CAPSTONE PROJECT

Bike Sharing Demand Prediction

<u>TEAM MEMBERS</u>

DEBABRATA SAHOO

VINAY VIJAY LANJEWAR

CONTENT

- BUSINESS UNDERSTANDING
- DATA SUMMARY
- FEATURE ANALYSIS
- EXPLORATORYDATAANALYSIS
- DATA PREPROCESSING
- IMPLEMENTING ALGORITHMS
- CHALLENGES
- CONCLUSIONS

BUSINESS UNDERSTANDING

- Bike rentals have became a popular service in recent years and it seems people are using it more often. With relatively cheaper rates and ease of pick up and drop at own convenience is what making this business thrive.
- Mostly used by people having no personal vehicles and also to avoid congested public transport which that’s why they prefer rental bikes.
- Therefore, the business to strive and profit more, it has to be always ready and supply no. of bikes at different locations, to fulfil the demand.
- Our project goal is a pre planned set of bike count values that can be a handy solution to meet all demands.

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
8755	30/11/2018	1003	19	4.2	34	2.6	1894	-10.3	0.0	0.0	0.0	Autumn	No Holiday	Yes
8756	30/11/2018	764	20	3.4	37	2.3	2000	-9.9	0.0	0.0	0.0	Autumn	No Holiday	Yes
8757	30/11/2018	694	21	2.6	39	0.3	1968	-9.9	0.0	0.0	0.0	Autumn	No Holiday	Yes
8758	30/11/2018	712	22	2.1	41	1.0	1859	-9.8	0.0	0.0	0.0	Autumn	No Holiday	Yes
8759	30/11/2018	584	23	1.9	43	1.3	1909	-9.3	0.0	0.0	0.0	Autumn	No Holiday	Yes

- **This Dataset contains 8760 lines and 14columns.**
- **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.**

INSIGHTS FROM OUR DATASET

- There are No Missing Values present
- There are No Duplicate values present
- There are No null values.
- And finally we have 'rented bike count' variable which we need to predict for new observations
- The dataset shows hourly rental data for one year (1December 2017 to 31 November(2018)(365 days).we consider this as a single year data
- 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'

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 Hrs),Fun(Functional Hrs)

Data Description

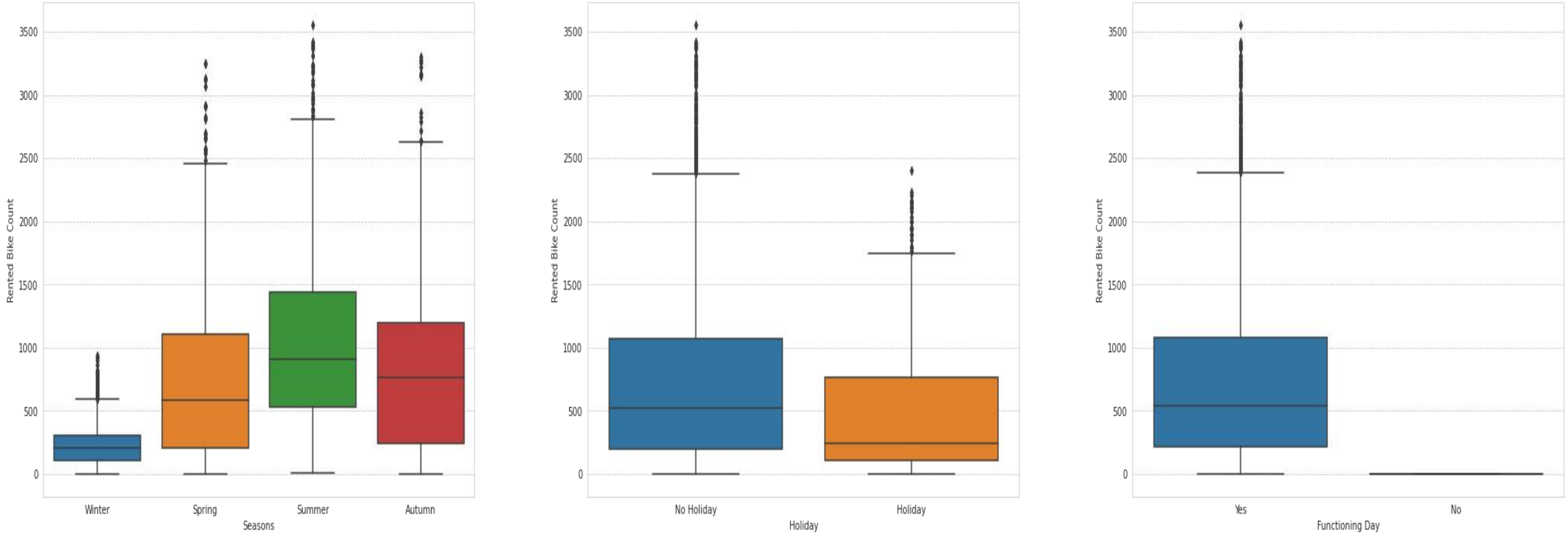
Dependent variable:

- Rented Bike count - Count of bikes rented at each hour

Independent variables:

- Date : year-month-day
- Hour - Hour of he day
- Temperature-Temperature in Celsius
- Humidity - %
- Windspeed - m/s
- Visibility - 10 m
- 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)

EDA (contd...)

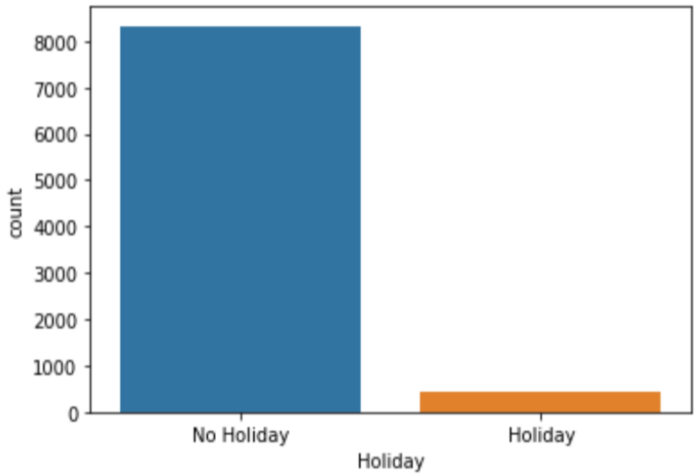


- Less demand on winter seasons
- Slightly Higher demand during Non holidays
- Almost no demand on non functioning day

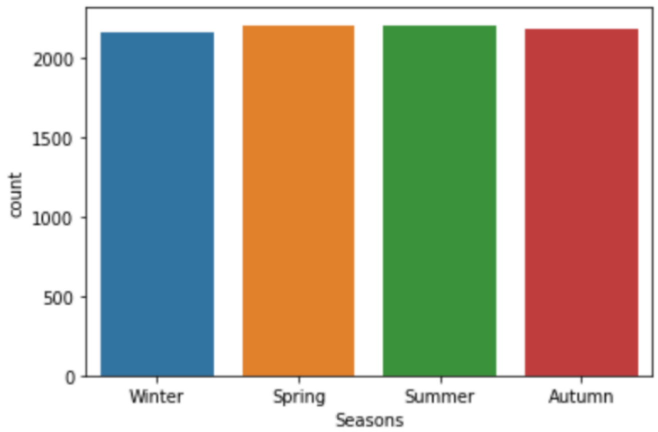
EXPLORATORY DATA ANALYSIS

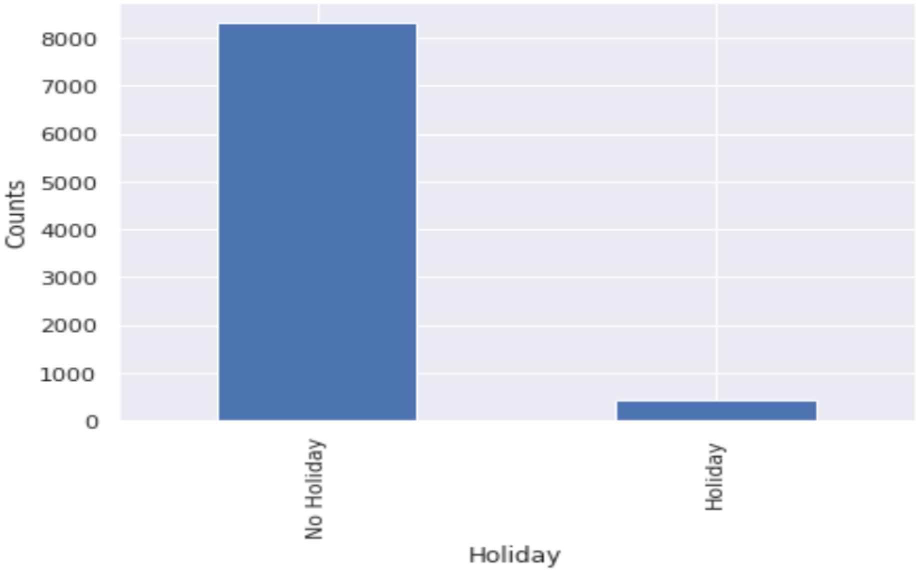
- We are given the data of one year which include many weather factors such as seasons , holiday etc.
- We can say that From the below data large number of bikes are being rented when there is a working day/No regardless of the seasons.

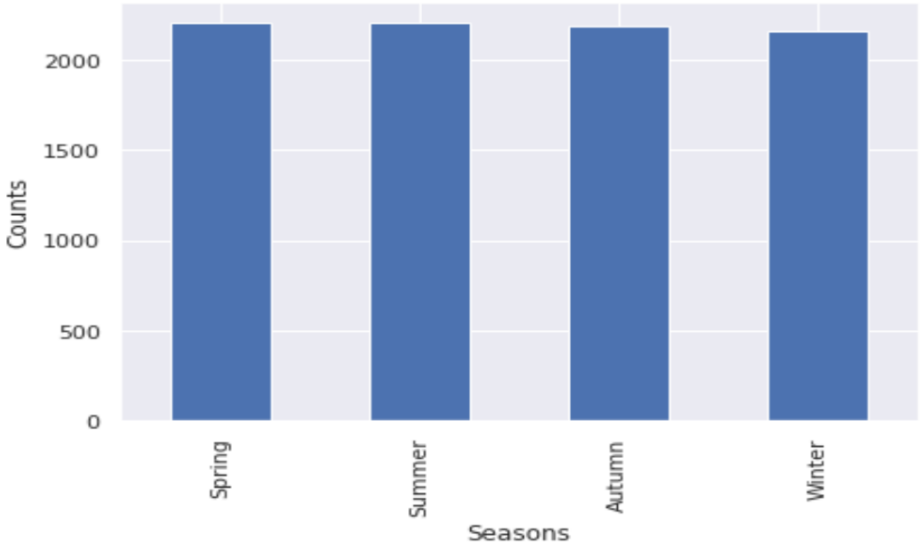
<matplotlib.axes._subplots.AxesSubplot at 0x7fc4f2d57cd0>



<matplotlib.axes._subplots.AxesSubplot at 0x7f3da72f28d0>

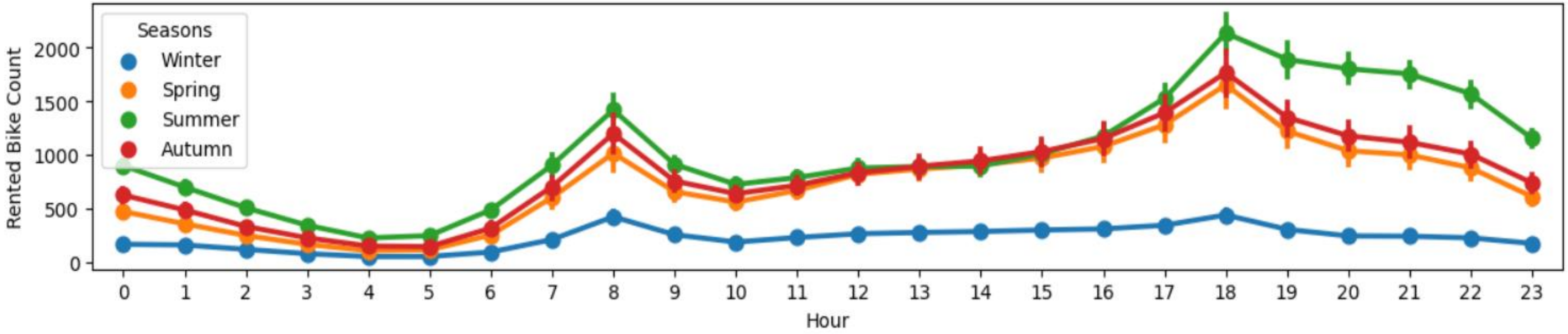




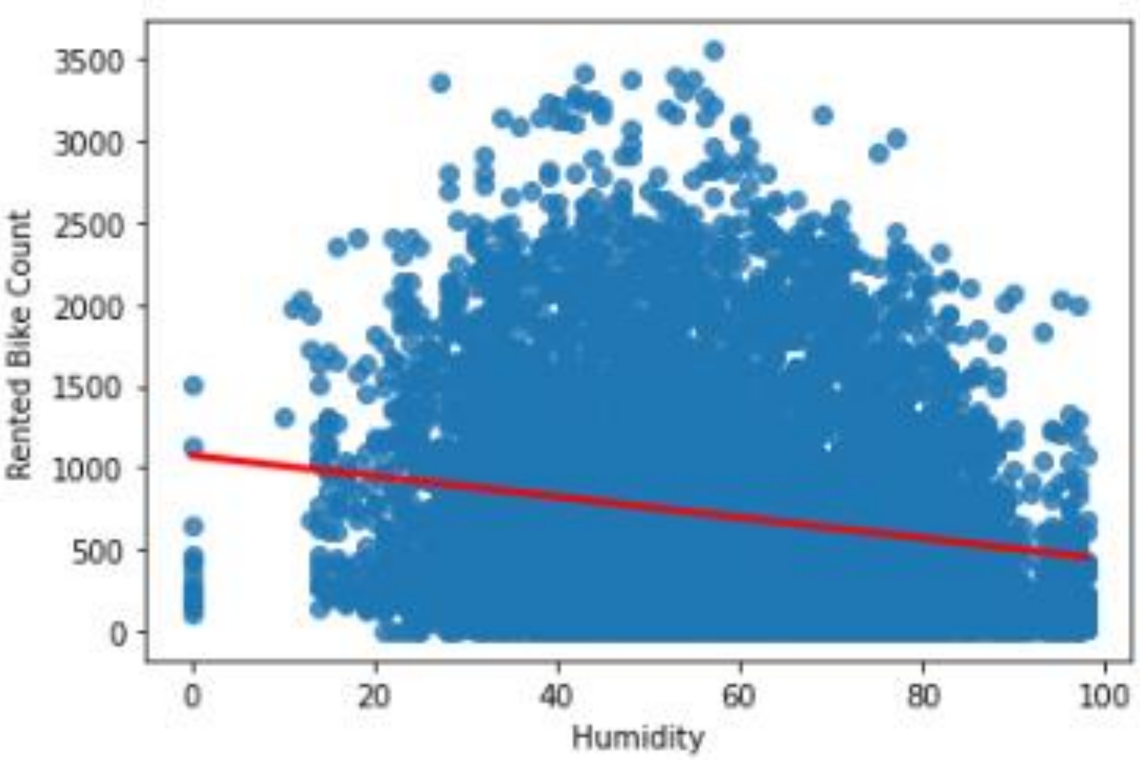


- **We are given the data of one year which include many weather factors such as seasons , humidity etc.**
- **From the above data, we observe that large number of bikes are being rented when there is a working day/No Holiday and more often in summer season. Even in general also, bikes are being rented more in the working day itself regardless of the seasons.**

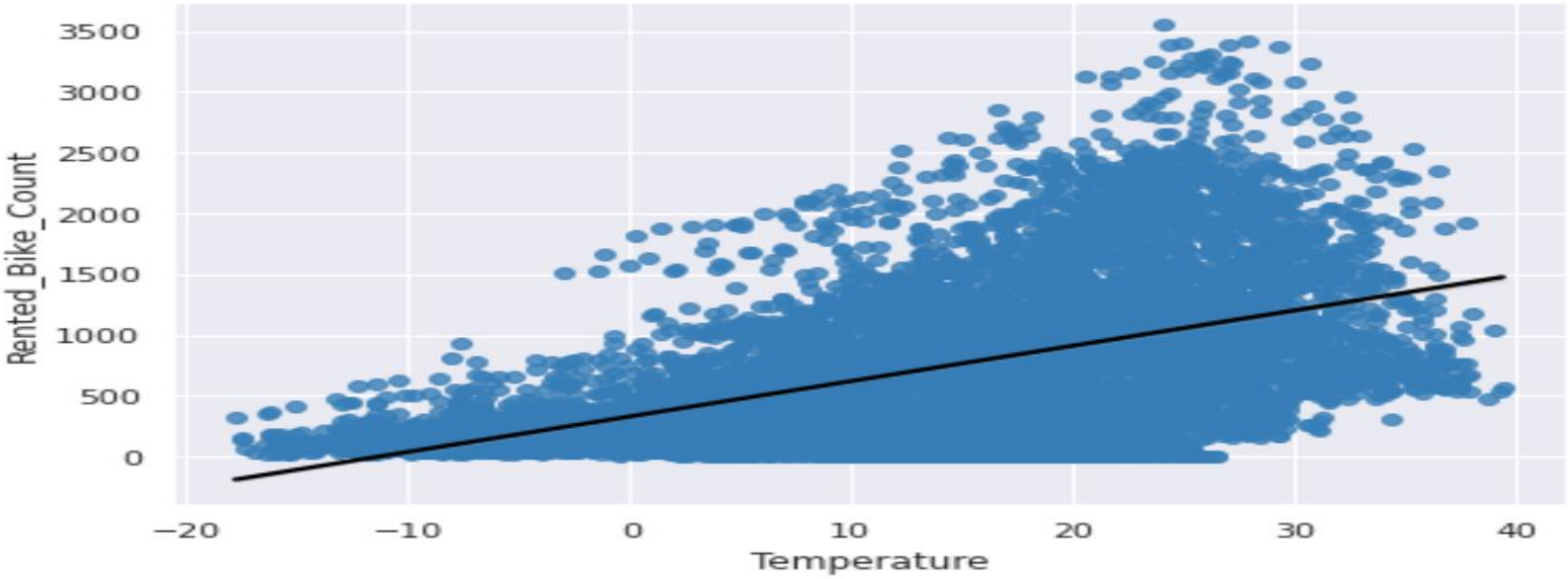
Hourly based bike counts with season



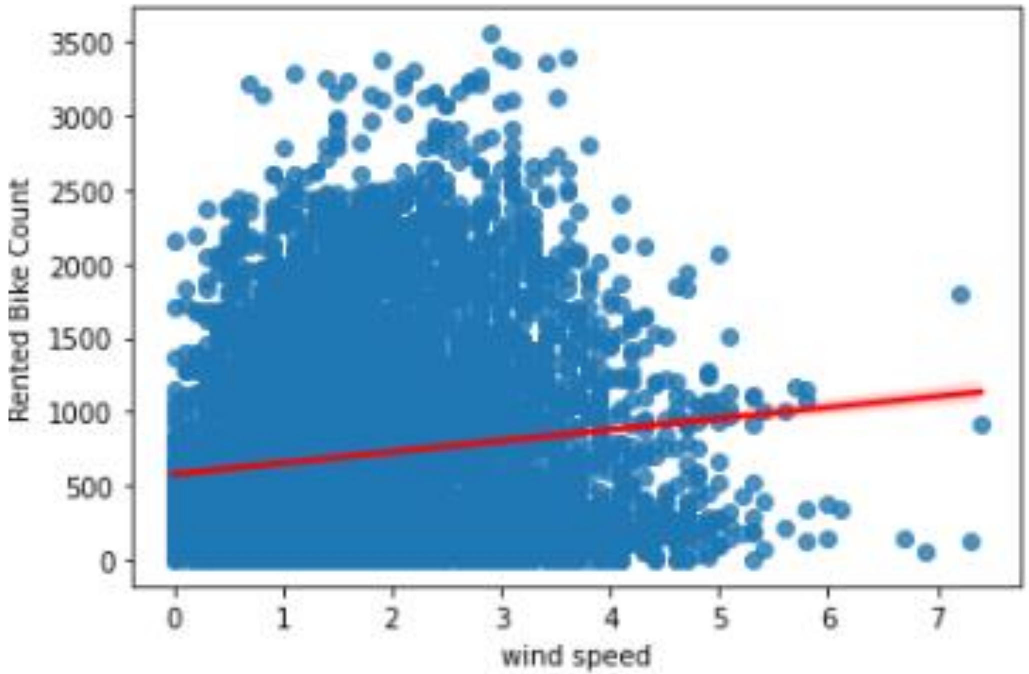
- **From the above subplot of the hourly based bike count we observed that during the period of 7 to 9 & 17 to 20 there is spike the bike count.**
- **Which clearly indicates that regardless the season people use the bike to travel to work place.**
- **In case of the winter season due to heavy snowfall the bike count decreses.**



➢ **Humidity acts as a deterrent to a bike ride. The bike count decreases when the humidity increases**.

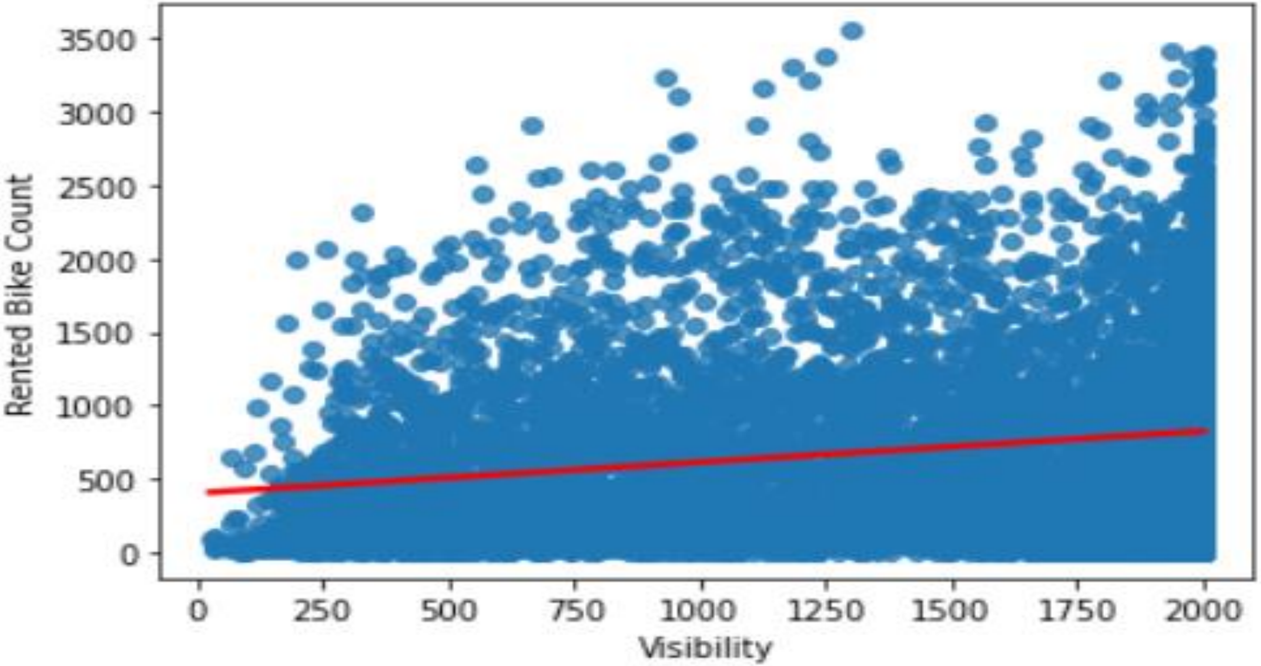


➢ **In general, temperature has negative correlation with the bike demands. So, as the temperature increases, the bike count also increases.**



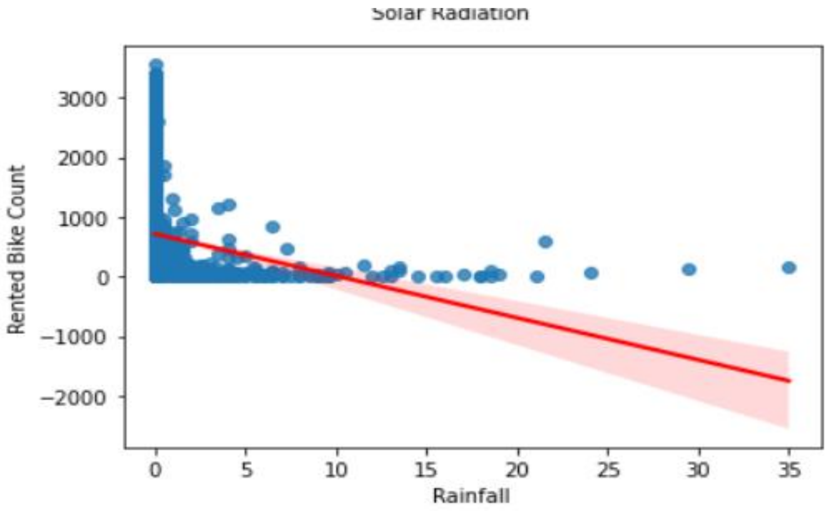
Wind Speed:

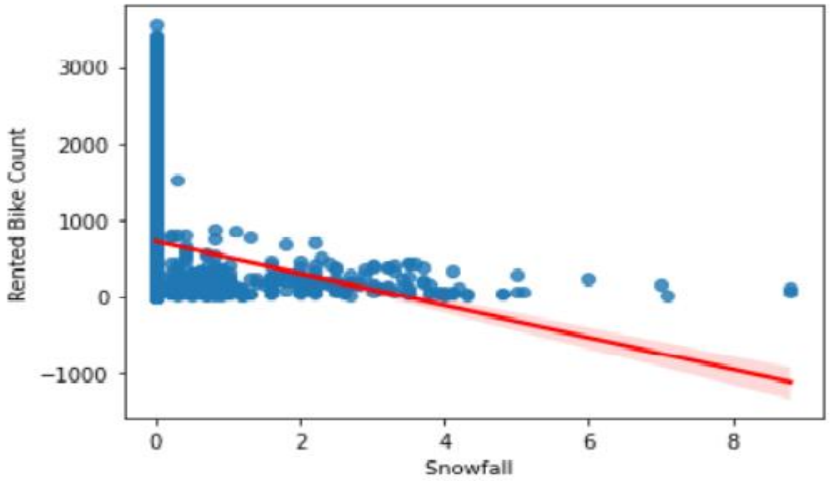
- **Due to Wind speed , there is certain increase in the bike count but the change is very small**.



Visibility:

- **If there is low visibility, people won't prefer to ride the bike. So,as the visibility increases , the number of bike count also increases.**

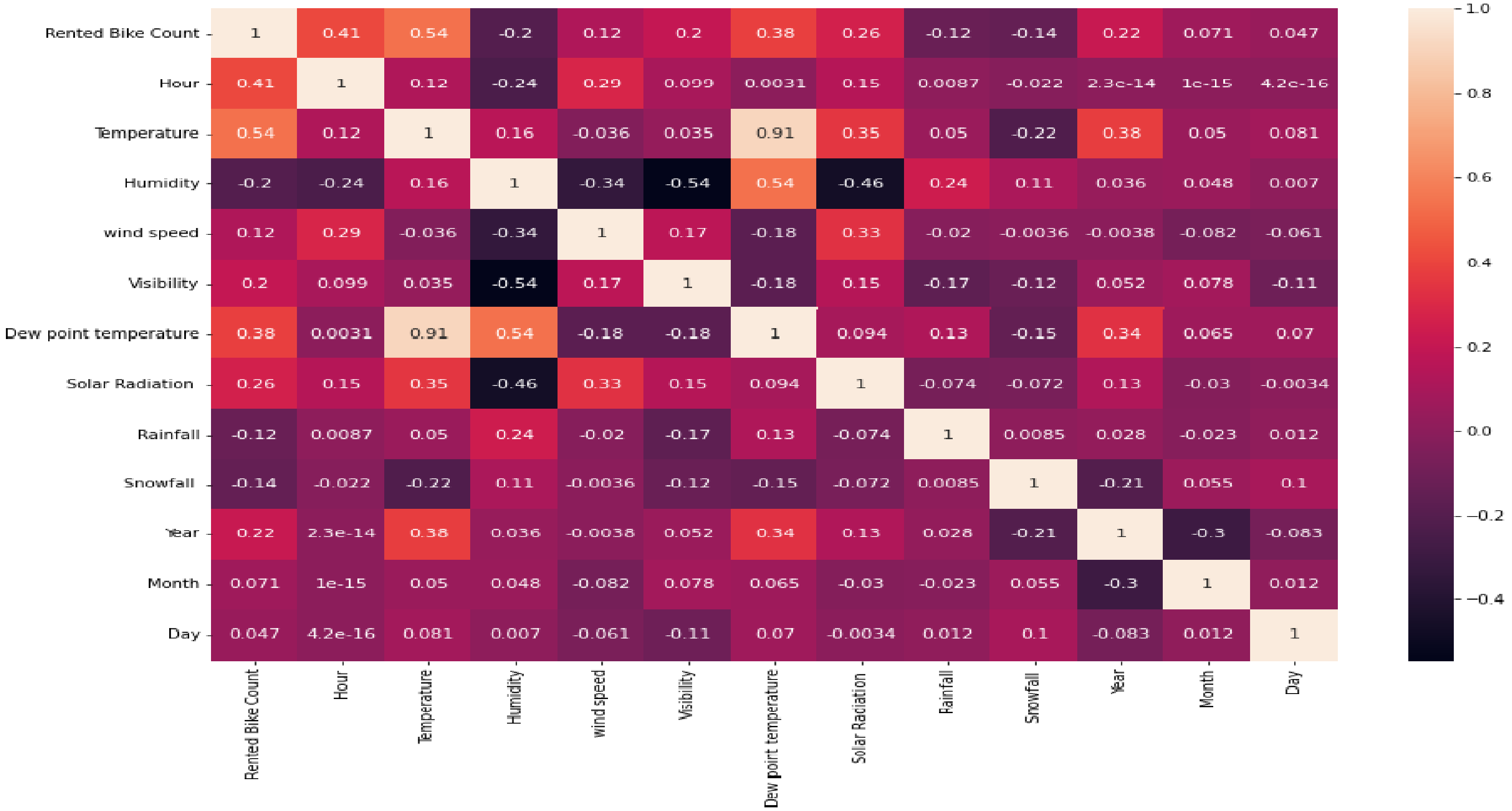




Rainfall and Snowfall:

➢ **If there is rainfall/Snowfall, people don't prefer to travel out. And, hence the bike count decreases.**

Correlation heatmap



REGRESSION PLOT FOR NUMERICAL VARIABLE

- From the above regression plot of all numerical features we see that the columns 'Temperature', 'Wind_speed','Visibility', 'Dew_point_temperature', 'Solar_Radiation' are positively relation to the target variable.
- which means the rented bike count increases with increase of these features.
- 'Rainfall','Snowfall','Humidity' these features are negatively related with the target variable which means the rented bike count decreases when these features increase.

Numerical Feature vs Rented Bike Count

- **Observations from above plot:-**

- **As the visibility & Temperature increase the Bike Count increases , which shows that they have positive correlation w.r.t. target variable rented bike count.**
- **And as the Rainfall & snowfall increase the bike count decreases , which shows that they have negatively correlation w.r.t. target variable rented bike count.**

MODEL BUILDING

- LINEAR REGRESSION
- LASSO REGRESSION
- RIDGE REGRESSION
- DECISION TREES REGRESSOR
- RANDOM FOREST REGRESSOR
- GRADIENT BOOSTED REGRESSOR
- GRADIENT BOOSTING REGRESSOR WITH GRIDSEARCHCV

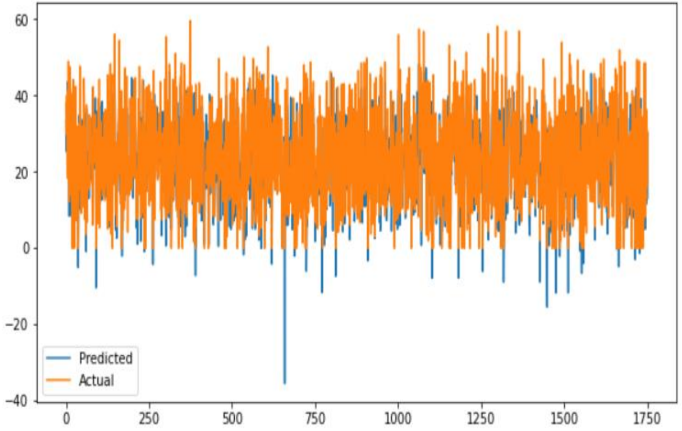
Linear Regression

Train Set Metrics

MSE : 37.0311260552138
RMSE : 6.08532053841158
R2 : 0.7603855187296256
Adjusted R2 : 0.7536318516121987

Test Set Metrics

MSE : 37.882593399150124
RMSE : 6.154883703137706
R2 : 0.7584668027178505
Adjusted R2 : 0.7516590555249303



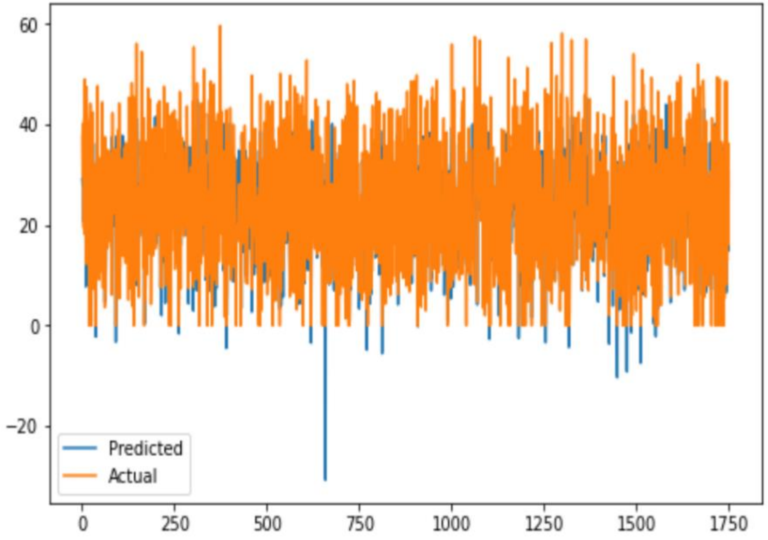
Lasso Regression

Train Set Metrics

```
MSE : 43.36534936956268
RMSE : 6.5852372295584525
R2 : 0.7193991433367898
Adjusted R2 : 0.7114902524854485
```

Test Set Metrics

```
MSE : 43.4651785011789
RMSE : 6.592812639623463
R2 : 0.722873156459647
Adjusted R2 : 0.7150621825959143
```



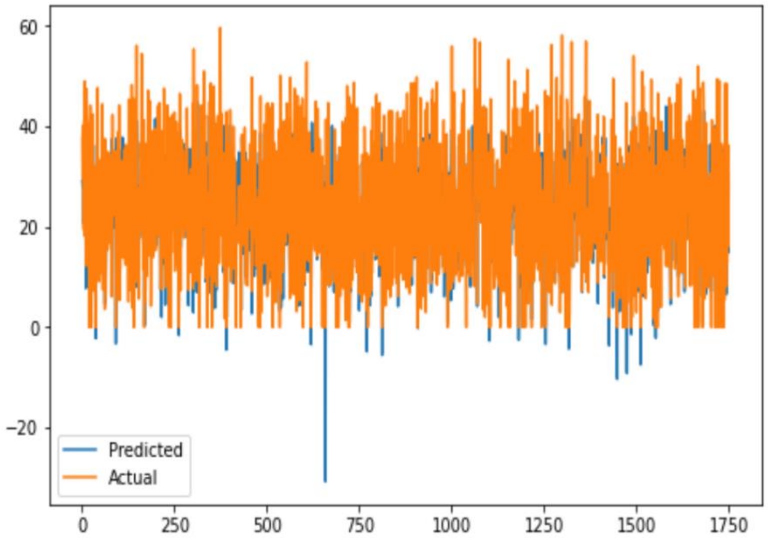
Ridge Regression

Train Set Metrics

```
MSE : 37.0311377420355
RMSE : 6.085321498658513
R2 : 0.7603854431086003
Adjusted R2 : 0.7536317738597529
```

Test Set Metrics

```
MSE : 37.882268046900464
RMSE : 6.154857272666886
R2 : 0.7584688771103895
Adjusted R2 : 0.751661188385374
```



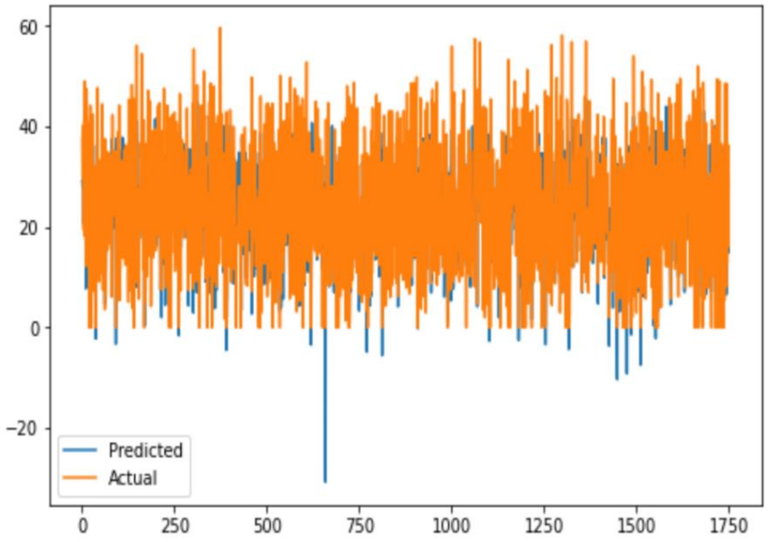
ElasticNet Regression

Train Set Metrics

```
MSE : 59.33563528401667
RMSE : 7.702962760134354
R2 : 0.6160614330704103
Adjusted R2 : 0.6052399115127941
```

Test Set Metrics

```
MSE : 59.61213050635593
RMSE : 7.720889230286621
R2 : 0.6199228409128946
Adjusted R2 : 0.60921015527802
```



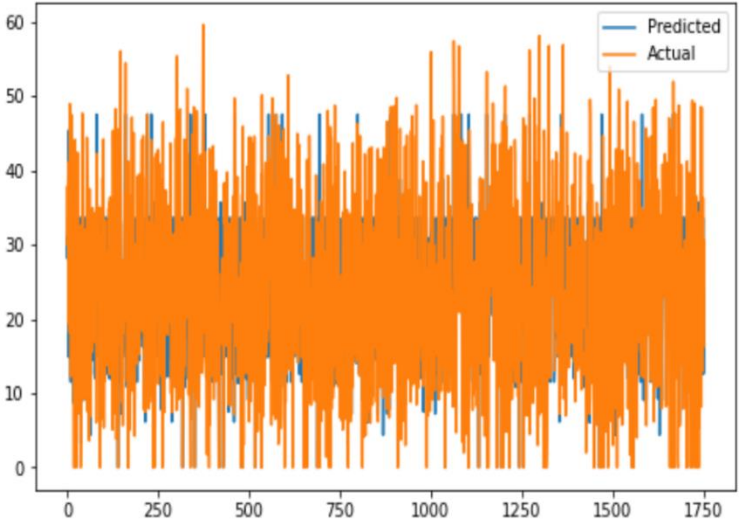
Decision Tree

Train Set Metrics

```
Model Score: 0.6926505638240831
MSE : 47.49919810224951
RMSE : 6.891966200022277
R2 : 0.6926505638240831
Adjusted R2 : 0.6839877494163062
```

Test Set Metrics

```
MSE : 55.98283569878599
RMSE : 7.482167847541646
R2 : 0.6430626288760792
Adjusted R2 : 0.6330021510052934
```



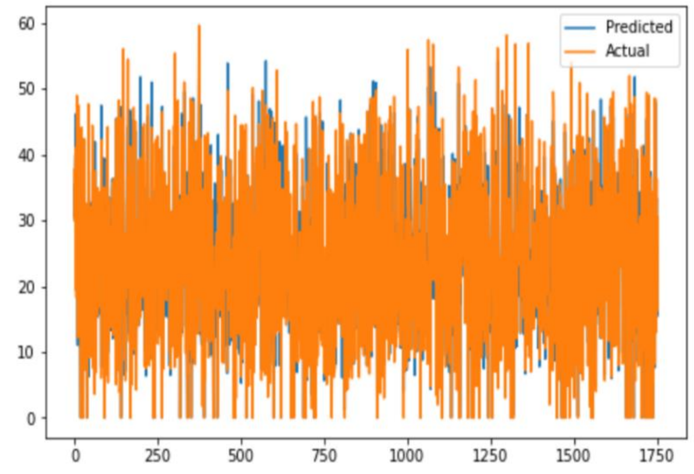
RANDOM FORREST REGRESSION

Train Set Metrics

```
Model Score: 0.8627849600092803
MSE : 21.205844553418405
RMSE : 4.604980407495606
R2 : 0.8627849600092803
Adjusted R2 : 0.8589174779660891
```

Test Set Metrics

```
Model Score: 0.8627849600092803
MSE : 24.45482735319885
RMSE : 4.945182236601483
R2 : 0.8440800349288231
Adjusted R2 : 0.839685344192818
```



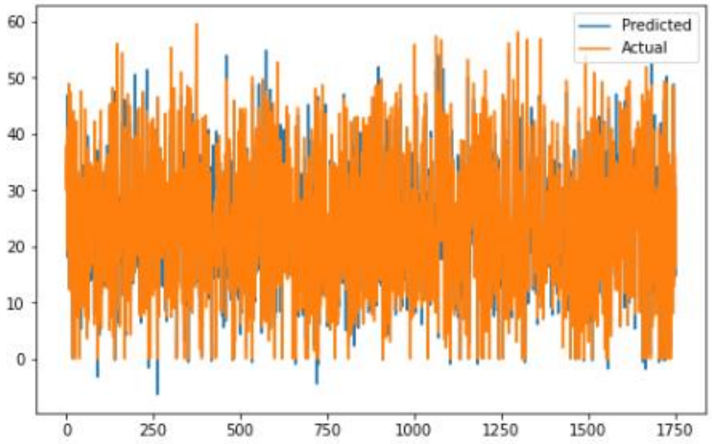
XGBOOST REGRESSION

Train Set Metrics

```
Model Score: 0.9674298955235844
MSE : 5.033534025586255
RMSE : 2.2435538829246457
R2 : 0.9674298955235844
Adjusted R2 : 0.9665118890556643
```

Test Set Metrics

```
Model Score: 0.9674298955235844
MSE : 17.314957566633634
RMSE : 4.161124555298675
R2 : 0.8896026727154431
Adjusted R2 : 0.8864910627861073
```



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 R2score.

CONCLUSION

- ‘Hour’ of the day holds the most important feature.
- Bike rental count is mostly correlated with the time of the day as it is peak at 10 am morning and 8 pm at evening.
- We observed that bike rental count is high during working days than non working day.
- The Rented bike Count has been increased from 2017-2018. No Overfitting is seen
- When we compare the root mean squared error and mean absolute error of all the models, the XGBoost model has less root mean squared error and mean absolute error, ending with the accuracy of 87% . So, finally this model is best for predicting the bike rental count on daily basis.

THANK YOU