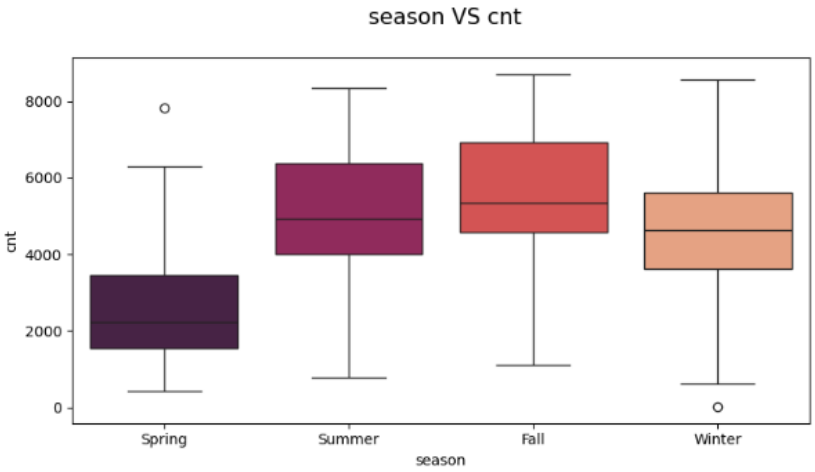
Assignment based Subjective Questions

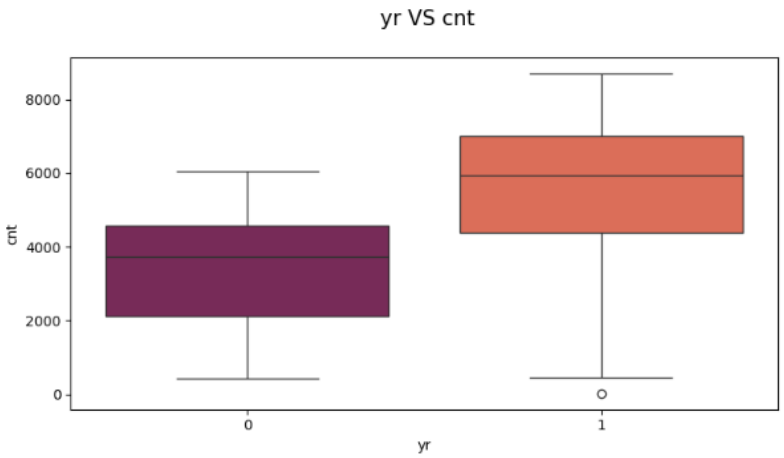
Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer :

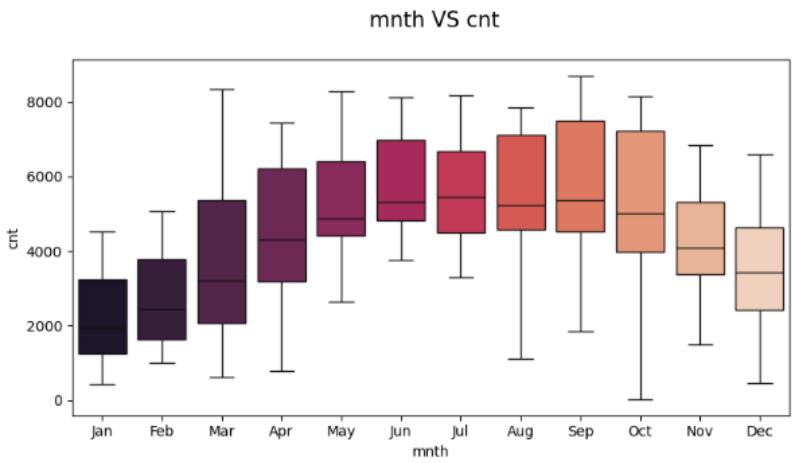
The dataset included categorical variables such as season, year, month, holiday, weekday, working day, and weather situation. A boxplot was used to visualize these variables, highlighting their influence on the dependent variable (count) in the following ways:



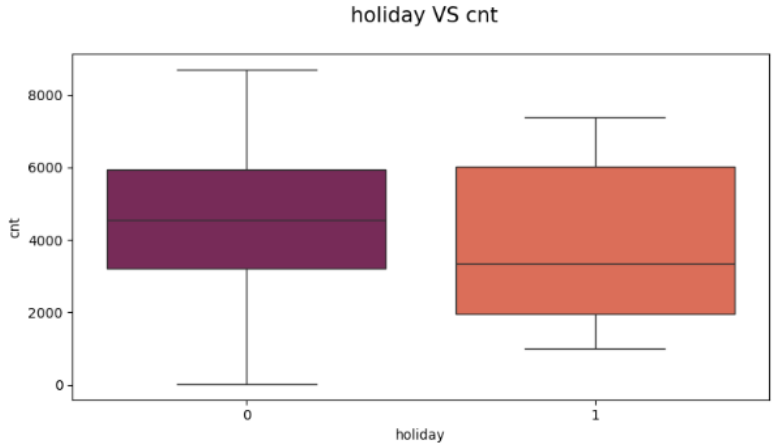
- The count of bikes rented is lesser in Spring as compared to other seasons



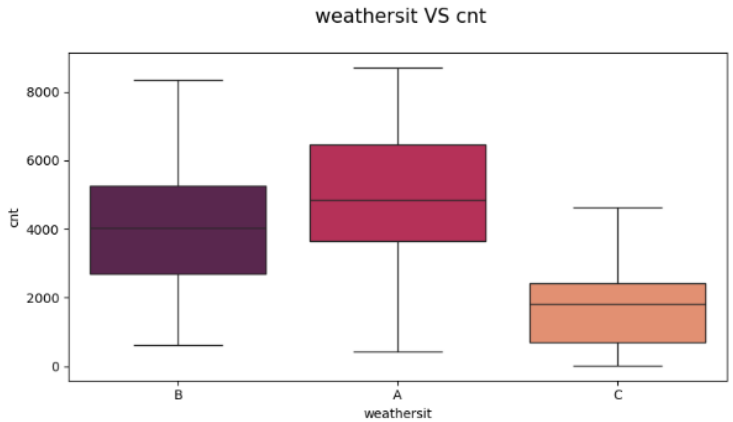
- The count of bikes rented is significantly higher in 2019 compared to 2018.



- The count of bikes rented gradually increases from January to September, reaching a peak in September. Then, the count decreases steadily until December.



- The median count of bikes rented is significantly lower on holidays.



- Very few bikes are rented during weather situation C (i.e., light snow, light rain, thunderstorms, scattered clouds). In poor weather conditions, the number of rented bikes decreases.

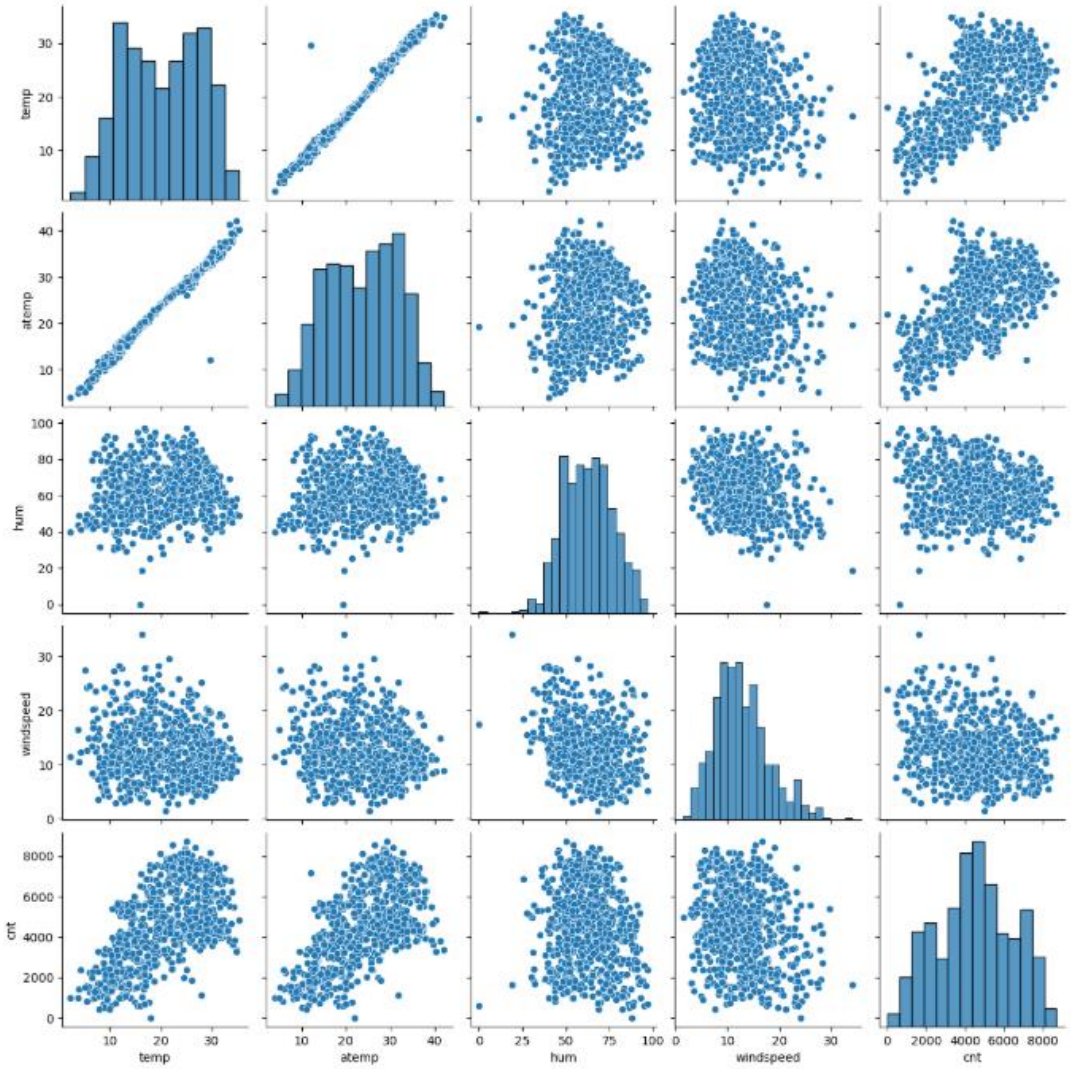
Q2. Why is it important to use drop_first=True during dummy variable creation?

Answer:

- When creating dummy variables, each categorical variable is converted into multiple binary (0/1) variables, one for each category.
- If you keep all the dummy variables, they become linearly dependent on each other. For example, if you have three categories (A, B, C), knowing the values of two dummy variables automatically determines the third. This causes multicollinearity, which can lead to unstable estimates in regression models.
- Removing one dummy variable reduces the number of columns in the dataset, leading to a more compact representation of the model without losing any information.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer:



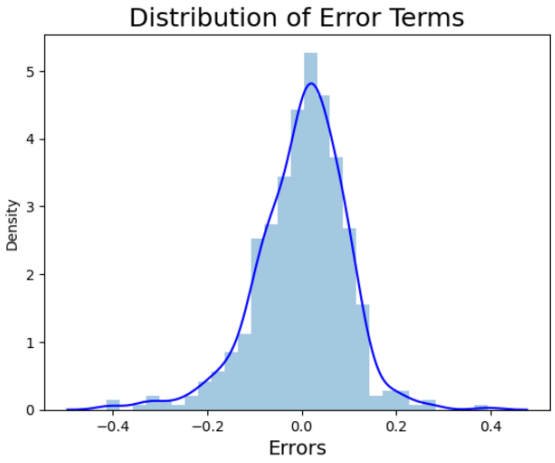
- The variables *'temp'* and *'atemp'* are having the highest correlation with the target variable *'cnt'* when compared to the other features in the dataset.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

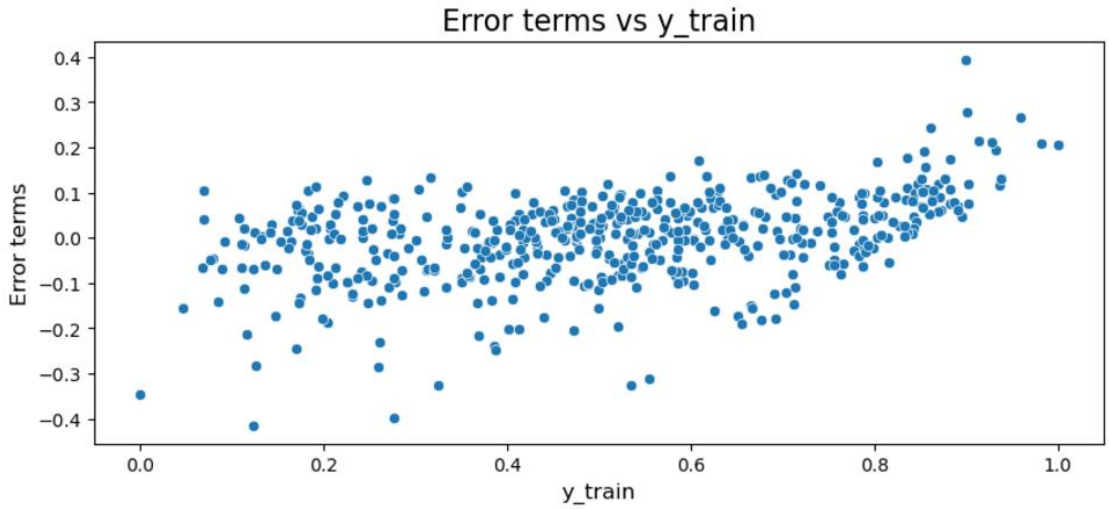
Answer:

Linear regression models are validated based on several key assumptions:

1. There should be a linear relationship between target variable and predictor variables
2. Error terms should be normally distributed with mean zero.
3. Error terms should be independent of each other.
4. Error terms should be exhibiting homoscedasticity (constant variance).



- Here we can see that Residuals are normally distributed and are symmetrically centered around zero.



- There is no observable pattern among the error terms, indicating that the error terms are independent of one another.
- The error terms demonstrate constant variance, indicating that they exhibit homoscedasticity.

5. Also the absence of multicollinearity among the independent variables (Low VIF).

These assumptions are critical for ensuring the reliability, interpretability, and accuracy of the model's estimates and predictions.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer:

Based on the regression results, the contribution of features can be analysed by considering magnitude of coefficient.

1. **Temperature (temp):**
   Coefficient: 0.3507
   - Temperature has the highest positive impact on demand. Higher temperatures (within the comfortable range) increase bike usage.
2. **Year (yr):**
   Coefficient: 0.2368
   - The demand increased significantly in 2019 compared to 2018, indicating a strong growth trend over time.
3. **Weather-Light sow, Light Rain & Scattered Clouds (weathersit_C):**
   Coefficient: -0.2762
   - Adverse weather (light snow, thunderstorms, scattered clouds) significantly reduces bike demand.

The most significant factors driving shared bike demand are **temperature, year**, and **adverse weather conditions (weathersit_C)**. BoomBikes can leverage these insights to optimize their services, marketing strategies, and resource allocation to maximize demand.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Answer:

Linear regression is a statistical method used to establish a relationship between a dependent variable (the target) and one or more independent variables (predictors). It helps predict the value of the target variable based on the input variables.

In its simplest form, **simple linear regression**, there is only one independent variable, and the relationship is represented by the equation:

$y=\beta 0+\beta 1x+\epsilon$

Here:

- y is the predicted value (dependent variable).
- β0 is the intercept (the value of y when x = 0).
- β1 is the slope (rate of change in y with respect to x).
- ϵ is the error term, accounting for deviations in data.

For **multiple linear regression**, the model extends to include multiple independent variables:

$y=\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta nxn+\epsilon y$

The goal is to find the values of β0, β1 ,… βn that minimize the difference between the predicted and actual values. This is usually achieved using **Ordinary Least Squares (OLS)**, which minimizes the sum of squared differences between observed and predicted values (called residuals).

Linear regression can model relationships that are either:

- **Positive linear**: As the independent variable increases, the dependent variable also increases.
- **Negative linear**: As the independent variable increases, the dependent variable decreases.

This algorithm is widely used because of its simplicity and ease of interpretation. However, it works best when the following assumptions are met:

1. The relationship between variables is linear.
2. The residuals (errors) are normally distributed.
3. The error variance is constant (homoscedasticity).

Linear regression provides insights into the factors influencing the dependent variable, making it a valuable tool in predictive modelling and analysis.

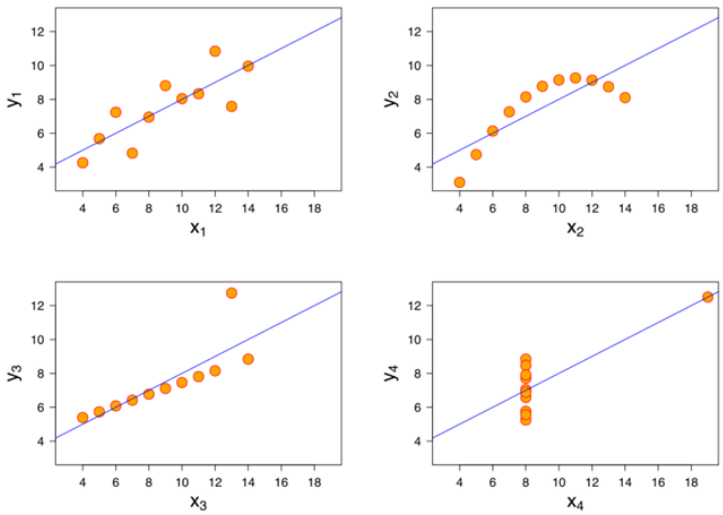
Q2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's Quartet is a collection of four datasets that share almost identical key statistical properties, such as mean, variance, and correlation, but are visually and structurally very different when plotted. Introduced by statistician Francis Anscombe in 1973, it emphasizes the importance of visualizing data before relying on statistical summaries or building models.

Each dataset in Anscombe's Quartet consists of 11 data points with two variables (xx and yy), but their relationships differ dramatically. This highlights how summary statistics alone can be misleading and why scatter plots or other visualizations are crucial for data analysis.

Characteristics of the Quartet:



Dataset 1:

Exhibits a clear linear relationship between xx and yy.

Fits a linear regression model well, as expected from the correlation coefficient (r=0.816r = 0.816).

Dataset 2:

Appears to follow a **nonlinear relationship**, resembling a curve.

Despite having the same correlation as Dataset 1, it does not suit a linear regression model.

Dataset 3:

Includes an **outlier** that heavily influences the regression line.

The outlier gives the false impression of a linear relationship in summary statistics but would mislead any model built on this data.

Dataset 4:

Contains a significant **high-leverage point** (a single extreme value for xx).

While the other data points form a near-constant y, this point skews the correlation and statistical summaries.

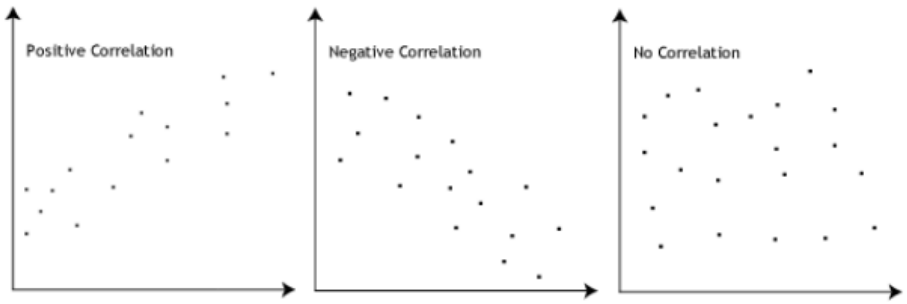
Anscombe's Quartet is a powerful reminder that data visualization is an essential step in any analysis. It helps uncover nuances, ensuring better decision-making and avoiding misinterpretations caused by relying solely on numerical summaries.

Q3. What is Pearson's R?

Answer:

Pearson's R, or the **Pearson correlation coefficient**, is a statistic used to measure the strength and direction of the **linear relationship** between two continuous variables. Developed by Karl Pearson, it provides a value between −1-1 and 11, where:

- **r = 1** : Perfect positive linear correlation (as one variable increases, the other also increases proportionally).
- **r = –1** : Perfect negative linear correlation (as one variable increases, the other decreases proportionally).
- **r = 0** : No linear correlation (no consistent relationship).
- **0 < r < 1** : Positive linear correlation (the closer to 1, the stronger the relationship).
- **-1 < r < 0** : Negative linear correlation (the closer to -1, the stronger the relationship).



Formula for Pearson's R:

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

Where:
xi and yi: Individual data points for the two variables x and y.
$\bar{x}$ and $\bar{y}$: Mean values of x and y, respectively.

The numerator calculates the covariance between x and y, while the denominator normalizes it using their standard deviations, ensuring r stays between -1 and 1.

Key Features of Pearson's R:

Direction:

1. Positive (r>0r > 0): Both variables move in the same direction.
2. Negative (r<0r < 0): As one variable increases, the other decreases.

Strength:

1. Strong relationship: r is close to 1 or -1.
2. Weak relationship: r is close to 0.

Limitations:

- Pearson's R only measures linear relationships and does not work well for non-linear patterns.
- It is sensitive to **outliers**, which can skew the results.
- A high r **value** does not imply causation; it only indicates association.

In summary, **Pearson's R** is a useful tool for understanding linear relationships, but it is essential to visualize data and check assumptions to avoid misleading interpretations.

Q 4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:

Scaling is a data pre-processing technique used to adjust the values of features in a dataset to fit within a specific range or distribution. This ensures that all features have comparable magnitudes, which is crucial for many machine learning algorithms to function properly and scaling helps prevent one feature from dominating others.

There are several reasons to use scaling:

1. **Consistency:** Features often vary in magnitude, units, and ranges. Scaling ensures they are on the same scale.
2. **Improved Algorithm Performance**: Scaling helps algorithms converge faster and produce more accurate results.
3. **Preventing Bias**: Without scaling, algorithms may weigh features with higher magnitudes more heavily, leading to incorrect modelling.

Types of Scaling

- **Normalized Scaling (Normalization)**:

  Adjusts feature values to a fixed range, typically **[0, 1]** or **[-1, 1]**.

  Formula :

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Pros: Maintains the original data distribution and ensures comparability of features.

Cons: Highly sensitive to outliers, as extreme values can skew the scaling.
Use Case: When features are measured on different scales but their distribution is not normal.
Implemented using MinMaxScaler in Python (Scikit-Learn).

- **Standardized Scaling (Standardization)**:
  Centres data around a mean of **0** and scales to a standard deviation of **1**.
  Formula:

$$z = \frac{x - \mu}{\sigma}$$

  Where μ is the mean and σ is the standard deviation.
  Pros: Reduces the effect of outliers by focusing on deviations from the mean.
  Cons: Does not bound the data to a specific range.
  Use Case: When the dataset follows a normal distribution or in algorithms like PCA.
  Implemented using StandardScaler in Python (Scikit-Learn).

In summary, Scaling is a crucial step in pre-processing to ensure effective and fair training of machine learning models. The choice between normalization and standardization depends on the data's characteristics and the requirements of the machine learning algorithm.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer:

The Variance Inflation Factor (VIF) is a measure of multicollinearity in a regression model. It quantifies how much the variance of a regression coefficient is inflated due to the correlation among independent variables.

VIF becomes infinite when there is **perfect multicollinearity**, meaning one variable is an exact linear combination of one or more other variables. In such cases, the regression model cannot distinguish between the perfectly correlated variables, leading to computational issues.

Formula for VIF:

$$VIF_i = \frac{1}{1 - R_i^2}$$

Here, R^2 is the coefficient of determination when the variable is regressed on all other predictors.

- If $R^2 = 1$, the denominator becomes zero, resulting in VIF = ∞

Example :
If $X_2 = 2 \times X_1$ then X_1 and X_2 are perfectly collinear, and VIFVIFVIF for both variables becomes infinite.

Causes of Infinite VIF:

- **Highly Correlated Predictors:**
  When one variable is a perfect function of others (e.g., duplicate columns, transformations like X2 = aX1 + bX2).

- **Dummy Variable Trap:**
  In categorical data, including all dummy variables without dropping one reference category leads to perfect collinearity.
- **Coding or Data Errors:**
  Mistakes like duplicated variables or incorrect pre-processing can introduce perfect multicollinearity.

In summary, Infinite VIF values highlight severe multicollinearity, making regression results unreliable. Addressing these issues ensures a robust and interpretable model, improving the accuracy of predictions and insights.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer:

A **Q-Q (Quantile-Quantile) Plot** is a graphical tool used to compare the distribution of a dataset to a reference distribution (commonly a normal distribution). It plots the quantiles of the observed data against the quantiles of the reference distribution.

If the data follows the reference distribution, the points on the Q-Q plot align closely along a straight diagonal line.

Deviations from the line indicate departures from the reference distribution, such as skewness, heavy tails, or outliers.

In linear regression, one of the key assumptions is that the residuals (errors) follow a normal distribution. A Q-Q plot helps evaluate this assumption by visualizing the distribution of the residuals.

Uses in Linear Regression:

Assessing Normality of Residuals:

A Q-Q plot of the residuals can confirm whether they are approximately normally distributed.

Normally distributed residuals are crucial for valid statistical inferences (e.g., confidence intervals and hypothesis tests).

Detecting Deviations:

Outliers: Points far from the straight line suggest potential outliers.

Skewness: A curved pattern indicates skewed residuals.

Heavy Tails: Points diverging at the extremes show heavy tails.

Comparing Datasets:

In scenarios with train and test datasets, Q-Q plots can verify whether both datasets come from populations with the same distribution, ensuring consistent model performance.

Advantages of Q-Q Plots:

Versatile: Can be used with small or large datasets.

Comprehensive: Detects shifts in location, scale, symmetry, and presence of outliers.

Visual Insight: Offers a clear visual summary of distributional characteristics.

Example: Interpreting a Q-Q Plot

Straight Line: Residuals follow a normal distribution.

S-shaped Curve: Indicates skewness in the data.

Deviations at Ends: Suggests heavy tails or outliers.

In summary, Q-Q plots are an essential diagnostic tool in linear regression to validate the normality assumption of residuals. Ensuring normal residuals enhances the reliability of regression models, enabling robust predictions and accurate statistical conclusions.