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)**

   Answer-The categorical variables in the dataset were season,mnth,holiday,weathersit,yr,weekday,workingday. I analyzed them using Boxplot.

   a. **Season**- Among all the seasons the ride count was minimum for Spring and maximum for Fall season.

   b. **Mnth**- September month had maximum number of rides whereas January had minimum number of rides.

   c. **Holiday**- Average number of rides were more when it was not holiday.

   d. **Weathersit**- Clear, Few clouds, Partly cloudy, Partly cloudy weather situation saw maximum number of rides whereas no rides were done in Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog weather situation.

   e. **Weekday**- Sunday saw minimum number of rides.

   f. **Workingday**- Workingday has slightly more rides than non-workingday.

   g. **Yr**- Year 2019 saw significantly higher number of rides compared to 2018, so it is a growing business.

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

   Answer- drop_first=True helps in reducing correlation among the dummy variables by removing the one extra column which is obvious. For example if we have a categorical variable which takes two values--1 or 0. If the value is not 1 then it is obviously 0 ,so we do not need the column for 0. Hence it can be dropped safely. If we don`t drop the first column then dummy variables will be correlated and it can affect the model performance and the feature importance can`t be computed correlectly.

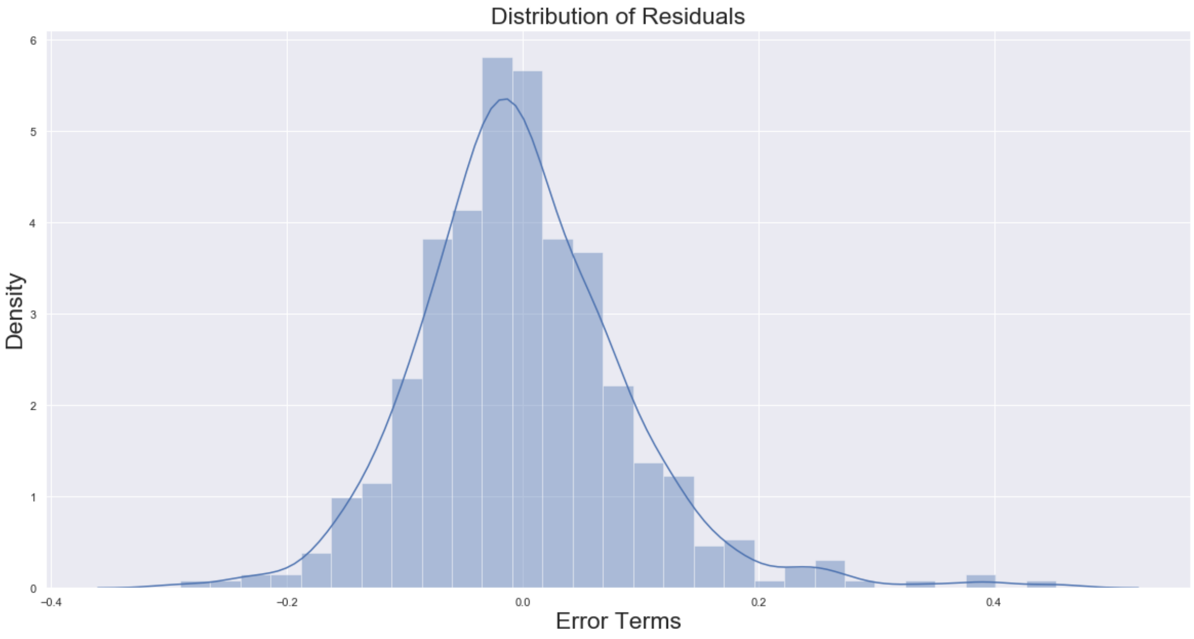
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

   Answer- ‘temp’ and ‘atemp’ have highest correlation with target variable(‘cnt’)

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

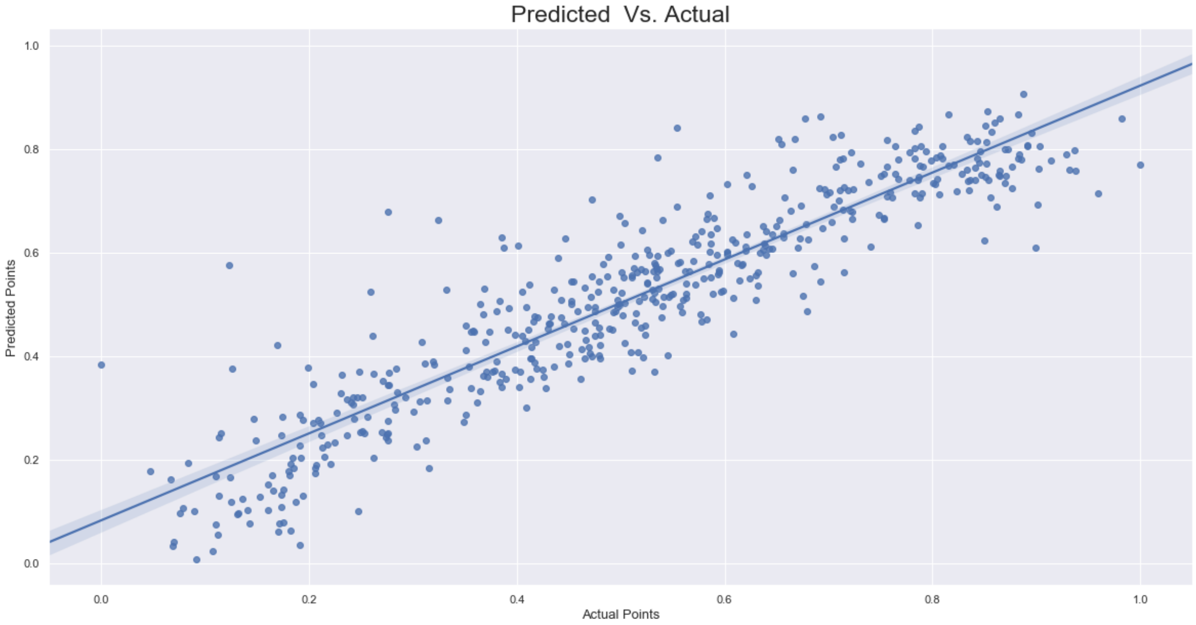
   Answer-

   a. **Error terms are normally distributed** – I plotted distribution plot for Residuals or Error terms and it was found to follow normal distribution



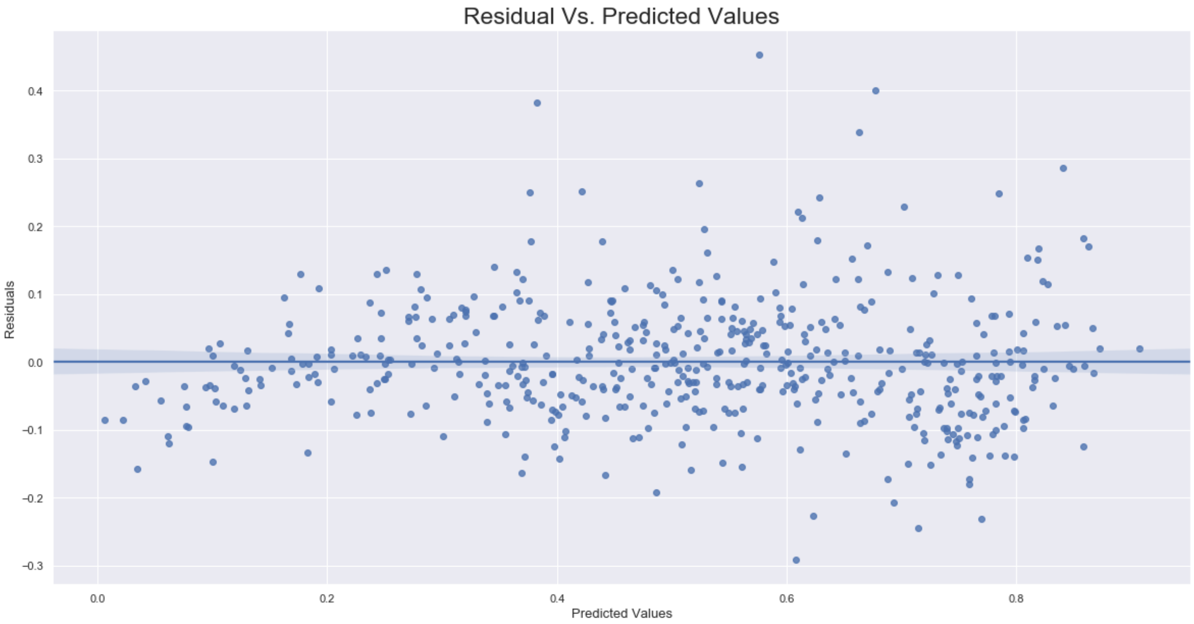
Distribution of Residuals

b. **Error terms have constant variance(Homoscedasticity) –** I plotted regression plot between predicted and actual values and it was found to follow constant variance.



Predicted Vs. Actual

c. **Error terms are independent-** I plotted regression plot between Residuals vs. Predicted values and they were found to be independent.



d. **Multicollinearity-** There should not be significant multicollinearity among variables.

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

Answer- The equation of best fitted line for our model was—

cnt = 0.211 + **yr** x 0.234 + **temp** x 0.474 - **windspeed** x 0.169 + **season_Summer** x 0.057 + **season_Winter** x 0.121 - **mnth_Dec** x 0.058 - **mnth_Feb** x 0.059 - **mnth_Jan** x 0.089 - **mnth_Jul** x 0.035 - **mnth_Nov** x 0.050 + **mnth_Sep** x 0.074 - **weekday_Sun** x 0.047 - **weathersit_Light_SnowRainThunder** x 0.293 - **weathersit_Mist** x 0.081

The top 3 features contributing significantly are—

a. **Temp**- Coefficient = 0.474
b. **Yr**-Coefficient = 0.234
c. **Weathersit_Light_SnowRainThunder**- Coefficient= -0.293

General Subjective Questions

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

Answer-Linear Regression is defined as a statistical model that analyzes linear relationship between dependent and one/many independent variables. Mathematically it is represented by "**y =mX + c**". Here we have

a. **independent variable(X)**

b. **dependent variable(y)**

c. **Y-intercept (c)** i.e value of y when X=0.

d. **m** is the slope of the regression line.

Linear Regression is of following types—

a. **<u>Simple Linear Regression-</u>** Dependent variable is predicted using only one dependent variable. Formula

$$Y = \beta_0 + \beta_1 X$$

b. **<u>Multiple Linear Regression-</u>** Dependent variable is predicted using multiple independent variables. Formula

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

The beta values are the co-efficient of the independent variables.

Assumptions of Linear Regression—

a. Linear relationship between X and Y.

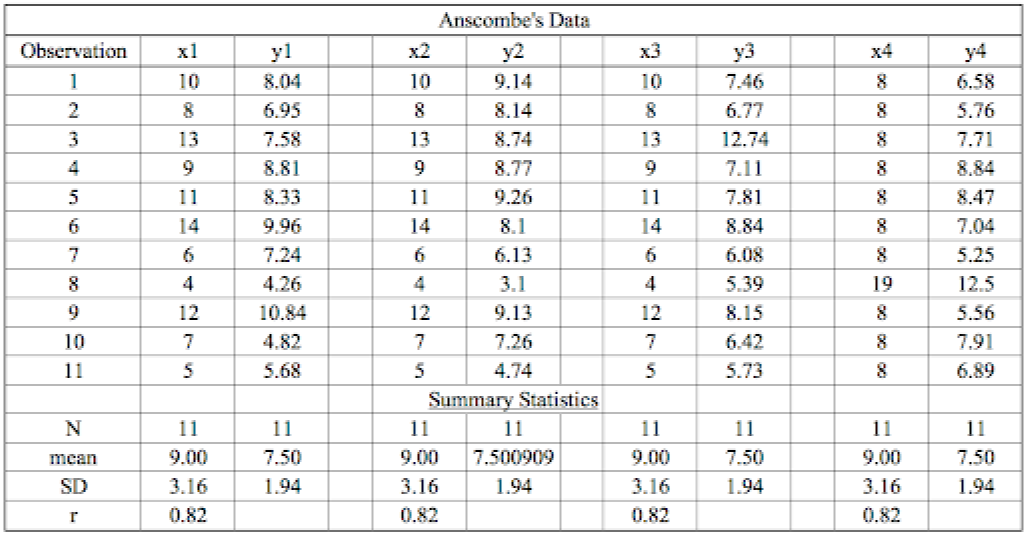
b. Error terms are normally distributed.

c. Error terms are independent of each other.

d. Error terms have constant variance (homoscedasticity).

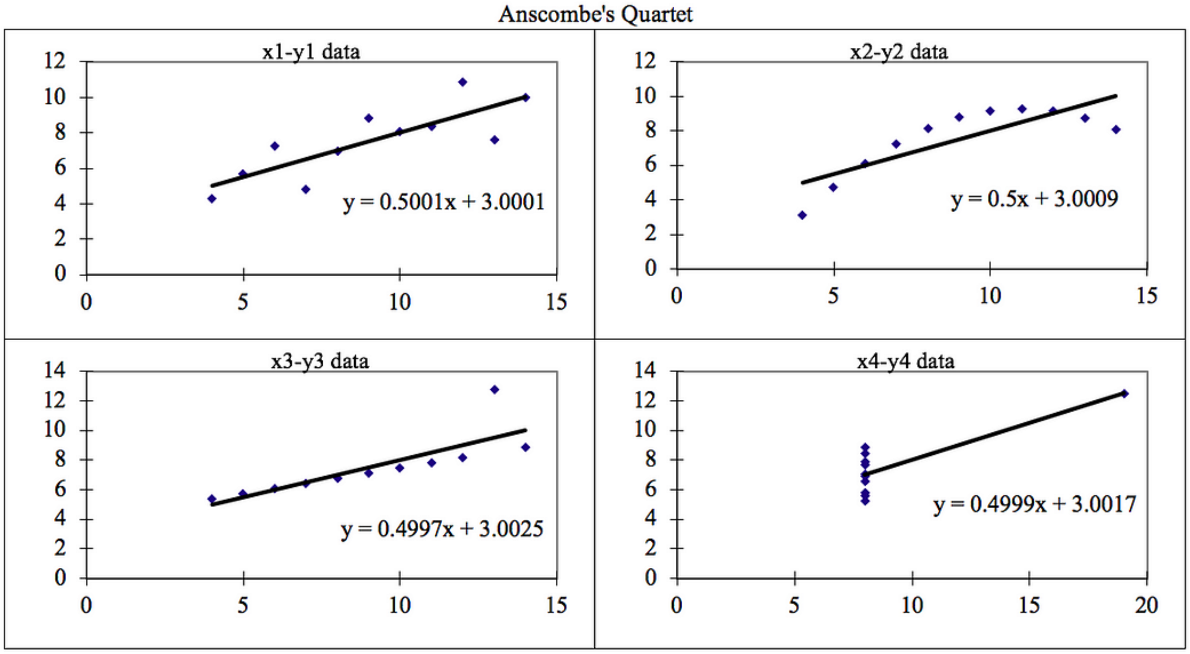
2. Explain the Anscombe's quartet in detail. (3 marks)

Answer- Anscombe's quartet was developed by Francis Anscombe to illustrate the importance of plotting data before analyzing it. It also stresses on the importance of identifying anomalies in the data (for example-outliers), linear separability etc. in the data. These four datasets have clearly identical simple descriptive statistics, however they look very different from each other when plotted on the graph.



					Anscombe's Data						
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
					Summary Statistics						
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

When these are plotted on scatter plot , they generate different kind of plot which is not interpretable by any regression algorithm.



- **Data Set 1:** fits the linear regression model pretty well.
- **Data Set 2:** cannot fit the linear regression model because the data is non-linear.
- **Data Set 3:** shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- **Data Set 4:** shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

We can consider linear relationship between data only after plotting and looking the data.

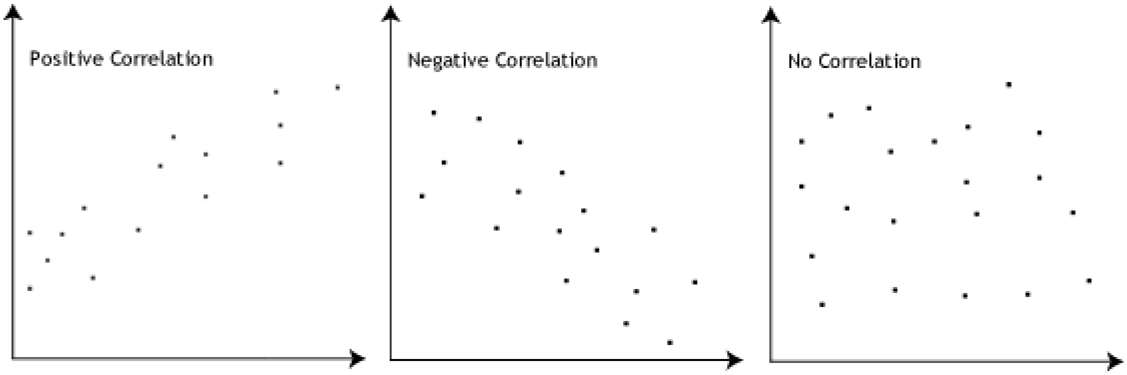
3. What is Pearson's R? (3 marks)

Answer- Pearson`s R is a measure of strength of linear association between variables. It takes a value from -1 to +1.

a. **Positive value –** A positive value indicates that increase in value of one variable causes an increase in the value of another variable. Value of 1 indicates perfect positive relationship.

b. **Negative value -** A negative value indicates that increase in value of one variable causes a decrease in the value of another variable. Value of -1 indicates perfect negative relationship.

c. **Zero value-**There is no relationship between two variables.



Positive Correlation

Negative Correlation

No Correlation

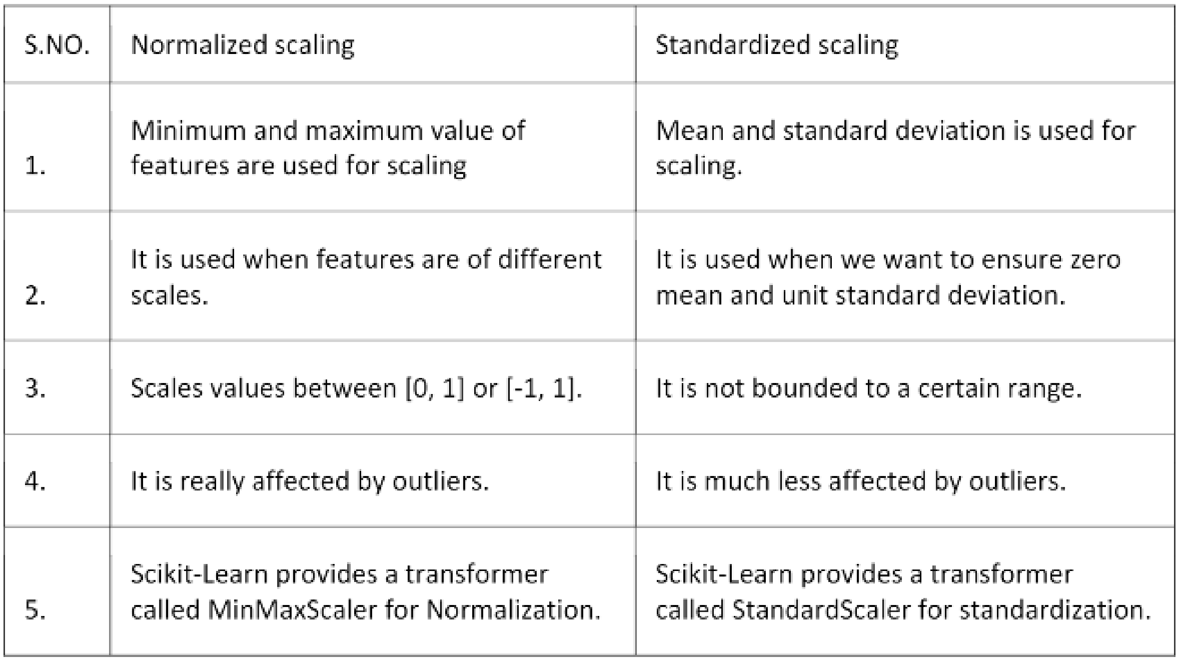
4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**

Answer- When we have many independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. For example- A feature using grams as unit and having value 1000 grams might be considered to be more impactful than another feature using Kg as unit and having value as 2 Kg but actually it is not. So we need to scale features because of two reasons:

a. Ease of interpretation

b. Faster convergence for gradient descent methods .

The below two types are the popular choice for Feature Scaling-



S.NO.	Normalized scaling	Standardized scaling
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

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

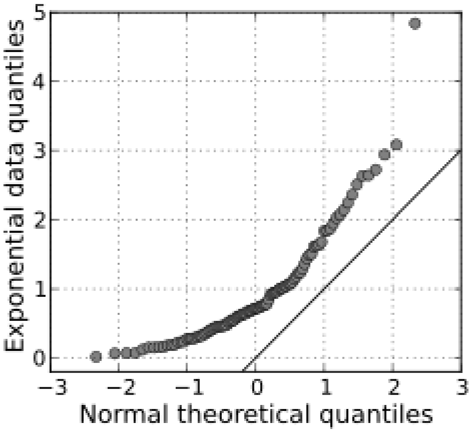
Answer- **$VIF_i=1/(1-R_i^2)$**

If there is perfect correlation, then VIF = infinity. It shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R^2 =1, which leads to $1/(1-R^2)$ = infinity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Answer- Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, 0.4 quantile is the point at which 40% of the data fall below and 60% of the data fall above that value. A 45-degree reference line is plotted on the Q Q plot. If the two data sets come from a common distribution, the points will fall along that reference line.



The **importance** and purpose of Q Q plots is to find out if two sets of data come from the same distribution. If two samples differ, it also helps in understanding the differences. The Q-Q plot helps in better understanding the nature of differences. It is also used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.