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)

    - season: most usage during summer and fall
    - yr: 2019 has seen much better bike usage compared to 2018
    - weathersit: higher value when weather is clear or cloudy. No usage of bikes during thunderstorm
    - weathersit + holiday: During holiday with light rain bike usage is very less. May be people prefer to stay at home.
    - season + holiday: Less usage of bike during Spring holiday, may be they are out of town.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
    - With true-false combination, we can calculate the first variable (i.e. when all others are false). So we can drop first dummy variable and still not lose any data. We can create our model using 1 less variable.

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

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
    - Check that residual errors have a mean value of zero.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
    - Weather: Clear, Few clouds, partly cloudy
    - Yr
    - Windspeed (explains negative demand)

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
    - Linear regression algorithm provides a linear relationship between an independent variable and a dependent variable. This is used to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis.
    - linear regression is a supervised learning algorithm that simulates a mathematical relationship between variables and makes predictions

2. Explain the Anscombe’s quartet in detail. (3 marks)
    - Anscombe’s quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build your model. These four data sets have nearly the same statistical observations, which provide the

same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.

- Anscombe’s quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data

3. What is Pearson’s R? (3 marks)
   - The Pearson correlation coefficient (r) is the most common way of 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)
   - scaling is a method to standardize or normalize the independent variables so that they have similar ranges or distributions
   - Scaling is performed to bring all the variables to similar scale so that it is comparable while building a model. It helps algorithms converge faster during training
   - Standardization centres data around a mean of zero and a standard deviation of one, while normalization scales data to a set range (0, 1) by using the minimum and maximum value

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

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
   - s Q-Q plots or Quantile-Quantile plots, plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.