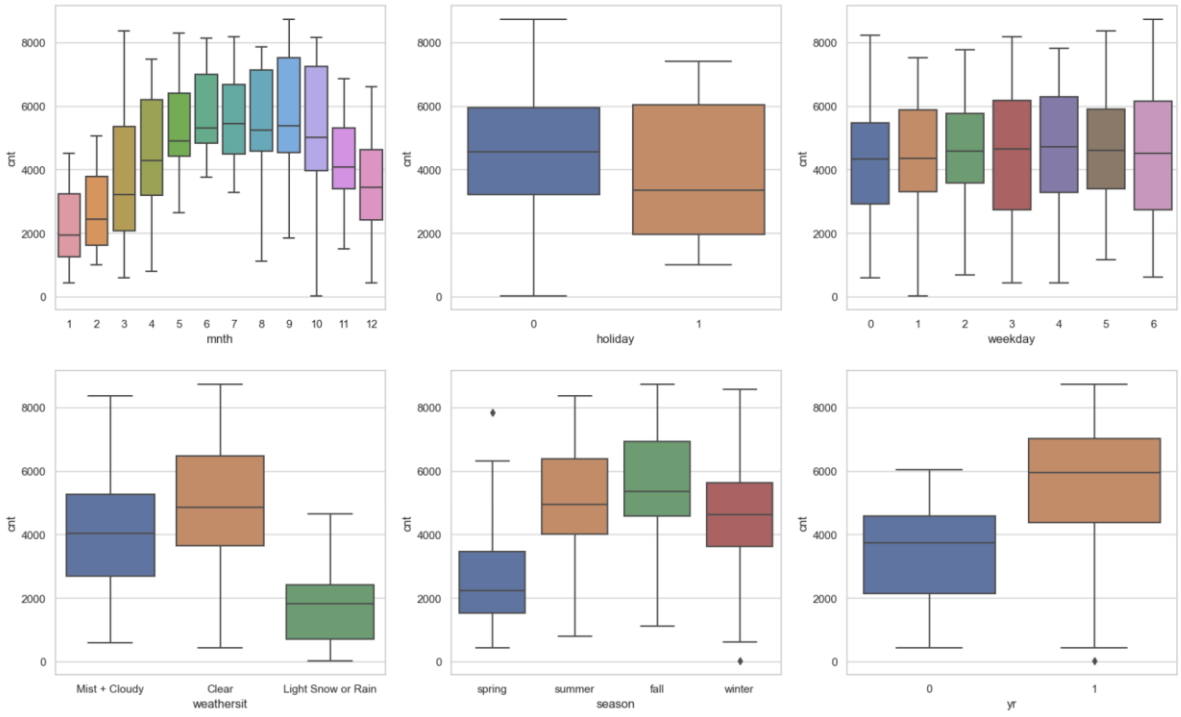
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)



1. Month - Plots shows that demand is low during the winter season.
2. Holiday - Usage during holiday period is higher on an average.
3. Weekday - Usage is same across the weeks on an average.
4. Weathersit – Demand is higher during clear weather.
5. Season – Demand is higher during summer and fall
6. Yr – demand has increased during 2019 compared to the prev year

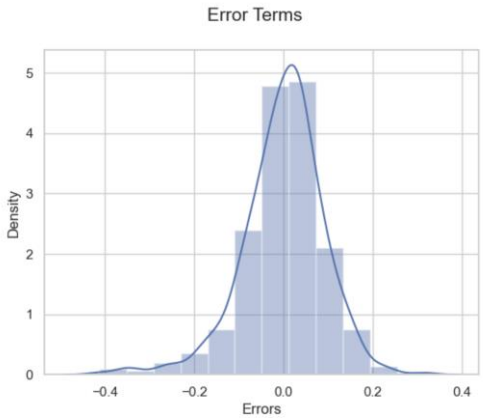
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

   It helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
   Dropping your first categorical variable is possible because if every other dummy column is 0, then this means your first value would have been 1.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

   Temp and atemp has the high correlation with target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)



Residual Analysis of the train data
- Look for the error between a predicted value and the observed actual value.
- Plot the data and ensure errors are independent and normally distributed.
- A normally distribute curve satisfies the assumptions of the normality of residuals

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
    a. Temp - 0.549936
    b. Light Snow and Rain - -0.288021
    c. Yr - 0.233056

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
    - Basic concept is based on Dependent and Independent Variables. Dependent Variable (Y) is the variable we are trying to predict. Independent Variable(s) (X) are the variable(s) used to predict the dependent variable. The general form of a linear equation with one independent variable is: Y=mX+b
    - Objective is to minimize the difference between the actual values and the predicted values by adjusting the parameters (slope and intercept) of the linear equation.
    - Using the cost function, measure the difference between the predicted values and the actual values. The most common cost function for linear regression is the Mean Squared Error (MSE)
    - Iteratively try different options for X to update the parameters until convergence (when the cost function reaches a minimum value).
    - Once the model is trained, use the learned parameters to make predictions on new data.

2. Explain the Anscombe’s quartet in detail.

    - Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation) but differ significantly when graphically depicted. Anscombe's quartet is used to emphasize the limitations of

relying solely on numerical summaries and the value of data visualization in statistical analysis.

3. What is Pearson's R?

- Statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. Denoted by r
- Range : -1 =< r >= 1
  r=1 indicates a perfect positive linear relationship.
  r=−1 indicates a perfect negative linear relationship.
  r=0 indicates no linear correlation
  Closer to 0 indicate weaker Correlation

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
   - Scaling is used for transforming the values of variables to a specific range or distribution.
   - The primary goal of scaling is to bring all the features of a dataset onto a similar scale, making it easier to compare and analyse them.

	Normalized	**Standardized**
Range	Transforms values to a specific range (e.g., 0 to 1)	Centres values around the mean, with a standard deviation of 1
Sensitivity to Outliers	Sensitive to outliers because it is based on the minimum and maximum values	Less sensitive to outliers because it uses the mean and standard deviation
Preservation of Distribution Shape	Preserves the shape of the original distribution	May change the shape of the distribution, especially if the data is not approximately normally distributed
Algorithms	Min-Max Scaling	Z-score Scaling

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
   - It occurs is equal to 1. This situation occurs when One or more independent variables are exact linear combinations of other variables in the model
   - It can be solved by removing redundant variables or combining variables into a composite variable

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

- (Q-Q) plot is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution.
- It compares the quantiles of the observed data with the quantiles expected from the theoretical distribution.
- It helps to understand,
  - Normality Assumption
  - Identification of Outliers

    - Residual Analysis
    - Model Validity
- It is used for assessing the normality of residuals and identifying potential issues with the model's assumptions. They provide a visual representation of how well the residuals conform to the expected distribution.