**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)
    - *Based on the regression model seasonal and weather-related categorical variables have the strongest effect on bike demand. Demand increases significantly during summer and fall and decreases poor weather days (cloudy or rainy). So, weather plays a crucial role for good business.*
    - *The year variable also has a positive and substantial impact, indicating that overall bike usage increased from 2018 to 2019.*
    - *Other categorical variables such as weekday, holiday, and working day show weaker or inconsistent effects once season and weather variables are included.*
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
    - *It prevents multicollinearity which makes our regression model stable.*
    - *The dropped category acts as a benchmark where each dummy coefficient shows how much that category differs from baseline.*
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
    - Temperature (temp/atemp) has highest corelation with target variable cnt.
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
    - *After building model with training dataset, we validated linearity by plotting residuals vs fitted values where residuals distributed around Zero.*
    - *With a histogram plotted of residuals shows a bell-shaped distribution which confirms normality of residuals.*
    - As the residuals appeared evenly spread which indicates constant variance.
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
    - Based on the regression model, the top 3 most influential features are determined by looking at the largest-magnitude coefficients in the final model which are season_winter, mnth_9 (September), mnth_3 (March).

**General Subjective Questions**

1. Explain the linear regression algorithm in detail. (4 marks)
    - Linear Regression draws the best straight line that minimizes the total squared distance between actual and predicted values. It assumes linear relationships, learns coefficients via least squares and checks model validity through multiple diagnostic plots and statistics. We follow it via line equation Y = mX + C, where C is the intercept and m is a Coefficient or called slope as well.
    - Linear regression relies on the following assumptions:
        i. Linearity, Y is a linear combination of the predictors.
        ii. Independence of errors, Errors are not correlated with each other (checked by Durbin–Watson test).
        iii. Homoscedasticity, Constant variance of errors across all fitted values.
        iv. Normality of errors, Residuals are normally distributed (checked with histogram/Q-Q plot).

v. No multicollinearity, Predictors should not be highly correlated with one another (checked using VIF).

2. Explain the Anscombe's quartet in detail. (3 marks)
    - Anscombe's Quartet is a set of four small datasets that share almost identical summary statistics—such as mean, variance, correlation, and even the same linear regression line—but look completely different when plotted. Created by statistician Francis Anscombe, the quartet demonstrates that relying only on numerical summaries can be misleading as datasets with the same statistics can have very different patterns, trends, outliers, or shapes. The key message is that data visualization is essential before performing any analysis or building models, because plots reveal insights that summary statistics alone cannot.
    - In Summary though 4 datasets has identical summary but after plotting these datasets they look completely different So crucial lesson we have is to always visualize your data for a confirmation.
3. What is Pearson's R? (3 marks)
    - Pearson's R (Pearson correlation coefficient) measures the strength and direction of a linear relationship between two numerical variables. Its value ranges from –1 to +1.
    - +1 → perfect positive linear relationship
      –1 → perfect negative linear relationship
      0 → no linear relationship
    - It shows how strongly two variables move together. Example: Suppose temperature and ice-cream sales have a Pearson's R of +0.85. This means that as temperature increases, ice-cream sales also increase strongly in a linear pattern.
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
    - Scaling is the process of transforming numerical features so that they are on a similar range or scale.so scaling ensures all variables contribute fairly.

    - Scaling is of 2 Types
      1. Normalized Scaling (Min–Max Scaling)
         Converts values to range 0 to 1 and Sensitive to outliers

      2. Standardized Scaling

         Converts values to have mean = 0 and standard deviation = 1 and Better when data contains outliers.

    - Scaling is important because:
      1. Prevents domination by large-range features
         Variables like home price (0–100000) can dominate variables like home age (0–100).
      2. Speeds up model training
         Gradient-based algorithms converge faster.
      3. Essential for distance-based algorithms
         KNN, K-means, PCA, SVM heavily rely on scaled data.
      4. Improves model stability and performance
         Reduces numerical instability and avoids exploding coefficients.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
    - The Variance Inflation Factor (VIF) becomes infinite when one predictor variable is perfectly correlated (or almost perfectly correlated) with one or more other predictor variables. It means one feature can be expressed as an exact linear combination of others so the regression cannot distinguish between them.
    - Typical reasons as below:
        - Dummy variables created without drop_first=True
        - Repeated columns (accidental duplicates)
        - Features that are mathematical combinations of others
        - Including both a variable and its scaled/encoded version
        - Highly correlated seasonal or weekday dummies
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
    - A Q-Q plot shows whether the residuals in your regression model behave like a normal distribution. If the points lie on a straight line, the assumption is satisfied; if not, the model may need improvement.
    - If the points fall roughly along a straight 45° line, the data is approximately normally distributed.