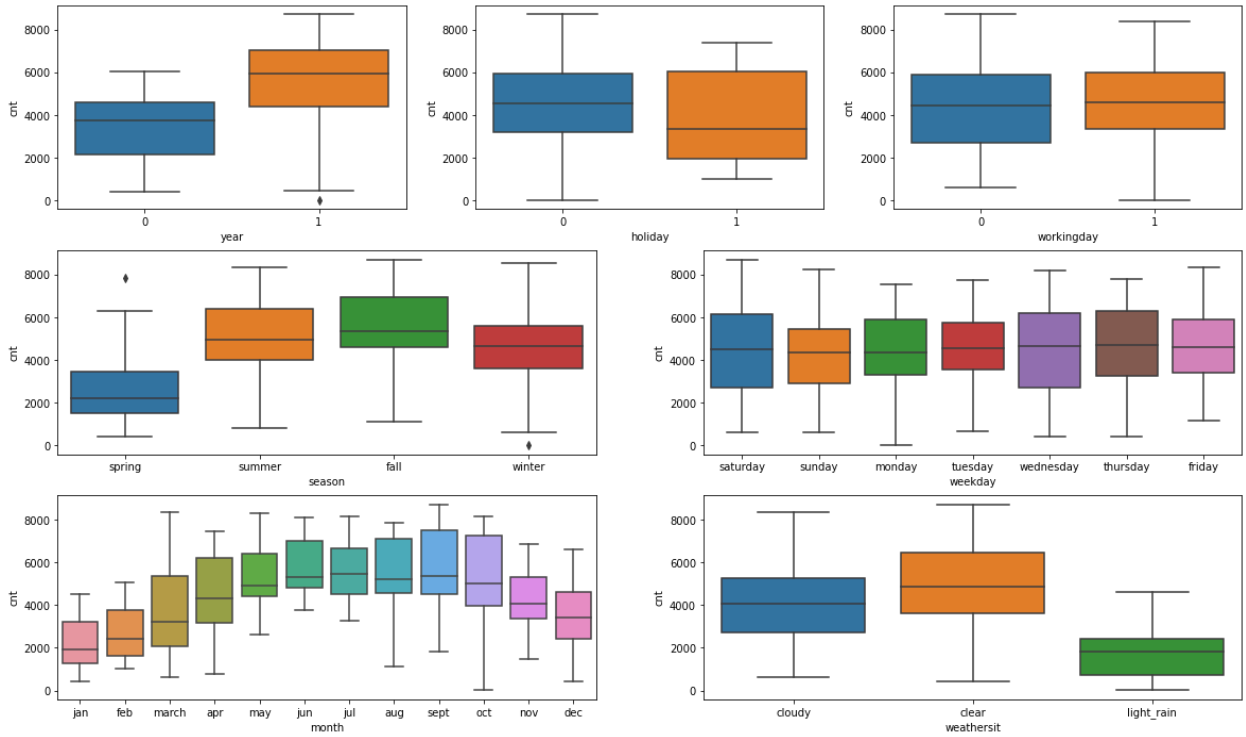
Assignment-based Subjective Questions

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:



   a. For the `year` 2019, usage of bike sharing is more than previous year.
   b. For `holidays` spread between the 25th percentile and 75th percentile is more.
   c. For `workingday` and `weekends` 50th percentile is almost same.
   d. During summer, fall and winter `season` usage of bike sharing seems more than spring.
   e. Usage of shared bikes is more during `month` march to oct.
   f. When there is heavy_rain no one is using shared bikes and for light_rain usage seems to be decreased than cloudy and clear `weathersit`.

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

   When we use the get_dummies method it creates a dummy column for each n-categorical value from the column. As each row will have only one of the categorical value from that column, to represent it in newly created dummy columns, for each row any one of the dummy columns will have value 1 for the category it represents and rest dummy columns will have value 0 for that row. But

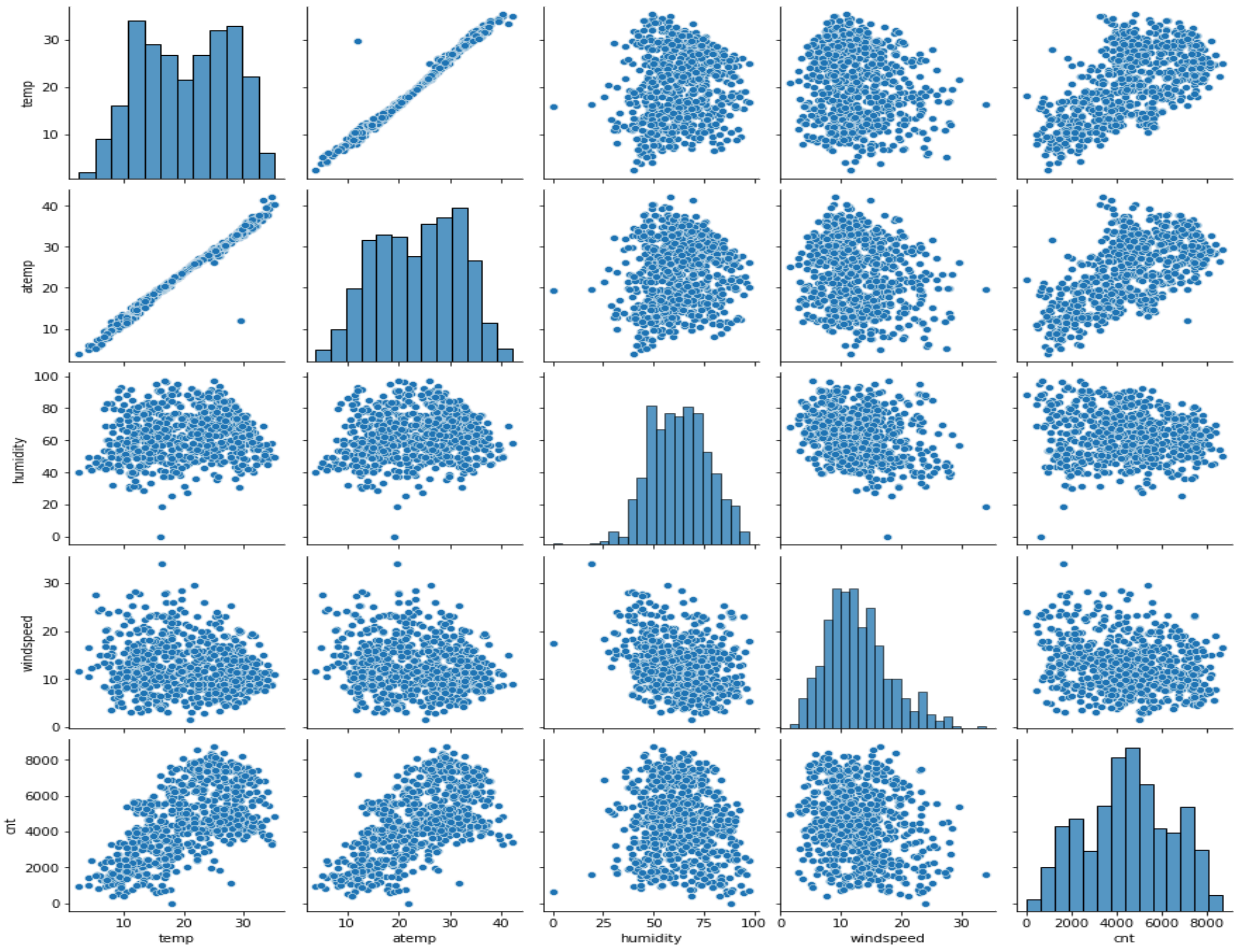
if we drop one of the column one category can be identified with all dummy categorical columns having value 0 and instead of using n column we can reduce it to n-1 columns.
For example:

Season:

with n features	**with n-1 features**
1000: spring	000: spring
0100: summer	100: summer
0010: fall	010: fall
0001: winter	001: winter

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

   'temp' and 'atemp' 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? (3 marks)**
   Answer:
   I have validated the assumption of Linear Regression model based on below 5 assumptions:
   - Normality of error terms
     - Error terms should be normally distributed
   - Multicollinearity check
     - There should be insignificant multicollinearity among variables
   - Homoscedasticity
     - There should be no visible pattern in residual values
   - Independence of residuals
     - No auto-correlation

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:
   Following are the top 3 features contributing significantly towards explaining the demand of the shared bikes:
   1. year
   2. workingday
   3. windspeed

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**
   Answer:
   Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data.
   When there is only one independent feature, it is known as Simple Linear Regression, and when there are more than one feature, it is known as Multiple Linear Regression.
   Similarly, when there is only one dependent variable, it is considered Univariate Linear Regression, while when there are more than one dependent variables, it is known as Multivariate Regression.

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

   Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

   The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

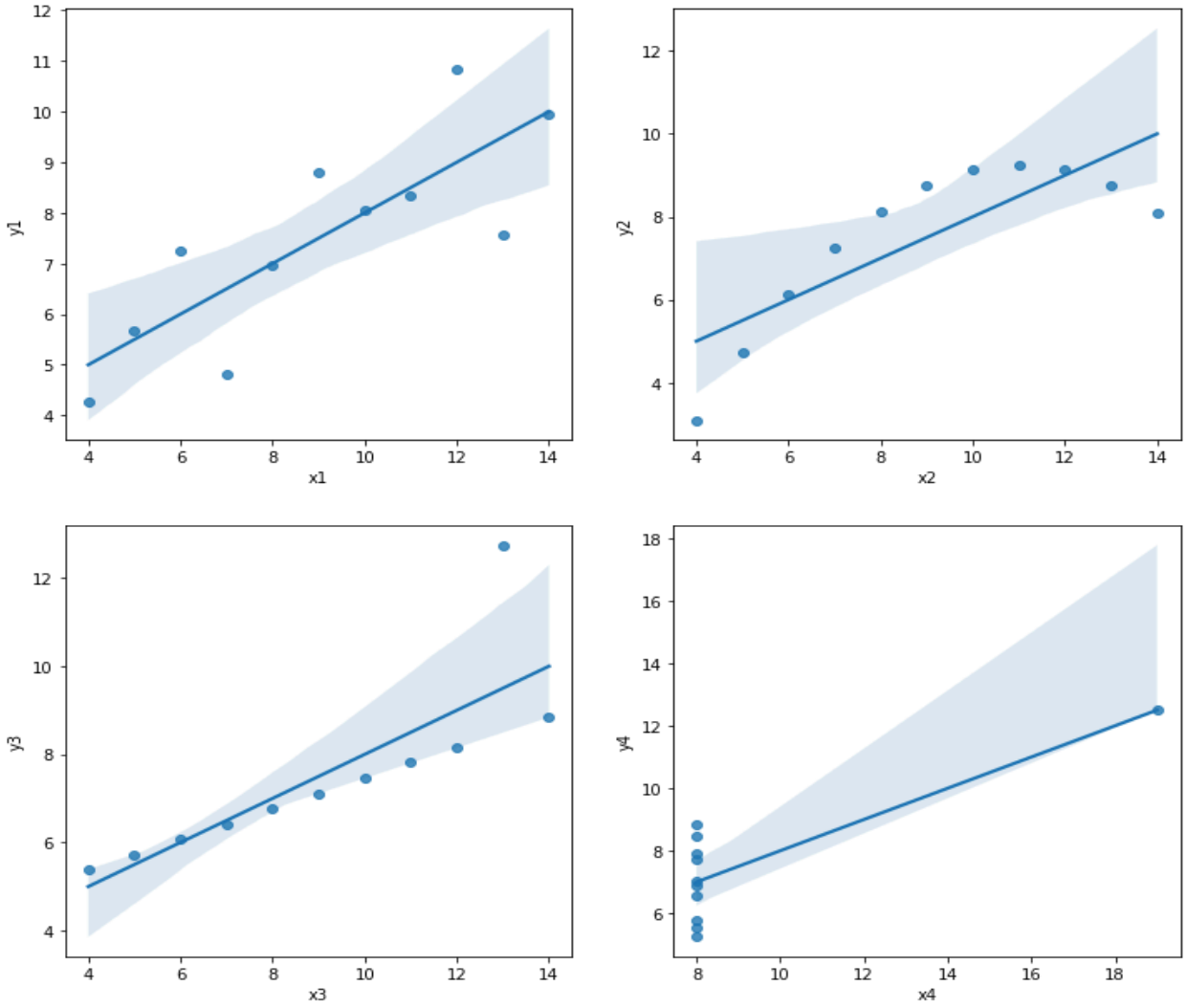
   The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

   Following is the observations for above data

	x1	x2	x3	x4	y1	y2	y3	y4
	10	10	10	8	8.04	9.14	7.46	6.58
	8	8	8	8	6.95	8.14	6.77	5.76
	13	13	13	8	7.58	8.74	12.74	7.71
	9	9	9	8	8.81	8.77	7.11	8.84
	11	11	11	8	8.33	9.26	7.81	8.47
	14	14	14	8	9.96	8.1	8.84	7.04
	6	6	6	8	7.24	6.13	6.08	5.25
	4	4	4	19	4.26	3.1	5.39	12.5
	12	12	12	8	10.84	9.13	8.15	5.56
	7	7	7	8	4.82	7.26	6.42	7.91
	5	5	5	8	5.68	4.74	5.73	6.89
sum	99	99	99	99	82.51	82.51	82.5	82.51
avg	9	9	9	9	7.5	7.5	7.5	7.5
stdev	3.32	3.32	3.32	3.32	2.03	2.03	2.03	2.03

- Mean of X is 9 and Y is 7.5 for each dataset.
- Stdev of X is 3.32 and Y is 2.03 for each dataset.

When we plot these four datasets on x and y axis data points shows different patterns.



- 1st data set fits linear regression model as it is showing linear relationship between X and Y
- 2nd data set does not show any linear relationship between X and Y, and data points are not normally distributed
- 3rd data set shows linear relationship between X and Y. but due to outlier there is not best fit for linear regression model
- 4th data set has a high leverage point means it produces a high correlation coefficient.

As we can see, the above datasets have the same statistical summary but different visualization and may not always have the best fit linear regression model.

3. **What is Pearson's R? (3 marks)**
   Answer:

   The Pearson correlation coefficient (r) also known as Pearson's 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.

   - $r = 1$: Perfect positive linear relationship
   - $r = -1$: Perfect negative linear relationship
   - $r = 0$: No linear relationship

   The formula is:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

   where X_i and Y_i are individual data points, and $\bar{X}$ and $\bar{Y}$ are the means of the variables.

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

Scaling is a procedure through which we draw an object that is proportional to the actual size of the object. It is performed to convert a highly varying magnitude date to the same unit / specific scale. If scaling is not performed then the algorithm tends to weight high values magnitudes and ignore other parameters which will result in incorrect modeling.

Sr. no	Min-Max Scaling	Standardization
1	Data compressed between 0 to 1.	Not bound to specific range (mean = 0, std = 1).
2	Min and max values are used for scaling.	Mean and Standard deviation is used for scaling.
3	It is affected by outliers.	It is much less affected by outliers.
4	It is called scaling normalization.	It is called as Z score normalization.
5	It is used when features are at different scale.	It is used to ensure zero mean and unit standard deviation.

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

Answer:

VIF (Variance Inflation Factor) indicates multicollinearity of the independent variables. It explains the relationship of one independent variable with all other independent variables. If any feature has VIF >=5 we remove that feature from the model.

Formula for VIF:

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

As you can see, in the above formula if any independent variable has coefficient of determination (R^2) value equal or more than 0.8 which results VIF >=5. R^2 can have max value 1. R^2=1 means an independent variable is 100% correlated with all other independent variables, which results in VIF=infinite(∞) which is practically not possible. But in multilinear regression any feature having VIF very high is considered VIF is infinite.

example :

	feature	vif
3	temp	379.53
4	atemp	368.80
5	humidity	31.84

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

Answer:

The quantile-quantile(q-q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came 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. They are commonly used in statistics, data analysis, and quality control to check assumptions and identify departures from expected distributions.

Use of Q-Q plot:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below

the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.