Capstone Project

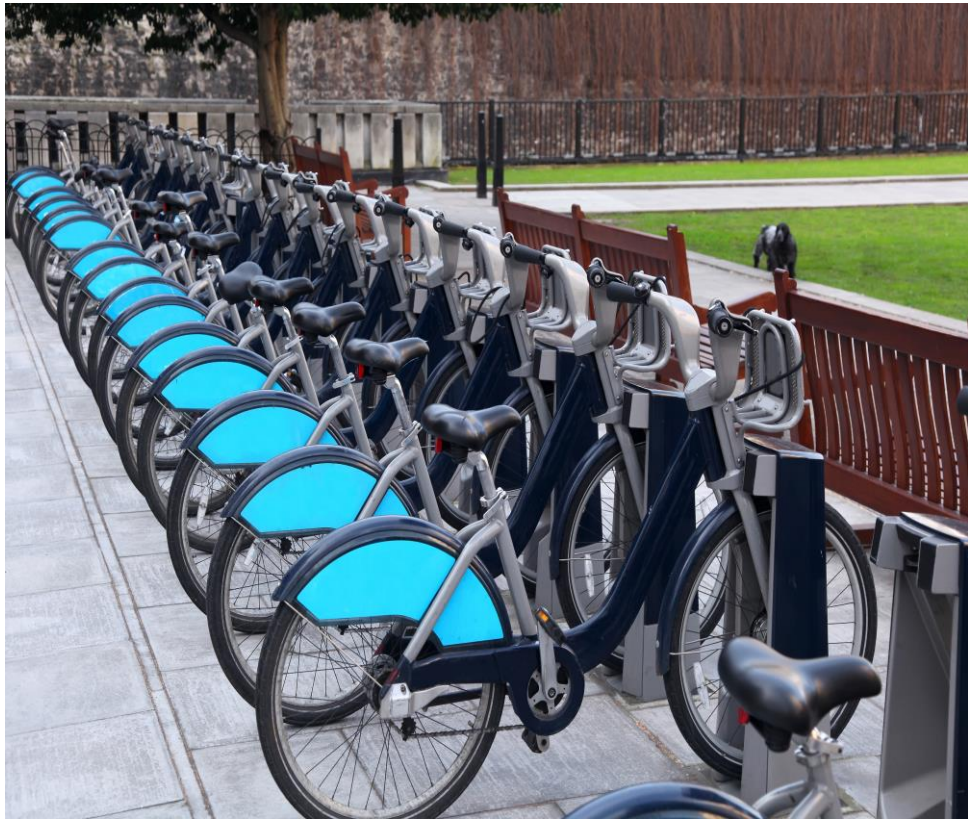
Seoul Bike sharing- Prediction

Team members
Mohd Sharik

Introduction

- Seoul's bike sharing system was set up in 2015 . This service was first introduced in Seoul in October 2015 in select areas of the right bank of the Han River.

- Seoul’s bike sharing service can be used by downloading the Seoul Bike smartphone application and by registering or logging in with Naver, Facebook or Kakao accounts. The app is available in both Korean and English.



Problem Statement

- The problem statement that we are given is as follows-

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.

Project Overview

We have divided this entire project in 7 sections –

1. Loading and reading the dataset :

In this section we simply have loaded the dataset and take a look on the basic info. Such as features, data types and data summary.

2. Univariate analysis :

In this section we have performed univariate analysis by plotting distribution plot and have observed many data insights

3. Bi-Variate analysis

in this Section we will look at bi-variate analysis of features we will majorly focus on the relation between target column and other features.

continued……..

4. Predictions and it's results:

In this section we will apply Suitable algorithm. And finally we will compare the results with actual values.

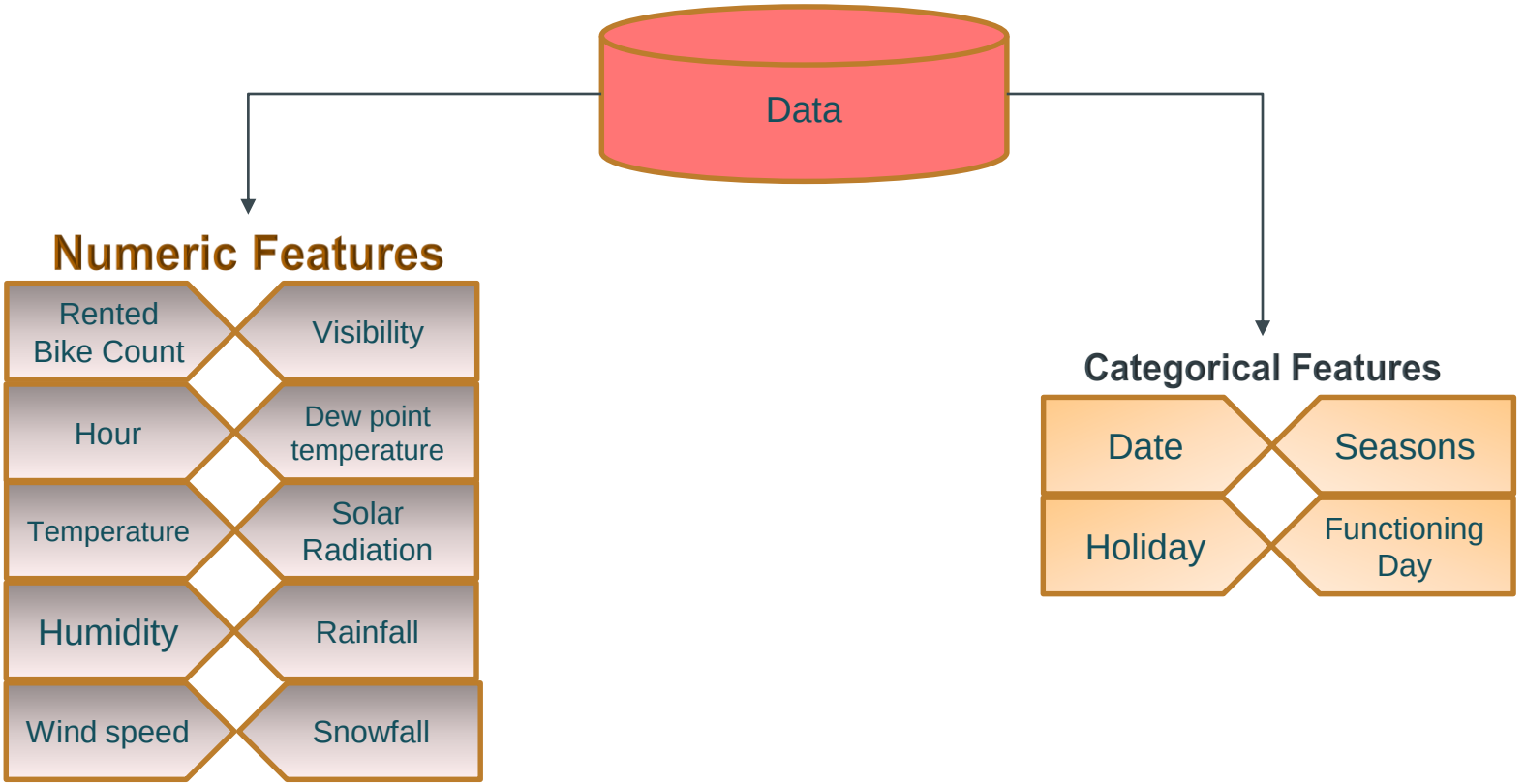
We will also talk about feature importance.

5. Challenges: In this section we will talk about challenges that we faced.

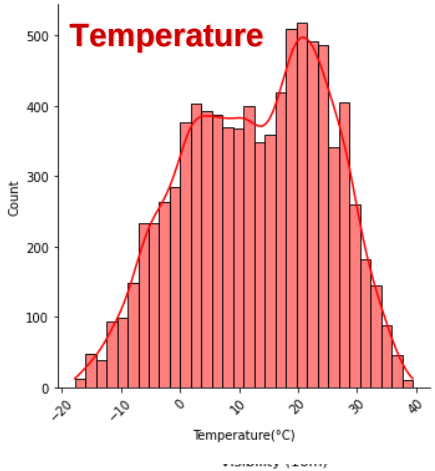
6. Future scope: here we will discuss about future scope on same dataset.

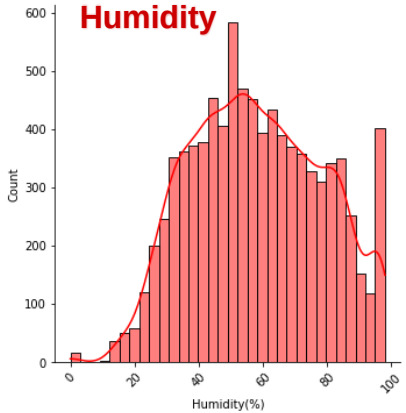
7. Summary: A quick summary of the entire session

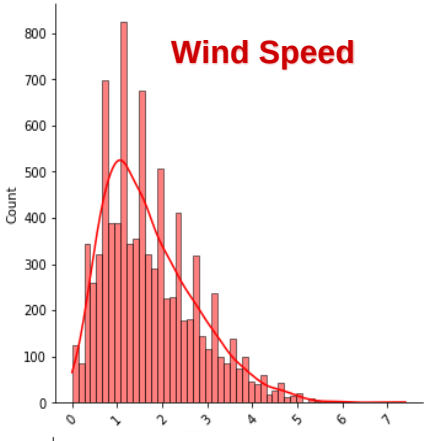
Data Summary

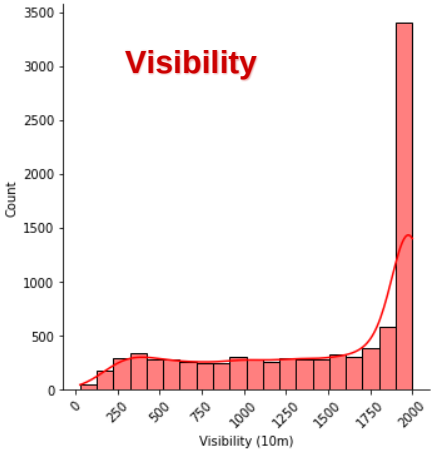


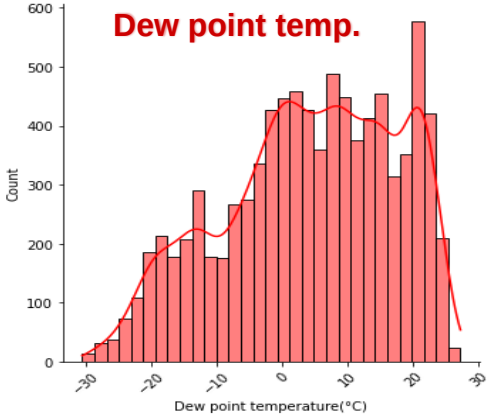
Univariate analysis

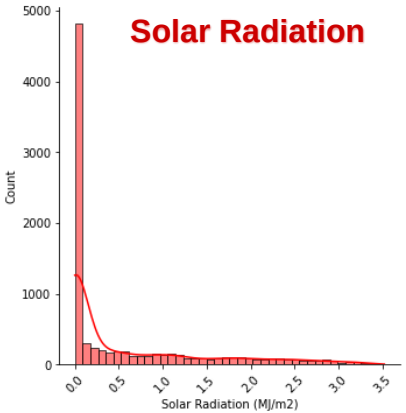


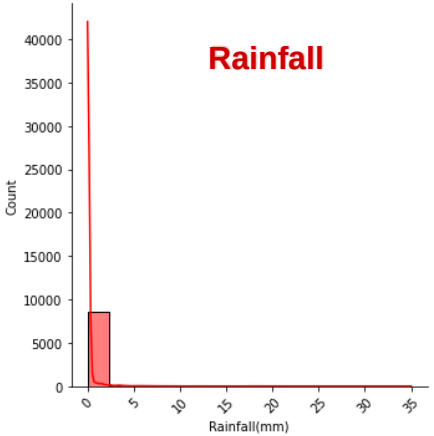


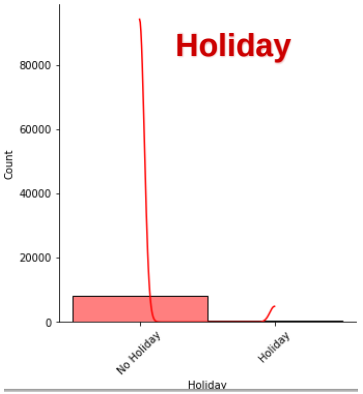






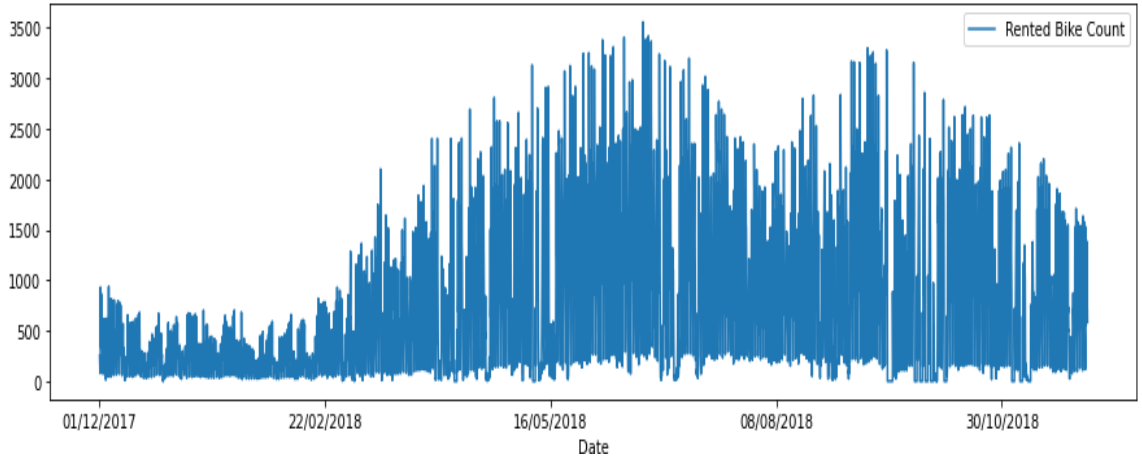


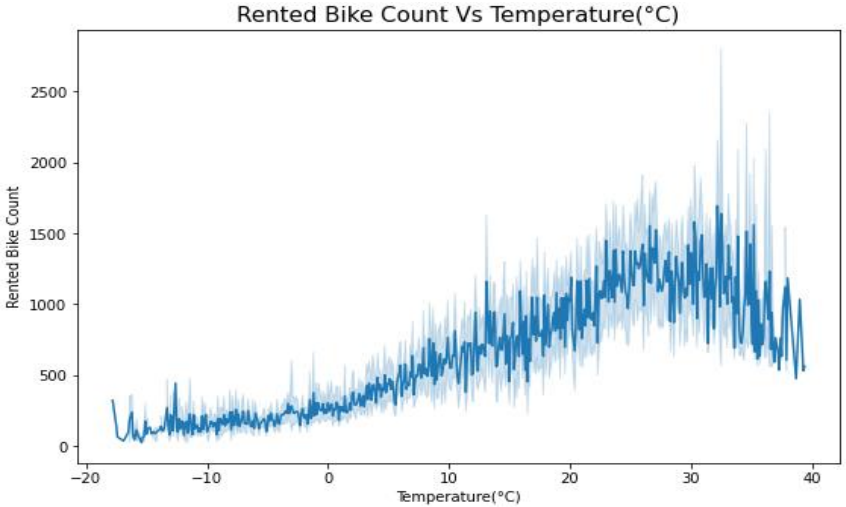




Bi-Variate analysis

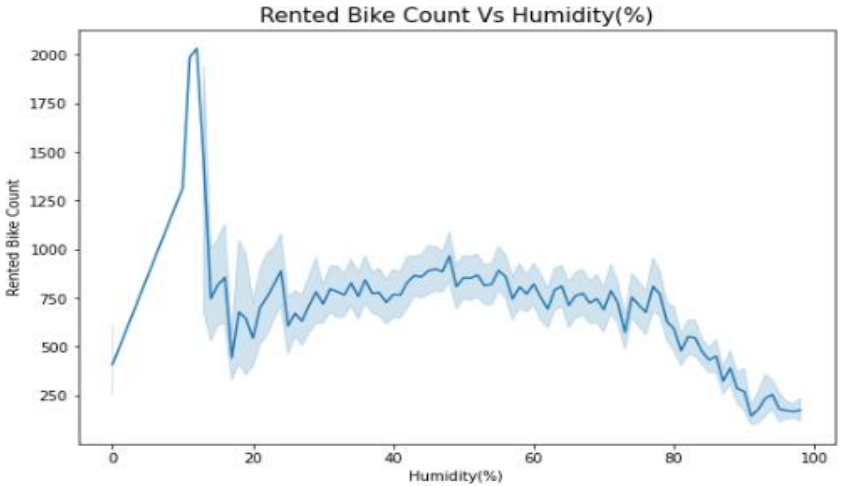
No. of rented bikes is very high in 2018 in comparison of 2017 which slowed down a bit in Dec. One of the reason could be that we are given data which is biased towards 2018. i.e data is majorly of 2018





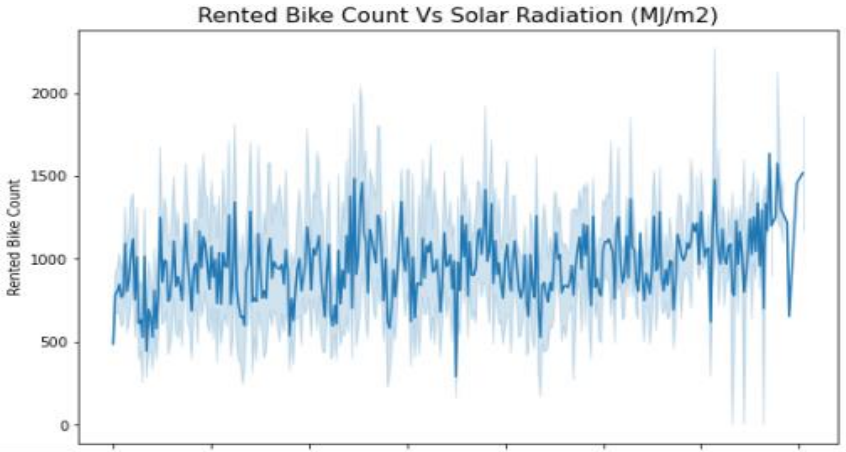
We noticed that as the temperature approaches to 33C no. of rented bike increases and as soon as temperature further increases bike counts get lower down.

Bi-Variate analysis



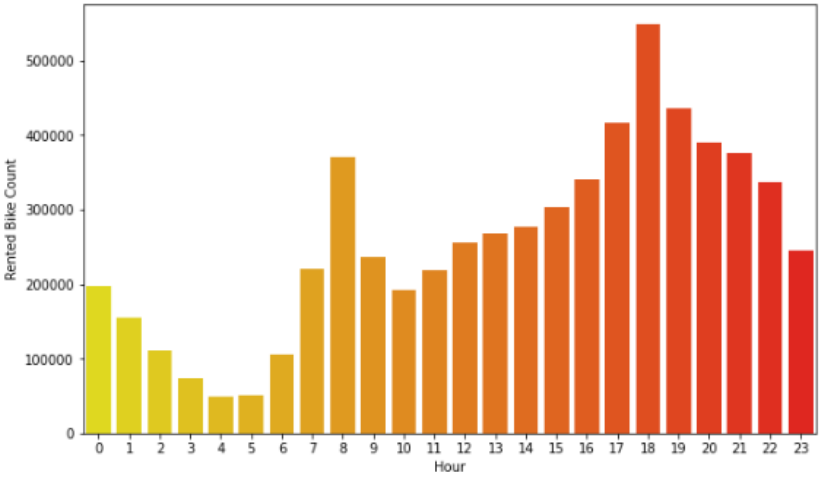
- One more interesting insight we have noticed that there is no impact of solar radiation on bike counts

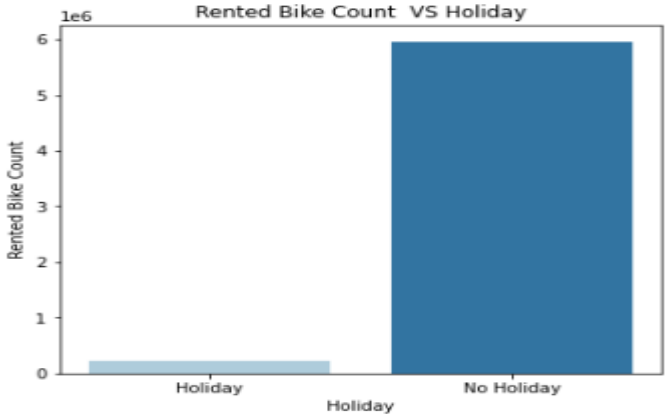
- We also have noticed that bike counts remain almost un-effective when humidity level is between 20 to 60 and when it further increases then we notice a linear decrease in counts.



Bi-Variate analysis

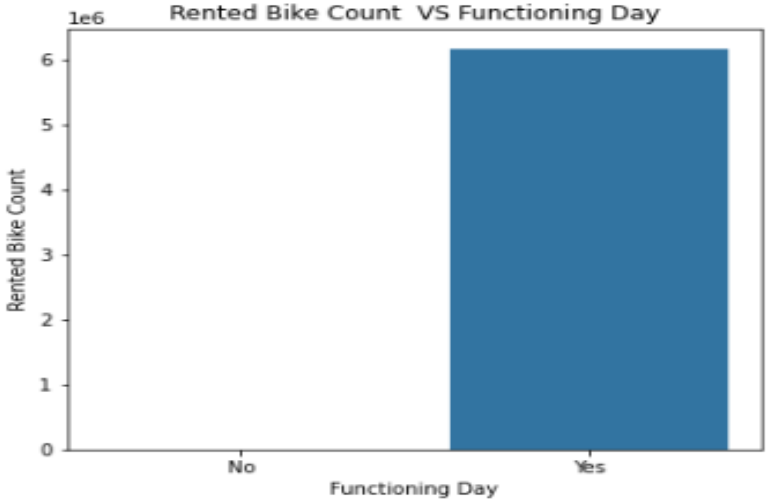
We noticed that bike counts is low in first half of a day and comparatively quite high in second half. Even if we take a look at evening hours 5PM to 10PM then there is a high rush.





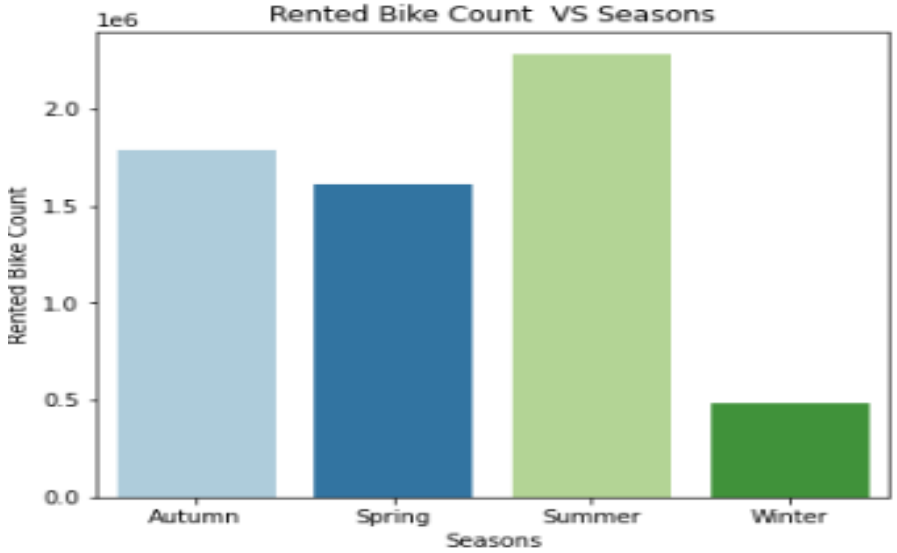
	Holiday	Rented Bike Count
0	Holiday	215895
1	No Holiday	5956419

Bi-Variate analysis



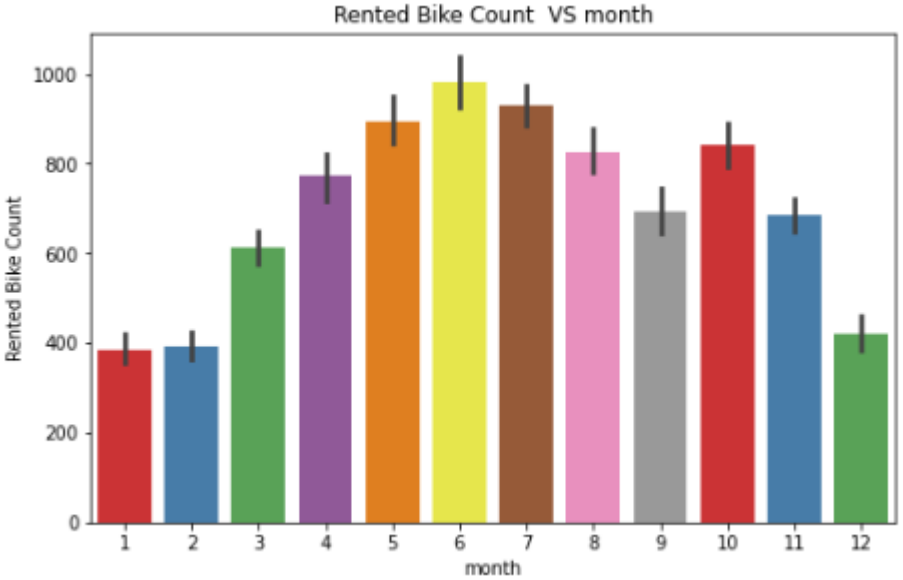
	Functioning Day	Rented Bike Count
0	No	0
1	Yes	6172314

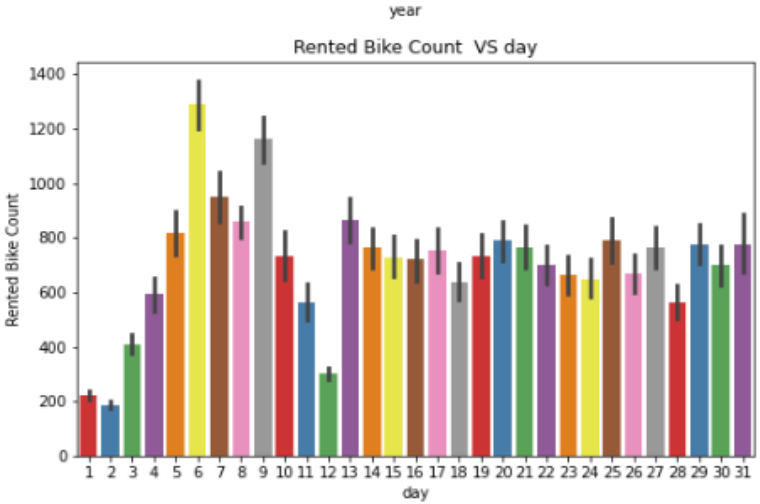
	Seasons	Rented Bike Count
0	Autumn	1790002
1	Spring	1611909
2	Summer	2283234
3	Winter	487169



Bi-Variate analysis

Further we observed that Bike counts is very low in Jan, Feb and Dec months whereas in May, Jun and July it's quite high.



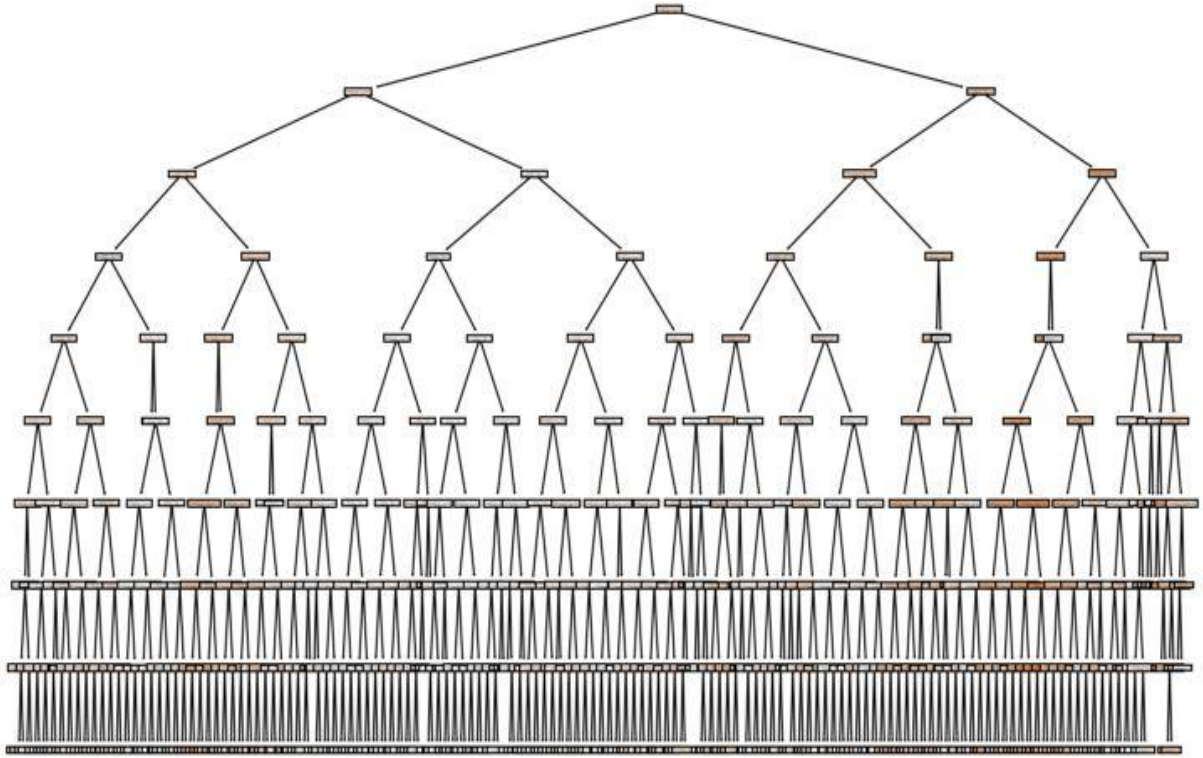


An interesting anomaly was seen in days of the month. In starting 3 days bike count was low and then it increases up to 10th day. Finally remain approx. constant through out the month

4. Predictions and it’s results :

```
R2 : 0.793471882002239
Adjusted R2 :  0.7917595346397517
```

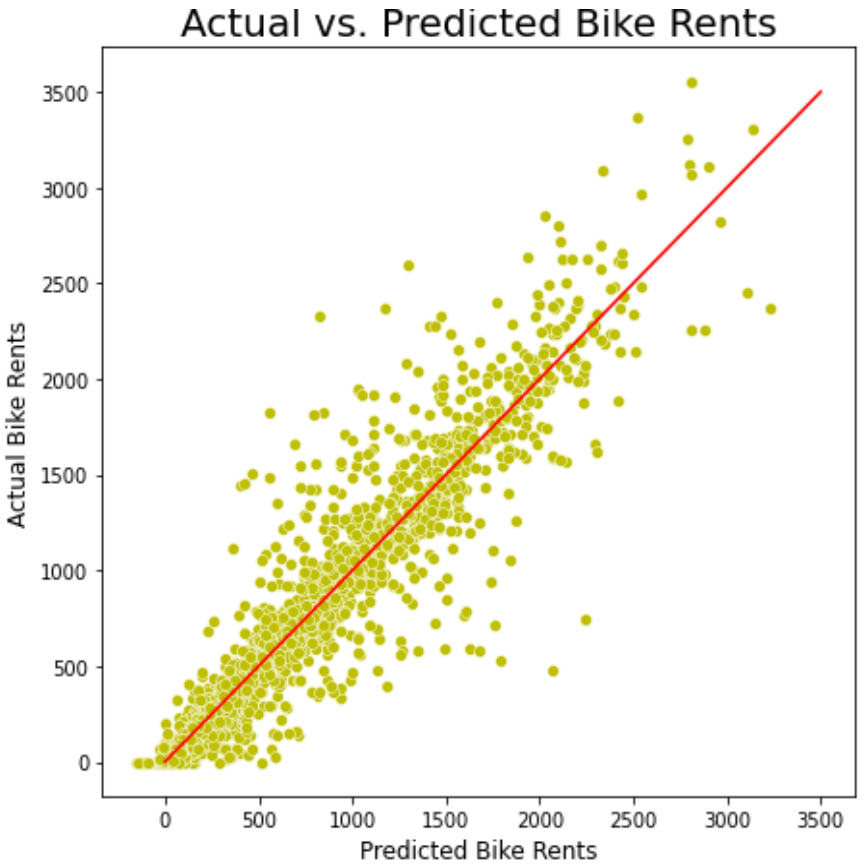
We have tried Decision tree algorithm how ever results were not very good. As we can see that the r^2 value is under 0.80 and the tree that takes place in this algorithm is also given in this slide



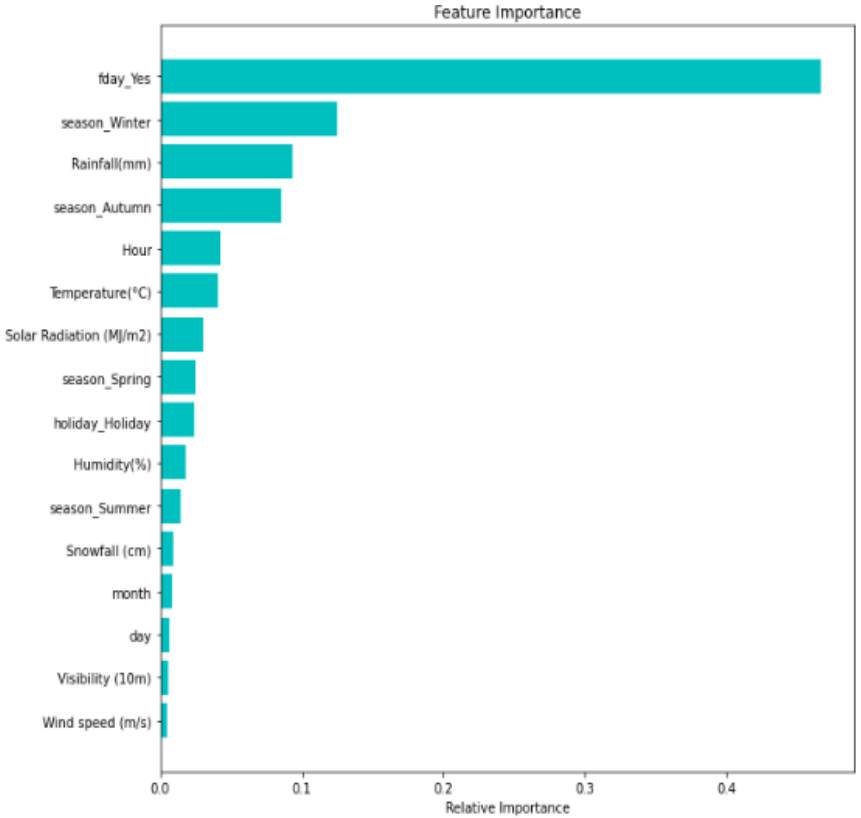
4. Predictions and it’s results :

```
R2 on test data : 0.8876661400961339
Adjusted R2 on test data :  0.8867347676971151
R2 on train data : 0.9994354970815345
Adjusted R2 on train data :  0.9994339460126087
```

In order to get better predictions we had to use a black-box model and hence we used XGBoost regressor and using this we got quite good results which we can see in the adjacent slide



Predictions and it’s results



In this picture we can see what feature contributed the most in our predictions. This can be noticed that non-functional day is the most important column for our predictions followed by winter season and a functional and non-holiday day are least contributed columns.

Challenges

- *In order to use a parametric algorithm and upon analyzing features we found that deciding the transformation was not so easy as we have noticed various distributions.*

- *Deciding a non-parametric algorithm was a challenge as we were not getting very good results using some of the algorithms.*

- *Bi-Variate analysis section was a bit difficult as there was lot of importance and we were supposed get best of it.*

Future Scope

- Since this project was focused on regression hence we have worked majorly in a way so that we can achieve better accuracy hence we didn’t dive into exploratory data analysis if we would have much time to complete this project so we had go for in-depth analysis.

- We could have spend more time on hyper parameter tuning as XGBoost model was taking a lot time in running and finding best hyper parameters.

- Using above two approaches there was a possibility of getting much better results.

Summary

- *First we have introduced the Seoul Bike sharing service*
- *Then we had a look at the data that we were given*
- *Next we have used univariate analysis to understand the feature distributions and found that there a many different types distributions hence we have decided to use non-parametric algorithm*
- *Next we looked at bi-variate analysis where we find many interesting insights*

Summary

- *Next we have seen results of decision tree which were not very effective*
- *Then we have looked at our final results which were obtained using XGBoost regressor*
- *Next we have seen feature importances.*
- *Finally we had talked about project challenges and future scope.*

Thank You