**Bike Rental Prediction**

*Allwyn Joseph*

*06-09-2019*

Table of Contents

[**Introduction** 3](#_Toc18711539)

[**Problem Statement** 3](#_Toc18711540)

[**Data** 3](#_Toc18711541)

[**Methodology** 4](#_Toc18711542)

[**Pre-Processing** 4](#_Toc18711543)

[Exploring the target variable 4](#_Toc18711544)

[Exploring the numerical variables 5](#_Toc18711545)

[Exploring relationships between data 6](#_Toc18711546)

[Dealing with datatypes 9](#_Toc18711547)

[Feature Engineering 10](#_Toc18711548)

[Missing Value Analysis 11](#_Toc18711549)

[Outlier Analysis 11](#_Toc18711550)

[Feature Selection 13](#_Toc18711551)

[**Conclusion** 14](#_Toc18711552)

[**Model Selection** 14](#_Toc18711553)

[Multiple Linear Regression 14](#_Toc18711554)

[Decision Tree Regression 17](#_Toc18711555)

[Random Forest Regression 18](#_Toc18711556)

[**Model Evaluation** 19](#_Toc18711557)

[**Model Selection** 21](#_Toc18711558)

Chapter 1

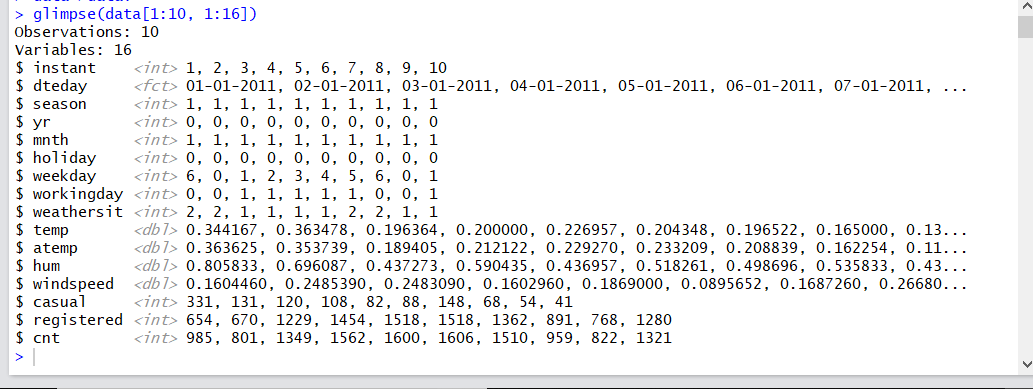
# **Introduction**

## **Problem Statement**

For a bike rental venture, it is important to understand and estimate the count of bicycles being rented. The objective of this Case is to Predication of bike rental count on daily based on the environmental and seasonal settings.

## **Data**

Given below is a sample of the data set that we are using to predict the count of users. The data contains 731 observations and 16 variables.



Chapter 2

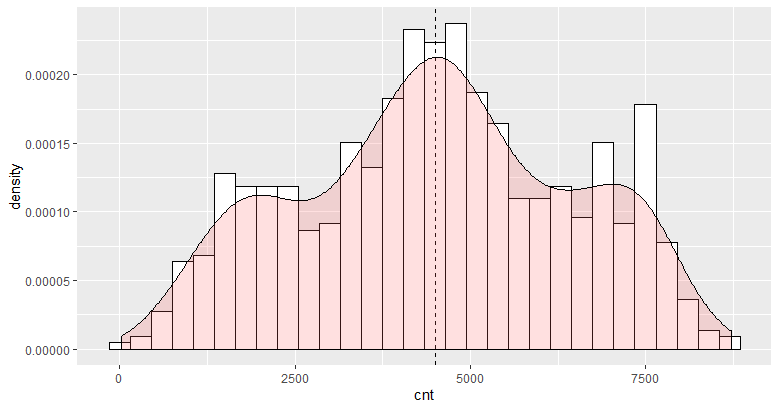
# **Methodology**

## **Pre-Processing**

Before starting to build a model to predict the target variable, it is necessary to understand what kind of data we are dealing with. The relationships between the independent variables and with the dependent variable are analysed. The datatypes of the necessary variables which need to be converted are done accordingly. Pre-processing of data involves steps like missing value analysis, outlier analysis and normalisation.

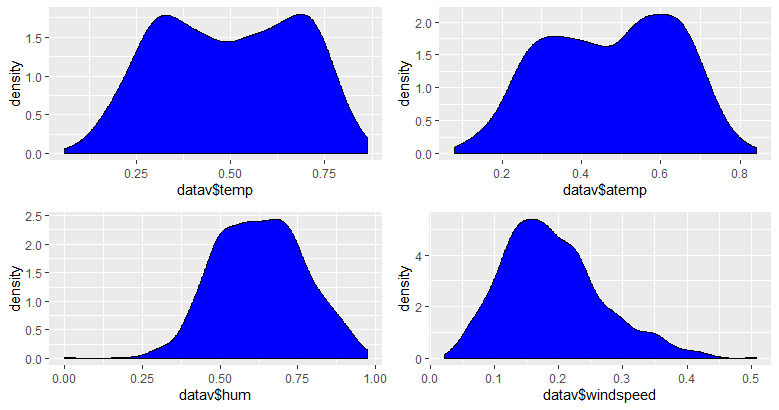
### Exploring the target variable

‘cnt’ is the target variable. According to the histogram with density plot below, we see that the variable ‘cnt’ is normally distributed.



### Exploring the numerical variables

‘hum’, ‘temp’, ‘atemp’ and ‘windspeed’ are the numerical features available in this dataset. By looking at the density plots of numerical variables, we can see that most of the variables are normally distributed.

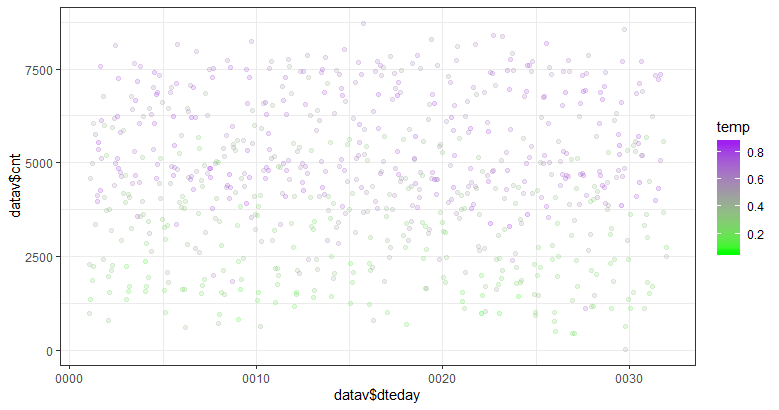


### Exploring relationships between data

Let’s see how the variables perform against ‘cnt’ variable

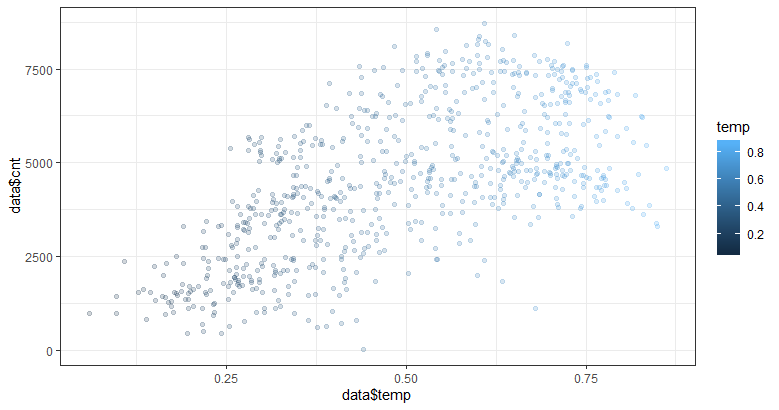
1. Cnt vs dteday

The below graph shows that the count of bike rentals has increased over time. Thus, indicating a linear relationship.



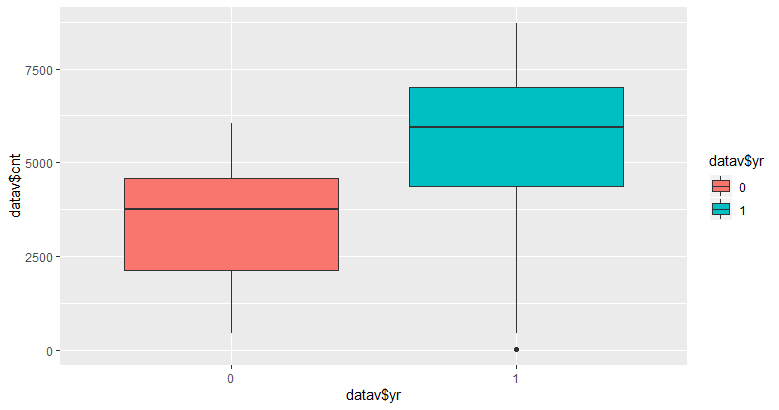
1. Cnt vs temp

From the below graph, we can see that there is an increase in bike rentals as temperature increses.



1. Cnt vs yr

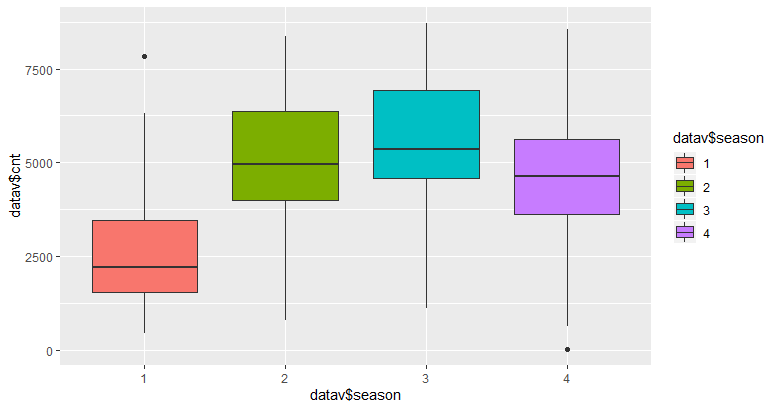
We can say that the count of bike rentals has increased in the second year.



1. Cnt vs season

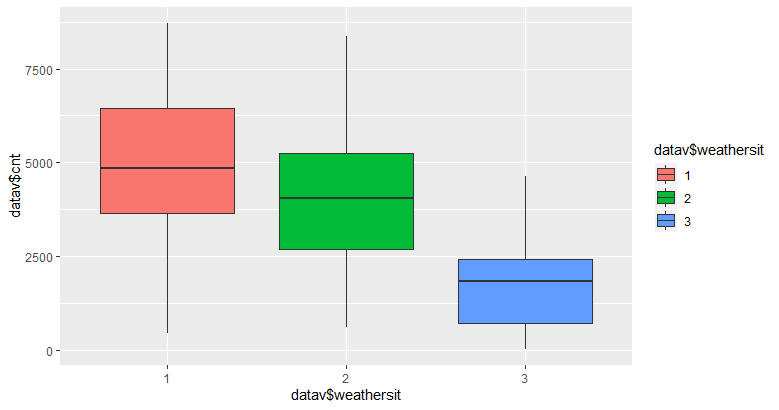
The four seasons mentioned below are spring, summer, fall and winter.

The below plot shows that the number of bike rentals is considerably more during fall.



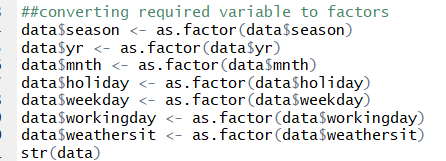
1. Cnt vs weather situation

Based on weather situation we can say that the count of bike rentals is more when it is the 1st weather situation i.e. Clear, Few clouds, Partly cloudy, Partly cloudy .

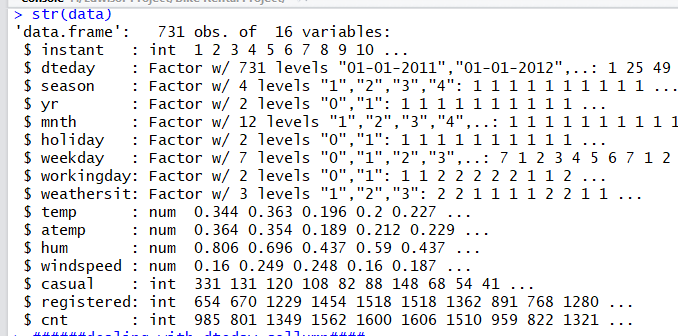


### Dealing with datatypes

Variables ‘season’, ‘yr’, ‘mnth’, ‘holiday’, ‘weekday’, ‘workingday’ and ‘weathersit’ have data type as integer. These are converted into factors using the below code.

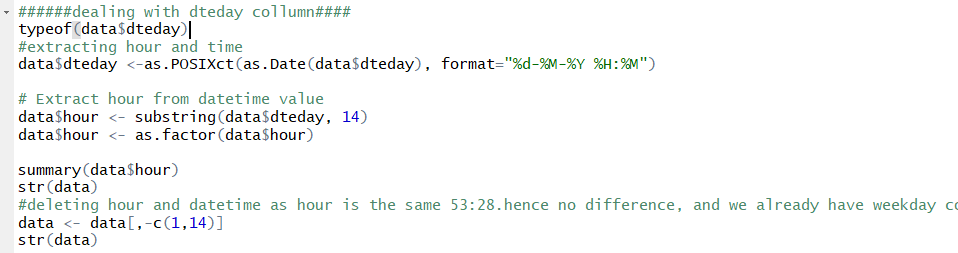


Structure of data after converting data type of variables is as below.

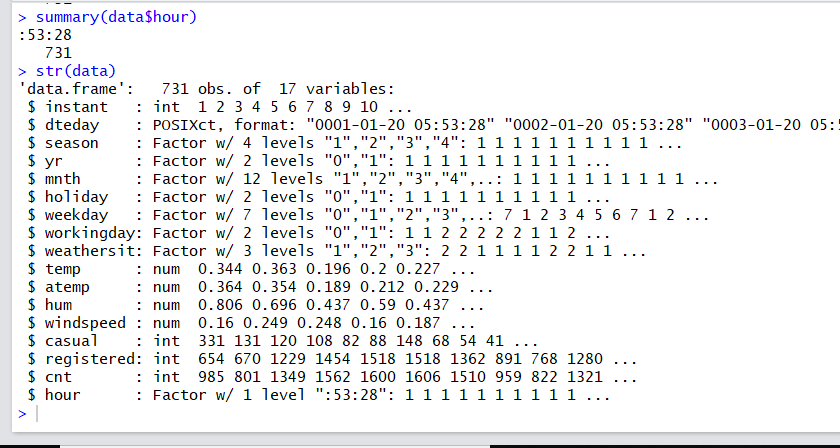


### Feature Engineering

The given data contains variable ‘dteday’. Let’s try extracting hour value from date as I believe hour can be an important factor affecting bike rental count.

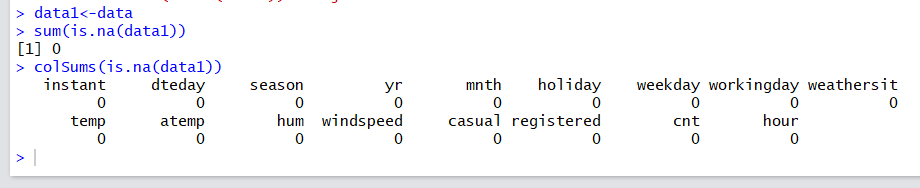


‘dteday’ is converted into datetime format and the hour variable is ceated. However, we can see from the summary of the variable that all the 731 observations have the same hour value i.e 53:28. This means that we cannot consider the hour variable. Hence, we can delete the ‘dteday’ variable as we already have other date information in ‘mnth’ and ‘yr’ variables.



### Missing Value Analysis

There are no missing values in the data.

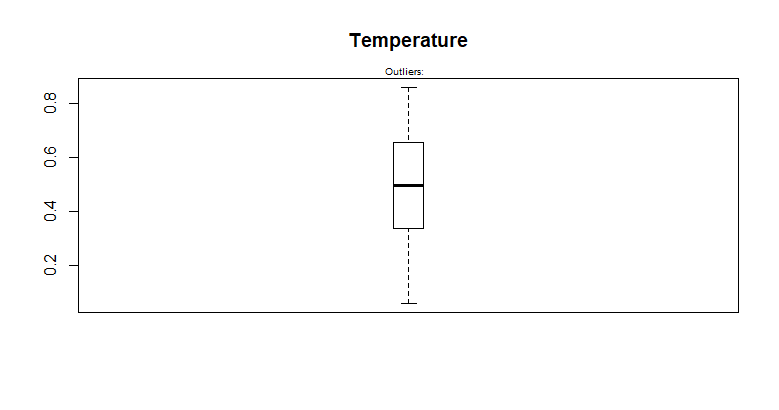


### Outlier Analysis

Checking if there are abnormal values in the numerical data variables which have cause large errors during model building.

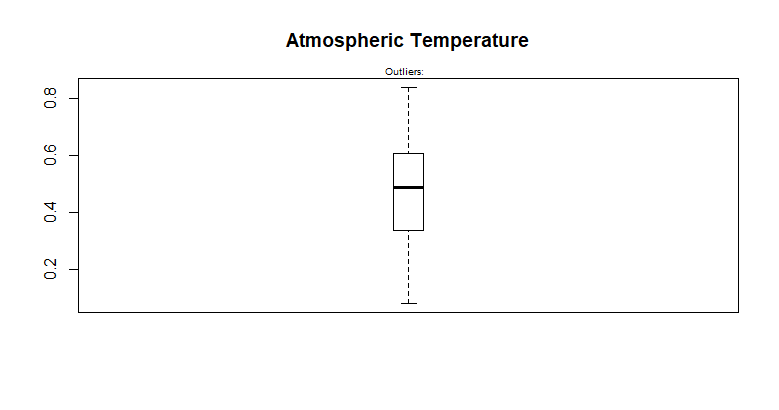
1. TEMP

There are no outliers in ‘temp’ variable.



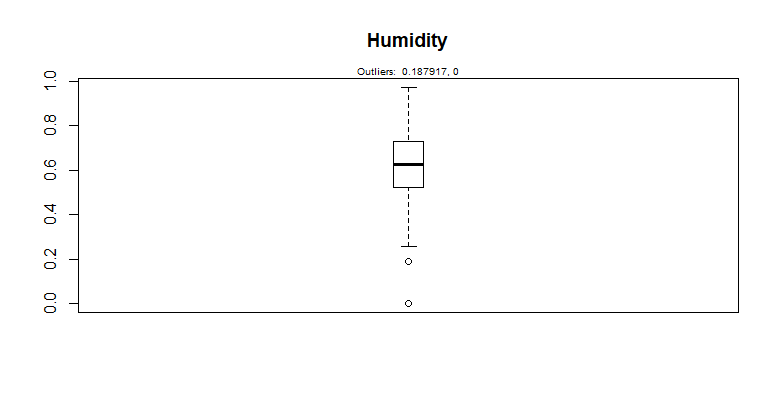
1. ATEMP

There are no outliers in ‘atemp’ variable.



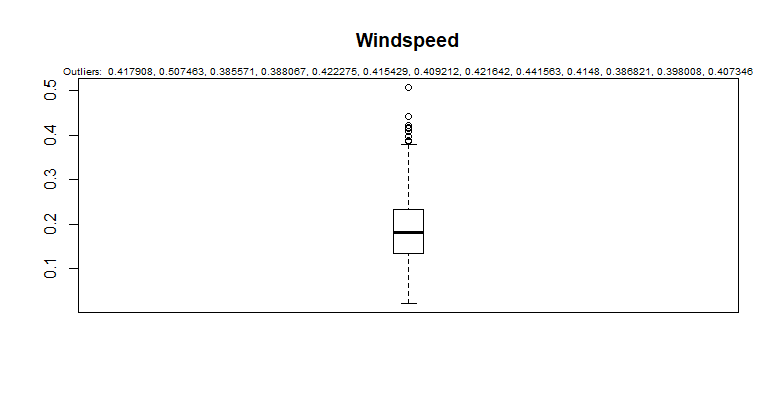
1. HUM

There are 2 outliers in ‘hum’ variable.



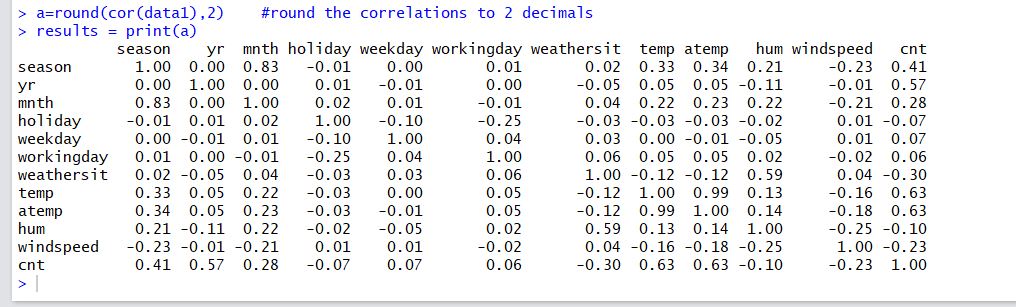
1. WINDSPEED

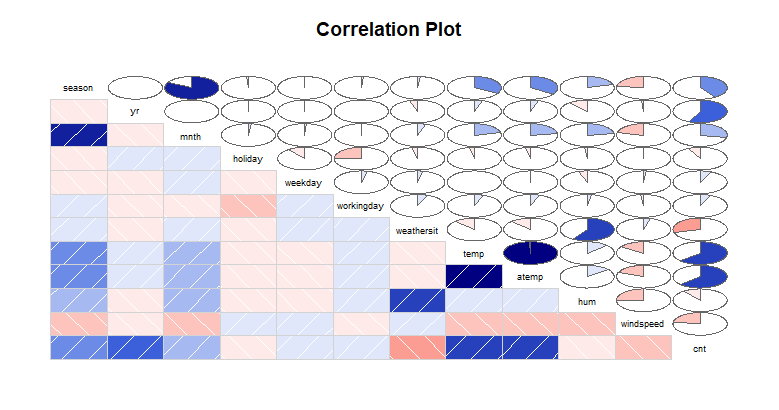
The ‘windspeed’ variable has some outliers present.



### Feature Selection

Correlation matrix: Using the correlation matrix and the correlation plot, we can identify if any variables in the data have high correlation amongst themselves. The presence of such variables might hamper the results. From the matrix and the plot, we can say that ‘atemp’ has a very high correlation with ‘temp’. Hence, we wiill be removing ‘atemp’ from the data.





Chapter 3

# **Conclusion**

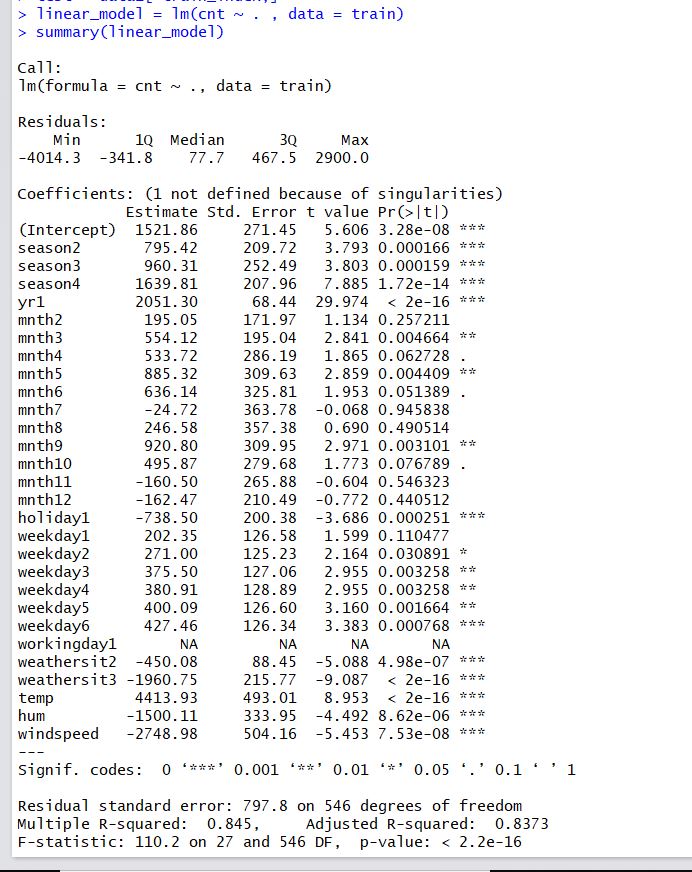
## **Model Selection**

We have completed pre-processing of the data by removing outliers, performing feature selection. Since the available numerical data is in the normalised form, the data has not further been normalised. Now, this data is available for model building.

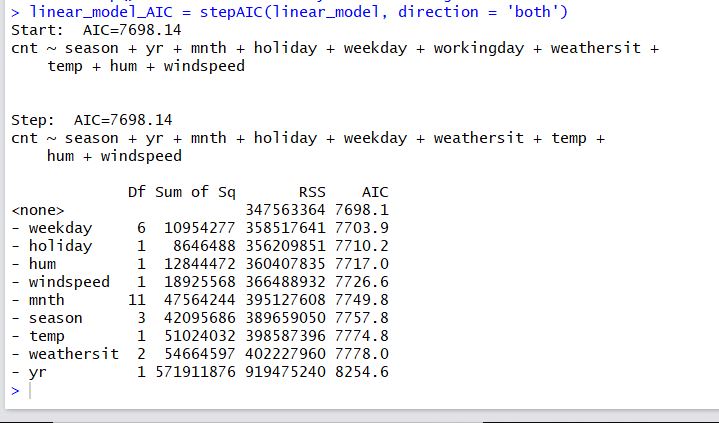
We have used 3 types of models, namely linear regression model, decision tree regression model and the random forest regression model. We have used 80% data for training the model and 20% data for prediction.

### Multiple Linear Regression

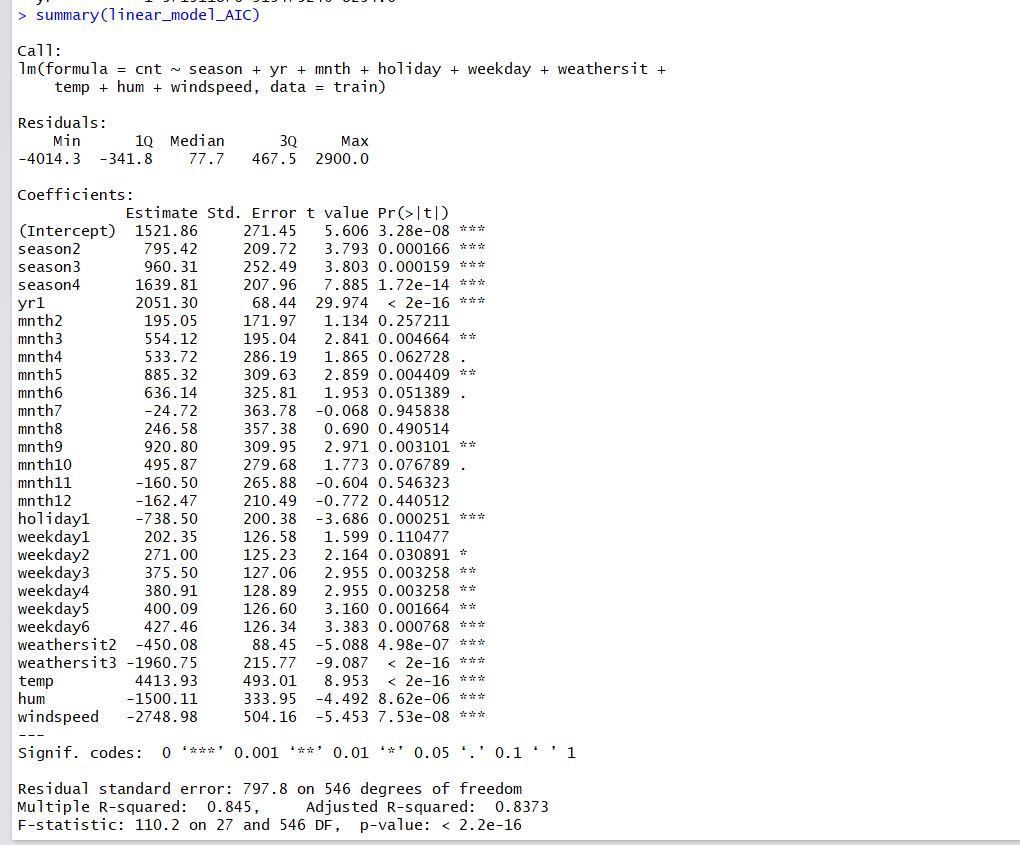
1. Multiple Linear regression(without log())



We can explain about 84% of the data using multiple linear regression. Applying a stepwise selection method with AIC we get the below model

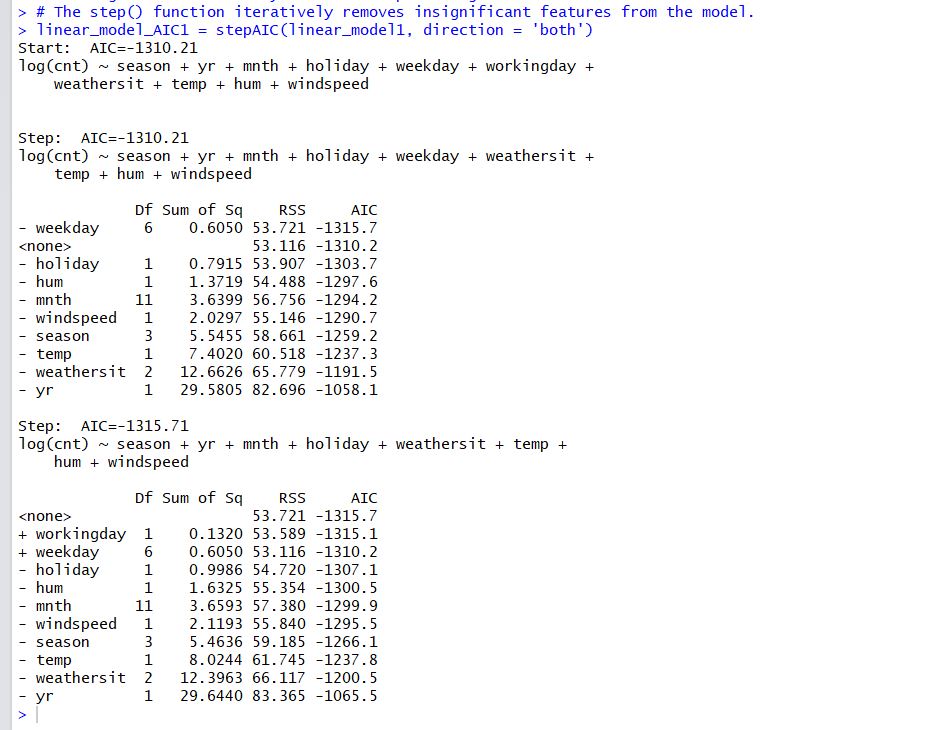


The summary statistics for the stepwise model is as follows:

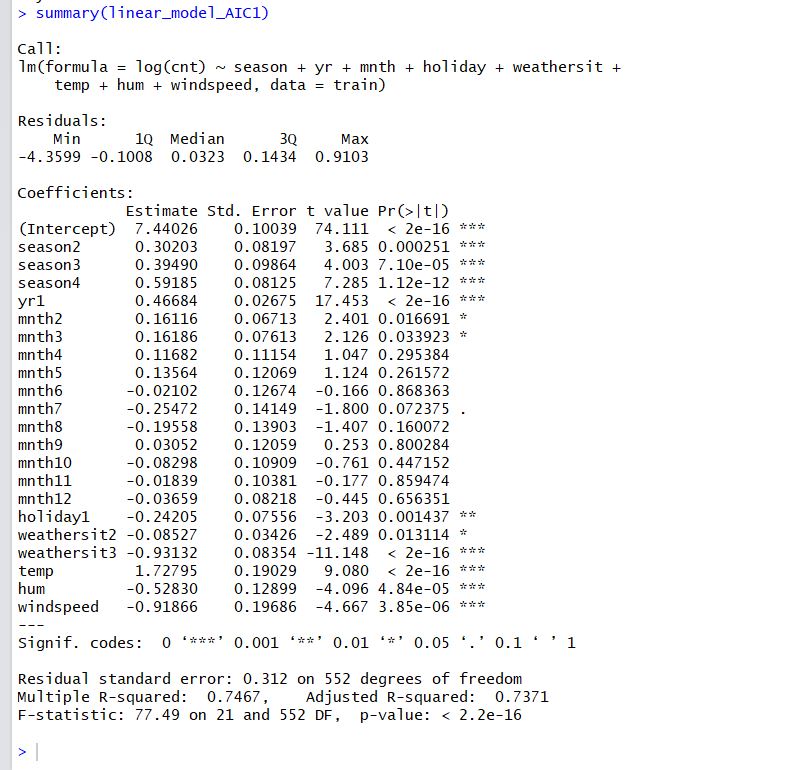


1. Multiple Linear Regression(With Log())

Using the log function in linear regression, we get the below model.

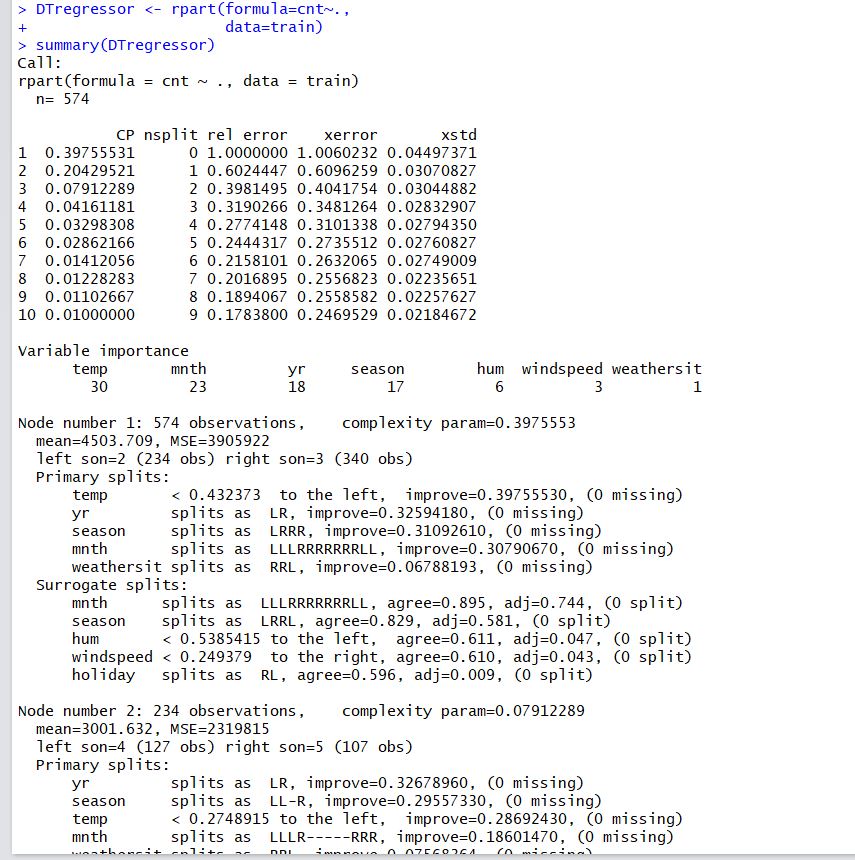


The summary statistics is as follows:



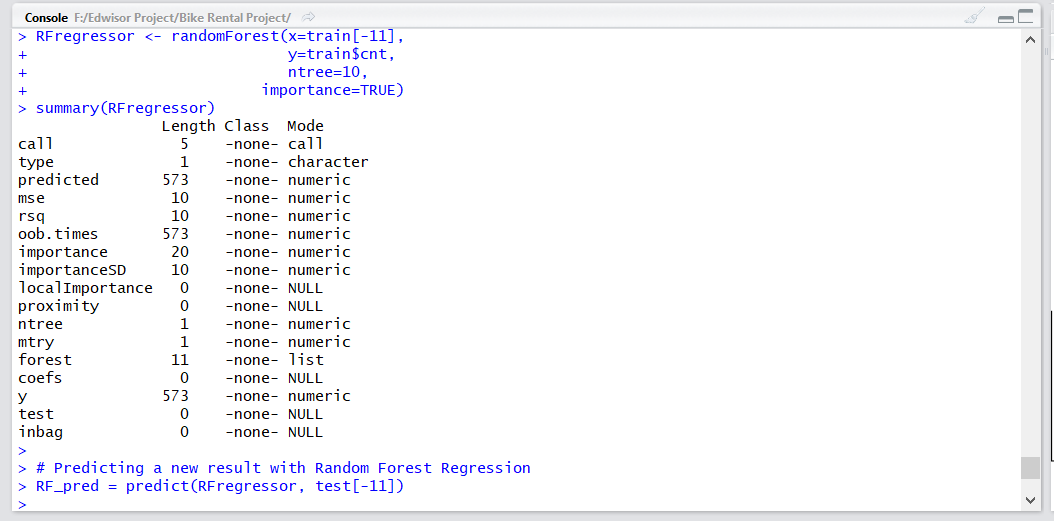
### Decision Tree Regression

The model and the summary statistics for the decision tree regression are as below. According to the model, we can see that tmp, mnth, yr have higher importance.

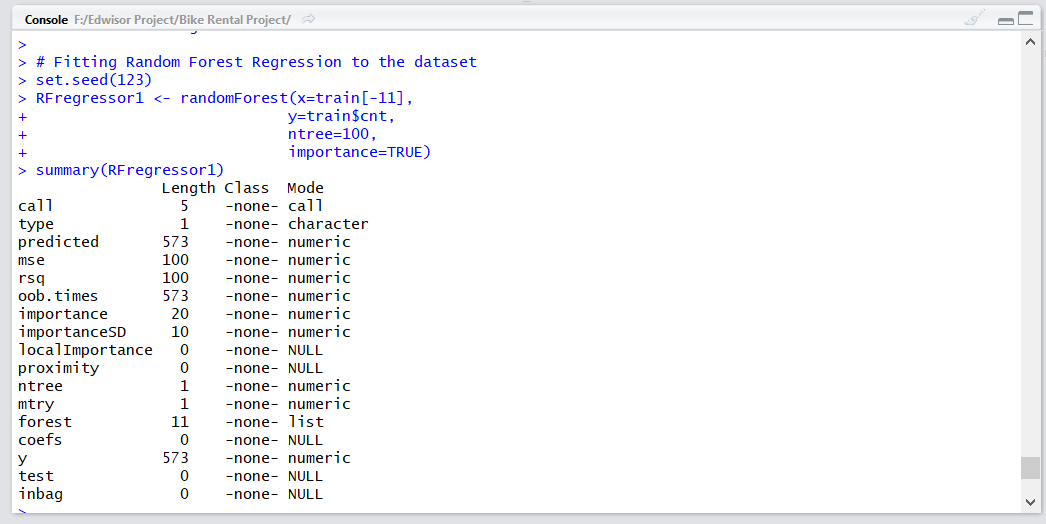


### Random Forest Regression

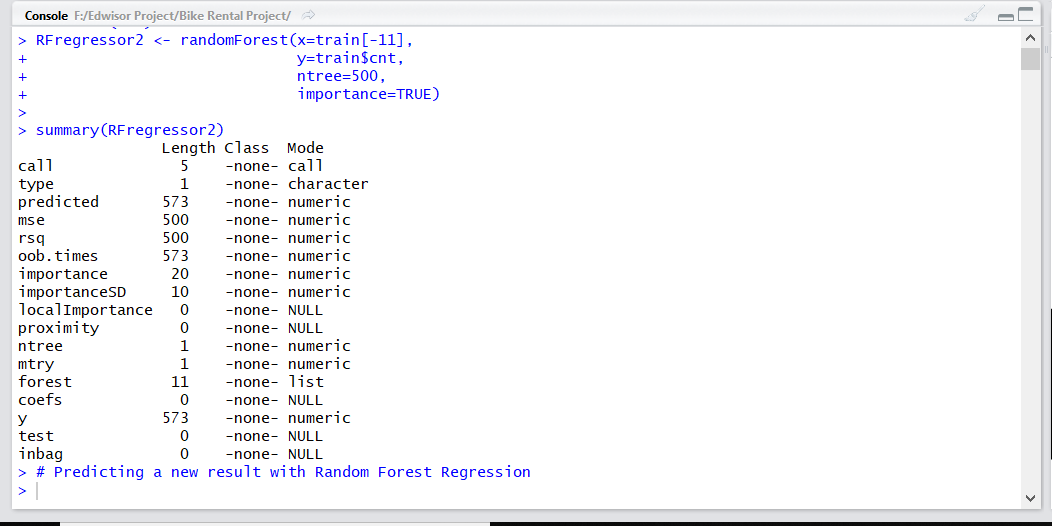
1. Random forest regression with (ntrees=10)



1. Random forest regression with (ntrees=100)

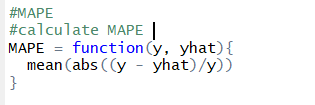


1. Random forest regression with (ntrees=500)



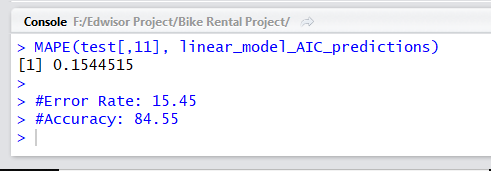
## **Model Evaluation**

We’ll be using Mean Absolute Percentage Error(MAPE) to evaluate the models.Mean Absolute Percentage Error is calculated using the below function.

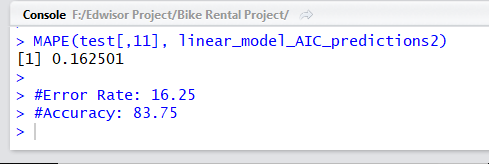


MAPE values for the various models are as follows:

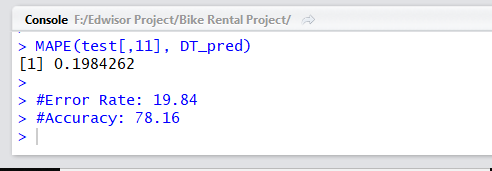
1. Linear Regression (without log)



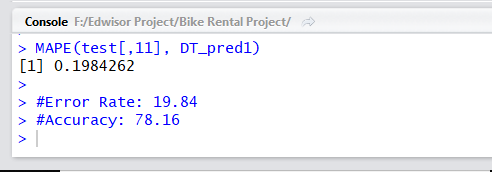
1. Linear Regression (without log)



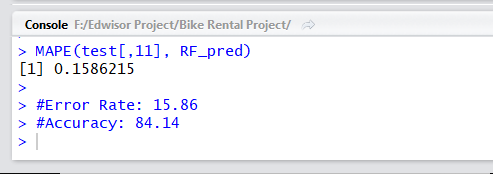
1. Tree Regression



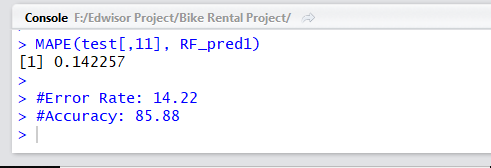
1. Decision Tree Regression(min split=2)



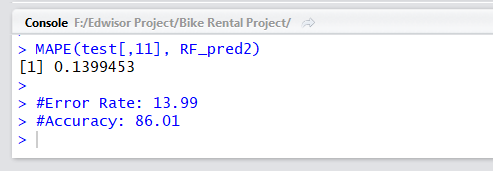
1. Random Forests Regression(ntree=10)



1. Random Forests Regression(ntree=100)



1. Random Forests Regression(ntree=500)



## **Model Selection**

Based on the evaluation of all the models discussed above, we can conclude that the Random Forests regression model with ntree=500 has performed better than the rest giving a better accuracy ,less MAPE and less RMSE values. Hence, we can proceed with the Random forests regression model for predicting the count of bike rentals.