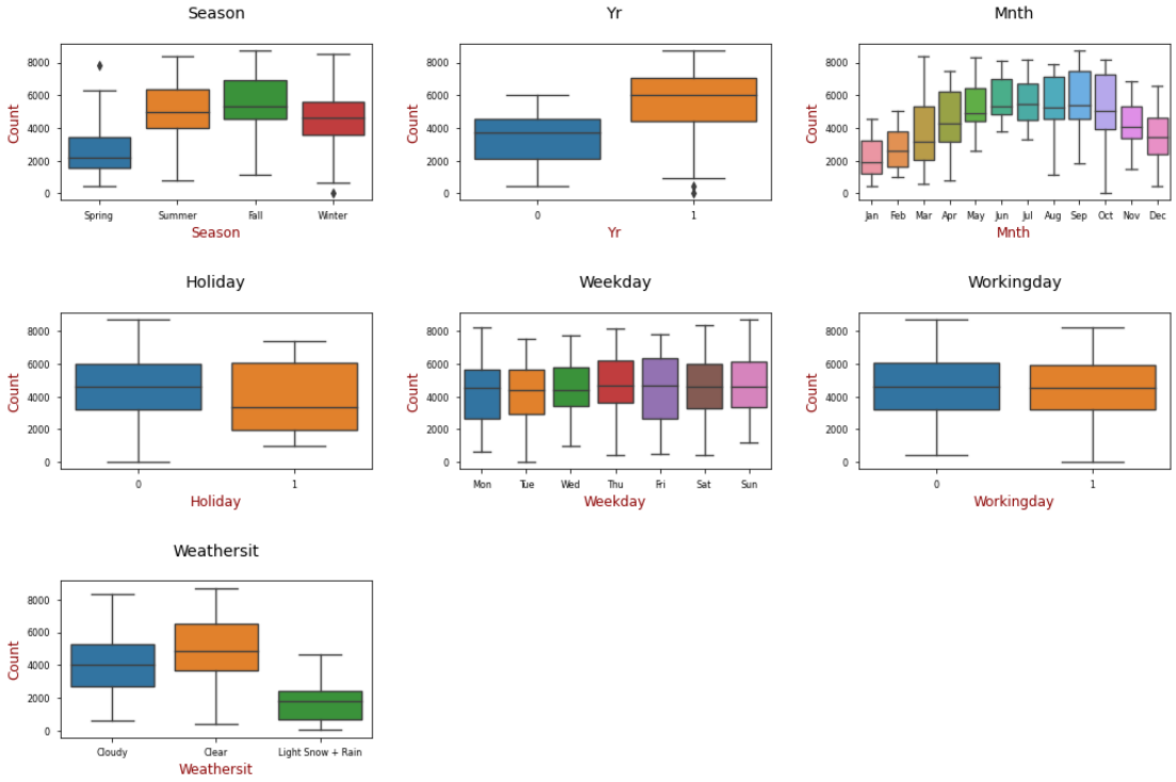
Name: Gauri Bhale

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?
   Answer: The categorical variables in the data set are **season**, **year**, **month**, **holiday**, **weekday**, **workingday** and **weathersit**.



1. **Season**: The demand for shared bikes is most likely to be higher in fall and least likely to be in spring
2. **Year**: Demand of shared bikes has increased in 2019 than the previous year
3. **Month**: Demand of shared bikes is less in the month of January, February, November and December than rest of the year. September has the highest demand
4. **Holiday**: Bike demand is higher on non-holidays than on holidays
5. **Weekday**: Bike demand is higher on Saturday and Sunday than on any other weekday.
6. **Workingday**: Bike demand is very similar whether its working day or not, working day has slightly less demand
7. **Weathersit**: Highest demand of shared bikes is seen during clear weather whereas Light Snow + Rain resulted in low demand. There is no demand during Heavy Rain + Thunderstorm.

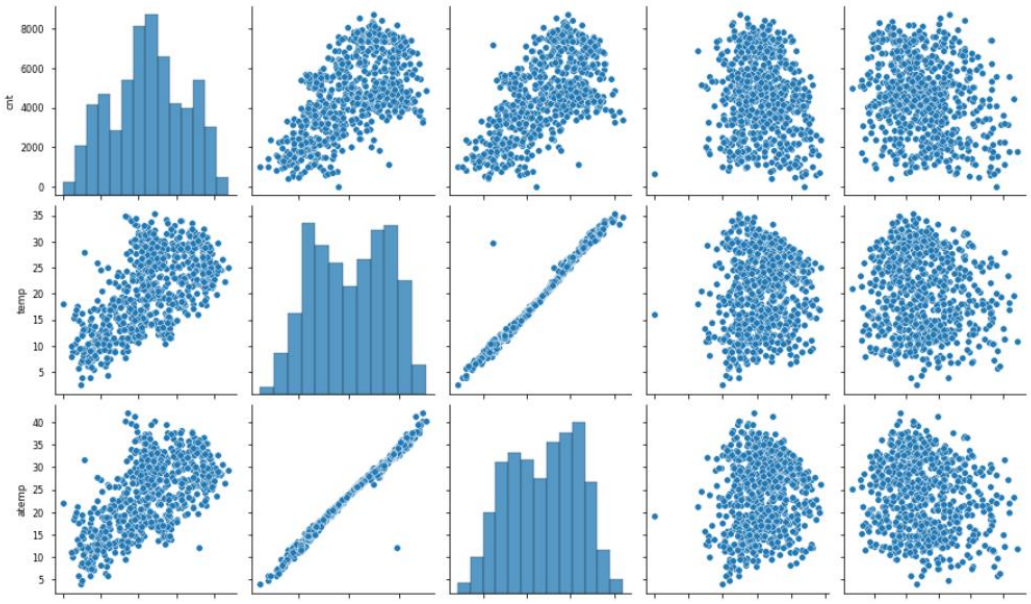
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

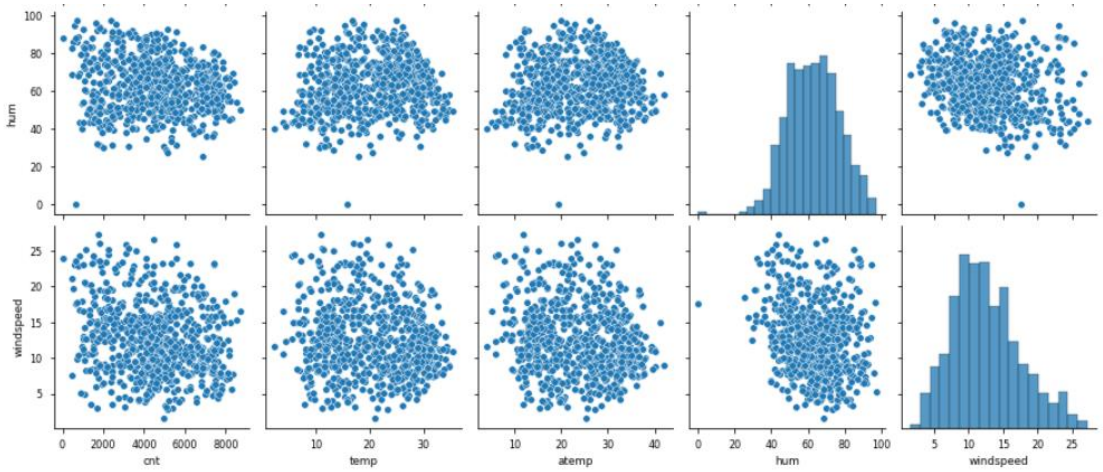
   Answer:

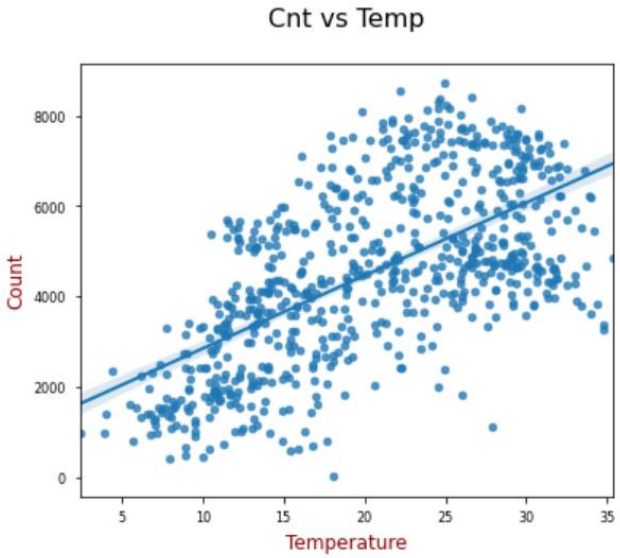
   1. drop_first=True helps in removing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables
   2. For example, if we have 3 types of values in categorical column, say, furnished, semi-furnished and unfurnished and created dummy variable 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. Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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

   Answer:





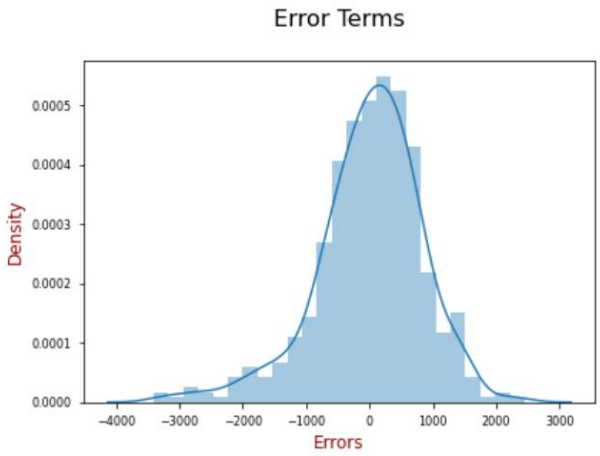


Temp variable has the highest correlation with the target variable cnt with coefficient of correlation 0.63. As it can be seen from the plot that as the temperature increase, the demand for shared bike increases.

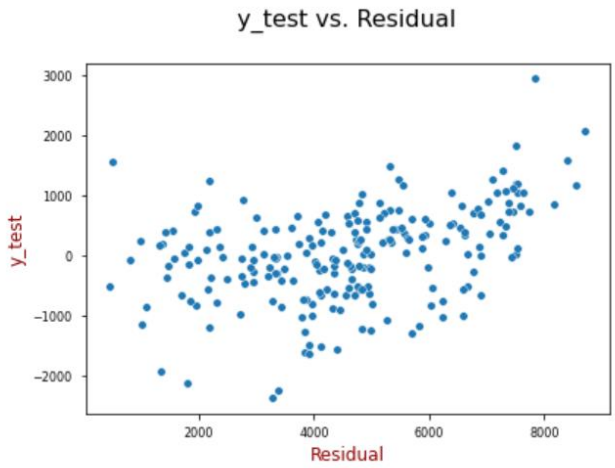
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

   Answer:

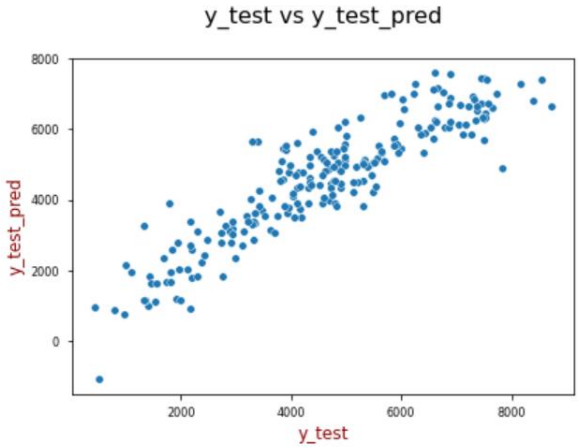
   1. Error terms are normally distributed with mean zero (not X, Y):



2. Error terms are independent of each other:



3. Error terms have constant variance (homoscedasticity):



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:

1. **Temperature** (4879.31): Based on the value of coefficient, there is linear relationship between temperature and demand. As temperature increases, demand of bikes shared increases.
2. **Weathersit_Light Snow + Rain** (-2419.48): It is cleared form the value of coefficient that Light Snow + Rain is affecting the demand of shared bikes inversly.
3. **Year** (1928.18): As the year has progressed, the demand of shared bikes has increased.

General Subjective Questions:

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

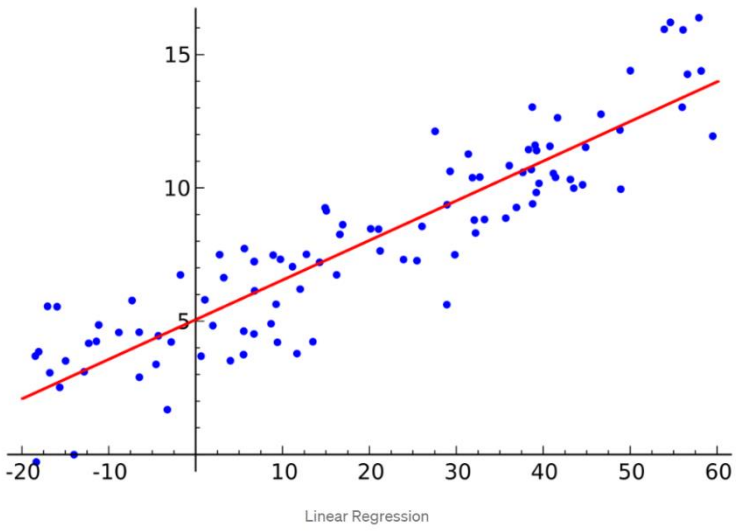
   Answer:

   Regression:

   - Regression is a supervised learning technique that supports finding the correlation among variables. A regression problem is when the output variable is a real or continuous value.
   - In other words, **Regression shows a line or curve that passes through all the data points on a target-predictor graph in such a way that the vertical distance between the data points and the regression line is minimum.** It is used principally for prediction, forecasting, time series modelling, and determining the causal-effect relationship between variables.

   Linear Regression:

   - Linear regression is a simple statistical regression method used for predictive analysis and shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression.
   - If there is a single input variable (x), such linear regression is called **simple linear regression**. And if there is more than one input variable, such linear regression is called **multiple linear regression**.
   - The linear regression model gives a sloped straight line describing the relationship within the variables.



   - The above graph presents the linear relationship between the dependent variable and independent variables. When the value of x (independent

variable) increases, the value of y (dependent variable) is likewise increasing. The red line is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best.

- To calculate best-fit line linear regression uses a traditional slope-intercept form.

$$y = mx+b \implies y = a_0+a_1x$$

y= Dependent Variable.
x= Independent Variable.
a0= intercept of the line.
a1 = Linear regression coefficient.

- A regression line can be a Positive Linear Relationship or a Negative Linear Relationship.

Multiple Linear Regression:

- Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.
- Multiple regression is an extension of linear (OLS) regression that uses just one explanatory variable.
- Formula and Calculation of Multiple Linear Regression

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$$

where, for i=n observations:

y_i = dependent variable

x_i = explanatory variables

β_0 = y-intercept (constant term)

β_p = slope coefficients for each explanatory variable

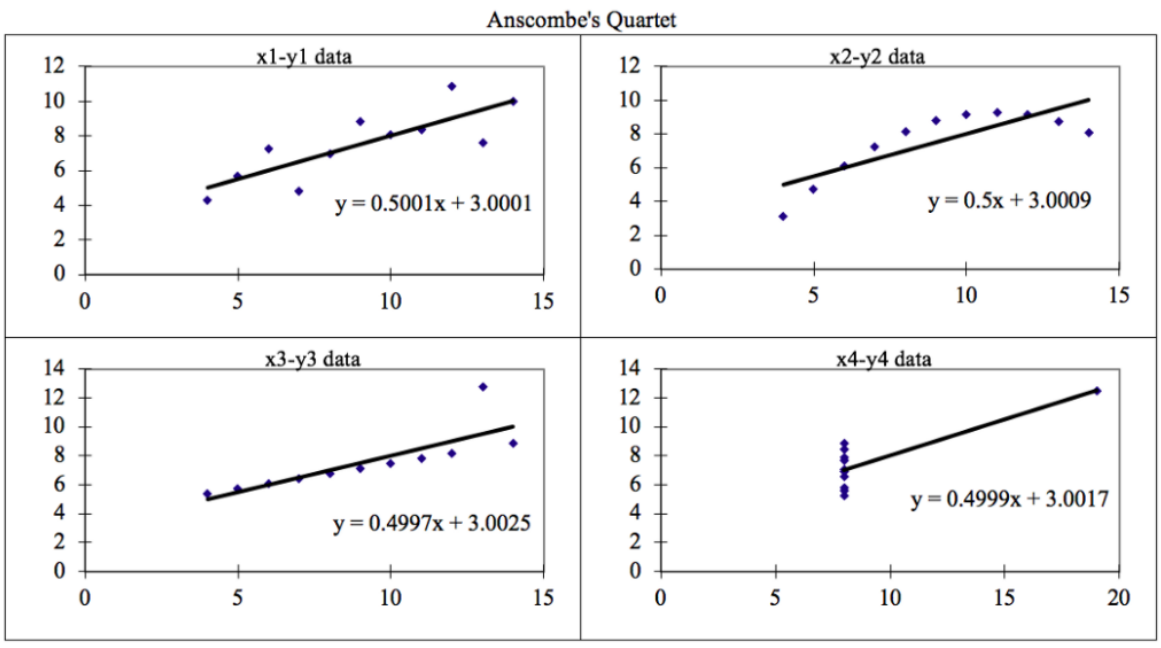
ϵ = the model's error term (also known as the residuals)

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 some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.
- This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. These four plots can be defined 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	



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

   Answer:

   - In Statistics, the **Pearson's Correlation Coefficient** is also referred to as Pearson's R, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.
   - It cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.
   - Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a **product moment**, that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

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

   Answer:

   - **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.

   - Min-Max Normalization: This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

   - Standardization: It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$x_{new} = \frac{x - \mu}{\sigma}$$

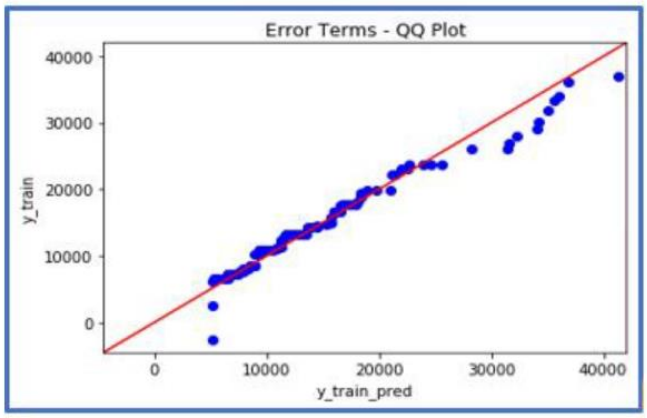
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
   Answer:

   - If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables.
   - In the case of perfect correlation, we get R2 = 1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
   - An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

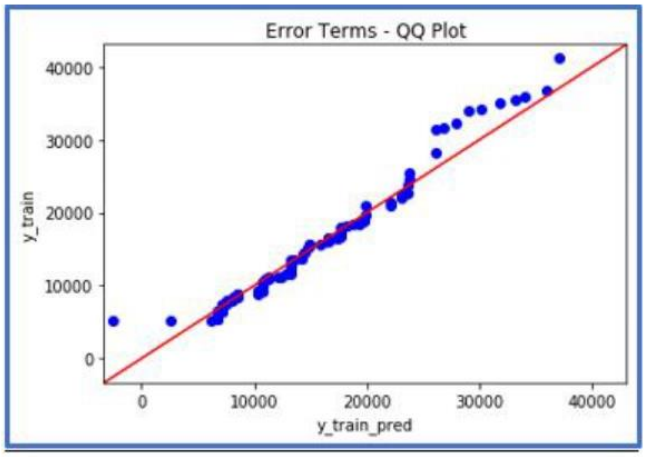
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
   Answer:

   - **Quantile-Quantile (Q-Q) plot** is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.
   - This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.
   - <u>Few advantages</u>:
     a. It can be used with sample sizes also
     b. Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
     c. It is used to check following scenarios: If two data sets —
        i. come from populations with a common distribution
        ii. have common location and scale
        iii. have similar distributional shapes
        iv. have similar tail behaviour
   - <u>Interpretation:</u>
     A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Below are the possible interpretations for two data sets.
     a. <u>Similar distribution:</u> If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b. Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c. X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d. Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis