## Assignment-based Subjective Questions

- From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)
    - From the analysis of the categorical variables in the dataset, we can make the following inferences about their effects on the dependent variable, cnt (bike rentals):
        1. **Season (season_Spring, season_Summer, season_Winter)**:
            i. **Spring** has a negative effect on bike rentals, with **892.67** fewer rentals compared to the reference season (likely fall). Spring is generally a slower season for bike rentals, possibly due to unpredictable weather.
            ii. **Summer** and **Winter** have positive effects on bike rentals, with **300.99** and **625.41** additional rentals respectively. Summer sees more outdoor activities, and winter may reflect a stable demand due to commuting needs despite cold conditions.
        2. **Weather Situation (weathersit_Mist)**:
            i. Misty or slightly cloudy weather significantly reduces bike rentals by **543.31**. This indicates that bad weather conditions discourage people from using bikes, as safety or comfort concerns may arise.
        3. **Holiday (holiday)**:
            i. Holidays reduce bike rentals by **613.68**. This implies fewer people use shared bikes on holidays, possibly because regular commuting declines, and people may choose alternative forms of transportation or leisure activities.

        These variables play a crucial role in predicting bike demand, as they influence user behavior.

- Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
    - Using drop_first=True during dummy variable creation helps avoid the dummy variable trap, a situation where multicollinearity arises because the dummy variables are perfectly correlated. By dropping the first category, we ensure that one category acts as the reference level, preventing redundancy and ensuring the

model can uniquely estimate the effects of each category without multicollinearity issues.

- Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
    - From the pair-plot among the numerical variables, temperature (temp) shows the highest correlation with the target variable (cnt), the number of bike rentals. As temperature increases, the number of bike rentals tends to increase as well, indicating a positive correlation.
- How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
    - After building the model, we validated the assumptions of linear regression by:
        - Residual Plot: Checking for randomness in the residuals versus fitted values to confirm linearity and homoscedasticity.
        - VIF (Variance Inflation Factor): Ensuring no multicollinearity among the independent variables.
        - Normality: Using the residuals' distribution to check for approximate normality, ensuring model assumptions were met.
- 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 final model, the top 3 features that contribute significantly to explaining the demand for shared bikes are:
        - Year (yr): Each year increase adds 2047.30 bike rentals, reflecting growing popularity.
        - Season (season_Winter): Winter leads to an increase of 625.41 rentals, indicating consistent demand in colder months.
        - Weather Situation (weathersit_Mist): Misty weather reduces bike rentals by 543.31, highlighting its significant negative impact.

## General Subjective Questions

- Explain the linear regression algorithm in detail. (4 marks)
    - Linear Regression is a supervised learning algorithm used to model the relationship between a dependent variable (target) and one or more independent variables (features). The goal is to find the best-fitting straight line (for simple linear regression) or a hyperplane (for multiple regression) that minimizes the distance between the actual data points and the predicted values.
        - The linear regression equation can be written as:

          y = b0 + b1x1 + b2x2 + …….Bnxn + e

        - Where: y is the dependent variable.
        - x1, x2, ….xn are the independent variables.
        - b0 is the intercept, and
        - b1, b2, b3….bn are the coefficients that represent the contribution of each independent variable.
        - e is the error term (residuals).
        - The coefficients are estimated by minimizing the sum of squared residuals (errors), which is typically solved using the Ordinary Least Squares (OLS) method.
    - **Assumptions of Linear Regression:**
        - Linearity: The relationship between the independent variables and the dependent variable is linear.
        - Independence: Observations are independent of each other.
        - Homoscedasticity: Constant variance of the residuals across all levels of

the independent variables.
        - Normality of Residuals: The residuals (errors) should be normally distributed.
        - No Multicollinearity: Independent variables should not be highly correlated with each other.
    - Meeting these assumptions is crucial for reliable and valid predictions in linear regression.
- Explain the Anscombe's quartet in detail. (3 marks)
    - Anscombe's Quartet is a set of four datasets that appear nearly identical when analyzed using basic statistical methods (e.g., mean, variance, correlation, and linear regression), but are drastically different when visualized. Created by statistician Francis Anscombe in 1973, the quartet demonstrates the importance of visualizing data rather than relying solely on summary statistics.
    - Each dataset in the quartet has the same linear regression equation, correlation coefficient (~0.82), and similar descriptive statistics. However, plots reveal differences such as non-linear relationships, outliers, and unusual data patterns, emphasizing the necessity of data visualization in analysis.
- What is Pearson's R? (3 marks)
    - Pearson's R (Pearson's correlation coefficient) is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to +1:
        - +1 indicates a perfect positive linear relationship,
        - -1 indicates a perfect negative linear relationship,
        - 0 suggests no linear relationship.
    - Pearson's R is calculated as the ratio of the covariance of the variables to the product of their standard deviations, making it a normalized value of linear dependence.
- What is scaling? Why is scaling performed? What is the difference between normalized scaling

and standardized scaling? (3 marks)

- Scaling is the process of adjusting the range of feature values in a dataset to ensure consistency, often necessary for algorithms sensitive to feature magnitudes (e.g., regression, SVM).
- Scaling is performed to prevent features with larger ranges from dominating the model and to improve convergence in gradient-based algorithms.
  - Normalized Scaling (Min-Max scaling): Rescales data to a fixed range, usually [0, 1]. It preserves the relative distances between values.
  - Standardized Scaling (Z-score scaling): Centers the data around 0 with a standard deviation of 1, ensuring a normal distribution for better handling of outliers and different feature distributions.

- You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
  - A VIF (Variance Inflation Factor) value becomes infinite when there is perfect multicollinearity among the independent variables, meaning one variable is an exact linear combination of others. This results in the denominator of the VIF formula ($1 - R^2$) becoming zero, causing division by zero, which leads to an infinite VIF. It indicates highly redundant features in the model.
- What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
  - A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, typically the normal distribution. In linear regression, a Q-Q plot is essential for checking the assumption of normality of residuals. If the residuals follow a straight line in the Q-Q plot, they are approximately normally distributed. Deviations from the line indicate non-normality, which may suggest the need for transformation or indicate problems with the model's assumptions.