Capstone Project

Seoul Bike Sharing Demand Prediction

Team members

Kajal Dhun

Navinkumar Sambari

Tanu Rajput

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.

Content

- Data description
- Exploratory data analysis
- Correlation Analysis
- Multicollinearity Detection
- All models Evaluation Metrics
- Model Selection
- Challenges faced
- Conclusion

Data Description

Seoul Bike Data

Date and Time	Weather	Others
Date	Temperature	Rented Bike Count
Hour	Humidity	Holiday
	Dew Point Temperature	Functional Day
	Visibility	
	Snowfall	
	Rainfall	
	Windspeed	
	Solar Radiation	
	Seasons	

A glance at the dataset

AI

- This dataset contains 8760 rows and 14 columns.
- There are no null values in any feature.
- If we observe the date column, in the dataset, it begins from 1-12-2017 to 30-11-2018. That means, we have exact 1 year of seoul bike sharing demand data.
- From 14 features our target feature is Rented Bike Count and rest are independent features.

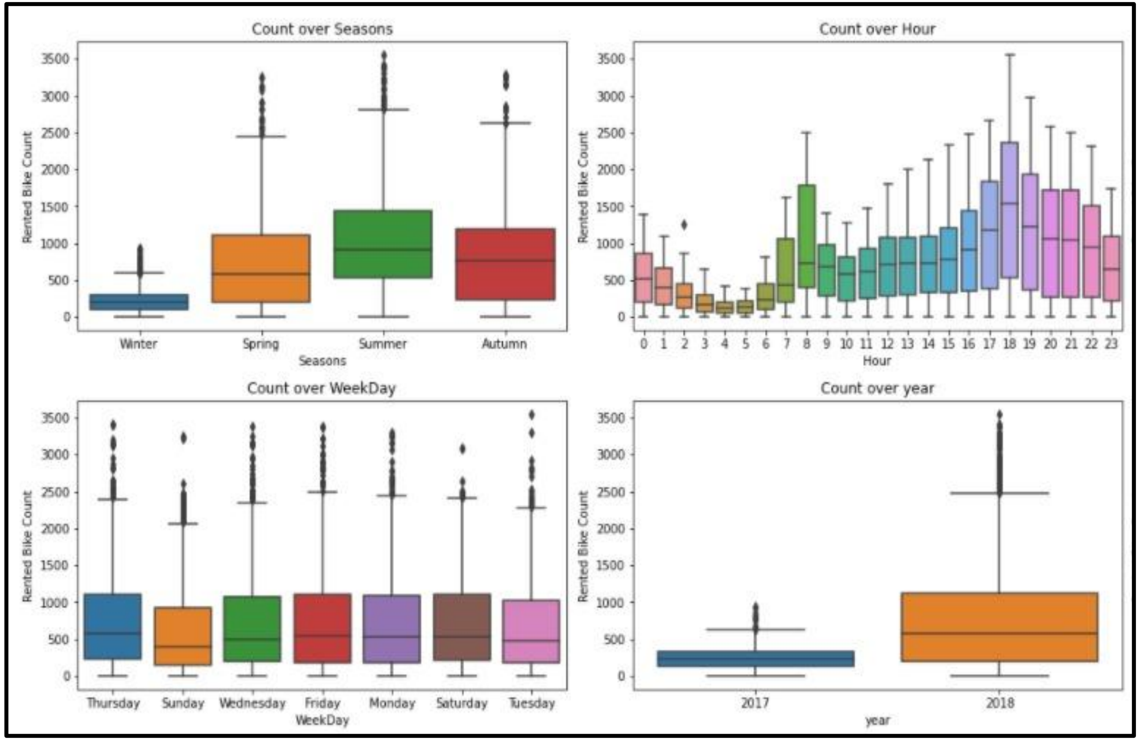
```
# Data information
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8760 entries, 0 to 8759
Data columns (total 14 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Date                       8760 non-null   object
 1   Rented Bike Count          8760 non-null   int64
 2   Hour                       8760 non-null   int64
 3   Temperature(°C)            8760 non-null   float64
 4   Humidity(%)                8760 non-null   int64
 5   Wind speed (m/s)           8760 non-null   float64
 6   Visibility (10m)           8760 non-null   int64
 7   Dew point temperature(°C)  8760 non-null   float64
 8   Solar Radiation (MJ/m2)    8760 non-null   float64
 9   Rainfall(mm)               8760 non-null   float64
 10  Snowfall (cm)              8760 non-null   float64
 11  Seasons                    8760 non-null   object
 12  Holiday                    8760 non-null   object
 13  Functioning Day            8760 non-null   object
dtypes: float64(6), int64(4), object(4)
memory usage: 958.2+ KB
```

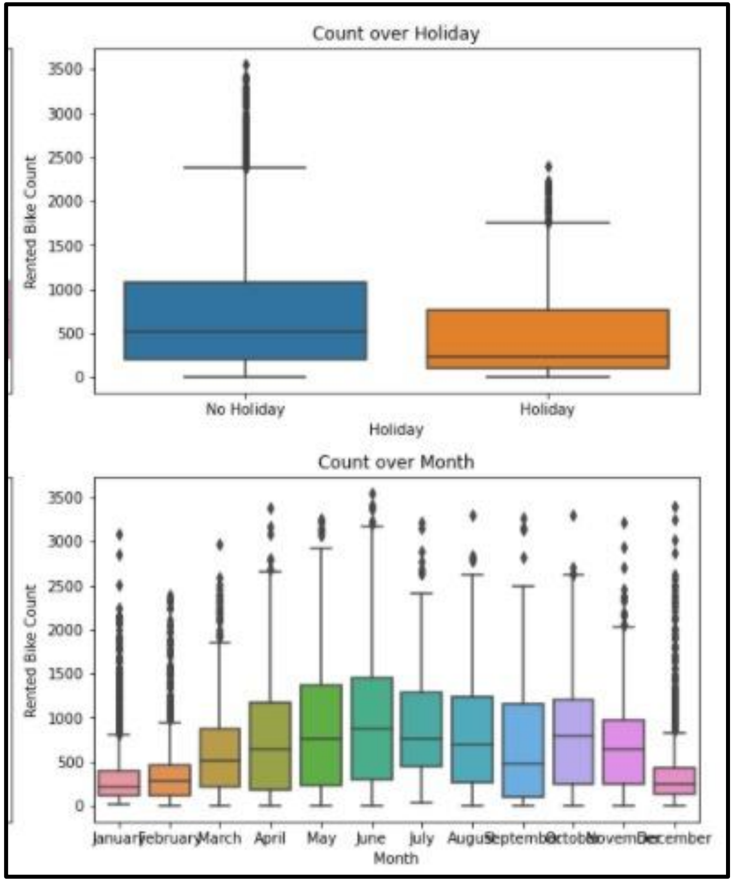
Exploratory Data Analysis

Boxplots on Rental Bike Count

- In Count over Seasons, the demand for bike in Winter is less than compare to summer and other seasons
- In Count over Hour, if we observe during the day, the demand for bikes is high from morning 8 am and from evening 6pm

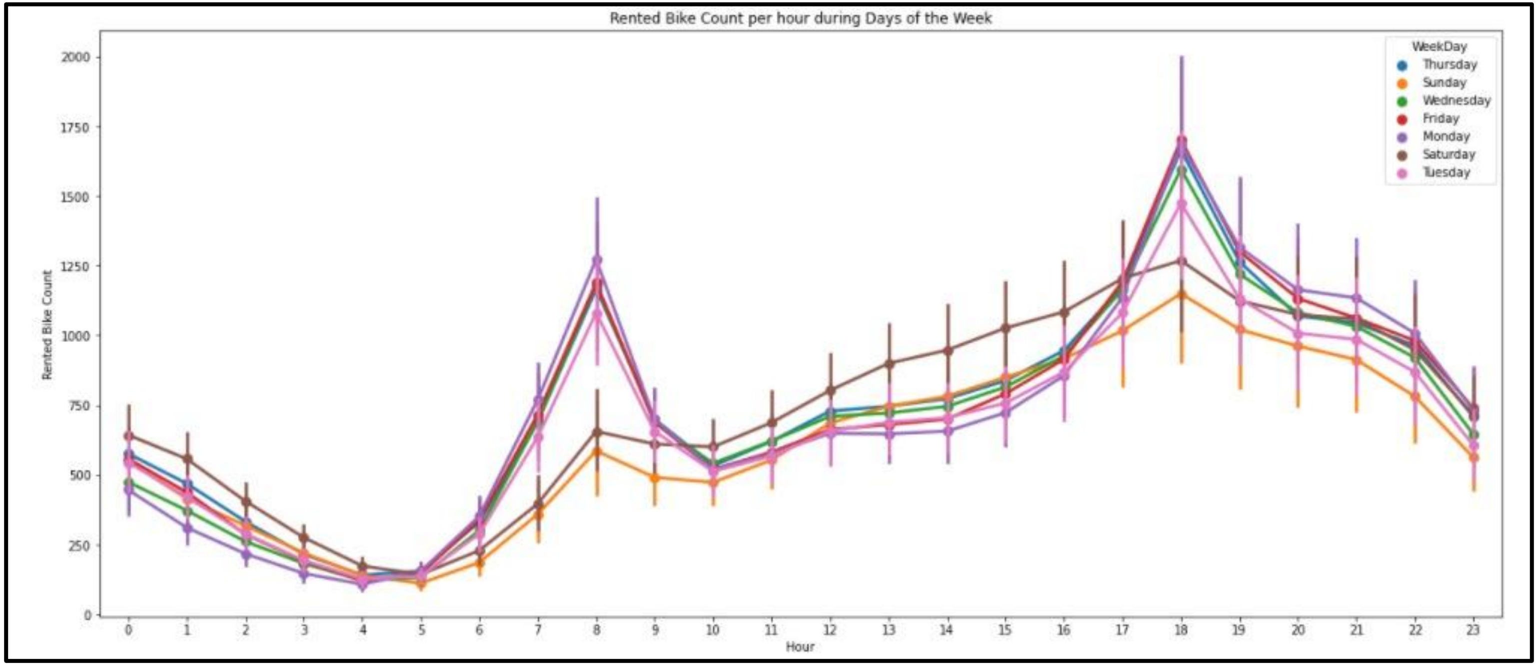


- Demand for the rented bike during No Holiday is higher than the Holiday.
- Now In Count over Month graph, if we observe carefully, the demand for the bike is lesser in the months which are December, January, February as at that time it is the winter season.
- In the months such as April, May, June , the demand for bike is higher because these months are fall in Summer seasons.



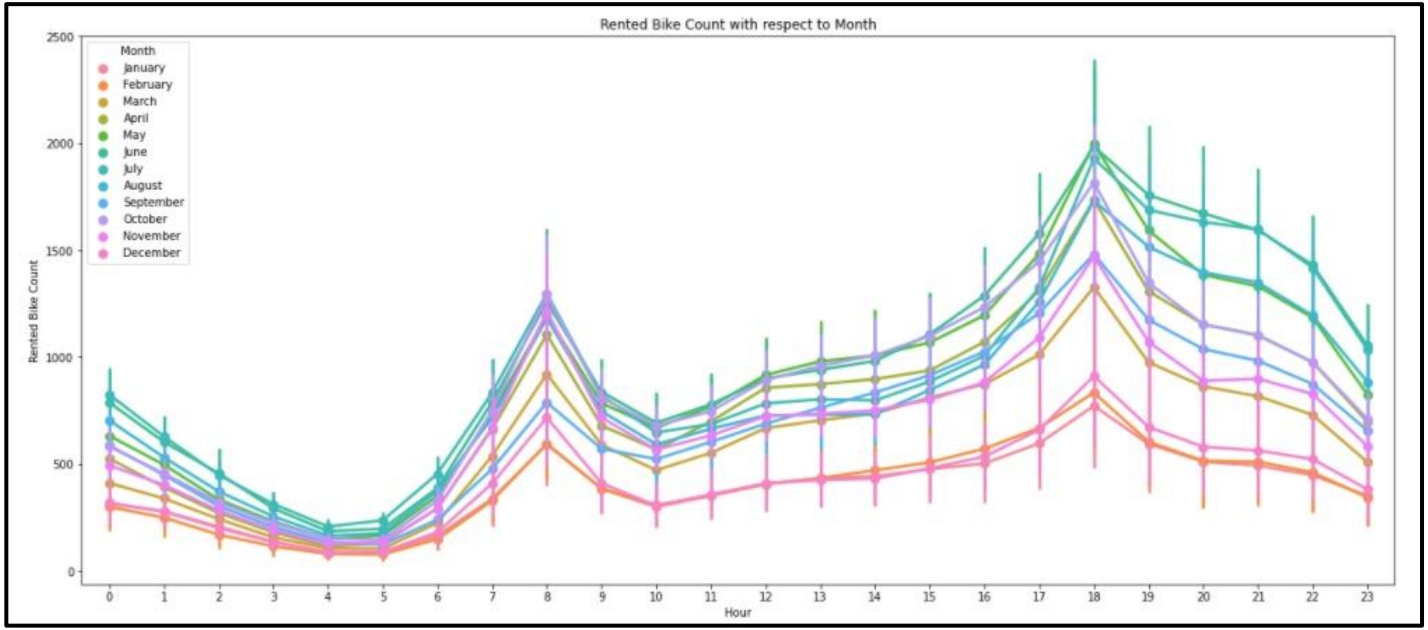
Rented Bike Count per hour during "Weekdays"

- ➢ Here, From Monday to Friday we consider as a Weekdays while Saturday, Sunday considered as Weekends.
- ➢ If we closely look into this pointplot, either its weekdays or weekend, the demand for rented bike count approx starts from morning 6 am. At 8am it is high and also from 6pm.
- ➢ The bike count is high in weekdays than weekend



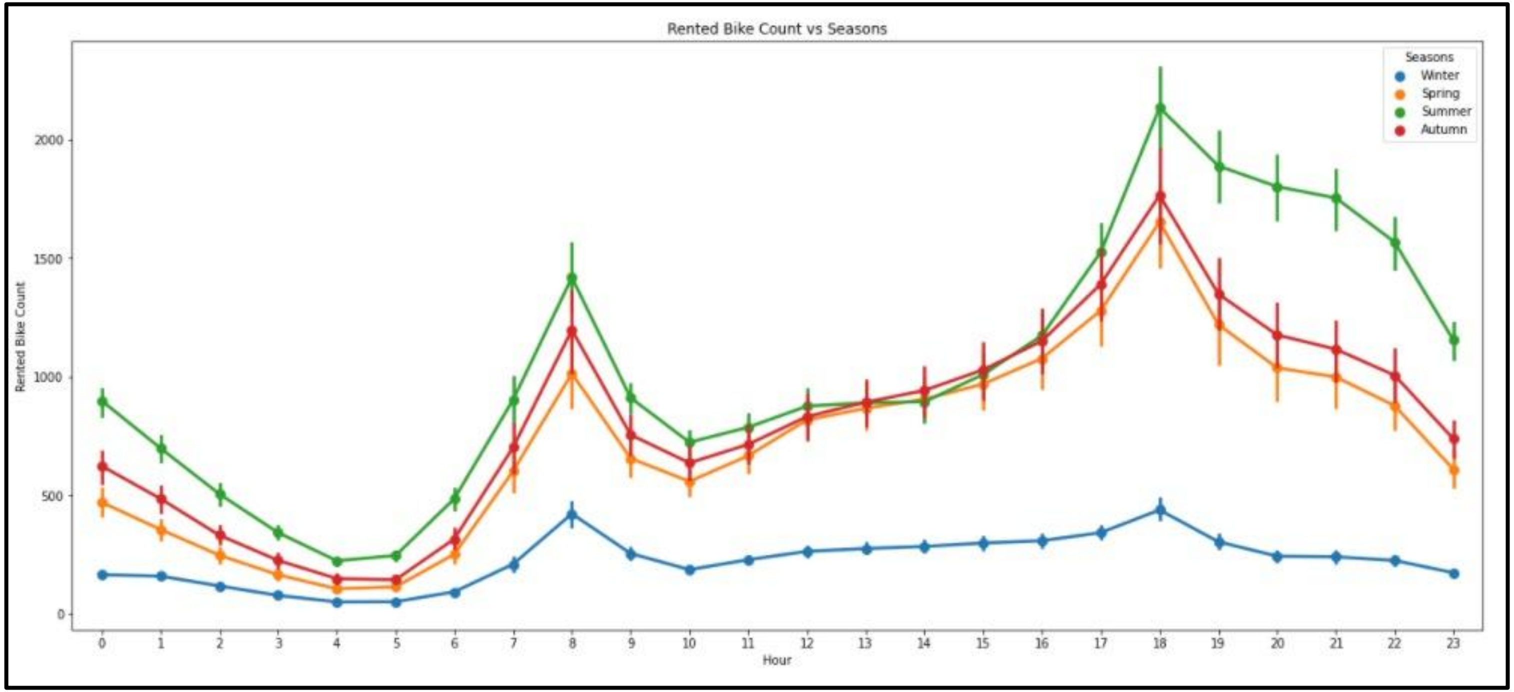
Rented Bike Count per hour wrt. "Month"

- **By observing, we get to know that in the month of December, January, February the demand for bike is less due to cold weather.**
- **Although the pattern is same with respect to hour, as demand gets peak at 8am and 6am.**



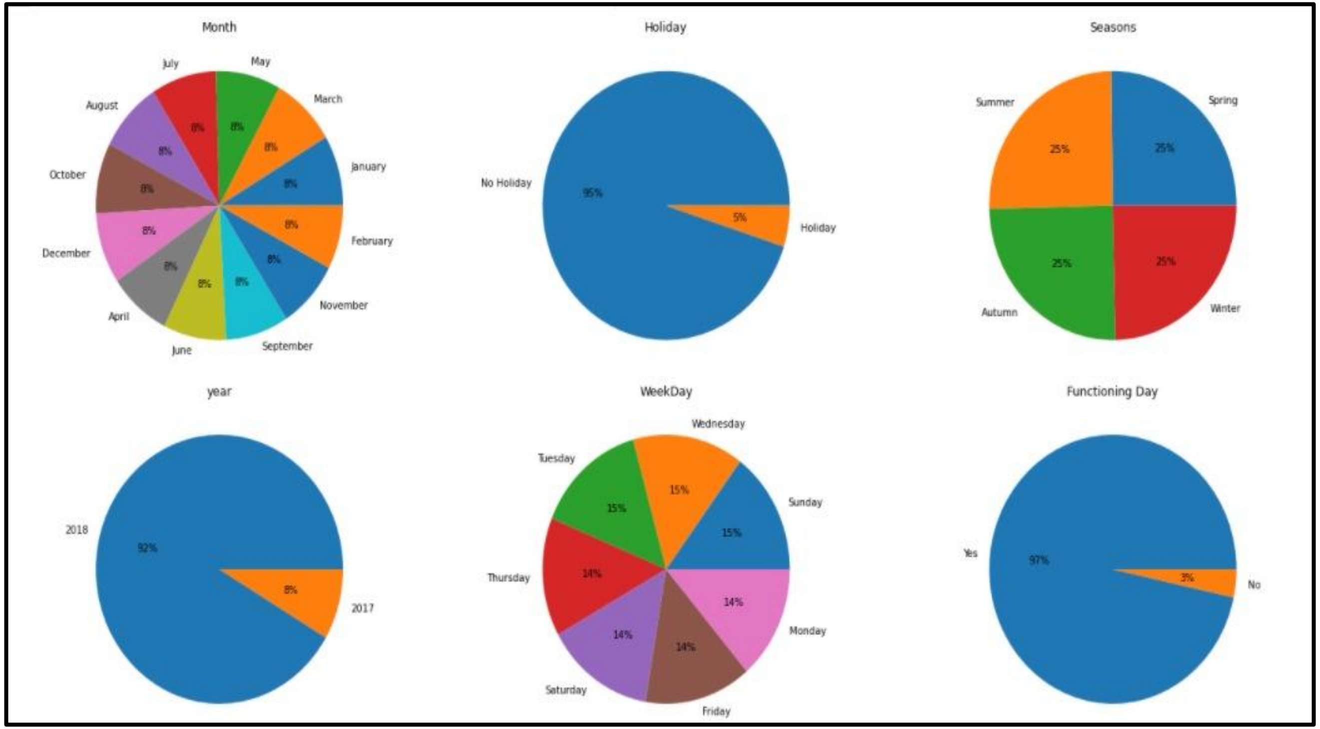
Rented Bike Count per hour wrt. “Seasons”

➢ We already seen before in boxplots, that the demand for bike in summer is high and in winter is low.



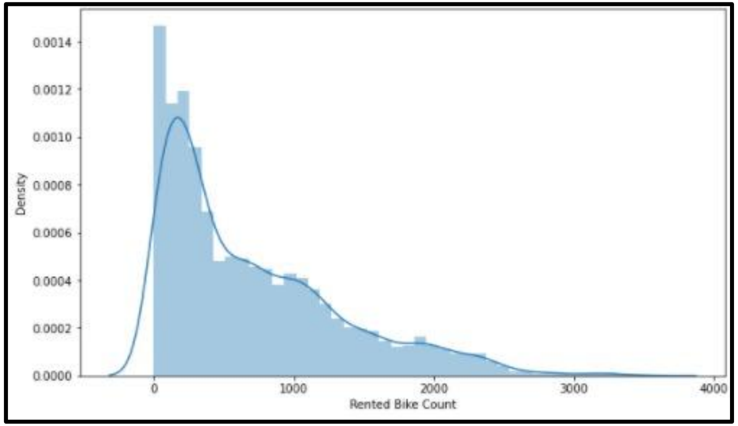
Visualizing % data distribution of Categorical features

- ➢ Month feature is equally distributed.
- ➢ In holiday features, No holiday is 95% distributed and 5% of holiday
- ➢ In season column, all season labels is 25% distributed equally.
- ➢ In year column,
  2017 = 8%
  2018 = 92%
- ➢ Functioning day,
  Yes = 97%
  No = 03%

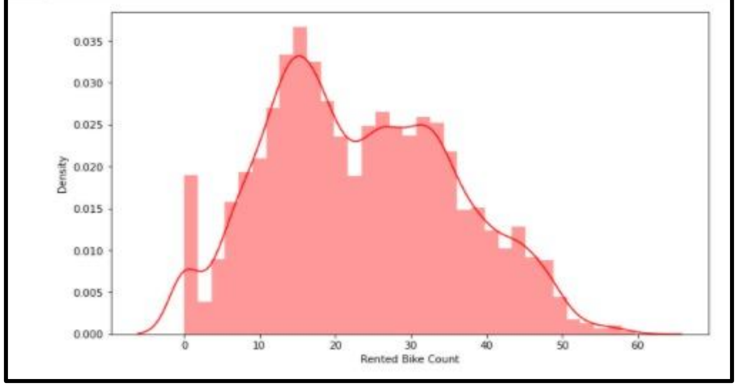


Distribution of Target Column

- ➢ The shape of the Rented Bike Count feature is RIGHTLY SKEWED.
- ➢ We have to transform this distribution into approx normal distribution using appropriate transformation techniques.
- ➢ We used square root transformation, as it transforming this skewed distribution into normal.

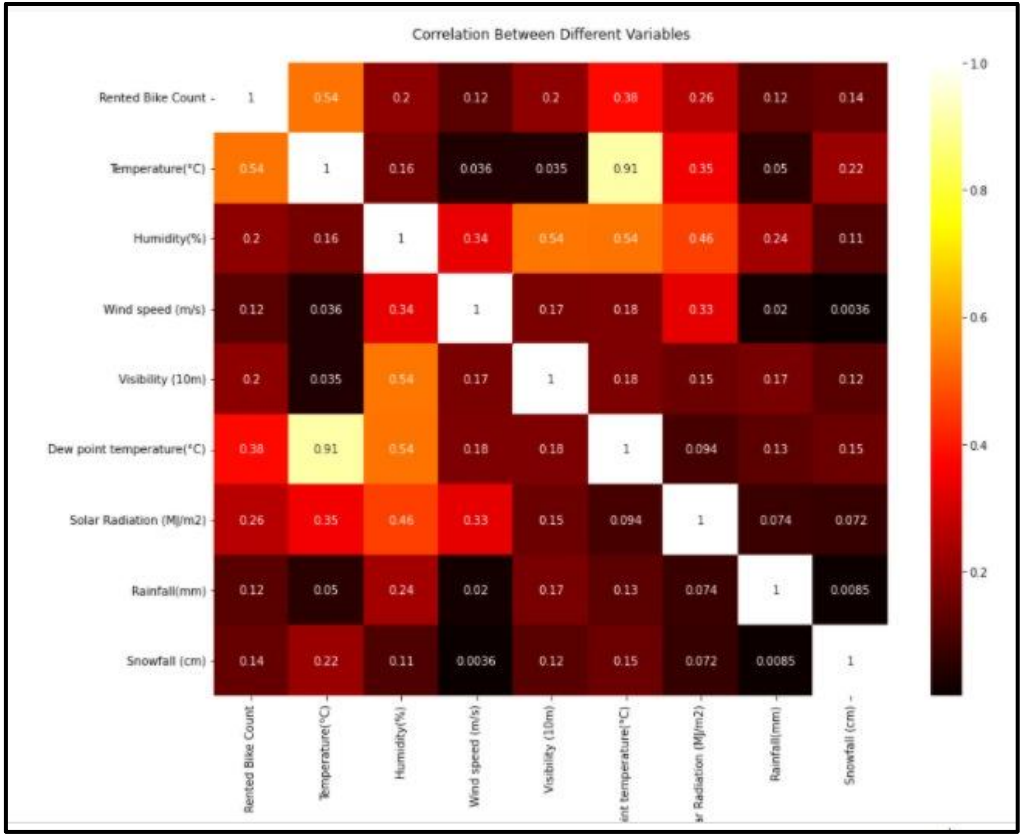



```
sns.distplot(np.sqrt(df['Rented Bike Count']),color='r')
plt.show()
```



Correlation matrix

- ➢ From this correlation matrix, we can easily say that the Temperature and Dew point has higher correlation between them, i.e., 0.91 which is good but it will badly affect while training the model and doing prediction.
- ➢ This type of high correlation is also called as multicollinearity.
- ➢ We used VIF technique to detect multicollinearity separately and then we decided to remove one of the column which is Dew Point Temperature.



Multicollinearity Detection

	feature	Variance Inflation Factor
0	Temperature(°C)	29.075866
1	Dew point temperature(°C)	15.201989
2	Humidity(%)	5.069743
3	Wind speed (m/s)	4.517664
4	Visibility (10m)	9.051931
5	Solar Radiation (MJ/m2)	2.821604
6	Rainfall(mm)	1.079919
7	Snowfall (cm)	1.118903

Before

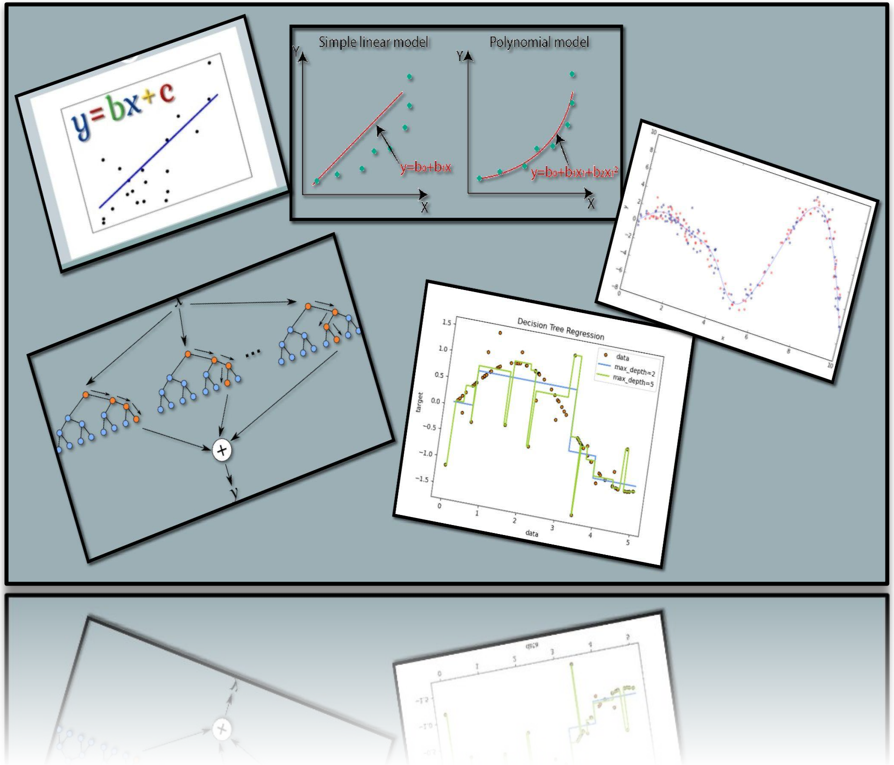
	feature	Variance Inflation Factor
0	Temperature(°C)	3.166007
1	Humidity(%)	4.758651
2	Wind speed (m/s)	4.079926
3	Visibility (10m)	4.409448
4	Solar Radiation (MJ/m2)	2.246238
5	Rainfall(mm)	1.078501
6	Snowfall (cm)	1.118901

After

Modeling

List of algorithms used:

- Linear Regression
- Ridge
- Lasso
- Polynomial
- Decision Tree Regressor
- Random Forest Regressor
- Gradient Boosted Regressor
- Extra Trees Regressor



Metrics dataframe of all models before Hyperparameter tuning:

	TRAININGSCORE	MSE	RMSE	R2	ADJ_R2
Linear Regression	0.795529	30.516904	5.524211	0.798363	0.793360
Ridge	0.663220	49.511948	7.036473	0.672855	0.664738
Lasso	0.795528	30.517397	5.524255	0.798359	0.793356
Decision_Tree	1.000000	28.869927	5.373074	0.809245	0.804512
Random_Forest	0.986000	14.660766	3.828938	0.903131	0.900727
Gradient_boost	0.862500	22.563778	4.750135	0.850912	0.847213
ExtraTreeReg	1.000000	13.136375	3.624414	0.913203	0.911049

➔ RandomForest, Gradient Boost and ExtraTreesReg giving best ADJ_R2 score. But there are overfitting in them. So, Hyperparameter tuning is must.

Metrics dataframe of all models after Hyperparameter tuning:

	TRAININGSCORE(ht)	MSE(ht)	RMSE(ht)	R2(ht)	ADJ_R2(ht)
ExtraTreesReg	0.987206	13.383054	3.658286	0.911573	0.909379
Random_Forest	0.918456	17.557124	4.190122	0.883993	0.881115
Decision_Tree	0.891449	23.800368	4.878562	0.842742	0.838840
Gradient_boost	0.829335	26.876589	5.184264	0.822416	0.818009
Linear Regression	0.764982	110046.223964	331.732157	0.731070	0.724398
Lasso	0.764974	110233.803977	332.014765	0.730612	0.723928
Ridge	0.764959	110260.060533	332.054304	0.730548	0.723862
Polynomial	0.935143	45983.513185	214.437667	0.887626	0.650587

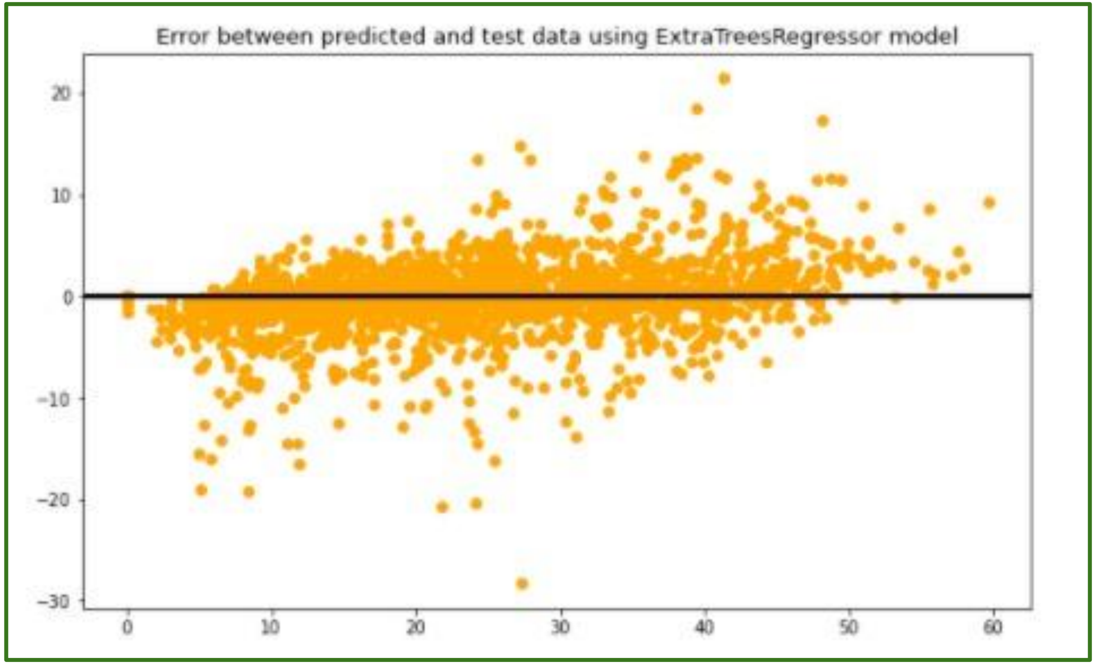
➔ **Even after hyperparameter tuning, dataframe showing the three best models, viz., ExtraTreesReg, Random_Forest, Decision_Tree**

Deciding best Model Selection:

1) According to Model Evaluation metrics dataframe, Linear Regression and polynomial is not giving best results.
2) Decision Tree & Gradient Boost have performed approximately equally good in terms of ADJ_R2 and R2.
3) So, the best results that we getting from RandomForest and ExtraTreesRegressor.
4) But we are selecting the ExtraTreesRegressor for model selection and prediction

Visualizing the error of a best model:

- After seeing all model's error b/w test and predicted data. So, among all of them, extratreesregressor gives less error compare to others.
- So this is the error scatterplot by using ExtraTreesRegressor.



Challenges faced:

- We felt little challenging when we start working on different algorithms and its metrics, choosing quite number of algorithms to work upon.
- As dataset was quite big enough which led more computation time.
- Also, deciding about the best model for prediction.

Conclusion:

1) We observed that bike rental count is high during weekdays than weekend days.
2) The rental bike counts is at its peak at 8 AM in the morning and 6pm in the evening.
3) We observed that people prefer to rent bikes during moderate to high temperature.
4) Highest rental bike count is during Autumn and summer seasons and the lowest in winter season.
5) Comparing the Adjusted R2 among all the models, ExtraTreesRegressor gives the highest Adjusted R2 score that is 0.908699 and Training score is 0.987167. Therefore, this model is the best for predicting the bike rental count on hour basis

THANK YOU