I have selected the following three data sets

1. **500 Person Gender-Height-Weight-Body Mass Index**

<https://www.kaggle.com/yersever/500-person-gender-height-weight-bodymassindex>

For long time, doctors, fitness centers and even the military collect Males and Females Height Weight and Body mass index. Rather than using the body measurement to determine if a person is within the normal weight scale, let's use that data to predict gender

> mydata<-read.csv("bmi.csv")

> names(mydata)

[1] "Gender" "Height" "Weight" "Index"

> str(mydata)

'data.frame': 500 obs. of 4 variables:

$ Gender: Factor w/ 2 levels "Female","Male": 2 2 1 1 2 2 2 2 2 1 ...

$ Height: int 174 189 185 195 149 189 147 154 174 169 ...

$ Weight: int 96 87 110 104 61 104 92 111 90 103 ...

$ Index : int 4 2 4 3 3 3 5 5 3 4 ...

> summary(mydata)

Gender Height Weight Index

Female:255 Min. :140.0 Min. : 50 Min. :0.000

Male :245 1st Qu.:156.0 1st Qu.: 80 1st Qu.:3.000

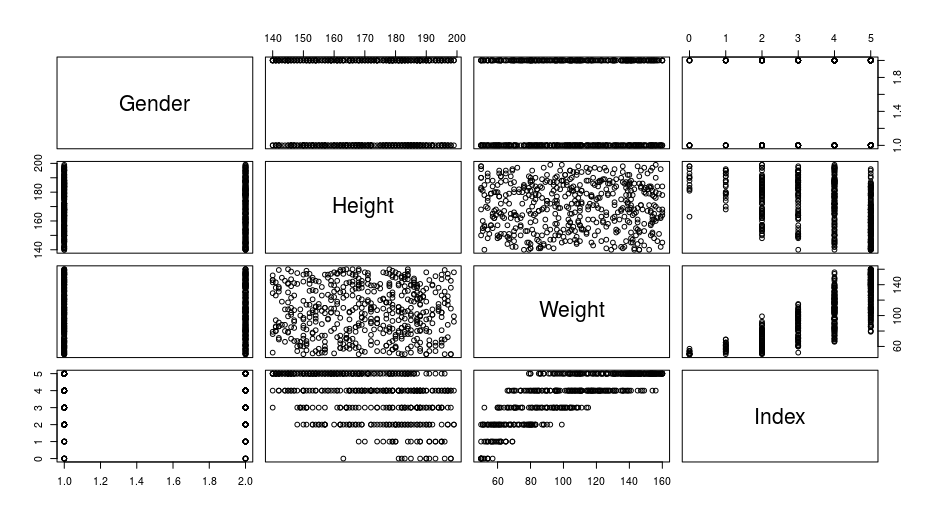
Median :170.5 Median :106 Median :4.000

Mean :169.9 Mean :106 Mean :3.748

3rd Qu.:184.0 3rd Qu.:136 3rd Qu.:5.000

Max. :199.0 Max. :160 Max. :5.000

> plot(mydata)



1. **Salary - Salary against years of exp**

<https://www.kaggle.com/rsadiq/salary>

> mydata<-read.csv("Salary.csv")

> names(mydata)

[1] "YearsExperience" "Salary"

> str(mydata)

'data.frame': 35 obs. of 2 variables:

$ YearsExperience: num 1.1 1.3 1.5 2 2.2 2.9 3 3.2 3.2 3.7 ...

$ Salary : int 39343 46205 37731 43525 39891 56642 60150 54445 64445 57189 ...

> summary(mydata)

YearsExperience Salary

Min. : 1.100 Min. : 37731

1st Qu.: 3.450 1st Qu.: 57019

Median : 5.300 Median : 81363

Mean : 6.309 Mean : 83946

3rd Qu.: 9.250 3rd Qu.:113224

Max. :13.500 Max. :139465

> plot(mydata)

> model<-lm(Salary~YearsExperience, data=mydata)

> model

Call:

lm(formula = Salary ~ YearsExperience, data = mydata)

Coefficients:

(Intercept) YearsExperience

28860 8732

> summary(model)

Call:

lm(formula = Salary ~ YearsExperience, data = mydata)

Residuals:

Min 1Q Median 3Q Max

-8178.8 -5026.5 303.9 4098.0 13350.6

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 28859.5 2092.8 13.79 2.99e-15 \*\*\*

YearsExperience 8731.9 288.8 30.24 < 2e-16 \*\*\*

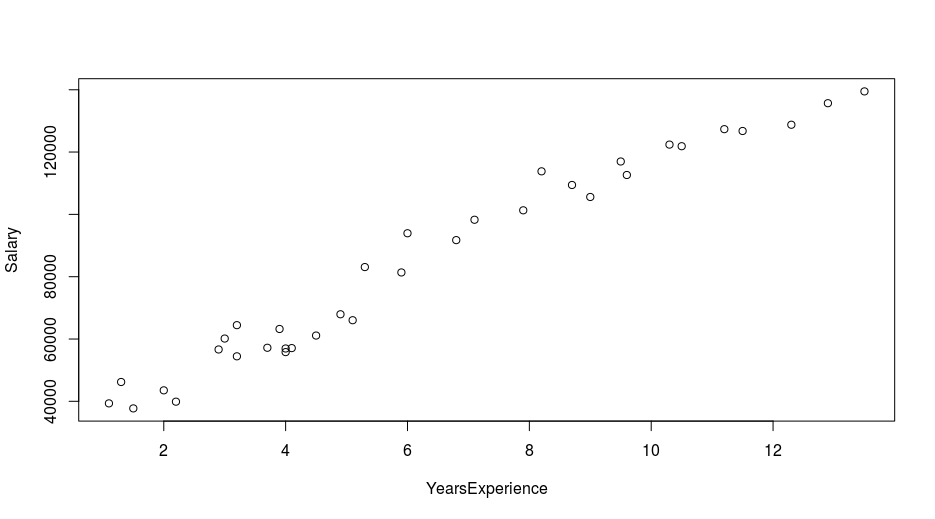
---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6093 on 33 degrees of freedom

Multiple R-squared: 0.9652, Adjusted R-squared: 0.9641

F-statistic: 914.3 on 1 and 33 DF, p-value: < 2.2e-16



1. **The Heart Disease Dataset**

<https://www.kaggle.com/johnsmith88/heart-disease-dataset#heart.csv>

Dataset:

The Heart Disease Dataset is downloaded from the link given below. The data set is downloaded in csv file and analyzed using the R programming.

<https://www.kaggle.com/johnsmith88/heart-disease-dataset#heart.csv>

The data set have the following variables and their description

* age
* sex
* chest pain type (4 values) - cp
* resting blood pressure - trestbps
* serum cholesterol in mg/dl - chol
* fasting blood sugar > 120 mg/dl -fbs
* resting electrocardiographic results (values 0,1,2) -restecg
* maximum heart rate achieved -thalach
* exercise induced angina - exang
* old peak = ST depression induced by exercise relative to rest -oldpeak
* the slope of the peak exercise ST segment - slope
* number of major vessels (0-3) coloured by fluoroscopy -ca
* thal: 0 = normal; 1 = fixed defect; 2 = reversable defect - thal

Introduction:

Heart diseases are the significant reason for human death rate. Right analysis and treatment at a beginning time will spare individuals from coronary illness and will diminish mortality rate because of heart issue. Since ten years different information mining procedures have been utilized to encourage the forecast of heart diseases .when all is said in done forecast calculations for prepared with tremendous, known dataset to show up at a classifier which at that point predicts the diseases for obscure information with the assistance of ordering traits. These characteristics additionally called as highlights. In this work significant highlights are resolved for coronary illness expectation with known dataset utilizing relationship measures. The outcomes are introduced

Research Question :

The main idea of this project to model for predicting the person who is already diagnosed heart disease by using the pattern from the 14 variable descriptive data. The task comprises of two stages. Stage I centers around information preprocessing and investigation, as shrouded in this report. The model structure, approval and expectation are introduced in Phase II. The remainder of this report is composed as follow. Segment 2 portrays the informational collections and their properties. segment 3 spreads information pre-handling. In segment 4, we investigate each characteristic and their between connections. In the last we summaries the result.

There may be many attributes related to a given prediction problem. But not all the attributes have strong association with the prediction. Hence finding the relevant attributes for a given prediction problem is important. In this work, relevant attributes for heart disease prediction are determined using correlation measure. As mentioned above, from link it is found that the 13 attributes (thal, ca, exang, oldpeak, thalach, cp, slope, sex, age, restecg, trestbps, chol, fbs) are beingused while predicting heart diseases. In order to find the weight or rank of these attributes an experiment has been conducted. In this experiment the correlation between each attribute and class label is found out.

Packages :

library(knitr) library(readr)

library(dplyr) library(ggplot2)

library(mlr) library(cowplot)

library(tidyverse) library(corrplot)

library(qgraph) library(jtools)

library(caret) library(DataExplorer)

library(funModeling) library(ggm)

Data Analysis

> tbl\_df(heart)

# A tibble: 1,025 x 14

age sex cp trestbps chol

*<int>* *<int>* *<int>* *<int>* *<int>*

1 52 1 0 125 212

2 53 1 0 140 203

3 70 1 0 145 174

4 61 1 0 148 203

5 62 0 0 138 294

6 58 0 0 100 248

7 58 1 0 114 318

8 55 1 0 160 289

9 46 1 0 120 249

10 54 1 0 122 286

# … with 1,015 more rows, and 9 more

# variables: fbs *<int>*,

# restecg *<int>*, thalach *<int>*,

# exang *<int>*, oldpeak *<dbl>*,

# slope *<int>*, ca *<int>*, thal *<int>*,

# target *<int>*

> str(heart)

'data.frame': 1025 obs. of 14 variables:

$ age : int 52 53 70 61 62 58 58 55 46 54 ...

$ sex : int 1 1 1 1 0 0 1 1 1 1 ...

$ cp : int 0 0 0 0 0 0 0 0 0 0 ...

$ trestbps: int 125 140 145 148 138 100 114 160 120 122 ...

$ chol : int 212 203 174 203 294 248 318 289 249 286 ...

$ fbs : int 0 1 0 0 1 0 0 0 0 0 ...

$ restecg : int 1 0 1 1 1 0 2 0 0 0 ...

$ thalach : int 168 155 125 161 106 122 140 145 144 116 ...

$ exang : int 0 1 1 0 0 0 0 1 0 1 ...

$ oldpeak : num 1 3.1 2.6 0 1.9 1 4.4 0.8 0.8 3.2 ...

$ slope : int 2 0 0 2 1 1 0 1 2 1 ...

$ ca : int 2 0 0 1 3 0 3 1 0 2 ...

$ thal : int 3 3 3 3 2 2 1 3 3 2 ...

$ target : int 0 0 0 0 0 1 0 0 0 0 ...

There are 14 variables and 1025 observations and 13 variables are integers and 1 variable is numeric

corr <- cor(heart)

corrplot(corr,method = "number",type="lower",order = "hclust",title = "Correlations between Variables ")

> corr

age sex cp trestbps chol

age 1.00000000 -0.10324030 -0.07196627 0.27112141 0.21982253

sex -0.10324030 1.00000000 -0.04111909 -0.07897377 -0.19825787

cp -0.07196627 -0.04111909 1.00000000 0.03817742 -0.08164102

trestbps 0.27112141 -0.07897377 0.03817742 1.00000000 0.12797743

chol 0.21982253 -0.19825787 -0.08164102 0.12797743 1.00000000

fbs 0.12124348 0.02720046 0.07929359 0.18176662 0.02691716

restecg -0.13269617 -0.05511721 0.04358061 -0.12379409 -0.14741024

thalach -0.39022708 -0.04936524 0.30683928 -0.03926407 -0.02177209

exang 0.08816338 0.13915681 -0.40151271 0.06119697 0.06738223

oldpeak 0.20813668 0.08468656 -0.17473348 0.18743411 0.06488031

slope -0.16910511 -0.02666629 0.13163278 -0.12044531 -0.01424787

ca 0.27155053 0.11172891 -0.17620647 0.10455372 0.07425934

thal 0.07229745 0.19842425 -0.16334148 0.05927618 0.10024418

target -0.22932355 -0.27950076 0.43485425 -0.13877173 -0.09996559

fbs restecg thalach exang oldpeak

age 0.121243479 -0.13269617 -0.390227075 0.08816338 0.20813668

sex 0.027200461 -0.05511721 -0.049365243 0.13915681 0.08468656

cp 0.079293586 0.04358061 0.306839282 -0.40151271 -0.17473348

trestbps 0.181766624 -0.12379409 -0.039264069 0.06119697 0.18743411

chol 0.026917164 -0.14741024 -0.021772091 0.06738223 0.06488031

fbs 1.000000000 -0.10405124 -0.008865857 0.04926057 0.01085948

restecg -0.104051244 1.00000000 0.048410637 -0.06560553 -0.05011425

thalach -0.008865857 0.04841064 1.000000000 -0.38028087 -0.34979616

exang 0.049260570 -0.06560553 -0.380280872 1.00000000 0.31084376

oldpeak 0.010859481 -0.05011425 -0.349796163 0.31084376 1.00000000

slope -0.061902374 0.08608609 0.395307843 -0.26733547 -0.57518854

ca 0.137156259 -0.07807235 -0.207888416 0.10784854 0.22181603

thal -0.042177320 -0.02050406 -0.098068165 0.19720104 0.20267203

target -0.041163547 0.13446821 0.422895496 -0.43802855 -0.43844127

slope ca thal target

age -0.16910511 0.27155053 0.07229745 -0.22932355

sex -0.02666629 0.11172891 0.19842425 -0.27950076

cp 0.13163278 -0.17620647 -0.16334148 0.43485425

trestbps -0.12044531 0.10455372 0.05927618 -0.13877173

chol -0.01424787 0.07425934 0.10024418 -0.09996559

fbs -0.06190237 0.13715626 -0.04217732 -0.04116355

restecg 0.08608609 -0.07807235 -0.02050406 0.13446821

thalach 0.39530784 -0.20788842 -0.09806817 0.42289550

exang -0.26733547 0.10784854 0.19720104 -0.43802855

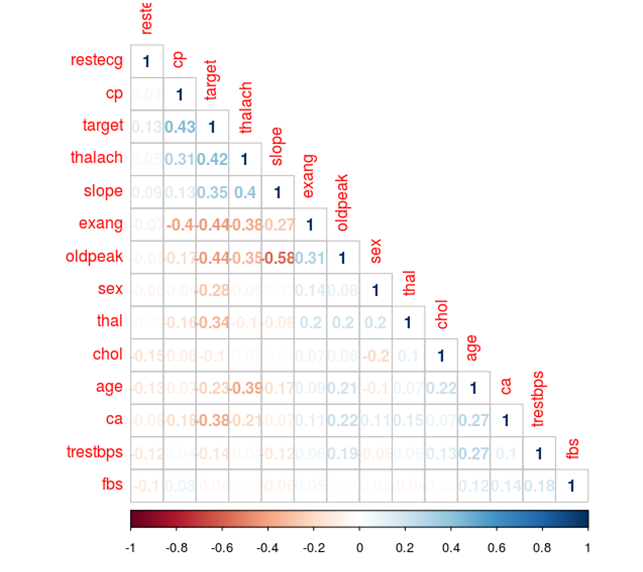
oldpeak -0.57518854 0.22181603 0.20267203 -0.43844127

slope 1.00000000 -0.07344041 -0.09409006 0.34551175

ca -0.07344041 1.00000000 0.14901387 -0.38208529

thal -0.09409006 0.14901387 1.00000000 -0.33783815

target 0.34551175 -0.38208529 -0.33783815 1.00000000



Oldpeak has the highest correlation coefficient of 0.58 which indicating that these variables have good association with each other. So we can say that depression induced by exercise relative to rest is associated with the slope of the peak exercise ST segment - slope

**heart$sex <- ifelse(heart$sex==1,"male","female")**

**heart$cp <- ifelse(heart$cp==1,"typical angina",ifelse(heart$cp==2,**

**"atypical angina",ifelse(heart==3, "non-anginal pain","asymptomatic")))**

**heart$restecg <- ifelse(heart$restecg==0, "normal", ifelse(heart$trestbps==1,**

**"ST-T abnormality", "hypertrophy"))**

**heart$exang <- ifelse(heart$exang==1,"yes","no")**

**heart$slope <- ifelse(heart$slope==1,"upsloping", ifelse(heart$slope== 2,**

**"flat","downsloping"))**

**heart$thal <- ifelse(heart$thal==3,"normal", ifelse(heart$thal==6,"fixed defect",**

**"reversable defect"))**

**heart$fbs <- ifelse(heart$fbs==1,"true","false") # fasting or not**

**heart$target <- ifelse(heart$target==0,"typical","worse")**

Numeric categories only

**str(heart)**

> str(heart)

'data.frame': 1025 obs. of 14 variables:

$ age : int 52 53 70 61 62 58 58 55 46 54 ...

$ sex : chr "male" "male" "male" "male" ...

$ cp : chr "asymptomatic" "asymptomatic" "asymptomatic" "asymptomatic" ...

$ trestbps: int 125 140 145 148 138 100 114 160 120 122 ...

$ chol : int 212 203 174 203 294 248 318 289 249 286 ...

$ fbs : chr "false" "true" "false" "false" ...

$ restecg : chr "hypertrophy" "normal" "hypertrophy" "hypertrophy" ...

$ thalach : int 168 155 125 161 106 122 140 145 144 116 ...

$ exang : chr "no" "yes" "yes" "no" ...

$ oldpeak : num 1 3.1 2.6 0 1.9 1 4.4 0.8 0.8 3.2 ...

$ slope : chr "flat" "downsloping" "downsloping" "flat" ...

$ ca : int 2 0 0 1 3 0 3 1 0 2 ...

$ thal : chr "normal" "normal" "normal" "normal" ...

$ target : chr "typical" "typical" "typical" "typical" ...

**> introduce(heart)**

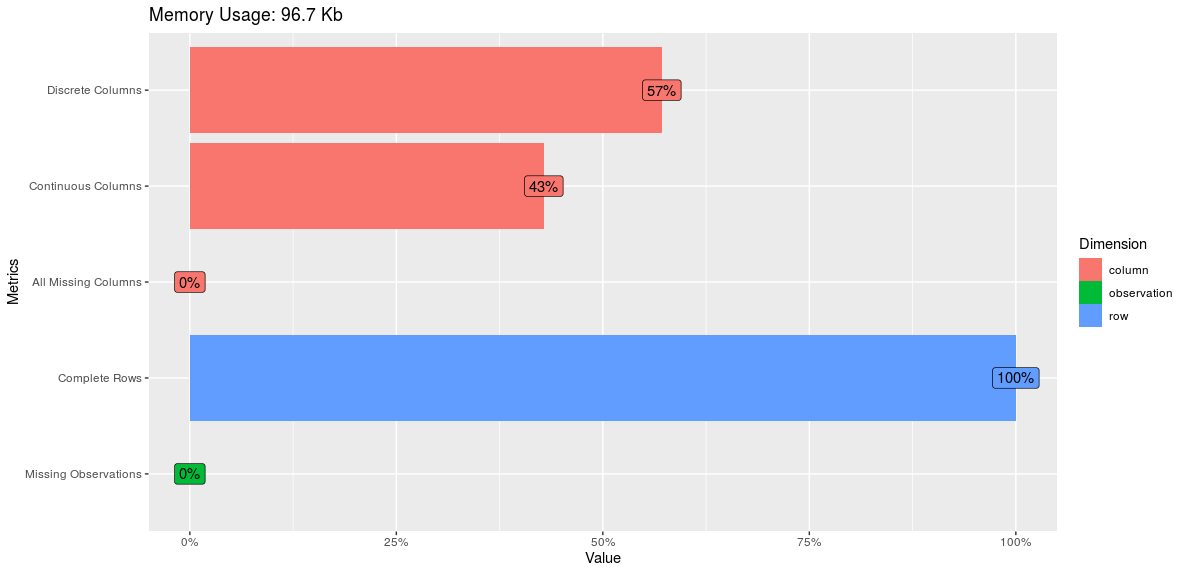
rows columns discrete\_columns continuous\_columns all\_missing\_columns

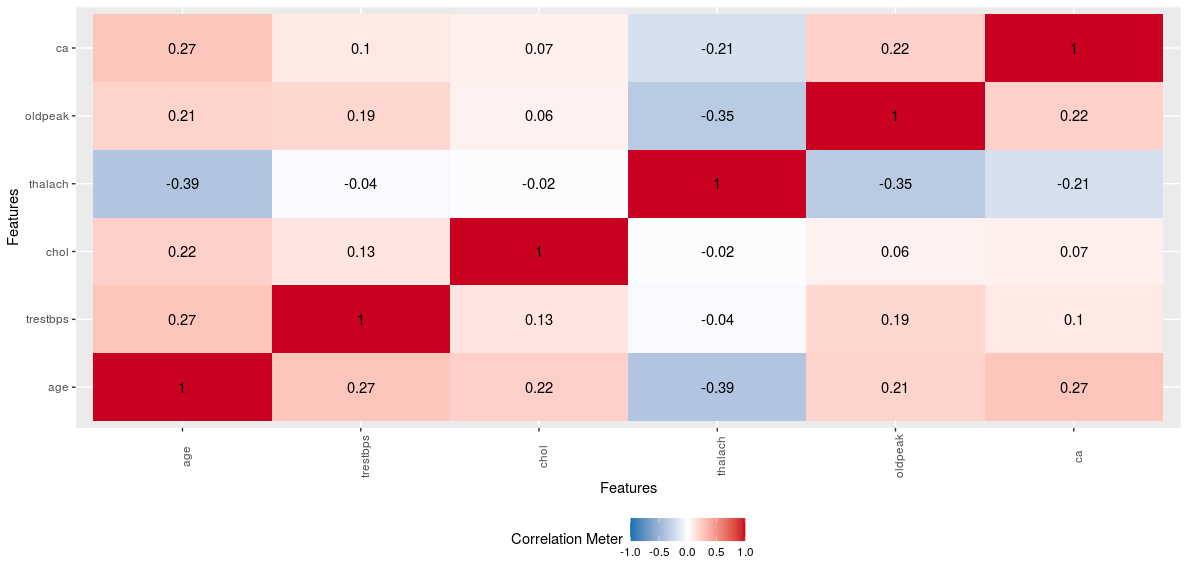
1 1025 14 8 6 0

total\_missing\_values complete\_rows total\_observations memory\_usage

1 0 1025 14350 99008

**plot\_intro(heart)**





**par(mfrow=c(2,2))**

**hist(heart$trestbps, col = "skyblue", xlab = "Blood Pressure", main =**

**"Histogram of trestbps: resting blood pressure")**

**hist(heart$age, col = "lightblue", xlab = "Age", main =**

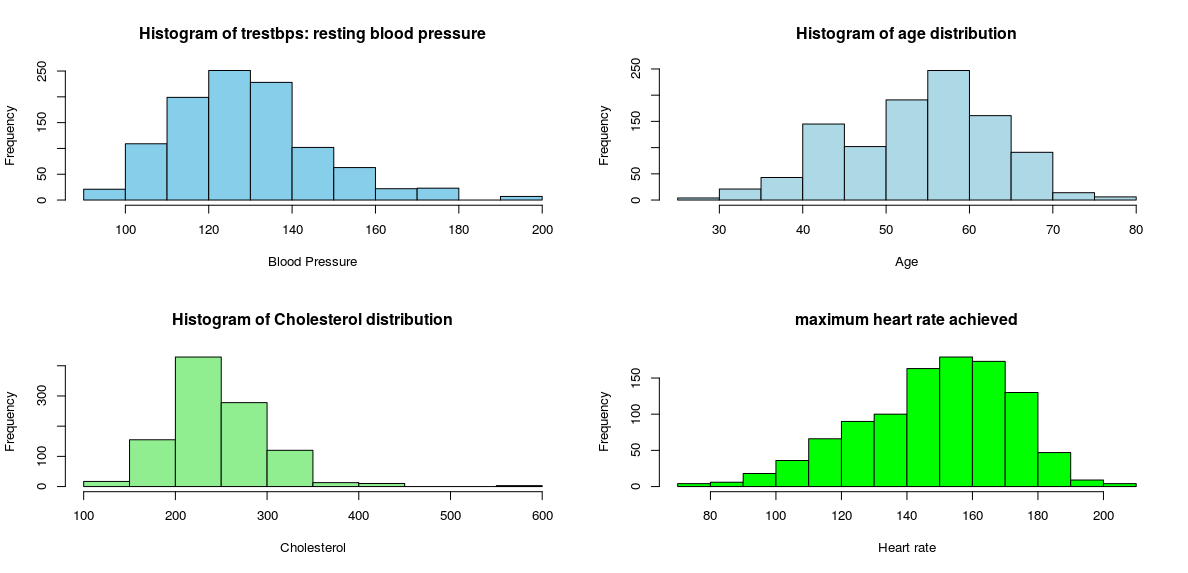
**"Histogram of age distribution")**

**hist(heart$chol, col = "lightgreen", xlab = "Cholesterol", main =**

**"Histogram of Cholesterol distribution")**

**hist(heart$thalach, col = "green", xlab = "Heart rate", main =**

**"maximum heart rate achieved")**

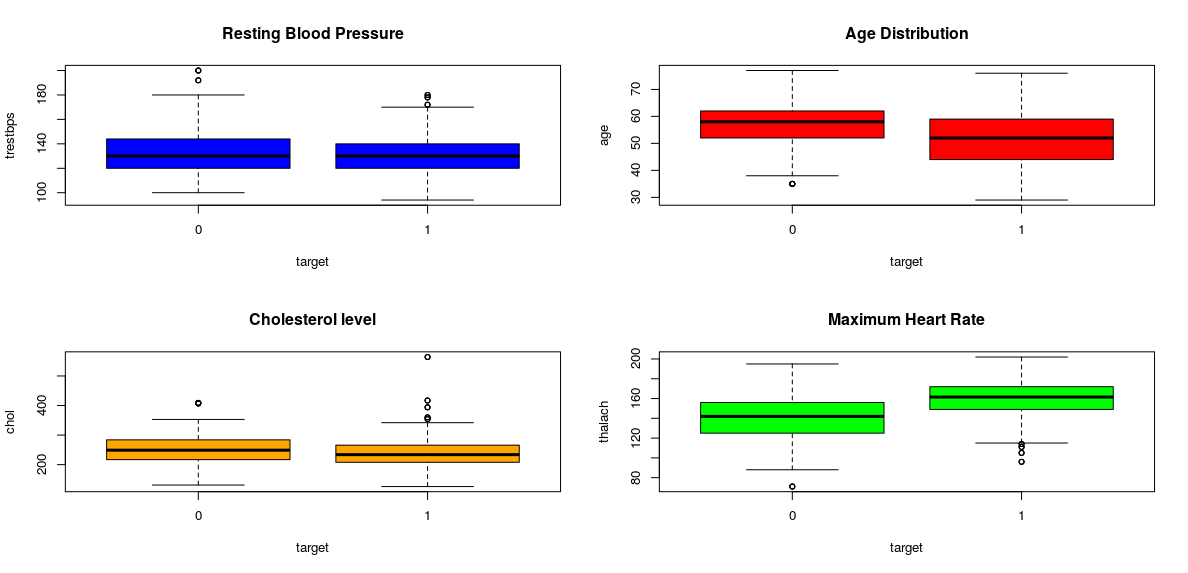


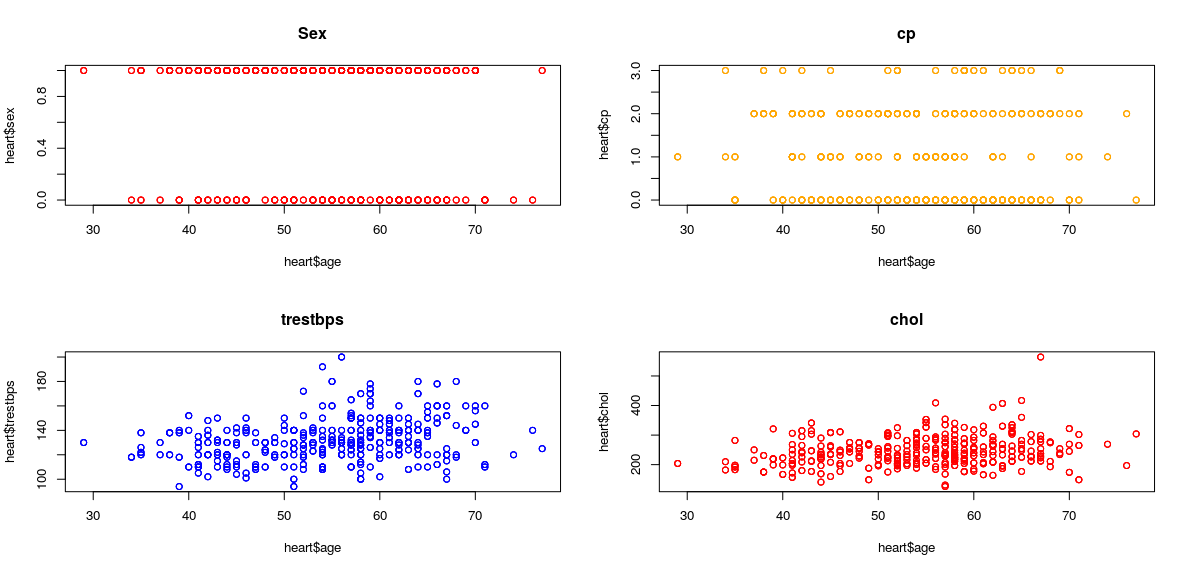
The Histogram of resting blood pressure is right skewed, showing that few patients have had an extremely higher blood pressure. When comparing Histograms separately for levels of target feature, we can see patients having a heat disease showed a higher blood pressure compared to patients not having a heart disease.

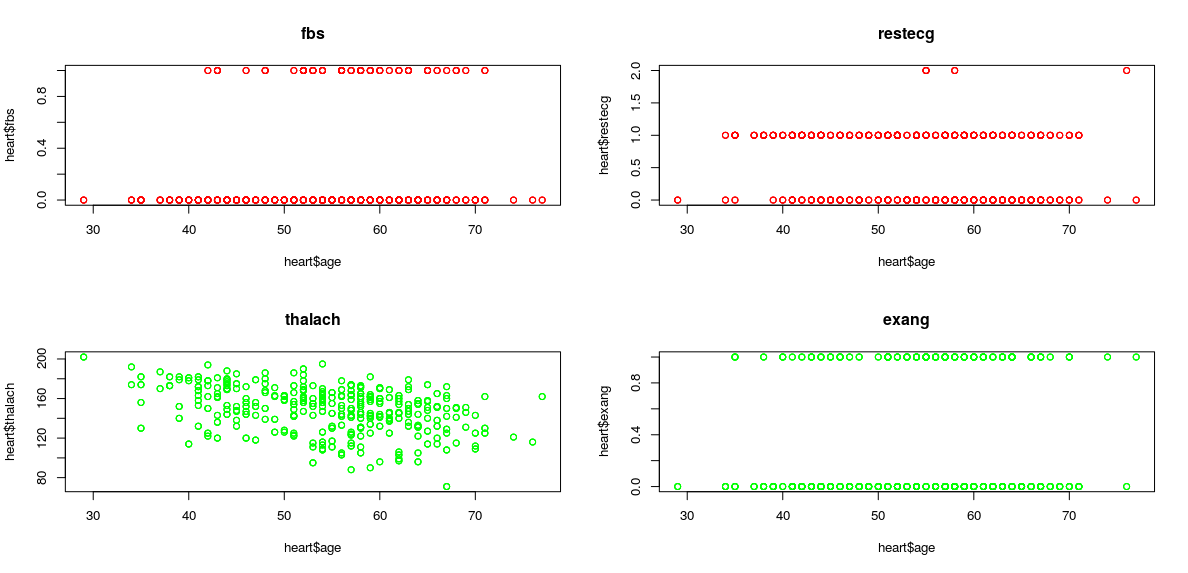
Patients from age 29 years to 77 years were included in this data set. The Histogram of age is little skewed to the left, showing the average age is little lower than the median age. Around 50% of patients’ age was in between 45-65 years. There is no visualized difference in ages for patients with or without heart disease. That is age may not be a major factor to diagnose a hear disease.

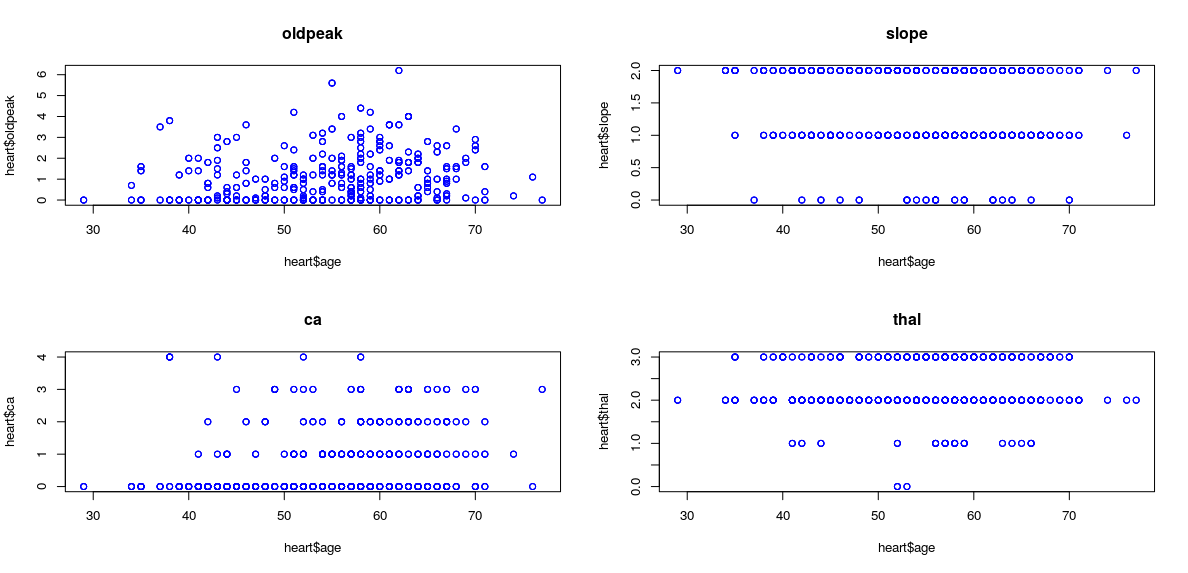
The distribution of patients’serum cholesterol level is highly right skewed, showing that few patients have had extremely high cholesterol levels. When we compare this distribution separately for patients with a heart disease and patients without a heart disease, the healthy patients’ distribution is leptokurtic. That means, there were many healthy peaple who had there cholesterol level around 200-220 mg/fl than patients with a hear disease.

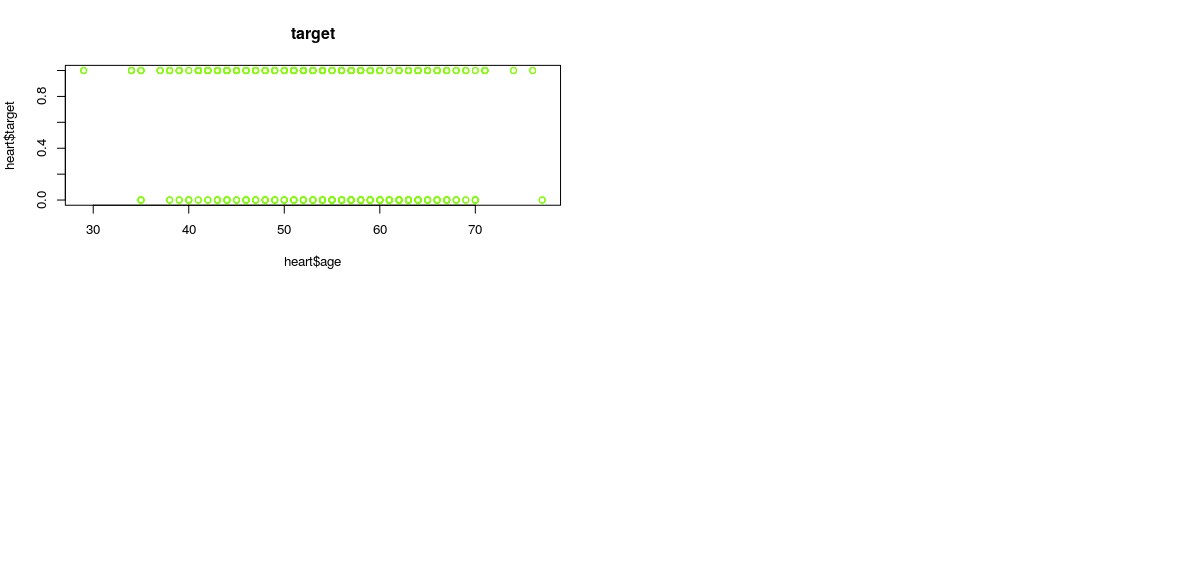
The histogram for maximum heart rate achieved by patients is left skewed as few patients showed a comparatively low heart rate. The separate Histograms for two levels of target feature show healthy people have had a quite higher maximum heart rate (around 160) compared to the maximum heart rate (150) of patients with a heart disease. Further, the Histogram for health people is leptokurtic (There are many people around the peak point).





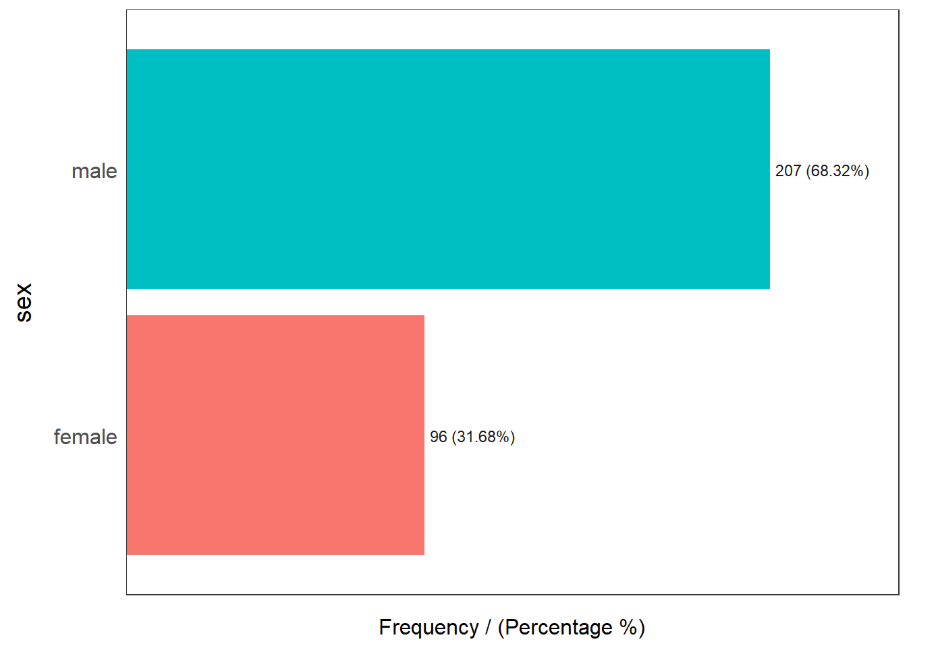


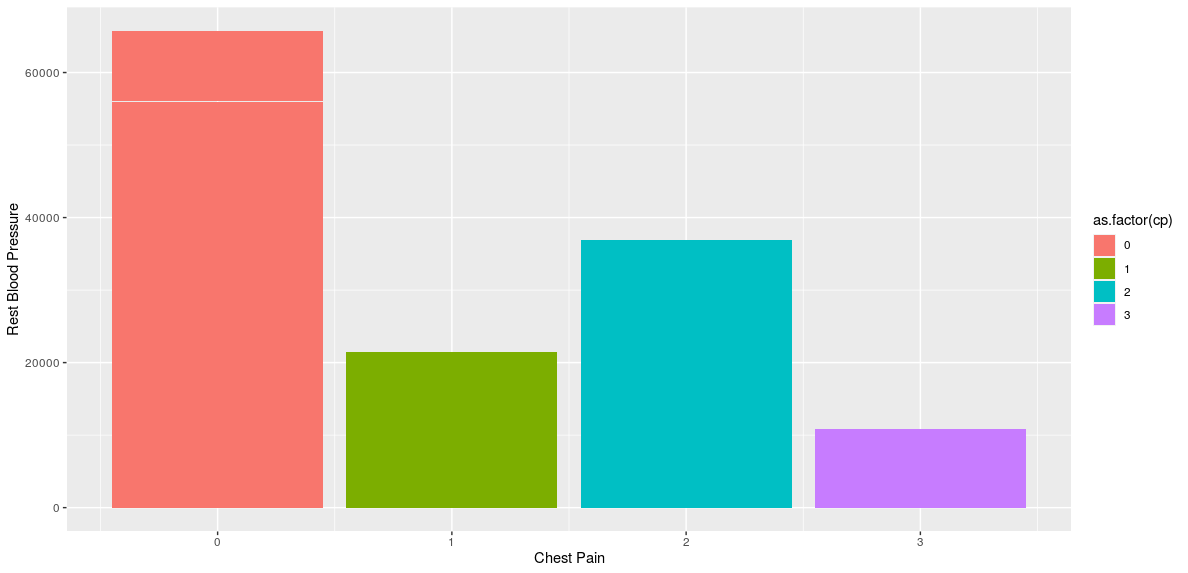




Age, Resting Blood Pressure (TRESTBPS),Serum Cholesterol (CHOL),Maximum Heart Rate Achieved (THALACH) and ST Depression Induced by Exercise Relative to Rest (OLDPEAK) There’s no significant linear correlation between any two of features. Blood pressure, cholesterol level and ST depression shows a weak positive linear correlation with age while maximum heart rate achieved shows a weak negative correlation with age. Maximum heart rate achieved also shows a weak negative linear relationship with ST depression induced by exercise relative to rest.

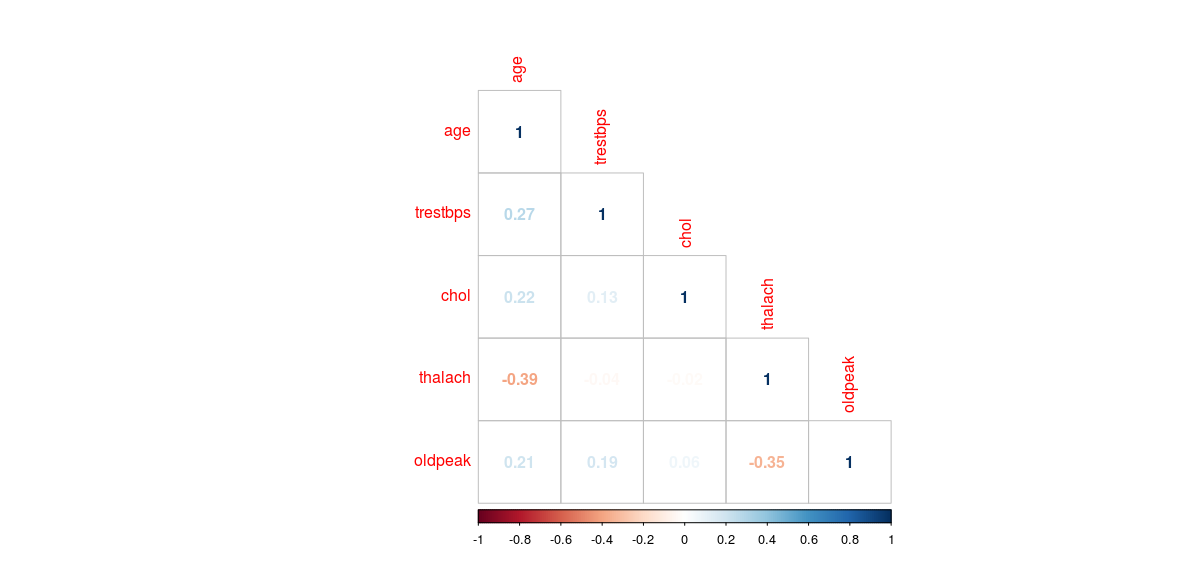
**freq(heart)**

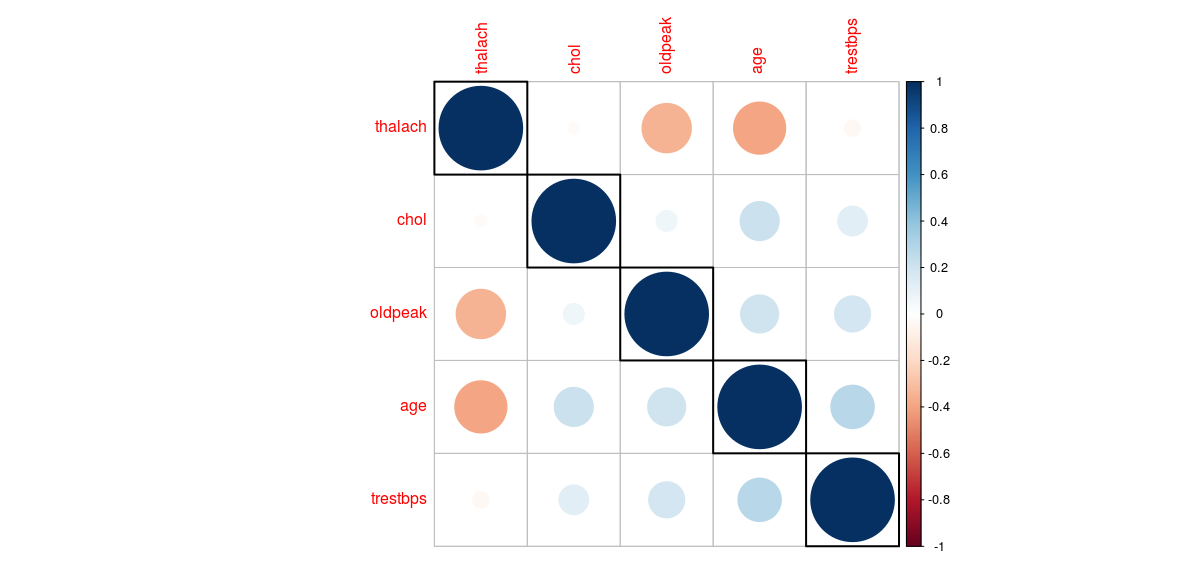




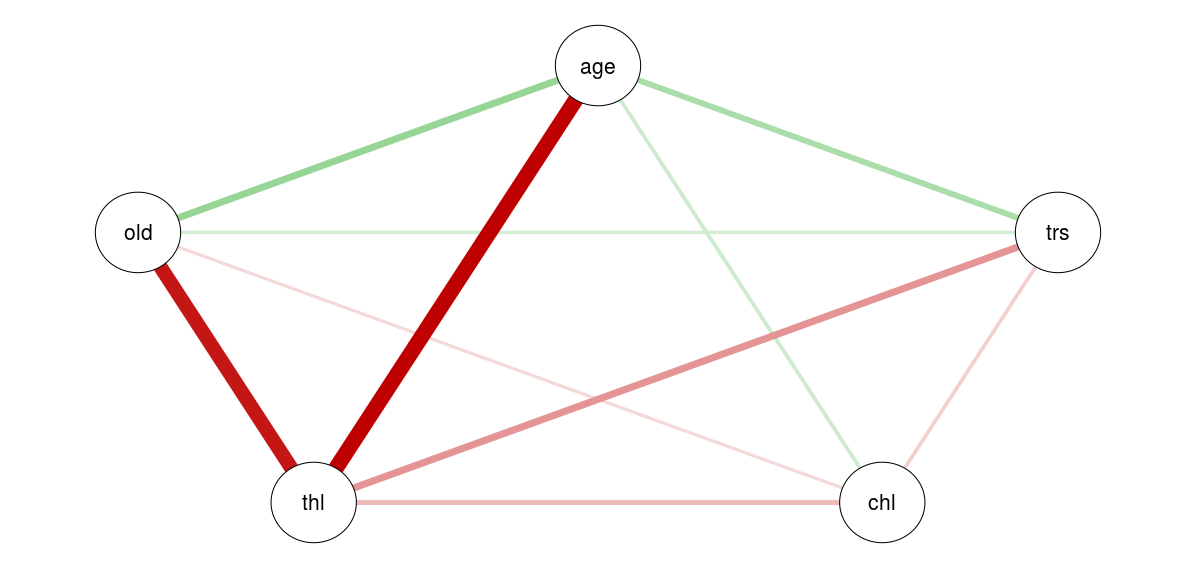
**corr\_mat <- cor(heart[,c(1,4,5,8,10)]) # numberic**

**corrplot(cor(heart[, c(1,4,5,8,10)]), type = "lower", method = "number") # excluding Outcome**





**qgraph(cor(corr\_mat))**

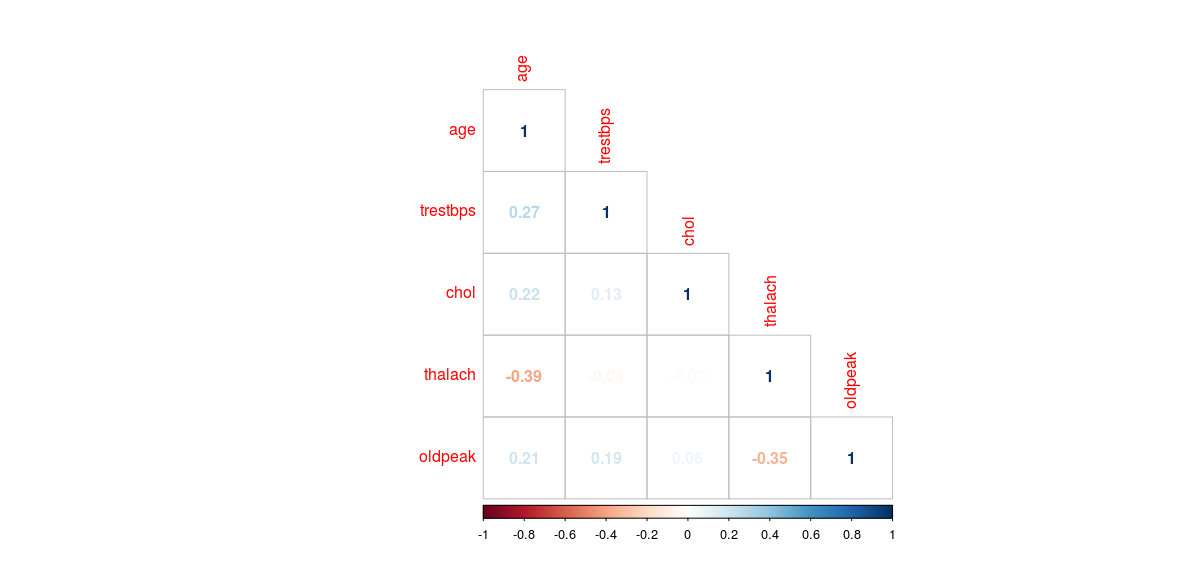


|  |
| --- |
| > cov(heart[c(1,4,5,8,10)])  age trestbps chol thalach oldpeak  age 82.306450 43.085733 102.890625 -81.446089 2.218825  trestbps 43.085733 306.835410 115.657227 -15.822822 3.857971  chol 102.890625 115.657227 2661.787109 -25.841797 3.933301  thalach -81.446089 -15.822822 -25.841797 529.263325 -9.456022  oldpeak 2.218825 3.857971 3.933301 -9.456022 1.380750  > var(heart[c(1,4,5,8,10)])  age trestbps chol thalach oldpeak  age 82.306450 43.085733 102.890625 -81.446089 2.218825  trestbps 43.085733 306.835410 115.657227 -15.822822 3.857971  chol 102.890625 115.657227 2661.787109 -25.841797 3.933301  thalach -81.446089 -15.822822 -25.841797 529.263325 -9.456022  oldpeak 2.218825 3.857971 3.933301 -9.456022 1.380750  > cor(heart[c(1,4,5,8,10)])  age trestbps chol thalach oldpeak  age 1.0000000 0.27112141 0.21982253 -0.39022708 0.20813668  trestbps 0.2711214 1.00000000 0.12797743 -0.03926407 0.18743411  chol 0.2198225 0.12797743 1.00000000 -0.02177209 0.06488031  thalach -0.3902271 -0.03926407 -0.02177209 1.00000000 -0.34979616  oldpeak 0.2081367 0.18743411 0.06488031 -0.34979616 1.00000000 |
|  |
| |  | | --- | | > | |

Correlations between numeric variables

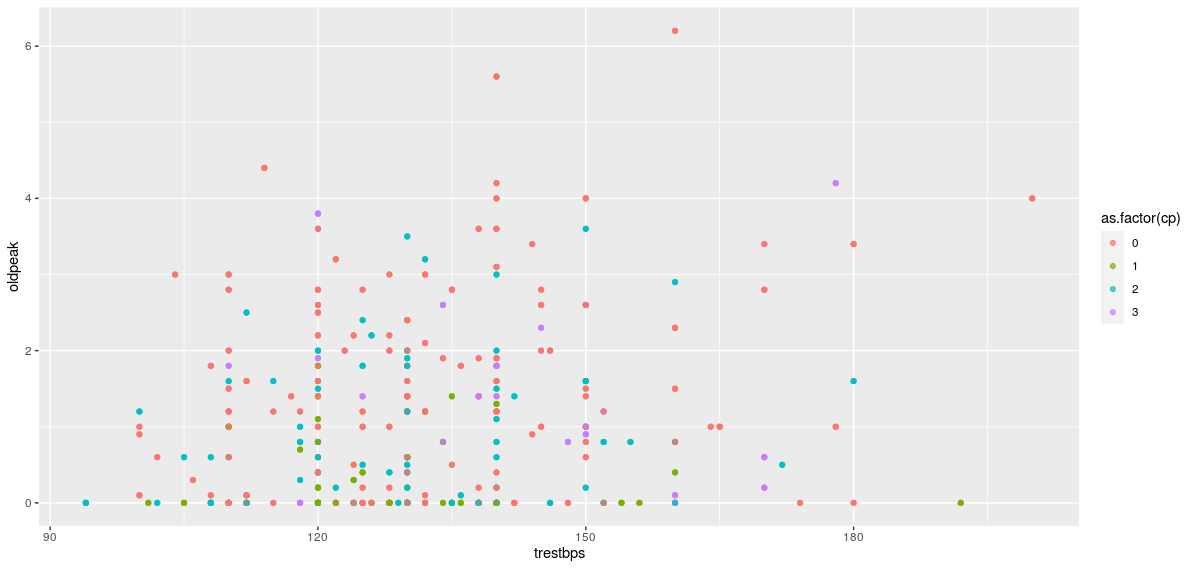
corr\_mat <- cor(heart[,c(1,4,5,8,10)])

corrplot(cor(heart[, c(1,4,5,8,10)]), type = "lower", method = "number")

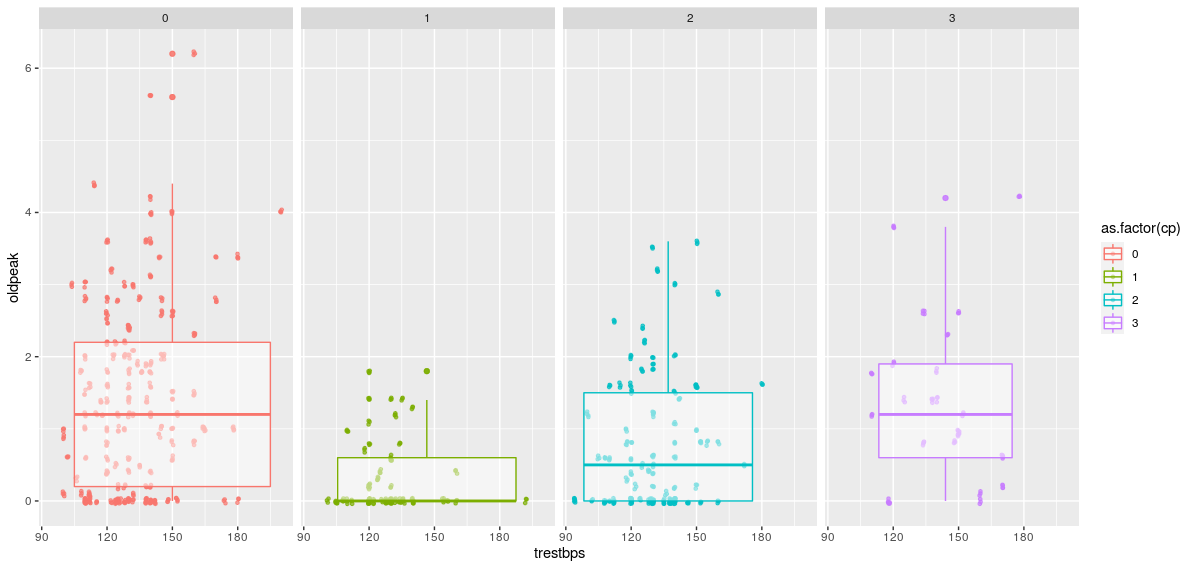


**p <- ggplot(heart, aes(x = trestbps, y = oldpeak,col=as.factor(cp)))**

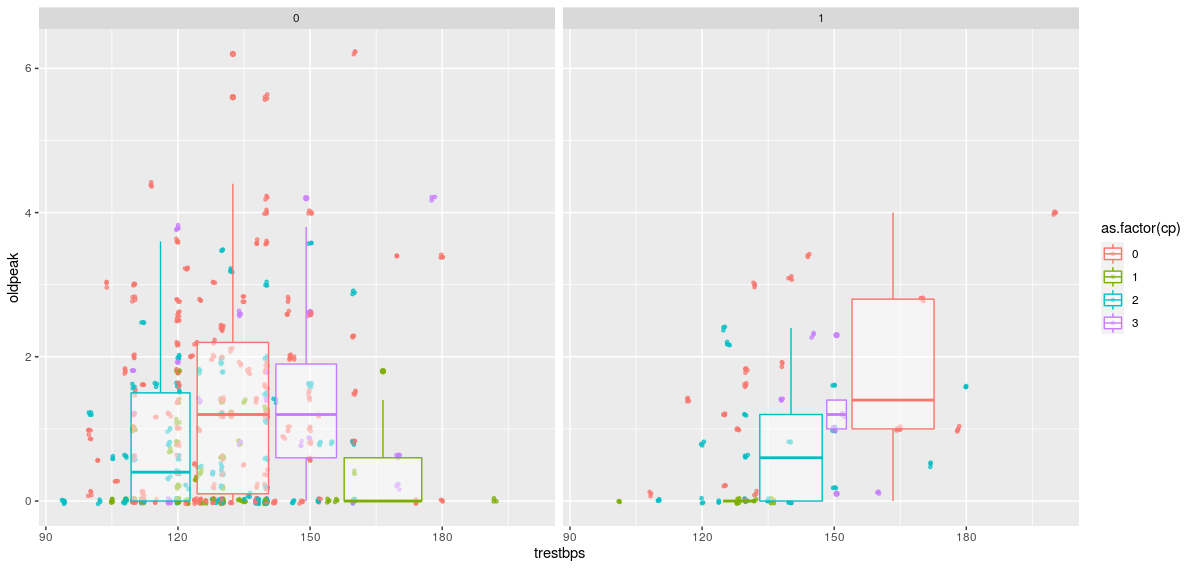
**p + geom\_point(alpha = 0.7, size = 1.5)**

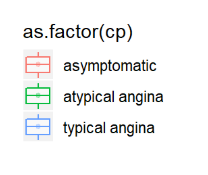


**p + geom\_jitter(alpha = 0.7, size = 1) + stat\_boxplot(alpha = 0.5) + facet\_grid(.~ cp)**



**p + geom\_jitter(alpha = 0.7, size = 1) + stat\_boxplot(alpha = 0.5) + facet\_grid(.~ fbs )**





As per the three dimensional bar chart depicted below there is no men in the data set with resting ECG result level 1 and having a exercise induced angina.The most of female who’s not having exercise induced angina haven’t had a heart disease for almost all the levels of resting ECG results. Some numerical features with few values (2-4) were converted to factors to improve the convenience of the analysis. Two categorical features ca and thal had few missing values and due to that reason it cardinality also get wrong. We estimated values for missing cells and regenerated the cardinality too. The target feature was num with 4 cardinalities. We regenerated this feature with two cardinalities and renamed it as target. From the data exploration, we found that sex,serum cholesterol level, maximum heart rate achieved, chest pain type, resting ECG result, exercise induced angina,ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment and The heart status as retrieved from Thallium test may be potentially useful features in predicting the diagnosis of patient’s heart disease.