Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

From the dataset (bikesharing.csv), dependent variable is 'cnt'. ('registered' and 'casual' not considered as 'cnt' = 'registered' + 'casual').

Categorical variables are 'season', 'yr', 'mnth', 'holiday', 'weekday', 'workingday', 'weathersit'

From the visualization of the original dataset, these insights are made on categorical variables:

a) 'season' - Fall has highest demand for bike sharing as compared to Summer, Winter and Spring
b) 'yr' - Demand for rental bikes is higher in 2019 compared to 2018
c) 'mnth' - Demand is continuously increasing each month from Jan-Jun. September has the highest demand. Sep to Dec, demand decreases. Start of year and End of year, demand seems to be less. It could be possibly because of extreme weather conditions.
d) 'holiday' - On holidays, demand is less.
e) ' Weekday' - Demand is almost alike for weekday or weekend
f) 'workingday' - Demand on working day & non- working is almost similar
g) 'weathesit' - Highest demand on days with Clear weather. Lowest demand when there is light snow or light rain.

After multiple regression, following inferences about categorical variables:

- Summer and Winter have moderate impact on demand of bikes. Winter has more demand of bikes than Summer. Summer and Winter has more demand for bikes compared to the Fall and Spring.
- Weather conditions list Misty or Light rain or Snow, Windy day reduces the demand for bikes.
- Month of September has a moderate positive impact on demand for rental bikes. September seems to have higher usage of rental bikes compared to other months.
- Year by year also the demand is expected to grow.
- Saturdays and Working days are also expected to have slightly higher demand of rental bikes compared to holidays.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

If there are N categories for a categorical variable, then Number of dummy variables should be N-1. So, to drop the first category (represented by 0/1), we use **drop_first=True** during dummy variable creation.

For e.g.: variable 'season' has 4 categories [1-Spring, 2- Summer, 3-Fall,4-Winter]

When we create dummy variables
- Spring gets represented as - 000
-Summer gets represented as - 001
-Fall gets represented as - 010
-Winter gets represented as- 001
We will drop Spring (000) using **drop_first=True**, which means that if not summer /fall/winter then it represents Spring

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

The numerical variables in the dataset are :temp, atemp, hum, windspeed, casual, registered.

temp- temperature

atemp – feeling temperature

hum- humidity

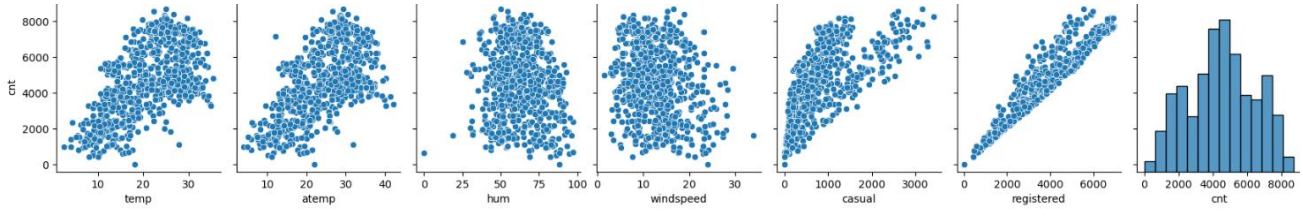
windspeed – speed of wind

casual – count of casual users

registered – count of registered users

cnt – count of total users (casual + registered)

The pair plot of each of these numerical variables with cnt is shown below:



In the above plot, registered and casual have highest correlation with target variable 'cnt'.

Also temp and atemp are similar values, hence we will drop one of these.

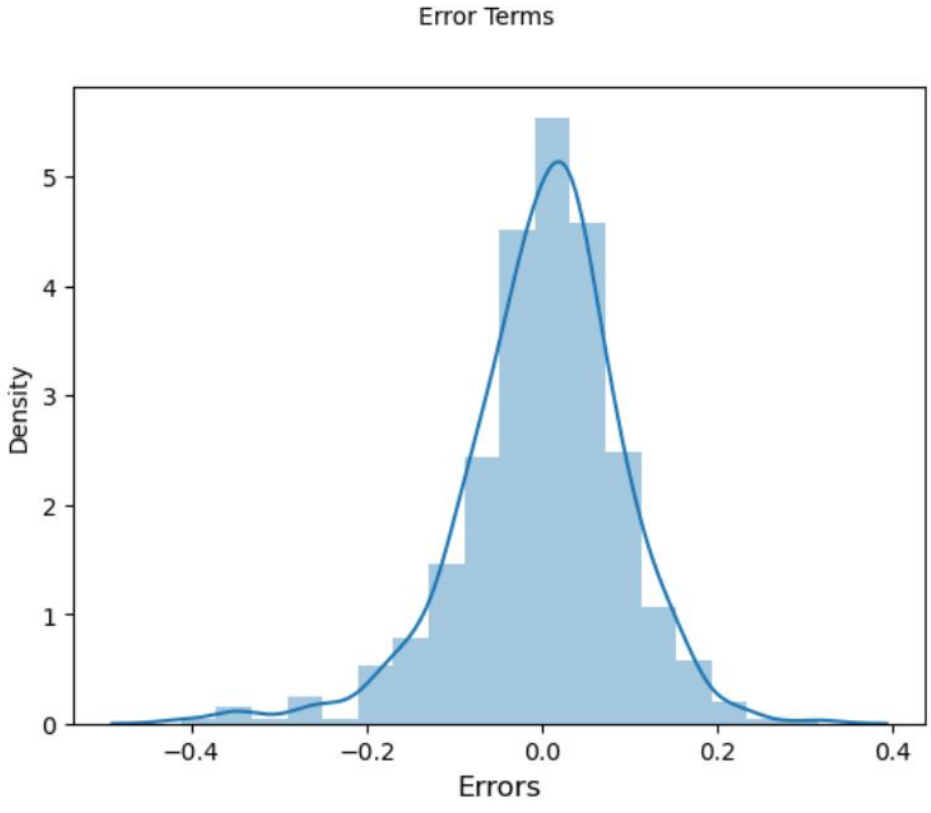
For analysis, we will drop casual, registered and atemp.

From the remaining numerical variables, temp seems to have highest correlation with cnt.

4.How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

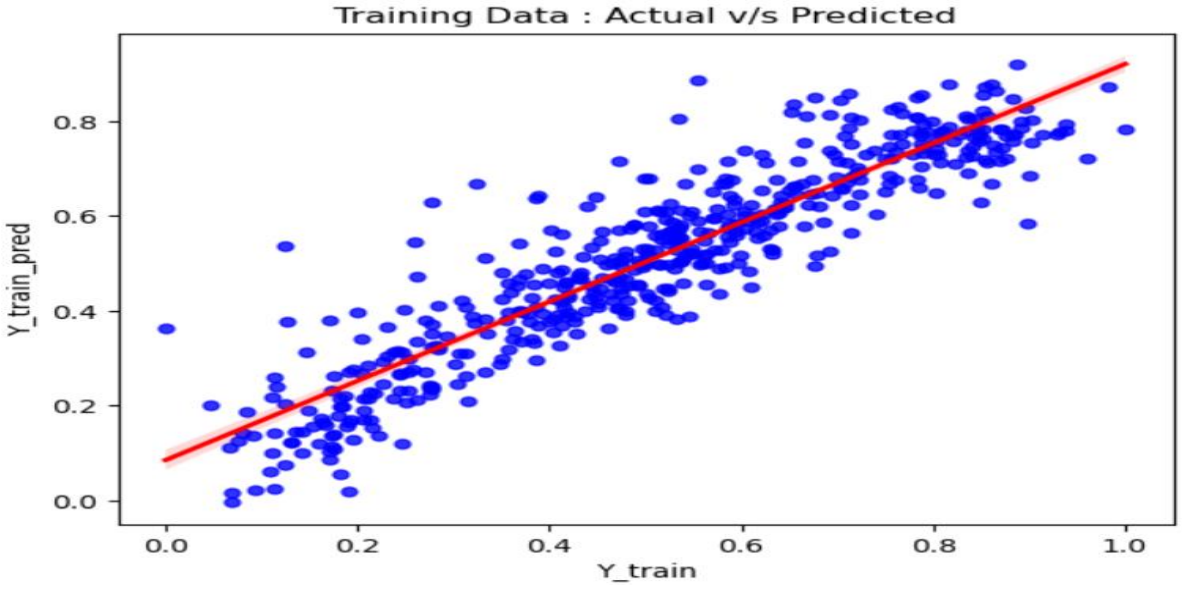
a) **Assumption of Normality: [Error terms are normally distributed with mean = 0]**

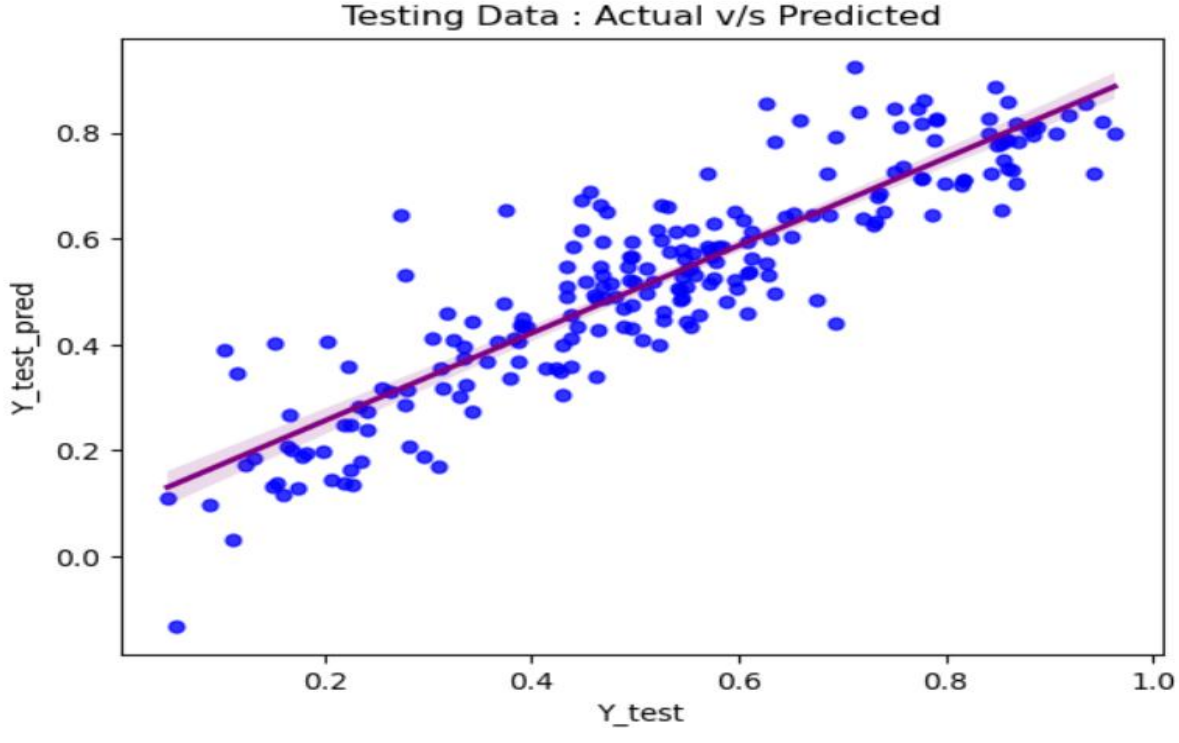
Computing Residuals (Y_pred – Y_actual) and plotting a histogram of residual or error terms . The plot shows a normalized distribution of error terms with mean = 0.



b) **Assumption of Linearity – There is Linear relationship between X and Y to fit the best linear model**

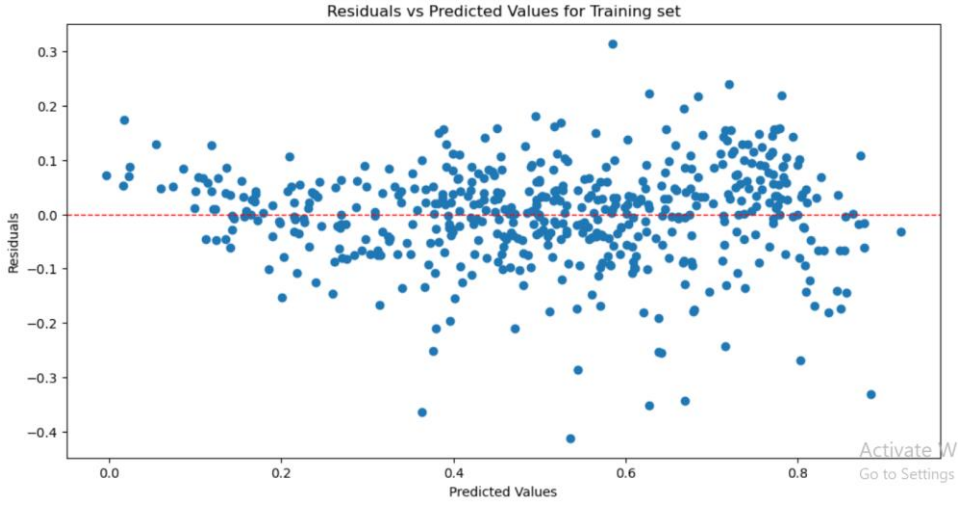
Used regplot to visualize the relationship between Predicted values and Actual values and to visualize the linear regression fit of the model on the training and testing data. The plot shows a linear relationship, and the line fits well along the data.

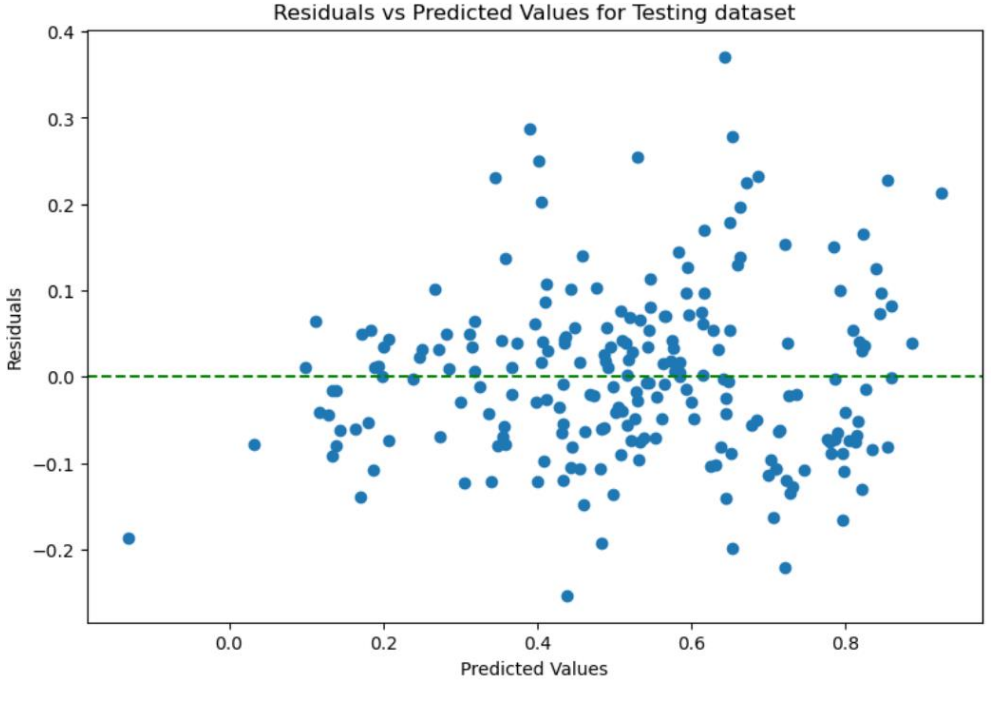




c) **Assumption of Independence of Errors and Homoscedasticity: The error terms are independent of each other and have constant variance**

Used scatter plot to visualize the relationship between residuals vs. predicted values. The plot does not any show any specific pattern, it is random. This explains Homoscedasticity ie, error terms have constant variance. should also show no patterns; a random scatter indicates linearity. No funnel shape.





5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

The multiple linear regression formula is :

cnt = 0.08 + (0.55 * temp) + (0.23 * yr) + (0.13 * Winter) + (0.10 * Sep) + (0.09 * Summer) + (0.07 * Saturday) + (0.06 * workingday) - (0.08 * Misty) - (0.16 * windspeed) - (0.29 * Light_Rain_Snow)

Based on final model, top 3 features are 'temp', 'yr' and 'winter'. It indicates that as these values increases, there will be significant increases in demand for shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

The Linear regression algorithm has the following steps

Step 1- Importing the required libraries – numpy, pandas, sklearn, statsapi, matlabplot, seaborn etc.

Step2 – Reading, Understanding and Visualizing the Data.

- Reading the data to a dataframe
- Looking at the data size, datatypes, statistical values (min, max etc.)
- Understanding the data dictionary

Step3 – Exploratory Data Analysis

- Visualize the data – plotting pair plots, heatmaps, boxplots for categorical variables and make some analysis from the data

Step 4- Preparation of Data

- Drop columns which are not necessary for analysis
- Handling missing or NULL values if any by filling or imputing
- Creating dummy variables for variables that have more than 2 categories
- Split the data into training (70%) and test set(30%) (df_train, df_test)
- Rescale the numerical variables using either MinMaxScaler (values in range 0-1) or Standardization Scaler (normalizing data with mean = 0, std deviation =1)

Step5 – Building the Model

Create X_train (which has all the independent variables) and Y_train(which has the target variable)

- Perform Linear Regression
    - You can follow three methods here:
        - Add variables one by one at a time to build the model until you get best model parameters (low p_value < 0.05, Low VIF < 5)
        - Add all variables -> Build the model -> Keep removing one variable at a time till the model has best parameters (low p_value < 0.05, Low VIF < 5)
        - Use RFE (Recursive feature elimination) which is an automated approach:

- Build the model -> Remove features one at a time until you get the best model parameters (low p_value < 0.05, Low VIF < 5)
- End when you have a final model decided

Step 6 – Residual Analysis

- Calculate the Residual or Error Terms for training set (Residual = Predicted – Actual)
- Visualize the data to evaluate the model
  - Linearity - Linear relationship between X and Y and best Line fit to represent the model
  - Normalized distribution for error values
- Check R-squared and Adjusted R-squared value if they are good (r-squared > 80% is good)
- Check Prob (F-statistic)
- Check mean square error on training set

Step 7– Predict and Evaluate the Model on Testing dataset

- Apply scaling on test sets (same scaling which was applied on training set)
- Predict test data using the model
- Evaluate the model
  - Calculate R2_score for Training and Test Dataset (Difference should be minimal)
  - Calculate mean square error for test data (should be close to)
- Visualize the Predicted test data and validate assumptions of regression
- Formulate the equation for the multiple regression as

Y= constant + coefficient_1* X1+ coefficient_2* X2 + ……. + coefficient_n* Xn

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet is a set of 4 small datasets, created by the statistician Francis Anscombe in 1973.

Features of Anscombe's quartet:

- Datasets with identical simple descriptive statistics (mean, variance, correlation)
  - Mean of x =9, Mean of y =7.5, Correlation = 0.816
- They have very different distributions and graphical representations
- The datasets have identical Linear regression properties ie, Same regression line when fitted with linear model
- Some peculiarities in the dataset fools the regression model if built.
- It is importance to visualize the data before drawing any conclusions from statistical analysis. Visualizing the data helps view the data patterns, trends, and outliers.

- Dataset (plotted with scatter plot):
  - **Dataset 1**: Shows linear relationship resembles a typical linear trend. This **fits** the linear regression model pretty well.

- **Dataset 2**: Shows a linear trend but with some points forming a curve. Does not fit the regression model well.
- **Dataset 3**: Shows a linear trend with an outlier that significantly affects the correlation. Outliers cannot be handled by Regression model
- **Dataset 4**: Shows non-linear relationship with a clear curve. Cannot be handled by Regression model

3. What is Pearson's R? (3 marks)

Pearson's R is known as the Pearson correlation coefficient. It is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables.

It ranges from -1 to 1.

- R=1: Indicates positive correlation
  - It means, when one variable increases, the other variable also increases.
    - 0<R<0.3: Weak positive correlation
    - 0.3<R<0.7: Moderate positive correlation
    - 0.7<R<1: Strong positive correlation
- R=−1: Indicates negative correlation
  - It means as one variable increases, the other variable decreases.
    - -0.3<R<0: Weak negative correlation
    - -0.7<R<-0.3: Moderate negative correlation
    - -1<R<-0.7: Strong negati correlation
- R=0: No linear correlation
  - It means the variables do not have a linear relationship

$$R = \sum (X_i - X') * (Y_i - Y') / SQRT (\sum (X_i - X')^2 * (Y_i - Y')^2)$$

X_i, $Y_{i\ are}$ sample data points

X', Y' are means of X and Y

Disadvantages of Pearson's R:

- Pearson's R captures only linear relationship
- It Does not explain causation. They do not explain variance of one variable wrt other

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is a data pre-processing method applied on independent variables (features) helps to represent the data distribution within a range. Scaling only affects the coefficients. It does not affect the statistical parameters like F-statistic, p-values, R-squared etc.

Scaling is needed due to these reasons

- **Handling Units of Data** – The multiple independent features can have varying values and units. If scaling is not done, the units are not considered by the algorithm and hence the model can be very inaccurate.
- **Handling Large Values of Data** – Some features can have high values which tend to influence the model, this can affect building a good model. When scaled, it helps to build an accurate and unbiased model.
- **Have an Optimized Algorithm** - The performance of the algorithm improves when data is scaled as compared to otherwise. This can also help the model converge faster.

Two types of Scaling used and their differences:

Normalized Scaling OR Min-Max Scaling

- Scales the feature to a fixed range between 0 and 1
- **Min Max Scaling: $X' = (X- X_{min}) / (X_{max}-X_{min})$**
- This scaling method has a disadvantage that it loses some information of data if there are outliers.

Standardized Scaling

- Scales the data into a standard normal distribution with mean = 0 and standard deviation =1.
- **Standardization Scaling: X' = (X- mean(X)) / std_dev(X)**

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

VIF is known as Variance Inflation Factor (VIF) which is a measure of multicollinearity between independent variables in a regression model.

VIF = infinite indicates that this variable has a perfect multicollinearity or correlation with multiple independent variables in a regression model.

In such cases, the model cannot determine the accurate value of coefficients.

To solve this problem, you have to either remove some correlated variable or combine some correlated variables in the data.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A Q-Q plot is also called a quantile-quantile plot. It is a graphical tool that helps to assess the distribution of the dataset and get the distribution parameters. It can compare two probability distributions by plotting their quantiles against each other. This helps linear regression where we have training and test data set received separately and Q-Q plot helps to confirm that both the data sets are from populations with same distributions.

It is used to check the following:

whether two data sets

i. come from populations with a common distribution

ii. they have common location and scale

iii. they have similar distributions

v. they have similar tail behavior

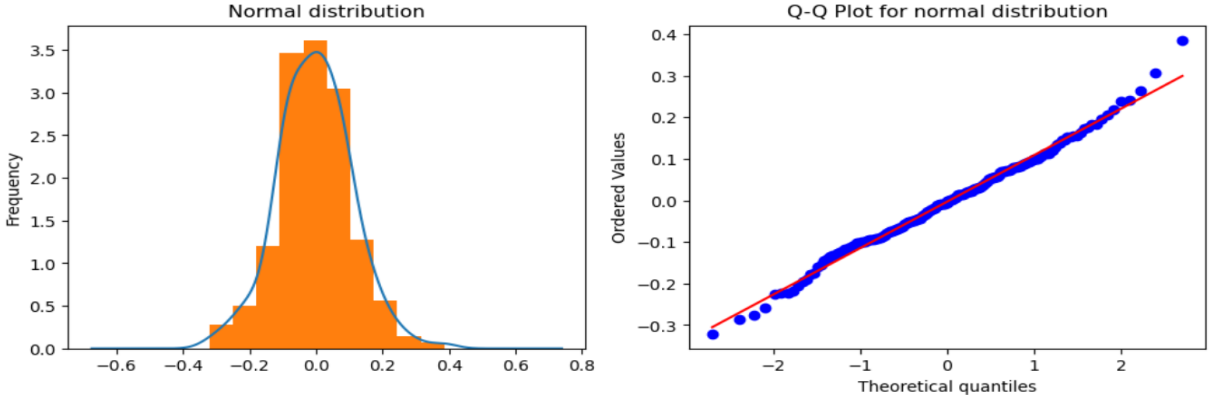
Advantages:

a) It can be used with any sample size

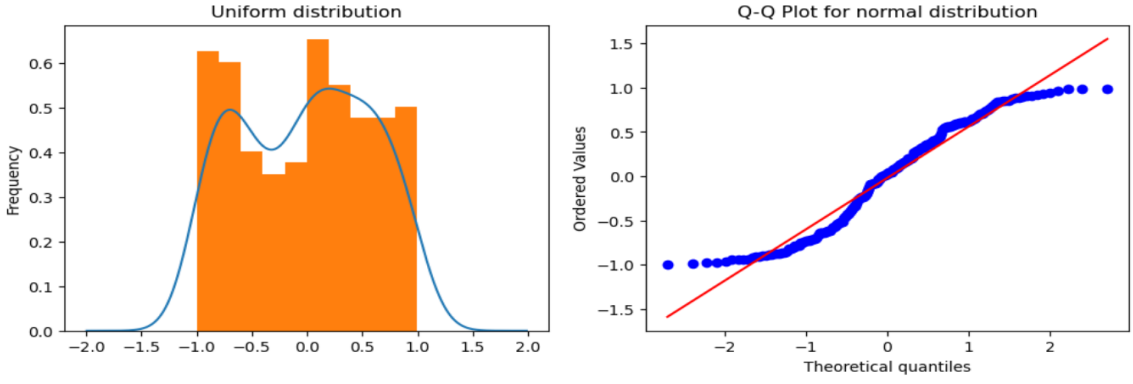
b) This plot gives information on the distributional aspects -shifts in location, shifts in scale, changes in symmetry, and presence of outliers

Example of Q-Q plots for various distributions

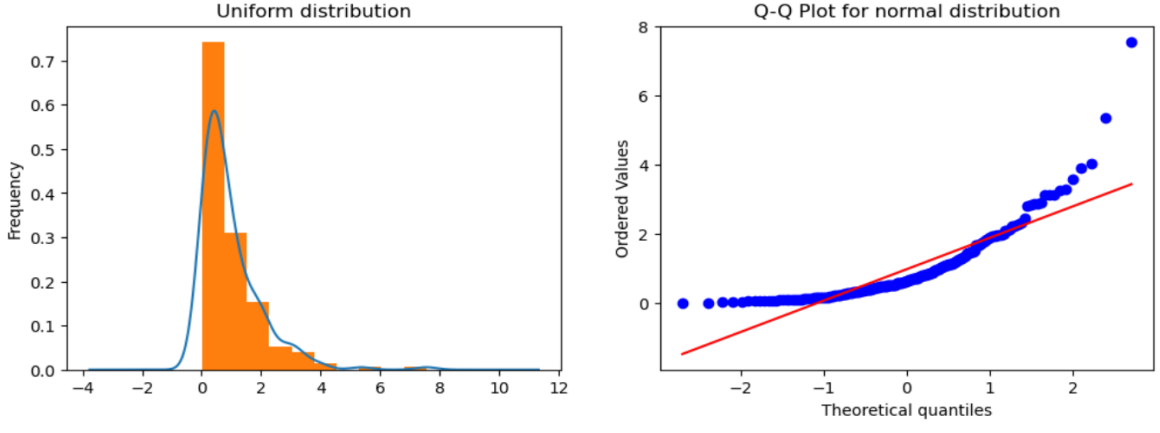
1. Normal distribution:



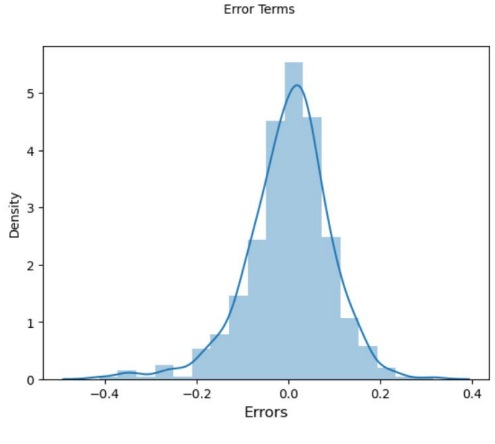
2. Uniform Distribution:



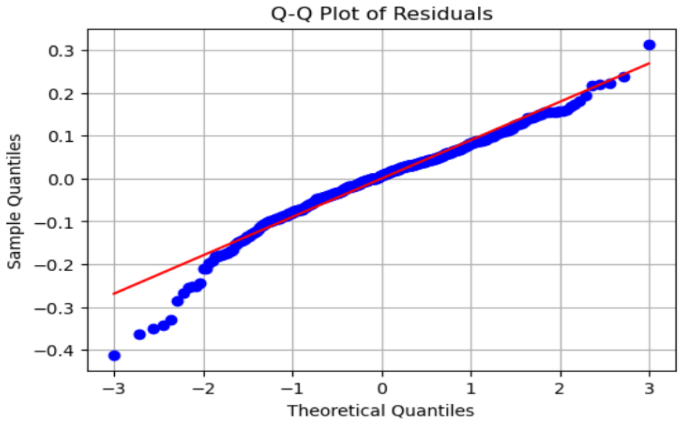
3,Exponential Distribution



In the Boom Bikes Bike sharing problem, after obtaining residual = Y_pred – Y_actual, the plot shows a normal distribution as shown below:



A Q-Q Plot was plotted for the same residual above:



The above plot explains the normal distribution of the residual data.