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)

- **Yr (Year):** There were more bookings in 2019 compared to 2018, showing positive business growth.
- **Season**: The fall season saw a notable rise in bookings, and overall, bookings increased across all seasons from 2018 to 2019.
- **Weathersit**: Clear weather conditions (labeled as "Good") significantly boosted bookings.
- **Mnth (Month)**: Most bookings happened in May, June, July, August, September, and October. The trend increased from the start of the year until mid-year, then decreased towards the end.
- **Holidays**: Booking counts are lower on non-holidays, likely because people spend time with family during holidays.
- **Weekdays**: More bookings occurred on Thursday, Friday, Saturday, and Sunday compared to the start of the week.

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 creation of dummy variables is a technique utilized in statistical modeling and machine learning that employs binary values (0 or 1) to symbolize categorical variables. This procedure involves generating new binary (dummy) variables for each category contained within the original categorical variable, thereby indicating the presence or absence of a specific category.

The number of dummy variables we need to create depends on how many categories the original variable has. If the variable has *n* categories, we usually make *n -1* dummy variables.

It is important to use because:

1. Reduce Redundancy: By dropping the first category, we avoid a dummy variable for every category, which otherwise would provide redundant information.
2. Prevent Multicollinearity: It helps in preventing multicollinearity, by including one dummy variable for every category would make the sum of dummy variables equal to one. Which would lead to perfect multicollinearity.
3. Simple Interpretation: It makes the interpretation of the regression coefficient easier. The dropped category is taken as the baseline, and the coefficient of the other dummy variables show the deviation from the baseline.

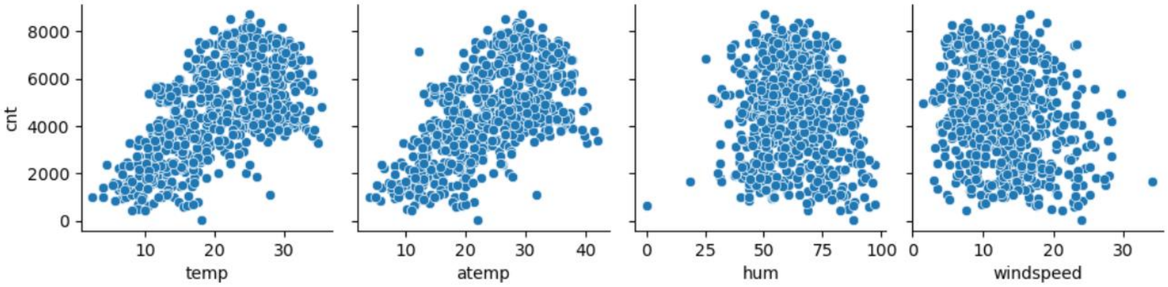
For example, if we have a categorical variable with three categories: A, B, and C, creating dummy variables with drop_first=True will result in two dummy variables (B and C), with A being the baseline.

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)

The variable 'temp' shows the highest correlation with the target variable cnt, as illustrated in the graph below. Since 'atemp' and 'temp' are redundant variables, only one is chosen when determining the best fit line.



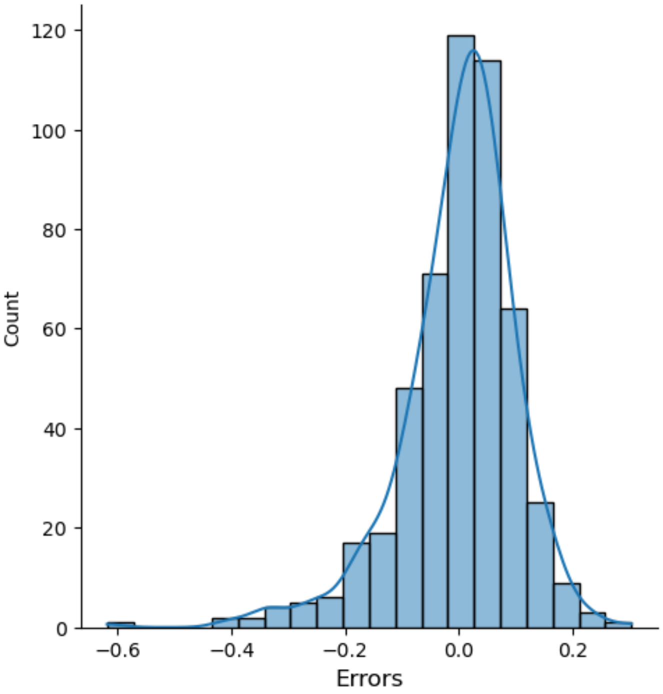
Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

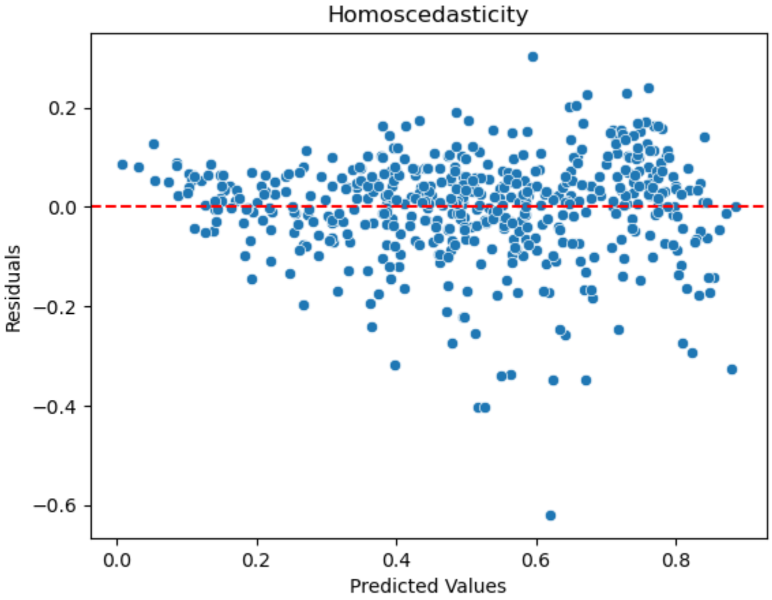
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

Validating the assumptions of Linear Regression is an important step to ensure the reliability of the model. The assumptions of Linear Regression were validated after building the model on the training set based on the following steps:

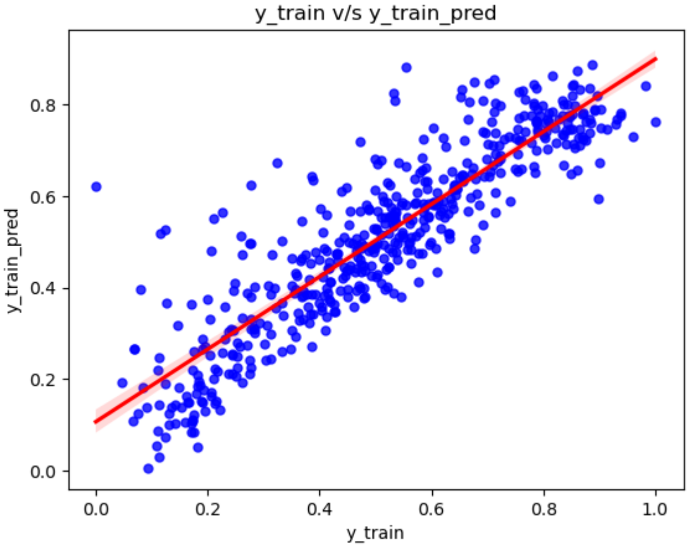
1. Residual Analysis: We have plotted a histogram to examine the residuals, which is the difference between the observed values and the predicted values. We have checked if the residuals are normally distributed and if there is any discernible patterns in the residual plot.



2. Homoscedasticity: We have plotted the residuals against the predicted values. We have checked if the spread of the residuals are roughly constant across the predicted values.



3. Linearity Relationship: We have plotted a scatterplot between the observed values and the predicted values from the model. We have checked if the points are falling along the diagonal line for indicating a linear relationship.



4. Multicollinearity: We have calculated the Variance Inflation Factor (VIF) for the predicted values to make sure that the VIF values are under the threshold (i.e. below 5) to ensure that there is no problematic multicollinearity.

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 towards explaining the demands of the shared bikes based on the final model can be determined from the equation of the best fit line.

The equation of the best fit line is:

cnt = 0.08 + 0.24 x yr + 0.05 x workingday + 0.55 x temp - 0.18 x windspeed + 0.09 X Summer + 0.12 x Winter + 0.06 X Saturday + 0.09 X September - 0.07 X Moderate

The top 3 features contributing towards explaining the demands of the shared bikes is:
- temp (Temperature)
- Winter (Winter Season)
- yr (Year)

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 understand the relationship between a dependent variable (the outcome we want to predict) and one or more independent variables (the predictors). The goal is to find the best-fitting line that represents this relationship.

The basic linear regression equation is represented as: $y = mx + b$

Where:

- "y" is the dependent variable,
- "x" is the independent variable,
- "m" is the slope of the line, and
- "b" is the y-intercept.

Types of Linear regression:

Single Linear regression: It is a basic form of linear regression that models the relationship between two variables: one dependent variable (y) and one independent variable (x). The relationship is modeled by using the equation:

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

where:

- y is the dependent variable,
- x is the independent variable,
- β0 is the y-intercept,
- β1 is the slope of the line, and
- ε is the error term.

Multi Linear regression: It is an extension of simple linear regression that models the relationship between a dependent variable and two or more independent variables. The relationship is modeled by using the equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$$

where:

- y is the dependent variable,
- x1, x2,.., xn are the independent variables,
- β0 is the y-intercept,

- β1, β2,…, βn is the coefficients, and
- ε is the error term.

How the Linear Regression works:

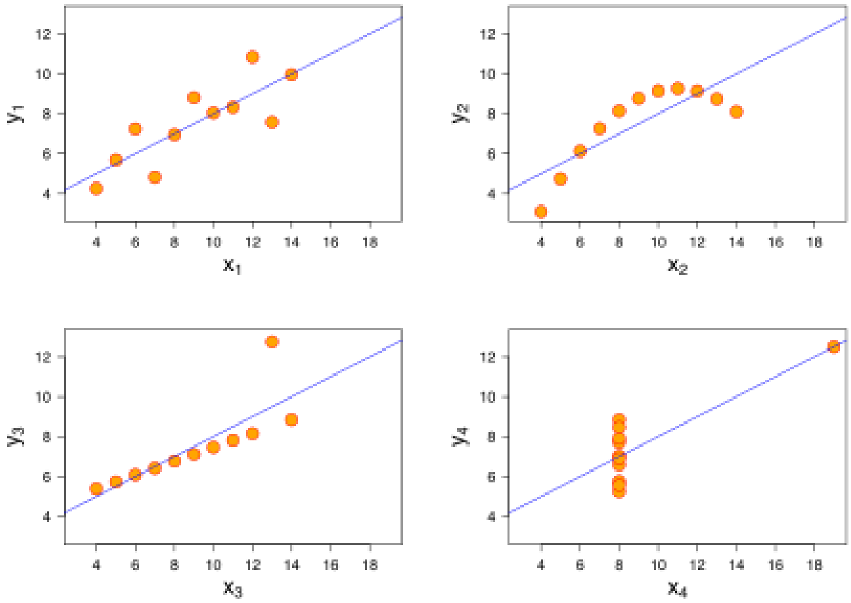
1. **Data Collection:** First, we need to gather data that has both the outcome we're attempting to predict (the dependent variable) and the variables we think affect it (the independent variables). We need to have a complete and relevant dataset because the quality of the data directly affects how well our model will perform.
2. **Data Processing:** Next, we prepare the data for analysis. This involves cleaning the data by handling any missing values and converting categorical variables into numerical values using techniques like one-hot encoding. We also standardize or normalize the independent variables if they have different scales to ensure they're on a similar level.
3. **Model fitting:** Now, we fit the linear regression model to our data. We do this by estimating the coefficients using the least squares method, which minimizes the sum of the squared differences between the observed values and the predicted values. Essentially, we're finding the best-fitting line that represents the relationship between our variables.
4. **Model Evaluation:** After fitting the model, we need to evaluate its performance. We use metrics like R-squared, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) to see how well our model predicts the dependent variable. A good model will have a high R-squared value and low MSE/RMSE values, indicating accurate predictions.
5. **Residual Analysis:** We then check the residuals, which are the differences between the observed and predicted values. We want to ensure these residuals are randomly distributed and don't show any patterns. This step helps us verify that the model assumptions hold true, such as linearity, independence, homoscedasticity, and normality of residuals.
6. **Prediction:** Finally, we use our fitted model to make predictions on new data. We apply the estimated coefficients to predict the dependent variable for new values of the independent variables. We can then compare these predictions with actual values (if available) to assess their accuracy.

Assumptions of Linear Regression

- **Linearity:** A linear relationship exists between independent and dependent variables.
- **Independence:** Data points are independent of each other.
- **Homoscedasticity:** The variance of the errors is constant across the range of the independent variable.
- **Normality:** The residuals (errors) follow a normal distribution.

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 a set of four datasets that have nearly identical simple descriptive statistics but differ significantly when graphed. This illustrates the importance of visualizing data and the limitations of relying solely on summary statistics. This quartet highlights the concept that datasets with similar statistical properties can exhibit diverse patterns when graphed. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties.



Graphical representa8on of Anscombe's quartet

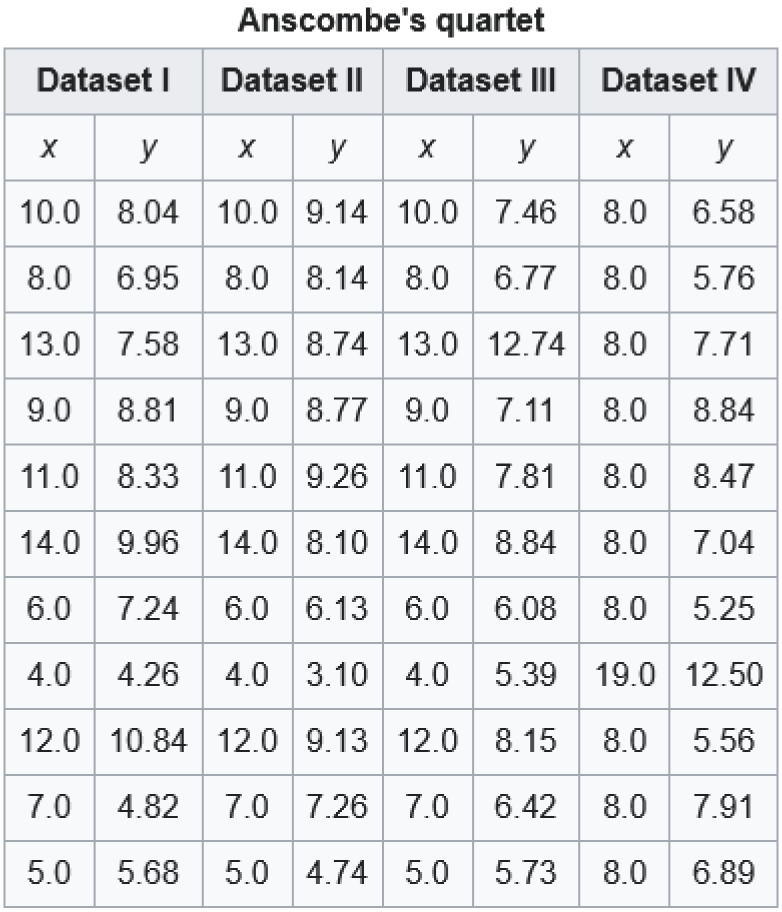
1. **First Scatter Plot (Top Left)**: Shows a simple linear relationship between two correlated variables, where y*y* can be modeled as Gaussian with a mean linearly dependent on x*x*.

2. **Second Scatter Plot (Top Right)**: Displays an obvious relationship between the variables, but it is not linear. The Pearson correlation coefficient is not relevant here; a more general regression and the corresponding coefficient of determination would be more appropriate.

3. **Third Scatter Plot (Bottom Left)**: Depicts a linear relationship, but the regression line should be different. A robust regression would be better, as the calculated regression is skewed by one outlier, reducing the correlation coefficient from 1 to 0.816.

4. **Fourth Scatter Plot (Bottom Right)**: Illustrates how a single high-leverage point can produce a high correlation coefficient, even though the other data points do not show any relationship between the variables.

The datasets are as follows. The x values are the same for the first three datasets.



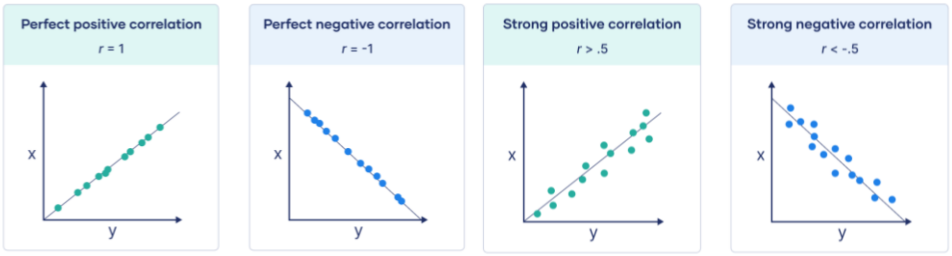
Anscombe's quartet

Dataset I		Dataset II		Dataset III		Dataset IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

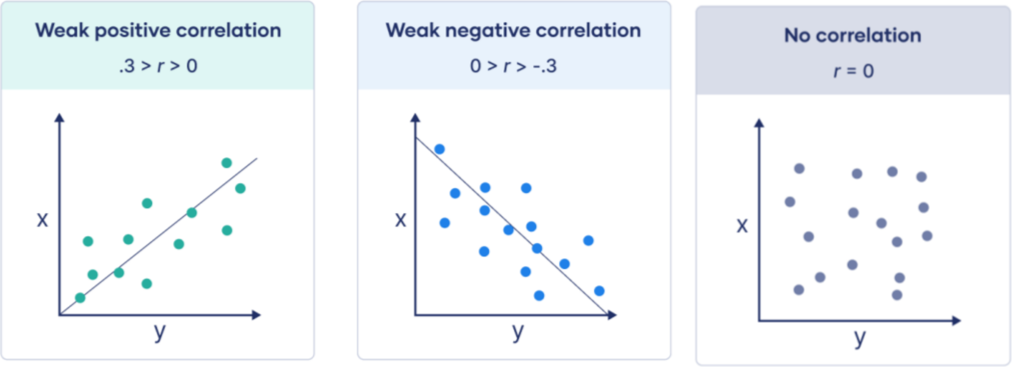
The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

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, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two variables. It quantifies the strength and direction of this relationship, providing a value between -1 and 1. The Pearson correlation coefficient also tells us whether the slope of the line of best fit is negative or positive. When the slope is negative, *r* is negative. When the slope is positive, *r* is positive.
When *r* is 1 or –1, all the points fall exactly on the line of best fit and r is greater than .5 or less than –.5, the points are close to the line of best fit:



When *r* is between 0 and .3 or between 0 and –.3, the points are far from the line of best fit and *r* is 0, a line of best fit is not helpful in describing the relationship between the variables:



Below is a formula for calculating the Pearson correlation coefficient (*r*):

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

where:

- *n* is the number of data points.
- *x* and *y* are the individual data points.

Pearson's correlation coefficient is commonly used in statistics to evaluate the strength and direction of the linear relationship between two variables. It's crucial to remember that correlation does not imply causation. Additionally, a correlation coefficient close to zero does not necessarily indicate the absence of a relationship; it only signifies the absence of a linear 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 data to fit within a specific range or distribution. This is particularly important in machine learning and statistical modeling, where different features may have different units or scales.

Scaling is performed to ensure that all features contribute equally to the model. Without scaling, features with larger ranges can dominate the learning process, leading to biased results. Scaling helps improve the performance and convergence speed of algorithms, especially those that rely on distance calculations, such as k-nearest neighbors (KNN) and support vector machines (SVM).
Difference between normalized scaling and standardized scaling:
Normalized Scaling(Min-Max Scaling):

- **Definition**: Normalization rescales the data to a fixed range, typically [0, 1] or [-1, 1].
- **Formula**:

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

where *x* is the original value, *min(x)* is the minimum value of the feature, and *max(x)* is the maximum value of the feature.

- Useful when we want to ensure that all features are on the same scale, especially when the data does not follow a Gaussian distribution.
- It uses minimum and maximum values of features for scaling
- It is sensitive to the outliers.

Standardized Scaling:

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

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

where x is the original value, *mean(x)* is the mean of the feature, and *sd(x)* is the standard deviation of the feature.

- Useful when the data follows a Gaussian distribution or when we want to ensure that the features have similar distributions.
- It uses mean and standard deviation for scaling.
- It is less sensitive to the outliers.

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)

The Variance Inflation Factor (VIF) is a measure used to assess multicollinearity in multiple regression analysis. It quantifies how much the variance of the estimated regression coefficients is increased due to multicollinearity.
The formula for VIF for a variable X_i is:

$$\mathrm{VIF}(X_i) = \frac{1}{1 - R_i^2}$$

where R_i^2 is the R^2 value obtained by regressing X_i against all other independent variables.
An infinite VIF value is a clear sign of perfect multicollinearity, which occurs when one or more independent variables in a regression model are exactly correlated with others. This means that certain variables contain redundant information, as one can be expressed as a precise linear combination of the others.

The issue arises during VIF computation, which requires matrix inversion. When perfect multicollinearity is present, the matrix becomes singular, making inversion impossible. As a result, the calculation fails, leading to an infinite VIF.

To resolve this, it's essential to detect and address multicollinearity in the dataset. This can be done by removing one of the highly correlated variables, merging them if they convey similar information, or applying dimensionality reduction techniques like PCA. Taking these steps not only eliminates the infinite VIF issue but also enhances the stability and clarity of the regression model.

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

How a Q-Q Plot Works

1. **Quantiles Calculation**: Quantiles are points taken at regular intervals from the cumulative distribution function (CDF) of a dataset. For a Q-Q plot, we calculate the quantiles of the sample data and the quantiles of the theoretical distribution we are comparing it to.
2. **Plotting the Quantiles**: The quantiles of the sample data are plotted on the y-axis, and the quantiles of the theoretical distribution are plotted on the x-axis. If the sample data follows the theoretical distribution, the points will lie approximately on a straight line.
3. **Line of Best Fit**: The 45-degree line (y = x) represents the ideal scenario where the sample distribution perfectly matches the theoretical distribution. Deviations from this line indicate deviations from the theoretical distribution.

Use and Importance in Linear Regression

1. **Normality Assumption**: One of the key assumptions in linear regression is that the residuals (errors) are normally distributed. This assumption is crucial because it affects the validity of hypothesis tests and confidence intervals. A Q-Q plot helps to visually check this assumption by comparing the distribution of residuals to a normal distribution.
2. **Identifying Deviations**: The Q-Q plot can reveal deviations from normality, such as skewness (asymmetry) or kurtosis (peakedness). If the points on the Q-Q plot deviate significantly from the line, it indicates that the residuals are not normally distributed. For example, if the points form an S-shaped curve, it suggests skewness in the data.
3. **Model Diagnostics**: By assessing the normality of residuals, you can diagnose potential issues with the model. If the residuals are not normally distributed, it may indicate the need for transformation of variables, inclusion of additional predictors, or addressing outliers. Ensuring that the residuals are normally distributed helps in making reliable inferences from the model.

Q-Q plot is a powerful diagnostic tool in linear regression that helps to assess the normality of residuals. By visually comparing the distribution of residuals to a normal distribution, it aids in identifying deviations from normality and diagnosing potential issues with the model. Ensuring that the residuals are normally distributed is crucial for making reliable inferences and improving the overall performance of the regression model.