Assignment-based Questions

1. Effect of categorical variables on dependent variable:

Categorical variables like season and weathersit affect bike demand by capturing different market conditions (seasonal effects and weather). These variables help model nonlinear relationships and group differences in demand behaviour that numeric variables alone cannot describe.

2. Importance of `drop_first = True` during dummy variable creation:

This avoids the dummy variable trap by dropping one dummy category to serve as a baseline. It prevents redundant features and allows the linear regression to estimate coefficients correctly.

3. Highest correlation numerical variable with target from pair-plot:

Typically, temperature-related variables (temp or atemp) could show the highest positive correlation with demand because favourable temperatures encourage bike rentals.

4. Validation of linear regression assumptions on training Assumptions

validated usually include linearity (scatter plots), normality of residuals (Q-Q plot), homoscedasticity (residuals vs fitted plot with constant variance), and independence of errors (Durbin-Watson test).

5. Top 3 significant features explaining bike demand:

Variables like year (yr), temperature (temp) or feeling temperature (atemp), and working day or holiday status often significantly explain variations in bike demand.

General Linear Regression Questions

1. Linear regression algorithm detail:

Linear regression models the relationship between dependent and one or more independent variables by fitting a linear equation to observed data. The model estimates coefficients minimizing sum of squared residuals to predict continuous outcomes.

2. Anscombe's quartet:

A set of four datasets with nearly identical simple statistics (mean, variance, correlation) but different distributions and relationships. It demonstrates the importance of graphical data analysis rather than just relying on statistics.

3. Pearson's R:

A measure of linear correlation between two variables ranging from -1 to +1. Values close to ±1 indicate strong linear relationships, while values near 0 indicate weak or no linear relationships.

4. Scaling purpose and normalization vs standardization:

Scaling ensures features have comparable ranges. Normalization rescales data to, while standardization transforms data to have mean zero and unit variance. Both help in model training convergence.

5. Reason for infinite VIF values sometimes:

Occurs due to perfect multicollinearity when a variable is an exact linear combination of others, making the variance inflation factor infinite.

6. Q-Q plot and its importance in linear regression:

Q-Q plot compares the distribution of residuals to a theoretical normal distribution to check normality assumption. Deviations from the straight line imply violation of normality in residuals.