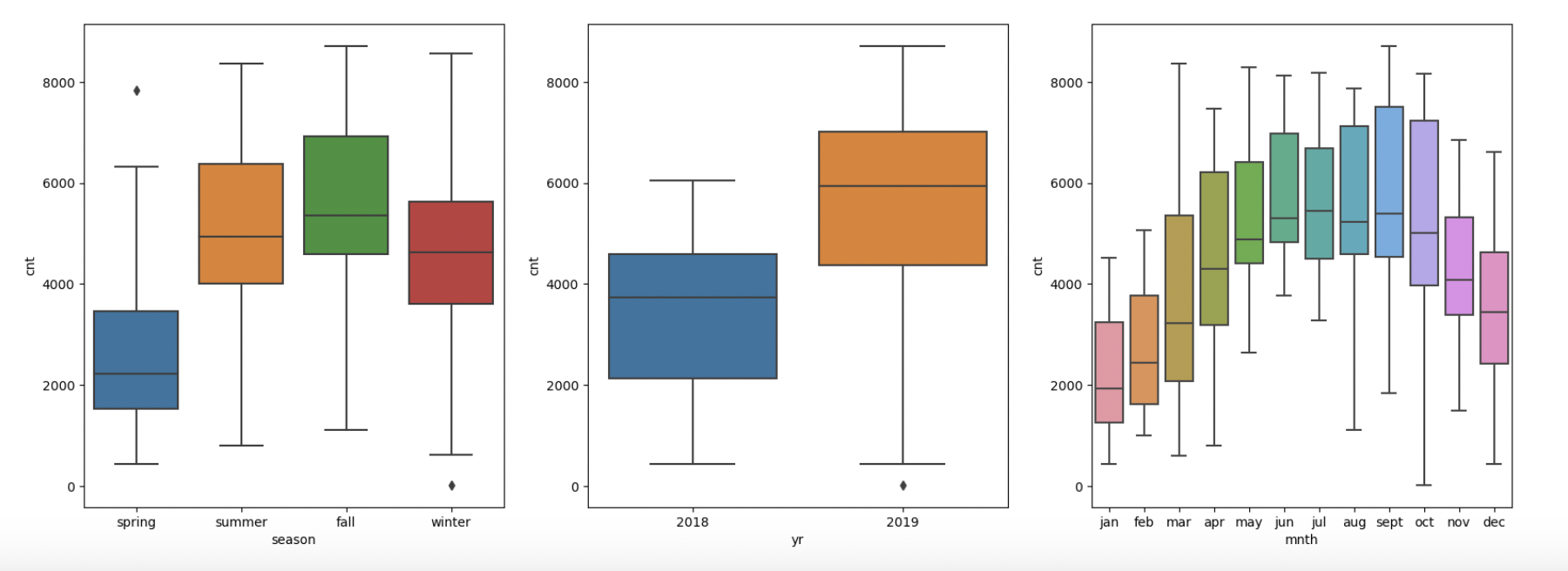
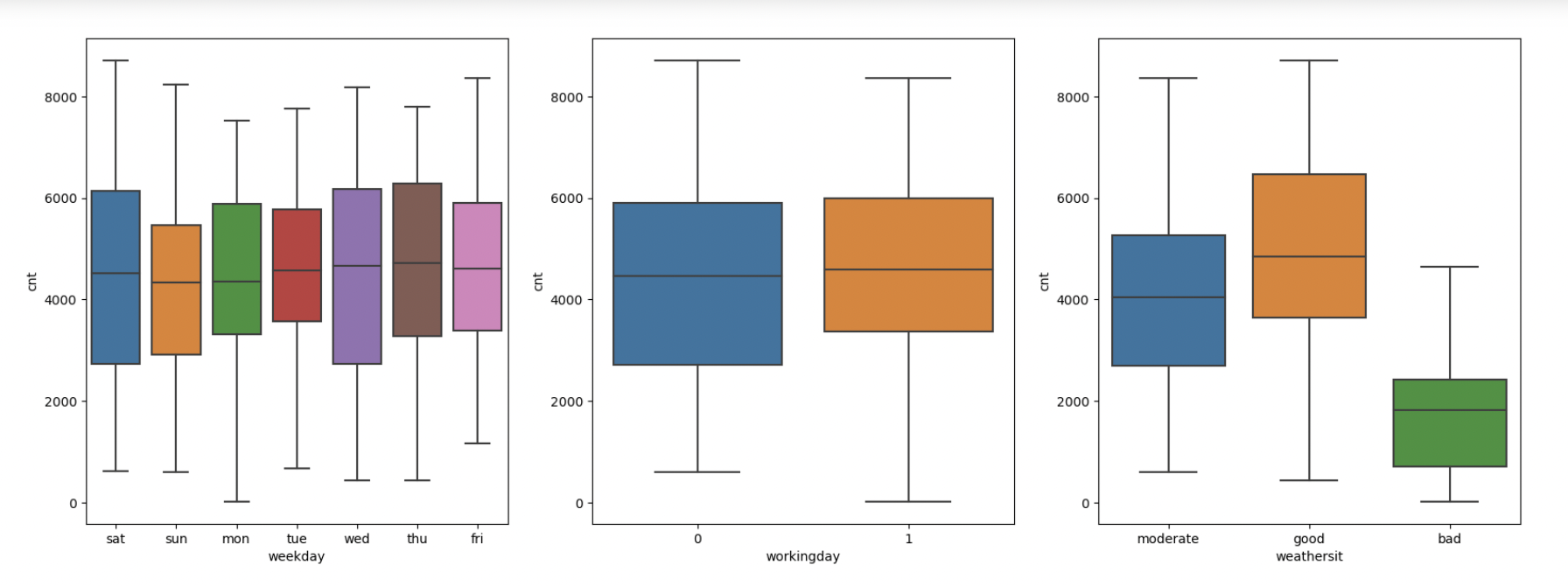
Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

ANSWER:





Observations:

1. SEASON AND CNT: Seasons have a significant effect on the target variable Cnt.

Cnt is highest in fall and lowest in spring

1. YEAR AND CNT: Cnt is more in 2019 than in 2018. It may be inferred that the demand is increasing with time.
2. MONTH AND CNT: Cnt is highest in sept, oct and lowest in jan, feb , march
3. WEEKDAYS AND CNT: No significant pattern observed.
4. WORKINGDAY AND CNT: More variation on non-working day but the median and highest demand is almost same.
5. WEATHERSIT AND CNT: CNT high on good-weathersit and low on bad-weathersit.

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

ANSWER: *The intention behind the dummy variable is that for a categorical variable with ‘n’ levels, we*

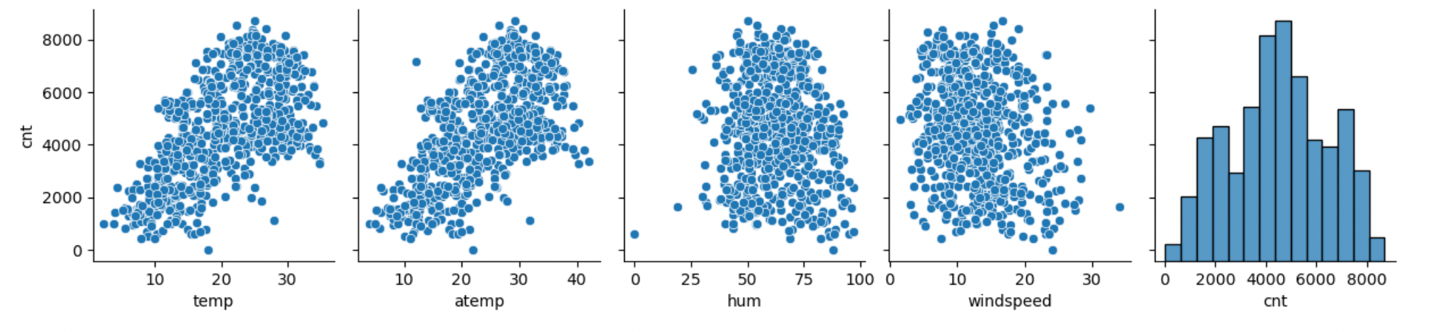
*create ‘n-1’ new columns each indicating whether that level exists or not using a zero or one.*

*Hence drop\_first=True is used so that the resultant can match up n-1 levels.*

*Eg: If there are 3 levels, the drop\_first will drop the first column.*

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

ANSWER: temp and atemp have highest correlation with cnt



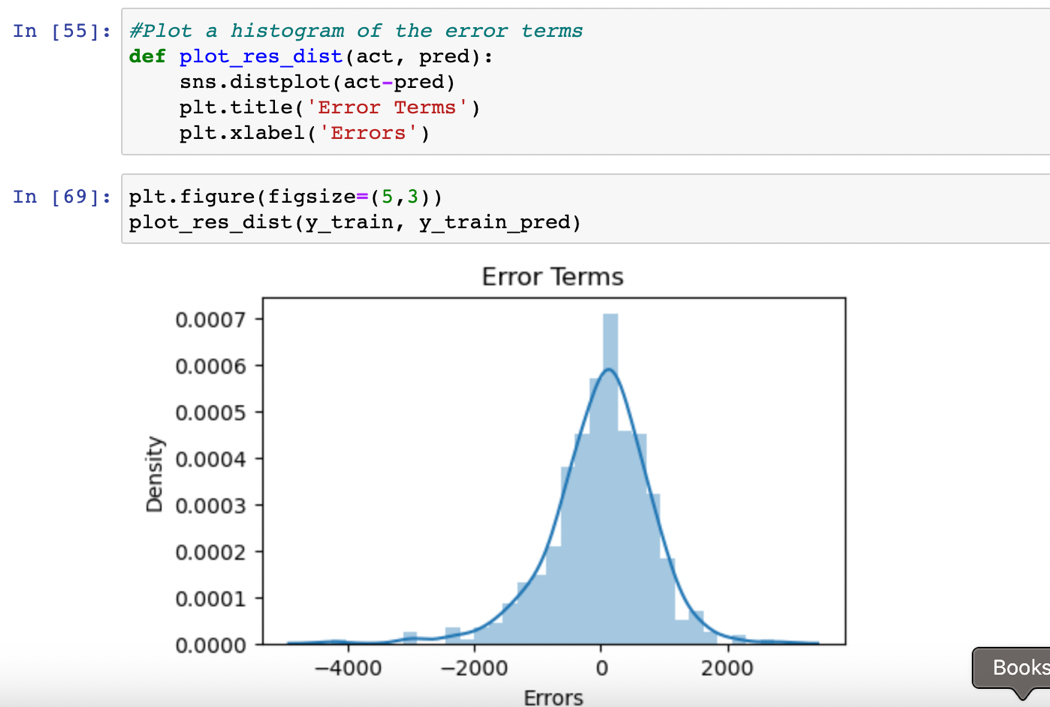
Q 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

*Multiple Linear Regression model assumptions*

1. *X and Y should have linear relationship*
2. *All independent variables should not be correlated with each other*
3. *Error terms should be normally distributed with mean at 0*
4. *Error terms should have constant variance*

*All independent variables should not be correlated with each other: This has been taken care by removing the variables with high VIF which signifies the correlation*

*Error terms are normally distributed has been verified during residual analysis:*



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

ANSWER: The top 3 features contributing significantly towards explaining the demand of the shared bikes are:

|  |
| --- |
| 1. temp |
| 1. yr |
| 1. season\_winter |