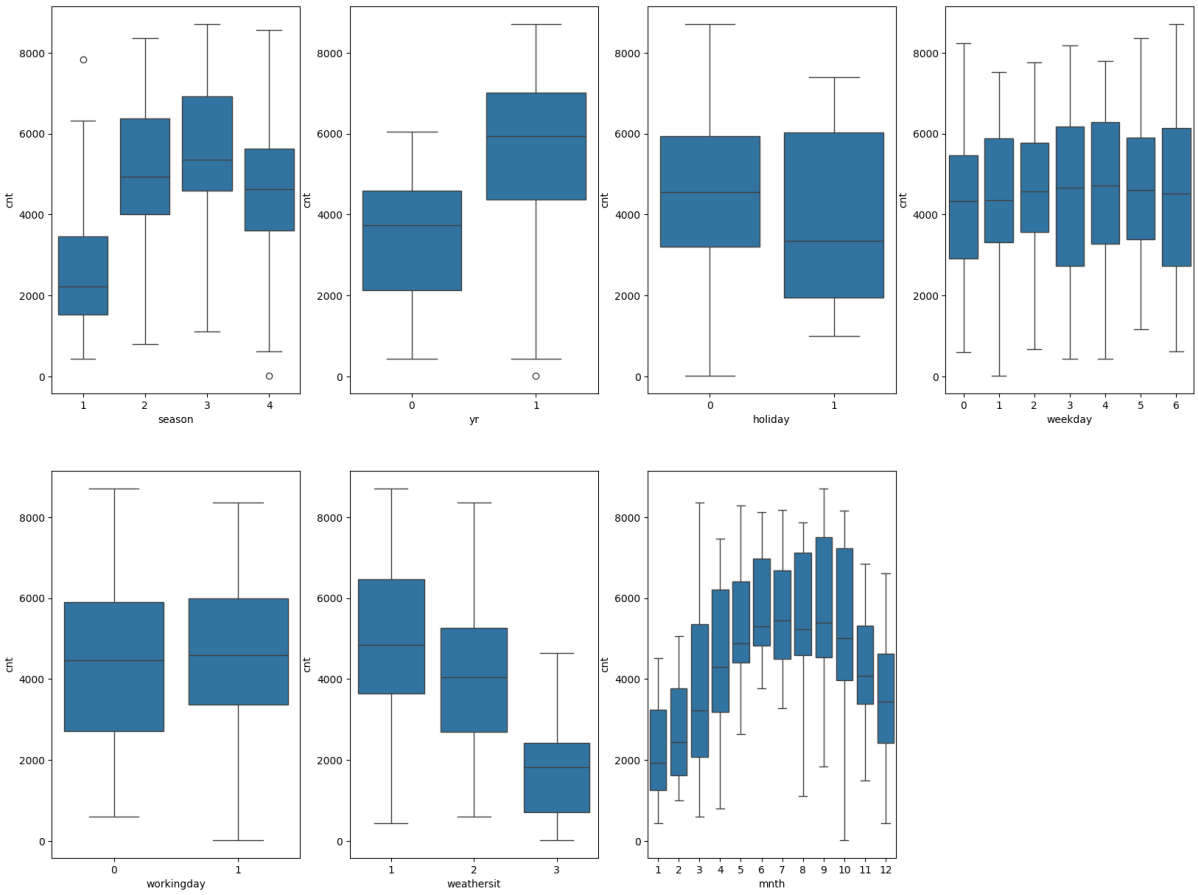
Assignment-based Subjective Questions

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



Categorical variables used in dataset: Season, Yr, Holiday Weekday, workingday, weathersit and mnth.

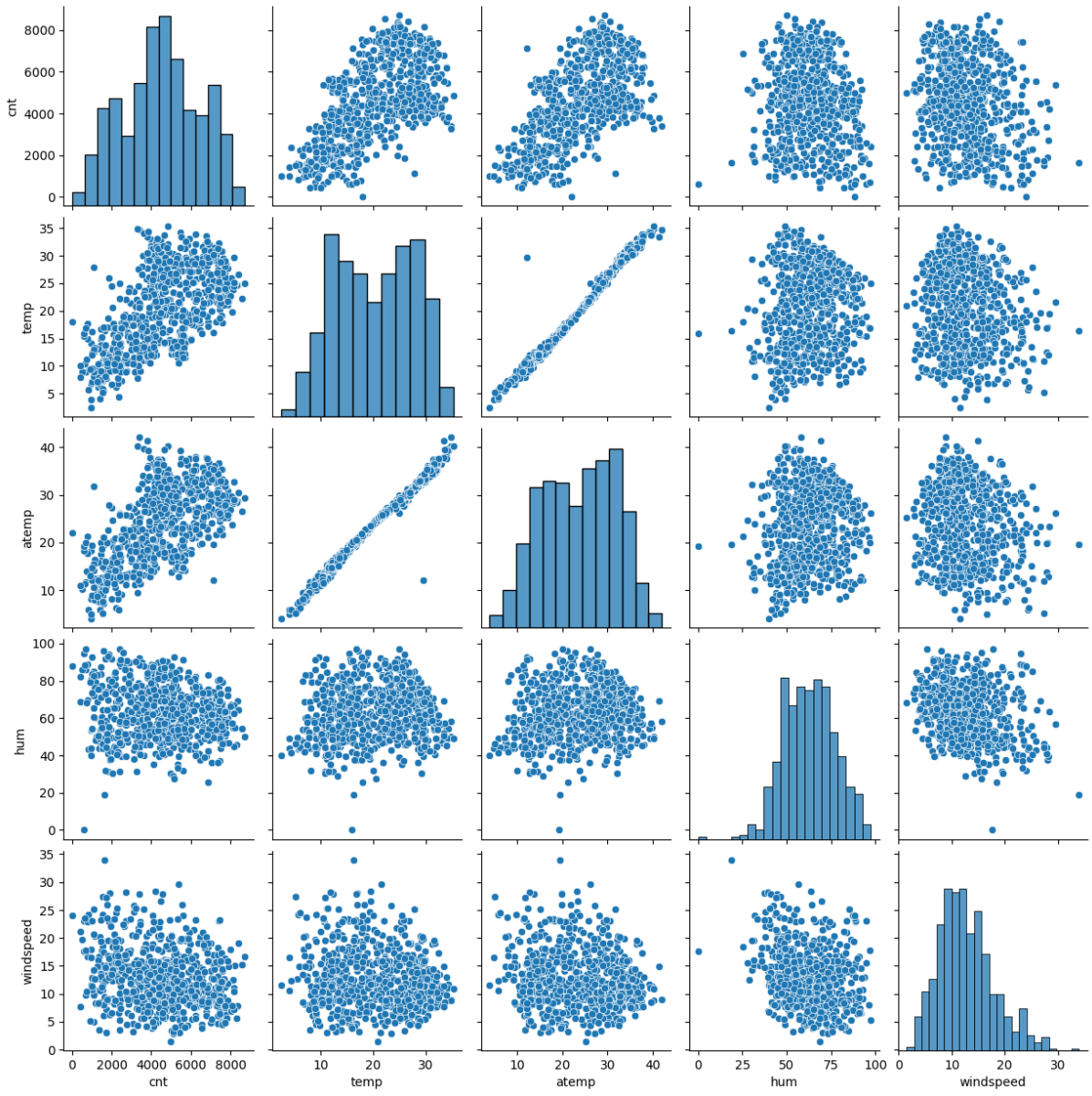
- Season: We can see that category 3: falls, Has highest median
- Yr – Year 2019 had higher counts compare to 2018
- Holiday – Rental reduced
- Weekday – Demand is almost constant
- Workingday – Maximum booking happened during 3000 – 4000. Median count of user is constant
- Weathersit – No users when there is heavy rain or snow
- Mnth – Highest count in September

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

Drop_first=True helps in reducing the extra columns created during dymmy variable creation. If we have categorical variable with n-levels, then we need to use n-1 columns.

Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

“Temp” and “Atemp” are highly correlated with target variable cnt



General Subjective Questions

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

Linear regression is a machine learning algorithm based on supervised learning. It performs a regression task, which means it predicts a continuous output variable (y) based on one or more input variables (x). It is mostly used for finding out the linear relationship between variables and forecasting.

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 line can be represented by an equation of the form:

$y = \theta 0 + \theta 1x$

where θ0 is the intercept (the value of y when x is zero) and θ1 is the slope (the change in y for a unit change in x). These are called the parameters or coefficients of the linear model.

To find the best values of θ0 and θ1, we need to define a cost function that measures how well the line fits the data. A common choice is the mean squared error (MSE), which is the average of the squared differences between the actual y values and the predicted y values:

$MSE = (1/n) * \sum(y - y')^2$

where n is the number of data points, y is the actual value, and y' is the predicted value.

The goal is to minimize the MSE by adjusting θ0 and θ1. There are different methods to do this, such as gradient descent, normal equation, or using libraries like scikit-learn.

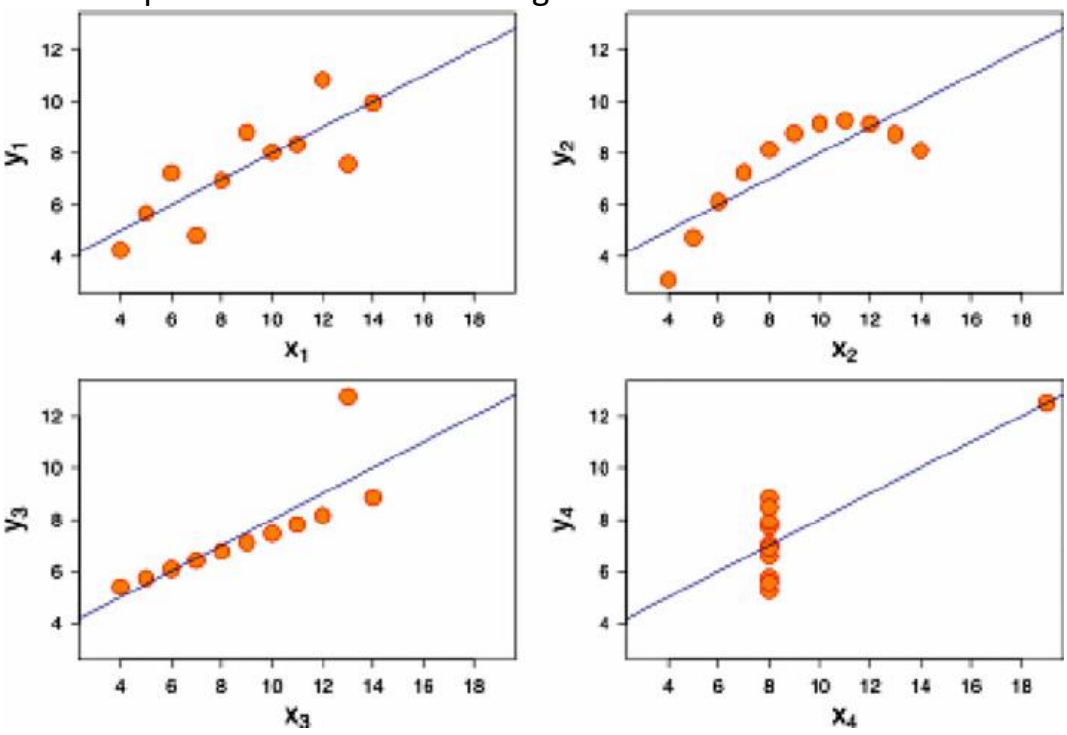
Linear regression can also be extended to multiple input variables (x1, x2, …, xn), in which case the equation becomes:

$y = \theta 0 + \theta 1x1 + \theta 2x2 + \ldots + \theta nxn$

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

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.

- First scatter plot appears to be simple linear
- Second graph is not distributed normally
- Third graph, Distribution is linear but should have a different regression line
- Fourth graph shows an example when one high-leverage point is enough to produce a high correlation coefficient



- **What is Pearson's R?**

It is a numerical summary of the strength of the linear association between the variables. It value ranges between -1 to +1. It sshhows the linear relationship betn two set of data.

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

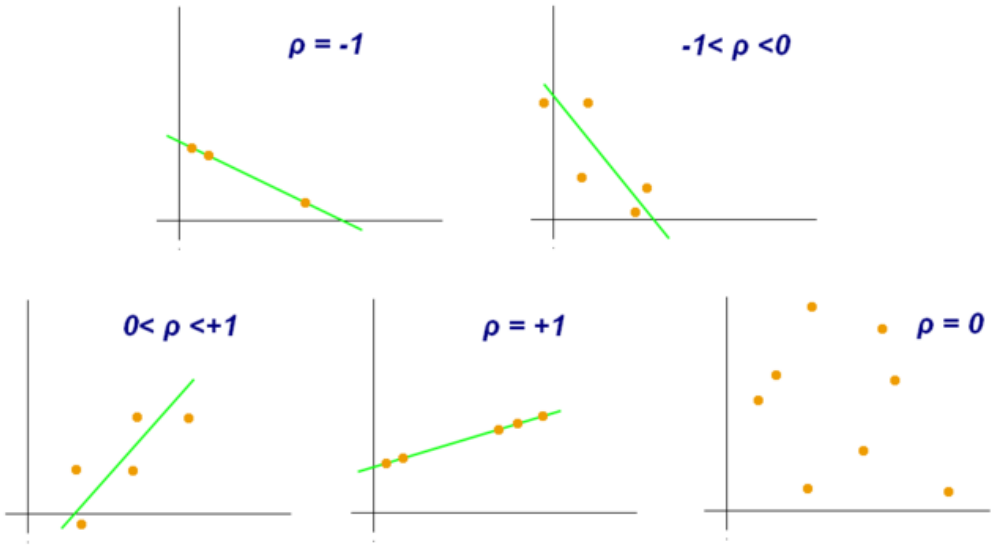
y_i = values of the y-variable in a sample

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

R=1 means data is perfectly linear with +ve slope

R=-1 means data is perfectly linear with -ve slope

R=0 means no linear associations



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

It's a method used to normalize the range of independent variables of data. It is performed during the data pre-procesing stage to deal with varying value in the dataset

If feature scaling is not done, Then machine learning algo tends to weigh greater value.

Normalization is generally used when we know that the distribution of our data does not follow gaussian distribution. This can be useful in algo that do not assume any distribution

Standardization: It can be helpful in cases where the data follows a gaussian distribution, However, does not have to be necessarily true. Unlike normalization, standardization does not have bounding range.

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

The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. If there is perfect correlation, Then VIF = infinity. It gives a basic quantitive idea bout how much the future variable are correlated with each other.

$$VIF = \frac{1}{1 - R^2}$$

Where R-1 is r-squared value of theat independent variable which we want to check how well this independent variable is explained well by other variables. If that independent variable can be explained perfectly by other independent variables, then it will have perfect corellation.

A rule of thumb for interpreting the variance inflation factor:

- 1 = not correlated
- Between 1-5 = moderately correlated
- Greater than 5 = highly corellated

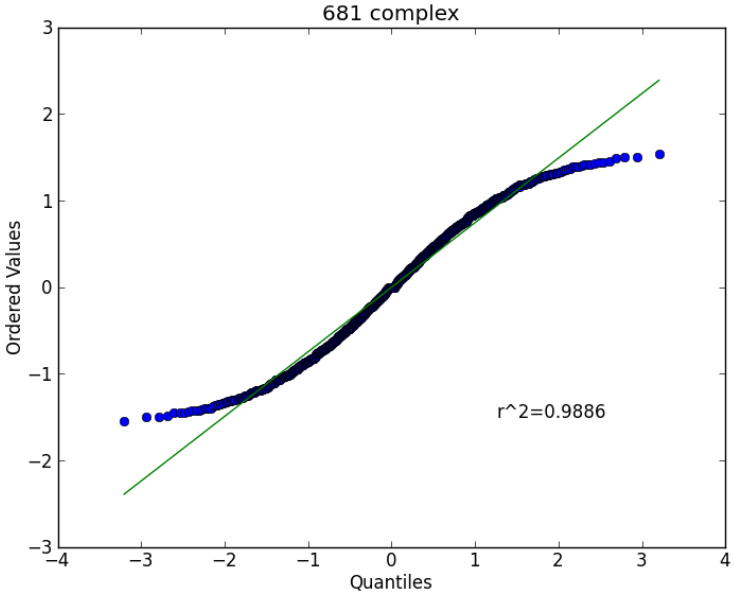
- **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**

It's a plot of the quantiles of the first data set against the quantiles of the second data set. A q-q plot is scatterplot created by plotting two sets of quantiles against one another.

If both sets of quantiless come from the same distribution, We should see the point sforming roughfly straight line.

Q-Q plot is used to answer question like:

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



-