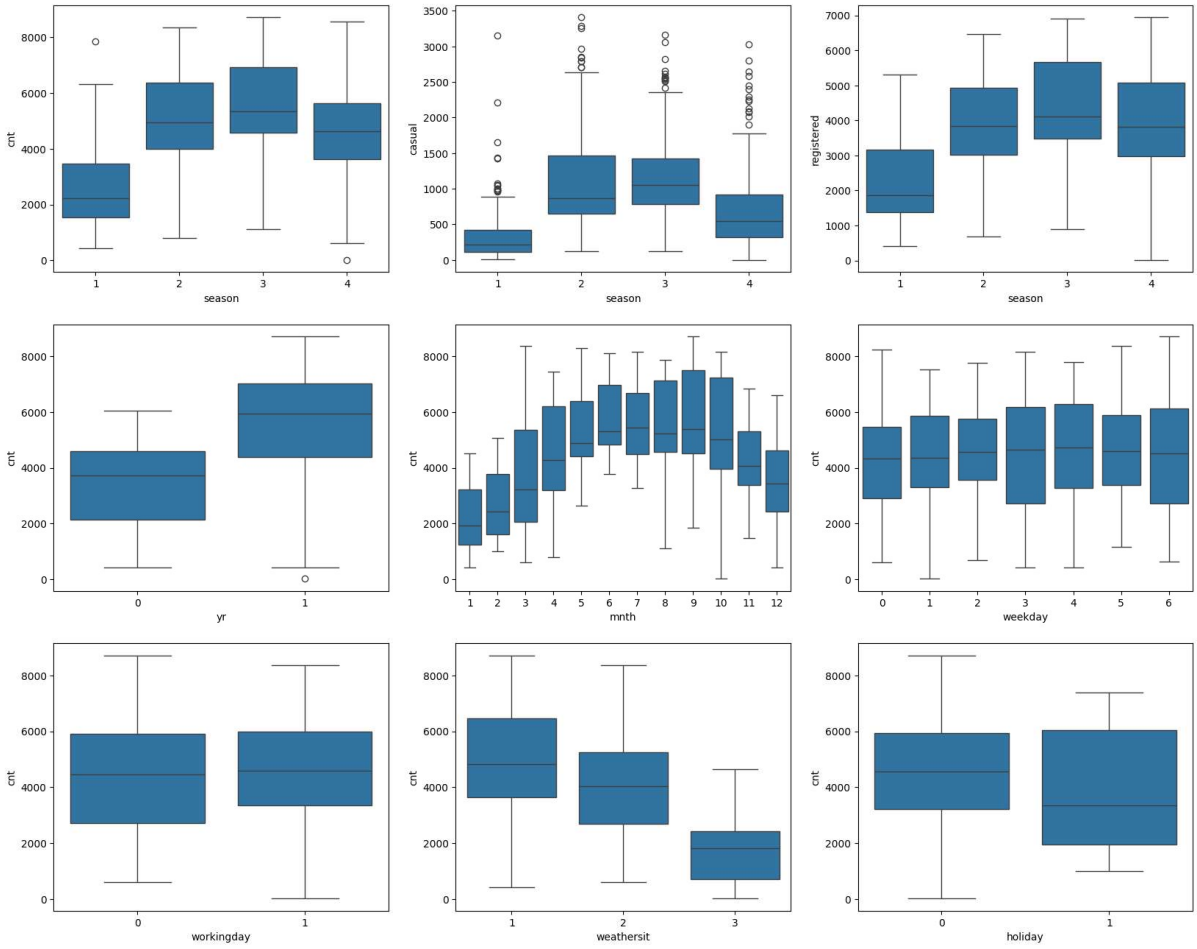
Assignment-based Subjective Questions

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)



1. Season 2 (Summer) & Season 3 (Fall) have more demand compared to winter and spring.
2. Year 2019 has more demand compared to 2018
3. Some months (6,7,8,9) have more demand compared to other months , inline with the season observation
4. Weatherset 1 (Clear weather) have more demand compared to other weather conditions.

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

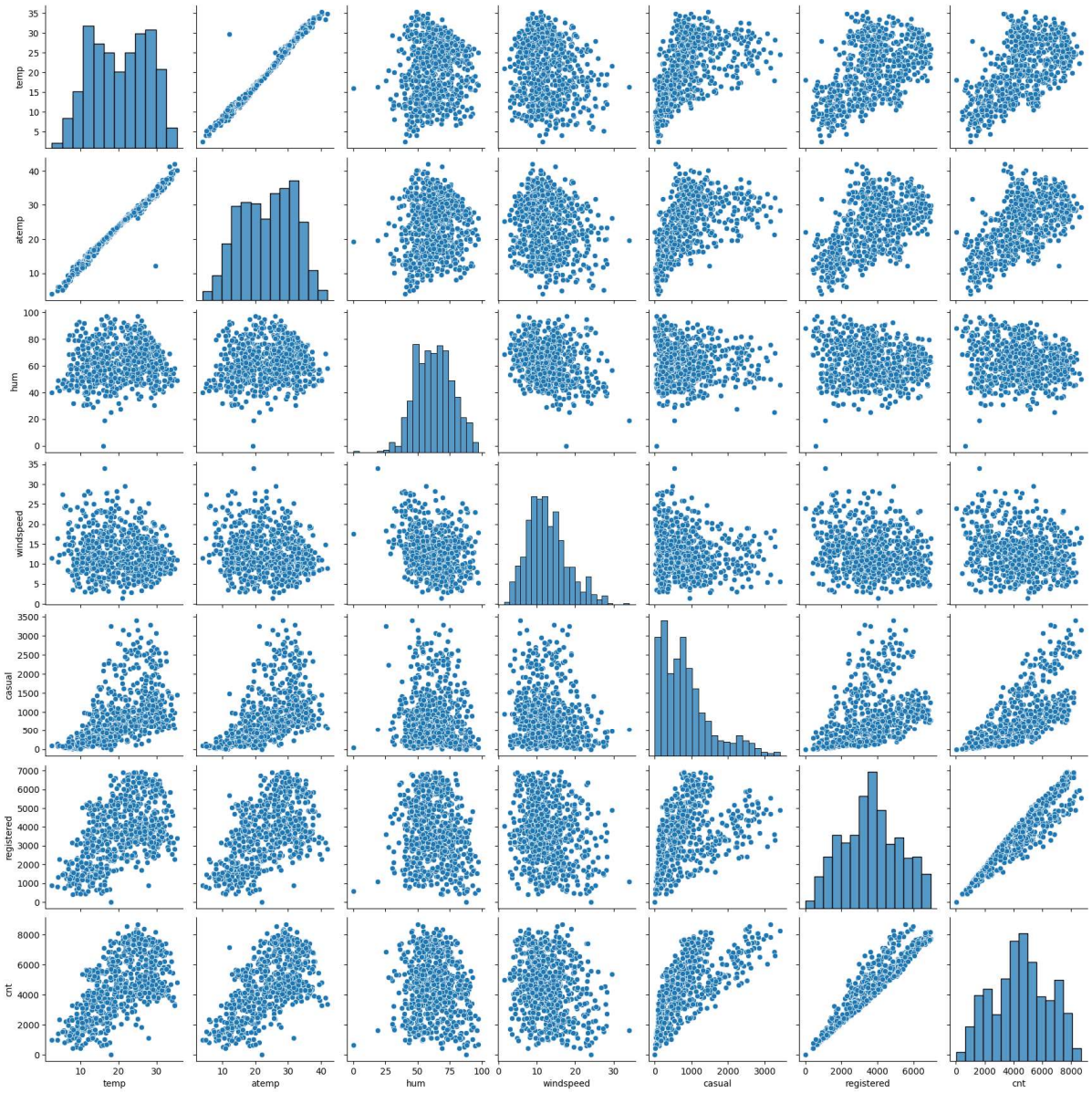
Total Marks: 2 marks (Do not edit)

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

We use this to reduce one column during dummy variable creation process to reduce the correlation

effects between dummy variables. As a rule of thumb , if we have 3 levels in a categorical variable , then we only need 3 – 1 = 2 columns and we drop the 3rd column created during dummy variable creation process.

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)

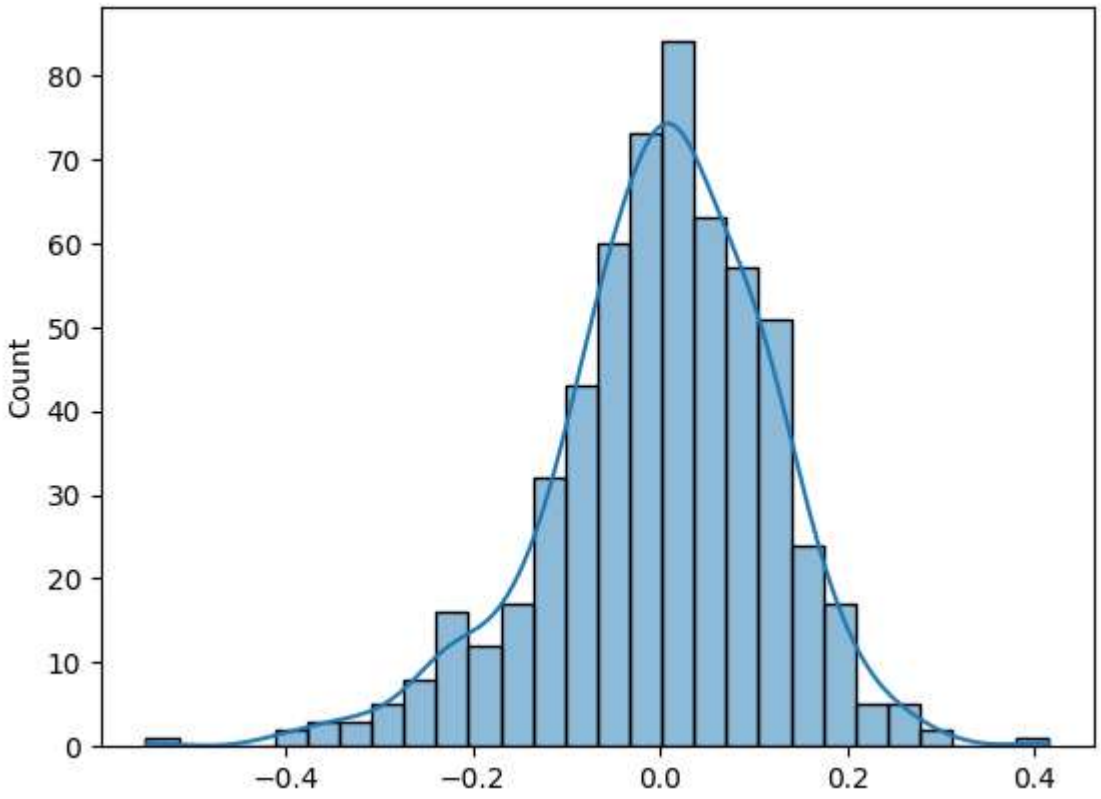


We can observe 'temp' and 'atemp' has more positive correlation with 'cnt' variable

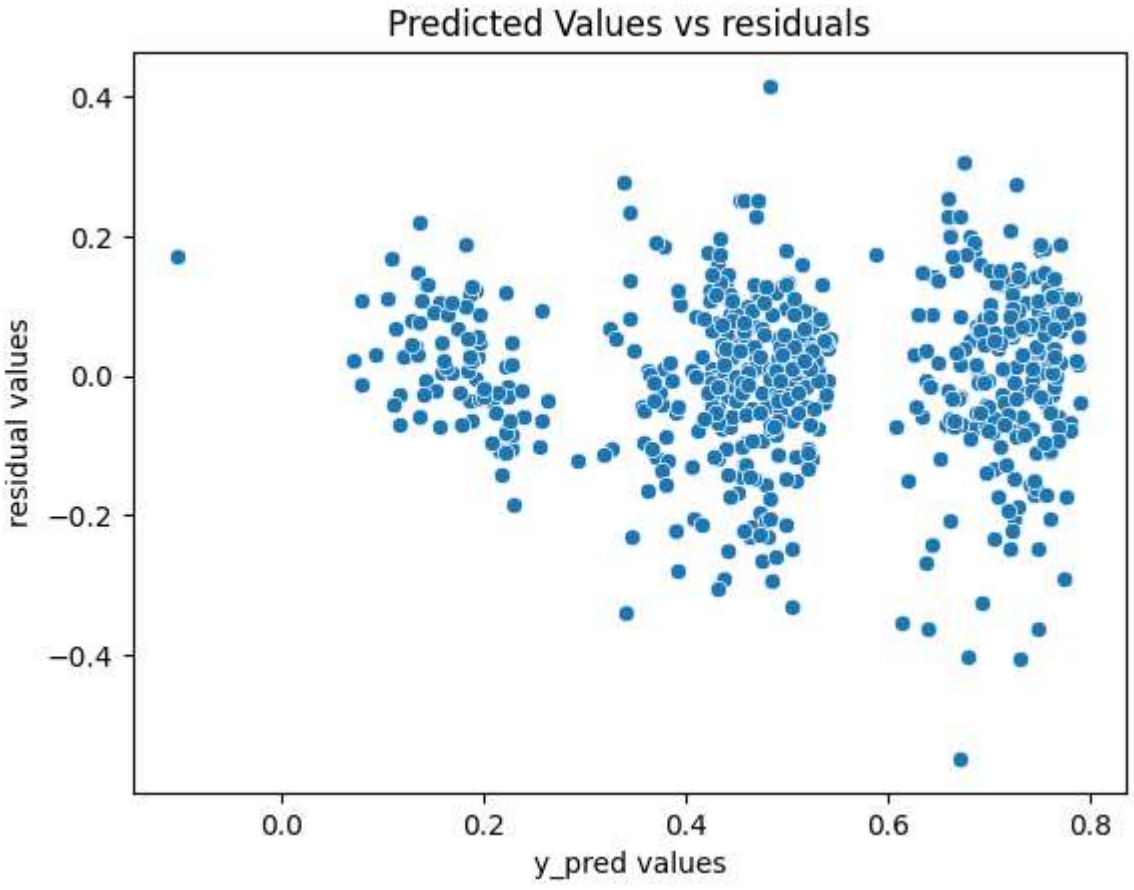
---

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)

From residual distribution graph , we could see that it follows normal distribution with mean centered at zero.



Also , the scatter plot between predicted value and the residual value showed no pattern.



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)

Top 3 contributing features are

1. Season_spring -> coefficient of -0.2986
2. Weathersit_light_rain -> coefficient of -0.2721
3. Yr -> coefficient of 0.2422

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)

<Your answer for Question 6 goes here>

Linear Regression is supervised learning type machine learning algorithm. This algorithm can used when the target variable is a continuous variable. It gives us the linear relationship between the target variable and independent variables.

The objective of using Linear regression is to find the best-fitting straight line with best possible prediction.

The general equation of simple linear regression is

$$Y = B_0 + B_1 * X$$

Where y is target variable , X is the independent variable.
B_0 is the intercept (value of Y , when x=0)
B_1 is the slope of the equation.

The algorithm gives the values of the intercept and slope.

We use scaling method to scale the continuous variables before making the model.

Then , we fit the model on train dataset and predict the values of y.

We calculate the residual , which is the difference between true y value and y predicted.

We validate out model by ensuring that the residual follow a normal distribution curve centered at zero.

Also , the scatter plot of residuals and y predicted should not have any visible pattern.

And finally , we can test the model on test data and calculate r2 score.

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)

<Your answer for Question 7 goes here>

Anscombe's Quartet is a set of four different datasets that have nearly identical statistical properties (such as mean, variance, correlation, and regression line) but appear very different when plotted. It was created by Francis Anscombe in 1973 to emphasize the importance of data visualization in statistical analysis.

It demonstrates how relying solely on summary statistics (mean, variance, correlation, etc.) can be misleading.

It highlights the importance of visualizing data before drawing conclusions.

It shows how outliers and patterns in data can significantly affect interpretations.

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)

<Your answer for Question 8 goes here>

Pearson's r is a numerical summary of the strength of linear relationship between the variables.
It's value lies between -1 and +1 .
The value of 1 indicates a strong positive relationship
The value of -1 indicates a strong negative relationship
The value of 0 indicates no relationship.

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)

<Your answer for Question 9 goes here>

Scaling is the process of transforming numerical features so that they have a specific range or distribution. This is crucial in machine learning because many algorithms are sensitive to the magnitude of numerical values, and unscaled data can lead to inefficient training, poor performance, or biased models.

Scaling is Performed to

1. Prevents Features from Dominating Others:
2. Ensures Equal Weightage for Features

Aspect	Normalization (Min-Max Scaling)	Standardization (Z-score Scaling)
Definition	Rescales values to a fixed range, typically **[0,1]** or **[-1,1]**.	Transforms data to have **zero mean** and **unit variance**.
Formula	$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$	$X' = \frac{X - \mu}{\sigma}$
Effect on Data	Retains relative distances but compresses the scale.	Centers the data around 0 with a standard deviation of 1.
Sensitive to Outliers?	**Yes** (since it depends on min/max values).	**No** (less sensitive due to mean and standard deviation).

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)

<Your answer for Question 10 goes here>

This happens when any of the independent variable is perfectly correlated with one or more of the other independent variables.
This could happen due to duplicate columns , redundant column not removed during dummy variable creation.

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)

<Your answer for Question 11 goes here>

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset to a theoretical distribution (commonly a normal distribution). It helps assess whether the

data follows a particular distribution by plotting the quantiles of the dataset against the quantiles of the reference distribution.

Linear regression makes several assumptions, one of which is that the residuals (errors) should be normally distributed. A Q-Q plot is crucial for checking this assumption