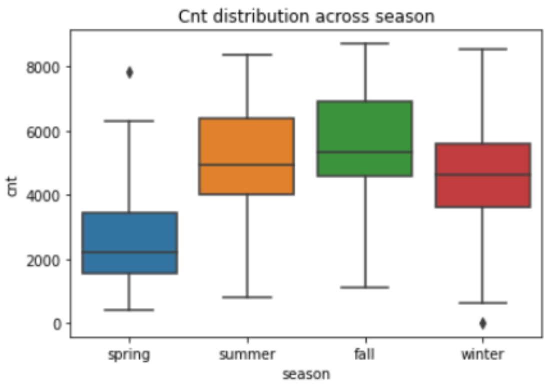
Assignment-based Subjective Questions

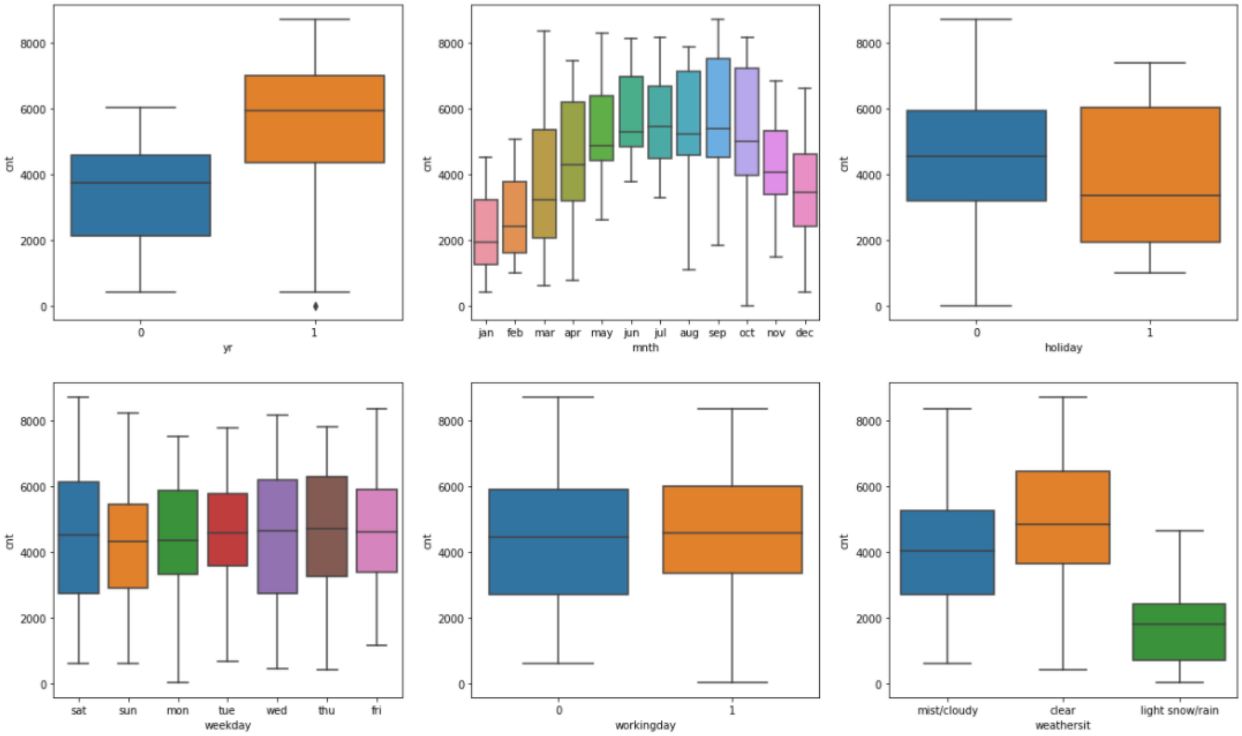
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans: Firstly, here are all the categorical variables: 'yr','holiday','workingday', 'season','holiday', 'mnth', 'weekday', 'weathersit'.

Based on the EDA of the data set and the individual variables, we could see that only the following variables have significant impact.

1. Season: spring has a significant drop in 'cnt' as compared to the other seasons
2. Month: most rentals are during the months in the summer and fall season
3. Holiday: the rentals decrease on Holidays
4. Year: there’s a year over year growth in the rentals
5. Weathersit: As expected the rentals are much higher on clear days as compared to cloudy days, with hardly any rentals when it's snowing or raining.
6. Weekday: As for the weekly patterns, there's a very slight difference on each consecutive day of the week with sunday being the lowest and saturday reaching the peak.



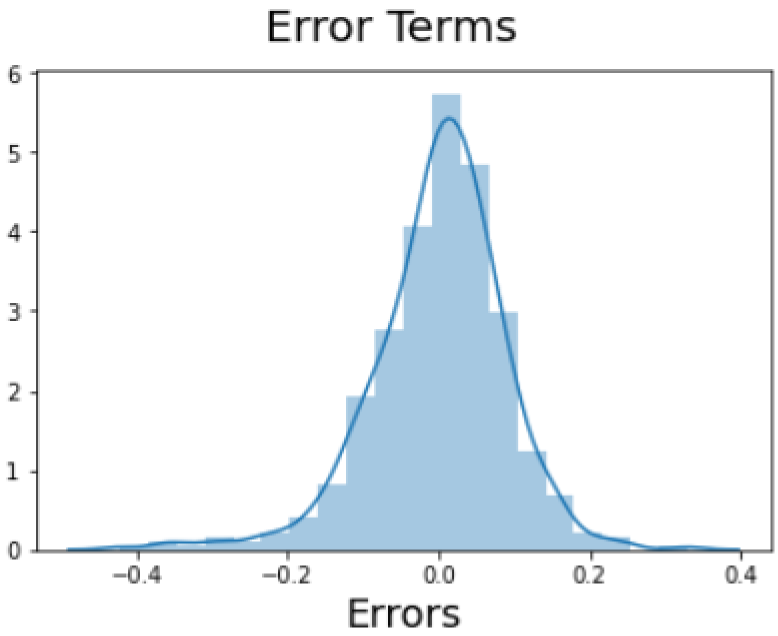


2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
Ans: Because the first dummy can be explained as the linear combination of the other dummies

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

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
Ans: The four assumptions of linear regression are:

- Linear relationship: There exists a linear relationship between the independent variable and the dependent variable. We checked this by performing EDA and looking at the correlation of variables.
- Independence: The residuals are independent. We performed residual analysis of the train data.
- Homoscedasticity: The residuals have constant variance at every level of x. We performed residual analysis of the train data.
- Normality: The residuals of the model are normally distributed. We performed residual analysis of the train data.



-

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
Ans: Temperature, year and winter

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)