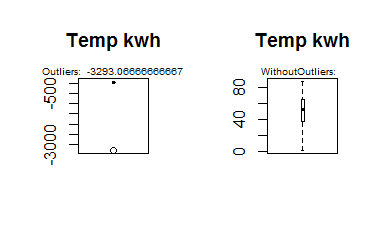
**Assignment -1-Energy Forecast**

**Team 9**

**Part1: Data wrangling and cleaning**

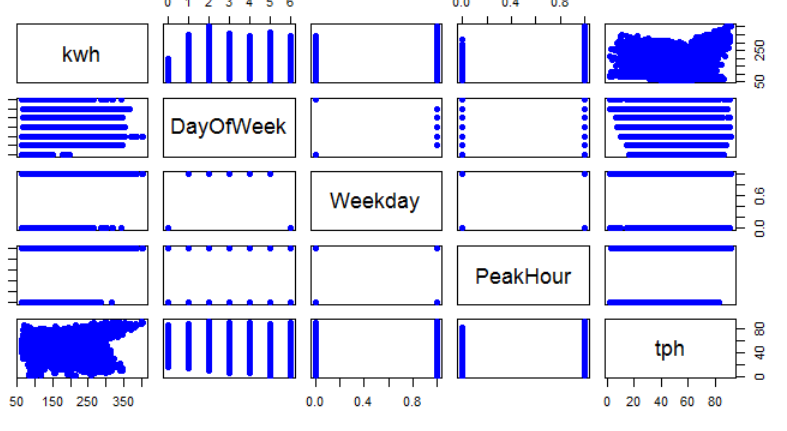
1. **Data is cleansed and temperature is merged using R into final file sampleformat.csv**
2. **Following Methods were used to transform the data and remove outliners**

* ***Histogram* of all variables were plotted to see the dependency of each predictors on output.**
* ***Scatter plot* was plotted to see the behavior of Kwh with all predictors.**
* ***BoxPlot* was plotted to see outliers and how the value varies.**
* **We also performed Influential of all variables through *Cook’s* *distance* so that we get a clear idea of the outliers and extremes in dataset.**

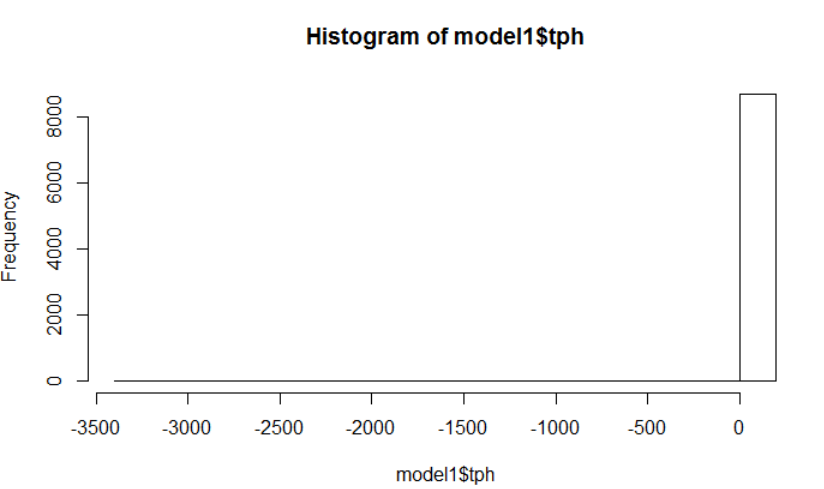
****

**Fig2: Box Plot**

**Fig 2: Description - For cleansing the data we plotted a box plot which shows the outliers in the data. After removing the outliers, we can see the box plot does not show any outliers.**

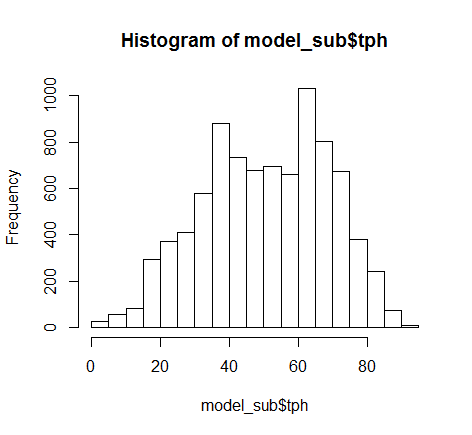
****

**Fig2: Scatter Plot**

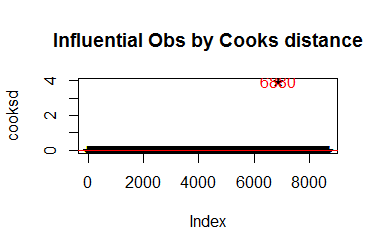


**Fig3: Histogram Plot**

**Fig 3 Description: For cleansing the data we also plotted a histogram to show the frequency of data which showed where most of the data actually lies and accordingly removed the outliers or so called data where frequency is too low. Below Graph shows Histogram when outliers were removed**

****

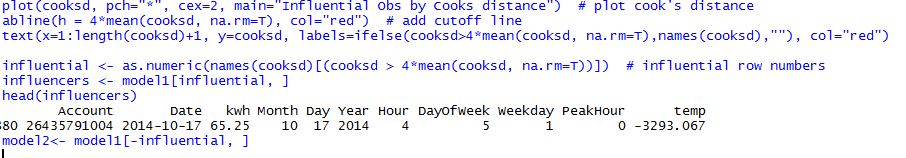
**Fig4: Histogram Plot after outliers removed**

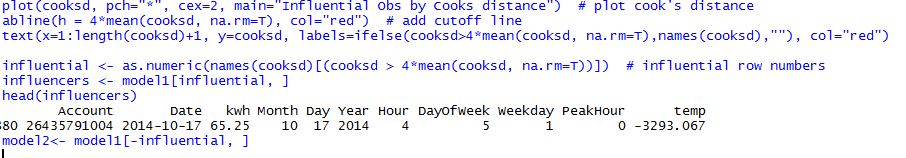
****

**Fig5: Cooks Distance**

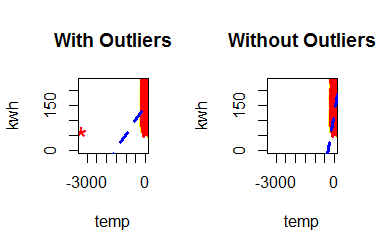
**Cooks distance gives the data points which are mostly influencing the model in a negative way i.e. we can set the threshold after looking at the data as in what part of the data we need to remove so as to build an effective model.**

**This plot showed the outliers and we removed those so as to cleanse the data.**

****

****

**Fig 6 Description : From the Below Plot we can see that the outlier influences our regression line shown in blue Color. Once the outlier is removed the regression line slope changes and its closer to residual points thus, minimizing the sum of residual errors.**

****

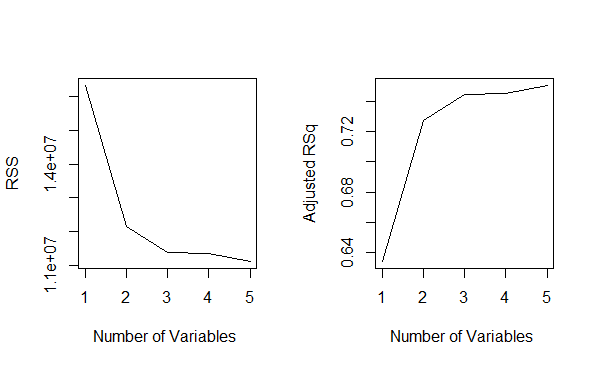
**Fig 6: Regression Line and Residual Points**

1. **Next we made subset of the dataset to get consistent data without outliners.**

**Part2: Multiple Linear Regression**

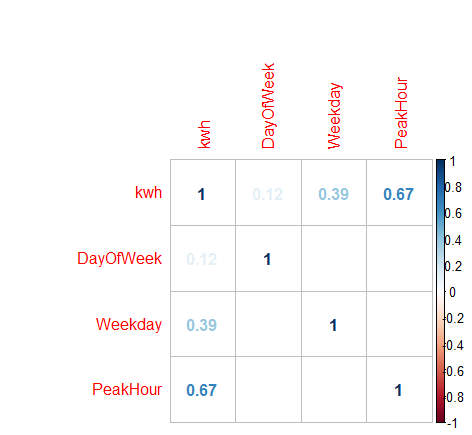
1. **Feature Selection was done for all the variables and we concluded that DayofWeek, PeakHour, Week are the variable with maximum influence to Kwh.**
2. **We used forward selection, backward Selection, Exhaustive search to infer this conclusion.**
3. **It was also observed that Temperature does not play much role when it comes to calculation but logically it plays an important factor to predict Kwh. *We Squared the value of Temperature to make it efficient predictor***
4. **Linear regression model was constructed with variables DayofWeek, PeakHour, Week and Temperature.**

**First we decided on number of variables to be included in our model and according to the below plot we can conclude that almost after 4 variables the R square and Adjusted R Square values are nearly constant.**

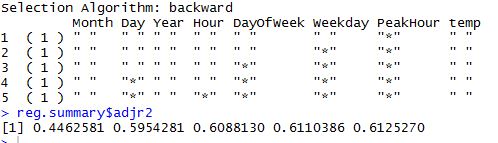
****

***Coefficient Matrix was plotted for evaluating the correlation between variables***

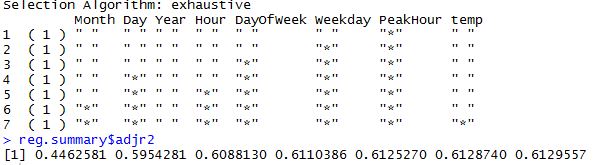
* **Once we decided on number of variables which is 4, we need to do feature selection on which 4 variables influences our model**

******

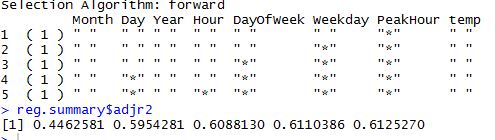
***Backward Algorithm:***

****

***Exhaustive Algorithm:***

****

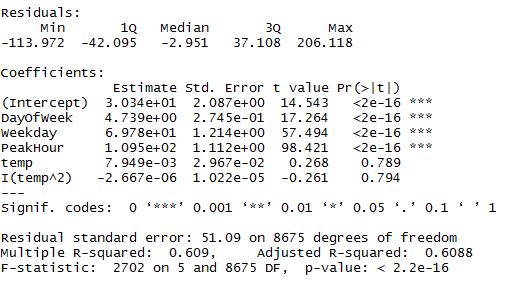
***Forward Algorithm:***

****

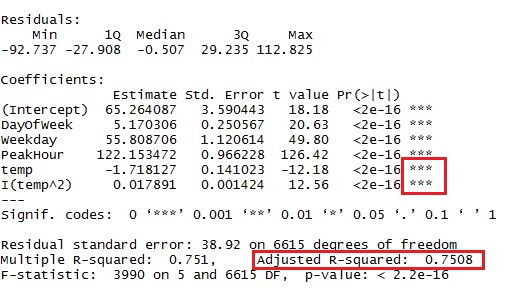
**# Feature Selection and Transformation**:

Transformation after data cooks cleansing, feature selection and feature transformation for temperature based on skew curve:

***Original Coefficients:***

****

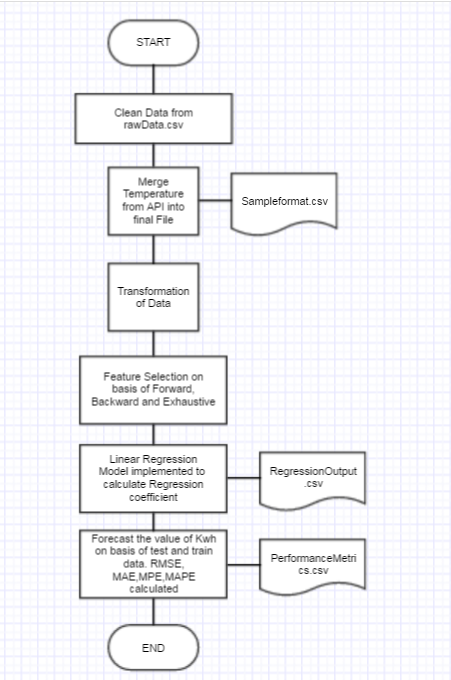
**RESULT: *Coefficients after Transformation:***

****

**Part 3: Forecast**

1. **The dataset was divided into 75% train and 25% test data to predict the Kwh.**
2. **Model was run on Test data and Kwh was forecasted using libraries.**
3. **The same was also forecasted against the data given by Professor.**

**FLOWCHART**

****