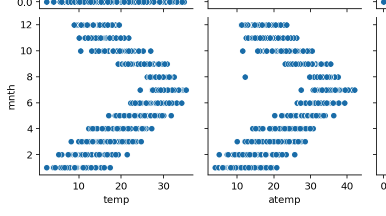
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

**Answer**: From the pairplot, the have the following findings:

The dependent variables “temp” and “atem” changes according to the categorical variable”mnth” with a clear patter.

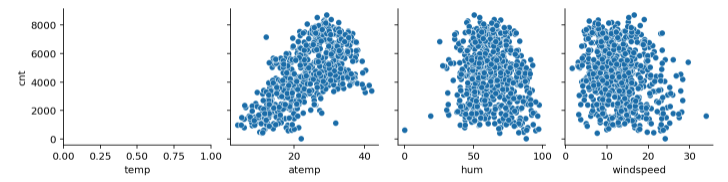


2. Why is it important to use drop\_first=True during dummy variable creation? (2 mark)

**Answer**: Because it can reduce the extra columns which is created when the dummy variable is created. The consequence is that it can help to reduce the correlations among the dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

**Answer**: the “atemp” 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? (3 marks)

**Answer**: I use a test set to validate the model of LR. The test set is complete not overlapped with the training set. Thus if the assumption is true, the model should be working well over the test set.

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

**Answer**: I analyze the coefficients of the LR model and find the coefficients of the following 3 variables has the largest absolute values of coefficients:

* If the times is a spring season
* If the time is a winter season
* If the day is a Monday

Thus I think these 3 features are the top 3 features contributing significantly towards explaining the demand of the shared bikes.