Assignment-based Subjective Questions

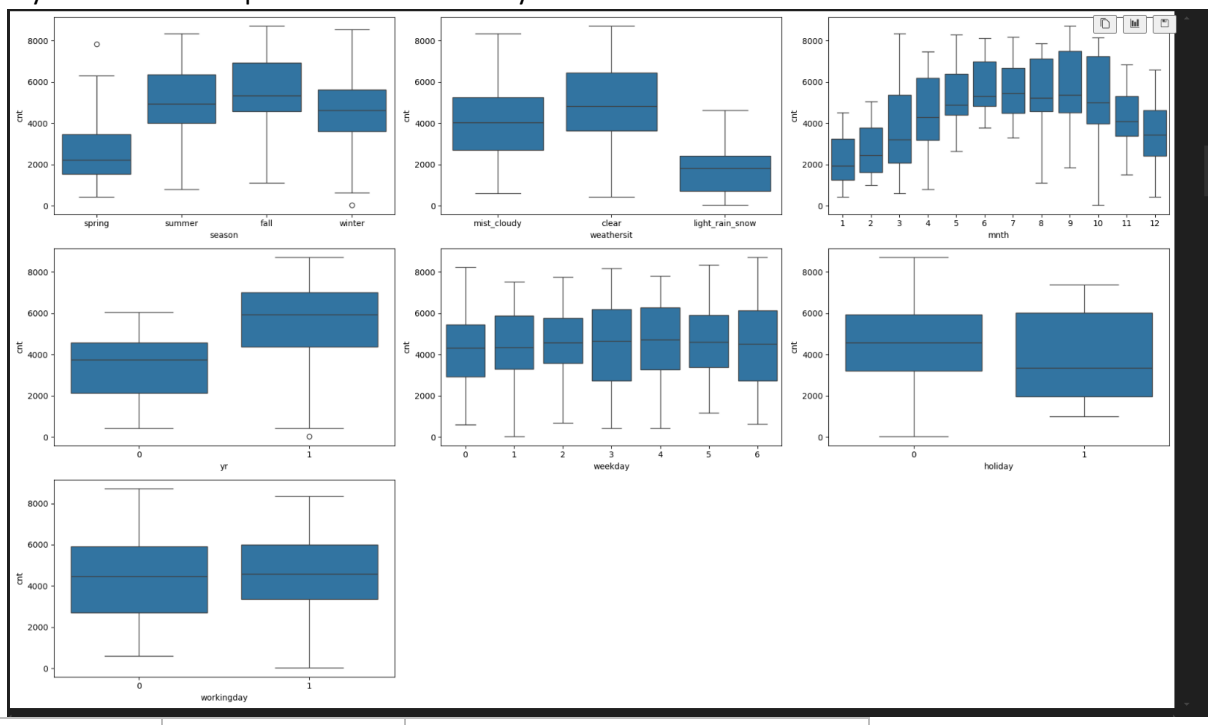
Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

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

Varriables like mnth, yr, weekday, holiday and workingday appear as vertical strips in scattered plot with dependent variable along with season and weathersit. These are categorical in nature and should be converted.

If you draw the box plot of these variables you see



Feature	Influence on cnt	Notes
weathersit	✅ Strong	Clear trend: Clear > Cloudy > Rain
season	✅ Strong	Summer/Fall best
mnth	✅ Strong	Mid-year peak
yr	✅ Moderate	2019 better than 2018
holiday	⚠️ Weak/Noisy	Needs further analysis
workingday	⚠️ Weak	Flat trend
weekday	⚠️ Weak	Minimal variation

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)

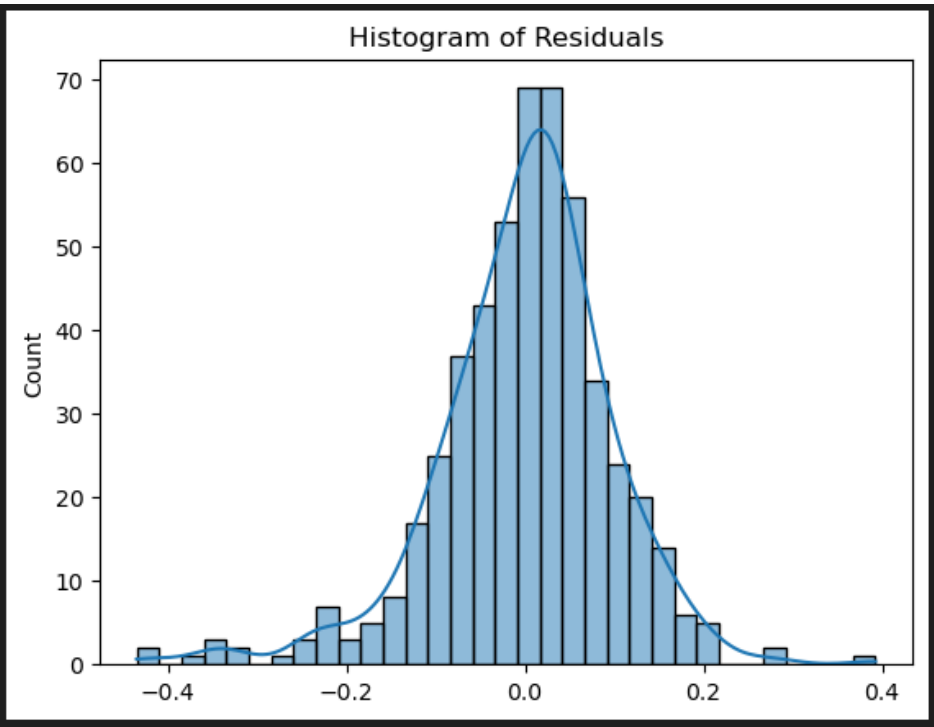
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

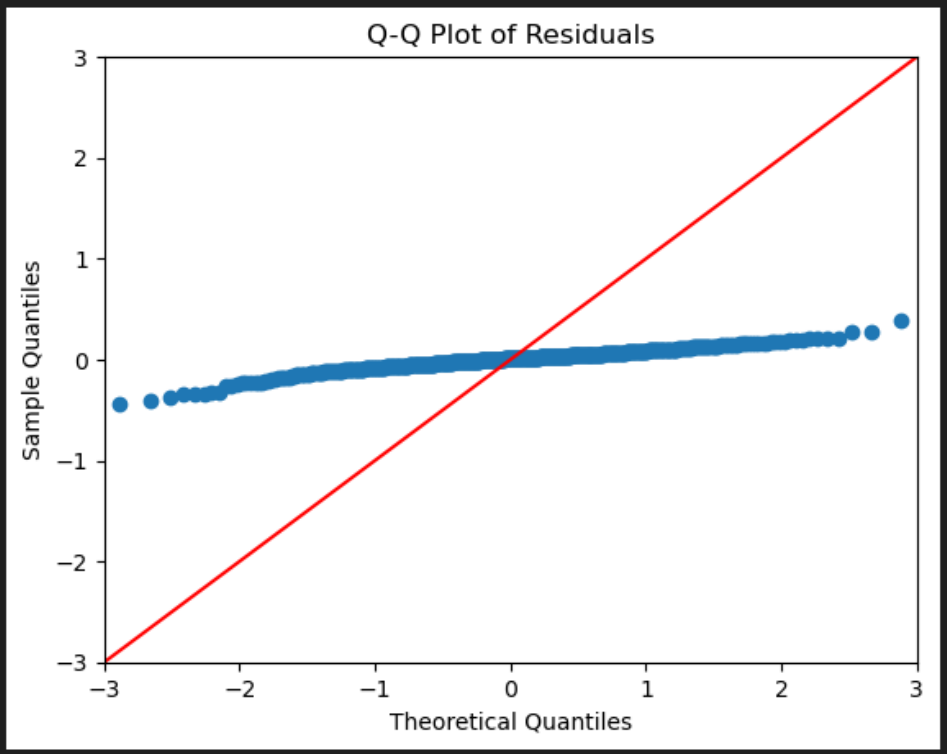
Always use drop_first=True when creating dummy variables for regression to avoid multicollinearity and the dummy variable trap.

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)
- temp and atemp are almost perfectly correlated.

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)
To validate linear regression assumptions:
1. Normality of residuals: Q-Q line is almost straight and histogram is bell curve
2. Multicollinearity (VIF): VIF for all the features is <5


Histogram of Residuals
Count
70
60
50
40
30
20
10
0
−0.4
−0.2
0.0
0.2
0.4


Q-Q Plot of Residuals
Sample Quantiles
Theoretical Quantiles
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

VIF Scores:

	Feature	VIF
0	const	46.070448
1	temp	3.288718
2	season_spring	4.455441
3	season_summer	2.000451
4	season_winter	2.968406
5	weathersit_light_rain_snow	1.036069
6	weathersit_mist_cloudy	1.030579
7	yr_1	1.019145

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)

Based on low Low P>|t| values and High absolute values of coefficients, here are the strong variables

1. temp: More warmth = more bike use
2. yr_1: Usage has grown over time (2019 vs 2018)
3. weathersit_light_rain_snow**:** Bad weather significantly reduces usage.

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>

Liner Regression is like drawing the best straight line through your data to predict future values. It is of two types

1. Simple Linear regression (one feature): $y=\beta0+\beta1x$
2. Multiple Linear regression (many features): $y=\beta0+\beta1x1+\beta2x2+\cdots+\beta nxn$

- $y \rightarrow$ predicted value (target)
- $x_1, x_2, ..., x_n \rightarrow$ input features
- $\beta_o \rightarrow$ intercept (constant)
- $\beta_1, ..., \beta_n \rightarrow$ coefficients

Goal of linear regression is to minimize the sum of squared errors (residuals)

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>

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>

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 This is useful when your data has numbers with very different sizes — like one column with values from 1 to 100, and another from 1 to 10,000. If you don't scale, some machine learning models may wrongly think that bigger numbers are more important — even if they are not.

- Use **normalization** when you want all values to be between **0 and 1**.
- Use **standardization** when you want the values to have a **center at 0** and be spread out in a standard way.

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>

It happens when the feature is perfectly corelated

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 shows if the residuals of your regression model are normally distributed.
If the points fall on a straight line, the assumption is valid

---