CAPSTONE PROJECT

BIKE SHARING DEMAND PREDICTION

Anil Bhattt

CONTENTS:

- Problem statement
- Introduction
- Data Summary
- Exploratory Data Analysis
- Model Building
- Evaluation
- Challenges
- Conclusion

PROBLEM STATEMENT :

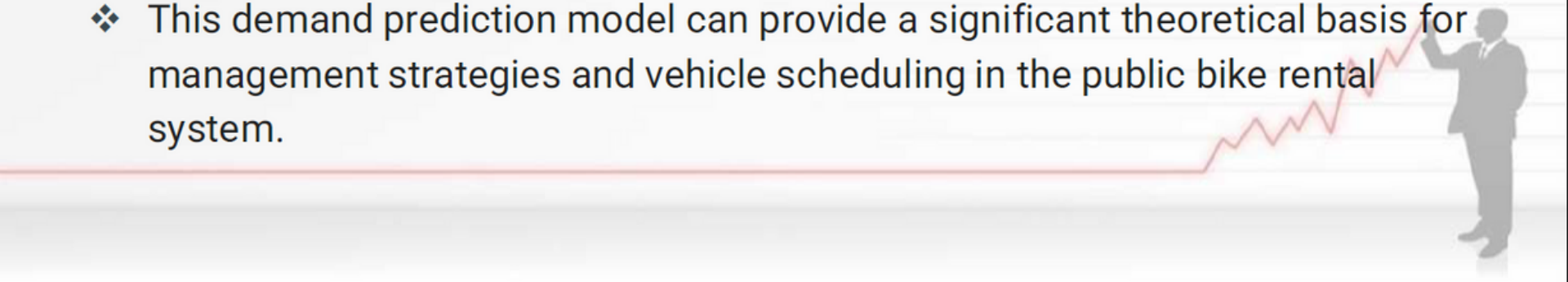
- Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort.
- It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time.
- Eventually, providing the city with a stable supply of rental bikes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

INTRODUCTION :

- Prediction of bike sharing demand can help bike sharing companies to allocate bikes better and ensure more sufficient circulation of bikes for customers.

- This presentation proposes a real-time method for predicting bike renting based on historical data, weather data, and time data.

- This demand prediction model can provide a significant theoretical basis for management strategies and vehicle scheduling in the public bike rental system.



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.

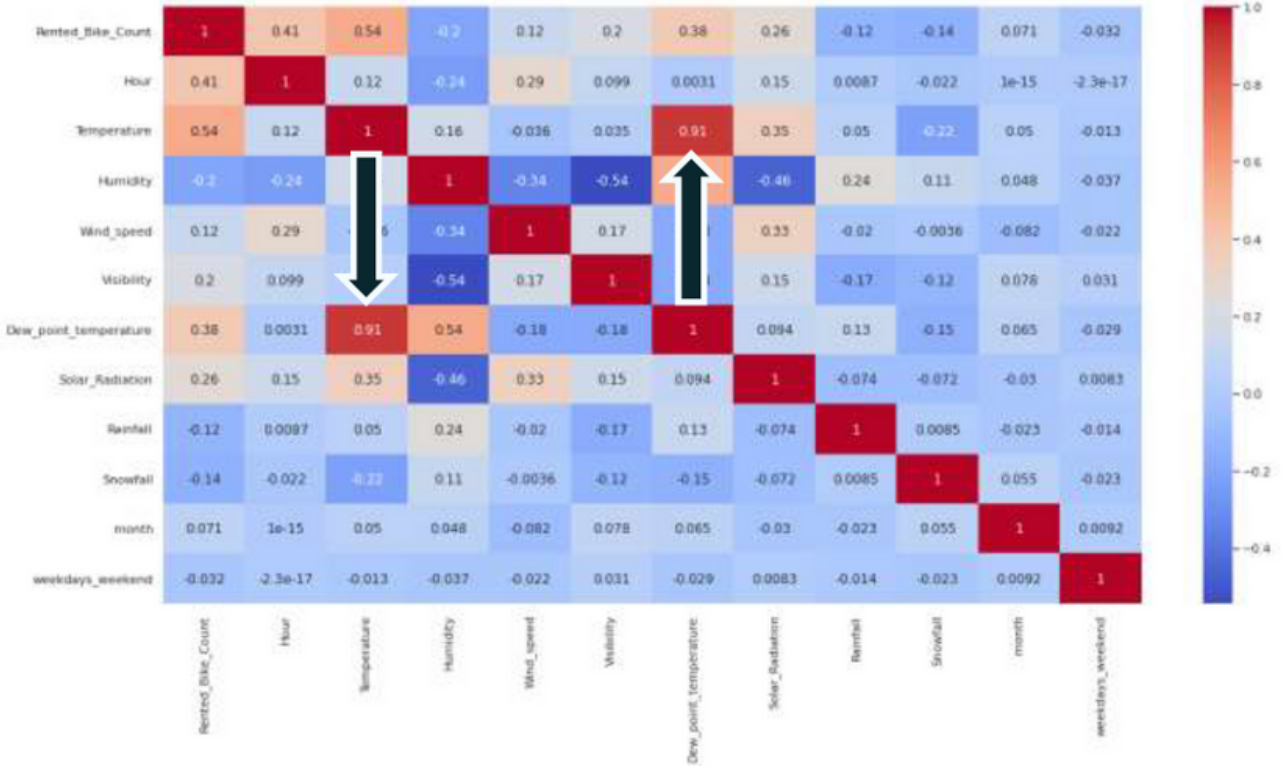
- Date: year-month-day
- Rented Bike count -Count of bikes rented at each hour
- Hour -Hour of the day
- Temperature-Temperature in Celsius
- Humidity -%
- Windspeed -m/s
- Visibility -10m
- Dew point temperature -Celsius
- Solar radiation -MJ/m2
- Rainfall -mm

DATA SUMMARY Cont. :

- Snowfall -cm
- Seasons -Winter, Spring, Summer, Autumn
- Holiday -Holiday/No holiday
- Functional Day -NoFunc(Non Functional Hours), Fun(Functional hours)

- 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 column -which we need to predict for new observations

CHECKING MULTICOLLINEARITY :



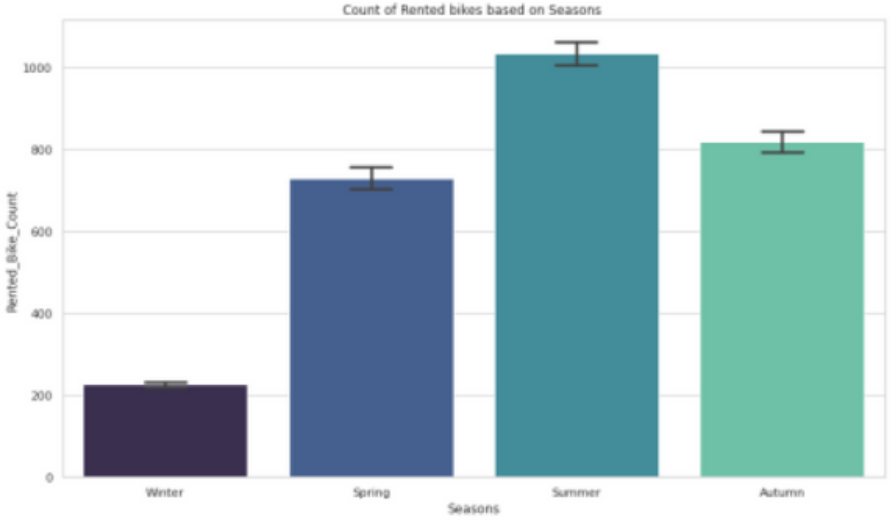
- Variables like Dew Point Temperature, and Temperature are highly correlated.
- We don't want Multicollinearity in our dataset it affects the performance metrics so we are dropping Dew Point Temperature.

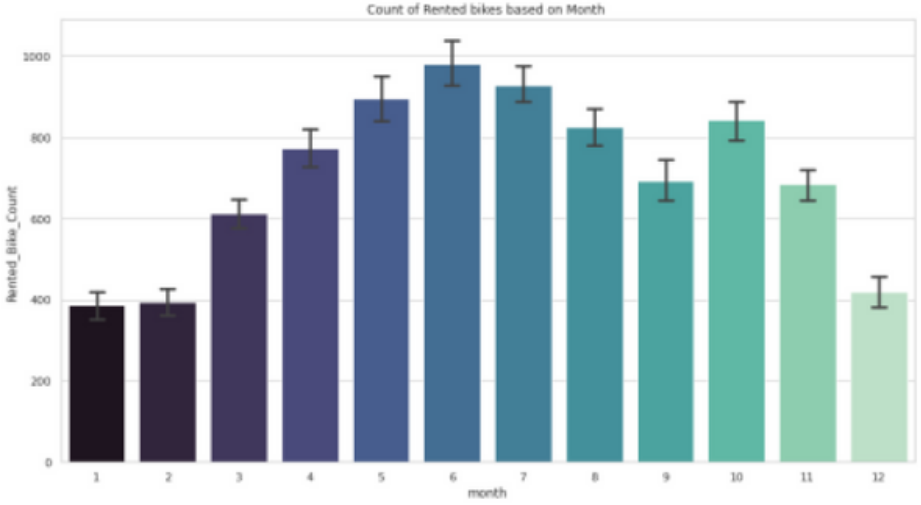
EXPLORATORY DATA ANALYSIS (EDA) :

- 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.

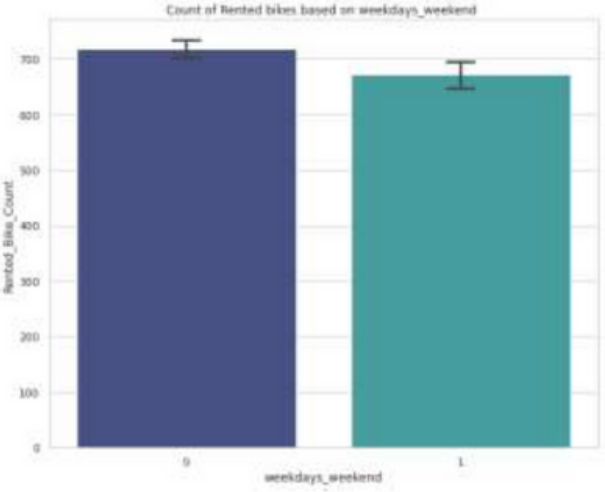
- EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

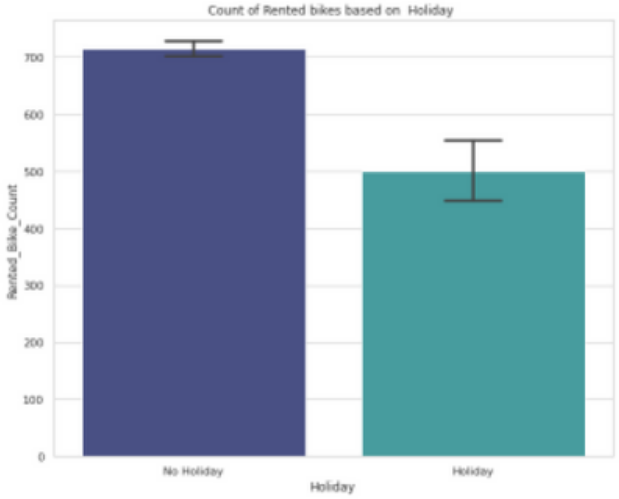
ANALYSIS BY DATA VISUALIZATION:

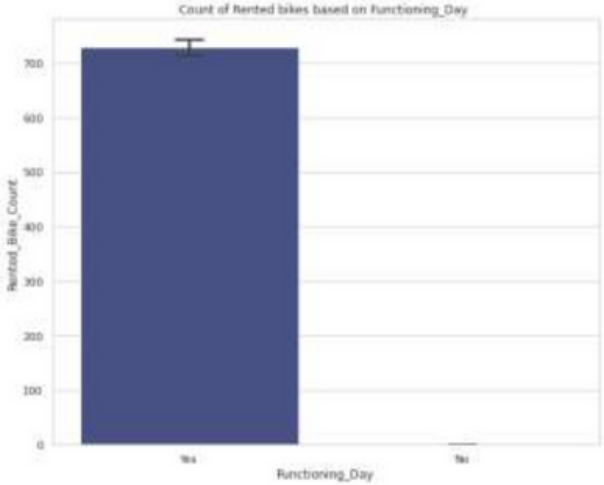


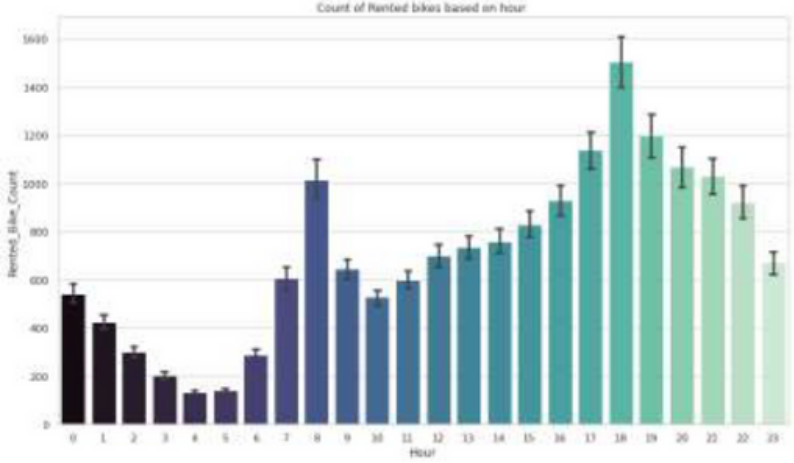


- In the summer season, the use of rented bikes is high whereas in the winter season the use of rented bikes is very low.
- The demand for the rented bike is high from the month 5 to 10



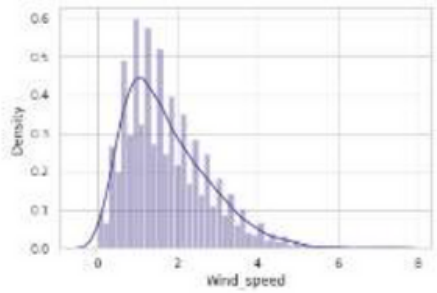


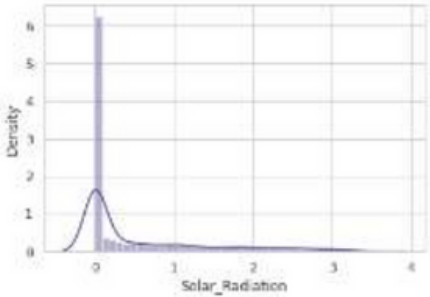


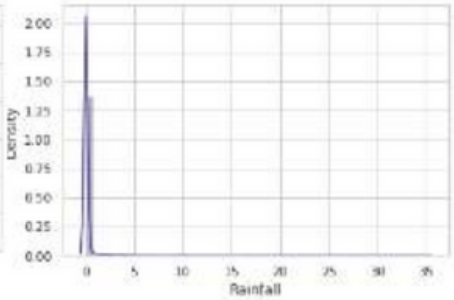


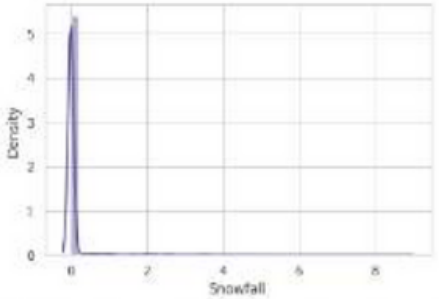
- By visualizing graph 1, People prefer to rent bikes on the weekdays but the difference is not significant.
- By visualizing graph 2, People prefer bikes to rent on non-holidays as compared to holidays.
- By visualizing graph 3, People use rented bikes only on functioning days.
- By visualizing graph 4, people generally use rented bikes during their working hours from 7 am to 9 am and 5 pm to 8 pm.

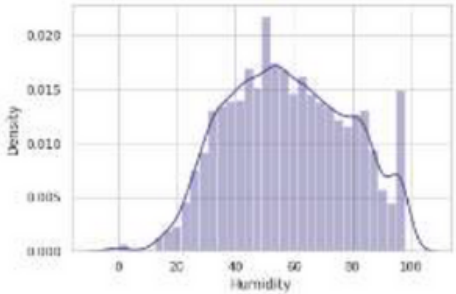
VISUALISING DISTRIBUTIONS :

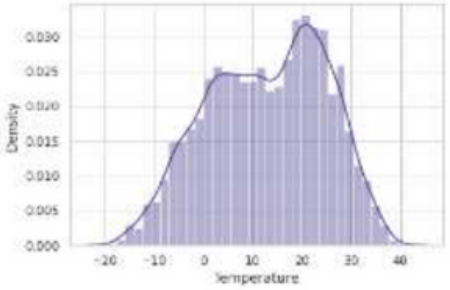


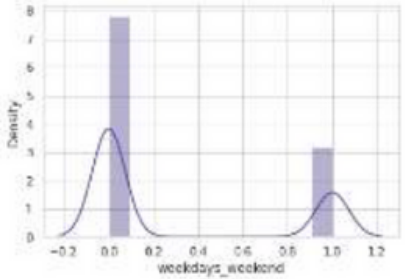


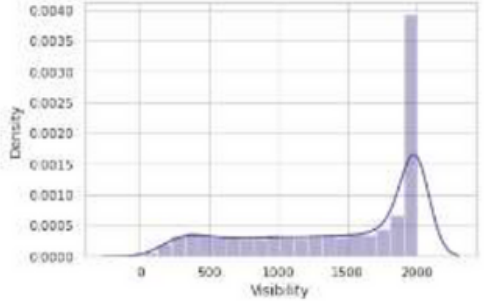


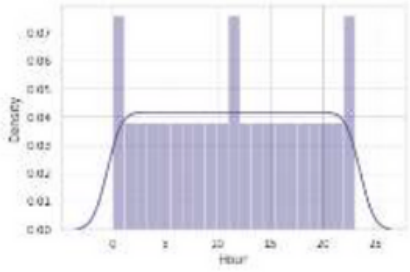


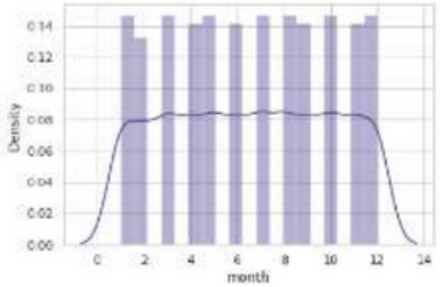








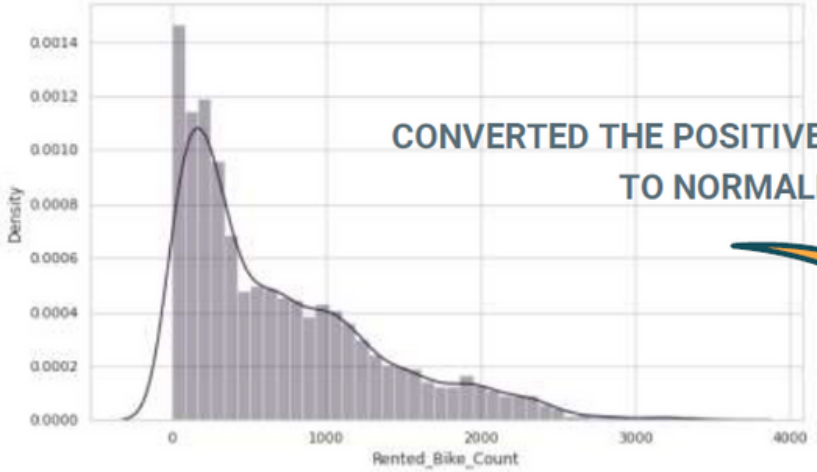




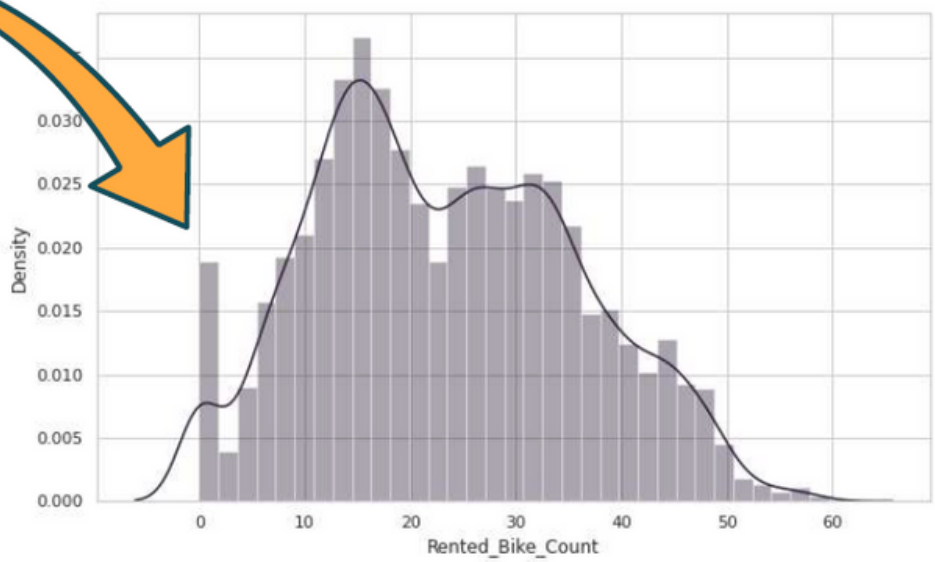
ANALYSIS FROM VISUALISING DISTRIBUTIONS :

- “Temperature”, “Hour”, “Month” and “Humidity” columns follow a uniform distribution.
- “Wind Speed”, “Solar Radiation”, “Rainfall” and “Snowfall” are having positively skewed distribution.
- “Visibility” column is negatively skewed.

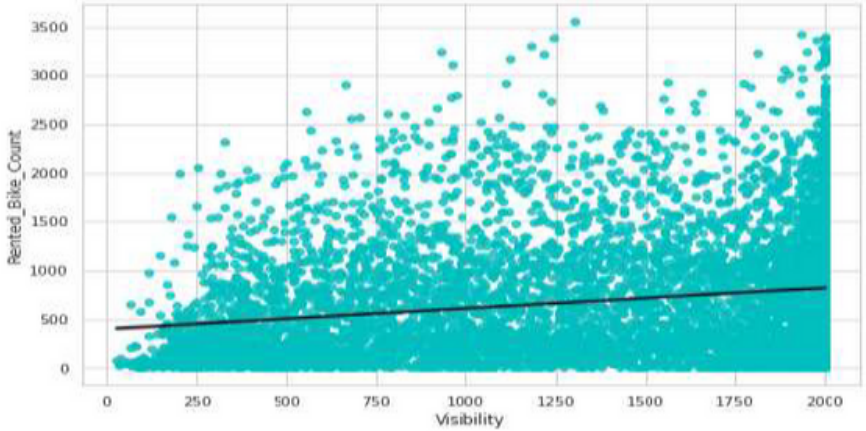
DISTRIBUTION OF TARGETED FEATURE :

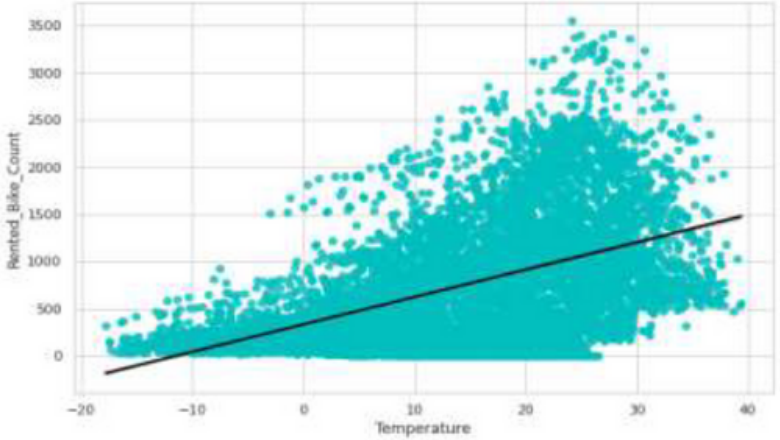


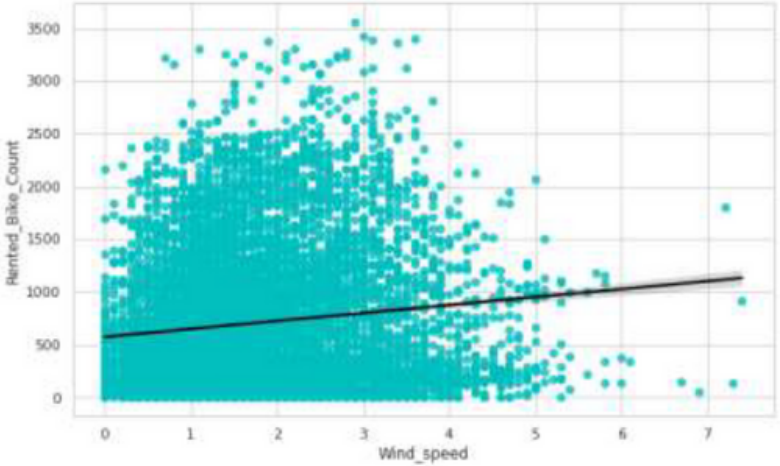
CONVERTED THE POSITIVELY SKEWED DATA
TO NORMALISED

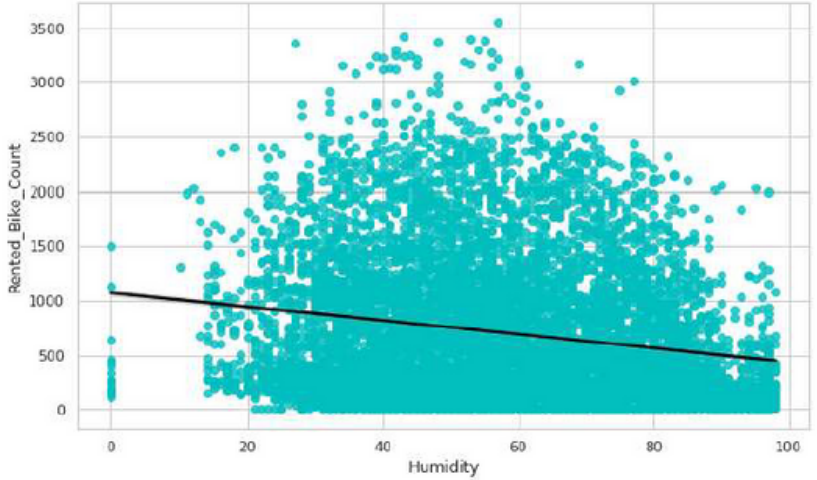


REGRESSION PLOTS : DEPENDENDENT FEATURE TO TARGETED FEATURE

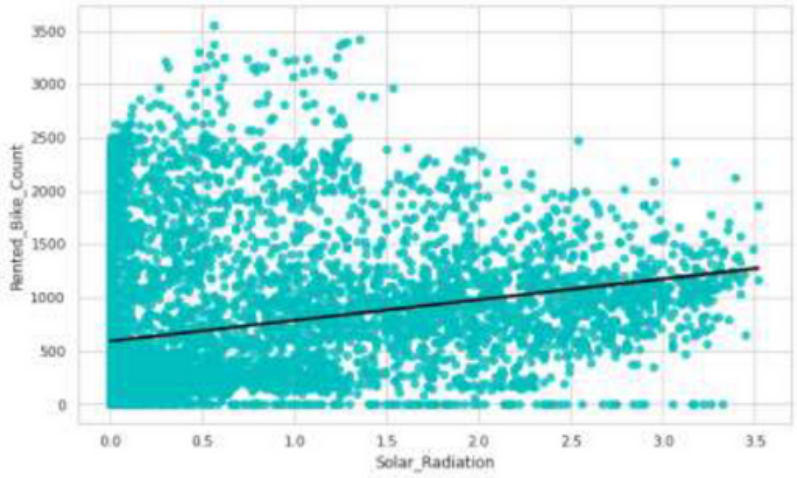


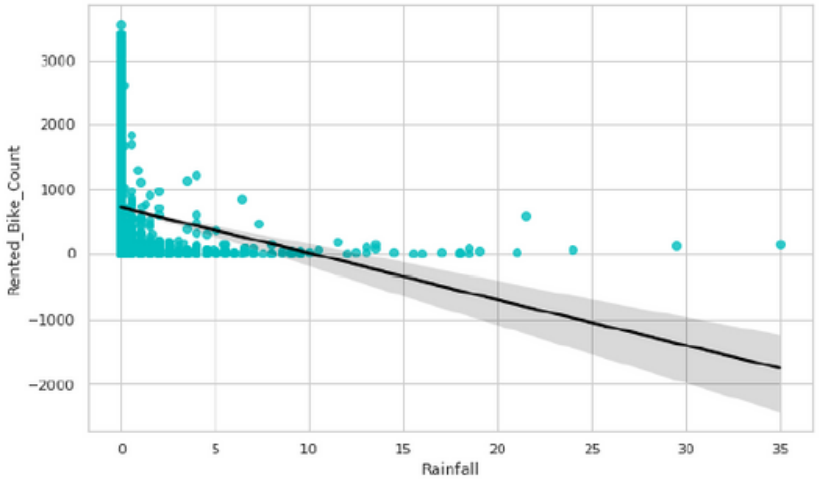


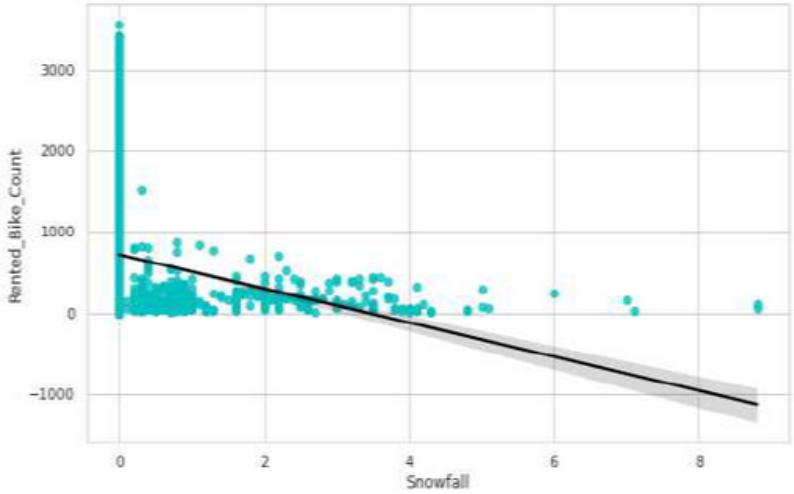




CONT....


Rented_Bike_Count
Solar_Radiation
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0
500
1000
1500
2000
2500
3000
3500


Rented_Bike_Count
Rainfall
0
5
10
15
20
25
30
35
−2000
−1000
0
1000
2000
3000


Rented_Bike_Count
Snowfall
0
2
4
6
8
−1000
0
1000
2000
3000

MODEL BUILDING :

❖ Linear regression model

Regularized linear regression:

- Lasso regression model
- Ridge regression model
- Elastic net regression model

❖ Decision tree regression model

Ensemble techniques:

- Random-forest regression model
- XG–Boost regression model
- XG–Boost GridsearchCV Regression model

EVALUATION OF MODELS :

		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	4.658	37.606	6.132	0.756	0.75
	2	Lasso regression	7.255	91.594	9.570	0.405	0.39
	3	Elasticnet regression	5.892	59.247	7.697	0.615	0.61
	4	Decision tree regression	5.166	51.274	7.161	0.667	0.66
	5	Random forest regression	0.946	2.074	1.440	0.987	0.99
	6	XG Boost Regression	3.460	21.073	4.591	0.863	0.86
	7	XG boost regg GridserachCV	1.378	4.060	2.015	0.974	0.97
Test set	0	Linear regression	4.658	36.645	6.053	0.768	0.76
	1	Ridge regression	4.659	36.647	6.054	0.768	0.76
	2	Lasso regression	7.456	96.775	9.837	0.387	0.37
	3	Elasticnet regression	6.011	61.651	7.852	0.610	0.60
	4	Decision tree regression	5.383	54.582	7.388	0.654	0.65
	5	Random forest regression	2.585	15.863	3.983	0.900	0.90
	6	XG Boost Regression	3.670	23.709	4.869	0.850	0.85
	7	XG boost regg GridserachCV	2.527	13.980	3.739	0.911	0.91

- Out of all the above models “Random forest Regressor” gives the highest Adj.R2 score of 99%.
- For the Train Set and “XG Boost Grid search CV” gives the highest Adj.R2 score of 91% for the Test set.
- No overfitting is seen.

CHALLENGES FACED

- Feature engineering.
- Feature selection.
- Dummy encoding in XG-boost regressor. (because it does not support sparse data type)
- Model Training and performance improvement.

CONCLUSIONS :

- No overfitting is seen.

- When we compare the root mean squared error and mean absolute error of all the models, the XG- boost grid search CV regression model has less root mean squared error and mean absolute error, ending with the Adj. R-squared of 91% in test data. So, finally, this model is best for predicting the bike rental count on daily basis.

- For all the models, temperature and Functioning days were ranked as the most influential variable to predict the rental bike demand at each hour.


thank you