**Bike Rental Prediction**

Description

**Problem Statement:**

In bike-sharing systems, the entire process from membership to rental and return has been automated. Using these systems, users can easily rent a bike from one location and return it to another. Hence, a bike rental company wants to understand and predict the number of bikes rented daily based on the environment and seasons.

**Objective:**The objective of this case is to predict bike rental counts based on environmental and seasonal settings with the help of a machine learning algorithm.

**Data Set:**day.csv

**Data Description**

|  |  |
| --- | --- |
| **Variable** | **Description** |
| 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) |
| holiday | Weather day is a holiday or not |
| weekday | Day of the week |
| workingday | Working day (1: neither weekend nor holiday, 0: other days) |
| weathersit | 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 |
| temp | Normalized temperature in Celsius; The values are divided into 41 (max) |
| atemp | Normalized feeling temperature in Celsius; The values are divided into 50 (max) |
| hum | Normalized humidity; The values are divided into 100 (max) |
| windspeed | Normalized wind speed; The values are divided into 67 (max) |
| casual | Count of casual users |
| registered | Count of registered users |
| cnt | Count of total rental bikes including both casual and registered |

**Steps to Perform:**

1.    Exploratory data analysis  
•    Load dataset and libraries  
•    Perform data type conversion of the attributes  
•    Carry out the missing value analysis  
2. Attributes distributions and trends  
•    Plot monthly distribution of the total number of bikes rented  
•    Plot yearly distribution of the total number of bikes rented  
•    Plot boxplot for outliers analysis  
3. Split the dataset into train and test dataset  
4. Create a model using the random forest algorithm  
5. Predict the performance of the model on the test dataset

library(readxl)

data <- read\_excel(file.choose())

View(data)

# To get the data type of each column

sapply(data,class)

# for data type conversion

data['mnth'] <- as.factor(data$mnth)

data['yr'] <- as.factor(data$yr)

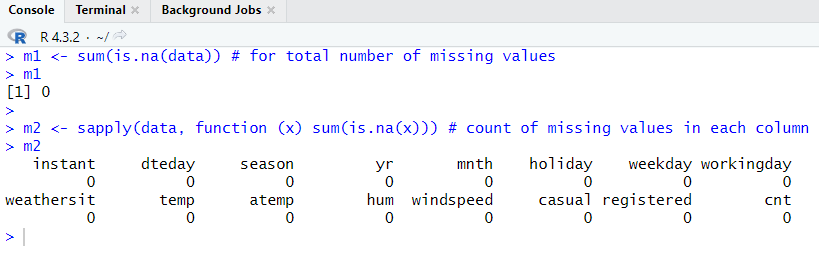
# To find missing values

m1 <- sum(is.na(data)) # for total number of missing values

m1

m2 <- sapply(data, function (x) sum(is.na(x))) # count of missing values in each column

m2



**# 1 Plot monthly distribution of the total number of bikes rented**

library(ggplot2)

library(RColorBrewer)

library(dplyr)

count <- data %>%

group\_by(mnth) %>%

summarise(Total\_bikes = sum(cnt))

count

c1 <- ggplot(data, aes(x =mnth, y = cnt)) +

geom\_bar(stat = "identity", fill = 'black') +

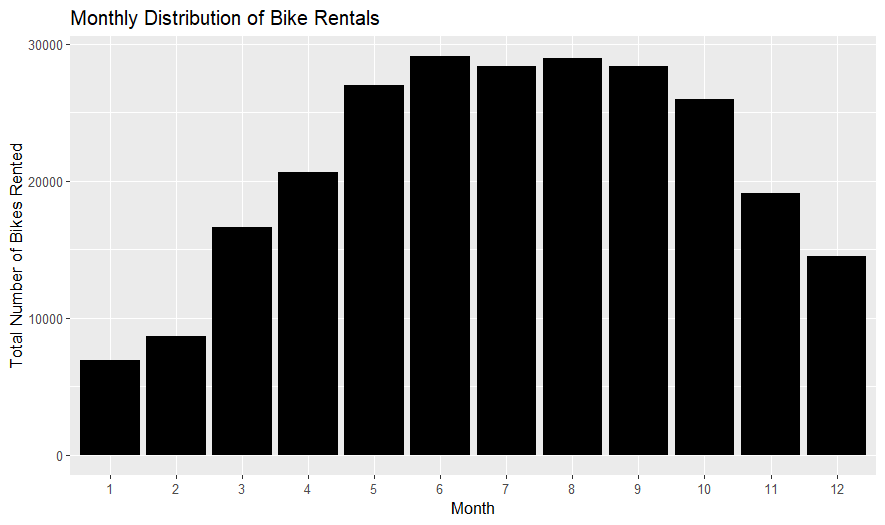
labs(title = "Monthly Distribution of Bike Rentals",

x = "Month",

y = "Total Number of Bikes Rented")+

scale\_y\_continuous(labels = function(x) format(x, scientific = F))

c1



**# 2 Plot yearly distribution of the total number of bikes rented**

yr\_count <- data %>%

group\_by(yr) %>%

summarise(Total\_bikes = sum(cnt))

yr\_count

c2 <- ggplot(data, aes(x =yr, y = cnt)) +

geom\_bar(stat = "identity", fill = 'skyblue') +

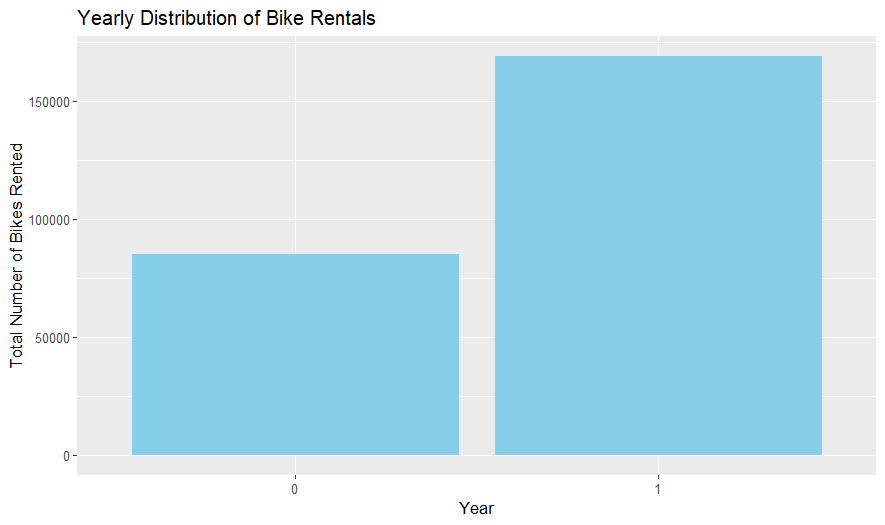
labs(title = "Yearly Distribution of Bike Rentals",

x = "Year",

y = "Total Number of Bikes Rented")+

scale\_y\_continuous(labels = function(x) format(x, scientific = F))

c2



**# 3 Plot boxplot for outliers' analysis**

# Monthly Boxplot for outliers' analysis

c3 <- ggplot(data, aes(x = mnth, y = cnt)) +

geom\_boxplot(fill = 'red', color = 'black') +

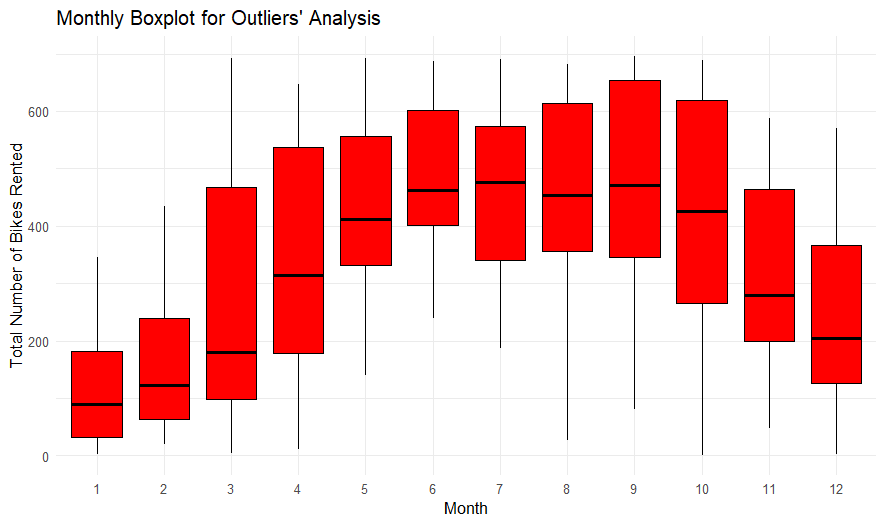
labs(title = "Monthly Boxplot for Outliers' Analysis",

x = "Month",

y = "Total Number of Bikes Rented")+

theme\_minimal()

c3



**# Yearly Boxplot for outliers' analysis**

c4 <- ggplot(data, aes(x =yr, y = cnt))+

geom\_boxplot(fill = 'skyblue', color = 'darkblue')+

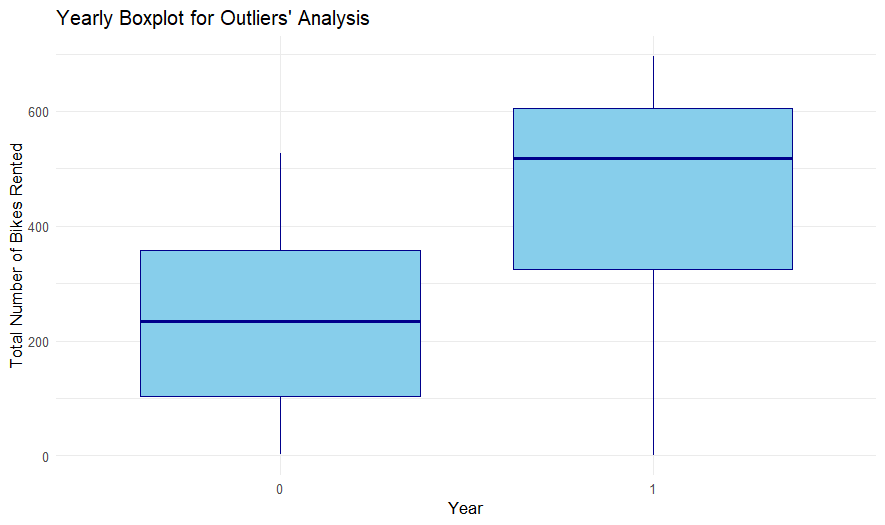
labs(title = "Yearly Boxplot for Outliers' Analysis",

x = "Year",

y = "Total Number of Bikes Rented") +

theme\_minimal()

c4



**# 3. Split the dataset into train and test dataset**

data$cnt <- as.factor(data$cnt)

set.seed(1234)

train\_ind <- sample(1:nrow(data), nrow(data)\*0.75)

train\_dt <- data[train\_ind,]

test\_dt <- data[-train\_ind,]

View(train\_dt)

View(test\_dt)

**# 4. Create a model using the random forest algorithm**

sum(is.na(test\_dt$cnt))

train\_dt$cnt <- as.numeric(train\_dt$cnt)

test\_dt$cnt <- as.numeric(test\_dt$cnt)

library(randomForest)

rf <- randomForest(cnt ~ .,

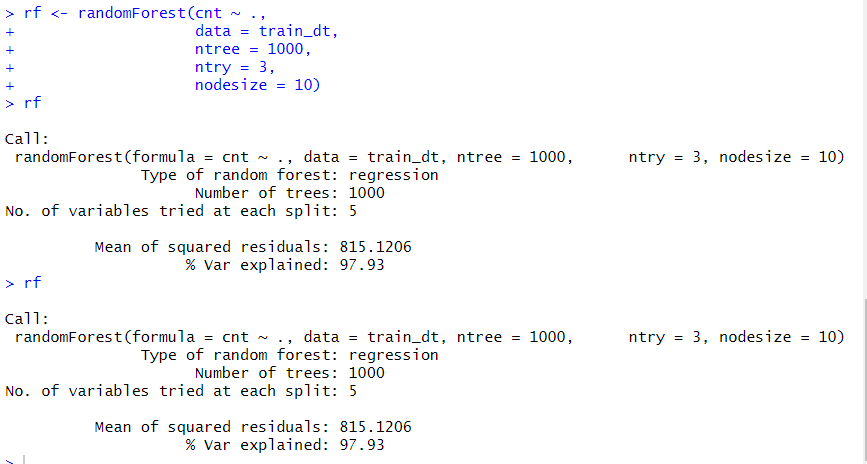
data = train\_dt,

ntree = 1000,

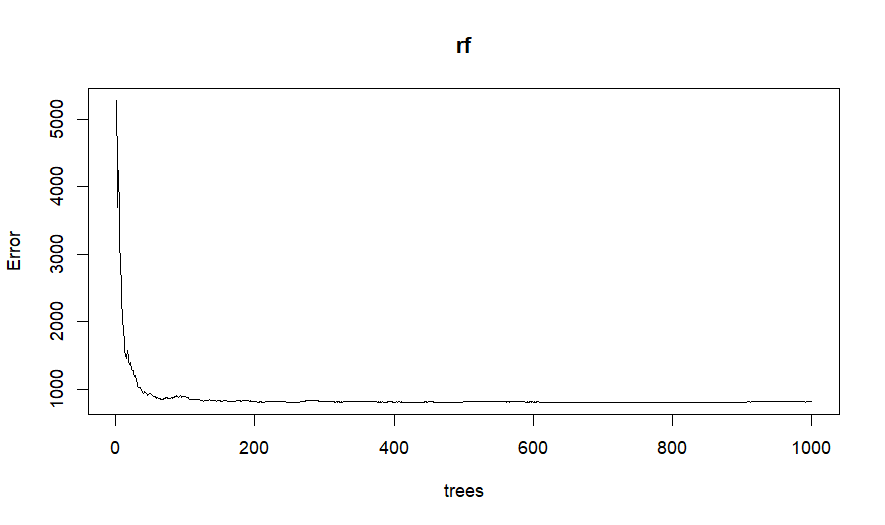
ntry = 3,

nodesize = 10)

rf



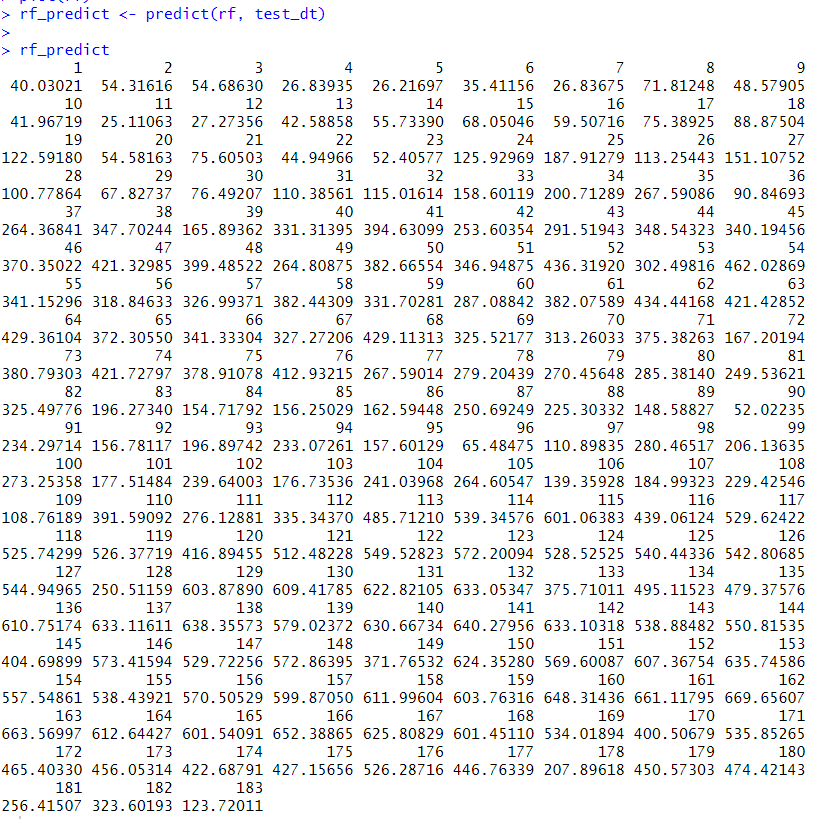
plot(rf)



**# 5. Predict the performance of the model on the test dataset**

rf\_predict <- predict(rf, test\_dt)

rf\_predict



The below File is the code:

