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)

    a. temp (Cofficent - 0.4893) : Shows that a unit increase in temp variable, increases the bike hire numbers by 0.4893 units.
    b. weathersit_3 (Cofficent - -0.2715): Shows that a unit increase in Weathersit_3 variable, decreases the bike hire numbers by 0.2715 units.
    c. yr (Cofficent - 0.2287): Shows that a unit increase in yr variable, increases the bike hire numbers by 0.2287 units.
    d. windspeed (Cofficent -0.1872): Shows that a unit increase in windspeed variable decreases the bike hire numbers by 0.1872 units.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
Ans - Will drop one column value so if there are two values in one categorical column (Men and Women) then it will create only one new column (Men or Women)

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
Ans - Linear relationship b/w "Temp", "Cnt", and "Atemp"

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
Ans – Errors must be 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 – Temperature, year, wind speed

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
Linear regression analysis is **used to predict the value of a variable based on the value of another variable**. The variable you want to predict is called the dependent variable.

2. Explain the Anscombe’s quartet in detail. (3 marks)
Anscombe's quartet is **a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different**. Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different. It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.

3. What is Pearson’s R? (3 marks)
The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by *r*. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two

variables, and the Pearson correlation coefficient, *r*, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks
In statistics, a Q–Q plot (quantile–quantile plot) is a probability plot, a graphical method for comparing two probability distributions by plotting their quantiles against each other.[1] A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). This defines a parametric curve where the parameter is the index of the quantile interval.