Assignment Based Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

A:

2. Why is it important to use drop_first=True during dummy variable creation?

A: Dropping One Dummy: By setting drop_first=True, you automatically drop one of the dummy variables, effectively breaking the perfect correlation and resolving multicollinearity. The remaining dummy variables still capture all the necessary information about the categorical variable. This makes your model more stable and interpretable, as it avoids issues caused by multicollinearity.

Key Benefits:

- Improved Model Performance: Prevents multicollinearity, leading to more reliable model estimates and predictions.
- Interpretability: Contributes to easier interpretation of model coefficients, as we are not dealing with redundant variables.
- Computational Efficiency: Reduces the number of variables, potentially improving model training and inference speed.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

A: “temp” variable i.e. temperature has the highest correlation with the target variable.

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

A:

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

A: Temperature, summer season & winter season are the top 3 features contributing significantly towards the demand of the shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

A: It is a statistical method for modeling the relationship between a dependent variable (y) and one or more independent variables (x). It assumes that this relationship is linear, meaning it can be represented by a straight line. It is commonly used for prediction and explanation of trends.

Types:

- Simple linear regression: Involves only one independent variable.
- Multiple linear regression: Involves multiple independent variables.

Working of this algorithm:

- Fitting a line: The algorithm finds the line that best fits the data points. It does this by minimizing the sum of the squared differences between the actual y values and the predicted y values (residuals). This is called the "least squares" method.
- Equation: The line is represented by the equation: y = b0 + b1x. b0 is the y-intercept (where the line crosses the y-axis). b1 is the slope of the line (how steep it is).
- Prediction: Once the line is fitted, it can be used to predict the value of y for any given value of x.

Example:

- Predicting house prices based on square footage:
- Independent variable (x): Square footage
- Dependent variable (y): House price

The model might find a relationship like: House price = 100,000 + 100 * Square footage

So, a house with 2000 square feet would be predicted to cost around $300,000.

2. Explain the Anscombe's quartet in detail.

A: Anscombe's quartet is a set of four datasets, crafted by statistician Francis Anscombe in 1973. While all four-share identical descriptive statistics (mean, variance, correlation, and even the linear regression line), they display radically different characteristics when visualized. This intentional design serves as a powerful reminder of the limitations of relying solely on numerical summaries and the importance of data visualization in statistical analysis.

The Datasets: Each dataset consists of 11 data points with an x-axis variable and a y-axis variable. The first three datasets share the same x-values, while the fourth uses a completely different set. The manipulation of the y-values creates the visual differences while maintaining identical statistics.

The Deceptions:

- Outlier Influence: One data point in the third dataset acts as a powerful outlier, skewing the relationship and suggesting a stronger connection than truly exists.
- Curved Relationship: While statistics imply a linear relationship, the fourth dataset showcases a clear parabolic curve, highlighting the danger of assuming linearity without visual confirmation.
- Irrelevant Variable: The fourth dataset demonstrates how even with identical statistics, including an irrelevant variable can create spurious correlations.

Consequences and Lessons:

- Statistical summaries can obscure crucial patterns and outliers. Plotting the data is essential for detecting these subtleties and understanding the true nature of the relationship.
- Not to blindly rely on assumptions of linearity or normality. Always visually explore the data before drawing conclusions.
- A single outlier can significantly impact statistics and distort interpretations. Be vigilant in identifying and handling outliers appropriately.

3. What is Pearson’s R?

A: Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the linear relationship between two continuous variables. It is a numerical value ranging from -1 to 1, where:

1: Indicates a perfect positive linear relationship. As one variable increases, the other increases proportionally.

0: Indicates no linear relationship. Changes in one variable do not affect the other.

-1: Indicates a perfect negative linear relationship. As one variable increases, the other decreases proportionally.

The closer the absolute value of R is to 1 (either positive or negative), the stronger the linear relationship between the variables. Values closer to 0 suggest a weaker or no linear relationship. However, it is important to remember that R only measures linear relationships. Non-linear relationships, like curves or parabolas, will not be captured by Pearson's R.

Applications:

Pearson's R has diverse applications across various fields:

- Finance: Analyzing correlations between stock prices, market trends, and economic indicators.
- Healthcare: Understanding relationships between risk factors and diseases, assessing treatment effectiveness.
- Education: Exploring connections between study habits and exam scores, identifying factors influencing student performance.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

A: It is a data preprocessing technique that involves transforming features (independent variables) in a dataset to a common scale. It aims to bring all features to a comparable range, typically between 0 and 1 or with a mean of 0 and a standard deviation of 1.

Scaling is performed to:

- Improve model performance: Many machine learning algorithms are sensitive to feature scales. Large-scale features can dominate model learning and diminish the impact of smaller-scale features. Scaling ensures all features contribute more equally.
- Enhance model convergence: Gradient descent-based algorithms converge faster when features are scaled. It prevents large-scale features from creating large gradients that slow down convergence.
- Increase distance-based algorithm accuracy: Algorithms like k-nearest neighbors and support vector machines rely on distance calculations. Scaling ensures distances are meaningful and not skewed by feature ranges.

Types of Scaling:

- Normalization (Min-Max Scaling): Shifts and rescales features to a range between 0 and 1.

  Formula: x_new = (x - x_min) / (x_max - x_min)

- Standardization (Z-Score Normalization): Shifts the mean of each feature to 0 and scales to a standard deviation of 1.

  Formula: x_new = (x - mean(x)) / std(x)

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

A: The VIF measures the degree of multicollinearity in a regression model, which essentially means how much information about one independent variable can be predicted from the information in other independent variables.

The VIF values can be infinite under certain circumstances:

- Perfect Multicollinearity: When there's perfect multicollinearity, meaning one independent variable is an exact linear combination of other independent variables, the VIF becomes infinite. This happens because Colinear variables have identical correlation coefficients with the dependent variable. The denominator of the VIF calculation involves subtracting these correlation coefficients from 1. When the coefficients are identical, they cancel each other out in the subtraction, leaving 0 in the denominator. Dividing by 0 results in an infinitely large VIF.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A: Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess whether a set of data follows a particular theoretical distribution, like a normal distribution. It is also helpful in comparing the distribution of one set of data to another.

How it works:

- Quantiles: Both datasets are divided into equal percentiles, representing different proportions of the data (e.g., 25%, 50%, 75%).
- Paired points: The quantiles from each dataset are then plotted against each other.
- Straight line: If the data follows the theoretical distribution, the points will roughly fall along a straight line. Deviations from the line indicate departures from the assumed distribution.

Use in Linear Regression:

- In linear regression, several assumptions need to be met for the model to be reliable. Q-Q plots help assess two crucial assumptions:
- Normality of residuals: The residuals (differences between predicted and actual values) should be normally distributed. If the points in the Q-Q plot deviate significantly from the straight line, especially in the tails, it suggests non-normality.
- Homoscedasticity: The variance of the residuals should be constant across all values of the independent variable. Non-constant variance is visualized by increasing or decreasing "spread" of the points around the line at different locations.

Importance of Q-Q Plots: Violations of these assumptions can lead to unreliable estimates of coefficients, confidence intervals, and p-values, affecting the overall validity of the model. Q-Q plots provide a visually intuitive way to detect these issues, prompting further investigation and potential remedies.