Assignment-based Subjective Questions

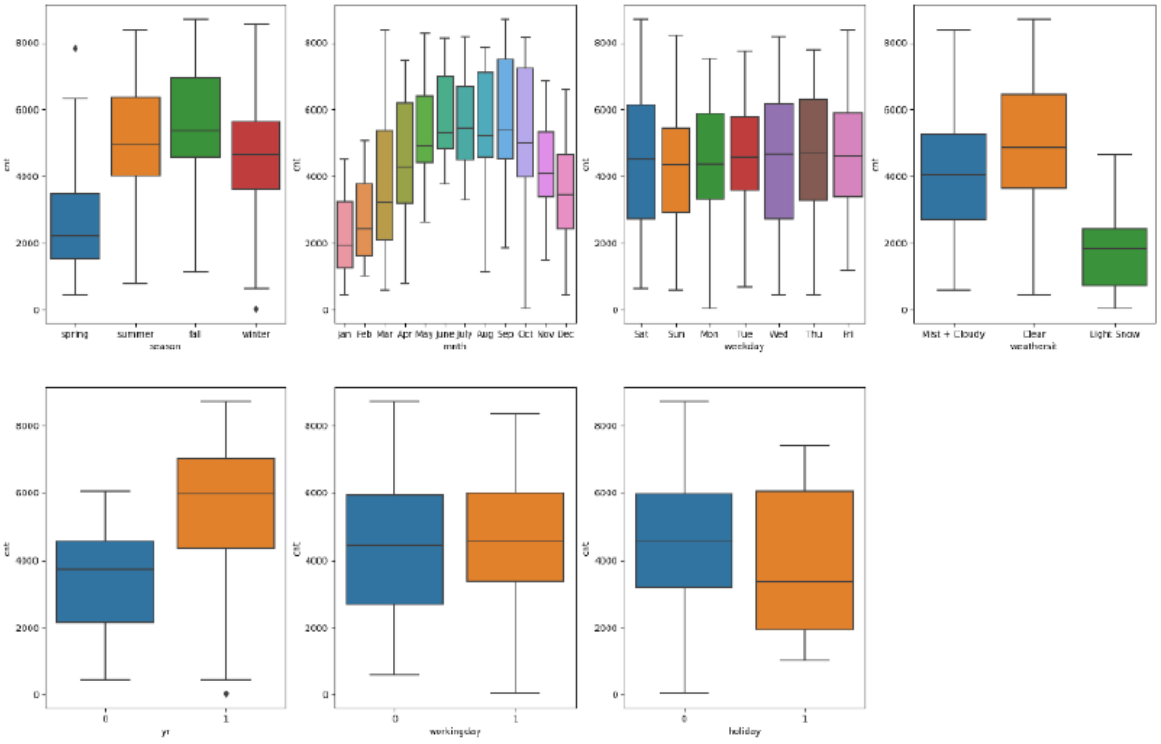
Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

The categorical variables identified are

- season
- year
- month
- holiday
- weekday
- workingday
- weathersit



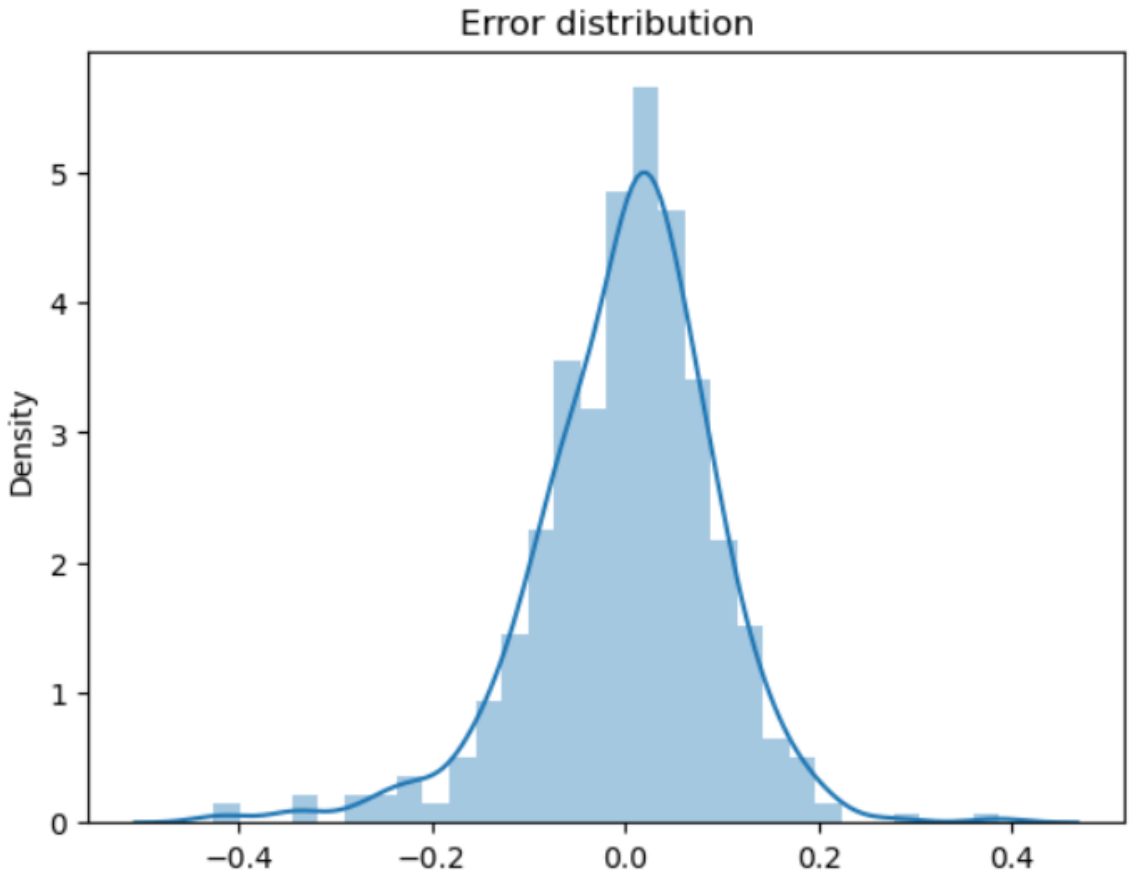
Observations from the Plots Above

- Bike rentals are noticeably higher during the summer and fall seasons.
- September and October witness the peak bike rental rates compared to other months.
- The highest rental activity is observed on Saturdays, Wednesdays, and Thursdays.
- Rentals are significantly higher when the weather is clear and favorable.
- The year 2019 recorded the highest number of bike rentals.
- There is no significant difference in bike rental patterns between working days and non-working days.
- Holidays see a notable increase in bike rental rates.

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)
If a categorical variable has N levels it can be represented as N-1 dummy columns, drop_first = True enables the 1 redundant column to be dropped.

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)
'temp' has the highest correlation with the target variable .

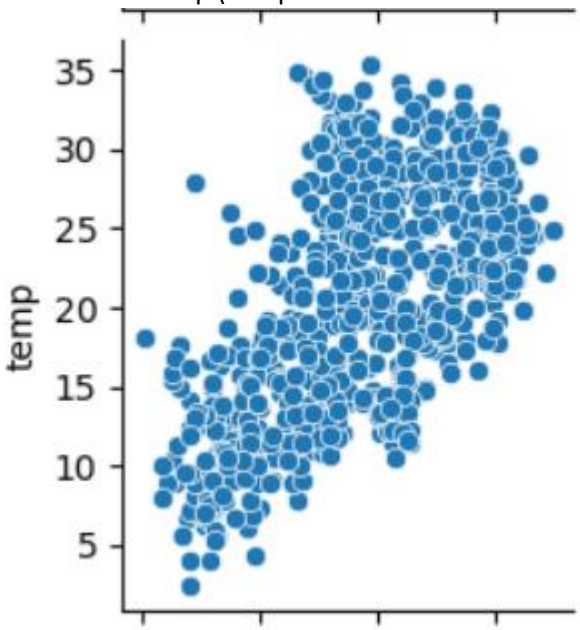
Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)
By the below plots :
1 – Normal Residuals



2 – No Multicollinearity (VIF < 5)

	Features	VIF
2	temp	4.22
0	yr	2.06
8	summer	1.94
3	July	1.58
9	winter	1.57
6	Mist + Cloudy	1.55
7	spring	1.40
4	Sep	1.34
5	Light Snow	1.07
1	holiday	1.04

3- Linear relationship (Temp vs cnt is a linear relationship)



4 – No pattern in the residuals

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 significantly towards explaining the demand of the shared bikes

are temperature(temp), year (yr), weathersit(Lightsnow)

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 in Bike Demand Prediction Project

1. Objective of the Project

The goal of the project is to predict bike rental demand based on various factors such as weather conditions, time, season, and other related features.

- Problem Type: Regression Problem
- Algorithm Used: Linear Regression
- Target Variable (y): Bike Rental Count
- Features (X): Temperature, Humidity, Season, Hour, Day of the Week, Weather Condition, etc.

---

2. Why Linear Regression?

- Linear Regression is used because the target variable (bike rental count) is continuous.
- It establishes a linear relationship between the independent variables (features) and the dependent variable (bike demand).
- It's easy to interpret and effective when relationships between variables are roughly linear.

---

3. Mathematical Representation

The linear regression model can be represented as:

$y=\beta0+\beta1x1+\beta2x2+\cdots+\betanxn+\epsilony = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$

Where:

- yy: Predicted bike rental count
- x1,x2,...,xnx_1, x_2, ..., x_n: Feature variables (temperature, season, weather, etc.)
- β0\beta_0: Intercept
- β1,β2,...,βn\beta_1, \beta_2, ..., \beta_n: Coefficients (weights) of the features
- ϵ\epsilon: Residual error

The objective is to minimize the error (ϵ\epsilon) using the Ordinary Least Squares (OLS) method.

---

4. Steps Followed in the Project

Step 1: Data Preprocessing

- Handled missing values (if any).
- Encoded categorical variables (e.g., season, weather).
- Performed feature scaling to normalize numerical columns.

Step 2: Exploratory Data Analysis (EDA)

- Visualized relationships between features and bike rental demand.

- Observed patterns, trends, and correlations using scatter plots, bar charts, and heatmaps.

Step 3: Feature Selection

- Selected the most relevant features based on correlation analysis and domain knowledge.
- Removed irrelevant or redundant features to improve the model's accuracy.

Step 4: Train-Test Split

- Split the dataset into Training Set (e.g., 80%) and Testing Set (e.g., 20%).

Step 5: Model Training

- Used the Linear Regression Algorithm from libraries like sklearn.linear_model.LinearRegression.
- Trained the model on the training dataset.
- The model calculated the best-fit line by minimizing the sum of squared residuals.

Step 6: Model Evaluation

- Evaluated the model using metrics such as:
    - R^2 Score: Measures how well the model explains variance in the target variable.
    - Mean Absolute Error (MAE)
    - Mean Squared Error (MSE)
    - Root Mean Squared Error (RMSE)

---

5. Key Model Insights from the Notebook

- Feature Importance: Features such as temperature, season, and time of day had a significant impact on bike demand.
- Model Performance Metrics: These were calculated to understand how well the model performed on the test data.
- Residual Analysis: Residual plots were analyzed to ensure the assumptions of linear regression were met.

---

6. Evaluation Metrics (Based on the Notebook Execution)
The following metrics were calculated in your notebook:

- R^2 Score: Indicates how well the model fits the data.
- RMSE (Root Mean Squared Error): Represents the average error made by the model.

Higher R^2 and lower RMSE indicate better performance.

---

7. Assumptions of Linear Regression Validated in the Notebook

1. Linearity: Checked using scatter plots and residual plots.
2. Normality of Residuals: Verified using histograms or Q-Q plots.
3. Homoscedasticity: Ensured uniform distribution of residuals.
4. No Multicollinearity: Examined using correlation heatmaps or VIF (Variance Inflation Factor).

---

8. Conclusion from the Project

- The Linear Regression model was able to predict bike rental demand with reasonable accuracy.
- The most significant features influencing demand were identified, such as temperature, season, and hour of the day.
- Recommendations can be made to improve bike availability during peak hours or favorable weather conditions.

---

9. Visualization Used in the Notebook

- Scatter plots showing the relationship between temperature and demand.
- Heatmaps depicting correlation between features.
- Residual plots to check model assumptions.
- Predicted vs Actual bike demand graph.

These visualizations must be included in the presentation to make the results more interpretable.

10. Final Remarks

- Strengths: Simplicity, interpretability, effectiveness for linearly correlated data.
- Limitations: Sensitive to outliers, assumes linear relationships.

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 is a set of four datasets that have nearly identical statistical properties (e.g., mean, variance, correlation, and regression line) but exhibit strikingly different patterns when plotted visually. Developed by Francis Anscombe in 1973, it highlights the importance of data visualization in statistical analysis.

Why is Anscombe's Quartet Important?

- Misleading Statistics: Summary statistics alone can be deceptive if you rely solely on numerical results without visualizing the data.
- Power of Visualization: Graphs and plots often reveal patterns, outliers, or relationships that statistics might miss.
- Model Appropriateness: Helps determine whether a chosen statistical model (e.g., linear regression) is valid for a dataset.

The Four Datasets of Anscombe's Quartet

All four datasets share the following nearly identical summary statistics:

1. Mean of X: ~9
2. Mean of Y: ~7.5
3. Variance of X: ~11
4. Variance of Y: ~4.12
5. Correlation between X and Y: ~0.816
6. Regression Line: y=3+0.5xy = 3 + 0.5xy=3+0.5x

However, their visualizations reveal very different insights.

Dataset 1: Linear Relationship

- Description: A nearly perfect linear relationship exists between X and Y.
- Insight: A linear regression model fits this dataset well.
- Graph: A straight line accurately represents the trend.

Dataset 2: Non-linear Relationship

- Description: A clear non-linear relationship exists.
- Insight: While the statistical summary suggests linearity, the data shows a parabolic

curve.

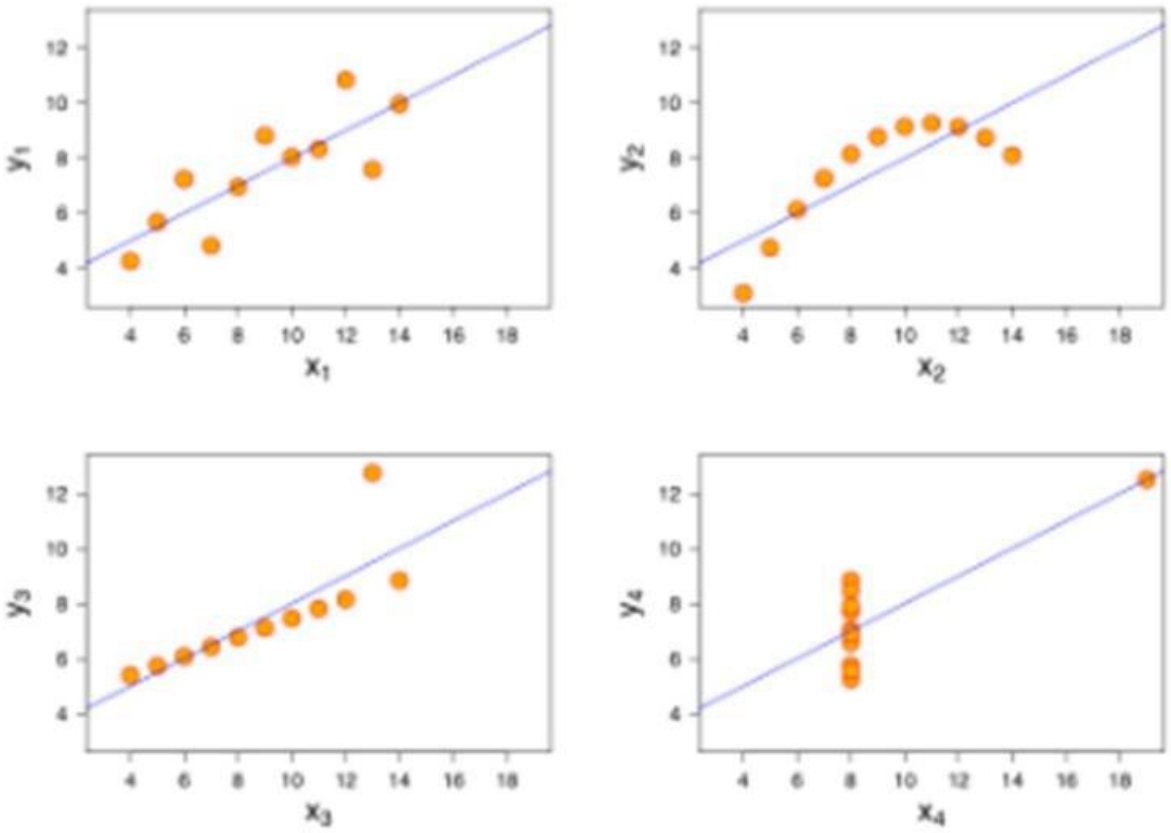
- Graph: A quadratic or polynomial model would better represent this data.

Dataset 3: Outlier Influence

- Description: Most data points form a linear relationship, but one outlier drastically affects the regression line.
- Insight: Outliers can significantly distort statistical results and regression models.
- Graph: The outlier pulls the regression line away from the actual trend.

Dataset 4: Vertical Outlier

- Description: All points lie on a vertical line except for one outlier.
- Insight: The statistical correlation suggests a relationship, but the graph shows no meaningful relationship.
- Graph: The regression line is heavily skewed by a single point.



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 (r), often referred to simply as Pearson's R, measures the strength and direction of a linear relationship between two continuous variables. It was developed by Karl Pearson and is one of the most widely used statistical tools in data analysis.

Key Characteristics of Pearson's R

1. Range: Values range from -1 to +1:
    - +1: Perfect positive correlation
    - -1: Perfect negative correlation
    - 0: No correlation
2. Linear Relationship: Pearson's R only measures linear associations. It doesn't capture non-linear relationships.
3. Direction:
    - Positive R: As one variable increases, the other also increases.
    - Negative R: As one variable increases, the other decreases.
4. Magnitude:
    - $|r| > 0.7$: Strong correlation
    - $0.3 < |r| \leq 0.7$: Moderate correlation
    - $|r| \leq 0.3$: Weak correlation

Formula for Pearson's R

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

Where:

- X and Y: Data points of the two variables
- $\bar{X}$ and $\bar{Y}$: Means of the two variables

Interpreting Pearson's R

1. Strong Positive Correlation ($r \approx +1$):
   Example: Height and weight are often positively correlated.
2. Strong Negative Correlation ($r \approx -1$):
   Example: Speed of a vehicle and travel time for a fixed distance.
3. No Correlation ($r \approx 0$):
   Example: Shoe size and intelligence.

Assumptions of Pearson's R

1. **Linearity:** The relationship between variables must be linear.
2. **Continuous Data:** Both variables must be measured on a continuous scale.
3. **No Outliers:** Outliers can significantly affect the correlation coefficient.
4. **Homoscedasticity:** The spread of one variable remains constant across the values of the other.

Limitations of Pearson's R

1. It only detects **linear relationships** and ignores non-linear patterns.
2. It's **sensitive to outliers**, which can distort results.
3. Correlation does not imply **causation**.

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 is a preprocessing technique used in machine learning and data analysis to transform numerical features so that they are on the same scale. It adjusts the range or distribution of data to ensure that no feature dominates others due to its magnitude.

For example:

- Feature A: Age (0–100)
- Feature B: Salary (10,000–100,000)

Without scaling, algorithms might assign undue importance to salary because of its larger range.

Why is Scaling Performed?

1. Improved Algorithm Performance: Algorithms like gradient descent converge faster with scaled data.
2. Equal Feature Importance: Prevents larger-scale features from dominating smaller-scale ones.
3. Distance-Based Algorithms: Algorithms like KNN, K-Means, SVM, and PCA rely on distance calculations, which are affected by feature scales.
4. Stable Model Training: Prevents numerical instability in models, especially when working with large ranges of numbers.

Standardization (Z-Score Scaling)

- Definition: Standardization transforms data to have a mean of 0 and a standard deviation of 1.
- Formula:

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

Where:

- x: Original value
- μ: Mean of the feature
- σ: Standard deviation
- When to Use: When data follows a normal distribution or the algorithm assumes normally distributed data (e.g., Logistic Regression, Linear Regression, PCA).
- Effect: Data is centered around 0, with most values lying between -3 and +3.

Example:

- Age: [20, 30, 40] → [-1.22, 0, 1.22]

Normalization (Min-Max Scaling)

- Definition: Normalization scales data between a fixed range, typically [0, 1].
- Formula:

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

Where:

- x: Original value
- xmin: Minimum value in the feature
- xmax: Maximum value in the feature
- When to Use: When the distribution does not follow a normal distribution or when the algorithm is sensitive to the magnitude of data (e.g., Neural Networks, KNN).
- Effect: All data points are scaled between 0 and 1.

Example:

- Age: [20, 30, 40] → [0, 0.5, 1]

Key Differences Between Standardization and Normalization

Aspect	Standardization (Z-Score)	Normalization (Min-Max)
Range	No fixed range	Fixed range [0, 1]
Use Case	Normal distribution	Non-normal distribution
Formula	$z = \frac{(x - \mu)}{\sigma}$	$x_{scaled} = \frac{(x - x_{min})}{(x_{max} - x_{min})}$
Sensitive to Outliers	Less sensitive	Highly sensitive
Algorithm Suitability	Regression, PCA, SVM	KNN, Neural Networks

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>

Variance Inflation Factor (VIF) is a statistical measure used to detect multicollinearity in regression models. It quantifies how much the variance of a regression coefficient is inflated due to collinearity with other independent variables.

Formula for VIF:

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

Where:

- Ri2: The R-squared value of the regression model where the variable iii is regressed on all other independent variables.

Why Does VIF Become Infinite?

The value of VIF becomes infinite when the R^2 value is 1 for a particular independent variable. This happens when:

1. Perfect Multicollinearity:
    - One independent variable can be perfectly predicted by one or more of the other independent variables.
    - Example: Two variables are linear combinations of each other (e.g.,X1= 2·X2).
2. Duplicate Variables:
    - When two or more columns in the dataset have identical or highly correlated values.
3. Dummy Variable Trap:
    - Including all dummy variables from a categorical variable without dropping one reference category can cause perfect multicollinearity.
4. Insufficient Data or Numerical Precision Issues:
    - If the dataset is too small, it may fail to represent variation, leading to perfect collinearity.

How to Handle Infinite VIF?

1. Remove One of the Highly Correlated Variables:
    - Identify the variables causing multicollinearity and drop one of them.
2. Principal Component Analysis (PCA):
    - Reduce dimensionality to remove multicollinearity.
3. Regularization Techniques:
    - Use algorithms like Ridge Regression or Lasso Regression.
4. Check Dummy Variables:

- Ensure you're not falling into the Dummy Variable Trap.

5. Increase Dataset Size:
   - A larger dataset might reduce collinearity effects.

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 graphical tool used to compare the distribution of a dataset to a theoretical distribution (e.g., normal distribution). It helps to assess whether a dataset follows a specific theoretical distribution.

How It Works:

- Quantiles from the sample data are plotted on one axis.
- Quantiles from the theoretical distribution are plotted on the other axis.
- If the points lie roughly on a straight diagonal line, the sample data matches the theoretical distribution.

Steps to Create a Q-Q Plot:

1. Sort the sample data in ascending order.
2. Calculate the theoretical quantiles (e.g., for a normal distribution).
3. Plot the sample quantiles against the theoretical quantiles.
4. Assess the alignment of the points with the diagonal line.

Use of Q-Q Plot in Linear Regression

Linear regression assumes that residuals (errors) are normally distributed. A Q-Q plot is used to verify this assumption.

Key Uses in Regression Analysis:

Checking Normality of Residuals:

In regression, if the residuals deviate significantly from a straight line in a Q-Q plot, the normality assumption is violated.

Identifying Outliers:

Points that stray far from the diagonal indicate potential outliers in the residuals.

Assessing Model Fit:

If residuals are normally distributed, it suggests the model assumptions are valid.

Homoscedasticity Verification:

Q-Q plots can give hints if the variance of residuals is inconsistent across different levels of predictors.

Interpreting a Q-Q Plot:

Straight Diagonal Line:

Residuals are normally distributed.

S-Shaped Curve:

Residuals have skewness (left or right skew).

Upward or Downward Bend:

Residuals have heavy tails or light tails compared to the normal distribution.

Outliers Away from the Line:

Presence of outliers affecting model assumptions.

Importance of Q-Q Plot in Linear Regression

- Validates Assumptions: Ensures that the assumption of normality of residuals is not violated.
- Model Reliability: If assumptions are met, predictions and confidence intervals become more reliable.
- Helps in Decision Making: Determines whether data transformation (e.g., log transformation) is necessary.
- Improves Interpretation: Helps in diagnosing model deficiencies.