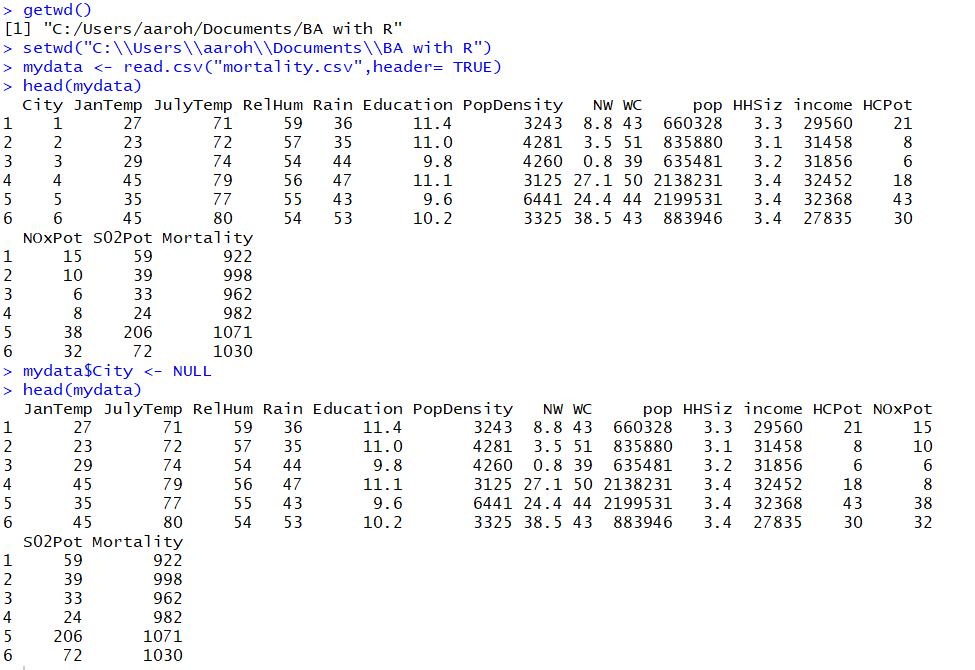
**Exercise 1: Linear Regression & Principal Components Analysis**

Linear regression is used to predict a relationship between dependent variable and predictors.

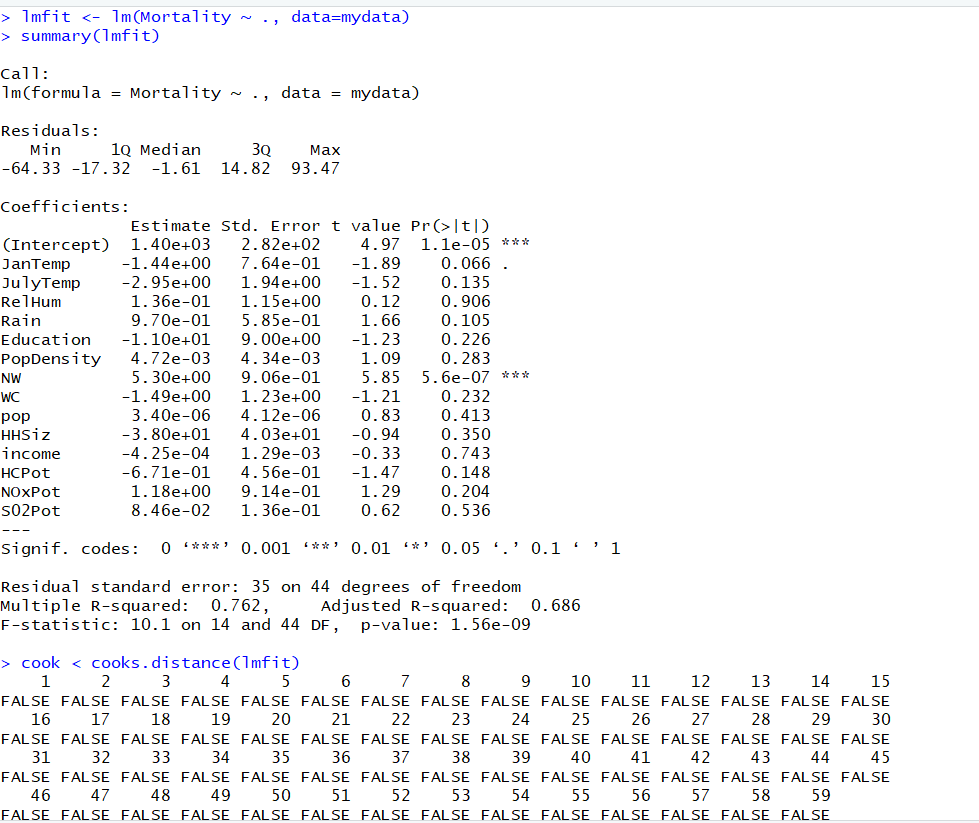
Below is a step by step approach to linear regression. As mentioned in the assignment the City column is removed so that it does not interfere with the regression and does not have any impact on the final results.

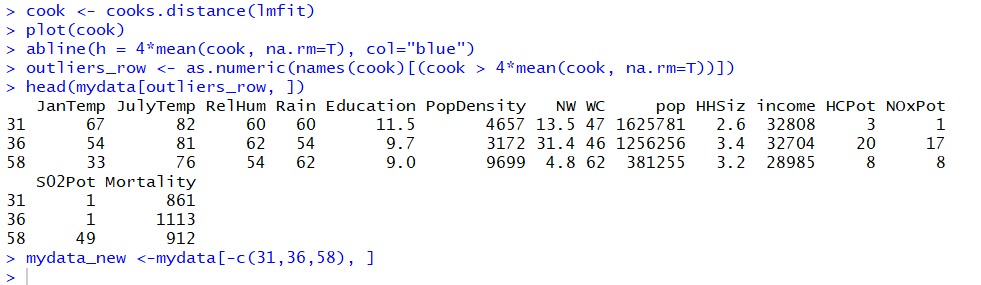


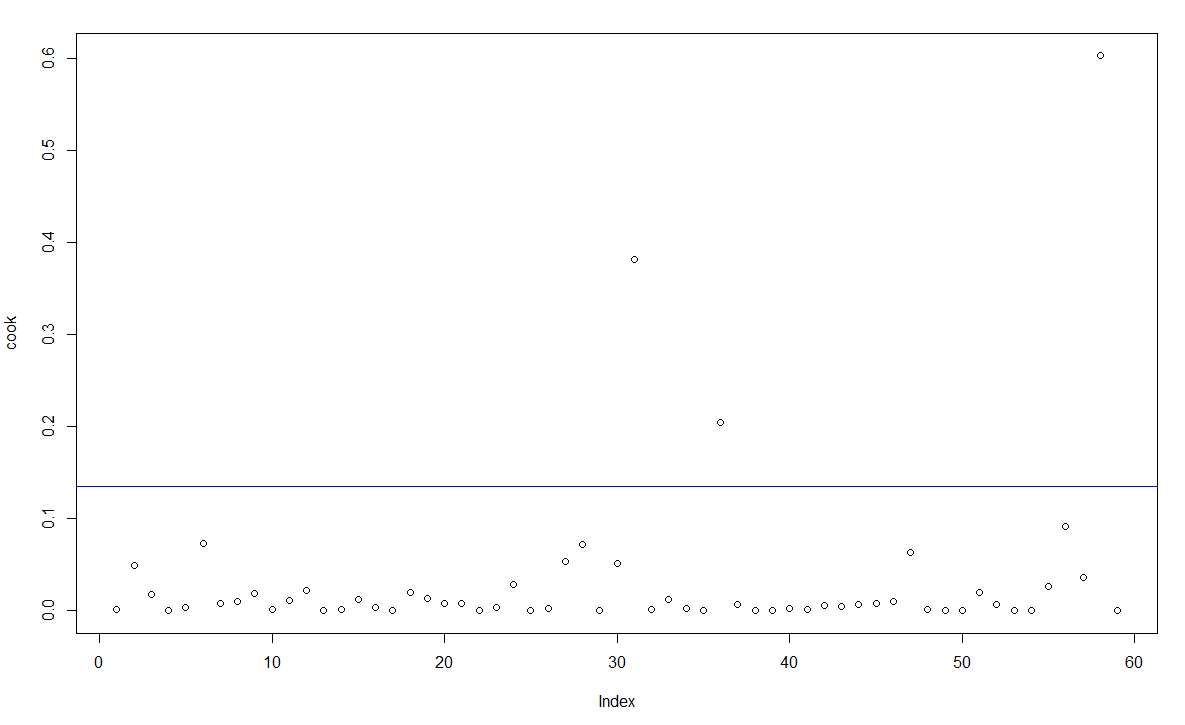
The next step followed is the removal of outliers. The outlier treatment is important because it can drastically impact the estimates and predictions due to extreme values, thus, not giving accurate results.

To decide if the observation is an outlier or not, we will plot Cook’s distance.

For plotting the Cook’s distance, we first need to predict the model using the **lm** function for the entire dataset. Once the results are obtained, the Cook’s distance is plotted for the model using cooks.distance function as shown below. Those observations that have a cook’s distance greater than 4 times the mean may be classified as influential. In the below graph, we see 3 observations which we can treat as outliers. We then obtain the rows of these 3 observations and remove them using the complement function. In this model we have outliers at rows 31, 36, 58.





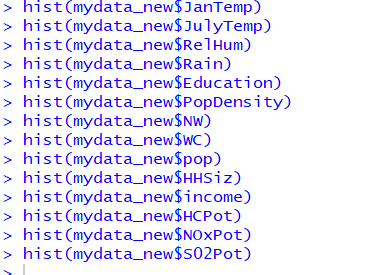


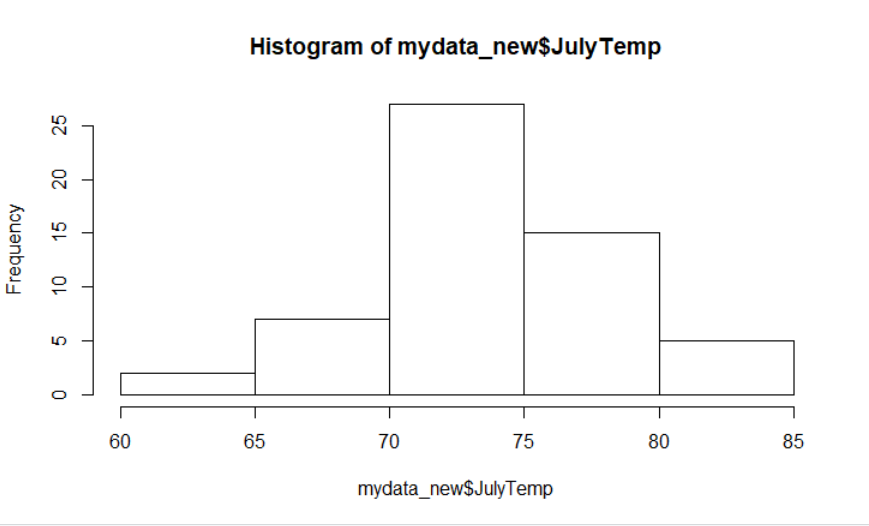
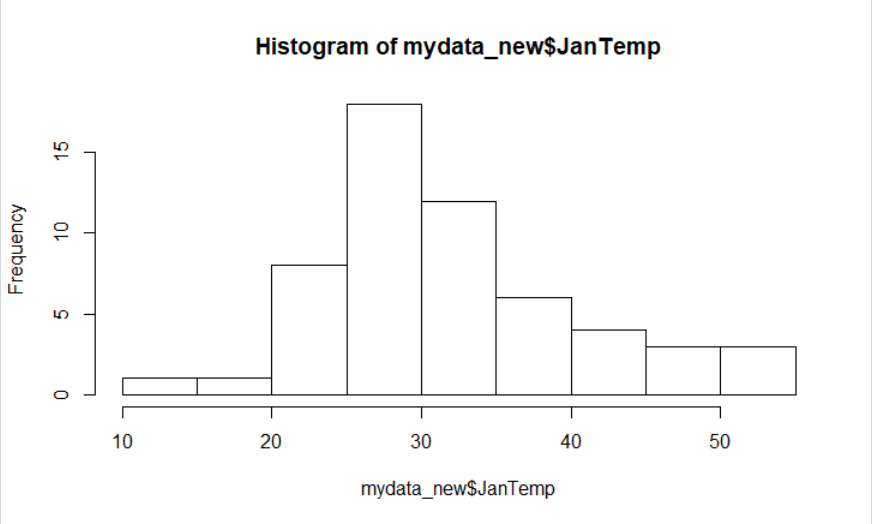
After removing the outliers, we check by the count of the total number of rows in the new dataset to validate that the outliers were deleted.

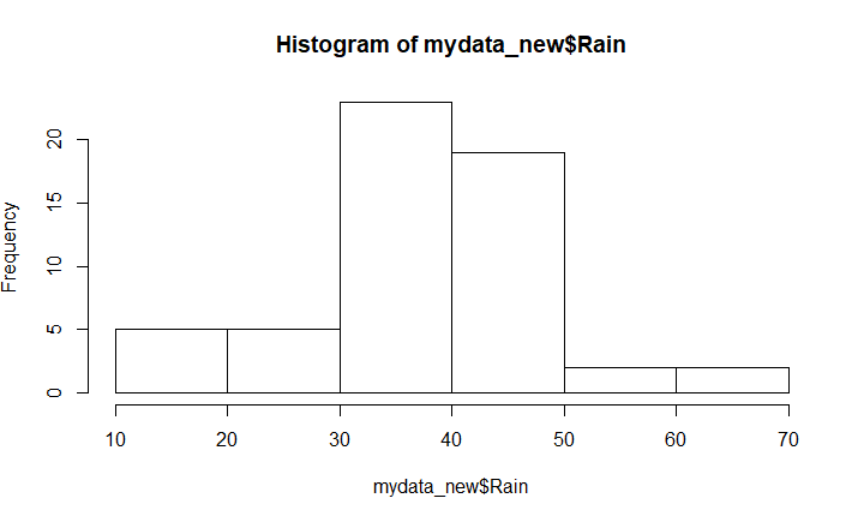
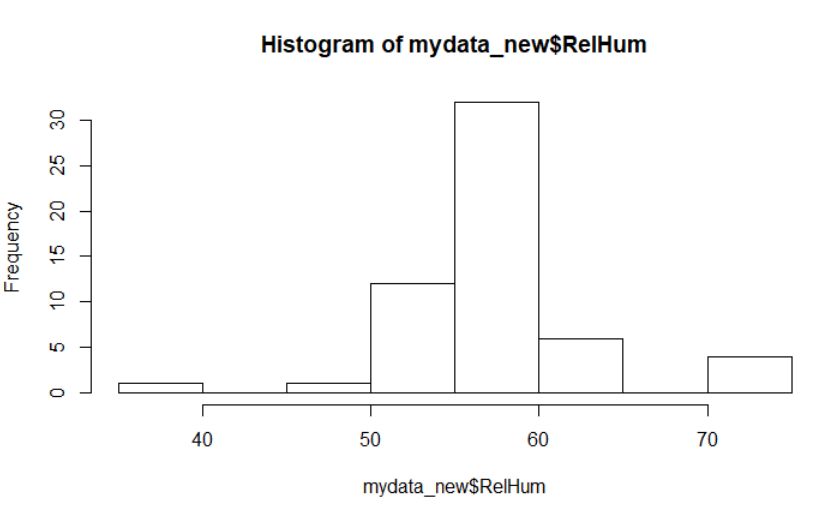


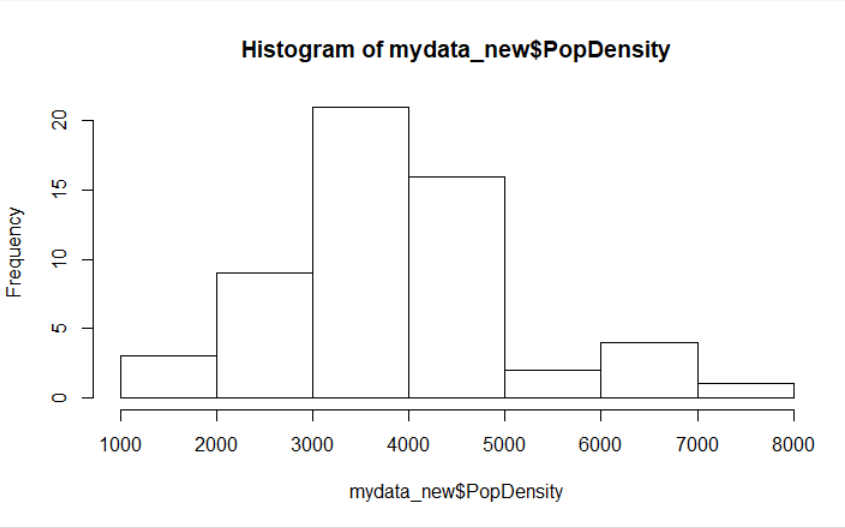
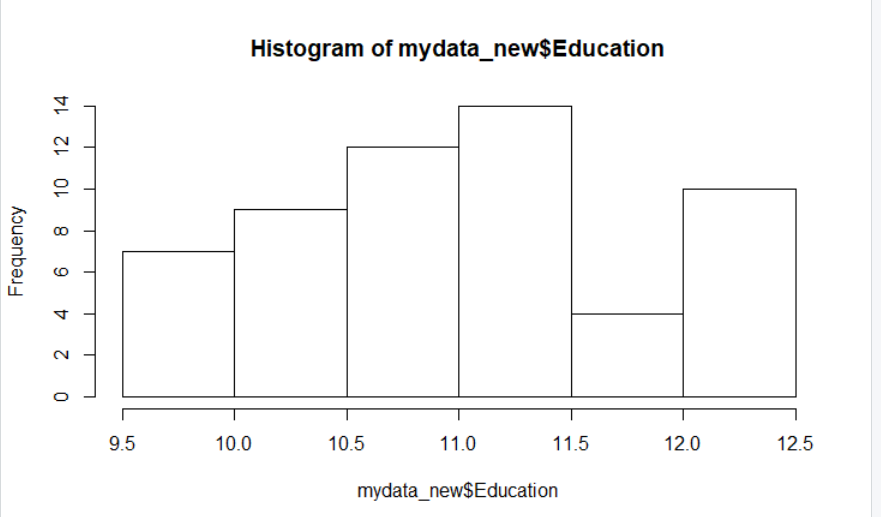
After removing the outliers, the next step is to check if the data is normalized or not. If the data is skewed, it needs to be normalized else the estimation will not give proper results and might be biased.

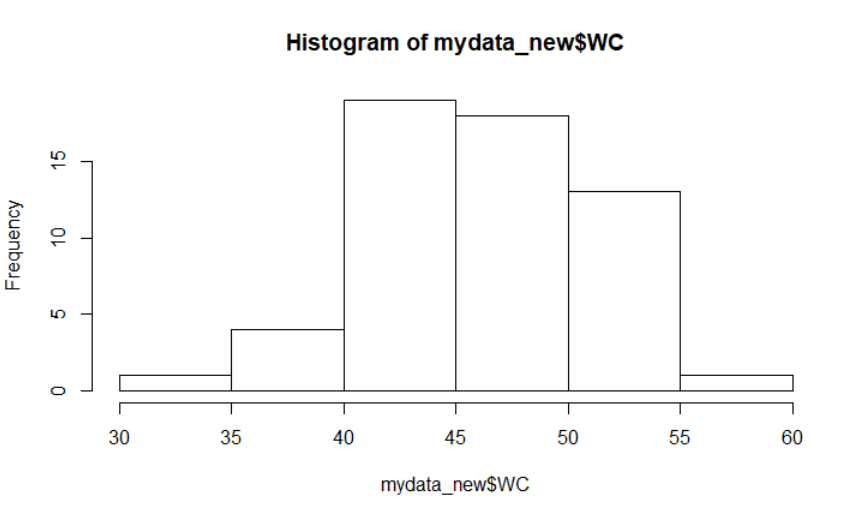
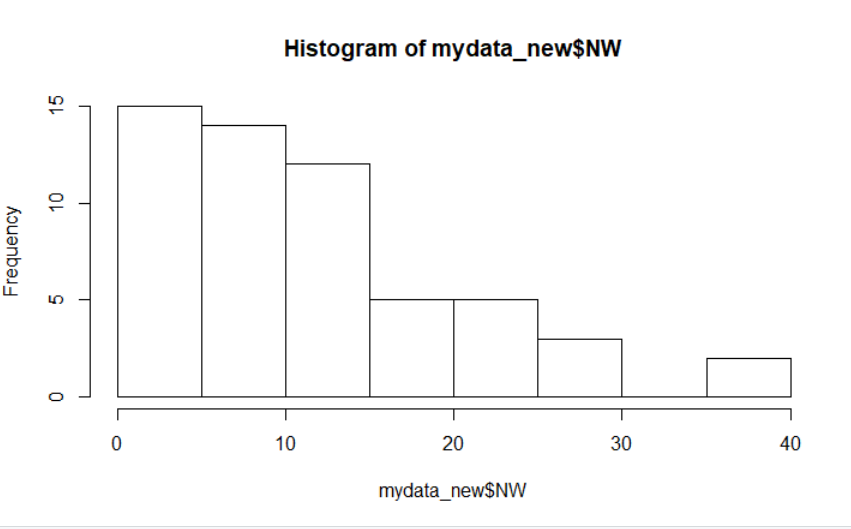
To check if the data is skewed or not, histogram is plotted for all the columns.

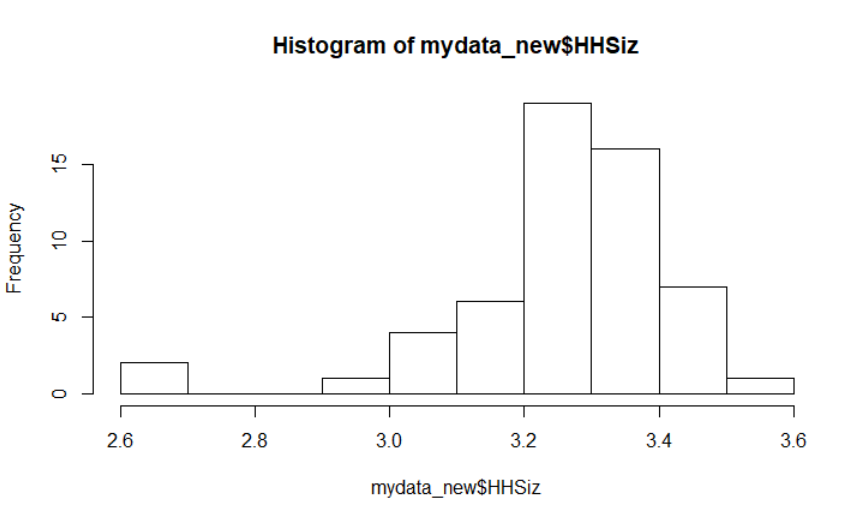
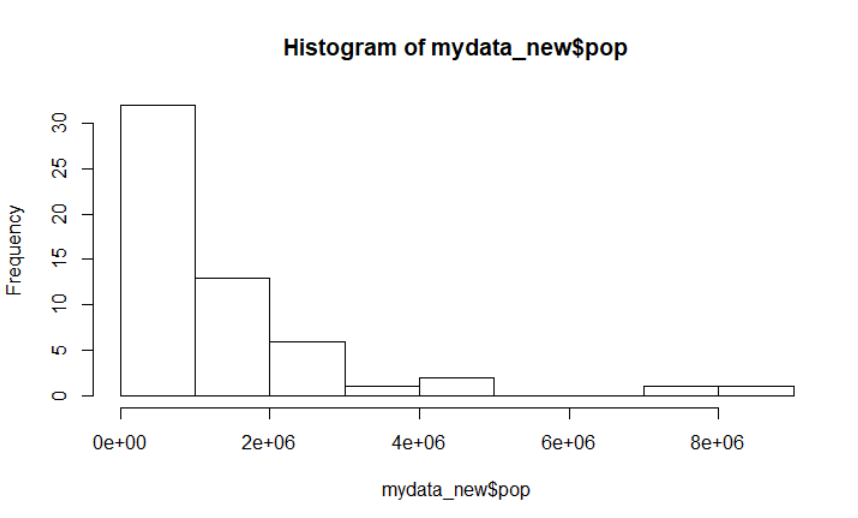


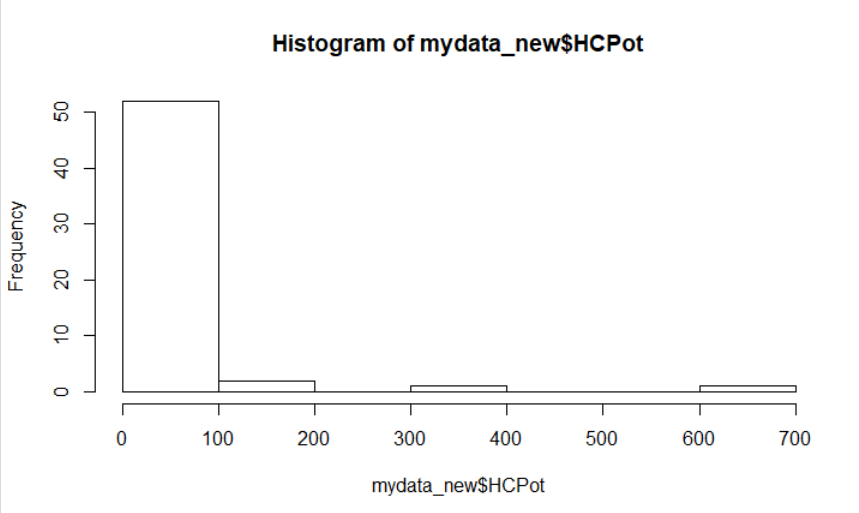
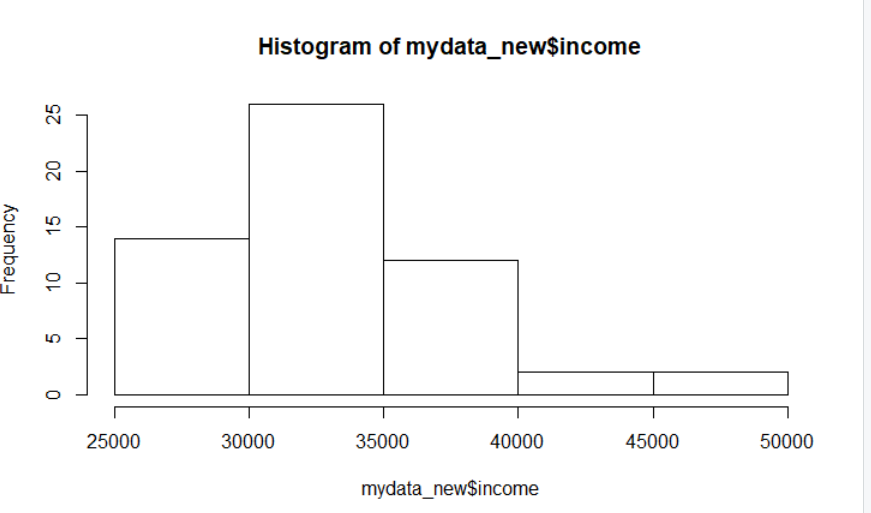


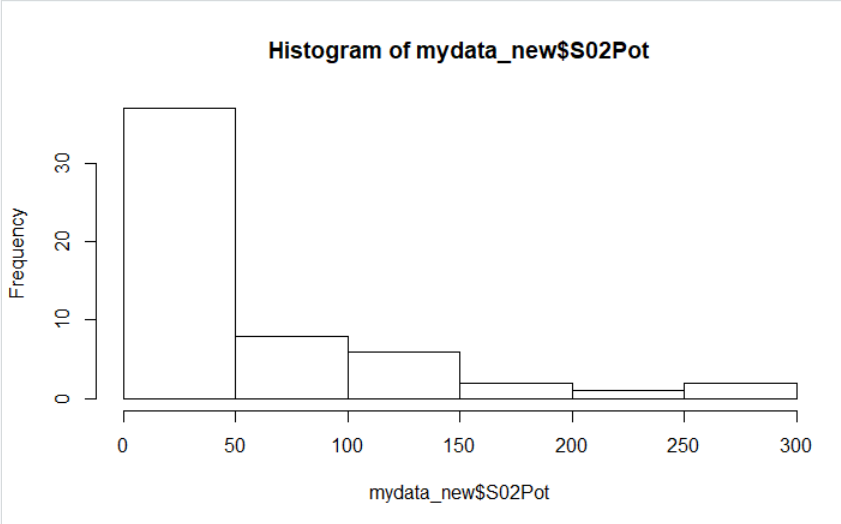
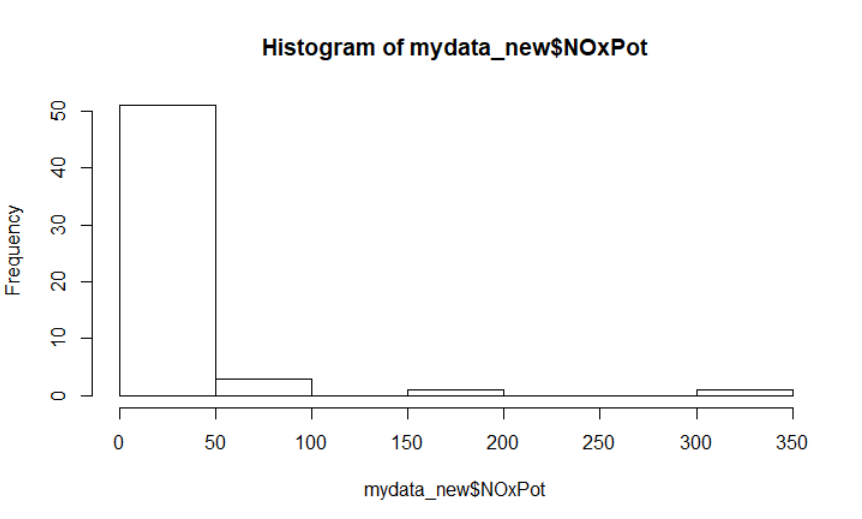






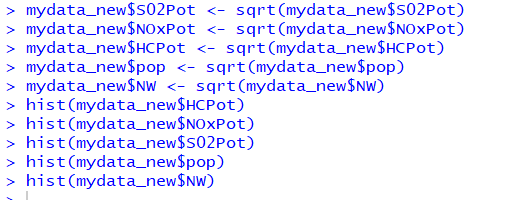


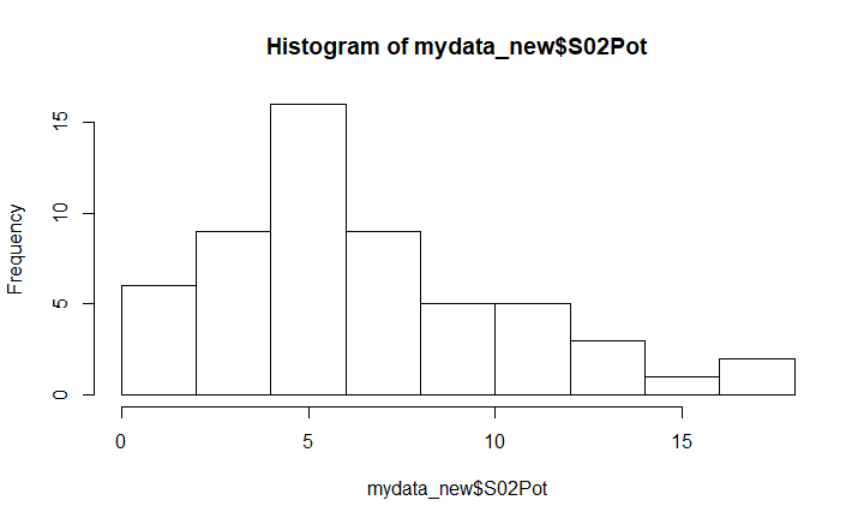
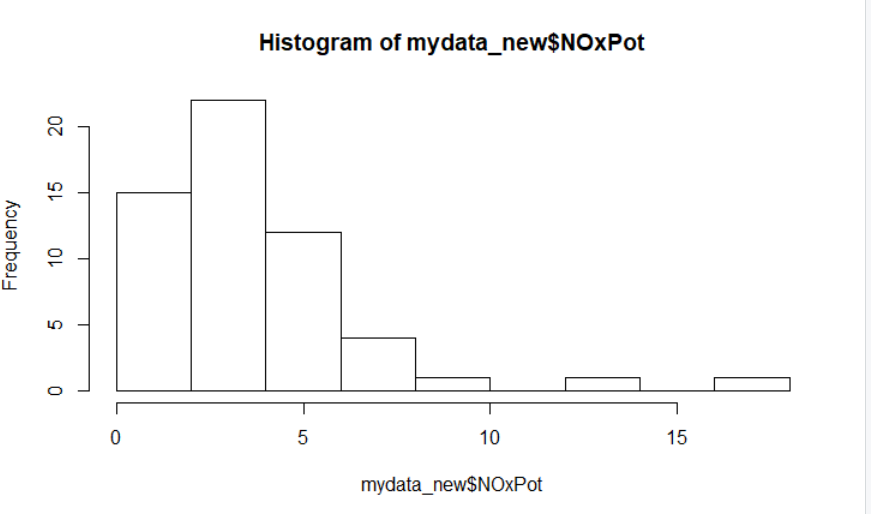


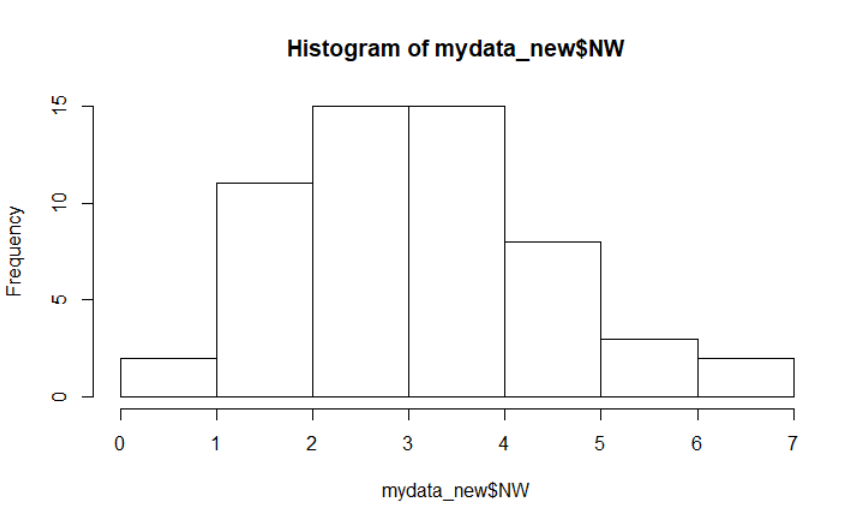
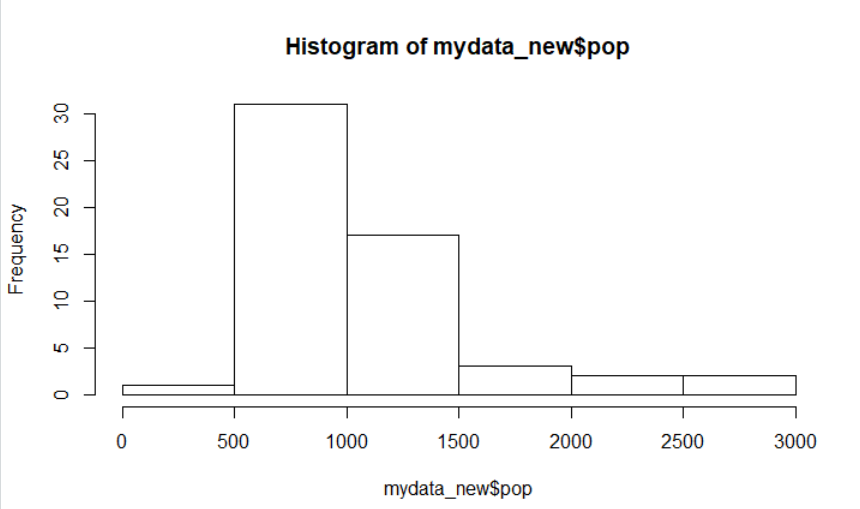


As shown in the above histograms, it is observed that for some columns the data is highly skewed for some columns and it needs to be normalized. To normalize the data, the square root of the impacted columns is taken and then these new updated values are used to predict the model.

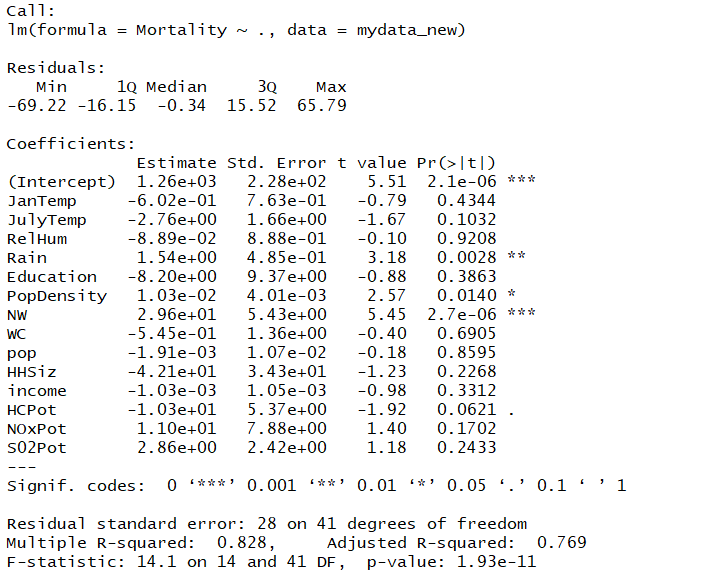
Below are the columns after normalizing the columns.







After normalizing the data, the linear regression is performed and below are the results.



The summary gives the coefficients of dependent variables. Here we see while predicting the dependent variable Mortality using the predictors mentioned we find that the predictors Rain, PopDensity, NW are significant while HCPot is marginally significant. Three asterisks mean 99.99% significant (If it is less than 0.01 means variable is 99.99% significant).

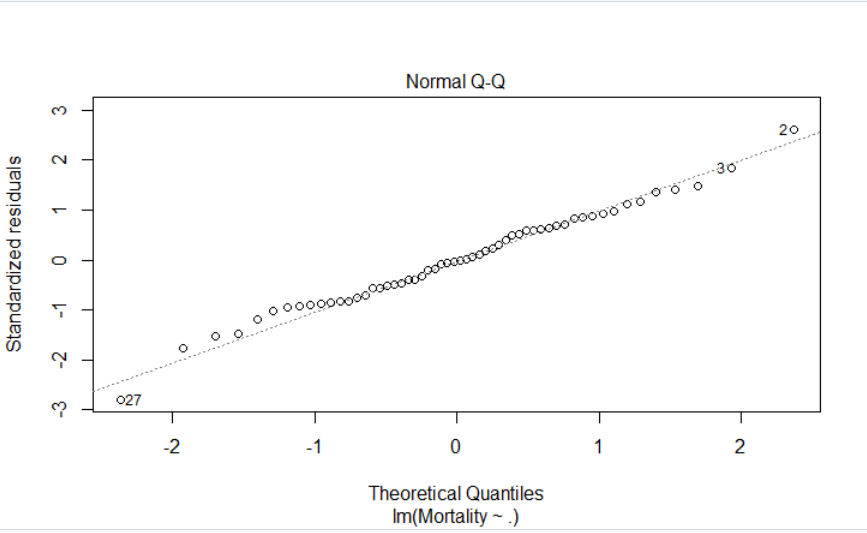
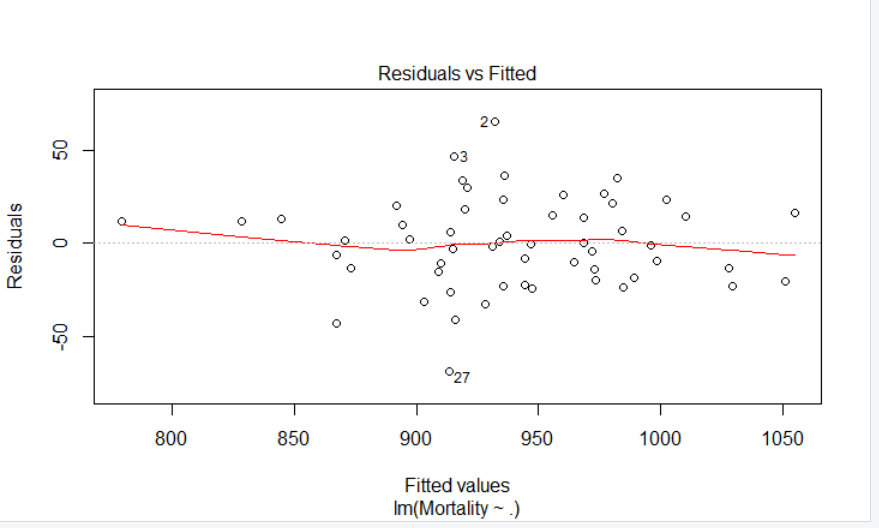


Figure 2.1 Figure 2.2

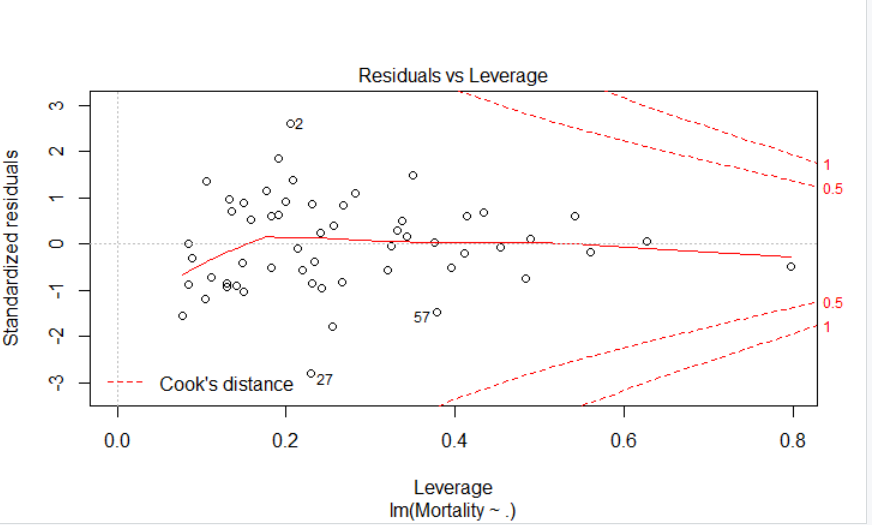
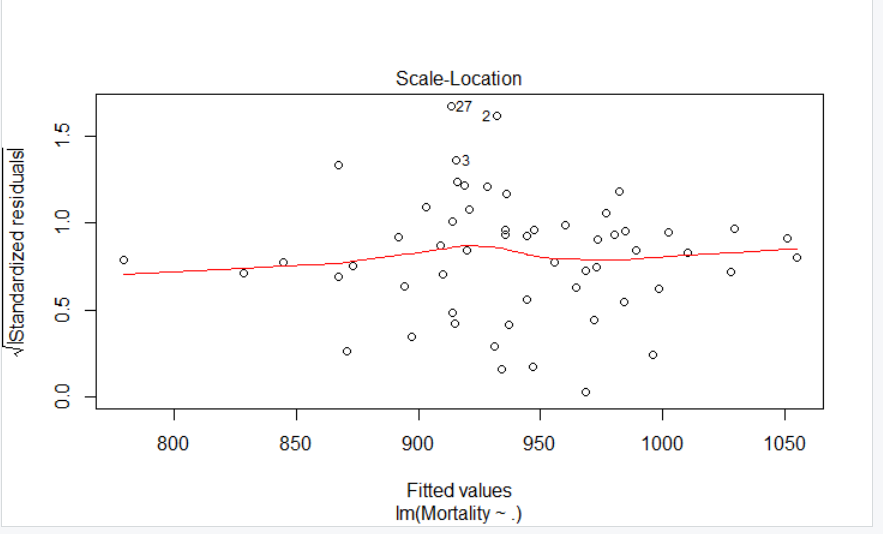


Figure 2.3 Figure 2.4

The plot in the Figure 2.1 shows the residual errors plotted versus their fitted values. The residuals should be randomly distributed around the horizontal line representing a residual error of zero; that is, there should not be a distinct trend in the distribution of points. The plot in the Figure 2.2 is a standard Q-Q plot, which should suggest that the residual errors are normally distributed. The scale-location plot in the Figure 2.3 shows the square root of the standardized residuals as a function of the fitted values. Figure 2.4 shows each point leverage, which is a measure of its importance in determining the regression result.

5a)

For the model analyzed above, R-squared is 0.828 and adjusted R-squared is 0.769

R-Squared is the statistical measure that defines how close the data is to the fitted line. It is the percentage of the variable variation explained by the linear model. In simple terms we can say the percentage of change in the dependent variable explained by the percentage change in the independent variable. R- squared provides an estimate of the strength of the model and response variable. Here it shows that 82% variation in the dependent variable is explained by 82% variation in the independent variable.

Adjusted R-squared is a better indicator of the performance of the model since it considers only the variables that are significant to the model and penalizes the inclusion of predictors which are insignificant. The adjusted R-squared increases only when the new predictor improves the model. When considered the case here the value shows the model is 76% fitted.

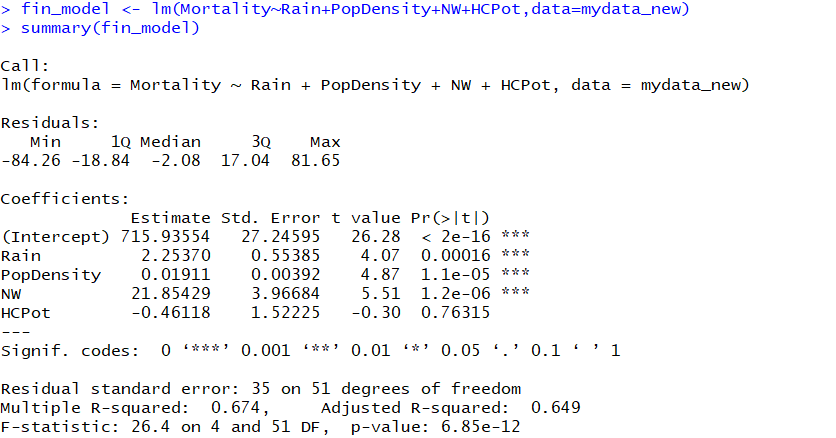
5b)

When we look at the t value statistics and the associated probabilities we see that variables Rain, PopDensity, NW and HCPot are significant Pr>|t| values as 0.0028, 0.0140, 2.7e-06 and 0.0621 respectively. These values show that these predictors are 99.99% significant.

6a)

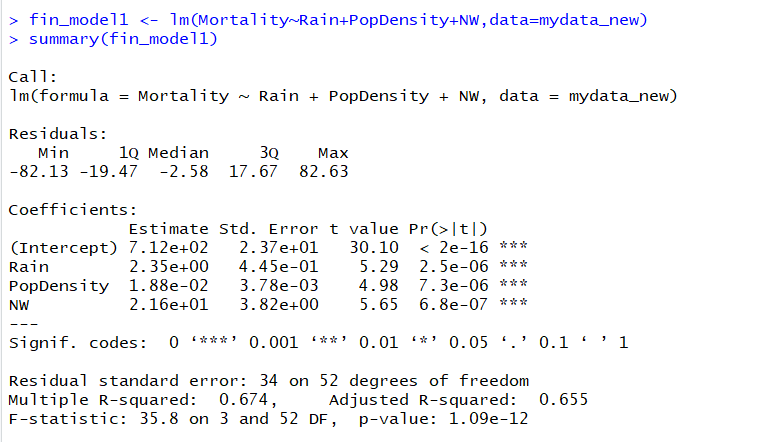
The regression coefficients might be unstable because of multicollinearity with many variables. Multicollinearity is a situation where two or more predictors are highly linearly related. Regression coefficients are more stable with parsimonious models. This is the reason why running a regression model with only the significant variables is an important step.

6b)



For the new model analyzed above, R-squared is 0.674 and adjusted R-squared is 0.649. It means that 67.4% variation in dependent variable is explained by 67.4% variation is independent variable the model is 64.9% fit.

However, we see that HCPot which was previously marginally significant is not significant now. So, we remove this variable and perform the regression again and find below results.



We see that when we remove the independent variable HCPot there is hardly any noticeable change in R-squared and adj R-squared. Hence, we find that this model is better fitted as compared to the previous models. Here R-squared and adj. R-squared being 0.674 and 0.655 respectively.

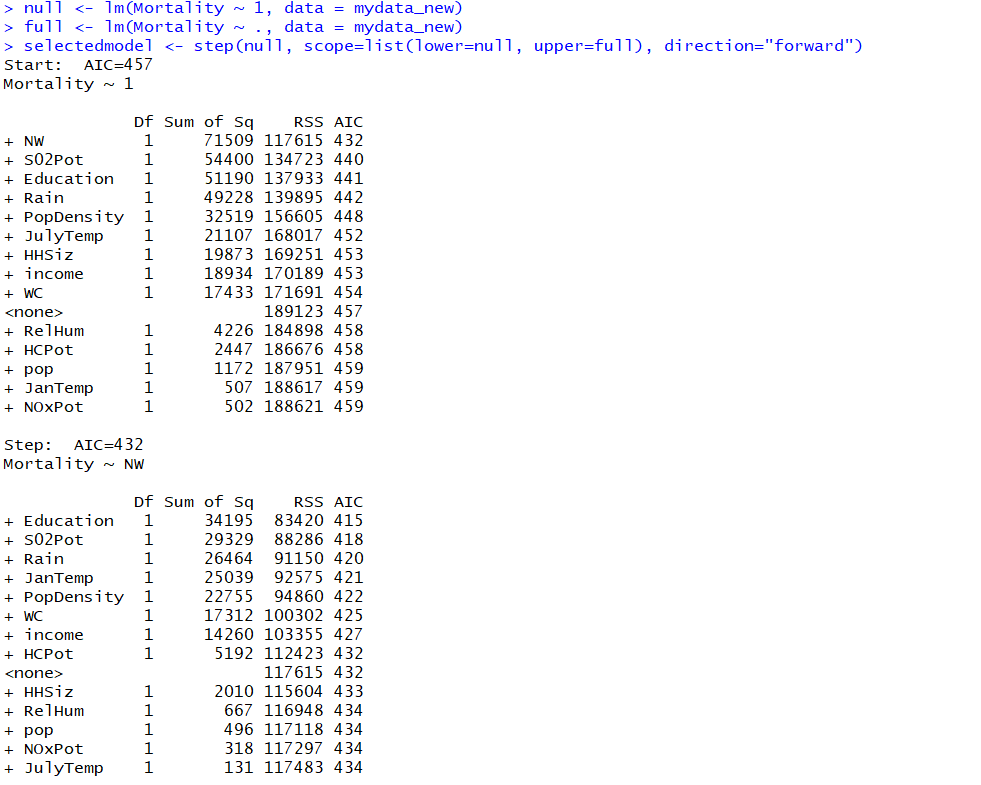
6c)

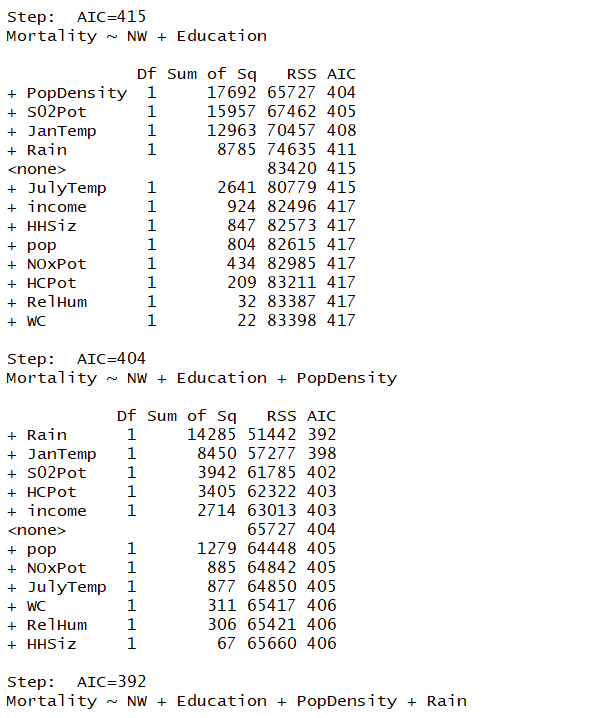
We observe that independent variables Rain (Pr(>|t|= 2.5e-06), PopDensity (Pr(>|t|= 7.3e-06) and

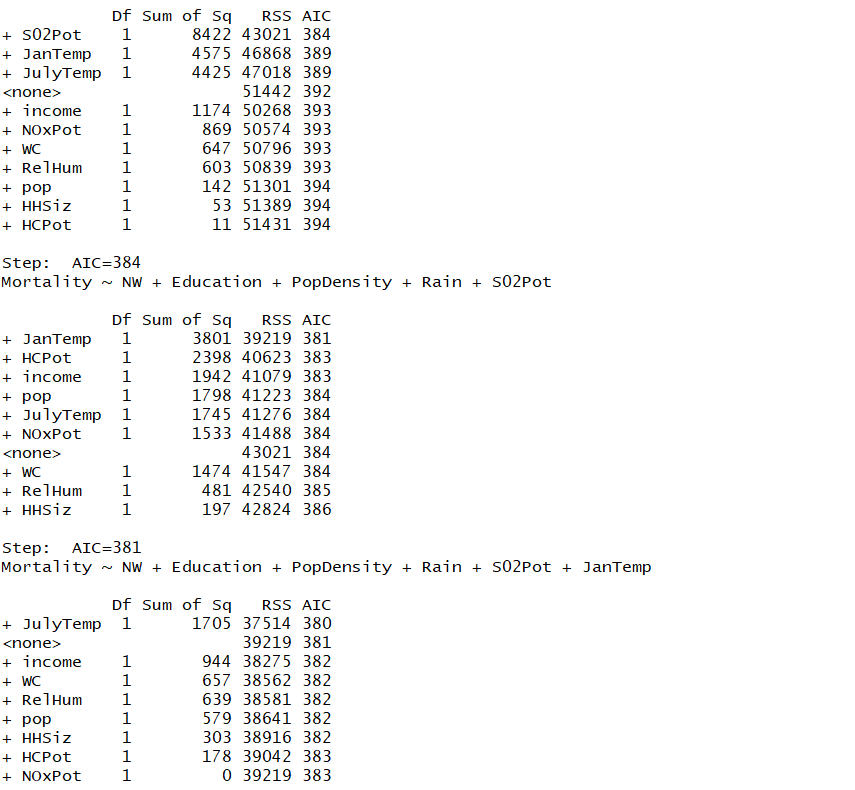
NW (Pr(>|t|= 6.8e-07) are significant since their significance is 99.99% as shows by the R significant codes.

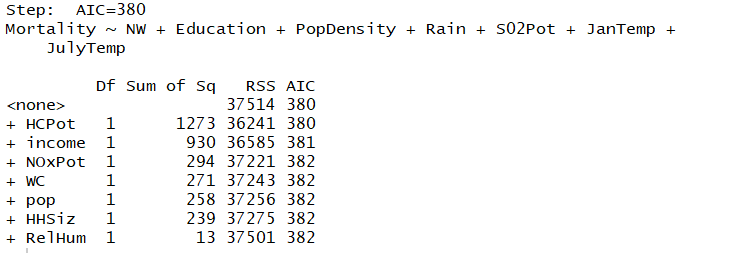
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

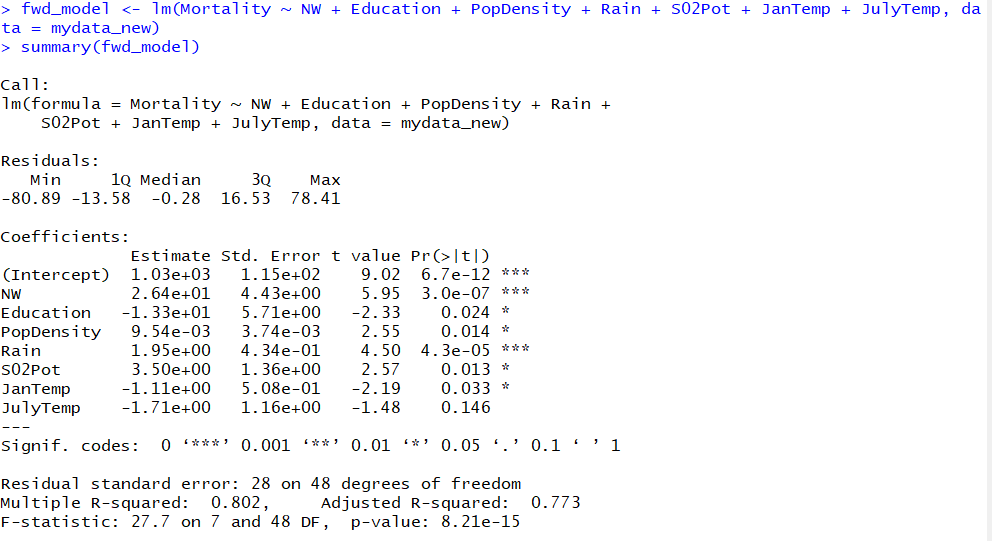
There is another method known as stepwise regression method which gives the list of significant variables. Below is forward regression sample to obtain significant independent variables. In forward selection, we start with no predictors and then add predictors one by one. Each predictor added is the one that has the largest contribution to R-squared on top of the predictors that are already in it. The algorithm stops when the contribution of additional predictors is not statistically significant.



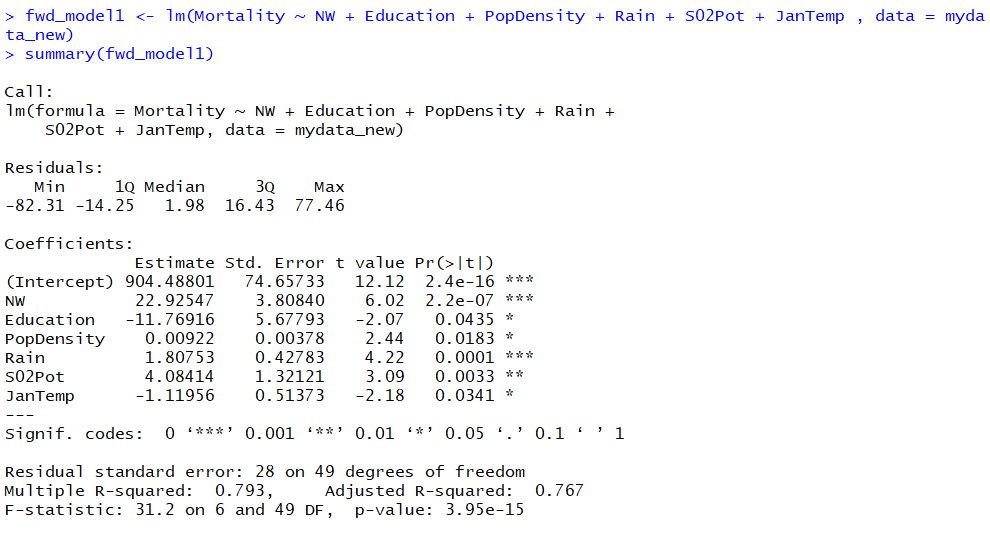




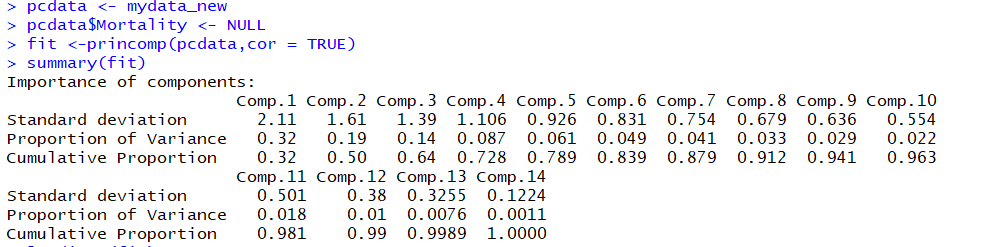


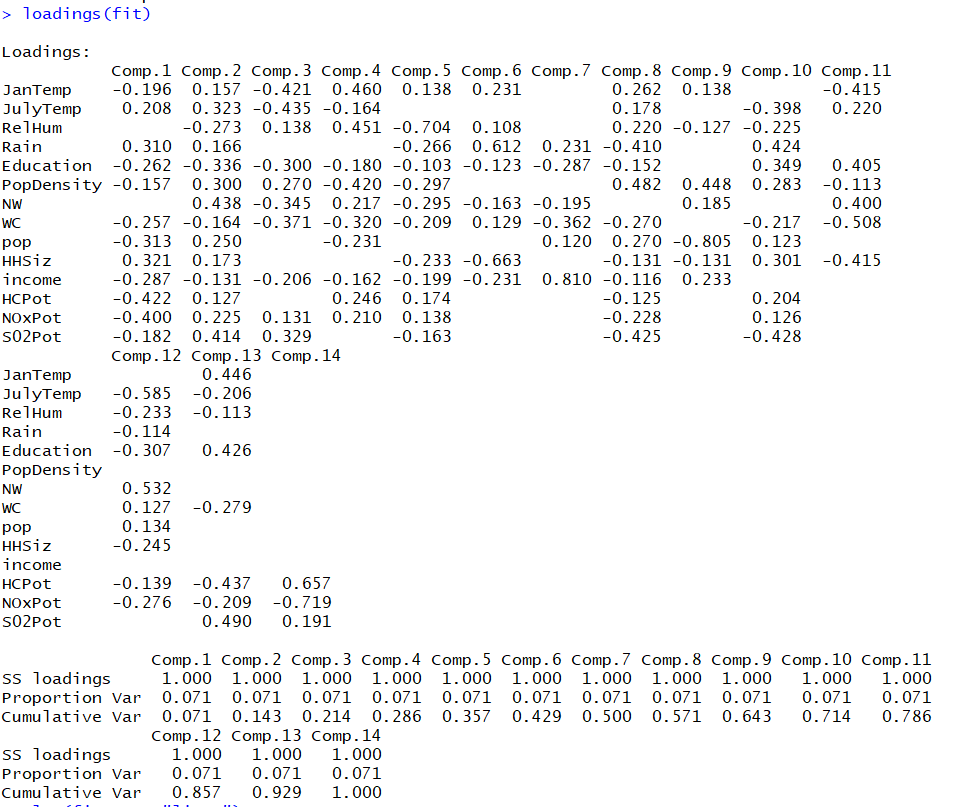


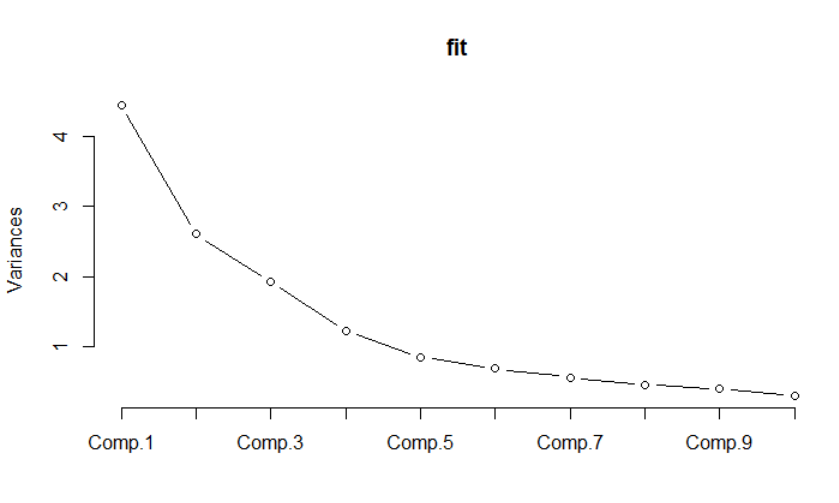
The above model as suggested by forward regression shows R-squared as 0.802 and adj R-squared to be 0.773. However, we see that independent variable JulyTemp has become insignificant thus removing that we find the model below which does not show much variation in adj.R-squared with lesser number of predictors.



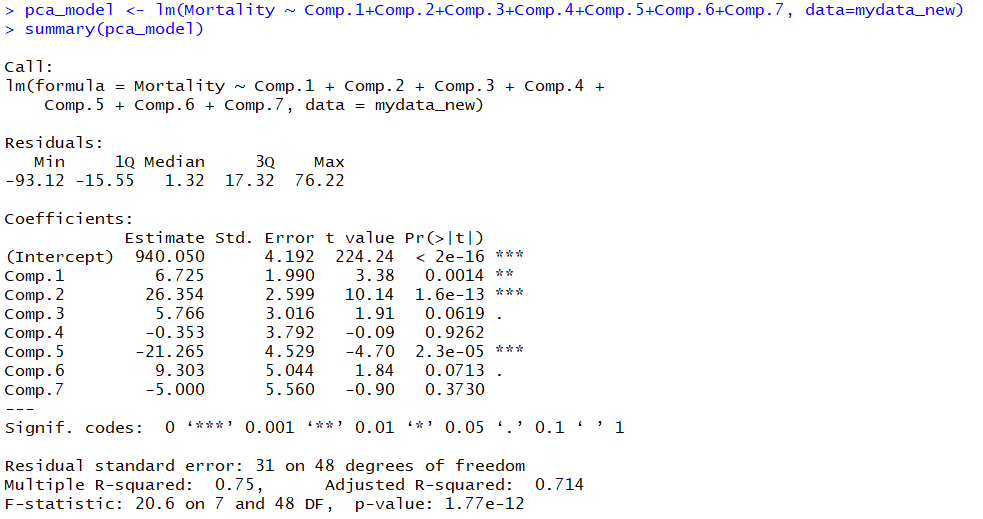
8)







Given the scree plot generated, we will select 1st seven (7) components. This is because after the 7th component we are not able to see much variation in the eigen values and the plot becomes parallel to the x-axis.



10a)

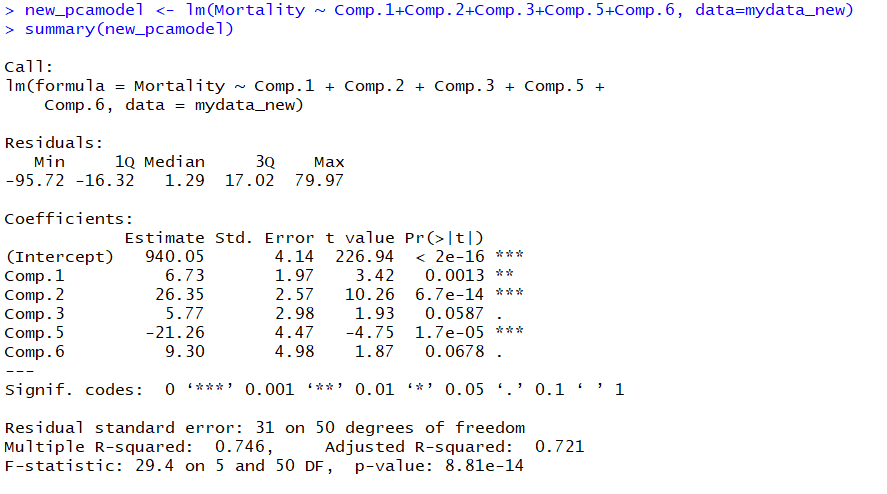
R- squared provides an estimate of the strength of the model and response variable. Here it shows that 75% variation in the dependent variable is explained by 75% variation in the independent variable.

Adjusted R-squared is a better indicator of the performance of the model since it considers only the variables that are significant to the model and penalizes the inclusion of predictors which are insignificant. The adjusted R-squared increases only when the new predictor improves the model. When considered the case here the value shows the model is 71% fitted.

10b)

From the above model we observe that Comp.1, Comp.2 and Comp.6 are 99.9% significant since their respective Pr|t| values are 0.0014, 1.6 e-13 and 2.3 e-05. Alongwith, the significant variables we have Comp.3 and Comp.6 being marginally significant, their Pr|t| being 0.0619 and 0.0713 respectively.

11a)



When the regression was executed with the significant variables from previous model, we see that after removing the insignificant variables the R-squared is 0.746 and adj. R-squared is 0.721. Hence the model is 74% strong and 72% fitted.

11b)

Comp.1, Comp.2, Comp.5 are highly significant with Pr|t| being 0.0013, 6.7 e-14, 1.7 e-05 and Comp.3 and Comp.6 are marginally significant with Pr|t| values being 0.0587 and 0.0678

12)

We find the models generated show different characteristics when different predictors are used. The previously generated models are not as stable as the one generated after PCA. And, the PCA model is more fitted as compared to the others. The model is 72% fitted and 74% strong. We also see that p-value is 8.81 e-14 which is substantially low and shows the model is a good fit.