Capstone Project – Regression

Project - Seoul Bike Sharing Demand Prediction

Team Members

Sanchit Misra

Mohit Jain ,

Tushar Hande,

Problem Statement:- Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually, providing the city with a stable supply of rental bikes becomes a major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.

Project description:-

- Bike rental businesses give customers—who are often, but not necessarily, tourists—bicycles for a short period. Bikes are generally rented for a few hours to recreationally explore the locality. But the customer base might also consist of college students on campus or others who rent for practical reasons.

- City bike rentals are particularly popular among tourists who like to explore their destination by bicycle. Usually, the customers of these businesses are most interested in an efficient, comfortable, and safe way of commuting from one place to another. Depending on the destination, weather conditions or the business can be seasonal. However, due to very seasonal industry it can be negatively affected by environmental forecasts and various other variables.

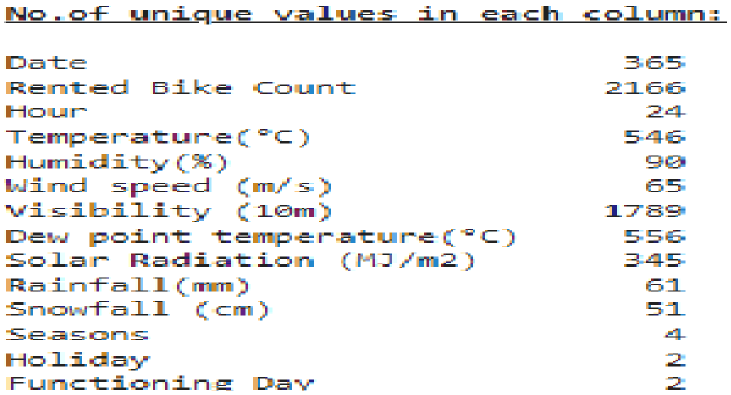
- Bike sharing is increasingly attracting more riders in cities around the world for its benefits regarding the urban environment and public health. One critical issue that Seoul is currently facing is the serious air pollution levels. The city’s PM10 and PM2.5 levels maintained considerably high levels in the past few years.

Data Description:-The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

Attribute Information:

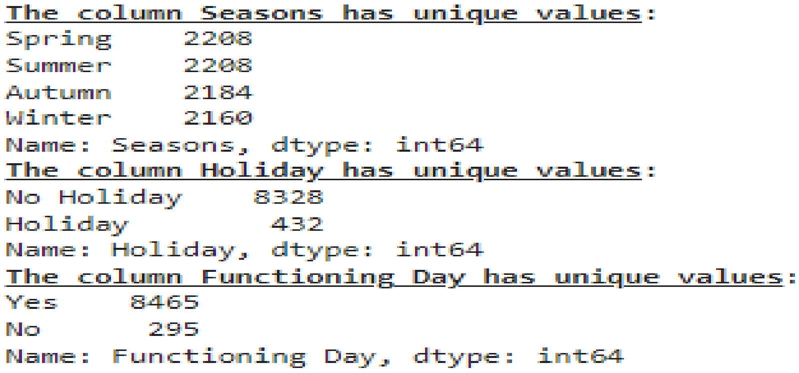
1. Date : year-month-day
2. Rented Bike count - Count of bikes rented at each hour
3. Hour - Hour of the day
4. Temperature-Temperature in Celsius
5. Humidity - %
6. Windspeed - m/s
7. Visibility - 10m
8. Dew point temperature - Celsius
9. Solar radiation - MJ/m2
10. Rainfall - mm
11. Snowfall – cm
12. Seasons - Winter, Spring, Summer, Autumn
13. Holiday - Holiday/No holiday
14. Functional Day – No_Func(Non Functional Hours), Fun(Functional hours)

Data Inspection:-



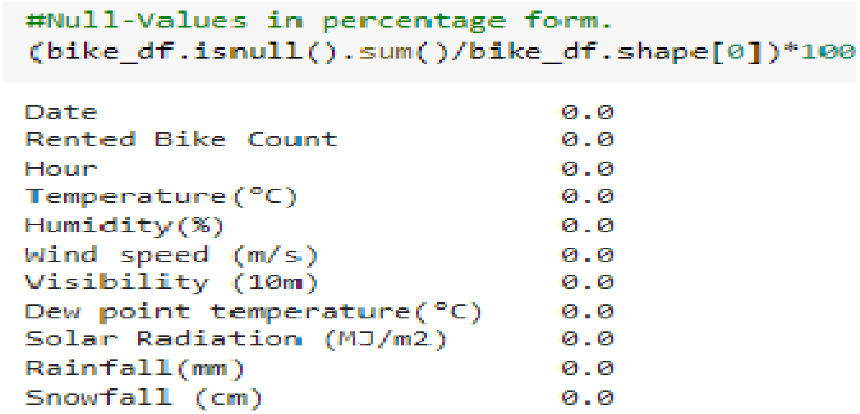
```
No.of unique values in each column:

Date                                365
Rented Bike Count                  2166
Hour                                 24
Temperature(°C)                     546
Humidity(%)                          90
Wind speed (m/s)                     65
Visibility (10m)                   1789
Dew point temperature(°C)           556
Solar Radiation (MJ/m2)             345
Rainfall(mm)                         61
Snowfall (cm)                        51
Seasons                               4
Holiday                               2
Functioning Day                       2
```

Checked unique value of data set



```
The column Seasons has unique values:
Spring    2208
Summer    2208
Autumn    2184
Winter    2160
Name: Seasons, dtype: int64
The column Holiday has unique values:
No Holiday    8328
Holiday        432
Name: Holiday, dtype: int64
The column Functioning Day has unique values:
Yes    8465
No      295
Name: Functioning Day, dtype: int64
```

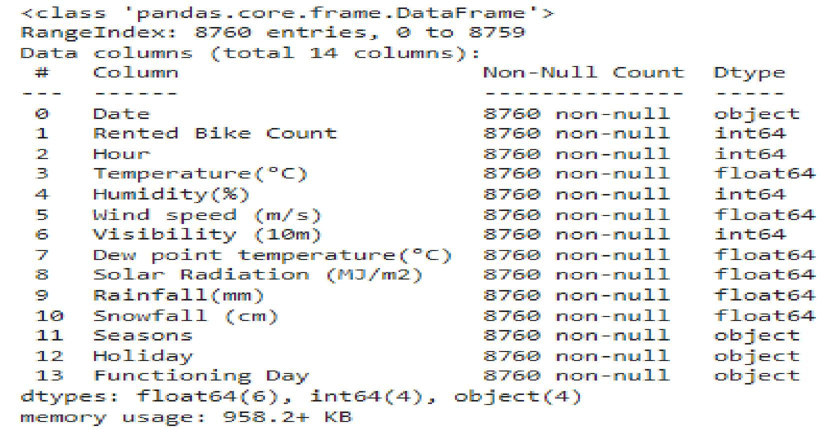
Checked unique value of Categorical data set.



```
#Null-Values in percentage form.
(bike_df.isnull().sum()/bike_df.shape[0])*100
```

```
Date                         0.0
Rented Bike Count            0.0
Hour                         0.0
Temperature(°C)              0.0
Humidity(%)                  0.0
Wind speed (m/s)             0.0
Visibility (10m)             0.0
Dew point temperature(°C)    0.0
Solar Radiation (MJ/m2)      0.0
Rainfall(mm)                 0.0
Snowfall (cm)                0.0
```

Checked Null value of data set



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8760 entries, 0 to 8759
Data columns (total 14 columns):
 #   Column                     Non-Null Count  Dtype  
---  ------                     --------------  -----  
 0   Date                       8760 non-null   object 
 1   Rented Bike Count          8760 non-null   int64  
 2   Hour                       8760 non-null   int64  
 3   Temperature(°C)            8760 non-null   float64
 4   Humidity(%)                8760 non-null   int64  
 5   Wind speed (m/s)           8760 non-null   float64
 6   Visibility (10m)           8760 non-null   int64  
 7   Dew point temperature(°C)  8760 non-null   float64
 8   Solar Radiation (MJ/m2)    8760 non-null   float64
 9   Rainfall(mm)               8760 non-null   float64
 10  Snowfall (cm)              8760 non-null   float64
 11  Seasons                    8760 non-null   object 
 12  Holiday                    8760 non-null   object 
 13  Functioning Day            8760 non-null   object 
dtypes: float64(6), int64(4), object(4)
memory usage: 958.2+ KB
```

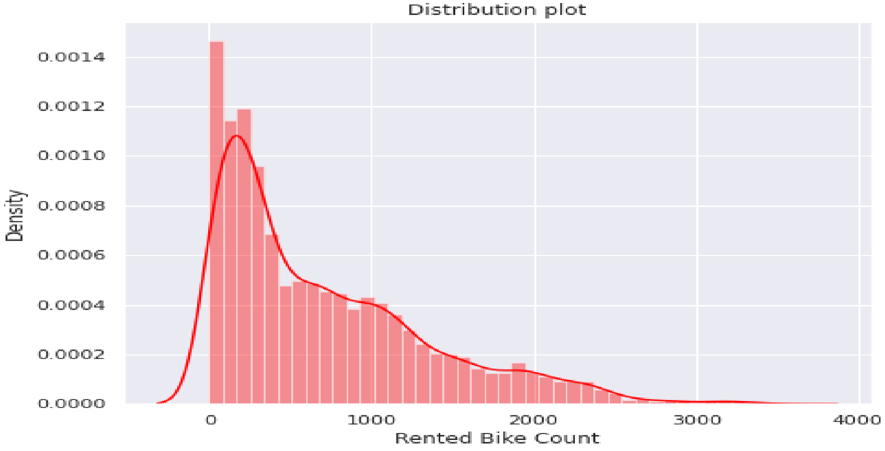
Information of whole data set

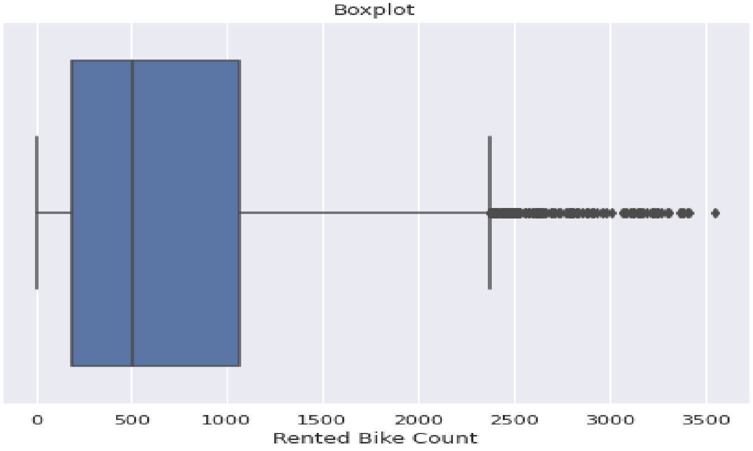
Exploratory Data Analysis (EDA)

An EDA is a detailed analysis designed to reveal a data set's underlying structure. It is significant for a business because it identifies trends, patterns, and linkages that are not intuitively clear

Univariate Analysis:- On Dependent Variable Rented Bike Count

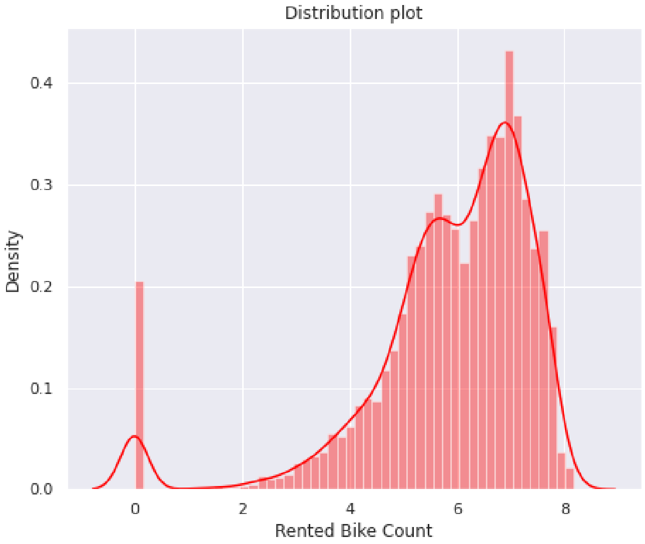
- Provides summary statistics for each field in the raw data set or summary only on one variable.
- The ultimate purpose of a Univariate analysis is to simply explain the data and look for patterns therein.





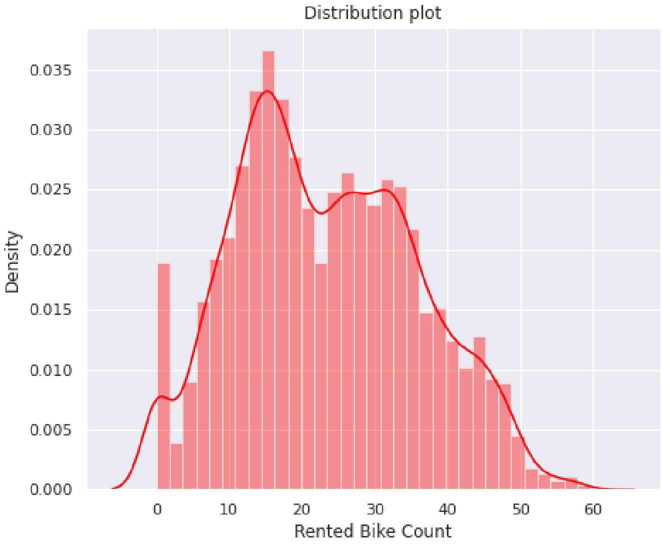
Transformation of data to Normal distribution by various method:-

- Log Transformation



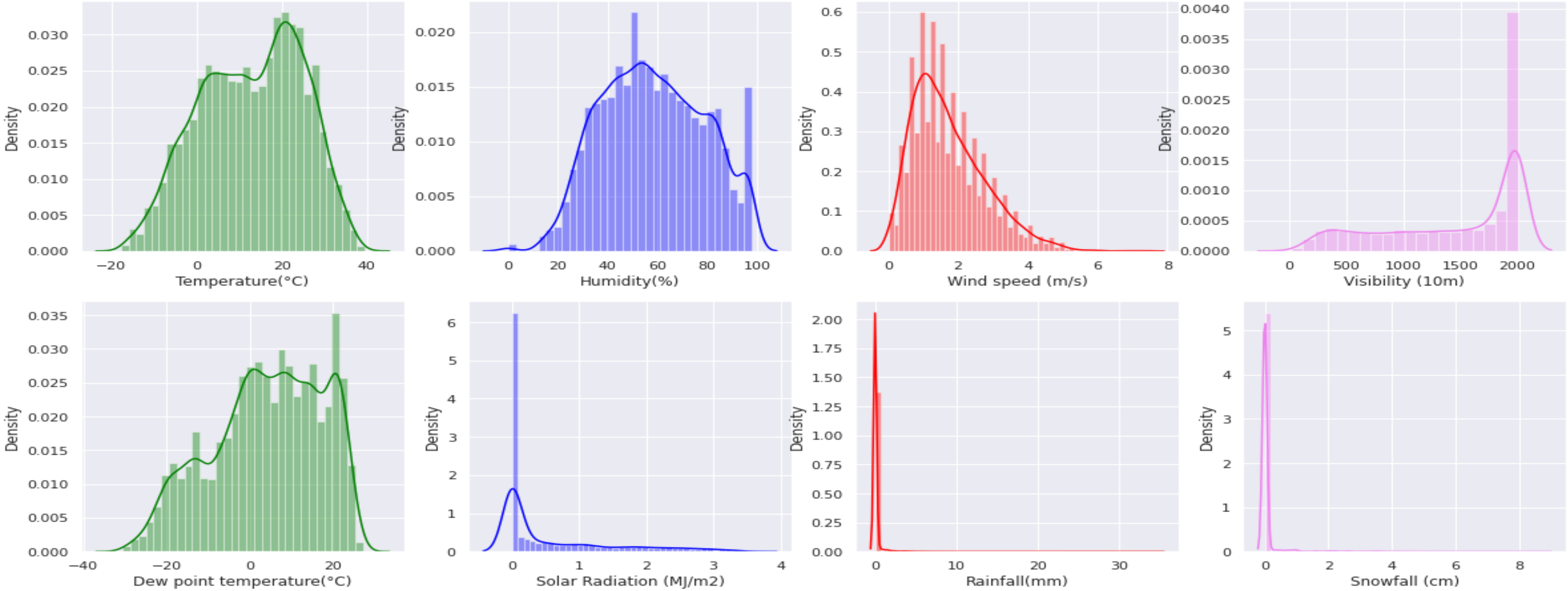
The log transformation gives us a left skewed distribution.

- Square root Transformation.



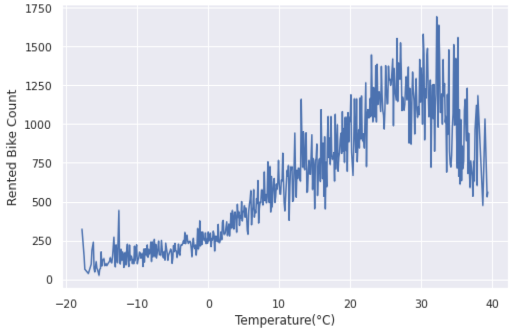
So, we used square root transformation on the dependent variable which gives us distribution which is almost normal in nature

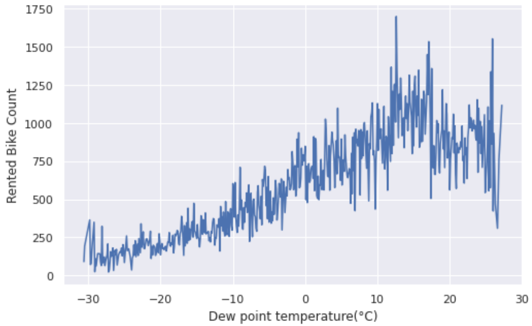
Distribution of all numerical independent Variable in the data set:-

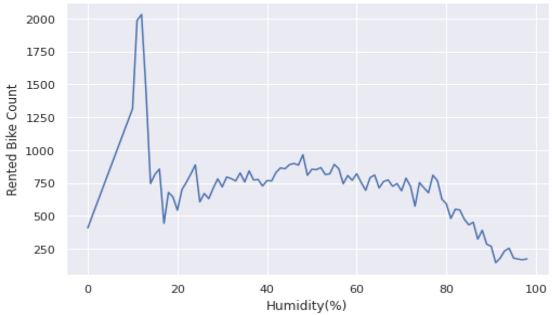


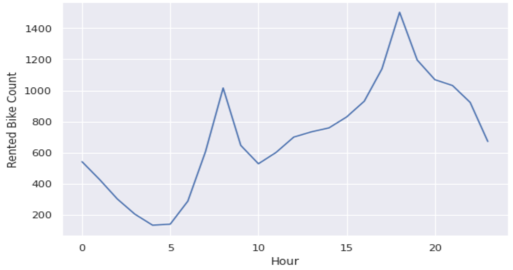
- Interpretation:- We can see from the preceding dist plots that **"Temperature"**, **"Dew point temperature"** and **"Humidity"** follows approximately normal distribution.
- The distribution of all other variables are either left- or right-skewed.

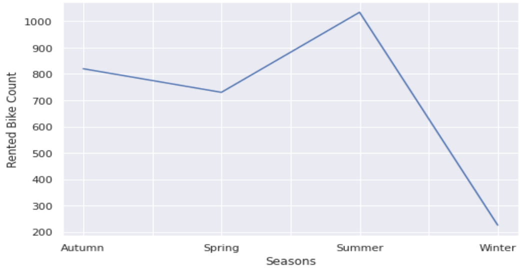
Bivariate Analysis:- It is performed to find the relationship between each variable in the dataset and the target variable of interest or using 2 variables and finding the relationship between them.

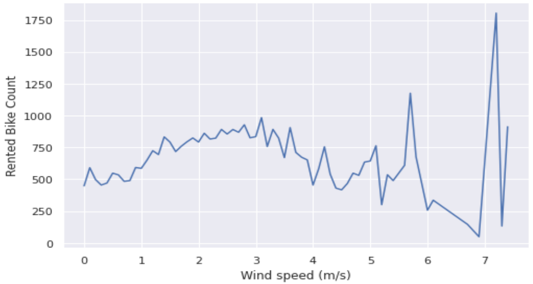












Some of the following observations can be interpreted based on the above analysis:

•The trend for bike rentals grows from **temperature 20 to 30 degrees** Celsius.

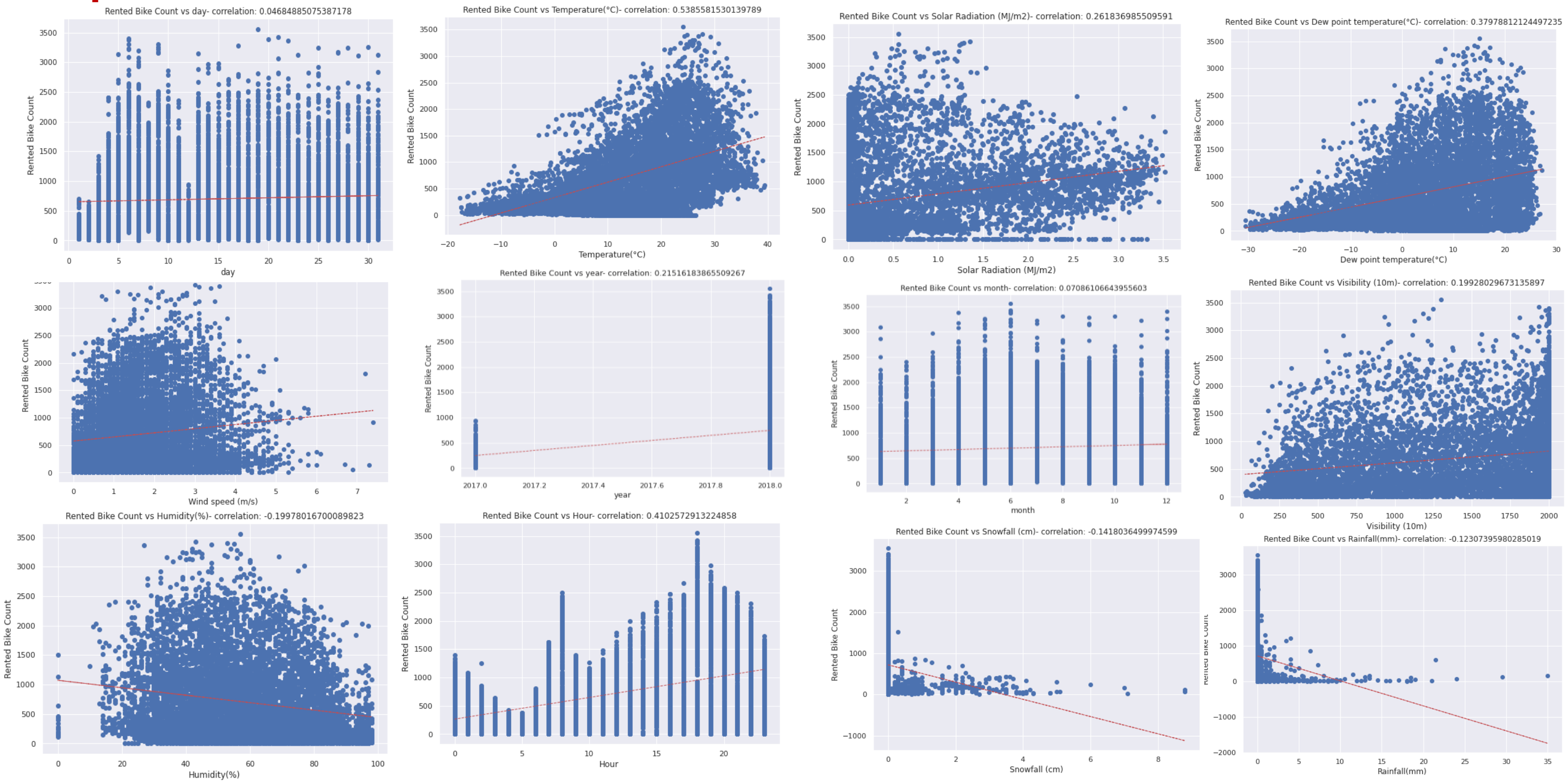
•The dew point temperature behaves similar to the temperature for the dependent variable.

•**Lower humidity** tends to increase demand for bike rentals.

•The bike usage is high in peak office hours that is **7-9 AM in morning and 6-8 PM in evening**.

•People prefer to drive in the **summer season** when the **wind is stronger**.

Used regression scatterplot for relation between independent and dependent variable:-


Rented Bike Count vs day- correlation: 0.04684885075387178

Rented Bike Count vs Temperature(°C)- correlation: 0.5385581530139789

Rented Bike Count vs Solar Radiation (MJ/m2)- correlation: 0.261836985509591

Rented Bike Count vs Dew point temperature(°C)- correlation: 0.37978812124497235

Rented Bike Count vs year- correlation: 0.21516183865509267

Rented Bike Count vs month- correlation: 0.07086106643955603

Rented Bike Count vs Visibility (10m)- correlation: 0.19928029673135897

Rented Bike Count vs Humidity(%)- correlation: -0.19978016700089823

Rented Bike Count vs Hour- correlation: 0.4102572913224858

Rented Bike Count vs Snowfall (cm)- correlation: -0.1418036499974599

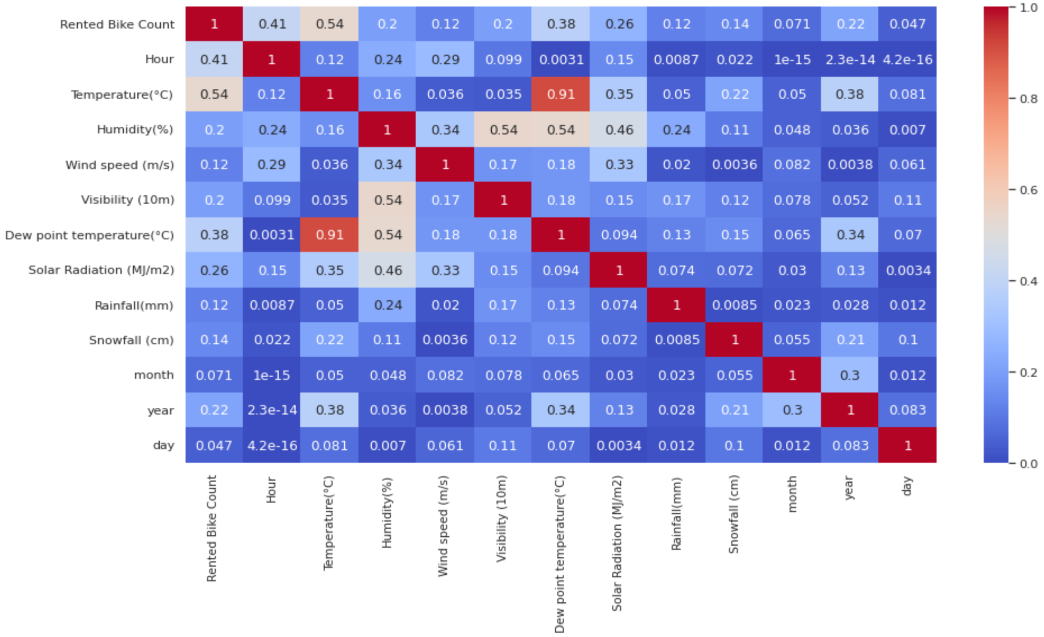
Rented Bike Count vs Rainfall(mm)- correlation: -0.12307395980285019

Identifying Multicollinearity:-

A correlation could be positive, meaning both variables move in the same direction, or negative, meaning that when one variable’s value increases, the other variables’ values decrease. Correlation can also be neutral or zero, meaning that the variables are unrelated to each other.
Positive Correlation: both variables change in the same direction.
Neutral Correlation: No relationship in the change of the variables.
Negative Correlation: variables change in opposite directions.
The performance of some algorithms can deteriorate if two or more variables are tightly related, this condition called multicollinearity

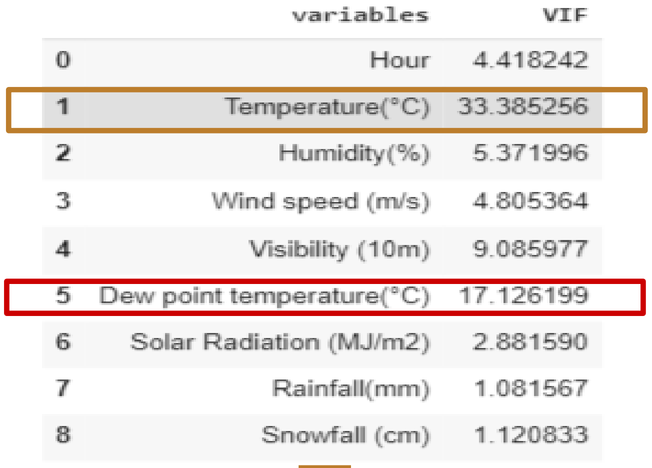
From the Heatmap we can see that:

- The **Dew point temperature** attributes have a high correlation with the **Temperature** attributes. As a result, it can be dropped.

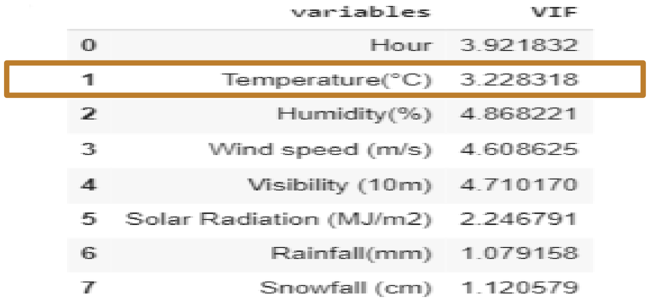


	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	month	year	day
Rented Bike Count	1	0.41	0.54	0.2	0.12	0.2	0.38	0.26	0.12	0.14	0.071	0.22	0.047
Hour	0.41	1	0.12	0.24	0.29	0.099	0.0031	0.15	0.0087	0.022	1e-15	2.3e-14	4.2e-16
Temperature(°C)	0.54	0.12	1	0.16	0.036	0.035	0.91	0.35	0.05	0.22	0.05	0.38	0.081
Humidity(%)	0.2	0.24	0.16	1	0.34	0.54	0.54	0.46	0.24	0.11	0.048	0.036	0.007
Wind speed (m/s)	0.12	0.29	0.036	0.34	1	0.17	0.18	0.33	0.02	0.0036	0.082	0.0038	0.061
Visibility (10m)	0.2	0.099	0.035	0.54	0.17	1	0.18	0.15	0.17	0.12	0.078	0.052	0.11
Dew point temperature(°C)	0.38	0.0031	0.91	0.54	0.18	0.18	1	0.094	0.13	0.15	0.065	0.34	0.07
Solar Radiation (MJ/m2)	0.26	0.15	0.35	0.46	0.33	0.15	0.094	1	0.074	0.072	0.03	0.13	0.0034
Rainfall(mm)	0.12	0.0087	0.05	0.24	0.02	0.17	0.13	0.074	1	0.0085	0.023	0.028	0.012
Snowfall (cm)	0.14	0.022	0.22	0.11	0.0036	0.12	0.15	0.072	0.0085	1	0.055	0.21	0.1
month	0.071	1e-15	0.05	0.048	0.082	0.078	0.065	0.03	0.023	0.055	1	0.3	0.012
year	0.22	2.3e-14	0.38	0.036	0.0038	0.052	0.34	0.13	0.028	0.21	0.3	1	0.083
day	0.047	4.2e-16	0.081	0.007	0.061	0.11	0.07	0.0034	0.012	0.1	0.012	0.083	1

After Removing Multicollinearity on the basis of VIF score and Heat MAP:-

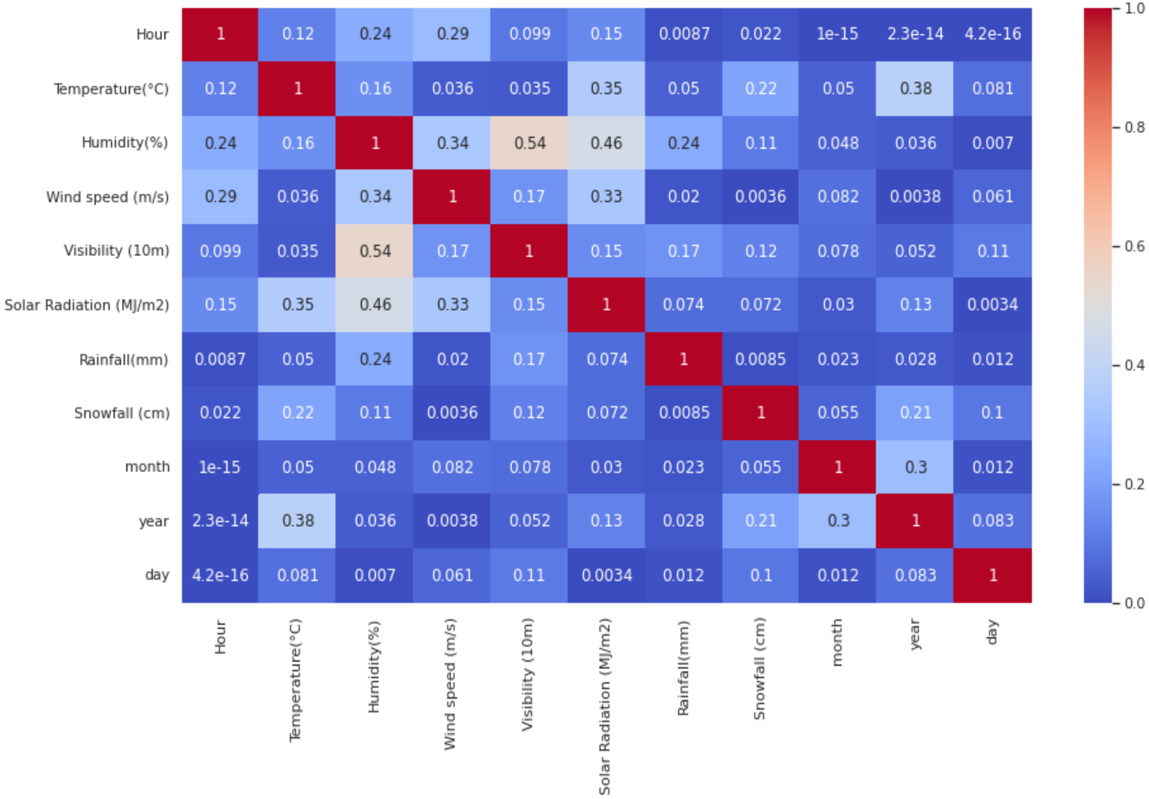


	variables	VIF
0	Hour	4.418242
1	Temperature(°C)	33.385256
2	Humidity(%)	5.371996
3	Wind speed (m/s)	4.805364
4	Visibility (10m)	9.085977
5	Dew point temperature(°C)	17.126199
6	Solar Radiation (MJ/m2)	2.881590
7	Rainfall(mm)	1.081567
8	Snowfall (cm)	1.120833



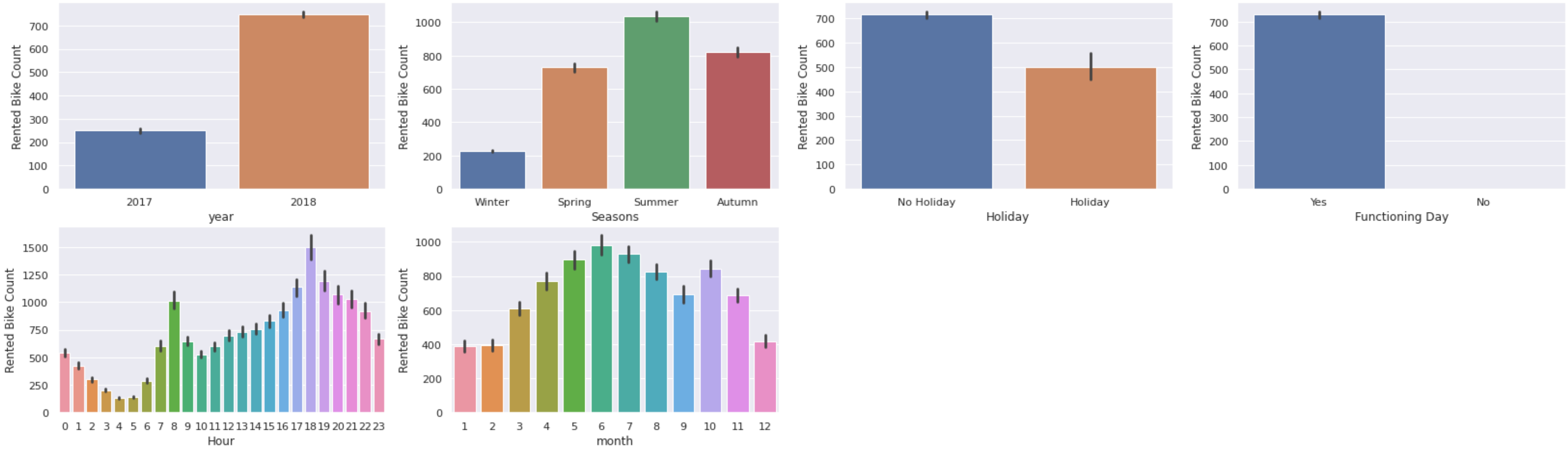
	variables	VIF
0	Hour	3.921832
1	Temperature(°C)	3.228318
2	Humidity(%)	4.868221
3	Wind speed (m/s)	4.608625
4	Visibility (10m)	4.710170
5	Solar Radiation (MJ/m2)	2.246791
6	Rainfall(mm)	1.079158
7	Snowfall (cm)	1.120579

Modified HEAT MAP



	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	month	year	day
Hour	1	0.12	0.24	0.29	0.099	0.15	0.0087	0.022	1e-15	2.3e-14	4.2e-16
Temperature(°C)	0.12	1	0.16	0.036	0.035	0.35	0.05	0.22	0.05	0.38	0.081
Humidity(%)	0.24	0.16	1	0.34	0.54	0.46	0.24	0.11	0.048	0.036	0.007
Wind speed (m/s)	0.29	0.036	0.34	1	0.17	0.33	0.02	0.0036	0.082	0.0038	0.061
Visibility (10m)	0.099	0.035	0.54	0.17	1	0.15	0.17	0.12	0.078	0.052	0.11
Solar Radiation (MJ/m2)	0.15	0.35	0.46	0.33	0.15	1	0.074	0.072	0.03	0.13	0.0034
Rainfall(mm)	0.0087	0.05	0.24	0.02	0.17	0.074	1	0.0085	0.023	0.028	0.012
Snowfall (cm)	0.022	0.22	0.11	0.0036	0.12	0.072	0.0085	1	0.055	0.21	0.1
month	1e-15	0.05	0.048	0.082	0.078	0.03	0.023	0.055	1	0.3	0.012
year	2.3e-14	0.38	0.036	0.0038	0.052	0.13	0.028	0.21	0.3	1	0.083
day	4.2e-16	0.081	0.007	0.061	0.11	0.0034	0.012	0.1	0.012	0.083	1

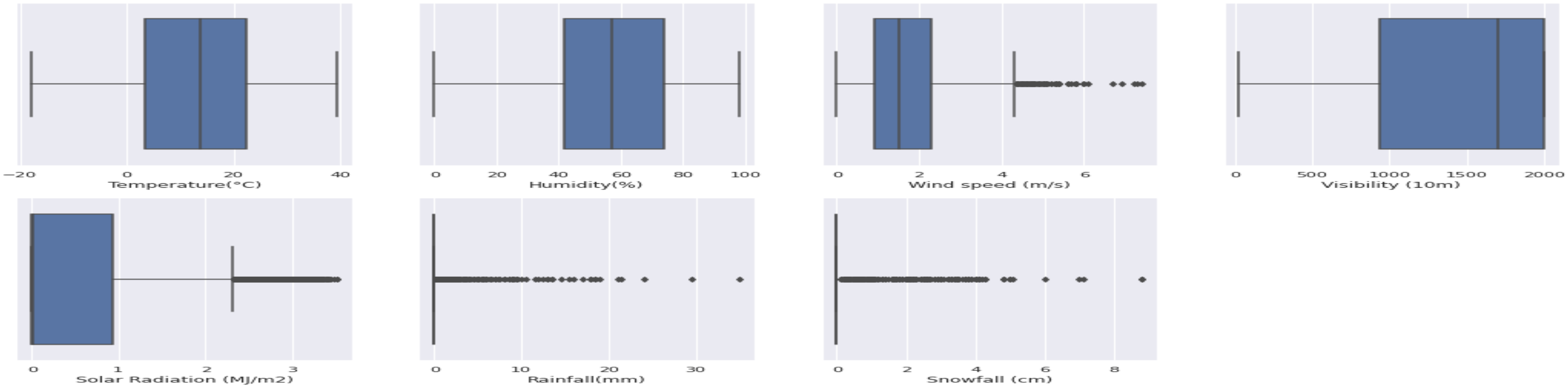
Bar plots between dependent and independent feature:-



We may interpret the following information based on the above plots:

- During **peak office hours**, it appears that rented bikes are in high demand.
- As it was seen in the preliminary analysis, riding a bike is prioritized during the **summer months.**
- On **non-holiday days**, consumption for bikes is **prominent**.
- It is intuitive that, on **functioning days**, there is a **significant demand** for bikes.
- The month of **June** appears to be the most demanding one overall.

Identification of Outlier in the dataset feature by using Box Plot:-



Interpretation:-

- The highest number of outliers among all are found in snowfall and rainfall.
- Both wind speed and solar radiation constitute a substantial quantity of outliers.

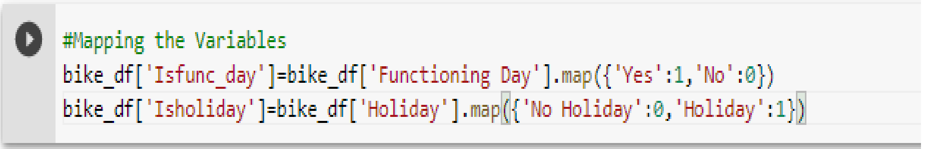
Action:- We have capped both upper and lower value of all the feature with the below formula.

- Upper Limit cap with :- Q3 + 1.5*IQR
- Lower Limit cap with :- Q1 - 1.5*IQR

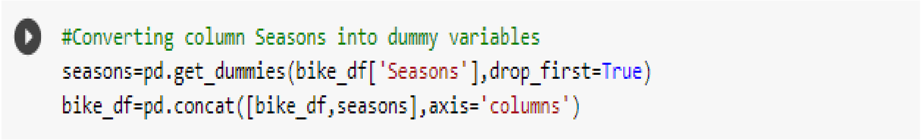
Feature Engineering And Selection:-

Feature engineering is **a machine learning technique that leverages data to create new variables that aren't in the training set**. It can produce new features for both supervised and unsupervised learning, with the goal of simplifying and speeding up data transformations while also enhancing model accuracy.

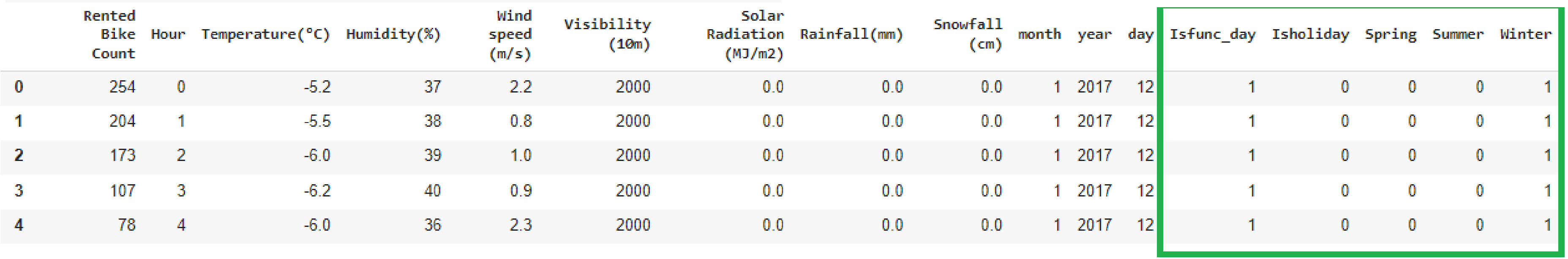
We have used pandas pd.get_dummies method to convert categorical data to numerical data.



```
#Mapping the Variables
bike_df['Isfunc_day']=bike_df['Functioning Day'].map({'Yes':1,'No':0})
bike_df['Isholiday']=bike_df['Holiday'].map({'No Holiday':0,'Holiday':1})
```



```
#Converting column Seasons into dummy variables
seasons=pd.get_dummies(bike_df['Seasons'],drop_first=True)
bike_df=pd.concat([bike_df,seasons],axis='columns')
```



	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	month	year	day	Isfunc_day	Isholiday	Spring	Summer	Winter
0	254	0	-5.2	37	2.2	2000	0.0	0.0	0.0	1	2017	12	1	0	0	0	1
1	204	1	-5.5	38	0.8	2000	0.0	0.0	0.0	1	2017	12	1	0	0	0	1
2	173	2	-6.0	39	1.0	2000	0.0	0.0	0.0	1	2017	12	1	0	0	0	1
3	107	3	-6.2	40	0.9	2000	0.0	0.0	0.0	1	2017	12	1	0	0	0	1
4	78	4	-6.0	36	2.3	2000	0.0	0.0	0.0	1	2017	12	1	0	0	0	1

Model Building:-

A machine learning model is a program that can find patterns or make decisions from a previously unseen data set. The process of running a machine learning algorithm on a dataset (called training data) and optimizing the algorithm to find certain patterns or outputs is called model training. The resulting function with rules and data structures is called the trained machine learning model.

We have used below some of the model for our project to achieve high accuracy result of bike sharing demand.

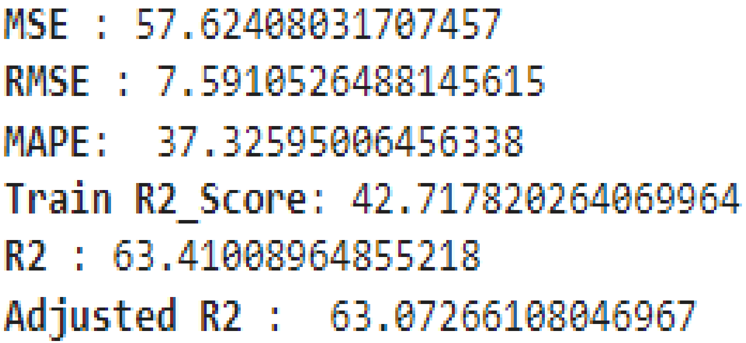
- **Linear Regression:-**
- **a. Lasso Regression**
- **b. Ridge Regression.**
- **c. Elastic net Regression**
- **Decision tree**
- **Random Forest Regressor**
- **XG Boosting**
- **Cat Boosting**

Linear Regression:-

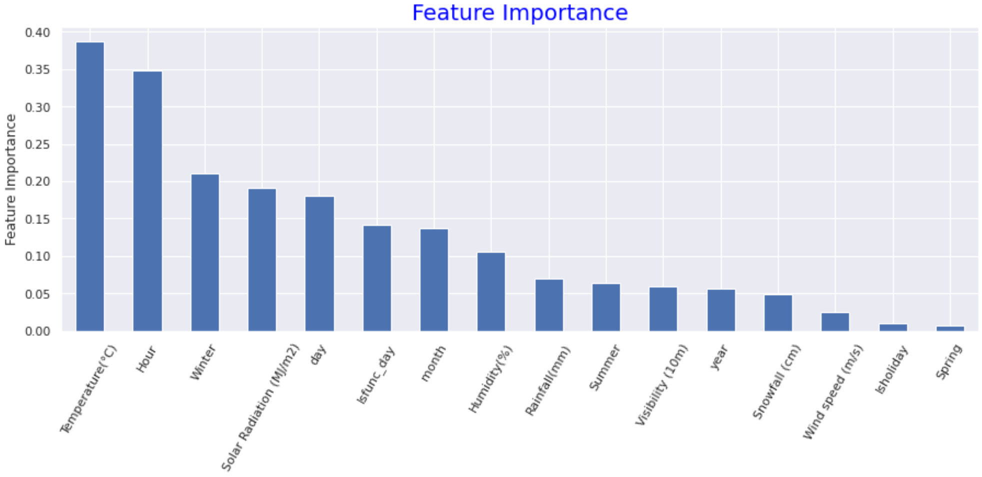
Linear regression is used to identify relationships between the variable of interest and the inputs, and predict its values based on the values of the input variables.

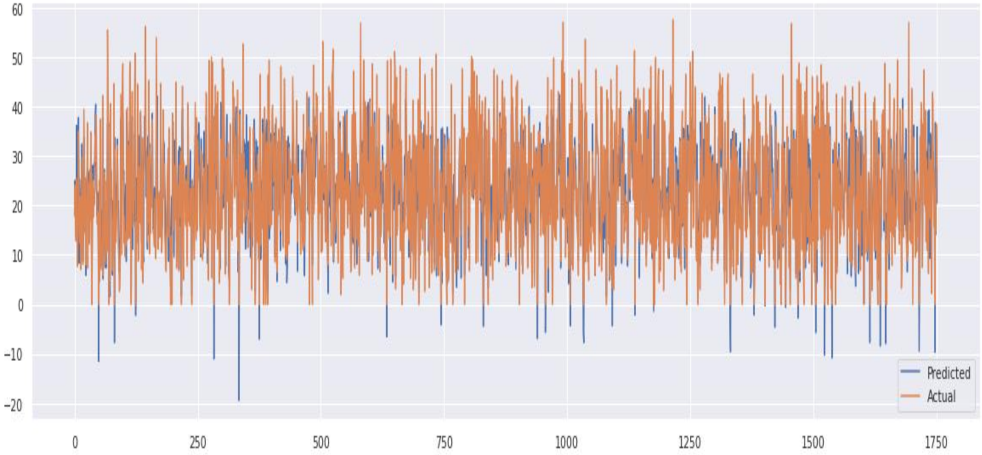
Observations:

- After implementing the linear regression algorithm, we obtain a R2 score of 63.41 %, which is considered appropriate but we have also seen MAPE as 37.32%, which needs to be reduce further. Therefore, we will now attempt to maximize our R2 score and reduce MAPE% by introducing more regression algorithms to our dataset.
- For a linear regression model, the **Temperature** feature seems to be very significant following **Hour** and **Winter**.



```
MSE : 57.62408031707457
RMSE : 7.591052648814561 5
MAPE:  37.32595006456338
Train R2_Score: 42.71782026406996 4
R2 : 63.41008964855218
Adjusted R2 :  63.07266108046967
```



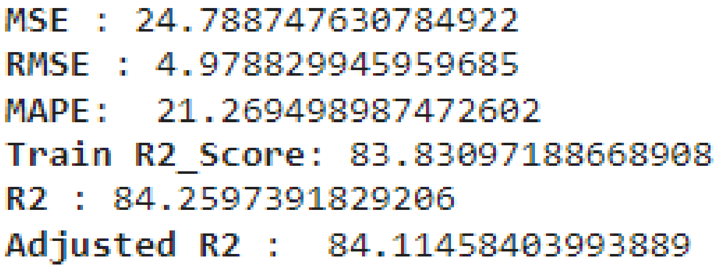


Decision Tree:-

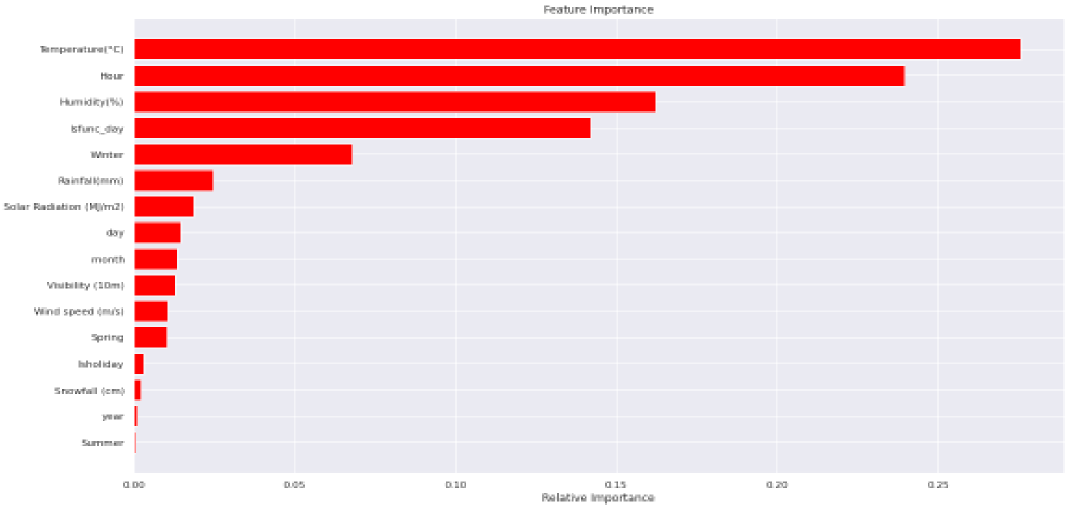
A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes).

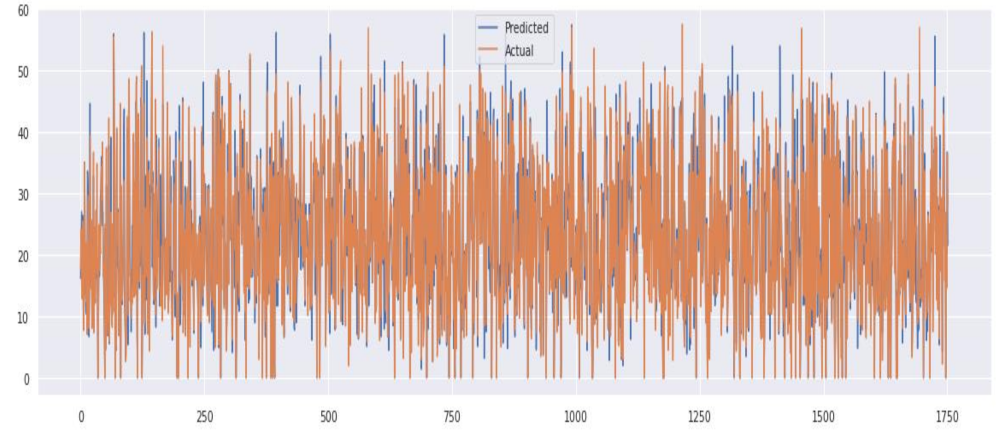
Observations:

- After implementing the Decision Tree algorithm, we obtain a R2 score of 84.25 % and MAPE as 21.26%, both are improved majorly. We will now attempt to maximize our R2 score and minimize MAPE% by introducing more tree-based and boosting algorithms to our dataset.
- For a Decision Tree model, the **Temperature** feature seems to be very significant following **Hour** and **Humidity**.



```
MSE : 24.788747630784922
RMSE : 4.978829945959685
MAPE:  21.269498987472602
Train R2_Score: 83.83097188668908
R2 : 84.2597391829206
Adjusted R2 :  84.11458403993889
```



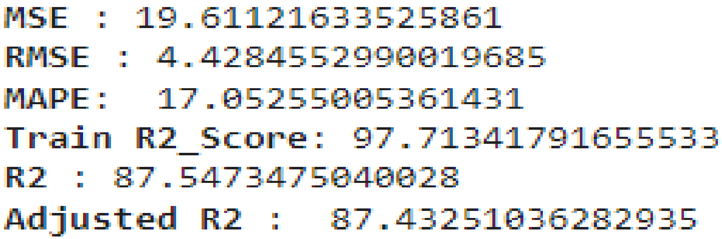


Random Forests:-

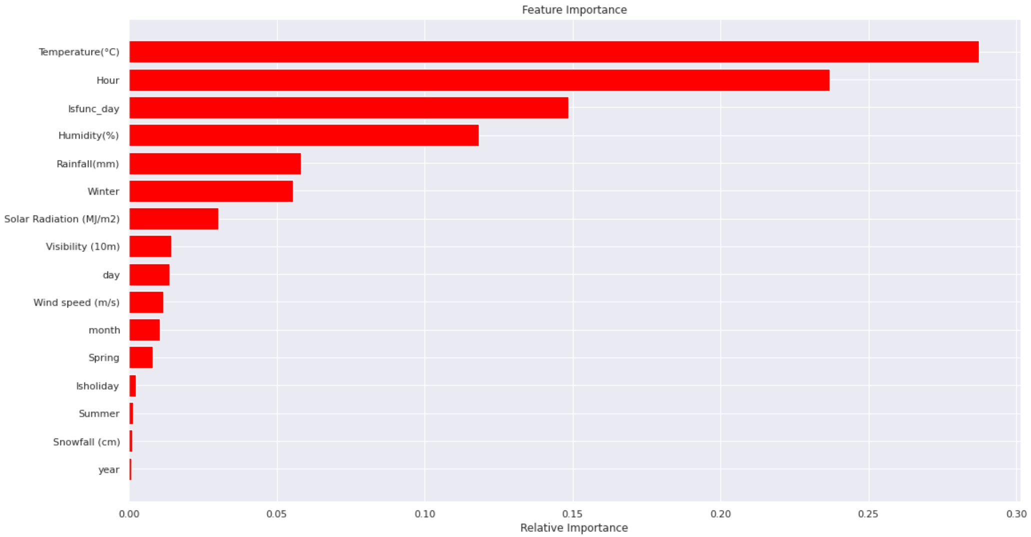
Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time.

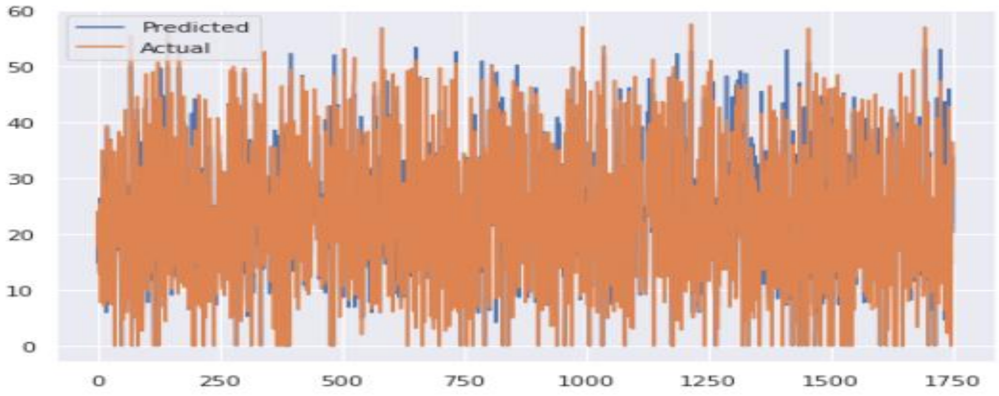
Observations:

- After implementing the Random Forest algorithm, we obtain a R2 score of 87.54 % and MAPE as 17.05%, both can be considered as better. We will now attempt to maximize our R2 score and minimize MAPE% by introducing more boosting algorithms to our dataset.
- For a Random Forest model, the **Temperature** feature seems to be very significant following **Hour** and **Isfunc_day.**



```
MSE : 19.61121633525861
RMSE : 4.428455299001968‍5
MAPE:  17.05255005361431
Train R2_Score: 97.71341791655533
R2 : 87.54734750400 28
Adjusted R2 :  87.43251036282935
```



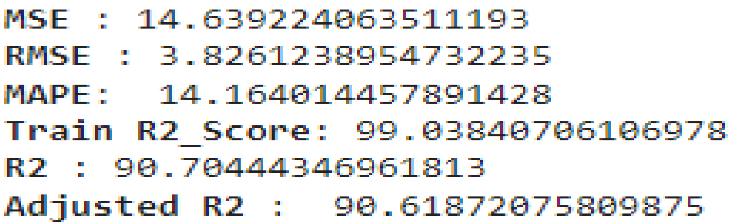


XG BOOST:-

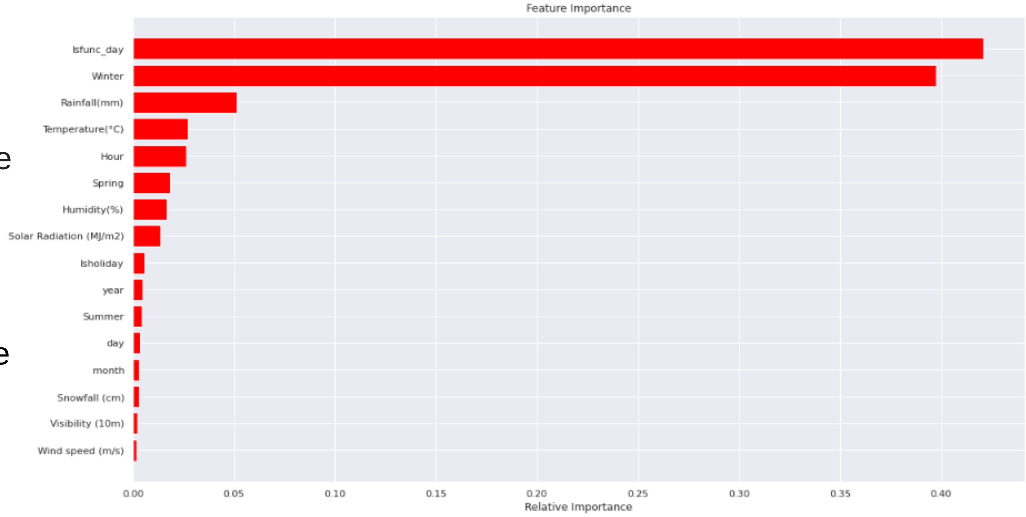
XG Boost is a scalable and highly accurate implementation of gradient boosting that pushes the limits of computing power for boosted tree algorithms, being built largely for energizing machine learning model performance and computational speed.

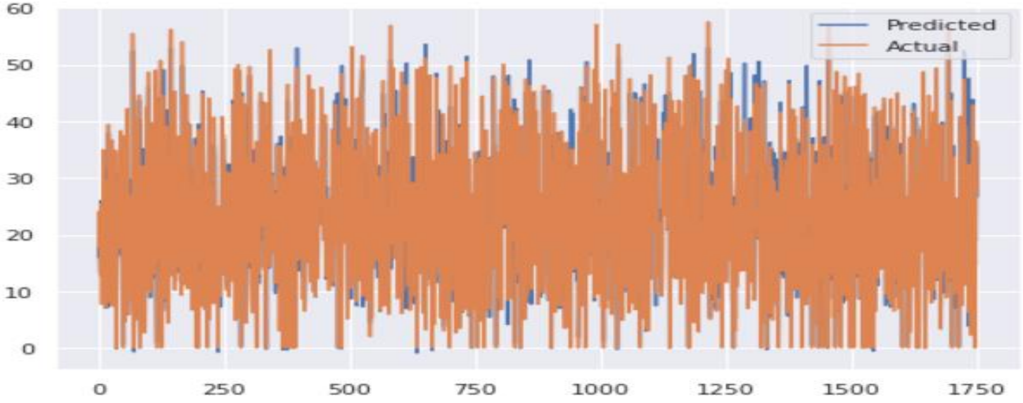
Observations:

- After implementing the XG Boosting algorithm, we obtain a R2 score of 90.70 % and MAPE as 14.16%, both are best till now. We will now attempt to maximize our R2 score and try to reduce MAPE% more by introducing other boosting algorithms to our dataset.
- For a XG Boosting model, the **Isfunc_day** feature seems to be very significant following **Winter** and **Rainfall**



```
MSE : 14.639224063511193
RMSE : 3.8261238954732235
MAPE:  14.164014457891428
Train R2_Score: 99.03840706106978
R2 : 90.70444346961813
Adjusted R2 :  90.61872075809875
```



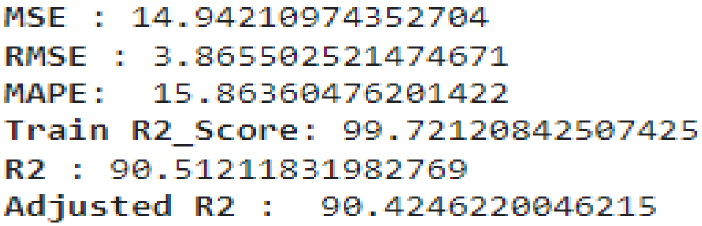


CAT BOOSTING:-

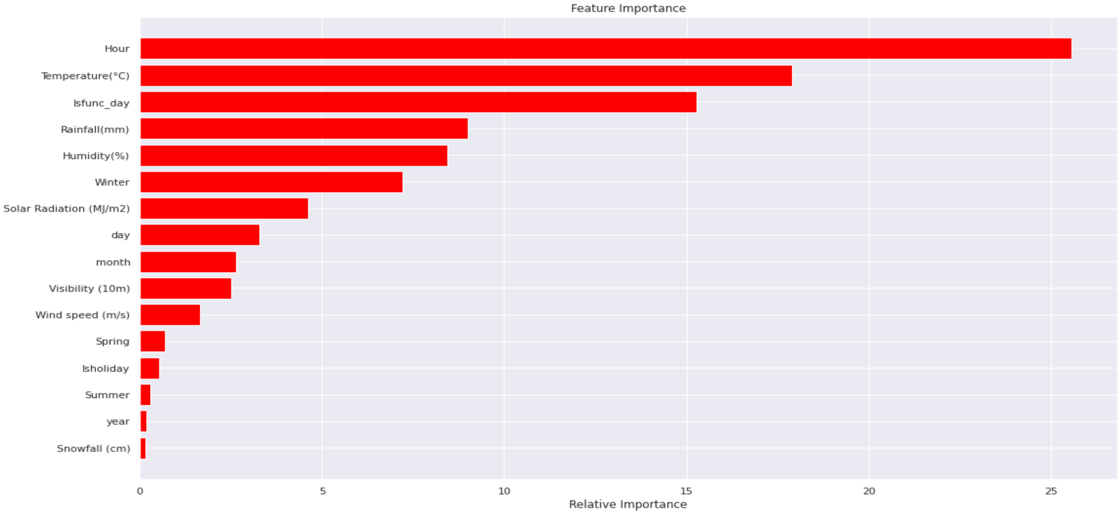
Cat Boost is an open-source software library developed by Yandex. It provides a gradient boosting framework which among other features attempts to solve for Categorical features using a permutation driven alternative compared to the classical algorithm.

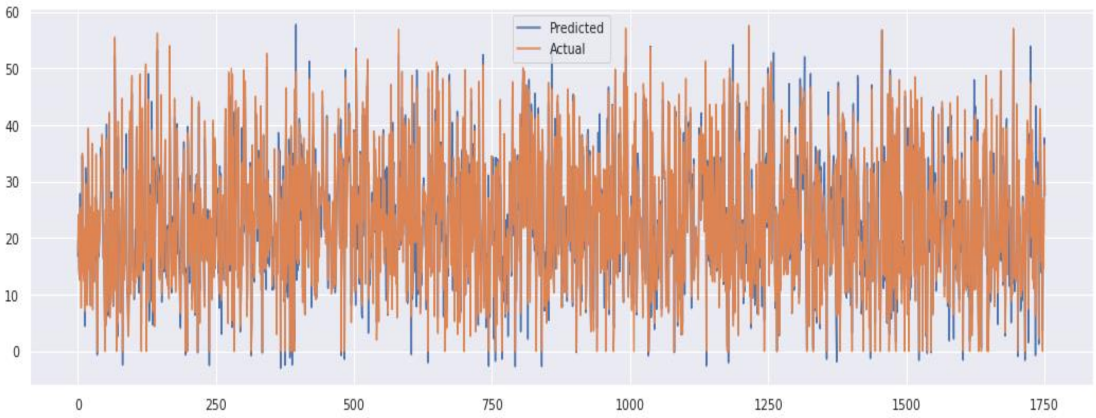
Observations:

- After implementing the Cat Boosting algorithm, we obtain R2 score of 90.51 % and MAPE 15.86%, which is considered good. Therefore, we will end our analysis here.
- For a Cat Boosting model, the **Hour** feature seems to be very significant following **Temperature** and **Isfunc_day.**

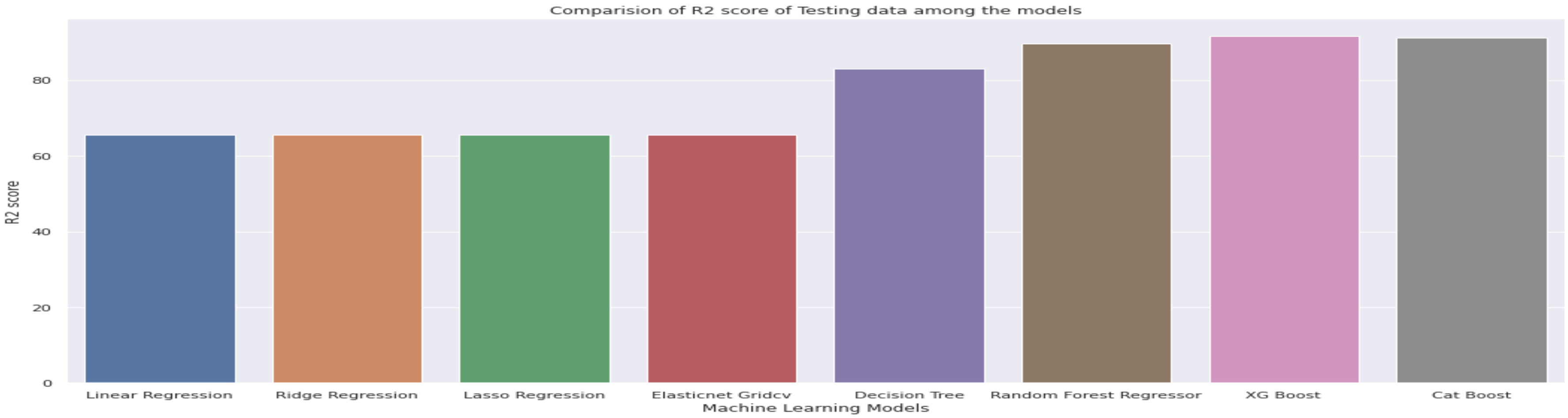


```
MSE : 14.94210974352704
RMSE : 3.865502521474671
MAPE:  15.86360476201422
Train R2_Score: 99.72120842507425
R2 : 90.51211831982769
Adjusted R2 :  90.4246220046215
```



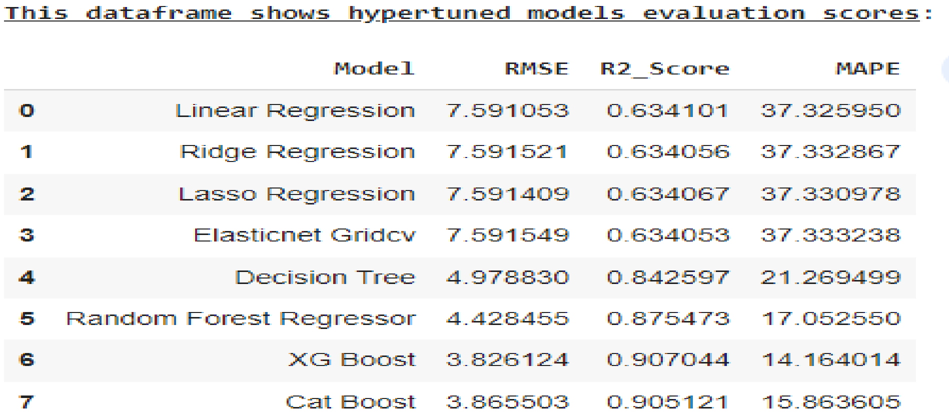


Comparing evaluation metrics of the models being used:-



Observation:-

The XG Boost model, which predicted the data more accurately, came out on top, followed by Cat Boost and Random Forest for the hyper tuned models.


This dataframe shows hypertuned models evaluation scores:

	Model	RMSE	R2_Score	MAPE
0	Linear Regression	7.591053	0.634101	37.325950
1	Ridge Regression	7.591521	0.634056	37.332867
2	Lasso Regression	7.591409	0.634067	37.330978
3	Elasticnet Gridcv	7.591549	0.634053	37.333238
4	Decision Tree	4.978830	0.842597	21.269499
5	Random Forest Regressor	4.428455	0.875473	17.052550
6	XG Boost	3.826124	0.907044	14.164014
7	Cat Boost	3.865503	0.905121	15.863605

Conclusion:-

•The results clearly suggest that Cat Boost is the best model for predicting bike sharing demand, as the performance measure (MAPE, RMSE) is lower and (R2, adjusted_R2) value is greater for XG Boost and Cat Boost.
•We may say that the **XG Boost** model helps us to anticipate the number of rental bikes more accurately.
•**Temperature** is regarded as the most important variable for the linear regression & tree based models.
•The Cat Boosting algorithm considered **Hour** feature to be its most significant factor.
•We discovered through EDA that people ride bikes more frequently in the **summer** season and also when the **winds** are strong.
•The months of **May, June, and July** can be considered to have a larger demand for rented bikes.
•There is a considerable demand for the rented bikes during the **peak office hours**. Therefore base stations can be setup near office buildings.
•In order to reduce public waiting times, the number of bikes should be raised during the summer. As a result, they can easily rent bikes whenever they need to or at any time

Problems faced during the project:

•Transforming the target variable, which was previously nonlinear, into a normal distribution with square root transformation.
•Handling the outliers.
•We transformed several categorical data into numerical features to feed them into the model and in order to train them more easily.
•Initially we got low R2 score and unsatisfied MAPE% with linear regression model, so we used some advanced tree-based and boosting models to improve our accuracy

THANK YOU