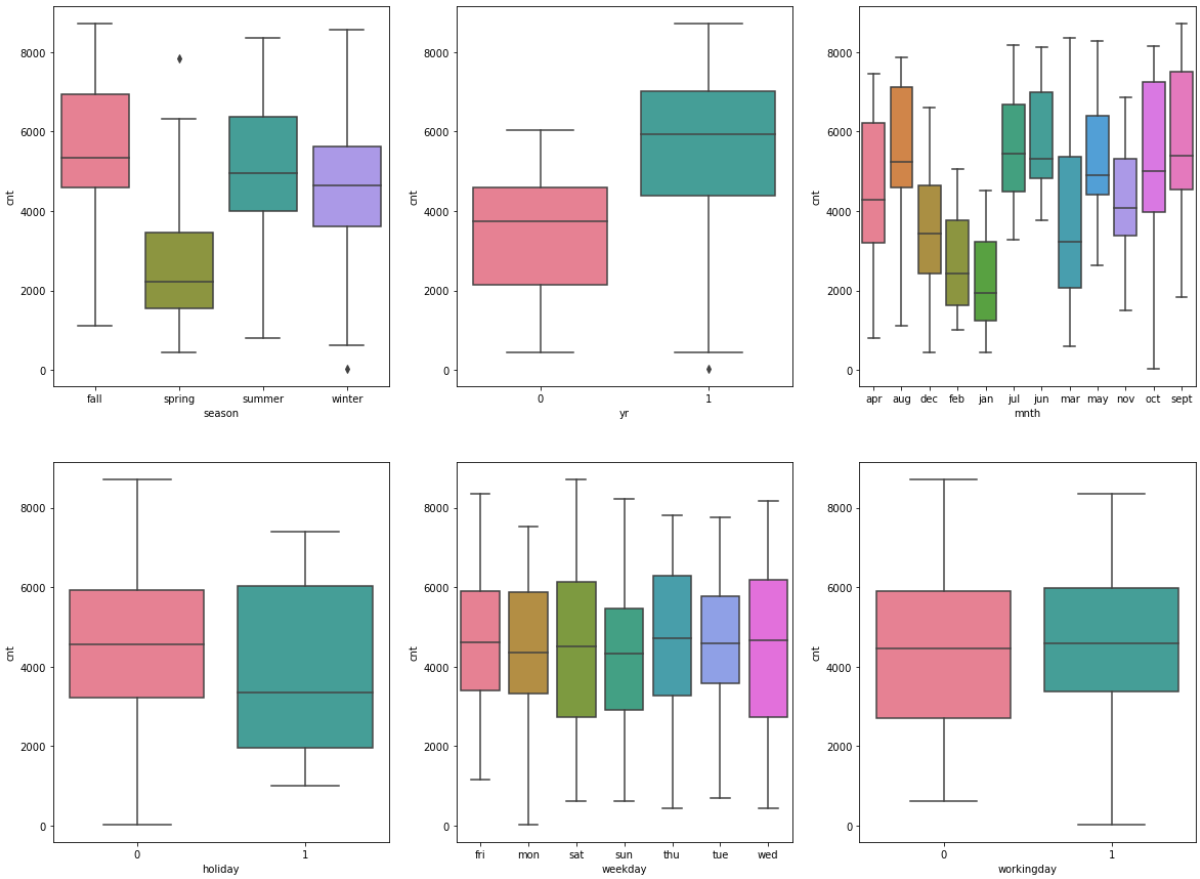
Assignment-based Subjective Questions

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

Answer:

The categorical variables are : 'mnth', 'weekday', 'season' , 'weathersit' ,'yr' and 'holiday'. Based on the boxplot analysis we can infer following:



• **Yr:** Bike bookings are higher in 2019 as compared to 2018, it might be due to the fact bike rentals are getting popular and people are becoming more aware about environment.

• **season:** Highest booking happening in season3(fall) with a median of over 5000 booking. This was followed by season2(summer) & season4(winter) of total booking.

• **mnth:** Bike booking is quite high in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.

• **weathersit:** Almost 67% of the bike booking were happening during 'weathersit1 with a median of close to 5000 booking. This was followed by weathersit2. Clear weather is most optimal for bike renting.

• **holiday:** The bike booking were happening mostly when it is not a holiday.

• **weekday:** weekday variable shows very close trend. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.

• **workingday:** Median is quite close, does not have much impact.

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

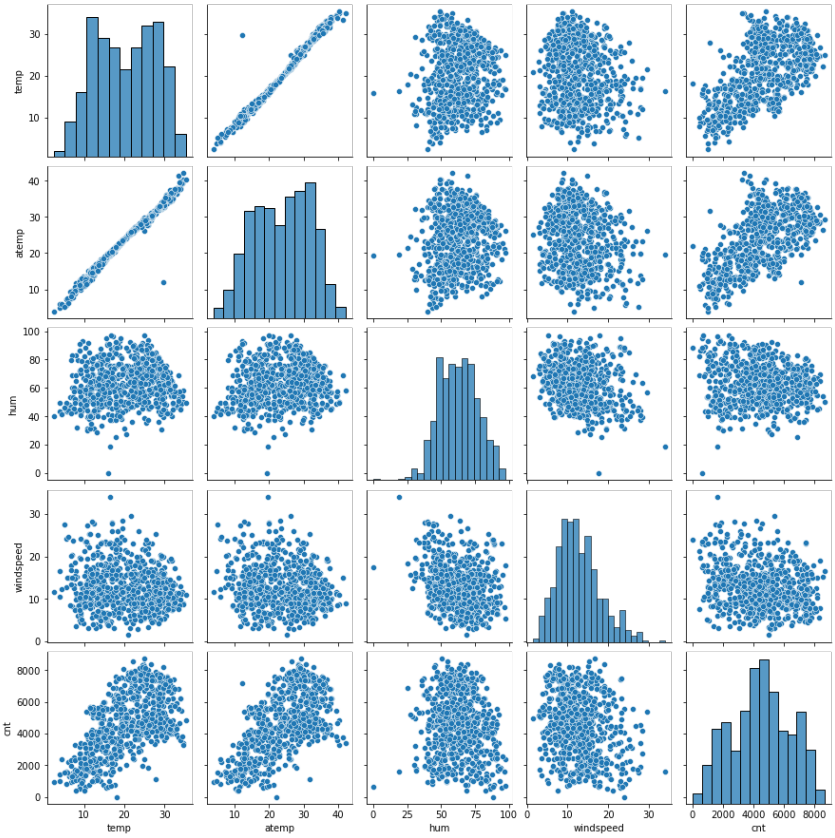
Answer:

Reasons for dropping first dummy variable :
1. To reduce multicollinearity : If we do not drop the first column during dummy variable creation , it will lead to high multicollinearity between the dummy variables and will adversely affect the model.
2. To reduce redundancy. For eg for a variable gender , both male and female dummy are not required.
Male=0 will be a female.However , sometimes it depends on the number of values in a categorical variables. For a categorical variable with large number of values, the drop_first could be avoided to see the effect of all the values of the variable.For eg a categorical variable=Month , here all values Jan-Dec should be created as dummy column

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

Answer:

Temp[temperature] has the highest correlation with the target variable

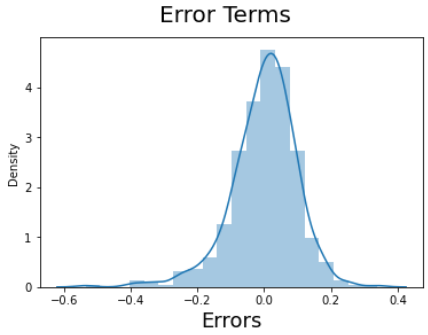


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer:

a. Error terms are normally distributed with mean zero(not X,Y)

i.



Residual distribution should follow normal distribution and centred around 0. We confirm this by plotting a distplot of the residuals.

b. Independence of residuals can be calculated via the Durbin-Watson value of the model

c. The Linear distribution of the independent variables(temp,hum,windspeed) can be scatter plot with the target variable(cnt)

d. Validating Homoscedasticity i.e the residuals should not follow a pattern of distribution.

i.