Capstone Project

On

Seoul Bike Sharing Demand Prediction

By

Premanand Gaikwad
Raushan Kumar

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

Data Summary

- Bike sharing has been gaining importance over the last few decades. More and more people are turning to healthier and more livable cities where activities like bike sharing are easily available. there are many benefits from bike sharing, such as environmental benefits. It was a green way to travel

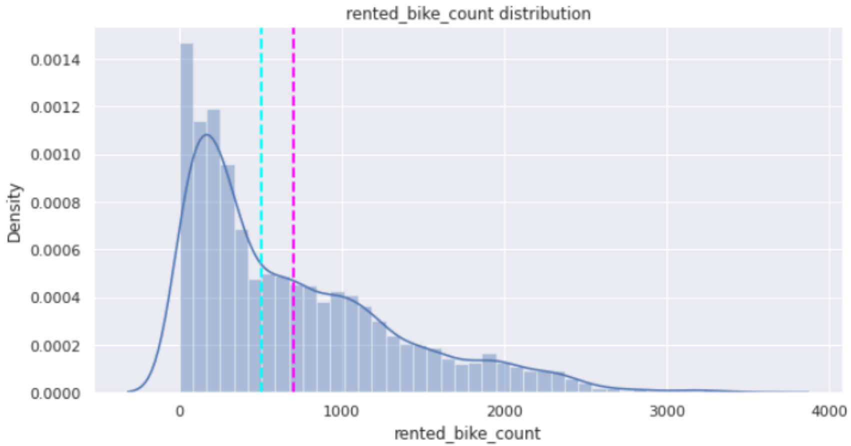
- The dataset contains weather information (Temperature, Humidity, Wind speed, Visibility, Dew point, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

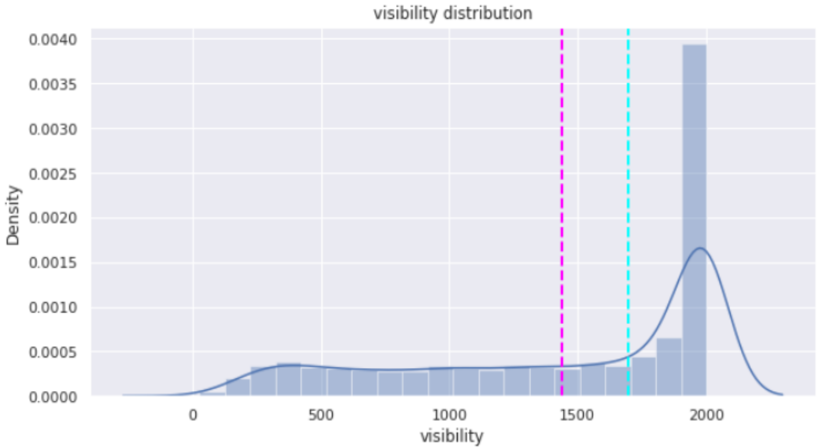
- This dataset contains the hourly and daily count of rental bikes between years 2017 and 2018 in Seoul bike share system with the corresponding weather and seasonal information. The dataset contains 8760 rows (every hour of each day for 2017 and 2018 i.e. 375days * 24 Hr) and 14 columns (the features which are under consideration).

Exploratory Data Analysis (EDA)

Univariate Analysis :

1. Distribution of numerical features

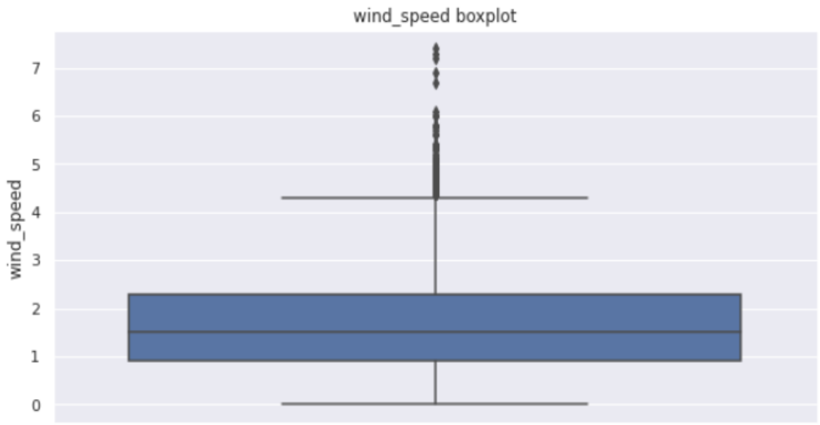


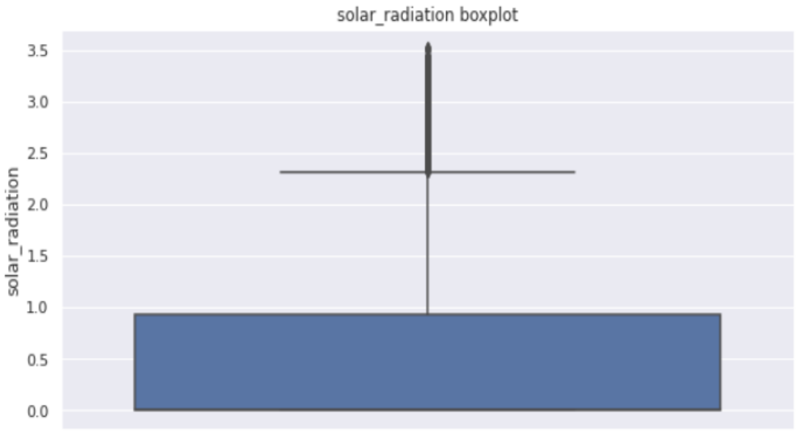


From above distribution of the feature it is seen that some feature are skewed

- ✓ Right skewed columns are Rented Bike Count (Its also our Dependent variable), Wind speed (m/s), Solar Radiation (MJ/m2), Rainfall(mm), Snowfall (cm),
- ✓ Left skewed columns are Visibility (10m), Dew point temperature(°C)

2. Distribution of features by using boxplot

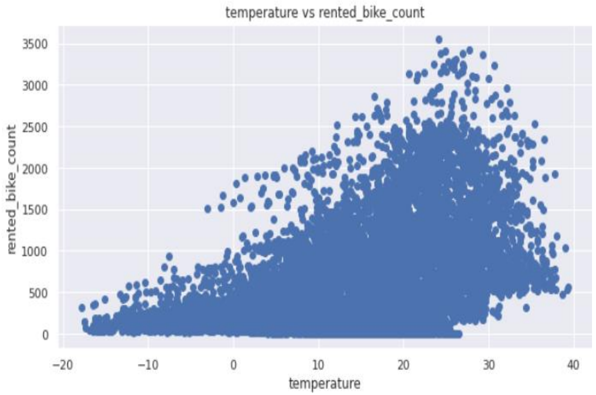


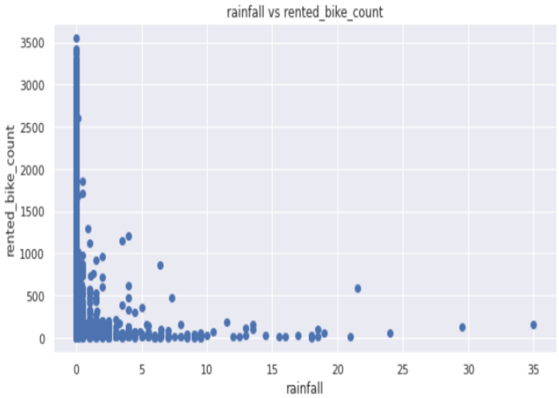


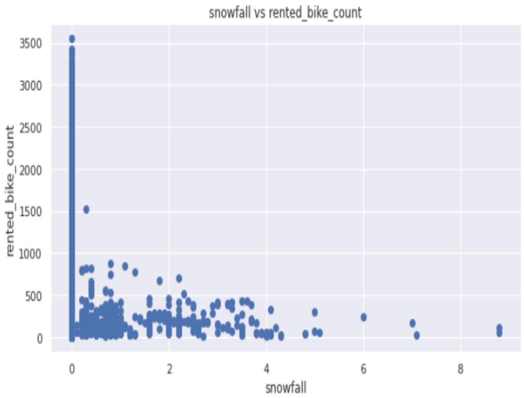
From above it is seen that some of the features have outliers. So that we will remove them later

Bivariate Analysis :

1.Analyzing the relationship between the dependent variable and the continuous variables in the data

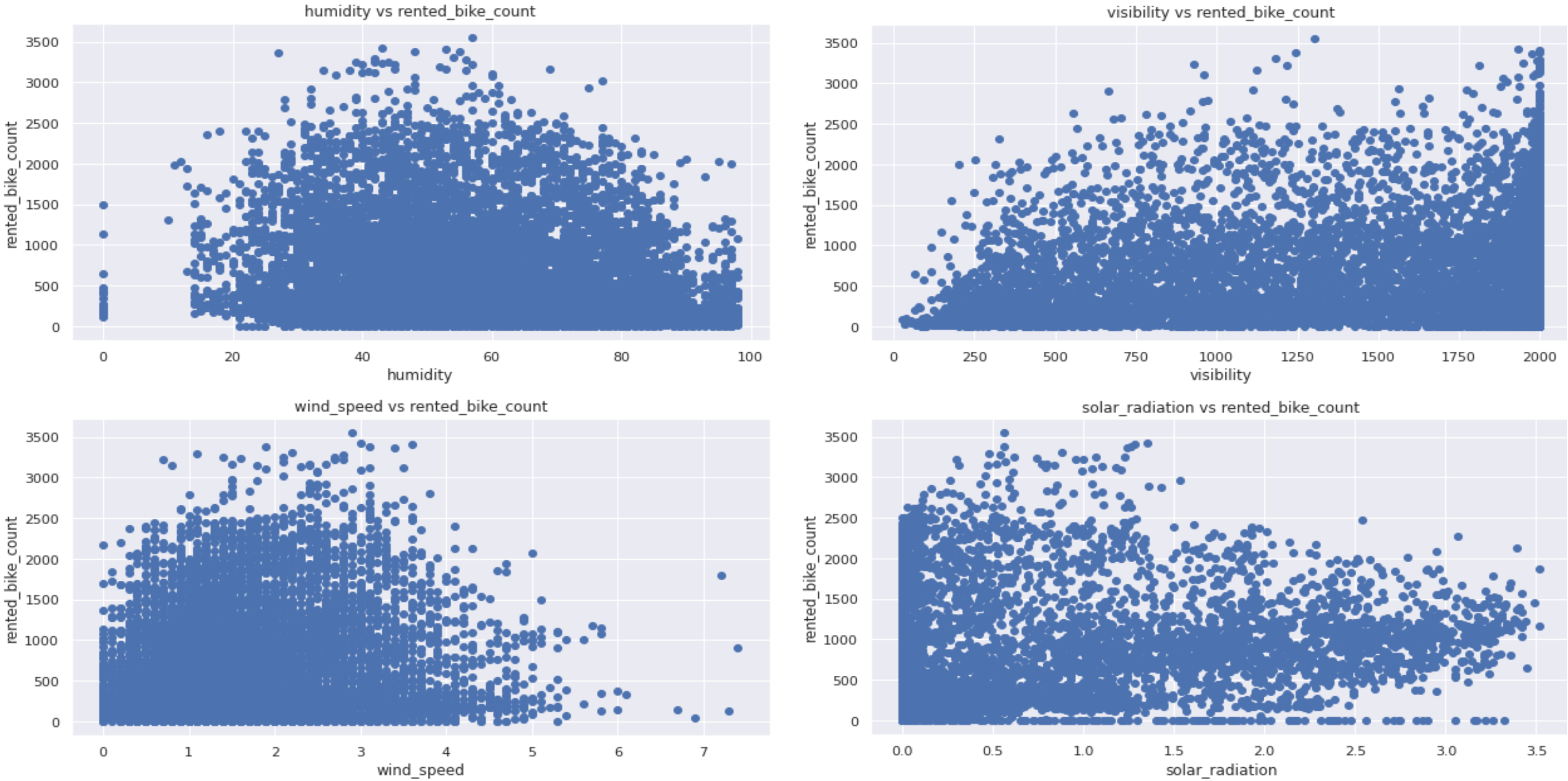




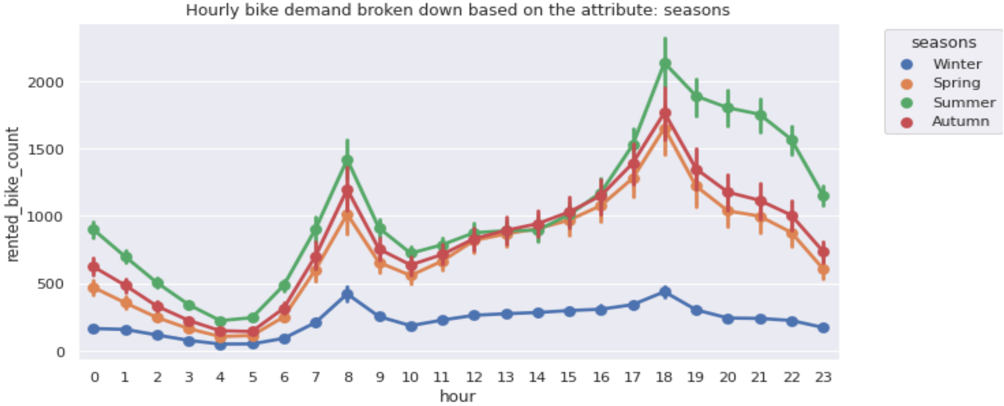


- ✓ Temperature, with the room temperature range, bike demand is higher than the extreme low and high temperature range.
- ✓ Rainfall, demand is high when there are no rainfall because bikes are open and chance of steep in rainfall.
- ✓ Snowfall, bike demand is same in snowfall as in rainfall.

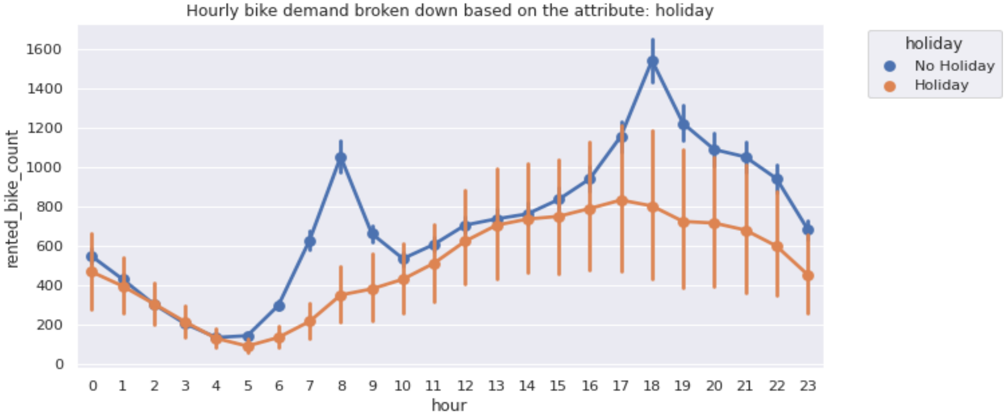
Factor by which bike demand is varies with very less amount are humidity, wind speed, visibility, solar radiation.



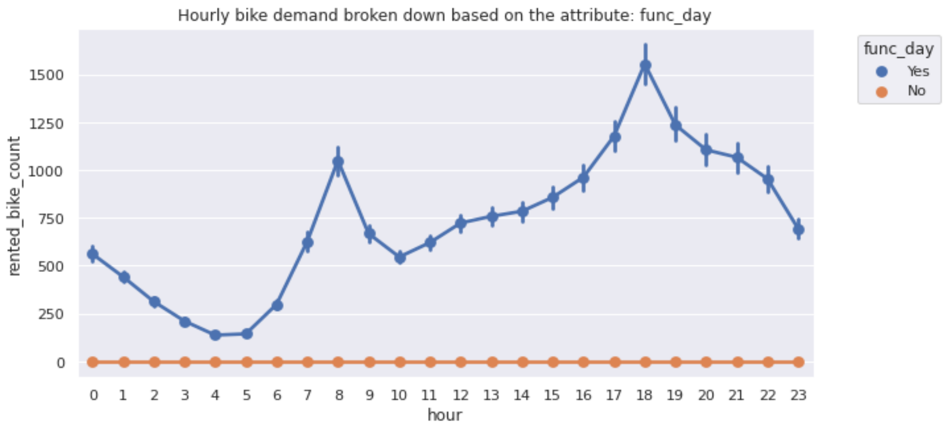
2.Analyzing the relationship between the dependent variable and the categorical variables in the data



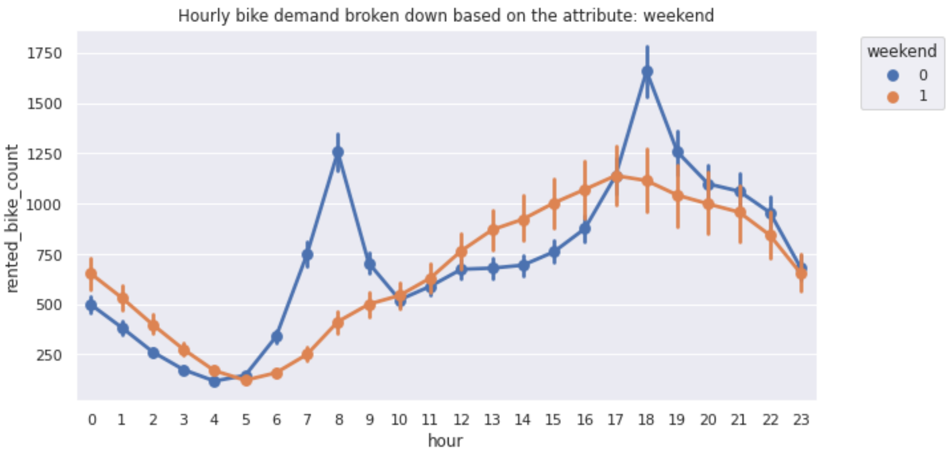
- ✓ Seasons, Demand is high in summer and then spring, autumn have same demand moderate and then lowest in winter.



- ✓ Holiday, High demand on regular day i.e. no holiday and less on holiday.



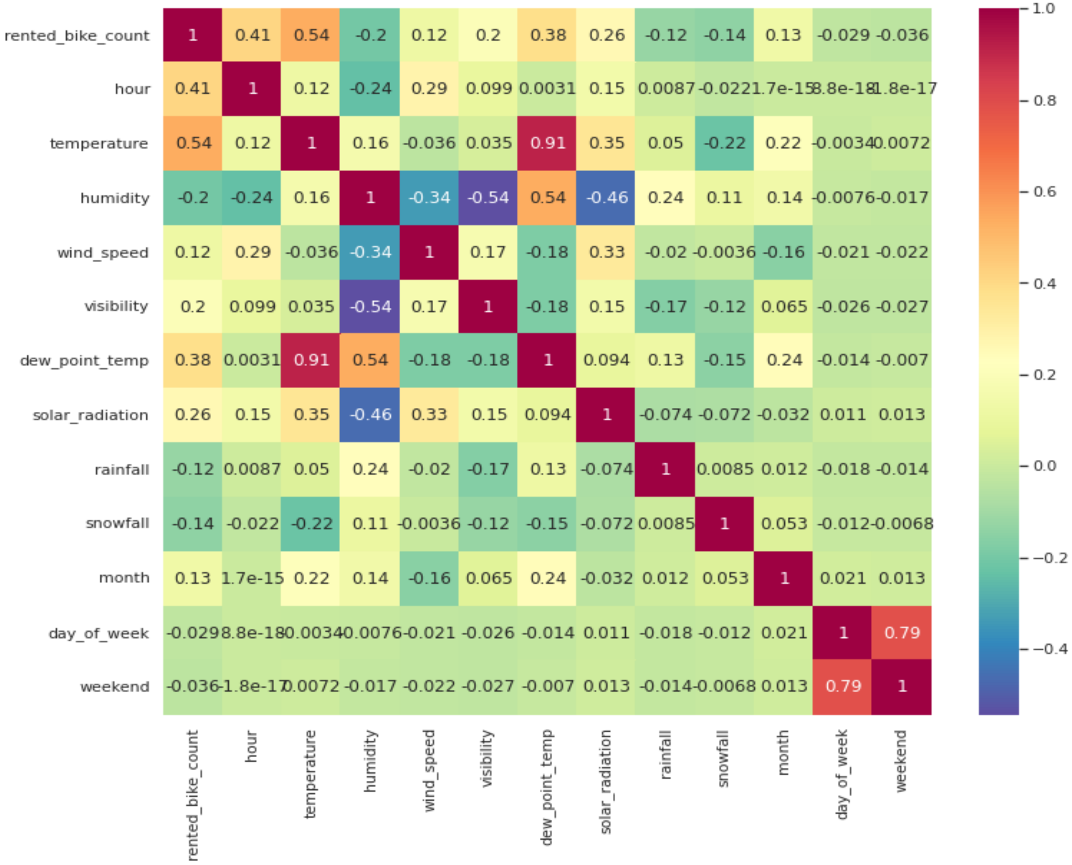
✓ Functional day, zero demand on non- functional day



✓ On weekend demand of the bike remain less than regular week day

Multivariate Analysis:

From the correlation graph with Heat map we saw that dew point temp and temperature is highly correlated. Then we checked VIF and concluded that these two features are affecting VIF score also. so we decided to drop one of these feature and to do this we checked which feature is least correlated with Dependent variable and we identified it to be Dew point temperature and therefore we dropped the Dew point temperature.

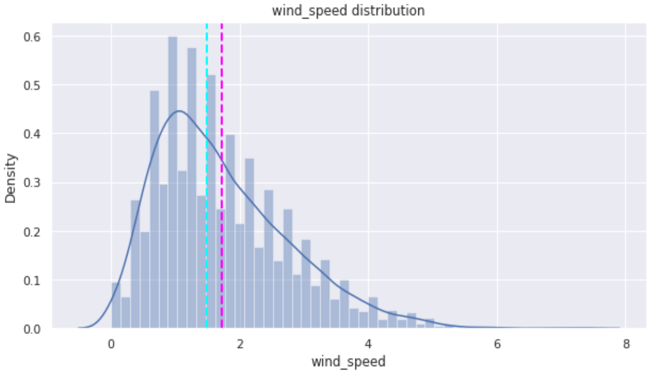


Data Pre-Processing

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model by following processes:

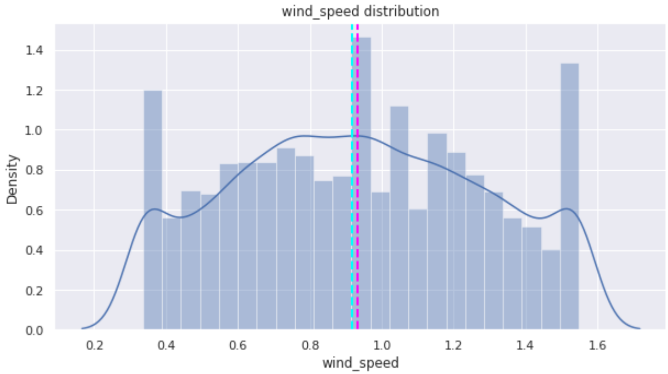
- ✓ Handling The Outlier by capping, we have capped the outlier having values greater than 95 percentile at higher level i.e. at 95 Percentile and outlier having value greater than 5 Percentile capped to the lower level i.e.at 5 percentile.

- ✓ Skewness reduction by using log transformation



Before

After



- ✓ One hot encoding to produce binary integers of 0 and 1 to encode our categorical features, because categorical features that are in string format cannot be understood by the machine and needs to be converted to numerical format

- ✓ Multicollinearity, removing the feature that are correlate to each other.

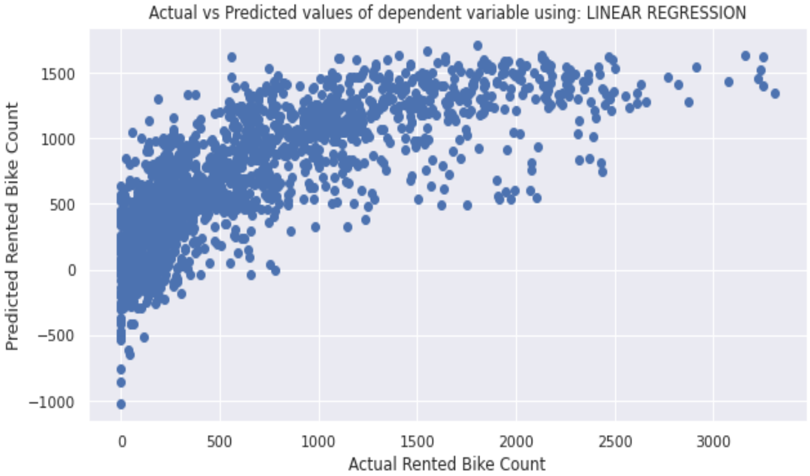
Regression Analysis

Result of the Regression Models:

Actual rented bike demand vs. predicted rented bike demand

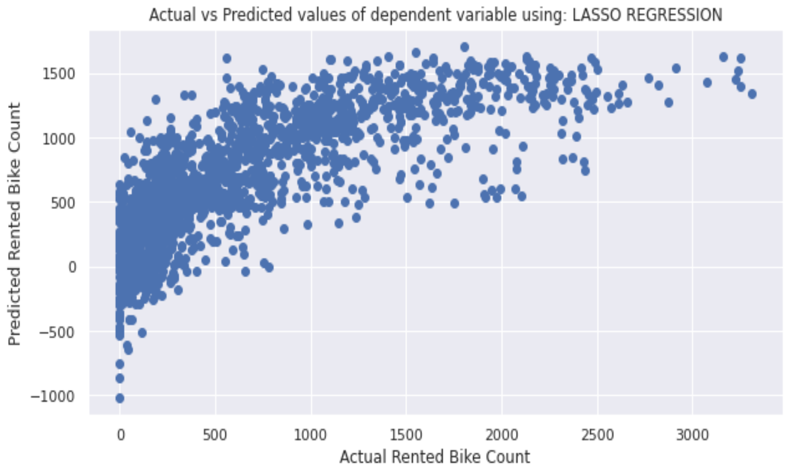
We have taken same scale on both the axis so scatter plot points are plotted by taking the intersection of the actual and predicted demand values , so that if scatter plot is seen to be linear means that model is predicted as per the actual demand i.e. well doing , and if plot is non-linear means that model is not predicting as per the actual demand i.e. not doing well

1. Linear Regression Model



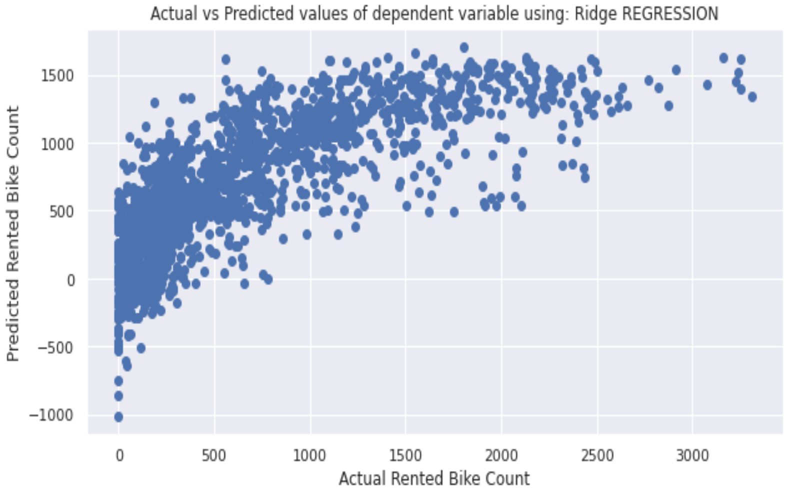

```
MAE : 315.65378662628666
MSE : 178982.72248926878
RMSE : 423.0634969945632
SCORE : 0.5723463262753175
```

2. Regularization Lasso Regression



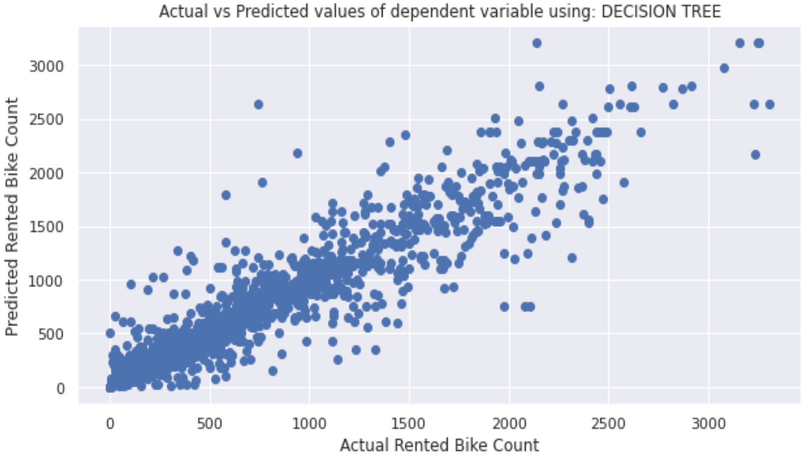

```
MAE : 315.6237405031461
MSE : 179010.78856067243
RMSE : 423.0966657404339
SCORE : 0.5722792664028569
```

3. Regularization Ridge Regressor



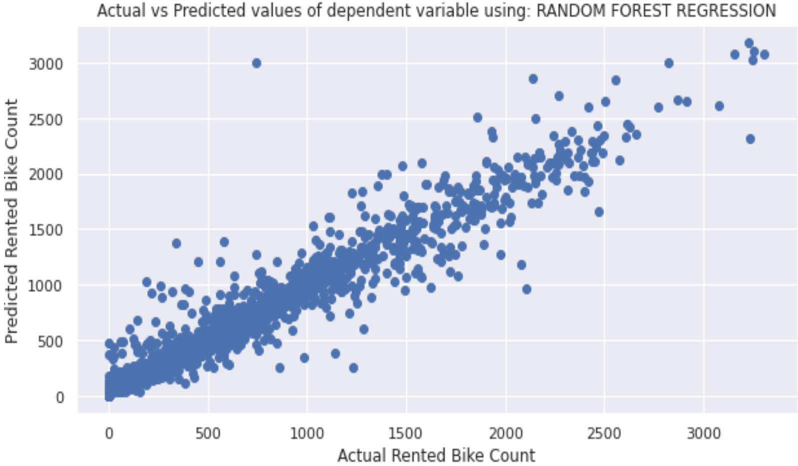

```
MAE : 315.63330067837205
MSE : 178985.81677752748
RMSE : 423.06715398093417
SCORE : 0.5723389329150926
```

4. Decision Tree Regression



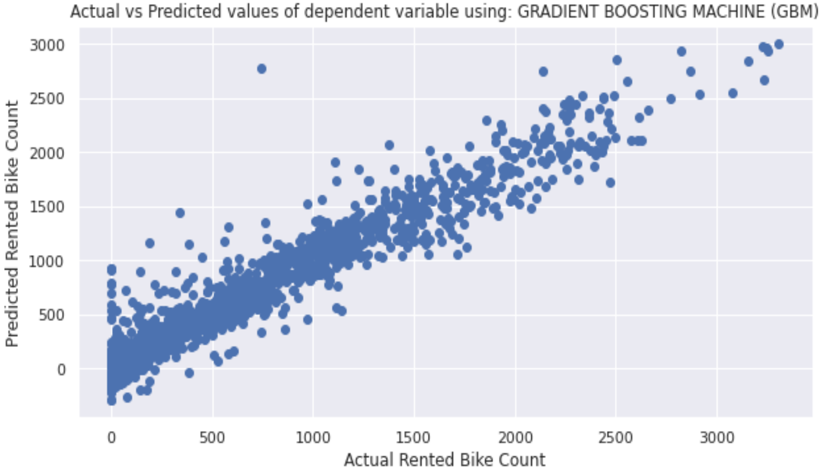

```
MAE : 125.98140393547858
MSE : 48313.3345358122
RMSE : 219.80294478421393
SCORE : 0.8845621816632695
```

5. Random Forests Regression



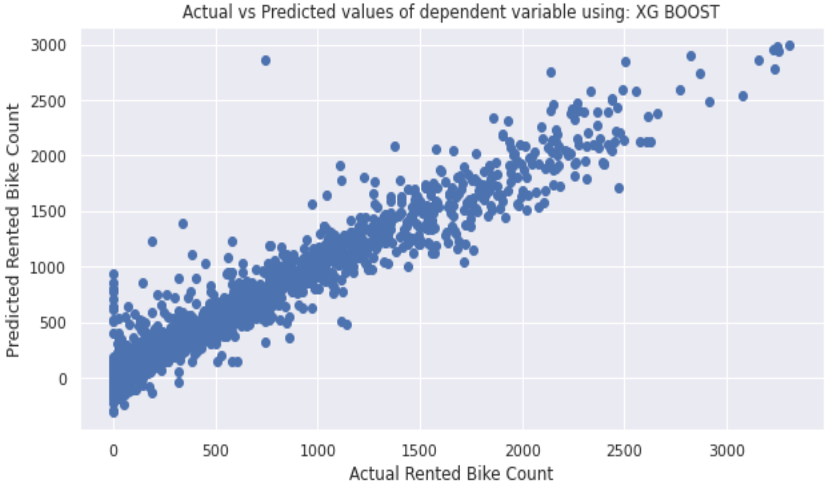

```
MAE : 99.92153538812785
MSE : 30397.33501569635
RMSE : 174.34831520750737
SCORE : 0.9273699058204801
```

6. Gradient Boosting Regression



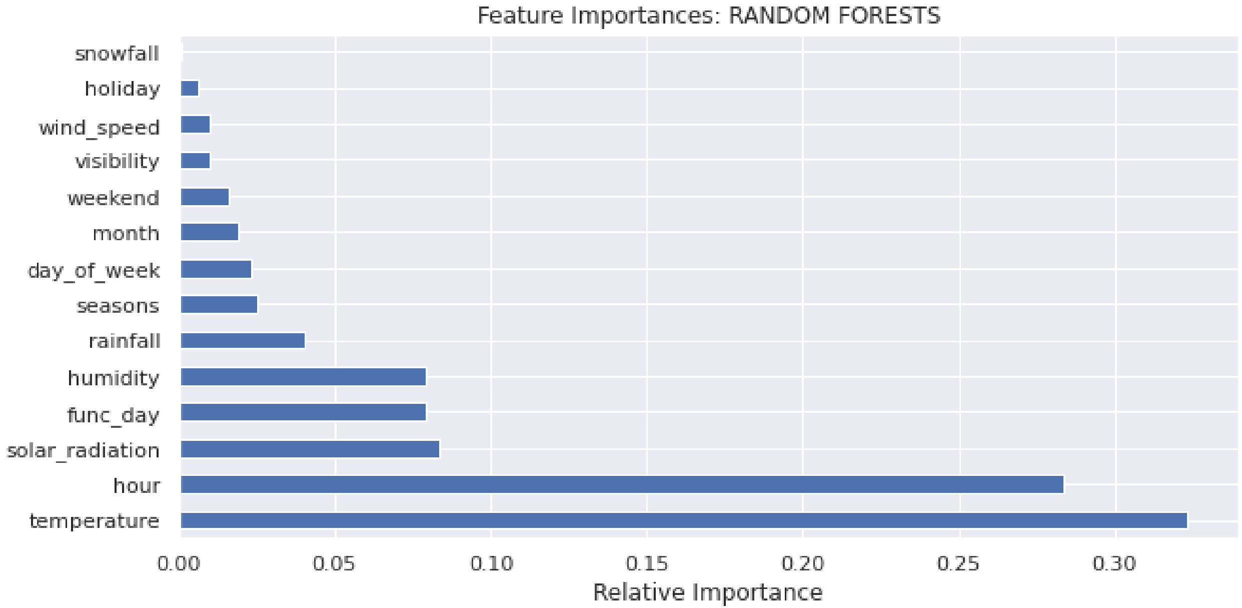

```
MAE : 117.37733972107534
MSE : 33170.17759688237
RMSE : 182.1268173468212
SCORE : 0.9207445941702144
```

7. XGBoost Regression




```
MAE : 115.99107459864524
MSE : 32560.165930979685
RMSE : 180.44435688316685
SCORE : 0.9222021299943989
```

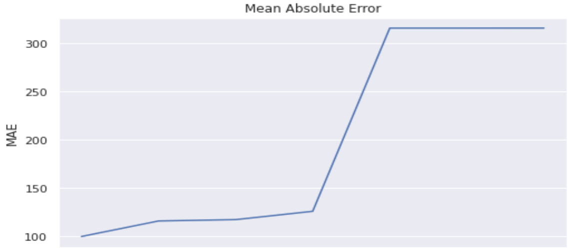
Features Importance

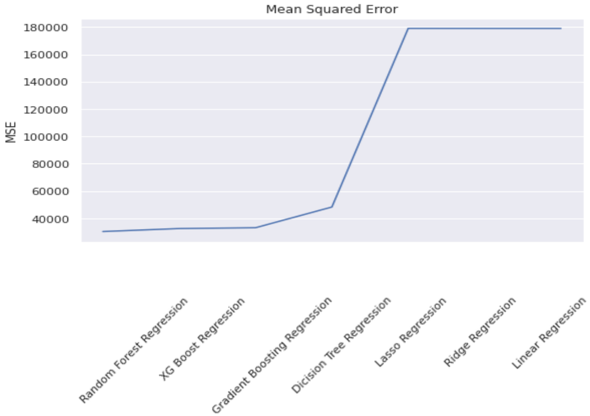


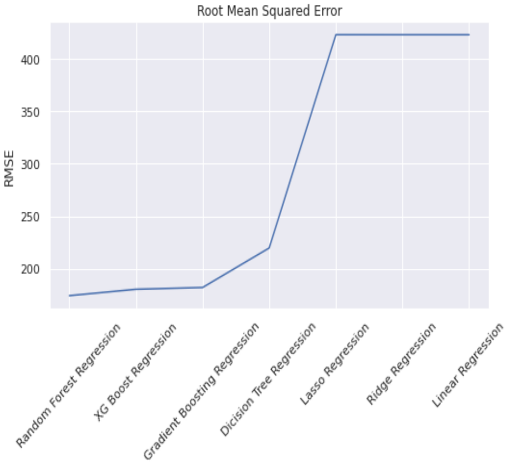
- ✓ Top five important features as per the highest performing model among executed model i.e. random forest , features are temperature, hour of the day, solar radiation, functional day, humidity.

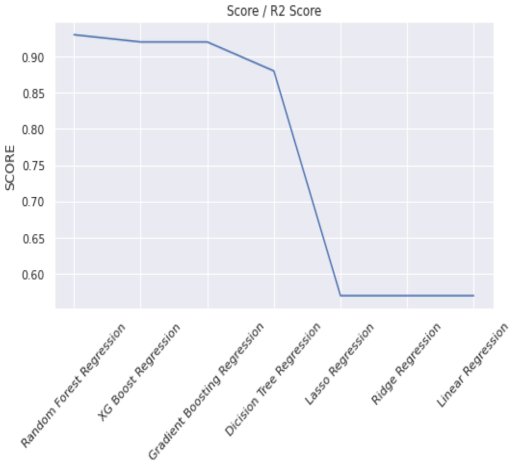
Models Performance Metrics

	Model_Name	MAE	MSE	RMSE	SCORE
4	Random Forest Regression	99.92	30397.34	174.35	0.93
6	XG Boost Regression	115.99	32560.17	180.44	0.92
5	Gradient Boosting Regression	117.38	33170.18	182.13	0.92
3	Dicision Tree Regression	125.98	48313.33	219.80	0.88
1	Lasso Regression	315.62	179010.79	423.10	0.57
2	Ridge Regression	315.63	178985.82	423.07	0.57
0	Linear Regression	315.65	178982.72	423.06	0.57









Conclusion

- ✓ The target variable i.e. dependent variable (count of bike sharing demand) is highly dependent on input variables i.e. independed variables.

- ✓ Linear regression did not give an excellent result. Ridge regression shrunk the parameters to reduce complexity and multicollinearity but ended up affecting the evaluation metrics and ended up giving up worse results than lasso regression. These three models gave almost the same results.

- ✓ Decision tree gave a moderate result than the previous three models but not enough score with 0.88. Gradient Boosting and XG Boost regression gave the same result about 0.92 score.

- ✓ Random Forest regression gives the highest result about 0.93 score with minimum error than all other implemented models, so we can use the random forest regressor model for further prediction.

- ✓ As we have seen above while selecting a model should have well explainability and less complexibility. As per the result, we have all three models with higher accuracy and less error are black box models so that less explainable, but in this case, accuracy is more important so that our final model can be the random forest regression.

Thank You