Homework_2 :Working With Healthcare Data

Pavithra Senthilkumar

2024-02-08

DATA LOADING AND CLEANING

Q_1.1 Getting the working Directory:

```
getwd()
```

```
## [1] "C:/Users/Pavithra/Documents/Notebooks/RScripts"
```

Setting up Working directory:

```
setwd("C:/Users/Pavithra/Documents/Notebooks/RScripts")
```

Reading the Dataset:

```
data <-read.csv("synthetic_health_data.csv",sep=",")
```

Q_1.2 Printing data

```
head(data)
```

```
##   PatientID Age    Sex BloodPressure Cholesterol Diabetes SmokingStatus
## 1         1  48 Female            87         215       No Former Smoker
## 2         2  68 Female           132         215       No Former Smoker
## 3         3  31 Female           101         151       No        Smoker
## 4         4  84   Male           121         220       No Former Smoker
## 5         5  59   Male           102         234       No    Non-Smoker
## 6         6  67   Male           122         236       No    Non-Smoker
##   HeartDisease
## 1          Yes
## 2          Yes
## 3           No
## 4           No
## 5           No
## 6           No
```

Dimensions of the Data:

```
dim(data)
```

```
## [1] 1000    8
```

The dataset has 1000 entries/rows and has 8 features/columns.

```
summary(data)
```

```
##    PatientID           Age            Sex            BloodPressure
##  Min.   :   1.0   Min.   :18.00   Length:1000        Min.   : 63.0
##  1st Qu.: 250.8   1st Qu.:35.75   Class :character   1st Qu.:109.0
##  Median : 500.5   Median :53.00   Mode  :character   Median :120.0
##  Mean   : 500.5   Mean   :53.68                      Mean   :119.7
##  3rd Qu.: 750.2   3rd Qu.:71.00                      3rd Qu.:129.0
##  Max.   :1000.0   Max.   :90.00                      Max.   :178.0
##   Cholesterol      Diabetes         SmokingStatus      HeartDisease
##  Min.   :112.0   Length:1000        Length:1000        Length:1000
##  1st Qu.:182.0   Class :character   Class :character   Class :character
##  Median :200.0   Mode  :character   Mode  :character   Mode  :character
##  Mean   :200.5
##  3rd Qu.:220.0
##  Max.   :283.0
```

It is noted that there are 4 numeric variables and 4 character variables in the dataset.

The average age of the population is 53.68, with older people belonging to 90 years, and the younger people belonging to 18 years of age.

```
str(data)
```

```
## 'data.frame':    1000 obs. of  8 variables:
##  $ PatientID    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Age          : int  48 68 31 84 59 67 60 31 42 86 ...
##  $ Sex          : chr  "Female" "Female" "Female" "Male" ...
##  $ BloodPressure: int  87 132 101 121 102 122 118 141 119 112 ...
##  $ Cholesterol  : int  215 215 151 220 234 236 204 206 198 204 ...
##  $ Diabetes     : chr  "No" "No" "No" "No" ...
##  $ SmokingStatus: chr  "Former Smoker" "Former Smoker" "Smoker" "Former Smoker" ...
##  $ HeartDisease : chr  "Yes" "Yes" "No" "No" ...
```

Integer Variables in Data : PatientID,Age,BloodPressure,Cholesterol

Character Variables in Data: Sex, Diabetes, SmokingStatus,HeartDisease

Q_1.3 Identifying Missing Values

This approach loops over each column and the function is.na() returns the missing values count.

```
for(i in 1:length(data))
{
  print(which(is.na(data[i])))
}
```

```
## integer(0)
## integer(0)
## integer(0)
## integer(0)
## integer(0)
## integer(0)
## integer(0)
## integer(0)
```

There are no missing values in the dataset.

Q_1.4 Converting Categorical to Factor Variables:

The character variables that are categorical are converted to factors using the lapply & sapply functions.

```
cols<-c("Sex","Diabetes","SmokingStatus","HeartDisease")

data[cols]<-lapply(data[cols],factor)

sapply(data,class)
```

```
##     PatientID           Age           Sex BloodPressure   Cholesterol
##     "integer"     "integer"      "factor"     "integer"     "integer"
##      Diabetes SmokingStatus  HeartDisease
##      "factor"      "factor"      "factor"
```

Encoding Variables to numeric:

Since the categorical variables are in character categories, encoding them to numeric would be easier for further numerical analysis.

Encoding Sex by creating a new column Gender

Having Females as 1 and males as 0.

```
data$Gender<- data$Sex
data$Gender<-as.character(data$Gender)

data$Gender[data$Gender=="Female"] <- 1
data$Gender[data$Gender=="Male"] <- 0

data$Gender<-as.numeric(data$Gender)
```

Encoding Diabetes

Absence of diabetes is marked as 0, presence as 1.

```
data$Diabetes<-as.character(data$Diabetes)
data$Diabetes[data$Diabetes=="No"] <-0
data$Diabetes[data$Diabetes=="Yes"] <-1

data$Diabetes<-as.numeric(data$Diabetes)
```

Encoding Heart Disease

Absence of diabetes is marked as 0, presence as 1.

```
data$HeartDisease<-as.character(data$HeartDisease)
data$HeartDisease[data$HeartDisease=="No"] <-0
data$HeartDisease[data$HeartDisease=="Yes"] <-1

data$HeartDisease<-as.numeric(data$HeartDisease)

head(data)
```

```
##   PatientID Age    Sex BloodPressure Cholesterol Diabetes SmokingStatus
## 1         1  48 Female            87         215        0 Former Smoker
## 2         2  68 Female           132         215        0 Former Smoker
## 3         3  31 Female           101         151        0        Smoker
## 4         4  84   Male           121         220        0 Former Smoker
## 5         5  59   Male           102         234        0    Non-Smoker
## 6         6  67   Male           122         236        0    Non-Smoker
##   HeartDisease Gender
## 1            1      1
## 2            1      1
## 3            0      1
## 4            0      0
## 5            0      0
## 6            0      0
```

Checking Distinct values in SmokingStatus Column:

```
sort(unique(data$SmokingStatus),decreasing=TRUE)
```

```
## [1] Smoker        Non-Smoker    Former Smoker
## Levels: Former Smoker Non-Smoker Smoker
```

Encoding Smoking Status:

Creating different columns for each smoking category.

```r
library(dplyr)
```

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

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

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

```r
data<- data %>%

  mutate(Smoker = ifelse(SmokingStatus == "Smoker", 1, 0),
         Non_Smoker = ifelse(SmokingStatus=="Non-Smoker",1,0),
         Former_Smoker = ifelse(SmokingStatus=="Former Smoker",1,0))
```

Converting Columns to integer types:

```r
data$Gender <-as.integer(data$Gender)
data$Diabetes <-as.integer(data$Diabetes)
data$HeartDisease <-as.integer(data$HeartDisease)
data$Smoker<-as.integer(data$Smoker)
data$Non_Smoker<-as.integer(data$Non_Smoker)
data$Former_Smoker<-as.integer(data$Former_Smoker)

str(data)
```

```
## 'data.frame':    1000 obs. of  12 variables:
##  $ PatientID    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Age          : int  48 68 31 84 59 67 60 31 42 86 ...
##  $ Sex          : Factor w/ 2 levels "Female","Male": 1 1 1 2 2 2 1 2 2 2 ...
##  $ BloodPressure: int  87 132 101 121 102 122 118 141 119 112 ...
##  $ Cholesterol  : int  215 215 151 220 234 236 204 206 198 204 ...
##  $ Diabetes     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ SmokingStatus: Factor w/ 3 levels "Former Smoker",..: 1 1 3 1 2 2 1 2 3 3 ...
##  $ HeartDisease : int  1 1 0 0 0 0 0 0 0 0 ...
##  $ Gender       : int  1 1 1 0 0 0 1 0 0 0 ...
##  $ Smoker       : int  0 0 1 0 0 0 0 0 1 1 ...
##  $ Non_Smoker   : int  0 0 0 0 1 1 0 1 0 0 ...
##  $ Former_Smoker: int  1 1 0 1 0 0 1 0 0 0 ...
```

```r
head(data)
```

```
##   PatientID Age    Sex BloodPressure Cholesterol Diabetes SmokingStatus
## 1         1  48 Female            87         215        0 Former Smoker
## 2         2  68 Female           132         215        0 Former Smoker
```

```
## 3         3  31 Female           101         151        0        Smoker
## 4         4  84   Male           121         220        0 Former Smoker
## 5         5  59   Male           102         234        0    Non-Smoker
## 6         6  67   Male           122         236        0    Non-Smoker
##   HeartDisease Gender Smoker Non_Smoker Former_Smoker
## 1            1      1      0          0             1
## 2            1      1      0          0             1
## 3            0      1      1          0             0
## 4            0      0      0          0             1
## 5            0      0      0          1             0
## 6            0      0      0          1             0
```

Selecting/Creating only numeric columns data:

```
data_numeric<-data[,c("Age","Gender","Diabetes",
  "HeartDisease","Smoker","Non_Smoker","Former_Smoker",
  "Cholesterol","BloodPressure")]
```

Data Exploration and Transformation

Q_2.1 Creating a new HighBP Column:

Creating a new variable named HighBP that has an indication of high bloodpressure level.

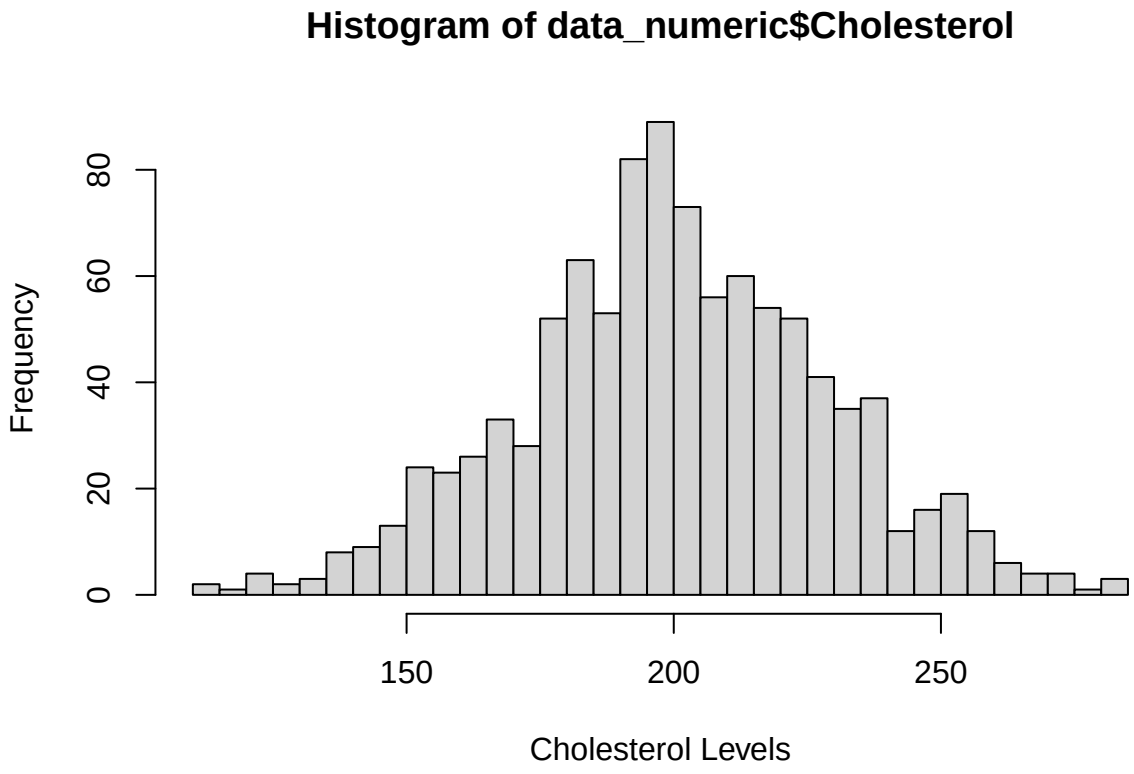
```
library(dplyr)

upper_quantile_Bp<-quantile(data$BloodPressure,0.75)

data_numeric<- data_numeric %>%

  mutate(HighBP = ifelse(BloodPressure>=upper_quantile_Bp, 1, 0))

head(data_numeric)
```

```
##   Age Gender Diabetes HeartDisease Smoker Non_Smoker Former_Smoker Cholesterol
## 1  48      1        0            1      0          0             1         215
## 2  68      1        0            1      0          0             1         215
## 3  31      1        0            0      1          0             0         151
## 4  84      0        0            0      0          0             1         220
## 5  59      0        0            0      0          1             0         234
## 6  67      0        0            0      0          1             0         236
##   BloodPressure HighBP
## 1            87      0
## 2           132      1
## 3           101      0
## 4           121      0
## 5           102      0
## 6           122      0
```

The threshold chosen to indicate a High Blood Pressure is the upper Quartile (Q3), which represents values that are 75th percentile and more. For which, the values are above the median and mean value of Bloodpressure.

Q_2.2 Histogram for Cholesterol,

```
hist(data_numeric$Cholesterol, xlab="Cholesterol Levels",breaks=30)
```

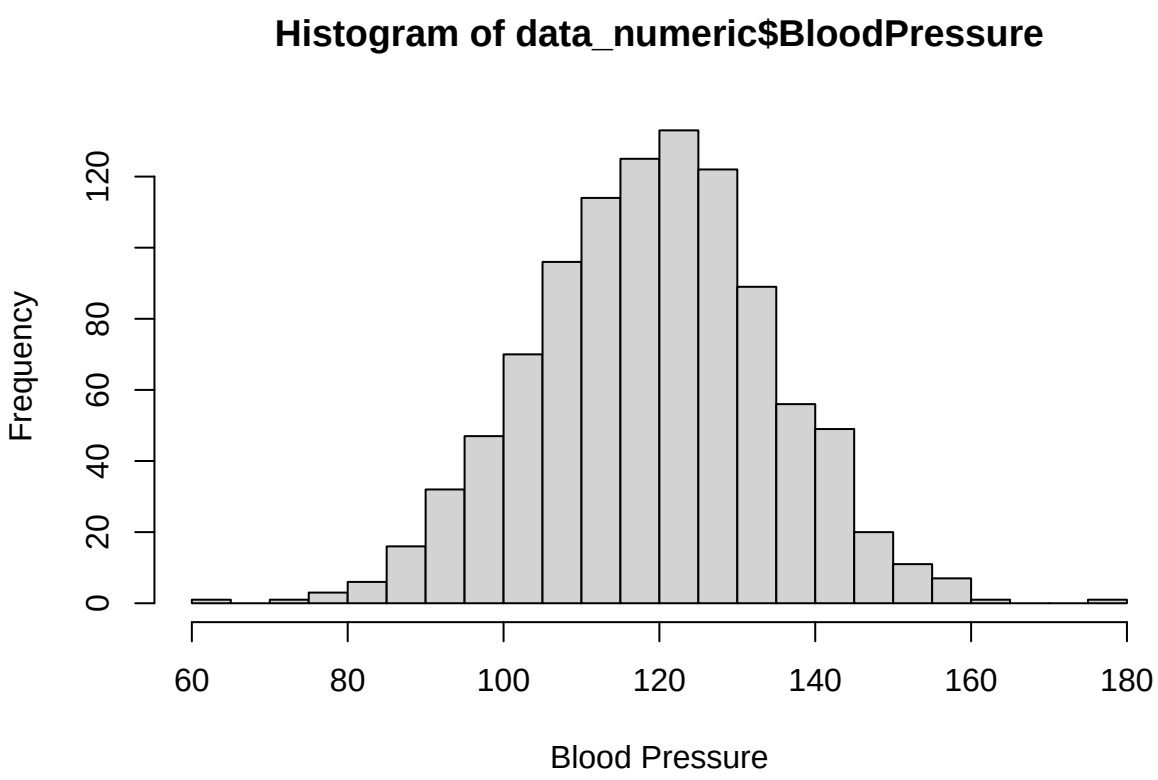


The distribution of cholestrol levels follows a Normal Distribution. The median value lies around 200.

The bin size or breaks was chosen as 30, since it could accommodate more entries together and gives a big picture of whole population.

Histogram for Blood Pressure,

```
hist(data_numeric$BloodPressure, xlab="Blood Pressure",breaks=20)
```



The distribution of Blood Pressure also follows a Normal Distribution. The median value lies around 120.

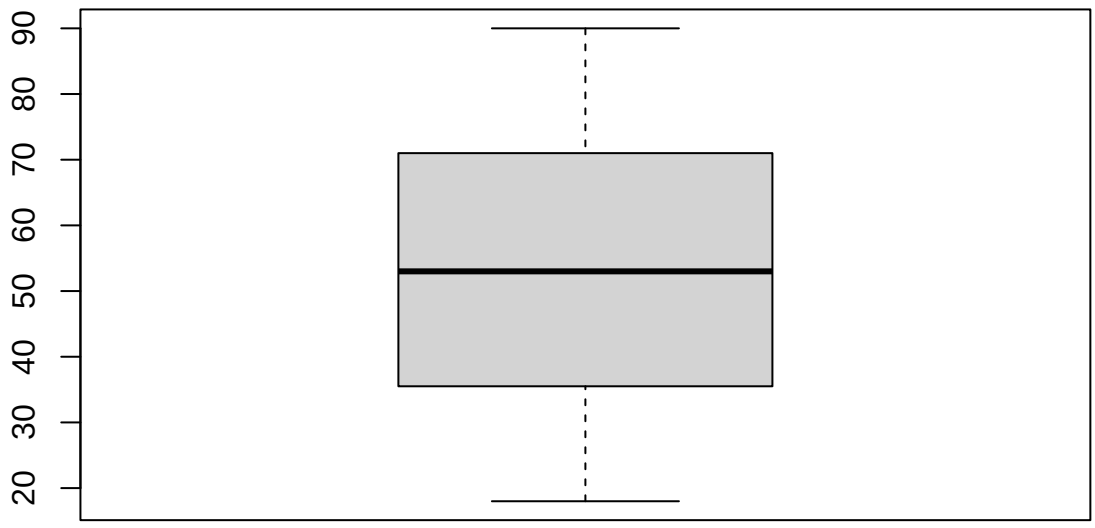
The bin size or breaks was chosen as 20, since it could accommodate more entries together and gives a big picture of whole population.

By having the bin size as 20, the histogram is not limited to the amount of individual points it could represent.

This bin size gives a better representation of the population's Blood Pressure.

Examining with Boxplots,

```
boxplot(data_numeric$Age, xlab="Age")
```

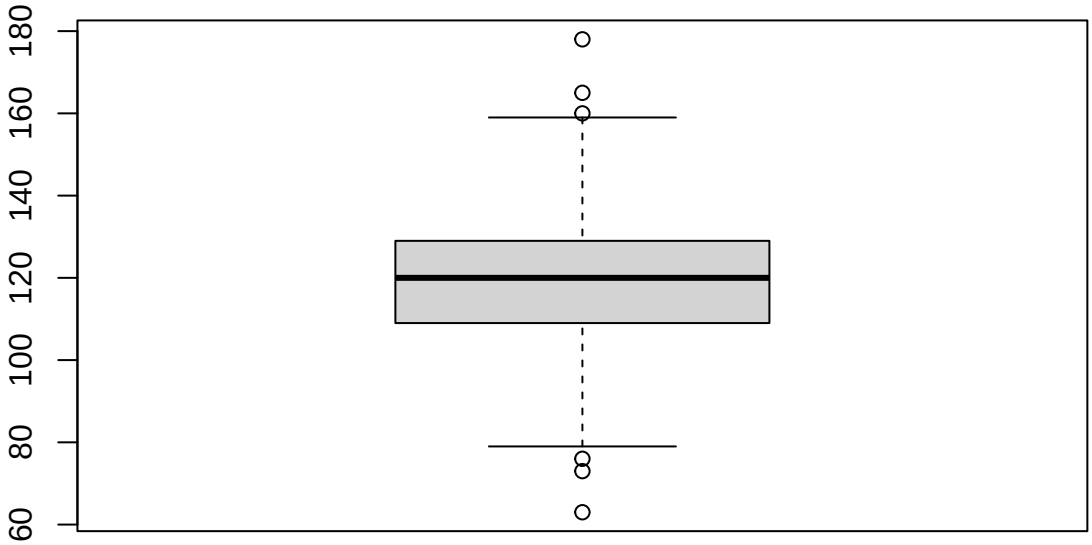


The median age or the 50th percentile happens to be around 53, with lower quartile around 35 years and upper quartile around 71.

There are no outliers in this column.

Boxplot for BloodPressure

```
boxplot(data_numeric$BloodPressure, xlab="BloodPressure")
```



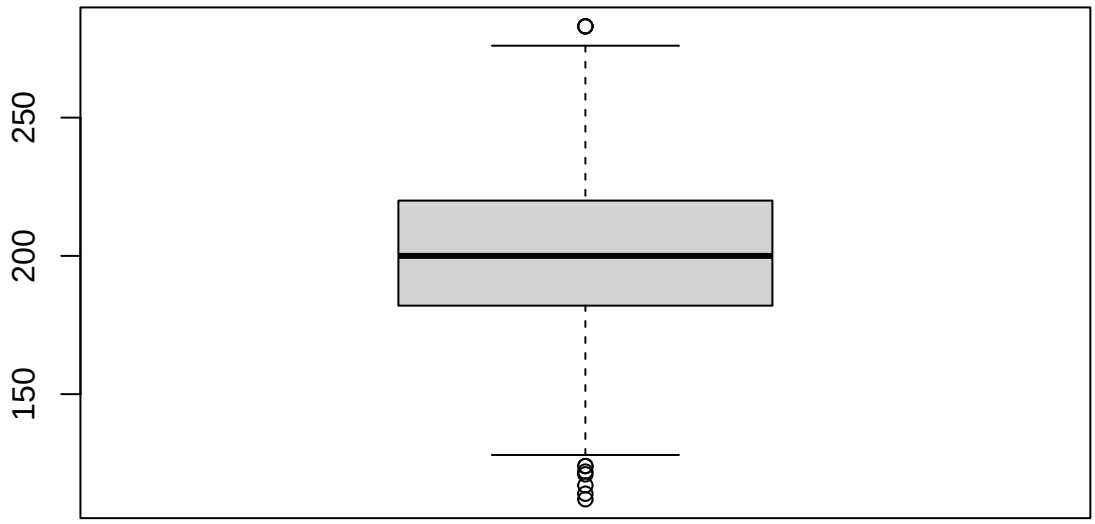
The blood pressure has a narrow IQR, with lower quartile of 109 and upper quartile of 129.

The data has outliers/extreme values in both beyond the upper and lower regions of the quartiles.

Extreme values with Highest BP value of : 178 Lowest BP value of : 63

Boxplot for Cholesterol

```
boxplot(data_numeric$Cholesterol, xlab="Cholesterol")
```

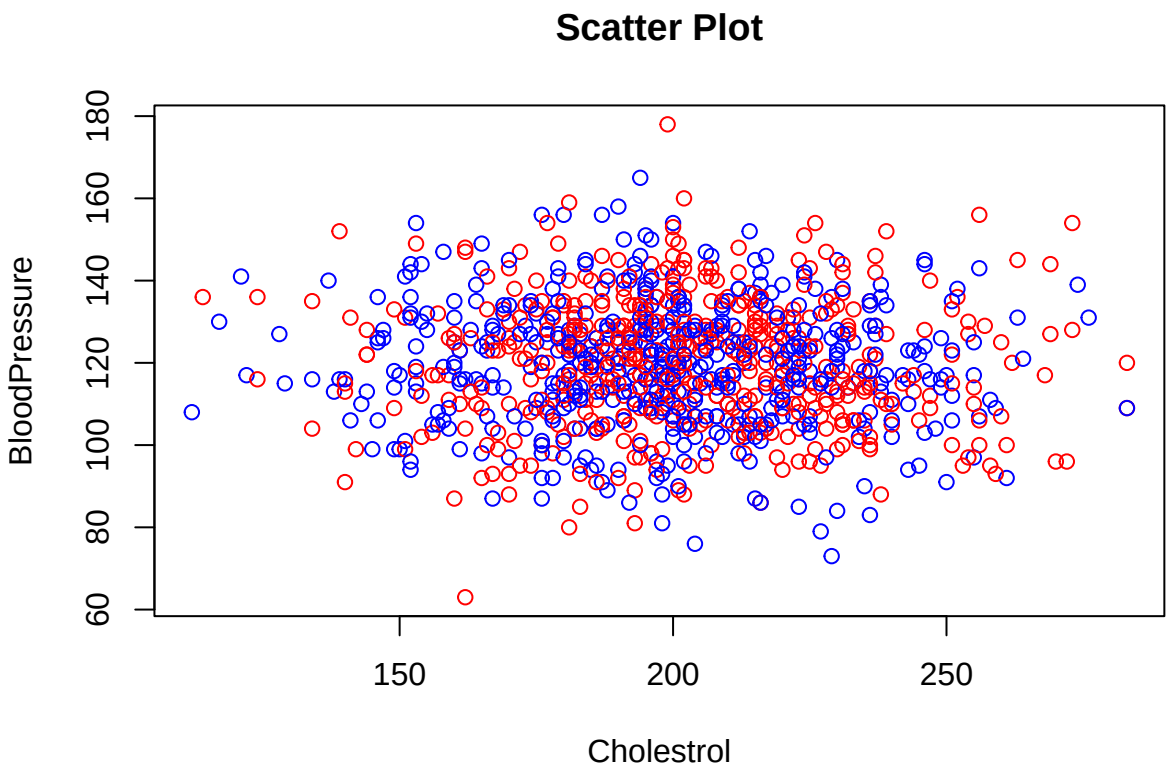


The average value of the cholestrol is nearly 200, with lower quartile of 122 and upper quartile of 220. The cholestrol variable too has extreme values too, with lowest value being 112 and highest being 283.

Q_2.3 Scatter Plots to understand relationship between variables

Cholesterol Vs BloodPressure

```
cholest_scat<-data_numeric$Cholesterol
bloodpressure_scat <-data_numeric$BloodPressure

plot(cholest_scat,bloodpressure_scat,main="Scatter Plot",xlab="Cholestrol",ylab="BloodPressure",col=c("
```



As indicated by the plot the relationship between cholesterol and blood pressure is slightly inversely correlated.

The blue dots (Cholesterol) in the left is higher and towards the right the cholesterol reduces and blood pressure increases.

```
print(cor(data_numeric$BloodPressure,data_numeric$Cholesterol))
```

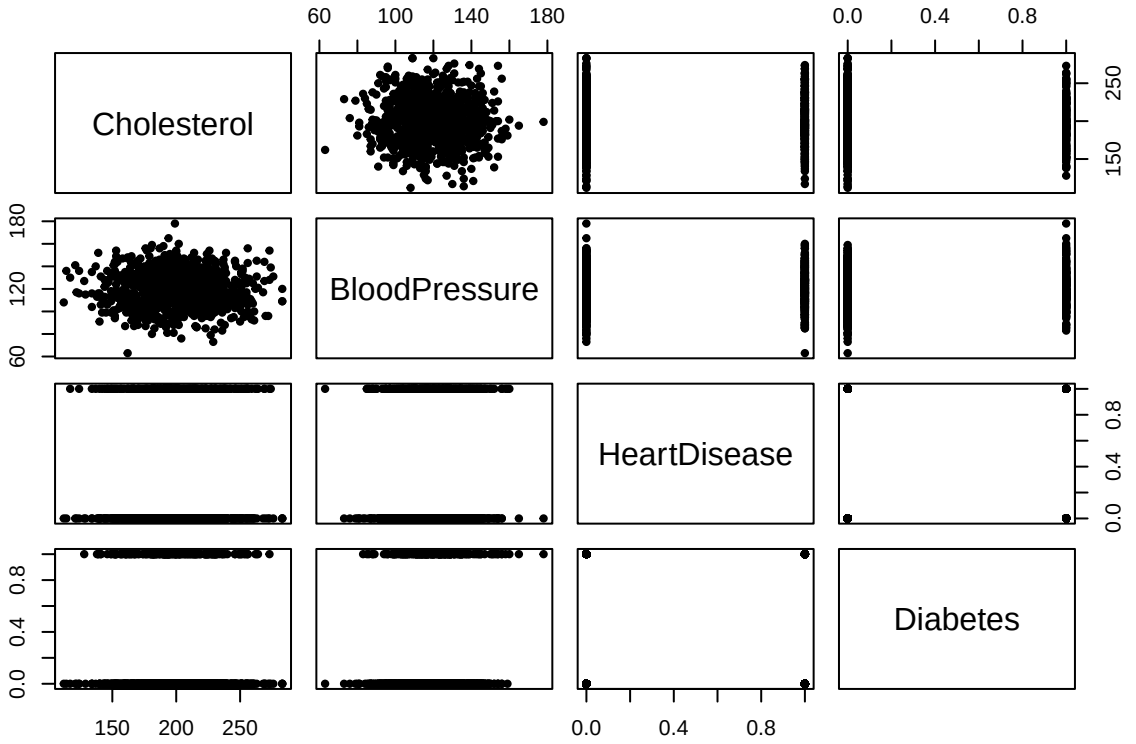
```
## [1] -0.03881453
```

Also the correlation coefficient, it is assumed that these two variables have a negative correlation of -0.03.

Since the cholesterol and bloodpressure should be positively correlated in reality, which could be due to the outliers

Pairs Plot

```
pairs(data_numeric[,c("Cholesterol","BloodPressure","HeartDisease","Diabetes")],pch=20)
```

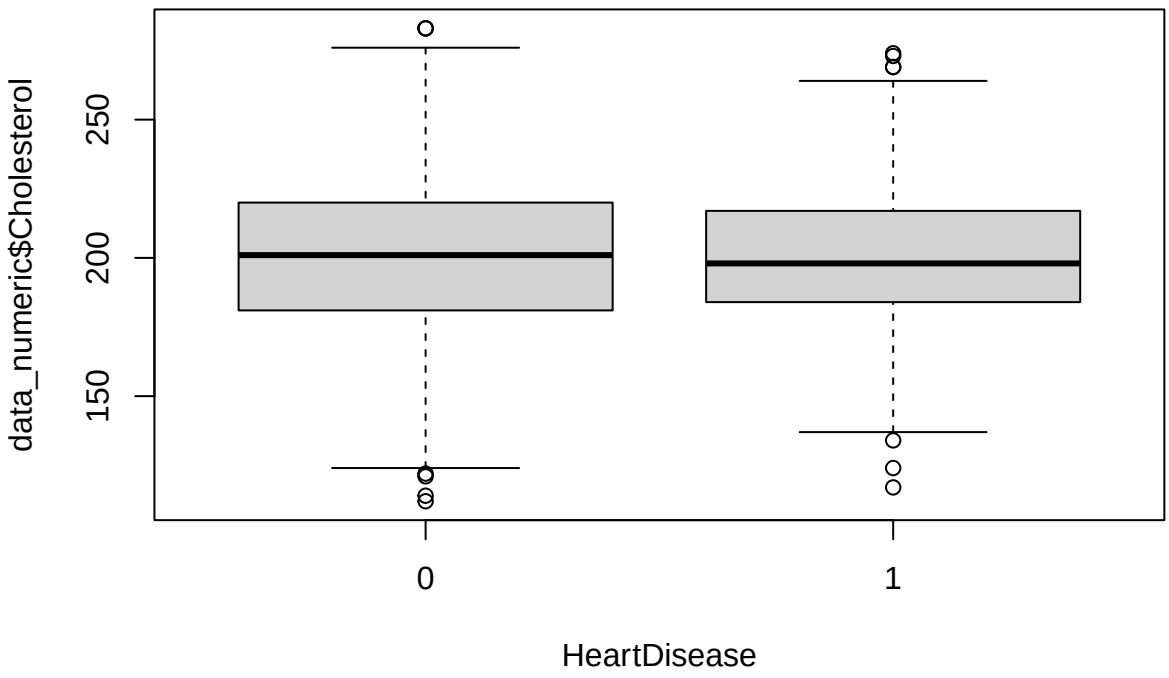


This plot gives the relationship plot between the blood pressure, cholesterol, heartdisease and Smoker.

Examining two variables at a time for more clarity - Identifying linearly Correlated variables:

Distribution of Cholesterol Vs Presence of HeartDisease

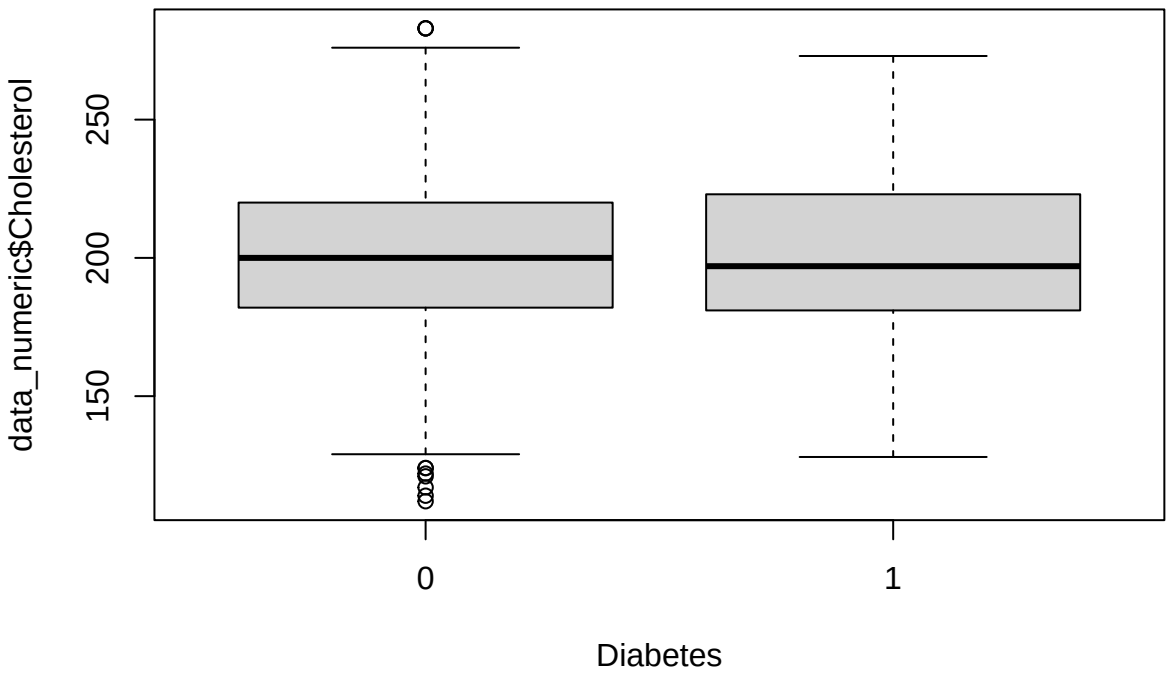
```
boxplot(data_numeric$Cholesterol ~ HeartDisease, data = data_numeric)
```



It is noted that both the presence and absence of Heart Disease cohort has outliers. For population without the heart disease, it is seen that there are outliers in the lower extreme (less than minimum value). These population have a very low cholesterol level. The maximum value of cholesterol in people with heart disease is higher than those without.

Distribution of Cholesterol Vs Presence of Diabetes

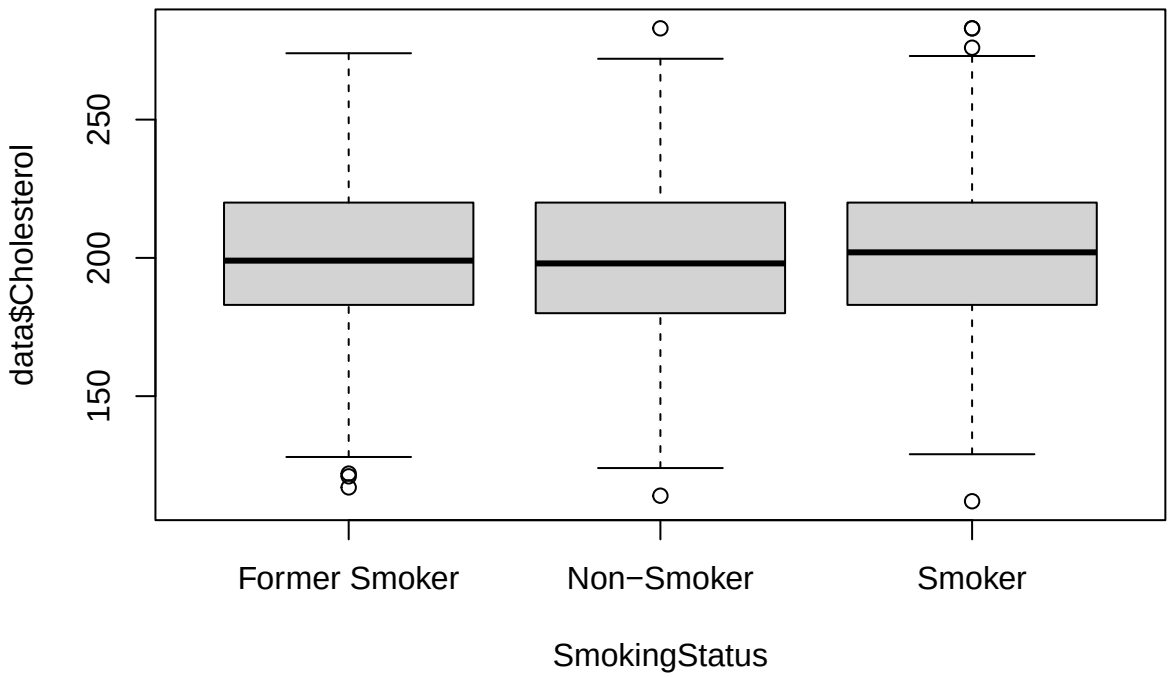
```
boxplot(data_numeric$Cholesterol ~ Diabetes, data = data_numeric)
```



It is seen that people with diabetes have no outliers in the levels of cholesterol, unlike those who don't have diabetes. People with diabetes are seen to have a higher Q3 values compared to those who don't have. The lower cholesterol levels are 128 and higher levels are 273, which is higher than the non diabetic cohort.

Smoking Vs Cholesterol Levels

```
boxplot(data$Cholesterol ~ SmokingStatus, data = data)
```

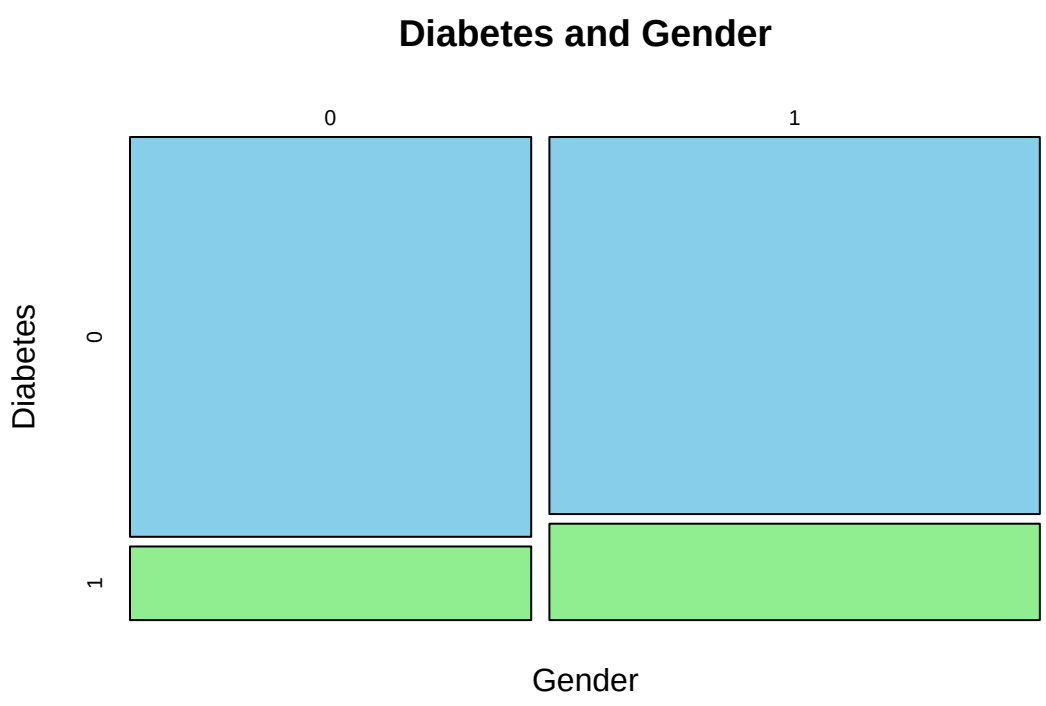


It is noted that the smokers have extreme outliers with high levels of cholesterol. Also the smokers have a higher value of 50th percentile compared to the non-smokers and former smokers. Former-smokers do have a higher median cholesterol level compared to the non-smokers.

So, Smoking and Cholesterol levels could be related.

Diabetes Vs Gender - How is it distributed across genders?

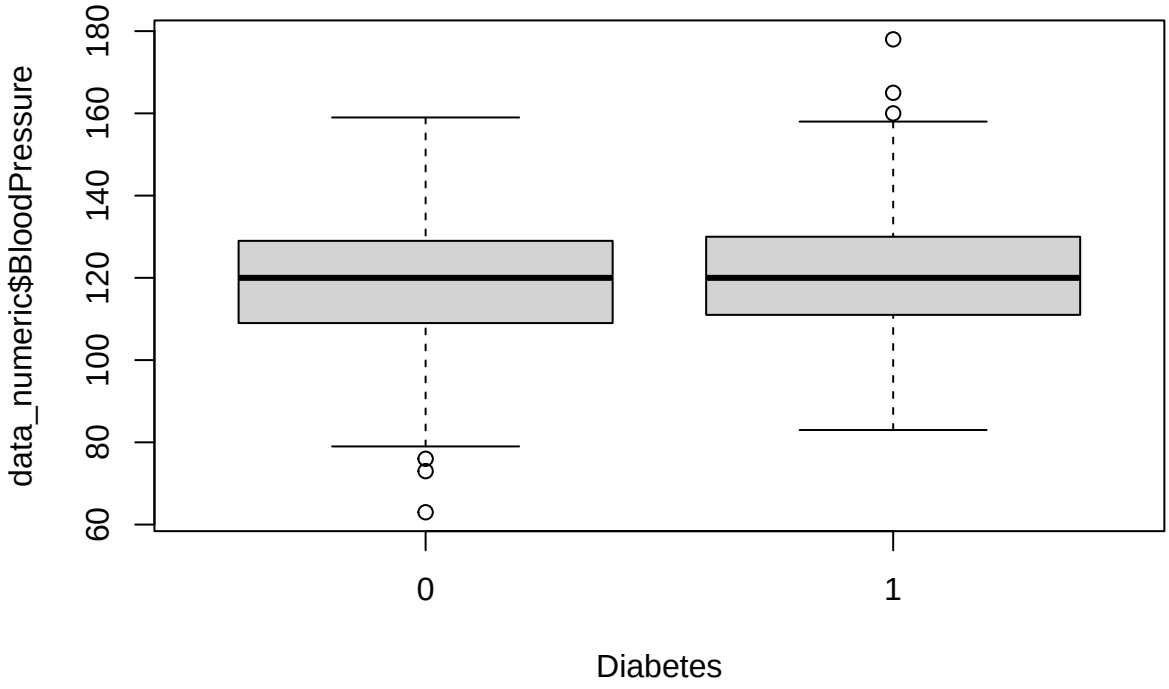
```
mosaicplot(table(data_numeric$Gender, data_numeric$Diabetes),
           color = c("skyblue", "lightgreen" ),xlab="Gender",ylab="Diabetes",
           main = "Diabetes and Gender")
```



From the plot, it is seen that the majority proportion of people with diabetes are females.

Diabetes vs BloodPressure

```
boxplot(data_numeric$BloodPressure ~ Diabetes, data = data_numeric)
```

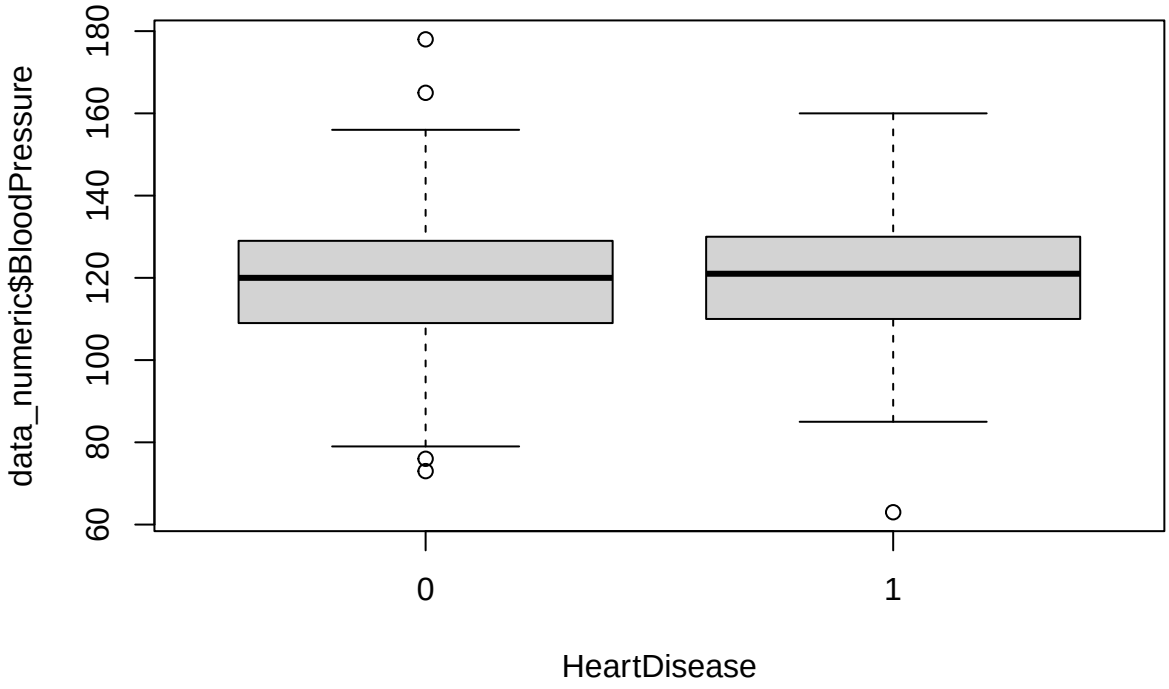


The presence of diabetes does tend to show an impact on the blood pressure levels. With the presence of diabetes, the blood pressure average is higher. The minimum value of bloodpressure level is higher in cohorts with diabetes than without. There are also outliers beyond the maximum value.

Diabetes are Bloodpressure could be related.

BloodPressure Vs Heart Disease

```
boxplot(data_numeric$BloodPressure ~ HeartDisease, data = data_numeric)
```

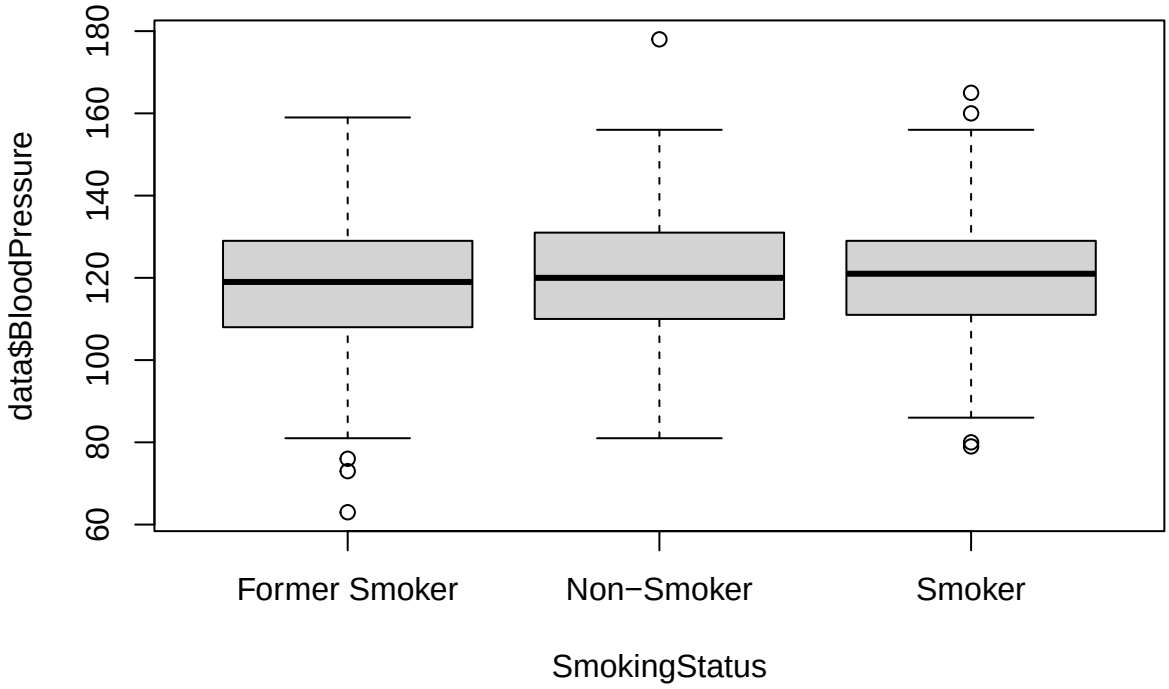


Though, the absence of heartdisease cohort has outliers, the presence of heart disease shows a slightly higher lower quartile and a higher maximum levels of blood pressure.

Hence, bloodpressure could be an indicator of heart disease.

BloodPressure Vs Smoking Status

```
boxplot(data$BloodPressure ~ SmokingStatus, data = data)
```



From the boxplot, it is seen that the smokers have a higher median compared to the non smoers and former smokers. the former smokers tend to show higher maximum levels of bloodpressure.

So, smoking and Bloodpressure might have an influence on each other.

The key variables/metrics with relationship are:

- **Cholesterol** and BloodPressure -> Inversely Related
- **Smoker** and cholesterol -> Positively related
- **HeartDisease** and BloodPressure -> Positively related
- **Gender** and Diabetes -> positively related, More Females are found to be diabetic.
- **BloodPressure** and Smoker/ Former_Smoker-> Positively related

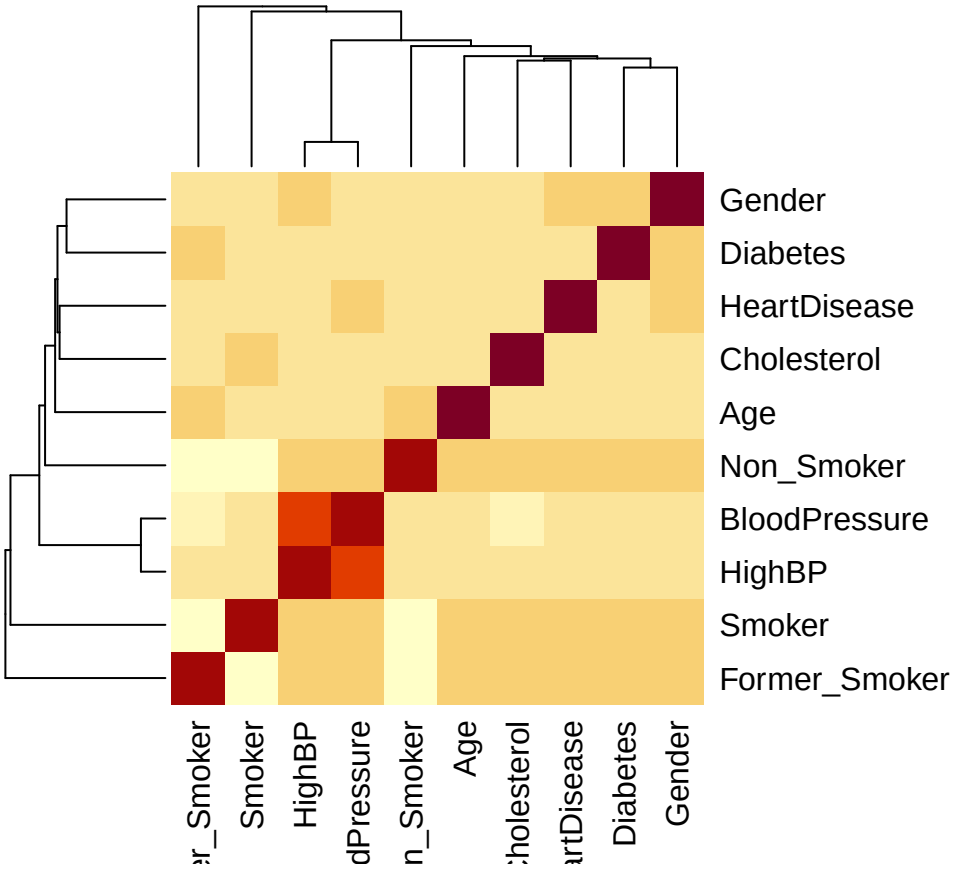
Q_1.2.4 Correlation between variables:

```
cor_matrix<-cor(data_numeric,method="spearman")

cor_matrix
```

```
##                     Age       Gender    Diabetes HeartDisease       Smoker
## Age          1.000000000 -0.011309360 -0.04648421 -0.016377118 -0.056845805
## Gender      -0.011309360  1.000000000  0.06199362  0.060046702 -0.009449073
## Diabetes    -0.046484207  0.061993622  1.00000000 -0.017322110 -0.010898504
## HeartDisease -0.016377118  0.060046702 -0.01732211  1.000000000  0.022315093
## Smoker      -0.056845805 -0.009449073 -0.01089850  0.022315093  1.000000000
```

```
## Non_Smoker     0.024633763 -0.005831848 -0.02057522 -0.005404110 -0.471648477
## Former_Smoker  0.031564179  0.014875258  0.03059013 -0.016528605 -0.518104503
## Cholesterol   -0.025947056  0.003617622 -0.02296748 -0.003878064  0.039251838
## BloodPressure -0.003896019  0.020793084  0.02182144  0.034335231  0.033371776
## HighBP        -0.002769953  0.048393331  0.01712278  0.001592913 -0.014834759
##                 Non_Smoker Former_Smoker  Cholesterol BloodPressure
## Age            0.024633763    0.03156418 -0.025947056  -0.003896019
## Gender        -0.005831848    0.01487526  0.003617622   0.020793084
## Diabetes      -0.020575224    0.03059013 -0.022967481   0.021821438
## HeartDisease  -0.005404110   -0.01652861 -0.003878064   0.034335231
## Smoker        -0.471648477   -0.51810450  0.039251838   0.033371776
## Non_Smoker     1.000000000   -0.50984421 -0.024722289   0.031271380
## Former_Smoker -0.509844207    1.00000000 -0.014313722  -0.062890021
## Cholesterol   -0.024722289   -0.01431372  1.000000000  -0.042353451
## BloodPressure  0.031271380   -0.06289002 -0.042353451   1.000000000
## HighBP         0.037662489   -0.02205922 -0.037303904   0.773561402
##                     HighBP
## Age           -0.002769953
## Gender         0.048393331
## Diabetes       0.017122782
## HeartDisease   0.001592913
## Smoker        -0.014834759
## Non_Smoker     0.037662489
## Former_Smoker -0.022059224
## Cholesterol   -0.037303904
## BloodPressure  0.773561402
## HighBP         1.000000000
```

```
heatmap(cor_matrix)
```



With the correlation coefficients and heatmap, it is seen that though the positive or negative relation is not that strong enough, there is a weak (positive or negative) relationship that exists among the key variables listed above.

Q_1.2.5 Creating a subset Data for Patients with Diabetes

```
data_numeric_diab<- data_numeric[data_numeric$Diabetes==1,]

head(data_numeric_diab)
```

```
##    Age Gender Diabetes HeartDisease Smoker Non_Smoker Former_Smoker Cholesterol
## 16  59      1        1            0      0          0             1         223
## 28  70      0        1            0      1          0             0         206
## 29  24      0        1            0      0          0             1         195
## 37  30      1        1            1      0          0             1         246
## 39  55      1        1            0      0          1             0         163
## 41  58      0        1            0      1          0             0         226
##    BloodPressure HighBP
## 16           116      0
## 28           126      0
## 29           139      1
## 37           103      0
## 39           128      0
## 41           138      1
```

```
summary(data_numeric_diab)
```

```
##       Age            Gender          Diabetes  HeartDisease        Smoker
##  Min.   :18.00   Min.   :0.0000   Min.   :1   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:35.00   1st Qu.:0.0000   1st Qu.:1   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :50.00   Median :1.0000   Median :1   Median :0.0000   Median :0.0000
##  Mean   :51.60   Mean   :0.6154   Mean   :1   Mean   :0.2802   Mean   :0.3132
##  3rd Qu.:66.75   3rd Qu.:1.0000   3rd Qu.:1   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :90.00   Max.   :1.0000   Max.   :1   Max.   :1.0000   Max.   :1.0000
##    Non_Smoker     Former_Smoker     Cholesterol    BloodPressure
##  Min.   :0.0000   Min.   :0.0000   Min.   :128.0   Min.   : 83.0
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:181.0   1st Qu.:111.0
##  Median :0.0000   Median :0.0000   Median :197.0   Median :120.0
##  Mean   :0.2967   Mean   :0.3901   Mean   :199.2   Mean   :121.1
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:223.0   3rd Qu.:129.8
##  Max.   :1.0000   Max.   :1.0000   Max.   :273.0   Max.   :178.0
##      HighBP
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2912
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

```
summary(data_numeric)
```

```
##       Age            Gender         Diabetes      HeartDisease       Smoker
##  Min.   :18.00   Min.   :0.00   Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:35.75   1st Qu.:0.00   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000
##  Median :53.00   Median :1.00   Median :0.000   Median :0.000   Median :0.000
##  Mean   :53.68   Mean   :0.55   Mean   :0.182   Mean   :0.297   Mean   :0.324
##  3rd Qu.:71.00   3rd Qu.:1.00   3rd Qu.:0.000   3rd Qu.:1.000   3rd Qu.:1.000
##  Max.   :90.00   Max.   :1.00   Max.   :1.000   Max.   :1.000   Max.   :1.000
##    Non_Smoker    Former_Smoker    Cholesterol    BloodPressure
##  Min.   :0.000   Min.   :0.000   Min.   :112.0   Min.   : 63.0
##  1st Qu.:0.000   1st Qu.:0.000   1st Qu.:182.0   1st Qu.:109.0
##  Median :0.000   Median :0.000   Median :200.0   Median :120.0
##  Mean   :0.317   Mean   :0.359   Mean   :200.5   Mean   :119.7
##  3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:220.0   3rd Qu.:129.0
##  Max.   :1.000   Max.   :1.000   Max.   :283.0   Max.   :178.0
##      HighBP
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.275
##  3rd Qu.:1.000
##  Max.   :1.000
```

Insights between Diabetes Vs Regular Cohort

1. From the summaries, it is noted that the **population with diabetes, show a higher minimum level of bloodpressure**.

2. Likewise, the population with diabetes tend to show a **higher minimum cholesterol level and higher maximum cholesterol level compared to the whole population**.
3. It is seen that the mean gender is higher for the diabetic population compared to overall gender mean; **which means more females have diabetes, than males**.
4. Surprisingly, **people with diabetes, show less heart disease frequency** than people in the whole population.
5. There are **more people with highBlood Pressure in the diabetes cohort**, than in the overall population.

Linear Regression Analysis

Q_2.1.1 Model Formulation - Fitting the linear Regression Model

Predicting BloodPressure

From the correlation of variables, it is seen that the variables HighBP, HeartDisease, Smoker, Cholesterol, Diabetes shows a relation with the bloodpressure.

- HighBP has higher correlation with Bloodpressure (Since it is a Derived Column)
- HeartDisease has a positive correlation(though slightly but more compared to others) with bloodpressure.
- Smoker has a positive correlation with Bloodpressure too.
- Diabetes has the a very mild positive correlation of value 0.021 with Bloodpressure.
- For this population. cholesterol has an inverse effect on the Bloodpressure, which could be a good representing factor

Correlation levels with BloodPressure:

- Gender : 0.020793084
- Diabetes : 0.021821438
- HeartDisease : 0.034335231
- Smoker : 0.033371776
- Non_Smoker : 0.031271380
- Former_Smoker : -0.062890021
- Cholesterol : -0.042353451
- BloodPressure : 1.000000000
- HighBP : 0.773561402
- Age : -0.003896019

Q_2.1.2 Equation of the Model

```
library(tinytex)
```

$$Y = b_0 + b_1 * (diabetes) + b_2 * (heartdisease) + b_3 * (smoker) + b_4 * (cholesterol) + b_5 * (highBP) + \varepsilon$$

Y is the dependent variable that determines blood pressure. The b0 is the intercept, which is the value of the bloodpressure when all the other metrics are zero.

b1, b2, b3, b4 and b5 are the slopes(explains change in Y with unit change in the corresponding independent variable) or coefficients of the model.

And epsilon is the error term associated with the model.

Q_2.1.3 What part of the model remains in knowable and unknowable world?

Before fitting in the model, we have the information regarding the independent variables which we chose. Since this is present in the data, this is considered as the knowable information. However, te dependent variable's prediction from the model(model performance) is in the unknowable world.

But Post fitting in the model, the parts of the model that are in the knowable world are:

- Intercept
- Slopes/coefficients
- Y-> dependent variable(predictions)

Parts of the model that are in the unknowable world are:

- Errors/ Residuals (Difference between actual and predicted)
- The overall population as a whole is unknown (we generally estimate the population using the current results)
- Unseen Data that could affect dependent variable

Model Fitting and Interpretation

Q_2.2.1 Fit the linear model and show the summary

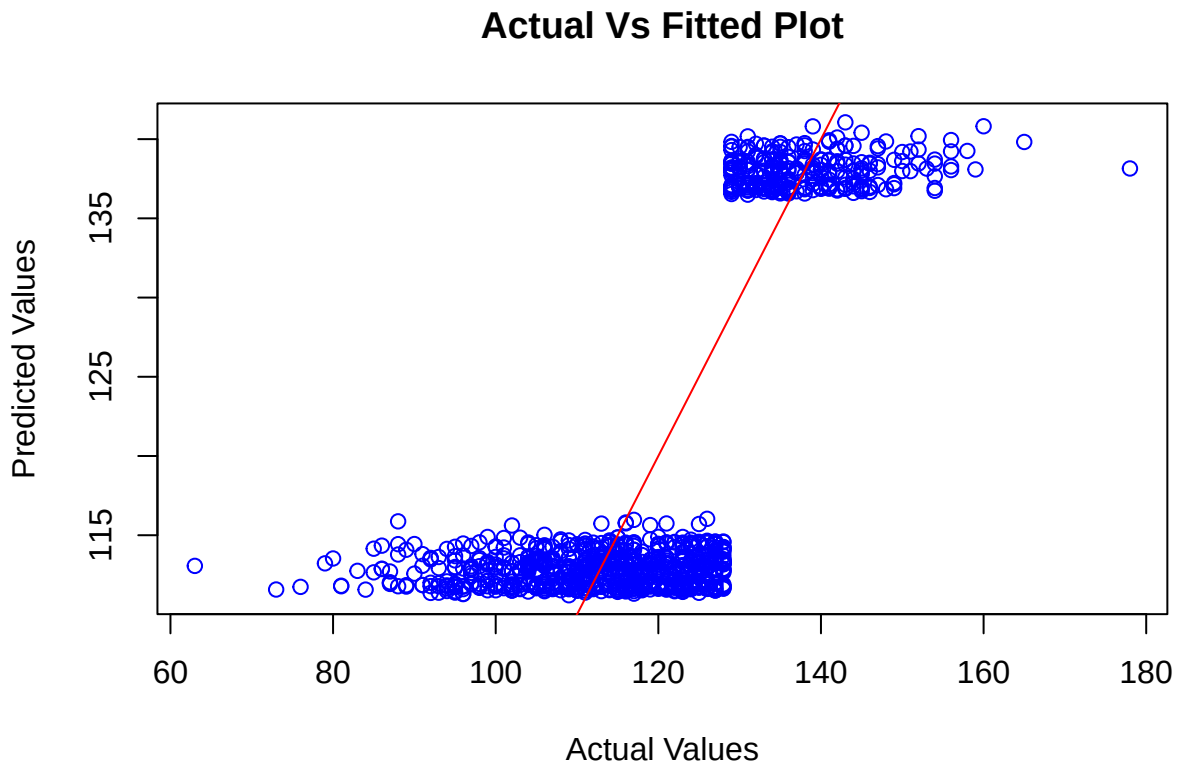
```
model_1 <-lm(data_numeric$BloodPressure ~ data_numeric$Diabetes+
             data_numeric$HeartDisease+
             data_numeric$Smoker+
             data_numeric$Cholesterol+
             data_numeric$HighBP,
           data=data_numeric)
summary(model_1)
```

```
##
## Call:
## lm(formula = data_numeric$BloodPressure ~ data_numeric$Diabetes +
##     data_numeric$HeartDisease + data_numeric$Smoker + data_numeric$Cholesterol +
##     data_numeric$HighBP, data = data_numeric)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

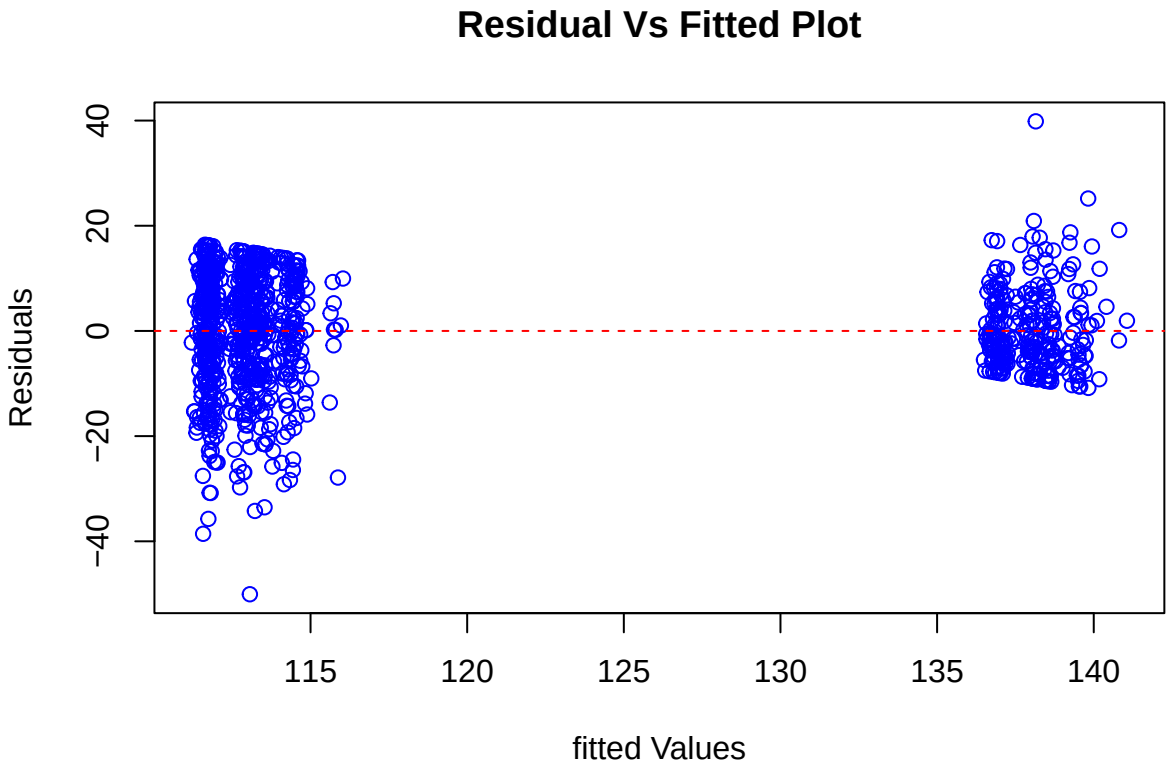
```
## -50.063  -6.358   0.323   7.876  39.851
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                113.100822   2.317077  48.812   <2e-16 ***
## data_numeric$Diabetes        1.225756   0.838646   1.462    0.144
## data_numeric$HeartDisease    1.044339   0.708106   1.475    0.141
## data_numeric$Smoker          1.639169   0.691799   2.369    0.018 *
## data_numeric$Cholesterol    -0.006678   0.011273  -0.592    0.554
## data_numeric$HighBP         25.151479   0.724986  34.692   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.23 on 994 degrees of freedom
## Multiple R-squared:  0.5504, Adjusted R-squared:  0.5482
## F-statistic: 243.4 on 5 and 994 DF,  p-value: < 2.2e-16
```

Plotting the Actual Vs Predicted

```
pred <-predict(model_1)
plot(data_numeric$BloodPressure,pred,xlab="Actual Values",
     ylab="Predicted Values",main="Actual Vs Fitted Plot",col="blue")

abline(a=0, b=1, col="red")
```



```r
plot(model_1$fitted.values,model_1$residuals,xlab="fitted Values",
     ylab="Residuals",main="Residual Vs Fitted Plot",col="blue")

abline(h=0,col="red",lty=2)
```



Probable Model Inefficiency Reasons:

From the actual vs predicted plot, it is seen that the following are violated:

- The distribution is *non linear*, which means the predicted values are biased by a specific independent variable.
- *There could be multicollinearity*, since there is *HIGHBP* variable that has a threshold value, and is a derived column, so that directly influences bloodpressure, thereby suppressing the contribution of other variables.
- The model might have been overfitted, leading to outcomes that are unreliable

Fitting the Model again

Removing Residuals from Cholesterol

If there are extreme values, the maximum value is assigned to it. If there are lower values, the minimum value is assigned to it.

```r
library(dplyr)

upper_chol<-max(data$Cholesterol)
lower_chol<-min(data$Cholesterol)


data_numeric<- mutate(data_numeric,
Cholesterol_1 = ifelse(Cholesterol>upper_chol, upper_chol, ifelse(Cholesterol<lower_chol,lower_chol,Cho

head(data_numeric)
```

```
##   Age Gender Diabetes HeartDisease Smoker Non_Smoker Former_Smoker Cholesterol
## 1  48      1        0            1      0          0             1         215
## 2  68      1        0            1      0          0             1         215
## 3  31      1        0            0      1          0             0         151
## 4  84      0        0            0      0          0             1         220
## 5  59      0        0            0      0          1             0         234
## 6  67      0        0            0      0          1             0         236
##   BloodPressure HighBP Cholesterol_1
## 1            87      0           215
## 2           132      1           215
## 3           101      0           151
## 4           121      0           220
## 5           102      0           234
## 6           122      0           236
```

Dropping the HighBP variable are including the next positive correlated variable 'Gender', and adding former smoker variable, and removing smoker since it has a better correlation, shows a relationship (negative) with blood pressure than smoker.

```r
model_2 <-lm(data_numeric$BloodPressure ~ data_numeric$Diabetes+
             data_numeric$HeartDisease+
             data_numeric$Former_Smoker+
             data_numeric$Cholesterol_1+
             data_numeric$Gender+
               data_numeric$Age,
           data=data_numeric)
summary(model_2)
```

```
##
## Call:
## lm(formula = data_numeric$BloodPressure ~ data_numeric$Diabetes +
##     data_numeric$HeartDisease + data_numeric$Former_Smoker +
##     data_numeric$Cholesterol_1 + data_numeric$Gender + data_numeric$Age,
##     data = data_numeric)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.174 -10.400   0.312   9.815  56.585
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)                123.610008   3.710771  33.311   <2e-16 ***
## data_numeric$Diabetes        1.722012   1.249102   1.379   0.1683
## data_numeric$HeartDisease    1.052656   1.053052   1.000   0.3177
## data_numeric$Former_Smoker  -2.363243   1.002028  -2.358   0.0185 *
## data_numeric$Cholesterol_1  -0.021191   0.016724  -1.267   0.2054
## data_numeric$Gender          0.727431   0.968996   0.751   0.4530
## data_numeric$Age             0.003792   0.023039   0.165   0.8693
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.18 on 993 degrees of freedom
## Multiple R-squared:  0.01051,    Adjusted R-squared:  0.004531
## F-statistic: 1.758 on 6 and 993 DF,  p-value: 0.1047
```

Q_2.2.2 Interpretation of Model Summary

The estimated parameters are the coefficients of each of the independent variable that could influence the predictor/dependent variable.

- Here the intercept has the coefficient of 123.610008
- Diabetes : 1.722012
- HeartDisease : 1.052656
- Former_Smoker : -2.363243
- Cholesterol : -0.021191
- Gender : 0.727431
- Age : 0.003792

According to the significance level, it is noted that the variable Former smoker has a p value less than 0.05 which indicates a good significance level associated with the variable.

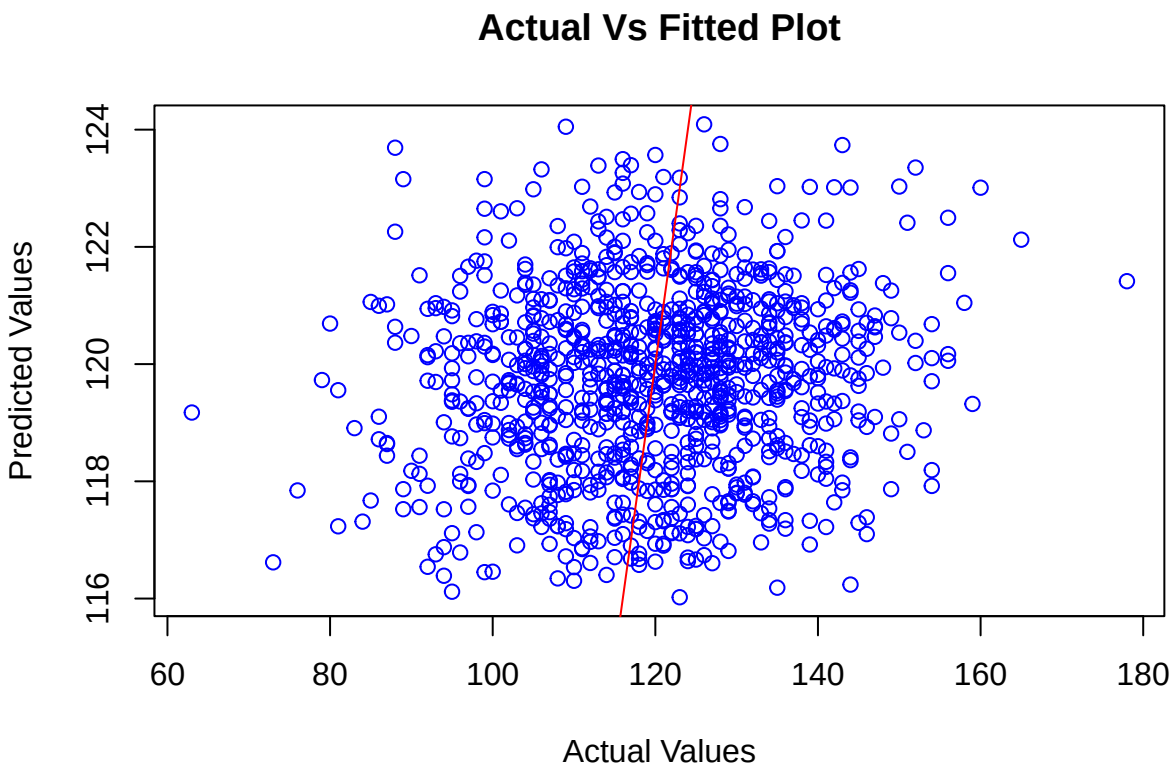
The intercept has a 3 star (p value less than 0.001) with high significance in contributing to the predictor.

From the health outcome perspective, the variables former smoker is most contributing to the values of blood pressure. In other terms, the probability of having another variable with the absence of former smoker is 0.001.

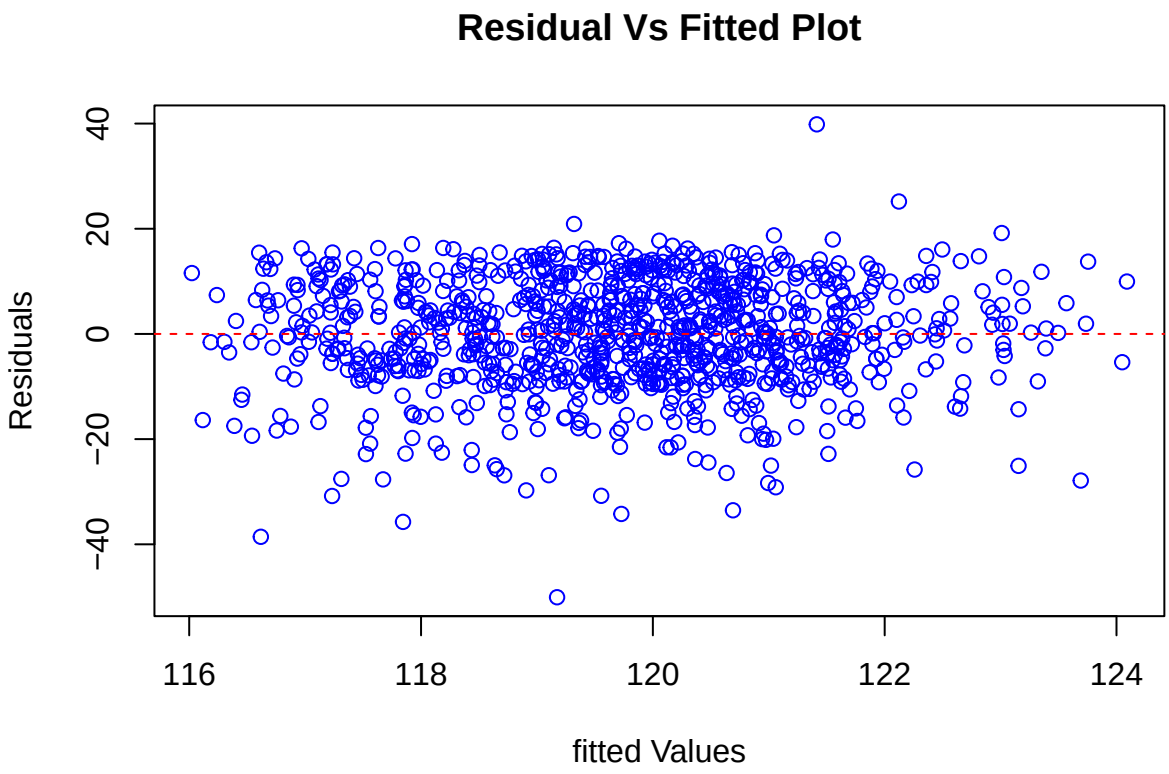
The model has an R square value of 0.0105, which isn't a great performing model.

Q_2.2.3 Assesing the model using plots

```
pred <-predict(model_2)
plot(data_numeric$BloodPressure,pred,xlab="Actual Values",
     ylab="Predicted Values",main="Actual Vs Fitted Plot",col="blue")

abline(a=0, b=1, col="red")
```




```
plot(model_2$fitted.values,model_1$residuals,xlab="fitted Values",
     ylab="Residuals",main="Residual Vs Fitted Plot",col="blue")

abline(h=0,col="red",lty=2)
```



From the fitted vs actual graph, it is seen that the regression line is passing through the data in such a way that it isn't capturing much of the data.

The residuals = (actual-predicted), are plotted vs the predicted values.

It is noted that the residuals are spread and scattered around the line, indicating that it is capturing the relationship well.

It could also be inferred that there is no overfitting in the model and no multicollinearity.

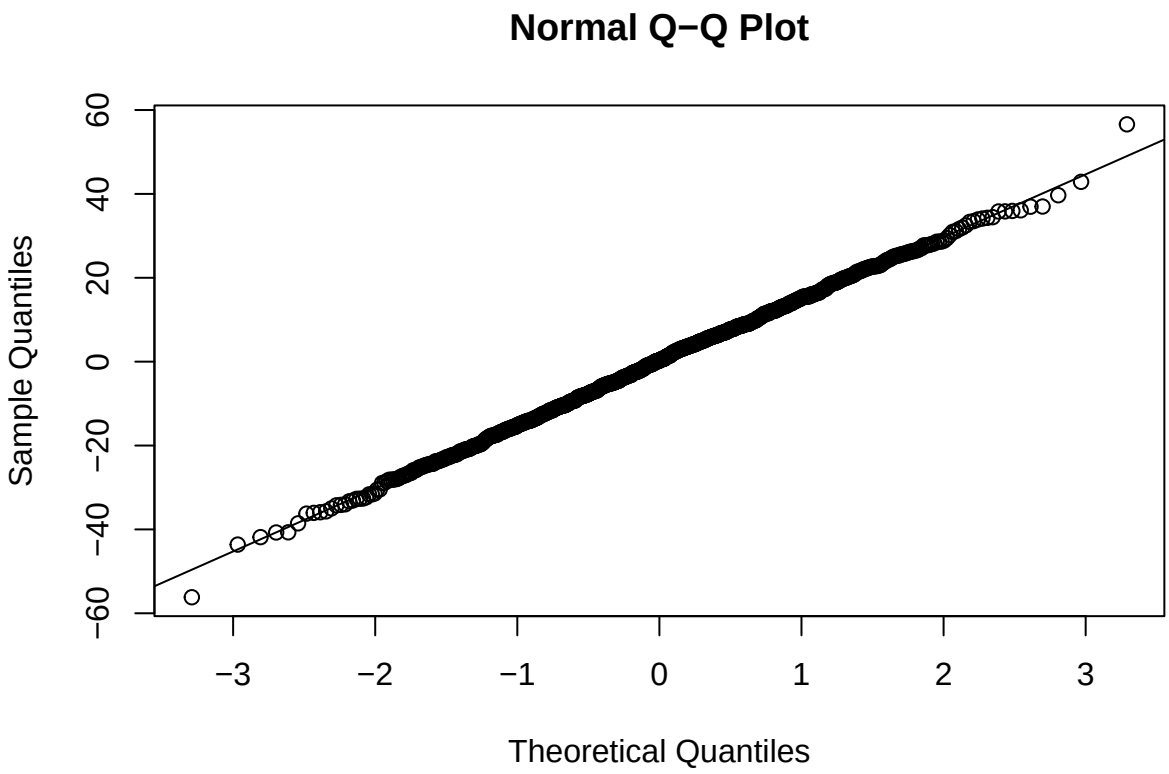
There is a linearity in the model.

Model Evaluation and Hypothesis Testing

Q_2.3.1 Hypothesis Testing

Plotting quantiles to check if residuals are normally. If not, they could affect the hypothesis tests, which are actually based on the residuals.(t values)

```
qqnorm(model_2$residuals)
qqline(model_2$residuals)
```



It is noted that the residuals are normally distributed.

```
library(tinytex)
```

Let the chosen variable be Former Smoker, whose coefficient is indicated by b_3.

$$H_0 : b_3 = 0 H_1 : b_3 \neq 0$$

The null hypothesis states that predictor coefficient is zero, and there is no relationship with the predictor and the target.

The alternate hypothesis states that there is a effect on target by this predictor.

```
hypothesis<-summary(model_2)$coefficients["data_numeric$Former_Smoker",
                                          c("t value","Pr(>|t|)")]
hypothesis
```

```
##     t value    Pr(>|t|)
## -2.35845927  0.01854396
```

```
#Setting the alpha/significance level as 0.05

sig_level<-0.05

if(hypothesis['Pr(>|t|)']<sig_level)
```

```
{
  print("Reject the null hypothesis")
}else
{
  print("Do not reject the null hypothesis")
}
```

```
## [1] "Reject the null hypothesis"
```

Since, the p value is less than the significance level, we reject the null hypothesis. The alternate hypothesis is accepted, therefore there is a relationship between the predictor and the target.

Type 1 Error: (False Positive)

In this case the true null hypothesis is rejected, where actually, there is a no relationship between former_smoker and blood pressure.

The result of which could be faulty inclusion of the variable in the model, assuming that there is a relationship, where there is actually no relation.

This could affect the linearity of the model.

Type 2 Error: (False Negative)

In this case, the null hypothesis is accepted, when there is actually a relationship between the predictor and the target.

In this case, we ignore the presence of the predictor and the way it could have helped in the prediction of target.

Q_2.3.2 Assesing the Overall Fit

- The bloodpressure variable(target), was attempted to be explained by the independent variables heartdisease, diabetes, age, gender, cholesterol and former_smoker
- According to the model's residual vs fitted values plot, it is noted that the model has only partially captured the data. The linear pattern is not well studied, with many data points uncaptured, hence has a low R square ratio.
- However, from the residual plot, it is interpreted that the model has no over fitting, has homoskedascity throughout (similiar variance for all predictors). It also implies there is a linear relationship.
- Among the variables that contribute to the target variable, the variable former_smoker has more significance than the other variables.
- The Multiple R square value is 0.0105, and the maximum residual value is 56.585 and minimum residual value is -56.174.
- The residuals follow a normal distribution
- With R square of 0.0105, the model could explain only around 1.05 percent of the variation in the target variable

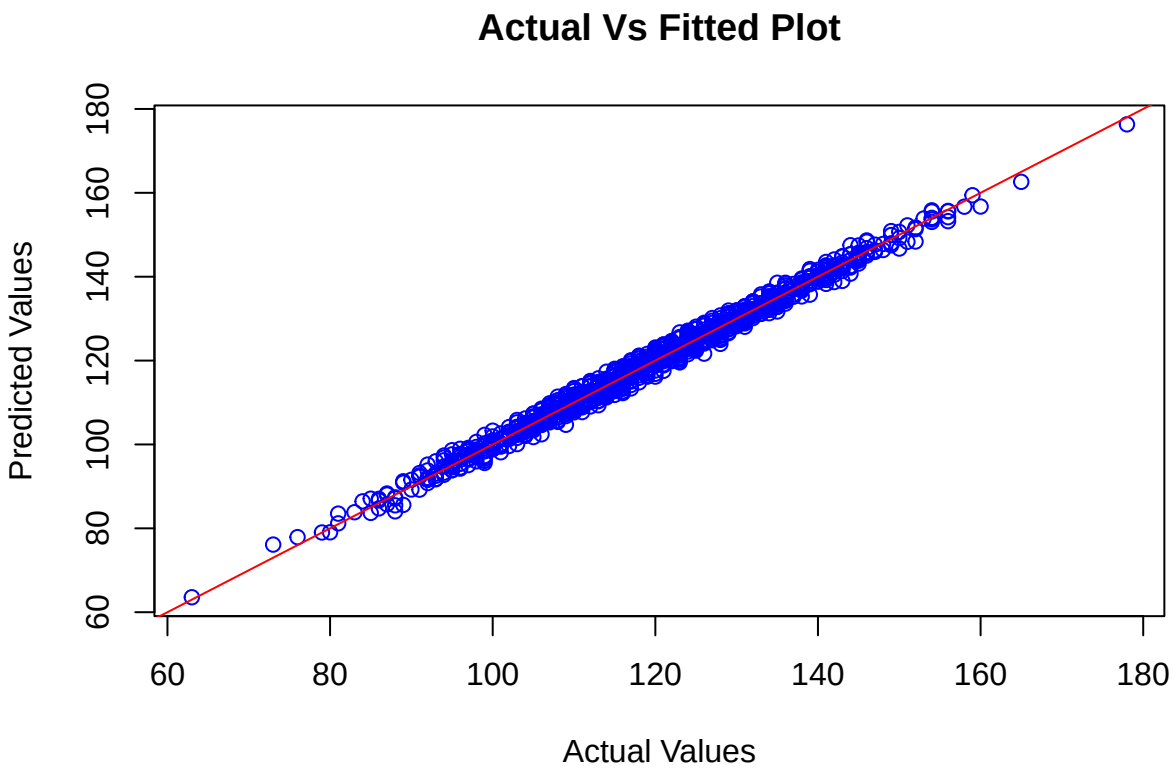
Q_2.3.3 Plotting the residuals back to the model

```
model_residual <-lm(data_numeric$BloodPressure ~ model_2$residuals)

summary(model_residual)
```

```
##
## Call:
## lm(formula = data_numeric$BloodPressure ~ model_2$residuals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7204 -1.1182  0.0977  1.0401  4.3473
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.197e+02  4.935e-02  2426.4   <2e-16 ***
## model_2$residuals 1.000e+00  3.262e-03   306.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.561 on 998 degrees of freedom
## Multiple R-squared:  0.9895, Adjusted R-squared:  0.9895
## F-statistic: 9.396e+04 on 1 and 998 DF,  p-value: < 2.2e-16
```

Fitting the actual vs Predicted for the residual fitted model.

```
pred <-predict(model_residual)
plot(data_numeric$BloodPressure,pred,xlab="Actual Values",
     ylab="Predicted Values",main="Actual Vs Fitted Plot",col="blue")

abline(a=0, b=1, col="red")
```



- If the residuals are plotted to the model, the model generally captures the error terms and generally performs well.
- Since it captures the model differences between actuals and predicted, it tends to **overfit**
- The model has a R square of 0.98 which actually explain 98 percent of the changes in the target variable. This implies overfitting of the model
- If the model is exposed to unseen data, it is unlikely that the model would give the same good performance.

Q_3.Fun with Linear Model

```
n <- 100
p <- 95
x <- rnorm(n*p)
dim(x) <-c(n,p)
y <- x[,1] - 1.2*x[,2] + rnorm(n)
fit.lm = lm(y ~ x)
```

What is this code doing? What is the Purpose?

n is the number of rows p is the number of columns

X is a data set of normal distribution, with n rows and p columns.

The dimensions of the data set x are set to 100 rows and 95 columns.

Then, a vector Y is created by the following computation:

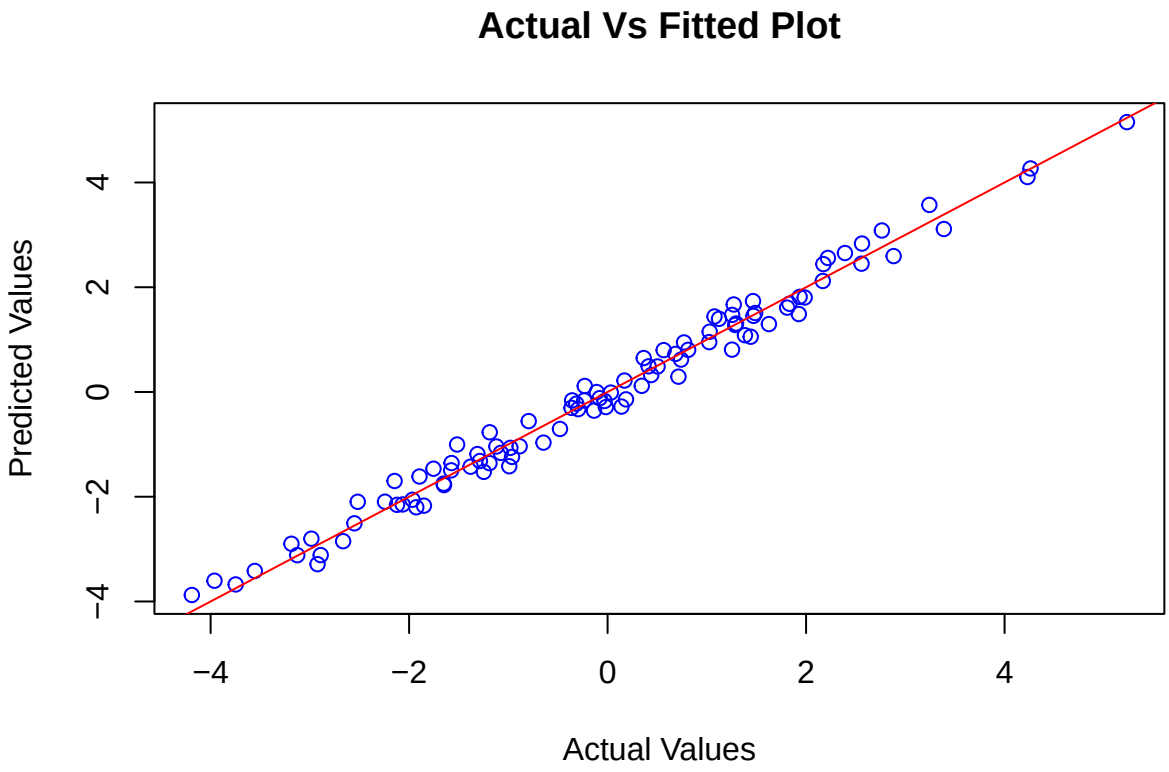
1.2 times of second column of X is subtracted from the first column. Then added with a linear combination of random normal numbers of size 100, which is equal to the rows.

The random distributed values in the dataset X, entirely is used to predict the variable Y which has a sum of different set of random variable and a difference between the existing columns of X.

The purpose of this linear model is to predict a variable Y which is a combination of already existing columns and a sum random normal distribution.

Interpret the output fit.lm in light of the input you created. Is it what you expected; why or why not?

```
pred_norm <-predict(fit.lm)
plot(y,pred_norm,xlab="Actual Values",ylab="Predicted Values",main="Actual Vs Fitted Plot",col="blue")

abline(a=0, b=1, col="red")
```



The ideal output of the model has to be normally distributed too, and should be able to accurately predict the variable Y.

The expected output of the model is a normal distribution since the target variable y is a combination of normally distributed variables.

```r
summ <- summary(fit.lm)
print(summ$r.squared)
```

```
## [1] 0.9849227
```

The predictions of this model is quite accurate.

Why do you think this question might have been called "fun"?

This could have been called fun because the variable creation Y (dependent variable) was done using X's two variables, and then added a random list of numbers to it, and this variable was taken as a target.

Also, it is fun as the rnorm() function generates a new set of random variables everytime the model is run. Also, the values of Y is different everytime it is executed.