QUANTITATIVE DATA ANALYSIS

[Document subtitle]

Noor Mohammad Atapoor (ID: 2590537)

NOVEMBER 13, 2023

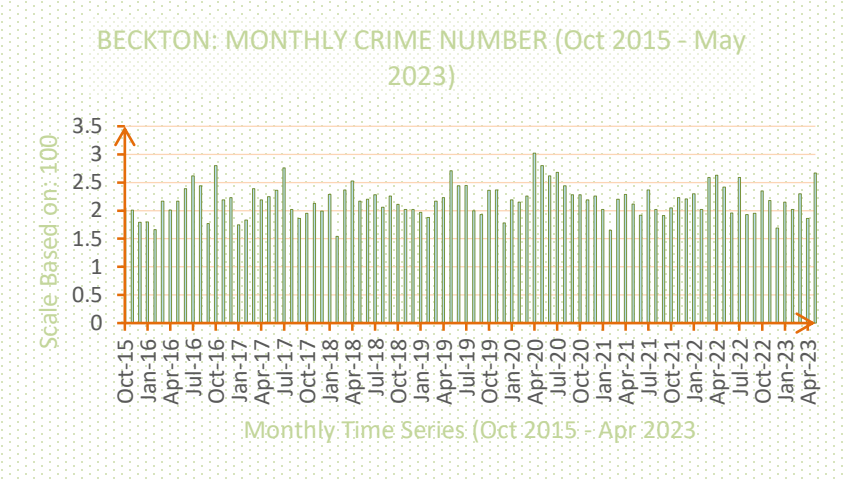
UNIVERSITY OF EAST LONDON, SCHOOL OF COMPUTRING AND ENGINEERING
Dockland Campus, London, England

Table of Contents

1 Session-01: Trend, Distribution, Standardization, and Indexing

1.1 Task-01: Beckton Monthly Crime Graph

The graph shows that there is some variability in the number of crimes in each specific month in different years, but in some specific months the number of crimes in each almost remains constant. For example, the month January show the lowest number of crimes each year with low variability in the time series; other months does not show a pattern, in each year different number of crimes occurs in them.

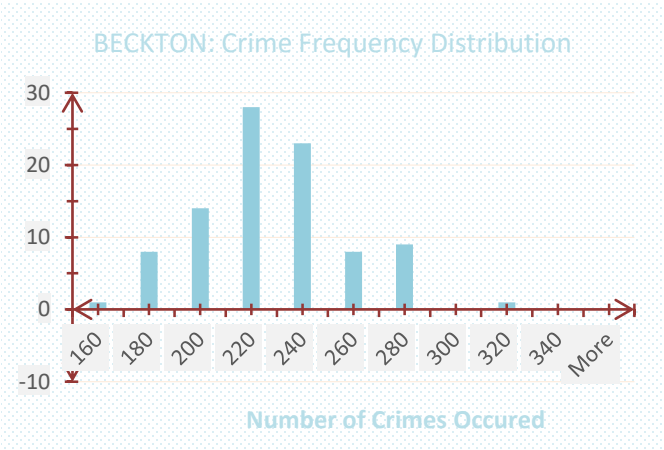


1.2 Task-03: Beckton Crime Distribution, and Summary Statistics

In this part the statistical summary of crimes in the Beckton has been calculated and presented in the table. The distribution of Beckton crimes also presented visually by using a histogram.

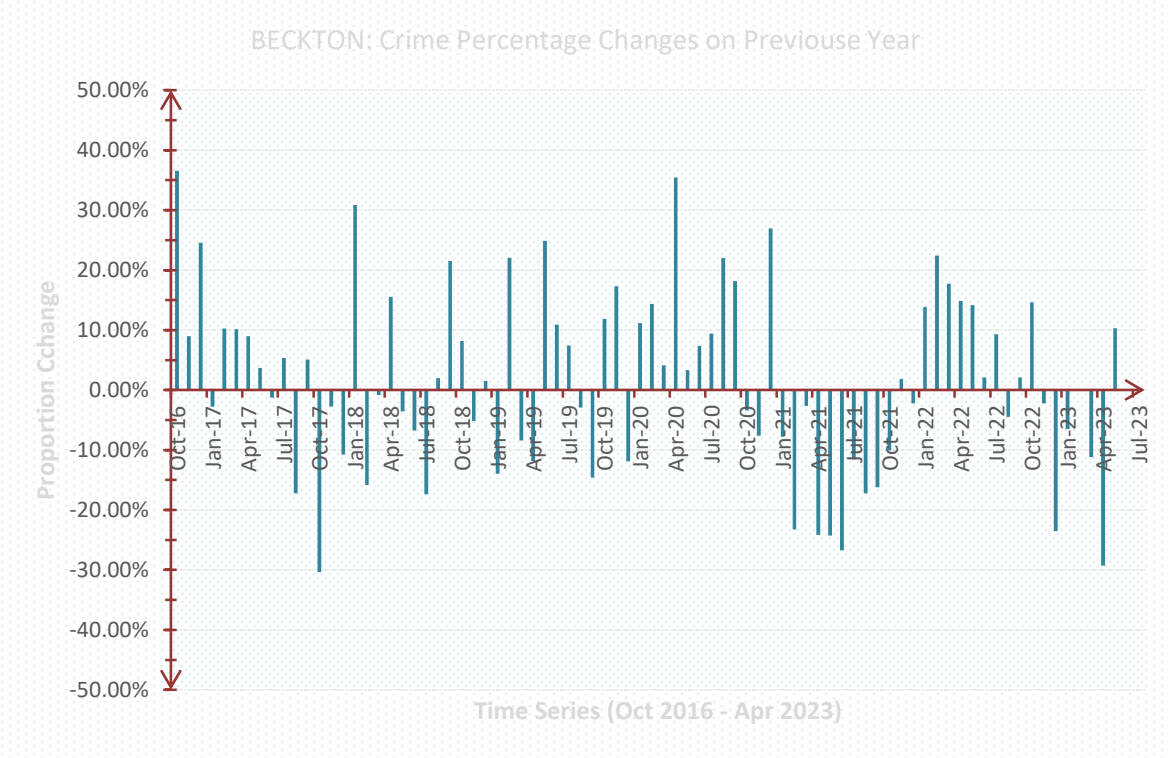
The crimes distribution is normal as clearly can be seen from the mean and median on the summary table. The histogram also shows a normal distribution as the mean point is in the middle and other data frequency are almost equally around the mean.

SUMMARY STATISTICS AND CENTRAL TENDENCY MEASURES		
Minimum:	154	*= MIN(B2:B93)*
Maximum:	302	*= MAX(B2:B93)*
Range:	148	*= MAX – MIN: B97 – B96)*
Average (Mean):	219.011	*= AVERAGE(B2:B93)*
Standard Deviation:	28.9363	*= STDEV(B2:B93)*
Median:	219	*= MEDIAN(B2:B93)*
Mode:	202	*= MODE(B2:B93)*



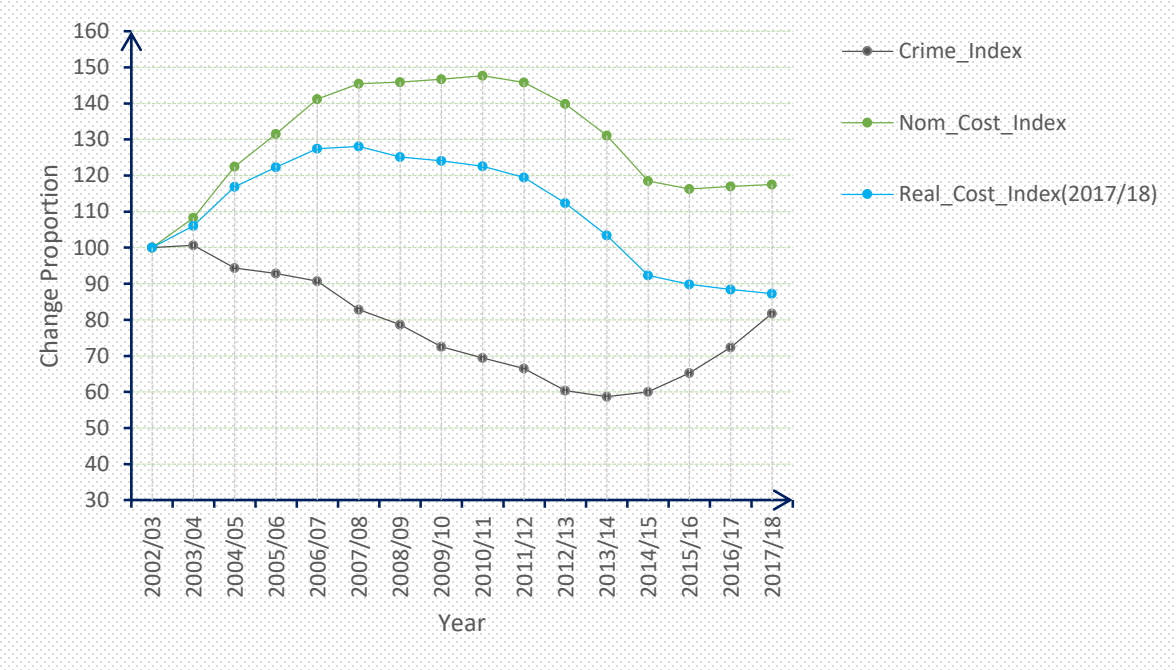
1.3 **Task-04:** Beckton's Crime Percentage Changes on Previous Years

In this part, the number of monthly crimes in a time-series have been presented in terms of percentage of corresponding previous months. Presenting the data in terms of percentage give much more clear vision about their actual changes. In the Task-01 the number of crimes between April 2020, 2021 and 2022 are not so different; but, in their percentage presentation, they are totally different, and give a different insight. For example, among the three consecutive years, April 2020 shows the highest number of crimes, while in the April 2021 the number of crimes are in its lowest level. Therefore, to get an actual insight about the trend of data, its presentation in terms of percentage is much more useful and reliable.



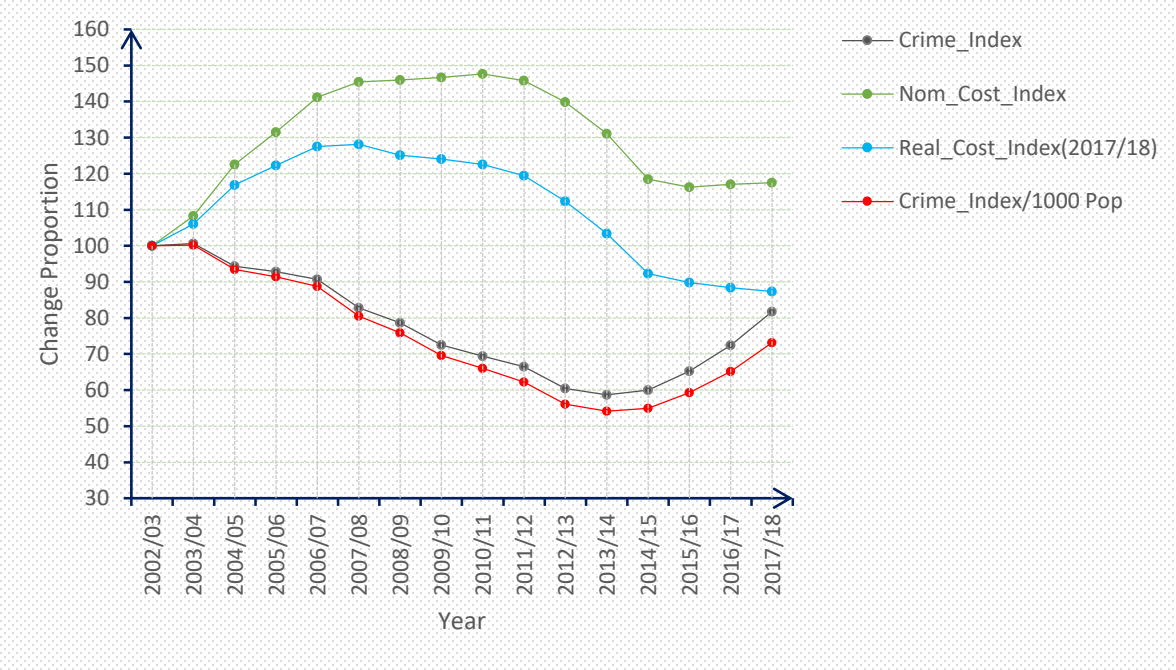
1.4 **Task-05:** Standardizing Crime by Indexing

One way to get an realistic insight from the data is to standerdize it by indexing. The following graphs indicate the nominal costs, real cost, and index of the crimes dueing years 2002 to 2018. As the graph shows, there are remarkable differences between the nominal cost and real cost which has been calculated by indexing. The crime index (balck line) shows that from 2002 to 2014, the crime number are continuesly decreasing proportionally, and accordingly the crime real costs are also prportionally decreasing in a rapid pace compared to the crime nominal costs. From 2015 to 2018, despite the cirme number are increasing, but the real cost are decreasing which is totally different from the nominal cost showing increas considering the mentioed time period. It means the apparent costs of the crimes due to inflation can be deciptive if are not calculated through indexing, and we cannot take a reali insight about the trend of the change.



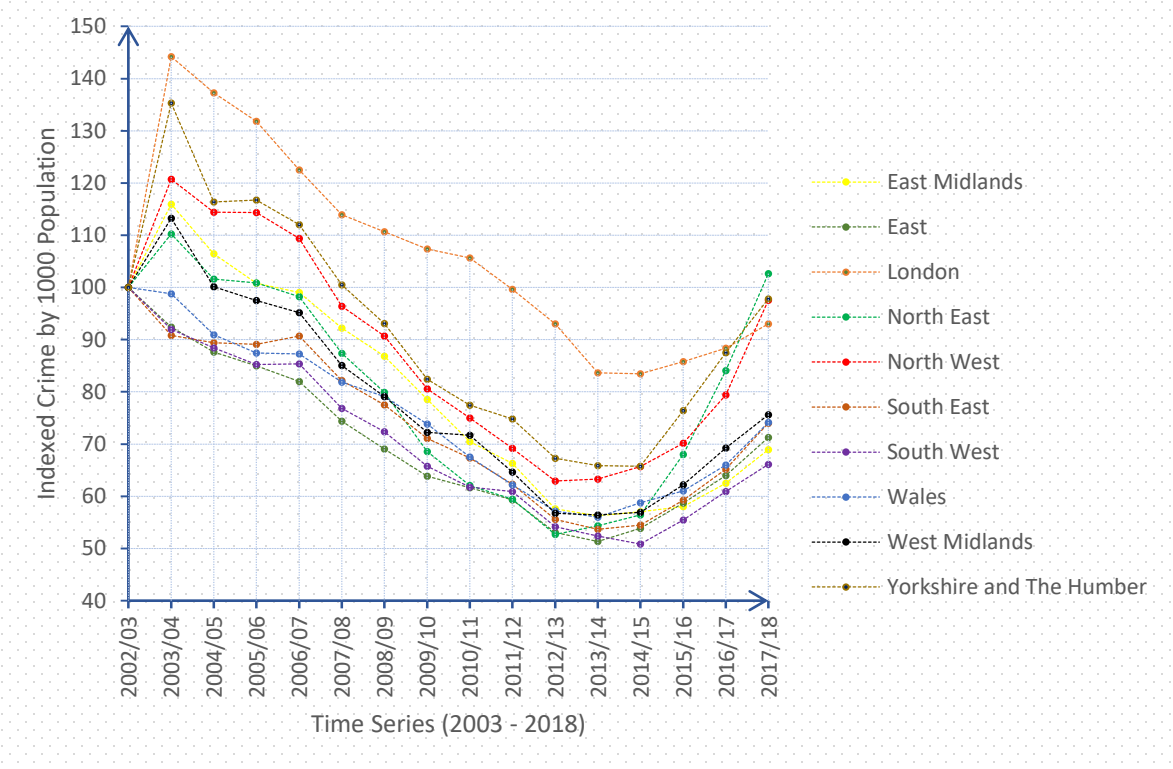
1.5 **Task-06:** Standardizing Crimes by Population

The crime index presented in terms of population shows a different change path as indicated in the following graph. It indicates (read line) that if the crime is presented in terms of population, the number of crimes have more decreased.



1.6 **Task-07:** Indexing Crimes Per Thousand Population

The following graph shows the crime change trends in terms of their index based on population in a time series. Based on this presentation, in 2003/04, the proportion of the crime have been highest in London, and lowest in South East. Although there are crime number proportionally different between different region in the time series, but the overall the crimes have proportionally been in decreasing till 2014. From 2014 onward, the crime proportionally increasing where in 2018 it is the North East region that shows the highest proportion of crime instead of London. One of the reason in proportion of crime increment might be due to influx of migrants in the country.



2 Session-03: Database

2.1 Task-04: Extract Data from SOIL_ID Table

In this task, data is extracted from all attributes of tables SOIL_ID based on the condition of the value of *'ph-val'* attribute is bigger than five, and finally is grouped by the *'ph_val'* attribute. By executing one of the following query, we will get the result presented in the right hand table:

Soil_id	Soil_type	Ph-val	Sur_id	assessed
213	32	5.12	1	1991-12-24
222	26	5.13	3	1990-05-19
225	17	5.34	4	1992-07-04
141	17	5.43	3	1990-11-22
212	27	5.44	3	1992-07-04
245	21	5.74	4	1990-11-06
193	17	6.44	4	1990-05-19
170	30	6.54	4	1989-09-12
179	17	6.54	1	1992-09-17

```
/* TASK-04: We can write the query in different form as following: */


SELECT SOIL_ID.soil_id, SOIL_ID.soil_type, SOIL_ID.ph_val, SOIL_ID.sur_id, SOIL_ID.assessed
    FROM SOIL_ID WHERE SOIL_ID. ph_val > 5 ORDER BY SOIL_ID.ph_val;

                    /* OR */

SELECT SO.soil_id, SO.soil_type, SO.ph_val, SO.sur_id, SO.assessed
    FROM SOIL_ID as SO WHERE SO. ph_val > 5 ORDER BY SO.ph_val;

                    /* OR */

SELECT * FROM SOIL_ID AS SO WHERE SO.ph_val > 5 ORDER BY SO.ph_val;
```

2.2 Task-05: Extract Data from SOIL_ID and SURVEYOR Tables

In this task, data is extracted from two tables, SOIL_ID and SURVEYOR, as they are joined based on their '*sur_id'* attributes. The values from fields *'soil_name', 'series', and 'ph_val'* are filtered based on the surveyor '*surveyor = N.Brown '*, and finally is displayed in ascendant form. By executing one of the following query, we will get the result presented in the right hand table:

```
/* TASK-05: Query for the Task 05, Session 03 */


SELECT SOIL_ID.soil_id, SOIL_ID.ph_val, SOIL_ID.assessed, SURVEYOR.surveyor
    FROM SOIL_ID JOIN SURVEYOR ON SOIL_ID.sur_id = SURVEYOR.sur_id
        WHERE SURVEYOR.surveyor = 'N.Brown' ORDER BY SOIL_ID.ph_val DESC

                /* OR  */

SELECT SO.soil_id,SO.ph_val, SO.assessed, SU.surveyor
    FROM SOIL_ID AS SO  JOIN SURVEYOR AS SU ON SO.sur_id = SU.sur_id
        WHERE SU.surveyor = 'N.Brown' ORDER BY SO.ph_val DESC
```

Soil_id	Ph_val	Assessed	surveyor
216	4.73	1990-11-06	N.Brown
274	4.56	1990-05-19	N.Brown
194	4.52	1990-11-06	N.Brown
232	4.32	1990-05-19	N.Brown
254	4.32	1991-12-24	N.Brown
182	3.34	1991-12-24	N.Brown
261	3.29	1992-07-04	N.Brown
157	2.36	1989-07-12	N.Brown

2.3 **Task-06:** Extract Data from SOIL_TYPE and SOIL_ID Tables

In this task, data is extracted from two tables, SOIL_TYPE and SOIL_ID, as they are joined based on their '*soil_name'* attributes. The values from fields *'soil_name', 'series', and 'ph_val'* are filtered based on the condition '*series = 3*', and finally is displayed in ascendant form. By executing one of the following query, we will get the result presented in the right hand table:

```
/* Task-06: SQl Query fo the Taks 06, Session 03 */

SELECT SOIL_TYPE.soil_name, SOIL_TYPE.series, SOIL_ID.ph_val
    FROM SOIL_TYPE JOIN SOIL_ID ON SOIL_TYPE.soil_type = SOIL_ID.soil_type
        WHERE SOIL_TYPE.series = 3 ORDER BY SOIL_TYPE.soil_name ASC

                        /* OR */

SELECT STP.soil_name, STP.series, SO.ph_val
    FROM SOIL_TYPE AS STP JOIN SOIL_ID AS SO ON STP.soil_type = SO.soil_type
        WHERE STP.series = 3 ORDER BY STP.soil_name ASC
```

Soil_name	series	Ph_val
Clay	3	2.9
Clay	3	3.43
Clay	3	3.44
Clay	3	4.52
Clay	3	4.65
Gravel	3	2.36
Gravel	3	3.54
Gravel	3	4.56
Gravel	3	6.54
Sand	3	2.22
Sand	3	3.33

3 **Session-04:** Data Exploration and Graphics

3.1 **Task_01 - Section_02:** Population Aged 16 or Over in London

The function, *creat.boxPlot(ds)*, was defined to generally create a box plot for a given data set as its argument. The function not only creat box plot, also draws values of quartiles, outliers and their row numbers on the given data set. Finally by calling the function, it is executed and draw the box plot for the given data set (*KS13N.London$P16plus*).

```
# define a function creating boxplot using a data set as its argument;
creat.boxPlot <- function(data_set){
  op <- par(mar = c(5, 8, 4, 2) + 0.1)
  textLabel = c("Min(Q0):", "Q1:", "Median(Q2):", "Q3:", "Max(Q4):")
  boxData <- boxplot(data_set, xlab = "London Population Aged 16 or Over",
                     ylab = "Number of Population (16+)", col = "beige", border = "cyan4"
                     ,varwidth = TRUE, whisklty = 3,frame.plot = FALSE,cex.axis = 0.5, las = 1)

  for( i in 1:length(boxData$stats)){
    if(i <= length(boxData$group)){
      text(boxData$group[i], boxData$out[i],
           paste("Outlier:[ Row:", which(data_set == boxData$out[i]),"; Data point:",
                 boxData$out[i],"]"), pos = 4, cex = 0.6, col = "brown4")
    }
    text(boxData$group, boxData$stats[i] + 220, paste(textLabel[i], boxData$stats[i]),
         pos = 4, cex = 0.6, col = "darkgoldenrod")
  }
  par(op)
}
# The function is called by providing with the 'KS13N.London$P16plus' data set as its argument
creat.boxPlot(KS13N.London$P16plus)
```

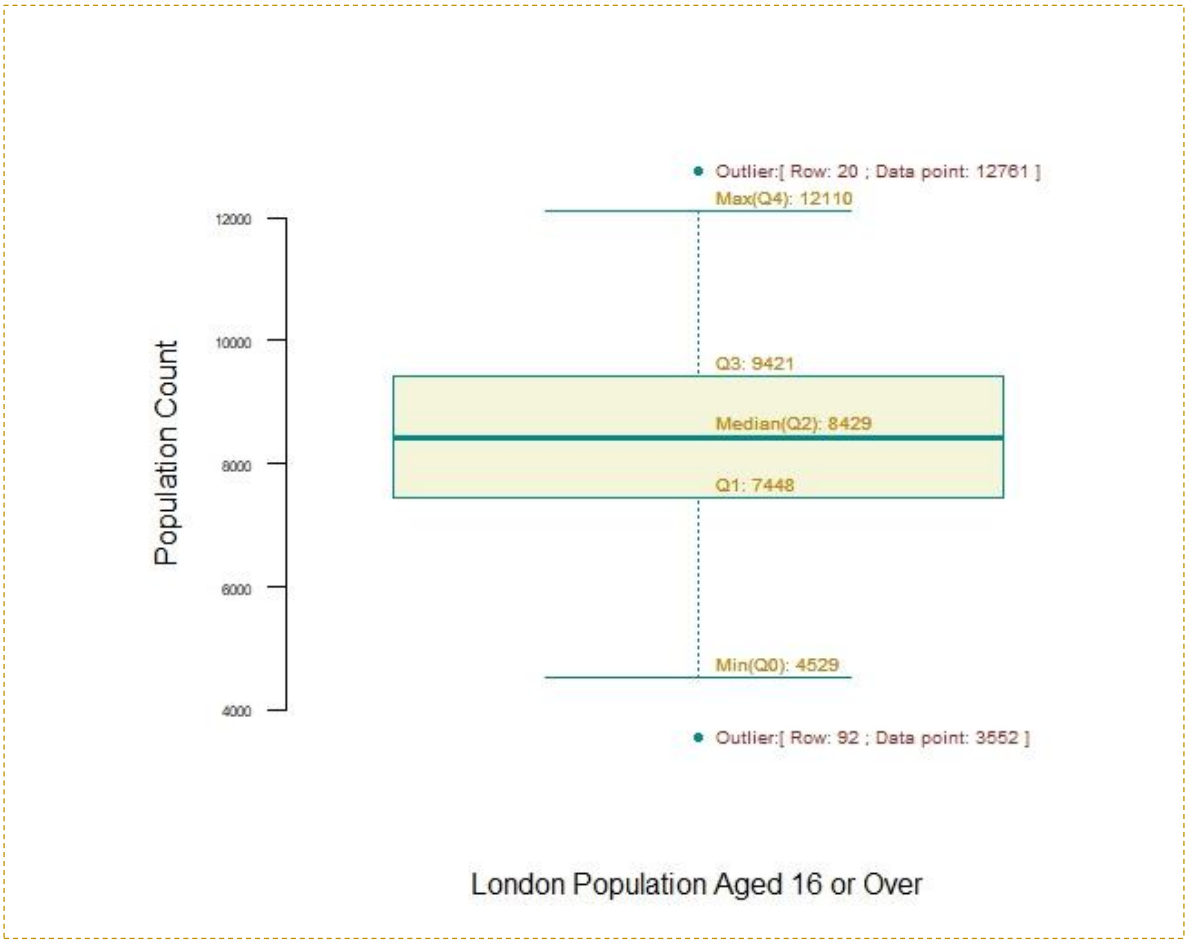
The box plot that has been drawn, present the following points about the statistical characteristics of the KS13N.London$P16plus data set:

```
KS13N.London[92,]
KS13N.London[20,]
```

```
   Region   Sub_region Borough   Code   Ward P16plus
92 London Outer London Bromley 00AFGQ Darwin    3552
```

```
   Region   Sub_region Borough   Code        Ward P16plus
20 London Outer London  Barnet 00ACFZ Childs Hill   12761
```

- Overall, the london population aged 16 or over are normally distirbuted in Wards;
- There are two data points (3552 and 12761) were detected as lower and uper outliers respectively;
- The lower and uper outliers correspond to the rows number 92 and 20 respectively;
- The subsequent assessment of the data set as shown in the above boxes, revealed that both outliers belongs to the Outer London sub-region; the lower outlier belongs to the Darwing ward in Bromely borough, and uper outlier belongs to Childs Hill ward in Barnet borough;



3.2 **Task_01- Section_06:** Create Box Plots for 6 Qualifications Percentage

Utilizing the function *within ()*, as shown below, after calculating the percentage of population in each educational qualification we can add them as new variables in the KS13N.London data set. To make work easy, we can use the *attach()* function to make all the data set variable directly accessible.

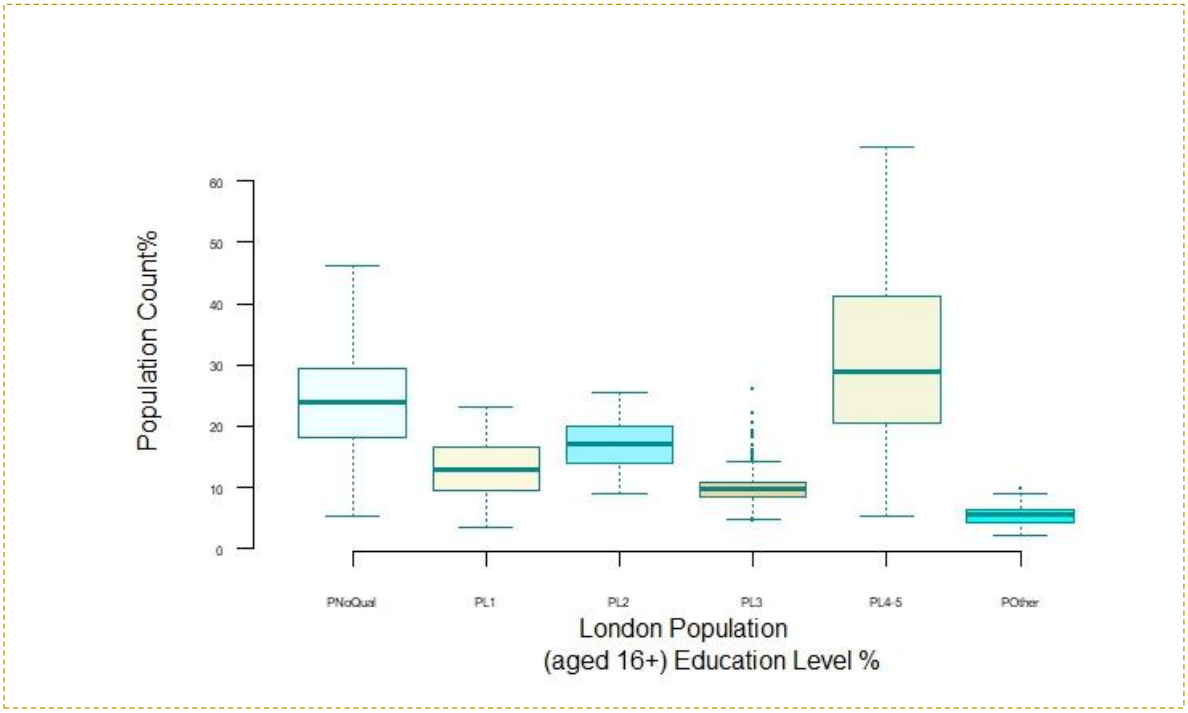
```
# Create percentage of population (p16plus) variables
KS13N.London <- within (KS13N.London, pNoQual <- (NoQual / P16plus)*100)
KS13N.London <- within (KS13N.London, pLevel1 <- (Level1 / P16plus)*100)
KS13N.London <- within (KS13N.London, pLevel2 <- (Level2 / P16plus)*100)
KS13N.London <- within (KS13N.London, pLevel3 <- (Level3 / P16plus)*100)
KS13N.London <- within (KS13N.London, pLevel4_5 <- (Level4_5 / P16plus)*100)
KS13N.London <- within (KS13N.London, pOther <- (Other / P16plus)*100)
attach (KS13N.London)
```

The following code snip draw box plots for the percentages of the all educational qualifications.

```
# Create a boxplot of all six percentage variables
bg <- c("azure", "cornsilk","cadetblue1", "burlywood1","beige","cyan")
boxplot(pNoQual, pLevel1, pLevel2, pLevel3, pLevel4_5, pOther, xlab = "London Population
        (aged 16+) Education Level %", ylab = "Population Count%",
        col = bg, border = "cyan4", varwidth = TRUE, outcex = .5,outpch = 20,
        whisklty = 3,frame.plot = FALSE,cex.axis = 0.5, las = 1,
        names=c("PNoQual", "PL1", "PL2", "PL3", "PL4-5", "POther")
        )
```

After executing the above code piece, the following box plots are created. As we see, there are some difference between the distributions of educational qualifications and their percentages:

- Qualifications such as PNoQual, PL1, PL2 distributions are more normalized compared to their non-percentage counterparts;
- The PL2 has no outlier as it was in the L2; On the POthers outlier appeared, and it skewed negatively, its counterparts, Others, is more normalized and has no outlier.



3.3 **Task_01 - Section_06:** Box Plot of No-Qualification Percentage by the London Sub-Region

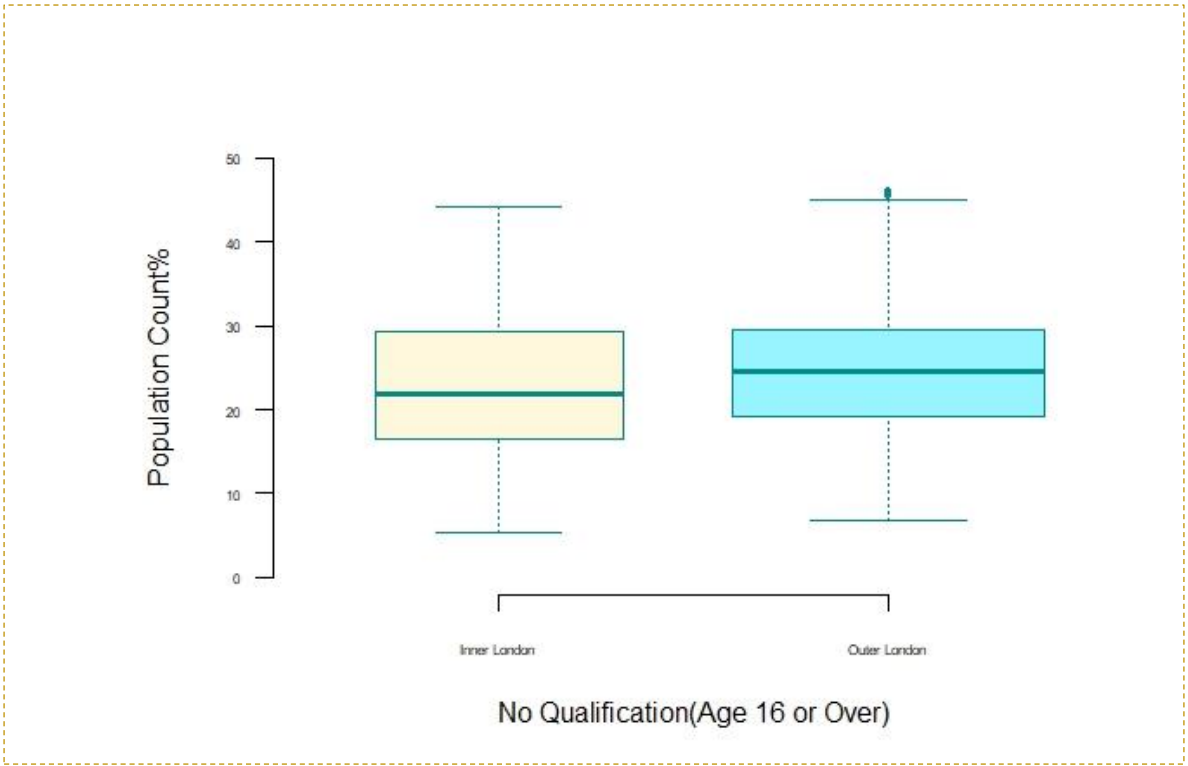
The following code snip creates box plots for the percentage of population having no-qualification by the London Sub-region:

```
# Create a boxplot of pNoQual by factor Sub_region
bg <- c("cornsilk","cadetblue1")
boxplot(pNoQual ~ Sub_region, xlab = "No Qualification(Age 16 or Over)", ylab = "Population Count%",
        col = bg, border = "cyan4", varwidth = TRUE, outcex = .8,outpch = 20, whisklty = 3,
        frame.plot = FALSE,cex.axis = 0.5, las = 1, names=c("Inner London","Outer London" ),ylim=c(0,50)
       )
```

Executing the above code piece, the following box plots are created for the percentage of population with no qualification in the London's sub-regions. Looking at the box plots, we can understand the following points about the statistical characteristics of the percentage of population distributions in to sub-regions of the London:

- The percentage of population having no-qualification is higher in Outer London than Inner London;
- Although, there are a number of outlier in the Outer London, the percentage of population having no-qualification are more normally distributed compared to the Inner London;

- The distribution in Inner London shows positive skewness which such case is not remarkable in the Outer London;



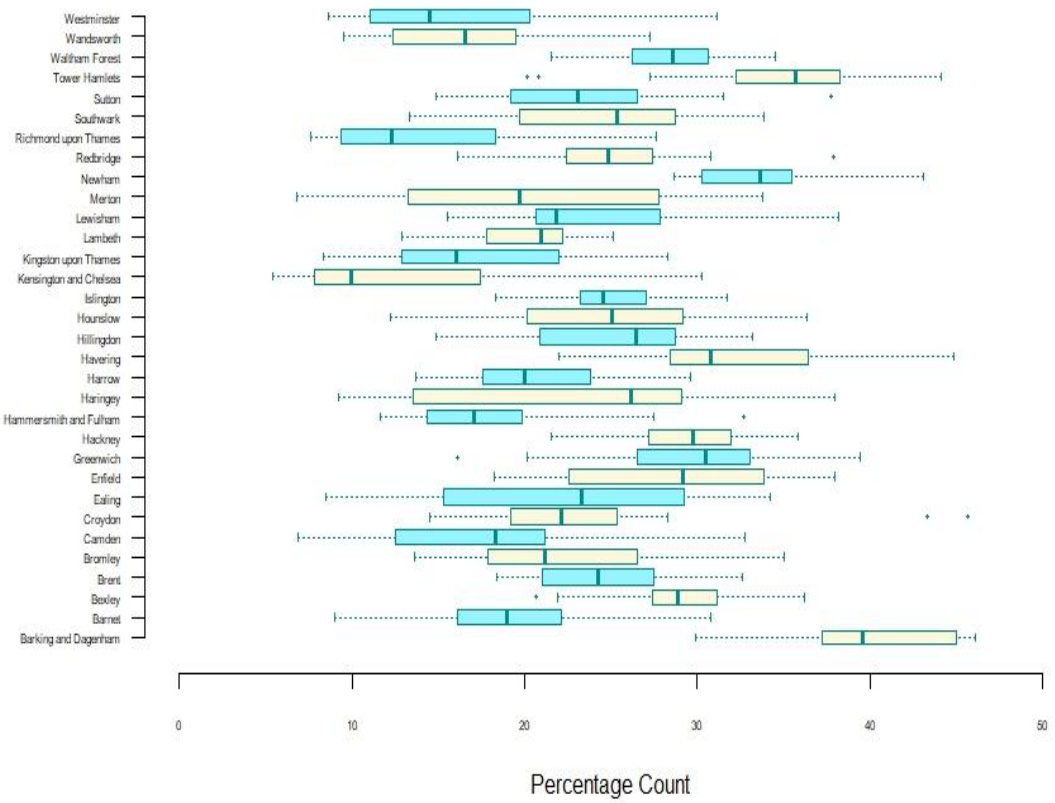
3.4 **Task_01-Section_06:** Box Plot for Percentage of No-Qualification by Borough

The following code snip creates box plots for the percentage of population having no qualification by London's Borough:

```
# Create a boxplot of pNoQual by factor of Borough
bg <- c("cornsilk","cadetblue1")
op <- par(mar = c(5, 8, 4, 2) + 0.1)
boxplot(pNoQual ~ Borough, xlab = "Percentage Count", ylab = "",col = bg, border = "cyan4",
        varwidth = TRUE, outcex = .6,outpch = 20,whisklty = 3,frame.plot = FALSE,cex.axis = 0.5,
        las = 1, horizontal = TRUE, ylim = c(0,50)
        )
par(op)
```

Executing the above code piece, the following box plots are drawn for all the population having no qualification and living in each London's Boroughs. The box plots tell us the following points on the distribution of population in London's Boroughs:

- The populations having no qualification are not distributed equally, and there are high distribution variabilities among the Boroughs;
- Within each borough, also the populations are not distributed normally, and there can be seen data dispersion, skewness, and one sided tails. Some of them have a number of high extreme values or outliers.

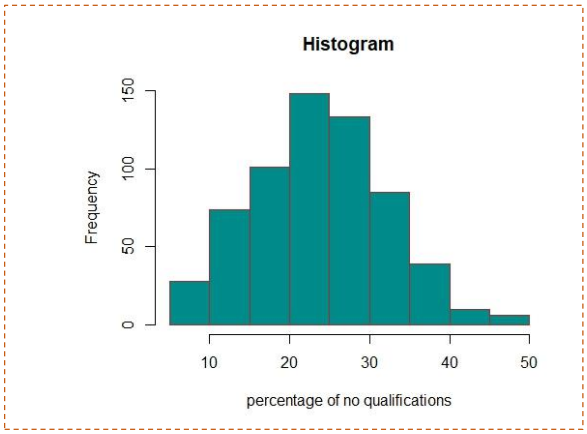


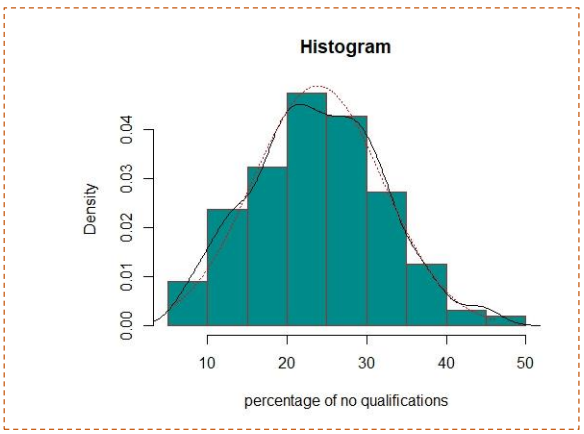
3.5 Task-01(Section_07): Frequency and Probability Density Histogram

The frequency histogram shows there are 150 wards that 20 to 25 percent of their population aged 16 or over have no qualifications; and 140 wards that 25 to 30 percent of their population aged 16 or over have no qualification. Accordingly, there are about 10 wards that their 45 to 50 percent of their population aged 16 or over have no qualification. Although, the distribution has been skewed a little to the right, but overall it has normally distributed; and the probability density line is somehow different from the normal distribution line.

```
op <- par(mar = c(5, 8, 4, 2) + 0.1)
hist(pNoQual, col = "cyan4", border = "brown", freq = T,ylim = c(0,150),
     xlab = "percentage of no qualifications", main = "Histogram")
```

```
op <- par(mar = c(5, 8, 4, 2) + 0.1)
hist(pNoQual, col = "cyan4", border = "brown", freq = F,
     xlab = "percentage of no qualifications", main = "Histogram")
lines (density(sort(pNoQual)))
xfit <- seq(from = min(pNoQual), to = max(pNoQual), by = 0.1)
yfit = dnorm(xfit, mean(pNoQual), sd(pNoQual))
lines(xfit, yfit, lty = "dotted", col = "red", cex = 0.8)
```

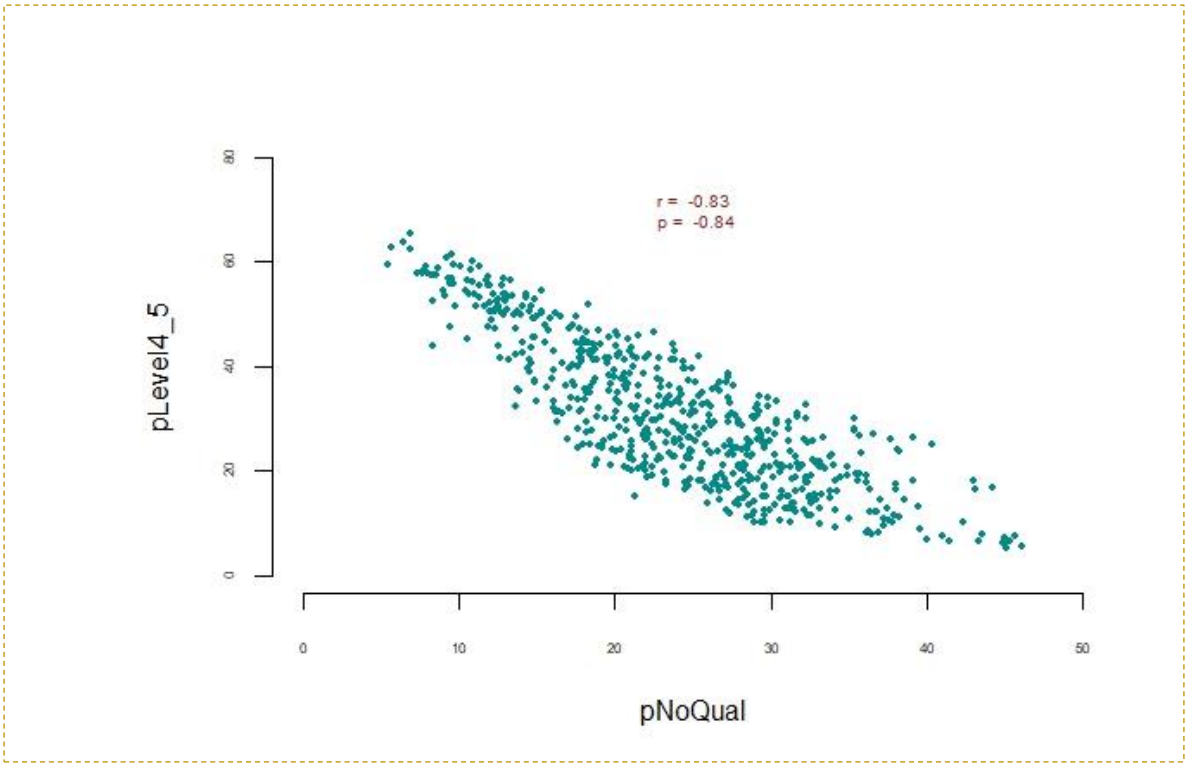




3.6 Task-01(Section-08): Scatter Plot of Percentage of No Qualification against Qualification Level 4 & 5

The following code snip create scatter plot of pNoQual against pLevel4&5, and also calculate the Pearson (p) and Spearman(r) correlation coefficient and labels them on the graph as shown below. The size of both correlation coefficients are negatively huge; it means that the both variables have negatively correlated; if the number of proportion of people education in level 4 and 5 increases then the proportion of people with no qualification drastically are decreased.

```
op <- par(mar = c(5, 8, 4, 2) + 0.1)
plot(pNoQual, pLevel4_5, pch = 16,xlim = c(0,50),ylim = c(0,80),cex.axis = 0.5, col = "cyan4",
   frame.plot = FALSE, cex = 0.5)
r <- cor(pNoQual, pLevel4_5, method = "spearman")
p <- cor(pNoQual, pLevel4_5, method = "pearson")
text(25,70,paste(" r = ",round(r,2),"\n p = ",round(p,2)),cex = 0.6, col="brown4")
```



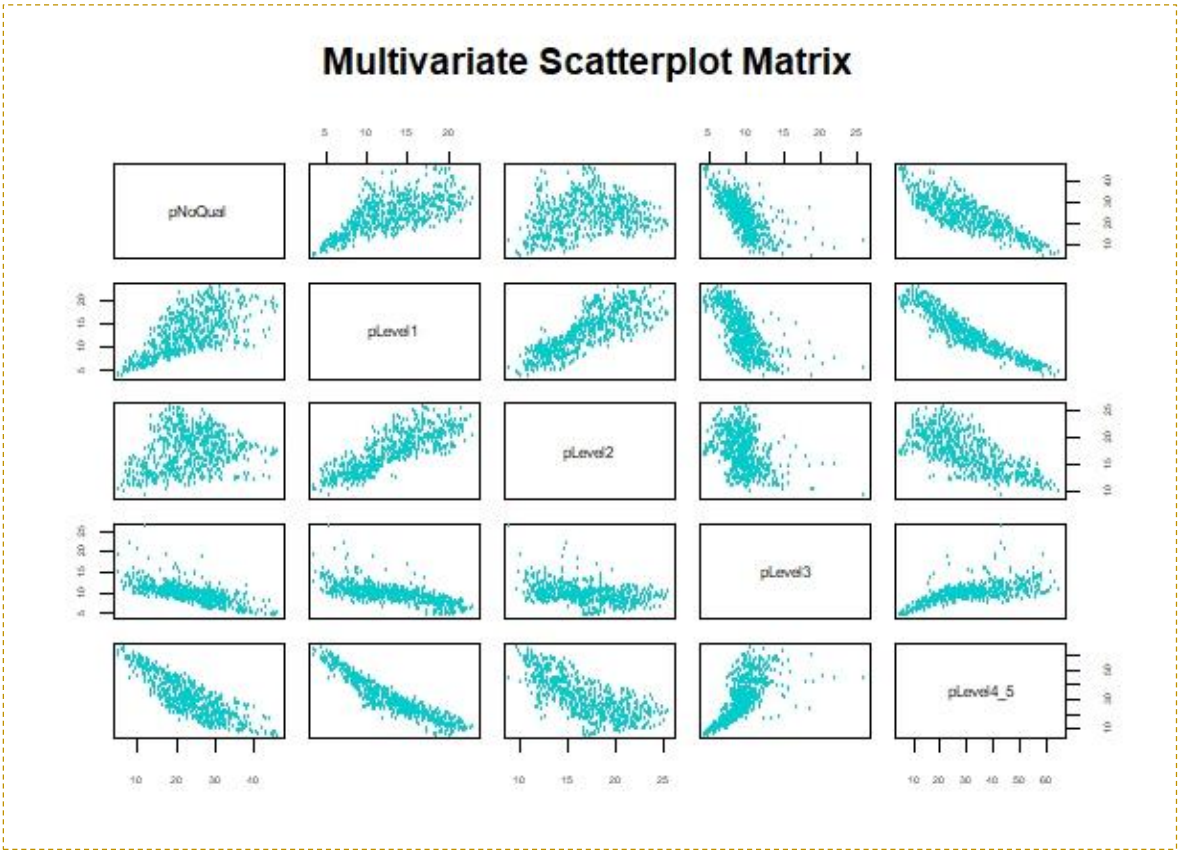
3.7 Task-01(Section-10): Multivariate Scatter Plot for the Various Qualification Levels

The scatter plots and the correlation matrix indicates that strongest positive correlations are between plevel 1 and 2. It means that the increase of one causes the increase of others and vice versa. The highest negative correlation is between pLevel 1 and 4_5. It means the increase of one causes the rapid decreases of the other. However, all the relations between percentage qualification can be investigated and analyzed from the correlation matrix and multivariate scatter plots depicted here.

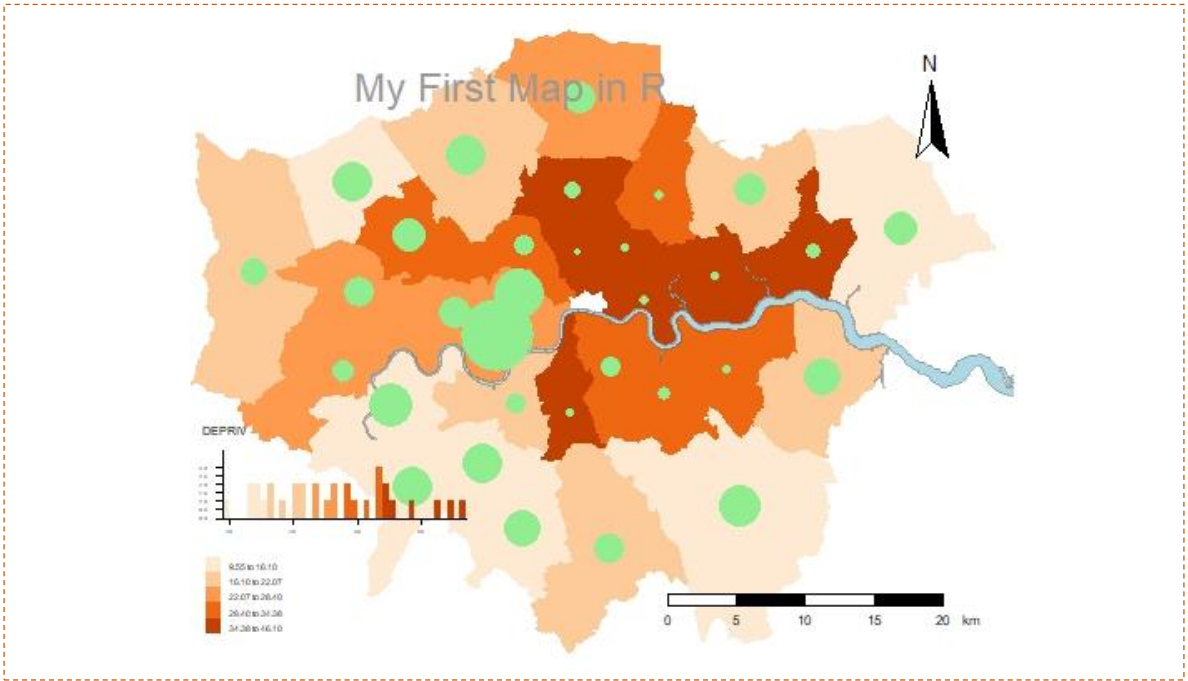
```
pairs(~ pNoQual + pLevel1 + pLevel2 + pLevel3 + pLevel4_5,  data = KS13N.London,
      main = "Multivariate Scatterplot Matrix",cex.labels = 0.8,pch = 1,
      cex = 0.5,cex.axis = 0.5, col = "cyan3")
```

```
# Sub set of KS13N.London data set
Sub.set <- data.frame(pNoQual, pLevel1, pLevel2, pLevel3, pLevel4_5)

# add column names on each data frame variables;
colnames(Sub.set) <- c("PNoQual", "PL1", "PL2", "PL3", "PL4_5")

# Basic correlation matrix on each pairs of variables;
Sub.cor <- cor(Sub.set, method = "spearman")
round(Sub.cor, digits = 2)
```

```
         PNoQual   PL1   PL2   PL3 PL4_5
PNoQual     1.00  0.66  0.26 -0.74 -0.83
PL1         0.66  1.00  0.83 -0.71 -0.95
PL2         0.26  0.83  1.00 -0.39 -0.71
PL3        -0.74 -0.71 -0.39  1.00  0.74
PL4_5      -0.83 -0.95 -0.71  0.74  1.00
```



3.8 **Task-02(Section-04):** Male Life Expectancy and Deprivation



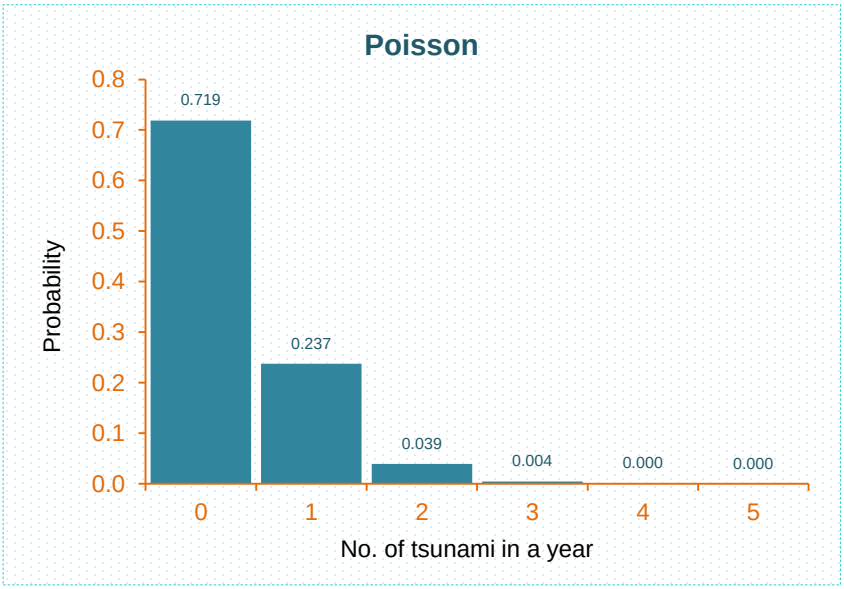
4 SESSION-05: Probability Distribution

4.1 Task-01: Binomial Distribution

Given the probability of boy (0.5) in each pregnancy, the probability that the couple have 7 boys out to 10 pregnancies is 0.117. The most likely outcome for this family is 5 boys in 10 pregnancies. Because, as the following probability distribution shows, the highest probability distribution belongs to frequency 5.

4.2 Task-02: Poison Distribution

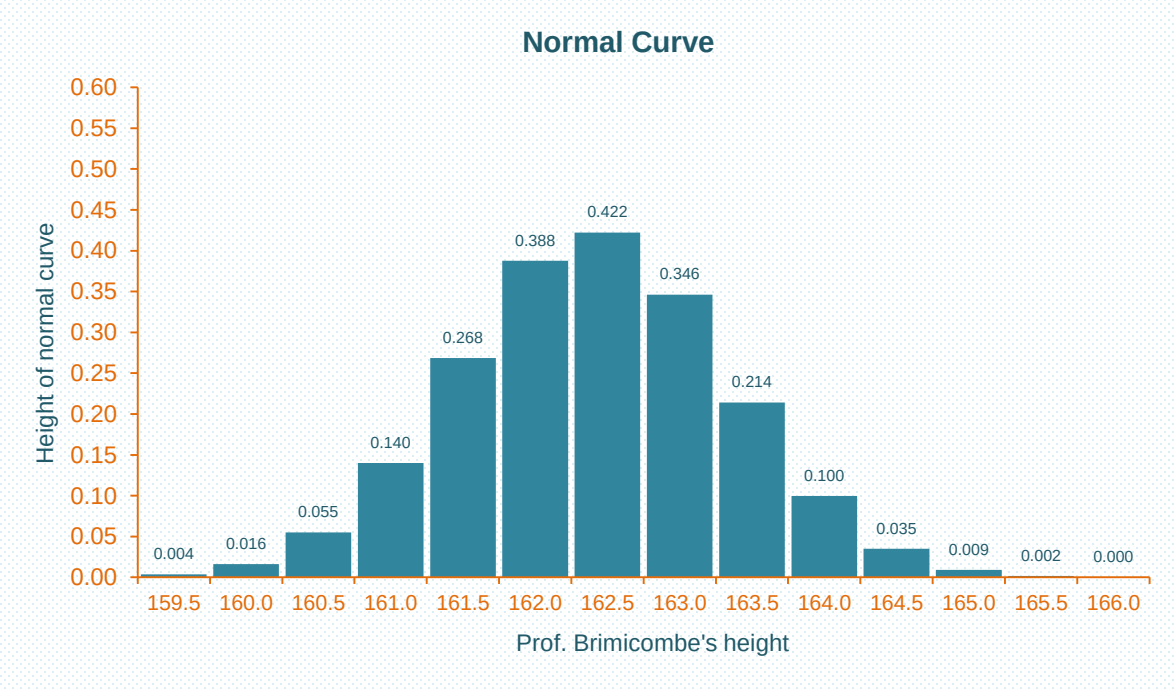
Given one Tsunami in each three years, the probability of one Tsunami in a year is 0.237, or 23.7 %. The probability of 0 tsunami is 0.719 which is the highest probability in this scenario. Therefore, Jack can feel himself very safe while going in diving holiday in the pacific.

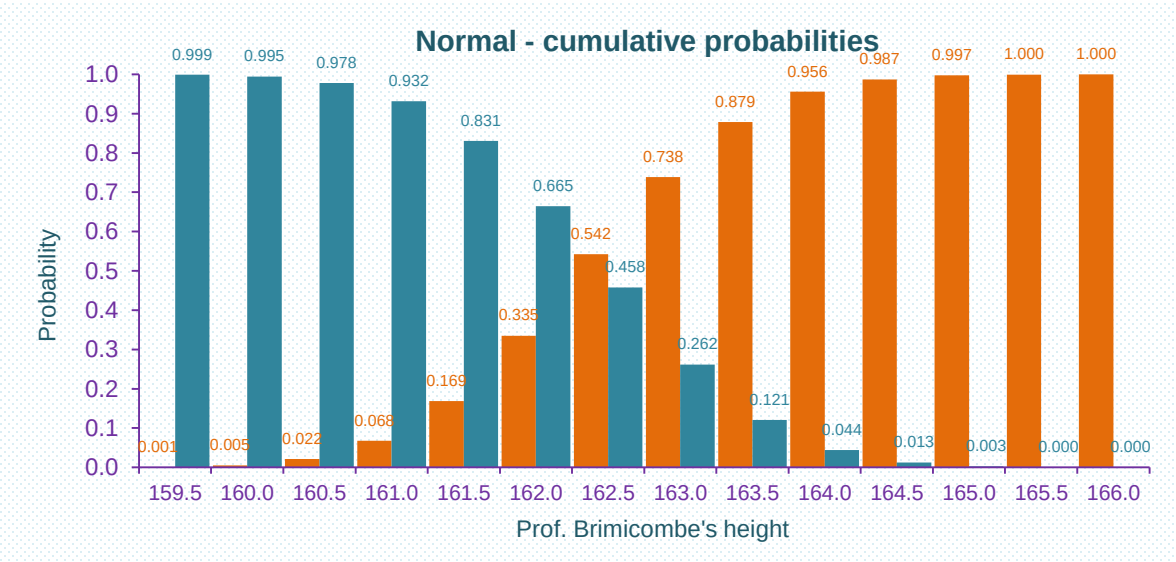


4.3 **Task-03:** Normal Distribution

The probability that the Prof. Brimicombe is taller than 64cm is 0.046; it comes from the cumulative probabilities heights bigger than 64cm. As shown below, considering the case, the cumulative probability is mathematically calculated as following:

$P(height > 64cm) = \sum P(64.5cm) + P(165cm) + P(165.5cm) + P(166cm)$ = **0.046cm**





5 SESSION-06: Hypothesis Testing (Non-Parametric)

I have developed two functions called *getData ( )* and *makeData( )* for accomplishing some initial necessary steps of reading data from the csv files.

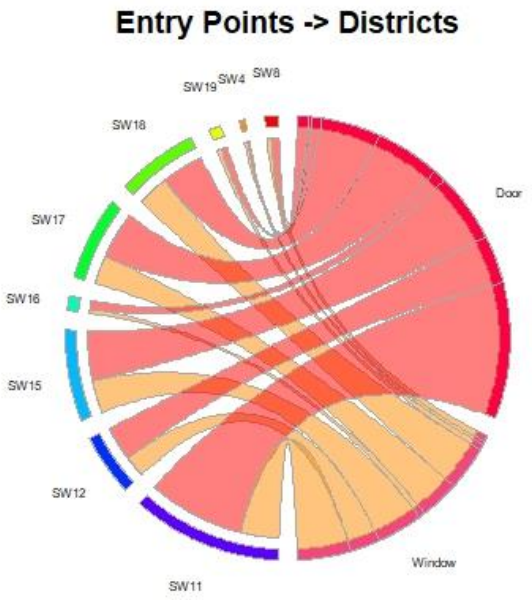
```
# function for getting data address and reading it and returning the result;
getData <- function(dataFile){
  setwd(dirname(file.choose()))
  getwd()
  data_set <- read.csv(dataFile, stringsAsFactors = TRUE)

  return(data_set)

}

# function for structuring the data based on certain rows and columns and returning the result;
makeData <- function(mydata,var1,var2,entry1, entry2 ){
  crosstab <- xtabs(~ var1 + var2, exclude = "", data=mydata)
  datatable <- as.data.frame(cbind(crosstab[,entry1],crosstab[,entry2]))
  names(datatable)[1] <- entry1
  names(datatable)[2] <- entry2

  return(datatable)
}

# two functions are called;
data_set <- getData("6.4- burglary-chi.csv")
myData <- makeData(data_set,data_set$district,data_set$entry_pn,"Door","Window")
```

5.1 Circular Plot: Districts by Entry Points (Door & Window)

By calling and executing the "getData (), and makeData()" functions, now we have the structured data object (myData) in our disposal, and can plot them out by executing the following code snip:

```
# Create circular plot of the districts by burglars entry points (Door, Window);
op <- par(mar = c(5, 8, 4, 2) + 0.1)
PlotCirc(t(as.matrix(myData)), main = "Entry Points -> Districts",cex = 0.5 )
par(op)
```

Executing the above piece of code, we have the following circular plot. Looking at the plot, it is clearly seen that the burglars mostly used the doors as their entry points into houses in each districts. No district can be found on the plot indicating equal proportion of doors and windows as entry points used by the burglars.



5.2 **Chi – Square Test:** Non-Parametric Hypothesis Test

To make sure whether the unequal proportion between doors and windows as entry points for burglars have actually been by chance or it is real, we need to test it. As the data set we are going to run the test is categorical data, thus, we need to use Chi-Square test for examining test hypothesis as defined below:

- **Null Hypothesis:** it is equally likely that a burglar enters house through window or door;

$$H_0: \quad P_D = P_W \qquad (1)$$

- **Alternative Hypothesis:** the probability that burglars enter houses through window or door differs;

$$H_a: \quad P_D \neq P_w \qquad (2)$$

In examining the null hypothesis, the following series of tests have been performed to pave the ground for the interpretation about the validity of the null hypothesis:

```
# Simple chi-squared test on 'myData' data set;
chisq.test(myData)
```

```
        Pearson's Chi-squared test

data:  myData
X-squared = 54.193, df = 8, p-value = 6.333e-09
```

```
# Calculating the critical value for the test;
qchisq(p = .05, df = 8, lower.tail = FALSE)
```

```
[1] 15.50731
```

The Chi-Square test was conducted to test whether the probabilities of

burglars enter houses through doors and windows were identical for all districts. As has shown below, the results were significant ($X^2(8) = 54.193, P < 0.05$), suggesting that burglaries did not select entry points randomly, and most likely they use doors to enter houses; therefore, the null hypothesis is rejected.

From the critical value (CV) which is calculated by the 'qchisq ( )', function, also we can judge about the null hypothesis. In the situation where $CV < X^2$, we can conclude that null hypothesis need to be rejected. In our case, there are a big difference between the critical value and X-Squared value, as CV = 15.5, and X-Squared = 54.19; therefore, it suggests that the test null hypothesis definitely be rejected.

5.2.1 Chi-Square Test Using Monte Carlo Simulation

Conducting the test with Monte Carlo Simulation, we come to a result suggesting that the null hypothesis be rejected. Although, using this simulation, the p-value is bigger than it was in simple chi-square test, but still far smaller than significant level ($\alpha = 0.05$) was chosen for the test. Therefore, it also rejects the null hypothesis.

```
# Chi-squared test using Monte Carlo simulation
chisq.test(myData, correct = FALSE,
           p = rep(1/length(myData), length(myData)), rescale.p = FALSE,
           simulate.p.value = TRUE, B = 2000)
```

```
        Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  myData
X-squared = 54.193, df = NA, p-value = 0.0004998
```

```
# calculate effect size
CramerV(myData, conf.level = 0.95)
```

```
  Cramer V      lwr.ci      upr.ci
0.11777425 0.07705922 0.14219813
```

The small size of the Cramer's V as measure of the effect size indicates a weak association between test variables: Door and Window.

5.3 **Circular Plot:** Dwelling by Entry Points (Door & Window)

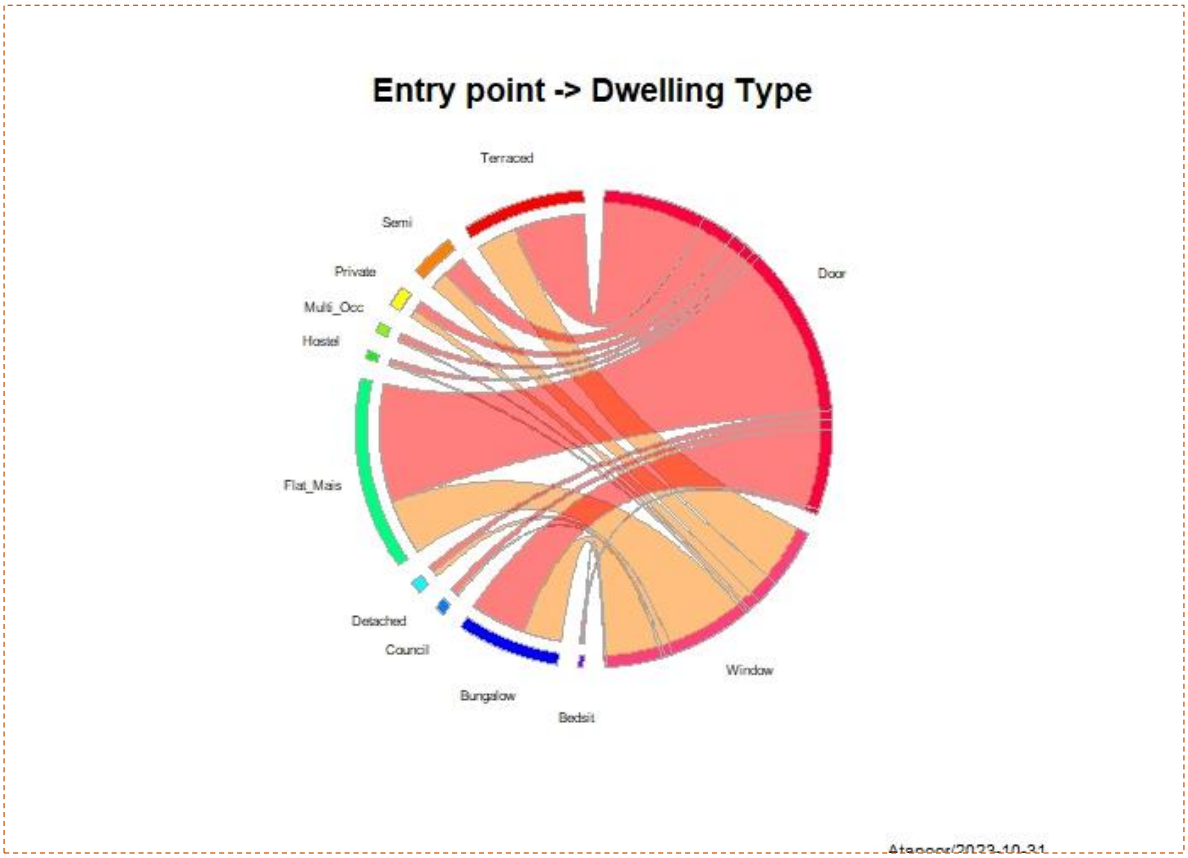
The functions created to make the work easy, again are called here to get the required data

```
data_set <- getData("6.4- burglary-chi.csv")
myData <- makeData(data_set,data_set$dwell_ty,data_set$entry_pn,"Door","Window")
```

Having the data set (*myData*) from calling the functions, we can easily plot Dwelling types by entry points; executing the following code snip, we will get the following circular plot:

```
op <- par(mar = c(5, 8, 4, 2) + 0.1)
PlotCirc(t(as.matrix(myData)), main = "Entry point -> Dwelling Type", cex = 0.5)
par(op)
```

The circular plot clearly shows that the burglars have used doors more than windows to enter dwellings. It means, that the likelihood of doors as entering points for each dwelling types are much higher than windows. To examine whether the higher likelihood of doors as entry points for burglars are by chance, or it is realistic, we will go through a series of testing.



5.4 **Chi-Squared Test:** Dwelling by Entry Points (Door & Window)

Despite the data visualization clearly showed that burglars are using Doors more than Windows as their entry points in different types of houses. To know whether this differences in our data is by chance, a non-parametric test is required by defining the following hypothesis:

- **Null Hypothesis:** it is equally likely that a burglar enters dwellings through the windows and doors;

$$H_0: \quad P_D = P_W \qquad (3)$$

- **Alternative Hypothesis:** the probability that burglars enter dwellings through window differs from door;

$$H_a: \quad P_D \neq P_w \qquad (4)$$

The Chi-Square test was conducted to test whether the probabilities of burglars enter houses through doors and windows were identical for all dwelling types. As has shown below, the results were significant ($X^2(9) = 36.866, P < 0.05$), suggesting that burglaries did not select entry points randomly, and

```
# Chi Square test;
chisq.test(myData)
```

```
        Pearson's Chi-squared test

data:  myData
X-squared = 36.866, df = 9, p-value = 2.78e-05
```

```
# Obtaining the critical value for the test;
qchisq(p = .05, df = 9, lower.tail = FALSE)
```

```
[1] 16.91898
```

most likely they use doors to enter houses; therefore, the null hypothesis is safely rejected.

Conducting the test with Monte Carlo Simulation, we come to a result suggesting that the null hypothesis be rejected. Although, using this simulation, the p-value is bigger than it was in simple chi-square test, but still far smaller than significant level ($\alpha = 0.05$) was chosen for the test. Therefore, it also rejects the null hypothesis.

```
# Chi-Square test using the Monte Carlo simulation based on 2000 replication;
chisq.test(myData, correct = false,
           p = rep(1/length(myData), length(myData)), rescale.p = FALSE,
           simulate.p.value = TRUE, B = 2000)
```

```
        Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  myData
X-squared = 36.866, df = NA, p-value = 0.0004998
```

```
# Determining the effect size;
CramerV(myData, conf.level = 0.95)
```

```
  Cramer V      lwr.ci      upr.ci
0.09786797 0.05192032 0.11983891
```

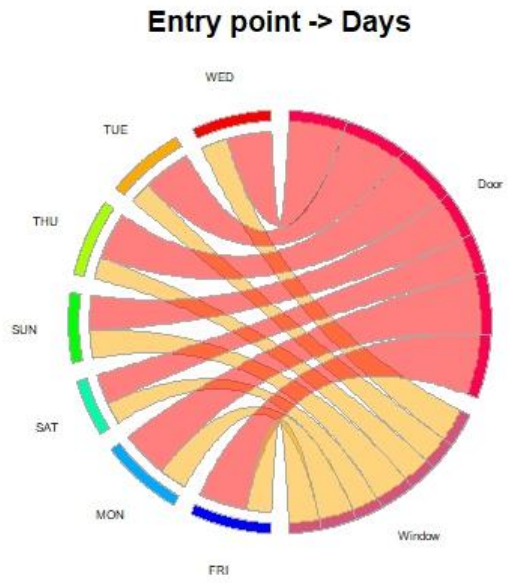
The Cramer's V effect size also small and indicates a weak association between test variables.

5.5 **Circular Plot:** Days by Entry Points (Door & Window)

The circular plot shows clear difference between entry points used by burglars in each days. In each day, burglars have entered houses through doors more than windows. Sunday and Saturday indicates few burglaries in comparison to other days which is intuitively justifiable. Because, on Sunday and Saturday as weekend, people are staying at home.

```
data_set <- getData("6.4- burglary-chi.csv")
myData <- makeData(data_set,data_set$day,data_set$entry_pn,"Door","Window")
```

```
op <- par(mar = c(5, 8, 4, 2) + 0.1)
PlotCirc(t(as.matrix(myData)), main = "Entry point -> Days", cex = 0.5)
par(op)
```



5.6 **Chi-Squared Test:** Days by Entry Points (Door and Window)

To know whether the differences between doors and windows used as entry point by burglars every day are by chance or it is actual, we need to test the data by defining the following hypothesis:

- **Null Hypothesis:** it is equally likely that burglars enters dwellings each day through the windows and doors;

$$H_0: \quad P_D = P_W \qquad (5)$$

- **Alternative Hypothesis:** the probability that burglars enter dwellings through window differs from door;

$$H_a: \quad P_D \neq P_w \qquad (6)$$

The Chi-Square test was conducted to test whether the probabilities of burglars enter houses through doors and windows were identical for all dwelling types. As has shown below, the results were significant ($X^2(6) = 30.957, P < 0.05$), suggesting that burglaries did not select entry points randomly, and most likely they use doors to enter houses; therefore, the null hypothesis is safely rejected.

```
# Chi Square test;
chisq.test(myData)
```

```
        Pearson's Chi-squared test

data:  myData
X-squared = 30.957, df = 6, p-value = 2.583e-05
```

```
# Obtaining the critical value for the test;
qchisq(p = .05, df = 6, lower.tail = FALSE)
```

```
[1] 12.59159
```

```
# Chi-Square test using the Monte Carlo simulation based on 2000 replication;
chisq.test(myData, correct = false,p = rep(1/length(myData), length(myData)),
           rescale.p = FALSE, simulate.p.value = TRUE, B = 2000)
```

```
       Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  myData
X-squared = 30.957, df = NA, p-value = 0.0004998
```

Conducting the test with *Monte Carlo Simulation*, we come to a result suggesting that the null hypothesis be rejected. Although, using this simulation, the p-value is bigger than it was in simple chi-square test, but still far smaller than significant level ($\alpha = 0.05$) was chosen for the test. Therefore, it also rejects the null hypothesis.

```
# Determining the effect size;
CramerV(myData, conf.level = 0.95)
```

```
  Cramer V      lwr.ci      upr.ci
0.08901443 0.04797034 0.11400732
```

The Cramer's V indicating the effect size is also small and suggest weak association between testing variables.

5.7 **Wilcoxon Signed Rank Test:** Male and Female (Two Dependent Group) in AandE

For testing two dependent categorical variables, we use the Wilcoxon Signed Rank Test. The test is done on male and female which are categorical dependent variables in the AandE data set. To know whether the means

```
# Read data from a csv file and assign it to an data AandE;
#then, in second line, make the in_out column as factor;
AandE <- read.csv("6.5- AandE.csv", stringsAsFactors = FALSE)
AandE$in_out <- factor(AandE$in_out)
```

of the two dependent variables are equal or significantly differs from each other, we need define the hypothesis and test it. Therefore, the following hypothesis are developed for the test:

- **Null Hypothesis:** the male and female admitted into AandE unit are equal in average;

$$H_0: \quad \mu_f = \mu_m \qquad (7)$$

- **Alternative Hypothesis:** the number of male admitted AandE unit differs in average;

$$H_a: \quad \mu_f \neq \mu_m \qquad (8)$$

The Wilcoxon Test result (V = 81.5, P < 0.05) suggest that it is statistically significant, and the null hypothesis is safely rejected. Therefore, the means of both male and female admitted in Accident and Emergency Units in London is not equal.

```
# Dependent 2-group Wilcoxon Signed Rank Test;
WRank.Test <- wilcox.test(AandE$male, AandE$female, paired=TRUE)
WRank.Test
```

```
        Wilcoxon signed rank test with continuity correction

data:  AandE$male and AandE$female
V = 81.5, p-value = 0.0003899
alternative hypothesis: true location shift is not equal to 0
```

Checking the means of both female and male variables by the summary () function also confirm that there is remarkable difference between the two means:

```
# Check the means of two group and think about the null hypothesis;
summary(AandE$male)
summary(AandE$female)
```

```
> summary(AandE$male)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    204    5499    6907    6835    8301    9813
> summary(AandE$female)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    189    5732    7355    7198    8895   10621
```

5.8 Mann-Whitney U Test (Two Independent Group):

The Mann-Whitney U Test is used for testing the two group of independent categorical variables. Here the total of both male and female admitted into Accident and Emergency Unites into London Sub-Regions of the AandE data set are tested. The mathematical presentation of the test hypotheses is as following:

```
# Independent 2-group Mann-Whitney U Test
Mann.UTest <- wilcox.test(AandE$total ~ AandE$in_out)
Mann.UTest
```

```
        Wilcoxon rank sum exact test

data:  AandE$total by AandE$in_out
W = 139, p-value = 0.8434
alternative hypothesis: true location shift is not equal to 0
```

$$H_0: \quad \mu_{inner} = \mu_{outer} \qquad (7)$$
$$H_a: \quad \mu_{inner} \neq \mu_{outer} \qquad (8)$$

The Mann-Whitney Test result (W = 139, P > 0.05) suggest that it is not statistically significant, and the null hypothesis is maintained. But, the statistical summary shows difference in means of total patients in London Sub-Region.

```
# Check the means of two group and think about the null hypothesis;
mydata.inner<-AandE[AandE$in_out=="Inner",]
mydata.outer<-AandE[AandE$in_out=="Outer",]
summary(mydata.inner$total)
summary(mydata.outer$total)
```

```
 summary(mydata.inner$total)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   393   12204   14978   13811   17087   18805
 summary(mydata.outer$total)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  9249   11152   14081   14197   16659   20434
```

5.9 **Wilcoxon Signed Rank Test:** Proportional Dependent Variables

For the sake of brevity, I put the mathematical form of the test hypothesis as show in the following:

H_0: $\mu_f = \mu_m$ (1)
H_a: $\mu_f \neq \mu_m$ (2)

The Wilcoxon Test result (V = 340, P > 0.05) on proportional data suggest that it is not statistically significant, and the null hypothesis is not rejected. Therefore, the proportional means of both male and female admitted in Accident and Emergency Units in London is equal. This equality is confirmed by the statistical summary of the proportional variables.

```
# create proportional variables (per thousand population)
AandE <- within (AandE, pMale <- (male / popM) * 1000)
AandE <- within (AandE, pFemale <- (female / popF) * 1000)
AandE <- within (AandE, pTotal <- (total / popT) * 1000)
```

```
# Dependent 2-group Wilcoxon Signed Rank Test on Proportional Data;
WRank.Test2 <- wilcox.test(AandE$pMale, AandE$pFemale, paired=TRUE)
WRank.Test2
```

```
        Wilcoxon signed rank exact test

data:  AandE$pMale and AandE$pFemale
V = 340, p-value = 0.296
alternative hypothesis: true location shift is not equal to 0
```

```
# Check why the result of hypothesis test
summary(AandE$pMale)
summary(AandE$pFemale)
```

```
summary(AandE$pMale)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
40.30   58.46   65.81   64.93   70.99   85.57
summary(AandE$pFemale)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
38.63   58.45   64.65   64.07   68.68   83.01
```

This result is completely different from the result of the same test performed on non-proportional male and female dependent variables. In that test, the test result was statistically significant, and the null hypothesis was rejected.

5.10 **Mann-Whitney U Test:** Proportional Independent Variables

The mathematical forms of the test hypothesis are as following:

H_0: $\mu_{inner} = \mu_{outer}$ (1)
H_a: $\mu_{inner} \neq \mu_{outer}$ (2)

The Mann-Whitney Test result on independent proportional variables (W = 203, P < 0.05) suggest that it is statistically significant, and the null hypothesis is safely rejected. The test result which is rejecting the null hypothesis also is confirmed by the statistical summary. The means of patients admitted in Accident and Emergency Units in London Sub-Regions are remarkably different. However, it is concluded that in order to get the right result close to the actuality of the data, proportional variable is much more reliable.

```
# Independent 2-group Mann-Whitney U Test on Proportional Data;
Mann.UTest2 <- wilcox.test(AandE$pTotal ~ AandE$in_out)
Mann.UTest2
```

```
        Wilcoxon rank sum exact test

data:  AandE$pTotal by AandE$in_out
W = 203, p-value = 0.009935
alternative hypothesis: true location shift is not equal to 0
```

```
# Check why the result of hypothesis test
mydata.inner<-AandE[AandE$in_out=="Inner",]
mydata.outer<-AandE[AandE$in_out=="Outer",]
summary(mydata.inner$pTotal)
summary(mydata.outer$pTotal)
```

```
summary(mydata.inner$pTotal)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
49.66   67.11   68.48   68.48   72.26   84.24
summary(mydata.outer$pTotal)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
39.44   58.16   62.45   61.54   66.16   76.03
```

6 Session-07: Hypothesis Testing (Parametric)

In this session, the parametric test is performed on a data set containing a number of socioeconomic variables. To get a sense of normality of each variables, the statistical summary of each variable has been calculated, and each variable presented visually in terms of box plots and Q-Q plots.

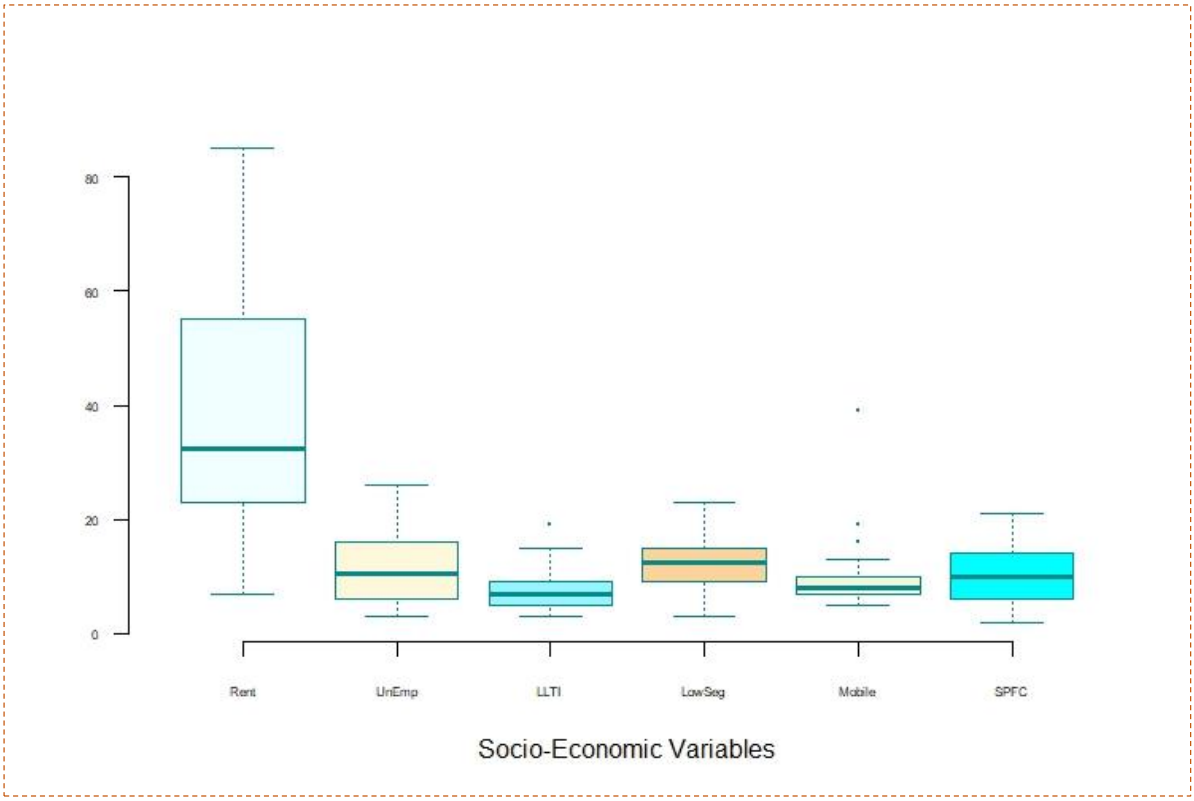
Based on the statistical summary, box plots, and Q-Q plots, the 'llti', 'lowseg', and 'spfc' variables have relatively normal distributions. The remaining socioeconomic variables such as 'rent', 'unemp', and 'mobile' have relatively no-normal distributions.

6.1 Normality Check of Parametric Data

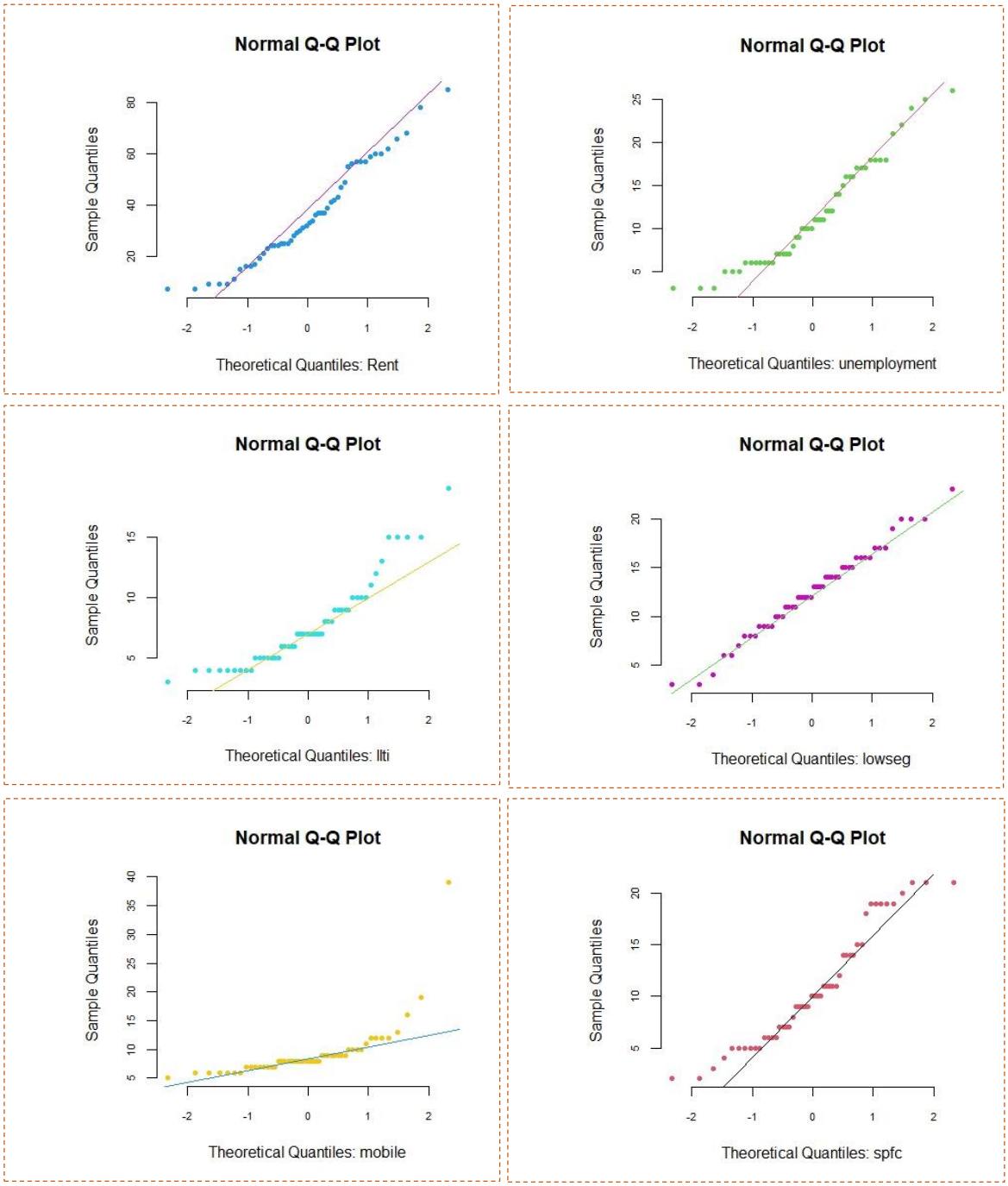
6.1.1 **Section-02:** Summary Statistic

```
> summary(pcs.sample[-1])
      rent           unemp            llti           lowseg          mobile           spfc
 Min.   : 7.00   Min.   : 3.00   Min.   : 3.00   Min.   : 3.00   Min.   : 5.00   Min.   : 2.0
 1st Qu.:23.25   1st Qu.: 6.25   1st Qu.: 5.00   1st Qu.: 9.25   1st Qu.: 7.00   1st Qu.: 6.0
 Median :32.50   Median :10.50   Median : 7.00   Median :12.50   Median : 8.00   Median :10.0
 Mean   :35.94   Mean   :11.58   Mean   : 7.74   Mean   :12.44   Mean   : 9.34   Mean   :10.7
 3rd Qu.:53.50   3rd Qu.:16.00   3rd Qu.: 9.00   3rd Qu.:15.00   3rd Qu.: 9.75   3rd Qu.:14.0
 Max.   :85.00   Max.   :26.00   Max.   :19.00   Max.   :23.00   Max.   :39.00   Max.   :21.0
```

6.1.2 Section -03: Box Plots



6.1.3 Section-04: Q-Q Plots



6.1.4 Section-05: Normality Test:

Rent Variable:

The D (0.098) P > 0.05 suggest that the 'rent' sample has been taken from a population with a normal distribution.

```
ks.test(rent,"pnorm", mean(rent), sd(rent))
```

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  rent
D = 0.098432, p-value = 0.7178
alternative hypothesis: two-sided
```

```
shapiro.test(rent)
```

```
        Shapiro-Wilk normality test

data:  rent
W = 0.95519, p-value = 0.05596
```

Unemployment Variable:

The normality checks on the 'unemp' variable suggest that the data is normally distributed.

```
ks.test(unemp,"pnorm", mean(unemp), sd(unemp))
```

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  unemp
D = 0.13595, p-value = 0.3138
alternative hypothesis: two-sided
```

```
shapiro.test(unemp)
```

```
        Shapiro-Wilk normality test

data:  unemp
W = 0.93564, p-value = 0.009089
```

Limiting Long-Term Illness Variable:

The normality checks on the 'llti' variable suggest that the data is normally distributed.

```
ks.test(llti,"pnorm", mean(llti), sd(llti))
```

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  llti
D = 0.18178, p-value = 0.07345
alternative hypothesis: two-sided
```

```
shapiro.test(llti)
```

```
        Shapiro-Wilk normality test

data:  llti
W = 0.88672, p-value = 0.0001801
```

Low Segment Employment:

The normality checks on the 'lowseg' variable suggest that the data is normally distributed.

```
ks.test(lowseg,"pnorm", mean(lowseg), sd(lowseg))
```

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  lowseg
D = 0.060699, p-value = 0.9928
alternative hypothesis: two-sided
```

```
shapiro.test(lowseg)
```

```
        Shapiro-Wilk normality test

data:  lowseg
W = 0.98734, p-value = 0.8658
```

Single Parent Families with Children:

The normality checks on the 'spfc' variable suggest that the data is normally distributed.

```
ks.test(spfc,"pnorm", mean(spfc), sd(spfc))
```

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  spfc
D = 0.1384, p-value = 0.2937
alternative hypothesis: two-sided
```

```
shapiro.test(spfc)
```

```
        Shapiro-Wilk normality test

data:  spfc
W = 0.92981, p-value = 0.005437
```

Having Recently Moved Location:

The normality checks on the 'mobile' variable suggest that the data highly departures from normal distribution.

```
ks.test(spfc,"pnorm", mean(spfc), sd(spfc))
```

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  mobile
D = 0.26737, p-value = 0.001572
alternative hypothesis: two-sided
```

```
shapiro.test(mobile)
```

```
        Shapiro-Wilk normality test

data:  mobile
W = 0.53263, p-value = 2.197e-11
```

Summary:

VARIABLE	TEST	TEST VALUE	P-VALUE	Normality
Rent	Kolmorov – Smirnov	D = 0.098	0.7178	Confirmed
	Shapiro - Wilk	W = 0.955	0.0559	Confirmed
Umemp	Kolmorov – Smirnov	D = 0.135	0.3138	Confirmed
	Shapiro - Wilk	W = 0.935	0.0090	Not Confirmed
Ltti	Kolmorov – Smirnov	D = 0.181	0.0734	Confirmed
	Shapiro - Wilk	W =0.886	0.00018	Not Confirmed
Lowseg	Kolmorov – Smirnov	D = 0.060	0.9928	Confirmed
	Shapiro - Wilk	W = 0.987	0.8658	Confirmed
spfc	Kolmorov – Smirnov	D = 0.138	0.2937	Confirmed
	Shapiro - Wilk	W = 0.929	0.0054	Not confirmed
mobile	Kolmorov – Smirnov	D = 0.267	0.0015	Not confirmed
	Shapiro - Wilk	W = 0.532	0.0000000000219	Not confirmed

6.2 Section-06: T-Test (Comparing Two Means)

6.2.1 T-Test (rent and umemp):

As the check shows that there is huge difference between variance of these two variables; therefore, the ***Welch test*** is appropriate for the paired test.

```
> var(rent)
[1] 384.0576
> var(lowseg)
[1] 19.88408
```

The Welch paired test (T(49) = 9.7344, P < 0.05) suggest that the result is statically significant, and the null hypothesis is safely rejected.

To understand that whether this difference between the means of these two variable are also practically significant, the *Cohens'D* effect size is calculated. The effect size also indicate that this effect is practically significant.

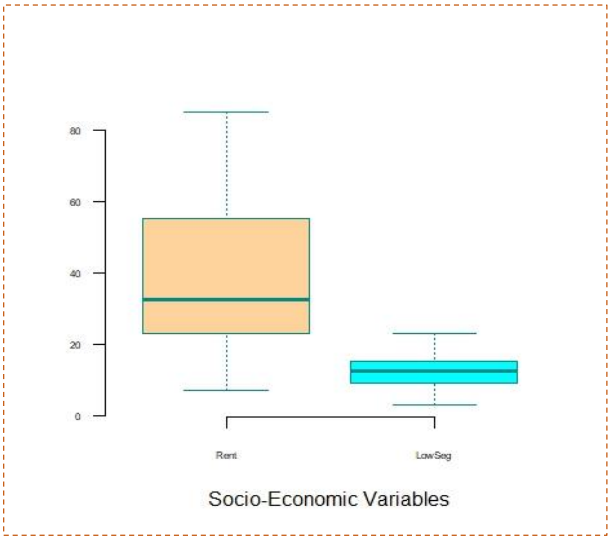
The box plots drawn from the two variables also indicates that there is huge difference between the two variable means.

Therefore, it is concluded that the two sample has not been taken randomly from populations with the same means.

```
t.test(rent, lowseg, paired=TRUE, var.equal = FALSE)
```

```
        Paired t-test

data:  rent and lowseg
t = 9.7344, df = 49, p-value = 4.882e-13
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 18.64865 28.35135
sample estimates:
mean difference
          23.5
```

```
> cohensD(rent, lowseg,  method = "unequal")
[1] 1.653574
```



6.2.2 T-Test (unemp and fcfc):

The checks show that there are differences between variances of the unemp and spfc variables. Therefore, the Welch test is preferred for testing.

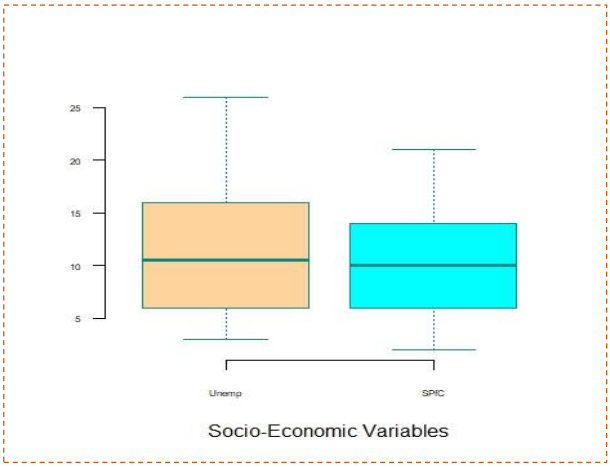
The Welch test on the paired variables shows that the test result (T(49) = 1.659, P > 0.05) is statistically not significant, and the null hypothesis is not rejected. The effect size (d = 0.1519) also confirm that the effect size is small, and there is no practically significant difference between the means.

The box plots drawn from the two variables also show that despite of differences in variances, there are no remarkable difference between means of two samples.

```
> var(unemp)
[1] 36.45265
> var(spfc)
[1] 30.66327
```

```
        Paired t-test

data:  unemp and spfc
t = 1.6591, df = 49, p-value = 0.1035
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -0.1859064  1.9459064
sample estimates:
mean difference
          0.88
```

```
> cohensD(unemp, spfc,  method = "unequal")
[1] 0.1519095
```



6.3 Wilcoxon Signed Rank Test: Non-Normal Two Dependent Group

6.3.1 Wilcoxon Signed Rank Test (rent and lowseg)

The Wilcoxon Signed Rank test result show that test result (V =1222, P < 0.05) are statistically significant, and the null hypothesis is safely rejected. The T-Test on these variables produced the same result in which the null hypothesis rejected.

```
wilcox.test(rent, lowseg, paired = TRUE)
```

```
        Wilcoxon signed rank test with continuity correction

data:  rent and lowseg
V = 1222, p-value = 1.371e-09
alternative hypothesis: true location shift is not equal to 0
```

6.3.2 Wilcoxo Signed Rank Test (unemp and spfc)

The Wilcoxon Signed Rank test result (V = 684, P > 0.05) on two variables, unemp & spfc, suggest that it is not statistically significant, therefore, the null hypothesis is to be retained. The similar result was produced by the T-Test in previous section.

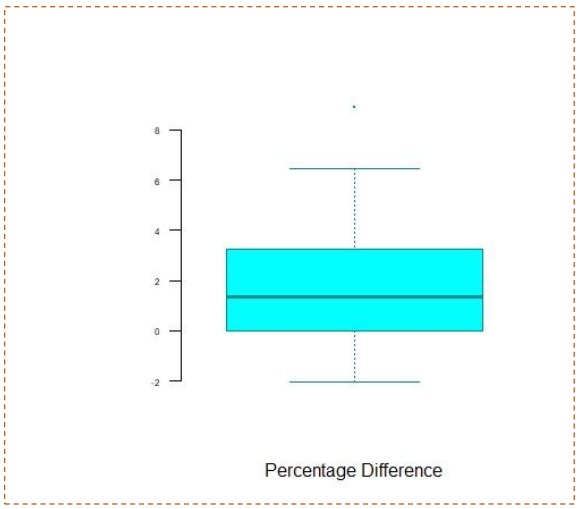
```
wilcox.test(unemp, spfc, paired = TRUE)
```

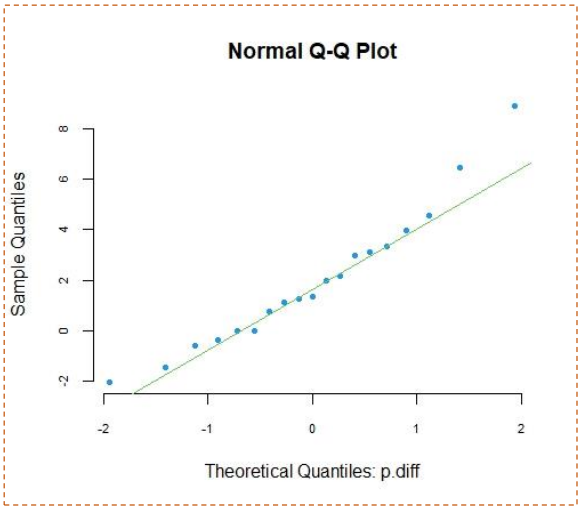
```
        Wilcoxon signed rank test with continuity correction

data:  unemp and spfc
V = 648, p-value = 0.1404
alternative hypothesis: true location shift is not equal to 0
```

7 Session 08: ANOVA

After calculating the percentage of **the** blood-test differences, 'diff', variable based on the 'b-freez' variable, the percentage differences, p-diff, got in more extent normally distributed as the following box plot and the Q-Q plot indicate. The box plot shows that there is only one outlier in the p-diff data points, and the distribution a little bit skewed to the right and it has longer right tail than the left; the Q-Q plot also confirming the normality of the p-diff variable distribution despite of showing the outlier and the skewness on the right side of the distribution.

```
# Calculate a new column for percentage difference and inspect for normality
blood.tests <- within(blood.tests, p.diff <- (blood.tests$diff / b_freeze)*100)

op <- par(mar = c(5, 8, 4, 2) + 0.1)
boxplot(blood.tests$p.diff, xlab="Percentage Difference",
        col = "cyan", border = "cyan4",varwidth = TRUE, outcex = .5,
        outpch = 20, whisklty = 3, frame.plot = FALSE,cex.axis = 0.5, las = 1
        )
par(op)

qqnorm(blood.tests$p.diff, xlab = "Theoretical Quantiles: p.diff", col = 4,
       frame.plot = FALSE, pch = 20, cex.axis = 0.7)
qqline(blood.tests$p.diff, col=3)  # red color
```





Test for Normality:

The null hypothesis for the KS test is that the data set has normally been distributed. The test result is statistically not significant, and P > 0.05; therefore, the null hypothesis is confirmed.

```
ks.test(blood.tests$p.diff, "pnorm", mean(blood.tests$p.diff), sd(blood.tests$p.diff))
```

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  blood.tests$p.diff
D = 0.11711, p-value = 0.9568
alternative hypothesis: two-sided
```

One Sample T-Test on the p-diff:

Despite, the normality of the p-diff distribution is confirmed by different plots and KS test, but the one-sample T-Test [t (3.1637), df = 18, $p < 0.05$) suggest that the result is significant, and the null hypothesis is rejected;

```
t.test(blood.tests$p.diff, mu = 0)  # One-Sample T test for mean=0
```

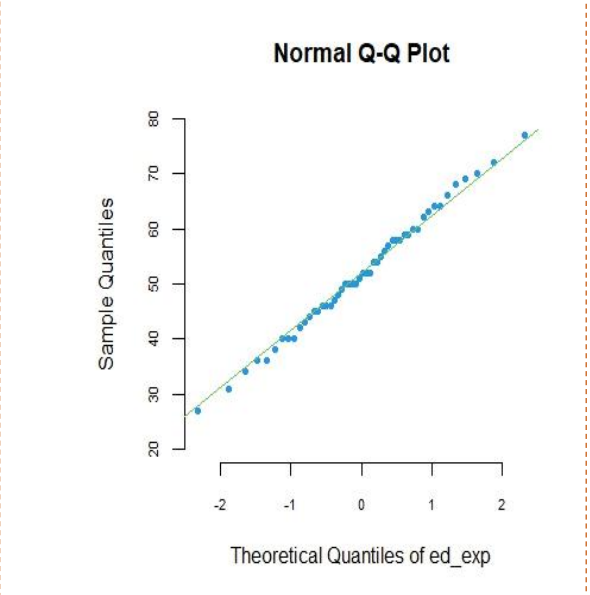
```
        One Sample t-test

data:  blood.tests$p.diff
t = 3.1637, df = 18, p-value = 0.005374
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.6646709 3.2925655
sample estimates:
mean of x
 1.978618
```

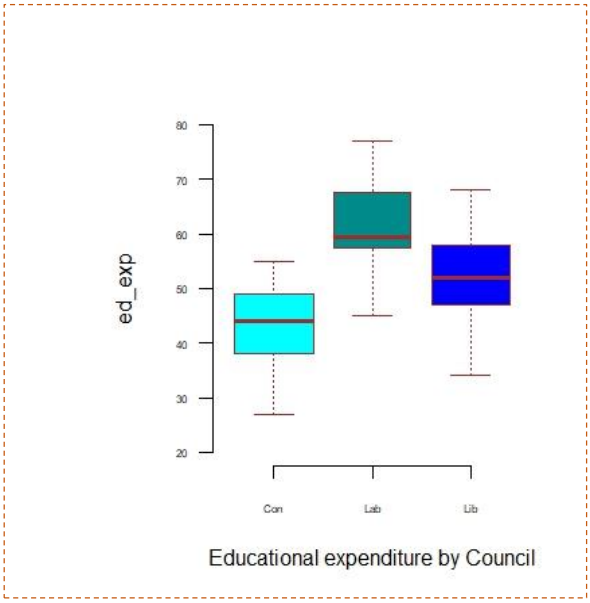
ANOVA Test:

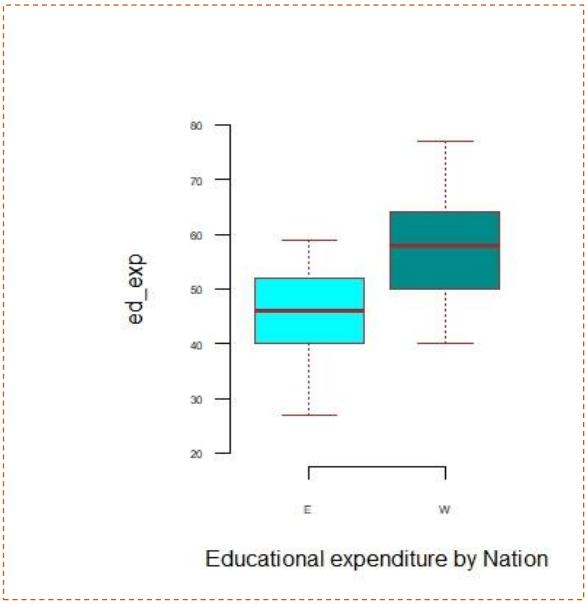
To compare the means of more than two treatment groups, we are using ANOVA test. In this example we are testing the influence of groups of independent variables such Wals and England, Lib, Lab, and Con councils on the education expenditures, ed_exp, dependent variables. First of all a descriptive statistics is performed on the data set by drawing different box plots, and Q-Q plots for checking the statistical characteristics of the data set. The following code snip were used for this purpose.

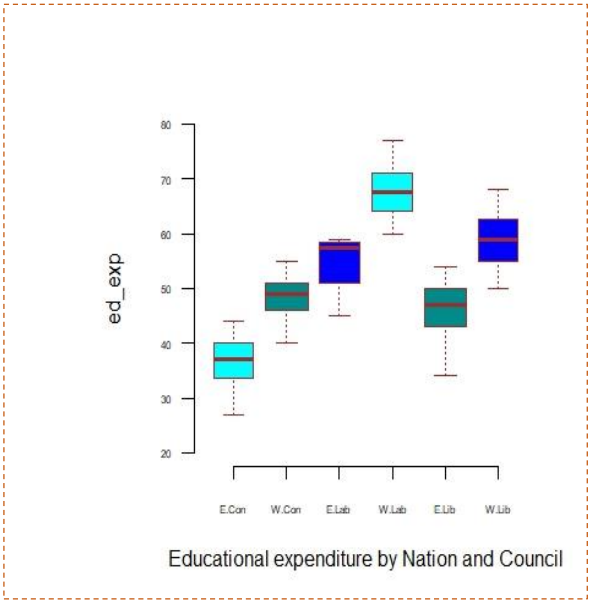
The box plots drawn show that there are lots of variability in terms of educational expenditures between nations (England and Wales), and between councils. the means of all the treatments groups are far different from each other. The Q-Q plot indicates that the educational expenditures (ed_exp) overall has normal distribution.

```
op <- par(mar = c(5, 8, 4, 2) + 0.1)
boxplot(ed_exp ~ eng_wal, data = Johnston.80, xlab = "Educational expenditure by Nation",
        col = c("cyan","cyan4"), border = "brown",varwidth = TRUE, outcex = .5,ylim = c(20,80),
        outpch = 20, whisklty = 3, frame.plot = FALSE,cex.axis = 0.5, las = 1)

boxplot(ed_exp ~ council, data = Johnston.80, xlab = "Educational expenditure by Council",
        col = c("cyan","cyan4","blue"), border = "brown",varwidth = TRUE, outcex = .5,
        ylim = c(20,80), outpch = 20, whisklty = 3, frame.plot = FALSE,cex.axis = 0.5, las = 1)

boxplot(ed_exp ~ eng_wal + council, data = Johnston.80,
        xlab = "Educational expenditure by Nation and Council",col = c("cyan","cyan4","blue"),
        border = "brown",varwidth = TRUE, outcex = .5, ylim = c(20,80),outpch = 20, whisklty = 3,
        frame.plot = FALSE,cex.axis = 0.5, las = 1)

qqnorm(ed_exp, xlab = "Theoretical Quantiles of ed_exp", col = 4, frame.plot = FALSE, pch = 20,
       cex.axis = 0.7,ylim = c(20,80))
qqline(ed_exp, col=3) # red color
par(op)
```









One-Way ANOVA Test of Education Expenditures by Councils:

Aside from the visualization which indicates remarkable differences in means of educational expenditures in different councils, a One-Way ANOVA test performed on the 'ed_exp" variable by councils considering the following hypothesis:

```
# One Way Anova (parametric test) education expenditure by council

fit1 <- aov(ed_exp~council)
summary(fit1)
```

```
> summary(fit1)
            Df Sum Sq Mean Sq F value   Pr(>F)
council      2   2837  1418.5   20.83 3.34e-07 ***
Residuals   47   3201    68.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Null Hypothesis: there are no any difference in means of educational expenditures in each council;

- Alternative Hypothesis: at least the mean of educational expenditures of one group is different from the others;

To present the above hypothesis mathematically:

$$H_0: \quad \mu_{Lib} = \mu_{Lab} = \mu_{Con} \qquad (1)$$

$$H_a: at\ least \quad \mu_{Lib} \neq \mu_{Lab}; or\ \mu_{Lib} \neq \mu_{Con}; or\ \mu_{Lab} \neq \mu_{Con} \qquad (2)$$

As the One-Way ANOVA test result [F(2) = 20.83, P < 0.05) on the ed_exp by the Councils show that it is statistically significant; therefore, the null hypothesis of the test is safely rejected.

Running Post-Hoc Test by Tukey Honestly Significant Difference:

The post-hoc test was performed by the TukeyHSD () function confirms the alternative hypothesis that there are differences between educational expenditures means of all the councils.

```
# Tukey Honestly Significant Difference
TukeyHSD(fit1)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = ed_exp ~ council)

$council
             diff        lwr       upr     p adj
Lab-Con 18.544118  11.587412 25.500823 0.0000002
Lib-Con  9.470588   2.620098 16.321078 0.0045357
Lib-Lab -9.073529 -16.030235 -2.116824 0.0077054
```

One-Way ANOVA Test of Education Expenditures by Council and Nations:

Aside from the visualization which indicates remarkable differences in means of educational expenditures in different councils, a One-Way ANOVA test performed on the 'ed_exp" variable by councils considering the following hypothesis:

- Null Hypothesis: there are no any difference between means of educational expenditures in each council, and nations;
- Alternative Hypothesis: at least the mean of educational expenditures of one of group is different from the others treatments;

```
# One Way Anova (parametric test) of Education Expenditures by Council and Nation;
fit2 <- aov(ed_exp ~ council + eng_wal)
summary(fit2)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
council      2   2837  1418.5   51.44 1.86e-12 ***
eng_wal      1   1932  1932.4   70.07 8.45e-11 ***
Residuals   46   1269    27.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Tukey Honestly Significant Difference
TukeyHSD(fit2)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = ed_exp ~ council + eng_wal)

$council
             diff        lwr      upr     p adj
Lab-Con 18.544118  14.114139 22.97410 0.00e+00
Lib-Con  9.470588   5.108246 13.83293 1.09e-05
Lib-Lab -9.073529 -13.503508 -4.64355 2.95e-05

$eng_wal
        diff      lwr      upr p adj
W-E 12.41882 9.428956 15.40869     0
```

As the One-Way ANOVA test result [F(2) = 20.83, P < 0.05) on the ed_exp by the Councils show that it is statistically significant; therefore, the null hypothesis of the test is safely rejected.

8 Session-09: Factor Analysis and Cluster Analysis

8.1 Task-02 (Section-04): Test Correlation of Dependent and Independent Variables

The results of correlation tests between dependent variable (Life_Male) and independent variables have been summarized in the following table:

For the correlation test, the following hypothesis was developed:

- **Null hypothesis (H0):** there is no correlation between dependent variable and independent variables, and the correlation value is zero;
- **Alternative Hypothesis (Ha):** there are correlations between dependent and independent variables and the correlation values are smaller or bigger than zero;

$$H_0: \ r_s = 0 \qquad (1)$$
$$H_a: \ r_s \neq 0 \qquad (2)$$

```
#Ttest correlation of dependent variable with all independent variables
cor.test(London.dis$Life_Male, London.dis$Dom_Build, method = "spearman")
cor.test(London.dis$Life_Male, London.dis$Smoking, method = "spearman")
cor.test(London.dis$Life_Male, London.dis$Obese, method = "spearman")
cor.test(London.dis$Life_Male, London.dis$Episodes, method = "spearman")
cor.test(London.dis$Life_Male, London.dis$Benefits, method = "spearman")
cor.test(London.dis$Life_Male, London.dis$Crime, method = "spearman")
```

Dep_Var	Indep_Var	Test Value	P-Value	Null Hypothesis
Life_Male	Dom_Build	6990.7	0.1189	Not rejected
Life_Male	Smoking	8782.4	0.0002121	Rejected
Life_Male	Obese	6545.9	0.273	Rejected
Life_Male	Episodes	7180.9	0.07794	Rejected
Life_Male	Benefits	10202	1.016e-10	Not rejected
Life_Male	Crime	8789.7	0.0002036	Not rejected

	Life_Male	Dom_Build	Smoking	Obese	Episodes	Benefits	Crime
Life_Male	1.00	-0.28	-0.61	-0.20	-0.32	-0.87	-0.61
Dom_Build	-0.28	1.00	0.11	-0.43	-0.41	0.43	0.61
Smoking	-0.61	0.11	1.00	0.23	0.32	0.54	0.36
Obese	-0.20	-0.43	0.23	1.00	0.67	0.12	-0.17
Episodes	-0.32	-0.41	0.32	0.67	1.00	0.21	-0.14
Benefits	-0.87	0.43	0.54	0.12	0.21	1.00	0.81
Crime	-0.61	0.61	0.36	-0.17	-0.14	0.81	1.00

8.2 **Task-02(Section-05):** Test Partial Correlation of Dependent and Independent Variables

From the previous section we know that there relatively strong negative correlation between dependent variable (Life_Male), and independent variables (smoking, Benefits, Crime). The correlation between dependent variable and each of the dependent variable might be affected by the third independent variable. To identify it, we need to perform partial correlation test. The null hypothesis for this partial correlation test is defined as following:

Null Hypothesis (H0): there is no statistically significant correlation between the dependent (Life_Male) and the independent variable while controlling for the third independent variable (x);

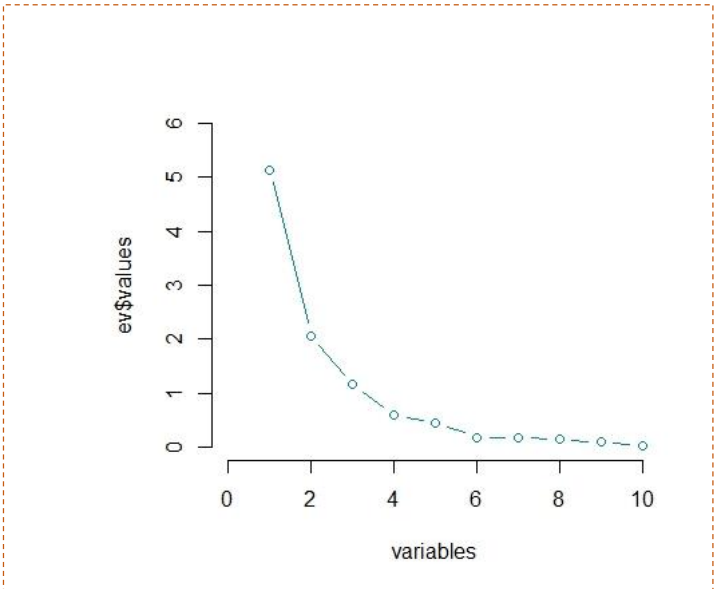
```
# Calculate partial correlation using Pearson and then Spearman
pcor.test(London.dis$Life_Male, London.dis$Benefits, London.dis$Smoking)
pcor.test(London.dis$Life_Male, London.dis$Smoking, London.dis$Benefits)
pcor.test(London.dis$Life_Male, London.dis$Benefits, London.dis$Smoking, method="spearman")
pcor.test(London.dis$Life_Male, London.dis$Smoking, London.dis$Benefits, method="spearman")
```

Dep_Var	Indep_Var	Cont_Var	PCor	P-Value	Method	Null Hypo
Life_Male	Benefits	Smoking	-0.679	2.59e-05	Pearson	Rejected
Life_Male	Smoking	Benefits	-0.351	0.052	Pearson	Not Rejected
Life_Male	Benefits	Smoking	-0.809	3.39e-08	Spearman	Rejected
Life_Male	Smoking	Benefits	-0.334	0.066	Spearman	Not Rejected

Discussion: there are to different result between correlation and partial correlation test on Life_Male and Benefits variable. The correlation test result says there are no significant correlation between these two variables, but, the partial correlation test result says that there is significant negative correlation between the two variables if the effect of the smoking variable is controlled. Therefore, to get a realistic understanding of correlation between two variables, it would be good the we assess their correlation using the partial correlation test. Because, there are a number of variables that affecting the correlation of the two variables.

8.3 **Task-03:** Exploratory Factor Analysis

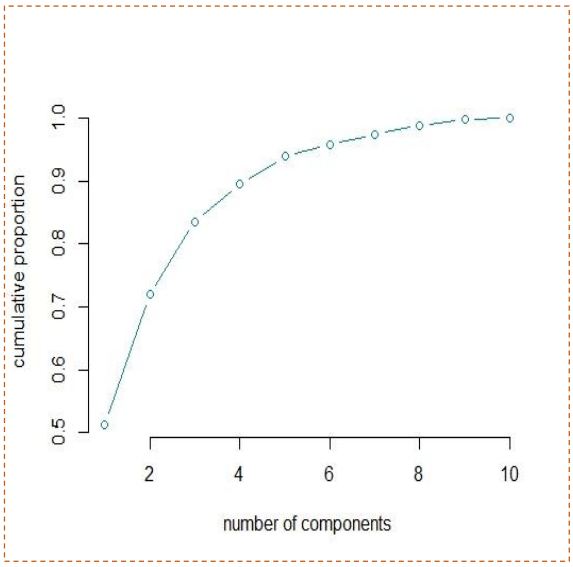
The Scree plot of eigenvalue calculated from correlation matrix of the "London.dis3" data set shows that we can define 4 factors to explain a number of variables in our data set.



The plot drawn from the Cumulative Percentage Variance Explained from the eigenvalue indicates that about 90% of the variance can be explained by the 4 components. Therefore, it also confirms that 4 factor/components are sufficient for defining the variables.

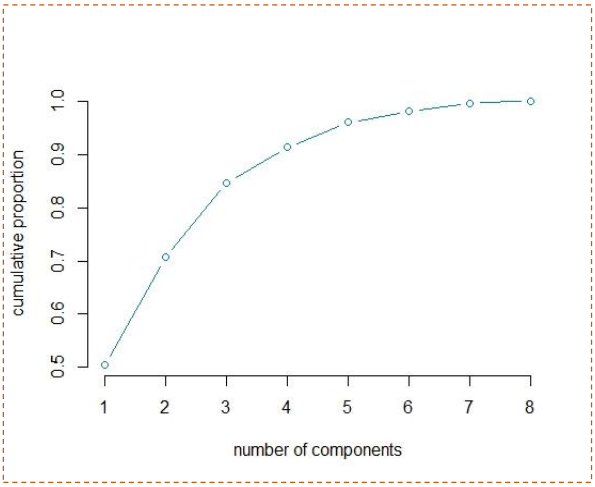
In the first rotated matrix of correlation, there are two variables (*Binge_Drink, Benefits*) which take small value and do not satisfy the threshold value, and they need to be removed.

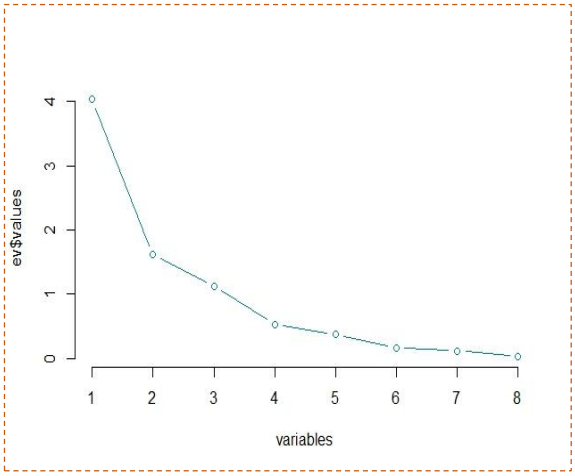
After removing the two variables, and calculating the eigenvalue based on the correlation matrix of new variables, the scree plot is created, and there is no change in the number of factors suggested by the plot eigenvalues.




```
Principal Components Analysis
Call: principal(r = London.dis3, nfactors = 4, rotate = "varimax")
Standardized loadings (pattern matrix) based upon correlation matrix
               RC1   RC3   RC4   RC2   h2    u2 com
Dom_Build     0.91  0.08 -0.27 -0.03 0.91 0.086 1.2
NonDom_Build  0.60  0.71 -0.31  0.09 0.97 0.031 2.4
Dom_Gardens   0.11 -0.84  0.35 -0.28 0.92 0.079 1.6
Greenspace   -0.90 -0.24  0.14 -0.16 0.91 0.093 1.3
Smoking       0.13  0.18  0.21  0.92 0.93 0.067 1.2
Obese        -0.20 -0.18  0.92  0.06 0.93 0.074 1.2
Episodes     -0.31 -0.20  0.73  0.37 0.79 0.205 2.1
Crime         0.45  0.86  0.00 -0.02 0.94 0.059 1.5
```

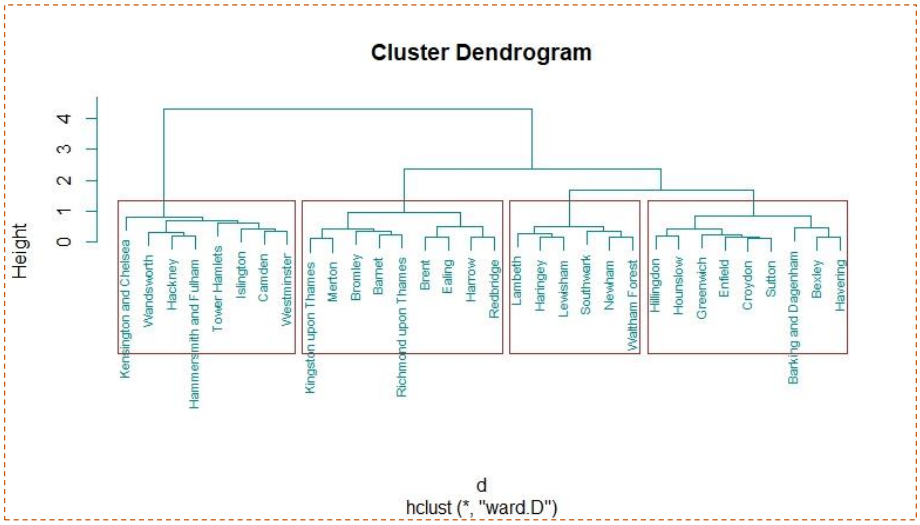
```
Principal Components Analysis
Call: principal(r = London.dis3, nfactors = 4, rotate = "varimax")
Standardized loadings (pattern matrix) based upon correlation matrix
               RC1   RC3   RC2   RC4   h2    u2 com
Dom_Build     0.90  0.03 -0.32 -0.03 0.92 0.082 1.2
NonDom_Build  0.61  0.69 -0.33  0.08 0.96 0.039 2.5
Dom_Gardens   0.07 -0.86  0.31 -0.28 0.92 0.082 1.5
Greenspace   -0.90 -0.21  0.15 -0.12 0.90 0.100 1.2
Smoking       0.14  0.16  0.19  0.91 0.91 0.093 1.2
Binge_Drink   0.64  0.40 -0.48  0.33 0.91 0.090 3.2
Obese        -0.19 -0.20  0.88  0.05 0.86 0.140 1.2
Episodes     -0.25 -0.17  0.79  0.34 0.83 0.173 1.7
Benefits      0.60  0.42  0.38  0.40 0.85 0.152 3.4
Crime         0.47  0.82 -0.05 -0.02 0.90 0.103 1.6
```



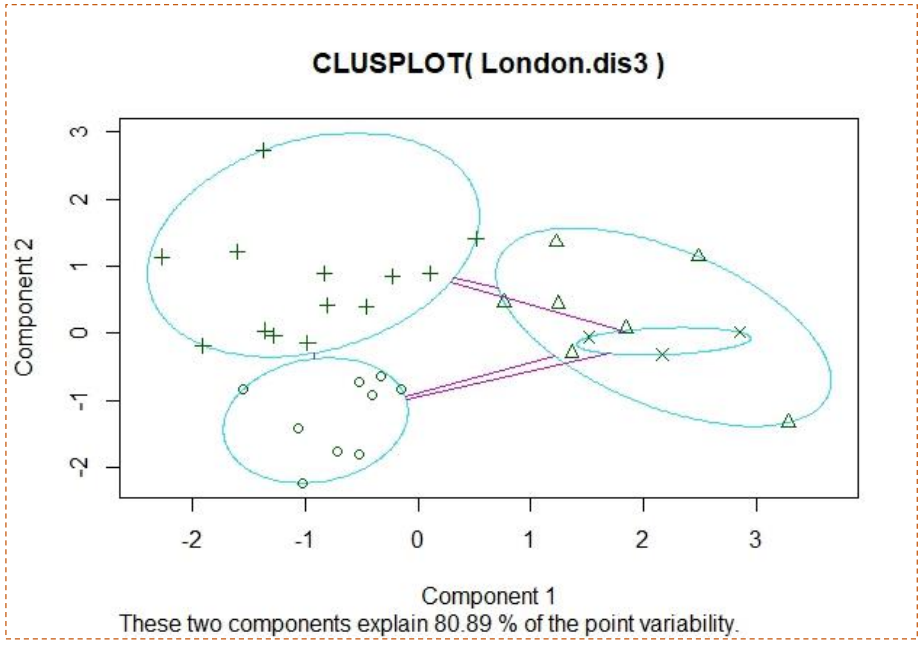


8.4 **Task-04:** Clustering (Wards Hierarchical Clustering and K-Means)

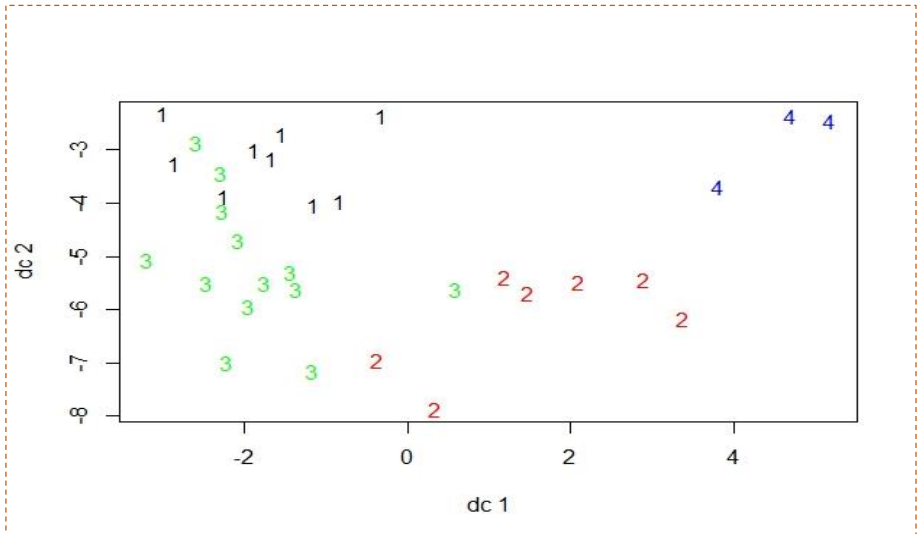
Wards Hierarchical Clustering:



Cluster Plot from K-Means (library (cluster)):



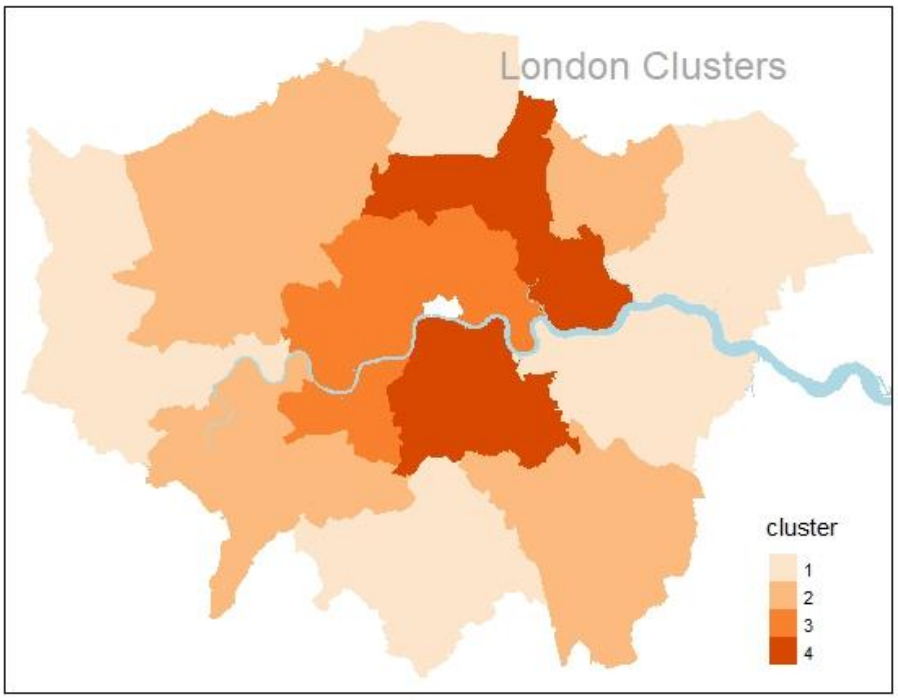
Cluster Plot K-Means (library(fpc))



Matrix of Cluster Means from K-Mean:

```
> aggregate(London.dis3, by=list(fit$cluster),FUN = mean)
  Group.1 Dom_Build   Smoking NonDom_Build     Obese
1       1 0.2741053 0.2556818    0.1406461 0.5231481
2       2 0.7050432 0.5787338    0.4918302 0.3500000
3       3 0.2624689 0.6389860    0.1988695 0.7333333
4       4 0.4038912 0.5454545    0.7582535 0.1166667
```

London Map Clustered:



9 Session – 10: Regression Modeling

9.1 Task-02 (Section-02):

Correlation, and Normality Test: The null hypothesis for the test:

$\boldsymbol{H_0}$: the correlation between the two variable is zero;

The test result (S = 9843.8, P < 0.05) is statistically significant; therefore, the null hypothesis is safely reject, and there is correlation between 'Deprivation' and 'Life_Male';

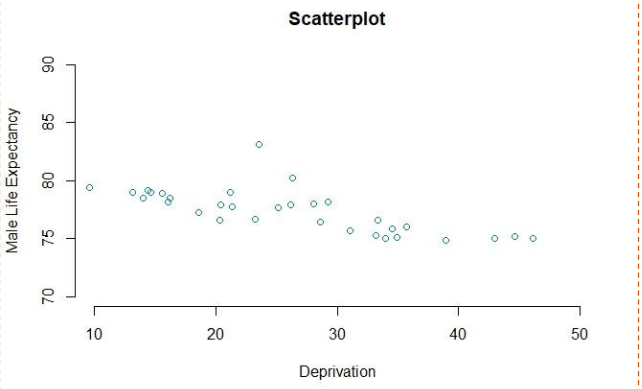
Scatter Plot: the scatter plot also shows that there is a negative correlation between the variables 'Deprivation' and 'Life_Male', and confirm the correlation test result.

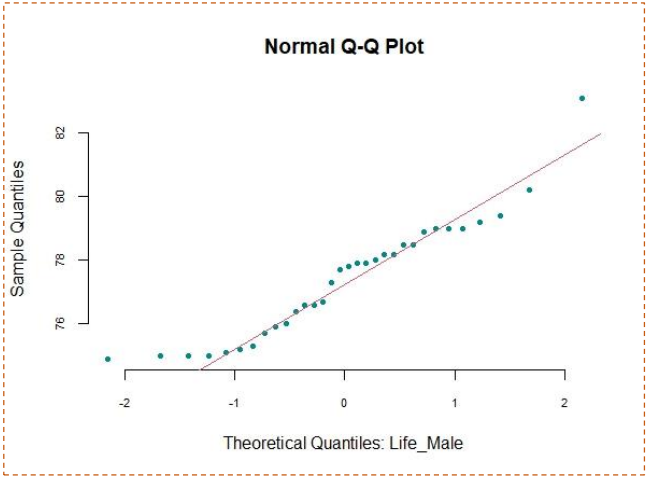
Normality Check: The Q-Q Plot indicate that the dependent variable 'Life_Male' distribution is somehow normal as most of the data points positioned close the normal line.

Kolmogorov-Smirnov test (KS-Test) also show that the test result (D = 0.09, P > 0.05) is not significant, therefore, the null hypothesis (*H0: The Life_Male variable have normally distributed*) of the test is maintained.

```
cor.test(Deprivation, Life_Male, method = "spearman")
```

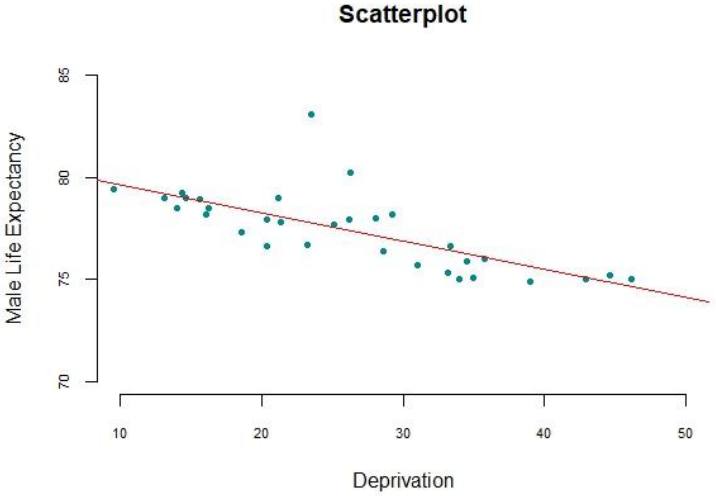
```
        Spearman's rank correlation rho

data:  Deprivation and Life_Male
S = 9843.8, p-value = 2.938e-08
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.8042207
```





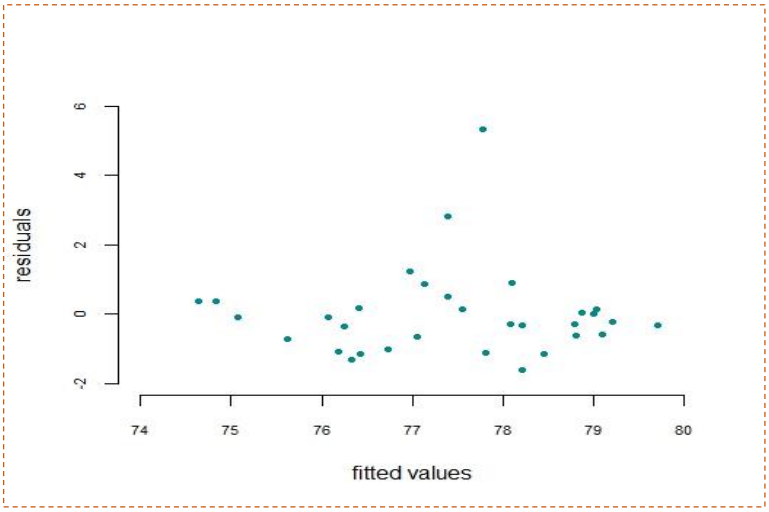

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  Life_Male
D = 0.091833, p-value = 0.9501
alternative hypothesis: two-sided
```

Regression Model: the regression model line is mapped on the scatter plot of the two variables: *Life_Male* and Deprivation. The statistical summary of the regression model shows that the model is fitting about 50% to the relevant data distribution. The F-Test p-value shows statistically significant, and the null hypothesis of the test is rejected; it means that the regression model does not fully fits the data distribution.




```
Call:
lm(formula = Life_Male ~ Deprivation)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6092 -0.6776 -0.2439  0.2342  5.3304

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 81.02974    0.67302 120.398  < 2e-16 ***
Deprivation -0.13867    0.02419  -5.733 2.95e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.324 on 30 degrees of freedom
Multiple R-squared:  0.5228,    Adjusted R-squared:  0.5069
F-statistic: 32.86 on 1 and 30 DF,  p-value: 2.95e-06
```

Residual Plot: the residual scatter plot shows that the variability of the residual is low around the fitted values. It means the homoscedasticity is persevered across the data points distribution. Only there are two residual points appeared to have extreme values.



The KS-Test of the Residuals: the test result (D = 0.20279, P > 0.05) suggest that it is significant, and the null hypothesis of the test is maintained. Therefore, the residual has normally been distributed.

```
        Exact one-sample Kolmogorov-Smirnov test

data:  model1$residuals
D = 0.20279, p-value = 0.1248
alternative hypothesis: two-sided
```

9.2 **Task-02(Section-04):** Multiple Regression Model

```
# Model with all variables
model2 <- lm(Life_Male ~ Dom_Build  + NonDom_Build + Dom_Gardens + Greenspace
             + Smoking + Binge_Drink + Obese + Episodes + Benefits + Crime)
summary(model2)
```

```
Call:
lm(formula = Life_Male ~ Dom_Build + NonDom_Build + Dom_Gardens +
    Greenspace + Smoking + Binge_Drink + Obese + Episodes + Benefits +
    Crime)

Residuals:
     Min       1Q   Median       3Q      Max
-2.02272 -0.44289  0.00305  0.70921  1.38197

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  84.9067214  5.5176720  15.388 6.57e-13
Dom_Build     0.2820230  0.1241387   2.272   0.0337
NonDom_Build  0.3465355  0.2185814   1.585   0.1278
Dom_Gardens  -0.0553691  0.0534437  -1.036   0.3120
Greenspace    0.0343311  0.0433610   0.792   0.4374
Smoking      -0.0748256  0.0652076  -1.147   0.2641
Binge_Drink  -0.4092845  0.2385194  -1.716   0.1009
Obese         0.0928405  0.1143755   0.812   0.4261
Episodes      0.0003746  0.0170236   0.022   0.9826
Benefits     -0.0332028  0.0061391  -5.408 2.30e-05
Crime        -0.0084368  0.0110463  -0.764   0.4535
```

Model2: the summary statistic of the multiple regression model suggests that there are a number of independent variables that are more significant than the others. It means that they have strong correlation with the dependent variable (Life_Male). The P-Value of the variable determine the correlation of the independent variable with the dependent variable. When the p-value of an independent variable is smaller than the defined confident level, it rejects the null hypothesis of the correlation test of the that variable which means the correlation exist. The significant independent variables which have strong correlation with the 'Life_Male' dependent variable have been circled above.

Detecting Multicollinearity: the multicollinearity of the independent variables are detected by calculating the variance inflation factor. The result of the variance inflation factor shows that there are no multicollinearity between the independent variables of the regression model.

```
# Calculate variance inflation factor
library(car)
vif(model2)
sqrt(vif(model2)) > 2  # if > 2 vif too high
```

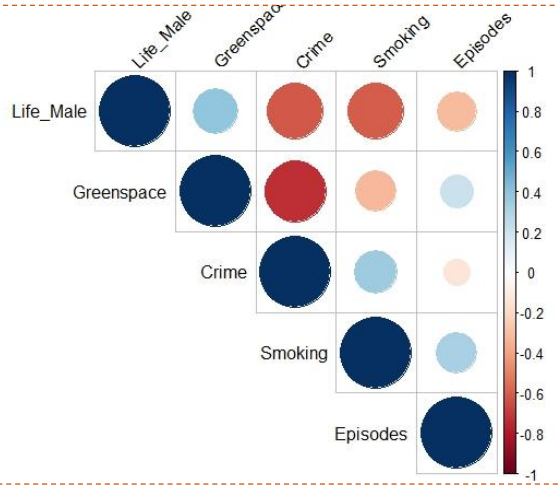
```
> vif(model2)
   Dom_Build NonDom_Build  Dom_Gardens   Greenspace
    5.826126    23.190030     4.242418     9.289247
     Smoking  Binge_Drink        Obese     Episodes
    2.135680     7.082530     3.899891     4.169436
    Benefits        Crime
    3.490845     6.406399
> sqrt(vif(model2)) > 2  # if > 2 vif too high
   Dom_Build NonDom_Build  Dom_Gardens   Greenspace
        TRUE         TRUE         TRUE         TRUE
     Smoking  Binge_Drink        Obese     Episodes
       FALSE         TRUE        FALSE         TRUE
    Benefits        Crime
       FALSE         TRUE
```

Check Correlation: in this part, correlation of the five variables which were in factor analysis component are determined. This correlation also plotted. As the correlation plot also show, two variables (Greenspace and Crime) is relatively strongly correlated.

```
# model with four variables representing components from factor analysis
cor3 <- cor(London.dis2[, c(5,4,12,7,10)], method = "spearman")
round(cor3, 2)
corrplot(cor3, type = "upper", tl.col = "black", tl.srt = 45)
```

```
> round(cor3, 2)
            Life_Male Greenspace Crime Smoking Episodes
Life_Male        1.00       0.40 -0.61   -0.61    -0.32
Greenspace       0.40       1.00 -0.74   -0.32     0.22
Crime           -0.61      -0.74  1.00    0.36    -0.14
Smoking         -0.61      -0.32  0.36    1.00     0.32
Episodes        -0.32       0.22 -0.14    0.32     1.00
```

Model-03: the third regression model is created based 'Greenspace', 'Crime', 'Smoking', and 'Episodes' independent variable to check their effect on the 'Life_Male' dependent variable. The summary statistic of the model along the multicollinearity of independent variables have been calculated as shown in the picture below.




```
Call:
lm(formula = Life_Male ~ Greenspace + Crime + Smoking + Episodes)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6418 -0.7812 -0.1888  0.6973  3.3464

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 89.400085   3.020137  29.601  < 2e-16 ***
Greenspace   0.043878   0.025460   1.723  0.09625 .
Crime       -0.005213   0.008138  -0.641  0.52717
Smoking     -0.177092   0.074085  -2.390  0.02407 *
Episodes    -0.044568   0.014397  -3.096  0.00454 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.368 on 27 degrees of freedom
Multiple R-squared:  0.5415,    Adjusted R-squared:  0.4736
F-statistic: 7.973 on 4 and 27 DF,  p-value: 0.0002226
```

The summary statistic shows that the two variables (Smoking, Episodes) have the smallest p-value; it means they are more correlated to the dependent variable. The R-squared showing the size of fitness of the model is small in this regression model. The F-statistic suggest that overall there are correlation between dependent and independent variable as its p-value (0.0002226) is much more small then the default confidence level (0.05).

```
> sqrt(vif(model3)) > 2
Greenspace      Crime    Smoking   Episodes
     FALSE      FALSE      FALSE      FALSE
```

From the calculation of the variance of inflation factor it is concluded that the multicollinearity do not exist between the independent variable of this regression model.

```
pcor.test(Life_Male, Crime, Greenspace)
pcor.test(Life_Male, Greenspace, Crime)
```

```
> pcor.test(Life_Male, Crime, Greenspace)
     estimate   p.value  statistic  n gp  Method
1 -0.02933233 0.8755304 -0.1580274 32  1 pearson
> pcor.test(Life_Male, Greenspace, Crime)
  estimate   p.value statistic  n gp  Method
1  0.21626 0.2426071  1.192823 32  1 pearson
```

Partial Correlation Test: the test between 'Life_Male' and 'Crime' while controlling the 'Greenspace' independent variables is statistically insignificant, and null hypothesis is maintained. Therefore, the test suggests that there is no correlation between 'Life_Male' and 'Crime'. The Test of 'Life_Male' and 'Greenspace' while controlling the 'Crime' variable also says that it is not significant, and the null hypothesis is maintained.

Model3a: again another regression model is created based on the 'Life_Male' dependent variable, and 'Greenspace', 'Smoking', and 'Episodes' independent variables. The summary statistic of the model suggest that these three independent variables have effective correlation with the dependent variable. Because, the p-value of each of these variables are much smaller than the default confidence level (0.05); therefore, each of them are rejecting their associated null hypothesis.

The variance inflation factor calculation of the model shows that there are no multicollinearity among the independent variables.

```
Call:
lm(formula = Life_Male ~ Greenspace + Smoking + Episodes)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6361 -0.7323 -0.2256  0.6953  3.5249

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 88.31182    2.47066  35.744  < 2e-16 ***
Greenspace   0.05122    0.02249   2.277  0.03059 *
Smoking     -0.18946    0.07077  -2.677  0.01227 *
Episodes    -0.04161    0.01349  -3.084  0.00456 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.354 on 28 degrees of freedom
Multiple R-squared:  0.5346,    Adjusted R-squared:  0.4847
F-statistic: 10.72 on 3 and 28 DF,  p-value: 7.298e-05

> sqrt(vif(model3a)) > 2
Greenspace    Smoking   Episodes
     FALSE      FALSE      FALSE
```

Relative Importance of Variables: in the ‘Model3a’ we found that all the three independent variables are important for the explaining the ‘*Life_Male’* dependent variable. Now, it is determined that which is relatively more important. The result indicates that among all the three variables, ‘*Smoking*’ are the relatively more important variable for describing the dependent variable, and the ‘*Greenspace*’ is the least important predictor among all.

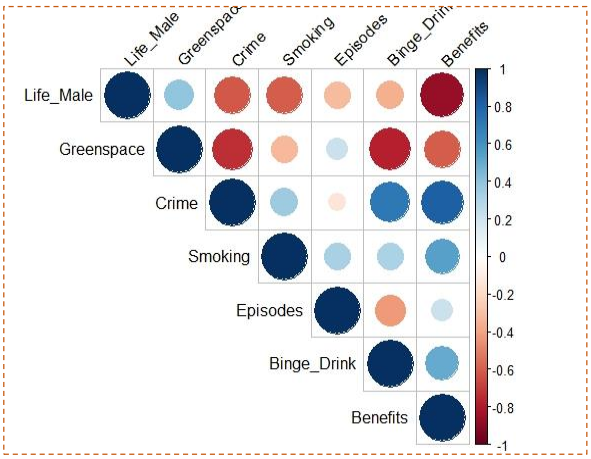
```
# relative importance of variables
library(relaimpo)
calc.relimp(model3a, type = c("lmg"), rela = TRUE)
```

```
Response variable: Life_Male
Total response variance: 3.556613
Analysis based on 32 observations

3 Regressors:
Greenspace Smoking Episodes
Proportion of variance explained by model: 53.46%
Metrics are normalized to sum to 100% (rela=TRUE).

Relative importance metrics:

                lmg
Greenspace 0.1687922
Smoking    0.4577703
Episodes   0.3734376

Average coefficients for different model sizes:

                      1X          2Xs          3Xs
Greenspace  0.04367634  0.04799786  0.05122375
Smoking    -0.29247074 -0.25995405 -0.18946138
Episodes   -0.04354081 -0.04278427 -0.04161196
```

Stepwise Approach for the Best Model: another multiple regression model (*model3b*) is created from a number of variables which are not in the component factor analysis. The correlations of these variables are plotted below.

```
model3b <- lm (Life_Male ~ Greenspace + Crime + Smoking + Episodes +
               Binge_Drink + Benefits)
summary(model3b)
sqrt(vif(model3b)) > 2
calc.relimp(model3b, type = c("lmg"), rela = TRUE)
```

```
Call:
lm(formula = Life_Male ~ Greenspace + Crime + Smoking + Episodes +
    Binge_Drink + Benefits)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8013 -0.5450 -0.1715  0.5892  3.0345

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 88.933737   4.076275  21.817  < 2e-16 ***
Greenspace  -0.024638   0.027122  -0.908 0.372325
Crime        0.009846   0.007213   1.365 0.184414
Smoking     -0.086204   0.064657  -1.333 0.194471
Episodes    -0.015411   0.015032  -1.025 0.315071
Binge_Drink -0.105941   0.191090  -0.554 0.584229
Benefits    -0.028909   0.006606  -4.376 0.000188 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.057 on 25 degrees of freedom
Multiple R-squared:  0.7465,    Adjusted R-squared:  0.6857
F-statistic: 12.27 on 6 and 25 DF,  p-value: 2.033e-06
```



There is no multicollinearity among the independent variables in the model.

```
> sqrt(vif(model3b)) > 2
 Greenspace       Crime     Smoking    Episodes Binge_Drink    Benefits
      FALSE       FALSE       FALSE       FALSE       FALSE       FALSE
```

The ‘*Benefits*’ is relative important independent variable for the dependent variable, and ‘*Smoking*’ and ‘*Episodes*’ stand in the second and third positions respectively in terms of importance.

```
> calc.relimp(model3b, type = c("lmg"), rela = TRUE)
Response variable: Life_Male
Total response variance: 3.556613
Analysis based on 32 observations

6 Regressors:
Greenspace Crime Smoking Episodes Binge_Drink Benefits
Proportion of variance explained by model: 74.65%
Metrics are normalized to sum to 100% (rela=TRUE).

Relative importance metrics:

                  lmg
Greenspace  0.04946814
Crime       0.03491201
Smoking     0.19045203
Episodes    0.18668103
Binge_Drink 0.03731938
Benefits    0.50116741
```

The stepwise selection resulted in creation of ‘model4’ out of ‘model3b’:

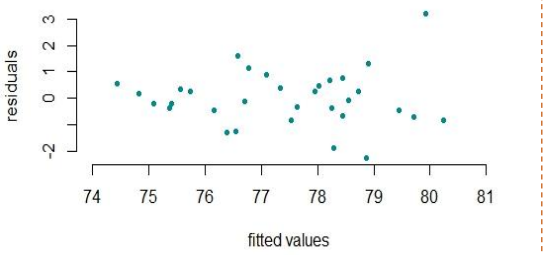
```
# forward stepwise selection
model4 <- stepwise(model3b, direction = "forward")
summary(model4)
```

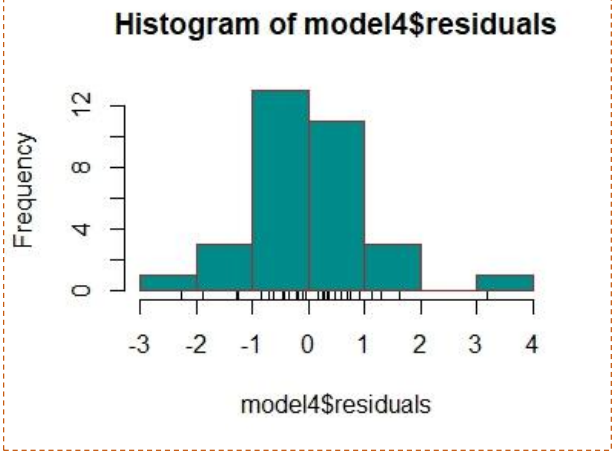
```
Call:
lm(formula = Life_Male ~ Benefits + Episodes)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2674 -0.5097 -0.0773  0.4953  3.1776

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 87.773485   1.862490  47.127  < 2e-16 ***
Benefits    -0.025528   0.003819  -6.684 2.49e-07 ***
Episodes    -0.028578   0.009691  -2.949  0.00624 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.077 on 29 degrees of freedom
Multiple R-squared:  0.6947,    Adjusted R-squared:  0.6737
F-statistic:    33 on 2 and 29 DF,  p-value: 3.372e-08
```

The summary statistic of ‘model4’ shows that only two significant independent variables (Benefits, Episodes) selected out of all the independent variables in ‘model3b’. Therefore, based on the stepwise selection for the best model, only two independent variables can well describe the dependent variable, ‘Life_Male’.

The histogram, rug, scatter plot, KS-Test, variance of inflation factors, and relative importance of the ‘model4’ also have been calculated.

```
hist(model4$residuals, col = "cyan4", border = "brown")
rug(model4$residuals)
plot(model4$residuals ~ model4$fitted.values,pch = 20,xlim = c(74,81), xlab = "fitted values",
     ylab = "residuals", frame.plot = FALSE, col = "cyan4")
ks.test(model4$residuals, "pnorm", mean(model4$residuals), sd(model4$residuals))
sqrt(vif(model4)) > 2
calc.relimp(model4, type = c("lmg"), rela = TRUE)
```






```
        Exact one-sample Kolmogorov-Smirnov test

data:  model4$residuals
D = 0.085178, p-value = 0.959
alternative hypothesis: two-sided
```

- the residual histogram and scatter plot, and the KS-Test suggest that the residual relatively has normally been distributed.
- The relative importance calculation says that the 'Benefits' independent variable is relatively more important.
- There variance inflation calculation says that multicollinearity does not exist between independent variable in the model.
- The ANOVA test between 'model2', and 'model4' says that there are no different between the two model, as its p-value is bigger than 0.05, and its null hypothesis retains.

```
Response variable: Life_Male
Total response variance: 3.556613
Analysis based on 32 observations

2 Regressors:
Benefits Episodes
Proportion of variance explained by model: 69.47%
Metrics are normalized to sum to 100% (rela=TRUE).

Relative importance metrics:

               lmg
Benefits 0.7725658
Episodes 0.2274342

Average coefficients for different model sizes:

                  1X         2Xs
Benefits -0.02813004 -0.02552817
Episodes -0.04354081 -0.02857761
```

```
> sqrt(vif(model4)) > 2
Benefits Episodes
   FALSE    FALSE
```

```
# test whether model2 and model4 are significantly different using F test
anova(model2, model4, test = "F")
```

```
Analysis of Variance Table

Model 1: Life_Male ~ Dom_Build + NonDom_Build + Dom_Gardens + Greenspace +
    Smoking + Binge_Drink + Obese + Episodes + Benefits + Crime
Model 2: Life_Male ~ Benefits + Episodes
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     21 19.055
2     29 33.658 -8   -14.602 2.0116 0.09536 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note: the remaining part of the session also can be done the same way. Therefore, I have stopped working to this point.