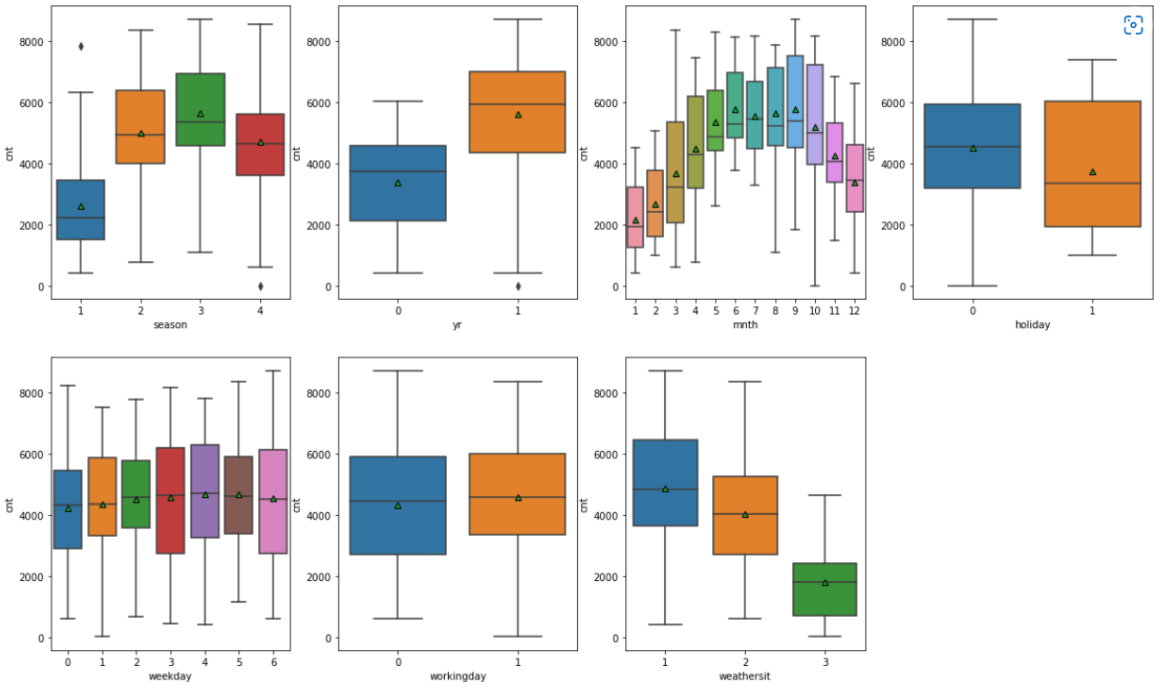
Assignment-based Subjective Questions

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

   Answer:

   The dataset contains the following categorical variables: ‘season' 'yr' 'mnth' 'holiday', 'weekday', 'workingday', and 'weathersit'.

   A box plot of this against target variable ‘cnt’ is below.



   Data shows below inferences on the categorical values against the total rentals

   - There is a significant difference in the rentals on different seasons. Highest rentals are on fall and lowest on sprint.
   - Rentals increased on year over year.
   - Rentals show similar patterns with seasons. Highest rentals are between June and September. Lowest on January.
   - Higher rentals on non-holidays.
   - There is not much difference on rentals on different days of the week.
   - There is not much difference on rentals on working day and non-working day.
   - There is a gradual decrease in rentals based on the bad weather. No rentals and weather situation category 4.

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

   Answer:

Dummy variable creation by default creates one variable per category for a categorical variable. For example, a categorical variable with three options, it creates three variables. Here only two variables are sufficient to represent all the categorical values for the variable. Hence, as a matter of brevity, it is important to use the option drop_first=True, which will drop the first generated variable and keep the rest of the variables as part of 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)
   Answer:
   atemp has highest correlation.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   Answer:
   Validated the assumptions of Linear regression are all met with below steps:
   - Checked highest contributing attributes has linear relation with the dependent variable.
   - Checked no autocorrelation in residuals. DW value is 2.055 which is closest to the ideal value 2.
   - Checked no hetroskedasticity.
   - Checked multicollinearity values are at moderate level by listing VIF value and fixing the values VIF <5.
   - Checked the residuals are normally distributed.

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

   Answer:

   The final model is:

   Y = 0.2911 + 0.2416 * yr_2019 - 0.0701 * holiday + 0.4089 * atemp - 0.1647 * windspeed - 0.1391 * spring - 0.0628 * weather_mist

The features with highest coefficient are atemp, yr_2019, and windspeed. Hence these features significantly explain the demand of the shared bikes.

General Subjective Questions

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

Linear regression algorithm is a statistical method to analyse the linear relationship between a variable and one or more variable. The target variable is called dependent variable and the variable/variables used find the target variable is called dependent variable/variables.

There are two types of linear regression: 1) Simple Linear Regression (SLR) and 2) Multiple linear regression (MLR).
The difference is in SLR there would be a single independent variable involved and in MLR two or more independent variables involved.

The linear relation can be positive or negative, means an increase in independent variable cause an increase in dependant variable in a positive linear relationship where as an increase in independent variable cause a decrease in dependant variable for negative relationship.

SLR can represented mathematically as $Y = mX + b$. Here, Y is dependant variable , X is independent variable and m is the coefficient and b is the constant/intercept.

MLR can be represented mathematically as $Yi = b0 + b1xi1 + b2xi2+ \ldots + bnxin + E$. Here , yi is the dependant variable, bo is the intercept, b1..bn are the coefficient and xi1..xin are independent variables.

2. Explain the Anscombe's quartet in detail. (3 marks)
   Statistician Francis Anscombe constructed a quartet to illustrate the importance of plotting data before build model or do some analysis. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.

3. What is Pearson's R? (3 marks)
   The Pearson Correlation Coefficient is the measure of change in one variable on another variable. The value range between -1 to 1. When the Pearson's R value is 0, then it implies that the variables has not correlation. Pearson's $R > 0$ means there is a positive correlation, or when one value increase, other also increase. Pearson's $R<0$ means there is a negative correlation or when once value increases other value decrease. The value 1 is the highest it can take either positive or negative way.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
   Scaling is a method of bringing the values of all variables withing the same range.
   Scaling brings the interpretation easy, and helps faster convergence of gradient descent.
   Scaling does not change model accuracy, p-values, Fstatistics, R-squared etc but just affects the coefficients.

There are two scaling methods: normalization and standardization.

Normalization brings the variables with in the range of zero to 1. The formula is (x – xmin)/(xmax -xmin).

Standardization makes the mean 0 and standard deviation 1. The formula is (x – xmean)/sd(x)

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

Variance Inflation Factor is used to detect the multicollinearity in regression analysis.
The VIF uses the formula 1/(1-R2). Hence, when R2 is 1 the VIF becomes infinite. This means there is a perfect correlation between two or more independent variables. The way to solve this is to drop the variable from your mode.

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

Quantile-Quantile uses to determine two set of data is from population with common distribution. This also helps to find the two data sets has common location and scale, similar distributional shape, similar tail behaviour.