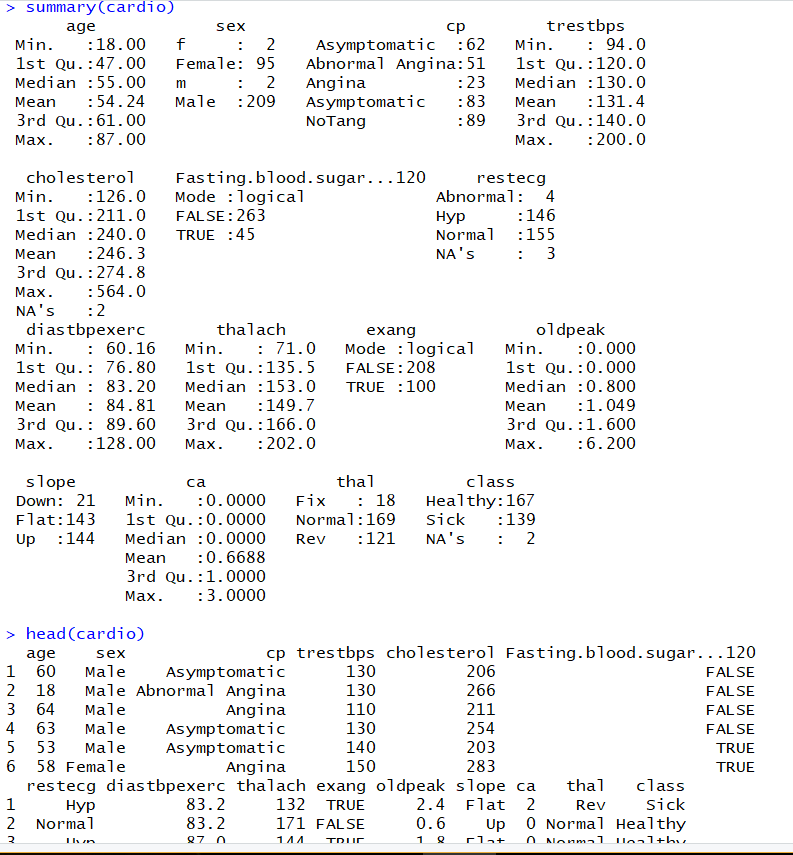
#readfile from remote location

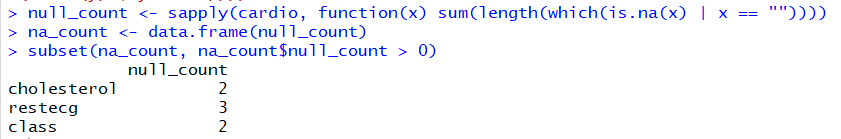
cardio <- read.table(file = "https://drive.google.com/uc?export=download&id=1aeJkrlDpK7LNptsqHdoHzwG2O4T7neDI", header=TRUE, sep =",", stringsAsFactors = TRUE)

#summary of cardio dataset

summary(cardio)



#getting missing values

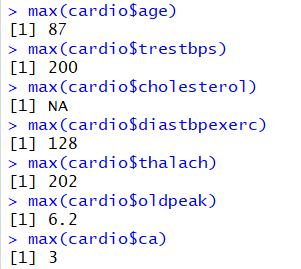


**1b** #percentage of missing values



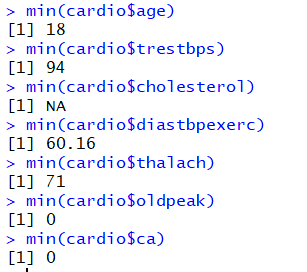
#as missing columns = 7, entries in data = 308 \*15, missing % = 7/4620 \*100 = 0.1515

**1c** #getting max for numeric values.

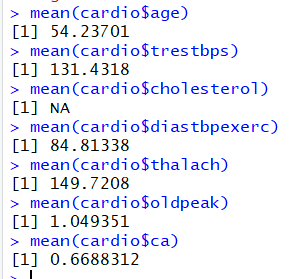


#max, min, sd for cholesterol is NA could be due to missing values

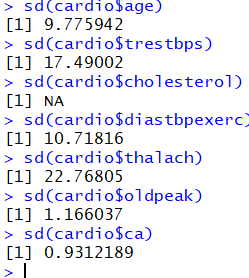
**1c** #getting min for numeric values



**1c** #Mean for numeric values



**1c** #median(sd) for numeric columns



**1c** #mode

**1d** #qqnorm normality test

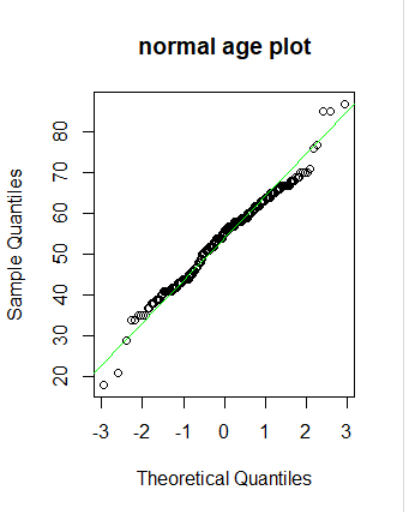
norm\_q <- function(dataIn, labelIn) {

qqnorm(dataIn, main = labelIn)

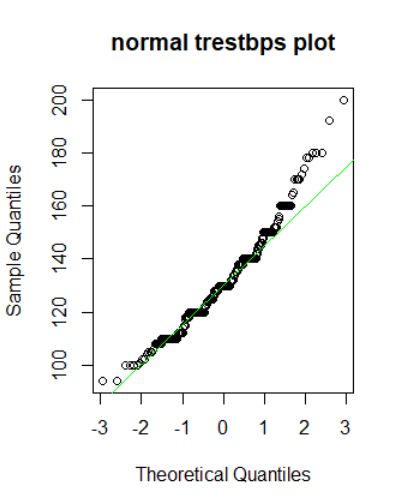
qqline(dataIn, col='green')

}

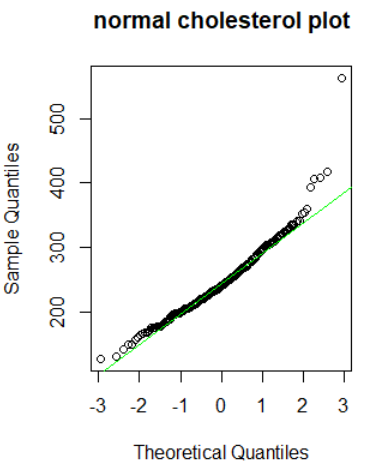
norm\_q(cardio$age, "normal age plot")



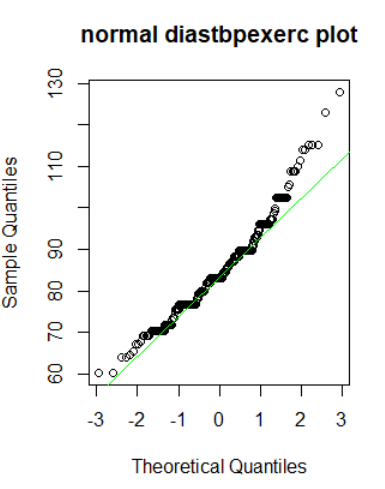
norm\_q(cardio$trestbps, "normal trestbps plot")



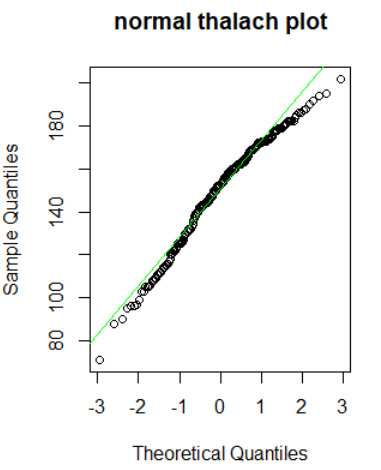
norm\_q(cardio$cholesterol, "normal cholesterol plot")



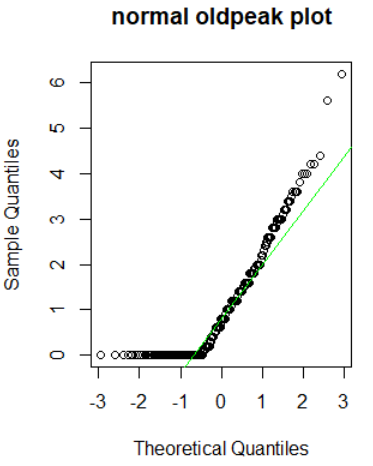
norm\_q(cardio$diastbpexerc, "normal diastbpexerc plot")



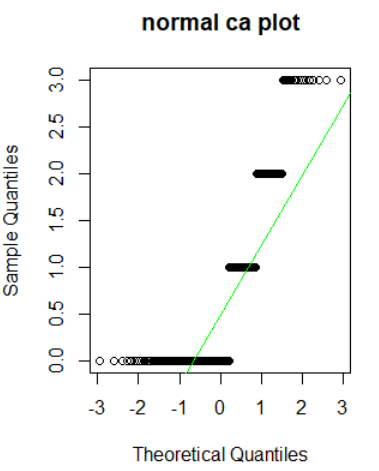
norm\_q(cardio$thalach, "normal thalach plot")



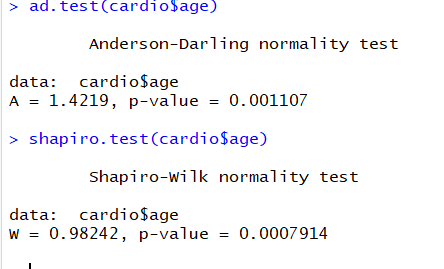
norm\_q(cardio$oldpeak, "normal oldpeak plot")



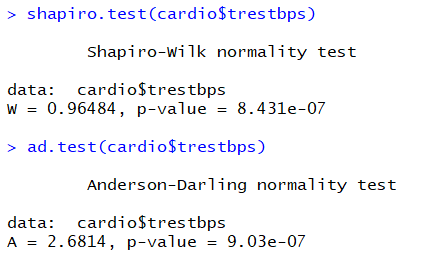
norm\_q(cardio$ca, "normal ca plot")



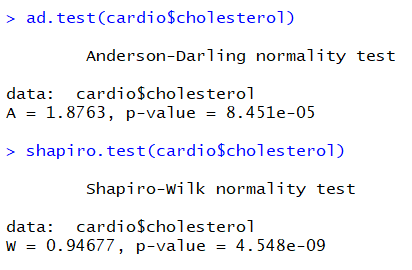
**1d** #Anderson-Darling and shapiro-wilk normality test for age



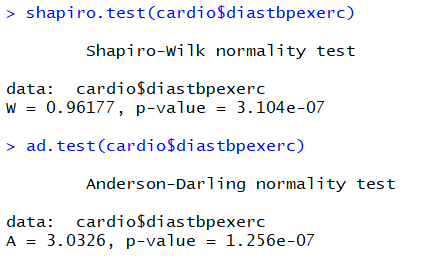
**1d** #Anderson-Darling and shapiro-wilk normality test for trestbps



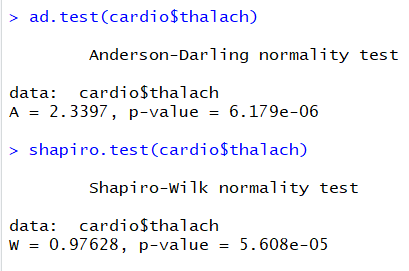
**1d** #Anderson-Darling and shapiro-wilk normality test for cholesterol



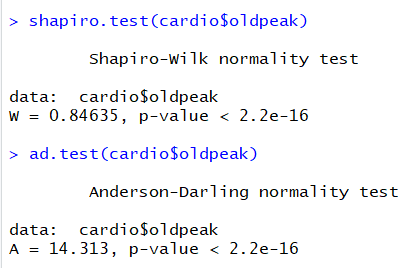
**1d** #Anderson-Darling and shapiro-wilk normality test for diastbpexerc



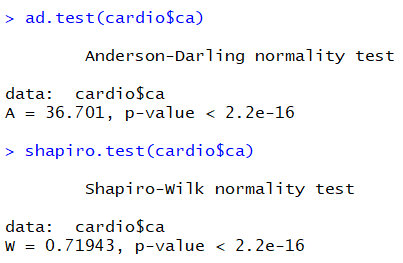
**1d** #Anderson-Darling and shapiro-wilk normality test for thalach



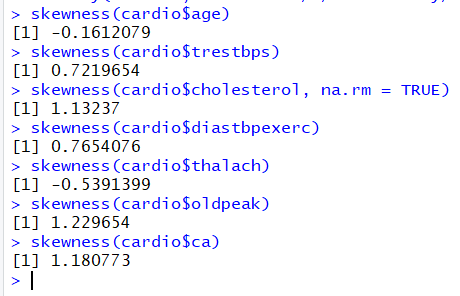
**1d** #Anderson-Darling and shapiro-wilk normality test for oldpeak



**1d** #Anderson-Darling and shapiro-wilk normality test for ca



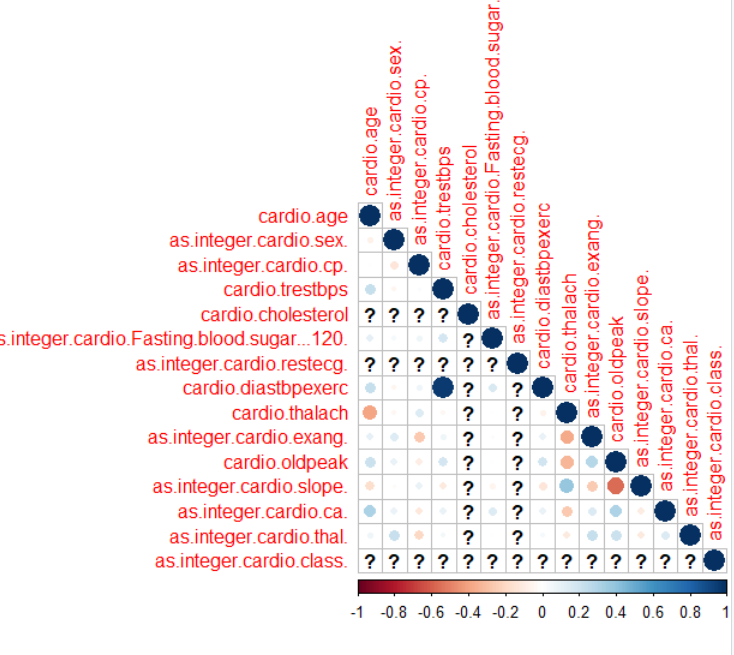
**1e** skewness



#trestbps, cholesterol, diastbpexerc, oldpeak and ca are positively skewed i.e. skewed to the right.

While age and thalach are negatively skewed and as such skewed to the left.

**1f** correlations



**#’f’ to female, ‘m’ to males**

> cardio[cardio=="f"]<-"Female"

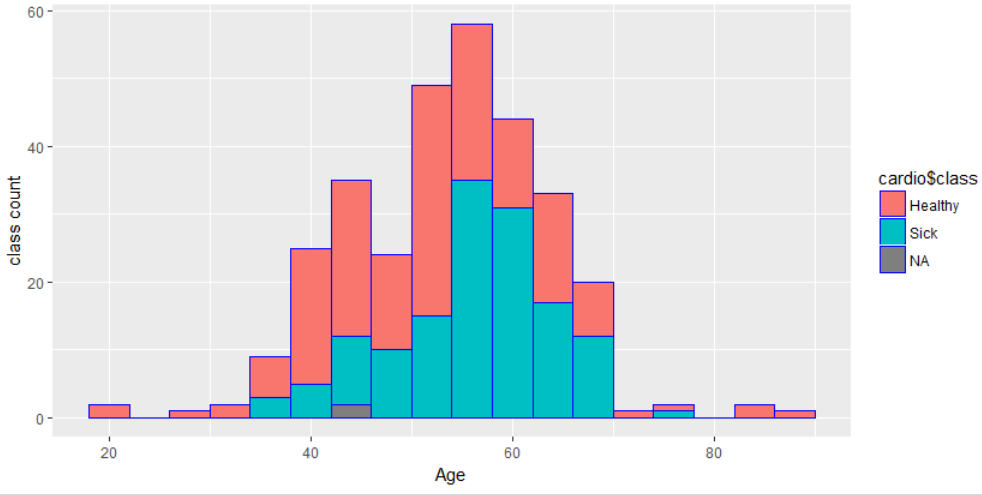
> cardio[cardio=="m"]<-"Male"

> cardio$sex <- factor(cardio$sex)

**2 histograms**

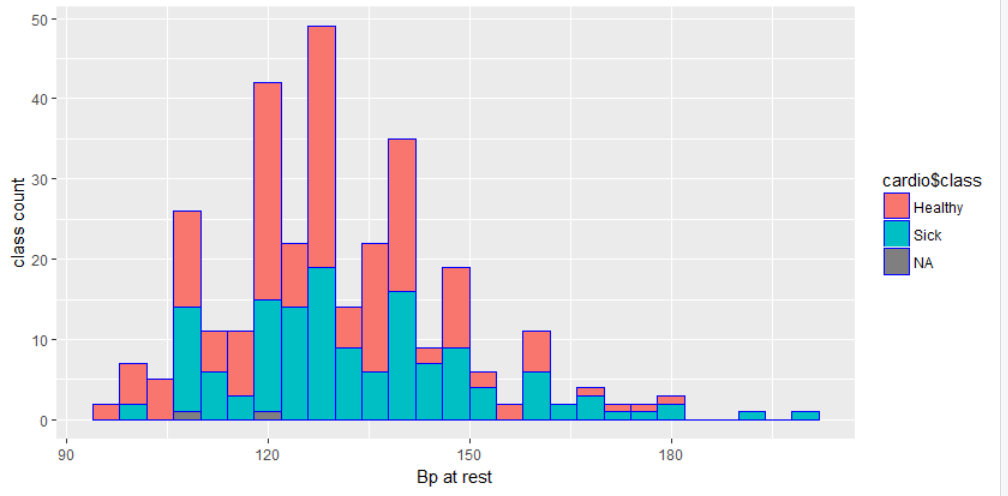
**#histogram for age by class**

ggplot(cardio, aes(x = cardio$age)) + geom\_histogram(colour = "blue", aes(fill = cardio$class), binwidth = 4) + labs(x = "Age", y = "class count")



**#histogram for trestbps by class**

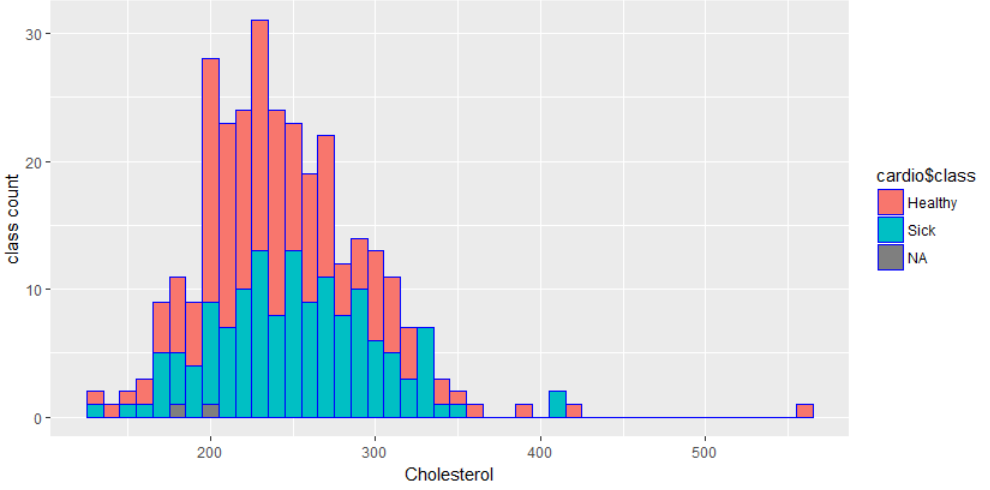
ggplot(cardio, aes(x = cardio$trestbps)) + geom\_histogram(colour = "blue", aes(fill = cardio$class), binwidth = 4) + labs(x = "Bp at rest", y = "class count")



**#histogram for cholesterol by class**

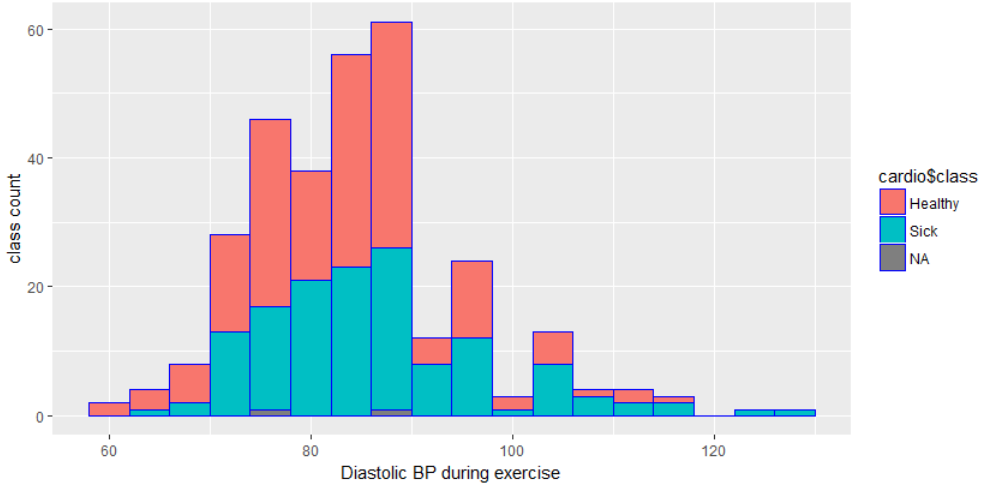
ggplot(cardio, aes(x = cardio$cholesterol)) + geom\_histogram(colour = "blue", aes(fill = cardio$class), binwidth = 10) + labs(x = "Cholesterol", y = "class count")

## Warning message:Removed 2 rows containing non-finite values (stat\_bin).



**#histogram for diastbpexerc by class**

ggplot(cardio, aes(x = cardio$diastbpexerc)) + geom\_histogram(colour = "blue", aes(fill = cardio$class), binwidth = 4) + labs(x = "Diastolic BP during exercise", y = "class count")

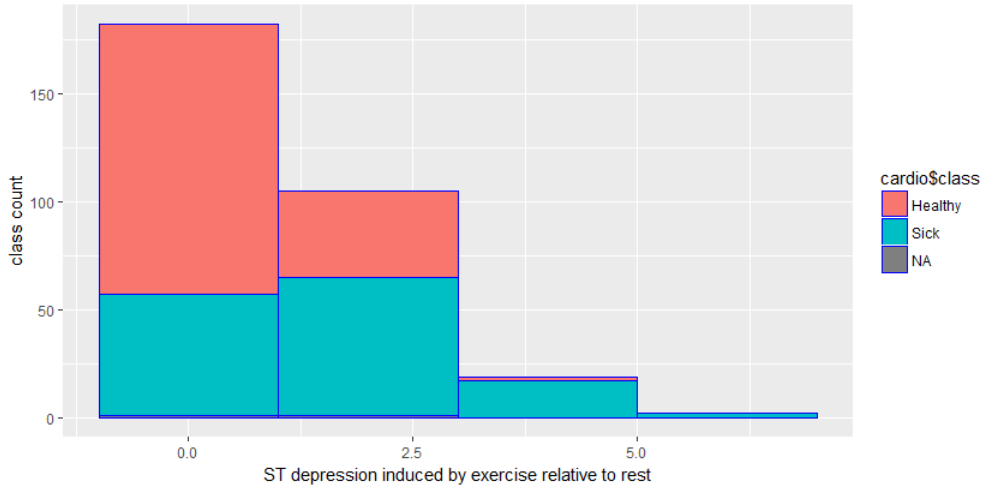


**#histogram for thalach by class**

ggplot(cardio, aes(x = cardio$thalach)) + geom\_histogram(colour = "blue", aes(fill = cardio$class), binwidth = 4) + labs(x = "Max heart rate achieved", y = "class count")

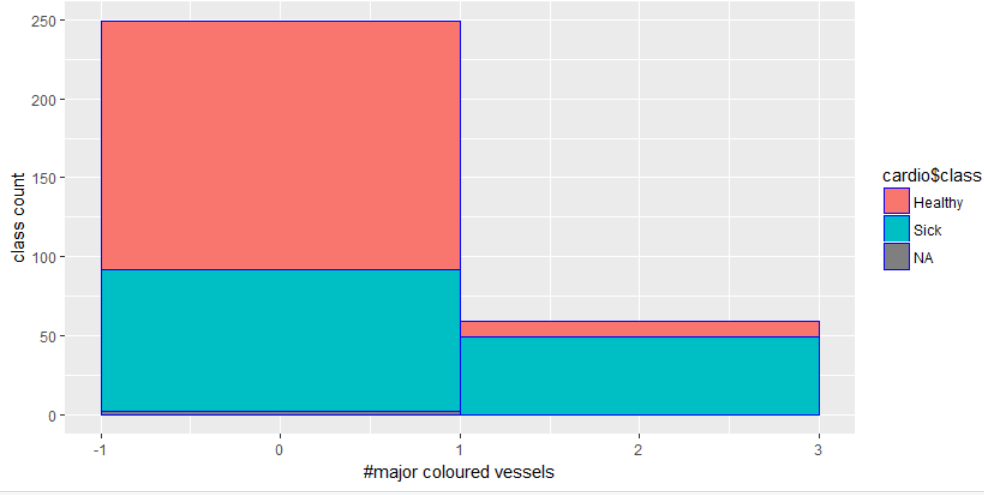


**#histogram for oldpeak by class**

ggplot(cardio, aes(x = cardio$oldpeak)) + geom\_histogram(colour = "blue", aes(fill = cardio$class), binwidth = 2) + labs(x = "ST depression induced by exercise relative to rest", y = "class count")

**#histogram for ca by class**

ggplot(cardio, aes(x = cardio$ca)) + geom\_histogram(colour = "blue", aes(fill = cardio$class), binwidth = 2) + labs(x = "#major coloured vessels", y = "class count")



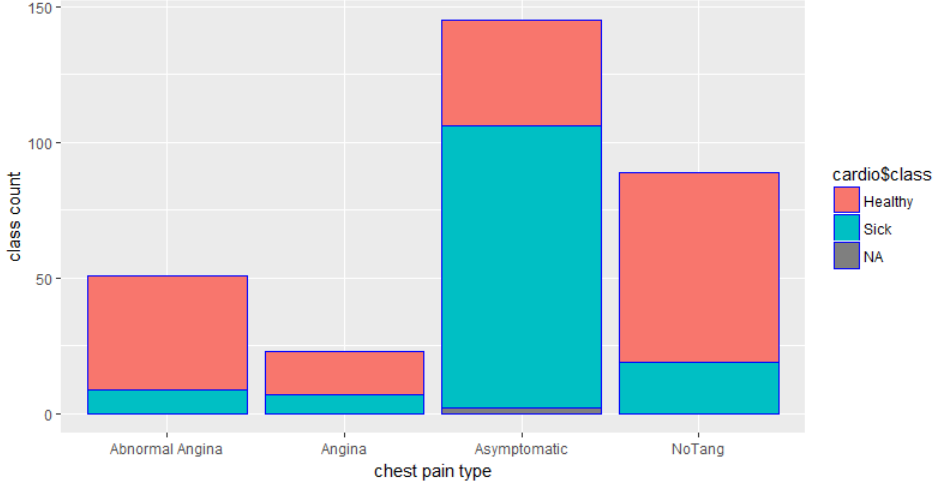
**3 Bar charts**

**#bar for cp by class**

#transfoming “ asymptomaic” to “Asymptomatic” to reveal 4 categories

cardio[cardio==" Asymptomatic"]<-"Asymptomatic"

ggplot(cardio, aes(x = cardio$cp)) + geom\_bar(colour = "blue", aes(fill = cardio$class)) + labs(x = "chest pain type", y = "class count")



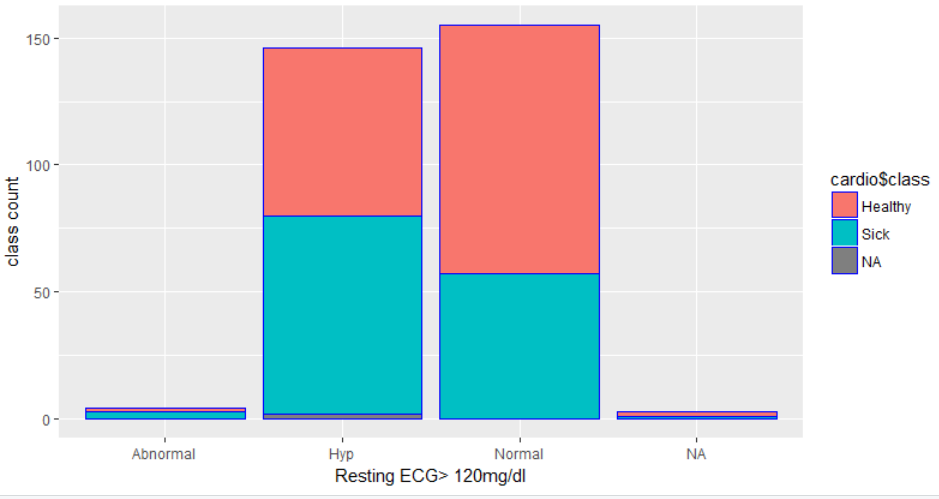
**#bar for Fasting blood sugar > 120mg/dl by class**

ggplot(cardio, aes(x = cardio$Fasting.blood.sugar...120)) + geom\_bar(colour = "blue", aes(fill = cardio$class)) + labs(x = "Fasting blood sugar > 120mg/dl", y = "class count")



**#bar for restecg by class**

ggplot(cardio, aes(x = cardio$restecg)) + geom\_bar(colour = "blue", aes(fill = cardio$class)) + labs(x = "Resting ECG> 120mg/dl", y = "class count")



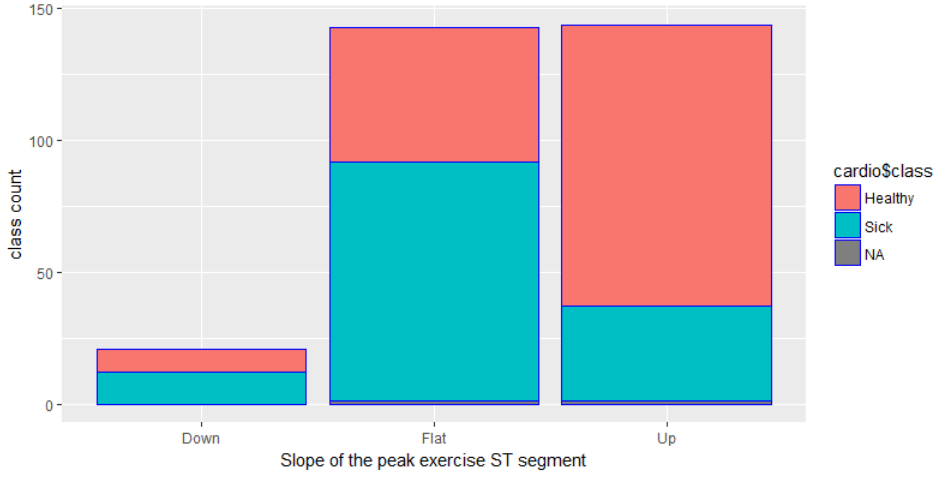
**#bar for exang by class**

ggplot(cardio, aes(x = cardio$exang)) + geom\_bar(colour = "blue", aes(fill = cardio$class)) + labs(x = "Exercise induced Angina", y = "class count")



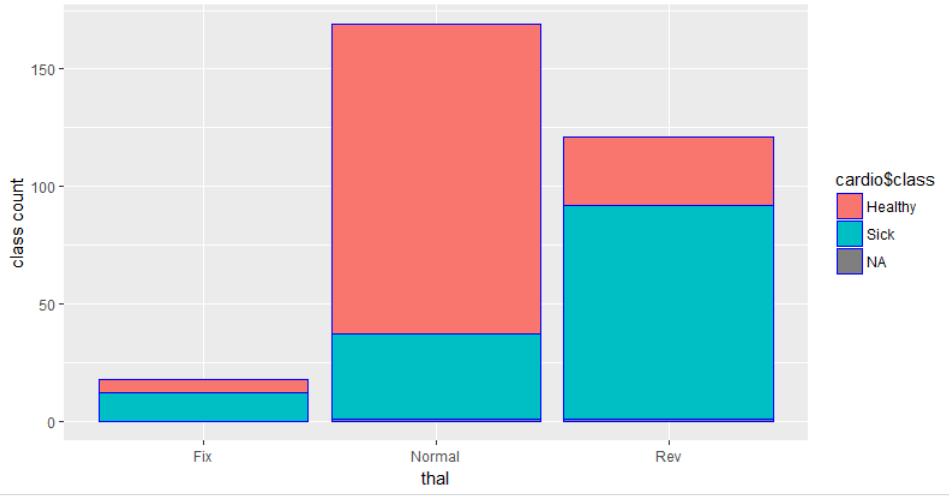
**#bar for slope by class**

> ggplot(cardio, aes(x = cardio$slope)) + geom\_bar(colour = "blue", aes(fill = cardio$class)) + labs(x = "Slope of the peak exercise ST segment

+ ", y = "class count")

**#bar for thal by class**

ggplot(cardio, aes(x = cardio$thal)) + geom\_bar(colour = "blue", aes(fill = cardio$class)) + labs(x = "thal", y = "class count")



**4 outliers**

#getting outliers by boxplot

> boxplot(cardio$age)$out

[1] 18 21 85 87 85

> boxplot(cardio$trestbps)$out

[1] 180 174 178 192 180 178 180 172 200

> boxplot(cardio$cholesterol)$out

[1] 417 407 564 394 409

> boxplot(cardio$diastbpexerc)$out

[1] 108.80 115.20 108.80 111.36 113.92 122.88

[7] 115.20 113.92 115.20 108.80 108.80 110.08

[13] 128.00

> boxplot(cardio$thalach)$out

[1] 88 71

> boxplot(cardio$oldpeak)$out

[1] 6.2 5.6 4.2 4.2 4.4

> boxplot(cardio$ca)$out

[1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3

#getting outliers using z\_score

z\_score <- function(x) {scale(x, center = TRUE, scale = TRUE) }

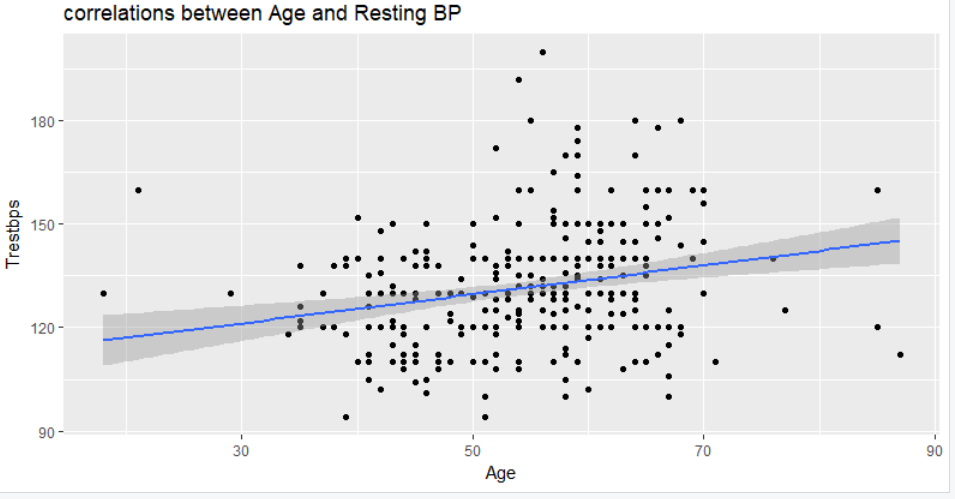
z\_range <- function(x){ table(z\_score(x) > -3 & z\_score(x) < 3) }

|  |
| --- |
| > z\_range(cardio$age) #found 5 outliers in age  FALSE TRUE  5 303  > z\_range(cardio$cholesterol) #found 4 outliers  FALSE TRUE  4 302  > z\_range(cardio$trestbps) #found 2 outliers  FALSE TRUE  2 306  > z\_range(cardio$diastbpexerc) #found 2 outliers  FALSE TRUE  2 306  > z\_range(cardio$thalach) #found 1 outlier  FALSE TRUE  1 307  > z\_range(cardio$oldpeak) #found 2 outliers  FALSE TRUE  2 306  > z\_range(cardio$ca) #found no(0) outliers  TRUE  308 |
|  |
| |  | | --- | |  | |

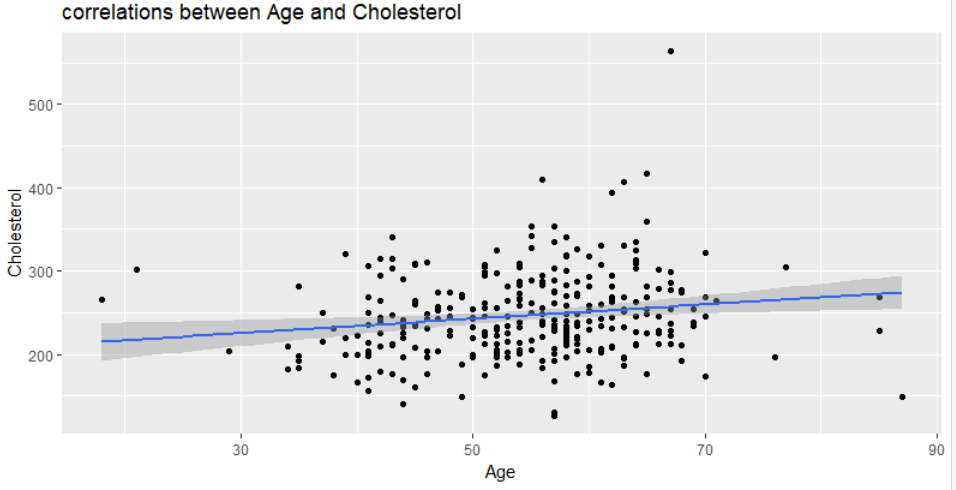
**5 correlation between numeric variables**

> PairedScatterPlot <- function(Xin, Yin, myX, myY,myTitle) { ggplot(cardio, aes(Xin, Yin )) + geom\_point() + labs(x = myX, y = myY, title = myTitle) + geom\_smooth(method = "lm") }

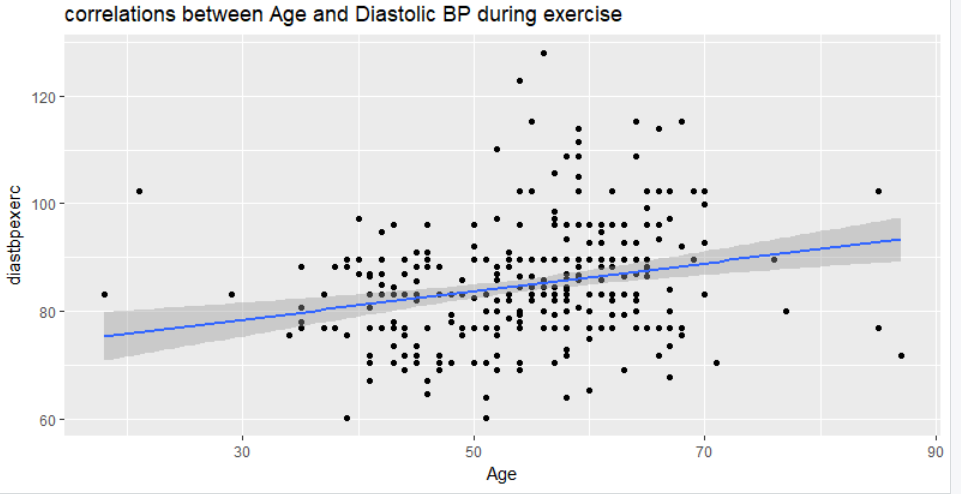
> PairedScatterPlot(cardio$age,cardio$trestbps, "Age", "Trestbps", "correlations between Age and Resting BP")



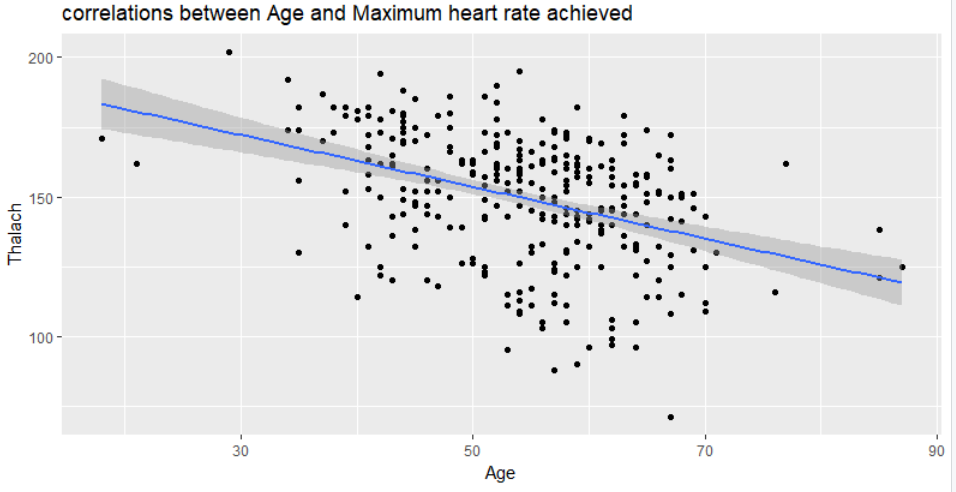
> PairedScatterPlot(cardio$age,cardio$cholesterol, "Age", "Cholesterol", "correlations between Age and Cholesterol")



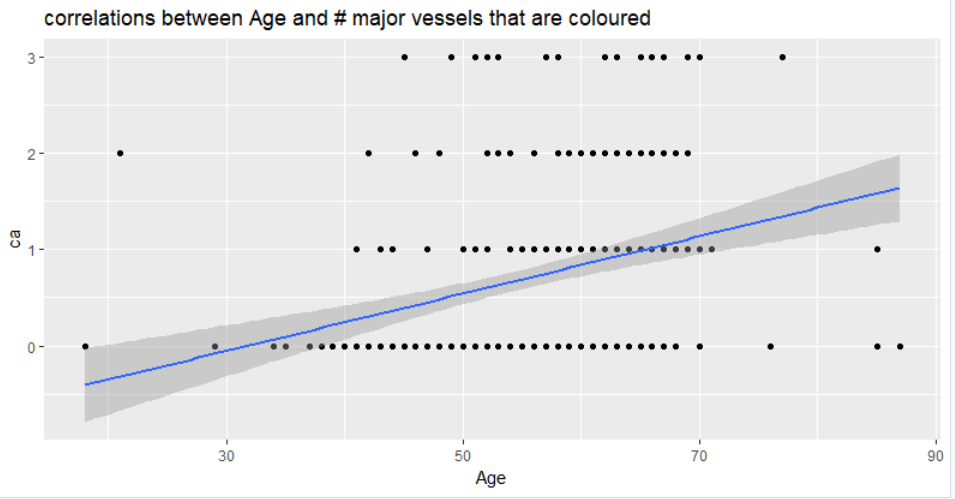
> ggplot(cardio, aes(cardio$age, cardio$diastbpexerc )) + geom\_point() + labs(x = "Age", y = "diastbpexerc", title = "correlation between Age and Diastolic BP during exercise")



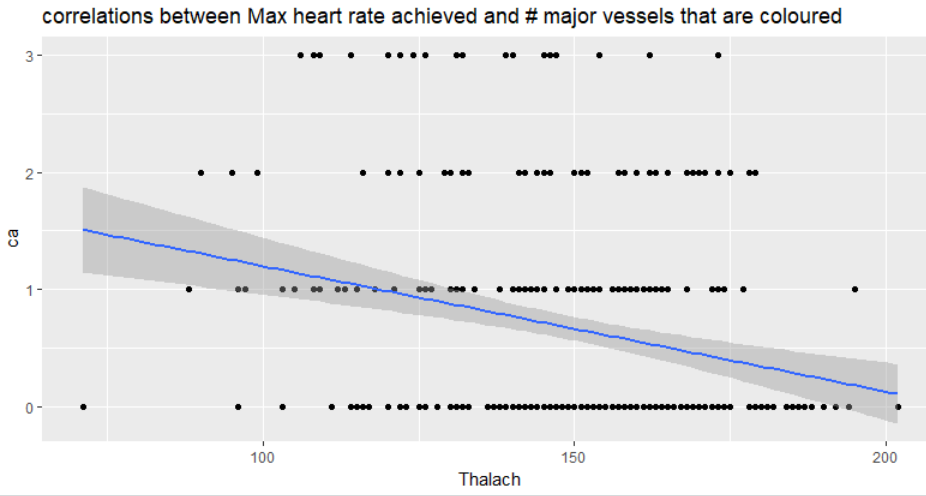
> PairedScatterPlot(cardio$age,cardio$thalach, "Age", "Thalach", "correlations between Age and Maximum heart rate achieved")



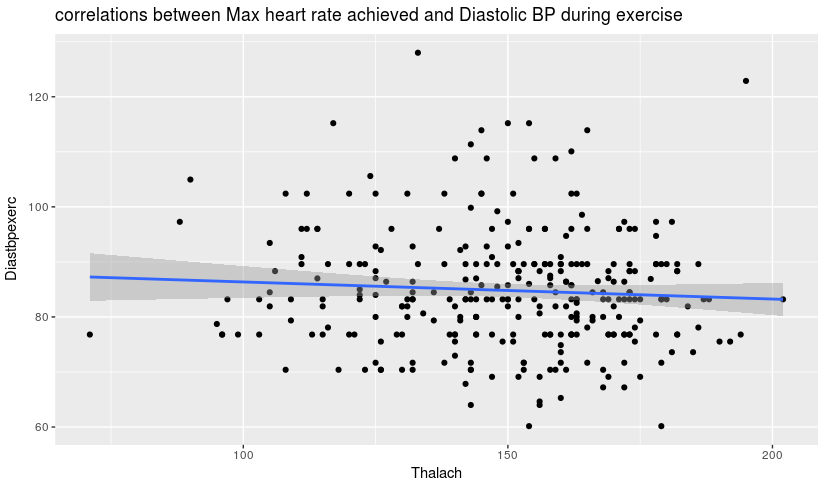
> PairedScatterPlot(cardio$age,cardio$ca, "Age", "ca", "correlations between Age and # major vessels that are coloured")

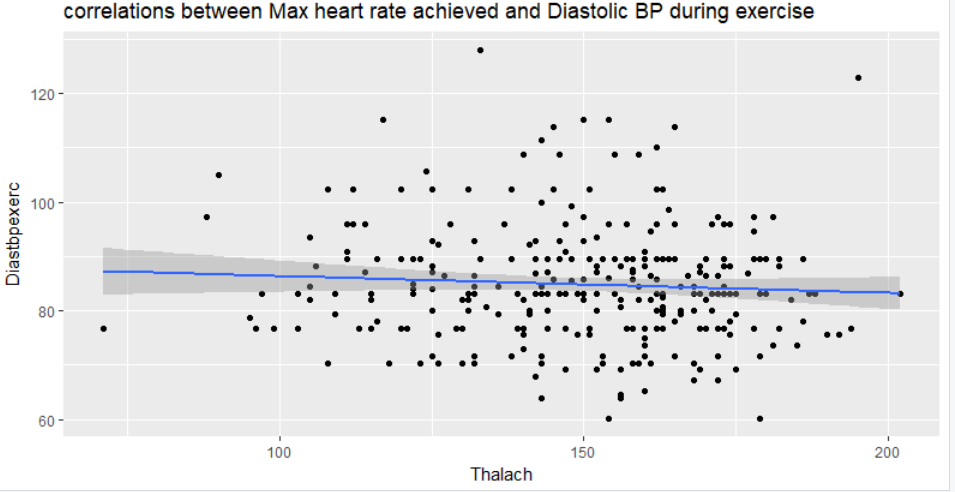


> PairedScatterPlot(cardio$thalach,cardio$ca, "Thalach", "ca", "correlations between Max heart rate achieved and # major vessels that are coloured")

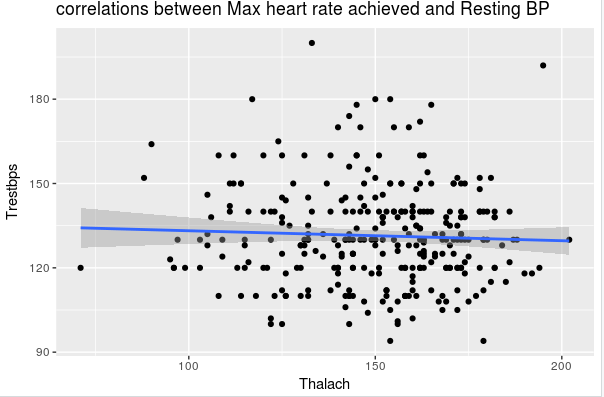


> PairedScatterPlot(cardio$thalach,cardio$diastbpexerc, "Thalach", "Diastbpexerc", "correlations between Max heart rate achieved and Diastolic BP during exercise")

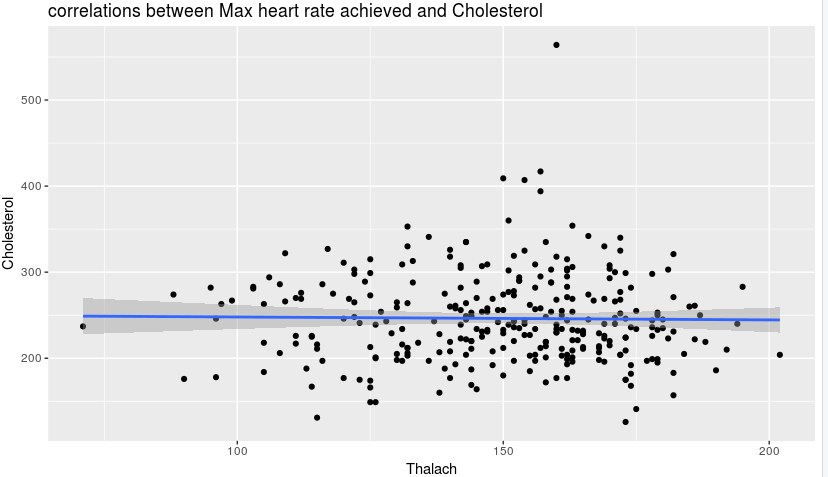




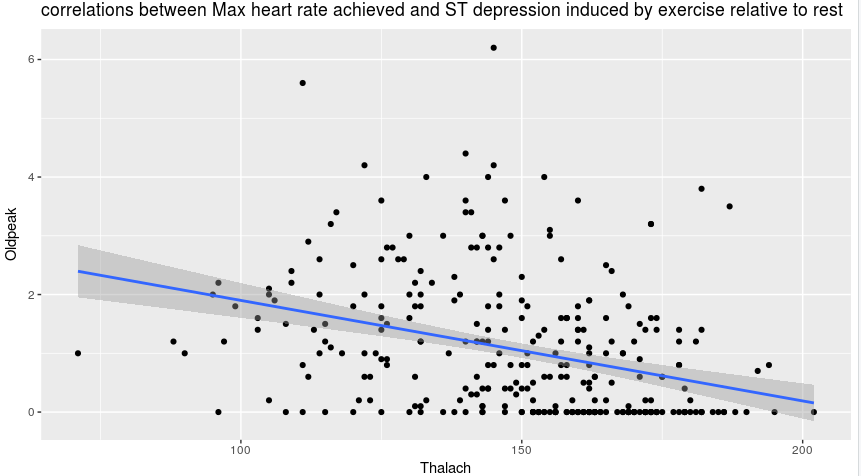
PairedScatterPlot(cardio$thalach,cardio$trestbps, "Thalach", "Trestbps", "correlations between Max heart rate achieved and Resting BP")



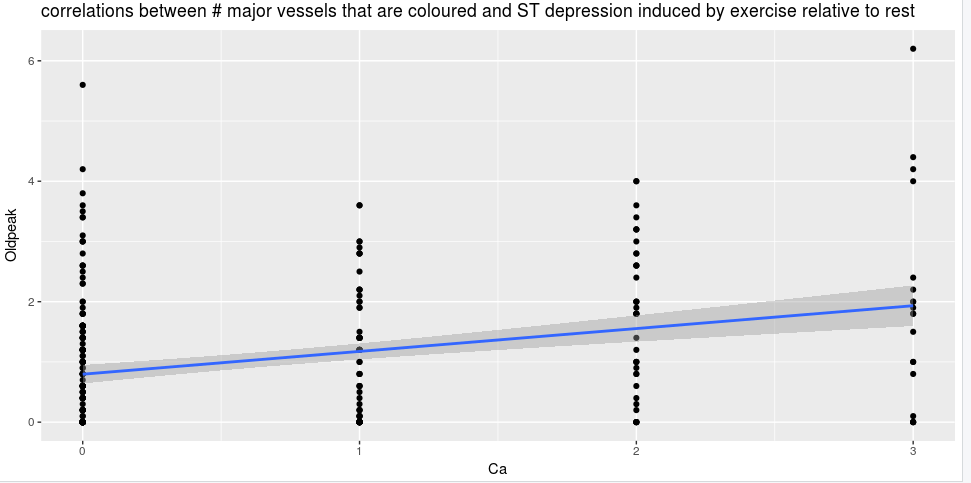
> PairedScatterPlot(cardio$thalach,cardio$cholesterol, "Thalach", "Cholesterol", "correlations between Max heart rate achieved and Cholesterol")



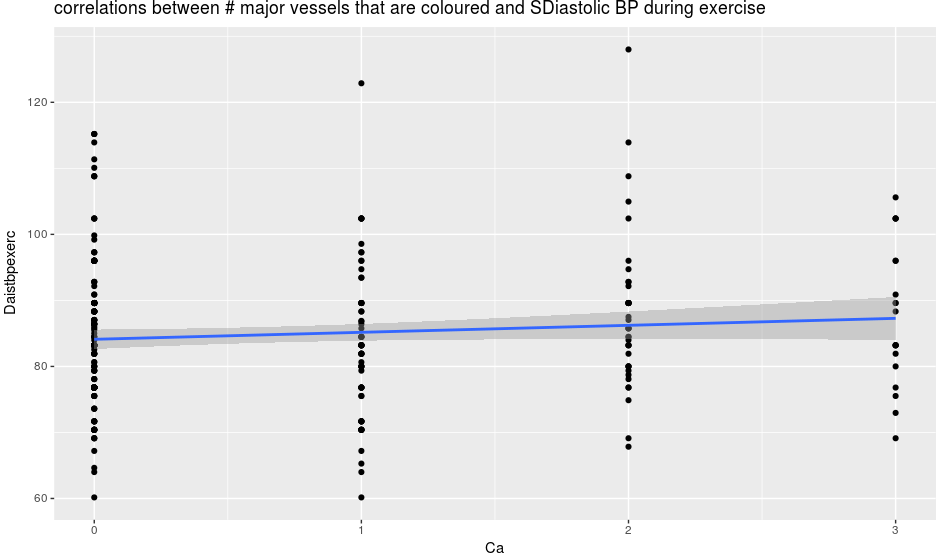
> PairedScatterPlot(cardio$thalach,cardio$oldpeak, "Thalach", "Oldpeak", "correlations between Max heart rate achieved and ST depression induced by exercise relative to rest")



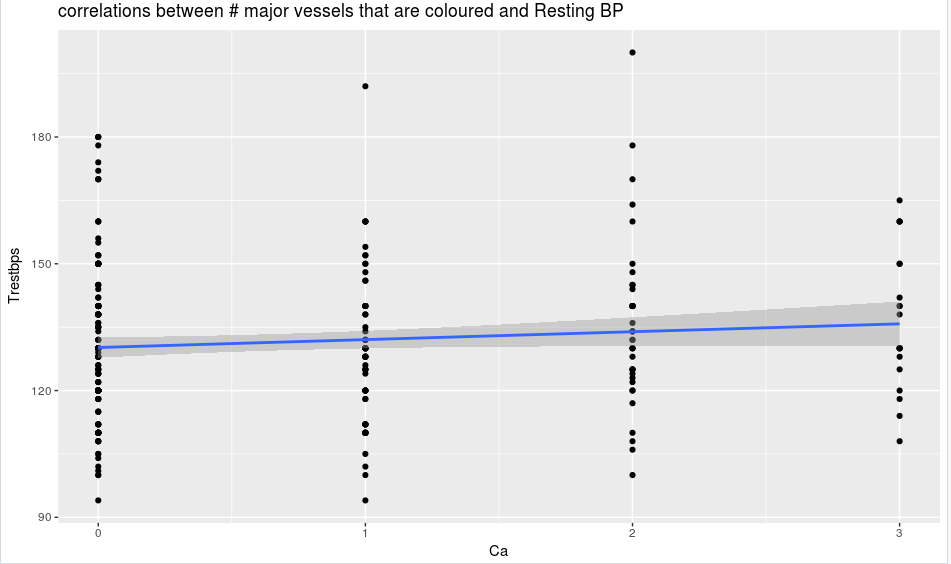
> PairedScatterPlot(cardio$ca,cardio$oldpeak, "Ca", "Oldpeak", "correlations between # major vessels that are coloured and ST depression induced by exercise relative to rest")



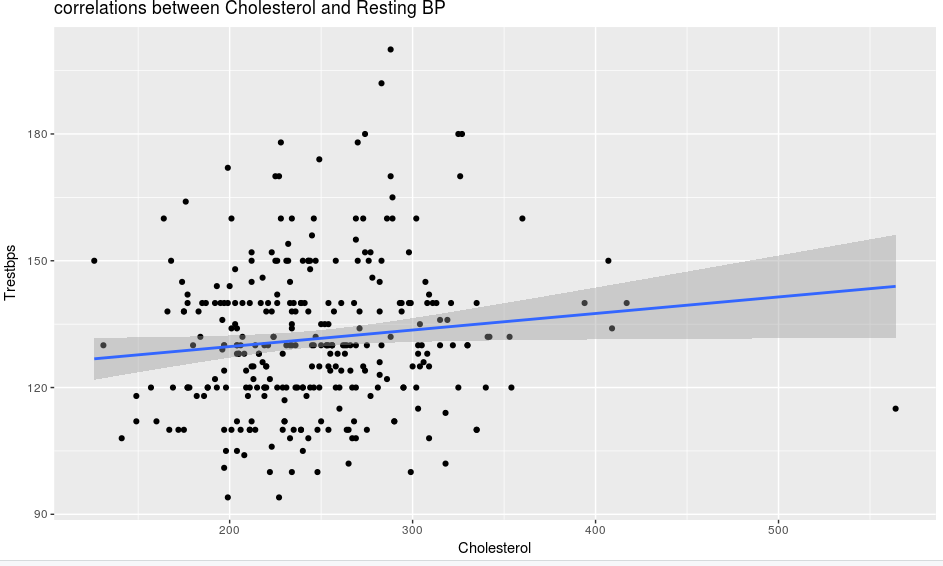
> PairedScatterPlot(cardio$ca,cardio$diastbpexerc, "Ca", "Daistbpexerc", "correlations between # major vessels that are coloured and SDiastolic BP during exercise")



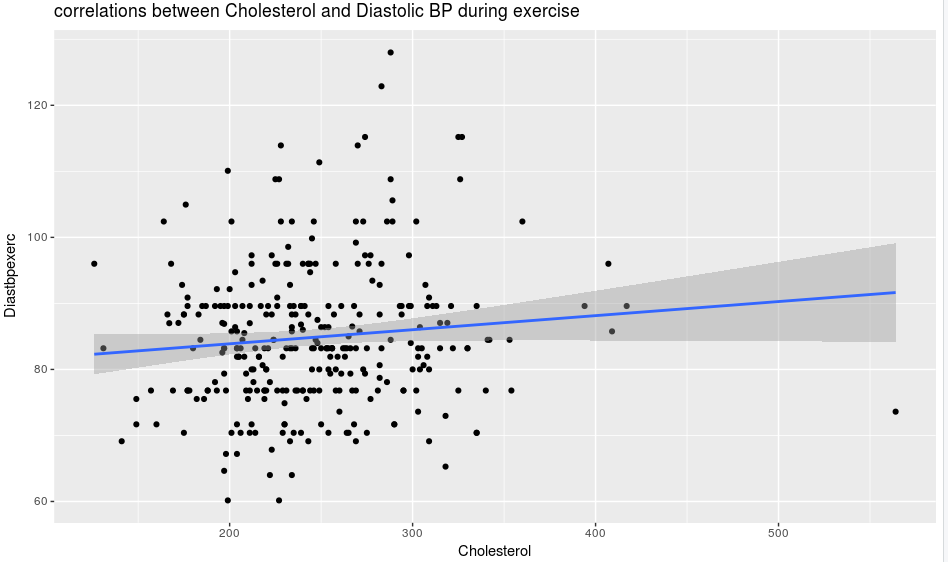
> PairedScatterPlot(cardio$ca,cardio$trestbps, "Ca", "Trestbps", "correlations between # major vessels that are coloured and Resting BP")



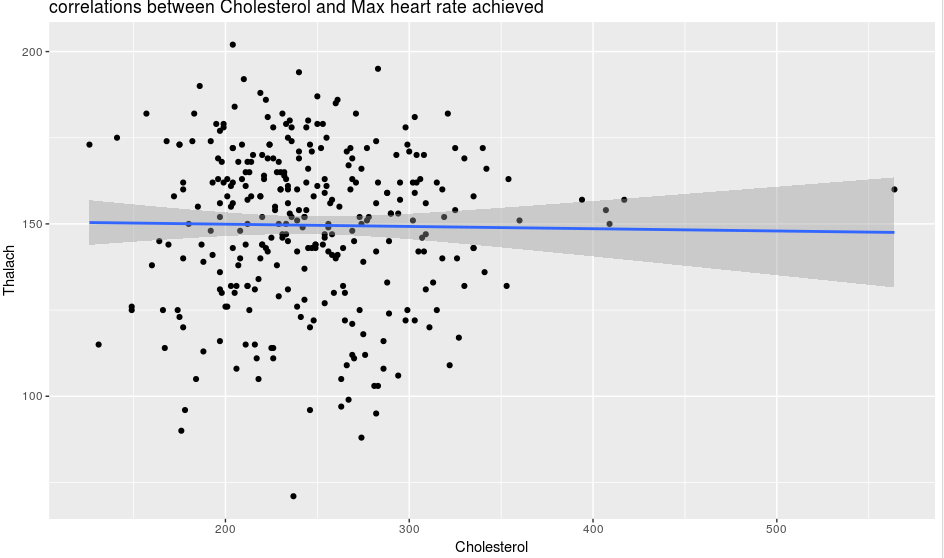
> PairedScatterPlot(cardio$cholesterol,cardio$trestbps, "Cholesterol", "Trestbps","correlations between Cholesterol and Resting BP")



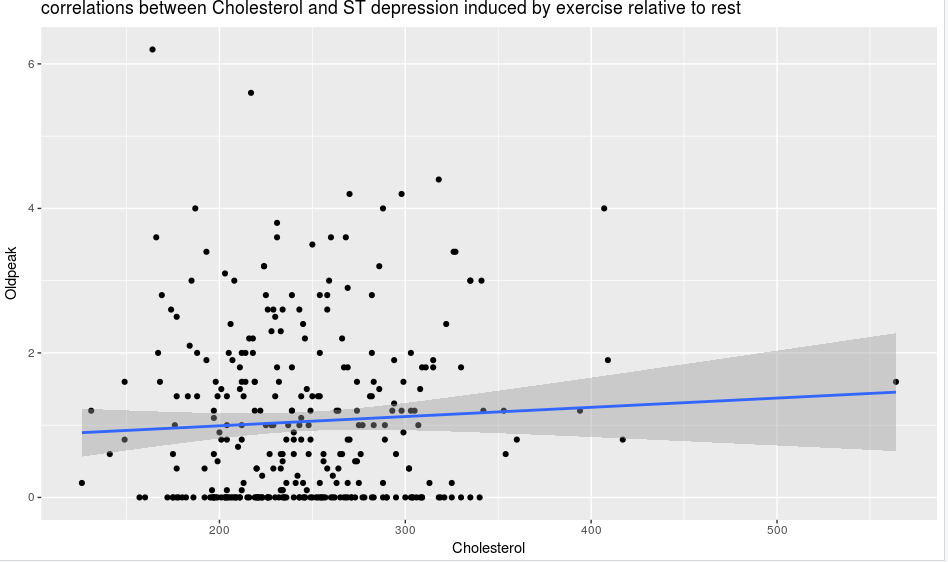
> PairedScatterPlot(cardio$cholesterol,cardio$diastbpexerc, "Cholesterol", "Diastbpexerc","correlations between Cholesterol and Diastolic BP during exercise")



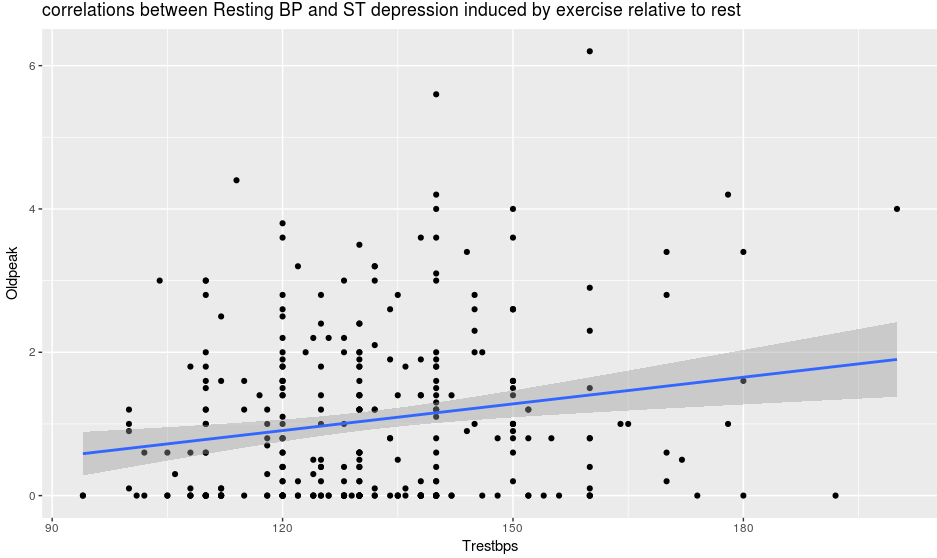
> PairedScatterPlot(cardio$cholesterol,cardio$thalach, "Cholesterol", "Thalach","correlations between Cholesterol and Max heart rate achieved")



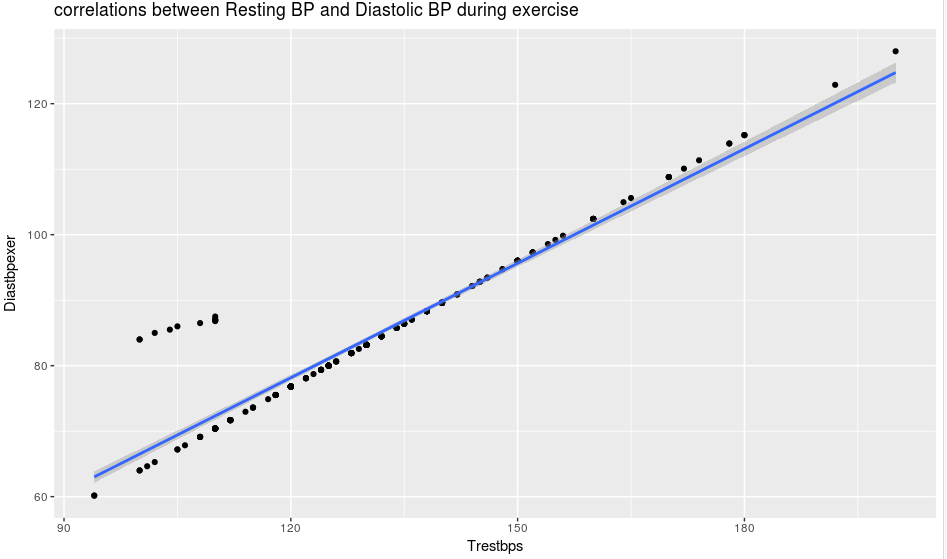
> PairedScatterPlot(cardio$cholesterol,cardio$oldpeak, "Cholesterol", "Oldpeak","correlations between Cholesterol and ST depression induced by exercise relative to rest")



> PairedScatterPlot(cardio$trestbps,cardio$oldpeak, "Trestbps", "Oldpeak","correlations between Resting BP and ST depression induced by exercise relative to rest")

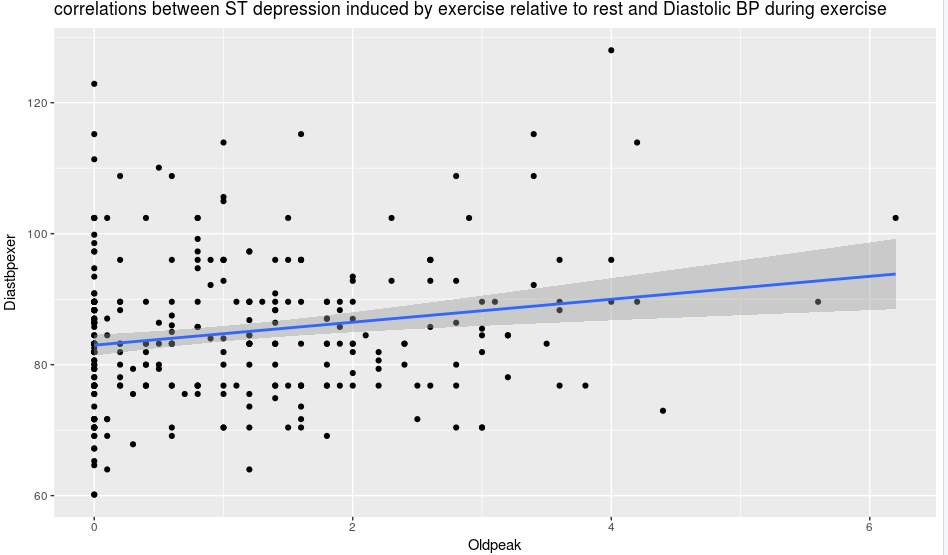


> PairedScatterPlot(cardio$trestbps,cardio$diastbpexerc, "Trestbps", "Diastbpexer","correlations between Resting BP and Diastolic BP during exercise")



very strong relationship betwwen resting BP and diastolic BP during exercise as majority of the plots are within the trendline with a few outliers. Thereby, denoting a raise in resting BP corresponds with a raise in Diastolic BP and vice versa.

> PairedScatterPlot(cardio$oldpeak,cardio$diastbpexerc, "Oldpeak", "Diastbpexer","correlations between ST depression induced by exercise relative to rest and Diastolic BP during exercise")



5b. Statistical verification of trestbps and diastbpexerc as these pair were the most correlated.

> cor(cardio$diastbpexerc, cardio$trestbps, method = 'spearman')

[1] 0.9285519

> cor(cardio$diastbpexerc, cardio$trestbps, method = 'kendall')

[1] 0.9127713

> cor(cardio$diastbpexerc, cardio$trestbps, method = 'pearson')

[1] 0.9505095

whereas correlations diastbpsexerc and thalach is low as shown below

> cor(cardio$diastbpexerc, cardio$thalach, method = 'pearson')

[1] -0.06591353

6.

> nbins <- 8

> classIntervals(cardio$trestbps, nbins, style = 'equal')

style: equal

one of 73,629,072 possible partitions of this variable into 8 classes

[94,107.25) [107.25,120.5) [120.5,133.75) [133.75,147) [147,160.25)

14 86 84 71 38

[160.25,173.5) [173.5,186.75) [186.75,200]

7 6 2

#kmeans cluster

> kmeans(cardio$trestbps,centers = nbins)$cluster

[1] 1 1 4 1 6 7 8 1 1 7 5 8 1 1 8 6 8 6 6 1 8 8 8 7 7 4 6 8 6 3 1 6 8 7 1 7 8 7 6 5

[41] 7 1 4 4 7 6 1 4 8 4 1 1 8 6 4 1 1 1 6 1 4 3 8 7 6 7 2 7 5 1 8 4 4 5 1 6 1 7 4 1

[81] 6 2 8 6 3 1 3 1 8 5 1 4 3 1 4 7 3 8 8 8 8 7 1 8 1 1 3 6 3 1 3 1 7 4 6 1 6 8 3 7

[121] 8 8 1 6 2 1 1 4 4 5 4 1 4 7 4 8 5 8 1 6 1 6 1 1 4 5 3 6 2 6 7 7 4 1 1 1 8 3 7 2

[161] 3 6 8 5 2 6 1 6 1 3 8 4 5 3 8 5 4 2 7 6 4 6 1 6 1 7 8 8 1 2 4 8 4 3 1 3 4 4 4 8

[201] 4 1 8 4 7 4 4 3 2 8 1 8 3 8 4 1 1 3 8 8 4 4 1 1 4 7 1 8 8 8 4 6 5 8 1 7 1 6 5 6

[241] 4 3 2 7 3 7 1 4 4 1 1 8 7 8 1 1 1 8 6 8 8 5 6 4 6 1 1 3 2 1 8 7 2 7 4 6 1 8 1 4

[281] 8 4 6 1 3 1 2 4 8 8 8 4 8 3 7 8 3 3 8 4 1 4 4 8 1 8 8 5

#data points per cluster

> table(kmeans(cardio$trestbps,centers = nbins)$cluster)

1 2 3 4 5 6 7 8

5 69 40 75 84 2 7 26

#summed square of distance from data point to cluster

> kmeans(cardio$trestbps,centers = nbins)$withinss

[1] 438.0000 2444.5299 136.7750 82.0000 1000.0000 61.0000 0.0000 370.9189

i would choose a lower withinss over a higher because the higher withinss indicates a need for more clusters or high rate of outliers

7. oldpeak was the most skewed with a reading of 1.229654(see 1e)

#using z\_score function created previously in 4(function(x) {scale(x, center = TRUE, scale = TRUE) }

> skewness(z\_score(cardio$oldpeak))

[1] 1.229654

> skewness(sqrt(cardio$oldpeak))

[1] 0.1501614