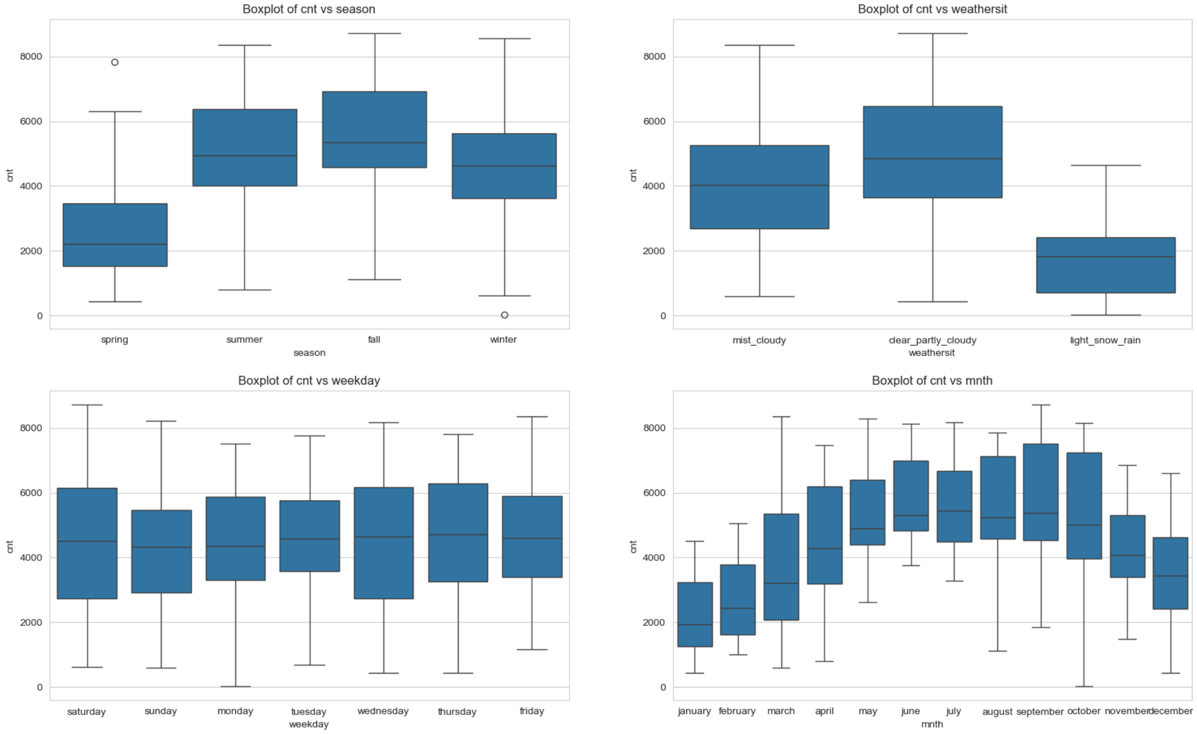
Assignment-based Subjective Questions

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)

Categorical variables like season, weather, and working day have a significant effect on the bike demand. For instance, the season influences demand due to weather preferences, with summer and fall showing higher demand. Similarly, non-working days tend to show higher demand as people use bikes for leisure activities. Weather conditions like "clear or partly cloudy" also positively influence demand, while "heavy rain or snow" reduces it.



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)

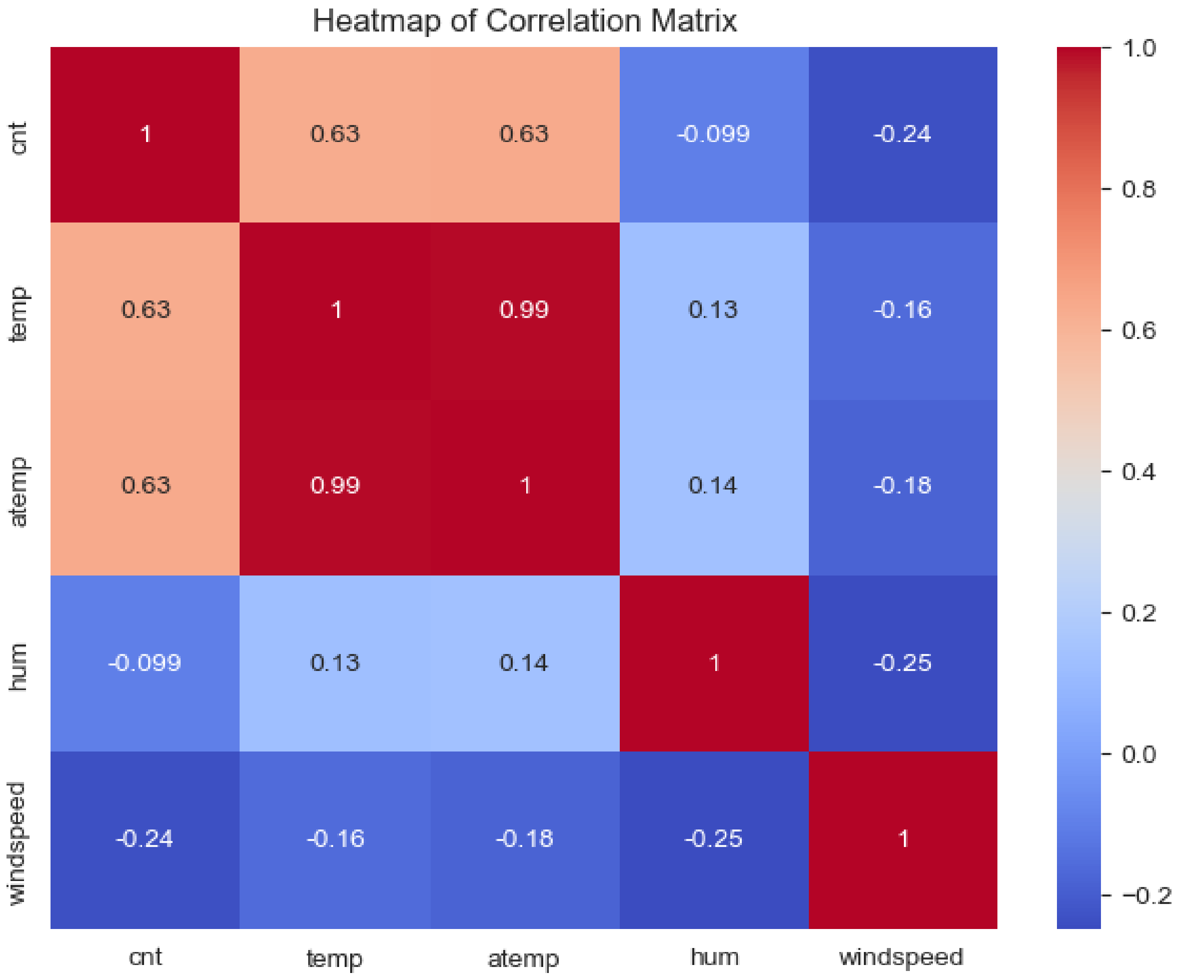
Using drop_first=True gets rid of multicollinearity by dropping one category of the dummy variables. Its prevents from creating extra variable. This makes sure the model doesn't have problems with perfectly collinear features.

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)

atemp and temp both have same correlation with target variable of 0.63 which is the loftiest among all numerical variables.



	cnt	temp	atemp	hum	windspeed
cnt	1	0.63	0.63	-0.099	-0.24
temp	0.63	1	0.99	0.13	-0.16
atemp	0.63	0.99	1	0.14	-0.18
hum	-0.099	0.13	0.14	1	-0.25
windspeed	-0.24	-0.16	-0.18	-0.25	1

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)

Model evolution

Checking linearity.
Checking Normality of residuals.
Checking VIF.
Checking R2 and Adjusted R2
RIF

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)

The top 3 features contributing significantly towards explaining the demand of the shared bikes are 'yr', 'temp' and 'weathersit_light_snow_rain'.

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 is a supervised learning algorithm used to model the relationship between a dependent variable (target) and one or more independent variables (predictors). It assumes a linear relationship between the variables and aims to find the best-fit line using the equation:

Y = mX + c
In this equation:

- Y represents the dependent variable under scrutiny for prediction.
- X stands for the independent variable employed for predictive purposes.
- m denotes the slope of the regression line, signifying the impact of X on Y.
- c is a constant known as the Y-intercept, indicating that when X equals 0, Y will equal c

Key steps include:

- Minimizing the sum of squared residuals (differences between observed and predicted values) using Ordinary Least Squares (OLS).
-Evaluating model fit through metrics like R-squared, adjusted R-squared, and RMSE.
-Validating assumptions such as linearity, normality of residuals, and absence of multicollinearity.

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 consists of four datasets with nearly identical statistical properties (mean, variance, correlation, and regression line) but vastly different distributions when visualized. It highlights the importance of data visualization in statistical analysis. Each dataset tells a different story despite similar summary statistics, emphasizing that relying solely on numbers can be misleading without visual context.

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 the Pearson correlation coefficient, measures the strength and direction of a linear relationship between two variables. Its value ranges from -1 to 1:

- R=1: Perfect positive correlation.
- R=−1: Perfect negative correlation.
- R=0: No linear correlation.
  It is calculated as the covariance of the variables divided by the product of their standard deviations.

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 adjusts feature values to a common scale to improve model performance and convergence. It is crucial for algorithms sensitive to feature magnitudes, such as regression and distance-based models.

- Normalization scales values to a [0, 1] range, useful when data needs to be bounded.
- Standardization transforms data to have zero mean and unit variance, ensuring features follow a standard normal distribution.

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 becomes infinite when two or more predictors are perfectly collinear, meaning one predictor can be exactly predicted as a linear combination of others. This leads to a division by zero in the VIF formula:

$\text{VIF}=1/(1-R2)$

Here, $R2=1$ for perfectly collinear predictors.

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)

A Q-Q (quantile-quantile) plot compares the distribution of residuals to a theoretical normal distribution. Points falling along the 45-degree line indicate normality. It is crucial in linear regression to validate the assumption of normally distributed residuals, ensuring reliable p-values and confidence intervals.
