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. The categorical variables 'season', 'year', and 'month' from the dataset has some strong correlation with the target variable 'cnt'. These features along with numeric column 'atemp' affects the model's performance strongly. These columns help us understand the underlying pattern of customer bike usage and help is making impactful business decisions.

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

Ans. It helps in reducing extra column created during dummy variable creation. Therefore, if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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

Ans. 'temp', and 'atemp' column has the highest correlation with target variable.

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

Ans. Residual analysis ensures reliable predicted values, residuals should be randomly scattered at zero without clear pattern., homoscedasticity – ensuring that spread of residuals is consistant across all predicted values., and by calculating models mean_squared_error and r2_score to evaluate training performance.

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. 'season', 'yr', and 'atemp' are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions

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

Ans. The basic assumption for linearity starts with the relationship between the input variables and the output variable must be linear. In simple linear regression only one independent variable is highly correlated with target variable, and many independent variables are correlated with target in multiple linear regression. The goal of the algorithm is to find weights that minimizes errors using Ordinary Least Squares method. Gradient descent is used for optimization or minimize the cost function. Metrices such as mean_squared_error, mean_absolute_error, and r2_score are popular evaluation metrics for linear regression.

2. Explain the Anscombe's quartet in detail. (3 marks)

Ans. Anscombe's quartet highlights the importance of combining statistical summaries with visualization of data. While summary statistics like mean, variance, and correlation provide valuable information, they are insufficient for fully understanding the data's structure. Visualization can reveal patterns, outliers, and relationships that are otherwise hidden, leading to better modeling choices and more accurate insights.

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

Ans. Pearson's R also known as Pearson correlation coefficient (r), is a measure of the linear correlation between two variables. It quantifies the strength and direction of a linear relationship between two continuous variables, ranging from -1 to 1., 1 indicates a strong positive linear relationship, -1 indicates a strong negative linear relationship, and 0 indicates no linear relationship

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

Ans. Scaling is a data preprocessing technique where values in a column are adjusted to a common scale without distorting the difference in range of values. Scaling is performed to improve algorithm convergence, enhance model's accuracy, and avoid bias. Normalization (or Min-Max Scaling) transforms the data to fit within a specific range, typically between 0 and 1. Standardization transforms the 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. Variance Inflation Factor (VIF) is a measure used in statistics to detect multicollinearity in regression analysis. An infinite VIF indicates perfect multicollinearity, meaning one predictor can be exactly predicted by other predictors. This often occurs due to redundant features, duplicate columns, or the inclusion of all dummy variables. To resolve infinite VIF, it's crucial to, handle dummy variables correctly, and check for linear dependencies among features.

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

(3 marks)

Ans. A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare the distribution of a dataset with a normal distribution. It plots the quantiles of the sample data against the quantiles of a reference distribution, helping to visually assess how closely the data follows that distribution. The Q-Q plot is primarily used to check the assumption that the residuals follow a normal distribution, which helps to evaluate normality of residuals, detect skewness and outliers, and improve model validity.