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)**

The categorical variable like seasons has strong significant with the dependent variable , similarly for the categorical variable weathersit the pairplot shows a good linear relationship and we can notice that the dependent variable is more affected when the weathersit = 3 .

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

When we create dummy variables for categorical values drop_first = True helps to reduce one of the significant columns for the dummy set .

For e.g if for gender we have values like **Male , Female and Others** then we don't need three distinct dummy columns . It is understood that if it is not male or female it will fall in category Others so we can remove one of the columns . drop_first = True ensures to eliminate the extra column.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

The field temp has the highest co-relation according to Pair-plots and heatmap.

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

Drawing pair-plots between various numeric variables of the data set we can notice that there Exists linear relationship between the dependent variable and all the other independent variables

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

1) seasons

2) temp

3) Weathersit

General Subjective Questions :-

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

Linear regression in statistical analysis between two or more variables which has a linear relationship .this is explained by a scalar(dependent variable) and its relationship with an independent variable)

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

This quartet comprises of nearly 4 data sets that have nearly same statistical descriptions but there exist peculiarities which will go against the regression analysis .One of the four data sets in the quartet will fit into the regression model but the other data sets may have different behaviour cannot be managed with the regression model.

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

The pearson R is a co-relation co-efficient which describes the strength and direction of a linear regression .the value of R lies between 1 and -1 .

If R = 1 or R= -1 it depicts a perfect Linear relation .

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

Scaling is a data processing algorithm which computes the values of each entry in the data frame on scale of [-1, 1] . For e.g for a dataframe column with categorical values may have many entries which has no values or extremely peculiar values amongst the category and this may affect the overall analysis by drastically affecting the co-efficent . scaling actually redefines the values of a particular variable for a better comparison .

Normalised scaling : 1) the scaling is calculated using the Min and max values on the below formulae . Outliers may have effect on the normalised scaling .

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

Standardized scaling : here the scaling is calculated using the mean and the standard Deviation . Outliers will not have any impact on the standardised scaling .

```
X_new = (X - mean)/Std
```

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

VIF will be infinite if its ideal or a perfect co-relation between the dependent and independent variables . The linear regression model with the R2 value equal to 1 will have an infinite VIF .

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

A distribution between two quantiles is called a QQ plot and both the quantiles are from two different distributions. In linear regression a QQ plot can be run between the train and the test data frames in order to validate that they belong to the same population .