Cardiovascular disease detection

Tathagata Bardhan

# Importing the dataset

`cardio\_train(project)` <- read.csv("C:/Users/user/OneDrive/Desktop/project-baan bapat/cardio\_train(project).csv")  
data=`cardio\_train(project)`  
head(data)

## id age gender height weight ap\_hi ap\_lo cholesterol gluc smoke alco active  
## 1 0 18393 2 168 62 110 80 1 1 0 0 1  
## 2 1 20228 1 156 85 140 90 3 1 0 0 1  
## 3 2 18857 1 165 64 130 70 3 1 0 0 0  
## 4 3 17623 2 169 82 150 100 1 1 0 0 1  
## 5 4 17474 1 156 56 100 60 1 1 0 0 0  
## 6 8 21914 1 151 67 120 80 2 2 0 0 0  
## cardio  
## 1 0  
## 2 1  
## 3 1  
## 4 1  
## 5 0  
## 6 0

# Data cleaning

*Eliminating patient ID column*

Mydata=data[,-1]

*Null value detection*

sapply(Mydata,function(x) sum(is.na(x)))

## age gender height weight ap\_hi ap\_lo   
## 0 0 0 0 0 0   
## cholesterol gluc smoke alco active cardio   
## 0 0 0 0 0 0

**So there are no null values in the dataset**

head(Mydata)

## age gender height weight ap\_hi ap\_lo cholesterol gluc smoke alco active  
## 1 18393 2 168 62 110 80 1 1 0 0 1  
## 2 20228 1 156 85 140 90 3 1 0 0 1  
## 3 18857 1 165 64 130 70 3 1 0 0 0  
## 4 17623 2 169 82 150 100 1 1 0 0 1  
## 5 17474 1 156 56 100 60 1 1 0 0 0  
## 6 21914 1 151 67 120 80 2 2 0 0 0  
## cardio  
## 1 0  
## 2 1  
## 3 1  
## 4 1  
## 5 0  
## 6 0

summary(Mydata)

## age gender height weight   
## Min. :10798 Min. :1.00 Min. : 55.0 Min. : 10.00   
## 1st Qu.:17664 1st Qu.:1.00 1st Qu.:159.0 1st Qu.: 65.00   
## Median :19703 Median :1.00 Median :165.0 Median : 72.00   
## Mean :19469 Mean :1.35 Mean :164.4 Mean : 74.21   
## 3rd Qu.:21327 3rd Qu.:2.00 3rd Qu.:170.0 3rd Qu.: 82.00   
## Max. :23713 Max. :2.00 Max. :250.0 Max. :200.00   
## ap\_hi ap\_lo cholesterol gluc   
## Min. : -150.0 Min. : -70.00 Min. :1.000 Min. :1.000   
## 1st Qu.: 120.0 1st Qu.: 80.00 1st Qu.:1.000 1st Qu.:1.000   
## Median : 120.0 Median : 80.00 Median :1.000 Median :1.000   
## Mean : 128.8 Mean : 96.63 Mean :1.367 Mean :1.226   
## 3rd Qu.: 140.0 3rd Qu.: 90.00 3rd Qu.:2.000 3rd Qu.:1.000   
## Max. :16020.0 Max. :11000.00 Max. :3.000 Max. :3.000   
## smoke alco active cardio   
## Min. :0.00000 Min. :0.00000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:1.0000 1st Qu.:0.0000   
## Median :0.00000 Median :0.00000 Median :1.0000 Median :0.0000   
## Mean :0.08813 Mean :0.05377 Mean :0.8037 Mean :0.4997   
## 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.00000 Max. :1.00000 Max. :1.0000 Max. :1.0000

dim(Mydata)

## [1] 70000 12

*Identifying duplicate rows*

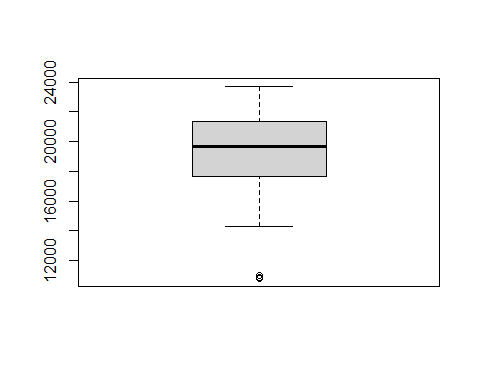
df=unique(Mydata)  
dim(df)

## [1] 69976 12

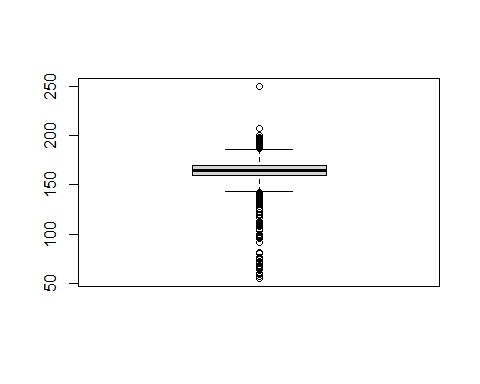
**So, there were 24 duplicate rows in the dataset**

*Outliers detection*

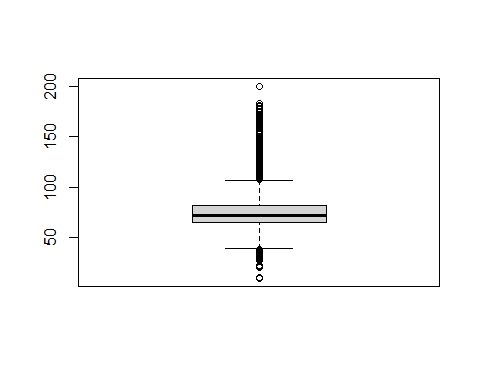
a=boxplot(df$age)$out



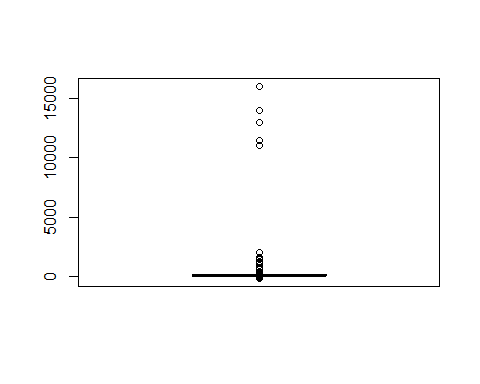
b=boxplot(df$height)$out



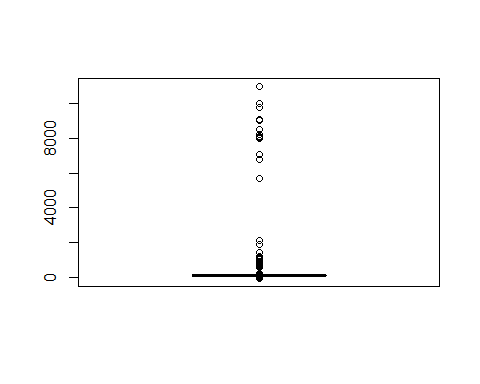
c=boxplot(df$weight)$out



d=boxplot(df$ap\_hi)$out



e=boxplot(df$ap\_lo)$out

 **So, there are 4 outliers in the variable ‘Age’, 519 outliers in variable ‘Height’, 1819 outliers in variable ‘Weight’, 1435 and 4632 outliers in variables ‘ap\_hi’ and ‘ap\_lo’ respectively.** **These are to be treated.**

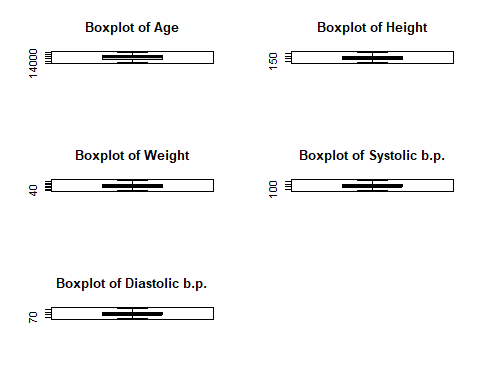
x=c()  
for(i in 1:length(a)){  
 x=c(x,which(df$age==a[i]))  
}  
y=c()  
for(i in 1:length(b)){  
 y=c(y,which(df$height==b[i]))  
}  
z=c()  
for(i in 1:length(c)){  
 z=c(z,which(df$weight==c[i]))  
}  
p=c()  
for(i in 1:length(d)){  
 p=c(p,which(df$ap\_hi==d[i]))  
}  
q=c()  
for(i in 1:length(e)){  
 q=c(q,which(df$ap\_lo==e[i]))  
}

data\_cardio=df  
data\_cardio$age[x]=min(data\_cardio$age[-x])  
  
data\_cardio$height[data\_cardio$height>(quantile(data\_cardio$height,0.75)+(1.5\*IQR(data\_cardio$height)))]=max(data\_cardio$height[-y])  
data\_cardio$height[data\_cardio$height<(quantile(data\_cardio$height,0.25)-(1.5\*IQR(data\_cardio$height)))]=min(data\_cardio$height[-y])  
  
data\_cardio$weight[data\_cardio$weight>(quantile(data\_cardio$weight,0.75)+(1.5\*IQR(data\_cardio$weight)))]=max(data\_cardio$weight[-z])  
data\_cardio$weight[data\_cardio$weight<(quantile(data\_cardio$weight,0.25)-(1.5\*IQR(data\_cardio$weight)))]=min(data\_cardio$weight[-z])  
  
data\_cardio$ap\_hi[data\_cardio$ap\_hi>(quantile(data\_cardio$ap\_hi,0.75)+(1.5\*IQR(data\_cardio$ap\_hi)))]=max(data\_cardio$ap\_hi[-p])  
data\_cardio$ap\_hi[data\_cardio$ap\_hi<(quantile(data\_cardio$ap\_hi,0.25)-(1.5\*IQR(data\_cardio$ap\_hi)))]=min(data\_cardio$ap\_hi[-p])  
  
data\_cardio$ap\_lo[data\_cardio$ap\_lo>(quantile(data\_cardio$ap\_lo,0.75)+(1.5\*IQR(data\_cardio$ap\_lo)))]=max(data\_cardio$ap\_lo[-q])  
data\_cardio$ap\_lo[data\_cardio$ap\_lo<(quantile(data\_cardio$ap\_lo,0.25)-(1.5\*IQR(data\_cardio$ap\_lo)))]=min(data\_cardio$ap\_lo[-q])

*For removal of outliers, for each variable, the values greater than Q3+1.5IQR are replaced by the maximum of the values other than the outliers and the values less than Q1-1.5IQR are replaced by the minimum of the values other than the outliers.*

**Verification**

par(mfrow=c(3,2))  
boxplot(data\_cardio$age,main="Boxplot of Age")  
boxplot(data\_cardio$height,main="Boxplot of Height")  
boxplot(data\_cardio$weight,main="Boxplot of Weight")  
boxplot(data\_cardio$ap\_hi,main="Boxplot of Systolic b.p.")  
boxplot(data\_cardio$ap\_lo,main="Boxplot of Diastolic b.p.")

 **We can ensure that there are no more outliers in the dataset**

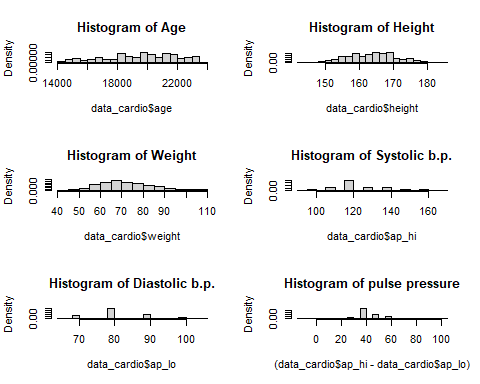
dim(data\_cardio)

## [1] 69976 12

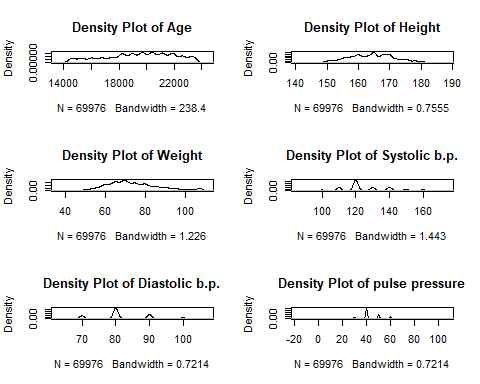
**Thus we have our cleaned dataset ‘data\_cardio’ which is free from missing values, duplicate values & outliers**

# Univariate Analysis

par(mfrow=c(3,2))  
hist(data\_cardio$age,main="Histogram of Age",prob=TRUE)  
hist(data\_cardio$height,main="Histogram of Height",prob=TRUE)  
hist(data\_cardio$weight,main="Histogram of Weight",prob=TRUE)  
hist(data\_cardio$ap\_hi,main="Histogram of Systolic b.p.",prob=TRUE)  
hist(data\_cardio$ap\_lo,main="Histogram of Diastolic b.p.",prob=TRUE)  
hist((data\_cardio$ap\_hi-data\_cardio$ap\_lo),main="Histogram of pulse pressure",prob=TRUE)

 **Height and weight are more or less symmetrically distributed and age is slightly negatively skewed**

par(mfrow=c(3,2))  
plot(density(data\_cardio$age),main="Density Plot of Age")  
plot(density(data\_cardio$height),main="Density Plot of Height")  
plot(density(data\_cardio$weight),main="Density Plot of Weight")  
plot(density(data\_cardio$ap\_hi),main="Density Plot of Systolic b.p.")  
plot(density(data\_cardio$ap\_lo),main="Density Plot of Diastolic b.p.")  
plot(density(data\_cardio$ap\_hi-data\_cardio$ap\_lo),main="Density Plot of pulse pressure")

 **Most of the patients are in the age of 18000 to 22000 days, are of height 160 to 170 cm, weight 60 to 80 kg & most of the patients have systolic bp 120 and diastolic bp 80 i.e. b.p. is in normal range**

# Bivariate Analysis

summary(data\_cardio)

## age gender height weight   
## Min. :14275 Min. :1.00 Min. :143.0 Min. : 40.00   
## 1st Qu.:17664 1st Qu.:1.00 1st Qu.:159.0 1st Qu.: 65.00   
## Median :19703 Median :1.00 Median :165.0 Median : 72.00   
## Mean :19469 Mean :1.35 Mean :164.4 Mean : 73.92   
## 3rd Qu.:21327 3rd Qu.:2.00 3rd Qu.:170.0 3rd Qu.: 82.00   
## Max. :23713 Max. :2.00 Max. :186.0 Max. :107.00   
## ap\_hi ap\_lo cholesterol gluc   
## Min. : 90.0 Min. : 65.00 Min. :1.000 Min. :1.000   
## 1st Qu.:120.0 1st Qu.: 80.00 1st Qu.:1.000 1st Qu.:1.000   
## Median :120.0 Median : 80.00 Median :1.000 Median :1.000   
## Mean :126.7 Mean : 81.78 Mean :1.367 Mean :1.227   
## 3rd Qu.:140.0 3rd Qu.: 90.00 3rd Qu.:2.000 3rd Qu.:1.000   
## Max. :170.0 Max. :105.00 Max. :3.000 Max. :3.000   
## smoke alco active cardio   
## Min. :0.00000 Min. :0.00000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:1.0000 1st Qu.:0.0000   
## Median :0.00000 Median :0.00000 Median :1.0000 Median :0.0000   
## Mean :0.08816 Mean :0.05379 Mean :0.8037 Mean :0.4998   
## 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.00000 Max. :1.00000 Max. :1.0000 Max. :1.0000

round((cor(data\_cardio[,-c(12)])),4)

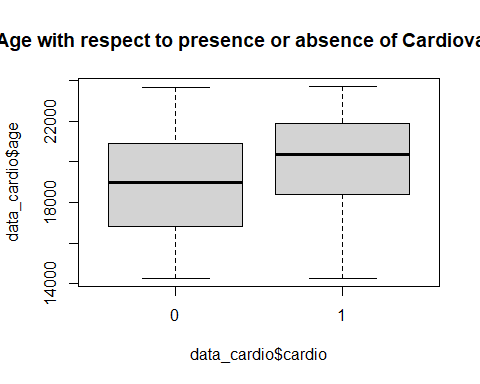
## age gender height weight ap\_hi ap\_lo cholesterol gluc  
## age 1.0000 -0.0228 -0.0866 0.0597 0.2093 0.1557 0.1544 0.0987  
## gender -0.0228 1.0000 0.5223 0.1630 0.0617 0.0670 -0.0359 -0.0206  
## height -0.0866 0.5223 1.0000 0.3169 0.0166 0.0305 -0.0547 -0.0207  
## weight 0.0597 0.1630 0.3169 1.0000 0.2741 0.2589 0.1438 0.1056  
## ap\_hi 0.2093 0.0617 0.0166 0.2741 1.0000 0.7257 0.1973 0.0926  
## ap\_lo 0.1557 0.0670 0.0305 0.2589 0.7257 1.0000 0.1660 0.0757  
## cholesterol 0.1544 -0.0359 -0.0547 0.1438 0.1973 0.1660 1.0000 0.4515  
## gluc 0.0987 -0.0206 -0.0207 0.1056 0.0926 0.0757 0.4515 1.0000  
## smoke -0.0477 0.3381 0.1963 0.0687 0.0272 0.0270 0.0103 -0.0048  
## alco -0.0298 0.1710 0.0980 0.0682 0.0324 0.0407 0.0357 0.0112  
## active -0.0099 0.0059 -0.0082 -0.0174 0.0021 0.0016 0.0099 -0.0068  
## smoke alco active  
## age -0.0477 -0.0298 -0.0099  
## gender 0.3381 0.1710 0.0059  
## height 0.1963 0.0980 -0.0082  
## weight 0.0687 0.0682 -0.0174  
## ap\_hi 0.0272 0.0324 0.0021  
## ap\_lo 0.0270 0.0407 0.0016  
## cholesterol 0.0103 0.0357 0.0099  
## gluc -0.0048 0.0112 -0.0068  
## smoke 1.0000 0.3401 0.0259  
## alco 0.3401 1.0000 0.0255  
## active 0.0259 0.0255 1.0000

cor.test(data\_cardio$ap\_hi,data\_cardio$ap\_lo)$p.value

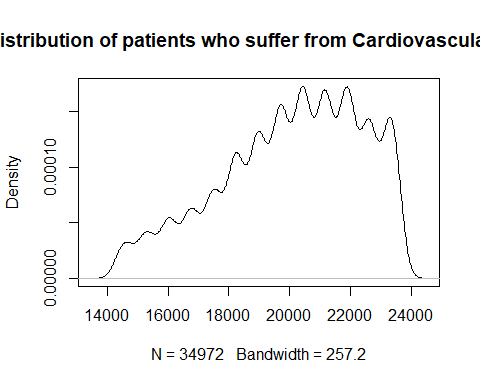
## [1] 0

**So there is significant correlation between Systolic and Diastolic b.p.**

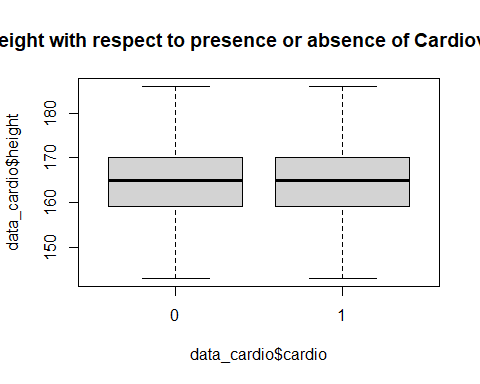
par(mfrow=c(1,1))  
boxplot(data\_cardio$age~data\_cardio$cardio,main="Boxplot of Age with respect to presence or absence of Cardiovascular disease")



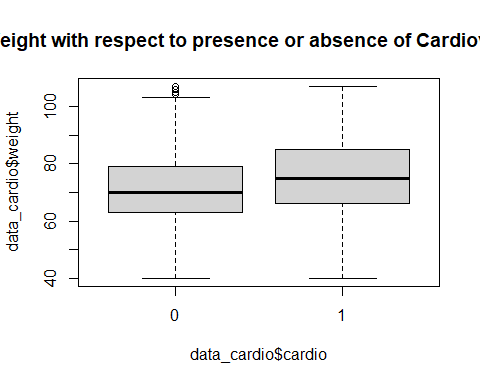
plot(density(data\_cardio$age[data\_cardio$cardio==1]),main="Age distribution of patients who suffer from Cardiovascular disease")

 **Thus on an average, people older in age suffer more heart disease than those who are younger. More precisely patients having age between 20,000-22,000 days i.e. between 55 to 60 years, suffer mostly with cardiovascular disease**

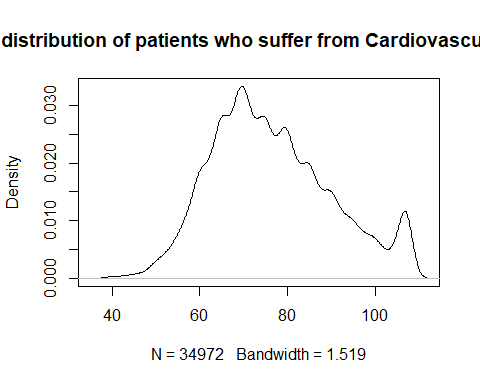
par(mfrow=c(1,1))  
boxplot(data\_cardio$height~data\_cardio$cardio,main="Boxplot of Height with respect to presence or absence of Cardiovascular disease")

 **Thus on an average, people who don’t suffer from Cardiovascular disease have same height with those who suffer from the disease. Thus height of a person doesn’t play major role in detection of the disease**

par(mfrow=c(1,1))  
boxplot(data\_cardio$weight~data\_cardio$cardio,main="Boxplot of Weight with respect to presence or absence of Cardiovascular disease")



plot(density(data\_cardio$weight[data\_cardio$cardio==1]),main="Weight distribution of patients who suffer from Cardiovascular disease")

 **Thus, on an average, people who have suffered from Cardiovascular disease have more weight than those who didn’t, infact it is mostly occuring for patients whose weights are within 65-75 kgs**