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)
Ans -

1. **Seasonality and Weather**: Rentals peak in summer and clear weather; drop in winter and adverse weather.
2. **Time Trends**: Rentals increase year-over-year, remain consistent throughout the week, slightly lower on Mondays and holidays.
3. **Working vs Non-Working Days**: Rentals are marginally higher on non-working days, likely due to leisure biking.

These insights aid in planning and optimizing bike rental services by considering seasonal trends, weather, and user behavior.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
Ans -
Using drop_first=True during dummy variable creation is important for:

1. **Avoiding Multicollinearity**: Ensures dummy variables are linearly independent, crucial for model stability and interpretability.
2. **Providing a Reference Category**: Makes coefficient interpretation clearer, with each representing the effect relative to the omitted reference category.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
Ans -
The scatterplot shows a strong linear relationship between cnt and registered, while temp, atemp, and casual also correlate positively but less strongly with cnt.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
Ans -
I validated the assumptions of Linear Regression by checking for multicollinearity using VIF, examining residual plots for homoscedasticity and independence, and ensuring normality of residuals using Q-Q plots. Additionally, I verified the linear relationship between predictors and the target variable.

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

Temp, Hum and Windspeed are the top 3 features contributing significantly towards explaining the demand of the shared bike.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
Ans -
Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The algorithm minimizes the sum of squared differences between the observed and predicted values (errors) to find the best-fitting line. This line is represented by the equation $y = mx + b$, where m is the slope and b is the y-intercept. The coefficients m and b are calculated using techniques like Ordinary Least Squares (OLS).

2. Explain the Anscombe's quartet in detail. (3 marks)
Ans -
Anscombe's quartet consists of four datasets that have nearly identical simple statistical properties, such as mean, variance, and correlation, but have very different distributions and appearances when graphed. This illustrates the importance of graphing data to understand its underlying structure and not relying solely on summary statistics. Each dataset demonstrates unique patterns and outliers that are not evident through statistical summaries alone. This highlights the potential for misleading conclusions if data is not visualized properly.

3. What is Pearson's R? (3 marks)
Ans -
Pearson's R, or Pearson correlation coefficient, is a statistical measure that quantifies the linear relationship between two variables. It ranges from -1 to 1, where 1 indicates a perfect positive linear correlation, -1 indicates a perfect negative linear correlation, and 0 signifies no linear correlation. This coefficient helps in understanding the strength and direction of the relationship between the variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Ans -
Scaling is the process of adjusting the range of data values to a standard range. It's performed to improve the performance and accuracy of machine learning models. Normalized scaling rescales the data to a fixed range, usually [0, 1], while standardized scaling adjusts 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? (3 marks)
Ans -

The Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity among the predictor variables. This means one predictor is an exact linear combination of others, leading to a division by zero in the VIF calculation.

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 a graphical tool to assess if a dataset follows a specified distribution, often a normal distribution. In linear regression, it helps check the normality of residuals, which is an important assumption for the validity of inference tests and confidence intervals. Ensuring residuals are normally distributed supports the model's reliability and accuracy.