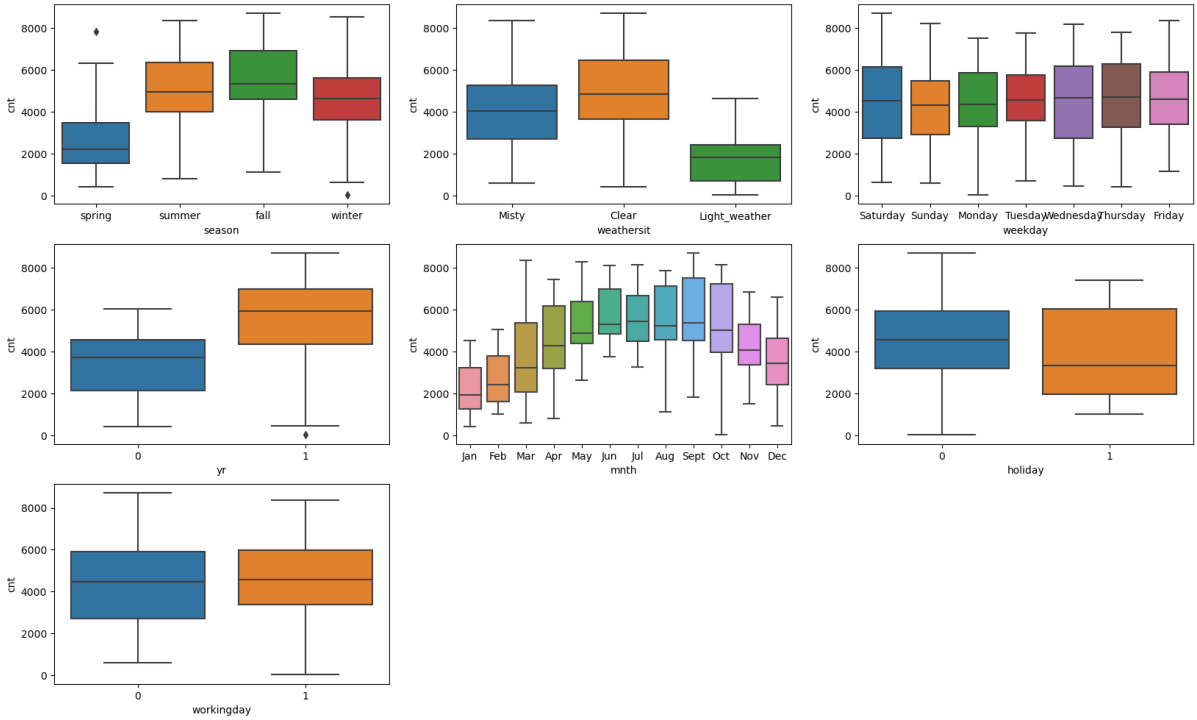
Assignment-based Subjective Questions

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)



There is more demand during fall, during clear weather and during months from June to Sept. Median is almost similar on all weekdays. There was more demand in 2019 as compared of 2018 which shows more demand with time.

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 is important to use during dummy variable creation as it drops the first dummy variable to reduce redundancy and multicollinearity.

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

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)

We validated the assumptions of Linear Regression after building the model on the training set by

first plotting histogram to check whether the residuals are approximately normally distributed and found they are. We then applied the same model to test set and checked the R square and R square adjusted for both train and test set and they came approximately near values.

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)
`Atemp: Feeling temperature in Celsius`
Yr: 2019 saw more demands than 2018
During Clear weather: Few clouds, Partly cloudy, Partly cloudy there was less demand for bike rentals
Winter season saw increase in bike rentals

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 a statistical model used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship of the form:
y = mx + c + e where y is the dependent variable, x is the independent variable, m is the coefficient, c is the constant and e is error terms. The onjective is to find coefficients using the Ordinary Least Squares method which minimizes the sum of squared 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>
Anscombe's Quartet consist of four datasets that have nearly identical simple descriptive statistics but they have very different distributions and visual appearances. The goal is to show the importance of visualizing data before drawing conclusions since statistics alone can be misleading

---

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 measure of the linear relationship between two variables. It quantifies how well the changes in one variable correspond to changes in another. It is used for determining the strength and direction of a linear relationship between two continuous variables.

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 variables to a specific range so they have similar magnitude and units. There are two types of Scaling: Normalized scaling or Min-Max Scaling and Standardised Scaling using z-score

---

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>
VIF becomes infinite when one predictor can be perfectly predicted from the others which makes it impossible to separate their individual effects in the model. This means that we need to either remove or combine the collinear variables to improve the model.

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 is a graphical tool to assess if a dataset follows a specific theoretical distribution, such as the normal distribution. A Q-Q plot helps visualize the linear regression assumption that the residuals of the model are normally distributed. It indicates outliers or extreme values in the data and hence useful for ensuring reliable and accurate model estimates.
