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)

   Answer:
   - **Season :** Bike demand is **high in fall** and Bike demand is **low in spring**
   - **Yr:** Bike demand **is high in the year 2019.**
   - **Weathersit:** Bike demand is **more it is clear, few clouds, partly cloudy** or has Mist **and** Bike demand is **low if it is Light Snow, Rainy**
   - Bike demand **does not differ** much based on **holidays** or **day of the week**

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

   Answer: "drop_first=True" is important because while creating dummy variables for column, it makes sure it only creates **n-1 dummies** for the column in order to **reduce the correlations** created among dummy variables. **For example:** while creating dummies for the column "**furnished/semi-furnished/unfurnished**", it only creates dummies for furnished and semi-furnished, so that if both of them have 0's then it means that entry has "unfurnished" as value.

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

   Answer: The highest correlation is **0.63** and the variable that are highly correlated with **"cnt"** are the fields **"temp"** and **"atemp",** followed by the field **"yr"** with correlation of **"0.53"**

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

   Answer: By plotting a histogram to make sure **error terms are normally distributed, independent and have constant variance** and also plotting scatter plot/pair plot to make sure they **have linear relationship**

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

   Answer:
   1. "**temp**" with a positive coefficient
   2. "**yr**" with a positive coefficient
   3. "**weathersit_3**" with a negative coefficient

General Subjective Questions:

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

   Answer:
   - Linear Regression is a machine learning algorithm, in specific a supervised learning algorithm.
   - Linear regression is used to analyze one variable against other variable to know how well those other variables describe it.
   - The variable taken for analysis is “Target Variable” or “Dependent Variable”
   - Other variables that describe the target variable are “Independent Variable”
   - Simple linear regression model tries to fit all data points into a sloped straight line representing the relationship between the variables. Multiple linear regression fits a hyper plane instead of a line
   - It is used to estimate the coefficients of the linear equation, that explains how the other variables describe the target variable
   - This helps in predicting the value of the target variable
   - Examples: Predict the sales of a company for the next year, Forecasting demand for product in the market in the next year.

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

   Answer:
   - Anscombe’s quartet comprises of four datasets that may appear different when graphed but still have nearly identical statistical properties
   - It was constructed in the year 1973 by the statistician “Francis Anscombe”.
   - It shows the importance of graphing /data visualization.
   - It explains the effect that outliers can have on the statistical properties

3. What is Pearson’s R? (3 marks)

   Answer:
   - Pearson’s R is a measure of linear correlation between two sets of data
   - It is the covariance of two variables, divided by the product of their standard deviations
   - It has a value between -1 and 1
   - It is also called as the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation

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

   Answer:

   - Scaling is a technique applied to independent variables to normalize the data within a particular range
   - It is done in data Pre-Processing stage
   - Scaling is performed to bring all the variables to the same level of magnitude
   - Scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc
   - Scaling just changes the range of the data whereas normalization is more radical transformation
   - Normalization used min and max value whereas standardized scaling uses mean and standard deviation
   - Standardized Scaling is much less affected by outliers compared to normalization

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

   Answer:

   - If VIF is infinite it means that there is **a perfect correlation** between two variables.
   - In this case the R-squared value will be **1**, which makes the VIF value infinity
   - This is the case of perfect multicollinearity
   - To solve this we might have to drop one of the two perfectly correlated variables

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

   Answer:

   - The Quantile - Quantile plot is a graphical method for determining whether two samples of data came from the same population or not
   - It is used to see if two samples have the same tail, same distribution shape
   - The purpose of Q-Q plots is to find out if two sets of data come from the same distribution.
   - A 45 degree angle is plotted on the Q-Q plot, if the two data sets come from a common distribution, the points will fall on that reference line.
   - If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x
   - In Q-Q plots sample size need not to be equal for both samples and Since we need to normalize the dataset, we don’t need to care about the dimensions of values