Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dset, what could you infer about their effect on the dependent variable? (3 marks)

Ans – Season, yr are highly corelated with cnt and humidity is negatively corelated.

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

Ans – We need n-1 clumns to describe n categorical 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

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

Ans- We use the same scaling on the test data set and then apply the model on test data set without using the fit() function. We see the r2score is similar for both train and test data set.

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- weathersit, season and humidity

General Subjective Questions

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

Ans – a model that describes a linear relationship between the input variables (x – predictor) and the single output variable (y - target). More specifically, that y can be calculated from a linear combination of the input variables (x).

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

Ans – It has four datasets and uses different distributions to describe. They created very different graphs when ploted.

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

Ans- Pearson correlation coefficient is the bivariate correlation coefficient — is a measure of linear correlation between two sets of data. Calulates values between -1 to 1.

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

Ans – Scaling is to fit all the data in similar range, like Min-Max or standardizastion . So that model will prdict the data correctly. Min-max will scale the values betw 0-1 .

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

Ans – If we didn't use drop_first=true to drop one column for dummy variables

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

Ans – It is a graphical tool to help us to view if a set of data came from some theoretical distribution such as a Normal, exponential or Uniform distribution.

It helps to determine if two data sets come from populations with a common distribution or not.
Also can be used in sample sizes.