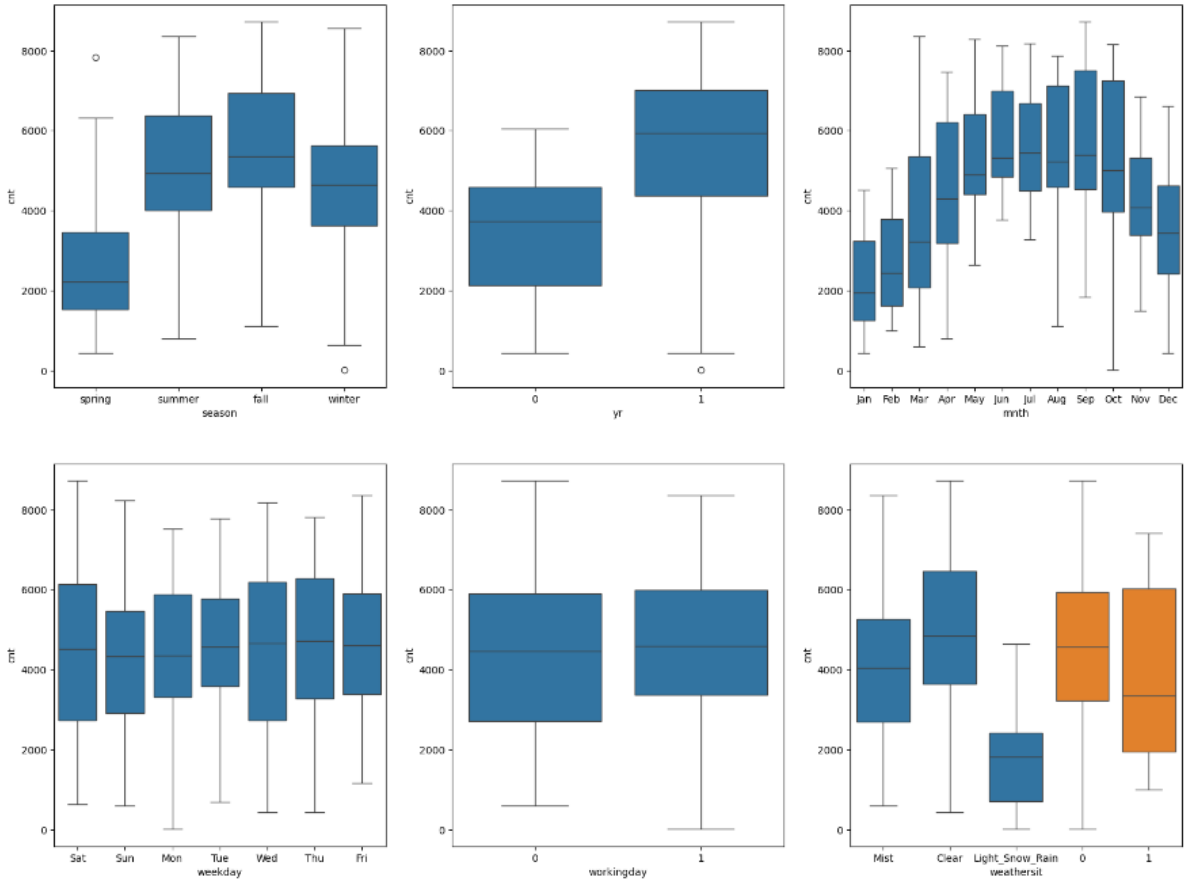
Assignment-based Subjective Questions

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

From the Categorical Variables, below are the boxplots where:

- Working day which is actual working days, those counts are high.
- When weather is clear, counts are high.
- Weekdays, month, season are more or less has same counts.



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

- It helps in reducing the first or extra column created during dummy variable creation.

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' both has highest correlation with the target variable. So we can consider 'temp' variable.

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

- There is a normal distribution in Error Terms.
- Linearity should be visible among variables.

- There is no visible patterns in residual values.
- There is no auto-correlation.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

   Three Features:
   - Temp
   - Year
   - Season

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

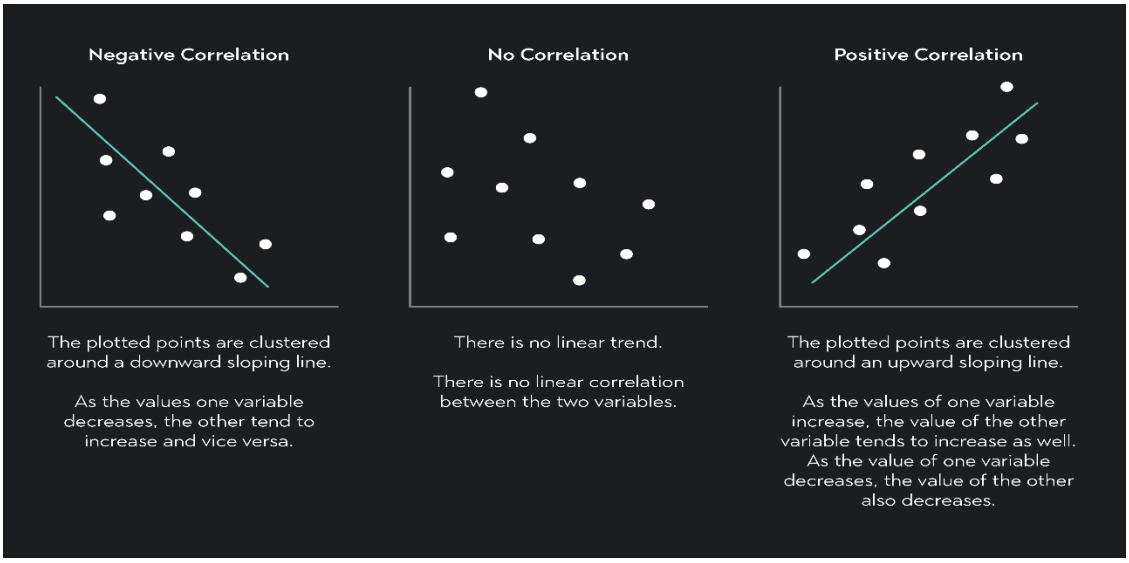
   - Linear Regression is a data analysis technique which predicts the value of unknown data by using the related another and known data. Example: We can predict the sales of the goods from the past sales data which is known.
   - It generally tells the relationship between the dependent (target variable) and independent variables.
   - This relationship can be both Positive linear relationship and Negative linear relationship.
   - If both independent and dependent variable increases then positive relationship
   - If independent increases and dependent decreases then negative relationship.

2. Explain the Anscombe's quartet in detail. (3 marks)

   - It generally is used to illustrate the importance of EDA and drawbacks of depending only on statistical summary.
   - Anscombe's Quartet shows how four entirely different data sets can be reduced down to the same summary metrics

3. What is Pearson's R? (3 marks)

   - The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation.
   - It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

- It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
- If scaling is not performed than algorithm tends to weigh high values magnitudes and ignore other parameters which will result in incorrect modeling.

Difference between Normalizing Scaling and Standardize Scaling:

- In normalized scaling minimum and maximum value of features being used whereas in Standardize scaling mean and standard deviation is used for scaling.
- Normalized scaling is used when features are of different scales whereas standardized scaling is used to ensure zero mean and unit standard deviation.
- Normalized scaling scales values between (0,1) or (-1,1) whereas standardized scaling is not having or is not bounded in a certain range.
- Normalized scaling is affected by outliers whereas standardized scaling is not having any effect by outliers.
- Normalized scaling is used when we don't know about the distribution whereas standardized scaling is used when distribution is normal

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

- VIF(Variance Inflation Factor) basically helps explain the relationship of one independent variable with all the other independent variables.
- The formulation of VIF is given below:
  - A VIF value of greater than 10 is definitely high, a VIF of greater than 5 should also not be ignored and inspected appropriately.
  - A very high VIF value shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.

To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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 probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data possibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution
- QQ plot can also be used to determine whether or not two distributions are similar or not. If they are quite similar you can expect the QQ plot to be more linear. The linearity assumption can best be tested with scatter plots. Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot.
- Importance of QQ Plot in Linear Regression : In Linear Regression when we have a train and test dataset then we can create Q-Q plot by which we can confirm that both the data train and test data set are from the population with the same distribution or not.