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: The categorical variables season, holiday, weekday, working day, weather sit and winds peed has positive effect on dependent variable.

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

Answer: drop first=True is used to remove redundant variable after creation of dummy variable from original variable.

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 has 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)

Answer: We validated the assumptions of Linear Regression by performing Residual Analysis and R2 Score analysis.

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: The top 3 features contributing significantly towards explaining the demand of the shared bikes are **temp, season, weekday .**

General Subjective Questions

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

Answer: Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

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

Answer: Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

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

Answer: Pearson R statistical test, measures the strength between the different variables and their relationships. Therefore, whenever any statistical test is conducted between the two variables, it is always a good idea for the person analysing to calculate the value of the correlation coefficient to know how strong the relationship between the two variables is.

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

Answer: It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done, then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. sklearn.preprocessing. MinMaxScaler helps to implement normalization in python.

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

Answer: The value of VIF is calculated by the below formula:

$$VIF_i = \frac{1}{1-R_i^2}$$

Where, 'i' refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

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

Answer: The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.