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 : 'casual','atemp','yr','season' are the variables which impact the dependent variable others I found to either have no effect of negative effect. Registered has linearly correlated hence should be ignored.

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

Ans : it helps in reducing the extra column created during dummy variable creation. Hence 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 : registered

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

Ans : r2 , adjusted r2 values as well as p values

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 : 'casual','atemp','season'

General Subjective Questions

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

Ans : It is used for data with linear relationships. Objective is to find the coefficients of the linear equation so that future data points can be used on the same equation to get dependent values(which are projection or predictions)

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

Ans : Anscombe’s Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

3. What is Pearson’s R? (3 marks)

Ans : It is a statistic that measures the linear correlation between two variables. Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables

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

Ans: Some numerical variables are larger compared to others which will impact the coefficients hence we need to bring them to similar scale as other variables which is referred as scaling.

Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation

Normalization or Min-Max Scaling is used to transform features to be on a similar scale.

```
X_new = (X - X_min)/(X_max - X_min)
```

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
Ans : This shows a perfect correlation between two independent variables

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Ans : The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set