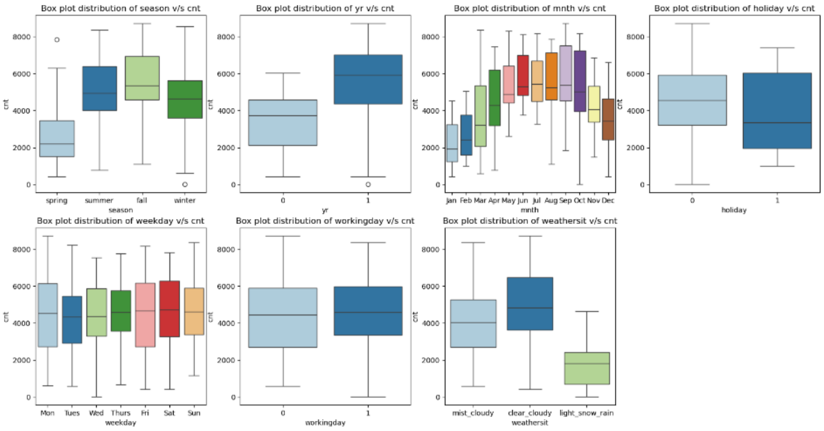
Assignment-based Subjective Questions

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

   The different categorical variables, season, year, month, holiday, weekday, workingday and weathersit have different impact on the dependent variable. The plot below shows the distribution of their values based on the bike demand cnt. The major categorical variables which can be used to predict bike demand are:

   - yr: Increase in year, implies increase in bike sharing demand.
   - season:
     - winter: When the season is winter, it implies increase in bike sharing demand.
     - summer: When the season is summer, the bike sharing demand is higher but it has lesser positive impact in comparison to winter season
     - spring: When the season is spring, it implies decrease in bike sharing demand.
   - weekday: When the day is Monday, the bike sharing demand increases.
   - month: When the month is September, the bike sharing demand increases; when it is July, Nov, Jan the bike sharing demand decreases.
   - weathersit: When the weather is “Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist” or “Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds” the bike sharing demand decreases.



2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
   To quantify the categorical variables, we create dummies for them. In the case of a categorical variable with n levels, the fact that we can predict what the final value for the variable will be using (n-1) levels instead of n levels, helps in avoiding redundant information and redundant variables in the model.
   For example,
   If a categorical variable grade has three possible values: A, B and C, the final value can be predicted as below:

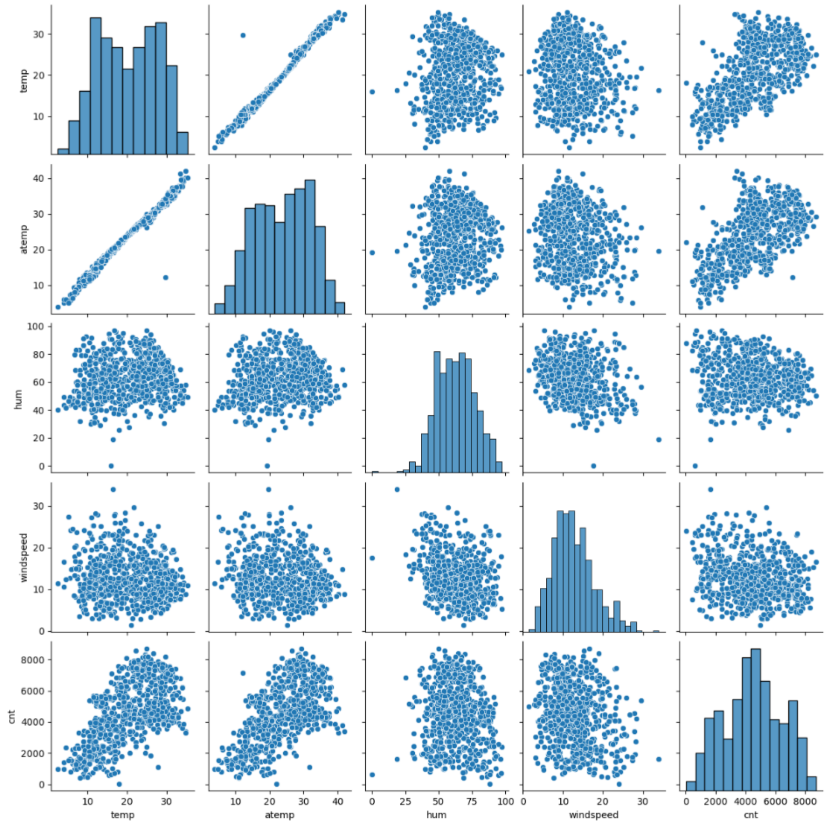
A	**B**	**C**	**Final Value**
1	0	0	A
0	1	0	B
0	0	1	C

The above final value can also be predicted without actually having the dummy level A since when B and C are 0 or False, it implies the value of the variable is A.

B	**C**	**Final Value**
0	0	A
1	0	B
0	1	C

In effect, we can drop any of the n levels, and keep in total (n-1) levels for the categorical variable. In pandas, the get_dummies function supports drop_first parameter which drops the first level to create dummies. Hence it is convenient to create (n-1) levels for an n level categorical variable using drop_first = True.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   The highest correlation with the dependent variable cnt for bike sharing demand is of the numerical variable “temp” and “atemp”. The correlation is positive in nature for temp v/s cnt and atemp v/s cnt.



4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   Linear Regression models can be validated based on:

- Linear relationship between X (independent variables) and Y (dependent variable)
- Error terms are normally distributed (not X, Y)
- Error terms are independent of each other
- Error terms have constant variance (homoscedasticity)

The Linear Regression model built on the training set was verified against the above assumptions and it satisfied all the conditions. Hence, we can conclude that the linear relationship exists between the independent and 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)

   The top 3 features impacting the demand of shared bikes are:

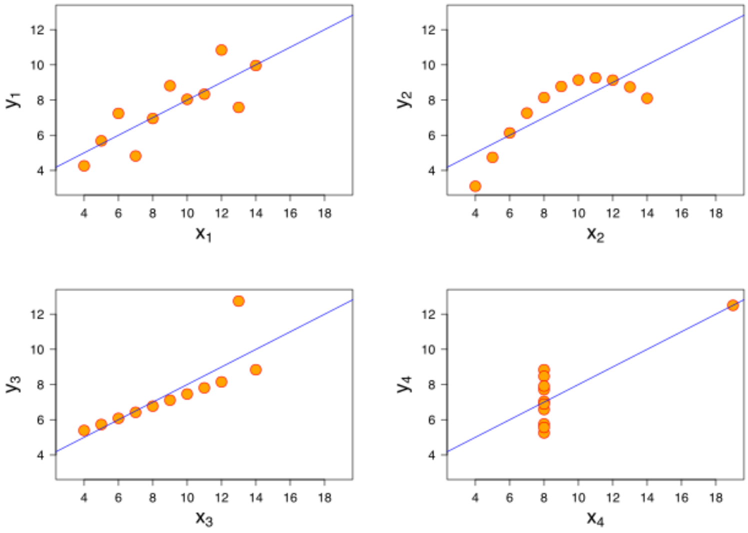
   - temp: Positively correlated with the bike sharing demand
   - yr: Positively correlated with the bike sharing demand
   - season: Winter and summer have positive impact on the bike sharing demand, while spring has negative impact on the bike sharing demand

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   - Linear regression is a supervised machine learning algorithm that can be used to identify the relationship between a dependent variable (target) and one or more independent features variables (predictors).
   - It involves finding the best-fitting line that minimizes the differences between predicted and actual values for the target variable.
   - The equation for linear regression can be represented as:
     $y = \beta_0 + \beta_1X_1 + \beta_2X_2 + … + \beta_nX_n$
     where y is the target independent variable,
     $X_1, X_2, …, X_n$ are the dependent feature variables,
     $\beta_1, \beta_2, … + \beta_n$ are the coefficients of the dependent feature variables $X_1, X_2, …, X_n$ respectively
     β_0 is the intercept
   - Examples:
     - Housing Price Prediction: The prediction of the price of a house based on various input feature variables like number of bedrooms, area in square feet, locality etc is an example of a linear regression problem.
     - Sales Prediction: The prediction of the sales of a company based on various marketing mediums for advertisement, TV, radio, newspaper etc is an example of a linear regression problem.
   - Assumptions for Linear Regression model:
     - Linearity: There should be a linear relationship between the independent variable X_i's and the dependent variable y.
     - Independence: Residuals (differences between predicted and actual values) should be independent.
     - Normality: The residuals of the model should follow a normal distribution.

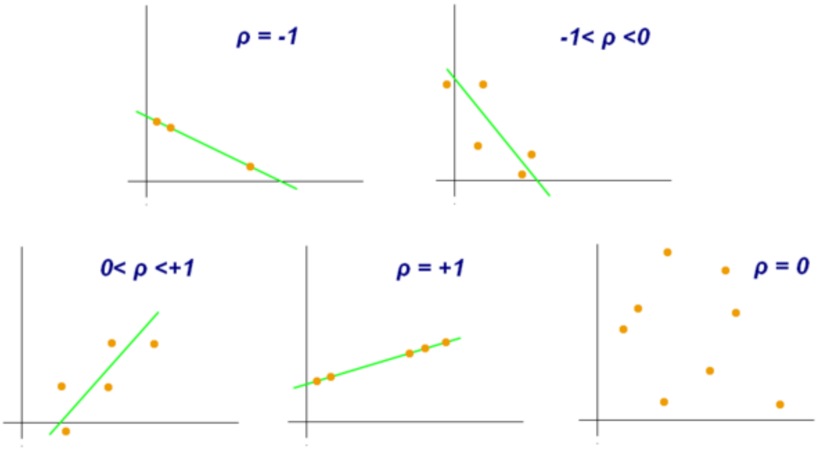
- Homoscedasticity: The variance of residuals should be constant across all levels of X_i's.

2. Explain the Anscombe's quartet in detail. (3 marks)
   - Anscombe's quartet consists of four distinct datasets.
   - These datasets have nearly identical descriptive statistics such as mean, standard deviation, and regression line.
   - It emphasizes the significance of visualizing data before analysing it statistically since quite different distributions can have the same statistical values.
   - It highlights the importance of data visualization over direct use of statistical values.
   - The 4 distributions:
     i. The dataset represents linear relationship and can be modelled using linear regression.
     ii. The dataset does not represent linear relationship and cannot be modelled using linear regression.
     iii. The dataset consists of outliers that can disrupt the fit of a linear model.
     iv. The dataset has even more outliers making it unsuitable to fit and predict for a linear regression model.



3. What is Pearson's R? (3 marks)
   - Pearson's R is a numerical summary of the strength of the linear association between the variables.
   - If the variables tend to go up and down together, the correlation coefficient will be positive.
   - If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.
   - The values of Pearson's R can range from -1 to +1.
   - Pearson's R value significance:
     - When the value is between -1 to 0, it implies negative association between the variables.

- When the value is 0, it implies there is no association between the two variables.
- When the value is between 0 to +1, it implies positive association between the variables.



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes of feature variable values. A feature having high magnitude values might have smaller coefficient than the feature having values in smaller ranges, even though their impact might be same or even opposite i.e. one having smaller coefficient might have more impact than one with larger coefficient. Scaling is done to bring all the feature variables in a similar range so that their coefficients can be used to make inferences of their impact on the target variable.

Normalized Scaling	**Standardized Scaling**
Minimum and maximum value of features are used for scaling.	Mean and standard deviation is used for scaling.
Affected by outliers.	Less impact of outliers.
Scales values between [0, 1] or [-1, 1]	Scale values are not bounded by range. It ensures zero mean and unit standard deviation.
sklearn's MinMaxScaler can be used normalized sclaing.	sklearn's StandardScaler can be used for standardized scaling.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
VIF = infinity implies a very high correlation and it implies the respective variable should definitely be dropped from the model and retrained to reduce multicollinearity in the model. VIF is represented the formula mentioned below:

$$\boxed{\begin{array}{l} VIF_i = 1 / (1-R_i)^2 \\ \text{where i refers to the } i^{th} \text{ variable.} \end{array}}$$

If R squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

In general, Variance Inflation Factor (VIF) helps to identify the variance, the relationship of one independent variable with all the other independent variables. The acceptable VIF values below 5 are acceptable in industry. VIF value above 5 and below 10 means that VIF is high and should be inspected. VIF > 10 implies that the respective variable whose VIF is high is highly correlated with any of the independent variables and dropping the same can help improve the model and avoid multicollinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
   - Quantile-Quantile (Q-Q) plot, is a graphical technique to determine if a dataset follows a certain probability distribution or whether two samples of data are coming from the same population or not. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution.
   - A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate).
   - If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the identity line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.
   - Advantages of Q-Q plot:
     - Q-Q plots can be used to compare datasets of different sizes without requiring equal sample sizes.
     - Q-Q plots are dimensionless, making them suitable for comparing datasets with different units or scales.
     - They provide a clear visual representation of data distribution compared to a theoretical distribution.
     - The plots detect deviations from assumed distributions, helping in identifying data discrepancies.
     - Q-Q plots help in assessing distributional assumptions, identifying outliers, and understanding data patterns.