Capstone Project II

Project Title : Seoul Bike Sharing Demand Prediction

SUBMITTED BY : Sonika Baheti

CONTENTS

- Problem Statement
- Objective
- Data Overview
- Data Wrangling
- Exploratory Data Analysis (EDA)
- Models Implemented
- Model Evaluation
- Model Explainability
- Challenges faced
- Conclusion

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.

OBJECTIVE

The main objective of my study is to predict the Number of rented bikes required as per given weather conditions in the city of Seoul using different Machine Learning algorithms.

DATA OVERVIEW

Dependent variable:

- Rented Bike count - Count of bikes rented at each hour

Independent variables:

- Date : year-month-day
- Hour - Hour of he day
- Temperature-Temperature in Celsius
- Humidity - % • Windspeed - m/s
- Visibility - 10 m
- Dew point temperature – ˚ Celsius
- Solar radiation - MJ/ m2
- Rainfall - mm
- Snowfall - cm
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday/No holiday
- Functioning Day – Yes/No

Overview and Understanding of Dataset

-> info(): It informs about data columns and data types.

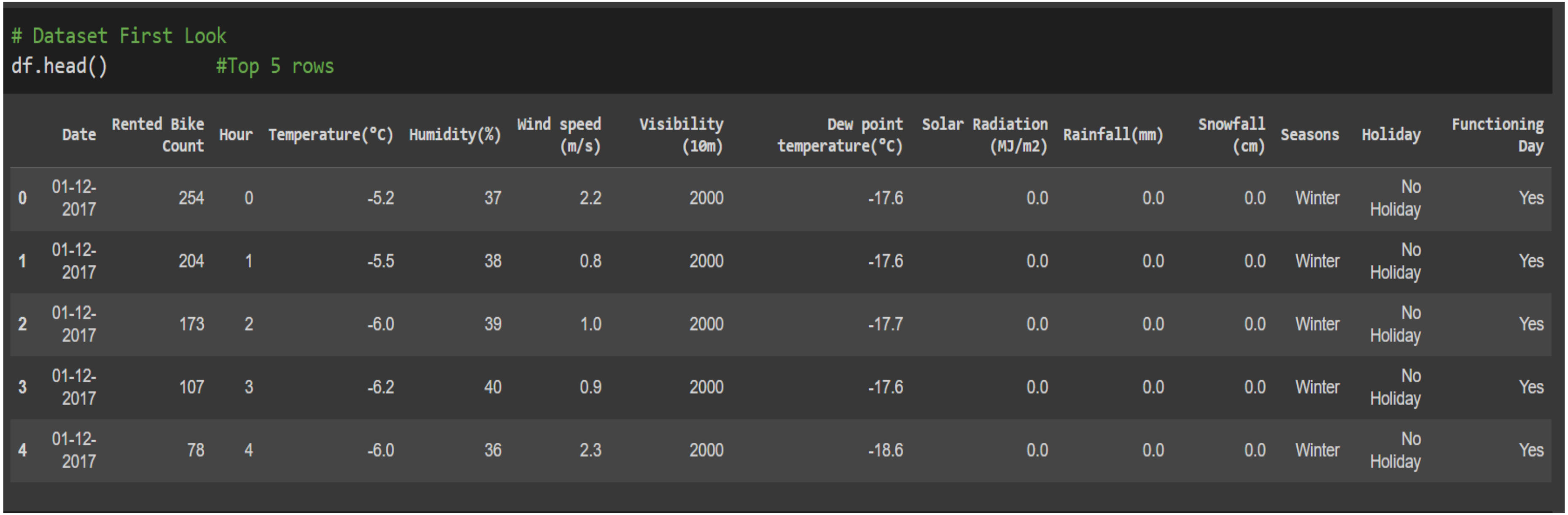
-> head(): It returns the first five data.

-> tail(): It returns the last five data.

-> columns : It returns data columns

-> shape : It gives number of rows and columns in a tuple.

Head of the Dataset



```
# Dataset First Look
df.head()          #Top 5 rows
```

	Date	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)	Seasons	Holiday	Functioning Day
0	01-12-2017	254	0	-5.2	37	2.2	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
1	01-12-2017	204	1	-5.5	38	0.8	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
2	01-12-2017	173	2	-6.0	39	1.0	2000	-17.7	0.0	0.0	0.0	Winter	No Holiday	Yes
3	01-12-2017	107	3	-6.2	40	0.9	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
4	01-12-2017	78	4	-6.0	36	2.3	2000	-18.6	0.0	0.0	0.0	Winter	No Holiday	Yes

DATA WRANGLING

- ✓ Changed the datatype to ‘Date’ column to ‘datetime’.
- ✓ Created New columns ‘Month’ and ‘weekdays_weekend’ to have more deeper analysis.
- ✓ Dropped the column ‘Date’ because I won’t be using this.

Here I’m considering the following columns as categorical columns :
['Seasons’,'Holiday', 'Functioning Day', 'month', 'weekdays_weekend’]

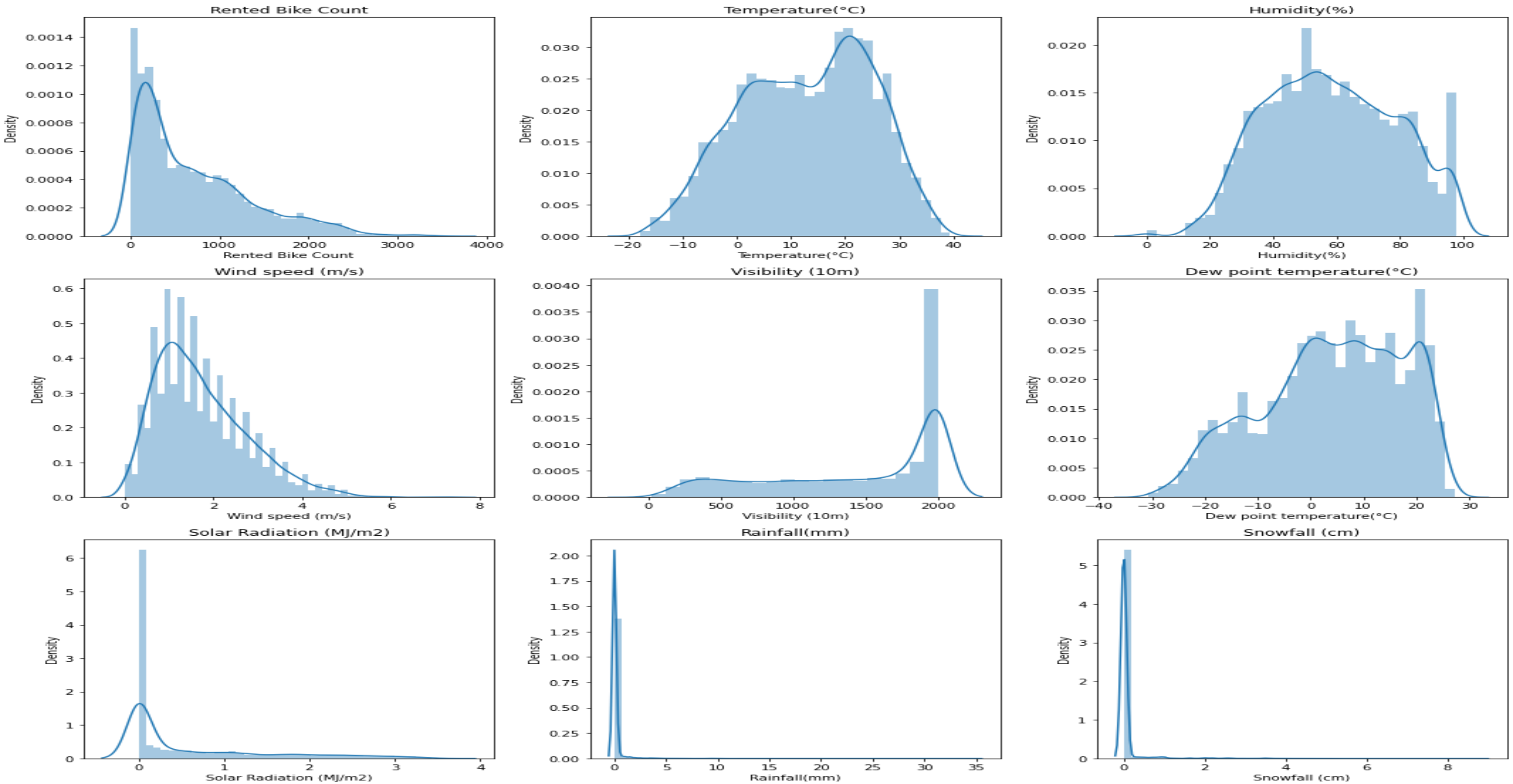
And others as numerical columns:
['Rented Bike Count', 'Temperature(°C)', 'Humidity(%)', 'Wind speed (m/s)’, ‘ Visibility (10m)', 'Dew point temperature(°C)','Solar Radiation (MJ/m2)’,'Rainfall(mm)' , 'Snowfall (cm)']

EXPLORATORY DATA ANALYSIS (EDA)

Data is MONEY
Information is POWER

1. Univariate Analysis

1.1 Distribution plot of numerical columns



✓ INFERENCE FROM DISTRIBUTION PLOT

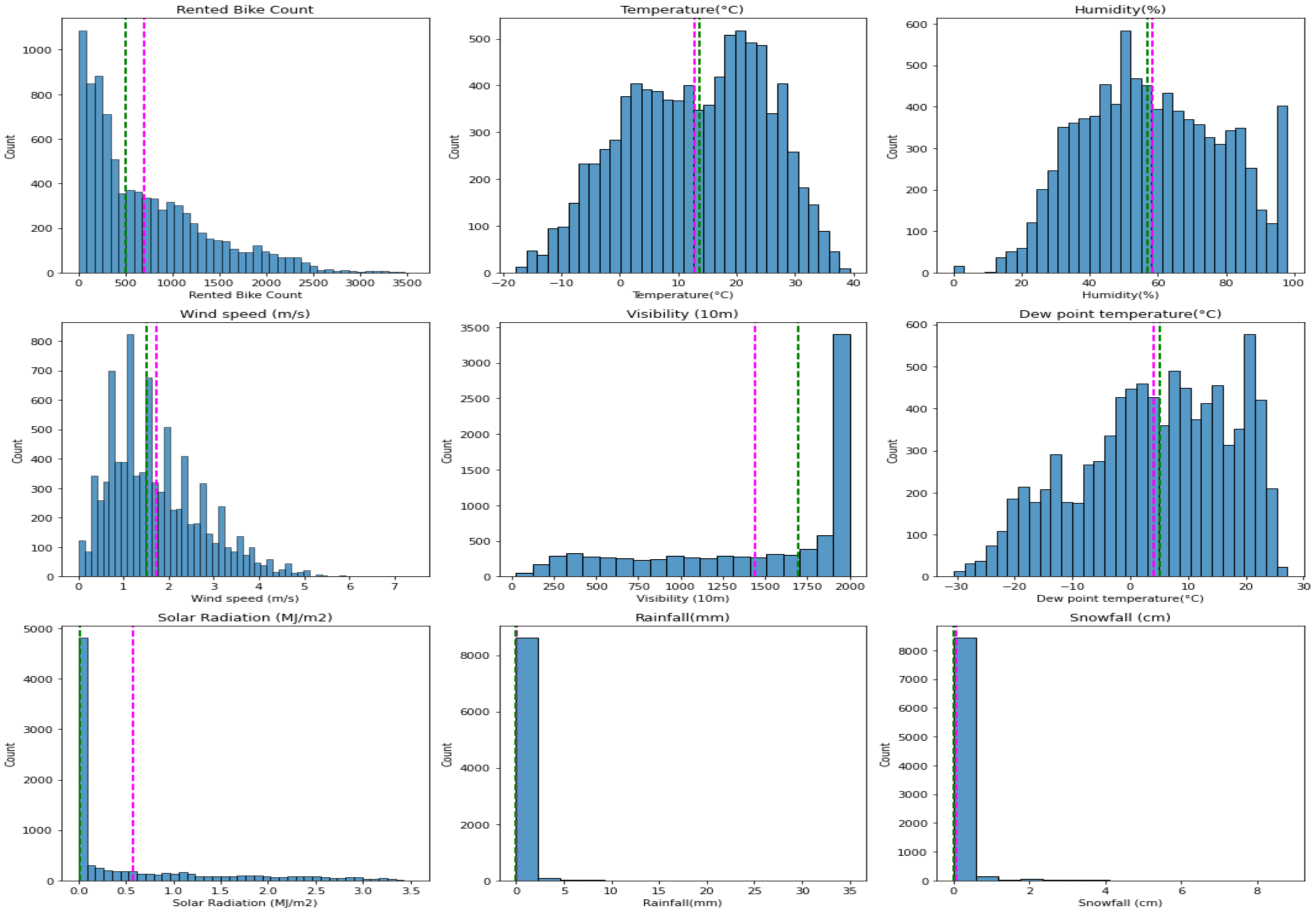
Here it is clearly visible that some of the columns are not normally distributed. These are: ['Rented Bike Count', 'Wind speed (m/s)', 'Visibility (10m)','Solar Radiation (MJ/m2)', 'Rainfall(mm)', 'Snowfall (cm)']. These columns have skewed distribution.

Approximately normally distributed : ***Temperature, Dew point temperature & Humidity***

Right skewed distributed : ***Wind speed, Solar Radiation, Rainfall, Snowfall***

Left skewed : ***Visibility***

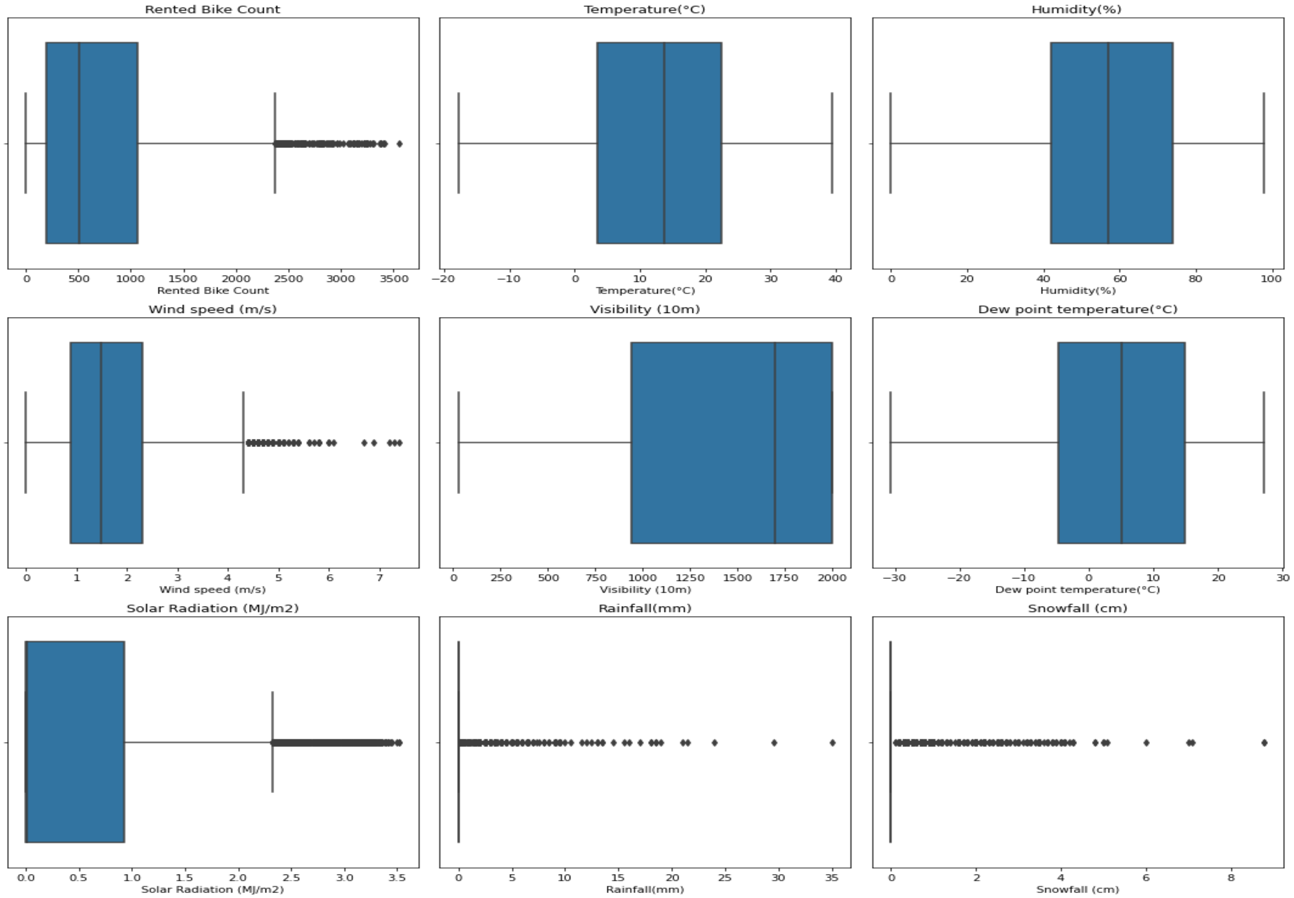
1.2 Histogram with mean and median axes



INFERENCES

Here Rented bike count, wind speed, visibility and solar radiation columns do not have mean and median on same axes.

1.3 Boxplot of numerical columns



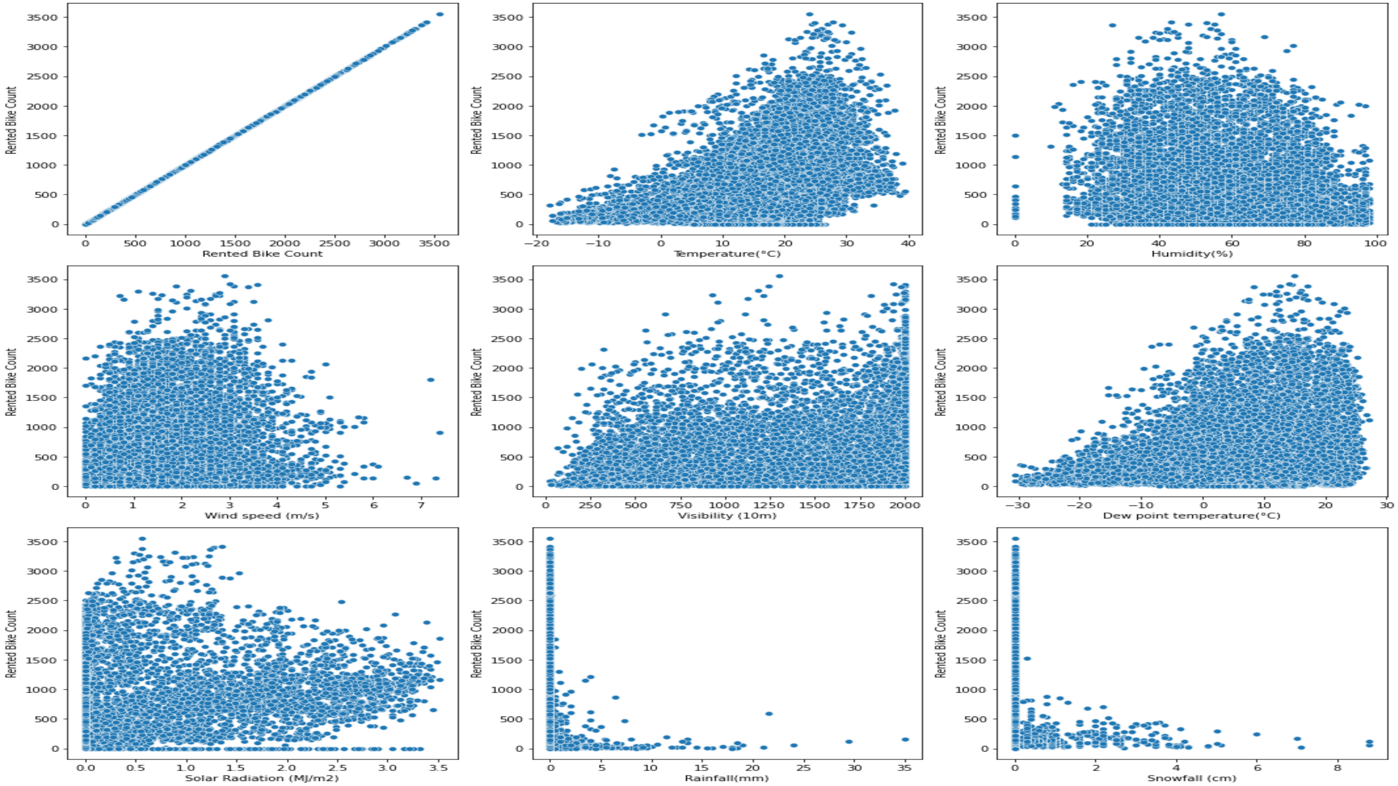
INFERENCE

Outliers present in some columns. These are:

['Rented Bike Count' , 'Wind speed (m/s)' , 'Solar Radiation (MJ/m2)' , 'Rainfall(mm)' , 'Snowfall (cm)']

2. Bivariate Analysis

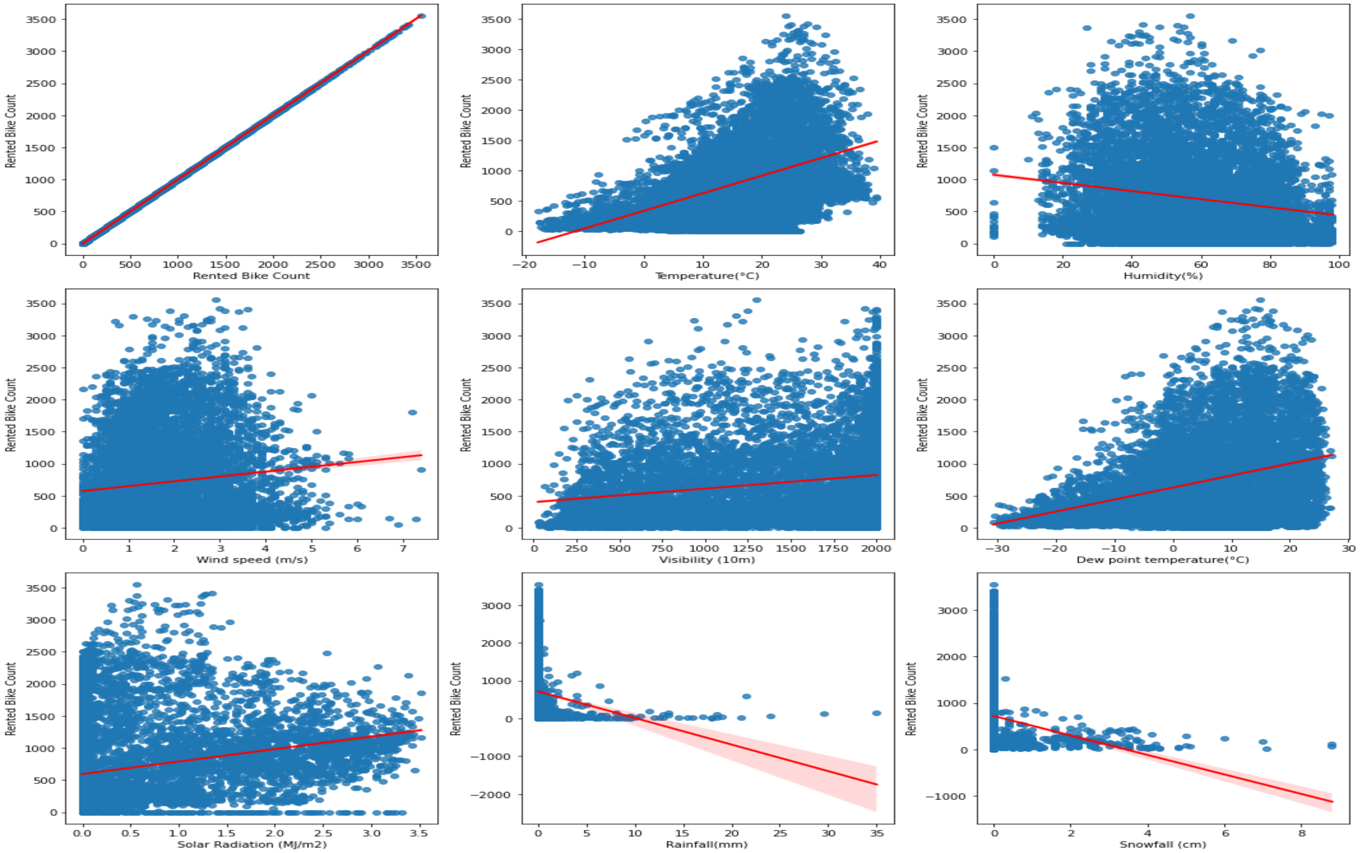
2.1 Scatter plot between Numerical Independent variables and Dependent variable



INFERENCES:

We can see the relationship between Dependent variable('Rented Bike Count') and other independent variables.

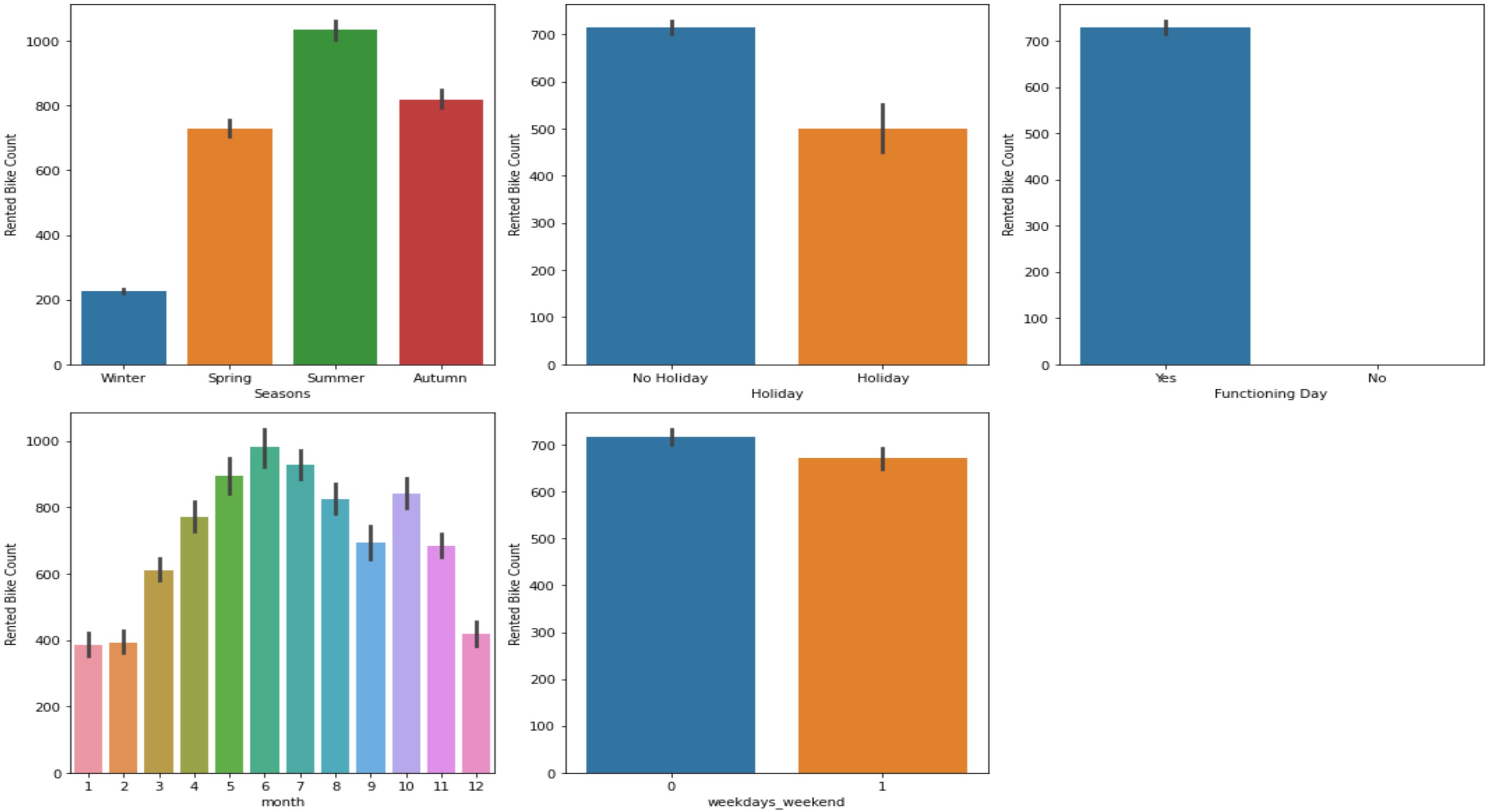
2.2 Scatter plot with line plot



INFERENCE:

Here some of the variables have positive effect on dependent variable while some have negative impact.

2.3 Bar plot of Categorical columns



INFERENCE

1.Seasons : Of all the four seasons Rented bikes are mostly used in summer season followed by Autumn season. Which means that people used to ride bike in pleasant weather.

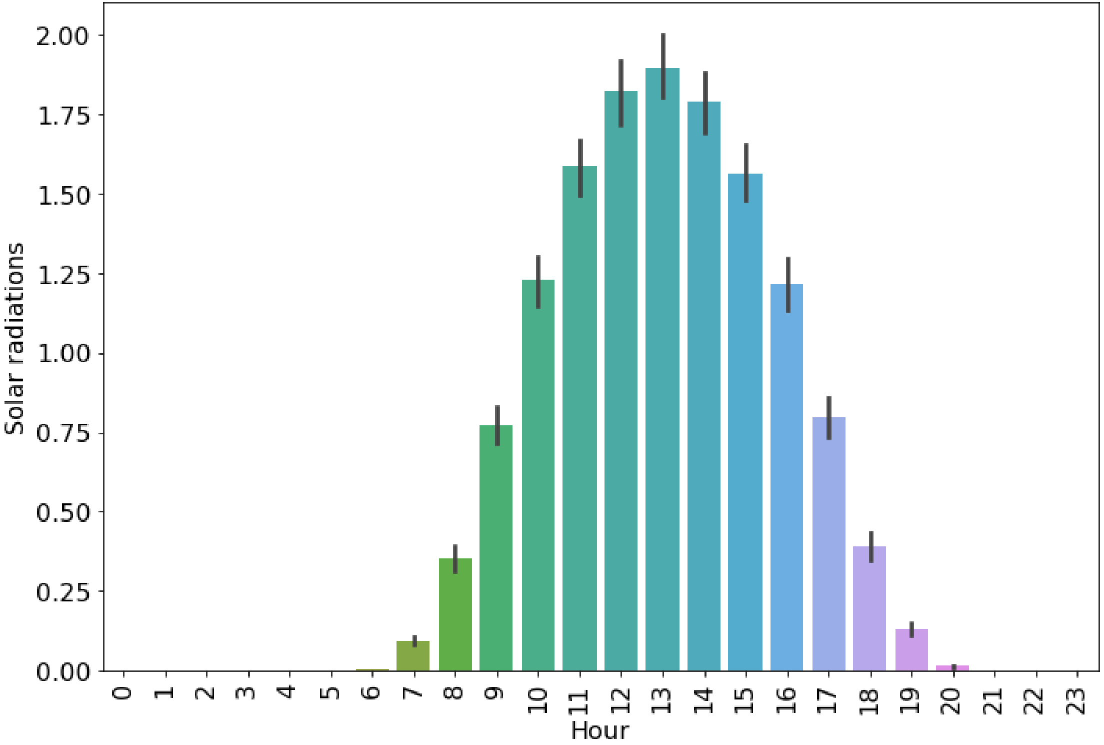
2.Holiday : Count of rented bikes is more during non Holiday. Hence bikes are used for work or office purpose.

3.Functioning Day : Rented bikes are mostly used during functioning days. Hence used mostly for office or work purpose.

4.Month : Rented bikes are less used during December, January & February ,i.e., Winters. and mostly used in the month of May, June ,July & October.

5.Weekends : During weekends bikes are comparatively less used than weekdays.

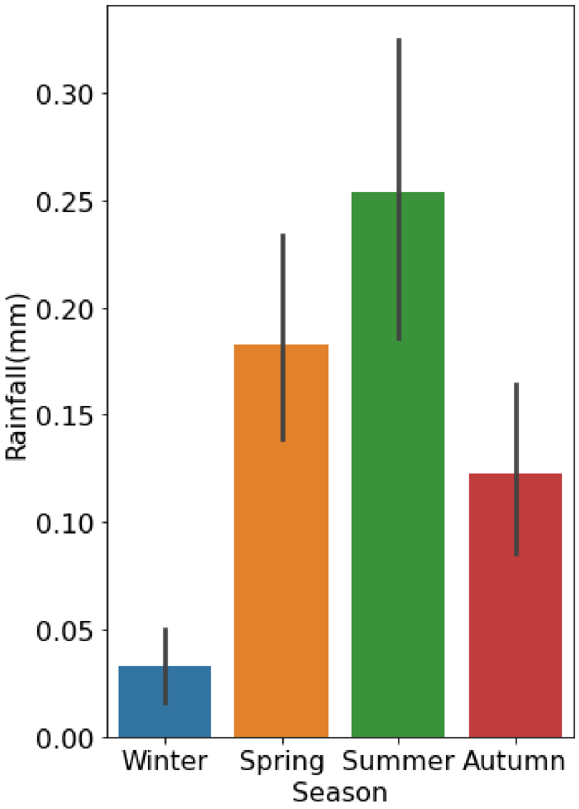
2.5 Solar radiations and hour:

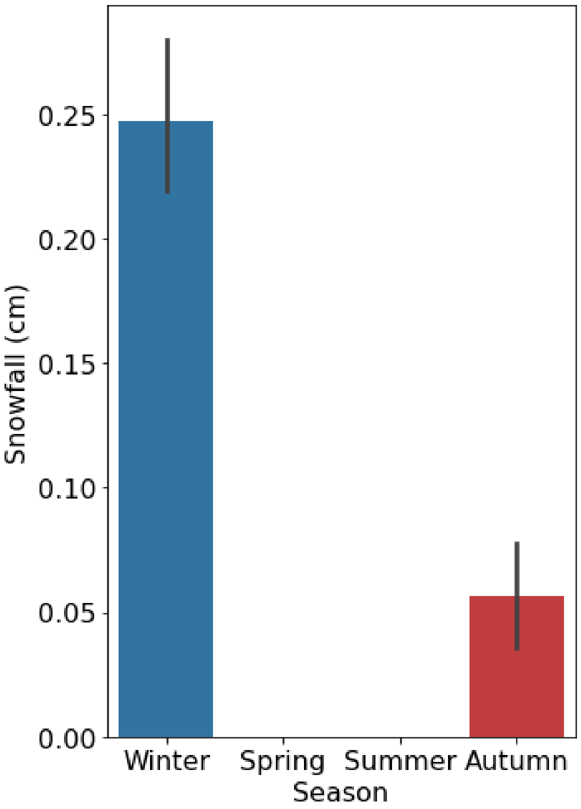


INFERENCE:

Sun hours in the day are lesser than non-sun hours. Hence Solar radiation is not an outlier.

Does Rainfall & Snowfall depends upon Season?



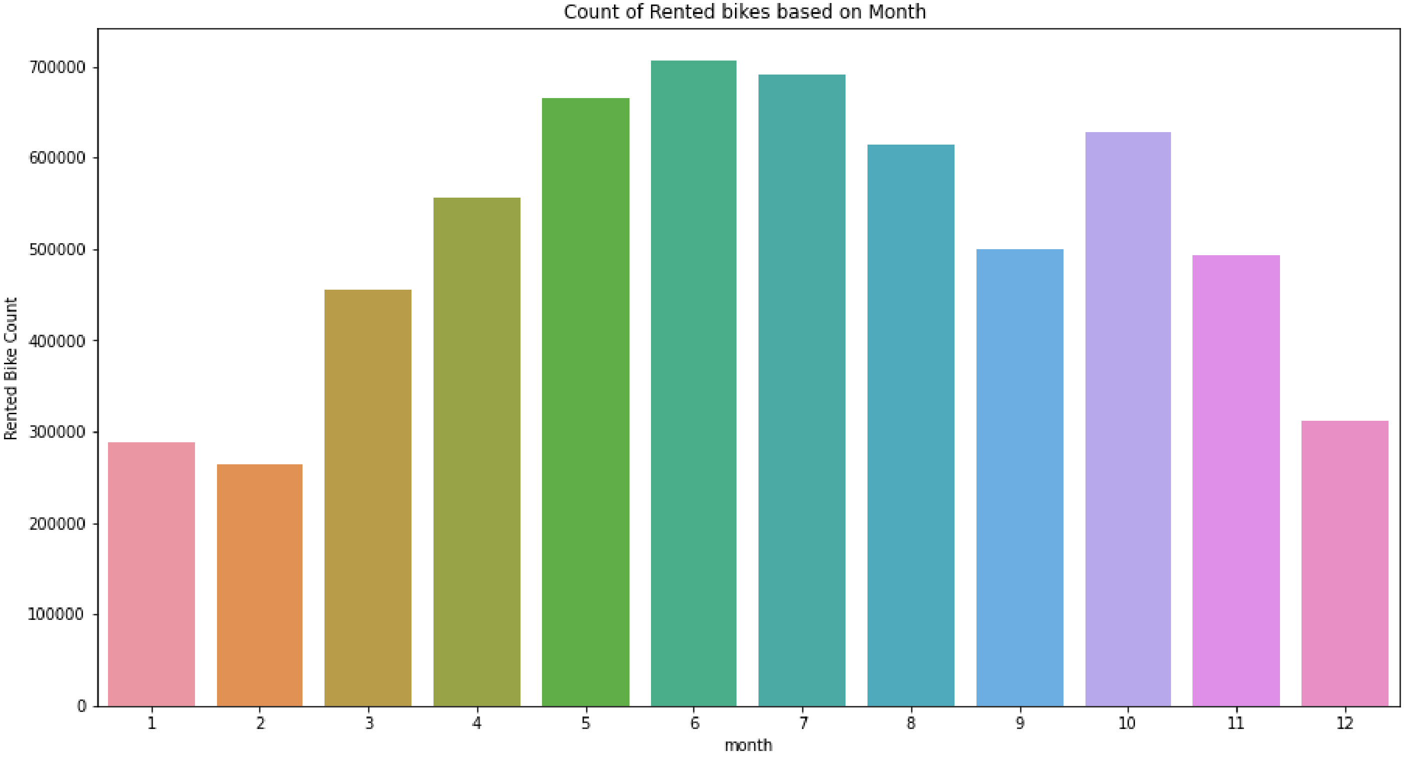


INFERENCE;

Clearly , Rainfall & Snowfall is dependent upon season. Because very little rainfall in winters is observed. And there is almost no snowfall in Spring & Summer.

Hence these columns are not outliers.

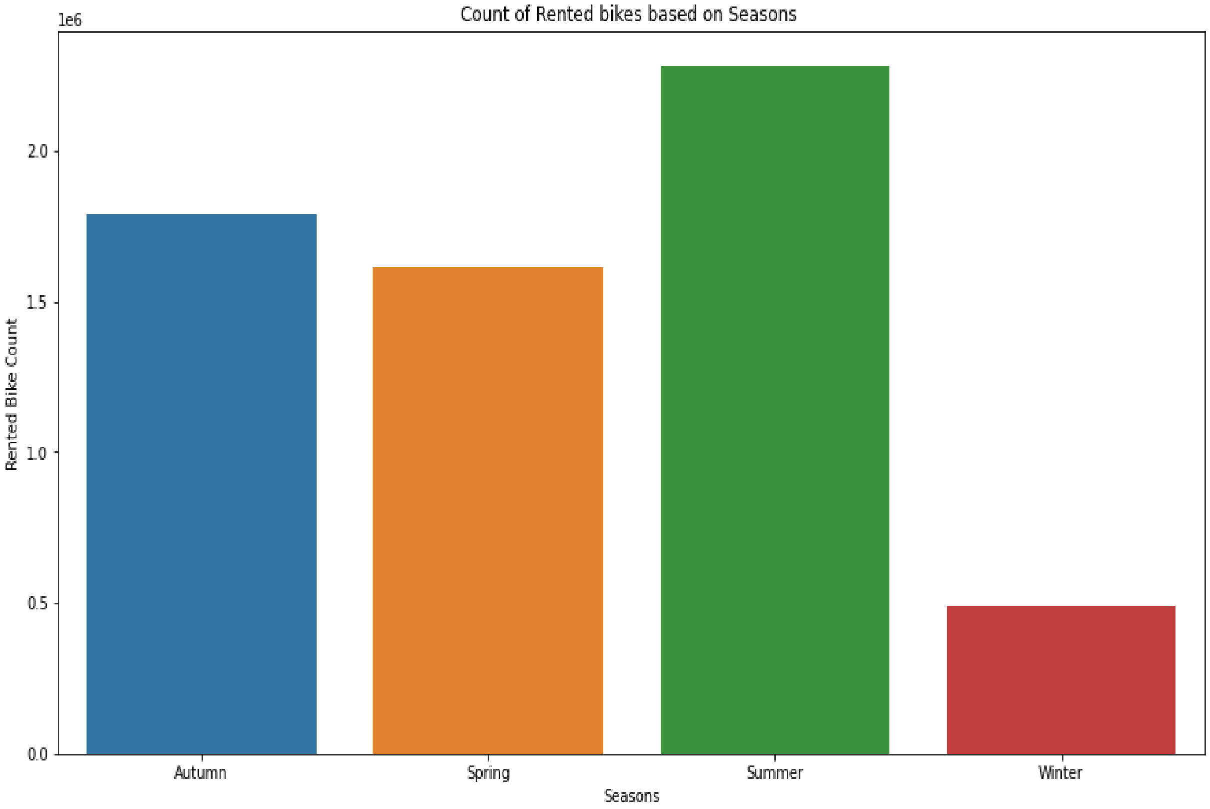
Which month have the most demand for rented bike count?



INFERENCE:

The demand for rented bike count is highest in the month of May, June, July and lowest in the month of December, January, February.

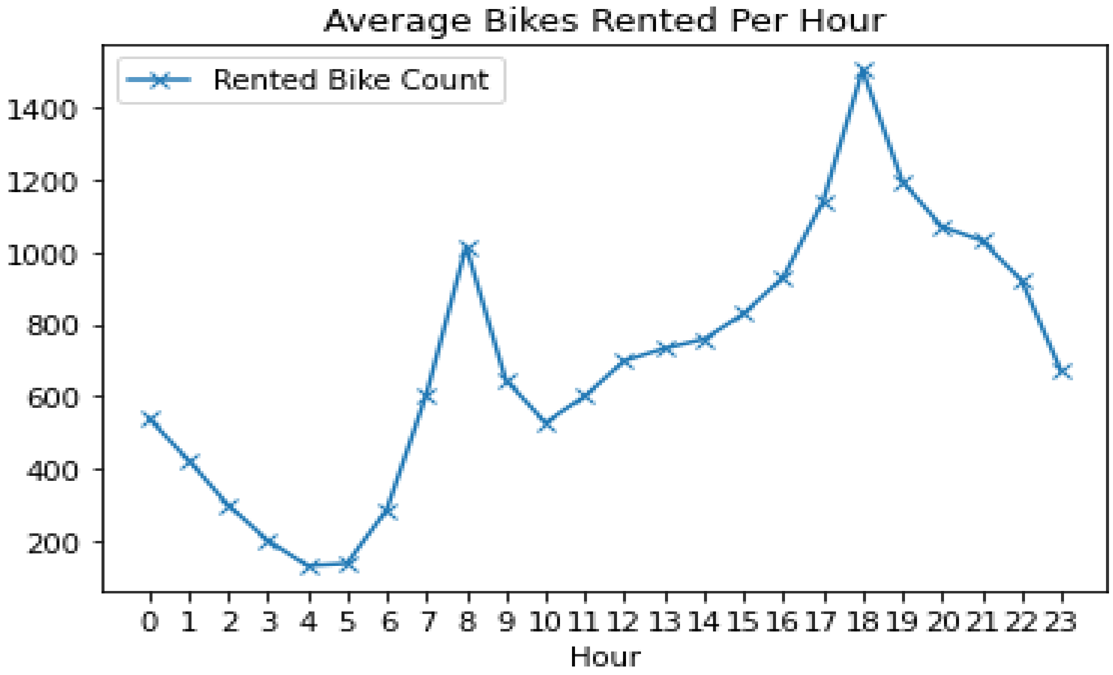
Demand of Rented Bikes for different Seasons



INFERENCE:

The demand for rented bike is high in summer season while the demand for rented bike is low in winter season. It means that people like to have bike ride in convincing season.

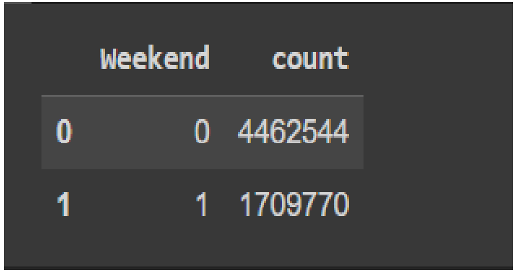
Average count of rented bike in each hour



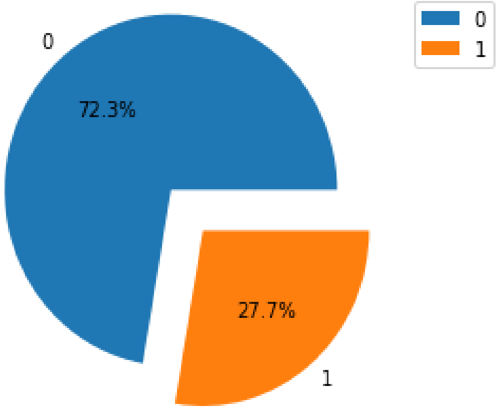
INFERENCE:

High rise of Rented Bikes in the morning from 7:00 to 9:00 am & in the evening from 5:00 pm to 9:00 pm means people prefer rented bike during rush hour.

Effect of weekends on demand of bike



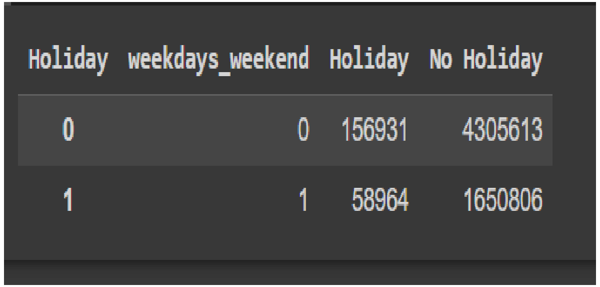
	Weekend	count
0	0	4462544
1	1	1709770



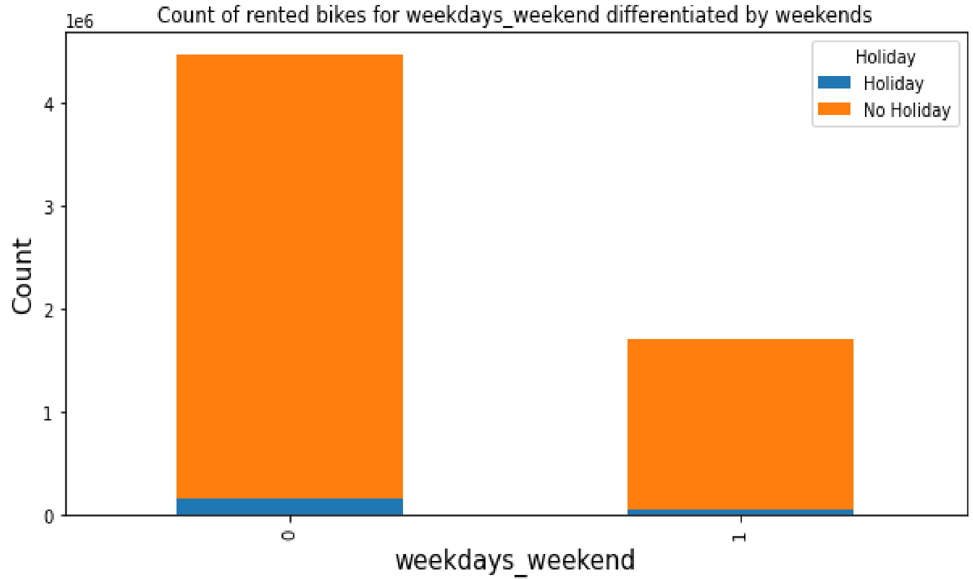
INFERENCE:

Rented bikes are approximately 30% used by riders on weekends. now find out rented bikes count if weekend was holiday or not.

Rented Bike in weekend holidays



Holiday	weekdays_weekend	Holiday	No Holiday
0	0	156931	4305613
1	1	58964	1650806



INFERENCE:

Rented bikes are mostly used in non weekend. During holidays on weekends people generally do not use much rented bikes.

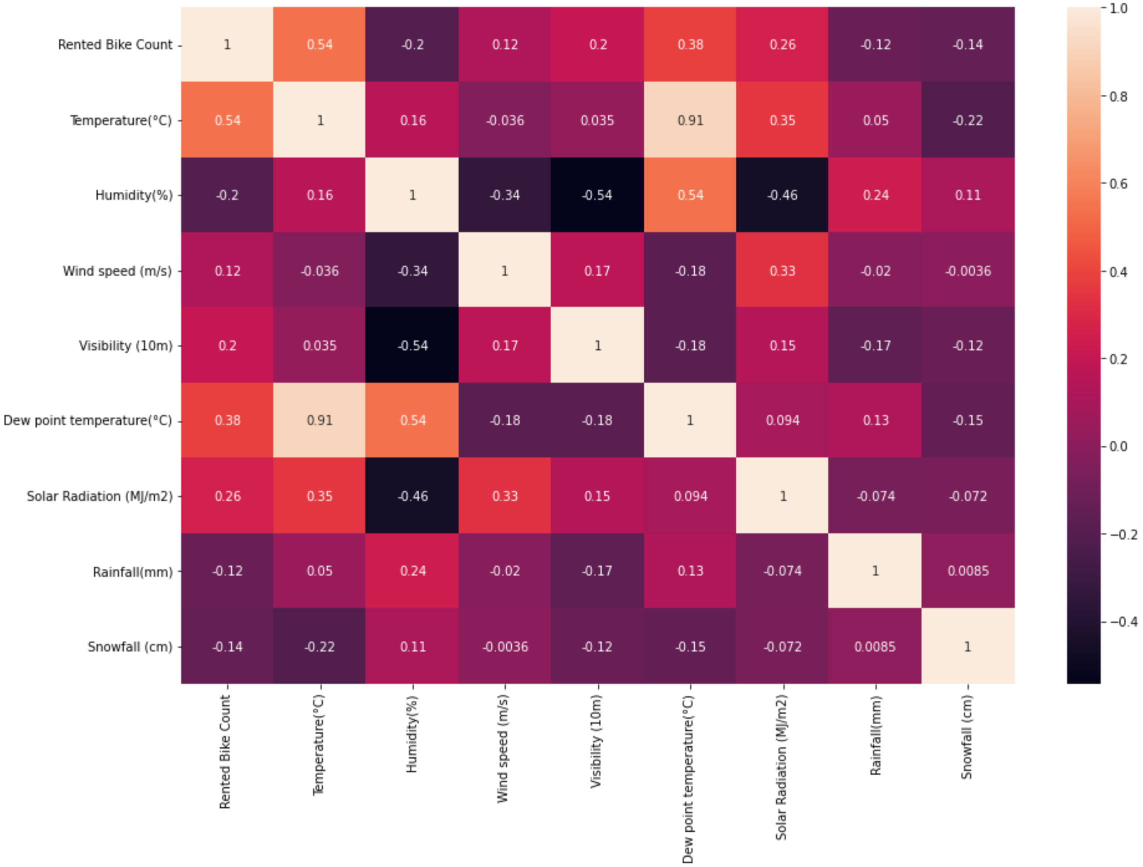
Which means that people generally use rented bike for travelling to office.

3. Multivariate Analysis

3.1 Correlation

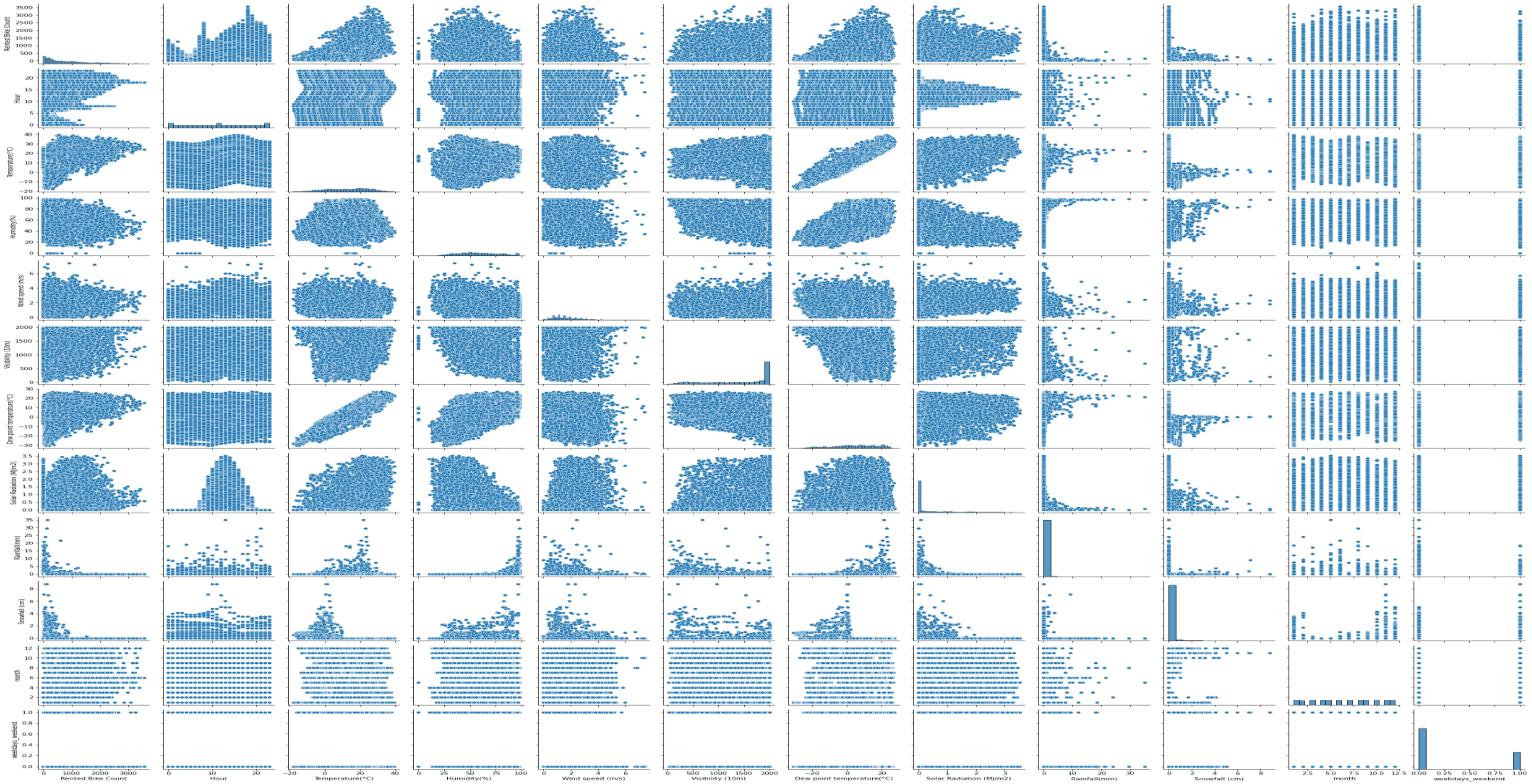
INFERENCE:

There is high correlation between Dew point temperature and Temperature , i.e., 0.91



	Rented Bike Count	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)
Rented Bike Count	1	0.54	-0.2	0.12	0.2	0.38	0.26	-0.12	-0.14
Temperature(°C)	0.54	1	0.16	-0.036	0.035	0.91	0.35	0.05	-0.22
Humidity(%)	-0.2	0.16	1	-0.34	-0.54	0.54	-0.46	0.24	0.11
Wind speed (m/s)	0.12	-0.036	-0.34	1	0.17	-0.18	0.33	-0.02	-0.0036
Visibility (10m)	0.2	0.035	-0.54	0.17	1	-0.18	0.15	-0.17	-0.12
Dew point temperature(°C)	0.38	0.91	0.54	-0.18	-0.18	1	0.094	0.13	-0.15
Solar Radiation (MJ/m2)	0.26	0.35	-0.46	0.33	0.15	0.094	1	-0.074	-0.072
Rainfall(mm)	-0.12	0.05	0.24	-0.02	-0.17	0.13	-0.074	1	0.0085
Snowfall (cm)	-0.14	-0.22	0.11	-0.0036	-0.12	-0.15	-0.072	0.0085	1

3.2 Pairplot



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
weekdays_weekend

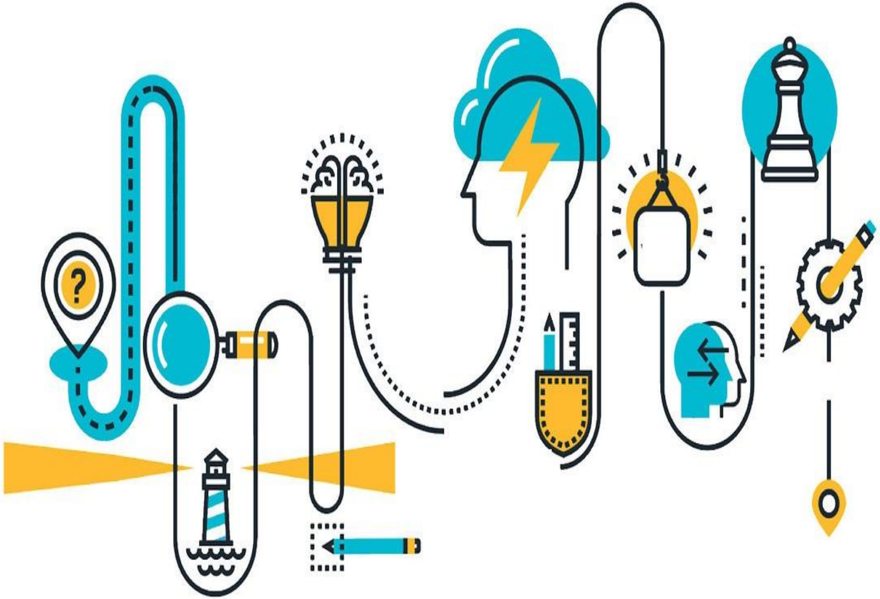
MODELS IMPLEMENTATION

A. LINEAR MODELS

1. Simple Multiple Linear Regression
2. Ridge Regression
3. Lasso Regression
4. Elastic net Regression
5. Polynomial Regression

B. TREE BASED MODELS

1. Decision Tree Regressor
2. Random Forest Regressor
3. Extreme Gradient Boosting



Data Preparation for Evaluating Models

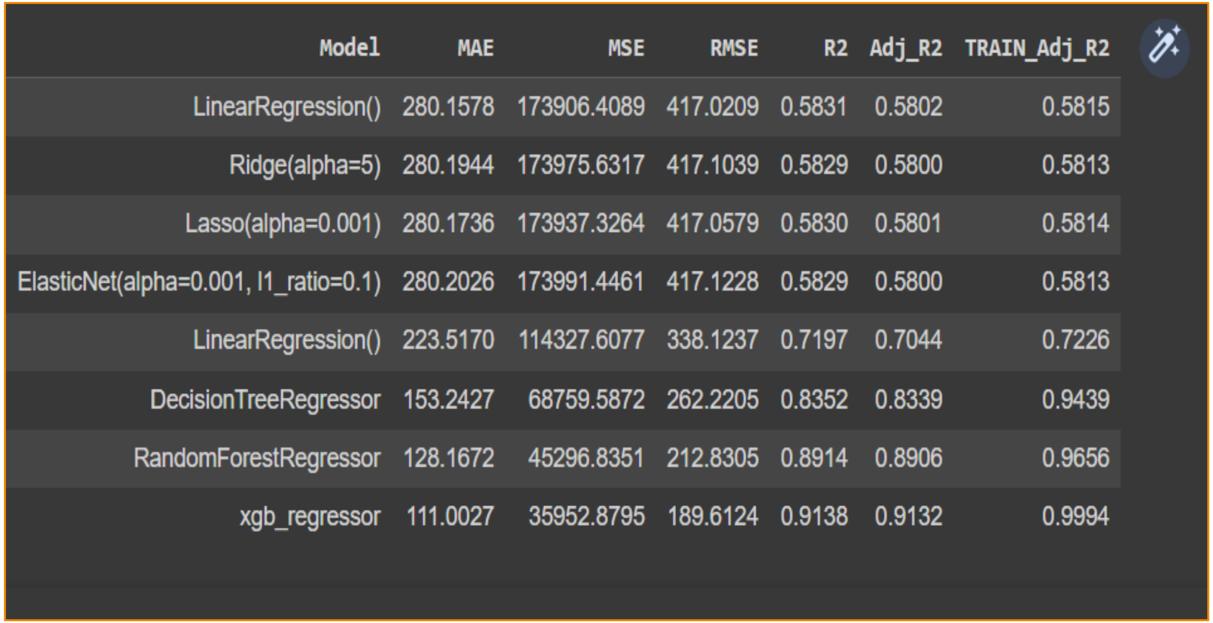
- ❑ Checking for Multi-collinearity : Using variance inflation factor and found high multicollinearity between ‘Temperature’ and ‘Dew point temperature’. So dropped ‘Dew point temperature’.

- ❑ One hot encoding One hot encoding of categorical columns ‘Seasons’, ‘Holiday’, ‘Functioning Day’.

- ❑ Transformation: Square root transformation of Dependent variable to normalize it.
  StandardScalar transformation of independent variables due to presence of outliers.

- ❑ Splitting data : Splitting data to train and test with test size of 0.25.

MODEL EVALUATION



Model	MAE	MSE	RMSE	R2	Adj_R2	TRAIN_Adj_R2
LinearRegression()	280.1578	173906.4089	417.0209	0.5831	0.5802	0.5815
Ridge(alpha=5)	280.1944	173975.6317	417.1039	0.5829	0.5800	0.5813
Lasso(alpha=0.001)	280.1736	173937.3264	417.0579	0.5830	0.5801	0.5814
ElasticNet(alpha=0.001, l1_ratio=0.1)	280.2026	173991.4461	417.1228	0.5829	0.5800	0.5813
LinearRegression()	223.5170	114327.6077	338.1237	0.7197	0.7044	0.7226
DecisionTreeRegressor	153.2427	68759.5872	262.2205	0.8352	0.8339	0.9439
RandomForestRegressor	128.1672	45296.8351	212.8305	0.8914	0.8906	0.9656
xgb_regressor	111.0027	35952.8795	189.6124	0.9138	0.9132	0.9994

Root Mean squared error is lowest in XGB regressor, also
Best adjusted R^2 value is obtained in extreme gradient boosting regressor.

It indicates that XGB is performing best among all the other models.

What does different matrices imply?

1. **MAE (Mean Absolute Error):** It is the average of absolute differences of actual and predicted value.

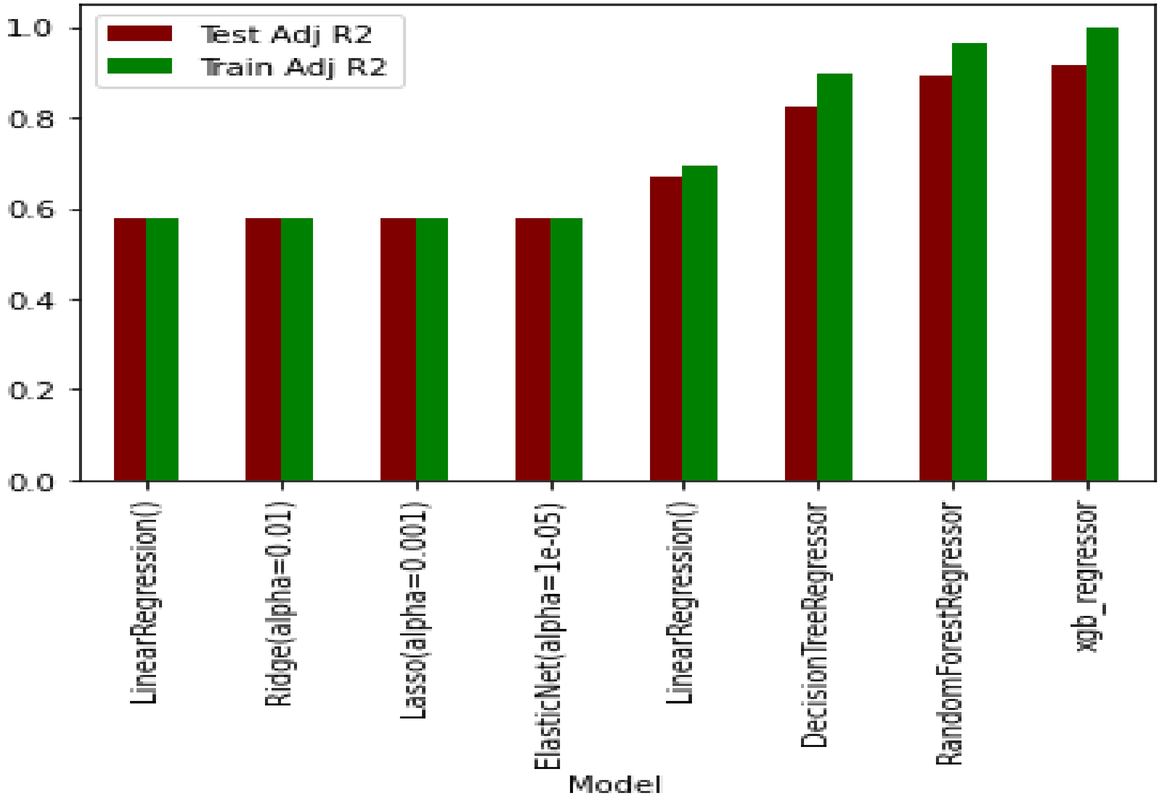
2. **MSE (Mean Squared Error):** It is the average of square of difference of actual and predicted value.

3. **RMSE (Root mean squared error):** It is the square root of mean squared error.

4. **R2 (R square) :** It determines the proportion of variance in the dependent variable that can be explained by the independent variable. It is also called as the coefficient of determination.

5. **Adjusted R2 :** Adjusted R Square determines the extent of the variance of the dependent variable, which the independent variable can explain.

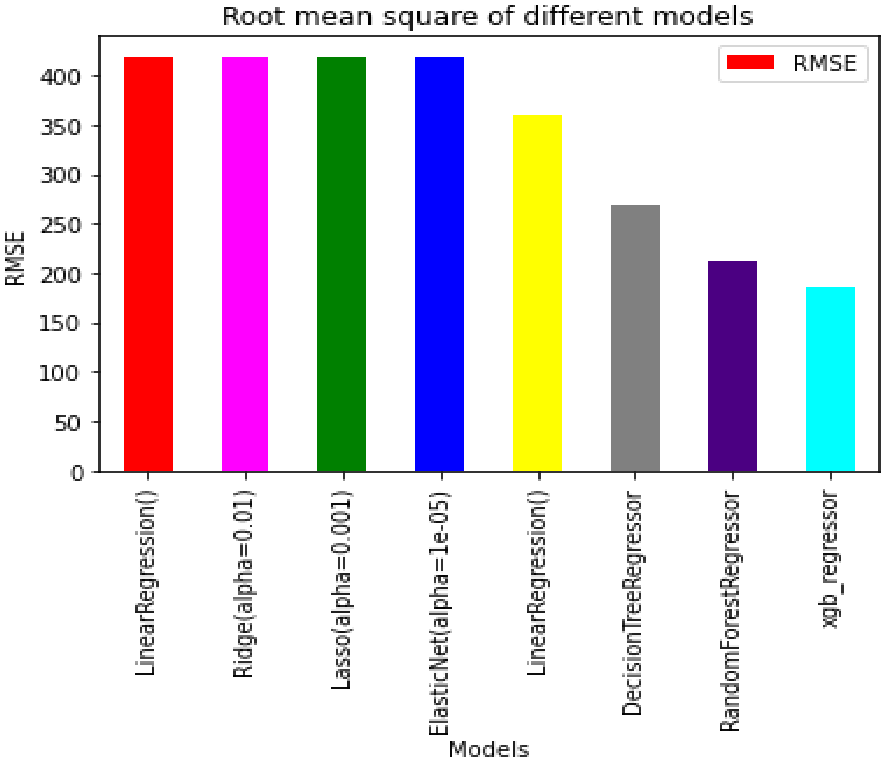
Visual comparison of Test adjusted R^2 AND Train adjusted R^2



INFERENCE:

Extreme gradient boosting is performing best in training and testing data.

Visual comparison of Root Mean Squared Error of different models



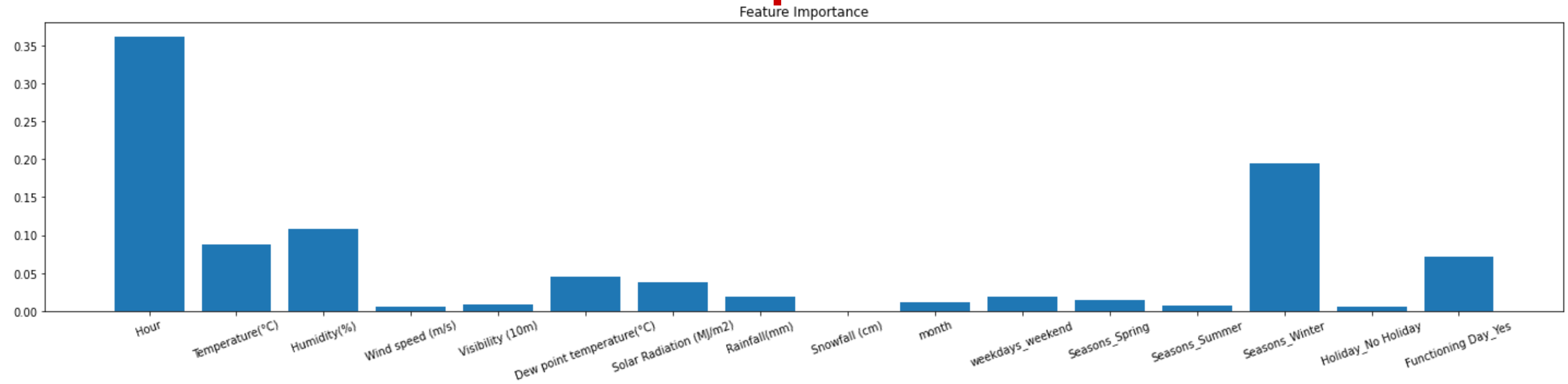
INFERENCE:

It is clearly visible that Root mean squared value of XGB is lowest among all other models

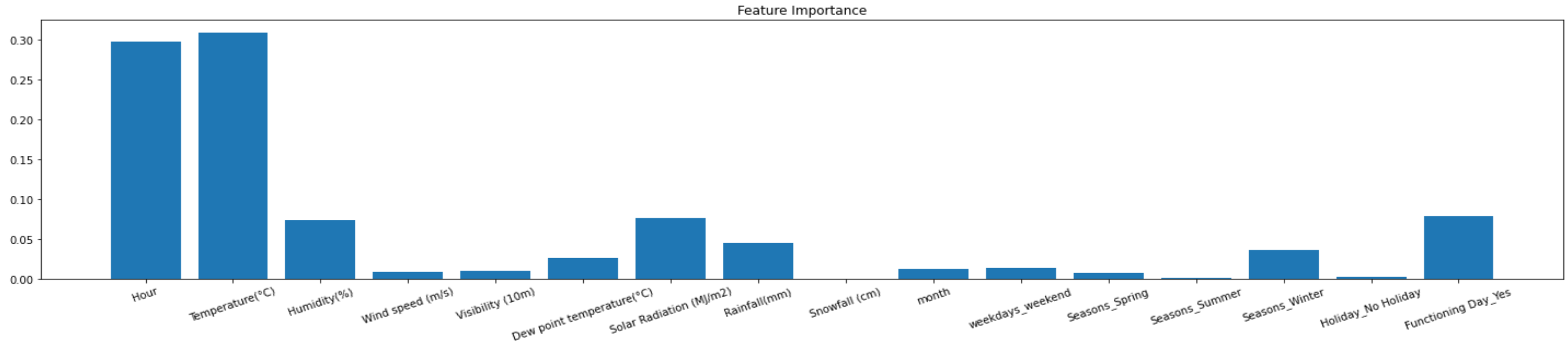
Model Validation $ Selection

- ***Observation 1***: As seen in the Model Evaluation Matrices table, Linear Regression (Multiple and also regularized) is not giving great results because of lack of linear dependency among variables.
- ***Observation 2***: Random forest & Decision Tree have performed equally good in terms of adjusted r2.
- ***Observation 3***: We are getting the best results from XGBoost.

Feature importance

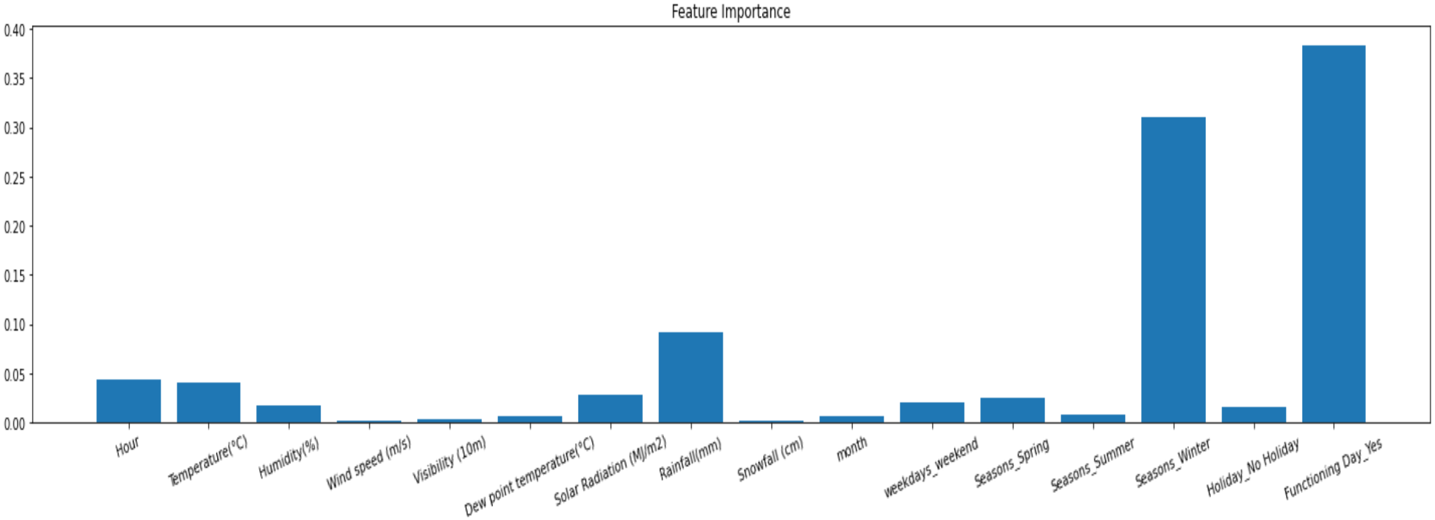


Decision Tree Regressor



Random Forest Regressor

Feature importance



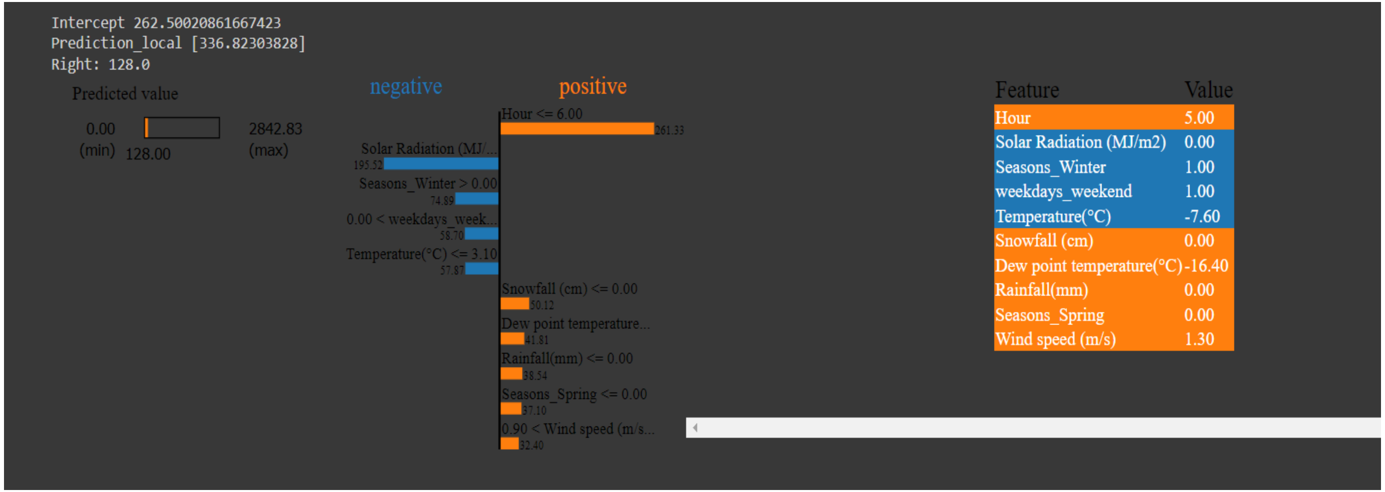
Extreme Gradient Boosting Regressor

INFERENCE:

- Temperature and hour are the most important features observed in Decision Tree and Random Forest Regressor.
- While functioning Day is the most important feature obtained in XGB model.

MODEL EXPLAINABILITY

1. Decision Tree Regressor

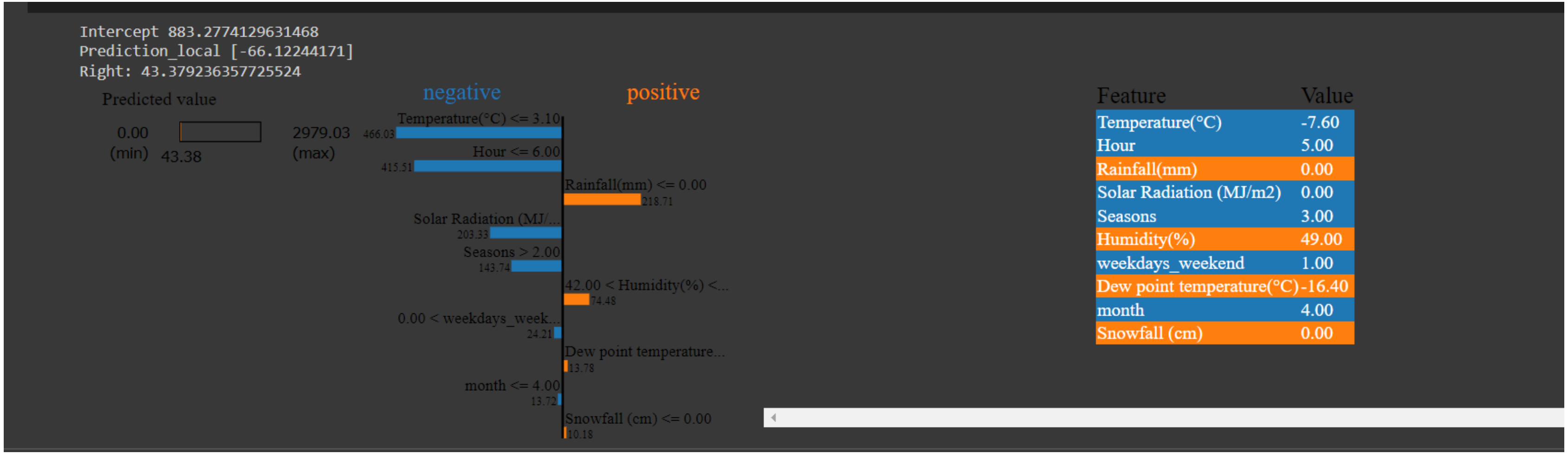


```
Intercept 262.50020861667423
Prediction_local [336.82303828]
Right: 128.0
```

Feature	Value
Hour	5.00
Solar Radiation (MJ/m2)	0.00
Seasons_Winter	1.00
weekdays_weekend	1.00
Temperature(°C)	-7.60
Snowfall (cm)	0.00
Dew point temperature(°C)	-16.40
Rainfall(mm)	0.00
Seasons_Spring	0.00
Wind speed (m/s)	1.30

For the particular observation no 17, I used model explainability technique (LIME). In Decision tree, Hour is positively affecting the rented bike count while solar radiation is negatively affecting the rented bikes.

2. Random Forest Regressor

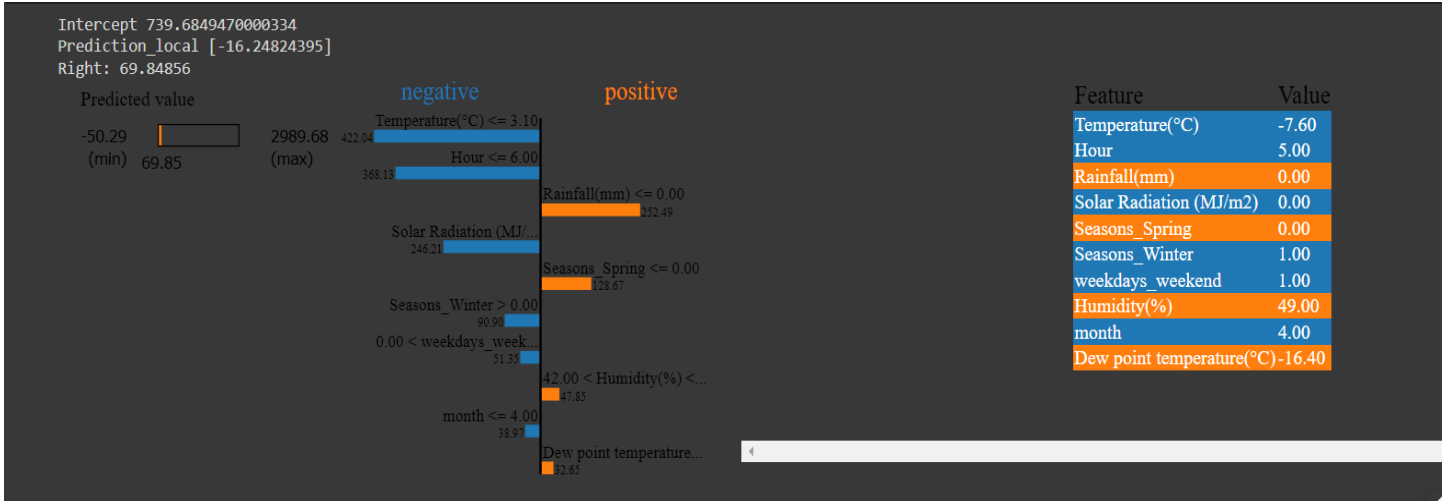


```
Intercept 883.2774129631468
Prediction_local [-66.12244171]
Right: 43.379236357725524
```

Feature	Value
Temperature(°C)	-7.60
Hour	5.00
Rainfall(mm)	0.00
Solar Radiation (MJ/m2)	0.00
Seasons	3.00
Humidity(%)	49.00
weekdays_weekend	1.00
Dew point temperature(°C)	-16.40
month	4.00
Snowfall (cm)	0.00

For the particular observation no 17, I used model explainability technique (LIME). In Random forest regressor, Temperature, Hour is negatively affecting the rented bike count while rainfall is positively affecting the rented bikes.

3. Extreme Gradient Boosting Regressor



```
Intercept 739.6849470000334
Prediction_local [-16.24824395]
Right: 69.84856
```

Feature	Value
Temperature(°C)	-7.60
Hour	5.00
Rainfall(mm)	0.00
Solar Radiation (MJ/m2)	0.00
Seasons_Spring	0.00
Seasons_Winter	1.00
weekdays_weekend	1.00
Humidity(%)	49.00
month	4.00
Dew point temperature(°C)	-16.40

For the particular observation no 17, I used model explainability technique (LIME). In XGB regressor, Temperature, Hour is negatively affecting the rented bike count while rainfall is positively affecting the rented bikes. Solar radiations are also affecting negativel for this observation.

CONCLUSION

1. FROM EDA

- ➢ Some of the columns are not normally distributed. These are: ['Rented Bike Count', 'Wind speed (m/s)', 'Visibility (10m)','Solar Radiation (MJ/m2)', 'Rainfall(mm)', 'Snowfall (cm)']. These columns have skewed distribution.
- ➢ Rented bike count, wind speed, visibility and solar radiation columns do not have mean and median on same axes.
- ➢ Outliers present in some columns. These are: 'Rented Bike Count' , 'Wind speed (m/s)' , 'Solar Radiation (MJ/m2)' , 'Rainfall(mm)' , 'Snowfall (cm)'.
- ➢ No linear relationship between Dependent variable('Rented Bike Count') and other independent variables is observed.
- ➢ Some variables have positive effect on dependent variable while some have negative effect.

1.Seasons : Of all the four seasons Rented bikes are mostly used in summer season followed by Autumn season.

2.Holiday : Count of rented bikes is more during non Holiday. Hence bikes are used for work or office purpose.

3.Functioning Day : Rented bikes are mostly used during functioning days. Hence used mostly for office or work purpose.

4Month : Rented bikes are less used during December, January & February ,i.e., Winters. and mostly used in the month of May, June ,July & October

5.Weekends : During weekends bikes are comparatively less used than weekdays.

2. From Modelling:

- Since there do not seem any linear relationship between independent and dependent variables, hence linear models are not giving great results
- From model evaluation, adjusted R^2 of extreme gradient boosting is maximum among all other models.
- Root mean squared error of XGB regressor is also minimum among all models.

Hence XGB regressor is the best performing model.
If the model interpretability is important to the stakeholders, we can choose deploy XGB model.

CHALLENGES FACED

- Comprehending the problem statement, and understanding the business implications.
- Feature engineering – deciding on which features to be dropped / kept / transformed so that assumptions of linear models is satisfied.
- Choosing the best visualization to show the trends among different features clearly in the EDA phase.
- Deciding on how to handle outliers.
- Choosing the ML models to make predictions.
- High computational time in hyper parameter tuning for ensemble models.

THANK YOU!