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?

Categorical variables can be inferred to have the following effects on the dependent variables.

Poor weather conditions negatively impact bike rental demand. The 'weathersit_2' dummy variable indicates a 0.2394 unit decrease in rentals during misty or cloudy weather, while 'weathersit_3' leads to a significant 1.0738 unit decrease during light rain or snow. This highlights the importance of weather in determining bike rental demand.

Bike rentals vary with the seasons, as positive coefficients of 'season' dummy variables show. Rentals increase during 'season_2' (summer), 'season_3' (fall), and 'season_4' (winter), with the most significant increase during winter. This suggests that people rent more bikes in later seasons due to favorable weather and outdoor conditions.

Bike rental demand is affected by weather and seasons. Companies can use this info to plan operations and marketing strategies, capitalizing on high demand during good weather and minimizing low demand during less favorable times.

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

When creating dummy variables for a categorical feature with 'n' levels, 'n' new features are generated, each representing a level of the original feature. This means one variable can be accurately predicted from others. However, including all 'n' variables in the model could lead to multicollinearity when highly correlated predictor variables.

More dummy variables can make interpreting the model's estimates challenging and cause imprecise coefficient estimates. To avoid this, we use **drop_first=True**, which removes the first dummy variable, reducing the number from 'n' to 'n-1'. This eliminates perfect correlation and multicollinearity, making our results more accurate.

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

According to the analysis, the 'atemp' feature, which represents the feeling temperature, strongly correlates with the target variable 'cnt'. The correlation value is around 0.630685. This means a positive correlation exists between 'atemp' and 'cnt', indicating that the total number of rented bikes ('cnt') also tends to increase as the feeling temperature increases.

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

Two assumptions of Linear Regression were checked:

1. **No Multicollinearity**: Multicollinearity happens when independent variables are strongly correlated with each other. To check this, we used the Variance Inflation Factor (VIF). If a variable has a high VIF (usually > 5), it means it's highly correlated with other variables. In the code we provided, we removed variables with high VIF to avoid multicollinearity.
2. **Normality of Error Terms**: To ensure the error terms are normally distributed, we check by plotting a histogram of the residuals. If the histogram takes the shape of a bell curve, it confirms that the assumption is met.

The other main assumptions of linear regression, such as Linearity, Independence of error terms, and Homoscedasticity, were not checked.

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

The top three features that significantly explain the demand for shared bikes are 'weathersit_3', 'yr', and 'season_4'.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is a crucial algorithm in machine learning and statistics utilized to forecast a continuous outcome variable (known as the dependent variable) based on one or more predictor variables (known as the independent variables). The algorithm assumes a linear correlation between the dependent and independent variables, which is why it is referred to as "linear" regression.

2. Explain the Anscombe’s quartet in detail.

Anscombe’s quartet is a set of four datasets that have nearly identical statistical properties, but when graphed, they look very different. The statistician Francis Anscombe created the quartet in 1973 to demonstrate the importance of graphing data before analyzing it and to show how misleading summary statistics can be.

3. What is Pearson’s R?

Pearson's R, also called Pearson's correlation coefficient, is a statistical tool that measures the intensity and direction of a linear connection between two continuous variables. The correlation coefficient is represented by the letter 'r' and can vary from -1 to +1.

It's important to note that Pearson's correlation coefficient can only measure linear relationships between two variables. If there is no linear relationship, a correlation of zero does not necessarily mean that there is no relationship at all.
To calculate Pearson's correlation coefficient, the covariance of the two variables is standardized by dividing it by the product of their standard deviations.

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

Scaling is an important data preprocessing step that involves adjusting the range of values for numeric variables in a dataset. The goal is to bring all variables to a similar scale. It's essential to ensure that no particular variable dominates others, which can happen if variables have different units or vastly different ranges.

Scaling is particularly crucial in many machine learning algorithms, including distance measures like K-nearest neighbors (K-NN) and K-means clustering, and optimization algorithms like gradient descent. These algorithms can only function effectively if features are on relatively similar scales.

There are several methods of scaling data, but two of the most common ones are normalization and standardization.

1. **Normalization (Min-Max Scaling)**
2. **Standardization (Z-score Normalization)**

In summary, the choice between normalization and standardization depends on the algorithm used, the distribution of the data, and whether or not outliers are important to the analysis.

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

The Variance Inflation Factor (VIF) is used to measure the extent of multicollinearity among multiple regression variables. It is a mathematical calculation that determines the ratio of the overall model variance to that of a model that includes only a single independent variable. This calculation is performed for each independent variable, and a high VIF indicates a strong correlation with other variables, meaning that the others highly influence the variable.

Ideally, the VIF value should be as close to 1.0 as possible, indicating no correlation with other predictors. However, a VIF value exceeding 5 or 10 is typically considered high and suggests a high level of multicollinearity between the independent variable and the others.

Perfect multicollinearity occurs when a variable can be predicted linearly from others, causing the VIF to become undefined or infinite. To avoid this, eliminate one of the highly correlated variables.