Capstone Project

SEOUL BIKE SHARING DEMAND PREDICTION

By

SARANYA N

EMAIL ADDRESS: snk44111@gmail.com

Key steps

- PROBLEM STATEMENT
- DATA SUMMARY
- FEATURE ANALYSIS
- EXPLORATORY DATA ANALYSIS
- DATA PREPROCESSING
- IMPLEMENTING ALGORITHMS
- CHALLENGES
- 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.

Data summary

	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

- **This Dataset contains 8760 lines and 14 columns.**
- **Three categorical features ‘Seasons’, ‘Holiday’, & ‘Functioning Day’.**
- **One Datetime features ‘Date’.**
- **We have some numerical type variables such as temperature, humidity, wind, visibility, dew point temp, solar radiation, rainfall, snowfall which tells the environment conditions at that particular hour of the day.**

Feature Summary

Date : Year-Month-Day
Rented Bike Count - Count of bikes rented at each hour
Hour - Hour of the day
Temperature - Temperature in Celsius
Humidity - %
Wind Speed - 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 and functional hours)

EXPLORATORY DATA ANALYSIS:

EDA is used for analyzing what the data can tell us before the modeling or by applying any set of instructions /code.

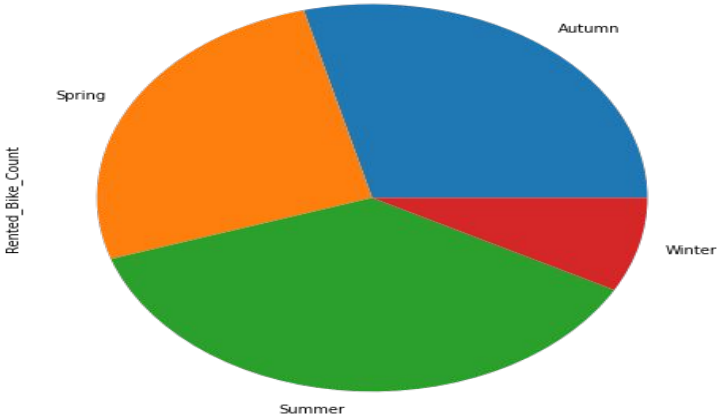
After loading and reading the dataset in notebook, we performed EDA. Comparing target variable which is bike rentals counts with other independent variables. This process helped us figuring out various aspects and relationships among the target and the independent variables and also we observed the distribution of variables. It gave us a better idea that how feature behaves with the target variable.

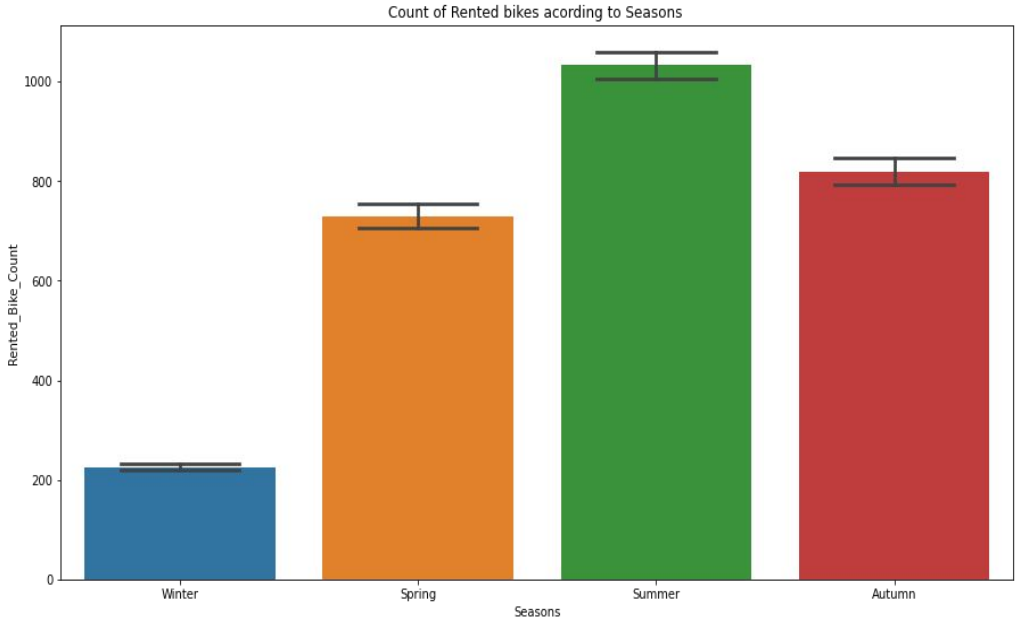
Insights From Our Dataset

- There are No Missing Values .
- There are No Duplicate values present.
- There are No null values.
- The dataset shows hourly rental data for one year (1 December 2017 to 31 November(2018)(365 days).we consider this as a single year data
- So we convert the "date" column into 3 different column i.e "year","month","day".
- We change the name of some features for our convenience , they are as below

'Rented_Bike_Count', 'Hour', 'Temperature', 'Humidity', 'Wind_speed', 'Visibility', 'Dew_point_temperature', 'Solar_Radiation', 'Rainfall', 'Snowfall', 'Seasons', 'Holiday', 'Functioning_Day', 'month', 'weekdays weekend'

Analysis of season variable



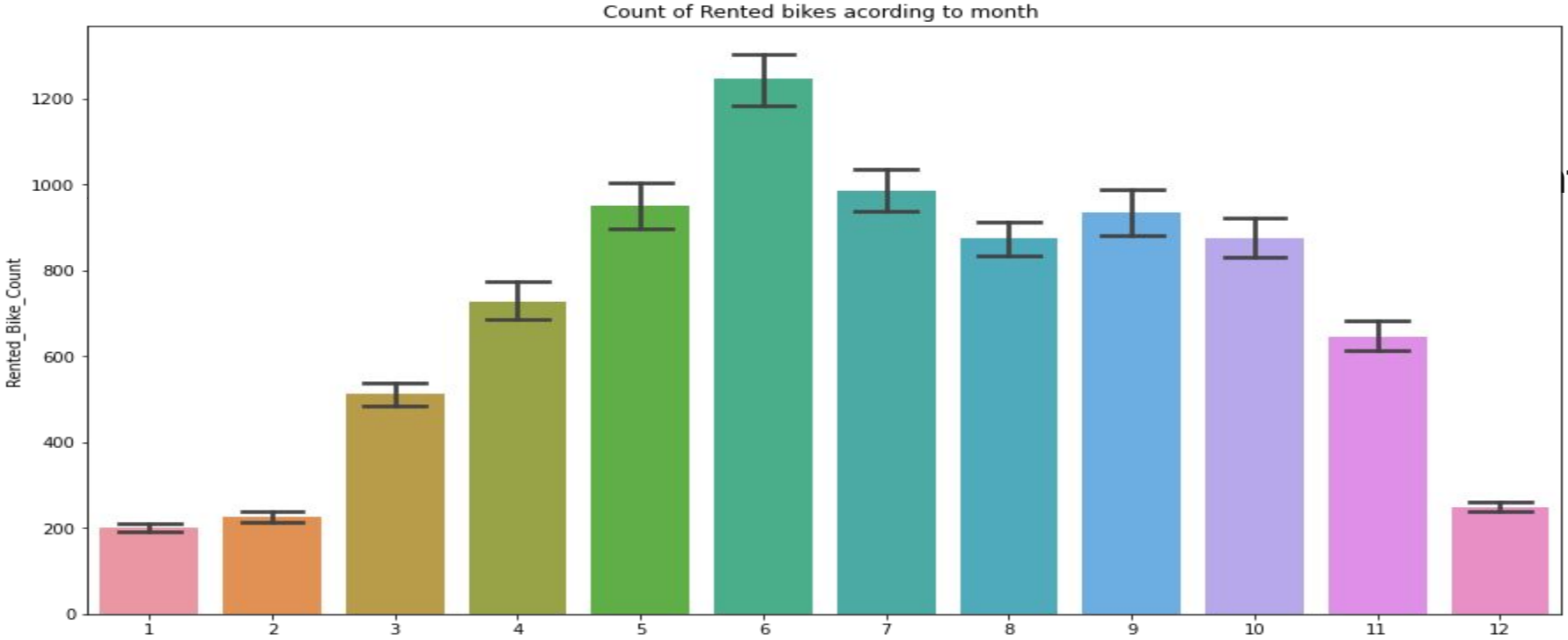


From the above pie and barplot we can understand that usage of rented bike in those 4 seasons

In summer season use of rented bike is very high

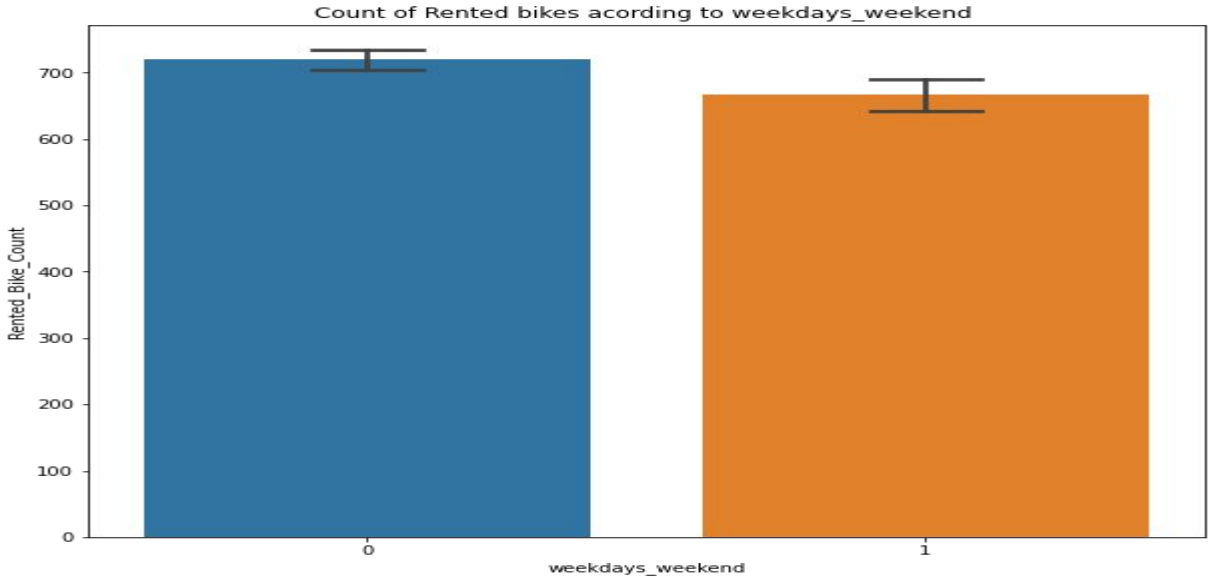
In winter season the use of rented bike is very low

Analysis of month variable



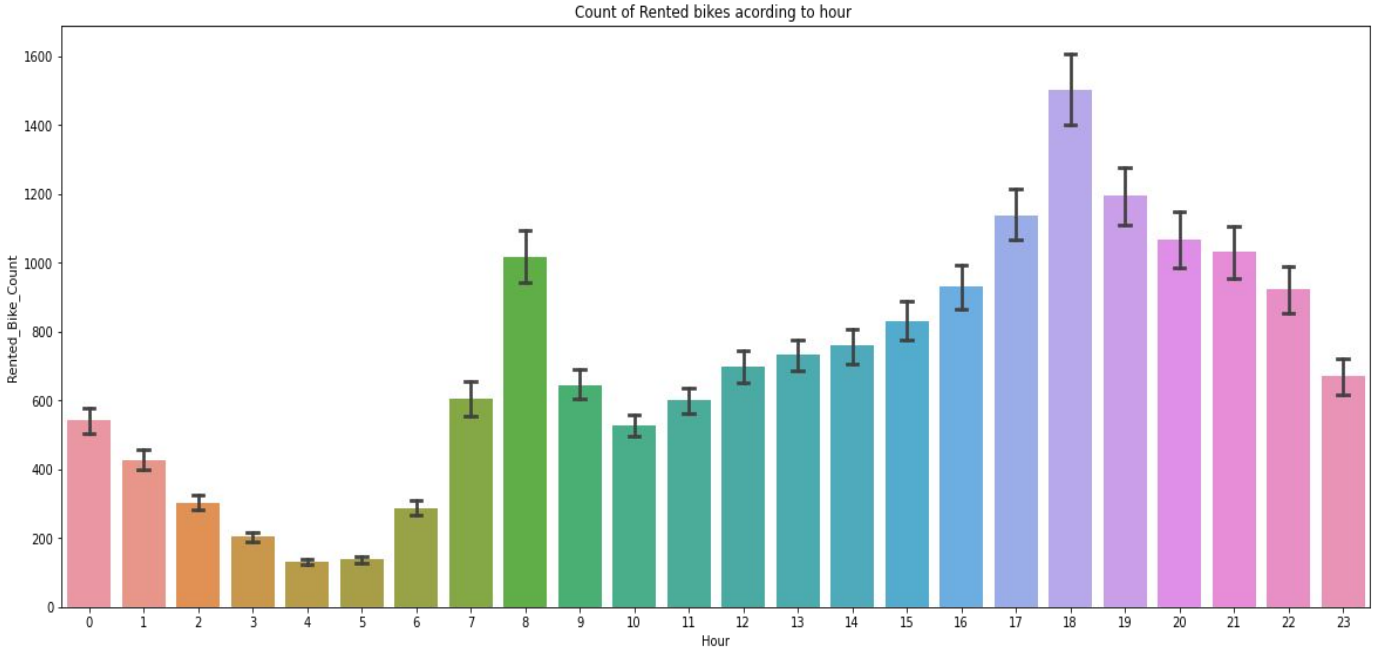
From this month and bike count bar plot we can say that on June ,July , august ,September , October , the use of rented bike is high as compared to other month

Analysis of weekdays_weekend variables



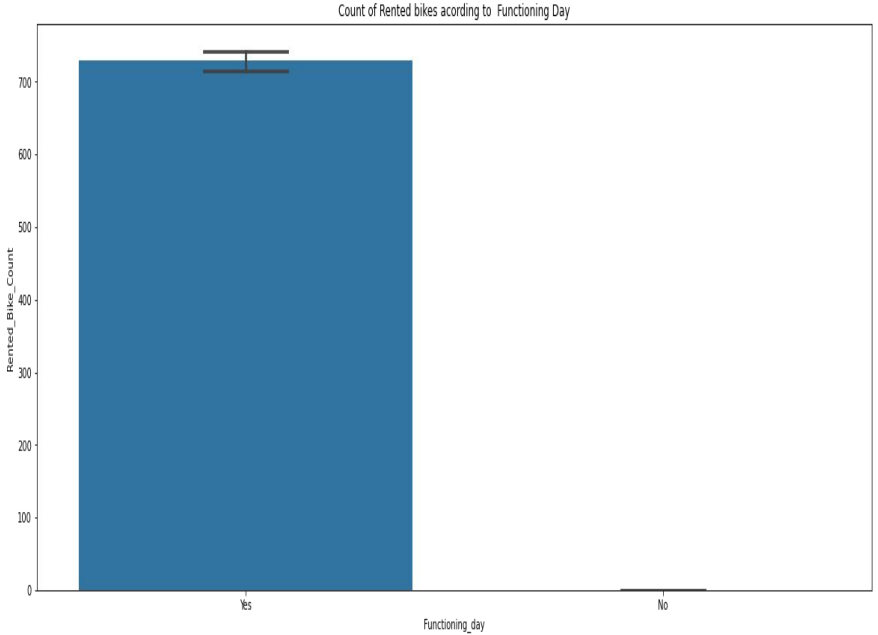
From the above graph we observe that demand of bike is high in weekdays as compaired to weekend . it may because of the office . peoples using rented bike for reaching their office .

Analysis of hour variables



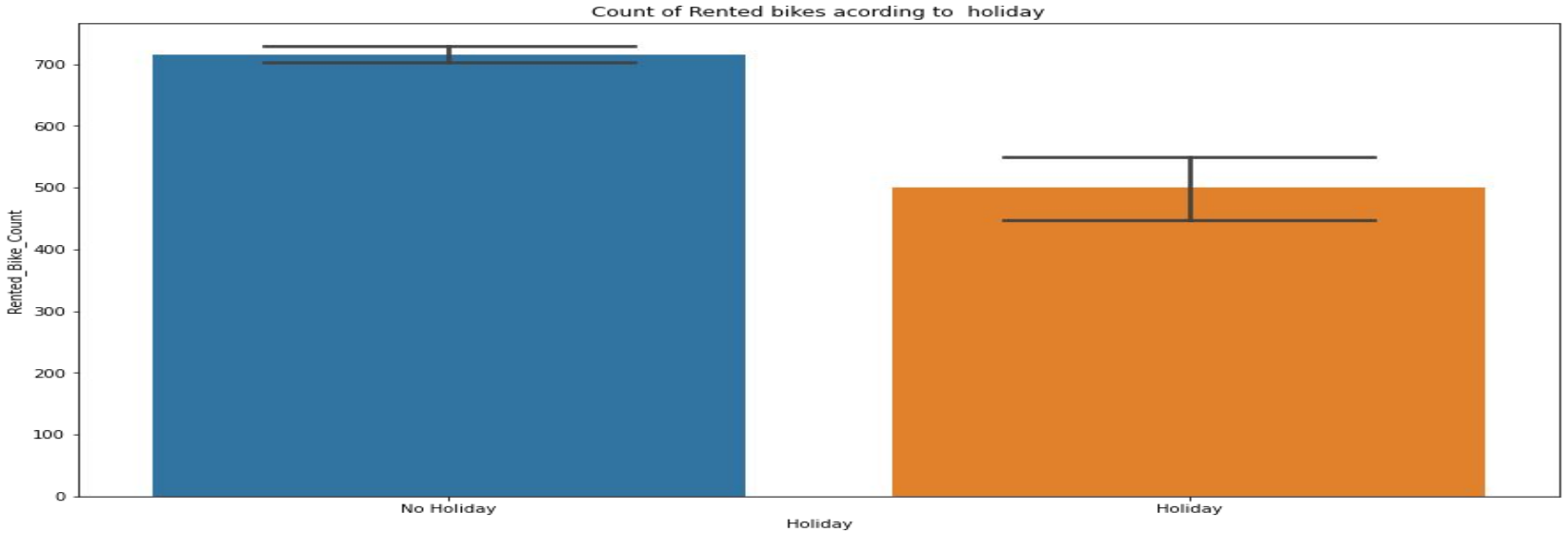
- From this graph we observe that people are using rented bike between 6AM to 9AM And 5pm to 7PM ,that means they use bikes for reaching their office

Analysis of functioning day variable



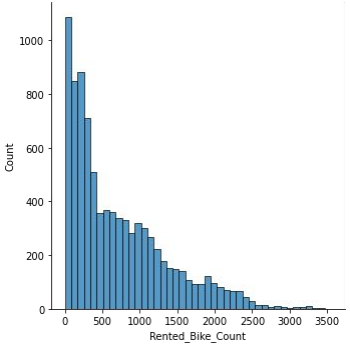
The above graph represent if the rented bike is using in functing day or not ** **Here we understand that people does not use bike in non functioning day

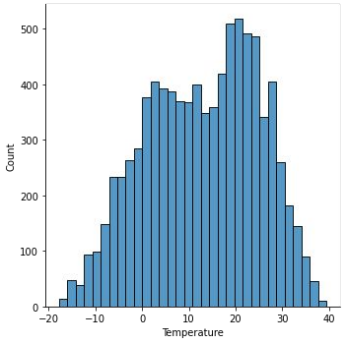
Analysis of holyday variable

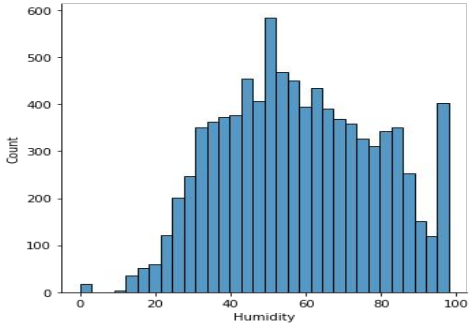


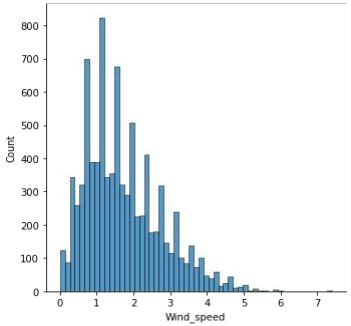
This graph shows the use of rented bike in holiday
We can understand from the graph that in nonholidays people use rented bike more as compared to holidays

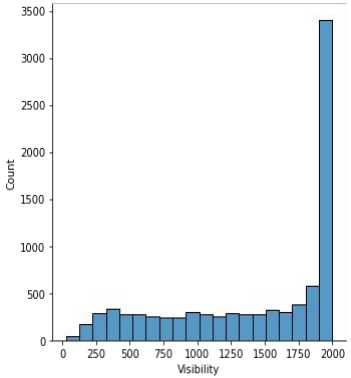
Rented bike count versus numerical variable

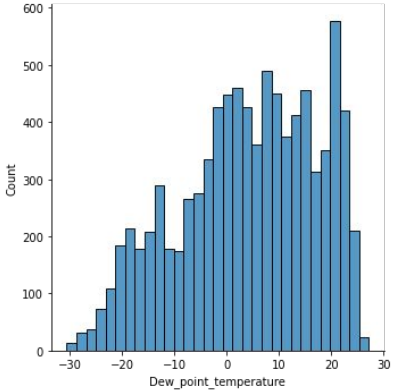




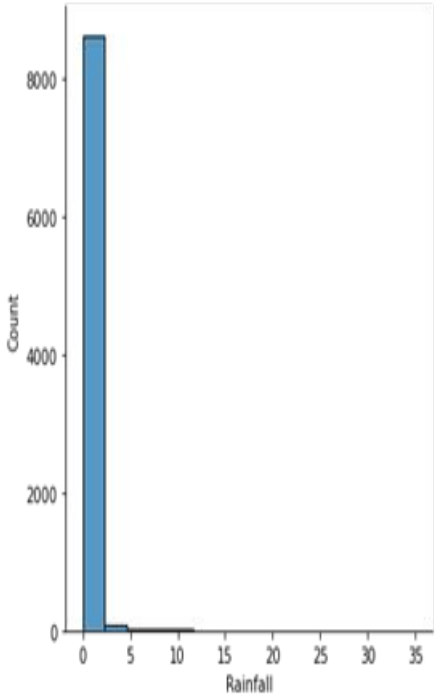


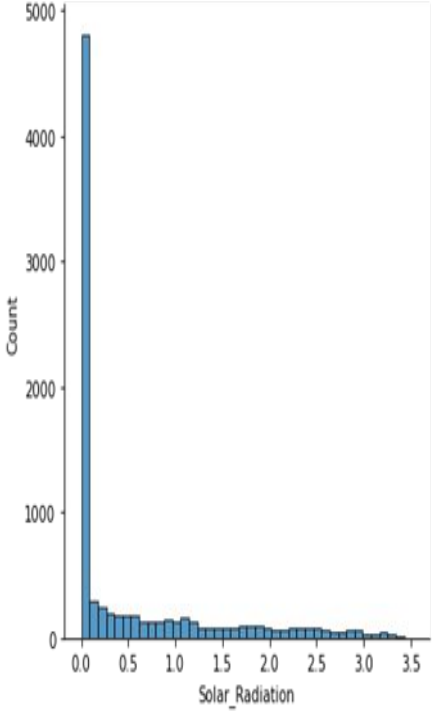


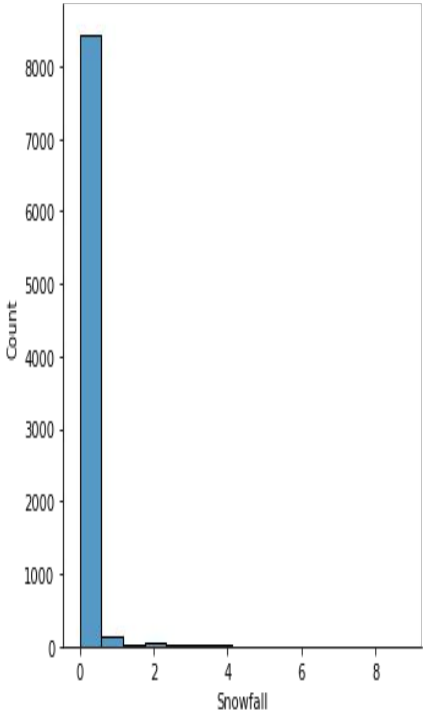




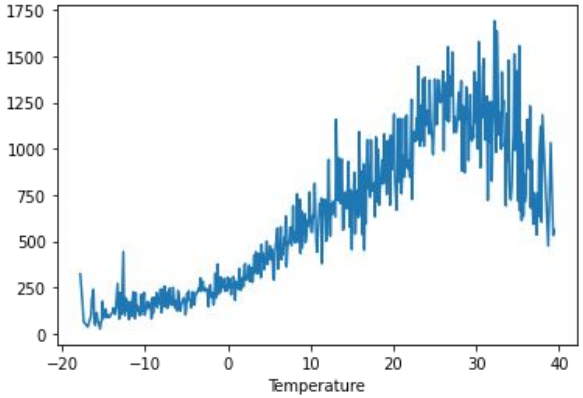
Rented bike count versus numerical variable

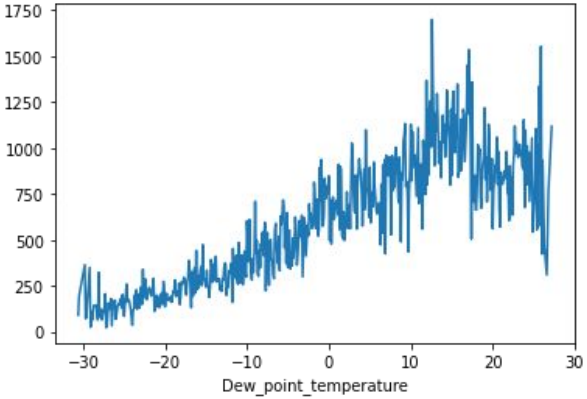


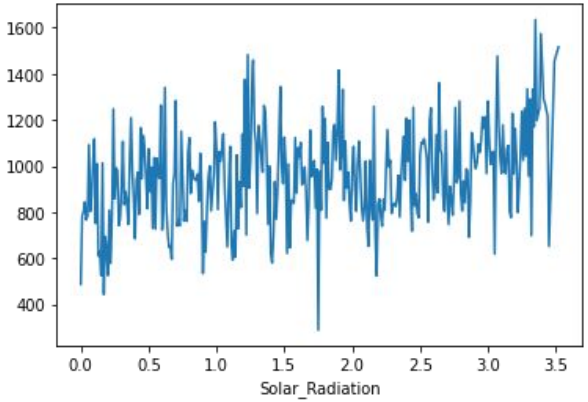




NUMERICAL VS RENTED BIKE COUNT

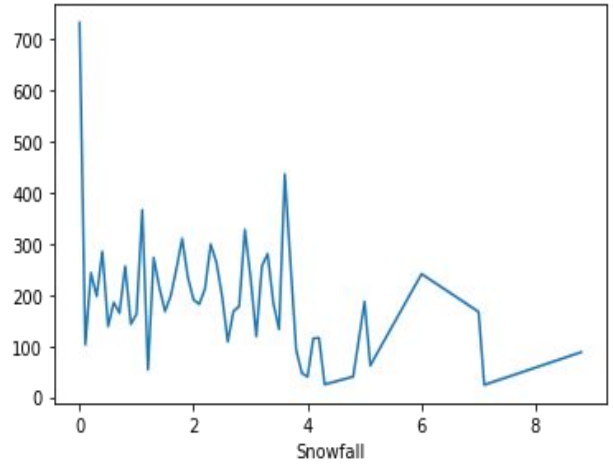


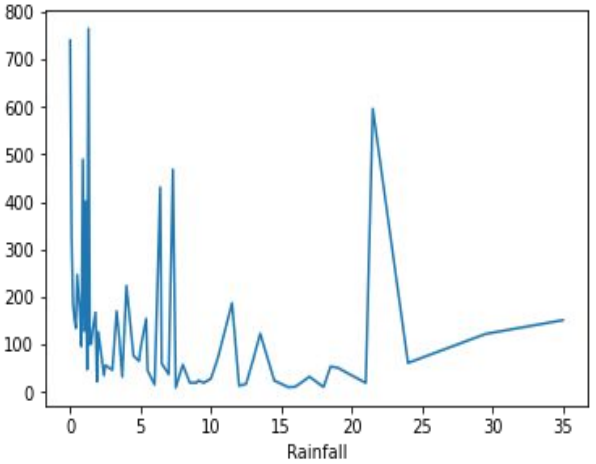


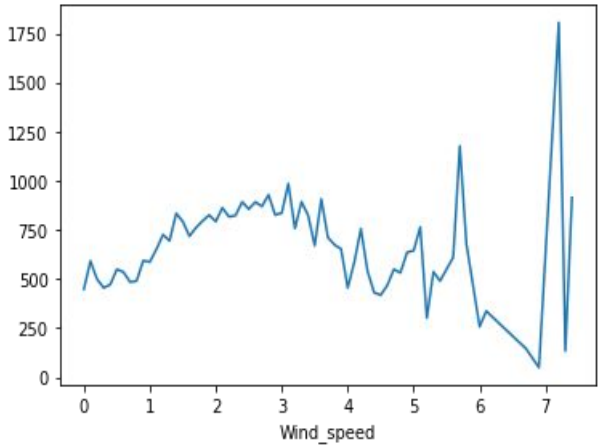


- From this plot we see that people like to ride bikes when I*t is around 25°C* people are using less number of rented bike around-20°C
- This plot shows dew point temperature is similar to temperature graph people use bike around the temperature 25°C
- from the above plot we see that, the amount of rented bikes is huge, when there is solar radiation, the counter of rents is around 1000

NUMERICAL VS RENTED BIKE COUNT

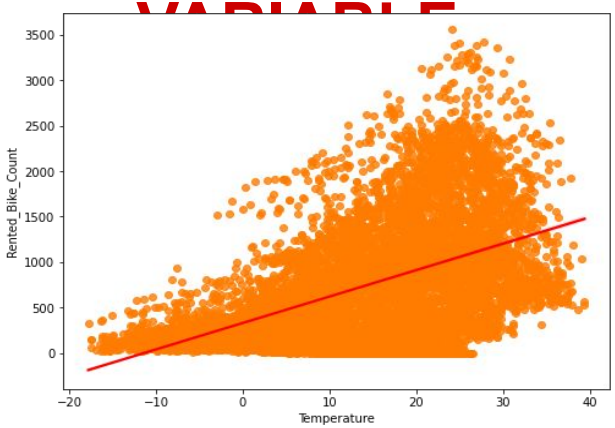


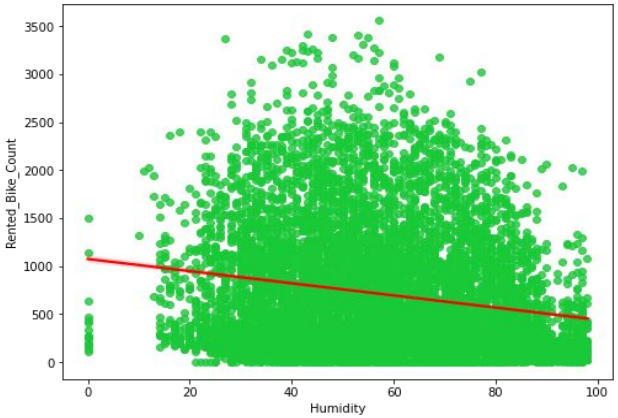


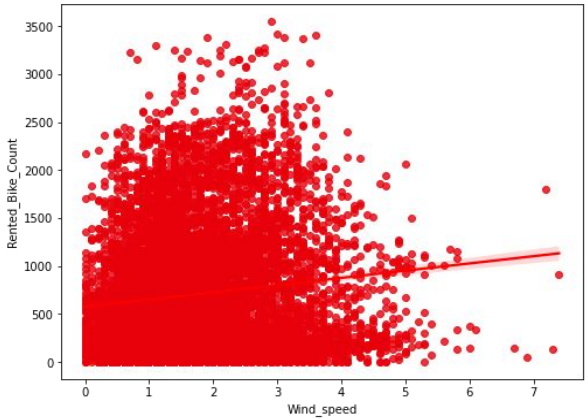


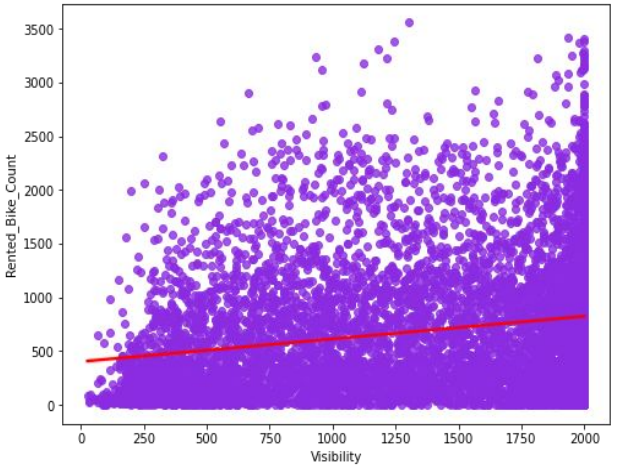
- From this above plot it shows that when the snowfall is more than 4cm the usage of rented bike is less
- From the above graph we can understand that bike renting is *not affected by rainfall , because demand of rent* bikes is not decreasing when there is 20mm of rainfall
- Above plot shows that the demand of rented bike is uniformly distributed because when the speed of wind was 7 m/s then the demand of bike also increase.

REGRESSION PLOT FOR NUMERICAL VARIABLE

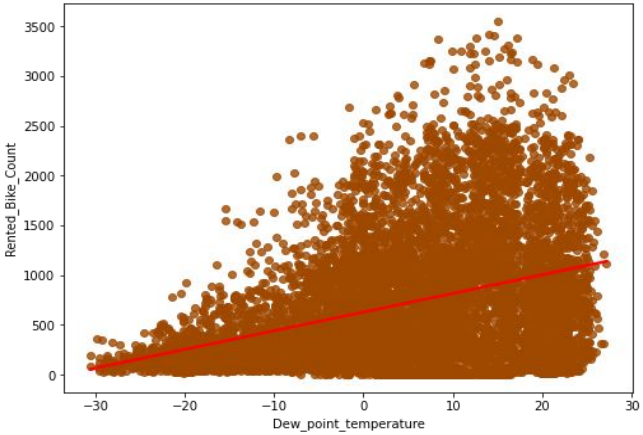


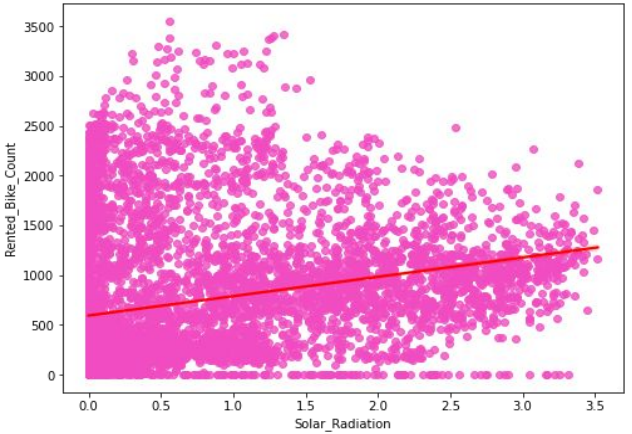


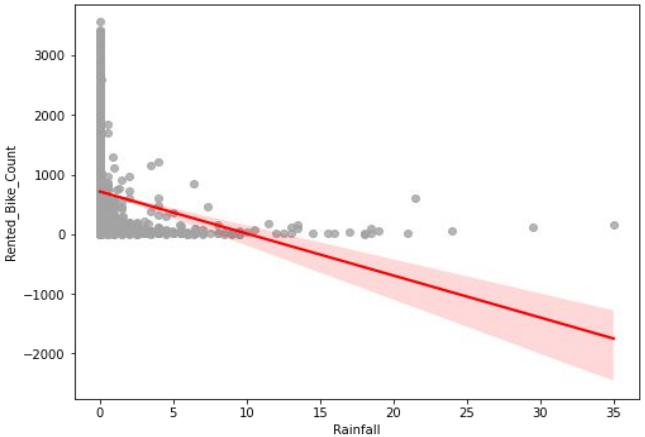


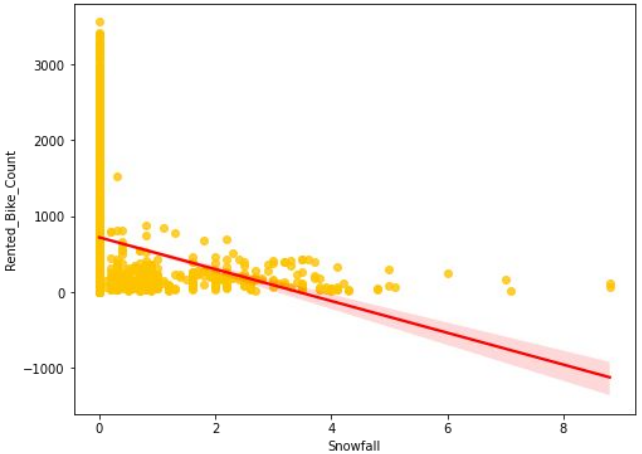


REGRESSION PLOT FOR NUMERICAL VARIABLE







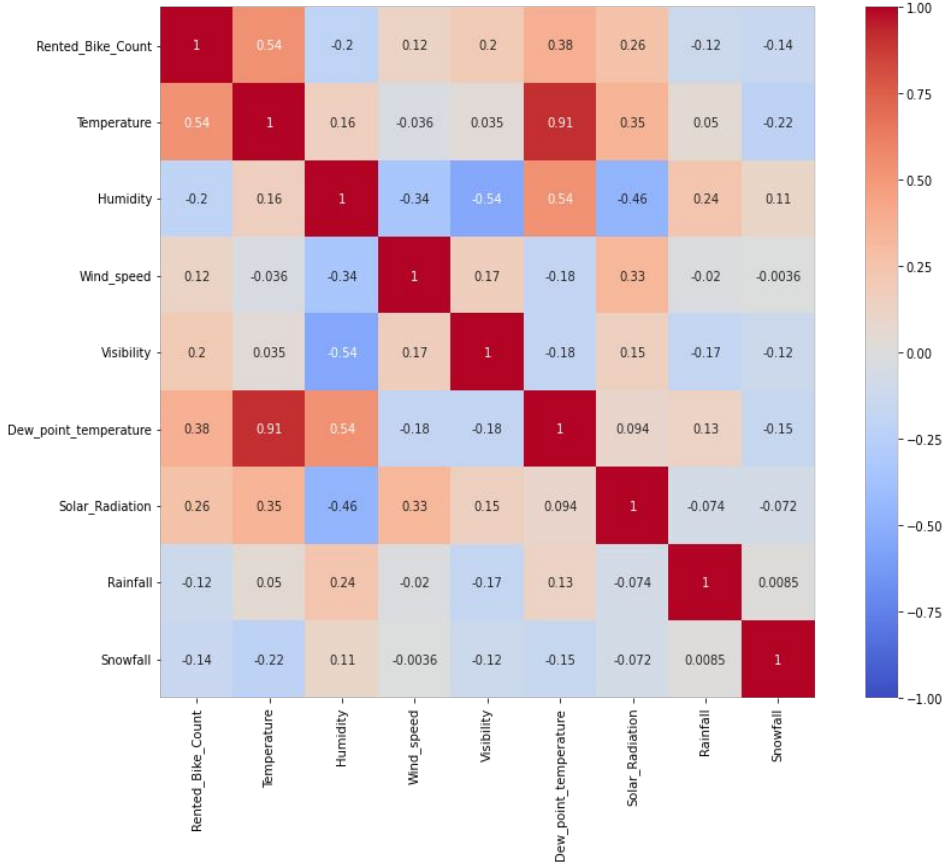


OBSERVATIONS

- Regression plot of numerical featuressee 'Temperature', 'Wind_speed','Visibility', 'Dew_point_temperature', 'Solar_Radiation‘ we can understand that rented bike count increases with these features

- 'Rainfall' ,'Snowfall' ,'Humidity' these features are negatively related with the rented bike count which means the rented bike count decreases when these features increase.

CORRELATION MATRIX



Highly correlated

- the temperature
- the dew point temperature
- the solar radiation

Negatively correlated

- Humidity
- Rainfall

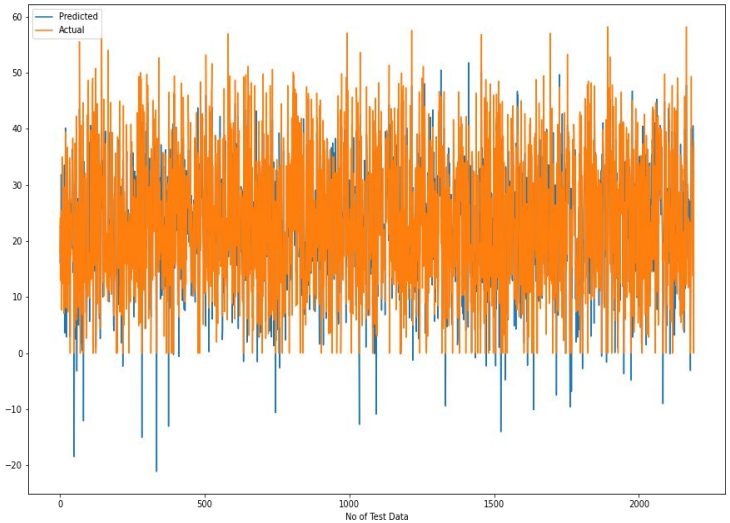
From the above correlation heatmap, We see that there is a positive correlation between columns 'Temperature' and 'Dew point temperature' i.e 0.91 so even if we drop this column then it does not affect the outcome of our analysis. And they have the same variations.. so we can drop the column 'Dew point temperature.

MODEL BUILDING

IN THIS PROJECT WE ARE USING 7 MODEL ON OUR DATA SET FOR GETTING BEST PERFORMANCE

- **LINEAR REGRESSION**
- **LASSO REGRESSION**
- **RIDGE REGRESSION**
- **ELASTICNET REGRESSION**
- **GRADIENT BOOSTING**
- **DECISION TREE**
- **RANDOM FORREST**
- **XGBOOST**

LINEAR REGRESSION



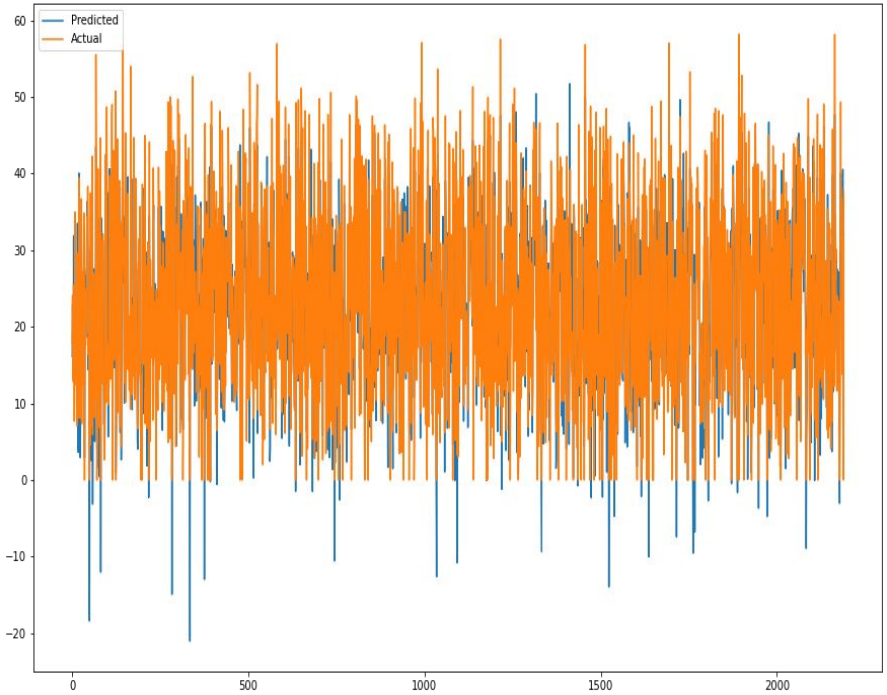
We train model by linear regression and we get results as follows:

R Squared for Training Data: 0.754
R Squared for Testing Data: 0..766

RMS for Training Data: 319.390
RMS for Testing Data: 312.062

MAE for Training Data: 217.536
MAE for Testing Data: 211.726

LASSO REGRESSION



We train model by lasso regression and we get results as follows:

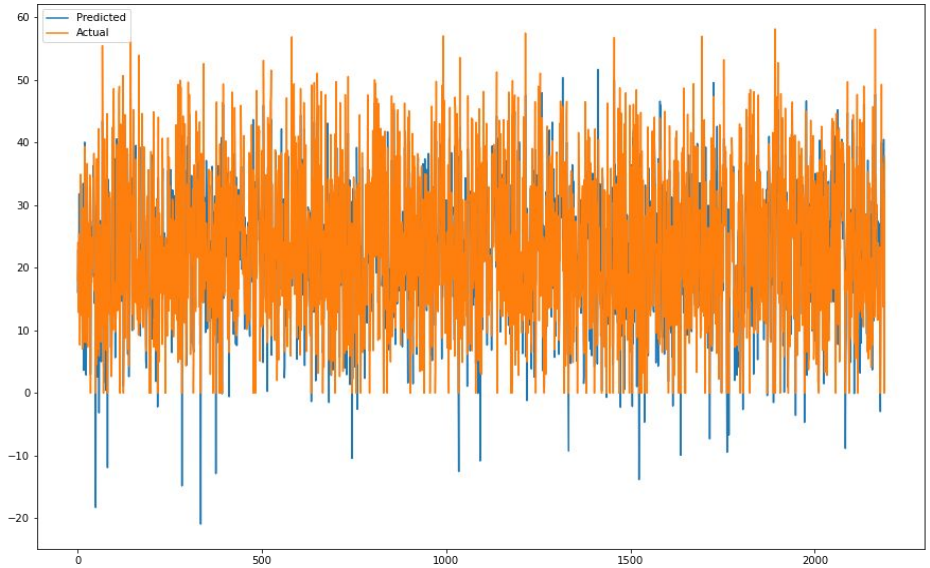
- **R Squared for Training Data: 0.754**
- **R Squared for Testing Data: 0..766**

- **RMS for Training Data: 319.390**
- **RMS for Testing Data: 312.062**

- **MAE for Training Data: 217.536**
- **MAE for Testing Data: 211.726**

RIDGE REGRESSION

By performing ridge regression we get the results are as follows :

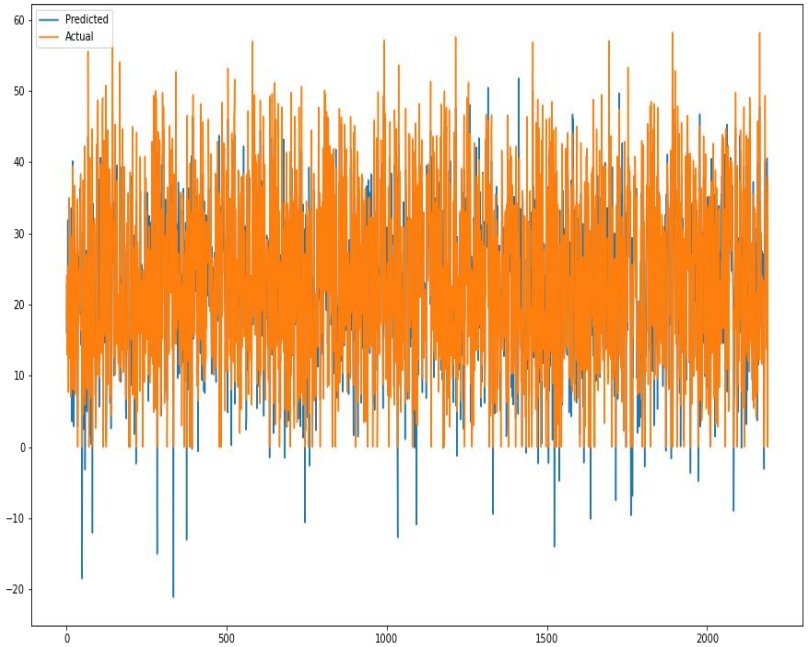


- **R Squared for Training Data: 0.7539**
- **R Squared for Testing Data :0.7904**

- **RMS for Training Data: 319.735**
- **RMS for Testing Data: 312.434**

- **MAE for Training Data: 217.696**
- **MAE for Testing Data: 211.917**

ELASTIC NET REGRESSION



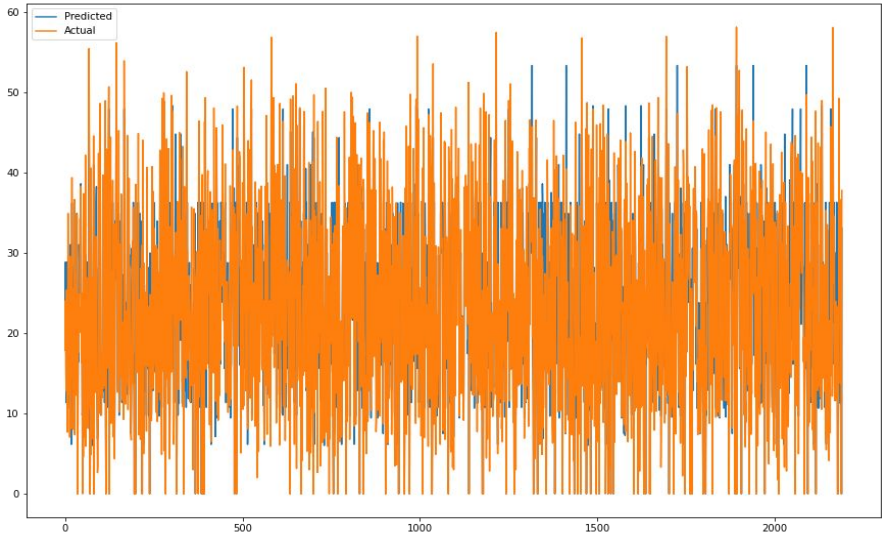
We train model by Elastic regression and we get results as follows:

- **R Squared for Training Data: 0.754**
- **R Squared for Testing Data: 0..766**

- **RMS for Training Data: 319.402**
- **RMS for Testing Data: 312.075**

- **MAE for Training Data: 217.542**
- **MAE for Testing Data: 211.733**

DECISION TREE



We train model by Decision tree and we get results as follows:

- **R Squared for Training Data: 0.689**
- **R Squared for Testing Data: 0.651**

- **RMS for Training Data: 359.273**
- **RMS for Testing Data: 381.502**

- **MAE for Training Data: 240.805**
- **MAE for Testing Data: 253.517**

RANDOM FOREST

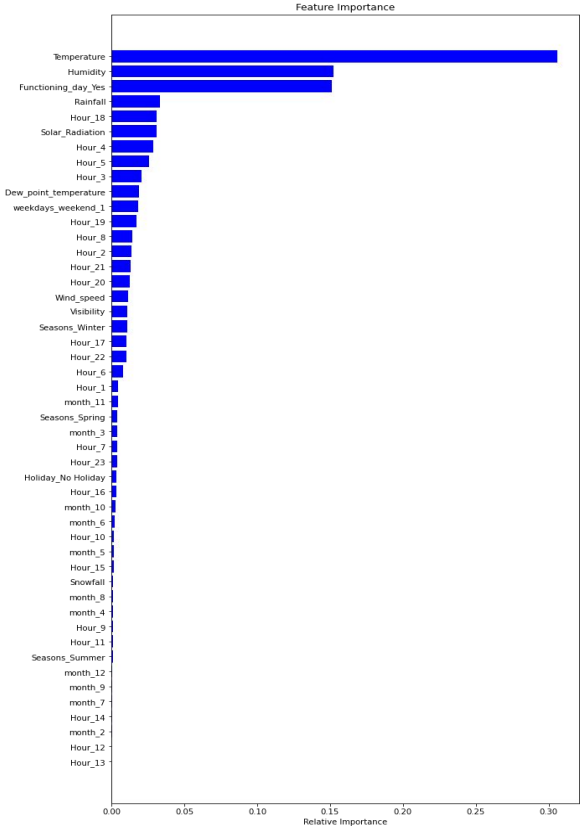
We train model by Random forest and we get results as follows:

- **R Squared for Training Data: 0.989**
- **R Squared for Testing Data: 0.906**

- **RMS for Training Data: 67.364**
- **RMS for Testing Data: 197.244**

- **MAE for Training Data: 40.276**
- **MAE for Testing Data: 112.071**

	Feature	Feature Importance
0	Temperature	0.31
35	Functioning_day_Yes	0.15
1	Humidity	0.15
25	Hour_18	0.03
5	Solar_Radiation	0.03
6	Rainfall	0.03
12	Hour_5	0.03
11	Hour_4	0.03
10	Hour_3	0.02
26	Hour_19	0.02
47	weekdays_weekend_1	0.02
4	Dew_point_temperature	0.02
9	Hour_2	0.01
2	Wind_speed	0.01
33	Seasons_Winter	0.01
28	Hour_21	0.01



GRADIENT BOOSTING

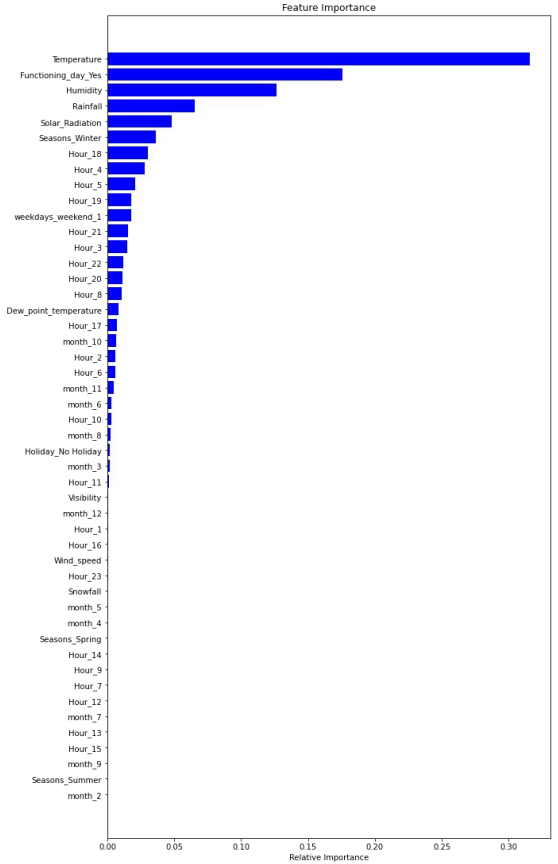
We train model by Gradient boosting and we get results as follows:

R Squared for Training Data: 0.845
R Squared for Testing Data: 0.835

RMS for Training Data: 253.748
RMS for Testing Data: 261.964

MAE for Training Data: 168.44
MAE for Testing Data: 175.007

	Feature	Feature Importance
0	Temperature	0.32
35	Functioning_day_Yes	0.18
1	Humidity	0.13
6	Rainfall	0.07
5	Solar_Radiation	0.05
33	Seasons_Winter	0.04
11	Hour_4	0.03
25	Hour_18	0.03
28	Hour_21	0.02
26	Hour_19	0.02
12	Hour_5	0.02
47	weekdays_weekend_1	0.02
10	Hour_3	0.01
9	Hour_2	0.01
44	month_10	0.01
13	Hour_6	0.01
29	Hour_22	0.01
4	Dew_point_temperature	0.01
27	Hour_20	0.01



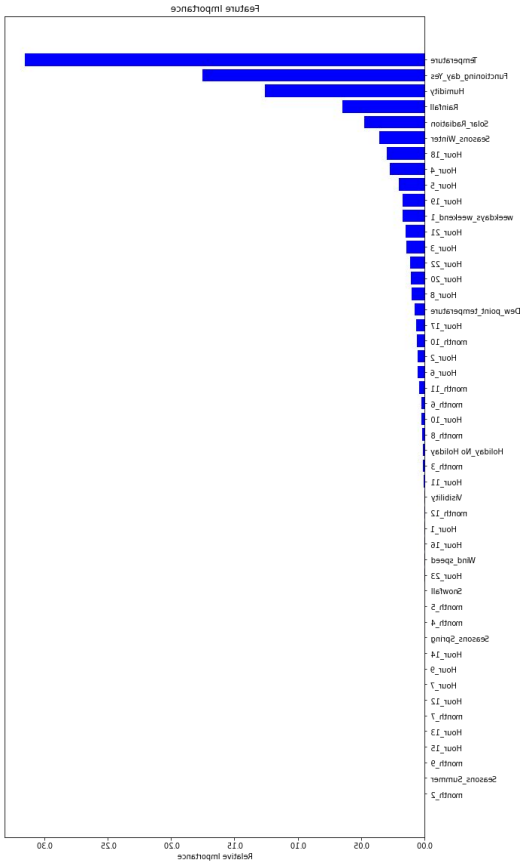
GRADIENT BOOSTING REGRESSOR WITH GRIDSEARCHCV

We train model by Gradient boosting and we get results as follows:

- **R Squared for Training Data: 0.946**
- **R Squared for Testing Data: 0.923**

- **RMS for Training Data: 148.92**
- **RMS for Testing Data:188.80**
- **MAE for Training Data: 91.965**
- **MAE for Testing Data:115.46**

	Feature	Feature Importance
0	Temperature	0.31
35	Functioning_day_Yes	0.16
1	Humidity	0.15
5	Solar_Radiation	0.04
6	Rainfall	0.04
11	Hour_4	0.03
25	Hour_18	0.03
33	Seasons_Winter	0.02
26	Hour_19	0.02
12	Hour_5	0.02
47	weekdays_weekend_1	0.02
10	Hour_3	0.02
4	Dew_point_temperature	0.02
44	month_10	0.01



CHALLENGES

- Large Dataset to handle.
- Needs to plot lot of Graphs to analyse.
- Feature engineering
- Feature selection
- Optimising the model
- Carefully tuned Hyperparameters as it affects the R2 score.
- Execution Takes Time

CONCLUSION

- We train the dataset to predict the number of rented bike is used in the given weather conditions .
- initially we did EDA on all the features of our dataset ,analyze dependent variable and categorical variables.
- Implemented 7 machine learning algorithms Regression,lasso,ridge,elasticnet,decission tree, Random Forest and XGBoost. All algorithms performed really well on both training dataset and testing dataset so we can say that variance is less and no issues of overfittings are present.
- "Random forest regression(90%)" and "Gradient Boosting regression(gridsearch cv) has highest R2 score(84%).
- DECISION TREE algorithm has comparatively less R2 score(65%)

CONCLUSION

		Model	MAE	MSE	RMSE	R2_score	Adjusted R2
Training set	0	Linear regression	211.726	97382.918	312.062	0.767	0.790
	1	Lasso regression	211.865	97574.975	312.370	0.766	0.790
	2	Ridge regression	211.917	97615.595	312.435	0.790	0.760
	3	Elastic net regression Test	211.734	97391.114	312.075	0.767	0.761
	4	Decision tree test	253.517	145543.947	381.502	0.651	0.640
	5	Random forest regression	112.072	38905.493	197.245	0.907	0.900
	6	Gradient boosting regression	168.449	64388.434	253.749	0.845	0.840

Test set	0	Linear regression	211.726	97382.918	312.062	0.767	0.790
	1	Lasso regression	211.865	97574.975	312.370	0.766	0.790
	2	Ridge regression	211.917	97615.595	312.435	0.790	0.760
	3	Elastic net regression Test	211.734	97391.114	312.075	0.767	0.761
	4	Decision tree test	253.517	145543.947	381.502	0.651	0.640
	5	Random forest regression	112.072	38905.493	197.245	0.907	0.900
	6	Gradient boosting regression	175.007	68625.633	261.965	0.835	0.830
	7	Gradient boosting regression	175.007	68625.633	261.965	0.835	0.830
	8	Gradient Boosting gridsearchcv	115.463	35646.927	188.804	0.924	0.910

However ,this results are not the ultimate . as this data is time dependent , the values for variables like temperature, solar_radiation, wind_speed etc., Will not always be consistent. Therefore, there will be scenarios where the model might not perform well. As machine learning is an exponentially evolving field, we will have to be prepared for all contingencies and also keep checking our model from time to time. Therefore, having a quality knowledge and keeping pace with the ever evolving ML field would surely help one to stay a step ahead in future


THANK YOU