**Project Name - Bike Renting**

*Akhilesh Sayana*

*21 November 2018*

**Contents**

**1 Introduction 2**

1.1 Problem Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2

1.2 Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2

**2 Methodology 4**

2.1 Pre Processing . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

2.1.2 Feature Selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

2.2 Modeling . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

**Ordinary Least Squares**

**Random Forest**

**XGBoost**

**Cross Validation**

**Model Evaluation**

**Conclusion**

**Important Features**

**Chapter 1**

**Introduction**

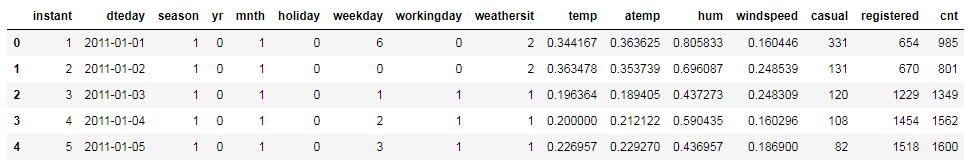
* 1. **Problem Statement**

‘Bike Sharing and Renting’ is a growing trend in the cities creating huge data associated with the climatic conditions and several other factors.

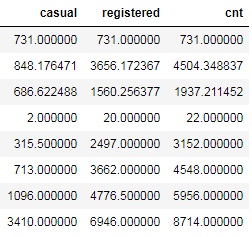
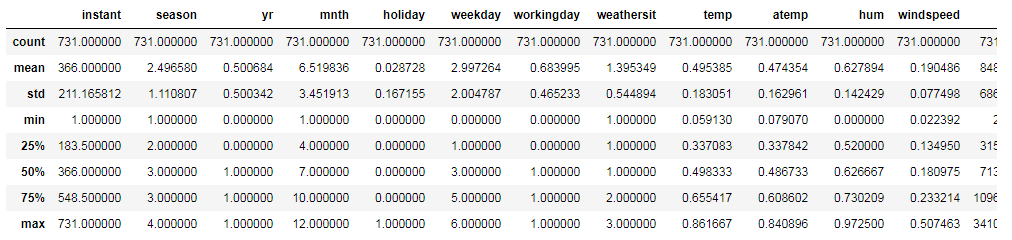
The aim of the project is to drive the insights for finding the count of bikes rented for a particular day with the given seasonal and environmental settings

**1.2 Data**

Our task is to build a prediction model which will predict the no. of bikes rented. The dataset for the same has been shared. The sample of the data and its description is shows below.



**FIG:** Sample Data



**FIG:** Description of Data

Dataset has 731 rows and 16 columns

**1.3 Attribute Information:**

* instant: Record index
* dteday: Date
* season: Season (1:springer, 2:summer, 3:fall, 4:winter)
* yr: Year (0: 2011, 1:2012)
* mnth: Month (1 to 12)
* hr: Hour (0 to 23)
* holiday: weather day is holiday or not (extracted fromHoliday Schedule)
* weekday: Day of the week
* workingday: If day is neither weekend nor holiday is 1, otherwise is 0.
* weathersit: (extracted fromFreemeteo)
* 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: Normalized temperature in Celsius. The values are derived via

(t-t\_min)/(t\_max-t\_min), || t\_min=-8, t\_max=+39 (only in hourly scale)

* atemp: Normalized feeling temperature in Celsius. The values are derived via

(t-t\_min)/(t\_maxt\_min), t\_min=-16, t\_max=+50 (only in hourly scale)

* hum: Normalized humidity. The values are divided to 100 (max)
* windspeed: Normalized wind speed. The values are divided to 67 (max)
* casual: count of casual users
* registered: count of registered users
* cnt: count of total rental bikes including both casual and registered

**Chapter 2**

**Methodology**

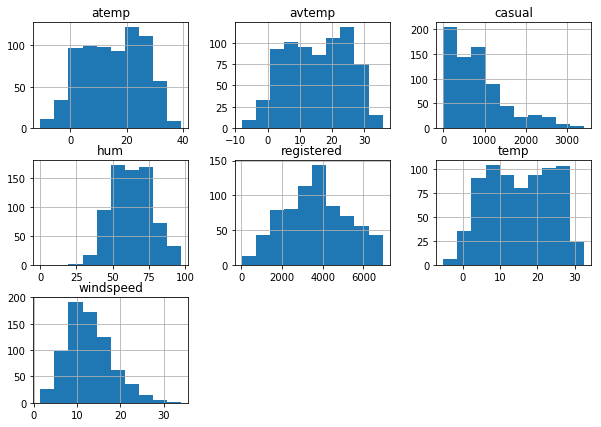
**2.1 Pre Processing**

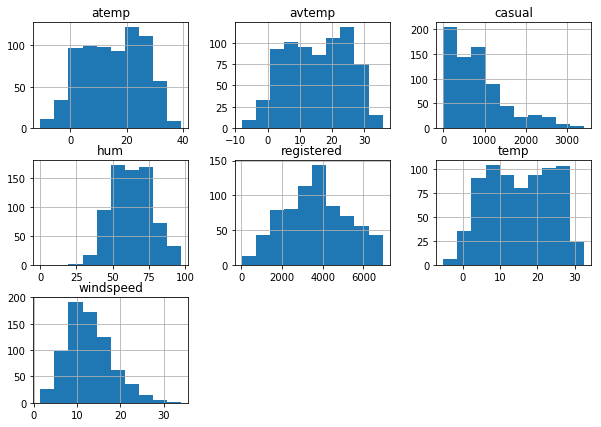
Any predictive modeling requires that we look at the data before we start modeling.

Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as Exploratory Data Analysis.

To start the process we will first try and look at all the probability distributions of the variables. Most analysis like regression, assume the data to be normally distributed.

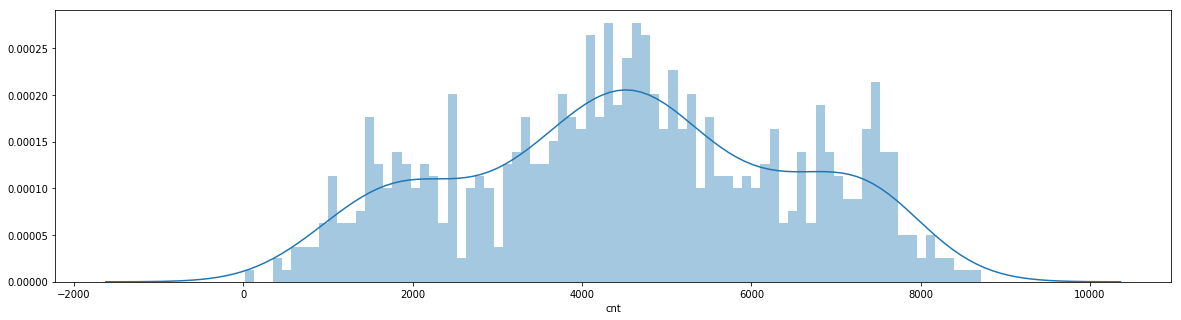
We can visualize that in a glance by looking at the distribution of several predictors:





The corresponding names of the continuous predictors are :. Temp, atemp, hum, windspeed, casual and registered.

Also we are interested in the variable which is directly related to the bike rental count – ‘cnt’.



The distribution of the target variable ‘cnt’ is nearly normal.

**2.1 EDA : Exploratory Data Analysis**

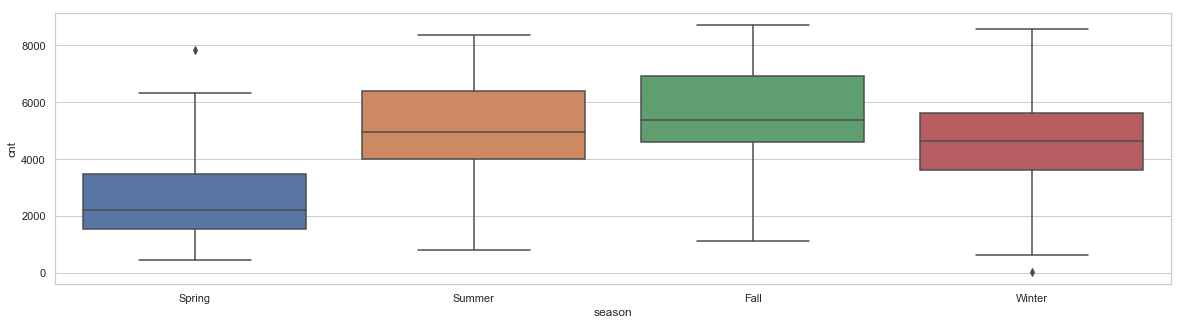
After having a look at the data we updated certain attributes in the data:

* There are temp and atemp, bot of the predictors are in the normalized form.
* We have converted the variables to their actual values.
* Also we have created a new variable ‘avtemp, which is an average of the two variables ‘temp’ and ‘atemp’.

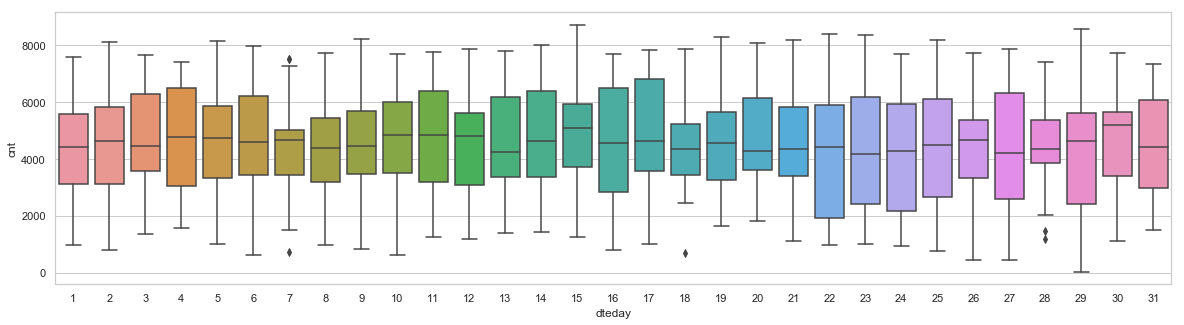
**2.4 Missing Value Treatment:**

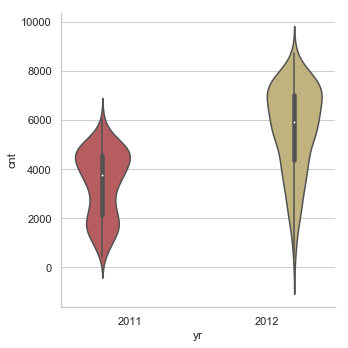
* This is a clean dataset and doesn’t have missing values

**2.5 Catagorical Data:**

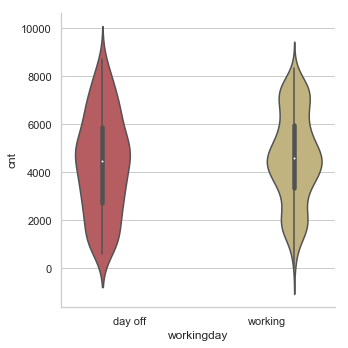
****

From the above figure it can be seen that the number of bikes rented was higher in the ‘summer’ and ‘Winter’ season. And was lowest in the ‘Spring

The distribution of ‘cnt’ seems to be constant throughout the month w.r.t the variable ‘dteday’.



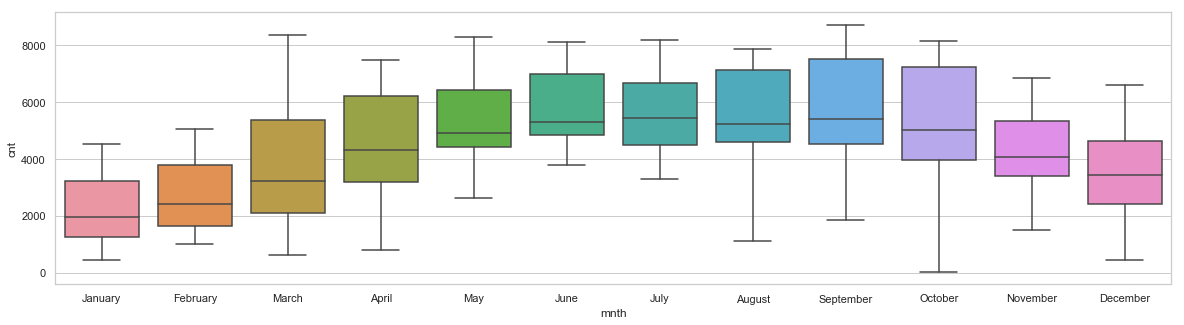
The count of rented bikes has increased in 2012 with respect to 2011 which can be observed using the above violin plot.

****

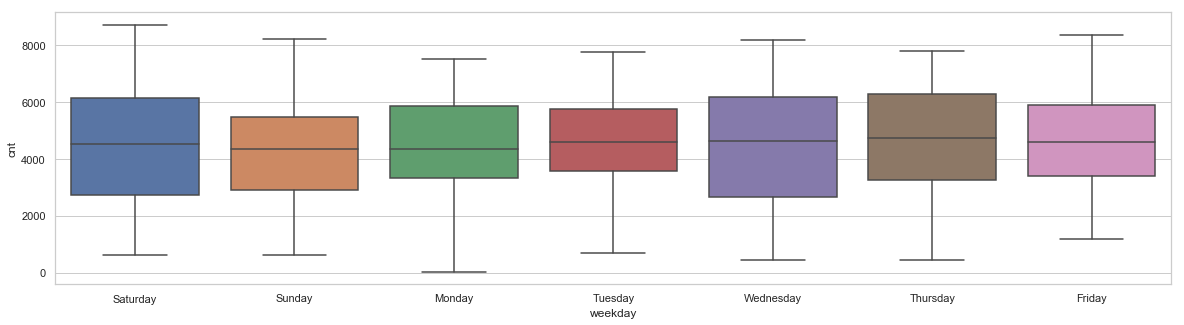
It has been analyzed that there is no overlap between holiday and week-offs. Holidays, working days and week-offs (Sat Sun) are independent of each other

Month wise speculation of the count of bike rentals- we can see that the months from May through October has seen the rentals with average above 4000.

And remained the lowest in the month of January with an average of less than 2000.



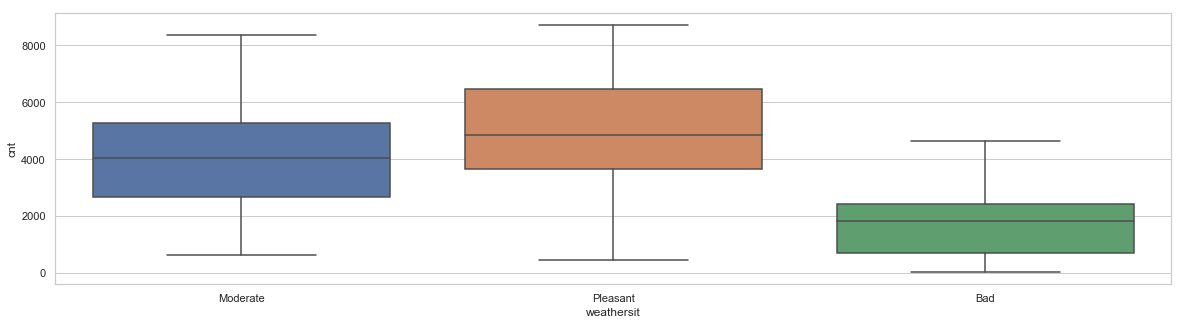
Week wise distribution. There is not a great difference in the distribution.



FINALLY, season has played an important role. The weather had not remained extreme.

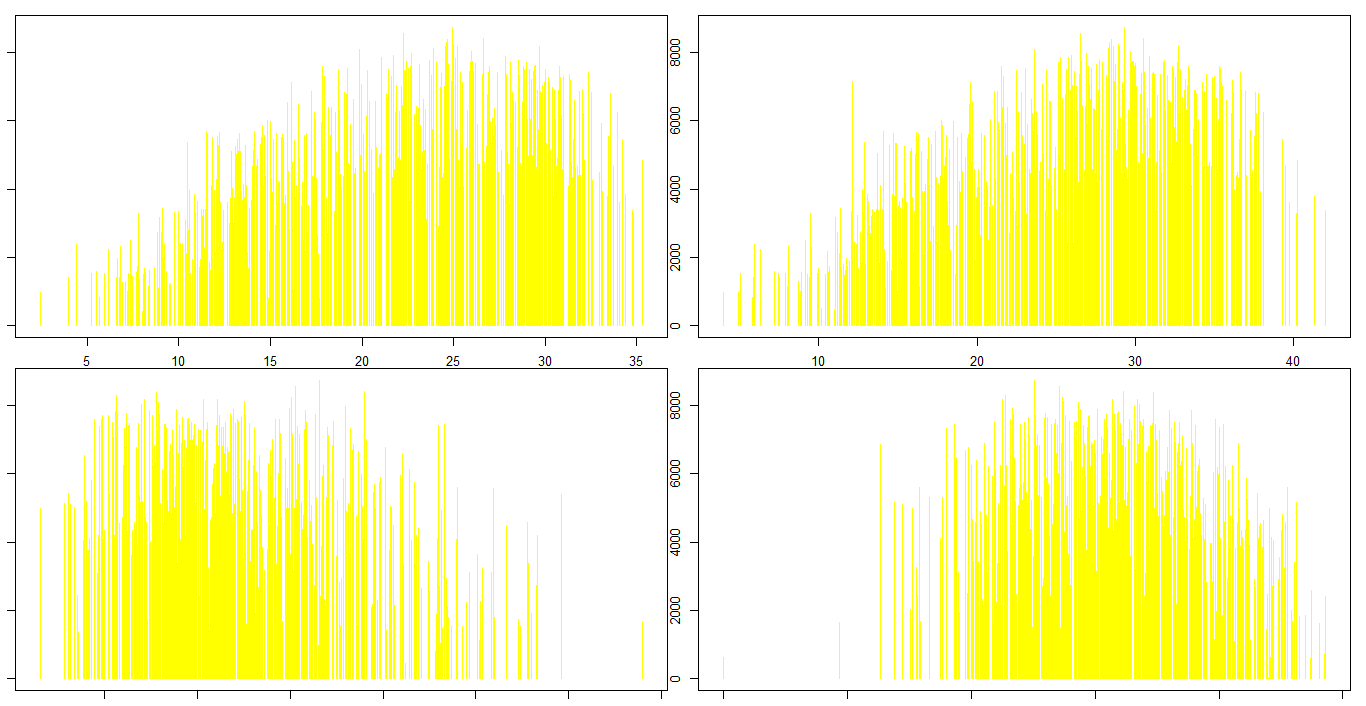
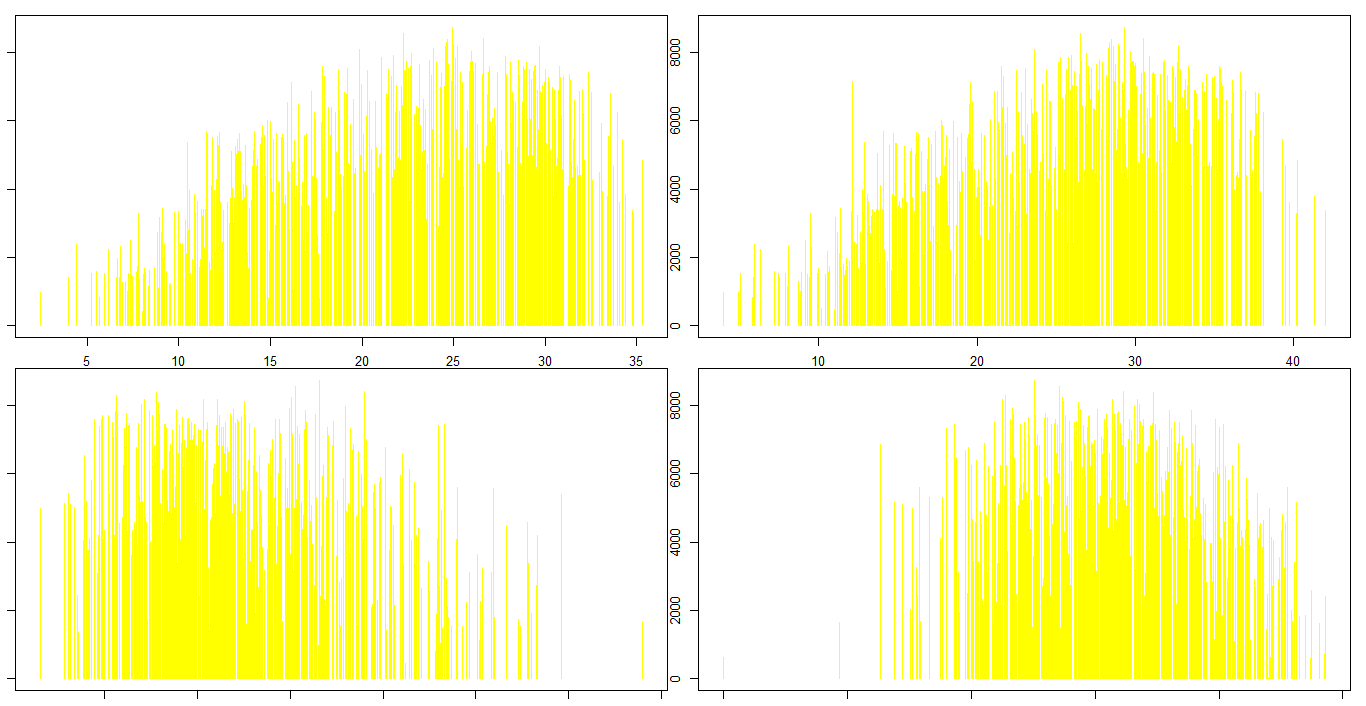
However, we can see that the count is highest when the weather was ‘pleasant’ with an average roughly around 5000 and around 4000 when the weather was ‘moderate’.

Also, we can see that the count average dropped down to less than 2000 when the weather conditions were not in favor.

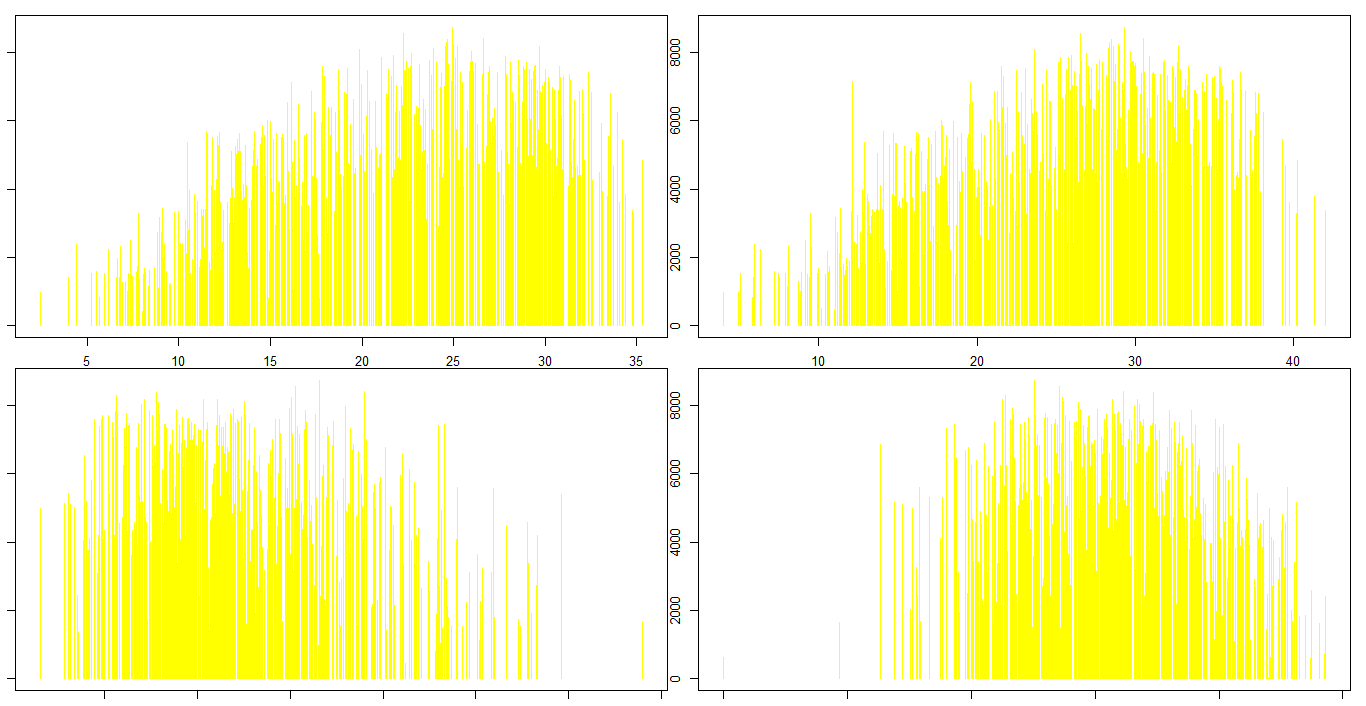
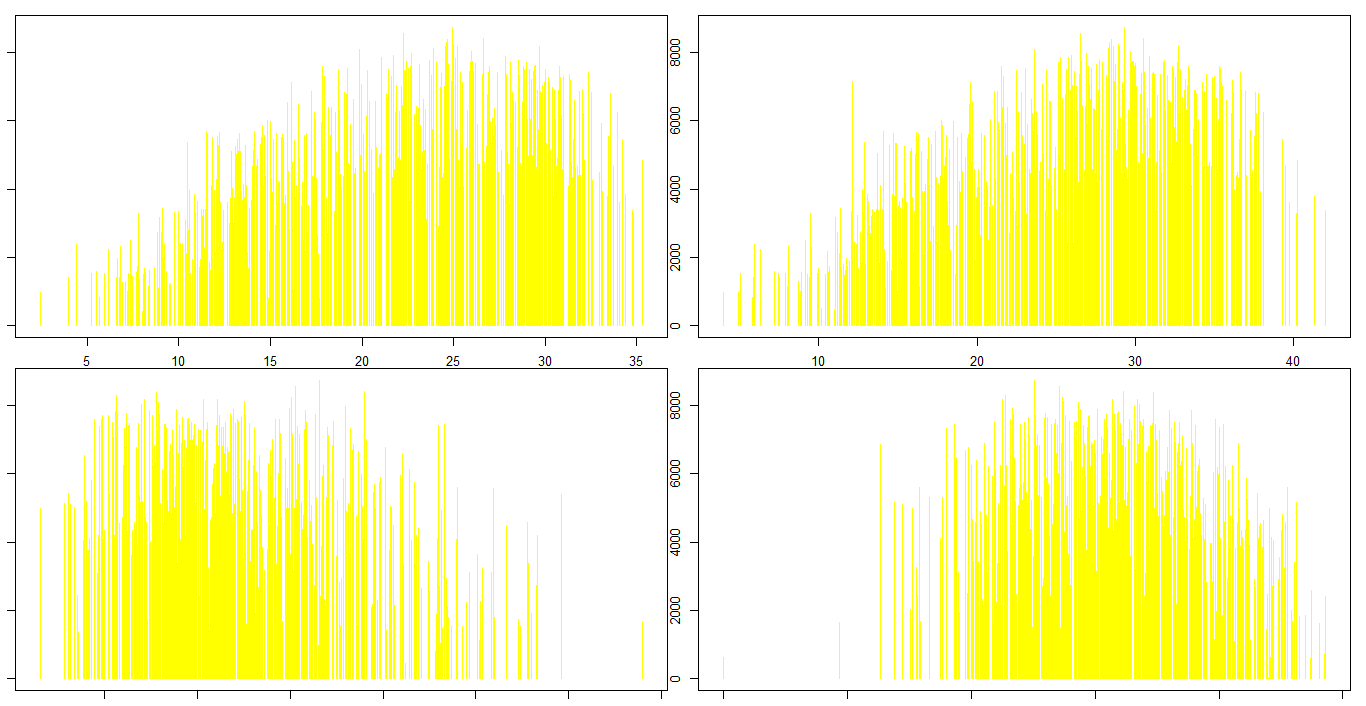


**Linear Relationship**

1. **Plot1 : Actual Temp vs bike count 2) Plot2 : Actual feel temp vs bike count**

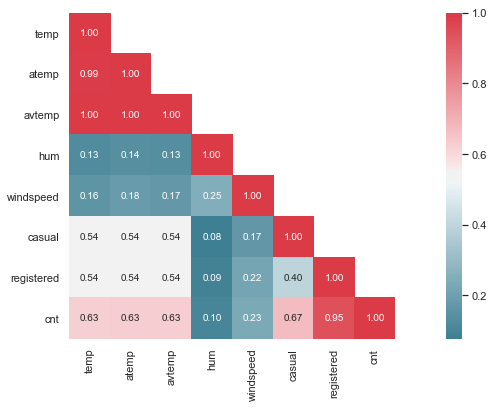


**3) Plot1 : Actual windspeed vs bike count 4) Plot2 : Humidity vs bike count**



Next>> FEATURE SELECTION

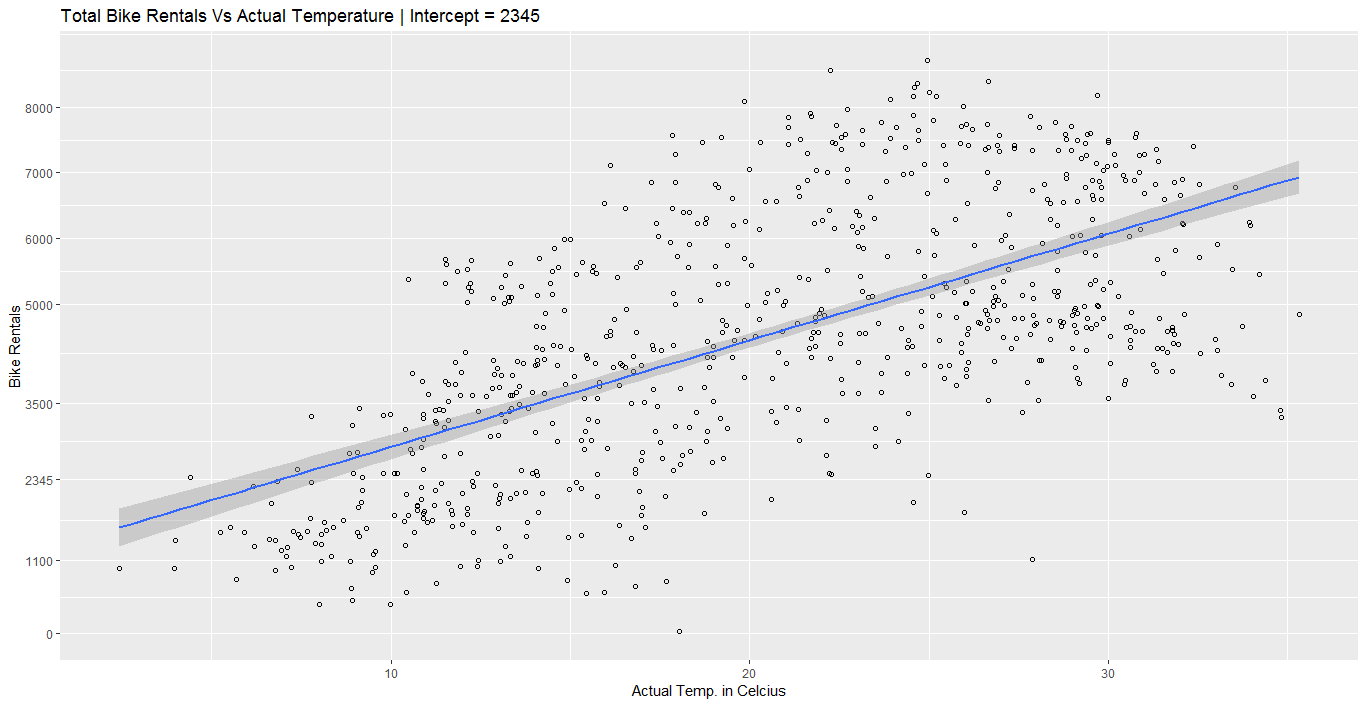
**3) FEATURE SELECTION**



The above correlation heatmap shows that casual and registered are highly correated.

Also, avtemp is derived from temp and atemp so these are also correlated which is obvious.

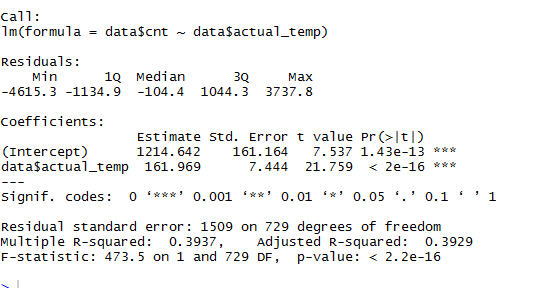
So, the major predictors here are, windspeed, hum and avtemp.

**SCATTER PLOT BETWEEN BIKE RENTALS and ACTUAL TEMPERATURE**

**4) MODEL BUILDING:**

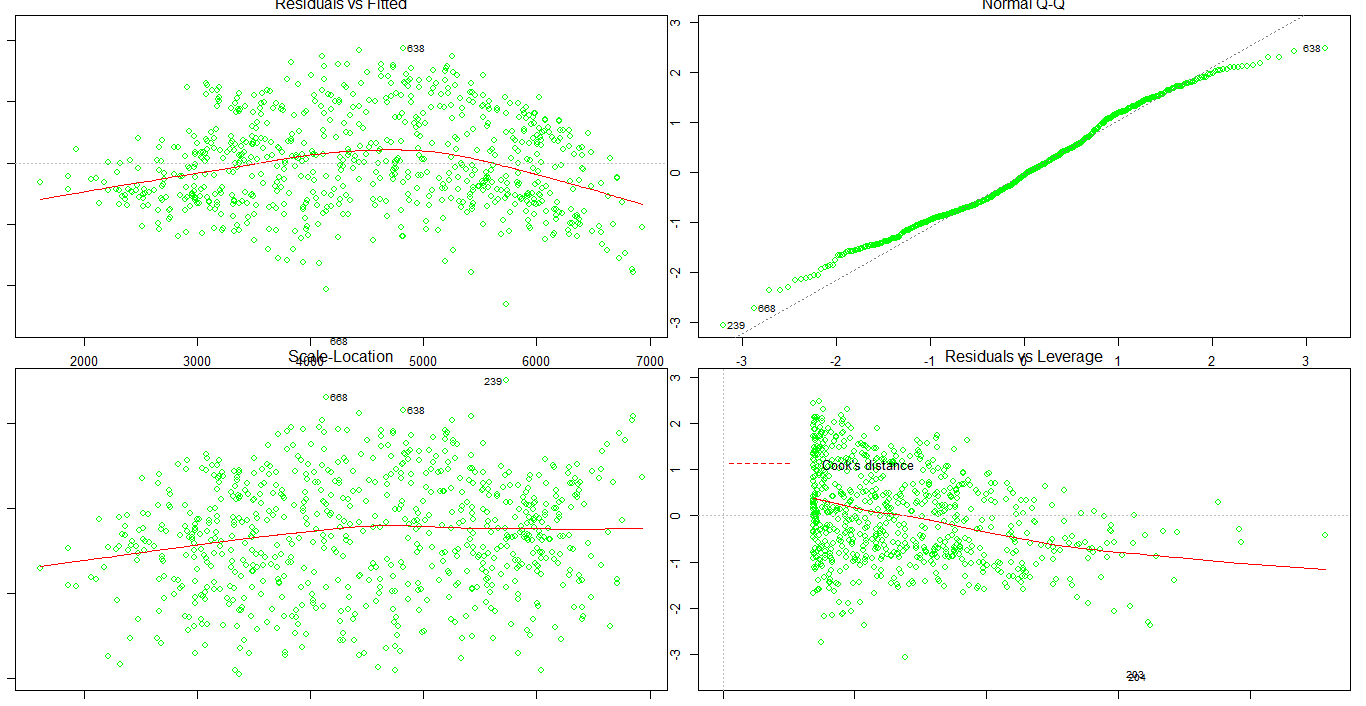
\*OLS Regression with actual temp as predictor:

MODEL 1

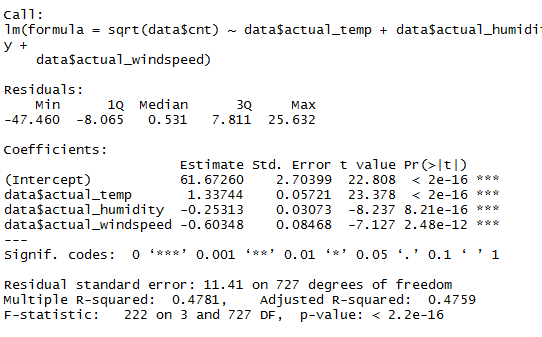


\*\*The model is significant with R-squared: 0.3937, coefficient 161.969,

intercept 1214.642 and the p-value: < 2.2e-16



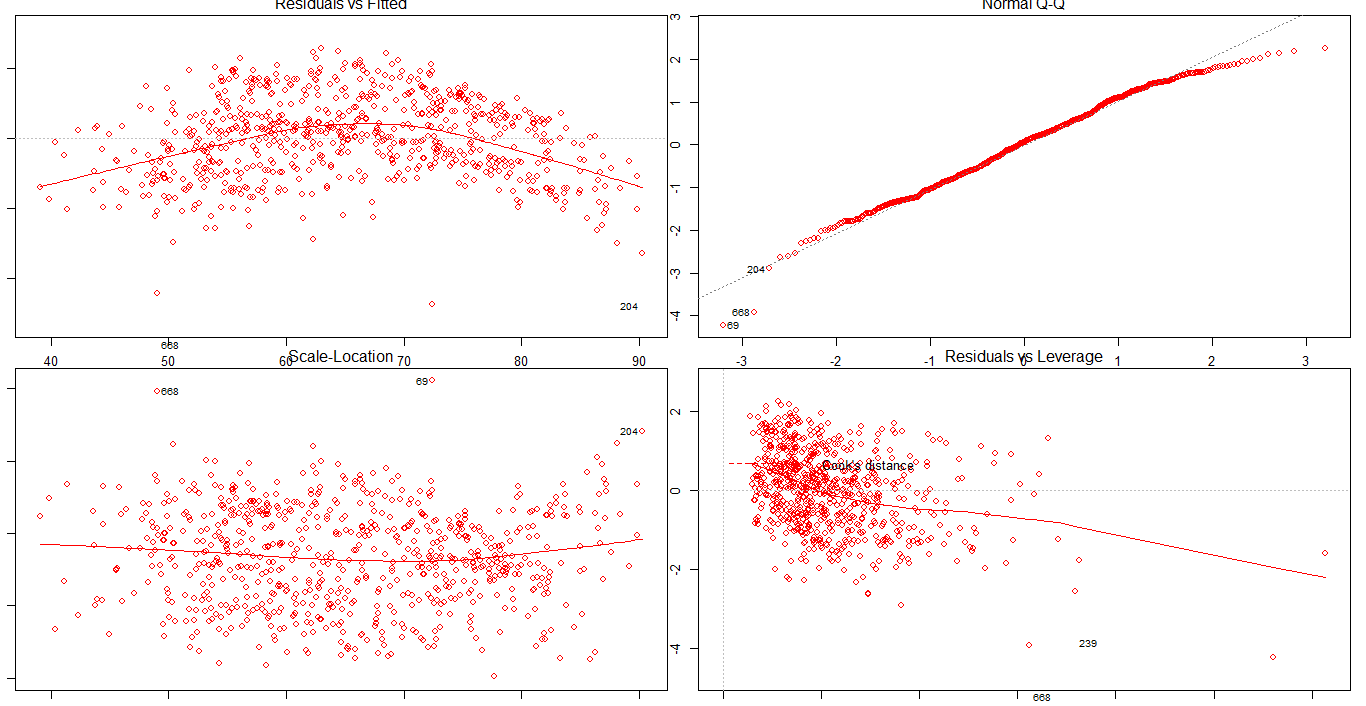
MODEL 2



Checked with square root of the target and actual\_temp and actual\_humidity as predictors:

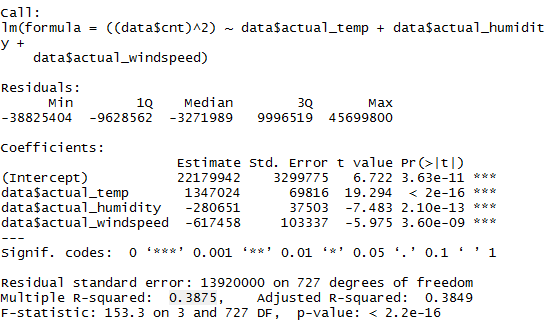
\*\*\* The model is significant at r squared 0.4781 and p-value < 2.2e-16

Also all p-value for three variables were significant.



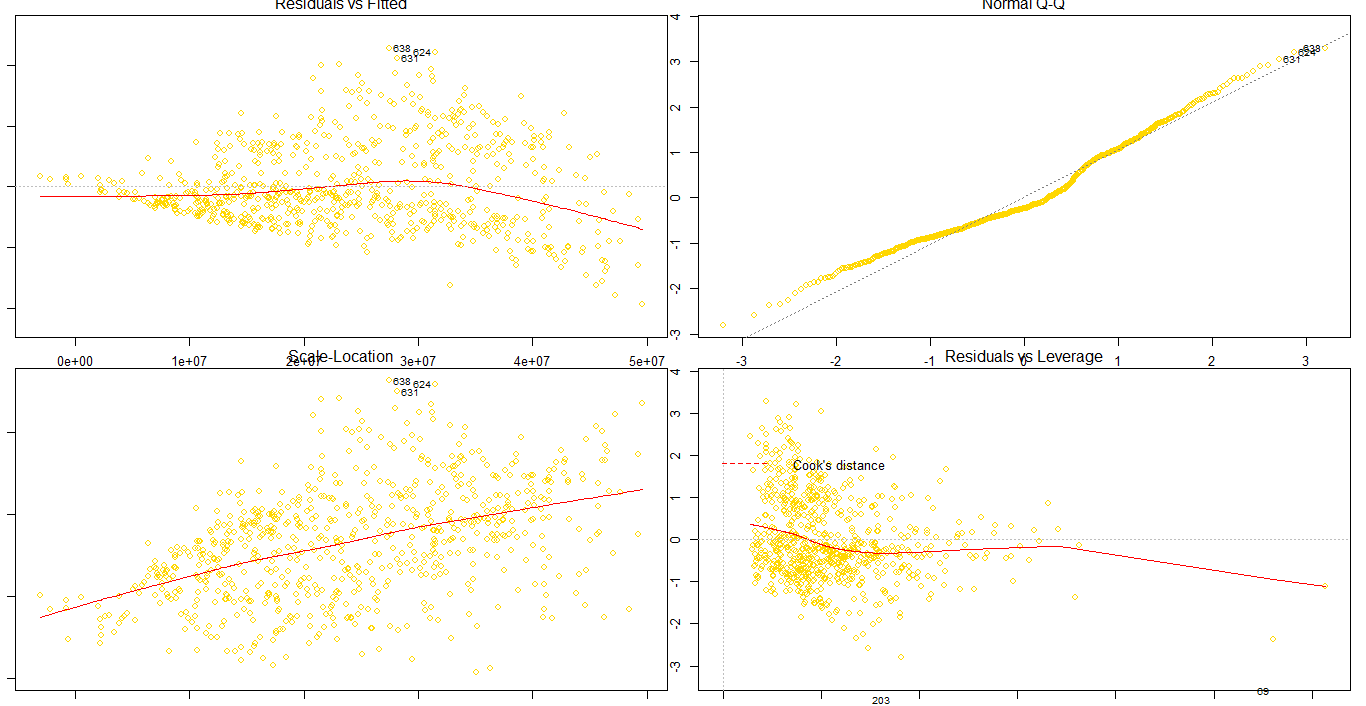
MODEL 3

Checked with square of the target and actual\_temp, windspeed and actual\_humidity as predictors:



\*\*\* The model is significant at r squared 0.3875 and p-value < 2.2e-16

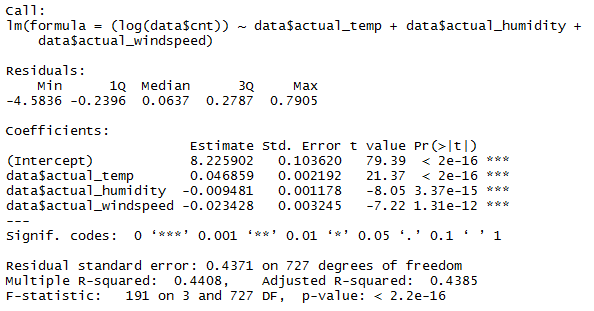
Also all p-value for three variables were significant.

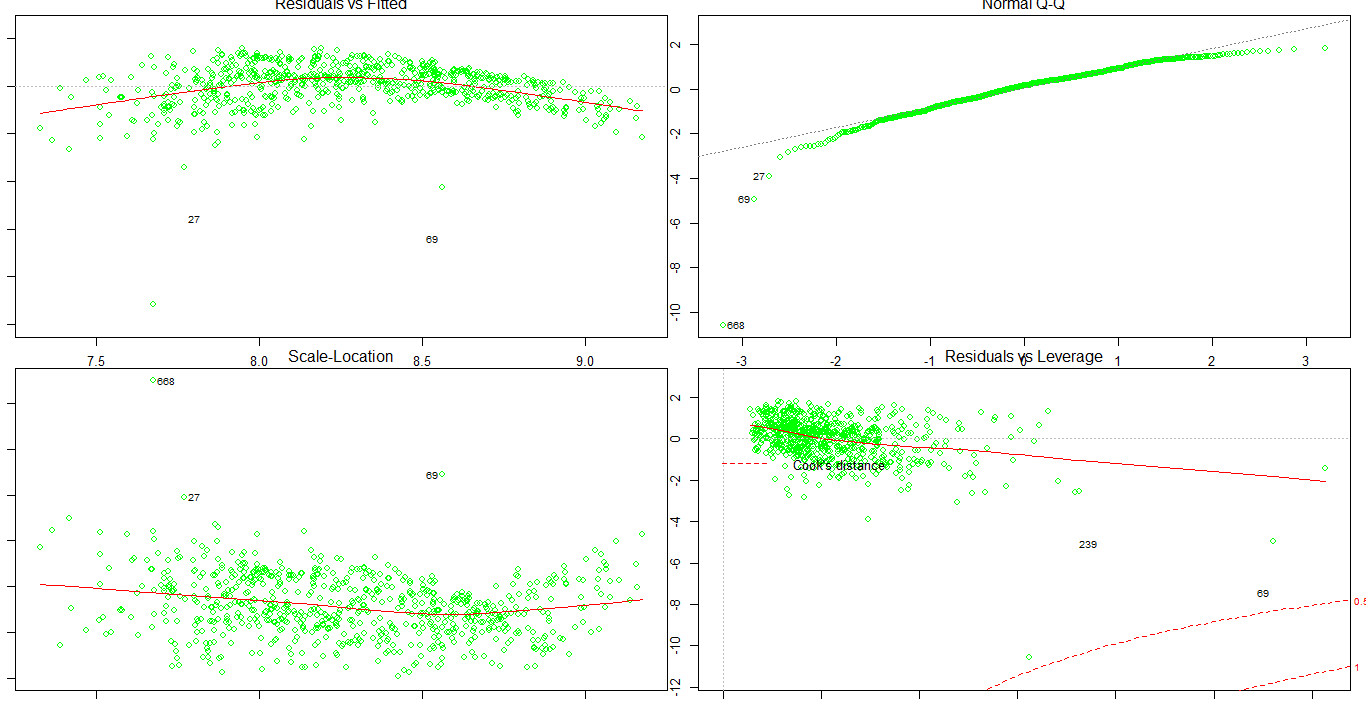


MODEL 4

Checked with OLS model with the target against actual\_temp, windspeed and actual\_humidity as predictors:

Also, all p-value for three variables were significant.

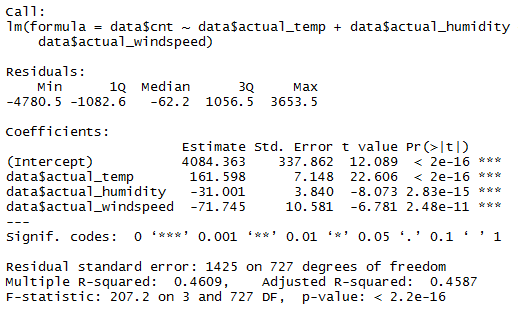


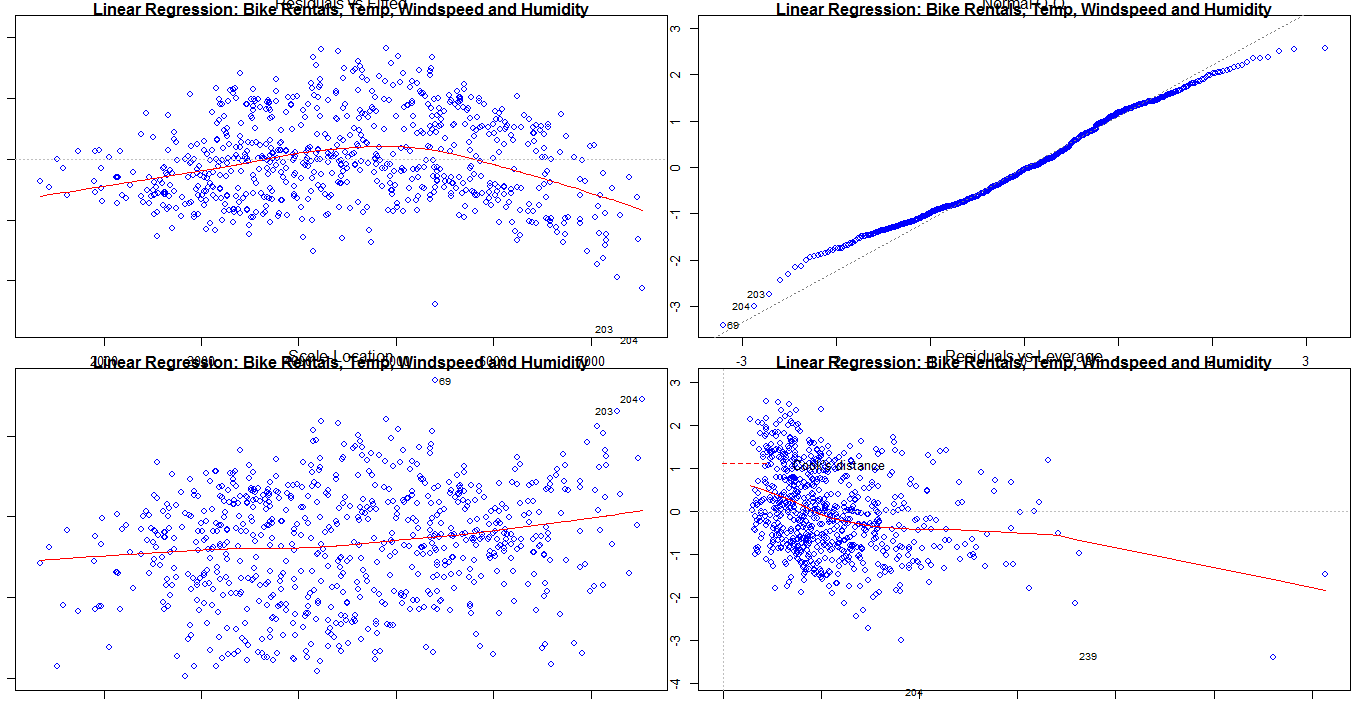


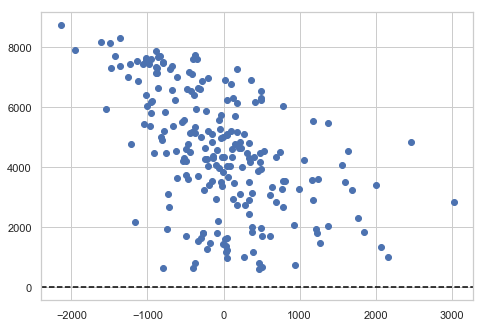
\*\*\* The model is significant at r squared 0.4371 and p-value < 2.2e-16

MODEL FINAL

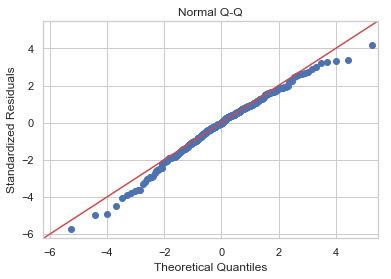
\*\*\* The final model is significant at r squared 0.4609 and p-value < 2.2e-16







**FIG : Residual vs Fitted Plot**



**CONCLUSION LINEAR REGRESSION:**

As we found the correlation plots against bike rentals with humidity and windspeed were slightly related, we created a linear model and found the R-Squared value at 46% and all p-value for three variables were significant.

Though, checking the residual plot and QQ plot, we can see that the residuals have a pattern, and are not normally distributed, which means the linear model doesn’t fit the data so well.

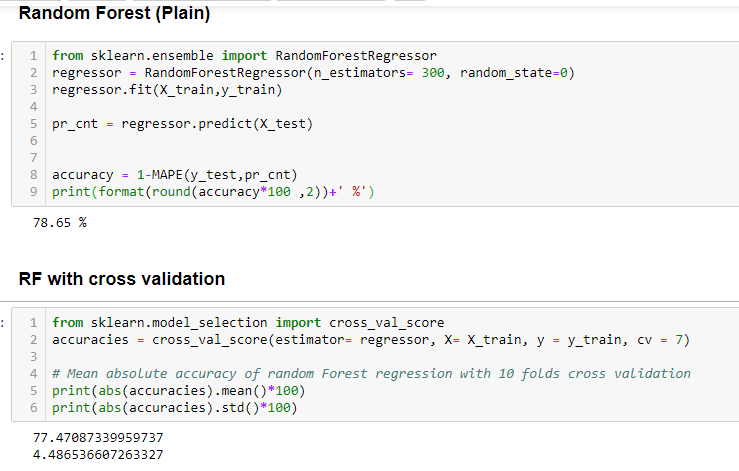
**RANDOM FOREST:**



Random Forest is built on the top of decision trees. Decision tree is the basic unit of Random Forest. Intuition is decision trees are weak learners. Random Forest is an ensemble technique, that learns on each iteration through multiple decision trees.

Why exactly is a random forest better than a single decision tree? We can think about it terms of having hundreds of humans make estimates for the max temperature problem: by pooling predictions, we can incorporate much more knowledge than from any one individual. Each individual brings their own background experience and information sources to the problem.

Why the name ‘random forest?’ Well, much as people might rely on different sources to make a prediction, each decision tree in the forest considers a random subset of features when forming questions and only has access to a random set of the training data points.



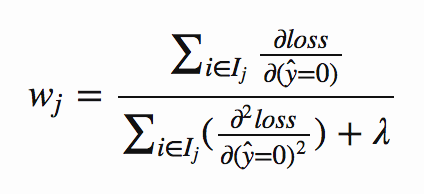
The accuracy recorded with Random Forest without CV is 78.65%

* On cross validating with 7 samples over Random Forest, Accuracy ~ 77.47% with standard deviation of 4.4

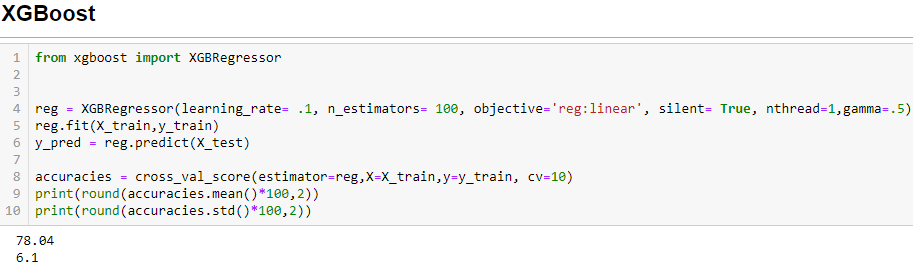
\*XGBoost

Finally, XGboost is fitted to the data with nrounds = 500.

It is an powerful, and lightning fast machine learning library. The idea behind the concept is to find the local minima or the global minima using partial derivatives.



Using the hyperparameters, the gradient boost readjusts the weights and minimizes the loss function.



We have fitted xgboost regression model with K-fold cross-validation with K=10

The accuracy measure went up to ~ 78.04% with the standard deviation of 6.1

**Conclusion**

As the linear model doesn’t meet the assumptions, we have rejected the model.

Now we will look into other models we have created.

We performed Random Forest Regression and the XGboost. Both of these models do not work on the Euclidean distances.

With XGBoost we have got an accuracy better than Random Forest, we will use the XGBoost model for the future data.

With this model implemented, we have got the important features as shown below

