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

- **Season**: The different seasons had a notable effect on bike demand. For instance, spring had a negative coefficient, indicating lower demand during this season than others. In contrast, winter showed a positive coefficient, suggesting higher demand.
- **Weather Situation**: Weather conditions also significantly impacted demand. Poorer weather conditions, such as light snow/rain or mist, were associated with reduced demand, as indicated by their negative coefficients.
- **Year (yr)**: The year variable was highly significant, with a positive coefficient. This suggests that bike-sharing demand increased in 2019 compared to 2018, likely reflecting the service's growing popularity.

2. Why is it important to use drop_first=True during dummy variable creation?
Using `drop_first=True` avoids the **dummy variable trap**. The dummy variable trap occurs when the created dummy variables are perfectly collinear, which can lead to issues in regression models. By dropping the first category, we reduce redundancy and ensure that multicollinearity among dummy variables does not affect the model.

3. Looking at the pair-plot among the numerical variables, which has the highest correlation with the target variable?
Among the numerical variables, the temperature (`atemp`) had the highest correlation with the target variable (`cnt`). This suggests that perceived temperature plays a crucial role in determining the demand for shared bikes, with higher temperatures generally leading to higher demand.

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

- **Linearity**: The assumption was checked by plotting the residuals against the fitted values. A random scatter without a distinct pattern suggested that the relationship between the predictors and the target variable was approximately linear.
- **Normality of Residuals**: The normality of residuals was validated using a Q-Q plot (Quantile-Quantile plot). If the residuals are normally distributed, the points on the plot should ideally fall on the line.
- **Homoscedasticity**: Homoscedasticity was checked by plotting the residuals against the fitted values. The residuals should be evenly spread without any funnel shape or pattern.
- **Multicollinearity**: Variance Inflation Factors (VIF) were calculated to ensure that the predictors were not multicollinear. A VIF value below 5 is typically acceptable.

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

- **Year (yr)**: This variable had the highest positive impact on bike demand, showing that demand significantly increased in 2019 compared to 2018.
- **Perceived Temperature (atemp)**: This was the most significant environmental factor, with warmer perceived temperatures leading to higher bike demand.

- **Spring Season (season_spring)**: This variable was one of the top features, but it had a negative coefficient, indicating that demand was lower during the spring season compared to other seasons.

General Subjective Questions

Explain the linear regression algorithm in detail.
Linear regression is a statistical method for modelling the relationship between a dependent variable (target) and one or more independent variables (predictors). The goal is to find the best-fitting linear equation that predicts the target variable based on the predictors.

Steps in Linear Regression:

1. **Data Collection**: Gather data on the dependent variable and independent variables.
2. **Data Preparation**: Clean and preprocess the data, handle missing values, and encode categorical variables.
3. **Model Fitting**: The linear model is fitted to the data by estimating the coefficients. This is done by minimizing the sum of the squared differences between the observed and predicted values (Ordinary Least Squares—OLS).
4. **Model Evaluation**: Evaluate the model's performance using metrics like R-squared, adjusted R-squared, and analyzing residuals.
5. **Prediction**: Use the fitted model to make predictions on new data.

Explain the Anscombe's quartet in detail.

Anscombe's quartet is a collection of four datasets with nearly identical simple descriptive statistics, yet they have very different distributions and appear very different when graphed. The four datasets have the same:

- Mean of x and y.
- Variance of x and y.
- Correlation between x and y.
- Linear regression line.

However, when you plot these datasets, they reveal very different patterns:

- The first dataset appears as a simple linear relationship.
- The second dataset shows a parabolic relationship.
- The third dataset has a linear relationship but with one outlier.
- The fourth dataset shows a vertical cluster with one outlier.

Importance: Anscombe's quartet demonstrates the importance of graphing data before analyzing it. Relying solely on summary statistics can be misleading.

What is Pearson's R?

Pearson's R, or Pearson correlation coefficient, measures the linear correlation between two variables, X and Y. 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.

What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Scaling refers to the process of adjusting the range of independent variables so that they fit within a specific range or distribution. This is crucial in machine learning because many algorithms are sensitive to the scale of the data.
Why Scaling is Performed:

- To ensure that each feature contributes equally to the model.
- To speed up the convergence of gradient-based algorithms.
- To avoid situations where variables with more extensive ranges dominate the model.

Types of Scaling:

- **Standardized Scaling (Z-score normalization)**: Transforms the data with a mean of 0 and a standard deviation of 1.
- **Normalized Scaling (Min-Max Scaling)**: Rescales the data to a fixed range, usually [0, 1].

Difference:

- **Standardization** is proper when the data follows a normal distribution and is sensitive to outliers.
- **Normalization** is functional when you want to bound your data within a specific range, typically in cases like image processing.

You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Variance Inflation Factor (VIF) measures the extent of multicollinearity in a regression model.
An infinite VIF occurs when one predictor variable is a perfect linear combination of the other predictors. This situation indicates perfect multicollinearity, where the model cannot distinguish between the predictors.

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 for assessing whether a dataset follows a given distribution, typically the normal distribution. In the context of linear regression, it’s used to check whether the residuals (differences between observed and predicted values) are typically distributed.
How it Works:

- The plot compares the quantiles of the residuals with the quantiles of a standard normal distribution.
- The points should lie approximately along the reference line if the residuals are normally distributed.

Importance in Linear Regression:

- Validating the normality of residuals is crucial because one of the assumptions of linear regression is that the residuals should follow a normal distribution. Deviations from normality can indicate model misspecification or the presence of outliers.