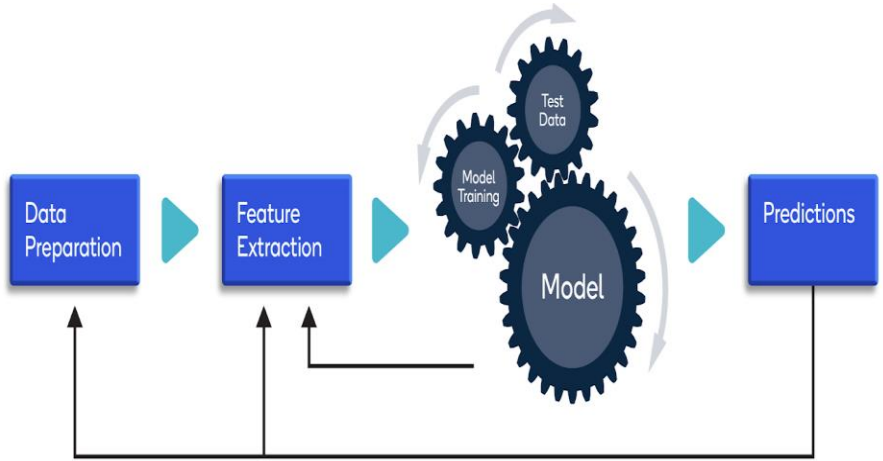
Capstone Project

Bike Sharing Demand Prediction

Content:

1. Problem Statement
2. Introduction
3. Data Summary
4. Hypothesis
5. Exploratory Data Analysis
6. Model Building
7. Evaluation
8. Challenges
9. Conclusion
10. Q&A



1.Problem Statement

Predict the demand of Rent Based Bike on historical usage over different factors such as seasons,weather,temperature,humidity etc. where there is hourly rent data for one year 2017-2018.

2. Introduction

- This bike sharing demand prediction is useful for companies who works in bike sharing business to allocate bikes better and ensure more sufficient circulation of bikes for customer.
- This prediction gives information about how many bikes required hourly which is helpful for company to distribute bikes wisely.
- This presentation proposes a real time method for predicting bike renting based on historical data, weather data and time data.
- Model is evaluated by comparing Root Mean Squared error with other models RMSE.

3.Data summary

- Date: year-month-day
- Rented Bike count :count bike rented per hour
- Hour : hour of day
- Temperature : Temperature in Celsius
- Humidity : Humidity %
- Wind speed : in m/s
- Solar Radiation : MJ.m2
- Rainfall :mm
- Snowfall :cm
- Seasons : Winter,Spring,Summer,Autumn
- Holiday : Holiday/Non Holiday
- Function Day : Yes/No
- Dew point temperature: in Celsius
- Visibility : 10m

4.Hypothesis

1. The dataset provides information about Hourly requirement of Rented bikes in one year (1 December 2017 to 31 November 2018).
2. There may be high demand on office hours from 8 am to 6 pm. And less demand during 10 pm to 4 am.
3. There is high demand on Non Holidays and low demand on Holidays.
4. Bike rented demand is positively correlated with temperature.
5. People mainly rent bikes on nice day and nice temperature.
6. There is zero demand on Non Functioning Day.

5.Exploratory Data Analysis

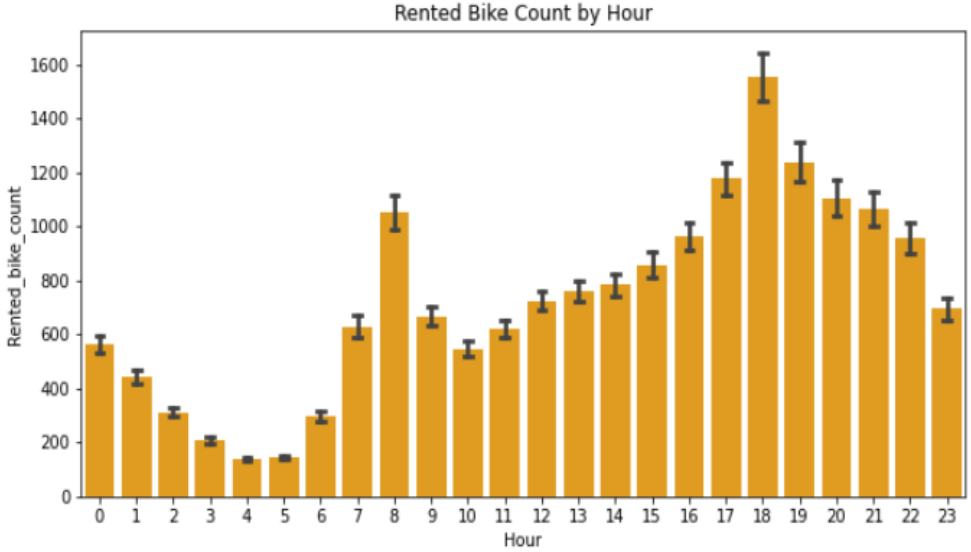
Overview of Dataset :

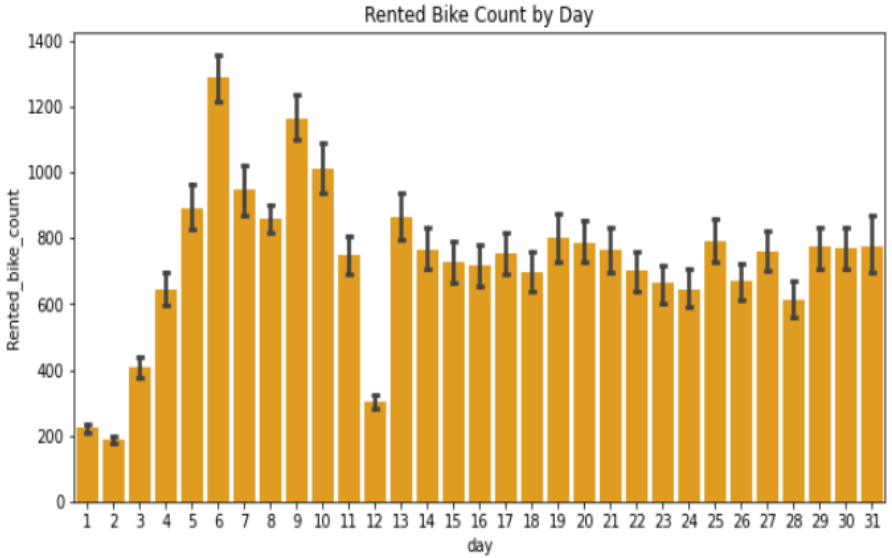
- Dataset consist of 14 columns and 8760 rows.
- In that Rented Bike count is Dependent Variable and remaining are independent variables.

```
df.head()
```

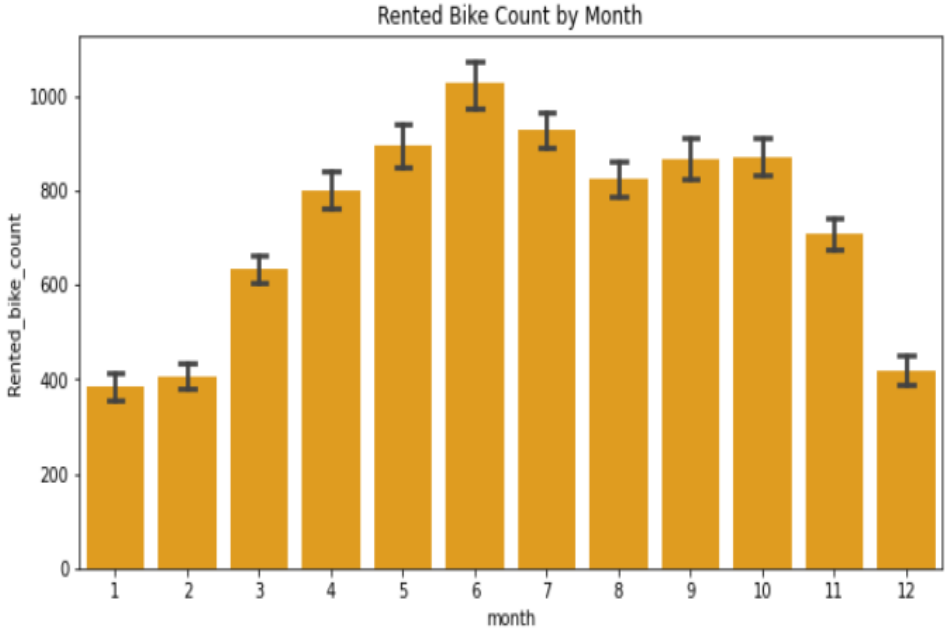
	Date	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Seasons	Holiday	Functioning Day
0	01/12/2017	254	0	-5.2	37	2.2	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
1	01/12/2017	204	1	-5.5	38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

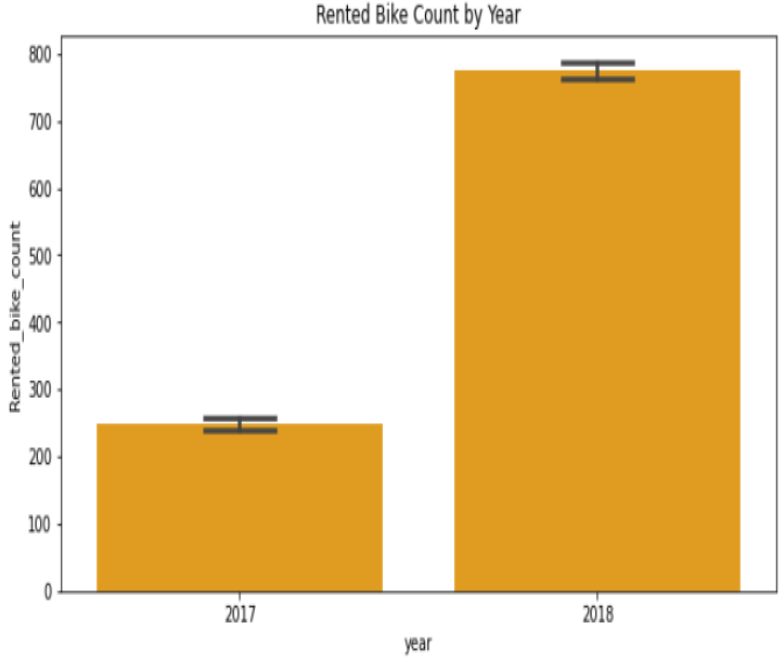
Date wise Trend





…Date wise Trend

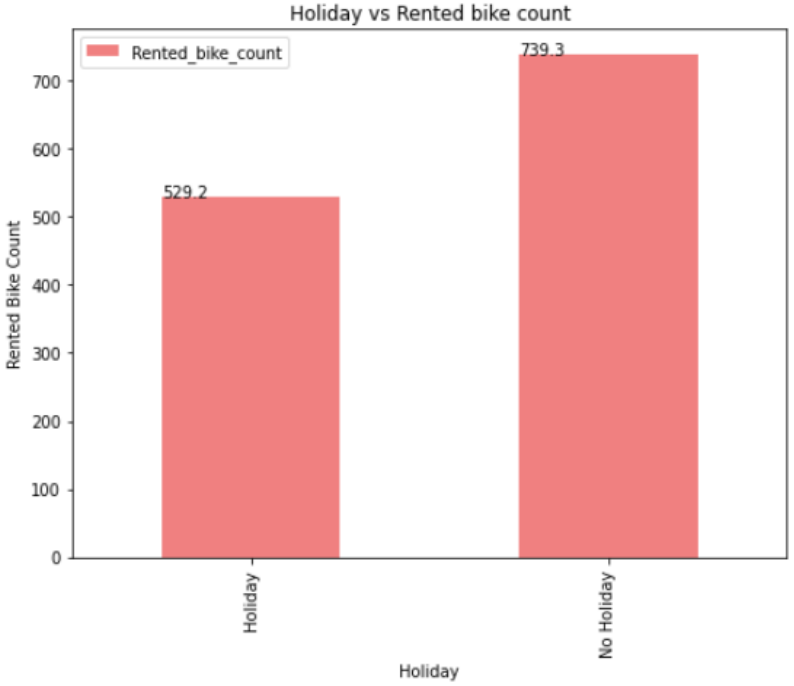


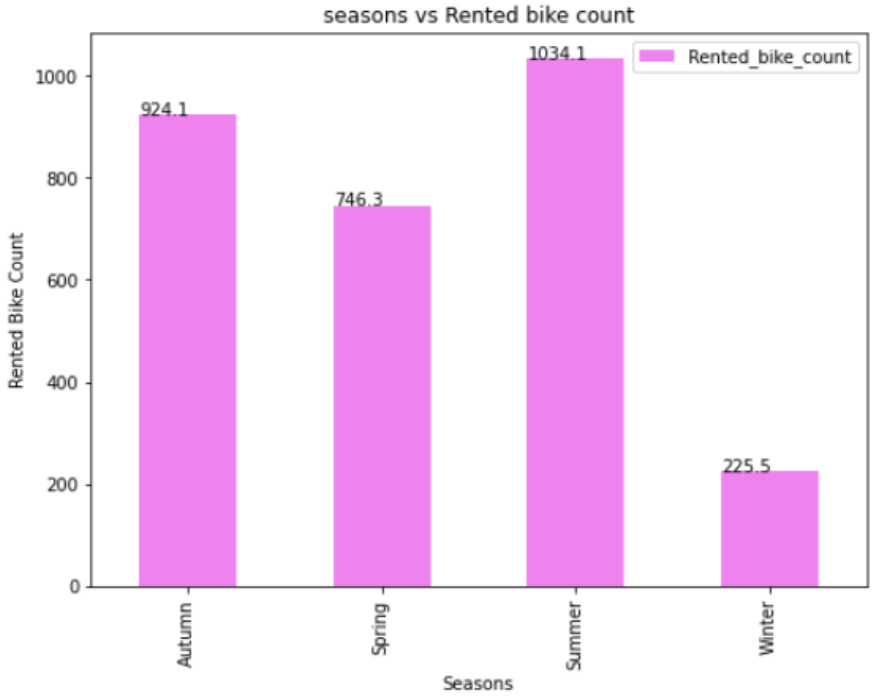


…Date wise Trend

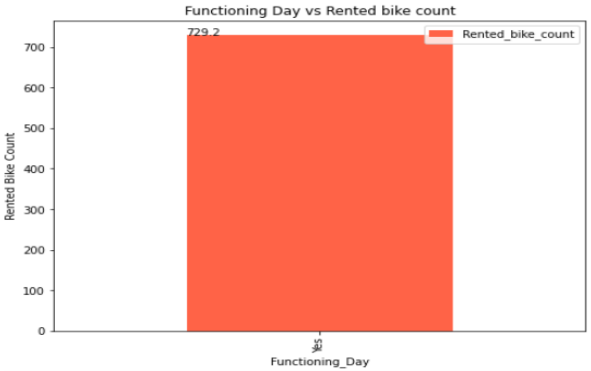
- In Morning Bike demand is low as compared to whole day. There is high demand between 3 pm to 10 pm.
- At evening 6 pm bike demand is highest in a day.
- At starting of month bike demand is low otherwise demand is same in all over month.
- At starting and ending of the year bike demand decreases. In mid year demand is high.

Categorical Variables



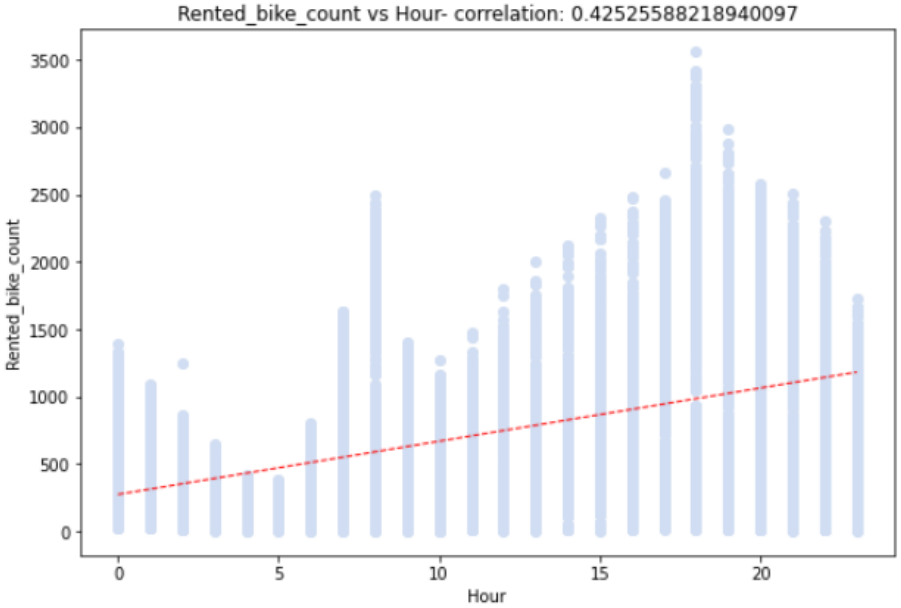


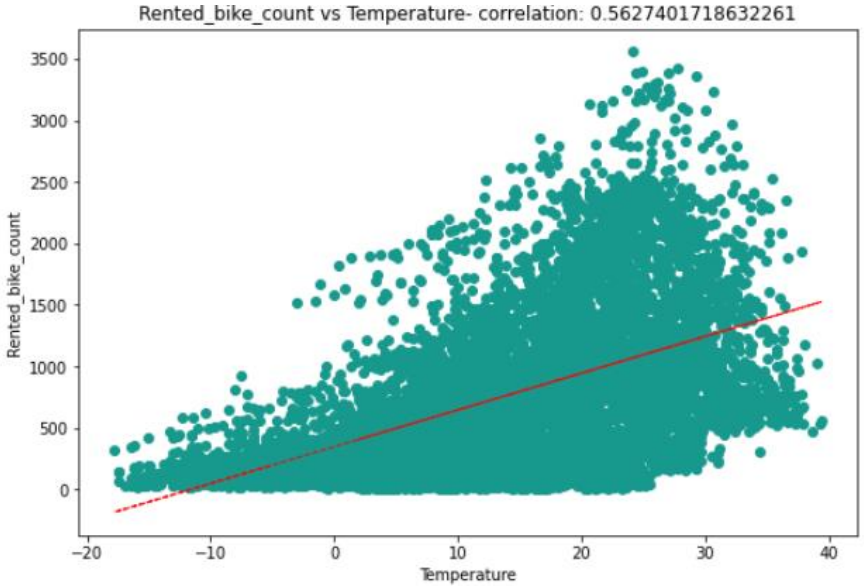
Categorical Variables



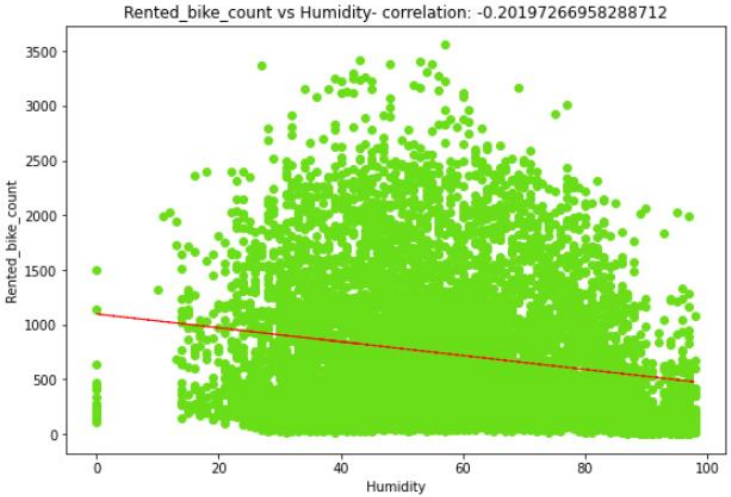
- On Holidays bike demand decreases lightly.
- In summer bike demand is highest and lowest in winter.
- In Autumn demand is relatively high as compared to spring.
- There is zero demand on Non functioning day.

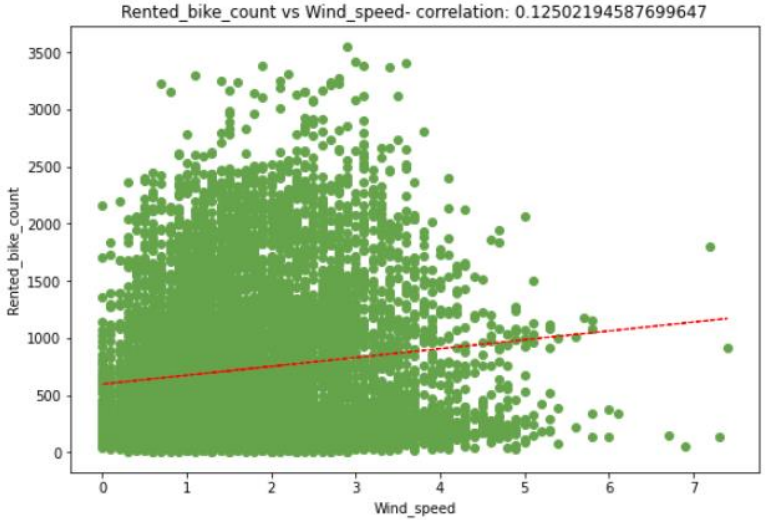
Data Processing Assumption check



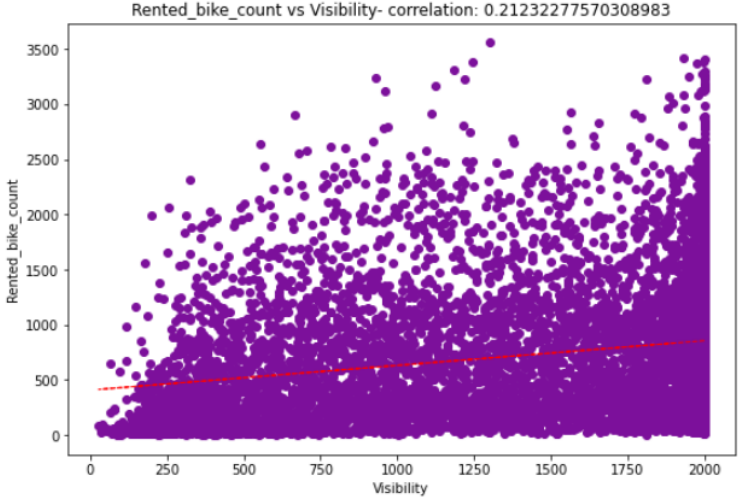


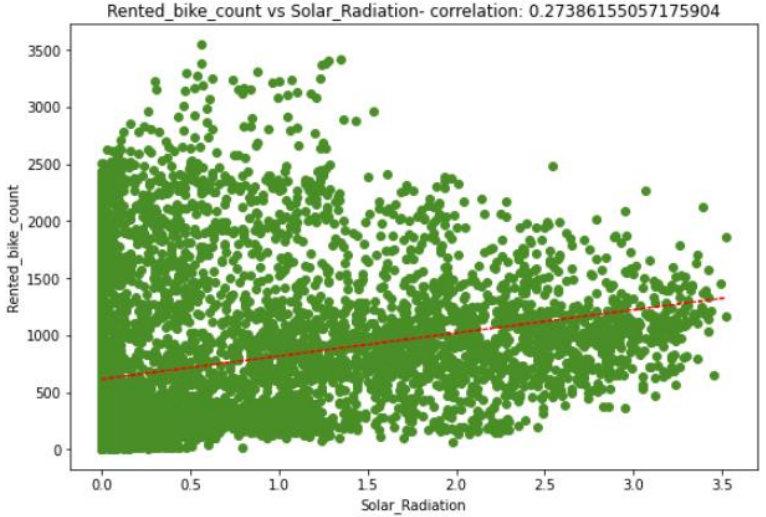
Data Processing Assumption check



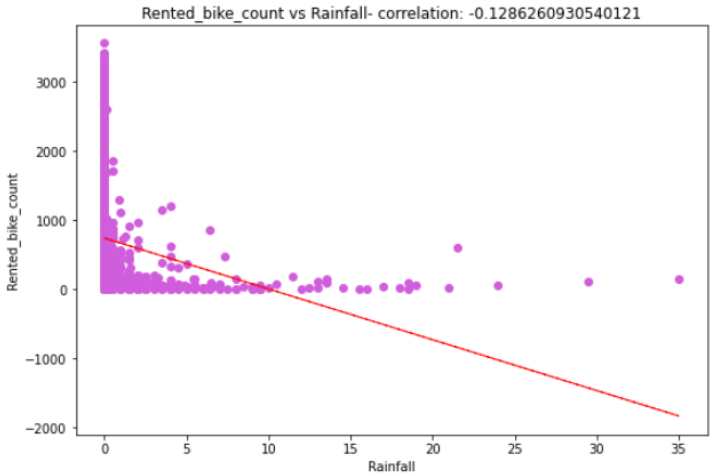


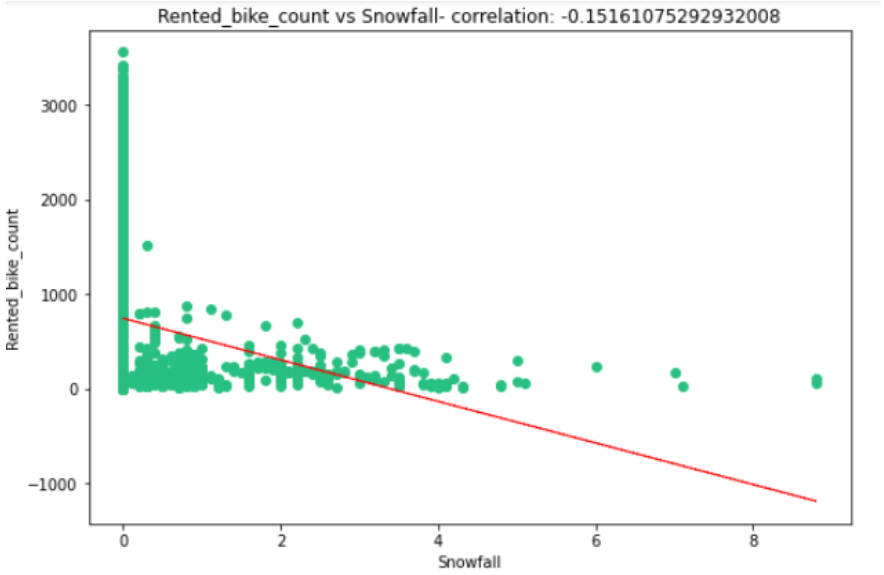
Data Processing Assumption check





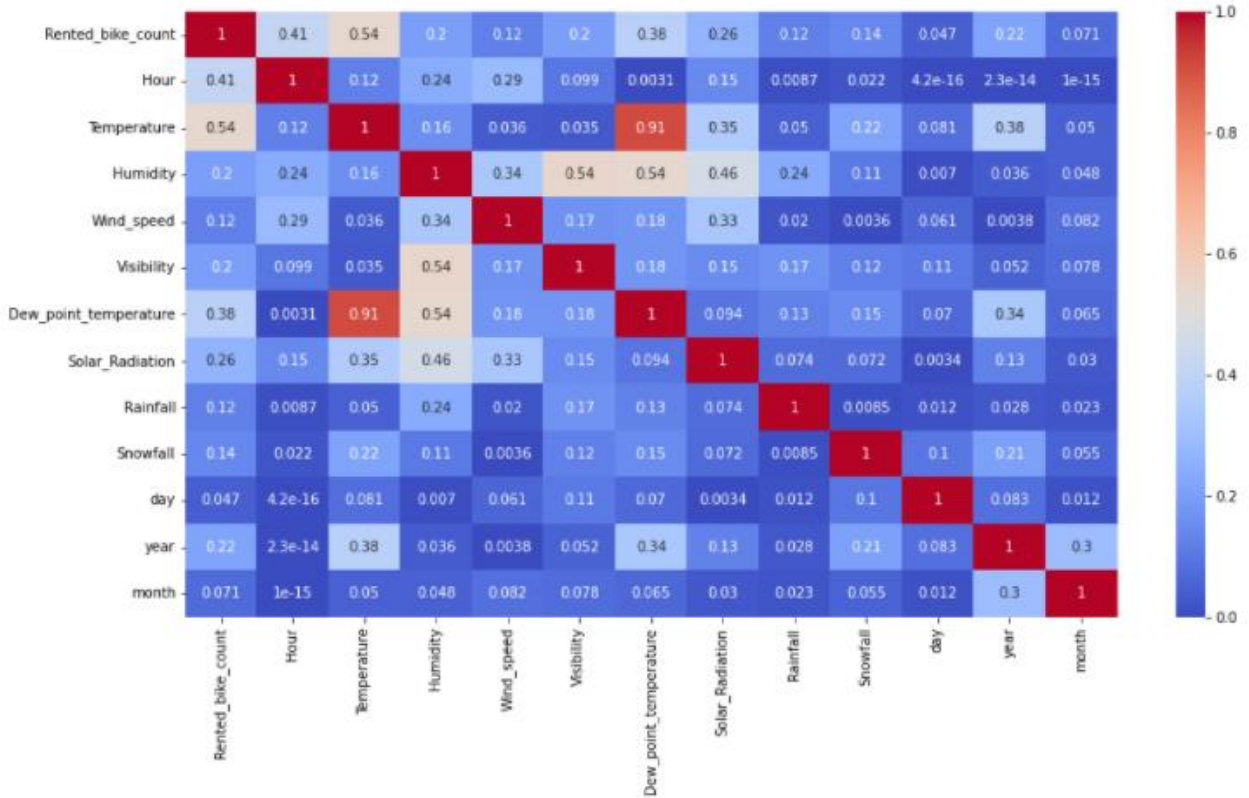
Data Processing Assumption check





Multicollinearity

- Variables like Dew point temperature and Temperature are highly correlated.

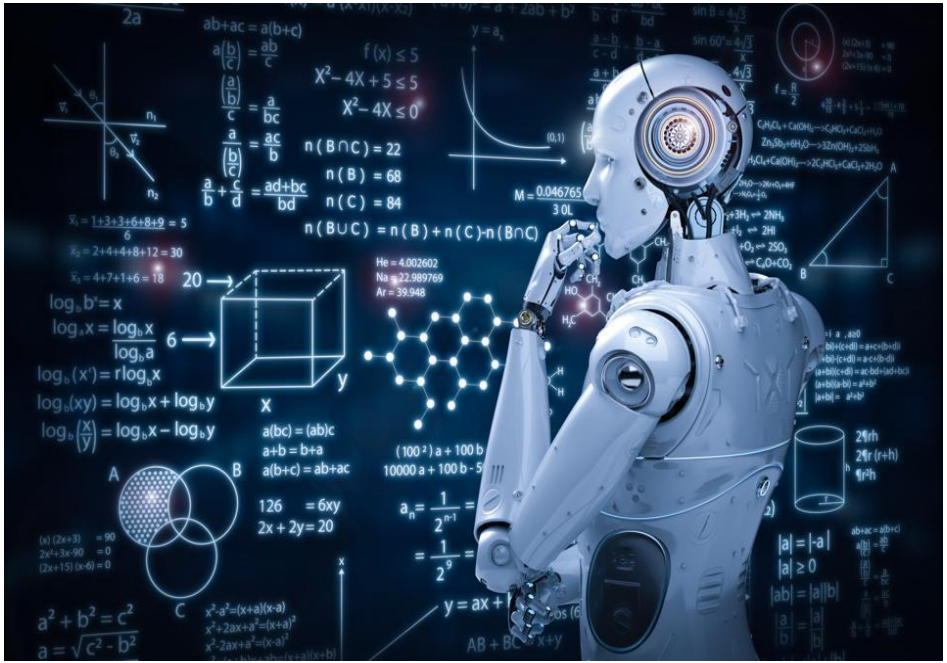


Data Pre-processing

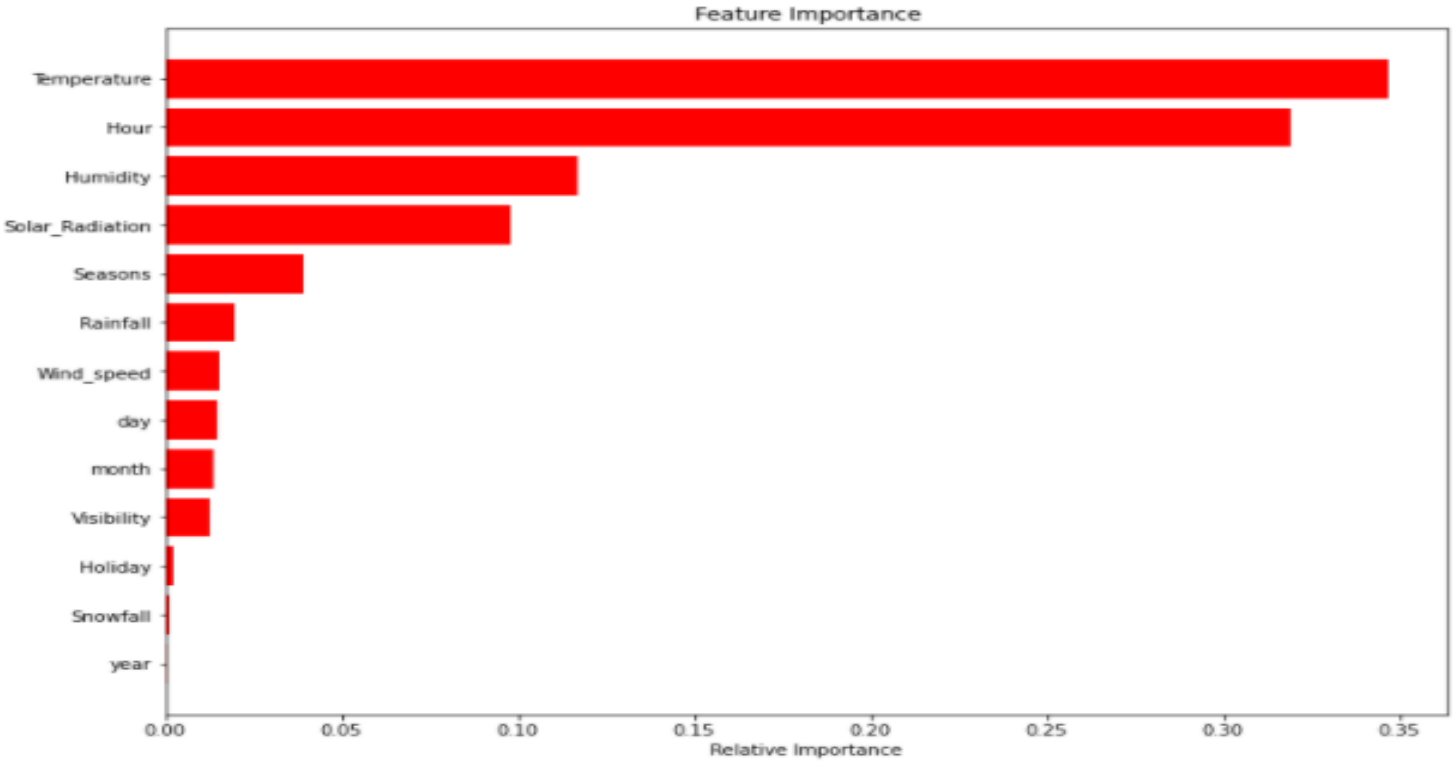
- On Non functioning day bike demand is zero so removed observations where it was ‘Nonfunctioning day’.
- As Dew point temperature and Temperature are highly correlated we drop ‘Dew point Temperature’ column.
- Also we have dropped Date column as it will not helping to give good prediction to model.

6.Model Building

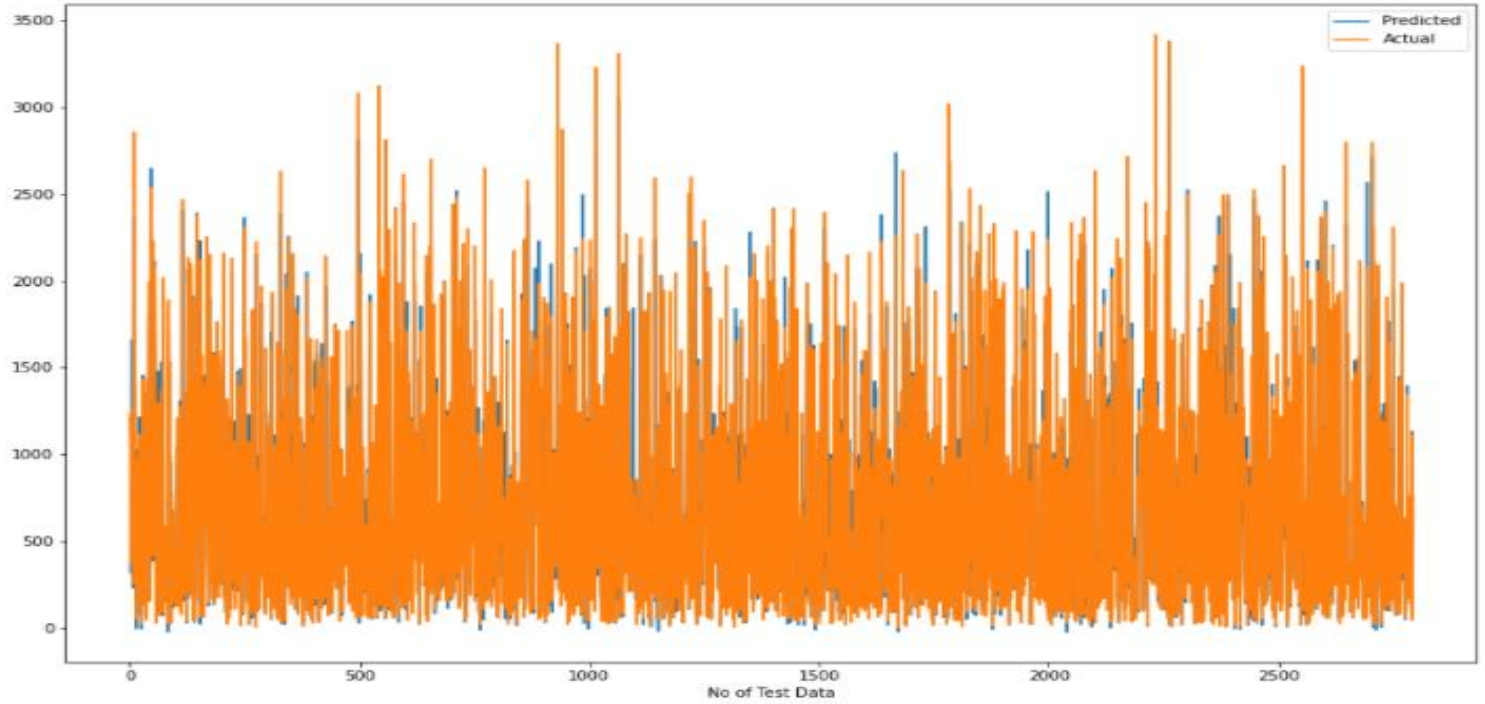
1. Linear Regression – Lasso, Ridge
2. Elastic Net
3. Decision Tree
4. Random Forest
5. Gradient Boosting
6. XGBoost



Feature Importance



Actual vs. Predicted



7.Evaluation

Model Name	r2 score	Adj r2score	MSE	RMSE	MAE
Linear Regression	0.511274	0.508253	197192.376070	444.063482	331.253420
Lasso Regression	0.511272	0.508251	197193.107677	444.064306	331.252991
Ridge Regression	0.511165	0.508143	197236.305917	444.112943	331.242356
Elasctic Net	0.509782	0.506752	197794.291337	444.740701	331.545728
Decision Tree	0.721158	0.719434	112507.837034	335.421879	191.725555
Random Forest	0.856134	0.855244	58047.504856	240.930498	145.181809
Random Forest Optimal	0.868217	0.867403	53171.925450	230.590385	144.585213
GradientBoosting	0.836202	0.835189	66089.644672	257.079063	170.897387
XGBoost	0.885368	0.884832	46519.012744	215.682667	129.518299
XGBoost Optimal	0.968070	0.967921	12957.512862	113.831072	68.452805

8.Challenges

- Feature engineering.
- Feature selection.
- Model Training and Performance Improvement.

9.Conclusion

- **The Rented Bike count is highly correlated with temperature i.e. in summer more no. of bikes get rented as compared to winter. Whereas it seems that the rentals are independent of the wind speed and the humidity, because they are almost constant over the months. So people mainly rent bikes on nice days and nice temperature.**
- **we use different algorithms to build model with high accuracy to predict count of rented bikes. so by comparing results from different algorithm we found that XGBoost algorithm gives best results.XGBoost has highest accuracy 96% and lowest RMSE,MAE with respect to other algorithms. So finally this model is best for predicting the bike rental count on daily basis.**

Q & A