**Project Name –Prediction Of Bike Rental Counts**

**GAURI KANDLIKAR ON 25TH JANUARY 2020**

**Contents**

1. Introduction ……………………………………………………………………………………………………………………………1

1.1 Problem Statement …………………………………………………………………………………………………………….1

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

2. Methodology …………………………………………………………………………………………………………………………..5

2.1 Pre-Processing ………………………………………………………………………………………………………………………5

2.2 Distribution of continuous variables ……………………………………………………………………………………7

2.3 Distribution of categorical variables ……………………………………………………………………………………8

2.4 Relationship of Continuous variables against bike count …………………………………………………..10

2.5 Detection of outliers ………………………………………………………………………………………………………….12

2.6 Feature Selection ……………………………………………………………………………………………………………..13

3. Modelling…………………………………………………………………………………………………………………………15

3.1 Model Selection ……………………………………………………………………………………………………………16

3.2 Multiple Linear Regression ……………………………………………………………………………………………..16

3.3 Decision Tree ……………………………………………………………………………………………………………………… 17

4. Conclusion……………………………………………………………………………………………………………………………….17

4.1 MAE ……………………………………………………………………………………………………………………………………….18

5.1 Figures …………………………………………………………………………………………………………………………………21

6. R code……………………………………………………………………………………………………………………………………….23

**Chapter 1**

**Introduction**

* 1. **Problem Statement**

Problem statement - The objective of this Case is to Predication of bike rental count on daily based on the environmental and seasonal settings. The details of data attributes in the dataset are as follows -

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\_max- t\_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.

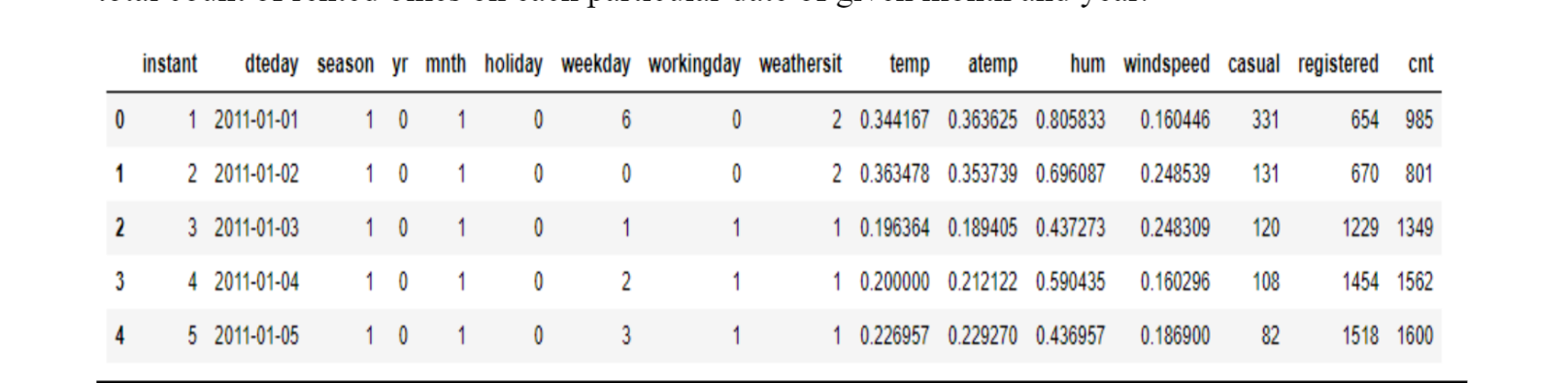
The values are divided to 67 (max) casual: count of casual users

registered: count of registered users cnt: count of total rental bikes including both casual and registered

The goal of this problem is to predict the bike rental counts with consideration of weather, season and some environmental conditions. By predicting the counts it would be convenient to get prepared and ready for high demand of bikes during the peak hours.

* 1. **Data**

The goal of this dataset is to build a regression model and analyse the given bike rental dataset. The snapshot of the given dataset is as follows.



The dataset consists of 16 variables and around 756 observations.

Out of which I have selected some variables for the project analysis.

**Chapter 2**

**Methodology**

**2.1 Pre Processing**

Any predictive modelling requires that we look at the data before we start modelling. However, in data mining terms looking at data refers to so much more than just looking. 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 this process we will first try and look at all the probability distributions of the variables. Most analysis like regression, require the data to be normally distributed. We can visualize that in a glance by looking at the probability distributions or probability density functions of the variable. In Fig we have plotted some distribution curves.

Bar plots, Box Plots etc. corrplot with the help of Exploratory Data Analysis.

**Variable** **Identification**:-In order to perform Analysis on the dataset we have to first identify dependent and independent variables.

In the given dataset dependent i.e target variable is counts, and the remaining season, holiday ,weekday, weathersit ,..etc are the independent variables.

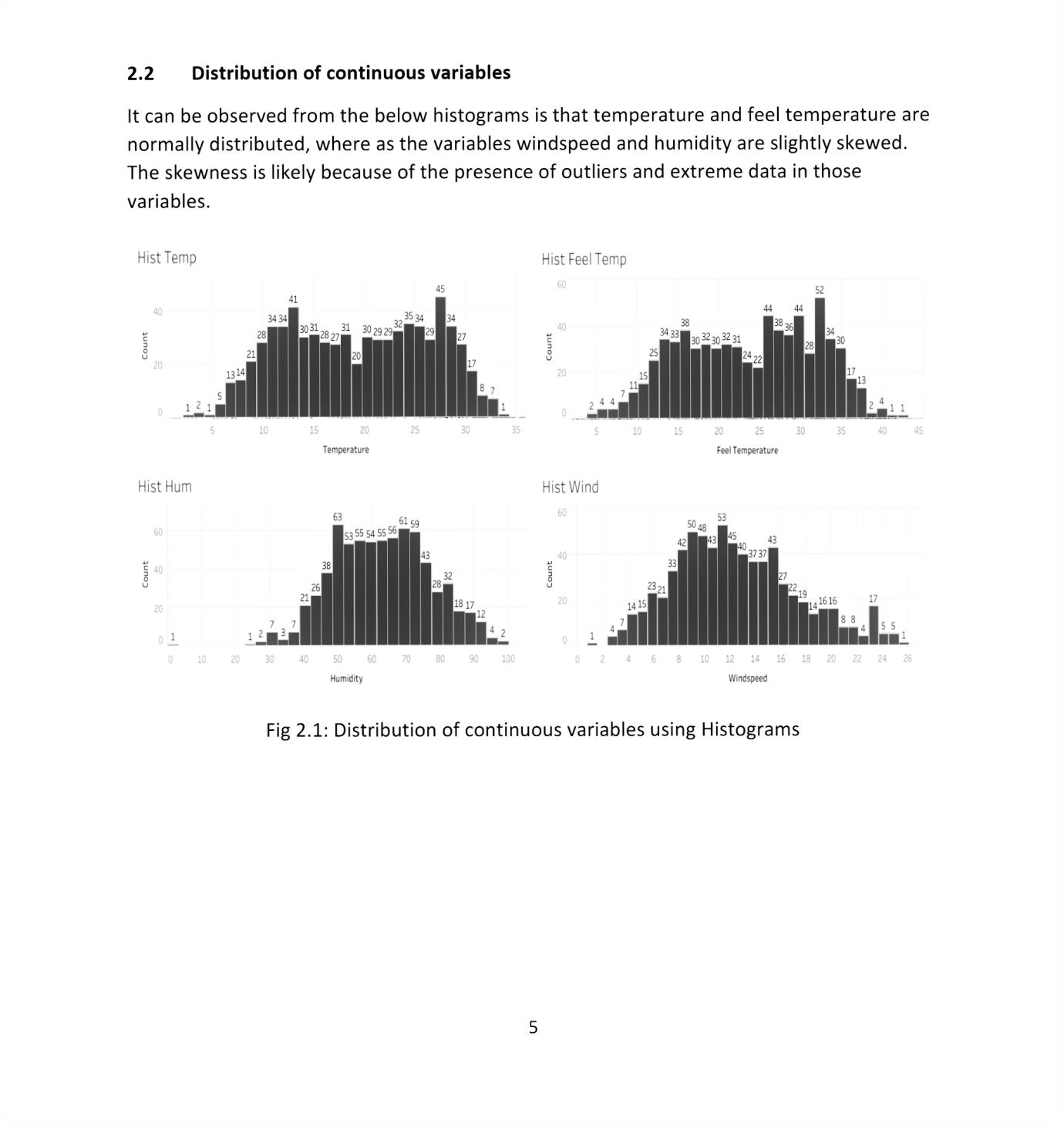
**Missing Value Analysis:-** We checked the given data whether there is a presence of missing values in the data or not.

But there are no missing values present in the data.

One of the other steps of pre-processing apart from checking for normality is the presence of outliers. We visualize the outliers using boxplots. In figure 2.3 we have plotted the boxplots of the continuous predictor variables with respect to each quality value ranging from 3 to 8. A lot of useful inferences can be made from these plots. First as you can see, we have outliers present in humidity and windspeed variables.

**2.2 Data Types:-** The variables in this data set are in type of character,numeric,factor..etc

|  |
| --- |
| str(day)  'data.frame': 731 obs. of 16 variables:  $ instant : int 1 2 3 4 5 6 7 8 9 10 ...  $ dteday : Factor w/ 731 levels "01-01-2011","01-01-2012",..: 1 25 49 73 97 121 145 169 193 217 ...  $ season : int 1 1 1 1 1 1 1 1 1 1 ...  $ yr : int 0 0 0 0 0 0 0 0 0 0 ...  $ mnth : int 1 1 1 1 1 1 1 1 1 1 ...  $ holiday : int 0 0 0 0 0 0 0 0 0 0 ...  $ weekday : int 6 0 1 2 3 4 5 6 0 1 ...  $ workingday: int 0 0 1 1 1 1 1 0 0 1 ...  $ weathersit: int 2 2 1 1 1 1 2 2 1 1 ...  $ temp : num 0.344 0.363 0.196 0.2 0.227 ...  $ atemp : num 0.364 0.354 0.189 0.212 0.229 ...  $ hum : num 0.806 0.696 0.437 0.59 0.437 ...  $ windspeed : num 0.16 0.249 0.248 0.16 0.187 ...  $ casual : int 331 131 120 108 82 88 148 68 54 41 ...  $ registered: int 654 670 1229 1454 1518 1518 1362 891 768 1280 ...  $ cnt : int 985 801 1349 1562 1600 1606 1510 959 822 1321 ... |
|  |
| |  | | --- | | So changing their types into our required format.  ##converting some numeric variables into catrgorical variables.  day$dteday=as.Date(day$dteday)  day$season=as.factor(day$season)  day$yr=as.factor(day$yr)  day$mnth=as.factor(day$mnth)  day$holiday=as.factor(day$holiday)  day$workingday=as.factor(day$workingday)  day$weathersit=as.factor(day$weathersit) | |



The distribution of categorical variable is as follows.

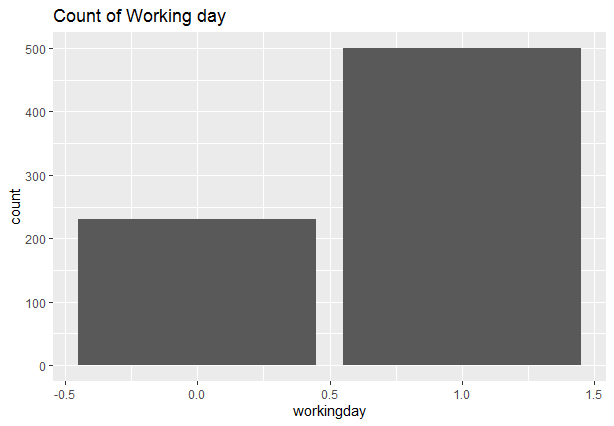
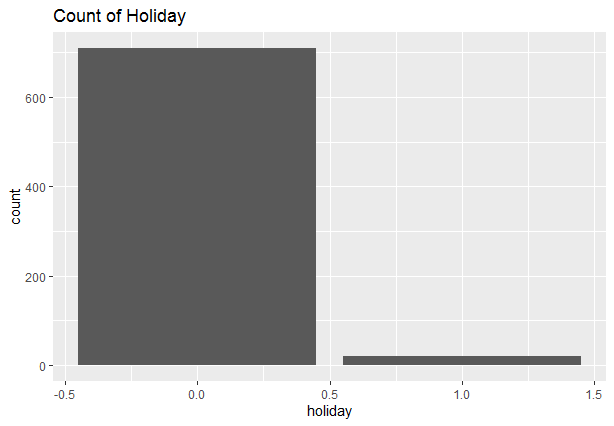
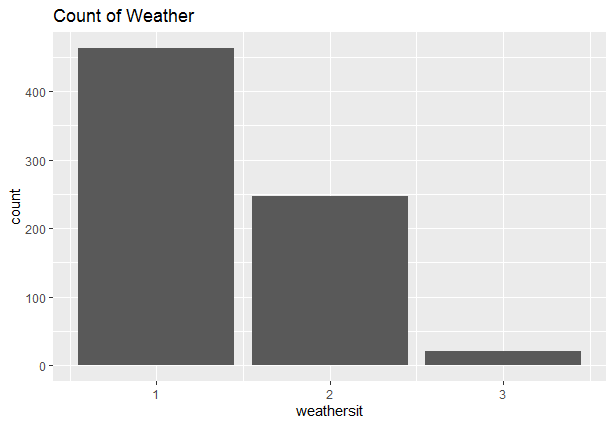
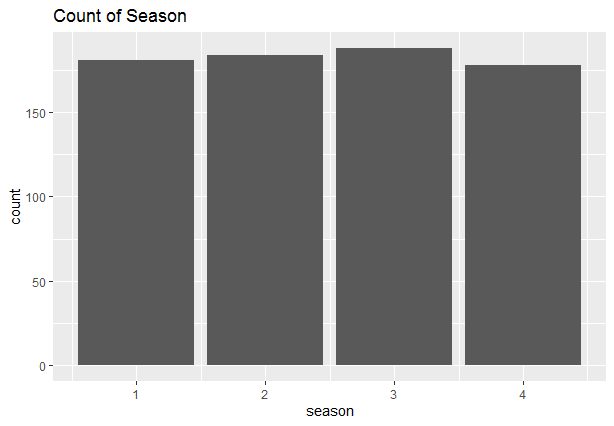
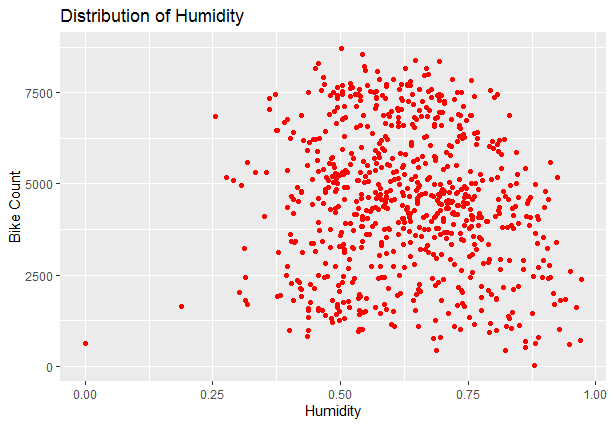
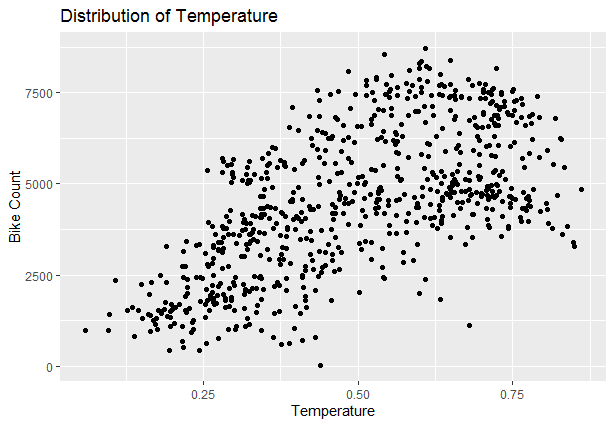
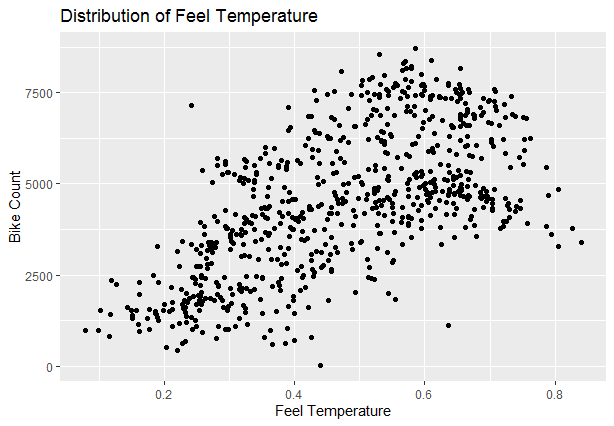


Fig 2.2 Distribution Of Categorical Variables Using Barplots

2.3 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 variable’s humidity and windspeed with the bike rental counts .

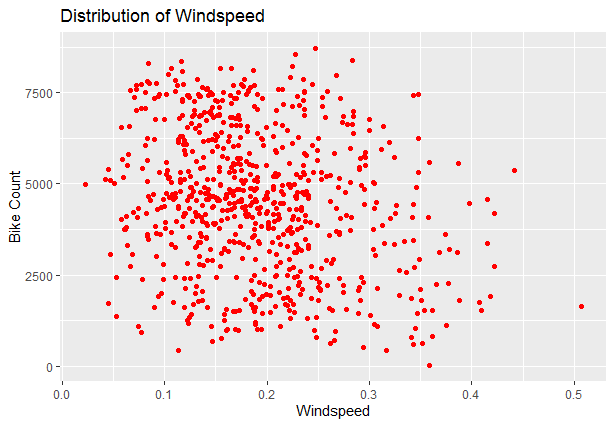
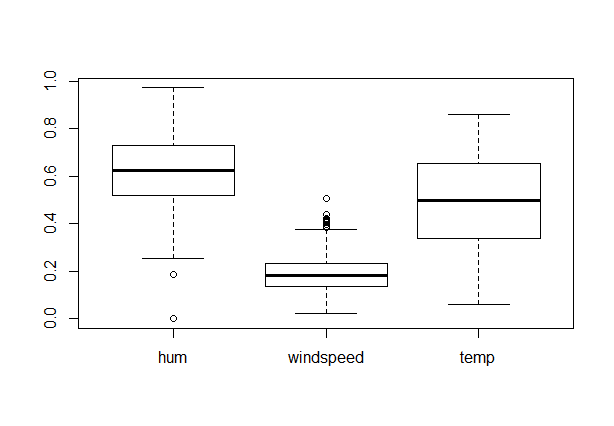
**

Fig 2.3: Scatter plot for continuous variables

**2.5: Detection of outliers:** Outliers are detected using boxplots. Below figure illustrates the boxplots for all the continuous variables.

Outliers are the anomalies in the data, generated due to improper handling of the data.

Outliers affect the accuracy of the model by impacting on the distribution of the data.



**Fig2.4 Boxplot For Continuous Variables**

Outliers can be removed using the Boxplot stats method, wherein the Inter Quartile Range (IQR) is calculated and the minimum and maximum value are calculated for the variables. Any value ranging outside the minimum and maximum value are discarded.

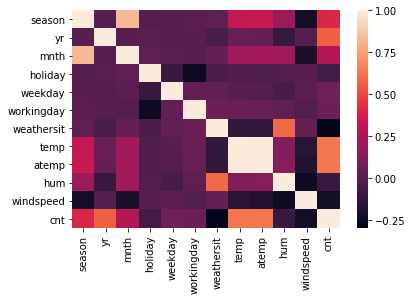
Wherever there are outliers in humidity and windspeed we first replace those with the NA values.

After replacing imputed with mean of that variable.

After that we can cross check those variables.

**2.5 Feature Selection**

Feature Selection reduces the complexity of a model and makes it easier to interpret. 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.



**Fig 2.6: Correlation plot of all the variables**

**H**ere the temp and atemp variables are highly positively correlated,so we will skip variable atemp from our final analysis.

**Chapter 3: Modelling**

3.1 Model Selection The dependent variable in our model is a continuous variable i.e., Count of bike rentals. Hence the models that we choose are Linear Regression, Decision Tree and Random Forest. The error metrics chosen for the problem statement is Mean Absolute Error (MAE).

3.2 Multiple Linear Regression Multiple linear regression is the most common form of linear regression analysis. Multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables. The independent variables can be continuous or categorical.

Linear Regression measures the impact of change in the dependent variables with the change in independent variables.

The summary of the model is as follows

Call:

lm(formula = cnt ~ ., data = train\_data)

Residuals:

Min 1Q Median 3Q Max

-4116.7 -468.9 28.0 558.0 2912.2

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1601.56 257.42 6.222 9.51e-10 \*\*\*

season 539.60 60.09 8.980 < 2e-16 \*\*\*

yr 2054.40 72.84 28.205 < 2e-16 \*\*\*

mnth -52.09 18.59 -2.802 0.005254 \*\*

holiday -657.78 242.83 -2.709 0.006956 \*\*

weekday 69.91 17.97 3.891 0.000112 \*\*\*

workingday 113.02 79.60 1.420 0.156194

weathersit -647.13 86.33 -7.496 2.51e-13 \*\*\*

temp 4986.52 216.15 23.069 < 2e-16 \*\*\*

hum -647.16 345.72 -1.872 0.061723 .

windspeed -2849.77 507.27 -5.618 3.02e-08 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 871.8 on 573 degrees of freedom

Multiple R-squared: 0.8420, Adjusted R-squared: 0.8355

F-statistic: 129.7 on 10 and 573 DF, p-value: < 2.2e-16

As you can see the Adjusted R-squared value, we can explain 83.55% of the data using our multiple linear regression model. By looking at the F-statistic and combined p-value we can reject the null hypothesis that target variable does not depend on any of the predictor variables. This model explains the data very well and is considered to be good. Even after removing the non-significant variables, the accuracy, Adjusted R-squared and Fstatistic change very negligibly, hence the accuracy of this model is chosen to be final. Mean Absolute Error (MAE) = 511.43.

MAPE = 12.63%.

**In python**:- ##conclusion :-RMSE MAE MAPE

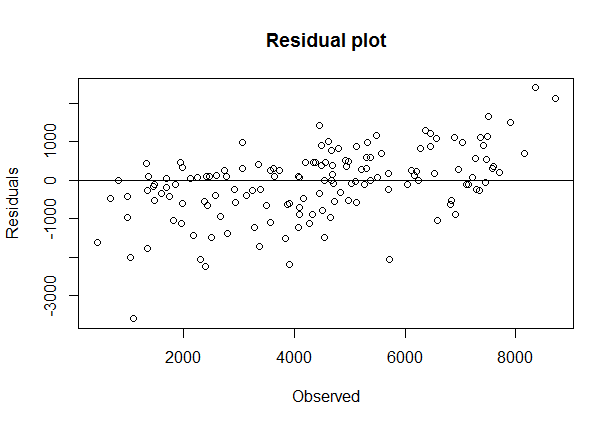
##Linear Regression 583.59 312.39 16.22

##Decision Tree 756.380 445.10 24.01

##Comparing two models Linear Regression Model has less error metrics

##Hence Linear Regression is good model for prediction of bike rental counts

**Residual Plot:-**



Hence from the above Residua Plot we can say that error points are spread randomly.

They are not identical.This is said to be a good fit.

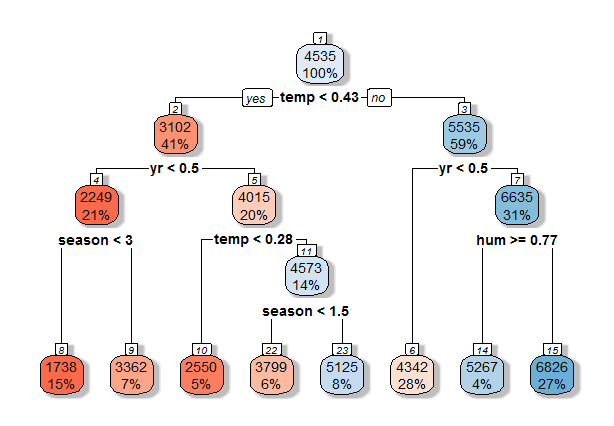
Hence the accuracy of this model= 87.83%. This model performs very well for this test data.

**3.3 Decision Tree**: A decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions. Using decision tree, we can predict the value of bike count. MAE for this mode= 609.

MAE=609

MAPE=27.81

Hence Accuracy of Decision Tree Model=72.81%



**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.

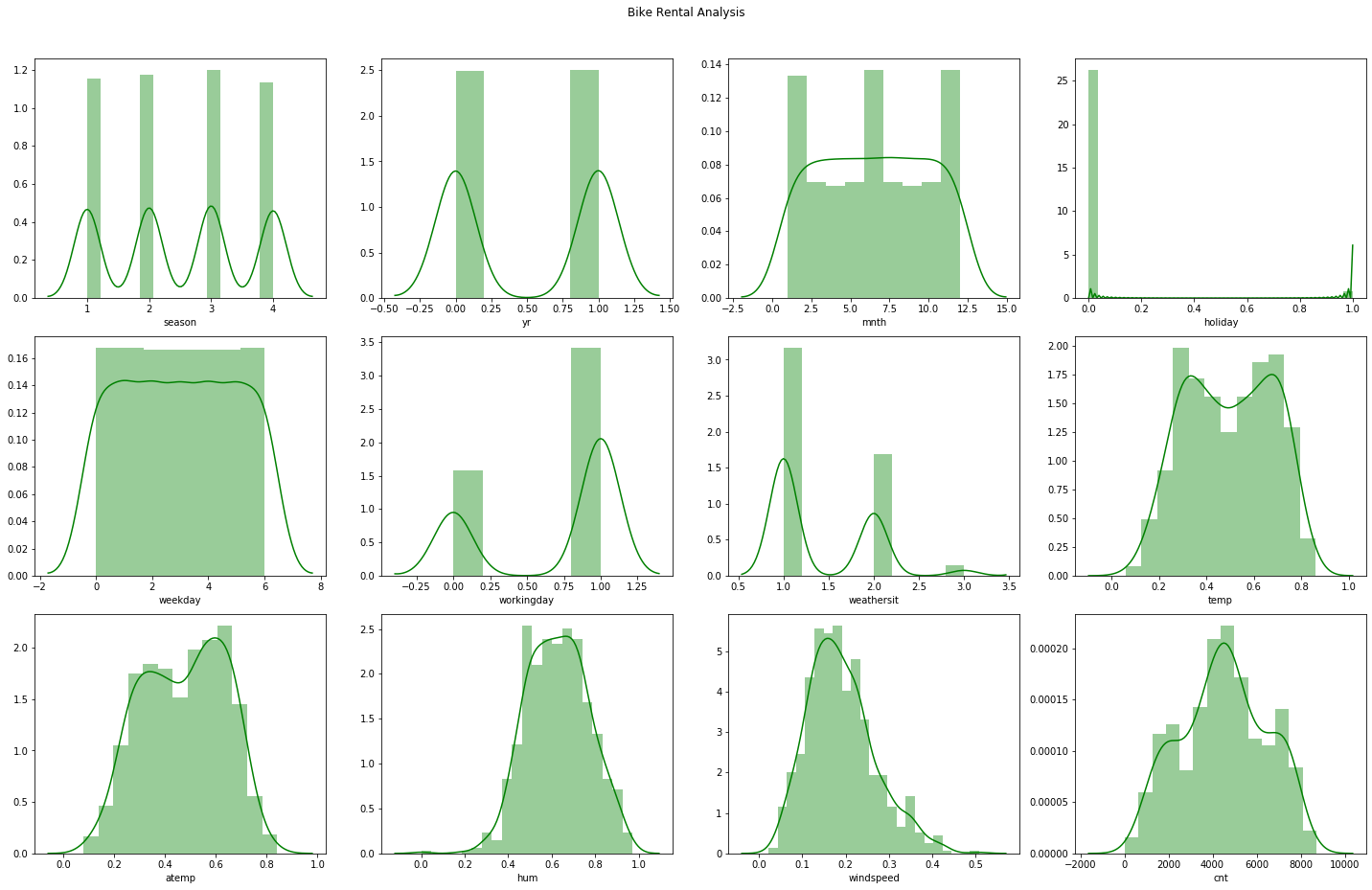
4.1 Mean Absolute Error (MAE) MAE is one of the error measures used to calculate the predictive performance of the model. We will apply this measure to our models that we have generated in the previous section. MAE <- function (actual, pred) { print(mean (abs (actual - pred))) } Linear Regression Model: MAE = 511.43

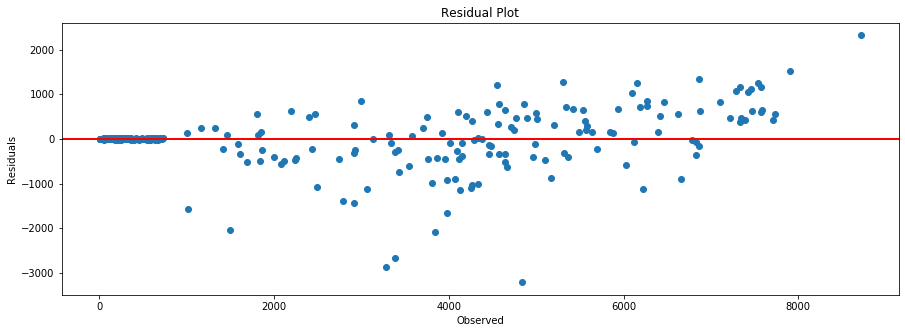
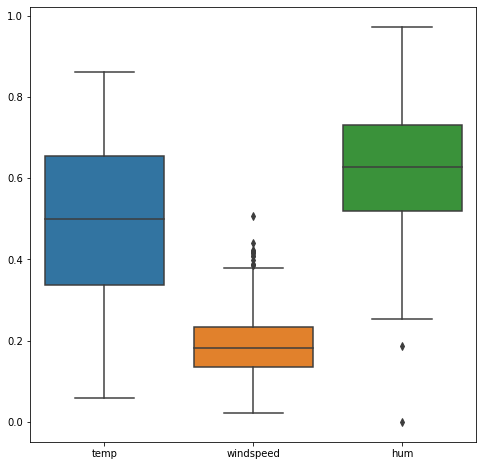
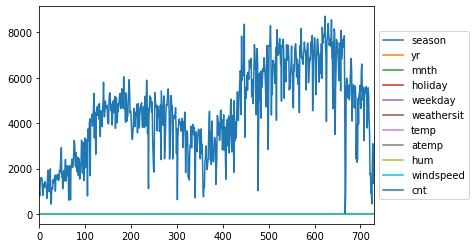
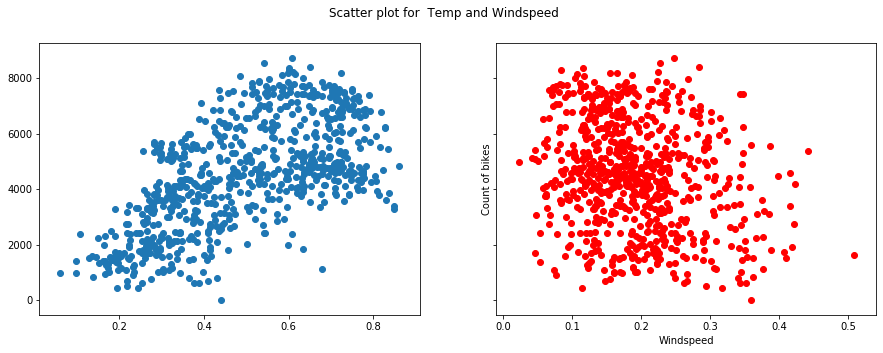
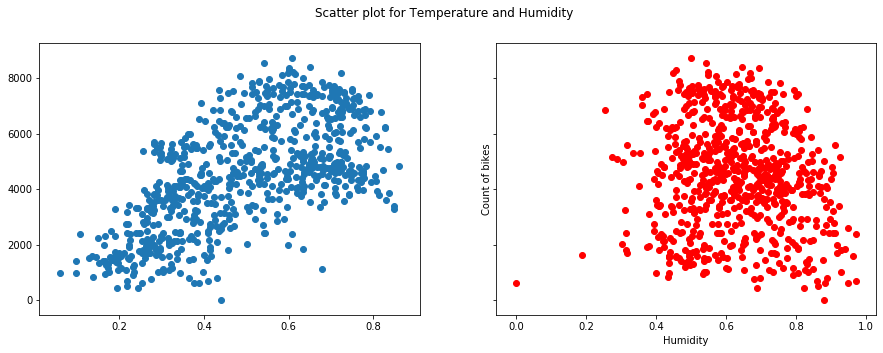
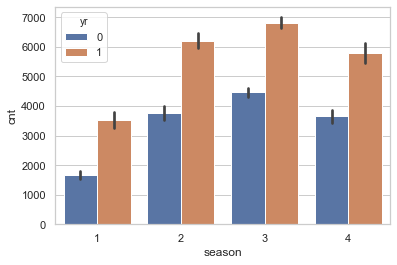
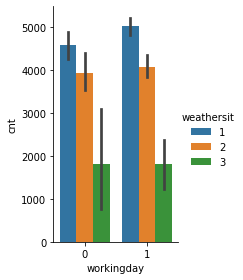
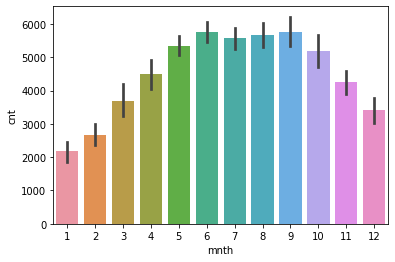
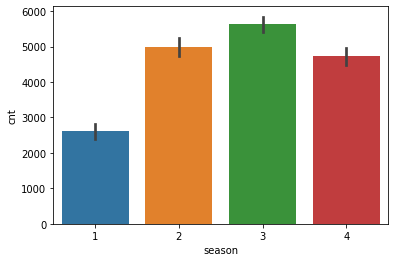
Decision Tree: MAE = 609.

Based on the above error metrics, Linear Regression is the better model for the bike rental counts

Hence we will chose the linear Regression model for prediction of bike rental counts.

**Some important Graphs:-**

****

****

**Complete R Code File**

**rm(list = ls())**

**library(ggplot2 )**

**library(grid)**

**library("tidyverse")**

**install.packages("randomForest")**

**library('randomForest')**

**install.packages("rpart")**

**library(rpart)**

**install.packages('rpart.plot')**

**library(rpart.plot)**

**library(randomForest)**

**install.packages("caret")**

**library("caret")**

**install.packages("DMwR")**

**library(DMwR)**

**setwd("D:\\edwisor\\bike\_rental")**

**#######Reading CSV file#######################**

**day=read.csv("bikee.csv")**

**View(day)**

**dim(day)##(731,16)**

**summary(day)#from this function we can say that there are no NA values in this dataset**

**str(day)**

**######converting some numeric variables into catrgorical variables.########**

**day$dteday=as.Date(day$dteday)**

**day$season=as.factor(day$season)**

**day$yr=as.factor(day$yr)**

**day$mnth=as.factor(day$mnth)**

**day$holiday=as.factor(day$holiday)**

**day$workingday=as.factor(day$workingday)**

**day$weathersit=as.factor(day$weathersit)**

**str(day)**

**unique(day$holiday)**

**#######################visualization###################################**

**##to visualize the distribution of the continuous variables**

**v1=ggplot(day,mapping=aes(day$temp))+geom\_histogram(bins = 25)+geom\_density()##bimodal**

**v2=ggplot(day,mapping=aes(day$atemp))+geom\_histogram(bins = 25)+geom\_density()**

**v3=ggplot(day,mapping=aes(day$hum))+geom\_histogram(bins = 25)+geom\_density()##left skewed**

**v4=ggplot(day,mapping=aes(day$windspeed))+geom\_histogram(bins = 25)+geom\_density()##right skewed**

**v1**

**v2**

**v3**

**v4**

**##In brief all the continuous variables selected are normally distributed.**

**##########Check the distribution of categorical Data using bar graph####################**

**bar1 = ggplot(data = day, aes(x = season)) + geom\_bar() + ggtitle("Count of Season")**

**bar2 = ggplot(data = day, aes(x = weathersit)) + geom\_bar() + ggtitle("Count of Weather")**

**bar3 = ggplot(data = day, aes(x = holiday)) + geom\_bar() + ggtitle("Count of Holiday")##**

**bar4 = ggplot(data = day, aes(x = workingday)) + geom\_bar() + ggtitle("Count of Working day")##more no of bike counts on working day ,here 0-holiday 1-workingday**

**bar1**

**bar2**

**######################## ##BIVARIATE AND MULTIVARIATE ANALYSIS#######################**

**###################To check the relationship between dependent and independent variables using scatterplot**

**p1 = ggplot(data = day, aes(x =temp, y = cnt)) + ggtitle("Distribution of Temperature") + geom\_point() + xlab("Temperature") + ylab("Bike Count")##from this scatter plot we can say that temp and bike counts are positively correlated**

**p2 = ggplot(data = day, aes(x =hum, y = cnt)) + ggtitle("Distribution of Humidity") + geom\_point(color="red") + xlab("Humidity") + ylab("Bike Count")##there is good amount of correlation between humidity and bike counts**

**p3 = ggplot(data = day, aes(x =atemp, y = cnt)) + ggtitle("Distribution of Feel Temperature") + geom\_point() + xlab("Feel Temperature") + ylab("Bike Count")##the variable atemp and bike counts are highly correlated**

**p4 = ggplot(data = day, aes(x =windspeed, y = cnt)) + ggtitle("Distribution of Windspeed") + geom\_point(color="red") + xlab("Windspeed") + ylab("Bike Count")## windspeed and bike counts are correlated**

**#bar plot for season wise monthly distribution of counts:-**

**ggplot(day, aes(season,fill=mnth) ) +**

**geom\_bar(position = 'dodge')**

**##From above we can say that there is decreasing trend of bike counts in season 1**

**##whereas in other three seasons bike counts are initially increasing and after that it is decreasing**

**ggplot(day,aes(x=yr,y=cnt,fill=yr))+geom\_col()+theme\_bw()+**

**labs(x='Year',y='Total\_Count',title='Year wise distribution of counts')**

**##We can say that there is highest bike counts in the year 2012 as compared to year 2011.**

**###Distribution of Bike Counts during Holiday################################**

**ggplot(day,aes(x=holiday,y=cnt,fill=season))+geom\_col()+theme\_bw()+**

**labs(x='holiday',y='Total\_Count',title='distribution of counts During Holidays')**

**##From the above plot we can say that when there is no holiday the bike counts are highest**

**##whereas when there is a holiday the bike counts are negligible as compared to the non holiday**

**##here 0- workingday,1-holiday**

**#####################Distribution of counts During Workingday:-######################################**

**ggplot(day,aes(x=workingday,y=cnt,fill=season))+geom\_col()+theme\_bw()+**

**labs(x='workingday',y='Counts',title='Workingday wise distribution of counts')**

**##Here 1-workingday,0-holiday**

**##from the above graph we can say that there are highest no of bike counts on the working day as comapared to the holiday**

**##In the season of summer and fall on the workingady there is more demand of bikes**

**###############Distribution of counts During Weather conditions:-############################**

**ggplot(day,aes(x=weathersit,y=cnt,fill=season))+geom\_col()+theme\_bw()+**

**labs(x='Weather\_condition',y='counts',title=' distribution of counts in Different Weather conditions')**

**##Here 1: Clear, Few clouds, Partly cloudy, Partly cloudy 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist 3: Light Snow, Light Rain +**

**##From the above plot we can clearly say that when the weather is clear, partly cloudy,the bike counts are highest**

**##In the same weather conditions when there is summer and fall season the bike counts are always high irrespective of the other factors.**

**#############Select the continuous variables to check the outliers#########**

**cont\_var=subset(day,select = c('hum','windspeed','temp'))**

**View(cont\_var)**

**boxplot(cont\_var)**

**##From the above boxplot we can see that the outliers are present in humidity and windspeed**

**sum(is.na(day))**

**########## Replace and impute the outliers:-########**

**#create subset for windspeed and humidity variable:-**

**cont\_var<-subset(cont\_var,select=c('windspeed','hum'))**

**View(cont\_var)**

**#column names of wind\_hum:-**

**cnames<-colnames(cont\_var)**

**for(i in cnames){**

**val=cont\_var[,i][cont\_var[,i] %in% boxplot.stats(cont\_var[,i])$out] #separating outlier values in val variable.**

**cont\_var[,i][cont\_var[,i] %in% val]= NA # Replacing outliers with NA**

**}**

**########## ##crosschecking#########**

**sum(is.na(cont\_var))##15 NA values**

**##################Imputating the missing values using mean imputation method############**

**cont\_var$windspeed[is.na(cont\_var$windspeed)]=mean(cont\_var$windspeed,na.rm=T)**

**cont\_var$hum[is.na(cont\_var$hum)]=mean(cont\_var$hum,na.rm=T)**

**######### #Remove the windspeed and humidity variable in order to replace imputed data:-########**

**df<-subset(day,select=-c(windspeed,hum))**

**#Combined df and cont\_var data frames:-**

**df1=cbind(df,cont\_var)**

**head(df1,5)**

**sum(is.na(df1))**

**##############################FEATURE SELECTION#########################################**

**install.packages("corrgram")**

**library(corrgram)**

**corrgram(day, order = F, upper.panel=panel.pie, text.panel=panel.txt, main = "Correlation Plot")**

**##From the above coorrplot we can temp and atemp are highly positively correlated,so we will skip atemp ,some variables are positively correlated with each other & some are nagatively correlated with each other**

**View(df1)**

**#####Selecting the Required Columns for Model Building##################**

**new\_df1=subset(df1,select = c('season','yr','mnth','holiday','weekday','workingday','weathersit','temp','hum','windspeed','cnt'))**

**#########################Divide the data into train and test#############################**

**set.seed(123)**

**train\_index = sample(1:nrow(new\_df1), 0.8 \* nrow(new\_df1))**

**train\_data = new\_df1[train\_index,]**

**test\_data = new\_df1[-train\_index,]**

**###################Train the data using linear regression####################################**

**lr\_model = lm(formula = cnt~., data = train\_data)**

**######Check the summary of the model#####################**

**summary(lr\_model)###Adjusted R-squared: 0.8355**

**## R-squared: 0.842,F statistics:129.7**

**#######################Predict the test cases#############################**

**lr\_predictions = predict(lr\_model, test\_data[,1:10])**

**##############################Create dataframe for actual and predicted values########################**

**df = data.frame("actual"=test\_data[,11], "pred"=lr\_predictions)**

**head(df)**

**rmse<-RMSE(lr\_predictions, test\_data$cnt)**

**print(rmse)##692.3745**

**#######################Mean squared error:-##########################################**

**mae<-MAE(lr\_predictions, test\_data$cnt)**

**print(mae)##511.43**

**#################################calculate MAPE######################################**

**MAPE = function(actual, pred){**

**print(mean(abs((actual - pred)/actual)) \* 100)**

**}**

**MAPE(test\_data$cnt, dt\_predictions)**

**####################################Residual plot:-###########################**

**y\_test<-test\_data$cnt**

**residuals<-y\_test-lr\_predictions**

**plot(y\_test,residuals,xlab='Observed',ylab='Residuals',main='Residual plot')**

**abline(0,0)**

**###########From the above graph we can say that the error is scattered randomly which is said to be a good fit.**

**dt\_model = rpart(cnt ~ ., data = train\_data, method = "anova")**

**summary(dt\_model)**

**###############Visualize the learned decision tree model:-########################**

**rpart.plot(dt\_model, box.palette="RdBu", shadow.col="gray", nn=TRUE,roundint=FALSE)**

**############################Predict the test cases#########################**

**dt\_predictions = predict(dt\_model, test\_data[,-11])**

**##################Create dataframe for actual and predicted values########**

**df = data.frame("actual"=test\_data[,11], "pred"=dt\_predictions)**

**head(df)**

**rmse<-RMSE(dt\_predictions, test\_data[,11])**

**print(rmse)##852.8797**

**mae<-MAE(dt\_predictions, test\_data$cnt)**

**print(mae)**

**#calculate MAPE**

**MAPE = function(actual, pred){**

**print(mean(abs((actual - pred)/actual)) \* 100)**

**}**

**MAPE(test\_data$cnt, dt\_predictions)##27.8162**

**##############comparing the two models##############**

**##When we compare two models Linear Regression and Decision Tree**

**##RMSE(LINEAR REGRESSION)=692.37 RMSE(DECISION TREE)=852.87**

**##MSE(LINEAR REGRESSION)=511.43 MAE(DECISION TREE)=609**

**###When we compare the root mean squared error and mean absolute error of 2 models, the Linear Regression model has less root mean squared error and mean absolute error. So, the Linear Regression model is best for predicting the bike rental count on daily basis.**

**Bike\_prediction=data.frame(y\_test,lr\_predictions)**

**write.csv(Bike\_prediction,'Bike\_Renting\_R.CSV',row.names=F)**

**Bike\_prediction**

**str(day)**