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?

The analysis of the categorical variables from the dataset infer as below about their effect on the dependent variable:

1. **Clear Weather (Binary Categorical Variable):**
    - The variable 'clear weather' is a binary variable representing clear (1) or cloudy (0) weather.
    - Inference: A likely inference is that clear weather may positively impact the daily wage ('dwage') compared to cloudy weather. Clear weather conditions could lead to increased productivity or demand for certain types of work.
2. **Handling Categorical Variables in Analysis:**
    - The encoding process may involve assigning numerical values to categorical variables, allowing them to be incorporated into mathematical models for predicting 'dwage'.
3. **Predicted Wage (Numeric Variable):**
    - The variable 'predicted wage' is mentioned as a numerical variable calculated by the model.
    - Inference: 'Predicted wage' could serve as an important variable in understanding the model's predictions for 'dwage' based on the independent variables.

Additionally, it reveals a positive correlation between clear weather and heightened bike demand, characterized by increased variability and more extreme cases.

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

The parameter **drop_first=True** in the context of creating dummy variables is associated with the concept of multicollinearity, and its importance can be explained as follows:

1. **Multicollinearity Mitigation:** The parameter helps to mitigate multicollinearity, which occurs when two or more independent variables in a regression model are highly correlated. By dropping one level of the categorical variable, we reduce the risk of multicollinearity issues in the model.
2. **Simplicity and Efficiency:** It contributes to model simplicity by avoiding redundancy and unnecessary complexity introduced by including all levels of a categorical variable as dummy variables. This not only enhances the interpretability of the model but also improves computational efficiency, particularly when dealing with many categorical levels.

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

The windspeed has the highest correlation with the target variable.

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

In validating the assumptions of linear regression after constructing the model on the training set, several key steps were undertaken.

Assignment-based Subjective Questions

- The linearity assumption was addressed by reviewing scatter plots of each independent variable against the dependent variable, ensuring that the observed relationships appeared linear.
- The independence of residuals was validated through the examination of the residuals plot, verifying the absence of systematic patterns.
- Homoscedasticity was assessed by plotting residuals against predicted values, confirming that the spread of residuals exhibited a consistent variance across all levels of the predicted variable.
- Normality of residuals was checked using both histogram and Q-Q plot visualizations, supplemented by statistical tests such as the Shapiro-Wilk test.
- Multicollinearity was scrutinized by calculating the variance inflation factor (VIF) for each independent variable, identifying and mitigating high VIF values indicative of potential multicollinearity.

Additionally, the absence of perfect multicollinearity was ensured by reviewing correlations between independent variables through correlation matrices or heatmaps. The validation process incorporated both graphical methods and statistical tests to comprehensively assess the adherence of the model to key linear regression assumptions, facilitating a robust evaluation of its reliability and predictive performance.

5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes?

According to the final model, the top three features that significantly contribute to explaining the demand for shared bikes are as follows:

1. **Year (yr_1):** The coefficient for the 'yr_1' feature is approximately 1.20e+12, indicating a substantial positive impact on bike demand. This suggests that being in the year '1' (presumably representing a specific year) significantly contributes to increased bike demand.

2. **Month:** The 'month' variable has a coefficient of approximately 9.83e+08, implying a considerable positive influence on bike demand. This suggests that specific months contribute significantly to variations in bike demand.

3. **Season (season_4):** The 'season_4' feature, representing the fourth season, has a coefficient of approximately 1.61e+03. This suggests a positive impact on bike demand during this particular season, contributing significantly to the overall demand for shared bikes.

These coefficients provide insights into the relative importance of each feature in explaining the variation in bike demand, with higher coefficients indicating a stronger influence.