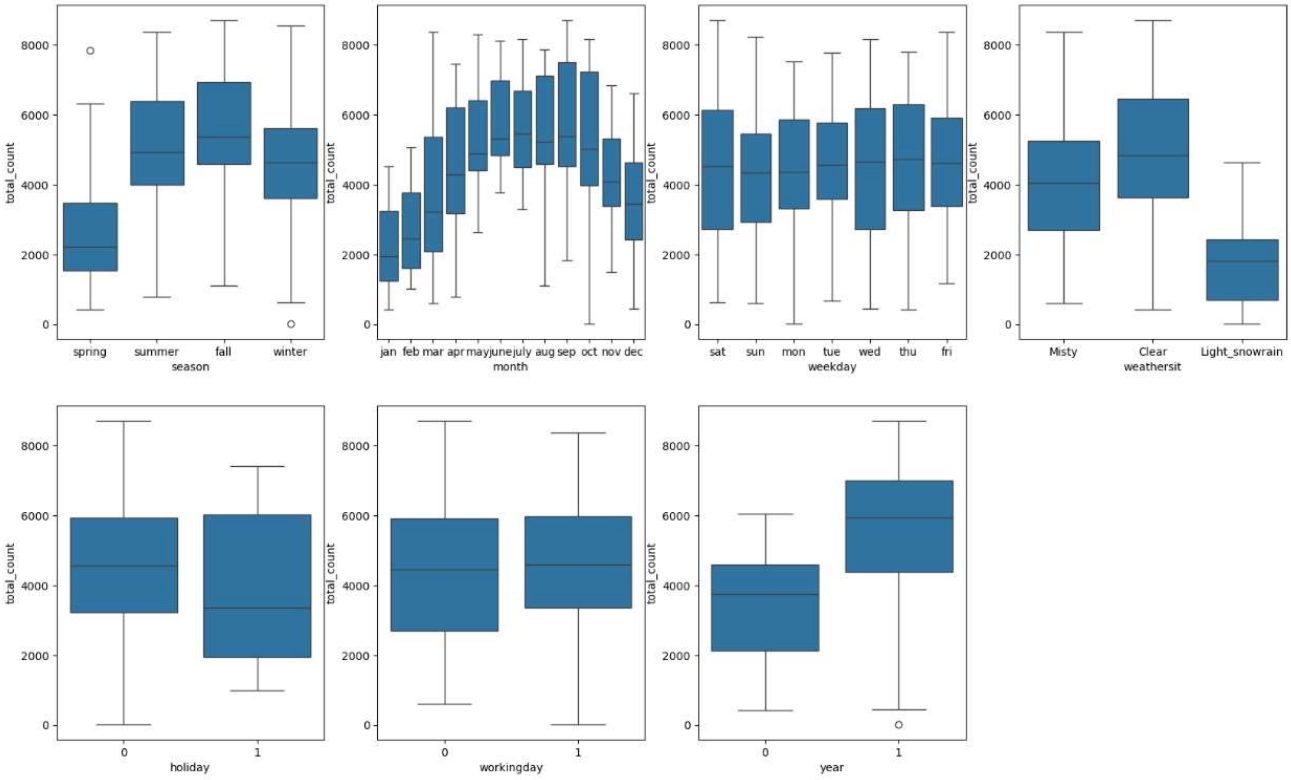
Assignment-based Subjective Questions

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:

<Figure size 2000x1800 with 0 Axes>



Insights from the above box plots for categorical variables:

1. season: The majority of users prefer to rent bikes during the fall season (season_3). This is likely because fall, being a transitional season between summer and winter, offers moderate temperatures, making it ideal for biking.
2. month: The highest number of bike rentals occurs in August and September, making these the peak months for bike rentals compared to others.
3. weekday: On day 3 (Wednesday), bike rentals are typically higher than on other weekdays. This could be because Wednesday is the mid-week day, and many people may rent bikes to commute or enjoy a break from their routine.
4. weathersit: Bike rentals are more frequent when the weather is clear, with few clouds, or partly cloudy, suggesting that good weather conditions encourage more people to rent bikes.
5. holiday: Bike rentals are higher during holidays, as people take the opportunity to enjoy leisure time with family and friends.
6. workingday: Bike rentals are generally higher on non-working days compared to working days, as people have more free time to rent bikes.
7. year: More users rented bikes in 2019 compared to 2018. This is typical in the initial years of a company's growth, where the number of users tends to increase significantly over time.

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: Using drop_first=True in dummy variable creation removes a redundant column, avoiding multicollinearity among variables. This streamlines the model while preserving essential information, with the dropped category acting as the reference. It also enhances interpretability by expressing coefficients relative to the baseline category.

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)

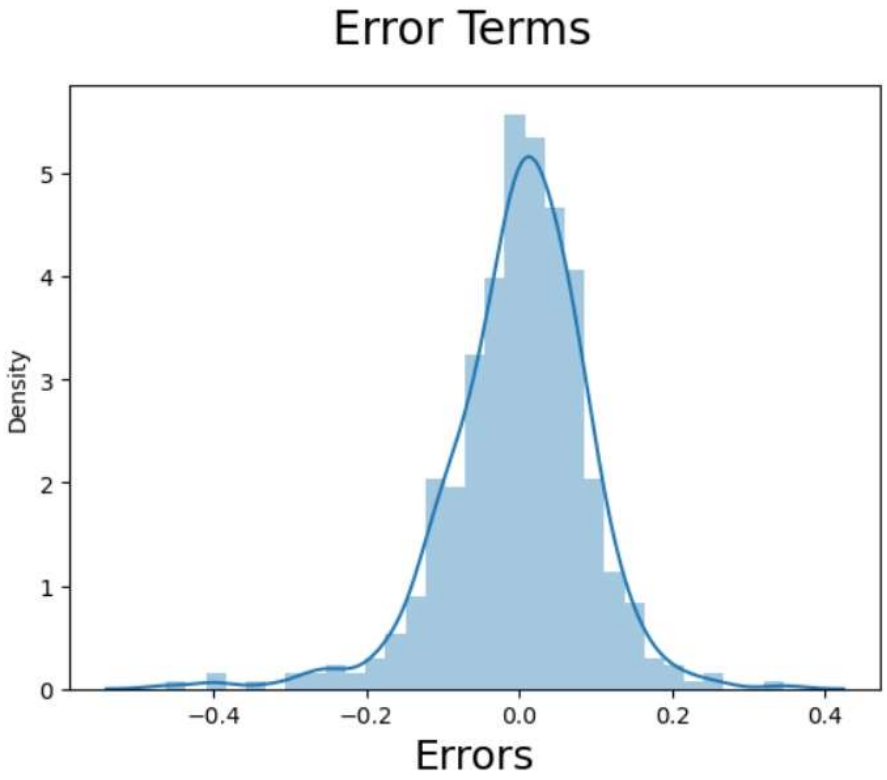
Answer: From the pair plot of numerical variables, we observe that 'temp' and 'atemp' have the strongest correlation with the target variable, total_count. Both variables exhibit similar patterns and display a linear relationship with total_count.

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: The validation of Linear Regression assumptions after building the model on the training set is as follows:

1. One key assumption of Linear Regression is that the error terms should follow a normal distribution with a mean of 0. To verify this, we perform residual analysis on the training set, where the residuals are the differences between the actual and predicted y_train values.

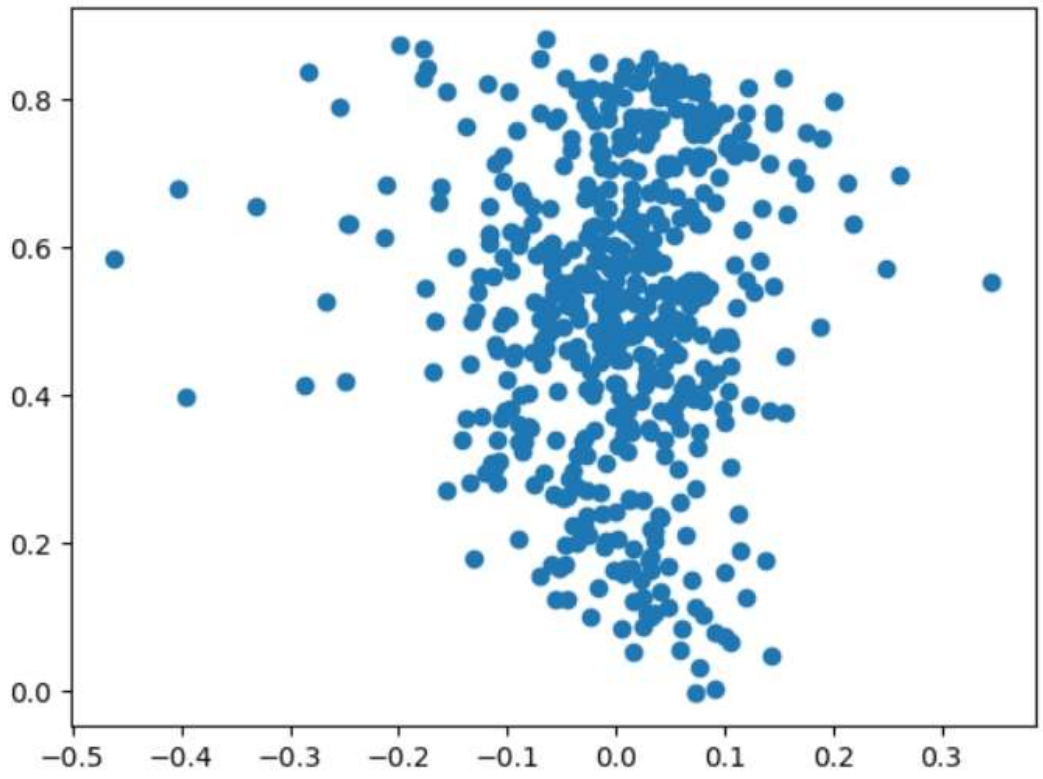


2. Another assumption is the absence of multicollinearity, meaning there should be little to no

correlation between independent features. This is checked by examining the Variance Inflation Factor (VIF) of the variables in the final model, with a VIF below 5 considered acceptable.

	Features	VIF
1	atemp	4.77
2	windspeed	4.12
4	season_winter	2.58
3	season_spring	2.36
0	year	2.06
11	month_nov	1.79
9	month_jan	1.65
7	weathersit_Misty	1.54
8	month_dec	1.46
10	month_july	1.35
12	month_sep	1.21
5	weekday_sun	1.17
6	weathersit_Light_snowrain	1.09

3. The third assumption is homoscedasticity, which requires that there should be no pattern when plotting residuals against fitted values. The plot below confirms the validity of the homoscedasticity assumption.



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: Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes

1. atemp
2. winter
3. sep

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Linear Regression is one of the simplest and most widely used algorithms in machine learning and statistics for predicting a continuous target variable based on one or more input features. It assumes a linear relationship between the independent variables (features) and the dependent variable (target).

Types of Linear Regression
There are two main types of linear regression:

1. **Simple Linear Regression**

This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable. The equation for simple linear regression is:
$y=\beta0+\beta1X$*y=β0+β1X*
where:

- Y is the dependent variable
- X is the independent variable
- β0 is the intercept
- β1 is the slope

2. **Multiple Linear Regression**

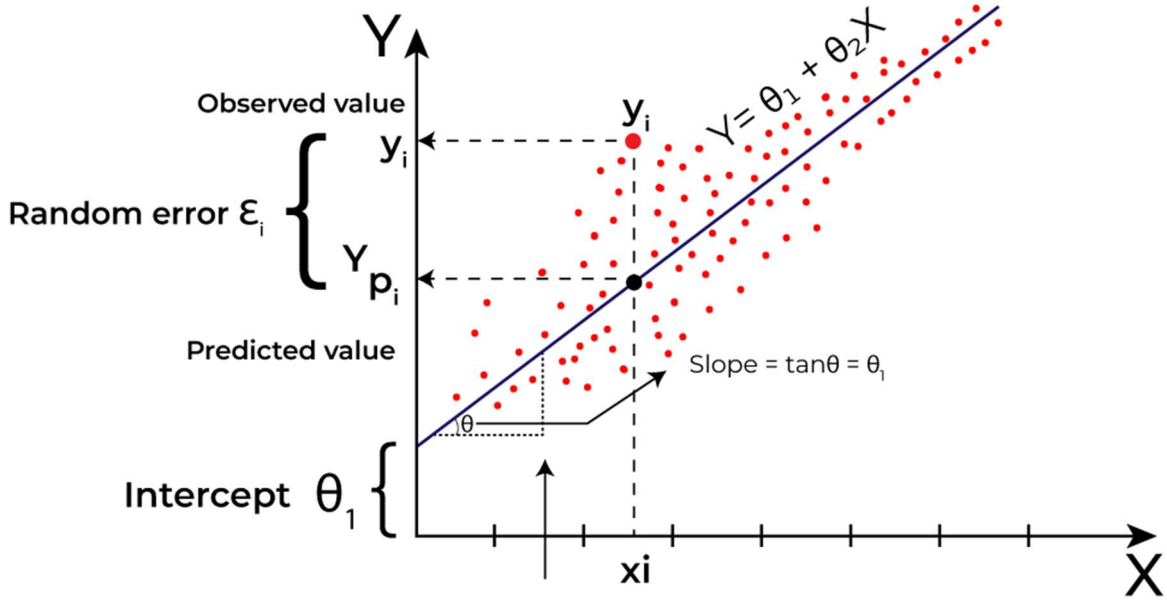
This involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:
$y=\beta0+\beta1X1+\beta2X2+\ldots\ldots\ldots\beta nXn$
where:

- Y is the dependent variable
- X1, X2, …, Xn are the independent variables
- β0 is the intercept
- β1, β2, …, βn are the slopes

The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables.

Our primary objective while using linear regression is to locate the best-fit line, which implies that the error between the predicted and actual values should be kept to a minimum. There will be the

least error in the best-fit line.
The best Fit Line equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variables.



Assumptions of Linear Regression:
Linear Regression relies on several key assumptions:

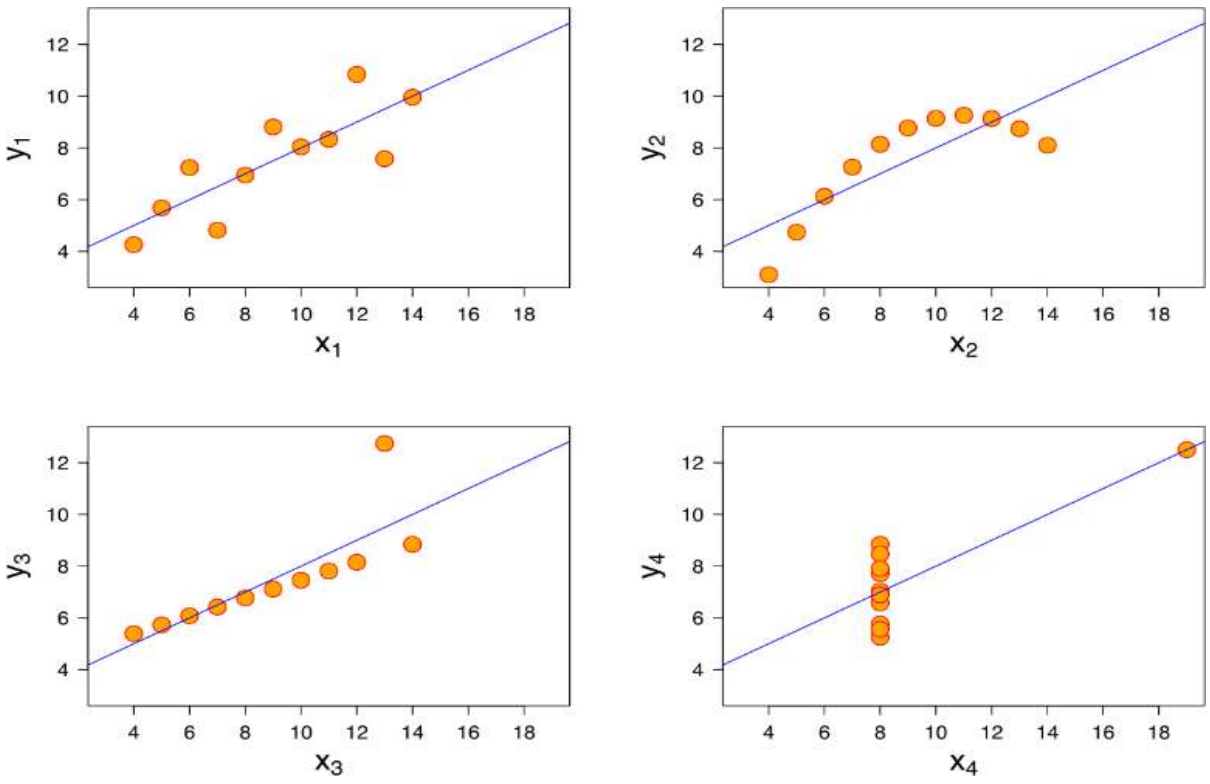
1. **Linearity**: The relationship between the dependent and independent variables is linear.
2. **Independence**: The residuals (errors) are independent of each other.
3. **Homoscedasticity**: The variance of the residuals is constant across all values of the independent variables.
4. **Normality of Residuals**: The residuals should be normally distributed.
5. **No Multicollinearity**: The independent variables should not be highly correlated with each other.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Anscombe's Quartet, created by statistician Francis Anscombe, consists of four datasets, each containing eleven (x,y)(x, y)(x,y) pairs. Remarkably, these datasets share identical descriptive statistics. However, when visualized, the graphs reveal drastically different patterns, each telling a unique story despite their identical summary statistics.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

- The mean of xxx is 9, and the mean of yyy is 7.50 in each dataset.
- The variance of xxx is 11, and the variance of yyy is 4.13 for all datasets.
- The correlation coefficient between xxx and yyy, indicating the strength of their relationship, is consistently 0.816.

When these datasets are plotted on an xxx-yyy coordinate plane, they show the same regression lines. However, each dataset conveys a completely different story, highlighting the importance of visualization in data analysis.



1. **Dataset I** exhibits a clean and well-fitting linear relationship.

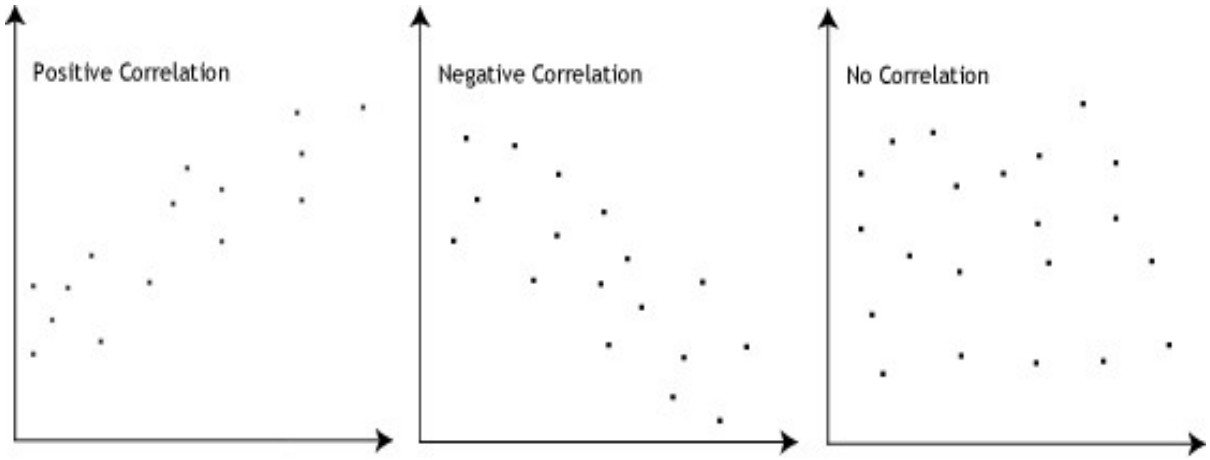
2. **Dataset II** deviates from a normal distribution.
3. **Dataset III** has a linear distribution, but an outlier significantly skews the regression line.
4. **Dataset IV** demonstrates how a single outlier can result in a high correlation coefficient.

This quartet underscores the crucial role of visualization in data analysis. Visualizing the data unveils its structure and provides a clearer understanding of the dataset.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Pearson's rrr is a numerical measure that quantifies the strength of the linear relationship between two variables. A positive correlation coefficient indicates that the variables increase or decrease together. Conversely, a negative correlation coefficient suggests that one variable tends to increase while the other decreases.
The Pearson correlation coefficient rrr ranges from -1 to +1:

- r=0r = 0r=0: No linear association between the variables.
- r>0r > 0r>0: Positive association, where an increase in one variable corresponds to an increase in the other.
- r<0r < 0r<0: Negative association, where an increase in one variable corresponds to a decrease in the other.

This relationship is visually represented in the diagram below.



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:
What is Scaling?
Scaling is the process of transforming data so that it fits within a specific range or distribution. It ensures that all features contribute equally to the model, preventing features with larger numerical ranges from dominating those with smaller ranges.

Why is Scaling Performed?

1. **Ensures Model Stability**: Many machine learning algorithms (e.g., gradient descent, SVMs, KNNs) are sensitive to the magnitude of features.
2. **Improves Convergence Speed**: Scaling speeds up optimization in algorithms like logistic

regression or neural networks.

3. **Balances Feature Importance**: Prevents features with larger ranges from overshadowing smaller-ranged features.
4. **Enables Better Performance**: Ensures algorithms like PCA and clustering that rely on distance metrics work effectively.

Aspect	Normalization	Standardization
Definition	Rescales data to a range (e.g., [0, 1]).	Centers data with mean 0 and standard deviation 1.
Formula	$X' = (X_{max} - X_{min}) / (X - X_{min})$	$X' = (X - \mu) / \sigma$
Sensitivity to Outliers	Highly sensitive to outliers.	Less sensitive to outliers.
Range	Fixed (e.g., [0, 1]).	No fixed range, but mean is 0 and standard deviation is 1.
Use Cases	Algorithms sensitive to scale without distribution assumptions (e.g., KNN, neural networks).	Algorithms assuming normality or using distance metrics (e.g., PCA, SVM).

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: The value of the **Variance Inflation Factor (VIF)** can become infinite when there is **perfect multicollinearity** among the independent variables in a dataset. Perfect multicollinearity occurs when one independent variable is an exact linear combination of one or more other independent variables.

If there is perfect correlation between variables, the VIF becomes infinite. A high VIF indicates multicollinearity among variables, with larger values reflecting greater inflation of the variance of model coefficients. For instance, a VIF of 4 means the variance of a coefficient is inflated by a factor of 4 due to multicollinearity.

An infinite VIF occurs when two independent variables are perfectly correlated, resulting in $R2=1$. This leads to the formula $1 / 1-R2$ approaching infinity. To address this issue, the variable causing perfect multicollinearity should be removed from the dataset**.**

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: A Quantile-Quantile (Q-Q) Plot is a graphical tool used to compare the distribution of a dataset to a theoretical distribution (typically a normal distribution). It plots the quantiles of the dataset against the quantiles of the theoretical distribution.

- If the data follows the theoretical distribution closely, the points on the Q-Q plot will align approximately along a straight 45-degree diagonal line.
- Deviations from this line indicate departures from the assumed distribution.

Use of Q-Q Plot in Linear Regression

In linear regression, one of the key assumptions is that the residuals (differences between actual and predicted values) follow a normal distribution. The Q-Q plot is used to **check the normality assumption** of these residuals.

1. **Assessing Normality**:
    - If the residuals are normally distributed, the points in the Q-Q plot will lie on or near the diagonal line.
    - Significant deviations indicate a violation of the normality assumption.
2. **Identifying Skewness**:
    - A curve away from the diagonal line in the Q-Q plot suggests skewness in the data.
        - Points above the line in the left tail indicate left-skewness.
        - Points above the line in the right tail indicate right-skewness.
3. **Detecting Outliers**:
    - Points that deviate significantly from the line, particularly at the tails, indicate the presence of outliers.

Interpreting the Q-Q Plot

- **Straight Line**: The residuals are approximately normally distributed.
- **S-Shaped Curve**: Indicates skewness in the data.
- **Upward or Downward Curves at the Ends**: Suggests heavy tails (kurtosis) or the presence of outliers.
