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?

Categorical variables like `season`, `weathersit`, and `weekday` can significantly influence bike demand. For example, the `season` variable might show higher bike usage during spring and summer compared to winter. Similarly, `weathersit` can affect demand, with fewer bikes rented during bad weather conditions like rain or snow.

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

Using `drop_first=True` avoids the dummy variable trap by eliminating one category from each categorical variable. This prevents multicollinearity, which occurs when one predictor variable can be linearly predicted from the others. For example, if a `season` variable is encoded into four dummy variables (spring, summer, fall, winter), dropping one (e.g., winter) avoids redundancy in the model.

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

Typically, the `temp` variable (or adjusted temperature `atemp`) shows a high positive correlation with `cnt`, indicating that bike demand increases with temperature.

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

Assumptions were validated by:

1. Checking for linearity by plotting residuals versus predicted values.
2. Assessing homoscedasticity through a plot of residuals versus fitted values, ensuring there's no pattern.
3. Checking for normality of residuals using a Q-Q plot.
4. Ensuring multicollinearity is low by calculating VIF for each predictor.

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

The top 3 features are likely to be `temp` (temperature), `hum` (humidity), and `weathersit` (weather situation), as they directly influence the comfort and feasibility of biking.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression predicts the value of a dependent variable (Y) based on one or more independent variables (X). The model assumes a linear relationship between X and Y, represented as \(Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_nX_n + \epsilon \). The algorithm minimizes the sum of squared residuals to find the best-fit line. It's used in scenarios where the relationship between variables is linear, like predicting house prices based on size and location.

2.Explain the Anscombe's quartet in detail.

Anscombe's quartet consists of four datasets with nearly identical statistical properties (mean, variance, correlation) but vastly different distributions when graphed. It highlights the importance of visualizing data before drawing conclusions, showing how different trends and relationships can be hidden or misrepresented by relying solely on summary statistics.

3.What is Pearson's R?

Pearson's R is the correlation coefficient that measures the linear relationship between two variables, ranging from -1 to 1. A value of 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no linear relationship. For example, a Pearson's R of 0.8 between temperature and ice cream sales indicates a strong positive correlation.

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

Scaling adjusts the range of features in the dataset. Normalized scaling (Min-Max scaling) rescales features to a fixed range, typically [0, 1], while standardized scaling centers the data to a mean of 0 and standard deviation of 1. Scaling is crucial when using algorithms like SVM or k-NN, where distances between points are important.

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

VIF becomes infinite when there is perfect multicollinearity, meaning one predictor variable is an exact linear combination of others. This usually occurs when dummy variables are not dropped appropriately or if features are linearly dependent.

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

A Q-Q plot (Quantile-Quantile plot) compares the distribution of residuals to a normal distribution. If the points lie on the 45-degree line, the residuals are normally distributed, validating the assumption of normality in linear regression. It is crucial for ensuring that confidence intervals and hypothesis tests are reliable.