Seoul Bike Sharing Demand Prediction

Dhanush Kumar

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. Bi-variate Analysis of Linearity in Data
9. Correlation Heat map
10. Model Building Prerequisites
11. Model Implementation
12. Model Summary
13. 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.

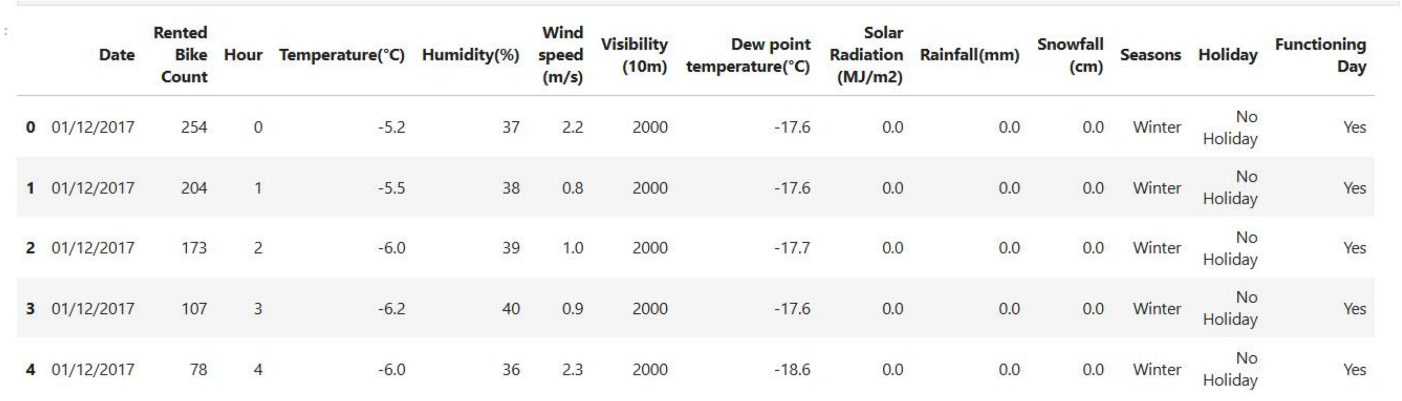
Bike sharing programs can become more successful if the limitation can be overcome, such limitations are :

- 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

This Data set contains 8760 rows and 14 columns.

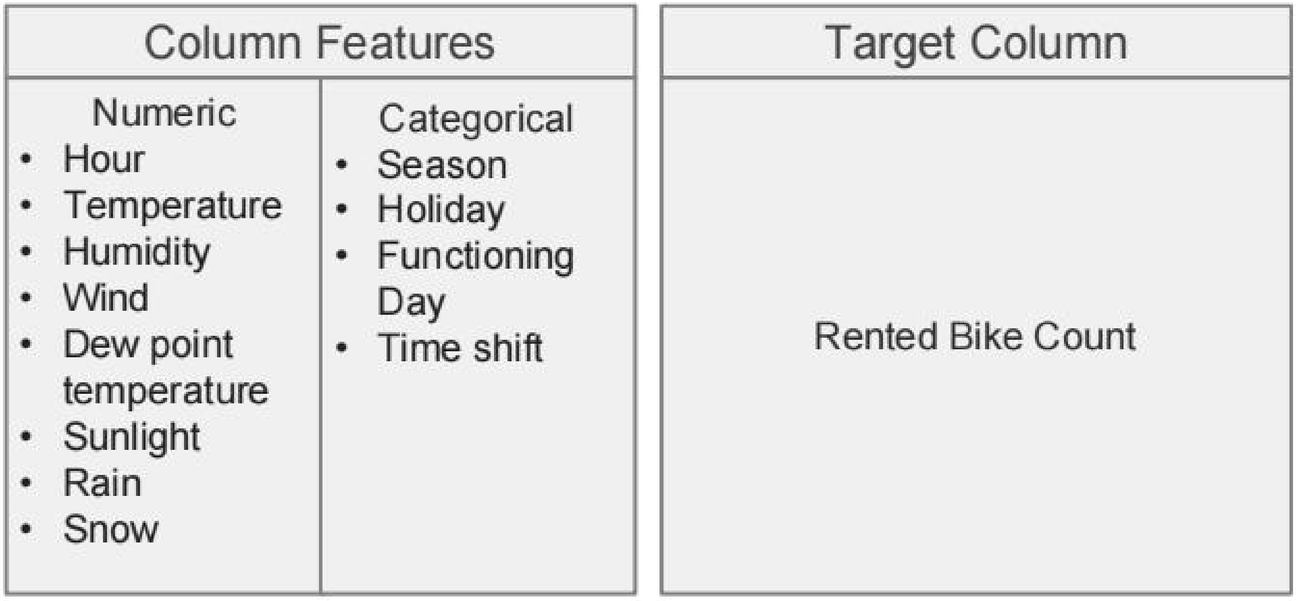
- Three categorical features ‘Seasons’, ‘Holiday’, & ‘Functioning Day’.
- One Datetime column ‘Date’.
- We have some numerical type variables such as temperature, humidity, wind, visibility, dew point temp, solar radiation, rainfall, and snowfall which show the environmental conditions for that particular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



	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

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 data set shows hourly rental data for one year (1 December 2017 to 31 November 2018) (365 days).

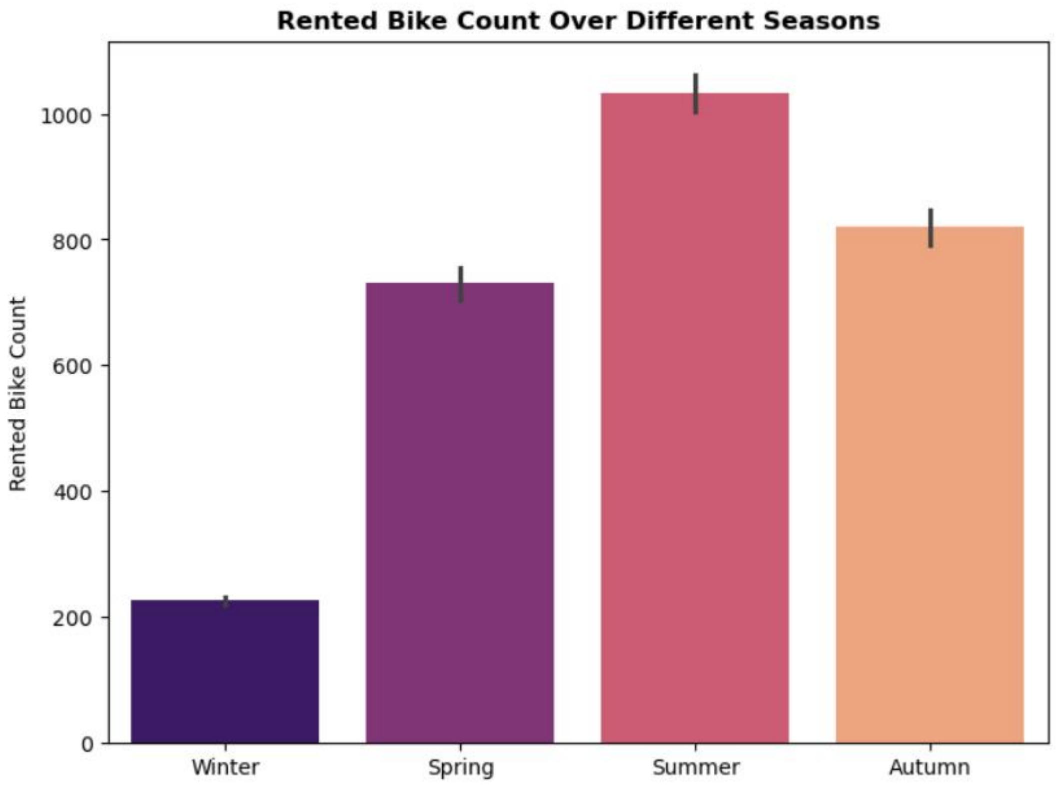


Column Features		Target Column
Numeric • Hour • Temperature • Humidity • Wind • Dew point temperature • Sunlight • Rain • Snow	Categorical • Season • Holiday • Functioning Day • Time shift	Rented Bike Count

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)

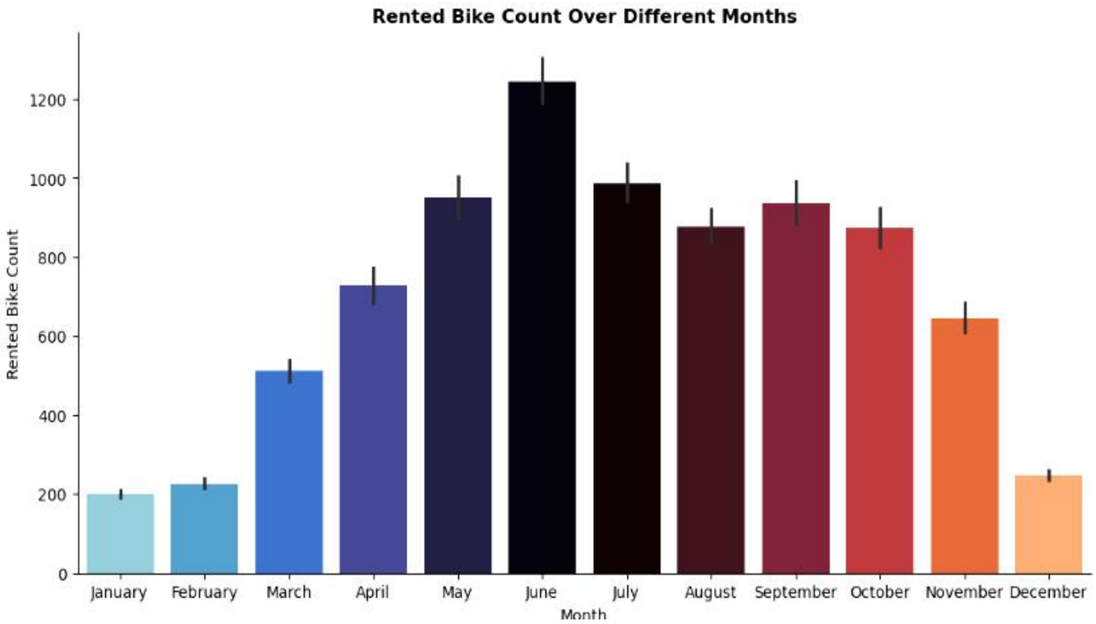


- The study of the season’s column determines whether seasons have greater and lower rental bike counts.

- Rented Bike Count is lowest in the winter season.

- Rented Bike Count is highest in the Summer season.

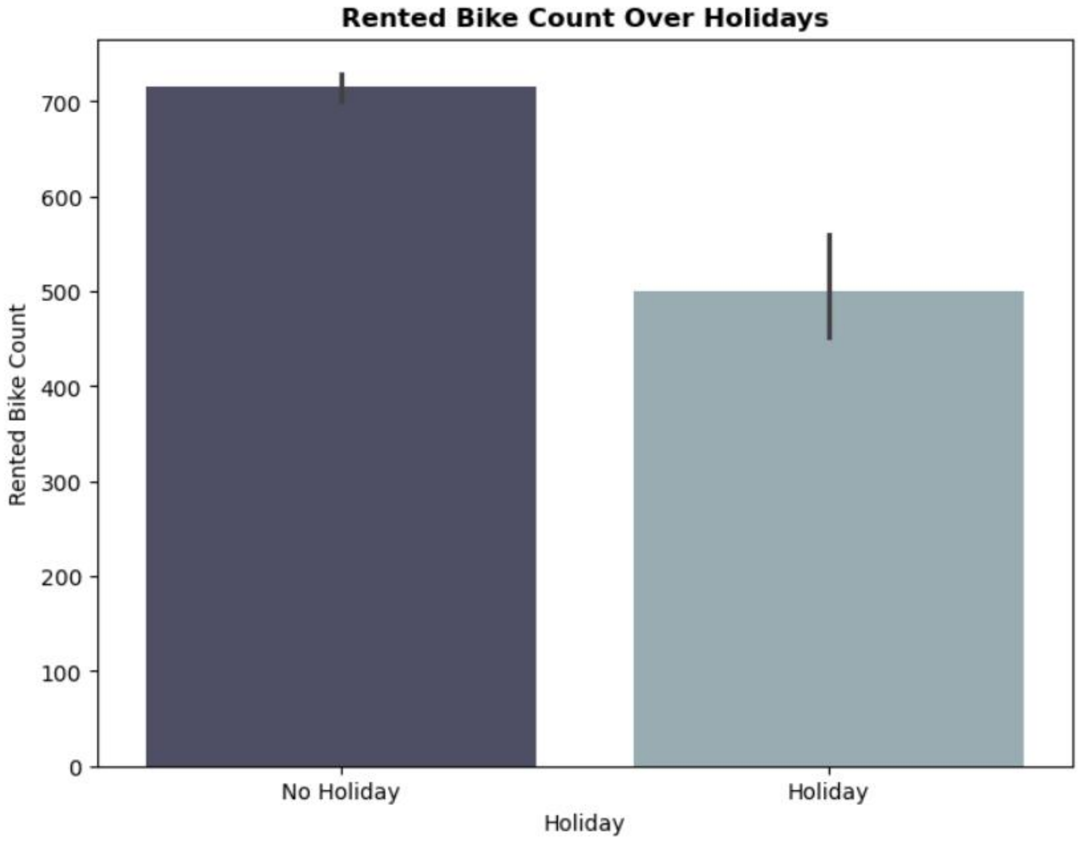
Exploratory Data Analysis(EDA) Contd...



- 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

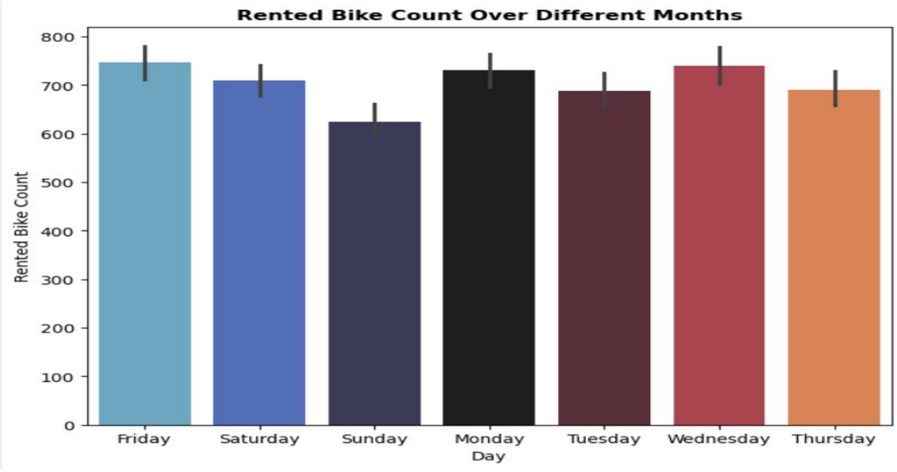
Exploratory Data Analysis(EDA) Contd...

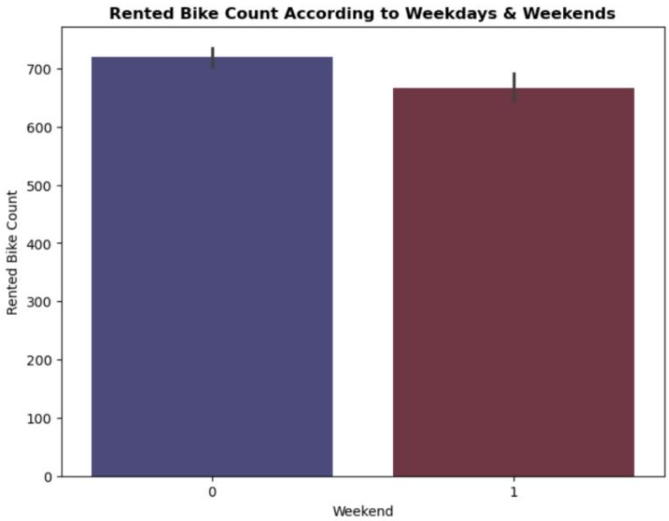


- As when there is a holiday the demand for rented bikes reduces.

- When it is a functioning day or no holiday the demand raises for the rented bike.

Exploratory Data Analysis(EDA) Contd...

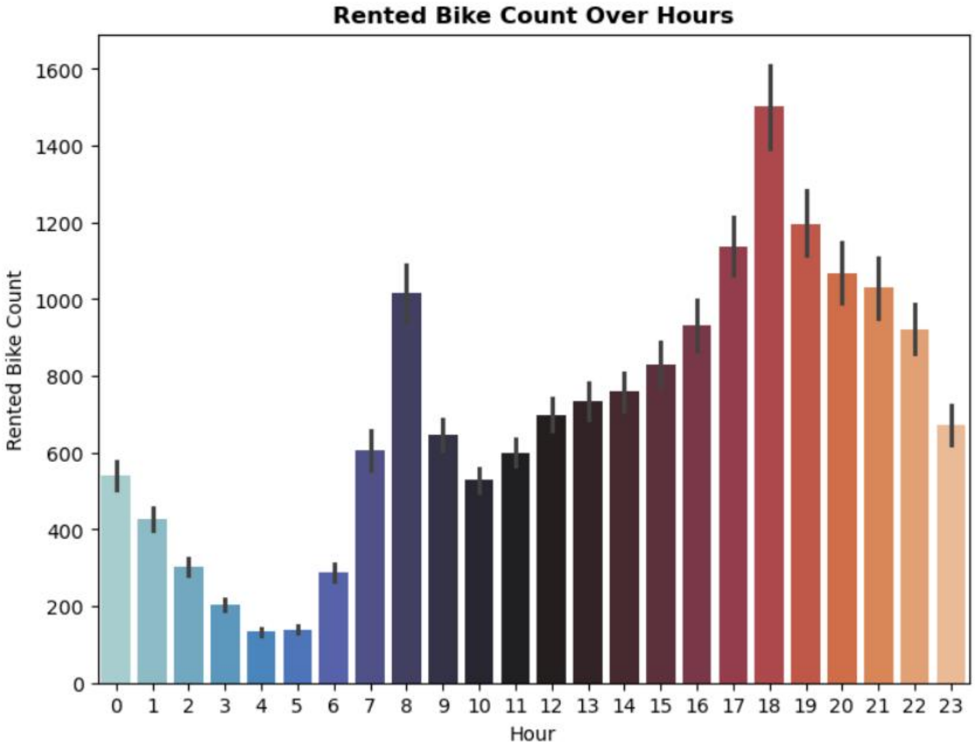




- We can see in the weekly graph that the demand for the rented bike is least on Sunday as it is a Holiday.

- 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 weekday

Exploratory Data Analysis(EDA) Contd…

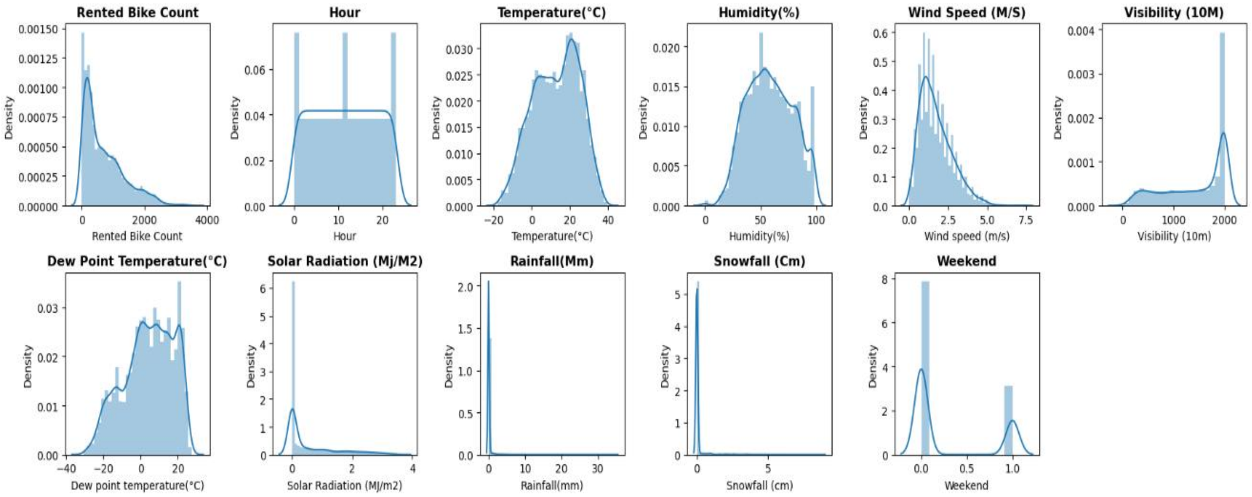


- The demand for the bikes rises the most in the evening around 5 -7 pm and the demand is highest at 6 pm.

- 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

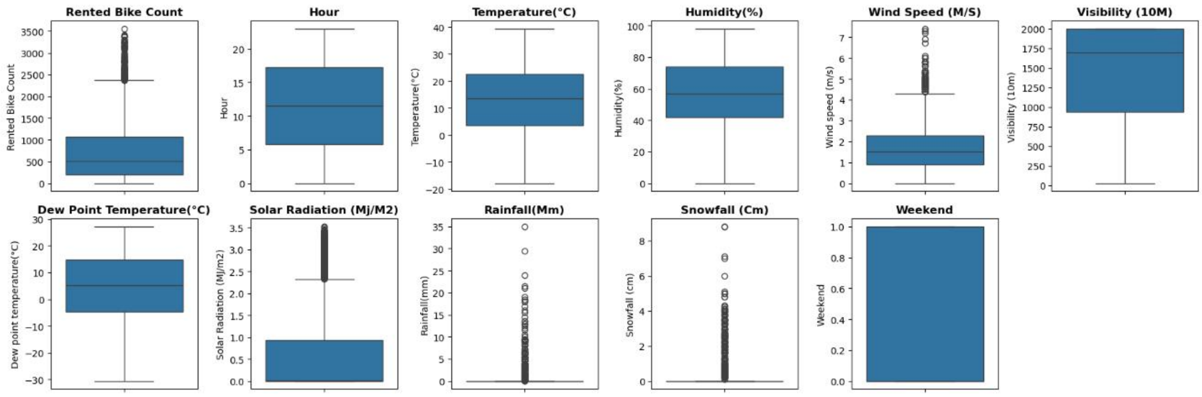


- Data distribution is checked on each column to see the skewness of the data and how much the data is normally distributed.

- 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.

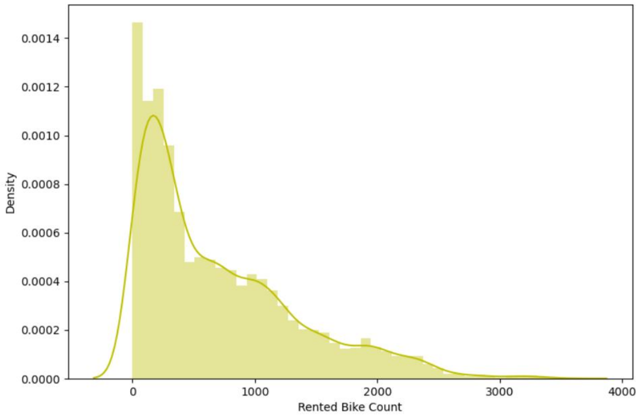
Visualizing Outliers

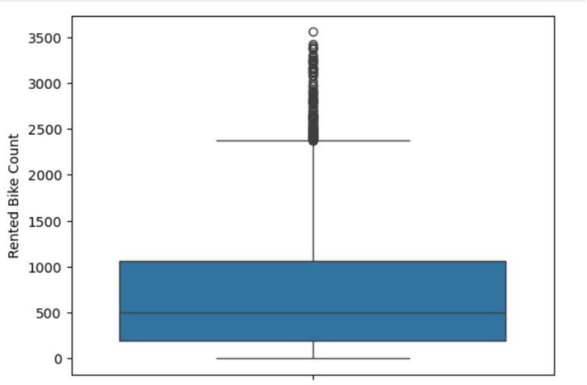


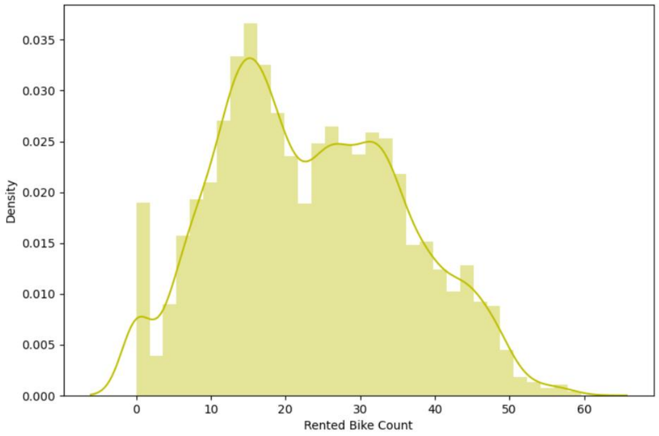
- We see outliers in some columns like Sunlight, Wind, Rainfall, and Snowfall but let’s not treat them because they may not be outliers as snowfall, rainfall, etc. themselves are rare events in some countries.

- We treated the outliers in the target variable by capping with IQR limits.

Handling Outliers



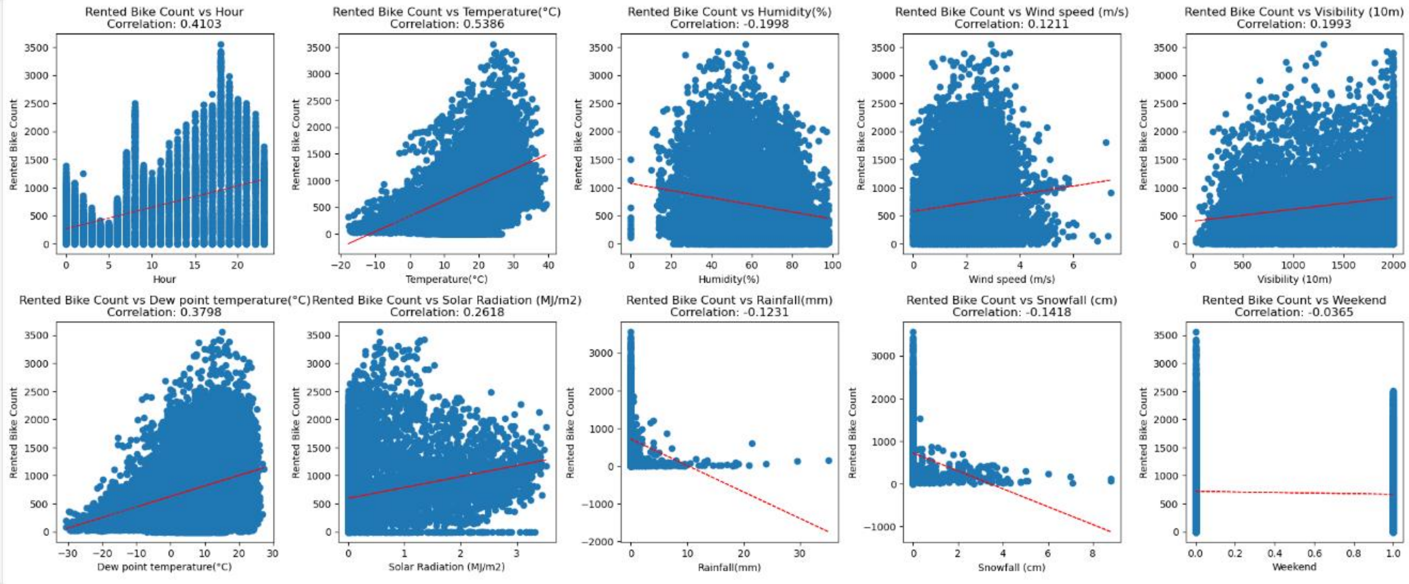




- Earlier the distribution of the target variable was positively skewed with a skewness value of 1.15. We tried to make this distribution somewhat close to normal distribution.

- 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.

Bi-variate Analysis of Linearity in Data

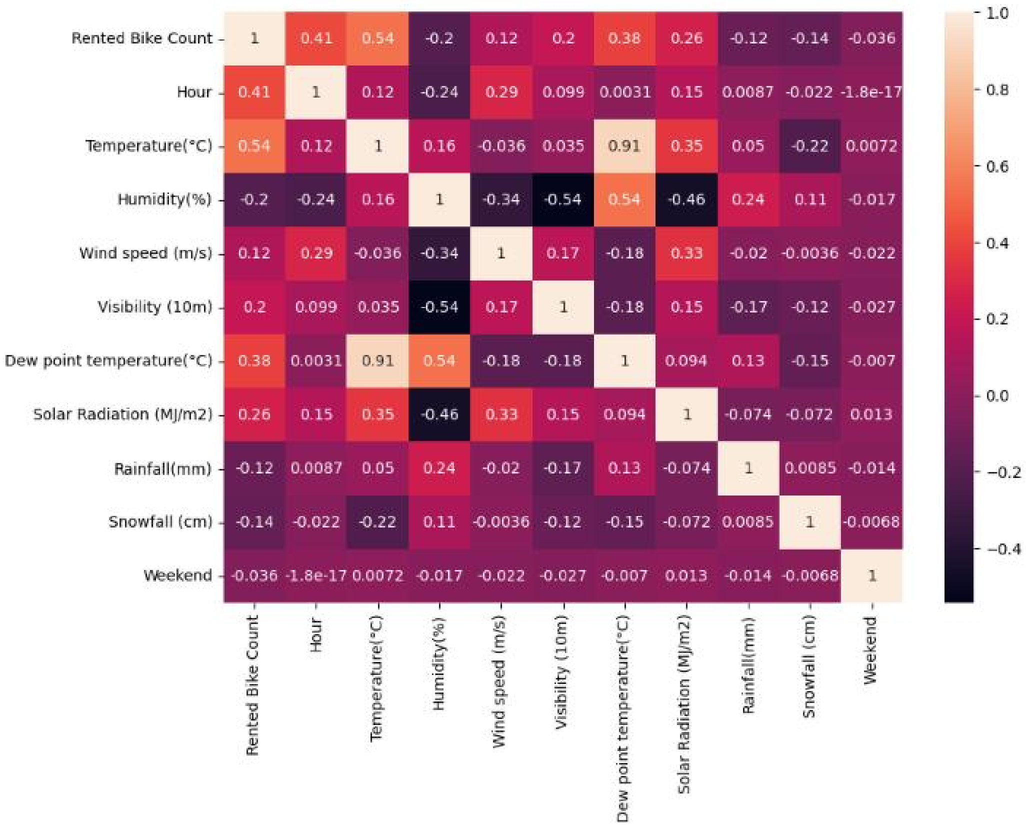


- From the visualizations we observed that hour, temp, sunlight, and dew temp are positively correlated with the bike count.

- 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 Heat-map



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

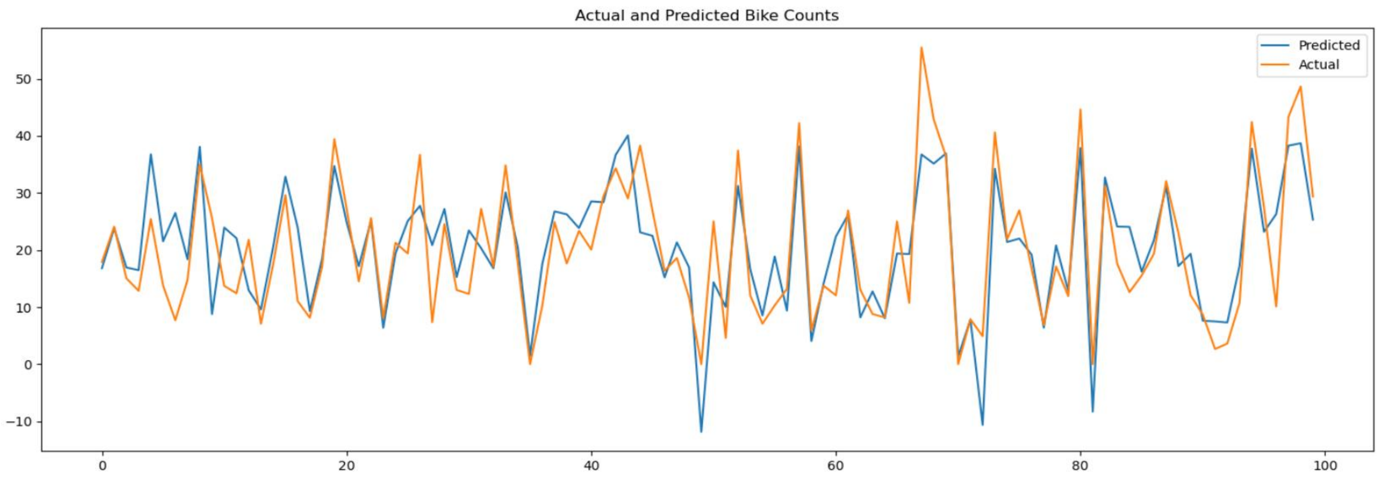
Model Implementation

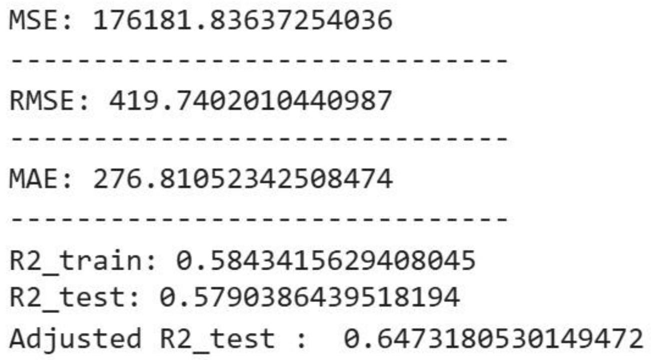
Linear Regression:

- 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:





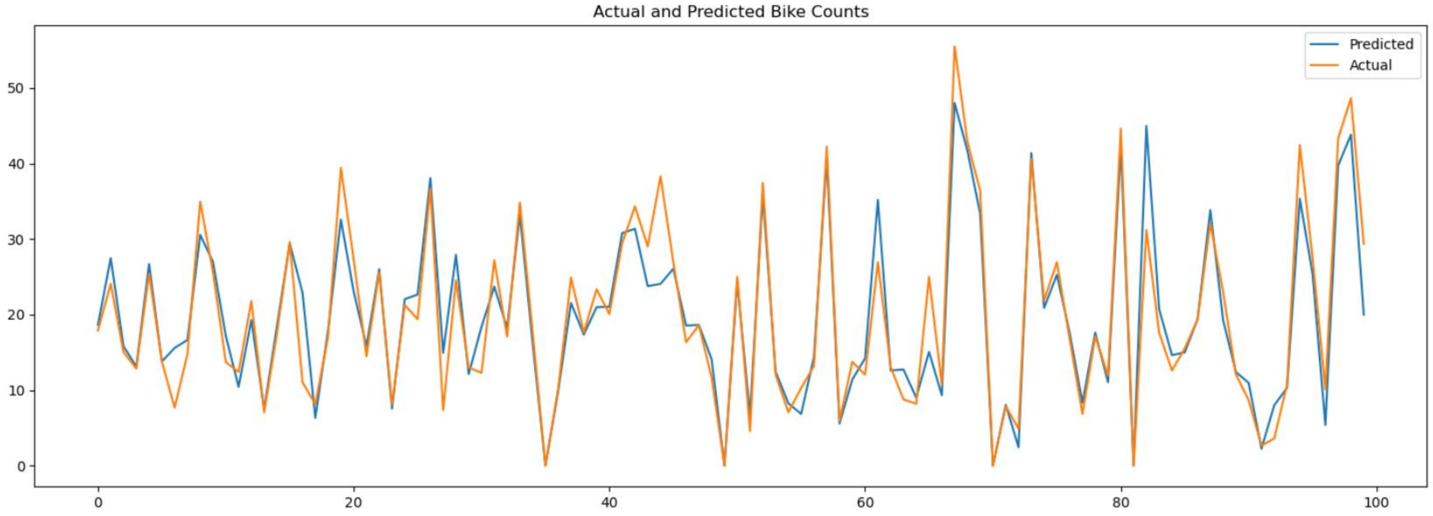
```
MSE: 176181.83637254036
------------------------------
RMSE: 419.7402010440987
------------------------------
MAE: 276.81052342508474
------------------------------
R2_train: 0.5843415629408045
R2_test: 0.5790386439518194
Adjusted R2_test :  0.6473180530149472
```

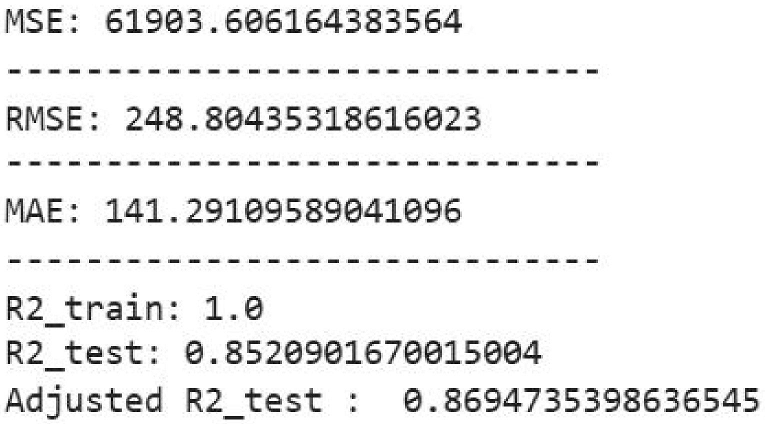
Model Implementation

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:



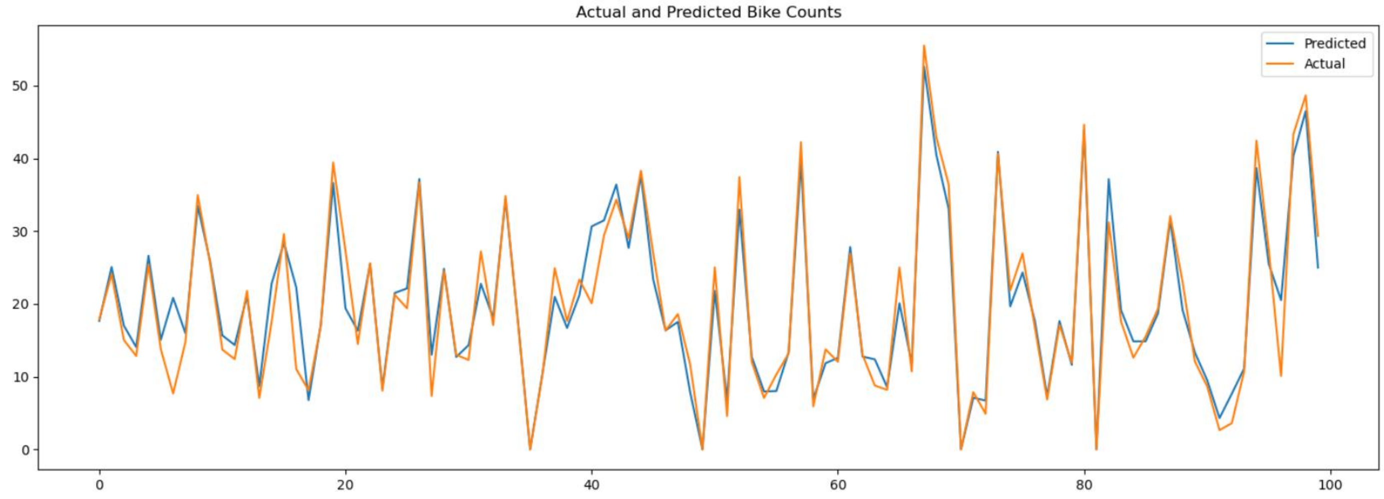


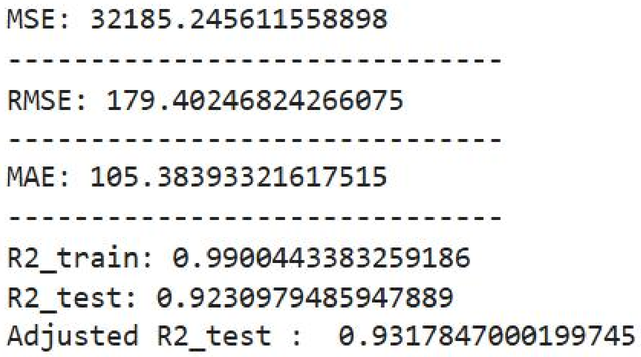
```
MSE: 61903.606164383564
------------------------------
RMSE: 248.80435318616023
------------------------------
MAE: 141.29109589041096
------------------------------
R2_train: 1.0
R2_test: 0.8520901670015004
Adjusted R2_test :  0.8694735398636545
```

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



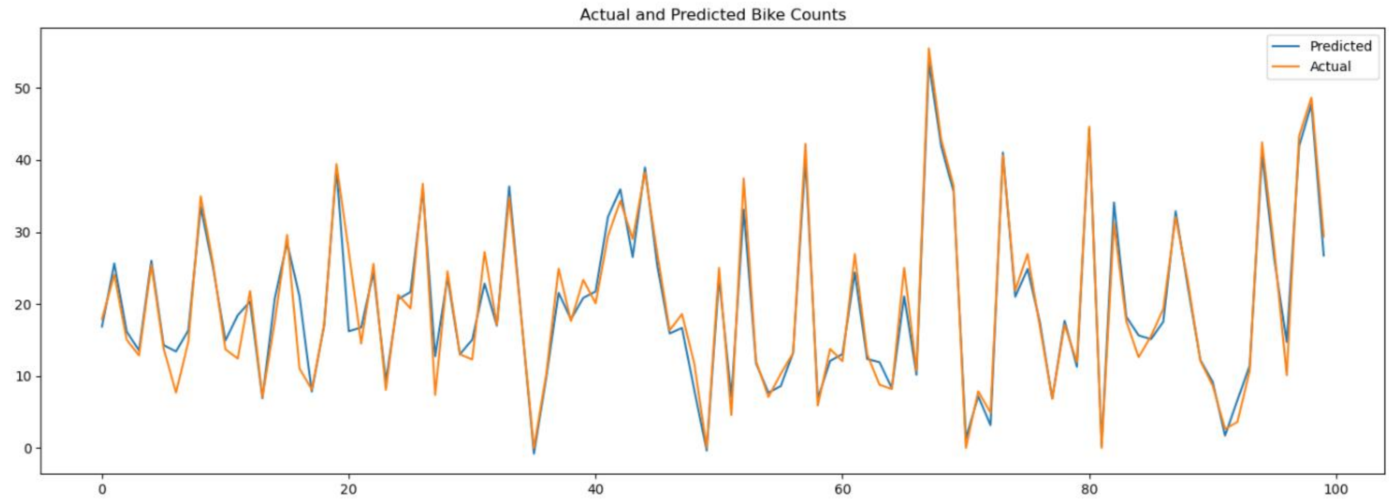


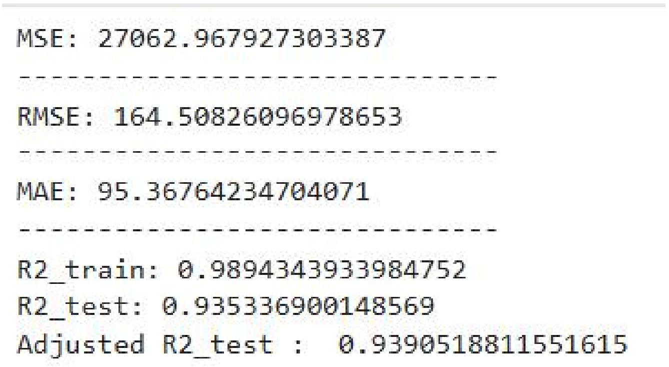
```
MSE:  32185.245611558898
------------------------------
RMSE:  179.40246824266075
------------------------------
MAE:  105.38393321617515
------------------------------
R2_train:  0.9900443383259186
R2_test:  0.9230979485947889
Adjusted R2_test :   0.9317847000199745
```

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 = 0.1, max_depth = 8, n_estimators = 200) the data performed quite well and given the best result among all the algorithms.





```
MSE: 27062.967927303387
------------------------------
RMSE: 164.50826096978653
------------------------------
MAE: 95.36764234704071
------------------------------
R2_train: 0.9894343933984752
R2_test: 0.935336900148569
Adjusted R2_test :  0.9390518811551615
```

Model Summary

We began with exploratory data analysis and then pre-processed the data, converting the category columns 'Seasons,' 'Functional day,' and 'Holiday' into numerical columns.

- **Linear Regression :** The model explains approximately 58% of the variability in the test data (R2_test: 0.579), indicating a moderate predictive performance. The plot shows that the predictions generally follow the trend of the actual bike counts, capturing the main fluctuations. However, a key weakness is that the model predicts negative counts, which is illogical for real-world data like bike rentals.

- **Decision Tree:** The model is severely overfitting the data, as shown by the perfect training R-squared score of 1.0, which means it has memorized the training set. Despite this, it performs strongly on unseen test data, explaining about 85% of the variability in bike counts (R2_test: 0.852). Visually, the model's predictions track the actual counts much more closely than the previous model and do not appear to make impossible negative predictions.

- **Random Forest**: The model demonstrates excellent predictive power, explaining over 92% of the variability in the test data (R2_test: 0.921). However, it shows strong signs of overfitting, as the near-perfect training R-squared of 0.998 indicates it has almost completely memorized the training data. Visually, this is the best-performing model so far, with predictions that track the actual bike counts with very high accuracy.

- **XGBoost:** This XGBoost model is the most effective in all, explaining 93.5% of the variability in the test data (R2_test: 0.935) and achieving the lowest error rates of all the models tested. While it also shows signs of overfitting with a training R-squared of 0.998, its performance on unseen data is superior. Overall, this model provides the most accurate predictions, as the plot visually confirms the tightest alignment between predicted and actual bike count

Finally, after reviewing all of the model scores, we can conclude that **XGBoost is the best model** to use in the future.

Conclusion

As indicated in the problem description, the company was founded in 2017. As a result, the quantity of motorcycles rented in 2017 was insufficient.

- We can observe that the number of rental bikes in 2018 was 5986984, which was more than in 2017.

- We may say that the number of rental bikes is substantially larger on non-holiday days than on holidays.

- We may say that the number of rental bikes is substantially larger on non-holiday days than on holidays.

- The number of business hours each day and the demand for leased motorcycles were the most closely associated.

- The most bikes were rented at the 18th hour of the day.

- After attempting various feature combinations with linear regression, the model was found to be under fit. Because data is dispersed so widely, it became evident. Fitting a line didn't seem realistic.

- The most critical factors for forecasting the number of bikes needed were the hour, temperature, and solar radiation.

- Rainfall and snowfall have a significant impact on the quantity of bikes leased, with a relatively high downfall..

- With good model performance and low RMSE, the Random Forest Regressor outperforms linear regression.

- The **Top performing model was XgBoost** , which outperformed trees algorithms. **XGBoost has lower error (lower RMSE/MAE) and a higher R-squared score**, meaning its predictions are more accurate and it explains more of the variance in the data

- In 2018, the number of bicycle rentals climbed considerably. Demand fell in the most recent month of 2018, after initially growing towards the end of 2017. This is because demand began to rise dramatically in 2017 and has continued to rise in the early months of 2018. There is a decrease at the end of the year. This could possibly be due to the cold weather.

- The demand surge began towards the end of 2017, during the winter season. An observer might find it strange that demand declined at the end of 2018. Indeed, it can be argued that the company's growth in this situation increased considerably from April 2017 to April 2018. As a result, while demand increased throughout the winter of 2017, it still fell short of its full potential. Using simple heuristics, we can forecast that demand will fall in December, but in proportion to demand for the entire year, assuming all other independent factors remain constant.

----- END OF THE REPORT ------