SUMMARISING DATA

CASE STUDY SCENARIO:

Heart Disease Detection: UCI Irvine Repository

Attribute Information:

1. age
2. sex (0:Female 1:Male)
3. chest pain type (4 values) Value 1: typical angina (Definite heart discomfort) Value 2: atypical angina (Probable) Value 3: non-anginal pain Value 4: asymptomatic (no symptoms)
4. resting blood pressure
5. serum cholestoral in mg/dl HDL+LDL+Triglycerides Normal is below 200.Above 400 is high risk.
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2) Normal ,Having ST T wave abnormality, showing probable or definite left ventricular hypertrophy by Estes' criteria. Higher levels higher risk of heart attack
8. maximum heart rate achieved (Used by doctors for heart risks)
9. exercise induced angina an indicator for heart disease related to obstruction of artery
10. oldpeak = ST depression induced by exercise relative to rest (slope (up, flat, down)
11. the slope of the peak exercise ST segment Heart rate slope calculated while excercising. The higher the slope the more chances of heart attack
12. number of major vessels (0-3) colored by flourosopy Detecting coronary calcification which is an indicator of heart disease
13. thal: 3 = normal; 6 = fixed defect; 7 = reversable

QUANTITATIVE DATA

Measures of Central Tendency:

Mean : the average (Most typical value) of the given numerical values. All the observations will have to be added and divided by number of observations

```
setwd("~/Desktop/Stat/Stat135/data")
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 3.4.4
```

```
HeartAttack <- read_excel("HeartAttack.xlsx")
```

```
## Warning in strptime(x, format, tz = tz): unknown timezone 'zone/tz/2018c.
## 1.0/zoneinfo/America/Los_Angeles'
```

```
Resting_Blood_Pressure <- HeartAttack$Resting_Blood_Pressure
mean(Resting_Blood_Pressure)
```

```
## [1] 131.3444
```

Median: Is the midmost value. It is called the 50th percentile.

```
median(Resting_Blood_Pressure)
```

```
## [1] 130
```

MEASURES OF DISPERSION Standard Deviation: the square root of the variance Variance : the average of square of deviations therefore the unit of variance is a squared unit of the original data wheras the data has an unsquared unit

```
sd(Resting_Blood_Pressure)
```

```
## [1] 17.86161
```

```
var(Resting_Blood_Pressure)
```

```
## [1] 319.0371
```

First Quartile: 25 th Percentile. Splits the data into lower 25% of the observations and higher 75% of the data

Second Quartile: Is the median of the observations Splits the data into lower 50% of the observations and higher 50% of the data

Third Quartile: 75 th Percentile. Splits the data into lower 25% of the observations and higher 75% of the data

```
quantile(Resting_Blood_Pressure)
```

```
##   0%  25%  50%  75% 100%
##   94  120  130  140  200
```

CATEGORICAL DATA

Categorical Variables can be summarized by the use of Counts, Proportions and Relative Frequencies. Relative Frequencies provide a good measure if there are unequal values in the given categories.

The Counts Percentages generated by R : Variable used are Pain Type and Gender.

Following these measures the Crosstabulated Data (Contingency Table) is given for these two variables. Pain Type Sex 1:Male 0:Female

```
PainType <- table(HeartAttack$Chest_Pain_Type)
prop.table(PainType)*100
```

```
##
##         1         2         3         4
##  7.407407 15.555556 29.259259 47.777778
```

```
addmargins(prop.table(PainType)*100)
```

```
##
##          1          2          3          4        Sum
##   7.407407  15.555556  29.259259  47.777778 100.000000
```

```
Gender <- table(HeartAttack$Sex)
Gender
```

```
##
##   0   1
##  87 183
```

```
addmargins(prop.table(Gender)*100)
```

```
##
##         0         1       Sum
##  32.22222  67.77778 100.00000
```

```r
PainTypeGender <- table(HeartAttack$Chest_Pain_Type,HeartAttack$Sex)

addmargins(PainTypeGender)
```

```
##
##          0   1 Sum
##   1      4  16  20
##   2     16  26  42
##   3     32  47  79
##   4     35  94 129
##   Sum   87 183 270
```

```r
addmargins(prop.table(PainTypeGender)*100)
```

```
##
##                 0          1        Sum
##   1      1.481481   5.925926   7.407407
##   2      5.925926   9.629630  15.555556
##   3     11.851852  17.407407  29.259259
##   4     12.962963  34.814815  47.777778
##   Sum   32.222222  67.777778 100.000000
```

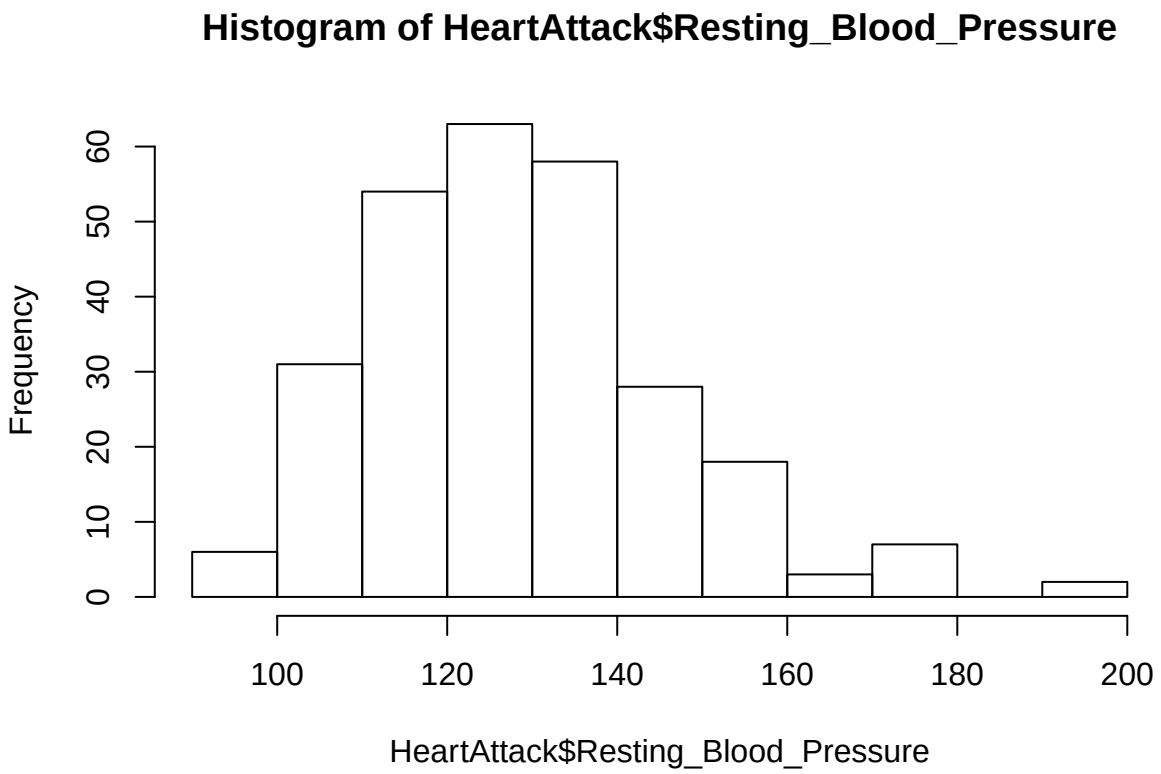
GRAPHICAL SUMMARIES

Graphical summary Enables us to visually ascertain the mean, median, mode (measures of central tendencies), standard deviation (spread/dispersion/deviation) and the shape (skewed or bell shaped) of the distribution.

Histogram

Consists of parallel vertical bars that depict the frequency distribution of the quantitative data set. The height of the bar is the frequency of that data interval.

```r
hist(HeartAttack$Resting_Blood_Pressure)
```



Slightly Right Skewed

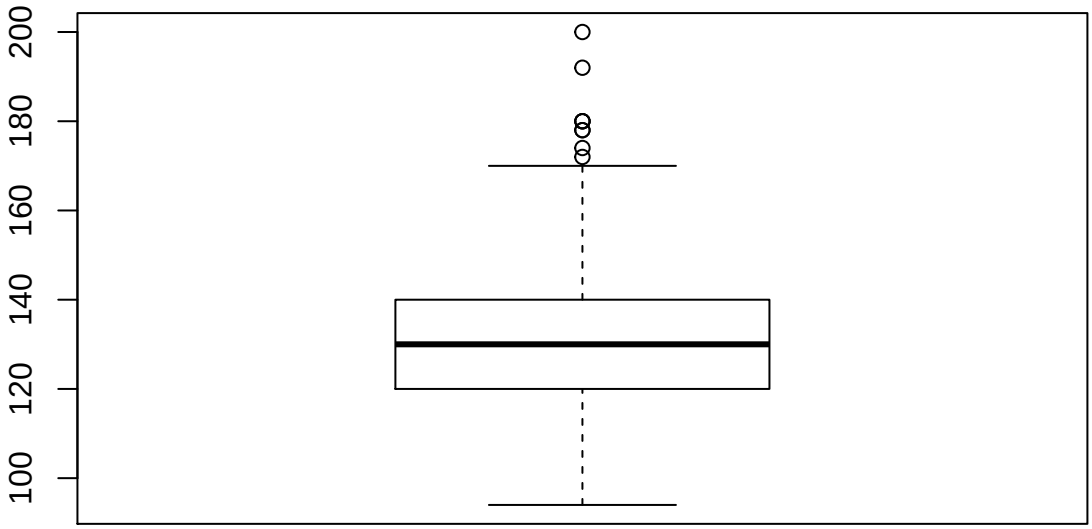
Stem and leaf a graphical display that displays quantitative data according to the most significant digit in the data set. Each row is placed adjacent to the next according to the ascending order.

```
stem(HeartAttack$Resting_Blood_Pressure)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##    9 | 44
##   10 | 000012245556888888
##   11 | 00000000000000002222222255578888888
##   12 | 000000000000000000000000000000022234444455555555556668888888889
##   13 | 00000000000000000000000000000022222244445555556668888888888
##   14 | 000000000000000000000000000022245555568
##   15 | 0000000000000002222568
##   16 | 000000000005
##   17 | 002488
##   18 | 000
##   19 | 2
##   20 | 0
```

Boxplot a visual display of quantitative data and is shows the division in terms of Min, Q1, Q2, Q3, Max, and outliers

```
boxplot(HeartAttack$Resting_Blood_Pressure)
```



Slightly right skewed. The lower 25% of the data set is less dispersed ie the BP of the first 25% of the data has less variation as compared to the highest 25% of the dataset. Lowest 25% of the data lie between 94—120 (Difference of 18 points) Highest 25% of the data lie between 140—170 (Difference 30 points)

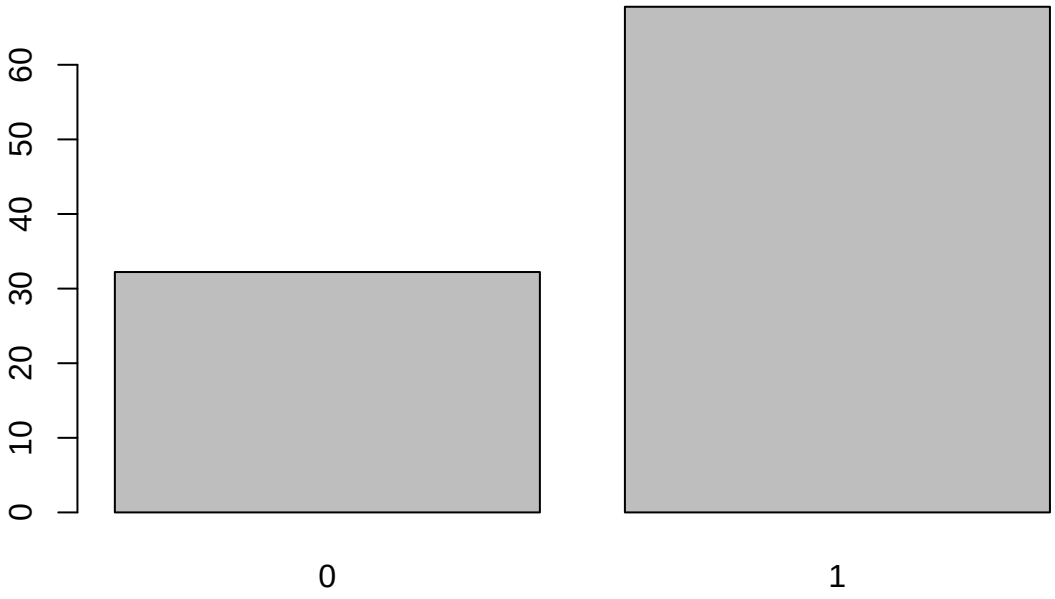
CATEGORICAL

A bar diagram represents the frequency of occurrences of different values of X.The bar diagram helps differentiate between the independent variables and the dependent variables.

```
prop.table(Gender)*100
```

```
##
##        0        1
## 32.22222 67.77778
```

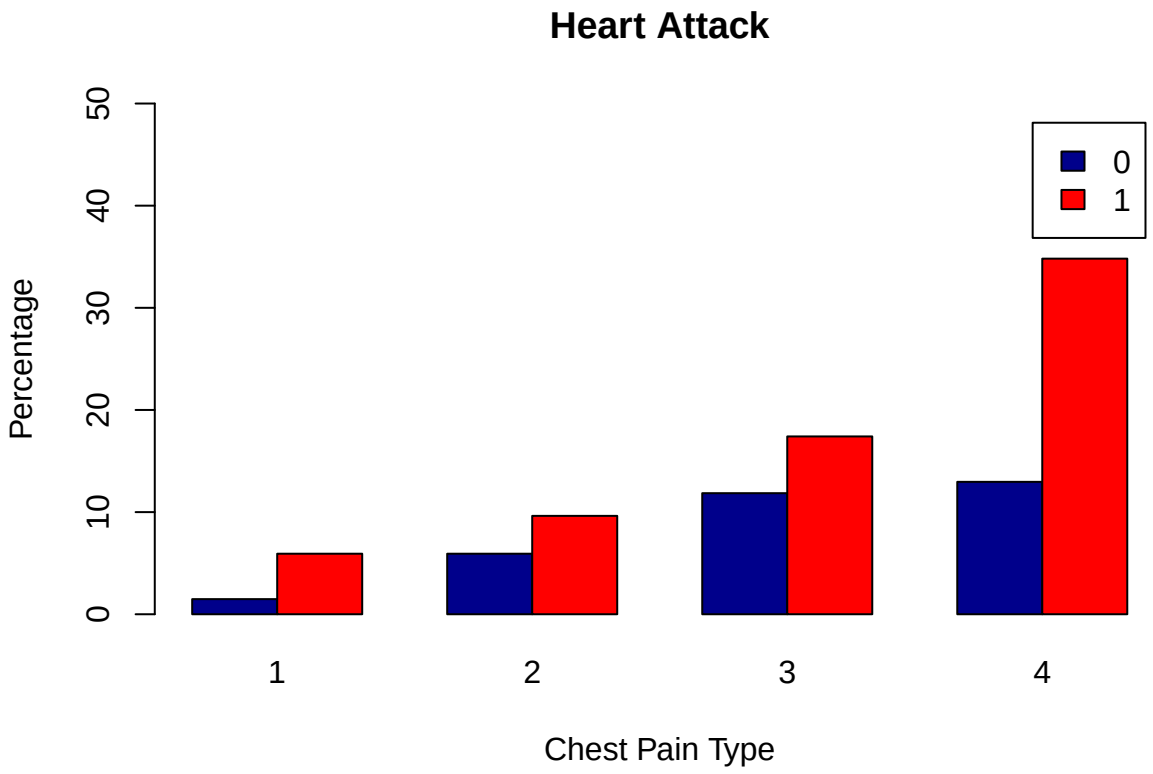
```
Gender <- table(HeartAttack$Sex)
barplot(prop.table(Gender)*100)
```



```
Gender_Chest_Pain_Type <- table(HeartAttack$Sex,HeartAttack$Chest_Pain_Type)
prop.table(Gender_Chest_Pain_Type )
```

```
##
##              1          2          3          4
##   0 0.01481481 0.05925926 0.11851852 0.12962963
##   1 0.05925926 0.09629630 0.17407407 0.34814815
```

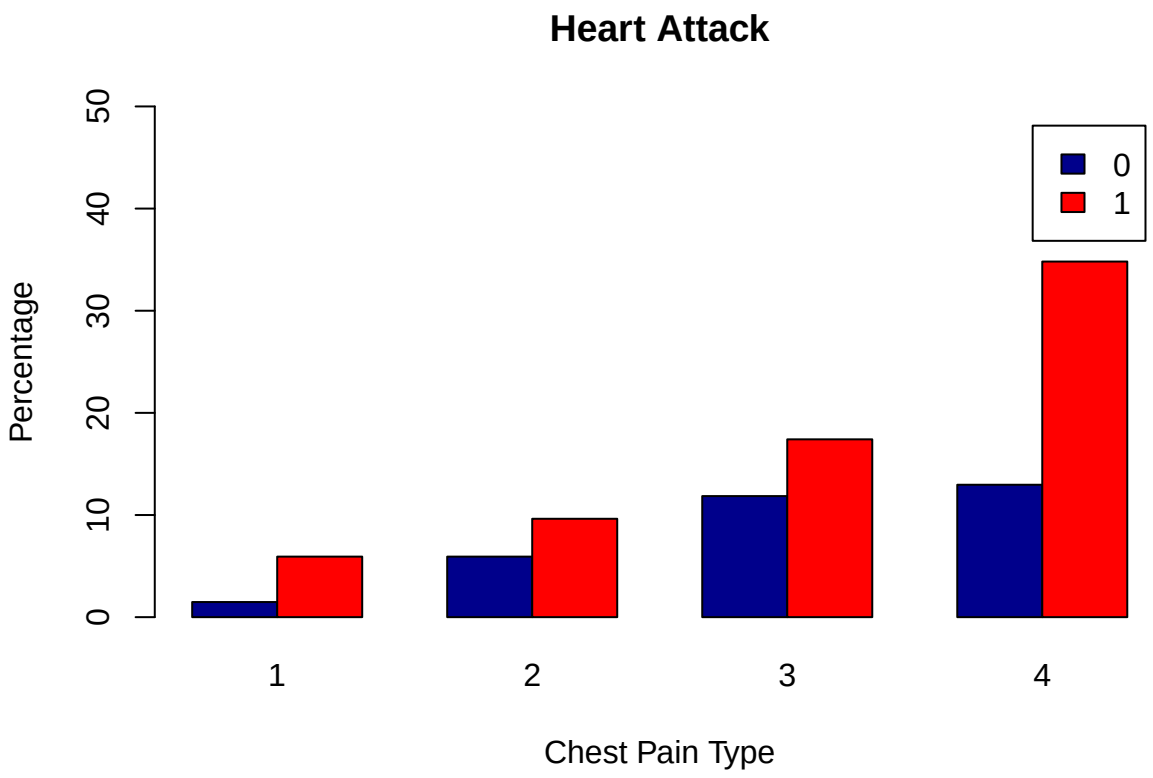
```
barplot(prop.table(Gender_Chest_Pain_Type)*100, main="Heart Attack", xlab="Chest Pain Type",ylab="Perce
```




```
Gender_Chest_Pain_Type <- table(HeartAttack$Sex,HeartAttack$Chest_Pain_Type)
prop.table(Gender_Chest_Pain_Type )
```

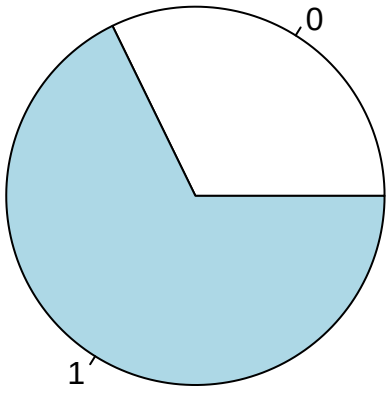
```
##
##              1          2          3          4
##   0 0.01481481 0.05925926 0.11851852 0.12962963
##   1 0.05925926 0.09629630 0.17407407 0.34814815
```

```
barplot(prop.table(Gender_Chest_Pain_Type)*100, main="Heart Attack", xlab="Chest Pain Type",ylab="Perce
```



It can be clearly seen that approximately 6% of the individuals experienced Type 1 pain. Out of these 1% were women and 5% were men

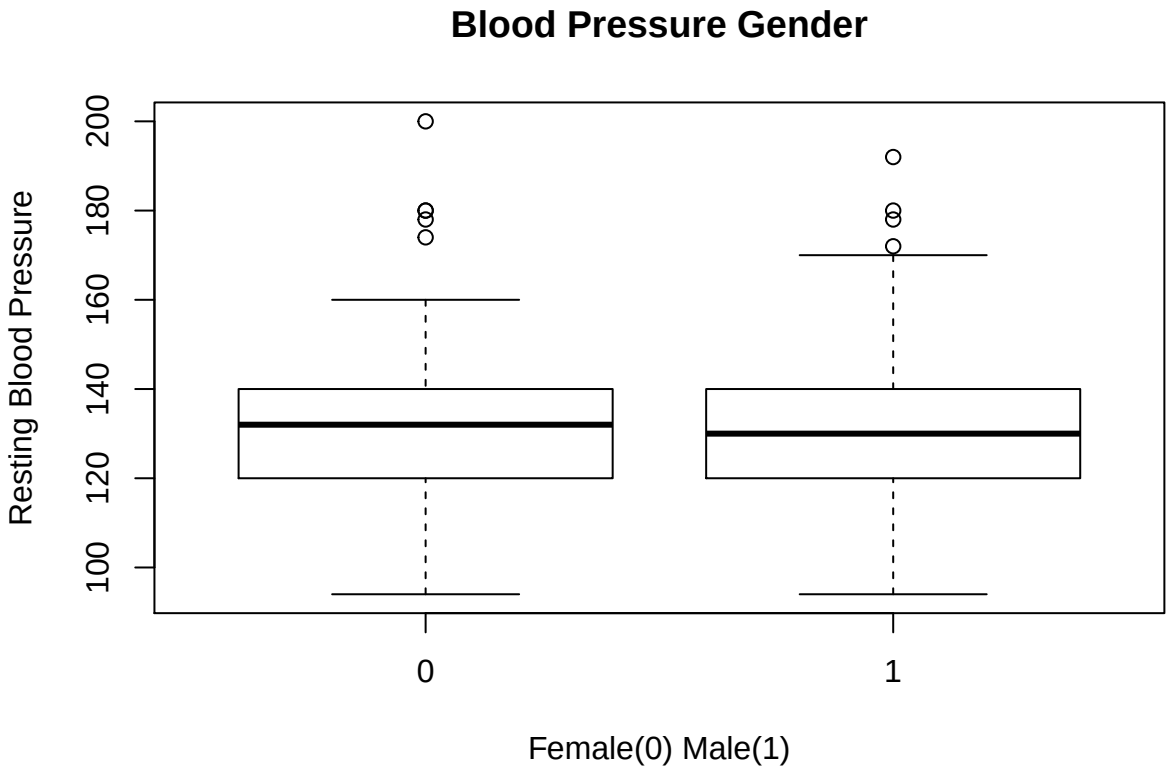
```
pie(prop.table (Gender)*100)
```



QUANTITATIVE GROUPED BY CATEGORICAL

Boxplots provides the graphical display of Resting Blood Pressure grouped by Gender

```
boxplot(Resting_Blood_Pressure~Sex,data=HeartAttack, main="Blood Pressure Gender",
xlab="Female(0) Male(1)", ylab="Resting Blood Pressure")
```



It can be clearly seen that overall Female have a slightly higher median Blood pressure. The lower 25% of both males and females have the same variability/range of blood pressures but for the upper most 25% of the data males have a higher variability as compared to females. They both have around 4 outliers and female have the most extreme outlier. Both distributions are slightly right skewed.

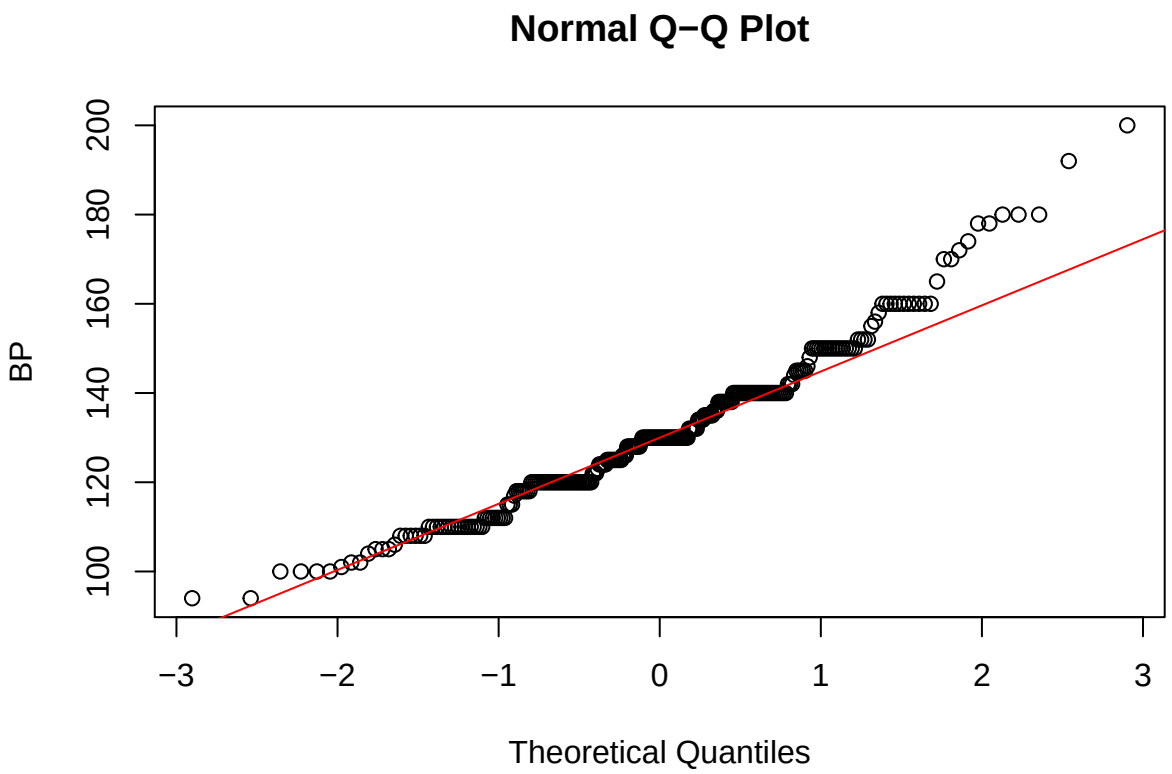
Quantile Quantile plots used to check the validity of the distributional shape of a data set. If the dataset follows a theoretical distribution then it will follow the straight diagonal line y=x.

If the data is curved like a C facing upwards close to and cutting the line the line y=x then it is right skewed (maybe will fit lognormal distribution) wheras if it is C facing downwards close to and cutting the line it is left skewed. If it follows the line it is normally distributed.

If the data is s shaped about the line y=x and has short tails ie few points deviate away from the tails and heads then it is not normal

If the data is s shaped about the line y=x and has long tails ie a large number of points deviate away from the tails and heads then it is not normal

```
qqnorm(HeartAttack$Resting_Blood_Pressure, ylab = "BP");qqline(HeartAttack$Resting_Blood_Pressure,col=2)
```



The distribution is clearly right skewed as the right tail deviates from the line on the right