Assignment-based Outcomes

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

➔ The variables yr has maximum correlation with the target variable cnt. We have seen that the month Aug and Sept has also shown an increase in cnt count with holiday and rain having a negative coefficient

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

➔ If you have a small number of dummies, i suggest removing the first dummy. For example, if you have a variable gender, you don't need both a male and female dummy. Just one will be fine. If male then the person is a male and if male = 0 then the person is female. However if you have a category with hundreds of values, not dropping the first column. That will make it easier for the model to "see" all the categories quickly during learning (and the adverse effects are negligible).

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

➔ atemp has the highest correlation followed by yr column

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

➔ By checking if the **error terms** are also normally distributed and by plotting y_test and y_pred to understand the spread and calculating the r2_score

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

➔ Yr and atemp explains the demand of the shared bikes significantly along with rain with negative coefficient