BoomBikes Bike Sharing Assignment

Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 Marks)**
   Answer:

   - The year 2019 witnessed a higher number of bookings compared to the previous year, indicating positive progress in terms of business.

   - The fall season has experienced a notable increase in bookings. Additionally, across all seasons, there has been a substantial rise in booking counts from 2018 to 2019.

   - On non-holidays, the booking count tends to be lower, which is reasonable as people may prefer spending time at home with family during holidays.

   - It's evident that clear weather conditions (labelled as Good in the notebook) played a significant role in attracting more bookings.

   - There appears to be a relatively equal distribution of bookings between working days and non-working days.

   - Bookings were more prevalent on Thursday, Friday, Saturday, and Sunday compared to the early days of the week.

   - Majority of bookings occurred in May, June, July, August, September, and October. The trend exhibited an increase from the beginning of the year until mid-year, followed by a decrease towards the year's end.

2. **Why is it important to use drop_first=True during dummy variable creation?**
   Answer: (2 Marks)

   Dummy variable creation is a technique used in statistical modelling and machine learning to represent categorical variables with binary values (0 or 1). It involves creating new binary (dummy) variables for each category of the original categorical variable. These dummy variables serve as indicators for the presence or absence of a specific category.

   The number of dummy variables to create depends on the number of categories in the original categorical variable. For a categorical variable with n categories, you typically create $n - 1$ dummy variables.

Here's the rationale:

1. $n - 1$ Dummy Variables Rule: Creating n dummy variables introduces multicollinearity because the information in one category can be predicted from the others. This can lead to problems in the estimation of coefficients in regression models. By creating $n - 1$ dummy variables, you avoid perfect multicollinearity, as the information about the omitted category is implicitly captured.

2. Avoiding Redundancy: The information about the omitted category is implicitly captured by the constant term in the model. Including all dummy variables would introduce redundancy.

3. Enhancing Interpretability: The coefficients of the dummy variables represent the change in the response variable compared to the omitted category. This makes the interpretation of the model more straightforward.

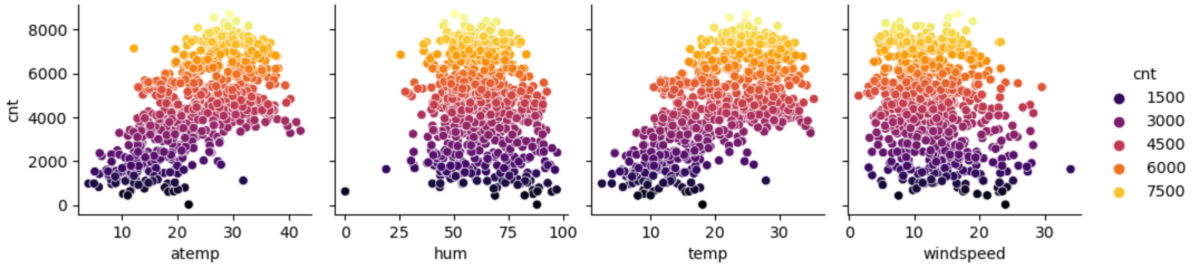
For example, if you have a variable "Color" with categories "Red," "Blue," and "Green," you would create two dummy variables, say "Is_Blue" and "Is_Green." The absence of both dummy variables implies that the color is "Red."

In Python, when using libraries like pandas, we can set `drop_first=True` while creating dummy variables to automatically drop one of the dummy variables to adhere to this $n - 1$ rule.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 Mark)**
Answer:

The variable 'temp' exhibits the strongest correlation with the target variable, as depicted in the graph below. Given that 'atemp' and 'temp' are redundant variables, only one of them is selected during the determination of the best fit line.

$$\begin{aligned} \text{cnt} = 4491.30 &+ 998.75 \times \text{yr} + 178.28 \times \text{workingday} + 1174.49 \times \text{temp} \\ &- 429.07 \times \text{hum} - 349.15 \times \text{windspeed} + 344.84 \times \text{Summer} \\ &+ 526.80 \times \text{Winter} + 234.70 \times \text{September} + 159.98 \times \text{Sunday} \end{aligned}$$



Bivariate Analysis of cnt with numerical features

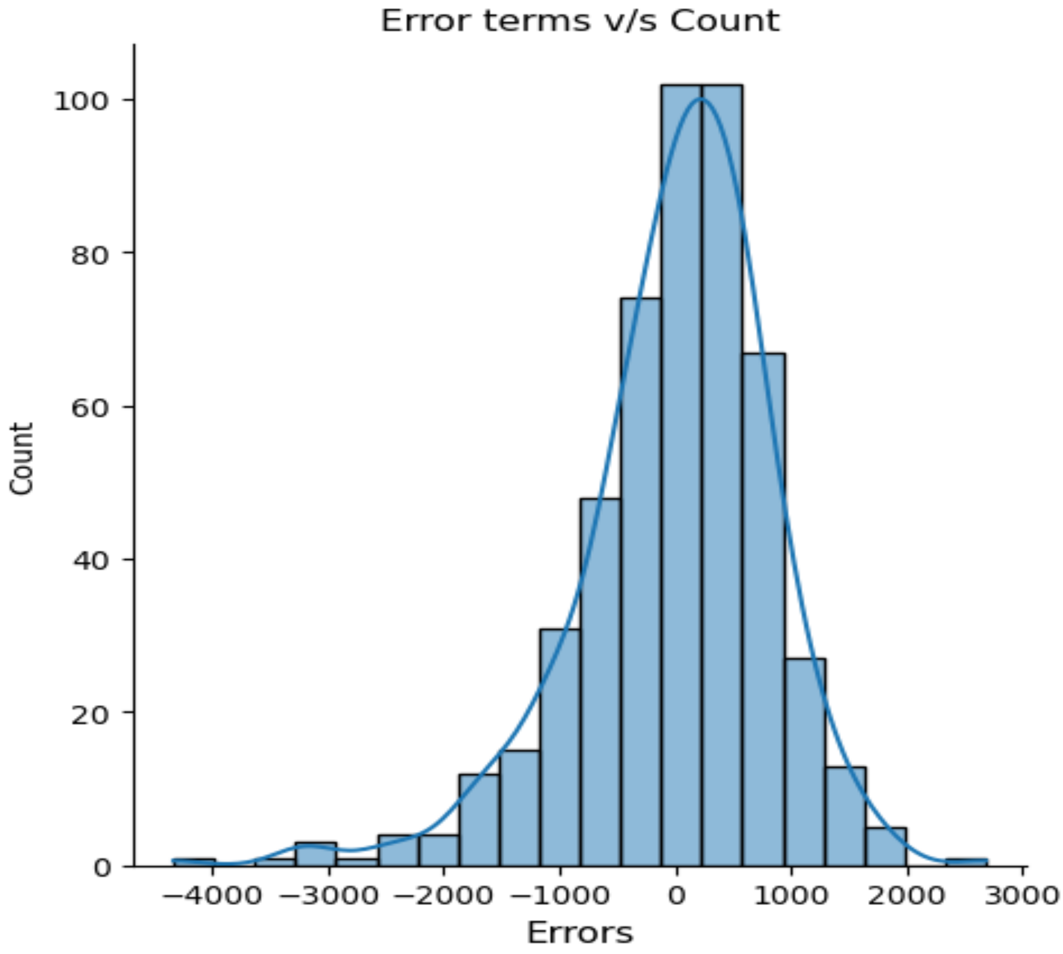
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 Marks)

Answer:

Validating the assumptions of linear regression is a crucial step to ensure the reliability of the model. After building the model on the training set, here are the steps I followed to validate the assumptions:

<u>1. Residual Analysis:</u>

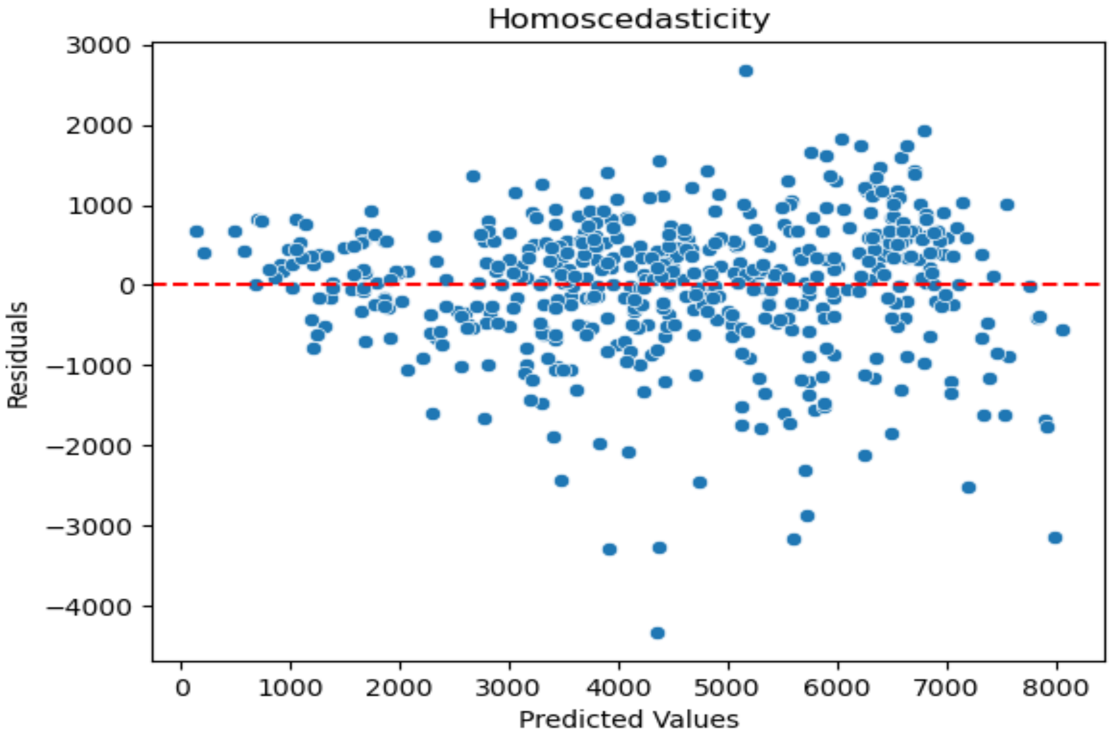
- Process: Examine the residuals (the differences between observed and predicted values).
- Check: Residuals should be approximately normally distributed, and there should be no discernible patterns in the residual plot.



Error counts with normal distribution curve

<u>2. Homoscedasticity (Constant Variance):</u>

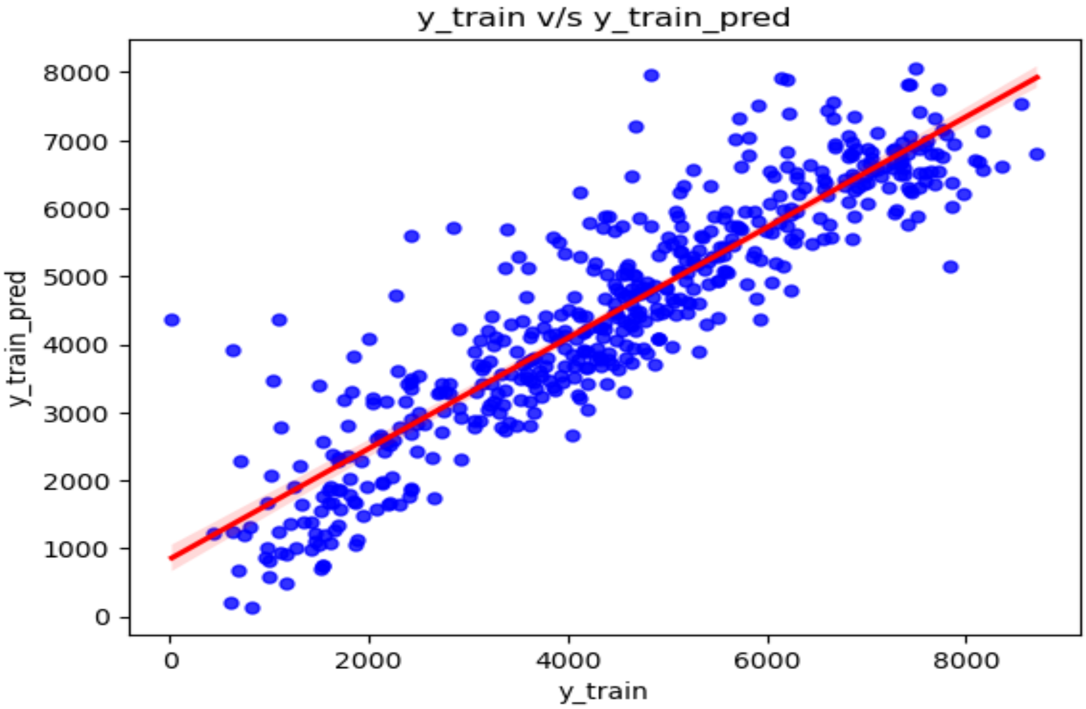
- Process: Plot residuals against predicted values.
- Check: The spread of residuals should be roughly constant across all levels of the predicted values.



Homoscedasticity

3. Linearity:

- Process: Create a scatterplot of observed vs. predicted values.
- Check: The points should fall approximately along a diagonal line, indicating a linear relationship.



Linearity of training set

4. Independence of Residuals:
- Process: Examine residuals for autocorrelation.
- Check: There should be no discernible pattern in the residuals when plotted against time or other relevant variables.

5. Multicollinearity:
- Process: Calculate Variance Inflation Factors (VIF) for predictor variables.
- Check: VIF values should be below a certain threshold (commonly 5 or 10) to ensure no problematic multicollinearity.

6. Cross-Validation:
- Process: Validate the model on a test set or through cross-validation.
- Check: Assess the model's performance on new data to ensure generalizability and consistency.

7. Check for Overfitting:
- Process: Evaluate model performance on a test set.
- Check: Ensure that the model generalizes well to new, unseen data without overfitting the training set.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 Marks)**
Answer:

From the equation of the best fit line:

$$\begin{aligned}\text{cnt} = 4491.30 + 998.75 \times \text{yr} + 178.28 \times \text{workingday} + 1174.49 \times \text{temp} \\ - 429.07 \times \text{hum} - 349.15 \times \text{windspeed} \\ + 344.84 \times \text{Summer} + 526.80 \times \text{Winter} \\ + 234.70 \times \text{September} + 159.98 \times \text{Sunday}\end{aligned}$$

The following three features significantly contribute to explaining the demand for shared bikes:

- Temperature (temp)
- Winter season (winter)
- Calendar year (year)

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 Marks)**
Answer:

Linear regression is a statistical method used for modelling the relationship between a dependent variable and one or more independent variables. It is widely used for predicting the value of the dependent variable based on the values of one or more independent variables. The basic idea is to find the best-fitting line (or hyperplane in the case of multiple independent variables) that minimizes the sum of the squared differences between the observed and predicted values of the dependent variable.

Linear regression algorithm follows following steps:

1. Model Representation:

Simple Linear Regression: In the case of a single independent variable, the model is represented as:

$$y = b_0 + b_1 \cdot x + \ \varepsilon$$

where:
- y is the dependent variable,
- x is the independent variable,
- b_0 is the y-intercept (constant term),
- b_1 is the slope of the line, and
- ε represents the error term.

Multiple Linear Regression: When there are multiple independent variables, the model is extended to:

$$y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \cdots + b_n \cdot x_n$$

where:
- $(x_1, x_2, \ldots, x_n)$ are the independent variables, and
- $(b_0, b_1, b_2, \ldots, b_n)$ are the coefficients.

2. Objective Function:

The goal is to find the values of $(b_0, b_1, b_2, \ldots, b_n)$ that minimize the sum of the squared differences between the observed and predicted values. This is often expressed as the sum of squared errors (SSE) or mean squared error (MSE):

$$\text{MSE} = \frac{1}{2m} \sum_{i=1}^{m} (y_i - \widehat{y_i})^2$$

where $(\ m\ )$ is the number of data points, (y_i) is the observed value, and $(\widehat{y_i})$ is the predicted value.

3. Minimization:

To find the optimal values of the coefficients, the algorithm uses optimization techniques such as gradient descent. The objective is to iteratively update the coefficients in the direction that minimizes the cost function.

4. Training the Model:

The model is trained on a dataset, where the algorithm learns the values of the coefficients that best fit the data. This involves feeding the algorithm input-output pairs and adjusting the coefficients until the model produces predictions close to the actual outcomes.

5. Prediction:

Once the model is trained, it can be used to make predictions on new, unseen data. The predicted values are obtained by plugging the new input values into the learned regression equation.

6. Evaluation:

The model's performance is assessed using metrics such as (R^2) (coefficient of determination), MSE, or other relevant metrics, depending on the context.

7. Assumptions:

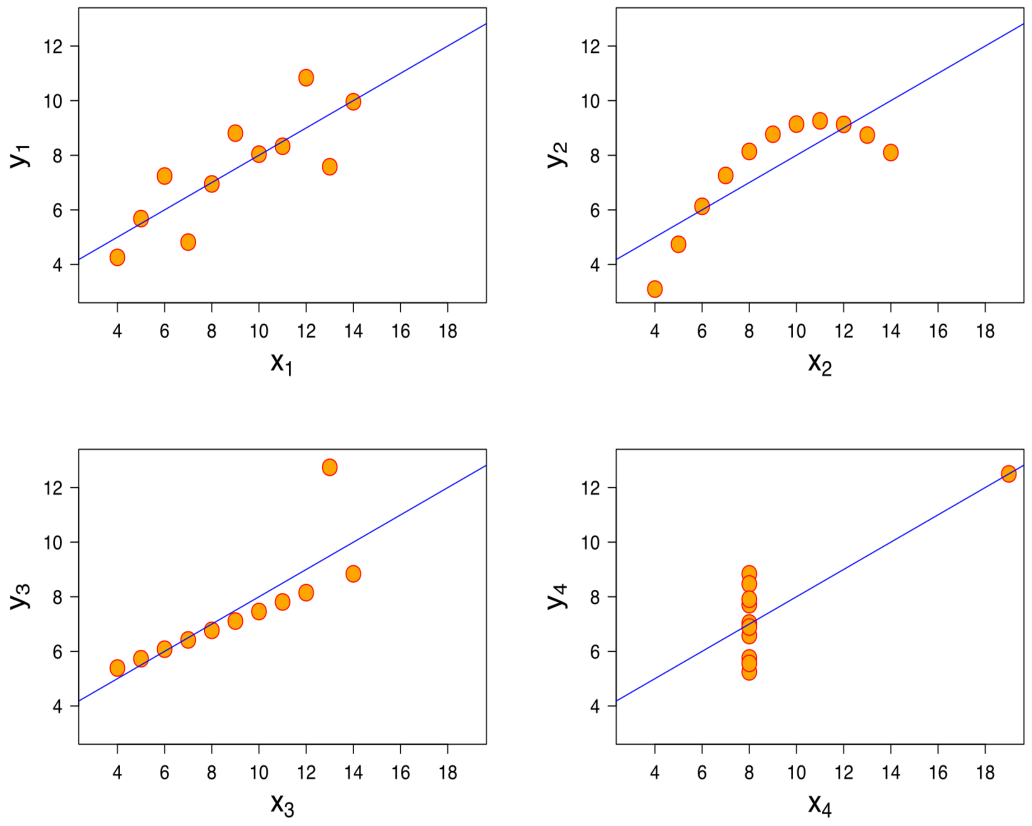
Linear regression relies on the assumption of a linear relationship between independent and dependent variables, normally distributed errors, constant error variance (homoscedasticity), and the absence of perfect multicollinearity, ensuring that there is no perfect linear relationship among the predictors.

Linear regression is a versatile and widely used algorithm, but it's important to check whether its assumptions hold in each dataset and consider more advanced techniques when those assumptions are violated.

2. Explain the Anscombe’s quartet in detail. (3 Marks)

Answer:

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 analysing it, and the effect of outliers and other influential observations on statistical properties.



Graphical representation of Anscombe's quartet

All four sets are identical when examined using simple summary statistics but vary considerably when graphed.

From the above diagram:

- The initial scatter plot (top left) suggests a straightforward linear relationship, depicting two correlated variables, where y could be characterized as Gaussian with a mean linearly dependent on x.

- In the second graph (top right), although a relationship between the variables is evident, it is not linear, rendering the Pearson correlation coefficient irrelevant. A more general regression and the corresponding coefficient of determination would be more suitable.

- Moving to the third graph (bottom left), the modelled relationship is linear, but a different regression line is warranted (considering a robust regression). The calculated regression is skewed by a single outlier, significantly reducing the correlation coefficient from 1 to 0.816.

- Lastly, the fourth graph (bottom right) exemplifies a scenario where a lone high-leverage point can yield a high correlation coefficient, even when the other data points fail to indicate any relationship between the variables.

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

I		II		III		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	12.5
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.

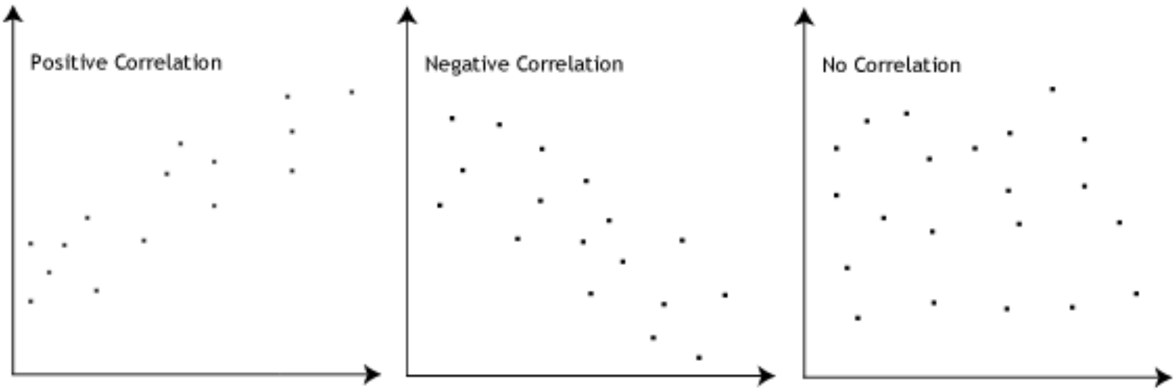
3. What is Pearson's R? (3 Marks)
Answer:

Pearson's correlation coefficient, often denoted as r, is a measure of the linear relationship between two variables. It quantifies the strength and direction of a linear association between two continuous variables. The coefficient takes values between -1 and 1, where:

- $r = 1$: Perfect positive linear correlation.
- $r = -1$: Perfect negative linear correlation.
- $r = 0$: No linear correlation.

A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

This is shown in the diagram below:



Correlation Graphs

The formula for Pearson's correlation coefficient between two variables, X and Y, with n data points, is given by:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Here:
- X_i and Y_i are the individual data points.
- $\bar{X}$ and $\bar{Y}$ are the means of X and Y.
- The numerator represents the covariance between X and Y.
- The denominator is the product of the standard deviations of X and Y.

Pearson's correlation coefficient is widely used in statistics to assess the strength and direction of the linear relationship between two variables. It's important to note that correlation does not imply causation, and a correlation coefficient close to zero does not necessarily mean the absence of a relationship; it only indicates the absence of a linear relationship.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 Marks)**
   Answer:

   Scaling in the context of data pre-processing refers to the process of transforming the values of variables to a specific range or distribution. The goal is to bring all variables to a similar scale, making them comparable and preventing one variable from dominating others.

Advantages of Scaling:

1. Equal Weightage: Scaling ensures that all variables contribute equally to the analysis, preventing variables with larger magnitudes from disproportionately influencing the results.

2. Convergence: Many machine learning algorithms, particularly those based on distances or gradients (e.g., k-nearest neighbours, support vector machines, gradient descent-based algorithms), perform better when features are on a similar scale. Scaling aids in faster convergence during the optimization process.
3. Interpretability: It improves the interpretability of coefficients in linear models, as the coefficients represent the change in the dependent variable for a one-unit change in the predictor variable.

Differences between Normalized Scaling and Standardized Scaling:

1. Normalized Scaling (Min Max Scaling):

- Range: Scales the values of a variable to a specific range, usually [0, 1].
- Formula:

$$X_{\text{normalized}} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

- Advantages: Useful when the distribution of the variable is unknown or not Gaussian.
- Disadvantages: Sensitive to outliers.

2. Standardized Scaling (Z-score normalization):

- Mean and Standard Deviation: Scales the values to have a mean of 0 and a standard deviation of 1.
- Formula:

$$X_{\text{standardized}} = \frac{X - \text{mean}\ (X)}{\text{std}(X)}$$

- Advantages: Less sensitive to outliers; preserves the shape of the distribution.
- Disadvantages: Assumes that the variable follows a Gaussian distribution.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 Marks)**
Answer:

The Variance Inflation Factor (VIF) is a measure used to assess multicollinearity in a 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:

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

where R_i^2 is the R^2 value obtained by regressing X_i against all other independent variables.

When the value of VIF is infinite, it usually indicates perfect multicollinearity. Perfect multicollinearity occurs when one or more independent variables in a regression model are perfectly correlated (linearly dependent) with other variables. In such cases:
1. There is redundant information - One variable can be expressed as a perfect linear combination of others.
2. Matrix Inversion Issues - In the computation of the VIF, there's an attempt to invert a matrix, and perfect multicollinearity leads to the matrix being singular (non-invertible).

When the matrix is singular, it means that one or more variables can be predicted exactly from the others, and as a result, the computation of the VIF becomes problematic, leading to an infinite VIF value.

To address this issue, it's crucial to identify and handle multicollinearity in the dataset. This can involve removing one of the perfectly correlated variables, combining them, or using dimensionality reduction techniques. Addressing multicollinearity not only resolves the infinite VIF problem but also improves the stability and interpretability of the regression model.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 Marks)**
Answer:

A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. It compares the quantiles of the observed data against the quantiles of the expected distribution. If the points in the Q-Q plot approximately fall along a straight line, it suggests that the data is well-modelled by the chosen theoretical distribution.

Use and Importance of Q-Q Plot in Linear Regression:

1. Normality Assessment:
- Use: In linear regression, it is often assumed that the residuals (the differences between observed and predicted values) are normally distributed. Q-Q plots are valuable for checking this assumption.
- Importance: If the residuals deviate significantly from normality, it can affect the reliability of statistical inferences made from the regression model.

2. Identifying Outliers:
- Use: Outliers in the residuals can be detected by examining points that deviate from the expected straight line in the Q-Q plot.

- Importance: Outliers can influence the estimation of model parameters and may indicate data points that are not well-captured by the regression model.

3. Model Fit Assessment:

- Use: Q-Q plots provide a visual assessment of how well the residuals conform to a normal distribution.
- Importance: A good model fit is crucial for accurate predictions, and departures from normality in residuals may suggest inadequacies in the regression model.

4. Validity of Statistical Tests:

- Use: When conducting hypothesis tests or constructing confidence intervals, the assumption of normality in residuals is important.
- Importance: Violations of this assumption can lead to inaccurate p-values and confidence intervals, affecting the validity of statistical inferences.

Interpretation of Q-Q Plots:

- If the points in the Q-Q plot closely follow a straight line, it suggests that the residuals are approximately normally distributed.
- Deviations from the straight line indicate departures from normality.

Q-Q plots are powerful diagnostic tools in linear regression for assessing the normality of residuals, identifying outliers, and ensuring the validity of statistical inferences. They provide a visual and intuitive way to check the assumptions underlying the regression model.