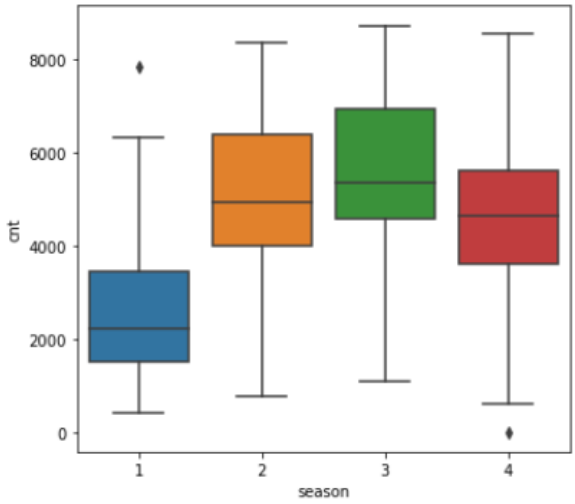
Assignment-based Subjective Questions

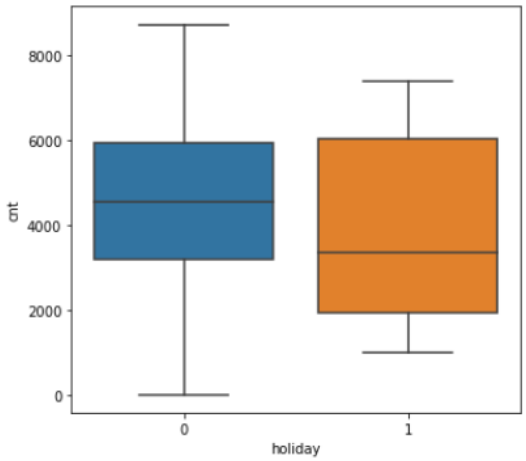
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

   Ans:

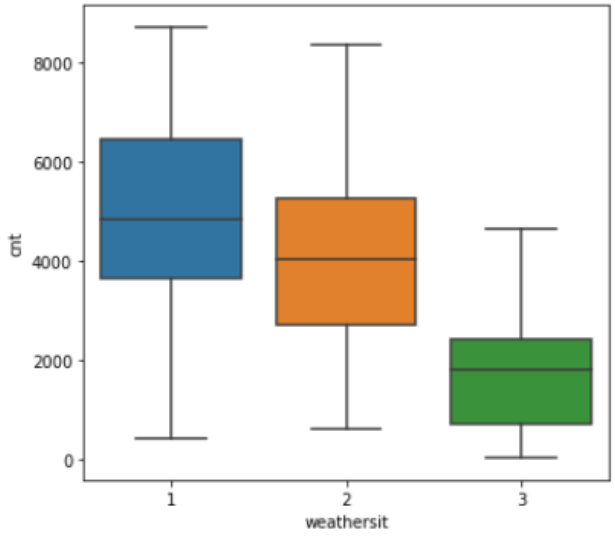
   Seasons: Count is higher in fall (highest), summer and winter and less in spring



   Holiday: When its is a non-holiday day, count is higher



   Weather: when weather is clear count is highest, and in heavy rain there is no bike rented counted



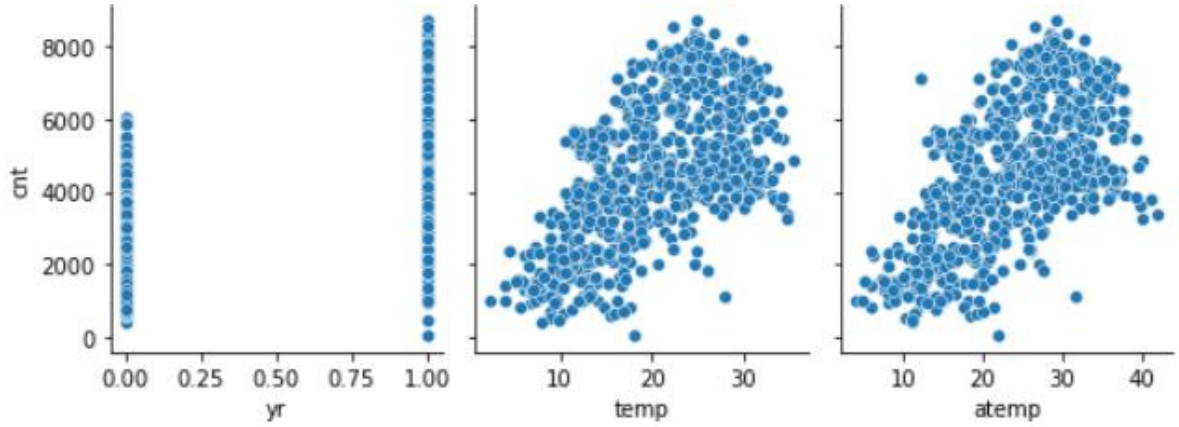
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

   Ans: N number of categorical variables can be explained by N-1 dummy values

   eg: Male is 1000 which can be explained by 000 as well, so we can drop the 1st column, reducing extra category column.

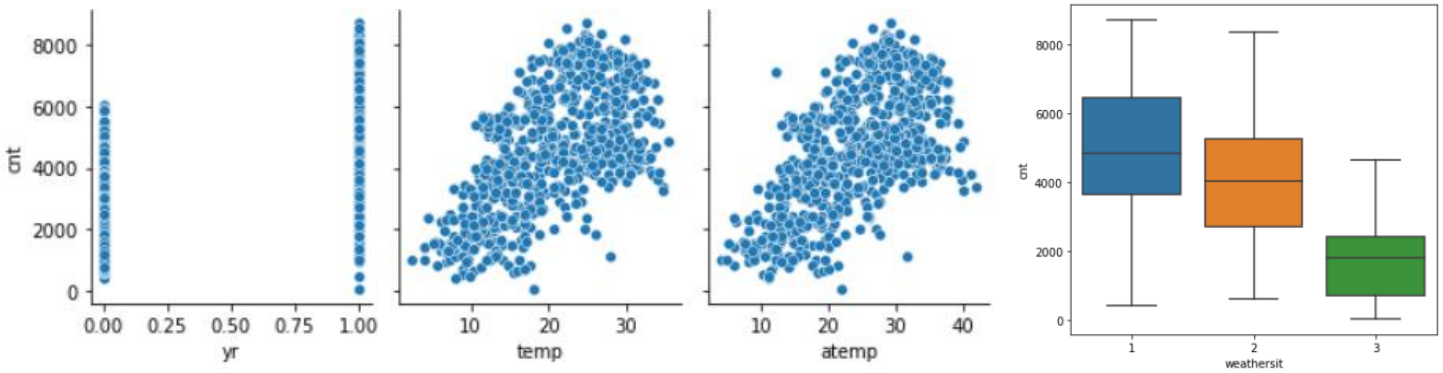
Gender	Male	Female	Transgender	Other
Male	1	0	0	0
Female	0	1	0	0
Transgender	0	0	1	0
Other	0	0	0	1

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



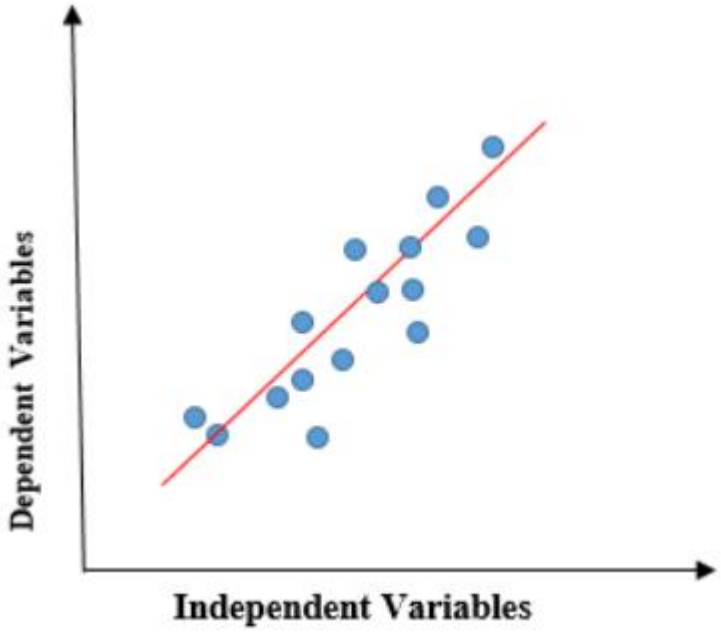
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   Ans: 1. VIF values to remove multicollinearity
   2. Residual errors are homoscedastic
   3. Residuals follow normal or almost normal distribution

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: based on summary and univariate analysis of features, temp, year and weather has the influence on the count of bike sharing.



General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   Ans: Linear regression is simple statistical regression method for predictive analysis to show the relationship between dependent and independent variables. It shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis). If there is only one single independent variable to define dependent variable then it is **simple linear regression** if there are more than one independent variable to define target variable then it is called **multiple linear regression**.



To calculate best-fit line linear regression uses a traditional slope-intercept form.

$$y = mx+b \quad \Longrightarrow \quad y = a_0+a_1x$$

y= Dependent Variable.
x= Independent Variable.
a0= intercept of the line.
a1 = Linear regression coefficient.

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)