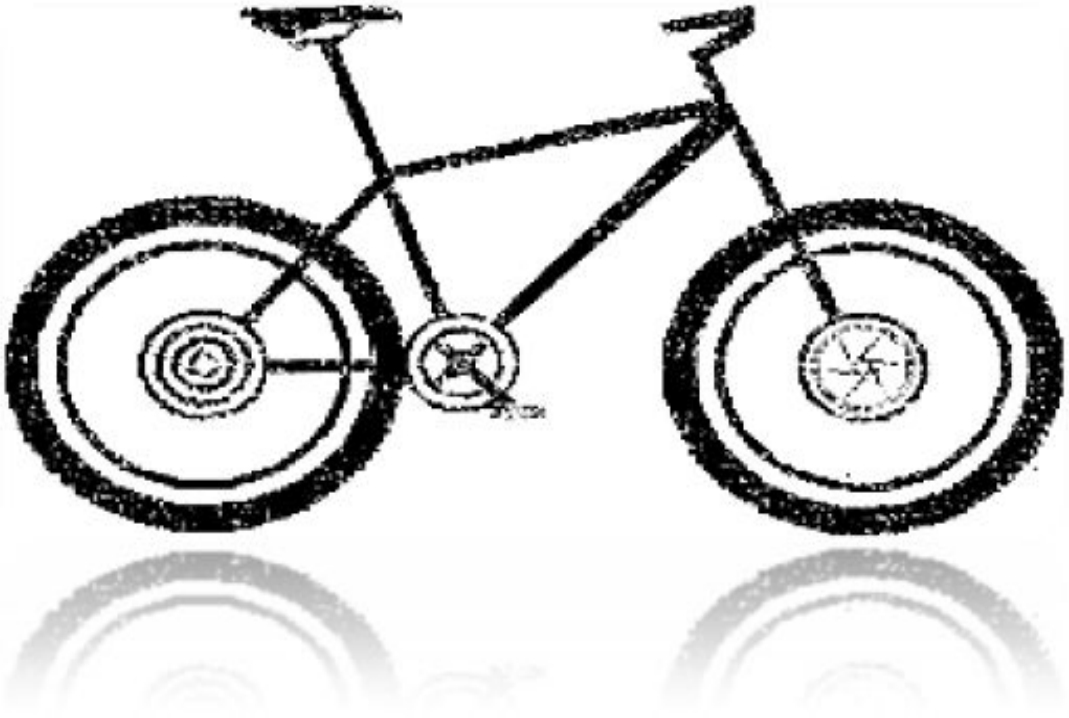
BIKE RENTING
DAILY COUNT PREDICTION



Prediction and Analysis Done in R and Python

ANTONY M THOPPIL

15th October 2018

Chapter-1: Introduction

Biking is an excellent way to stay in shape while exploring local areas and communing with nature. With many biking enthusiasts eager to find new paths to explore in and around their local area . Our case study is regarding an organization who lends rental bike. According to

business sense the demand of bikes for a particular day depends upon several factors like weather situation, season, holiday etc. It is important to know the demand of a particular day beforehand, so that they can meet the demand smoothly

Problem Statement

The objective of this case is to prediction of bike rental count on daily based on the environmental and seasonal settings.

Data

The details of data attributes in the dataset are as follows:-

The following are the dependent variables

Variables	:	Description
Instant	:	Record index
Dteday	:	Date (Ranging from 1st Jan 2011 to 31st Dec 2012)
season	:	Season (1 : Spring , 2 : summer, 3 : fall, 4 : winter)
yr	:	Year (0 : 2011, 1: 2012)
mnth	:	Month (1 to 12)
Hoiliday	:	Weather day is holiday or not (Extracted from holiday schedule) (0 : Not Holiday, 1: Holiday)
Weekday	:	Day of week
Workingday	:	If day is neither weekend or holiday : 1 otherwise : 0
Weathersit	:	Situation of weather (extracted from Freemeteo) 1: Clear,Few clouds, Partly cloudy, Partly cloudy 2: Mist + Cloudy, Mist + Broken Clouds, Mist + Few Clouds, Mist 3: Light Snow, Light Rain + Thunderstrom + Scattered Clouds 4 : Heavy Rain + ICE Pallets + Thunderstrom, 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
atemp	:	Normalized feeling temperature in Celsius. The values are derived via (t – t_min) / (t_max / t_min)) t_min = -16, t_max = +50
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

The independent variable is as follows:

cnt	:	Count of total rental bikes including both casual and registered Basciall it is (casual + registered)

According to the problem statement we have to predict bike rental count on a daily basis on the environmental and seasonal settings.

Chapter-2: EDA and Data Pre-processing

Hypothesis:

The following are the hypotheses made before doing analysis on the data:

1. Number of bikes rented would be more during spring, fall and winter season as it would be tough to ride the bicycle at higher temperatures
2. Total number of bikes rented would increase with year as more people would adopt environmental friendly activities, more people would be aware of the service and for reasons like lower cost
3. Registered users would be using the rented bike more during working days as they would choose it as their mode of transportation to the workplace whereas casual users would use it more during weekends and holidays as a leisure activity
4. No of bikes would be less during working days
5. Temperature would have a negative impact on the number of bikes rented
6. Humidity would have a negative impact on the number of bikes rented
7. Number of registered users would increase with year

Methodology of the project:

- Exploratory Data Analysis (Exploring data, Distribution of data, Visualization, Univarate, bivariate, Multivariate analysis)
- Preprocessing Data (Outliers in data, Dependencies among variables (Correlation // Anova // Chi-square // Multicollinearity), Sampling, dummies for categorical data in case of Statistical models)
- Basic Modeling & K-Fold Validation (Linear Regression, Decision Tree, Random Forest, SVR)

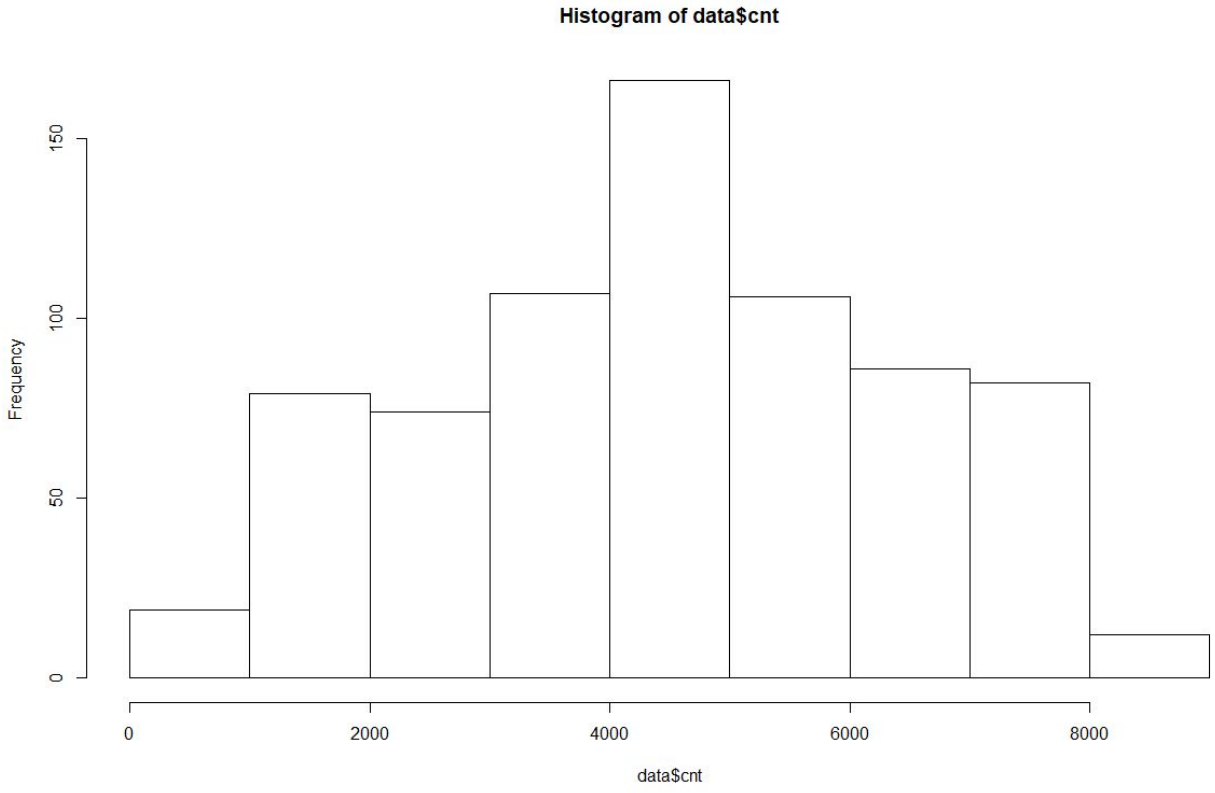
- Evaluation & Optimization of Final Model (Evaluating performances and tuning parameters for final model)

Exploratory Data Analysis:

Univariate Analysis :

Distribution of the Target variable

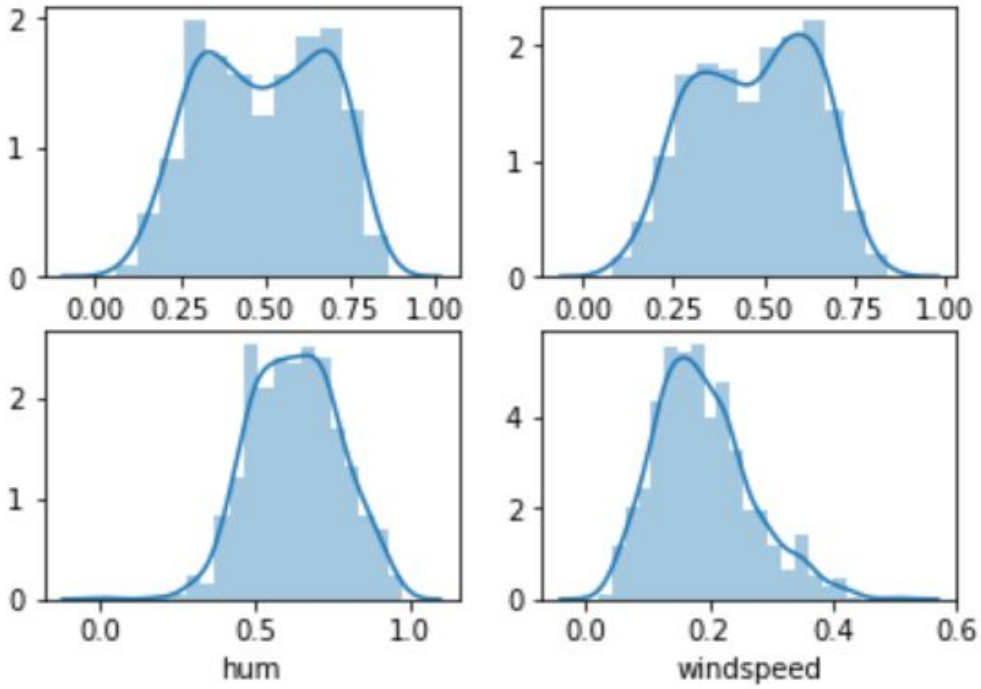
The Target variable is “cnt” , which is the total count of bikes rented during a given day. It is the arithmetic sum of registered users and casual users.



Distribution of other numerical variables:

The other numerical variables present in the dataset are “temp”(air temperature of the day), “atemp”(feel like temperature of the day), “hum”(humidity in the air during the day), “windspeed”(average windspeed during the day).

The distribution of the numerical variables are plotted with the help of histogram and is as shown below:



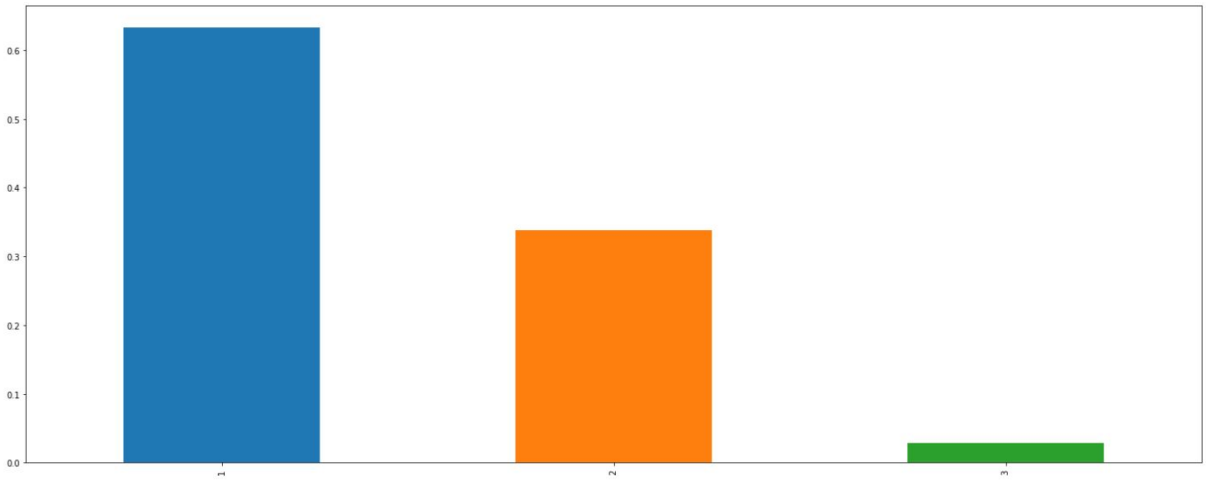
All the numerical variables are uniformly distributed and therefore scaling of variables needn't be done

Distribution of variable "weathersit"

Variable "weathersit" stands for weather situation of the day. The weather situations are broadly classified into four as :

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

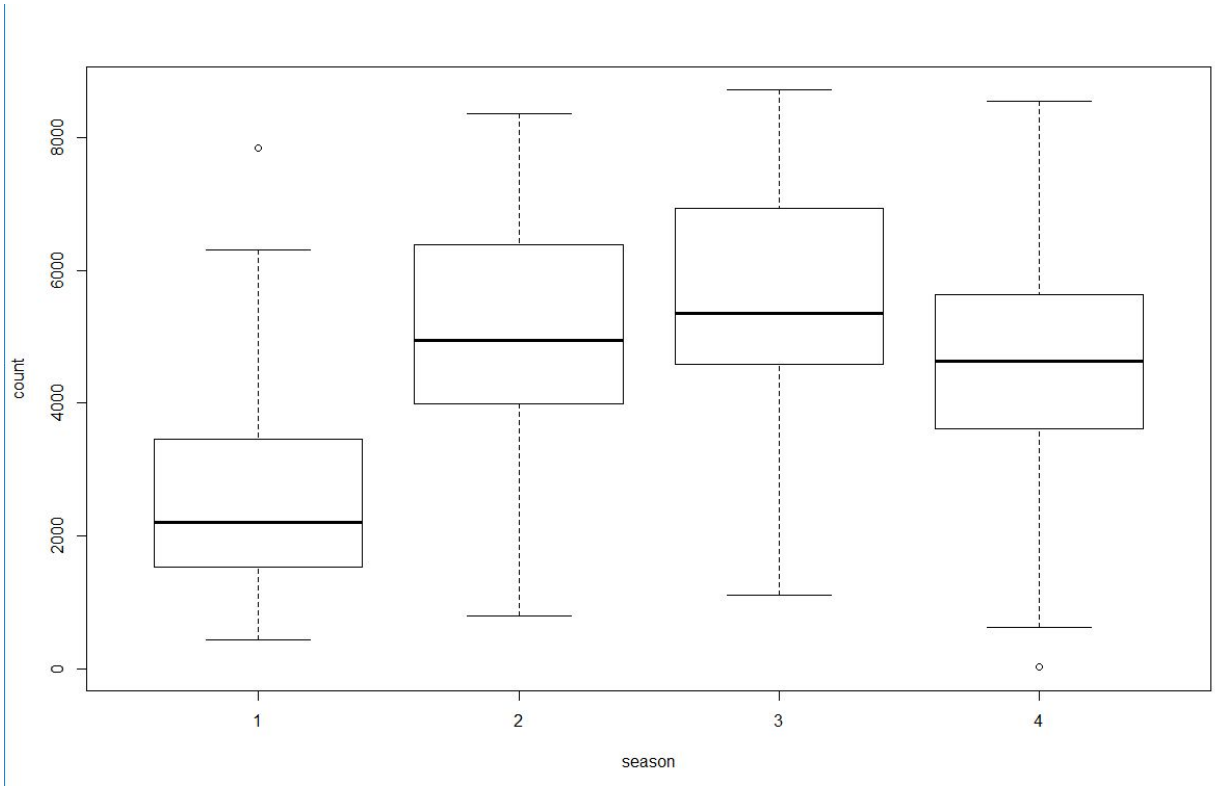
The distribution of different weather condition are as plotted below. The plot is a bar graph:



Multivariate Analysis:

Count of bikes rented with respect to different seasons:

We will plot a box plot to show the distribution of bikes rented in different seasons are plotted as shown below



From the plot we could conclude the following :

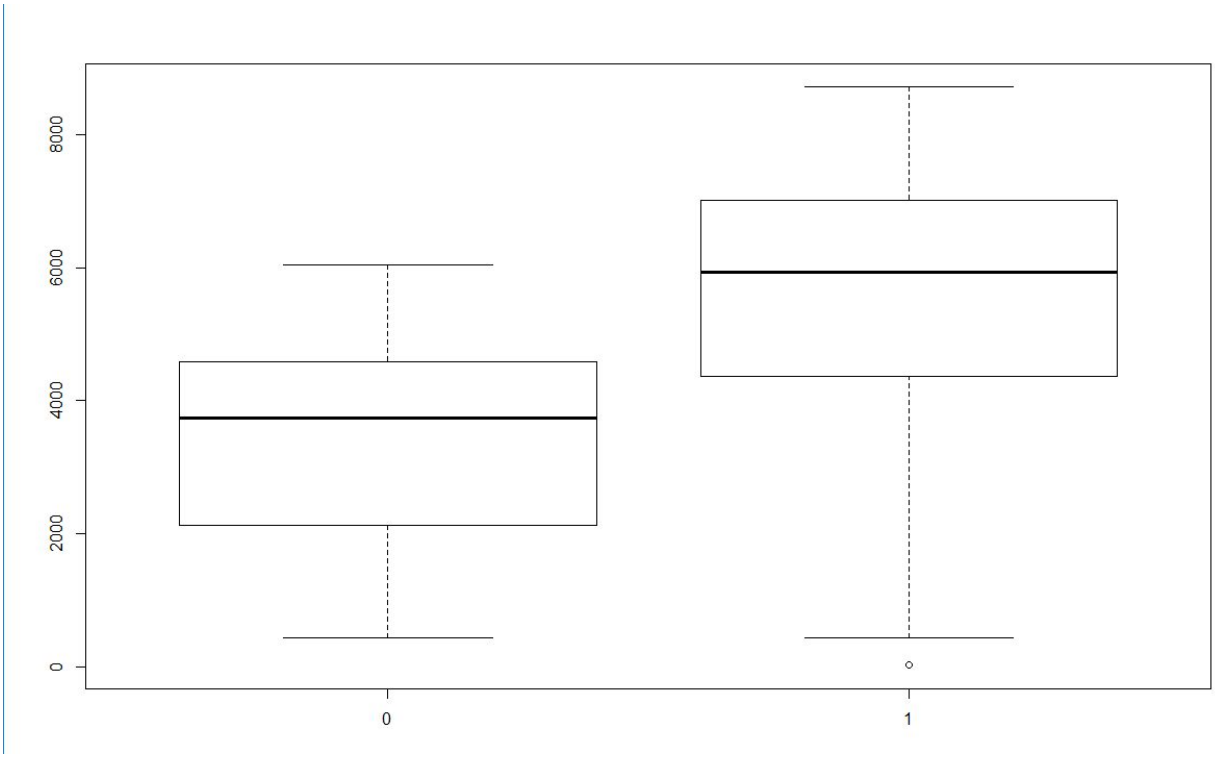
- Maximum number if bikes are rented during season 3(fall)
- Followed by season 2(summer)
- Followed by season 4(winter) and least during season1(spring)

This differs from our first hypotheses which states that the number of bikes rented would decrease with increase in temperature or less during summers. But the number of bikes rented are high during the summer season. One of the possible explanation for this observation is that the dataset available may be an arctic region, where the temperature is usually lower and thus people prefer to use bike during the summer seasons.

Count of bikes rented with respect to year.

Dataset only contains the number of bikes rented during two years. We had made an hypotheses earlier that the number of bikes rented would increase with year owing to the more adoption rate of eco friendly or more people came to know about the service and started to adopted it.

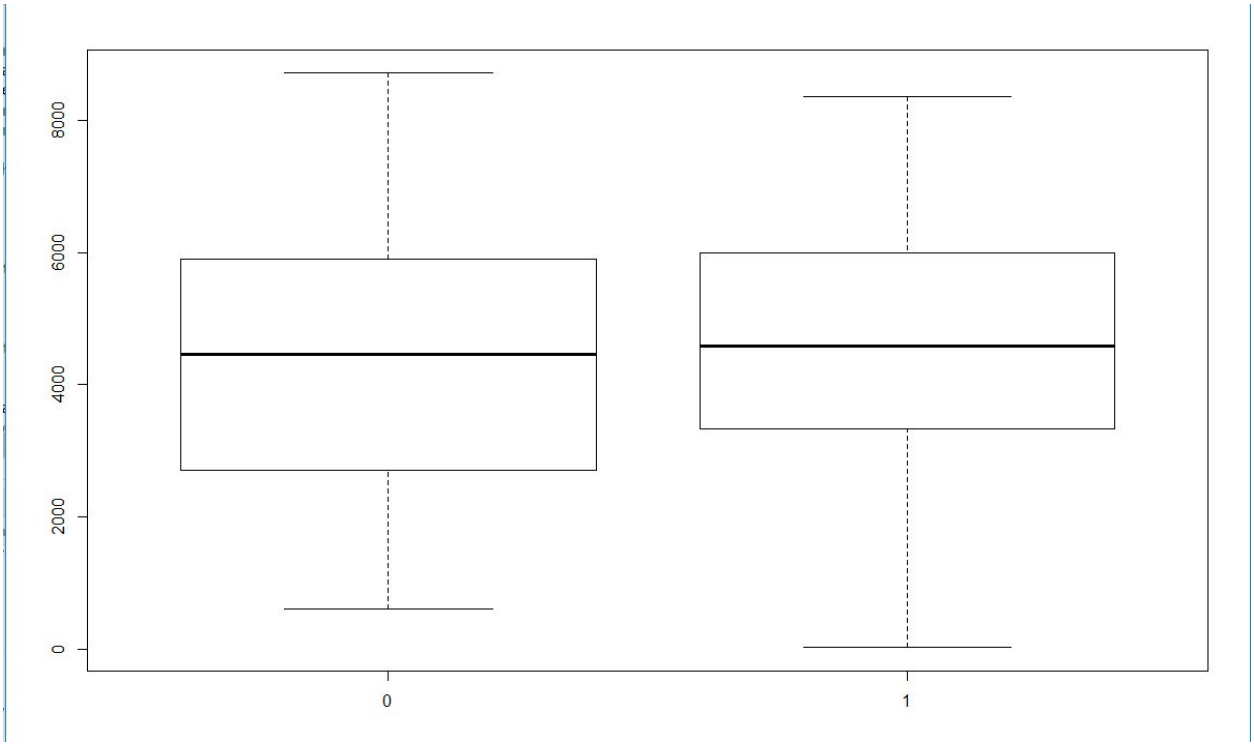
The boxplot of the number of bikes rented is plotted with respect to year below:



From the plot above we can conclude that the number of bikes rented is increasing with year and hence it proves our hypothesis.

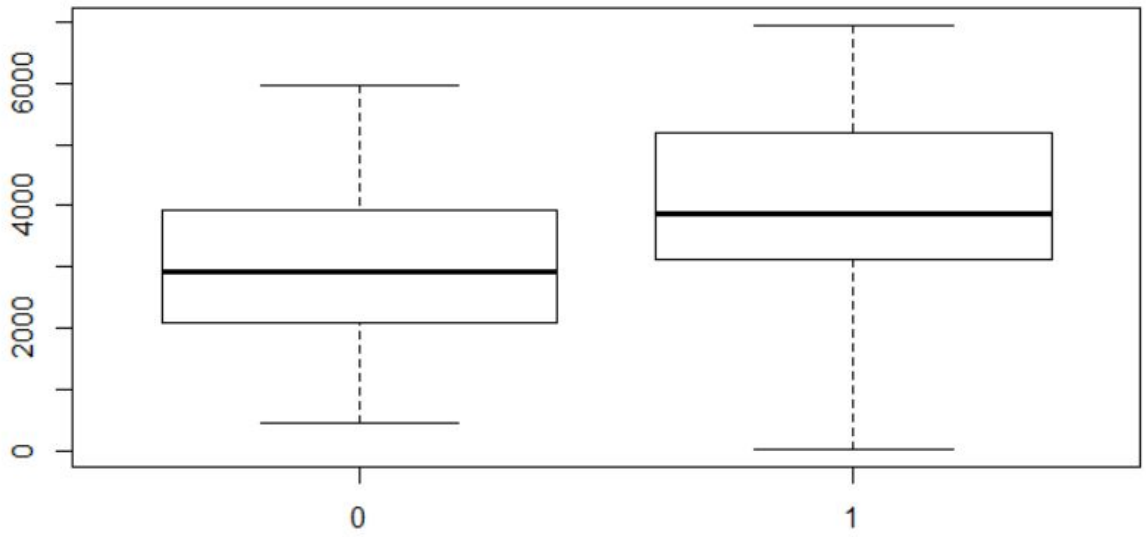
Distribution of bikes rented on working days and holidays

The number of bikes rented on working days and holidays are plotted with the help of a box plot as shown below:



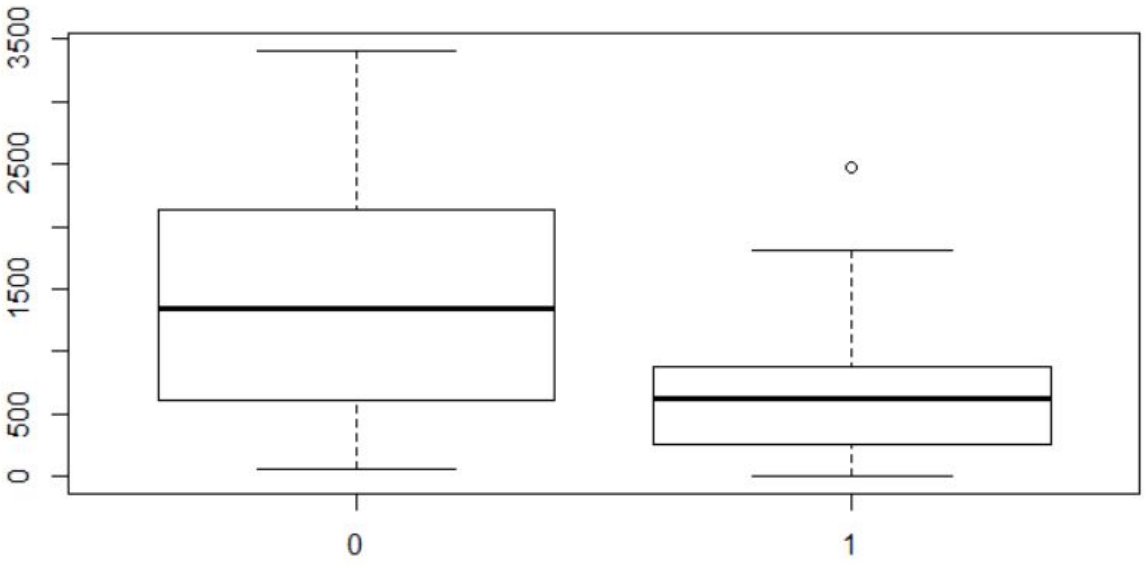
From the plot above we can see that the median for the number of bikes rented on holidays and number of bikes rented on working days are more less similar.

To find further insights into the data we plot another box plot of the bikes rented by the registered users during working days. It is also one of our hypotheses that the number of registered users using the bikes on the working day would be more as more registered users would use the bike for commute to work place in workingdays and casual users would use the bikes more on holidays:



Form the above box plot we can conclude that more bikes are rented on holidays by registered users as compared to working days which contradicts our hypotheses.

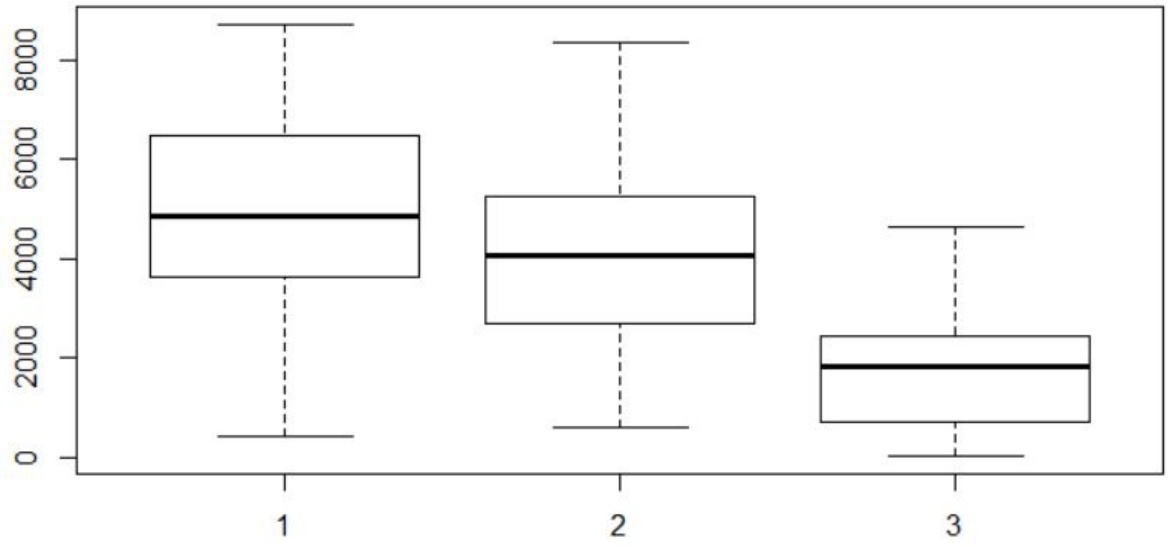
Box plot to show the distribution of bikes rented by casual users on working days and holidays.



From the above box plot we can conclude that casual users rent the bike more during working days compared to holidays.

Distribution of bikes rented with weather situation.

Box plot to show the distribution of bikes rented with respect to weather situation is plotted below:

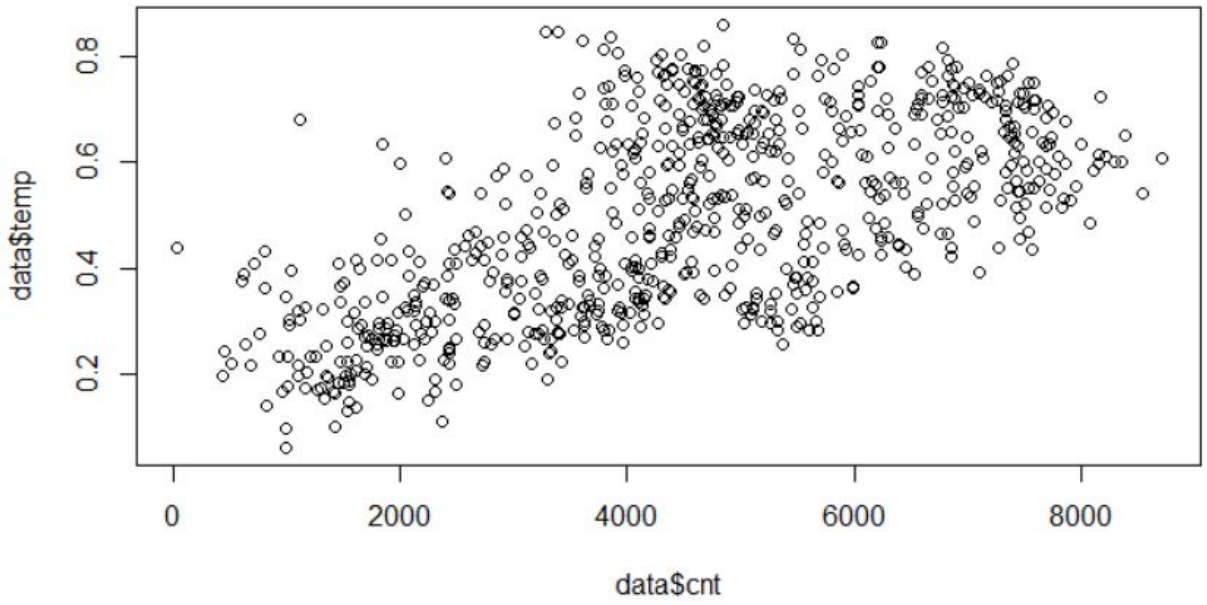


From the boxplot above we can conclude the following:

- More bikes are rented during weather situation 1 and 2 and followed by 3 and none during 4
- None during 4 proves our hypotheses that rain effects the number of bikes rented.

Relationship between temperature and bikes rented.

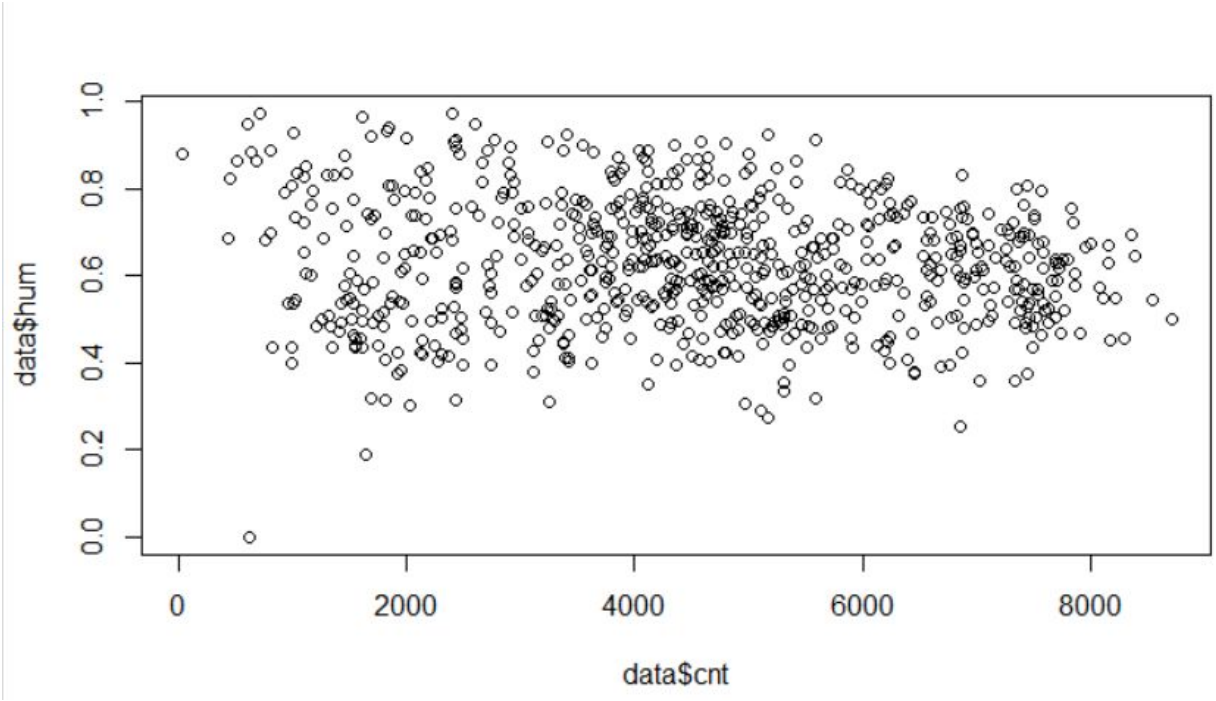
Both variables temperature and bikes rented are numerical variables and therefore we would use scatter plot to see the relationship between the two. The scatter plot is as shown below:



From the plot above we could see that there is a positive relationship between temperature and the number of bikes rented. This would be due to our earlier explanation that the dataset being used here is from a arctic country and therefore people prefer to bike during hotter temperatures.

Relationship between humidity and bikes rented.

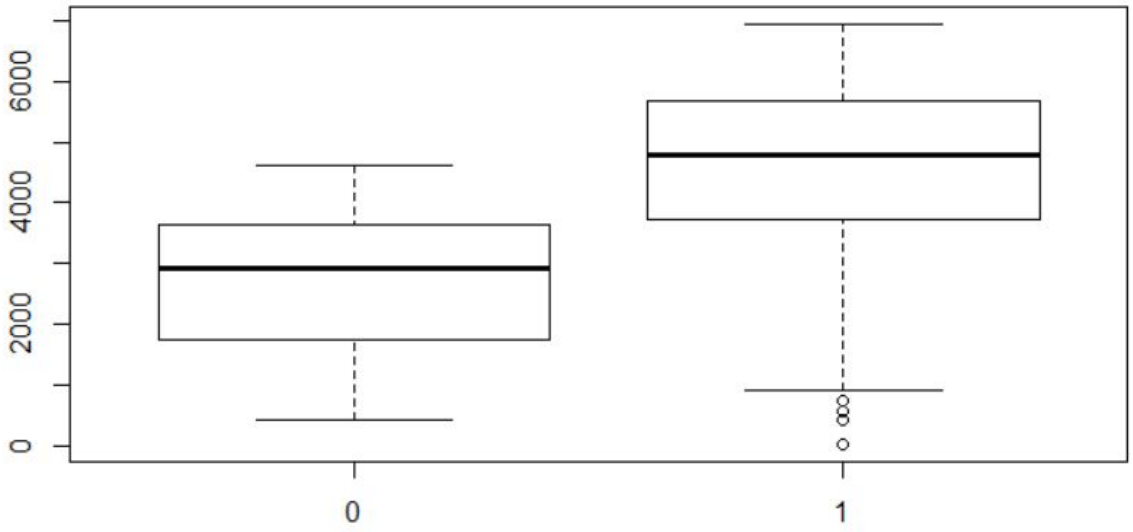
As the one above both humidity and bikes rented are numerical variables and therefore scatter plot is used to observe the relationship between the two. The scatter plot is shown below:



From the plot above we can observe that there is no relationship between humidity and the number of bikes rented.

Number of registered users with year.

In one of our hypotheses we had stated that the number of registered users should increase with year. We would use a boxplot to plot the relationship between registered users and year as shown below:

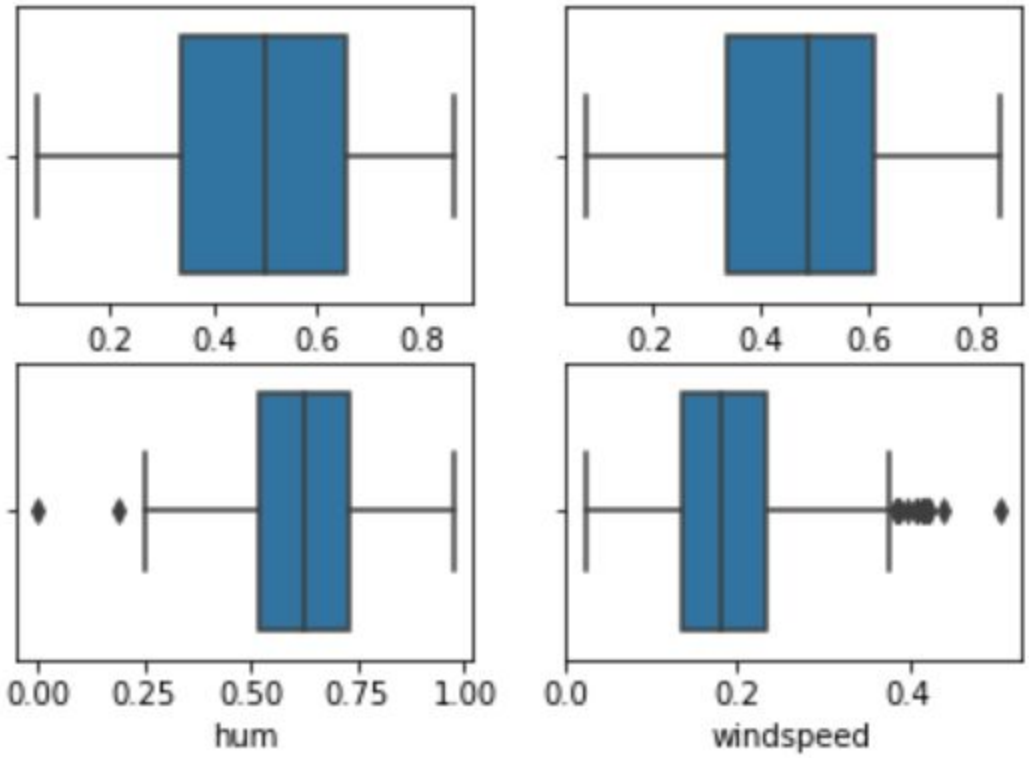


From the boxplot above we can conclude that the number of registered users have increased with year.

Missing Value and Outlier Analysis

There are no missing values in the dataset.

Outlier analysis is done to variables ‘temp’, ’atemp’, ‘hum’ and ‘windspeed’. The boxplot of the four numeric variables is as shown below:



From the plot above we could see that there are outliers in humidity and windspeed.
We need to check if the outliers are natural outliers or error outliers. In case of natural outliers we need to standardize the variables using log transformation or by caping the corresponding variables and in case of error outliers we need to remove the values and impute by using mean, median, mode or knn imputation.
The values of humidity which are outliers are as follows: 0.19 and 0.00. The value of humidity cannot be 0 and hence it is an error outlier. Therefore we would be deleting the outlier values and imputing using knn imputation.

In case of windspeed to check the type of outlier we check the weather situation when the outliers occurs. If the weather situation is raining or during thunderstorm we could say that windspeed could be higher than normal and hence the outliers are not error outliers. But the outlier windspeed doesnt occur during either of the weather situation and hence it needs to be treated as an error outliers and thus should be deleted and imputed by knn imputaiton

Knn imputation : KNN is an algorithm that is useful for matching a point with its closest k neighbors in a multi-dimensional space. It can be used for data that are continuous, discrete, ordinal and categorical which makes it particularly useful for dealing with all kind of missing data.

Hence the missing values and outliers are treated and the dataset is ready for feature engineering which is the last step of EDA and data pre-processing.

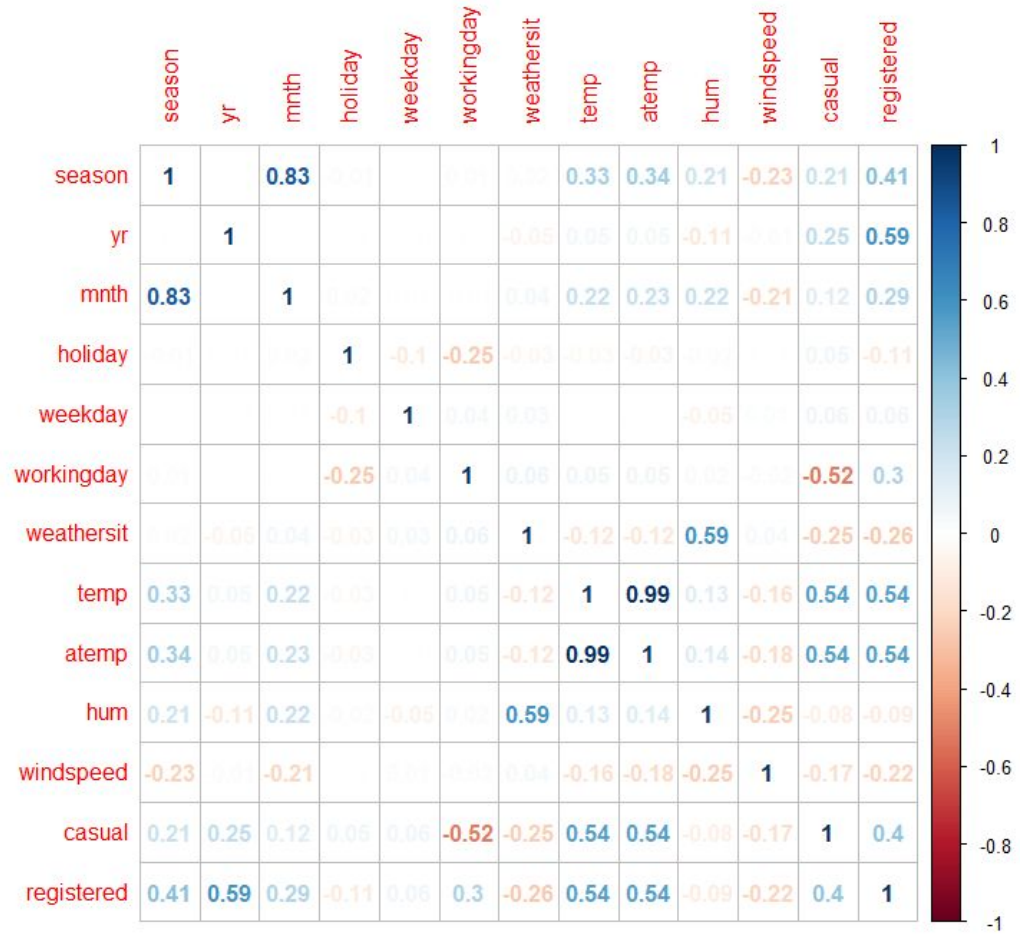
Feature Engineering

Feature engineering is that part of data pre-processing in which new variables are created and one which are not useful deleted from the dataset.

The variables ‘instant’ and ‘dteday’ are removed as they are unique for each observation and thus it needs to be removed.

We will plot the correlation plot of all the variables in the dataset to see the correlation between various variables. Variables that are highly correlated needs to be removed as it may lead to overfitting of the trained model to a particular class of variables.

The correlation plot of the given dataset is as follows:



From the above plot we could see that:

- Season and month variables are correlated and thus one of the two variables needs to be removed. We would remove month
- Temp and atemp the two variables are highly correlated and infact it represents the same data. So one of the variables needs to be removed. We will remove atemp

Hence the data is finally ready for model training and implementation.

Chapter-3: Model Training and Implementation

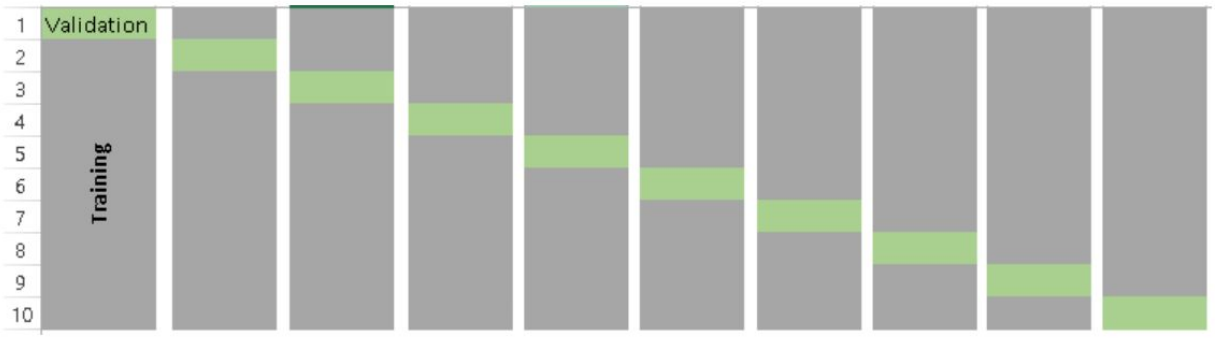
We would train the model on training set and test the data on validation set which is a smaller section of the original train data available with us.
There are many ways of cross validation, which are:

- The Validation set Approach
- Leave out one cross validation (LOOCV)
- k-fold cross validation
- Stratified k-fold cross validation
- Adversarial validation
- Cross validation for time series
- Custom cross validation techniques

We would be using k-fold cross validaiton.

K-fold Cross validation: Methods undertaken for k-fold cross validation are as follows :

1. Randomly split your entire dataset into k"folds"
2. For each k-fold in your dataset, build your model on k – 1 folds of the dataset. Then, test the model to check the effectiveness for *kth* fold
3. Record the error you see on each of the predictions
4. Repeat this until each of the k-folds has served as the test set
5. The average of your k recorded errors is called the cross-validation error and will serve as your performance metric for the model



The Metric used for comparison between moodels are RMSE

RMSE: Root Mean Square Error
is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit

$$\mathrm{RMSE}_{fo} = \left[\sum_{i=1}^{N} (z_{f_i} - z_{o_i})^2 / N\right]^{1/2}$$

Lower the value of RMSE the better the model

Models being implemented :

Linear Regression:

Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Every value of the independent variable x is associated with a value of the dependent variable y. In other terms, MLR examines how multiple independent variables are related to one dependent variable. Once each of the independent factors have been determined to predict the dependent variable, the information on the multiple variables can be used to create an accurate prediction on the level of effect they have on the outcome variable. The model creates a relationship in the form of a straight line (linear) that best approximates all the individual data points.

The model for multiple linear regressions is:

y = B0 + B1x1 + B2x2 + ... + Bnxn + E

Y = Target Variable
B0 = Intercept
B1 = regression coefficient that measures a unit change in the dependent variable when x1 changes – change in y
B2 = coefficient value that measures a unit change in the dependent variable when x2 changes – change in y
x1, x2, …, xn = Predictors
E = random error in prediction, that is variance that cannot be accurately predicted by the model. Also known as residuals.

The model is implemented and the error metric RMSE is : 882.07

Decision Tree

Decision tree belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms decision trees can also be used for solving regression and classification problem. The general motive of decision tree is to create a training model which can be used to predict class or value of target variables by learning decision rules inferred from training data.

Basically Decision tree is a rule based approach and it uses tree like structured. Each internal node of the tree corresponds to an attribute, and each leaf node corresponds to a class label.

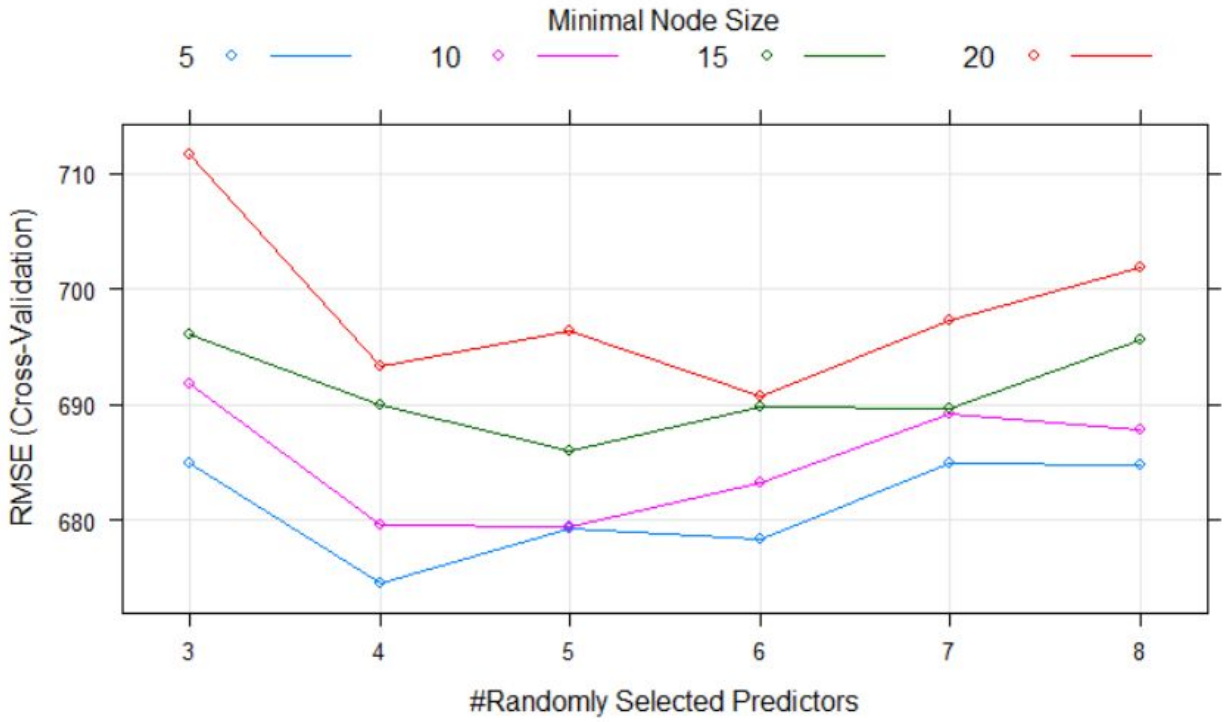
The model was implemented
The error metrics RMSE is : 1155.98

Random Forest

A group of decisions Trees is random forest. The Random forest model is a type of additive model that makes prediction by combining decisions from a sequence of base models.

In case of regression while predicting the output we go for mean of all the favorable rule case values.

The model was implemented and plotted as shown below



From the plot above we could see that mtry=4 and minimal node size of 5 is the best parameter for random forest model

The model is implemented
And the error metrics RMSE is : 682.00

Conclusion:

The RMSE score which is the error metrics in consideration in this problem statement, we could see that the RMSE score is least in case of random forest. Therefore it is selected as our final model

Final model : Random Forest.