Assignment-based Subjective Questions

Q1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

- Categorical variables show varying effects on the dependent variable. For instance, season and weather significantly impact bike demand, while holiday and working day have lesser influence.

Q2. **Why is it important to use drop_first=True during dummy variable creation? (2 marks)**

- Using drop_first=True in dummy variable creation prevents multicollinearity and the dummy variable trap, ensuring independence among variables.

Q3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

- Among numerical variables, 'temp' shows the highest correlation with the target variable, indicating a strong relationship.

Q4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

- Assumptions of Linear Regression were validated by examining residuals for normality, checking for homoscedasticity through residual plots, and verifying independence of residuals.

Q5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

- The top 3 features contributing significantly to bike demand are 'temp', 'yr', and 'weathersit', suggesting that weather conditions and year have a substantial impact.

General Subjective Questions

Q1. **Explain the linear regression algorithm in detail. (4 marks)**

- Linear regression predicts the relationship between independent and dependent variables by fitting a linear equation to observed data, minimizing the sum of squared differences between observed and predicted values.

Q2. **Explain the Anscombe’s quartet in detail. (3 marks)**

- Anscombe’s quartet consists of four datasets with different distributions but identical statistical properties, emphasizing the importance of data visualization in understanding relationships.

Q3. **What is Pearson’s R? (3 marks)**

- Pearson's R is a measure of the linear correlation between two variables, ranging from -1 to 1. It indicates the strength and direction of the linear relationship between variables.

Q4. **What is scaling? Why is scaling performed? What is the difference between normalised scaling and standardised scaling? (3 marks)**

- Scaling standardises the range of independent variables to ensure fair comparison and improve convergence in optimization algorithms. Normalised scaling constrains variables to a fixed range, while standardised scaling transforms variables to have a mean of 0 and standard deviation of 1.

Q5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

- Infinite VIF occurs when a predictor variable is a perfect linear combination of other variables, leading to multicollinearity issues, where one variable can be perfectly predicted by others.

Q6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

- A Q-Q plot compares the distribution of a sample to a normal distribution, assessing whether the data are normally distributed. In linear regression, Q-Q plots help validate the assumption of normality in residuals, crucial for model accuracy.