ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

Ques1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: Effect of categorical variables on dependent variables are:

- **yr:** 20% hike in rental bikes with median count of 4000 in "2018" to median count of 6000 in "2019".
- **season:** 32% bike are used in "fall" with median count of over 5000 followed by "summer" and "winter" with 27% and 25% respectively. Potential predictor for dependent variable.
- **month:** maximum rental bikes used in "may", "june", "july", "august", "September" and "october" with median count of over 4000. Potential predictor for dependent variable.
- **workingday:** there is no such effect of working day on rental bikes as median count is almost same between 4000 & 5000. Let model decide whether it is potential predictor or not.
- **holiday:** when there is no holiday then median count is over 4000, there is some impact of holiday on count of rental bikes.
- **weekday:** each weekday has median count of 4000 or over. So this variable has some or no impact on dependent variable.
- **weathersit:** "clear" weathersit have maximum rental bikes count with median of approx. 5000 followed by "cloud/mist" with median count of 4000 and then "thunderstorm/rain" with median count of below 2000. Potential predictor for dependent variable.

Ques2: Why is it important to use **drop_first=True** during dummy variable creation?

Ans: "drop_first = True" is important to use, as it helps in reducing the extra column created during dummy variable creation. It reduces the correlations created among dummy variables.

For example: we have 4 types of value in a categorical column “**season**”: “spring”, “summer”, “fall” & “winter” and created dummy variables for that column. If one variable does not come under “summer”, “fall” and “winter” category then it is obviously “spring”. So, we don’t need fourth variable to identify “spring” category.

Hence, if we have categorical column with n-levels, then we need to use n-1 columns to represent the dummy variables.

Ques3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: Looking at the pair-plot among the numerical variables, **“temp”** has the highest correlation with the target variable as it shows a good linear relationship with target variable with value of 0.63(inferred from heatmap).

Ques4: How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: Assumptions of Linear Regression are:

- Linear relationship between X and y means $\mu = 0$.
- No multicollinearity between independent variables.
- Error terms are normally distributed.
- Error terms are independent of each other.
- Homoscedasticity means error terms have constant variance.

Multicollinearity can be checked by variables VIF value less than 5 and p-value less than 0.05.

After building the model on training set, we will do Residual Analysis, which will validate other assumptions.

First, we will get prediction values of y in training set. Then, we will calculate difference of y train and y train predicted, which gives the **residual error** values.

Residual error = y_train – y_train_pred

Now, plot “**distplot**” of residual error values and check if mean is zero, graph formed is normally distributed or not.

Then, plot “**scatterplot**” to check Homoscedasticity & independency of error terms. This will validate the assumptions of Linear Regression.

Ques5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

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

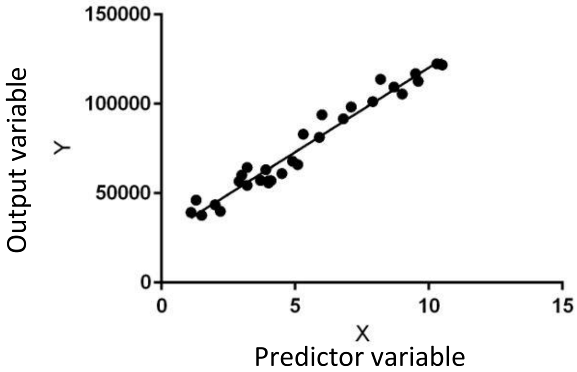
1. **temp** (temperature) with coefficient value = 0.4422 indicates positive linear relationship. A unit increase in temp, increases the bike demand by 0.4422 units.
2. **yr** (year) with coefficient value = 0.2358 indicates positive linear relationship. A unit increase in yr increases the bike demand by 0.2358 units.
3. **thunderstorm/rain** (weathersit_3) with coefficient value = -0.2280 indicates negative linear relationship. A unit increase in weathersit_3 decreases the bike demand by 0.2280 units.

GENERAL SUBJECTIVE QUESTIONS

Ques1: Explain the linear regression algorithm in detail.

Ans: Linear Regression Algorithm is a Machine Learning Algorithm based on supervised learning. Linear regression is a part of regression analysis. Regression analysis is a technique of predictive modelling that helps in finding relationship between several independent variables and dependent variable.

- Independent variables are known as **Predictor variables**.
- Dependent variable is known as **Output Variable**.



Linear regression is one of the basic forms of machine learning where we train a model to predict the behavior of a dependent/output(y) variable in data based on some independent/predictor variables(X). In linear regression, two variables should be linearly correlated which are on X-axis and y-axis.

Example: You are running a sales promotion and expecting a certain number of count of customers to be increased. Now you can look previous promotions and try to see whether there is increment into the number of customers when you rate the promotion. With help of previous historical data try to estimate count for current promotion. This will give you an idea to do the planning in a much better way. Here is the idea to estimate future value based on historical data by learning behavior or patterns from historical data.

In some cases, value will be linearly upward which means whenever X increases, y also increases means Positive Linear Relationship or vice versa means they have correlation or linear downward relationship means Negative Linear Relationship.

There are two types of Linear Regression:

- **Simple Linear Regression:** If there is only one predictor/independent variable then we use simple linear regression.

  Mathematical expression for simple linear regression is:

$$\mathbf{y = \beta_0 + \beta_1 X}$$

where,

y = Output variable

X = predictor variable

β_0 = intercept

β_1 = coefficient of X variable

- **Multiple Linear Regression:** If there is more than one predictor/independent variable the multiple regression is used.

  Mathematical expression for multiple linear regression is:

$$\mathbf{y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots + \beta_n X_n}$$

where,

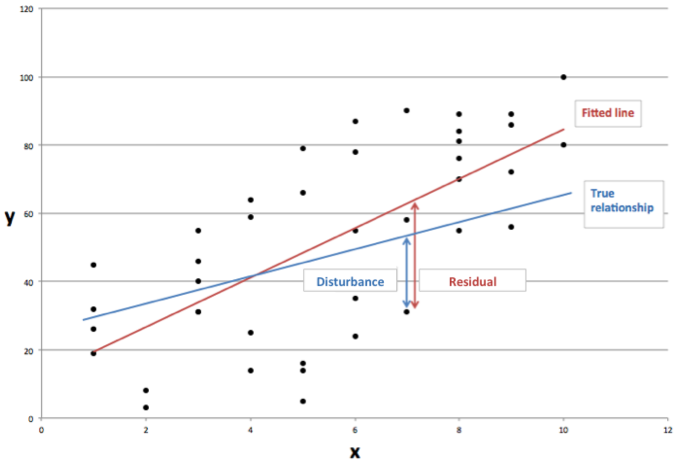
y = Output variable,

β_0 = intercept or constant,

$\beta_{1,2,..,n}$ = variables coefficients or slope,

$X_{1,2,..,n}$ = independent variables

The goal of linear regression algorithm is to get the best values for β_0 & $\beta_{1,.,n}$ to find the **Best Fit Line** and Best Fit Line should have least error.



Best-Fit Line

To calculate Best Fit Line, we use **Cost Function**. Cost Function optimizes the coefficients and used to find accuracy of mapping function that maps input variable to the output variable. **Mean squared Error** (MSE) cost function is used to find best possible values for β_0 & $\beta_{1,.,n,}$ which provides best fit line for data points.

MSE formula: $(1/n) * \Sigma(y_i - \bar{y}_i)^2$ where,

Y_i = actual values of y & **$\bar{y}_i$** = predicted values of y

Using MSE function we change the value of β_0 & $\beta_{1,.,n}$ such that the MSE value settles at minima. **Gradient Descent** method is used for updating β_0 & $\beta_{1,.,n}$ to minimize Cost Function (MSE).

Gradient Descent Formula: $\Theta_j - (\alpha/n) * \Sigma\ [(h_\Theta(x_i) - y_i)\ x_i]$ where,

Θ_j = wights of hypothesis,

α = Learning rate of gradient descent,

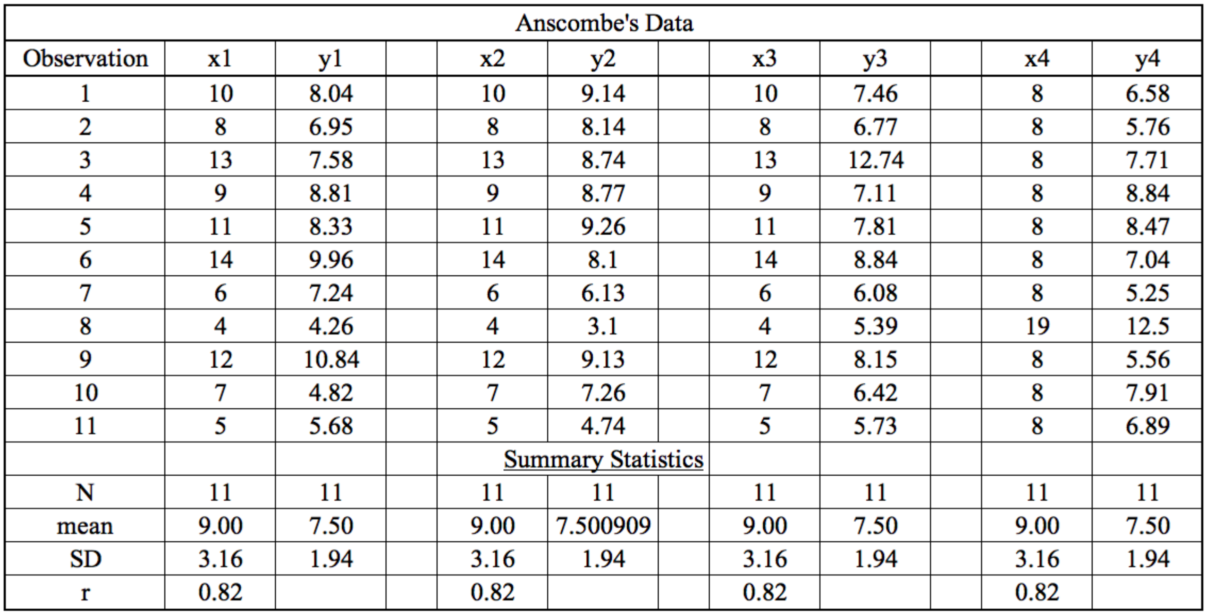
$h_\Theta(x_i)$ = predicted values of y

Ques2: Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet can be defined as the group of 4 data sets which are identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have different distributions and appear differently when plotted on scatter plots. Each dataset consists of 11 (x,y) points.

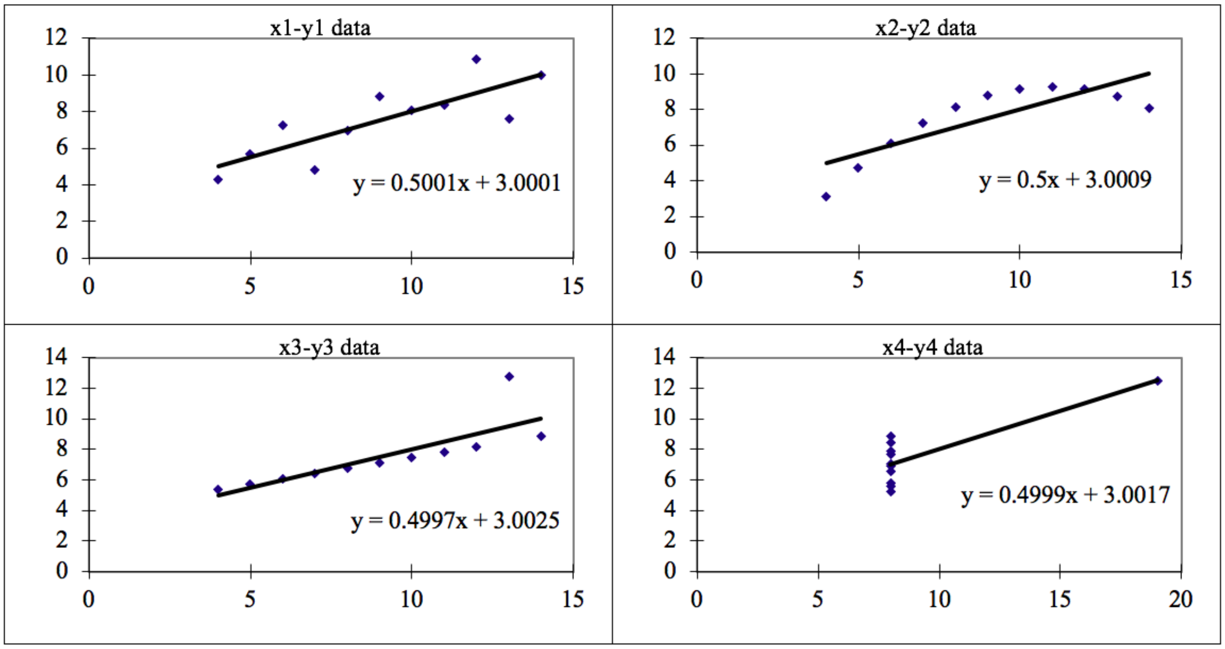
They were constructed by statistician Francis Anscombe to demonstrate both, importance of graphing data before analyzing it and the effect of outliers on statistical properties.

This tells us about the importance of visualizing the data, before applying various algorithms, to build models, 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. In addition, the Linear Regression only be considered a fit for the data with linear relationships and is incapable of handling other kind of datasets. These four plots defined as follows with summary statistics:



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	

When these models plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm, which is fooled by these peculiarities and can be seen as follows:



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

Hence, all the important features visualized in the dataset before implementing machine-learning algorithm, which will help to make a good fit model.

Ques3: What is Pearson's R?

Ans: Pearson's R or Pearson correlation coefficient is a measure of linear correlation between variables. It is the ratio between the covariance of two variables by product of their standard deviations. It is essentially a normalized measurement of covariance, such that result values lies between -1 and 1. It cannot capture non-linear relationships between two variables and cannot differentiate between dependent and independent variables.

There are certain requirements for Pearson's R:

- Scale of measurement should be interval or ratio.
- Variables should be approximately normally distributed.
- The association should be linear.
- There should be no outliers in the data.

Formula of Pearson's R:

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2- (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where,

N = number of paired values,

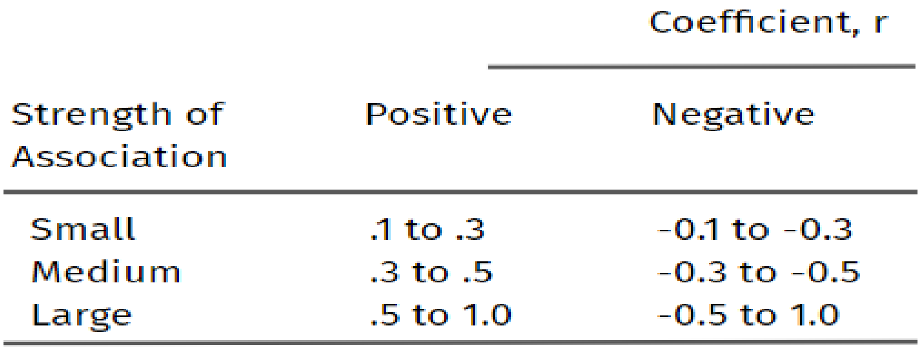
Σxy = sum of product of paired values,

Σx = sum of x values, Σy = sum of y values

Σx^2 = sum of squared x values, Σy^2 = sum of squared y values

The more inclined the value of Pearson’s R to -1 and 1 the stronger is the relationship between two variables.

Below are guidelines to interpret the Pearson’s R:



	Coefficient, r	
Strength of Association	Positive	Negative
Small	.1 to .3	-0.1 to -0.3
Medium	.3 to .5	-0.3 to -0.5
Large	.5 to 1.0	-0.5 to 1.0

A notable point is that strength of variables depend on sample size.

Ques4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: Scaling is a method used to normalize the range of independent variables or features of data. It is generally performed during data preprocessing step. It also helps in speeding up the calculations in algorithms.

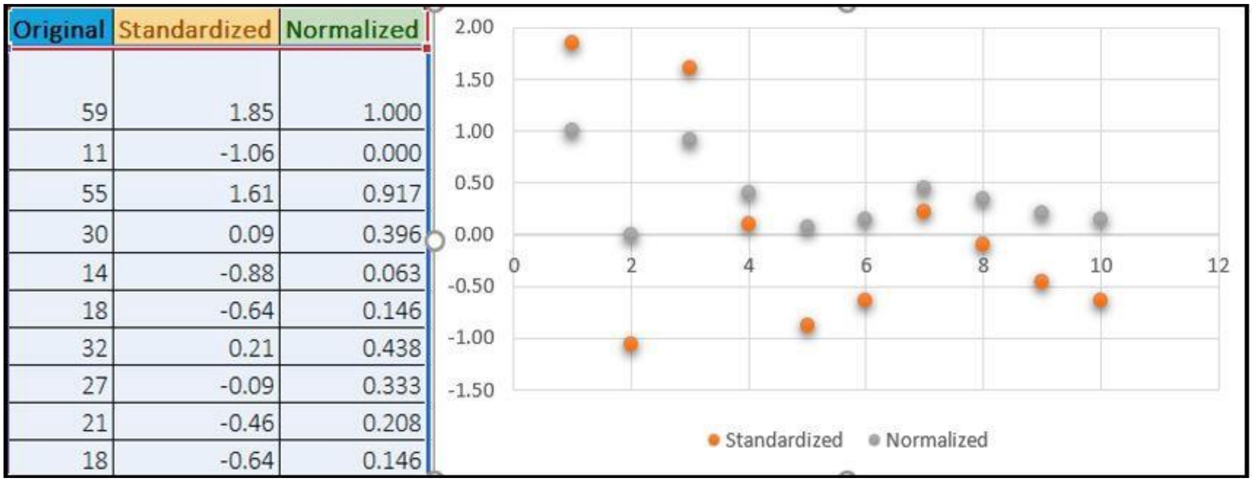
Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done, then the algorithm only takes magnitude in account and not units, which results in incorrect modelling. To solve this issue, scaling is done to bring all variables to the same level of magnitude. It is used in faster convergence for gradient descent method.

It is important to note that scaling just affects the coefficients and none other parameters like t-statistic, f-statistic, p-values, R-squared, etc.

There are two types of scaling:

1. **Min-Max Scaling:** Also known as **normalization**. It brings all the data in the range of 0 and 1.
   MinMax Scaling: x = (x - min(x)) / (max(x) - min(x))

2. **Standardization:** It replaces the values by their z-scores. It brings all of the data into a standard normal distribution which has mean (μ) = 0 and standard deviation (σ) = 1.
   Standardization: x = (x – mean(x)) / sd(x)



Original	Standardized	Normalized
59	1.85	1.000
11	-1.06	0.000
55	1.61	0.917
30	0.09	0.396
14	-0.88	0.063
18	-0.64	0.146
32	0.21	0.438
27	-0.09	0.333
21	-0.46	0.208
18	-0.64	0.146

Example of Standardized and Normalized scaling on original values.

Normalization is used when distribution of data does not follow Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-nearest.

Standardization can be useful in cases where data follows a Gaussians distribution. However, this does not have to be necessarily true. In standardization, values are not restricted to a particular range so if you have outliers in your data, they will not be affected by it.

Ques5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: Multicollinearity refers to the problem when independent variables are collinear. Collinearity refers to a linear relationship between two explanatory variables.

Two variables are perfectly collinear if there is an exact relationship between two variables. If the independent variables are perfectly collinear, then our model becomes singular and it would be possible to uniquely identify the model coefficients mathematically.

VIF (Variance Inflation Factor) is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables. A large value of VIF indicates that there is a correlation between the variables.

If there is perfect correlation, the VIF is infinite. This shows a perfect correlation between two independent variables. In a case of perfect correlation, we get R-Squared = 1, which lead to $1/(1 – R^2)$ infinity.

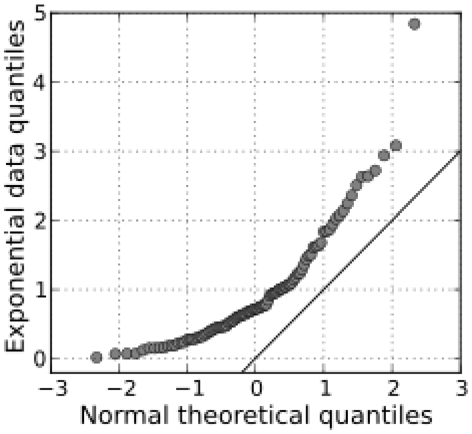
To solve this problem we need to drop one of the variables from dataset that is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

Ques6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans: A quantile is a fraction where certain values fall below that quantile. Q-Q (Quantile-Quantile) plot is a graphical tool to help assess if a set of data came from some theoretical distribution such as Normal exponential or uniform distribution. It also determines if two data sets come from populations with a common distribution.

The purpose of Q-Q plot is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on Q-Q plot, if two data sets come from a common distribution, the points will fall on that reference line.



Q-Q plot showing 45-degree reference line

If two distributions being compared are similar, the points in the Q-Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q-Q plot will approximately lie on a line, but not necessarily on the line y = x. Q-Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q-Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale and skewness are similar or different in the two distributions.

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.

Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and presence of outliers can all be detected from this plot.

It is used to check following scenarios: If two data sets —

- come from populations with a common distribution.
- have common location and scale.
- have similar distributional shapes.
- have similar tail behavior.