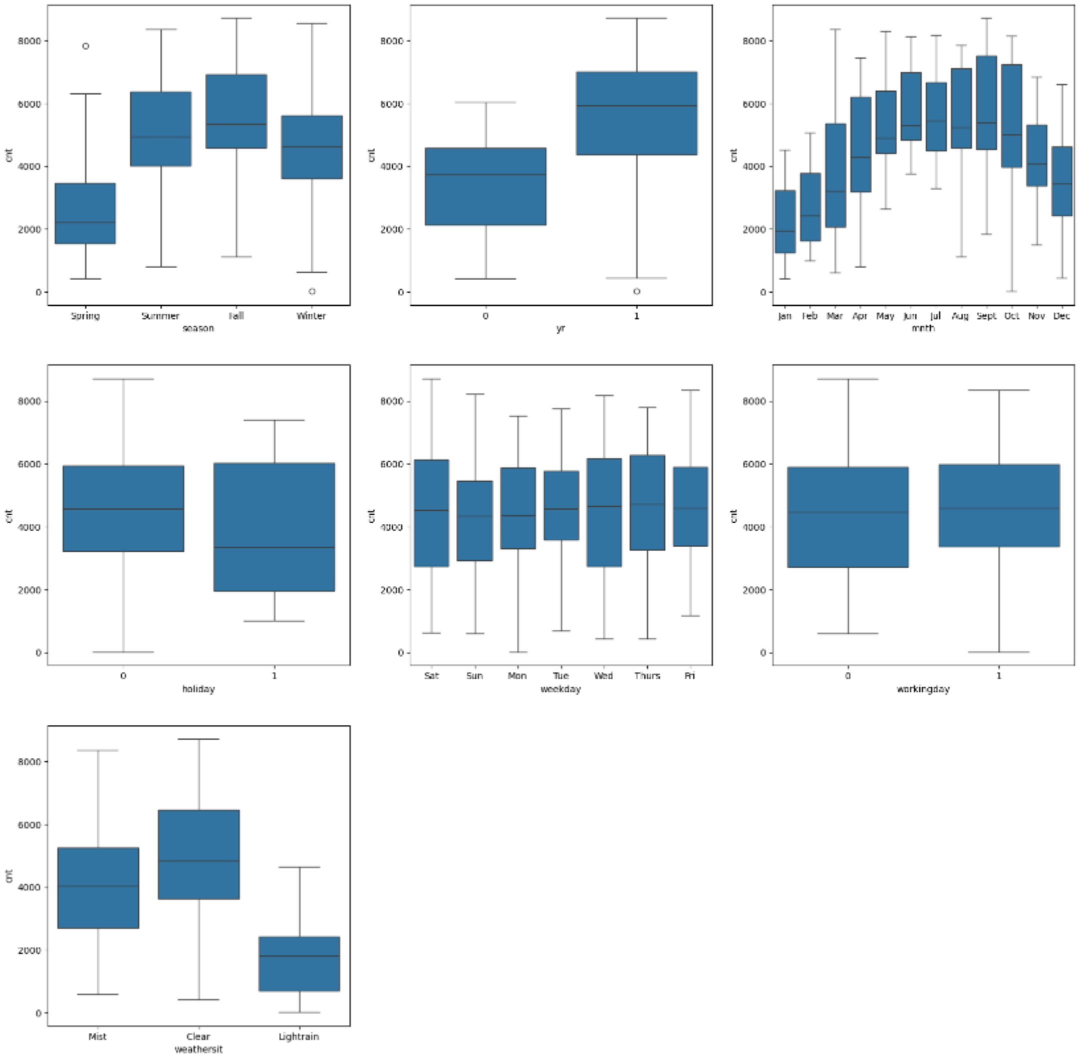
Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

- **Season**: Rental counts tend to be highest in fall and summer, with lower counts in winter and spring.
- **Year**: Rentals generally increased in 2019 compared to 2018.
- **Month**: There's a seasonal trend with higher rentals around mid-year (June to September) and lower counts at the beginning and end of the year.
- **Holiday**: Rental counts are slightly lower on holidays compared to non-holiday days.
- **Weekday**: Rental count is overall same across all days of the week.
- **Working Day**: Rental count is not much affected by working or non-working days.
- **Weather Situation**: Clear, Few clouds, Partly cloudy, Partly cloudy days have the highest rental counts, while days with Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog have lower counts, especially in severe weather conditions.



From the above analysis we can infer that that holiday, weekday and working day do not seem to be good differentiators of the dependent variable.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

drop_first=True removes the first column of the dummy variable.
N-1 dummy variable can be used to describe a categorical variable with N levels.
E.g: The categorical variable ‘season’ with four categories (spring, summer, fall, winter), creating dummies without drop_first=True would produce four columns.

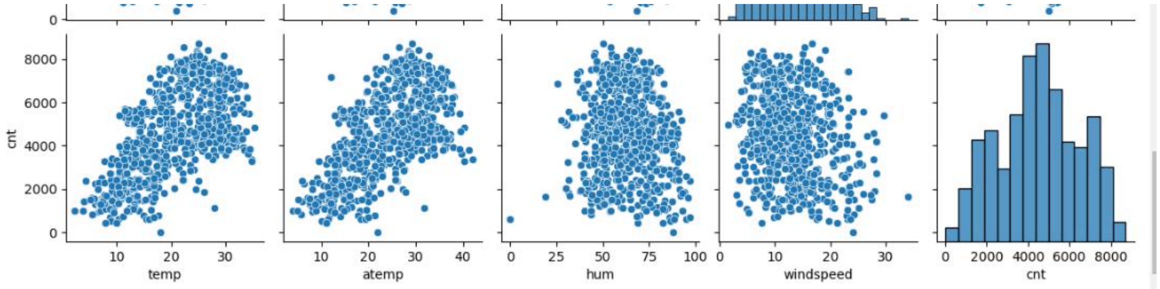
Season_Spring	Season_Summer	Season_Fall	Season_Winter
1	0	0	0
0	0	1	0
0	0	0	1

But we can interpret Season_Spring is 1 on 0 based on the other variables. Hence only 4-1 are required to explain the four categories.

This will also help us to reduce multicollinearity.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

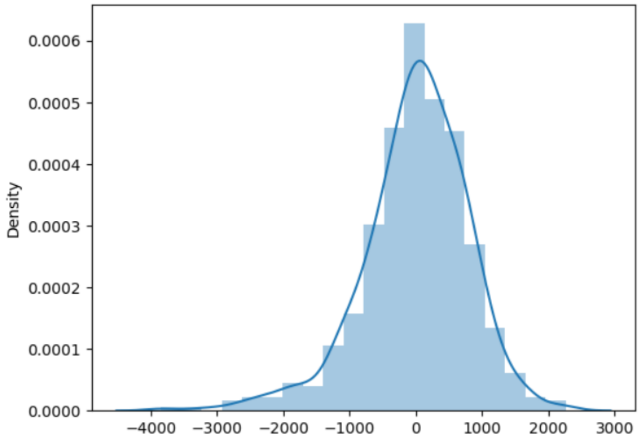
Looking at the pairplot we can say temp and atemp variables have the highest correlation with the target variable.



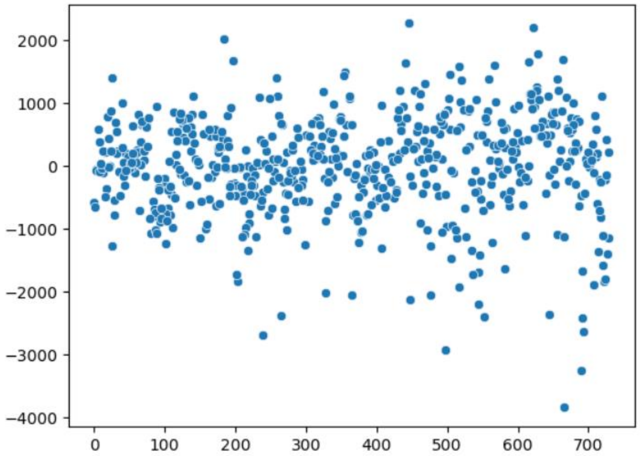
Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

We validate the assumption of linear regression by doing residual analysis on the training data set

Creating a barplot for all the error terms and checking if they are normally distributed with mean=0



Created a scatter plot of the error terms to check if the error terms are independent of each other and have a constant variance.



Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

All the features selected in the model are significant as they have a p-value of 0.
However top three features which are most significant can be decided based on the coefficient as all the features were scaled between 0 to 1 .

1. Temperature (temp)
2. weathersit (weathersit_Lightrain)
3. Year (yr)

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Linear regression is a supervised learning algorithm used for predicting a continuous outcome variable (dependent variable) based on one or more predictor variables (independent variables). The algorithm models the relationship between the dependent and independent variables by fitting a linear equation to the observed data. The basic idea is to find a line (or a hyperplane for multiple variables) that best describes the relationship between the variables.

Steps in Linear Regression:
Defining the Model: For a simple linear regression with one independent variable, the model is:

$$y = \beta_0 + \beta_1 x + \epsilon$$

where:

y is the dependent variable,
x is the independent variable,
$\beta 0$ is the y-intercept,
$\beta 1$ is the slope of the line (change in y for a unit change in x),
ϵ represents the error term, capturing the difference between predicted and actual values.

For multiple linear regression with multiple independent variables, the model becomes:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

Objective: The primary goal of linear regression is to find the values of $\beta 0$, $\beta 1$, $\beta 2$, etc., that minimize the difference between predicted and actual values of. This is achieved by minimizing the sum of squared residuals (least squares method), which measures the squared differences between the actual and predicted values.

Calculating Coefficients: The regression coefficients are estimated using methods like the Ordinary Least Squares (OLS), which aims to minimize the sum of the squared differences between observed values and predicted values.

Model Evaluation: After fitting the model, it's evaluated to check its accuracy and effectiveness. Key metrics for this include:

1. R-squared: Measures the proportion of variance in the dependent variable explained by the model. A higher R-squared indicates a better fit.
2. Mean Squared Error (MSE): The average of squared differences between actual and predicted values.
3. Adjusted R-squared: Adjusts R-squared for the number of predictors, particularly useful in multiple regression.

Assumptions:

1. Linearity: Assumes a linear relationship between the dependent and independent variables.
2. Independence of errors: Residuals should not be correlated.
3. Normality of errors: Residuals should follow a normal distribution.
4. Homoscedasticity: The variance of residuals should be constant across all levels of the independent variables.
5. No multicollinearity: Independent variables should not be highly correlated with each other.

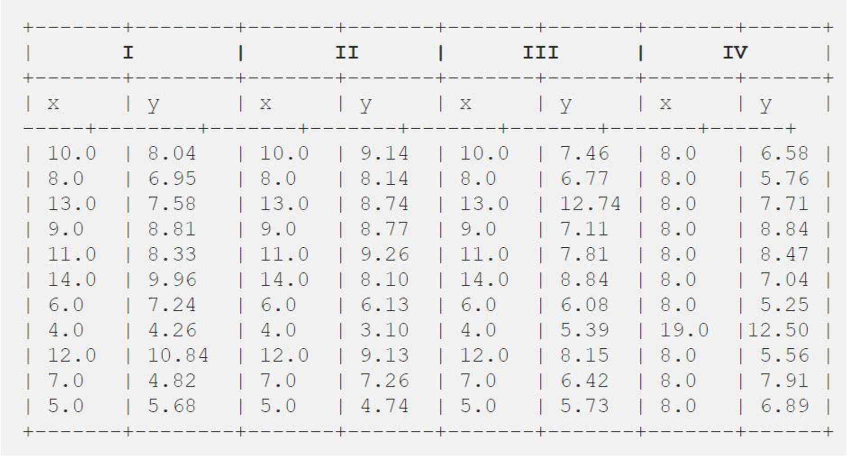
Application: Linear regression is widely used for predictive modeling, trend forecasting, and analyzing the relationship between variables across various domains such as finance, economics, and healthcare. The simplicity of the model makes it easy to interpret while providing a good baseline model for further complex analyses.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.
The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.
The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient,

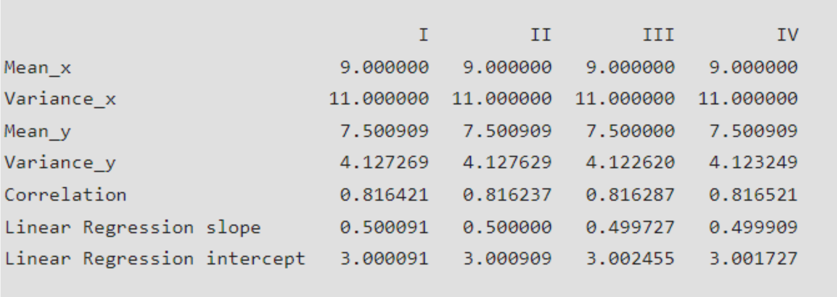
and linear regression line.

The four datasets of Anscombe’s quartet -



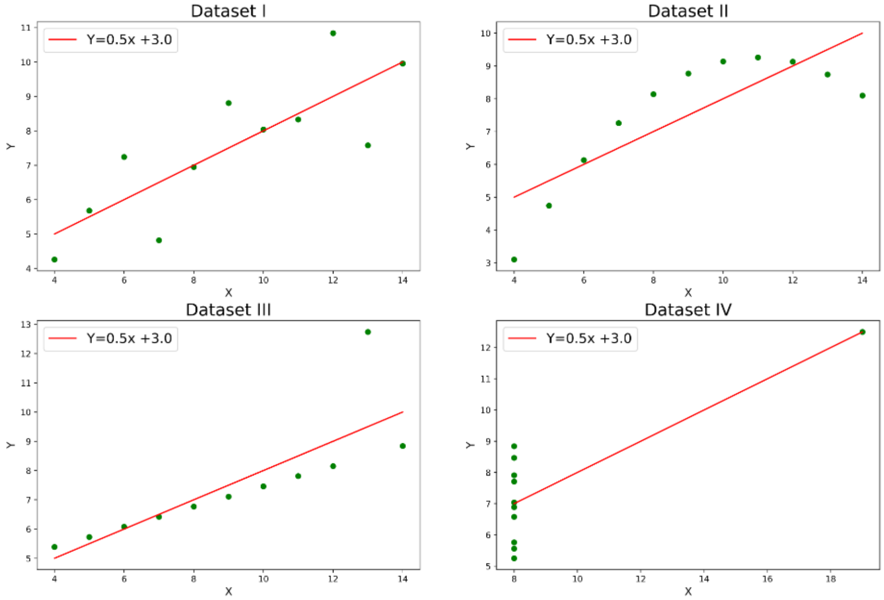
```
+-------+--------+-------+-------+-------+-------+-------+------+
|       I        |      II       |      III      |      IV      |
+-------+--------+-------+-------+-------+-------+-------+------+
| x     | y      | x     | y     | x     | y     | x     | y    |
-----+--------+--------+--------+--------+-------+--------+------+
| 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 |
+-------+--------+-------+-------+-------+-------+-------+------+
```

On taking the descriptive statistics of these 4 datasets, we can see that these statistics turn out to be identical leading one to believe that the datasets are essentially the same.



```
                                    I          II         III          IV
Mean_x                       9.000000    9.000000    9.000000    9.000000
Variance_x                  11.000000   11.000000   11.000000   11.000000
Mean_y                       7.500909    7.500909    7.500000    7.500909
Variance_y                   4.127269    4.127629    4.122620    4.123249
Correlation                  0.816421    0.816237    0.816287    0.816521
Linear Regression slope      0.500091    0.500000    0.499727    0.499909
Linear Regression intercept  3.000091    3.000909    3.002455    3.001727
```

These four datasets that have nearly identical simple statistical properties appear very different when graphed.



- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.

- In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

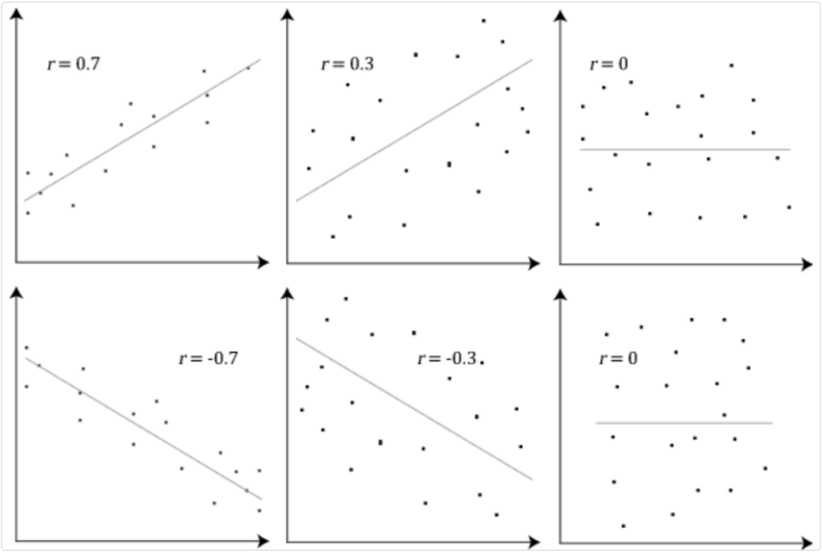
This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's R, also known as the Pearson correlation coefficient, is a statistic that measures the linear correlation between two variables. It indicates both the strength and direction of the linear relationship. The value of Pearson's R ranges from -1 to 1, where:
R=1: Perfect positive linear relationship (as one variable increases, the other also increases proportionally).
R=−1: Perfect negative linear relationship (as one variable increases, the other decreases proportionally).
R=0: No linear correlation (no linear relationship between the variables).

Pearson's R is calculated as:

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



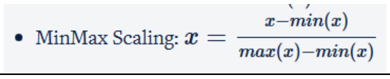
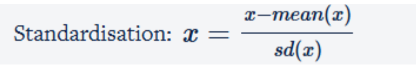
Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Scaling ensures that all features in your data have a similar range. This is important because some machine learning algorithms are sensitive to the scale of the data. For instance, if one feature has a very large range compared to others, it can dominate the model and make it difficult to learn from other features.

Some of the benefits of scaling are-

- Improves the performance of machine learning algorithms: By scaling the features, you can ensure that all features contribute equally to the model, leading to better performance.
- Reduces the risk of overfitting: Overfitting occurs when a model memorizes the training data too well and fails to generalize to unseen data. Scaling can help reduce the risk of overfitting by making the features more comparable.
- Speeds up the training process: Some machine learning algorithms, such as gradient descent, can converge faster when the features are scaled.

There are two common methods of scaling: normalization and standardization.

Sr. No.	Normalization	Standardization
1	This method scales the model using minimum and maximum values.	This method scales the model using the mean and standard deviation.
2	MinMax scaling,brings all of the data in the range of 0 and 1.	Standardisation brings all of the data into a standard normal distribution with mean zero and standard deviation one. .
3	Normalization is useful when you need to scale features to a specific range as in image processing	Standardization,is often better for distance-based algorithms and dimensionality reduction (e.g., clustering, PCA)
4	Helpful when feature distribution is unclear	Helpful when feature distribution is consistent
5	Formula –  • MinMax Scaling: $x = \frac{x-min(x)}{max(x)-min(x)}$	Formula –  Standardisation: $x = \frac{x-mean(x)}{sd(x)}$

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Perfect multicollinearity is the most common reason for infinite VIF. It occurs when one predictor variable can be expressed as an exact linear combination of one or more other predictor variables. For example, if you have two variables X1 and X2 and X2 = 2 * X1, then the variance of the estimated coefficients will inflate to infinity, leading to an infinite VIF for one or both variables.
In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) to be infinity.
To solve this, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
Addressing multicollinearity is crucial for obtaining reliable estimates and making valid inferences in regression analysis.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

A Q-Q plot, or quantile-quantile plot, is a graphical tool used to assess if a dataset follows a particular theoretical distribution, typically the normal distribution. It plots the quantiles of the sample data against the quantiles of the specified theoretical distribution.

Use of Q-Q plot –
1. Assessing Normality of Residuals:
The primary use of a Q-Q plot in linear regression is to check whether the residuals are normally distributed. This is a key assumption of linear regression that underpins the validity of statistical inference (e.g., hypothesis testing, confidence intervals).

2. **Model Diagnostics:**
Q-Q plots can be used as a diagnostic tool to evaluate the performance of the regression model. By plotting the residuals against the theoretical quantiles, analysts can visually inspect whether the model is capturing the underlying data structure accurately.

3.Identifying Outliers:
The Q-Q plot can help identify outliers in the residuals. Points that deviate significantly from the straight line indicate potential outliers that could influence the model's performance and accuracy.

4. Guiding Model Refinement:
If the Q-Q plot shows that the residuals do not follow a normal distribution, it may prompt analysts to consider transforming the response variable, using different modeling techniques, or incorporating additional predictors to improve the model.

Importance of Q-Q plot –

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.