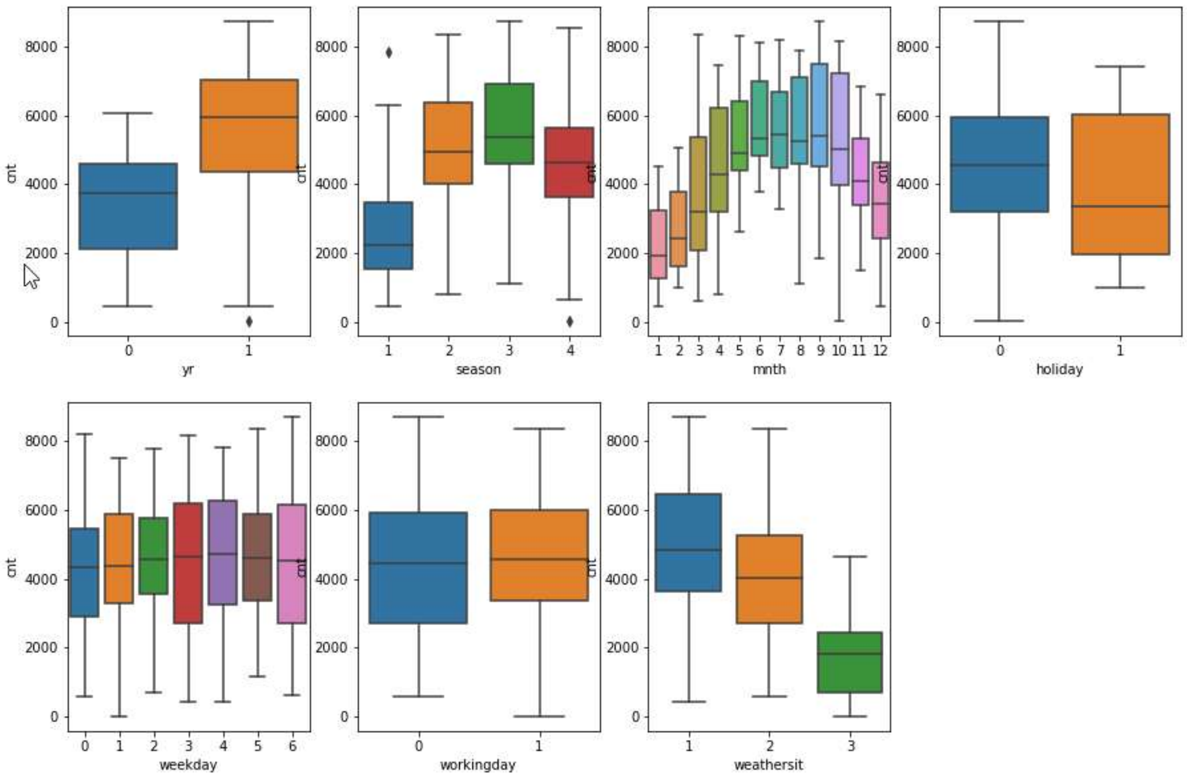
Assignment-based Subjective Questions

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: There are 7 categorical variables as shown in the below boxplots. They are yr, season, mnth, holiday, weekday, workingday and weathersit



- Yr – 2019 recorded more rentals compared to 2018
- season – Fall had highest booking and spring the least
- mnth – September month recorded highest bookings and december the least
- holiday – rental booking was less on a holiday
- weekday – Not major difference in rental volume observed through out the week
- workingday – the median of rental booking is same on workingday and non-working day
- weathersit – People preferred Clear, Few clouds, Partly cloudy, Partly cloudy weather for riding bikes and avoided during Heavy rain or thunderstorm or snow

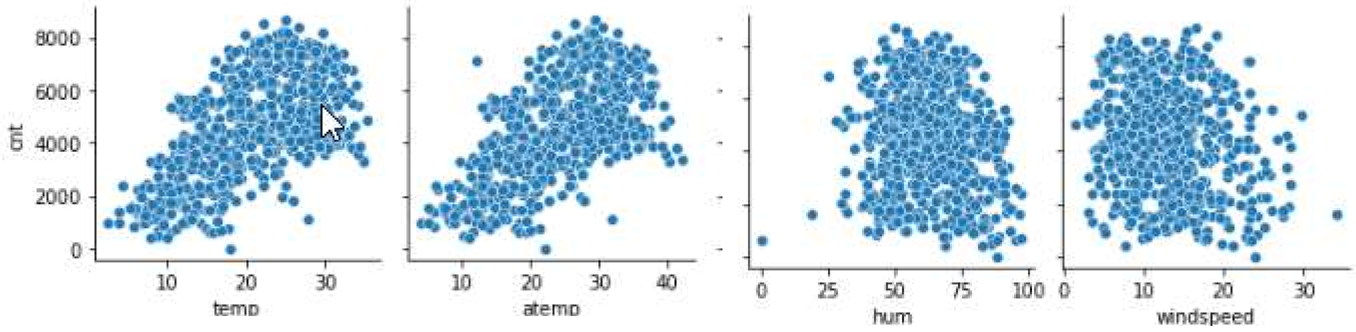
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Answer : drop_first=True helps in reducing the extra column created during dummy variable creation. It reduces the correlations created among dummy variables.

For example – Consider 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.

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

Answer : As per graph shown below, temp variable (Temperature) has highest co-relation 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 :

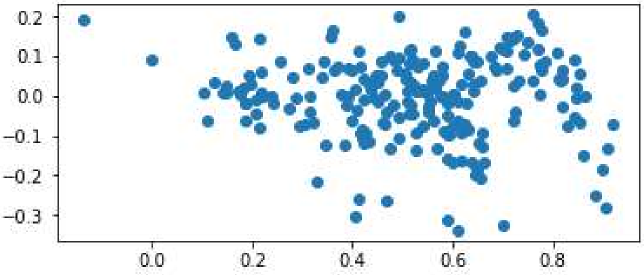
Following assumptions were validated after building the model :

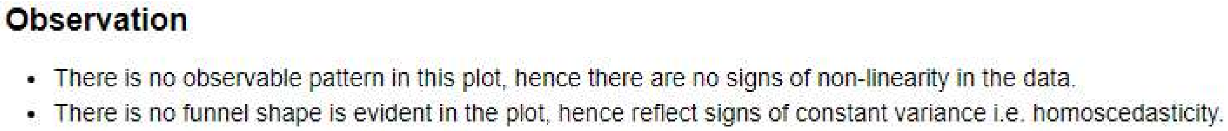
a. Linearity and Additivity of the relationship between dependent and independent variables

We plotted a graph of dependent and independent numeric variables and found them to have linear relationship.

b. Homoscedasticity (constant variance) of the errors

We plotted a scatter plot with y_test_pred and residuals and noticed that there were no observable pattern nor funnel shape is evident in the plot. The plot shows signs of constant variance i.e. homoscadasticity



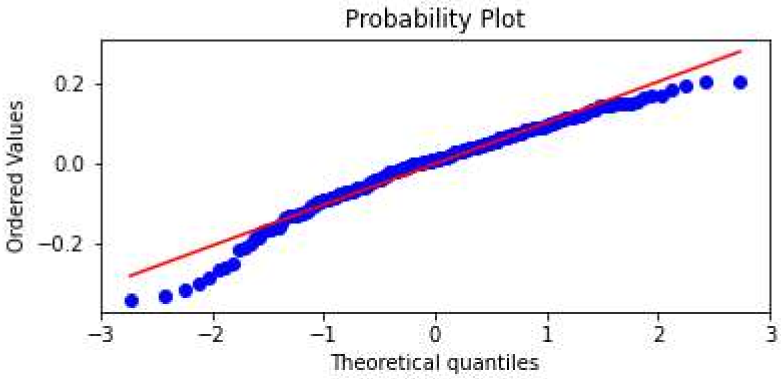


Observation

- There is no observable pattern in this plot, hence there are no signs of non-linearity in the data.
- There is no funnel shape is evident in the plot, hence reflect signs of constant variance i.e. homoscedasticity.

c. Normality of the error distribution

We plotted a q-q plot to ascertain normal distribution



Observation

- Plot show a fairly straight line. Hence data comes from a normal distribution.

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 features that contributed significantly are -

1. temp (temperature) with coeff of 0.609,

2. yr_1 (Year 2019) with coeff of 0.227and

3. humidity with coeff of -0.254.

Temp and year contributed positively whereas humidity contributed negatively.

General Subjective Questions

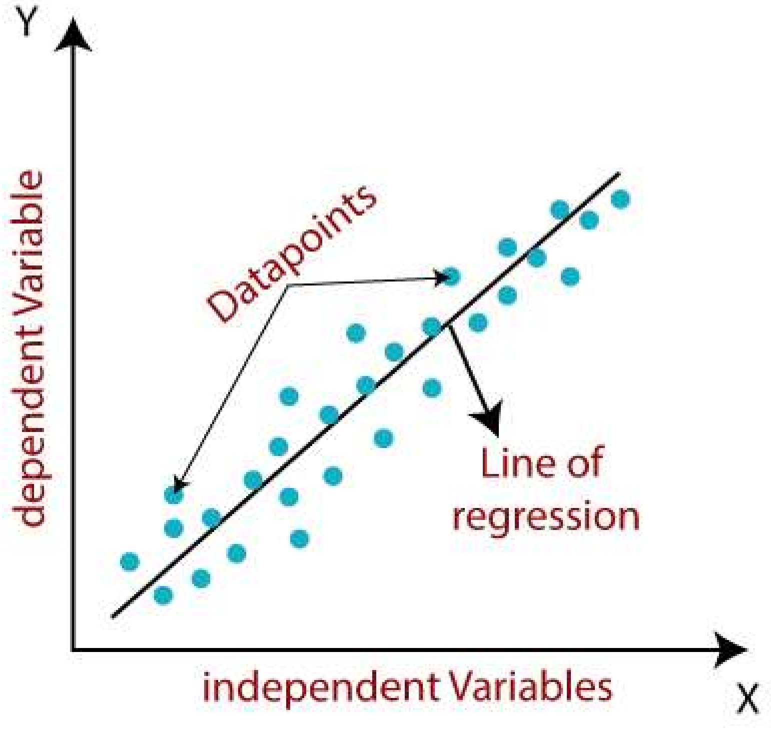
1. Explain the linear regression algorithm in detail. (4 marks)

Answer:

Linear regression is the most popular Machine Learning algorithms. It is a statistical method used for predictive analysis. Linear regression makes predictions for continuous/numeric variables like sales, price and age etc.

Linear regression algorithm shows a linear relationship between a dependent and one or more independent variables. It finds how the value of the dependent variable changes with the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables.



Regression is divided in to simple linear regression and multiple linear regression

1. Simple linear regression – This is used when the dependent variable is predicted using only one independent variable. The equation is as follows -

$$y = \beta_0 + \beta_1 X + \epsilon$$

- y is the predicted value of the dependent variable

- B0 is the intercept
- B1 is the regression coefficient
- x is the independent variable
- e is the error of the estimate

2. Multiple linear regression – This is used when the dependent variable is predicted using multiple independent variables.
The equation is as follows -

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \epsilon$$

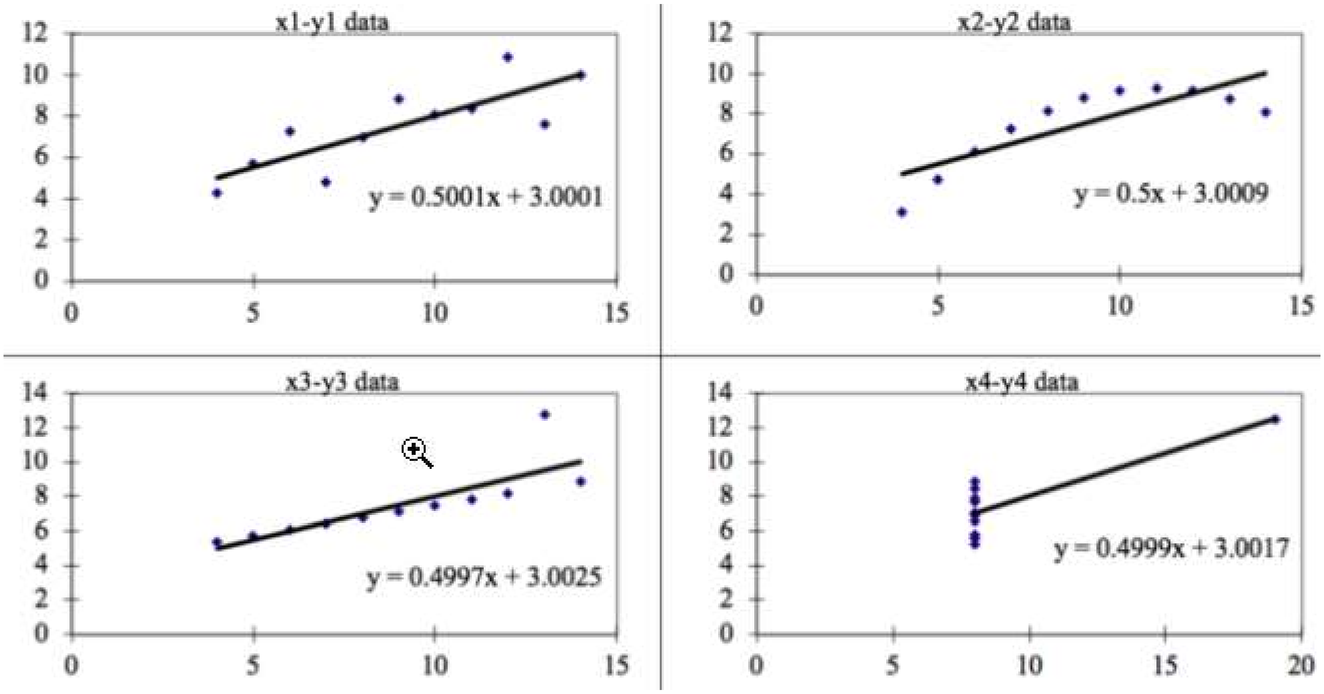
- y is the predicted value of the dependent variable
- B0 is the y-intercept
- B1 is the regression coefficient of the first independent variable X1
- Bn is the regression coefficient of the last independent variable
- ϵ is the model error

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

Answer:

Anscombe’s Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are peculiarities in the dataset that fools the regression model. They have different distributions and appear differently when plotted on scatter plots.

It was developed to emphasize the importance of graphing data before analysing it and the effect of outliers and other observations on statistical properties



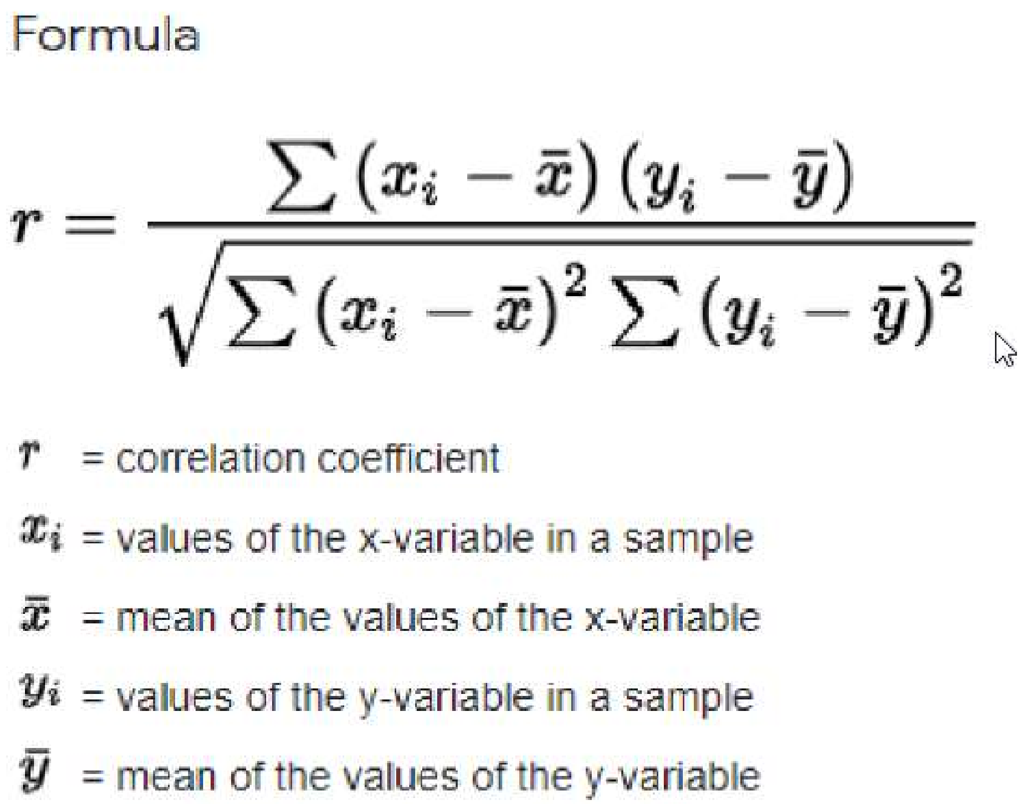
The four datasets can be described as:

1. Dataset 1: this fits the linear regression model pretty well.
2. Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
3. Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
4. Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

3. What is Pearson’s R? (3 marks)

Answer:

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. In simple terms, it tells us “ can we draw a line graph to represent the data? “



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

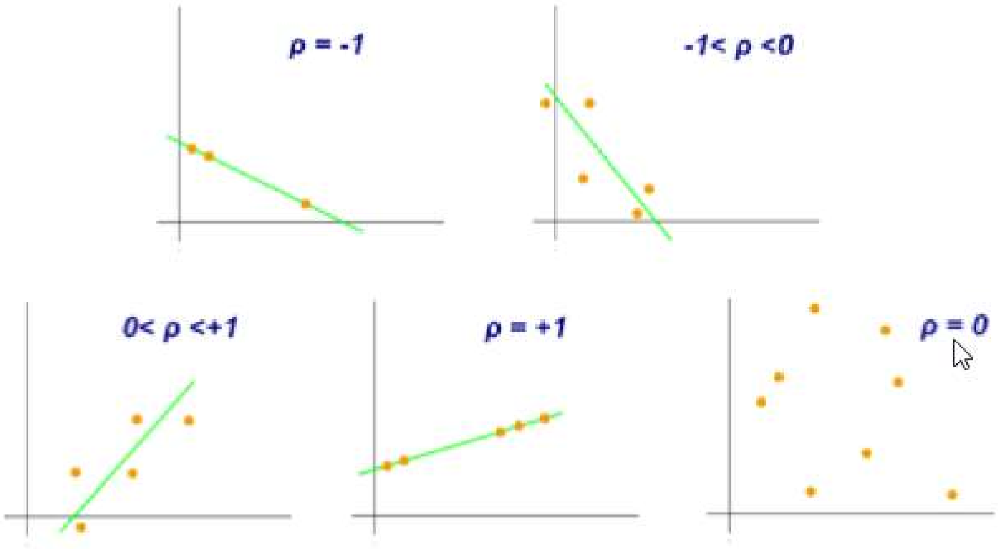
x_i = values of the x-variable in a sample

$\bar{x}$ = 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

As can be seen from the graph below, r = 1 means the data is perfectly linear with a positive slope r = -1 means the data is perfectly linear with a negative slope r = 0 means there is no linear association



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

Answer:

Scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data preprocessing stage to deal with varying values in the dataset. 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, irrespective of the units of the values.

The difference between normalised and standardized scaling are as follows -

Normalised	**Standardized**
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
Scales values between [0, 1] or [-1, 1].	No bounds
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
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:

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 independent 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.

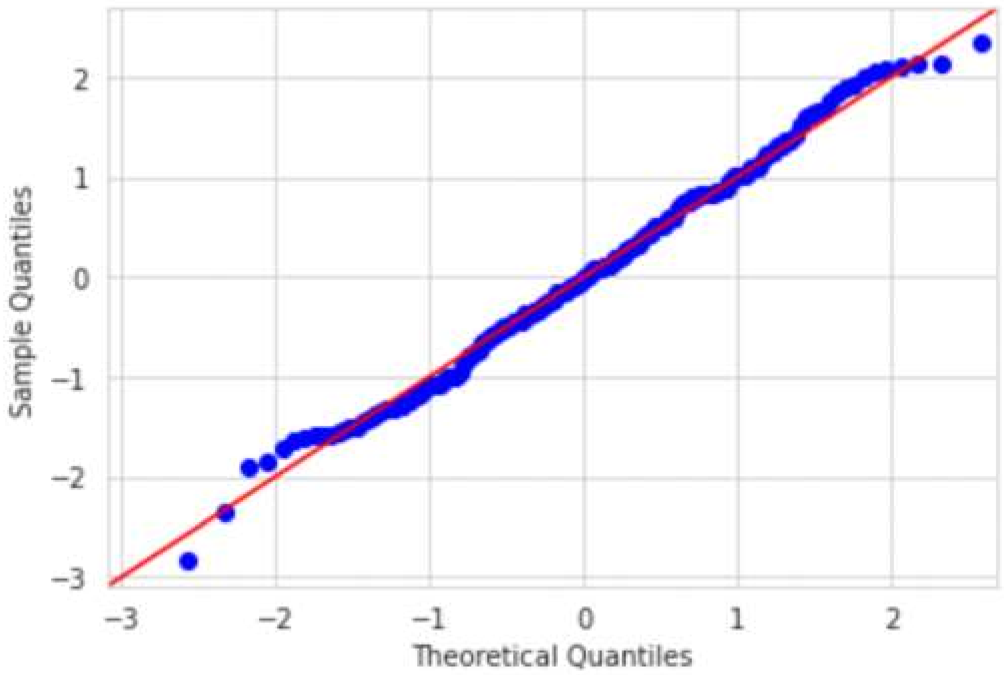
A rule of thumb for interpreting the variance inflation factor:

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

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

Answer:

The Q-Q plot is a graphical plotting of the quantiles of two distributions with respect to each other. In other words we can say it is a graphical technique for determining if two data sets come from populations with a common distribution. For example: If the datasets are distributed similarly then we would get a straight line as shown below.



Use of Q-Q plot :

Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution or even Pareto Distribution, etc. It is used to check following scenarios:

If two data sets —

1. come from populations with a common distribution
2. have common location and scale
3. have similar distributional shapes
4. have similar tail behavior

Importance of Q-Q plot :

Probability distributions are essential in data analysis and decision-making. Some machine learning models work best under some distribution assumptions. Knowing which distribution we are working with can help us select the best model.Hence understanding the type of distribution of feature variables is key to building robust machine learning algorithms. Q-Q plots can help us identify the distribution types by summarising any distribution visually.