BIKE RENTAL

SAMEER PANDEY

APRIL 05, 2019

Contents

“Many many thanks to EDWISOR for teaching me and making me able to do projects in data science”

Chapter 1: Introduction

1.1 Problem Statement

The objective of this Case is to Predication of bike rental count on daily based on the environmental and seasonal settings.

1.2 Variables

There are 731 Observations and 16 variables in our data in which 13 are independent variables and 3 (Casual, Registered and Count) are dependent variables. Since the type of target variable is continuous, this is a regression problem.

Variable Information:

1. **instant**: Record index
2. **dteday**: Date
3. **season**: Season (1:springer, 2:summer, 3:fall, 4:winter)
4. **yr**: Year (0: 2011, 1:2012)
5. **mnth**: Month (1 to 12)
6. **holiday**: weather day is holiday or not (extracted from Holiday Schedule)
7. **weekday**: Day of the week
8. **workingday**: If day is neither weekend nor holiday is 1, otherwise is 0.
9. **weathersit**: (extracted from Freemeteo) **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
10. **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)
11. **atemp**: Normalized feeling temperature in Celsius. The values are derived via

    (t-t_min)/(t_maxt_min), t_min=-16, t_max=+50 (only in hourly scale)
12. **hum**: Normalized humidity. The values are divided to 100 (max)
13. **windspeed**: Normalized wind speed. The values are divided to 67 (max)
14. **casual**: count of casual users
15. **registered**: count of registered users
16. **cnt**: count of total rental bikes including both casual and registered

1.3 Sample Data

instant	dteday	season	yr	mnth	holiday	weekday
1	01-01-2011	1	0	1	0	6
2	02-01-2011	1	0	1	0	0
3	03-01-2011	1	0	1	0	1
4	04-01-2011	1	0	1	0	2
5	05-01-2011	1	0	1	0	3
6	06-01-2011	1	0	1	0	4

Table 1·1: Bike Rental Sample Data (Columns: 1-7)

workingday	weathersit	temp	atemp	hum	windspeed
0	2	0.344167	0.363625	0.805833	0.160446
0	2	0.363478	0.353739	0.696087	0.248539
1	1	0.196364	0.189405	0.437273	0.248309
1	1	0.2	0.212122	0.590435	0.160296
1	1	0.226957	0.22927	0.436957	0.1869
1	1	0.204348	0.233209	0.518261	0.0895652

Table 1·2: Bike Rental Sample Data (Columns: 8-13)

casual	registered	cnt
331	654	985
131	670	801
120	1229	1349
108	1454	1562
82	1518	1600
88	1518	1606

Table 1·3: Bike Rental Sample Data (Columns: 14-16)

1.4 Unique count

Below figure shows the unique count of all the variables present in the data.

List of columns and their number of unique values -

```
instant       731
dteday        731
season          4
yr              2
mnth           12
holiday         2
weekday         7
workingday      2
weathersit      3
temp          499
atemp         690
```

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. In this project we look at the distribution of categorical variables and continuous variables. We also look at the missing values in the data and the outliers present in the data.

Remember the quality of our inputs decide the quality of your output. So, once we have got our business hypothesis ready, it makes sense to spend lot of time and efforts here. Data exploration, cleaning and preparation can take up to 70% of our total project time.

Below are the steps involved to understand, clean and prepare our data for building model:

1. Variable Identification
2. Univariate Analysis
3. Bi-variate Analysis
4. Missing values treatment
5. Outlier treatment
6. Feature Selection
7. Feature scaling

2.2 Variable Identification

From EDA we have concluded that there are 7 continuous variables (Including Target) and 7 categorical variables and one continuous target variable.

Target Variable = Casual, registered and cnt

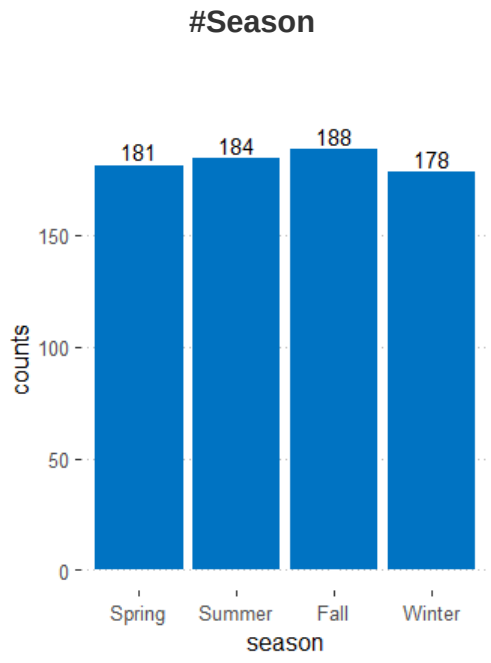
Continuous variables	Categorical variables
temp	season
atemp	yr
hum	mnth
windspeed	holiday
casual	weekday
registered	workingday
count	weathersit

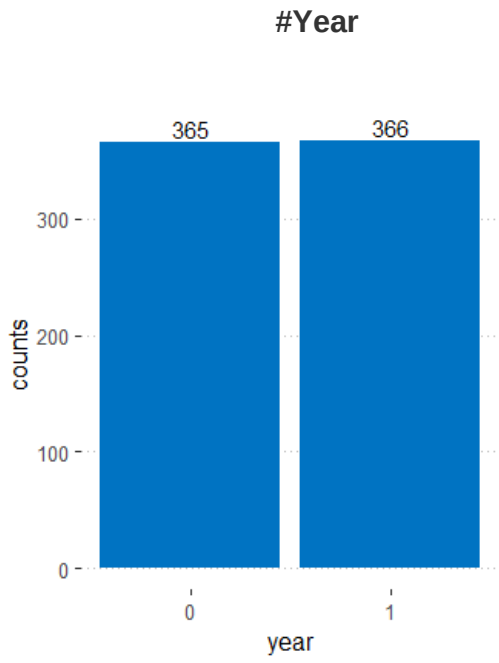
Table 1·4: Employee Absenteeism Variable Category

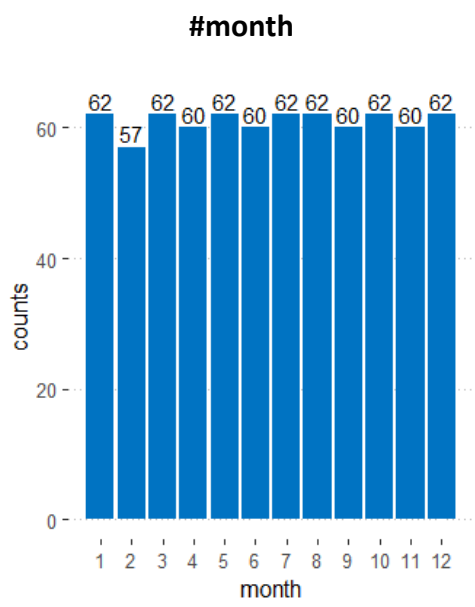
2.3 Univariate Analysis

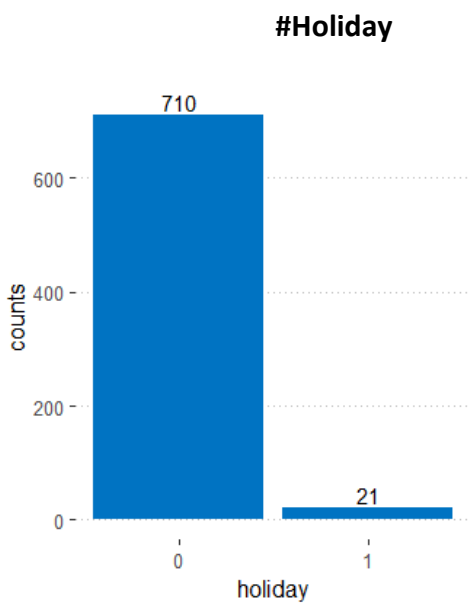
At this stage, we explore variables one by one. Method to perform uni-variate analysis will depend on whether the variable type is categorical or continuous. Let's look at these methods and statistical measures for categorical and continuous variables individually:

2.3.1 Categorical Variables

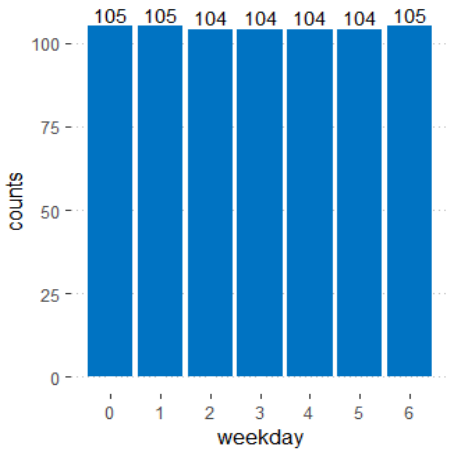




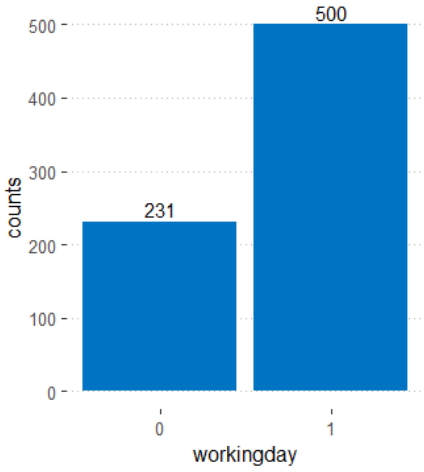




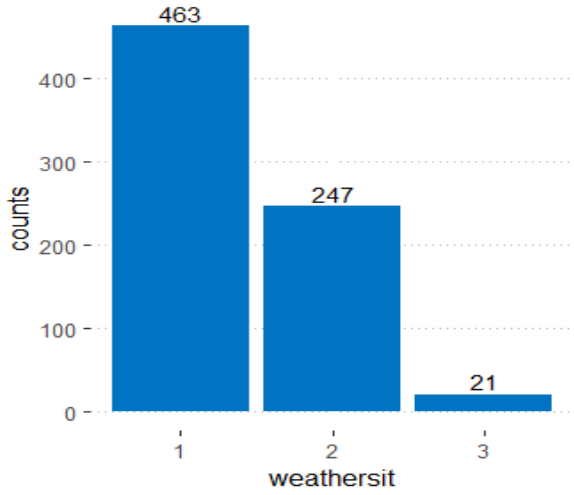
#Weekday



#Working Day

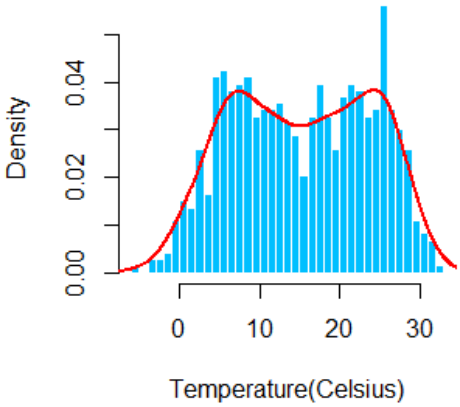


#Weather Situation

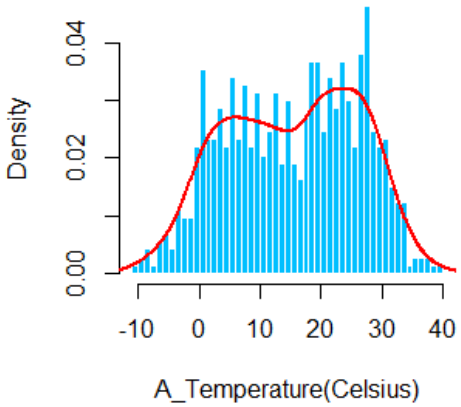


2.3.2 Continuous Variables

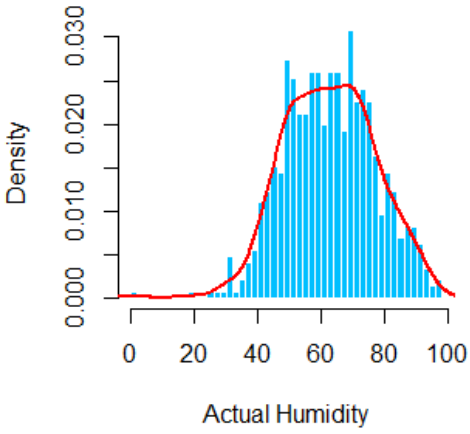
#Temp



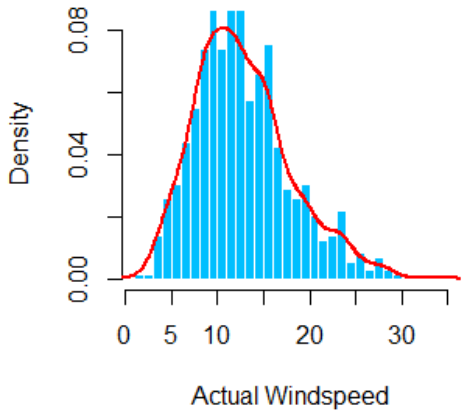
#atemp



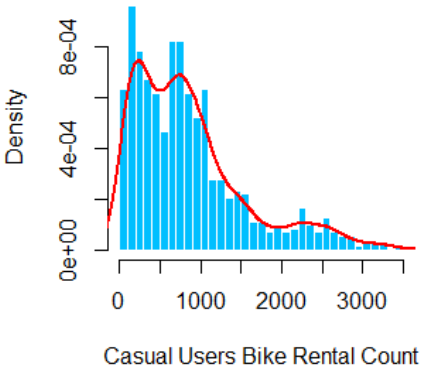
#hum



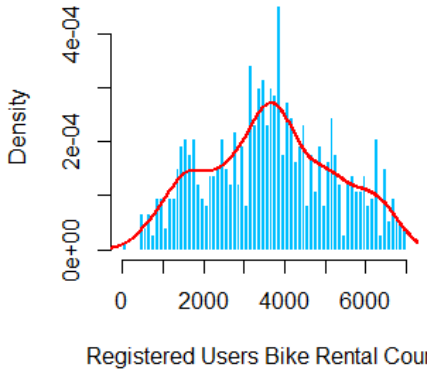
#windspeed



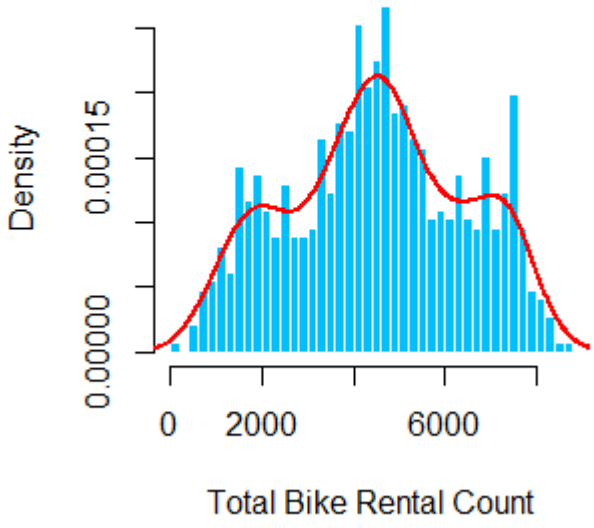
#casual



#registered

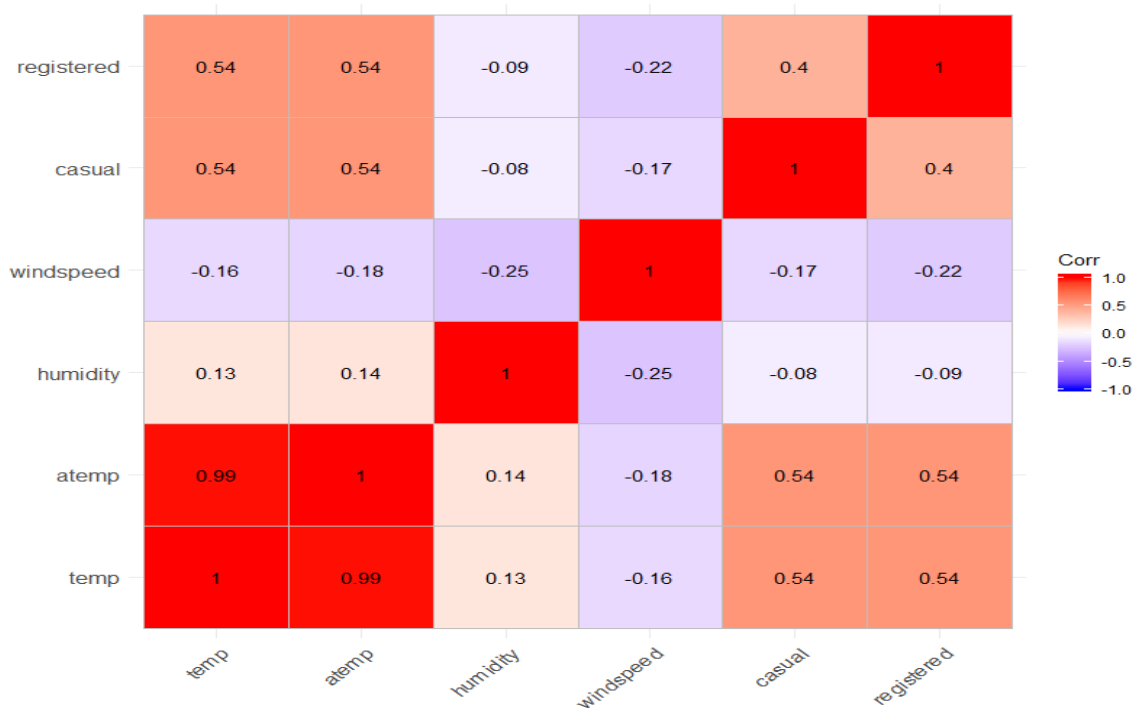


#cnt



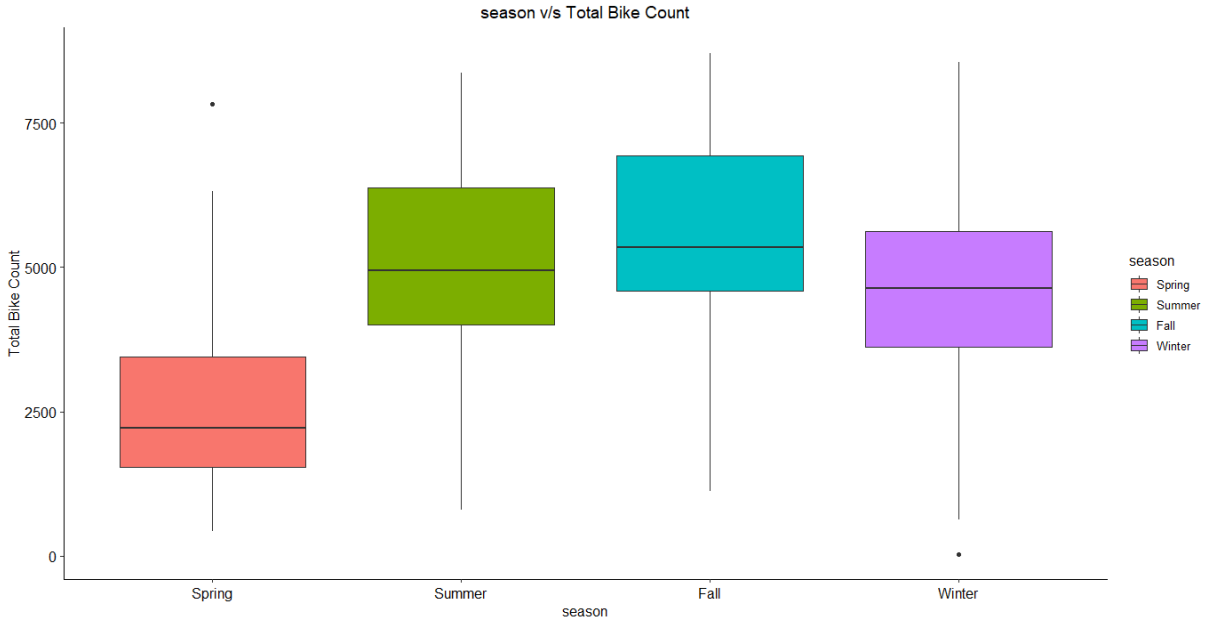
2.4 Bi-variate Analysis

2.4.1 Correlation Plot of Continuous Variables

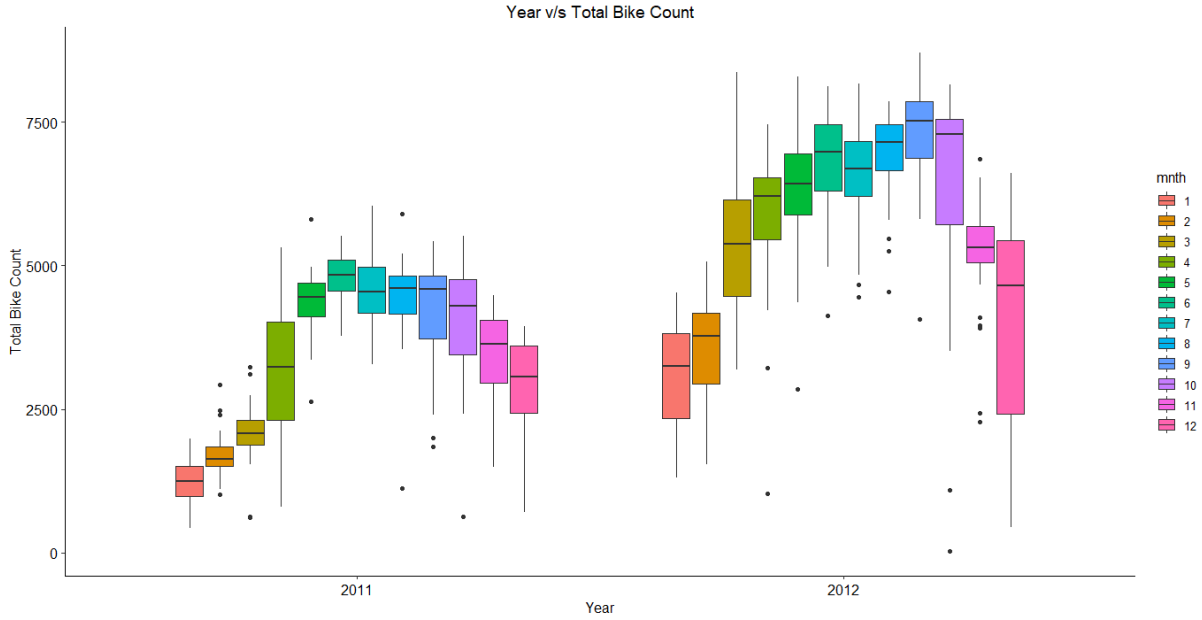


2.4.2 Box Plot B/W Categorical Predictor and Target Variable

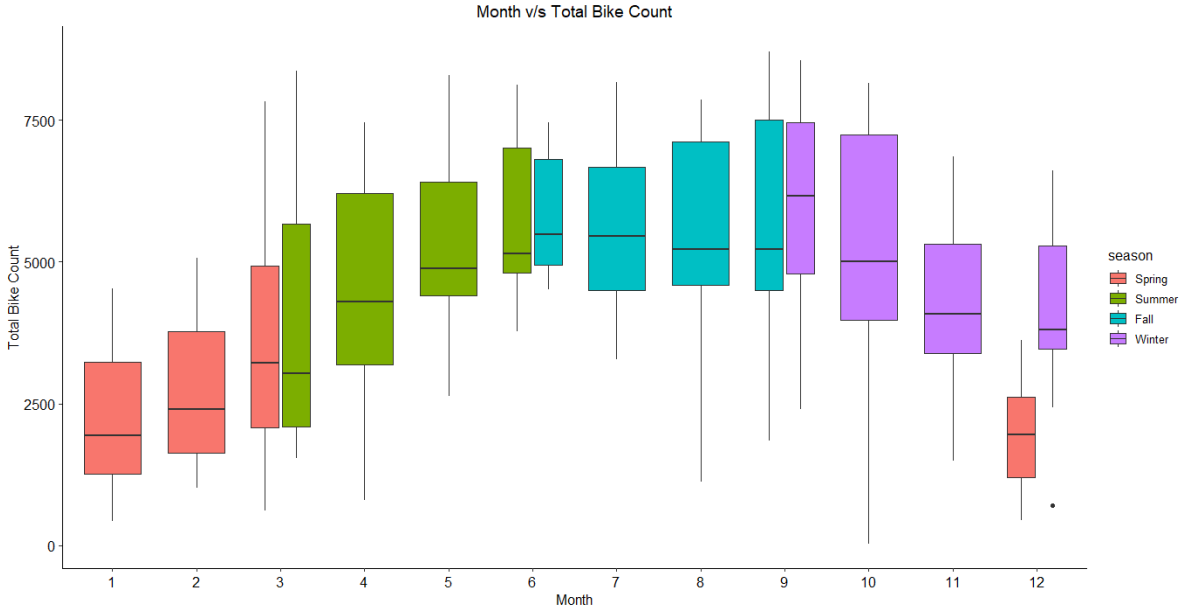
#Season v/s Total Bike rental count



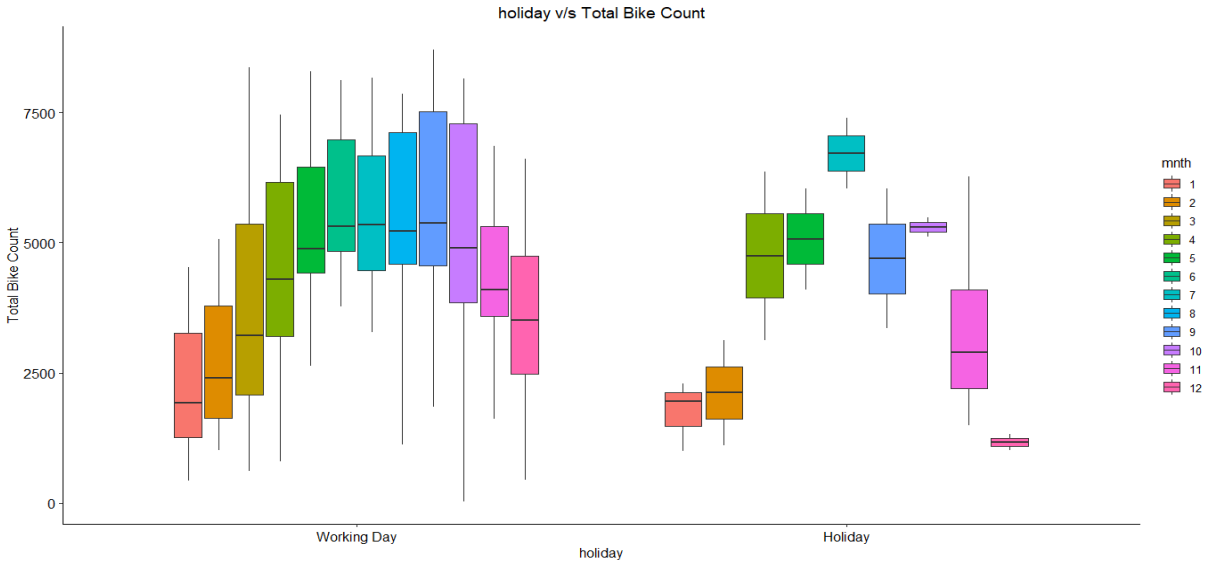
#Year v/s Total Bike rental count



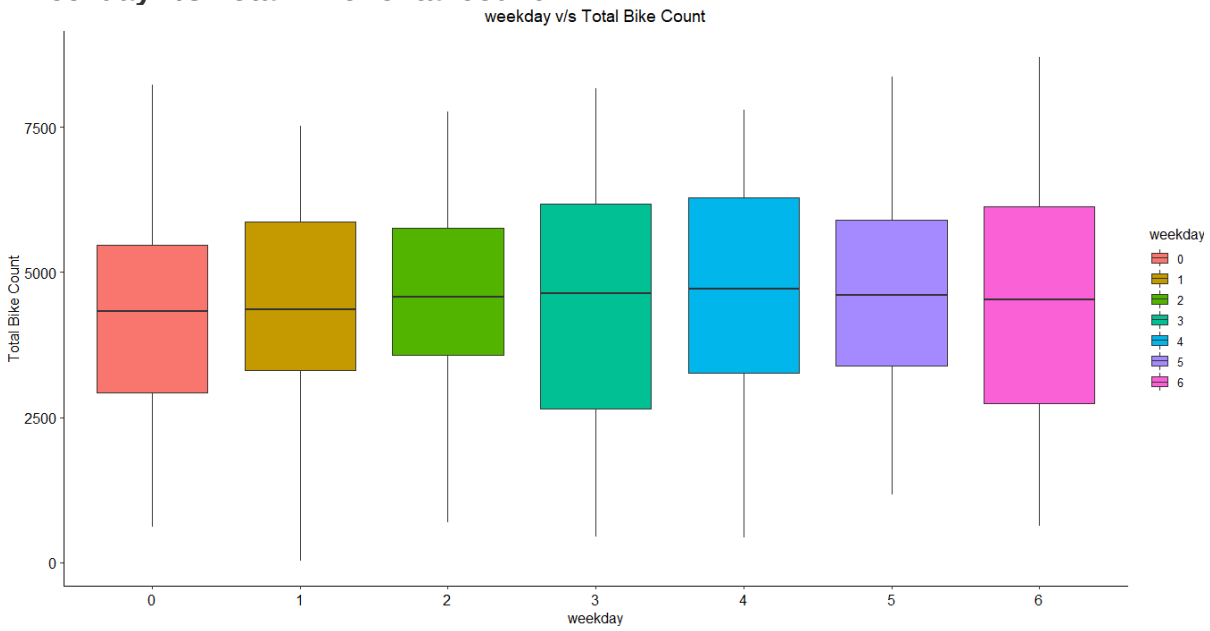
#Month v/s Total Bike rental count



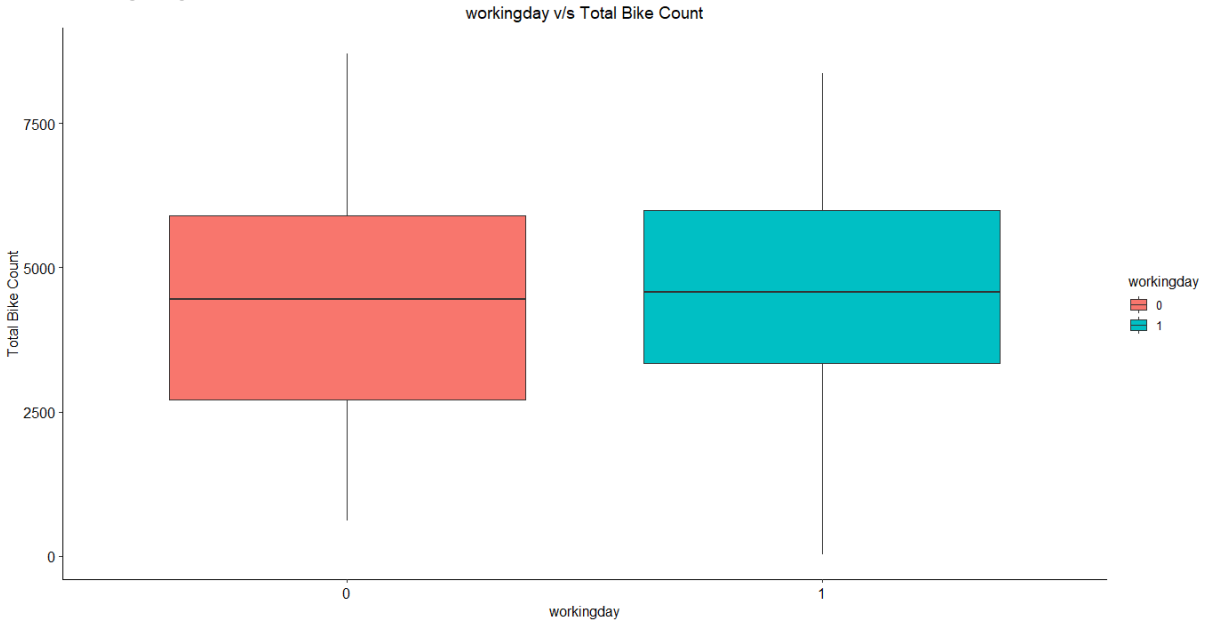
#Holiday v/s Total Bike rental count



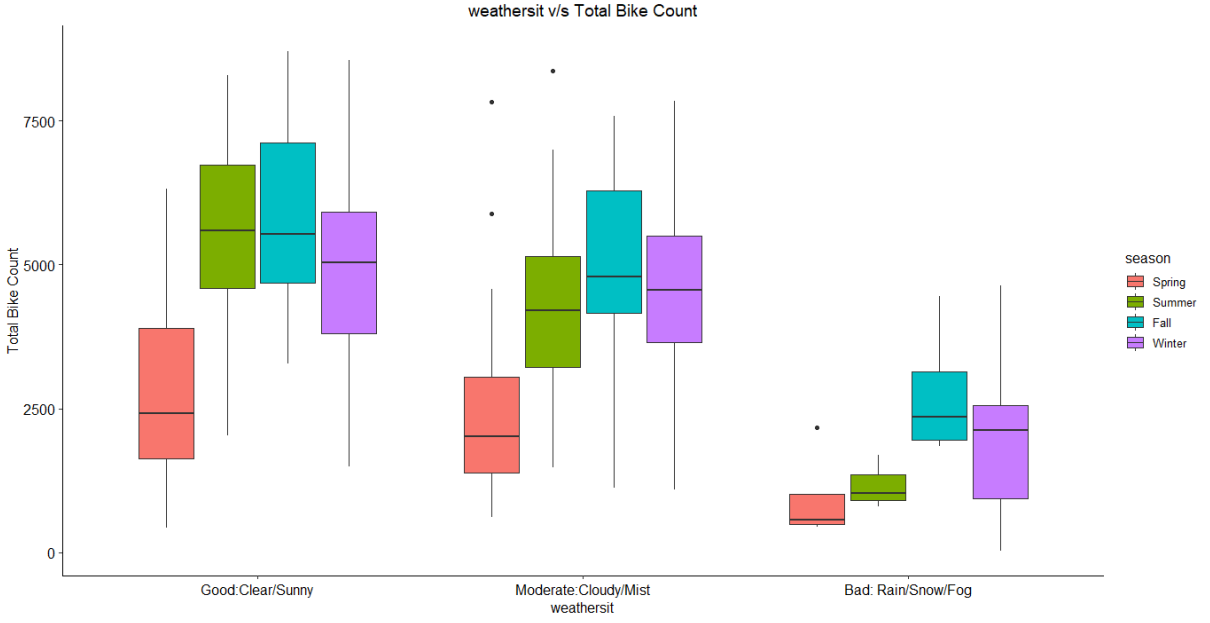
#Weekday v/s Total Bike rental count



#Workingday v/s Total Bike rental count

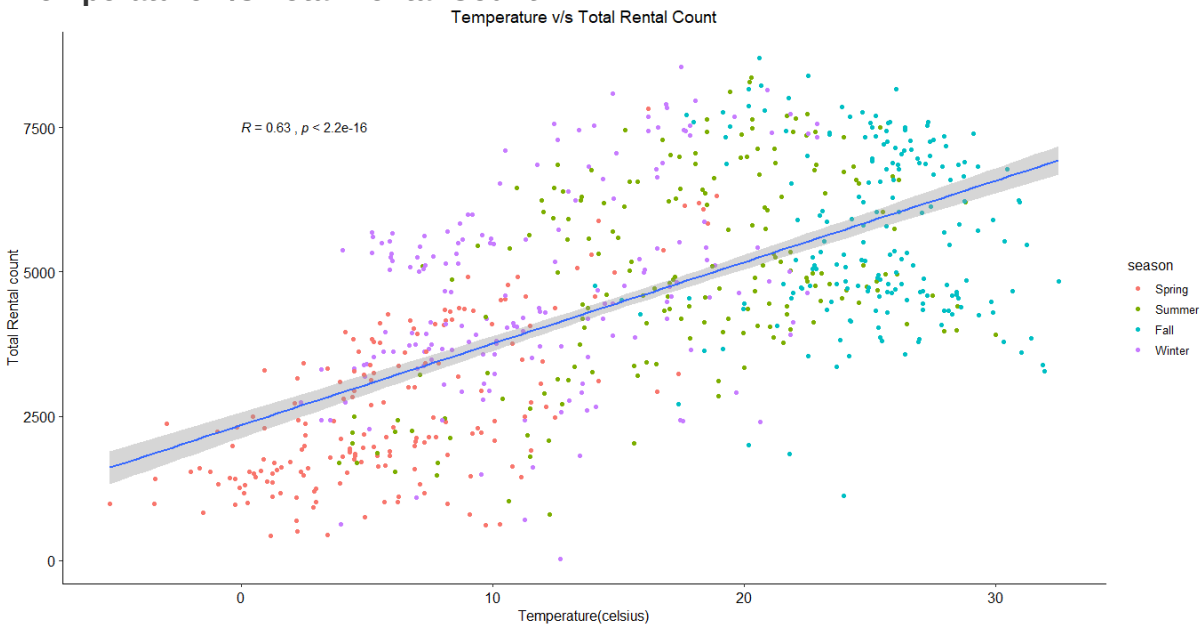


#Weathersit v/s Total Bike rental count

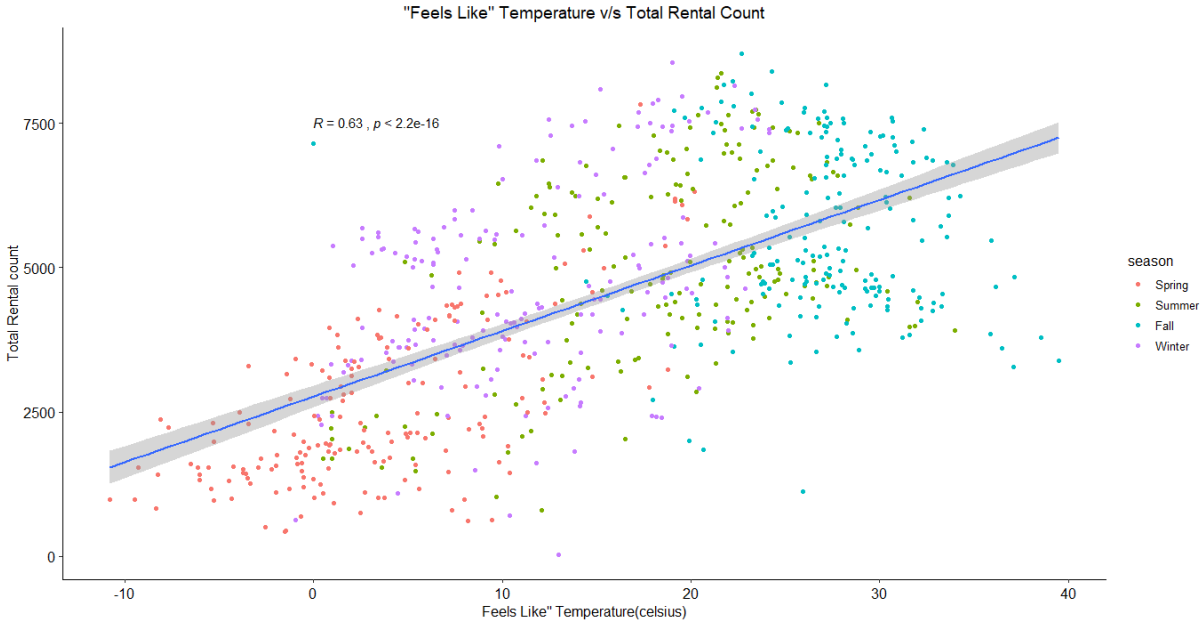


2.4.3 Scatter Plot B/W Continuous Predictor and Target Variable

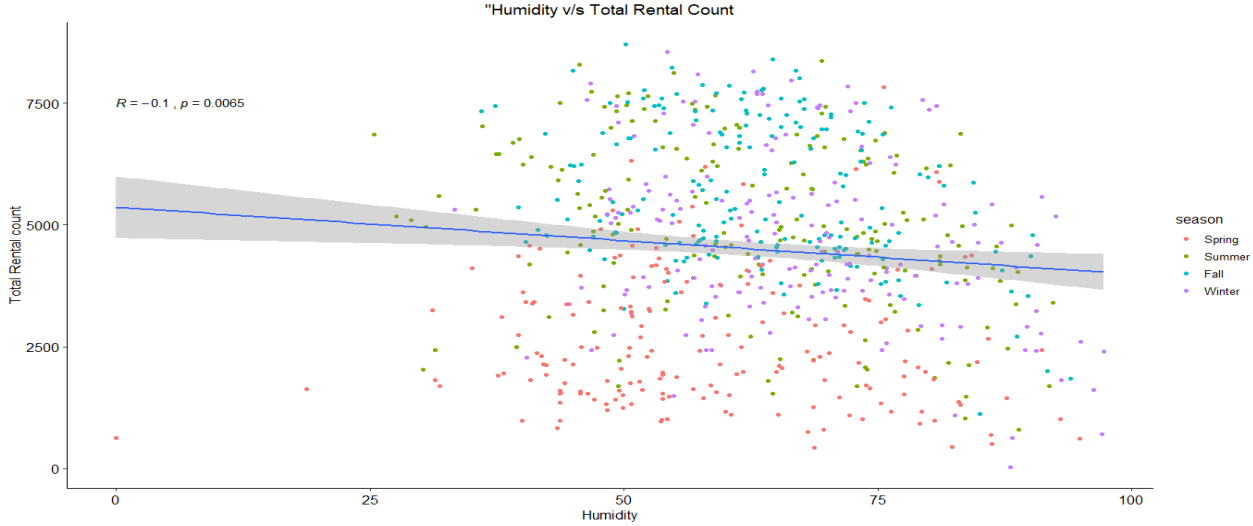
#Temperature v/s Total Rental Count



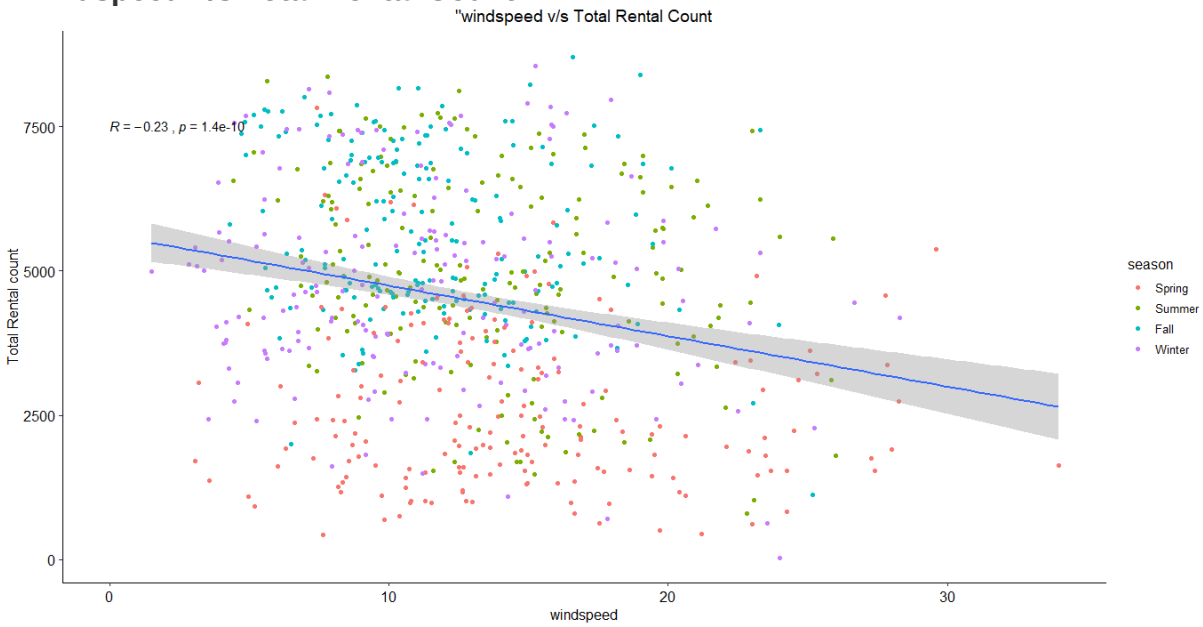
#"Feels Like" Temperature v/s Total Rental Count



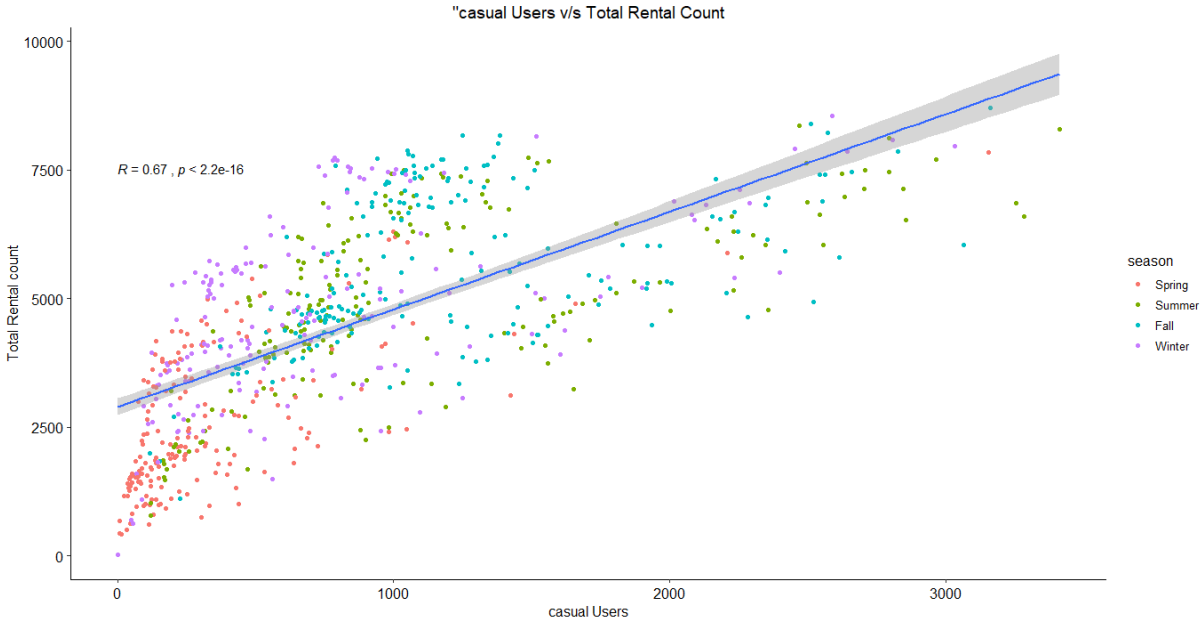
#Humidity v/s Total Rental Count



#Windspeed v/s Total Rental Count



#Casual v/s Total Rental Count



#Registered v/s Total Rental Count

2.5 Missing Value Analysis

In statistics, missing data or missing values occur when no data value is stored for the variable in an observation. Missing values are a common occurrence in data analysis. These values can have a significant impact on the results or conclusions that would be drawn from these data. If a variable has more than 30% of its values missing, then those values can be ignored, or the column itself is ignored. In our case, **none of the columns have missing values**.

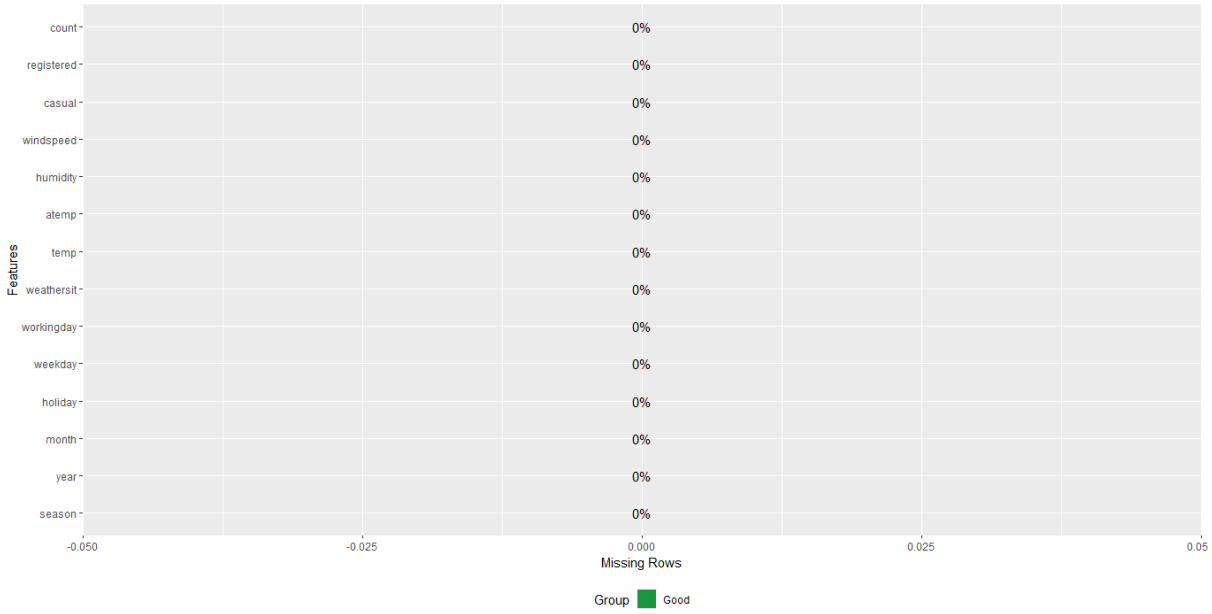
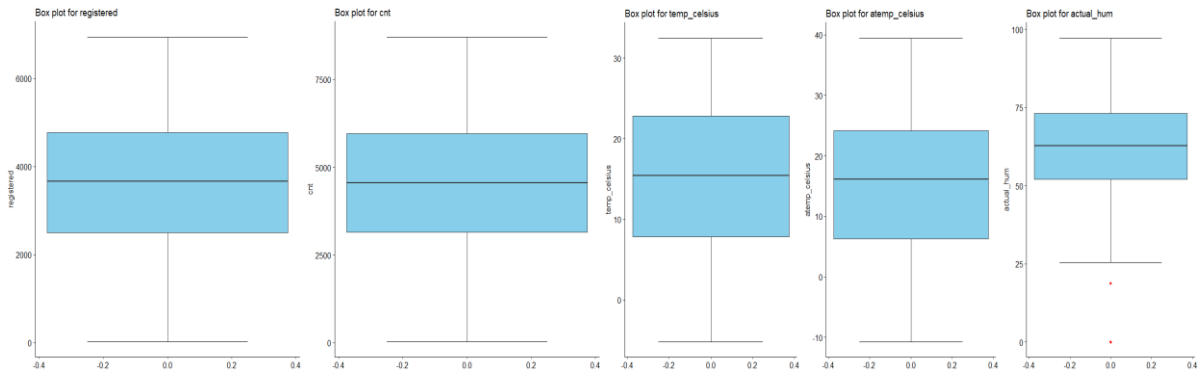


Figure: Missing Values by percentage in each column

2.6 Outlier Analysis

One of the steps in pre-processing involves the detection and removal of such outliers. In this project, we use boxplot to visualize and remove outliers. Any value lying outside of the lower and upper whisker of the boxplot are outliers. We have natural outliers in our dataset. Hence we won't remove them

In figure we have plotted the boxplots of all continuous variables.



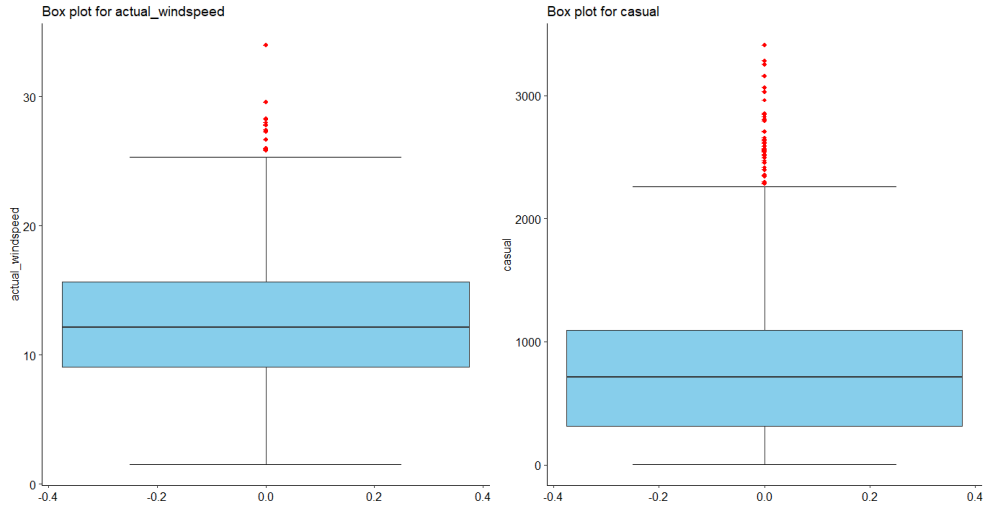


Figure – Boxplots of continuous variables with outliers

2.7 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. Also from our business understanding we also remove those columns which might not be contributing to the dataset for example 'instant' and 'dteday'.
From correlation analysis we have found that 'temp' and 'atemp' has high correlation (>0.9), so we have excluded the 'atemp' column from our dataset.

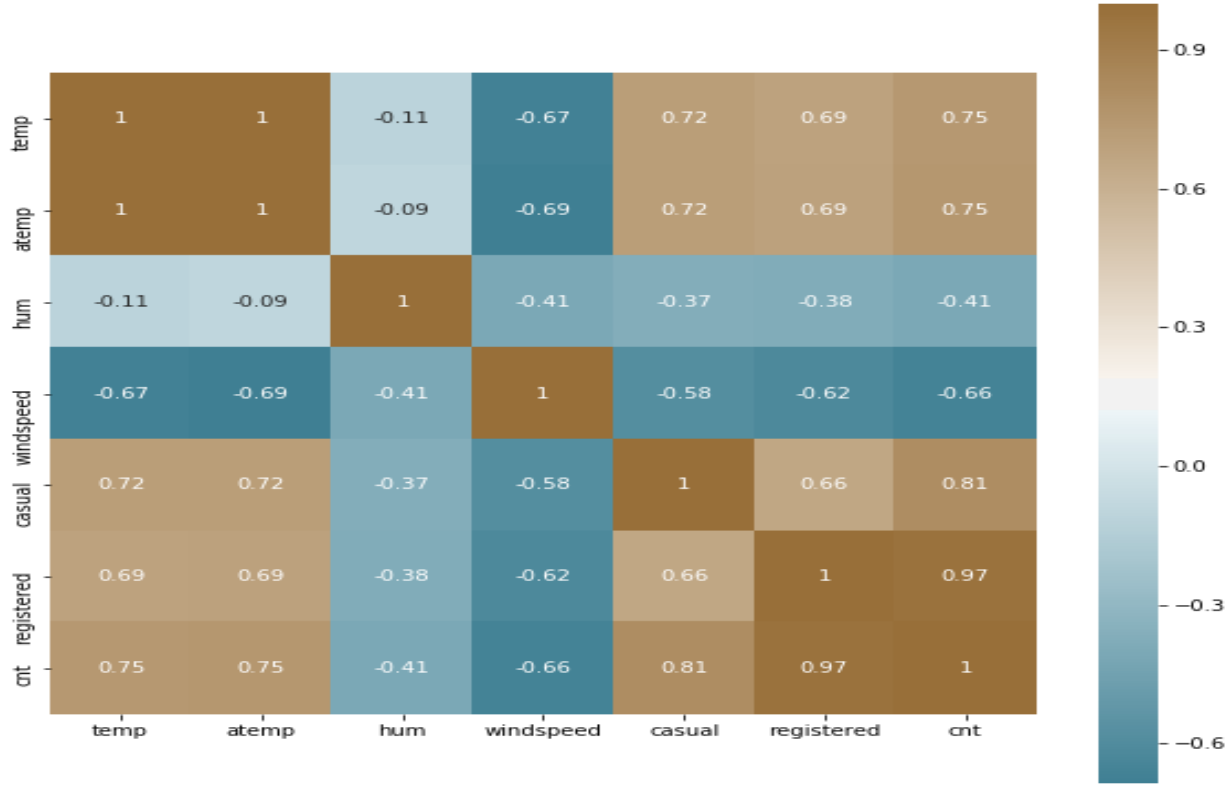


Figure – Correlation plot of Continuous variables

2.8 Feature Scaling

Feature scaling is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step.

Most classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this feature. Therefore, the range of all features should be scaled so that each feature contributes proportionately to the model and our result is not biased towards the variable greater in magnitude.
In our dataset all the predictor variables are in comparable magnitude i.e. they have range in same scale or they are already scaled.
So **feature scaling is not required for this dataset**.

	instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
0	1	2011-01-01	1	0	1	0	6	0	2	0.344167	0.363625	0.805833	0.160446	331	654	985
1	2	2011-01-02	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131	670	801
2	3	2011-01-03	1	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
3	4	2011-01-04	1	0	1	0	2	1	1	0.200000	0.212122	0.590435	0.160296	108	1454	1562
4	5	2011-01-05	1	0	1	0	3	1	1	0.226957	0.229270	0.436957	0.186900	82	1518	1600

Table – Scaled Continuous variables

Chapter 3: Modelling

3.1 Model Selection

After a thorough pre-processing we will be using some regression models on our processed data to predict the target variable. The target variable in our model is a continuous variable i.e., Total Bike rental count 'cnt', Count of casual users 'casual' and Count of registered user 'registered'. Hence the models that we choose are Linear Regression, Decision Tree, Random Forest and XG Boost. The error metric chosen for the given problem statement is Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE). We will select our best fit model by comparing both the statistical values.

3.2 Decision Tree

Decision Tree algorithm belongs to the family of supervised learning algorithms. Decision trees are used for both classification and regression problems.

A decision tree is a tree where each node represents a feature (attribute), each link (branch) represents a decision (rule) and each leaf represents an outcome (categorical or continues value). The general motive of using Decision Tree is to create a training model which can be used to predict class or value of target variables by learning decision rules inferred from prior data (training data).

3.2.1 Decision Tree for 'cnt'

DECISION TREE	MAPE	RMSE	MSE	R^2
R	7.40	27.05	10866.09	Not Calculated
PYTHON	Not Calculated	57.487	3304.795	0.9991

3.2.2 Decision Tree for 'registered'

DECISION TREE	MAPE	RMSE	MSE	R^2
R	0.583	48.507	639.601	Not Calculated
PYTHON	Not Calculated	78.72	6198.40	0.9975

3.2.3 Decision Tree for 'casual'

DECISION TREE	MAPE	RMSE	MSE	R^2
R	2.934	24.487	163.046	Not Calculated
PYTHON	Not Calculated	25.19	634.69	0.9984

3.3 Random Forest

Random Forest is a supervised learning algorithm. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. It can be used for both classification and regression problems. The method of combining trees is known as an ensemble method. Ensemble is nothing but a combination of weak learners (individual trees) to produce a strong learner.
The number of decision trees used for prediction in the forest is 100.

3.3.1 Random Forest for 'cnt'

DECISION TREE	MAPE	RMSE	MSE	R^2
R	8.664	70.95	1368.269	Not Calculated
PYTHON	Not Calculated	17.841	318.330	0.9999

3.3.2 Random Forest for 'registered'

DECISION TREE	MAPE	RMSE	MSE	R^2
R	0.756	18.113	328.080	Not Calculated
PYTHON	Not Calculated	18.351	336.77	0.9998

3.3.3 Random Forest for 'casual'

DECISION TREE	MAPE	RMSE	MSE	R^2
R	0.858	9.593	92.041	Not Calculated
PYTHON	Not Calculated	28.855	832.657	0.9979

3.4 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.

3.4.1 Linear Regression for 'cnt'

DECISION TREE	MAPE	RMSE	MSE	R^2
R	7.71	185.206	9324.123	Not Calculated
PYTHON	Not Calculated	1.317×10^{-5}	1.736×10^{-10}	1

3.4.2 Linear Regression for 'registered

DECISION TREE	MAPE	RMSE	MSE	R^2
R	5.519	137.364	5128.953	Not Calculated
PYTHON	Not Calculated	1.56×10^{-5}	6.67×10^{-11}	1

3.4.3 Linear Regression for 'casual'

DECISION TREE	MAPE	RMSE	MSE	R^2
R	11.948	81.215	1792.995	Not Calculated
PYTHON	Not Calculated	7.132×10^{-6}	5.087×10^{-11}	1

3.5 XG Boost

XG Boost (Extreme Gradient Boosting) is an advanced and more efficient implementation of Gradient Boosting Algorithm.

Advantages over Other Boosting Techniques

- It is 10 times faster than the normal Gradient Boosting as it implements parallel processing. It is highly flexible as users can define custom optimization objectives and evaluation criteria, has an inbuilt mechanism to handle missing values.
- Unlike gradient boosting which stops splitting a node as soon as it encounters a negative loss, XG Boost splits up to the maximum depth specified and prunes the tree backward and removes splits beyond which there is an only negative loss.

Extreme gradient boosting can be done using the XG Boost package in R and Python.

3.5.1 XG Boost for 'cnt'

DECISION TREE	MAPE	RMSE	MSE	R^2
R	5.263	161.427	7083.153	Not Calculated
PYTHON	Not Calculated	22.491	505.880	0.9998

3.5.2 XG Boost for 'registered'

DECISION TREE	MAPE	RMSE	MSE	R^2
R	0.464	38.859	410.466	Not Calculated
PYTHON	Not Calculated	16.695	278.775	0.9998

3.5.3 XG Boost for 'casual'

DECISION TREE	MAPE	RMSE	MSE	R^2
R	1.625	20.072	402.907	Not Calculated
PYTHON	Not Calculated	23.755	564.344	0.9986

Chapter 4: Conclusion

4.1 Model Evaluation

In the previous chapter we have seen the Mean Absolute Percentage error (MAPE), Root Mean Square Error (RMSE), Mean Squared Error (MSE) and R-Squared Value of different models.

4.1.1 Root Mean Square Error (RMSE)

It is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are, RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.

1. RMSE express average model prediction error in units of the variable of interest. RMSE can range from 0 to ∞ and are indifferent to the direction of errors. They are negatively-oriented scores, which means lower values are better.

2. Taking the square root of the average squared errors has some interesting implications for RMSE. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE should be more useful when large errors are particularly undesirable.

4.1.2 Mean Absolute Percentage Error (MAPE)

The mean absolute percentage error (MAPE) is a statistical measure of how accurate a forecast system is. It measures this accuracy as a percentage, and can be calculated as the average absolute percent error for each time period minus actual values divided by actual values. The mean absolute percentage error (MAPE) is the most common measure used to forecast error, and works best if there are no extremes to the data (and no zeros).

4.1.3 Mean Squared Error (MSE)

MSE basically measures average squared error of our predictions. For each point, it calculates square difference between the predictions and the target and then average those values.
The higher this value, the worse the model is. It is never negative, since we're squaring the individual prediction-wise errors before summing them, but would be zero for a perfect model.

4.1.4 R-squared

It explains how well your selected independent variable(s) explain the variability in your dependent variable(s) R-squared gives you the percentage variation in y explained by x-variables. The range is 0 to 1 (i.e. 0% to 100% of the variation in y can be explained by the x-variables.
Variance—in terms of linear regression, **variance** is a measure of how far observed values differ from the average of predicted values, i.e., their difference from the **predicted value mean**. The goal is to have a value that is low

4.2 Model Selection

Here we will select best fit model for each of our target variables by analysing MAPE, RMSE and R^2 values. For this dataset each of the model is giving high accuracy hence we might face the problem of overfitting. To avoid this we will select model which is moderately fitting the dataset among all the models i.e. neither too high nor low accuracy.

4.2.1 Model selection for 'cnt' variable

We will select 'Random Forest' as our Model for 'cnt' variable because it has low **RMSE (70.95)** and **MSE (1368.269)** values. **MAPE** is **8.66%** which is a good fit model.

4.2.2 Model selection for 'registered' variable

We will select 'Random Forest' as our Model for 'registered' variable because it has low **RMSE (18.113)** and **MSE (328.080)** values. **MAPE** is **0.756%** which is a good fit model.

4.2.3 Model selection for 'casual variable

We will select 'XG Boost' as our Model for 'casual' variable because it has low **RMSE (20.072)** and **MSE (402.907)** values. **MAPE** is **1.625%** which is a good fit model.

4.3 Key Prediction

I. Number of registered user count is mostly greater than casual user count throughout the year across all season.
II. There is 64.87% increase in total bike rental count from year 2011 to 2012.
III. Season wise Analysis of total bike rental count.

Season	2011	2012	Total Count	Percentage
1.Spring	150,000 (12.06%)	321,348 (15.68%)	471,348	14.32
2.Summer	347,316 (27.94%)	571,273 (27.87%)	918,589	27.89
3.Fall	419,650 (33.75%)	641,479 (31.30%)	1,061,129	32.22
4.Winter	326,137 (26.24%)	515,476 (25.15%)	841,613	25.56
Total	1,243,103	2,049,576	3,292,679	100

About 60 % of total bikes are rented in summer and fall season together throughout the year.

IV. Weather situation wise analysis.

Weather Situation	Total Rental Count	Percentage Count
1. Good: Clear/Sunny	2,257,952	68.58%
2. Moderate: Cloudy/Mist	996,858	30.27%
3. Bad: Rain/Snow/Fog	37,869	1.15%
4. Worse: Heavy Rain/Snow/Fog	NIL	NIL

As soon as weather is clear and cloudy people are renting bike more. But as rain comes in bike rental gets less and less as the condition get worse.

V. Total bike rental on non-working days is about 30.38% as compared to rentals on working days which is 69.62% of total bike rentals i.e. people are renting bikes more to commute to office.
VI. Weekday analysis of total bike rental

Day of week	Total Rental Count
0	444,027
1	455,503
2	469,109
3	473,048
4	485,395
5	487,790
6	477,807

Bike rentals is almost similar through the days of week.

Chapter 5: R Code (Double-click to open)

```
#Remove all objects stored
rm(list = ls())



#set working directory
setwd("E:/R/Project_2")

#Loading dataset in csv format
library(readr)
df <- read_csv("day.csv", col_names = T)
View(df)
str(df)
head(df)
nrow(df)
ncol(df)
dim(df)
data.class(df)
names(df)

#Removing unnecessary variables from our dataframe,which might not be useful for our analysis
  #Removing "instant" Variable as it contains recorded index numbers i.e 1st Column
  #Removing "dteday" as we already have month year and weekday and columns for our analysis.
df_new <- df[,c(-1,-2)]
View(df_new)
str(df_new)
dim(df_new)
names(df_new)
#Changing Column names
library(data.table)
setnames(df_new, old=c("yr","mnth", "hum", "cnt"), new=c("year", "month", "humidity","count"))
```

References

1. For Data Cleaning and Model Development – https://edwisor.com/career-data-scientist

2. For R code https://datacamp.com

3. For Visualization and other information – Analytics Vidya, Youtube, Udemy, Stackoveflow