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? (3 marks)

- **Season**: Different seasons have varying impacts on bike demand. For example, bike demand increases significantly during summer and fall, while spring and winter show lower demand due to temperature variations.
- **Weather Situation**: Clear weather (weathersit=1) increases bike usage, while poor weather conditions (e.g., mist, rain, or snow) lead to lower bike rentals.
- **Year**: The data shows that bike demand increased from 2018 (yr=0) to 2019 (yr=1), indicating a rising trend in bike-sharing popularity.

These categorical variables have distinct impacts on bike demand, allowing the company to predict and plan for fluctuations throughout the year.

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

Using drop_first=True prevents multicollinearity by avoiding the "dummy variable trap." In regression models, if all categories are represented by dummy variables, one category can be perfectly predicted from the others, causing multicollinearity. By dropping one category, we can avoid this issue, ensuring that the model can run efficiently without redundant information.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

The numerical variable **temperature (temp)** typically has the highest correlation with the target variable cnt (total bike rentals), as bike usage increases on warmer days.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

To validate the assumptions of linear regression, the following checks were performed:

- **Linearity**: Checked through the residual plot to ensure that residuals are randomly scattered.
- **Homoscedasticity**: Verified using the residual plot to check that residuals have constant variance.
- **Multicollinearity**: Checked using the VIF (Variance Inflation Factor) to ensure that independent variables are not highly correlated.
- **Normality of Residuals**: Assessed using a Q-Q plot to ensure that the residuals follow a normal distribution.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand for shared bikes? (2 marks)

The top 3 features contributing significantly towards explaining bike demand are:

- **Temperature**: A key factor affecting bike rentals, with warmer weather increasing demand.
- **Season (Summer/Fall)**: Bike rentals tend to be higher during these seasons.
- **Year**: An increase in demand from 2018 to 2019 indicates the growing popularity of bike-sharing.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression is a statistical method used to model the relationship between a dependent variable (Y) and one or more independent variables (X). In the case of multiple linear regression, the formula is:

Y=β0+β1X1+β2X2+⋯+βnXn+ϵY = \beta_0 + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_nX_n + \epsilonY=β0+β1X1+β2X2+⋯+βnXn+ϵ

Where:

- YYY is the dependent variable,
- X1,X2,…,XnX_1, X_2, \dots, X_nX1,X2,…,Xn are the independent variables,
- β0,β1,…,βn\beta_0, \beta_1, \dots, \beta_nβ0,β1,…,βn are the coefficients representing the effect of each independent variable on YYY,
- ϵ\epsilonϵ is the error term.

Linear regression works by minimizing the sum of squared residuals (the difference between observed and predicted values) to find the best-fit line.

2. Explain the Anscombe’s quartet in detail. (3 marks)

Anscombe’s quartet consists of four datasets that have nearly identical statistical properties (mean, variance, correlation, and regression line), yet they have very different distributions when graphed. The quartet demonstrates the importance of visualizing data before drawing conclusions from summary statistics, as different data patterns can yield similar statistical metrics but may require different interpretations.

3. What is Pearson’s R? (3 marks)

Pearson's R (also called the Pearson correlation coefficient) measures the linear relationship between two continuous variables. Its value ranges from -1 to 1, where:

- **+1**: Perfect positive correlation,
- **0**: No correlation,
- **-1**: Perfect negative correlation.

A high positive or negative value indicates a strong linear relationship, while a value close to zero suggests weak or no linear association.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling adjusts the range of independent variables to ensure that they have comparable units, which is crucial for models like linear regression, where unscaled data can result in biased or inaccurate predictions.

- **Normalized Scaling** (Min-Max Scaling): Rescales features to a fixed range, usually [0, 1], using the formula:
  $$X_{\text{scaled}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$
- **Standardized Scaling** (Z-score scaling): Centers the data around the mean and scales it based on the standard deviation, using the formula:
  $$X_{\text{scaled}} = \frac{X - \mu}{\sigma}$$

Normalization is used when data has varying ranges, while standardization is used when data has outliers.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

A VIF value becomes infinite when one or more independent variables are perfectly collinear, meaning they are exact linear combinations of other variables. This perfect multicollinearity causes the regression model to break down, making it impossible to compute a unique solution.

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 used to compare the distribution of a dataset against a theoretical distribution (e.g., normal distribution). In linear regression, Q-Q plots are often used to check if the residuals follow a normal distribution. If the points on the Q-Q plot lie along the reference line, it suggests that the residuals are normally distributed, which is an assumption for valid linear regression analysis.