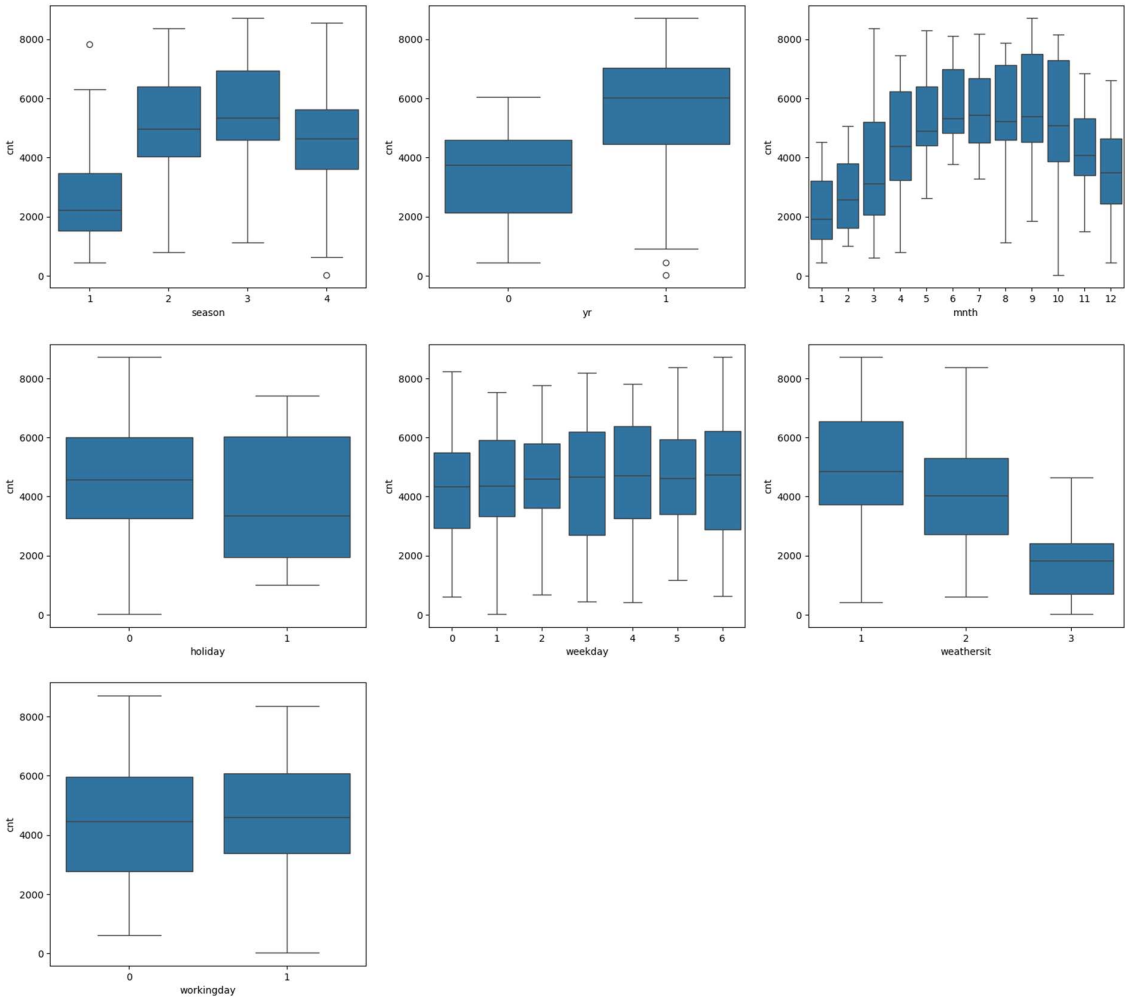
Answers- Assignment-based Subjective Questions

Q 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- From the boxplot for categorical variables we infer the following-



For seasons the value 3(Fall) has the highest median cnt

For yr 2019(which is coded as 1) has the highest median cnt

Among months july seems to have the highest median

median cnt is higher for non-holidays

For weathersit the category 1(Clear, Few clouds, Partly cloudy, Partly cloudy) has the highest median cnt

Working day seems to have the same impact as no working day on median cnt

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

Answer- Using `drop_first=True` when creating dummy variables in one-hot encoding is important for the following reasons:

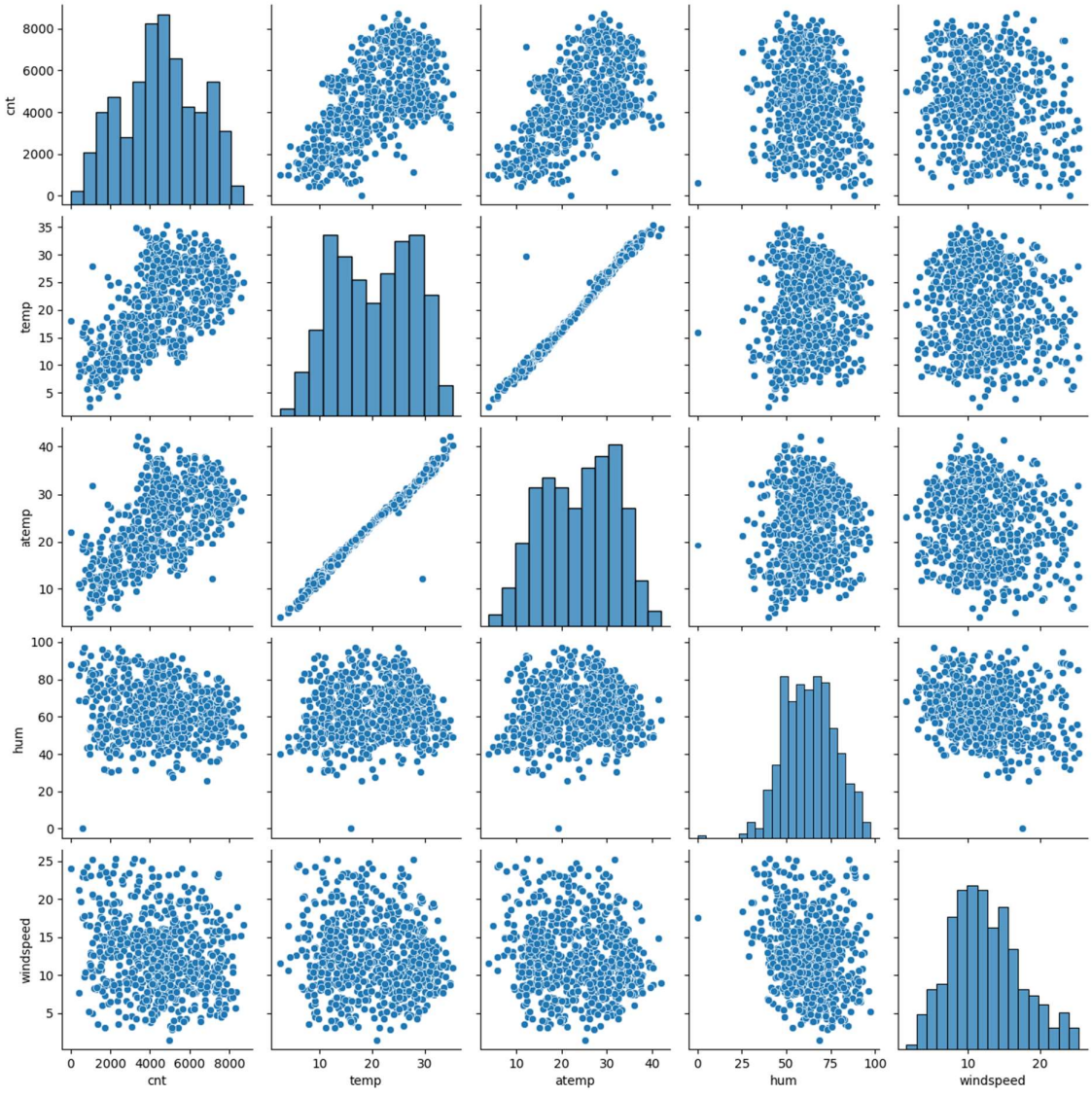
1. Prevents Multicollinearity: When all dummy variables are included, one can be perfectly predicted by the others, causing multicollinearity. Dropping the first dummy variable avoids this issue, as the dropped category becomes the baseline, and the remaining categories provide all necessary information.

2. Reduces Dimensionality: Dropping the first dummy variable helps to reduce the number of features in the dataset, which can simplify the model and slightly improve performance, especially with large numbers of categories.

Q 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation

with the target variable? (1 mark)

Answer-



From the above pair plot, we can clearly verify that the correlation between cnt and temp along with correlation between cnt and atemp is very high.

Q 4. How did you validate the assumptions of Linear Regression after building the model on the

training set? (3 marks)

Answer- To validate the assumptions of a Linear Regression model, several key checks are performed. Here's a summary of how this can be done after building the model on the training set:

1. Linearity:

- How to Check: Plot the predicted values vs. actual values or residuals vs. actual values. There should be no clear pattern in the residual plot, indicating a linear relationship between the predictors and the target.

- Action: If non-linearity is detected, consider feature transformations (e.g., log, square root) or adding polynomial terms.

2. Homoscedasticity (Constant Variance of Errors):

- How to Check: Plot residuals vs. predicted values. The spread of residuals should be consistent across all levels of the predicted values (i.e., no funnel shape).

- Action: If heteroscedasticity is present, try transformations or use weighted least squares regression.

3. Normality of Residuals:

- How to Check: Create a Q-Q plot (quantile-quantile plot) or histogram of the residuals. Residuals should be normally distributed.

- Action: If residuals deviate from normality, use a transformation on the target variable or consider a different model.

4. No Multicollinearity:

- How to Check: Calculate the Variance Inflation Factor (VIF). VIF values greater than 5-10 suggest multicollinearity.

- Action: If multicollinearity is found, remove or combine highly correlated features.

5. Independence of Errors:

- How to Check: For time-series data, plot residuals vs. time or use the Durbin-Watson test. For cross-sectional data, examine residuals for patterns of dependency.

- Action: If errors are not independent, consider using time-series models or adjusting for dependencies in the data.

By performing these checks, the assumptions of Linear Regression can be validated, and any potential violations can be addressed.

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

Answer- The features impacting the demand of bike positively are

1. yr
2. atemp
3. season_winter

Whereas the features affecting the demand adversely are

1. season_spring
2. workingday
3. weathersit_cloudy

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

Answer- Linear regression is a fundamental algorithm in machine learning and statistics used to model the relationship between a dependent variable (target) and one or more independent variables (predictors). Here's a detailed explanation:

1. Model Assumption Linear regression assumes a linear relationship between the independent variables X and the dependent variable y. The general form of the model is:

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

Where: y is the dependent variable (target). β_0 is the intercept (bias term). $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients (weights) corresponding to each independent variable. $x_1, x_2, \ldots, x_n$ are the independent variables (features). ϵ is the error term (residual) representing the difference between the observed and predicted values.

2. Objective The objective of linear regression is to find the values of the coefficients $\beta_0, \beta_1, \ldots, \beta_n$ that minimize the difference between the predicted and actual values of the target variable. This is done by minimizing the sum of squared errors (SSE) or residual sum of squares (RSS):

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

Where y_i is the actual value, and $\hat{y_i}$ is the predicted value. Minimizing SSE ensures the model fits the data as closely as possible.

3. Fitting the Model (Using Ordinary Least Squares - OLS) The linear regression algorithm uses the Ordinary Least Squares (OLS) method to estimate the coefficients. The OLS method calculates the values of β that minimize the sum of squared differences between the actual and predicted values of y.

Mathematically, the coefficients can be estimated using the following formula (in matrix form):

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

Where: X is the matrix of input features (including a column for the intercept). y is the vector of observed target values. $\hat{\beta}$ is the vector of estimated coefficients.

4. Model Interpretation Intercept (β_0): Represents the predicted value of y when all the independent variables are zero. Coefficients ($\beta_1, \beta_2, \ldots, \beta_n$): Represent the change in the predicted value of y for a one-unit increase in the corresponding independent variable, assuming all other variables are constant.

5. Assumptions of Linear Regression Linearity: The relationship between the independent and dependent variables is linear. Independence: Observations are independent of each other. Homoscedasticity: The variance of the residuals is constant across all levels of the independent variables. Normality of Errors: The

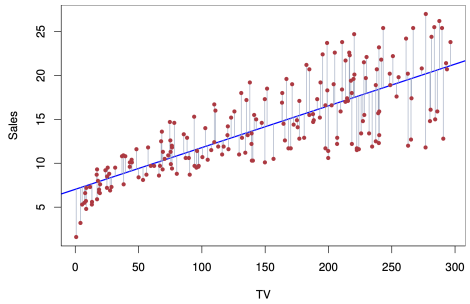


Figure 1: Linear Regression

residuals should be normally distributed. No Multicollinearity: Independent variables should not be highly correlated with each other.

6. Performance Metrics The model's performance can be evaluated using various metrics such as: R-squared: Indicates the proportion of variance in the dependent variable that can be explained by the independent variables. Mean Squared Error (MSE): Measures the average of the squared differences between actual and predicted values. Adjusted R-squared: Similar to R-squared but adjusted for the number of predictors in the model.

These steps define how linear regression works, from modeling the relationship to fitting the coefficients and validating its assumptions.

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

Answer- Anscombe's Quartet is a collection of four different datasets created by statistician Francis Anscombe in 1973. These datasets are used to illustrate the importance of visualizing data before relying solely on summary statistics. Despite having nearly identical summary statistics, the four datasets exhibit very different data distributions when plotted. Here's a detailed explanation:
1. Purpose of Anscombe's Quartet The primary goal of Anscombe's quartet is to demonstrate how relying on summary statistics (like mean, variance, correlation, and linear regression coefficients) can be misleading if you don't visualize the data. Anscombe created these datasets to show that datasets with identical statistical properties can still have vastly different structures, patterns, and outliers when graphed.
2. Statistical Properties of the Quartet Each dataset in Anscombe's quartet shares the following nearly identical summary statistics:
Mean of X: 9
Mean of Y: 7.50
Variance of X: 11
Variance of Y: 4.12
Correlation (Pearson's r) between X and Y: 0.816
Linear regression equation: $y = 3.00 + 0.500x$
Despite having these identical values, the underlying data in each dataset is very different, as revealed by visualizing them.

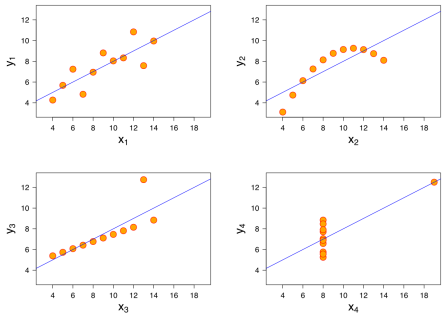


Figure 2: Anescombe's quartet

3. The Four Datasets

Dataset 1: This is a standard linear relationship. The points follow a roughly linear trend, and the linear regression model fits the data well. This dataset behaves as expected based on the summary statistics.

Dataset 2: This dataset is clearly non-linear. The data points form a curve, and while the summary statistics suggest a linear relationship, a linear regression line is not appropriate. This highlights the danger of relying solely on numerical correlation measures without examining the actual data.

Dataset 3: Here, most of the points form a perfect horizontal line except for one outlier. This outlier dramatically affects the summary statistics, making it seem like there's a moderate linear relationship, when in reality, the majority of the data shows no relationship at all. The outlier distorts the results of both the correlation and regression analysis.

Dataset 4: This dataset consists of nearly all points having the same x-value (except one point), resulting in a vertical line of data. The summary statistics still indicate a linear relationship, but it is meaningless since the x-values are essentially constant, and the regression line is determined mostly by one data point.

4.Key Takeaways from Anscombe's Quartet: Importance of Visualization: It emphasizes the importance of visualizing data through scatterplots or other graphical tools to understand its structure and behavior.

Misleading Summary Statistics: Even when datasets share the same statistical properties, the underlying data can tell very different stories. Summary statistics are not always sufficient for data analysis.

Outliers and Non-Linearity: Outliers and non-linear relationships can dramatically affect statistics like correlation, regression coefficients, and variance, leading to potentially incorrect conclusions if not properly considered.

Anscombe's Quartet serves as a classic example in data science and statistics that underscores the importance of combining data visualization with statistical analysis to avoid misleading interpretations.

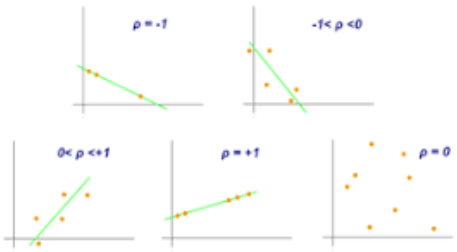


Figure 3: Pearson's r

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

Answer- Pearson's R (Pearson's Correlation Coefficient) is a measure of the linear relationship between two continuous variables. It quantifies the strength and direction of the relationship. Here's a breakdown of what it is and how it's used:

1.Definition and Formula Pearson's R is a statistic that measures the degree of correlation between two variables X and Y. The value of Pearson's R ranges from -1 to $+1$, where:

$+1$ indicates a perfect positive correlation (as one variable increases, the other increases). -1 indicates a perfect negative correlation (as one variable increases, the other decreases). 0 indicates no linear relationship between the variables.

The formula for Pearson's R is:

$$r = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2}\sqrt{\sum(Y_i - \overline{Y})^2}}$$

Where: X_i and Y_i are the individual data points of variables X and Y, $\overline{X}$ and $\overline{Y}$ are the means of the variables X and Y.

2. Interpretation $r = 1$: Perfect positive correlation. $r = -1$: Perfect negative correlation. $r = 0$: No correlation. $0 < r < 0.3$: Weak positive correlation. $0.3 \leq r < 0.7$: Moderate positive correlation. $r \geq 0.7$: Strong positive correlation (similar logic applies for negative values).

It's important to note that Pearson's R only measures linear relationships; it does not account for non-linear relationships.

3. Assumptions Pearson's R relies on the following assumptions: Linearity: The relationship between the two variables should be linear. Normality: The variables should be normally distributed. Homoscedasticity: The variance of the two variables should be equal across the range of values. Scale: The variables should be measured on an interval or ratio scale.

Pearson's R is a widely used correlation metric in statistics, machine learning, and data analysis to understand the strength of linear relationships between two continuous variables.

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

1. What is Scaling?

Scaling is the process of transforming the features of a dataset so that they lie within a certain range or follow a specific distribution. This is particularly important in algorithms that are sensitive to the magnitude of the data, such as those involving distance calculations (e.g., k-nearest neighbors, support vector machines) or gradient-based methods (e.g., linear regression, logistic regression).

2. Why is Scaling Performed?

Scaling is performed for the following reasons:

Improving model performance : Many machine learning algorithms perform better when features are on a similar scale because algorithms like gradient descent converge faster when the range of values across features is limited. Handling distance-based algorithms : In distance-based models (e.g., KNN, SVM), the distance between points is affected by the scale of the features. Without scaling, features with larger ranges may dominate distance computations. Avoiding biases : In unscaled data, features with larger values might disproportionately affect the outcome, causing biased model predictions.

3. Difference Between Normalized Scaling and Standardized Scaling

Both normalization and standardization are forms of scaling, but they differ in how they transform the data.

a. Normalization (Min-Max Scaling) : Definition : Normalization, also known as min-max scaling, rescales the feature values to a specific range, usually [0, 1]. It transforms data based on the minimum and maximum values in the feature. Formula:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Where X_{min} and X_{max} are the minimum and maximum values of the feature.

When to use : Normalization is useful when you want your data to lie within a specific range (e.g., for algorithms that rely on bounded values like neural networks).

Effect: The data is scaled proportionally, but outliers may affect the scaling because the min and max values drive the transformation.

b. Standardization (Z-score Scaling): Definition: Standardization transforms the data to have a mean of 0 and a standard deviation of 1 . This technique is often called Z-score normalization because it shifts and scales the data based on its statistical properties. Formula :

$$X_{std} = \frac{X - \mu}{\sigma}$$

Where μ is the mean of the feature and σ is the standard deviation.

When to use : Standardization is commonly used in algorithms where the assumption of normally distributed features is made (e.g., linear regression, logistic regression, SVM, PCA).

Effect : It's less affected by outliers compared to normalization and works well for data that follows a Gaussian distribution.

Summary : Normalization scales features to a range, often [0, 1], making it useful for algorithms where values need to be bounded or proportional. Standardization rescales data to have a mean of 0 and a standard deviation of 1, useful for algorithms that assume normally distributed data or that are sensitive to the variance.

Both methods are essential for improving model performance but should be chosen based on the specific algorithm and the nature of the data.

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

Answer- Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in regression models, specifically to check how much the variance of a regression coefficient is inflated due to multicollinearity with other features. A high VIF indicates that a variable is highly correlated with other predictors, leading to potential problems in the model.

1. Why Does VIF Become Infinite?
The VIF value can become infinite when a feature is a perfect linear combination of other features in the dataset. This means there is perfect multicollinearity—one predictor variable can be expressed as an exact linear function of others. In such a case, the regression model cannot estimate the coefficients of these variables because it's mathematically impossible to separate their effects on the dependent variable.

In this scenario, the denominator in the VIF calculation approaches zero, leading to an infinite value.

2. VIF Calculation and Perfect Multicollinearity VIF for a given feature X_i is calculated as:

$$VIF(X_i) = \frac{1}{1 - R_i^2}$$

Where: - R_i^2 is the coefficient of determination from the regression of X_i on all other independent variables.

If $R_i^2 = 1$, it means that X_i is perfectly explained by the other independent variables, indicating perfect multicollinearity. In such a case, the denominator becomes zero, and VIF becomes infinite:

$$VIF(X_i) = \frac{1}{1 - 1} = \infty$$

3. Why This Happens in Practice There are several reasons why you might encounter an infinite VIF value: Duplicate variables : If you accidentally include duplicate or near-duplicate columns in your dataset, they will be perfectly collinear, leading to infinite VIF. Exact linear relationships : If one or more features are exact linear combinations of others (e.g., summing two variables to create a third), multicollinearity will be perfect. Dummy variable trap : In cases where you create dummy variables for categorical data but do not drop one category, perfect multicollinearity can arise, leading to infinite VIF.

Solution : To deal with infinite VIF values, you should identify and remove the source of perfect multicollinearity, which might involve: Removing one of the perfectly collinear variables. Avoiding the dummy variable trap by excluding one category when creating dummy variables.

Q 6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer- A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a dataset follows a specific theoretical distribution, most commonly the normal distribution. In a Q-Q plot, the quantiles of the sample data are plotted against the quantiles of a theoretical distribution. Here's a detailed explanation of its use and importance in linear regression:

1. Definition and Construction of Q-Q Plot A Q-Q plot consists of points that represent the quantiles of the sample data on the y-axis and the quantiles of the theoretical distribution (often normal) on the x-axis.

Construction Steps: (i). Calculate Quantiles : For the sample data, compute the quantiles (e.g., 0.01, 0.02, ..., 0.99). (ii). Theoretical Quantiles : Calculate the corresponding quantiles for the chosen theoretical distribution (e.g., a normal distribution). (iii). Plot the Points : Each point in the plot corresponds to a pair of quantiles (sample quantile, theoretical quantile).

If the points fall approximately along a straight line (typically the 45-degree line), this indicates that the sample data follows the theoretical distribution.

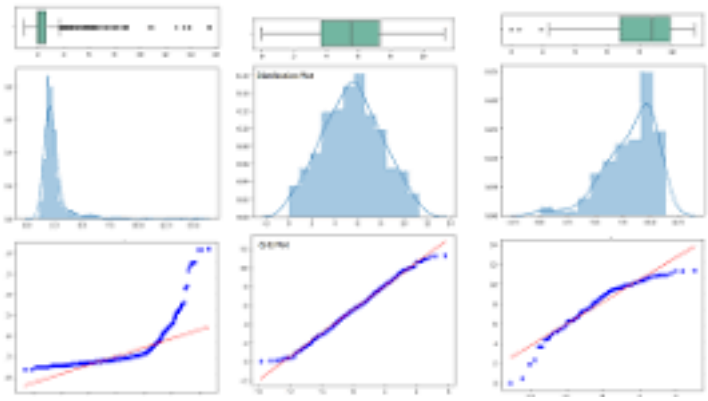
2. Use of Q-Q Plot in Linear Regression In the context of linear regression, Q-Q plots are primarily used to assess the normality of residuals (the differences between observed and predicted values). The assumptions of linear regression include:

Linearity : The relationship between independent and dependent variables is linear. Independence : Observations are independent. Homoscedasticity : Constant variance of residuals. Normality : The residuals of the regression model should be normally distributed.

The Q-Q plot helps in verifying the normality assumption. If the residuals are normally distributed, the points in the Q-Q plot will align closely with the diagonal line. Deviations from this line suggest departures from normality, which may indicate issues such as skewness or the presence of outliers.

3. Importance of Q-Q Plot in Linear Regression The importance of a Q-Q plot in linear regression lies in its ability to:

Diagnose Model Fit : By checking the normality of residuals, Q-Q plots provide insights into whether the linear regression model is appropriate for the data. Detect Outliers : Points that fall far from the diagonal line can indicate outliers or influential data points that may disproportionately affect the regression results. Improve Model Validity : If the residuals are not normally distributed, it might suggest that the model is not adequately capturing the underlying data patterns. This could lead to incorrect conclusions about the relationships between variables, necessitating model refinement or transformation of the dependent variable.



In summary, Q-Q plots are essential for validating the assumptions of linear regression, diagnosing potential issues in the model, and ensuring the reliability of the regression analysis.