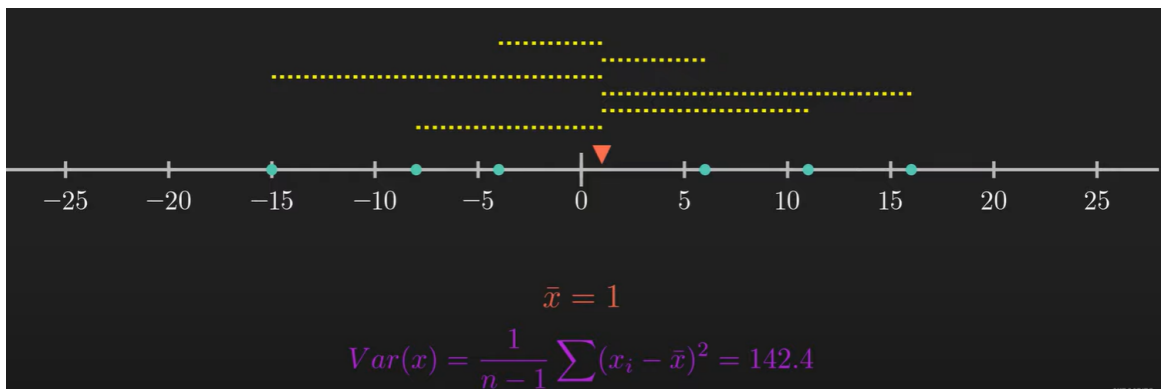


Difference between covariance and correlation

Covariance and correlation are two statistical measures used to describe the relationship between two or more variables in a dataset. They help us understand how changes in one variable relate to changes in another. While they are related concepts, they serve slightly different purposes and have different interpretations.

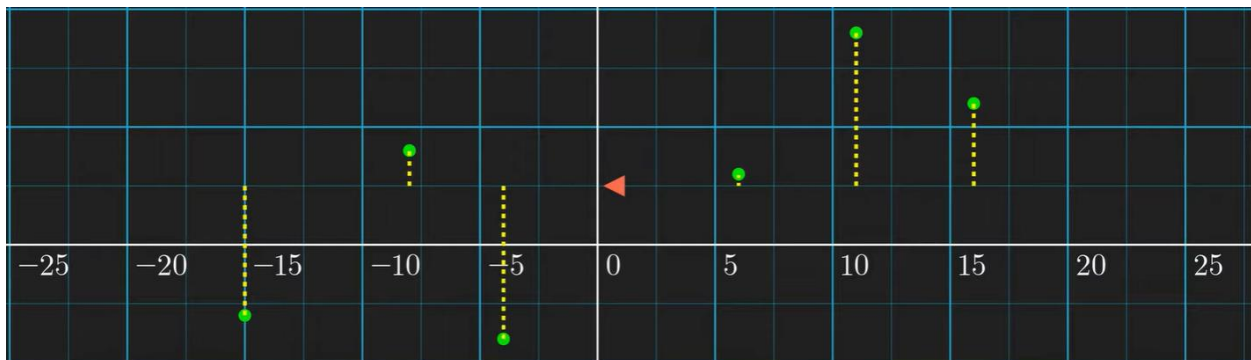
Before we understand covariance let's understand **variance**.

- Variance is a measure of how concentrated or scattered the data points are around the mean.
- In the below picture a 1 dimensional data set is presented and we calculate and visualize the variance along with the formula. The coordinates of the data points are not given but we know the variance between them in 1 dimension. It is called the X- Variance or Variance in the X-axis.

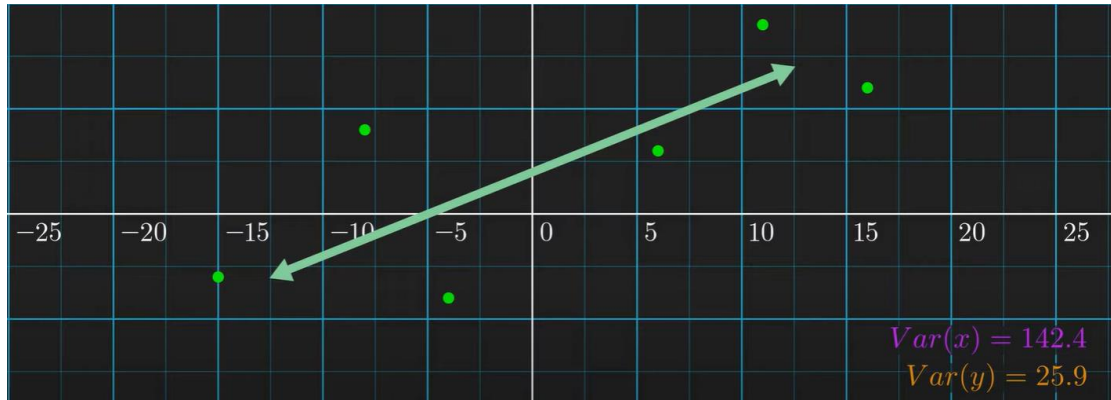


(Here X-Variance is 142.4, The red triangle denotes the average of the data)

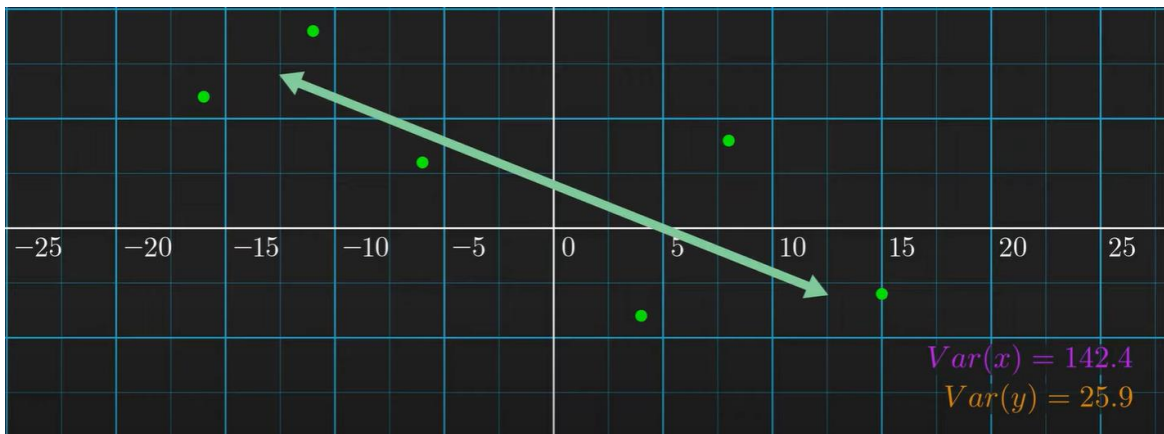
- Now let's look at a 2 dimensional dataset. And in this data set we can calculate the variance in the y-axis or Y-Variance using the same method. In this case the Y- variance is : 25.9, and the Red Triangle denotes the average of the data in the Y-axis.



- However, these two measures alone doesn't say much about the distribution of the data in 2-dimensions. Because upon looking at the dataset we can clearly say that if X increases Y also increases. But we can't infer this relation from the X-Variance or Y-Variance values.



- Also, if we alter the direction of the trend here the X-Variance and Y-variance remains the same but the directionality of the data is different. Look at below figure.



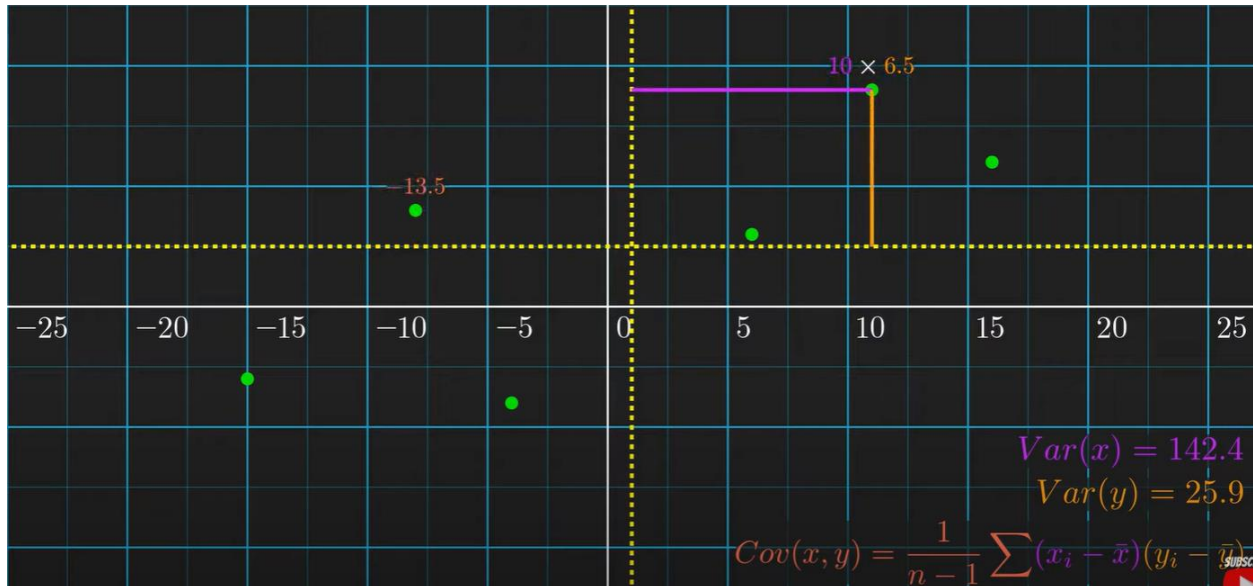
- Hence to resolve this issue we need to calculate covariance which can help in determining the direction of the trend.

Covariance:

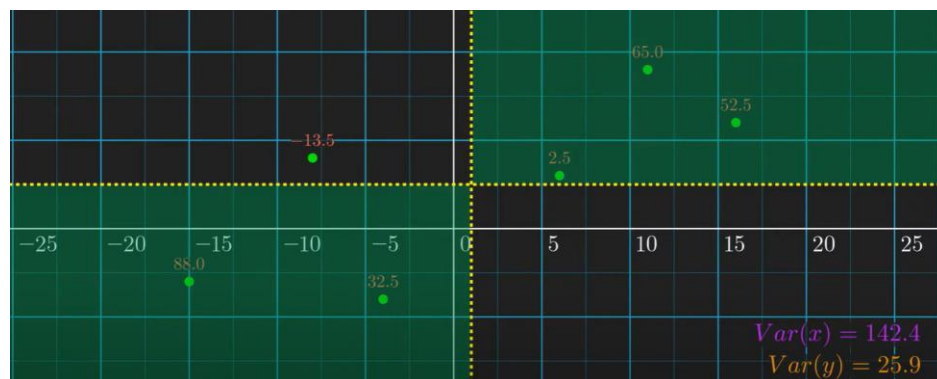
- Covariance measures the degree to which two variables change together. If the variables tend to increase or decrease at the same time, the covariance is positive. If one variable tends to increase when the other decreases, the covariance is negative.

- The formula for calculating the covariance between two variables X and Y in a dataset with n data points is:

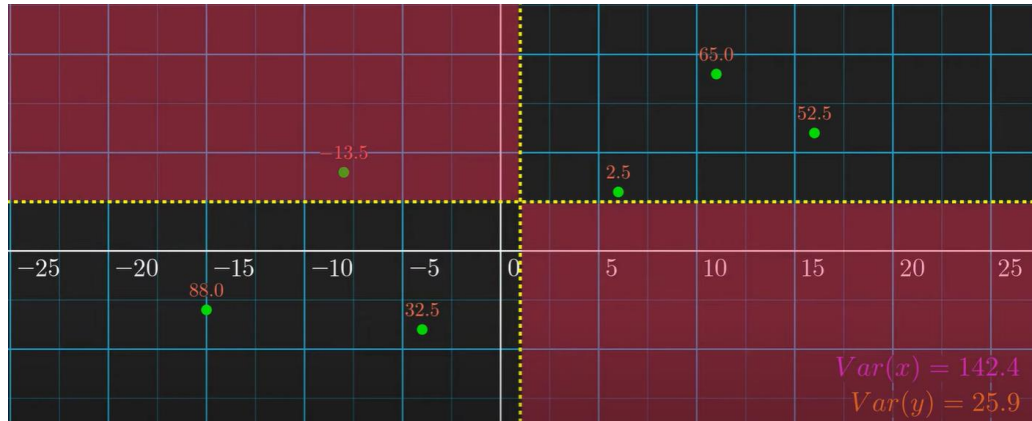
$$\text{Cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$



- Here, \bar{X} and \bar{Y} are the means (averages) of X and Y , respectively.
- In the above diagram if the sum of all these rectangular area is positive (i.e 1st Quadrant and 3rd Quadrant) then the data has a positive trend. i.e if X -increase then Y -increase and vice versa.
- If the sum of all the above rectangular areas is negative then the data has a negative trend. i.e If X increase Y - decreases and vice versa.



- The green region is the positive region and the data has a positive trend of most of the data lies in the green region and vice versa. In our case we have more points in the Green region than in red so our sum will be positive and the Dataset has a positive trend.



- We usually write covariance with a covariance matrix in the following form.

$$\begin{bmatrix} Cov(x, x) & Cov(x, y) \\ Cov(x, y) & Cov(y, y) \end{bmatrix}$$

In this case our covariance matrix will look like

$$\begin{bmatrix} 142.4 & 45.4 \\ 45.4 & 25.9 \end{bmatrix}$$

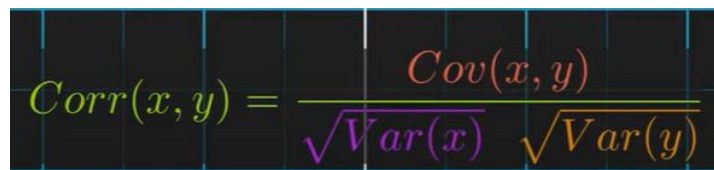
- Covariance has some limitations:
 - Covariance quantifies the direction of the trend but it doesn't say anything about the strength of the relationship.
 - Co variance is very sensitive to scaling . Hence the range of possible values for covariance is $(-\infty \text{ to } +\infty)$.
 - It is not standardized, so its value can be difficult to interpret on its own.
 - The scale of the covariance depends on the units of the variables, making comparisons between datasets with different units problematic.

Correlation:

- Correlation is a standardized measure of the strength and direction of the linear relationship between two variables. It ranges from -1 to 1, with -1 indicating a perfect negative linear relationship, 1 indicating a perfect positive linear relationship, and 0 indicating no linear relationship.
- The most commonly used correlation coefficient is the Pearson correlation coefficient, denoted as 'r'.
- The formula for calculating the Pearson correlation coefficient is:

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

Or


$$Corr(x, y) = \frac{Cov(x, y)}{\sqrt{Var(x)} \sqrt{Var(y)}}$$

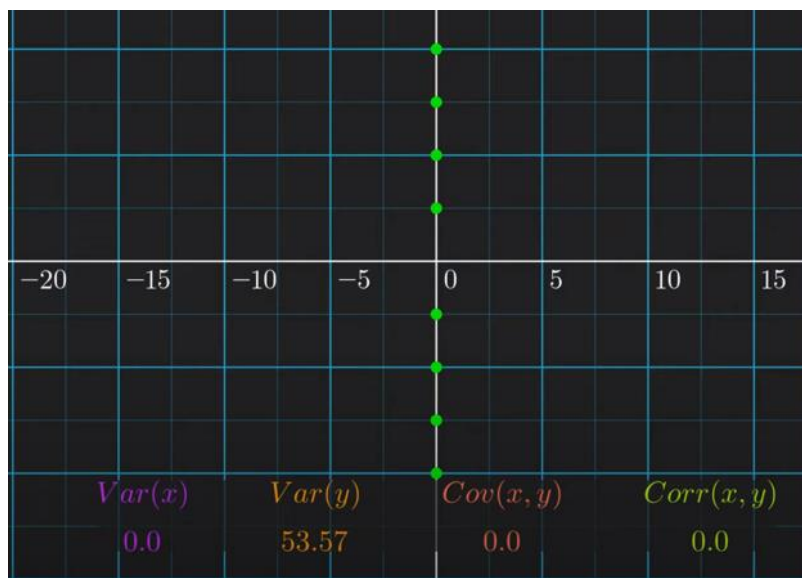
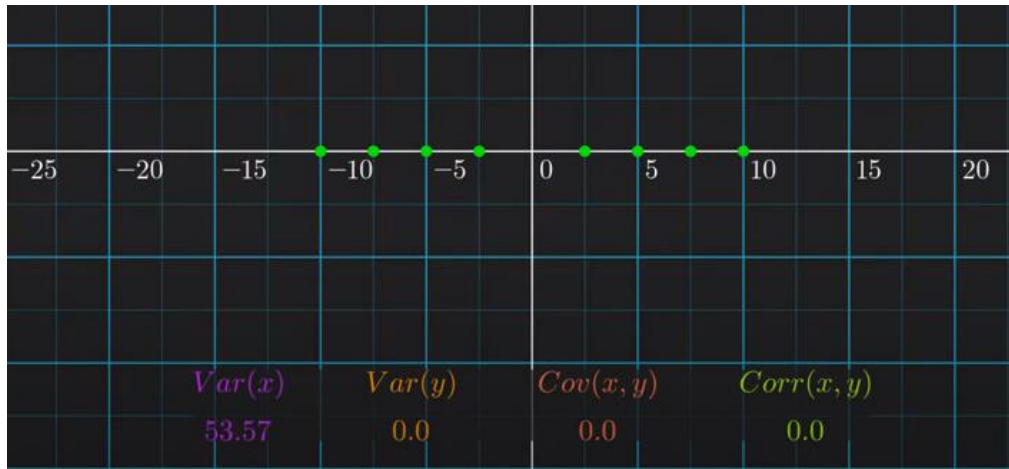
- Correlation is useful because it is unit less and standardized, making it easier to interpret and compare across different datasets.
- In our case the correlation coefficient is 0.75. Hence we have a strong positive trend.

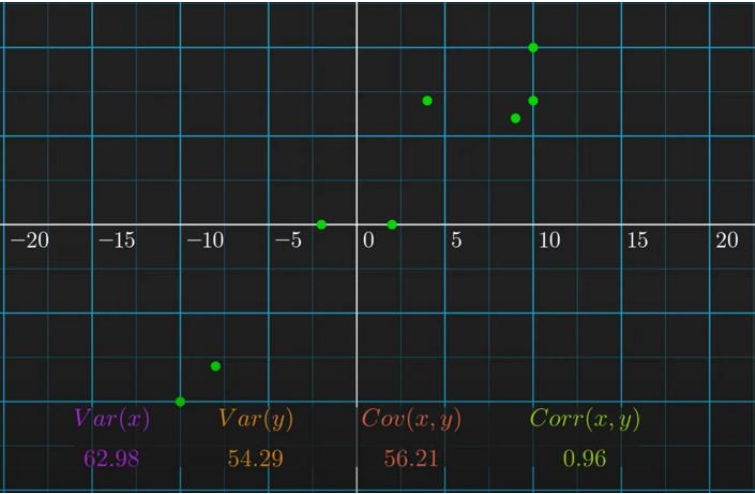
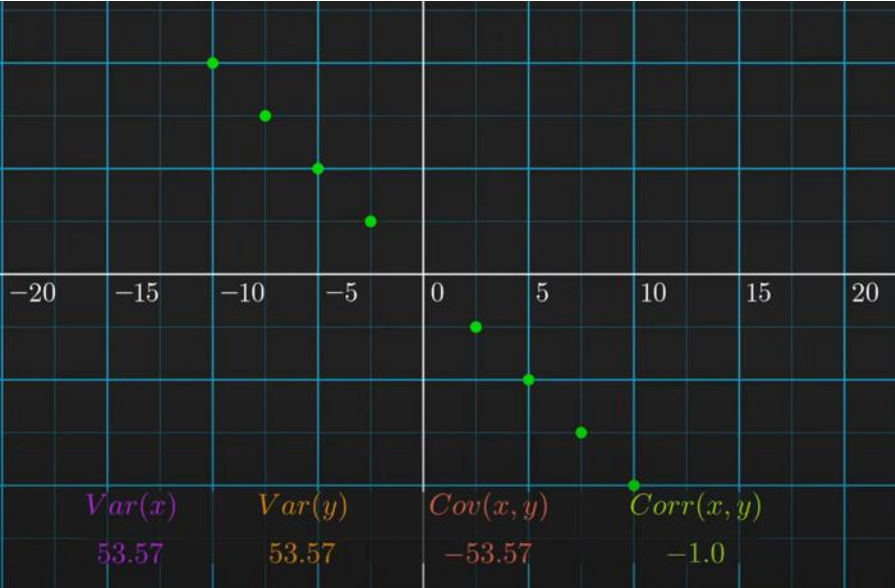
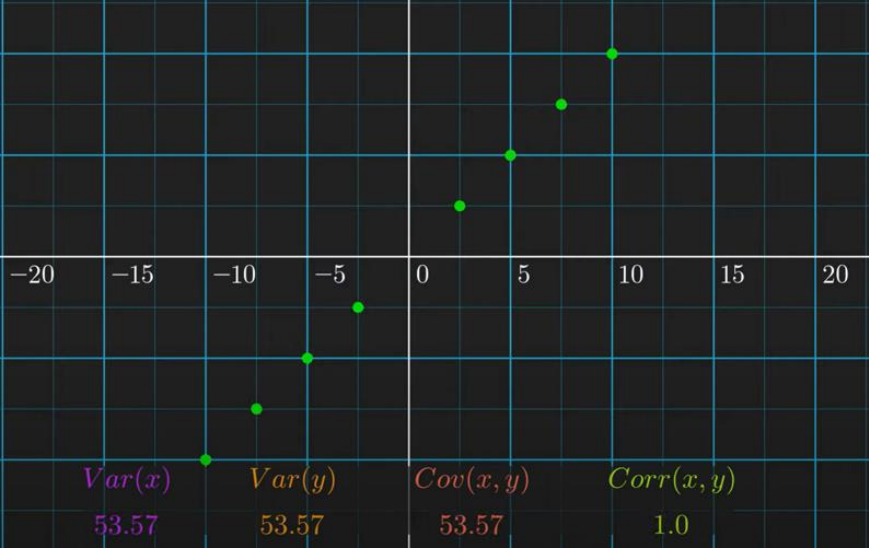
Key differences between covariance and correlation:

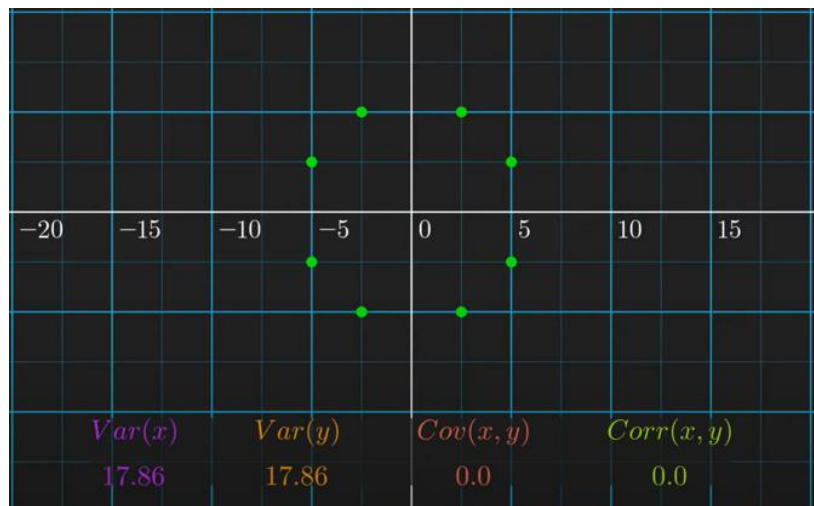
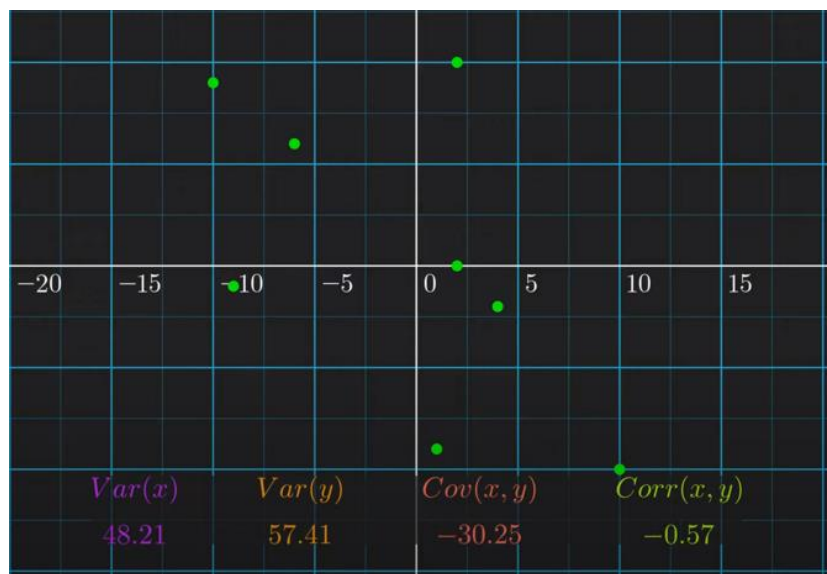
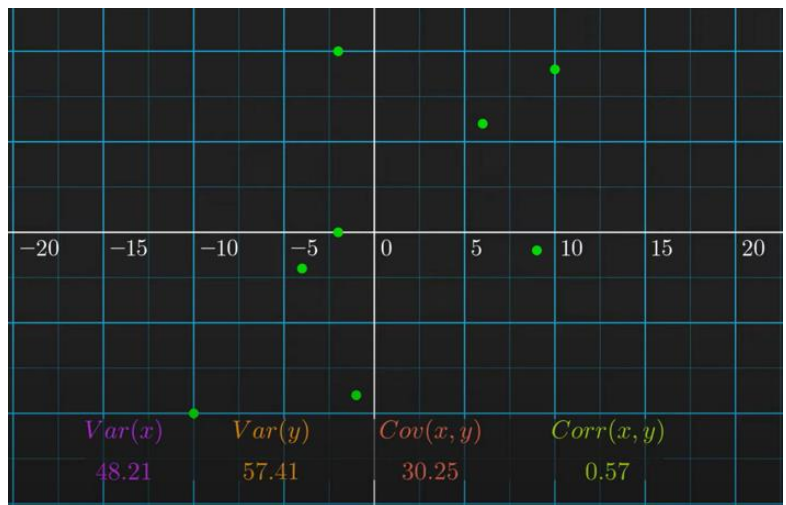
- Covariance can take any real value, while correlation is bounded between -1 and 1.
- Correlation standardizes the measure, making it easier to compare across different datasets.
- A correlation of 0 does not necessarily mean no relationship between variables, just no linear relationship. There could still be other types of relationships.

- Covariance can be positive, negative, or zero, but correlation is always between -1 and 1.
- Correlation is more commonly used when comparing the strength and direction of relationships between variables.

Here are some examples of how covariance and correlation changes in different data trends







In summary, covariance measures the degree to which two variables change together, while correlation quantifies the strength and direction of the linear relationship between them in a standardized manner. Correlation is generally preferred when analyzing relationships between variables because of its interpretability and consistency.