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)

Inferences from Categorical Variable Analysis:

The analysis of categorical variables provided valuable insights into their impact on bike demand. Key observations include:

- **Season:** Rentals are highest in summer and fall, indicating that these seasons positively affect bike rentals, while spring has the lowest rentals, showing a negative effect.
- **Year:** There is a significant increase in bike rentals in 2019 (yr = 1) compared to 2018 (yr = 0). This indicates a positive trend over time, possibly due to increased popularity of the bike-sharing service, improved infrastructure, or growing awareness among users.
- **Month:** Peak rentals occur during June, July, August, and September while lower rentals are observed in November, December, January, and February, highlighting a pattern influenced by the time of year.
- **Holiday:** Rentals are significantly lower on holidays compared to non-holidays, indicating that holidays have a noticeable negative impact on bike rentals.
- **Weekday:** Rental counts vary across weekdays, with Thursday having the highest and Sunday the lowest, indicating that weekday patterns, likely driven by work schedules or recreational activities, influence bike rentals. Bookings were also more frequent on Wednesday through Saturday compared to earlier in the week.
- **Working Day:** Rentals on working days are notably higher than on non-working days, with working days contributing substantially more to bike rentals.
- **Weather:** Clear weather conditions played a significant role in driving bike rentals, with the highest number of rentals occurring during these conditions.

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)

It is important to use **drop_first=True** during dummy variable creation to avoid the dummy variable trap, which occurs when there is multicollinearity between the dummy variables. By dropping the first category, we prevent redundancy between the intercept and the dummy variables, ensuring that each feature contributes uniquely to the model.

For example, consider the season variable with four categories (1: Spring, 2: Summer, 3: Fall, 4: Winter). If we create dummy variables without drop_first=True, we would generate four dummy variables: Spring, Summer, Fall, and Winter. However, to avoid multicollinearity, we only need to create n-1 dummy variables for n categories (3 in this case). These variables would be Spring, Summer, and Winter. Dropping Fall ensures that the absence of all three variables implicitly represents Fall.

This approach ensures that the model's coefficients are interpretable and that there is no redundancy in the features, allowing for more accurate predictions.

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)

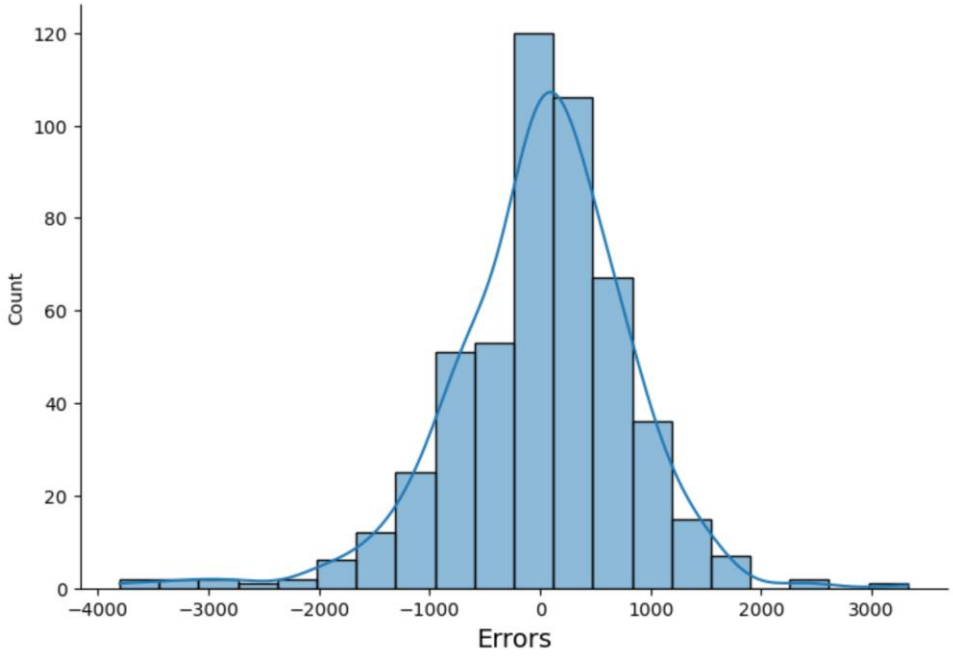
Based on the pair-plot, 'atemp' (feels like temperature) appears to have the highest correlation with the target variable 'cnt'. The scatter plot for 'atemp' vs. 'cnt' shows a clear positive trend, with a higher density of points towards the upper right corner, indicating a strong relationship.

While 'atemp' shows a strong correlation visually, the correlation coefficient with 'cnt' (0.6307) is very close to that of 'temp' (0.6270). Given the redundancy between 'atemp' and 'temp', 'temp' was chosen in the feature selection process.

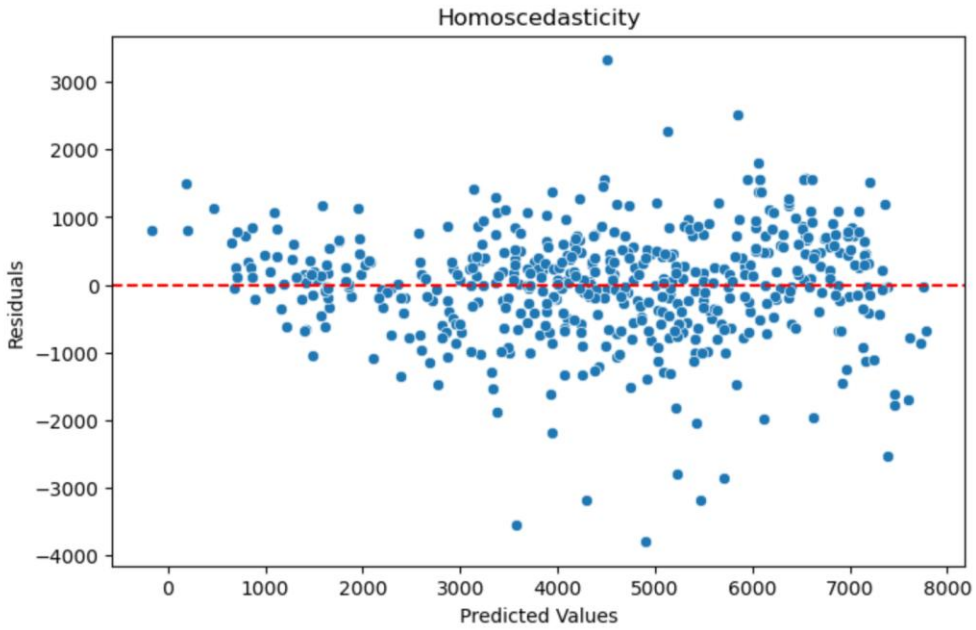
Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

Validating the assumptions of linear regression is a critical step to ensure the model's accuracy and reliability. After building the model on the training set, I followed these steps to validate the assumptions:

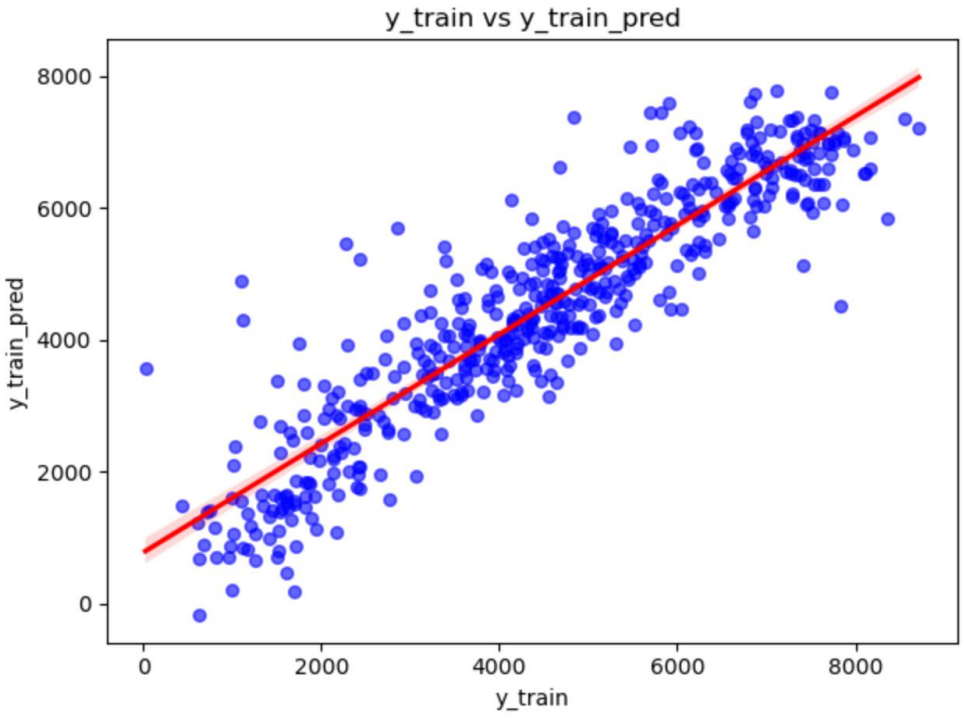
1. **Residual Analysis**:
    - **Process**: Examine the residuals, which represent the differences between the observed and predicted values.
    - **Check**: Ensure that the residuals are approximately normally distributed with mean zero, and with no discernible patterns in the residual plot. This confirms that the model is fitting the data well and the errors are random.



2. **Homoscedasticity (Constant Variance)**:
    - **Process**: Plot the residuals against the predicted values.
    - **Check**: The residuals should have a consistent spread across all levels of the predicted values, indicating that the variance of errors is constant and not influenced by the magnitude of predictions.



3. **Linearity**:
    - **Process**: Create a scatter plot of the observed vs. predicted values.
    - **Check**: The points should align closely along a diagonal line, suggesting a linear relationship between the independent variables and the dependent variable.



4. **Independence of Residuals**:
    - **Process**: Check for autocorrelation in the residuals.
    - **Check**: There should be no patterns or correlations when residuals are plotted against time or any other relevant variables, indicating that each residual is independent of the others.
5. **Multicollinearity**:
    - **Process**: Calculate the Variance Inflation Factor (VIF) for each predictor variable.
    - **Check**: VIF values should be below a commonly accepted threshold (typically 5), ensuring there is no problematic multicollinearity between the independent variables.
6. **Cross-Validation**:
    - **Process**: Perform cross-validation or validate the model using a separate test set.
    - **Check**: Evaluate the model's performance on unseen data to ensure it generalizes well and performs consistently across different datasets.
7. **Check for Overfitting**:
    - **Process**: Assess the model's performance on a test set that was not used during training.
    - **Check**: Verify that the model is not overfitting the training data by performing well on new data, indicating that it generalizes effectively without memorizing the training set.

By following these steps, I ensured that the assumptions of linear regression were met, making the model more reliable and robust for prediction.

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)

Based on the equation of the best-fit line:

cnt = 3477.19 + 2027.95 x yr + 481.86 x workingday + 923.13 x temp − 970.07 x spring + 491.99 x winter − 571.82 x July + 541.30 x September + 555.89 x Saturday − 703.49 x Cloudy − 2645.69 x Rainy

The following three features have the most significant impact on explaining the demand for shared bikes:

1. **Year** (2027.95): The "yr" coefficient of 2027.95 indicates a strong positive effect on bike demand.
2. **Temperature** (923.13): The "temp" coefficient of 923.13 suggests a positive correlation between temperature and bike demand, meaning higher temperatures lead to higher demand.

3. **Saturday** (555.89): The "Saturday" coefficient of 555.89 shows a substantial increase in bike demand on Saturdays.

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 supervised learning algorithm used for modelling the relationship between a dependent variable (target) and one or more independent variables (predictors). It is one of the simplest yet powerful methods for regression tasks. It is widely applied for **predicting the value** of the dependent variable based on the given independent variables. The main objective of linear regression is to find the **best-fitting line or hyperplane** (in the case of multiple independent variables) that minimizes the differences between observed and predicted values of the dependent variable.

Steps in the Linear Regression Algorithm

1. Model Representation

- **Simple Linear Regression**: In the case of one independent variable, the model is represented as:

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

  Where:

  - y: Dependent variable (response).
  - x: Independent variable (predictor).
  - β_0: Y-intercept (constant term).
  - β_1: Coefficient representing the slope of the line.
  - ε: Error term (residual).

- **Multiple Linear Regression**: When there are multiple independent variables, the model extends to:

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

  Where:

  - $x_1, x_2, \ldots, x_n$: Independent variables.
  - $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$: Coefficients of the model.

2. Objective Function

The goal of linear regression is to find the values of $\beta_0, \beta_1, \ldots, \beta_n$ that minimize the **sum of squared differences** between the actual and predicted values. This is known as the **Mean Squared Error (MSE)**:

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

Where:

- n: Number of data points.
- y_i: Actual observed value.
- $\hat{y}_i$: Predicted value from the model.

3. Minimization (Optimization)

To minimize the MSE, the algorithm optimizes the coefficients (β) using techniques like:

- **Gradient Descent**:
  - Iteratively adjusts coefficients to minimize the cost function.
  - Update rule: $\beta_j = \beta_j - \alpha \frac{\partial MSE}{\partial \beta_j}$

    Where α is the learning rate.
- **Normal Equation**:
  - Directly calculates the coefficients using matrix algebra: $\beta = (X^T X)^{-1} X^T y$

    Where X is the matrix of input features and y is the target variable.

4. Training the Model

- The model is trained on a dataset of input-output pairs.
- The algorithm adjusts coefficients iteratively (or via closed-form solution) to minimize the error.

5. Prediction

- Once trained, the model predicts values for new input data using the regression equation:

  $\hat{y} = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_n \cdot x_n$

6. Evaluation

The model's performance is assessed using metrics such as:

- **R-squared (R^2)**: Proportion of variance in the dependent variable explained by the model.
- **Mean Squared Error (MSE)**: Measures the average squared difference between actual and predicted values.

7. Key Assumptions

Linear regression relies on the following assumptions:

1. **Linearity**: Relationship between independent and dependent variables is linear.
2. **Independence**: The error terms are independent of each other.
3. **Homoscedasticity**: Residuals have constant variance across all levels of predicted values.
4. **Normality**: Residuals are normally distributed with mean zero.
5. **No Perfect Multicollinearity**: Independent variables are not highly correlated.

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: Explanation

Anscombe's Quartet is a set of four datasets devised by statistician Francis Anscombe in 1973. It demonstrates the importance of visualizing data before performing statistical analysis. Although these datasets have nearly identical summary statistics, they reveal strikingly different relationships when graphed. This highlights how relying solely on numerical summaries can be misleading.

Key Features of Anscombe's Quartet

1. **Identical Summary Statistics**:
    - **Mean of x**: 9 (approximately the same for all datasets).
    - **Mean of y**: 7.5 (approximately the same for all datasets).
    - **Variance of x and y**: Similar across all datasets.
    - **Correlation (r)**: Approximately 0.816 in all cases.
    - **Regression line**: y=3+0.5x (same for all datasets).

Dataset 1		**Dataset 2**		**Dataset 3**		**Dataset 4**	
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

2. **Dramatically Different Distributions**:
    - Despite having the same numerical properties, the datasets show diverse patterns when plotted.

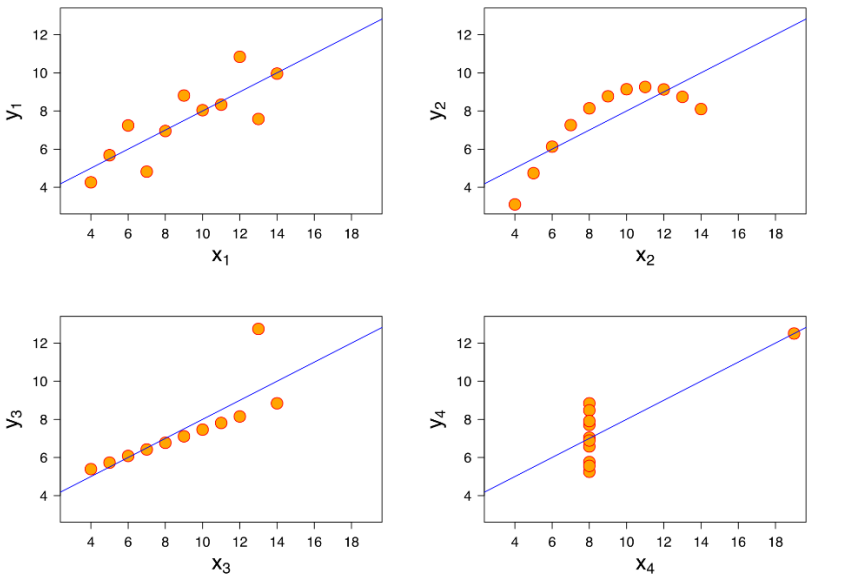


Fig: Graphical representation of Anscombe's quartet

Description of Each Dataset

1. **Dataset 1**:
   - Points are linearly distributed with some random noise.
   - Follows the regression line closely.
2. **Dataset 2**:
   - Points follow a nonlinear (curved) relationship.
   - Summary statistics are misleading as they suggest linearity.
3. **Dataset 3**:
   - Includes an outlier that influences the summary statistics and regression line significantly.
4. **Dataset 4**:
   - Almost all points are identical except for one influential outlier, which heavily skews the regression line.

Importance of Anscombe's Quartet

1. **Visual Analysis**:
   - Emphasizes the need to visualize data (e.g., scatterplots) to identify patterns or anomalies that summary statistics cannot reveal.
2. **Misleading Averages**:
   - Demonstrates how relying solely on averages, variances, or correlation coefficients can obscure the true nature of the data.
3. **Influence of Outliers**:
   - Highlights how outliers can disproportionately affect regression lines and statistical measures.

Conclusion

Anscombe's Quartet underscores the importance of combining statistical analysis with data visualization to fully understand the underlying patterns and relationships in a dataset. It remains a fundamental lesson in data analysis, teaching us not to rely on numbers alone.

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: Explanation

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. It is widely used in data analysis and research to determine how closely two variables are related.

Formula

The formula for Pearson's R is:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \cdot \sum(y_i - \bar{y})^2}}$$

Where:

- r: Pearson correlation coefficient.
- x_i: Values of the first variable.
- y_i: Values of the second variable.
- $\bar{x}$: Mean of the first variable.
- $\bar{y}$: Mean of the second variable.

Properties

1. **Range**:
    - r ranges from -1 to +1.
    - r = +1: Perfect positive linear relationship.
    - r = -1: Perfect negative linear relationship.
    - r = 0: No linear relationship.
2. **Sign**:
    - The sign of r indicates the direction of the relationship:
        - Positive r: As one variable increases, the other increases.
        - Negative r: As one variable increases, the other decreases.
3. **Magnitude**:
    - The closer r is to +1 or -1, the stronger the linear relationship.

Assumptions

1. **Linearity**: Assumes the relationship between the variables is linear.
2. **Continuous Data**: Variables should be continuous and normally distributed.
3. **Homogeneity of Variance**: Variability of one variable is constant across values of the other.

Use Cases

- Measuring relationships between variables in fields like economics, biology, and social sciences.
- Evaluating the effectiveness of predictive models.
- Identifying multicollinearity in regression analysis.

Limitations

1. **Only Measures Linear Relationships**: Cannot capture nonlinear relationships.
2. **Sensitive to Outliers**: Outliers can distort the value of r.
3. **Does Not Imply Causation**: A high r does not mean one variable causes change in the other.

Conclusion

Pearson's R is a versatile and widely used measure to determine the strength and direction of linear relationships between variables. However, it should be used alongside visualizations and domain knowledge to ensure accurate interpretation.

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)

1. Scaling in Linear Regression
Scaling is the process of adjusting the range of feature values in a dataset to a common scale without distorting the differences in the ranges of values. This transformation ensures that each feature contributes proportionately to the analysis.

2. Scaling is performed to:
- **Ensure Fair Contribution**: Prevent features with larger ranges from dominating those with smaller ranges, ensuring all features contribute equally to the model.
- **Improve Algorithm Performance**: Enhance the efficiency and convergence speed of algorithms like gradient descent used in linear regression.
- **Maintain Numerical Stability**: Reduce numerical issues in calculations, leading to more accurate and reliable model coefficients.

3. Difference Between Normalized Scaling and Standardized Scaling
- **Normalized Scaling (Min-Max Scaling)**:
    - **Definition**: Transforms features to a fixed range, typically [0, 1].
    - **Formula**: $X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$
    - **Use Case**: Suitable when the distribution of data is not Gaussian and when you need a bounded range.
    - **Example**: Transforming feature values like income, which may range from 10,000 to 100,000, to a range of [0, 1].

- **Standardized Scaling (Z-score Scaling)**:
    - **Definition**: Transforms features to have a mean of 0 and a standard deviation of 1.
    - **Formula**: $X_{scaled} = \frac{X-\mu}{\sigma}$

      Where μ is the mean and σ is the standard deviation.
    - **Use Case**: Ideal when the data follows a Gaussian distribution and when you want to center the data.
    - **Example**: Converting exam scores with a mean of 70 and a standard deviation of 10 into Z-scores, such as $\frac{85-70}{10}$ = 1.5.

Example:
Normalized Scaling:

- Original values: [1000, 20000]
- Scaled values: [0, 1]

Standardized Scaling:

- Original values: Mean = 50, Standard Deviation = 5
- Scaled values: $\frac{value-50}{5}$

By applying scaling, especially standardized scaling, the linear regression model can perform more effectively, ensuring that all features are on a comparable scale and contributing appropriately to the analysis or prediction task.

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)

Infinite VIF: Explanation and Causes

What is VIF?
Variance Inflation Factor (VIF) measures the extent of multicollinearity in regression analysis. It indicates how much the variance of a regression coefficient is inflated due to collinearity with other independent variables.

The formula for VIF is:

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

Where R_i^2 is the coefficient of determination when the i-th independent variable is regressed against all other independent variables.

Why Does VIF Become Infinite?

VIF becomes infinite when:

1. **Perfect Multicollinearity Exists**:

- If an independent variable is a perfect linear combination of one or more other independent variables, R_i^2 = 1.
- Substituting R_i^2 = 1 into the formula results in $VIF_i = \frac{1}{1-1} = \infty$

2. **Duplicate or Highly Correlated Features**:
   - This occurs when one feature is a duplicate or nearly identical to another. For instance:
     - Including a variable like "total sales" and another like "sales in thousands" (scaled versions of each other).
3. **Incorrect Data Preparation**:
   - Poor preprocessing, such as including categorical variables without proper one-hot encoding or adding highly correlated features like "price" and "discounted price."
   - **Scaling Dummy Variables**: Dummy variables represent categorical features and need not be scaled. Applying scaling to dummy variables can exaggerate relationships and result in perfect multicollinearity, causing VIF to become infinite.

Consequences of Infinite VIF:

- Infinite VIF indicates that the regression coefficients cannot be uniquely estimated due to redundancy among independent variables.
- It leads to unstable models where small changes in data result in large changes in coefficients.

Solution:

- Identify and remove one or more highly correlated variables.

Example:

If a dataset contains two variables, A and B, where B = 2×AB, their R_i^2 will be 1, causing VIF for both to be infinite. Dropping one variable resolves the issue.

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, often the normal distribution. It plots the quantiles of the observed data against the quantiles of the theoretical distribution to visually assess whether the data follows the expected distribution.

In the plot:

- If the points fall approximately along a straight diagonal line, the data conforms to the theoretical distribution.

- Deviations from the line indicate departures from the expected distribution.

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

In linear regression, the Q-Q plot is used to validate one of the critical assumptions: **normality of residuals**.

1. **Why Normality Matters**:
    - Linear regression assumes that the residuals (errors) are normally distributed. This is important for accurate hypothesis testing, reliable confidence intervals, and valid p-values.
2. **How Q-Q Plot Helps**:
    - **Identifying Normality**: By comparing residuals to a normal distribution, a Q-Q plot visually indicates whether the residuals are approximately normal.
    - **Detecting Outliers**: Points deviating significantly from the line in the Q-Q plot suggest outliers or heavy tails.
    - **Assessing Skewness or Kurtosis**: Systematic deviations like an S-shaped curve may indicate skewness, while deviations at the tails suggest kurtosis issues.
3. **Decision Making**:
    - If the Q-Q plot reveals significant deviations, transformations (e.g., log or square root) or alternative modeling approaches may be necessary to meet the assumptions of linear regression.

Why It's Important:
The Q-Q plot provides a simple yet effective way to diagnose potential issues with the normality assumption, ensuring the reliability and validity of linear regression results.
