Capstone Project-2

BIKE SHARING DEMAND PREDICTION- SEOUL

(supervised Machine Learning Regression)

Batch- Cohort Seattle

Presented By:

Vaitul Sidhdhapara

Drashti Shah

CONTENTS:

INTRODUCTION:

- 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

PROBLEM STATEMENT:

- We are tasked with predicting the number of bikes rented each hour so as to make an approximate estimation of the number of bikes to be made available to the public given a particular hour of the day.

DATA SUMMARY:

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:**

- 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

DATA SUMMARY:

- 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 SUMMARY:

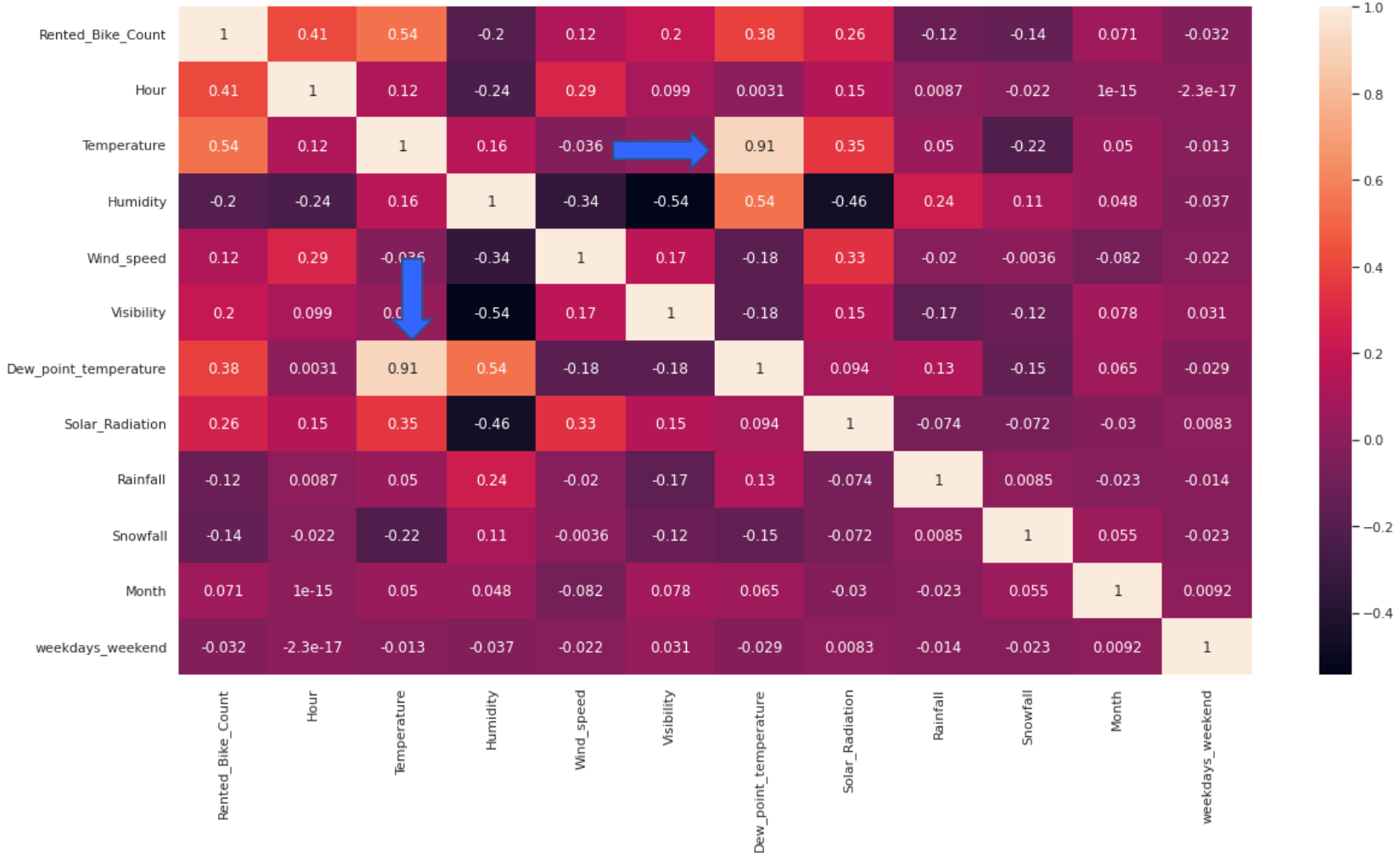
- This dataset contains 8760 rows and 14 columns
- **Numerical variables :** temperature, humidity, wind, visibility, dew point temp, solar radiation, rainfall, snowfall
- **Categorical variables :** seasons, holidays and functioning day
- Rented bike count column (Dependent Variable) : which we need to predict for new observations

AI

EXPLORATORY DATA ANALYSIS:

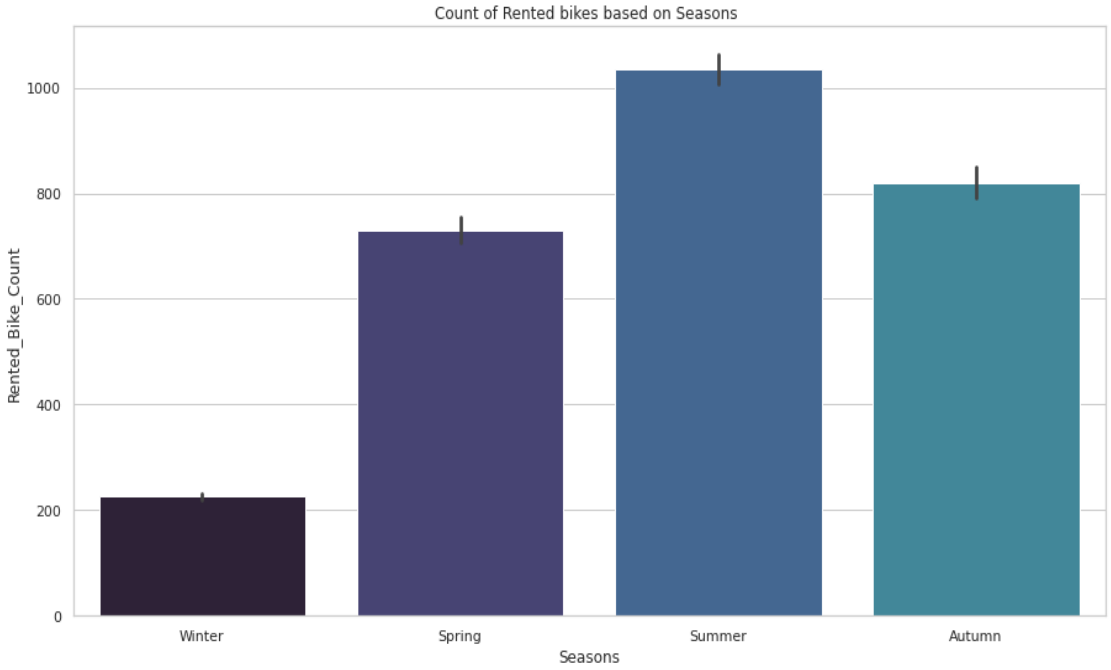
- Exploratory Data Analysis refers to the critical process of performing initial investigations on data to discover patterns, spot anomalies, test hypotheses, and check assumptions with the help of summary statistics and graphical representations.
- EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

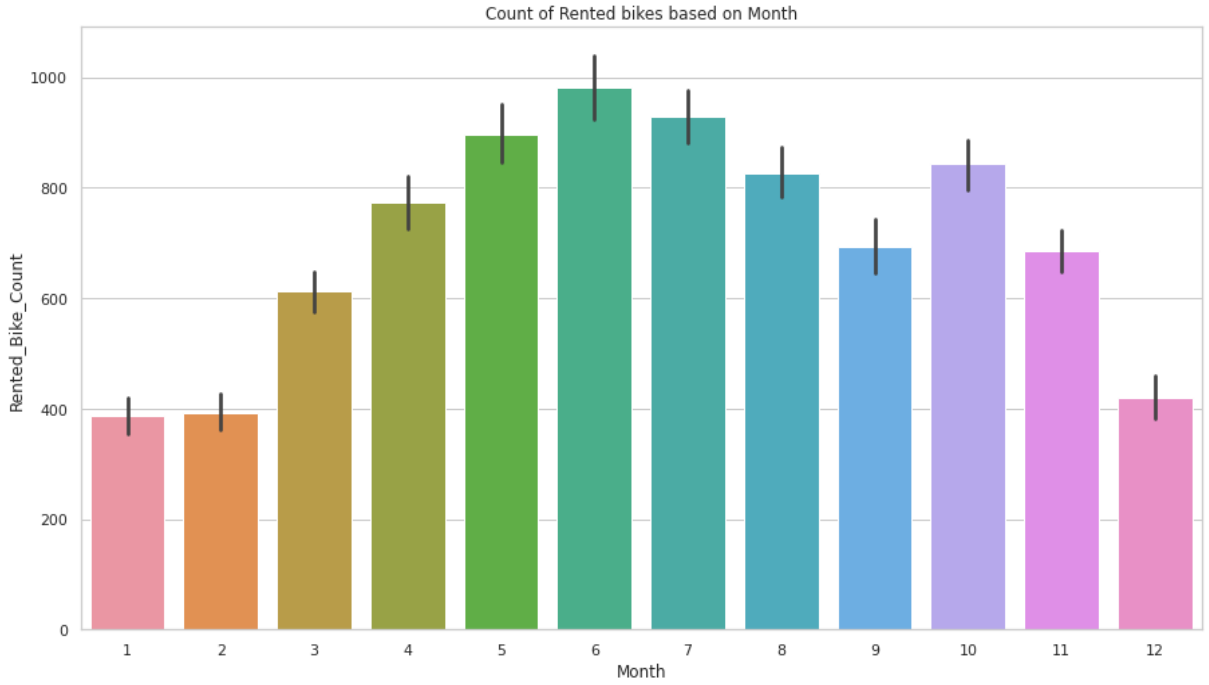
EXPLORATORY DATA ANALYSIS:



- Dew Point Temperature and Temperature are highly correlated.
- We don’t want Multicollinearity in our dataset it affects the performance metrics so we are dropping Dew Point Temperature.

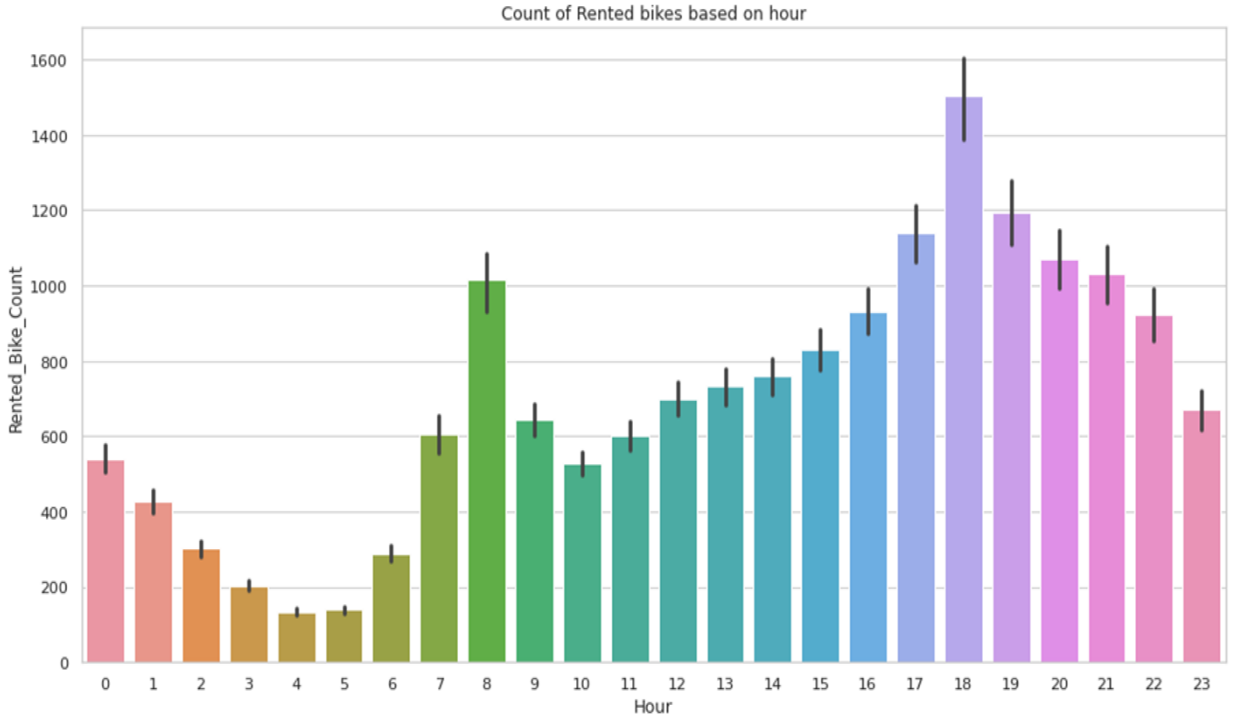
DATA VISUALIZATION:

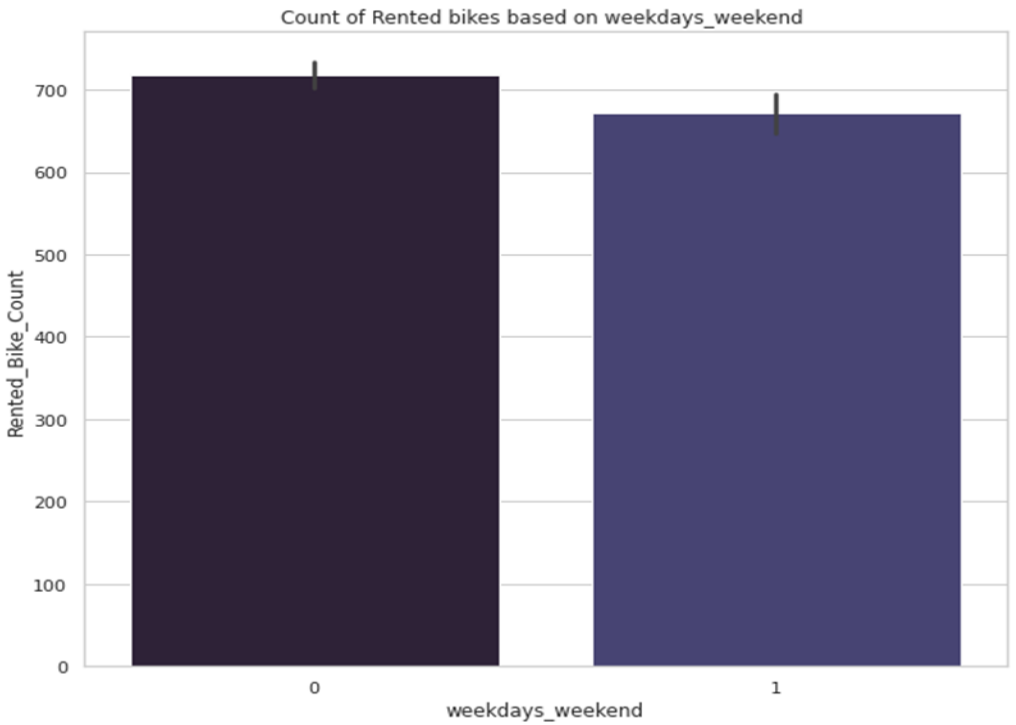




- Maximum number of bike rentals in Summer season, while in winter season the bike rentend bike count is low.
- The demand of the rented bike is high between 5th to 10th months as compare to other months.

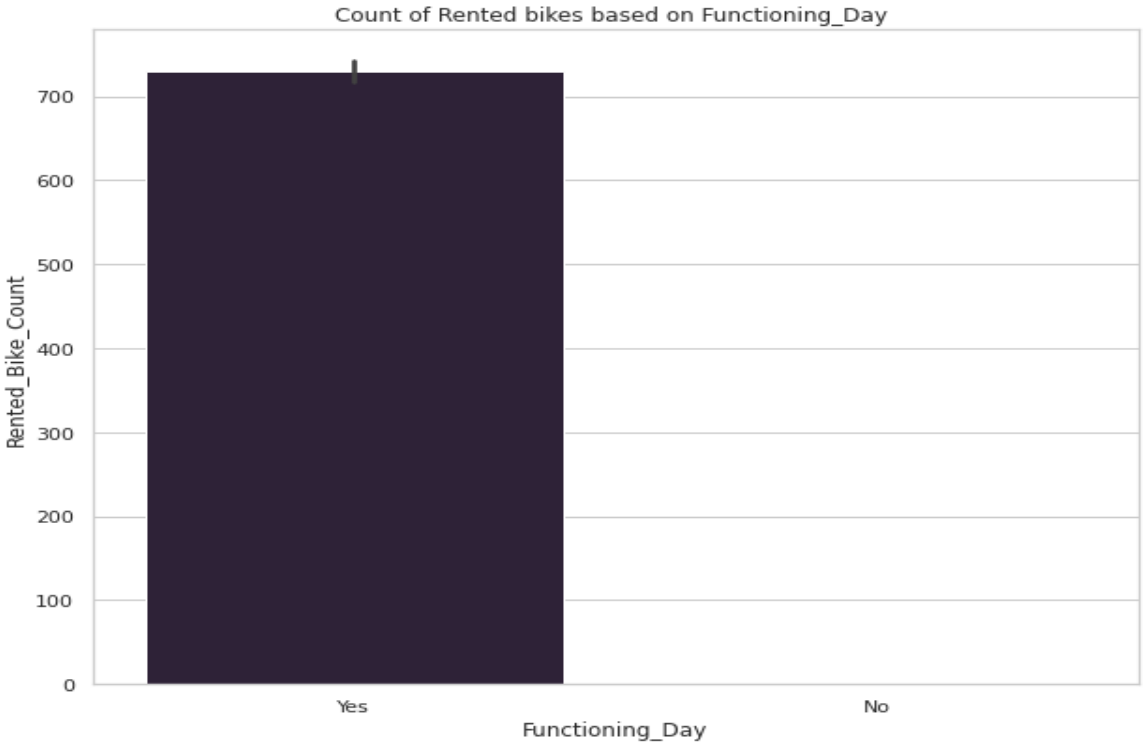
DATA VISUALIZATION:

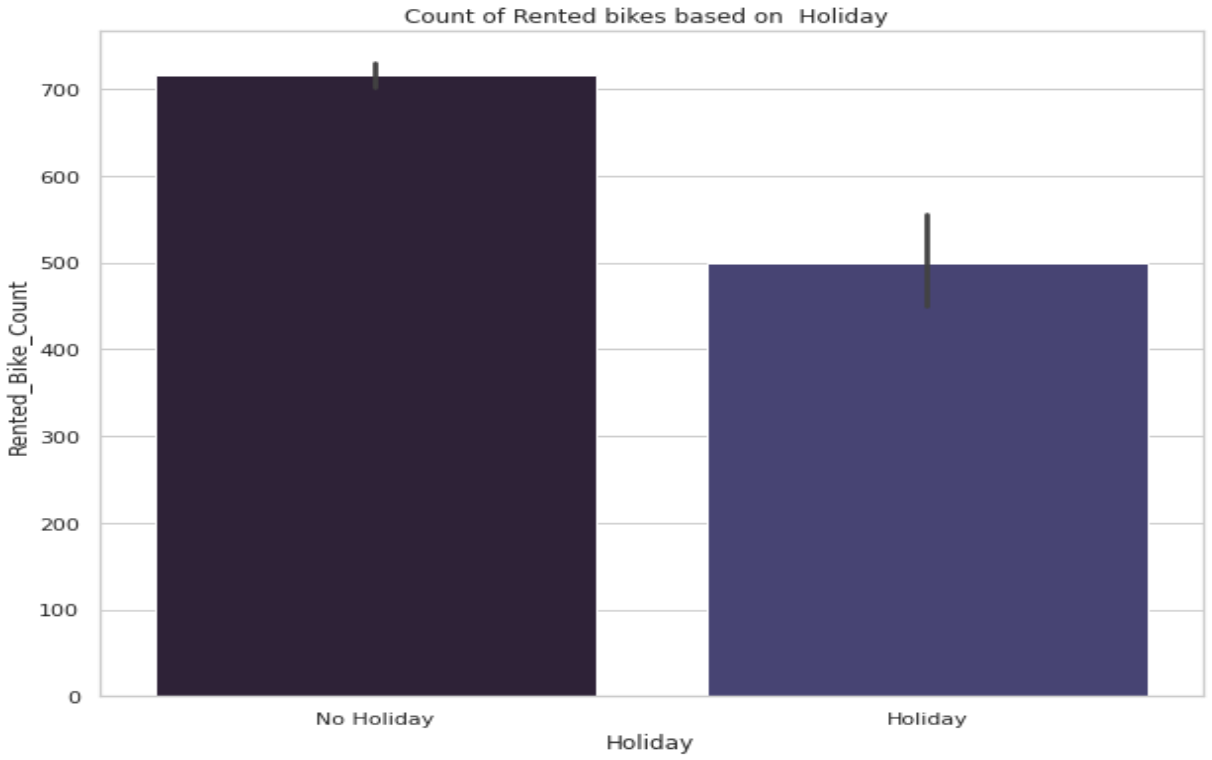




- People generally use rented bikes during their working hour from 7 AM to 9 AM and 5 PM to 8 PM.
- In the week days the demand of the bike is higher because of the office as compare to the weekend.

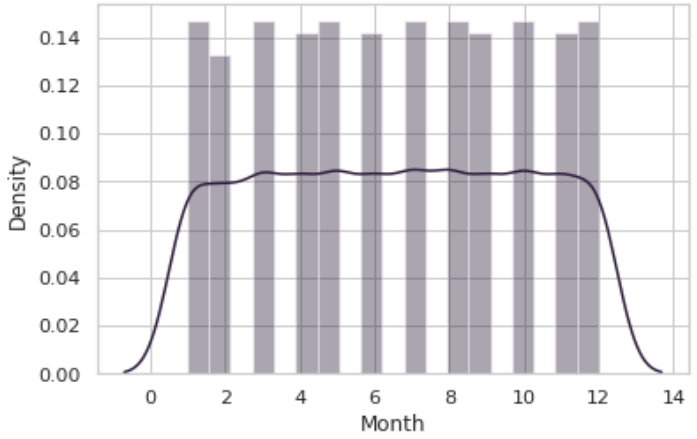
DATA VISUALIZATION:

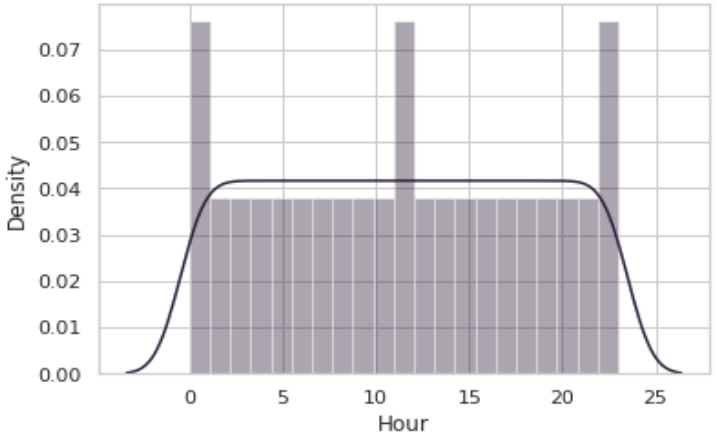


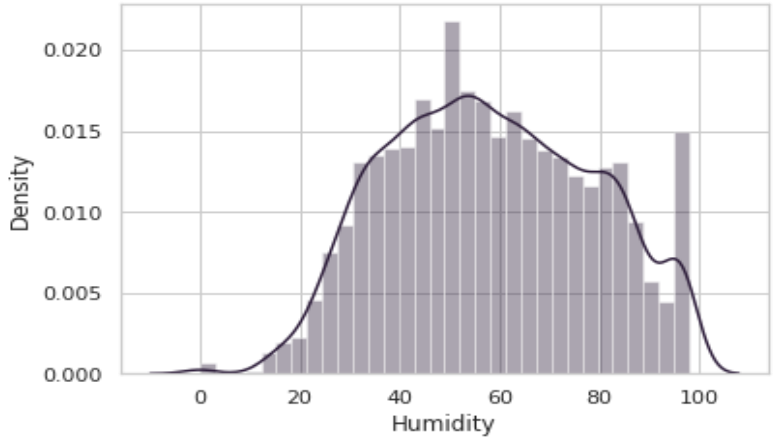


- People use rented bike only in functioning day
- Use of rented bike is more on 'No holiday' means on working days as compare to 'Holiday'.

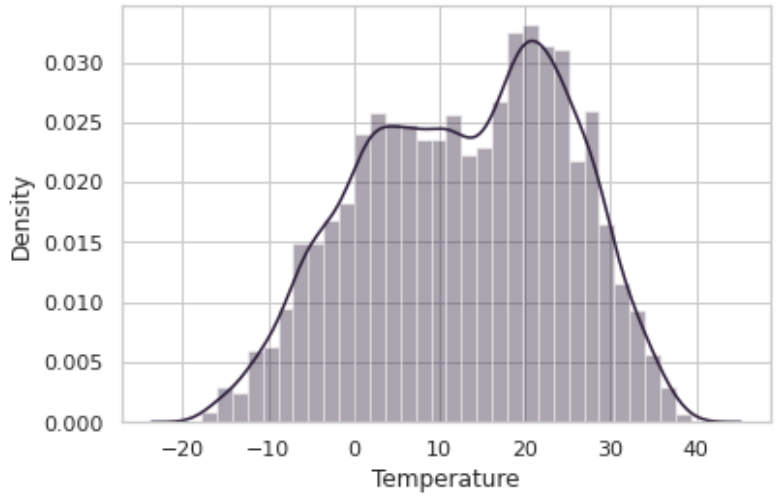
VISUALISING DATA DISTRIBUTIONS:



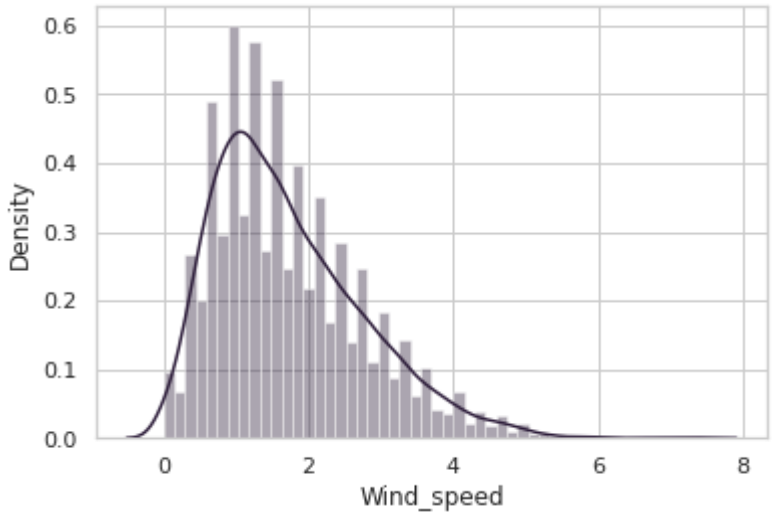


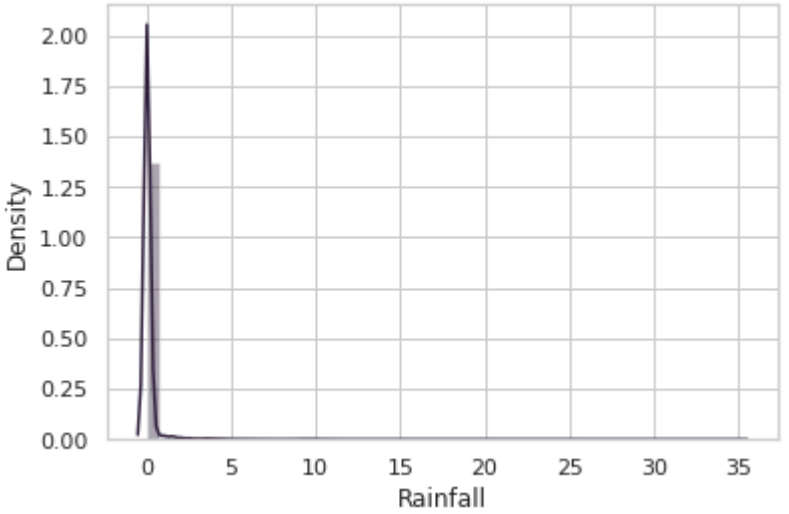


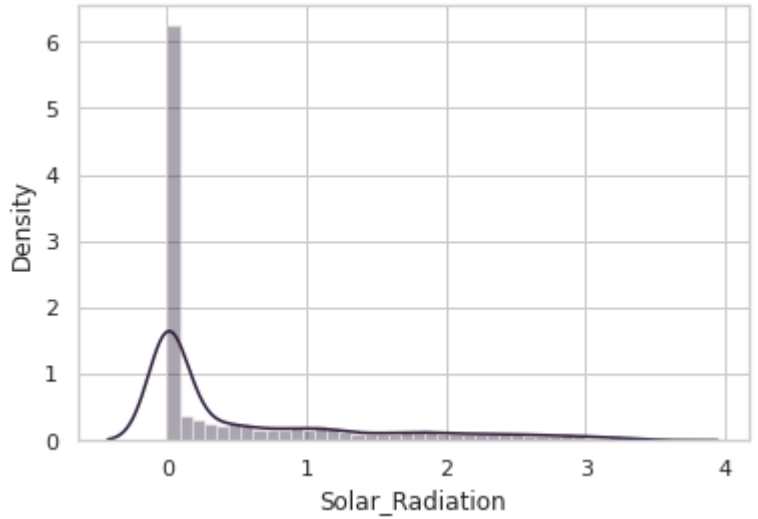
- From the graphs we can say that “Temperature” and “Humidity” are normally distributed.
- “Hour” and “Month” columns follow a uniform distribution.



VISUALISING DATA DISTRIBUTIONS:

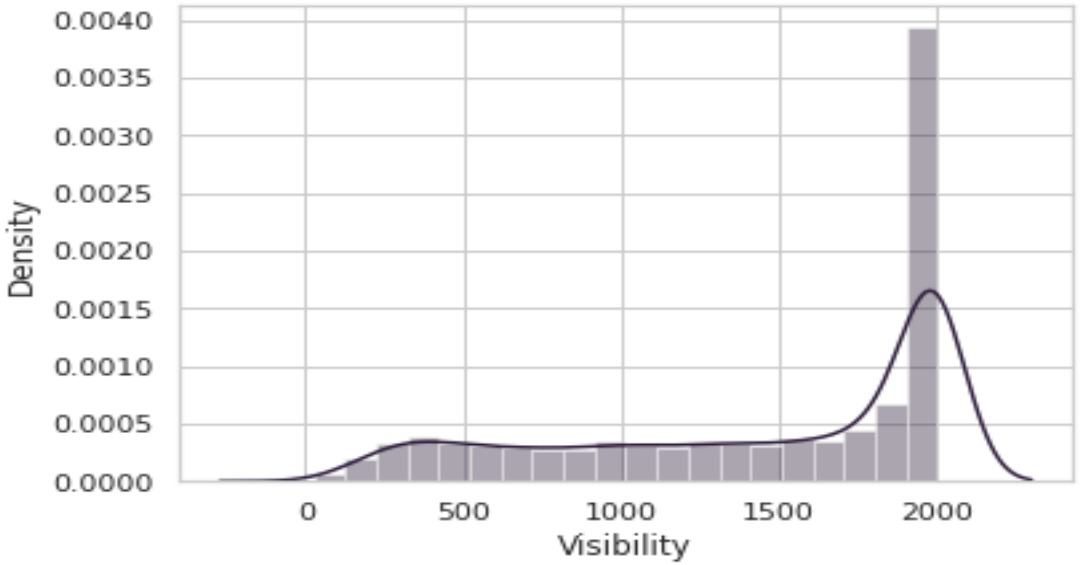


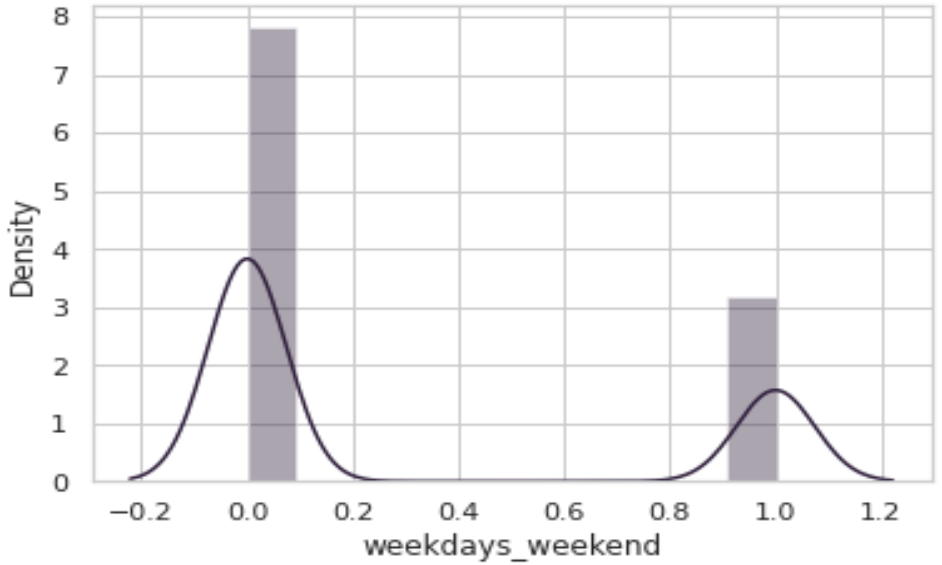




- From the graphs we can say that “Wind Speed”, “Solar Radiation”, “Rainfall” and “Snowfall” are having positively skewed distribution.

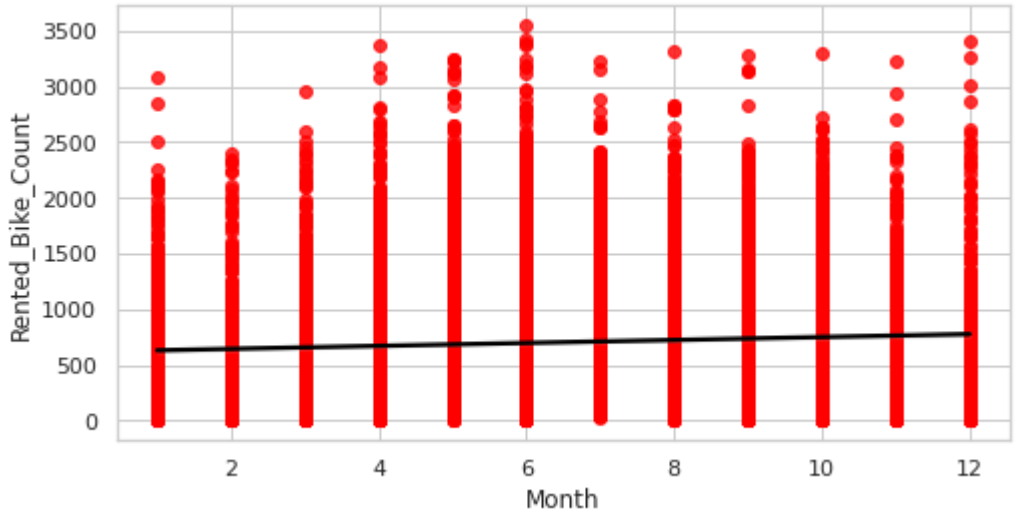
VISUALISING DATA DISTRIBUTIONS:

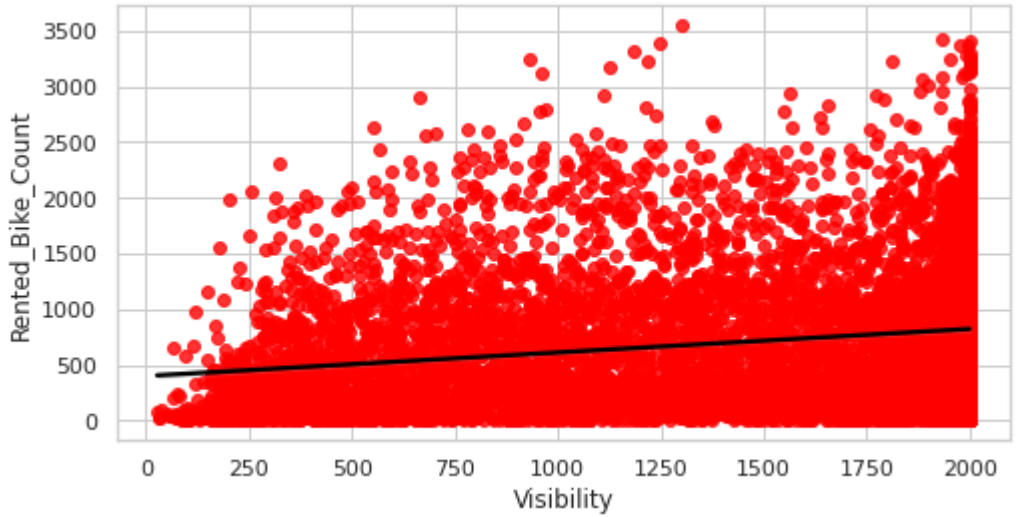


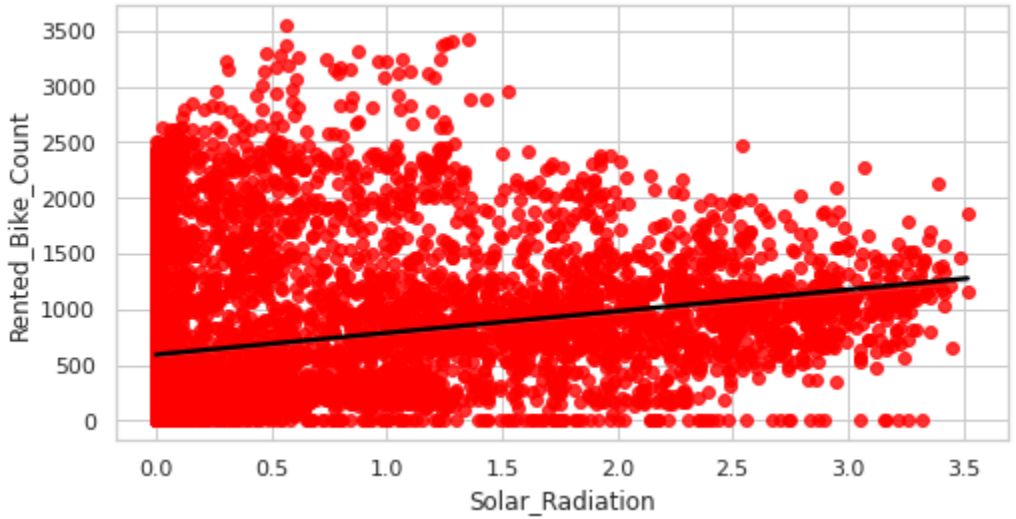


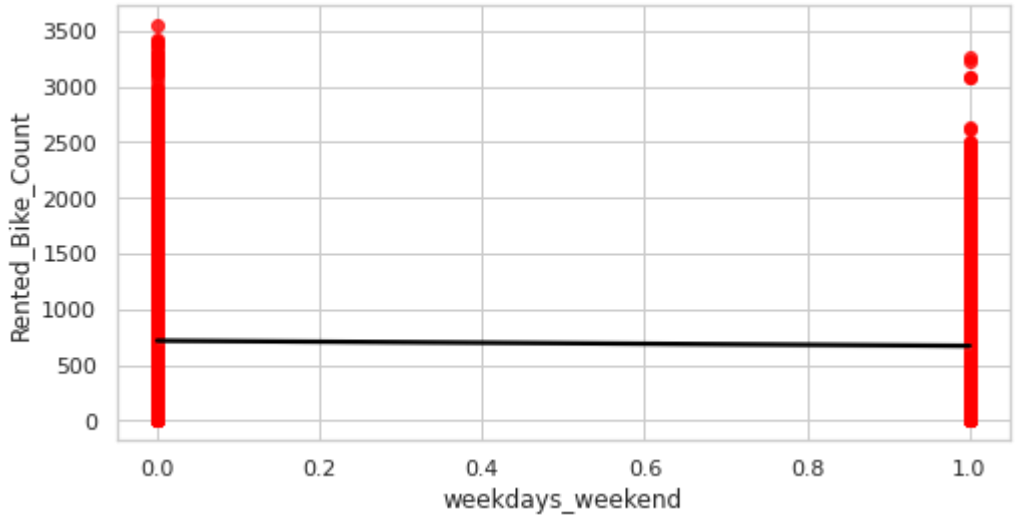
- Here graph shows that “Visibility” column is negatively skewed.

REGRSSION PLOTS (FEATURES v/s TARGET VARIABLES) :





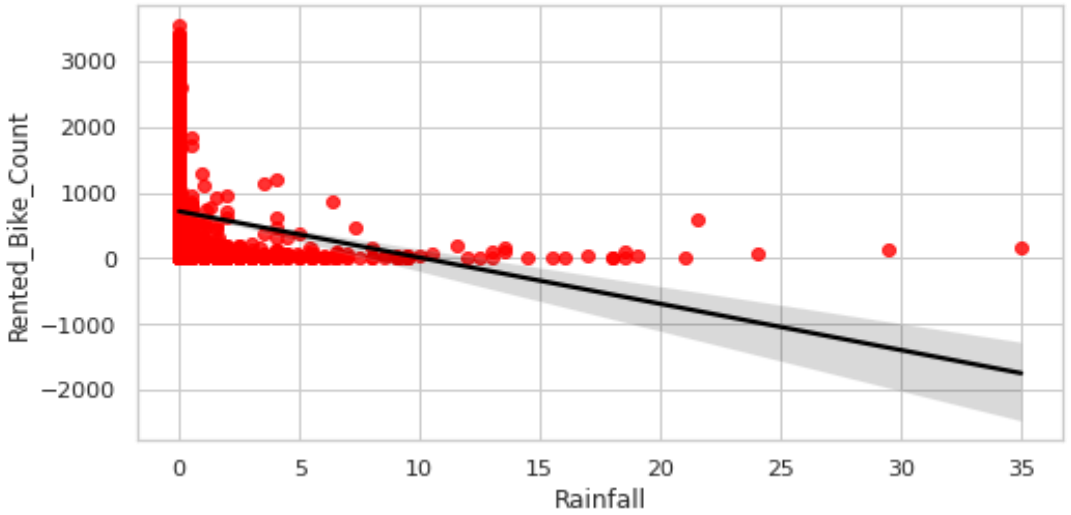


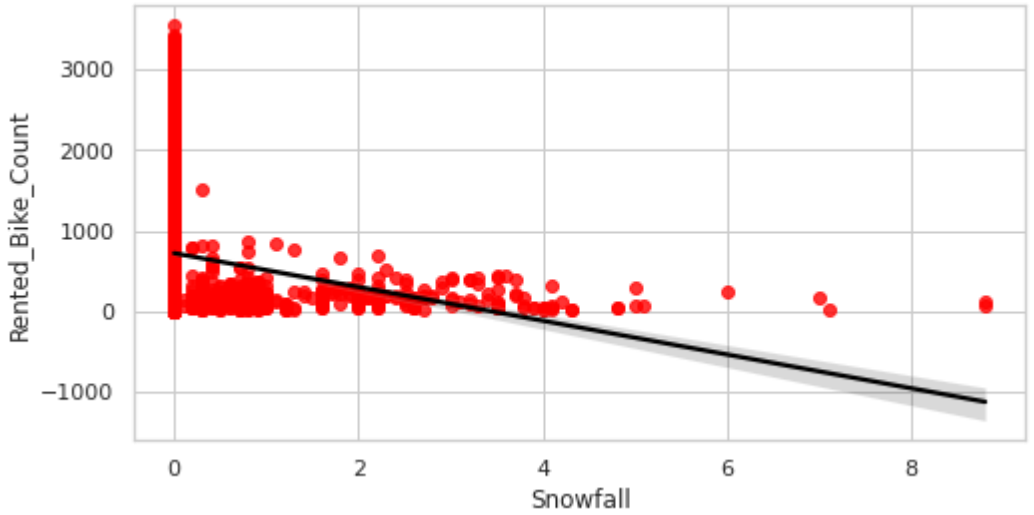


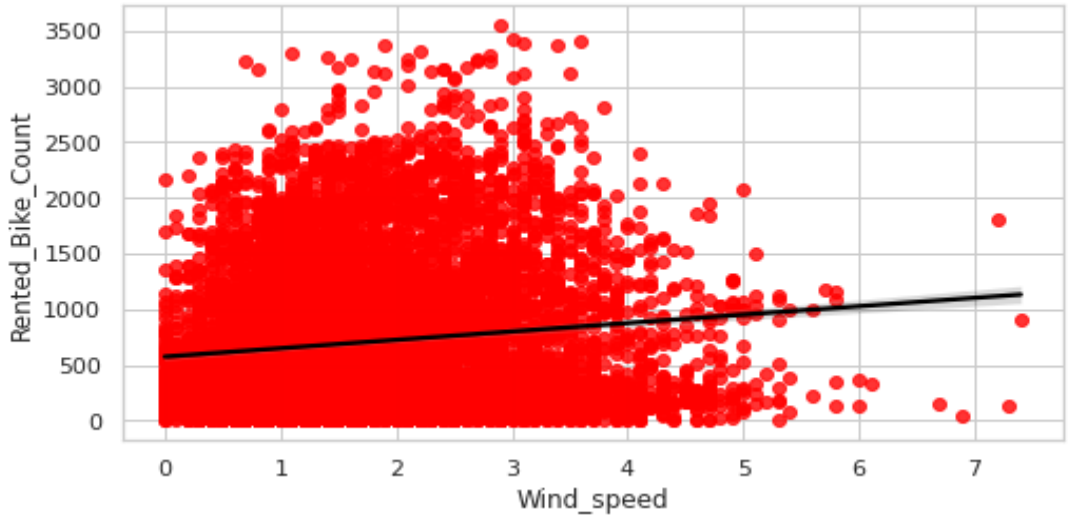
REGRSSION PLOTS (FEATURES v/s TARGET VARIABLES) ANALYSIS:

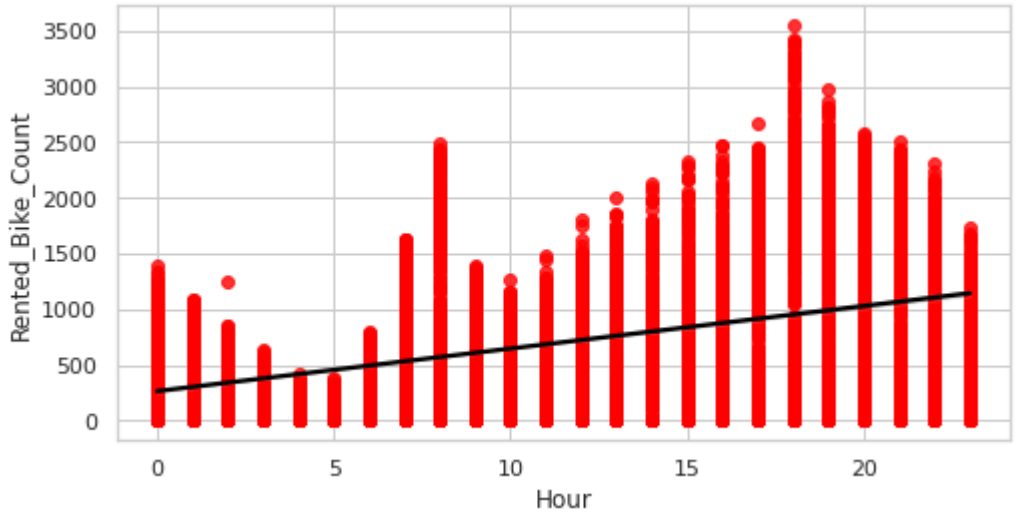
- The columns 'Hour', 'Temperature', 'Windspeed', 'Visibility' and 'Solar_Radiation' are positively related to the dependent variable. Which means that the rented bike count increases with increase of these features.
- Whereas, the columns 'Rainfall', 'Snowfall', 'Humidity' are those features which are negatively related with the dependent variable, which implies that the rented bike count decreases when these features increases.

REGRSSION PLOTS (FEATURES v/s TARGET VARIABLES) :





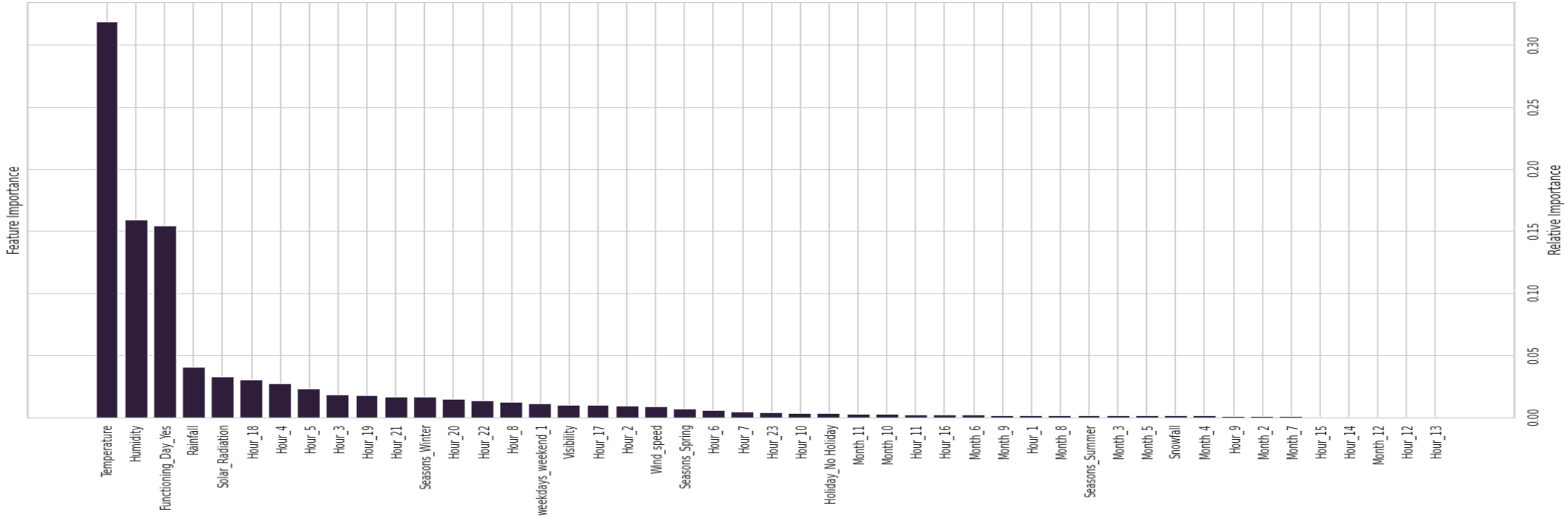




MODEL BUILDING:

- Linear Regression
- Ridge Regression
- Lasso Regression
- Elastic Net Regression
- Decision Tree Regression
- Random Forest Regression
- Gradient Boosting Regression

FEATURE IMPORTANCE:



- Best model is Gradient Boosting Regression. For this model very important feature is ‘temperature’ which we can see in the ‘Feature Importance Graph’.

EVALUATION MATRICS:

		Model	MAE	MSE	RMSE	R2	Adj_R2
Training set	**0**	Linear regression	4.658	37.606	6.132	0.756	0.75
	1	Ridge regression	4.659	37.608	6.133	0.756	0.75
	2	Lasso regression	4.658	37.608	6.133	0.756	0.75
	3	Elasticnet regression	4.660	37.611	6.133	0.756	0.75
	4	Decision tree regression	4.043	31.921	5.650	0.793	0.79
	5	Random forest regression	3.552	23.851	4.884	0.845	0.84
	6	Gradient Boosting Regression	1.331	3.688	1.921	0.976	0.98
Test set	**0**	Linear regression	4.658	36.645	6.053	0.768	0.76
	1	Ridge regression	4.661	36.664	6.055	0.768	0.76
	2	Lasso regression	4.659	36.647	6.054	0.768	0.76
	3	Elasticnet regression	4.662	36.674	6.056	0.768	0.76
	4	Decision tree regression	4.731	44.565	6.676	0.718	0.71
	5	Random forest regression	3.877	28.994	5.385	0.816	0.81
	6	Gradient Boosting Regression	2.546	14.252	3.775	0.910	0.91

1. Out of all above models Gradient Boosting Regressor gives the highest R2 score of 98% for Train Set and 91% for Test set.
2. No overfitting is seen.
3. We can deploy Gradient Boosting Regressor model.

CHALLANGES:

- Pre-processing the data was one of the challenges we faced which includes removing highly correlated variables from the data so as to not hinder the performance of our regression model.
- Selecting the appropriate models to maximize the accuracy of our predictions was one of the challenges faced.
- A huge amount of data needed to be dealt while doing the project which is quite an important task and also even small inferences need to be kept in mind.
- Feature engineering, Feature selection.
- Model Training and performance improvement.

CONCLUSION:

- Out of all above models Gradient Boosting Regression gives the highest R^2 score of 98% for Train Set and 91% for Test set.
- No overfitting is seen.
- We used 7 Regression Models to predict the bike rental count at any hour of the day - Linear Regression, Ridge, Lasso, Random Forest, Elastic Net, Gradient Boost and Decision Tree Regression

Thank you