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. From my Analysis I can infer that

- Bike Rents are increasing year on year
- People rent more on working days which also means that they rent more on non-holidays
- People also rent more in clear weather when temperature and humidity are less

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

Ans. It helps in reducing the extra column that was created during dummy variable creation. As a result, it reduces the correlations created among 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 has the highest correlation with count(i.e. bike renting count – target variable) followed by humidity

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

Ans.

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.

General Subjective Questions

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

Ans. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable,

such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables.

To calculate best-fit line, linear regression uses a traditional slope-intercept form.

Linear Regression equation -> y=mx+c → B0 + B1 * C

Where y= Dependent Variable and x= Independent Variable and

B0= intercept of the line

B1 = Linear regression coefficient.

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

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

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

Ans. Scaling is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

There are following types of scaling

- Normalization/Min-Max Scaling: It brings all of the data in the range of 0 and 1.
- 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)

Ans. If there is perfect correlation, then VIF = infinity. Which means there is perfect correlation between two independent variables. In this case, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this issue, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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) plot, is a helps us assess if a set of data came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.