Capstone Project – II
Supervised ML Regression

Bike Sharing Demand Prediction

Contents

- ☐ Evaluation matrices
- ☐ Graph of Actual v/s Predicted values of bike rent count prediction
- ☐ Model 4 – Ridge Regression
- ☐ Evaluation matrices
- ☐ Graph of Actual v/s Predicted values of bike rent count prediction
- ☐ Model 5 – Decision Tree
- ☐ Evaluation matrices
- ☐ Graph of Actual v/s Predicted values of bike rent count prediction
- ☐ Model 6 – Random Forest Regression
- ☐ Evaluation matrices
- ☐ Graph of Actual v/s Predicted values of bike rent count prediction
- ☐ Feature Importances for predicting bike rent count
- ☐ Conclusion of Project

Problem Statement

Bike sharing systems have become a popular mode of transportation in many cities, providing a low-cost and environmentally-friendly option for commuters. However, bike sharing companies face the challenge of predicting the demand for bikes at different times and locations, in order to optimize their inventory and ensure that bikes are available when and where they are needed.

The problem statement is to develop a model that accurately predicts the demand for bikes based on historical data on factors such as time, weather, and location. This model will help bike sharing companies to make data-driven decisions about inventory management and pricing, and ultimately improve the user experience for riders.

Objective

The goal of predicting bike sharing demand is to accurately forecast the number of bikes that will be rented out at different times and locations, based on historical data and other relevant factors. This will help bike sharing companies to optimize their inventory and pricing strategies, and ensure that bikes are available when and where they are needed. Ultimately, the goal is to improve the user experience for riders by providing a reliable and convenient mode of transportation that meets their needs.

Data Summary

- The dataset contains the following weather information: temperature, humidity, wind speed, visibility, dew point, solar radiation, snowfall, and rainfall.
- Additionally, the dataset includes information on the number of bikes rented per hour and date.
- The Bike sharing dataset covers rental bike counts for every hour and day from 2017 to 2018 within the Capital bike share system.
- The dataset also provides corresponding weather and seasonal information.
- In total, the dataset contains 8,760 rows, representing every hour of each day in 2017 and 2018.
- The dataset has 14 columns, representing the features that are being considered.

DataSet Description

Date	Year-month-day
Rented Bike count	Count of bikes rented at each hour
Hour	Hour of he day (0 to 23)
Temperature	Temperature of the day in degree Celsius
Humidity	Humidity measurement in %
Wind speed	Wind speed in m/s
Visibility	Visibility measurement around 10 meter
Dew point temperature	Dew point measurement in degree Celsius
Solar radiation	Solar radiation measurement in MJ/m2 (i.e. Mega Jules per meter square)
Rainfall	Rainfall measurement in mm
Snowfall	Snowfall measurement in cm
Seasons	Winter, Spring, Summer, Fall or Autumn
Holiday	Holiday/No holiday
Functional Day	No Func(Non Functional Hours), Fun(Functional hours)

Exploratory Data Analysis

```
In [3]: bike_df.head()
```

Out[3]:

	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	2017-01-12	254	0	-5.2	37	2.2	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
1	2017-01-12	204	1	-5.5	38	0.8	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
2	2017-01-12	173	2	-6.0	39	1.0	2000	-17.7	0.0	0.0	0.0	Winter	No Holiday	Yes
3	2017-01-12	107	3	-6.2	40	0.9	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
4	2017-01-12	78	4	-6.0	36	2.3	2000	-18.6	0.0	0.0	0.0	Winter	No Holiday	Yes

Checking for Null Values

```
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
dtype: int64
```

No null or missing values found in dataset

Check for Duplicated Values

```
In [20]: bike_df.duplicated().sum()
Out[20]: 0
```

Description of Numerical/Categorical analysis

```
In [31]: numeric_features
```

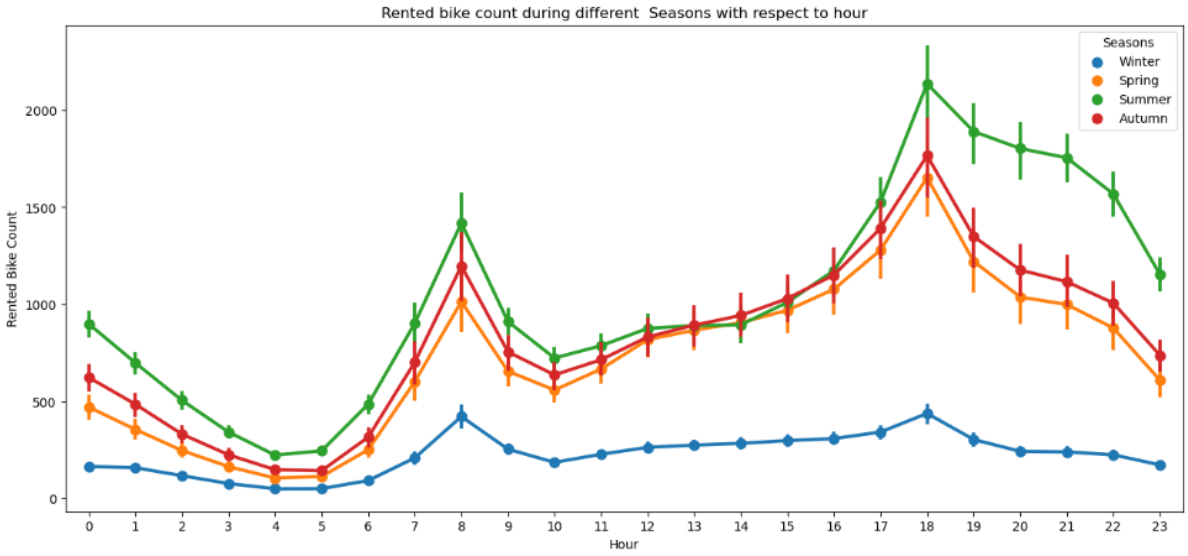
```
Out[31]: Index(['Rented Bike Count', 'Temperature(°C)', 'Humidity(%)',
                'Wind speed (m/s)', 'Visibility (10m)', 'Dew point temperature(°C)',
                'Solar Radiation (MJ/m2)', 'Rainfall(mm)', 'Snowfall (cm)'],
               dtype='object')
```

```
In [32]: categorical_features
```

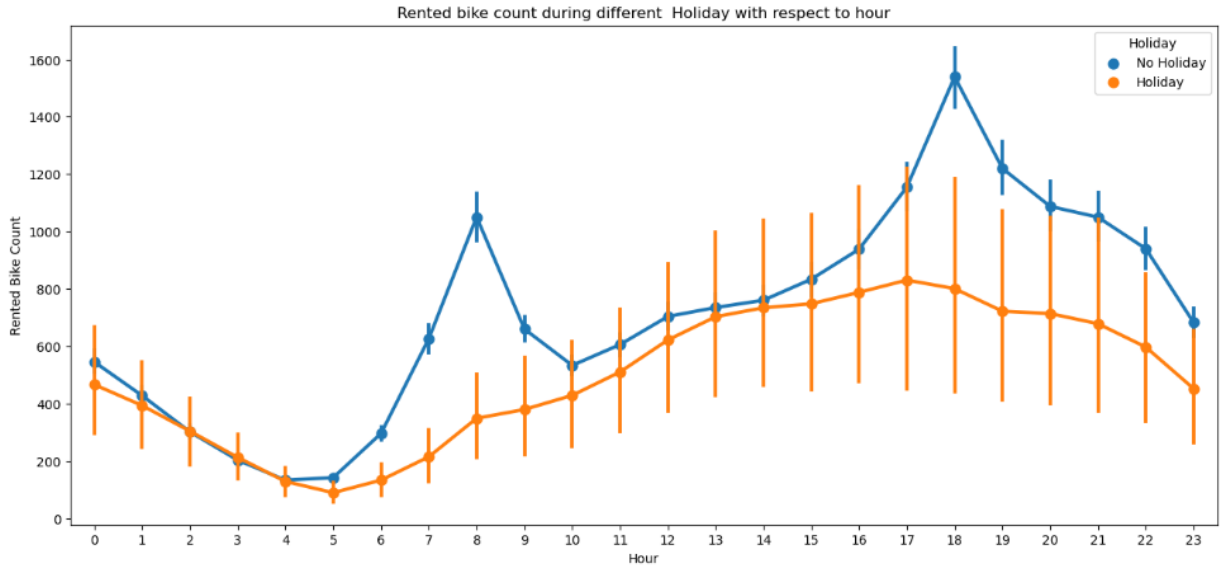
```
Out[32]: Index(['Hour', 'Seasons', 'Holiday', 'Functioning Day', 'year', 'month',
                'day'],
               dtype='object')
```

Visualizing Rented Bike Count, Hour with Respect to different categorical Feature

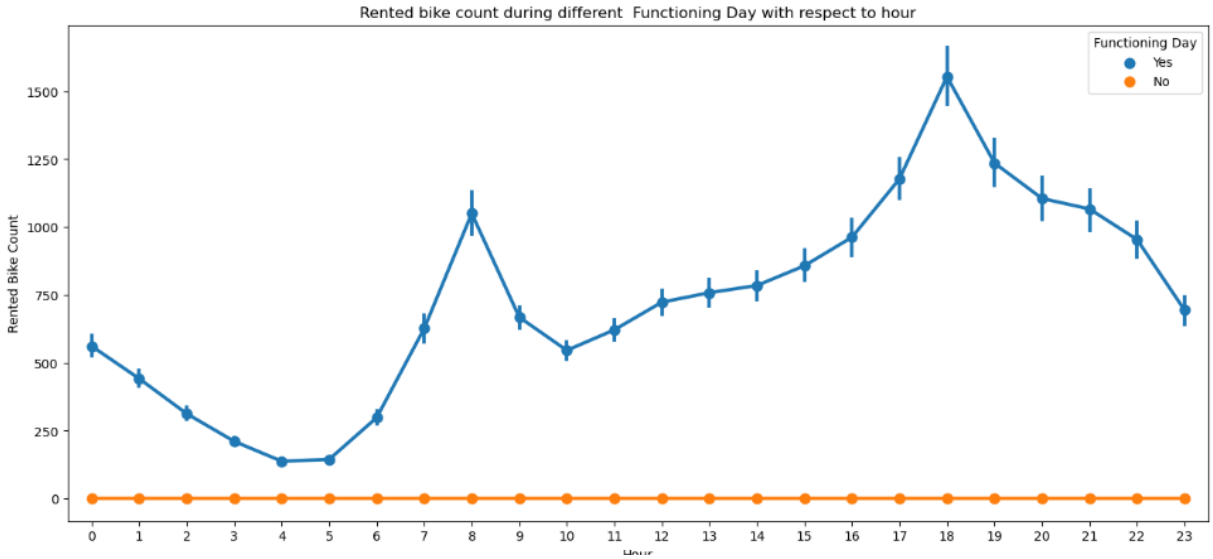
Session



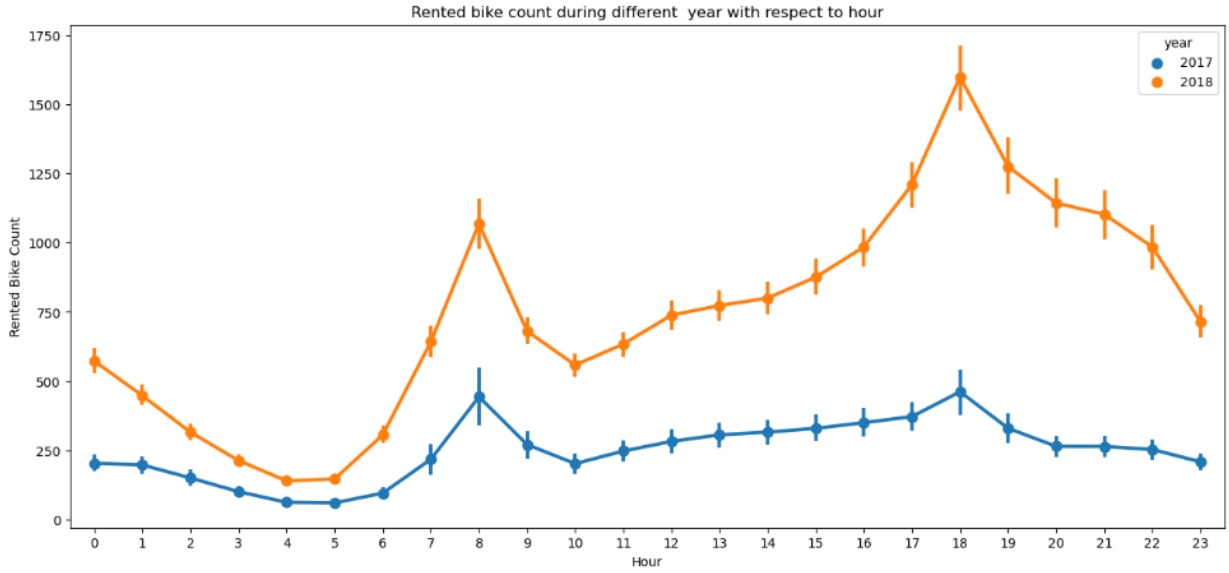
Holiday



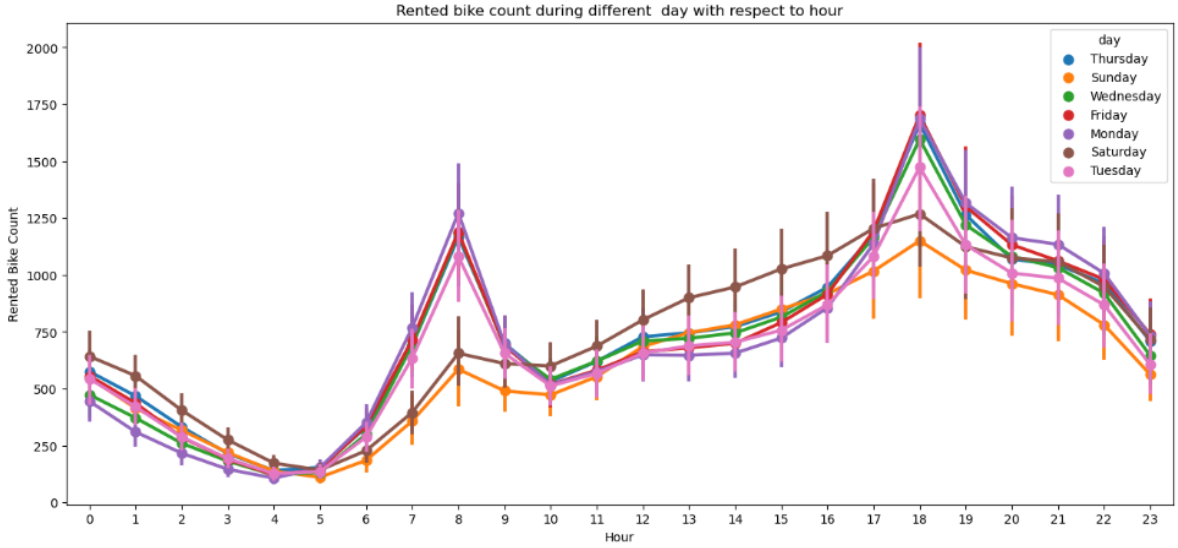
Functioning Day



Year



Day



Observation

Seasonal demand for bikes is lower during winter and higher during summer, as evident from the Season column.

Demand during holidays is lower compared to non-holidays, possibly due to people using bikes for commuting to work, as stated in the Holiday column.

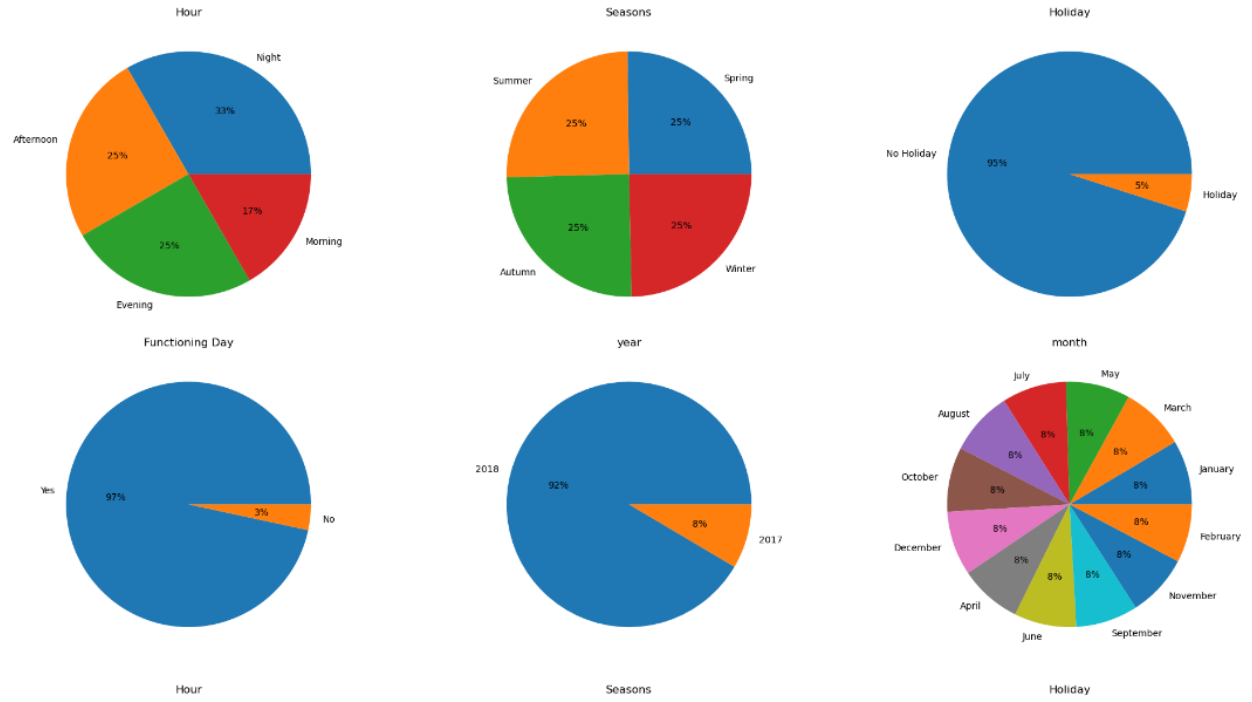
No demand exists on days when there is no Functioning Day, as mentioned in the Functioning Day column.

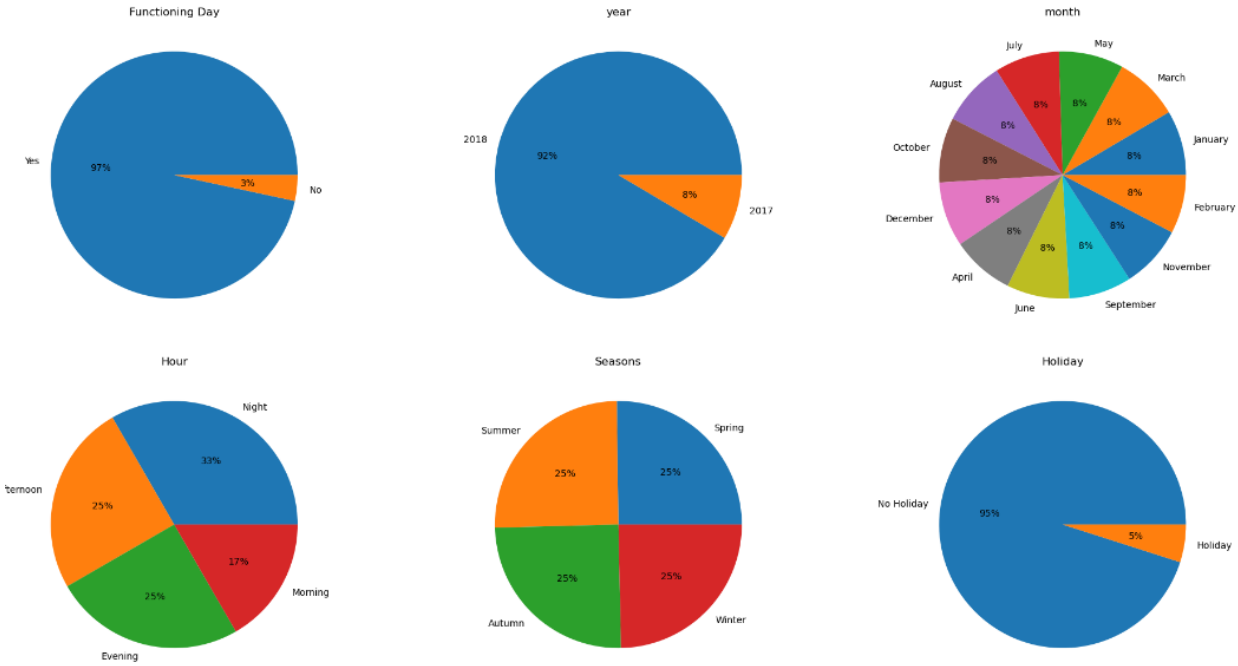
Demand for rented bikes increased in 2018 as compared to 2017, which could be attributed to increased awareness about rented bike facilities in 2017.

The Days of week column indicates a different demand pattern on weekdays and weekends. Afternoon demand on weekends is higher, while office hours show higher demand during weekdays.

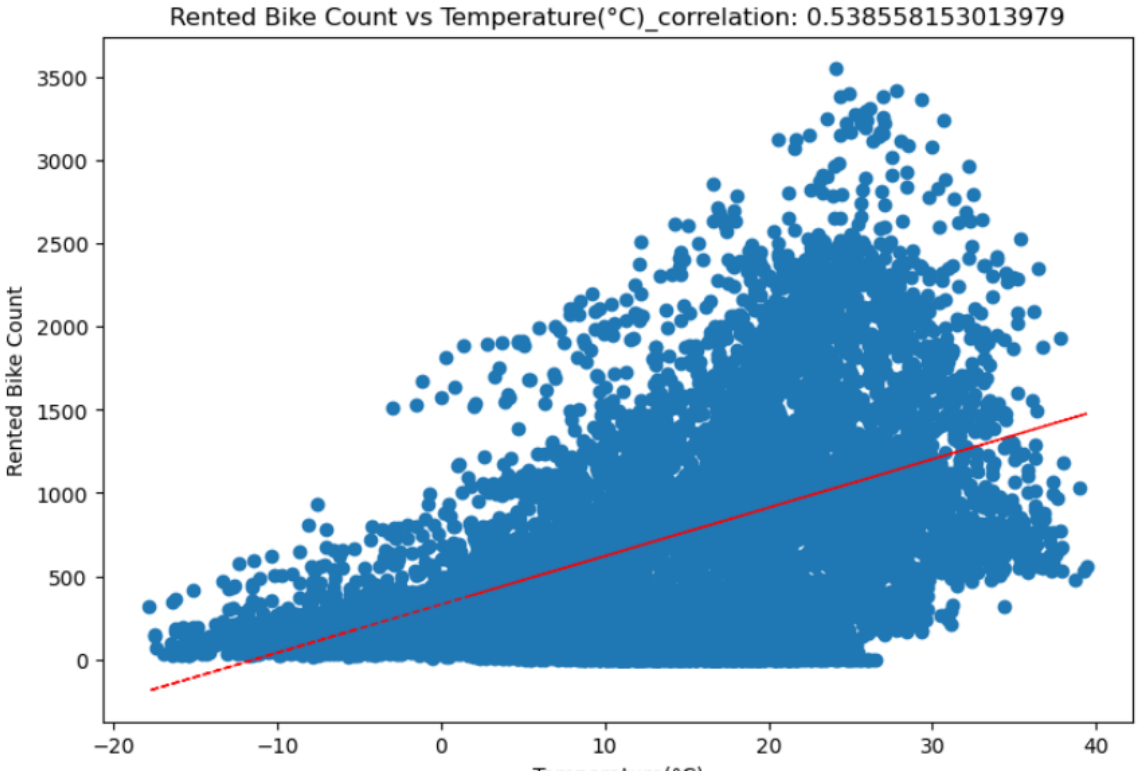
Demand is lower in December, January, and February as these months are cold and the Season column already established low demand during winters. This information is clearly visible in the Month column.

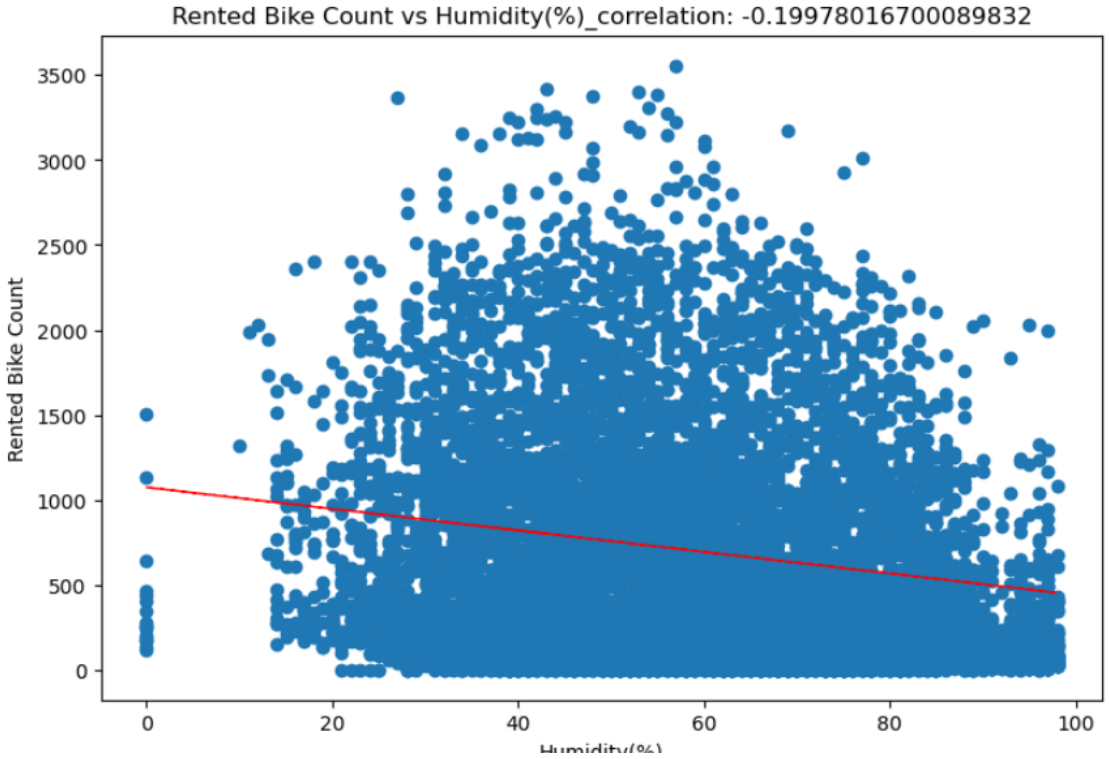
Visualizing Value Count of Categorical Features

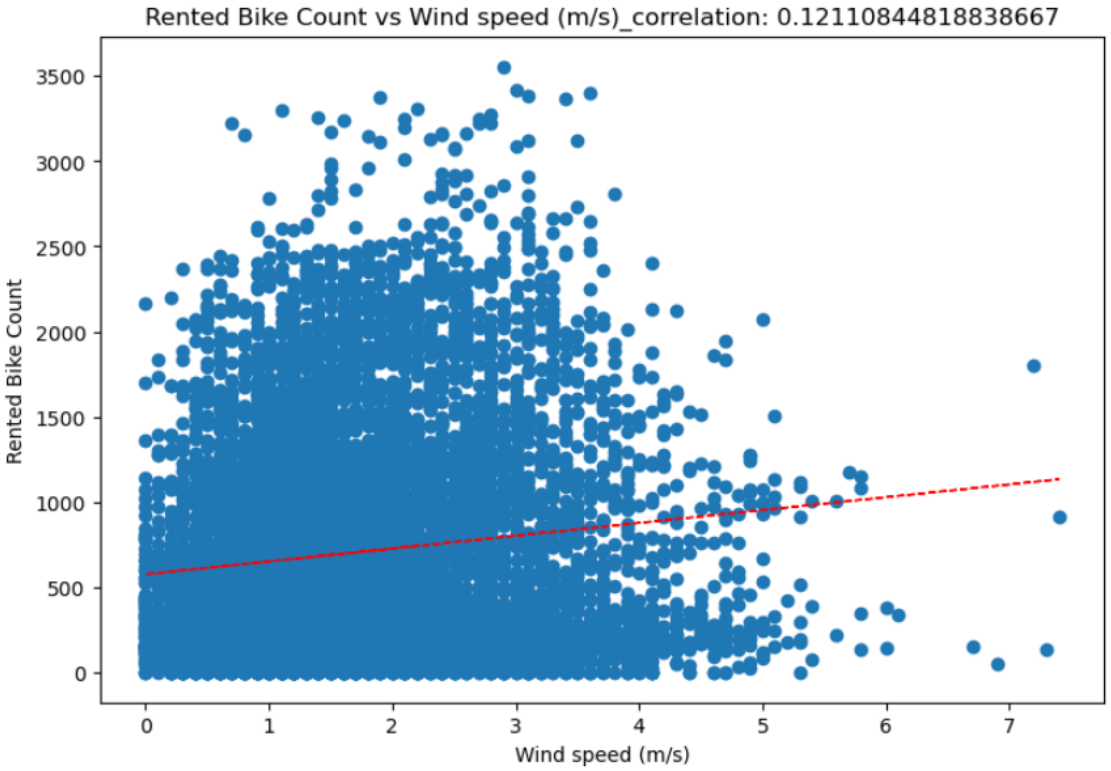


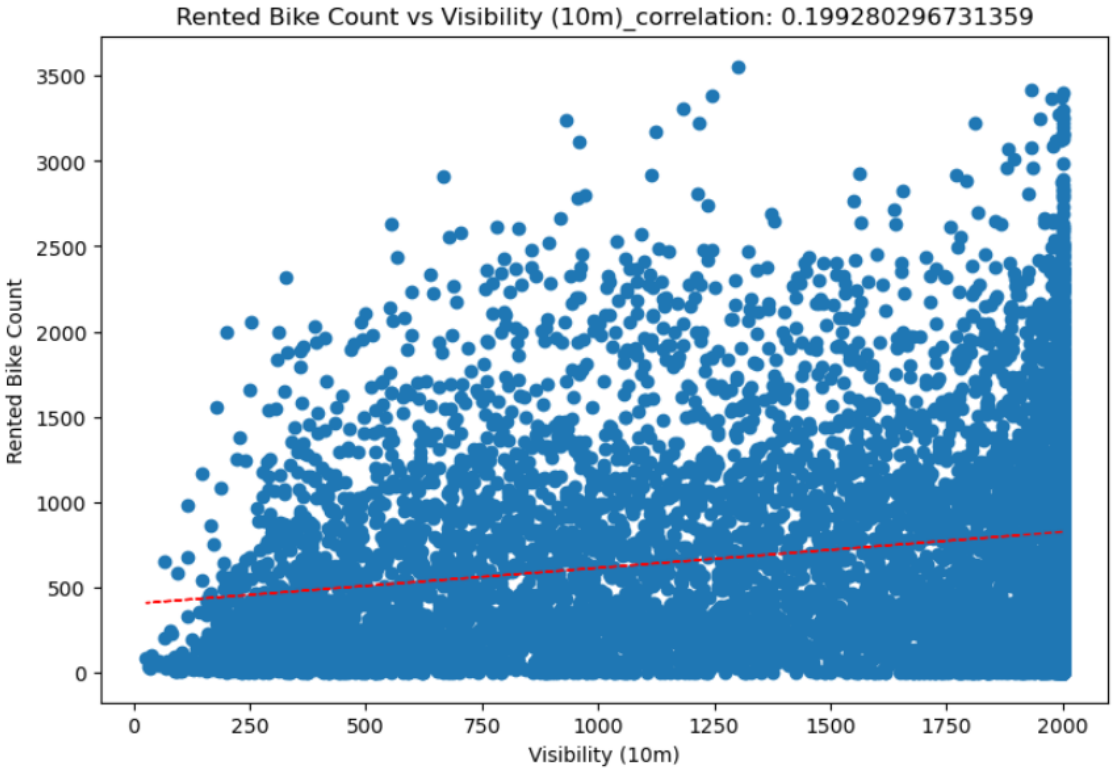


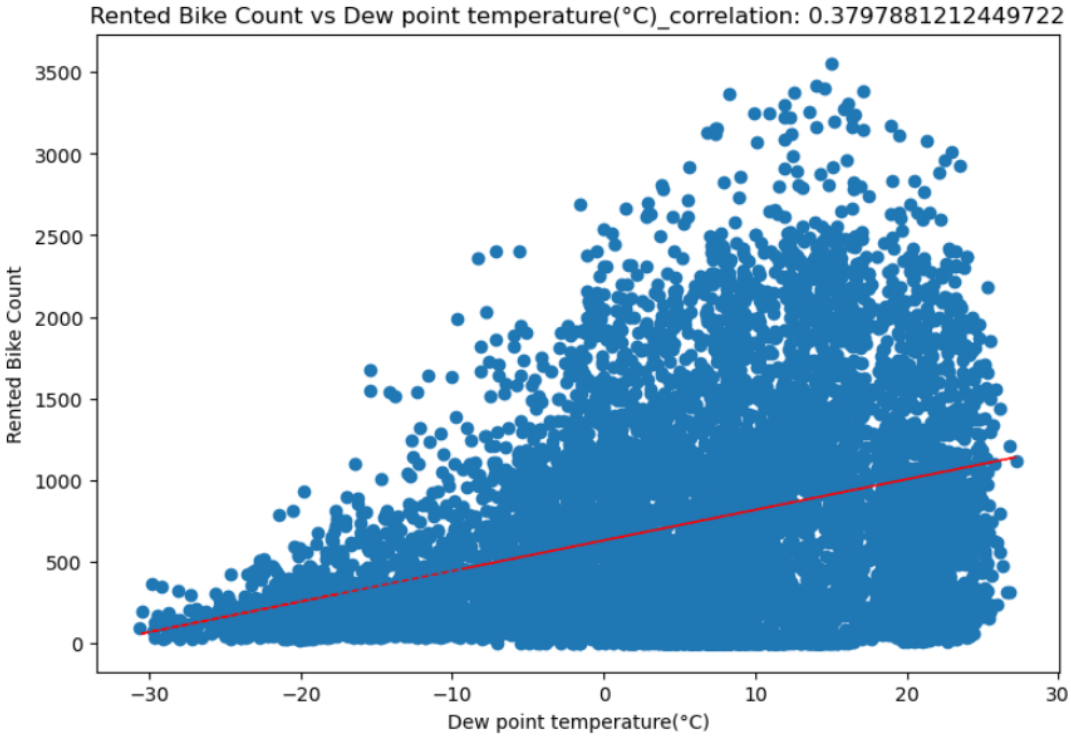
Visualizing how Numerical features correlated with Bike rented count

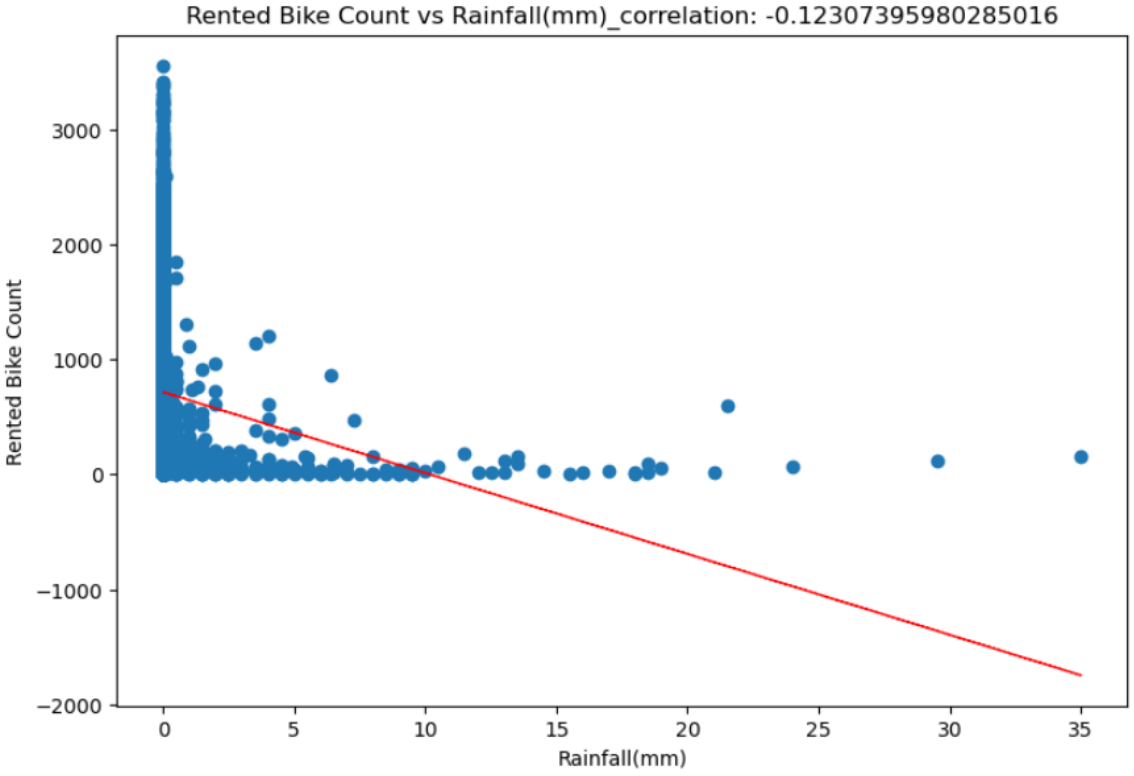

Rented Bike Count vs Temperature(°C)_correlation: 0.538558153013979
Rented Bike Count
3500
3000
2500
2000
1500
1000
500
0
−20
−10
0
10
20
30
40


Rented Bike Count vs Humidity(%)_correlation: -0.19978016700089832
Rented Bike Count
3500
3000
2500
2000
1500
1000
500
0
0
20
40
60
80
100


Rented Bike Count vs Wind speed (m/s)_correlation: 0.12110844818838667
Rented Bike Count
Wind speed (m/s)
0
500
1000
1500
2000
2500
3000
3500
0
1
2
3
4
5
6
7


Rented Bike Count vs Visibility (10m)_correlation: 0.199280296731359
Rented Bike Count
Visibility (10m)
0
500
1000
1500
2000
2500
3000
3500
0
250
500
750
1000
1250
1500
1750
2000


Rented Bike Count vs Dew point temperature(°C)_correlation: 0.3797881212449722
Rented Bike Count
Dew point temperature(°C)
3500
3000
2500
2000
1500
1000
500
0
−30
−20
−10
0
10
20
30


Rented Bike Count vs Rainfall(mm)_correlation: -0.12307395980285016
Rented Bike Count
Rainfall(mm)
3000
2000
1000
0
−1000
−2000
0
5
10
15
20
25
30
35

Model Implementation
Mode 1 : Linear Regression

Linear regression analysis is a statistical method used to make predictions about the value of a variable, based on the value of another variable. The variable that is being predicted is known as the dependent variable, while the variable used to make the prediction is referred to as the independent variable.

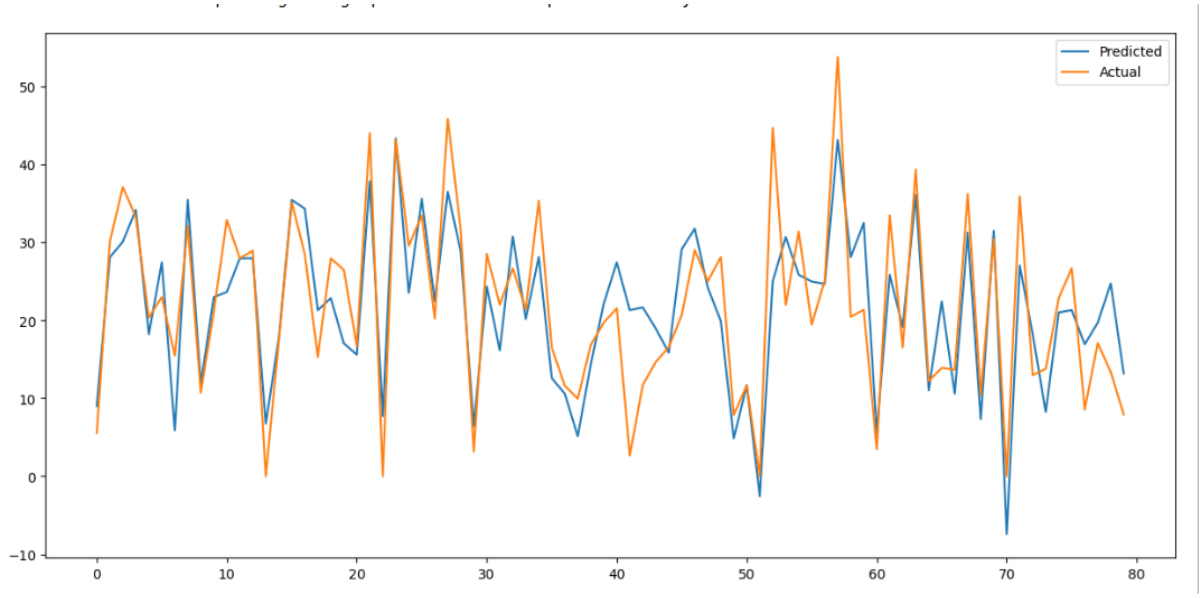
Evaluation Matrices

```
Training score  = 0.7031168479271276
MAE : 5.256832985963906
MSE : 44.35376150715582
RMSE : 6.659861973581421
R2 : 0.7161594175703074
Adjusted R2 :  0.7115468021854953

**************************************************************************************
coefficient
 [ 5.14940824e-01 -1.60847945e-01  1.58427497e-01  4.73815016e-04
 -4.10264161e-01 -1.53176051e+00  1.92997496e-01  7.31743738e+00
  3.49826484e+00 -3.20286654e+00 -2.99804972e+00 -3.65497680e+00
 -7.77576392e+00  2.71595480e+00  2.80909776e+01 -2.32980211e+00
 -8.22718544e-01  1.64703992e-01 -4.94031472e-01  4.38354159e-01
 -1.80821191e-01  4.28441282e+00  3.96743044e-01  1.72452283e+00
  6.05512430e-01  2.08197536e+00 -4.89976098e-02 -6.51584135e-01]


Intercept  =  -0.1911980167566938
```

Graph of Actual vs predicted



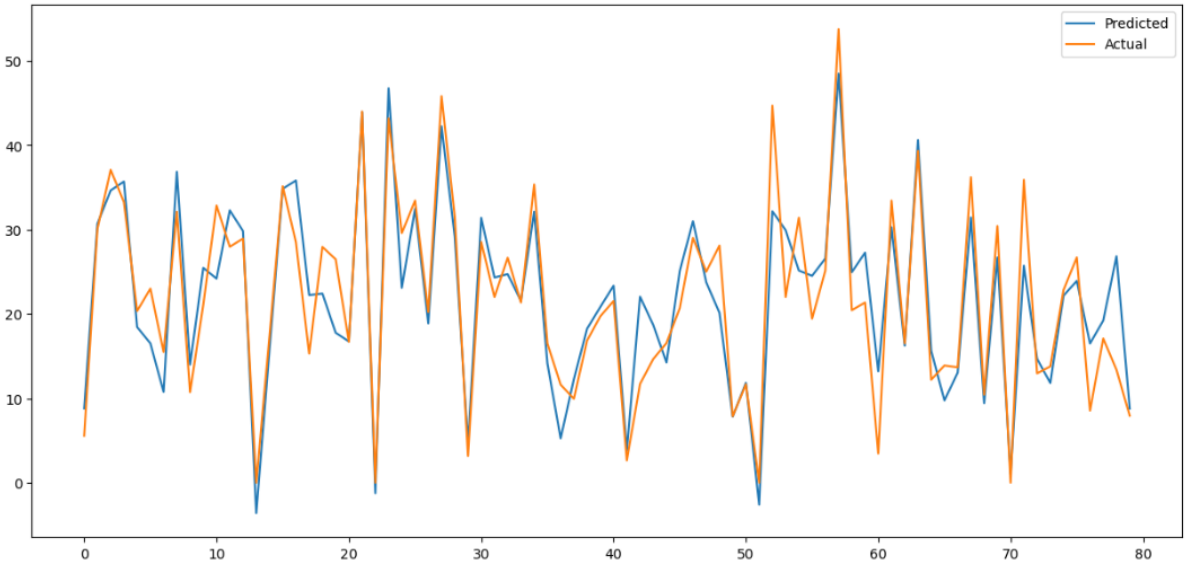
Mode 2 : Linear Regression model by using polynomial features

Evaluation Matrices

```
Training score  = 0.8514176744024103
MAE : 3.7032465524929217
MSE : 24.626439154483258
RMSE : 4.962503315312067
R2 : 0.8424038323863426
Adjusted R2 :  0.7903108742465699

**********************************************************************************
coefficient
```

Graph of Actual vs predicted



Mode 3 : Lasso Regression

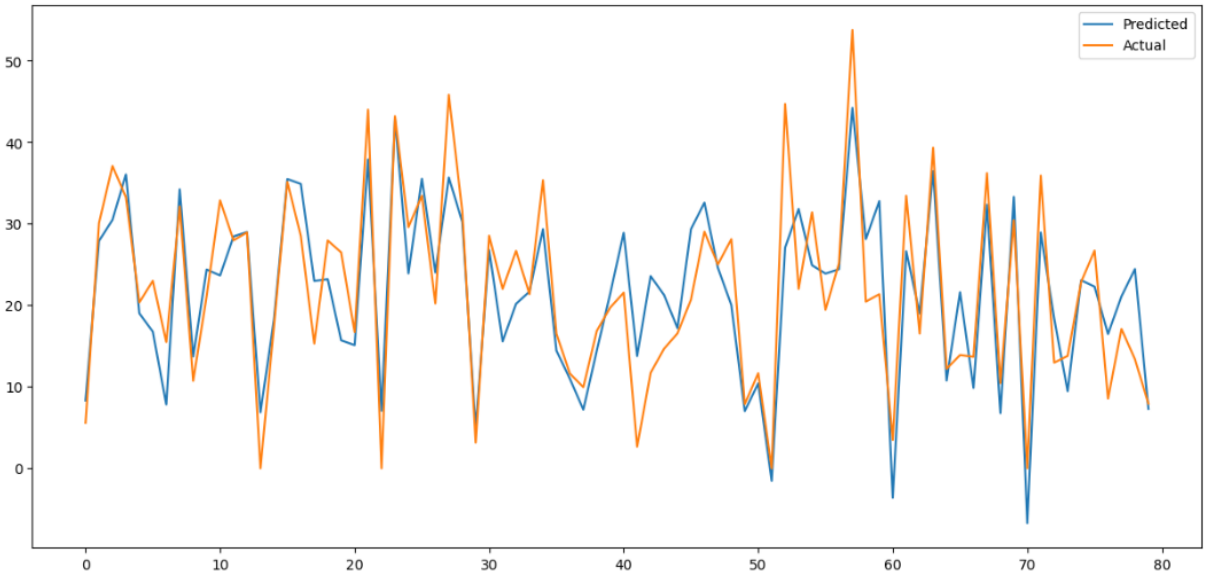
Evaluation Matrices

```
Training score  = 0.7358435727410344
The best parameters found out to be :{'alpha': 0.01}
where model best score is:  0.7329299487993713

MAE : 5.061698499519291
MSE : 41.512360555936745
RMSE : 6.443008657136566
R2 : 0.734342878758378
Adjusted R2 :  0.7300257578095879

**********************************************************************
```

Graph of Actual vs predicted



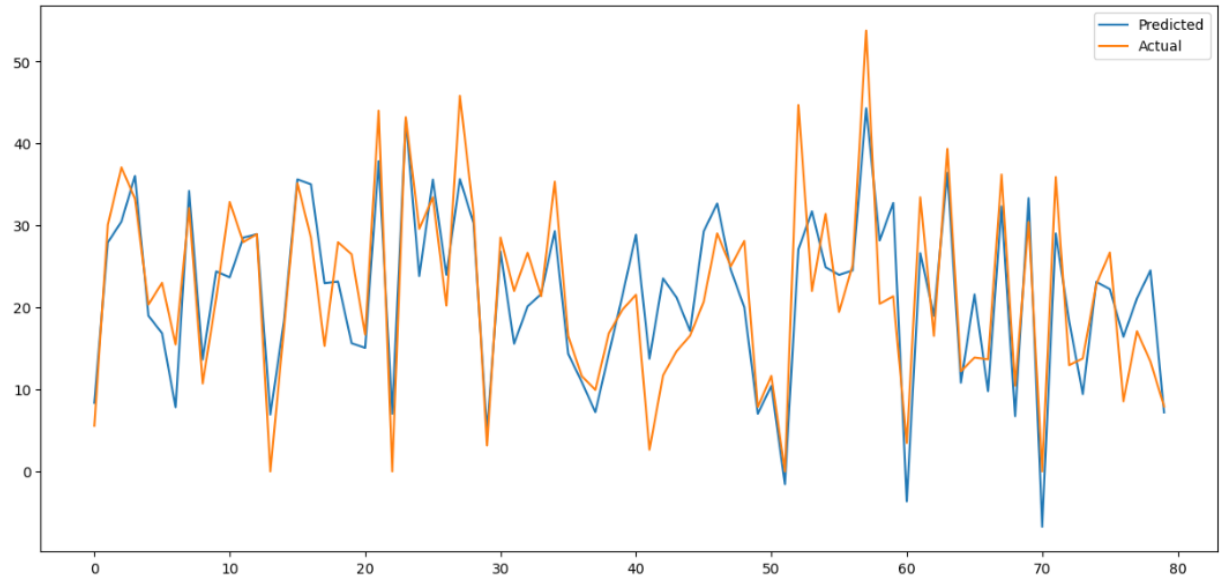
Mode 4 : Ridge Regression

Evaluation Matrices

```
Training score  = 0.7358671502026775
The best parameters found out to be :{'alpha': 10}
where model best score is:  0.732882411744482

MAE : 5.062807136059968
MSE : 41.52017837206753
RMSE : 6.4436153184425535
R2 : 0.7342928488757146
Adjusted R2 :  0.7299749149050355
```

Graph of Actual vs predicted



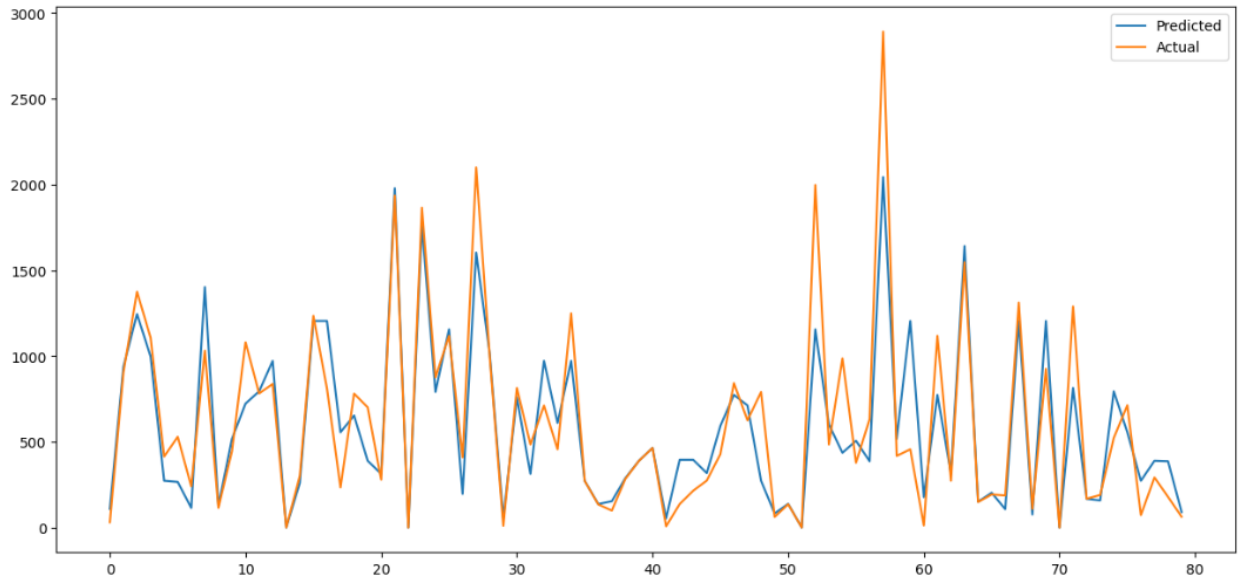
Mode 5 : Decision Tree

Evaluation Matrices

```
Training score  = 0.8246305633079585
The best parameters found out to be :{'criterion': 'mse', 'max_depth': 10, 'max_features': 24, 'min_samples_spli
t': 50, 'splitter': 'best'}
where model best score is:  0.7594795895227828

MAE : 194.7278665632221
MSE : 85363.18966798844
RMSE : 292.16979595431906
R2 : 0.7897762389551226
Adjusted R2 :  0.7862358852557606
```

Graph of Actual vs predicted



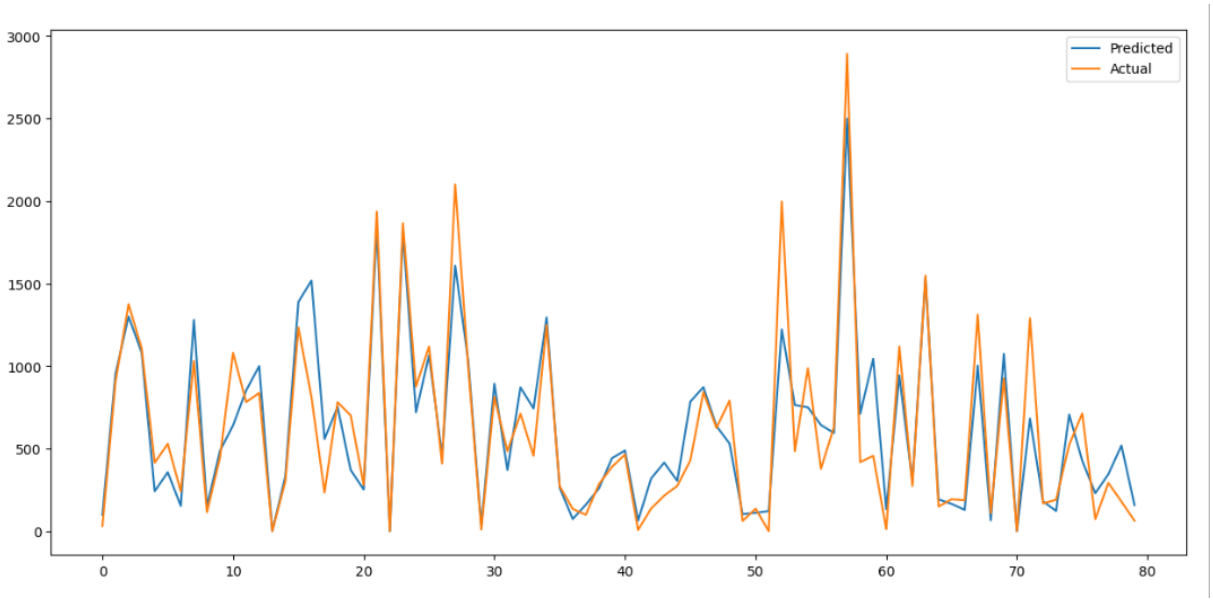
Mode 6 : Random Forest

Evaluation Matrices

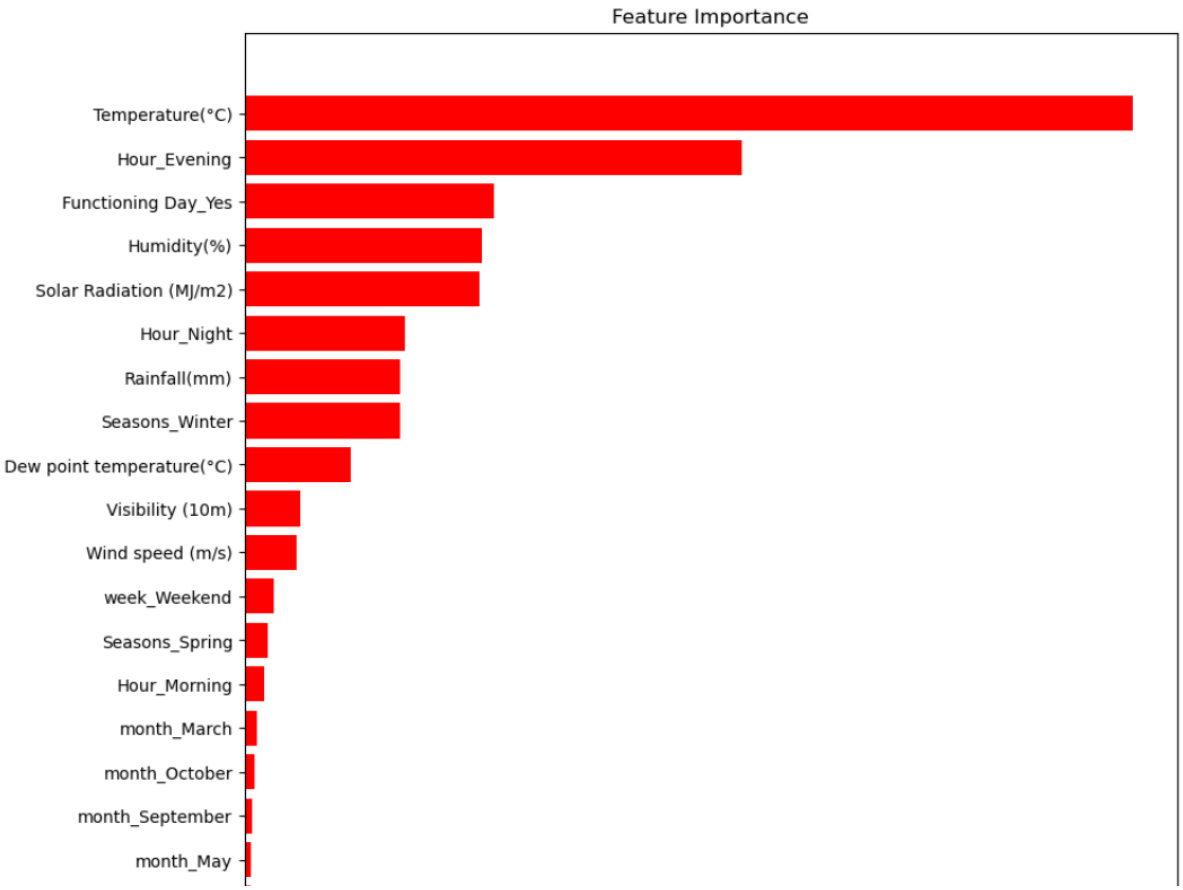
```
Training score  = 0.9462605232175477
The best parameters found out to be :{'max_depth': 25, 'max_features': 24, 'min_samples_split': 10, 'n_estimator
s': 150}
where model best score is:  0.8366238371741636

MAE : 163.50523456322918
MSE : 64531.58907766364
RMSE : 254.03068530723536
R2 : 0.8410781811823942
Adjusted R2 :  0.8384017974740837
```

Graph of Actual vs predicted



Feature Importance



Conclusion

From analyzing the model insights, we can conclude that the Random Forest Regression model performs the best in predicting bike rental counts with an R2 score of 0.842402, while the Linear Regression model performs the worst with an R2 score of 0.7161594175703074. Visualizations of actual versus predicted values have been created for all six models, and feature importance graphs have been used to explain each model. Furthermore, temperature and hour are the two most significant factors in predicting bike rental counts according to all six models, making them useful features for modeling purposes.