# Portfolio

**Name:** Cibi Aswanth.V.S

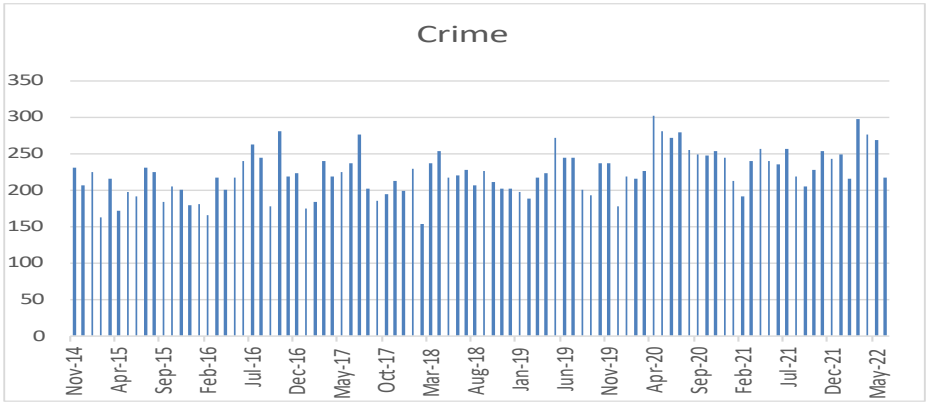
---

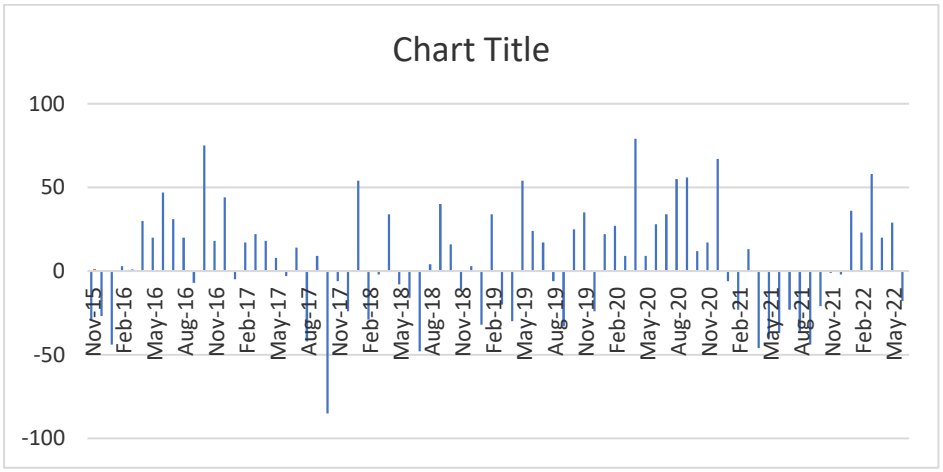
## SESSION-1

### Task 1: Indexing ,Trend, Standardization, Distribution

**Question:** Time Series Graph

**OUTPUT:**



**Interpretation:**

First the Dataset " 01_Crime_teach.xls" to be downloaded, from the link (https://www.police.uk/pu/your-area/metropolitan-police-service/beckton/ ) and then crime dataset is to be opened through Excel and where it contains three Sheets[Becton_crime, Crime_index, Becton_Crime_tool].

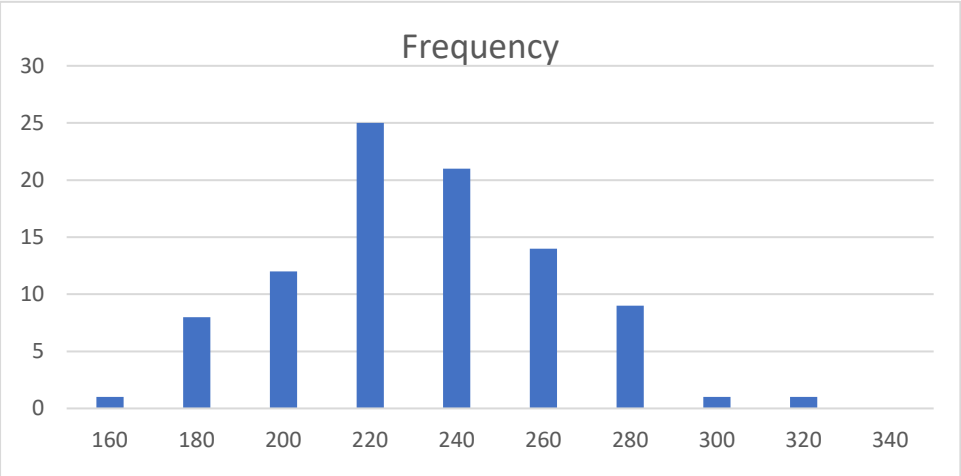
From Above Chart , our data shows that the crime rate varies over time. For instance, the crime rate was higher in January 2018 and lower in July 2018. By comparing crime rates for several months and years, it is possible to identify trends in crime and examine their causes, such as Covid or economic downturns.

## Task 3: The graphed distribution and together with the values for the summary statistics:

**Question:** There is a distribution of monthly crime values from minimum to maximum with an average somewhere in the middle. What does that distribution look like?

**Output:**

| *Bin* | *Frequency* |
|---|---|
| 160 | 1 |
| 180 | 8 |
| 200 | 12 |
| 220 | 25 |
| 240 | 21 |
| 260 | 14 |
| 280 | 9 |
| 300 | 1 |
| 320 | 1 |
| 340 | 0 |
| More | 0 |



**Interpretation:**

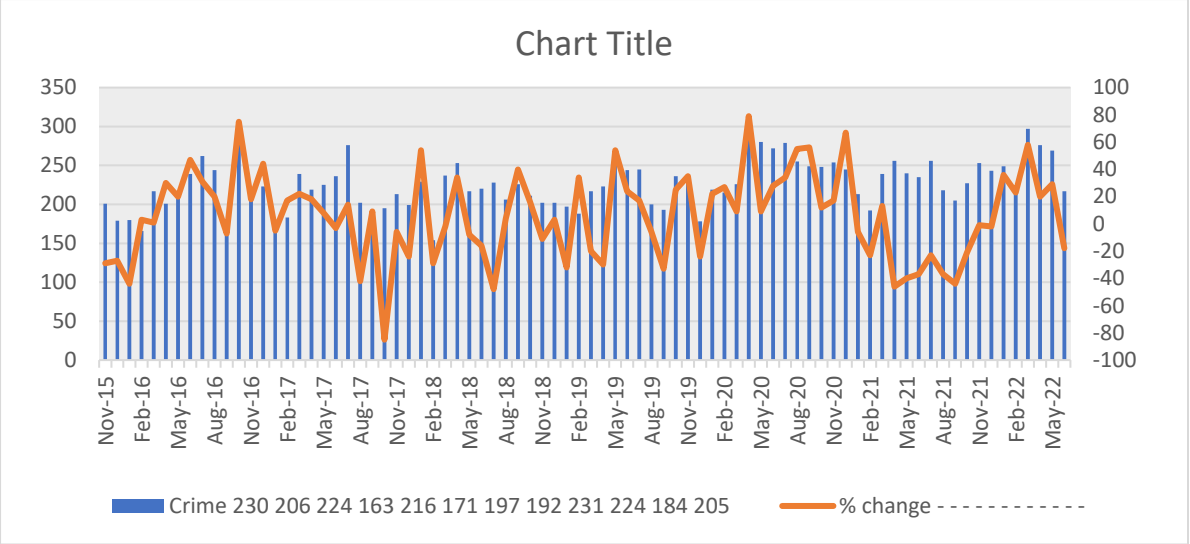
Using a data analysis tool, a bin and Frequency table was produced, and then histogram was Created. The formula in the excel was used to identify the Maximum Value , Minimum Value ,range and mean of the crime ratings, and the standard deviation was used to determine the dispersion of the set of values based on the crime counts.

It can be determined how frequently or infrequently a crime occurred based on the month by looking at the frequency of crimes that occurred during that month.

**Task 4: The second graph (percentage annual change):**

**Question: What can you say about the patterns of change?**

**Output:**



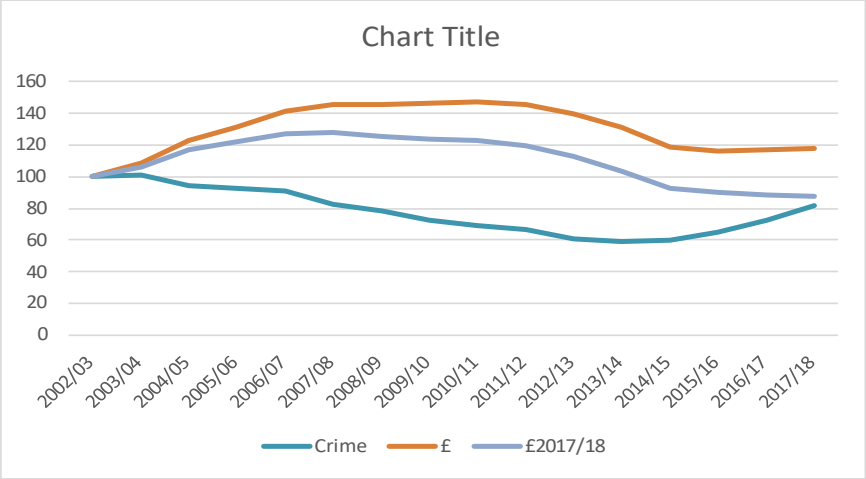

**Interpretation:**

Crime count frequency and percentage have been calculated by comparing crime rates based on various years, which is the outcome of the Percentage annual change. Overall, the layout of the crime data base based on years and count has been identified by comparing or creating bin-frequency,% change, and time-period.

**Task 5:**

**Question:The line chart:**

**Output:**



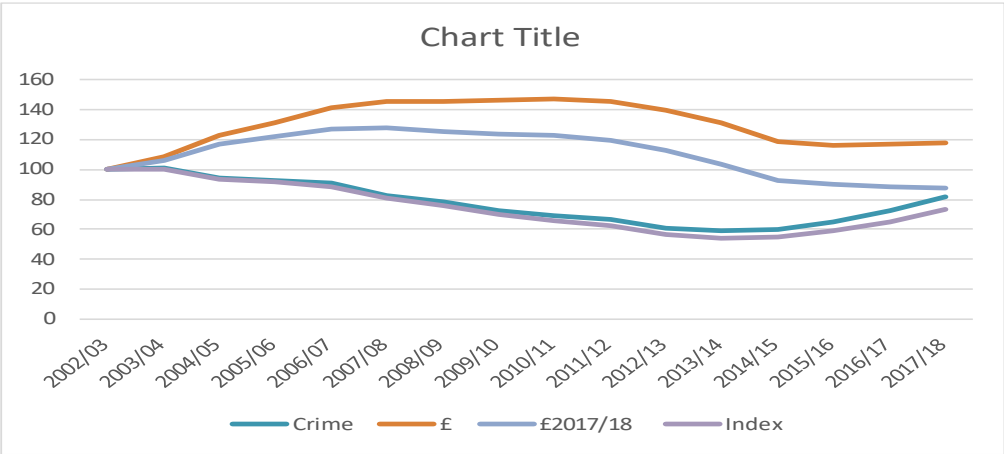

**Interpretation:**

The time series of crime reviewed the data of what is occurring to crime in England and Wales by shifting to the crime index sheet. Regarding the missing values of crime counts and pounds, the information was gathered from the Region Crime Tool Sheet and adjusted to make it equivalent to the information for England and Wales using paste option. Column C shows the annual cost of policing work, while column E shows that the values have risen over time. So through indexing By adjusting for crime, pounds and policy costs, the first year is made equal to 100. By doing this, we may see how the crime rate has changed over time. The Same way the analysis has been done for £2017/18 and pounds. According to the first output graph, it is Estimated policing costs in million pounds are 17% higher in 2017–18 than they were a decade and a half earlier, but in real terms, they are 13% lower.

**Task 6:**

**Question:Line chart**

**Output:**



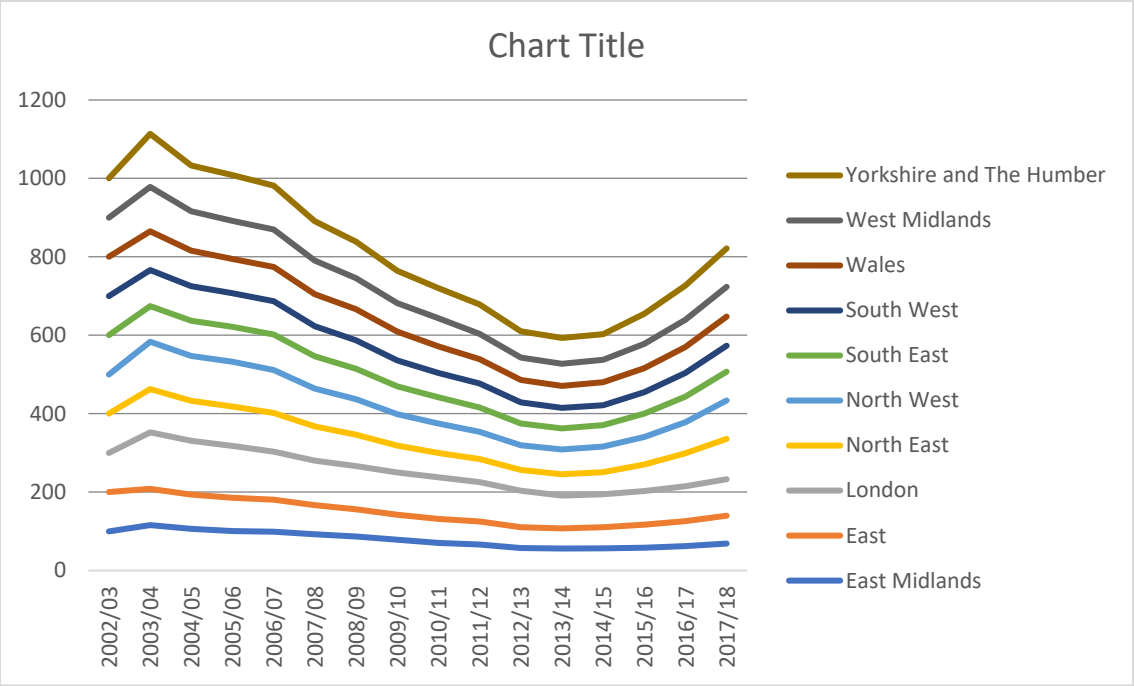
**Interpretation:**

According to the task 5 line chart and the aforementioned finding, the crime rate is declining year by year.

**Task 7:**

**Question:The line graph:**

**Output:**




**Interpretation:**

The graph covers the data of Yorkshire and The Humber , West Midlands , Wales, South West, South East, North West, North East, London, East and East Midlands from the years 2002 to 2018 , where there is regulate ups and downs in index Crime.

## SESSION-3

## Task:4(The SQL Queries):

## Question :Highlight the whole query where ph_value>5

**Output:** Query: 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

| | soil_id | soil_type | soil_id | ph_val | sur_id | assessed |
|---|---|---|---|---|---|---|
| 1 | 213 | 32 | 213 | 5.12 | 1 | 1991-12-24 |
| 2 | 222 | 26 | 222 | 5.13 | 3 | 1990-05-19 |
| 3 | 225 | 17 | 225 | 5.34 | 4 | 1992-07-04 |
| 4 | 141 | 17 | 141 | 5.43 | 3 | 1990-11-22 |
| 5 | 212 | 27 | 212 | 5.44 | 3 | 1992-07-04 |
| 6 | 245 | 21 | 245 | 5.74 | 4 | 1990-11-06 |
| 7 | 193 | 17 | 193 | 6.44 | 4 | 1990-05-19 |
| 8 | 179 | 17 | 179 | 6.54 | 1 | 1992-09-17 |
| 9 | 170 | 30 | 170 | 6.54 | 4 | 1989-09-12 |

**Interpretation:**

I have Download the SQL and then create a new data base and add the three csv files. Before that copy the values to the notepad and add commas to separate the values and arrange it to particular attributes and then save it. Then save it file name as Soil_id, Soil_Type, Surveyor in csv format. Then load it to sql in the new data base what we have created. In the above output the values are showed where ph_value > 5.

## Task:5(The SQL Queries):

## Question: Construct a query to show the soil_id, ph-val and assessed for N.Brown in descending order of data assessed.

**Output:** 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.assessed DESC

| | soil_id | ph_val | assessed | surveyor |
|---|---|---|---|---|
| 1 | 261 | 3.29 | 1992-07-04 | N.Brown |
| 2 | 182 | 3.34 | 1991-12-24 | N.Brown |
| 3 | 254 | 4.32 | 1991-12-24 | N.Brown |
| 4 | 194 | 4.52 | 1990-11-06 | N.Brown |
| 5 | 216 | 4.73 | 1990-11-06 | N.Brown |
| 6 | 232 | 4.32 | 1990-05-19 | N.Brown |
| 7 | 274 | 4.56 | 1990-05-19 | N.Brown |
| 8 | 157 | 2.36 | 1989-07-12 | N.Brown |

**Interpretation:**

Here Select clause is used to select the relevant table with relevant attributes like soil_id, ph_val, assessed from the Soil_id table and then joining two tables by using join clause with relevant attribute Surveyor from the Surveyor table with Soil_id table in the task. Then we used order clause to put Assessed value in the Descending Order by using key word DESC , then we get N.Brown Values for Surveyor in a table.

## Task:6(The SQL Queries):

**Question: Construct a query to list ph_val and soil_name for series = 3 ordered by soil_name in alphabetical order.**

**Output:**

SELECT Soil_id.ph_val, soil_type.soil_name, soil_type.series

FROM Soil_id JOIN soil_type on Soil_id.soil_type = soil_type.soil_type

WHERE soil_type.series= 3 ORDER by soil_type.soil_name ASC

| | ph_val | soil_name | series |
|---|---|---|---|
| 1 | 2.9 | Clay | 3 |
| 2 | 3.43 | Clay | 3 |
| 3 | 3.44 | Clay | 3 |
| 4 | 4.52 | Clay | 3 |
| 5 | 4.65 | Clay | 3 |
| 6 | 4.32 | Glacial Sand | 3 |
| 7 | 5.34 | Glacial Sand | 3 |
| 8 | 5.43 | Glacial Sand | 3 |
| 9 | 6.44 | Glacial Sand | 3 |
| 10 | 6.54 | Glacial Sand | 3 |
| 11 | 2.36 | Gravel | 3 |
| 12 | 3.54 | Gravel | 3 |
| 13 | 4.56 | Gravel | 3 |
| 14 | 6.54 | Gravel | 3 |
| 15 | 3.34 | River Alluvium | 3 |
| 16 | 5.74 | River Alluvium | 3 |
| 17 | 2.22 | Sand | 3 |
| 18 | 3.33 | Sand | 3 |

**Interpretation:**

The ph_val, soil_name and series=3 to be listed in a table with Ascending order by joining the two tables Soil_id and soil_type with “join” clause.

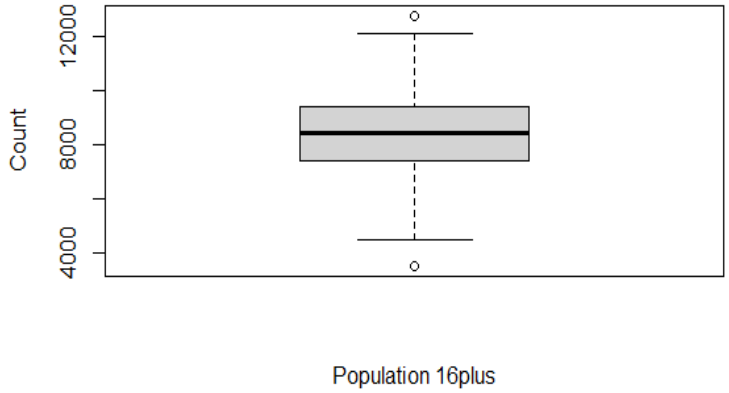
## SESSION-4

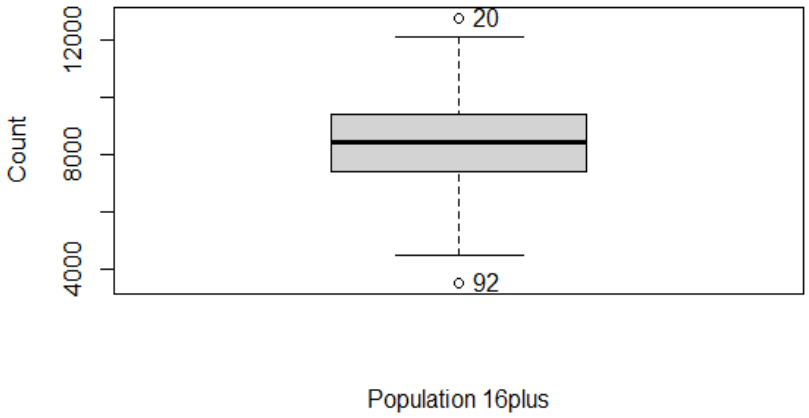
## Task 1: Data Exploration and Graphics:

## Section2:

**Question: Create a simple boxplot of variable P16plus which is the population aged 16 and over for each Ward. What do you find? Are there any outliers...are they errors? Label the outliers so you know which cases they are and then print the data for these two cases. Which wards are they?**

**Output:(Boxplot of population 16plus):**



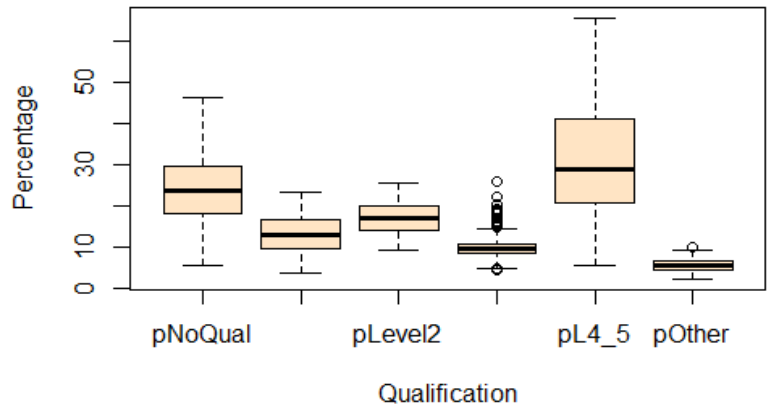
**Interpretation:**

First the 04_KS13N_London.csv File is loaded into the R Studio, where the csv files contains 12 variables with 624 objects. The 12 Variables are Region , Sub_region, Borough, Code, Ward, P16plus, NoQual, Level1, Level2, Level3,Level4_5, Other. Then we run the box plot for the variable "p16plus" is nothing but population 16plus.There are outliers in the Minimum Score and Maximum Score Area of the box plot. In the minimum Area the count of outliers are 92 and Maximum Area its 20. The outlier 92 were Bromley which is in outer London with the population count of 3552 which comes under Darwin ward. The outlier 20 were in Barnet which is in outer London with the population count of 12761 which comes under Hill ward.

**Section6:**

**Question: Create boxplots to explore all six percentage variables. What do you note? Now do the same for pNoQual by Sub_region and then by Borough. What seems to be the pattern of pNoQual across London?**

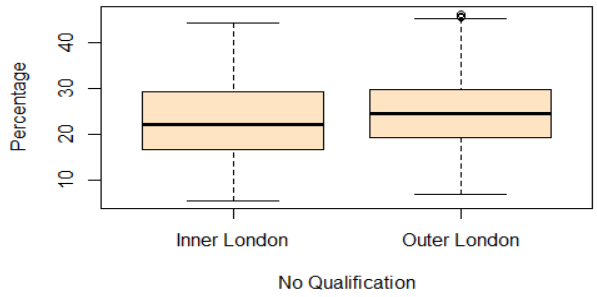
**Output: [Boxplot for All Six Percentage Variables]**




**Interpretation:**

In the Above Boxplot of All Six Percentage Variables , the pLevel3 has more outliers in the Maximum range area and few outlier in Minimum Area of plevel3 Boxplot.

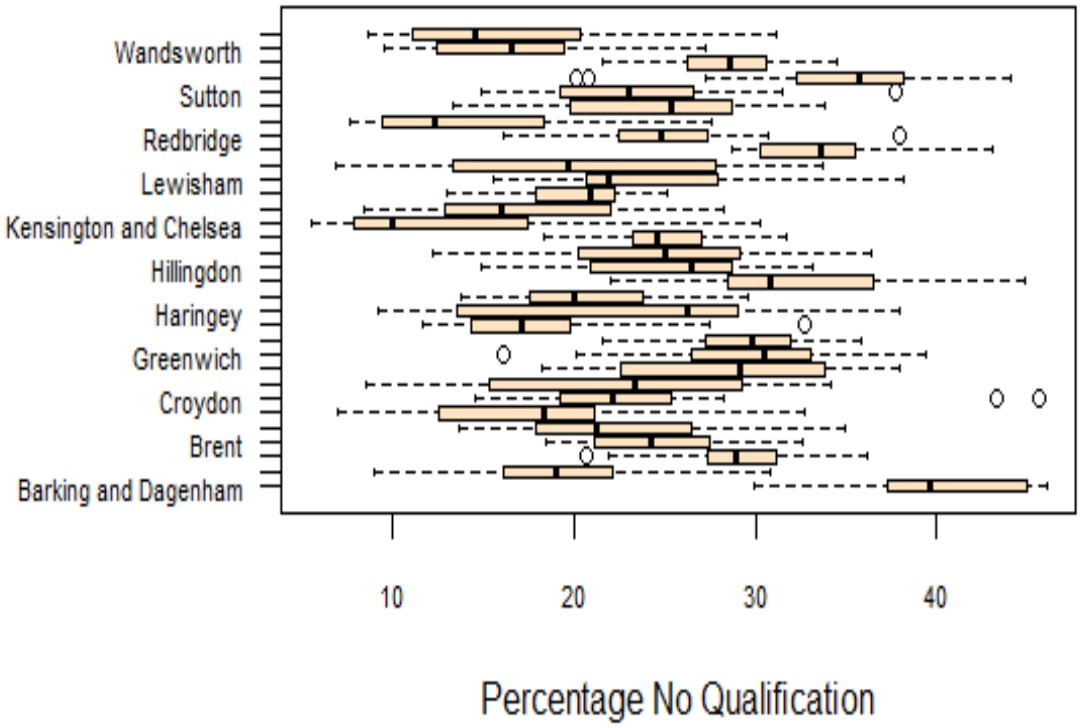
**[Box plot for PNoQual by Sub_region]**




**Interpretation:**

In PNoQual Sub_Region both Inner London and Outer London Are Equal but the outer London has few outliers in its Maximum Range Area which is above Q3.

**[Box plot for PNoQual by Borough]**



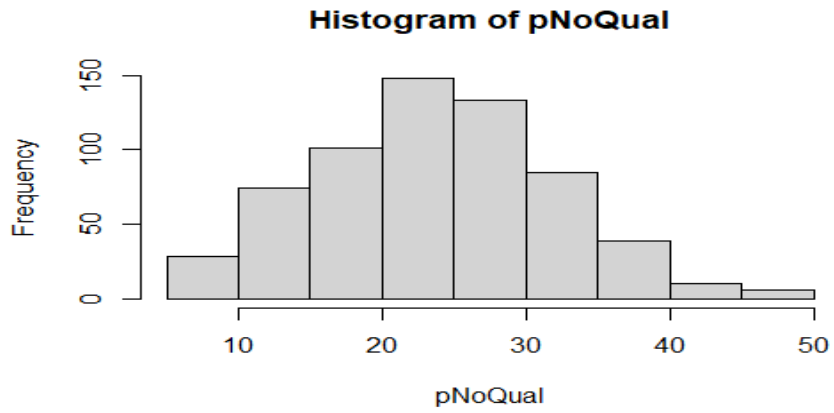

**Interpretation:**

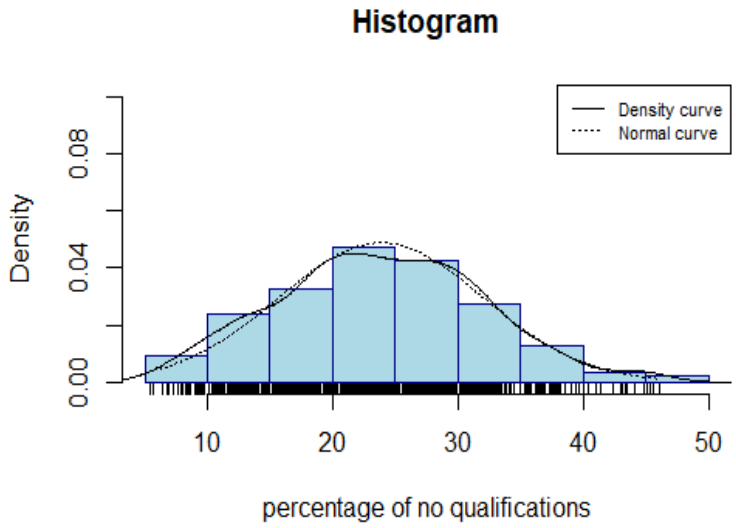
In Above boxplot it shows London Boroughs Qualification population Percentage which is stored in Borough Variable , Where Barking and Dagenham has large Amount of population with No Qualification which is 40%, the Kensington and Chelsea has small amount of population with no Qualification.

**Section7:**

**Question: Create histograms (frequency and probability density) for pNoQual. Add the density curve to the probability density histogram and the normal curve for comparison**.

**Output: Histograms (frequency and probability density) for pNoQual:**




**Interpretation(Histogram frequency):**

Using hist() function , the histogram is plotted for pNoQual Variable with Frequency in the plot. The percentage of no qualification over the community frequency can be calculated using the aforementioned histogram. For instance, 20–25% of the community in close to 150 has no qualifications.

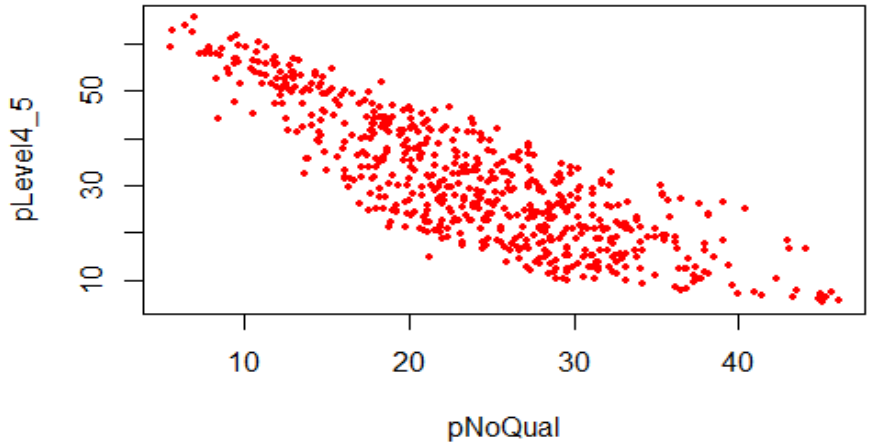
**Interpretation(Histogram Probability Density for pNoQual):**

The density for no qualification can be determined from the aforementioned density histogram; the dotted curve denotes the normal deviation, while the line curve denotes the density, which also explains why both were variable at some locations despite having the same mean and standard deviation.

**Section8:**

**Question: Plot a basic scatter graph of pNoQual and pLevel4_5 and a second variant of it. What sort of relationship do these two variables have?**

**Output:[scatter plot]:**



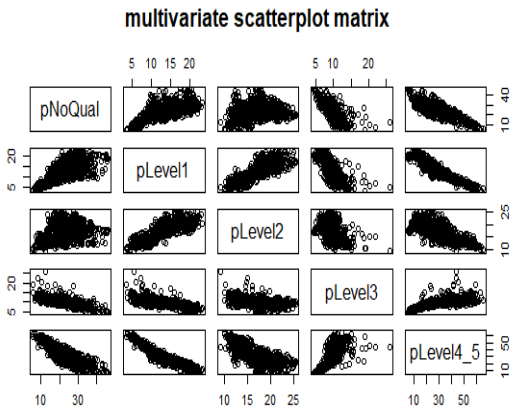

**Interpretation:**

The Co-relation between two variables can be seen by using Scatter Plot. There are More people with the University Qualification and there are Less People With No Qualification.

**Section10:**

**Question:Create a multivariate scatter plot for the various qualification levels. Inspect the strength of relationship between each variable and the others. Which pair of variables seems to have the strongest positive correlation and which have the strongest negative correlation?**

**Output:**




```
> # basic correlation matrix
> cor(London2, method = "spearman")
            pNoQual    pLevel1    pLevel2    pLevel3  pLevel4_5
pNoQual   1.0000000  0.6578097  0.2564265 -0.7430328 -0.8256637
pLevel1   0.6578097  1.0000000  0.8339219 -0.7135164 -0.9457736
pLevel2   0.2564265  0.8339219  1.0000000 -0.3871209 -0.7110420
pLevel3  -0.7430328 -0.7135164 -0.3871209  1.0000000  0.7376875
pLevel4_5 -0.8256637 -0.9457736 -0.7110420  0.7376875  1.0000000
> London_cor <- cor(London2, method = "spearman")
> round(London_cor, digits = 2)
          pNoQual pLevel1 pLevel2 pLevel3 pLevel4_5
pNoQual      1.00     0.66     0.26    -0.74     -0.83
pLevel1      0.66     1.00     0.83    -0.71     -0.95
pLevel2      0.26     0.83     1.00    -0.39     -0.71
pLevel3     -0.74    -0.71    -0.39     1.00      0.74
pLevel4_5   -0.83    -0.95    -0.71     0.74      1.00
>
```

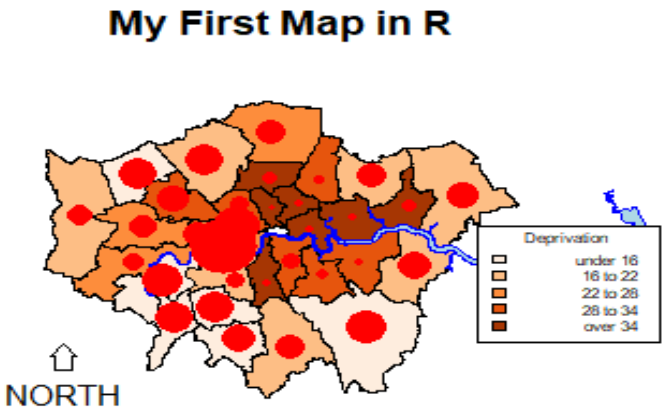
**Interpretation:**

In the Multivariate Scatter Plot, the Co relaltion matrix was used to find the Relationship Between the Variables. The pLevel1 and pLevel2 Variables  are positively highly Co-related.

**Task 2: Data Exploration and Graphics:**

**Section5: Plot LIFE_MALE as proportional circles on top. What do you conclude about the spatial relationship of deprivation to male life expectancy?**

**Output:**




**Interpretation:**

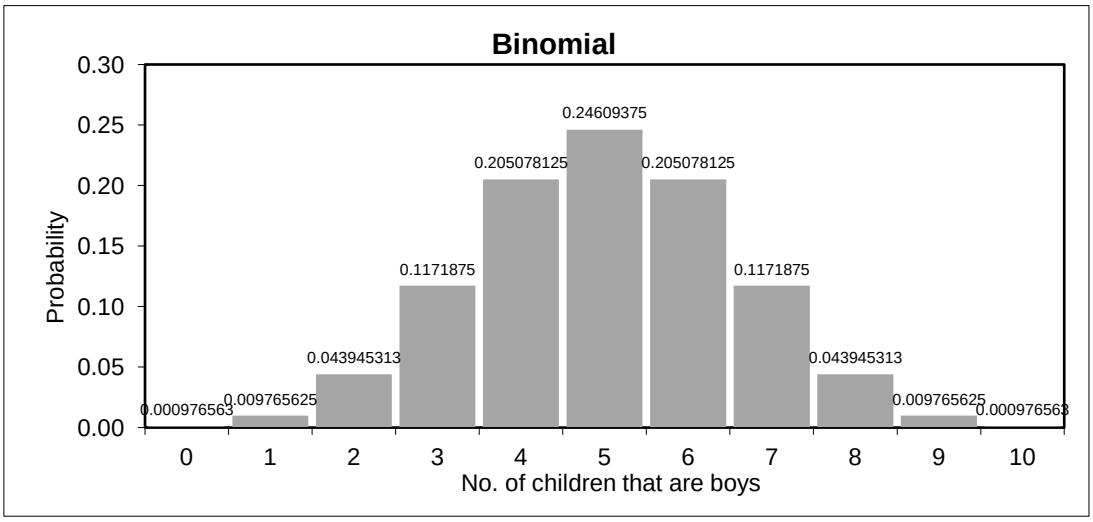
The thematic map was executed  by including variable and adjusting cex value. From the above map we can say that the circle represents the male life expectancy. The small red circle represents low life expectancy and large red circle represents the high life expectancy for male.

## SESSION-5

**Task1: Probability distributions:**

**Question: Jack and Jill grow up, fall in love and elope. They decide to have 10 children. If the probability of each pregnancy resulting in a boy is 0.5, what is the probability that they will have 7 boys?**

**Output:**



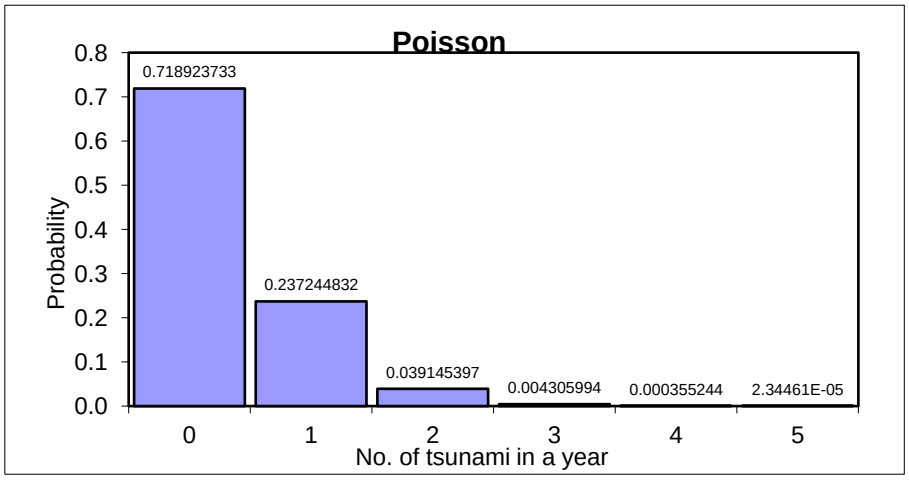

**Interpretation:**

First we load the data into Excel Sheet , after that we use Binomial Distribution Concept to find the probability of having a boy in 10 Pregnancies. The Probability to have 5 Boys has the High Possibility value of 0.246. The Probability to have 7 boys is 0.1171875 as mentioned in the above Graph. Hence the final or high chance of probability would be to have 5 Girls and 5 boys in the family.

**Task2: Probability distributions:**

**Question: Jack wants to go on a diving holiday in the Pacific. If there is on average one tsunami every three years, what is the probability of one tsunami in any year?**

**Output:**



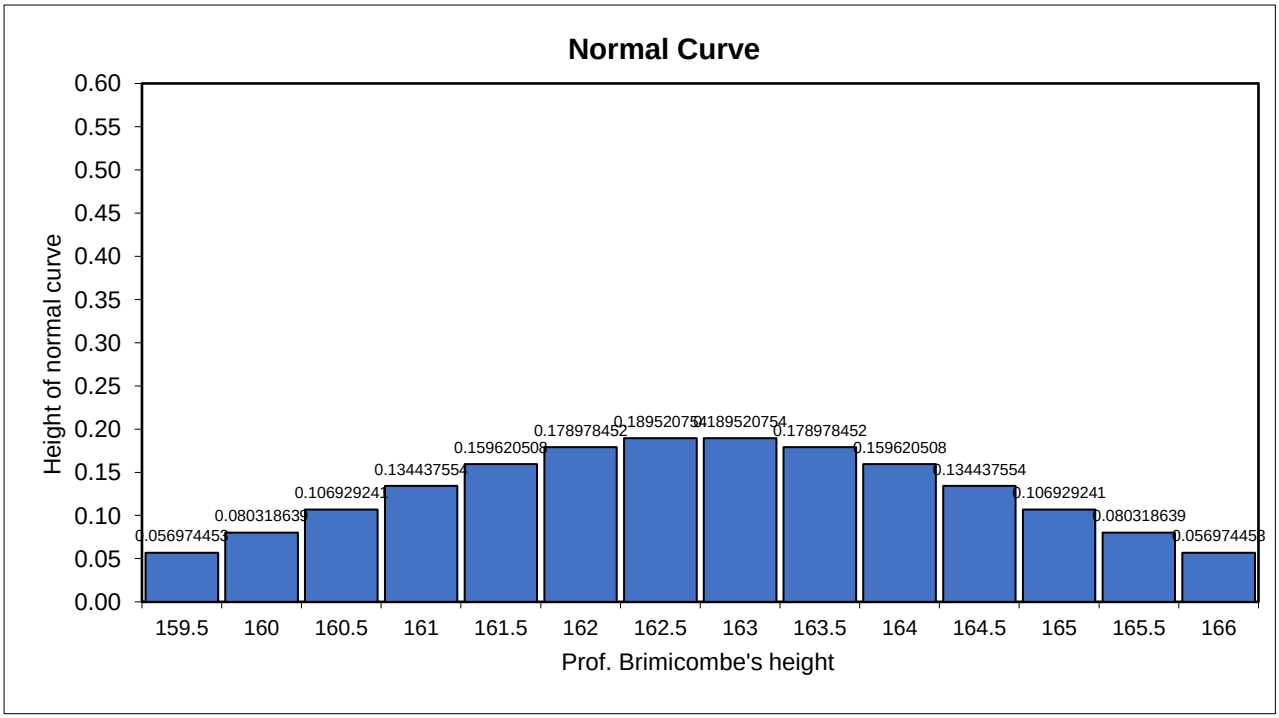
**Interpretation:**

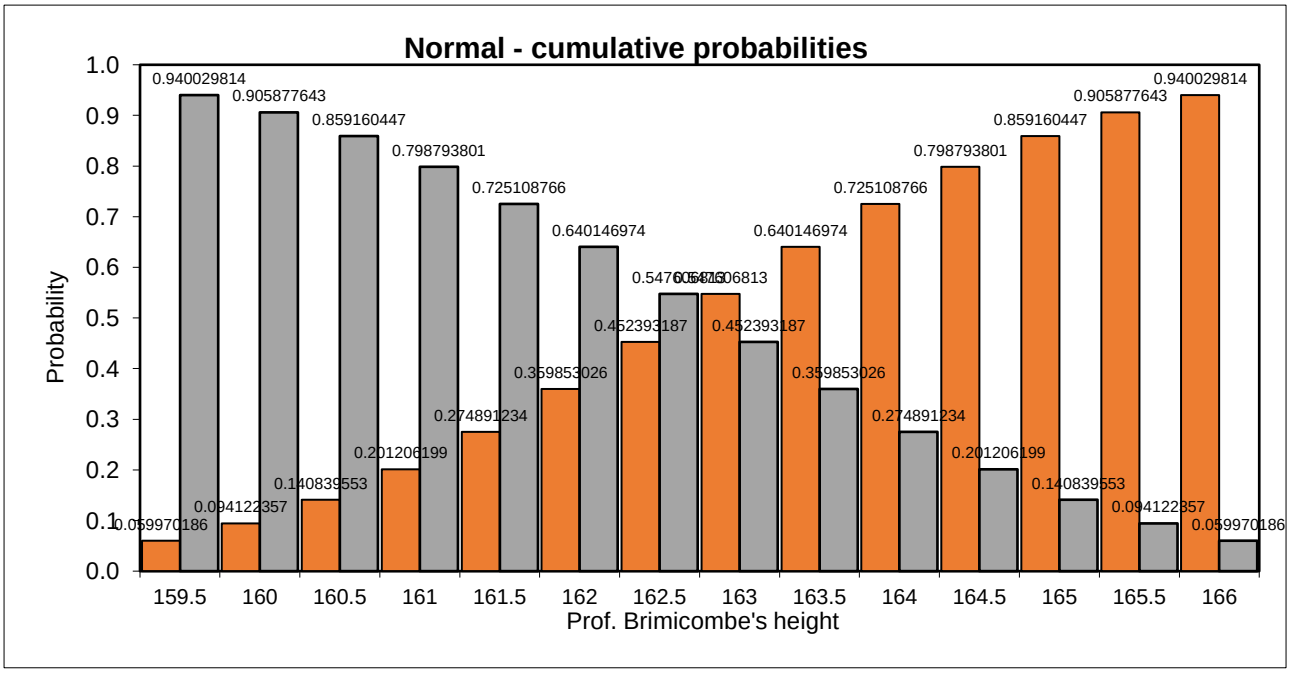
Here in this case Occurrence of Tsunami and Where it comes under the Rare Events, So in rare Event Cases we use Poisson Distribution Concept in Excel to find the probability of 1 Tsunami in one year. The Probability to have No Tsunami in a year is 0.7189. The Probability to have 1 Tsunami in a year is 0.2372. Finally the conclusion part is to have No Occurrence of Tsunami in a year because it has High Probability in the Poisson chart.

**Task3: Probability distributions:**

**Question: What is the probability that he is taller than 164cm?**

**Output:**



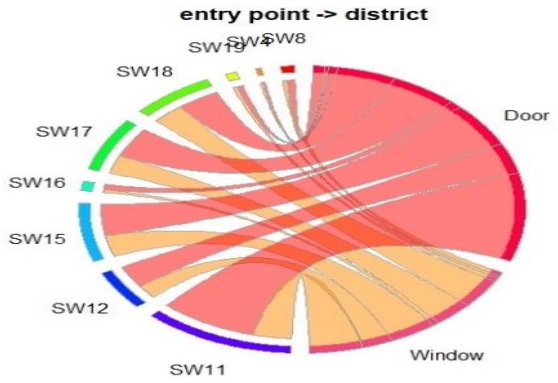
**Interpretation:**

The probability of taller than 164 , among the four possible outcomes where among those four possibilities, the 164.5 has the highest possible of outcome , where it holds value of 0.201 which greater than other outcomes.

## SESSION-6

## Section1&Section2:[ Hypothesis Testing using R, using Chi-Squared, Wilcoxon, and Mann-Whitney U]:

## Output:




RStudio Source Editor

burglary.chi

Filter

| | id | day | district | dwell_ty | entry_pn |
|---|---|---|---|---|---|
| 1 | 1 | THU | SW18 | Semi | Other |
| 2 | 2 | THU | SW17 | Flat_Mais | Door |
| 3 | 3 | THU | SW8 | Flat_Mais | Window |
| 4 | 4 | FRI | SW15 | Council | Door |
| 5 | 5 | FRI | SW15 | Flat_Mais | Door |
| 6 | 6 | FRI | SW17 | | Door |
| 7 | 7 | FRI | SW15 | Bungalow | Door |
| 8 | 8 | FRI | SW18 | Bungalow | Window |
| 9 | 9 | FRI | SW11 | Bungalow | Door |
| 10 | 10 | FRI | SW15 | | |
| 11 | 11 | FRI | SW15 | Detached | Door |
| 12 | 12 | FRI | SW18 | Flat_Mais | Window |
| 13 | 13 | FRI | SW18 | Terraced | Window |
| 14 | 14 | FRI | SW15 | Flat_Mais | Door |
| 15 | 15 | FRI | SW15 | Terraced | Door |
| 16 | 16 | SAT | SW18 | Bungalow | Door |
| 17 | 17 | SAT | SW18 | Terraced | Door |
| 18 | 18 | SAT | SW15 | Terraced | |
| 19 | 19 | SAT | SW17 | | |
| 20 | 20 | SAT | SW17 | Flat_Mais | Door |
| 21 | 21 | SUN | SW11 | Council | Door |
| 22 | 22 | SUN | SW11 | Flat_Mais | Door |
| 23 | 23 | SUN | SW11 | Multi_Occ | Door |

**Interpretation:**

Download the data files 06_burglary-chi.csv and 06_A&E_2003.csv and the script file 06_Hypothesis testing (non-parametric).Load the 06_burglary-chi.csv into R Script file using R Studio. Then we inspect the top or head of data's in the Csv file. The door and window columns were added to the table by district column after the cross table for entry pn was built, revealing the basic connection between entry pn and district. Then we use Package called Desc Tools to Visualize. The Above Mentioned diagram is Circular Plot. In Circular plot Sw11 Holds the Maximum doors in District Entry. Sw17 Holds the Minimum doors in the District Entry.

## Section3:[ Hypothesis Testing using R, using Chi-Squared, Wilcoxon, and Mann-Whitney U]:

## Output:

```
> PlotCirc(t(as.matrix(mytable)), main = "day -> district")
> # simple chi-squared test
> chisq.test(mytable)

        Pearson's Chi-squared test

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

> # simple chi-squared with additional outputs
> chisq_out <- chisq.test(mytable)
> chisq_out

        Pearson's Chi-squared test

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

> chisq_out$observed
```

**Interpretation (Chi-Squared Test ):**

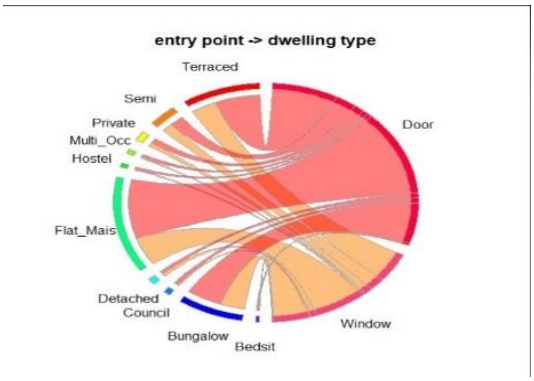
The Results of hypothesis could be found from the Z-squared values of DF=8,p_Value=6.33 e-09, by using Chi-Squared Test. Monte Carlo simulation was Executed results in giving the same output to basic method based on rescaling. Then Effect Size was Calculated to find Cramers V Size. As from the output it Clearly Says the p value is relatively higher and so we can Reject the Null Hypothesis. So this says there is huge important difference between window Entry and Door by District.

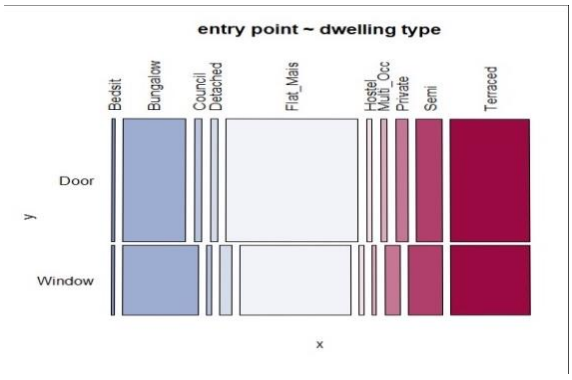
**Interpretation (Monte Carlo Simulation & Cramer's V):**

As per the Value P=0.0004998 ,smaller is compared to the basic method, results in chance of getting accepted is less.

## Section4:[ Hypothesis Testing using R, using Chi-Squared, Wilcoxon, and Mann-Whitney U]:

## Output:




## Output:(chi-squared test, Monte carlo &cramers'v):

```
Multi_Occ   50     17
Private     87     61
Semi       186    140
Terraced   569    322
> # visualise
> plotMosaic(t(as.matrix(mytable)), main = "entry point ~ dwelling type")
> plotCirc(t(as.matrix(mytable)), main = "entry point -> dwelling type")
> # simple chi-squared test
> chisq.test(mytable)

        Pearson's Chi-squared test

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

> # simple chi-squared with additional outputs
> chisq_out <- chisq.test(mytable)
> chisq_out$observed
          Door Window
```

## Output:

```
Private   -1.29729669  1.29729669
Semi      -2.65227043  2.65227043
Terraced   0.03687946 -0.03687946
> # chi-squared test using Monte Carlo simulation
> chisq.test(mytable, correct = false,
+            p = rep(1/length(mytable), length(mytable)), rescale.p = FALSE,
+            simulate.p.value = TRUE, B = 2000)

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

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

> # calculate effect size
> cramerv(mytable, conf.level = 0.95)
  Cramer V      lwr.ci      upr.ci
0.09786797 0.05192032 0.11983891
> rm(myxtab, mytable, chisq_out)
> fisher.test(mytable, simulate.p.value = TRUE)
```

### Interpretation of Circular plot & (Chi-Squared Test ):

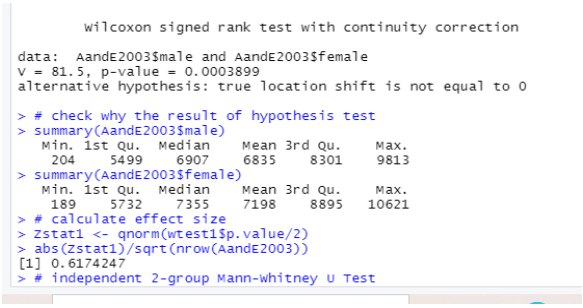
Here in the graph Flat_mais holds the major part in window Entry and Door Entry,where Terraced has medium Door entry and Window Entry when compared to flat_mais. In Chi squared test , the p value is Greater than 95 Percent , so the Null hypothesis is Rejected.Finally there is much noted Difference between Window and Door Entry in Dwelling Entry.

### Interpretation (Monte Carlo Simulation & Cramer's V):

As per the Value P=0.0004998 ,smaller is compared to the basic method, results in chance of getting accepted is less. Hence it is Rejected.

**Section6:(Wilcoxon signed rank test):**

**Output: (Wilcoxon Signed Rank Test ):**



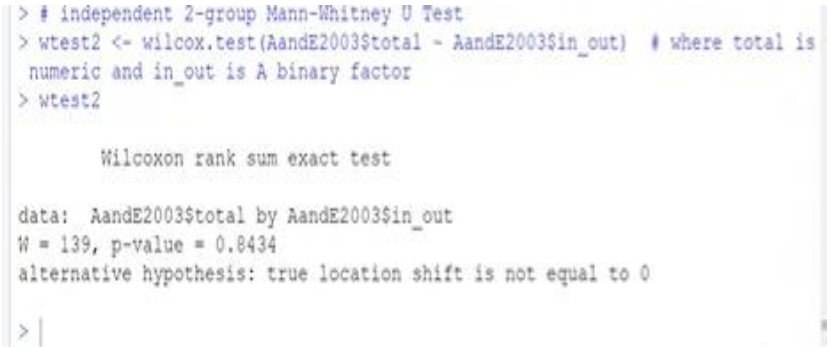
```
        Wilcoxon signed rank test with continuity correction

data:  AandE2003$male and AandE2003$female
V = 81.5, p-value = 0.0003899
alternative hypothesis: true location shift is not equal to 0

> # check why the result of hypothesis test
> summary(AandE2003$male)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    204    5499    6907    6835    8301    9813
> summary(AandE2003$female)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    189    5732    7355    7198    8895   10621
> # calculate effect size
> Zstat1 <- qnorm(wtest1$p.value/2)
> abs(Zstat1)/sqrt(nrow(AandE2003))
[1] 0.6174247
> # independent 2-group Mann-Whitney U Test
```

**Interpretation:**

First we have to load the load the 06_A&E Csv file into session 6 R script file using R studio and then Evaluation for male and female attendance that applied for the Wilcoxon signed rank test has been done. The, effect size was calculated by dividing the Standardized Z-score by Sq. of n. In the Above output p value is greater than 95percent , so it is safer to Reject the null hypothesis. By using Z-Score method the Effect Size was high , where the value is 0.6174247.

**Section7:(Mann-Whitney U Test):**

**Output:**



```
> # independent 2-group Mann-Whitney U Test
> wtest2 <- wilcox.test(AandE2003$total ~ AandE2003$in_out)  # where total is
 numeric and in_out is A binary factor
> wtest2

        Wilcoxon rank sum exact test

data:  AandE2003$total by AandE2003$in_out
W = 139, p-value = 0.8434
alternative hypothesis: true location shift is not equal to 0

>
```

**Interpretation:**

In the output of Mann-Whitney U test ,the Null Hypothesis is Accepted, Because the p value is 0.8434, where p value is High. So we Conclude that there is no big Difference Between the Outer and Inner London as Independent Variables.

## Section8:(Hypothesis with Proportional data):

## Output: (Wilson Signed Rank Test (female and male population):

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    204    5499    6907    6835    8301    9813
> summary(AandE2003$female)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    189    5732    7355    7198    8895   10621
> # calculate effect size
> Zstat1 <- qnorm(wtest1$p.value/2)
> abs(Zstat1)/sqrt(nrow(AandE2003))
[1] 0.6174247
> # independent 2-group Mann-Whitney U Test
> wtest2 <- wilcox.test(AandE2003$total ~ AandE2003$in_out)  # where total is numeric and
 in_out is A binary factor
> wtest2

        Wilcoxon rank sum exact test

data:  AandE2003$total by AandE2003$in_out
W = 139, p-value = 0.8434
alternative hypothesis: true location shift is not equal to 0
```

## Interpretation:

The Purpose to Choose the Proportional Data because data holds the Large Population and small population of London Boroughs. In the Above output the p value is smaller , where there is High Difference Between Female and Male Proportional Data. Since the female population was so large, there is no Much difference in proportion.

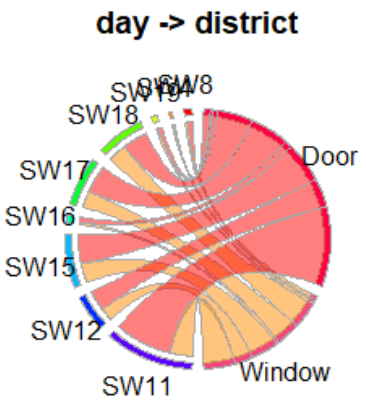
## Output:Mann-Whitney U Test (Inner and outer London):

```
> mydata.inner<-AandE2003[AandE2003$in_out=="Inner",]
> mydata.outer<-AandE2003[AandE2003$in_out=="Outer",]
> 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
> # calculate effect size
> Zstat4 <- qnorm(wtest4$p.value/2)
> abs(Zstat4)/sqrt(nrow(AandE2003))
[1] 0.4487897
>
```

## Interpretation:

The Effect Size Value is 0.04487897. We can Conclude that there is big Difference between Outer and Inner London by proportion , were the p value is Lesser than 1 From the output.

## Output:(Chi-Squared Test For Day of the Week):



**Interpretation:**

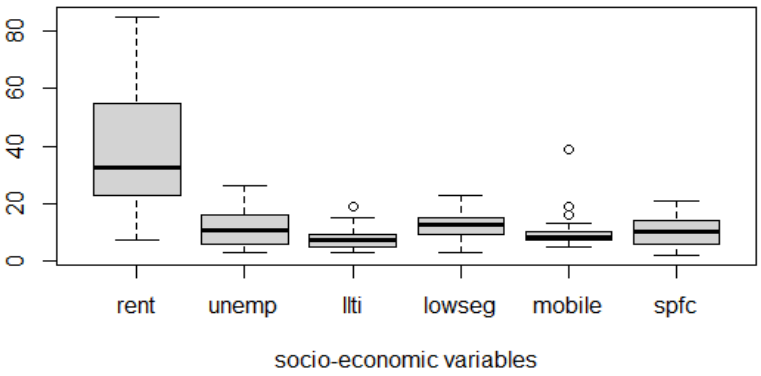
In the above circular plot we can conclude that the SW11 has maximum door entry and minimum window entry in a day of district, where S15 has the minimum door entry and minimum window entry in a day of district.

## SESSION 7

**Section 2&3:( Hypothesis testing (Parametric):**

**Question: Normality of all variables:**

**Output:(Boxplot for all the 6 variables):**



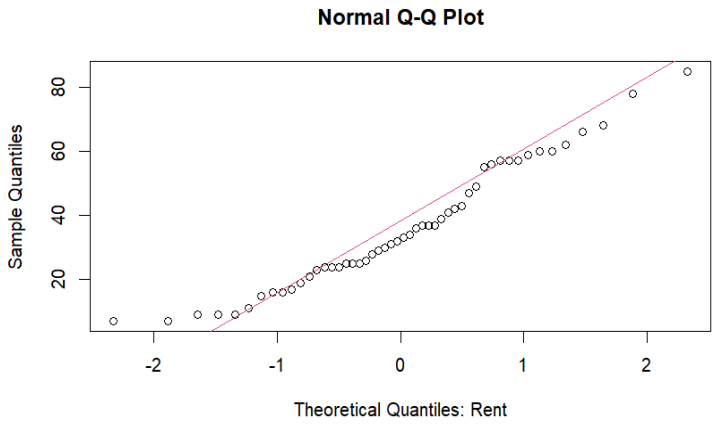

**Interpretation:**

First we have to upload the 07_pcs_sample.sav csv file into Session 7 R Script File using R studio. Where it Consist of 7 Variables with 50 Rows of data. The variables are the percentage of households having rented accommodation (rent), unemployment (unemp), limiting long-term illness (llti), low segment employment (lowseg), having recently moved location (mobile) and single parent families with children (spfc) for postcode areas in the north of England. (Area) Variable to be removed in the Early part of the stage. We make boxplot to check whether the variables of the data are normally Distributed. The variable rent is normally distributed ,where mobile variable has few outliers in the maximum area and above the maximum area in the boxplot and its not normally distributed . Unemp and spfc has no outliers in it. Llti variable have a outlier in the maximum part of the boxplot and its not normally distributed.

**Section 4:**

**Question: (QQ plot for all Six Variables):**

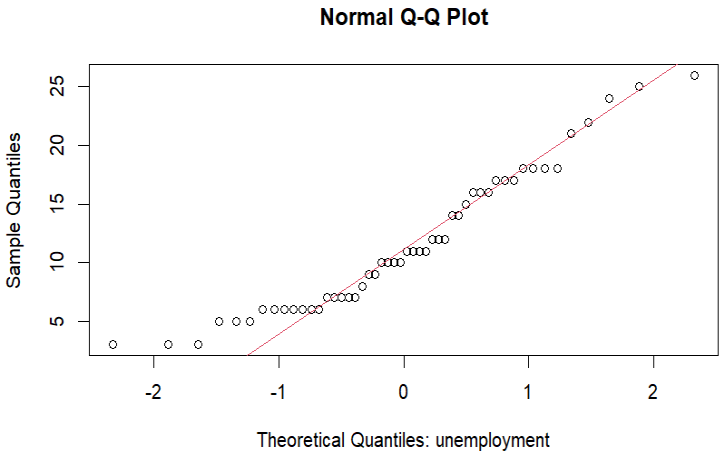
**Output: (QQ plot for Variable Rent):**




**Interpretation:**

The Dots are Data points in the QQ plot are Slightly Differ from the Diagonal line( Normal Distribution Line)and where Diagonal line is above 35 Degree which Says the Data in the Rent are Normally distributed.

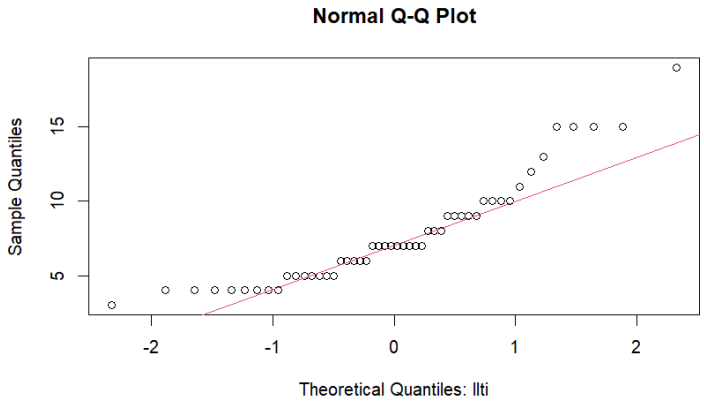
**Output: (QQ plot for Variable unemp):**




**Interpretation:**

The Dots are Data points in the QQ plot are Almost lies in the Diagonal line( Normal Distribution Line)and where Diagonal line is above 35 Degree which Says the Data in the Unemployment are Normally distributed.

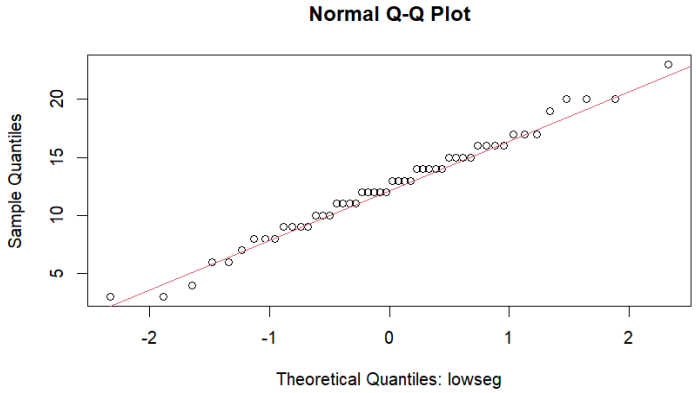
**Output: (QQ plot for Variable Llti):**




**Interpretation:**

The Dots are Data points in the QQ plot doesn't lies in the Diagonal line(Normal Distribution Line)and where Diagonal line is less than 35 Degree which Says the Data in the Llti are not Normally distributed. The Data points which differs from the diagonal line are said to be outliers.

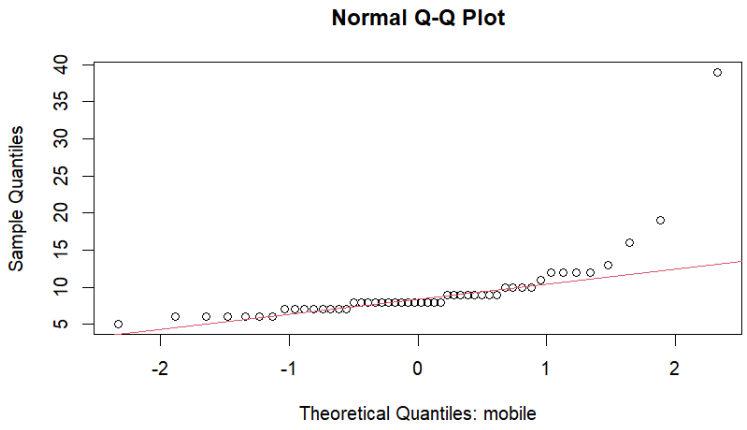
**Output: (QQ plot for Variable lowsleg):**




**Interpretation:**

The Dots are Data points in the QQ plot are Almost lies in the Diagonal line( Normal Distribution Line)and where Diagonal line is above 35 Degree which Says the Data in the lowsleg are Normally distributed.

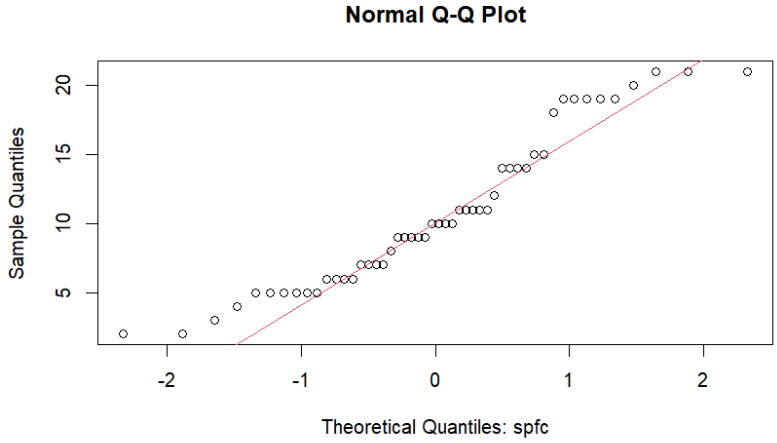
**Output: (QQ plot for Variable Mobile):**




**Interpretation:**

The Dots are Data points in the QQ plot where few data points lies in the Diagonal line(Normal Distribution Line)and where Diagonal line is less than 35 Degree which Says the Data in the mobile are not Normally distributed. The Data points which differs from the diagonal line are said to be outliers.

**Output: (QQ plot for Variable spfc):**



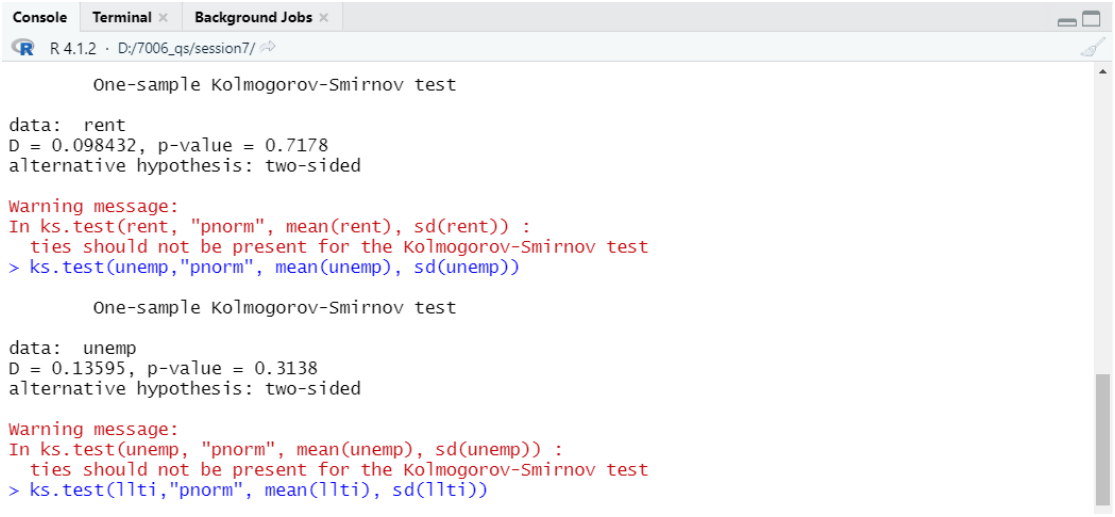

**Interpretation:**

The Dots are Data points in the QQ plot are Almost lies in the Diagonal line( Normal Distribution Line)and where Diagonal line is above 35 Degree which Says the Data in the spfc are Normally distributed.

**Section 5(Kolmogorov-Smirnov):**

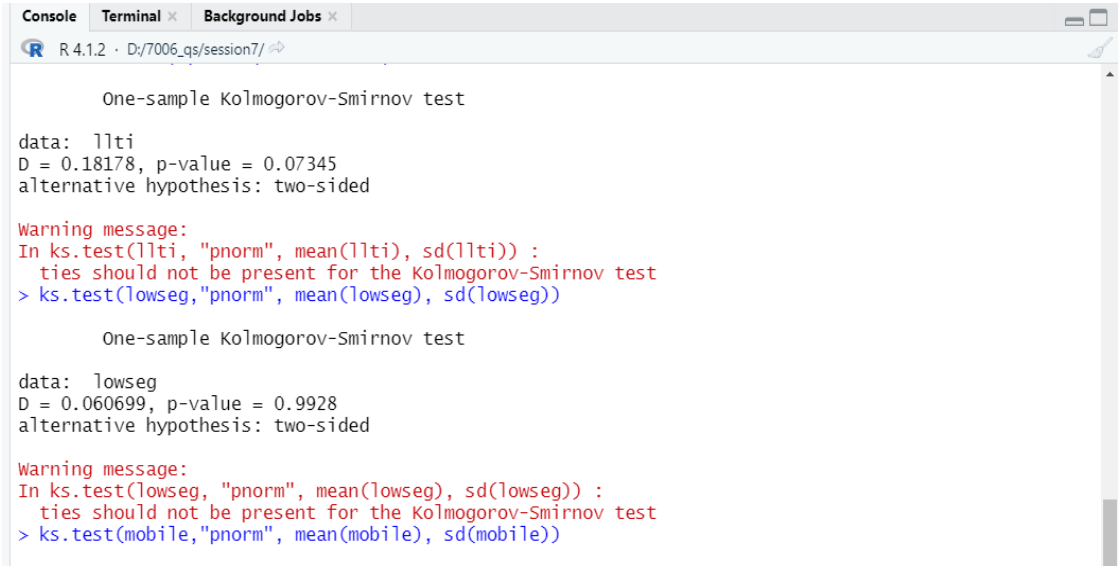
**Question: Which variable(s) can be safely rejected as not normally distributed and on what grounds? An alternative test is the Shapiro-Wilkes and should give you an equivalent result.**

## Output: (Ks test for Rent Var and Unemp) :



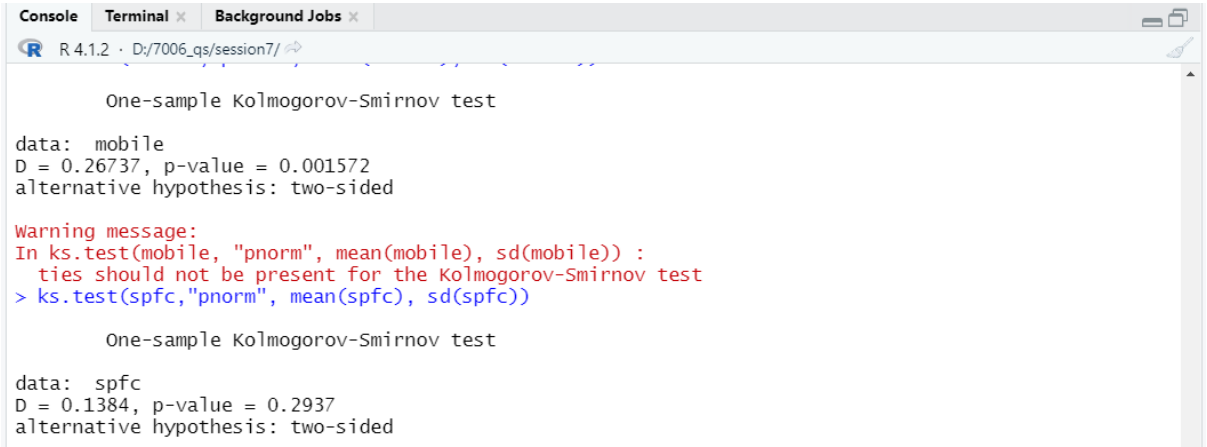
```
Console  Terminal  Background Jobs
R 4.1.2 · D:/7006_qs/session7/

        One-sample Kolmogorov-Smirnov test

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

Warning message:
In ks.test(rent, "pnorm", mean(rent), sd(rent)) :
  ties should not be present for the Kolmogorov-Smirnov test
> ks.test(unemp,"pnorm", mean(unemp), sd(unemp))

        One-sample Kolmogorov-Smirnov test

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

Warning message:
In ks.test(unemp, "pnorm", mean(unemp), sd(unemp)) :
  ties should not be present for the Kolmogorov-Smirnov test
> ks.test(llti,"pnorm", mean(llti), sd(llti))
```

## Output: (Ks test for Llti and lowseg):



```
Console  Terminal  Background Jobs
R 4.1.2 · D:/7006_qs/session7/

        One-sample Kolmogorov-Smirnov test

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

Warning message:
In ks.test(llti, "pnorm", mean(llti), sd(llti)) :
  ties should not be present for the Kolmogorov-Smirnov test
> ks.test(lowseg,"pnorm", mean(lowseg), sd(lowseg))

        One-sample Kolmogorov-Smirnov test

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

Warning message:
In ks.test(lowseg, "pnorm", mean(lowseg), sd(lowseg)) :
  ties should not be present for the Kolmogorov-Smirnov test
> ks.test(mobile,"pnorm", mean(mobile), sd(mobile))
```

## Output: (Ks test for Mobile and spfc):



```
Console  Terminal  Background Jobs
R 4.1.2 · D:/7006_qs/session7/

        One-sample Kolmogorov-Smirnov test

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

Warning message:
In ks.test(mobile, "pnorm", mean(mobile), sd(mobile)) :
  ties should not be present for the Kolmogorov-Smirnov test
> ks.test(spfc,"pnorm", mean(spfc), sd(spfc))

        One-sample Kolmogorov-Smirnov test

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

**Interpretation:**

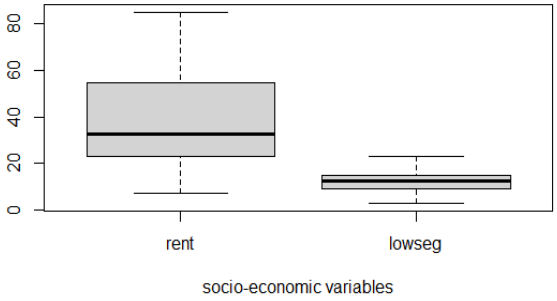
Kolmogorov-Smirnov test has been done for all the six variables.

| Variables | P-Values |
|---|---|
| Rent | 0.7178 |
| unemp | 0.3138 |
| Llti | 0.7345 |
| lowseg | 0.9928 |
| Mobile | 0.001572 |
| Spfc | 0.2937 |

By observing the p values in the above table we can say that Rent, unemp, Llti, lowseg, spfc variables has not much important difference , So its normally distributed. Hence we cant reject it. But by observing the p value of mobile we can say that there is important or high difference, so its not normally distributed. Hence we can conclude that its safe to reject mobile variable.

**Section 6(t-test on rent and lowseg):**

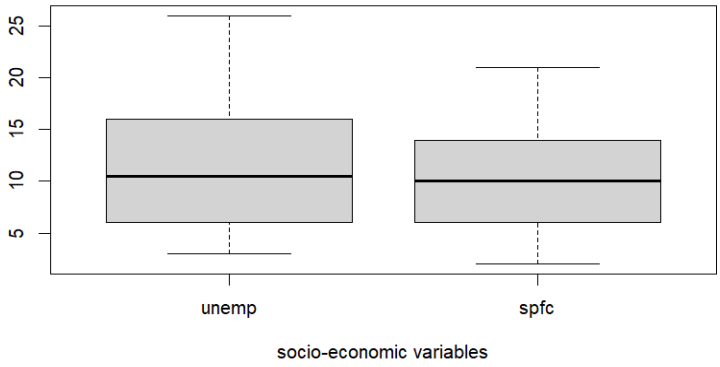
**Output:(Boxplot of rent and lowseg):**




**Output: (T-test of Rent and lowseg):**

```
Console   Terminal ×   Background Jobs ×
R 4.1.2 · D:/7006_qs/session7/

> # inspect if variance is equal
> var(rent)
[1] 384.0576
> var(lowseg)
[1] 19.88408
> # paired t test with unequal variance
> 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 difference in means is not equal to 0
95 percent confidence interval:
 18.64865 28.35135
sample estimates:
mean of the differences
                  23.5

> # check boxplots
> boxplot(rent, lowseg, names=c("rent", "lowseg"), xlab="socio-economic variables")
> library(lsr)
> cohensD(rent, lowseg, method = "unequal")
[1] 1.653574
```

**Output:(Boxplot of unemp and spfc):**




**Output: (T-test of unemp and spfc):**

```
Console   Terminal ×   Background Jobs ×
R 4.1.2 · D:/7006_qs/session7/
> # inspect if variance is equal
> var(unemp)
[1] 36.45265
> var(spfc)
[1] 30.66327
> # paired t test with unequal variance
> t.test(unemp, spfc, paired=TRUE, var.equal = FALSE)

        Paired t-test

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

> # check boxplots
> boxplot(unemp,spfc, names=c("unemp", "spfc"), xlab="socio-economic variables")
> # dependent 2-group Wilcoxon Signed Rank Test
> wilcox.test(rent, lowseg, paired = TRUE) # where rent and lowseg are numeric

        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
```

**Interpretation:**

Here in the output , the variance has important Difference after doing Paired t- test where Rent variable has a value 384.0576 and lowseg Variable has a value 19.88408. The mean difference for rent and lowseg is 23.5 during the check of Hypothesis outcome. The hypothesis was safely rejected because the p value is lesser than 95 percent. There is no much important difference for the variable unemp and spfc.

## Section 7:(2-group Wilcoxon Signed Rank Test):

## Output: (Wilcoxon Rank test for rent and lowsleg):

```
Console  Terminal  Background Jobs
R 4.1.2 · D:/7006_qs/session7/
> # dependent 2-group Wilcoxon Signed Rank Test
> wilcox.test(rent, lowseg, paired = TRUE) # where rent and lowseg are numeric

        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
```

## Output: (Wilcoxon Rank test for unemp and spfc):

```
Console  Terminal  Background Jobs
R 4.1.2 · D:/7006_qs/session7/
> wilcox.test(unemp, spfc, paired = TRUE)  # where unemp and spfc are numeric

        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
```

## Interpretation:

Rent and lowseg's hypothesis is rejected because it has low p value , so where it has important difference. In other pairs of unemp and spfc , cannot be rejected because it is not normally distributed where there is important Difference.

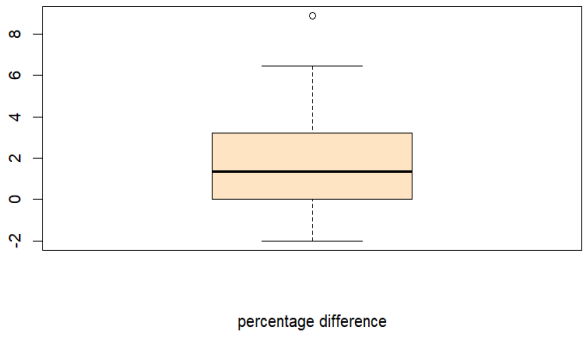
## SESSION 8

## Task2:( To carry out an investigation of appropriate data transformation in order to use a t-test correctly):

## Section 5:

## Question: (Create a further variable (p.diff) which is the percentage difference between b_freeze and a_freeze. Also boxplot and QQ plot to inspect for normality. Carry out a KS test of normality. Does this variable now appear to be normal?

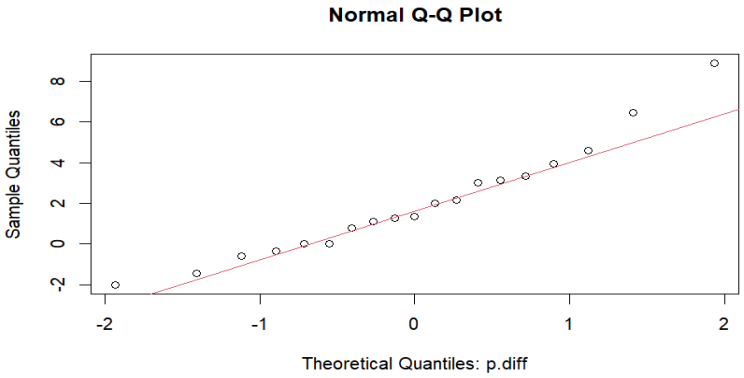
## Output: (Boxplot of Percentage Difference):



## Output:(bfreeze, afreeze, pdifference):

```
3       3    20.56    19.92   0.64
4       4  4076.00  3954.00 122.00
5       5    14.69    13.74   0.95
6       6    23.65    23.65   0.00
> boxplot(blood.tests$diff, xlab = "difference", col = "bisque")
> qqnorm(blood.tests$diff, xlab = "Theoretical Quantiles: diff")
> qqline(blood.tests$diff, col=2)  # red color
> blood.tests <- within(blood.tests, p.diff <- (blood.tests$diff / b_freeze)*100)
> head(blood.tests)
  sample b_freeze a_freeze   diff     p.diff
1      1     7.19     6.55   0.64  8.9012517
2      2    60.64    60.85  -0.21 -0.3463061
3      3    20.56    19.92   0.64  3.1128405
4      4  4076.00  3954.00 122.00  2.9931305
5      5    14.69    13.74   0.95  6.4669843
6      6    23.65    23.65   0.00  0.0000000
> boxplot(blood.tests$p.diff, xlab = "percentage difference", col = "bisque")
> qqnorm(blood.tests$p.diff, xlab = "Theoretical Quantiles: p.diff")
> qqline(blood.tests$p.diff, col=2)  # red color
> ks.test(blood.tests$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
```

## Output: (QQplot):




## Output(T-test & paired T-test):

```
> # T Tests
> # correct result
> 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
```

```
> t.test(a_freeze, b_freeze, paired = T)       # Paired Samples t-tests

        Paired t-test

data:  a_freeze and b_freeze
t = -1.0185, df = 18, p-value = 0.322
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -20.071319   6.965003
sample estimates:
mean difference
      -6.553158
```

**Interpretation:**

The 08_Blood_tests.csv File is uploaded to the session 8 r script file using r studio. The csv file contains 3 attributes with 19 rows. The attributes are sample, b_freeze and a_freeze. we analyse the b_freeze and a_freeze, where data's are not normally distributed and we check it by using boxplot. After that QQ plot was created for both Variables , where both variables data are not normally distributed. Then new variable Diff is added where, it is also not normally distributed. Then at last new variable is added to the data set where it holds the data percentage between b_freeze and a_freeze. We can conclude from the output that it is not normal and in the ks-test there is no big important Difference. In t-test, I clearly understood that it is normally distributed , where it is similar to ks test.

**Task3:( To carry out analysis of variance (ANOVA) to test for significant difference where an intervention has been applied**).

**Output:(Kruskal Wallis Anova test-1):**

```
          Df Sum Sq Mean Sq F value    Pr(>F)
eng_wal    1 1812.0 1812.02  65.704 2.066e-10 ***
council    2 2957.4 1478.69  53.617 9.554e-13 ***
Residuals 46 1268.6   27.58
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Output:(Kruskal Wallis Anova test-2):**

```
Residuals 46 1268.6   27.58
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> fit4 <- lm(ed_exp ~ eng_wal + council + eng_wal:council, data = Johnston.80)
> anova(fit4)
Analysis of Variance Table

Response: ed_exp
              Df  Sum Sq Mean Sq F value    Pr(>F)
eng_wal        1 1812.02 1812.02 63.0275 4.937e-10 ***
council        2 2957.39 1478.69 51.4334 3.045e-12 ***
eng_wal:council  2    3.63    1.81  0.0631    0.9389
Residuals     44 1264.99   28.75
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> kruskal.test(ed_exp ~ council, data = Johnston.80)
```

**Interpretation:**

Here in the Anova test p value is lesser than 95 percent, where it has much difference. The tukey hsd is bit similar to anova.

## SESSION 9(Factor Analysis and Cluster Analysis)

### Task2: To carry out correlations and partial correlations in R

### Section4&5:

### Output:(Table 1& Table2):

| | Life_Male | Conclusion |
|---|---|---|
| Dom_Build | P-value :0.1189<br>correlation: -0.2812846 | Not significant and weak negative correlation |
| Smoking | P-value: 0.0002121<br>correlation: -0.6096734 | Significant and strong negative correlation |
| Obese | P-value: 0.273<br>correlation: -0.1997613 | Not significant and weak negative correlation |
| Episodes | P-value: 0.07794<br>correlation: -0.316147 | Not significant and weak negative correlation |
| Benefits | P-value: 1.016e-10<br>correlation: -0.8699088 | Significant and strong negative correlation |
| Crime | P-value: 0.0002036<br>correlation: -0.6110095 | Significant and strong negative correlation |

| Smoking and Benefit | Smoking and Crime | Benefit and Crime |
|---|---|---|
| P-value: 0.001367<br>correlation: 0.5416209 | P-value: 0.04548<br>correlation: 0.356069 | P-value: 6.197e-07<br>correlation: 0.8148827 |
| Significant and positive correlation | A not significant and weak positive correlation | Significant and strong positive correlation |

### Interpretation:

The null hypothesis can be rejected but, in a different sense, the alternative hypothesis can be accepted if the P-value is less than 0.05, which indicates that there will not be a significant difference between the two variables. The alternative hypothesis is not supported if the P-value is larger than 0.05 since there will be a substantial difference between the two variables and we will be unable to accept the null hypothesis in such case. Even though two variables may have considerably distinct distributions and be different from one another, they may still be associated with one another since they co-vary. When one thing changes, another does too. I shall therefore determine the correlation between the two variables. Two variables will be included for further analysis if there is a substantial difference and a strong positive/negative correlation between them. As a result, there was a correlation between the dependent life male variable and the independent smoking, benefits, and crime variables. As a result, we'll incorporate these variables in our analysis.

In table 2 it shows the internally co-related variables using spearman test method. According to statistics, benefits are statistically connected with smoking, and benefits are co related with crime. According to the above table, it means that the benefit and crime variables have a very strong association, but the benefit and smoking have a weaker correlation. As a result, smoking is connected with crime through an internal correlation but not directly with crime. Table 1 revealed a negative correlation between Life male, advantages, and smoking. We will conduct a partial analytical test between Life male, benefit, and smoking because we do not yet know which one is actually correlated or which one is correlated as a result of internal correlation.

**Output:(Table 3& Table4 Partial correlation using Pearson):**

| | Life_Male | Conclusion |
|---|---|---|
| Benefits | P-value: 2.597152e-05<br>correlation: -0.6797932 | Significant and strong negative correlation |
| Smoking | P-value: 0.05226785<br>correlation: -0.3518209 | Not significant and weak negative correlation |

| | Life_Male | Conclusion |
|---|---|---|
| Benefits | P-value: 3.394579e-08<br>correlation: -0.809964 | Significant and strong negative correlation |
| Smoking | P-value: 0.06623432<br>correlation: -0.3340828 | Not significant and weak negative correlation |

**Interpretation:**

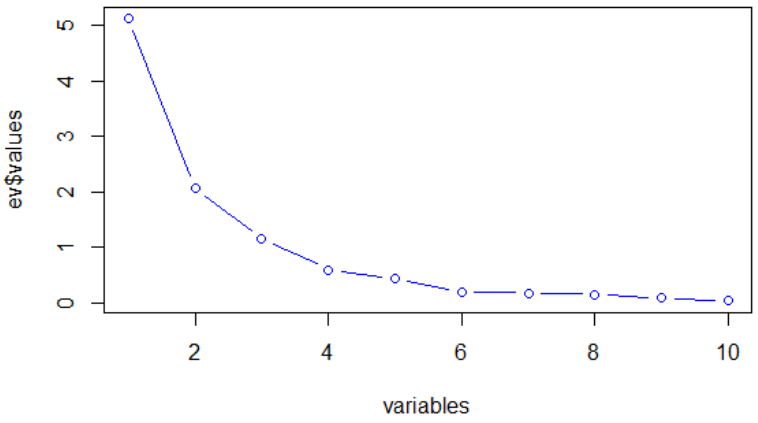
If we compare the Life Male and Benefit connection in tables 1 and 3, we find that it has reduced but is still pretty significant. This indicates that smoking has no inside effects. Once more, when comparing Tables 1 and 3's Life Male and Smoking correlations, the correlation has fallen by half. Due to the weak link between Life Male and Smoking, the strong, or dominant correlation is between Life Male and Benefit. Spearman's partial correlation test also supports it, so we may drop the smoking variable while keeping the dependent variables Life Male and Benefit. This is how we'll use partial correlation to reduce the variables.

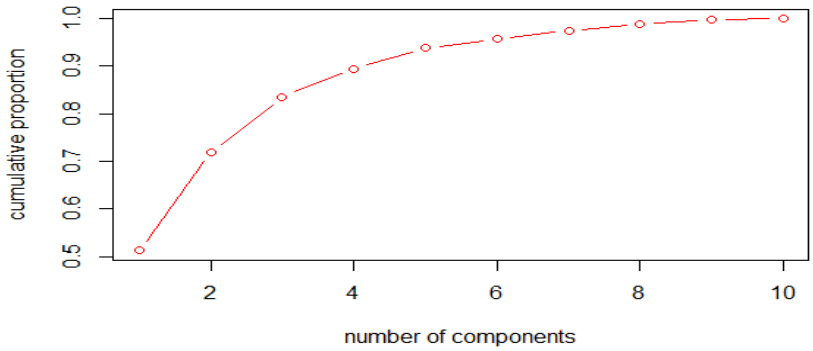
**Task3:(Screen plot):**

**Question: The screen plot and plot showing cumulative percentage variants explain, the final rotated component matrix and your interpretation of what the four component represent**

**Section7:**

**Output:**






**Interpretation:**

Load the 09_London_districts csv file in the R script using R studio and add all variables into new data frame except ten variables for factor analysis. The screen plot and plot with Eigen values with percentage variance is in the above output.

**Question: Carry out principle components analysis (pca) with varimax rotation of four components (nFactors=4) and inspect. Using a threshold of 0.7 are there any variables that do not exceed this in any of the components RC1 to RC4?**

**Section8&9:**

**Output:(Rotated matrix ):**

```
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

                       RC1  RC3  RC2  RC4
SS loadings           3.12 2.35 2.15 1.32
Proportion Var        0.31 0.24 0.22 0.13
Cumulative Var        0.31 0.55 0.76 0.89
Proportion Explained  0.35 0.26 0.24 0.15
Cumulative Proportion 0.35 0.61 0.85 1.00

Mean item complexity =  1.9
Test of the hypothesis that 4 components are sufficient.

The root mean square of the residuals (RMSR) is  0.04
 with the empirical chi square  5.14  with prob <  0.92

Fit based upon off diagonal values = 0.99
```

**Output:(Final Rotated matrix ):**

```
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

                       RC1  RC3  RC4  RC2
SS loadings           2.36 2.11 1.74 1.09
Proportion Var        0.30 0.26 0.22 0.14
Cumulative Var        0.30 0.56 0.78 0.91
Proportion Explained  0.32 0.29 0.24 0.15
Cumulative Proportion 0.32 0.61 0.85 1.00

Mean item complexity =  1.6
Test of the hypothesis that 4 components are sufficient.

The root mean square of the residuals (RMSR) is  0.04
 with the empirical chi square  3.44  with prob <  0.18

Fit based upon off diagonal values = 0.99
> View(London.dis3)
>
```

**Interpretation:**

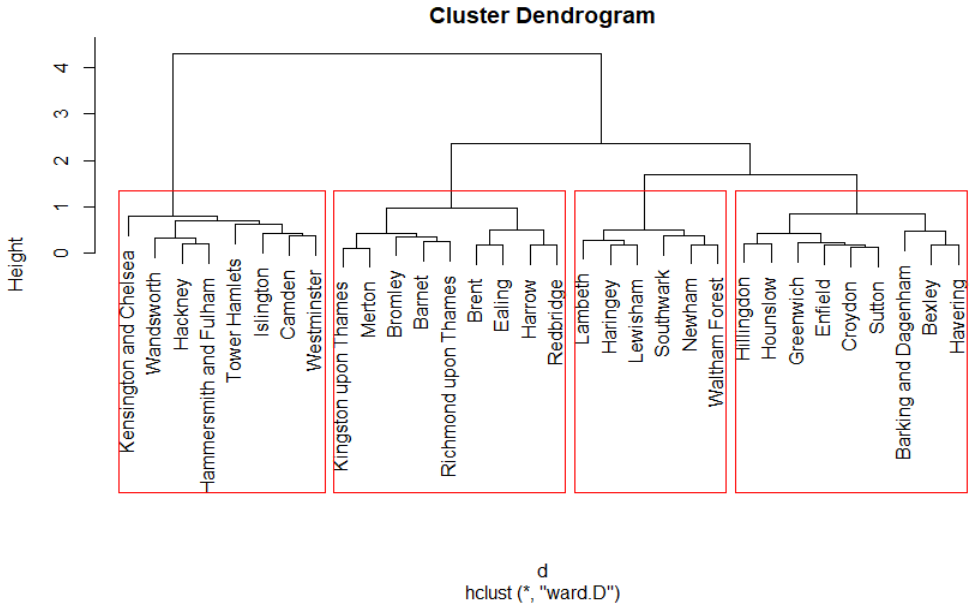
In the above output the mentioned factors(n=4) are Rc1, Rc2, Rc3, Rc4 , where Rc is nothing but rotation component and we can say it as dimensions too. Among all the dimensions in this case, the variable Benefits provides considerable complexity.

**Task4:**

**Question: Insert the dendrogram you produced for wards method, the cluster plot from k-means, the matrix of cluster means from k-means. What, if any, are the keys difference in cluster membership between wards method and k-means?**

**Section11:**

**Output:(Cluster dendrogram):**

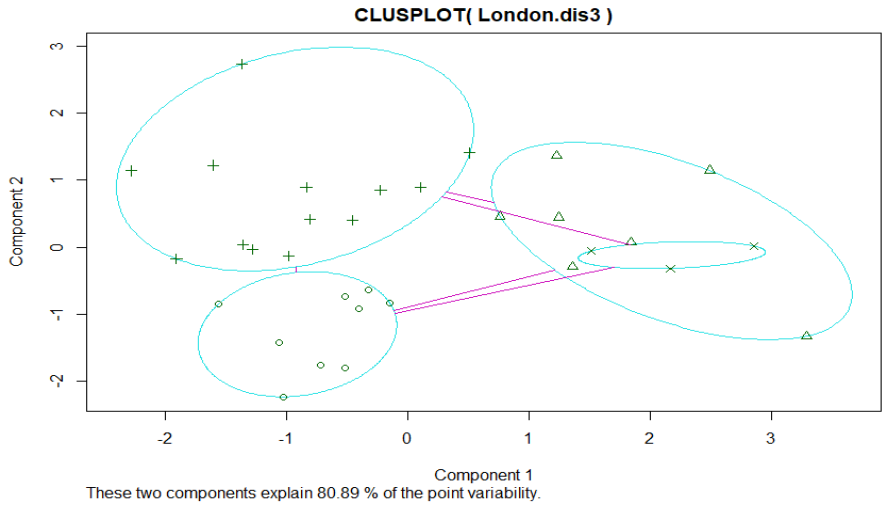



**Interpretation:**

First load the csv file into r script file using the r studio. Then select the four variable which would represent the Pc3. The four variables would be(dom_build, smoking, non_dom_build, obese)from rc1 to rc4 which is a key variable. The above cluster shows that its separated in four groups each group has splited through cluster pattern of height.

**Section12:**

**Output:(Cluster):**

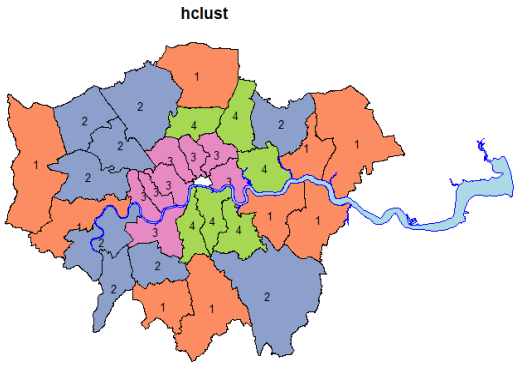



**Interpretation:**

The clusplot shown above has four groupings. The first group has a high average for residential and non-residential buildings that resembles outer London, as well as a high average for green space. Group 2 also has a high average for smoking and obesity-related home problems, and Group 3 has a high average for residential gardens, green space, and heavy smoking. Group 4 has non-dom buildings, an industry area with a high crime rate, and a significant area.

**Section13:**

**Output:(Cluster Map):**

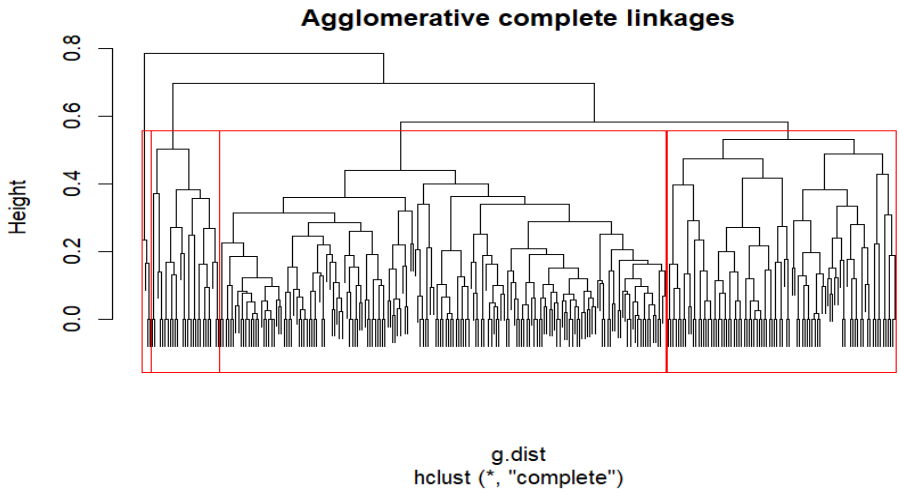



**Interpretation:**

For each area, the hclust above provides a numerical description. Numbers 2 and 3 denote Outer London, whereas number 4 signifies the area around Central London. Group 1 similarly denotes East London.

**Task5:(Cluster using goers dissimilarity)**

**Output:**




**Interpretation:**

The agglomerative links of h clust are full in the above. Group 1's alcohol cluster has seven issues. The majority of them respond positively. Similarly Group 2 contains 70 yes votes and 23 no votes. Then, in group 3, 16 people respond "Yes" and 11 respond "No." Likewise, in group 4, 0 people vote no, while 4 people vote yes.

## SESSION 10(Regression Modelling)

## Task2:(To carry out linear regression and multiple regression in R)

## Section2:

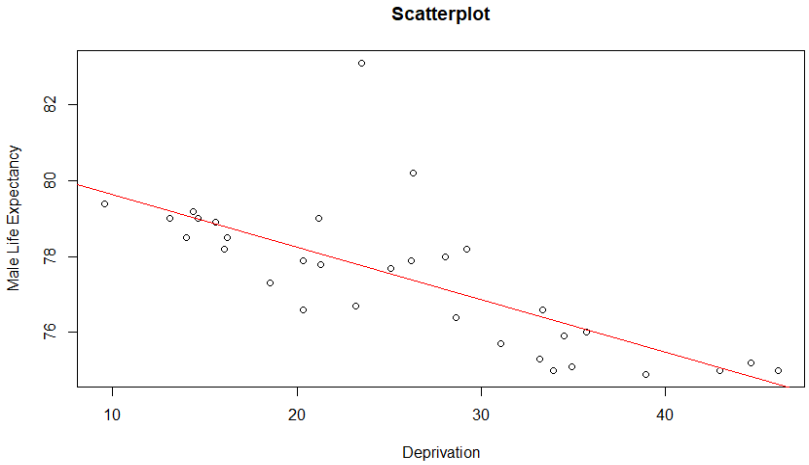
## Output:(Spearman Rank corelation):

```
Source
Console  Terminal  Background Jobs
R 4.1.2 · D:/7006_qs/session10/
 $ Crime        : num  118.9 79.6 69.5 107.5 79.1 ...
> attach(London.dis)
> 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
```

## Interpretation:

The Null hypothesis is Rejected.

## Output:(Male life expectancy vs deprivation):



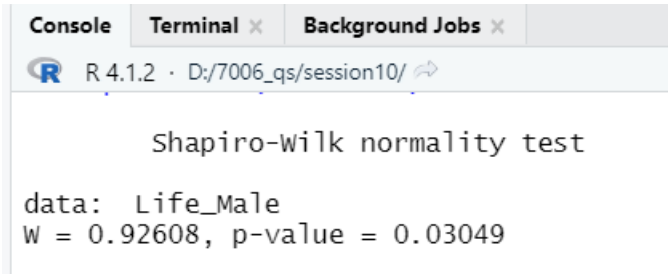

## Output:(Ks-test):

```
Console  Terminal  Background Jobs
R 4.1.2 · D:/7006_qs/session10/

        One-sample Kolmogorov-Smirnov test

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

**Interpretation:**

In the above outputs, the first one says about the male life expectancy vs deprivation where through red line that the data points are weak. In the ks test the p value is 0.95 , where the null hypothesis is accepted.

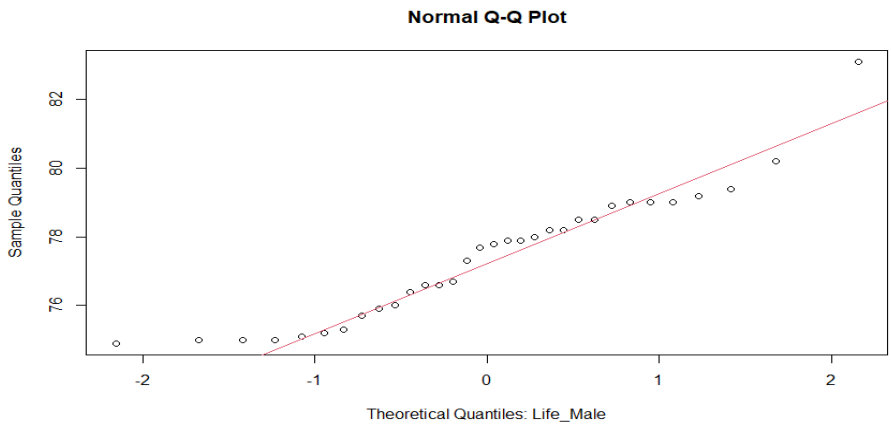
**Output:(Shapiro-Wilk normality test):**




**Interpretation:**

In the Shapiro Wilk normality test , the p value is lesser where p value is 0.03049.so the null hypothesis is rejected.

**Output:**




**Interpretation:**

The Above QQ plot of life of male says that the shapiro wilk normality test is normal.

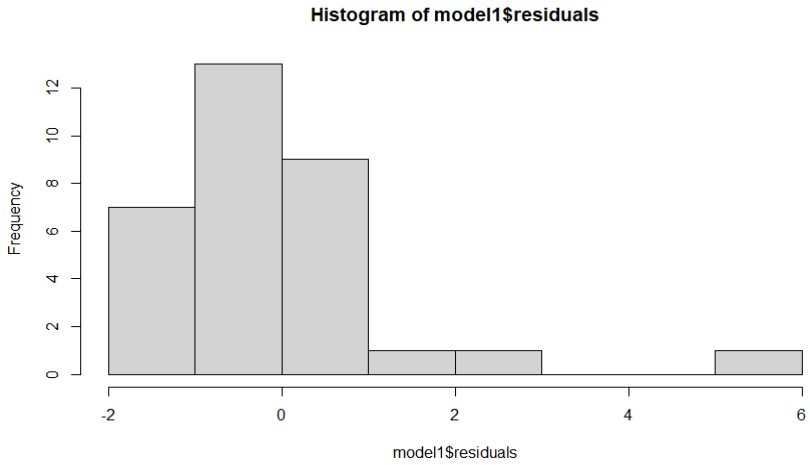
**Output:**

```
Console   Terminal ×   Background Jobs ×
R 4.1.2 · D:/7006_qs/session10/
data:  Life_Male
W = 0.92608, p-value = 0.03049

> # Linear Regression
> model1 <- lm(Life_Male ~ Deprivation)
> # add regression line to scatter plot
> plot(Deprivation, Life_Male, main = "Scatterplot",
+      xlab = "Deprivation", ylab = "Male Life Expectancy")
> abline(model1, col = "red")
> summary(model1)

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
```

**Interpretation:**

Here the null hypothesis is rejected.

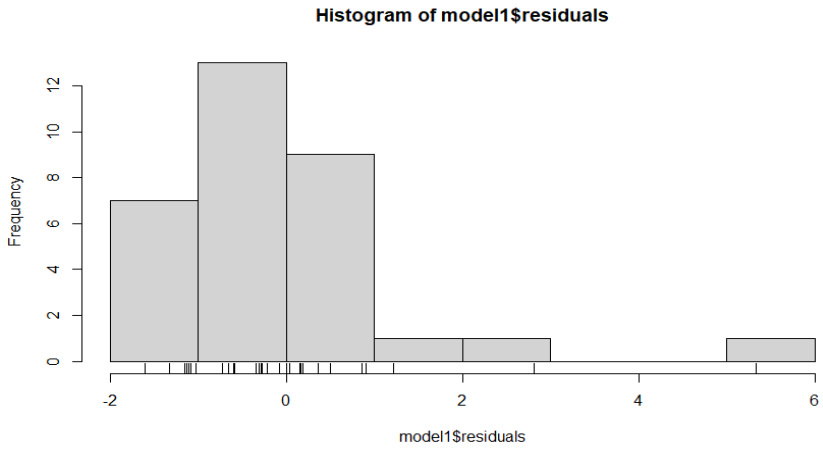
**Output:(model1$residuals):**




**Interpretation:**

From the above histogram the model 1residuals says that poor fit and skew is right.

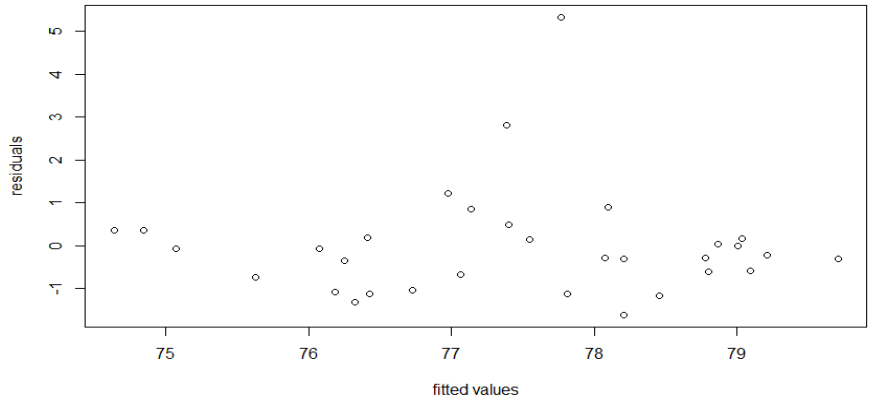
**Output:(model1$residuals):**




**Interpretation:**

From the above histogram it says the model 1 residuals has poor fit and skew is left.

**Output:(scatter plot)**




**Interpretation:**

From the above Scatterplot we can conclude that the data is not normally distributed and it has got much outliers in it.

**Output: (ks-test of model1$residuals):**

```
        One-sample Kolmogorov-Smirnov test

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

>
```

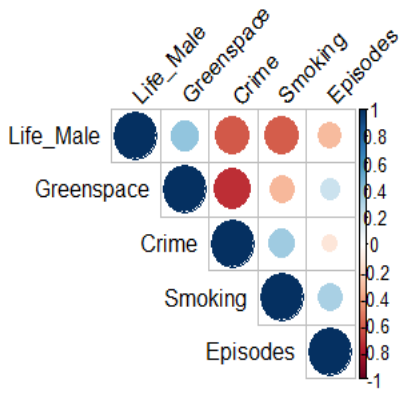
**Interpretation:**

From the above results, it says that p value is rejected.

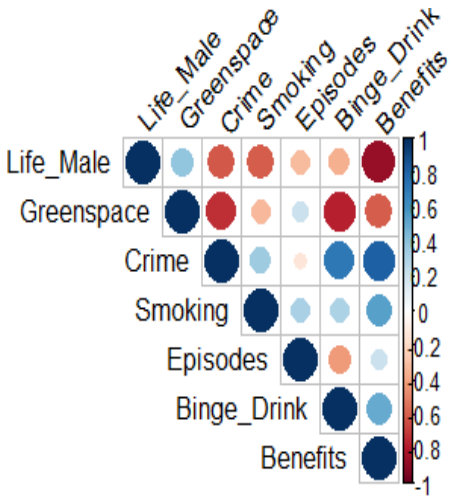
**Section4:**

**Question: Using Life_Male as the dependent variable, produce three multiple regression models to show the need to remove variables that are strongly correlated**:

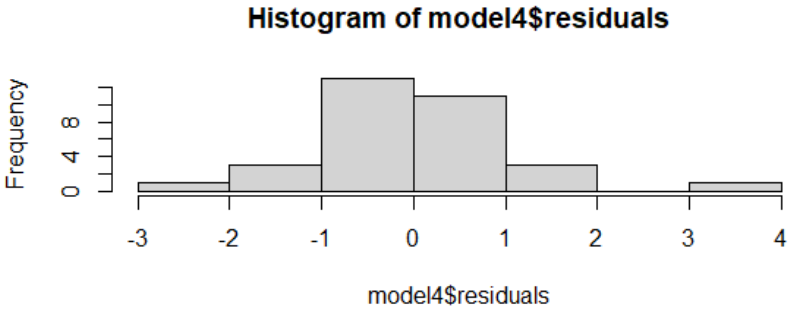
**Output:(corplot for selected Variable):**



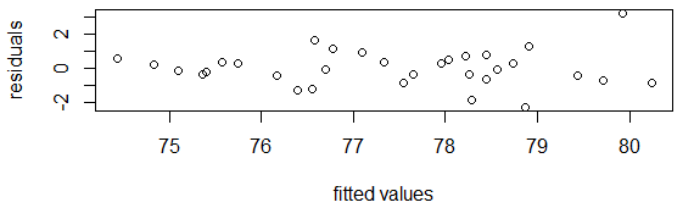

**Output:(corplot for Added Variable):**




**Output:(Histogram model$residuals):**



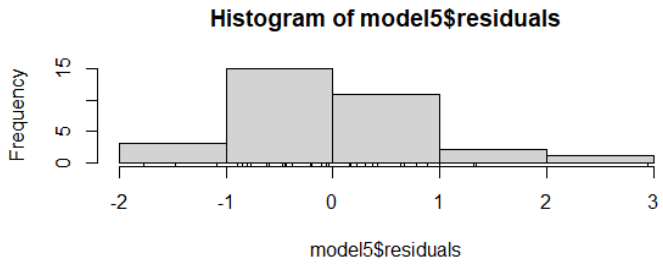
**Output:(residuals vs fitted values):**



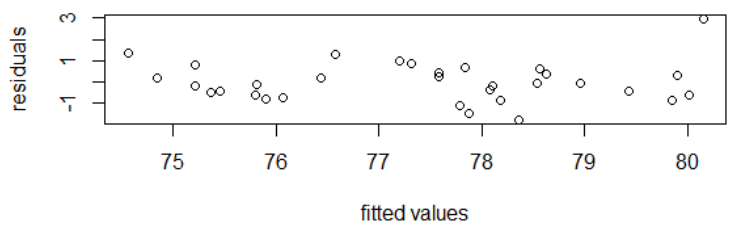

**Output:(Regression Model2):**

```
> vif(model2)
   Dom_Build NonDom_Build  Dom_Gardens   Greenspace      Smoking  Binge_Drink
    5.826126    23.190030     4.242418     9.289247     2.135680     7.082530
       Obese     Episodes     Benefits        Crime
    3.899891     4.169436     3.490845     6.406399
> sqrt(vif(model2)) > 2  # if > 2 vif too high
   Dom_Build NonDom_Build  Dom_Gardens   Greenspace      Smoking  Binge_Drink
        TRUE         TRUE         TRUE         TRUE        FALSE         TRUE
       Obese     Episodes     Benefits        Crime
       FALSE         TRUE        FALSE         TRUE
> # model with four variables representing components from factor analysis
> cor3 <- cor(London.dis2[, c(5,4,12,7,10)], method = "spearman")
> round(cor3, 2)
```

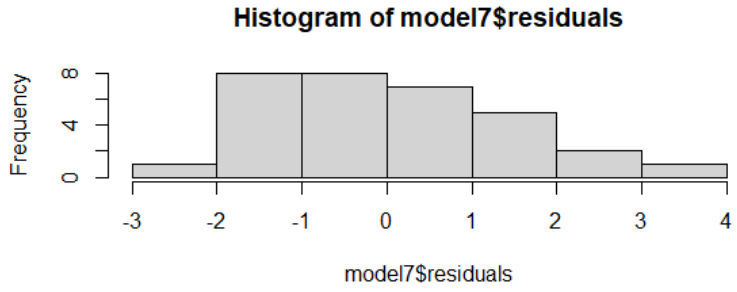
**Output:(histogram model5$residuals):**




**Output:(residuals vs fitted values):**



**Output:(Histogram model$residuals):**




**Output:(pcor test of normality):**

```
> 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
> model3a <- lm(Life_Male ~ Greenspace + Smoking + Episodes)
> summary(model3a)

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

>
```

**Interpretation:**

The high R-squared value of 0.7449, or over 75%, produced by the model 2 regression is regarded as being of low significance. There is a moderate amount of internal correlation (Vif). Although there is no internal correlation, the model 3 regression yields a low R-squared value of 0.4736, or roughly 47%, with a somewhat higher significance.

## Task3:(To carry out logistic regression in R):

## Question: To provide the correlation matrix, the result of the second round of logistic regression and an interpretation of the odds ratios. What is your conclusion about the risk factors in low birth weight?

## Output:(Round 1 co-relation):

```
> round(low.bwt.cor$correlations, 2)
        birth smoke ethnic m_age   mwt low_bwt
birth    1.00  0.03   0.00  0.64  0.30    0.09
smoke    0.03  1.00  -0.53 -0.04 -0.08    0.30
ethnic   0.00 -0.53   1.00 -0.15 -0.15   -0.32
m_age    0.64 -0.04  -0.15  1.00  0.32    0.09
mwt      0.30 -0.08  -0.15  0.32  1.00   -0.13
low_bwt  0.09  0.30  -0.32  0.09 -0.13    1.00
> # first round use all variables
> mylogit1 <- glm(low_bwt ~ birth + smoke + ethnic + m_age + mwt,
+                 data = low.bwt, family = "binomial")
> summary(mylogit1)

Call:
glm(formula = low_bwt ~ birth + smoke + ethnic + m_age + mwt,
    family = "binomial", data = low.bwt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4544  -0.8515  -0.6535   1.1618   2.3313

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.015638   0.720430   0.022 0.982682
birth2       0.232986   0.260393   0.895 0.370922
birth3       0.450347   0.336456   1.339 0.180733
birth4       0.343367   0.680500   0.505 0.613853
smoke1       0.508383   0.223412   2.276 0.022874 *
ethnic2     -0.727524   0.334386  -2.176 0.029578 *
ethnic3     -0.855269   0.256495  -3.334 0.000855 ***
m_age        0.021606   0.022955   0.941 0.346587
mwt         -0.010158   0.003904  -2.602 0.009263 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

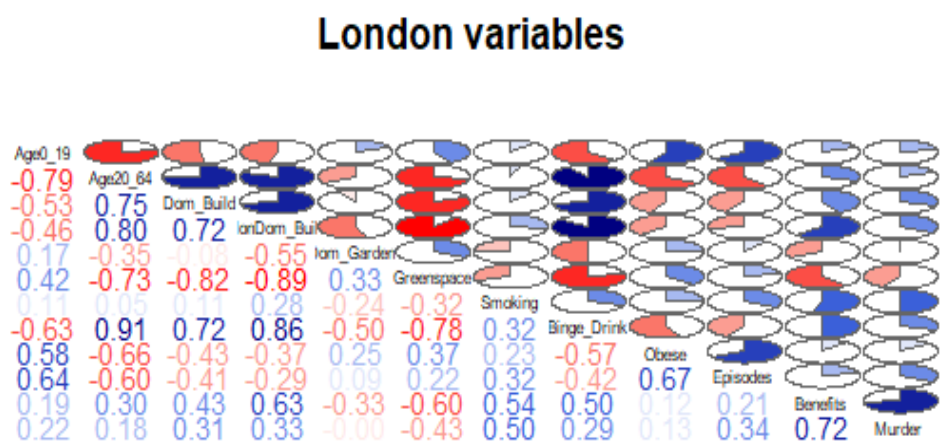
## Output:(Round 2 co-relation):

```
> # second round excluding not significant variables
> mylogit2 <- glm(low_bwt ~ smoke + ethnic + mwt,
+                 data = low.bwt, family = "binomial")
> summary(mylogit2)

Call:
glm(formula = low_bwt ~ smoke + ethnic + mwt, family = "binomial",
    data = low.bwt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2731  -0.8673  -0.6608   1.2017   2.0810

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.329802   0.544486   0.606 0.544706
smoke1       0.516914   0.218877   2.362 0.018193 *
ethnic2     -0.784194   0.325853  -2.407 0.016102 *
ethnic3     -0.845182   0.251175  -3.365 0.000766 ***
mwt         -0.006939   0.003466  -2.002 0.045300 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 603.79  on 487  degrees of freedom
Residual deviance: 568.98  on 483  degrees of freedom
AIC: 578.98

Number of Fisher Scoring iterations: 4

> sqrt(vif(mylogit2)) > 2
        GVIF    Df GVIF^(1/(2*Df))
smoke  FALSE FALSE           FALSE
ethnic FALSE FALSE           FALSE
mwt    FALSE FALSE           FALSE
> |
```

## Output:(Round 3 co-relation):

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> London.mur <- read.csv("10_London_poisson.csv", stringsAsFactors = FALSE)
> str(London.mur)
'data.frame':   32 obs. of  14 variables:
 $ Code         : chr  "00AB" "00AC" "00AD" "00AE" ...
 $ Name         : chr  "Barking and Dagenham" "Barnet" "Bexley" "Brent" ...
 $ Age0_19      : num  29.5 25 25.6 23.5 24.5 ...
 $ Age20_64     : num  58.1 61.2 58.4 64.5 58.9 ...
 $ Dom_Build    : num  8.19 8.31 7.03 11.59 5.35 ...
 $ NonDom_Build: num  5.91 2.88 3.39 6.87 1.63 ...
 $ Dom_Gardens : num  22.8 28.1 25.1 30.3 23.4 ...
 $ Greenspace   : num  33.6 41.3 31.7 21.9 57.8 ...
 $ Smoking      : num  32.1 17.9 27.8 18.6 21.9 23.5 23.2 18.6 23.5 26.6 ...
 $ Binge_Drink : num  11.4 11.1 10.7 12 10.7 15.3 11 12.5 11.7 12.6 ...
 $ Obese        : num  23.9 16.8 21.5 21.6 17.6 13.3 19.3 20 20.2 20.2 ...
 $ Episodes     : num  207 184 215 202 212 ...
 $ Benefits     : num  247 148 141 213 128 ...
 $ Murder       : int  3 1 2 5 4 2 6 3 3 6 ...
> head(London.mur)
  Code                 Name Age0_19 Age20_64 Dom_Build NonDom_Build Dom_Gardens
1 00AB Barking and Dagenham   29.54     58.12      8.19         5.91       22.76
2 00AC               Barnet   24.96     61.18      8.31         2.88       28.10
3 00AD               Bexley   25.62     58.35      7.03         3.39       25.15
4 00AE                Brent   23.48     64.52     11.59         6.87       30.34
5 00AF              Bromley   24.51     58.90      5.35         1.63       23.35
6 00AG               Camden   19.49     71.54     12.22        11.86       19.09
  Greenspace Smoking Binge_Drink Obese Episodes Benefits Murder
1      33.58    32.1        11.4  23.9   206.82   246.55      3
2      41.32    17.9        11.1  16.8   183.75   148.33      1
3      31.69    27.8        10.7  21.5   214.65   141.38      2
4      21.94    18.6        12.0  21.6   201.94   212.52      5
5      57.78    21.9        10.7  17.6   211.62   128.42      4
6      24.79    23.5        15.3  13.3   146.43   207.55      2
> # select variables by excluding those not required
> myvars <- names(London.mur) %in% c("Code", "Name")
> London.mur2 <- London.mur[!myvars]
> rm(myvars)
```

## Output:(Co-relation matrix (London Variables)):




## Output:(Odd ratios):

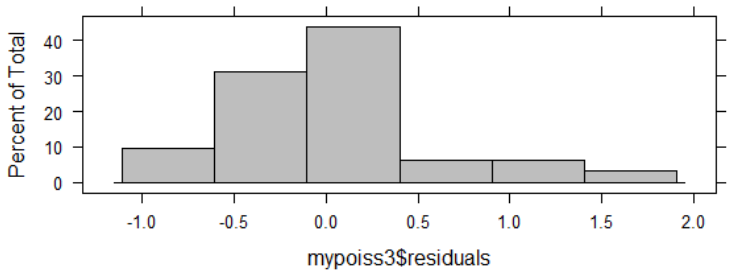
```
    Null deviance: 603.79  on 487  degrees of freedom
Residual deviance: 568.98  on 483  degrees of freedom
AIC: 578.98

Number of Fisher Scoring iterations: 4

> sqrt(vif(mylogit2)) > 2
        GVIF    Df GVIF^(1/(2*Df))
smoke  FALSE FALSE            FALSE
ethnic FALSE FALSE            FALSE
mwt    FALSE FALSE            FALSE
> # calculate Odds Ratio - Exp(b)
> exp(coef(mylogit2))
(Intercept)      smoke1     ethnic2     ethnic3         mwt
  1.3906929   1.6768442   0.4564874   0.4294791   0.9930852
> # calculate the 95% confidence intervals (2 tail)
> exp(cbind(OR = coef(mylogit2), confint(mylogit2)))
waiting for profiling to be done...
                   OR     2.5 %    97.5 %
(Intercept) 1.3906929 0.4842515 4.1107520
smoke1      1.6768442 1.0912538 2.5765994
ethnic2     0.4564874 0.2344355 0.8474666
ethnic3     0.4294791 0.2604481 0.6985763
mwt         0.9930852 0.9862028 0.9997272
>
```

**Interpretation:**

The low correlation explains the value of birth upto 1.00 ,smoke=0.03,Similiarly ethnic 0.00-0.53, age 0.64,mwt= 0.30, low bwt=0.09. The 8 variables are expecting the null hypothesis value. Null deviance: 603.79 on 487 degrees of freedom. Residual deviance: 563.18 on 479 degrees of freedom AIC: 581.18. The result shows sqrt of 1,2,3,4,5 variables falls negative. The above result shows that GVIF explains about birth, smoke, ethnic, age, mwt has false result. The ratio values that smoke1 per has more increase upto 1.6768442 it has most highest in 97.5%.Similiarly ethnic2 and ethnic3 values 0.4564874 and 0.4294791. Then mwt=0.9930852.

**Task4:(To carry out Poisson regression in R):**

**Output:(poisson regression):**




**Output:(plot for poisson):**


**Output:(Odd ratios for cor-relation):**

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.008260   0.921950   1.094   0.2741
Benefits      0.010034   0.002453   4.090 4.32e-05 ***
Greenspace   -0.030110   0.015760  -1.911   0.0561 .
NonDom_Build -0.122758   0.048892  -2.511   0.0120 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 61.559  on 31  degrees of freedom
Residual deviance: 27.874  on 28  degrees of freedom
AIC: 127.88

Number of Fisher Scoring iterations: 5

> # calculate Odds Ratio - Exp(b)
> exp(coef(mypoiss3))
 (Intercept)     Benefits   Greenspace NonDom_Build
   2.7408286    1.0100847    0.9703391    0.8844780
> # calculate the 95% confidence intervals (2 tail)
> exp(cbind(OR = coef(mypoiss3), confint(mypoiss3)))
Waiting for profiling to be done...
                    OR     2.5 %     97.5 %
(Intercept)  2.7408286 0.4405755 16.4634987
Benefits     1.0100847 1.0053326  1.0150596
Greenspace   0.9703391 0.9402068  1.0002203
NonDom_Build 0.8844780 0.8009171  0.9699181
> # variable relative importance for glm() models - absolute value of t
> library(caret)
> varImp(mypoiss3)
              Overall
Benefits     4.089753
Greenspace   1.910535
NonDom_Build 2.510769
> histogram(mypoiss3$residuals, col = "Grey")
> plot(London.mur2$Murder, mypoiss3$fitted.values)
```

**Interpretation:**

There are several outliers, the plots are not all in a straight line, and the data is not normally distributed. The histogram indicates that the residuals are abnormal. Similarly, skiw is on the histogram's right side. The estimate coefficients table shows that intercept p-value = 1.094 ,Similarly benefits value 4.090, Then Greenspace and Non dom build values are -1.911 and -2.511. Null deviance: 61.559 on 31 degrees of freedom Residual deviance: 27.874 on 28 degrees of freedom It can be conclude that all four variables except null hypothesis. The above result shows the odd ratio, similiarly,3 variables is the benefits of Non dom build and greenspace. It results that benefits has 1.0150596 has the confidence of 97.5. Similarly, Greenspace and Non Dom build has the values like 0.9703391 and 0.9699181 has the confidence of 97.5%. Overall benefit.