Capstone Project

Seoul Bike Sharing Demand Prediction

Anuj Menaria

Points to Discuss:

1. Agenda
2. Data Summary
3. Data Columns
4. Exploratory Data Analysis
5. Visualizing Distribution
6. Visualizing Outliers
7. Handling Outliers
8. Bivariate Analysis of Linearity in Data
9. Correlation Heatmap
10. Model Building Prerequisites
11. Model Implementation
12. Conclusion

Agenda

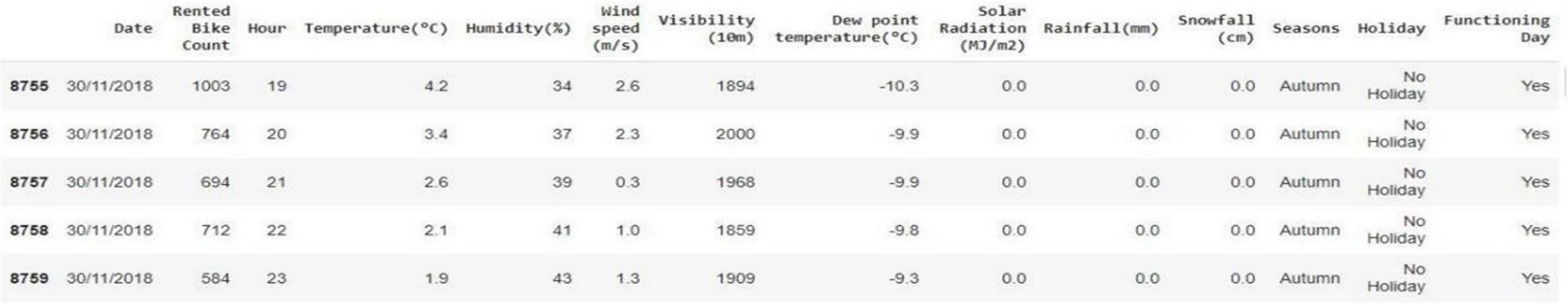
PROBLEM DESCRIPTION:

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 the bike count required at each hour for the stable supply of rental bikes.

Bike sharing programs can become more successful if the limitation can be overcome, such limitations are :

- Stable supply of bikes.
- Finding factors affecting shortage of bikes and time delay in availing bike.
- Maximizing the bike availability & Minimizing the waiting period.

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
8755	30/11/2018	1003	19	4.2	34	2.6	1894	-10.3	0.0	0.0	0.0	Autumn	No Holiday	Yes
8756	30/11/2018	764	20	3.4	37	2.3	2000	-9.9	0.0	0.0	0.0	Autumn	No Holiday	Yes
8757	30/11/2018	694	21	2.6	39	0.3	1968	-9.9	0.0	0.0	0.0	Autumn	No Holiday	Yes
8758	30/11/2018	712	22	2.1	41	1.0	1859	-9.8	0.0	0.0	0.0	Autumn	No Holiday	Yes
8759	30/11/2018	584	23	1.9	43	1.3	1909	-9.3	0.0	0.0	0.0	Autumn	No Holiday	Yes

- This Dataset contains 8760 rows and 14 columns.
- Three categorical features ‘Seasons’, ‘Holiday’, & ‘Functioning Day’.
- One Datetime column ‘Date’.
- We have some numerical type variables such as temperature, humidity, wind, visibility, dew point temp, solar radiation, rainfall, and snowfall which show the environmental conditions for that particular hour of the day.

Data Summary(contd..)

- There are No Missing Values present.
- There are No Duplicate values present.
- There are No null values.
- The dependent variable is 'rented bike count' which we need to make predictions on.
- The dataset shows hourly rental data for one year (1 December 2017 to 31 November 2018) (365 days).

Data Summary(contd..)

Column Features

Numeric

- Hour
- Temperature
- Humidity
- Wind
- Dew point temperature
- Sunlight
- Rain
- Snow

Categorical

- Season
- Holiday
- Functioning Day
- Time shift

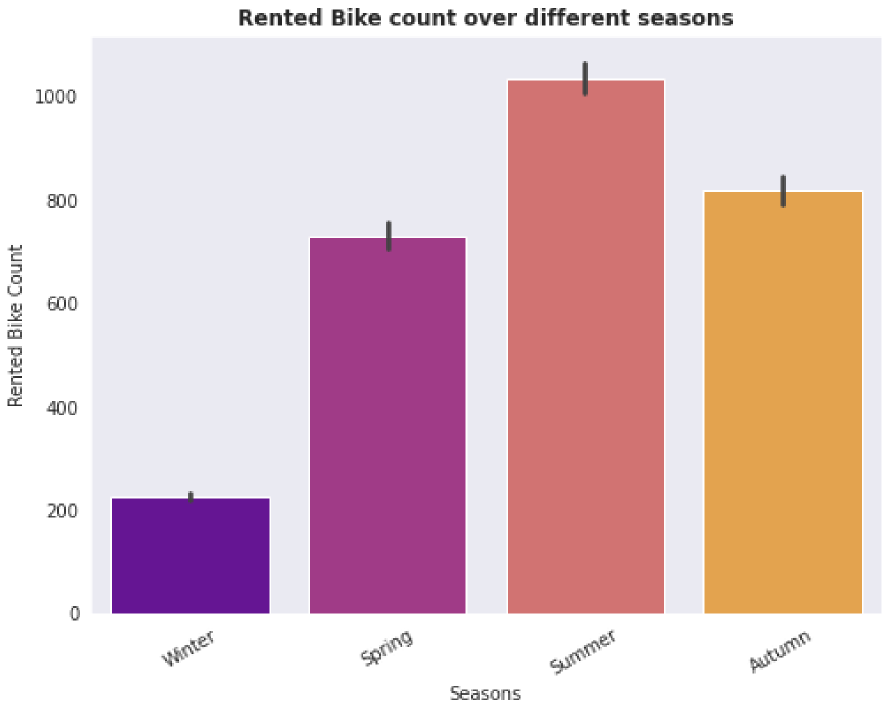
Target Column

Rented Bike Count

Data Columns

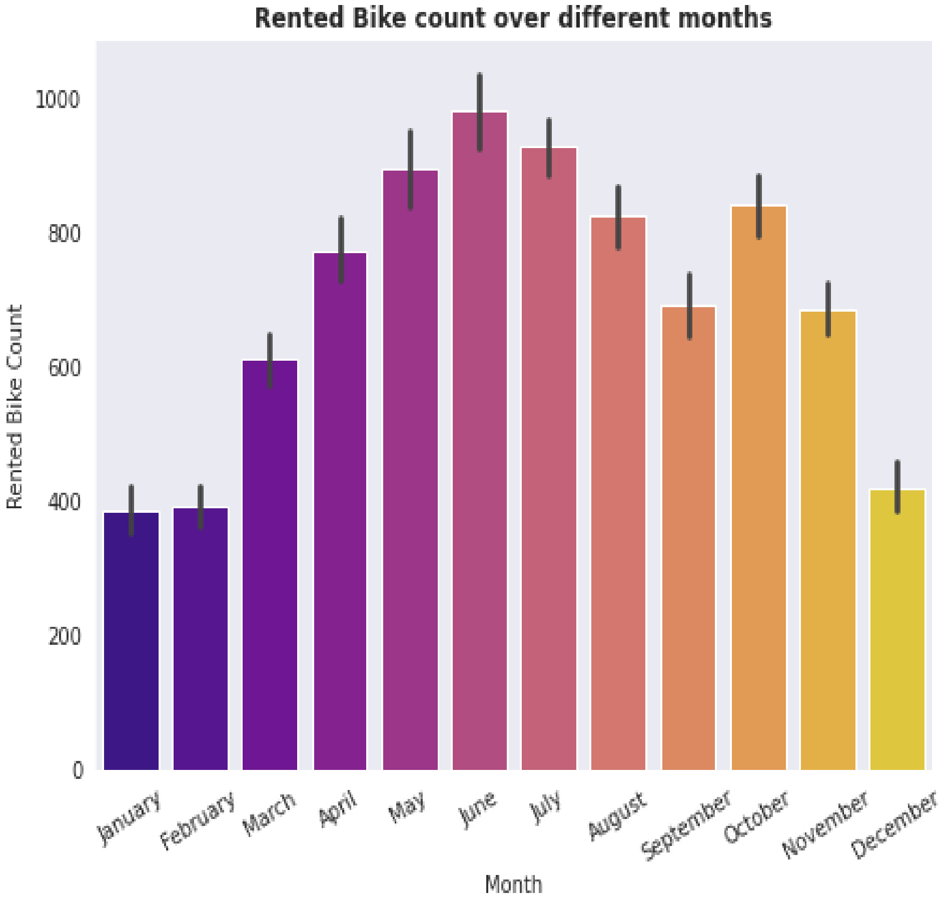
- **Date**: The date of the day, during 365 days from 01/12/2017 to 30/11/2018, formatting in DD/MM/YYYY, type: str.
- **Rented Bike Count**: Number of rented bikes per hour which is the dependent variable and we need to predict that, type: int
- **Hour**: The hrs. of the day, starting from 0-23 it's in digital time format, type: int
- **Temperature(°C)**: Temperature in Celsius, type: Float
- **Humidity(%)**: Humidity in the air in %, type: int
- **Wind speed (m/s)**: Speed of the wind in m/s, type: Float
- **Visibility (10m)**: Visibility in m, type: int
- **Dew point temperature(°C)**: Temp. at the beginning of the day, type: Float
- **Solar Radiation (MJ/m2)**: Sun contribution, type: Float
- **Rainfall(mm)**: Amount of rain in mm, type: Float
- **Snowfall (cm)**: Amount of snowing in cm, type: Float
- **Seasons**: Season of the year, type: str, there are only 4 season's in the data
- **Holiday**: If the day is a holiday period or not, type: str
- **Functioning Day**: If the day is a Functioning Day or not, type : str

Exploratory Data Analysis(EDA)



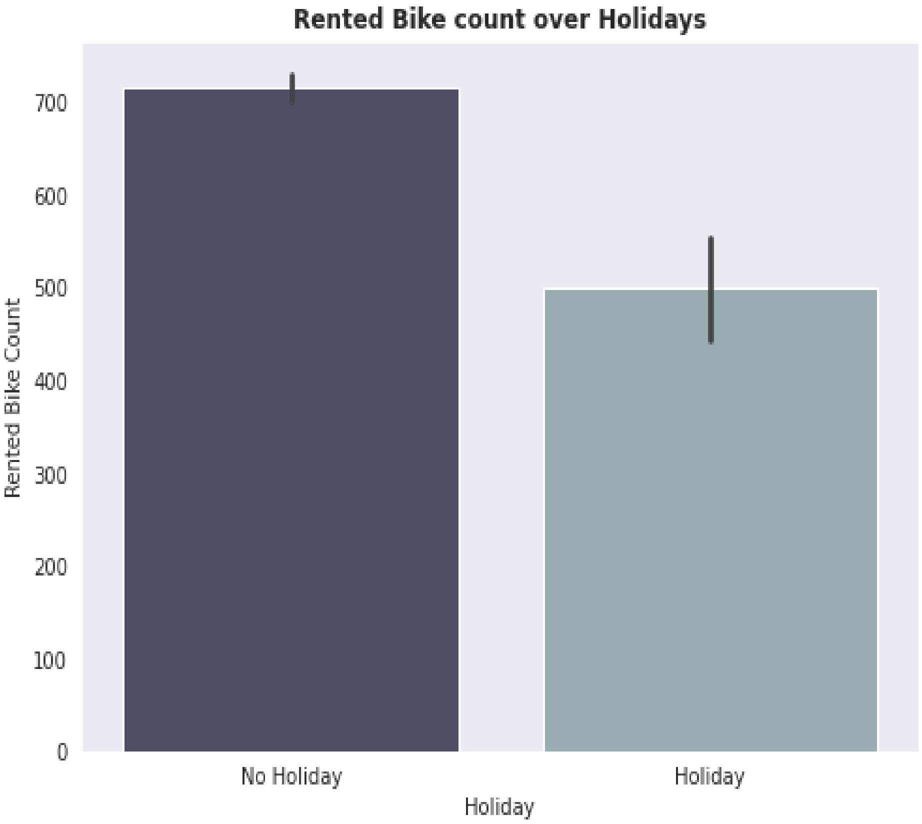
- The study of the season’s column determines whether seasons have greater and lower rental bike counts.
- Rented Bike Count is lowest in the winter season.
- Rented Bike Count is highest in the Summer season.

Exploratory Data Analysis(EDA) Contd…



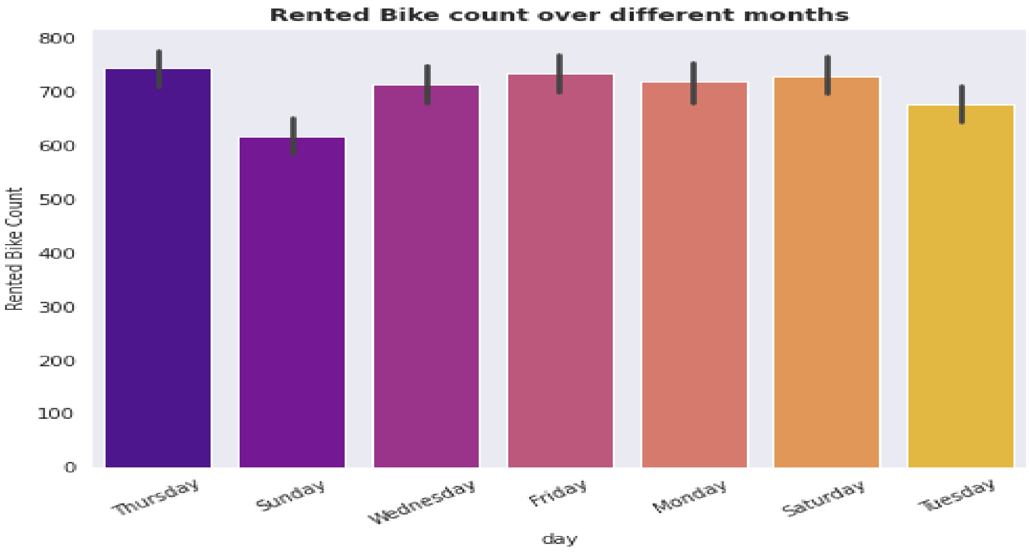
- Demand for bikes is at its peak in June.
- Least demand can be observed in January, February, and December which are also the month of the winter season.

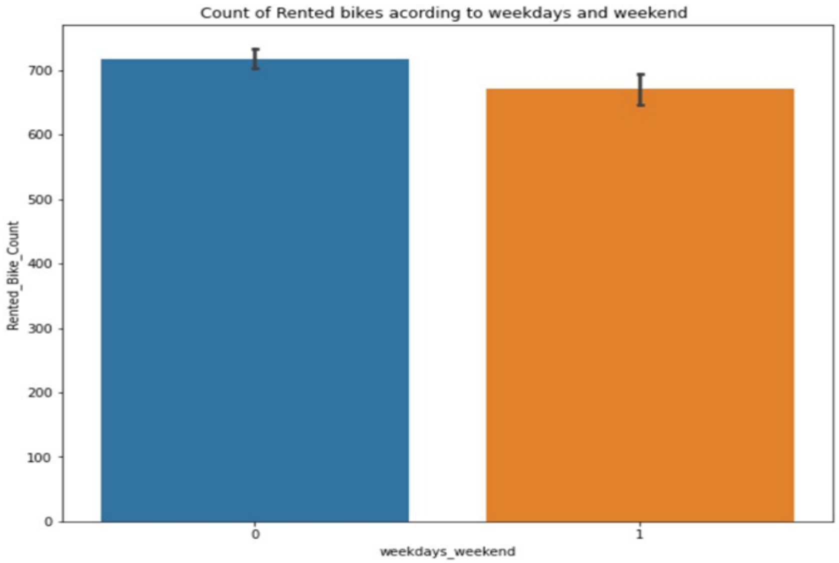
Exploratory Data Analysis(EDA) Contd…



- As when there is a holiday the demand for rented bikes reduces.
- When it is a functioning day or no holiday the demand raises for the rented bike.

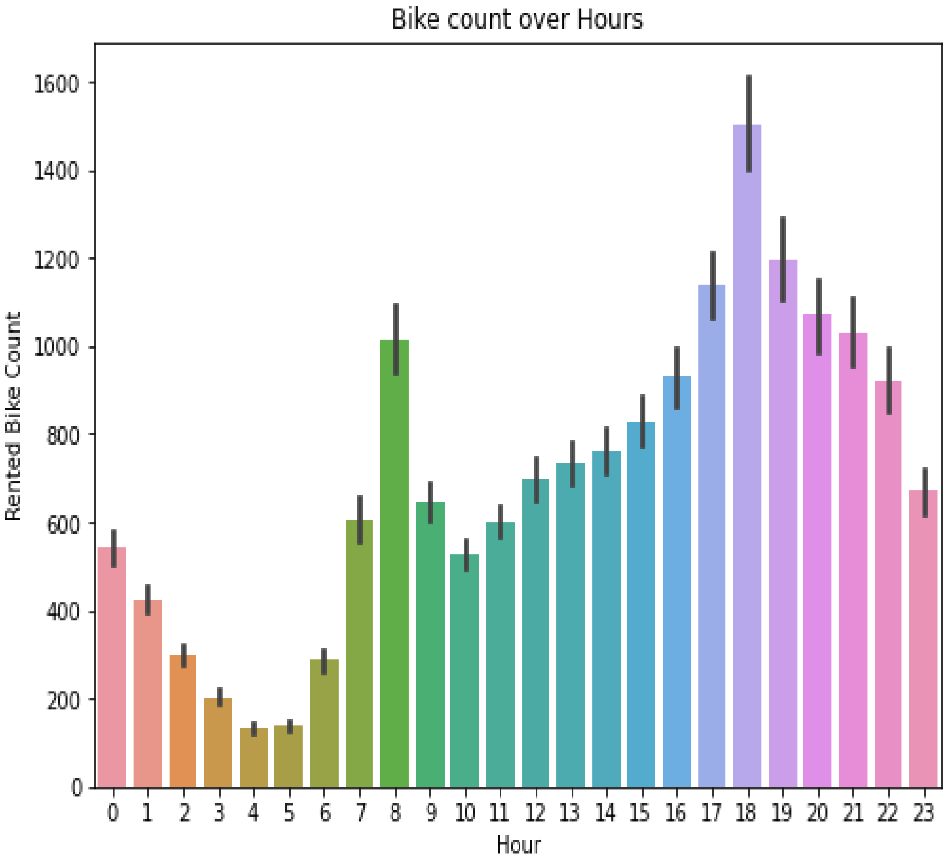
Exploratory Data Analysis(EDA) Contd…





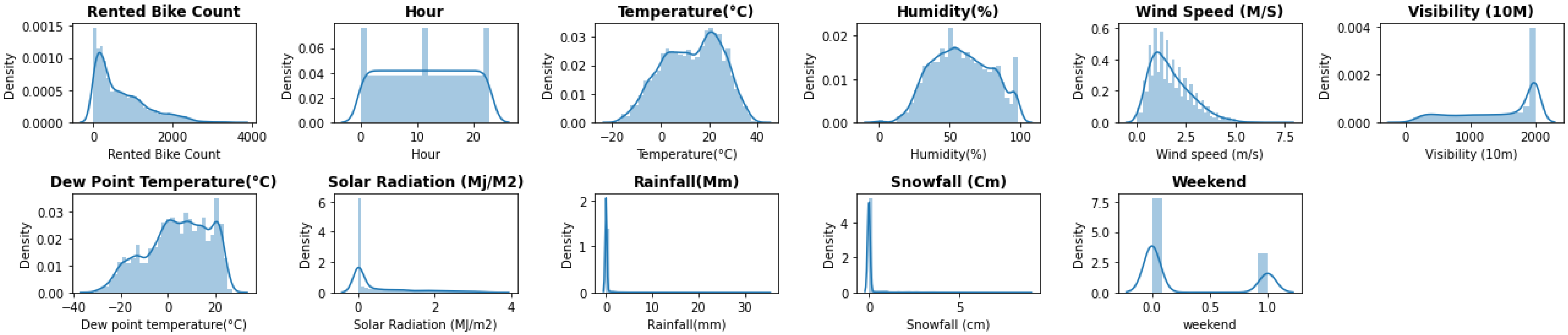
- We can see in the weekly graph that the demand for the rented bike is least on Sunday as it is a Holiday.
- In the second graph we can see the comparison between the weekdays and the weekend and the demand is higher on the weekdays as the commute is much more active on weekdays.

Exploratory Data Analysis(EDA) Contd…



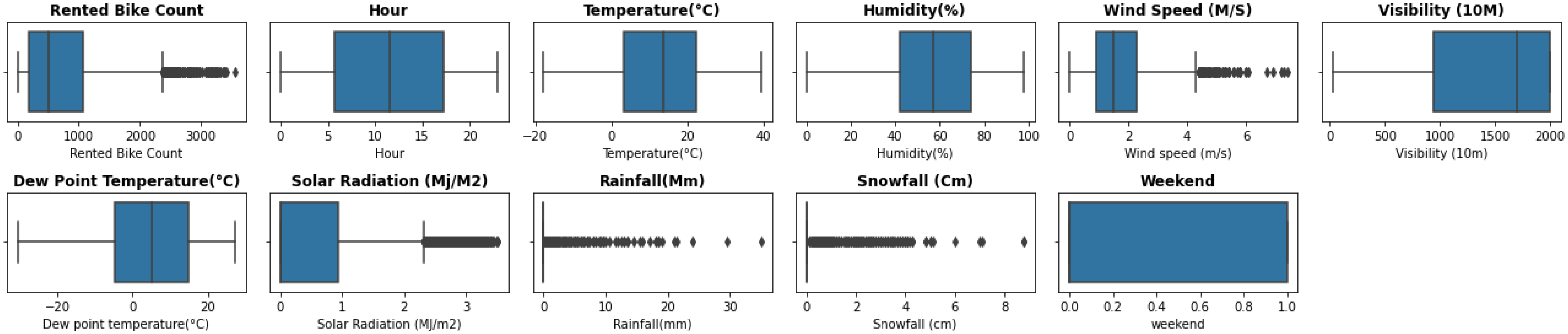
- The demand for the bikes rises the most in the evening around 5 -7 pm and the demand is highest at 6 pm.
- The demand for the bike in the 24 hours is least in the morning around 4 – 5 am.
- We can clearly see the pattern of the bike demand which is on the timings of the job commute which is around 8 am in the morning and 6 pm in the evening.

Visualizing Distribution



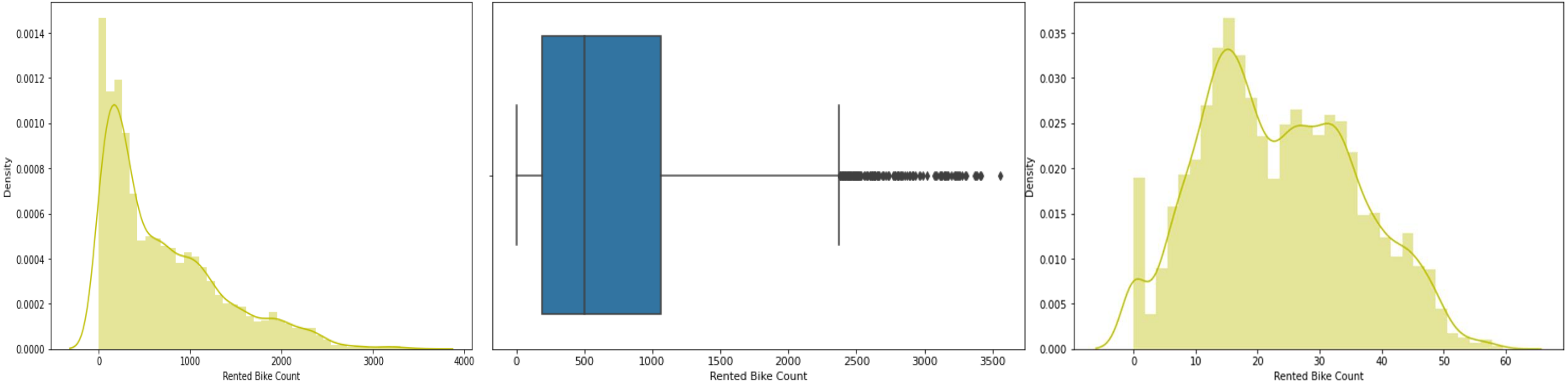
- Data distribution is checked on each column to see the skewness of the data and how much the data is normally distributed.
- It has been observed that the Hour, Temperature, Humidity, and Dew Point Temperature are quite normally distributed than the other columns.
- We can see that the Rented Bike Count, Solar Radiation, Rainfall, and Snowfall are right skewed data columns and the Validity Column is left Skewed.

Visualizing Outliers



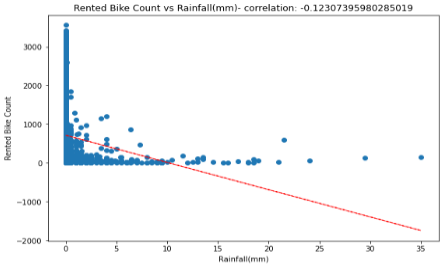
- We see outliers in some columns like Sunlight, Wind, Rainfall, and Snowfall but let’s not treat them because they may not be outliers as snowfall, rainfall, etc. themselves are rare events in some countries.
- We treated the outliers in the target variable by capping with IQR limits.

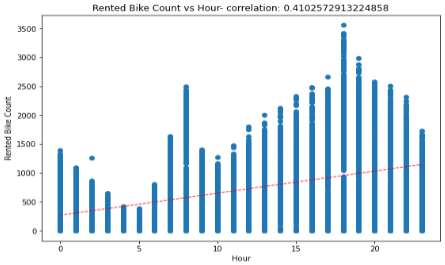
Handling Outliers

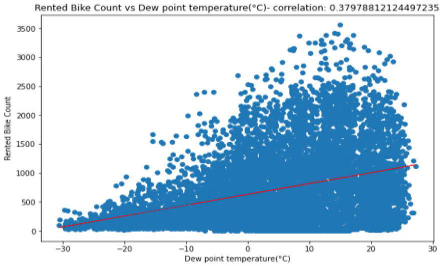


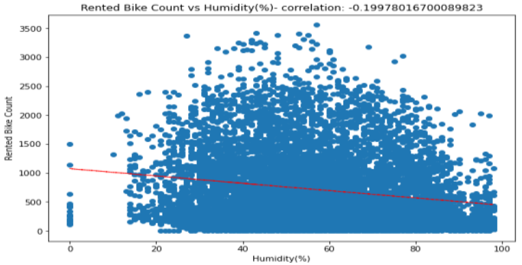
- Earlier the distribution of the target variable was positively skewed with a skewness value of 0.983. We tried to make this distribution somewhat close to normal distribution.
- First we applied log transformation, but it did not give the desired results, we finally applied square root transformation. We got favorable results, the skewness value was dropped to 0.2373, which is comparatively closer to the normal distribution.

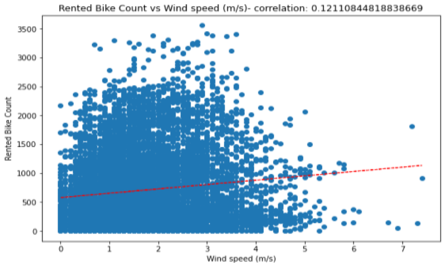
Bivariate Analysis of Linearity in Data

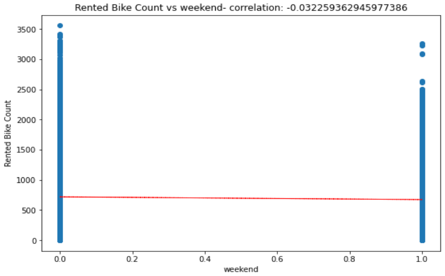

Rented Bike Count vs Rainfall(mm)- correlation: -0.12307395980285019

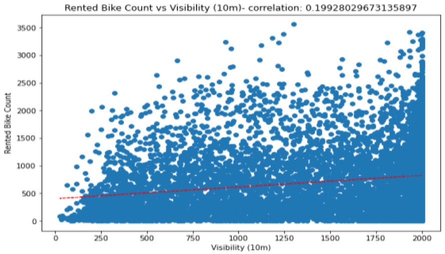

Rented Bike Count vs Hour- correlation: 0.4102572913224858

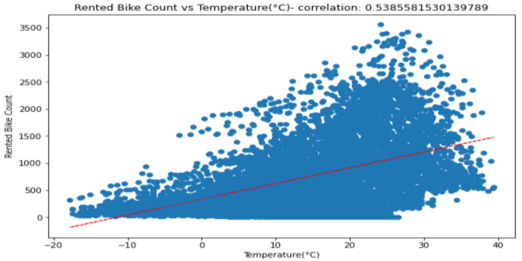

Rented Bike Count vs Dew point temperature(°C)- correlation: 0.37978812124497235

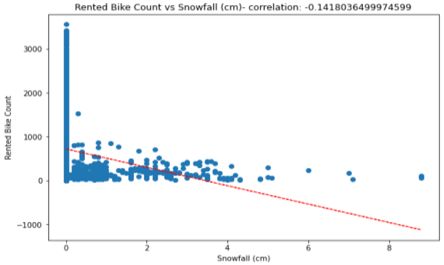

Rented Bike Count vs Humidity(%)- correlation: -0.19978016700089823

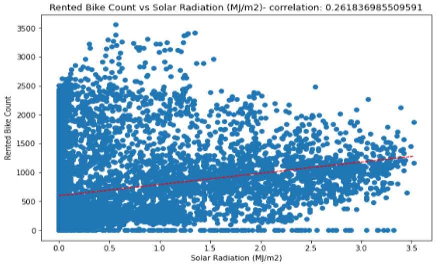

Rented Bike Count vs Wind speed (m/s)- correlation: 0.12110844818838669


Rented Bike Count vs weekend- correlation: -0.032259362945977386


Rented Bike Count vs Visibility (10m)- correlation: 0.19928029673135897

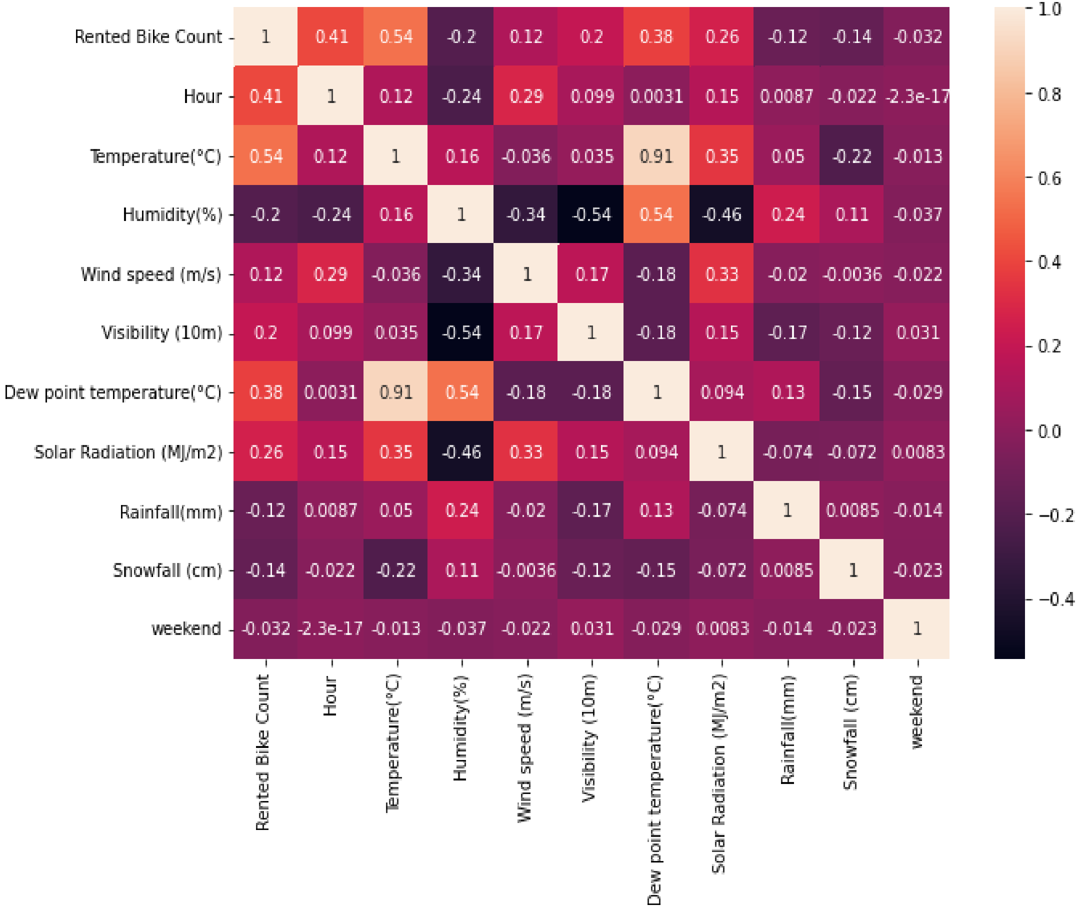

Rented Bike Count vs Temperature(°C)- correlation: 0.5385581530139789


Rented Bike Count vs Snowfall (cm)- correlation: -0.1418036499974599


Rented Bike Count vs Solar Radiation (MJ/m2)- correlation: 0.261836985509591

- From the visualizations we observed that hour, temp, sunlight, and dew temp are positively correlated with the bike count.
- Humidity, rain, snow, and winter features are having a negative correlation with the bike count.
- Some features are also showing close to zero correlation with the target variable as the regression line is not inclined.

Correlation Heatmap



	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)	weekend
Rented Bike Count	1	0.41	0.54	-0.2	0.12	0.2	0.38	0.26	-0.12	-0.14	-0.032
Hour	0.41	1	0.12	-0.24	0.29	0.099	0.0031	0.15	0.0087	-0.022	-2.3e-17
Temperature(°C)	0.54	0.12	1	0.16	-0.036	0.035	0.91	0.35	0.05	-0.22	-0.013
Humidity(%)	-0.2	-0.24	0.16	1	-0.34	-0.54	0.54	-0.46	0.24	0.11	-0.037
Wind speed (m/s)	0.12	0.29	-0.036	-0.34	1	0.17	-0.18	0.33	-0.02	-0.0036	-0.022
Visibility (10m)	0.2	0.099	0.035	-0.54	0.17	1	-0.18	0.15	-0.17	-0.12	0.031
Dew point temperature(°C)	0.38	0.0031	0.91	0.54	-0.18	-0.18	1	0.094	0.13	-0.15	-0.029
Solar Radiation (MJ/m2)	0.26	0.15	0.35	-0.46	0.33	0.15	0.094	1	-0.074	-0.072	0.0083
Rainfall(mm)	-0.12	0.0087	0.05	0.24	-0.02	-0.17	0.13	-0.074	1	0.0085	-0.014
Snowfall (cm)	-0.14	-0.022	-0.22	0.11	-0.0036	-0.12	-0.15	-0.072	0.0085	1	-0.023
weekend	-0.032	-2.3e-17	-0.013	-0.037	-0.022	0.031	-0.029	0.0083	-0.014	-0.023	1

A high correlation between the following variables-

- Dew point temperature with Temperature, 0.91
- Dew point temperature with Humidity,0.54
- Rented Bike count with Temperature, 0.54 & Dew point temperature 0.38
- Temperature with solar radiation 0.35 & snowfall - 0.22

Model Building Prerequisites

- Feature Scaling or Standardization: It is a step of Data Pre Processing that is applied to independent variables or features of data. It basically **helps to normalize the data within a particular range**. Sometimes, it also helps in speeding up the calculations in an algorithm.
- Here we used **OneHotEncoding** on certain Categorical columns to get the standardization in the data frame so the model performance can be increased and the time consumed by the model can be reduced.

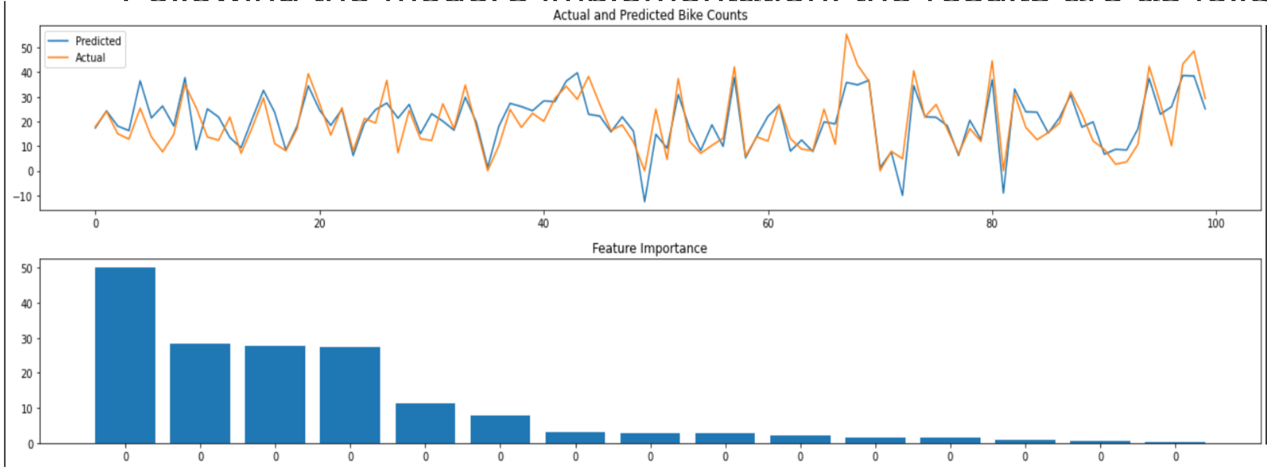
Models Used

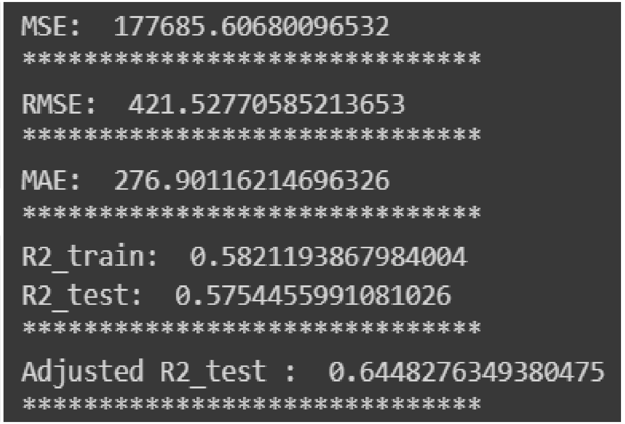
- Linear Regression
- Decision Tree Regressor
- Random Forest
- XGB Regressor

Model Implementation

Linear Regression:

- We presented the beta coefficients' absolute values, which can be observed parallel to how important a feature is for tree-based methods.
- Since the performance of the simple linear model is not so good. We experimented with some complex models.
- Following the model's implementation, the results are as follows:



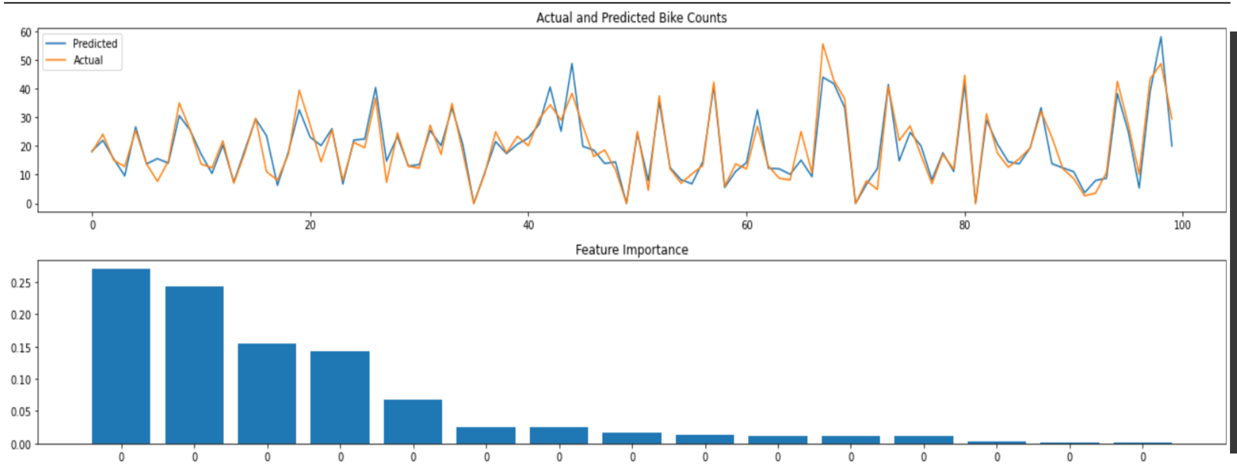


```
MSE:  177685.60680096532
******************************

RMSE:  421.52770585213653
******************************

MAE:  276.90116214696326
******************************

R2_train:  0.5821193867984004
R2_test:  0.5754455991081026
******************************

Adjusted R2_test :  0.6448276349380475
******************************
```

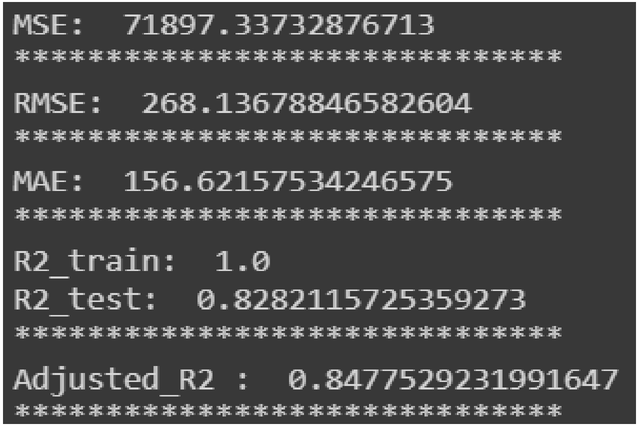
Model Implementation

Decision Tree Regressor:

- With an accuracy of more than 80%, decision trees perform better than linear regression.
- Following the model's implementation, the results are as follows:


Actual and Predicted Bike Counts

Feature Importance

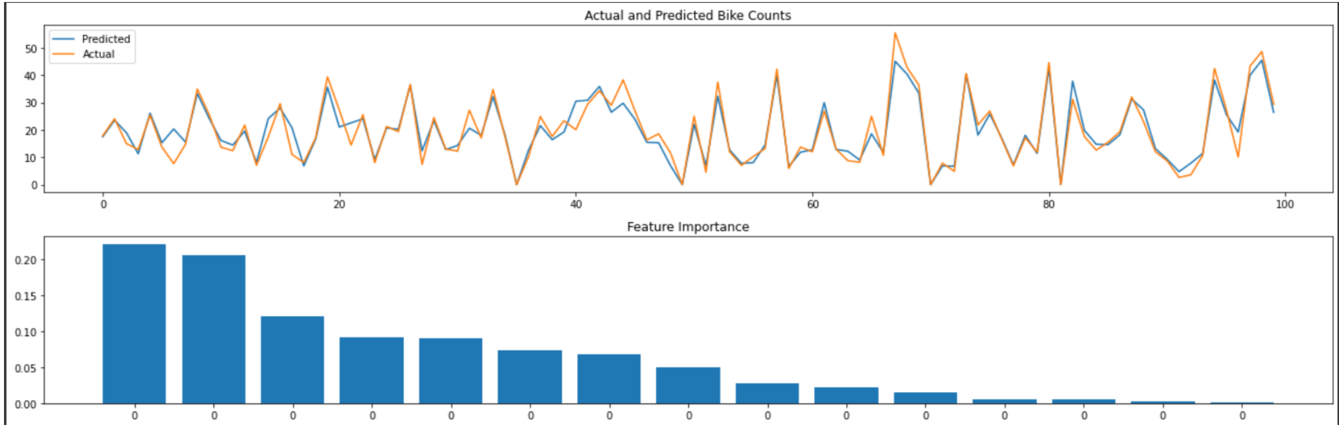


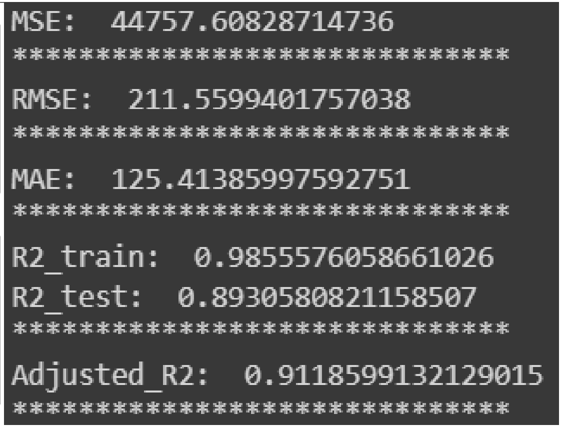
```
MSE:  71897.33732876713
******************************
RMSE:  268.13678846582604
******************************
MAE:  156.62157534246575
******************************
R2_train:  1.0
R2_test:  0.8282115725359273
******************************
Adjusted_R2 :  0.8477529231991647
******************************
```

Model Implementation

Random Forest Regressor:

- The system builds each tree randomly to encourage uncorrelated forests, which then employs the forecasting abilities of the forest to make informed decisions.
- With an accuracy of 90%, we can see that the random forest regressor is doing pretty well for the given situation.
- Following the model's implementation, the results are as follows:



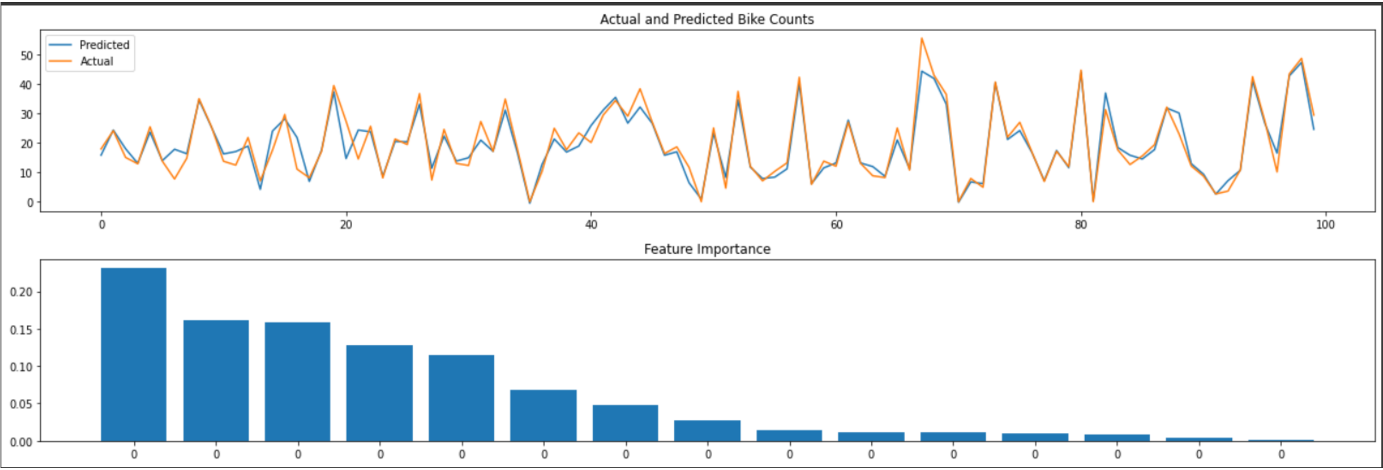


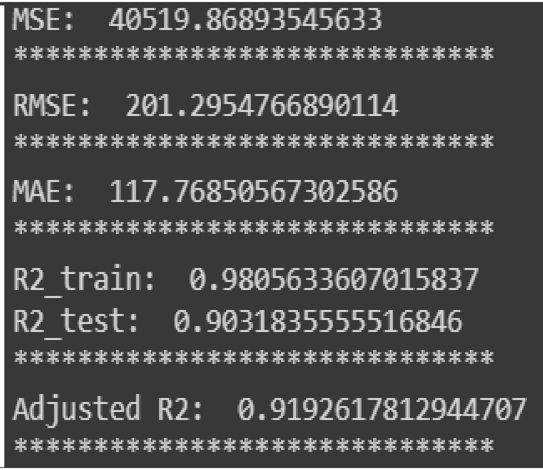
```
MSE:  44757.60828714736
******************************

RMSE:  211.5599401757038
******************************

MAE:  125.41385997592751
******************************

R2_train:  0.9855576058661026
R2_test:  0.8930580821158507
******************************

Adjusted_R2:  0.9118599132129015
******************************
```

Model Implementation

XG Boost Regressor:

- XG Boost Regressor emerges as the best model according to the evaluation matrix score in the train and test.
- With the help of hyperparameter tuning(eta = 0.05, max_depth = 8, n_estimators = 150) the data performed quite well and given the best result among all the algorithms.





```
MSE:  40519.86893545633
******************************

RMSE:  201.2954766890114
******************************

MAE:  117.76850567302586
******************************

R2_train:  0.9805633607015837
R2_test:  0.9031835555516846
******************************

Adjusted R2:  0.9192617812944707
******************************
```

Conclusion

- As it was stated in the problem statement, the business just started out in 2017. So the number of bikes rented in 2017 was too small.
- We can see in the year 2018 the rented bike count was 5986984 which is greater than in 2017.
- We can say on no holiday the rented bike count is much higher than on holiday.
- An ironic insight, all the holidays fall on functioning Days.
- We can say on no holiday the rented bike count is much higher than on holiday.
- The number of business hours of the day and the demand for rented bikes were most correlated. It's common sense too.
- Highest number of bikes rented at the 18th hour of the day.
- After trying combinations of features with linear regression the model under fitted. It seemed obvious because data is spread too much. It didn't seem practical to fit a line.
- Hour, temperature, and solar radiation were the most important features for predicting the count of bikes required.
- Rainfall and snowfall impact the number of bikes rented tremendously with a very high downfall.
- Random forest regressor performs really well when compared to linear regression with high model performance and low rmse.

- The quantity of bicycle rentals has significantly increased in 2018. Demand decreased in the most recent month in 2018, while it was seen to be increasing at the end of 2017. This is due to the fact that the demand began to increase significantly in 2017 and has continued to do so in the early months of 2018. At the conclusion of the year, there is a decline. This can also be a consequence of the cold months.

- The increase in demand began at the end of 2017, during the winter season. Because demand fell at the end of 2018, the observer would think it odd. In fact, it can be claimed that this situation's company growth from April 2017 to April 2018 surged dramatically. Therefore, we may conclude that while demand rose during the winter of 2017, it still fell short of its full potential. Using straightforward heuristics, we can predict that the demand will decline in December but proportionate to the demand for the entire year if all other independent factors remain constant.

Thank You