RANDOM FOREST

```
install.packages("mice")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```
install.packages("randomForest")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```
install.packages("cowplot")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

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

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```
library(randomForest)
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
library(dplyr)
```

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

```
## The following object is masked from 'package:randomForest':
##
##     combine
```

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

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

```
library(lattice)
library(caret)
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'ggplot2'

## The following object is masked from 'package:randomForest':
##
##     margin
```

```
library(cowplot)
library(caTools)
library(ggplot2)
library(randomForest)
library(mice)
```

```
##
## Attaching package: 'mice'

## The following object is masked from 'package:stats':
##
##     filter

## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
library(reshape2)
library(ggcorrplot)
```

```
df=read.csv("heart_disease_uci.csv")
```

```
#Structure of the datasset
str(df)
```

```
## 'data.frame':    920 obs. of  16 variables:
##  $ id      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ age     : int  63 67 67 37 41 56 62 57 63 53 ...
##  $ sex     : chr  "Male" "Male" "Male" "Male" ...
##  $ dataset : chr  "Cleveland" "Cleveland" "Cleveland" "Cleveland" ...
##  $ cp      : chr  "typical angina" "asymptomatic" "asymptomatic" "no
n-anginal" ...
##  $ trestbps: int  145 160 120 130 130 120 140 120 130 140 ...
##  $ chol    : int  233 286 229 250 204 236 268 354 254 203 ...
##  $ fbs     : logi  TRUE FALSE FALSE FALSE FALSE FALSE ...
##  $ restecg : chr  "lv hypertrophy" "lv hypertrophy" "lv hypertrophy"
 "normal" ...
##  $ thalch  : int  150 108 129 187 172 178 160 163 147 155 ...
##  $ exang   : logi  FALSE TRUE TRUE FALSE FALSE FALSE ...
##  $ oldpeak : num  2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...
##  $ slope   : chr  "downsloping" "flat" "flat" "downsloping" ...
##  $ ca      : int  0 3 2 0 0 0 2 0 1 0 ...
##  $ thal    : chr  "fixed defect" "normal" "reversable defect" "norma
```

```
l" ...
##  $ num     : int  0 2 1 0 0 0 3 0 2 1 ...

#Summary
summary(df)

##        id              age             sex                dataset
##  Min.   :  1.0   Min.   :28.00   Length:920         Length:920
##  1st Qu.:230.8   1st Qu.:47.00   Class :character   Class :character
##  Median :460.5   Median :54.00   Mode  :character   Mode  :character
##  Mean   :460.5   Mean   :53.51
##  3rd Qu.:690.2   3rd Qu.:60.00
##  Max.   :920.0   Max.   :77.00
##
##       cp                trestbps          chol          fbs
##  Length:920         Min.   :  0.0   Min.   :  0.0   Mode :logical
##  Class :character   1st Qu.:120.0   1st Qu.:175.0   FALSE:692
##  Mode  :character   Median :130.0   Median :223.0   TRUE :138
##                     Mean   :132.1   Mean   :199.1   NA's :90
##                     3rd Qu.:140.0   3rd Qu.:268.0
##                     Max.   :200.0   Max.   :603.0
##                     NA's   :59      NA's   :30
##    restecg             thalch        exang            oldpeak
##  Length:920         Min.   : 60.0   Mode :logical   Min.   :-2.6000
##  Class :character   1st Qu.:120.0   FALSE:528       1st Qu.: 0.0000
##  Mode  :character   Median :140.0   TRUE :337       Median : 0.5000
##                     Mean   :137.5   NA's :55        Mean   : 0.8788
##                     3rd Qu.:157.0                   3rd Qu.: 1.5000
##                     Max.   :202.0                   Max.   : 6.2000
##                     NA's   :55                      NA's   :62
##     slope                ca             thal                num
##  Length:920         Min.   :0.0000   Length:920         Min.   :0.00
```

```
00
##  Class :character   1st Qu.:0.0000   Class :character   1st Qu.:0.00
00
##  Mode  :character   Median :0.0000   Mode  :character   Median :1.00
00
##                     Mean   :0.6764                      Mean   :0.99
57
##                     3rd Qu.:1.0000                      3rd Qu.:2.00
00
##                     Max.   :3.0000                      Max.   :4.00
00
##                     NA's   :611
```

```
#Dimension
dim(df)
```

```
## [1] 920  16
```

```
#Checking missing values
sum(is.na(df))
```

```
## [1] 962
```

```
colSums(is.na(df))
```

```
##       id      age      sex  dataset       cp trestbps     chol
fbs
##        0        0        0        0        0       59       30
 90
##  restecg   thalch    exang  oldpeak    slope       ca     thal
num
##        0       55       55       62        0      611        0
  0
```

```
# handling missing value
df=complete(mice(df, method = "cart"))
```

```
##
##  iter imp variable
##   1   1  trestbps  chol  fbs  thalch  exang  oldpeak  ca
##   1   2  trestbps  chol  fbs  thalch  exang  oldpeak  ca
##   1   3  trestbps  chol  fbs  thalch  exang  oldpeak  ca
##   1   4  trestbps  chol  fbs  thalch  exang  oldpeak  ca
##   1   5  trestbps  chol  fbs  thalch  exang  oldpeak  ca
##   2   1  trestbps  chol  fbs  thalch  exang  oldpeak  ca
##   2   2  trestbps  chol  fbs  thalch  exang  oldpeak  ca
##   2   3  trestbps  chol  fbs  thalch  exang  oldpeak  ca
##   2   4  trestbps  chol  fbs  thalch  exang  oldpeak  ca
##   2   5  trestbps  chol  fbs  thalch  exang  oldpeak  ca
##   3   1  trestbps  chol  fbs  thalch  exang  oldpeak  ca
##   3   2  trestbps  chol  fbs  thalch  exang  oldpeak  ca
##   3   3  trestbps  chol  fbs  thalch  exang  oldpeak  ca
```

```
##   3   4  trestbps  chol  fbs  thalch  exang  oldpeak  ca
##   3   5  trestbps  chol  fbs  thalch  exang  oldpeak  ca
##   4   1  trestbps  chol  fbs  thalch  exang  oldpeak  ca
##   4   2  trestbps  chol  fbs  thalch  exang  oldpeak  ca
##   4   3  trestbps  chol  fbs  thalch  exang  oldpeak  ca
##   4   4  trestbps  chol  fbs  thalch  exang  oldpeak  ca
##   4   5  trestbps  chol  fbs  thalch  exang  oldpeak  ca
##   5   1  trestbps  chol  fbs  thalch  exang  oldpeak  ca
##   5   2  trestbps  chol  fbs  thalch  exang  oldpeak  ca
##   5   3  trestbps  chol  fbs  thalch  exang  oldpeak  ca
##   5   4  trestbps  chol  fbs  thalch  exang  oldpeak  ca
##   5   5  trestbps  chol  fbs  thalch  exang  oldpeak  ca

## Warning: Number of logged events: 6

df=subset(df, select = -ca)
colSums(is.na(df))

##       id      age      sex  dataset       cp trestbps     chol
fbs
##        0        0        0        0        0        0        0
  0
##  restecg   thalch    exang  oldpeak    slope     thal      num
##        0        0        0        0        0        0        0

df$dataset=as.factor(df$dataset)
df$restecg=as.factor(df$restecg)
df$thal=as.factor(df$thal)
df$num=as.factor(df$num)
df$slope=as.factor(df$slope)
df$exang=as.factor(df$exang)
df$fbs=as.factor(df$fbs)
df$sex=as.factor(df$sex)
df$cp=as.factor(df$cp)
summary(df)

##        id              age            sex                dataset
##  Min.   :  1.0   Min.   :28.00   Female:194   Cleveland    :304
##  1st Qu.:230.8   1st Qu.:47.00   Male  :726   Hungary      :293
##  Median :460.5   Median :54.00                Switzerland  :123
##  Mean   :460.5   Mean   :53.51                VA Long Beach:200
##  3rd Qu.:690.2   3rd Qu.:60.00
##  Max.   :920.0   Max.   :77.00
##                cp          trestbps          chol       fbs
##  asymptomatic   :496   Min.   :  0.0   Min.   :  0.0   0:768
##  atypical angina:174   1st Qu.:120.0   1st Qu.:177.8   1:152
##  non-anginal    :204   Median :130.0   Median :224.0
##  typical angina : 46   Mean   :132.2   Mean   :201.6
##                        3rd Qu.:140.5   3rd Qu.:269.0
##                        Max.   :200.0   Max.   :603.0
##              restecg        thalch      exang      oldpeak
```

```
##                  :  2   Min.   : 60.0   0:544   Min.   :-2.6000
##  lv hypertrophy  :188   1st Qu.:120.0   1:376   1st Qu.: 0.0000
##  normal          :551   Median :140.0           Median : 0.6000
##  st-t abnormality:179   Mean   :137.2           Mean   : 0.9052
##                         3rd Qu.:157.0           3rd Qu.: 1.5250
##                         Max.   :202.0           Max.   : 6.2000
##          slope                      thal     num
##             :309                      :486   0:411
##  downsloping: 63   fixed defect      : 46   1:265
##  flat       :345   normal            :196   2:109
##  upsloping  :203   reversable defect:192   3:107
##                                             4: 28
##
```

```
str(df)
```

```
## 'data.frame':    920 obs. of  15 variables:
##  $ id      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ age     : int  63 67 67 37 41 56 62 57 63 53 ...
##  $ sex     : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 2 1 1 2 2
 ...
##  $ dataset : Factor w/ 4 levels "Cleveland","Hungary",..: 1 1 1 1 1
1 1 1 1 1 ...
##  $ cp      : Factor w/ 4 levels "asymptomatic",..: 4 1 1 3 2 2 1 1 1
 1 ...
##  $ trestbps: num  145 160 120 130 130 120 140 120 130 140 ...
##  $ chol    : num  233 286 229 250 204 236 268 354 254 203 ...
##  $ fbs     : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 1 2 ...
##  $ restecg : Factor w/ 4 levels "","lv hypertrophy",..: 2 2 2 3 2 3
2 3 2 2 ...
##  $ thalch  : num  150 108 129 187 172 178 160 163 147 155 ...
##  $ exang   : Factor w/ 2 levels "0","1": 1 2 2 1 1 1 1 2 1 2 ...
##  $ oldpeak : num  2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...
##  $ slope   : Factor w/ 4 levels "","downsloping",..: 2 3 3 2 4 4 2 4
 3 2 ...
##  $ thal    : Factor w/ 4 levels "","fixed defect",..: 2 3 4 3 3 3 3
3 4 4 ...
##  $ num     : Factor w/ 5 levels "0","1","2","3",..: 1 3 2 1 1 1 4 1
3 2 ...
```

```
# Univariate analysis
ggplot(df,aes(x=age,fill=sex))+geom_histogram(col="white")+labs(title="
Distribution of Age")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

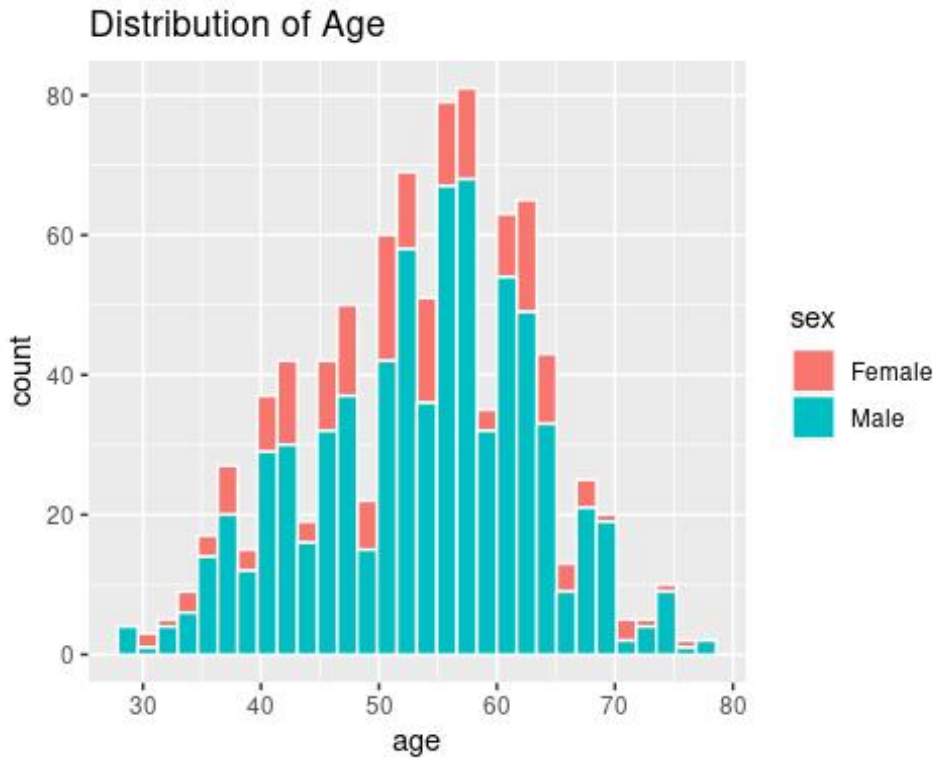


Fig 1.1

```
ggplot(df,aes(x=chol,fill=sex))+geom_histogram(col="white")+labs(title=
"Distribution of cholestrol mg/dl")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

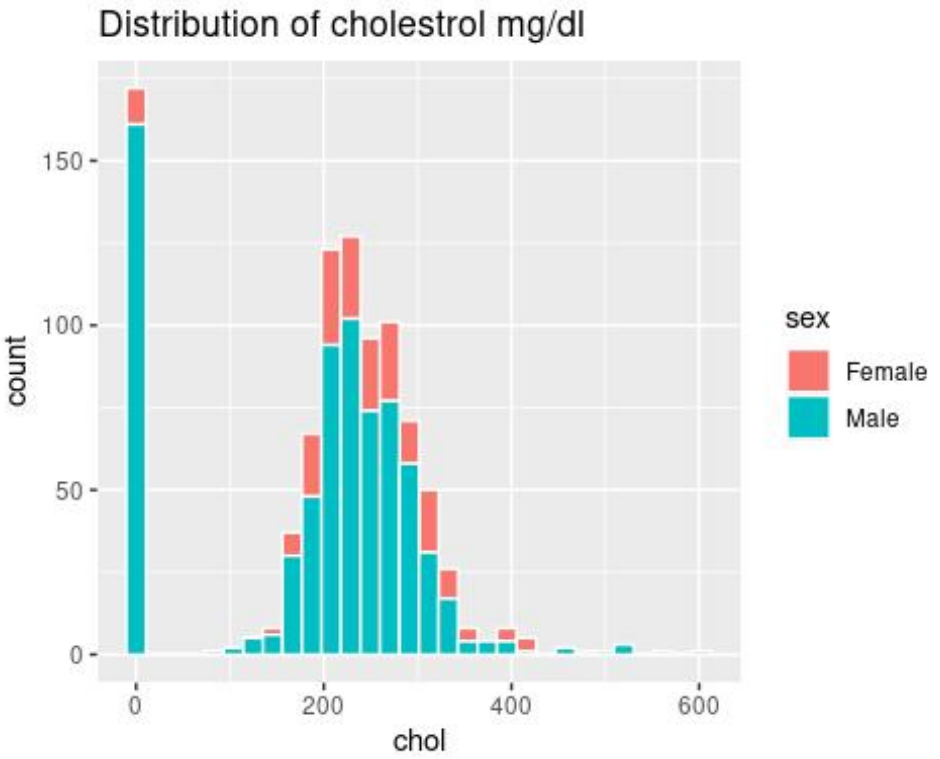


Fig 1.2

```
ggplot(df,aes(x=thalch,fill=sex))+geom_histogram(col="white")+labs(titl
e="Distribution of maximum heart rate ")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

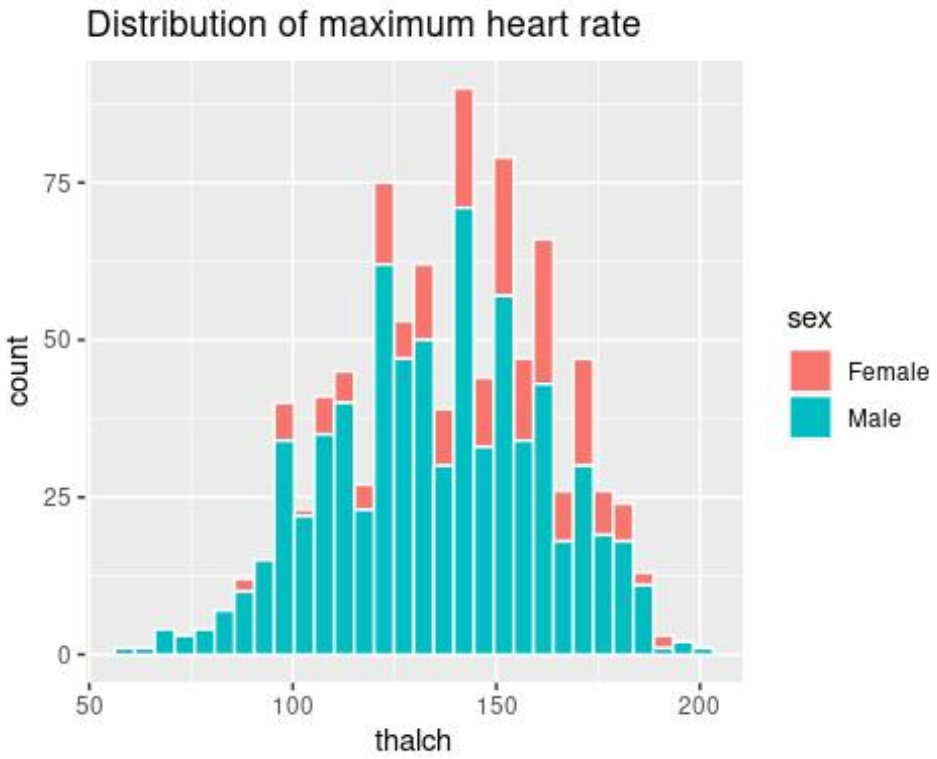


Fig 1.3

```
ggplot(df,aes(x=trestbps,fill=sex))+geom_histogram(col="white")+labs(ti
tle="Distribution of resting blood pressure")+xlim(75,200)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 2 rows containing non-finite outside the scale rang
e
## (`stat_bin()`).

## Warning: Removed 4 rows containing missing values or values outside
the scale range
## (`geom_bar()`).
```

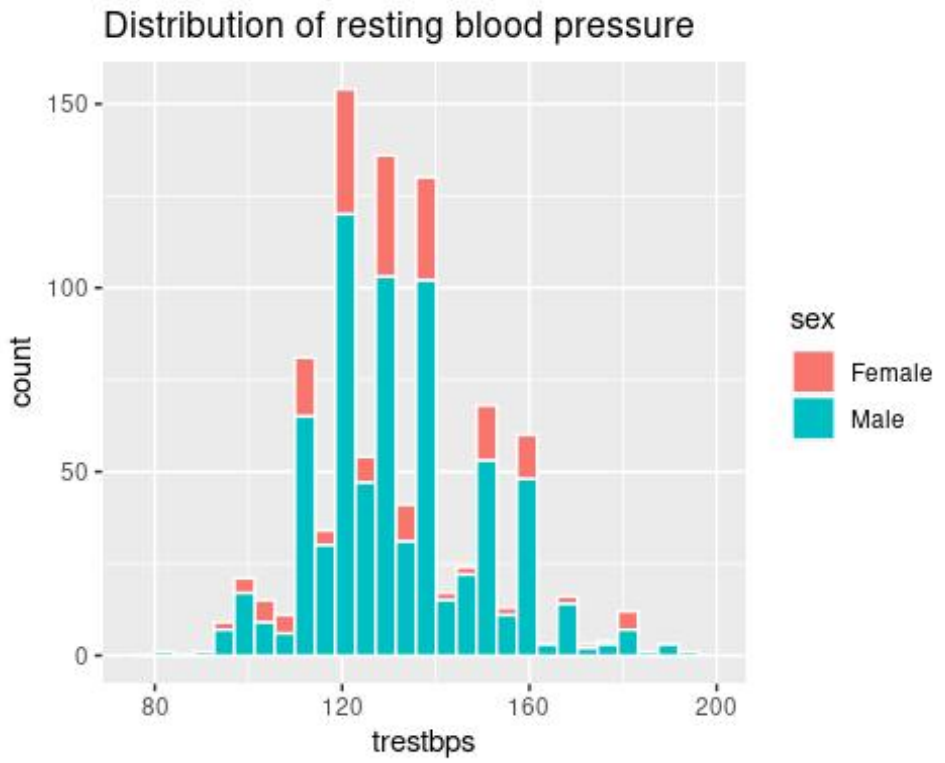


Fig 1.4

```r
# Bivariate analysis
ggplot(df,aes(x=cp))+geom_bar()+facet_wrap(~sex)+
  labs(title="Distribution of sex vs chest pain type")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

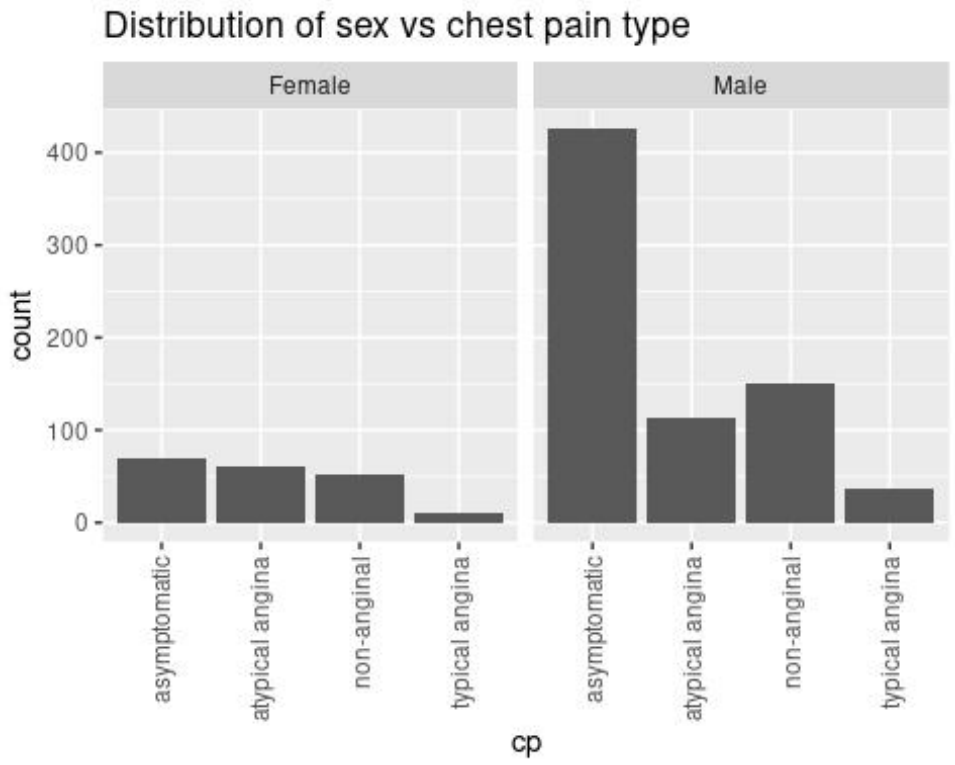


Fig 1.5

```
ggplot(df,aes(x=num,y=age))+geom_boxplot()+facet_wrap(~sex)+
  labs(title="Types of heart disease based on age and gender")
```

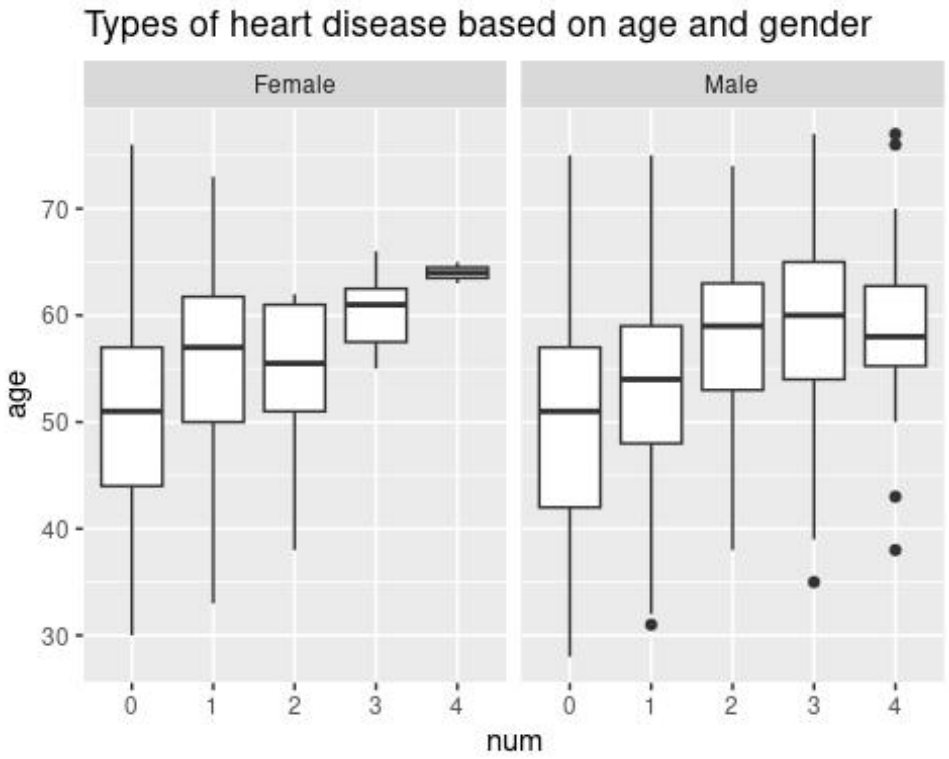


Fig 1.6

```
ggplot(df,aes(x=num))+geom_bar()+facet_wrap(~cp)+labs(title = "Distribu
tion of chest pain type with type of heart disease")
```

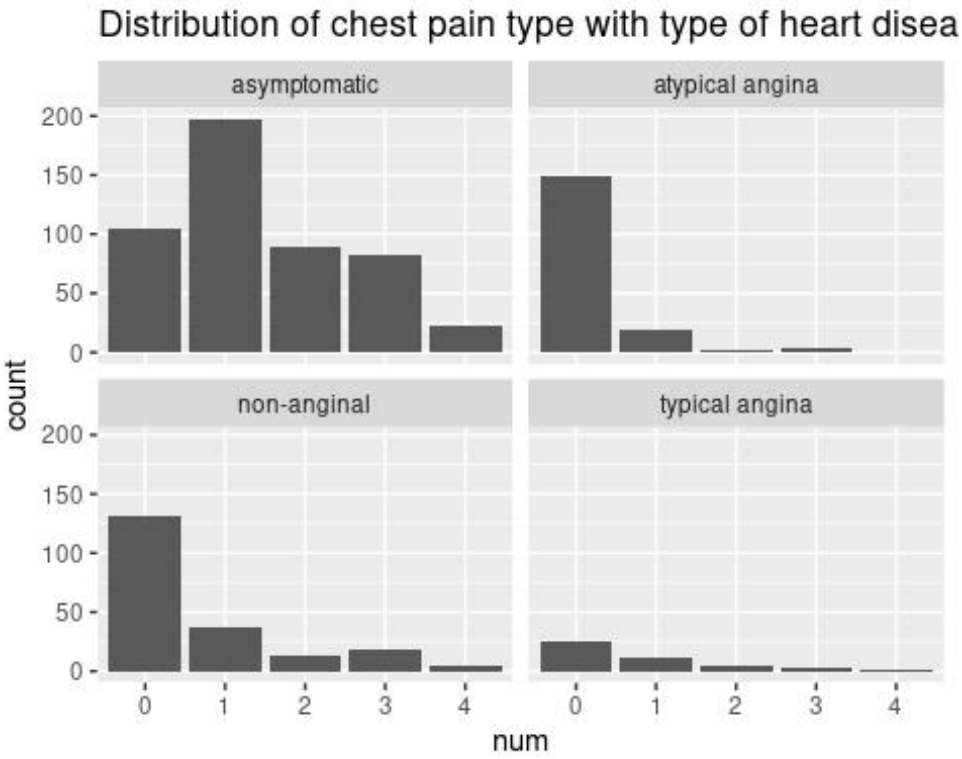


Fig 1.7

```
df%>%
  filter(trestbps>120&chol>100)%>%
  ggplot(aes(x=chol,y=trestbps))+geom_point()+labs("Relation btween res
ting BP and Cholestrol")
```

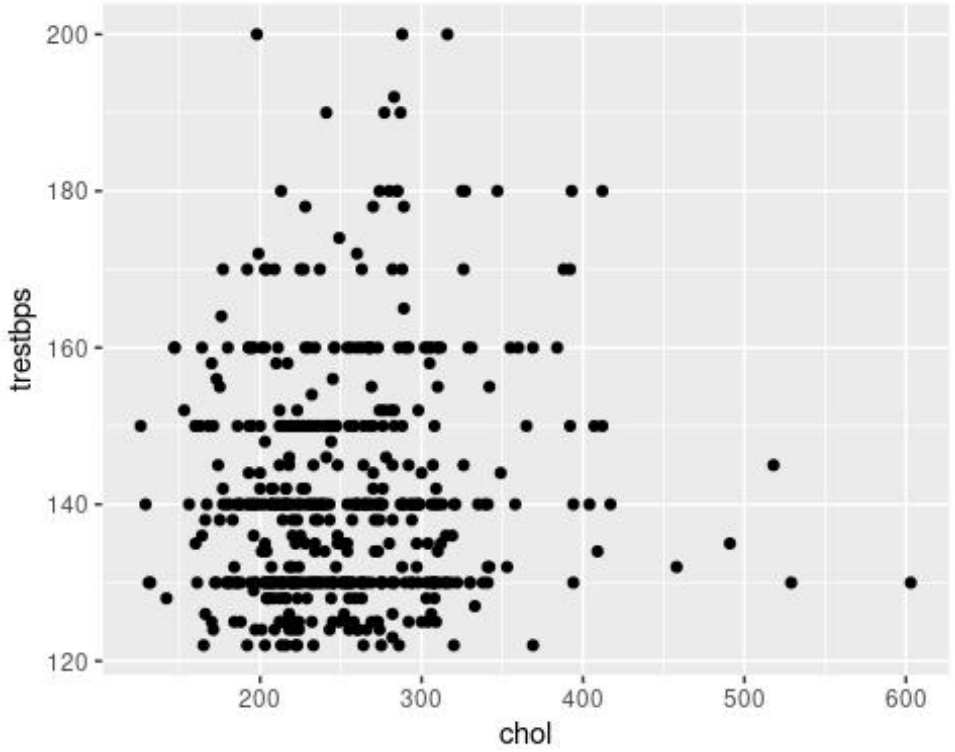


Fig 1.8

```
df%>%
  select(num,chol,trestbps,sex)%>%
  filter(trestbps>120)%>%
  ggplot(aes(x=num,y=trestbps))+geom_boxplot()+
  labs(title="Types of heart disease based on resting blood pressure")
```

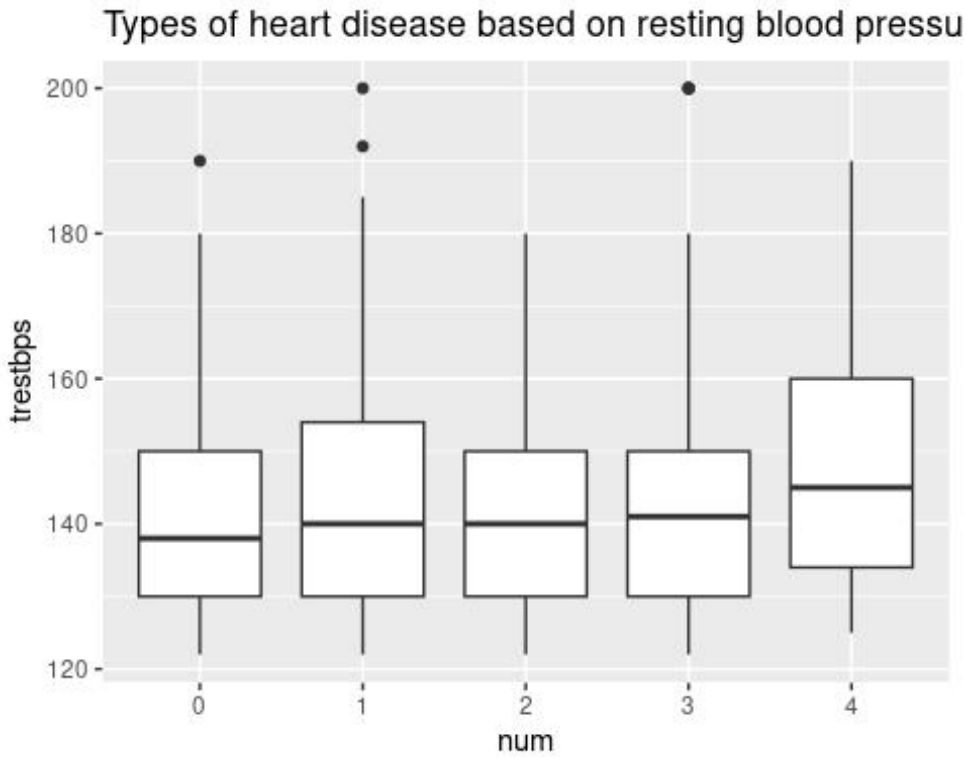


Fig 1.9

```
ggplot(df,aes(x=num))+geom_bar()+facet_wrap(~restecg)+
  labs(title="Types of heart disease based ECG result")
```

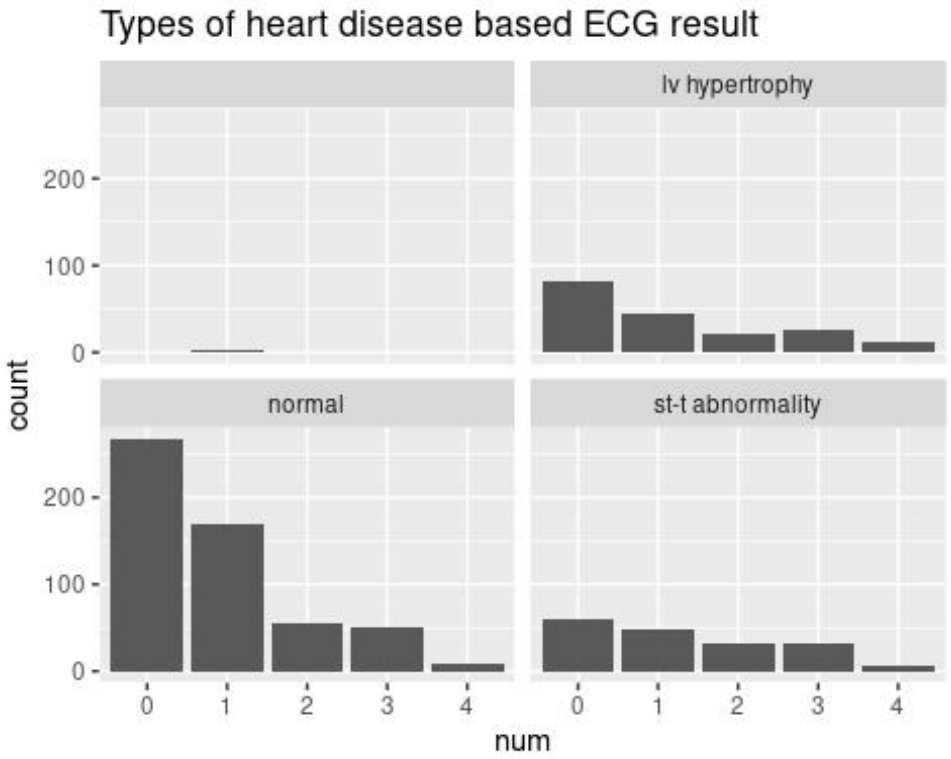


Fig 1.10

```
ggplot(df,aes(x=num))+geom_bar()+facet_wrap(~exang)+
  labs(title="Types of heart disease based on resting blood pressure")
```

Fig 1.11

```
#multivariate analysis
df[sapply(df, is.factor)] <- data.matrix(df[sapply(df, is.factor)])
data=cor(df[sapply(df, is.numeric)])
data1= melt(data)
ggcorrplot(data, hc.order = TRUE,lab = TRUE)
```

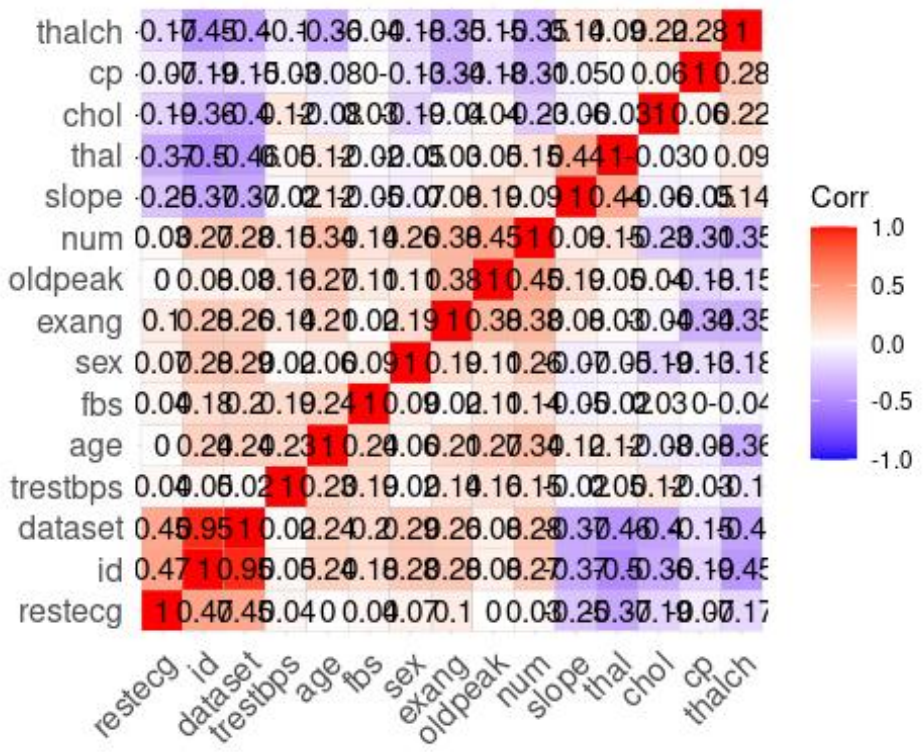


Fig 1.12

```
# Parametric test
# ANOVA test between Na_to_K
aov <- aov(df$age~factor(df$num))
p_value <- summary(aov)[[1]]$`Pr(>F)`[1]
if (p_value < 0.05) {
  # If p-value is significant, print ANOVA summary
  print("ANOVA is significant:")
} else {
  # If p-value is not significant, print a message
  print("ANOVA is not significant.")
}
```

```
## [1] "ANOVA is significant:"
```

```
data=table(df$cp,df$num)
test=chisq.test(data)
```

```
## Warning in chisq.test(data): Chi-squared approximation may be incorr
ect
```

```r
p_value_chi_square <- test$p.value
# Check if p-value is less than significance level (e.g., 0.05)
if (p_value_chi_square < 0.05) {
  # If p-value is significant, print chi-square test summary
  print("Chi-square test is significant:")
} else {
  # If p-value is not significant, print a message
  print("Chi-square test is not significant.")
}
```

```
## [1] "Chi-square test is significant:"
```

```r
aov <- aov(df$trestbps~factor(df$num))
p_value <- summary(aov)[[1]]$`Pr(>F)`[1]
if (p_value < 0.05) {
  # If p-value is significant, print ANOVA summary
  print("ANOVA is significant:")
} else {
  # If p-value is not significant, print a message
  print("ANOVA is not significant.")
}
```

```
## [1] "ANOVA is significant:"
```

```r
aov <- aov(df$chol~factor(df$num))
p_value <- summary(aov)[[1]]$`Pr(>F)`[1]
if (p_value < 0.05) {
  # If p-value is significant, print ANOVA summary
  print("ANOVA is significant:")
} else {
  # If p-value is not significant, print a message
  print("ANOVA is not significant.")
}
```

```
## [1] "ANOVA is significant:"
```

```r
data=table(df$sex,df$num)
test=chisq.test(data)
p_value_chi_square <- test$p.value
# Check if p-value is less than significance level (e.g., 0.05)
if (p_value_chi_square < 0.05) {
  # If p-value is significant, print chi-square test summary
  print("Chi-square test is significant:")
} else {
  # If p-value is not significant, print a message
  print("Chi-square test is not significant.")
}
```

```
## [1] "Chi-square test is significant:"
```

```r
data=table(df$restecg,df$num)
test=chisq.test(data)
```

```
## Warning in chisq.test(data): Chi-squared approximation may be incorr
ect
```

```
p_value_chi_square <- test$p.value
# Check if p-value is less than significance level (e.g., 0.05)
if (p_value_chi_square < 0.05) {
  # If p-value is significant, print chi-square test summary
  print("Chi-square test is significant:")
} else {
  # If p-value is not significant, print a message
  print("Chi-square test is not significant.")
}
```

```
## [1] "Chi-square test is significant:"
```

```
data=table(df$exang,df$num)
test=chisq.test(data)
p_value_chi_square <- test$p.value
# Check if p-value is less than significance level (e.g., 0.05)
if (p_value_chi_square < 0.05) {
  # If p-value is significant, print chi-square test summary
  print("Chi-square test is significant:")
} else {
  # If p-value is not significant, print a message
  print("Chi-square test is not significant.")
}
```

```
## [1] "Chi-square test is significant:"
```

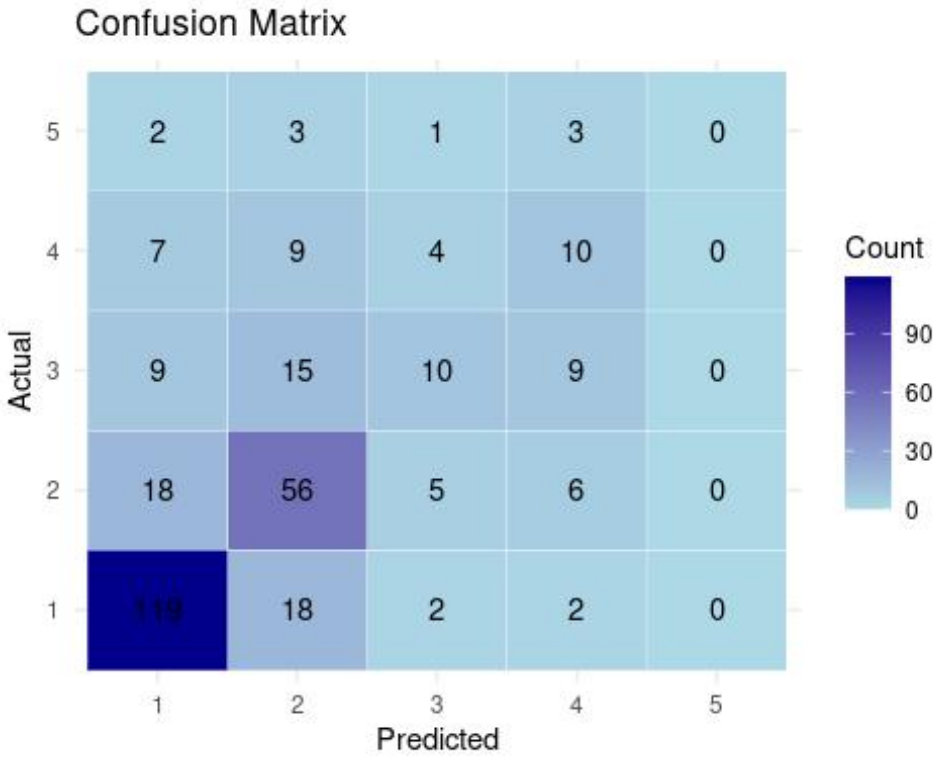
```
aov <- aov(df$thalch~factor(df$num))
p_value <- summary(aov)[[1]]$`Pr(>F)`[1]
if (p_value < 0.05) {
  # If p-value is significant, print ANOVA summary
  print("ANOVA is significant:")
} else {
  # If p-value is not significant, print a message
  print("ANOVA is not significant.")
}
```

```
## [1] "ANOVA is significant:"
```

```
# Build the Random Forest Model
df1=df%>%
  select(age,sex,cp,restecg,chol,exang,thalch,trestbps,num)
df1[sapply(df, is.factor)] <- data.matrix(df1[sapply(df, is.factor)])
# Train-Test Split
set.seed(123)
sample_data= sample.split(df, SplitRatio = 0.7)
train_data <- subset(df, sample_data == TRUE)
test_data <- subset(df, sample_data == FALSE)
train_data
```

```
##      id age sex dataset cp trestbps chol fbs restecg thalch exang ol
dpeak slope
## 1     1  63   2       1  4      145  233   2       2    150     1
  2.3     2
## 3     3  67   2       1  1      120  229   1       2    129     2
  ##     thal num
## 1      2   1
## 3      4   2
train_data$num <- as.factor(train_data$num)
rf_model <- randomForest(num~ ., data = train_data, ntree = 500)
fin = predict(rf_model, test_data)
#Evaluation metrics
table=table(test_data$num,fin)
acctest=sum(diag(table))/sum(table)
acctest

## [1] 0.6331169

#confusion matrix
conf_matrix=confusionMatrix(factor(fin), factor(test_data$num))

## Warning in levels(reference) != levels(data): longer object length i
s not a
## multiple of shorter object length

## Warning in confusionMatrix.default(factor(fin), factor(test_data$nu
m)): Levels
## are not in the same order for reference and data. Refactoring data t
o match.

cm_data <- as.data.frame(conf_matrix$table)
names(cm_data) <- c('Predicted', 'Actual', 'Count')
ggplot(cm_data, aes(x = Predicted, y = Actual, fill = Count)) +
  geom_tile(color = "white") +
  geom_text(aes(label = Count)) +
  theme_minimal() +
  scale_fill_gradient(low = "lightblue", high = "darkblue") +
  labs(title = "Confusion Matrix",
       x = "Predicted",
       y = "Actual")
```

Fig 1.13


Confusion Matrix
Actual
Predicted
Count
90
60
30
0
1
2
3
4
5
2 3 1 3 0
7 9 4 10 0
9 15 10 9 0
18 56 5 6 0
119 18 2 2 0