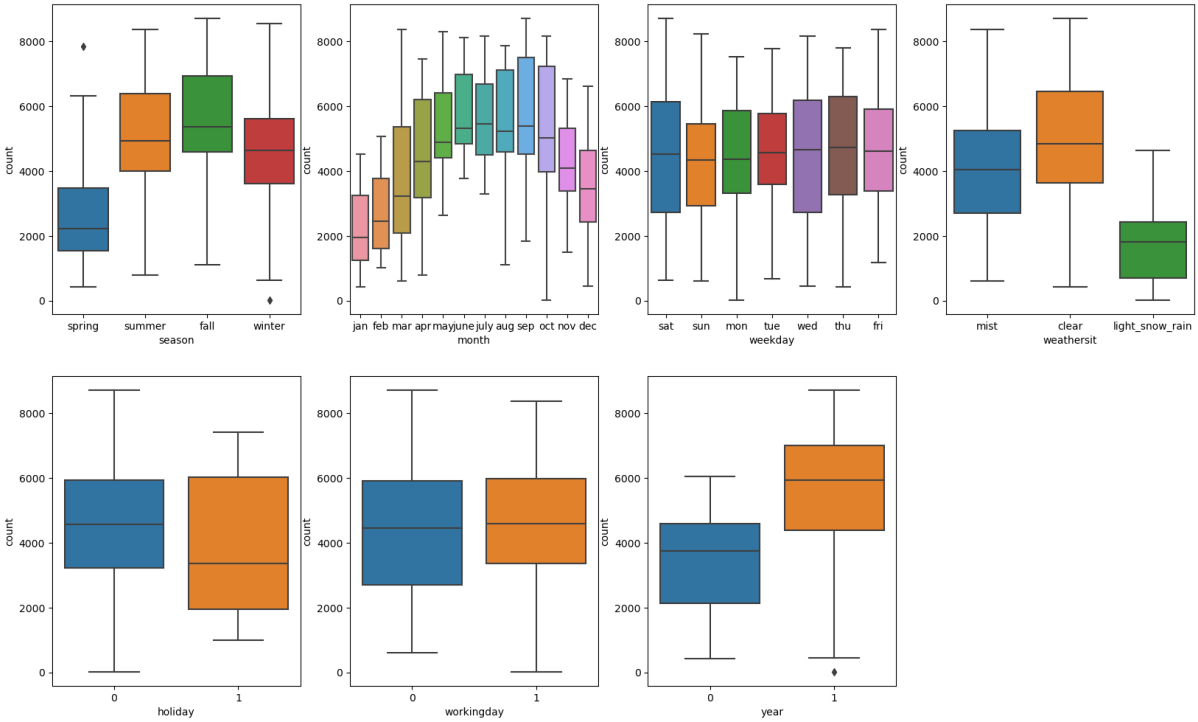
Assignment-based Subjective Questions

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: I have done analysis on categorical columns using the boxplot. Below are the few points we can infer from the visualization.



- Fall season seems to have attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019.
- Most of the bookings has been done during the month of May, June, July, Aug, Sep and Oct. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.
- Clear weather attracted more booking which seems obvious.
- Thu, Fri, Sat and Sun have a greater number of bookings as compared to the start of the week.
- When it's not holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.
- 2019 attracted a greater number of bookings from the previous year, which shows good progress in terms of business.
- Not able to infer anything from weekday data.

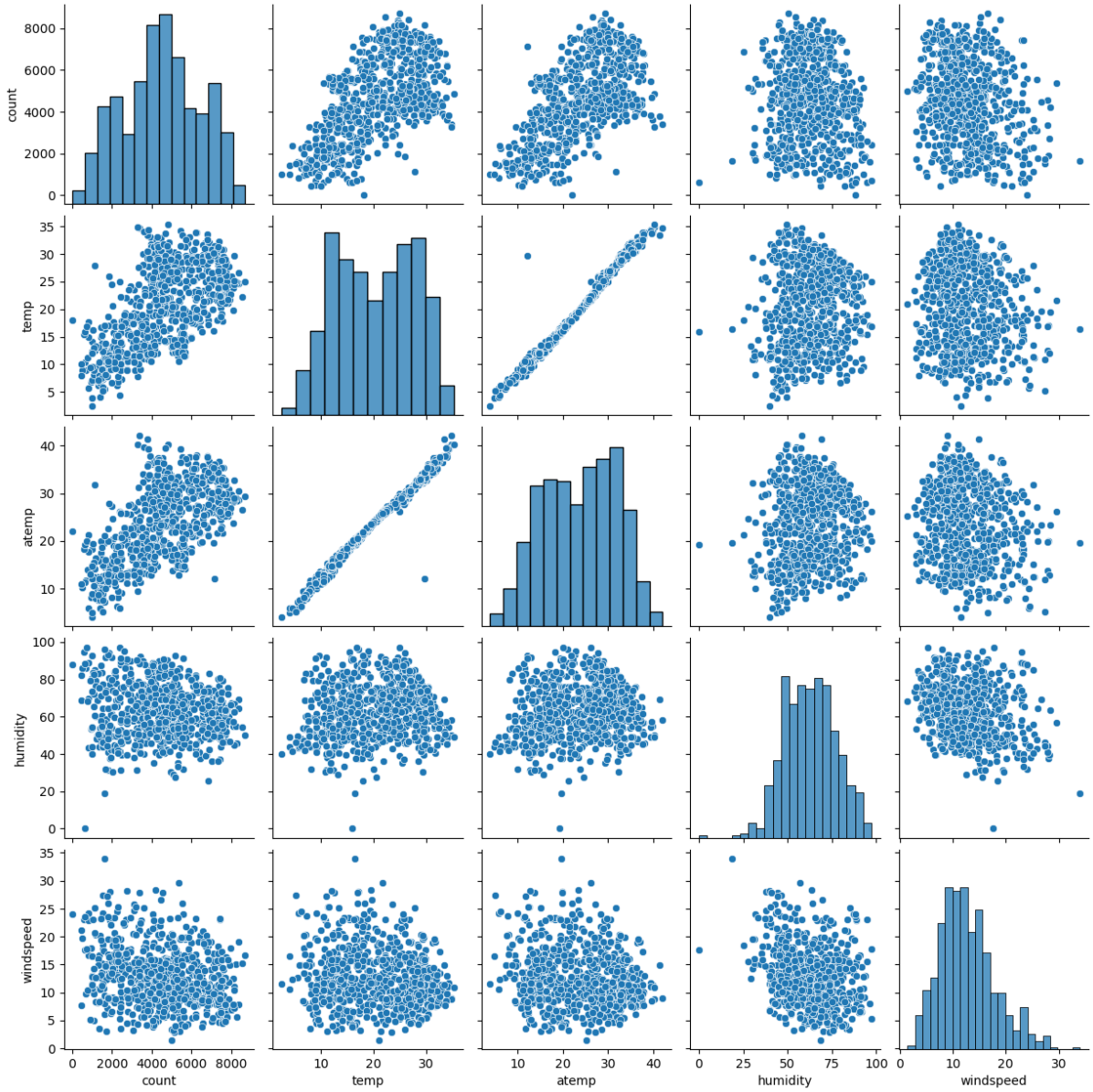
2. Why is it important to use drop_first=True during dummy variable creation? (2 marks)

Answer: If you set **drop_first = True**, then it will drop the first category. So, if you have K categories, it will only produce K – 1 dummy variables. Suppose you have a column for gender that contains 4

variables- "Male", "Female", "Other", "Unknown". So, a person is either "Male", or "Female", or "Other". If they are not either of these 3, their gender is "Unknown".

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

Answer:



The 'temp' and 'atemp' variables have highest correlation when compared to the rest with target variable as 'count'.

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

Answer: I have validated the assumption of Linear Regression Model based on below 5 assumptions.

- Normality of error terms – "Error terms are normally distributed".
- Multicollinearity check – "There are insignificant multicollinearity among variables".
- Linear relationship validation – "Linearity are visible among variables".

- Homoscedasticity – “There are no visible pattern in residual values”.
- Independence of residuals – “No autocorrelation”.

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: Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

- Year
- Holiday
- Temperature

General Subjective Questions

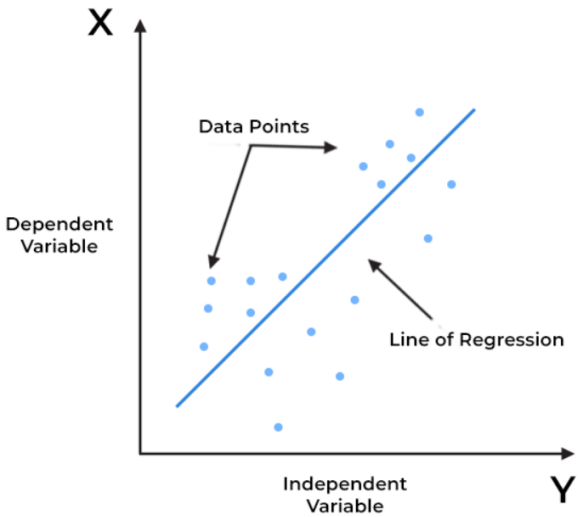
1. Explain the linear regression algorithm in detail. (4 marks)

Answer: Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis.

The independent variable is also the predictor or explanatory variable that remains unchanged due to the change in other variables. However, the dependent variable changes with fluctuations in the independent variable. The regression model predicts the value of the dependent variable, which is the response or outcome variable being analysed or studied.

Thus, linear regression is a supervised learning algorithm that simulates a mathematical relationship between variables and makes predictions for continuous or numeric variables such as sales, salary, age, product price, etc.

This analysis method is advantageous when at least two variables are available in the data, as observed in stock market forecasting, portfolio management, scientific analysis, etc.
A sloped straight line represents the linear regression model.



In the above figure,

X-axis = Independent variable

Y-axis = Output / dependent variable

Line of regression = Best fit line for a model

Here, a line is plotted for the given data points that suitably fit all the issues. Hence, it is called the 'best fit line.' The goal of the linear regression algorithm is to find this best fit line seen in the above figure.

Linear regression can be expressed mathematically as:

$y= \beta 0+ \beta\ 1x+ \varepsilon$

Here,

Y= Dependent Variable
X= Independent Variable
β 0= intercept of the line
β1 = Linear regression coefficient (slope of the line)
ε = random error
The last parameter, random error ε, is required as the best fit line also doesn't include the data points perfectly.

Types of Linear Regression
Linear Regression can be broadly classified into two types of algorithms:

1. Simple Linear Regression
A simple straight-line equation involving slope (dy/dx) and intercept (an integer/continuous value) is utilized in simple Linear Regression. Here a simple form is:

y=mx+c where y denotes the output x is the independent variable, and c is the intercept when x=0. With this equation, the algorithm trains the model of machine learning and gives the most accurate output

2. Multiple Linear Regression
When a number of independent variables more than one, the governing linear equation applicable to regression takes a different form like:

y= c+m1x1+m2x2… mnxn where represents the coefficient responsible for impact of different independent variables x1, x2 etc. This machine learning algorithm, when applied, finds the values of coefficients m1, m2, etc., and gives the best fitting line.

Assumptions of Linear Regression
Normally most statistical tests and results rely upon some specific assumptions regarding the variables involved. Naturally, if these assumptions are not considered, the results will not be reliable. Linear Regression also comes under same consideration. There are some common assumptions to be considered while using Linear Regression:

Linearity: The models of Linear Regression models must be linear in the sense that the output must have a linear association with the input values, and it only suits data that has a linear relationship between the two entities.

Homoscedasticity: Homoscedasticity means the standard deviation and the variance of the residuals (difference of (y-yˆ)2) must be the same for any value of x. Multiple Linear Regression assumes that the amount of error in the residuals is similar at each point of the linear model. We can check the Homoscedasticity using Scatter plots.

Non-multicollinearity: The data should not have multicollinearity, which means the independent variables should not be highly correlated with each other. If this occurs, it will be difficult to identify those specific variables which contribute to the variance in the dependent variable. We can check the data for this using a correlation matrix.

No Autocorrelation: When data are obtained across time, we assume that successive values of the disturbance component are momentarily independent in the conventional Linear Regression model. When this assumption is not followed, the situation is referred to be autocorrelation.

Not applicable to Outliers: The value of the dependent variable cannot be estimated for a value of an independent variable which lies outside the range of values in the sample data.

All the above assumptions are critical because if they are not followed, they can lead to drawing conclusions that may become invalid and unreliable.

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

Answer: Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyse it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.

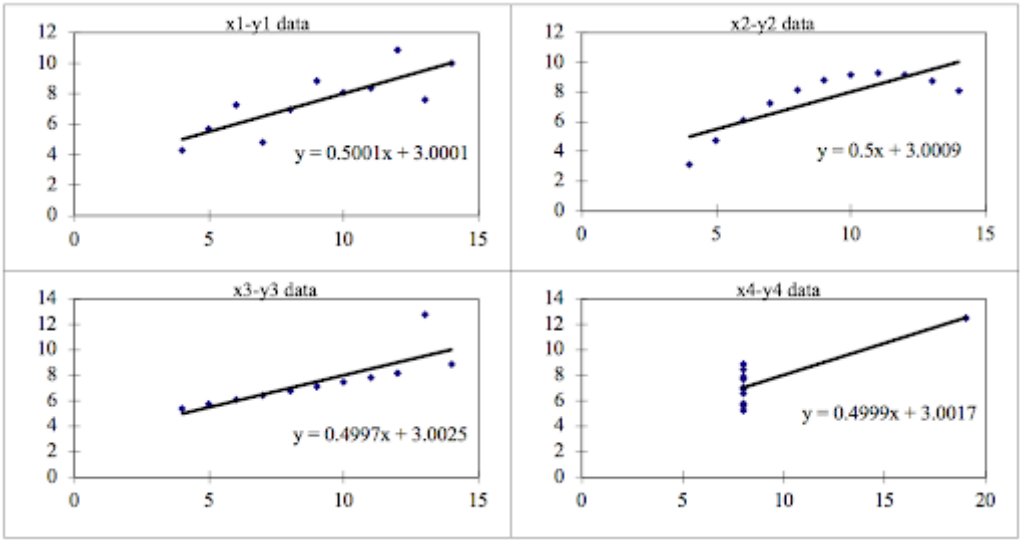
Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.
We can define these four plots as follows:

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

The statistical information for these four data sets is approximately similar. We can compute them as follows:

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	

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:



We can describe the four data sets as:

ANSCOMBE'S QUARTET FOUR DATASETS

- 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.

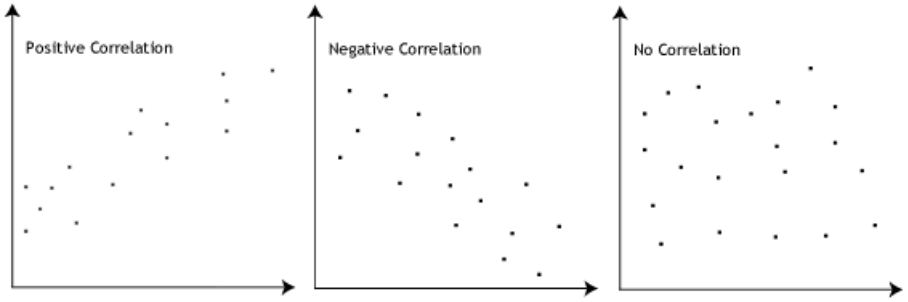
As we can see, Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data

or implement any machine learning algorithm, we first need to visualize the data set to help build a well-fit model.

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

Answer: Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be. positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



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

Answer: It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done, then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh. greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Example: If an algorithm is not using feature scaling method, then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus,

tackle this issue.

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Minimum and maximum value of features are used for scaling. It is used when features are of different scales. Scales values between [0, 1] or [-1, 1]. It is really affected by outliers.

Standardization Scaling:

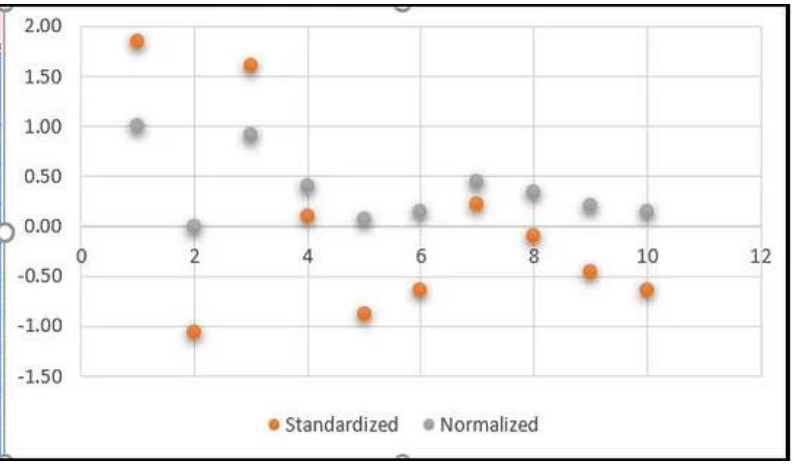
Standardization replaces the values by their Z scores. It brings all the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

sklearn.preprocessing.scale helps to implement standardization in python. One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers. Mean and standard deviation is used for scaling. It is used when we want to ensure zero. It is not bounded to a certain range.

Below shows example of Standardized and Normalized scaling on original values.

Original	Standardized	Normalized
59	1.85	1.000
11	-1.06	0.000
55	1.61	0.917
30	0.09	0.396
14	-0.88	0.063
18	-0.64	0.146
32	0.21	0.438
27	-0.09	0.333
21	-0.46	0.208
18	-0.64	0.146



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

Answer: If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

When the value of VIF is infinite it shows a perfect correlation between two independents variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. 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: The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.