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?**

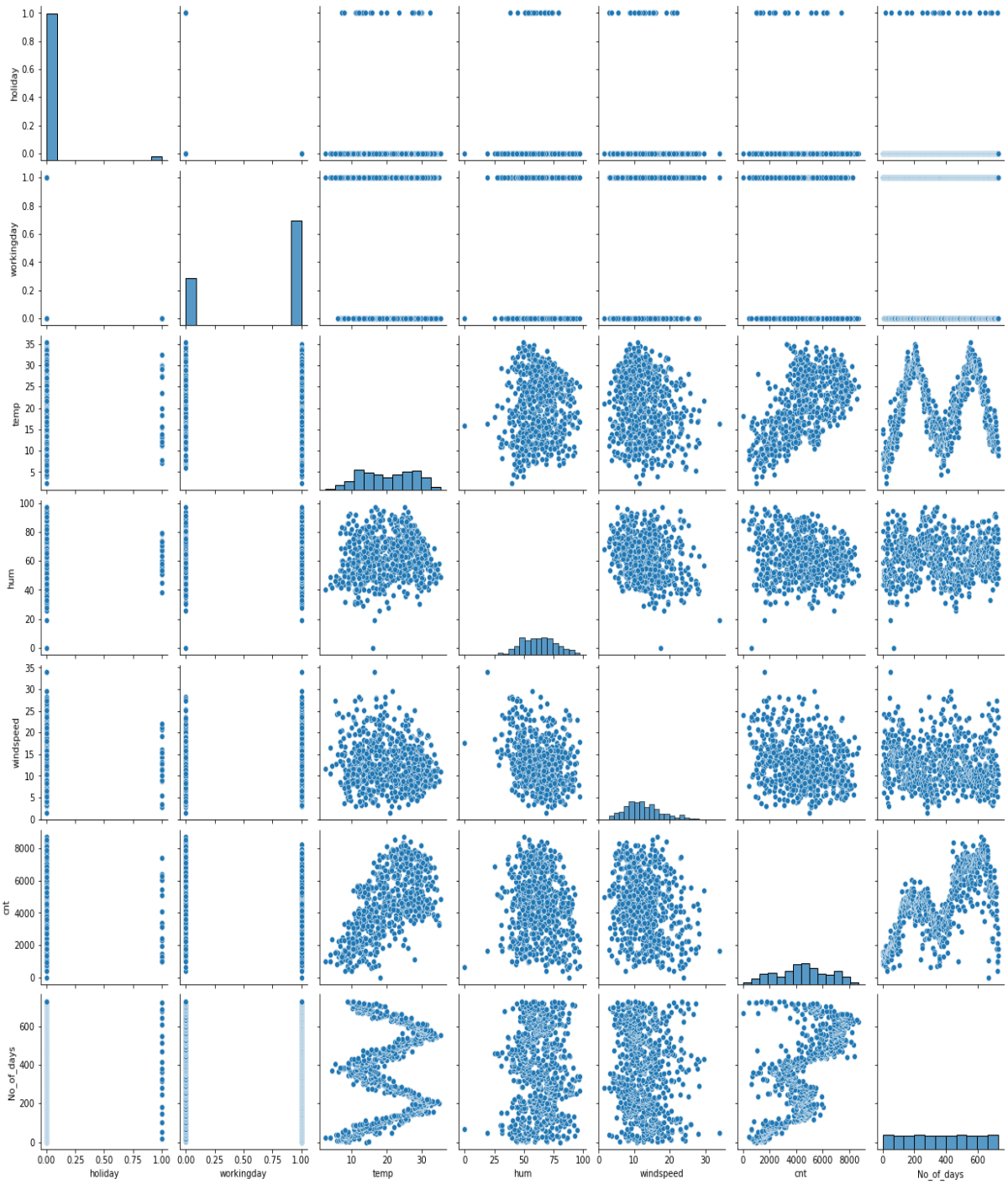
    - Fall Season has the highest demand for rental bikes followed by summer season.
    - Clear weather is most optimal for renting a bike, as temperature and humidity is less.
    - Year 2019 has the highest demand compared to 2018, it might be due to the fact that bike rentals are getting popular and people are aware of it.
    - Working and Non-Working days have almost the same median although spread is high for non-working days as people might have plans and do not want to rent bikes.
    - People rent bikes more on non-holidays compared to holidays, so reason might be they prefer to spend time with family.
    - Median across all days is same but spread for Saturday and Friday is high, may be evident that those who have plans for Saturday might not rent bikes as it is a non-working day.
    - Demand is continuously growing each month till June. September month has highest demand. After September, demand is gradually decreased.

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

    - A variable with 'n' values can be represented with 'n-1' dummy variables. So, if we remove the first column then also, we can represent the data.
    - For example, if we take season variable from the dataset, it has four values (spring, summer, fall and winter). Three dummy variables are enough to represent the data.
        - 000 will correspond to fall
        - 001 will correspond to winter
        - 010 will correspond to summer
        - 100 will correspond to spring
    - Hence the first column 'fall' can be dropped and from the table if the values for all the three dummy variables are zero then t is considered as fall
    - **Importance**: To avoid multicollinearity (if we don't drop, dummy variables will be correlated) and affects the model adversely.
    - We use drop_first=true to avoid redundant features.

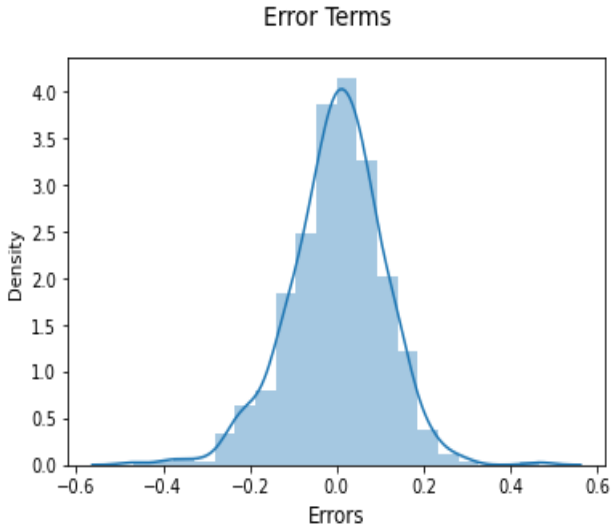
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

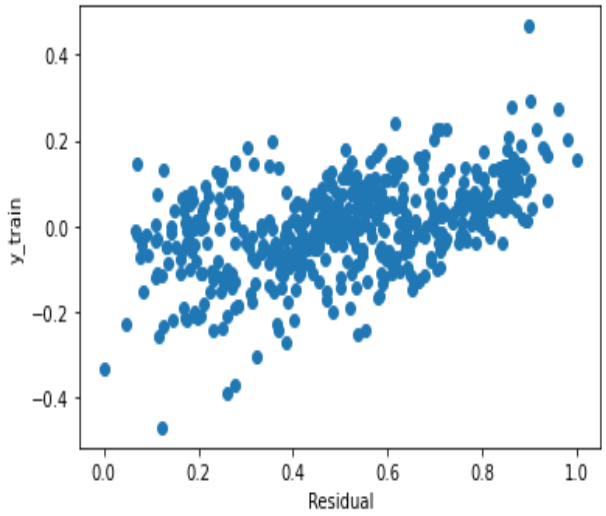
- Count (Target Variable) has significantly high correlation with temperature (temp) and No_of_days.



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

- Residual errors follow normal distribution with mean zero.
- No Multicollinearity
- No Heteroskedasticity
- No autocorrelation in residuals
- Linear relationship between dependent and independent variables





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

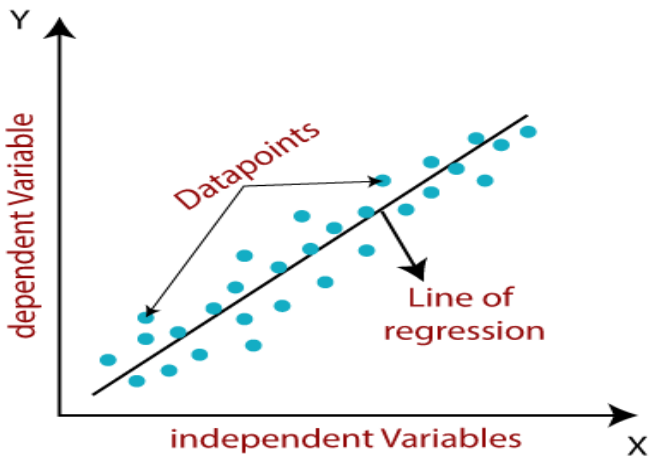
    - yr_2019 has the coefficient of 0.2470
    - mnth_Sep has the coefficient of 0.0723
    - weathersit_Light Snow has the negative correlation.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

    Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. It is mostly used for finding out the relationship between variables and forecasting. Regression models a target prediction value based on independent variables.

    Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.



y=b0+b1X+e

where y = dependent variable (target variable)

b0 = intercept of the line

b1 = Linear Regression Coefficient

X = independent variable (predictor variable)

e = random error

Linear Regression can be further divided into two types of algorithms.

- **Simple Linear Regression:** If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
- **Multiple Linear Regression:** If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Linear Regression Line: A linear line showing the relationship between the dependent and independent variables is called a regression line.

When working with linear regression, **our main goal is to find the best fit line** that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.

The different values for weights or the coefficient of lines (b0, b1) gives a different line of regression, so we need to calculate the best values for b0 and b1 to find the best fit line, so to calculate this we use cost function.

Cost function: Cost function is used to estimate the values of the coefficient for the best fit line. It optimizes the regression coefficients or weights. It measures how a linear regression model is performing.

We can use the cost function to find the accuracy of the mapping function, which maps the input variable to the output variable. This mapping function is also known as Hypothesis function.

we use the **Mean Squared Error (MSE)** cost function, which is the average of squared error occurred between the predicted values and actual values.

MSE = 1/N Sum(i=1 to n) (yi-ypred)^2

Where N = Total number of observations

Yi = Actual Value

Ypred = Predicted value

Residuals: The distance between the actual value and predicted values is called residual. If the observed points are far from the regression line, then the residual will be high, and so cost function will high. If the scatter points are close to the regression line, then the residual will be small and hence the cost function.

Gradient Descent: Gradient descent is used to minimize the MSE by calculating the gradient of the cost function. A regression model uses gradient descent to update the coefficients of the line by reducing the cost function. It is done by a random selection of values of coefficient and then iteratively update the values to reach the minimum cost function.

Model Performance: The Goodness of fit determines how the line of regression fits the set of observations. The process of finding the best model out of various models is called optimization. It can be achieved by R-squared method.

Assumptions of Linear Regression

- Linear Relationship between independent variable and dependent variable
- Error terms are normally distributed with mean zero
- Error terms are independent of each other
- Error terms have constant variance

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

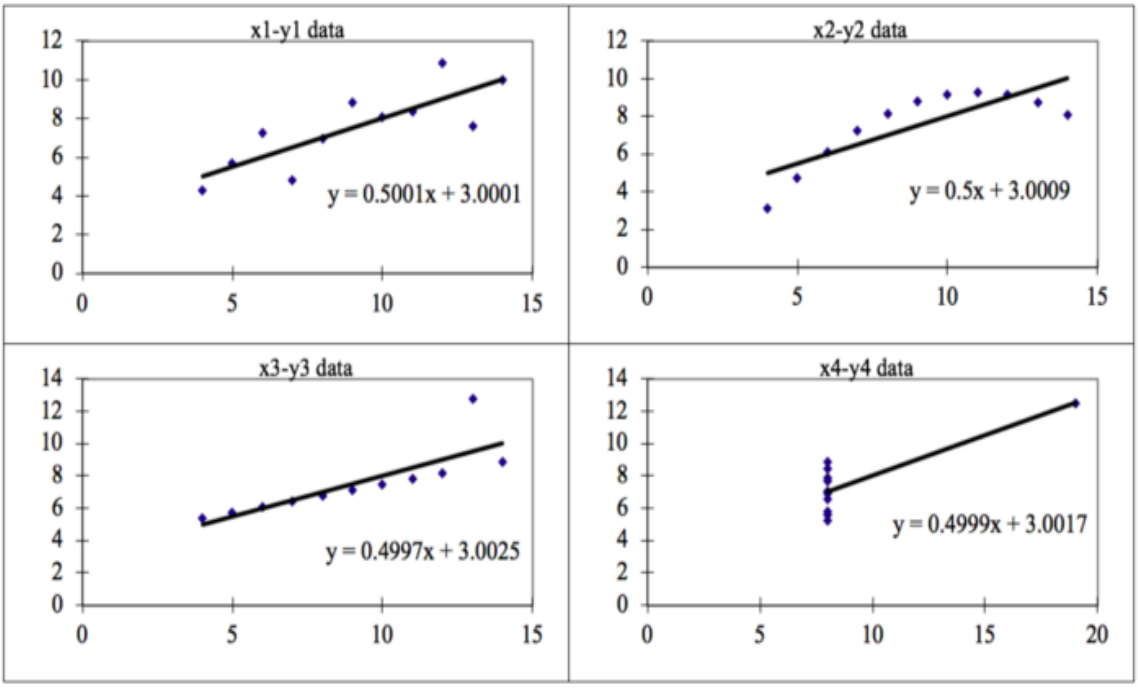
Example: Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.

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

Statistical information for all these four datasets is approximately similar and can be computed 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	

When these models are plotted on a scatter plot, all datasets generate 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:



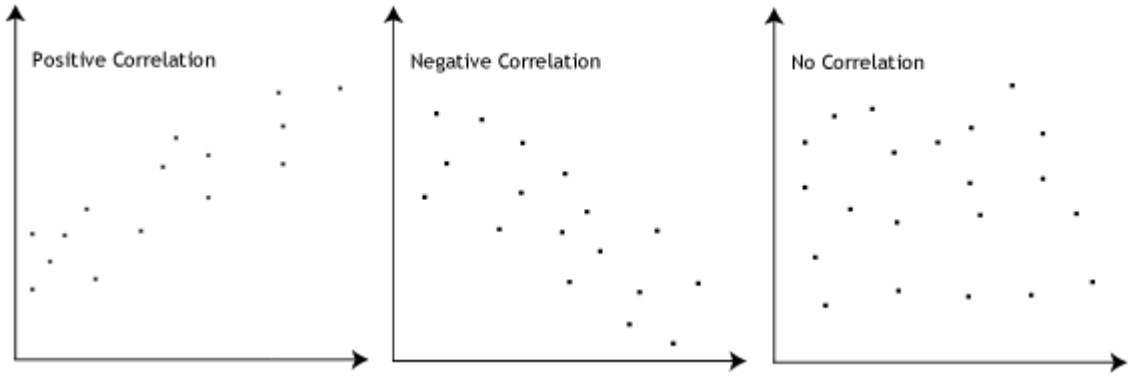
The four datasets can be described as:

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

3. What is Pearson's R?

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.
The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association



Pearson r Formula

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

where r = correlation coefficient
xi = values of the x-variable in a sample
x = mean of the values of the x-variable
yi = values of the y-variable in a sample
y = mean of the values of the y-variable

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

Scaling: It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

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

Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

S.No.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.

4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMax Scaler for Normalization.	Scikit-Learn provides a transformer called Standard Scaler for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is often called as Scaling Normalization	It is a often called as Z-Score Normalization.
9.	X= (x-min(x))/(max(x) – min(x))	X= (x-mean(x))/sd(x)

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

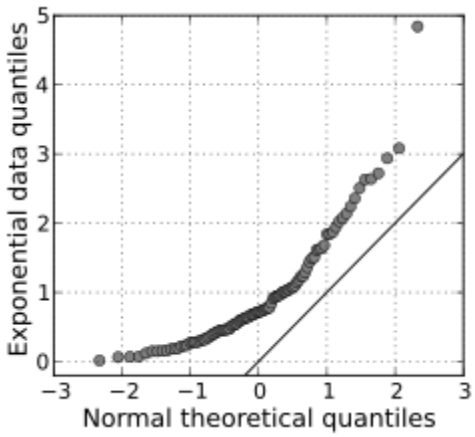
If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this 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

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of QQ plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the QQ plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A QQ plot showing the 45-degree reference line:



If the 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.