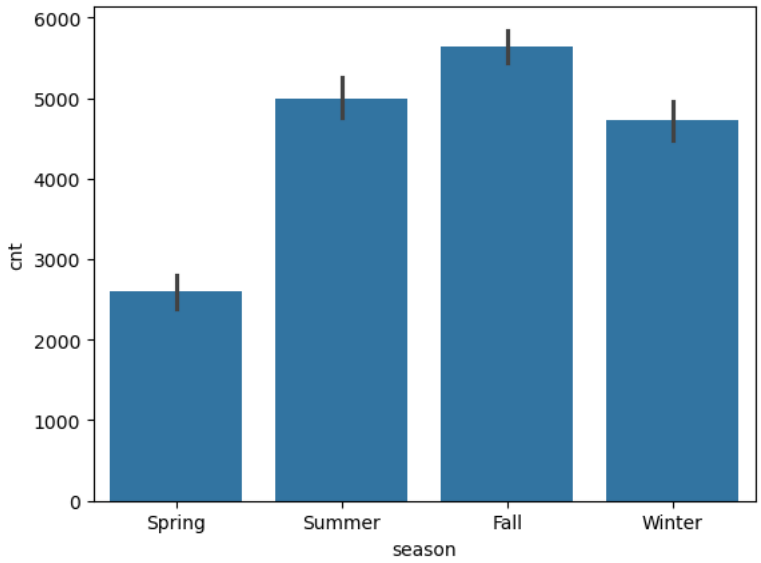
Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

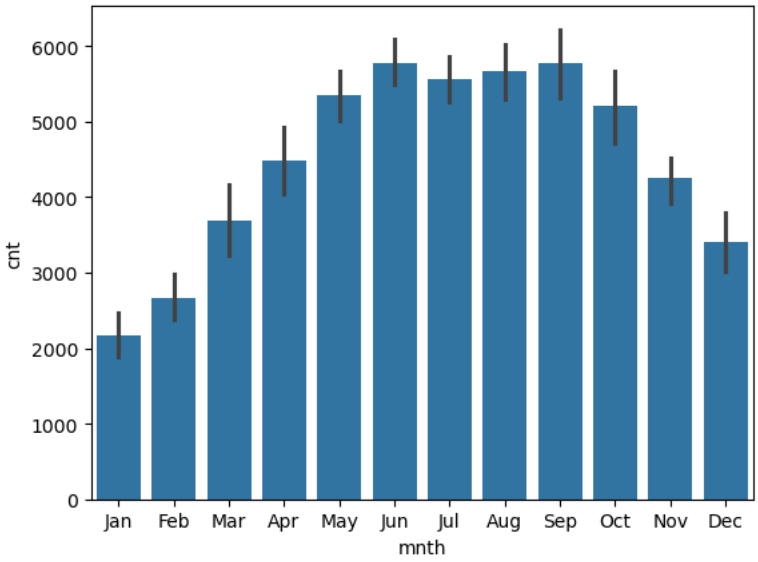
Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

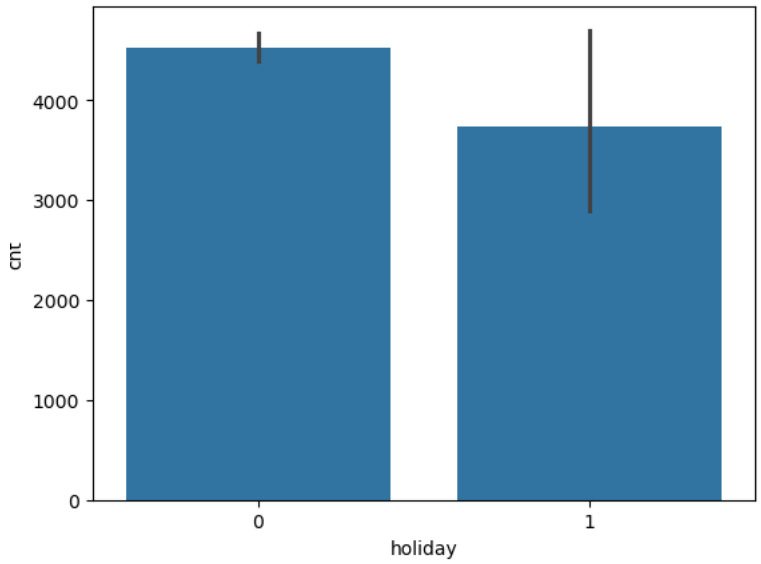
In Seasons – Fall has the highest count.



In month – Jan has lowest count and Jun has highest count.



Bike Share count is higher when there is no holiday



Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

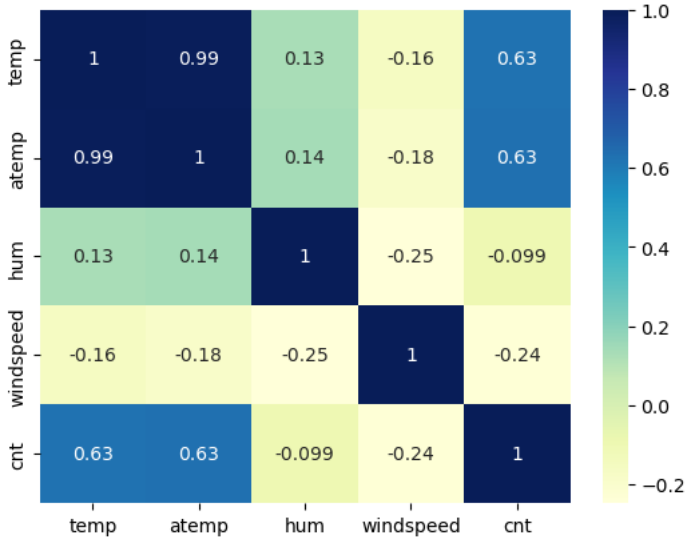
drop_first = True, will drop the extra column created during dummy variable creation. This helps in reducing the correlation among the dummy variables.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

Among the numerical variables, “temp” and “atemp” had the maximum correlation of 0.63



Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)
Features were selected using RFE (Automated) and p-value + VIF (Manual). Once we have all the features which met p-value < 0.05 and VIF <= 5, we calculated the R^2 value > 75%. Then we validated the model and checked the r-squared value.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)
From my model, the top 3 features are Windspeed, Workingday and Year.

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Linear Regression Algorithm is supervised ML model where the model finds the best fit between the independent and dependent variable. It finds the linear relationship between the dependent and independent variables.
Single Linear regression – one independent variable is present, and model has to find linear relationship with the dependent variable
Multiple Linear regression – more than one independent variable is present and model has to find linear relationship with the dependent variable >

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

< Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.
It is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. >

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

< Pearson's R or Pearson Correlation Coefficients are used to measure how strong a relationship is between two variables. The Pearson correlation coefficient, (r), is metric for assessing linear relationships between two variables. It yields a value ranging from –1 to 1, indicating both the magnitude and direction of the correlation. A change in one variable is mirrored by a corresponding change in the other variable in the same direction. >

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Scaling adjust the spread or variability of the data set hence making the analysis more accurate.
Normalization – Preserves the shape of the distribution but changes the scale
Standardized – Transforms data to have a mean of 0 and standard deviation of 1.>

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<VIF values shows infinite if there is perfect multicollinearity. This happens when two or more independent variables are perfectly linearly dependent. >

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

< Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution. Also, it helps to determine if two data sets come from populations with a common distribution**.** This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions. >
