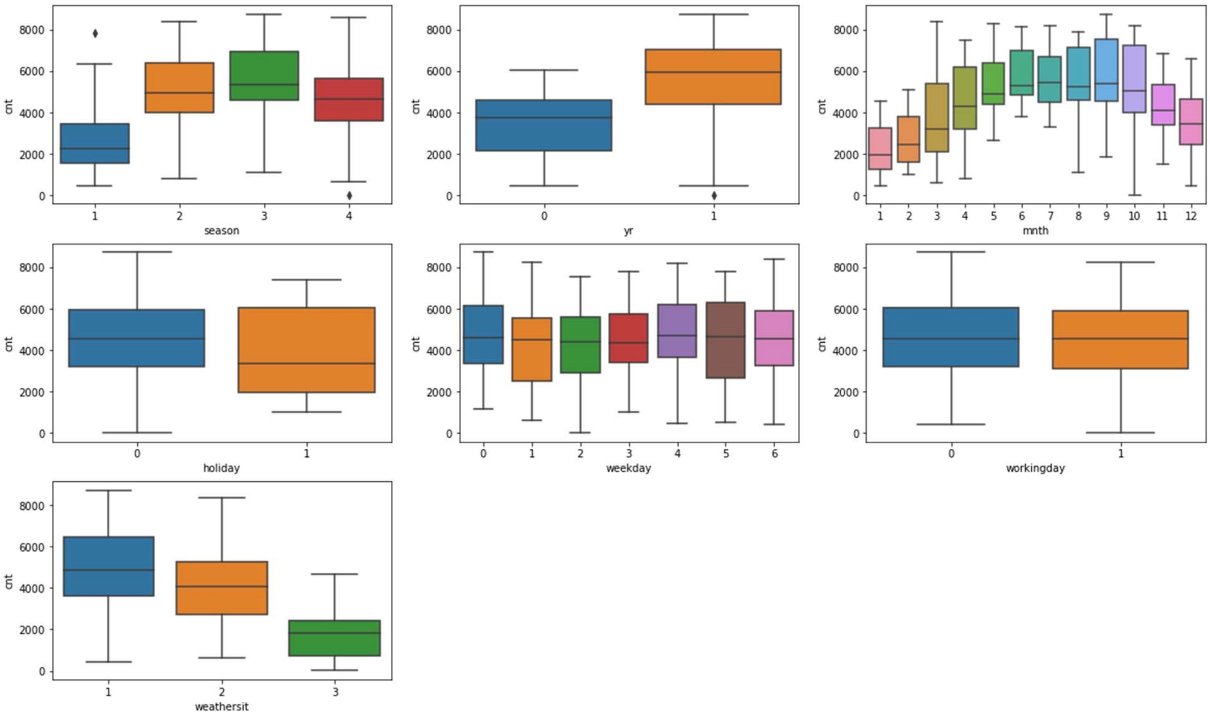
Linear Regression Answers

Chaitanya Yatin Shejwal

Assignment based subjective questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
Ans:



- **season:** the count of rental bikes increase when its not spring season
- **yr:** the count of rental bikes have increased over the year
- **mnth:** the count of rental bikes till September
- **holiday:** the median count of rental bikes is greater when there are no holidays. However, there is no clear distinction between the spread
- **workingday & weekday:** the count of rental bikes seems to be quite stable across the week
- **weathersit:** the count of rental bikes is 0 when there is heavy rain or thunderstorm. On the other hand the demand for rental bikes is highest when the weather is not snowy or rainy

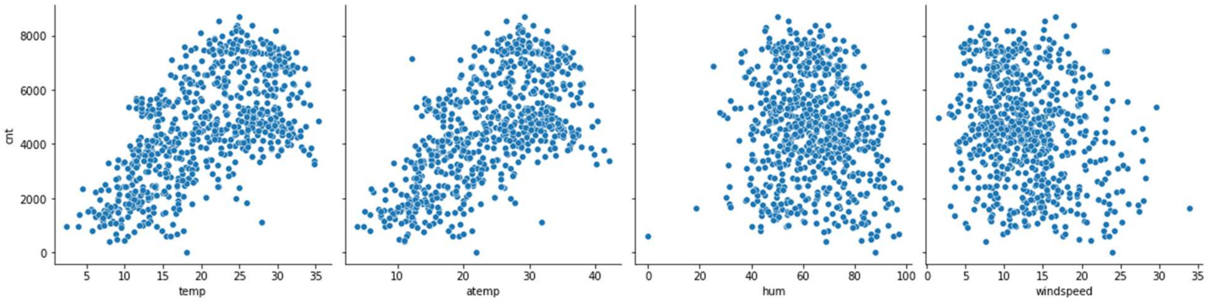
Q2. Why is it important to use drop_first=True during dummy variable creation?
Ans:

- Dummy variables are created to make the categorical values as the column value and their respective value a binary variable
- For explaining (n) categorical variables ideally we need (n-1) categorical columns. Eg:
  If there are 3 gender types: Male, Female, and Transgender. An individual can take only one gender value. After dummy variable creation we are left with two choices,
    a. Create three columns - Male, Female, and Transgender
    b. Or create two columns - Male and Female

  Now, if the individual is a transgender then the possible values are:
  Male - 0, Female - 0, and Transgender - 1, and
  Male - 0, Female - 0
  The above information provides the same conclusion. Hence, if the person is not a male and female then the person is definitely a transgender.
- Hence, to create (n-1) columns for (n) categories we use the drop_first=True argument.
- The argument basically doesn’t create a new column for the first categorical value

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans:

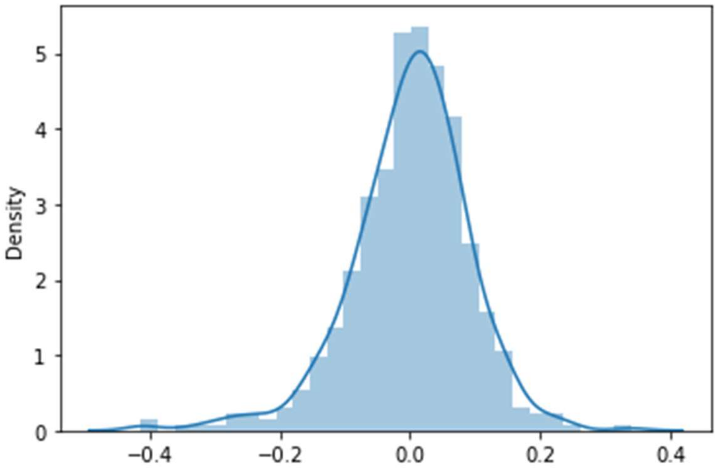


- Amongst the numerical variables the feature “registered” has the highest correlation with the dependent variable. However it is because the cnt feature is the sum of “registered” and “casual” feature. Hence it is evident that these features will be highly correlated
- On the other hand “temp” and “atemp” are highly correlated with the target variable
- Around 63% of the variance is explained by these two features
- It seems that as the temperature starts rising the number of rental bikes increase

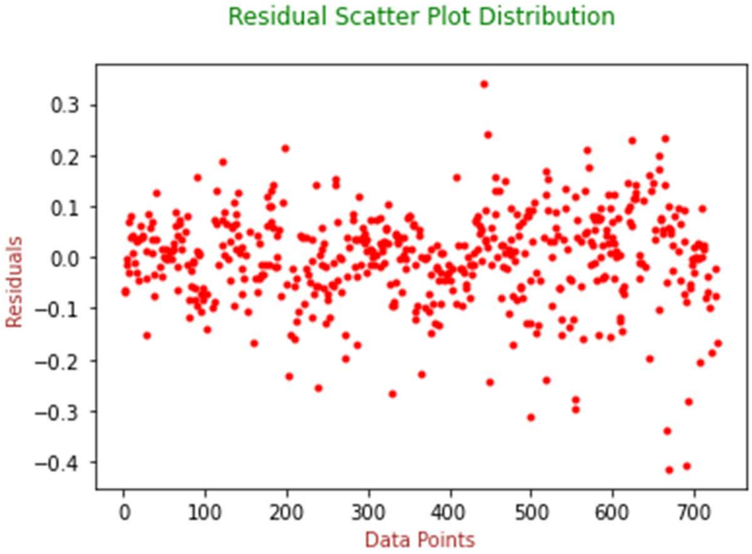
Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans:

- Assumption 1: Error terms are normally distributed. This assumption was validated by plotting a histogram of residuals (difference between the actual data point and predicted data point). As evident from the below graph the error terms are normally distributed around mean = 0



- Assumption 2: Error terms are independent of each other. This assumption was validated by plotting a scatter plot of residuals or the error terms
- Assumption 3: Error terms have constant variance. This assumption was validated by plotting a scatter plot of residuals and it was found that the variance and the scattered points don't create a pattern



Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Ans:

- **Weather**: In the dataset none of the bikes were rented when the weathersit category was 4. And from the model constructed if the weathersit category is 2 or 3 then the probability of renting bikes decreases. Hence, if the weather is Clear, Few clouds, Partly cloudy, Partly cloudy the probability of bikes rented increases
- **Temperature**: Higher the temperature higher the chance of bikes being rented
- **Season**: The probability of bikes being rented is higher if the season is either summer or winter

General subjective questions

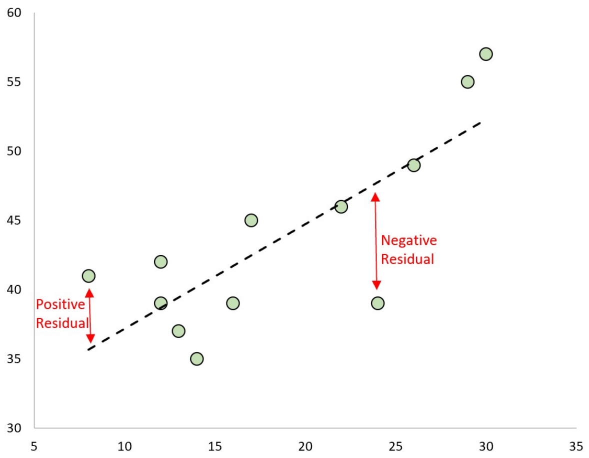
Q1. Explain the linear regression model in detail.
Ans:

- **Overview**
    - The linear regression model aims at explaining the relationship between a dependent variable and one or more independent variables. Often these dependent variable is also referred to as the target variable and the independent variables are referred to as the predictors.
    - The difference between correlation and regression is that correlation measures the strength of association between 2 different variables and regression quantifies this association by plotting a straight line
    - If one independent variable is used to predict the target variable then it is known as simple linear regression and if more than one independent variables is used then it is known as multiple linear regression
    - The simple linear regression equation is as follows:

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

      The multiple linear regression equation is as follows:

      $$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + \epsilon$$
    - Graphically, the dependent variable is plotted on the y-axis and independent variables are plotted on the x-axis
    - Generally all the data points will not fall on the predicted line hence we add a notion of error to the equation
    - The predicted values are generally represented as y-pred
- **Residuals**
    - The distance between the actual data point and the predicted data point is known as a residual.
    - Residuals can be positive or negative based on where the actual data point lie relative to the predicted data point.
    - General term used to calculate the residuals is known as Residual Sum of Squares. Squares are taken to nullify negative effect of residuals.



- **Cost function**
    - For a particular value of x there exists a range of values of y. Hence, it's a need that we select the optimal data point
    - This optimal data point can be selected by minimizing the overall residual distance
    - And hence to find the optimal solution the errors between intercepts and the respective slopes of the line needs to be minimized
    - This happens using two approaches:
        - Differentiation
        - Gradient descent
- **Assumptions**
    - Since we know that when our model was under creation there were different range of y values for corresponding x values. Hence it is quite possible that our model picked up some error. Hence, there are certain assumptions that are associated with our linear regression model
        - Linear relationship exists between x and y
        - Error terms are normally distributed
        - Error terms are independent of each other
        - Error terms have constant variance (homoscedasticity)
    - It is important to check for these assumptions because we are making inferences about the population using sample and hence just building a linear regression model isn't enough
- **Hypothesis test**
    - Since the data can be either randomly scattered or linearly scattered, it is important to cross check whether the slopes calculated are significant or not and hence hypothesis testing is important to be conducted
    - The null hypothesis is where the slopes are equal to zero and alternate hypothesis is where the slopes are not equal to zero

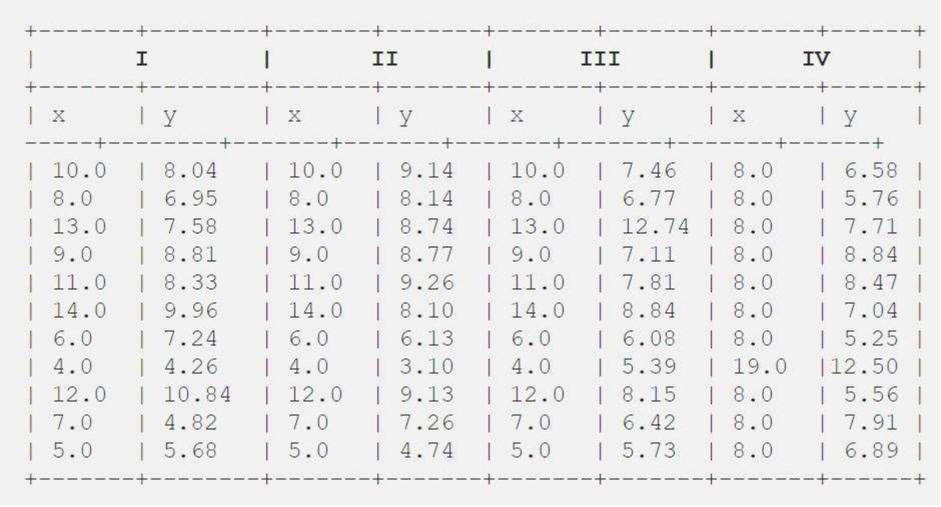
  - If we fail to reject H0 then it means the predictor-slope is insignificant and if we reject H0 then it means the fitted line is significant and not by chance
- **Important values to consider when evaluating linear regression model**
  - R-square: the R-square value signifies how good the model is able to explain the variance in the data
  - F-statistic and prob(F-statistic): this value signifies whether the overall model fit is significant or not. This is important to be taken into consideration because sometimes the slopes and intercepts are significant but the overall model fit happened just by chance
  - p-value: the p-value confirms the significance of the predictors
- **Multicollinearity**
  - It is important to be checked when building a linear regression model because it helps in finding the relationship between predictors
  - Multicollinearity is checked by calculating the metric - Variance Inflation Factor.

    If VIF is high then the variable can be explained by other variables as well and hence can be dropped
- **Other important points**
  - Linear regression guarantees interpolation of data and not extrapolation which means that the predicted value will lie within a certain range of the data we already have
  - Linear regression is a parametric model which means that it is model that has finite number of parameters
  - Bias-variance-trade off is a metric that is used to compare two different combinations of models because the more number of features we add to the model the better it explains the variance in the data. This extra effect of adding can be nullified by calculating the Adjusted R-square which takes into consideration the number of independent variables considered
  - Feature selection: It is important to select the right subset of predictors from the dataset so that the model is less complex, concise, and better. There are two methods to select the right features - manual feature elimination and automated feature elimination.

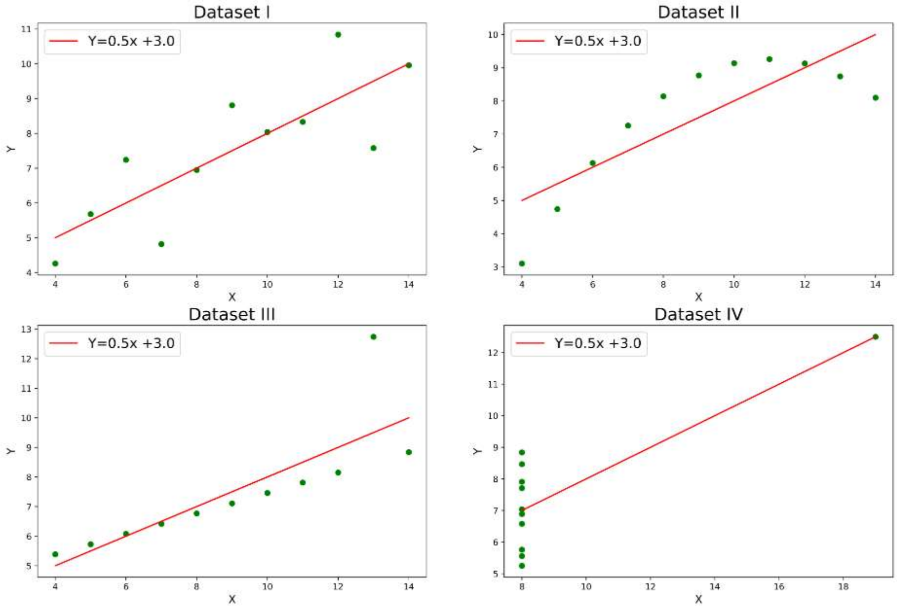
    The types of automated feature elimination are:
    - Recursive Feature Elimination
    - Forward/Backward/Stepwise selection
    - Regularization
    - Balanced approach (combination of both manual and automated selection techniques)

Q2. Explain the Anscombe's quartet in detail.
Ans:

- Francis Anscombe in 1973 constructed 4 datasets to showcase the importance of visualization of data and also to counter the point, *"numerical calculations are exact, but graphs are rough"*.
- The 4 datasets created by Anscombe have the same statistical properties in terms of mean, standard deviation, R-squared, correlations, and linear regression.
- The 4 datasets created by Anscombe have 11 x,y data points each set
    - The visualizations generated from the datasets is as below:



```
+-------+--------+-------+-------+-------+-------+-------+------+
|       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 |
+-------+--------+-------+-------+-------+-------+-------+------+
```



Dataset I

Dataset II

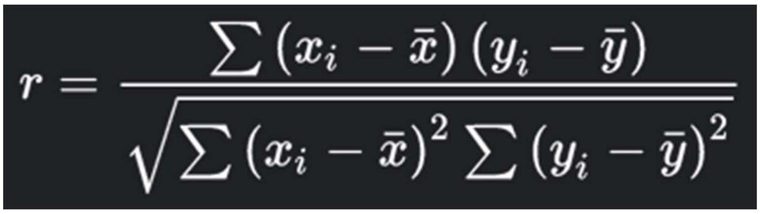
Dataset III

Dataset IV

- Since Anscombe created 4 datasets to explain the importance of EDA it is also known as Anscombe's quartet.

Q3. What is Pearson's R?
Ans:

- Pearson's R also known as the correlation coefficient explains the linear relationship between 2 variables by plotting them on x and y axis.
- Pearson's R only explains the relationship between the two variables however it doesn't explain causation between the two variables.
- The formula to calculate Pearson's R is:

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

  Where,
  x_i is the value of the variable x,
  x-bar is the average of the values of x
  y_i is the value of the variable y
  y-bar is the average of the values of y
- The value of Pearson's r is always in the range of -1 and +1
- Possible values of r and corresponding insights
  - If the value of r > 0.8 then a strong positive linear relationship exists between x and y
  - If the value of r < 0.8 and r > 0.3 then a weak positive linear relationship exists between x and y
  - If the value of r is close to 0 then there exists no relationship between x and y
  - If the value of r < -0.8 then a strong negative correlation exists between x and y
  - If the value of r > -0.8 and r < 0 then a weak negative correlation exists between x and y
- Advantages of r:
  - It is unit free
  - Any change in scale doesn't affect the value of r
  - It not only explains the relationship between two variables in magnitude but also in terms of direction

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Ans:

- The dataset contains different features/columns and each feature gathered has its own respective units. Eg: distance might be stored in km or miles, gender could be stored in binary variables, or age might be stored in days, yrs, or months.
- During model building these units and their respective values have an impact on the coefficients of the predictor variables
- Hence, in order to build an unbias and robust model these variables should be brought under same scales which helps in better and unbiased comparison
- It is important to note that scaling doesn't affect the original distribution of the dataset. It just shifts the data points.
- Scaling also doesn't change the value of R-square, F-statistic, prob(F-statistic), t-value, and p-value.
- Hence, the reason behind performing scaling is:
  - Helps with interpretation
  - Faster convergence of gradient descent which means that all the transformations performed by the gradient descent algorithm are not needed and are done quickly
- **Normalized scaling:**
  - Normalized scaling is also known as min-max scaling.
  - Performing normalized scaling on a variable scales its values in the range between 0 and 1
  - The formula of min-max scaling is as follows:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- **Standardized scaling:**
  - Standardized scaling scales the variables in such a way that the mean of the variables is 0 and separated by 1 standard deviation
  - The formula of standardized scaling is as follows:

$$X' = \frac{X - \mu}{\sigma}$$

- **Difference between normalized and standardized scaling**

Normalized scaling	Standardized scaling
Scaled values are in the range of 0 and 1	Scaled values have a mean of 0 and distributed by 1 standard deviation

More sensitive to outliers	Less sensitive to outliers
Good choice if the data distribution is not known	Good choice if the data distribution is known

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans:

- Variance Inflation Factor (VIF) is a measure to assess whether multicollinearity exists in the predictors. This approach is adopted mainly to understand the effect of a variable on the dependent variable.
- It is done by creating a model where one of the predictors now become the dependent variable and other predictors are the independent variables. It is important to note that the output variable is excluded in this process.
- The model built is evaluated and the respective R-square value is noted.
- The formula of VIF is as follows:

  $$VIF_i = \frac{1}{1 - R_i^2}$$

  Where VIF of i'th variable is calculated by dividing 1 by the difference between 1 and R-squared obtained.
- Now imagine if there are to predictor variables x1 and x2. These two variables are correlated with each other having r = 1. Hence, even if the linear regression equation is y = c + m1x1 + m2x2, it is difficult to know which variable is explaining the respective variation in y. Hence, if two variables have r = 1, their respective R-squared value will also be 1. In such a case the value of VIF goes infinite.
- Hence, if the value of VIF is infinite it means that the corresponding i'th variable is highly positively correlated with other predictor variables.
- **Actions to be taken if the value of VIF is infinite:**
  - Compare between the variables which one is best suit for the model from the perspective of business problem and market scenario and drop the variable which has less influence in the problem

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans:

- A quantile-quantile plot is a technique to check if two datasets come from the same population and possess the same distribution or not.
- This technique uses the concepts of graphical distribution wherein the dataset is breaken down into quantiles i.e., a fraction of value. Eg: if the quantile is 0.3 then 30% of the data below the point and 70% of the data above the point is taken into consideration.
- A 45-degree line is also plotted across the graph also known as the reference line
- If the two sets come from the same population with same distribution then the points should fall approximately near the reference line. If the distance between data points and the reference line increases then greater the evidence that the two datasets have come from populations with different data distributions
- Advantages:
  - No constraints on the sample size
  - Distributional aspects such as shift in location, shift in scale, etc can be tested
- Q-Q plots are generally used to find the distribution within the data for a random variable. It is done to check if the distribution is Gaussian, Uniform, Exponential, etc.
- Working:
  - Plot the standard theoretical quantiles on the x-axis and ordered distribution on the y-axis
  - If the points are scattered away from the line then the data doesn't follow a normal distribution, else it does.
- **Use and importance of Q-Q plot in Linear Regression:**
  - Since we know that Q-Q plot helps in concluding whether the two datasets come from the same population or not and follow the same distribution or not, it is used to test the same hypothesis on training and test datasets. Using Q-Q plots we can confirm whether the training and test datasets are from the same population distribution or not.

    Eg: if the test dataset contains 30 records and the training set contains 100 records, then it is possible to compare the distributions of these datasets and check if they are same or not
  - This helps in ensuring that the machine learning model is based on the right distribution
  - If the data points aren't doesn't lie on the line then the residuals do not follow gaussian distribution and so will the errors. This means that for such samples β estimator is invalid

- This also means that the standard confidence intervals and significance tests are invalid as well