1. From your analysis of the categorical variables from the dataset, what could you infer about
their effect on the dependent variable?

we can infer the following about the effect of each categorical variable on the dependent variable cnt:

1. **Season**:
    - The rental counts vary significantly across different seasons. Season 3 (fall) appears to have higher bike rental counts compared to other seasons.
2. **Year** :
    - There is a noticeable increase in bike rental counts from year 0 (2018) to year 1 (2019). This suggests a growing trend in bike usage over the years.
3. **Month** :
    - Bike rentals fluctuate by month, with higher counts observed during the summer months (June to August), indicating that more people rent bikes during warmer weather.
4. **Holiday**:
    - Non-holidays generally have higher rental counts compared to holidays. This may suggest that people use bikes more for commuting on regular working days than for leisure on holidays.
5. **Weekday**:
    - There is variation in rental counts across different days of the week. Weekends (Saturday and Sunday) tend to have lower rental counts compared to weekdays, indicating that bike rentals are more common on workdays.
6. **Working Day**:
    - Working days show higher rental counts compared to non-working days. This supports the inference that bikes are primarily used for commuting purposes.
7. **Weather Situation**:
    - Better weather conditions (weathersit 1: clear or partly cloudy) correlate with higher rental counts. As weather conditions worsen (weathersit 2: mist + cloudy/mist + broken clouds/mist + few clouds and weathersit 3: light snow, light rain + thunderstorm + scattered clouds, light rain + scattered clouds), the rental counts decrease significantly.

These observations suggest that bike rentals are influenced by seasonal changes, day of the week, weather conditions, and whether it is a working day or holiday. Understanding these patterns can help in planning and managing bike-sharing services more effectively.

2 Why is it important to use drop_first=True during dummy variable creation?
When creating dummy variables for categorical features, `drop_first=True` helps to avoid the dummy variable trap, which occurs due to multicollinearity. Multicollinearity is a situation where one predictor variable in a model can be linearly p
redicted from the others with a substantial degree of accuracy.

3 Looking at the pair-plot among the numerical variables, which one has the highest correlation
with the target variable?
The numerical variable with the highest correlation with the target variable (`cnt`) is `registered`, with a correlation coefficient of approximately 0.945. This indicates a very strong positive correlation between the number of registered users and the total bike demand.

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

We'll validate the assumptions of linear regression after building the model on the training set.

Steps

1. **Linearity**: Check if the relationship between the dependent and independent variables is linear.
2. **Normality**: Check if the residuals (errors) are normally distributed.
3. **Homoscedasticity**: Check if the residuals have constant variance.
4. **Multicollinearity**: Check for multicollinearity using VIF (Variance Inflation Factor).

5. Based on the final model, which are the top 3 features contributing significantly towards
explaining the demand of the shared bikes?

Season: This feature has the highest positive coefficient, indicating that it significantly affects bike demand.
Weather Situation (weathersit): This feature has a high negative coefficient, indicating adverse weather conditions reduce bike demand.
Windspeed: This feature also has a negative coefficient, suggesting higher wind speeds reduce bike demand

General Subjective Questions

1 Explain the linear regression algorithm in detail.

Linear regression is a fundamental statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (predictors). The goal of linear regression is to find the best-fitting line through the data points that can be used to predict the value of the dependent variable based on the independent variables.

2 Explain the Anscombe's quartet in detail.

Anscombe's quartet is a collection of four datasets that have nearly identical simple descriptive statistics, yet they have very different distributions and appear very different when graphed. This demonstrates the importance of graphing data before analyzing it and the limitations of simple descriptive statistics.

3 What is Pearson's R?

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two variables. It quantifies the degree to which a change in one variable is associated with a change in another variable. Pearson's R 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.

4 What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is a process used in data preprocessing to adjust the values of features so that they are on a comparable scale. This is important because many machine learning algorithms are sensitive to the scale of the data, which can affect their performance. Scaling ensures that no single feature dominates the model solely due to its magnitude.

Why is Scaling Performed?

1. **Algorithm Efficiency**: Some algorithms, such as gradient descent-based methods, converge faster with scaled data because they can move more efficiently through the parameter space.
2. **Feature Comparability**: Scaling ensures that features with larger ranges do not dominate those with smaller ranges, allowing the model to treat all features equally.

3. **Improved Performance**: Scaling can improve the performance of distance-based algorithms (e.g., k-nearest neighbors, k-means clustering) by ensuring that all features contribute equally to the distance computation.
4. **Stability**: Algorithms that involve matrix operations, such as Support Vector Machines (SVMs) and Principal Component Analysis (PCA), can benefit from scaling, as it improves numerical stability and prevents issues related to the varying magnitudes of features.

Difference Between Normalized Scaling and Standardized Scaling

There are two common methods of scaling: normalization and standardization.

Normalization (Min-Max Scaling)

Normalization scales the data to a fixed range, usually [0, 1] or [-1, 1]. This is done by subtracting the minimum value of each feature and dividing by the range (maximum value - minimum value).

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

The Variance Inflation Factor (VIF) measures how much the variance of a regression coefficient is inflated due to multicollinearity in the model. Specifically, VIF quantifies how much a feature is linearly related to the other features in the model.

When VIF Becomes Infinite

A VIF value becomes infinite when there is perfect multicollinearity, meaning that one feature can be perfectly predicted from the others. This situation occurs when:

1. **Exact Linear Dependence**: One or more features are exact linear combinations of other features in the dataset. For example, if feature X3X_3X3 can be expressed as a linear combination of features X1X_1X1 and X2X_2X2 (e.g., X3=2X1+3X2X_3 = 2X_1 + 3X_2X3=2X1+3X2), then the VIF for X3X_3X3 will be infinite.
2. **Duplicated Features**: When features are duplicated or one feature is an exact copy of another, perfect multicollinearity is present. For example, if X4X_4X4 is an exact duplicate of X5X_5X5 (e.g., X4=X5X_4 = X_5X4=X5), both X4X_4X4 and X5X_5X5 will have infinite VIF values.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess if a dataset follows a particular theoretical distribution, most commonly the normal distribution. It compares the quantiles of the sample data with the quantiles of a specified theoretical distribution.

How to Interpret a Q-Q Plot

- **Straight Line**: If the points fall approximately along a straight line (45-degree line), it indicates that the data follows the specified theoretical distribution (e.g., normal distribution).
- **S-shaped Curve**: If the points form an S-shaped curve, it indicates that the data has heavier or lighter tails than the specified distribution.
- **Other Deviations**: Deviations from the straight line suggest departures from the theoretical distribution, such as skewness or kurtosis.