**Predicting Bike Rental count**

Priyanka Chettri

28-Jul-2019

Contents

1. **Introduction**
   1. Problem Statement 3
   2. Data 3
2. **Methodology**
   1. Pre-Processing 5
   2. Distribution of Continuous variables 5
   3. Distribution of Categorical variables 6
   4. Relationship of Continuous variables against bike count 9
   5. Outlier Analysis 10
   6. Feature Selection 13
3. **Modelling**
   1. Model Selection 15
   2. Multiple Linear Regression 15
   3. Decision Tree 16
   4. Random Forest 17
4. **Conclusion**
   1. MAE 18
5. **R Code** 19

**Chapter 1**

**Introduction**

* 1. **Problem Statement**

The aim of this project is to predict the number of bikes rented based on seasonal and environmental settings. Bike rental has now become one of the most lucrative and fast-moving businesses. So, it is very important that the rental bikes always be available whenever a customer desire one. However, it is equally important the number of bikes do not exceed the demand by a huge amount as this would result in wastage of the client’s resources. This project aims to predict the desired number of bikes so the business can function at its utmost without any wastage or scarcity in times of peak demand.

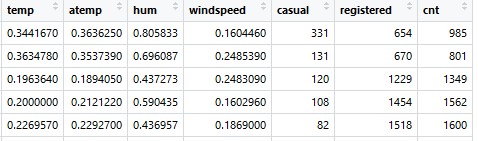
* 1. **Data**

Our goal is to build a regression model that will help in predicting the number of bikes used everyday based on seasonal and environmental factors. Given below is a sample of the data which we will use to build our model for predicting bike count.

Table 1.1: Bike Count sample data (Columns 1 through 9)



Table 1.2: Bike Count sample data (Columns 10 through 16)



As can be seen from the tables above, there are 13 variables using which we need to build our model for the predictions. The datatypes of these variables are as given below.

Table 1.3: Predictor variables

|  |  |
| --- | --- |
| S. No | Predictor Variables |
| 1 | Dteday |
| 2 | Season |
| 3 | Yr |
| 4 | Mnth |
| 5 | Holiday |
| 6 | Weekday |
| 7 | Workingday |
| 8 | Weathersit |
| 9 | Temp |
| 10 | aTemp |
| 11 | Hum |
| 12 | Windspeed |

**Chapter 2: Methodology**

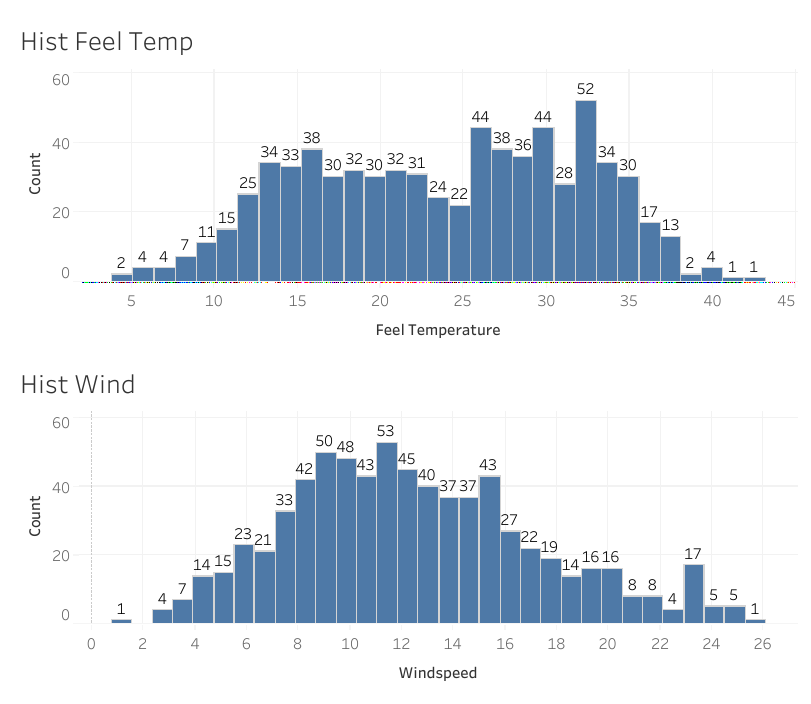
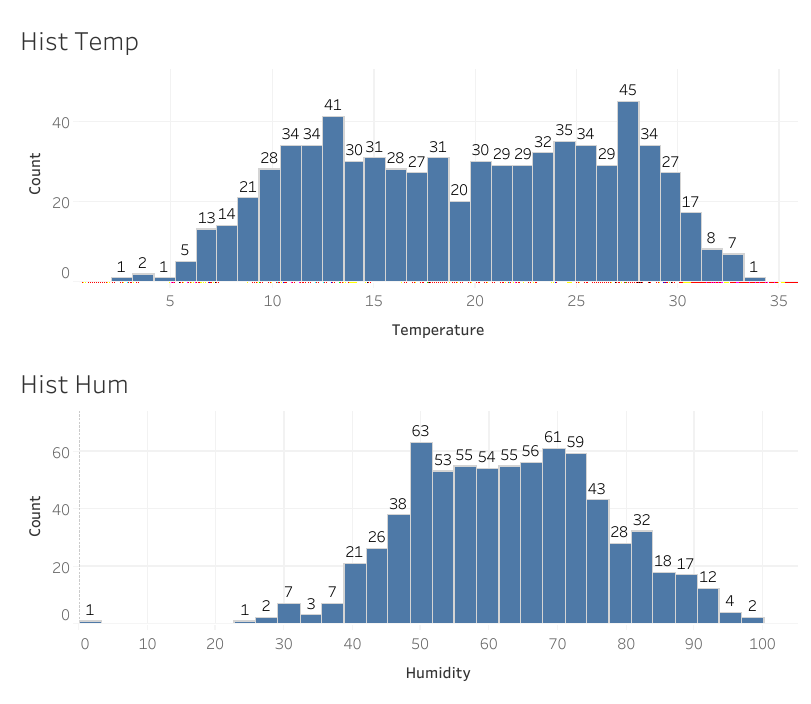
**2.1 Pre-Processing**

The first and foremost activity that is carried out in any data modelling is data pre- processing. This step is also the most important because we cannot feed the raw data directly to the model. The results would be erroneous. The process where data is made compatible to be fed to the model by removing all the unnecessary components while retaining the important deciding factors is called data pre-processing. We do this by visualizing data through graphs and plots and then cleaning it accordingly. This is also called exploratory data analysis.

**2.2 Distribution of continuous variables**

Most analysis like regression require the data to be normally distributed. So first we will have a look at the distribution of the numerical data through histograms. From the graphs below we can see that Temperature and Actual Temperature have normal distribution while humidity and Windspeed are slightly skewed. The skewness is likely because of the presence of outliers or extreme data points in those variables. These outliers should be removed to have a normal distribution for these variables as well.

Fig 2.1: Distribution of numerical variables using a histogram

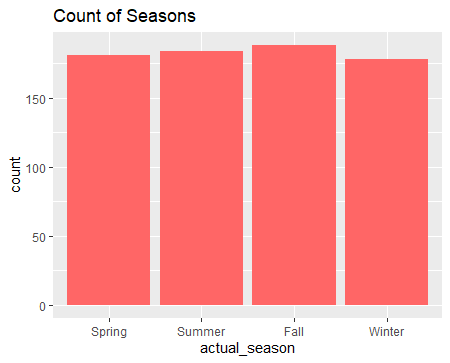


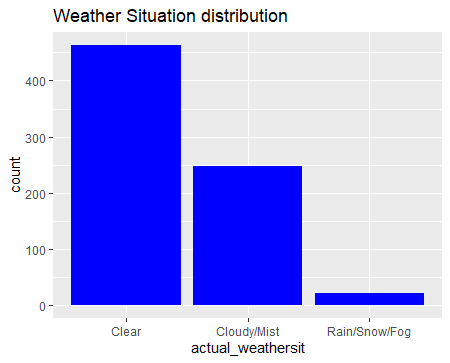
**2.3 Distribution of categorical variables**

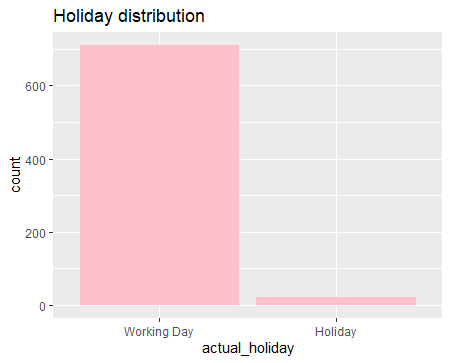
Numerical data was analysed based on histograms. Categorical data can be analysed using bar graphs or pie charts. For our case we will use bar graphs to see the distribution of the categorical variables.

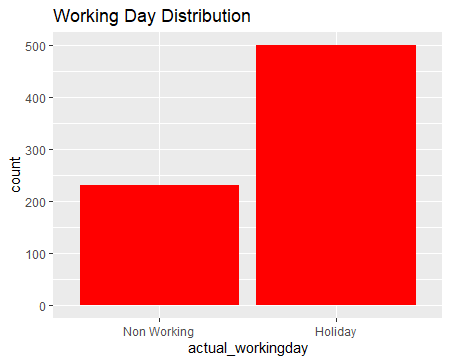
From the graphs we can see that season is evenly divided among the different levels, while weather, holiday and working day distribution seem to be more biased. It is only logical since a greater number of days are working day opposed to holidays. We can also conclude from that the weather has been clear for a major part of the year.

Fig 2.2: Distribution of categorical variables





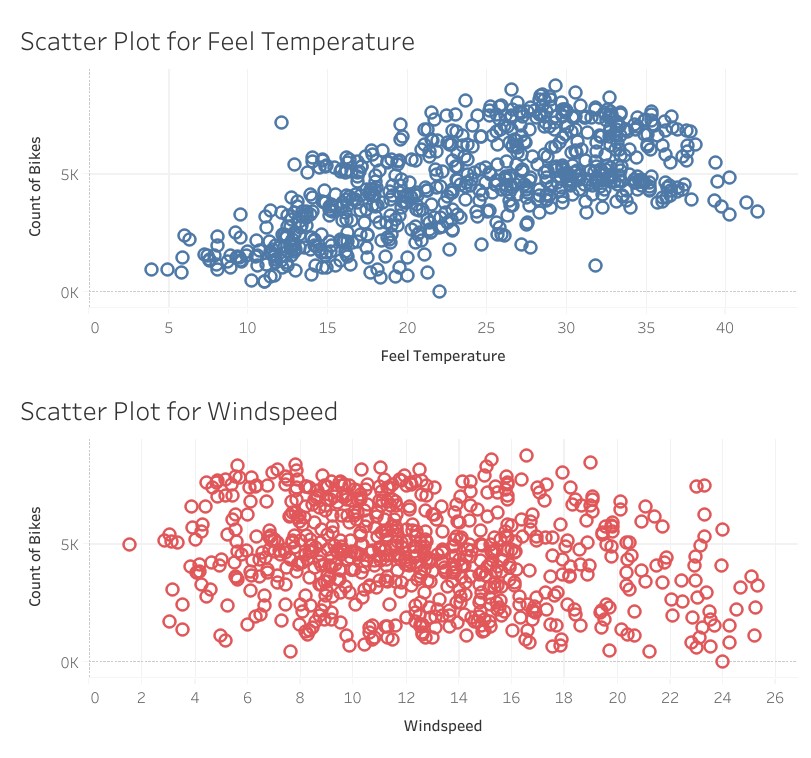
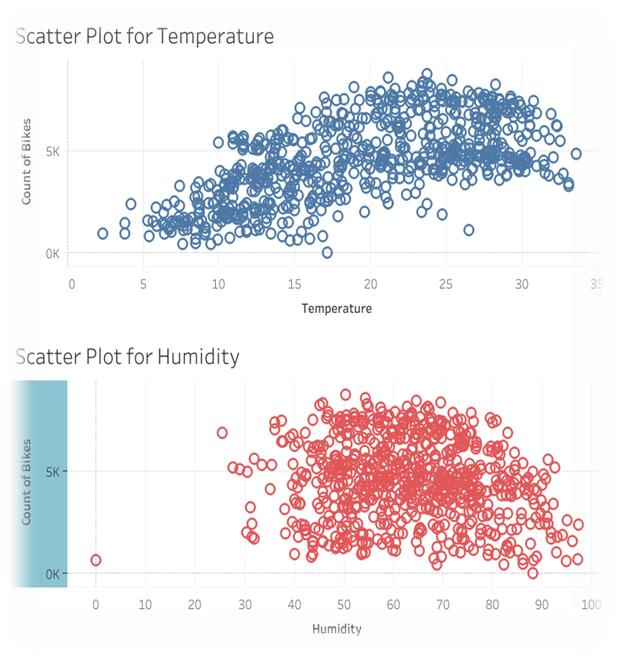




**2.4: Relationship of continuous variables against bike count**

The figure below shows the relationship of the target variable with the predictors using scatter plot. It can be observed from the graphs that there exists a linear positive relationship between the variables temperature and bike count and also actual feel temperature and bike count. However, the relationship is not purely linear. As the temperature increases above a certain level, there is a gradual drop in the bike count. There also exists a negative linear relationship between humidity and windspeed in regard to bike count. This is expected as people would prefer to bike when the weather is pleasant rather than hot, humid or windy.

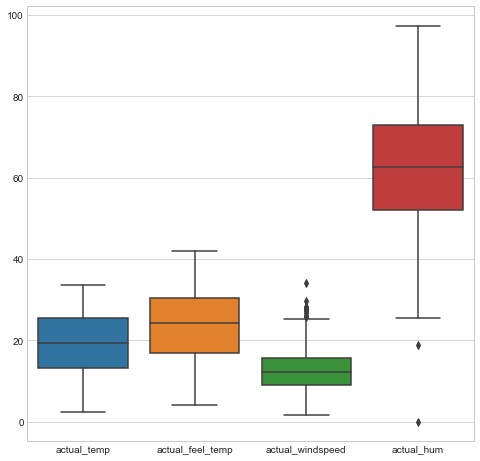
Fig 2.3: Scatter plots of continuous variables against bike count



**2.5: Outlier Analysis**

Outliers are extreme values that deviate from other observations in the mean. These outliers don’t contribute any special information to the model but rather affect it. Here we will check the outliers of the continuous variables and remove them.

Fig 2.4: Boxplot to check for outliers



As seen in the figure above, windspeed and humidity contain outliers. These outliers can be removed by using Boxplot stats method. Here, the inter quartile range(IQR) is obtained and minimum and maximum permissible values are calculated for the variable. All values lying outside this range are considered outliers and can be discarded.

IQR = q75 – q25, minimum = q25 – (1.5\*IQR), maximum = q75 + (1.5\*IQR)

The boxplot of above variables after removing outliers is shown below.

Fig 2.5: Boxplot after removal of outliers

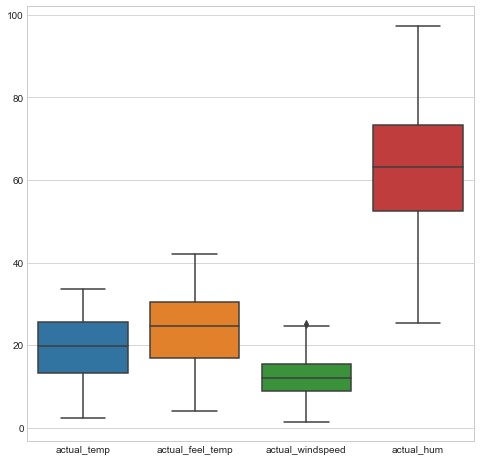
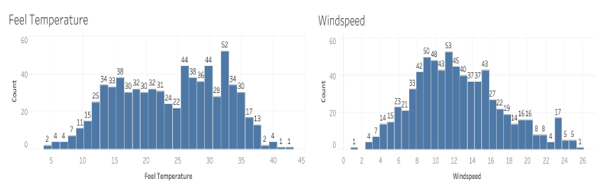


Fig 2.6: Histogram of variables after removal of outliers



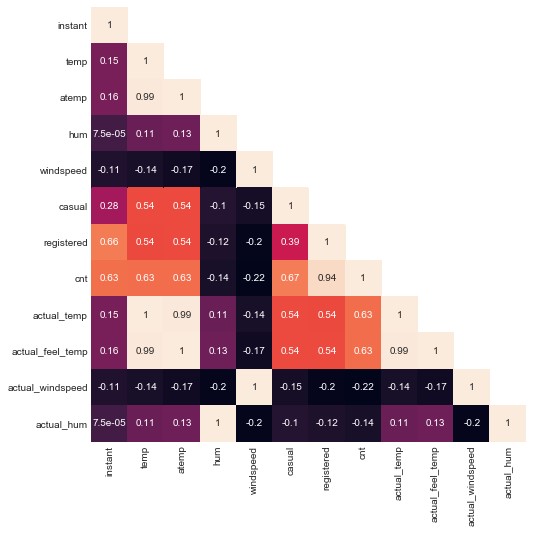
In comparison with fig 2.1 it can be seen that the skewness has been removed and the plot approaches normal distribution.

**2.6: Feature Selection**

Feature selection is the process of selecting a subset of relevant features for use in model construction. By selecting the relevant features we can remove the unwanted or redundant features that do not contribute to the accuracy of the model but rather hamper it. Fewer attributes are desirables because it reduces the complexity of the model, a simpler model is easier to understand and explain.

In order to determine the features to be dropped or retained we check for collinearity between the variables. The highly collinear variables are dropped, while the less collinear variables are retained. We achieve this using the correlation plot.

Fig 2.8: Correlation plot for all the variables



We can see that actual\_temp is highly correlated to temp. So, we can drop actual\_temp and just proceed with temp.

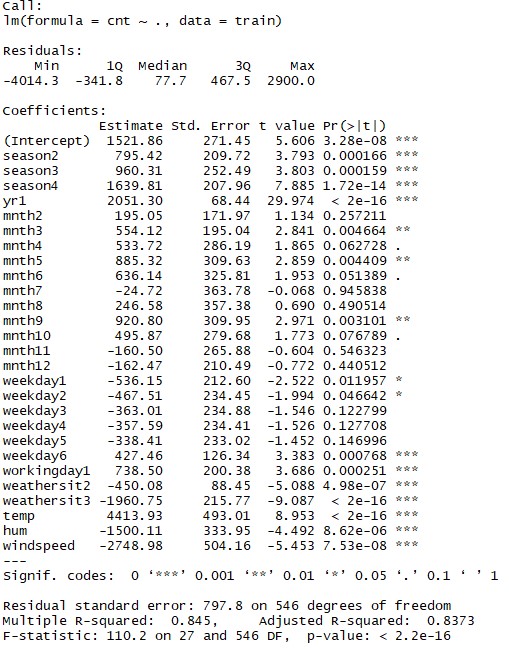
**Chapter 3: Modelling**

**3.1: Model Selection**

The dependent variable, i.e., count of bike rentals is continuous variable, therefore we use linear regression, decision tree and random forest models. The error metric chosen for the problem statement is Mean Absolute Error (MAE).

**3.2: Multiple Linear Regression**

Linear Regression is the model used when the target variable is numerical or continuous. Multiple linear regression is used when there are multiple predictors or independent variables involved in the prediction of the continuous target variable. The independent variables can be continuous or categorical.



As can be seen from the output of linear regression run on our data adjusted R-squared is 83.73% which is a measure of how close the data are to the fitted regression line.

Since the p-value is much lower than 0.05 we can safely reject the null hypothesis as well, which states that the target variable is not dependent on any if the predictor variables. This model explains the data well and is a good representation of it.

Even after removing the non-significant variables, the accuracy, Adjusted R-squared and Fstatistic do not change by much, hence the accuracy of this model is chosen to be final.

Mean Absolute Error (MAE) is calculated and found to be 494.

MAPE of this multiple linear regression model is 12.17%. Hence the accuracy of this model is 87.83%. This model performs very well for this test data.

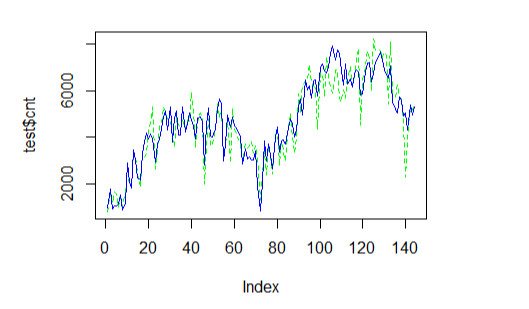
**3.3: Decision Tree**

Decision Tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

A decision tree is a flowchart-like structure in which each internal node represents a “test” on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

Using decision tree, we can predict the value of bike count. MAE for this model is 684. The MAPE for this decision tree is 17.47%. Hence the accuracy for this model is 82.53%.

Fig 3.1: Plot representing actual and predicted values. Actual(green) pred(blue)



**3.4: Random Forest**

Using Classification for prediction analysis in this case is not normal, though it can be done. The number of decision trees used for prediction in the forest is 500. MAE for this model is

392. Using random forest, the MAPE was found to be 10.68%. Hence the accuracy is 89.32%.

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=494**

**Decision Tree: MAE = 684.**

## Random Forest: MAE = 392

Based on the above error metrics, Random Forest is the better model for our analysis. Hence Random Forest is chosen as the model for prediction of bike rental count.

# Chapter 6: R code

####################EXPLORE USING GRAPHS######################

#CHECK THE DISTRIBUTION OF CATEGORICAL DATA USING BAR GRAPH

BAR1 = GGPLOT(DATA = DAY, AES(X = ACTUAL\_SEASON)) + GEOM\_BAR() + GGTITLE("COUNT OF SEASON")

BAR2 = GGPLOT(DATA = DAY, AES(X = ACTUAL\_WEATHERSIT)) + GEOM\_BAR() + GGTITLE("COUNT OF WEATHER")

BAR3 = GGPLOT(DATA = DAY, AES(X = ACTUAL\_HOLIDAY)) + GEOM\_BAR() + GGTITLE("COUNT OF HOLIDAY")

BAR4 = GGPLOT(DATA = DAY, AES(X = WORKINGDAY)) + GEOM\_BAR() + GGTITLE("COUNT OF WORKING DAY")

GRIDEXTRA::GRID.ARRANGE(BAR1,BAR2,BAR3,BAR4,NCOL=2)

#CHECK THE DISTRIBUTION OF NUMERICAL DATA USING HISTOGRAM

HIST1 = GGPLOT(DATA = DAY, AES(X =ACTUAL\_TEMP)) + GGTITLE("DISTRIBUTION OF TEMPERATURE") + GEOM\_HISTOGRAM(BINS = 25)

HIST2 = GGPLOT(DATA = DAY, AES(X =ACTUAL\_HUM)) + GGTITLE("DISTRIBUTION OF HUMIDITY") + GEOM\_HISTOGRAM(BINS = 25)

HIST3 = GGPLOT(DATA = DAY, AES(X =ACTUAL\_FEEL\_TEMP)) + GGTITLE("DISTRIBUTION OF FEEL

TEMPERATURE") + GEOM\_HISTOGRAM(BINS = 25)

HIST4 = GGPLOT(DATA = DAY, AES(X =ACTUAL\_WINDSPEED)) + GGTITLE("DISTRIBUTION OF WINDSPEED") +

GEOM\_HISTOGRAM(BINS = 25)

GRIDEXTRA::GRID.ARRANGE(HIST1,HIST2,HIST3,HIST4,NCOL=2)

#CHECK THE DISTRIBUTION OF NUMERICAL DATA USING SCATTERPLOT

SCAT1 = GGPLOT(DATA = DAY, AES(X =ACTUAL\_TEMP, Y = CNT)) + GGTITLE("DISTRIBUTION OF

TEMPERATURE") + GEOM\_POINT() + XLAB("TEMPERATURE") + YLAB("BIKE COUNT")

SCAT2 = GGPLOT(DATA = DAY, AES(X =ACTUAL\_HUM, Y = CNT)) + GGTITLE("DISTRIBUTION OF HUMIDITY") +

GEOM\_POINT(COLOR="RED") + XLAB("HUMIDITY") + YLAB("BIKE COUNT")

SCAT3 = GGPLOT(DATA = DAY, AES(X =ACTUAL\_FEEL\_TEMP, Y = CNT)) + GGTITLE("DISTRIBUTION OF FEEL

TEMPERATURE") + GEOM\_POINT() + XLAB("FEEL TEMPERATURE") + YLAB("BIKE COUNT") SCAT4 = GGPLOT(DATA = DAY, AES(X =ACTUAL\_WINDSPEED, Y = CNT)) + GGTITLE("DISTRIBUTION OF

WINDSPEED") + GEOM\_POINT(COLOR="RED") + XLAB("WINDSPEED") + YLAB("BIKE COUNT") GRIDEXTRA::GRID.ARRANGE(SCAT1,SCAT2,SCAT3,SCAT4,NCOL=2)

#CHECK FOR OUTLIERS IN DATA USING BOXPLOT

CNAMES =

COLNAMES(DAY[,C("ACTUAL\_TEMP","ACTUAL\_FEEL\_TEMP","ACTUAL\_WINDSPEED","ACTUAL\_HUM")]) FOR (I IN 1:LENGTH(CNAMES))

{

ASSIGN(PASTE0("GN",I), GGPLOT(AES\_STRING(Y = CNAMES[I]), DATA = DAY)+ STAT\_BOXPLOT(GEOM =

"ERRORBAR", WIDTH = 0.5) + GEOM\_BOXPLOT(OUTLIER.COLOUR="RED", FILL = "GREY" ,OUTLIER.SHAPE=18, OUTLIER.SIZE=1, NOTCH=FALSE) + THEME(LEGEND.POSITION="BOTTOM")+ LABS(Y=CNAMES[I]) + GGTITLE(PASTE("BOX PLOT FOR",CNAMES[I])))

}

GRIDEXTRA::GRID.ARRANGE(GN1,GN3,GN2,GN4,NCOL=2)

#REMOVE OUTLIERS IN WINDSPEED

VAL = DAY[,19][DAY[,19] %IN% BOXPLOT.STATS(DAY[,19])$OUT] DAY = DAY[WHICH(!DAY[,19] %IN% VAL),]

#CHECK FOR MULTICOLLINEARITY USING VIF

DF = DAY[,C("INSTANT","TEMP","ATEMP","HUM","WINDSPEED")] VIFCOR(DF)

#CHECK FOR COLLINEARITY USING CORELATION GRAPH

CORRGRAM(DAY, ORDER = F, UPPER.PANEL=PANEL.PIE, TEXT.PANEL=PANEL.TXT, MAIN = "CORRELATION PLOT")

#REMOVE THE UNWANTED VARIABLES

DAY <- SUBSET(DAY, SELECT = -

C(INSTANT,DTEDAY,ATEMP,CASUAL,REGISTERED,ACTUAL\_TEMP,ACTUAL\_FEEL\_TEMP,ACTUAL\_WINDSPEED,AC

TUAL\_HUM,ACTUAL\_SEASON,ACTUAL\_YR,ACTUAL\_HOLIDAY,ACTUAL\_WEATHERSIT))

#########################DECISION TREE#########################

#DIVIDE THE DATA INTO TRAIN AND TEST

SET.SEED(123)

TRAIN\_INDEX = SAMPLE(1:NROW(DAY), 0.8 \* NROW(DAY))

TRAIN = DAY[TRAIN\_INDEX,] TEST = DAY[-TRAIN\_INDEX,]

#RPART FOR REGRESSION

DT\_MODEL = RPART(CNT ~ ., DATA = TRAIN, METHOD = "ANOVA")

#PREDICT THE TEST CASES

DT\_PREDICTIONS = PREDICT(DT\_MODEL, TEST[,-11])

#CREATE DATAFRAME FOR ACTUAL AND PREDICTED VALUES

DF = DATA.FRAME("ACTUAL"=TEST[,11], "PRED"=DT\_PREDICTIONS) HEAD(DF)

#CALCULATE MAPE

REGR.EVAL(TRUES = TEST[,11], PREDS = DT\_PREDICTIONS, STATS = C("MAE","MSE","RMSE","MAPE"))

####################RANDOM FOREST################

#TRAIN THE DATA USING RANDOM FOREST

RF\_MODEL = RANDOMFOREST(CNT~., DATA = TRAIN, NTREE = 500)

#PREDICT THE TEST CASES

RF\_PREDICTIONS = PREDICT(RF\_MODEL, TEST[,-11])

#CREATE DATAFRAME FOR ACTUAL AND PREDICTED VALUES

DF = CBIND(DF,RF\_PREDICTIONS) HEAD(DF)

#CALCULATE MAPE

REGR.EVAL(TRUES = TEST[,11], PREDS = RF\_PREDICTIONS, STATS = C("MAE","MSE","RMSE","MAPE"))

#####################LINEAR REGRESSION################

#TRAIN THE DATA USING LINEAR REGRESSION

LR\_MODEL = LM(FORMULA = CNT~., DATA = TRAIN)

#CHECK THE SUMMARY OF THE MODEL

SUMMARY(LR\_MODEL)

#PREDICT THE TEST CASES

LR\_PREDICTIONS = PREDICT(LR\_MODEL, TEST[,-11])

#CREATE DATAFRAME FOR ACTUAL AND PREDICTED VALUES

DF = CBIND(DF,LR\_PREDICTIONS) HEAD(DF)

#CALCULATE MAPE

REGR.EVAL(TRUES = TEST[,11], PREDS = LR\_PREDICTIONS, STATS = C("MAE","MSE","RMSE","MAPE"))

#PREDICT A SAMPLE DATA

PREDICT(LR\_MODEL,TEST[2,])