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?**

- **Seasonal Trends**: Bike usage fluctuates across seasons, with peak demand typically observed in spring and summer. The warmer temperatures during these periods likely encourage more people to opt for cycling, whether for commuting or leisure.
- **Weather Conditions**: The prevailing weather has a noticeable effect on ridership. Clear, sunny days tend to see higher bike usage compared to days with precipitation or poor visibility, as adverse weather conditions can deter cyclists.
- **Holiday Effect**: Non-holiday periods often show increased bike usage, potentially due to regular commuting patterns. Conversely, holidays may see a dip in demand as fewer people need to travel to work or educational institutions.
- **Workday vs. Weekend**: Similar to the holiday effect, working days generally exhibit higher bike usage, likely attributed to commuters incorporating cycling into their daily routines.

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

When creating dummy variables, it's crucial to omit one category (usually achieved by setting `drop_first=True` in data preprocessing). This practice is vital for preventing the "dummy variable trap" - a scenario where perfect multicollinearity occurs due to the inclusion of all categories.

By excluding one category, we establish it as a reference point. This allows the model to interpret the effects of other categories relative to this baseline, avoiding redundancy and potential model instability. This approach ensures a more accurate and interpretable analysis of categorical variables' impact on bike-sharing demand.

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

Based on our correlation analysis, the variable showing the strongest relationship with overall bike usage (cnt) appears to be the number of registered users.

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

To ensure the validity of our linear regression model, we conducted several tests:

1. **Linearity**: We examined scatter plots of residuals versus predicted values to confirm linear relationships between independent and dependent variables.
2. **Constant Error Variance**: Residual plots were analyzed to check for homoscedasticity, ensuring errors maintain consistent variance across predictions.

3. **Error Distribution**: Quantile-Quantile (Q-Q) plots were used to verify that model residuals approximated a normal distribution.
4. **Error Independence**: The Durbin-Watson test was employed to check for autocorrelation in residuals, ensuring errors were not systematically related.

**5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
**Temperature**: Warmer conditions strongly correlate with increased bike usage, likely due to the comfort and appeal of cycling in pleasant weather.
**Temporal Trends**: We observed a notable increase in bike-sharing popularity from 2018 to 2019, indicating growing adoption of this transportation mode over time.
**Seasonal Effects**: Spring and summer seasons consistently show higher demand, aligning with the temperature findings and reflecting overall more favorable cycling conditions during these periods.

## General Subjective Questions

**1. Explain the linear regression algorithm in detail.**

Linear regression is a fundamental statistical and machine learning algorithm used for predicting a continuous outcome variable (dependent variable) based on one or more predictor variables (independent variables). The algorithm works by finding the best-fitting straight line (or hyperplane in higher dimensions) through the data points.

**2. Explain the Anscombe's quartet in detail**
Anscombe's quartet is a set of four datasets that have nearly identical simple statistical properties but appear very different when graphed.
It demonstrates the importance of visualizing data before analyzing it.

**3. What is Pearson's R?**
Pearson's correlation coefficient (Pearson's r) is a measure of linear correlation between two variables. It ranges from -1 to +1, where:

- +1 indicates a perfect positive linear correlation
- 0 indicates no linear correlation
- -1 indicates a perfect negative linear correlation

**4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
Scaling is the process of transforming numerical data to a common scale, without distorting differences in the ranges of values.

Why scaling is performed:

- To ensure all features contribute equally to the model
- To improve the convergence of gradient descent algorithms
- To prevent features with larger magnitudes from dominating the model

Differences between normalization and standardization:

Normalization (Min-Max scaling):

- Scales values to a fixed range, typically 0 to 1
- Formula: X_norm = (X - X_min) / (X_max - X_min)
- Preserves zero values and doesn't center the data

Standardization (Z-score normalization):

- Transforms data to have a mean of 0 and standard deviation of 1
- Formula: X_stand = (X - μ) / σ
- Centers the data around zero and handles outliers better

**5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
Variance Inflation Factor (VIF) can become infinite due to perfect multicollinearity among predictors. This occurs when:

- One predictor is an exact linear combination of others
- There's a one-to-one relationship between two or more variables
- The sample size is too small relative to the number of predictors

Infinite VIF indicates that the variable's variance is entirely explained by other predictors, making it redundant in the model. This can lead to unstable and unreliable coefficient estimates.

**6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
A Quantile-Quantile (Q-Q) plot is a graphical tool used to assess if a dataset follows a particular theoretical distribution, most commonly the normal distribution.

Use and importance in linear regression:

- Checks the assumption of normally distributed residuals
- Plots the quantiles of the observed data against the quantiles of the theoretical distribution
- If points roughly follow a straight line, it suggests the data follows the theoretical distribution

Interpreting Q-Q plots:

- Straight line: Data follows the theoretical distribution
- S-shaped curve: Data is skewed
- Curve at ends: Data has heavy or light tails compared to the theoretical distribution

Q-Q plots are crucial in linear regression because:

- They help validate the normality assumption of residuals
- They can identify outliers and potential influential points
- They guide decisions on data transformations or alternative modeling approaches if assumptions are violated

By using Q-Q plots, analysts can ensure the reliability and validity of their linear regression models.