CAPSTONE PROJECT

BIKE SHARING DEMAND PREDICTION

TEAM MEMBERS

KIRTESH VERMA

PRAVIN BEJJO

SAHIL PARDESHI

➔ STEPS

- ❖ Problem statement
- ❖ Overview of the data
- ❖ EDA
- ❖ Data Preprocessing
- ❖ Model building
- ❖ Evaluation of Models
- ❖ Conclusion

➔ PROBLEM STATEMENT

Rented bikes are introduced in many urban areas for the enhancement of comfort and mobility. Our key objective is to predict the count of bike required at each hour throughout the day so that bikes are available and accessible to the public at the right time and thus reducing the waiting time.

➔ DATA OVERVIEW

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

Columns:

- Date : year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of he day
- Temperature-Temperature in Celsius
- Humidity - %
- Windspeed - m/s
- Visibility - 10m
- Dew point temperature - Celsius
- Solar radiation - MJ/m2
- Rainfall - mm
- Snowfall - cm
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday/No holiday
- Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours)

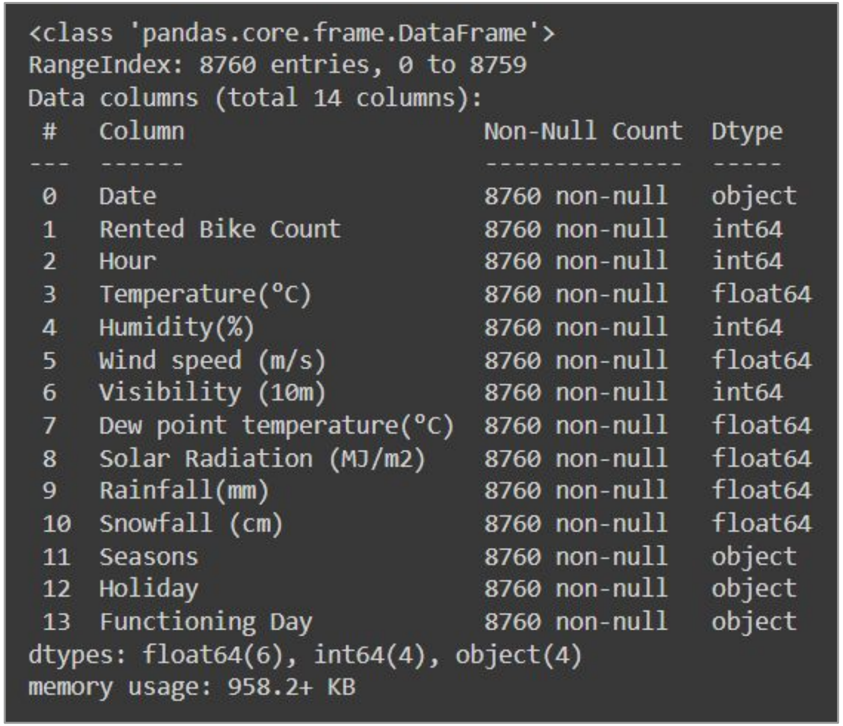
➔ DATA OVERVIEW (Conti…)

- **DATA INSIGHTS**

❖ The given dataset has 8760 entries and 14 columns.

❖ Seasons, Holiday, Functioning day and Date are the four categorical columns in our dataset.

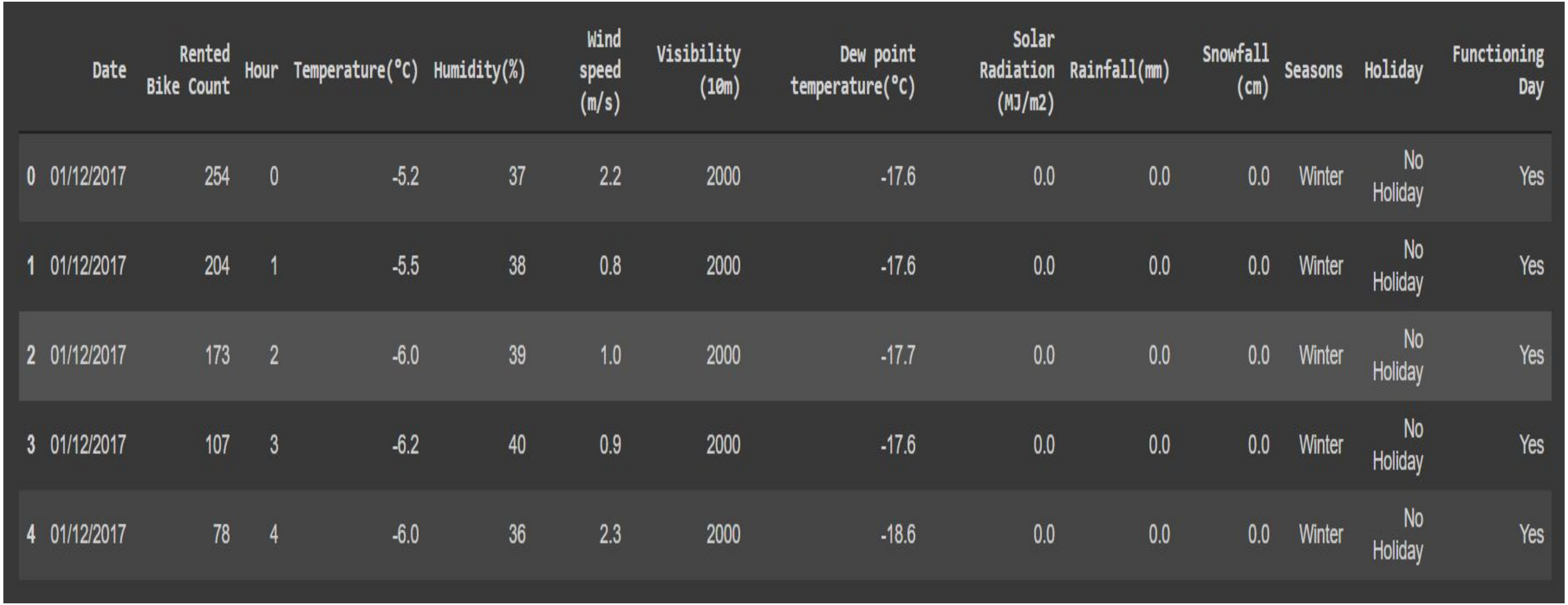
❖ There are no null or duplicate values.



```
<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
```

DATA OVERVIEW (Conti…)

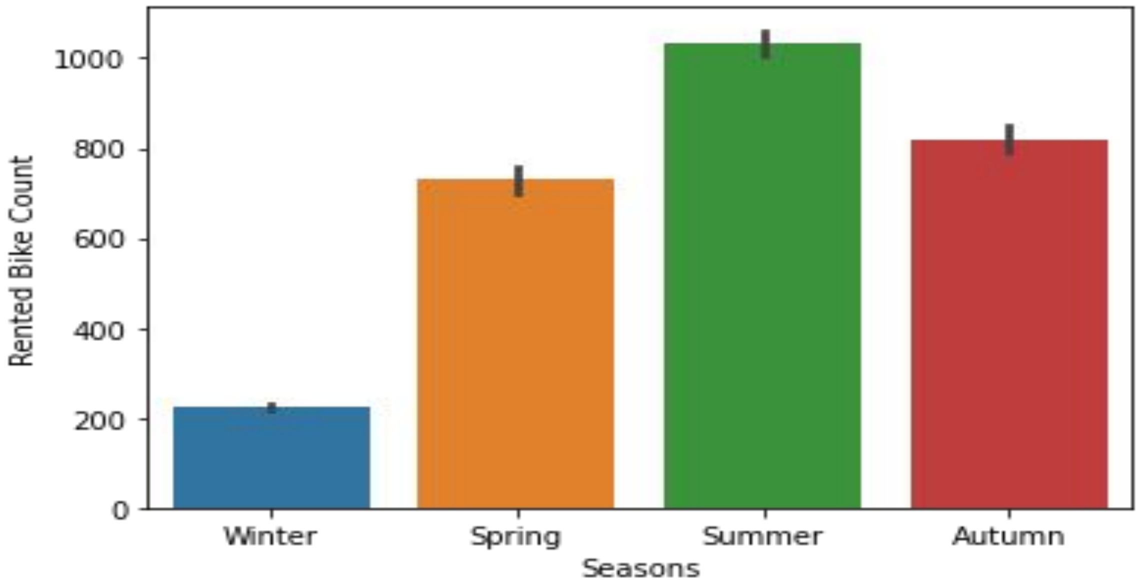
DATA SAMPLE



	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

➔ EXPLORATORY DATA ANALYSIS (EDA)

➔ SEASON-WISE OVERVIEW

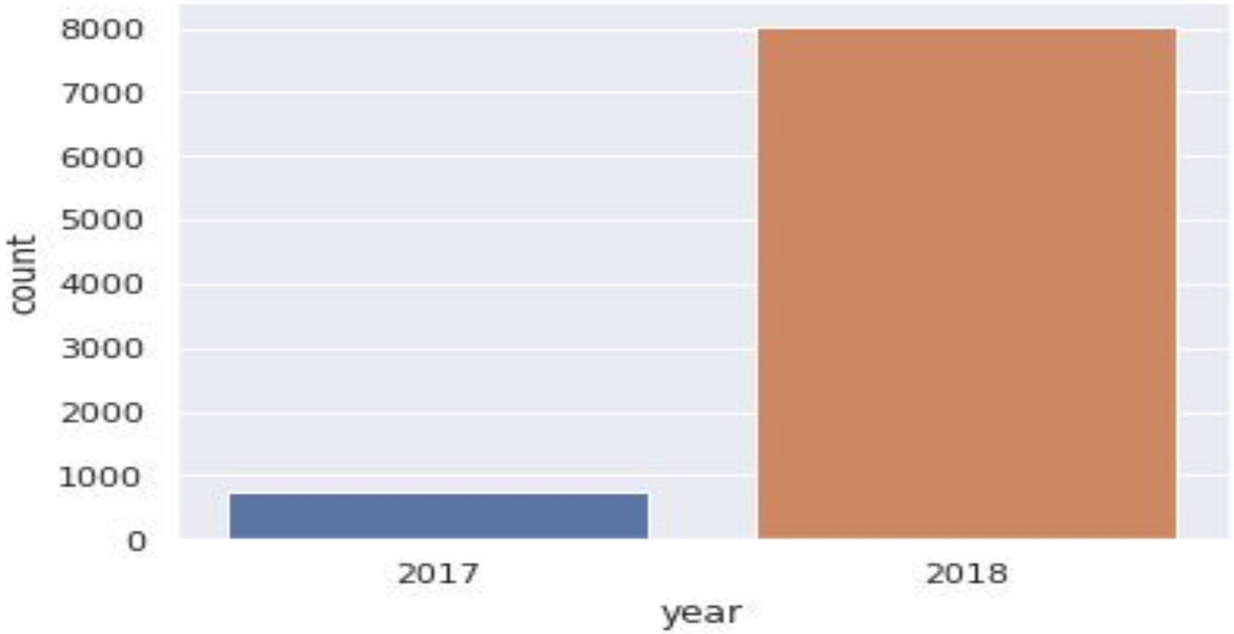


❖ The maximum numbers of bikes are rented during summer season.

❖ Least bikes are rented during winter season.

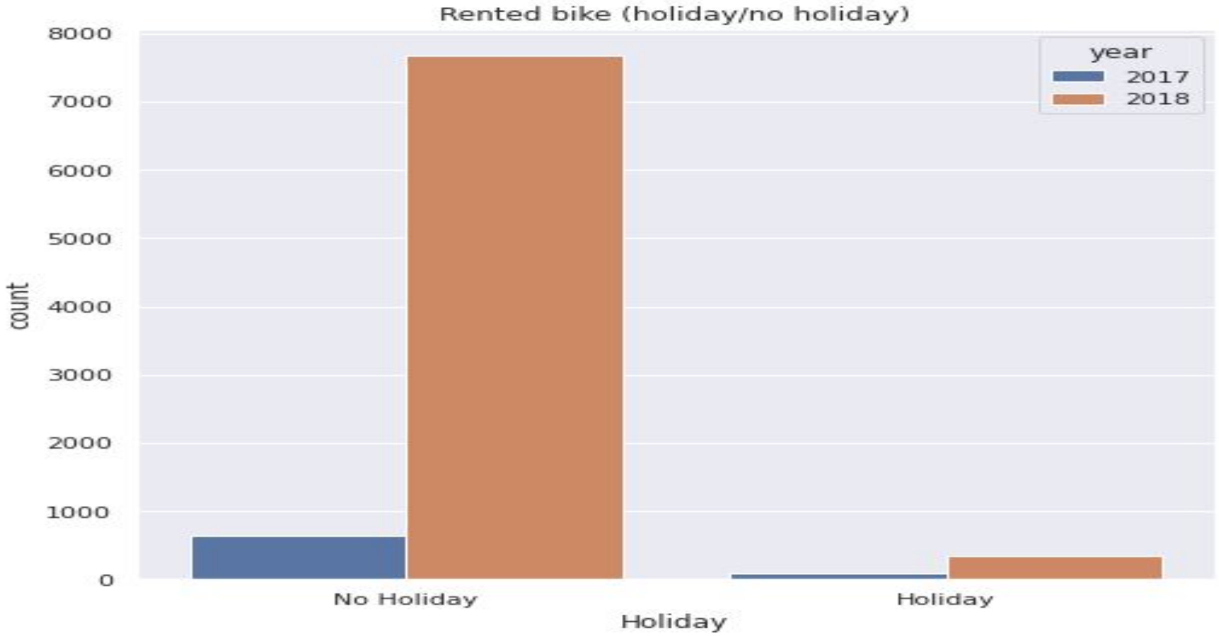
❖ Bikes rented during spring and autumn season are almost same.

➔ YEAR WISE OVERVIEW



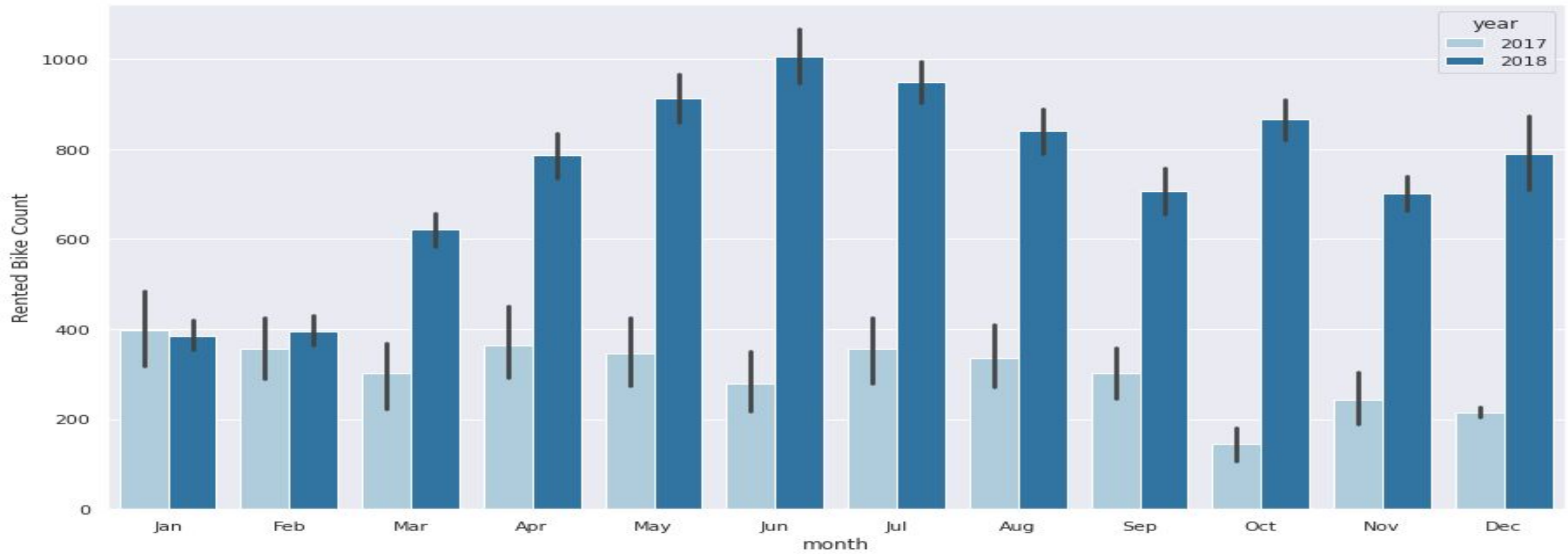
- ❖ Maximum numbers of bike are rented in the year 2018.
- ❖ The number of rented bike count in 2018 is 8x times higher than the previous year.

→ HOLIDAY / NO HOLIDAY



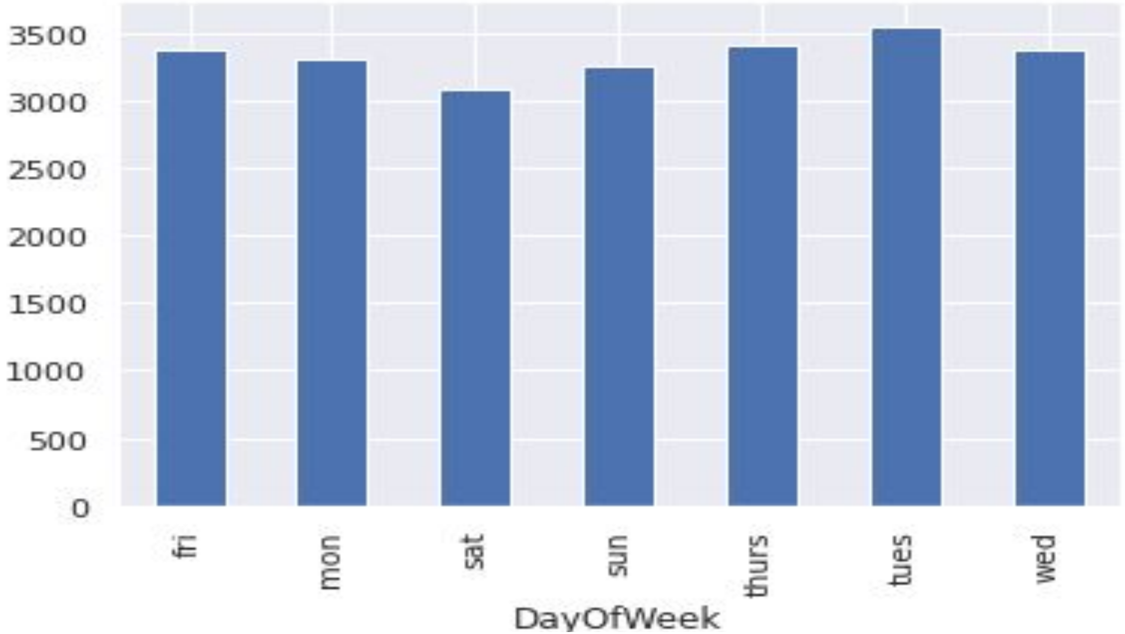
Majority of the bikes are rented during no holiday time in both the years.

➔ MONTH WISE OVERVIEW



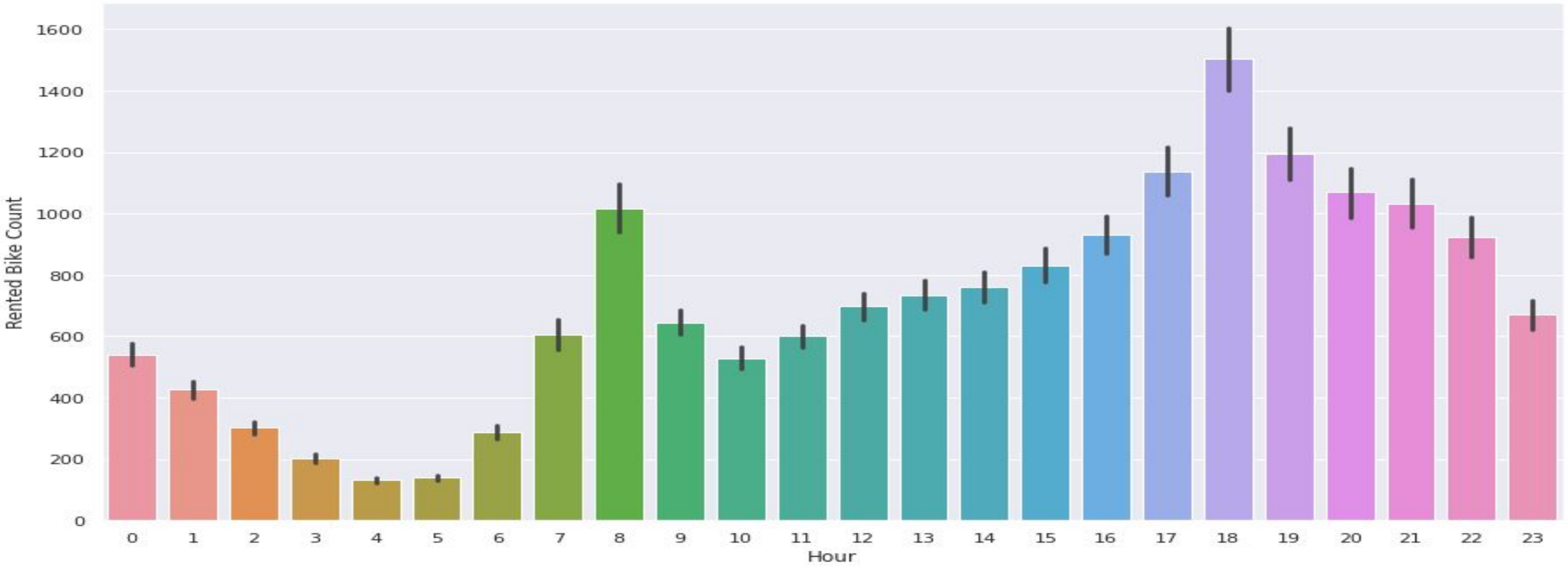
- ❖ Number of rented bikes starts increasing as we approach the summer seasons.
- ❖ Months of April, May and June falls under summer season and the number of rented bikes are highest during this time.
- ❖ Number of rented bikes starts decreasing during July, August and september which falls under rainy season.
- ❖ Winter season is the one were least numbers of bikes were rented

➔ DAY OF THE WEEK



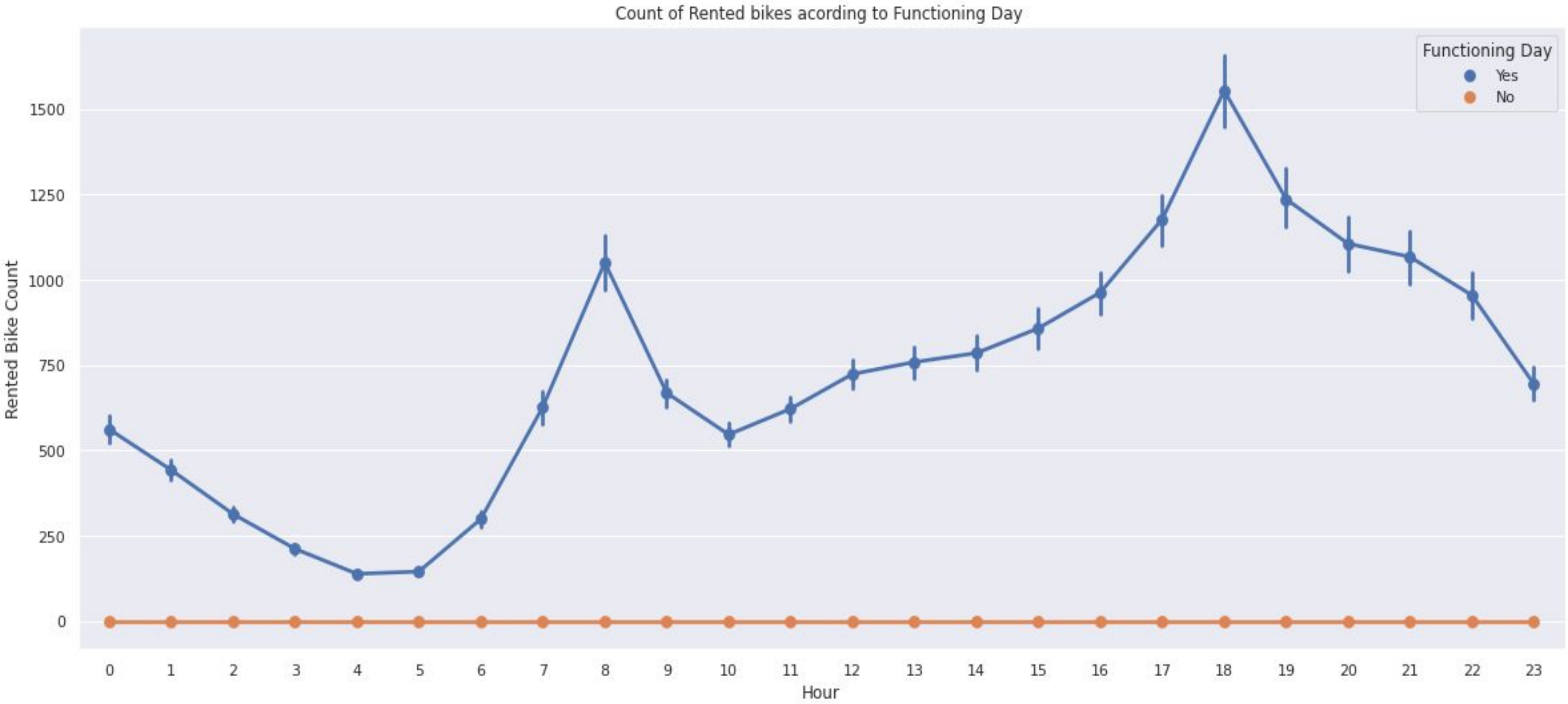
- ❖ 1.Least numbers of bike are rented on weekend day i.e. Saturday and Sunday
- ❖ 2.Week days are the working day hence number of rented bike count is maximum

➔ HOUR



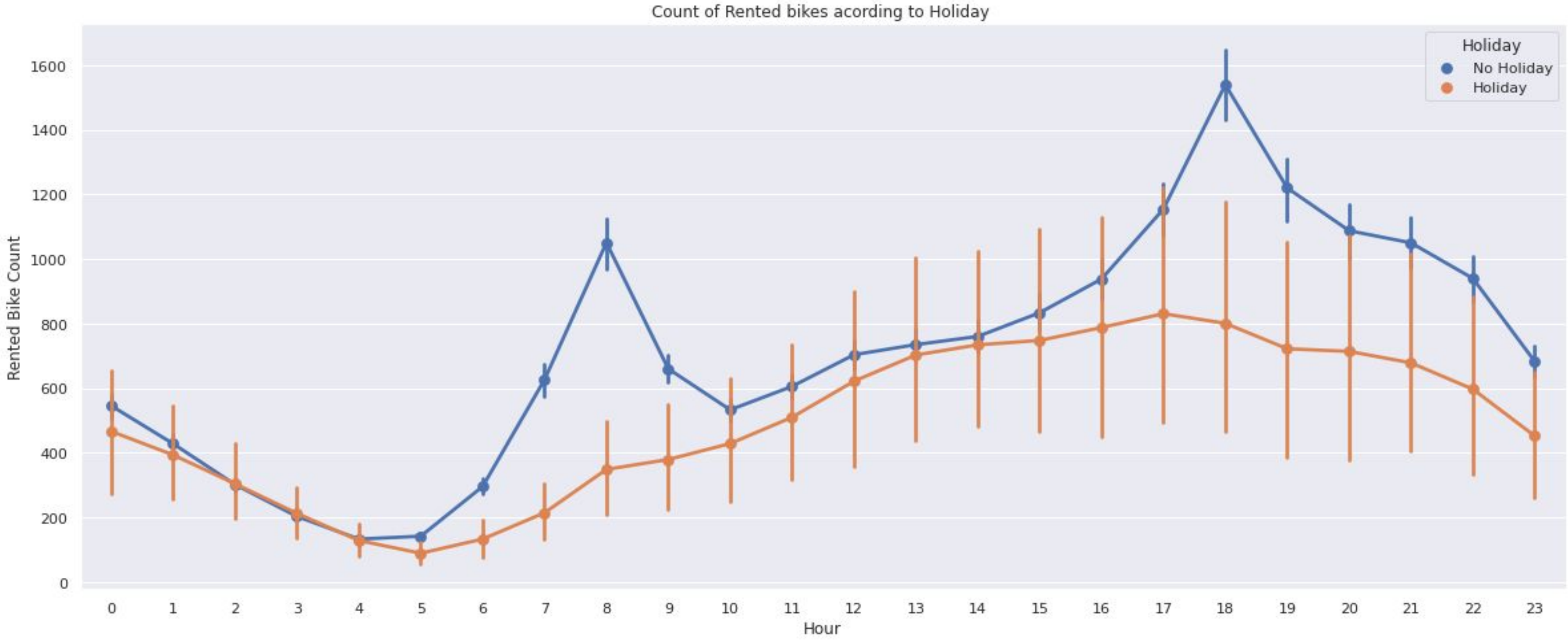
- ❖ Rented bike demand is at peak during 7 to 9am in the morning and 5 to 7pm in the evening.
- ❖ Least number of bikes are rented during early morning hours i.e. between 3 to 6 am.

➔ FUNCTIONING DAY



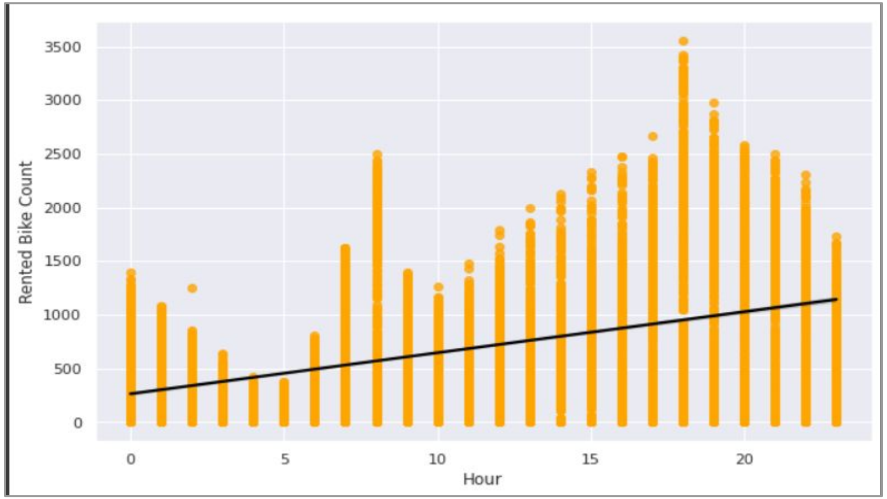
Functioning day's are the days when people usually goes to work.Hence count of rented bike is at its peak between 7 to 9 am in the morning and 5 to 7 in the evening which is generally considered as the office timing.

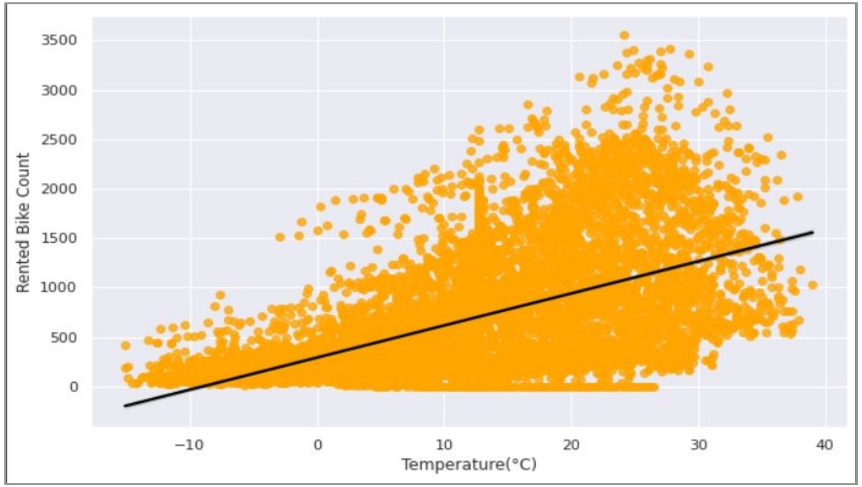
➔ HOLIDAY

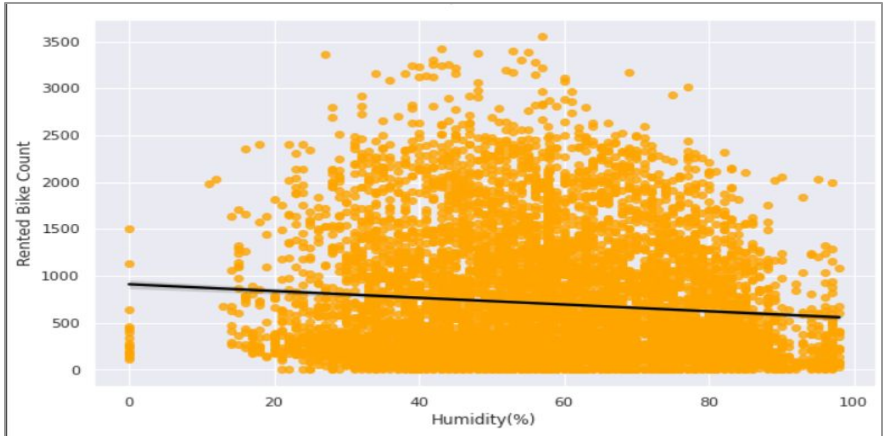


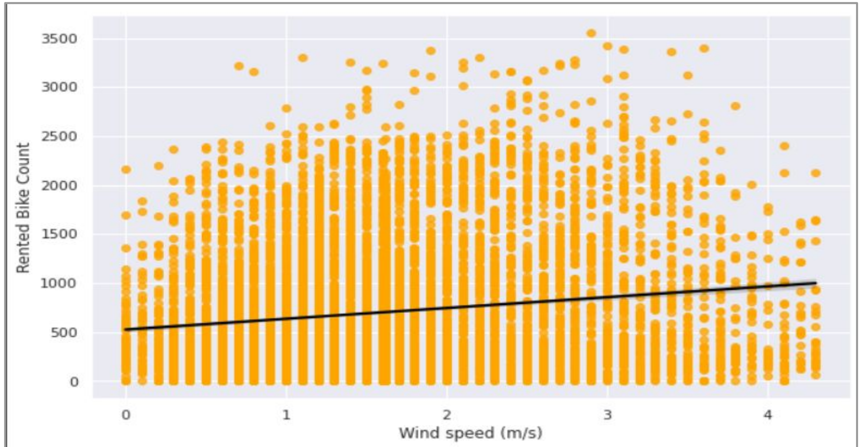
During the holidays the demand is very low especially in the morning and only starts increasing in the evening.

REGRESSION PLOT

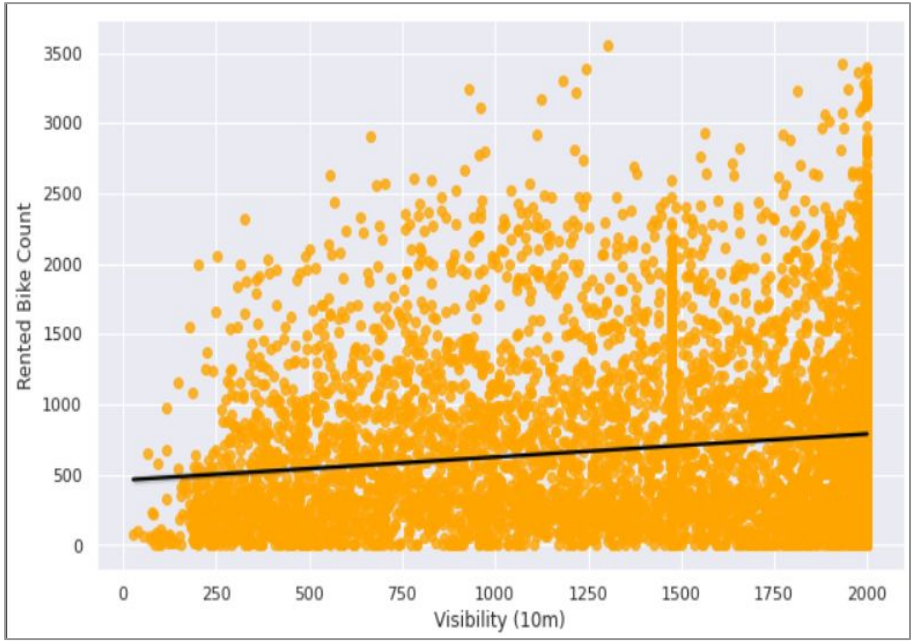


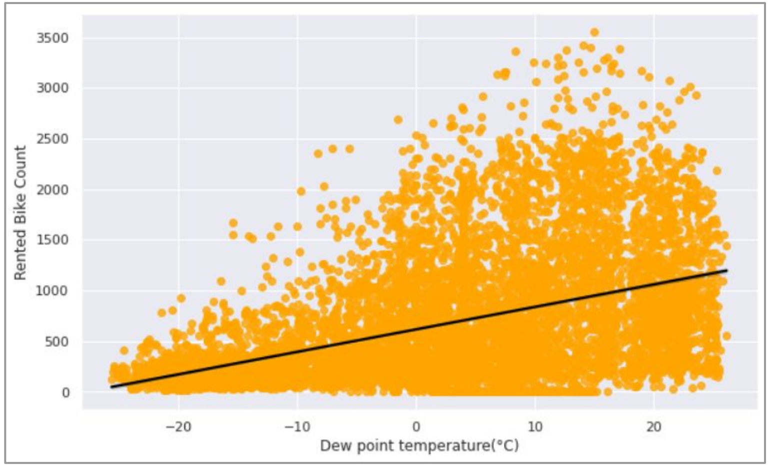


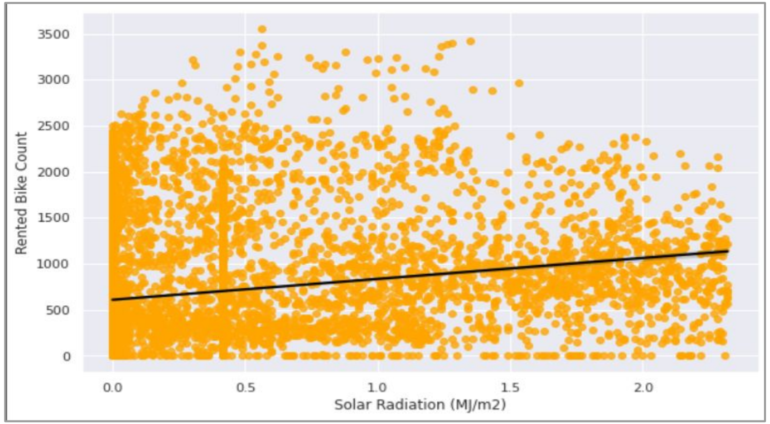




➔ REGRESSION PLOT(Conti…)



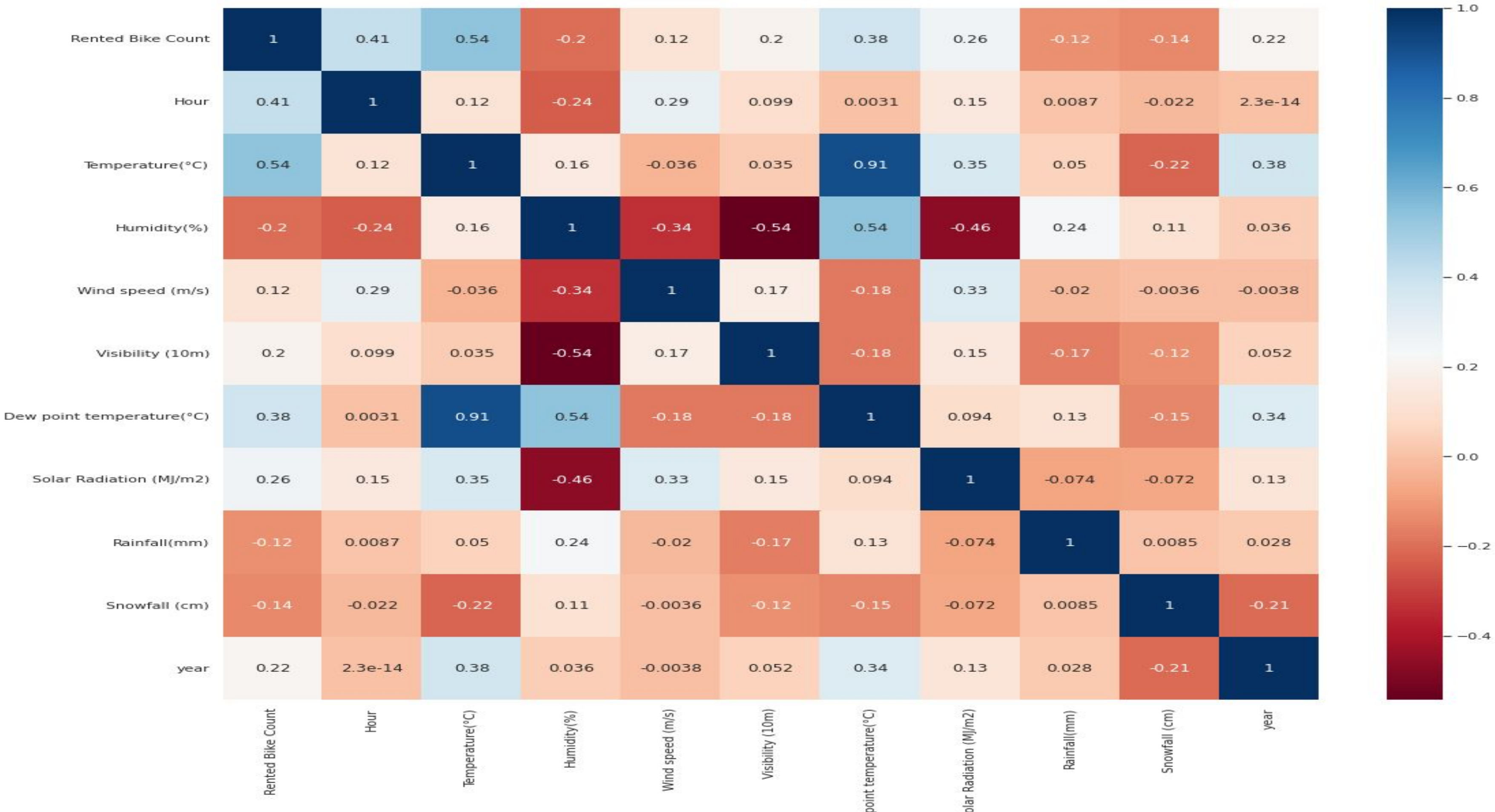




➔ REGRESSION PLOT(OBSERVATIONS)

- ❖ Temperature, Dew point Temperature, Hour,visibility, Windspeed and Solar radiation shows positive correlation with the target variable which means target variable increases with the increase of this features.
- ❖ Humidity shows negative correlation i.e. our target variable decreases when Humidity increases.

➔ HEAT MAP



	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	point temperature(°C)	olar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	year
Rented Bike Count	1	0.41	0.54	-0.2	0.12	0.2	0.38	0.26	-0.12	-0.14	0.22
Hour	0.41	1	0.12	-0.24	0.29	0.099	0.0031	0.15	0.0087	-0.022	2.3e-14
Temperature(°C)	0.54	0.12	1	0.16	-0.036	0.035	0.91	0.35	0.05	-0.22	0.38
Humidity(%)	-0.2	-0.24	0.16	1	-0.34	-0.54	0.54	-0.46	0.24	0.11	0.036
Wind speed (m/s)	0.12	0.29	-0.036	-0.34	1	0.17	-0.18	0.33	-0.02	-0.0036	-0.0038
Visibility (10m)	0.2	0.099	0.035	-0.54	0.17	1	-0.18	0.15	-0.17	-0.12	0.052
Dew point temperature(°C)	0.38	0.0031	0.91	0.54	-0.18	-0.18	1	0.094	0.13	-0.15	0.34
Solar Radiation (MJ/m2)	0.26	0.15	0.35	-0.46	0.33	0.15	0.094	1	-0.074	-0.072	0.13
Rainfall(mm)	-0.12	0.0087	0.05	0.24	-0.02	-0.17	0.13	-0.074	1	0.0085	0.028
Snowfall (cm)	-0.14	-0.022	-0.22	0.11	-0.0036	-0.12	-0.15	-0.072	0.0085	1	-0.21
year	0.22	2.3e-14	0.38	0.036	-0.0038	0.052	0.34	0.13	0.028	-0.21	1

→ DATA PREPROCESSING

➔ DATA PREPROCESSING

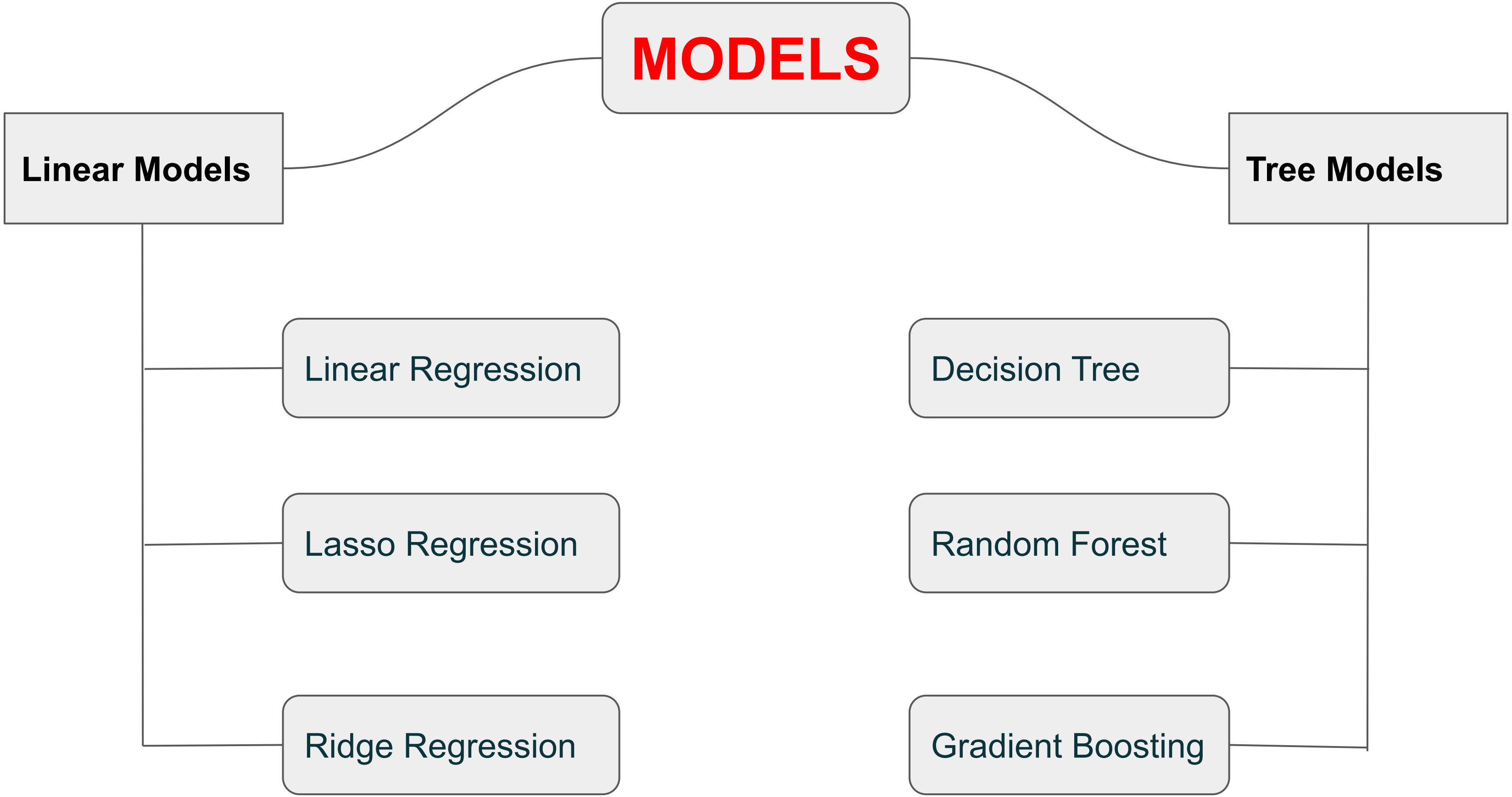
- We extracted Month, Day of the week and year from the Date column by first converting the Date column into a date format.
- Categorical columns present in the dataset after feature extraction are Season, Holiday, Functioning day, Month, Dayofweek and year.
- We created the dummy variables for all the categorical columns
- We removed the outliers from the numerical features and filled the NaN values with their respective mean

➔ DATA PREPROCESSING (Conti…)

- **Dropping highly correlated columns**

❖ Dew point temperature has high positive correlation (0.91) with Temperature so we dropped it.

❖ Rainfall and snowfall has NaN values after removing the outliers so we dropped them too.



MODELS

Linear Models

Linear Regression

Lasso Regression

Ridge Regression

Tree Models

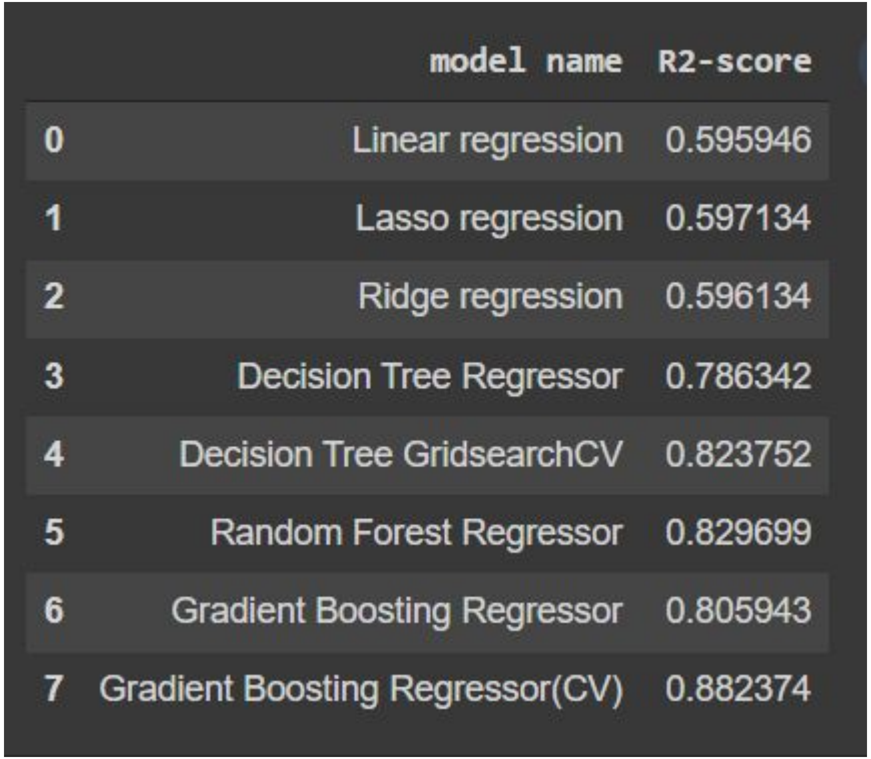
Decision Tree

Random Forest

Gradient Boosting

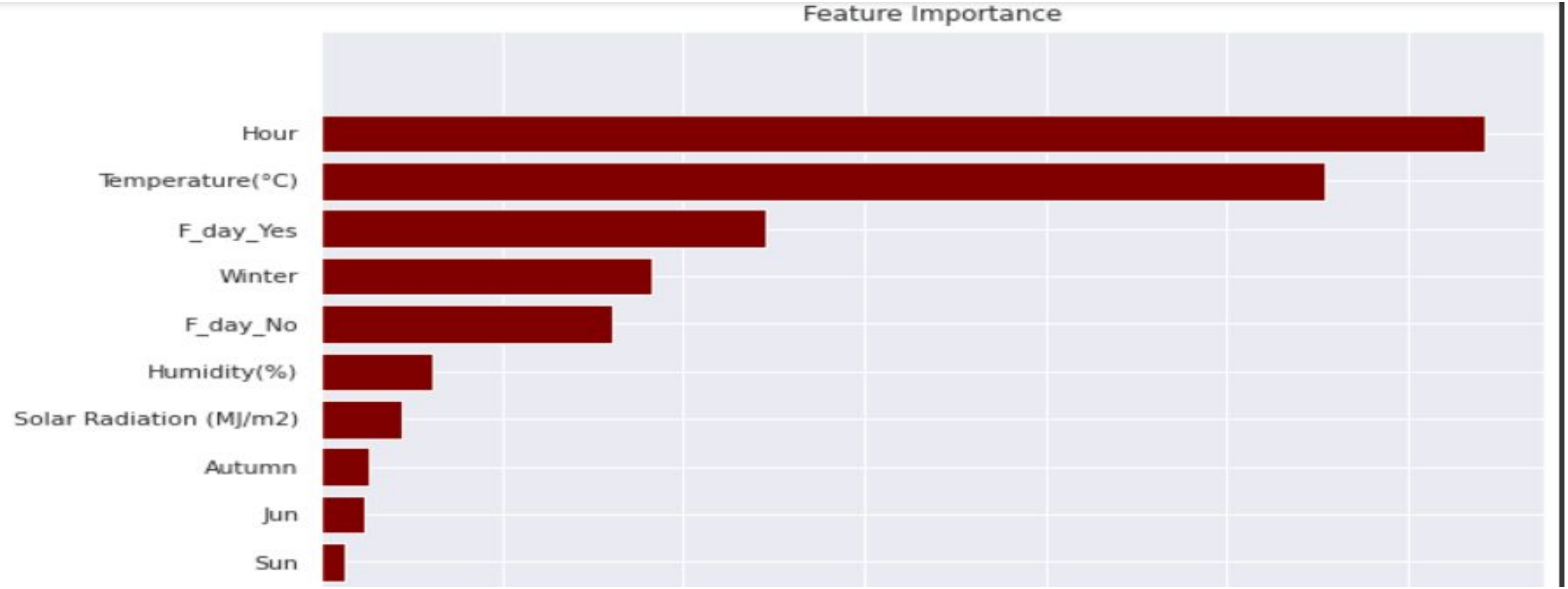
→ EVALUATION

- The model that performed best on the given dataset with the hyper-parameter tuning and cross validation is Gradient Boost with an r2 score of 0.88
- The model that performed best without hyper-parameter tuning is Random Forest with an r2 score of 0.829 (≈ 0.83)
- Linear regression performed the least. It has an r2 score of 0.595



	model name	R2-score
0	Linear regression	0.595946
1	Lasso regression	0.597134
2	Ridge regression	0.596134
3	Decision Tree Regressor	0.786342
4	Decision Tree GridsearchCV	0.823752
5	Random Forest Regressor	0.829699
6	Gradient Boosting Regressor	0.805943
7	Gradient Boosting Regressor(CV)	0.882374

➔ FEATURE IMPORTANCE(GRADIENT BOOST)



- Among the top 10 features, ‘Hour’ is the most important one followed by ‘Temperature’.
- F_day_Yes, winter ,F_day_No , humidity are some of the other important features

➔ CONCLUSION

- Demand was high during springs and summer and autumn and very low during winters.
- Number of bikes rented during the year 2018 is highest.
- Demand of rented bikes is high during no holiday and functioning day.
- Peak hours for rented bike demand is between 7 to 9 am in the morning and 5 to 7 pm in the evening which suggest that the bikes are rented mostly by the office going people
- Demand for bikes got higher when the temperature and hour values were more.
- Demand was high for low values of Humidity and solar radiation.
- Random forest Regressor and Gradient Boosting gridsearchcv gives the highest R2 score of 82% and 88% respectively.
- We can deploy Random forest Regressor and Gradient Boosting model.
- The most important features which had a major impact on the Gradient Boosting model predictions were; hour, temperature, Humidity, solar-radiation, and Winter.

THANK YOU!