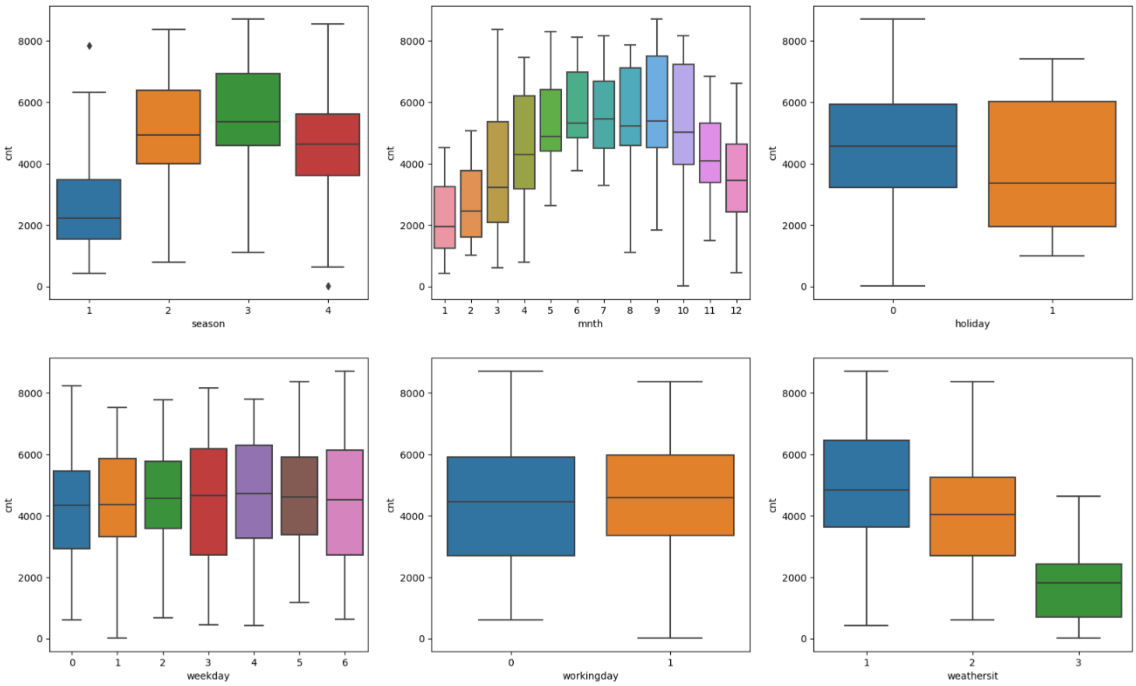
Assignment-based Subjective Questions

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

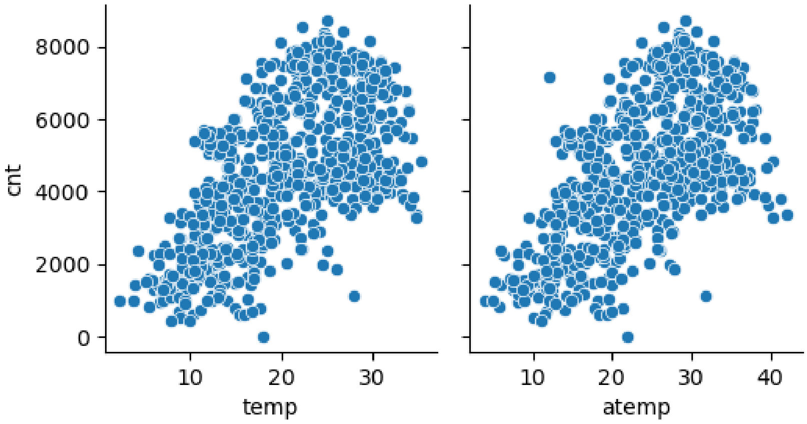


- The weather situation value is never 4, It can indicate either missing data or the weather was never type 4 (Heavy Rain + Ice Pellets + Thunderstorm + Mist, Snow + Fog) in the recorded two years dataset.
- Fall seems to have the highest rentals across the season. Which also correlates to the corresponding months having higher rentals
- There are more rentals recorded when the weather is clear.
- Holidays have slightly lower avg rentals than non holidays. Which shows similar pattern in working days having slightly higher rentals than that of non working days.
- The avg rentals are more or less similar across the days of the week.

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

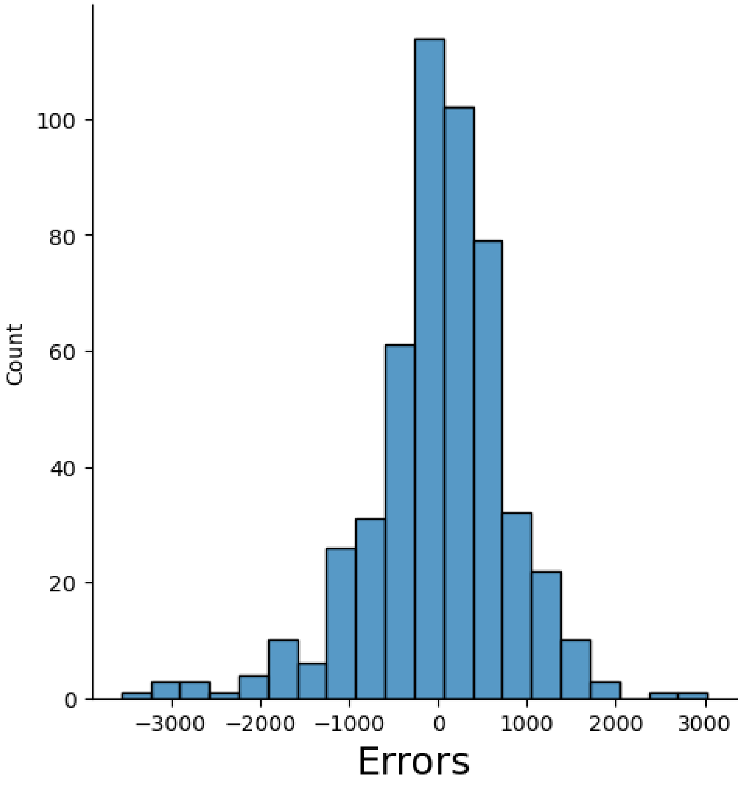
If we don't use drop_first we get the number of columns the same as the possible values for the categorical column. Which can be reduced to n-1 by dropping the first column, When all the other columns are `0` it can be inferred as the dropped column. Model would still be able to learn the effect of that column this way.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Temperature variables seems to have a good correlation with the target variable.



On checking the correlation values within the training data, `***atemp***` seems to have slightly higher as 0.65.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)



As part of residual analysis, I plotted the errors resulting in a normal distribution as expected for the linear regression model.
A scatter plot of residuals can’t be made because there are multiple variables in the X_train.

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

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

2. Explain the Anscombe's quartet in detail. (3 marks)

3. What is Pearson's R? (3 marks)

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
(3 marks)

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(3 marks)