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)

Based on the variances of the categorical variables in the dataset, we can infer the following about their potential effects on the dependent variable.

Effect of All Categorical Variables:

1. **Season:**
   Across all seasons, the average count of bike demand in 2019 is higher than in 2018.
   Summer and fall seasons have higher demand counts compared to winter and spring seasons which are the lowest.
   There's a notable increase in demand in the fall season for 2019, indicating possibly higher usage in that period compared to 2018.

2. **Month**:
   Bike demand peak around mid-year (May to August) for both years, which is typical for warmer months. In 2019, the demand counts are consistently higher throughout the year compared to 2018.
   A drop in demand is seen in December and Jan-March for both years, likely due to colder weather.

3. **Holiday:**
   There is no significant difference between holidays and non-holidays in terms of the average count of bike demand.
   However, the variability in the demand counts on holidays is higher, suggesting that while the median/average might not differ much, the usage pattern can vary more on holidays.

4. **Weekday:**
   The demand counts across weekdays show slight variations. Both years have similar distributions across the weekdays, with 2019 having generally higher median/average demand.
   The demand counts seem to be more consistent across the week.

5. **Working days**
   Demands on working days are slightly lower in terms of median/average count compared to non-working days for both years. 2019 has higher counts on both working and non-working days compared to 2018.

6. **Weather situation (weathersit)**
   `Weather situation 1 (clear or partly cloudy)` has the highest average demand, followed by `situation 2 (misty)`.
   `Situation 3 (light rain or snow)` has significantly lower demand counts.

For all weather situations, 2019 shows higher demand counts compared to 2018.

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)

The dummy variable trap is a situation in statistics and machine learning where one or more dummy variables (categorical variables that are encoded into binary form) are redundant. This can lead to multicollinearity, making it difficult to estimate the relationship between variables and interpret regression coefficients.

When you have a categorical variable with n categories, and you create n dummy variables, the sum of these variables will always equal 1 (perfect multicollinearity). This redundancy confuses regression models because one variable can always be predicted from the others.

To avoid the dummy variable trap, we drop one of the dummy variables (called the "reference category"). This way, we use n−1 dummy variables for a categorical variable with n categories. The dropped category becomes the baseline, and the coefficients of the remaining dummy variables are interpreted relative to this baseline. So we use the drop_first=True parameter in get_dummy, which automatically drop the first category of each categorical variable to prevent this issue.

Example
Suppose you have a categorical variable, Seasons, with four categories: season : season (1:spring, 2:summer, 3:fall, 4:winter).

Step 1: Create dummy variables
We encode these categories as dummy variables:

Seasons	spring (Dummy)	summer (Dummy)	fall (Dummy)	winter(dummy)
spring	1	0	0	0
summer	0	1	0	0
fall	0	0	1	0
winter	0	0	0	1

If you include all four dummy variables in a regression model, their sum will always be 1:

spring + summer + fall + winter =1

This redundancy causes multicollinearity.
To avoid the dummy variable trap, we drop one column (e.g., spring) and use n−1=3 dummy variables:

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)

There is a very strong positive linear relationship between **registered** and **cnt** (95). But there is multicollinearity in **cnt** and **registered** hence **temp and atemp** should be considered highly correlated variables after r**egistered and casual** with target variable.

This indicates that the majority of bike demand comes from registered users, and as the count of registered users increases, the total bike demand (cnt) also increases proportionally.

"temp" and "atemp": Both temperature and "atemp" (apparent feeling temperature) have strong positive correlations with "cnt" (0.63 and 0.63 respectively), implying that higher temperatures are associated with higher demand (count).

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)

- **Residual Analysis:** After fitting the linear regression model, examine the residuals (the differences between the actual values and the predicted values).
- **Visual Inspection:** Create plots to visually inspect the residuals:
    - **Residual Plot:** Plot the residuals against the predicted values or the independent variables. Look for patterns, such as a funnel shape (indicating heteroscedasticity), or a curved pattern (suggesting non-linearity).
    - **Histogram of Residuals:** Check if the residuals are approximately normally distributed. This is a key assumption of linear regression.
- **Statistical Tests:**
    - Multicollinearity Test using Variance Inflation Factor (VIF):
    - Autocorrelation Test using Durbin-Watson Test:
    - Overall Model Fit using F-test:
    - Significance of Individual Coefficients using t-test for each coefficient:
    - Goodness of Fit Measures using R-squared and Adjusted R-squared:

Assumptions of Linear Regression and Validation Tests

1. **Linearity:** The relationship between the independent and dependent variables is linear.
   Test: Residual plot (scatter plot of residuals vs. predicted values) should show no pattern.
   Steps:
   Fit the linear regression model.
   Plot residuals against predicted values.
   Check for any systematic pattern.
2. **Independence of Errors: Residuals should be independent.**
   Test: Durbin-Watson test.
   Steps:
   Perform the Durbin-Watson test on the residuals.
   A value close to 2 indicates no autocorrelation.
3. **Homoscedasticity: Constant variance of errors.**

Test: Breusch-Pagan test or White's test.
Steps:
Perform the Breusch-Pagan test on the residuals.
If the p-value is greater than 0.05, homoscedasticity is not violated.

4. **Normality of Errors: Residuals should follow a normal distribution.**
   Test: Shapiro-Wilk test, Kolmogorov-Smirnov test, or Q-Q plot.
   Steps:
   Perform the Shapiro-Wilk test on the residuals.
   A p-value greater than 0.05 suggests normality.
   Alternatively, use a Q-Q plot to visually assess normality.
5. **No Multicollinearity: Independent variables should not be highly correlated.**
   Test: Variance Inflation Factor (VIF).
   Steps:
   Calculate VIF for each independent variable.
   VIF values greater than 5 or 10 indicate high multicollinearity.

- **Overall Analysis:**
  - The residuals are following the normally distributed with a mean 0.
  - We are confident that the model fit isn't by chance, and has decent predictive power. The normality of residual terms allows some inference on the coefficients.
  - The residuals appear to be randomly scattered around the horizontal line at zero. There isn't a clear pattern or trend visible in the scatter of points.

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)

Temperature (0.5211):** For every unit increase in temperature, bike demand increases by 0.5211 units, holding all other factors constant. (Strong positive impact) bike-sharing service.
Year (0.2328): Demand increases significantly in 2019 compared to the reference year (2018), suggesting increased bike usage in the second year. with an increase of 0.2328 unit

Winter (0.1374): Winter has a slightly higher positive impact compared to summer, with demand increasing by 0.1374 units.

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)

Linear regression is a statistical method used to model the relationship between one dependent variable (also called the target or response variable) and one or more independent variables (also called predictors or features). The goal is to find the best-fit line (or hyperplane in higher

dimensions) that minimizes the difference between the predicted values and the actual values of the dependent variable.

Types of Linear Regression

Simple Linear Regression: Models the relationship between one dependent variable and one independent variable using a straight line.

$y=\beta_0+\beta_1 x+\epsilon$

where:
y is the dependent variable.
x is the independent variable.
β_0 is the intercept (value of y when x=0).
β_1 is the slope (rate of change of y with respect to x).
ϵ is the error term.

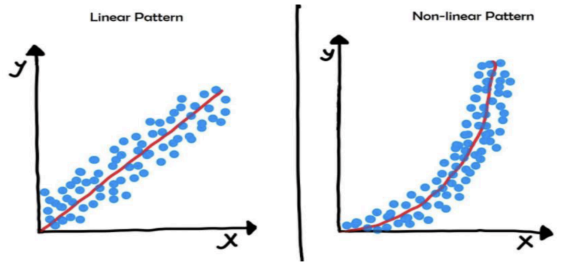
Multiple Linear Regression: Extends the simple linear regression by using multiple independent variables.

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon$

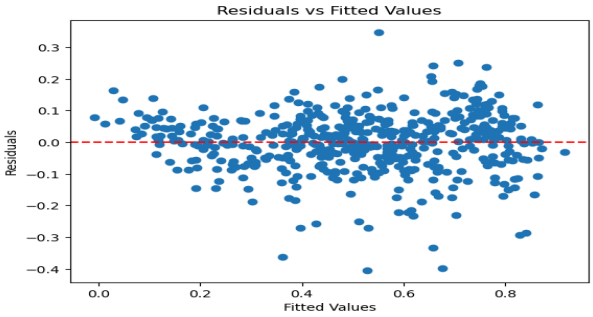
y: Dependent variable
$X_1, X_2, ..., X_n$: Independent variables
β_0: Intercept
$\beta_1, \beta_2, ..., \beta_n$: Slopes
ε: Error term

Assumptions of Linear Regression

Linear regression relies on several key assumptions:
Linearity: The relationship between the dependent variable and independent variables is linear.



Independence: Observations are independent of each other.
Homoscedasticity: The variance of residuals (errors) is constant across all levels of the independent variables.



Normality: Residuals are normally distributed.

No Multicollinearity: Independent variables are not highly correlated with each other.

Steps in the Algorithm:

a) Formulating the Hypothesis

The model assumes a linear relationship between the predictors and the target:

$y=X\beta+\epsilon$

where:

y is the vector of observed values for the dependent variable.

X is the matrix of independent variables (including a column of 1s for the intercept).

β is the vector of coefficients (parameters to be estimated).

ϵ is the vector of residuals.

b) Cost Function (Ordinary Least Squares)

The most common method for fitting the linear regression model is Ordinary Least Squares (OLS). This minimizes the sum of the squared residuals:

$$\text{Cost} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where:

y_i is the actual value.

$\hat{y}_i$ is the predicted value ($X\beta$).

c) Parameter Estimation

The coefficients β are estimated to minimize the cost function.The closed - form solution for β is:

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

where:

X^T X is the transpose of X

$(X^T X)^{-1}$ is the inverse of $X^T X$.

d) Making Predictions

Once the coefficients are estimated, predictions for new inputs can be made as:

$\hat{y} = X\beta$

e) Evaluation Metrics:

R-squared: Measures the proportion of variance in the dependent variable explained by the model.

Adjusted R-squared: Penalizes the model for including unnecessary variables.

Mean Squared Error (MSE): Measures the average squared difference between actual and predicted values.

Root Mean Squared Error (RMSE): Square root of MSE, in the same units as the target variable.

Regularization (Optional)

When the model has too many features or multicollinearity, regularization methods can help by adding penalties to the cost function:

Lasso Regression (L1 Regularization): Adds a penalty proportional to the absolute value of coefficients.

$$\text{Cost} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Ridge Regression (L2 Regularization): Adds a penalty proportional to the square of coefficients.

$$\text{Cost} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Elastic Net: Combines L1 and L2 penalties.

Interpreting the Coefficients

The sign and magnitude of coefficients indicate the direction and strength of the relationship between predictors and the target.

Example:
A positive coefficient indicates an increase in the predictor variable increases the dependent variable.
A negative coefficient indicates an inverse relationship.

Scaling and Preprocessing

Standardization or normalization is often recommended for linear regression when predictors are on different scales (e.g., temperature in Celsius and windspeed in km/h).

Steps in Building a Linear Regression Model:

Data preparation: Load, clean, and prepare the data.
Model building: Select variables, add a constant, create and fit the model.
Model evaluation: Check coefficients, R-squared, F-statistic, and residuals.
Model refinement: Feature selection, transformations, regularization.
Model validation/evaluation: Predict on test data, evaluate performance.

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)

Anscombe's Quartet is not a formal statistical test but rather a set of four datasets designed by statistician Francis Anscombe in 1973 to demonstrate the importance of graphical representations in

data analysis. These datasets have nearly identical summary statistics, including the mean, variance, correlation, and linear regression line, yet they appear very different when graphed.

Purpose of Anscombe's Quartet:
The purpose of Anscombe's Quartet is to illustrate the importance of visualizing data before analyzing it. It demonstrates that relying solely on summary statistics can be misleading, and different datasets with the same statistical properties can have very different distributions and patterns.

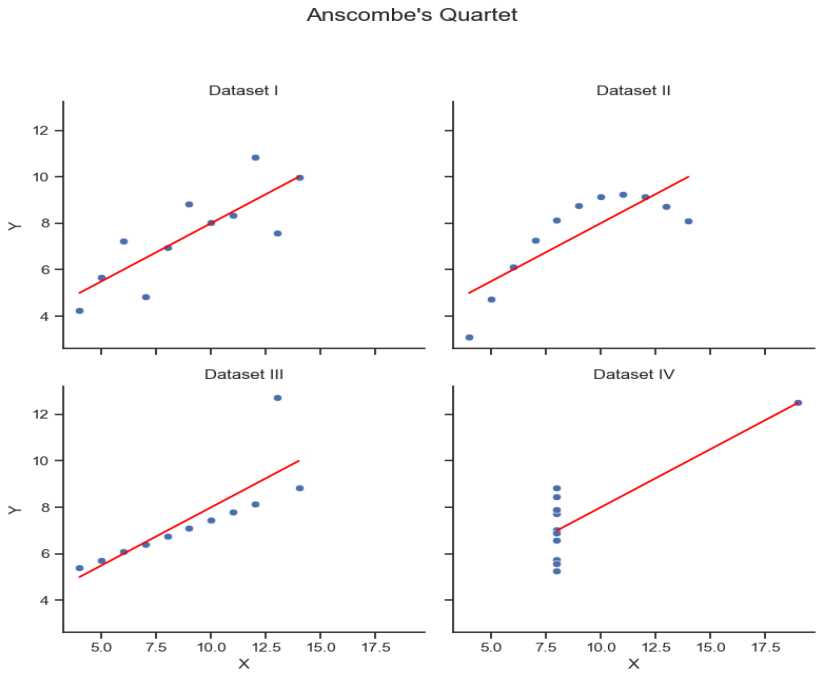
Key Insights from Anscombe's Quartet:

Graphical Analysis: Even though the four datasets have similar statistical summaries, their scatter plots reveal very different relationships.
Linear Fit: A simple linear regression may not be appropriate for all datasets, as seen in the quartet.
Outliers and Non-linearity: Outliers or non-linear patterns can drastically affect the regression line and other statistical measures, which may not be apparent from summary statistics alone.

The Four Datasets: Each dataset has:
The same mean of x and y.
The same variance of x and y.
The same correlation coefficient between x and y (~ 0.816).
The same linear regression line (y = 3 + 0.5x).

Dataset Examples:

Dataset	X Mean	Y Mean	X Variance	Y Variance	Correlation (r)
Dataset 1	9.00	7.50	11.00	4.12	0.816
Dataset 2	9.00	7.50	11.00	4.12	0.816
Dataset 3	9.00	7.50	11.00	4.12	0.816
Dataset 4	9.00	7.50	11.00	4.12	0.816



Visual Interpretation:

Dataset 1: Linear relationship

A well-fitted linear regression line.
Data points align closely to the regression line.

Dataset 2: Non-linear relationship
A parabolic shape.
The regression line does not fit well.

Dataset 3: Linear relationship with an outlier
Most points align with the regression line.
A single outlier skews the fit and affects statistical measures.

Dataset 4: Vertical outlier
The x-values are nearly identical except for one outlier.
The regression line is highly influenced by this outlier.

What It Demonstrates:
Visualizing Data: Always plot your data to understand its distribution and potential anomalies.
Contextualizing Statistics: Summary statistics should be complemented with visual exploration to ensure that the statistical models being applied are appropriate.

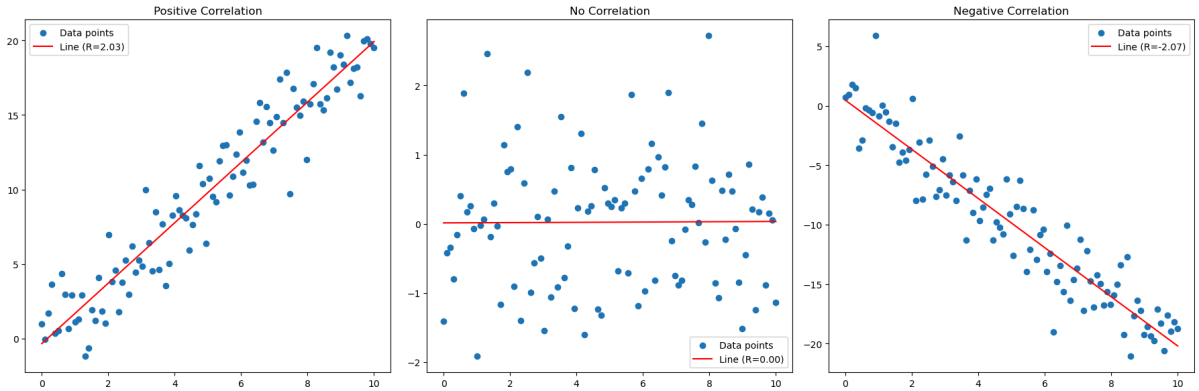
Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's R (Pearson's correlation coefficient) is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. It ranges from -1 to +1, with the following interpretations:

+1: Perfect positive correlation – As one variable increases, the other variable increases in perfect proportion.
0: No correlation – There is no linear relationship between the variables.
-1: Perfect negative correlation – As one variable increases, the other variable decreases in perfect proportion.

Interpretation of Pearson's R:

r > 0: Positive relationship (as one variable increases, the other tends to increase).
r < 0: Negative relationship (as one variable increases, the other tends to decrease).
r = 0: No linear relationship between the variables.



Strength of the correlation:

0.1 to 0.3: Weak positive correlation.
0.3 to 0.5: Moderate positive correlation.
0.5 to 1.0: Strong positive correlation.
-0.1 to -0.3: Weak negative correlation.
-0.3 to -0.5: Moderate negative correlation.
-0.5 to -1.0: Strong negative correlation.

Pearson's R is a measure of the linear association between two continuous variables, meaning it is best used when both variables show a roughly straight-line relationship.

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)

Scaling is the process of transforming the features (variables) of a dataset so that they are on a similar scale or within a specific range. This is particularly important in machine learning algorithms, as many algorithms (like gradient descent, support vector machines, and k-nearest neighbors) perform better when the data is scaled properly.

Why is Scaling Performed?

1. **Improves Algorithm Performance:** Many machine learning algorithms are sensitive to the scale of the data. If features have very different ranges, the algorithm may be biased towards variables with larger magnitudes. Scaling ensures that no variable dominates the others due to its larger scale.
2. **Speeds up Convergence:** Some optimization algorithms, especially those involving gradient descent, converge faster when the data is scaled. Features with large ranges can cause slow convergence or make it difficult for the algorithm to reach the optimal solution.
3. **Distance-based Algorithms:** Algorithms like k-nearest neighbors (KNN) and k-means clustering rely on distance metrics (e.g., Euclidean distance). If features have different scales, the distance calculation will be dominated by the features with larger values. Scaling ensures that each feature contributes equally to the distance computation.

Types of Scaling:

1. Normalized Scaling (Min-Max Scaling):

- **Definition:** Normalization transforms data into a fixed range, typically [0, 1]. It is done by subtracting the minimum value of the feature and dividing by the range (max - min).
- **Formula:**

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

- **When to use:** It is useful when you know that your data should be bounded within a specific range, or when you're dealing with algorithms that assume data within a specific range (e.g., neural networks).
- **Effect:** Normalized data is rescaled to the range [0, 1], which works well when features need

to be within a particular scale, and there are no outliers or they are not too extreme.

- **Example:**
  A feature with values ranging from 50 to 100 is normalized to a range of 0 to 1:
  - Original value: 75
  - Normalized value: (75−50)/(100−50)=0.5(75 - 50) / (100 - 50) = 0.5(75−50)/(100−50)=0.5

2. Standardized Scaling (Z-score Scaling):

- **Definition:** Standardization transforms the data so that it has a **mean of 0** and a **standard deviation of 1**. This is done by subtracting the mean of the feature and dividing by the standard deviation.
- **Formula:**

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

  Where:
  - μ is the mean of the feature.
  - σ sigma is the standard deviation of the feature.
- **When to use:** It is useful when the data does not need to be within a specific range or when you are dealing with algorithms that assume data follows a Gaussian (normal) distribution. It is particularly important for algorithms that assume normally distributed data, such as linear regression, logistic regression, and PCA.
- **Effect:** Standardization doesn't bound the data within a fixed range but adjusts it based on the distribution, which is helpful for algorithms that are sensitive to the variance or outliers in the data.
- **Example:**
  If a feature has values with a mean of 100 and a standard deviation of 20:
  - Original value: 120
  - Standardized value: (120−100)/20=1.0(120 - 100) / 20 = 1.0(120−100)/20=1.0

Key Differences between Normalized and Standardized Scaling:

Aspect	Normalized Scaling (Min-Max)	Standardized Scaling (Z-score)
Range	[0, 1] or [a, b]	Any range, but typically [-∞, ∞]
Formula	x_normalized = (x - min(x)) / (max(x) - min(x))	x' = (x - mean(x)) / std(x)
Mean & Standard Deviation	Not guaranteed to have any specific mean or standard deviation	Mean = 0, Standard deviation = 1
Sensitive to Outliers	Yes, because outliers affect min and max values	Yes, because outliers affect the standard deviation
When to Use	When you want a bounded range (e.g., neural networks)	When the distribution is important or when data has varying ranges but you want to preserve variance

Summary:

- **Normalized Scaling** is used to transform data into a specific range (usually [0, 1]), suitable for algorithms that are sensitive to the magnitude of data.
- **Standardized Scaling** transforms the data to have a mean of 0 and a standard deviation of 1, making it useful for algorithms that assume the data is normally distributed and for those that rely on the variance or covariances in the data.

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)

Variance Inflation Factor (VIF) is used to measure how much the variance of a regression coefficient is inflated due to multicollinearity with other predictor variables. In other words, it quantifies how much a variable is correlated with other variables in the model

Why Can VIF Be Infinite?

VIF can be infinite in cases of **perfect multicollinearity**. This happens when one or more predictor variables in the regression model are perfectly (or nearly perfectly) correlated with other predictor variables. Here's why:

Key Factors Leading to Infinite VIF:

1. **Perfect Correlation Between Variables:**
    - When one predictor is a perfect linear function of another, the VIF for that predictor will be infinite.
2. **Redundant Variables:**
    - Including variables that are redundant or nearly identical (e.g., highly correlated features) can cause multicollinearity and lead to high or infinite VIF values.

How to Handle Infinite VIF:

1. **Remove or Combine Collinear Variables:** Identify and remove one of the variables that are highly correlated with others. Alternatively, you can combine them into a single composite variable if they provide similar information.
2. **Use Principal Component Analysis (PCA):** PCA can be used to reduce multicollinearity by transforming the features into a new set of uncorrelated variables (principal components).
3. **Check for Data Errors:** Sometimes, perfect multicollinearity is caused by data issues, such as duplicate columns or improperly scaled variables.

Example:

VIF = inf (infinity) indicates perfect multicollinearity. Since casual, registered, and cnt are closely related (as cnt is the sum of casual and registered), they exhibit perfect multicollinearity. Including all three in a regression model will cause issues, as the model cannot distinguish their individual effects.

Features		VIF
12	casual	inf
13	registered	inf
14	cnt	inf
0	instant	2211.20
2	yr	842.23
3	mnth	643.00
9	atemp	578.78
8	temp	500.15
10	hum	28.20
1	season	24.82
7	weathersit	14.58
6	workingday	9.46
11	windspeed	5.51
5	weekday	3.26
4	holiday	1.11

Summary:

Infinite VIF occurs when there is perfect multicollinearity, meaning one predictor variable is perfectly correlated with others in the model. This results in an undefined VIF because the R-squared value of the predictor in question is 1, leading to division by zero in the VIF formula. Identifying and addressing this issue (by removing or combining collinear variables) is essential for improving the stability of regression models.

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)

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset with a theoretical distribution, typically a normal distribution. It helps to assess whether the data follows a particular distribution by plotting the quantiles of the dataset against the quantiles of the theoretical distribution.

How a Q-Q Plot Works:
X-axis: Quantiles from the theoretical distribution (e.g., normal distribution).
Y-axis: Quantiles from the sample data.
If the data follows the theoretical distribution closely, the points on the Q-Q plot will approximately lie on a straight diagonal line (the 45-degree line).
Use and Importance of a Q-Q Plot in Linear Regression:

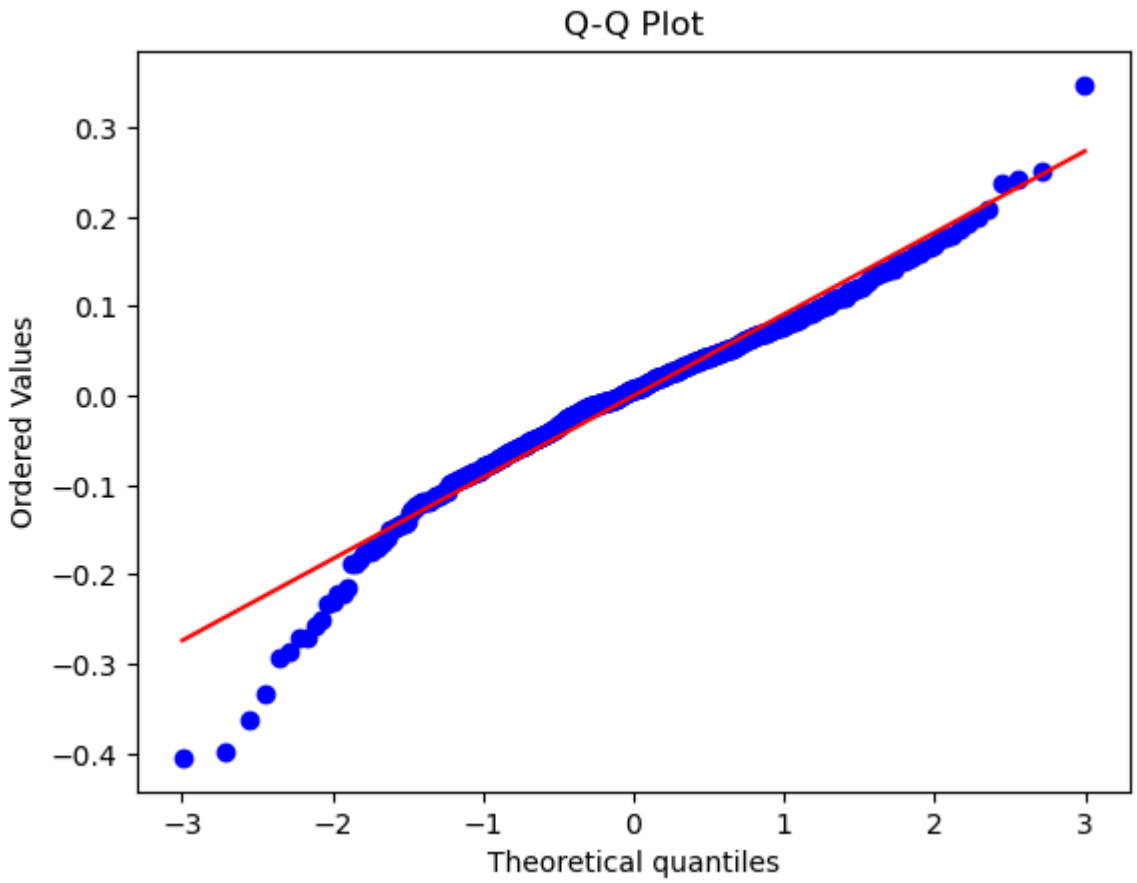
Assessing Normality of Residuals:
One of the key assumptions in linear regression is that the residuals (errors) are normally distributed. A Q-Q plot can be used to check this assumption.
If the residuals are normally distributed, the points in the Q-Q plot will fall close to the straight line. If they deviate significantly, it indicates that the residuals may not be normally distributed, which can affect the validity of hypothesis tests and confidence intervals in the regression model.

Identifying Outliers:
Outliers appear as points that deviate significantly from the line in the Q-Q plot. This helps to detect unusual data points that could disproportionately influence the regression model.

Checking for Heavy Tails:
If the data has heavy tails (more extreme values than expected under normality), the Q-Q plot will show points deviating from the line at the ends (high quantiles).
Conversely, light tails (fewer extreme values than expected) will show the points pulling inward at the ends of the plot.

Model Diagnostic:
The Q-Q plot is part of the diagnostic tools used to assess the quality of a linear regression model. It helps to determine whether the normality assumption of the residuals holds, which is crucial for making valid inferences from the model.

Example:



In a Q-Q plot for residuals of a linear regression model:
If the points lie along the 45-degree line, the residuals are approximately normally distributed, supporting the normality assumption.
If the points systematically deviate from the line, it suggests a departure from normality, indicating potential issues with the regression model, such as skewness, kurtosis, or the presence of outliers.

Summary:
The Q-Q plot is a simple yet powerful tool for visually assessing whether the residuals in a linear

regression model follow a normal distribution. This check is essential because many inferential statistics in regression rely on the assumption of normality of residuals. Detecting deviations early helps to ensure the accuracy and reliability of the regression analysis.