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?

   Effect of Categorical Variables on the Dependent Variable "Count of Total Rental Bikes" (cnt):

   - **Season**: Seasons play a significant role in bike rentals. For example, rentals in summer and fall are typically higher due to better weather, as indicated by the distribution in the dataset.
   - **Weather Situation**: Clear weather or partly cloudy days (coded as 1) are likely to have a positive impact on bike rentals, while mist, cloudy, or worse weather reduces rentals.
   - **Holiday and Working Day**: Holidays typically experience lower bike rentals, while working days, especially in the summer and fall, may lead to more consistent rentals for commuting purposes.

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

   When performing regression analysis or machine learning with categorical variables, we use **dummy variables** to represent those categories numerically. Here's why drop_first=True is important:

   - **Multicollinearity**: Multicollinearity happens when independent variables in the model are highly correlated with each other. This creates instability in the model, leading to unreliable coefficients.
   - **Dummy Variable Trap**: If you create dummy variables for all categories of a categorical variable (for example, for the season variable which has four categories), the model will include redundant information. Essentially, you can predict one category from the others, which violates the assumption of independent predictors.
   - **Baseline Reference**: By dropping the first category (using drop_first=True), you create a **baseline reference**. The coefficients of the other categories will represent their impact relative to the dropped (baseline) category. For example, if spring is dropped from the season dummy variables, the coefficients of summer, fall, and winter will tell us how much bike rentals increase or decrease relative to spring.

   In summary, drop_first=True is critical to avoid multicollinearity and to interpret the model's coefficients meaningfully.

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

After analyzing the correlation among the numerical variables to identify which ones are most strongly associated with bike rentals (cnt).

Based on the initial analysis of the heatmap:

- **Temperature (temp)**: The variable with the highest positive correlation with bike rentals is **temperature**. As temperature rises, more people are likely to rent bikes, especially for leisure activities.
    - The correlation between temperature and total rentals is strong because biking becomes more appealing in mild weather conditions. During very cold or very hot weather, rentals might drop off, but generally, moderate temperatures lead to more rentals.

Other factors like **humidity** and **windspeed** have a weaker, or even negative, correlation with the number of rentals.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
    - **Linearity**:
        - **Scatter plot between temperature and total rentals**: The scatter plot between temperature and bike rentals shows a clear positive linear trend, which validates the assumption of linearity between the dependent variable and at least some independent variables.
    - **Multicollinearity**:
        - **Dummy variable creation with drop_first=True**: By dropping one category from each categorical variable, we avoided multicollinearity, ensuring that the predictors are independent of each other.
    - **Homoscedasticity**:
        - This assumption requires the residuals (differences between observed and predicted values) to have constant variance. Although the residuals plot wasn't explicitly shown, this could be checked by plotting the residuals against predicted values. If there's no pattern, the assumption holds.
    - **Normality of Residuals**:
        - We can validate this by plotting a **Q-Q plot** or by checking the distribution of residuals. A normal distribution of residuals would confirm that the errors are normally distributed.

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

Based on the linear regression model and correlation analysis, the top 3 features that explain the demand for bike rentals are:

1. **Temperature (temp)**: As discussed, higher temperatures generally lead to more rentals. This is because warmer weather is more conducive to biking, especially for leisure.
2. **Season**: Seasons like summer and fall are expected to contribute positively to bike rentals, while winter may have a negative impact. Summer is a key season for biking due to better weather and more daylight hours.
3. **Weather Situation (weathersit)**: Clear weather conditions (coded as 1) contribute significantly to higher bike rentals. Adverse weather conditions such as heavy rain or snow discourage people from renting bikes, leading to a negative impact on demand.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The main idea is to fit a linear equation to the observed data.

Key Concepts:

- **Dependent Variable (Y)**: The outcome or target variable we are trying to predict.
- **Independent Variables (X)**: The features or predictors that influence the dependent variable.
- **Linear Equation**: The relationship is modeled as:

$$y = b_o + b_1x_1 + b_2x_2 + b_3x_3 \ldots . + b_nx_n$$

where:

- Y is the predicted value.
- b0 is the intercept (the value of Y when all Xs are 0).
- b1,b2,...,bn are the coefficients of the independent variables.

Steps Involved:

1. **Data Collection**: Gather the dataset with the dependent and independent variables.
2. **Model Training**: Use the method of least squares to find the best-fitting line. This minimizes the sum of squared differences between the observed values and the values predicted by the model.
3. **Evaluation**: Assess the model's performance using metrics such as R-squared, Mean Squared Error (MSE), and residual analysis.
4. **Prediction**: Use the trained model to predict new data points.

Assumptions of Linear Regression:

- Linearity: The relationship between the independent and dependent variables is linear.
- Independence: The residuals (errors) are independent.
- Homoscedasticity: Constant variance of residuals.
- Normality: The residuals are normally distributed.

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics yet have very different distributions and relationships when plotted. It was created to demonstrate the importance of graphing data before analyzing it and to illustrate how statistical properties can be misleading.

The Datasets:

1. **Dataset 1**: A linear relationship with positive slope.
2. **Dataset 2**: A linear relationship with a positive slope but with one outlier.
3. **Dataset 3**: A nonlinear relationship (quadratic).
4. **Dataset 4**: A linear relationship with a high degree of clustering and an outlier.

Visual Representation:

When plotted, each dataset shows different patterns, despite having the same mean, variance, and correlation coefficient.

Importance:

Anscombe's Quartet highlights:

- The necessity of visualizing data to understand its structure.
- The potential pitfalls of relying solely on summary statistics.

3. What is Pearson's R?

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the degree of linear relationship between two variables. It is widely used in statistics and data analysis to determine how closely two variables move in relation to one another.

$$\rho_{X,Y} = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{\mathbb{E}[X^2] - (\mathbb{E}[X])^2}\,\sqrt{\mathbb{E}[Y^2] - (\mathbb{E}[Y])^2}}.$$

- r = Pearson correlation coefficient

- n = number of data points
- x and y = the two variables being compared
- ∑xy = sum of the product of paired scores
- ∑x = sum of x values
- ∑y = sum of y values
- ∑x2 = sum of squared x values
- ∑y2 = sum of squared y values

Pearson's R (or Pearson correlation coefficient) quantifies the linear relationship between two continuous variables, indicating how strongly they are correlated. R ranges from -1 to 1: a value of 1 signifies a perfect positive correlation (as one variable increases, the other does too), -1 indicates a perfect negative correlation (as one variable increases, the other decreases), and 0 suggests no correlation. It is calculated using the covariance of the variables normalized by the product of their standard deviations, making it a standardized measure of correlation. Pearson's R assumes a linear relationship and is sensitive to outliers, which can significantly affect its value.

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

Scaling refers to the process of adjusting the range of independent variables or features in the dataset. It is crucial in many machine learning algorithms to ensure that the features contribute equally to the distance calculations.

Reasons for Scaling:

- **Improve convergence**: Helps algorithms converge faster.
- **Handle different units**: Ensures that features measured on different scales do not disproportionately affect the model.
- **Enhance performance**: Improves the performance of gradient descent and other optimization algorithms.

Types of Scaling:

1. **Normalized Scaling (Min-Max Scaling)**:
    - Transforms features to a range between 0 and 1.

      X_new = (X - X_min)/(X_max - X_min)

2. **Standardized Scaling (Z-score Normalization)**:
    - Centers the data around the mean and scales based on standard deviation.

      X_new = (X - mean)/Std

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

Variance Inflation Factor (VIF) is a measure used to quantify how much the variance of a regression coefficient is inflated due to multicollinearity (correlation between independent variables).

When VIF is Infinite:

- VIF becomes infinite when one independent variable is a perfect linear combination of other independent variables (perfect multicollinearity). This means that the independent variables are not providing any new information about the dependent variable, leading to instability in the model.

Implication:

When VIF is infinite, it is an indication that the dataset should be re-evaluated, and redundant variables should be removed or combined.

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

A **Quantile-Quantile (Q-Q) plot** is a graphical tool used to compare the distribution of a dataset against a theoretical distribution, typically the normal distribution.

Purpose:

- To assess whether the data follows a specific distribution (e.g., normality).
- To identify deviations from the expected distribution.

How it Works:

- The quantiles of the dataset are plotted against the quantiles of the theoretical distribution.
- If the points fall approximately along the reference line (y = x), the data is likely normally distributed.

Importance in Linear Regression:

- **Normality of Residuals**: In linear regression, one of the assumptions is that the residuals should be normally distributed. A Q-Q plot helps visualize this.

- **Model Validation**: It aids in validating the appropriateness of the linear regression model, ensuring the validity of statistical inferences made from the model.