heart disease.

12. Ca: number of major vessels (0-3) colored by fluoroscopy
13. Thal: 3 = normal; 6 = fixed defect; 7 = reversible defect

    Thalassemia is an inherited blood disorder characterized by less hemoglobin and fewer red blood cells in your body than normal. The type of thalassemia you have depends on the number of gene mutations you inherit from your parents and which part of the hemoglobin molecule is affected by the mutations. The more mutated genes, the more severe your thalassemia. Heart problems such as congestive heart failure and abnormal heart rhythms (arrhythmias) may be associated with severe thalassemia.

14. Target: 1 or 0

    The “target” field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) and 1 (presence). Target is the dependent variable and rest all the variable are the independent variable.

Tools and packages used for analysis

```
install.packages("ggplot2")
install.packages("gridExtra")
install.packages("caTools")
```

```
library(tidyverse)
library(corrplot)
library(gridExtra)
library(caTools)
library(dplyr)
```

```
── Attaching packages ─────────────────────────────────────── tidyverse 1.2.1 ──
✔ ggplot2 3.1.1       ✔ purrr   0.3.2
✔ tibble 2.1.1        ✔ dplyr   0.8.0.1
✔ tidyr  0.8.3        ✔ stringr 1.4.0
✔ readr  1.3.1        ✔ forcats 0.4.0
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
```

Loading the dataset

```
input <- read.csv("~/Desktop/Project/heart.csv")
dim(input)
str(input)
```

```
[1] 303  14

'data.frame': 303 obs. of  14 variables:
 $ age     : int  63 37 41 56 57 57 56 44 52 57 ...
 $ sex     : int  1 1 0 1 0 1 0 1 1 1 ...
 $ cp      : int  3 2 1 1 0 0 1 1 2 2 ...
 $ trestbps: int  145 130 130 120 120 140 140 120 172 150 ...
 $ chol    : int  233 250 204 236 354 192 294 263 199 168 ...
 $ fbs     : int  1 0 0 0 0 0 0 0 1 0 ...
 $ restecg : int  0 1 0 1 1 1 0 1 1 1 ...
 $ thalach : int  150 187 172 178 163 148 153 173 162 174 ...
 $ exang   : int  0 0 0 0 1 0 0 0 0 0 ...
 $ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
 $ slope   : int  0 0 2 2 2 1 1 2 2 2 ...
 $ ca      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ thal    : int  1 2 2 2 2 1 2 3 3 2 ...
 $ target  : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
head(input)
```

	age <int>	sex <int>	cp <int>	trestbps <int>	chol <int>	fbs <int>	restecg <int>	thalach <int>	exang <int>
1	63	1	3	145	233	1	0	150	0
2	37	1	2	130	250	0	1	187	0
3	41	0	1	130	204	0	0	172	0
4	56	1	1	120	236	0	1	178	0
5	57	0	0	120	354	0	1	163	1
6	57	1	0	140	192	0	1	148	0

Data pre-processing

We changed the columns sex and target to their corresponding values. Changed the integer value of sex from 0/1 to male/female and change the integer value of target from 0/1 to string - “No Heart Disease/ Heart Disease”

```
input$sex <- ifelse((input$sex == 1), 'male','female')
input$target <- ifelse((input$target == 1), 'HeartDisease','No Heart Disease')
```

```
# Select the processed columns to verify the changes
select_cols <- select(input, "age","sex","target","chol","trestbps","fbs")
select_cols
```

age <int>	sex <chr>	target <chr>	chol <int>	trestbps <int>	fbs <int>
63	male	HeartDisease	233	145	1
37	male	HeartDisease	250	130	0
41	female	HeartDisease	204	130	0
56	male	HeartDisease	236	120	0
57	female	HeartDisease	354	120	0
57	male	HeartDisease	192	140	0
56	female	HeartDisease	294	140	0
44	male	HeartDisease	263	120	0
52	male	HeartDisease	199	172	1
57	male	HeartDisease	168	150	0

Descriptive Analysis

First we used the dataset to gain Insights into what is happening using visualizations.

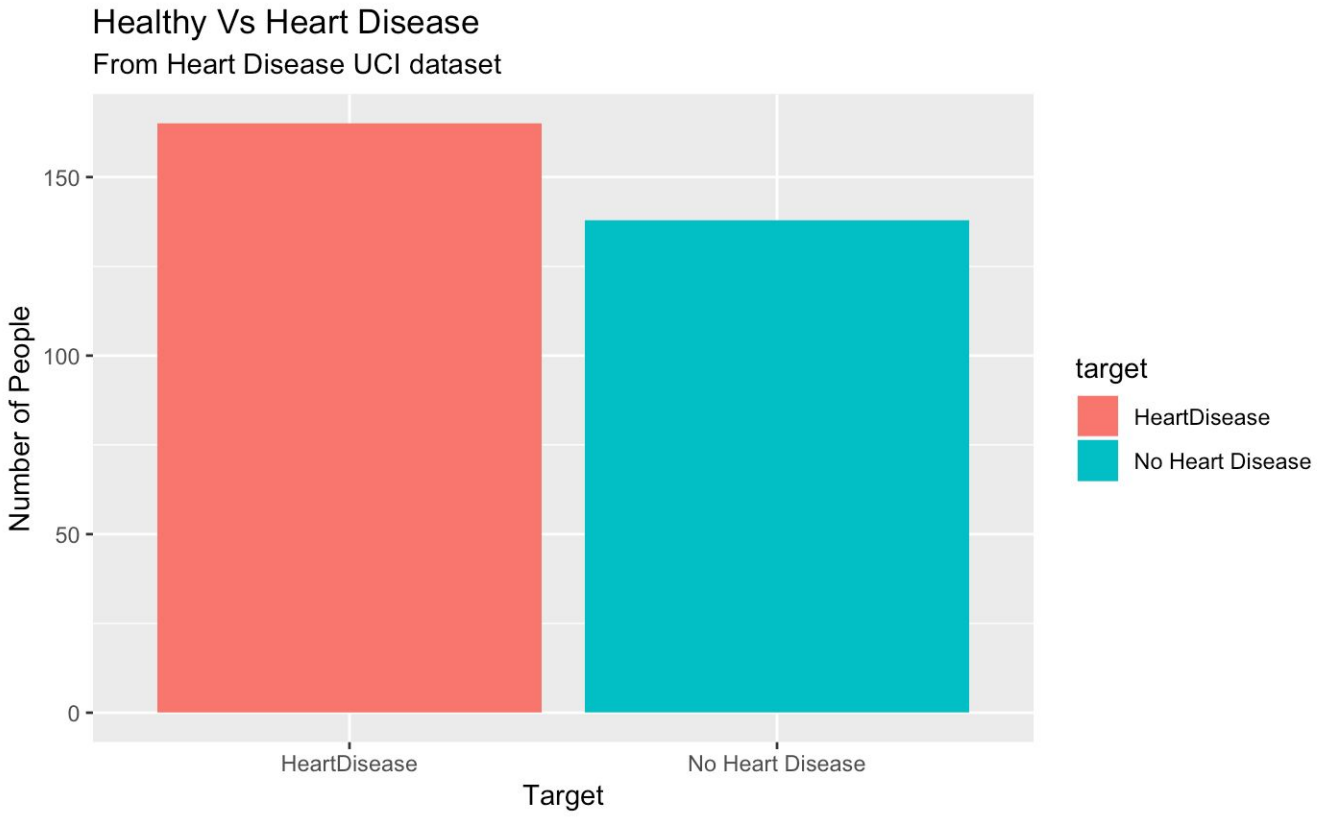
We looked at the number of observations and created a bar plot to determine the observations with no heart disease vs presence of heart disease. Within the dataset we have 55% patients

had heart disease and 45% did not. Hence, we have good percentage of data to look for insights.

```
table(input$target)
```

```
Heart Disease       No Heart Disease
      165                    138
```

```
## Create Bar Plot

ggplot(input,aes(target, fill=target)) + geom_bar(stat="count") +
ggtitle("Healthy Vs Heart Disease", subtitle="From Heart Disease UCI dataset ") +
xlab("Target") + ylab("Number of People")
```

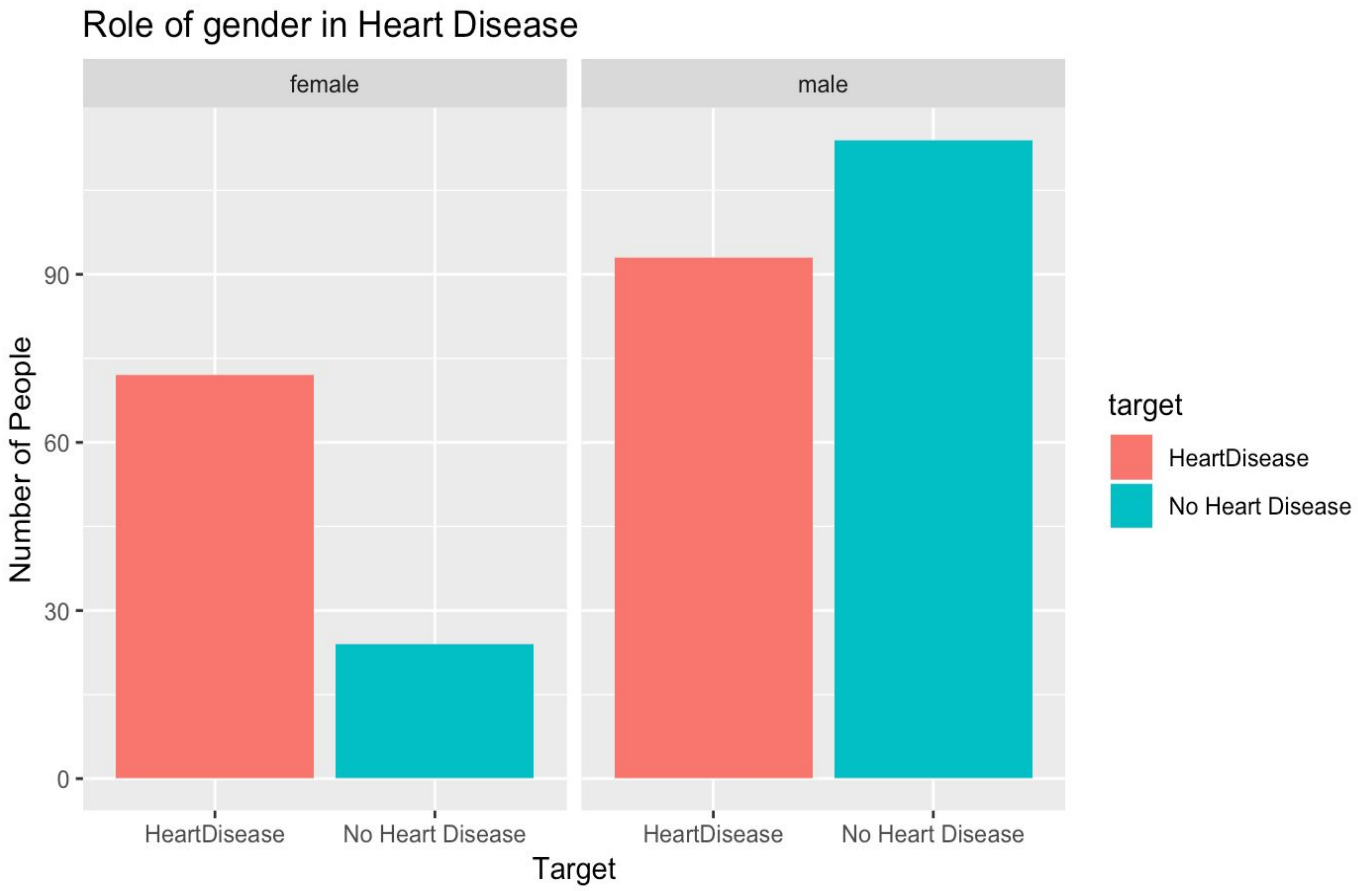


We looked at how does gender play a role in heart-disease? We created a bar plot to determine the number of observations between male and female who have no heart disease vs presence of heart disease. Within the dataset we have we found that more males are healthy within their category, and more females have heart disease within their category.

```
table(input$target, input$sex)
```

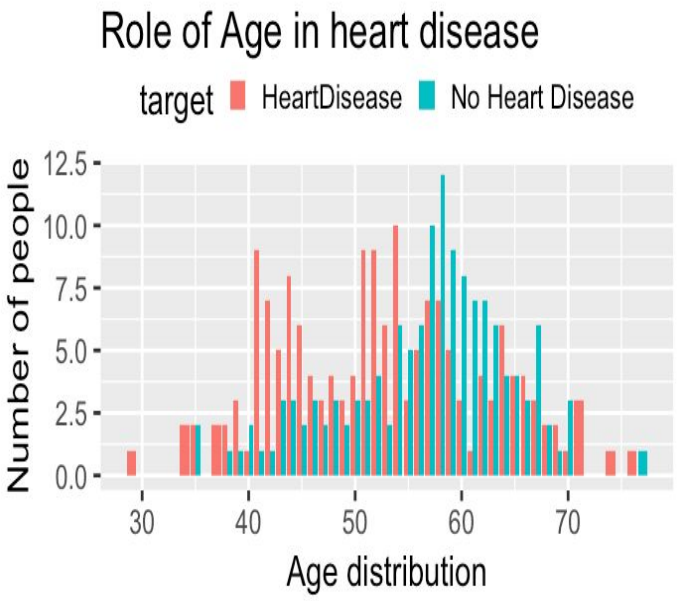
```
                female male
HeartDisease          72   93
No Heart Disease      24  114
```

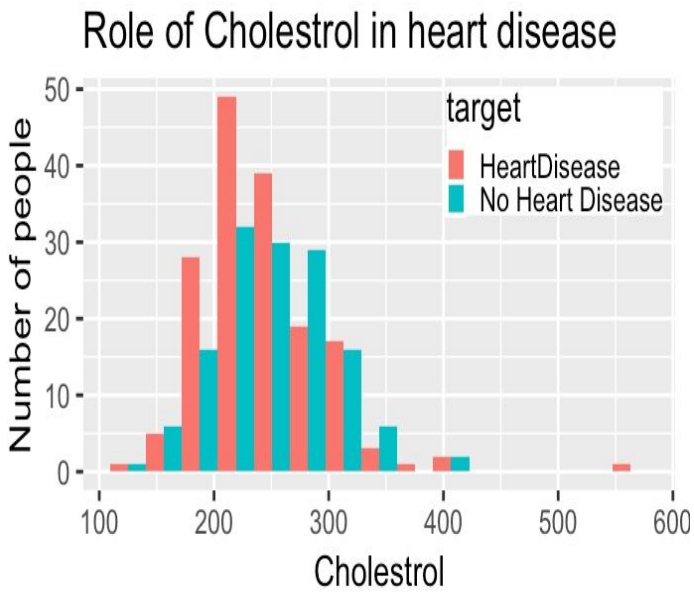
```
## Create Bar Plot

ggplot(input,aes(target, fill=target)) + geom_bar(stat="count") +
ggtitle("Role of gender in Heart Disease") + xlab("Target") + ylab("Number of People") +
facet_wrap(~sex, ncol=2,scale="fixed")
```

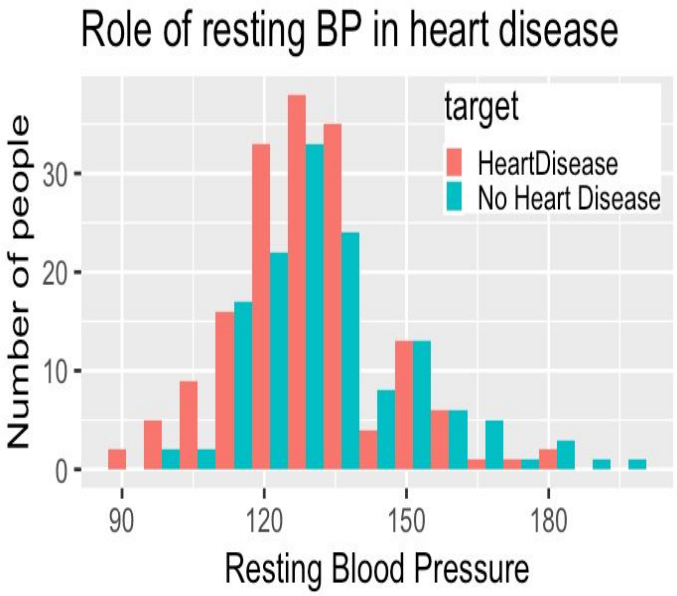


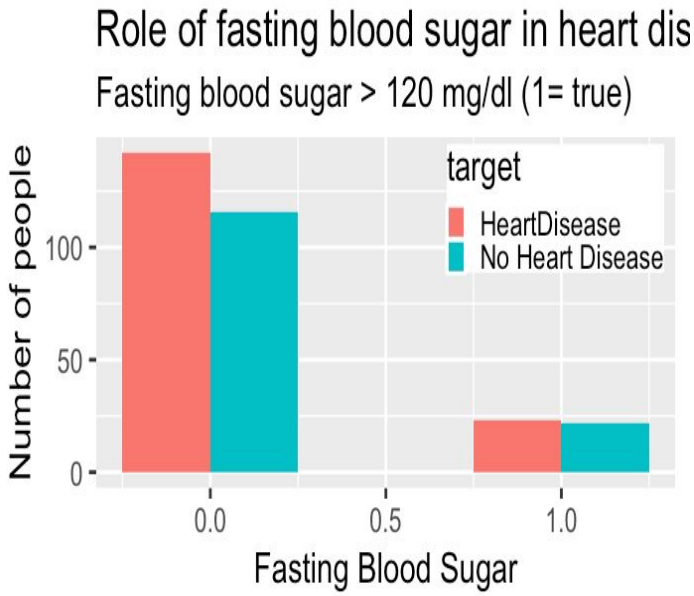
We looked at other known risk factors like age, cholesterol, blood pressure and fasting blood sugar to determine any correlation in predicting heart disease? We created bar plot and histogram to gain some insights.

```
# Role of Age

p1 <- ggplot(data = input) + geom_bar(mapping = aes(x = age ,fill = target), position =
"dodge") + labs(title = "Role of Age in heart disease") + xlab("Age distribution") +
ylab("Number of people") + theme( legend.position ="top", legend.margin=margin(t = 0,
unit='cm'), legend.key.size = unit(0.25, "cm"))

# Role of Cholesterol

p2 <- ggplot(data = input) + geom_histogram(mapping = aes(x =  chol, fill = target), position
= "dodge", bins = 15) + labs(title = "Role of Cholestrol in heart disease") + xlab("Cholestrol")
+ ylab("Number of people") + theme(legend.position = c(0.98, 0.98), legend.justification =
c("right", "top"), legend.margin=margin(t = 0, unit='cm'), legend.key.size = unit(0.25, "cm"))

# Role of resting BP

p3 <- ggplot(data = input) + geom_histogram(mapping = aes(x =  trestbps, fill = target),
position = "dodge", bins = 15) + labs(title = "Role of resting BP in heart disease") +
xlab("Resting Blood Pressure") + ylab("Number of people") + theme( legend.position =
c(0.98, 0.98), legend.justification = c("right", "top"), legend.margin=margin(t = 0, unit='cm'),
legend.key.size = unit(0.25, "cm"))

# Role of fasting blood sugar
# fasting blood sugar > 120 mg/dl = 1
# fasting blood sugar < 120 mg/dl = 0

p4 <- ggplot(data = input) + geom_histogram(mapping = aes(x =  fbs, fill = target), position =
"dodge", bins =3, xlim=c(0,1)) + labs(title = "Role of fasting blood sugar in heart disease",
subtitle="Fasting blood sugar > 120 mg/dl (1= true)") + xlab("Fasting Blood Sugar") +
ylab("Number of people") + theme( legend.position = c(0.98, 0.98), legend.justification =
c("right", "top"), legend.margin=margin(t = 0, unit='cm'), legend.key.size = unit(0.25, "cm"))

## Create a multi-plot with all the 4 plots
grid.arrange(p1,p2,p3,p4, nrow = 2, ncol = 2)
```





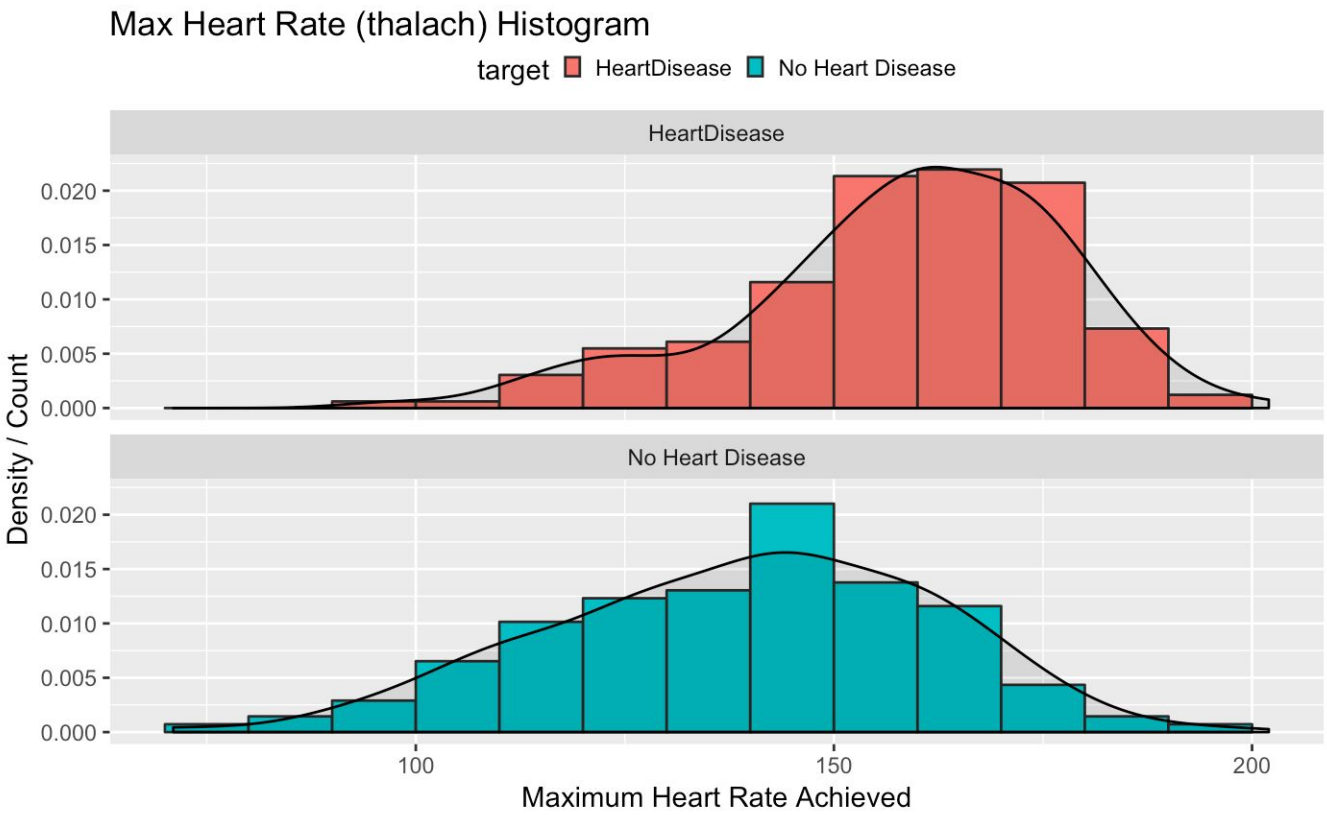




For age, we see a higher no. of patients have heart disease in middle age compared to the older age patients, so we can't make any judgement based on age for prediction of heart disease. For Cholesterol, higher value of cholesterol does not seem to indicate more patients with heart disease. For resting Blood Pressure, high BP doesn't seem to indicate increase in patients with heart disease. We also see that the patients with heart disease have fasting blood sugar < 120 mg/dl, so having high blood sugar (>120 mg/dl) does not seem to infer much in predicting heart disease.

Next, we looked at how max heart rate differ in healthy vs patients with heart disease? Maximum heart rate achieved is measured by the attribute: thalach. We created a histogram to determine the differences. Within the dataset we have we found that patients with heart disease have higher maximum heart rate than patients with no heart disease.

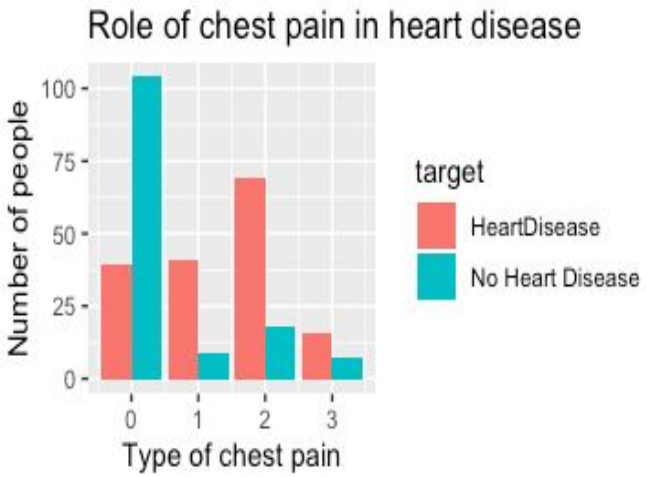
```
## Create histogram
ggplot(input,aes(thalach, fill=target)) +
  geom_histogram(aes(y=..density..),breaks=seq(70, 205, by=10), color="grey17") +
  geom_density(alpha=.1, fill="black") +
  facet_wrap(~target, ncol=1,scale="fixed") +
  ggtitle("Max Heart Rate (thalach) Histogram") + xlab("Maximum Heart Rate Achieved") +
ylab("Density / Count") + theme(legend.position ="top", legend.margin=margin(t = 0,
unit='cm'), legend.key.size = unit(0.25, "cm"))
```

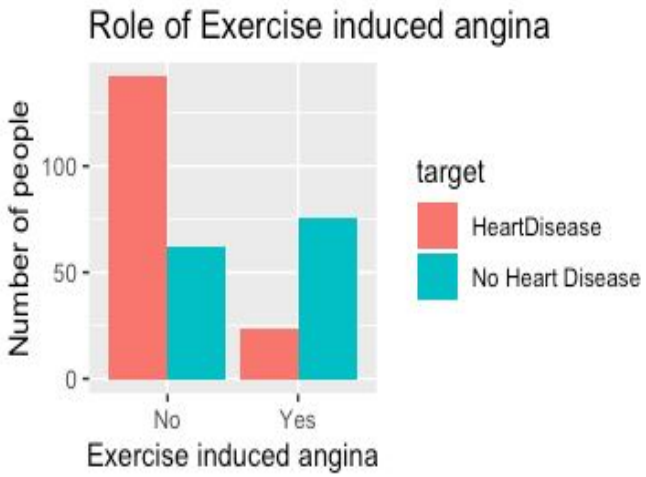


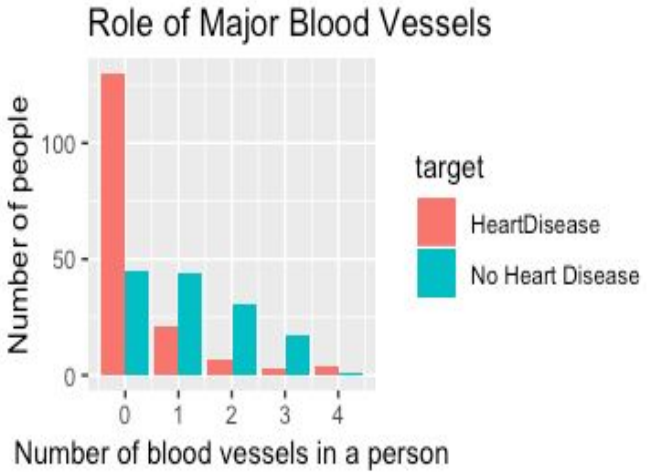
Next, we tried to analyze if symptoms like chest pain (of different types - typical angina, atypical angina, non-angina pain, asymptotic pain), exercise induced angina, number of major blood vessels, slope and genetic blood disorder (thalassemia) can be a cause of heart disease?

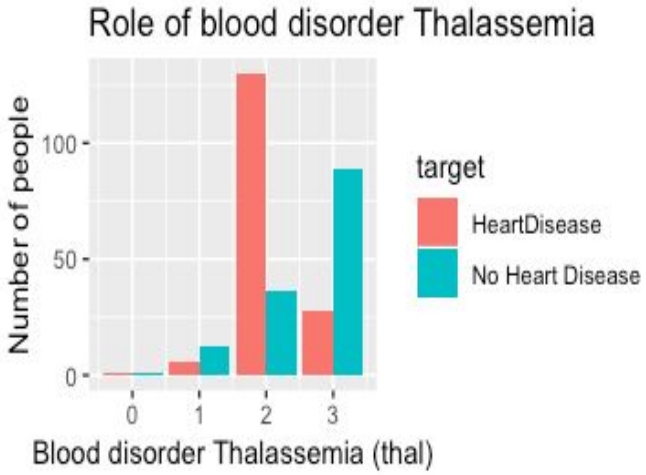
```
# Chest pain type
```

```
p1 <- ggplot(data = input) + geom_bar(mapping = aes(x = cp ,fill = target), position = "dodge") + labs(title = "Role of chest pain in heart disease") + xlab("Type of chest pain") + ylab("Number of people") + theme( legend.position = "top", legend.margin=margin(t = 0, unit='cm'), legend.key.size = unit(0.25, "cm"))

# thalassemia

p2 <- ggplot(data = input) + geom_bar(mapping = aes(x = thal ,fill = target), position = "dodge") + labs(title = "Role of blood disorder thalassemia") + xlab("Thalassemia") + ylab("Number of people") + theme( legend.position ="top", legend.margin=margin(t = 0, unit='cm'), legend.key.size = unit(0.25, "cm"))

# Major blood vessels

p3 <- ggplot(data = input) + geom_bar(mapping = aes(x =  ca, fill = target) , position = "dodge") +  labs(title = "Role of Major Blood Vessels") + xlab("Number of blood vessels in a person") + ylab("Number of people")  + theme( legend.position = c(0.98, 0.98), legend.justification = c("right", "top"), legend.margin=margin(t = 0, unit='cm'), legend.key.size = unit(0.25, "cm"))

# slope

p4 <- ggplot(data = input) + geom_bar(mapping = aes(x = slope ,fill = target), position = "dodge") + labs(title = "The ST segment/heart rate slope as a predictor of CAD") + xlab("Slope") + ylab("Number of people") + theme( legend.position = "top", legend.margin=margin(t = 0, unit='cm'), legend.key.size = unit(0.25, "cm"))

# create multi-plot with all the 4 plots
grid.arrange(p1,p2,p3,p4, nrow = 2, ncol = 2)
```









From the above plots, we can analyse that chest pain type 0,1,2 result in a higher number of patients with heart disease, hence we can infer that cp does have correlation with heart disease. Absence of exercise induced angina result in higher number people with heart disease. Number of major blood vessels play a important role, presence of less no. of blood vessels does predict steep increase in patients with heart disease. Also, Blood disorder thalassemia type 2 can lead to heart disease.

Predictive Analysis

After analyzing all the attributes in detail, we use this data set to create a model using logistic regression to predict whether a person has heart disease or not.

1. We use the split function to divide the data into 80% training data and 20% testing data.

```
split <- sample.split(heartData, SplitRatio = 0.8)
trainingSet <- subset(heartData, split == "TRUE")
testingSet <- subset(heartData, split == "FALSE")

print(paste("Data ", nrow(heartData)))
print(paste("Training Data: ", nrow(trainingSet)))
print(paste("Testing Data: ", nrow(testingSet)))
```

2. Before we create the model, we need to munge the data, convert the categorical variables to factor.

```
heartData$sex <- as.factor(heartData$sex)
heartData$cp <- factor(heartData$cp)
heartData$restecg <- factor(heartData$restecg)
heartData$fbs <- factor(heartData$fbs)
heartData$exang <- factor(heartData$exang)
heartData$ca <- factor(heartData$ca)
heartData$thal <- factor(heartData$thal)
heartData$target <- factor(heartData$target)

str(heartData)  ## structure of heartData
```

```
'data.frame': 303 obs. of  14 variables:
 $ age     : int  63 37 41 56 57 57 56 44 52 57 ...
 $ sex     : Factor w/ 2 levels "0","1": 2 2 1 2 1 2 1 2 2 2 ...
 $ cp      : Factor w/ 4 levels "0","1","2","3": 4 3 2 2 1 1 2 2 3 3 ...
 $ trestbps: int  145 130 130 120 120 140 140 120 172 150 ...
 $ chol    : int  233 250 204 236 354 192 294 263 199 168 ...
 $ fbs     : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 2 1 ...
 $ restecg : Factor w/ 3 levels "0","1","2": 1 2 1 2 2 2 1 2 2 2 ...
 $ thalach : int  150 187 172 178 163 148 153 173 162 174 ...
 $ exang   : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 1 1 1 ...
 $ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
 $ slope   : int  0 0 2 2 2 1 1 2 2 2 ...
 $ ca      : Factor w/ 5 levels "0","1","2","3",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ thal    : Factor w/ 4 levels "0","1","2","3": 2 3 3 3 3 2 3 4 4 3 ...
 $ target  : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
```

3. We use glm (general linear model) function to model the data and training set as the input. We first take all the attributes into consideration and created the model.

```
logisticModel <- glm(target ~ age + sex + cp + trestbps + chol + fbs + restecg + thalach + exang + oldpeak + slope + ca + thal, data = trainingSet, family = 'binomial')

summary(logisticModel)
```

```
Call:
glm(formula = target ~ age + sex + cp + trestbps + chol + fbs +
    restecg + thalach + exang + oldpeak + slope + ca + thal,
    family = "binomial", data = trainingSet)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4583  -0.3805   0.1711   0.5854   2.5486

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.4945754  2.9595220   1.181 0.237686
age          0.0001219  0.0265851   0.005 0.996341
sex         -1.7214756  0.5155962  -3.339 0.000841 ***
cp           0.8319555  0.2134252   3.898 9.69e-05 ***
trestbps    -0.0165166  0.0113820  -1.451 0.146747
chol        -0.0094447  0.0049185  -1.920 0.054828 .
fbs         -0.2088238  0.5963395  -0.350 0.726207
restecg      0.0918914  0.3952476   0.232 0.816157
thalach      0.0254869  0.0118256   2.155 0.031143 *
exang       -0.9269533  0.4925628  -1.882 0.059850 .
oldpeak     -0.3687352  0.2353809  -1.567 0.117221
slope        0.7741934  0.3843939   2.014 0.044003 *
ca          -0.8017445  0.2136802  -3.752 0.000175 ***
thal        -0.9531914  0.3337998  -2.856 0.004296 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

4. As we can see there are certain attributes whose p value > 0.05 and those parameters are insignificant in creating the model, we can eliminate them and create a new model using step function.

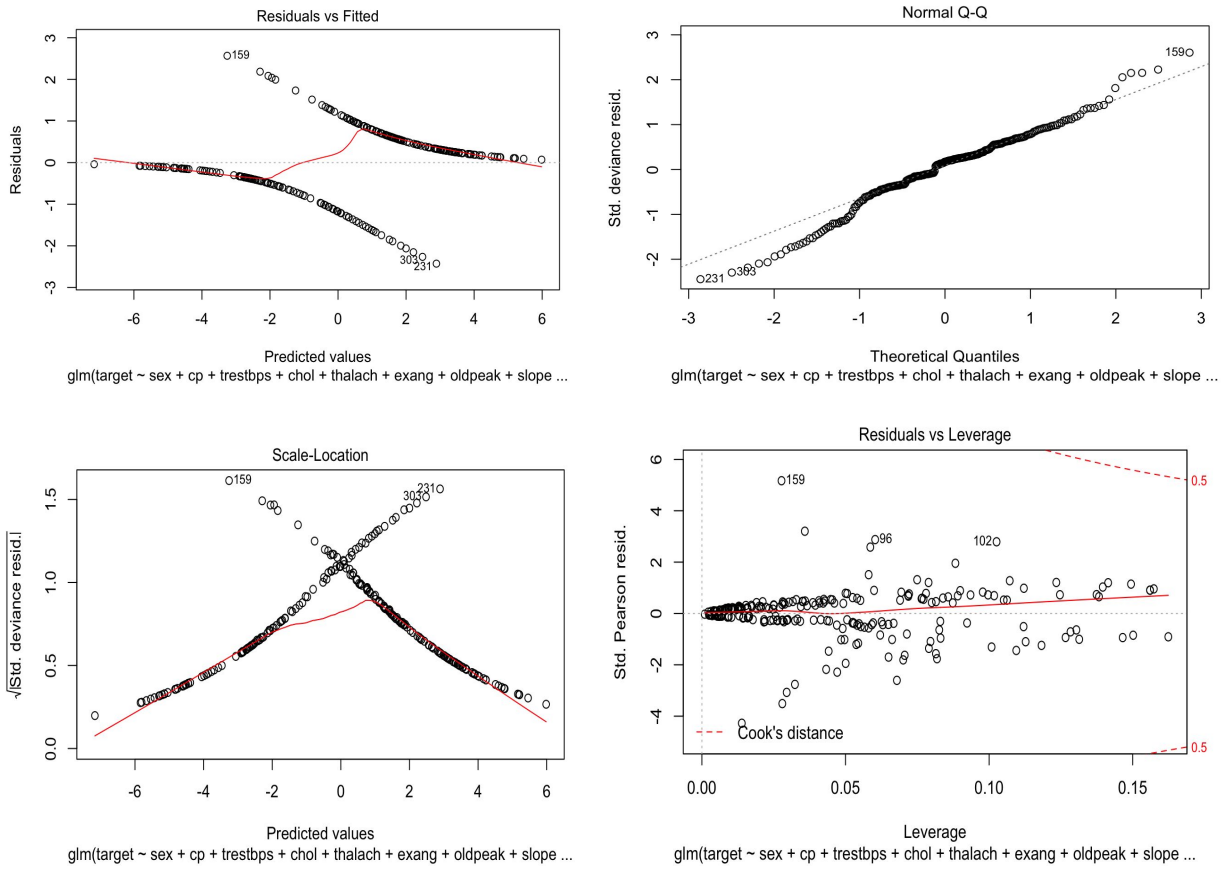
```
newModel <- step(logisticModel, direction = "backward")
summary(new Model)
```

```
Call:
glm(formula = target ~ sex + cp + trestbps + chol + thalach +
    exang + oldpeak + slope + ca + thal, family = "binomial",
    data = trainingSet)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4272  -0.3943   0.1701   0.5774   2.5673
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.612641   2.403945   1.503 0.132891
sex         -1.739158   0.505508  -3.440 0.000581 ***
cp           0.813103   0.205945   3.948 7.88e-05 ***
trestbps    -0.016946   0.011017  -1.538 0.124007
chol        -0.009754   0.004730  -2.062 0.039196 *
thalach      0.025429   0.010971   2.318 0.020455 *
exang       -0.942404   0.487262  -1.934 0.053103 .
oldpeak     -0.364433   0.234478  -1.554 0.120128
slope        0.785171   0.384216   2.044 0.040996 *
ca          -0.804130   0.211612  -3.800 0.000145 ***
thal        -0.921687   0.325484  -2.832 0.004629 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

5. Plot the new model, visual display of regression residuals

```
plot(newModel)
```



6. After the model is created, we test and evaluate the model. We use the testing data to run through the model. As logistic regression is a sigmoid function, if the value is greater than 0.5, we consider it as 1 else 0.

```
result <- predict(newModel, newdata = testingSet, type='response')
result <- ifelse(result > 0.5, 1, 0)
```

7. To validate the model, we compare the actual and predicted results of testing set and find the percentage of the correct result

```
linearPredCorrect <- data.frame(actual=testingSet$target, predicted=result, match =
(testingSet$target == result))

accuracy = length(linearPredCorrect$match[linearPredCorrect$match==TRUE]) /
count(testingSet) * 100

print(paste("Logistic regression has ", round(accuracy, 2), "% accuracy"))
```

```
"Logistic regression has  86.15 % accuracy"
```

We try different combinations of training and testing sets on an average, we got accuracy from 80-90%

Conclusion

After analysing all the attributes in detail, we can infer that the attributes such as age, sex, cholesterol, higher sugar level, high BP does not show much correlation with heart disease.

However, the attributes like chest pain type, number of blood vessels, max heart rate, exercise induced angina, exercise induced ST depression (in treadmill ECG test), thalassemia blood disorder lead to more probability of heart disease.

The model created using logistic regression can also be used to predict the probability of having having a heart disease.

References

https://www.kaggle.com/ronitf/heart-disease-uci
https://www.cdc.gov/heartdisease/facts.htm
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4958709/
https://www.sciencedirect.com/science/article/abs/pii/0002870386902656