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?

Analyzing categorical variables is crucial to understand their impact on the dependent variable, which in this case is likely to be the count of bike rentals ('cnt'). Here's what can be inferred about the effect of categorical variables based on the analysis:

Season: The 'season' variable, even though numerically encoded, doesn't necessarily imply an order. Instead, it represents different seasons (1:spring, 2:summer, 3:fall, 4:winter). Its effect on bike rentals might vary based on weather preferences during different seasons. For instance, rentals might be higher in summer and fall due to more favorable weather conditions.

Month: The 'mnth' variable, representing months from 1 to 12, might have a varying effect on bike rentals. Seasonal trends or events specific to each month could influence rental demand. For instance, summer months or months coinciding with holidays might witness higher bike rentals.

Holiday: The 'holiday' variable (0:non-holiday, 1:holiday) could significantly impact bike rentals. Holidays might lead to more leisure activities, increasing bike usage, while non-holiday periods might have more regular commuting patterns.

Weekday: Days of the week ('weekday') can have distinct effects on rental patterns. Weekdays might see higher rentals during commuting hours, while weekends (Saturday and Sunday) could have increased leisure bike usage.

Working Day: The 'workingday' variable (0:non-working day, 1:working day) might indicate varying rental patterns based on work schedules. Working days might have more structured usage, focusing on commuting, while non-working days might witness more recreational bike usage.

Year: The 'yr' variable (0:2018, 1:2019) suggests an increasing trend in bike rentals over the two years. This variable could capture the overall growth or popularity of bike-sharing services year on year.

Inference: Each categorical variable represents distinct aspects influencing bike rentals. Factors such as weather, seasonality, holidays, days of the week, and the year itself play significant roles in determining the demand for shared bikes. Understanding these categorical variables can aid in creating a predictive model that incorporates these influences to better estimate bike rental demand.

2 WHY IS IT IMPORTANT TO USE DROP_FIRST=TRUE DURING DUMMY VARIABLE CREATION?

Response: In the context of creating dummy variables from categorical data, setting **drop_first=True** is a strategy used to prevent multicollinearity and reduce redundancy in the dataset when creating dummy variables.

When creating dummy variables for categorical features, if you have **N** categories, typically, you create **N-1** dummy variables. This is because if you create dummies for all categories and include them all in the model, it introduces perfect multicollinearity, which can lead to issues in regression analysis. By setting **drop_first=True**, it automatically drops the first category during dummy creation. The dropped category becomes the reference category, and the model includes dummy variables only for the remaining categories. This approach helps to mitigate the multicollinearity issue.

Ultimately, using **drop_first=True** in dummy variable creation helps in improving the model's performance, reducing computational complexity, and making the interpretation of the model coefficients more straightforward by avoiding multicollinearity among the predictor variables.

3 LOOKING AT THE PAIR-PLOT AMONG THE NUMERICAL VARIABLES, WHICH ONE HAS THE HIGHEST CORRELATION WITH THE TARGET VARIABLE?

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

Among the numerical variables ('temp', 'atemp', 'hum', 'windspeed', 'casual', 'registered') plotted against the target variable ('cnt'), 'registered' tends to have the highest correlation. The scatterplot of 'registered' against 'cnt' shows a stronger linear relationship compared to the other variables, indicating a relatively higher correlation.

4 BASED ON THE FINAL MODEL, WHICH ARE THE TOP 3 FEATURES CONTRIBUTING SIGNIFICANTLY TOWARDS EXPLAINING THE DEMAND OF THE SHARED BIKES?

Identifying the top significant features contributing to the demand for shared bikes can be determined by examining the coefficients or importance scores from the final model. In a linear regression model, the magnitude of coefficients indicates the impact of each feature on the target variable. In other models like Random Forest or Gradient Boosting, feature importance scores are used.

The top three features contributing significantly to explaining the demand for shared bikes are determined by their highest coefficients (in linear regression) or highest importance scores (in other models) in the final model. These features include variables like 'registered users,' 'temperature,' 'season,' 'humidity,' 'weathersit,' or 'windspeed.' The three features contributing significantly towards explaining the demand for shared bikes will be the ones with the highest coefficients (in linear regression) or the highest importance scores (in other models) in the final model.