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?

Season: The season variable showed a significant effect on bike demand. For instance, the demand was higher in summer and fall compared to winter and spring.
Weather Situation: Clear weather conditions had a positive impact on bike demand, while adverse weather conditions like heavy rain or snow reduced the demand.
Year: The year variable indicated an increasing trend in bike demand from 2018 to 2019, suggesting growing popularity and usage of bike-sharing services over time.

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

Using drop_first=True helps to avoid the dummy variable trap, which occurs when the dummy variables are highly collinear. By dropping the first category, we prevent multicollinearity and ensure that the model can uniquely identify each category without redundancy.

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

The variable temp (temperature) showed the highest correlation with the target variable cnt (total rental bikes). Higher temperatures generally led to increased bike rentals.

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

Linearity: Checked by plotting the residuals against the predicted values. The residuals should be randomly scattered around zero.
Homoscedasticity: Ensured by examining the residual plot for constant variance.
Normality: Validated using a Q-Q plot to check if the residuals follow a normal distribution.
Independence: Assumed based on the nature of the data collection process.

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

Temperature (temp): Higher temperatures positively influenced bike demand.
Feeling Temperature (atemp): Similar to actual temperature, the perceived temperature also had a significant positive effect.
Year (yr_2019): The year 2019 showed a higher demand compared to 2018, indicating an increasing trend in bike usage.

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 goal is to find the best-fitting line (or hyperplane in higher dimensions) that minimizes the sum of the squared differences between the observed and predicted values. The equation of the line is given by:
$y = \beta0 + \beta1x1 + \beta2x2 + ... + \betanxn + \varepsilon$
where y is the dependent variable, x1, x2, ..., xn are the independent variables, β0 is the intercept, β1, β2, ..., βn are the coefficients, and ε is the error term.

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, etc.) but have very different distributions and appear very different when graphed. It demonstrates the importance of visualizing data before analyzing it, as relying solely on summary statistics can be misleading.

3. What is Pearson's R?

Pearson's R, also known as the Pearson correlation coefficient, measures the linear correlation between two variables. It ranges from -1 to 1, where 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship.

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

Scaling is the process of transforming the features to a common scale without distorting differences in the ranges of values. It is performed to ensure that all features contribute equally to the model and to improve the convergence of optimization algorithms.
Normalized Scaling: Rescales the data to a range of [0, 1] or [-1, 1].
Standardized Scaling: Transforms the data to have a mean of 0 and a standard deviation of 1.

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

An infinite VIF value occurs when there is perfect multicollinearity, meaning one predictor variable is a perfect linear combination of other predictor variables. This makes it impossible to estimate the regression coefficients uniquely.

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

A Q-Q (quantile-quantile) plot is a graphical tool to assess if a dataset follows a particular distribution, typically the normal distribution. In linear regression, it is used to check the

normality of residuals. If the residuals follow a straight line in the Q-Q plot, it indicates that they are normally distributed, which is an assumption of linear regression.