Probability Project - Spring 2022

Heart Disease Prediction Analysis

F190126 BILAL AFZAL

F190130 UZAIR RAMZAN

F190170 DANISH AHMAD

F190228 MUHAMMAD ZAIN

Introduction:

For this project, we searched the different datasets for health domain. After a detailed searched we have selected **HEART FAILURE PREDICTION DATA SET** from KAGGLE. It contains the information about AGE, SEX, CP [Chest pain], TRTBPS [The person's resting blood pressure], CHOL [cholesterol], FBS [fasting blood sugar], RESTECG, THALACHH, EXNG, OLDPEAK, SLP, CAA THALL, and OUTPUT.

We applied different statistical methods, graphical representations, and probability methods using these attributes.

Summary of Data:

We applied following statistical methods to understand data so that it can be helpful in later methods to conclude results.

1. Min
2. First Quartile
3. Median
4. Mean
5. Third Quartile
6. Max

age	sex	cp	trtbps	chol	fbs	restecg
Min. :29.00	Min. :0.0000	Min. :0.000	Min. : 94.0	Min. :126.0	Min. :0.0000	Min. :0.0000
1st Qu.:47.50	1st Qu.:0.0000	1st Qu.:0.000	1st Qu.:120.0	1st Qu.:211.0	1st Qu.:0.0000	1st Qu.:0.0000
Median :55.00	Median :1.0000	Median :1.000	Median :130.0	Median :240.0	Median :0.0000	Median :1.0000
Mean :54.37	Mean :0.6832	Mean :0.967	Mean :131.6	Mean :246.3	Mean :0.1485	Mean :0.5281
3rd Qu.:61.00	3rd Qu.:1.0000	3rd Qu.:2.000	3rd Qu.:140.0	3rd Qu.:274.5	3rd Qu.:0.0000	3rd Qu.:1.0000
Max. :77.00	Max. :1.0000	Max. :3.000	Max. :200.0	Max. :564.0	Max. :1.0000	Max. :2.0000

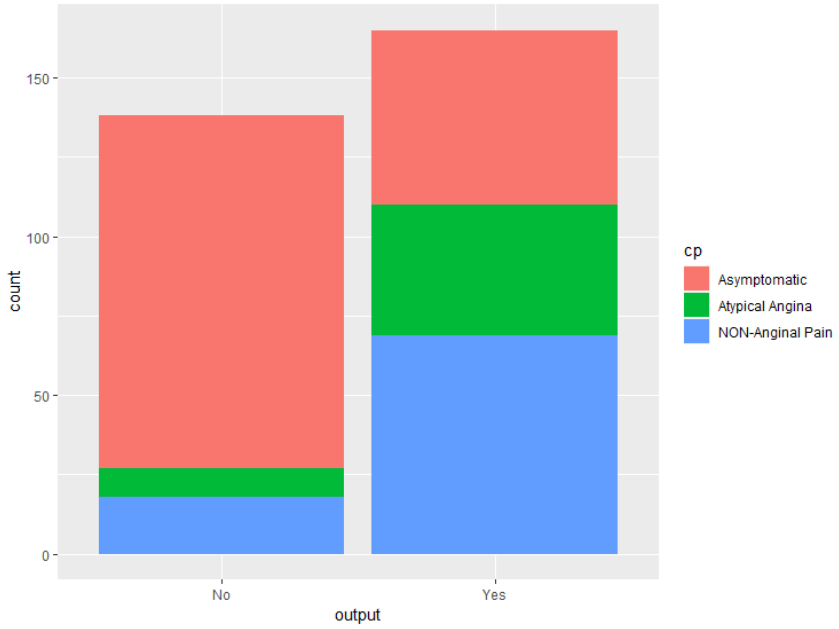
thalachh	exng	oldpeak	slp	caa	thall	output
Min. : 71.0	Min. :0.0000	Min. :0.00	Min. :0.000	Min. :0.0000	Min. :0.000	Min. :0.0000
1st Qu.:133.5	1st Qu.:0.0000	1st Qu.:0.00	1st Qu.:1.000	1st Qu.:0.0000	1st Qu.:2.000	1st Qu.:0.0000
Median :153.0	Median :0.0000	Median :0.80	Median :1.000	Median :0.0000	Median :2.000	Median :1.0000
Mean :149.6	Mean :0.3267	Mean :1.04	Mean :1.399	Mean :0.7294	Mean :2.314	Mean :0.5446
3rd Qu.:166.0	3rd Qu.:1.0000	3rd Qu.:1.60	3rd Qu.:2.000	3rd Qu.:1.0000	3rd Qu.:3.000	3rd Qu.:1.0000
Max. :202.0	Max. :1.0000	Max. :6.20	Max. :2.000	Max. :4.0000	Max. :3.000	Max. :1.0000

~FIG 1

Heart Attack due to chest pain:

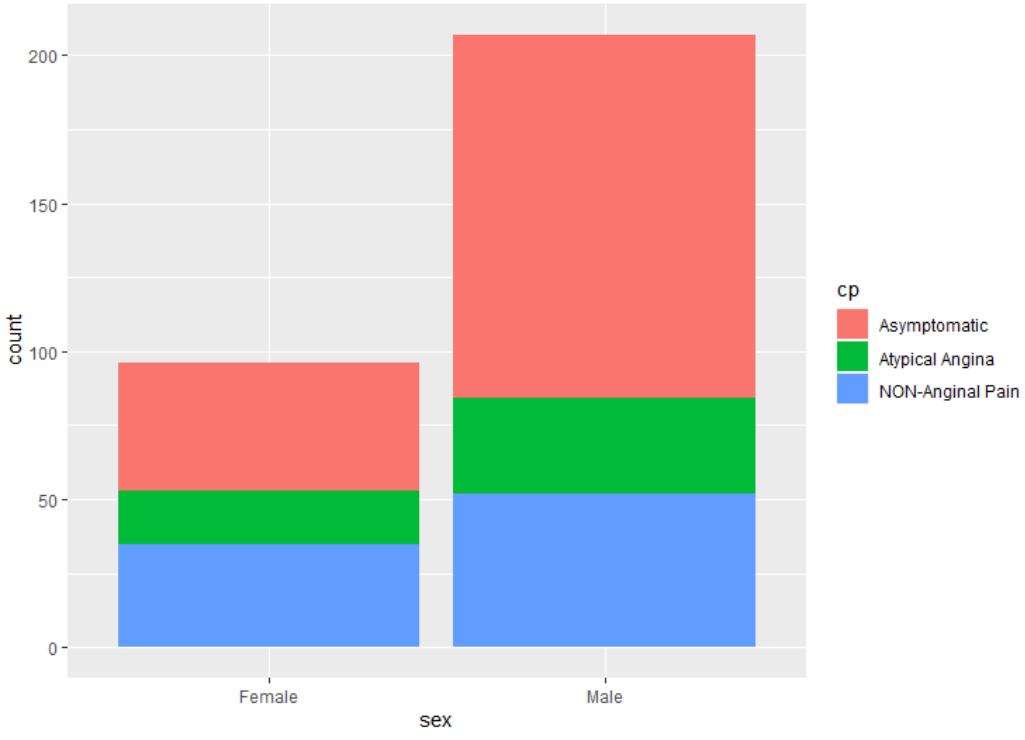
After analyzing the required attributes i.e., CP and Output, we concluded the following results:

1: Most of the people had heart attack due to **non-anginal chest pain.**
2: Very few got heart attack due to **asymptomatic pain**.
3: Rate of heart attack in **Atypical Angina** was moderate.



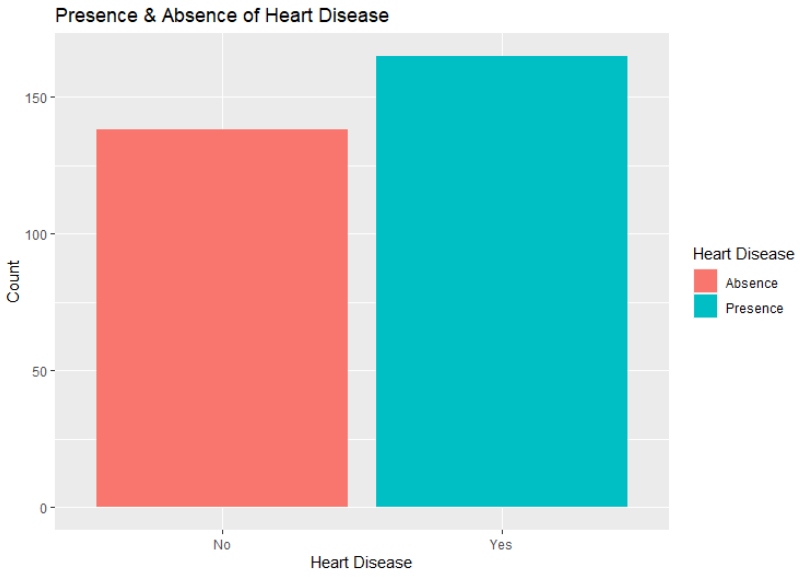
Chest pain in male & female:

As the graph clearly shows that the ratio of asymptomatic, atypical angina and non-anginal pain is greater in male then female.



Presence & Absence of Heart Disease

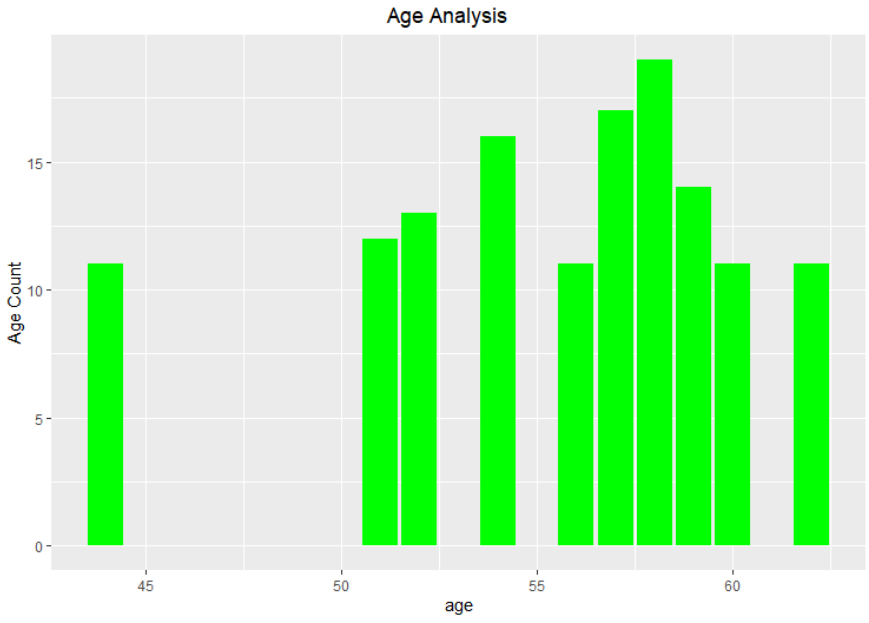
The result of graph clearly states that the probability of heart patient is 0.54 and non- heart patient is 0.45.



Age Analysis Graph

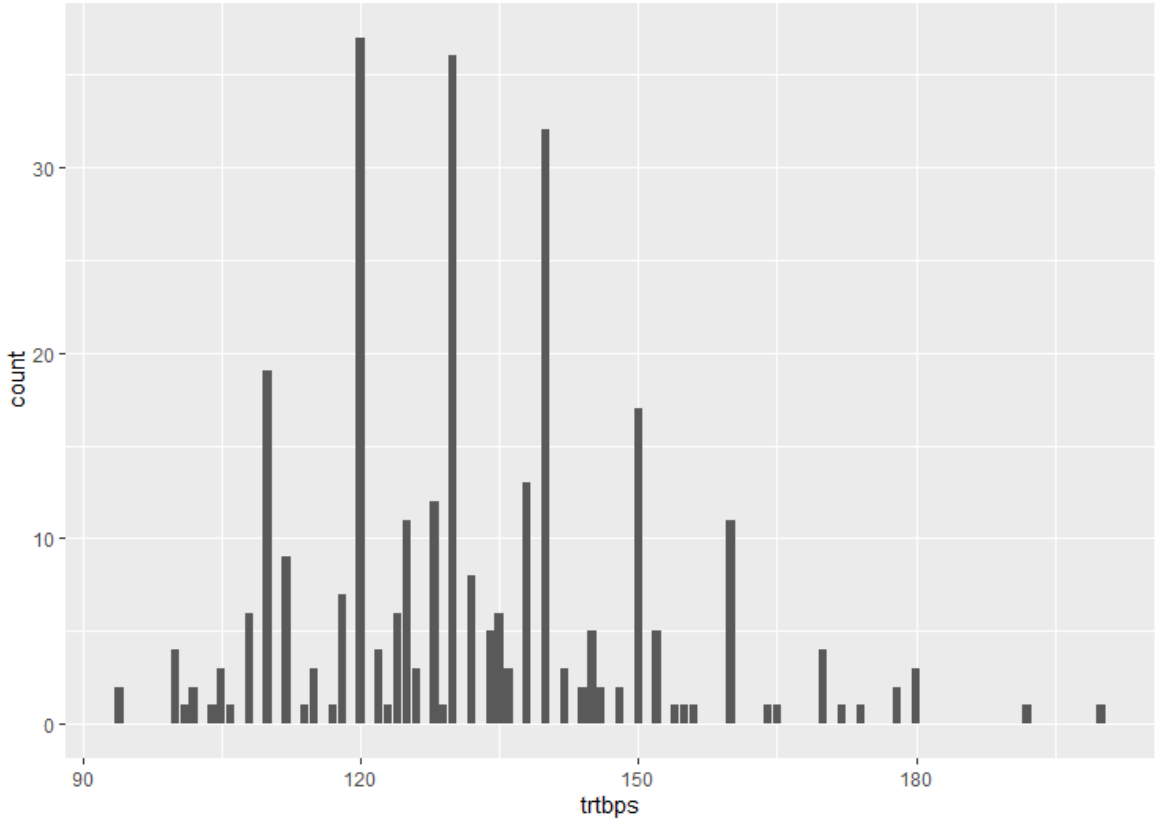
The age analysis graph shows the following results:

- 19 people got heart attack at the age of 58
- 11 people had heart attack at the age below 45
- The rate of heart attack between the age of 45-50 was none
- The heart attack rate of the age above 50 was the highest



Distribution of Resting Blood Pressure

Most of the people had blood pressure rate ranges between 120 to 140.
The blood pressure of rest of the people are scattered below 120 and above 140.



Ratio b/w Male and Female patients

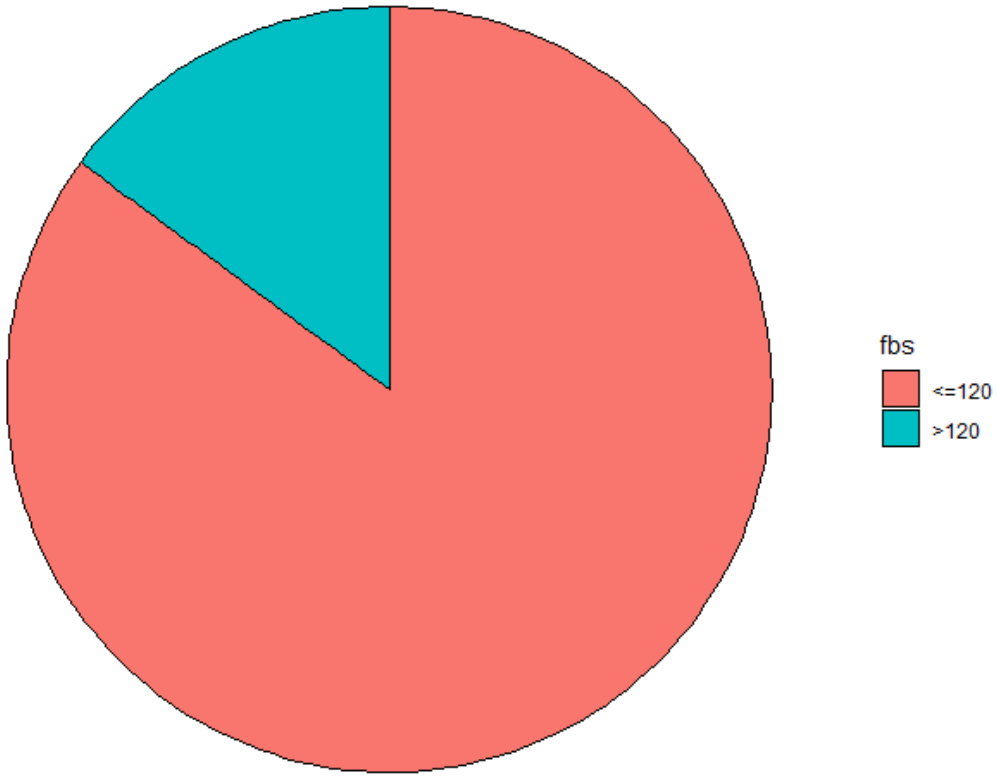
Following are the results between male and female patients had heart attack.

The ratio of heart attack between male and female is approximately 1/3.

	sex	n	prop	lab.ypos
1	Male	207	68.3	34.15
2	Female	96	31.7	84.15

Fasting blood sugar level:

Data for examining sugar level of people was analyzed and it shows the following results:



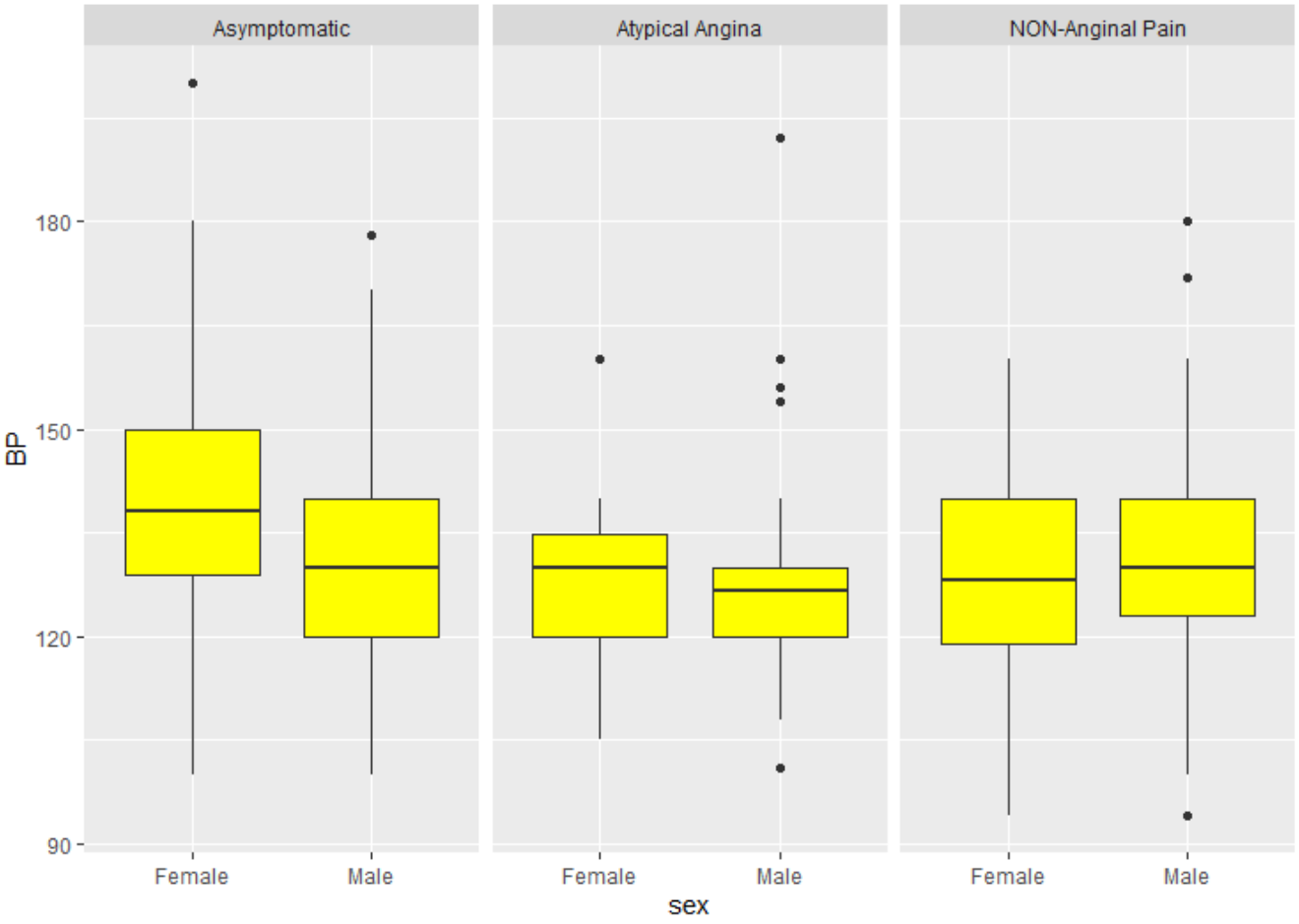
	fbs	n	prop	lab.ypos
1	>120	45	14.9	7.45
2	<=120	258	85.1	57.45

As it can be seen from the above table and chart, most of the people have low sugar level during fasting.

Comparing bp across with chest pain according to the gender:

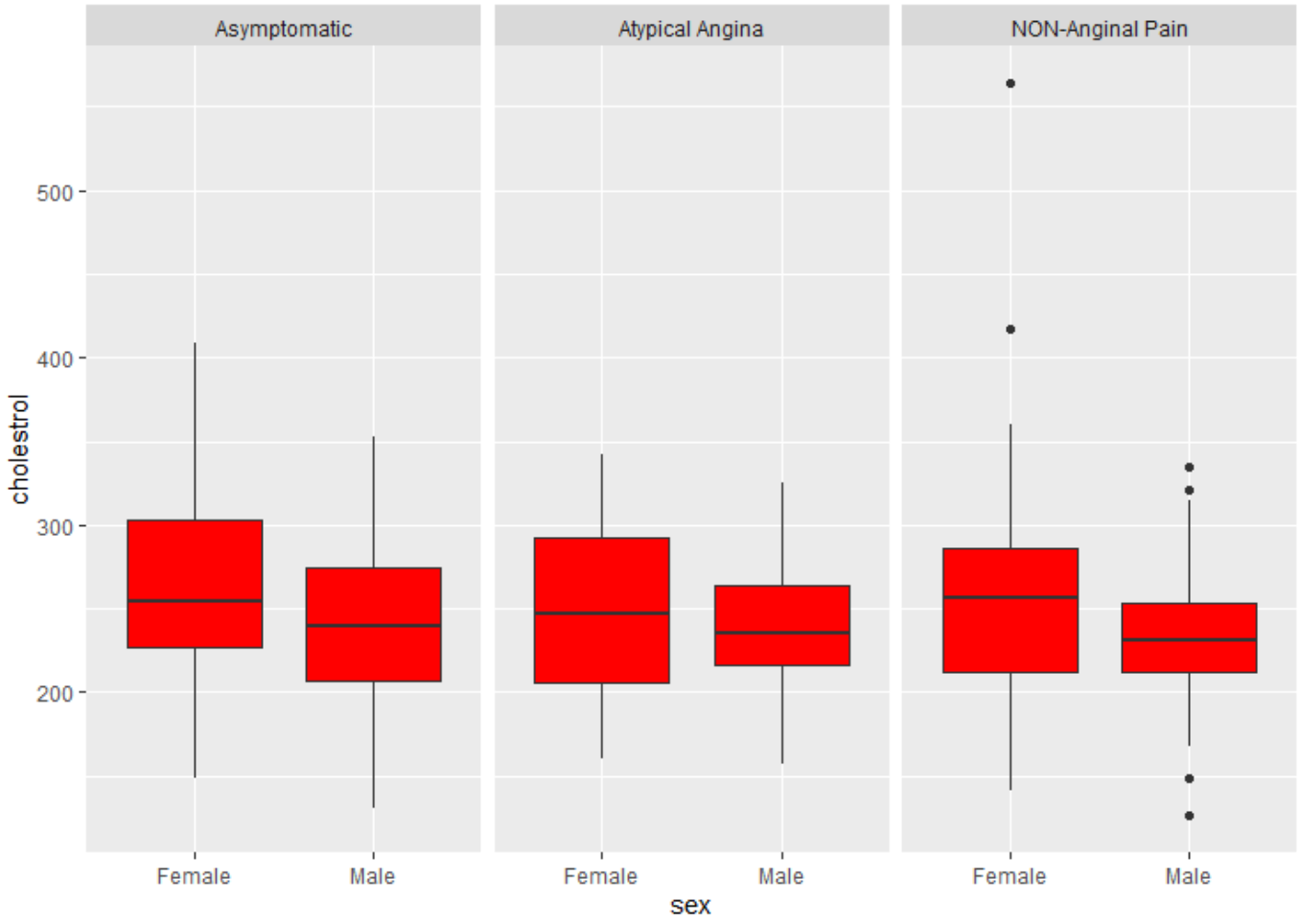
Gender data is given to the box plot and the results are shown below:

In asymptotic most of the female lies near 130-150 and males lie between 120-140.
In atypical angina females lie between 120-140 and males between 120-130.
In NON- anginal pain females lie between 120-140 and males between 125-140.



Comparing cholestrol across with chest pain according to the gender

Gender data is given to the box plot to analyze cholesterol level and the results are shown below:

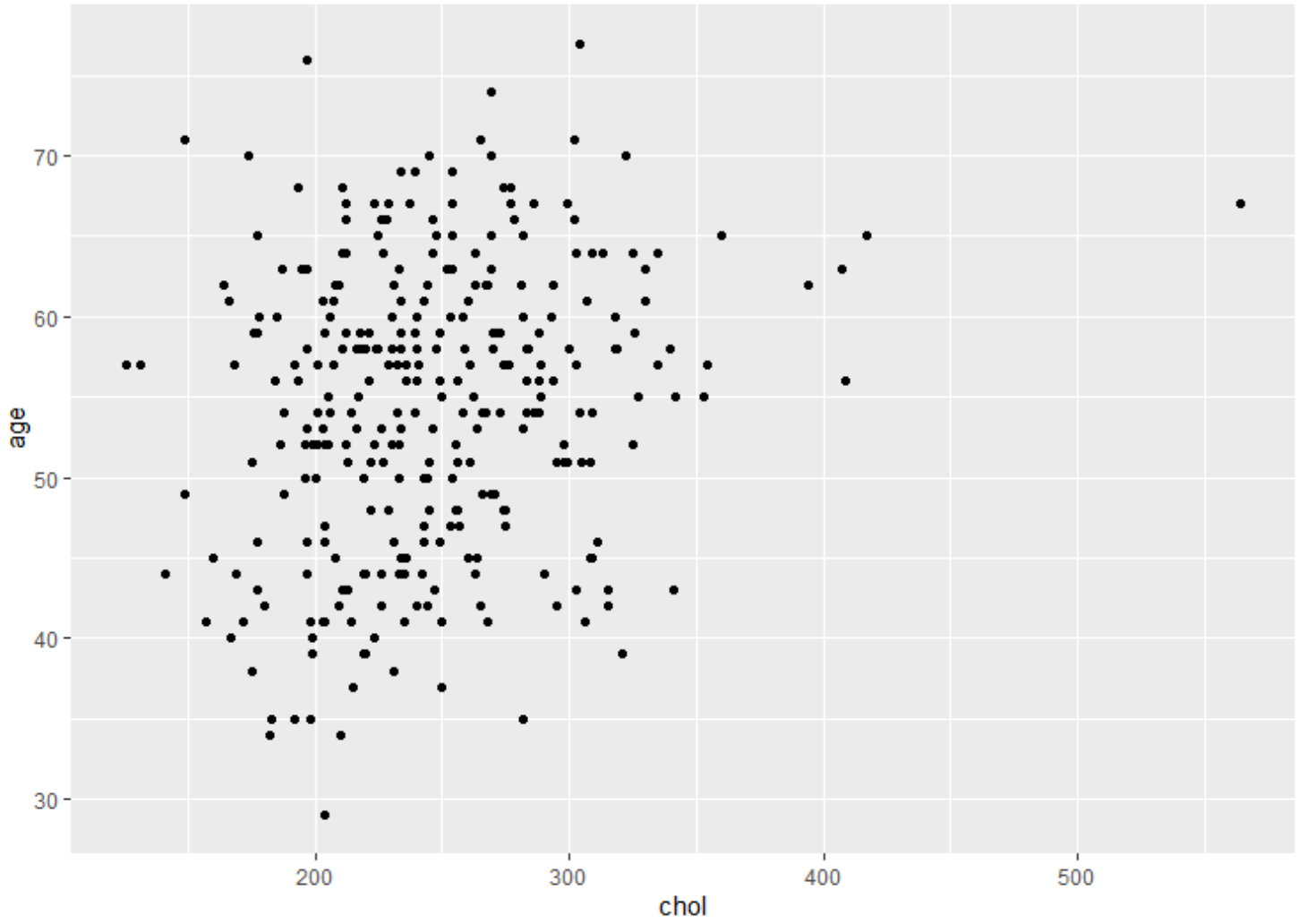


In asymptotic most of the female lies near 210-300 and males lie between 205-260.
In atypical angina females lie between 200-290 and males between 210-260.
In NON- anginal pain females lie between 210-280 and males between 210-250.

Scatter plot:

Below is the scatter plot which is drawn using age and cholesterol level is analyzed.

From the graph above, the trend is seen as the age increases the cholesterol level is also increasing.



SOURCE CODE

```
library(ggplot2)
library(dplyr)
library(tidyverse)
mydata <- read.csv(file.choose(),header = TRUE)

View(mydata)

library(GGally)
```

```
summary(mydata)

mysampledata <- mydata %>%
  mutate(sex = if_else(sex == 1, "Male", "Female"),
         fbs = if_else(fbs == 1, ">120", "<=120"),
         exng = if_else(exng == 1, "Yes", "No"),
         cp = if_else(cp == 1, "Atypical Angina",
                      if_else(cp == 2, "NON-Anginal Pain",
"Asymptomatic")),
         restecg = if_else(restecg == 1, "Normal",
                      if_else(cp == 2, "Abnormality", "Definite")),
         slp = as.factor(slp),
         caa = as.factor(caa),
         thall = as.factor(thall),
         output = if_else(output == 1, "Yes","No")
         )%>%
  mutate_if(is.character, as.factor)%>%
  dplyr::select(output, sex, fbs, exng, cp, restecg, slp, caa,
thall, everything())
  library(repr)

options(repr.plot.width = 6, repr.plot.height = 3)

view(mysampledata)

ggparcoord(mysampledata, columns = 1:2 , groupColumn = 5, showPoints
= TRUE,
           alphaLines = 0.3,scale = "std")

#
ggplot(mysampledata,aes(x = output,fill = cp))+geom_bar(position
="stack")

#
ggplot(mysampledata,aes(x = sex,fill = cp))+geom_bar(position
="stack")


#Presence & Absence of Heart Disease
ggplot(mysampledata, aes(x=output, fill = output))+
  geom_bar()+
  xlab("Heart Disease")+
  ylab("Count")+
  ggtitle("Presence & Absence of Heart Disease")+
  scale_fill_discrete(name = "Heart Disease", labels = c("Absence",
"Presence"))
```

```
prop.table(table(mysampledata$output))


#Age Analysis
mysampledata %>%
  group_by(age)%>%
  count()%>%
  filter(n>10) %>%
  ggplot()+
  geom_col(aes(age, n), fill = "green")+
  ggtitle("
Age Analysis")+
  xlab("age")+
  ylab("Age Count")

#distribution of resting blood pressure
ggplot(mysampledata, aes(x = trtbps))+
  geom_bar()


# ratio of male and female patient
plotdata <-mysampledata %>%
  count(sex) %>%
  arrange(desc(sex)) %>%
  mutate(prop = round(n * 100 / sum(n), 1),
         lab.ypos = cumsum(prop) - 0.5 *prop)

ggplot(plotdata,
       aes(x="",
           y = prop,
           fill = sex))+
  geom_bar(width = 1,
           stat = "identity",
           color = "black")+
  coord_polar("y",
              start = 0,
              direction = -1)+
  theme_void()


#fasting blood sugar level
plotdata1 <-mysampledata %>%
  count(fbs) %>%
  arrange(desc(fbs)) %>%
  mutate(prop = round(n * 100 / sum(n), 1),
         lab.ypos = cumsum(prop) - 0.5 *prop)

ggplot(plotdata1,
       aes(x="",
```

```
           y = prop,
           fill = fbs))+
  geom_bar(width = 1,
           stat = "identity",
           color = "black")+
  coord_polar("y",
              start = 0,
              direction = -1)+
  theme_void()

view(plotdata1)

#comparing bp across with chest pain according to the gender
mysampledata %>%
  ggplot(aes(x=sex, y=trtbps))+
  geom_boxplot(fill = "yellow")+
  xlab("sex")+
  ylab("BP")+
  facet_grid(~cp)

#comparing cholestrol across with chest pain according to the gender
mysampledata %>%
  ggplot(aes(x=sex, y=chol))+
  geom_boxplot(fill = "red")+
  xlab("sex")+
  ylab("cholestrol")+
  facet_grid(~cp)

#scatterplot
ggplot(mysampledata,
       aes(x=chol,
           y=age))+
  geom_point()
```