**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?**
    - Season : Spring (category 1) would likely have the lowest demand compared to Summer and Fall, which would show a significant positive effect on cnt due to favourable weather.
    - Year (yr): The year 2019 (yr=1) would show a significant positive coefficient compared to 2018 (yr=0), indicating the growth in popularity of the bike-sharing system.
    - Weather (weathersit): The worst weather conditions (weathersit = 3: Light Snow/Rain) would show a significant negative effect on demand compared to clear weather (weathersit = 1).

2. **Why is it important to use drop_first=True during dummy variable creation?**
    - Using drop_first=True prevents the Dummy Variable Trap, a situation of perfect multicollinearity.
    - For a categorical variable with N levels (e.g., 4 seasons), N-1 dummy variables are sufficient for the model to capture all the information. The Nth category is implicitly represented when all N-1 dummy variables are set to 0. Keeping all N variables would make the feature matrix linearly dependent, leading to an inability to calculate a unique solution for the regression coefficients.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
    - Among the numerical variables (temp, atemp, hum, windspeed) and the target variable cnt, the feeling temperature (atemp) is expected to have the highest positive correlation with cnt. The total number of bike rentals is highly dependent on how comfortable people feel outdoors. temp would be a very close second, but atemp is often slightly more correlated.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
    - The model's assumptions (known as LINE) were validated by analysing the model's residuals:
        - Linearity: Check a scatter plot of Predicted Values vs. Actual Values (or a scatter plot of Residuals vs. Predicted Values to ensure the mean of residuals is near zero across the range).
        - Equal Variance (Homoscedasticity): Checked using a scatter plot of Residuals vs. Predicted Values. There should be no discernible pattern (like a funnel or cone shape), and the spread of residuals should be constant.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
    - While the actual model must be run, the most likely top 3 contributors would be:
        1. atemp / temp (Feeling/Actual Temperature): The most significant factor influencing outdoor activity.
        2. yr (Year 2019): Capturing the year-on-year growth and increasing popularity.
        3. A dummy variable representing Good Weather (weathersit_1 or a similar variable from season): Indicating the strong positive effect of clear, pleasant weather.

---

## General Subjective Questions

1. **Explain the linear regression algorithm in detail.**
    - Definition: Linear Regression is a statistical algorithm used to model the linear relationship between a dependent variable (Y) and one or more independent variables (X). For multiple variables, the model equation is:

      $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$

      where $\beta_0$ is the intercept, $\beta i$ are the coefficients (slopes), and $\varepsilon$ is the error term.
    - **Objective:** The goal is to determine the coefficient values $\beta i$ that minimize the difference between the actual value (Y) and the predicted value (Y^).
    - **Method:** The most common method is **Ordinary Least Squares (OLS)**. OLS minimizes the **Sum of Squared Errors (SSE)**, which is the sum of the squared differences between the observed and predicted values.
2. **Explain the Anscombe's quartet in detail.**
    - Anscombe's Quartet is a set of four distinct datasets that were created by statistician Frank Anscombe .
    - Key Insight: All four datasets share nearly identical basic statistical properties, including: the mean of the x-values, the mean of the y-values, the variance of x, the variance of y, the correlation coefficient (Pearson's r), and the equation for the linear regression line (and thus the R^2).
    - Importance: When plotted, however, the scatter plots for the four datasets show them to be dramatically different (one is linear, one is curved, one is linear with an outlier, and one has a vertical cluster). The quartet is used to emphasize that one must always visualize data and residuals to confirm assumptions and validate a model, rather than relying solely on summary statistics.
3. **What is Pearson's R?**
    - Pearson's R is a measure of the linear correlation between two sets of data.
    - It measures both the strength and the direction of a linear relationship.
    - The value of Pearson's R ranges from -1.0 to +1.0:
        - +1.0 indicates a perfect positive linear correlation (as one variable increases, the other increases).

- **-1.0** indicates a perfect negative linear correlation (as one variable increases, the other decreases).
- **0** indicates no linear correlation.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
    - Scaling is a data preprocessing technique used to transform the range of independent variables.
    - Why Performed? Scaling is essential for machine learning algorithms that are sensitive to the magnitude of the features (e.g., Linear Regression with Gradient Descent, K-Nearest Neighbors). It ensures that no single feature dominates the model simply because it has a larger range of values, leading to fairer and faster model convergence.
    - Normalization (Min-Max Scaling): Rescales the data to a fixed range, usually [0, 1]. It is calculated as:$X' = \{X - X_{\{min\}}\}/\{X_{\{max\}} - X_{\{min\}}\}$
    - Standardization (Z-score Scaling): Rescales the data so that it has a mean of 0 and a standard deviation of 1. It is calculated as: $X' = \frac{X - \mu}{\sigma}$
    - . Standardization is less affected by extreme outliers than Normalization.
5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
    - **VIF (Variance Inflation Factor)** measures how much the variance of an estimated regression coefficient is inflated due to **multicollinearity** (correlation among independent variables).
    - The VIF formula for a variable i is: VIFi = \frac{1}{1 - R_i^2}, where R_i^2 is the R^2 value from a regression of X_i against all other independent variables.
    - VIF becomes **infinite** when the denominator is zero, meaning R_i^2 = 1. This occurs when one independent variable can be **perfectly predicted** by a linear combination of the other independent variables. This is the case of **perfect multicollinearity**, which makes the calculation of unique regression coefficients mathematically impossible.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
    - Q-Q Plot (Quantile-Quantile Plot): A Q-Q plot is a graphical tool used to check if a set of data follows a specific theoretical distribution, most commonly the Normal Distribution.
    - Importance in Linear Regression: One of the key assumptions of Linear Regression is that the model's residuals (errors) are normally distributed. The Q-Q plot is used to visually test this assumption. If the residuals are normally distributed, the points on the plot will fall approximately along a straight diagonal line. Any significant departure from this line suggests a violation of the normality assumption.

**submitted by: Abhisek Gupta(abhisek.gupta@sap.com)**