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)

   Answer: From categorical variables, it is observed that-

   a. The count of bike rentals in higher in the year 2019.

   b. The count of bike rentals start lowest in Spring season and rise through Summer and Fall to come back down in Winter season.

   c. The count of bike rentals are higher in Clear weather with few clouds and Misty and Cloudy weather.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
   Answer: Since the same information can be represented with n-1 variables when a feature has n levels. The first dummy variables out of n can be dropped. The amount of information will remain the same.
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   Answer: Temp variable seems to have highest correlation with target variable cnt in pair plot.
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   Answer: By doing residual analysis, plotting error terms to see if they fall on a normal distribution. Also checked VIF, to eliminate highly correlated features to avoid model overfitting. Checked R2 score of test data.
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
   Answer: Season, Weathersit and Windspeed

General Subjective Questions

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

   Answer: A supervised linear regression algorithm analyzes data from labelled datasets to find linear relationship between independent and dependent variables. It is used in machine learning to make predictions about future datsets.

   When there's only one independent feature, it is called Simple Linear Regression (SLR). And when there are multiple independent features, it is called Multiple Linear Regression (MLR).
2. Explain the Anscombe's quartet in detail. (3 marks)
   Answer: Anscombe's quartet refers to four datasets having same statistical metrics like mean, variance, r-squared, correlations but having different qualitative data when plotted on graph. It emphasizes on exploring data visually rather than relying solely on statistical metrics about data.
   Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be

obvious from summary statistics alone.

3. What is Pearson's R? (3 marks)

   Answer: The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. The Pearson correlation coefficient also tells you whether the slope of the line of best fit is negative or positive. When the slope is negative, r is negative. When the slope is positive, r is positive.

   When r is 1 or –1, all the points fall exactly on the line of best fit.

   .When r is 1, it signifies positive correlation, when it is 0, it means no correlation and negative correlation, when it is -1.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

   Answer: Scaling is a method to standardize the values in independent variables in a fixed range. It is done at data pre-processing step. It is used to handle highly varying magnitude of values in data. If scaling is not done, machine learning algorithm weighs bigger values higher and smaller values lower regardless of their units.

   Normalized scaling scales the values of independent variables in the range of 0 to 1 by subtracting the mean value from the entry and then diving it by difference of min and max values.

   Standardized scaling is based on central tendencies and variance of the data. Each entry is subtracted from the mean value and then divided the standard deviation.

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

   Answer: An infinite VIF indicates a variables can exactly be expressed as a linear combination of other variables. It happens when one variable is perfectly collinear with other variables.

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

   Answer: Q-Q plot also know Quantile-Quantile plots. They're used to check validity of assumptions made for linear regression model like if residuals are normally distributed. It is a scatter plot created by plotting two different quantiles against each other. For example, if you are testing if the distribution of age of employees in your team is normally distributed, you are comparing the quantiles of your team members' age vs quantile from a normally distributed curve. If two quantiles are sampled from the same distribution, they should roughly fall in a straight line.