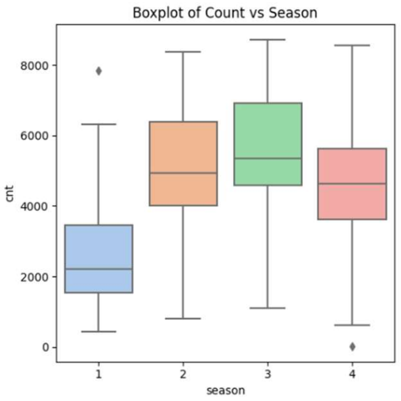
Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

Answer:

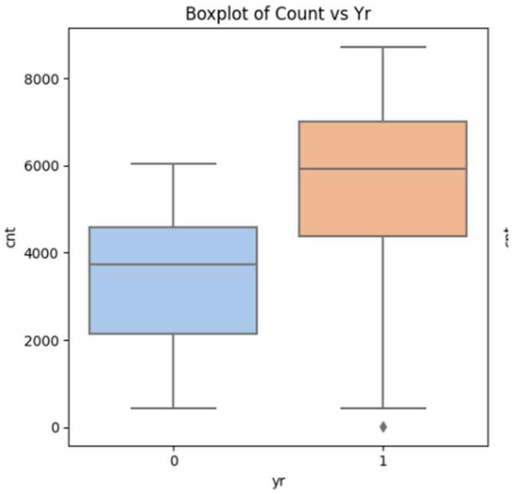
Variable 1: Season

There are 4 seasons: season 1 (spring), season 2(summer), season 3 (fall), season 4 (winter).



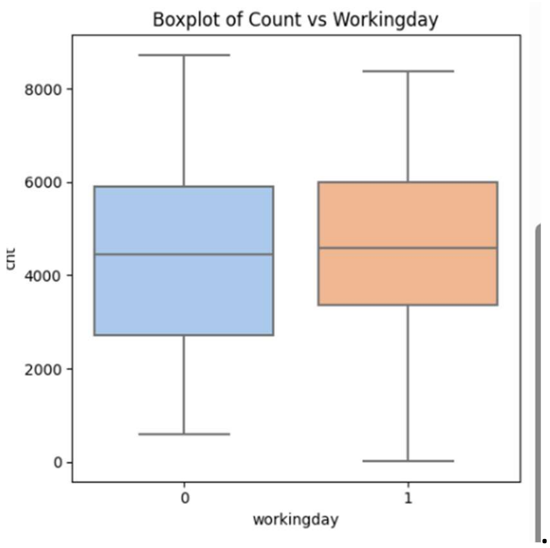
From the box plot, it can be seen that demand for bikes is less for spring, then increases for summer with maximum for fall season and then again decreases for winter.

Variable 2: year



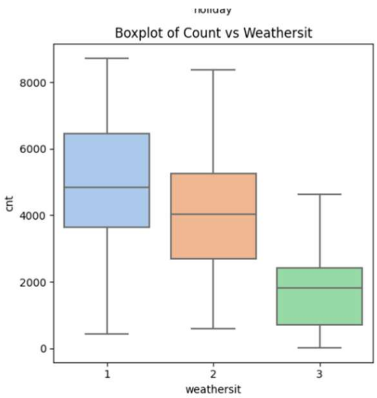
The bike demand for year 2019 is greater than year 2018.

Variable 3: working day



The bike demand for working day is greater than that of holiday.

Variable 4: weather_sit



1: Clear, Few clouds, partly cloudy, Partly cloudy

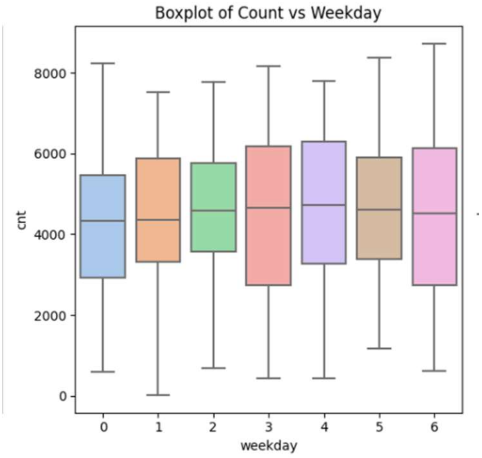
2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

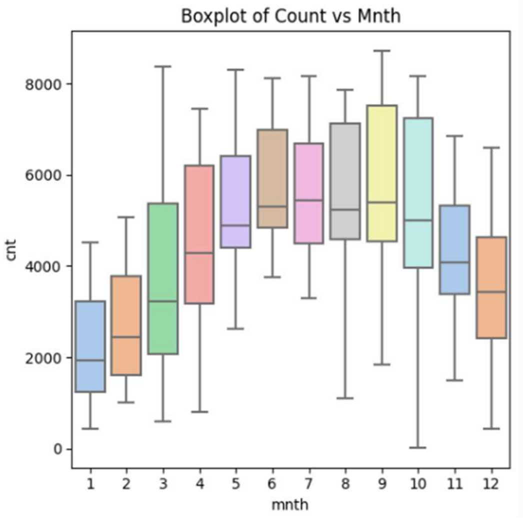
It can be seen that demand of bikes is highest for clear or parly cloudy weather while minimum for weather with light snow/light rain + thunderstorm. Worsening of weather condition decreases the demand for bikes.

Variable 5: weekday



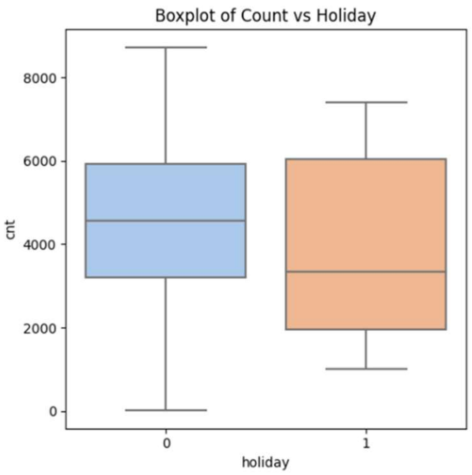
The count of registered bikes is almost same for all days of week.

Variable 6: Month



The demand for bikes is minimum for month of January and maximum for month of June. The demand gradually increases from January to June, then it more or less remains constant from July to September while again shows a decreasing trend from October till December.

Variable 7: Holiday



It can be seen that demand for bike is less for holidays than that when it is not an holiday.

Q2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

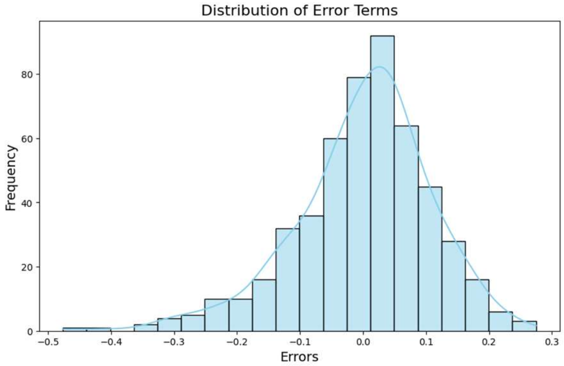
Using **drop_first=True** during dummy variable creation is important to prevent multicollinearity in regression models. Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated. This can lead to issues in interpreting the individual effects of each predictor, as well as instability in the estimated coefficients. When creating dummy variables for categorical variables with **k** levels, it is common to create **k-1** dummy variables. This is because if you have **k** dummy variables, the information from the **k**-th dummy variable can be perfectly predicted from the information in the first **k-1** dummies. This creates a multicollinearity problem. By setting **drop_first=True**, we are essentially excluding one of the levels of the categorical variable from the model. This eliminates the issue of perfect multicollinearity and ensures that the model can be estimated without encountering problems.

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

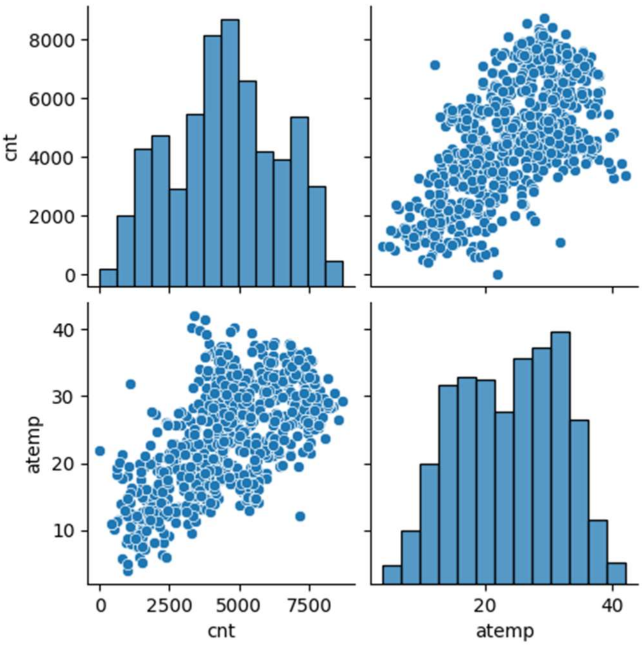
From the pair plot, we can see that temp and atemp has highest correlation with “cnt” variable.

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

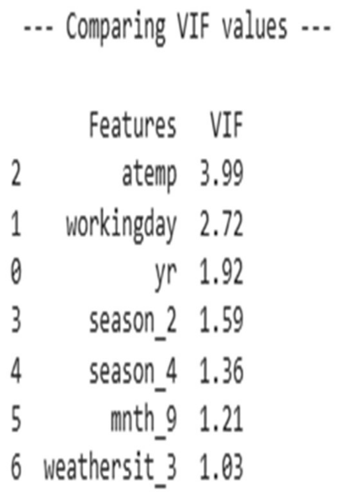
1. Errors are normally distributed with mean around zero.



2.Atemp and temp are linearly related with target variable “cnt”.

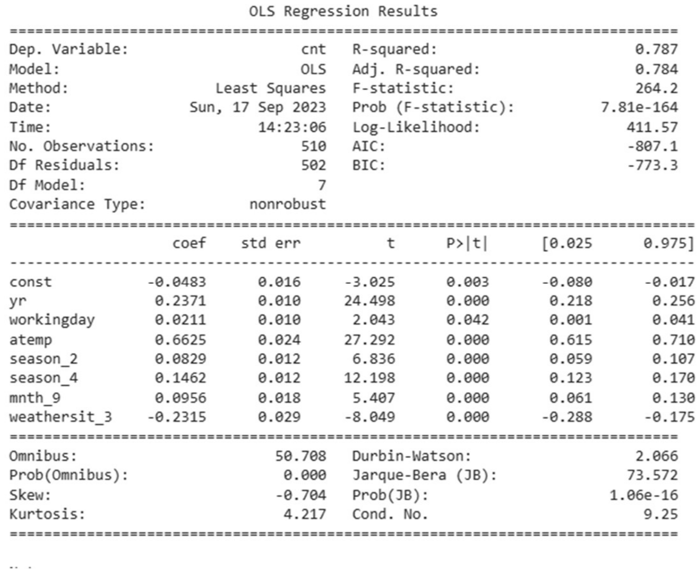


3.Variance inflation factors of all the features are less than 5 and p values are less than 0.05



```
--- Comparing VIF values ---

     Features   VIF
2       atemp  3.99
1  workingday  2.72
0          yr  1.92
3    season_2  1.59
4    season_4  1.36
5      mnth_9  1.21
6 weathersit_3 1.03
```



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.787
Model:                            OLS   Adj. R-squared:                  0.784
Method:                 Least Squares   F-statistic:                     264.2
Date:                Sun, 17 Sep 2023   Prob (F-statistic):          7.81e-164
Time:                        14:23:06   Log-Likelihood:                 411.57
No. Observations:                 510   AIC:                            -807.1
Df Residuals:                     502   BIC:                            -773.3
Df Model:                           7
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           -0.0483      0.016     -3.025      0.003      -0.080      -0.017
yr               0.2371      0.010     24.498      0.000       0.218       0.256
workingday       0.0211      0.010      2.043      0.042       0.001       0.041
atemp            0.6625      0.024     27.292      0.000       0.615       0.710
season_2         0.0829      0.012      6.836      0.000       0.059       0.107
season_4         0.1462      0.012     12.198      0.000       0.123       0.170
mnth_9           0.0956      0.018      5.407      0.000       0.061       0.130
weathersit_3    -0.2315      0.029     -8.049      0.000      -0.288      -0.175
==============================================================================
Omnibus:                       50.708   Durbin-Watson:                   2.066
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               73.572
Skew:                          -0.704   Prob(JB):                     1.06e-16
Kurtosis:                       4.217   Cond. No.                         9.25
==============================================================================
```

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

1. atemp

2.year

3. weather_Sit3

General Subjective Questions

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

Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (features). It assumes a linear relationship between the predictor variables and the target variable.

Here is a detailed explanation of the algorithm:

1. **Model Representation:**

Linear regression aims to find the best-fitting line (or hyperplane in multiple dimensions) that minimizes the sum of squared differences between the actual and predicted values. Mathematically, this is represented as:

$$y = \beta0 + \beta_1x_1 + \beta_2x_2 + ... + \beta_nx_n + \varepsilon$$

- \`y\` is the dependent variable (target).
- \`$x_1, x_2, ..., x_n$ are the independent variables (features).
- \`$\beta_0, \beta_1, ..., \beta_n$ are the coefficients or weights that the algorithm learns during training.
- \`ε\` represents the error term, which accounts for unexplained variability in \`y\`.

2. **Objective:**

The goal of linear regression is to find the values of \`β0, β1, ..., βn\` that minimize the sum of squared errors (SSE). This is achieved through a process called "ordinary least squares" (OLS) regression.

3. **Fitting the Model:**

-During training, the algorithm adjusts the coefficients (\`β\` values) to minimize the SSE. This is done by iteratively updating the coefficients based on the gradient of the loss function.

4. **Loss Function:**

The loss function in linear regression is typically the Mean Squared Error (MSE), which is the average of the squared differences between predicted and actual values.

5. **Assumptions:**

Linear regression assumes that there is a linear relationship between the independent and dependent variables.

It assumes that the errors (`ε`) are normally distributed with mean 0.

There should be little to no multicollinearity among the independent variables.

6. **Interpretation of Coefficients:**

The coefficients (`β` values) represent the change in the dependent variable for a one-unit change in the corresponding independent variable, while holding other variables constant.

7. **Predictions:**

Once the model is trained, it can be used to make predictions on new data by substituting the feature values into the regression equation.

8. **Evaluation:**

The performance of a linear regression model is typically assessed using metrics like R-squared (proportion of the variance explained by the model) and root mean squared error (RMSE) to quantify prediction accuracy.

In summary, linear regression is a fundamental algorithm used for modelling the relationship between variables. It aims to find the best-fitting line that minimizes prediction errors, making it a widely used tool in statistics and machine learning.

2. **Explain the Anscombe’s quartet in detail. (3 marks)**

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, and linear regression lines), but appear very different when graphed. It was created by the statistician Francis Anscombe in 1973 to illustrate the importance of graphing data before analysing it and to demonstrate the limitations of summary statistics.

Here's a detailed explanation of Anscombe's quartet:

1. Datasets:

- Anscombe's quartet consists of four datasets, each containing 11 (x, y) pairs. These datasets are denoted as I, II, III, and IV.

2. Descriptive Statistics:

- Despite having very different distributions and relationships between x and y, the four datasets share nearly identical summary statistics, including:

  - Mean of x: ~ 9.0
  - Mean of y: ~ 7.5
  - Variance of x: ~ 11.0
  - Variance of y: ~ 4.12
  - Correlation between x and y: ~ 0.816
  - Linear regression line: y = 3 + 0.5x

3. Graphical Representation:

- When the datasets are graphed, they reveal strikingly different patterns:

  - Dataset I: A linear relationship with some variance.
  - Dataset II: A curved relationship that can be well approximated by a quadratic model.
  - Dataset III: A perfectly linear relationship except for one outlier that has a significant impact on the linear regression line.
  - Dataset IV: No apparent relationship, but a high correlation coefficient.

4. Implications:

- Anscombe's quartet emphasizes the importance of visualizing data before drawing conclusions based on summary statistics alone. It shows that different datasets can lead to the same summary statistics but have fundamentally different underlying structures.

5. Statistical Analysis:

- The quartet demonstrates that summary statistics can be misleading and that graphical representation provides a more comprehensive understanding of the data.

In summary, Anscombe's quartet serves as a powerful reminder of the limitations of summary statistics and the importance of data visualization in statistical analysis. It highlights the need to explore and understand the structure of data graphically, rather than relying solely on numerical summaries.

3. What is Pearson’s R? (3 marks)

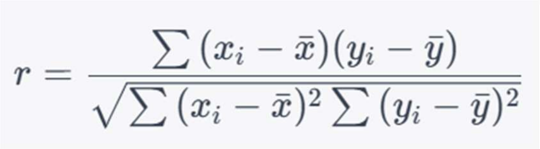
Pearson's correlation coefficient, often denoted as r is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1, where:

- r=1 indicates a perfect positive linear relationship,
- r=-1 indicates a perfect negative linear relationship, and
- r =0 indicates no linear relationship.

Here's a detailed explanation of Pearson's correlation coefficient:

1. Formula:

- Pearson's correlation coefficient (\(r\)) is calculated using the following formula:



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

x_i and y_i are the individual data points,

2. Interpretation:

- r = 1 indicates a perfect positive linear relationship. This means that as one variable increases, the other also increases proportionally.
- r = -1 indicates a perfect negative linear relationship. This means that as one variable increases, the other decreases proportionally.
- r = 0 indicates no linear relationship. However, it's important to note that there could still be other types of relationships.

3. Strength of Relationship:

The absolute value of r indicates the strength of the linear relationship.

- r ~1 suggests a strong linear relationship.
- r ~0 suggests a weak or no linear relationship.

4. Direction of Relationship:

- The sign of r indicates the direction of the relationship.

- r>0 indicates a positive linear relationship.
- r<0 indicates a negative linear relationship.

5. Assumptions:

Pearson's correlation assumes a linear relationship between the variables. It may not be sensitive to non-linear relationships.

It also assumes that the relationship is approximately bivariate normal (for reliable interpretations).

6. Limitations:

- It may not capture non-linear relationships.
- It is affected by outliers.

In summary, Pearson's correlation coefficient is a widely used statistical measure for quantifying the linear relationship between two continuous variables. It provides insights into the strength and direction of the relationship, but it's important to note its limitations, especially in the presence of non-linear relationships.

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

Scaling is a process in data preprocessing that involves transforming the values of variables into a specific range. It is done to ensure that all variables contribute equally to the analysis and to make the data more suitable for machine learning algorithms.

Why Scaling is Performed:

1. Magnitude Consistency: Scaling ensures that all variables are on a similar scale, preventing variables with larger magnitudes from dominating the analysis.

2. Algorithm Sensitivity: Many machine learning algorithms are sensitive to the scale of the input variables. Scaling can improve the performance of these algorithms.

3. Gradient Descent Convergence: Algorithms like gradient descent converge faster when variables are within a similar range.

4. **Distance-Based Methods:** Scaling is crucial for distance-based algorithms (e.g., k-nearest neighbours) where the distance between points is a key factor.

5. **Regularization:** Techniques like Lasso and Ridge regression are sensitive to the scale of variables, so scaling helps in applying these methods effectively.

Normalized Scaling vs. Standardized Scaling:

1. Normalized Scaling:

- Also known as Min-Max scaling.
- Involves transforming data to a specific range, usually [0, 1].

Formula: $\frac{x - \min(x)}{\max(x) - \min(x)},$

- Preserves the relative relationships between data points.
- Suitable when the distribution of the data is not assumed to be Gaussian.

2. Standardized Scaling:

- Also known as Z-score scaling or zero-mean normalization.
- Involves transforming data such that it has a mean of 0 and standard deviation of 1.
- Formula: $\frac{x - \mathrm{mean}(x)}{\mathrm{std\ dev}(x)}$
- Suitable for data with a Gaussian distribution.
- Preserves the shape of the original distribution.

3.Key Differences:

-Range:

- Normalized Scaling: Transforms data to a specific range (e.g., [0, 1]).
- Standardized Scaling: Adjusts data to have a mean of 0 and standard deviation of 1.

Impact on Distribution:

- Normalized Scaling: Preserves relative relationships between data points.
- Standardized Scaling: Preserves the shape of the original distribution.

Suitability:

- Normalized Scaling: Suitable when the distribution of the data is not assumed to be Gaussian.

- Standardized Scaling: Suitable for data with a Gaussian distribution.

In summary, scaling is performed to standardize the range of variables, making them suitable for machine learning algorithms. Normalized scaling transforms data to a specific range, while standardized scaling adjusts data to have a mean of 0 and standard deviation of 1, making it suitable for Gaussian-distributed data.

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

The occurrence of an infinite VIF (Variance Inflation Factor) typically indicates a problem known as perfect multicollinearity within the dataset. Perfect multicollinearity arises when two or more independent variables in a regression model are perfectly linearly dependent, meaning one of them can be exactly predicted from a linear combination of the others.

Here's why infinite VIF occurs:

1. Mathematical Issue:
- The formula for calculating VIF includes division by zero, which leads to an undefined or infinite result when perfect multicollinearity is present.

2. Perfect Linear Relationship:
- In a case of perfect multicollinearity, one or more independent variables can be expressed as an exact linear combination of the others. For example, if variable A can be expressed as 2 * Variable B they are perfectly linearly dependent.

3. Singular Matrix:
- In matrix algebra, perfect multicollinearity leads to a singular matrix, which causes the mathematical operations involved in VIF calculation to fail.

4. Regression Coefficients Cannot be Estimated:
- When perfect multicollinearity is present, it's not possible to uniquely estimate the regression coefficients for the affected variables. This is because there are infinite solutions to the regression equation.

5. Practical Implications:
- In practical terms, this means that the affected variable(s) cannot be included in the regression model, as their coefficients cannot be estimated.

6. Common Causes:
- Perfect multicollinearity can occur when there is redundancy in the set of independent variables, or when a variable is a linear combination of others.

7. Identifying and Addressing:
- Detecting perfect multicollinearity is important in regression analysis. It may require re-evaluating the set of independent variables, checking for data errors, or using techniques like variable selection or dimensionality reduction.

In summary, infinite VIF values indicate the presence of perfect multicollinearity, which is a critical issue in regression analysis. It requires careful examination and often necessitates the removal or modification of variables to address the problem.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
A Q-Q plot, or Quantile-Quantile plot, is a graphical tool used in statistics to assess whether a given dataset follows a particular probability distribution, such as the normal distribution. It helps in comparing the distribution of a dataset to a theoretical distribution.

Here's an explanation of the use and importance of a Q-Q plot in linear regression:

Use of Q-Q Plot:

1. Comparing Distributions:
- A Q-Q plot is used to visually compare the distribution of a dataset to a known theoretical distribution (e.g., normal distribution).

2. Identifying Departures from Assumed Distribution:
- It helps in identifying departures from the assumed distribution. If the points on the plot deviate significantly from the diagonal line, it suggests that the data does not follow the assumed distribution.

3. Checking for Normality:
- In linear regression, one of the key assumptions is that the residuals (errors) should be normally distributed. A Q-Q plot of residuals can help verify this assumption.

Importance of Q-Q Plot in Linear Regression:

1. Assumption Checking:
- Linear regression models make certain assumptions about the underlying data, including the normality of residuals. A Q-Q plot of residuals helps assess whether this assumption is met.

2. Detecting Skewness or Kurtosis:
- Deviations from a straight diagonal line in the Q-Q plot can indicate skewness or kurtosis in the data. This information is important for understanding the distribution of residuals.

3. Validity of Inferences:
- If the residuals are not normally distributed, it can impact the validity of statistical inferences, such as hypothesis testing and confidence intervals based on the regression model.

4. Guidance for Model Improvement:
- If the Q-Q plot reveals non-normality, it may suggest the need for transformations or the use of robust regression techniques to improve the model's performance.

5. Residual Diagnostics:

- Along with other diagnostic plots, a Q-Q plot provides insights into the behaviour of residuals, helping to identify potential issues with the model.

6. Model Interpretation:

- A normally distributed set of residuals simplifies the interpretation of coefficients and prediction intervals in the regression model.

In summary, a Q-Q plot is an essential tool in linear regression for assessing the assumption of normality in the residuals. It provides valuable insights into the distribution of the data and helps ensure the validity of inferences made from the regression model.