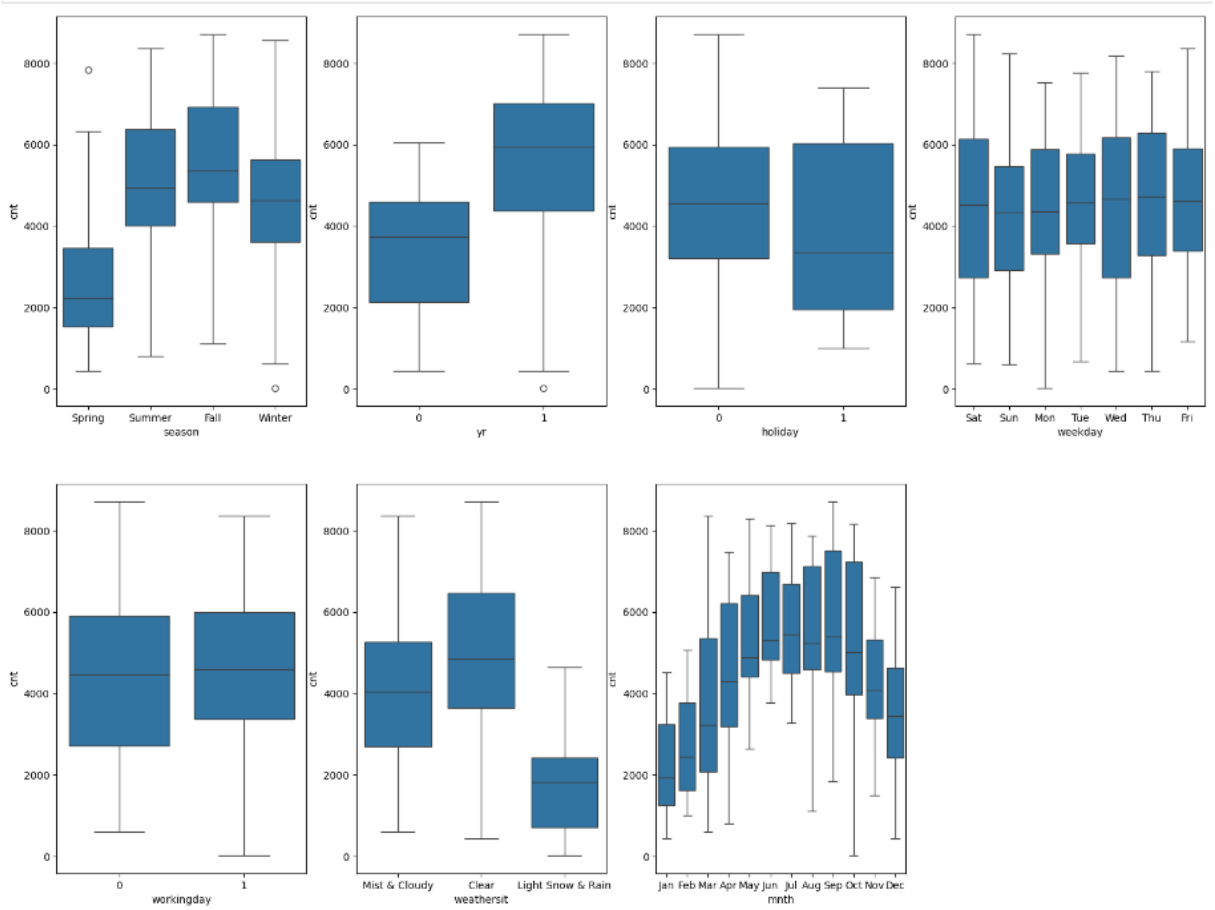
Assignment-based Subjective Questions

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



The categorical variables were visualized using boxplots. The inferences from these variables are:

i. Season - For the variable Season we can see that Fall has the highest median, which means the demand was high during this season. It is the lowest in 1: Spring.

ii. Yr - The count of users is more in the year 2019 than 2018.

iii. Holiday - For the variable Holiday, the median is higher on the days when it is not a holiday.

iv. Weekday - The count of users is almost same throughout the week as observed for the variable Weekday.

v. WorkingDay - From WorkingDay we can see that the median is almost similar. So the count of users was not affected whether it was working day or not.

vi. WeatherSit - From the variable WeatherSit, we can see that there are no users for heavy rain. And the count is higher when the day is clear.

vii. Mnth - The number of rentals peaked in September. This observation is consistent with the observations made regarding the weather. As a result of the typical substantial snowfall in December, rentals may have declined.

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

   When you have a categorical variable with say 'n' levels, the idea of dummy variable creation is to build 'n-1' variables, indicating the levels. For a variable say, 'Relationship' with three levels namely, 'Single', 'In a relationship', and 'Married', you would create a dummy table like the following:

Relationship Status	Single	In a relationship	Married
Single	1	0	0
In a relationship	0	1	0
Married	0	0	1

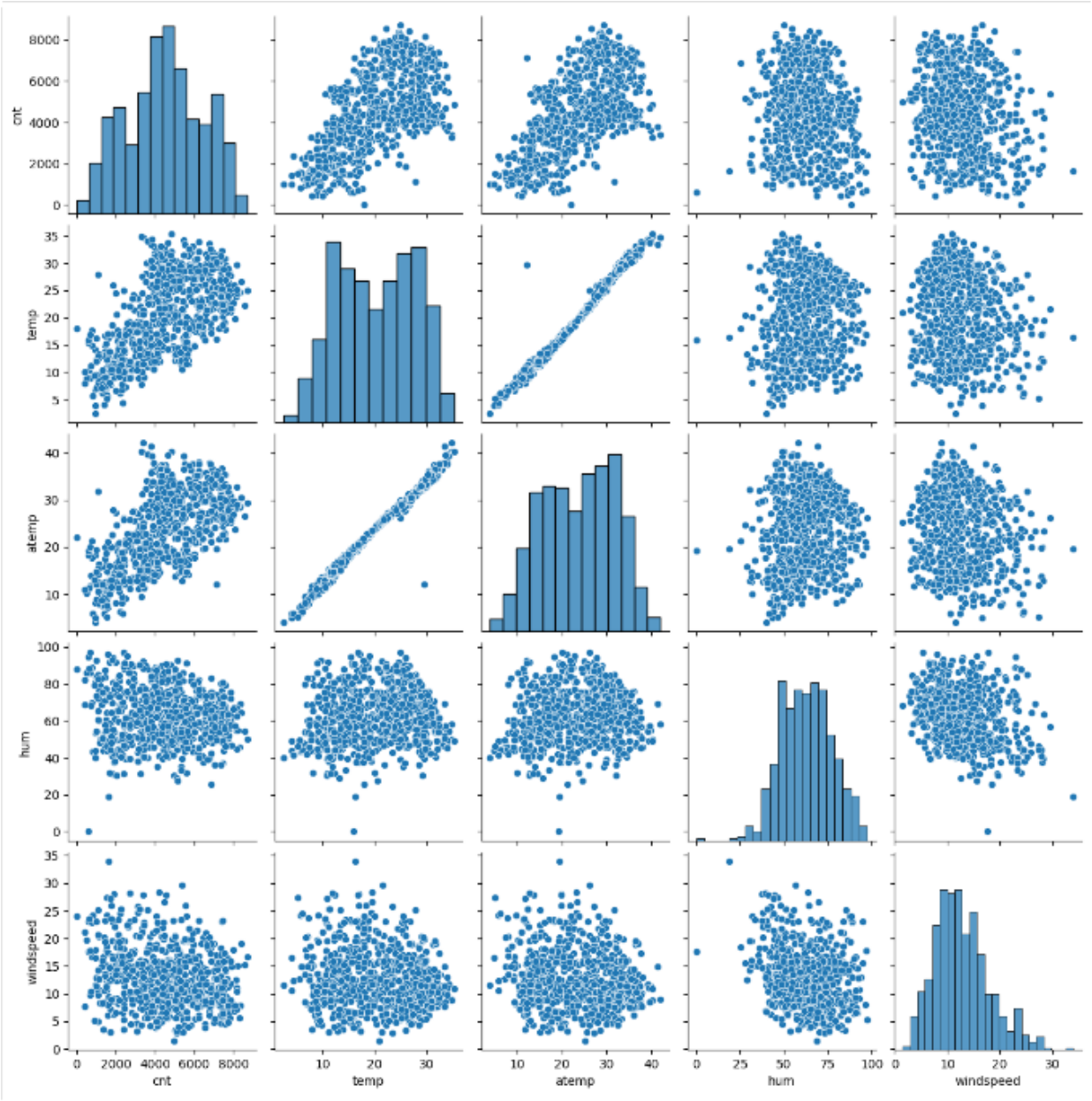
   But you can clearly see that there is no need of defining **three** different levels. If you drop a level, say 'Single', you would still be able to explain the three levels.

   Let's drop the dummy variable 'Single' from the columns and see what the table looks like:

Relationship Status	In a relationship	Married
Single	0	0
In a relationship	1	0
Married	0	1

   If both the dummy variables namely 'In a relationship' and 'Married' are equal to zero, that means that the person is single. If 'In a relationship' is one and 'Married' is zero, that means that the person is in a relationship and finally, if 'In a relationship' is zero and 'Married' is 1, that means that the person is married.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   'temp' and 'atemp' variables have 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?**
   i. One of the most important assumptions is that a linear relationship is said to exist between the dependent and the independent variables. We validated it by creating a pairplot x vs y. If the data points fall on a straight line in the graph, there is a linear relationship between the dependent and the independent variables, and the assumption holds.
   ii. Residuals distribution should follow normal distribution and centered around 0 (mean = 0). We validated this assumption about residuals by plotting a distplot of residuals and saw if residuals are following normal distribution or not.
   iii. The independent variables shouldn't be correlated. If multicollinearity exists between the independent variables, it is challenging to predict the outcome of the model. We validated it by calculating the VIF to quantify how strongly the features are associated with each other.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   The based on the final model, top three features contributing significantly towards explaining the demand are:
   - Temprature (0.490988)
   - weathersit_Light Snow & Rain (-0.284199)
   - Year (0.233570)

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear regression is a machine learning algorithm based on supervised learning A simple linear regression model attempts to explain the relationship between a dependent and an independent variable using a straight line.
The independent variable is also known as the **predictor variable**. And the dependent variables are also known as the **output variables**.

The basic idea of linear regression is to find a line that best fits the data points, such that the distance between the line and the data points is minimized.
The standard equation of the regression line is given by the following expression:
$\mathbf{Y = \beta_o + \beta_1 X}$
where β_o is the intercept and β_1 is the slope. These are called the parameters or coefficients of the linear model.

The best-fit line is found by minimizing the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable.

The strength of the linear regression model can be assessed using 2 metrics:

i. R^2 or Coefficient of Determination
ii. Residual Standard Error (RSE)

Limitations are –

i. It assumes a linear relationship between the input variables and the output variable, which may not always be the case.
ii. It may be sensitive to outliers or multicollinearity.

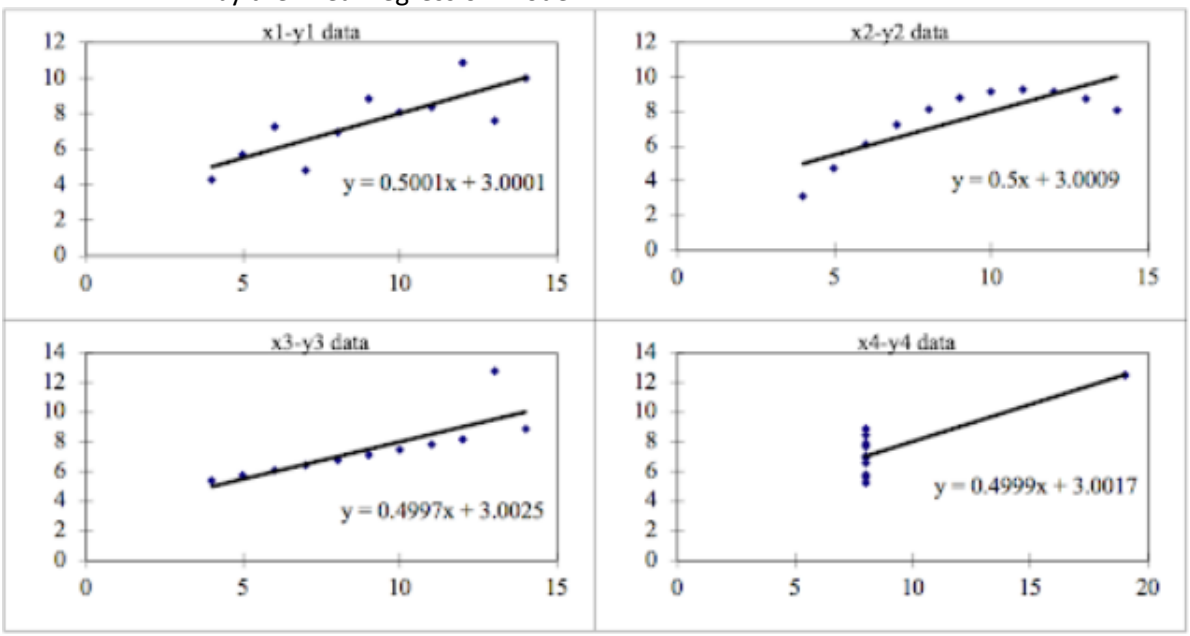
2. **Explain the Anscombe's quartet in detail.**

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.).

Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.

Anscombe's Quartet Four Datasets:

i. Data Set 1: It fits the linear regression model well.
ii. Data Set 2: It cannot fit the linear regression model because the data is non-linear.
iii. Data Set 3: It shows the outliers involved in the dataset, which cannot be handled by the linear regression model.
iv. Data Set 4: It shows the outliers involved in the data set, which also cannot be handled by the linear regression model.



3. What is Pearson's R?

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.

Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Pearson correlation coefficient (*r*)	Correlation type	Interpretation	Example
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is **no relationship** between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.	Elevation & air pressure: The higher the elevation, the lower the air pressure.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data pre-processing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

There are two major methods to scale the variables, i.e. standardization and MinMax scaling. Standardization basically brings all of the data into a standard normal distribution with mean zero and standard deviation one. MinMax scaling, on the other hand, brings all of the data in the range of 0 and 1.

- Standardisation: $x = \frac{x - mean(x)}{sd(x)}$
- MinMax Scaling: $x = \frac{x - min(x)}{max(x) - min(x)}$

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

   If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

   The VIF is given by:

$$VIF_i = \frac{1}{1-R_i^2}$$

where 'i' refers to the i-th variable which is being represented as a linear combination of rest of the independent variables. You'll see VIF in action during the Python demonstration on multiple linear regression.

The common heuristic we follow for the VIF values is:

- **10:** Definitely high VIF value and the variable should be eliminated.
- **5:** Can be okay, but it is worth inspecting.
- **< 5:** Good VIF value. No need to eliminate this variable.

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

   A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It is used to compare the shapes of distributions. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. The Q-Q plot is used to answer the following questions:

   - Do two data sets come from populations with a common distribution?
   - Do two data sets have common location and scale?
   - Do two data sets have similar distributional shapes?
   - Do two data sets have similar tail behavior?