Capstone Project

Bike Sharing Demand Prediction

By: Vikash Kumar

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.

Steps performed

1. Data cleaning
2. Data visualizations
3. Data preprocessing
4. Model Implementation
5. Evaluation metrics

Data Description

The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour, and date information.

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

Data Cleaning

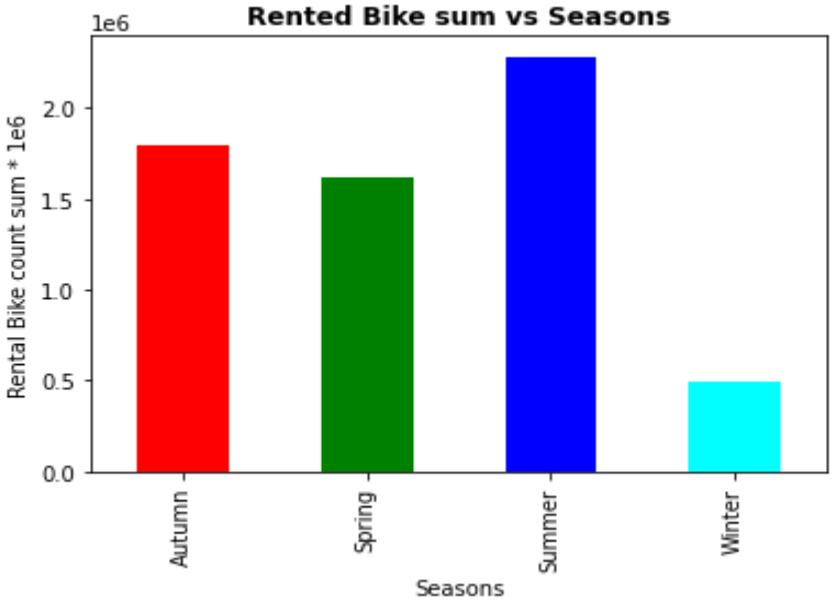
Prior to EDA, cleaning the data is essential since it will get rid of any ambiguous information that can have an impact on the results.

Since null values don't even exist, there's no reason to remove them.

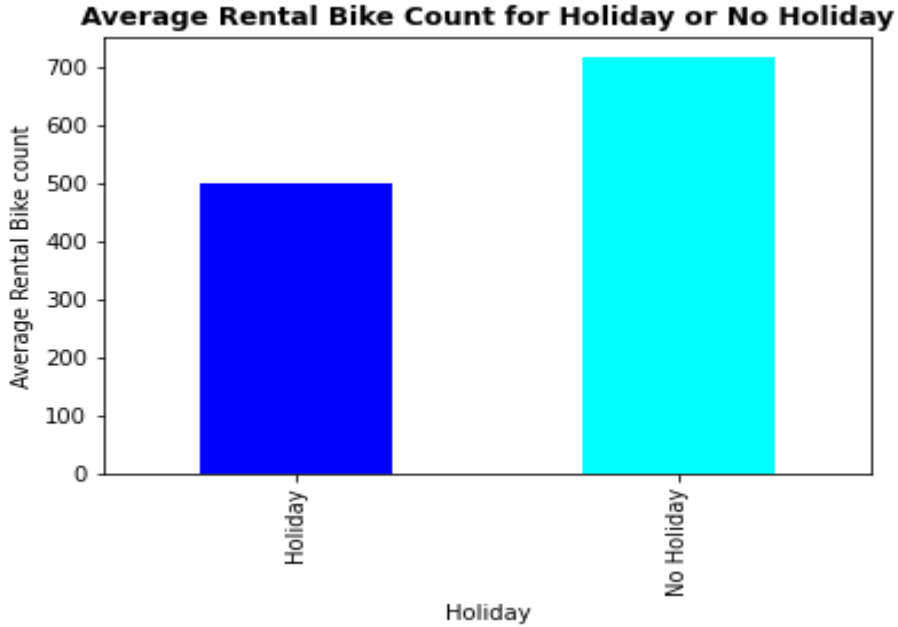
data.isna().sum()

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

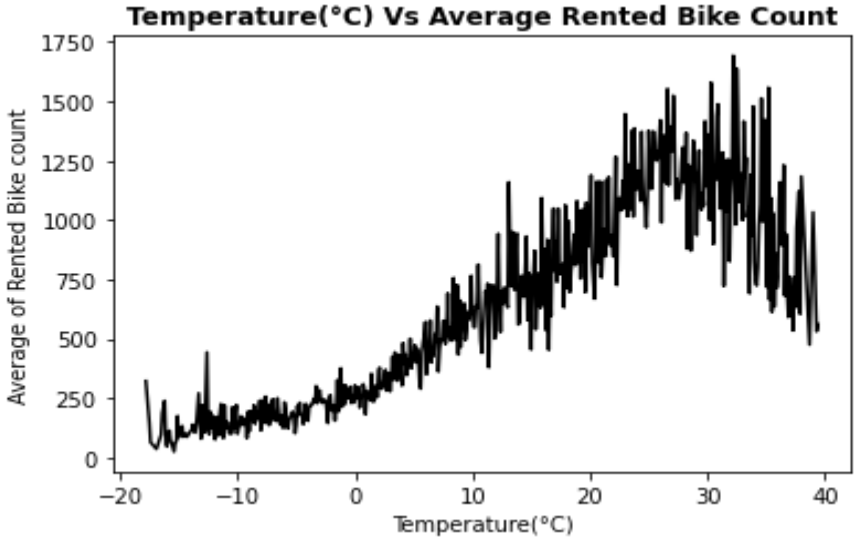
Exploratory Data Analysis

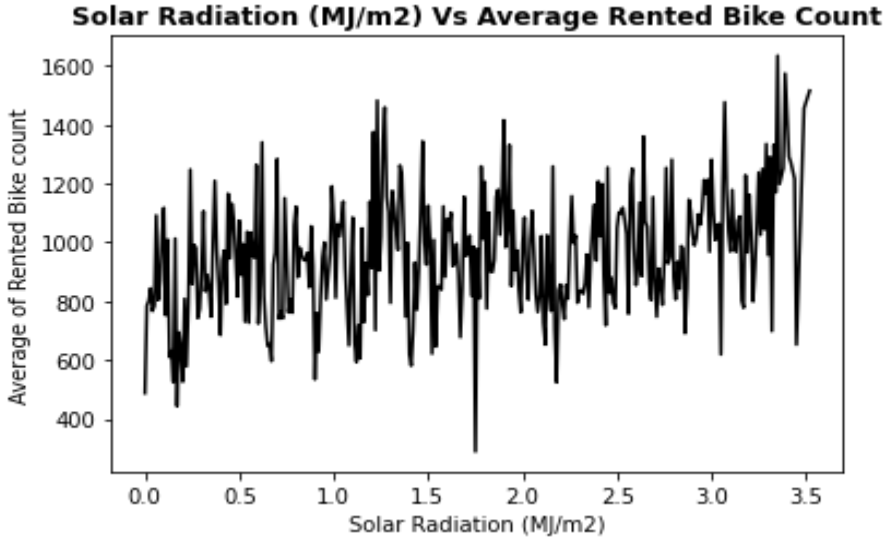


During the summer, a lot of bikes are rented as compared to other seasons.



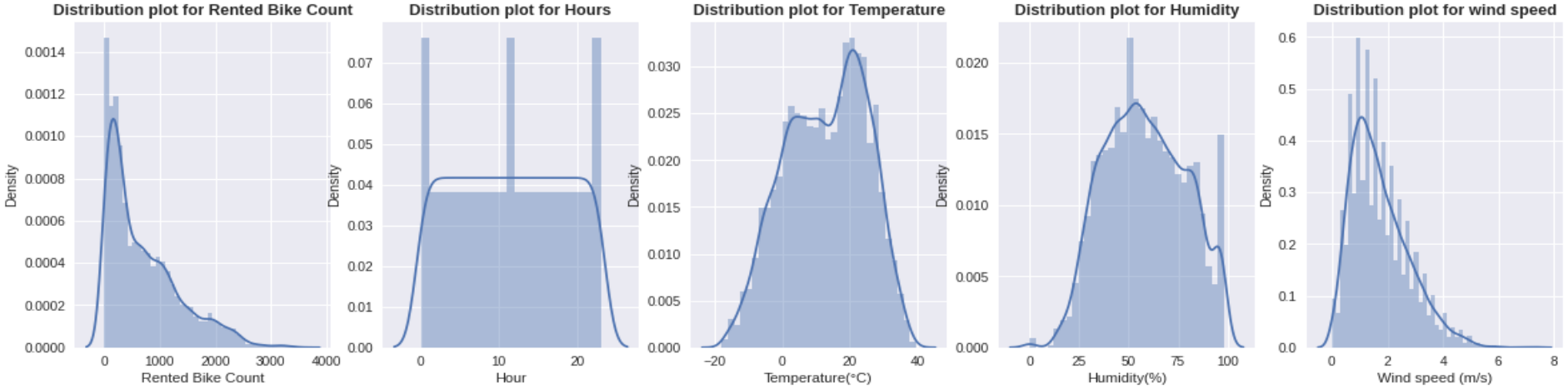
The average number of rental bikes is high outside of holidays. That suggests a large proportion of people ride rental bikes to work.

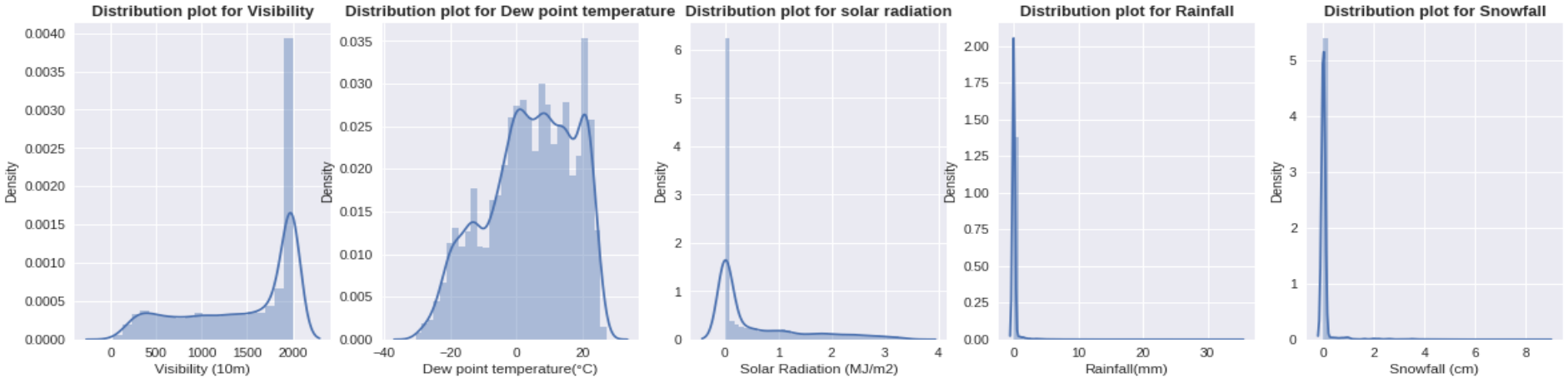


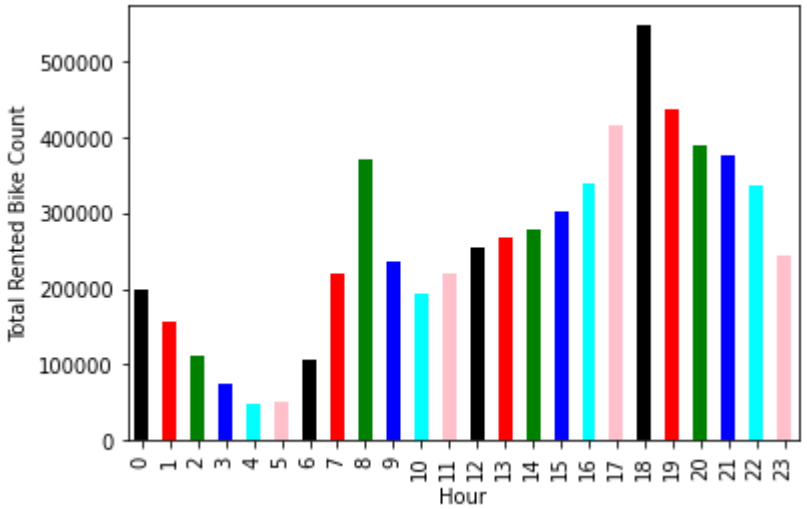


The number of rented bikes grows as the temperature rises, as seen by the line graph. However, the number of rental bikes starts to decline beyond 30 degrees Celsius.
But there is no effect of Solar Radiation.

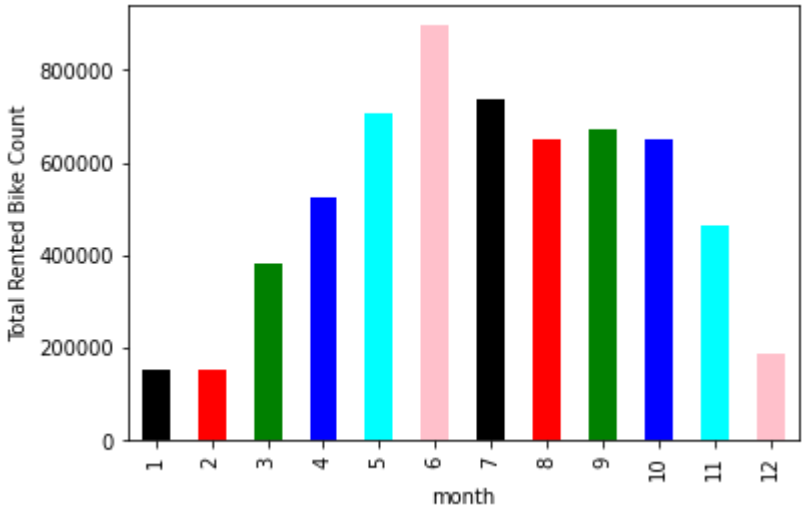
Distribution plots for numerical features (Independent and Dependent features)



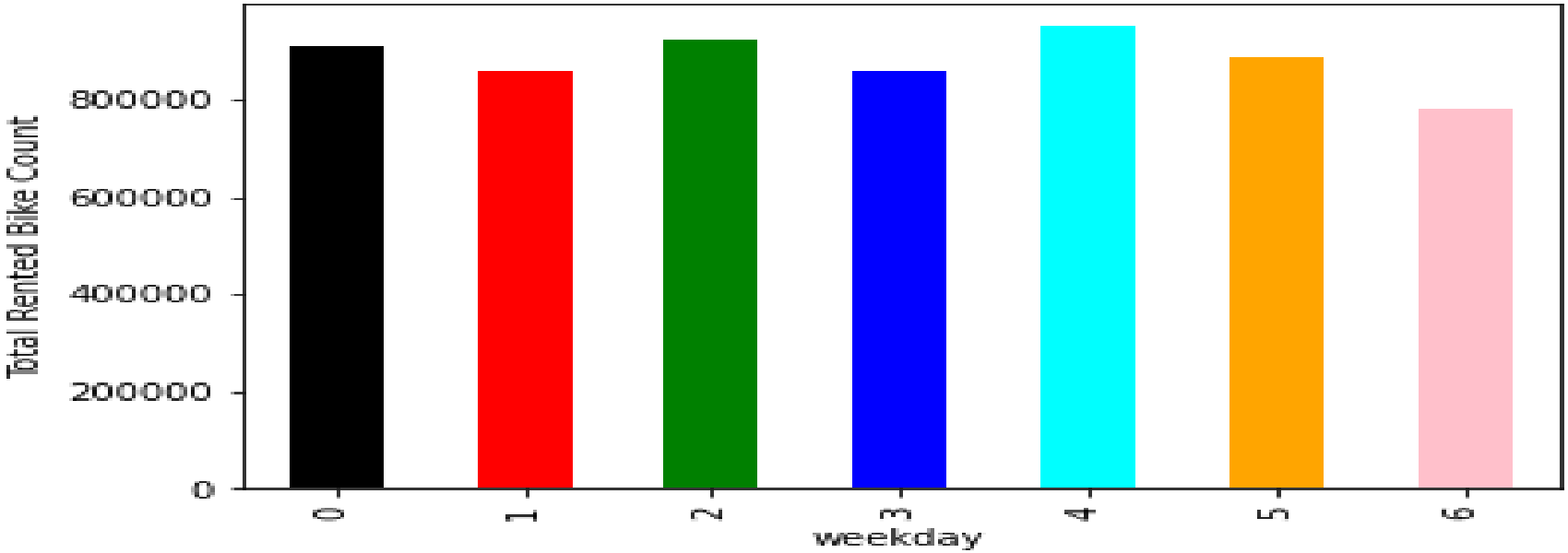




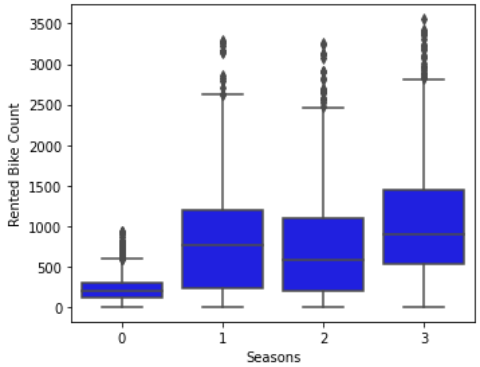
People travel to or from work primarily around morning 9 AM and evening 7 PM. The increased rental bike count during those hours may therefore be due to this. Additionally, more people prefer to hire bikes when they are leaving work than when they are going.

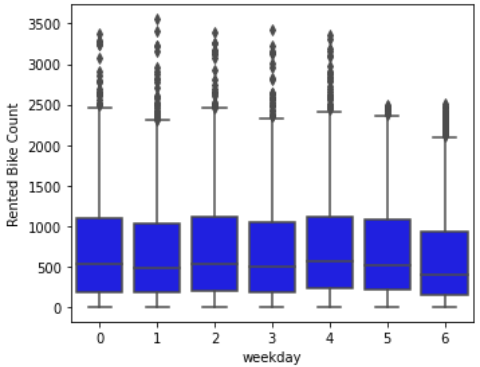


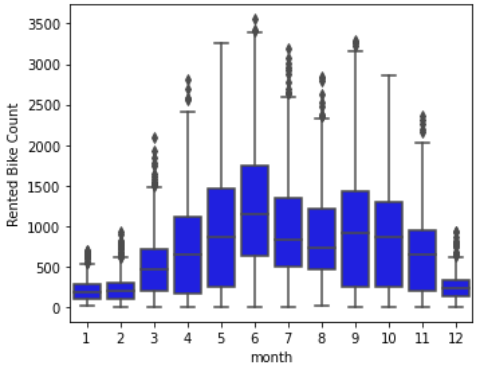
The most bikes are rented in the month of June. Additionally, the lowest number of bikes are rented in January and February.

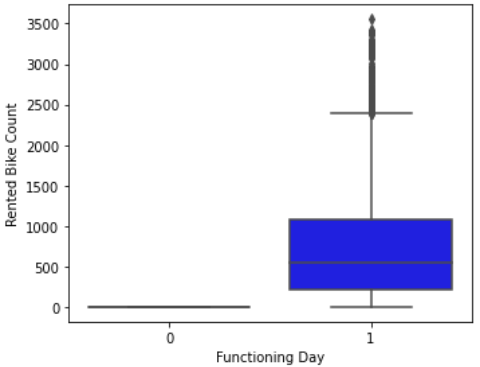


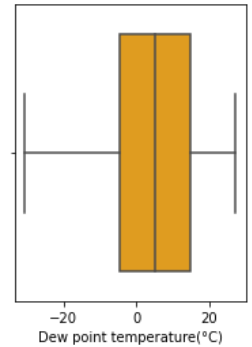
For Sundays the rented bikes are used less as compared to other days

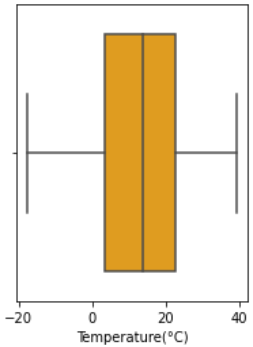


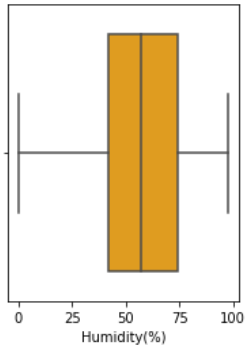


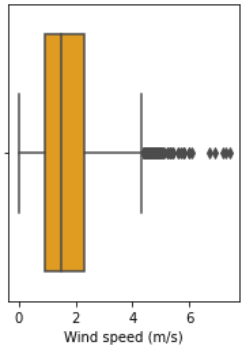


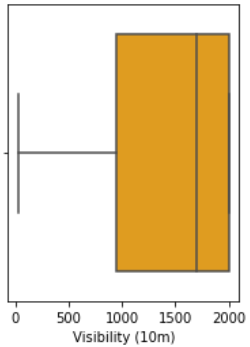


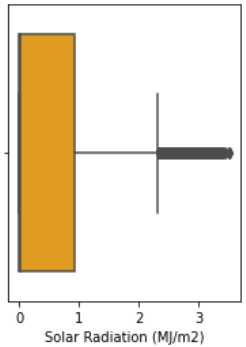


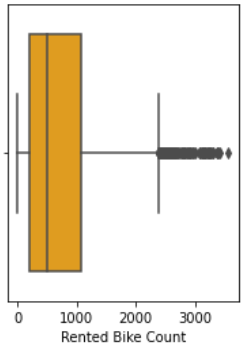




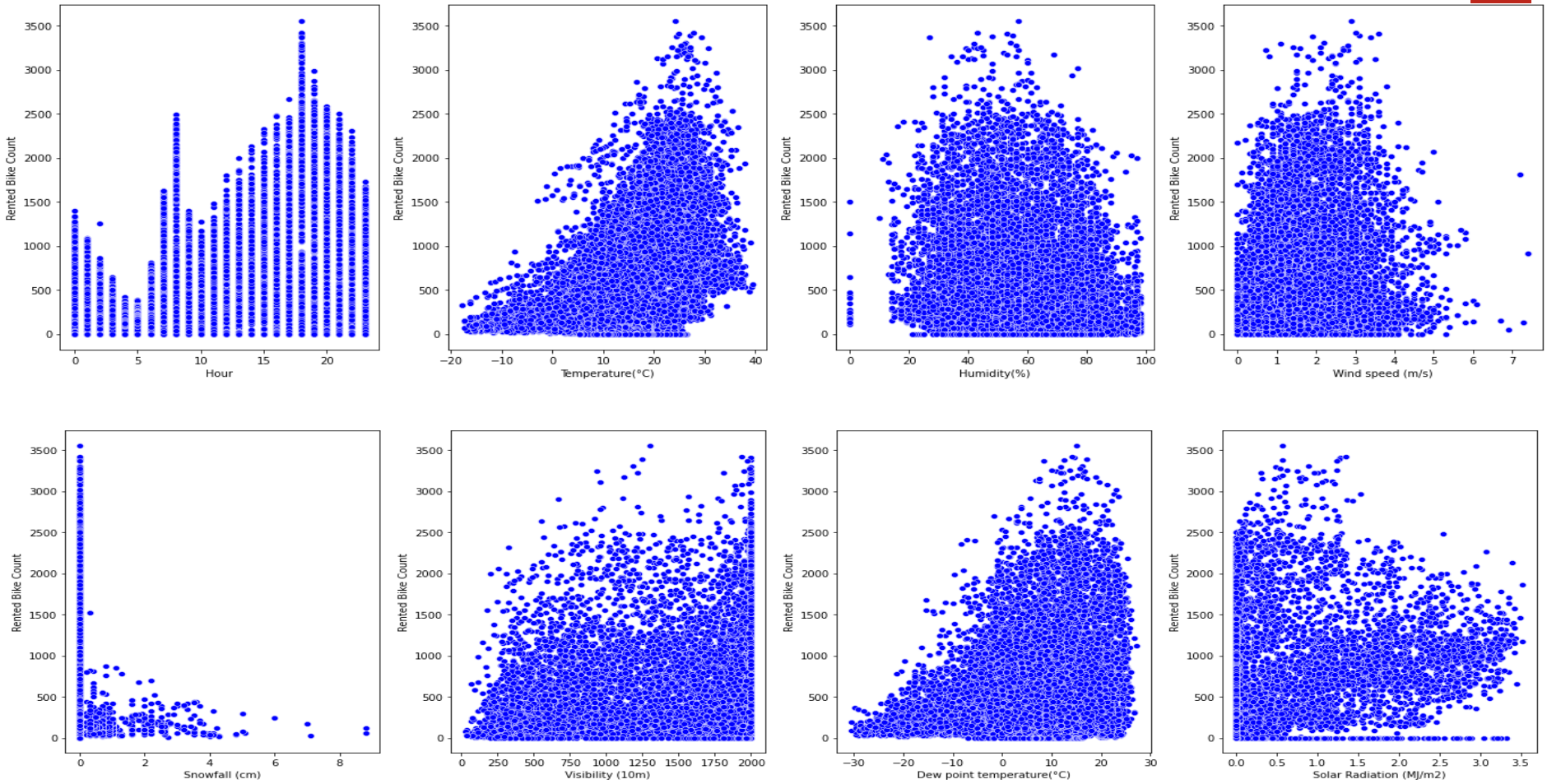


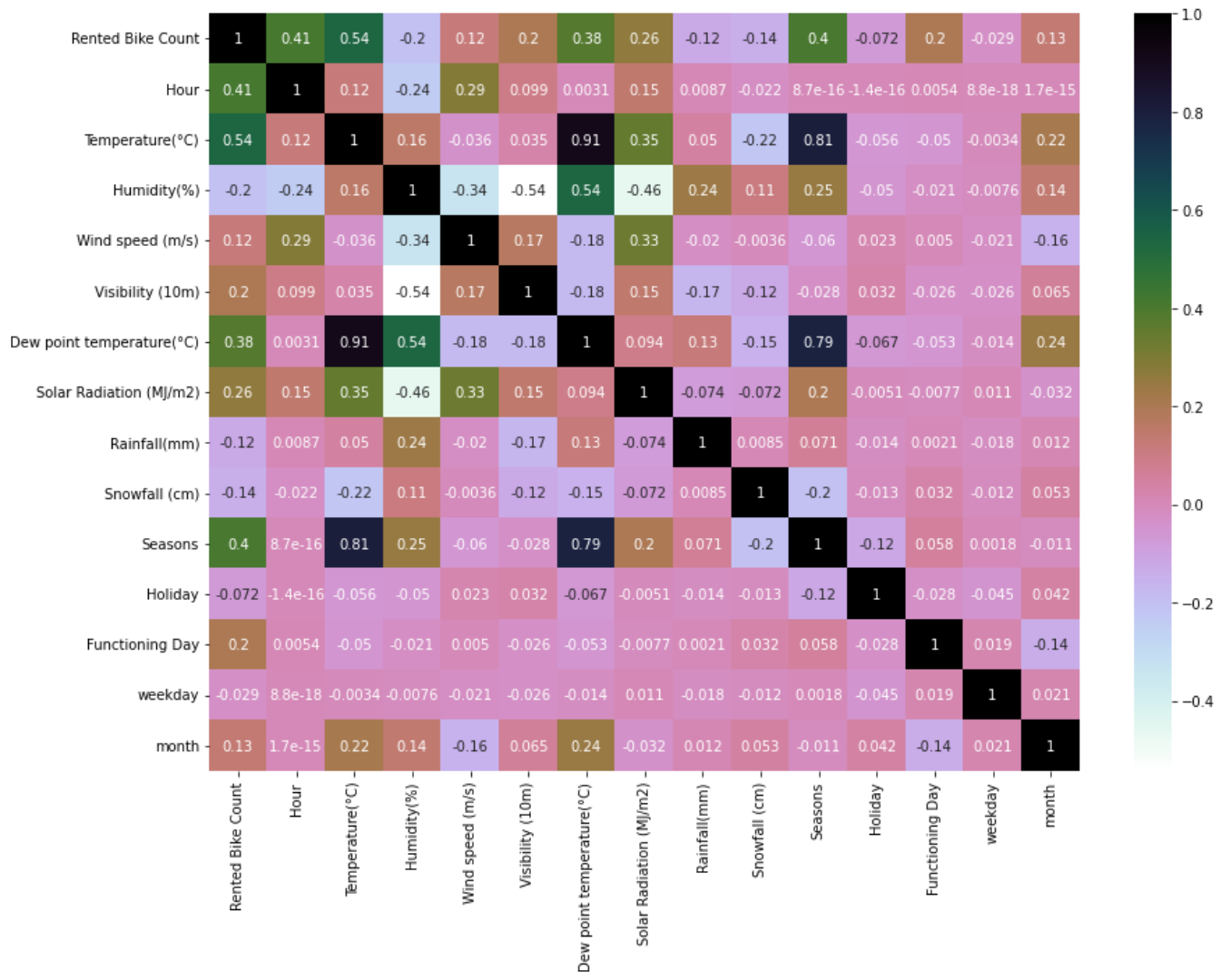






Here 'hours', 'temperature', 'humidity', 'visibility', and 'due point temperature' has no outliers.
Not deleting these outliers because occasionally doing so results in considerable data loss.


Rented Bike Count
Hour
Temperature(°C)
Humidity(%)
Wind speed (m/s)
Snowfall (cm)
Visibility (10m)
Dew point temperature(°C)
Solar Radiation (MJ/m2)



Heatmap helps to find the correlation between the features. While implementing Linear Regression, features having high collinearity will be removed.

From the correlation heatmap, it is clear that only Temperature, Hour, Dew point Temperature, and solar radiation are highly correlated with the Rented bike. All other features are less correlated. The temperature has a very high correlation with dew point temperature and hence both have the same variation so dew point temperature can be eliminated.

Data preparation

The high correlation between :

1. Temperature and Seasons
2. Temperature and Dew Temperature
3. Dewpoint Temperature and seasons
4. Humidity and dew point temperature

Dropping Dew Point Temperature, Humidity, and Seasons columns from dataset.

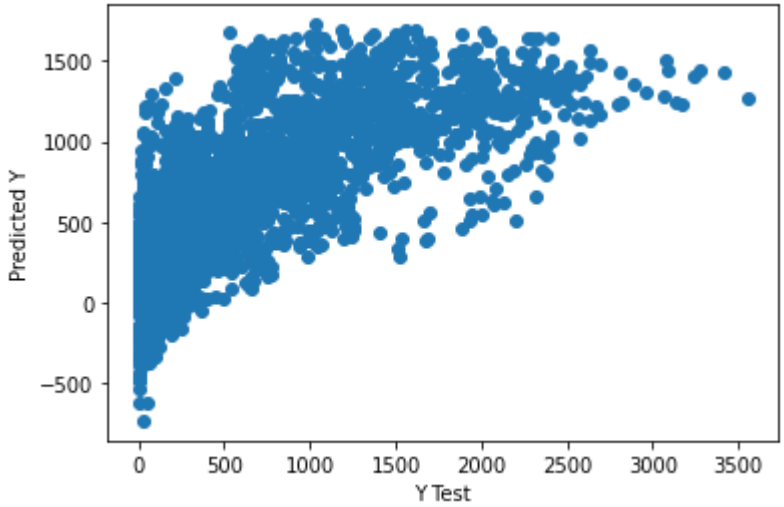
Added the columns like date, weekday, and month from date columns.

Linear Regression

Evaluation metrics for training data
MSE is 201743.44262392577
RMSE is 449.15859406664566
R2 square is 0.5132089109583456
MAE is 332.4687428359205
Adjusted R2 score is2.1352286675216368

Evaluation metrics for test data
MSE is 210880.57353383838
RMSE is 459.2173489033688
R2 square is 0.4973096526324337 6
MAE is 340.3131776303789
Adjusted R2 score is1.2153216276756231

scatterplot of the real test values versus the predicted values



Conclusion

Interpreting the coefficients:

1. Holding all other features fixed, a 1 unit increase in Hour is associated with an increase of 30.41 Rented Bike Count.

2. Holding all other features fixed, a 1 unit increase in Temperature(°C) is associated with an increase of 26.54 Rented Bike Count.

3. Holding all other features fixed, a 1 unit increase in Wind speed (m/s) is associated with an increase of 14.70 Rented Bike Count.

4. Holding all other features fixed, a 1 unit increase in Visibility (10m) is associated with an increase of 0.14 Rented Bike Count.

5. Holding all other features fixed, a 1 unit increase in Solar Radiation (MJ/m2) is associated with an increase of 5.24 Rented Bike Count.

6. Holding all other features fixed, a 1 unit increase in Rainfall (mm) is associated with an increase of -72.23 Rented Bike Count.

7. Holding all other features fixed, a 1 unit increase in Time on Snowfall (cm) is associated with an increase of -28.80 Rented Bike Count.

8. Holding all other features fixed, a 1 unit increase in Holiday is associated with an increase of -139.94 Rented Bike Count.

9. Holding all other features fixed, a 1 unit increase in Functioning Day is associated with an increase of 843.44 Rented Bike Count.

10. Holding all other features fixed, a 1 unit increase in weekdays is associated with an increase of -12.56 Rented Bike Count.

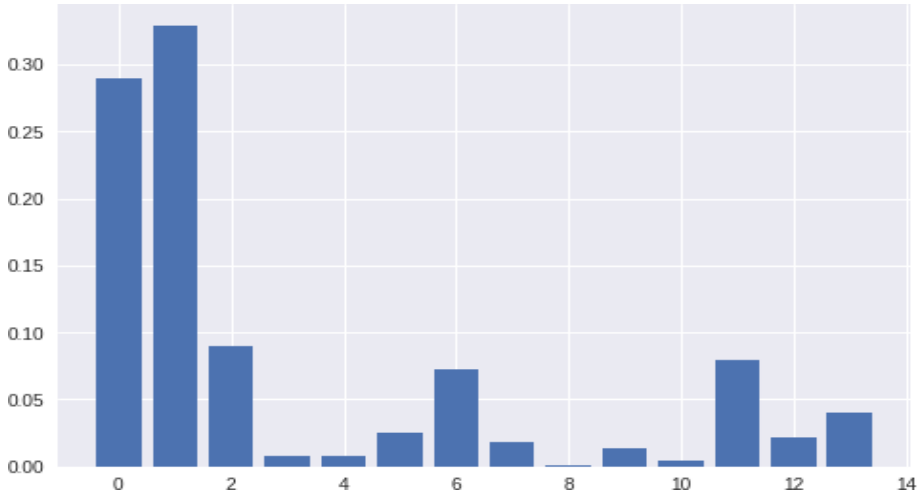
11. Holding all other features fixed, a 1 unit increase in the month is associated with an increase of 12.06 Rented Bike Count.

Decision Tree Model

Feature Importances

```
Feature: 0, Score: 0.28233
Feature: 1, Score: 0.34606
Feature: 2, Score: 0.09697
Feature: 3, Score: 0.00321
Feature: 4, Score: 0.00415
Feature: 5, Score: 0.02088
Feature: 6, Score: 0.07751
Feature: 7, Score: 0.01524
Feature: 8, Score: 0.00079
Feature: 9, Score: 0.01571
Feature: 10, Score: 0.00244
Feature: 11, Score: 0.08176
Feature: 12, Score: 0.01909
Feature: 13, Score: 0.03387
```

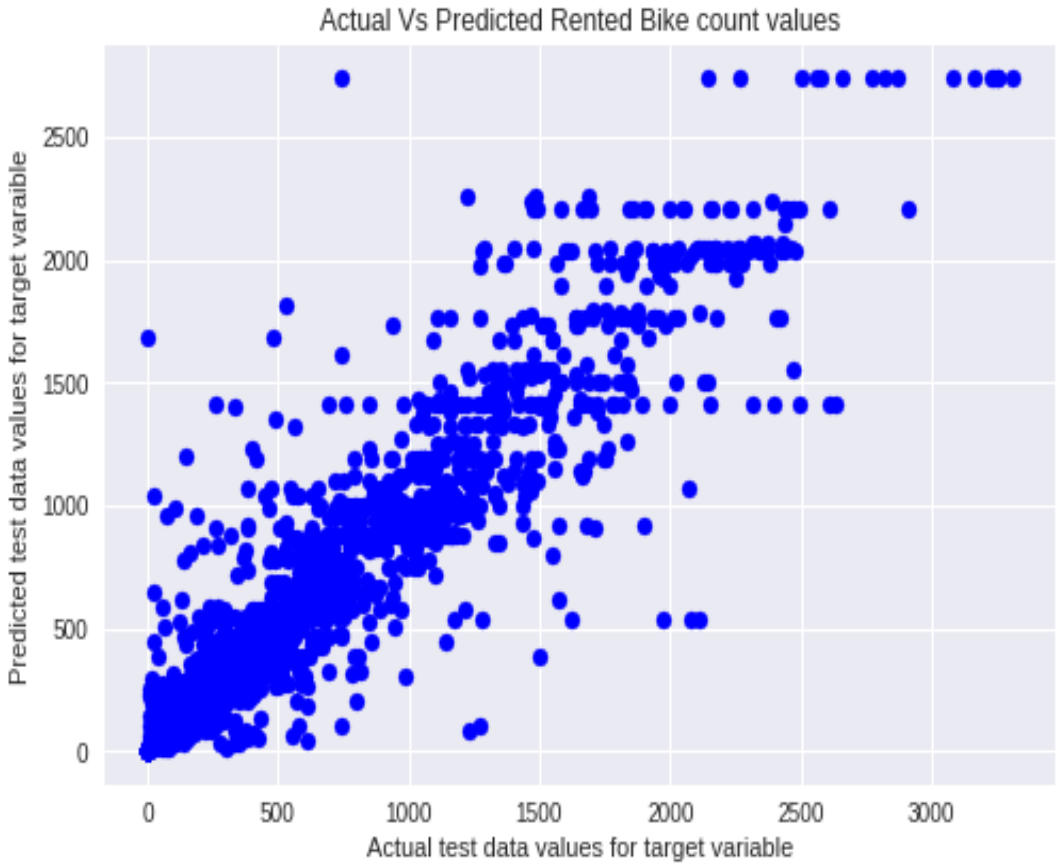
Effects of various variables on the anticipated value in the application of the decision tree model



Feature 0 :'Hour', Feature 1: 'Temperature(℃)',Feature 2: 'Humidity(%)',Feature 3:'Wind speed (m/s)',Feature 4: 'Visibility (10m)', Feature 5: 'Dew point temperature(℃)', Feature 6:'Solar Radiation (MJ/m2)',Feature 7:'Rainfall(mm)',Feature 8: 'Snowfall (cm)', Feature 9:'Seasons',Feature 10:'Holiday',Feature 11: 'Functioning Day', Feature 12:'month', Feature 13:'weekday'

Evaluation metric for training data
MSE is 7614.5078413785595
RMSE is 87.26114737601472
R2 square is 0.9816267903610103
MAE is 45.56648184387911
Adjusted R2 score is1.042847526929116

Evaluation metric for testing data
MSE is 59180.02197752832
RMSE is 243.26944316442277
R2 square is 0.8589285617609046
MAE is 140.31751014713342
Adjusted R2 score is1.0604263277766353



Random Forest model

Evaluation metric for training data
MSE is 24438.768784439922
RMSE is 156.32904011871858
R2 square is 0.94103116950576
MAE is 98.27296734358215
Adjusted R2 score
is1.1375191706961525

Evaluation metric for test data
MSE is 44384.78898029837
RMSE is 210.676977812713
R2 square is 0.8941969636346702
MAE is 135.78959652014996
Adjusted R2 score
is1.045319513538516 5

Conclusion

We have analyzed Seoul city bike sharing dataset. Through analysis, we saw that in general, the number of bike rents in 2018 was more than in 2017. The highest number of bike rents occur in summer while the least bike rents occur in winter. On daily basis, the trend of bike rents is almost similar with slight peaking demands on Thursday while drops on Sunday. On hourly basis, the bike counts peak in the afternoon (from 15.00 to 20.00). There are two peak occurrences, at 7.00 and at 17.00, which are most likely to be caused by workers going to office in the morning and going back home in the afternoon.

The hourly movement of bike counts seems to correlate with temperature, visibility, windspeed, and humidity. The bike counts peak in the afternoon (from 15.00 to 20.00) when temperature is the highest, with the most visibility, windspeed, and least humidity. There were days on a weekday when the bike sharing facility was not functioning. However, during public holidays, the facility was still operating.

All measures, including MSE (Mean Squared Error), MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), R2 Score, and Adjusted R2 Score, were evaluated for each model.

Based on this analysis,
Future events can be predicted with 49.73% accuracy using linear regression.
With an accuracy rate of 85.69%, decision trees can forecast the future.
Random Forest has a 92.18% accuracy rate for future prediction.