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?

Inference about their effect on the dependent variable:

- **Season**: The dummy variables for seasons (spring, summer, fall) indicate how the bike demand varies across different seasons compared to the baseline season (winter). For instance, if the coefficients of these dummy variables are positive and significant, it implies higher demand in those seasons compared to winter.
- **Weather Situation**: The dummy variables for weather situations (clear, mist) reflect the impact of different weather conditions on bike demand. Clear weather typically has a positive impact on demand, while mist or worse weather conditions might reduce demand.
- **Year**: The dummy variable for the year (yr_2019) shows how bike demand changes between years (2018 and 2019). If the coefficient is positive, it indicates an increase in demand in 2019 compared to 2018.
- **Holiday**: The dummy variable for holidays indicates whether the demand is higher or lower on holidays compared to non-holidays.
- **Weekday**: The dummy variables for weekdays show how bike demand varies on different days of the week compared to the baseline day (e.g., Sunday if not included as a dummy variable).
- **Working Day**: The dummy variable for working days indicates the difference in bike demand on working days compared to non-working days.

2. Importance of drop_first=True during Dummy Variable Creation

Using drop_first=True is crucial because it helps:

- **Avoid Multicollinearity**: It prevents the dummy variable trap, where multicollinearity arises due to the inclusion of all categories of a categorical variable, leading to a perfect linear relationship among them.
- **Baseline Comparison**: By dropping one category, the remaining dummy variables compare their effect relative to the dropped (baseline) category. This simplifies the interpretation of the regression coefficients.

3. Highest Correlation with the Target Variable

To determine the numerical variable with the highest correlation with the target variable (cnt), you can create a pair-plot or correlation matrix. From typical bike-sharing datasets, temp often shows the highest positive correlation with cnt.

4. Validating Assumptions of Linear Regression

After building the model on the training set, the assumptions of Linear Regression were validated as follows:

- **Linearity**: Checked by plotting the predicted values against the actual values and residuals. A linear pattern indicates a good fit.

- **Homoscedasticity**: Ensured by plotting residuals vs. predicted values. Constant variance in residuals indicates homoscedasticity.
- **Normality of Residuals**: Verified using histograms or Q-Q plots of residuals. Residuals should be normally distributed.
- **Multicollinearity**: Assessed using Variance Inflation Factor (VIF). VIF values below 10 indicate no significant multicollinearity.
- **Independence of Errors**: Ensured by plotting residuals over time (if applicable) to check for any patterns or autocorrelation.

5. Top 3 Features Contributing Significantly

Based on the final model, the top 3 features contributing significantly towards explaining the demand for shared bikes are typically:

1. **Temperature (temp)**: Directly impacts user comfort and bike usage.
2. **hum (atemp)**: Similar to temperature, affects perceived comfort.
3. **Seasonal Dummy Variables (e.g., season_summer, season_fall)**: Indicate demand variations due to seasonal changes.

These features usually show strong coefficients and statistical significance in the regression model, indicating their substantial impact on bike demand.

1. Explain the Linear Regression Algorithm in Detail?

Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (predictors). The goal is to find the best-fitting line (or hyperplane in higher dimensions) that predicts the dependent variable based on the independent variables. Here's a detailed breakdown:

Basic Concept

- **Simple Linear Regression**: Involves one independent variable (X) and one dependent variable (Y). The relationship is modeled as:

$$Y=\beta0+\beta1X$$

Where:

  - Y is the dependent variable.
  - β0 is the y-intercept (constant term).
  - β1 is the slope of the line (coefficient for X).
  - X is the independent variable.

- **Multiple Linear Regression**: Involves multiple independent variables. The relationship is modeled as:

$$Y=\beta0+\beta1X1+\beta2X2+…+\beta nXn$$

Assumptions

1. **Linearity**: The relationship between the dependent and independent variables is linear.
2. **Independence**: The residuals (errors) are independent.
3. **Homoscedasticity**: The residuals have constant variance.
4. **Normality**: The residuals are normally distributed.

Steps in Linear Regression

1. **Data Collection**: Gather data for the dependent and independent variables.
2. **Model Specification**: Define the linear relationship.
3. **Parameter Estimation**: Use methods like Ordinary Least Squares (OLS) to estimate the coefficients (β).
4. **Model Fitting**: Fit the model to the data.
5. **Model Evaluation**: Evaluate the model using metrics like R-squared, Adjusted R-squared, Mean Squared Error (MSE), etc.
6. **Prediction**: Use the fitted model to make predictions.

Ordinary Least Squares (OLS)

OLS is the most common method to estimate the coefficients in linear regression. It minimizes the sum of the squared differences between the observed and predicted values:

$$\text{Minimize} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in}))^2$$

Evaluation Metrics

- **R-squared**: Proportion of the variance in the dependent variable that is predictable from the independent variables.
- **Adjusted R-squared**: Adjusted for the number of predictors in the model.
- **MSE (Mean Squared Error)**: Average of the squared differences between actual and predicted values.

2. Explain the Anscombe's Quartet in Detail

Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, regression line), but which have very different distributions and appear very different when graphed. The quartet was created by the statistician Francis Anscombe in 1973 to illustrate the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

Key Points

- **Identical Statistical Properties**: All four datasets have the same mean, variance, correlation, and linear regression line, but they are visually different.
    - Mean of x is 9 for all datasets.

- Variance of x is approximately 11 for all datasets.
- Mean of y is 7.5 for all datasets.
- Variance of y is approximately 4.12 for all datasets.
- Correlation between x and y is 0.816 for all datasets.
- The linear regression line is Y=3+0.5X.

Interpretation

- **Dataset 1**: A typical dataset with a linear relationship.
- **Dataset 2**: A dataset where the relationship is non-linear.
- **Dataset 3**: A dataset with an outlier that influences the regression line.
- **Dataset 4**: A dataset with a vertical pattern and one outlier that has a significant effect.

Importance

- **Graphical Analysis**: Emphasizes the necessity of visualizing data before performing statistical analysis.
- **Outliers and Patterns**: Highlights how outliers and patterns can influence statistical measures and misleading interpretations.

3. What is Pearson's R?

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear correlation between two variables X and Y. It quantifies the strength and direction of the linear relationship between the variables.

Formula

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

Where:

- Xi and Yi are the individual sample points.
- $\bar{X}$ and $\bar{Y}$ are the means of X and Y.
- n is the number of data points.

Interpretation

- **Range**: -1 to 1.
  - r=1: Perfect positive linear correlation.
  - r=−1: Perfect negative linear correlation.
  - r=0: No linear correlation.

Assumptions

1. **Linearity**: Assumes a linear relationship between the variables.
2. **Homogeneity of Variance**: The variance of one variable is stable across levels of the other variable.
3. **Normality**: The variables are normally distributed.

4. What is Scaling? Why is Scaling Performed? What is the Difference Between Normalized Scaling and Standardized Scaling?

Scaling

Scaling is the process of transforming the features of data to a specific range or distribution. It is a crucial preprocessing step in machine learning and data analysis to ensure that features contribute equally to the analysis.

Reasons for Scaling

- **Algorithm Efficiency**: Many machine learning algorithms perform better with scaled data.
- **Model Convergence**: Helps in faster convergence during model training.
- **Equal Contribution**: Ensures all features contribute equally to the model.

Types of Scaling

1. **Normalization (Min-Max Scaling)**
    - **Formula**:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

    - **Range**: Transforms data to a fixed range, typically [0, 1].
    - **Use Case**: When the data does not have outliers and you want to preserve the relationship among features.
2. **Standardization (Z-score Scaling)**
    - **Formula**:

$$X' = \frac{X - \mu}{\sigma}$$

    - **Mean**: 0
    - **Standard Deviation**: 1
    - **Use Case**: When the data has outliers and you want to center the data.

5. Why Does the Value of VIF Become Infinite Sometimes?

VIF (Variance Inflation Factor) measures the extent of multicollinearity in a set of multiple regression variables. A VIF of 1 indicates no correlation between a given predictor and any other predictor, while values greater than 1 indicate the presence of multicollinearity.

Reason for Infinite VIF

- **Perfect Multicollinearity**: Occurs when one predictor variable is an exact linear combination of one or more of the other predictor variables. This results in an undefined or infinite VIF value because the variance of the regression coefficient becomes infinitely large.

Example

- If two predictors X1 and X2 are perfectly correlated, the VIF calculation will involve division by zero, leading to an infinite value.

6. What is a Q-Q Plot? Explain the Use and Importance of a Q-Q Plot in Linear Regression.

Q-Q Plot (Quantile-Quantile Plot)

A Q-Q plot is a graphical tool to assess if a dataset comes from a theoretical distribution such as the normal distribution. It plots the quantiles of the dataset against the quantiles of the theoretical distribution.

Steps to Create a Q-Q Plot

1. **Calculate Quantiles**: Compute the quantiles of the sample data.
2. **Theoretical Quantiles**: Compute the corresponding quantiles from the theoretical distribution.
3. **Plotting**: Plot the sample quantiles against the theoretical quantiles.

Interpretation

- **Straight Line**: Data follows the theoretical distribution.
- **Deviation**: Indicates departures from the theoretical distribution (e.g., skewness, kurtosis).

Importance in Linear Regression

- **Normality Assumption**: Linear regression assumes that the residuals (errors) are normally distributed.
- **Residual Analysis**: Q-Q plots help in diagnosing if the residuals deviate from normality, indicating potential issues with the model or data.