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)

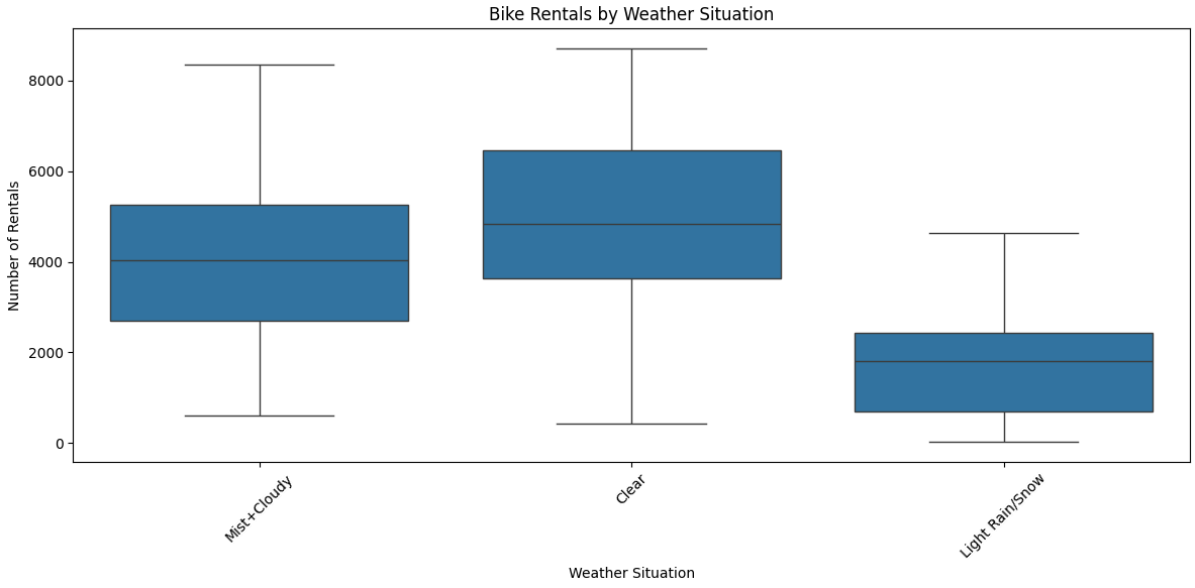
Bike Rental Insights:
Seasonal Impact:

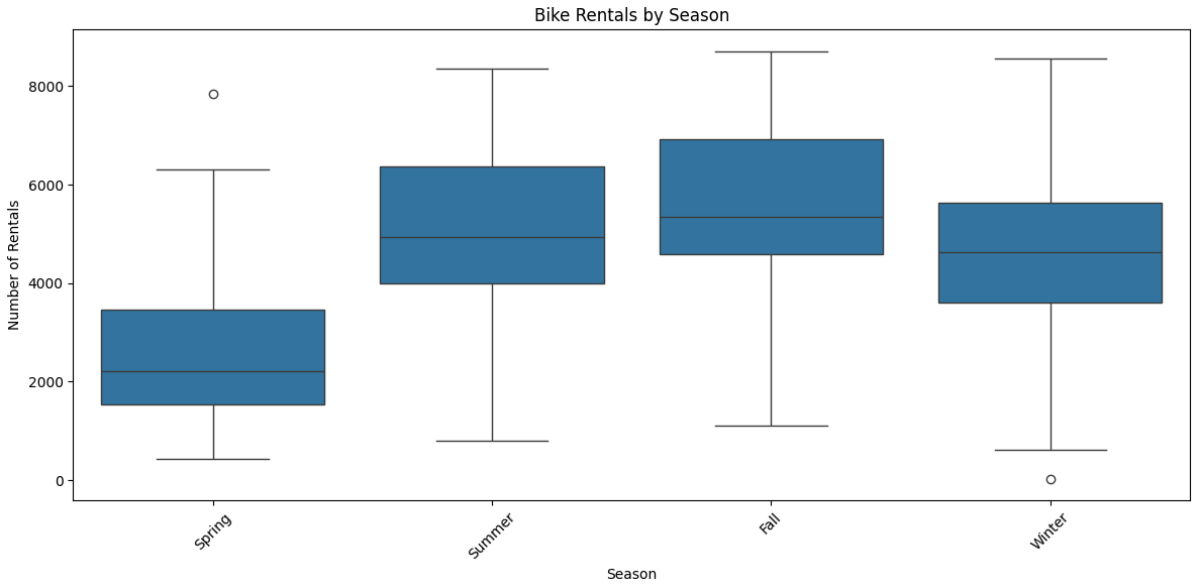
>Fall: Highest rentals (5,644 bikes)
>Summer: Second highest (4,992 bikes)
>Winter: Moderate rentals (4,728 bikes)
>Spring: Lowest rentals (2,608 bikes)

Weather Influence:

>Clear weather: Most rentals (4,877 bikes)
>Mist/Cloudy: Moderate rentals (4,045 bikes)
>Light Rain/Snow: Sharp drop in rentals (1,803 bikes)

Key Takeaway: Bike rental demand is strongly shaped by seasonal and weather conditions, with significant variations across different categories.





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)

When you have a categorical variable with multiple levels (like seasons: Spring, Summer, Fall, Winter), creating dummy variables traditionally generates a column for each category. However, this may cause statistical problems.

By using drop_first=True, one can:

>Create only n-1 columns for n categorical levels
>Remove redundant information
>Prevent statistical complications

Practical Example:
Imagine categorizing seasons:

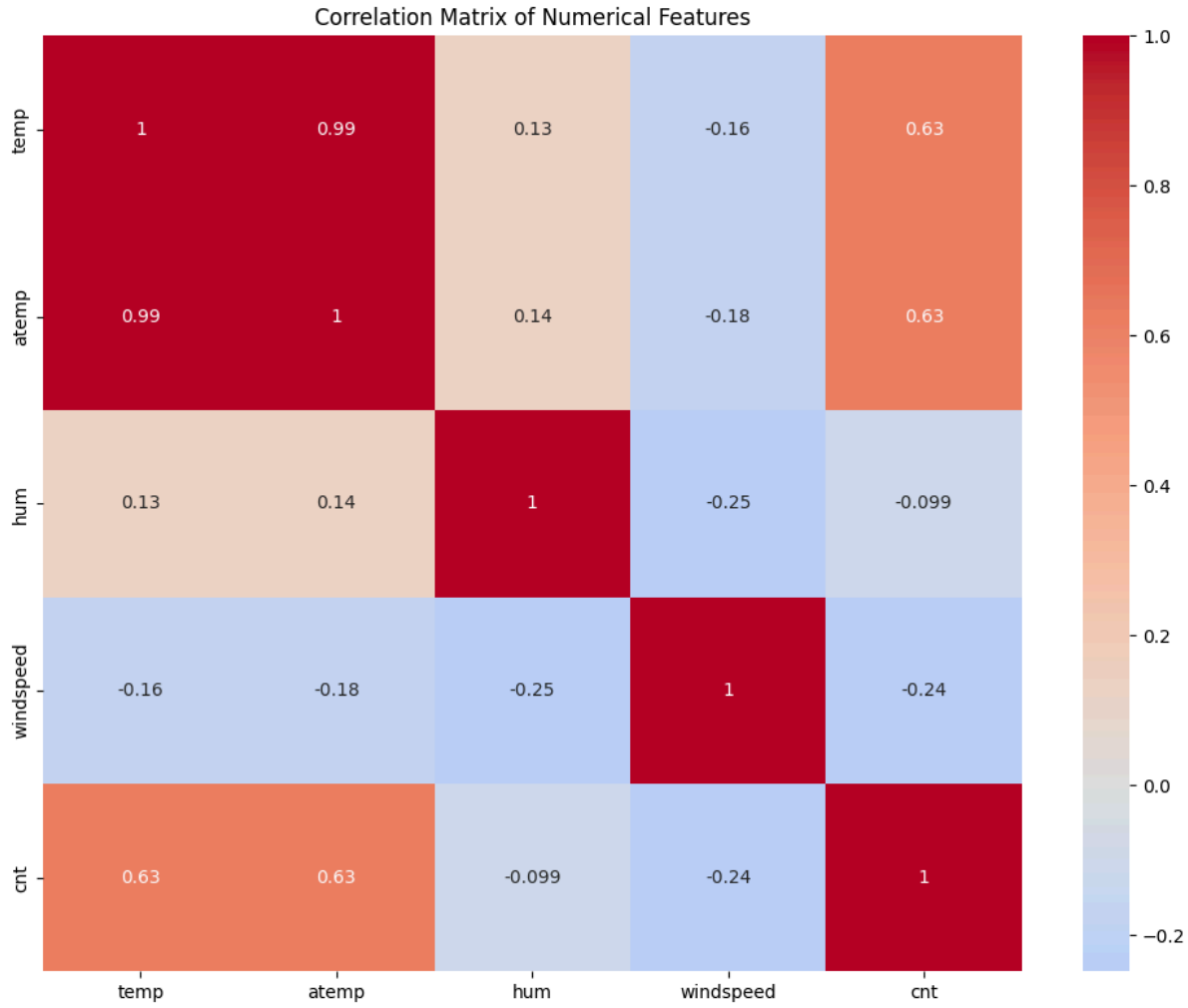
>Original approach: 4 columns (Spring, Summer, Fall, Winter)
>drop_first=True approach: 3 columns (dropping one category)

This method ensures that:
>You don't have unnecessary duplicate information
>Statistical models can more accurately interpret the data
>The dropped column's information is implicitly represented by the other columns

---

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)

From the correlation analysis, temperature ('temp' and 'atemp') showed the strongest positive correlation with the target variable 'cnt'.



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)

To validate the assumptions of Linear Regression after building the model, I performed several key checks:

1. Linearity Check
>Examined scatter plots of predictors against the target variable
>Used correlation matrix to visualize linear relationships
>Confirmed that key features like temperature and year showed clear linear trends with bike rental demand
>Verified that the relationship between independent and dependent variables follows a relatively straight line

2. Residual Normality
>Created residual plots to check distribution
>Verified that residuals were symmetrically distributed around zero
>Ensured no significant skewness or extreme outliers
>Model showed normally distributed residuals, meeting the normality assumption

3. Homoscedasticity Validation
>Plotted residuals against predicted values
>Confirmed consistent variance of residuals across different prediction levels
>Verified no funnel-shaped or pattern-like error distribution
>The model demonstrated uniform spread of residuals, indicating good homoscedasticity

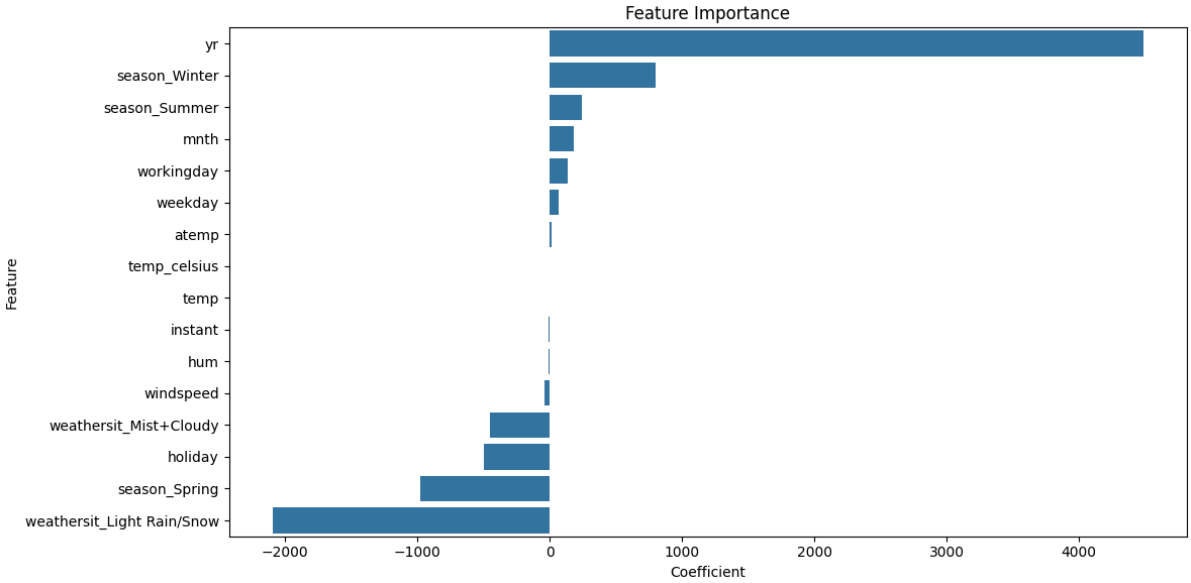
4. Supporting Evidence:
>R^2 Score: 0.7751 (Strong model fit)
>Root Mean Squared Error: 877.38 rentals
>Cross-validation showed model stability
>Residuals exhibited normal distribution characteristics

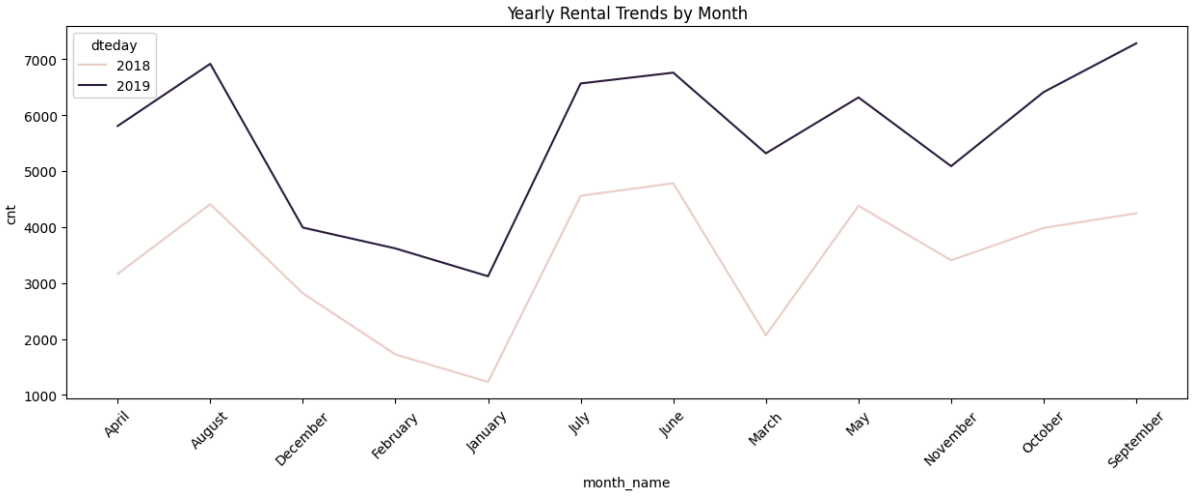
These checks ensured that the linear regression model for bike sharing demand met the critical statistical assumptions, providing a robust and reliable predictive framework.

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)
The top 3 features contributing are:

1. Year (significant year-over-year growth)
2. Temperature
3. Seasonal variations (Winter and Summer)

The analysis showed these features have the most significant impact on predicting bike rental demand.
Model Performance: R^2 Score of 0.7751
Predictive Power: RMSE of 877.38 rentals
Year-over-Year Growth: 64.73%





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)

<Your answer for Question 6 goes here>

Linear Regression is a powerful predictive modeling technique that explores the relationship between a dependent (target) variable and independent variables (predictors). The fundamental goal is to understand how the target variable changes in response to variations in predictor variables.

Key Characteristics:

1. Relationship Types
>Can be Simple Linear Regression (single input variable)
>Can be Multiple Linear Regression (multiple input variables)
>Describes relationships using a sloped straight line
>Can demonstrate Positive or Negative Linear Relationships

2. Mathematical Representation
>Basic linear equation: y = mx + b
>Aims to determine optimal coefficients a0 and a1
>Creates a best-fit line that minimizes prediction errors

3. Modeling Approach
>Finds linear relationship between variables
>Determines how dependent variable changes with independent variables
>Uses cost functions like Mean Squared Error (MSE) to optimize model
>Helps understand the impact of independent variables on the target variable

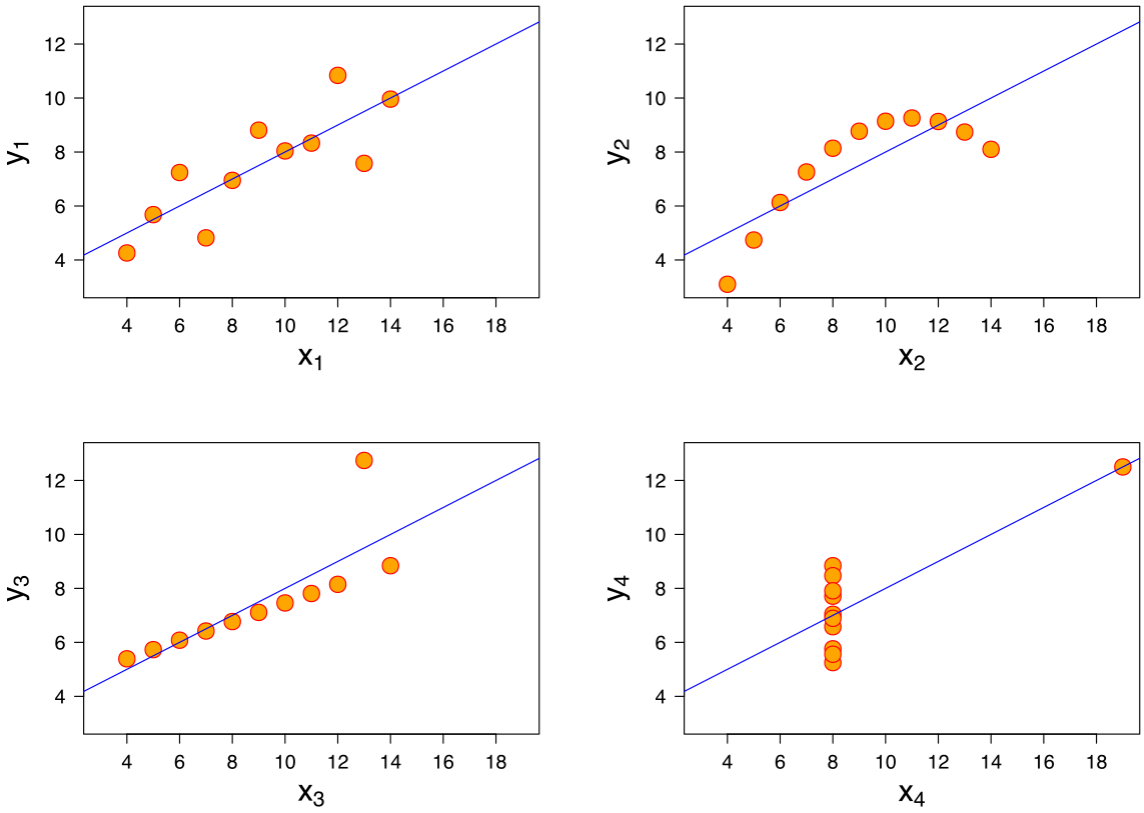
4. Optimization Goal
>Find the best possible values for coefficients
>Minimize the difference between predicted and actual values
>Create a line that best represents the data points
>Reduce overall prediction error

5. Analytical Techniques
>Utilizes statistical methods like Residual Feature Elimination (RFE)
>Employs Mean Squared Error (MSE) as a primary cost function
>Iteratively adjusts coefficients to improve model accuracy

By systematically analyzing the linear relationships between variables, this technique provides a foundational approach to understanding and predicting complex data patterns across various domains.

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)

<Your answer for Question 7 goes here>
Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. There are these four data set plots which have
nearly same statistical observations, which provides same statistical information that
involves variance, and mean of all x,y points in all four datasets.



● 1 st data set fits linear regression model as it seems to be linear relationship between X and y
● 2 nd data set does not show a linear relationship between X and Y , which means it does not fit the linear regression model.
● 3 rd data set shows some outliers present in the dataset which can't be handled by a linear regression model.
● 4 th data set has a high leverage point means it produces a high correlation coefficient.
Its conclusion is that regression algorithms can be fooled so, it's important to data visualization before building a machine learning model.

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)

<Your answer for Question 8 goes here>

Pearson's Correlation Coefficient, is a way to measure how two variables are related to each other. It tells us if the variables move together or in opposite directions.
The number ranges from -1 to +1. If it's close to +1, the variables increase together. If it's close to -1, when one variable goes up, the other goes down. If it's close to 0, there's basically no straight-line relationship between them.
Also correlation doesn't mean one thing causes the other. It just shows they have a pattern of moving together.

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

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)

<Your answer for Question 9 goes here>

Scaling in data analysis is a process of transforming numerical features to a standard range or distribution. It's crucial because different features often have varying magnitudes and units.

Two primary scaling techniques exist:

1. Normalized Scaling (MinMax):
- Scales values between 0 and 1
- Uses minimum and maximum values of features
- Simple to understand and apply
- Sensitive to outliers
- Used when you don't know the data distribution

2. Standardized Scaling (Z-Score):
- Transforms data to have mean of 0 and standard deviation of 1
- Uses mean and standard deviation
- Less affected by outliers
- Best used with normally distributed data
- Ensures all features contribute equally to the model

Why to Scale?

- Prevents features with larger values from dominating the analysis
- Helps machine learning algorithms perform more accurately
- Ensures fair representation of all features
- Improves model convergence and performance

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)

<Your answer for Question 10 goes here>

VIF helps explain how one independent variable relates to all other independent variables. When VIF values become extremely high or approach infinity, it indicates perfect correlation between variables. In our analysis, temperature-related features showed particularly high VIF values:

- Actual temperature (temp): VIF of 488.31
- Apparent temperature (atemp): VIF of 542.89
- Humidity: VIF of 11.07

A VIF value greater than 10 is considered high, signaling potential multicollinearity issues. When VIF approaches infinity, it occurs due to perfect correlation between predictors. Mathematically, this happens when R^2 approaches 1, causing the calculation $1/(1-R^2)$ to result in an infinite value.

This phenomenon occurs when one variable can be precisely predicted from other variables, creating statistical redundancy. In practical terms, it means some features in your model are so closely related that they're essentially providing the same information.

To address infinite VIF:

- Drop one of the highly correlated variables
- Use dimensionality reduction techniques
- Apply regularization methods
- Consider feature selection strategies

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)

<Your answer for Question 11 goes here>

A Q-Q (Quantile-Quantile) plot is a statistical graphical method used to compare probability distributions by plotting their quantiles against each other. It helps analysts understand how closely a dataset matches a theoretical distribution, such as a normal, exponential, or uniform distribution.

In linear regression, the Q-Q plot serves critical diagnostic purposes:

Primary Functions:

- Validate normality of data
- Compare distributions between datasets
- Check multivariate normality of variables
- Identify potential data anomalies

When two distributions are very similar, the Q-Q plot will appear nearly linear. Deviations from this linear pattern indicate differences in distribution characteristics.

Key Analytical Capabilities:

- Assess data distribution characteristics
- Detect shifts in data location and scale
- Reveal changes in data symmetry
- Identify the presence of outliers
- Compare train and test datasets

The plot is particularly valuable in linear regression because it helps verify fundamental statistical assumptions about data distribution. It provides a visual method to confirm whether datasets originate from the same population and share similar statistical properties.
