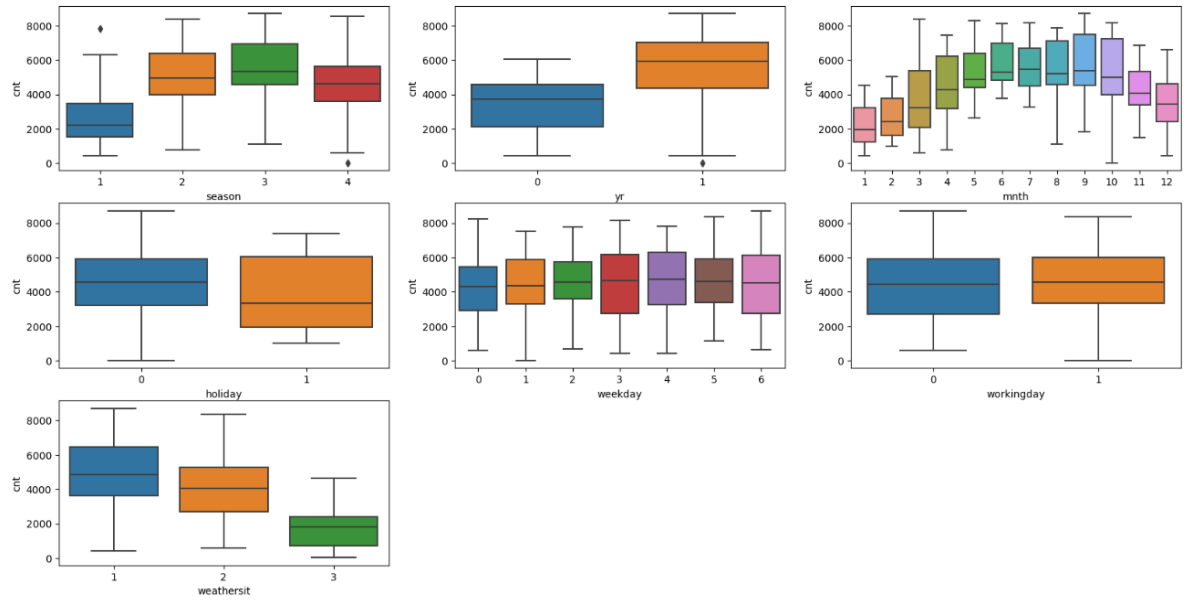
Assignment-based Subjective Questions

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

Ans: The boxplots for the different categorical variables are given below. From the Boxplots for the different categorical variables, we can infer the following:

a) The median value of the count follows the following trend for the seasons: Spring < winter < Summer < Fall
b) The count has increased significantly in year 2019 as compared to year 2018.
c) The trend of count against the months of the year is first increasing and then decreasing (highest median values are from June to September).
d) The count is lower on holidays as compared to regular days.
e) The count does not show any significant trend against the day of the week (weekday).
f) The count does not show any significant trend against the working day status.
g) The count shows a trend for the weather: Light snow/rain < Mist + Cloudy < Clear with few clouds



2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   Ans.: If we donot drop the first level of the dummy variable, then the problem of multi-collinearity will arise as the sum of the different levels of the dummy variable is always 1. So to avoid multi-collinearity, the first level has to be dropped. Also, dropping the first variable makes the interpretation of the coefficients easier.
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
   Ans.: Variable 'atemp' has the highest correlation with the target variable.
4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
   Ans.: (1) The errors follow normal distribution (verified using a distribution plot of the errors).
   (2) There is no multi-colinearity among the independent variables – verified using the VIF table (VIF < 5).
   (3) There is no specific trend of the error terms when plotted against the dependent variable.
5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**
   Ans.: Based on the final model, the top 3 features (having highest magnitude of coefficients):
   a) Temperature (coefficient = 0.63),
   b) Year (coefficient = 0.24), and
   c) Light snow or rain (coefficient = -0.28)

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**
   Ans: The goal of the linear regression algorithm is to find a linear equation relating the dependent variable to the independent (predictor) variables, such that the root mean square error (RMSE) between the predicted values and the actual values of the dependent variable is minimised. The different steps involved in the process are:
   (1) Calculate the mean values of X and Y variables (Xmean and Ymean)
   (2) Calculate the deviation of each value of X from the Xmean calculated above (Xdev).
   (3) Calculate the deviation of each value of Y form the Ymean calculated above (Ydev).
   (4) Calculate the product of each values of X deviation and Y deviation (= Xdev x Ydev) and calculate their sum (= ΣXdev.Ydev)
   (5) Calculate the sum of squares of Xdev (= $\Sigma Xdev^2$)
   (6) Coefficient of the predictor variable $\beta_1 = \Sigma Xdev.Ydev \ / \Sigma Xdev^2$
   (7) Constant term = Ymean - β_1 . Xmean

2. **Explain the Anscombe's quartet in detail. (3 marks)**
   Ans: The Anscombe's quartet are a group of datasets which are qualitatively different (look different when plotted in a scatter plot), but these datasets have the same mean and standard deviation and even the same regression line equation. The existence of such datasets indicates that we should be careful while interpreting the results of linear regression and always look for visualization of the predicted and actual datasets to confirm the goodness of fit.

3. **What is Pearson's R? (3 marks)**
   Ans.: Pearson's R indicates the strength and direction of the relation between two continuous variables. R = 0 indicates no correlation between the two variables, while R = 1 indicates positive perfect linear correlation between the variables. R = -1 indicates negative perfectly linear correlation (i.e. one variable increase linearly as the other variable decreases).

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

   Ans: Scaling is the process of normalizing the data so that values of the different predictor (independent) variables become comparable in magnitude and no single variable is able to dominate the distance calculations of the regression algorithm. Scaling also speeds up the process of minimizing the cost function and the algorithm can converge to the solution at a faster rate.

   Normalized scaling scales the data in the range of 0 to 1. On the other hand, the standardized scaling method scales the data as per the normal distribution (but there is no restriction on the maximum and minimum values, so it is not the preferred scaling algorithm).

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

(3 marks)

Ans.: The value of VIF increases as the relation between two predictor variables becomes more and more linear. So, VIF will tend to infinity if the correlation between the variables is perfectly linear (Pearson R = +1 or -1). In such case, a straight-line relationship can be found between the two predictor variables, and one of the variables can be replaced by a linearly scaled version of the other predictor variable.

6**. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**

(3 marks)

Ans.: The quantile - quantile plot is a scatter plot of 2 sets of quantiles (one set is the test dataset and the other set is a dataset from a theoretical distribution (like normal or other standard distributions). If the plot is close to a straight line, then it indicates that the test dataset follows the same theoretical distribution as used for the QQ plot. Hence the QQ plot helps in determining the underlying distribution of the dataset.