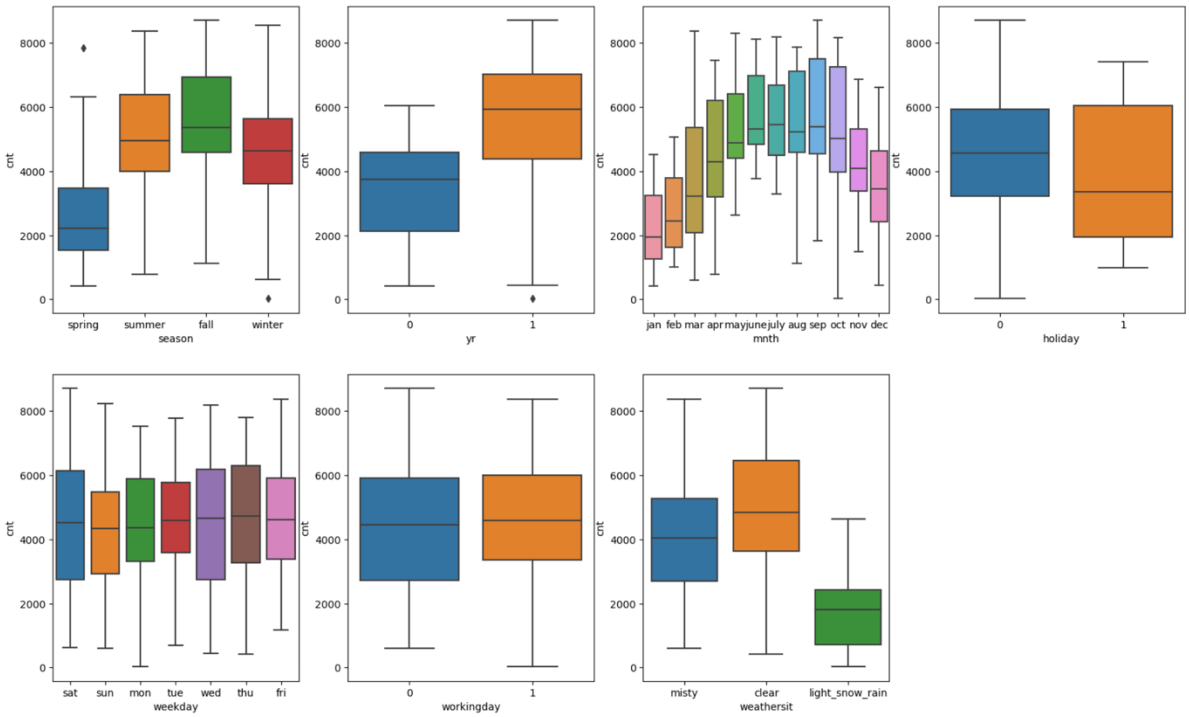
Assignment-based Subjective Questions

Question 1:
From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer:



Analysis as per variable is as below:

Season:

Business during fall is higher compared to other seasons while during spring is lowest

Year:

2019 year which is represented by 1 has more business than 2018 which is represented by 0, which means business has increased in a year which is sync with growing demand of the bike sharing

Month:

Most of the bookings has been done during the month of may, june, july, aug, sep and oct which is obvious at is summer and fall time in USA

holiday:

During holiday the usage of bike sharing seems more and is obvious as usually on holidays people go out for spending time and ma tend to use ride share more than during weekdays and it seems people are preferring less bike sharing during weekdays (office travel)

Month:

Business seems similar during weekdays but on Thu, Fir, Sat and Sun seems to have more number of bookings as compared to the start of the week.

Working Day:

Booking seemed to be almost equal either on working day or non-working day.

weathersit :

Clear weather attracted more booking which is obvious.
Lesser bookings during the light snow or rain
No data for the high rain or snow.

Question 2:
Why is it important to use drop_first=True during dummy variable creation?

Answer:

We use drop_first=True when creating dummy variables to avoid the "dummy variable trap" which refers to a situation where two or more dummy variables are highly correlated and can cause problems in regression analysis due to multicollinearity.

Without drop_first=True when you create the dummy variables for categorical feature with n values, it would generate n dummy variables (coumns).
With drop_first=True you will get n-1 dummy variables (columns).

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

Answer:

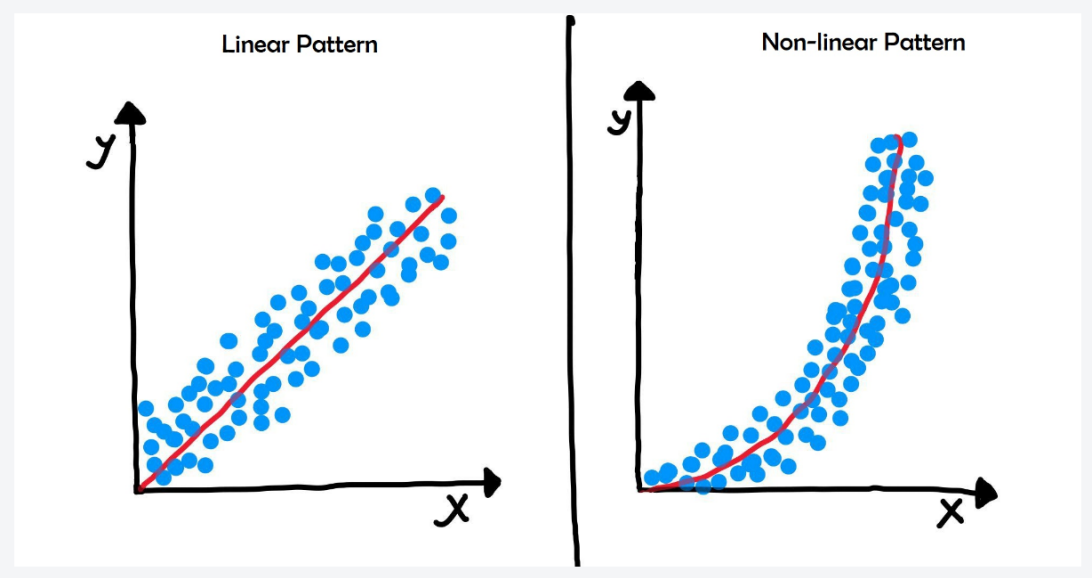
temp & atemp variables are having the highest correlation with target variable cnt.

Question 4:
How did you validate the assumptions of Linear Regression after building the model on the training set?

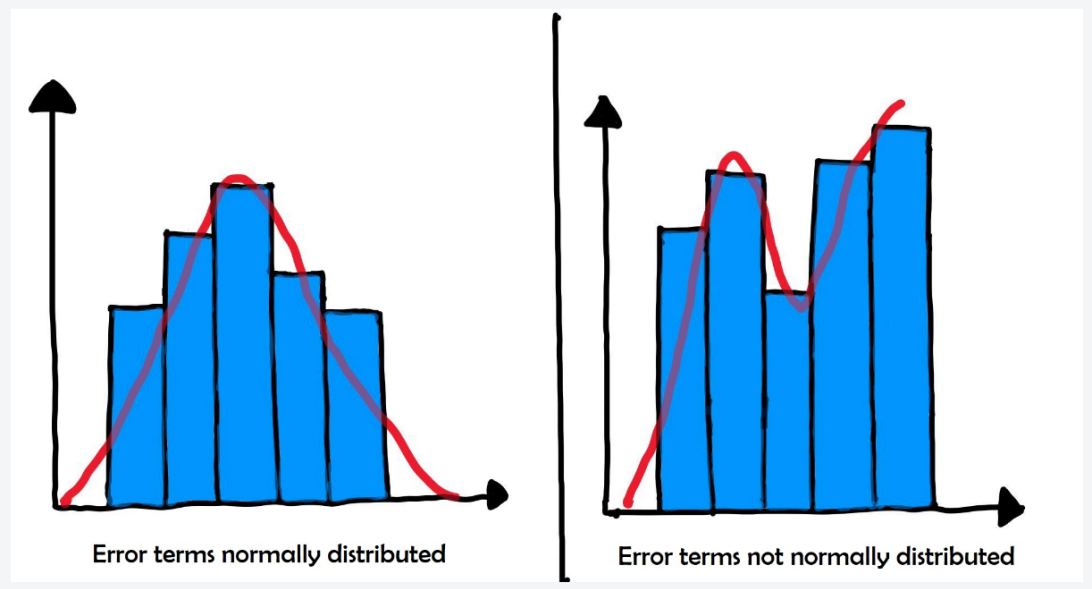
Answer:

After building a Linear Regression model on the training set, you can validate its assumptions to ensure that the model's underlying assumptions are met.

Linearity:
Check if the relationship between each independent variable and the dependent variable is approximately linear.
We can use scatter plots of observed vs. predicted values or residuals to assess linearity visually.

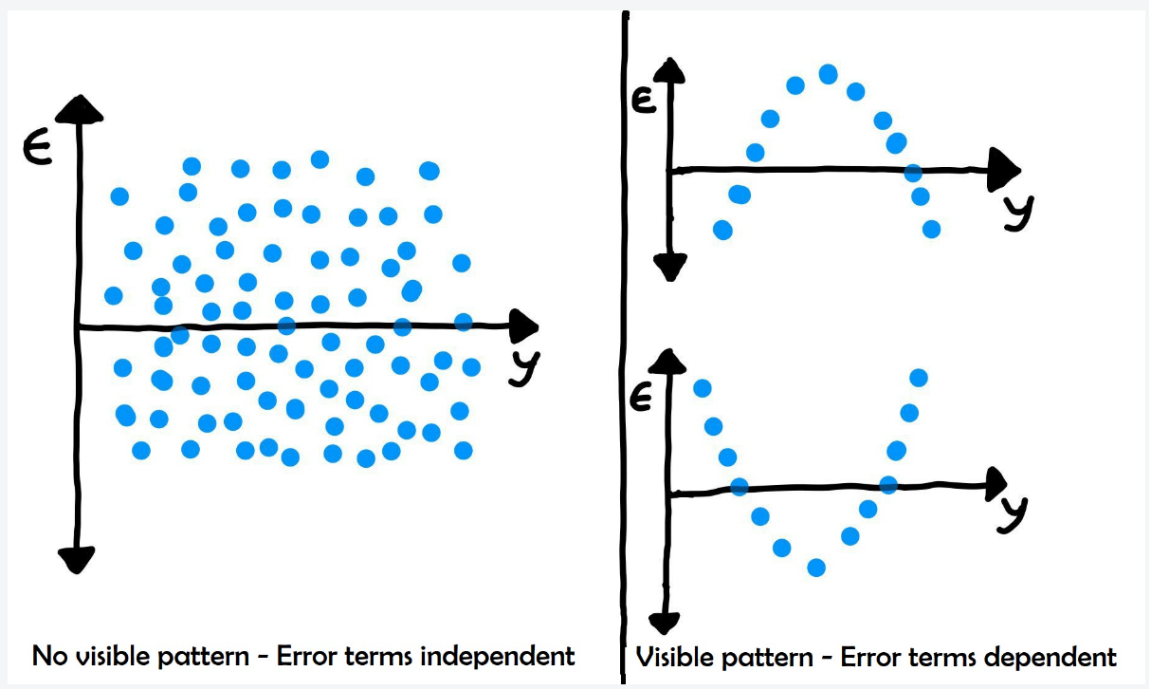


Normality of Error Terms:
Error terms should follow a normal distribution.
We can use a histogram to check the distribution of error terms.

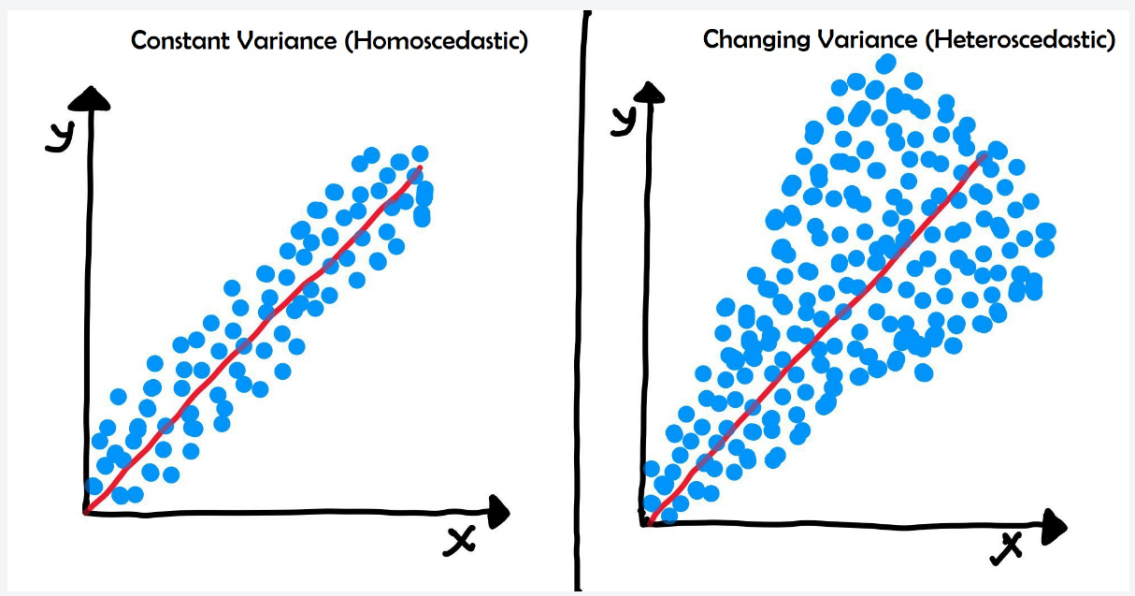


Independence of Errors:

Error terms should be independent of each other.
We can inspect residuals over time or across observations to identify patterns or correlations.



Homoscedasticity of error terms:
Error terms should have constant vaariance across not increase or decrease as error value changes.
A scatter plot of residuals vs. predicted values can help you identify any patterns.



No Multicollinearity:
Independent variables should not be correlated with each other, which can lead to multicollinearity.
We can calculate variance inflation factors (VIFs) to quantify the degree of multicollinearity among independent variables.

These validation steps help ensure that the assumptions of Linear Regression are satisfied, leading to a reliable and accurate model.

Question 5:
Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer:
Temp (temperature) +ve impact

light_snow_rain (weathersit being Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) -ve impact

wind speed -ve impact

Not considering year variable as its assumed bike sharing business is gaining popularity so business is assumed to increase, just highlighting what other factors may contribute towards demand of shared bikes.

General Subjective Questions

Question 1:
Explain the linear regression algorithm in detail.

Answer:

Linear Regression is a statistical technique used to model the relationship between a dependent variable (y) and one or more independent variables (x1, x2, x3, x4…).
The primary goal is to find the best-fitting linear equation that explains the variation in the dependent variable based on the independent variables.

Linear regression is of the following two types

1. Simple Linear Regression
2. Multiple Linear Regression

Basic formula:

$y = \beta_0 + \beta_1 x + \varepsilon$

Where:

- y is the dependent variable you want to predict.
- x is the independent variable.
- β_0 is the y-intercept (the value of y when x is 0).
- β_1 is the slope of the line (how much y changes for a unit change in x).
- ε represents the error term or residuals (unexplained variability).

In multiple linear regression, with multiple independent variables:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$

Furthermore, the linear relationship can be positive or negative in nature as explained below:

Positive Linear Relationship:
A linear relationship will be called positive if both independent and dependent variable increases.

Negative Linear relationship:
A linear relationship will be called positive if independent increases and dependent variable decreases.

Assumptions of linear regression:

1. There is linear regression between x & y or dependent and independent variables.
2. Error terms are normally distributed but not X & Y
3. Error terms are independent of each other
4. Error terms have constant variance

Error calculation methods

1. Order least square method (RSS)
2. Total sum of square (TSS)
3. Residual square error (RSE)

While performing Multiple regression method, we take care of

1. Overfitting
2. Multicollinearity

For model assessment we can use

1. R Square
2. Adjusted R square
3. F-Statistics

Question 2:
Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines, yet have very different distributions and appear very different when graphed.

They were constructed in 1973 by the statistician Francis Anscombe to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

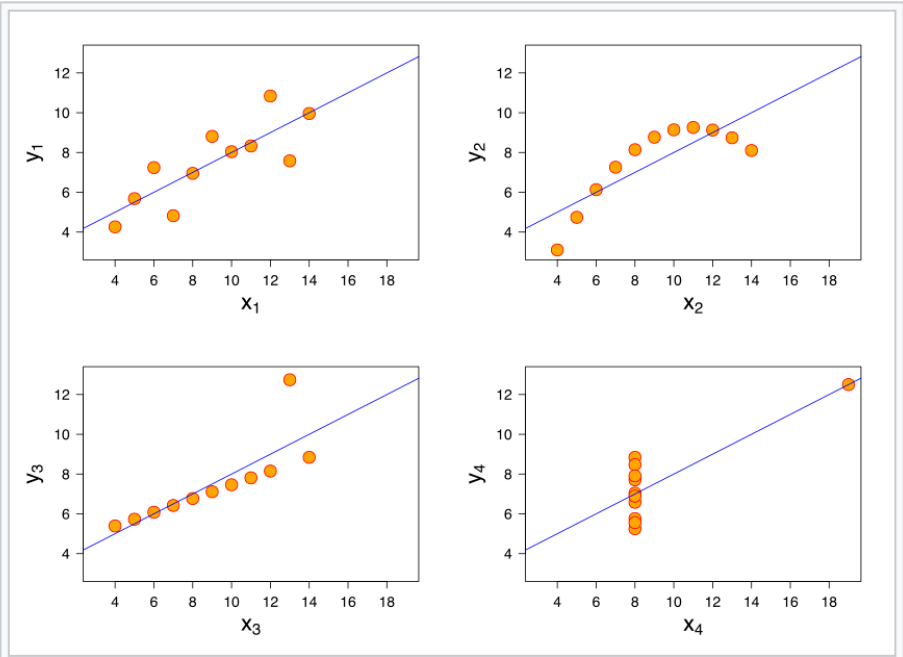
The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

The summary statistics show that the means and the variances were identical for x and y across the groups:

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:

- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

Question 3:
What is Pearson's R?

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.

It ranges between -1 and 1, where:

- r = 1 indicates a perfect positive linear correlation, meaning as one variable increases, the other variable increases proportionally.

- r = −1 indicates a perfect negative linear correlation, meaning as one variable increases, the other variable decreases proportionally.
- r = 0 indicates no linear correlation, implying that there is no consistent linear relationship between the variables.

Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Pearson's correlation coefficient is commonly used to assess the strength and direction of a linear relationship between two variables. However, it assumes that the relationship is linear and that the data follows a normal distribution. It might not capture non-linear relationships well and can be sensitive to outliers.

Question 4:

What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:

Scaling is a step in data preparation for modelling that involves transforming the values of features (variables) in a dataset to a specific range or distribution. The goal of scaling is to bring the features to a common scale, ensuring that no feature dominates others due to differences in their magnitudes.

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.

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.

There are two major methods to scale the variables, i.e. standardisation and MinMax scaling. Standardisation basically brings all of the data into a standard normal distribution with mean zero and standard deviation one.
MinMax scaling, on the other hand, brings all of the data in the range of 0 and 1.

The formulae in the background used for each of these methods are as given below:

- Standardisation: $x = \frac{x - mean(x)}{sd(x)}$
- MinMax Scaling: $x = \frac{x - min(x)}{max(x) - min(x)}$

Difference in scaling methods:

Normalized scaling	Standardized scaling
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

Question 5:

You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer:

When there is perfect correlation between independent variables, then VIF is 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.

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.

Question 6:

What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer:

A Quantile-Quantile (Q-Q) plot is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, such as a normal distribution. It's a powerful visualization technique that helps you compare the quantiles of your data with the quantiles of a theoretical 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 or 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.