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)

[Within the categorical variables Weather plays an important role , demand can drop during conditions like Light Snow , Rain and to some extend in Mist , Cloudy . The demand is low for winter months like Nov , Dec , Jan , Feb . However there is demand in other months of the year. On working days (Mon , Tues , Wed , Thus , Fri) demand is low , but on weekends demand it high. People also prefer less use of the service during spring season]

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

[This is done to drop any one category from the available categories in a categorical variable . If there are n categories then n-1 categories can explain the outcome for nth category]

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

[temp has the highest correlation with the target variable]

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

[Checked with P value and VIF value if coming in preferred range . 3 variables showed bit high VIF value but dropping it caused a significant dip in r2_score hence restored it as significant. Checked error term distribution and distribution was centered at 0.0]

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

[Top 3 features contributing significantly towards explaining the demand of the shared bikes are : 'temp' , 'yr' , 'weathersit = Light Snow + Rain']

General Subjective Questions

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

[Linear regression algorithm provides a linear relationship between a predictor/independent and the target/dependent variable. It is a supervised learning algorithm where we work on training the model with training data set and validate the r2_score using the test data set. The goal of the algorithm is to find the best fitted line that is not overfitting but at the same time can predict target variable value]

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

[Anscombe's quartet proves that although we can statistically get same results , from a visual representation it can give a different story. Anscombe's quartet consists of 4 datasets that end up giving same outcomes statistically but visualization differ for each dataset]

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

[Pearson's R measures the correlation between 2 continuous variables. The value can range between -1 to 1.

1 Indicates that the 2 variables increase proportionally in positive direction

0 Indicates no linear regression between the variables

-1 Indicates that if 1 variable increase , the other decreases proportionally.]

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

[Scaling refers to the process of shrinking the values of different features to a common scale before we start with model building.

Scaling is performed to overcome the issue of coefficients of variable that can range from very high to very low.

Normalized scaling includes doing a min max scaling and making all values range between 0 and 1. This is preferred when there are outliers in the dataset.

Standardized scaling is done when we know data is normally distributed]

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

[Infinite VIF occurs for an independent variable when it can be perfectly predicted by other variables in the model. The variable with infinite VIF indicates perfect multicollinearity.]

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

[Q-Q plot or Quantile Quantile plot is used to validate if graphically 2 sample datasets are being fetched from the same population or not. In linear regression Q-Q plot is used to check if the residuals are normally distributed]