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:

- Holiday – demand increases during Holiday.
- Working day – demand during Holidays is more than Normal Working day
- Yr – year 2019 had a significant growth in terms of business and demand increases significantly high
- Season – Summer and Fall demand increases for the bike
- Mnth – Demand continuously increases till the month of September and October
- Weekday – Demand for Bikes on Friday and Sunday is significant.

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

Ans :- This is an additional and redundant columns and doesn't add any significance in the model building.

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 highest correlation with the target

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

Ans: by checking P stats, VIF, error distribution and relationship between dependent variable and Target variable

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:

- Temp
- Holiday
- Year

General Subjective Questions

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

Ans:- Linear regression is a method of finding the best straight line fitting to the given data, i.e. finding the best linear relationship between the independent and dependent variables.

In technical terms, linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method

2. Explain the Anscombe's quartet in detail. (3 marks)
Ans:-
Anscombe's quartet consists of **four data sets** that have nearly identical simple descriptive statistics but have **very different distributions** and **appear very different when presented graphically**. Each dataset consists of eleven points.
The primary purpose of Anscombe's quartet is to illustrate the **importance of looking** at a **set of data graphically before beginning the analysis process** as the statistics merely does not give the an accurate representation of two datasets being compared

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

Ans: Pearson Correlation Coefficient is Linear relationship between 2 quantities. It indicates measure of strength between the two variables and value of coefficient can be between 1 and -1

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

Ans:- It's a preprocessing step performed during model building in machine learning to standardize independent features variables of a dataset on a fixed range.

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

Ans:- VIF Infinite indicate there is a perfect correlation between two independent variable. R-squared value is 1 in this case. This leads to VIF infinite (1/1-R2). This indicates multicollinearity and one of the variable need to be dropped for model to be working for regression.

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

1. Basically ensuring ML model for right distribution
2. Two population are of the same distribution
3. If residuals follow normal distribution, having normal error term is an assumption in an regression and we can verify it is met
4. Skewness of the distribution