Capstone Project-2

Bike Sharing Demand Prediction

Team Members:

Soni Rani

Vivek Kumar

Suraj Singh

CONTENT

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 a 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 public bike rental system.

Data Summary

- Date : year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of the day
- 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
- Functioning Day – Functioning day/Non Functioning day

Let's Look At our Dataset

- The original dataset has 14 columns and 8760 rows.
- The data types of various columns are Object , Float and Integer.
- Dependent variable being Rented Bike Count and all other variables are our feature or independent variables like Hour, Temperature etc.

```
#checking the top five row to take a glimpse of the data
bike_df.head()
```

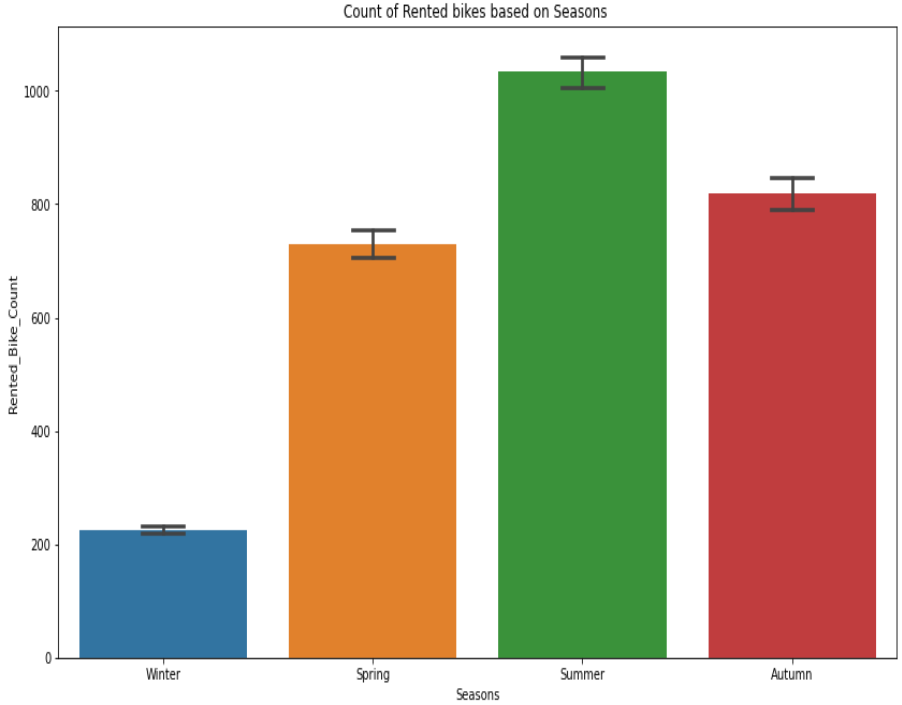
	Date	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Seasons	Holiday	Functioning Day
0	01/12/2017	254	0	-5.2	37	2.2	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
1	01/12/2017	204	1	-5.5	38	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

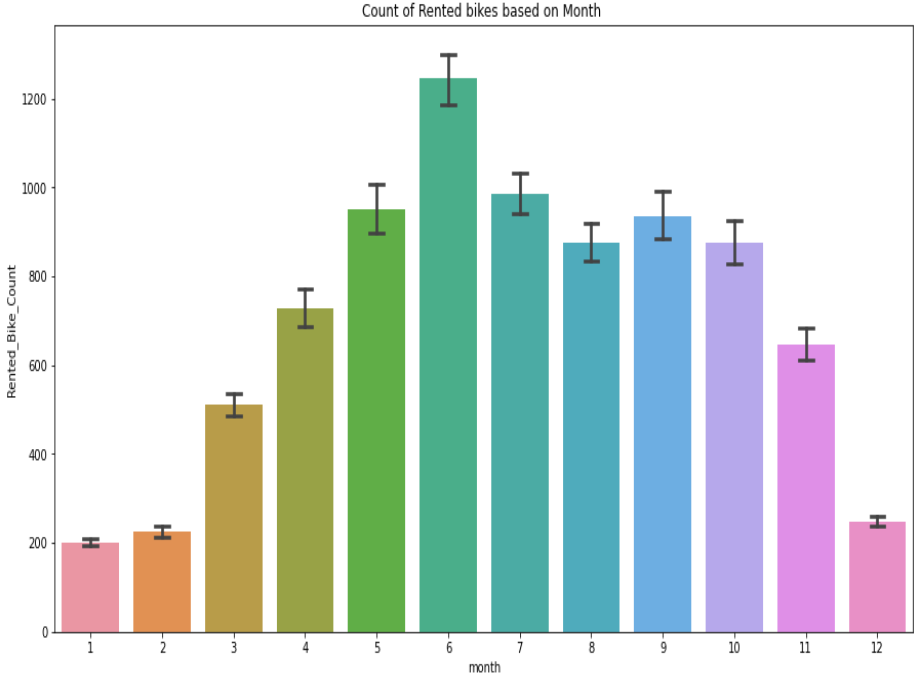
Exploratory data analysis

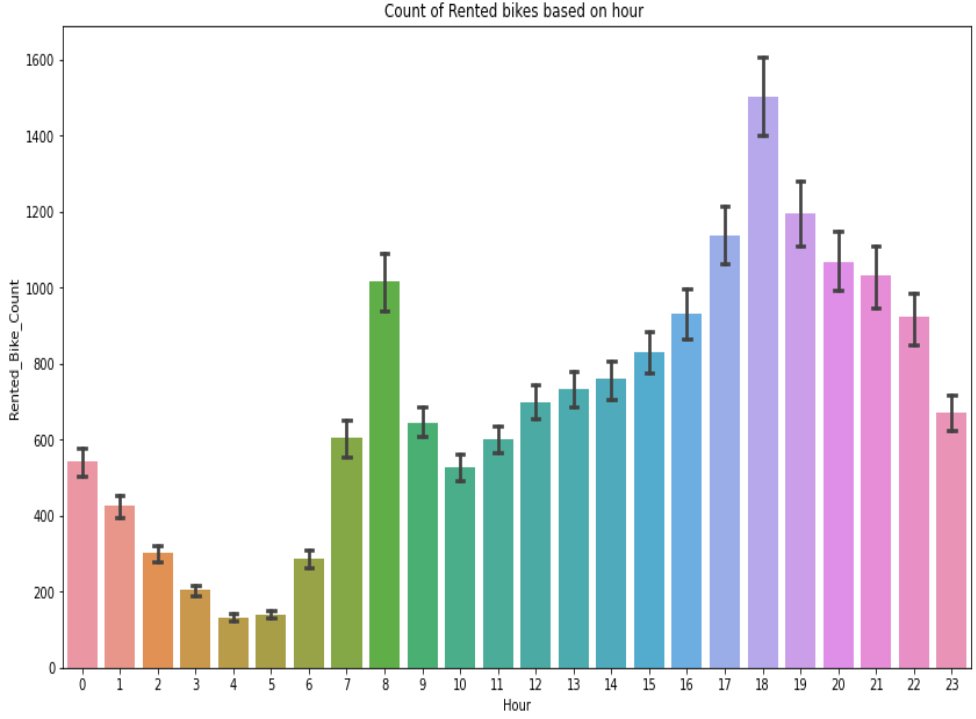
- Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies ,to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

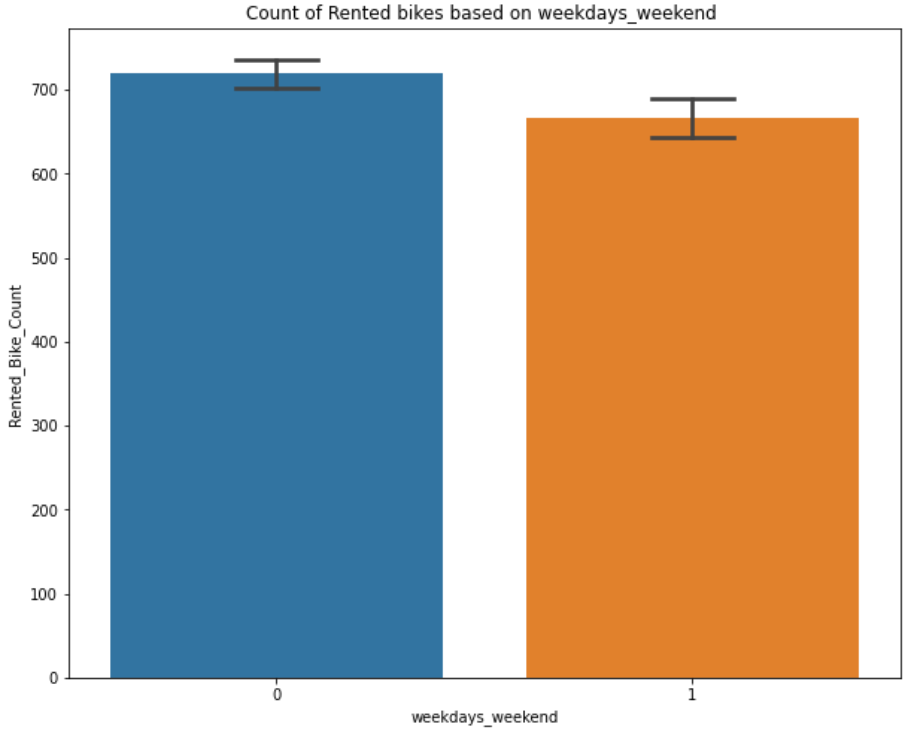
- EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

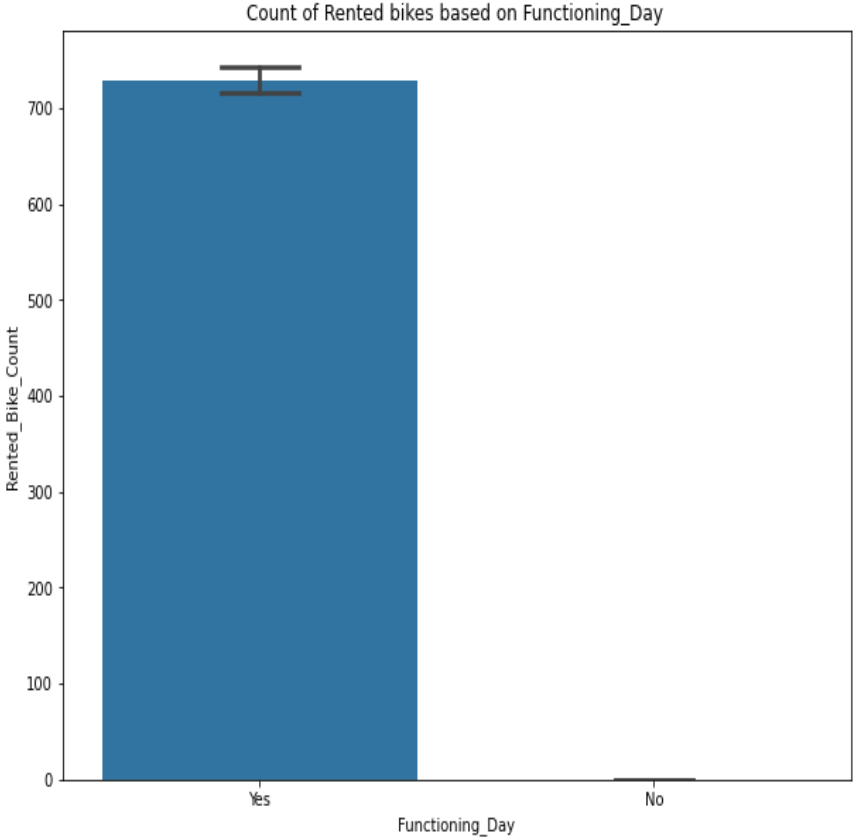
Analysis of data by visualization

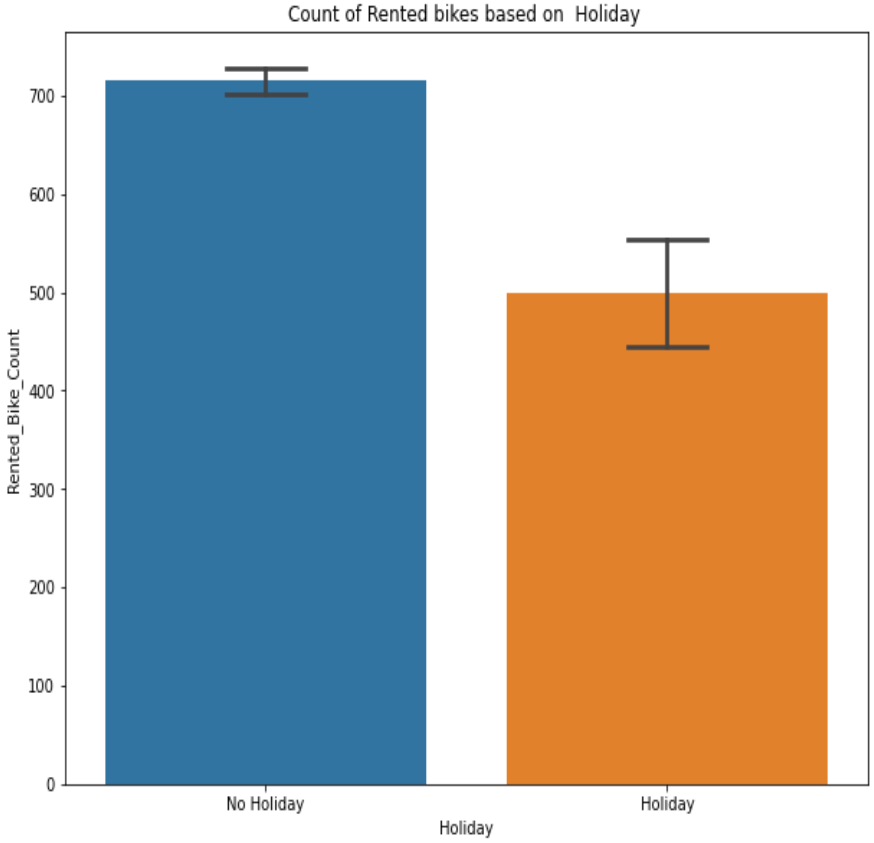





Count of Rented bikes based on hour
Rented_Bike_Count
Hour
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
0
200
400
600
800
1000
1200
1400
1600


Count of Rented bikes based on weekdays_weekend
Rented_Bike_Count
weekdays_weekend
0
1
0
100
200
300
400
500
600
700

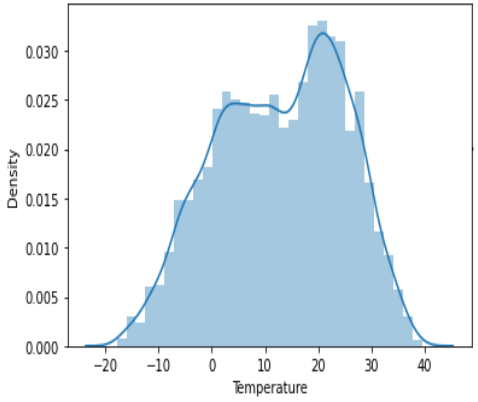

Count of Rented bikes based on Functioning_Day
Rented_Bike_Count
0
100
200
300
400
500
600
700
Yes
No
Functioning_Day

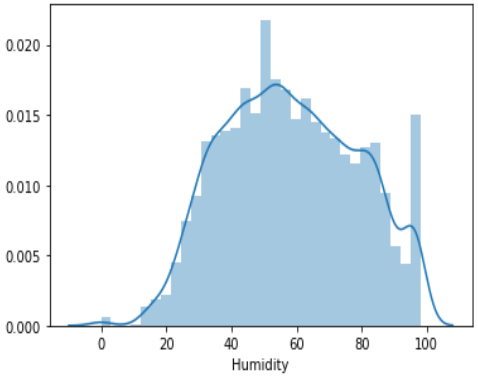

Count of Rented bikes based on Holiday
Rented_Bike_Count
0
100
200
300
400
500
600
700
No Holiday
Holiday
Holiday

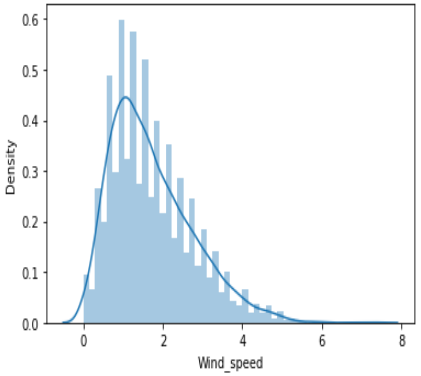
Hypothesis

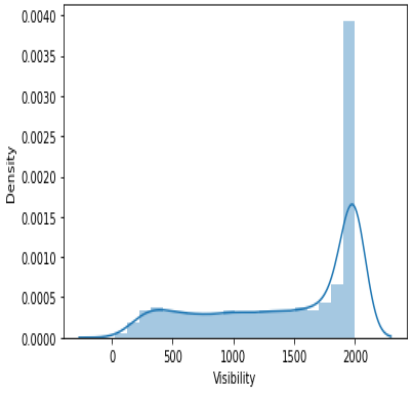
1. In summer season the use of rented bike is high whereas in winter season the use of rented bike is very low.
2. In the month 5 to 10 the demand of the rented bike is high as compare to other months.
3. People generally use rented bikes during their working hour from 7am to 9am and 5pm to 8pm.
4. In the week days the demand of the bike is higher because of the office as compare to the weekend.
5. People use rented bike only in functioning day.
6. Use of rented bike is more on no holiday as compare to holiday**.**

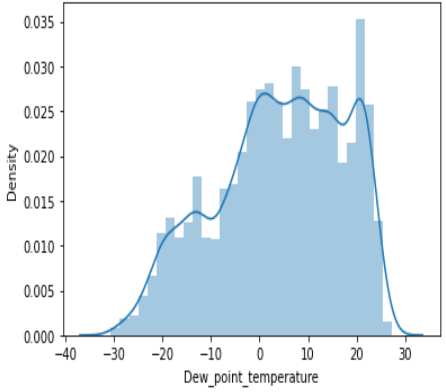
Visualizing Distributions

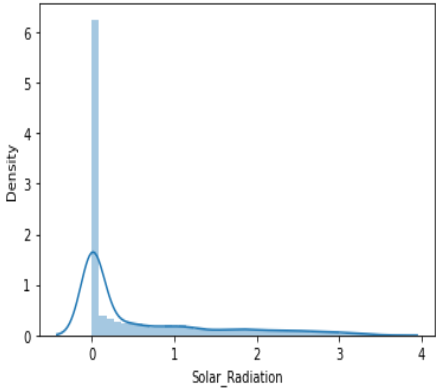


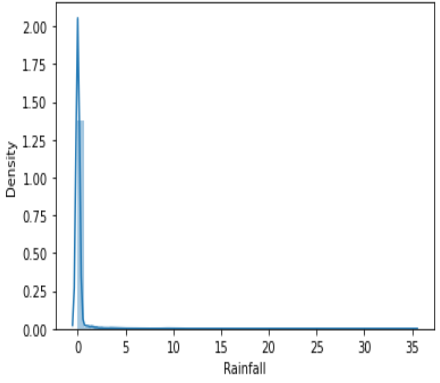


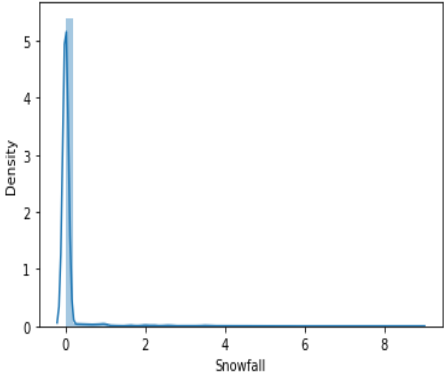












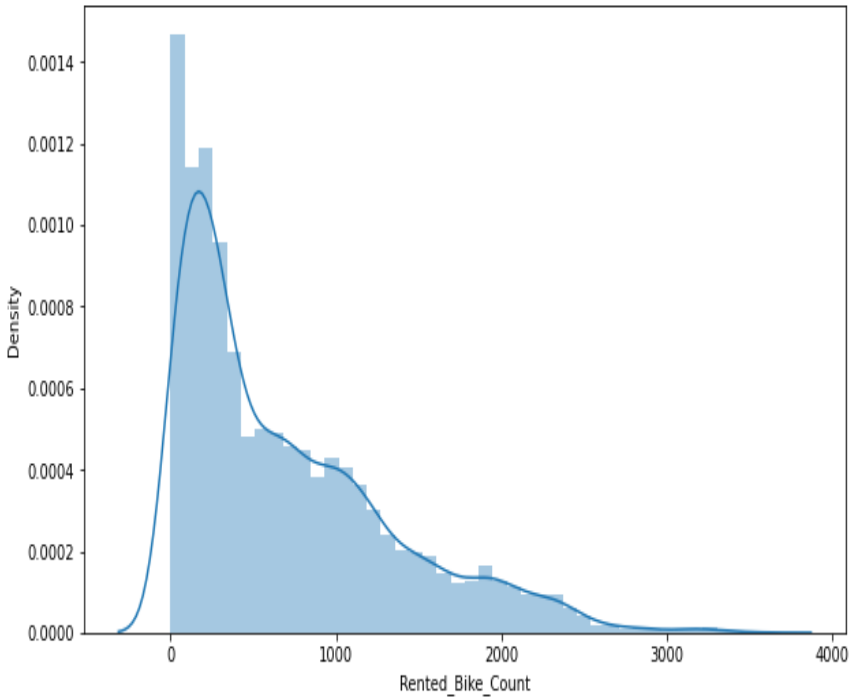
Analysis from Distributions

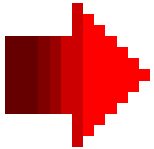
- “Temperature” and “Humidity” columns follows uniform distribution.

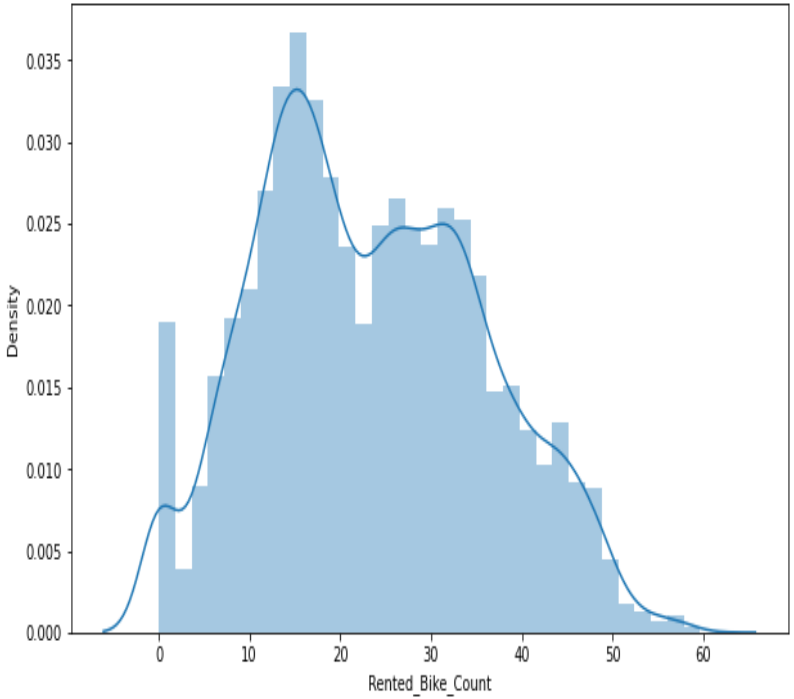
- “Wind Speed” , “Solar Radiation” , “Rainfall” and “Snowfall” are having positively skewed distribution.

- “Dew Point Temperature” and “Visibility” are negatively skewed.

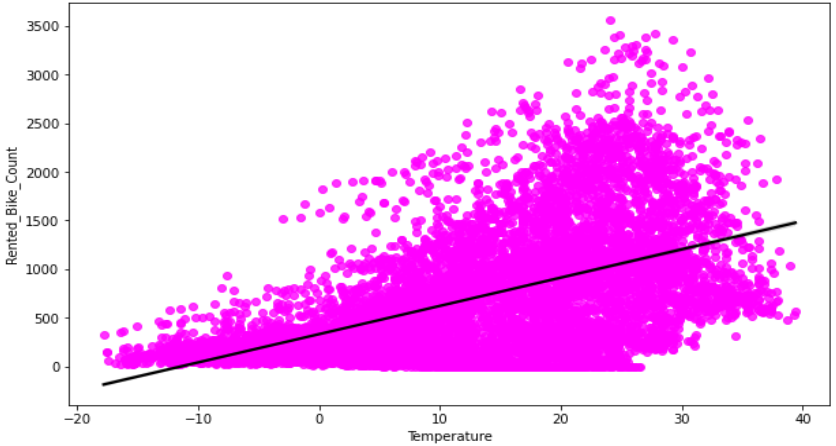
Distribution Of Dependent Variable

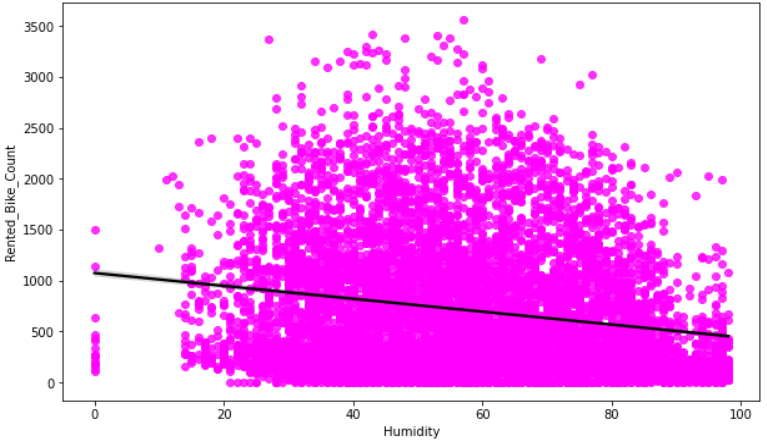


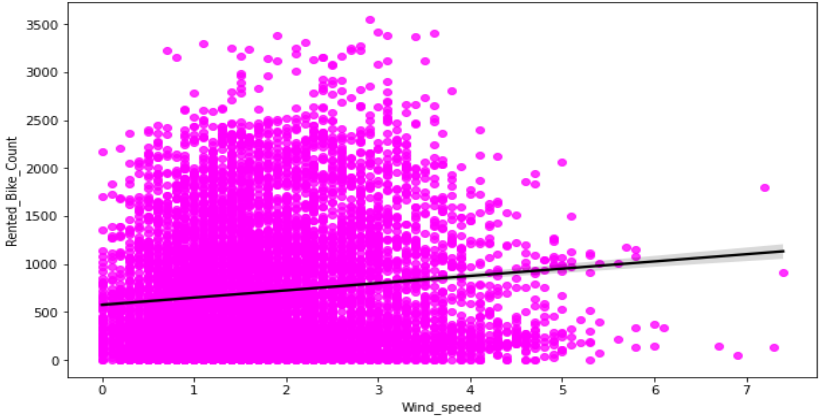


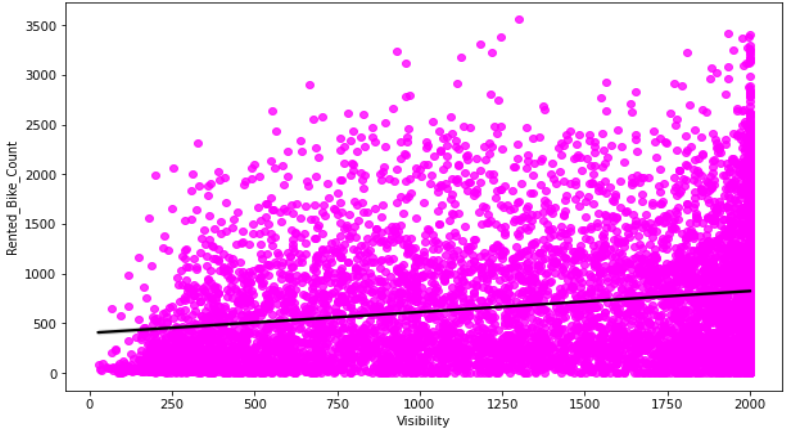


Data Preprocessing: Assumptions Check

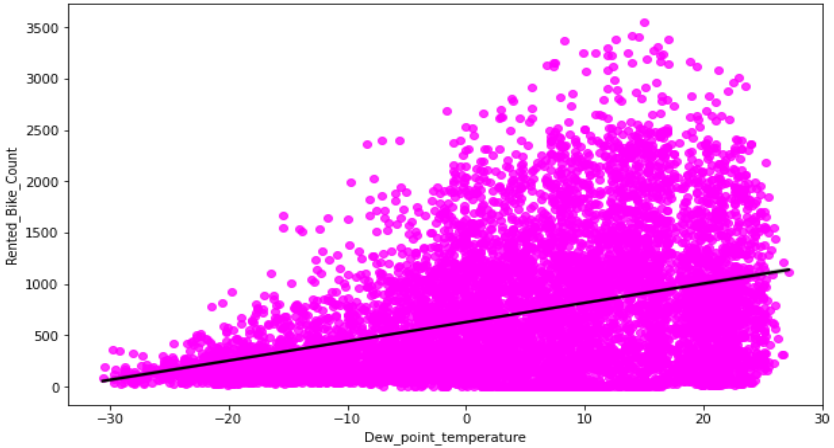


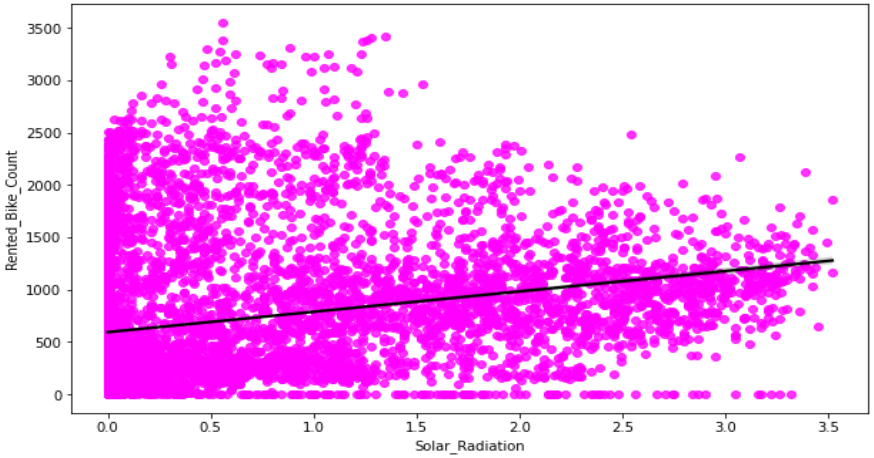


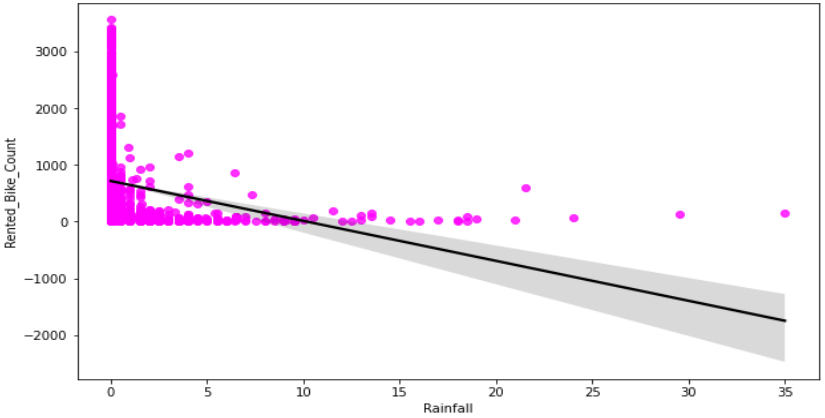


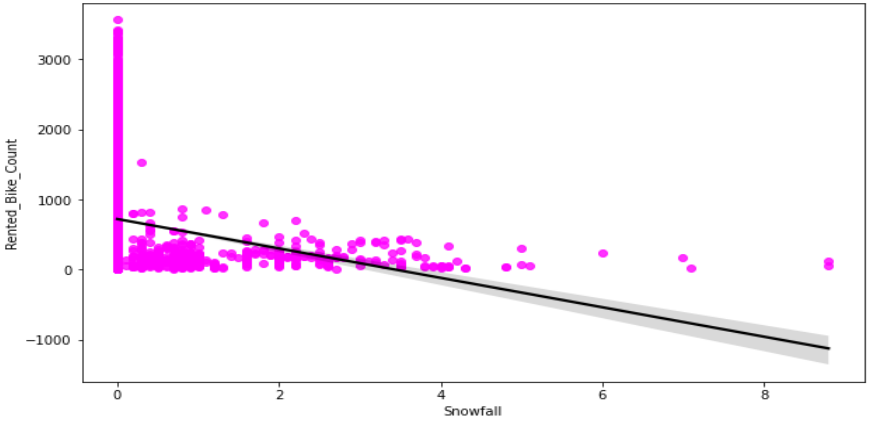


….Assumptions Check



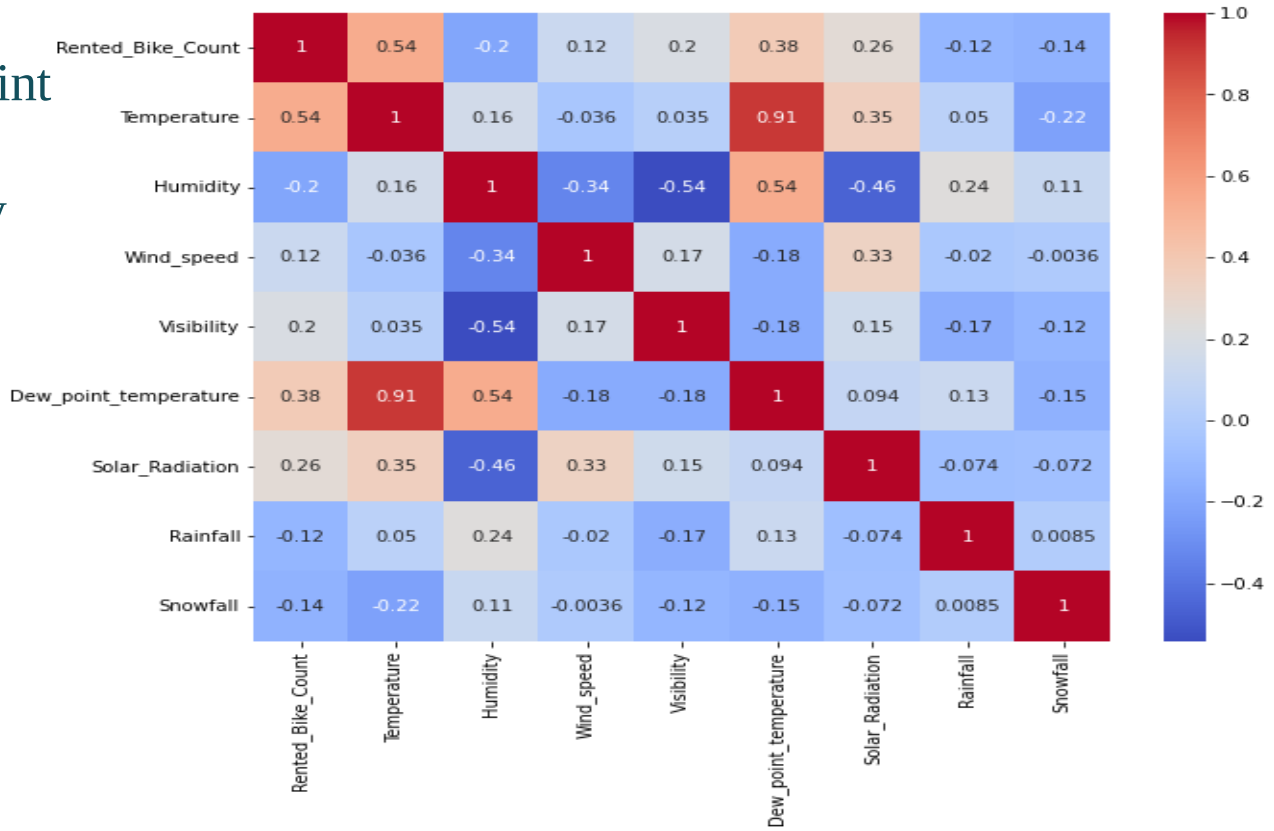






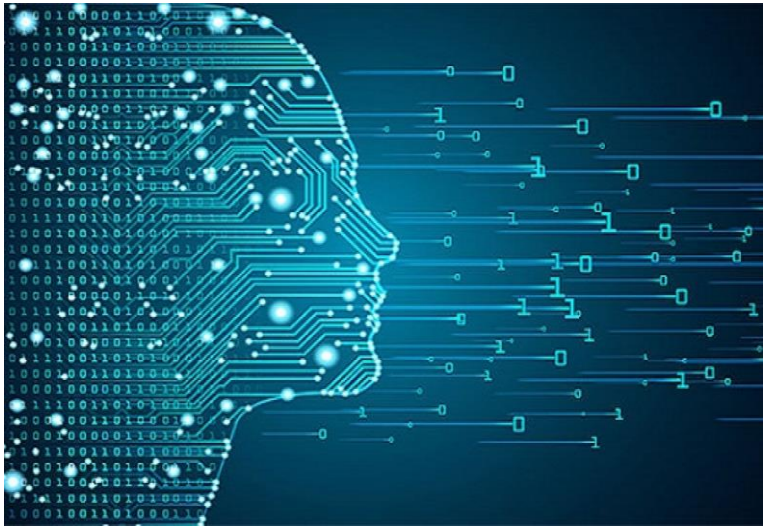
Multicollinearity

- Variables like Dew Point Temperature, and Temperature are highly correlated.



Model Building

- Linear regression model
- Lasso regression model
- Ridge regression model
- Elastic net regression model
- Decision tree regression model
- Random-forest regression model
- Gradient Boosting Regression model



Evaluation of models

- No overfitting is seen.
- Random forest Regressor gives the highest R2 score of 99% for Train Set and Gradient Boosting gridsearchcv gives the highest R2 score of 92% for Test set.

		Model	MAE	MSE	RMSE	R2_score	Adjusted R2
Training set	0	Linear regression	4.474	35.078	5.923	0.772	0.77
	1	Lasso regression	7.255	91.594	9.570	0.405	0.39
	2	Ridge regression	4.474	35.078	5.923	0.772	0.77
	3	Elastic net regression	5.792	57.574	7.588	0.626	0.62
	4	Dicision tree regression	5.025	47.213	6.871	0.693	0.69
	5	Random forest regression	0.800	1.552	1.246	0.990	0.99
	6	Gradient boosting regression	3.269	18.648	4.318	0.879	0.88
	7	Gradient Boosting gridsearchcv	1.849	7.455	2.730	0.952	0.95
Test set	0	Linear regression	4.410	33.275	5.768	0.789	0.78
	1	Lasso regression	7.456	96.775	9.837	0.387	0.37
	2	Ridge regression	4.410	33.277	5.769	0.789	0.78
	3	Elastic net regression Test	5.874	59.451	7.710	0.624	0.62
	4	Dicision tree regression	5.529	57.919	7.610	0.633	0.63
	5	Random forest regression	2.230	12.834	3.583	0.919	0.92
	6	Gradient boosting regression	3.493	18.648	4.318	0.865	0.86
	7	Gradient Boosting gridsearchcv	2.401	12.393	3.520	0.922	0.92

Challenges

- Feature engineering
- Feature selection
- Model Training and performance improvement

Conclusion

- No overfitting is seen.

- When we compare the root mean squared error and mean absolute error of all the models, the random forest regression model has less root mean squared error and mean absolute error, ending with the R-squared of 99% . So, finally this model is best for predicting the bike rental count on daily basis.

- For all the models, temperature or hour was ranked as the most influential variable to predict the rental bike demand at each hour.