# 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**: The categorical variables in the dataset such as season, weathersit, mnth, and weekday have specific impacts on the demand for shared bikes. For instance, the season variable indicates higher demand during summer and fall compared to winter and spring. Similarly, weathersit shows that clear or partly cloudy weather conditions have a higher bike demand compared to misty or rainy conditions. The mnth variable shows variations in demand across different months, and the weekday variable indicates how the demand fluctuates during weekdays and weekends. These inferences help in understanding the patterns and trends in bike-sharing demand.

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

**Answer**: Using drop_first=True during dummy variable creation helps in avoiding the dummy variable trap. The dummy variable trap occurs when the created dummy variables are highly correlated, leading to multicollinearity in the model. By dropping the first category, we reduce redundancy and ensure that the model remains interpretable and avoids issues related to multicollinearity.

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

**Answer**: Based on the pair-plot among the numerical variables, the variable registered shows the highest correlation with the target variable cnt (total count of rental bikes). This is expected as registered users typically constitute a significant portion of the total bike rentals.

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

**Answer**: The assumptions of Linear Regression were validated using the following methods:

   - **Linearity**: Scatter plots of predicted values vs. residuals were analysed to check for linear relationships.
   - **Independence**: Durbin-Watson test was used to check for independence of residuals.
   - **Homoscedasticity**: Plots of residuals vs. predicted values were used to verify constant variance (homoscedasticity).
   - **Normality**: Q-Q plots were used to check if residuals follow a normal distribution.

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**: The top 3 features contributing significantly towards explaining the demand for shared bikes, based on the final model, are:

1. **Temperature (temp)**: Higher temperatures generally lead to higher bike demand.
2. **Year (yr)**: The year 2019 shows an increase in bike demand compared to 2018.
3. **Season (season_summer and season_fall)**: Summer and fall seasons have higher bike demand compared to spring and winter.

# General Subjective Questions

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

**Answer**: Linear regression is a statistical method for modelling the relationship between a dependent variable and one or more independent variables. The algorithm aims to find the best-fitting line (regression line) that minimizes the sum of squared differences (residuals) between the observed values and the values predicted by the line. The regression line is defined by the equation:

$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta nXn + \epsilon$

where

- Y is the dependent variable, X1, X2, …, Xn are the independent variables,
- $\beta 0$ is the intercept
- $\beta 1, \beta 2, \ldots, \beta n$ are the coefficients, and
- $\epsilon$ is the error term.

The coefficients are estimated using the Ordinary Least Squares (OLS) method, which minimizes the sum of the squared residuals.

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

**Answer**: Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation) but appear very different when graphed. The quartet was created by Francis Anscombe to illustrate the importance of graphical data analysis and the effect of outliers on statistical properties. The four datasets show that relying solely on summary statistics can be misleading, and visualizing data is crucial for understanding its true nature and distribution.

3. **What is Pearson's R?** (3 marks)

**Answer**: Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear correlation between two variables. It ranges from -1 to 1, where:

- 1 indicates a perfect positive linear relationship,
- -1 indicates a perfect negative linear relationship,
- 0 indicates no linear relationship.

Pearson's R is calculated as the covariance of the two variables divided by the product of their standard deviations. It provides insights into the strength and direction of the linear relationship between variables.

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

**Answer**: Scaling is the process of adjusting the range of features in a dataset. It is performed to ensure that all features contribute equally to the model, especially in algorithms sensitive to feature magnitudes (e.g., k-nearest neighbours, gradient descent). There are two common types of scaling:

- **Normalized Scaling**: Rescales features to a fixed range, usually [0, 1], using the formula:

  $X' = (X - Xmin) / (Xmax - Xmin)$

- **Standardized Scaling**: Centers features around the mean and scales them to unit variance, using the formula:

  $X' = (X - \mu) / \sigma$

  where $\mu$ is the mean and $\sigma$ is the standard deviation.

Standardized scaling is useful for algorithms that assume a Gaussian distribution of features.

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

**Answer**: The Variance Inflation Factor (VIF) measures the extent of multicollinearity in regression analysis. A VIF value becomes infinite when there is perfect multicollinearity, meaning that one predictor variable is an exact linear combination of other predictor variables. This perfect collinearity makes it impossible to estimate the regression coefficients uniquely, resulting in an infinite VIF.

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

**Answer**: A Q-Q (quantile-quantile) plot is a graphical tool to compare the distribution of a dataset with a theoretical distribution, typically the normal distribution. It plots the quantiles of the dataset against the quantiles of the theoretical distribution. If the points on the Q-Q plot lie approximately on the reference line, it indicates that the data follows the theoretical distribution. In linear regression, Q-Q plots are used to check the normality assumption of residuals, which is important for the validity of statistical tests and confidence intervals.