HS631_Project_code

Name : Tejaswee Katanguri

Import libraries

```
library(tidyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 4.0.3
```

```
library(ggcorrplot)
```

```
## Warning: package 'ggcorrplot' was built under R version 4.0.3
```

```
library(lattice)
library(PASWR2)
```

Load the data

```
Dataset <- read.csv("C:\\Users\\nithi\\Downloads\\heart.csv")
```

View the first 6 rows of the dataset

```
head(Dataset)
```

```
##   ï..age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal
## 1     63   1  3      145  233   1       0     150     0     2.3     0  0    1
## 2     37   1  2      130  250   0       1     187     0     3.5     0  0    2
## 3     41   0  1      130  204   0       0     172     0     1.4     2  0    2
## 4     56   1  1      120  236   0       1     178     0     0.8     2  0    2
## 5     57   0  0      120  354   0       1     163     1     0.6     2  0    2
## 6     57   1  0      140  192   0       1     148     0     0.4     1  0    1
##   target
## 1      1
## 2      1
## 3      1
## 4      1
## 5      1
## 6      1
```

Exploratory data analysis

Missing values

```
sum(is.na(Dataset))
```

```
## [1] 0
```

Fortunately there are no missing values in the dataset

Summary statistics of the data

```
summary(Dataset)
```

```
##      ï..age           sex               cp           trestbps
##  Min.   :29.00   Min.   :0.0000   Min.   :0.000   Min.   : 94.0
##  1st Qu.:47.50   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:120.0
##  Median :55.00   Median :1.0000   Median :1.000   Median :130.0
##  Mean   :54.37   Mean   :0.6832   Mean   :0.967   Mean   :131.6
##  3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:140.0
##  Max.   :77.00   Max.   :1.0000   Max.   :3.000   Max.   :200.0
##       chol            fbs            restecg          thalach
##  Min.   :126.0   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0
##  1st Qu.:211.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.5
##  Median :240.0   Median :0.0000   Median :1.0000   Median :153.0
##  Mean   :246.3   Mean   :0.1485   Mean   :0.5281   Mean   :149.6
##  3rd Qu.:274.5   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:166.0
##  Max.   :564.0   Max.   :1.0000   Max.   :2.0000   Max.   :202.0
##      exang           oldpeak         slope             ca
##  Min.   :0.0000   Min.   :0.00   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000   Median :0.80   Median :1.000   Median :0.0000
##  Mean   :0.3267   Mean   :1.04   Mean   :1.399   Mean   :0.7294
##  3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :6.20   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   :2.314   Mean   :0.5446
##  3rd Qu.:3.000   3rd Qu.:1.0000
##  Max.   :3.000   Max.   :1.0000
```

Changing the labels for better interpretation and visualizations

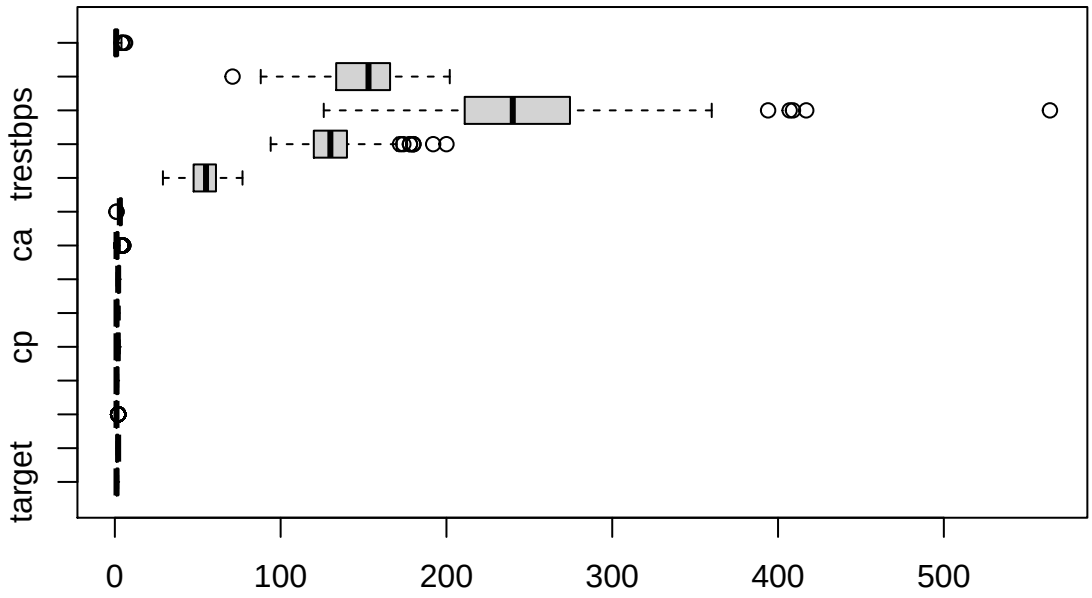
```
Dataset1 <- Dataset %>%
  rename("Age"=ï..age)
```

```
Heart_dataset <- Dataset1 %>%
  mutate(sex = ifelse(sex == 1, "Male", "Female"),
         fbs = ifelse(fbs == 1, ">120", "<=120"),
         exang = ifelse(exang == 1, "Yes","No"),
         restecg = ifelse(restecg == 0, "Normal",
                          ifelse(restecg == 1, "Abnormality", "Probable or Definite")),
         cp = ifelse(cp == 1, "Atypical Angina",
                     ifelse(cp == 2, "Non-Anginal Pain", "Asymptomatic")),
         target = ifelse(target == 1, "Heart Disease", "No Heart Disease"),
         slope = ifelse(slope == 1, "Flat",
```

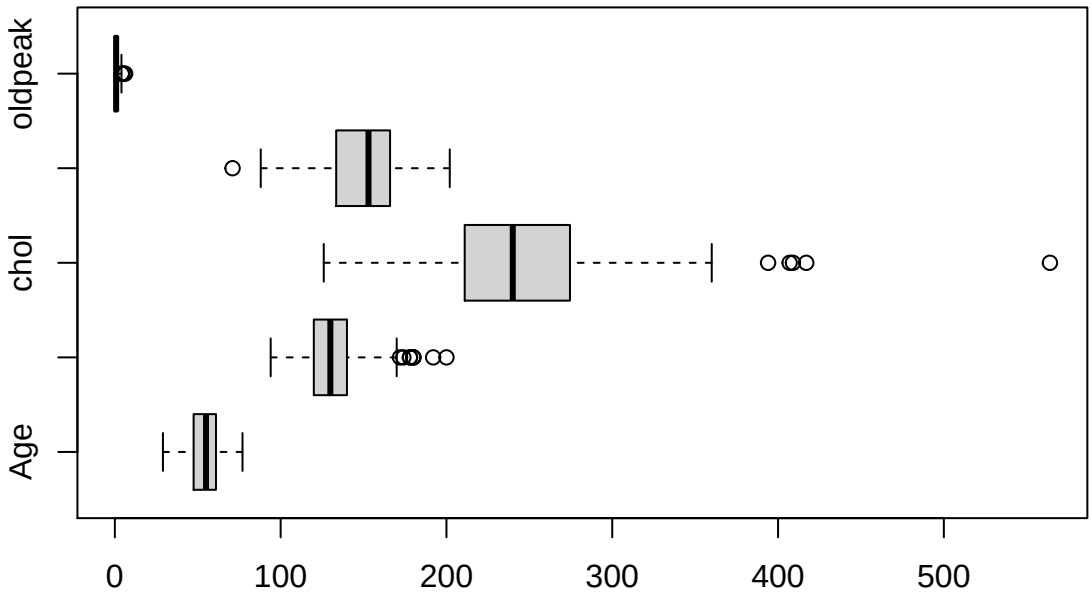
```
                        ifelse(slope == 0,"Downsloping", "Upsloping")))

Heart_dataset <- Heart_dataset %>%
  mutate(ca = as.factor(ca),
         thal = as.factor(thal)
         )%>%
  mutate_if(is.character, as.factor) %>%
  dplyr::select(target, sex, fbs, exang, cp, restecg, slope, ca, thal, everything())
```

BoxPlot to find Outliers

```
boxplot(Heart_dataset, horizontal = TRUE)
```



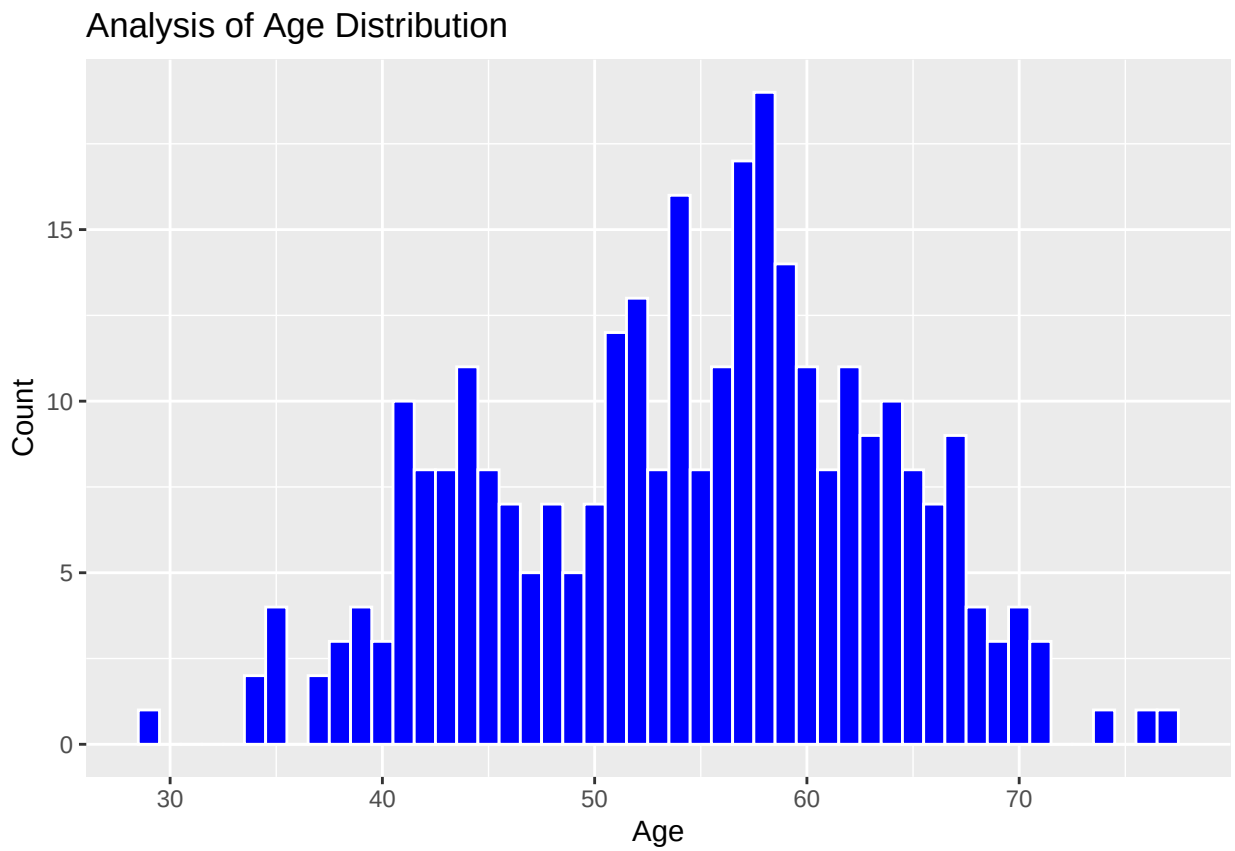

```
boxplot(Heart_dataset[, 10:14], horizontal = TRUE)
```



For the variable trestbps, number of outliers is 9 For the variable chol, number of outliers is 5 For the variable age, number of outliers is 0 For the variable oldpeak, number of outliers is 5 For the variable thalach , number of outliers is 1

Visualizations Distribution of Age, Cholestrol, thalach, oldpeak, trestbps Distribution of Age

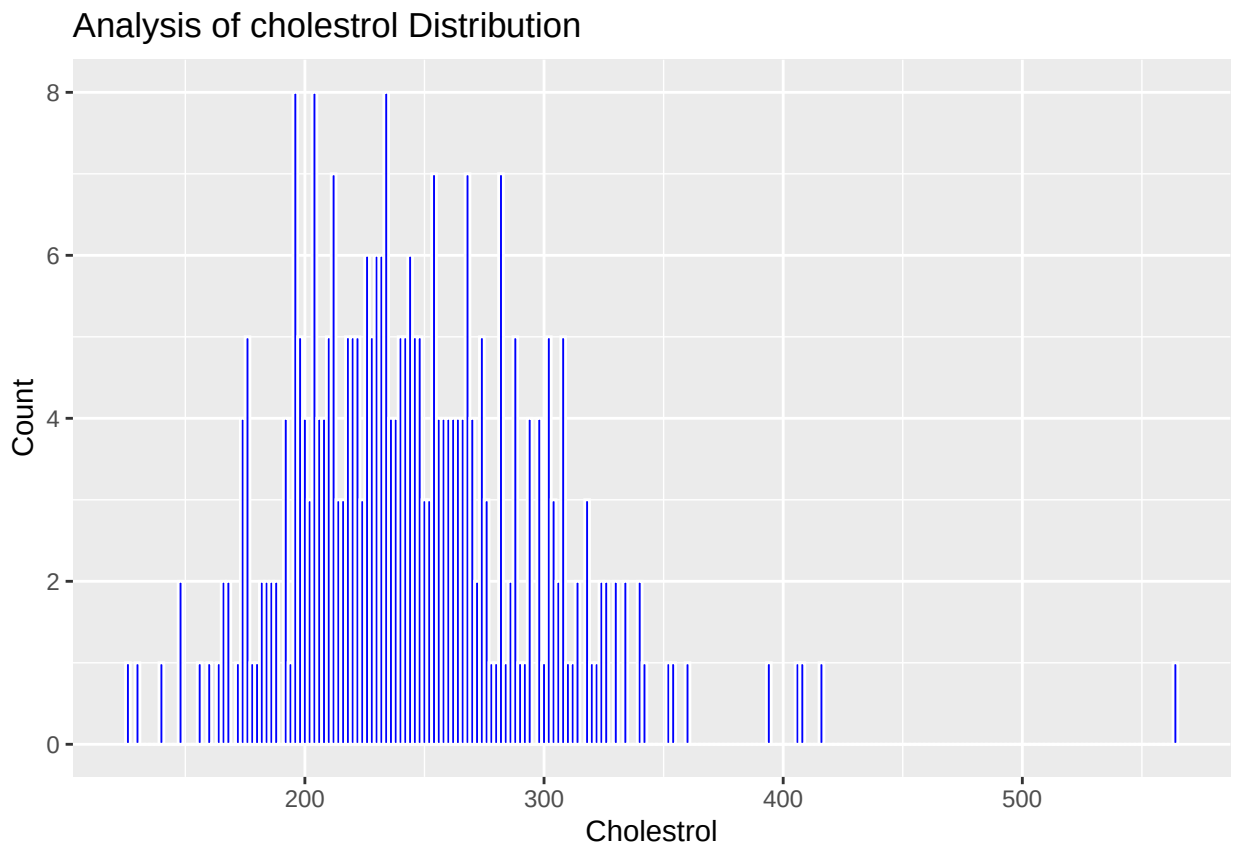
```
ggplot(Heart_dataset, mapping = aes(Age), fill = Age)+
  geom_histogram(binwidth = 1, color = "white", fill = "blue")+
  xlab("Age") +
  ylab("Count") +
  ggtitle("Analysis of Age Distribution")
```



Age distribution graph is Normally distributed and the mean age is 54 years

Distribution of Cholesterol

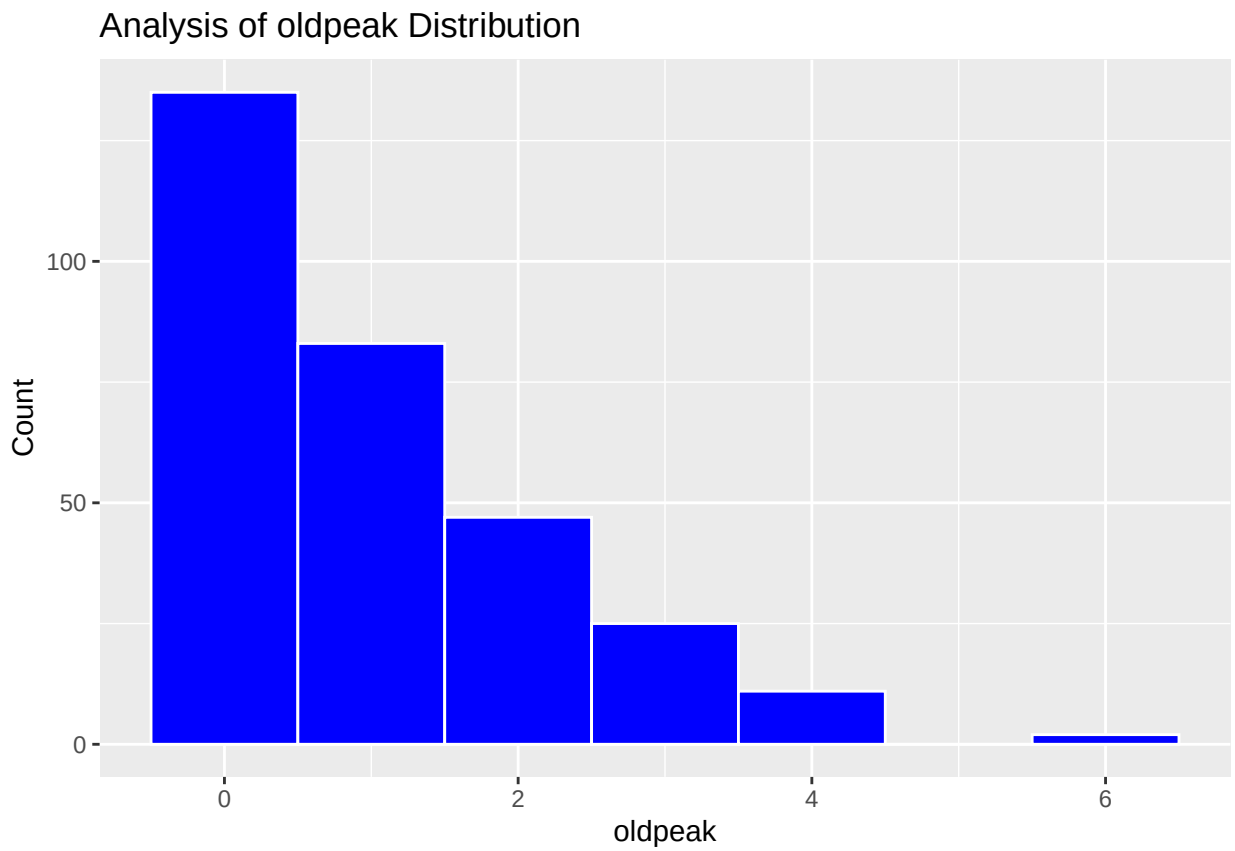
```
ggplot(Heart_dataset, mapping = aes(chol), fill = chol)+
  geom_histogram(binwidth = 2, color = "white", fill = "blue")+
  xlab("Cholestrol") +
  ylab("Count") +
  ggtitle("Analysis of cholestrol Distribution")
```



Cholesterol – Looks like it is almost Normally distributed

Distribution of oldpeak(ST depression induced by exercise relative to rest)

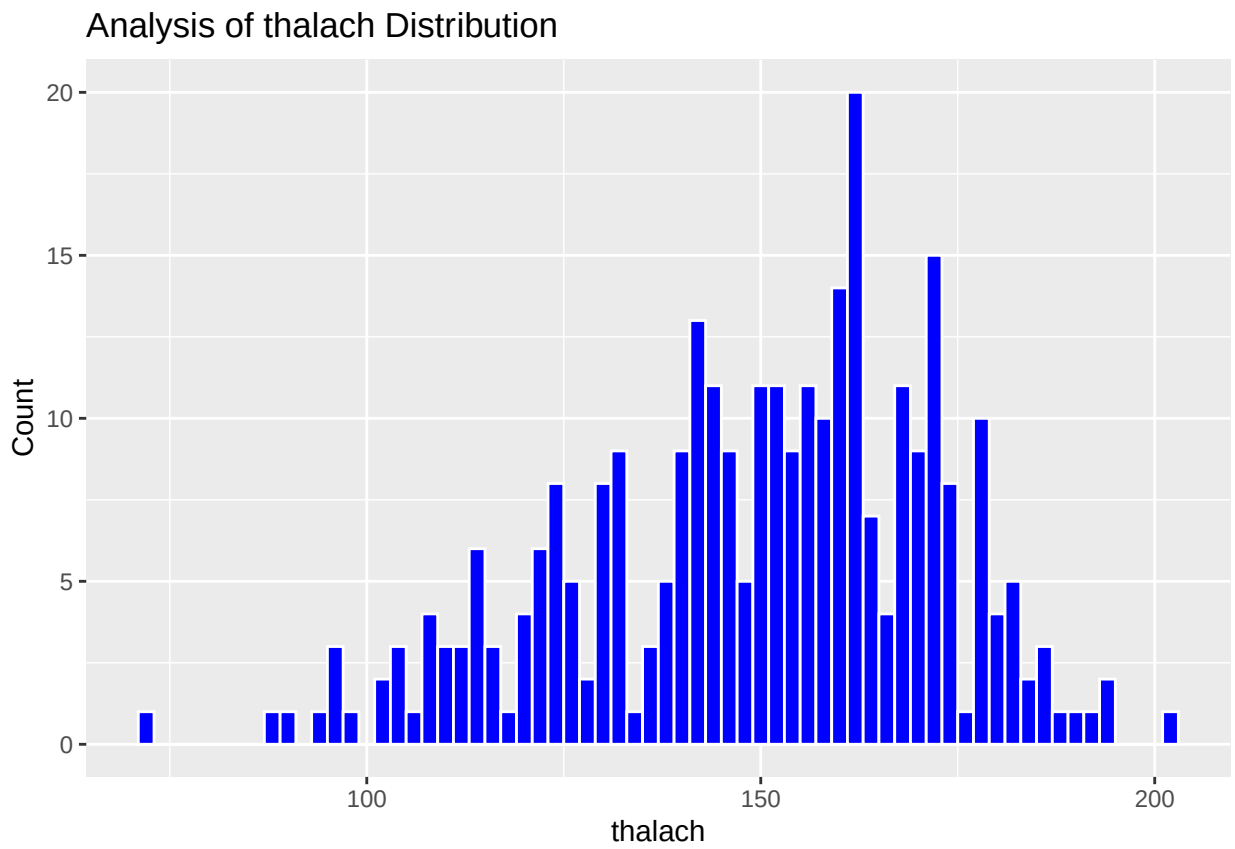
```
ggplot(Heart_dataset, mapping = aes(oldpeak), fill = oldpeak)+
  geom_histogram(binwidth = 1, color = "white", fill = "blue")+
  xlab("oldpeak") +
  ylab("Count") +
  ggtitle("Analysis of oldpeak Distribution")
```



Oldpeak is left skewed

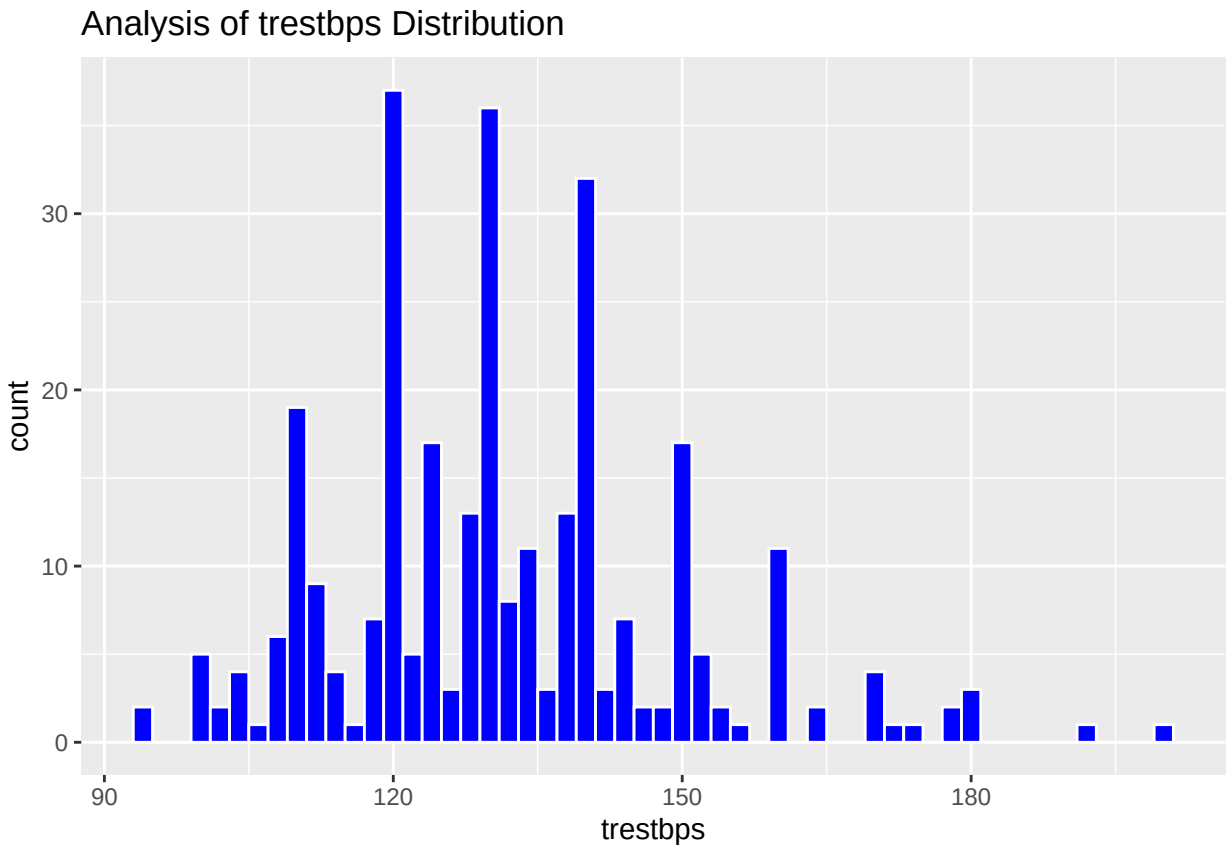
Distribution of thalach(maximum heart rate achieved)

```
ggplot(Heart_dataset, mapping = aes(thalach), fill = thalach)+
  geom_histogram(binwidth = 2, color = "white", fill = "blue")+
  xlab("thalach") +
  ylab("Count") +
  ggtitle("Analysis of thalach Distribution")
```



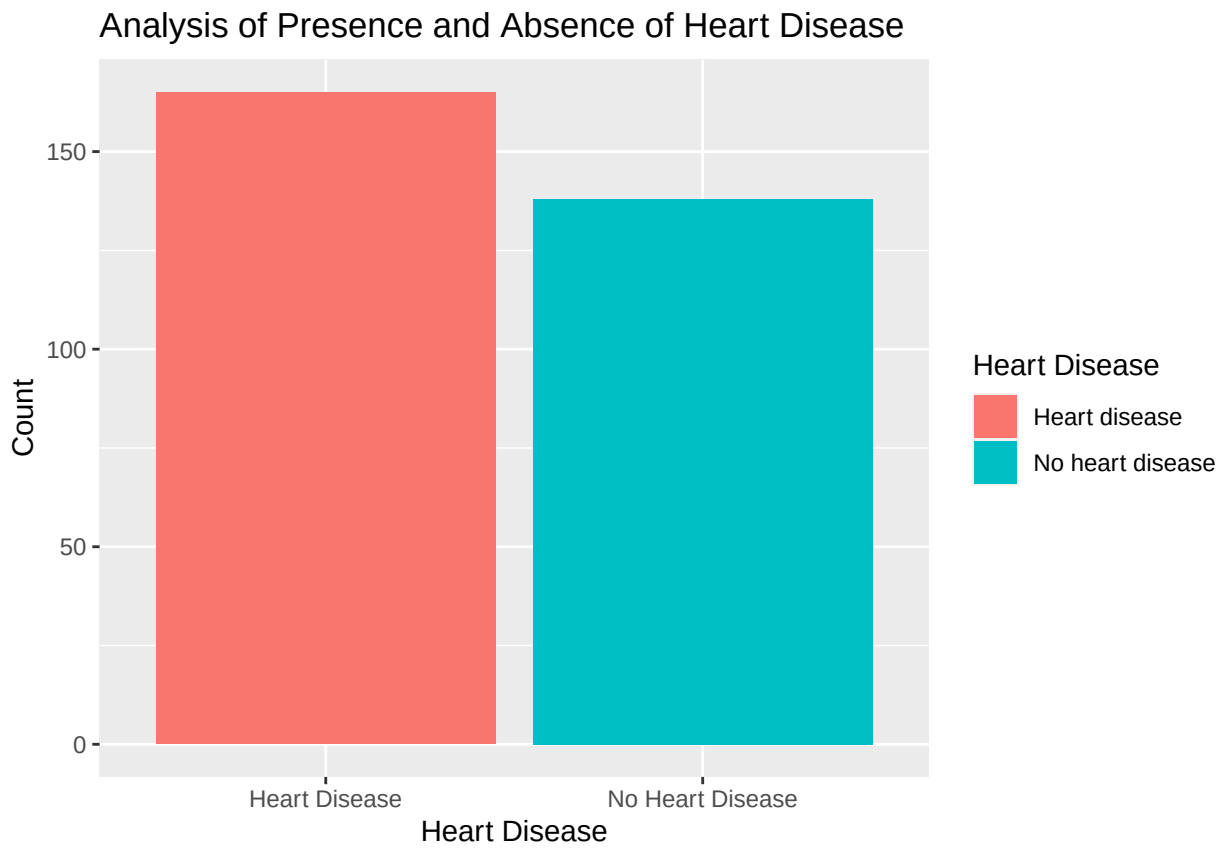
thalach- graph is right skewed Distribution of trestbps(resting blood pressure)

```
ggplot(Heart_dataset, mapping = aes(trestbps), fill = trestbps)+
  geom_histogram(binwidth = 2, color = "white", fill = "blue")+
  xlab("trestbps") +
  ggtitle("Analysis of trestbps Distribution")
```



Trestbps – graph is Normally distributed Distributions and relationships Analysis of Presence and Absence of Heart Disease

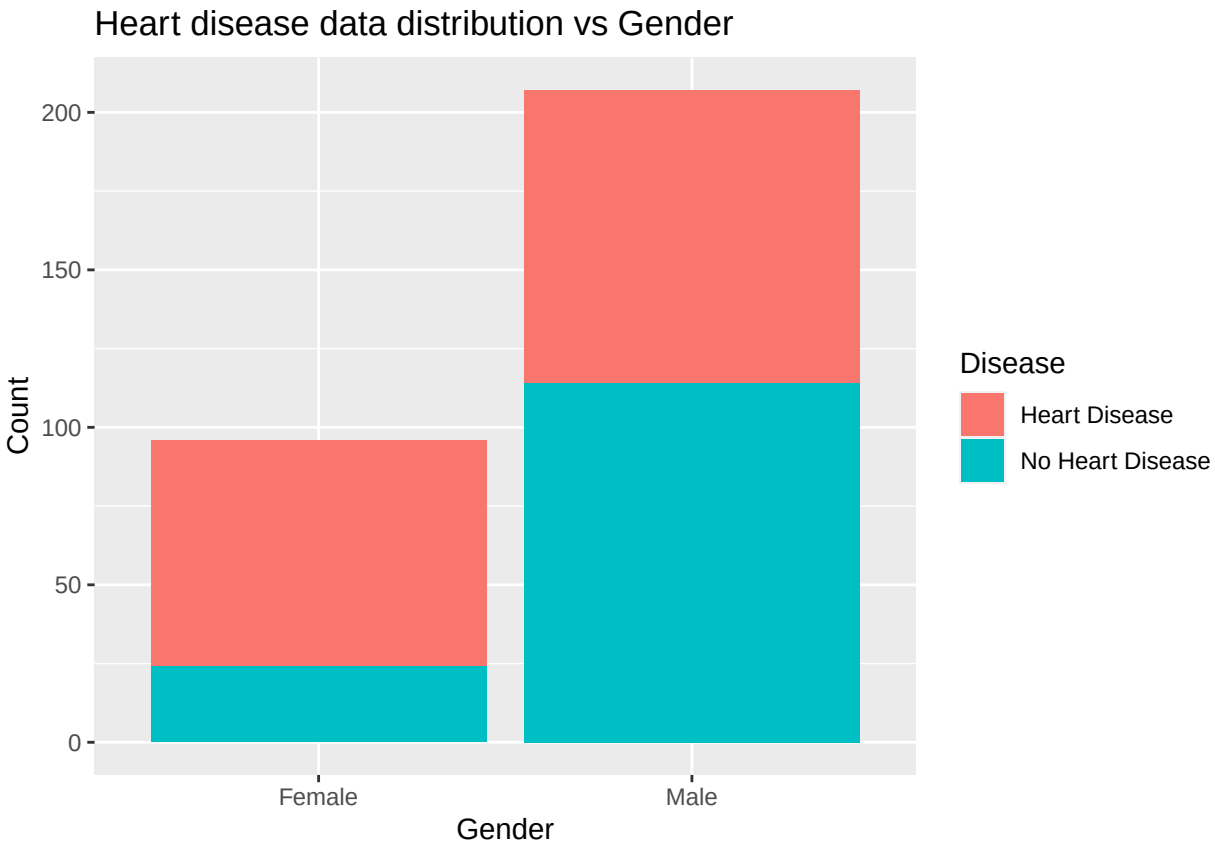
```
ggplot(Heart_dataset, aes(x=target, fill = target)) +
  geom_bar() +
  xlab("Heart Disease") +
  ylab("Count") +
  ggtitle("Analysis of Presence and Absence of Heart Disease") +
  scale_fill_discrete(name = "Heart Disease", labels = c("Heart disease", "No heart disease"))
```



There are more diseased patients than healthy in the dataset.

Heart disease data distribution vs Gender

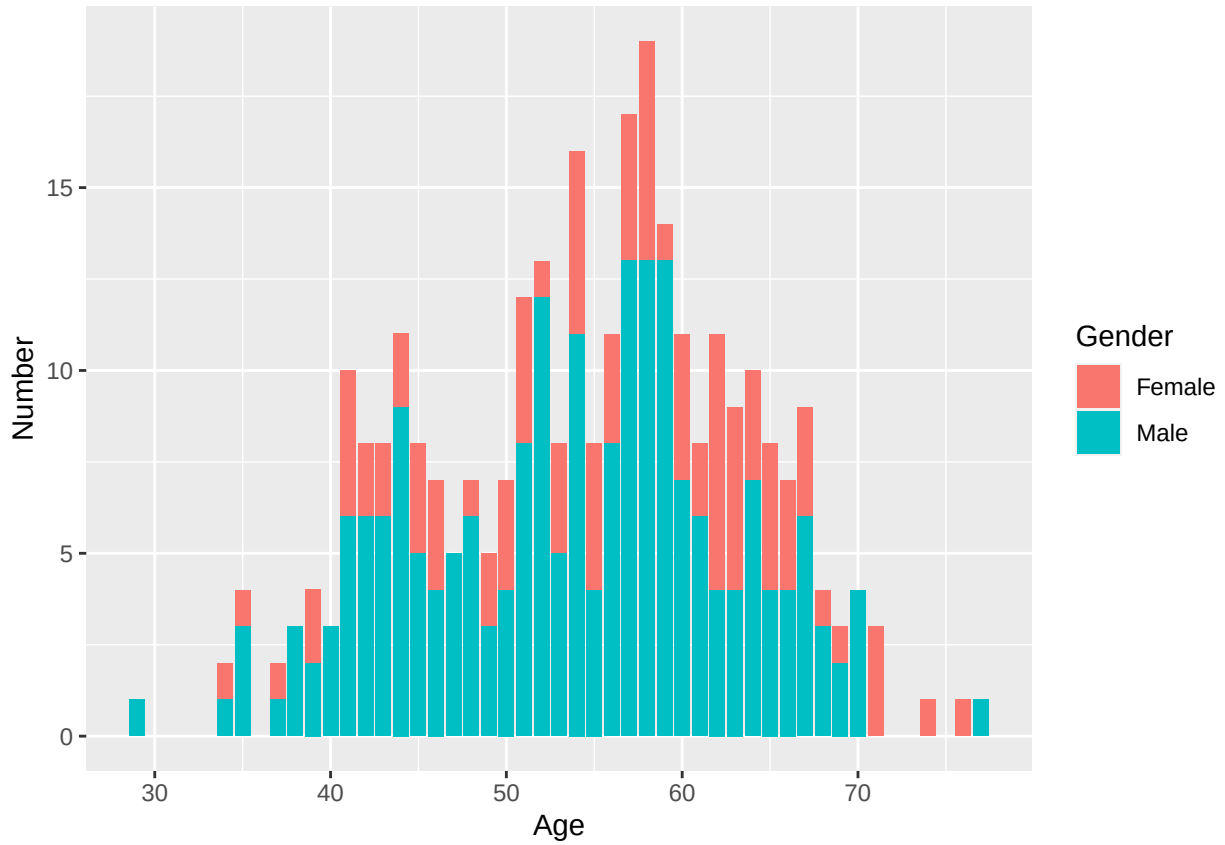
```
Heart_dataset %>%
  ggplot(aes(x=sex, fill = target))+
  geom_bar()+
  xlab("Gender") +
  ylab("Count")+
  guides(fill = guide_legend(title = "Disease"))+
  ggtitle("Heart disease data distribution vs Gender ")
```



Heart disease vs gender – From the graph, we can observe that among disease patients, male are higher than female. But the ratio of heart disease to no heart disease is higher with Females in the samples collected.

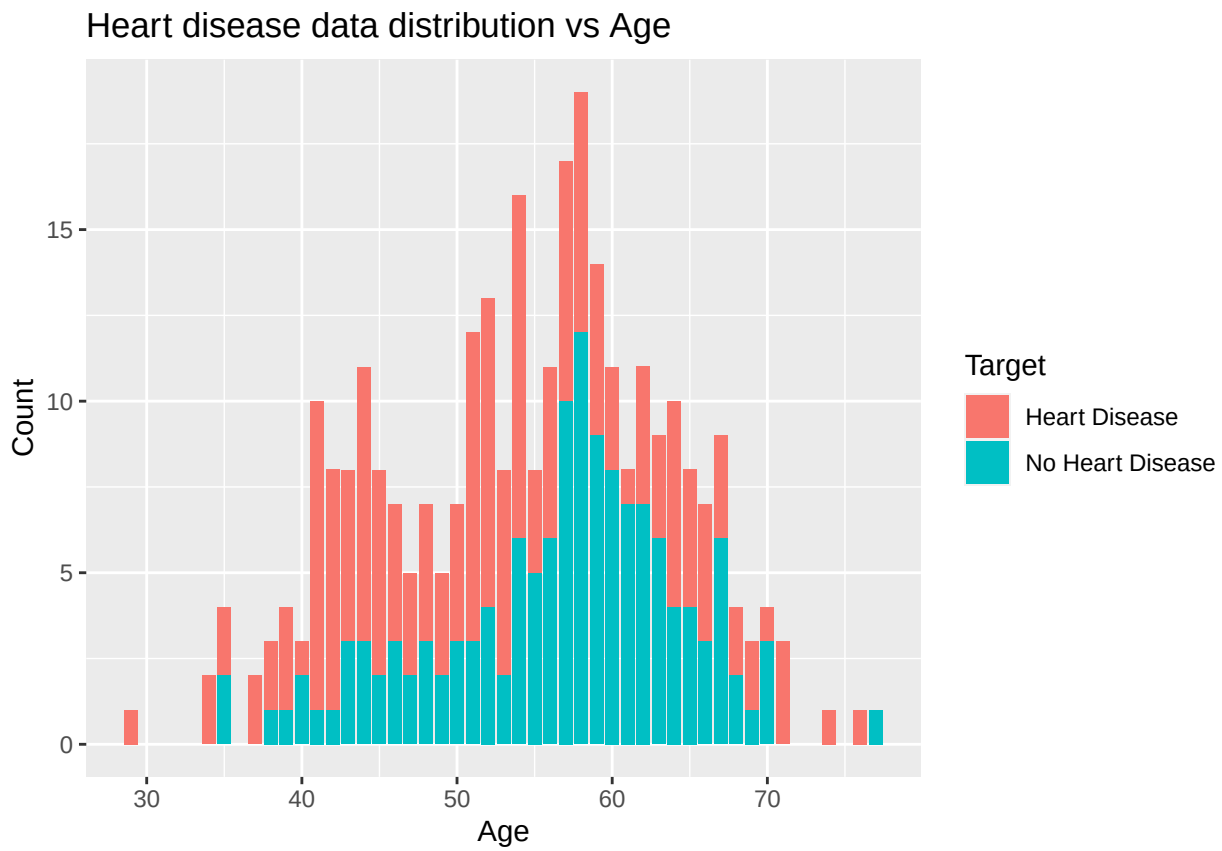
Heart disease data distribution vs Age/Gender

```
Heart_dataset %>%
  ggplot(aes(x=Age,fill=sex))+
  geom_bar()+
  xlab("Age") +
  ylab("Number")+
  guides(fill = guide_legend(title = "Gender"))
```



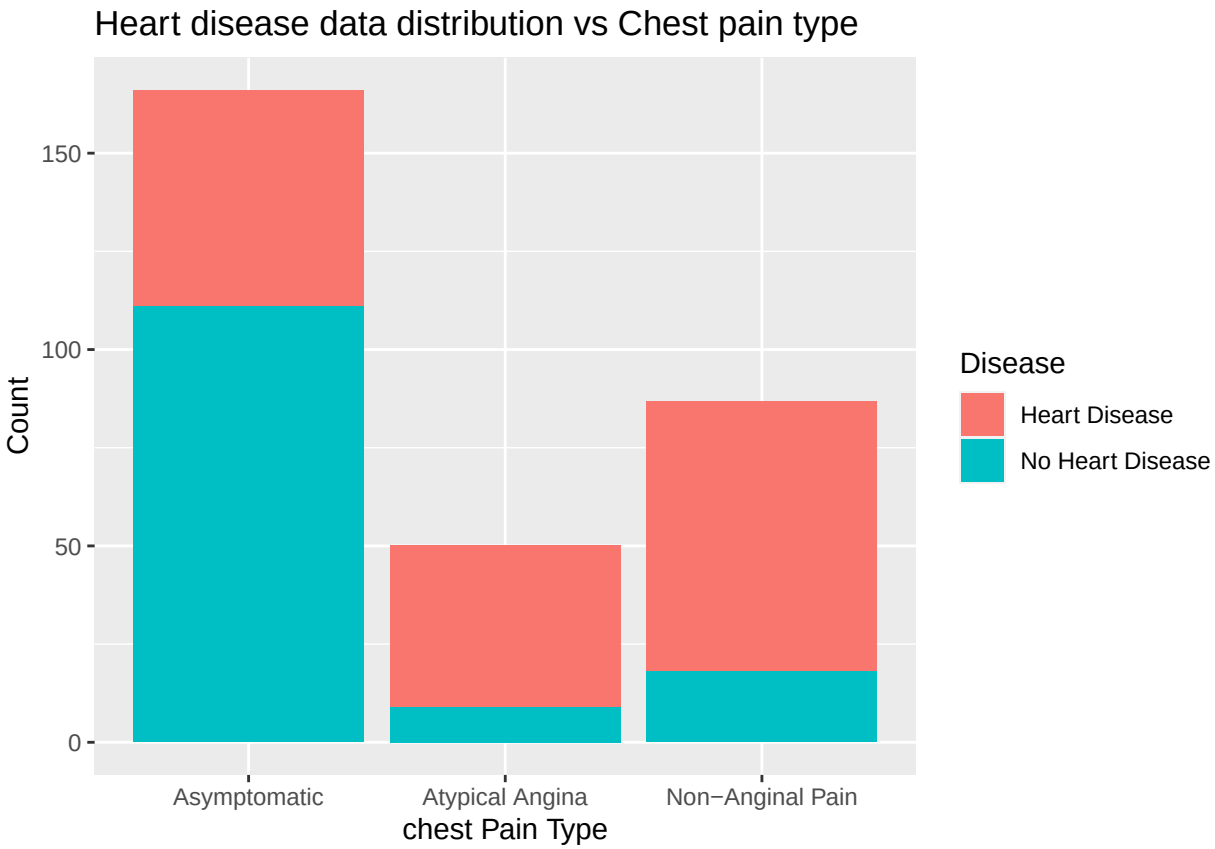
Heart disease data distribution vs Age

```
Heart_dataset %>%
  ggplot(aes(x=Age, fill = target))+
  geom_bar()+
  xlab("Age") +
  ylab("Count")+
  guides(fill = guide_legend(title = "Target"))+
  ggtitle("Heart disease data distribution vs Age")
```



Heart disease data distribution vs Chest pain type

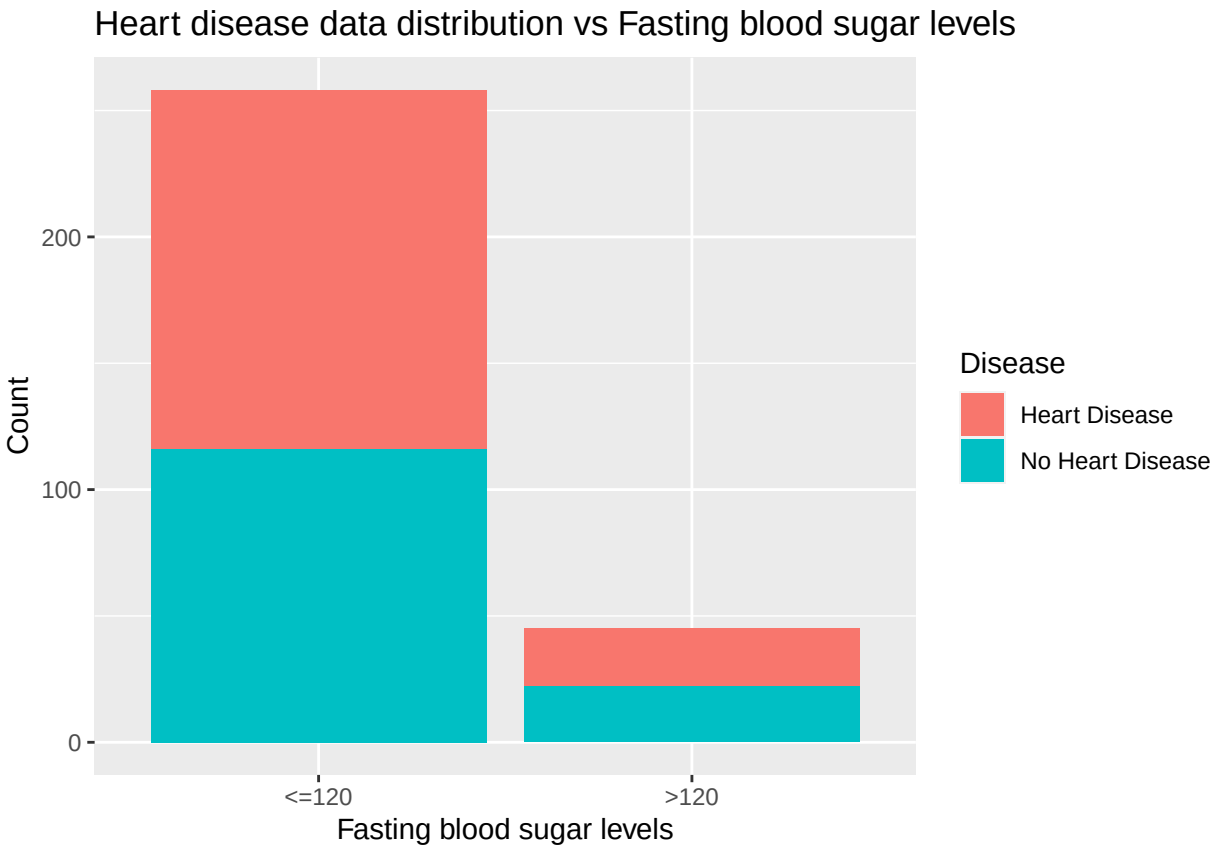
```
Heart_dataset %>%
  ggplot(aes(x=cp, fill = target))+
  geom_bar()+
  xlab("chest Pain Type") +
  ylab("Count")+
  guides(fill = guide_legend(title = "Disease"))+
  ggtitle("Heart disease data distribution vs Chest pain type")
```



Heart disease vs chest pain type – In the graph we can observe there are good number of heart disease patients with asymptomatic(without chest pain) condition and for the other two atypical and non-anginal pain, there are lot more patients with heart disease who are split almost evenly between the two..

Heart disease data distribution vs Fasting blood sugar levels

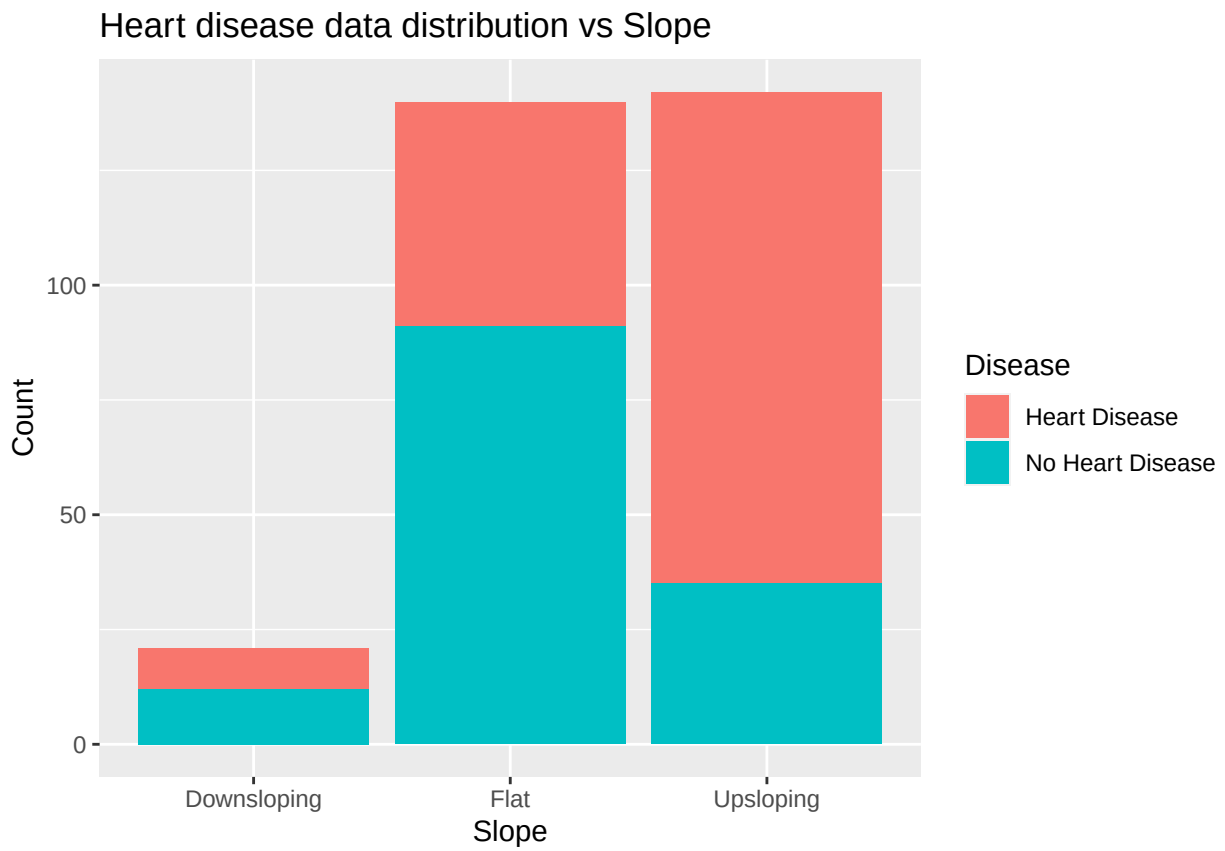
```
Heart_dataset %>%
  ggplot(aes(x=fbs, fill = target))+
  geom_bar()+
  xlab("Fasting blood sugar levels") +
  ylab("Count")+
  guides(fill = guide_legend(title = "Disease"))+
  ggtitle("Heart disease data distribution vs Fasting blood sugar levels")
```



Analysis of fasting blood sugar vs disease – fbs is a diabetes indicator with fbs >120 mg/d is considered diabetic. In the graph we observe that there are higher number of heart disease patient without diabetes.

Heart disease data distribution vs Slope

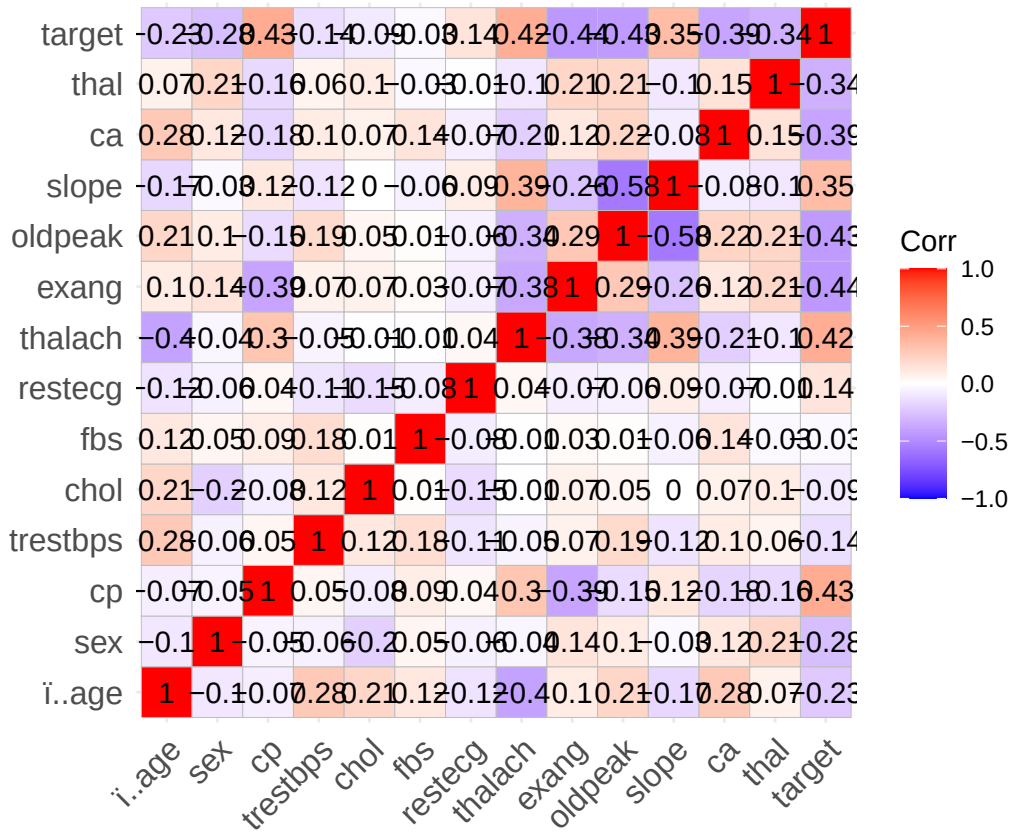
```
Heart_dataset %>%
  ggplot(aes(x=slope, fill = target))+
  geom_bar()+
  xlab("Slope") +
  ylab("Count")+
  guides(fill = guide_legend(title = "Disease"))+
  ggtitle("Heart disease data distribution vs Slope")
```



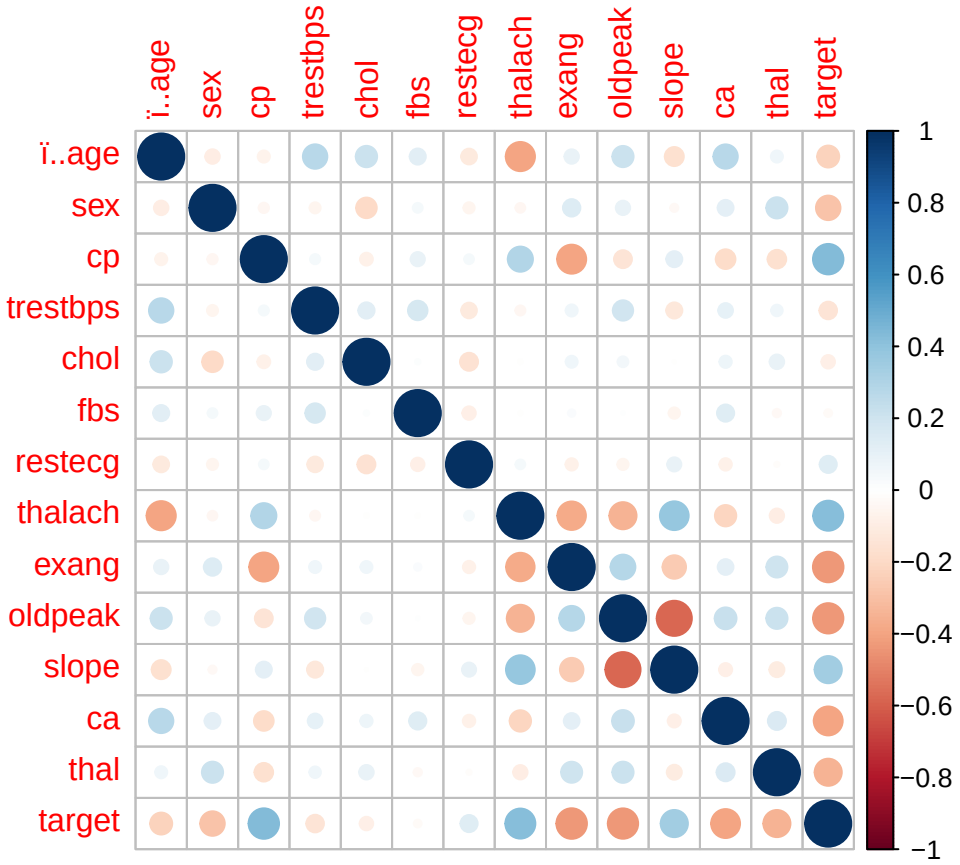
Heart disease vs slope – In the graph below we see total number of samples being fewer in downsloping case, in flat slope, we can see we have lot more patients without heart disease and upsloping shows the exact opposite behavior.

Correlation Matrix

```
correlation <- cor(Dataset[ , ])
ggcorrplot::ggcorrplot(correlation, lab = T)
```




```
corrplot::corrplot(correlation)
```



Correlation Matrix – From the matrix, thalach, cp, slope shows good positive correlation with target variable. Fbs, chol, trestbps, restecg has low correlation with our target variable.

Hypothesis Testing[one-sample] Claim 1: to test if a person has average BP value greater than 120 then it's most likely that person has heart disease.

Hypothesis: H0 (Null) : µBP <= 120
H1 (Alternate) : µBP > 120

```
total_rows <- nrow(Heart_dataset)
class(Heart_dataset$trestbps)
```

```
## [1] "integer"
```

```
Heart_dataset$trestbps <- as.numeric(Heart_dataset$trestbps)
class(Heart_dataset$trestbps)
```

```
## [1] "numeric"
```

```
null_variable <- 120

mean_sample <- replicate(100, mean(sample(Heart_dataset$trestbps, total_rows*0.70, replace = TRUE)))

sample_df <- data.frame(mean_sample, null_variable)

p_value <- mean(mean_sample >= null_variable)

t.test(Heart_dataset$trestbps, mu = null_variable, alternative = "greater")
```

```
##
```

```
##  One Sample t-test
##
## data:  Heart_dataset$trestbps
## t = 11.537, df = 302, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 120
## 95 percent confidence interval:
##  129.9614      Inf
## sample estimates:
## mean of x
##  131.6238
```

p-value <2.2e-16 Reject the Null hypothesis Conclusion: As the p value is than 0.05 we conclude that the average BP value is greater than 120, hence a person with BP value more than 120 is more likely to have heart disease.

Hypothesis Testing[Two_sample] Claim 2: to test if the average BP of a person with heart disease is greater than the average BP of the person without the heart disease. Hypothesis: H0 (Null) : µA – µB < = 0 H1 (Alternate) : µA - µB > 0

```
Heart_dataset$Age <- as.numeric(Heart_dataset$Age)
Heart_dataset$target <- as.numeric(Heart_dataset$target)
group_by(Heart_dataset, target)%>%
  summarise(count = n(),
            mean = mean(trestbps))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 3
##   target count  mean
##    <dbl> <int> <dbl>
## 1      1   165  129.
## 2      2   138  134.
```

```
Variance_test <- var.test(trestbps ~ target, data = Heart_dataset)
t.test(trestbps ~ target, data = Heart_dataset, var.equal = TRUE, alternative = "greater")
```

```
##
##  Two Sample t-test
##
## data:  trestbps by target
## t = -2.5413, df = 301, p-value = 0.9942
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -8.403782       Inf
## sample estimates:
## mean in group 1 mean in group 2
##        129.3030        134.3986
```

p-value = 0.9942 Fail to reject the null hypothesis Conclusion: As the p-value > 0.05 we failed to reject the hypothesis. So the claim of average BP of a person with heart disease is not necessarily greater than the average BP of the person with heart disease.

Data preprossing for anova creating a dataset with significant variables

```
Dataset1 <- Dataset %>%
  rename("Age"=ï..age)

Heart_disease1 <- Dataset1 %>%
```

```
  select(cp, chol, fbs, thalach, slope, target)
Heart_disease3 <- Dataset1 %>%
  select(cp, chol, fbs, thalach, slope, target, restecg)
```

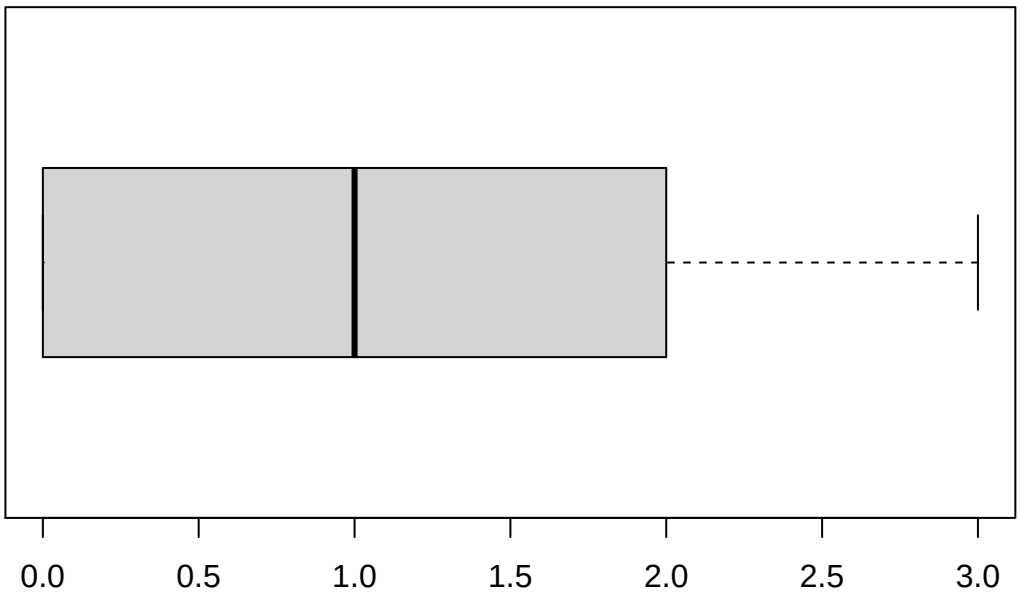
Checking for missing values No missing values found

```
sum(is.na(Heart_disease1))
```

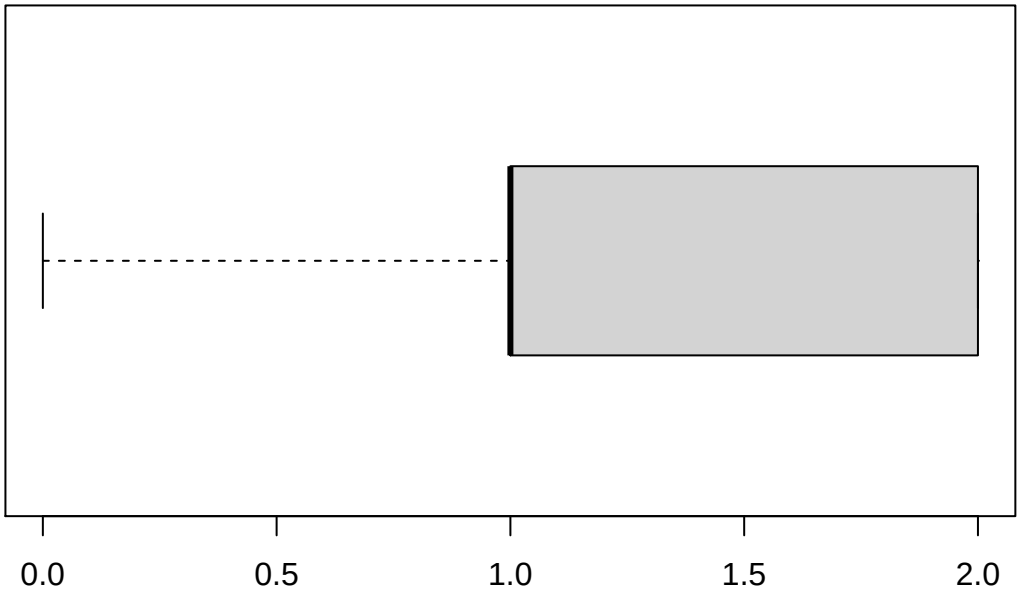
```
## [1] 0
```

Checking for Outliers in significant variables

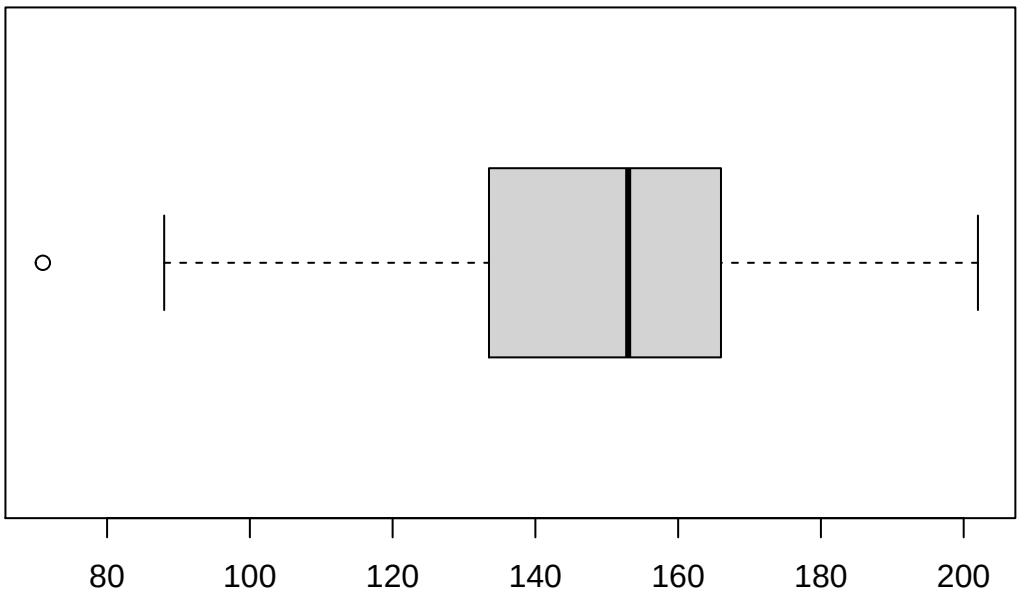
```
boxplot(Heart_disease1$cp, horizontal = TRUE)
```



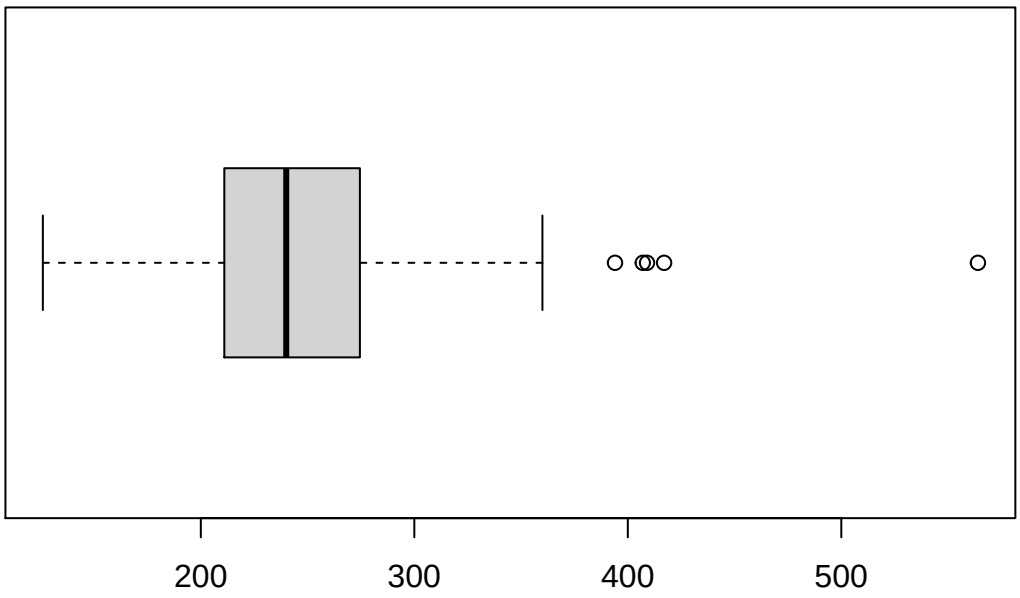

```
boxplot(Heart_disease1$slope, horizontal = TRUE)
```




```
boxplot(Heart_disease1$thalach, horizontal = TRUE)
```




```
boxplot(Heart_disease1$chol, horizontal = TRUE)
```



Checking for outliers in numerical variables with z-score method, grouped by whether the patient had heart disease

```
repl <- function(x) {replace(x, abs(scale(x))>3, mean(x))}
Heart_disease1$thalach <- ave(Heart_disease1$thalach, FUN = repl)
Heart_disease1 %>% group_by(target) %>%
  ungroup()  %>%
  select(cp) %>%
  scale() %>%
  abs() %>%
  (function(x) x>3) %>%
  sum()
```

```
## [1] 0
```

```
repl <- function(x) {replace(x, abs(scale(x))>3, mean(x))}
Heart_disease1$cp <- ave(Heart_disease1$cp, FUN = repl)
Heart_disease1 %>% group_by(target) %>%
  ungroup()  %>%
  select(cp) %>%
  scale() %>%
  abs() %>%
  (function(x) x>3) %>%
  sum()
```

```
## [1] 0
```

```
repl <- function(x) {replace(x, abs(scale(x))>3, mean(x))}
Heart_disease1$slope <- ave(Heart_disease1$slope, FUN = repl)
Heart_disease1 %>% group_by(target) %>%
  ungroup()  %>%
  select(slope) %>%
  scale() %>%
  abs() %>%
  (function(x) x>3) %>%
  sum()
```

```
## [1] 0
```

```
repl <- function(x) {replace(x, abs(scale(x))>3, mean(x))}
Heart_disease1$fbs <- ave(Heart_disease1$fbs, FUN = repl)
Heart_disease1 %>% group_by(target) %>%
  ungroup()  %>%
  select(fbs) %>%
  scale() %>%
  abs() %>%
  (function(x) x>3) %>%
  sum()
```

```
## [1] 0
```

```
repl <- function(x) {replace(x, abs(scale(x))>3, mean(x))}
Heart_disease1$chol <- ave(Heart_disease1$chol, FUN = repl)
Heart_disease1 %>% group_by(target) %>%
  ungroup()  %>%
  select(chol) %>%
  scale() %>%
  abs() %>%
  (function(x) x>3) %>%
  sum()
```

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

Replacing the outlier with mean

```
repl <- function(x) {replace(x, abs(scale(x))>3, mean(x))}
Heart_disease1$chol <- ave(Heart_disease1$chol, FUN = repl)
```

Checking that outliers were replaced

```
sum(abs(scale(Heart_disease1$chol))>3)
```

```
## [1] 0
```

Anova Hypothesis testing H0 (Null) : All means are equal H1 (Alternate) : atleast one or two means differ

```
ANOVA_1 <- aov(Heart_disease1$target ~ factor(Heart_disease1$cp))
summary(ANOVA_1)
```

```
##                             Df Sum Sq Mean Sq F value Pr(>F)
## factor(Heart_disease1$cp)    3  20.26   6.753   36.79 <2e-16 ***
## Residuals                  299  54.89   0.184
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
ANOVA_1b <- aov(Heart_disease1$thalach ~ factor(Heart_disease1$cp))
summary(ANOVA_1b)
```

```
##                             Df Sum Sq Mean Sq F value   Pr(>F)
## factor(Heart_disease1$cp)    3  22620    7540   17.39 1.95e-10 ***
## Residuals                  299 129617     434
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value is <0.05 which tells us that three gropus of chest pain means are different.

Regression Analysis

```r
model <- glm(target ~ ., data = Heart_disease1,family = binomial(link = "logit"))
model %>% summary()
```

```
##
## Call:
## glm(formula = target ~ ., family = binomial(link = "logit"),
##     data = Heart_disease1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5467  -0.7101   0.3723   0.8137   2.2061
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.733107   1.274436  -3.714 0.000204 ***
## cp           0.866167   0.148644   5.827 5.64e-09 ***
## chol        -0.006059   0.003321  -1.825 0.068033 .
## fbs         -0.396078   0.402772  -0.983 0.325420
## thalach      0.029093   0.007403   3.930 8.50e-05 ***
## slope        0.930271   0.244510   3.805 0.000142 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 303.28  on 297  degrees of freedom
## AIC: 315.28
##
## Number of Fisher Scoring iterations: 4
```

```r
model
```

```
##
## Call:  glm(formula = target ~ ., family = binomial(link = "logit"),
##     data = Heart_disease1)
##
## Coefficients:
## (Intercept)           cp         chol          fbs      thalach        slope
##   -4.733107     0.866167    -0.006059    -0.396078     0.029093     0.930271
##
## Degrees of Freedom: 302 Total (i.e. Null);  297 Residual
## Null Deviance:       417.6
```

```
## Residual Deviance: 303.3     AIC: 315.3
```

```
predict <- predict(model, type = "response")
prop.table(table(Heart_disease1$target, predict > 0.5))
```

```
##
##         FALSE      TRUE
##   0 0.3201320 0.1353135
##   1 0.1056106 0.4389439
```

p values for the significant variables are lower than 0.05 F1 score for the model is 80%