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)

Following is the analysis:

- The box plots show that the median number of bike rentals is highest in fall and lowest in spring.
- The blox plot suggests that the clear weather influences people for renting bikes compared to mist or rain.
- Marginal difference in bike rental on Workingday vs Holiday/Weekend.
- Bike rentals appear to be same on each day of the week.
- July and Sep months seem to have the highest median rental counts.
- Ceratinly year 2019 shows a better trend of bike rentals 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)

Using drop_first=True when creating dummy variables helps prevent multicollinearity, making your regression model more stable and easier to interpret. In the given dataset we have used dummies for Season, month, weathersit.

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 positive correlation between bike rentals and temperature and feeling temperature. As the temperature increases, the number of bike rentals tends to increase.

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)

- Create scatter plots of the predicted values versus the actual values to visually assess linearity.
- Use residual plots (predicted values vs. residuals) to check if the residuals show a random scatter around zero.
- No multi collinearity : Check the Variance Inflation Factor (VIF) for each predictor. VIF values greater than 10 suggest significant multicollinearity.
- Use a correlation matrix or heat-map to visually assess correlations between independent variables.

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)

Windspeed,

temperature,

Season

Weather

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 fundamental algorithm used in statistics and machine learning to model the relationship between a dependent variable and one or more independent variables. Here's a detailed explanation of the linear regression algorithm:

1. Concept

Linear regression assumes that there is a linear relationship between the dependent variable (Y) and the independent variables (X). The goal is to find the best-fitting line (or hyperplane in higher dimensions) that minimizes the difference between the predicted values and the actual values of (Y).

2. Mathematical Representation

For a simple linear regression with one independent variable, the relationship can be expressed as:

Y = beta_0 + beta_1 X + epsilon

- (Y) is the dependent variable, (X) is the independent variable, (beta_0) is the y-intercept (constant term), (beta_1) is the slope of the line (coefficient for (X)), (epsilon) is the error term (residuals).

For multiple linear regression with multiple independent variables, the equation expands to:

Y = beta_0 + beta_1 X_1 + beta_2 X_2 + ... + beta_n X_n + epsilon

3. Objective

The primary objective of linear regression is to estimate the coefficients (beta) that minimize the residual sum of squares (RSS):

RSS = sum (Y_i - Ycap_i)^2

where (Y_i) is the actual value, and (Ycap_i) is the predicted value based on the linear model.

4. Finding Coefficients

The coefficients can be estimated using various methods, with the most common being:

- Ordinary Least Squares (OLS): This method finds the coefficients by minimizing the RSS. The solution can be derived using calculus (taking derivatives) or matrix algebra.

- Gradient Descent: An iterative optimization algorithm that adjusts the coefficients to minimize the RSS. This is particularly useful for large datasets.

5. Assumptions

Linear regression relies on several key assumptions:

- Linearity: The relationship between the dependent and independent variables is linear.

- Independence: Observations are independent of each other.

- Homoscedasticity: The residuals (errors) have constant variance across all levels of the independent variables.

- Normality: The residuals are normally distributed (especially important for inference).

6. Evaluation Metrics

To assess the performance of a linear regression model, several metrics can be used:

- R-squared ((R^2)): Measures the proportion of variance in the dependent variable that can be explained by the independent variables.

- Adjusted R-squared: Adjusts (R^2) for the number of predictors, useful for comparing models with different numbers of predictors.

- Mean Absolute Error (MAE): The average absolute difference between predicted and actual values.

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

7. Applications

Linear regression is widely used in various fields such as economics, biology, engineering, and social sciences for tasks like:

- Predicting sales based on advertising spend.

- Estimating real estate prices based on features of the property.

- Analyzing the relationship between temperature and bike rentals.

8. Limitations

While linear regression is powerful, it has limitations:

- It may not perform well if the relationship is non-linear.

- Sensitive to outliers, which can disproportionately affect the results.

- Assumes that the relationship between variables is additive.

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 collection of four datasets created by statistician Francis Anscombe in 1973. Each dataset contains 11 pairs of (x) and (y) values, and they all share the same statistical properties, including means, variances, and correlation. However, they differ significantly in their graphical representations. Here’s a closer look at each dataset and its implications:

1. The Datasets
- Dataset I: This dataset has a clear linear relationship. When graphed, the points form a straight line, indicating a strong correlation between (x) and (y).

- Dataset II: Like the first, this dataset appears mostly linear, but it contains one outlier. This outlier can heavily influence the results of any statistical analysis, making it important to identify such anomalies.

- Dataset III: This dataset showcases a non-linear relationship, specifically a parabolic curve. As (x) increases, (y) increases at an increasing rate, suggesting that a simple linear regression would not accurately capture the relationship.

- Dataset IV: In this dataset, most points cluster together vertically, with one outlier far from the group. This creates a misleading impression that could suggest a strong relationship, despite the reality that the majority of data points are constant.

2. Statistical Properties
Despite their differing appearances, all four datasets yield similar statistical summaries, such as:
- Mean of (x): 9
- Mean of (y): 7.5
- Variance of (x): 11
- Variance of (y): 4.125
- Correlation coefficient: approximately 0.816

This means that if you only looked at the statistics, you might assume they all represent similar relationships.

3. Key Takeaways
- Importance of Visualization: The quartet highlights the critical need to visualize data. Graphing the datasets reveals distinct patterns that are not apparent from summary statistics alone.

- Misleading Statistics: Similar statistical measures can mask significant differences in the data's behavior. This can lead to incorrect conclusions if analysis is based solely on numerical summaries.

- Model Selection: Different datasets require different analytical approaches. For example, datasets with non-linear relationships (like Dataset III) would need non-linear modeling techniques.

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)

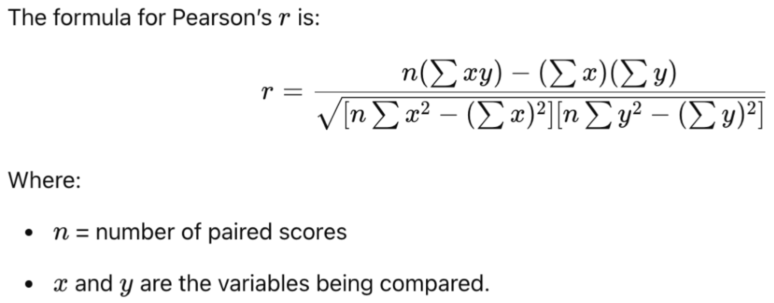
Pearson's (r), also known as Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. Here's a concise overview:

1. Definition

Pearson's (r) ranges from -1 to 1:

- - 1: Perfect positive correlation (as one variable increases, the other also increases).
- - -1: Perfect negative correlation (as one variable increases, the other decreases).
- - 0: No correlation (no linear relationship).

2. Calculation



The formula for Pearson's r is:

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

Where:

- n = number of paired scores
- x and y are the variables being compared.

3. Interpretation

- Strong correlation: Values closer to 1 or -1 indicate a strong relationship.

- Weak correlation: Values closer to 0 indicate a weak relationship.

4. Limitations

- Linearity: Pearson's (r) only measures linear relationships. Non-linear relationships may produce a low correlation coefficient.

- Sensitivity to Outliers: Outliers can greatly affect the value of (r), potentially skewing the results.

5. Usage

Pearson's (r) is widely used in statistics to assess relationships in various fields, such as psychology, finance, and biology, helping researchers understand how two variables may influence each other.

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)

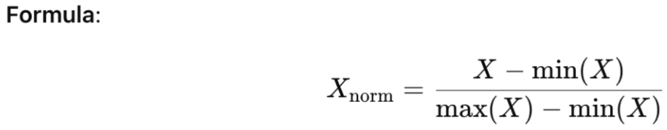
Answer: Please write your answer below this line. (Do not edit)

What is Scaling?
Scaling refers to the process of transforming the values of features in a dataset so that they fall within a specific range or have specific statistical properties. This is often necessary when working with machine learning algorithms, as features with different units or scales can adversely affect model performance.

Why is Scaling Performed?
1. Improves Model Performance: Many algorithms, particularly those based on distance metrics (like k-nearest neighbors or support vector machines), can be sensitive to the scale of the data.
2. Speeds Up Convergence: In optimization algorithms, such as gradient descent, scaling can lead to faster convergence.
3. Enhances Interpretability: Scaling can help in interpreting coefficients in regression models, especially when variables are on different scales.

Types of Scaling
The two common types of scaling are normalized scaling and standardized scaling:

1. Normalized Scaling (Min-Max Scaling)
- Definition: Normalization transforms the data to fit within a specified range, typically [0, 1].

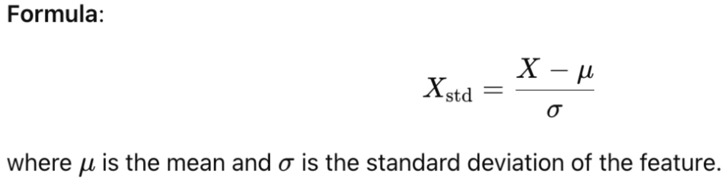


Formula:

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

- Use Case: It is useful when you need all features to contribute equally to the distance calculations and when the data is not normally distributed.

2. Standardized Scaling (Z-score Scaling)
- Definition: Standardization transforms the data so that it has a mean of 0 and a standard deviation of



Formula:

$$X_{\text{std}} = \frac{X - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation of the feature.

- Use Case: It is particularly useful when the data follows a normal distribution and is often required for algorithms that assume normally distributed data, such as linear regression.

Key Differences
- Range: Normalization rescales features to a fixed range [0, 1], while standardization centers the data around the mean with a unit standard deviation.

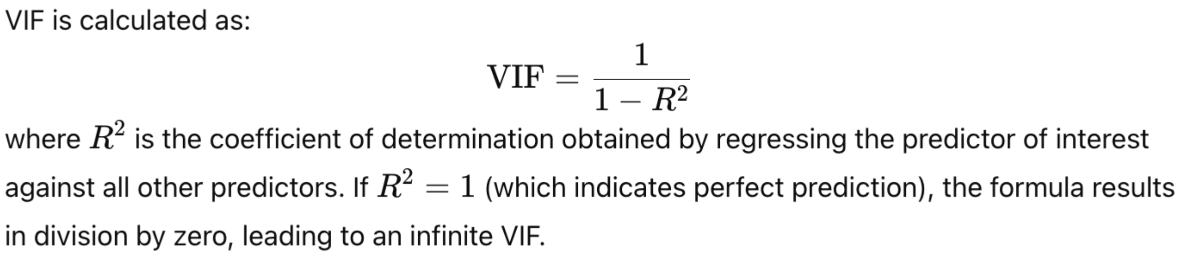
Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen?   (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

The Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in regression models. It quantifies how much the variance of a regression coefficient is inflated due to multicollinearity with other predictors. A VIF value greater than 10 is often considered problematic, but sometimes the VIF can be infinite. Here’s why that happens:

1. Perfect Multicollinearity
- Definition: Infinite VIF occurs when there is perfect multicollinearity between predictor variables. This means that one or more predictors can be expressed exactly as a linear combination of other predictors.
- Example: If you have two variables, (X_1) and (X_2), where (X_2) is simply (X_1) multiplied by a constant, the model cannot distinguish between the two variables, leading to an infinite VIF for one or both variables.

2. Redundant Features
- If a dataset includes duplicate or highly correlated features, it can lead to situations where the linear relationships between the variables are exact, resulting in infinite VIF values.

3. Mathematical Breakdown



VIF is calculated as:

$$\text{VIF} = \frac{1}{1 - R^2}$$

where R^2 is the coefficient of determination obtained by regressing the predictor of interest against all other predictors. If $R^2 = 1$ (which indicates perfect prediction), the formula results in division by zero, leading to an infinite VIF.

Conclusion
Infinite VIF indicates severe multicollinearity issues, which can complicate the interpretation of the regression coefficients and undermine the model's reliability. Addressing this typically involves removing or combining correlated variables to reduce redundancy.
- Sensitivity to Outliers: Normalization can be heavily affected by outliers, while standardization is more robust in this regard because it uses mean and standard deviation.

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, or Quantile-Quantile plot, is a graphical tool used to assess whether a dataset follows a particular distribution, most commonly the normal distribution. It compares the quantiles of the data against the quantiles of a theoretical distribution.

1. How a Q-Q Plot Works
- Axes: The x-axis represents the quantiles of the theoretical distribution (e.g., the standard normal distribution), while the y-axis represents the quantiles of the sample data.

- Interpretation: If the data follows the theoretical distribution, the points on the Q-Q plot will approximately lie along a straight line (y = x). Deviations from this line indicate departures from the expected distribution.

2. Uses of a Q-Q Plot

- Assessing Normality: One of the primary uses of a Q-Q plot is to check whether the residuals of a regression model are normally distributed. This is an important assumption for many statistical methods, including linear regression.
- Identifying Outliers: Points that lie far from the line can indicate outliers or unusual observations that may affect the model's performance.
- Comparing Distributions: Q-Q plots can be used to compare the distribution of two datasets to see if they come from the same distribution.

3. Importance of Q-Q Plots in Linear Regression

- Model Assumptions: In linear regression, one of the key assumptions is that the residuals (the differences between observed and predicted values) are normally distributed. A Q-Q plot helps validate this assumption.
- Robustness of Results: If the residuals are normally distributed, the statistical inferences made from the regression model (such as confidence intervals and hypothesis tests) are more reliable. If they are not, it may signal the need for model adjustments or transformations.
- Improving Model Fit: By identifying non-normal patterns in the residuals, practitioners can make informed decisions about potential transformations (like logarithmic or square root) or alternative modeling approaches (such as generalized linear models).