**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] Most of the metrics have some significant impact on the final dependent variable, cnt. Holiday & Working Day doesn't have significant impact.

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

[Ans] The problem of Multicollinearity will raise If drop_first=True is not used. This is because The value of the first column can be derived from the other n-1 columns created.

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.

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

[Ans] I checked if the error terms are also normally distributed, which is in-fact, one of the major assumptions of linear regression. Yes, The error terms are normally distributed.

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] Season_3, Year, season_4

**General Subjective Questions**

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

[Ans] Linear regression is a supervised machine learning method that provides a linear relation between independent variables and dependent variable. This is used to forecast upcoming occurrences. Of all statistical methods, linear regression analysis is the one that is most frequently utilised.

There are 2 types of Linear Regression

a. Simple Linear Regression – 1 dependent & 1 independent variable
b. Multiple Linear Rgression – 1 dependent & multiple (more than 1) independent variables

Model performance is evaluated by R2 & Adjusted R2.

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

[Ans] Anscombe's Quartet was devised by the statistician Francis Anscombe to illustrate how important it was to not just rely on statistical measures when analysing data. To explain this in detail, Francis created 4 data sets which would produce nearly identical statistical measures but different behaviours/ distributions/ patters.

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

[Ans] The Pearson correlation coefficient (r) is t is the test statistics that is sued for measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
[Ans] Scaling is a process of redistributing multiple variable in a single common smaller range. Every variable has its own range, few might have big range (Eg - prices) & few might have small range (Quantity). Scaling is performed to bring all the variables into a single range.
Both normalized scaling and standardized scaling are 2 different techniques used to scale variables. Standardized scaling is the process of scaling features into Gaussian Distribution (mean = 0 & SD = 1). Normalization often also simply called Min-Max scaling basically shrinks the range of the data such that the range is fixed between 0 and 1

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
[Ans] VIF denotes the multicollinearity between the independent variables. If the dependent variables are perfectly correlated with the other metrics, then VIF becomes infinity. VIF=1, of variables are orthogonal to each other.

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 graphical approach to determine if two datasets come from populations with a common distributions. This is done by comparing two probability distributions by plotting their quantiles against each other