Seoul Bike Sharing Demand Prediction

Submitted by:
Shahfaissal I Dharwad
Neha Gupta,
Ajinkya Jumde,

---**ABSTRACT**---

This study predicts the demand for shared bikes based on the different features which are either numerical or categorical. The detailed Exploratory Data Analysis followed by Feature Engineering and then fitting the dataset for regression model helps us to predict the demand for the Bikes. An analysis with variable importance was carried to analyze the most significant variables was carried of using OLS, Lasso and Ridge regression. The variable importance results have shown that Temperature and Hour of the day are the most influential variables in the hourly rental bike demand prediction.

Keywords: 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.

1. INTRODUCTION

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. This dataset contains the hourly and daily count of rental bikes between in 2017-2018 in in Seoul with the corresponding weather and seasonal information. Bike sharing is one of the ways to reduce urban traffic. It also reduces air pollution by reducing the number of cars on the road. The bike sharing system is a new generation of traditional bike rental systems, and the entire process has been automated. Users can borrow bicycles for free or for a fee and return them to another place.

The hypothesis in the research is that the bike sharing is highly related with the time of the day, season, and weather conditions. The research will try to predict the bike shares in the future. We are going to try to understand the factors on which the demand for these shared bikes depends.

There are few attributes of the dataset given below:

- Date: year-month-day
- Rented Bike count: Count of bikes rented at each hour
- Hour: Hour of the day
- Temperature: Temperature in Celsius
- Humidity %
- Windspeed-m/s
- Visibility-10m
- Dew point temperature-Celsius
- Solar radiation-MJ/m2
- Rainfall-mm
- Snowfall-cm
- Seasons: Winter, Spring, Summer, Autumn
- Holiday: Holiday/ No holiday
- Functional Day: No Functional (Nonfunctional Hours)
- Fun (Functional hours)

2. PROBLEM STATEMENT

Explore the given dataset and after proper EDA select the most relevant features in order to predict the demand for bike based on the atmosphere in the city, day type(weekend/weekday) and hour. The types of regression models used is OLS, Lasso and Ridge Regression

3. STEPS INVOLVED

a. Python Libraries:

NumPy: NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
Pandas: Pandas is an open-source library that provides high-performance data manipulation in Python.

Matplot: Matplotlib is a python library used to create 2D graphs and plots.
Seaborn: Seaborn is a library for making statistical graphics in Python.
Sklearn: Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.
Datetime: Datetime library which is generally used for calculating differences in dates and also can be used for date manipulations in Python. It is one of the easiest ways to perform date manipulations.

b. **Importing a Data frame:**

Data frame has been imported from google drive and read data frame applying read_csv.

c. **Recognizing the Data frame:**

Dealing with a huge data set is a time-consuming part. To minimize the workload and efforts we must have to distribute data and analyze the contents first.

d. **Dealing with Null Values and Outliers:**

As we have seen, our data frame has no null values so we don't need to deal with null values at the beginning of our project, Whereas the boxplot demonstrated a lot of outliers that may affect the functioning of the regression model and so as to deal with them we try to filter them and replace them with the relevant values.

e. **Analyzing the Data:**

Univariate Analysis: - ***Univariate Analysis*** is the key to understanding each and every variable in the data. Learn how to visualize and interpret ***univariate*** data.

Multivariate Analysis: ***multivariate*** data is to make a matrix scatterplot, showing each variable plotted against our target variable that is Rented Bike Count.

f. **Functions & Methods applied for Data frame:**

Using statistical data, we have represented the various types of graphs. We used Seaborn bar plots and box plots for the demonstration of our analysis.

As per the analysis we found that the distribution of the Rented Bike count was positively skewed and so as to normalize the given dataset, we used Square root function using Numpy library so as to normalize the distribution.

g. Creating heat map and find co-relation between different columns with each other:

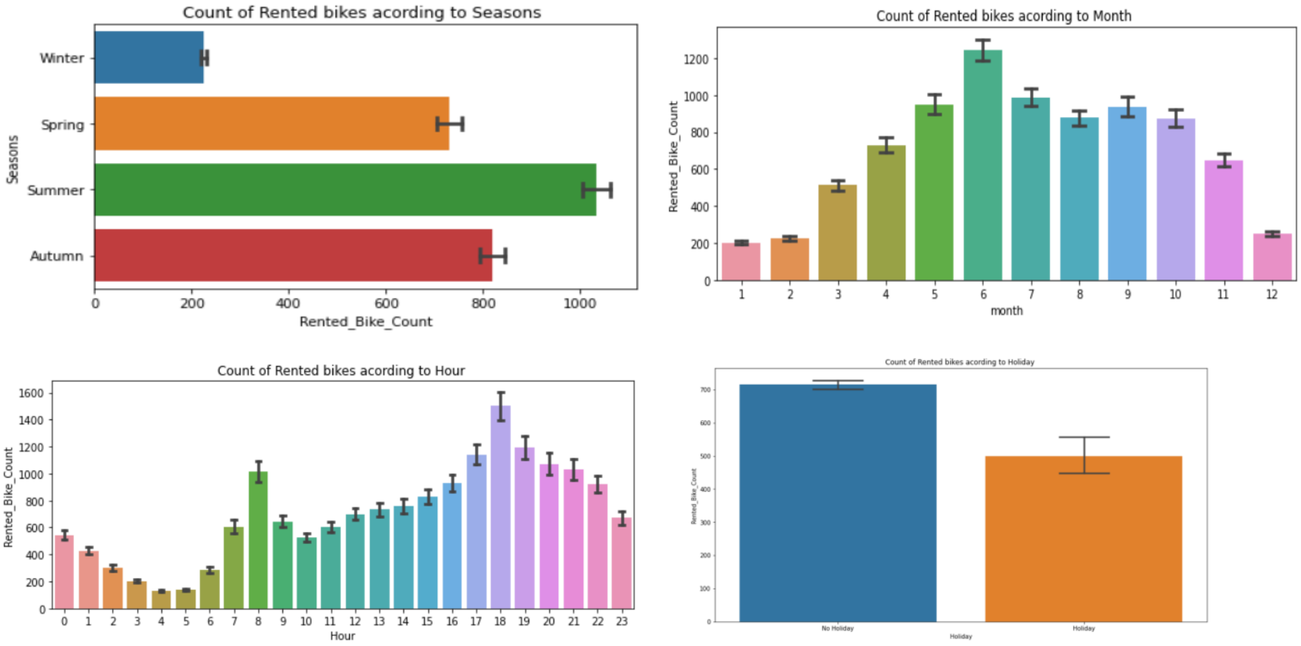
We have created a heat map between columns to find the co-relationship between all the columns with the help of correlation of statistical method. The Temperature and Dew point temperature were found to be highly correlated and so to avoid multicollinearity we dropped the second column from the dataset.

Moreover, we found the negatively correlated attributed in the heatmap which were not relevant for the regression model and so they were dropped from the further analysis. The negatively correlated features were Snowfall and Visibility.

4. Data Analysis:

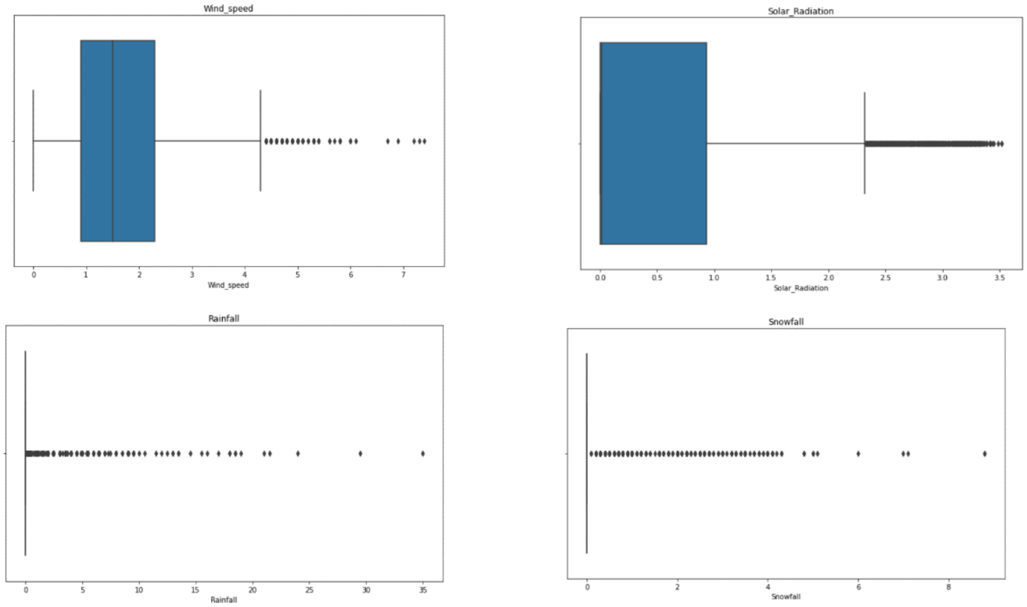
Eda is performed with the data frame between various independent variables against our dependent variable that is Rented Bike Count and visualization of the result is done using various plots such as scatter plot, boxplot, bar plot, histogram, heatmap, line chart, few of the important analysis is shown below.

1. **Effect of Season on Rented Bike Demand:** During the multivariate analysis using bar plot, it was found that the season has a significant effect on the Rented Bike Demand. It is highest during the Summer season whereas it is almost same during Autumn and Spring. Whereas during Winter, the demand is quite less.
2. **Effect of Month on Rented Bike Demand:** During the analysis it was found that the month plays a significant role in Rented Bike Demand. As clearly based on the analysis of seasons, we were able to conclude that the atmosphere is one of the key attribute in demand prediction, from the analysis based on month we can conclude that the demand is highest in the month of May, June, July.
3. **Effect of Hour on Rented Bike Demand:** During the analysis the bar plot plotted between the demand and hour clearly shows that the demand is at the peak in the late evening and it is quite high in the early morning that is during the office hours.
4. **Effect of holiday on Rented Bike Demand:** During the analysis the plot clearly shows that the demand is higher on working day as compared to holiday.

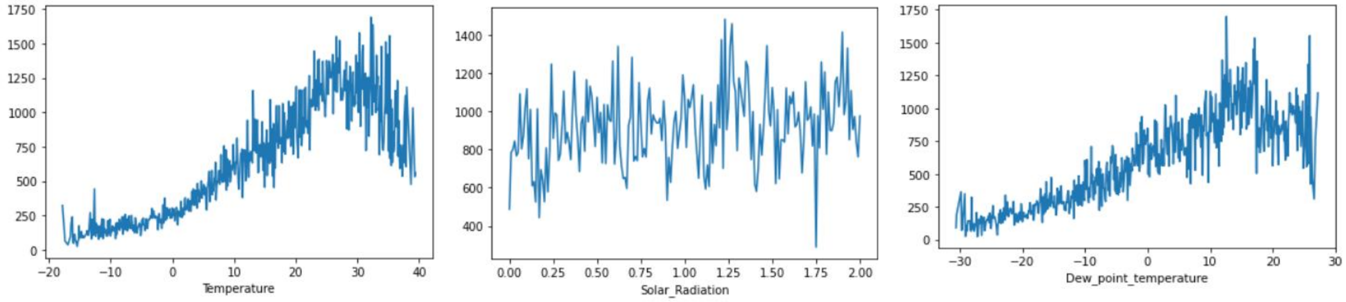


The above graphs show the multivariate analysis between the dependent variables with respect to the dependent variable

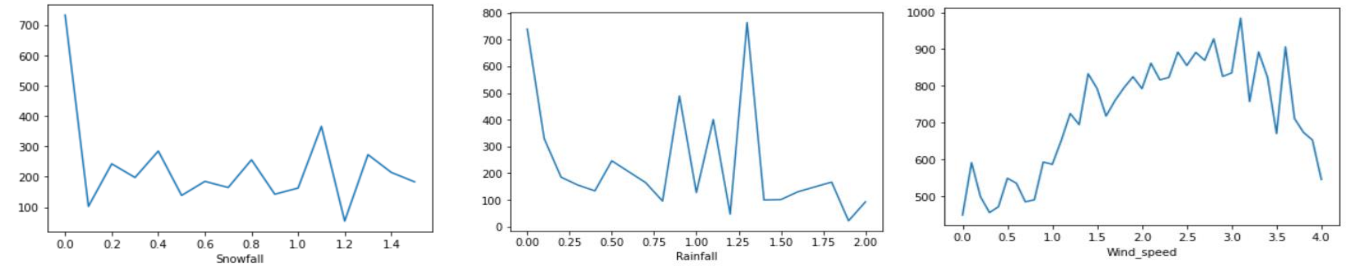
e. **Box plot analysis for outlier detection**: Based on the box plots plotted for the detection of outliers, it was found that some of the attributes were having outliers. To remove the outliers, we filter the data as per the standard values for the variables with outliers. The variables having outliers were Wind Speed, Solar Radiation, Rainfall, Snowfall. To remove these outliers we change the data at outlier points to the most standard data point available that is for example, in case of rainfall the standard range for moderate showers is maximum upto 2mm so we replace all the values greater that 2mm with 2 mm. Same is the case for Wind Speed where we replace the values greater than 4m/s with 4 m/s.



f. **Analysis of effect of independent variables on dependent variable:** After the Outlier detection and correction, we try to analyze the effect of theses features on our dependent variable.



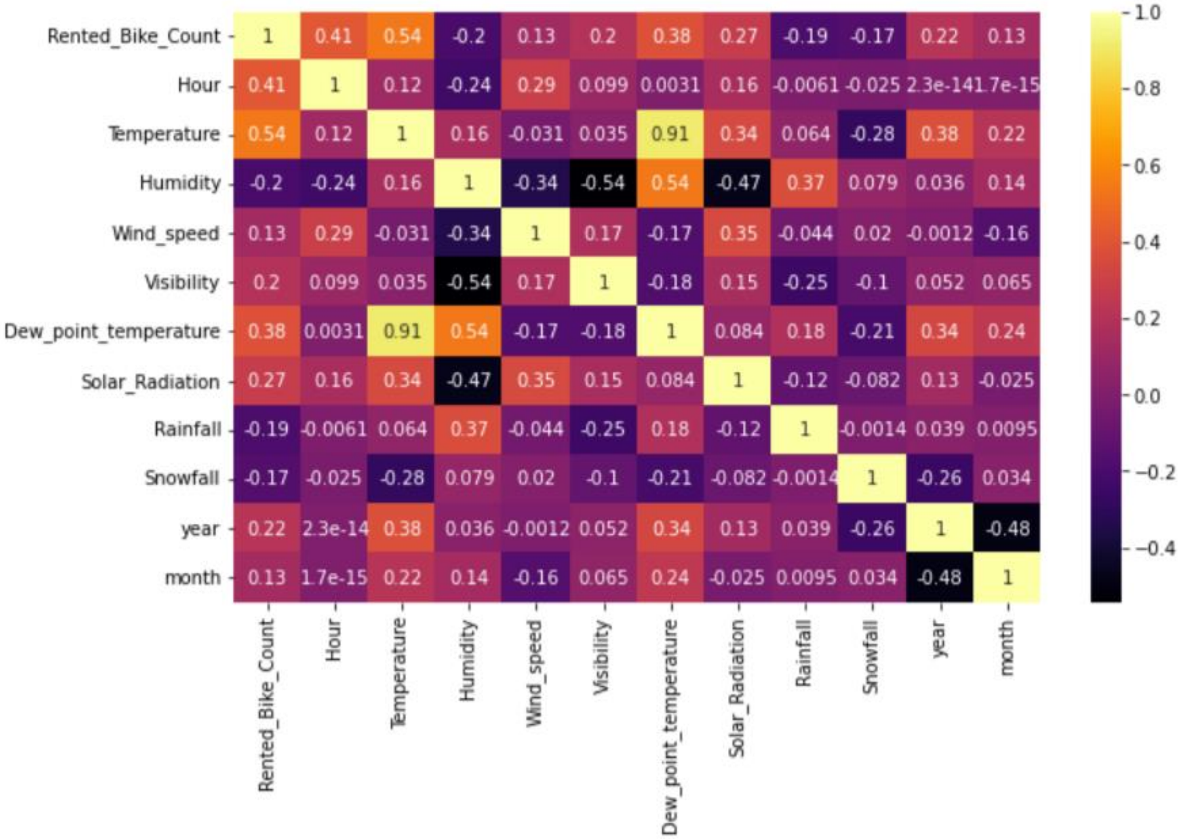
The above graphs



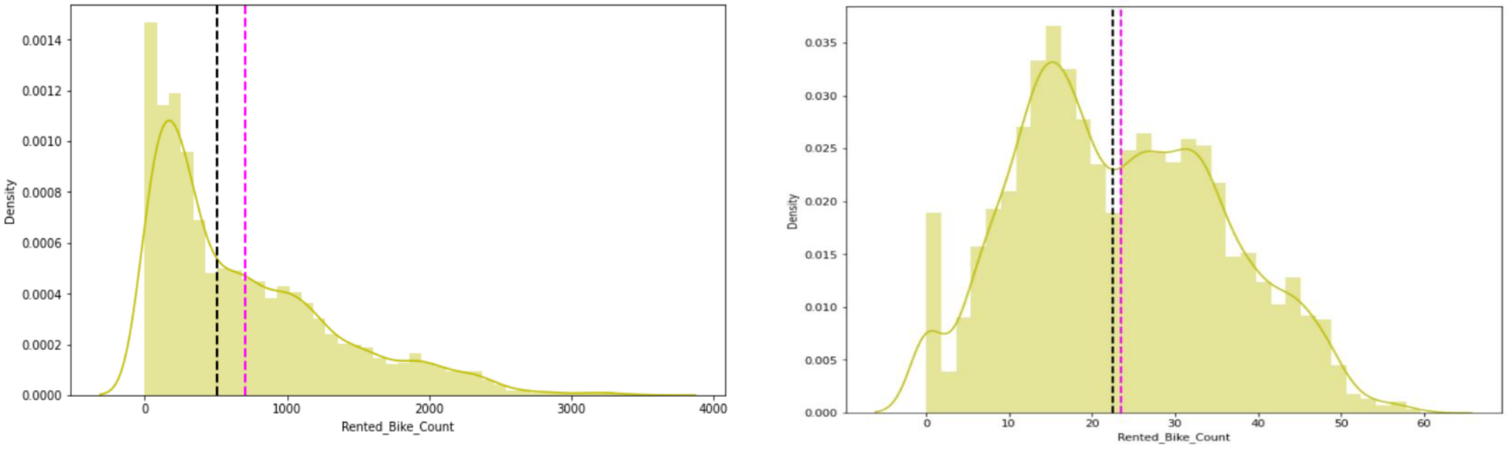
show the variation of dependend variable with respect to the numeric features in the given dataset.

g. **Detection of multicollinearity in the dataset:** After plotting the heatmap with respect to the collinearity in the given dataset. The highly collinear features were filtered by dropping one of the features. And the negatively correlated features were dropped as well. The Temperature and Dew point temperature were found to be highly correlated and so to avoid multicollinearity we dropped the second column from the dataset.

Moreover, we found the negatively correlated attributed in the heatmap which were not relevant for the regression model and so they were dropped from the further analysis. The negatively correlated features were Snowfall and Visibility

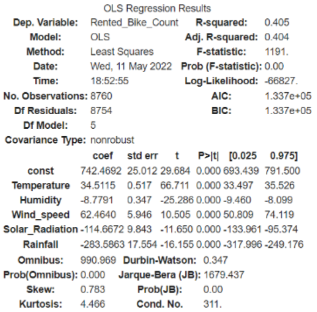


h. Analysis of Dependent Variable: Based on the histogram plot of the dependent variable, it is clear that the distribution is positively skewed. In order to normalize the distribution, we use sqrt function in numpy to normalize the variable distribution



The above graphs show the distribution of rented bike counts before and after applying square root function.

5. Linear Regression Model (OLS): After proper analysis of the data many features were dropped or modified in accordance with the regression model requirement. Now using the ‘statsmodels.api’ library, we assign the dependent variables and dependent variable as a dataset to X and Y. And we try to get the model summary using ‘model.summary’ method. And ‘.corr’ to verify the correlation.

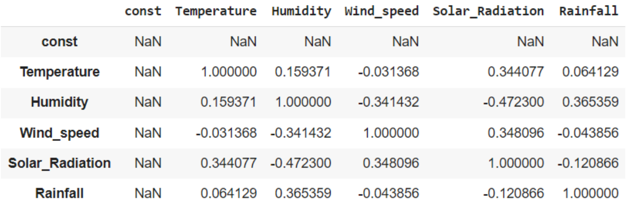


OLS Regression Results

Dep. Variable:	Rented_Bike_Count	**R-squared:**	0.405
Model:	OLS	**Adj. R-squared:**	0.404
Method:	Least Squares	**F-statistic:**	1191.
Date:	Wed, 11 May 2022	**Prob (F-statistic):**	0.00
Time:	18:52:55	**Log-Likelihood:**	-66827.
No. Observations:	8760	**AIC:**	1.337e+05
Df Residuals:	8754	**BIC:**	1.337e+05
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	742.4692	25.012	29.684	0.000	693.439	791.500
Temperature	34.5115	0.517	66.711	0.000	33.497	35.526
Humidity	-8.7791	0.347	-25.286	0.000	-9.460	-8.099
Wind_speed	62.4640	5.946	10.505	0.000	50.809	74.119
Solar_Radiation	-114.6672	9.843	-11.650	0.000	-133.961	-95.374
Rainfall	-283.5863	17.554	-16.155	0.000	-317.996	-249.176

Omnibus:	990.969	**Durbin-Watson:**	0.347
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1679.437
Skew:	0.783	**Prob(JB):**	0.00
Kurtosis:	4.466	**Cond. No.**	311.

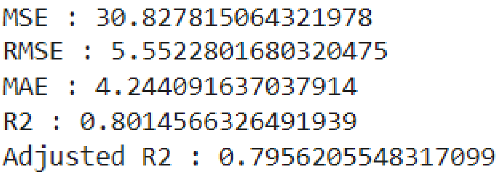


	const	Temperature	Humidity	Wind_speed	Solar_Radiation	Rainfall
const	NaN	NaN	NaN	NaN	NaN	NaN
Temperature	NaN	1.000000	0.159371	-0.031368	0.344077	0.064129
Humidity	NaN	0.159371	1.000000	-0.341432	-0.472300	0.365359
Wind_speed	NaN	-0.031368	-0.341432	1.000000	0.348096	-0.043856
Solar_Radiation	NaN	0.344077	-0.472300	0.348096	1.000000	-0.120866
Rainfall	NaN	0.064129	0.365359	-0.043856	-0.120866	1.000000

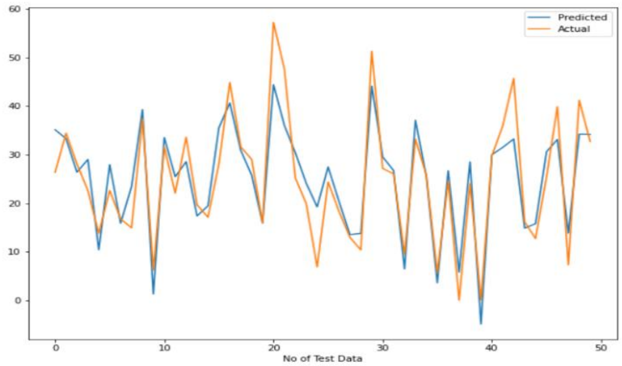
a. We assign the dependent variables to X and the dependent variable that is rented bike count to Y from the dataset. Later the splitting of data takes place and the data is split in the ration 80:20 for train test split.
b. The model is fit for linear regression in terms of X_train and y_train.
c. Subsequently, the coefficients are computer for each attribute of the dataset and based on the predicted values are computed on the test data.
d. The regression model output is then verified in order to compute MSE, MAE, RMSE, R2 and Adjusted R2.

e. The values show that the model is able to capture 80% of the variance and so we might consider this model as an efficient model to compute the dependent variable.

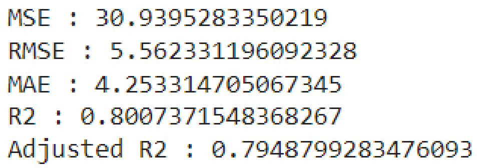
f. The plot between the predicted values and the actual values show that the model has predicted the rented bike count with high accuracy and with minimal residue.



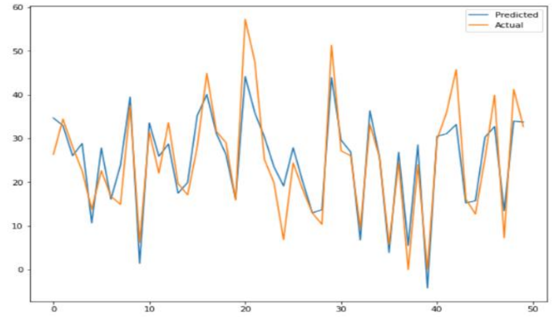
```
MSE : 30.827815064321978
RMSE : 5.5522801680320475
MAE : 4.244091637037914
R2 : 0.8014566326491939
Adjusted R2 : 0.7956205548317099
```



6. Lasso Regression: We try to apply lasso regression model to the same dataset and the values for the model were computed using ‘sklearn.metrics’ library. The R2 value signifies that the Lasso model was also able to capture the variance of the data efficiently and so there is high accuracy and low residue in the predicted data.

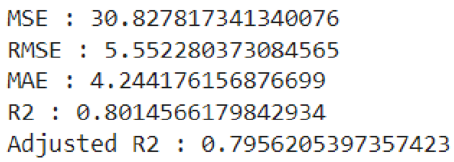


```
MSE : 30.9395283350219
RMSE : 5.562331196092328
MAE : 4.253314705067345
R2 : 0.8007371548368267
Adjusted R2 : 0.7948799283476093
```

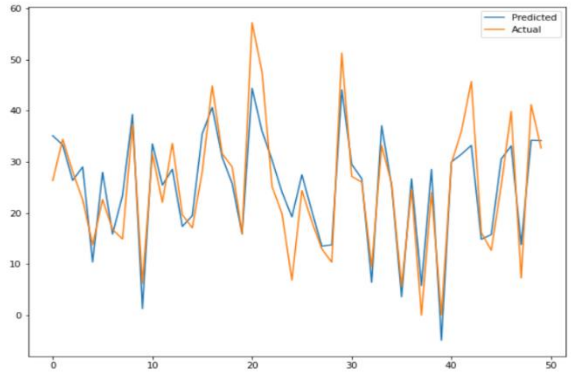


7. Ridge Regression Model: We try to apply Ridge Regression Model to the same dataset and the values for the model were computed using ‘sklearn.metrics’ library. The R2 value signifies that the Ridge Model was able to capture the variance of the data efficiently and so there is high accuracy and lower residue in the predicted data.

Conclusion: After the analysis we conclude that, this analysis will be helpful for the supplying entities in Seoul city to predict the demand for the rental bikes based on the primary and atmospheric data of the city. The predicted values were highly accurate and so the companies can rely on the model in order to predict the demand for the bikes on particular date based on temperature, humidity, wind speed, solar radiation, type of day (Holiday/No holiday), hour and month.



```
MSE : 30.827817341340076
RMSE : 5.552280373084565
MAE : 4.244176156876699
R2 : 0.8014566179842934
Adjusted R2 : 0.7956205397357423
```


Predicted
Actual