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

- Season : Season 1 has the lowest median, whereas season 2 and 3 are the most popular ones. They play a huge part in the total counts of bikes used.
- Weathersit : Weather 1 has the highest median, followed by 2 and 3. They play a huge part in the total counts of bikes used.
- Months : Months between 5-10, i.e. May-Oct are the most popular ones. They play a huge part in the total counts of bikes used.
- Holiday : Even though medians differ but the max values remain the same.
- Weekday : All medians remain the same.
- workingday : Seems like it has no effect on total counts of bikes used.

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

Because it is possible to have a dummy variable without the first column we often prefer to drop it, to reduce the number of correlations produced using dummy variables.

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

Temperature has the highest correlation with total counts of bikes.

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

After building a model on a train set, we check our assumptions : (i) If our residual errors are normally distributed, (ii) if our mean of residual errors is zero. These two assumptions were correct for our current model, thus we can confidently say that Linear Regression can be performed.

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

Temperature, Season (Summer and Winter), Weather (Positive: Clear, Few clouds, Partly cloudy, Partly cloudy ; Negative: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)