Assignment-based Subjective Questions:

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

From the analysis of the categorical variables, we can infer the following effects on the dependent variable (cnt, which represents the total count of bike rentals):

- **Season (season)**:
  The season plays a significant role in influencing bike demand. Typically, the demand is higher during summer and fall seasons due to favorable weather conditions, while it tends to drop in winter and spring. The transition from colder to warmer months sees an increase in bike usage as temperatures become more conducive to outdoor activities. Specifically, when dummy variables are created for season, we observe that summer and fall have a positive effect on bike rentals compared to winter (which is the baseline season when drop_first=True is used). This suggests that the likelihood of bike rentals is higher during these seasons, which might be due to better weather and longer daylight hours.
- **Year (yr)**:
  The yr variable, representing the years 2018 (0) and 2019 (1), indicates an increase in bike-sharing demand over time. This trend reflects the growing popularity and acceptance of bike-sharing programs. The binary nature of this variable allows us to see a clear year-on-year growth in demand. The model usually shows a higher coefficient for the year 2019, suggesting that demand was significantly higher in 2019 compared to 2018. This could be due to increased awareness, improvements in bike-sharing infrastructure, or social trends favoring eco-friendly transportation.
- **Month (mnth)**:
  The month variable captures seasonal trends within the year. Certain months, particularly those in summer (June, July, August), typically show higher bike demand. This can be attributed to the favorable weather and increased outdoor activities during these months. In contrast, colder months like January and February tend to see lower demand, reflecting the reduced willingness of people to use bikes in less favorable weather conditions.
- **Holiday (holiday)**:
  The holiday variable indicates whether a day is a public holiday. Generally, the demand for bikes tends to decrease on holidays, as many people may not commute to work or might use different modes of transportation for leisure activities. The model often shows a negative coefficient for holiday, indicating that bike rentals are lower on these days compared to regular working days.
- **Weekday (weekday)**:
  The day of the week influences bike demand, with weekends (Saturday and Sunday) generally showing different patterns compared to weekdays. On weekdays, bike demand is often higher due to commuting, whereas weekends might see more recreational usage, which could be lower overall. By converting this variable into dummy variables, we can observe specific trends, such as higher demand on certain weekdays (e.g., Friday) and lower on others (e.g., Sunday).
- **Workingday (workingday)**:
  This variable indicates whether a day is a working day (1) or not (0). Bike demand is typically higher on working days due to the need for commuting. The presence of this variable allows the model to differentiate between regular working days and weekends/holidays. The coefficient for workingday is usually positive, suggesting that bike rentals are higher on working days when people commute to work or school.
- **Weather Situation (weathersit)**:
  The weather situation has a strong impact on bike demand. Clear or partly cloudy days (weathersit = 1) are associated with higher bike rentals, while poor weather conditions such as heavy rain, snow, or fog (weathersit = 3 or 4) lead to a significant drop in demand. The model typically shows that as the weather condition worsens, the likelihood of bike rentals decreases, with the worst weather conditions showing the most negative impact.

The categorical variables provide crucial insights into factors that influence bike demand, such as seasonality, holidays, and weather conditions. These variables allow the model to capture patterns in bike rentals that are tied to specific times of the year, work schedules, and environmental conditions.

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

Using drop_first=True during dummy variable creation is important to avoid multicollinearity, a situation where one predictor variable in a model can be linearly predicted from the others with a substantial degree of accuracy.
Using drop_first=True during the creation of dummy variables is crucial to avoid the "dummy variable trap," which can cause multicollinearity in your regression model. Here's why it matters:

- **Dummy Variable Trap**: When you create dummy variables for a categorical feature with n categories, if you include all n dummies in the model, one of the categories can be perfectly predicted from the others. This perfect

multicollinearity causes issues because the regression model cannot distinguish between the categories effectively. For example, if you have a categorical variable for seasons (spring, summer, fall, winter), including all four dummy variables in the model would mean that the value of the fourth season is entirely predictable based on the other three.

- **Preventing Redundancy**: By using drop_first=True, you drop one category and use it as the baseline. The coefficients of the other dummy variables then represent the difference in the outcome variable relative to this baseline category. This avoids redundancy in the model, making it more stable and interpretable.
- **Simplified Interpretation**: With one category dropped, the model's coefficients for the remaining dummy variables indicate how each category differs from the baseline category. This makes the results easier to understand, as each coefficient directly compares its category to the baseline.

In summary, drop_first=True helps prevent multicollinearity, ensures the model is well-specified, and makes the interpretation of results more straightforward.

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

Among the numerical variables, **temperature (temp)** is the one most strongly correlated with bike demand (cnt). This makes intuitive sense:

- **Why Temperature?**: As the temperature rises, more people are likely to engage in outdoor activities, including biking. Warmer temperatures make biking more comfortable and enjoyable, leading to an increase in bike rentals. This positive correlation is visually evident in the pair-plot and is supported by the correlation coefficient between temp and cnt, which is typically high.
- **Impact on Modeling**: Including temperature as a predictor in the model helps capture this essential relationship, allowing the model to better predict bike rentals based on daily weather conditions.

In summary, temperature is a key factor in bike demand, and it's the most closely correlated numerical variable with the total count of bike rentals.

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

After building the linear regression model on the training set, the following steps were taken to validate the key assumptions of linear regression:

- **Linearity**: The assumption of linearity was checked by plotting the residuals (errors) against the fitted values (predicted values). In a well-fitting linear model, the residuals should show no clear pattern when plotted against the fitted values. If the plot reveals a random scatter of points, this indicates that the relationship between the independent variables and the dependent variable is likely linear. However, if there are patterns (e.g., a curve), this suggests non-linearity, indicating that the model may not be appropriate.
- **Normality of Residuals**: To check the normality of residuals, a Q-Q (Quantile-Quantile) plot was used. The Q-Q plot compares the quantiles of the residuals with the quantiles of a standard normal distribution. If the residuals are normally distributed, the points on the Q-Q plot will lie approximately along a straight line. Deviations from this line suggest that the residuals are not normally distributed, which could affect the reliability of hypothesis tests and confidence intervals in the model.
- **Homoscedasticity**: Homoscedasticity (constant variance of residuals) was checked by plotting the residuals against the fitted values. In a homoscedastic model, the spread of the residuals should be roughly constant across all levels of the fitted values. If the plot shows a funnel shape (i.e., the spread increases or decreases with fitted values), this indicates heteroscedasticity, which violates the assumption of constant variance and may lead to inefficient estimates.
- **Multicollinearity**: Multicollinearity among the independent variables was assessed using the Variance Inflation Factor (VIF). VIF quantifies how much the variance of a regression coefficient is inflated due to collinearity with other predictors. A VIF value greater than 10 is often considered a sign of high multicollinearity, suggesting that some variables may need to be removed or combined to improve model stability.

Summary: By validating these assumptions—linearity, normality of residuals, homoscedasticity, and multicollinearity—the reliability and validity of the linear regression model were assessed, ensuring that the model results are trustworthy and interpretable.

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

Based on the final model, the top 3 features contributing significantly to explaining the demand for shared bikes are:

- **Temperature (temp)**:
  Temperature consistently shows a strong positive effect on bike demand. Warmer temperatures are associated with higher bike rentals, as they make biking more comfortable and appealing. In the final model, temp often has one of the highest coefficients, indicating its significant role in driving bike demand. For every unit increase in temperature, the number of bike rentals increases, making temp one of the most influential factors in the model.

- **Humidity (hum):**
  Humidity generally has a negative effect on bike demand. Higher humidity levels can make outdoor activities less pleasant, reducing the number of people choosing to rent bikes. In the final model, hum often shows a significant negative coefficient, indicating its importance. As humidity increases, bike rentals decrease, making hum a critical factor in explaining variations in bike demand.

- **Year (yr):**
  The yr variable, which captures the year (2018 or 2019), typically has a positive coefficient in the model, reflecting the growth in bike-sharing popularity over time. The inclusion of this variable helps account for the overall increase in demand between the two years covered in the dataset. The year 2019 shows a higher demand for bikes compared to 2018, suggesting an increasing trend in the use of bike-sharing services.

The final model identifies temp, hum, and yr as the top three features contributing significantly to bike demand. These variables capture key aspects of weather conditions and temporal trends, which are crucial in predicting bike rentals.

General Subjective Questions:

Explain the linear regression algorithm in detail.

Linear regression is one of the most fundamental and widely used algorithms in statistics and machine learning. It's designed to model the relationship between a dependent variable (the outcome we're interested in) and one or more independent variables (the inputs or features). Here's a breakdown of how it works:

In linear regression, the idea is to find a straight line that best fits the data points in such a way that the differences (errors) between the actual data points and the predicted points on the line are minimized. This line is called the **regression line**.

Equation of the Line: For simple linear regression (one independent variable), the relationship is expressed as:

$$Y=\beta0+\beta1X+\epsilon$$

where:
- Y is the dependent variable (e.g., the number of bikes rented).
- X is the independent variable (e.g., temperature).
- β0 is the intercept (the value of Y when X is 0).
- B1 is the slope (how much Y changes for a one-unit change in X).
- ϵ represents the error term (the difference between the observed and predicted values).

Multiple Linear Regression: When there are multiple independent variables, the equation extends to:

$$Y=\beta0+\beta1X1+\beta2X2+\cdots+\beta nXn+\epsilon$$

Here, each independent variable Xi has its own coefficient βi, which indicates its contribution to predicting Y.

- **Objective**: The primary goal of linear regression is to find the values of β0, β1, ..., βn that minimize the sum of the squared differences between the observed values and the values predicted by the model. This method is known as **Ordinary Least Squares (OLS)**.
- **Assumptions**: For linear regression to be effective, several assumptions must hold:
  - **Linearity**: The relationship between the independent and dependent variables is linear.
  - **Independence**: The observations are independent of each other.
  - **Homoscedasticity**: The variance of the error terms is constant across all levels of the independent variables.
  - **Normality**: The residuals (errors) of the model are normally distributed.
- **Model Evaluation**: After fitting the model, we evaluate its performance using metrics like **R-squared**, which measures the proportion of variance in the dependent variable that is predictable from the independent variables. Other evaluation methods include residual analysis and checking for multicollinearity.

Summary: Linear regression is a straightforward yet powerful technique for modeling the relationship between variables. It's widely used in both research and industry for making predictions and understanding how different factors influence outcomes.

Explain the Anscombe's quartet in detail

Anscombe's quartet is a famous example in statistics that consists of four datasets that have nearly identical simple statistical properties—such as mean, variance, and correlation—but look very different when graphed. Introduced by statistician Francis Anscombe in 1973, the quartet illustrates the importance of data visualization.

- **The Four Datasets**: Despite having the same statistical properties (e.g., the same mean of XXX and YYY, the same correlation coefficient, and the same linear regression line), each dataset tells a different story when plotted:
  - **Dataset 1**: A typical linear relationship, where a straight line fits well.
  - **Dataset 2**: A non-linear relationship where the data forms a curve, which a straight line does not capture.
  - **Dataset 3**: A linear relationship with one influential outlier, which affects the slope of the regression line.
  - **Dataset 4**: A dataset with one vertical line of points and one extreme outlier, where the regression line is influenced heavily by the outlier.
- **Why It's Important:** Anscombe's quartet demonstrates that relying solely on summary statistics can be misleading. For instance, the correlation might suggest a strong relationship, but the actual data might be non-linear or driven by outliers. This example underscores the importance of visualizing data to understand its true nature before making any conclusions.

Summary: Anscombe's quartet is a compelling reminder that data visualization is crucial in statistical analysis. It shows that different datasets can have the same basic statistics but tell very different stories when plotted, emphasizing the need to visualize your data.

What is Pearson's R?

Pearson's R, or the Pearson correlation coefficient, is a measure that quantifies the linear relationship between two continuous variables. It gives us a number between -1 and 1, where:

- **+1**: Indicates a perfect positive linear relationship. As one variable increases, the other increases in a perfectly proportional way.
- **-1**: Indicates a perfect negative linear relationship. As one variable increases, the other decreases in a perfectly proportional way.
- **0**: Indicates no linear relationship between the variables. Changes in one variable do not predict changes in the other.
- **Use in Analysis**: Pearson's R is widely used in statistics to understand the strength and direction of the relationship between two variables. For example, in our bike-sharing data, we might use Pearson's R to measure how strongly temperature is related to bike rentals. A high positive R value would indicate that warmer temperatures are associated with more bike rentals.
- **Interpretation**: The closer the absolute value of Pearson's R is to 1, the stronger the linear relationship. If R is close to 0, it suggests that there's little to no linear relationship between the variables.

Summary: Pearson's R is a powerful tool for measuring the linear relationship between two variables. It helps us understand whether and how strongly two variables are related, making it a key part of data analysis.

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 bring all the features in a dataset to a comparable range, ensuring that no single feature dominates the others due to its scale.

- **Why Scaling is Important**:
  - **Equal Contribution**: In many machine learning models, especially those based on distance metrics (like k-nearest neighbors) or gradient descent (like linear regression), the model's performance can be skewed if the features are on different scales. Scaling ensures that each feature contributes equally to the model.
  - **Improved Convergence**: Scaling helps gradient descent-based algorithms converge faster because the cost function becomes smoother, preventing the algorithm from oscillating and speeding up the path to the minimum.
  - **Reducing Bias**: Features with larger scales can dominate the training process, leading to biased results. Scaling helps to mitigate this issue.
- **Types of Scaling**:
  - **Normalized Scaling (Min-Max Scaling)**:
    - **Definition**: Rescales the data to a fixed range, usually between 0 and 1.
    - **Use Case**: When you want to bring all features into the same range without assuming any distribution. This is particularly useful when features have different units.
    - **Example**: Normalizing image pixel values to fall between 0 and 1.
  - **Standardized Scaling (Z-Score Normalization)**:
    - **Definition**: Rescales the data so that it has a mean of 0 and a standard deviation of 1, essentially converting the data into a standard normal distribution.
    - **Use Case**: When you want to eliminate the effects of different units of measurement and ensure that the features follow a normal distribution.
    - **Example**: Standardizing exam scores to compare results across different tests.

Summary: Scaling is a crucial step in data preprocessing, ensuring that all features contribute equally to the model. Normalized scaling rescales data to a specific range, while standardized scaling adjusts data to have a mean of 0 and a standard deviation of 1, making it especially useful when normality is assumed.

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

The Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in a regression model. It quantifies how much the variance of a regression coefficient is inflated due to multicollinearity with other predictors.

- **Infinite VIF**:
  - **Perfect Multicollinearity**: VIF becomes infinite when there is perfect multicollinearity between variables, meaning one variable is an exact linear combination of others. This situation makes it impossible to estimate the regression coefficients uniquely because the predictors are not independent of each other.
  - **Example**: If you include both temp and atemp in a model, and they are nearly identical (since atemp is derived from temp), the VIF for these variables could become infinite because the model cannot separate their effects.
- **Implications**: An infinite VIF suggests that your model has redundant predictors, making the model unstable and unreliable. This often leads to large, erratic coefficient estimates and poor generalization to new data.

Summary: An infinite VIF indicates perfect multicollinearity, where one predictor can be perfectly predicted from others. This leads to instability in the model, making it difficult to estimate the true effect of each predictor. When VIF is infinite, it's a clear sign that you need to remove or combine the collinear variables to stabilize your model.

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

A Q-Q plot, or Quantile-Quantile plot, is a graphical tool used to compare the distribution of a dataset (often the residuals from a regression model) against a theoretical distribution, typically the normal distribution. It's a valuable diagnostic tool in linear regression.

- **How It Works**:
  - **Plotting Process**: The Q-Q plot orders the data (e.g., residuals) and plots these ordered values against the quantiles of a theoretical distribution. If the data follows the theoretical distribution, the points on the Q-Q plot will lie along a straight line.
  - **Interpretation**:
    - **Straight Line**: If the points closely follow the diagonal line, it indicates that the data is well-approximated by the theoretical distribution, which is often assumed to be normal in the context of linear regression.
    - **Deviations**: Deviations from the line suggest departures from the assumed distribution. For example, if the points curve away from the line, this might indicate skewness, while S-shaped curves could indicate heavy tails or a bimodal distribution.
- **Importance in Linear Regression**:
  - **Normality Check**: One of the key assumptions in linear regression is that the residuals (errors) are normally distributed. The Q-Q plot is an effective way to check this assumption visually.
  - **Model Validation**: If the residuals do not follow a normal distribution, the standard errors, confidence intervals, and hypothesis tests may be invalid, leading to unreliable inferences. The Q-Q plot helps detect such issues early in the modeling process.
- **Practical Example**:
  - In the bike-sharing case study, after fitting a linear regression model, you might plot a Q-Q plot of the residuals to ensure they follow a normal distribution. If the residuals deviate significantly from the straight line, it might indicate that the model isn't capturing all the variability in the data or that some assumptions of linear regression are violated.

Summary: A Q-Q plot is a crucial tool in linear regression, used to assess whether the residuals are normally distributed, which is a key assumption of the model. Ensuring normality through a Q-Q plot helps validate the model, making its predictions and inferences more reliable.