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?**
   Based on the coefficients here's that can be inferred about the effects on the dependent variable:
   1. Summer (season_summer): Has a positive effect (+0.0956) on the dependent variable
   2. Winter (season_winter): Also has a positive effect (+0.1119), with a larger impact than summer
   3. August (mnth_Aug): Slight positive effect (+0.0568)
   4. January (mnth_Jan): Slight negative effect (−0.0461)
   5. October (mnth_Oct): Small positive effect (+0.0394)
   6. September (mnth_Sep): Positive effect (+0.1162), indicating higher impact compared to other months
   7. Cloudy (weathersit_Cloudy): Negative effect (−0.0809)
   8. Rainy (weathersit_Rainy): Stronger negative effect (−0.2867)
   9. Windspeed: Has a negative effect (−0.1620) on the dependent variable, indicating that as windspeed increases, the dependent variable decreases.

2. **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, which occurs when dummy variables are perfectly correlated with each other. By dropping the first category, you prevent redundancy, as the dropped category's effect is implicitly captured by the remaining dummy variables. This ensures that the regression model has a proper baseline category and avoids issues with inflated variance in the coefficient estimates, leading to more stable and interpretable results.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   The numerical variables temp and atemp have the strongest correlation with the cnt target variable. They are also strongly associated with each other, implying a significant similarity.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
    - Linear Relationship
    - Homoscedasticity
    - Absence of Multicollinearity
    - Independence of residuals (absence of auto-correlation)
    - Residuals are normally distributed

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   The top 3 features are temp, yr and weathersit_Rainy.
    1. temp: With a coefficient of +0.4828, temperature has the most substantial positive impact on bike demand.
    2. yr: The year variable has a positive coefficient of +0.2341, indicating a notable increase in demand over time.
    3. weathersit_Rainy: This has a significant negative impact, with a coefficient of −0.2867, indicating that rainy weather strongly reduces bike demand.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**
   Linear Regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. The goal is to find the line (or hyperplane in higher dimensions) that best predicts the dependent variable from the independent variables, minimizing the sum of squared differences between the observed and predicted values. This is achieved through Ordinary Least Squares (OLS), which estimates the model's coefficients to achieve the best fit. The model's performance is typically evaluated using metrics such as R-squared and Mean Squared Error (MSE), with assumptions including linearity, independence of residuals, and constant variance.

2. **Explain the Anscombe’s quartet in detail.**
   Anscombe’s Quartet is a set of four datasets created by statistician Francis Anscombe to demonstrate the importance of graphical analysis in understanding data. Each dataset

contains the same statistical properties, such as mean, variance, and correlation, yet they exhibit very different distributions and relationships when plotted. This quartet illustrates that summary statistics alone can be misleading, highlighting the necessity of visualizing data through plots to reveal underlying patterns and anomalies. By showcasing how different datasets with identical descriptive statistics can have diverse graphical representations, Anscombe's Quartet underscores the critical role of exploratory data analysis in statistical practice.

3. **What is Pearson's R?**

   Pearson's R, or Pearson correlation coefficient, measures the strength and direction of the linear relationship between two continuous variables. Ranges from -1 to 1, where 1 indicates a perfect positive linear relationship, -1 signifies a perfect negative linear relationship, and 0 implies no linear relationship. It quantifies how well the data points fit a straight line, with positive values indicating a direct relationship and negative values indicating an inverse relationship. Pearson's R is widely used in statistics to assess the degree of linear association between variables.

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

   1. Scaling is a data preprocessing technique used to adjust the range and distribution of feature values so that they are on a similar scale. This is essential for many machine learning algorithms, which may perform poorly or converge slowly if features have different scales. By scaling features, you ensure that each one contributes equally to the model, improving accuracy and efficiency. Common methods include normalization, which rescales values to a specific range, and standardization, which adjusts values to have a mean of 0 and a standard deviation of 1.

   2. Scaling is performed to ensure that all features in a dataset contribute equally to the model, as differences in feature ranges can disproportionately affect the performance of machine learning algorithms. Many algorithms, such as those using distance metrics (e.g., k-nearest neighbors) or optimization techniques (e.g., gradient descent), are sensitive to the scale of input data. Scaling improves the accuracy and efficiency of these algorithms by normalizing the range of features, which helps in achieving faster convergence and more reliable results.

3. Normalized scaling (min-max scaling) transforms feature values to a specific range, typically [0, 1], by adjusting the data based on its minimum and maximum values. This ensures that all features have a bounded range. Standardized scaling (z-score normalization) adjusts feature values to have a mean of 0 and a standard deviation of 1, using the feature's mean and standard deviation. While normalization is useful when a bounded range is required, standardization is preferred for data that needs to be centered and scaled based on its statistical properties, especially when features are on different scales or follow different distributions.

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

   A VIF value can become infinite when there is perfect multicollinearity among the features, meaning one or more features are perfectly linearly dependent on others. This occurs when a feature can be exactly predicted from a linear combination of other features, leading to an undefined or infinitely large VIF. In such cases, the matrix used to compute VIF becomes singular, causing its inverse to be non-existent or infinitely large. This indicates that the feature is redundant and should be removed to resolve the multicollinearity issue.

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

   1. A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare the distribution of a dataset against a theoretical distribution, such as the normal distribution, to assess how well the data fits that distribution. It plots the quantiles of the dataset against the quantiles of the theoretical distribution. If the points lie approximately along a straight line, it suggests that the data follows the theoretical distribution closely. Deviations from this line indicate departures from the assumed distribution, helping to identify issues such as skewness or kurtosis in the data.

   2. In linear regression, a Q-Q (Quantile-Quantile) plot is used to assess whether the residuals are normally distributed, which is a key assumption for many statistical tests and confidence intervals in regression analysis. By comparing the quantiles of the residuals against the quantiles of a normal distribution, the Q-Q plot helps validate the model's assumptions and identify deviations from normality, such as skewness or

outliers. Ensuring normality of residuals through a Q-Q plot enhances the reliability of the model's statistical inferences and overall robustness.