Linear Regression Subjective Questions

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: Categorical variable have the same effect on dependent variable as any continuous/numeric variable will have on the dependent variable.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
Ans: This will lead to a multicollinearity trap where these variables will be correlated with each other if we don't use drop_first=True.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
Ans: 'temp' variable 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)
Ans: Error distribution is normal and error sum is close to zero.

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: 'yr' and 'season' contributed the most.

General Subjective Questions

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

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

2. Explain the Anscombe’s quartet in detail. (3 marks)
Ans: Anscombe’s quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the

distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

3. What is Pearson's R? (3 marks)
Ans: The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling
and standardized scaling? (3 marks)
Ans: Scaling of variable means bringing all variable values on a similar scale so that they can have the same effect on the dependent variable else the range and difference of coefficients of different variables will vary in range and will have and will have varying effects on the dependent variable.
Without this, some variables will have more weightage as compared to others which can lead to inconsistencies in prediction.

Normalized Scaling: This method scales the model using minimum and maximum values and final values lies between -1 and 1.

Standardized scaling: This method scales the model using the mean and standard deviation and values are not constrained to a particular range.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
Ans: When there is a strong correlation near or exactly 1, -1 then VIF will be infinite.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
Ans: Q-Q(quantile-quantile) plots play a very vital role in graphically analyzing and comparing two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x.

In linear regression, Q-Q plot answers the question is this variable normally distributed which is a key assumption for linear regression.