**Data Science Project Report**

**Topic: Bike Sharing Demand**

**Team Members:**

1. **Apurva Mandalika**
2. **Sindhu Krovvidi**
3. **R S D Pragnavi**

**Executive Summary:**

In this problem, the variation of historical usage patterns with weather, season, wind-speed, humidity, temperature and working day has to be determined, in order to forecast bike rental demand in the Capital Bike share program in Washington, D.C.

The given train dataset was splitinto training, validation and test datasets in the ratio 60:20:20. Then Data Mining methods namely Linear Regression and Random Forest were implemented on training and validation datasets to find the method that is more accurate.

Random Forest was found out to be more accurate, so it was applied on the test dataset.

**Problem Description:**

The total count of bikes rented during each hour covered by the test set must be predicted using only the information available prior to the rental period. This problem was previously addressed using Linear Regression, while we chose to use Random Forest to see we would get a more accurate result.

**Data Description:**

We are provided with hourly rental data spanning two years. The training set is comprised of the first 19 days of each month.

1. **Source:**

Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able rent a bike from a one location and return it to a different place on an as-needed basis. Currently, there are over 500 bike-sharing programs around the world.

The data generated by these systems makes them attractive for researchers because the duration of travel, departure location, arrival location, and time elapsed is explicitly recorded. Bike sharing systems therefore function as a sensor network, which can be used for studying mobility in a city.

1. **Data Fields** –

* datetime - hourly date + timestamp
* season -  1 = spring, 2 = summer, 3 = fall, 4 = winter
* holiday - whether the day is considered a holiday
* workingday - whether the day is neither a weekend nor holiday
* weather –

1: Clear, Few clouds, Partly cloudy, Partly cloudy   
2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds   
4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

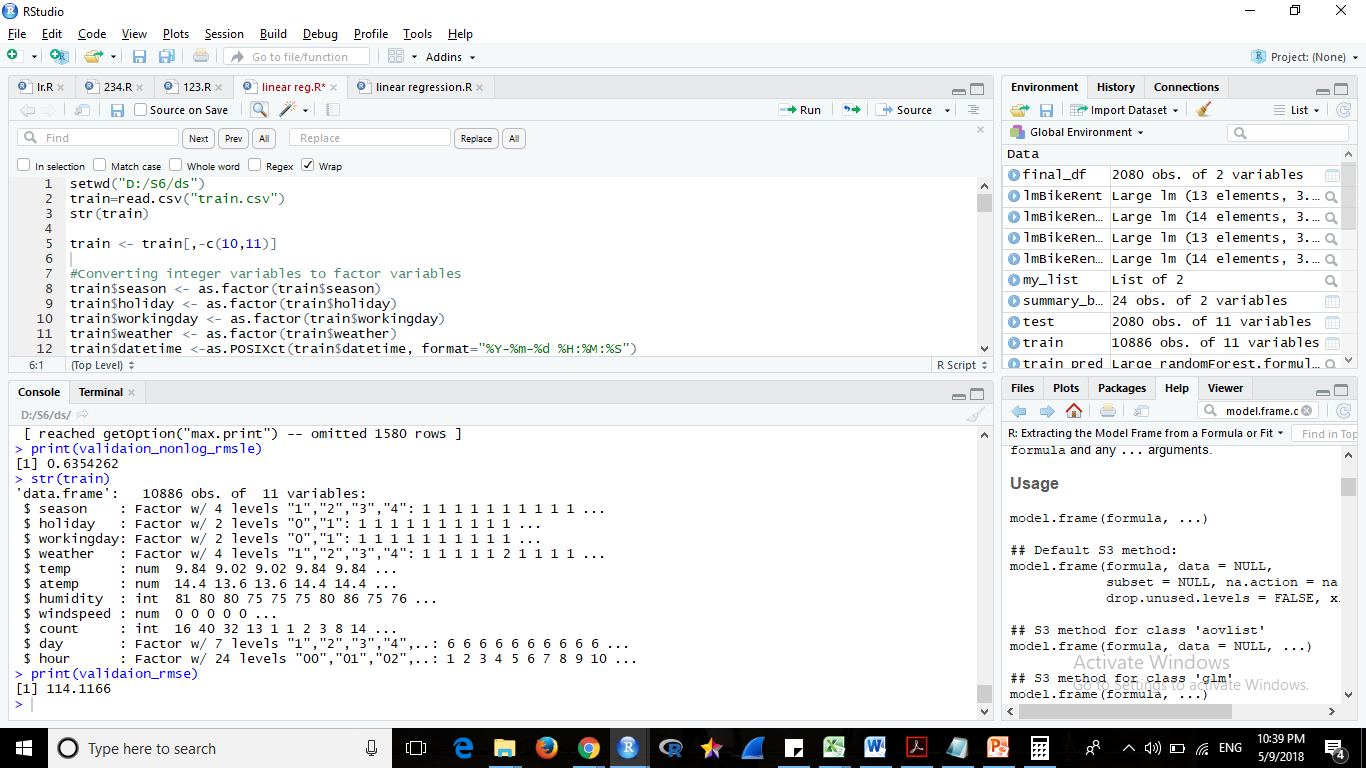
* temp - temperature in Celsius
* atemp - "feels like" temperature in Celsius
* humidity - relative humidity
* windspeed - wind speed
* casual - number of non-registered user rentals initiated
* registered - number of registered user rentals initiated
* count - number of total rentals

**Findings :**

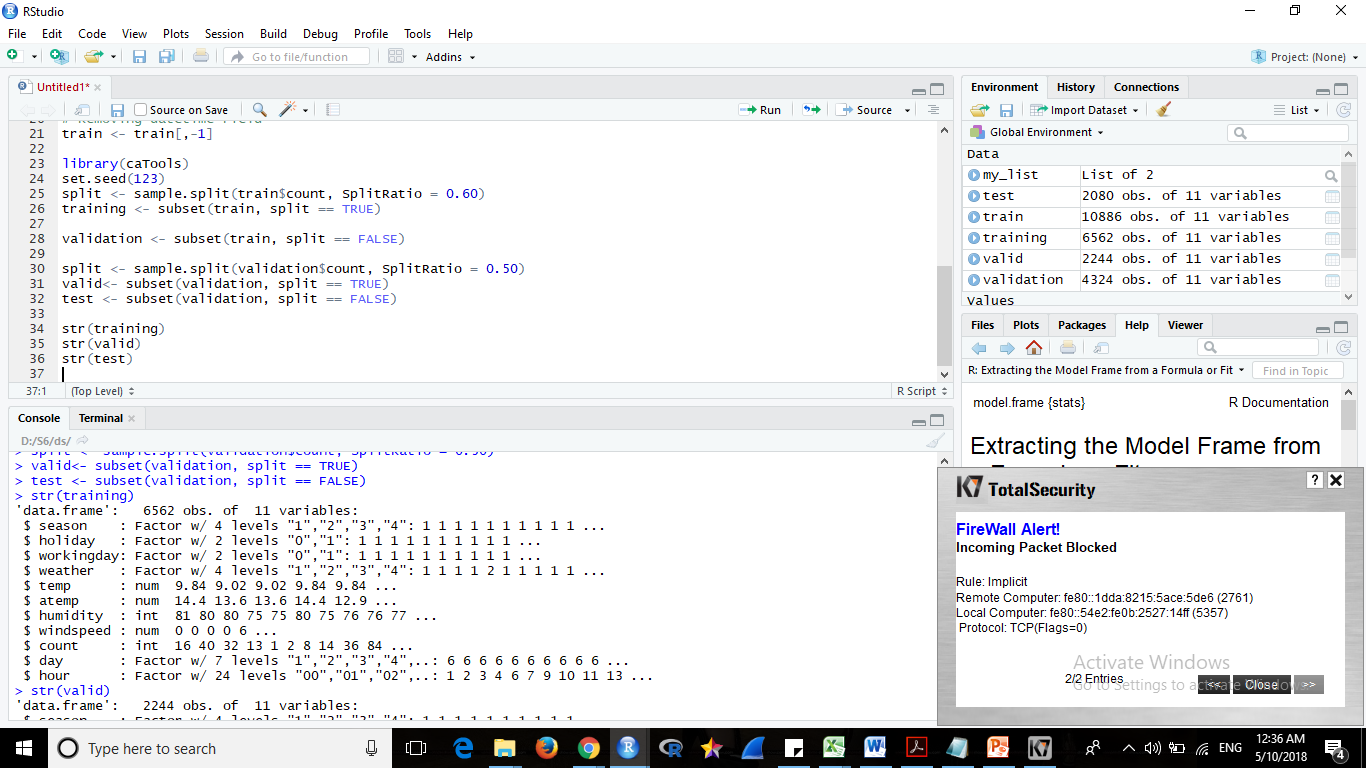
* Examined the dataset and found that ‘count’ column is the sum of ‘casual’ and ‘registered’ columns, therefore, ignored casual and registered columns. Converted the integer variables to factor variables. Extracted day and hour from the ‘datetime’ column . Split the dataset into ‘training’(train data) , ‘valid’(validation data) and ‘test’(test dataset) in the ratio 60:20:20 respectively. Then applied the following Data Mining models :
* Linear Regression
* Random Forest
* Common Implementation:
* **Read the csv files:**

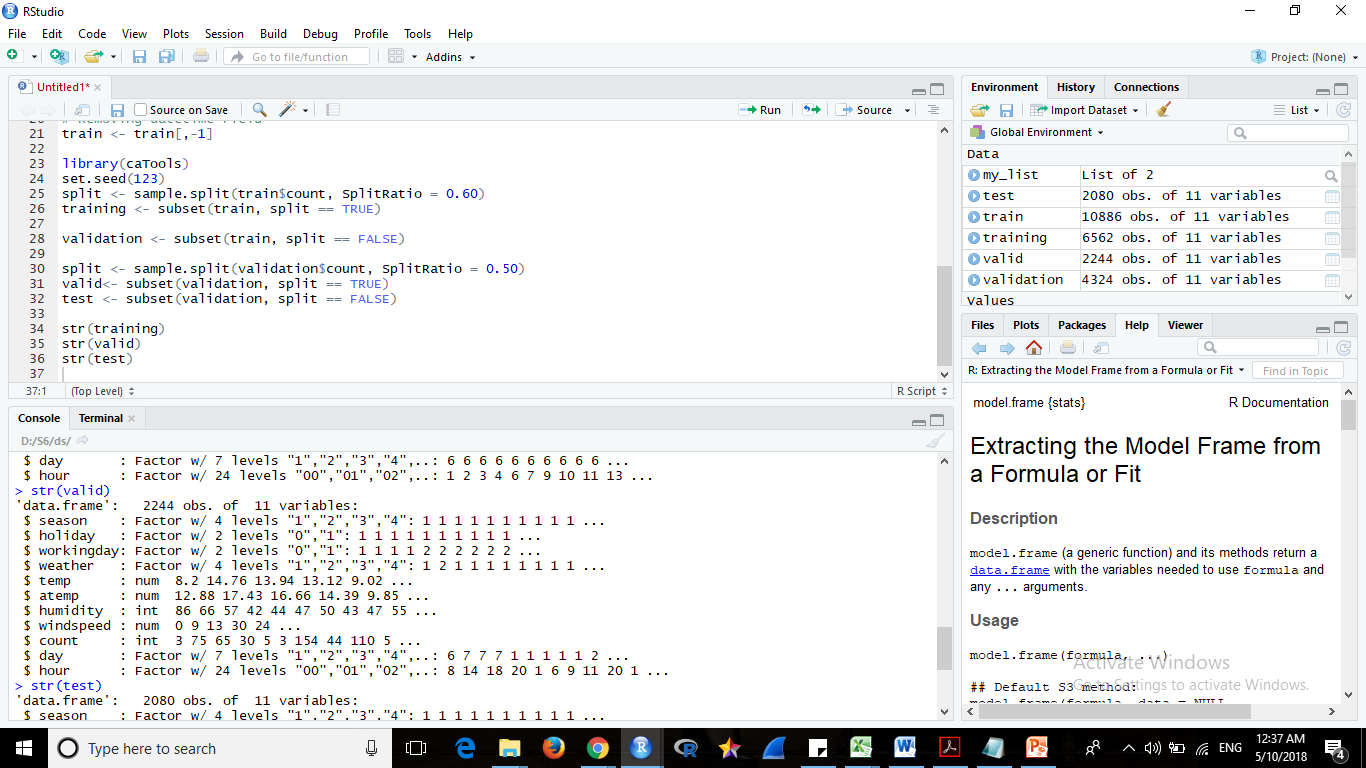
train=read.csv("train.csv")

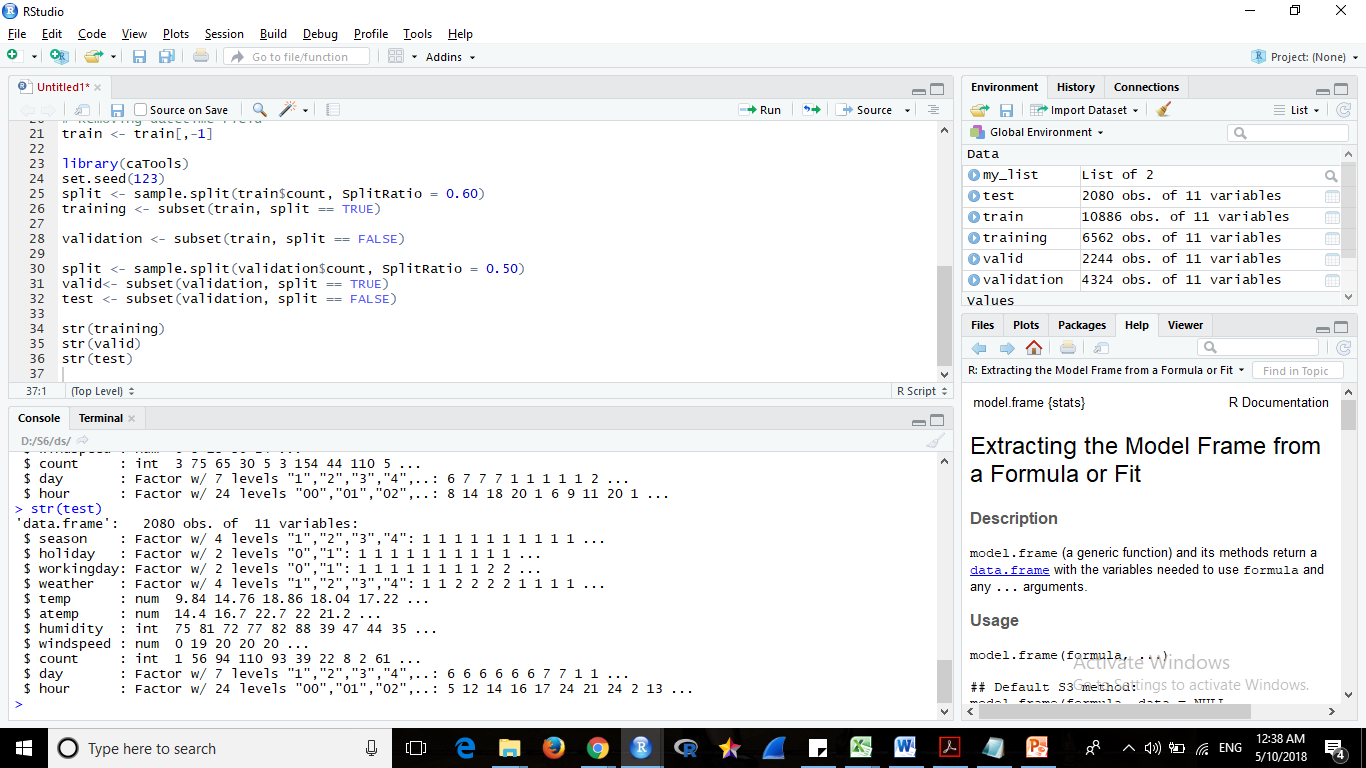
str(train)



* **Ignored the casual, registered columns as their sum is equal to count column.**
* **Split the Train dataset into training , valid and test datasets.**

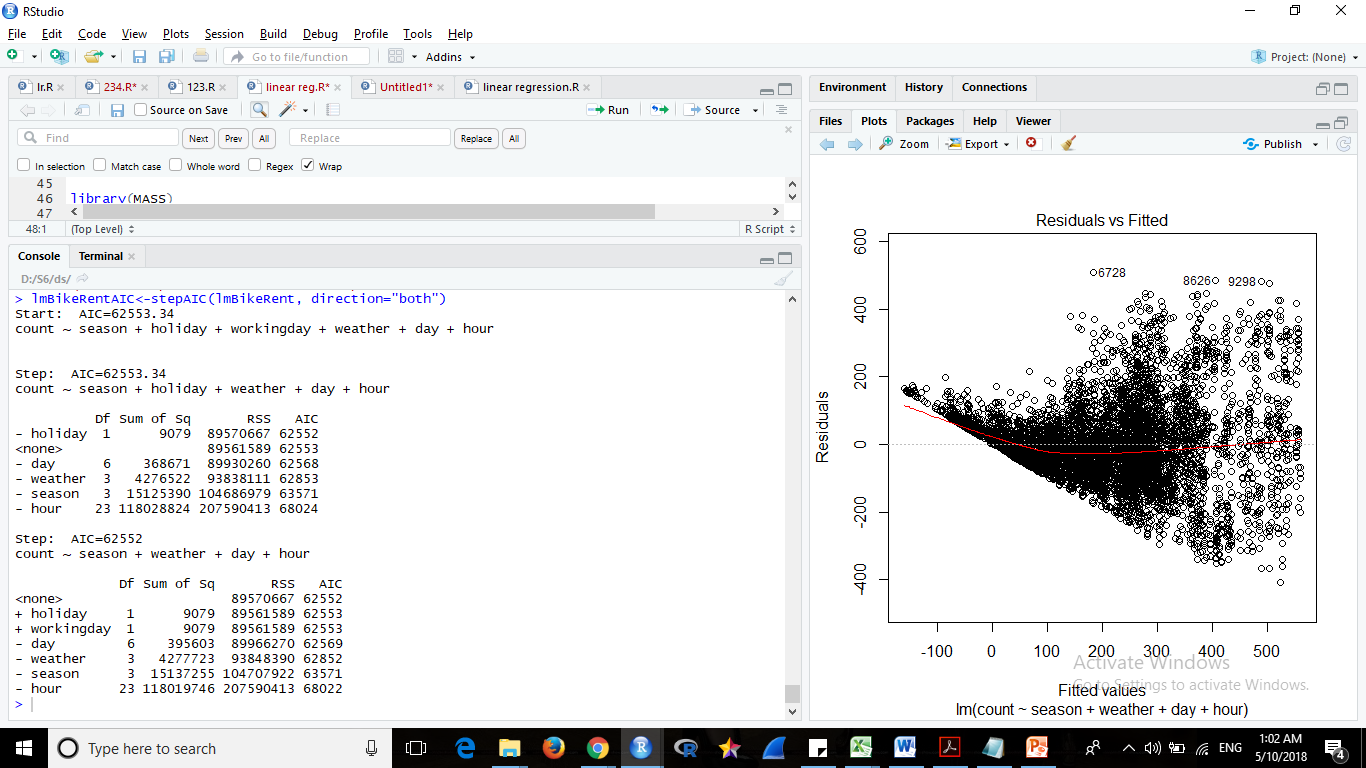




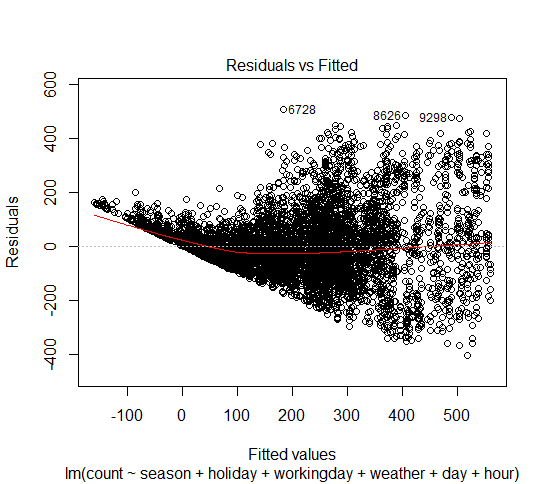


1. **Linear Regression Implementation :**

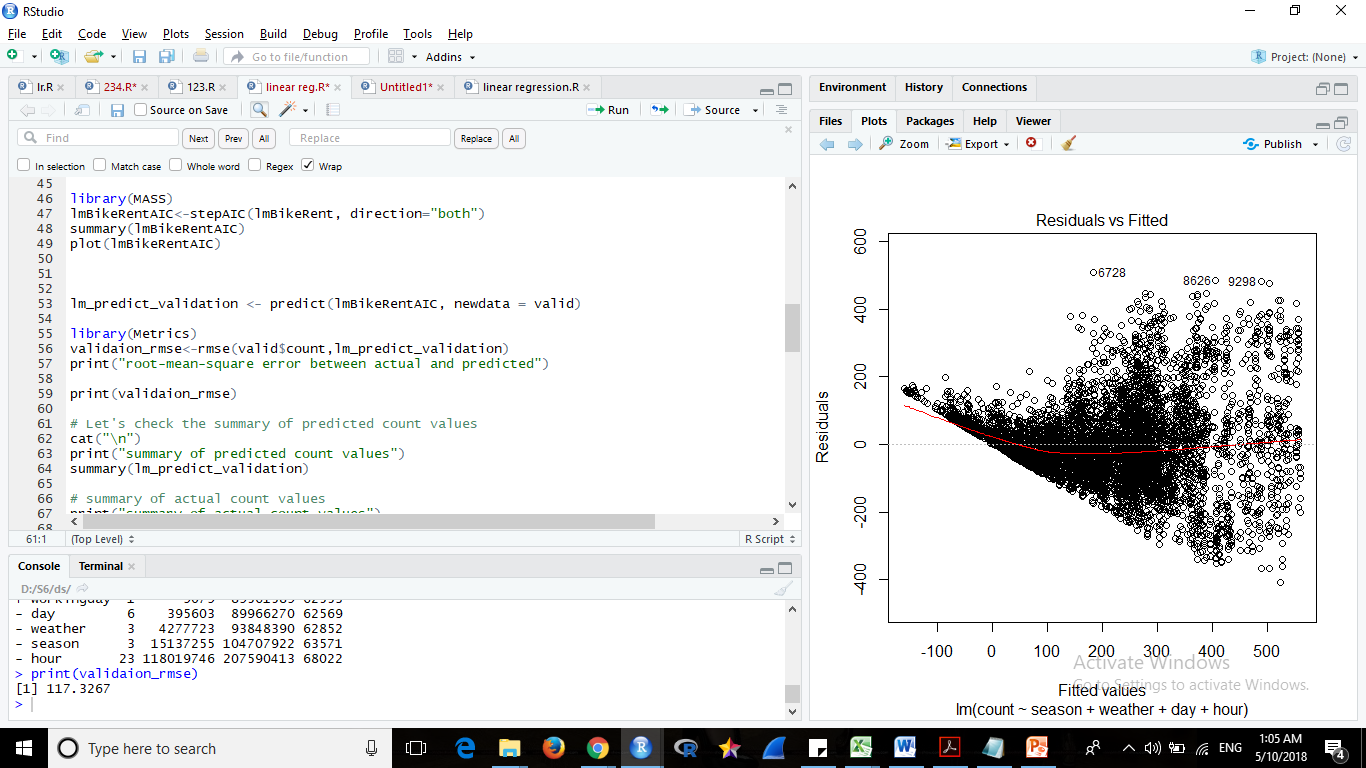
* **Applied Linear Regression model on the train dataset**
* **Performed step-wise model selection by AIC with both directions(Forward , Backward).**



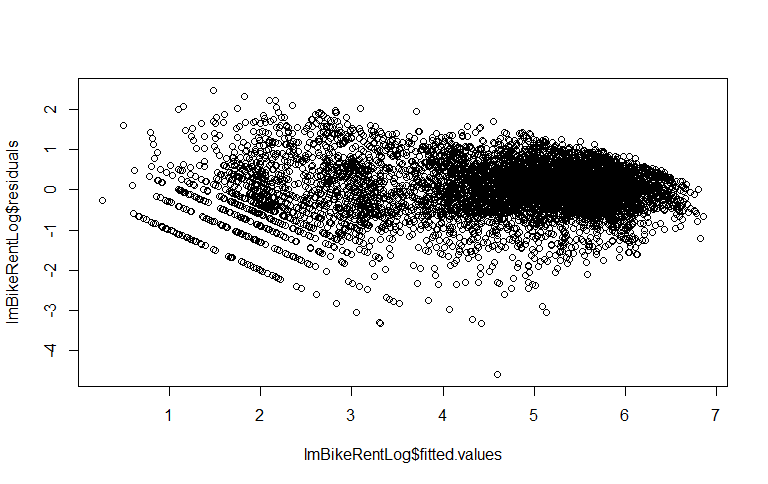
* **Plotted the model**



* **Applied prediction on validation set and computed the rmse value.**

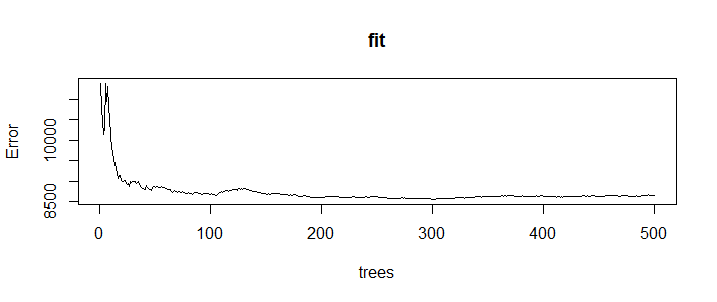


* **Plotted the residual vs fitted graph**

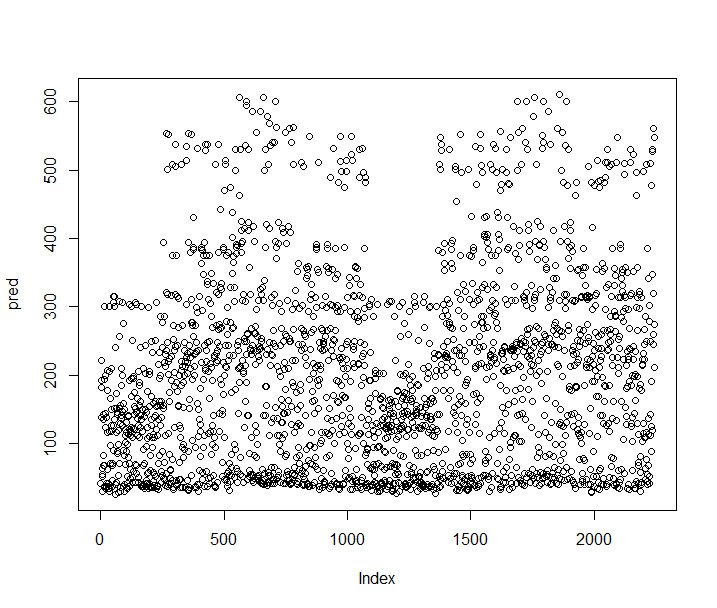
****

1. Random Forest Implementation :

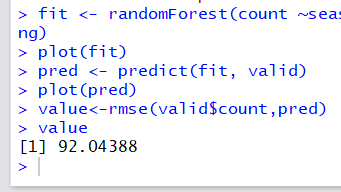
* **Applied** **Random Forest and plotted the model**



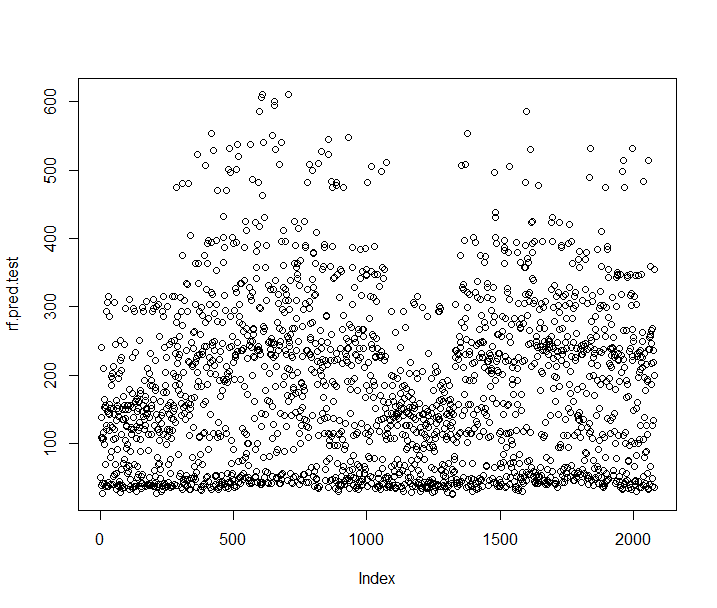
* **Applied model on validation set to get the count**



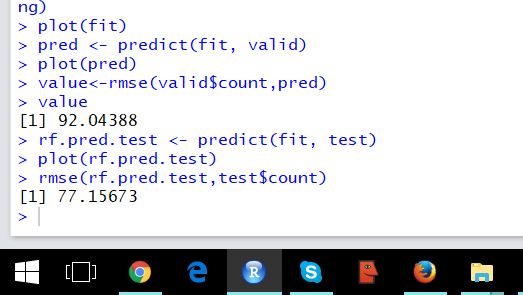
* **Calculated the rmse value**



* **Applied model on test set to get the count**



* **Calculated the rmse value**

****

**Conclusion**

* Linear regression and random forest methods were applied to the given dataset.
* The root mean square error of valid dataset was calculated on the models and the rmse value of linear regression model is found to be more than random forest.

linear regression rmse : 117.3267

random forest rmse : 92.04388

* So random forest was applied on the test dataset.
* The count of the rented bikes predicted using random forest was more accurate.