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?**

    - The categorical variables (`season`, `weathersit`, `mnth`, `weekday`) in the dataset, when encoded as dummy variables, show varying impacts on the dependent variable (`cnt`). For instance, seasons (`season_Summer`, `season_Winter`), certain months (`mnth_Aug`, `mnth_Dec`), and weather situations (`weathersit_Light Snow`, `weathersit_Mist`) significantly influence bike demand. The coefficients indicate the relative increase or decrease in bike demand when these categorical conditions are present.

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

    - Using `drop_first=True` during dummy variable creation is crucial to avoid the dummy variable trap, which occurs when one dummy variable can be perfectly predicted by others, leading to multicollinearity. Dropping the first category ensures that the model remains well-defined and the coefficients are interpretable.

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

    - Based on the pair-plot analysis among the numerical variables, **temperature (`temp`)** exhibits the highest correlation with the target variable (`cnt`). Despite having a high VIF (Variance Inflation Factor), indicating potential multicollinearity with other variables, `temp` remains crucial due to its significant impact on bike demand. It was retained in the final model alongside other important features, contributing substantially to explaining variations in bike rentals. Moreover, the overall VIF was effectively reduced to nearly 5 after selecting the final set of features, ensuring the model's interpretability.

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

    - *Linearity*: Checked through scatter plots of actual vs. predicted values to ensure a linear relationship.
    - *Normality of Errors*: Verified using a histogram of residuals to check for normal distribution.

- *Homoscedasticity*: Evaluated by plotting residuals vs. fitted values to ensure constant variance.
- *Multicollinearity*: Assessed using VIF values, leading to the removal of `temp` and `hum` due to high VIF.
- *Independence of Errors*: Residual plots test were used to check for autocorrelation.

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 provided model summary and considering the variables' coefficients and significance levels, the top 3 features contributing significantly towards explaining the demand for shared bikes are:

        1. **Temperature (temp):**
            - **Coefficient:** 0.5181
            - **P-value:** 0.000
            - **Interpretation:** Temperature has the highest coefficient among all variables, indicating the strongest positive impact on bike demand. As temperature increases, bike rentals tend to increase significantly.

        2. **Year (yr):**
            - **Coefficient:** 0.2324
            - **P-value:** 0.000
            - **Interpretation:** Year also has a substantial positive impact on bike demand. This suggests that over time, there has been an increasing trend in bike rentals, possibly due to growing popularity or improvements in bike-sharing services.

        3. **Weather Condition (weathersit_Light Snow):**
            - **Coefficient:** -0.2881
            - **P-value:** 0.000
            - **Interpretation:** Light snow has a significant negative impact on bike demand. When there is light snowfall, bike rentals decrease notably due to unfavorable weather conditions.

    These three variables stand out as the top features contributing significantly to explaining bike demand based on their coefficients and statistical significance (P-values). Temperature and year positively influence demand, while light snow negatively impacts it. These findings provide insights into the key factors driving variations in bike rentals in the model's context.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

- Linear regression is a statistical method for modeling the relationship between a dependent variable yyy and one or more independent variables XXX. The simplest form, simple linear regression, involves a single independent variable. The model assumes a linear relationship between XXX and yyy and can be represented as:

  y=β0+β1X+ϵy = \beta_0 + \beta_1 X + \epsilony=β0+β1X+ϵ

  Here, β0\beta_0β0 is the intercept, β1\beta_1β1 is the slope, and ϵ\epsilonϵ is the error term. The goal is to estimate β0\beta_0β0 and β1\beta_1β1 such that the sum of squared residuals (the differences between observed and predicted values of yyy) is minimized.

  For multiple linear regression, the model extends to multiple independent variables:

  y=β0+β1X1+β2X2+…+βpXp+ϵy = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilony=β0+β1X1+β2X2+…+βpXp+ϵ

  The estimation of parameters β\betaβ is typically done using the method of least squares, which minimizes the sum of the squared differences between observed values and the values predicted by the model. Mathematically, this is expressed as:

  minβ∑i=1n(yi−(β0+β1xi1+β2xi2+…+βpxip))2\min_{\beta} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}))^2minβ∑i=1n(yi−(β0+β1xi1+β2xi2+…+βpxip))2

  Key assumptions of linear regression include:

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

  Independence: Observations are independent of each other.

  Homoscedasticity: The residuals (errors) have constant variance.

Normality: The residuals of the model are normally distributed.

Model evaluation metrics include the R-squared value, which indicates the proportion of the variance in the dependent variable that is predictable from the independent variables, and various residual plots to assess the assumptions of the model.

2. Explain the Anscombe's quartet in detail.

○ Anscombe's quartet consists of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, and linear regression line) but differ greatly when graphed. It was created by Francis Anscombe in 1973 to demonstrate the importance of graphing data before analyzing it and the effects of outliers and non-linearity on statistical properties.

The four datasets each consist of eleven (x, y) points, and while they share similar summary statistics, their distributions and the relationships between the variables are markedly different when visualized:

i. Dataset 1: A typical dataset with a linear relationship.
ii. Dataset 2: A dataset with a clear non-linear relationship.
iii. Dataset 3: A dataset with a linear relationship but influenced by an outlier.
iv. Dataset 4: A dataset with two distinct clusters of points.

These examples highlight that relying solely on summary statistics without visualizing the data can be misleading, and different data distributions can yield the same statistical results.

3. What is Pearson's R?

- Pearson's R, also known as the Pearson correlation coefficient, measures the strength and direction of the linear relationship between two variables. It is denoted as rrr and 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.

Pearson's R is calculated using the formula:

r=∑(xi−x‾)(yi−y‾)∑(xi−x‾)2∑(yi−y‾)2r = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2}}r=∑(xi−x)2∑(yi−y)2∑(xi−x)(yi−y)

Where x‾\overline{x}x and y‾\overline{y}y are the means of the xxx and yyy variables, respectively. Pearson's R helps in understanding how changes in one variable are associated with changes in another, but it is sensitive to outliers and only measures linear relationships.

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

   - Scaling is the process of transforming data to fit within a specific range or distribution. It is essential in machine learning because many algorithms are sensitive to the magnitude of the data and perform better when features are on a similar scale.

     Normalized Scaling (Min-Max Scaling): This method transforms the data to fit within a specified range, typically 0 to 1. It is done using the formula:

     Xscaled=X−XminXmax−XminX_{\text{scaled}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}Xscaled=Xmax−XminX−Xmin

     Normalization ensures that all features contribute equally to the model, which is crucial for algorithms like k-nearest neighbors and neural networks.

     Standardized Scaling (Z-score Normalization): This method transforms the data to have a mean of 0 and a standard deviation of 1. It is done using the formula:

     Xscaled=X−μσX_{\text{scaled}} = \frac{X - \mu}{\sigma}Xscaled=σX−μ

     Where μ\muμ is the mean and σ\sigmaσ is the standard deviation of the feature. Standardization is essential for algorithms that assume normally distributed data, such as linear regression, logistic regression, and principal component analysis.

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

    - The Variance Inflation Factor (VIF) quantifies the degree of multicollinearity in a regression model. It measures how much the variance of an estimated regression coefficient increases due to collinearity. VIF is calculated as:

      VIF=11−Ri2\text{VIF} = \frac{1}{1 - R_i^2}VIF=1−Ri21

      Where Ri2R_i^2Ri2 is the R-squared value obtained by regressing the iii-th predictor on all other predictors. If Ri2=1R_i^2 = 1Ri2=1, indicating perfect multicollinearity, the denominator becomes zero, making VIF infinite. This occurs when a predictor is a perfect linear combination of other predictors, leading to an ill-conditioned regression model where it is impossible to estimate unique regression coefficients.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

    - A Q-Q (quantile-quantile) plot is a graphical tool to assess whether a dataset follows a specific distribution, typically the normal distribution. It plots the quantiles of the sample data against the quantiles of the theoretical distribution.

      In the context of linear regression, Q-Q plots are used to check the normality assumption of the residuals. The steps to create a Q-Q plot for residuals are:

        1. Compute the residuals from the regression model.
        2. Order the residuals.
        3. Plot the ordered residuals against the theoretical quantiles of the normal distribution.

      If the residuals are normally distributed, the points should lie approximately along a 45-degree reference line. Deviations from this line indicate departures from normality, which could suggest issues with model assumptions, such as non-linearity, heteroscedasticity, or the presence of outliers.

      The Q-Q plot is essential for diagnosing the adequacy of the linear regression model and ensuring the validity of statistical inferences based on the model.