SRI RAMACHANDRA
INSTITUTE OF HIGHER EDUCATION AND RESEARCH
(Category - I Deemed to be University) Porur, Chennai
SRI RAMACHANDRA ENGINEERING AND TECHNOLOGY

CSE-320 DATA MINING

CA-4

Submitted to

SRI RAMACHANDRA INSTITUTE OF HIGHER EDUCATION AND RESEARCH
SRI RAMACHANDRA ENGINEERING AND TECHNOLOGY

for the Award of the Degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

(Cyber Security and Internet of Things)

Shriram K.P (E0219007)
Umesh Kumar (E0219019)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SRET,PORUR, CHENNAI- 600116
SEPTEMBER 2021

INTRODUCTION:

Heart Dataset is from Kagle and has various factors and measures in heart. It has 303 rows and 14 col.

This report provides an analysis and evaluation of the factors that causes heart disease and how much prone it is if some of the factors is misleading and improper. So 14 different factors with proper information in dataset will help us to derive different types of conclusions.

AIM:

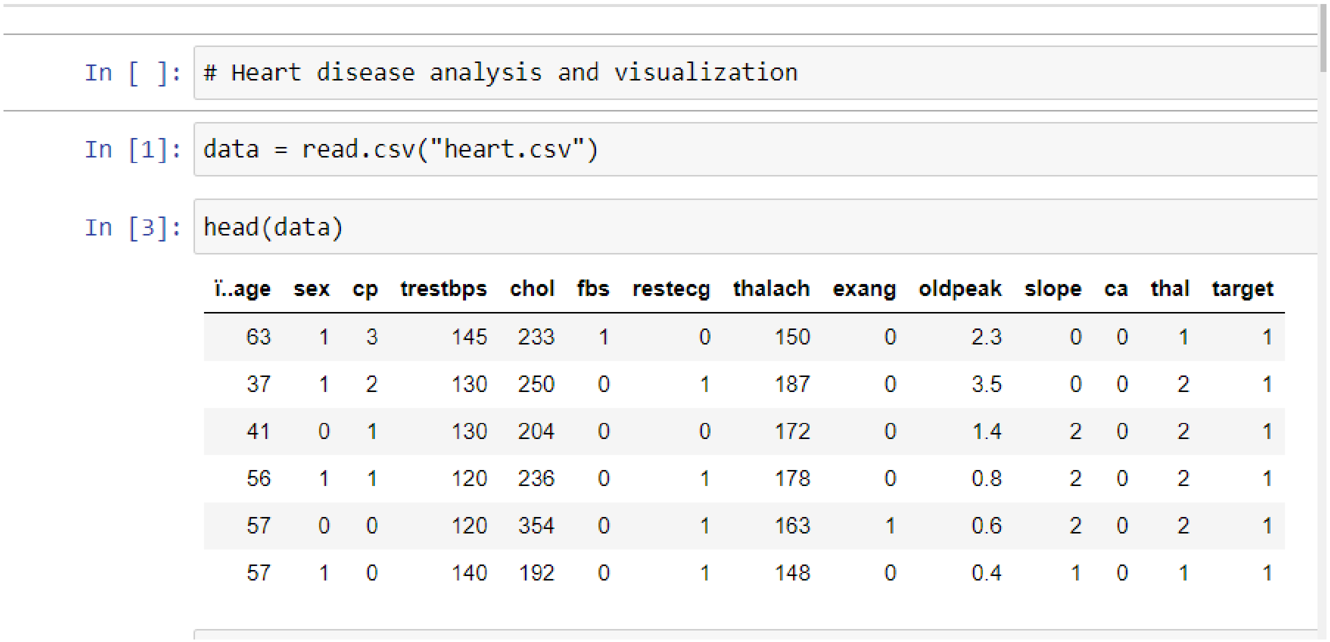
The goal of the project is to predict whether a person is prone to heart disease and why he or she having heart disease

PROBLEM STATEMENT:

The dataset had 14 variables which are listed in the later part of the report. Using R programming and powerful libraries like 〝tidyverse〞 is used to do proper analysis of dataset and derive proper conclusion.

ATTRIBUTES IN DATASET :

1. **AGE :** Age of the person
2. **SEX :** Gender of the person took test
3. **CP :** Chest pain type (0 – 4)
4. **TRESTPBS :** Resting blood pressure (blood pressure at resting position)
5. **CHOL :** Cholesterol Level
6. **FBS :** Fasting Blood Sugar (Sugar level after fasting) 120 > fbs – 0 120 < fbs - 1
7. **RESTECG :** Resting ElectroCardiographic result (0,1,2)
8. **THALACH :** Maximum heart rate achieved (Beats per minute)
9. **EXANG :** Exercise induced angina (exercise induced chest pain)
10. **OLDPEAK :** A measure of abnormality in electrocardiograms.
11. **SLOPE :** Quality of blood flow to heart
12. **CA :** Cardiography Results
13. **THAL :** Thallium stress test measuring blood flow to heart.
14. **TARGET :** Having heart disease or not (0 – negative 1 – positive)



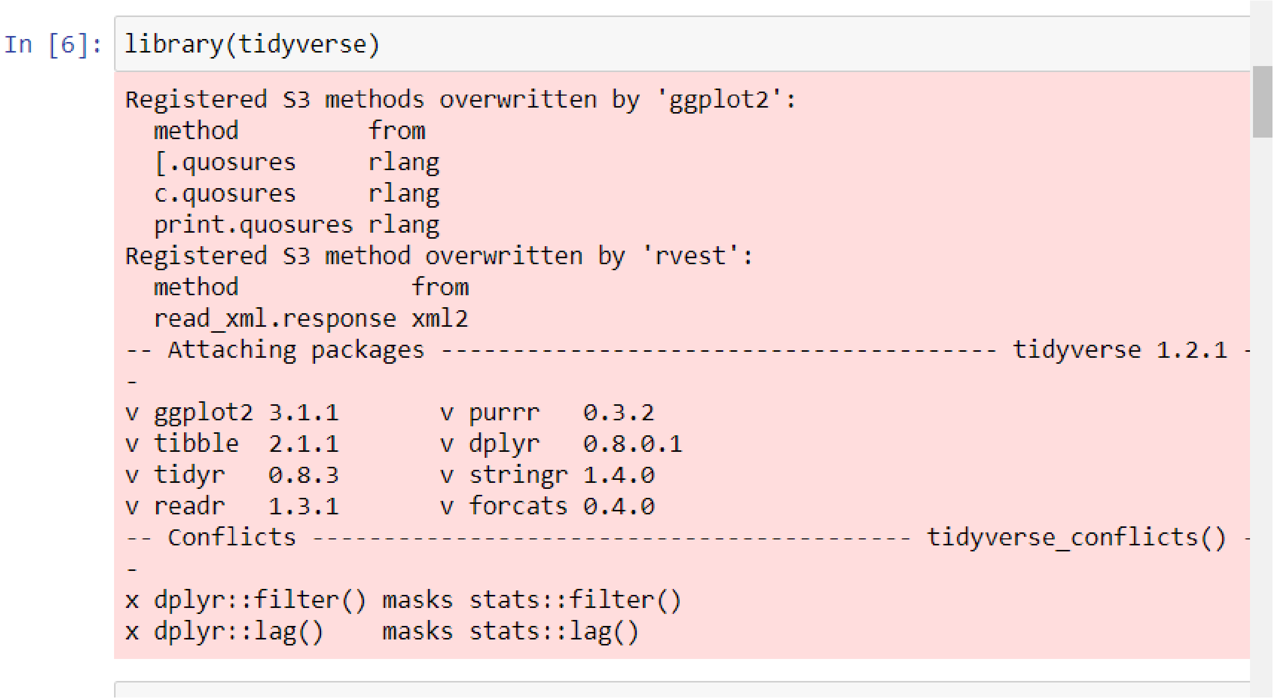
```
In [ ]: # Heart disease analysis and visualization
```

```
In [1]: data = read.csv("heart.csv")
```

```
In [3]: head(data)
```

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

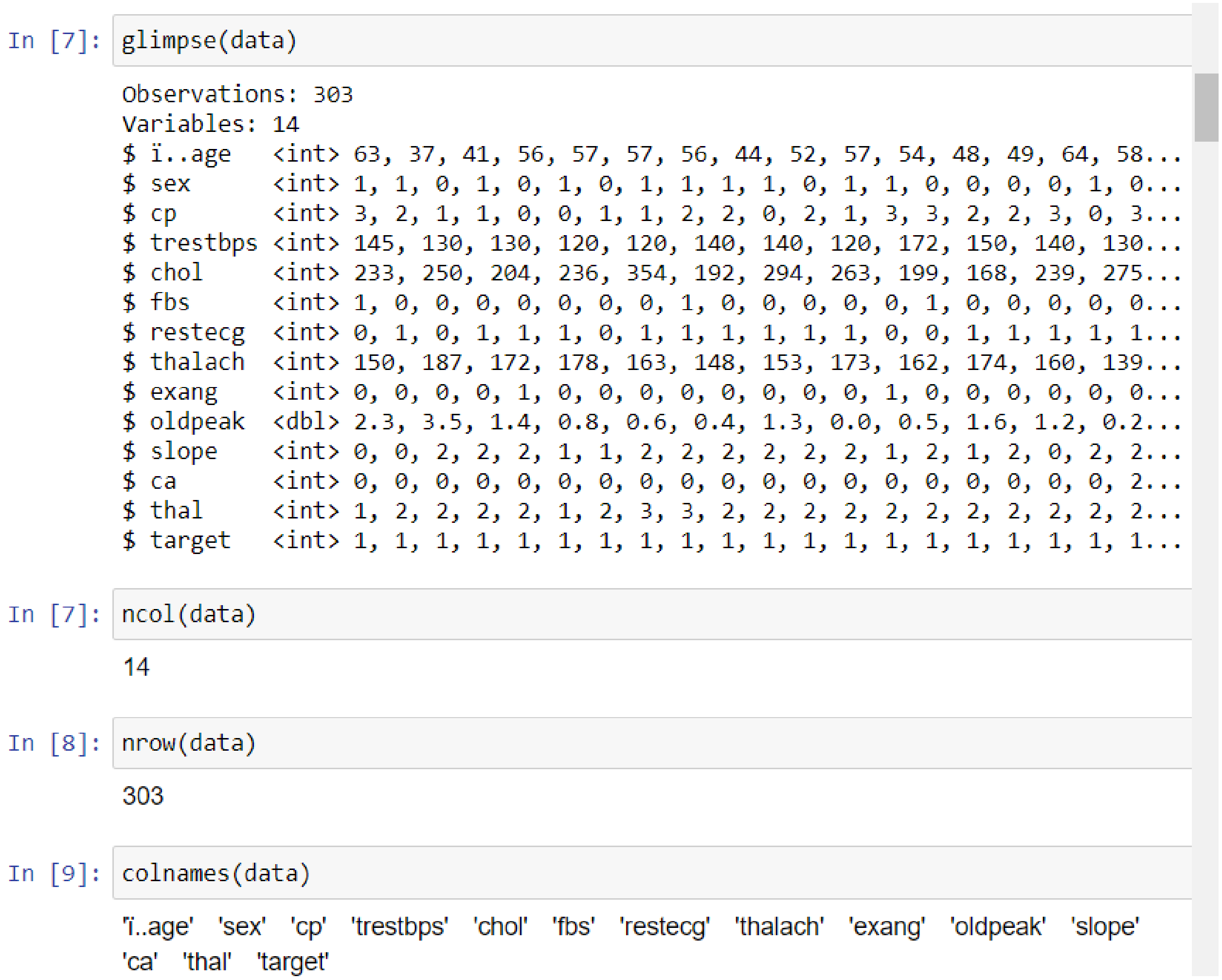
This is head of the dataset and values and attributes will be in this format.



```
In [6]: library(tidyverse)
```

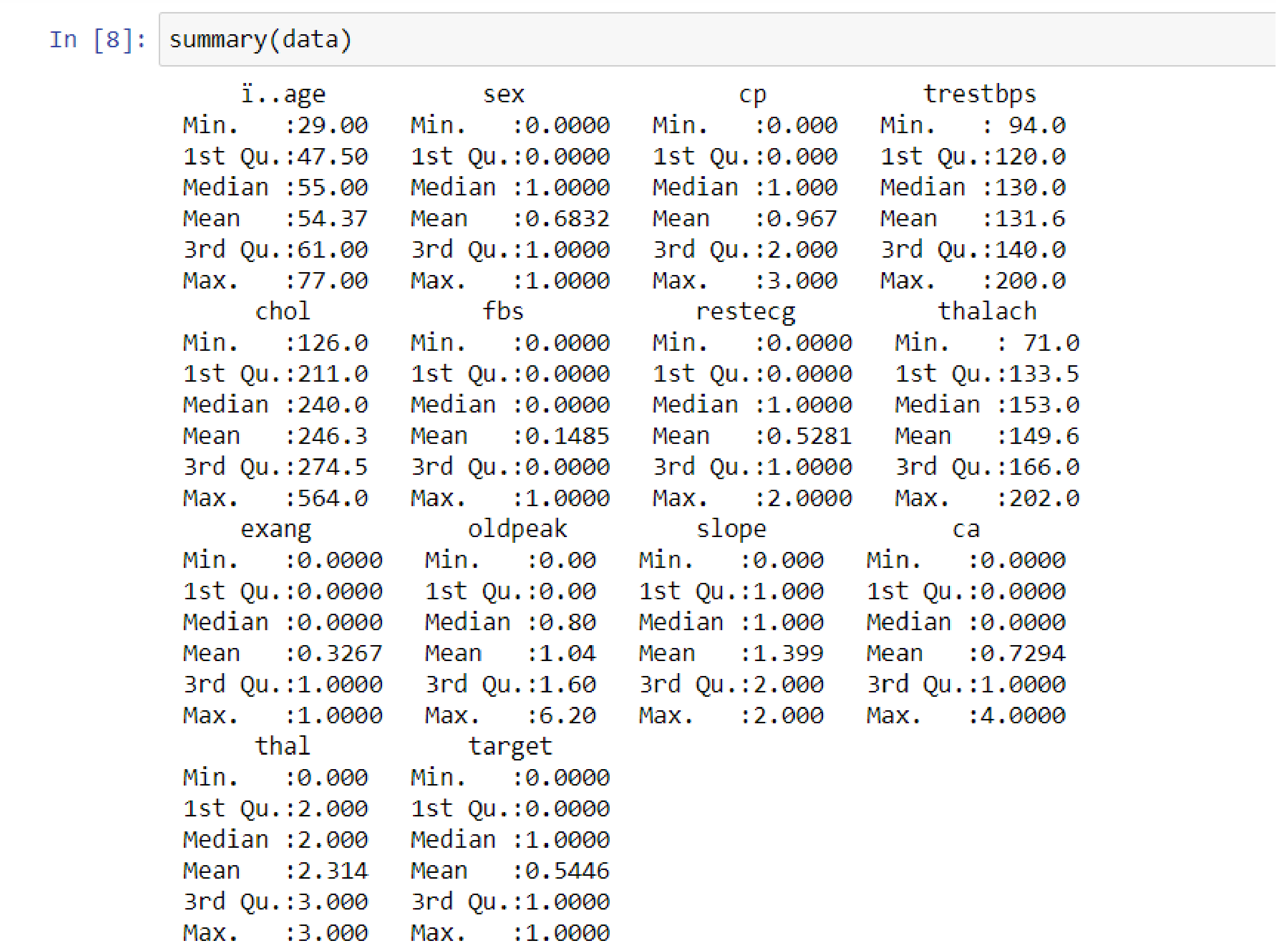
```
Registered S3 methods overwritten by 'ggplot2':
  method         from
  [.quosures     rlang
  c.quosures     rlang
  print.quosures rlang
Registered S3 method overwritten by 'rvest':
  method            from
  read_xml.response xml2
-- Attaching packages --------------------------------------- tidyverse 1.2.1 -
-
v ggplot2 3.1.1       v purrr   0.3.2
v tibble  2.1.1       v dplyr   0.8.0.1
v tidyr   0.8.3       v stringr 1.4.0
v readr   1.3.1       v forcats 0.4.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() -
-
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

Installing tidyverse library for enhanced analysis .



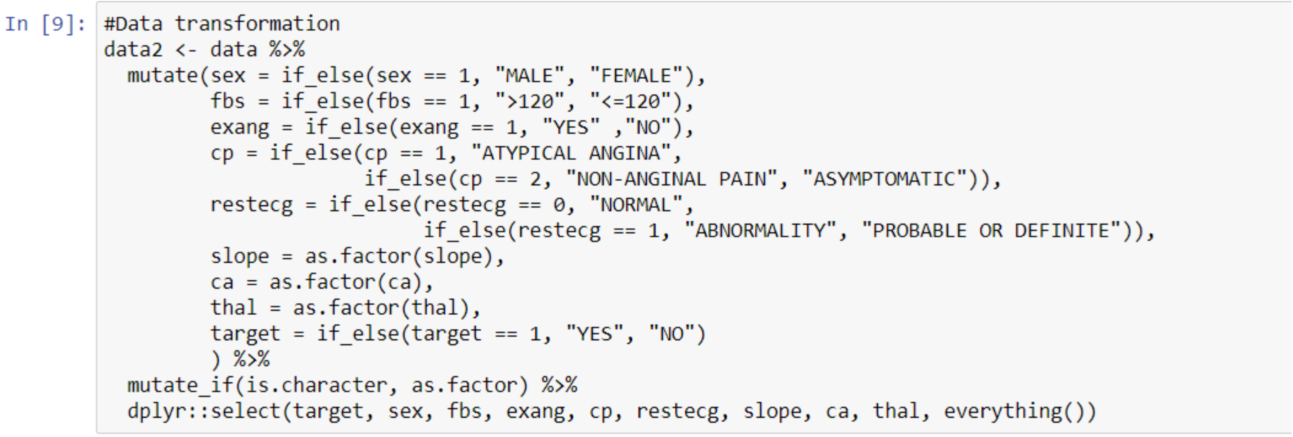
```
In [7]: glimpse(data)

Observations: 303
Variables: 14
$ ï..age   <int> 63, 37, 41, 56, 57, 57, 56, 44, 52, 57, 54, 48, 49, 64, 58...
$ sex      <int> 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0...
$ cp       <int> 3, 2, 1, 1, 0, 0, 1, 1, 2, 2, 0, 2, 1, 3, 3, 2, 2, 3, 0, 3...
$ trestbps <int> 145, 130, 130, 120, 120, 140, 140, 120, 172, 150, 140, 130...
$ chol     <int> 233, 250, 204, 236, 354, 192, 294, 263, 199, 168, 239, 275...
$ fbs      <int> 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0...
$ restecg  <int> 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1...
$ thalach  <int> 150, 187, 172, 178, 163, 148, 153, 173, 162, 174, 160, 139...
$ exang    <int> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0...
$ oldpeak  <dbl> 2.3, 3.5, 1.4, 0.8, 0.6, 0.4, 1.3, 0.0, 0.5, 1.6, 1.2, 0.2...
$ slope    <int> 0, 0, 2, 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 1, 2, 1, 2, 0, 2, 2...
$ ca       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2...
$ thal     <int> 1, 2, 2, 2, 2, 1, 2, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2...
$ target   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...

In [7]: ncol(data)

14

In [8]: nrow(data)

303

In [9]: colnames(data)

'ï..age'  'sex'  'cp'  'trestbps'  'chol'  'fbs'  'restecg'  'thalach'  'exang'  'oldpeak'  'slope'
'ca'  'thal'  'target'
```

Basic glimpse of the dataset and some info on number of rows and columns.



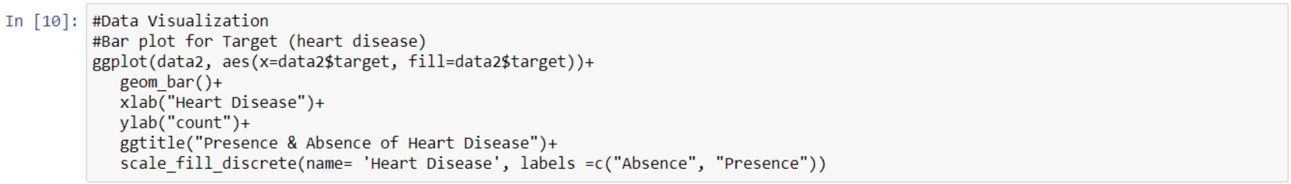
```
In [8]: summary(data)

     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  
```

Summary of the dataset and its attributes with mean and median and much more variables describing the dataset.

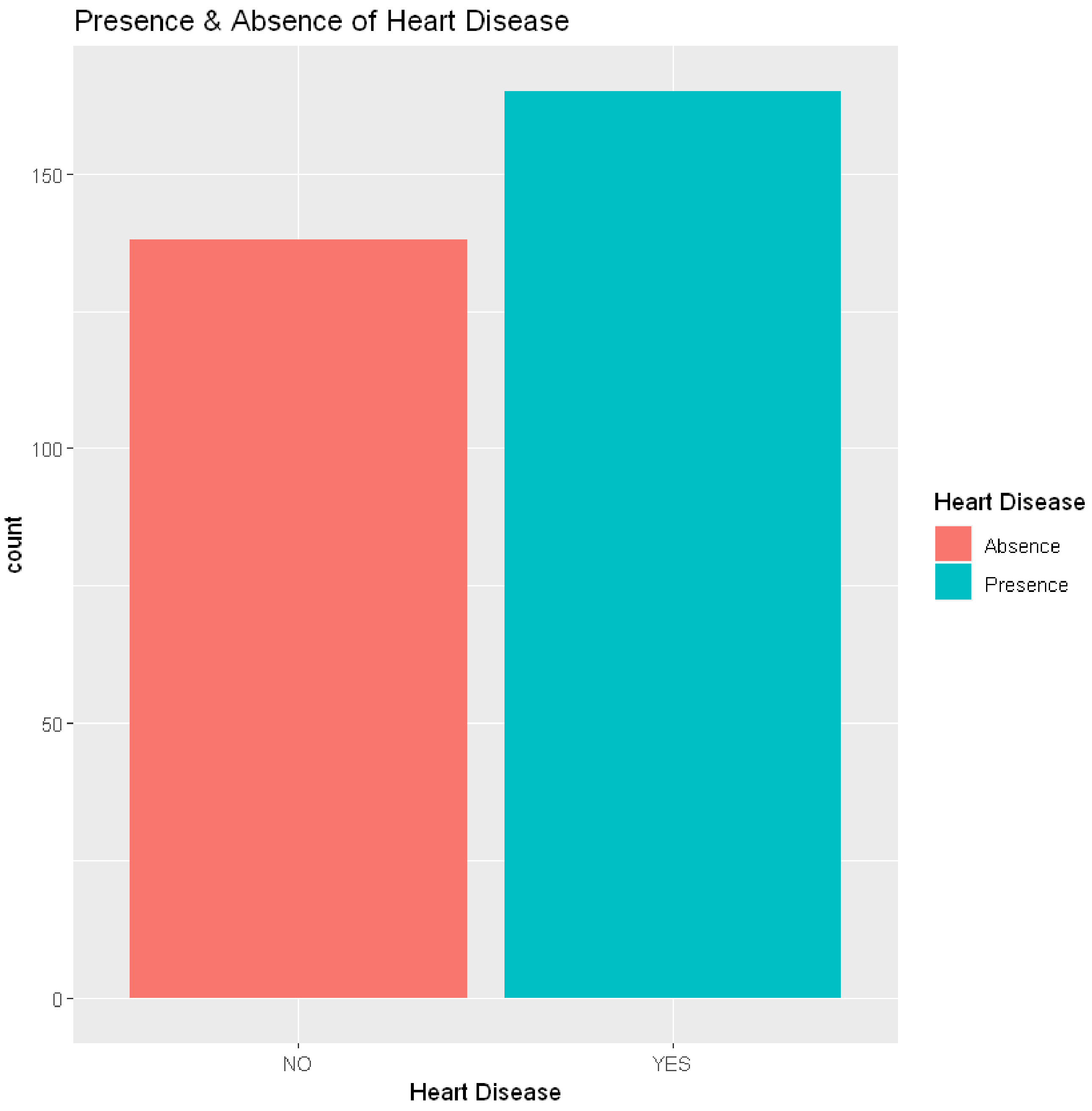


```
In [9]: #Data transformation
        data2 <- data %>%
          mutate(sex = if_else(sex == 1, "MALE", "FEMALE"),
                 fbs = if_else(fbs == 1, ">120", "<=120"),
                 exang = if_else(exang == 1, "YES" ,"NO"),
                 cp = if_else(cp == 1, "ATYPICAL ANGINA",
                              if_else(cp == 2, "NON-ANGINAL PAIN", "ASYMPTOMATIC")),
                 restecg = if_else(restecg == 0, "NORMAL",
                                   if_else(restecg == 1, "ABNORMALITY", "PROBABLE OR DEFINITE")),
                 slope = as.factor(slope),
                 ca = as.factor(ca),
                 thal = as.factor(thal),
                 target = if_else(target == 1, "YES", "NO")
                 ) %>%
          mutate_if(is.character, as.factor) %>%
          dplyr::select(target, sex, fbs, exang, cp, restecg, slope, ca, thal, everything())
```

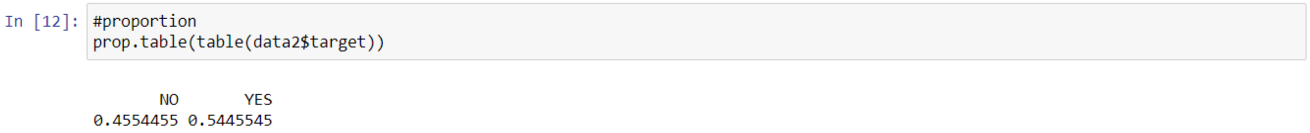
Data transformation for making user friendly analysis on further graphs .



```
In [10]: #Data Visualization
         #Bar plot for Target (heart disease)
         ggplot(data2, aes(x=data2$target, fill=data2$target))+
            geom_bar()+
            xlab("Heart Disease")+
            ylab("count")+
            ggtitle("Presence & Absence of Heart Disease")+
            scale_fill_discrete(name= 'Heart Disease', labels =c("Absence", "Presence"))
```

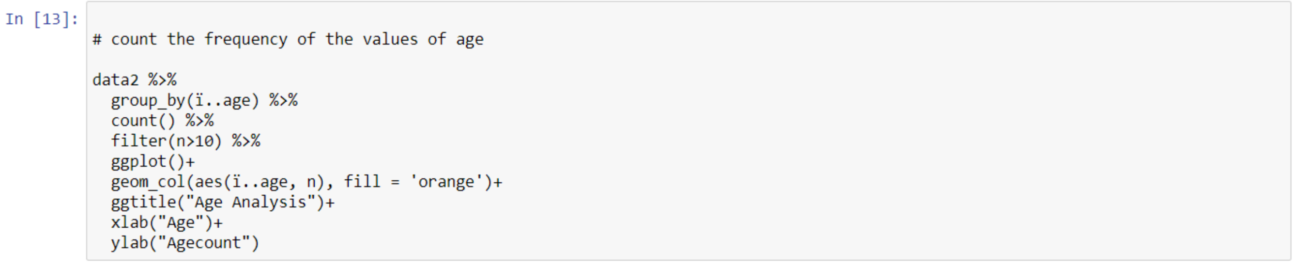


Presence of heart disease in this data set to see whether it is biased or natural and proper dataset.

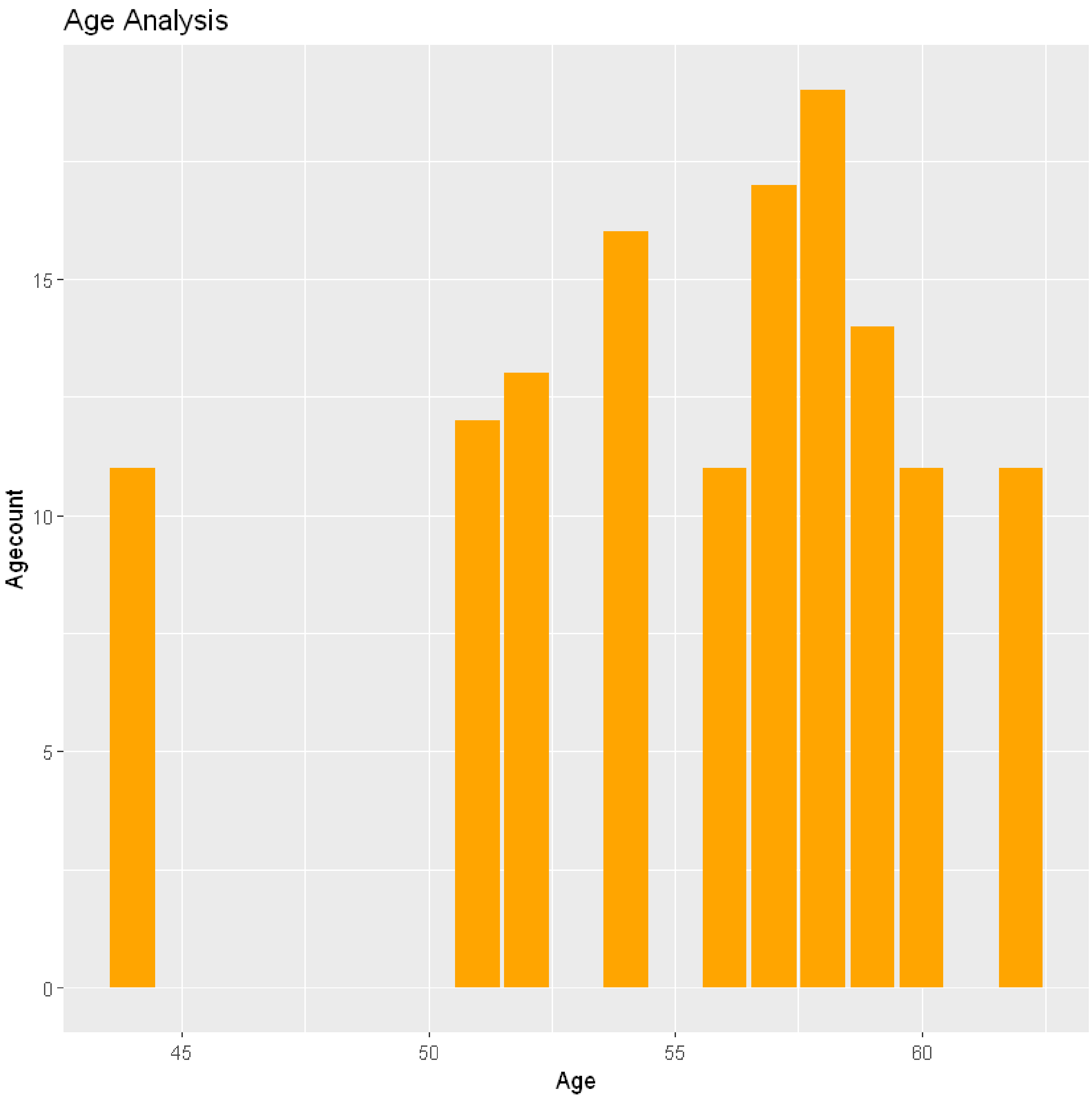


```
In [12]: #proportion
         prop.table(table(data2$target))

                NO       YES
         0.4554455 0.5445545
```

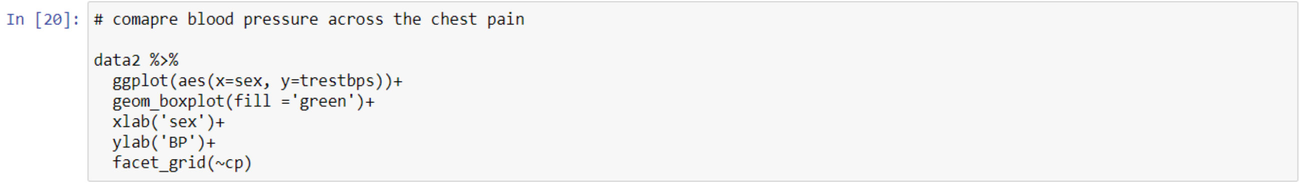
This proportion confirms our above statement.



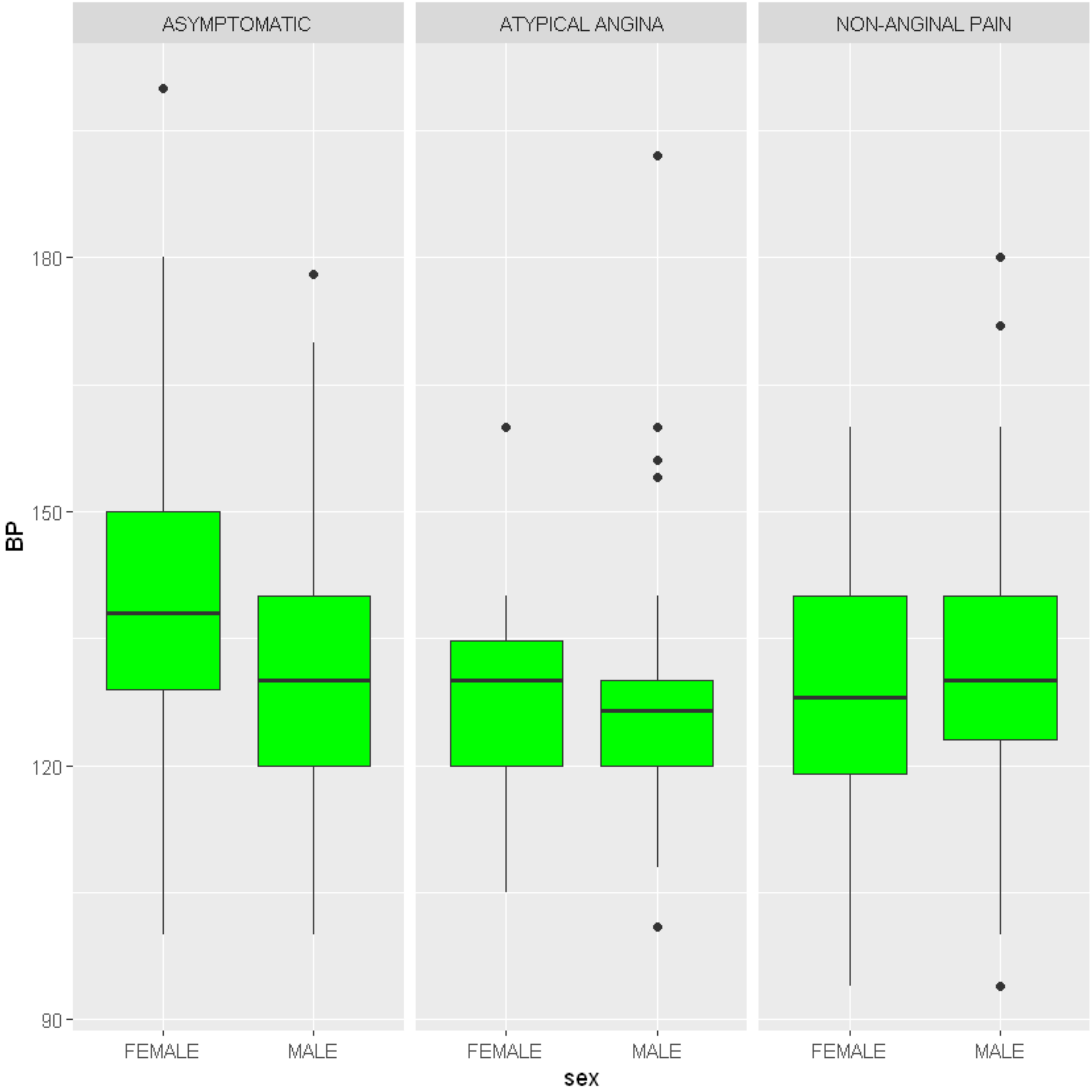
```
In [13]:
         # count the frequency of the values of age

         data2 %>%
           group_by(ï..age) %>%
           count() %>%
           filter(n>10) %>%
           ggplot()+
           geom_col(aes(ï..age, n), fill = 'orange')+
           ggtitle("Age Analysis")+
           xlab("Age")+
           ylab("Agecount")
```



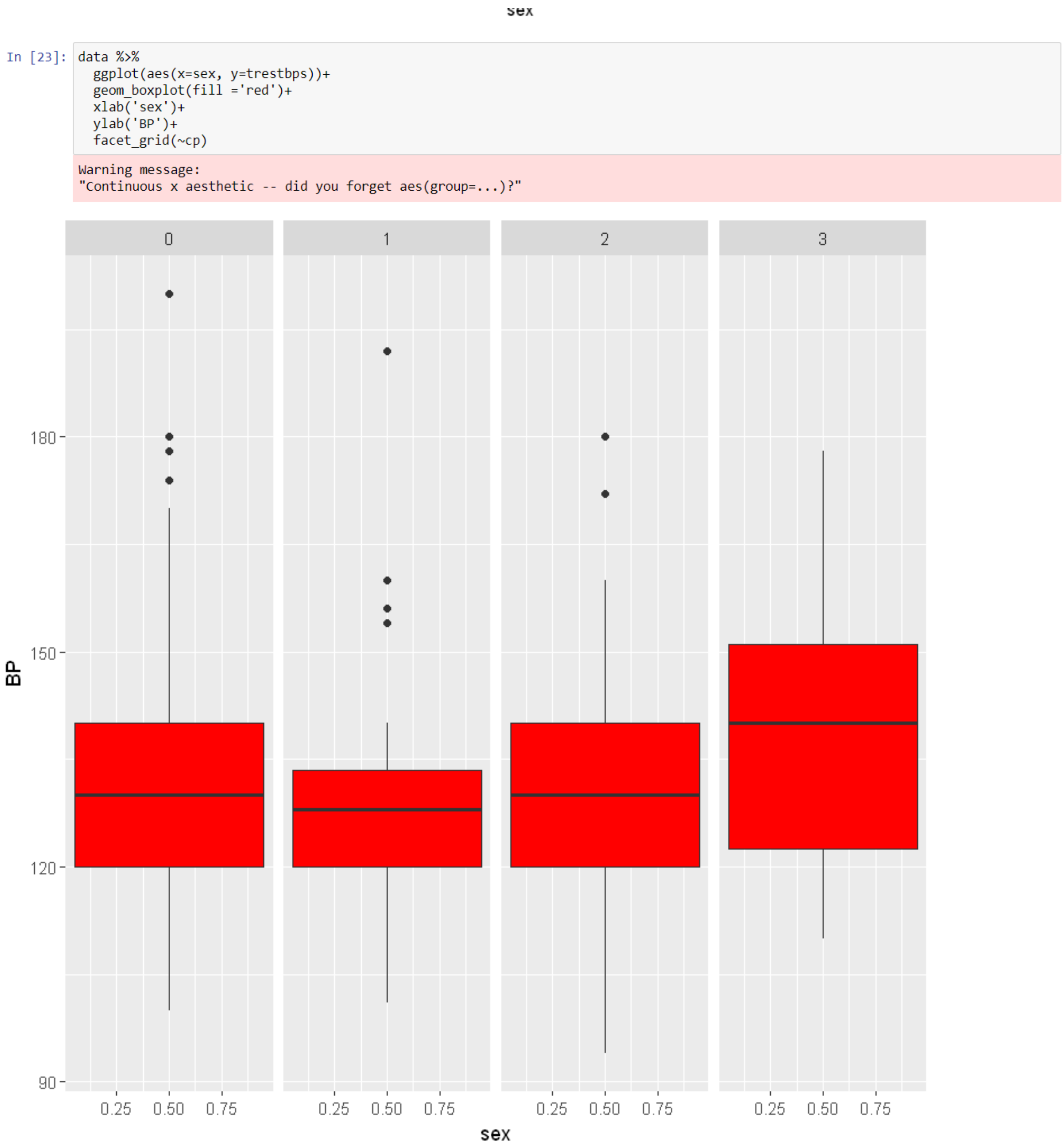
Above analysis shows how age factor affects the heart disease and we can see 45 to 50 has big gap not having any heart disease and below 30 lots of heart disease.



```
In [20]: # comapre blood pressure across the chest pain

         data2 %>%
           ggplot(aes(x=sex, y=trestbps))+
           geom_boxplot(fill ='green')+
           xlab('sex')+
           ylab('BP')+
           facet_grid(~cp)
```



Male having many outliers shows higher blood pressure for them and females are having not so higher than males so males are prone to heart disease .

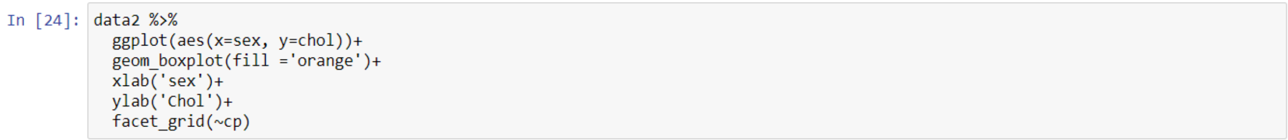


In [23]:

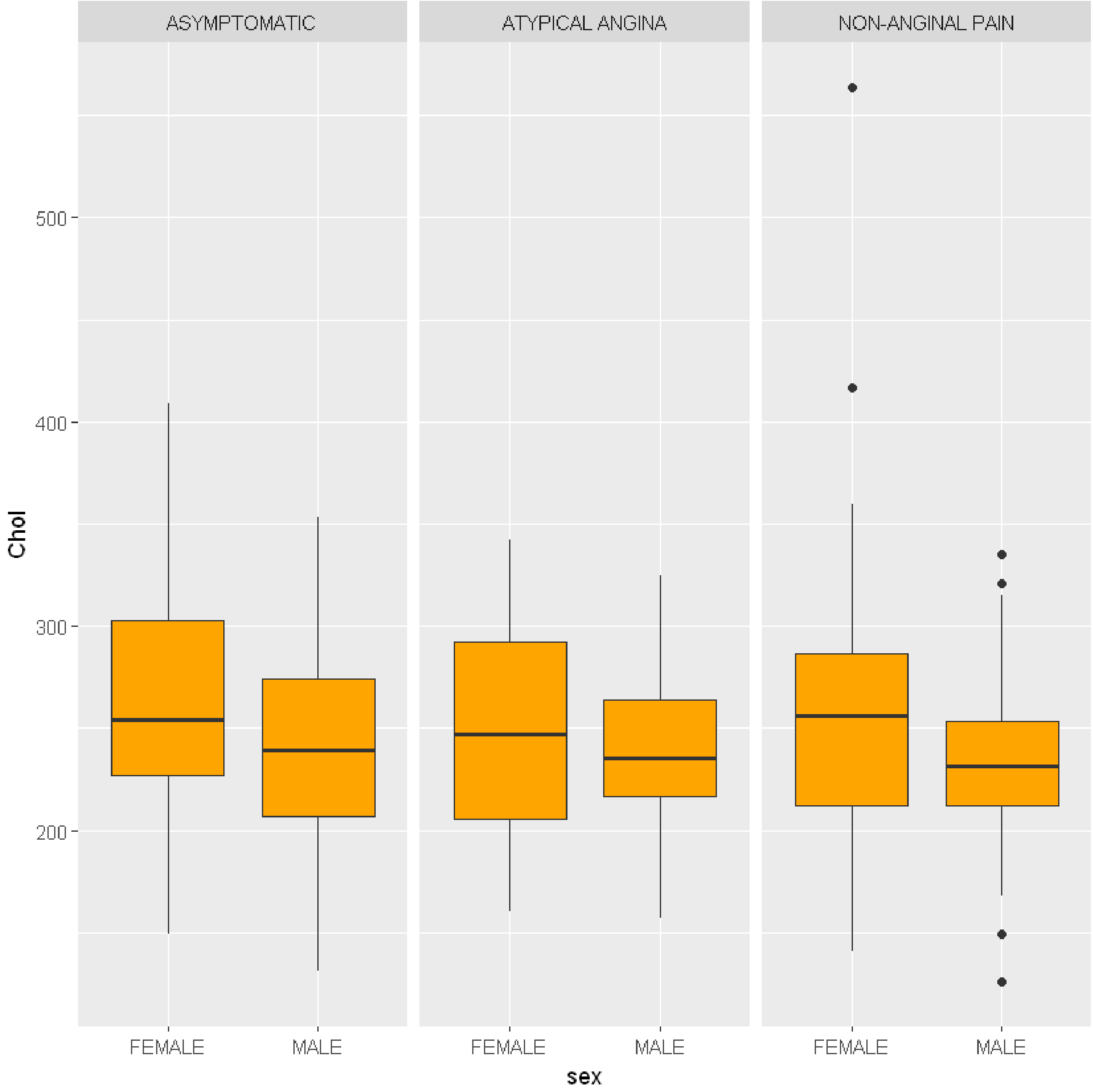
```
data %>%
  ggplot(aes(x=sex, y=trestbps))+
  geom_boxplot(fill ='red')+
  xlab('sex')+
  ylab('BP')+
  facet_grid(~cp)
```

```
Warning message:
"Continuous x aesthetic -- did you forget aes(group=...)?"
```

Here we are used data which is not transformed data , so from graph we can clearly see that without transforming how difficult is to do analysis and we cant able to derive any conclusions.

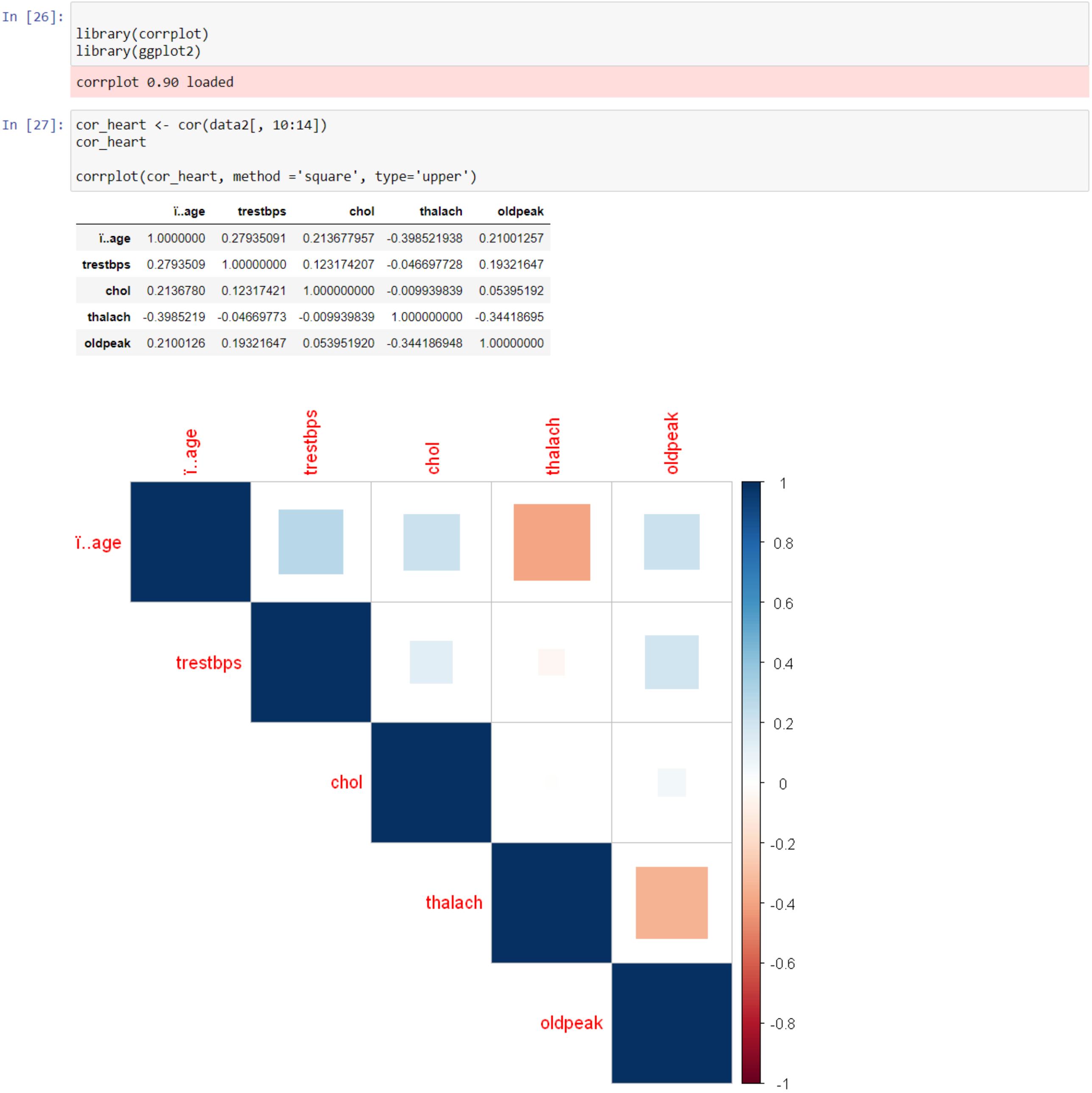


```
In [24]: data2 %>%
           ggplot(aes(x=sex, y=chol))+
           geom_boxplot(fill ='orange')+
           xlab('sex')+
           ylab('Chol')+
           facet_grid(~cp)
```



We can clearly see that female are showing higher spike than males which draws conclusion that they are prone to heart disease because of higher cholesterol. So they have to aware of that.

CORRELATION :



In [26]:

```
library(corrplot)
library(ggplot2)
```

```
corrplot 0.90 loaded
```

In [27]:

```
cor_heart <- cor(data2[, 10:14])
cor_heart

corrplot(cor_heart, method ='square', type='upper')
```

	ï..age	trestbps	chol	thalach	oldpeak
ï..age	1.0000000	0.27935091	0.213677957	-0.398521938	0.21001257
trestbps	0.2793509	1.00000000	0.123174207	-0.046697728	0.19321647
chol	0.2136780	0.12317421	1.000000000	-0.009939839	0.05395192
thalach	-0.3985219	-0.04669773	-0.009939839	1.000000000	-0.34418695
oldpeak	0.2100126	0.19321647	0.053951920	-0.344186948	1.00000000

As our dataset is small here the correlation not showing any significant relation between any of the attributes , may be in future dataset with more attributes may show relation between the attributes.

CONCLUSION :

From our dataset we have done some analysis on factors which leads to heart disease and how it affects male and female on different factors .
For example males are more prone because of high blood pressure and females are to higher cholesterol levels .
And at end we didn′ t see any significant relation between the attributes.