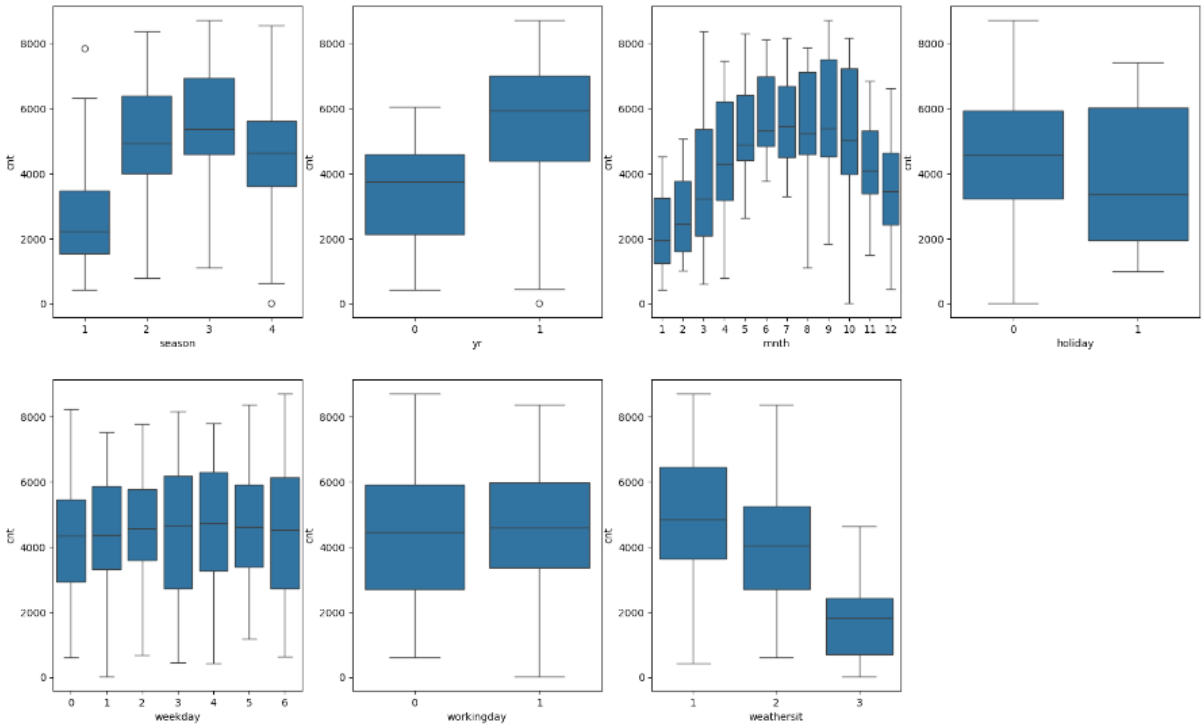
AI and ML Landscape - Graded Project 2

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)

I have done analysis on categorical columns using the boxplot and bar plot. Below are the few points we can infer from the visualization –

- Fall season seems to have attracted more bookings. And, in each season the booking count has increased drastically from 2018 to 2019.
- Bookings increased from may, june, july, aug, sep and oct, then it started decreasing as we approached the end of year.
- Clear weather attracted more booking which seems obvious.
- Thu, Fir, Sat and Sun have more bookings as compared to the start of the week.
- When it's not a holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.
- Booking seemed to be almost equal either on working day or non-working day.
- 2019 attracted more bookings from the previous year, which shows good progress in terms of business.

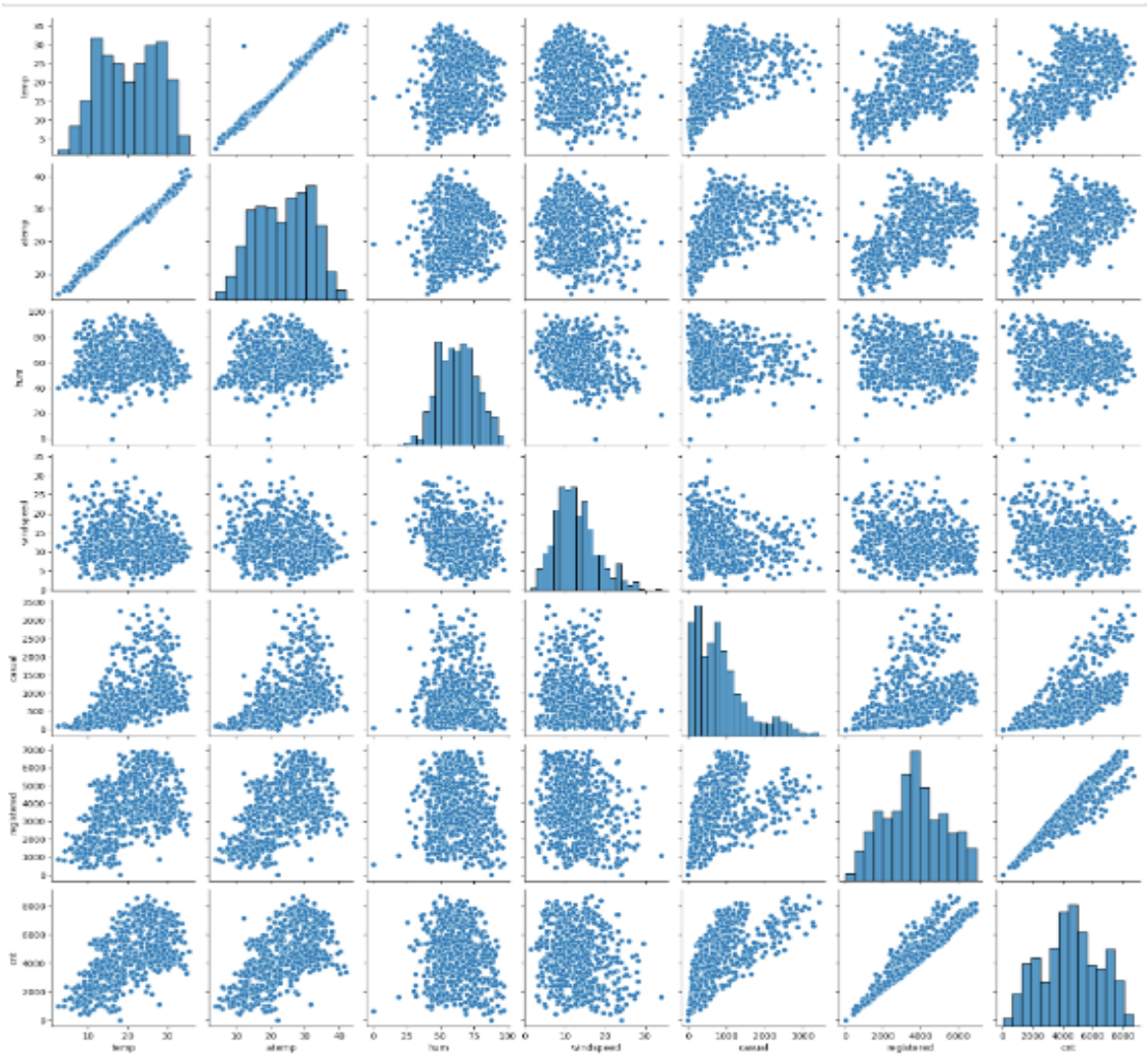


2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Using drop_first=True during dummy variable creation is important to avoid multicollinearity in regression models. Multicollinearity occurs when dummy variables are highly correlated with each other, which can lead to issues in estimating the model coefficients accurately. By dropping the first category, we create a reference group, preventing redundancy and ensuring that the dummy variables are independent. This results in a more stable and interpretable model, as the coefficients of the remaining dummy variables represent the effect relative to the dropped category.
Eg: If there are 3 levels, the drop_first will drop the first column.

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

The ‘temp’ and ‘atemp’ variables have highest correlation when compared to the rest with target variable as ‘cnt’.



4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

I have validated the assumption of Linear Regression Model based on below 5 assumptions -

- Normality of error terms
  - Error terms should be normally distributed
- Multicollinearity check
  - There should be insignificant multicollinearity among variables.
- Linear relationship validation
  - Linearity should be visible among variables
- Homoscedasticity
  - There should be no visible pattern in residual values.
- Independence of residuals
  - No auto-correlation

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –

- Temperature
- Year
- Weather

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 and one or more independent variables. The primary goal is to find the linear equation that best predicts the dependent variable based on the values of the independent variables. Here's a detailed explanation of the linear regression algorithm:

1. **Basic Concept**
   The linear regression equation is of the form:

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

where:

- y is the dependent variable.
- $x_1, x_2, \ldots \ldots x_n$ are the independent variables.
- β_0 is the y-intercept (the value of y when all x values are zero).
- $\beta_1, \beta_2, \ldots \ldots \ldots \ldots \beta_n$ are the coefficients (slopes) representing the relationship between each independent variable and the dependent variable.
- ϵ is the error term (the difference between the observed and predicted values).

2. Assumptions

Linear regression relies on several assumptions:

1. **Linearity**: The relationship between the dependent and independent variables is linear.
2. **Independence**: The observations are independent of each other.
3. **Homoscedasticity**: The residuals (errors) have constant variance at every level of the independent variables.
4. **Normality**: The residuals are normally distributed (especially important for inference).

3. Ordinary Least Squares (OLS)

The most common method for estimating the coefficients in linear regression is Ordinary Least Squares (OLS). OLS minimizes the sum of the squared differences between the observed values and the values predicted by the linear equation. The objective is to find $\beta_1, \beta_2, \ldots\ldots\ldots\ldots\ldots \beta_n$ that minimize the cost function:

$$Cost = \sum_{i-1}^{m} (yi - y_i)2$$

where yi is the actual value, y_i is the predicted value, and **m** is the number of observations.

4. Estimation of Coefficients

The coefficients can be estimated using the following matrix equation:

$$\beta = (X^T X)^{-1} - X_y^T$$

where:

- X is the matrix of independent variables (including a column of ones for the intercept).
- y is the vector of observed values.
- β is the vector of coefficients.

5. Model Evaluation

After fitting the model, it's important to evaluate its performance using metrics such as:

- **R-squared (R^2)**: Proportion of variance in the dependent variable that is predictable from the independent variables.
- **Adjusted R-squared**: Adjusted for the number of predictors in the model.
- **Mean Squared Error (MSE)**: Average of the squared differences between observed and predicted values.
- **Root Mean Squared Error (RMSE)**: Square root of the MSE.

6. Inference

Statistical tests (e.g., t-tests) can be performed to determine if the estimated coefficients are significantly different from zero, indicating a relationship between the independent and dependent variables.

7. Assumption Checks

It's crucial to check the assumptions of linear regression to ensure the validity of the model. This can be done using diagnostic plots such as:

- **Residual plots**: To check for homoscedasticity and linearity.
- **Q-Q plots**: To check for normality of residuals.
- **Variance Inflation Factor (VIF)**: To check for multicollinearity among independent variables.

8. Extensions

Linear regression can be extended to:

- **Multiple linear regression**: Involves more than one independent variable.
- **Polynomial regression**: Models non-linear relationships by including polynomial terms.
- **Regularized regression**: Includes techniques like Ridge and Lasso regression to handle multicollinearity and improve model generalization.

By understanding and applying these principles, linear regression can be a powerful tool for predicting and interpreting relationships between variables.

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

Anscombe's quartet is a collection of four datasets that have nearly identical simple descriptive statistics but have very different distributions and appear very different when graphed. It was constructed by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analyzing it and the effect of outliers and other anomalies on statistical properties. The quartet emphasizes that relying solely on summary statistics can be misleading.

Key Characteristics

Each of the four datasets in Anscombe's quartet has the following nearly identical properties:

- **Mean of x**: 9
- **Mean of y**: 7.5
- **Variance of x**: 11
- **Variance of y**: 4.12
- **Correlation between x and y**: 0.816
- **Linear regression line**: y=3+0.5x
- **Coefficient of determination (R^2)**: 0.67

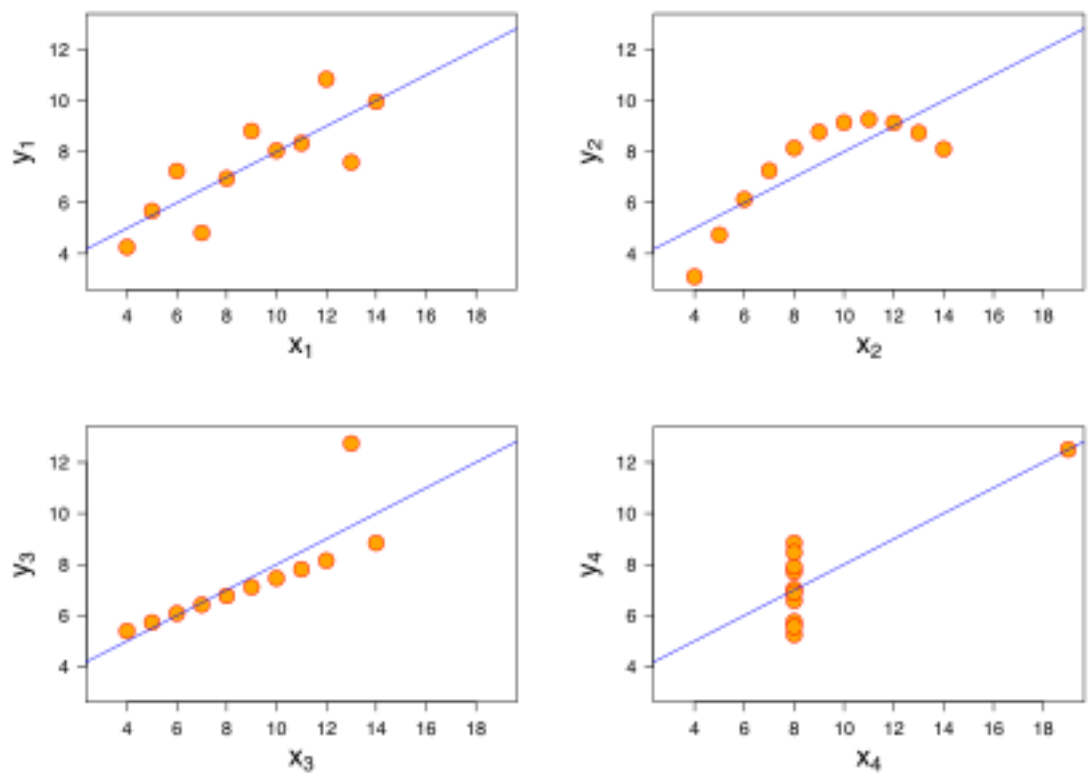
Despite these similarities in summary statistics, the datasets are vastly different in structure.

The Four Datasets

1. **Dataset I**:
    - This dataset is a classic example of a linear relationship between x and y.
    - The data points lie close to the line y=3+0.5x
2. **Dataset II**:
    - This dataset forms a perfect parabola.
    - Despite having a strong linear correlation, it clearly does not fit a linear model.
3. **Dataset III**:
    - In this dataset, almost all points are the same except one outlier.
    - The outlier has a significant effect on the correlation and the regression line.

4. **Dataset IV**:
    - This dataset includes an outlier in the x-values.
    - Most of the data points have the same y-value, and the outlier dictates the linear relationship.

Visual Representation



When graphed, these datasets reveal their differences clearly:

- **Dataset I**: The points align along a linear trend.
- **Dataset II**: The points form a curve, illustrating a non-linear relationship.
- **Dataset III**: The majority of points form a horizontal line, with one point far away, showing the impact of an outlier.
- **Dataset IV**: Most points are aligned vertically, with a single outlier influencing the linear trend.

Importance of Anscombe's Quartet

1. **Graphical Analysis**:
    - The quartet highlights the critical role of graphical analysis in data exploration.
    - Visualizing data can reveal patterns, trends, and anomalies that summary statistics might not.
2. **Statistical Insight**:
    - It demonstrates that different datasets can produce similar statistical measures, yet their actual distributions can be quite distinct.
    - It underscores the potential pitfalls of relying solely on numerical summaries.
3. **Outlier Influence**:
    - The quartet shows how outliers can disproportionately affect statistical properties and linear models.
    - Understanding and identifying outliers is crucial in data analysis.

Importance of Anscombe's Quartet

- **Always Plot Your Data**: Before jumping to conclusions based on summary statistics, visualize your data to understand its structure and distribution.
- **Beware of Outliers**: Recognize the influence of outliers and consider their impact on your analysis.
- **Comprehensive Analysis**: Use a combination of graphical and numerical methods to get a full picture of your data.
- **Model Appropriateness**: Ensure that the chosen model fits the data well; a linear model might not be suitable for all datasets even if the summary statistics suggest otherwise.

In summary, Anscombe's quartet serves as a powerful reminder of the importance of thorough and careful data analysis, highlighting the limitations of relying solely on summary statistics without graphical and more in-depth examination.

3. What is Pearson's R? (3 marks)

Pearson's R, also known as the Pearson correlation coefficient or simply the correlation coefficient, is a measure of the linear relationship between two variables. It quantifies the degree to which a pair of variables are linearly related. The value of Pearson's R ranges from -1 to 1, where:

- **1** indicates a perfect positive linear relationship.
- **-1** indicates a perfect negative linear relationship.
- **0** indicates no linear relationship.

Formula

Pearson's R is calculated using the following formula:

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

where:

- **n** is the number of data points.
- **x** and **y** are the individual data points.
- ∑ denotes the summation.

Interpretation

- **1**: There is a perfect positive linear relationship, meaning that as one variable increases, the other variable also increases proportionally.
- **-1**: There is a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases proportionally.
- **0**: There is no linear relationship between the variables.
- **0.5** or **-0.5**: Indicates a moderate positive or negative linear relationship, respectively.

Properties

- **Symmetry**: The correlation between x and y is the same as the correlation between y and x.
- **Range**: The value of Pearson's R is always between -1 and 1.

- **Unit-Free**: The correlation coefficient is a dimensionless number, meaning it does not depend on the units of the variables.

Assumptions

- **Linearity**: Pearson's R assumes a linear relationship between the variables.
- **Homoscedasticity**: The variability of the differences between the variables is constant.
- **Normality**: The variables should be approximately normally distributed (though this assumption is more relaxed when the sample size is large).

Use Cases

- **Correlation Analysis**: To understand the strength and direction of the linear relationship between two variables.
- **Regression Analysis**: To assess how well one variable can predict another.
- **Hypothesis Testing**: To test the null hypothesis that there is no linear relationship between the two variables.

Example

If we have two variables, **x** (hours studied) and **y** (test scores), and we calculate Pearson's R to be 0.85, this indicates a strong positive linear relationship: as the number of hours studied increases, test scores tend to increase as well.

In summary, Pearson's R is a valuable statistical tool for measuring the strength and direction of the linear relationship between two variables, providing insights into how changes in one variable are associated with changes in another.

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

Scaling is a data pre-processing technique used to adjust the range and distribution of numerical data features so they can be analysed more effectively by machine learning algorithms. Scaling helps ensure that features contribute equally to the model's performance and prevents features with larger magnitudes from dominating those with smaller magnitudes.

Why Scaling is Performed:

1. **Equal Contribution**: Ensures that all features contribute equally to the model, especially in algorithms that rely on distance measurements (e.g., k-nearest neighbours, support vector machines, and gradient descent-based methods).
2. **Convergence Speed**: Improves the convergence speed of gradient descent optimization algorithms by ensuring features are on a similar scale.
3. **Model Performance**: Enhances model performance by providing a better starting point for optimization, reducing bias, and improving numerical stability.

Types of Scaling:

1. **Normalized Scaling (Min-Max Scaling)**:
    - **Definition**: Rescales the feature values to a fixed range, usually [0, 1] or [-1, 1].
    - **Formula**:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

   - where **x** is the original value, $x_{\min}$ and $x_{\max}$ are the minimum and maximum values of the feature, respectively.
   - **Use Cases**: Suitable for algorithms that do not assume any particular distribution of the data and where the range of data needs to be preserved, such as image processing.
2. **Standardized Scaling (Z-score Normalization)**:
   - **Definition**: Rescales the feature values to have a mean of 0 and a standard deviation of 1.
   - **Formula**: $x' = \frac{x-\mu}{\sigma}$
   - where **x** is the original value, **μ** is the mean of the feature, and **σ** is the standard deviation.
   - **Use Cases**: Suitable for algorithms that assume normally distributed data, such as linear regression, logistic regression, and neural networks.

Differences Between Normalized and Standardized Scaling:

Aspect	Normalized Scaling	Standardized Scaling
Range	[0, 1] or [-1, 1]	Mean of 0 and standard deviation of 1
Formula	$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$	$x' = \frac{x - \mu}{\sigma}$
Assumptions	No assumption about data distribution	Assumes data is normally distributed
Effect on Data	Shrinks data to a specific range	Centers data and scales it based on variance
Use Cases	Image processing, where range matters	Algorithms assuming normal distribution, such as linear regression, logistic regression

Example:

Given a dataset with a feature age having values between 18 and 70:

- **Normalized Scaling**:
  - Original: [18, 35, 50, 70]
  - Normalized: [0, 0.28, 0.57, 1]
- **Standardized Scaling**:
  - Mean (μ): 43.25
  - Standard Deviation (σ): 20.52
  - Standardized: [-1.23, -0.40, 0.33, 1.30]

Summary

Scaling is a crucial pre-processing step to ensure that all features contribute equally to the model's performance. Normalized scaling is used to transform features to a fixed range, while standardized scaling centres the features around zero with a standard deviation of one. The choice of scaling method depends on the specific requirements and assumptions of the machine learning algorithm being used.

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

The value of VIF (Variance Inflation Factor) becomes infinite when there is **perfect multicollinearity** among the independent variables in a regression model. This means that one independent variable can be expressed as a perfect linear combination of the other independent variables.

Here's a breakdown of the reasons why VIF becomes infinite:

1. Underlying Reason:

- VIF is calculated by regressing each independent variable on all other independent variables in the model and then taking the reciprocal of 1 minus the R-squared (coefficient of determination) of this regression.
- In cases of perfect multicollinearity, the R-squared of the regression for a specific independent variable will be 1.

2. Mathematical Consequence:

- When R-squared is 1, it implies that the independent variable can be perfectly explained by the other independent variables in the model.
- The formula for VIF involves dividing 1 by (1 - R-squared). So, when R-squared is 1, the denominator becomes 0.
- Dividing by 0 results in an infinite value for VIF.

Why is Perfect Multicollinearity a Problem?

- It inflates the variance of the estimated regression coefficients, making them appear more significant than they truly are.
- This can lead to misleading interpretations of the model and difficulty in isolating the true effect of each independent variable.
- Additionally, with infinite VIF, it becomes impossible to determine the standard errors of the coefficients, further hindering accurate model assessment.

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 to compare the quantiles (percentiles) of two datasets. In the context of linear regression, it helps us assess whether the **residuals** (the difference between actual and predicted values) follow a normal distribution, which is a key assumption in linear regression.

Components of a Q-Q Plot:

1. **Quantiles:** These are the values that divide the data into equal-sized proportions. For example, the median divides the data into two halves, with 50% of the data points below it and 50% above it.
2. **Theoretical Distribution:** Usually, a normal distribution is used as the reference for comparison. The quantiles of the normal distribution are plotted on the x-axis.

3. **Sample Quantiles:** The quantiles of the data you're analyzing (in this case, the residuals) are plotted on the y-axis.

Interpretation:

- **Straight Line:** If the points in the Q-Q plot fall roughly along a straight line, it suggests that the residuals are **approximately normally distributed**. This satisfies an important assumption of linear regression and strengthens the validity of the model's results.
- **Deviations:** Deviations from a straight line indicate potential issues with the normality of the residuals. These deviations can reveal:
    - **Skewness:** If the points curve upwards, the residuals might be skewed to the right (positive skew).
    - **Heavy Tails:** If the points deviate from the line at the tails (either end), it might indicate heavier tails than a normal distribution (fatter tails).

Importance of Q-Q Plots in Linear Regression:

- **Assumption Checking:** Q-Q plots help identify potential violations of the normality assumption for residuals. This is crucial because non-normal residuals can lead to unreliable standard errors and hypothesis tests.
- **Model Diagnostics:** By visually inspecting the Q-Q plot, you can gain insights into the underlying distribution of the residuals and potentially diagnose issues like outliers or non-constant variance.
- **Model Improvement:** Based on the findings from the Q-Q plot, you can take corrective actions, such as data transformations (e.g., applying logarithms) or using robust regression techniques, to improve the model's performance and validity.