Normalized the following variables as they had outliers and were skewed:

JanTemp, NOxPot, SO2Pot, PopDensity, Pop, NW, Income, HCPot – Positively Skewed

HHsiz - Negatively skewed

Variables were considerably skewed and hence used logarithmic transformations

This was done using the **histogram** approach.

Removed the city identifier as it is nominal data

**Question 5**

Ran a forward linear regression model and found the following elements to be significant: **NW, JanTemp, Education, SO2Pot, Rain, WC, NOxPot**.

Residual standard error: 35.42 on 51 degrees of freedom

Multiple R-squared: 0.7169, Adjusted R-squared: 0.6781

F-statistic: 18.45 on 7 and 51 DF, p-value: 5.55e-12

This means that 67.81% of variance in the model is explained by the 7 chosen independent variables

**Question 6**

Eliminated **Education**, **SO2Pot** and **WC** in my refined model as they had very less t value and probability value. The following summary was observed:

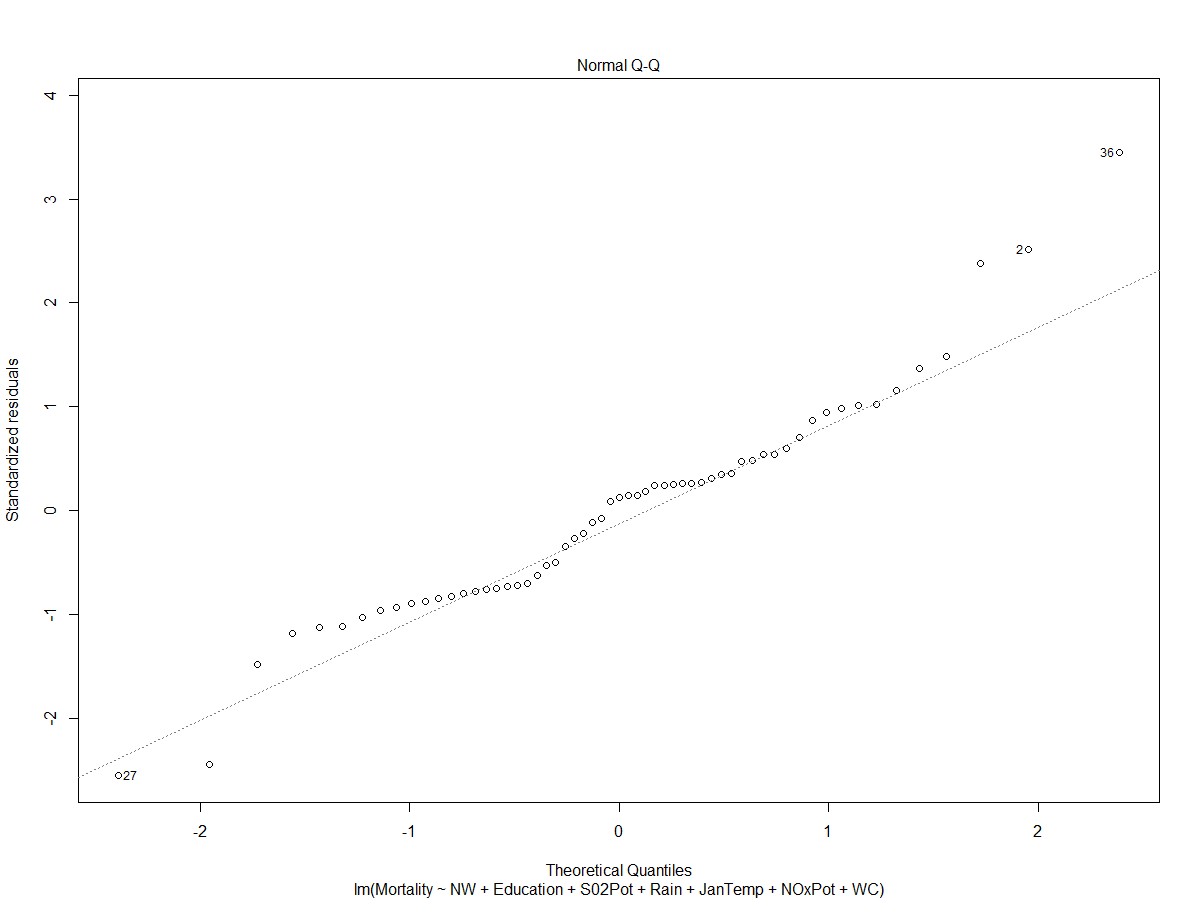
Residual standard error: 38.1 on 54 degrees of freedom

Multiple R-squared: 0.6531, Adjusted R-squared: 0.6274

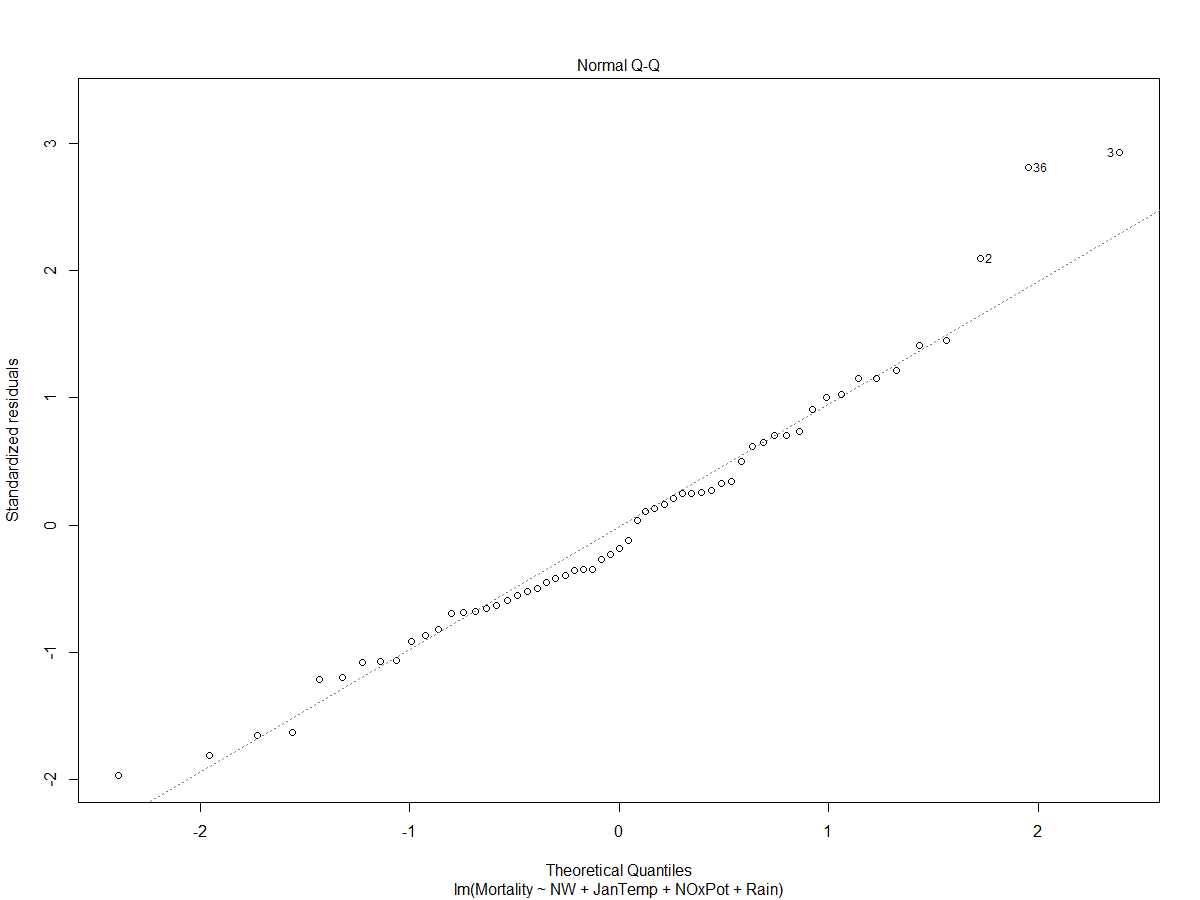
F-statistic: 25.41 on 4 and 54 DF, p-value: 7.187e-12

This means that 62.74% of variance in the model is explained by the 4 chosen independent variables

The p-value is almost same for both the models but the second model has a better F-statistic value for the same p-value. Any further reduction in the number of elements leads to a significant decrease in p-value and hence I have taken the second model with elements **NW, JanTemp, NOxPot and Rain as the best model. Also, comparing the Q-Q plot of the 2 models, the latter one had a better fit hence influencing the decision to consider the second model.**



**Model 2 plot:**



**Question 8**

The first five principal components are selected since they are the ones with eigen values greater than 1.

**Question 10**

Ran PCA with the 5 selected components.

**Components 1, 2 and 3** are significant based on p-value and t statistics

Residual standard error: 44.76 on 53 degrees of freedom

Multiple R-squared: 0.5301, Adjusted R-squared: 0.4858

F-statistic: 11.96 on 5 and 53 DF, p-value: 9.028e-08

This means that 48.58% of variance in the model is explained by the 5 chosen principal components

**Question 11**

Ran a linear regression with the 3 selected components and it is validated that they are significant to the model. For almost the same p-value, the F-statistic increases considerably and hence choosing this as the final model of the regression.

Residual standard error: 45.02 on 55 degrees of freedom

Multiple R-squared: 0.5067, Adjusted R-squared: 0.4798

F-statistic: 18.83 on 3 and 55 DF, p-value: 1.577e-08

This means that 47.98% of variance in the model is explained by the 3 chosen principal components

**Question 12**

**The linear regression model(stepwise forward) explains the variance better than the PCA analysis with more varuables**

**The linear model boasts of a better F-statistic and p-value compared to the PCA model making it a better model**

**Association Rules**

2.a. Two rules had the lift value of 1.7**.**

lhs rhs support confidence lift count

[1] {Greeting Cards} => {Candy Bar} 0.044 0.30 1.7 8732

[2] {Candy Bar} => {Greeting Cards} 0.044 0.26 1.7 8732

2.b. Lift value = Confidence/Benchmark confidence

For rule 1: In other words it is the (support of finding greeting cards and candy bar together)/(product of support of greeting cards and support of candy bar)

2.c.

lhs rhs support confidence lift count

[1] {Greeting Cards} => {Candy Bar} 0.044 0.30 1.7 8732

[2] {Candy Bar} => {Greeting Cards} 0.044 0.26 1.7 8732

[3] {Toothpaste} => {Candy Bar} 0.040 0.25 1.5 7956

[4] {Candy Bar} => {Toothpaste} 0.040 0.23 1.5 7956

[5] {Pencils} => {Candy Bar} 0.033 0.24 1.4 6603

Rule 1: 4.4 % of the transactions had both items. 30% of transactions that purchased greeting cards first purchased candy bars and lift of 1.7 indicates that the events are dependent on each other to a considerable extent

Rule 2: 4.4 % of the transactions had both items. 26% of transactions that purchased candy bars first purchased greeting cards and lift of 1.7 indicates that the events are dependent on each other to a considerable extent

Rule 3: 4.0 % of the transactions had both items. 25% of transactions that purchased toothpaste first purchased candy bars and lift of 1.5 indicates that the events are dependent on each other to a considerable extent

Rule 4: 4.0 % of the transactions had both items. 23% of transactions that purchased toothpaste purchased candy bars and lift of 1.5 indicates that the events are dependent on each other to a considerable extent

Rule 5: 3.3 % of the transactions had both items. 24% of transactions that purchased pencils purchased candy bars and lift of 1.4 indicates that the events are dependent on each other to a considerable extent

2.d.

* Magazines are mostly purchased as a standalone item with the exception of being purchased along with greeting cards.
* Candy Bars are present in most combinations(5) followed by greeting cards(4) meaning they are always bought in combination with other products
* Pencils are bought only along with candy bar predominantly compared to being bought separately or with any other item