Assignment-based Subjective Questions

1. **Effect of Categorical Variables on Dependent Variable**

    1. **Season**: Different seasons show varying levels of bike rentals.
    2. **Year (yr)**: The year variable (0 for 2018 and 1 for 2019) indicates an increase in bike rentals over time, suggesting a growing popularity of bike-sharing services.
    3. **Weather Situation (weathersit)**: Clear weather conditions lead to higher bike rentals, while misty or rainy conditions result in lower rentals.
    4. **Month (mnth)**, **Weekday (weekday)**, **Holiday (holiday)**, and **Working Day (workingday)**: These variables can affect daily bike rentals.

2. **Importance of Using `drop_first=True` During Dummy Variable Creation**

    1. Using `drop_first=True` prevents multicollinearity by dropping one dummy variable from each category, thus avoiding the dummy variable trap where one category can be perfectly predicted from the others. This ensures the linear regression model remains stable and the coefficients are interpretable.

3. **Highest Correlation with Target Variable**

    1. Based on the pair-plot among the numerical variables, `temp` (temperature) has the highest positive correlation with the target variable `cnt` (total bike rentals). This indicates that as the temperature increases, the number of bike rentals also tends to increase.

4. **Validating Assumptions of Linear Regression**

    1. **Linearity**: Checked by plotting residuals versus fitted values. If the plot shows a random scatter, it indicates linearity.
    2. **Homoscedasticity**: Assessed by the same residuals vs. fitted values plot. A constant spread of residuals across all levels of fitted values indicates homoscedasticity.
    3. **Normality of Residuals**: Validated using a Q-Q plot. If the residuals lie along the 45-degree line, they follow a normal distribution.
    4. **Independence of Residuals**: Verified by checking for patterns in the residuals. Randomly scattered residuals indicate independence.

5. **Top 3 Features Contributing to Bike Demand**

    1. **Temperature (temp)**: Higher temperatures lead to higher bike rentals.
    2. **Year (yr)**: Indicates growth in demand over time.
    3. **Humidity (hum)**: Lower humidity levels tend to have a positive effect on bike rentals.

General Subjective Questions

1. **Linear Regression Algorithm**

Definition: Linear regression is a supervised learning algorithm that models the relationship between a dependent variable (target) and one or more independent variables (predictors) by fitting a linear equation to observed data.

Equation: The linear equation is Y=β0+β1X1+β2X2+...+βnXnY = \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_nX_nY=β0+β1X1+β2X2+...+βnXn, where β\betaβ coefficients are estimated from the data.

Objective: The goal is to minimize the sum of squared residuals (differences between observed and predicted values), known as the least squares method.

Assumptions:

Linearity: The relationship between predictors and the target is linear.

Independence: Observations are independent of each other.

Homoscedasticity: Constant variance of residuals.

Normality: Residuals are normally distributed.

Evaluation: The performance of the model is evaluated using metrics like R-squared, Mean Squared Error (MSE), and Adjusted R-squared.

2. **Anscombe’s Quartet**

Definition: Anscombe’s quartet consists of four datasets with nearly identical simple descriptive statistics but with different distributions and graphs.

Purpose: Demonstrates the importance of graphical analysis of data and not relying solely on summary statistics.

Examples:

The first dataset shows a linear relationship.

The second shows a nonlinear relationship.

The third has a distinct outlier.

The fourth shows a nearly vertical line with one distinct outlier.

3. **Pearson’s R**

Definition: Pearson’s correlation coefficient (R) measures the linear relationship between two variables.

Range: R ranges from -1 to +1.

R=1R = 1R=1 indicates a perfect positive linear relationship.

R=−1R = -1R=−1 indicates a perfect negative linear relationship.

R=0R = 0R=0 indicates no linear relationship.

Formula: R=Cov(X,Y)σXσYR = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}R=σXσYCov(X,Y), where Cov(X, Y) is the covariance of X and Y, and σ\sigmaσ are their standard deviations.

4. **Scaling**

Definition: Scaling transforms the features into a specific range, often [0, 1] or with mean 0 and standard deviation 1.

Purpose: Ensures that all features contribute equally to the model, particularly important for algorithms sensitive to feature scales.

Normalized vs. Standardized Scaling:

Normalized Scaling: Rescales the data to a range of [0, 1] (Min-Max Scaling).

Standardized Scaling: Centers the data around the mean with a standard deviation of 1 (Z-Score Scaling).

5. **Infinite VIF**

Reason: Infinite VIF occurs when there is perfect multicollinearity, meaning one predictor variable is a perfect linear combination of one or more other predictors.

Consequence: The model cannot differentiate between the perfectly correlated predictors, leading to unstable estimates of the coefficients.

6. **Q-Q Plot**

Definition: A Q-Q (quantile-quantile) plot compares the quantiles of a variable’s distribution to the quantiles of a theoretical distribution (usually normal).

Purpose: Used to check the normality of residuals in linear regression.

Importance: If residuals are normally distributed, the Q-Q plot points will lie on the 45-degree line, validating the assumption of normality, which is crucial for making reliable inferences from the model.