# Prediction of Bike Rental Count

*Akash Ingole*

*13 March 2020*

## Contents

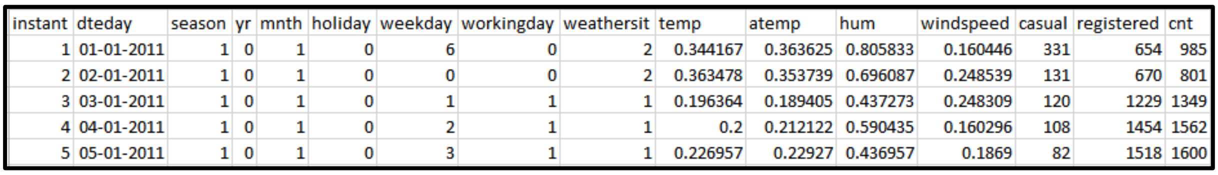
# Chapter 1 – Introduction

## 1.1 Problem Statement

The aim of this project is to predict the count of bike rentals based on the seasonal and environmental settings. By predicting the count, it would be possible to help accommodate in managing the number of bikes required on a daily basis, and being prepared for high demand of bikes during peak periods. This project also tries to answer the best algorithm that can work efficiently for this real world problem of bike rental prediction.

## 1.2 Data

The goal is to build regression models which will predict the number of bikes used based on the environmental and seasonal behaviour. Given below is a sample of the dataset that we are using to predict the number of bikes:



| instant | dteday | season | yr | mnth | holiday | weekday | workingday | weathersit | temp | atemp | hum | windspeed | casual | registered | cnt |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 01-01-2011 | 1 | 0 | 1 | 0 | 6 | 0 | 2 | 0.344167 | 0.363625 | 0.805833 | 0.160446 | 331 | 654 | 985 |
| 2 | 02-01-2011 | 1 | 0 | 1 | 0 | 0 | 0 | 2 | 0.363478 | 0.353739 | 0.696087 | 0.248539 | 131 | 670 | 801 |
| 3 | 03-01-2011 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0.196364 | 0.189405 | 0.437273 | 0.248309 | 120 | 1229 | 1349 |
| 4 | 04-01-2011 | 1 | 0 | 1 | 0 | 2 | 1 | 1 | 0.2 | 0.212122 | 0.590435 | 0.160296 | 108 | 1454 | 1562 |
| 5 | 05-01-2011 | 1 | 0 | 1 | 0 | 3 | 1 | 1 | 0.226957 | 0.22927 | 0.436957 | 0.1869 | 82 | 1518 | 1600 |

Table 1: Bike Prediction Sample Data

As you can see in the table below, we have following 13 variables, using which we have to correctly predict the count of bikes:

| Sr.No | Variable |
|---|---|
| 1 | Instant |
| 2 | dteday |
| 3 | Season |
| 4 | Yr |
| 5 | Month |
| 6 | Holiday |
| 7 | Weekday |
| 8 | Workingday |
| 9 | Weathersit |
| 10 | Temp |
| 11 | Atemp |
| 12 | Hum |
| 13 | Windspeed |

Table 2: Predictor variables

# Chapter 2 – Methodology

## 2.1 Pre-processing

A predictive model requires that we look at the data before we start to create a model. However, in data mining, looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is known as Exploratory Data Analysis.

## 2.2 Distribution of Continuous Variables

It can be observed from the below histograms, temperature and feel temperature are normally distributed, whereas, the variables, wind speed and humidity are slightly skewed. This skewness is likely of the presence of outliers and extreme data in those variables.

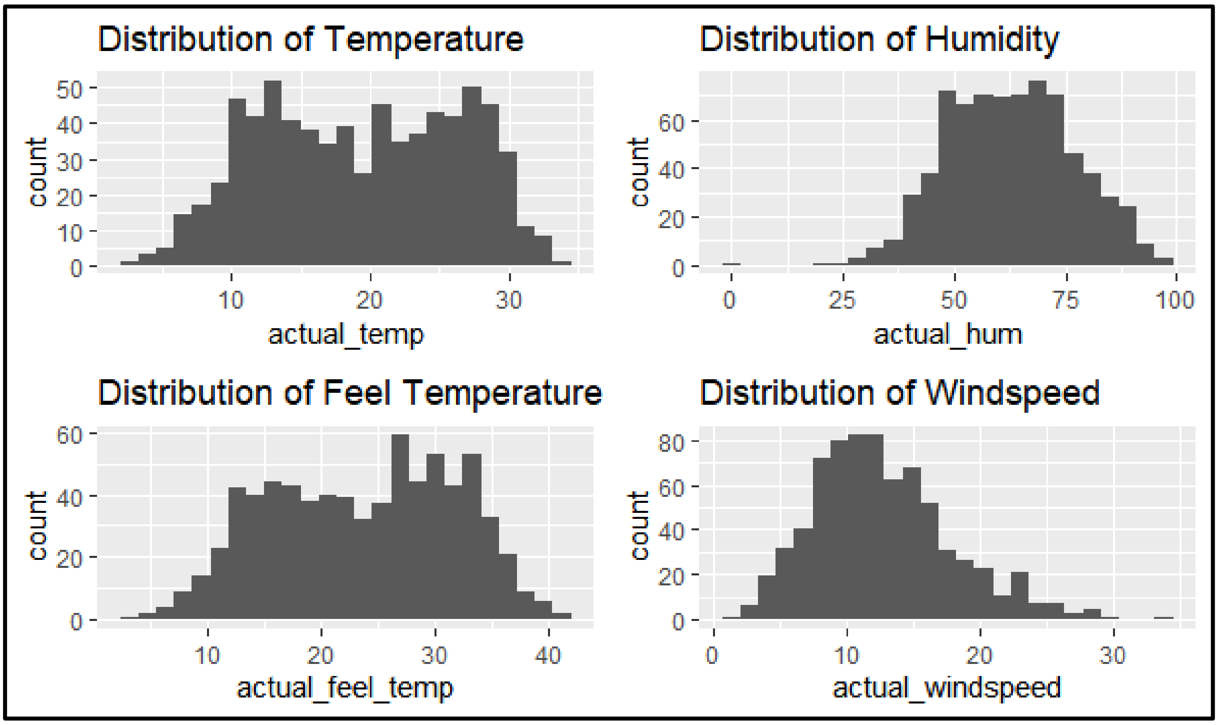


Figure 1: Distribution of continuous variables using histograms

## 2.3 Distribution of Categorical Variables

The distribution of categorical variables is shown in the below figure:

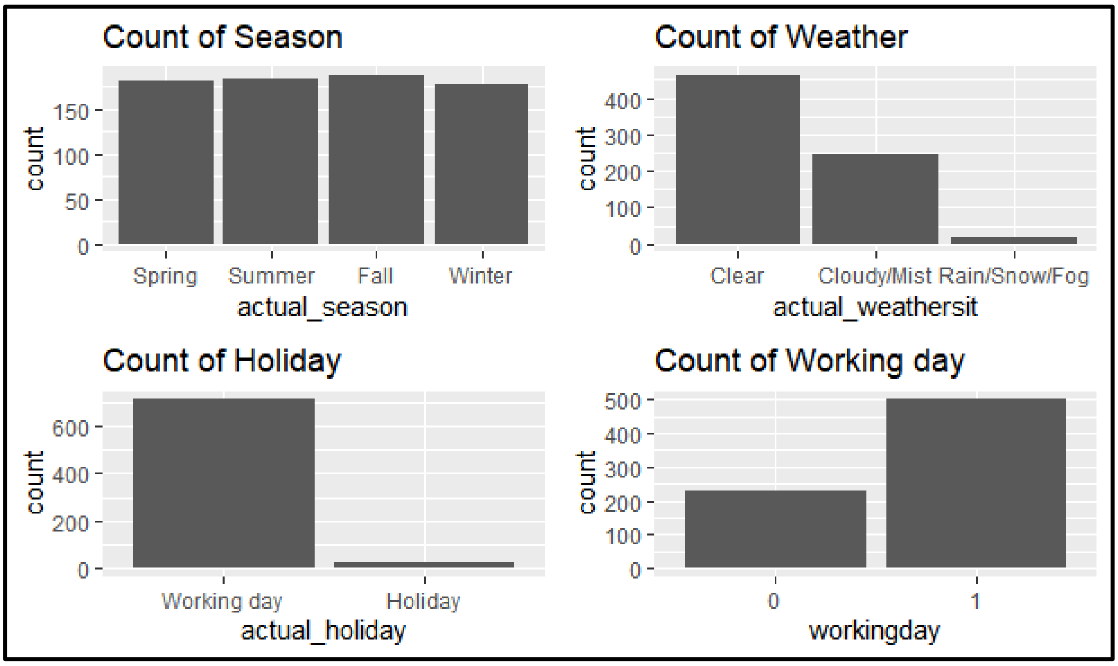


Figure 2: Distribution of categorical variables using bar plots

## 2.4 Relationship of continuous variables against bike count

The below figure shows the relationship between continuous variables and the target variable using scatter plot. It can be observed that there exists a linear positive relationship between the variables temperature and feel temperature with the bike rental count. There also exists a negative linear relationship between the variables humidity and wind speed with the bike rental count.

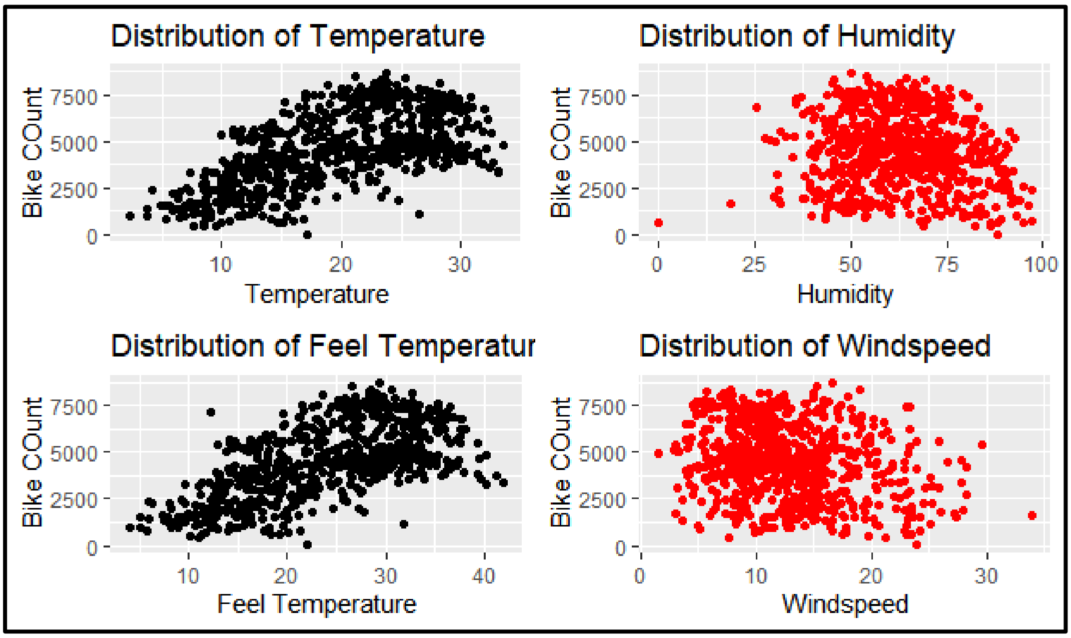


Figure 3: Scatter plots for continuous variables

## 2.5 Detection of Outliers

Outliers can be visually detected using box-whisker plot, or simply boxplot. Below figure illustrated the boxplot for all the continuous variables.

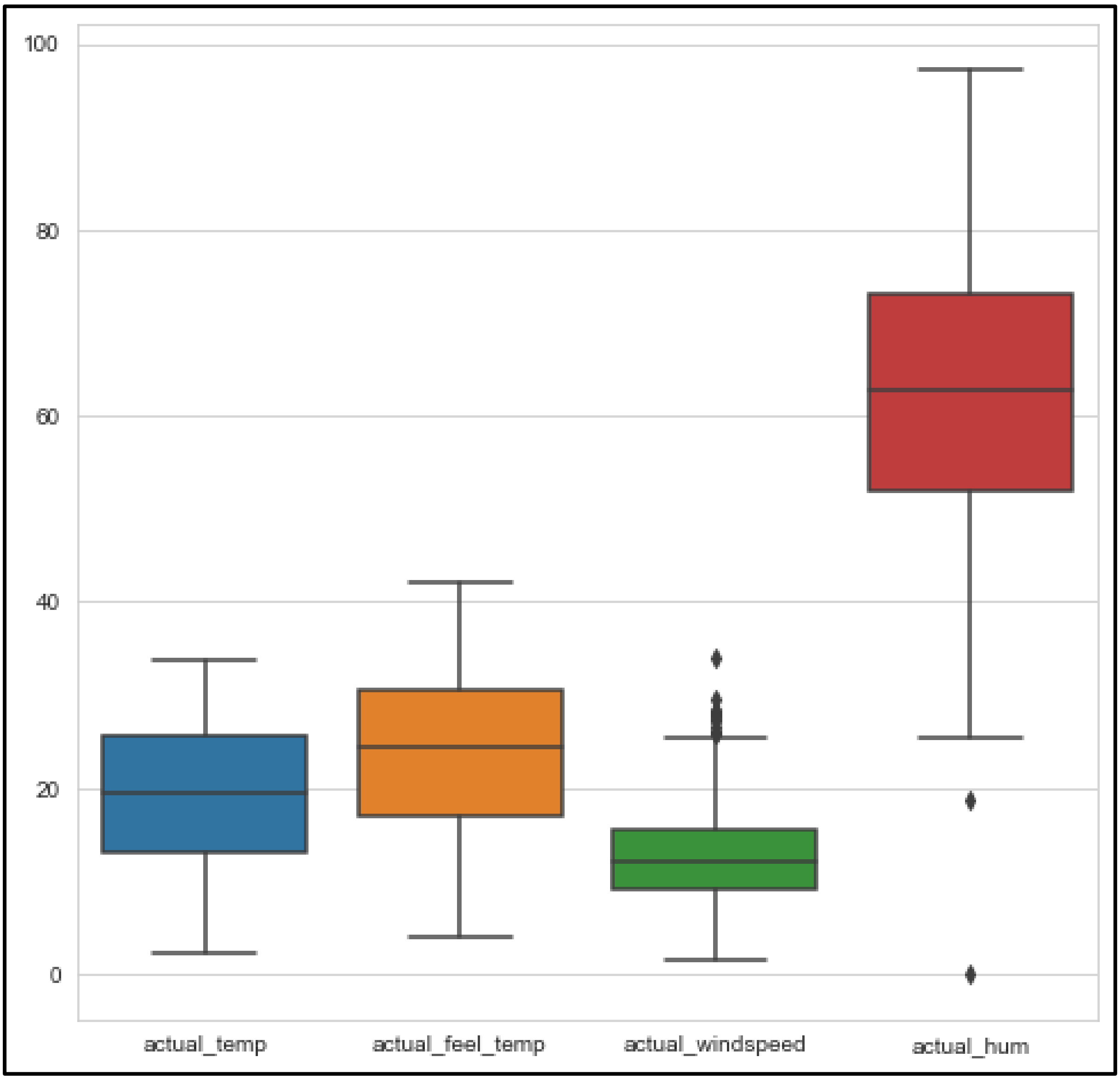


Figure 4: Boxplot for continuous variables

Outliers can be removed using boxplot stats method, wherein the interquartile range (IQR) is calculated and the minimum and maximum values for the variables are calculated. Any value ranging outside the minimum and maximum value is discarded.

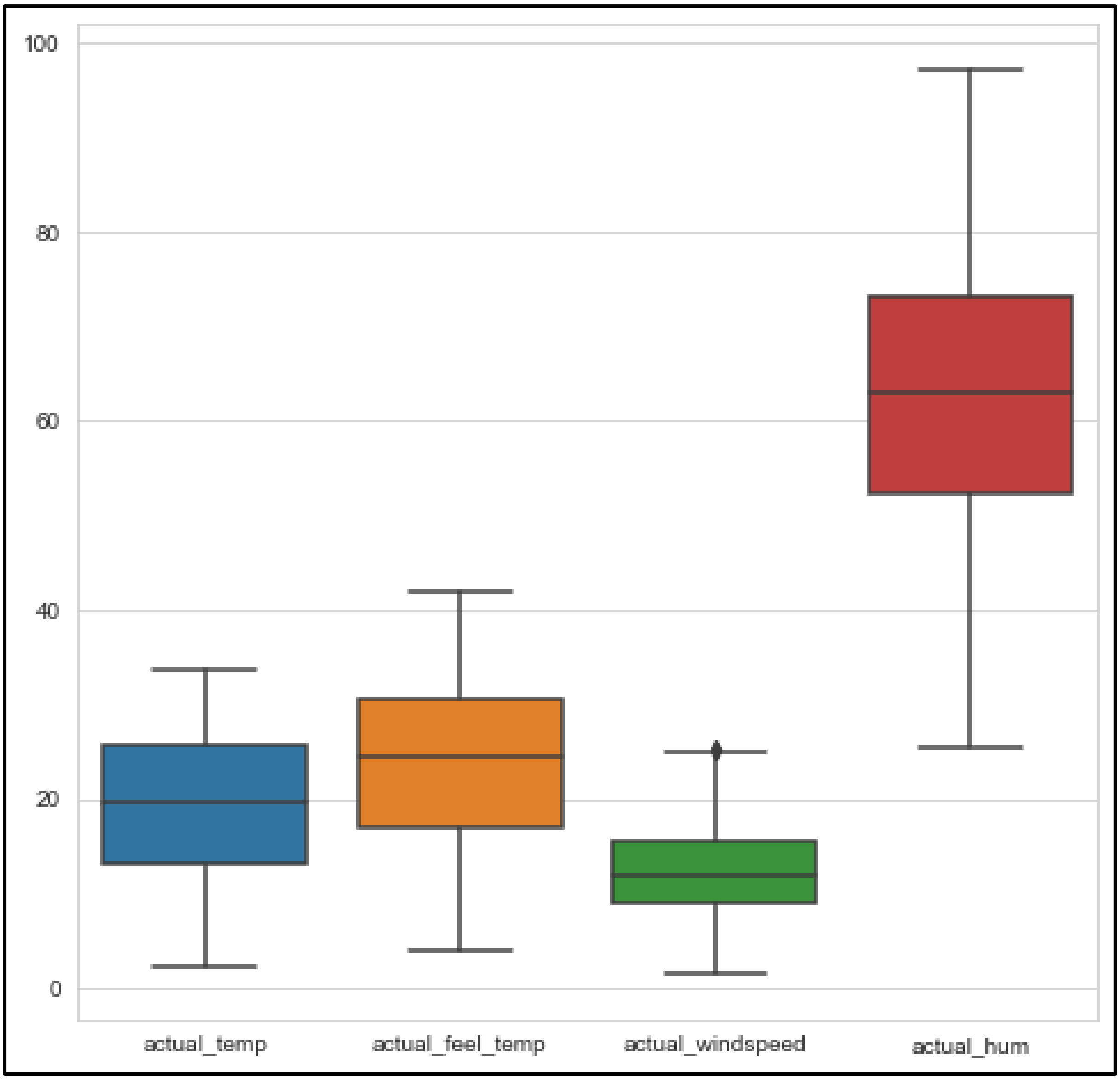


Figure 5: Boxplot of continuous variables after removing outliers

It can be observed from the distribution of wind speed and humidity after removal of outliers, the data is not skewed as much as before the removal of outliers.

## 2.6 Feature Selection

Feature selection reduces the complexity of a model and makes it easier to implement. It also reduces overfitting. Features are selected based on their scores in various statistical tests for their correlation with the outcome variable.

Correlation plot is used to find out if there is any multicollinearity between variables. The highly collinear variables are dropped and then the model is executed.

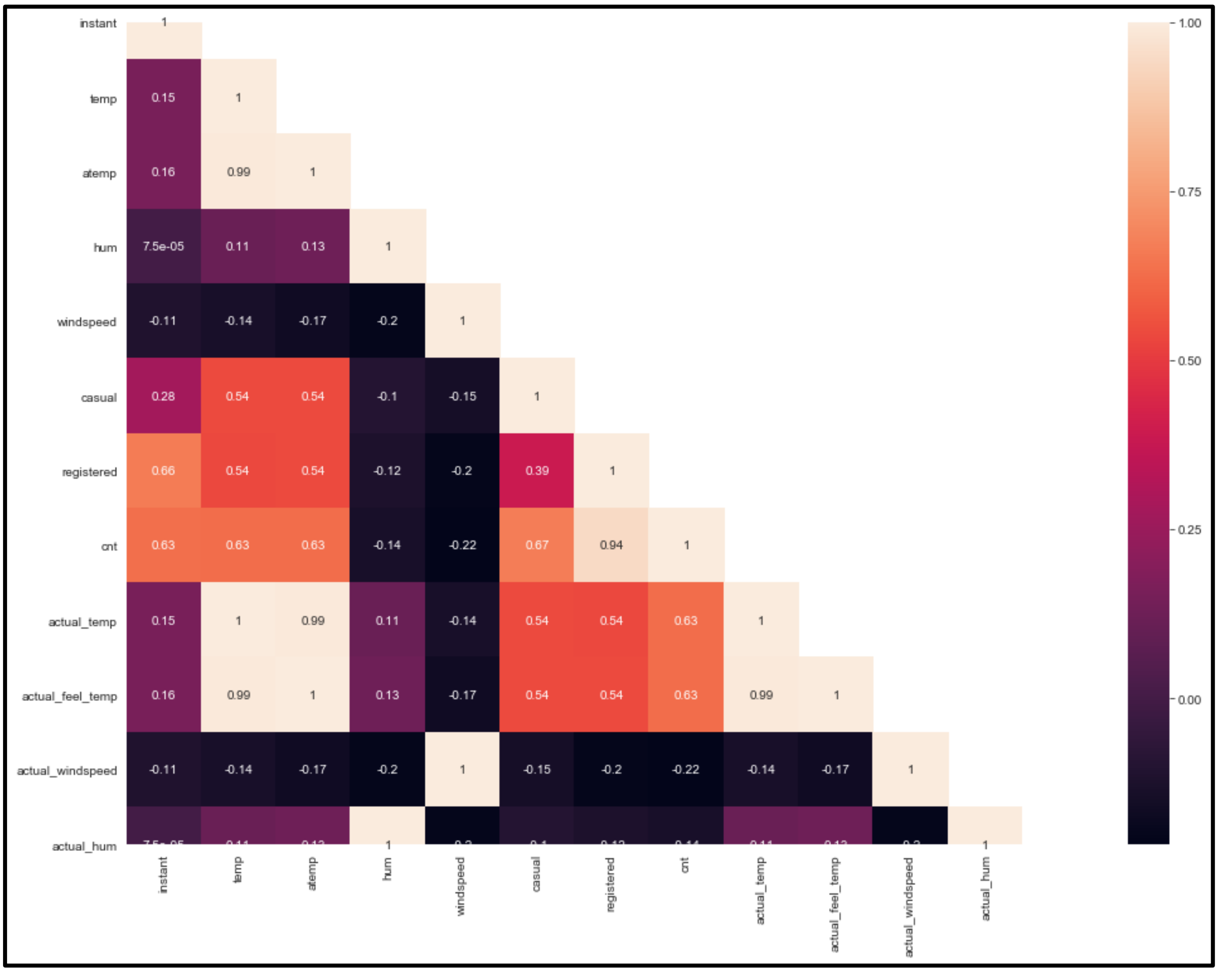


Figure 6: Correlation plot of all variables

# Chapter 3 – Modelling

## 3.1 Model Selection

The dependent variable in our data is a continuous variable, count of bike rentals. Hence the models we choose are linear regression, decision tree and random forest. The error metric chosen for the problem statement is Mean Absolute Error (MAE).

## 3.2 Decision Tree

A decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it used a tree-like model of decisions. Decision trees are computationally cheap to use, easy to understand and prone to overfitting.

Using decision tree algorithm, we can predict the value of bike count. The MAPE for this decision tree model is 18.4%. Hence the accuracy of this model is 81.60%.

## 3.3 Random Forest

Random forest or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees.

The number of trees used for the prediction using random forest is 500. The MAPE for this model is 13.10%. Hence the accuracy of this model is 86.90%

## 3.4 Linear Regression

Linear regression performs the task to predict a dependent variable based on the given set of independent values. So, the regression techniques find out the linear relationship between the two.

The MAPE for this model is 19.08%. Thus the accuracy for this model is 81.92%.

# Chapter 4 – Conclusion

Now that we have a few models for predicting the target variable, we need to decide which one to choose. There are several criteria that exist for evaluating and comparing models. We can compare the models using any of the following criteria:

1. Predictive Performance
2. Interpretability
3. Computational Efficiency

In our case of Bike count prediction Data, Interpretability and Computation Efficiency, do not hold much significance. Therefore, we will use Predictive performance as the criteria to compare and evaluate models.

Predictive performance can be measured by comparing Predictions of the models with real values of the target variables, and calculating some average error measure.

## 4.1 Mean Absolute Percentage Error (MAPE)

The mean absolute percentage error, also known as mean absolute percentage deviation, is the measure of prediction accuracy of a forecasting method in statistics, for example in trend estimation, also used in loss function for regression problems in machine learning.

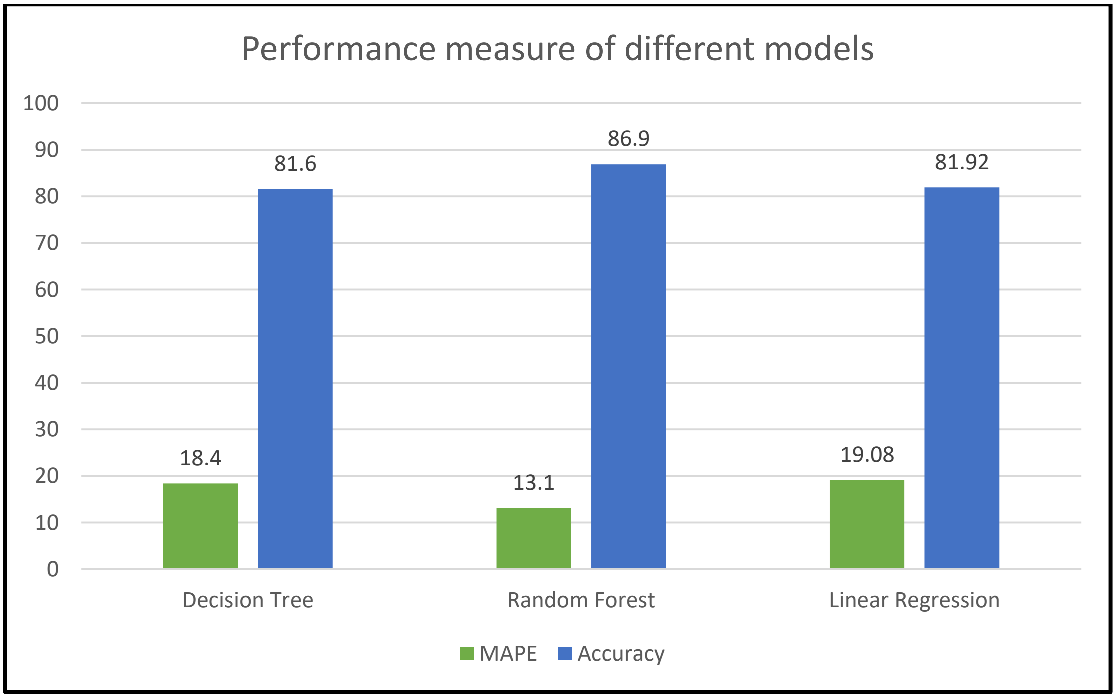


Figure 7: Performance comparison of different models

# Chapter 5 – Model Selection

As we can see from the above illustration, Random Forest performs better than the other two models. So we can select this algorithm for our prediction.