To save the public

HEART DISEASE PREDICTION

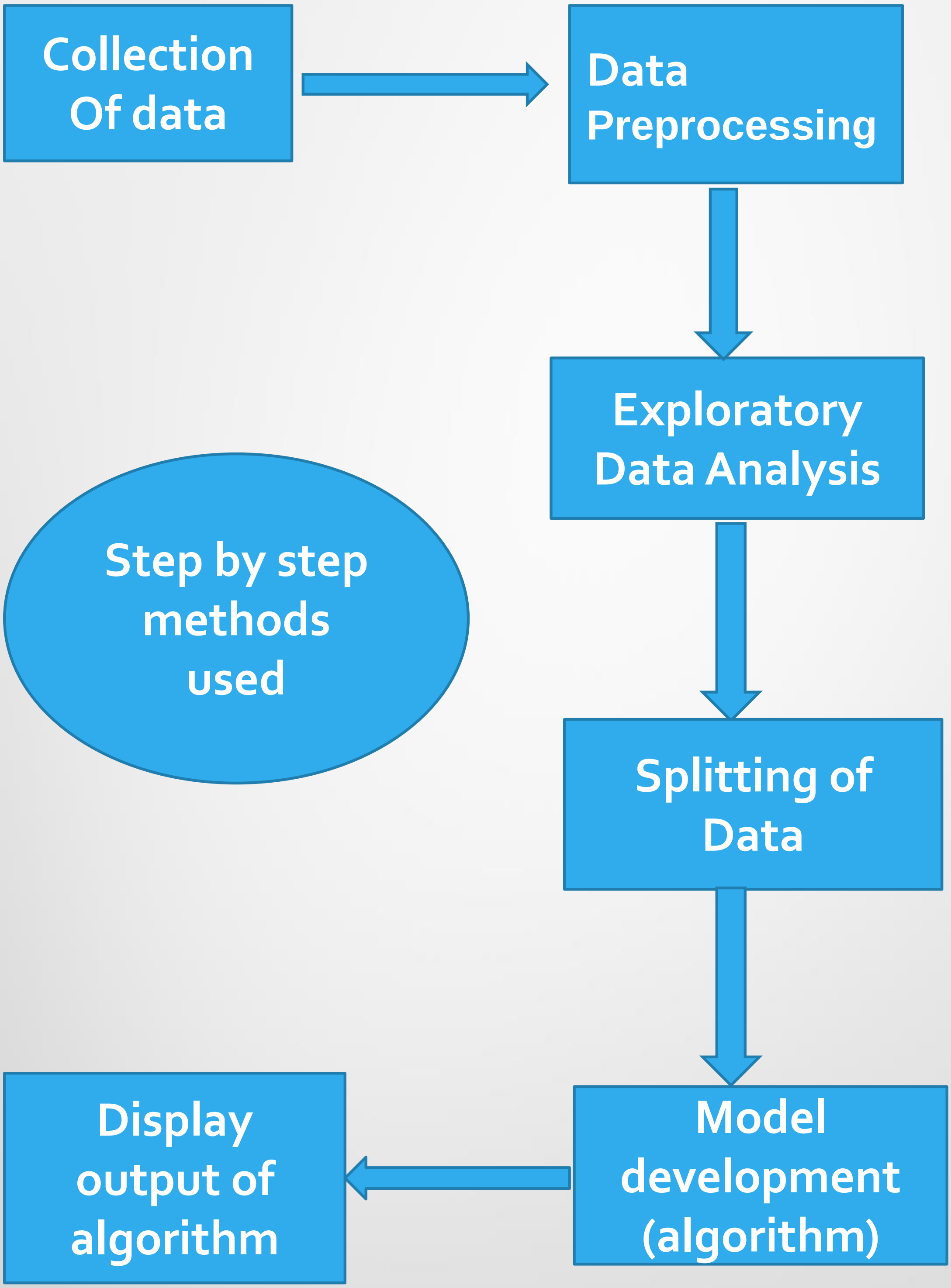
Analysis of Heart disease Using Model

Prepared by
Karthik.M (20-UST-021)
Akilan A (20-ust-018)
Tony Abinash
(20-UST-005)
Thirunavuk Arasu
(20-UST-061)
Akash (20-UST-075)

17 November 2022

TABLE OF CONTENTS

Methodology



Methodology

1. Collection of data

The process of gathering and analyzing accurate data from various sources to find answers to research problems, trends and probabilities, etc., to evaluate possible outcomes is Known as Data Collection. It includes understanding the data to study the hidden patterns and trends which helps to predict and evaluate the results

2. Data Preprocessing

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format. The data contain **missing Data, noisy Data, repeated data, etc.**
We can process the data using some respective data preprocessing method like **Ignore the tuples, Fill the Missing values, Regression, Clustering, ect.**

3. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary approach to analyze the data using visual and graphical representations. Also help us to understand the relationship between variables and show us maximize insight into a data set, uncover underlying structure, extract important variables, detect outliers and anomalies, test underlying assumptions, develop parsimonious models, and determine optimal factor settings.

Methodology

4. Splitting of Data

After data cleaning and pre-processing, the dataset becomes ready to train and test. In the train/split method, we split the dataset randomly into the training and testing set. For Training, we took 80% of the sample and for testing, we took 20% of the sample. It is called as cross validation

```
> #create split object
> train_test_split <-data %>% initial_split(prop = .8, strata = "Diagnosis_Heart_Disease")
> #pipe split obj to training() fcn to create training set
> train_tbl <- train_test_split %>% training()
> #pipe split obj to testing() fcn to create test set
> test_tbl <- train_test_split %>% testing()
> nrow(train_tbl)
[1] 236
> nrow(test_tbl)
[1] 60
```

5. Model development

Model development is an iterative process, in which many models are derived, tested and built upon until a model fitting the desired criteria is built and model are build through different algorithms. The model which performs best that algorithm will preferred for the same type of data for which the model built.

6. Display output of model

From the results of the models we can conclude which model is best to go with according to the Accuracy, Sensitivity, Specificity ,etc.

Introduction

- In this article, we will be closely working with the heart disease prediction and for that, we will be looking into the heart disease dataset from that dataset we will derive various insights that help us know the weightage of each feature and how they are interrelated to each other but this time our sole aim is to detect the probability of person that will be affected by a savior heart problem or not.
- I would like to explain the various data analysis operation, I have done on this data set and how to conclude or heart disease predict status of patients who undergone from surgery.
- First of all for any data analysis task or for performing operation on data we should have good domain knowledge so that we can relate the data features and also can give accurate conclusion. So, I would like to explain the features of data set and how it affects other feature.

Data Set Explanations

Initially, the dataset contains 76 features or attributes from 303 patients; however, published studies chose only 14 features that are relevant in predicting heart disease. Hence, here we will be using the dataset consisting of 303 patients with 14 features set.

Introduction

Attribute Information

1. **Age**: age of the patient [years].
2. **Sex**: sex of the patient [1: Male, 0: Female].
3. **Chest Pain Type**: chest pain type

[0 = Typical Angina, 1 = Atypical Angina,
2 = Non-Angina Pain, 3 = Asymptomatic] .
Angina pain is often described as squeezing, pressure, heaviness, tightness or pain in the chest.

4. **Resting BP**: resting blood pressure in mm Hg.
5. **Serum Cholesterol**: Serum cholesterol in mg/dl A person's serum cholesterol level represents the amount of total cholesterol in their blood.
6. **Fasting Blood Sugar**: Fasting blood sugar level relative to 120 mg/dl:[0 = fasting blood sugar <= 120 mg/dl,
   1 = fasting blood sugar > 120 mg/dl].
7. **Resting ECG**: Resting electrocardiographic results
   [0 = normal,
   1 = ST-T wave abnormality,
   2 = left ventricle hyperthrophy]
   An electrocardiogram records the electrical signals in the heart.
8. **Max Heart Rate Achieved**: Max heart rate of subject.
9. **Exercise Induced Angina**:[0 = no 1 = yes]
   Angina tends to appear during physical activity, emotional stress, or exposure to cold temperatures, or after big meals.

Introduction

10. ST Depression Induced by Exercise Relative to Rest:
ST Depression of subject.

11. Peak Exercise ST Segment:
[0 = Down-sloaping,
1 Flat,
2 = Up-sloaping,]

The ST segment shift relative to exercise-induced increments in heart rate, the ST/heart rate slope.

12. Number of Major Vessels (0-3) Visible on Flouroscopy:
Number of visible vessels under flouro.

13. Thalassemia : Form of thalassemia: [3]
[1 = fixed defect,
2 = normal,
3 = reversible defect]
Thalassemia is an inherited (i.e., passed from parents to children through genes) blood disorder caused when the body doesn't make enough of a protein called hemoglobin, an important part of red blood cells.

14. **Diagnosis of Heart Disease**: Indicates whether subject is suffering from heart disease or not:
[0= heart disease absence
1= heart disease present].

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a pre-processing step to understand the data. There are numerous methods and steps in performing EDA, however, most of them are specific, focusing on either visualization or distribution, and are incomplete. Therefore, here, I will walk-through step-by-step to understand, explore, and extract the information from the data to answer the questions or assumptions. There are no structured steps or method to follow, however, this project will provide an insight on EDA for you and my future self.

Operations

I had used R for this purpose as it has the rich collection of machine learning libraries and mathematical operation. I will mostly use common packages as **tidyverse, *scales*, *ROCR*, *gmodels*** *and* ***tidymodels*** which help me for mathematical operations and also plotting, importing and exporting of files.

```
> library(tidyverse)
> library(scales)
> library(gmodels)
> library(tidymodels)
> library(ROCR)
```

Exploratory Data Analysis

1. Import and get to know the data

```
> data<-read.csv(file.choose())
> head(data)
  age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal target
1  63   1  3      145  233   1       0     150     0     2.3     0  0    1      1
2  37   1  2      130  250   0       1     187     0     3.5     0  0    2      1
3  41   0  1      130  204   0       0     172     0     1.4     2  0    2      1
4  56   1  1      120  236   0       1     178     0     0.8     2  0    2      1
5  57   0  0      120  354   0       1     163     1     0.6     2  0    2      1
6  57   1  0      140  192   0       1     148     0     0.4     1  0    1      1
> dim(data)
[1] 303  14
> names <- c("Age",
+            "Sex",
+            "Chest_Pain_Type",
+            "Resting_Blood_Pressure",
+            "Serum_Cholesterol",
+            "Fasting_Blood_Sugar",
+            "Resting_ECG",
+            "Max_Heart_Rate_Achieved",
+            "Exercise_Induced_Angina",
+            "ST_Depression_Exercise",
+            "Peak_Exercise_ST_Segment",
+            "Num_Major_Vessels_Flouro",
+            "Thalassemia",
+            "Diagnosis_Heart_Disease")
> colnames(data) <- names
```

Here we have 303 rows with 14 variables and giving correct column names.

2. Data Preprocessing

The variables types are

•Binary: Sex, Fasting Blood Sugar, Exercise Induced Angina, Diagnosis Heart Disease.

•Categorical: Chest Pain Type, Resting ECG, Peak Exercise ST Segment , Num Major Vessels Flouro, Thalassemia.

•Continuous: Age, Resting Blood Pressure, Serum Cholesterol, Max Heart Rate Achieved, ST Depression Exercise.

Exploratory Data Analysis

```
> which(is.na(data))
integer(0)
> data %>%
+   drop_na() %>%
+   group_by(Thalassemia) %>%
+   count()
# A tibble: 4 × 2
# Groups:   Thalassemia [4]
  Thalassemia     n
        <int> <int>
1           0     2
2           1    18
3           2   166
4           3   117
>
> data %>%
+   drop_na() %>%
+   group_by(Num_Major_Vessels_Flouro) %>%
+   count()
# A tibble: 5 × 2
# Groups:   Num_Major_Vessels_Flouro [5]
  Num_Major_Vessels_Flouro     n
                     <int> <int>
1                        0   175
2                        1    65
3                        2    38
4                        3    20
5                        4     5
> data<-data %>%
+     mutate_at(c("Resting_ECG",
+                 "Fasting_Blood_Sugar",
+                 "Sex",
+                 "Diagnosis_Heart_Disease",
+                 "Exercise_Induced_Angina",
+                 "Peak_Exercise_ST_Segment",
+                 "Chest_Pain_Type",
+ "Thalassemia"), as_factor)%>%
+ filter(Thalassemia != 0) %>%
+ filter(Num_Major_Vessels_Flouro != 4) %>%
+     select(Age,
+            Resting_Blood_Pressure,
+            Serum_Cholesterol,
+            Max_Heart_Rate_Achieved,
+            ST_Depression_Exercise,
+            Num_Major_Vessels_Flouro,
+            everything())
```

The variables Num Major Vessels Flouroand and Thalassemia have unwanted variables, so i removed it and changing categorical into factor.

Exploratory Data Analysis

```
> data<-data %>%
+ mutate(Diagnosis_Heart_Disease= recode_factor(Diagnosis_Heart_Disease,
+  `0` = "absent", `1` = "present" ),
+ Sex = recode_factor(Sex, `0` = "female",`1` = "male" ),
+ Chest_Pain_Type = recode_factor(Chest_Pain_Type,
+ `0` = "typical",`1` = "atypical",`2` = "non-angina",`3` = "asymptomatic"),
+ Fasting_Blood_Sugar = recode_factor(Fasting_Blood_Sugar,
+ `0` = "<= 120 mg/dl",`1` = "> 120 mg/dl"),
+ Resting_ECG = recode_factor(Resting_ECG, `0` = "normal",
+ `1` = "ST-T abnormality",`2` = "LV hypertrophy"),
+ Exercise_Induced_Angina = recode_factor(Exercise_Induced_Angina,
+ `0` = "no",`1` = "yes"),
+ Peak_Exercise_ST_Segment = recode_factor(Peak_Exercise_ST_Segment,
+ `2` = "up-sloaping",`1` = "flat",`0` = "down-sloaping"),
+ Thalassemia = recode_factor(Thalassemia, `2` = "normal",
+ `1` = "fixed defect",`3` = "reversible defect")) %>%
+ select(Age,
+            Resting_Blood_Pressure,
+            Serum_Cholesterol,
+            Max_Heart_Rate_Achieved,
+            ST_Depression_Exercise,
+            Num_Major_Vessels_Flouro,
+            everything())
> glimpse(data)
Rows: 296
Columns: 14
$ Age                      <int> 63, 37, 41, 56, 57, 57, 56, 44, 52, 57, 54, 48, 4
$ Resting_Blood_Pressure   <int> 145, 130, 130, 120, 120, 140, 140, 120, 172, 150,
$ Serum_Cholesterol        <int> 233, 250, 204, 236, 354, 192, 294, 263, 199, 168,
$ Max_Heart_Rate_Achieved  <int> 150, 187, 172, 178, 163, 148, 153, 173, 162, 174,
$ ST_Depression_Exercise   <dbl> 2.3, 3.5, 1.4, 0.8, 0.6, 0.4, 1.3, 0.0, 0.5, 1.6,
$ Num_Major_Vessels_Flouro <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
$ Sex                      <fct> male, male, female, male, female, male, female, m
$ Chest_Pain_Type          <fct> asymptomatic, non-angina, atypical, atypical, typ
$ Fasting_Blood_Sugar      <fct> > 120 mg/dl, <= 120 mg/dl, <= 120 mg/dl, <= 120 m
$ Resting_ECG              <fct> normal, ST-T abnormality, normal, ST-T abnormalit
$ Exercise_Induced_Angina  <fct> no, no, no, no, yes, no, no, no, no, no, no, no,
$ Peak_Exercise_ST_Segment <fct> down-sloaping, down-sloaping, up-sloaping, up-slo
$ Thalassemia              <fct> fixed defect, normal, normal, normal, normal, fix
$ Diagnosis_Heart_Disease  <fct> present, present, present, present, present, pres
```

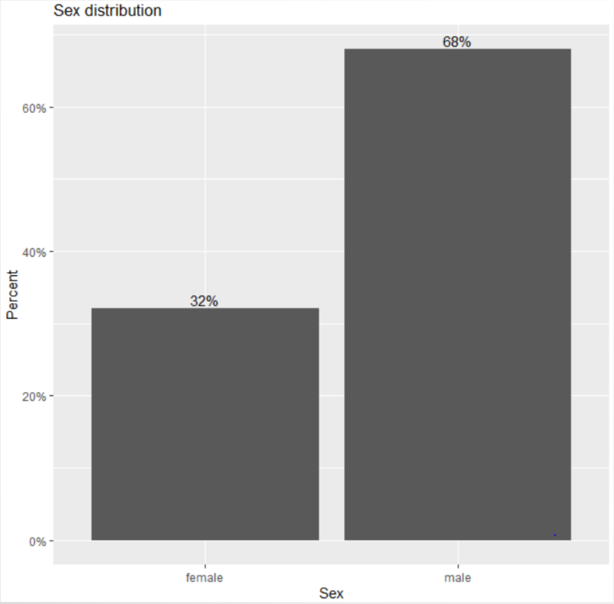
Changing dummy categorical variables values into respected names and understanding the data by **glimpse** function.

Exploratory Data Analysis

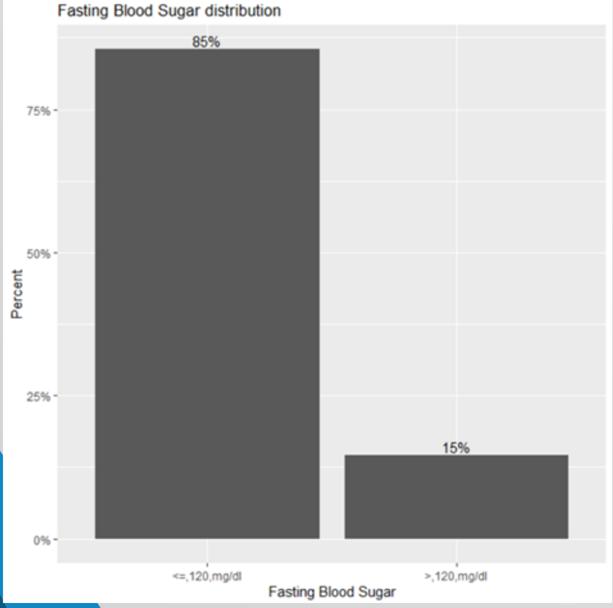
3. Univariate Analysis

Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable. It doesn’t deal with causes or relationships (unlike regression) and it’s major purpose is to describe; It takes data, summarizes that data and finds patterns in the data.

a) Categorical variables

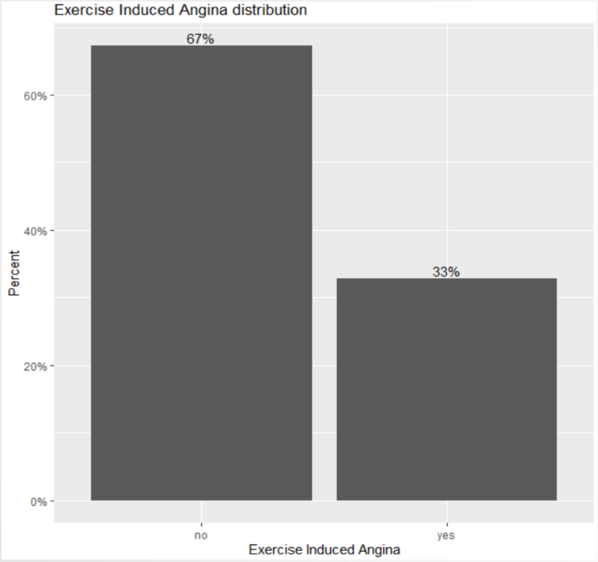


Sex variable is distributed into two categories Male and female and there respectively percentage are 68% and 32%. We can see that males are more than females.

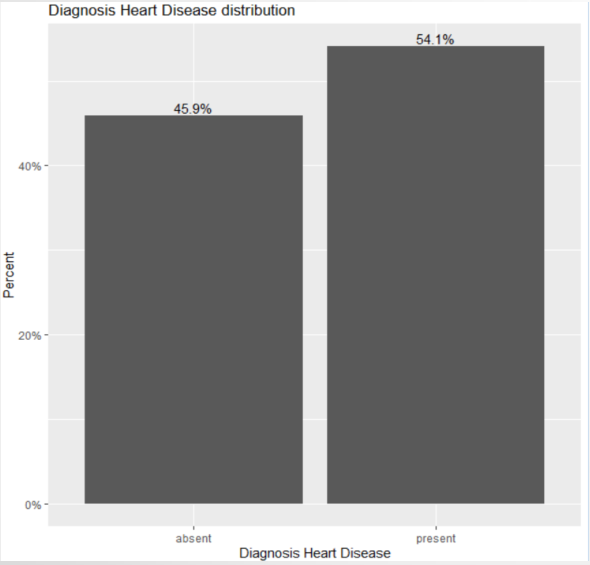


Fasting blood sugar is distributed into two categories <=120mg/dl and>120mg/dl and there respectively percentage are 85% and 15%. We can see that <=120mg/dl are more than >120mg/dl.

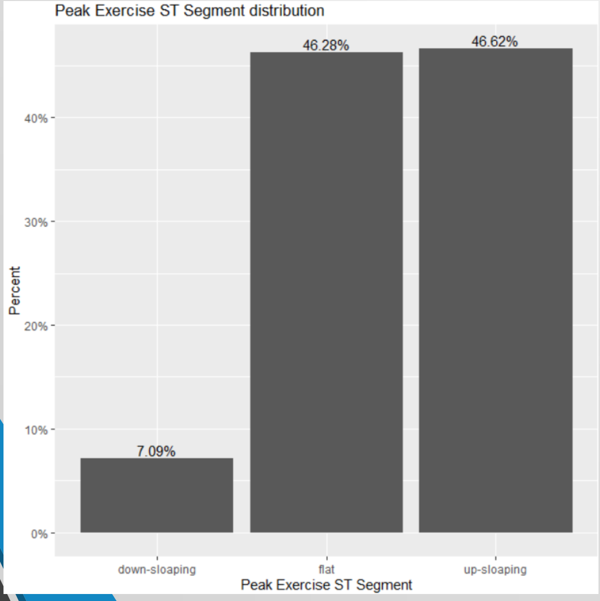
Exploratory Data Analysis



Exercise Induced angina is distributed into two groups are "yes" and "no" and there respectively percentage are 67% and 33%. The data says that majority of patients did not have Exercise Induced Angina.

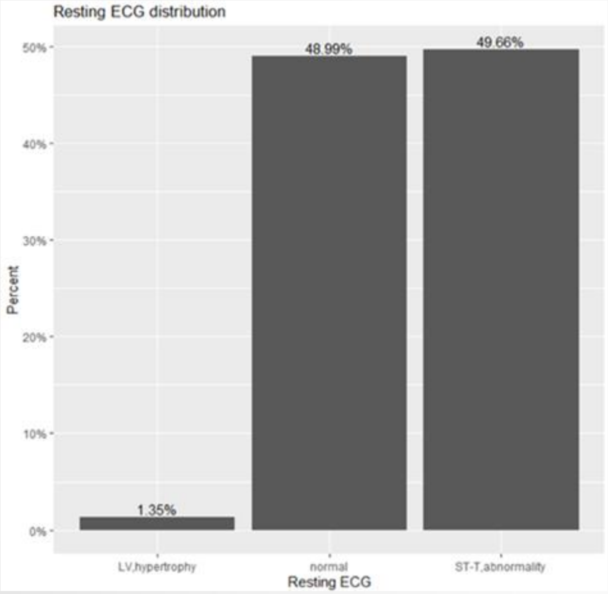


Diagnosis heart disease distribution into two groups are "Absent "and "present "and there respectively percentage are 45.9% and 54.1% .The data says that majority of them are present in Diagnosis heart disease.

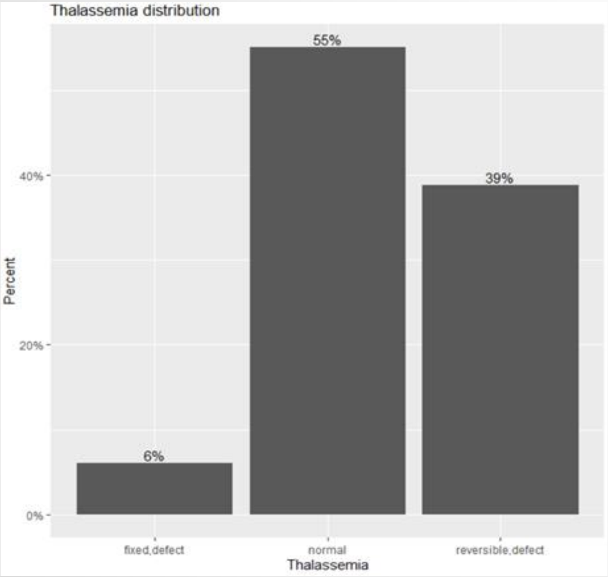


Peak exercise ST segment is distributed into three categories :Down sloaping has 7.09 Flat has 46.28Up sloping has 46.62aNow we know that up sloping is peaked up than Down sloaping and flat.

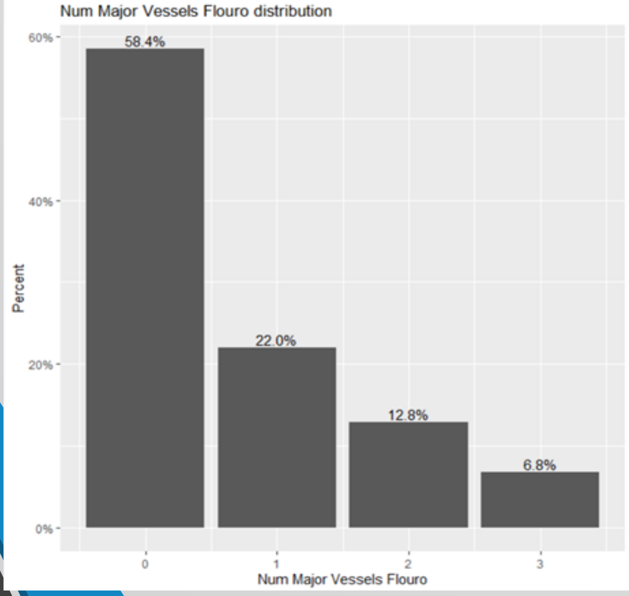
Exploratory Data Analysis



Resting_ECG are grouped into three LV hypertrophy, normal and ST-T abnorarmality and there respective percentage are 1.35%, 48.99% and 49.66% . The bar says that majority of them are in normal ST-T abnorarmality

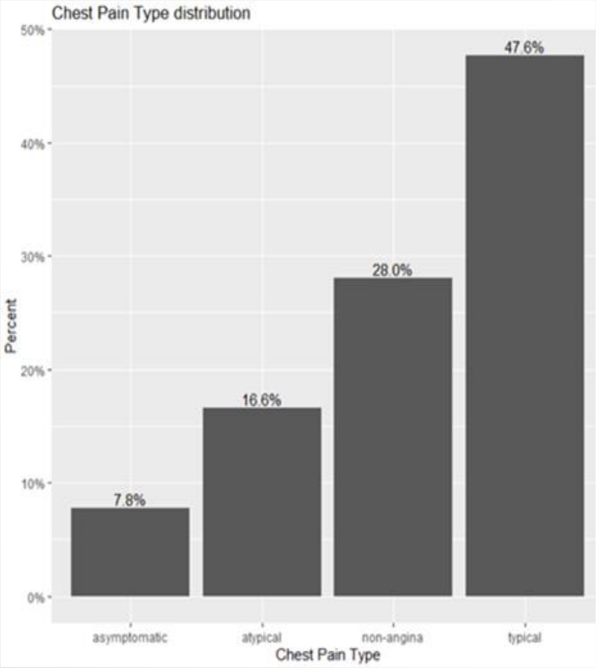


Thalassemia variable is distributed into three groups : fixed defect has 6% and normal has 55% and reversible has 38%. -Normal has higher percentage than fixed and reversible



Num major vessels flourois distributed into four types 0, 1, 2 and 3 and there respective percentage are 58.4%, 22.0%, 12.8%and 6.8% .The data says that majority of them are present in 0 type.

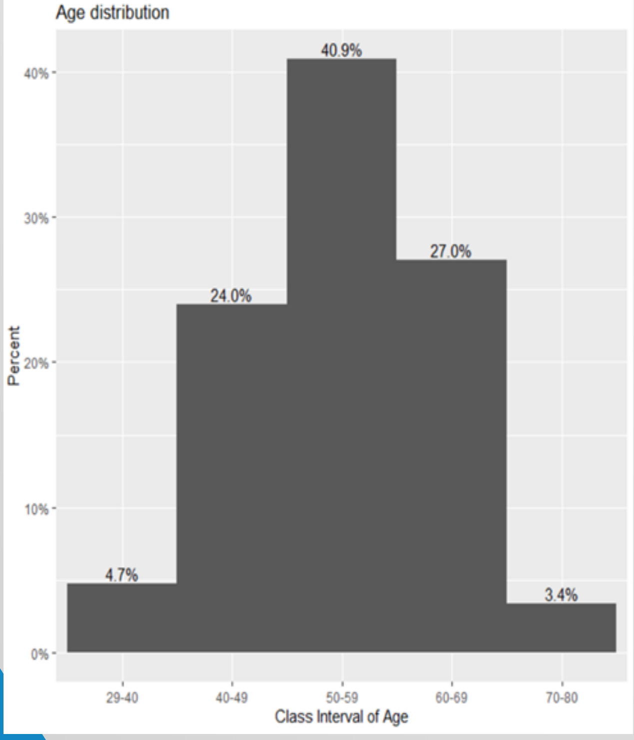
Exploratory Data Analysis



Chest pain is distributed into 4 types asymptomatic, atypical, non-agina and typical and there respective percentage are 7.8%, 16.6%, 28.0% and 47.6%. The data says that majority of them are present in typical chest pain type

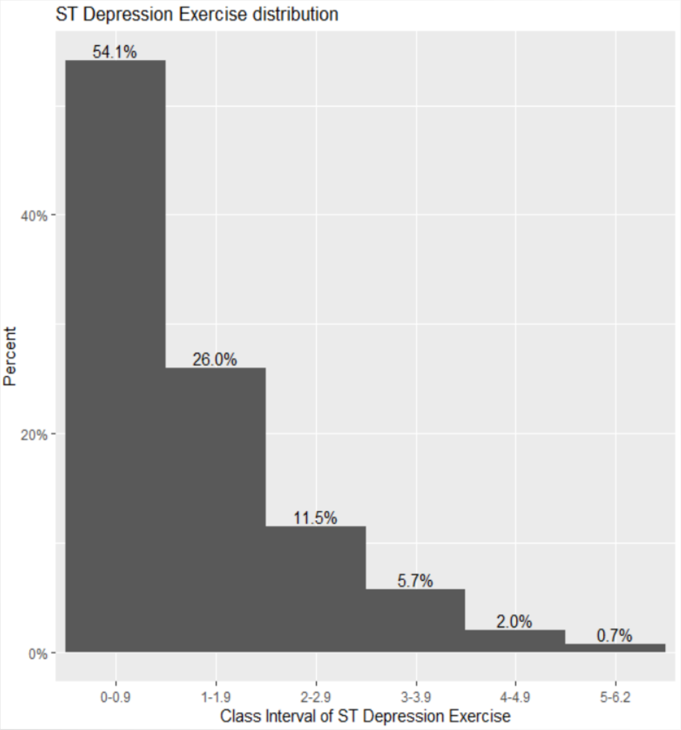
b) Continuous variables

Aggregating the new variable in class interval from the old continuous variables according to their values.

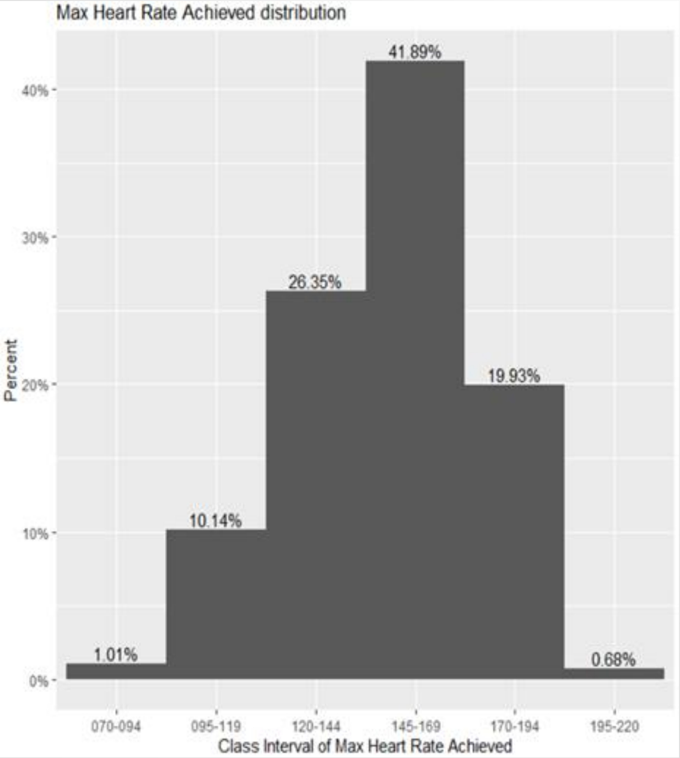


Patients age group are diverted into different class interval: 29-40 = 4.7%,
40-49 = 24.0%,
50-59 = 40.9%,
60-69 = 27.0% and
70-80 = 3.4%.
The age group between 50-59 are having highest patients.

Exploratory Data Analysis

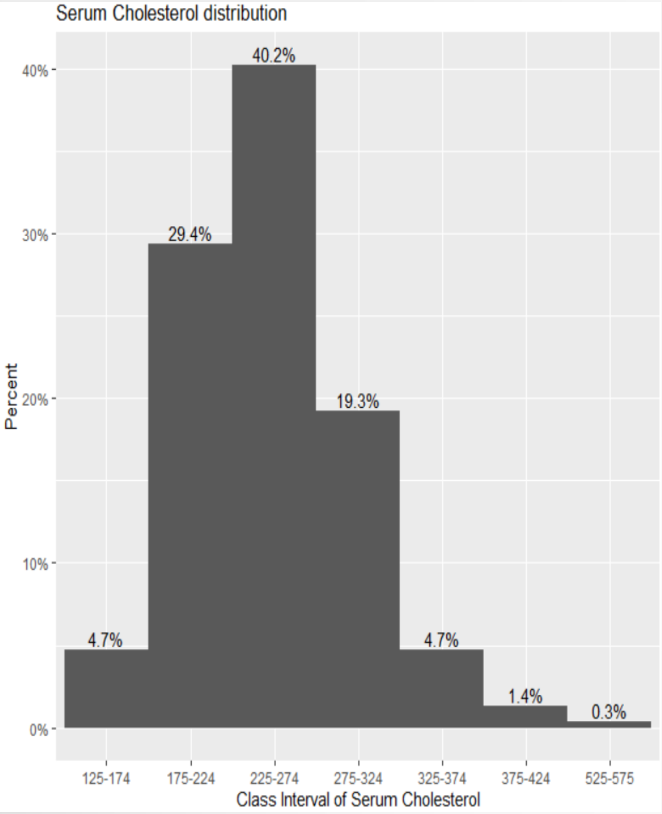


ST depression patients are grouped into different class interval.
0-0.9=54.1%, 11.9=20.0%, 2-2.9=11.5%, 3-3.9=5.7%, 4-4.9=2%,and
5-6.2=0.7%.
The group between 0-0.9 are having highest patients.

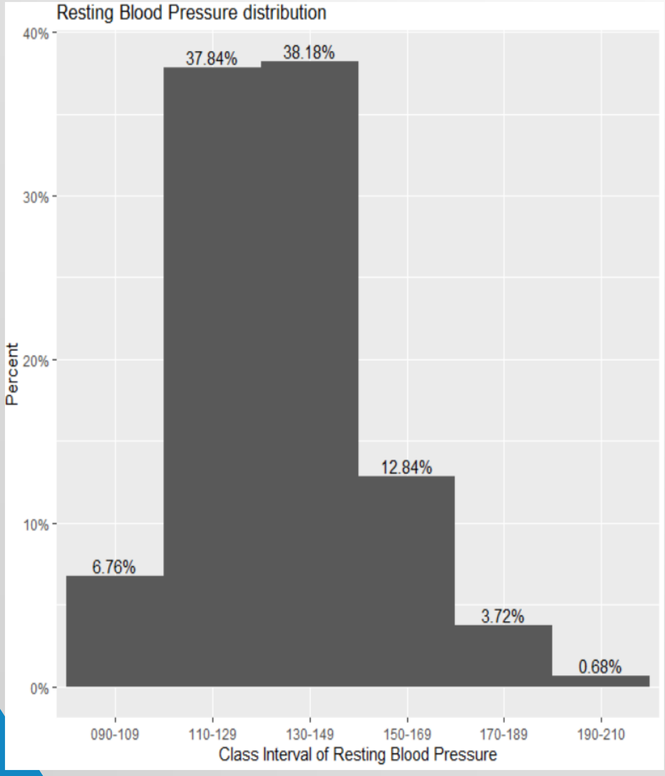


Max heart test are grouped into different class interval
070-094=1.01%,
095-119=10.14%,
120-144=26.35%,
145-169=41.89% ,
170-194=19.93% and
195-220=0.68%.
The group between 145-169 are having highest patients.

Exploratory Data Analysis



Serum cholesterol are grouped into different class interval
125-174=4.7%,
175-224=29.4%,
229-274=40.2%,
275-324=19.3%,
325-374=4.7%,
376-425=1.4%,
525-575=0.3%.
The group between 229-274 are having highest patients



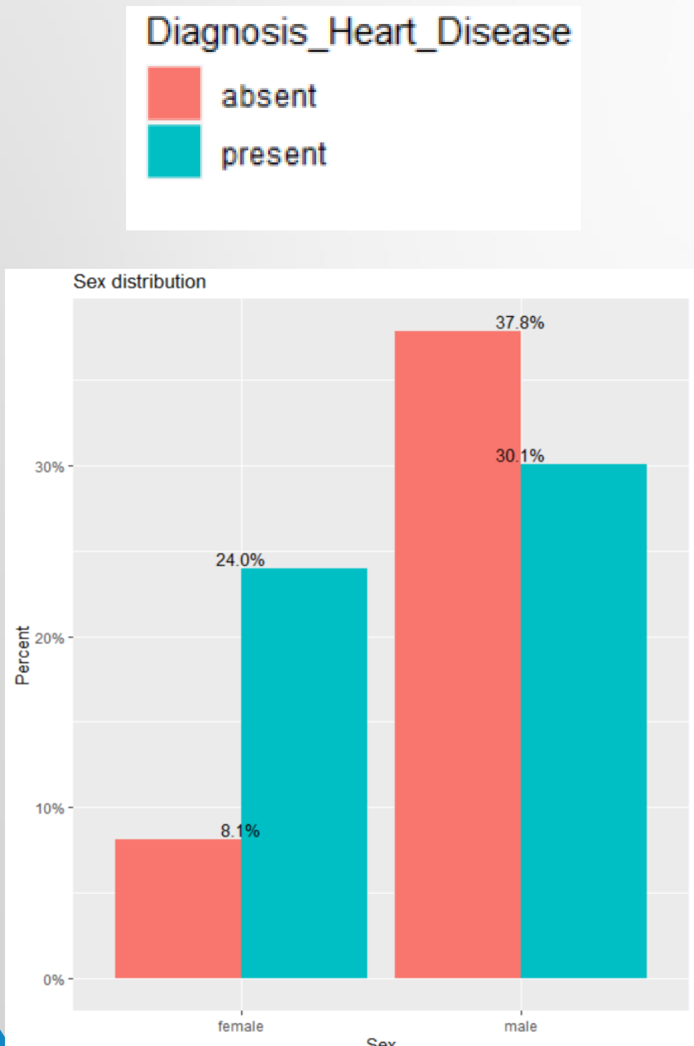
Resting blood pressure are grouped into different class interval
090-109=6.76%,
110-129=37.84%,
130-149=38.18%
150-169=12.84%,
170-189=3.72%
190-210=0.68%.
The group between 110-129,130-149 has highest percentage

Exploratory Data Analysis

4. Bivariate Analysis

Bivariate analysis is one of the statistical analysis where two variables are observed. One variable here is dependent while the other is independent. These variables are usually denoted by X and Y. So, here we analyse the changes occured between the two variables and to what extent.

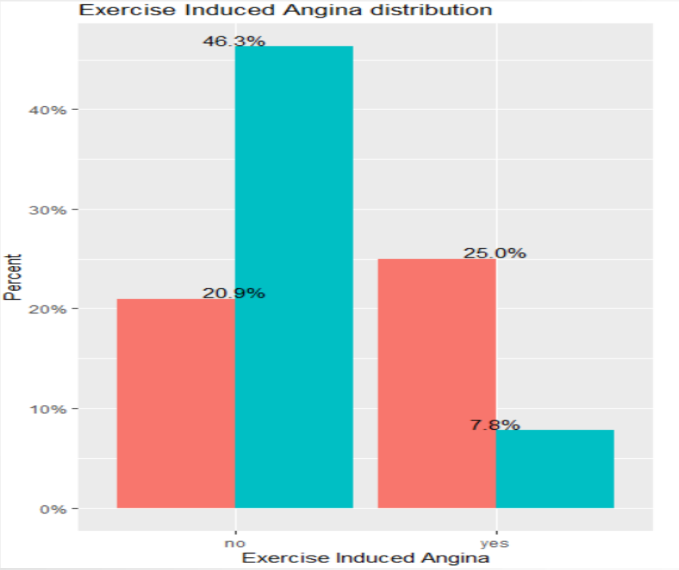
a) Categorical variables



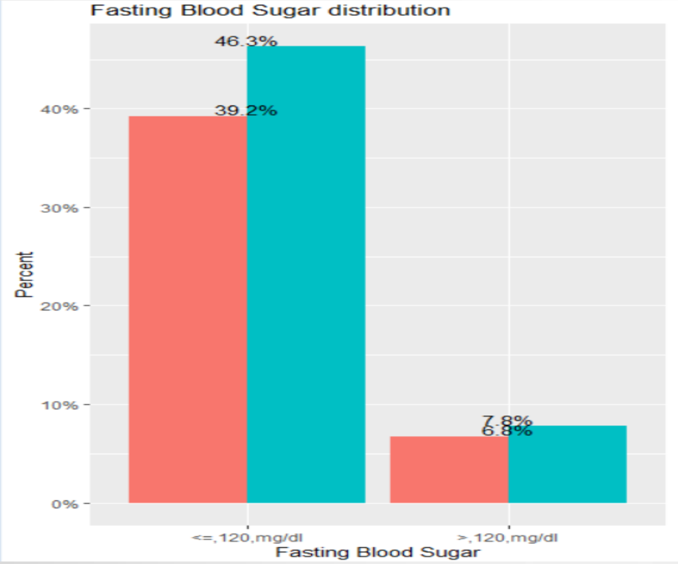
The p-value is too small and since it is less than 0.05 so we can conclude that gender of the patients is a significant variable.

p-value = 1.719e-06

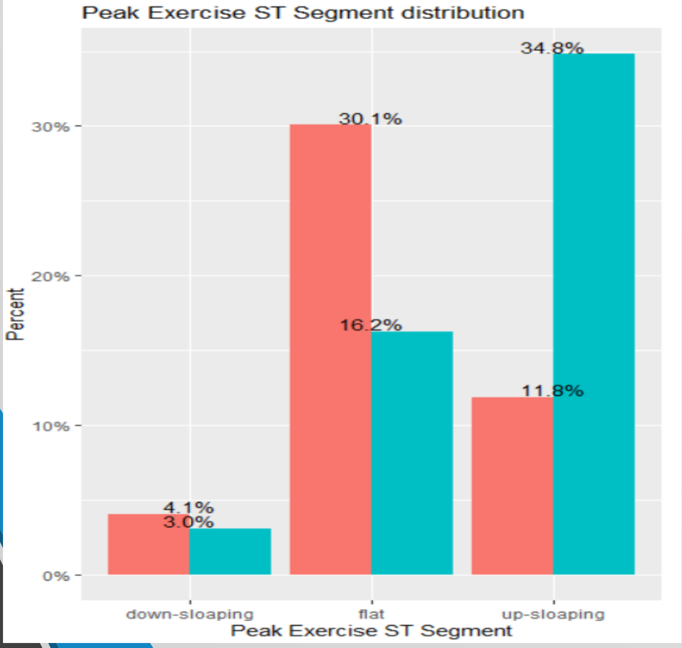
Exploratory Data Analysis



The p-value is too small and since it is less than 0.05 so we can conclude that Exercise Induced Angina of the patients is a significant variable.
p-value = 6.517e-13

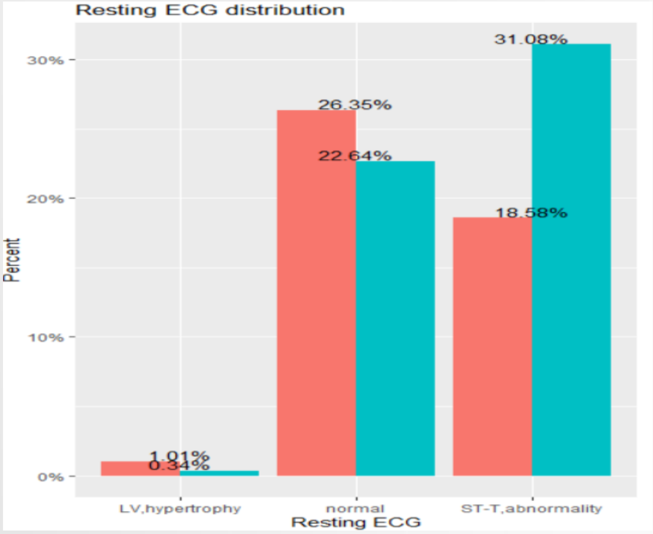


The p-value is large and since it is more than 0.05 so we can conclude that Fasting Blood Sugar of the patients is not a significant variable.
p-value = 1

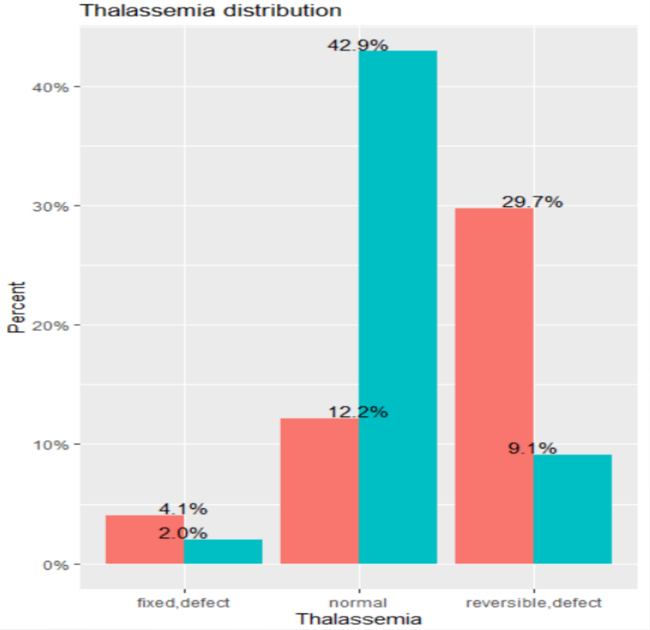


The p-value is too small and since it is less than 0.05 so we can conclude that Peak Exercise ST Segment of the patients is a significant variable.
p-value = 2.116e-10

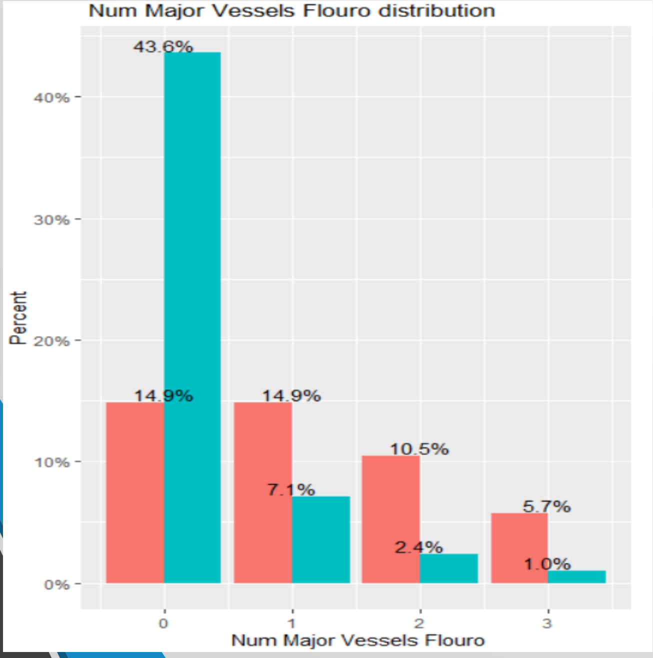
Exploratory Data Analysis



The p-value is too small and since it is less than 0.05 so we can conclude that Resting ECG of the patients is a significant variable.
p-value = 0.009743

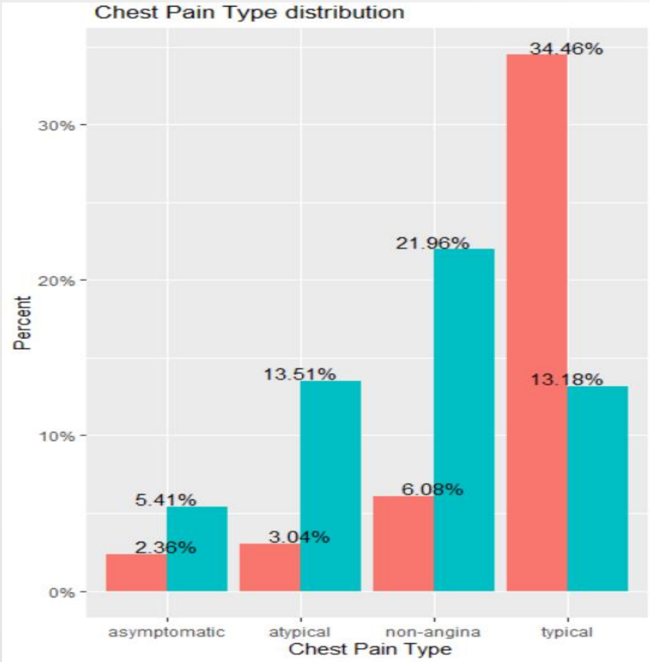


The p-value is too small and since it is less than 0.05 so we can conclude that Thalassemia of the patients is a significant variable.
p-value = 2.2e-16



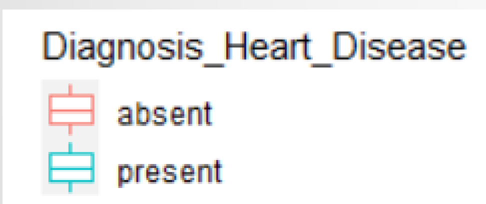
The p-value is too small and since it is less than 0.05 so we can conclude that Num Major Vessels Flouro of the patients is a significant variable.
p-value = 7.996e-16

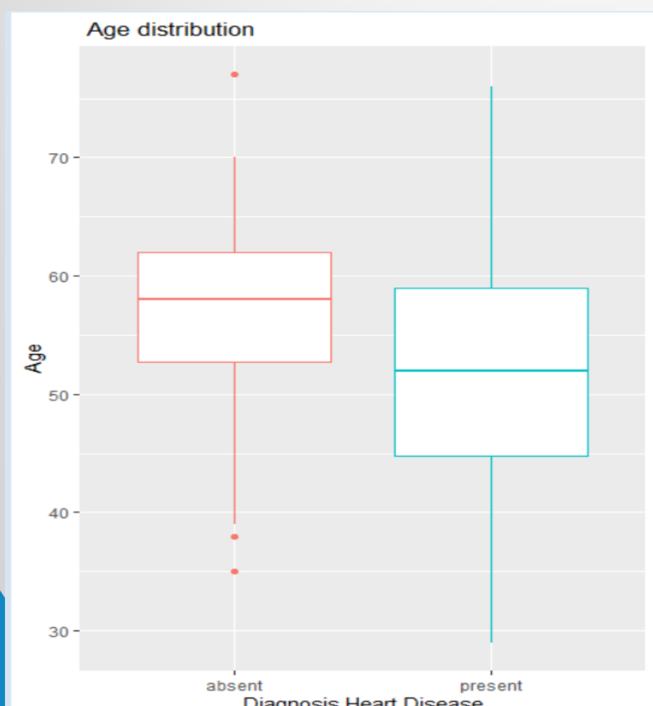
Exploratory Data Analysis



The p-value is too small and since it is less than 0.05 so we can conclude that Chest Pain Type of the patients is a significant variable.
p-value = 2.2e-16

b) Continuous variables





absent

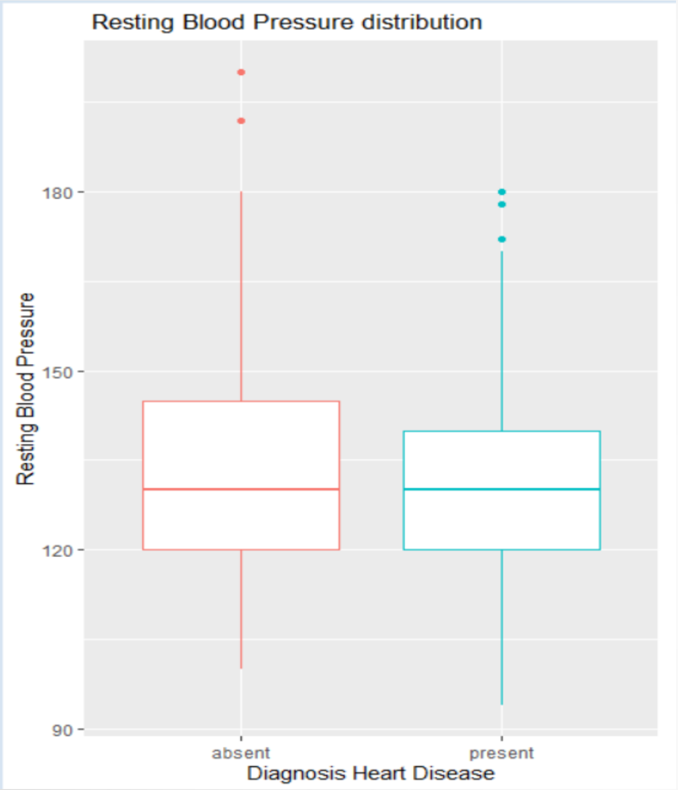
1st Qu.	Median	Mean	3rd Qu.
52.75	58.00	56.74	62.00

present

1st Qu.	Median	Mean	3rd Qu.
44.75	52.00	52.64	59.00

The p-value is too small and since it is less than 0.05 so we can conclude that Age of the patients is a significant variable.
p-value = 7.17e-05.There are very few outliers in the data.

Exploratory Data Analysis



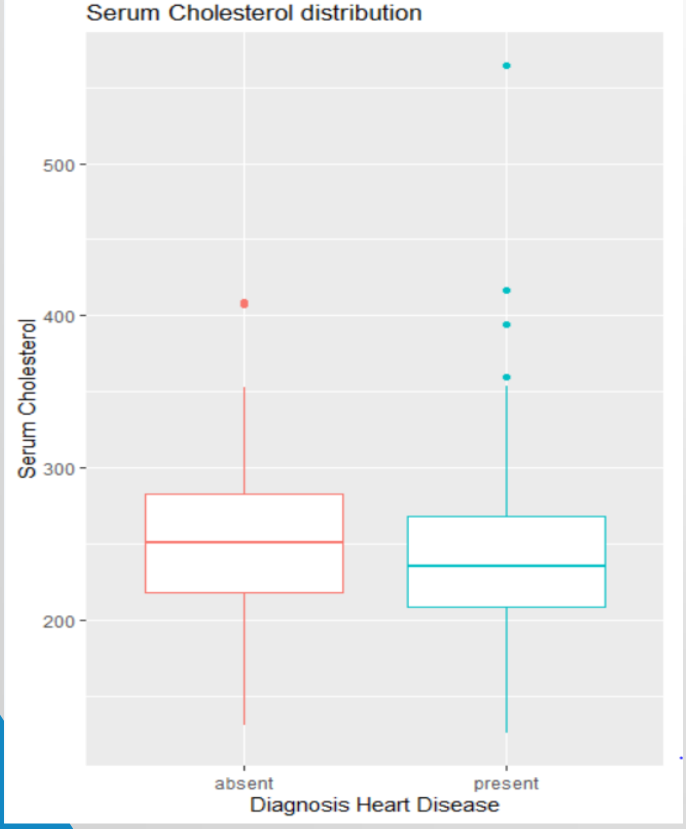
absent

1st Qu.	Median	Mean	3rd Qu.
120.0	130.0	134.5	145.0

present

1st Qu.	Median	Mean	3rd Qu.
120.0	130.0	129.2	140.0

The p-value is too small and since it is less than 0.05 so we can conclude that Resting Blood Pressure of the patients is a significant variable. p-value =0.01123.There are very few outliers in the data.



absent

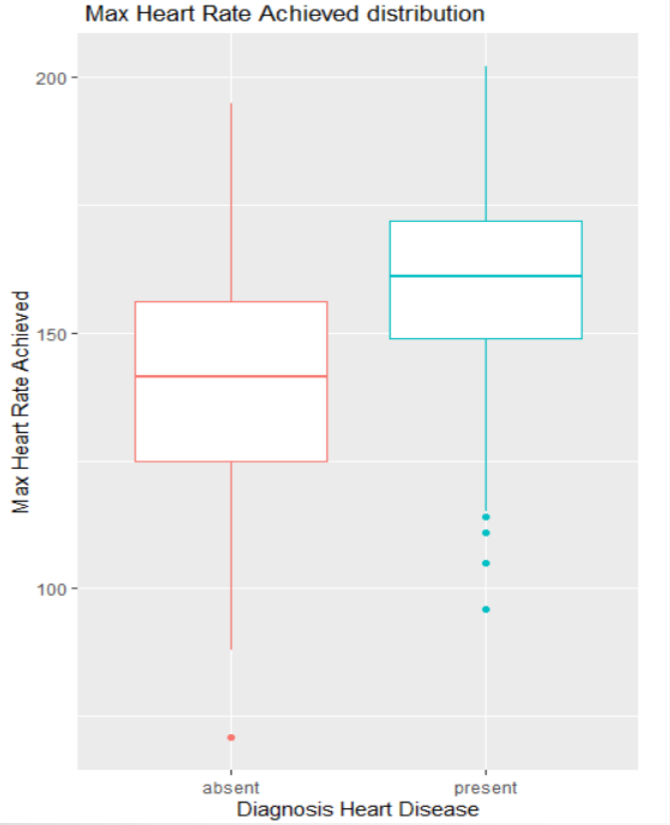
1st Qu.	Median	Mean	3rd Qu.
217.8	251.0	251.5	283.2

present

1st Qu.	Median	Mean	3rd Qu.
208.8	235.5	243.5	268.2

The p-value is large and since it is more than 0.05 so we can conclude that Serum Cholesterol of the patients is not a significant variable. p-value =0.1863.There are very few outliers in the data.

Exploratory Data Analysis



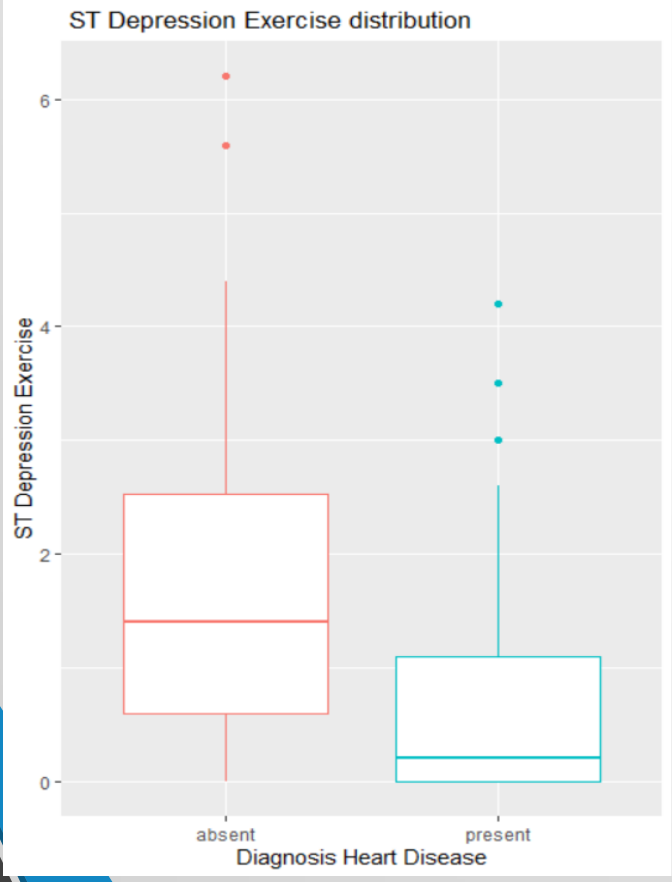
absent

1st Qu.	Median	Mean	3rd Qu.
125.0	141.5	138.9	156.2

present

1st Qu.	Median	Mean	3rd Qu.
149.0	161.0	158.6	172.0

The p-value is too small and since it is less than 0.05 so we can conclude that Max Heart Rate Achieved of the patients is a significant variable.
p-value =4.628e-14. There are very few outliers in the data.



absent

1st Qu.	Median	Mean	3rd Qu.
0.600	1.400	1.601	2.525

present

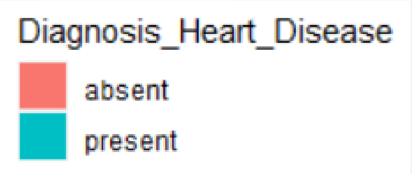
1st Qu.	Median	Mean	3rd Qu.
0.0000	0.2000	0.5988	1.1000

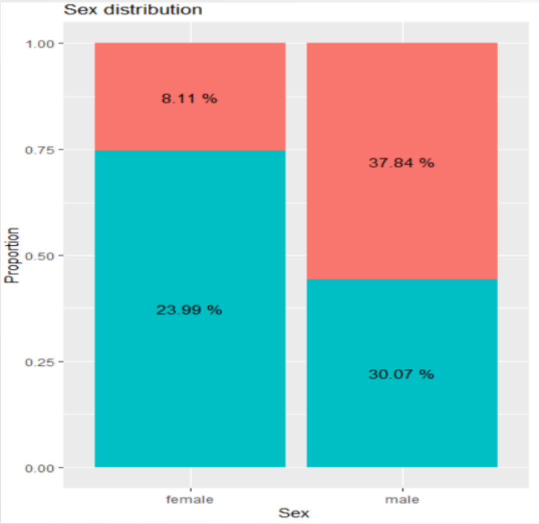
The p-value is too small and since it is less than 0.05 so we can conclude that ST Depression Exercise of the patients is a significant variable.
p-value =2.139e-13. There are very few outliers in the data.

Exploratory Data Analysis

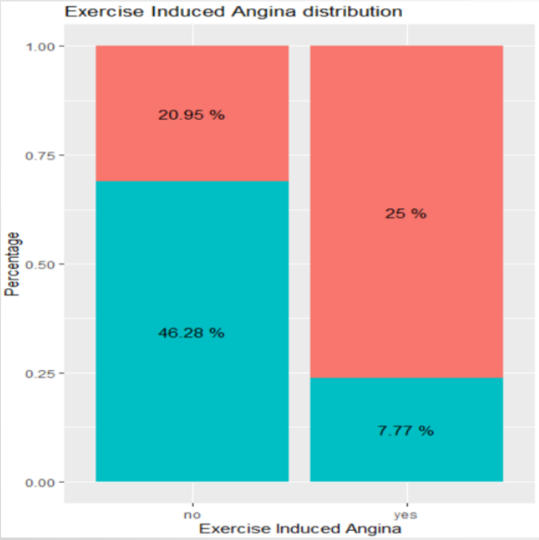
5. Cross Tabulation

a) Categorical variables

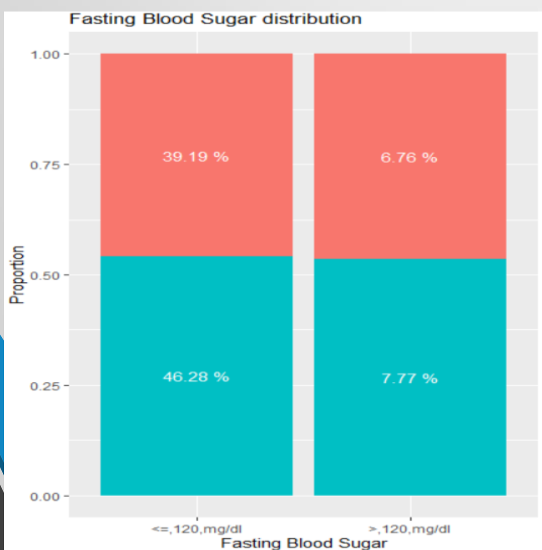




data$Sex	data$Diagnosis_Heart_Disease absent	present	Row Total
female	24	71	95
	8.845	7.518	
	0.253	0.747	0.321
	0.176	0.444	
	0.081	0.240	
male	112	89	201
	4.180	3.553	
	0.557	0.443	0.679
	0.824	0.556	
	0.378	0.301	
Column Total	136	160	296
	0.459	0.541	

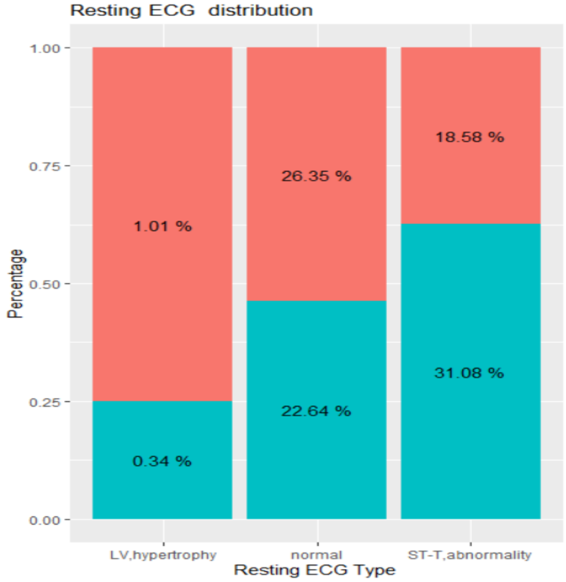


data$Exercise_Induced_Angina	data$Diagnosis_Heart_Disease absent	present	Row Total
no	62	137	199
	9.474	8.053	
	0.312	0.688	0.672
	0.456	0.856	
	0.209	0.463	
yes	74	23	97
	19.437	16.522	
	0.763	0.237	0.328
	0.544	0.144	
	0.250	0.078	
Column Total	136	160	296
	0.459	0.541	

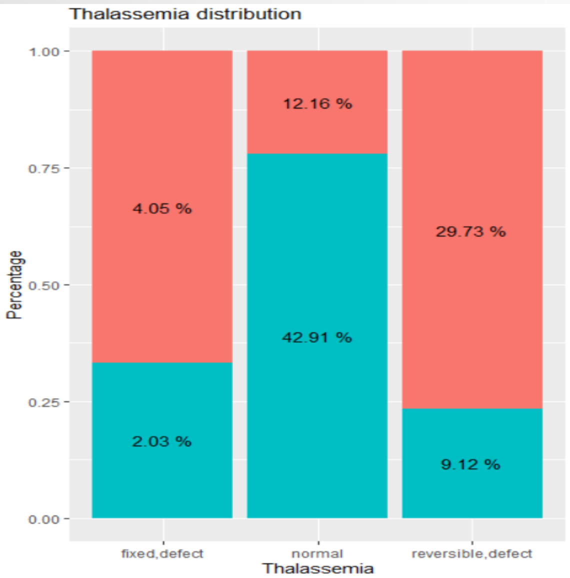


data$Fasting_Blood_Sugar	data$Diagnosis_Heart_Disease absent	present	Row Total
<=,120,mg/dl	116	137	253
	0.001	0.000	
	0.458	0.542	0.855
	0.853	0.856	
	0.392	0.463	
>,120,mg/dl	20	23	43
	0.003	0.003	
	0.465	0.535	0.145
	0.147	0.144	
	0.068	0.078	
Column Total	136	160	296
	0.459	0.541	

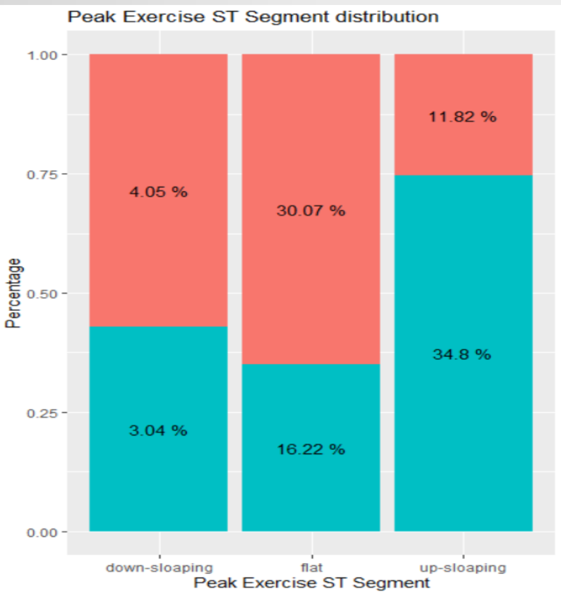
Exploratory Data Analysis




```
                  | data$Diagnosis_Heart_Disease
data$Resting_ECG  |    absent |   present | Row Total |
------------------|-----------|-----------|-----------|
   LV,hypertrophy |         3 |         1 |         4 |
                  |     0.735 |     0.625 |           |
                  |     0.750 |     0.250 |     0.014 |
                  |     0.022 |     0.006 |           |
                  |     0.010 |     0.003 |           |
------------------|-----------|-----------|-----------|
           normal |        78 |        67 |       145 |
                  |     1.943 |     1.652 |           |
                  |     0.538 |     0.462 |     0.490 |
                  |     0.574 |     0.419 |           |
                  |     0.264 |     0.226 |           |
------------------|-----------|-----------|-----------|
ST-T,abnormality  |        55 |        92 |       147 |
                  |     2.328 |     1.979 |           |
                  |     0.374 |     0.626 |     0.497 |
                  |     0.404 |     0.575 |           |
                  |     0.186 |     0.311 |           |
------------------|-----------|-----------|-----------|
     Column Total |       136 |       160 |       296 |
                  |     0.459 |     0.541 |           |
------------------|-----------|-----------|-----------|
```

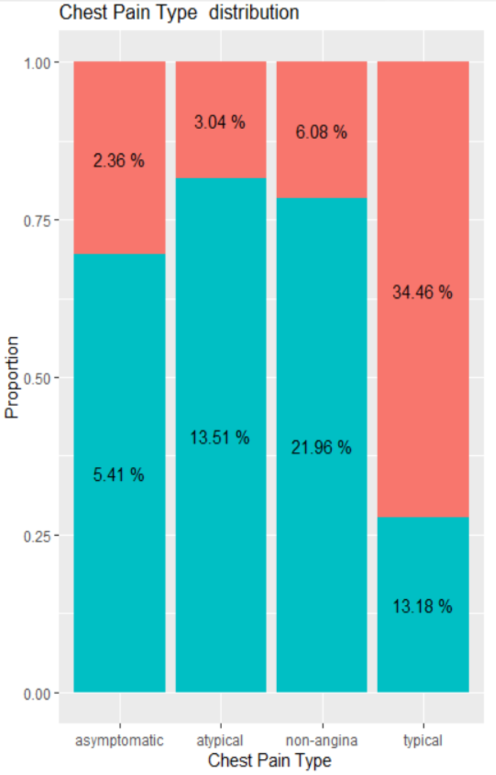



```
                   | data$Diagnosis_Heart_Disease
  data$Thalassemia |    absent |   present | Row Total |
-------------------|-----------|-----------|-----------|
      fixed,defect |        12 |         6 |        18 |
                   |     1.682 |     1.430 |           |
                   |     0.667 |     0.333 |     0.061 |
                   |     0.088 |     0.037 |           |
                   |     0.041 |     0.020 |           |
-------------------|-----------|-----------|-----------|
            normal |        36 |       127 |       163 |
                   |    20.197 |    17.167 |           |
                   |     0.221 |     0.779 |     0.551 |
                   |     0.265 |     0.794 |           |
                   |     0.122 |     0.429 |           |
-------------------|-----------|-----------|-----------|
 reversible,defect |        88 |        27 |       115 |
                   |    23.399 |    19.890 |           |
                   |     0.765 |     0.235 |     0.389 |
                   |     0.647 |     0.169 |           |
                   |     0.297 |     0.091 |           |
-------------------|-----------|-----------|-----------|
```

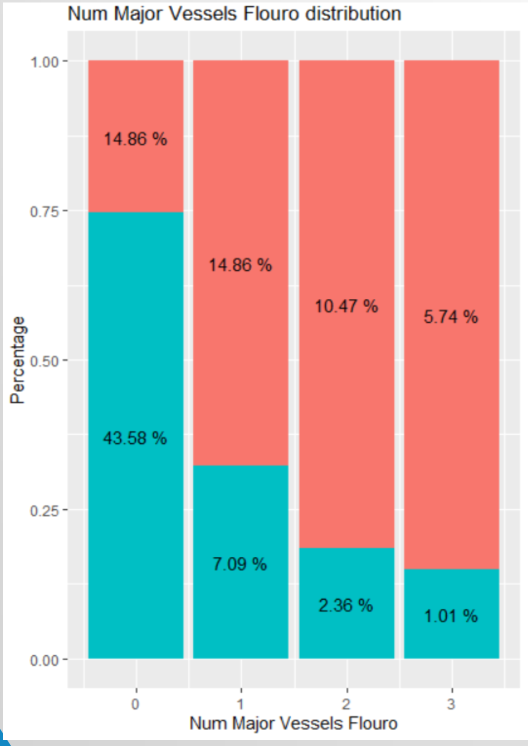



```
                              | data$Diagnosis_Heart_Disease
data$Peak_Exercise_ST_Segment |    absent |   present | Row Total |
------------------------------|-----------|-----------|-----------|
                 down-sloaping |        12 |         9 |        21 |
                              |     0.573 |     0.487 |           |
                              |     0.571 |     0.429 |     0.071 |
                              |     0.088 |     0.056 |           |
                              |     0.041 |     0.030 |           |
------------------------------|-----------|-----------|-----------|
                          flat |        89 |        48 |       137 |
                              |    10.784 |     9.166 |           |
                              |     0.650 |     0.350 |     0.463 |
                              |     0.654 |     0.300 |           |
                              |     0.301 |     0.162 |           |
------------------------------|-----------|-----------|-----------|
                   up-sloaping |        35 |       103 |       138 |
                              |    12.726 |    10.817 |           |
                              |     0.254 |     0.746 |     0.466 |
                              |     0.257 |     0.644 |           |
                              |     0.118 |     0.348 |           |
------------------------------|-----------|-----------|-----------|
                  Column Total |       136 |       160 |       296 |
                              |     0.459 |     0.541 |           |
------------------------------|-----------|-----------|-----------|
```

Exploratory Data Analysis



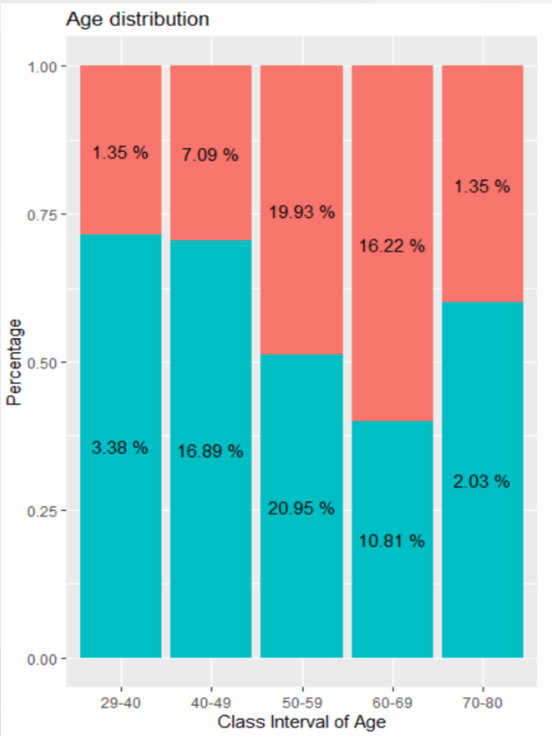
data$Chest_Pain_Type	data$Diagnosis_Heart_Disease absent	present	Row Total
asymptomatic	7	16	23
	1.204	1.024	
	0.304	0.696	0.078
	0.051	0.100	
	0.024	0.054	
atypical	9	40	49
	8.111	6.895	
	0.184	0.816	0.166
	0.066	0.250	
	0.030	0.135	
non-angina	18	65	83
	10.631	9.037	
	0.217	0.783	0.280
	0.132	0.406	
	0.061	0.220	
typical	102	39	141
	21.380	18.173	
	0.723	0.277	0.476
	0.750	0.244	
	0.345	0.132	
Column Total	136	160	296
	0.459	0.541	



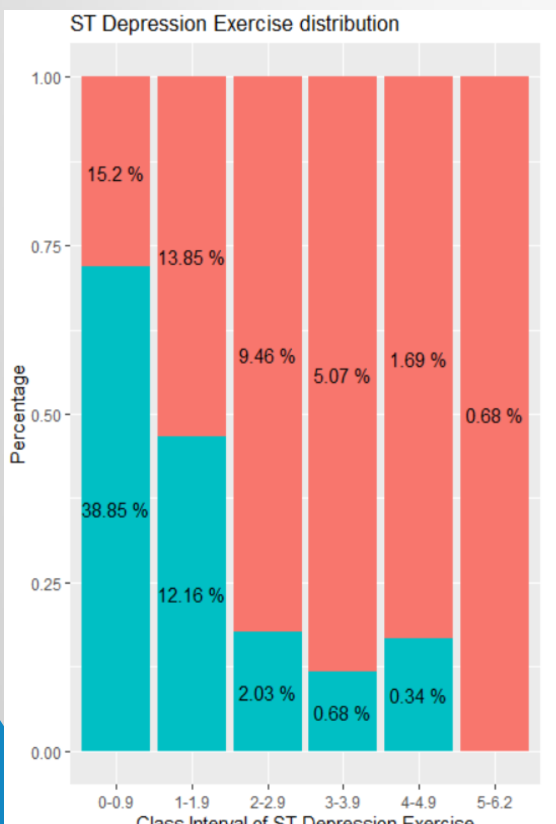
a$Num_Major_Vessels_Flouro	data$Diagnosis_Heart_Disease absent	present	Row Total
0	44	129	173
	15.843	13.466	
	0.254	0.746	0.584
	0.324	0.806	
	0.149	0.436	
1	44	21	65
	6.690	5.687	
	0.677	0.323	0.220
	0.324	0.131	
	0.149	0.071	
2	31	7	38
	10.501	8.926	
	0.816	0.184	0.128
	0.228	0.044	
	0.105	0.024	
3	17	3	20
	6.639	5.643	
	0.850	0.150	0.068
	0.125	0.019	
	0.057	0.010	
Column Total	136	160	296
	0.459	0.541	

Exploratory Data Analysis

b) Continuous variables

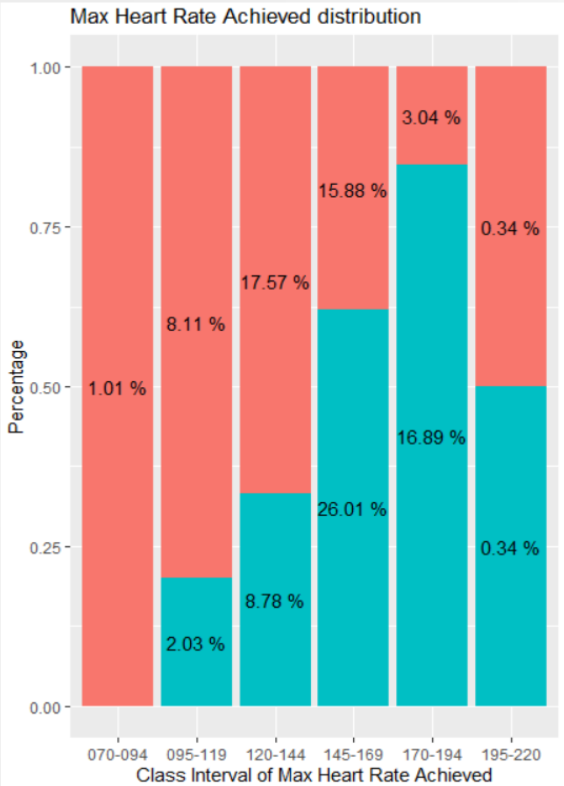


data$CI_Age	data$Diagnosis_Heart_Disease absent	present	Row Total
29-40	4	10	14
	0.920	0.782	
	0.286	0.714	0.047
	0.029	0.062	
	0.014	0.034	
40-49	21	50	71
	4.140	3.519	
	0.296	0.704	0.240
	0.154	0.312	
	0.071	0.169	
50-59	59	62	121
	0.209	0.177	
	0.488	0.512	0.409
	0.434	0.388	
	0.199	0.209	
60-69	48	32	80
	3.439	2.923	
	0.600	0.400	0.270
	0.353	0.200	
	0.162	0.108	
70-80	4	6	10
	0.077	0.065	
	0.400	0.600	0.034
	0.029	0.037	
	0.014	0.020	
Column Total	136	160	296
	0.459	0.541	

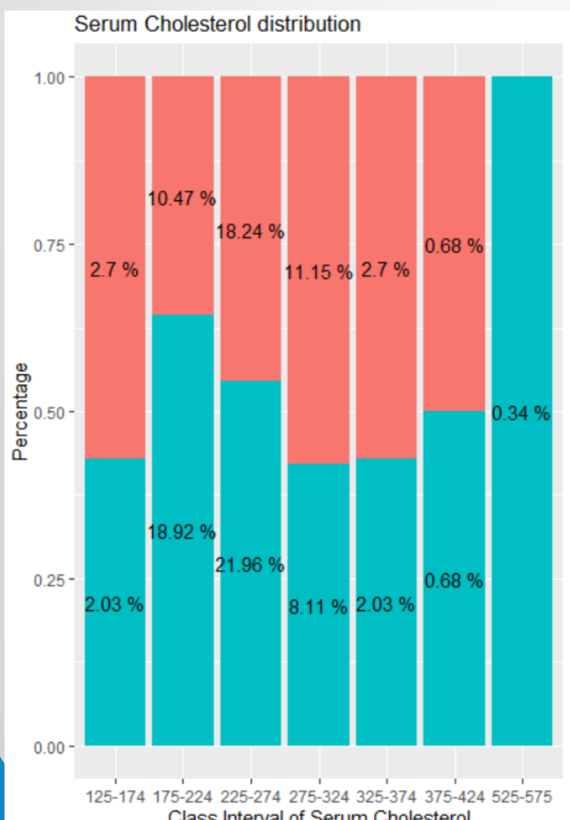


data$CI_STDE	data$Diagnosis_Heart_Disease absent	present	Row Total
0-0.9	45	115	160
	11.059	9.401	
	0.281	0.719	0.541
	0.331	0.719	
	0.152	0.389	
1-1.9	41	36	77
	0.893	0.759	
	0.532	0.468	0.260
	0.301	0.225	
	0.139	0.122	
2-2.9	28	6	34
	9.808	8.337	
	0.824	0.176	0.115
	0.206	0.037	
	0.095	0.020	
3-3.9	15	2	17
	6.617	5.624	
	0.882	0.118	0.057
	0.110	0.012	
	0.051	0.007	
4-4.9	5	1	6
	1.825	1.552	
	0.833	0.167	0.020
	0.037	0.006	
	0.017	0.003	
5-6.2	2	0	2
	1.272	1.081	
	1.000	0.000	0.007
	0.015	0.000	
	0.007	0.000	
Column Total	136	160	296
	0.459	0.541	

Exploratory Data Analysis

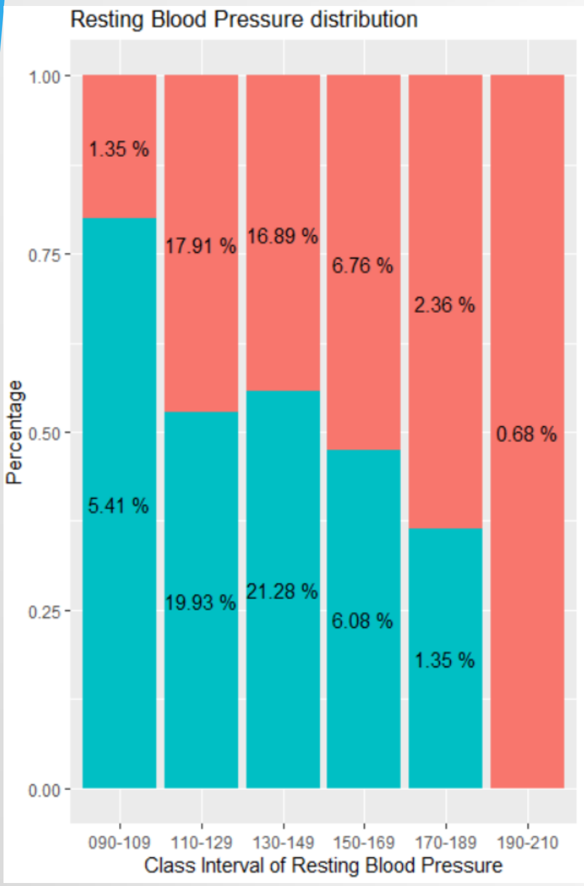


data$CI_MHRA	data$Diagnosis_Heart_Disease absent	present	Row Total
070-094	3	0	3
	1.908	1.622	
	1.000	0.000	0.010
	0.022	0.000	
	0.010	0.000	
095-119	24	6	30
	7.572	6.436	
	0.800	0.200	0.101
	0.176	0.037	
	0.081	0.020	
120-144	52	26	78
	7.289	6.195	
	0.667	0.333	0.264
	0.382	0.163	
	0.176	0.088	
145-169	47	77	124
	1.746	1.484	
	0.379	0.621	0.419
	0.346	0.481	
	0.159	0.260	
170-194	9	50	59
	12.096	10.282	
	0.153	0.847	0.199
	0.066	0.312	
	0.030	0.169	
195-220	1	1	2
	0.007	0.006	
	0.500	0.500	0.007
	0.007	0.006	
	0.003	0.003	
Column Total	136	160	296
	0.459	0.541	



data$CI_SC	data$Diagnosis_Heart_Disease absent	present	Row Total
125-174	8	6	14
	0.382	0.325	
	0.571	0.429	0.047
	0.059	0.037	
	0.027	0.020	
175-224	31	56	87
	2.014	1.712	
	0.356	0.644	0.294
	0.228	0.350	
	0.105	0.189	
225-274	54	65	119
	0.008	0.007	
	0.454	0.546	0.402
	0.397	0.406	
	0.182	0.220	
275-324	33	24	57
	1.771	1.506	
	0.579	0.421	0.193
	0.243	0.150	
	0.111	0.081	
325-374	8	6	14
	0.382	0.325	
	0.571	0.429	0.047
	0.059	0.037	
	0.027	0.020	
375-424	2	2	4
	0.014	0.012	
	0.500	0.500	0.014
	0.015	0.012	
	0.007	0.007	
525-575	0	1	1
	0.459	0.391	
	0.000	1.000	0.003
	0.000	0.006	
	0.000	0.003	
Column Total	136	160	296
	0.459	0.541	

Exploratory Data Analysis



	data$Diagnosis_Heart_Disease		
data$CI_RBP	absent	present	Row Total
090-109	4	16	20
	2.930	2.491	
	0.200	0.800	0.068
	0.029	0.100	
	0.014	0.054	
110-129	53	59	112
	0.046	0.039	
	0.473	0.527	0.378
	0.390	0.369	
	0.179	0.199	
130-149	50	63	113
	0.071	0.060	
	0.442	0.558	0.382
	0.368	0.394	
	0.169	0.213	
150-169	20	18	38
	0.370	0.314	
	0.526	0.474	0.128
	0.147	0.112	
	0.068	0.061	
170-189	7	4	11
	0.749	0.637	
	0.636	0.364	0.037
	0.051	0.025	
	0.024	0.014	
190-210	2	0	2
	1.272	1.081	
	1.000	0.000	0.007
	0.015	0.000	
	0.007	0.000	
Column Total	136	160	296
	0.459	0.541	

Model Development

Every machine learning algorithm works best under a given set of conditions. Making sure your algorithm fits the assumptions/requirements ensures superior performance. You can't use any algorithm in any condition. For example: Have you ever tried using linear regression on a **categorical dependent** variable? Don't even try! Because you won't be appreciated for getting extremely low values of adjusted R^2 and F statistic.
Since this data has dependent variables as **categorical dependent** we need to build the model using **Logistic Regression.**

Dependent variable in data is **"Diagnosis_Heart_Disease"** and rest of the 13 variables are independent or predictor variables.

Logistic Regression

Logistic Regression is a **classification algorithm**. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. To represent binary/categorical outcome, we use dummy variables. You can also think of logistic regression as a special case of linear regression when the outcome variable is categorical, where we are using log of odds as dependent variable. In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function.

Model Development

Model

\# A tibble: 19 × 6

	term	estimate	std.error	statistic	p.value	odds_ratio
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	(Intercept)	1.82	5.03	0.362	0.717	6.18
2	Resting_ECGnormal	0.859	3.34	0.257	0.797	2.36
3	Resting_ECGST-T,abnormality	0.725	3.33	0.217	0.828	2.06
4	Peak_Exercise_ST_Segmentup-sloaping	0.568	1.39	0.409	0.683	1.76
5	Chest_Pain_Typenon-angina	0.493	0.796	0.620	0.535	1.64
6	Age	0.0462	0.0324	1.43	0.154	1.05
7	Max_Heart_Rate_Achieved	0.0446	0.0150	2.96	0.00303	1.05
8	Serum_Cholesterol	-0.0108	0.00500	-2.16	0.0311	0.989
9	Resting_Blood_Pressure	-0.0331	0.0134	-2.47	0.0136	0.967
10	ST_Depression_Exercise	-0.103	0.281	-0.368	0.713	0.902
11	Fasting_Blood_Sugar>,120,mg/dl	-0.112	0.764	-0.147	0.883	0.894
12	Chest_Pain_Typeatypical	-0.148	0.952	-0.155	0.876	0.862
13	Peak_Exercise_ST_Segmentflat	-0.383	1.32	-0.290	0.772	0.682
14	Thalassemianormal	-0.783	0.916	-0.855	0.392	0.457
15	Exercise_Induced_Anginayes	-0.813	0.536	-1.52	0.129	0.444
16	Num_Major_Vessels_Flouro	-1.78	0.394	-4.51	0.00000647	0.169
17	Chest_Pain_Typetypical	-1.85	0.769	-2.41	0.0159	0.157
18	Sexmale	-2.09	0.688	-3.03	0.00241	0.124
19	Thalassemiareversible,defect	-2.13	0.931	-2.29	0.0221	0.119

```
Degrees of Freedom: 235 Total (i.e. Null);  217 Residual
Null Deviance:      325.5
Residual Deviance: 128.5        AIC: 166.5
```

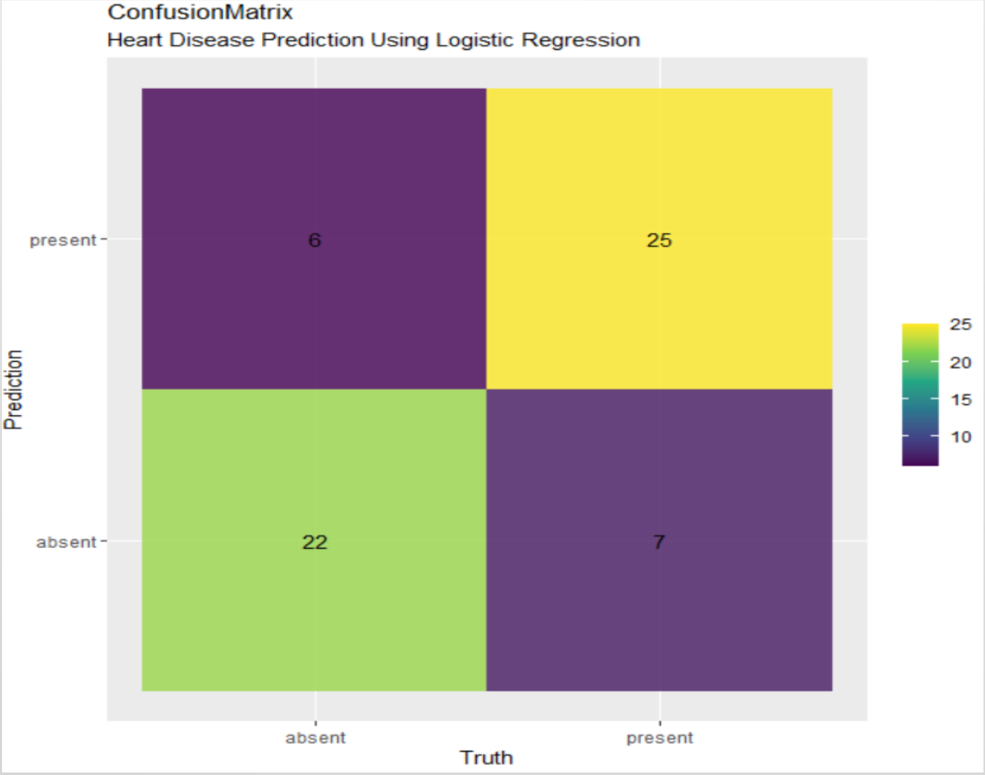
- **<u>AIC (Akaike Information Criteria)</u>** – The analogous metric of adjusted R^2 in logistic regression is AIC. AIC is the measure of fit which penalizes model for the number of model coefficients. Therefore, we always prefer model with minimum AIC value
- **<u>Null Deviance and Residual Deviance</u>** – Null Deviance indicates the response predicted by a model with nothing but an intercept. Lower the value, better the model. Residual deviance indicates the response predicted by a model on adding independent variables. Lower the value, better the model.

Model Development

- I've converted the estimate of the coefficient into the odds ratio. The odds ratio represents the odds that an outcome will occur given the presence of a specific predictor, compared to the odds of the outcome occurring in the absence of that predictor, assuming all other predictors remain constant. The odds ratio is calculated from the exponential function of the coefficient estimate based on a unit increase in the predictor

Confusion Matrix: It is nothing but a tabular representation of Actual vs Predicted values. This helps us to find the accuracy of the model and avoid overfitting. This is how it looks like:

	Prediction	Truth	n	outcome
1	absent	absent	22	true_negative
2	present	absent	6	false_positive
3	absent	present	7	false_negative
4	present	present	25	true_positive



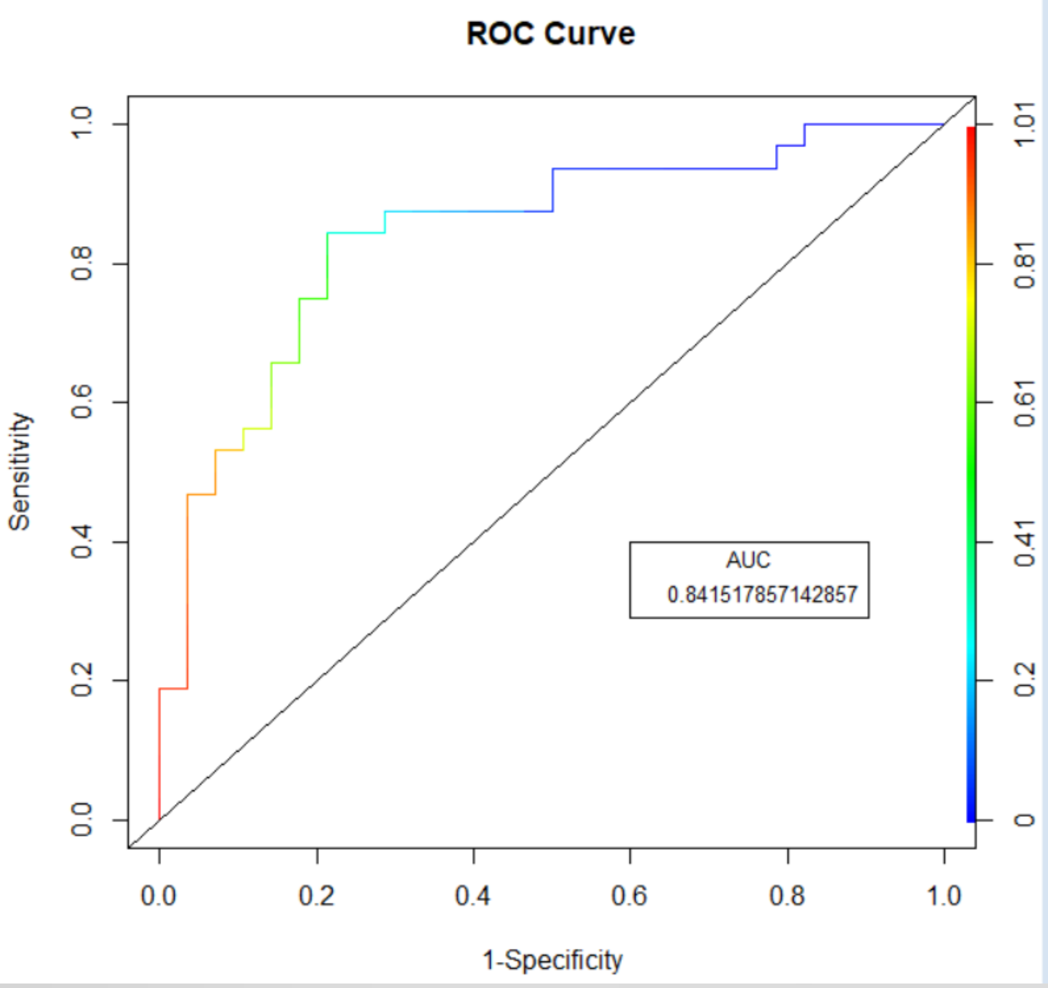
Model Development

- Accuracy represents the percentage of correct predictions.
- Sensitivity (true positive rate) refers to the probability of a positive test, conditioned on truly being positive.
- Specificity (true negative rate) refers to the probability of a negative test, conditioned on truly being negative
- Positive predictive value:It is the ratio of patients truly diagnosed as positive to all those who had positive test results (including healthy subjects who were incorrectly diagnosed as patient). This characteristic can predict how likely it is for someone to truly be patient, in case of a positive test result.
- Negative predictive value:It is the ratio of subjects truly diagnosed as negative to all those who had negative test results (including patients who were incorrectly diagnosed as healthy). This characteristic can predict how likely it is for someone to truly be healthy, in case of a negative test result.

	metric	estimate
1	accuracy	0.783
2	sensitivity	0.786
3	specificity	0.781
4	positive predictive value	0.759
5	negative predictive value	0.806

Model Development

<u>ROC Curve:</u> Receiver Operating Characteristic(ROC) summarizes the model's performance by evaluating the trade offs between true positive rate (sensitivity) and false positive rate(1- specificity). For plotting ROC, it is advisable to assume p > 0.5 since we are more concerned about success rate. ROC summarizes the predictive power for all possible values of p > 0.5. The **area under curve** (AUC), referred to as index of accuracy(A) or concordance index, is a perfect performance metric for ROC curve. Higher the area under curve, better the prediction power of the model. Below is a sample ROC curve. The ROC of a perfect predictive model has TP equals 1 and FP equals 0. This curve will touch the top left corner of the graph.



Conclusion

The project's objective was to develop a model that could identify patients with heart disease at high risk. Prediction of the risk of heart disease is a fairly complex task. This model can foresee whether the patient has heart disease present or absent, aiding specialists to ensure that the patient in need of heart surgery consideration can get on the schedule and also help anticipate the loss of human lives.

This project achieves this by analyzing many key variables which are called independent or predictor variables like Age, Sex, Chest Pain Type, Resting BP, etc using various models and retrospective analysis of the patient's medical records.