Capstone Project - 2

Bike Sharing Demand Prediction

Submitted by

Kousik Dutta

Contents

Problem Statement

- 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 excel. Mostly used by people having no personal vehicles and also to avoid congested public transport which follows its own time.

- 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
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

- This Dataset contains 8760 rows and 14 columns.
- The dataset shows hourly rental data for one year (1 December 2017 to 30 November 2018) (365 days).

Data Summary

Features

Categorical Column

- Seasons
- Holiday
- Functioning Day

Numerical Column

- Date
- Hour
- Temperature
- Humidity
- Windspeed
- Visibility
- Dew Point Temperature
- Solar Radiation
- Rainfall
- Snowfall

Target Variable

Rented Bike Count

Data Summary

Null Values

- There are No Missing Values present
- There are No Duplicate values present

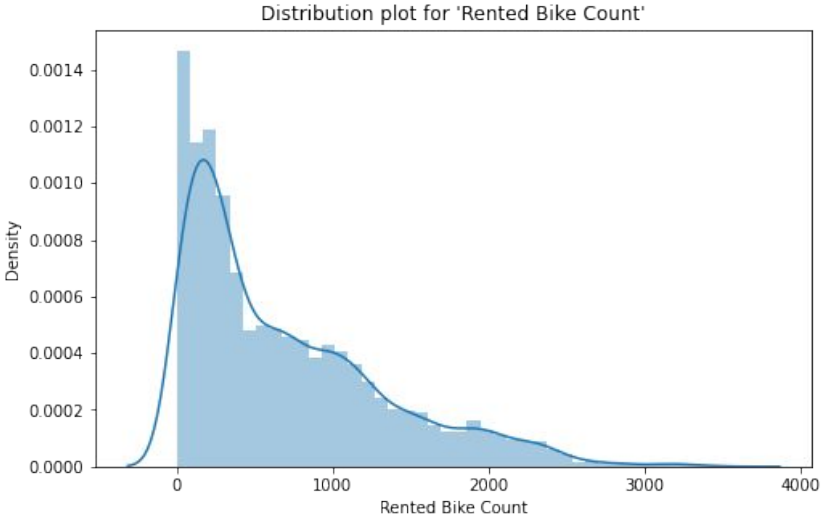
```
bike_df.isnull().sum()
```

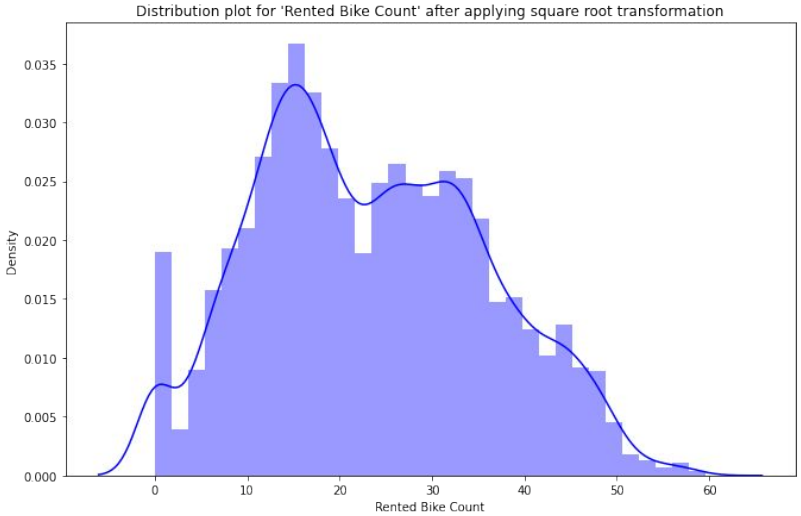
```
Date                          0
Rented Bike Count             0
Hour                          0
Temperature(°C)               0
Humidity(%)                   0
Wind speed (m/s)              0
Visibility (10m)              0
Dew point temperature(°C)     0
Solar Radiation (MJ/m2)       0
Rainfall(mm)                  0
Snowfall (cm)                 0
Seasons                       0
Holiday                       0
Functioning Day               0
dtype: int64
```

Exploratory Data Analysis

EDA (Univariate Analysis)

Rented Bike Count - Dependent Variable



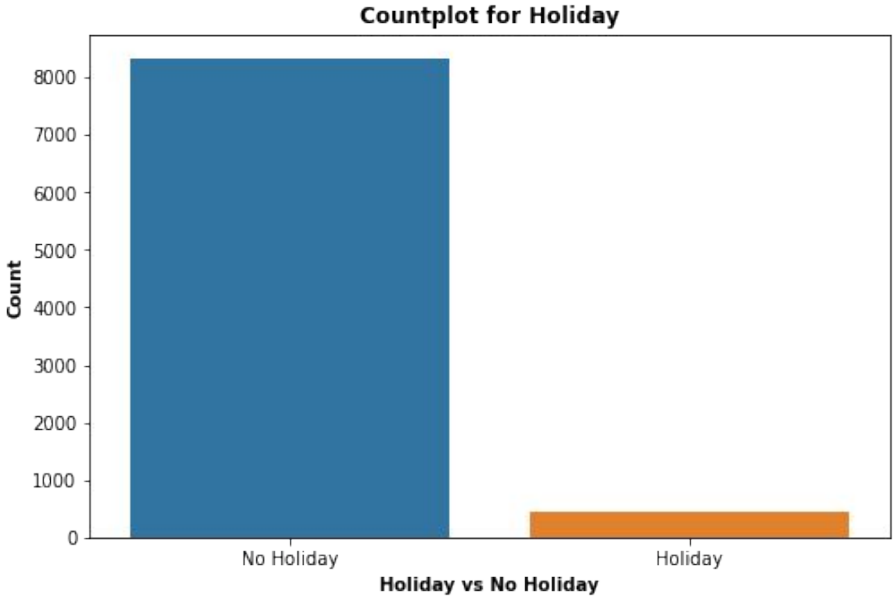


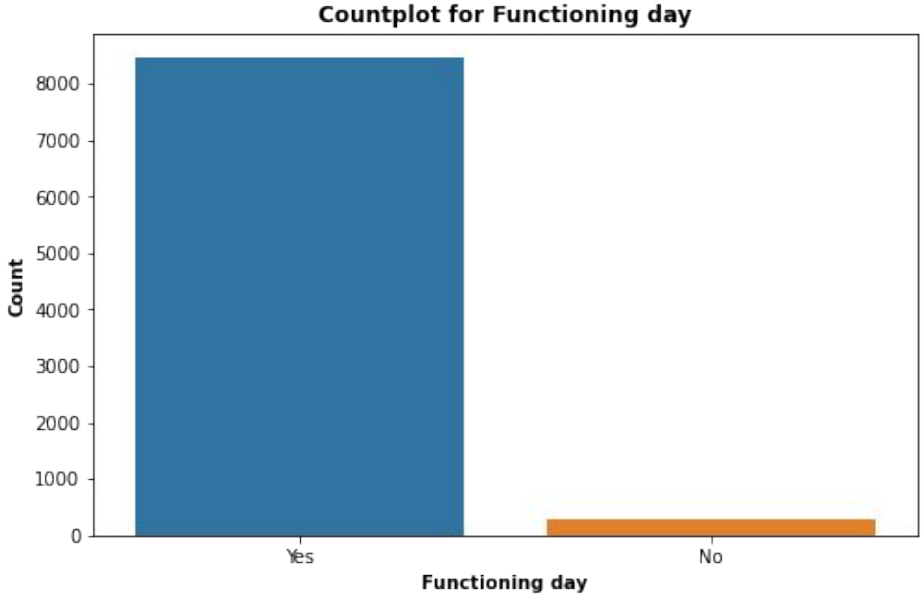
Skewness before : 1.153428
Kurtosis before : 0.853386

Skewness after transformation : 0.237362
Kurtosis after transformation : -0.657201

EDA (Univariate Analysis) - Categorical Variables

Functioning Day & Holiday

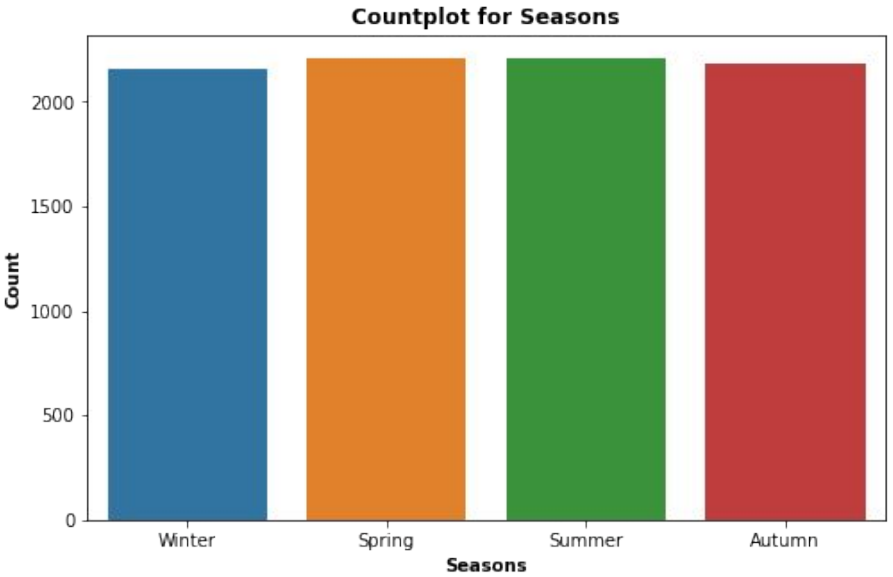


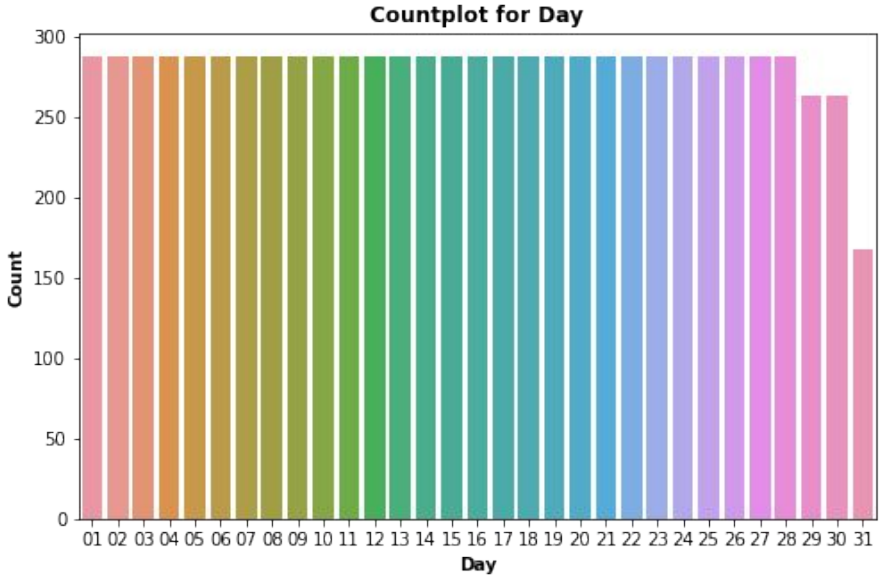


- Functioning day and holiday had majority of one class around 97% and 95%

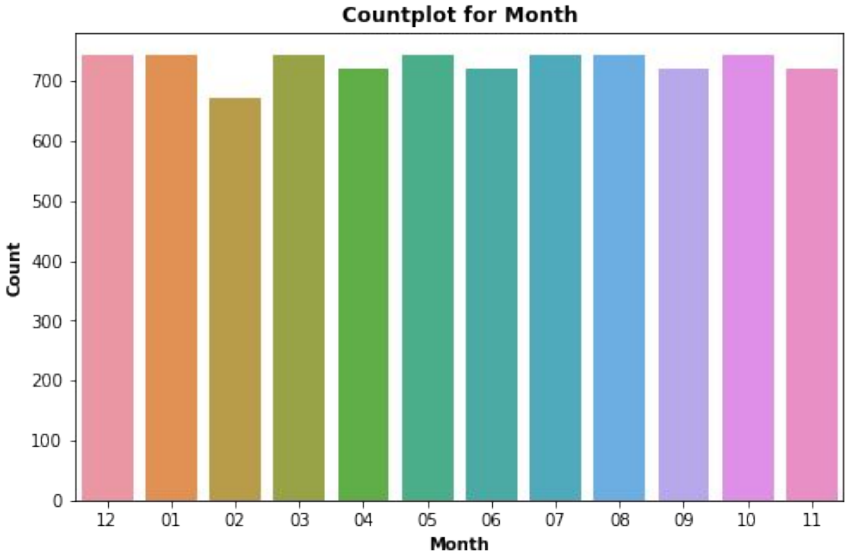
EDA (Univariate Analysis) - Categorical Variables

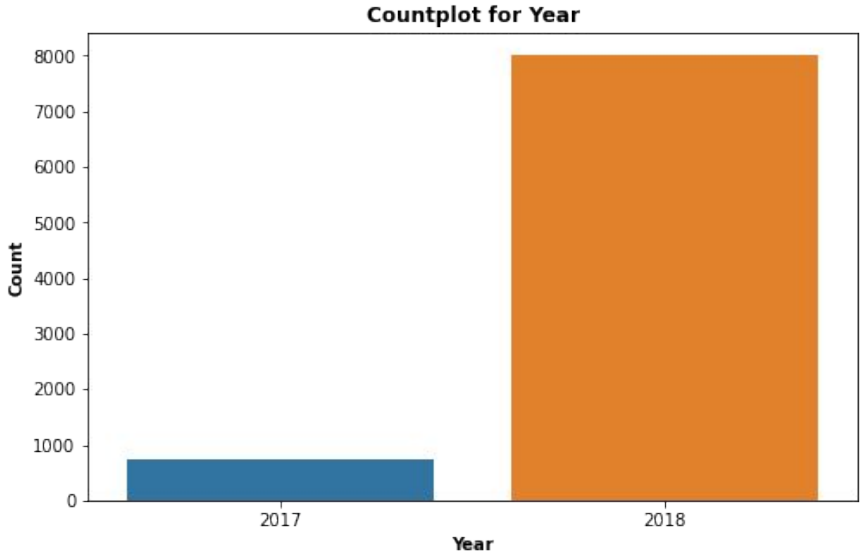
Seasons - Day - Month - Year





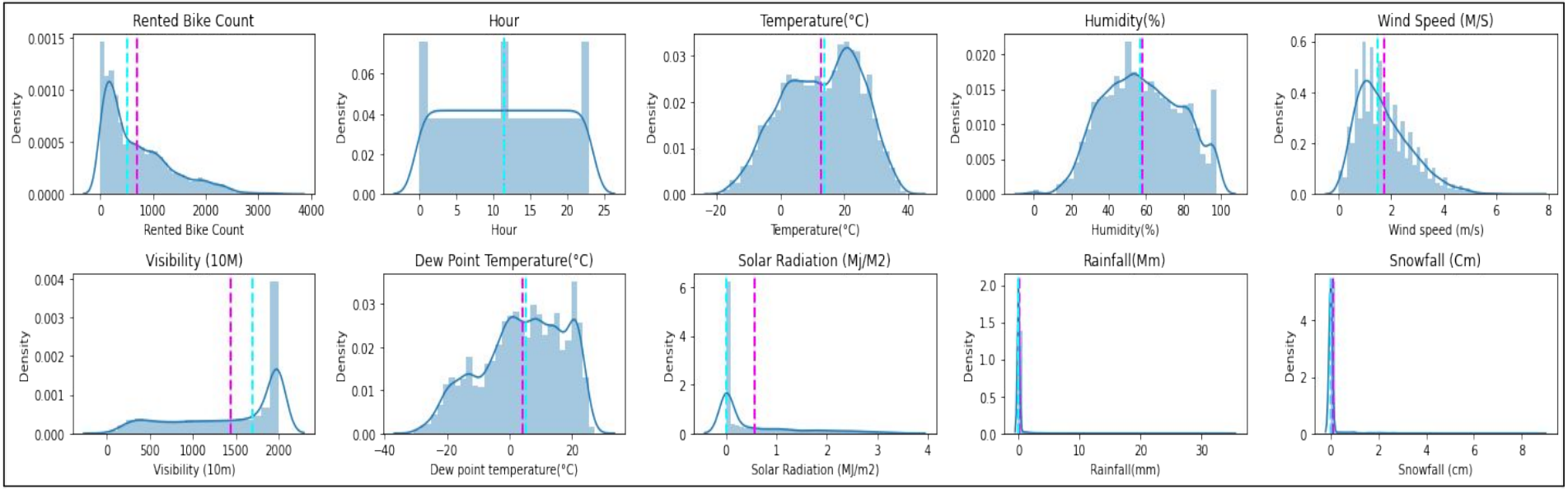
Seasons - Day - Month - Year (Contd.)





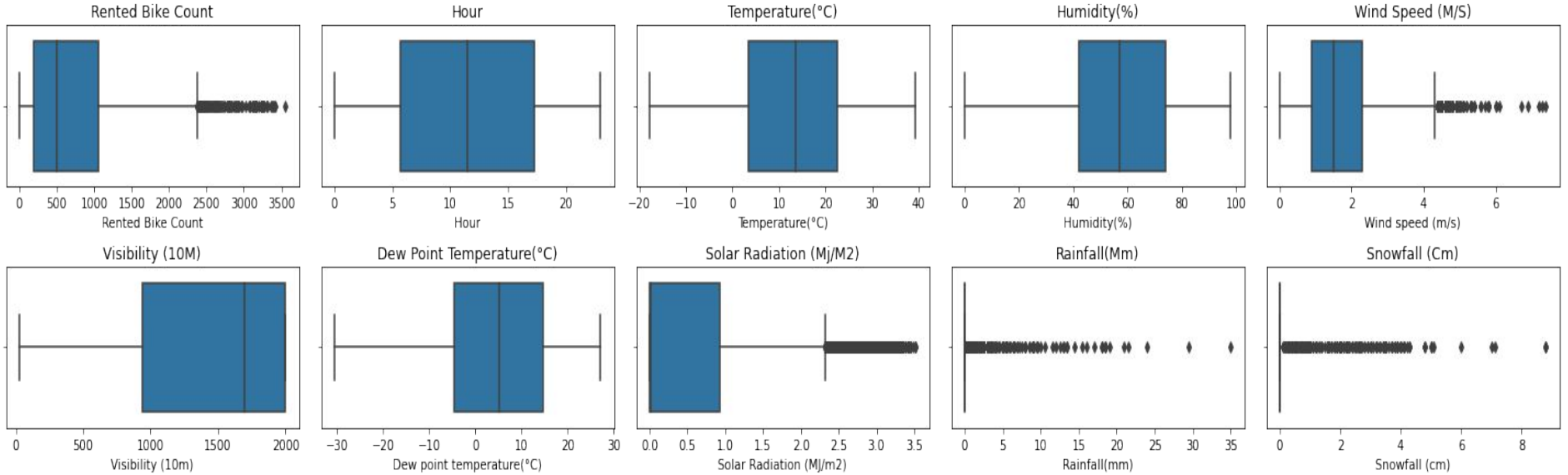
- Seasons, Day, Month had almost equal number of observations.
- Majority of observations were from 2018.

Visualizing Distributions of Numerical Variables



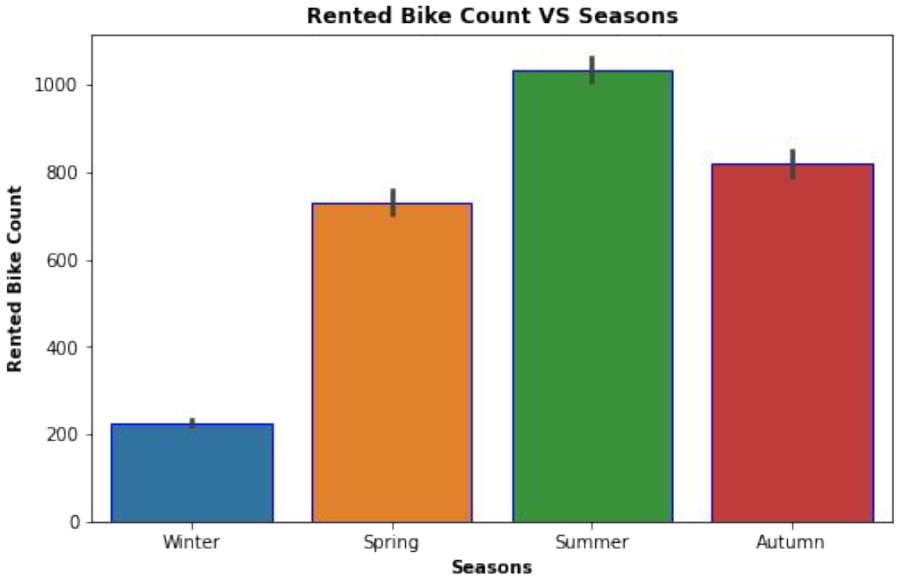
- Variables such as snowfall, rainfall has mostly zero values.

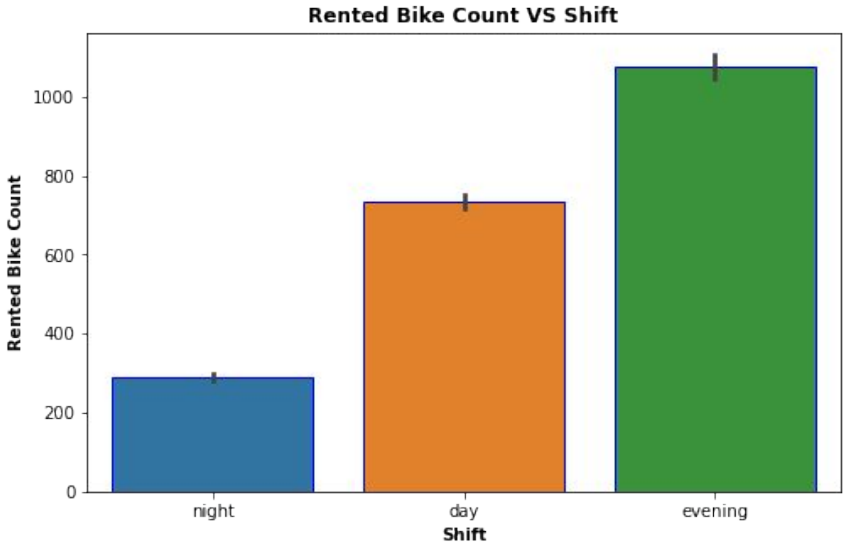
Outliers Analysis of Numerical Variables



- Outliers present in some columns like Solar Radiation , Wind Speed, Rainfall and Snowfall.
- We treated the outliers in the target variable by Transforming it.

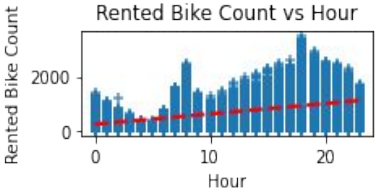
Multivariate Analysis

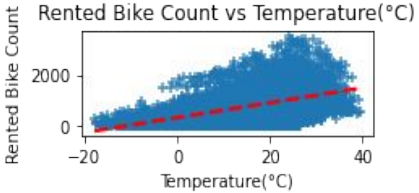


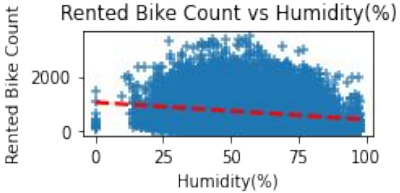


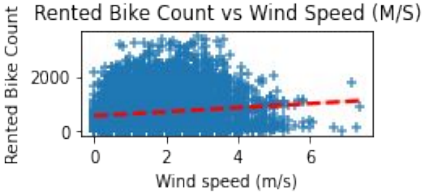
- Demand for bikes was high during summer compared to winter
- In Day and Evening time bike demands are very high.

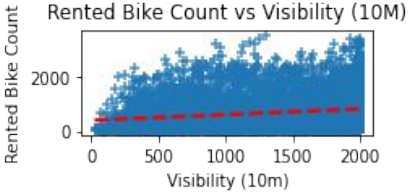
Checking Linearity

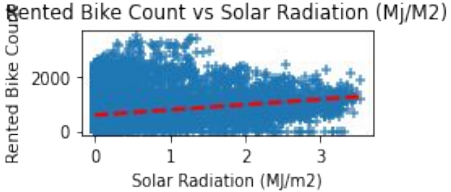


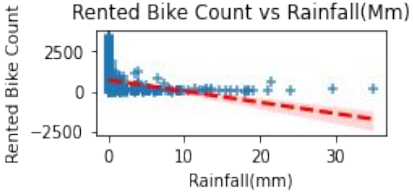


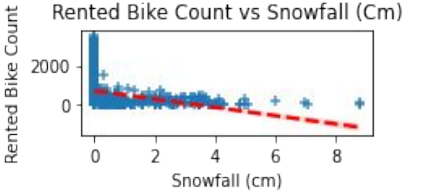


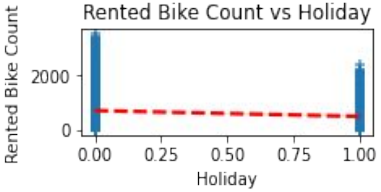


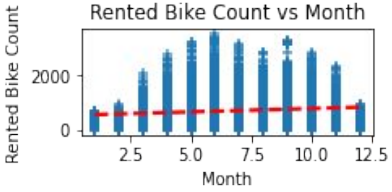


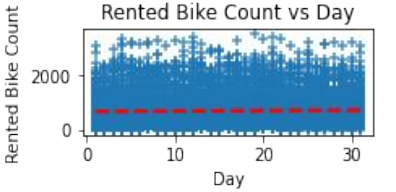


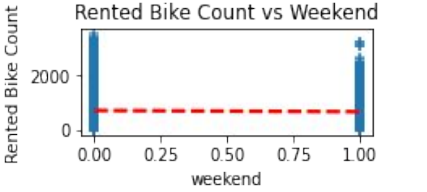


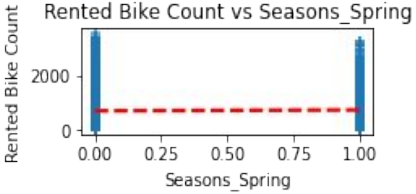




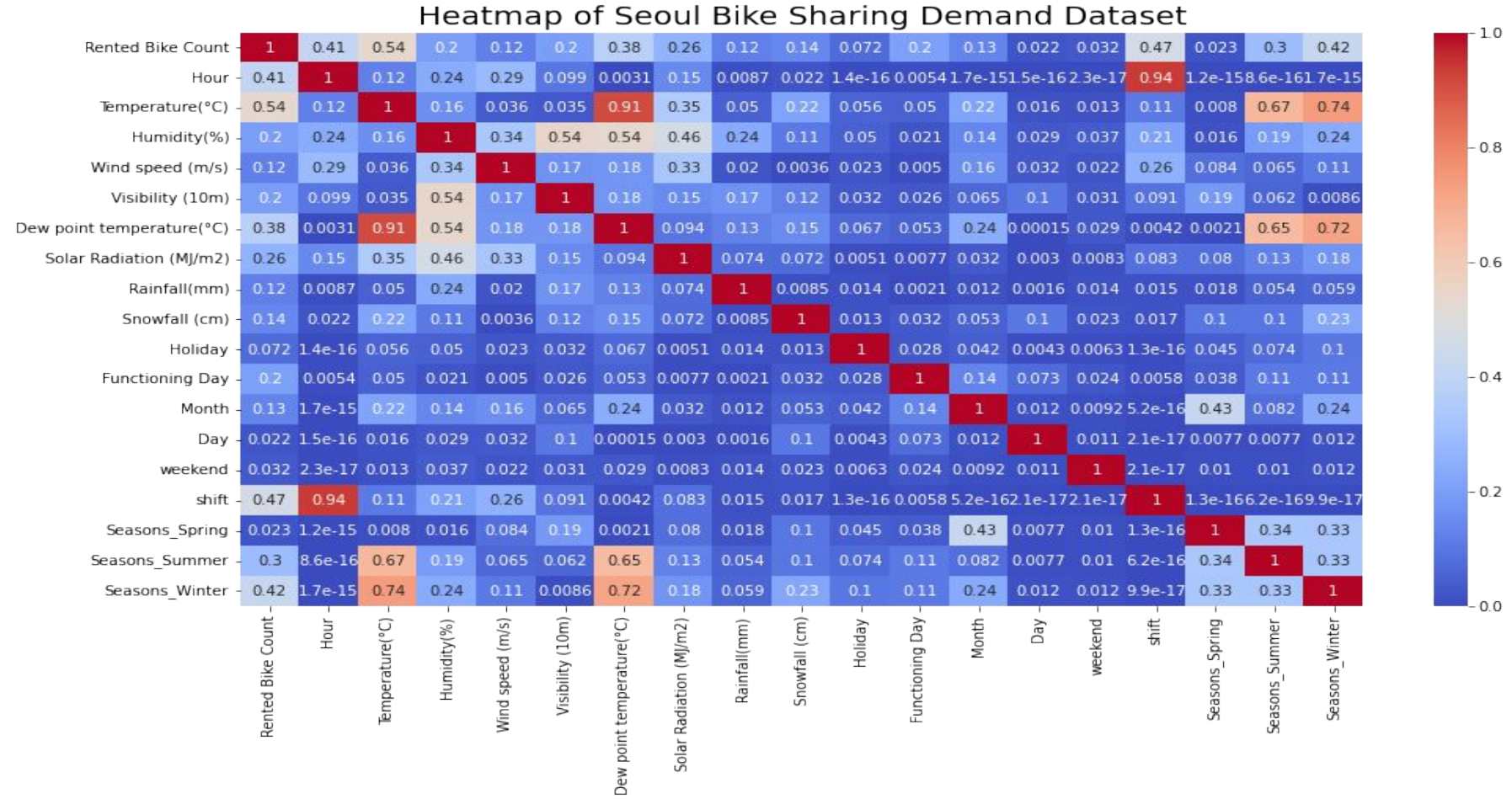








Multicollinearity Analysis

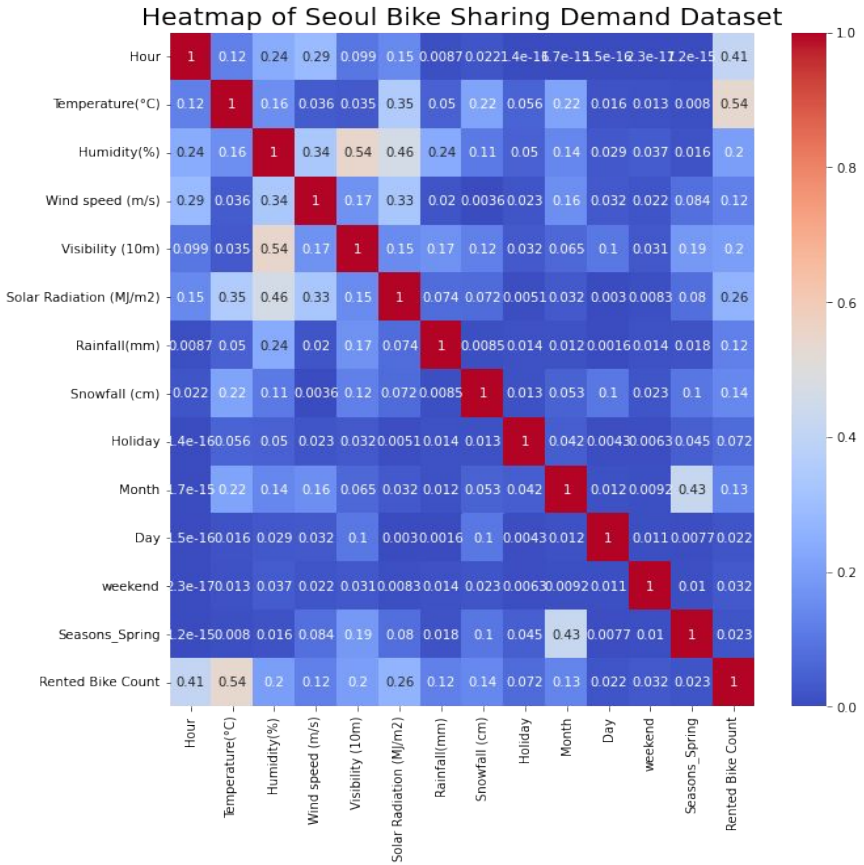


VIF Analysis

	variables	VIF
1	Temperature(°C)	53.971651
5	Dew point temperature(°C)	34.081319
0	Hour	33.015333
10	Functioning Day	30.694040
2	Humidity(%)	27.529223
14	shift	23.217307
4	Visibility (10m)	9.962396
11	Month	8.200745
17	Seasons_Winter	5.020106
3	Wind speed (m/s)	4.932561
12	Day	4.340426
16	Seasons_Summer	3.930104
15	Seasons_Spring	3.539509
6	Solar Radiation (MJ/m2)	2.971858
13	weekend	1.417917
8	Snowfall (cm)	1.165039
7	Rainfall(mm)	1.085800
9	Holiday	1.080635

	variables	VIF
2	Humidity(%)	7.833786
9	Month	5.914015
4	Visibility (10m)	5.544407
3	Wind speed (m/s)	4.707476
0	Hour	4.019816
10	Day	3.812845
1	Temperature(°C)	3.309920
5	Solar Radiation (MJ/m2)	2.335861
12	Seasons_Spring	1.647504
11	weekend	1.401594
7	Snowfall (cm)	1.155082
6	Rainfall(mm)	1.081764
8	Holiday	1.060761

Updated Heatmap



Machine Learning models

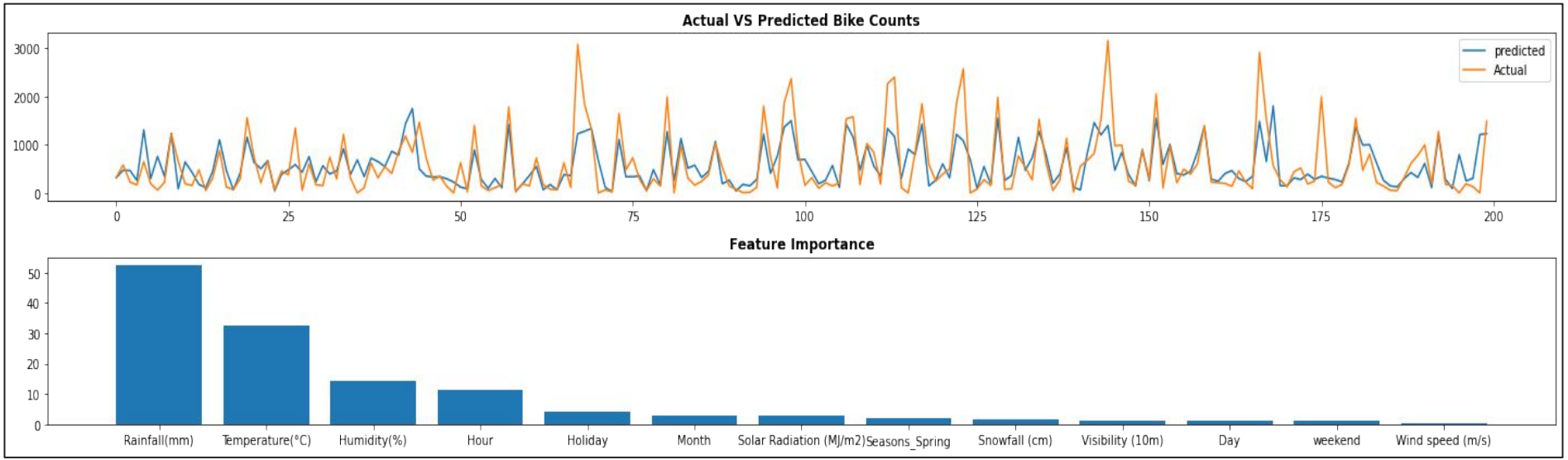
Selection of Models

- Rescale original variables to have equal range or variance.
- **Min - Max Scaler** - Standardized value using this method lies between 0 and 1.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- Since the data contains outliers and many categorical attributes, It won't be good if we fit linear models.
- Tree based algorithm will do good in case of noisy data.
- We can use **Decision tree** as a baseline model. Subsequently, to get better predictions, we can use ensemble models **Random forests**, **GBM**, **XGBoost**.
- Hyperparameter tuning is done to prevent overfitting, and the best parameters are chosen using **GridsearchCV**

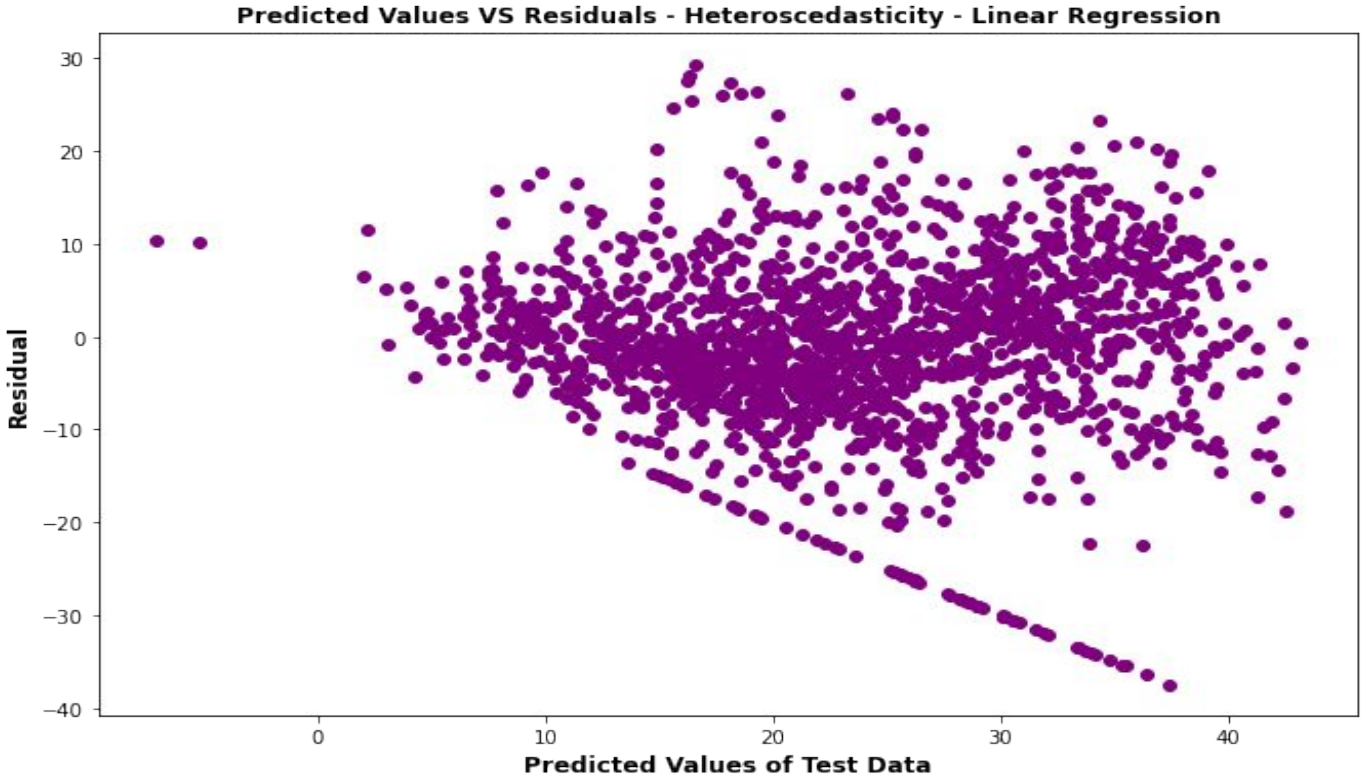
Linear Regression



- MAE of Train set : 316.71
- MSE of Train set : 213511.95
- R2 Score of Train set : 48.58
- RMSE of Train set : 462.07

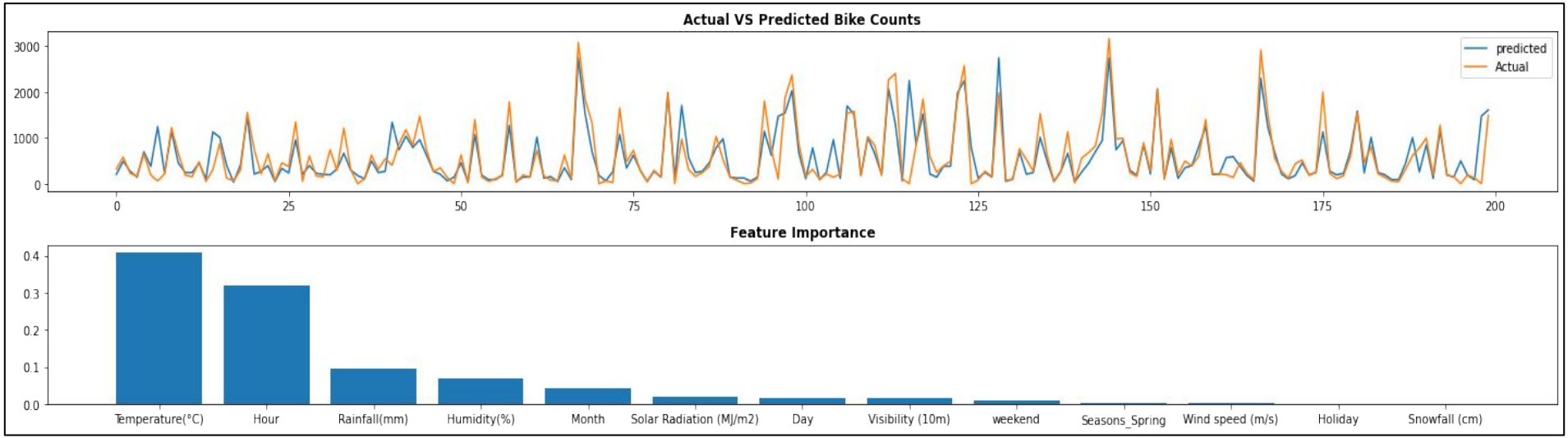
- MAE of Test set : 319.40
- MSE of Test set : 217765.87
- R2 Score of Test set : 47.96
- RMSE of Test set : 466.65

Linear Regression (Contd.)



- Linear Regression model has heteroscedasticity

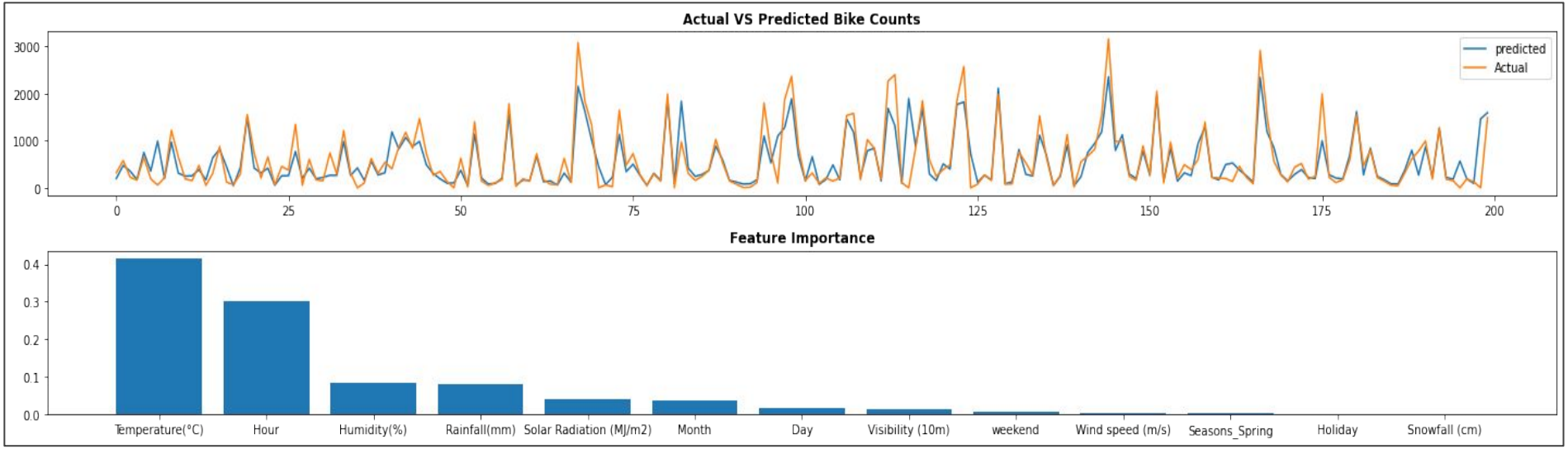
Decision Tree



- MAE of Train set : 191.81
- MSE of Train set : 92371.58
- R2 Score of Train set : 77.75
- RMSE of Train set : 303.92

- MAE of Test set : 221.51
- MSE of Test set : 128158.11
- R2 Score of Test set : 69.37
- RMSE of Test set : 357.99

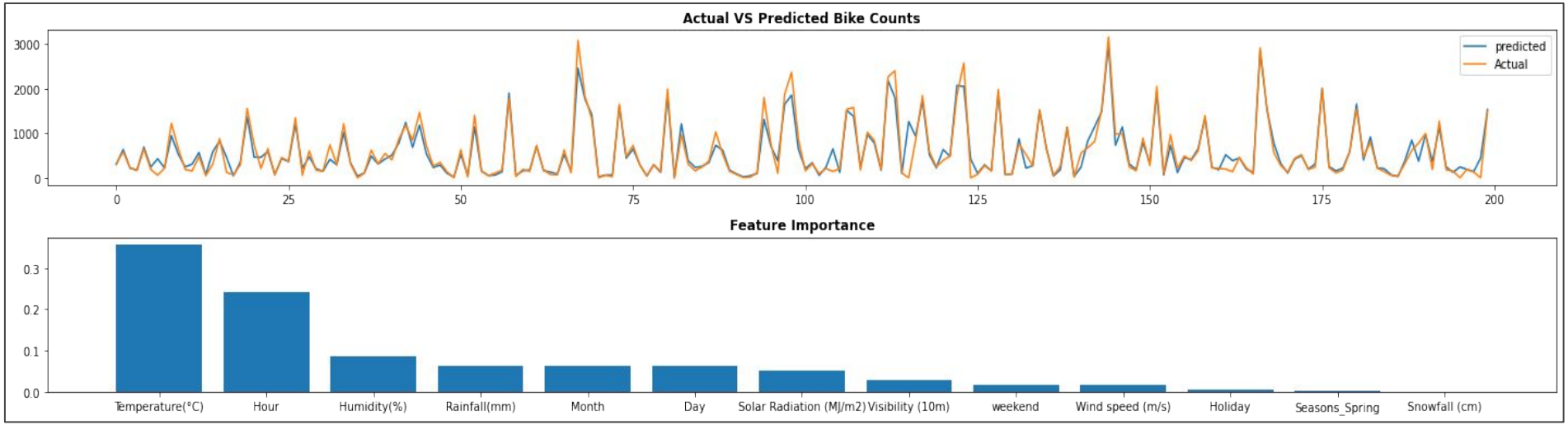
Random Forest



- MAE of Train set : 180.38
- MSE of Train set : 82050.48
- R2 Score of Train set : 80.24
- RMSE of Train set : 286.44

- MAE of Test set : 204.87
- MSE of Test set : 111067.29
- R2 Score of Test set : 73.46
- RMSE of Test set : 333.26

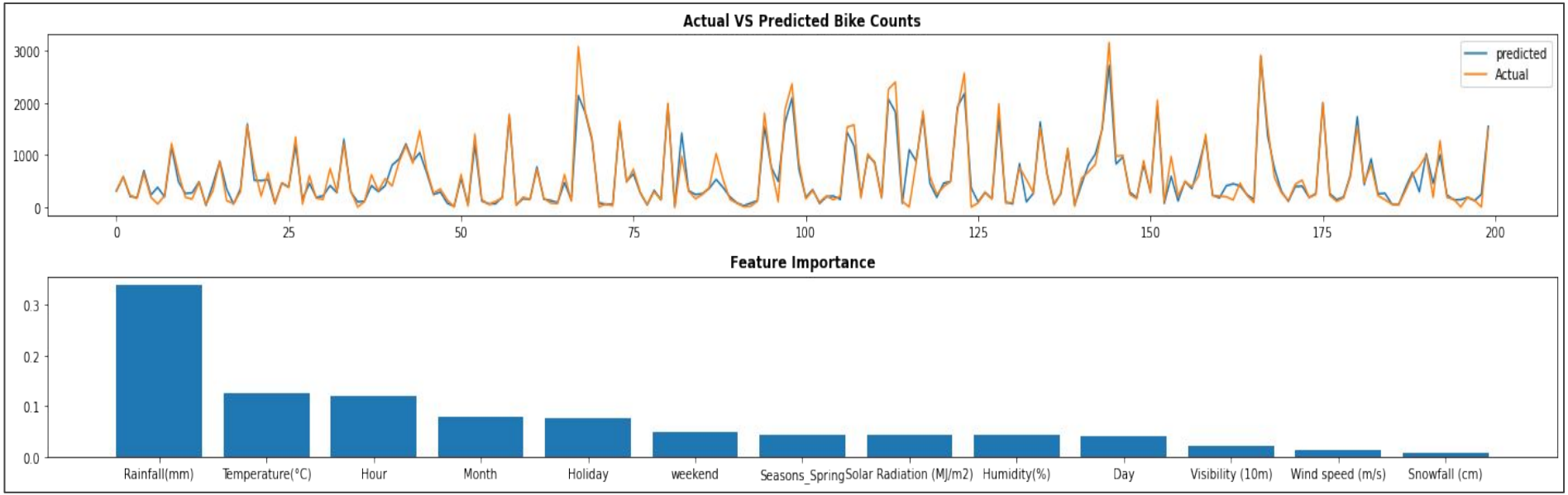
Gradient Boost



- MAE of Train set : 61.63
- MSE of Train set : 10341.46
- R2 Score of Train set : 97.50
- RMSE of Train set : 101.69

- MAE of Test set : 129.39
- MSE of Test set : 47605.83
- R2 Score of Test set : 88.62
- RMSE of Test set : 218.18

XGBoost



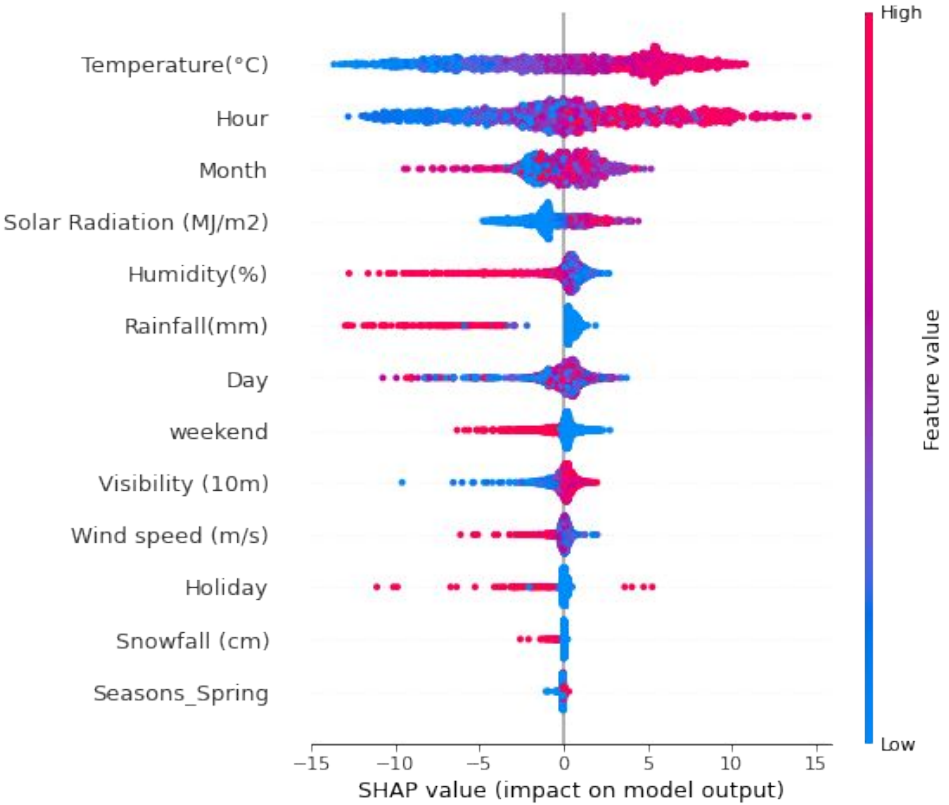
- MAE of Train set : 43.40
- MSE of Train set : 5370.92
- R2 Score of Train set : 98.70
- RMSE of Train set : 73.28

- MAE of Test set : 122.22
- MSE of Test set : 44354.86
- R2 Score of Test set : 89.40
- RMSE of Test set : 210.60

Model Summary

		Model	MAE	MSE	R2 Score	RMSE
Training Dataset Results	0	Linear Regression	316.72	213511.95	0.49	462.07
	1	Decision Tree	191.81	92371.59	0.78	303.93
	2	Random Forest Regressor	180.39	82050.48	0.80	286.44
	3	XGBoost Regressor	43.41	5370.92	0.99	73.29
	4	Gradient Boosting Regressor	61.64	10341.46	0.98	101.69
Test Dataset Results	0	Linear Regression	319.41	217765.88	0.48	466.65
	1	Decision Tree	221.51	128158.12	0.69	357.99
	2	Random Forest Regressor	204.87	111067.30	0.73	333.27
	3	XGBoost Regressor	122.23	44354.86	0.89	210.61
	4	Gradient Boosting Regressor	129.39	47605.84	0.89	218.19

Model Explanation



XGBoost - The most important features are Temperature, Hour, Humidity, Month.

Conclusion

- **Hour :** Demand for bikes is higher during morning hours and evening hours between 15 and 20. This maybe due to normal environment conditions when there is less heat and also starting and ending office hours.
- **Temperature :** People generally prefer to bike at moderate to high temperatures. We see highest rental counts between 32 to 35 degrees Celsius.
- **Humidity :** With increasing humidity, we see decrease in the number of bike rental count.
- **Weather :** As one would expect, we see highest number of bike rentals on a clear day and the lowest on a snowy or rainy day.
- **Windspeed :** Demand is high when wind speed is less.
- **Season :** Demand is high during summer time of the year.
- **Functioning VS Non Functioning Days :** Demand is very high on functioning days.
- As we can interpret from **machine learning models**, If we want **accuracy** - **XG Boost** has the lowest RMSE, MSE, MAE among all and with a R2 score of **0.89** which is highest among all, so we can consider XGBoost model as the best model.

Challenges Faced

- Pre - processing the data was one of the challenges we faced which includes removing highly correlated variables from the data so as to not hinder the performance of our regression model.
- Exploring all the columns and calculating VIF for multicollinearity was challenging it might decrease the models performance.
- Selecting the appropriate models to maximize the accuracy of our predictions was one of the challenges faced.

References

- MachineLearningMastery
- GeeksforGeeks
- Analytics Vidhya
- TowardsDataScience
- Stack Overflow

Thank You!