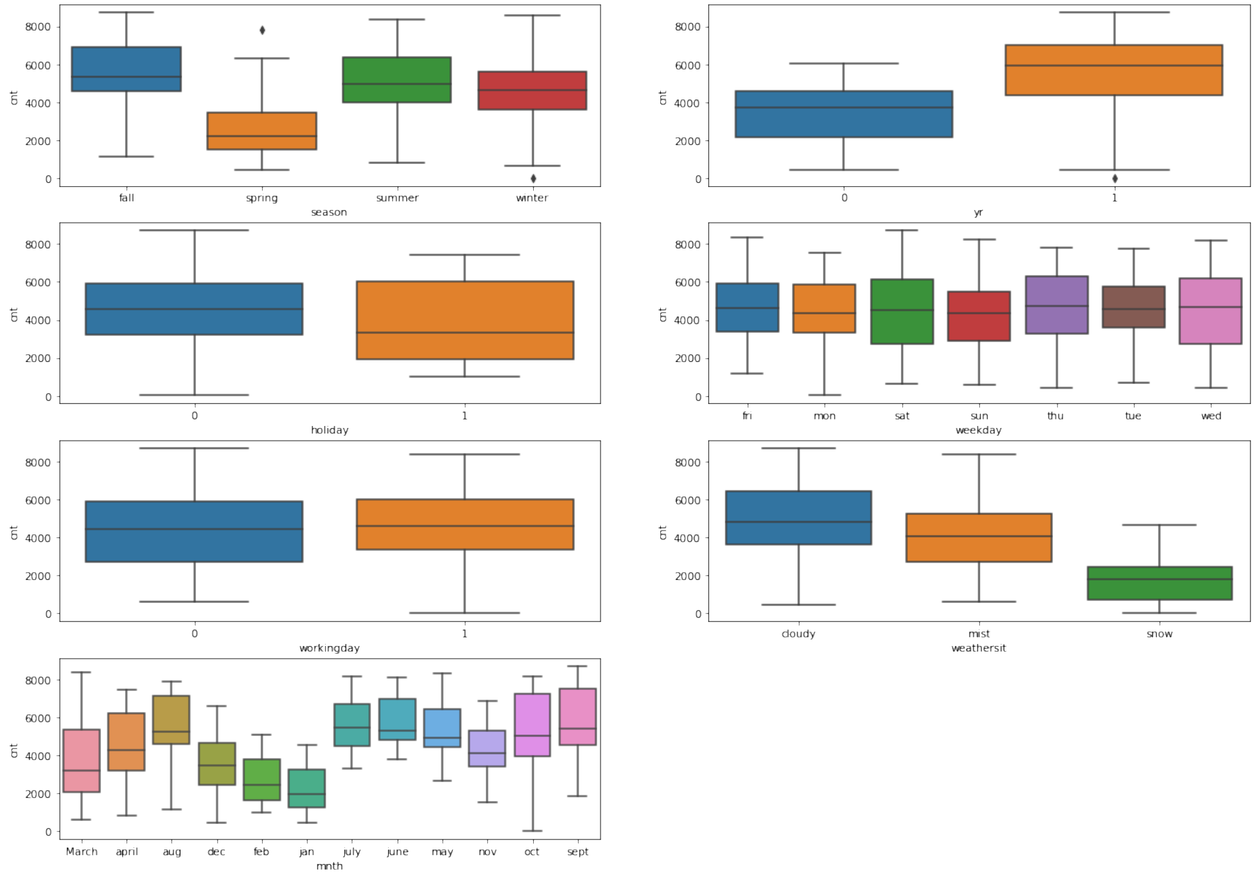
Assignment-based Subjective Questions

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

Answer 1 : In the analysis of categorical variables from the dataset, the following were considered as categorical variables – ***[“season”, “yr”, “holiday”, “weekday”, “workingday”, “weathersit”, “mnth”]***

The below fig show box plots were visualised and we’ve following observations :



Observations :

- Season -> Spring season has least value of cnt, Fall season has highest value of cnt. The other seasons showed similar values of cnt.
- Yr -> The number of rentals in 2019 are higher compared to 2018.
- Holiday -> During holidays, the rentals reduced.
- Weekday -> The rentals were nearly constant throughout the weekdays.

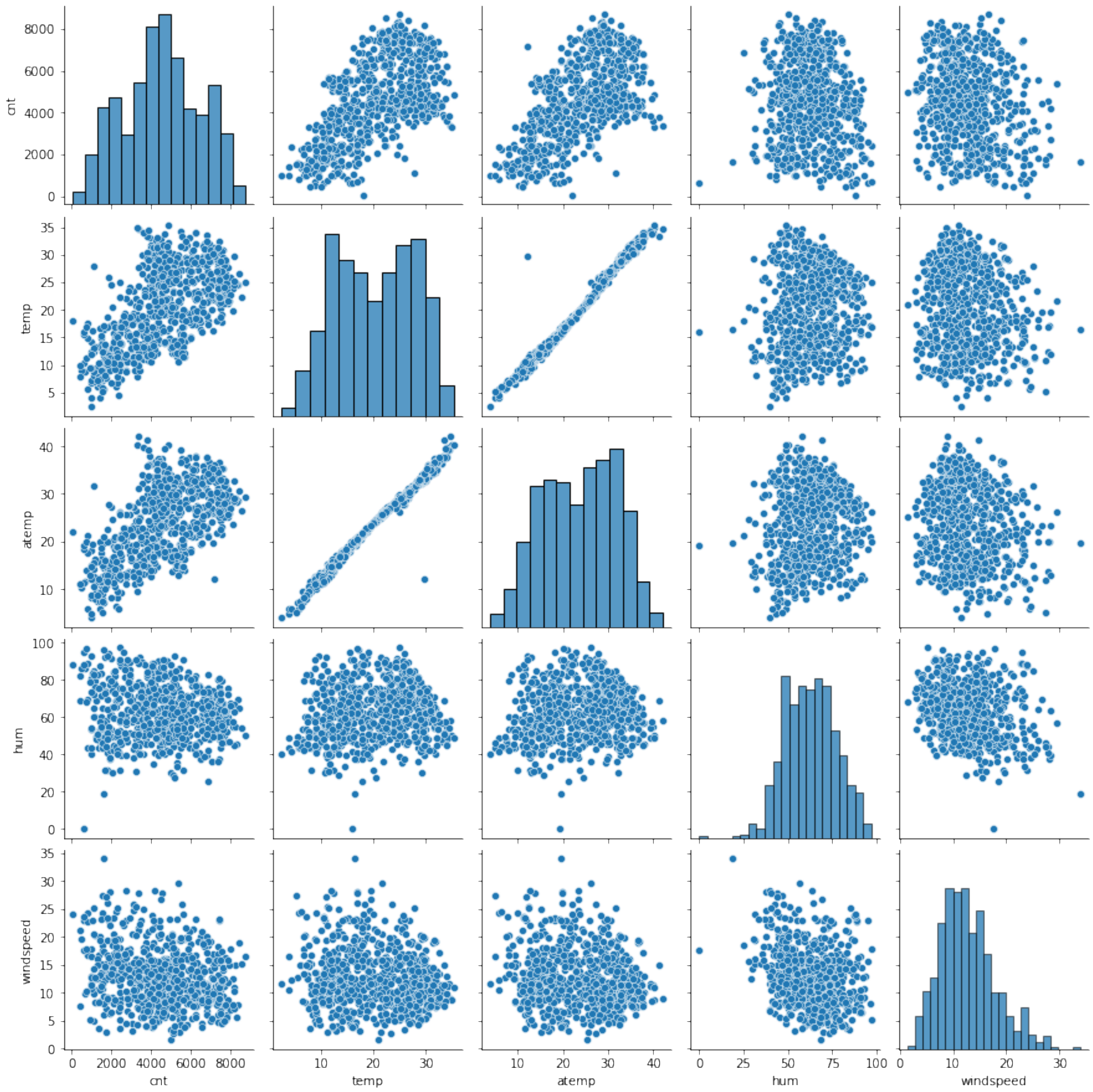
- Weather Sit -> Highest count is observed in cloudy or nearly clear skies, following mist and snowy conditions. It is highly expected behaviour that during misty or snowy conditions, the rentals must have dropped drastically. When the weather conditions are un-favourable, we have practically no or less users.
- Mnth -> September month saw highest number of rentals whilst December saw lower number of rentals.

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

Answer 2 : The main reason is to avoid over-correlation of dummy variables thereby affecting the model adversely and affect is stronger when cardinality is smaller. If these are considered, the training time for the model also increases thus reducing our expectancy in limited time frame.

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

Answer 3 : In my pair-plot consideration of numerical variables, ***["temp", "atemp"]*** values are highly correlated, even though this stage **atemp** is considered, the same is later pruned in modelling.



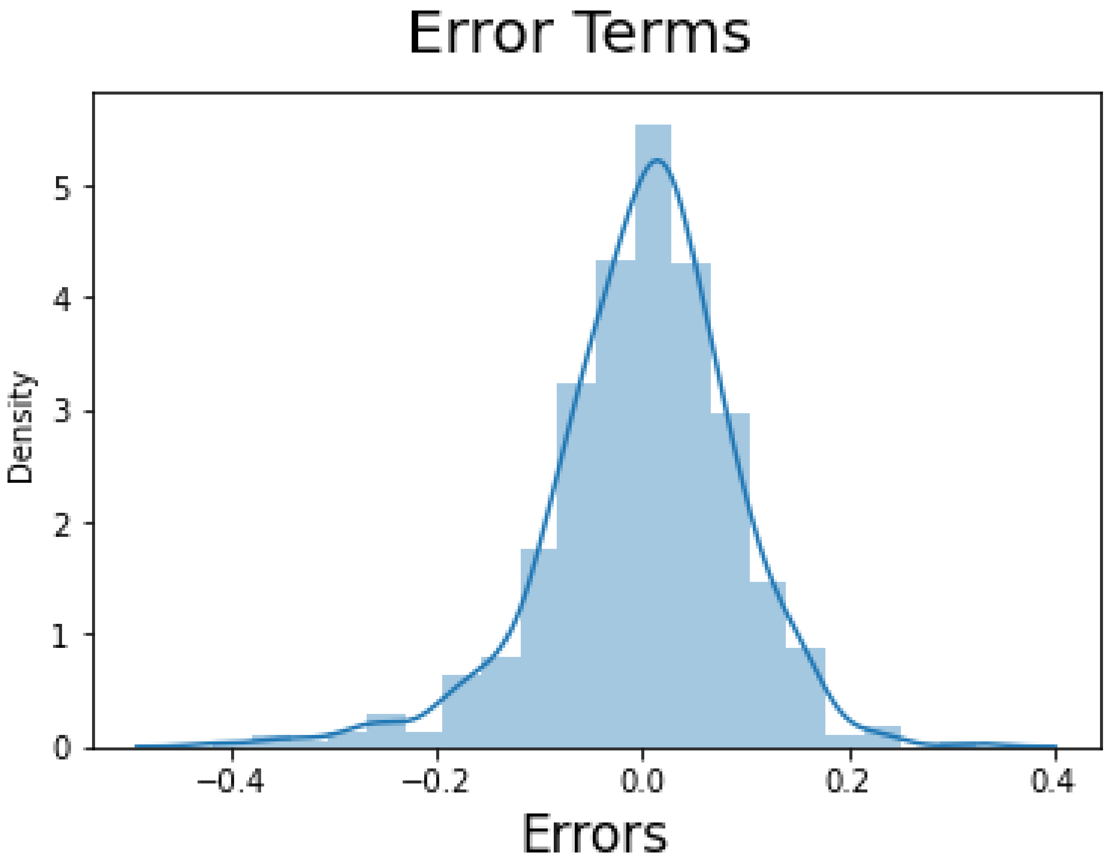
Question 4: How did you validate the assumptions of Linear Regression after building the model on the training set?

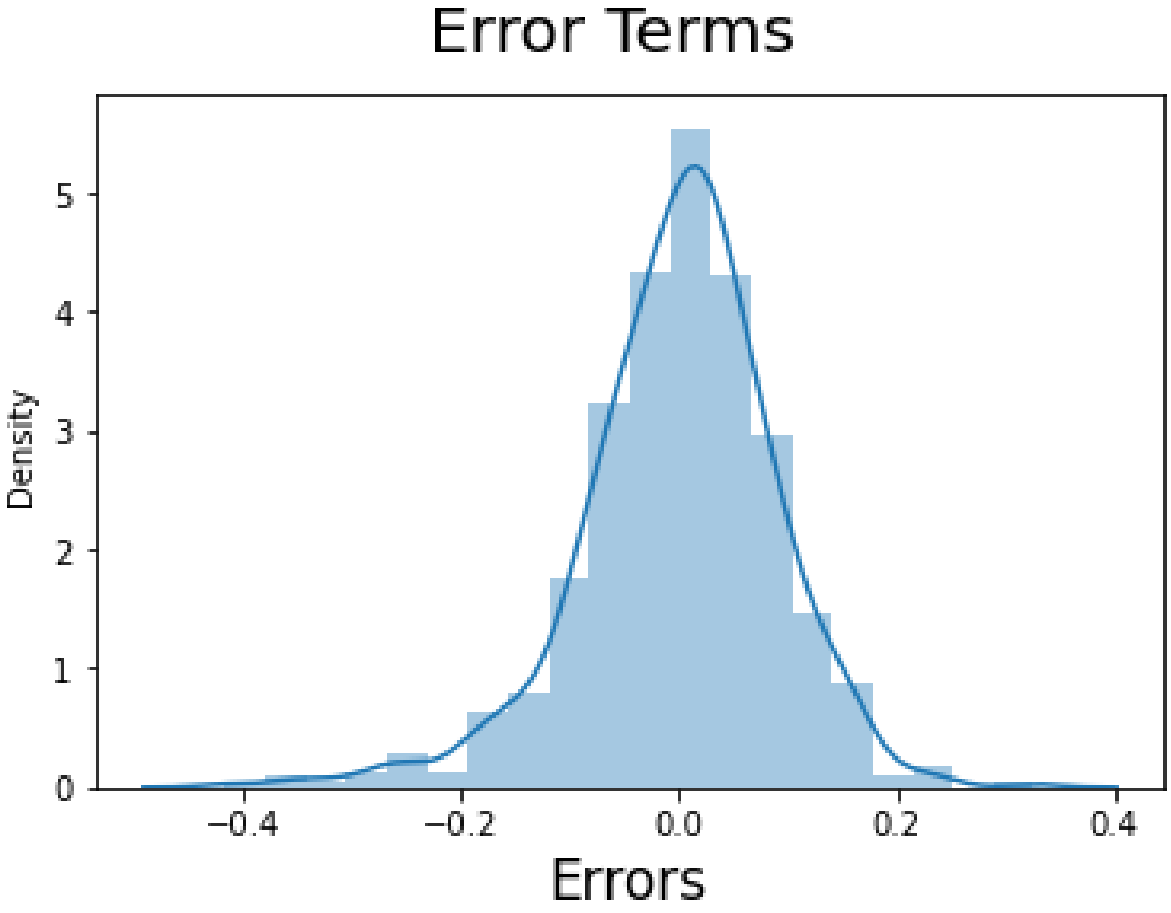
Answer 4:

- First is the visualisation of linearity existence among the dependent/independent features or variables.

We've plotted pair-plots to identify the existence of linearity in the variables.

- The second assumption is the residual distribution should be normal and mean centred at ~0 (mean=0).


Error Terms
Density
0
1
2
3
4
5
−0.4
−0.2
0.0
0.2
0.4
Errors


Error Terms
Density
0
1
2
3
4
5
−0.4
−0.2
0.0
0.2
0.4
Errors

- Third consideration is to verify if there is little or no multicollinearity in the data. Multicollinearity occurs when there is high correlation w.r.t each other. Such values are pruned by calculating VIF (Variance Influence Factor) and understand how these features are correlated.

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

Answer 5 : The following features/variables are contributing significantly towards the explanation of demand of the shared bikes i.e. ***[“tmp”, “yr”, “weathersit_snow”]***

General Subjective Questions

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

Answer: Linear regression is a statistical method used for modelling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. The goal of linear regression, is to find the best-fitting line that minimizes the sum of the squared differences between the observed and predicted values. The linear regression algorithm is commonly used for predictive modelling and understanding the relationship between variables.

In a simple linear regression, there is a single independent variable ***X,*** and a dependent variable ***Y***. The relationship between ***X and Y,*** is represented by the equation of a straight line:

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

Y is the dependent variable, X is the independent variable, $\boldsymbol{\beta_0}$ is the y-intercept, $\boldsymbol{\beta_1}$ is the slope(coefficient) of the line and $\in$ is the error.

The same equation is extended if multiple variables are involved as follows :

$$Y = \beta_0 + {}_{i=1}^{i=n}\Sigma\ \beta_i . X_i + \in$$

The y-intercept and slope is calculated as follows:

$$\boldsymbol{\beta_0 = \overline{Y} - \beta_1 * \overline{X}}$$

$$\beta_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X}) * (Y_i - Y)}{\sum_{i=1}^{n}(X_i - \overline{X})^{2}}$$

where X(bar), Y(bar) are X-mean and Y-mean i.e. mean values of X and Y respectively.

Assumptions, Interpretation, Prediction, Evaluation :

- Linear Regression makes assumption that there is linear relationship between variables, normality of the residues (i.e. mean = 0), homoscedasticity.
- The slope (beta-one) represents the change in the dependent variable X for a unit change in independent variable.
- This equation helps us in making predictions of new values of the independent variable.
- Model created based on linear regression is often assessed by metrics like Mean Square Error (MSE), R-Squared, or Adjusted R-Squared.
- Wide usage is seen in applied statistics and machine learning. Multiple Linear Regression

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

Answer : Anscombe's quartet, is defined as a group of four data sets that have nearly identical simple descriptive statistics but differ significantly when represented graphically. This quartet was created by statistician Francis Anscombe in the year 1973 to emphasize the importance of visualising data and not relying solely on summary stats. Each dataset consists of ***11(X,Y)*** points.

Summary Stats of 4 Dataset tabulated as follows :

Summary Variables	Values			
Mean of X	9.0	9.0	9.0	9.0
Variance of X	11.0	11.0	11.0	11.0
Mean of Y	7.5	7.5	7.5	7.5
Variance of Y	4.125	4.125	4.125	4.125
Correlation X and Y	0.816	0.816	0.816	0.816
Linear Equation	Y = 3 + 0.5*X	Y = 3 + 0.5*X	Y = 3 + 0.5*X	Y = 3 + 0.5*X

The descriptive stats for these datasets are nearly similar. Currently we're considering the case they're same but in reality they may be slightly varying. When these, are plotted graphically, we will observe very different and unravelling linear, non-linear, perfectly linear except for an outlier presence, perfectly linear for an influential outlier relationships.

The Anscombe's quartet thus, highlights the importance of visualising the dataset to understand the underlying patterns and relationships and sole dependence on relying on the statistics of the datasets might not be very useful and capture the complexity involved.

What is Pearson's R? (3 marks)

Answer: Pearson's correlation coefficient, often denoted as ***'r'*** or ***Pearson's r,*** is a measure of the strength and direction of a linear relationship between two continuous variables. It quantifies the degree to which a change in one variable is associated to the other variable, with a proportion ratio. The value of 'r' ranges between '-1' and '1',

Where

- -1 denotes perfectly negative linear relationship.
- 0 denotes no linear relationship.
- 1 denotes perfectly positive linear relationship.

For a given pandas data frame, we can find the correlation matrix using **df.corr()**

The Pearson's r coefficient is calculated as per this formula:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X}) * (Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^{2} \sum_{i=1}^{n}(X_i - \overline{X})^{2}}}$$

Furthermore, there are few points which are generally considered as properties of this coefficient such as

- Scale Invariance – If we add a constant or multiply to each of these variables, the value remains the same.
- Symmetry – The correlation between X and Y is same as Y and X.
- Range – The range as mentioned previously is between -1 and 1.
- Outlier Sensitivity – The coefficient can be influenced by the presence of outliers. If there are extreme data points, they may affect the correlation invariably.
- Linear Measurement Only – The correlation specifically measures the strength and direction of linear relationships.

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

Answer: Scaling refers to a process of transforming numerical variables to specific range or distribution. The goal is to be bring all numerical variables to comparable level thus preventing domination of individual variables or being biased by its scale compared to others. This is very important in machine learning that use distance-metrics, gradient based optimization or regularization.

Types of Scaling – 1) Normalised Scaling 2) Standardized Scaling.
The former is also known as min-max-scaling, which transforms variables to a specific range, usually 0-1.

Formula for Min,Max Scaling ->

$$X_{normalised} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

The latter is also known as z-score-scaling, tranforms variables to have a mean of 0 and standard deviation of 1.

The formula for this is

$$X_{zscore\ or\ standardised} = \frac{X - \mathrm{mean}(X)}{\mathrm{Std}(X)}$$

The following are the differences of Normalised and Standardised Scaling tabulated as follows:

	Normalised Scaling	Standardised Scaling
Alternative Name	Min-Max Scaling	Z-score Scaling

Range	The values range between 0-1	Mean = 0,
		SD = 1

Distribution Impact	Preserves relative distribution.	Preserves the shape of the distribution.
Outlier Sensitivity	More prone	Less prone
Suitability	For unknown distribution, this is applied in general	For known distribution like Gaussian, this is applied.

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

Answer: VIF stands for Variation Inflation Factor, is a measure that quantifies the severity of multicollinearity in a regression analysis. Multicollinearity occurs when two or more independent variables are highly correlated leading to disparities in correlation coefficients.

VIF Formula :

$$\mathrm{VIF}_i \quad = \quad \frac{\mathbf{1}}{(\mathbf{1}-\boldsymbol{R_i}^{\mathbf{2}})}$$

here the R_i^2 is the (R)sq value obtained by regressing the ith independent variable against all the other independent variable.

VIF value 1 implies no multicollinearity and higher value implies high multicollinearity. Generally VIF value greater than 5 or 10 is considered a cause for concern.

VIF value – ‘infinity’, implies the model is able to create multiple perfect regression models i.e. R(i) is 1. When divided by zero, its INF or infinity. The following are probable causes in the regression analysis

Perfect Multicollinearity, Redundant Data.

Probable solutions:

Identify and remove redundant variables, perform data cleaning, variable transformation or scaling, regularization by known techniques like Lasso or Ridge etc.

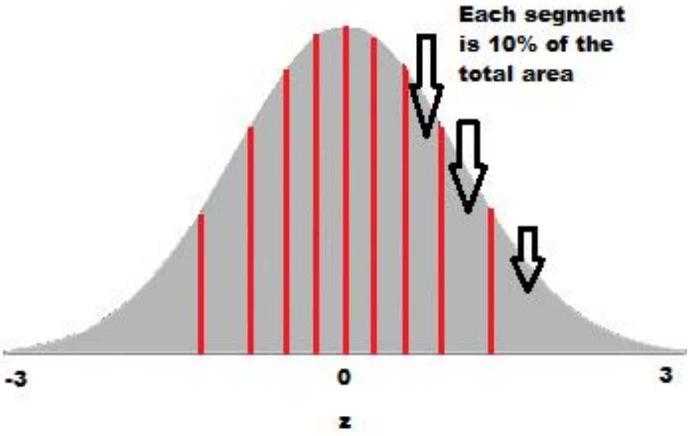
What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

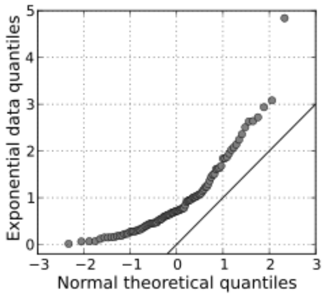
Answer: A Q-Q plot, shortly abbreviated, is popularly known as Quantile-Quantile plot. It's a graphical tool used in statistics to access whether a dataset follows a particular theoretical distribution. It compares the quantiles of the observed data with quantiles of specific probability distribution, usually normal distribution. Q-Q plots are very much helpful in checking and validating the assumption of normality in a dataset and widely used in linear regression and other statistical analysis.

There are three components viz Quantiles of Theoretical Distribution, Quantiles of Observed Data, and the diagonal line.

Usage of Q-Q plots: Normality check, Identifying skewness, Outlier detection, Model residual analysis.

How to interpret Q-Q plots? If the points are quite close to the diagonal line, it implies they follow expected distribution. Alternatively, if they're too far or deviation is observed, it signifies deviance from expected distribution. The nature of this deviance is to be visually assessed. If the distribution is close to expected distribution, the tails and central regions gives good insights of the data.





A sample graphs are taken as example. From tail ends ie -3, 0 (mean point), 3 points, we calculate z-value infer if its lying in middle or extreme left or extreme right.