Capstone Project

On

Seoul Bike Sharing Demand Prediction

By

Roshan Tile

CONTENT

- Problem Statement
- Data Summary
- Exploratory Data Analysis (EDA)
  - Univariate Analysis
  - Bivariate Analysis
  - Multivariate Analysis
- Data Pre-processing
- Regression Analysis
- Models Performance Metrics
- 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

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

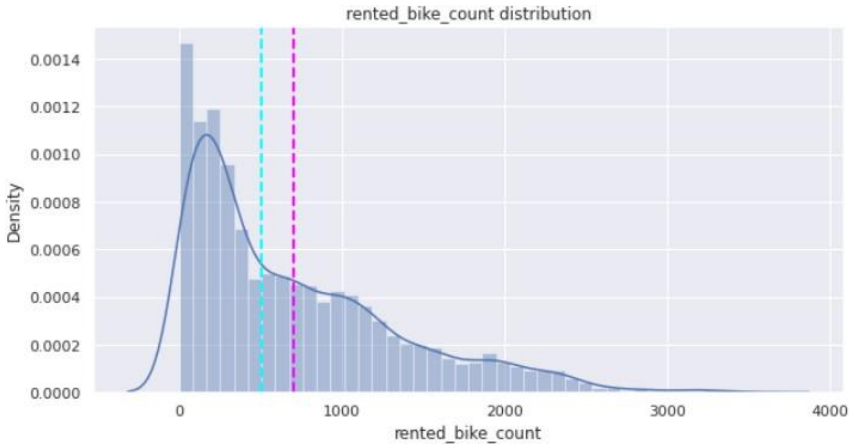
- The dataset contains weather information (Temperature, Humidity, Wind speed, Visibility, Dew point, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

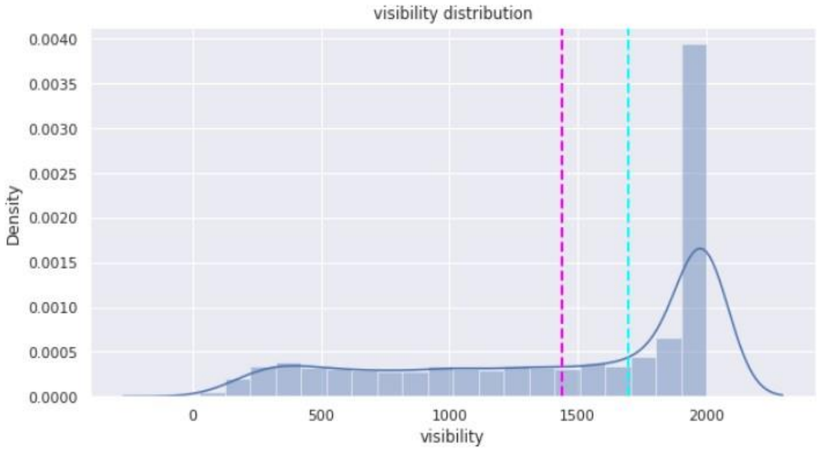
- This dataset contains the hourly and daily count of rental bikes between years 2017 and 2018 in Seoul bike share system with the corresponding weather and seasonal information. The dataset contains 8760 rows (every hour of each day for 2017 and 2018 i.e. 375days * 24 Hr) and 14 columns (the features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

Exploratory Data Analysis (EDA)

Univariate Analysis :

1. Distribution of numerical features

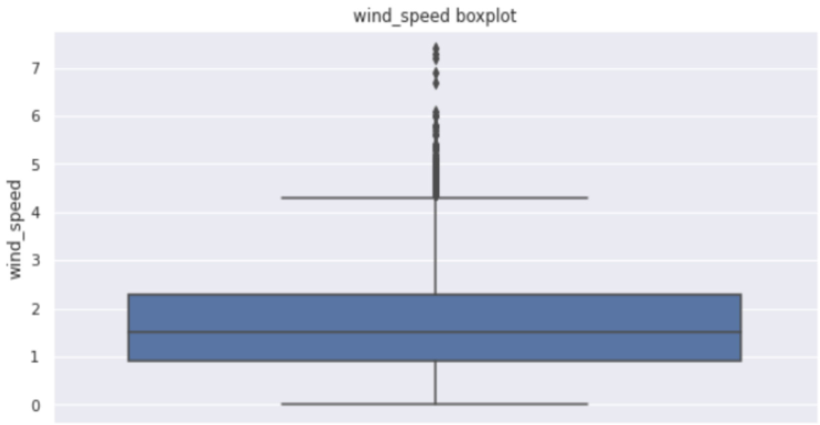


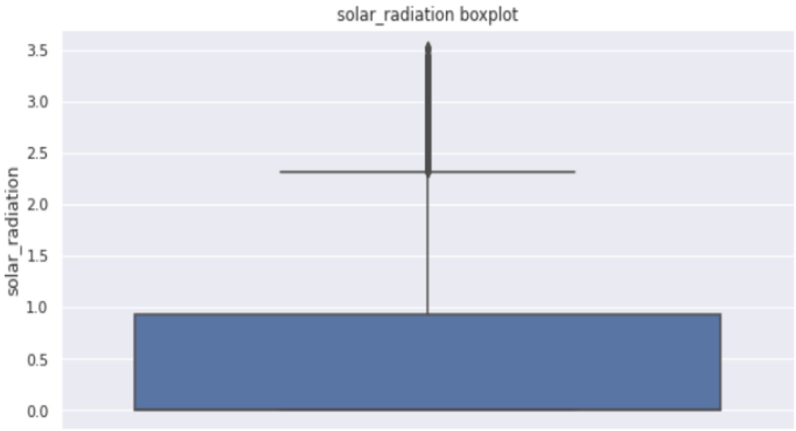


From above distribution of the feature, it is seen that some feature are skewed

- ✓ **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

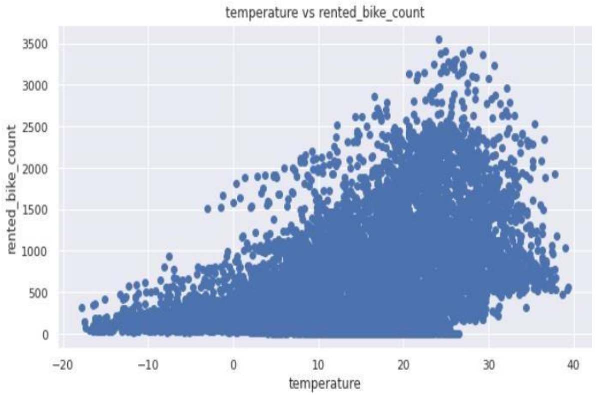


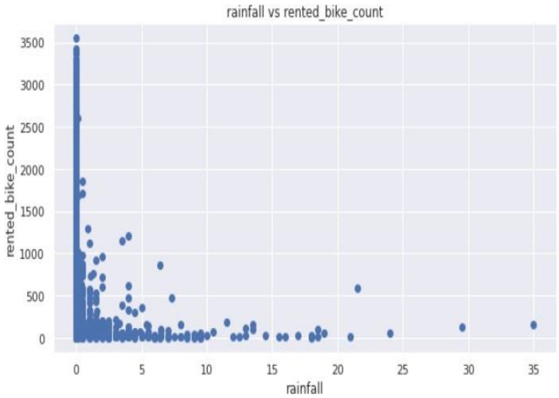


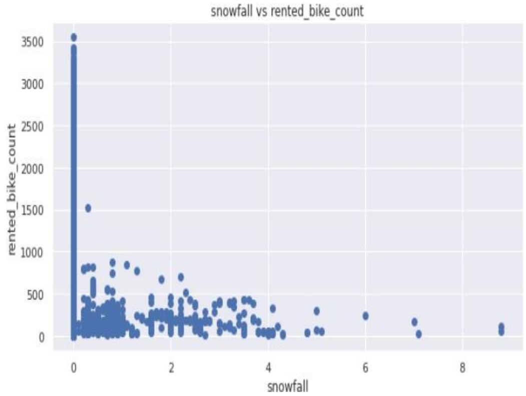
From above it is seen that some of the features have outliers, So that we will remove them later.

Bivariate Analysis :

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

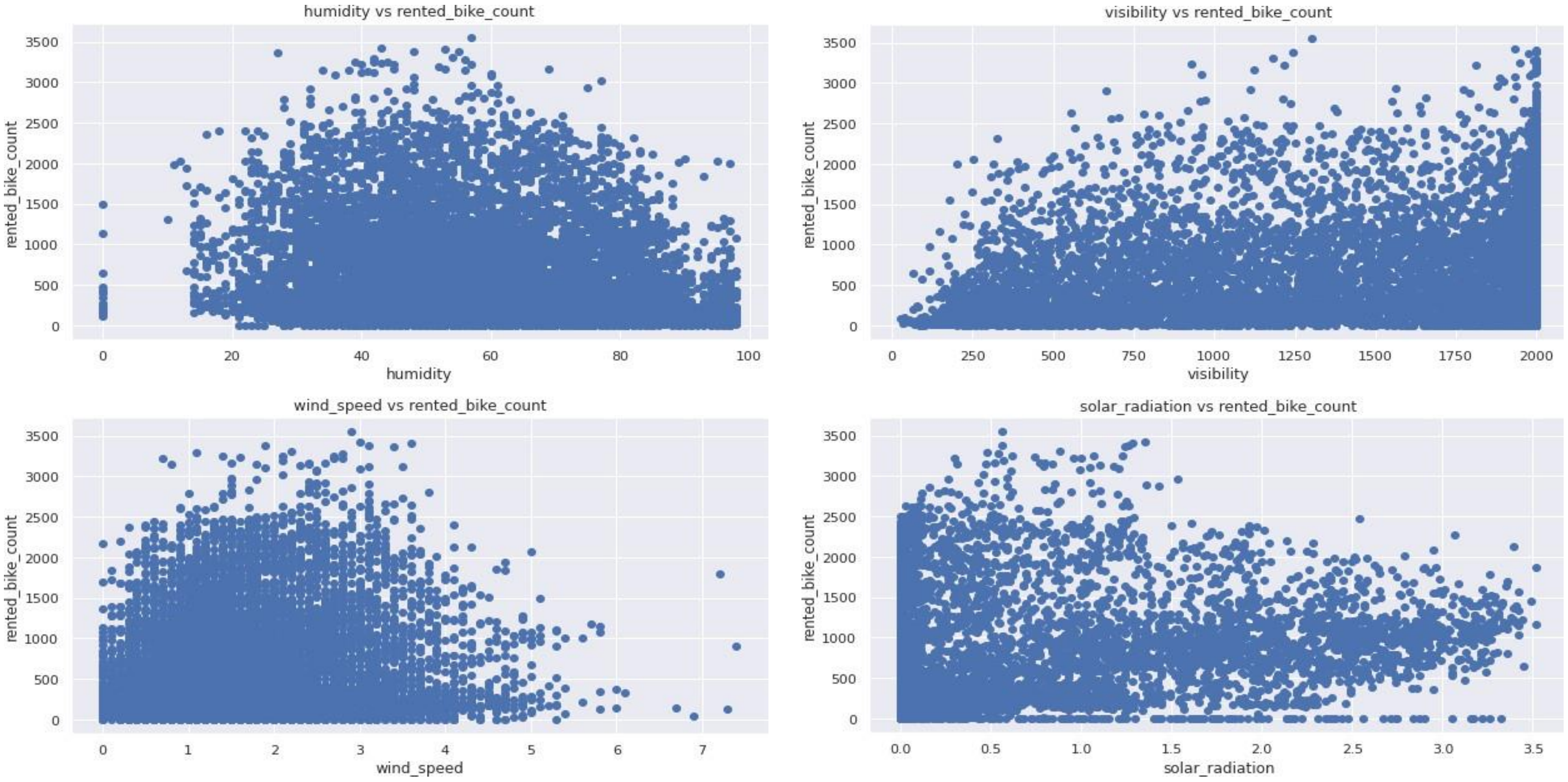




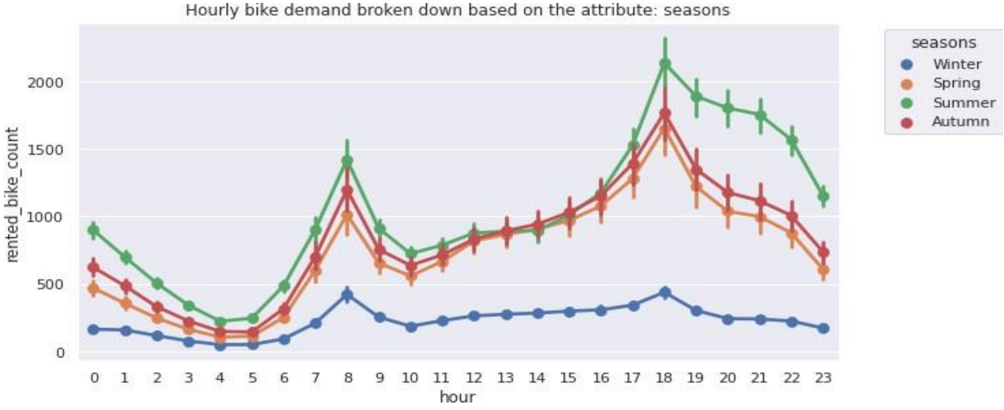


- ✓ Temperature, with the room temperature range, bike demand is higher than the extreme low and high temperature range.
- ✓ Rainfall, demand is high when there are no rainfall because bikes are open and chance of steep in rainfall.
- ✓ Snowfall, bike demand is same in snowfall as in rainfall.

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

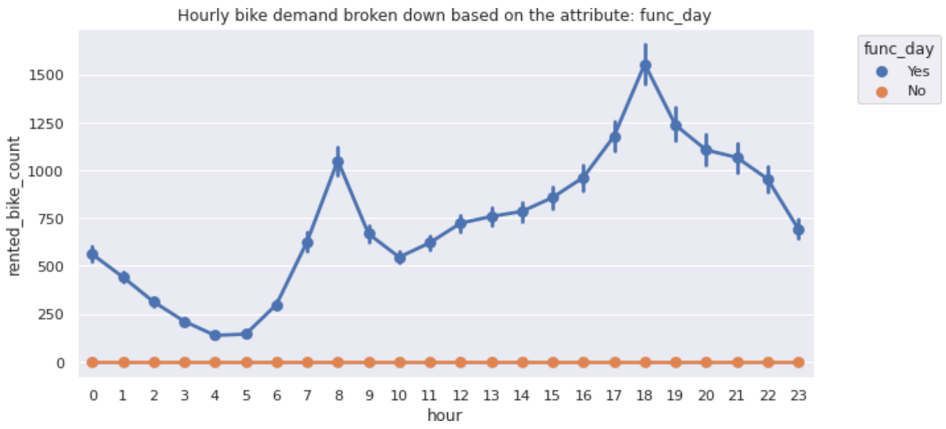


2.Analyzing the relationship between the dependent variable and the categorical variables in the data

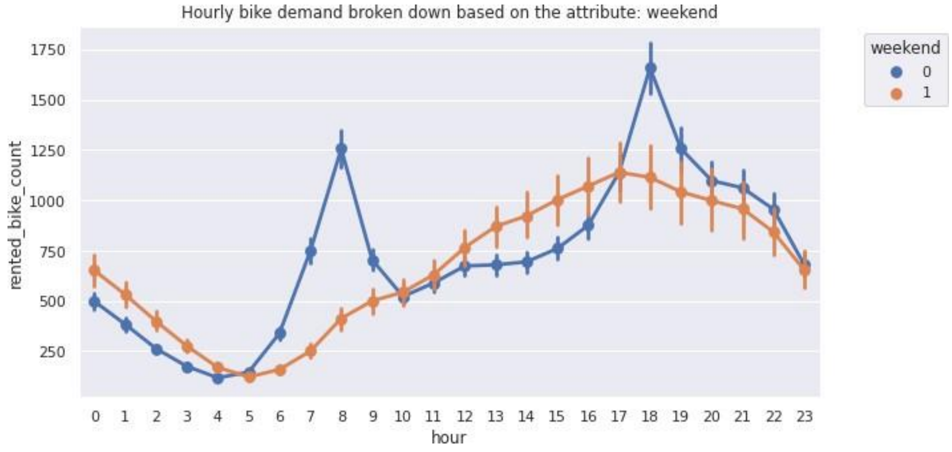


- ✓ **Seasons, Demand is high in summer and then spring, autumn have same demand moderate and then lowest in winter.**

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

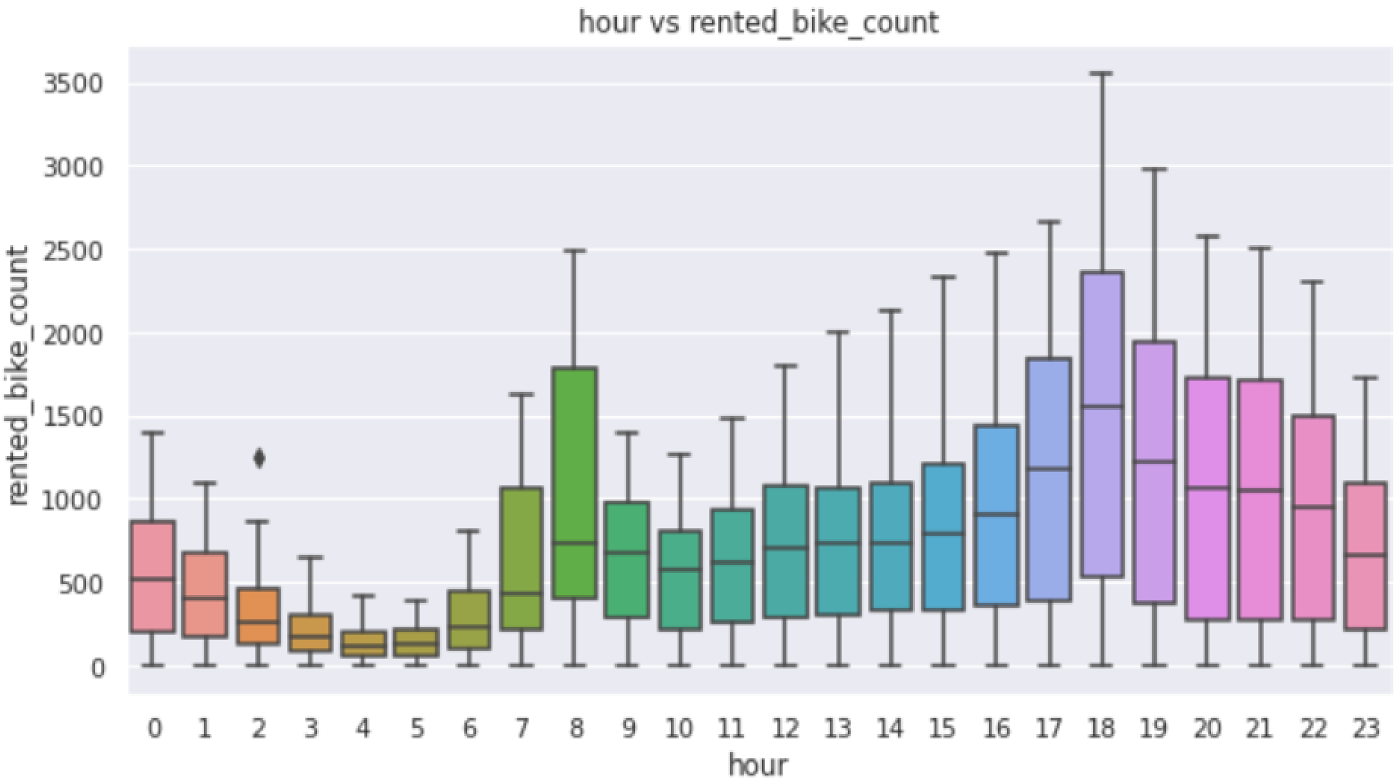


- ✓ **Functional day, zero demand on non-functional day**



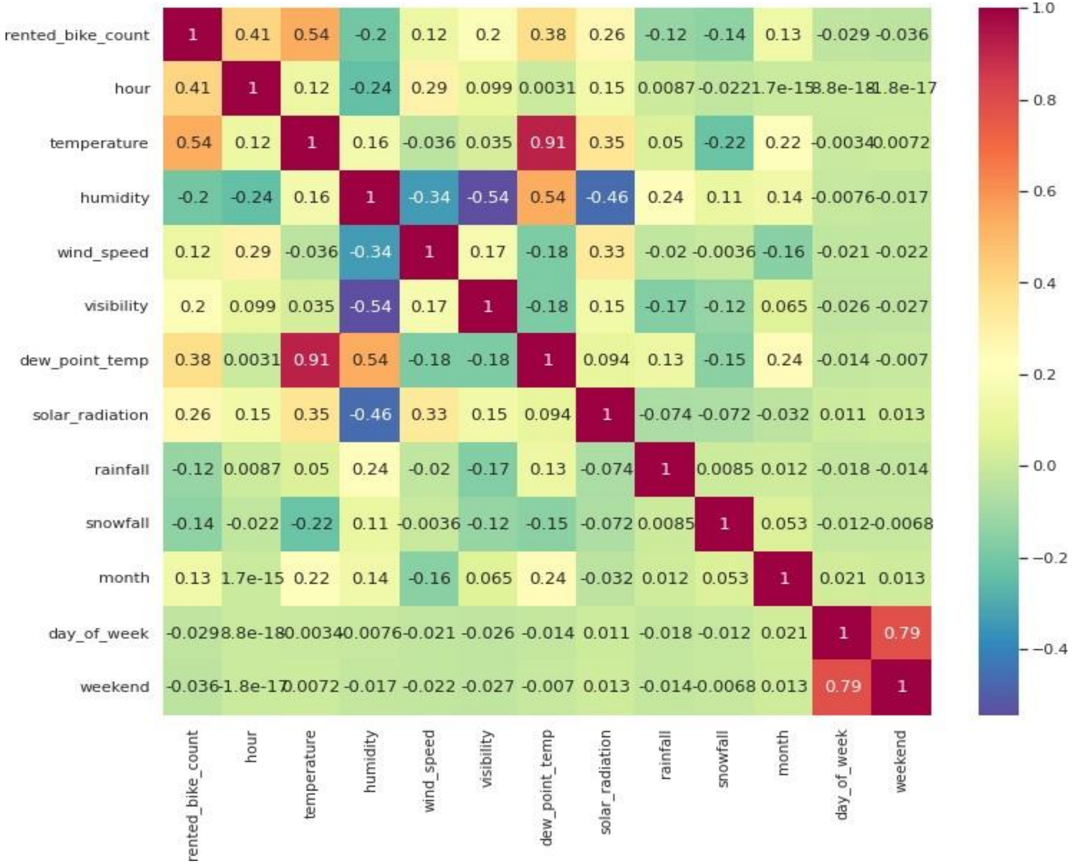
- ✓ **On weekend demand of the bike remain less than regular week day**

Hour vs rented_bike_count box plot



Multivariate Analysis:

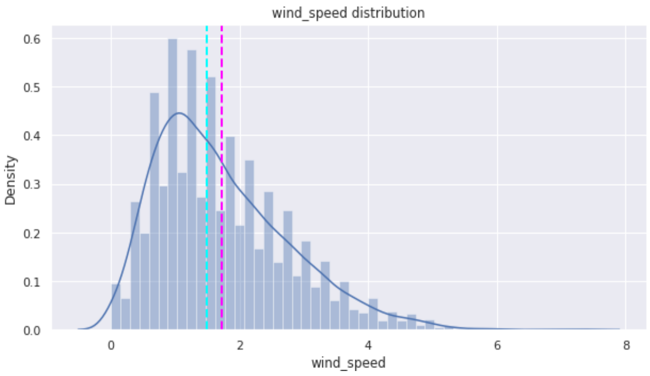
From the correlation graph with Heat map we saw that dew point temp and temperature is highly correlated. so we decided to drop one of these feature and to do this we checked which feature is least correlated with Dependent variable and we identified it to be Dew point temperature and therefore we dropped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



Data Pre-Processing

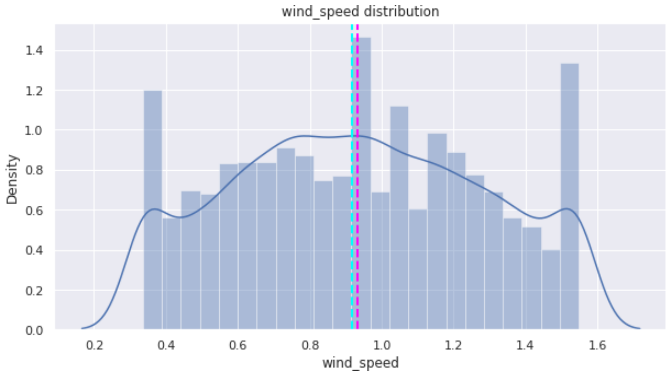
Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model by following processes:

- ✓ Handling The Outlier by capping.
- ✓ Skewness reduction by using transformation methods.



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.

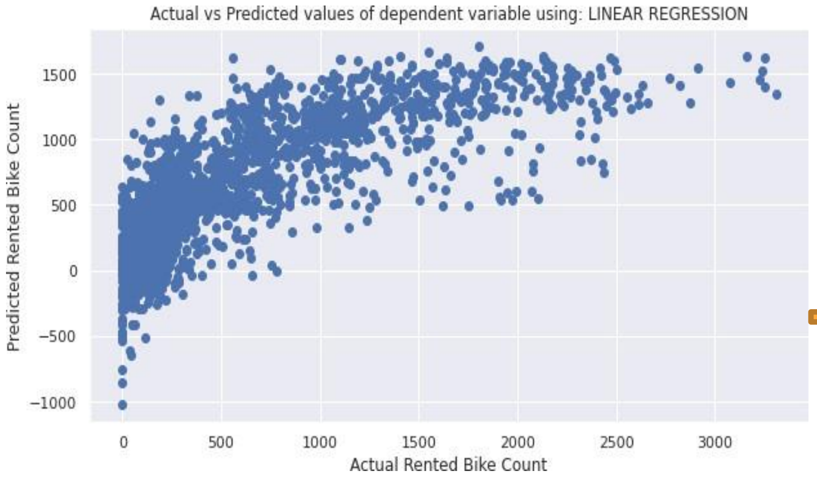
Regression Analysis

Result of the Regression Models:

Actual rented bike demand vs. predicted rented bike demand

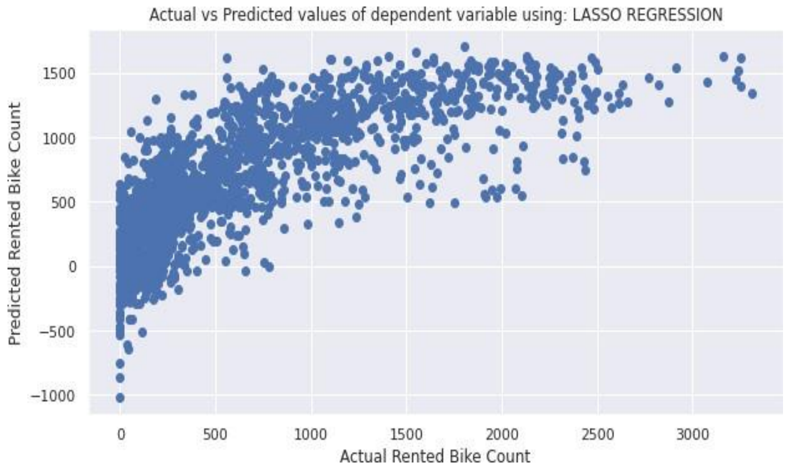
We have taken same scale on both the axis so scatter plot points are plotted by taking the intersection of the actual and predicted demand values , so that if scatter plot is seen to be linear means that model is predicted as per the actual demand i.e. well doing , and if plot is non-linear means that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

1. Linear Regression Model



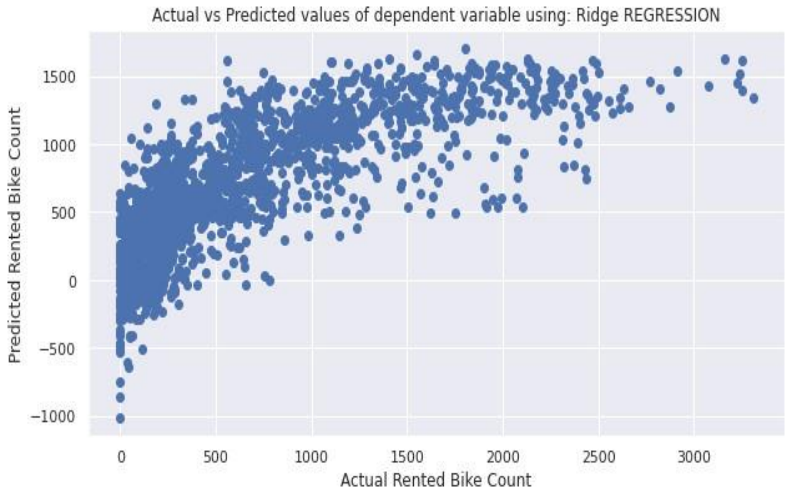

```
MAE : 318.234
MSE : 180279.308
RMSE : 424.593
SCORE : 0.569
```

2. Regularized Lasso Regression



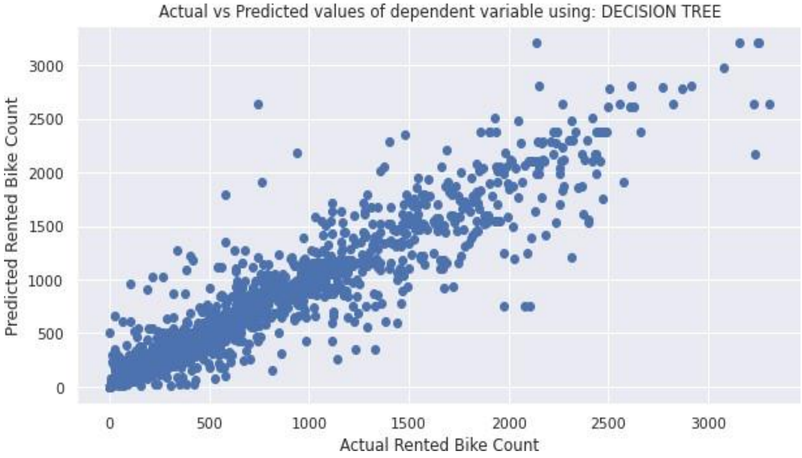
MAE : 318.241
MSE : 180291.149
RMSE : 424.607
SCORE : 0.569

3. Regularized Ridge Regression



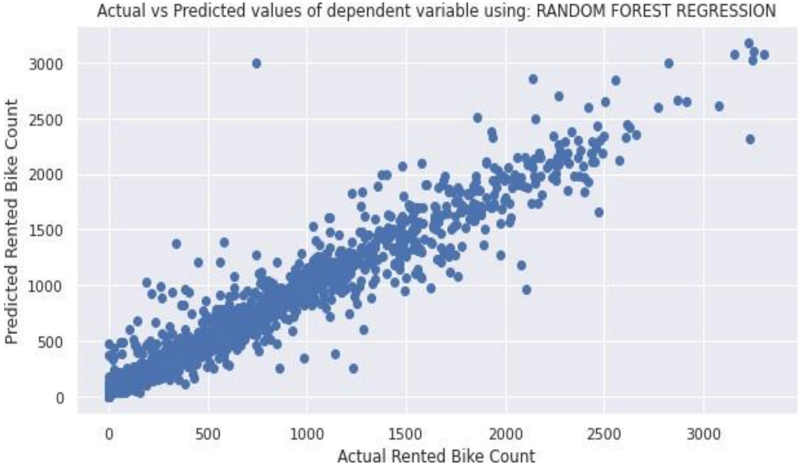
MAE : 318.256
MSE : 180307.249
RMSE : 424.626
SCORE : 0.569

4. Decision Tree Regression



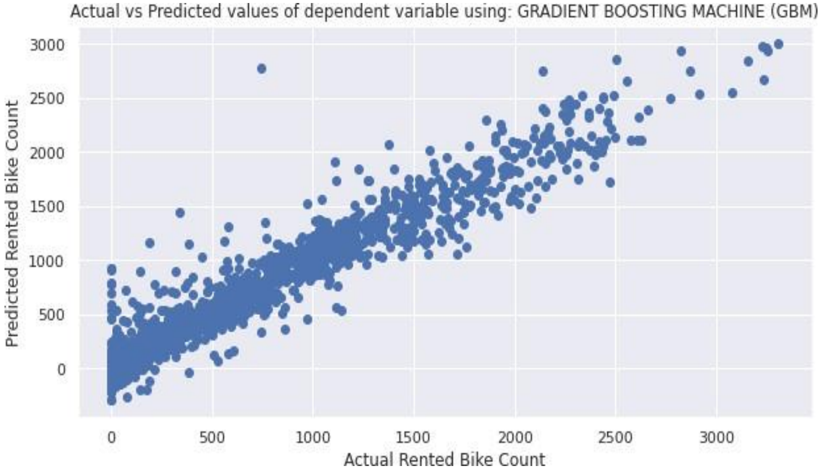
MAE : 126.739
MSE : 47806.117
RMSE : 218.646
SCORE : 0.886

5. Random Forests Regression



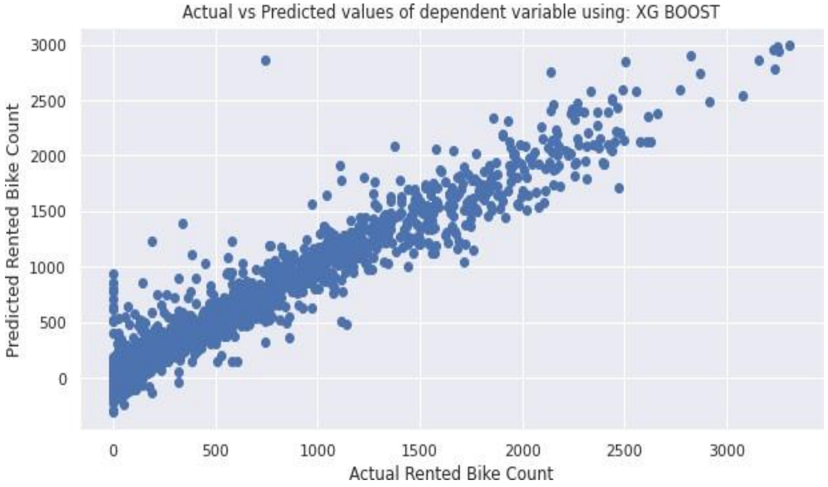
MAE : 100.262
MSE : 30596.819
RMSE : 174.919
SCORE : 0.927

6. Gradient Boosting Regression



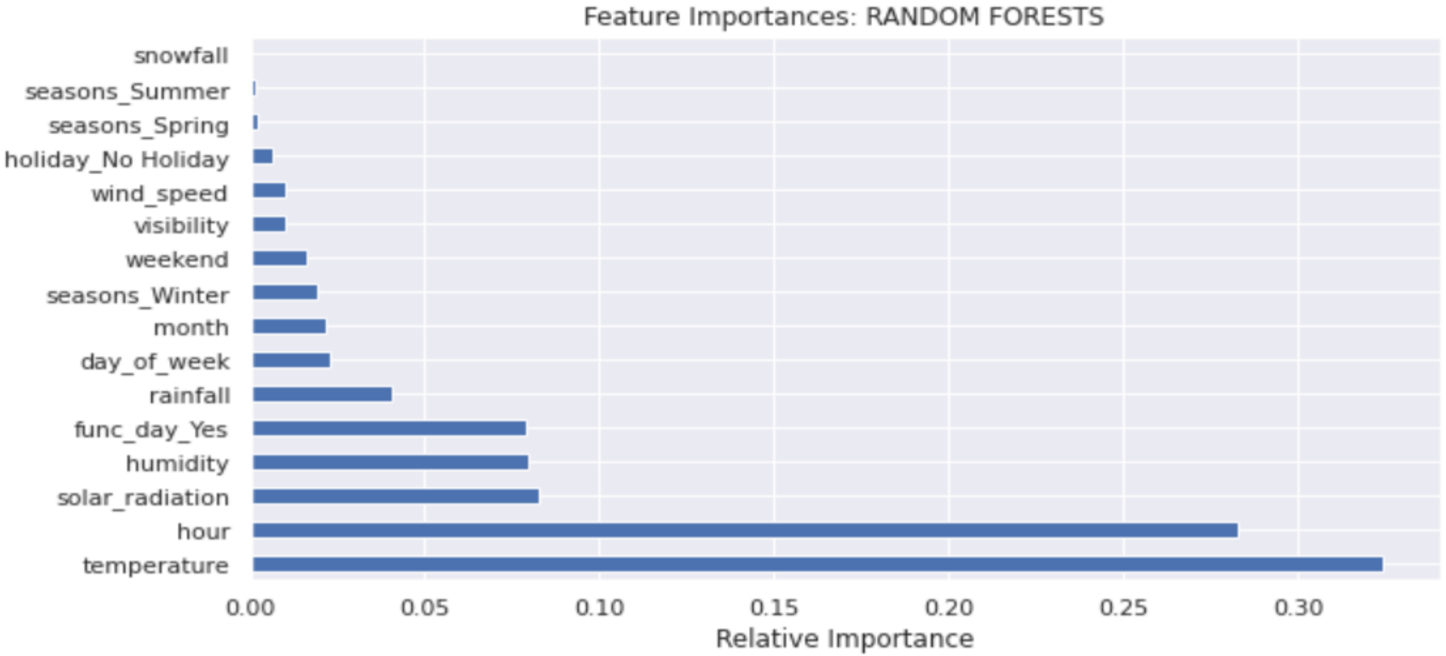

```
MAE : 115.65
MSE : 32808.552
RMSE : 181.131
SCORE : 0.922
```

7. XGBoost Regression




```
MAE : 114.164
MSE : 32133.1
RMSE : 179.257
SCORE : 0.923
```

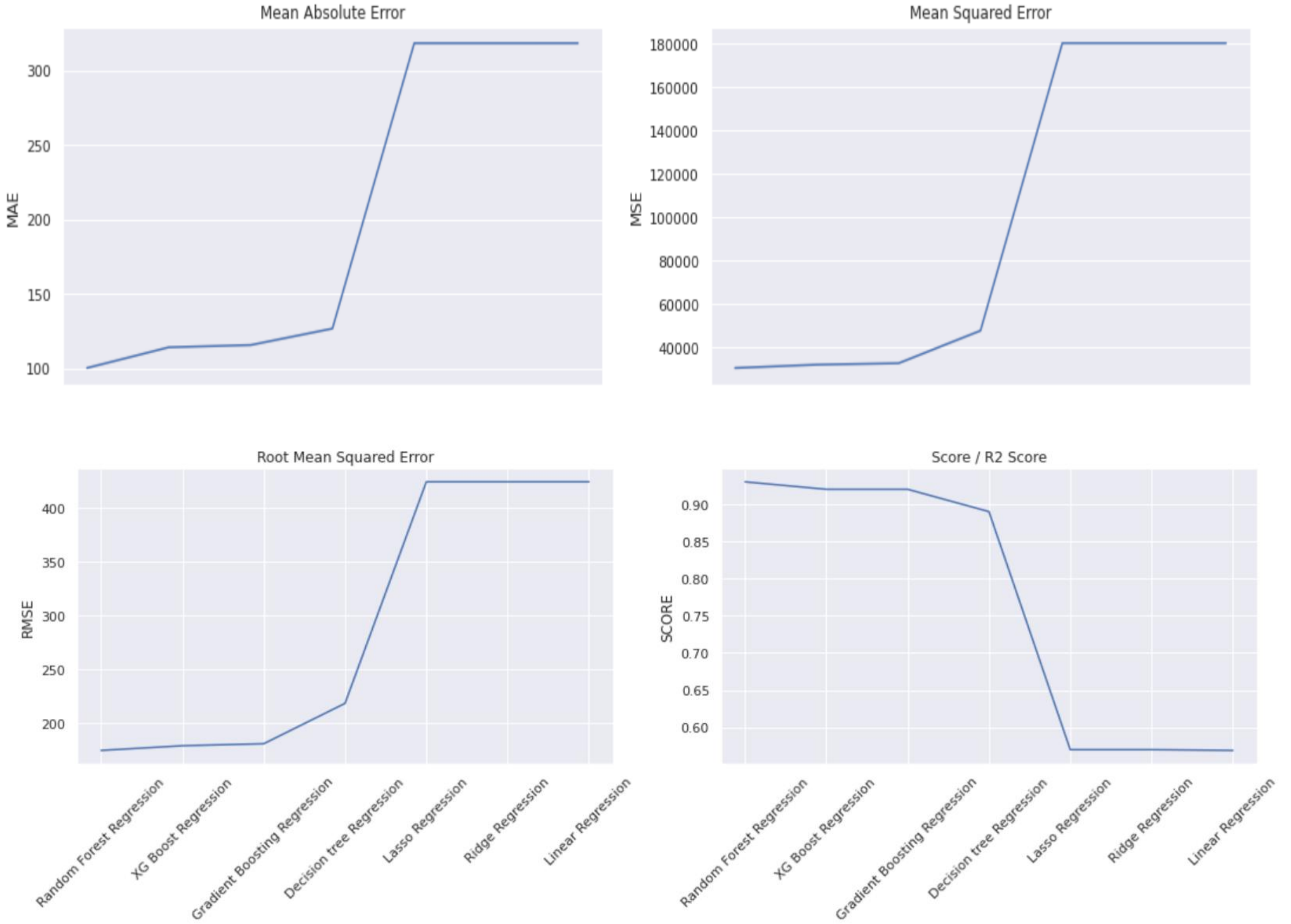
Features Importance



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

Models Performance Metrics

	Model_Name	MAE	MSE	RMSE	SCORE
4	Random Forest Regression	100.260	30596.820	174.920	0.930
6	XG Boost Regression	114.160	32133.100	179.260	0.920
5	Gradient Boosting Regression	115.650	32808.550	181.130	0.920
3	Decision tree Regression	126.740	47806.120	218.650	0.890
0	Linear Regression	318.234	180279.308	424.593	0.569
1	Lasso Regression	318.240	180291.150	424.610	0.570
2	Ridge Regression	318.260	180307.250	424.630	0.570


Mean Absolute Error
MAE
100
150
200
250
300
Mean Squared Error
MSE
40000
60000
80000
100000
120000
140000
160000
180000
Root Mean Squared Error
RMSE
200
250
300
350
400
Score / R2 Score
SCORE
0.60
0.65
0.70
0.75
0.80
0.85
0.90
Random Forest Regression
XG Boost Regression
Gradient Boosting Regression
Decision tree Regression
Lasso Regression
Ridge Regression
Linear Regression

Conclusion

- 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. 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.

- 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 they are less explainable, but in this case, accuracy is more important so that our final model is the random forest regression model.

Thank You