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. There are ‘season’, ‘mnth’, ‘weekday’,’weathersit’ categorical variables. According to these variables , the dependent variable ‘cnt’.
    b. A positive coefficient for 'spring' (season) suggests that bike rentals tend to be higher in spring compared to the reference category (e.g., winter).
    c. A negative coefficient for 'weathersit' (weathersit) suggests that bike rentals tend to decrease when the weather situation is misty and cloudy compared to the reference category (e.g., clear weather).
    d. This understanding can help in optimizing bike rental services by identifying the factors that influence demand and making informed decisions related to marketing, pricing, and inventory management.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
    a. Using drop_first=True during dummy variable creation is important to avoid multicollinearity and improve the interpretability of the model, especially when using linear regression or other linear models.

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

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
    a. Scatter plots – The data points should be scattered around the diagonal line which indicates that predictor and the target variable.
    b. By using MultiCollinearity (VIF)– Variance Inflation Factor value, a VIF value greater than 5 indicates high multicollinearity. And also used p-value greater than 0.5.

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

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
    a. Linear regression is a fundamental supervised learning algorithm used for predictive analysis. It models the relationship between a dependent variable (target) and one or more independent variables (features) by fitting a linear equation to the observed data. Mathematical Formulation:

$y = \beta 0 + \beta 1 \times x + \epsilon$, Where
y is the dependent variable (target).
x is the independent variable (feature).
β0 is the y-intercept.
β0 is the y-intercept.
ϵ is the error term (residuals), which captures the deviation of the observed values from the predicted values.

b. For multiple linear regression with 'n' independent variables, the equation is:
$y = \beta 0 + \beta 1 \times x1 + \beta 2 \times x2 + ... + \beta n \times xn + \epsilon$
Steps to perform Linear Regressions.
Step 1: Reading, understanding and visualize the data
Step 2: Plreparing the data for modelling (train-test split, rescaling etc)
Step 3: Splitting the Data into Training and Testing Sets
Step 4: Residual analysis
Step 5: Predictions and Evaluation on the test set

2. Explain the Anscombe's quartet in detail. (3 marks)
   a. Don't know

3. What is Pearson's R? (3 marks)
   a. Pearson's R is a measure of the linear relationship between two continuous variables. It quantifies the degree to which a relationship between two variables can be described by a straight line. The value of R ranges between -1 and 1, where:
   b. The formula for Pearson's R is given by:
   $R = \sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
   R=1 or R=−1 indicates a perfect linear relationship.
   R close to 0 indicates a weak or no linear relationship.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
   a. Scaling, in the context of machine learning and data preprocessing, refers to the process of transforming the numerical features of a dataset to a specific range or distribution. It involves adjusting the values of the features so that they can be compared on common grounds without affecting their original meaning.
   Scaling is performed for several reasons:
   Interpreting Coefficients
   Improving Convergence
   Enhancing Model Performance
   Avoiding Numerical Instabilities
   Normalized Scaling (Min-Max Scaling): $Xnorm = X - Xmin / Xmax - Xmin$

   Standardized scaling : $Xstd = X - \mu / \sigma$

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
    a. The Variance Inflation Factor (VIF) measures the multicollinearity between predictor variables in a regression model.
    b. One or more predictor variables can be expressed as a perfect linear combination of other predictor variables in the model. This situation is problematic because it violates the assumption of no multicollinearity, making it impossible to estimate the coefficients accurately.
    c. Including too many predictor variables relative to the number of observations (samples) can also lead to multicollinearity issues and result in high VIF values.

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