Bike Renting Prediction

Gursimran Singh

2019

Introduction

A decade ago, the condition of start-up ecosystem was very pitiable due to poor infrastructure, lack of support from the government and unavailability of funds from investors. Earlier, entrepreneurs were not used to receiving enough attention they needed from other stakeholders to sustain the start-up eco-system. But today, the scenario has changed dramatically and reflects the fact that how corporate giants and the government have come forward to boost the start-up ecosystem via mentoring, acquisition, funding, acceleration programmes and setting up of incubation centres.

One of the most demanding and successful start-up in present era is of Bike renting. A Bike renting is a service in which bikes are made available on daily rent basis to individuals on a short term basis for a price. There are many bike renting systems which allow people to borrow a bike from a "dock" and return it at another dock belonging to the same system. Docks are special bike racks that lock the bike, and only release it by computer control. The user enters payment information, and the computer unlocks a bike. The user returns the bike by placing it in the dock, which locks it in place.

Problem Statement

The objective of this dataset is to predict the bike rental count based on the environmental and seasonal settings, So that required bikes would be arranged and managed by the shops according to environmental and seasonal conditions.

Data

Our task is to a build regression model which will predict the count of the bikes rented depending on various other factors. Given below is a sample of the data set.. Sample from the whole dataset is shown below:-

Instant	Dteday	Season	Yr	mnth	holiday	weekday	workingday
1	01-01-2011	1	0	1	0	6	0
2	02-01-2011	1	0	1	0	0	0
3	03-01-2011	1	0	1	0	1	1
4	04-01-2011	1	0	1	0	2	1
5	05-01-2011	1	0	1	0	3	1

Table 1.1 : Sample Data(Columns 1-8)

Weathersit	Temp	Atemp	Hum	windspeed	casual	registered	cnt
2	0.344167	0.363625	0.805833	0.160446	331	654	985
2	0.363478	0.353739	0.696087	0.248539	131	670	801
1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
1	0.2	0.212122	0.590435	0.160296	108	1454	1562
1	0.226957	0.22927	0.436957	0.1869	82	1518	1600

Table 1.1 : Sample Data(Columns 9-16)

Descriptions of the attributes are given below:-

Attribute	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 holiday or not (extracted from Holiday Schedule)
weekday	Day of the week
workingday	If day is neither weekend nor holiday is 1, otherwise is 0.
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
Temp	Normalized temperature in Celsius.
atemp	Normalized feeling temperature in Celsius.
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

Table 1.2 : Attribute Descriptions

Methodology

Any predictive modelling requires to look at the data before 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 (EDA).

Exploratory data analysis (EDA) is a very important step which takes place after feature engineering and acquiring data and it should be done before any modelling. This is because it is very important for a data scientist to be able to understand the nature of the data without making assumptions.

The purpose of EDA is to use summary statistics and visualizations to better understand data, and find clues about the tendencies of the data, its quality and to formulate assumptions and the hypothesis of our analysis. EDA is not about making fancy visualizations or even aesthetically pleasing ones, the goal is to try and answer questions with data. A goal should be to be able to create a figure which someone can look at in a couple of seconds and understand what is going on. If not, the visualization is too complicated (or fancy) and something similar should be used.

EDA is also very iterative since we first make assumptions based on our first exploratory visualizations, and then build some models. We then make visualizations of the model results and tune our models.

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

Variable Identification

Variable identification is the first step in the exploratory data analysis. Identification of the variables in the dataset is totally dependent on the business requirements and need of the client. The main task here is to identify the Target variable on which future decision has to be made. Besides these identifying the independent variables are equally important because end result of the target variable are totally dependent on the independent/predictor variables. Brief introduction of predictor and target variables are given below:-

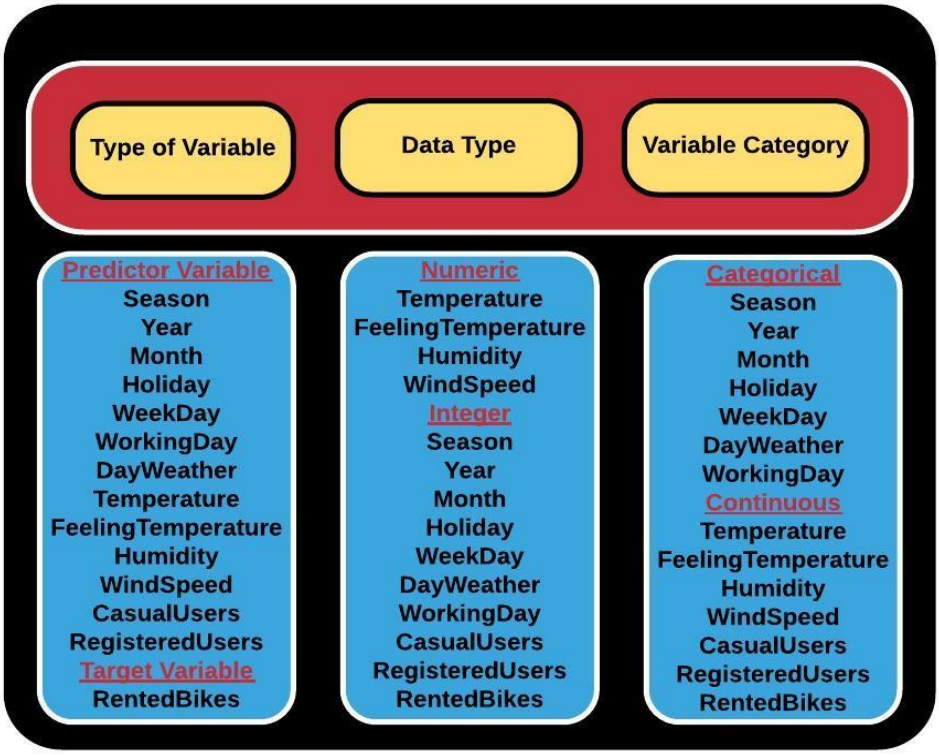
Predictor Variable

Predictor variables are those variables or attributes in the dataset on which the result of the target variable is totally dependent. These variables are those on which decisions are made by the clients to get the maximum profit from the business.

Target Variable

It is the variable or attribute in the whole dataset on which client is mostly interested. Based on the business requirements category of the target variable is identified by the data sciences experts.

Once the target variable and predictor variables are identified, our next task to identify the data types and categories of the variable. From analysing the dataset of bike renting company, the detailed description of all the variables are given below in the diagram on the different parameters.



In the further stages of exploratory data analysis process, we have to dive deep into the understanding of the each variable present in the dataset. From Business point of view each and every variable is crucial and even a minute mistake here can cause a loss of millions to your client. So to get the detailed summary of all the variables in the dataset on statistical parameters will help to better understanding of data to a data sciences expert. Detailed summary of all the variables are given below:-

Detailed Summary of Variables

	Season	Year	Month	Holiday	WeekDay	WorkingDay	DayWeather	Temperature
Nobs	731	731	731	731	731	731	731	731
NAs	0	0	0	0	0	0	0	0
Minimum	1	0	1	0	0	0	1	0.05913
Maximum	4	1	12	1	6	1	3	0.861667
1. Quartile	2	0	4	0	1	0	1	0.337084
3. Quartile	3	1	10	0	5	1	2	0.655417
Mean	2.49658	0.500684	6.519836	0.028728	2.997264	0.683995	1.395349	0.495385
Median	3	1	7	0	3	1	1	0.498333
Sum	1825	366	4766	21	2191	500	1020	362.1263
SE Mean	0.041085	0.018506	0.127674	0.006182	0.07415	0.017207	0.020154	0.00677
LCL Mean	2.415922	0.464353	6.269185	0.01659	2.851692	0.650213	1.355783	0.482093
UCL Mean	2.577238	0.537015	6.770487	0.040865	3.142836	0.717776	1.434915	0.508677
Variance	1.233892	0.250342	11.9157	0.027941	4.019171	0.216442	0.29691	0.033508
Stdev	1.110807	0.500342	3.451913	0.167155	2.004787	0.465233	0.544894	0.183051
Skewness	-0.00038	-0.00273	-0.00812	5.63104	0.00273	-0.7899	0.95346	-0.0543
Kurtosis	-1.34617	-2.00273	-1.21395	29.74932	-1.25869	-1.37795	-0.15154	-1.12456

	FeelingTemperature	Humidity	WindSpeed	CasualUsers	RegisteredUsers	RentedBikes
Nobs	731	731	731	731	731	731
NAs	0	0	0	0	0	0
Minimum	0.07907	0	0.022392	2	20	22
Maximum	0.840896	0.9725	0.507463	3410	6946	8714
1. Quartile	0.337842	0.52	0.13495	315.5	2497	3152
3. Quartile	0.608602	0.730208	0.233214	1096	4776.5	5956
Mean	0.474354	0.627894	0.190486	848.1765	3656.172	4504.349
Median	0.486733	0.626667	0.180975	713	3662	4548
Sum	346.7528	458.9906	139.2454	620017	2672662	3292679
SE Mean	0.006027	0.005268	0.002866	25.39565	57.70817	71.65035
LCL Mean	0.462521	0.617552	0.184859	798.3192	3542.879	4363.684
UCL Mean	0.486187	0.638236	0.196114	898.0337	3769.466	4645.014
Variance	0.026556	0.020286	0.006006	471450.4	2434400	3752788
Stdev	0.162961	0.142429	0.077498	686.6225	1560.256	1937.211
Skewness	-0.13055	-0.0695	0.674568	1.261261	0.04348	-0.04716
Kurtosis	-0.99211	-0.08029	0.390624	1.293082	-0.72267	-0.82055

NOTE: For my better understanding I have changed the names of the dataset variables.

Univariate Analysis

Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words data has only one variable. It doesn’t deal with causes or relationships (unlike regression) and it’s major purpose is to describe. It takes data, summarizes that data and finds patterns in the data.

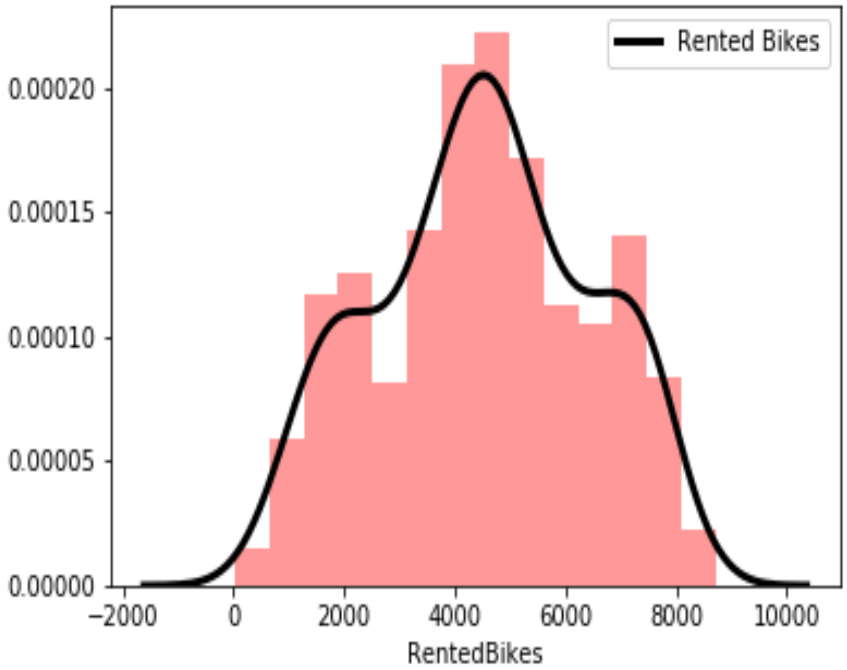
Method to perform univariate analysis will depend on whether the variable type is categorical or continuous.

Continuous Variables

A continuous variable is a variable that has an infinite number of possible values. In other words, any value is possible for the variable. All the continuous variables present in our dataset are analysed below:-

Rented Bikes

Rented bike is our target variable and most crucial variable from the business point of view. Whole business model will be revolving around this variable. Distribution of this variable is shown below in following plot:-

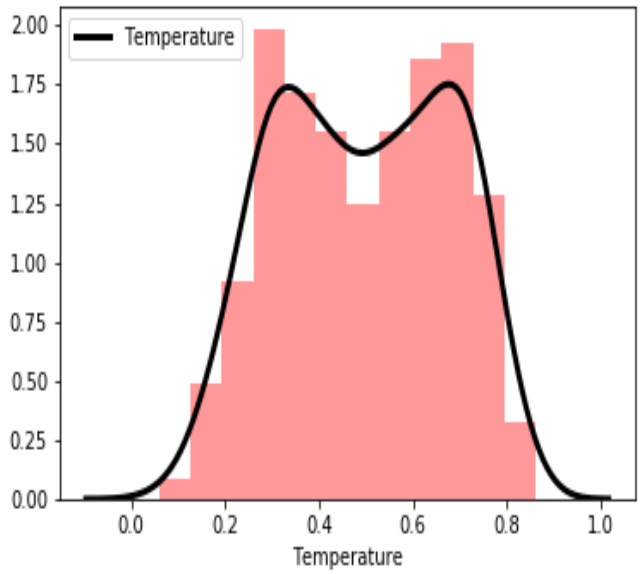


From the distribution of the Rented Bikes variable we came to knew that data of this variable is fairly symmetrical. Moreover, this variable is unimodal in nature. Skewness and Kurtosis of the Rented Bikes variable are given below:-

Skewness	**-0.047353**
Kurtosis	**-0.811922**

Temperature

Temperature variable is predictor variable in the dataset. Distribution of the temperature variable is shown below in the following plot:-

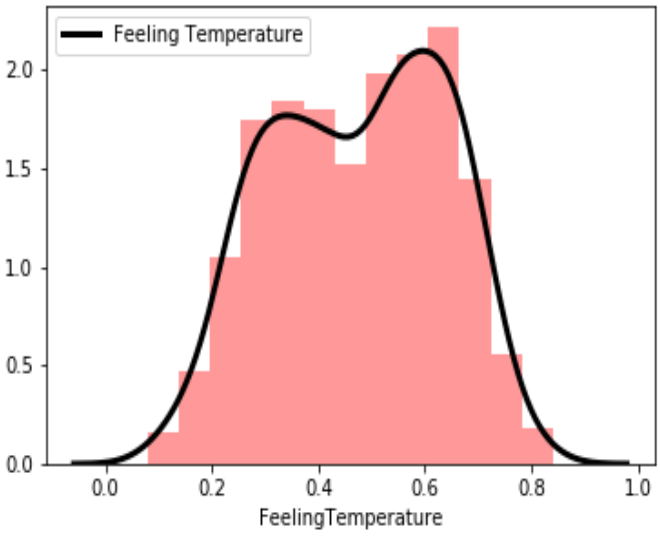


Temperature variable is bimodal in nature. It is a fairly symmetrical variable. Skewness and Kurtosis values for the above variable are given below:-

Skewness **-0.054521**
Kurtosis **-1.118864**

Feeling Temperature

Feeling temperature is a predictor variable in our dataset. Distribution of this variable is shown in the following plot:-

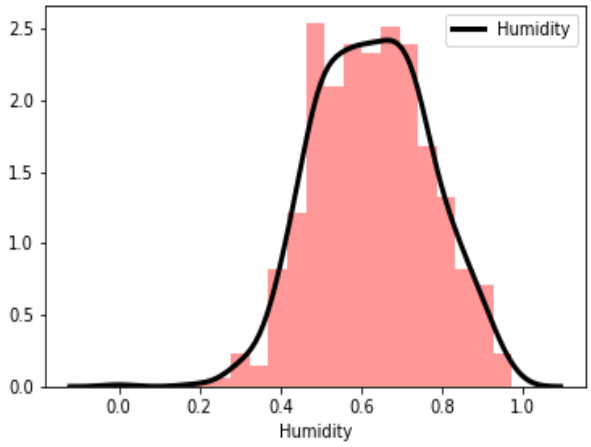


Feeling temperature variable is fairly symmetrical and bimodal in nature. Skewness and Kurtosis of this variable is shown below-

Skewness **-0.131088**
Kurtosis **-0.985131**

Humidity

Humidity is a predictor variable in our dataset. The distribution of this variable is shown below:-

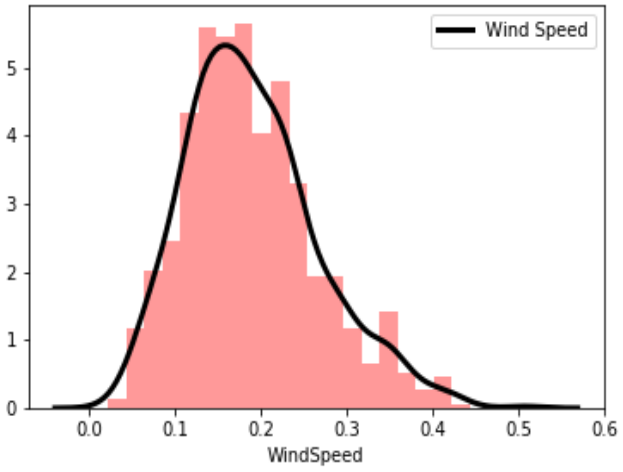


Humidity variable is unimodal in nature. Moreover this variable is fairly symmetrical. Skewness and Kurtosis values of this variable are given below:-

Skewness **-0.069783**
Kurtosis **-0.064530**

Wind Speed

Wind speed variable is predictor variable in our dataset. Distribution of this variable is shown below:-

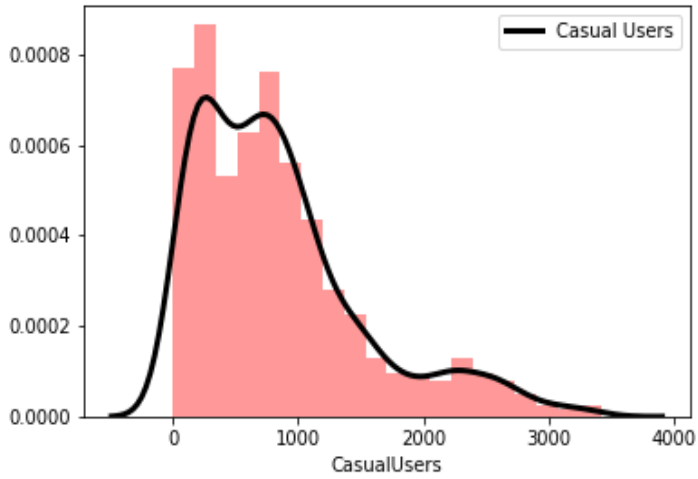


Wind speed variable is unimodal in nature. This variable is moderately skewed. Skewness and Kurtosis value of this variable are given below:-

Skewness	**0.677345**
Kurtosis	**0.410922**

Casual Users

Casual users variable are independent variable in nature. Distribution of the variable is shown below in the following plot:-

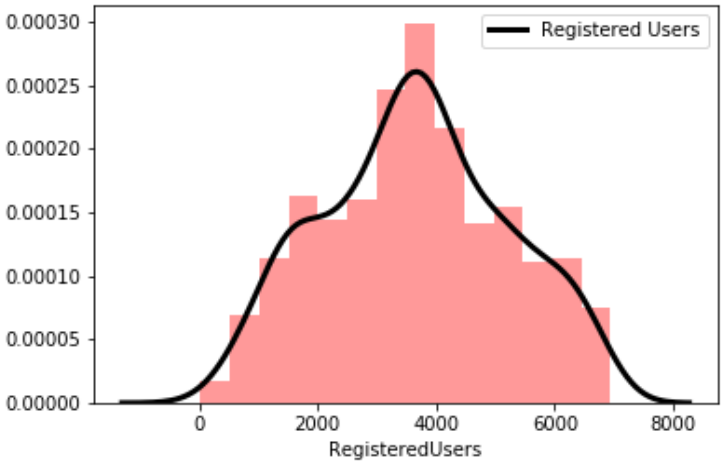


Casual users variable are bimodal in nature. This variable is highly skewed. Skewness and Kurtosis of this variable is given below:-

Skewness	**1.266445**
Kurtosis	**1.322074**

Registered Users

Registered users are independent variable. Distribution of the variable is shown below in the following plot:-

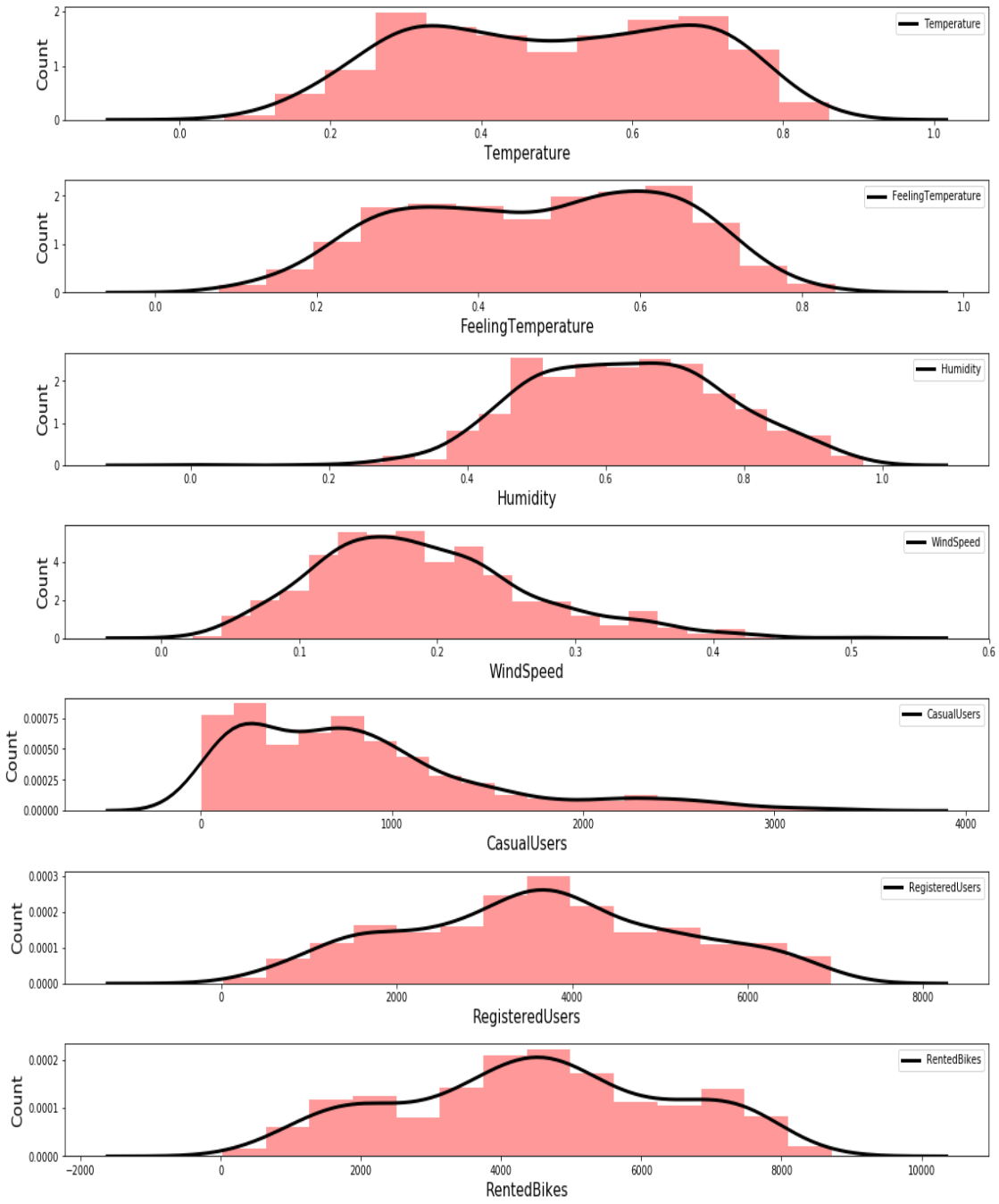


Registered users variable is uniformly distributed.This variable is fairly symmetrical. Skewness and Kurtosis values are shown below in the following table:-

Skewness	**0.043659**
Kurtosis	**-0.713097**

Distribution plot of all numerical Variables

The distribution plot is suitable for comparing range and distribution for groups of numerical data. Distribution of all the numeric variables present in our dataset in shown below:-



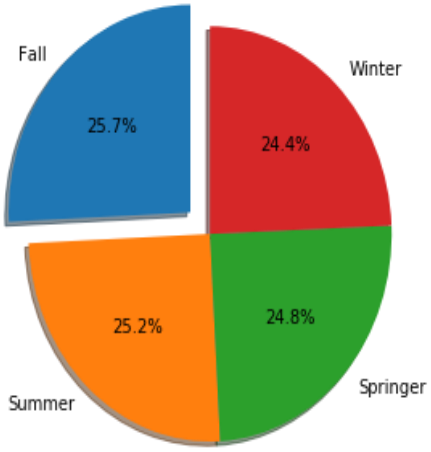
Dist plot of all the Numerical Variables

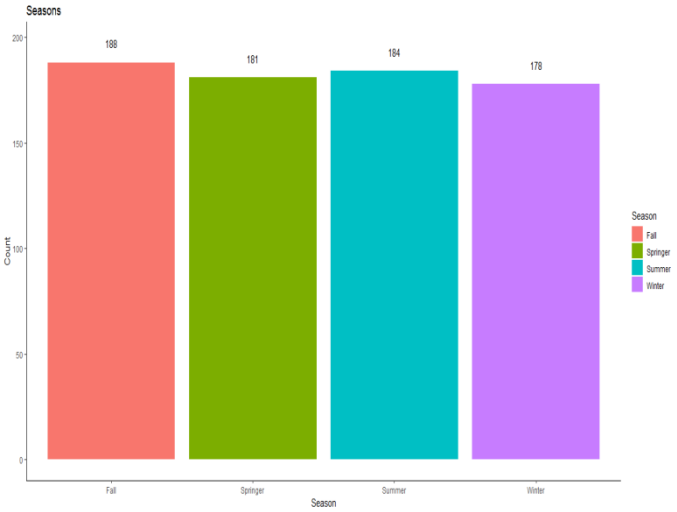
Categorical Variables

Categorical data are easier to interpret as compared to numerical variables. The best ways to analyze categorical variables are through bar graphs and pie charts. Bar graphs are mostly preferred to visualize the frequency of each category that falls into that variable, whereas pie charts are used to visualize the percentage of each category. Plots for analyzing categorical variables are shown below:-

Season

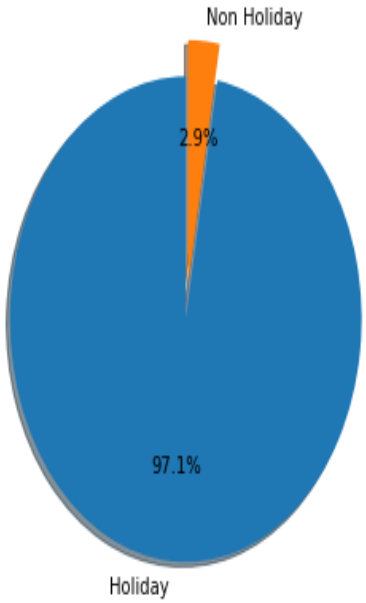
Season attribute is a predictor variable which falls into the category of categorical variables.

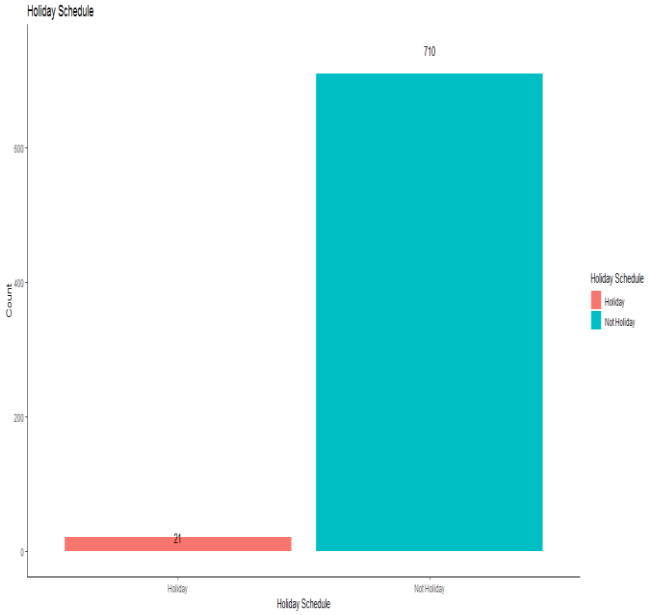




Holiday

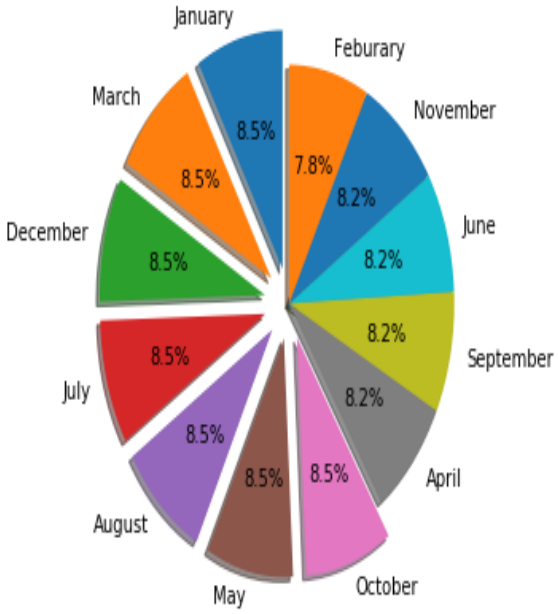
Holiday attribute is a predictor variable which falls into the category of categorical variables.

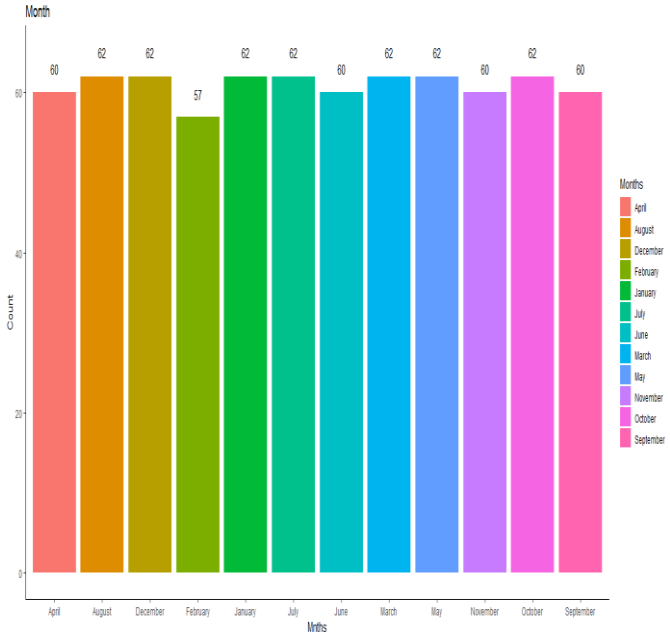




Month

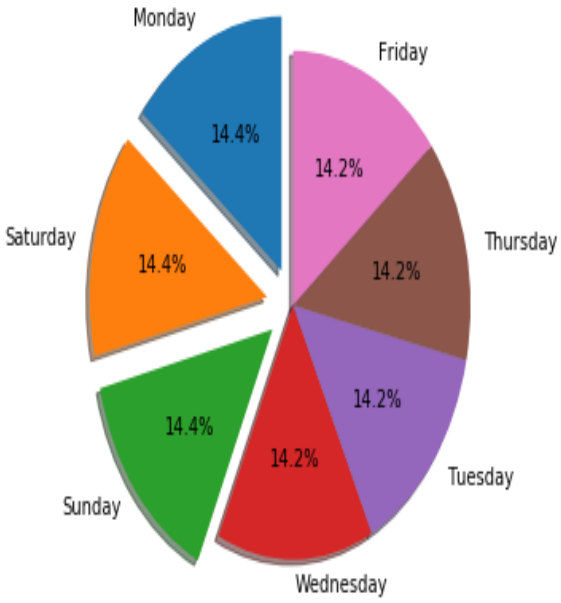
Month attribute is a predictor variable which falls into the category of categorical variables.

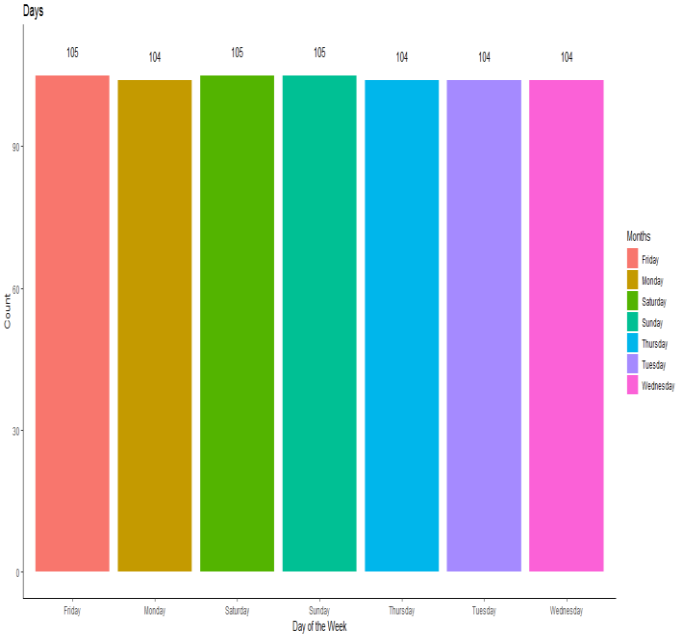




Days

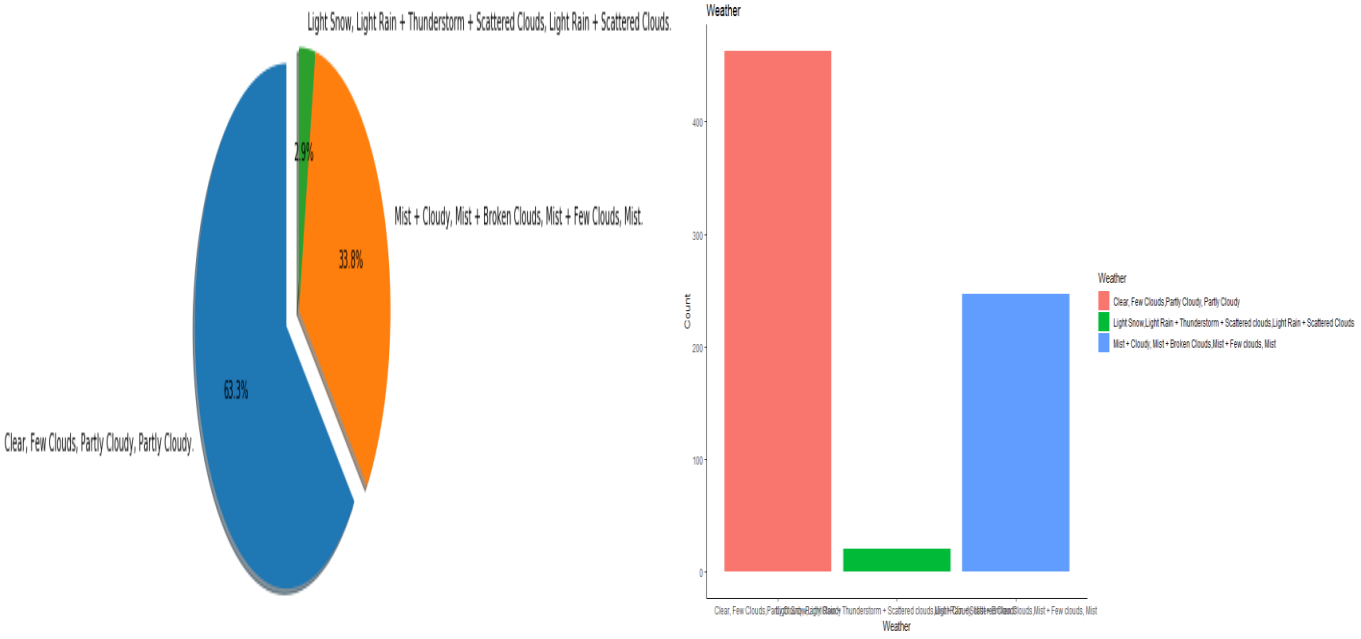
Days attribute is a predictor variable which falls into the category of categorical variables.





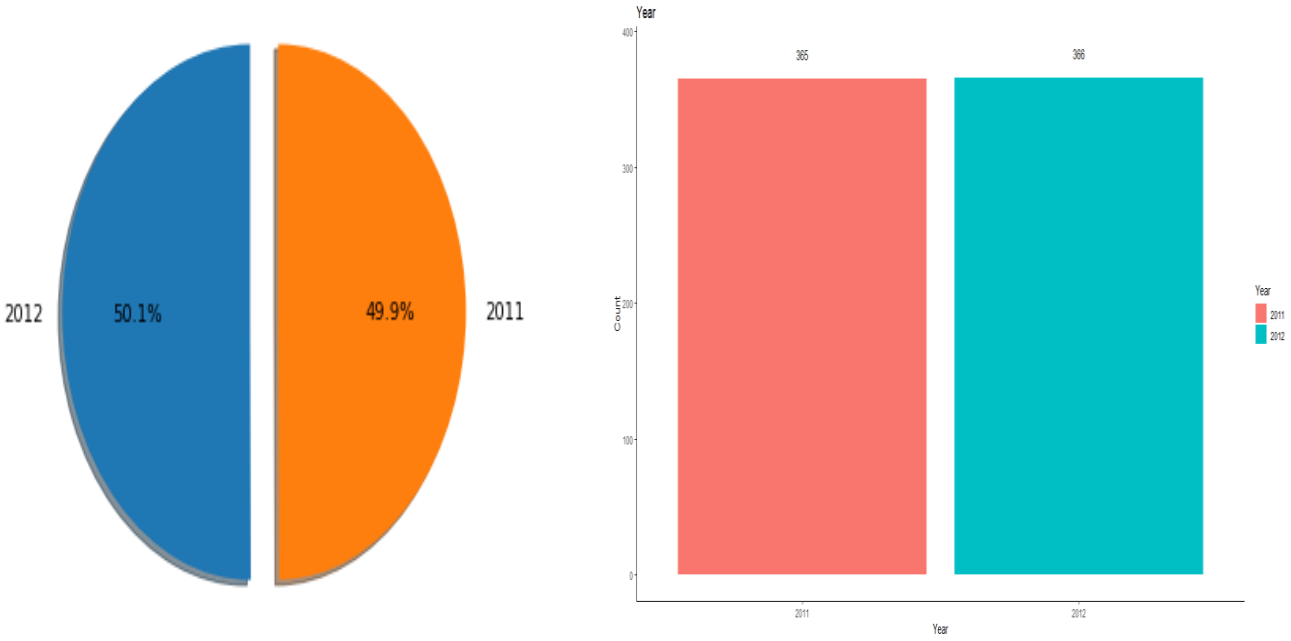
Weather

Weather attribute is a predictor variable which falls into the category of categorical variables.



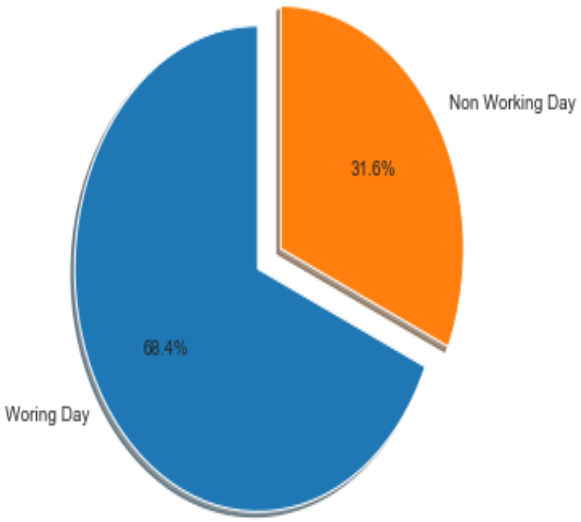
Year

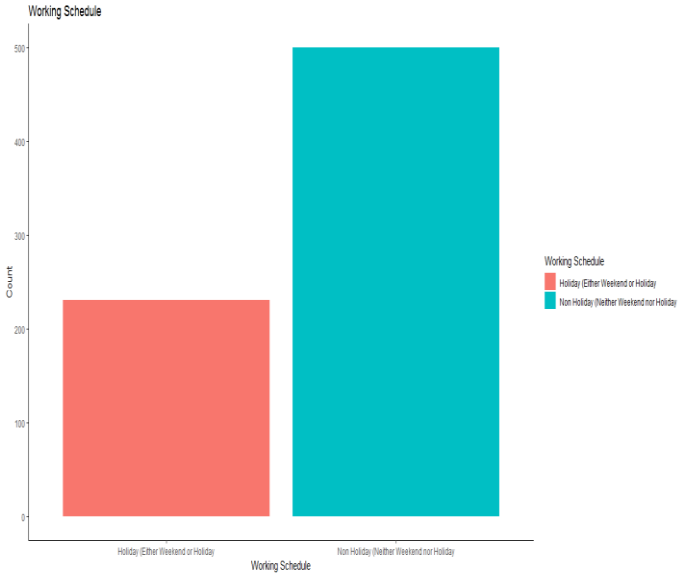
Year attribute is a predictor variable which falls into the category of categorical variables.



Working Schedule

Working Schedule attribute is a predictor variable which falls into the category of categorical variables.



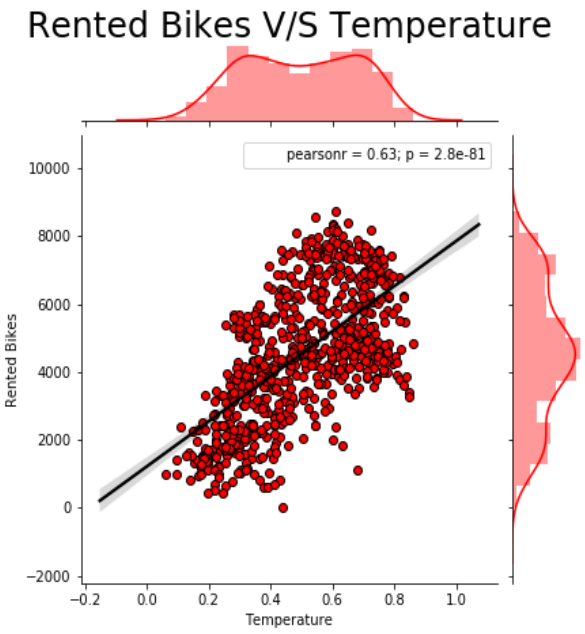


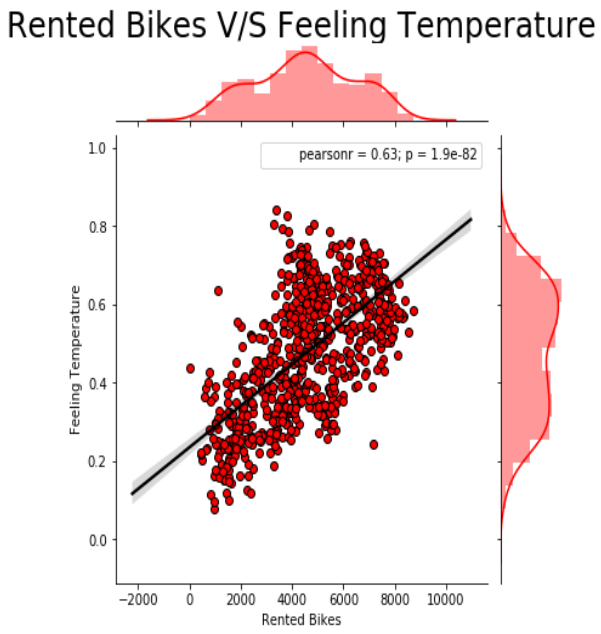
Bivariate Analysis

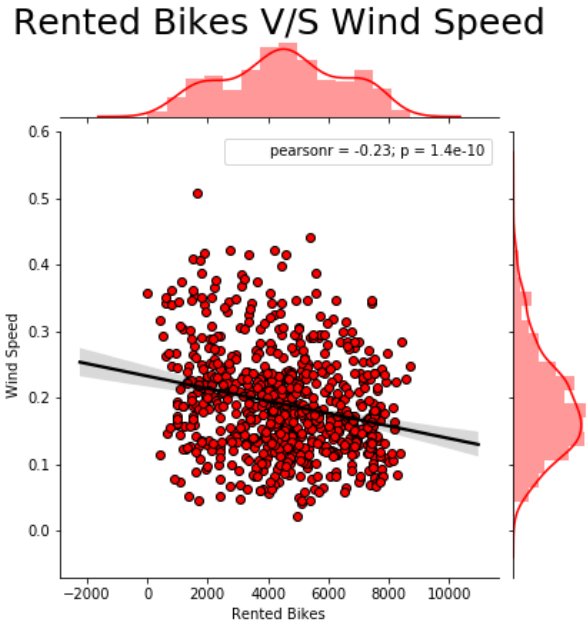
Bivariate analysis is the simultaneous analysis of two variables. It explores the concept of relationship between two variables, whether there exists an association and the strength of this association, or whether there are differences between two variables and the significance of these differences.

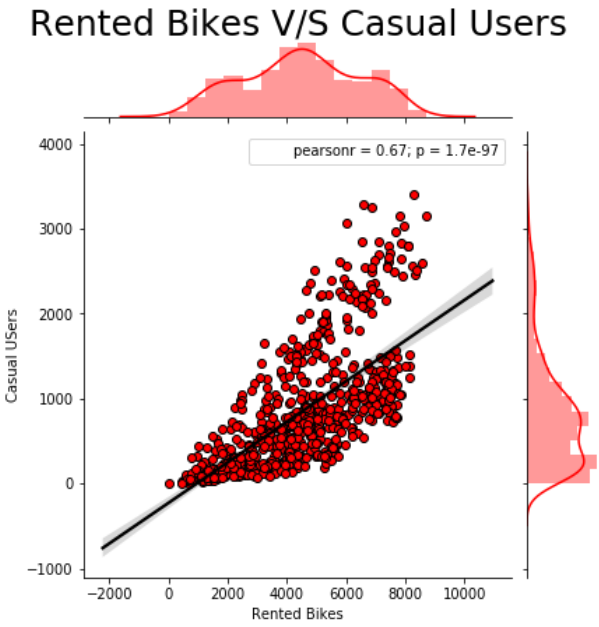
Continuous Variables

Scatter plot for all the continuous variables are shown below. Correlation value is shown in the legend and it describes the relationship between variables against which graphs are plotted.

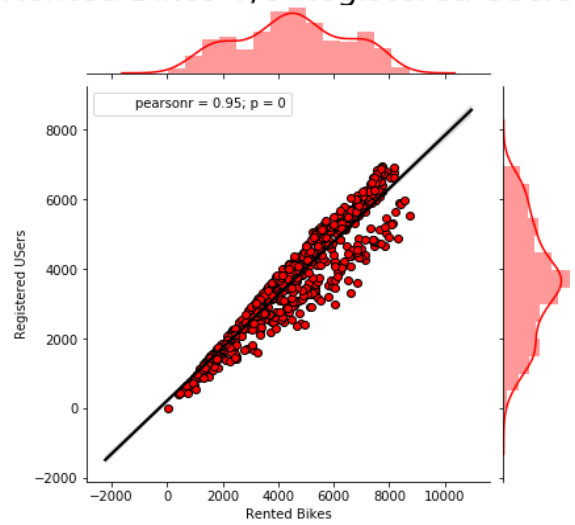




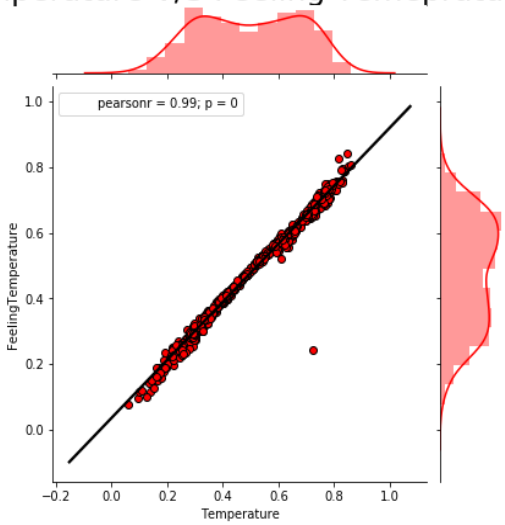




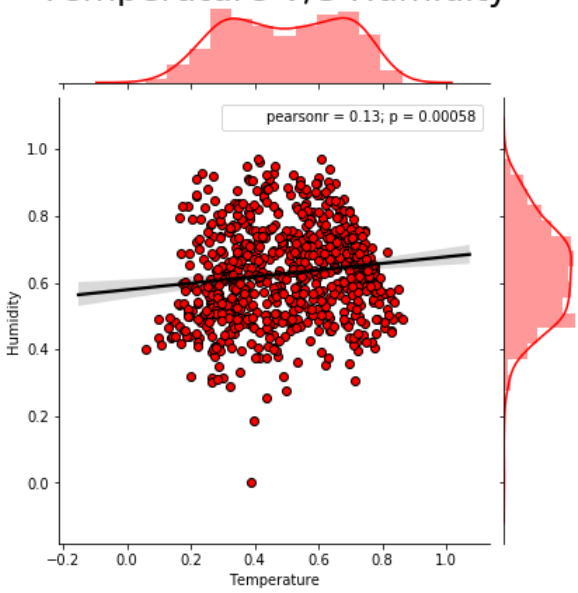
Rented Bikes V/S Registered Users



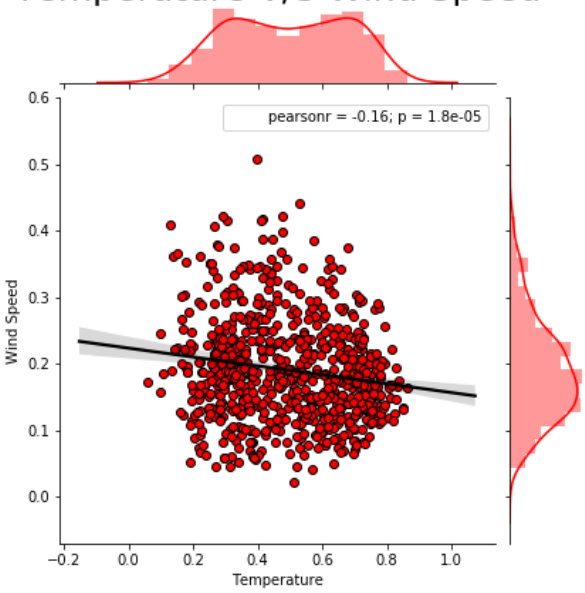
Temperature V/S Feeling Temeprature



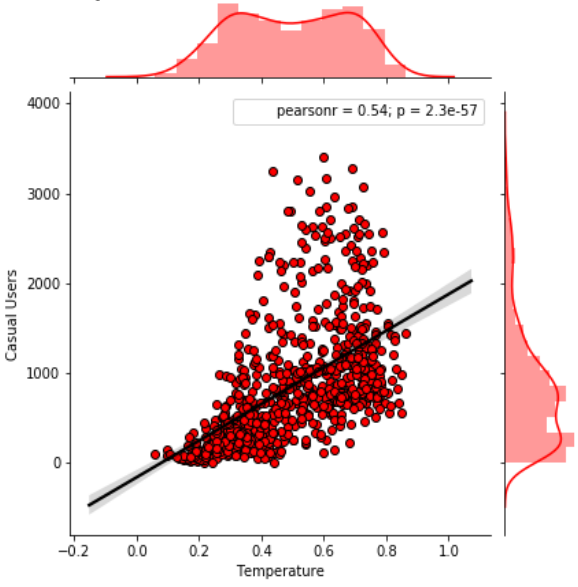
Temperature V/S Humidity



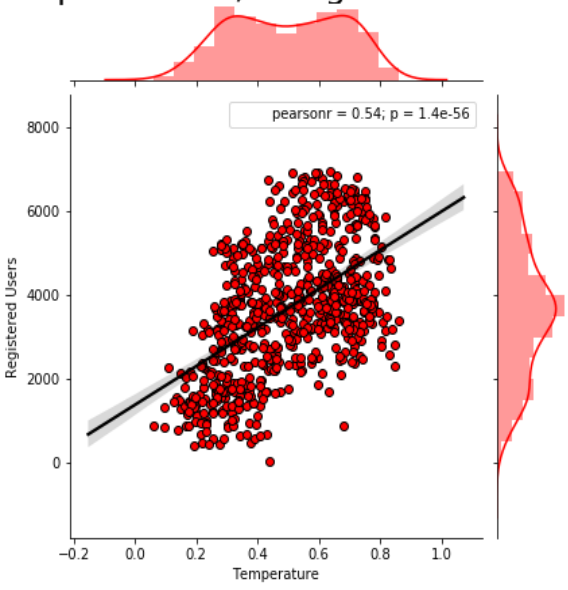
Temperature V/S Wind Speed



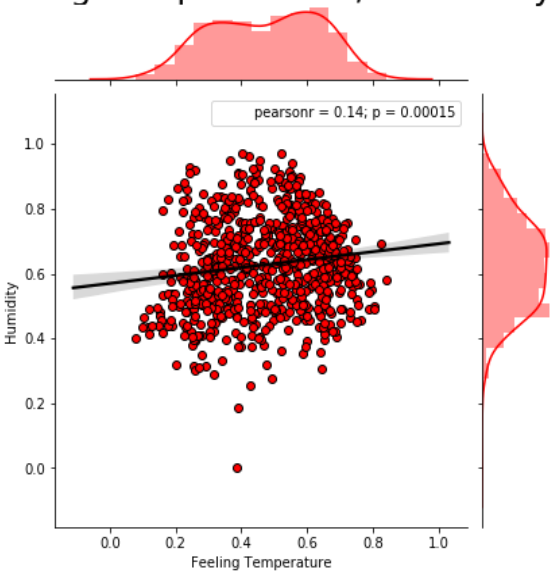
Temperature V/S Casual Users



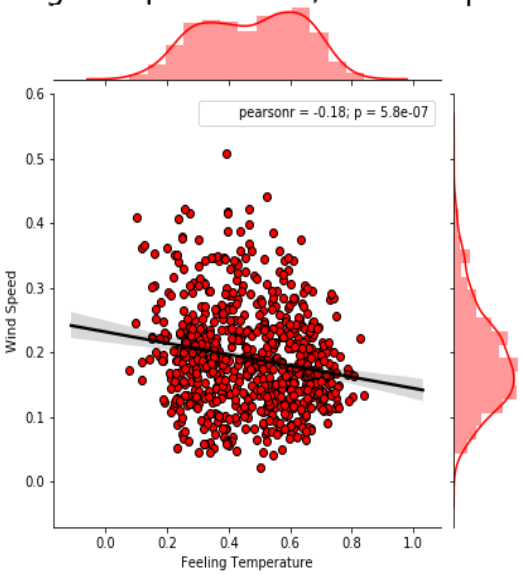
Temperature V/S Registered Users



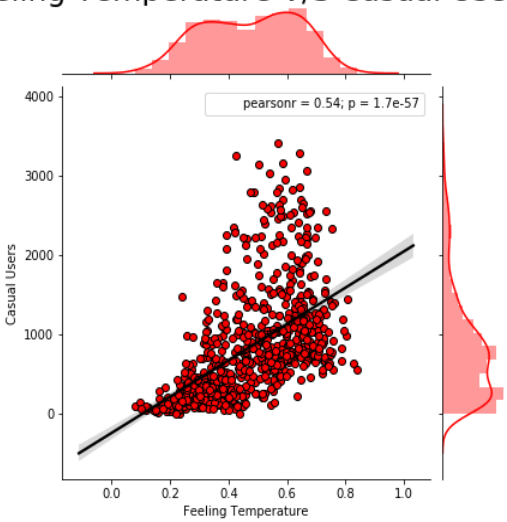
Feeling Temperature V/S Humidity



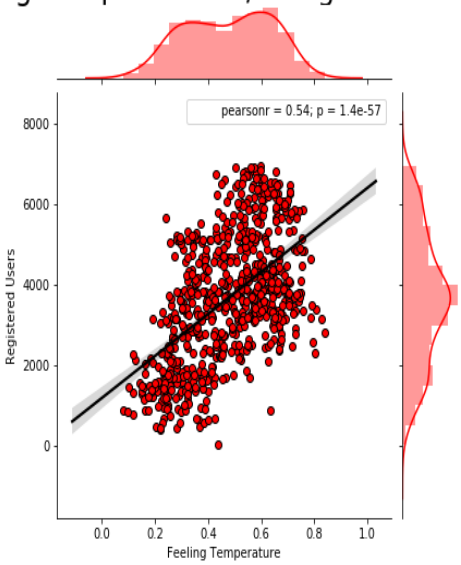
Feeling Temperature V/S Wind Speed



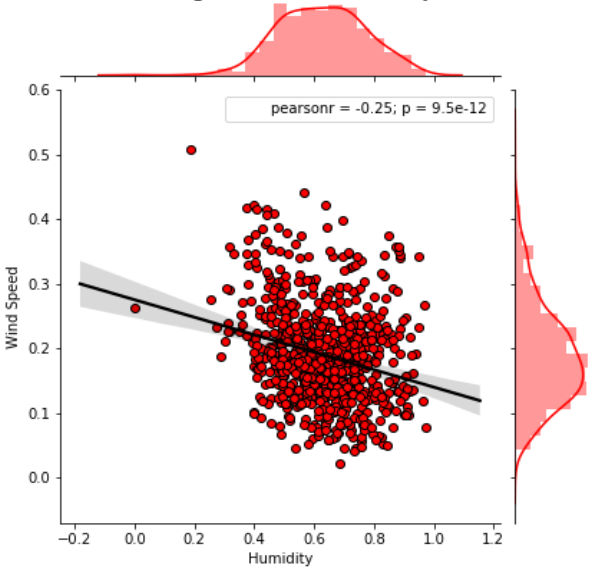
Feeling Temperature V/S Casual Users



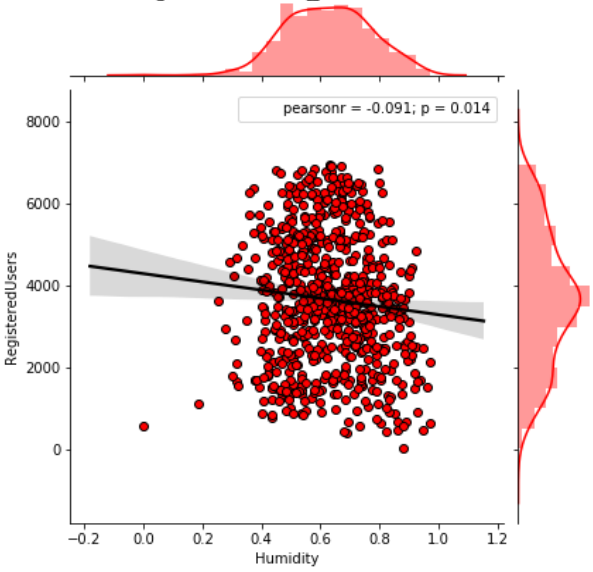
Feeling Temperature V/S Registered Users



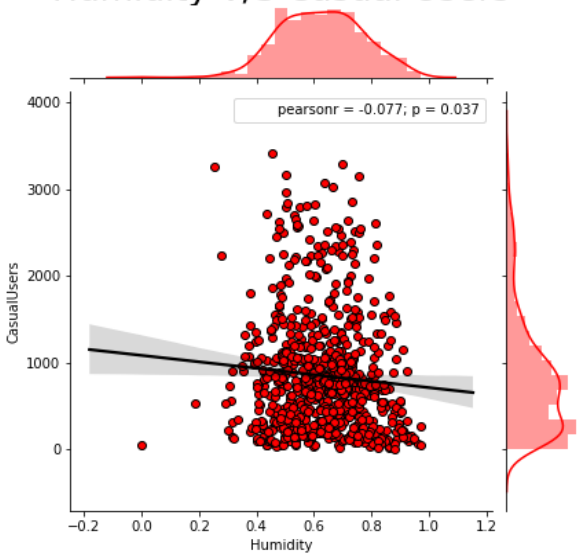
Humidity V/S Wind Speed



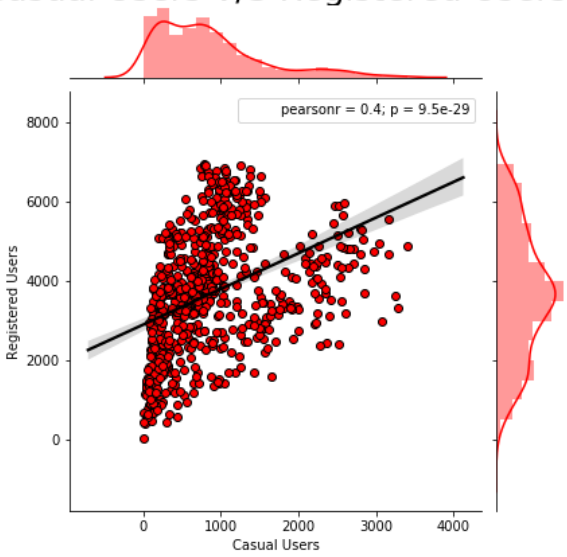
Humidity V/S Registered Users



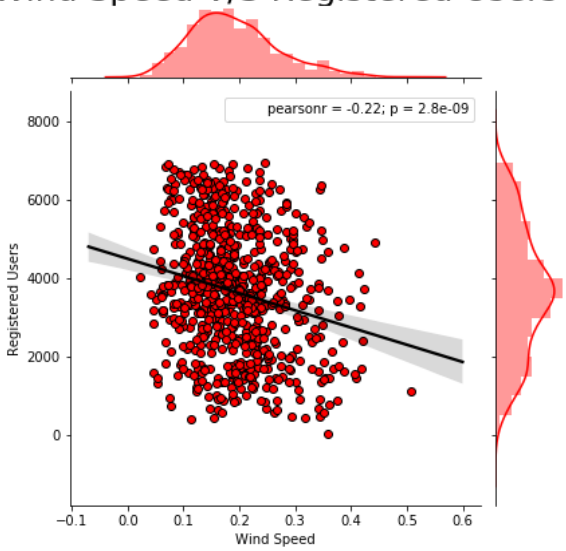
Humidity V/S Casual Users



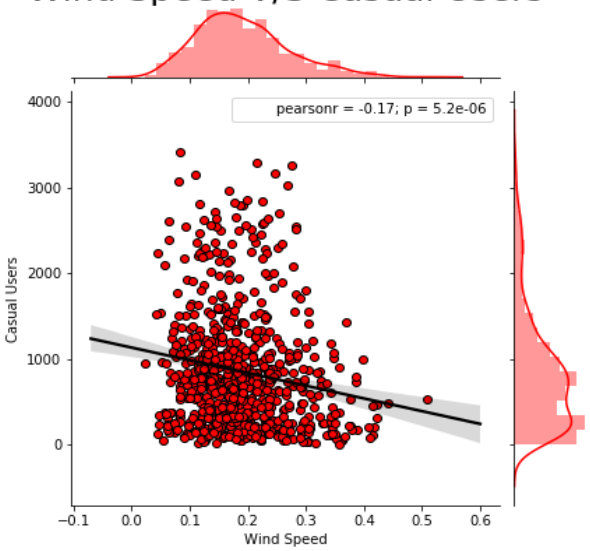
Casual Users V/S Registered Users



Wind Speed V/S Registered Users



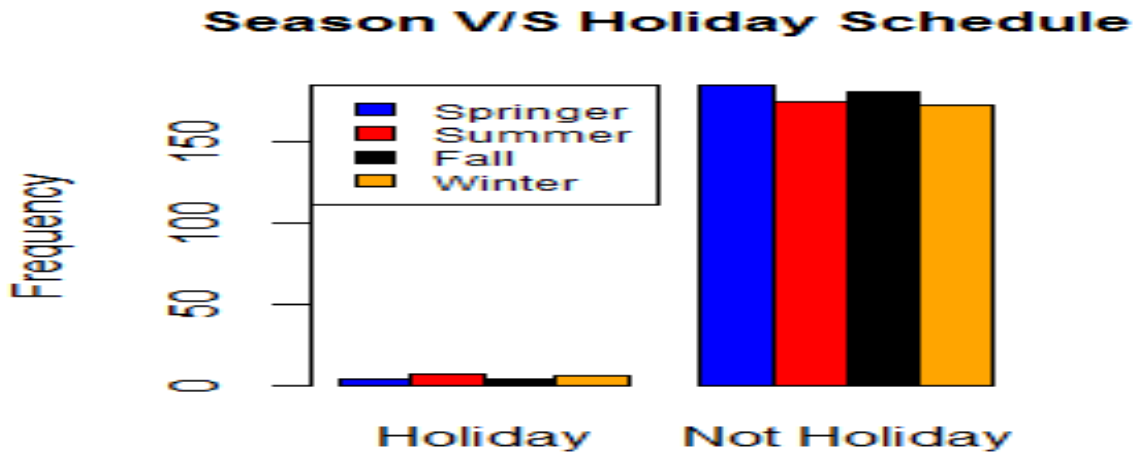
Wind Speed V/S Casual Users

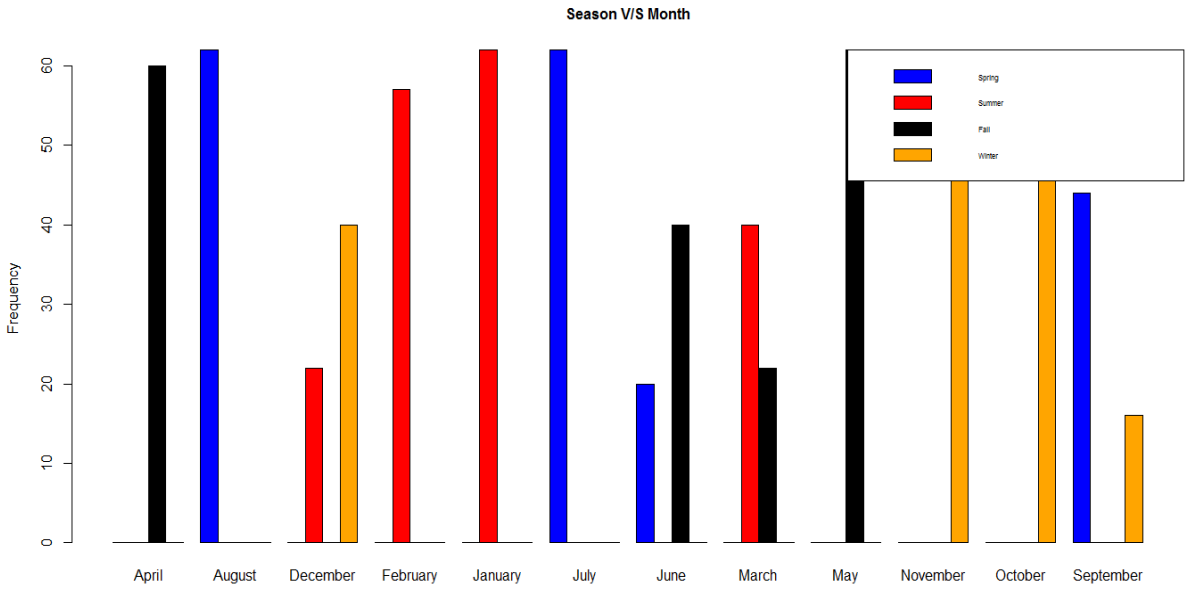


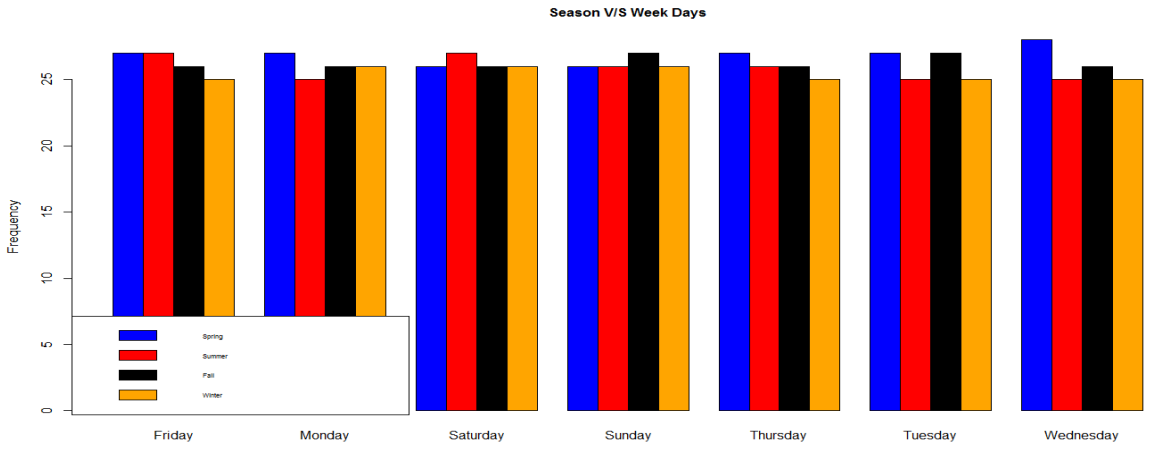
Pair Plot

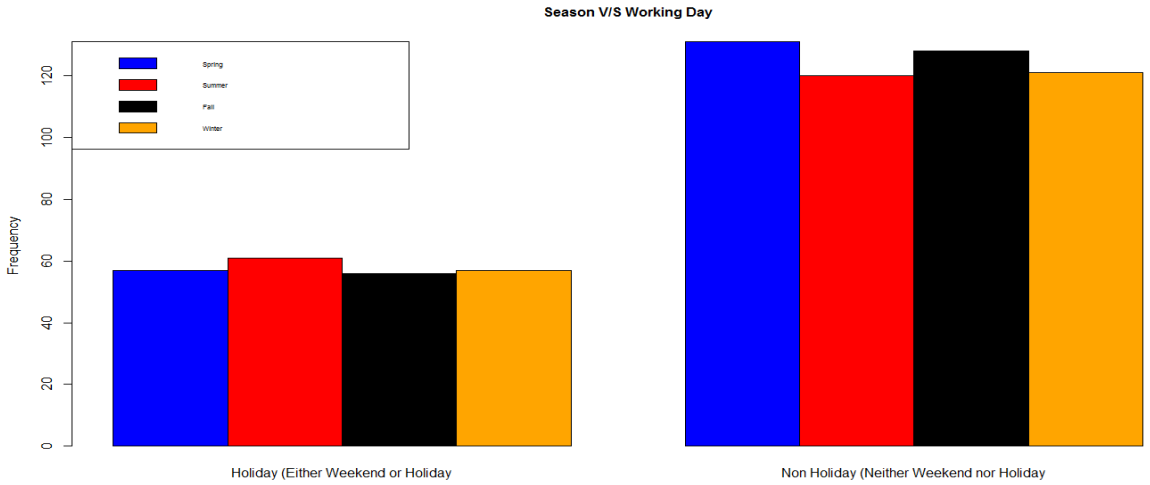
Pair plot is used to understand the best set of features to explain a relationship between two variables or to form the most separated clusters. It also helps to form some simple classification models by drawing some simple lines or make linear separation in our dataset. Pair plots for all our numerical variables are shown below:-

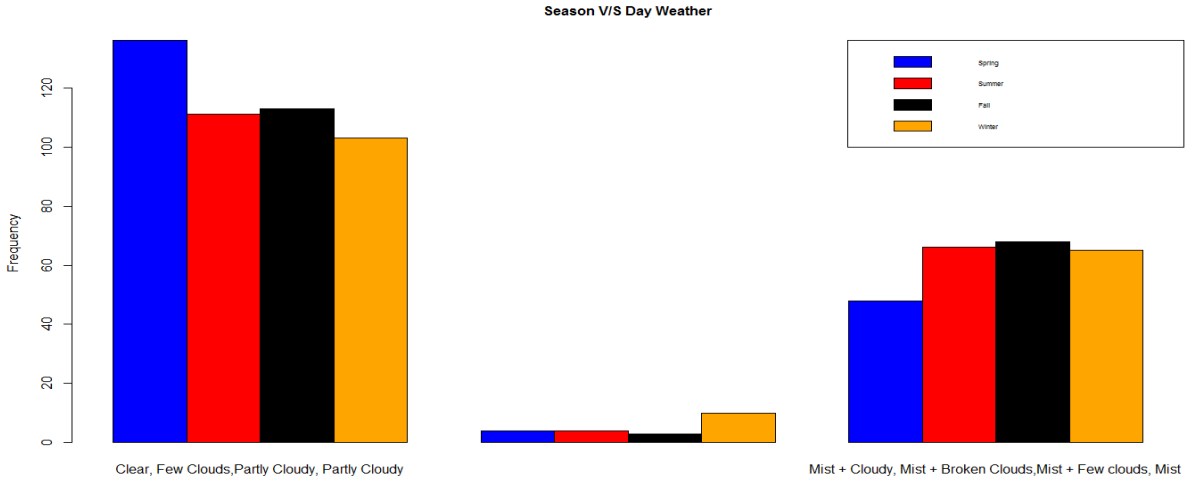
Categorical Variables

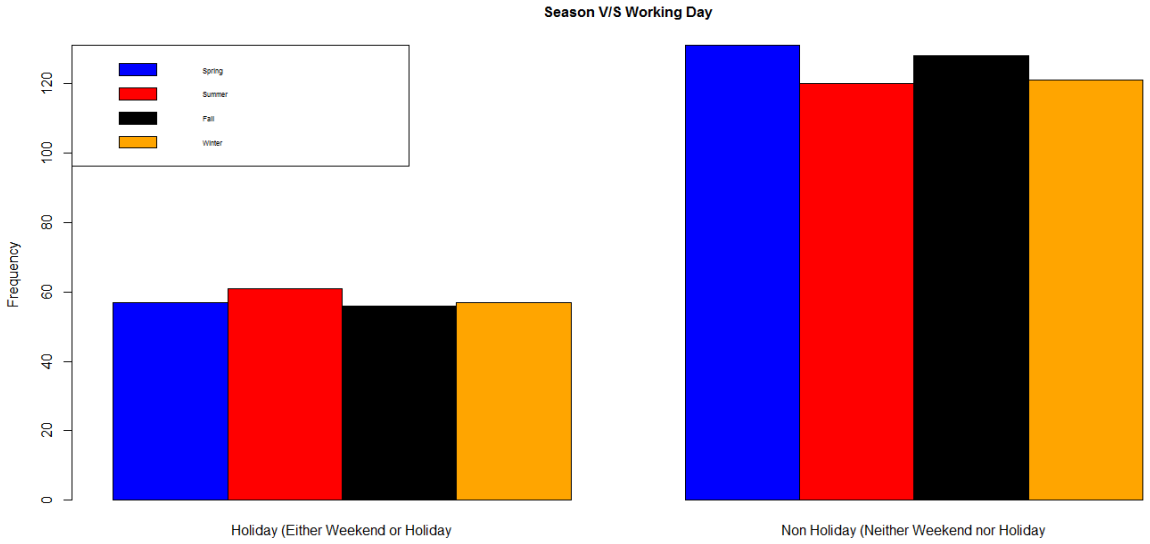


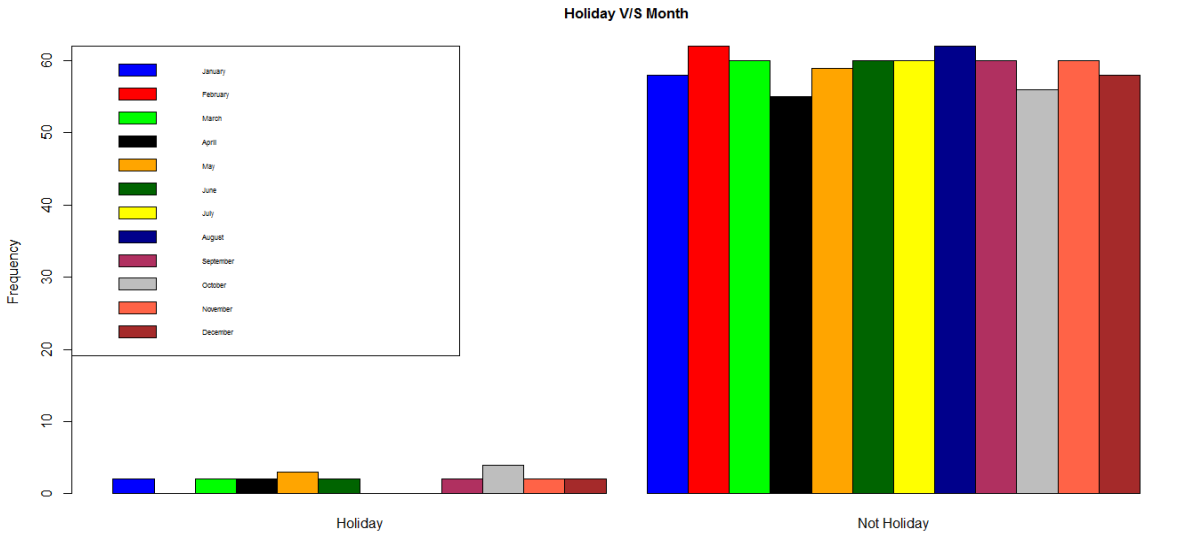


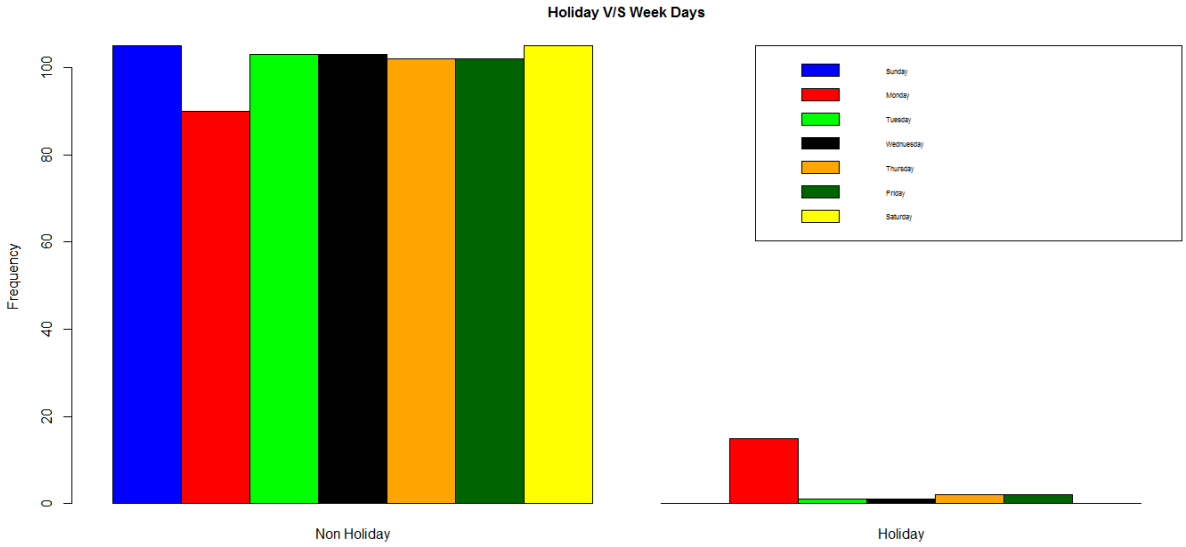


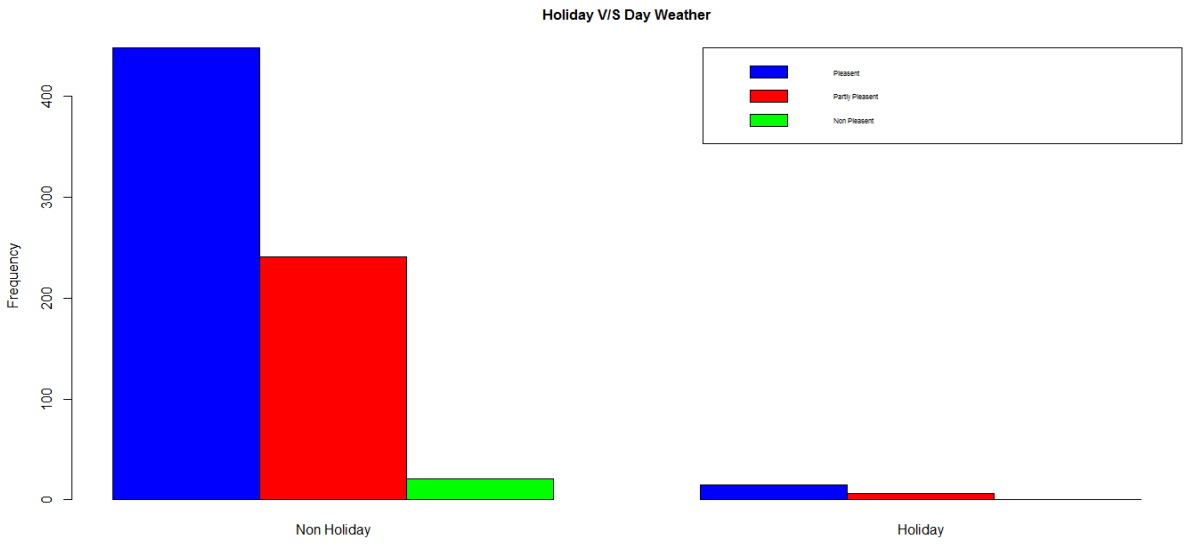

Season V/S Working Day
Spring
Summer
Fall
Winter
Frequency
0
20
40
60
80
100
120
Holiday (Either Weekend or Holiday
Non Holiday (Neither Weekend nor Holiday

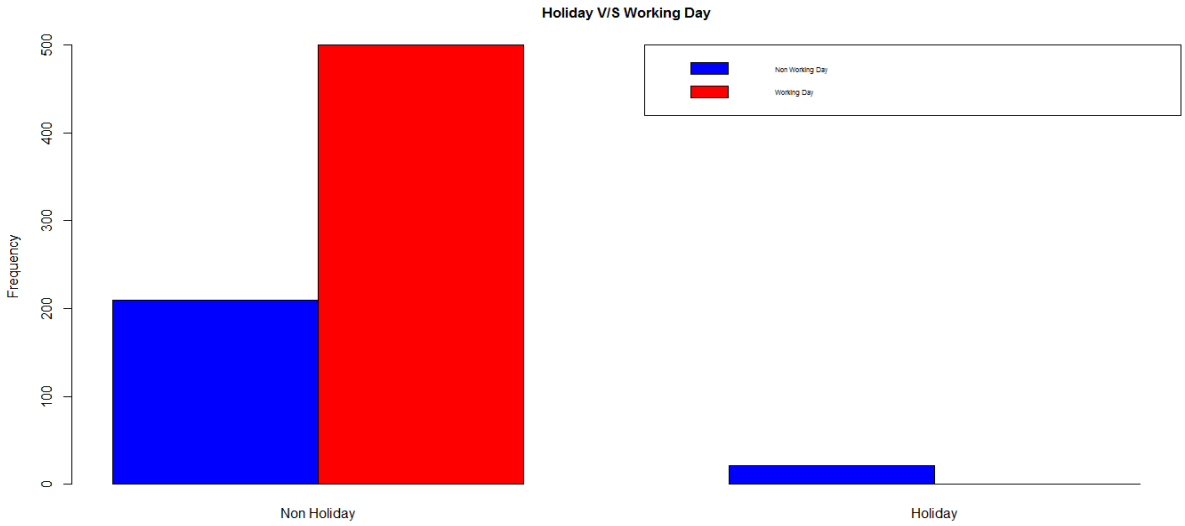

Season V/S Day Weather
Spring
Summer
Fall
Winter
Frequency
0
20
40
60
80
100
120
Clear, Few Clouds,Partly Cloudy, Partly Cloudy
Mist + Cloudy, Mist + Broken Clouds,Mist + Few clouds, Mist

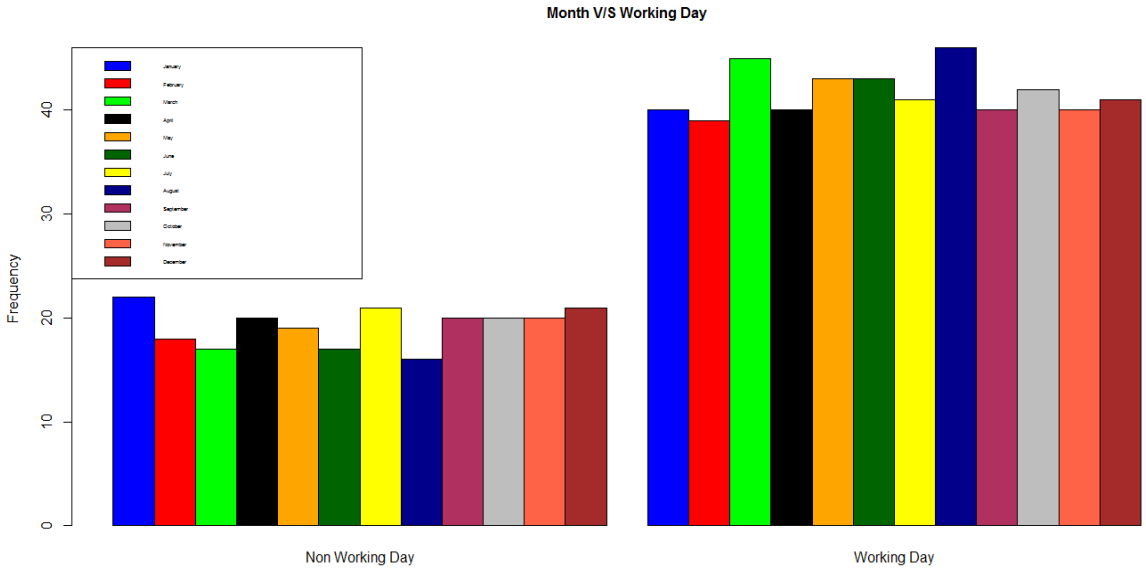

Season V/S Working Day
Spring
Summer
Fall
Winter
Frequency
0
20
40
60
80
100
120
Holiday (Either Weekend or Holiday
Non Holiday (Neither Weekend nor Holiday

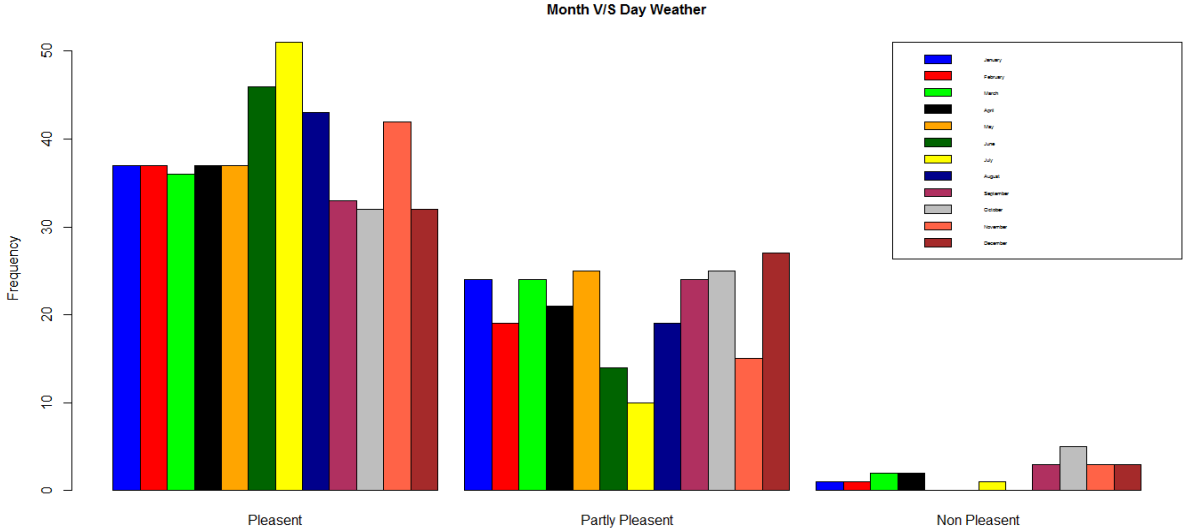

Holiday V/S Month
Frequency
0
10
20
30
40
50
60
January
February
March
April
May
June
July
August
September
October
November
December
Holiday
Not Holiday

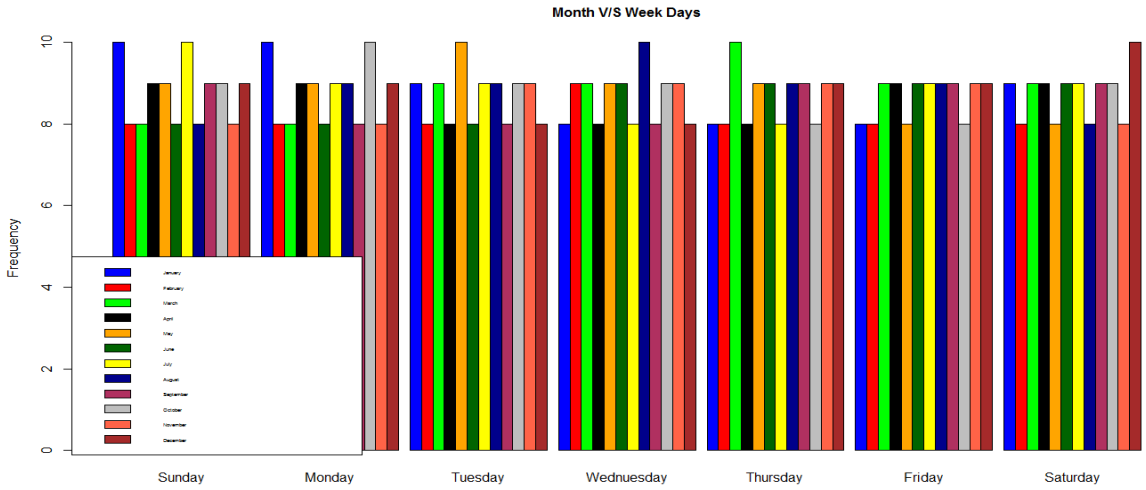

Holiday V/S Week Days
Frequency
0
20
40
60
80
100
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Non Holiday
Holiday

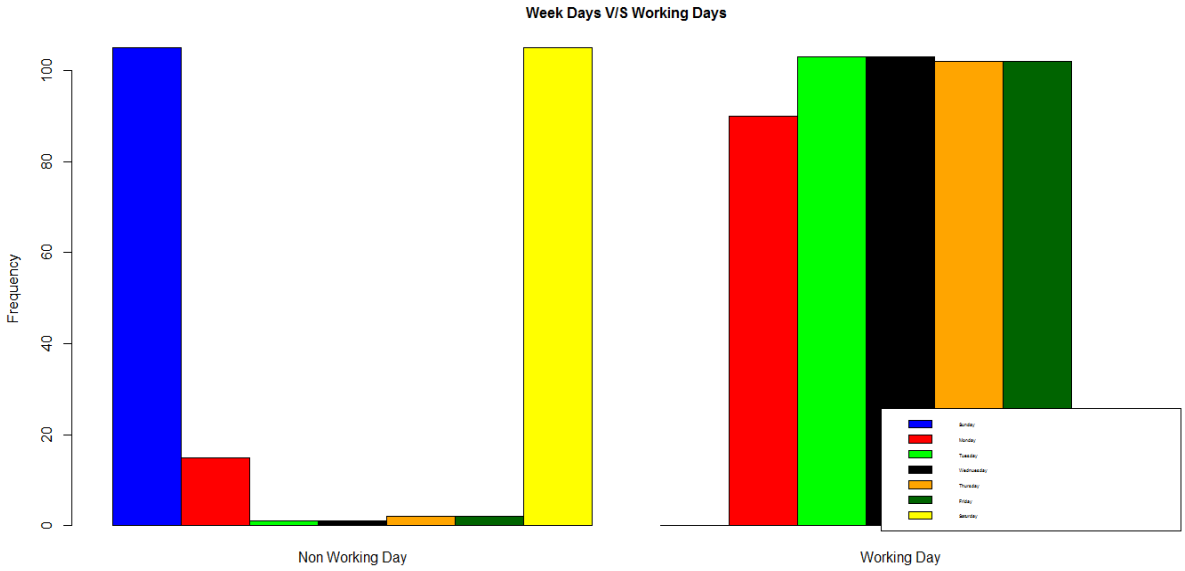

Holiday V/S Day Weather
Frequency
0
100
200
300
400
Pleasent
Partly Pleasent
Non Pleasent
Non Holiday
Holiday

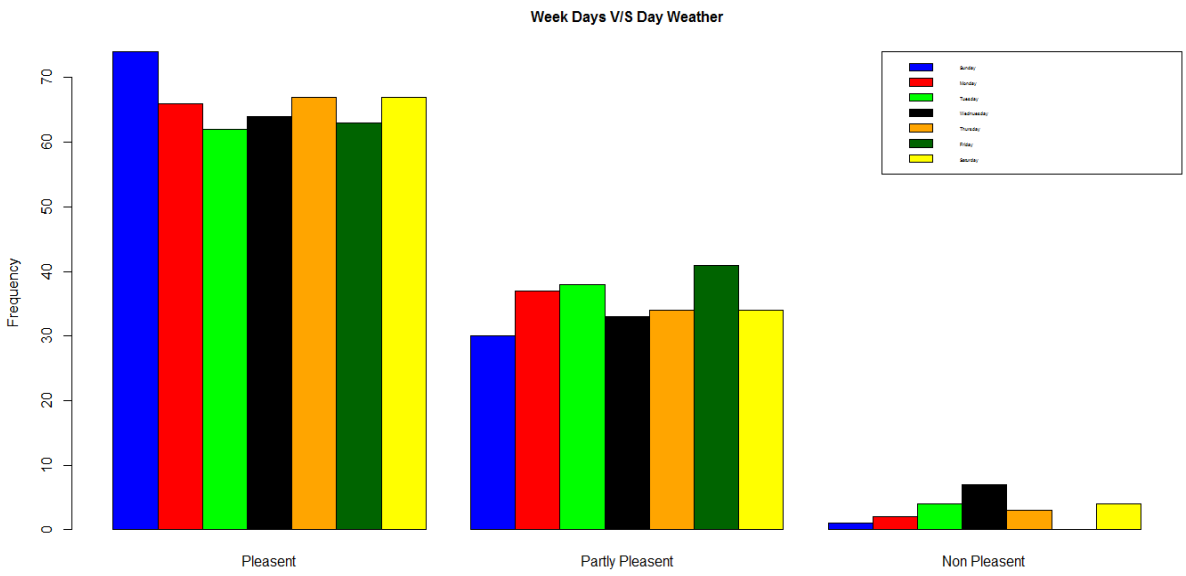

Holiday V/S Working Day
Frequency
0
100
200
300
400
500
Non Working Day
Working Day
Non Holiday
Holiday

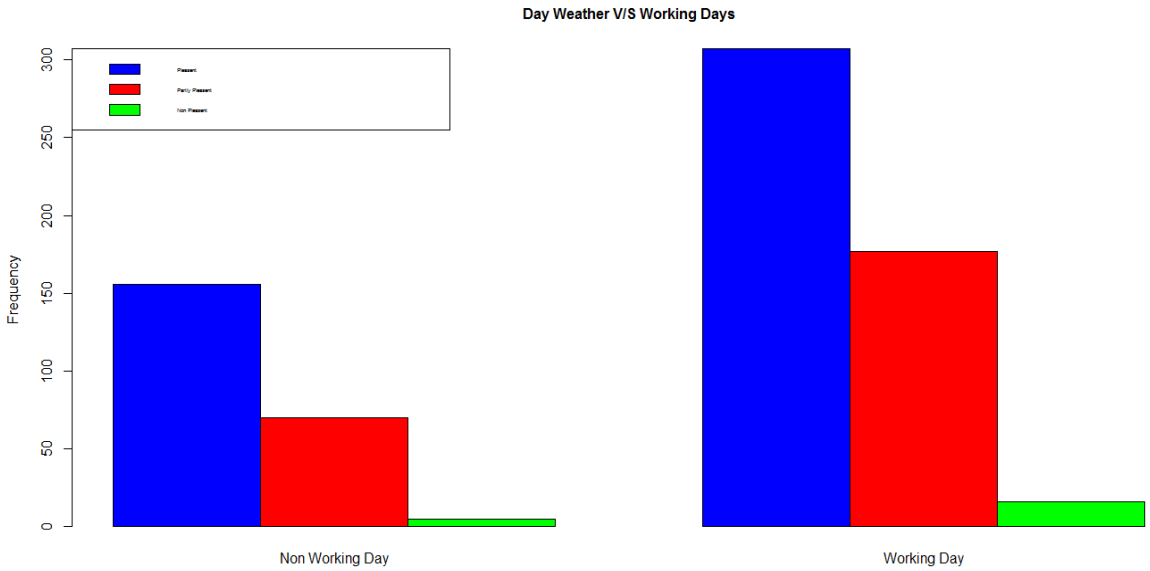

Month V/S Working Day
Frequency
0
10
20
30
40
January
February
March
April
May
June
July
August
September
October
November
December
Non Working Day
Working Day


Month V/S Day Weather
Frequency
0
10
20
30
40
50
Pleasent
Partly Pleasent
Non Pleasent


Month V/S Week Days
Frequency
0
2
4
6
8
10
Sunday
Monday
Tuesday
Wednuesday
Thursday
Friday
Saturday

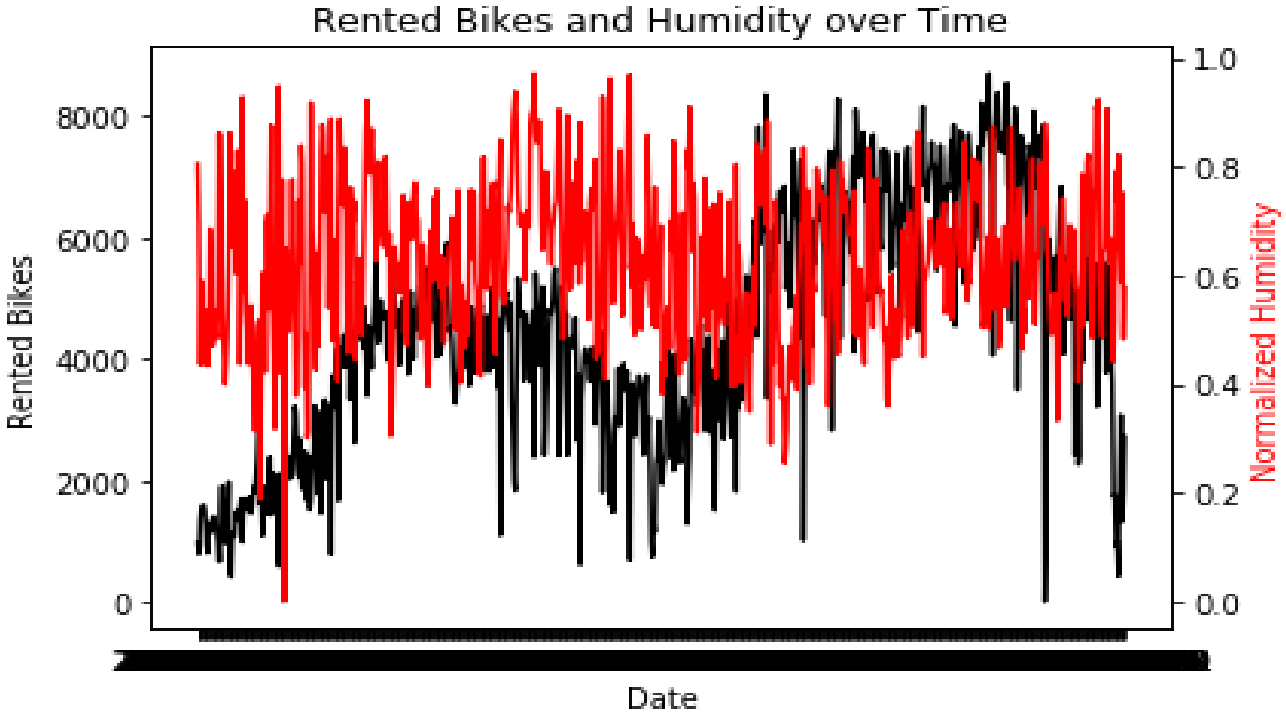

Week Days V/S Working Days
Frequency
0
20
40
60
80
100
Non Working Day
Working Day

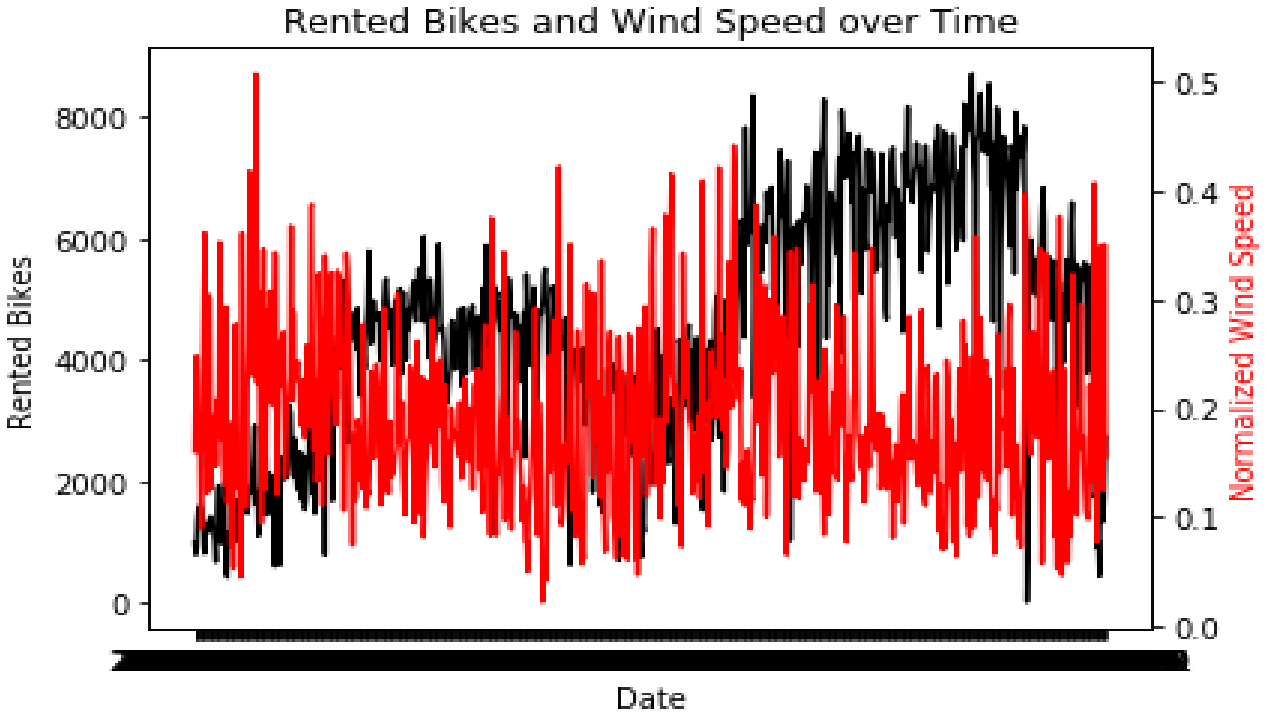

Week Days V/S Day Weather
Frequency
0
10
20
30
40
50
60
70
Pleasent
Partly Pleasent
Non Pleasent

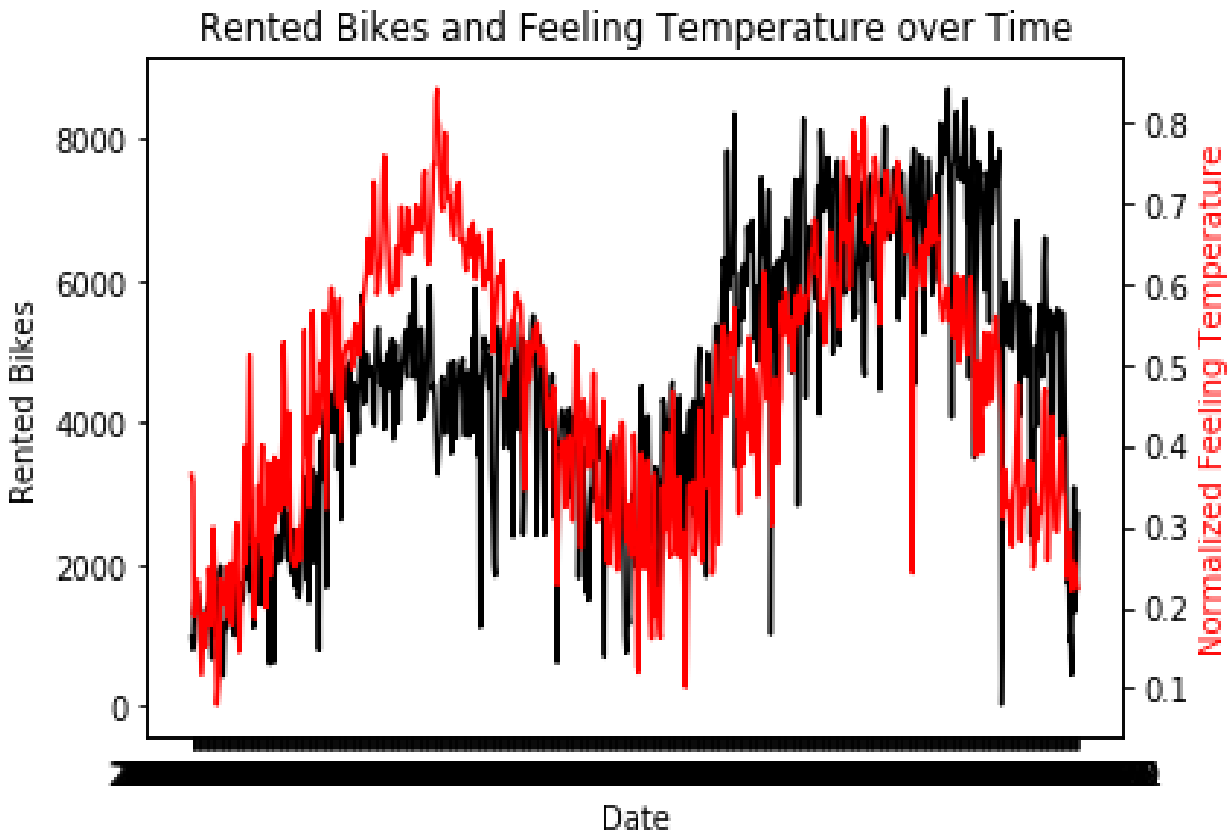

Day Weather V/S Working Days
Frequency
0
50
100
150
200
250
300
Non Working Day
Working Day

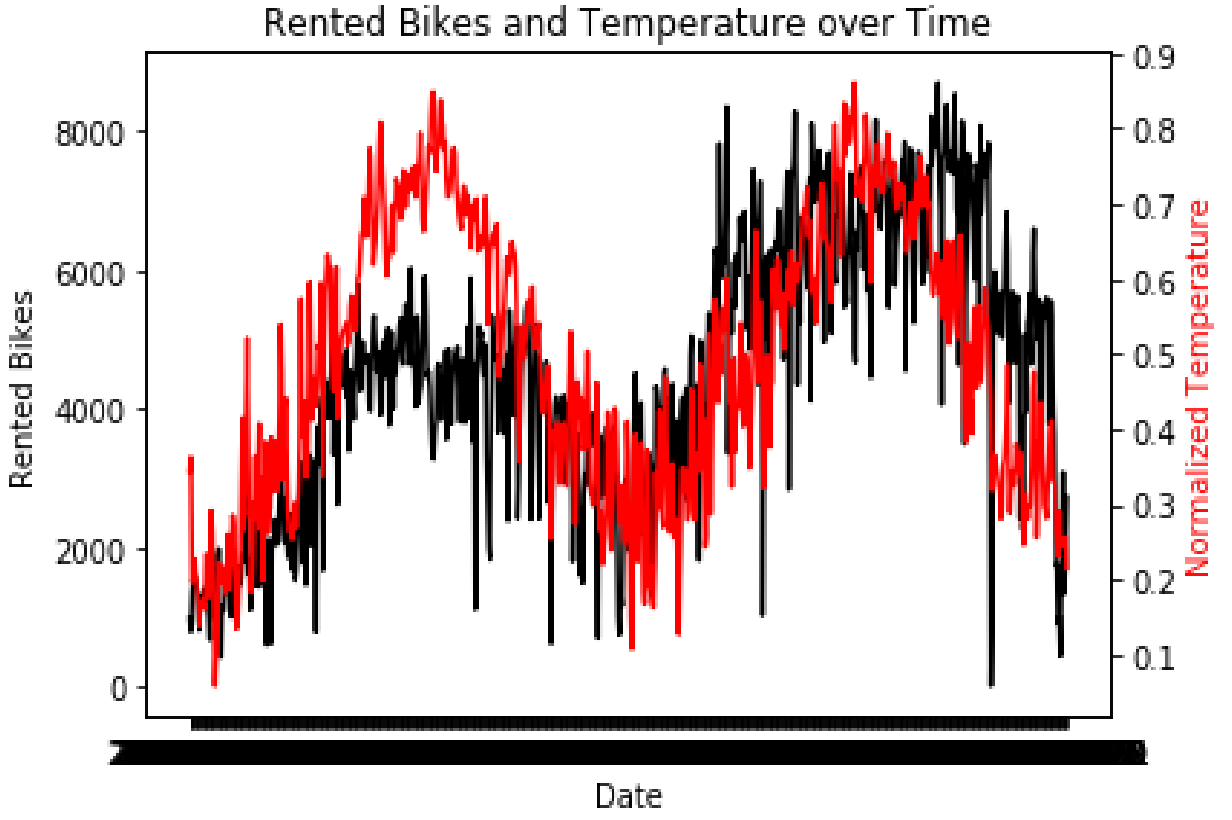
Multivariate Analysis

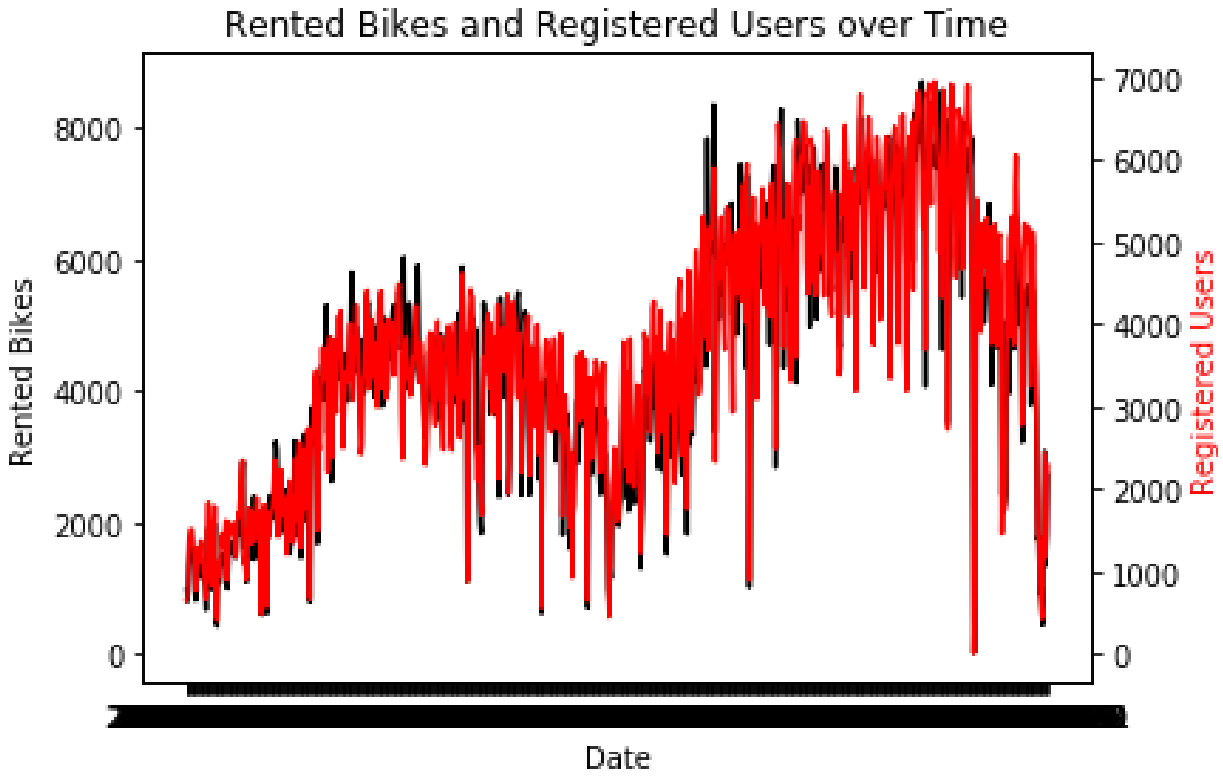
Multivariate analysis is conceptualized by tradition as the statistical study of experiments in which multiple measurements are made on each experimental unit and for which the relationship among multivariate measurements and their structure are important to the experiment's understanding. In order to understand the relationship between Rented Bikes and other numerical variables in the dataset over a period of time is shown below:-

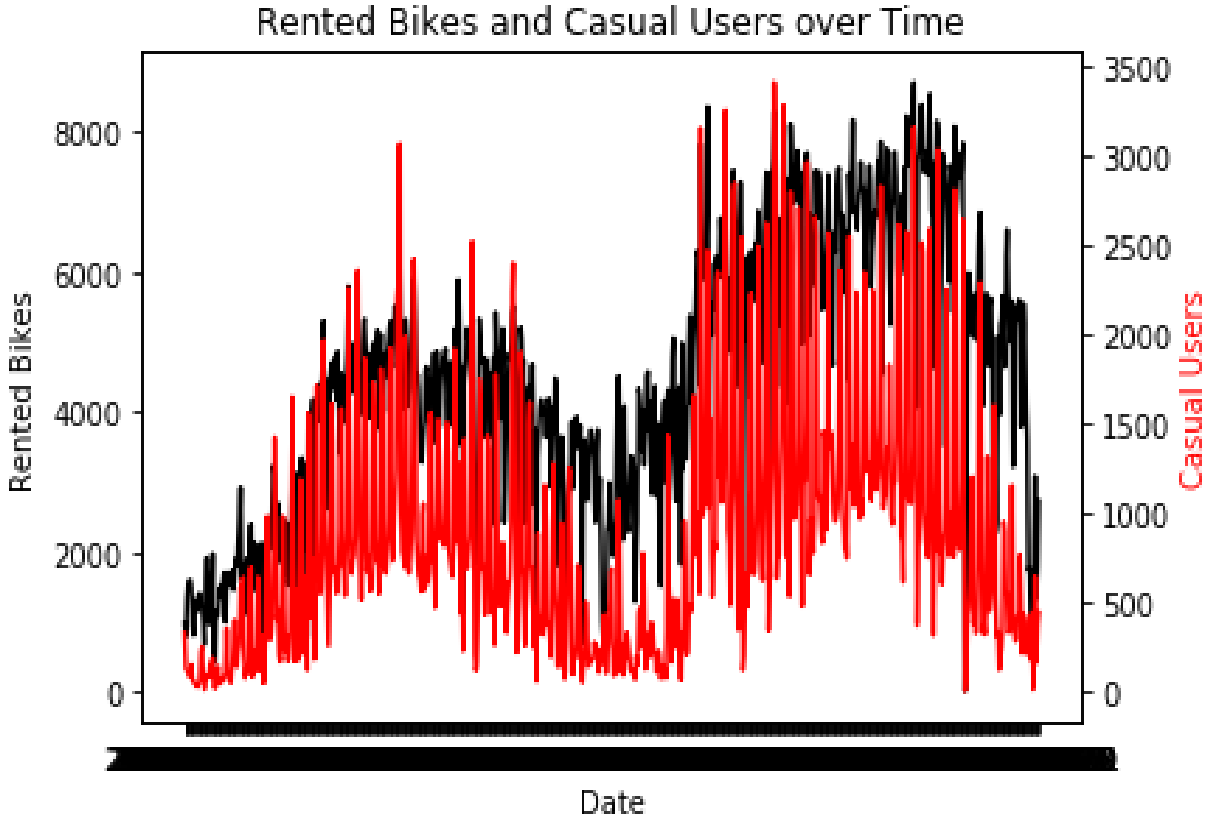





Rented Bikes and Feeling Temperature over Time
Rented Bikes
8000
6000
4000
2000
0
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Normalized Feeling Temperature
Date


Rented Bikes and Temperature over Time
Rented Bikes
8000
6000
4000
2000
0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Normalized Temperature
Date


Rented Bikes and Registered Users over Time
Rented Bikes
Registered Users
Date
0
2000
4000
6000
8000
0
1000
2000
3000
4000
5000
6000
7000


Rented Bikes and Casual Users over Time
Rented Bikes
Casual Users
Date
0
2000
4000
6000
8000
0
500
1000
1500
2000
2500
3000
3500

Missing Values

In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data.

Missing data can occur because of nonresponse: no information is provided for one or more items or for a whole unit ("subject"). Some items are more likely to generate a nonresponse than others

In our dataset, there is no missing value present. A table below will throw light on our dataset:-

Variables	Missing Values
Index	0
Date	0
Season	0
Year	0
Month	0
Holiday	0
WeekDay	0
WorkingDay	0
DayWeather	0
Temperature	0
FeelingTemperature	0
Humidity	0
WindSpeed	0
CasualUsers	0
RegisteredUsers	0
RentedBikes	0

Outliers

An Outlier is a rare chance of occurrence within a given data set. In Statistics and Data Science, an Outlier is an observation point that is distant from other observations. An Outlier may be due to variability in the measurement or it may indicate experimental error.

Outliers, being the most extreme observations, may include the sample maximum or sample minimum, or both, depending on whether they are extremely high or low. However, the sample maximum and minimum are not always outliers because they may not be unusually far from other observations.

While outliers are attributed to a rare chance and may not necessarily be fully explainable, Outliers in data can distort predictions and affect the accuracy, if you don't detect and handle them.

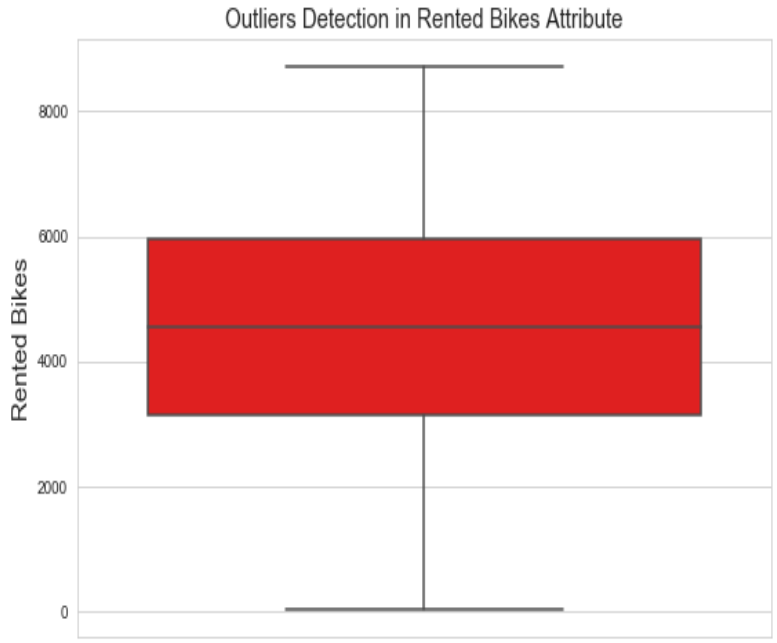
The contentious decision to consider or discard an outlier needs to be taken at the time of building the model. Outliers can drastically bias/change the fit estimates and predictions.

Detecting and Removing Outliers

Mostly outliers are present in the continuous variables and box plot method is best and easy way to detect and remove outliers. Moreover, our dataset contains categorical variables that are already encoded so we will perform outlier detections only on continuous variables.

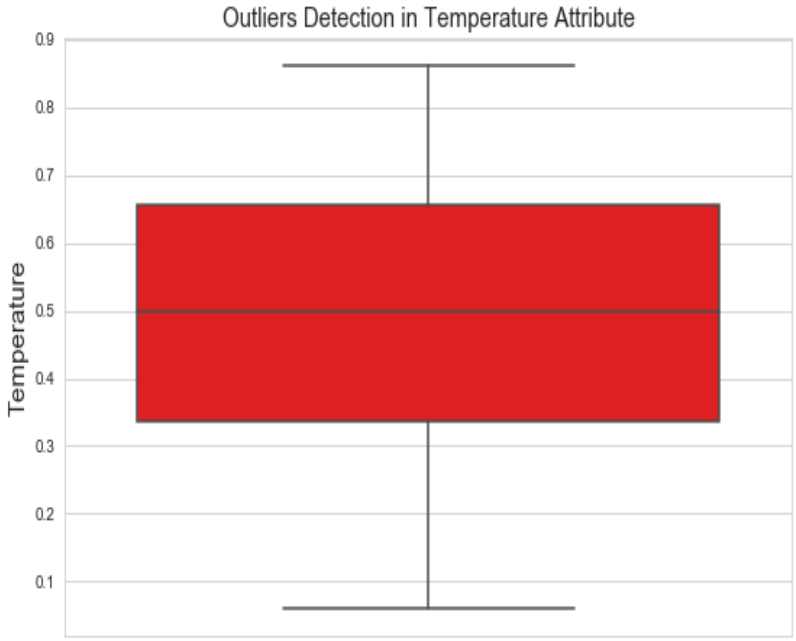
Rented Bikes

Rented Bikes is our target variable and it does not contain any outlier in it. Below visualization is showing that this variable is outlier free.



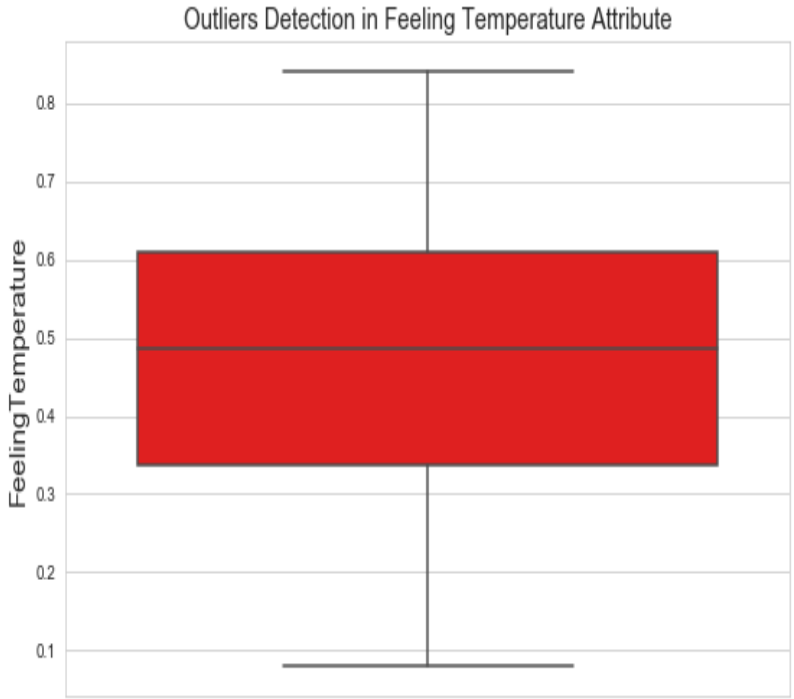
Temperature

Temperature is a predictor variable and it does not contain any outlier in it. Below visualization is showing that this variable is outlier free.



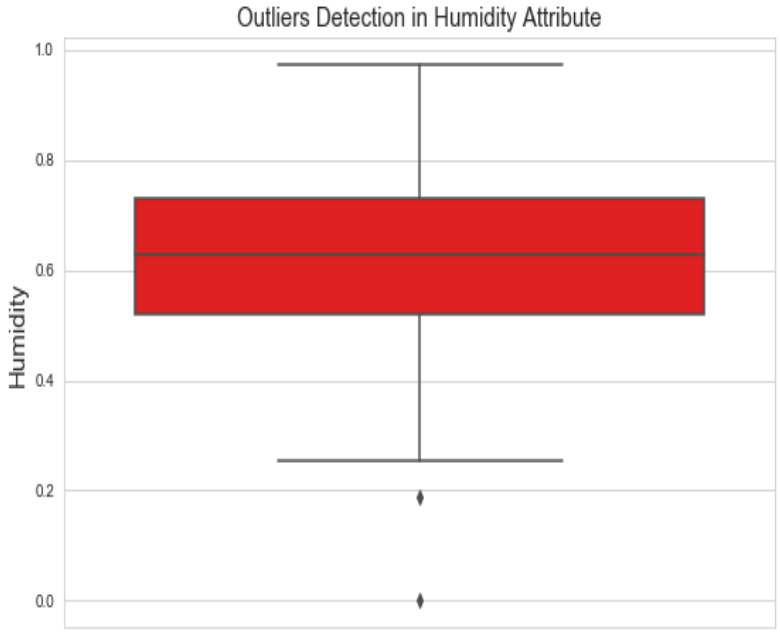
Feeling Temperature

Feeling temperature is a predictor variable and it does not contain any outlier in it. Below visualization is showing that this variable is outlier free.

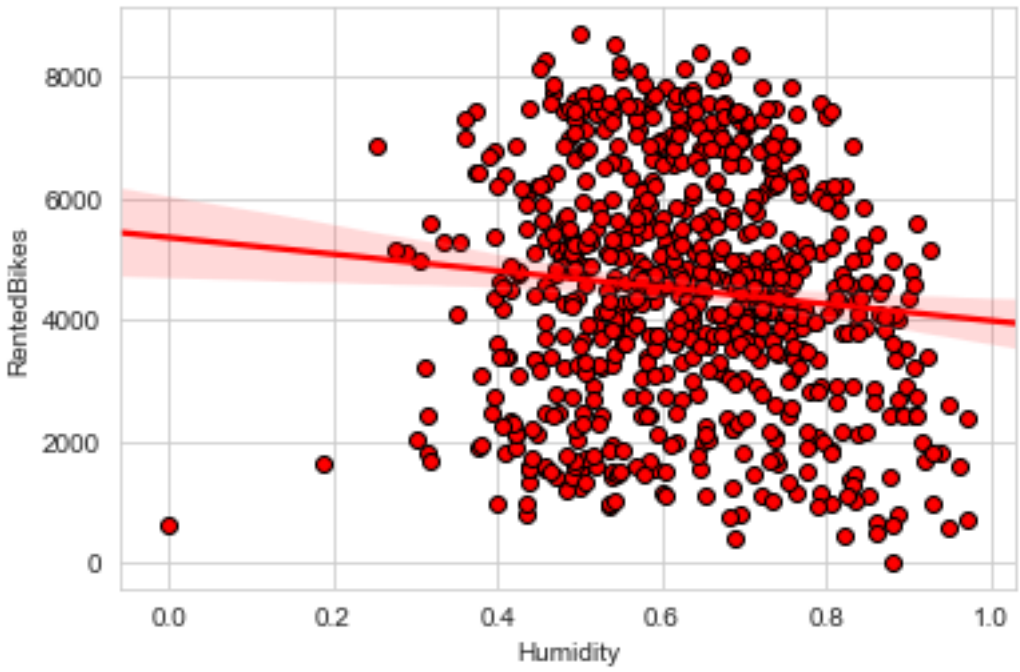


Humidity

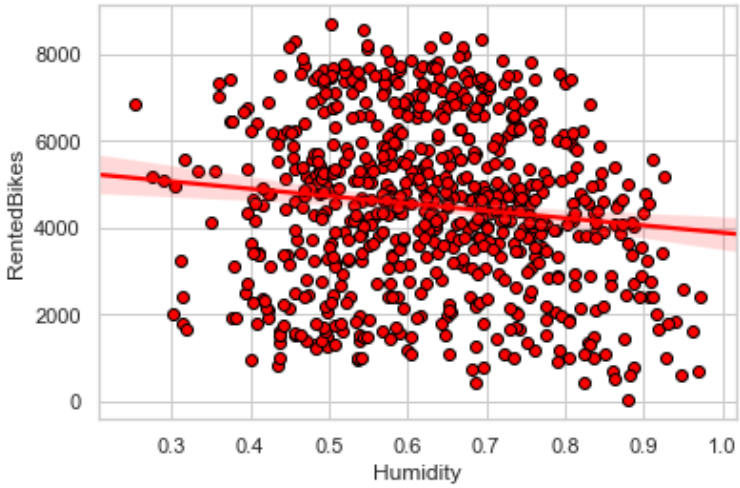
Humidity is a predictor variable in our dataset and outliers are present in this variable. Below visualization is clearly showing that outliers exist in this variable.



Before removing outliers from our dataset we have check relationship of humidity variable with our target variable. Below is a plot which helps us to understand more about relationship of these two variables:-



After removing outliers, extreme values are removed and there is a change in correlation values which are shown below:-

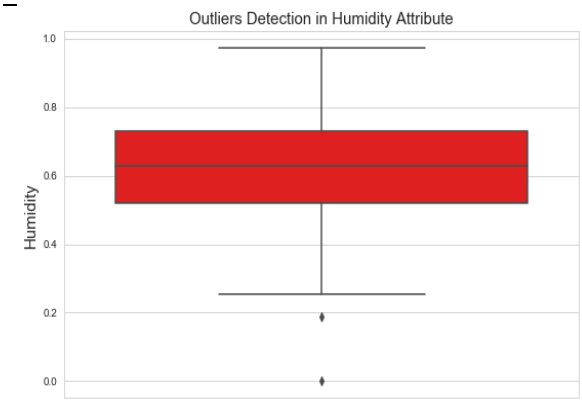


Change is correlation values are shown below:-

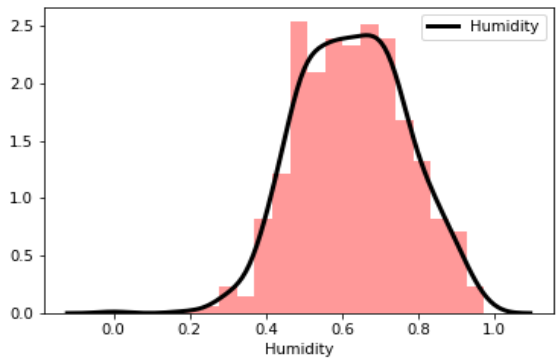
Correlation with Outliers	Correlation without Outliers
-0.10065856213715527	-0.12203925129240463

A change of distribution after removal of outliers is shown below:-

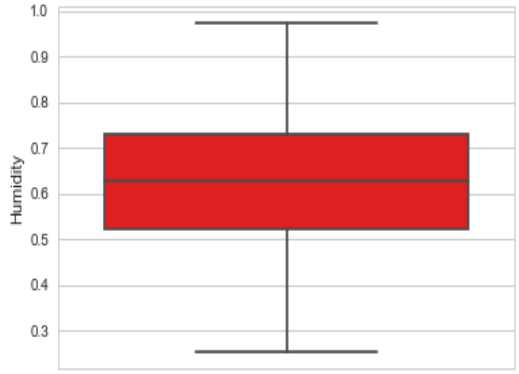
Box Plot with Outliers



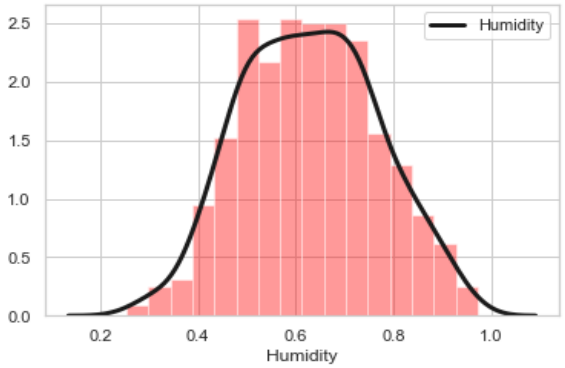
Distribution with Outliers



Box Plot with Outliers

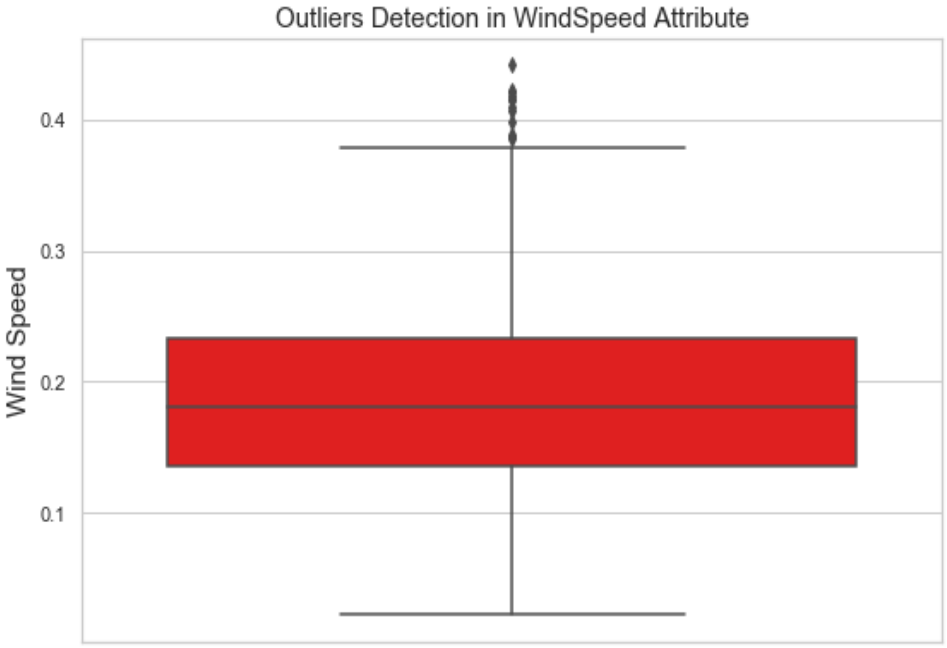


Distribution without Outliers

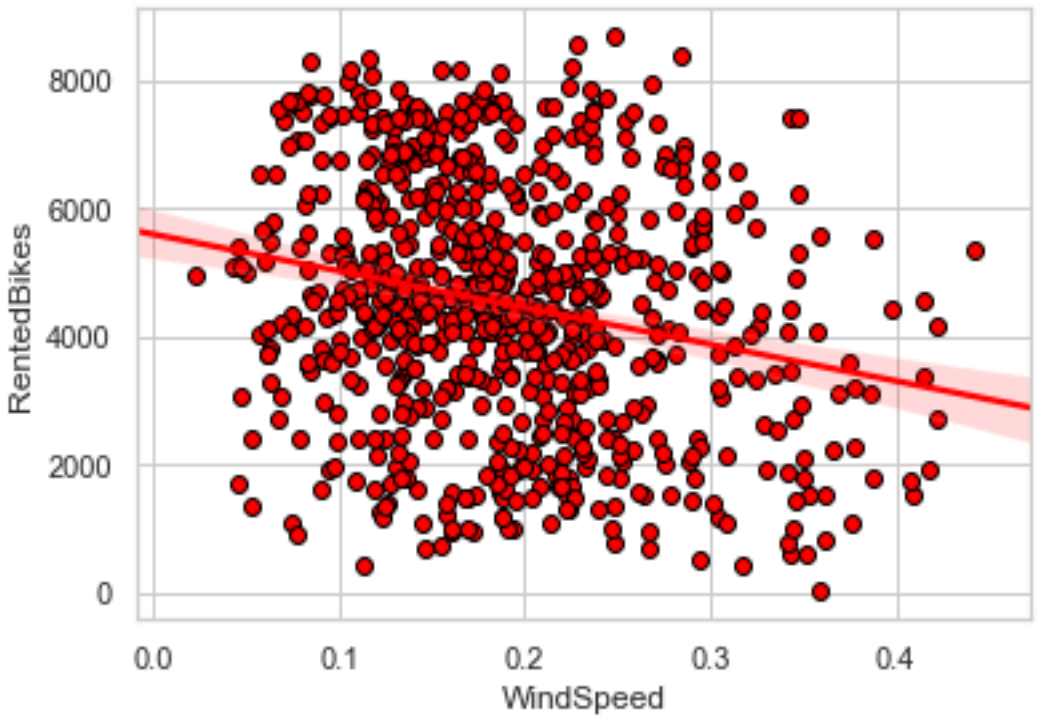


Wind Speed

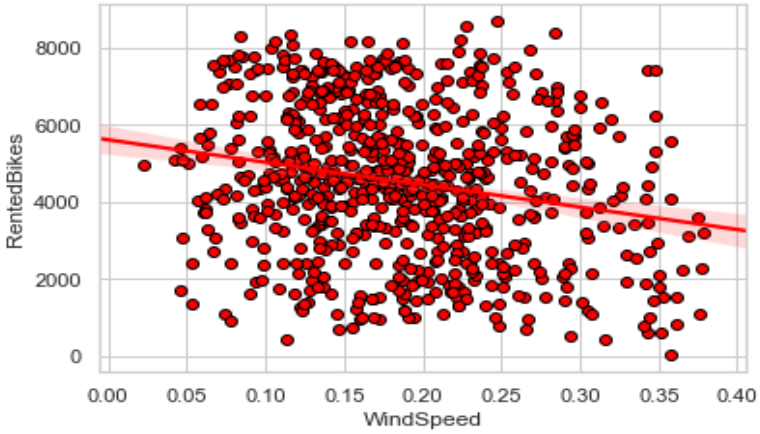
Wind Speed is a predictor variable in our dataset and there are outliers present in this variable. Below visualization is clearly showing that outliers exist in this variable.



Before removing outliers from our dataset we have check relationship of wind speed variable with our target variable. Below is a plot which helps us to understand more about relationship of these two variables:-



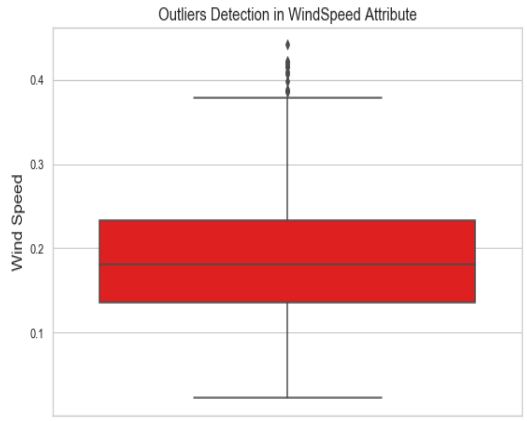
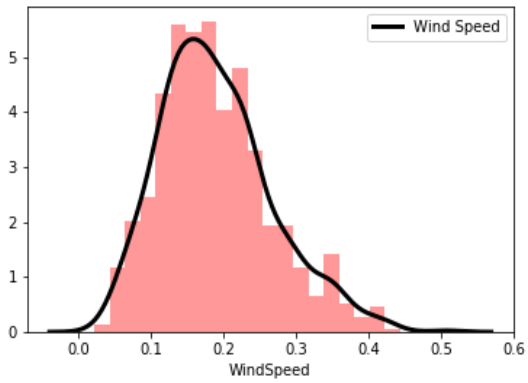
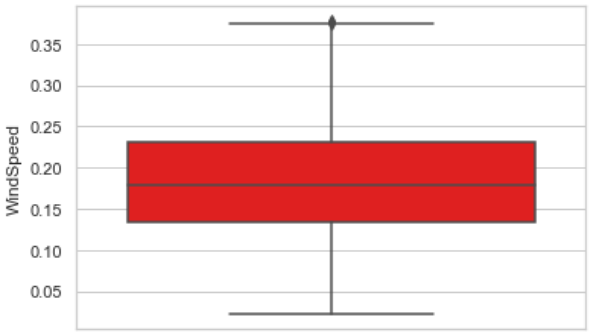
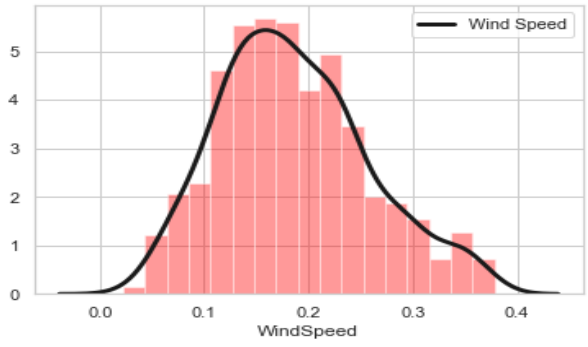
After removing outliers, extreme values are removed and there is a change in correlation values which are shown below:-



Change is correlation values are shown below:-

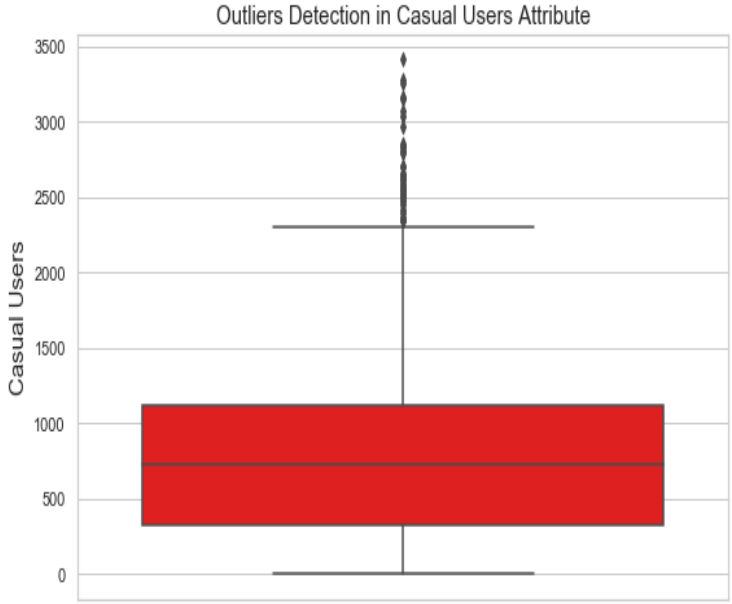
Correlation with Outliers	Correlation without Outliers
-0.2274047531235313	-0.2161932882927355

A change of distribution after removal of outliers is shown below:-

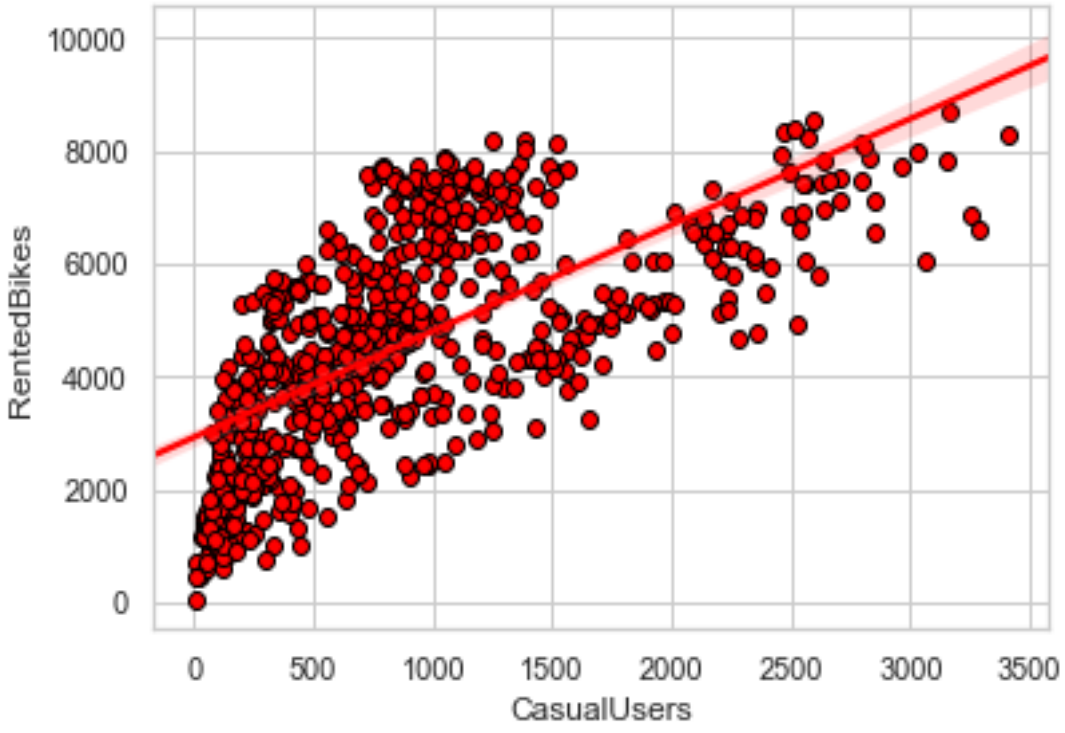
Box Plot with Outliers	Distribution with Outliers
	
Box Plot without Outliers	**Distribution without Outliers**
	

Casual Users

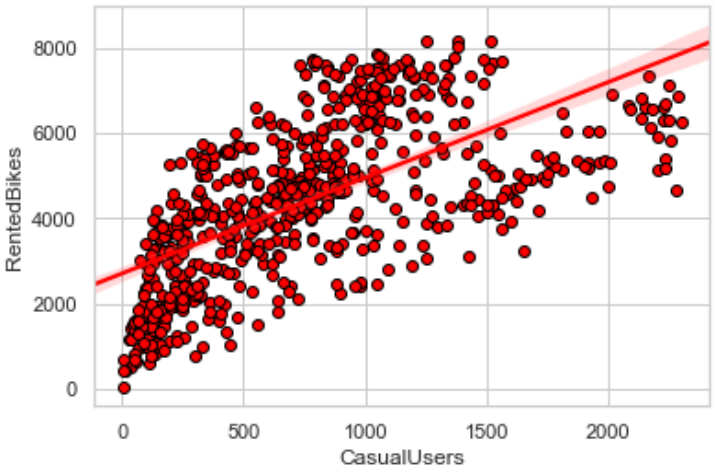
Casual Users is a predictor variable in our dataset and there are outliers present in this variable. Below visualization is clearly showing that outliers exist in this variable.



Before removing outliers from our dataset we have check relationship of casual user's variable with our target variable. Below is a plot which helps us to understand more about relationship of these two variables:-



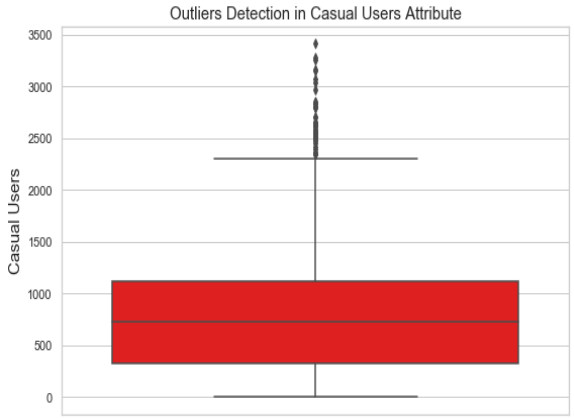
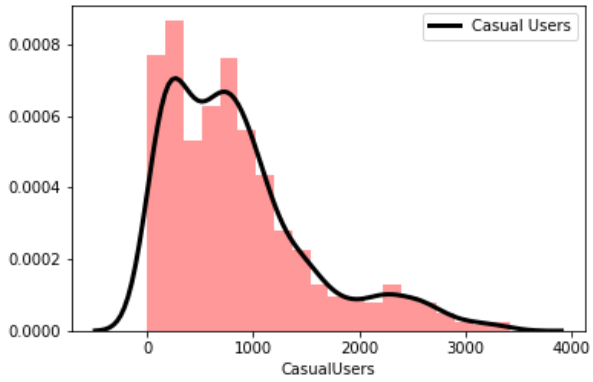
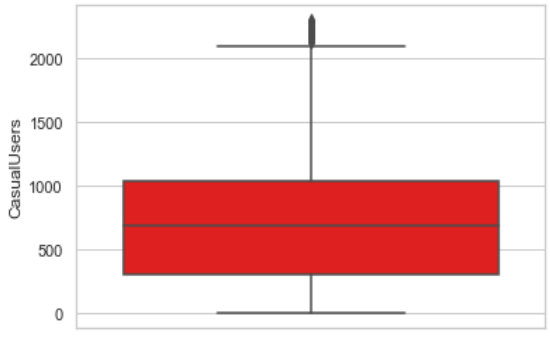
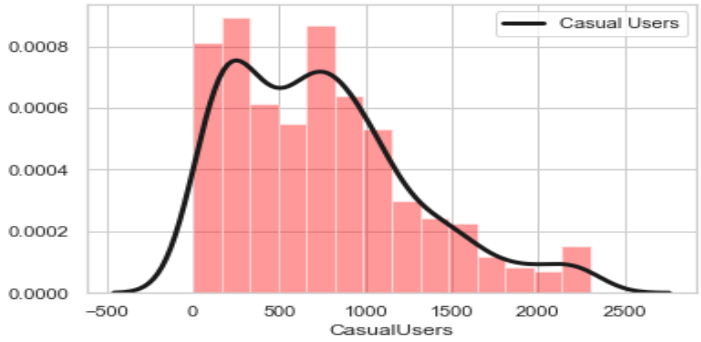
After removing outliers, extreme values are removed and there is a change in correlation values which are shown below:-



Change is correlation values are shown below:-

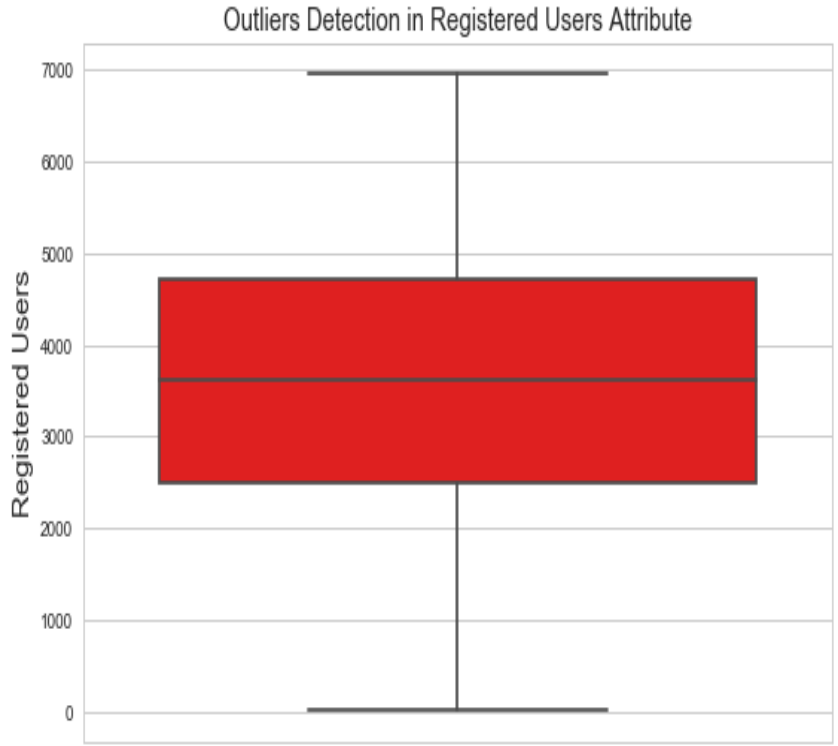
Correlation with Outliers	Correlation without Outliers
0.9999999999999998	0.6400803033164671

A change of distribution after removal of outliers is shown below:-

Box Plot with Outliers	Distribution with Outliers
	
Box Plot without Outliers	**Distribution without Outliers**
	

Registered Users

Registered Users is a predictor variable in our dataset and there is no outliers present in this variable. Below visualization is showing a box plot visualized on Registered user variable has no outlier:-



Feature Selection

Machine learning works on a simple rule – if we put garbage in, we will only get garbage to come out.

This becomes even more important when the number of features is very large. We need not use every feature at our disposal for creating an algorithm. We can assist our algorithm by feeding in only those features that are really important. Feature subsets giving better results than complete set of feature for the same algorithm or – "Sometimes, less is better!".

We should consider the selection of feature for model based on below criteria:-

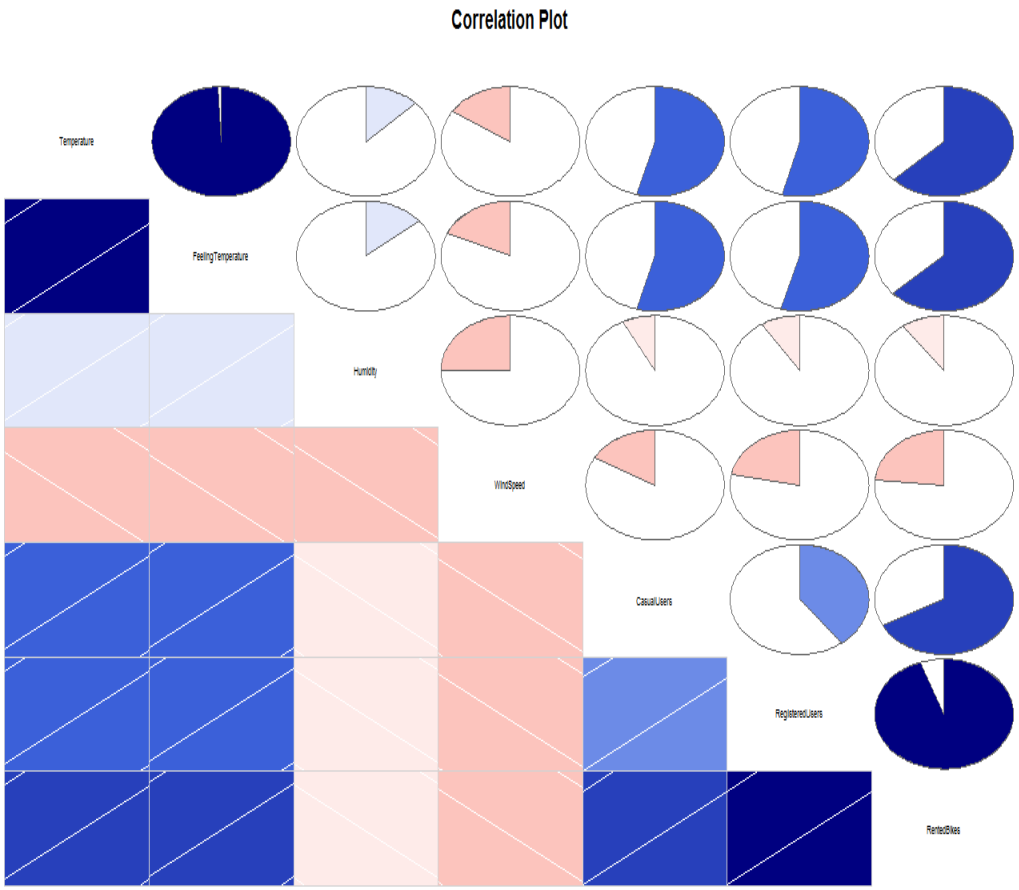
1. The relationship between two independent variable should be less and
2. The relationship between Independent and target variables should be high.

Below figure shows a graphical display of a correlation matrix, called a correlogram. The cells of the matrix are coloured to show the correlation value.

	Temperature	FeelingTemperature	Humidity	WindSpeed	CasualUsers	RegisteredUsers	RentedBikes
Temperature	1.0	0.99	0.13	-0.16	0.54	0.54	0.63
FeelingTemperature	0.99	1.0	0.14	-0.18	0.54	0.54	0.63
Humidity	0.13	0.14	1.0	-0.25	-0.077	-0.091	-0.1
WindSpeed	-0.16	-0.18	-0.25	1.0	-0.17	-0.22	-0.23
CasualUsers	0.54	0.54	-0.077	-0.17	1.0	0.4	0.67
RegisteredUsers	0.54	0.54	-0.091	-0.22	0.4	1.0	0.95
RentedBikes	0.63	0.63	-0.1	-0.23	0.67	0.95	1.0

In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model

Since our target variable is continuous variable so we will only going to check correlation values. Below diagram is showing an correlation plot of numeric variables:



From the above correlation plot, I have analyzed that Temperature and Feeling Temperature variable are highly correlated. So I have considered “Temperature” variable for training my

model and “Feeling Temperature” variable will not be considered as a feature while training modals.

Besides this correlation between Rented bikes and Humidity variable is very less and hence I will not be considering Humidity Variable while training my machine learning modals.

Feature Scaling

Feature scaling is done to reduce unwanted variation either within or between variables and to bring all of the variables into proportion with one another. I will use Normalization process to perform feature scaling. Formula for Normalization is given below:-

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Attribute before and after normalization are given below:-

Before Normalization	
CasualUsers	RegisteredUsers
331	654
131	670
120	1229
108	1454
82	1518
88	1518

After Normalization	
CasualUsers	RegisteredUsers
0.096538	0.091539
0.037852	0.093849
0.034624	0.17456
0.031103	0.207046
0.023474	0.216286
0.025235	0.216286

Modeling

Model Selection

In the case of this dataset we have to predict the count of total bike rented on basis of environmental and seasonal condition. The target variable here is a continuous variable and for a continuous variable we can use various Regression models. Trained model having less error rate and more accuracy will be our final model. Different machine learning methods which will be used to train our final model are mentioned below:-

1. Decision Tree Regression Model
2. Random Forest Model
3. Liner Regression Model

Final model with will be with the higher accuracy which we will able to decide at the end of the modelling process.

Decision Tree Regression Model

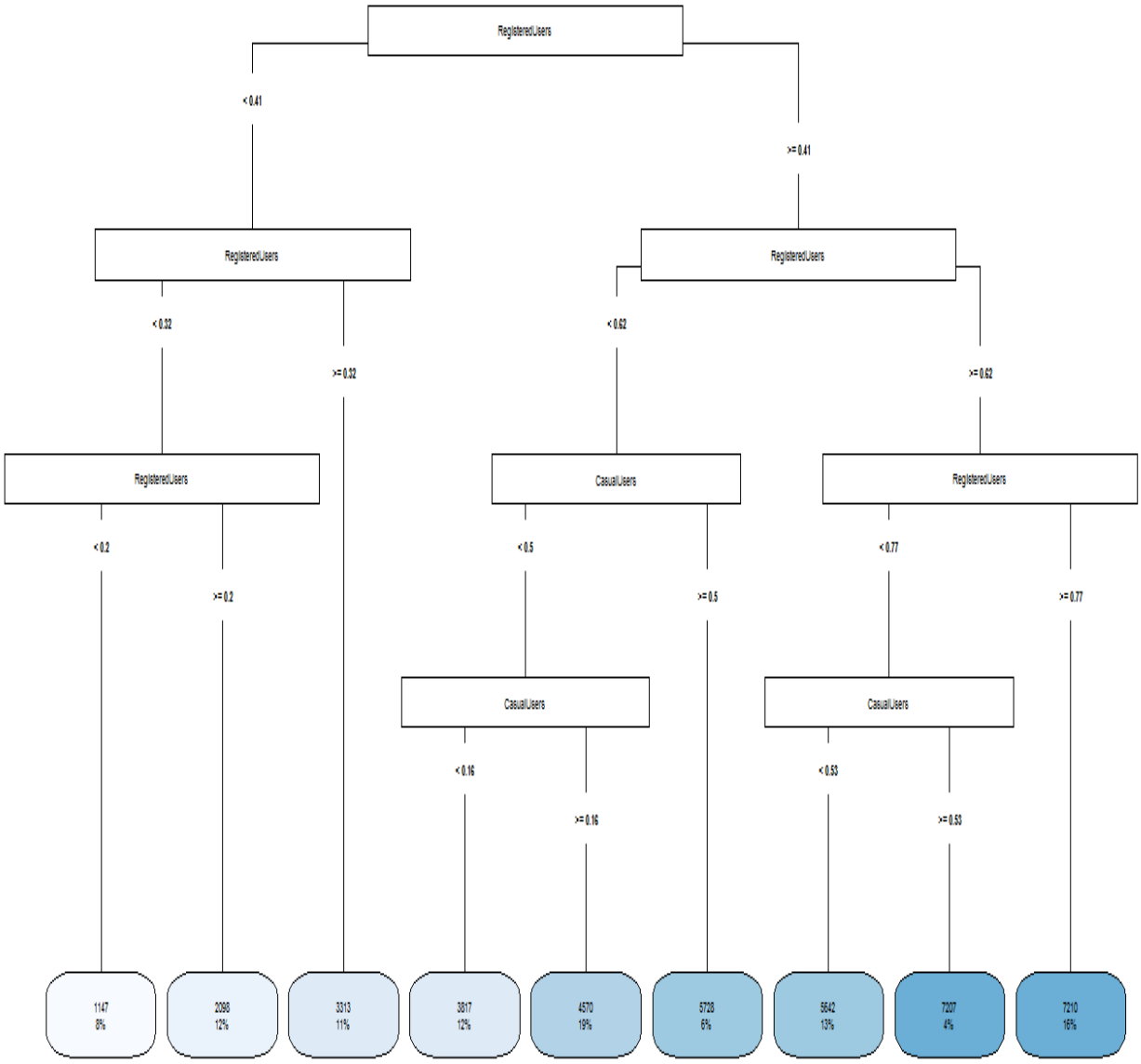
Decision tree builds regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

Trained Decision tree are shown below:-

```
root 584 2160583000 4524.541
   2) RegisteredUsers< 0.4127202 179  175337800 2307.011
     4) RegisteredUsers< 0.3164886 113   43455410 1719.204
       8) RegisteredUsers< 0.1960006 45    7909832 1147.244 *
       9) RegisteredUsers>=0.1960006 68   11082450 2097.706 *
     5) RegisteredUsers>=0.3164886 66   25991670 3313.409 *
   3) RegisteredUsers>=0.4127202 405  715987800 5504.635
     6) RegisteredUsers< 0.6155068 213  119084600 4519.948
      12) CasualUsers< 0.4970657 178   47914010 4282.382
        24) CasualUsers< 0.1600646 68    8142887 3816.691 *
        25) CasualUsers>=0.1600646 110   15907760 4570.264 *
      13) CasualUsers>=0.4970657 35   10034010 5728.143 *
     7) RegisteredUsers>=0.6155068 192  161261800 6597.021
      14) RegisteredUsers< 0.76581 98   62307970 6009.367
        28) CasualUsers< 0.5330106 75   13929460 5641.987 *
        29) CasualUsers>=0.5330106 23    5247255 7207.348 *
      15) RegisteredUsers>=0.76581 94   29827770 7209.681 *
```

From the trained model, I have analyzed that decision tree is not using all variable while training a model. It is only using RegisteredUsers and CasualUsers. So our mode is over fitted and biased towards two variables.

Graphical Visualization of a trained modal is shown below:



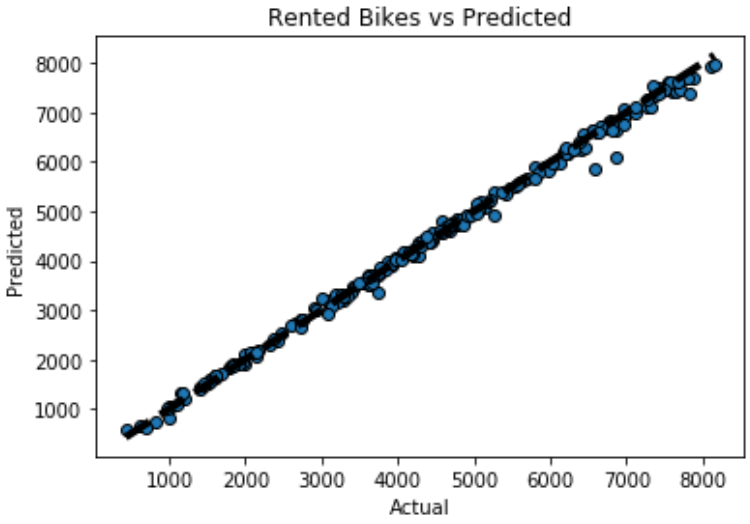
Accuracy and Error Rate of Decision Tree Model

Accuracy of the Model	Error Rate of the Model
96.52%	**3.48%**

Random Forest Model

Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because its simplicity and the fact that it can be used for both classification and regression tasks.

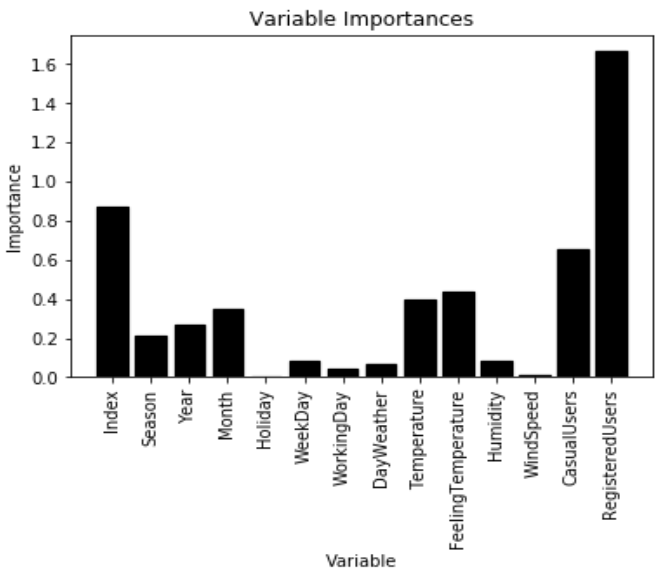
Visualization and Accuracy of the trained model is shown below:-



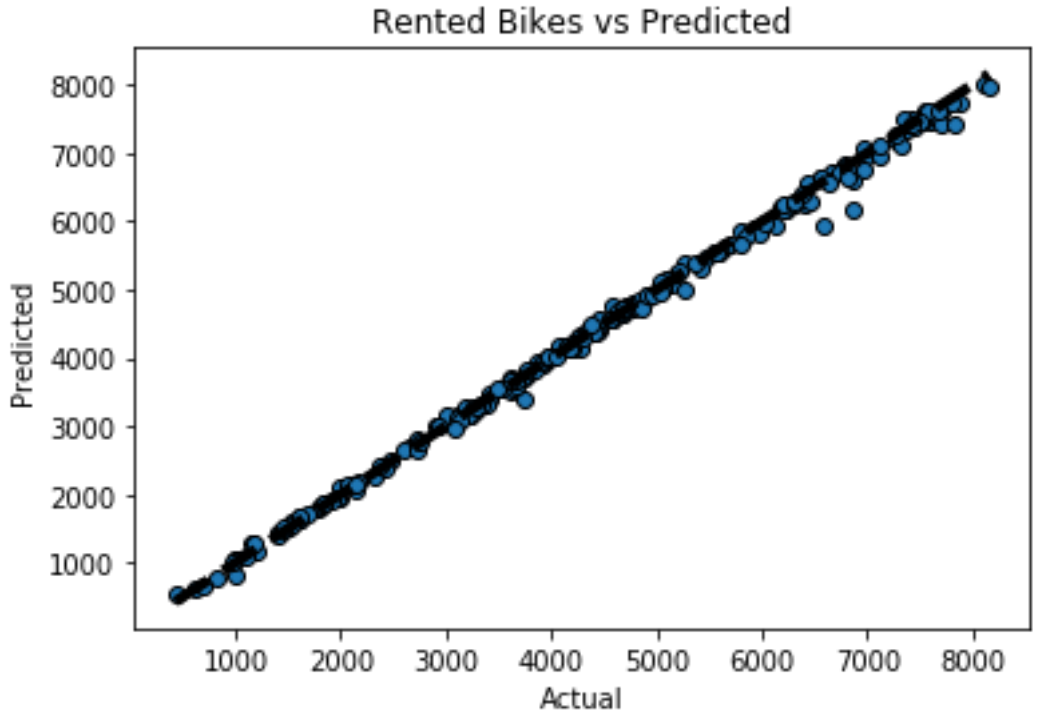
Accuracy of the Model	Error Rate of the Model
97.75%	**2.25%**

Over repetitions of iteration and training the model with all features were not increasing accuracy. So in order to increase the accuracy of the model we need to check importance of all features in the model then remove the features with least importance. Below graphs shows the importance of all the features present in the model.

	Importance
RegisteredUsers	0.907195
CasualUsers	0.085178
Index	0.002516
Temperature	0.001129
FeelingTemperature	0.000911
Humidity	0.000711
WindSpeed	0.000708
WeekDay	0.000523
Month	0.000393
WorkingDay	0.000313
Season	0.000163
DayWeather	0.000108
Year	0.000099
Holiday	0.000052



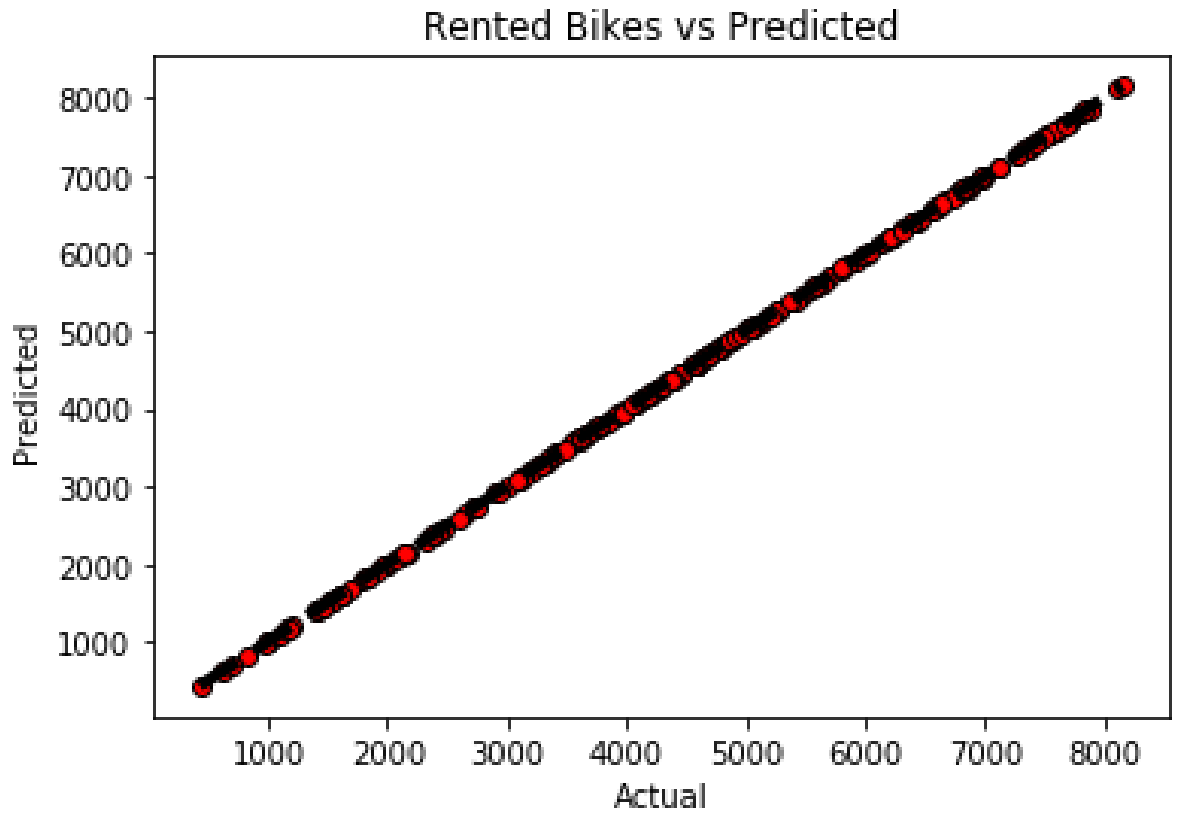
From the above graph and table, I came to conclusion that Holiday, WeekDay, DayWeather and WindSpeed variable are lest important for Random Forest model. So I remove all the four variables and trained the model again. Below mentioned figure gives clear visualization of final model:-



Accuracy of the Model	Error Rate of the Model
97.96%	**2.05%**

Linear Regression

Multiple Linear regression is the most common form of linear regression analysis. As a predictive analysis, the multiple linear regressions 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. The trained model is shown below:-



Accuracy of the Model	Error Rate of the Model
99.99% (Approx)	0.01% (Approx)

VIF values were user to select the features. Variable with the VIF values are shown below:-

```
---------- VIFs of the remained variables --------
          Variables      VIF
1            Season 3.745848
2              Year 2.544237
3             Month 3.082191
4           Holiday 1.134620
5           WeekDay 1.036123
6        WorkingDay 3.387797
7        DayWeather 1.275960
8       Temperature 2.358969
9         WindSpeed 1.086289
10      CasualUsers 3.678999
11 RegisteredUsers 5.759781
> |
```

ACCURACY OF REGRESSION MODEL IS HGH. SO LINEAR REGRESSION IS BEST FOR THIS DATASET.