Capstone Project – II
Supervised ML Regression

Seoul Bike Sharing Demand Prediction

Contributor ***:- Suraj Galande***

Contents

- ❖ Model (1) – Linear Regression
  - ✓ Evaluation Matrices
  - ✓ Graph of Actual v/s Predicted values of bike rent count prediction
- ❖ Model (2) – Linear Regression using polynomial
  - ✓ Evaluation matrices
  - ✓ Graph of Actual v/s Predicted values of bike rent count prediction
- ❖ Regularization Techniques –
- ❖ Model (3) – Lasso Regression
  - ✓ 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 Importance's for predicting bike rent count
- ❖ Conclusion of Project
  - ✓ From EDA
  - ✓ From Model Building

❖ Problem Statements

- 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.
- Bike sharing has been gaining importance over the last few decades. More and more people are turning to healthier and more livable cities where activities like bike sharing are easily available.
- There are many benefits from bike sharing, such as environmental benefits, Stopping Co2 emission, It was a green way to travel.

❖ *Data Summary*

- The dataset contains weather information's such as

(Temperature, Humidity, Wind speed, Visibility, Dew point, Solar radiation, Snowfall, Rainfall)

- And the number of bikes rented per hour and date information.
- This Seoul Bike sharing dataset contains the hourly and daily count of rental bikes between years 2017 and 2018 in Capital bike share system with the corresponding weather and seasonal information.
- The dataset contains 8760 rows (every hour of each day for 2017 and 2018) and 14 columns (the features which are under consideration).

❖ Dataset Description

Features	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 10meter
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)

❖ EDA

➢ Preprocessing The Data :-

✓ Looking for first 5 rows of Seoul bike sharing data frame

	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

✓ Null values in data frame are :-

```
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
```

✓So we see that no any null or missing values present in dataset.

➢ Checking duplicated values :-

- ✓ Since by observing and performing duplication method on data frame we can conclude that there are no duplicate values present.
- ✓ So there is no need of doing any operation to remove duplicate values to get good analysis of data frame.

➢ Separating Numerical and categorical features

- ✓ Numerical features = [Rented Bike Count, Hour , Temperatures, Humidity, Wind Speed, Visibility, Dew point Temperatures, solar radiation, rainfall, and snowfall.]

- ✓ Categorical features = [Season, Holliday, Functioning day]

➢Description of Numerical and Categorical features in dataset.

	count	unique	top	freq	first	last	mean	std	min	25%	50%	75%	max
Date	8760	365	2017-01-12 00:00:00	24	2017-01-12	2018-12-11	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Rented Bike Count	8760.0	NaN	NaN	NaN	NaT	NaT	704.602055	644.997468	0.0	191.0	504.5	1065.25	3556.0
Hour	8760.0	NaN	NaN	NaN	NaT	NaT	11.5	6.922582	0.0	5.75	11.5	17.25	23.0
Temperature(°C)	8760.0	NaN	NaN	NaN	NaT	NaT	12.882922	11.944825	-17.8	3.5	13.7	22.5	39.4
Humidity(%)	8760.0	NaN	NaN	NaN	NaT	NaT	58.226256	20.362413	0.0	42.0	57.0	74.0	98.0
Wind speed (m/s)	8760.0	NaN	NaN	NaN	NaT	NaT	1.724909	1.0363	0.0	0.9	1.5	2.3	7.4
Visibility (10m)	8760.0	NaN	NaN	NaN	NaT	NaT	1436.825799	608.298712	27.0	940.0	1698.0	2000.0	2000.0
Dew point temperature(°C)	8760.0	NaN	NaN	NaN	NaT	NaT	4.073813	13.060369	-30.6	-4.7	5.1	14.8	27.2
Solar Radiation (MJ/m2)	8760.0	NaN	NaN	NaN	NaT	NaT	0.569111	0.868746	0.0	0.0	0.01	0.93	3.52
Rainfall(mm)	8760.0	NaN	NaN	NaN	NaT	NaT	0.148687	1.128193	0.0	0.0	0.0	0.0	35.0
Snowfall (cm)	8760.0	NaN	NaN	NaN	NaT	NaT	0.075068	0.436746	0.0	0.0	0.0	0.0	8.8
Seasons	8760	4	Spring	2208	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Holiday	8760	2	No Holiday	8328	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Functioning Day	8760	2	Yes	8465	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN	NaN

✓**This displays summary of statistics of data frame.**

✓**Information of different descriptive statistics** :-

✓1) **Measures of Frequency** :- Count, Percent, Frequency.

✓2) **Measures of Central Tendency** :- Mean, Median, and Mode.

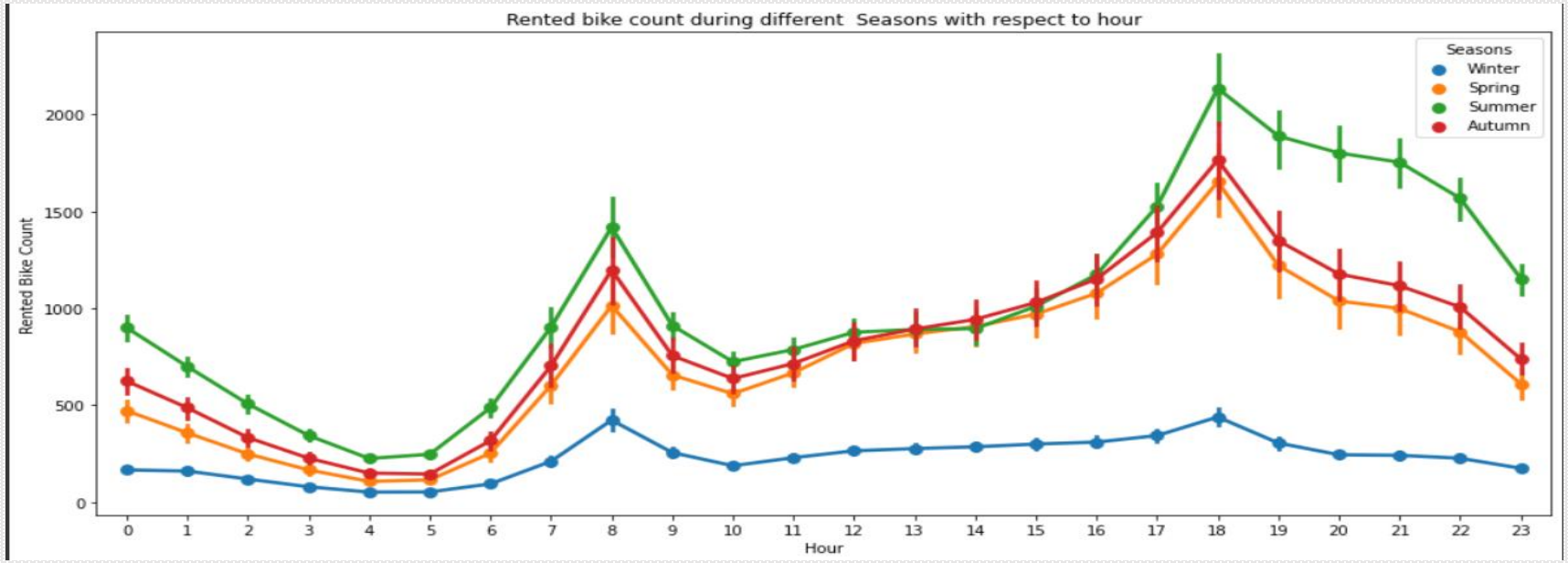
✓3)**Measures of Dispersion or Variation or spread** :- Range(max - min),Variance, Standard Deviation.

✓4) **Measures of Position** :- Percentile Ranks, Quartile Ranks.

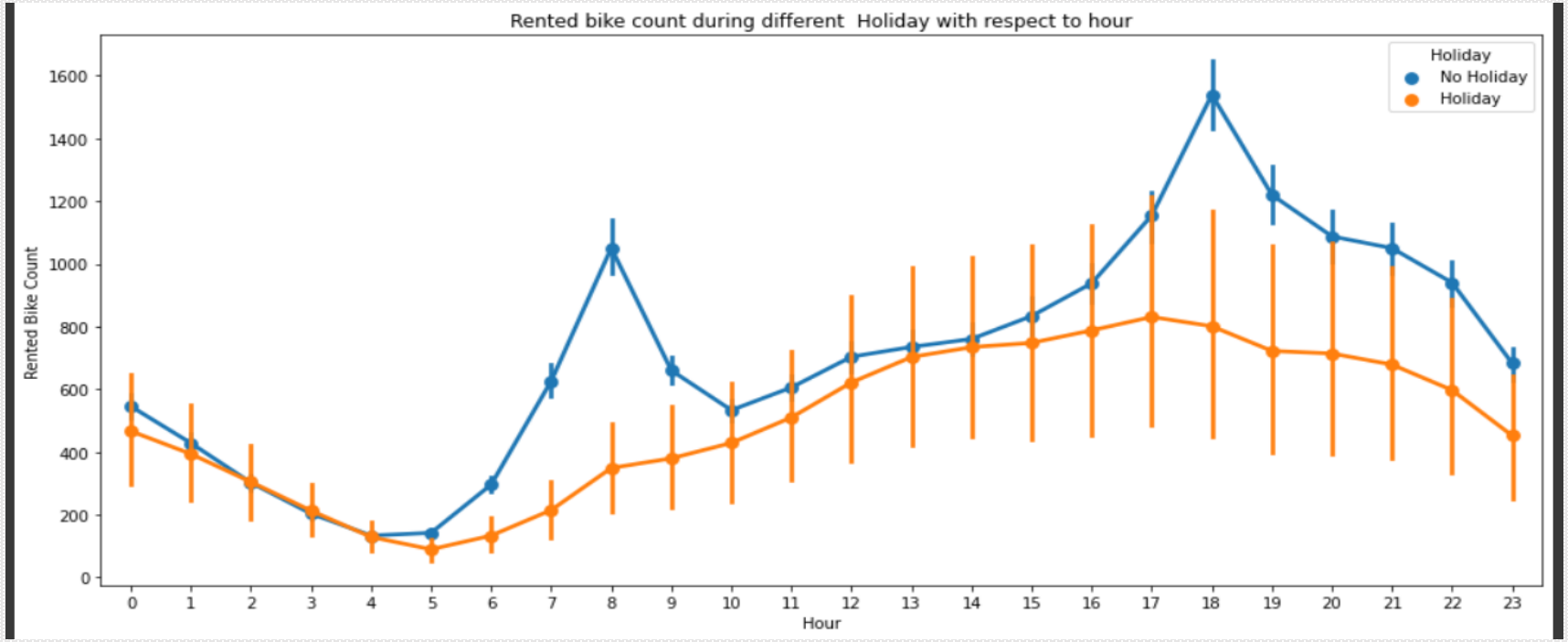
➢Visualizing Rented Bike Count, Hour with Respect to different categorical Feature :-

✓Here Categorical features are after performing feature engineering on data frame are - ['Hour', 'Seasons', 'Holiday', 'Functioning Day', 'year', 'month', 'day‘]

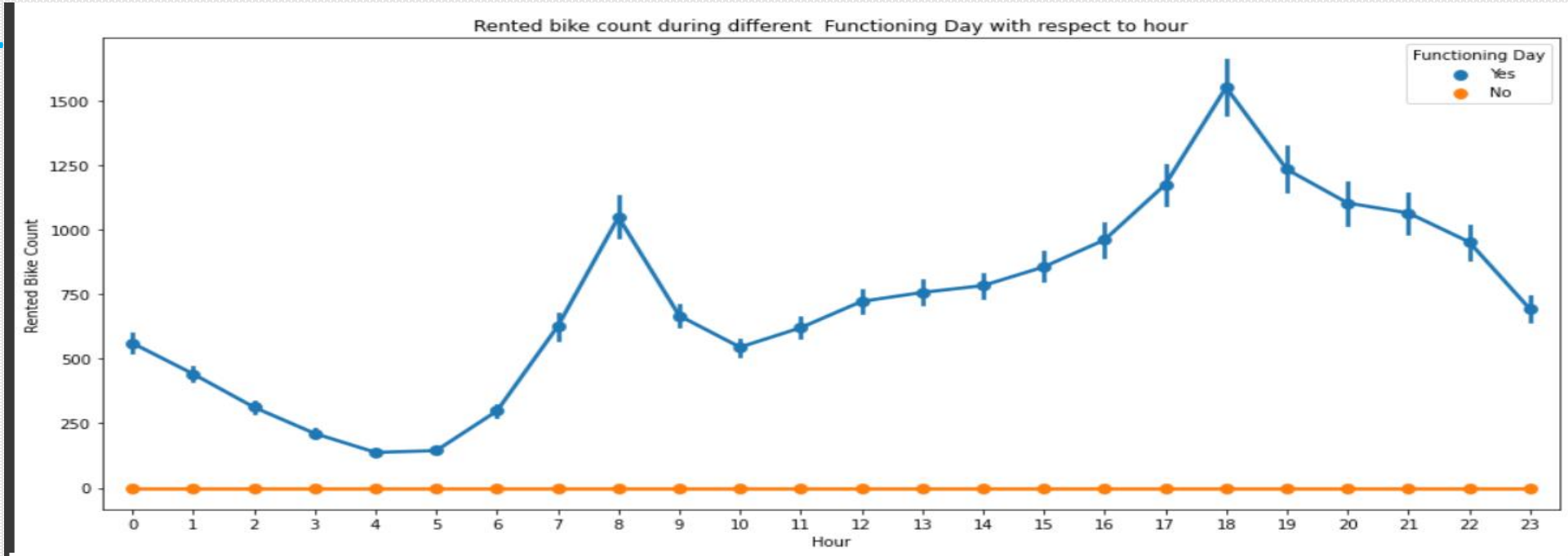
✓Season :-



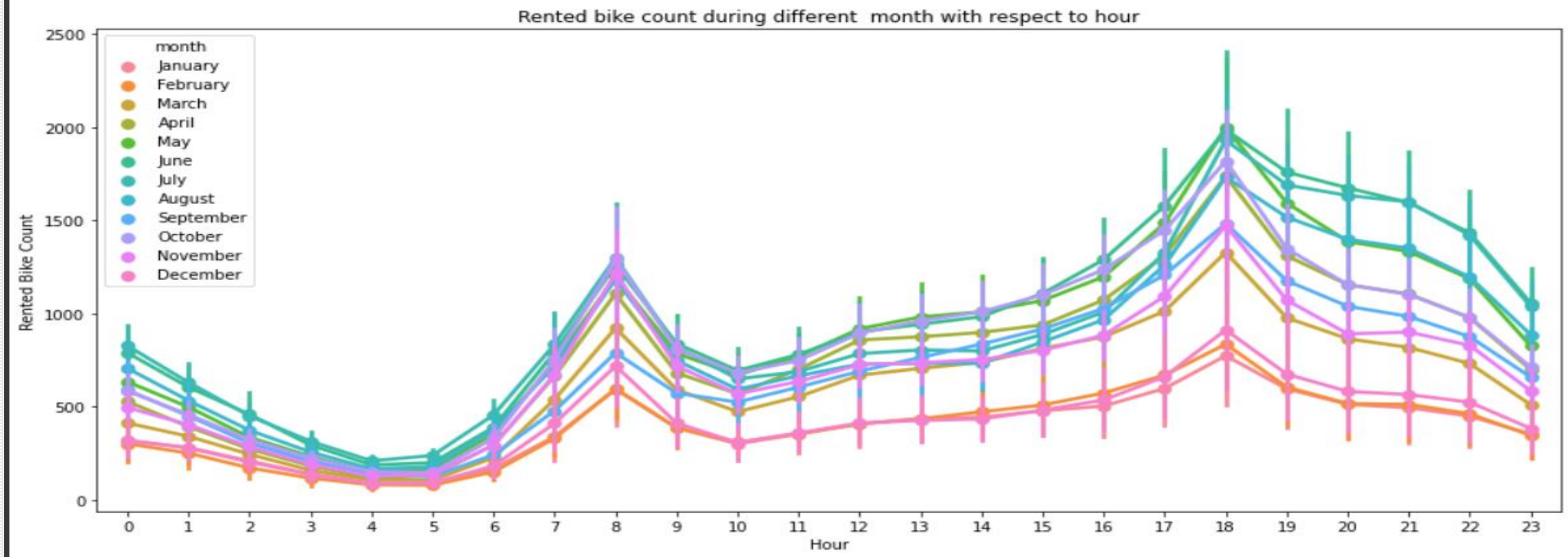
✓Holiday :-



✓Functioning day :-

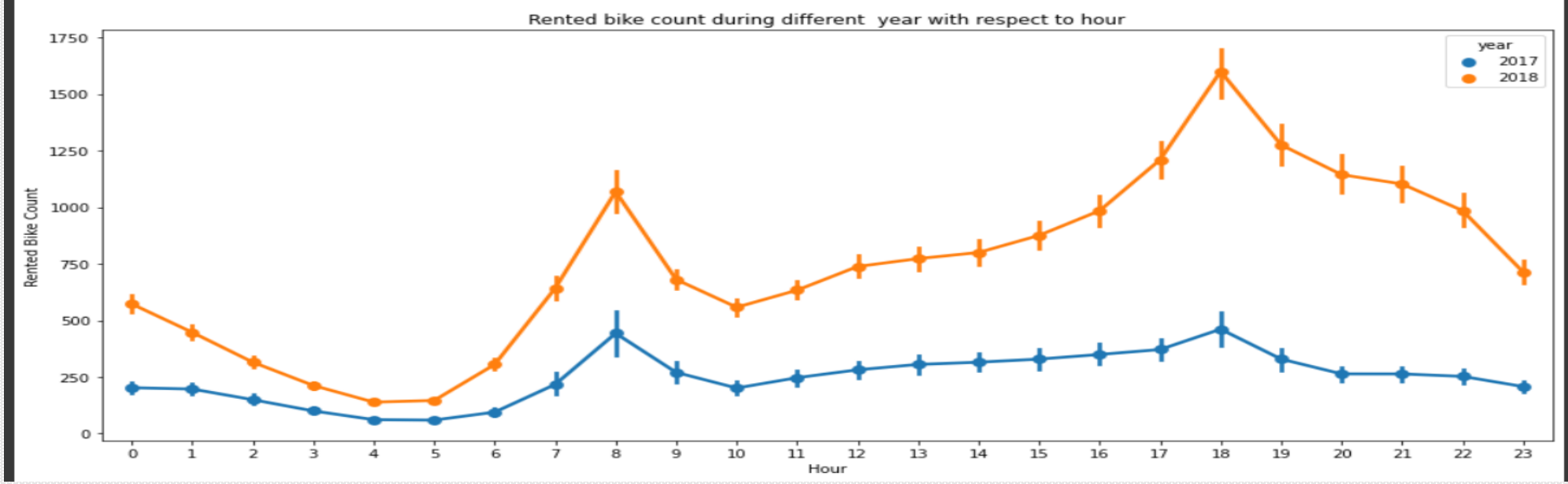


✓Month :-

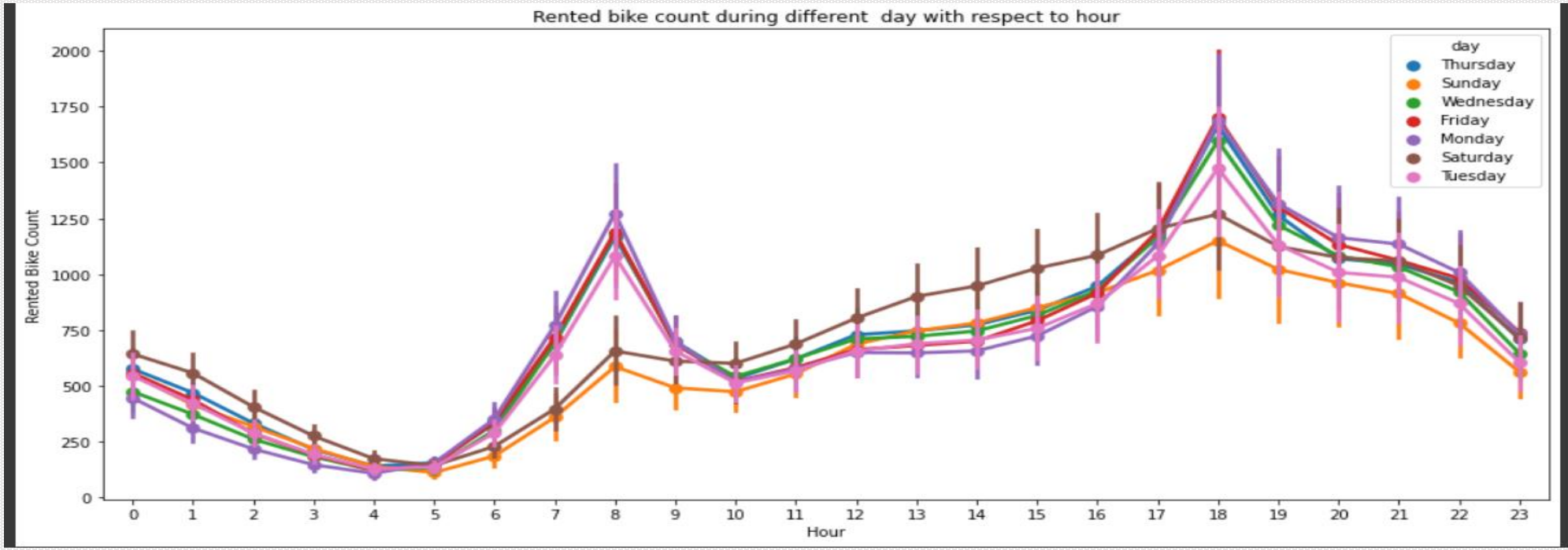


Rented bike count during different day with respect to hour

✓Year :-


Rented bike count during different year with respect to hour
Rented Bike Count
Hour
year
2017
2018

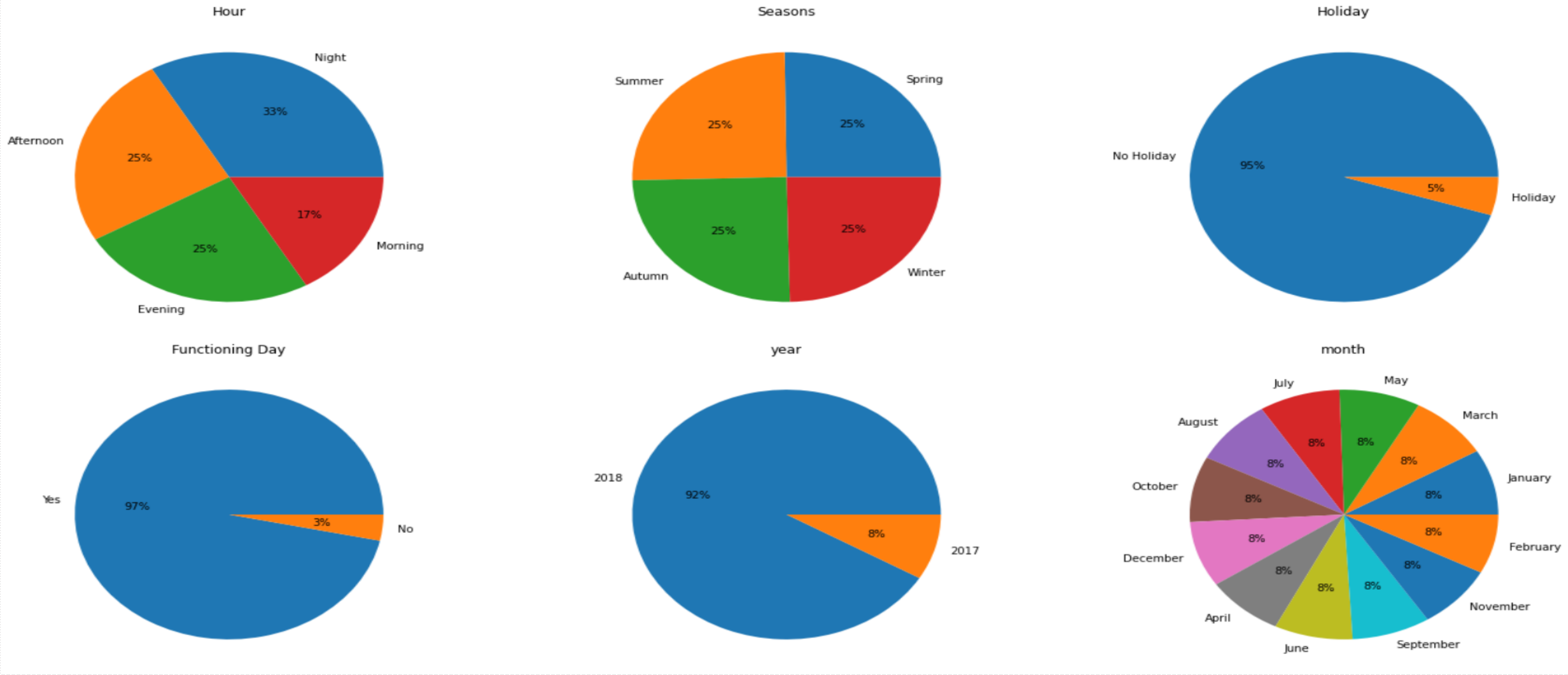
✓Day's :-


Rented bike count during different day with respect to hour
Rented Bike Count
Hour
day
Thursday
Sunday
Wednesday
Friday
Monday
Saturday
Tuesday

➢Conclusion from visualization -

- ✓Observation From all these point plot we have observed a lot from every column like :
- ✓ Season :- In the season column, we are able to understand that the demand is low in the winter season.
- ✓ Holiday :- In the Holiday column, The demand is low during holidays, but in non holidays the demand is high, it may be because people use bikes to go to their work.
- ✓ Functioning Day :- In the Functioning Day column, If there is no Functioning Day then there is no demand.
- ✓ month :- In the month column, We can clearly see that the demand is low in December January & February, It is cold in these months and we have already seen in season column that demand is less in winters.
- ✓year :- The demand was less in 2017 and higher in 2018, it may be because it was new in 2017 and people did not know much about it.
- ✓ day :- days of week. Further we divide them into weekdays and weekends

➢Visualizing Value count (in percentage) of Categorical Features :-



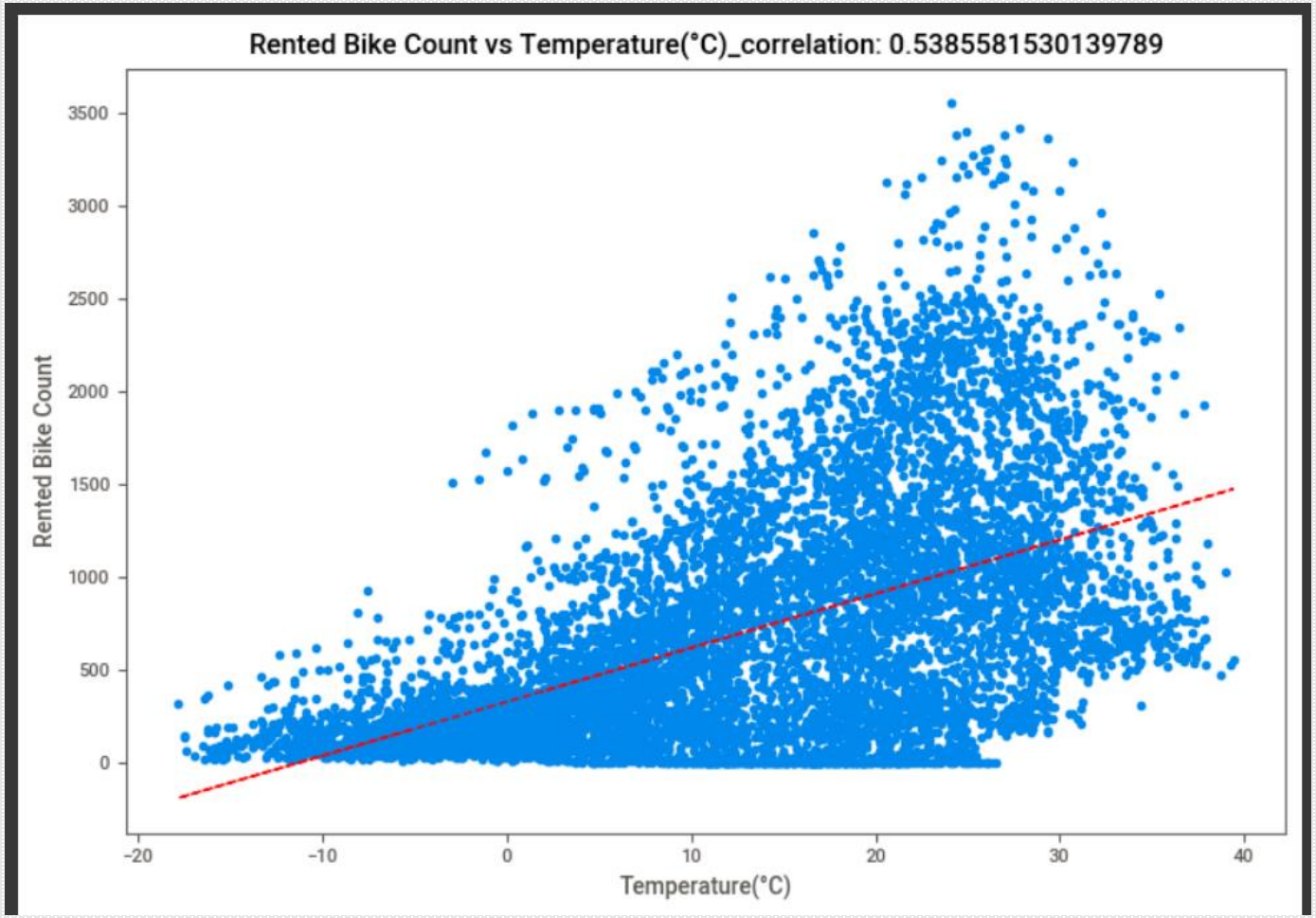
✓From above pie charts we can see contribution of each categorical features in dataset to predict the demand of bike sharing.

➢Visualizing how Numerical features correlated wrt Bike rented count :-

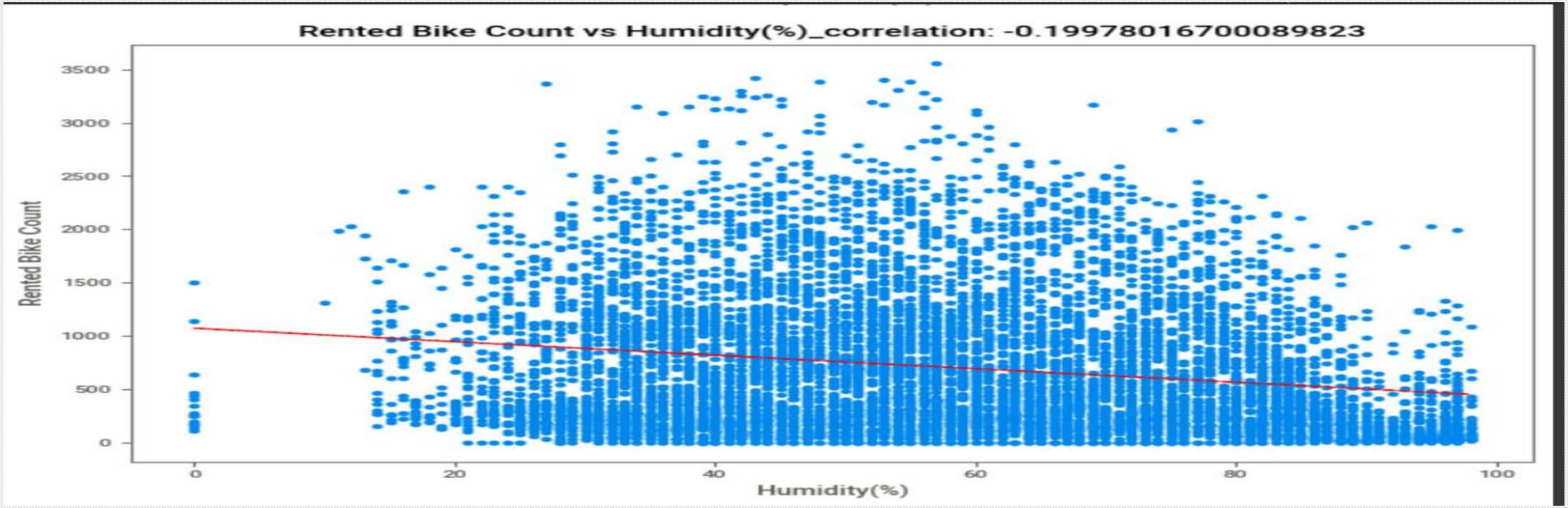
✓Here Numerical features after performing feature engineering are –

✓ ['Rented Bike Count', 'Temperature(°C)', 'Humidity(%)', 'Wind speed (m/s)', 'Visibility (10m)', 'Dew point temperature(°C)', 'Solar Radiation (MJ/m2)', 'Rainfall(mm)', 'Snowfall (cm)'].

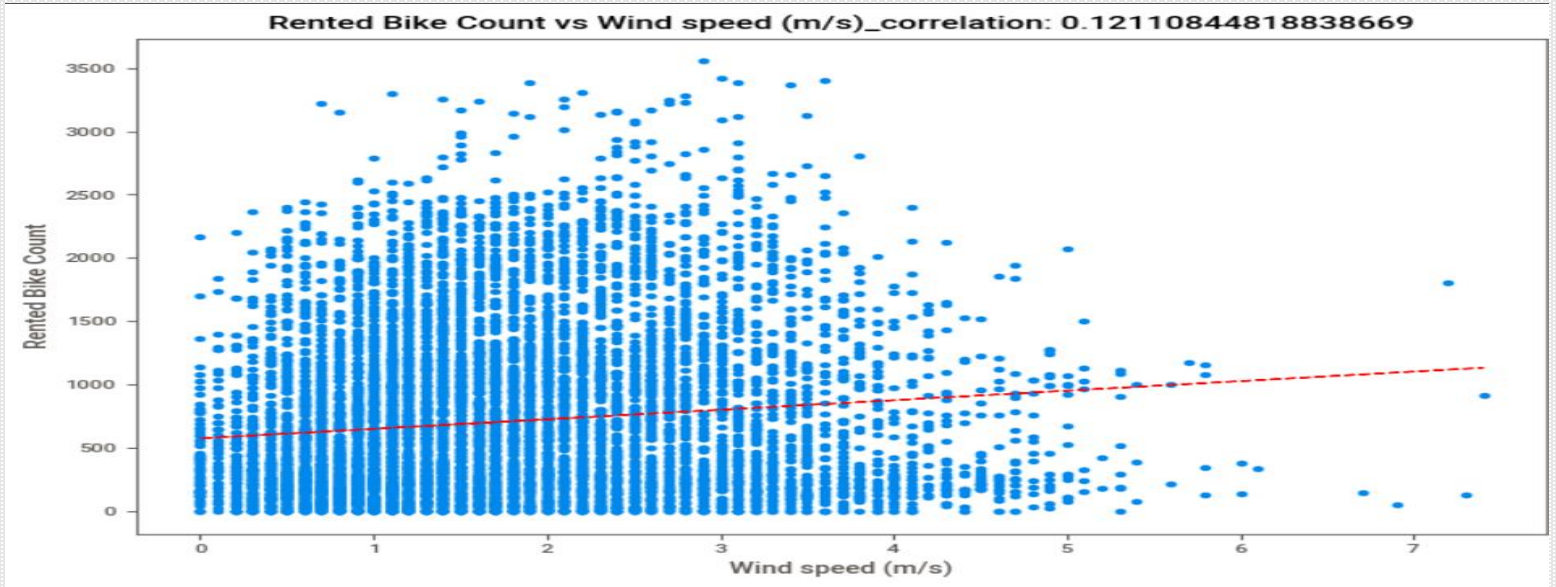
➢Temperature vs. Rented bike count



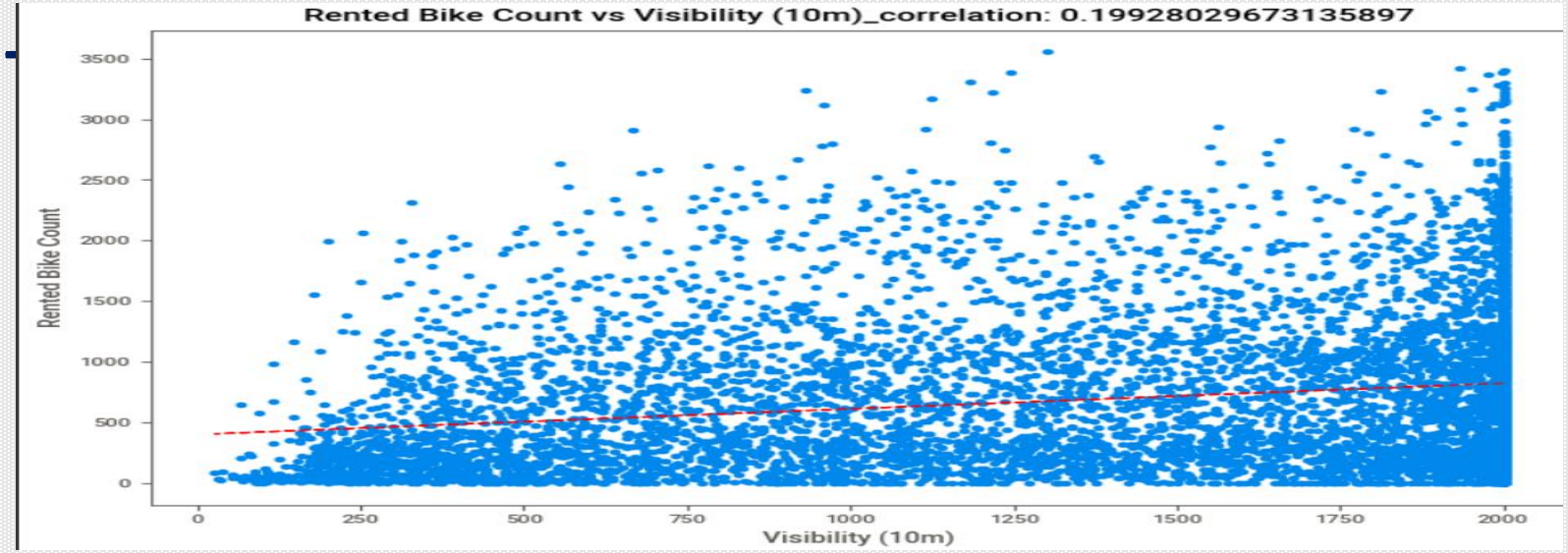
➢Humidity vs. Rented bike count :-



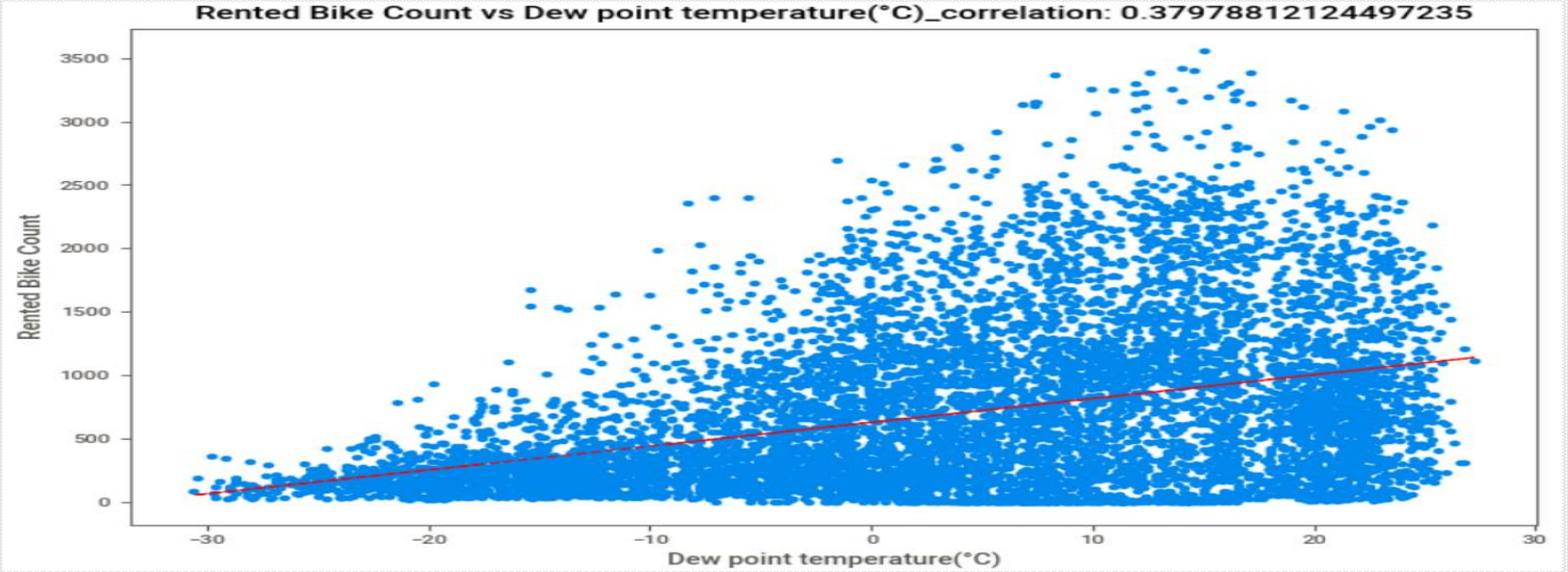
➢Wind speed vs. Rented bike



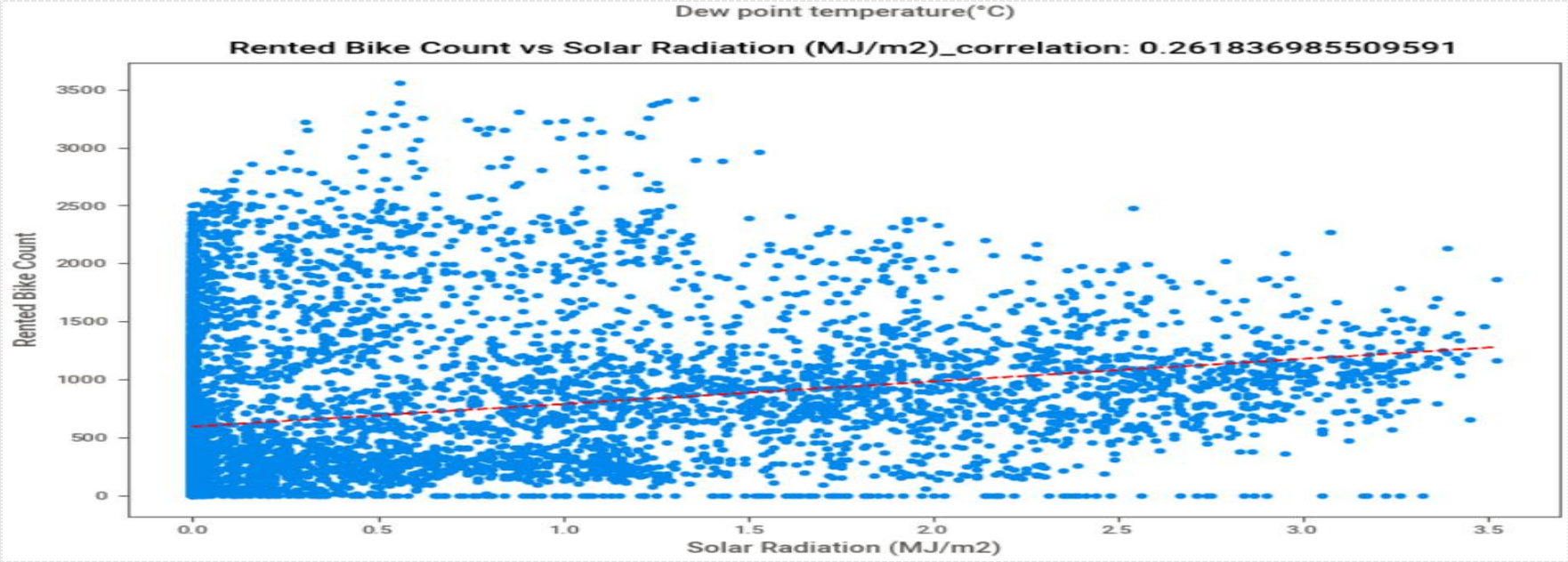
➢Visibility vs. Rented bike :-



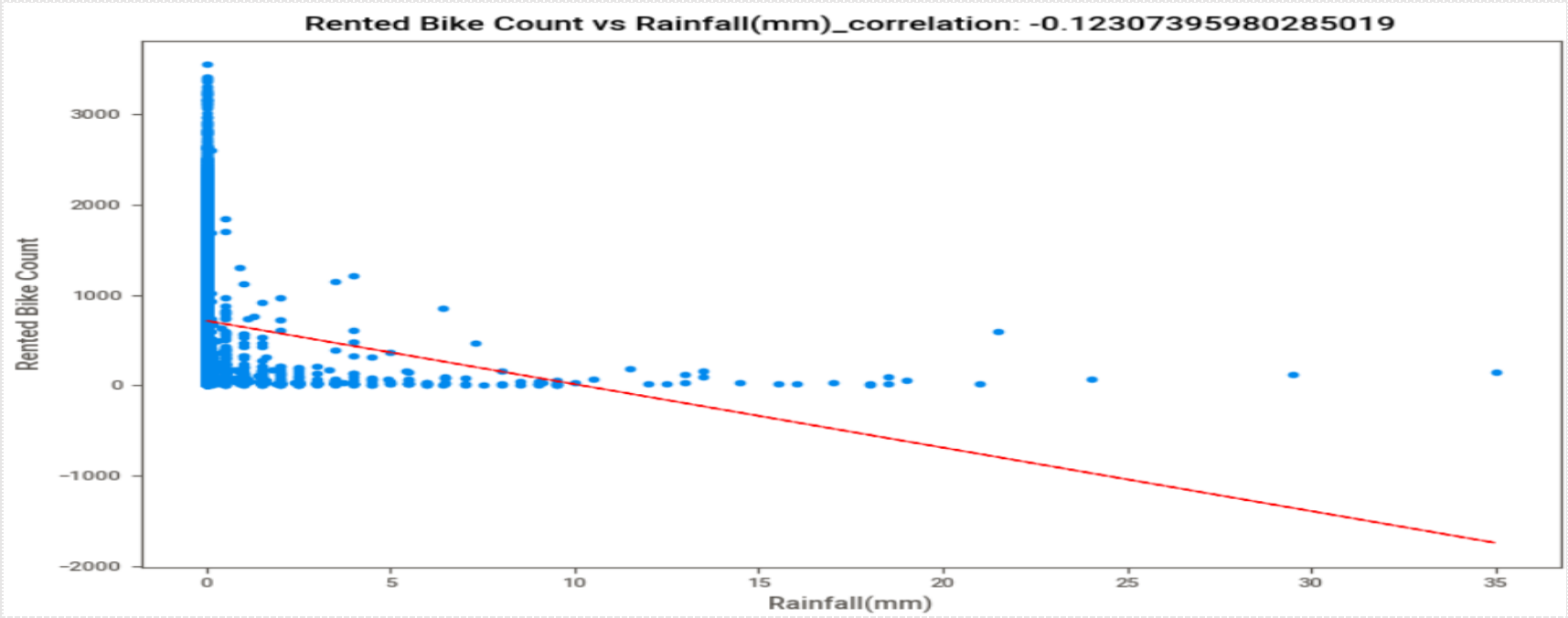
➢Dew point temp vs. Rented bike:-



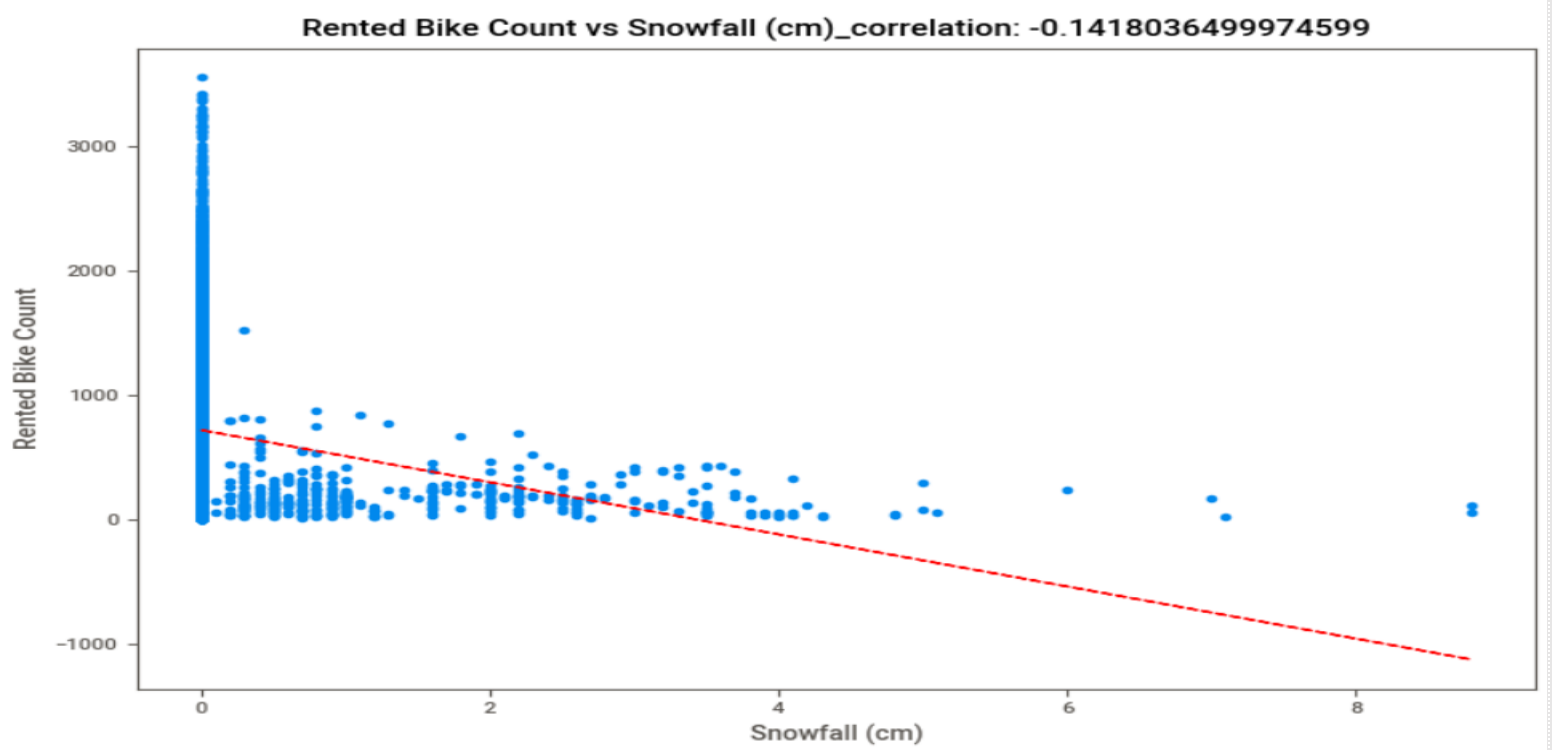
➢Solar radiation vs. Rented bike:-



➢ Rainfall vs. Rented bike



➢Snowfall vs. Rented bike:-



- ✓By all the above regression plots of numerical feature vs. Dependent feature we get clear idea about how these features related with bike rent count.
- ✓some numeric features has positive correlation with dependent variable and some has negative correlation.
- ✓**positive correlation features with Rented bike count**= Temperature, wind speed, Visibility, Dew point temperature, solar radiation.
- ✓**negative correlation features with Rented bike count** = Humidity, Rainfall, snowfall

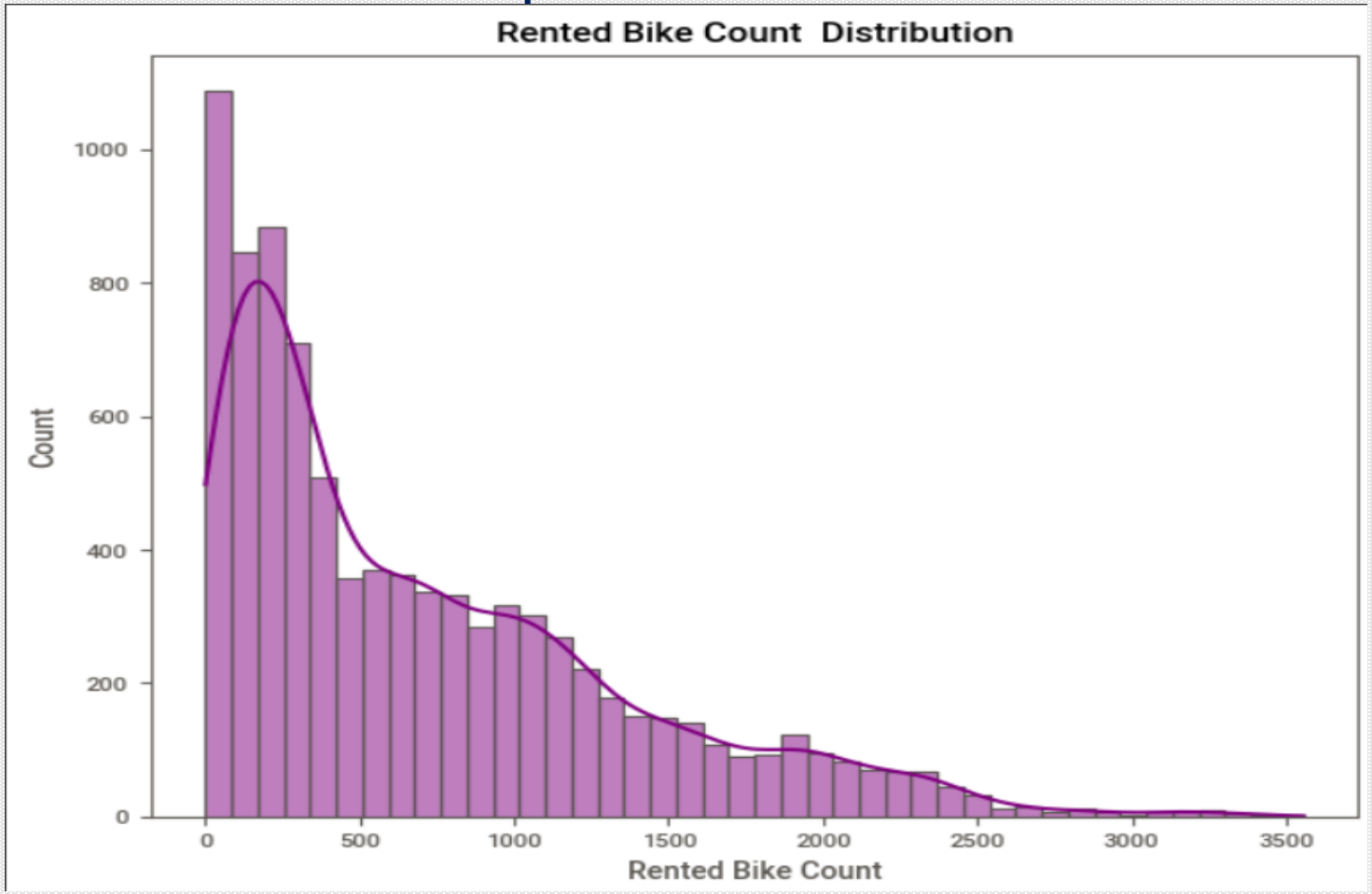
➢Correlation table for numerical features :-

	Rented Bike Count	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)
Rented Bike Count	1.00 %	0.54 %	-0.20 %	0.12 %	0.20 %	0.38 %	0.26 %	-0.12 %	-0.14 %
Temperature(°C)	0.54 %	1.00 %	0.16 %	-0.04 %	0.03 %	0.91 %	0.35 %	0.05 %	-0.22 %
Humidity(%)	-0.20 %	0.16 %	1.00 %	-0.34 %	-0.54 %	0.54 %	-0.46 %	0.24 %	0.11 %
Wind speed (m/s)	0.12 %	-0.04 %	-0.34 %	1.00 %	0.17 %	-0.18 %	0.33 %	-0.02 %	-0.00 %
Visibility (10m)	0.20 %	0.03 %	-0.54 %	0.17 %	1.00 %	-0.18 %	0.15 %	-0.17 %	-0.12 %
Dew point temperature(°C)	0.38 %	0.91 %	0.54 %	-0.18 %	-0.18 %	1.00 %	0.09 %	0.13 %	-0.15 %
Solar Radiation (MJ/m2)	0.26 %	0.35 %	-0.46 %	0.33 %	0.15 %	0.09 %	1.00 %	-0.07 %	-0.07 %
Rainfall(mm)	-0.12 %	0.05 %	0.24 %	-0.02 %	-0.17 %	0.13 %	-0.07 %	1.00 %	0.01 %
Snowfall (cm)	-0.14 %	-0.22 %	0.11 %	-0.00 %	-0.12 %	-0.15 %	-0.07 %	0.01 %	1.00 %

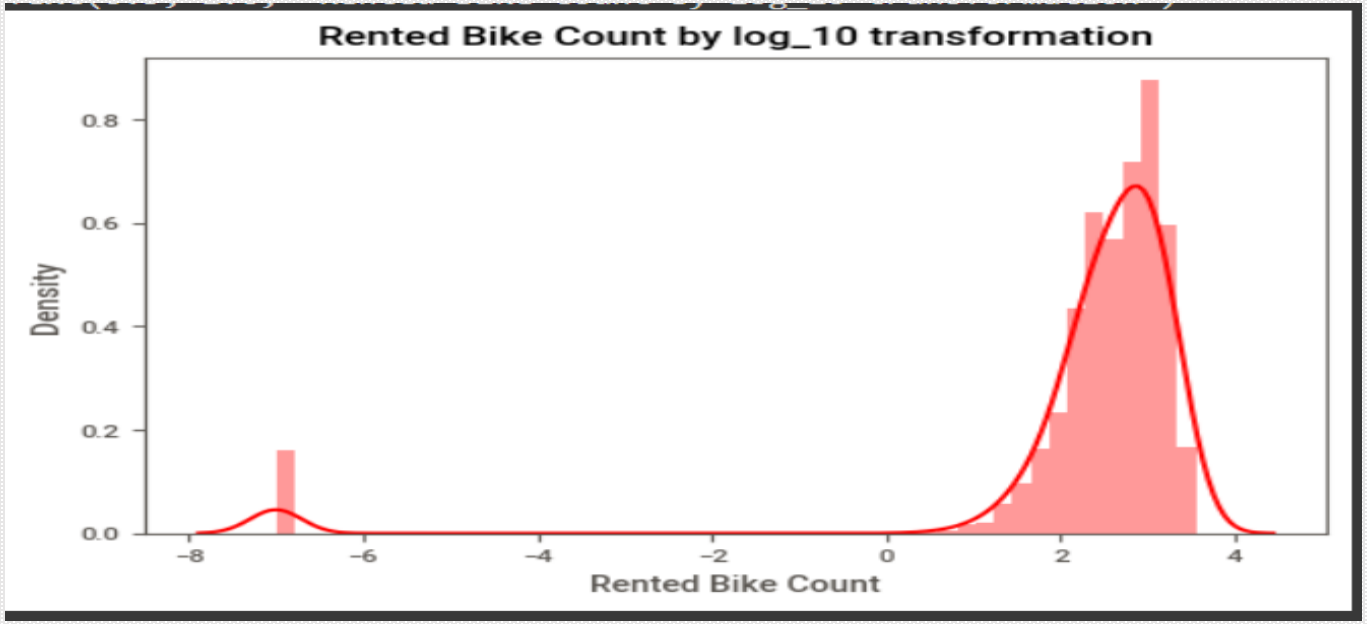
- ✓From this graph we are able to see that there is multicollinearity in temperature(°C) and dew point temperature(°C) column.
- ✓That is temp and dew point temp has 0.91% means there is lot of similarities between these features.

❖ *Model Implementation:-*

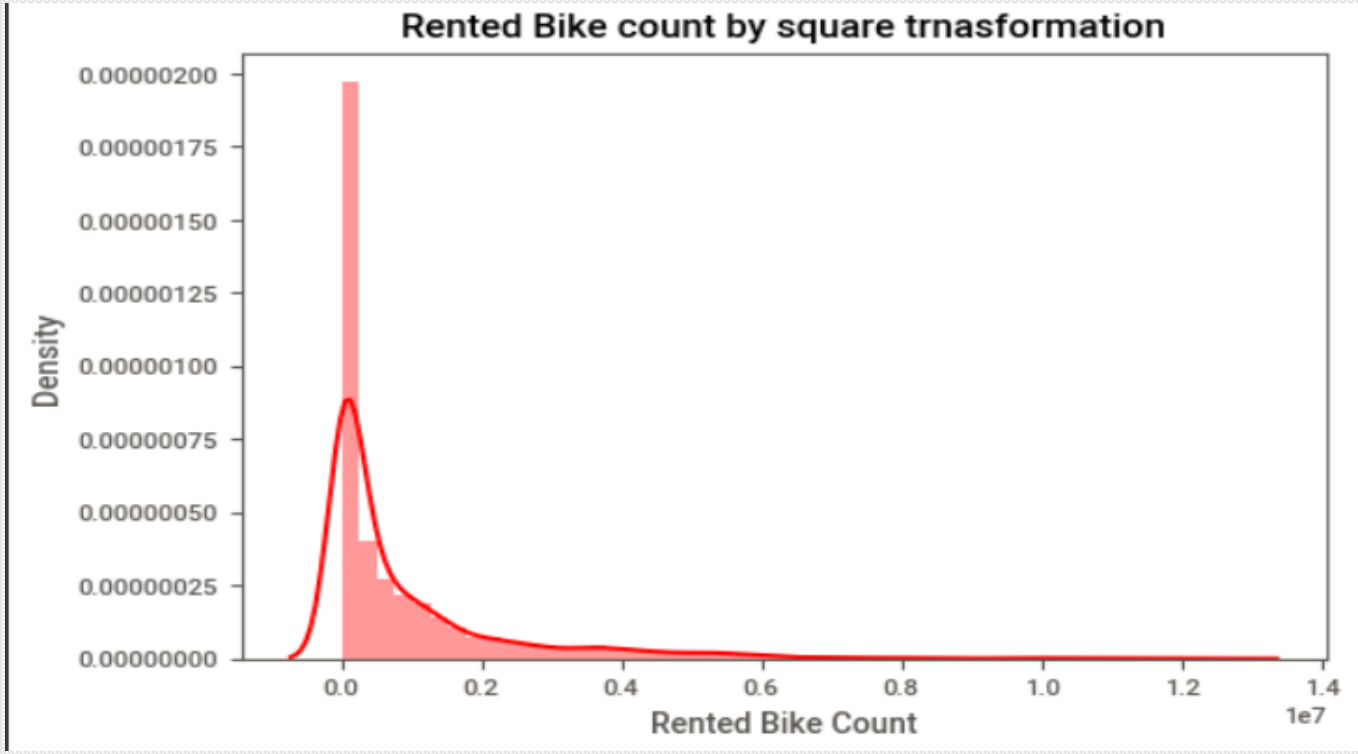
- Normalizing Dependent variable :- Since we know our Dependent feature is positively skewed so for the better model prediction we have to Normalize it.
- As shown above ->



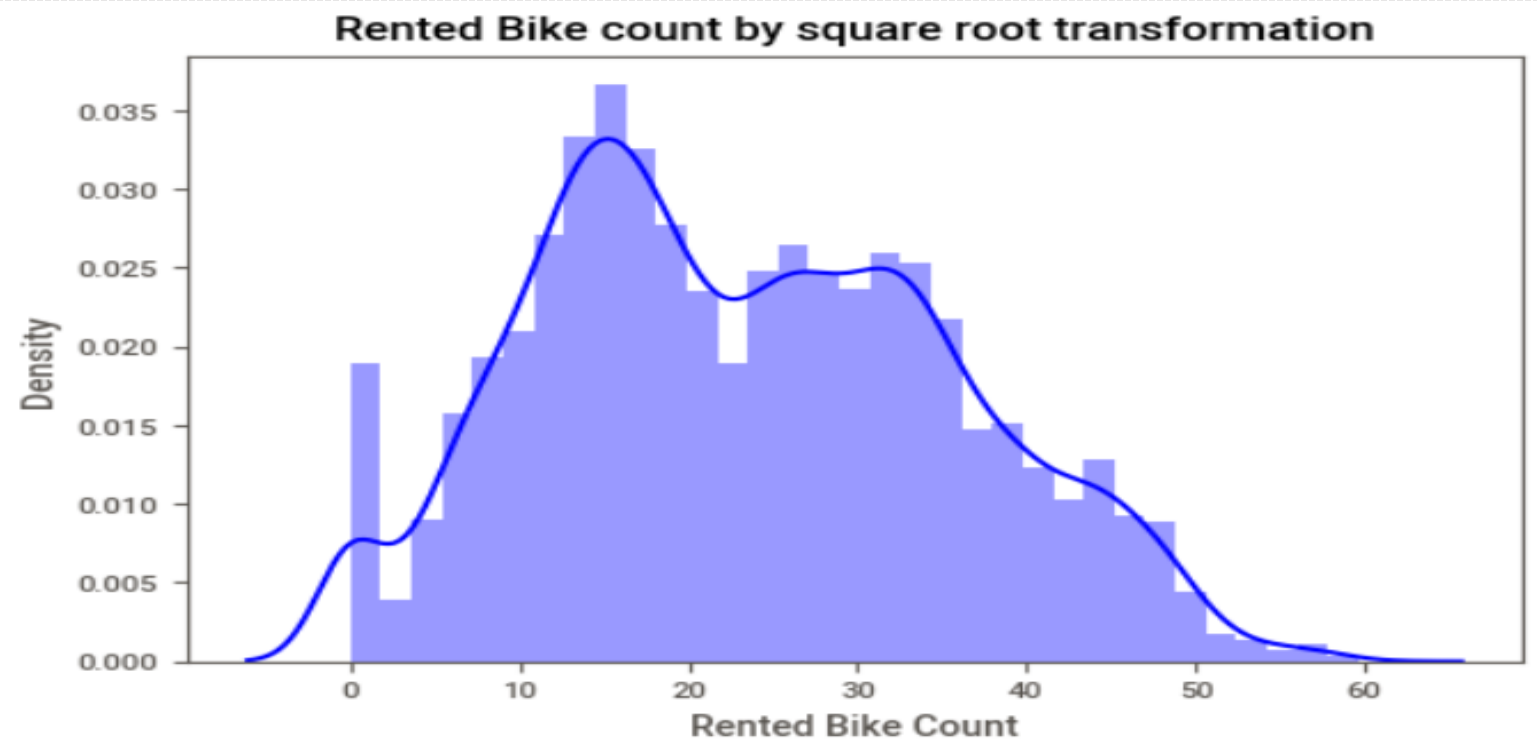
✓Method (1) – Log transformation



✓Method (2) – square transformation



✓Method (3) – Square root Transformation



✓So by comparing above three transformation methods we clearly see that **sqrt transformation** on dependent feature gives normal distribution.

➢ Separating Dependent and Independent features :-

✓ **Independent feature =** All features in data frame except “Rented bike count” after performing feature engineering.

✓ **Dependent Feature** = “Rented Bike count”

➢ Splitting Data for Training and testing the model :-

✓ Splitting 80% data from data frame for training model.

✓ And 20% data for testing the model, i.e. how much accurately we have predicted the bike sharing demand.

❖ *Model (1) – Linear Regression*

✓Linear regression analysis is **used to predict the value of a variable based on the value of another variable**. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

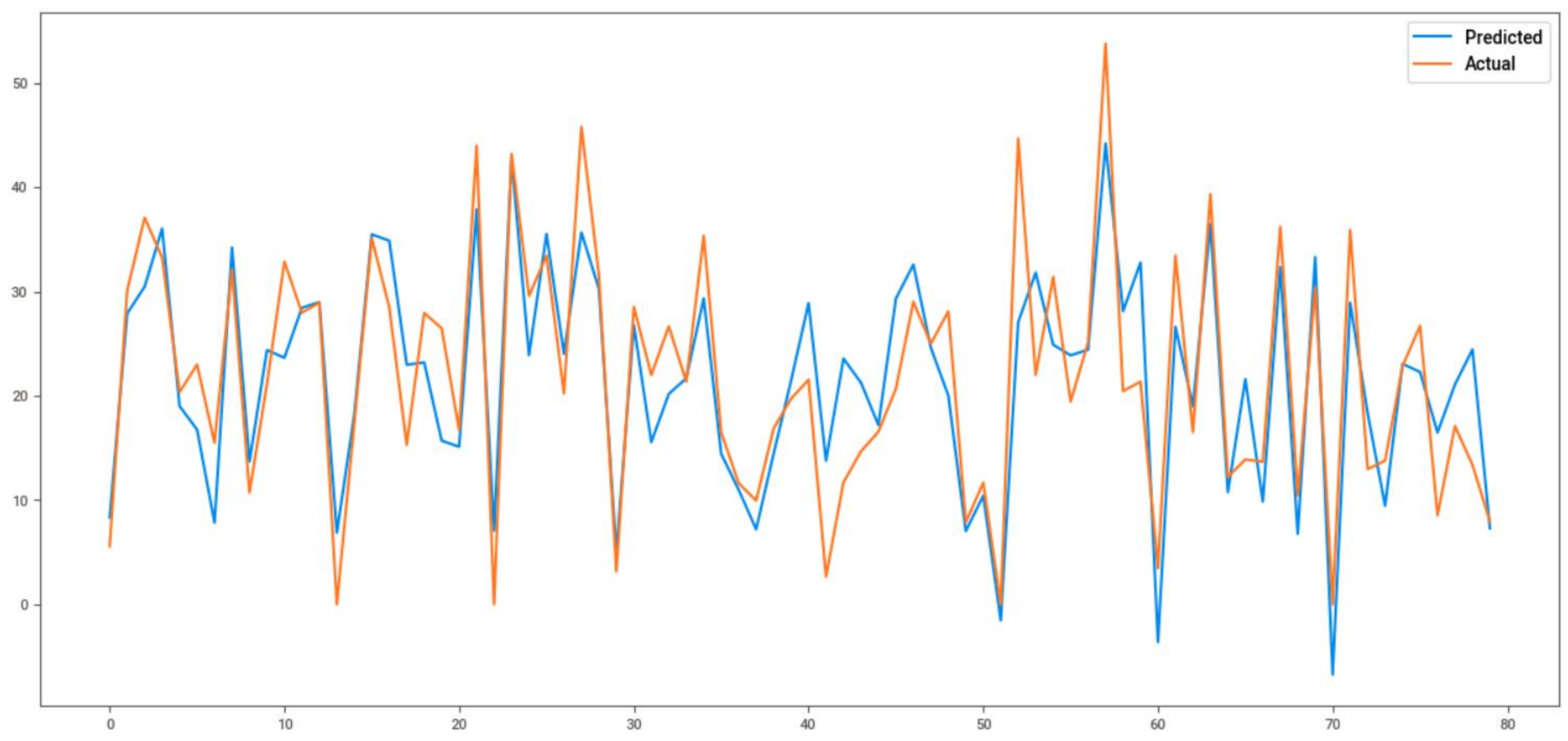
➢Evaluation Matrices :-

```
Training score  = 0.7031168479271276
MAE : 5.256832985963903
MSE : 44.35376150715583
RMSE : 6.659861973581422
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  =  -0.19119801675653747
```

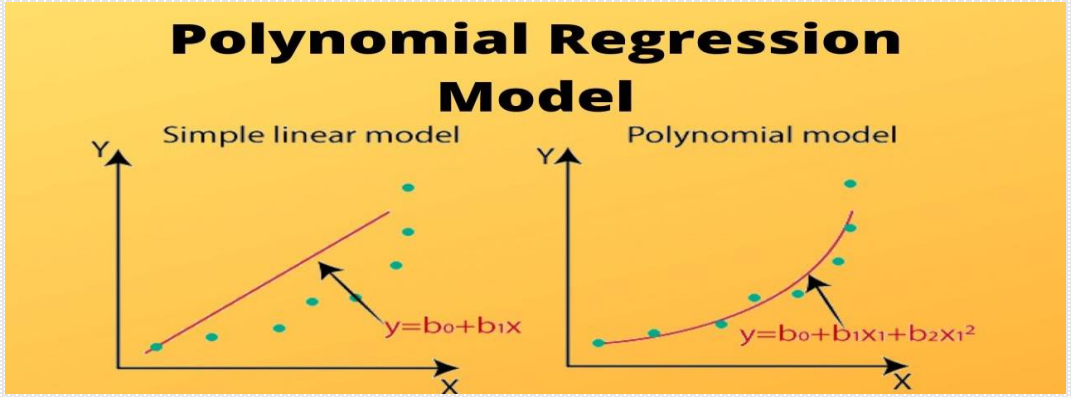
✓Graph of Actual v/s Predicted values of bike rent count prediction for LR model



❖ *Model (2) – Linear Regression using polynomial*

✓Polynomial regression is a form of Linear regression where only due to the Non-linear relationship between dependent and independent variables we add some polynomial terms to linear regression to convert it into Polynomial regression.

✓By using polynomial regression we can increase model performances.



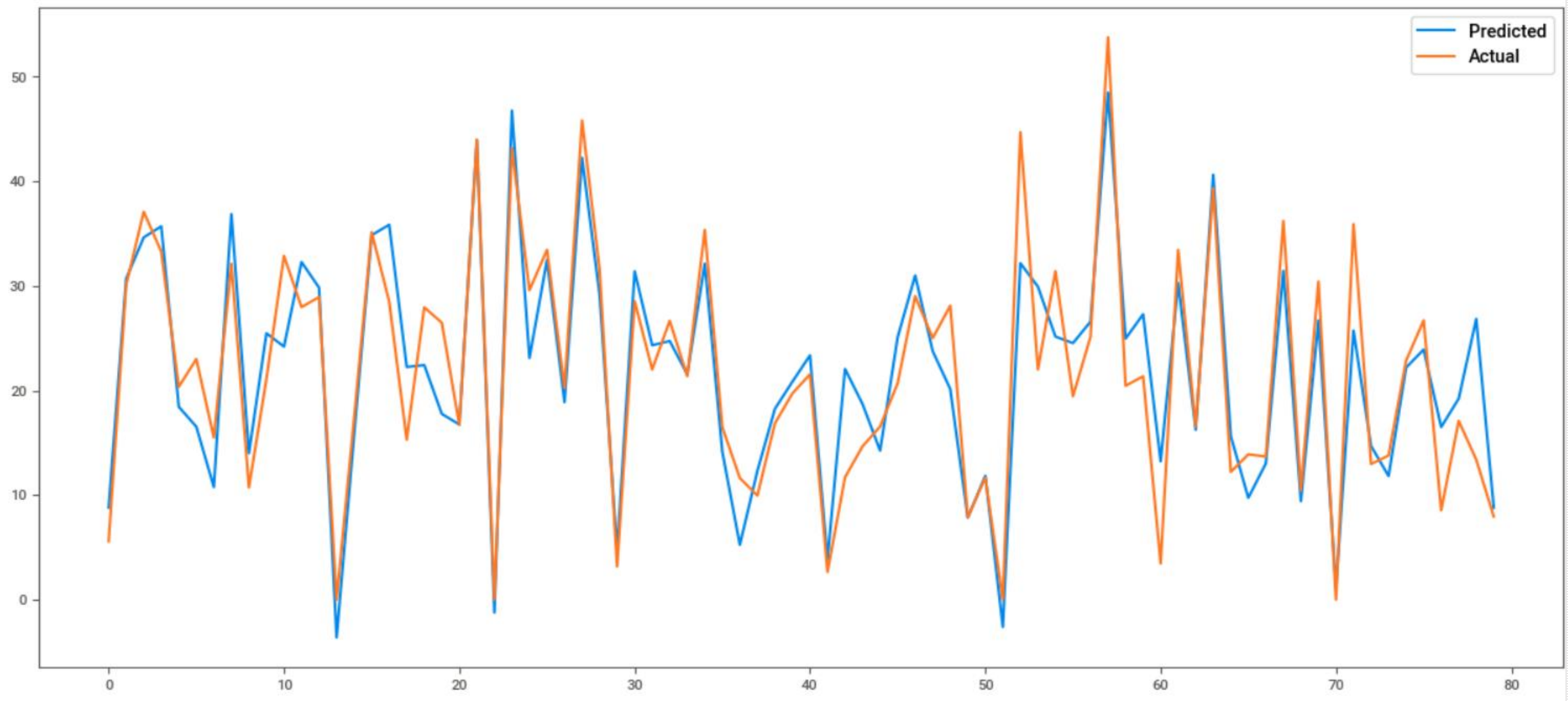
✓Evaluation matrices – As we can see that

Training accuracy of model increased as

Compared to simple linear regression model.

```
Training score  = 0.8514176478921974
MAE : 3.7037638918076135
MSE : 24.62886174549538
RMSE : 4.962747398920822
R2 : 0.8423883290869492
Adjusted R2 :  0.7902902463763283
```

✓Graph of Actual v/s Predicted values of bike rent count prediction by Polynomial regression model



❖ *Model (3) – Lasso Regression*

✓Regularization is one of the most important concepts of machine learning. It is a technique to prevent the model from over fitting by adding extra information to it.

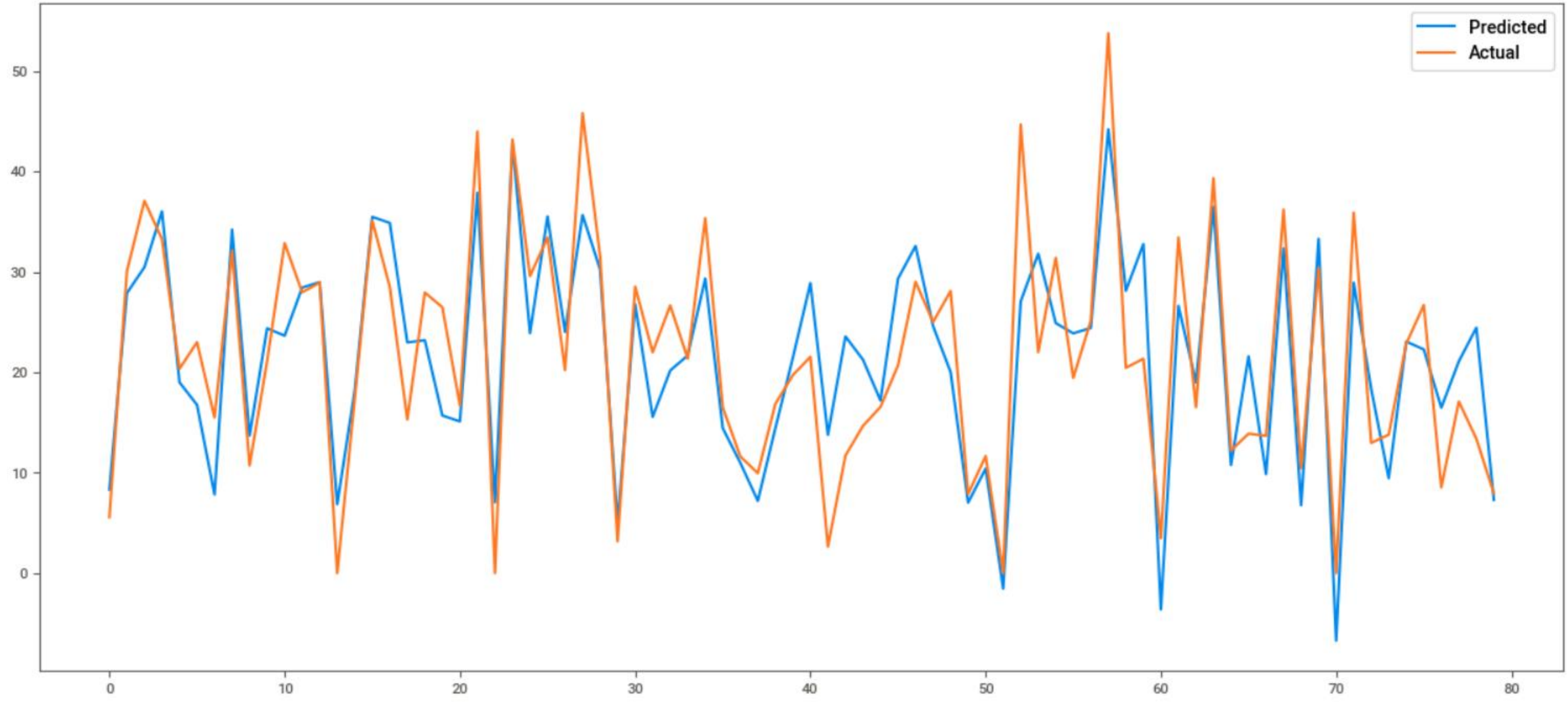
✓Two techniques of regularization are = 1) Lasso (l1 norm) and 2) Ridge regression (L2 norm)

✓Evaluation matrices for Lasso regression -

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

MAE : 5.06169849951929
MSE : 41.51236055593674
RMSE : 6.443008657136566
R2 : 0.7343428787583781
Adjusted R2 :  0.7300257578095879
```

✓Graph of Actual v/s Predicted values of bike rent count prediction by Lasso Regression model -



❖ *Model (4) – Ridge Regression*

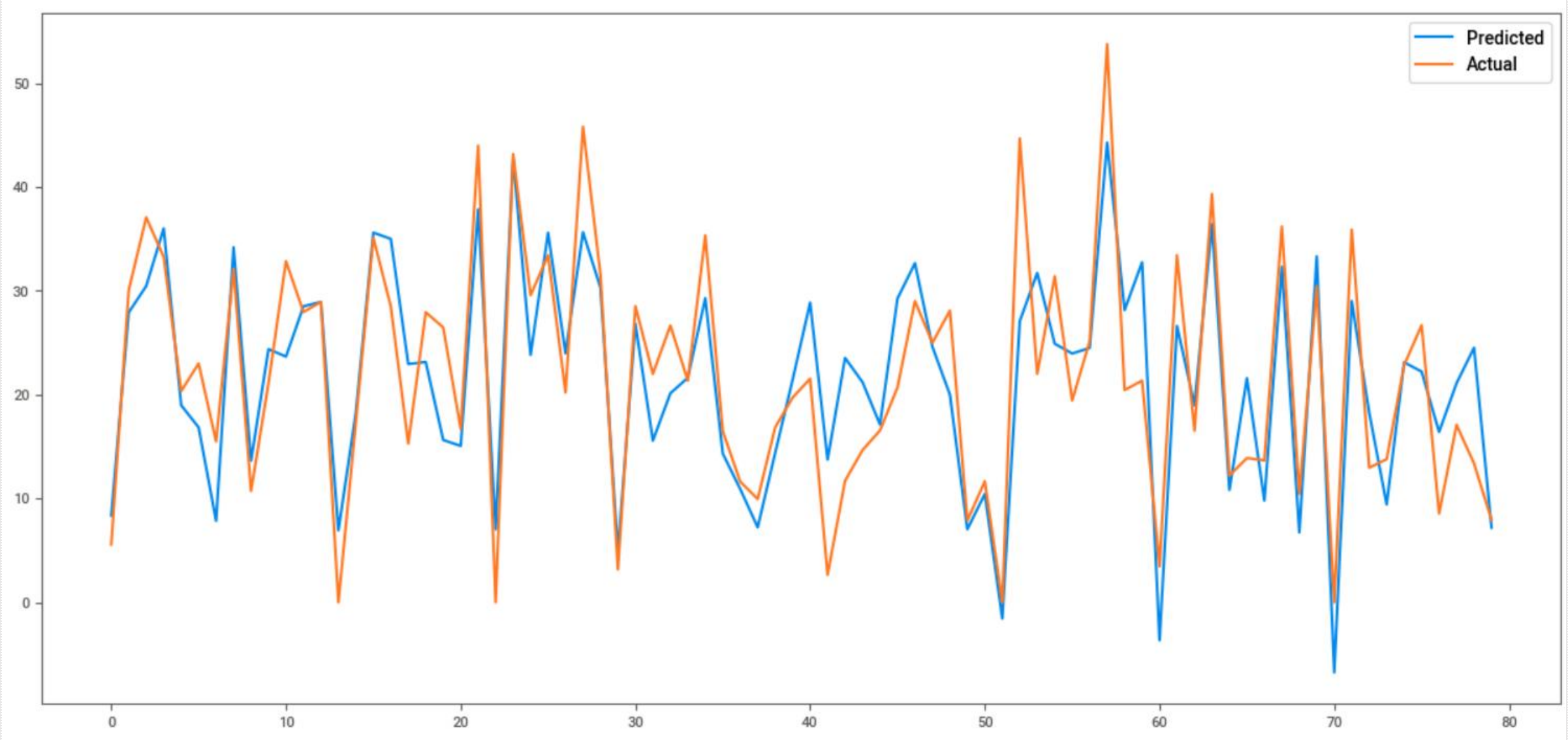
✓Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values.

✓Evaluation matrices

```
Training score  = 0.7358671502026775
The best parameters found out to be :{'alpha': 10}
where model best score is:  0.7328824117444819

MAE : 5.062807136059967
MSE : 41.52017837206752
RMSE : 6.443615318442553
R2 : 0.7342928488757146
Adjusted R2 :  0.7299749149050355
```

✓Graph of Actual v/s Predicted values of bike rent count prediction by Ridge Regression model -



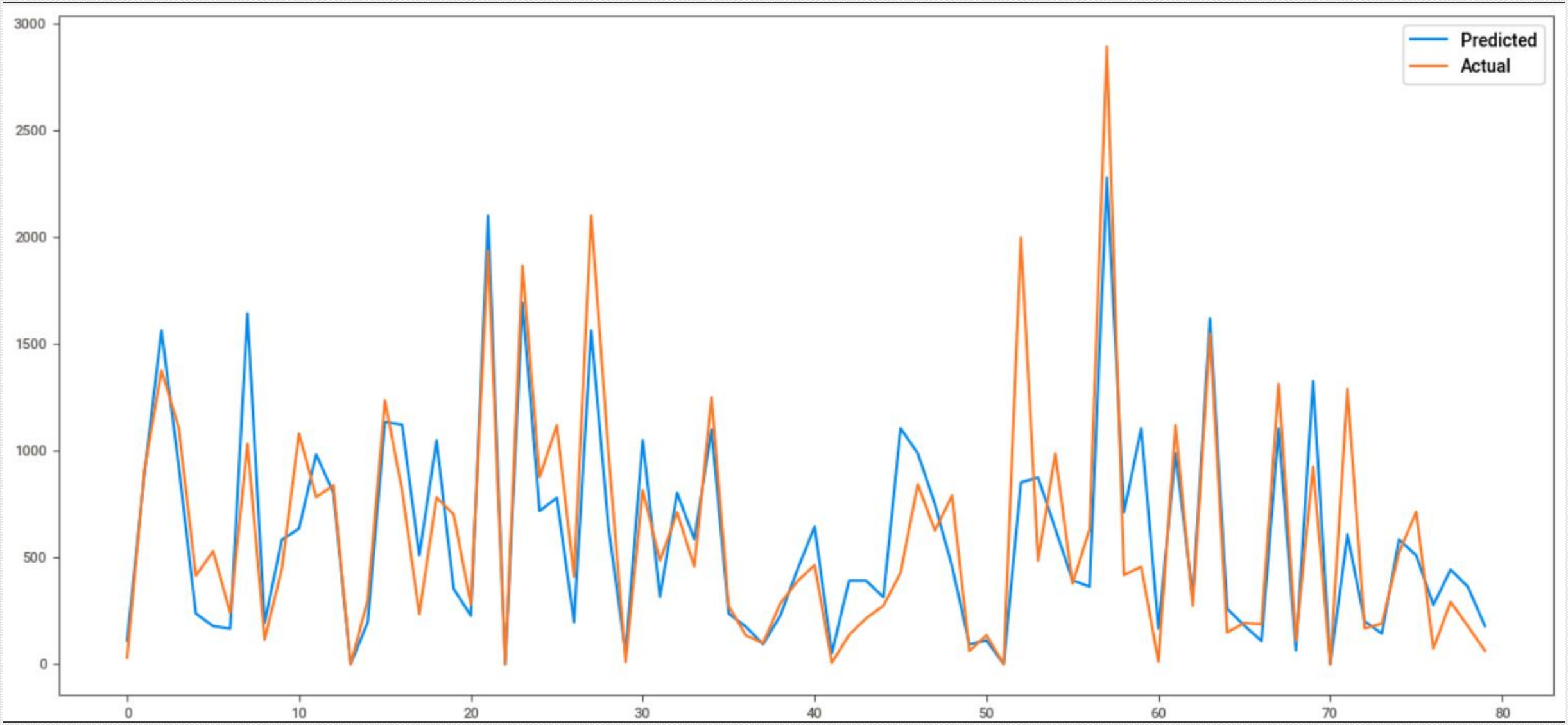
❖ Model (5) – Decision Tree

- ✓A decision tree is a type of supervised machine learning used to categorize or make predictions based on how a previous set of questions were answered. The model is a form of supervised learning, meaning that the model is trained and tested on a set of data that contains the desired categorization.
- ✓Since in decision tree multicollinearity of features does not affect the model accuracy. So in previous models we have removed multicollinear features (such as "Dew Point Temperature"). So here no need to remove any collinear features.
- ✓Evaluation matrices -

```
Training score  = 0.815011698329165
The best parameters found out to be :{'criterion': 'mse', 'max_depth': 15, 'max_features': 24, 'min_samples_split': 100, 'splitter': 'best'}
where model best score is:  0.7628017455589069

MAE : 205.27914126160096
MSE : 93623.91277816208
RMSE : 305.9802490000982
R2 : 0.7694325722303275
Adjusted R2 :  0.7655496132260763
```

✓Graph of Actual v/s Predicted values of bike rent count prediction by Decision tree algorithm -



❖ *Model (6) – Random Forest Regression*

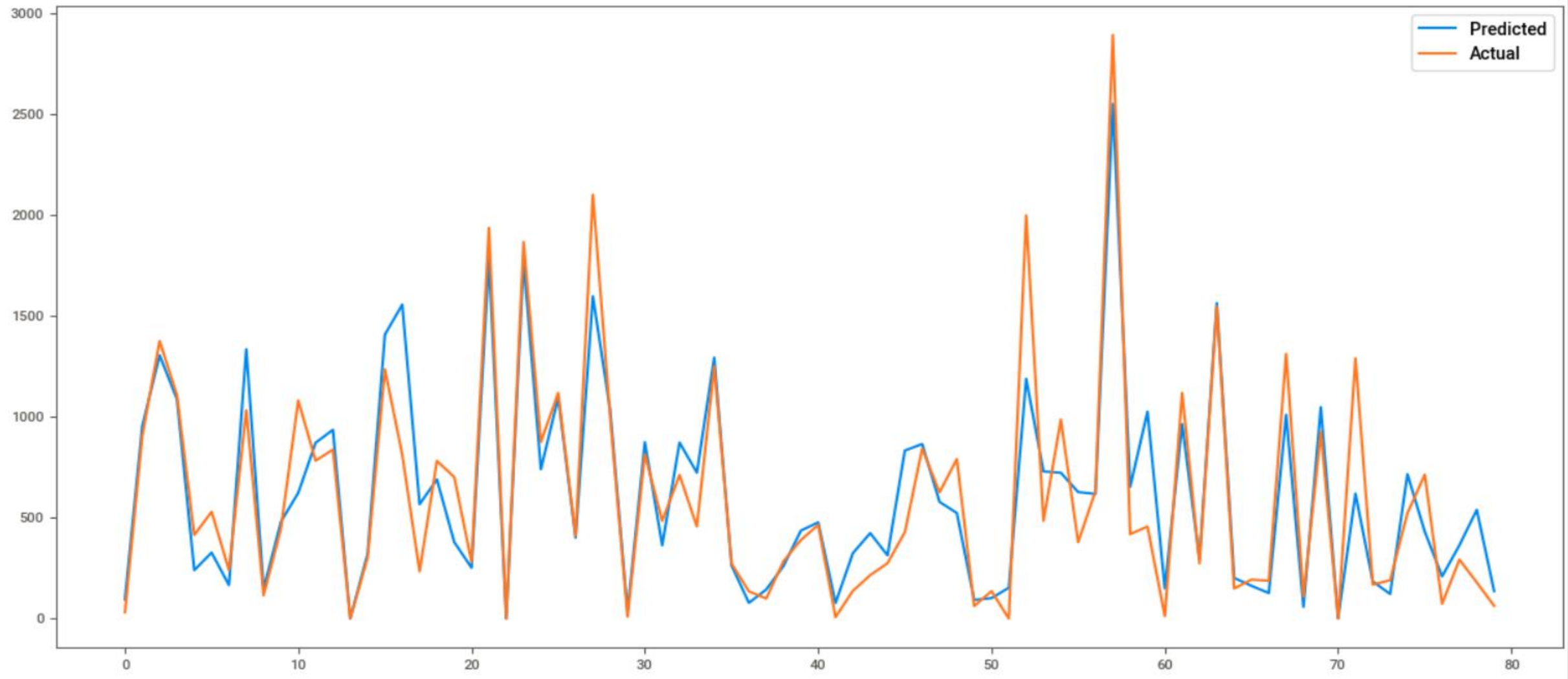
✓Random Forest is a powerful and versatile supervised machine learning algorithm that grows and combines multiple decision trees to create a “forest.” It can be used for both classification and regression problems in R and Python.

✓Evaluation matrices -

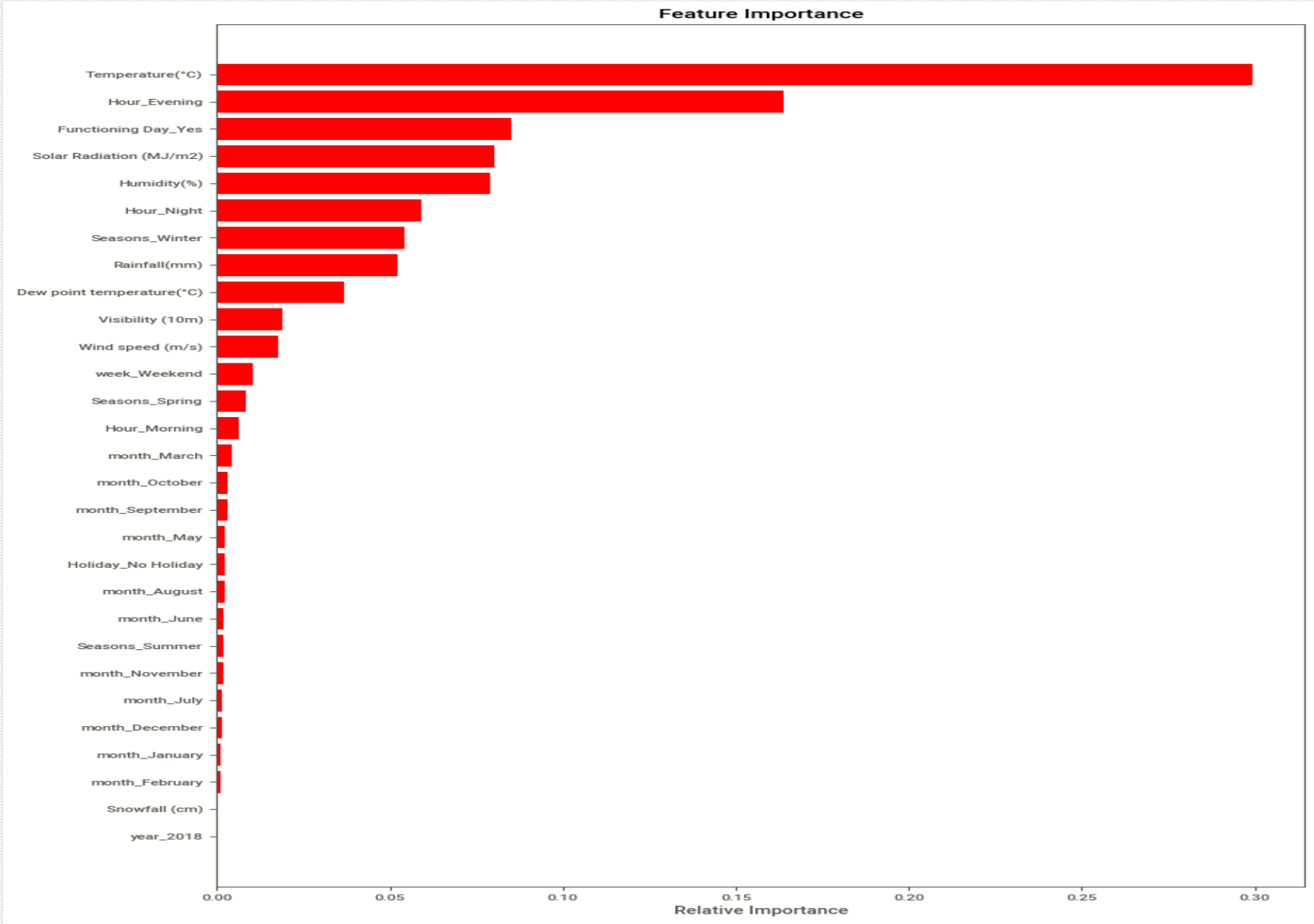
```
Fitting 5 folds for each of 180 candidates, totalling 900 fits
Training score  = 0.9458669379246406
The best parameters found out to be :{'max_depth': 25, 'max_features': 24, 'min_samples_split': 10, 'n_estimators': 150}
where model best score is:  0.8358874498856881

MAE : 163.05959090465572
MSE : 64255.2176672975
RMSE : 253.48612914180828
R2 : 0.841758800516775
Adjusted R2 :  0.8390938790388345
```

✓Graph of Actual v/s Predicted values of bike rent count prediction by Random Forest Regression -



❖ Feature Importance's for predicting bike rent count



❖ Conclusion of Project

➢ From EDA :-

- ✓ Most number of bikes are rented in the Summer season and the lowest in the winter season.
- ✓ 98% of the bikes are rented when there is non Holiday day present. That means Most of the user may use bike on rent to go there respective work places.
- ✓ Most number of bikes are rented in the temperature range of 15 degrees to 30 degrees.
- ✓ Most number of bikes are rented when there is no snowfall or rainfall.
- ✓ Majority of the bikes are rented for a humidity percentage range of 30 to 70.
- ✓ Gradual Increase in bike rent count is in morning 6 to 10 am i.e. it must be working time of employees. And after 10 am there slight decrease in count, And again start increasing count rate from 16 to 20 (4pm to 8pm) i.e. it must be leaving time of employees and they uses bike on rent to go there home.
- ✓ The highest number of bike rentals have been done in the 18th hour, i.e 6pm, and lowest in the 4th hour, i.e. 4am.
- ✓ Most of the bike rentals have been made when there is high visibility.
- ✓ In 2018 demand for Rented bikes is increased as compare to 2017 year. It may be because in 2017 people are aware about rented bike facility.

➢ *From Model Building :-*

- ✓ As seen in the Model Evaluation Matrices table, Linear Regression is not giving great results. i.e. we say that model is in over fitting condition.
- ✓ Over fitting – when model perform well with training data but fails to perform well with test data. i.e. LR model has Low Bias and High Variance.
- ✓ Random forest have performed well in terms of evaluation matrices score as compared to other models.
- ✓ So we can use Random forest regression model for better prediction of Bike sharing count which is our ultimate goal of this project.


THANK
YOU