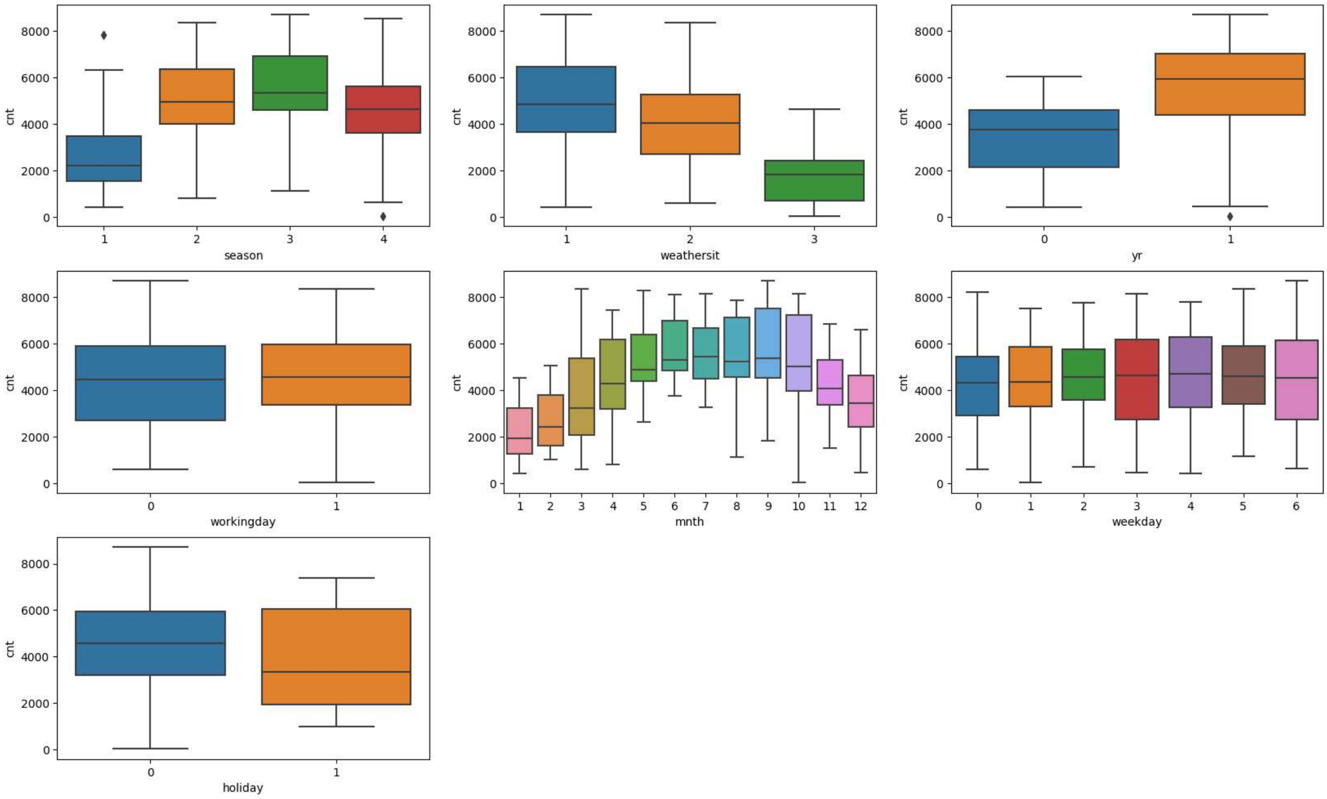
Assignment-based Subjective Questions

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:- Below box plots are created for categorial variables :

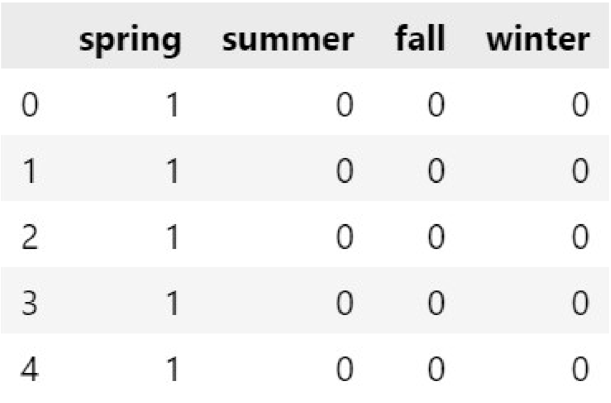


From this, it can be predicated that weathersit and yr have huge impact on cnt, while workingday didn't impact much.

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

Ans: - drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Ex: Incase of season, below dummy variables are created: -

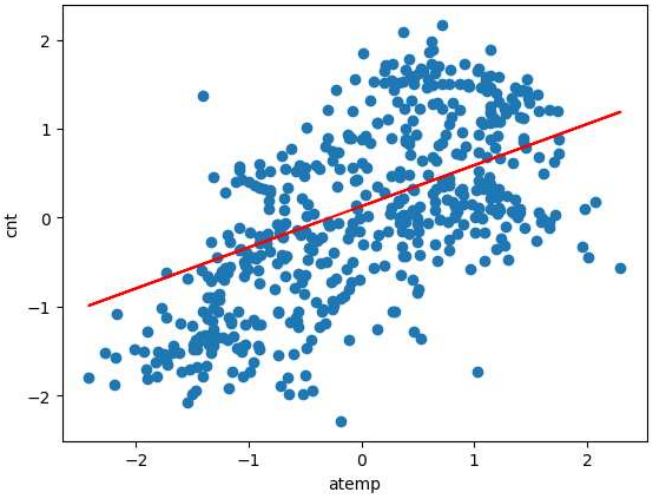


	spring	summer	fall	winter
0	1	0	0	0
1	1	0	0	0
2	1	0	0	0
3	1	0	0	0
4	1	0	0	0

Now, we don't need 4 columns. You can drop the `1` column, as the type of season can be identified with just the last 3 columns where —
- `000` will correspond to `1(spring)`
- `100` will correspond to `2(summer)`
- `010` will correspond to `3(fall)`
- `001` will correspond to `4(winter)`
So, using drop-first=True to drop 1(spring).

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

Ans: - atemp has the highest correlation with the target variable cnt



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

Ans: Validation of assumptions of linear regression can be done by either plotting scatter plot between 2 variables or using VIF (Variance Inflation Factor). VIF<=5 implies no multicollinearity, so it's better.

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: atemp(feeling temperature in Celsius), temp(temperature in Celsius) and season – 3(fall) are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions

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

Ans: Linear regression tells the relationship between one or more feature variables and one target variable. Linear regression is commonly used for predictive analysis and modelling. For example, in bike sharing system, target variable is cnt (count), so how other variables like temp, atemp, etc. can affect the demand(cnt), can be explained by linear regression.

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

Ans : Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points.

3.What is Pearson’s R? (3 marks)

Ans : The Pearson correlation coefficient also known as Pearson's R, is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations.Thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

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

Ans: When you have a lot of independent variables in a model, a lot of them might be on very different scales, having different range of values, ex : atemp and hum has different range of values, this will lead a model with very weird coefficients that might be difficult to interpret. So, we need to scale features to ease of interpretation.

- Normalization/Min-Max Scaling -> It brings all the data in the range of 0 and 1.
- Standardization Scaling -> Standardization replaces the values by their Z scores. It brings all the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

In this assignment, I tried to do scaling using normalization as well standardization, however result is more appropriate using standardization.

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

Ans: If there is perfect correlation, then VIF = infinity. This 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. In this assignment, it is observed that Thurs, Fri, Mon, holiday, Wed, Tue and workingday have infinite VIF. So, few of these variables are dropped.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

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

Ans: Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical. Also, it helps to determine if two data sets come from populations with a common distribution. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.