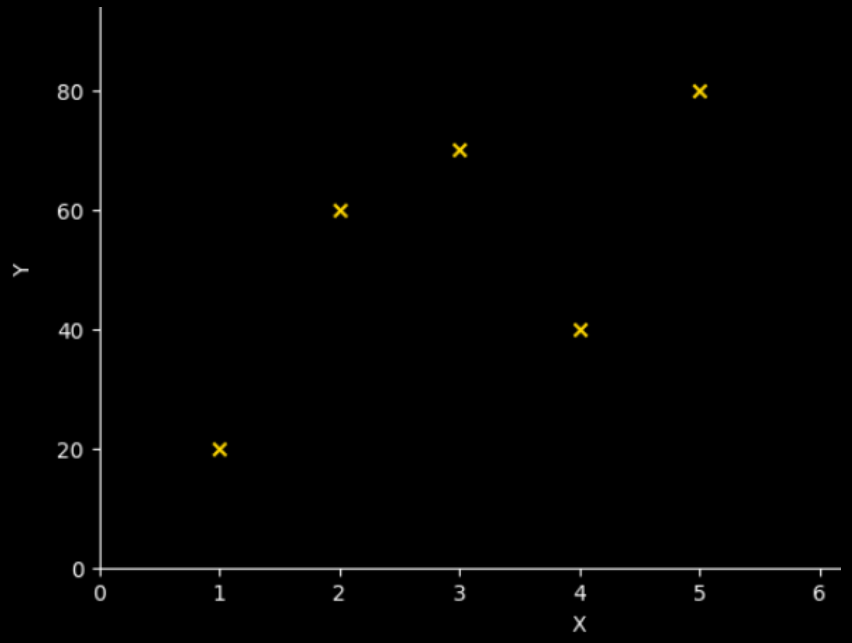
Covariance and Correlation

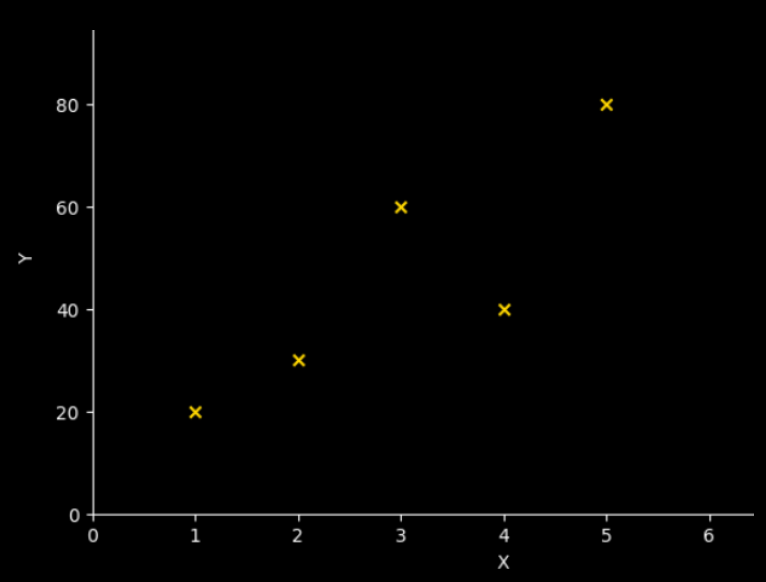
Problem: What type of relationship is between X and Y ? How would you measure it mathematically?

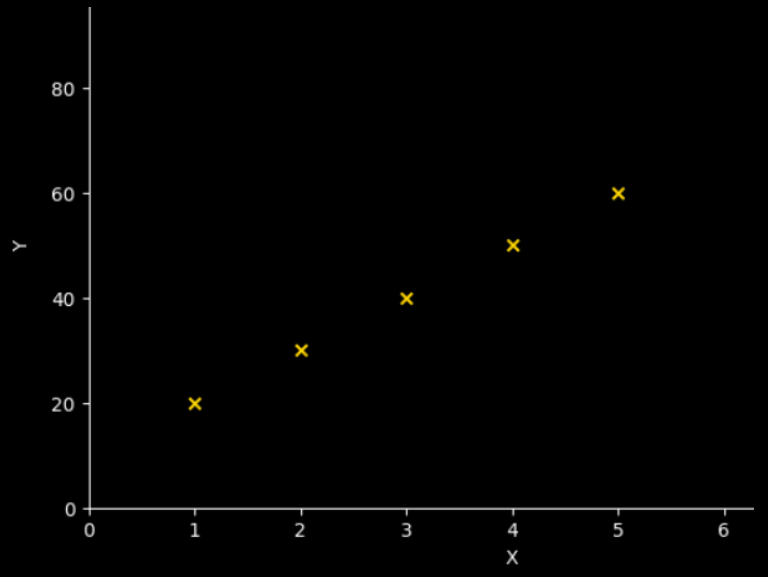
X	Y
1	20
2	60
3	70
4	40
5	80

X	Y
1	20
2	30
3	60
4	40
5	80

X	Y
1	20
2	30
3	40
4	50
5	60



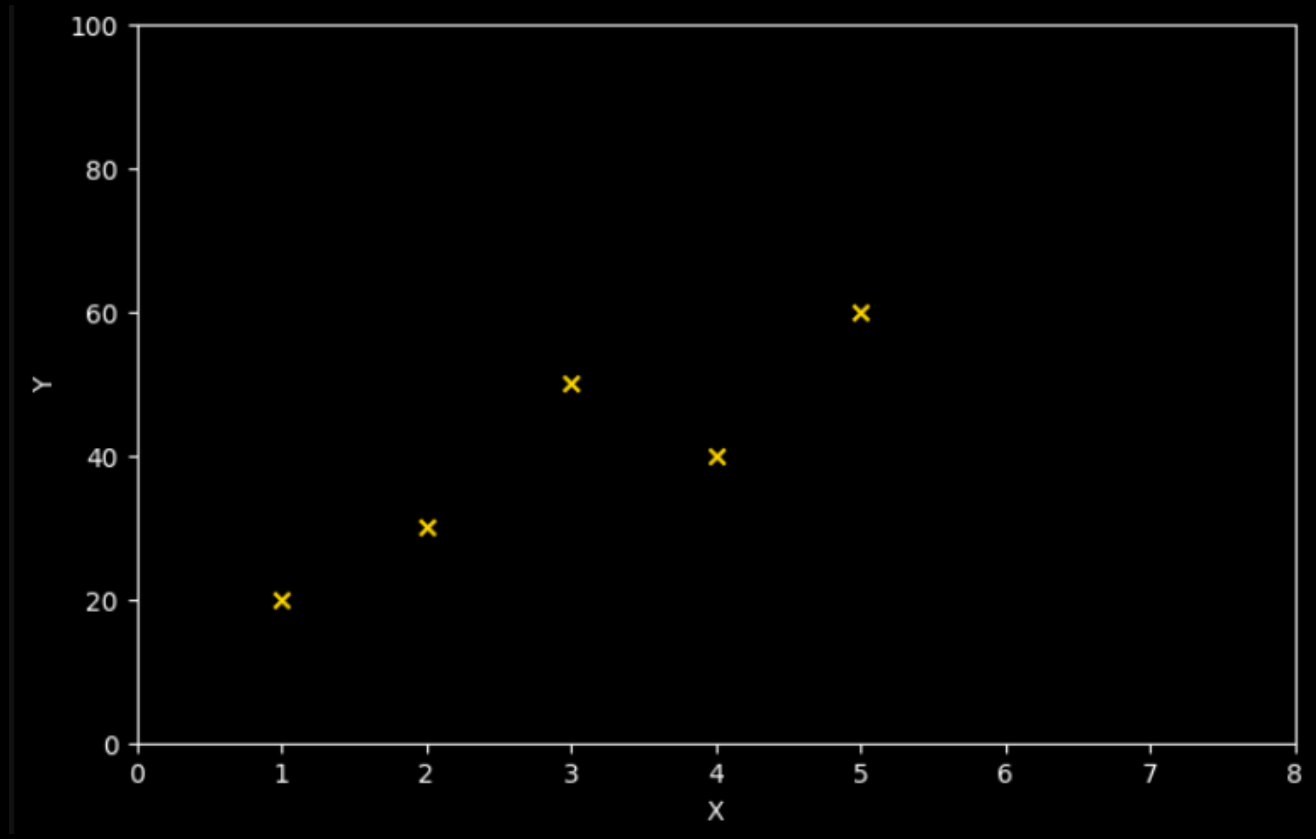




Covariance

Suppose you have following data(e.g. study hours and score):

X	Y
1	20
2	30
3	50
4	40
5	60

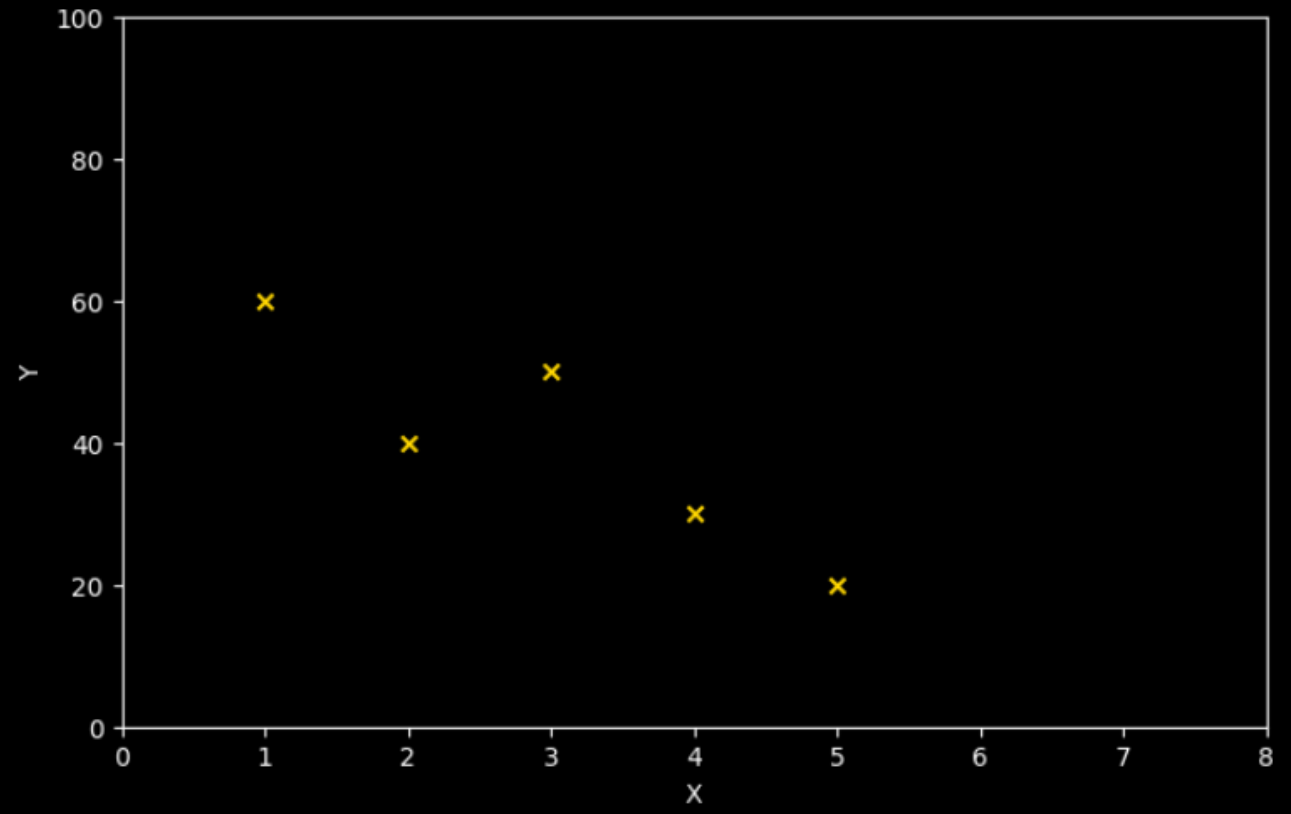


We see that as X increases, Y increases too.
Here X and Y are said to be **positively** correlated.

Covariance

Suppose you have following data (game-hours and score):

X	Y
1	60
2	40
3	50
4	30
5	20

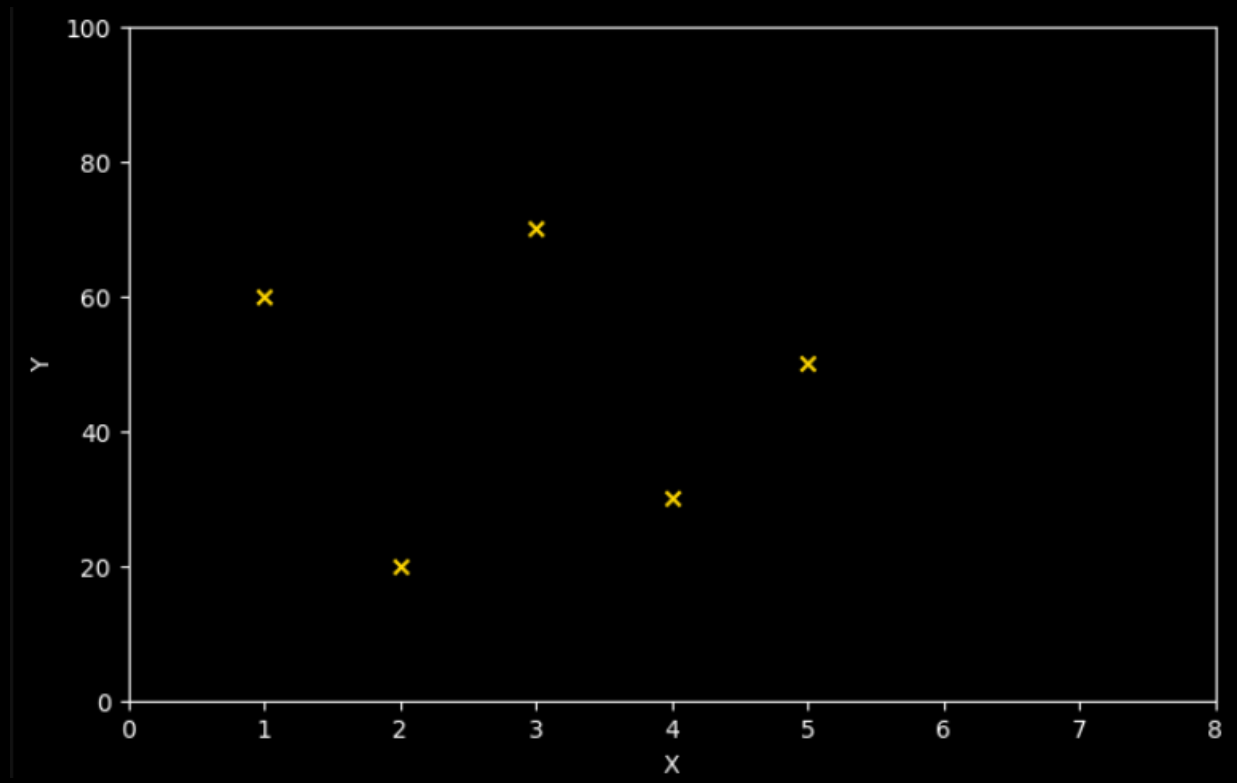


We see that as X increases, Y decreases.
Here X and Y are said to be **negatively** correlated.

Covariance

Suppose you have following data:

X	Y
1	60
2	20
3	70
4	30
5	50



Here as X increases, Y has no pattern.
X and Y are said to be **not** correlated.

Covariance

We use covariance to determine the **direction** of relation between X and Y:

$$\mathrm{Cov}(X,Y) = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})$$

- **Cov (X,Y) > 0**: X and Y move in same direction
- **Cov (X,Y) < 0**: X and Y move in opposite direction
- **Cov (X,Y) = 0**: X and Y are independent

Side Note: Above is the covariance of **samples,** not population. For population, n-1 → n

Covariance

Example1: Find covariance between X and Y

X	Y
1	20
2	30
3	50
4	40
5	60

$$\mathrm{Cov}(X,Y) = \frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})(y_k-\bar{y})$$

$$\bar{x} = \frac{1+2+3+4+5}{5} = 3$$

$$\bar{y} = \frac{20+30+50+40+60}{5} = 40$$

$$Cov(X,Y) = \frac{1}{n-1}\left[(x_1-\bar{x})(y_1-\bar{y}) + (x_2-\bar{x})(y_2-\bar{y}) + \cdots + (x_5-\bar{x})(y_5-\bar{y})\right]$$

$$Cov(X,Y) = \frac{1}{(5-1)}\left[(1-3)(20-40) + (2-3)(30-40) + (3-3)(50-40) + (4-3)(40-40) + (5-3)(60-40)\right]$$

$$= \frac{1}{4}\left[40 + 10 + 0 + 0 + 40\right]$$

$$= 22.5$$

Covariance

Example2: If I take same data as previous and divide Y by 10, scale it down, then COV(X,Y) = 2.25 (smaller by factor of 10)

X	Y
1	2
2	3
3	5
4	4
5	6

$$\mathrm{Cov}(X,Y) = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})$$

$$\bar{x} = \frac{1+2+3+4+5}{5} = 3$$

$$\bar{y} = \frac{2+3+5+4+6}{5} = 4$$

$$Cov(X,Y) = \frac{1}{n-1}\left[(x_1-\bar{x})(y_1-\bar{y}) + (x_2-\bar{x})(y_2-\bar{y}) + \cdots + (x_5-\bar{x})(y_5-\bar{y})\right]$$

$$Cov(X,Y) = \frac{1}{(5-1)}\left[(1-3)(2-4)+(2-3)(3-4)+(3-3)(5-4)+(4-3)(4-4)+(5-3)(6-4)\right]$$

$$= \frac{1}{4}\left[4 + 1 + 0 + 0 + 4\right]$$

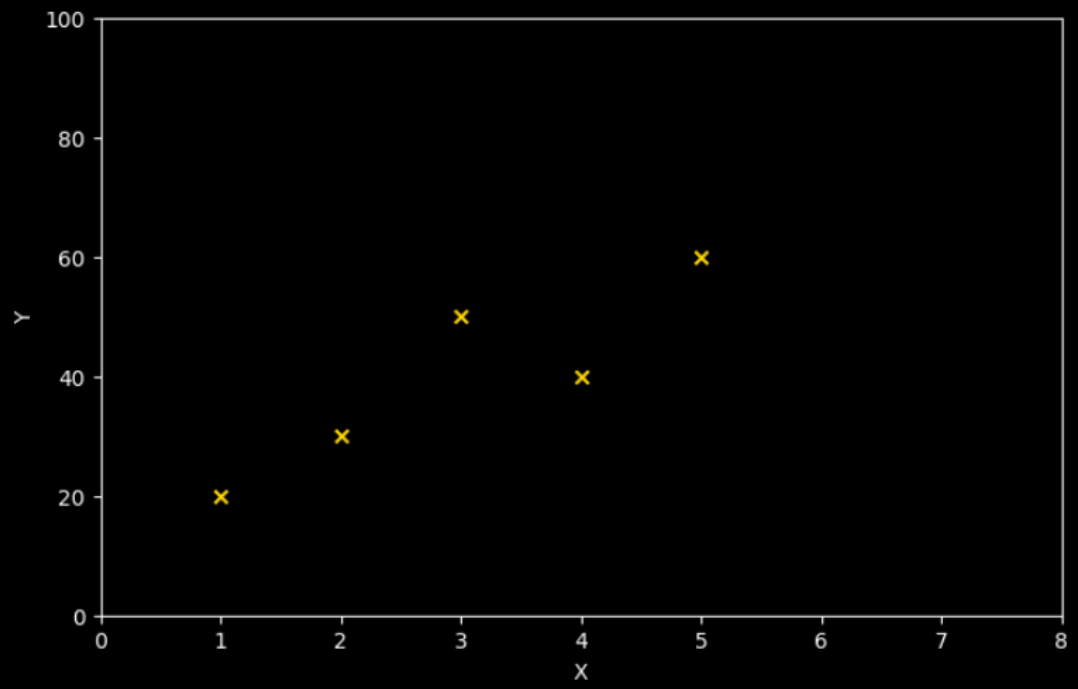
$$= 2.25$$

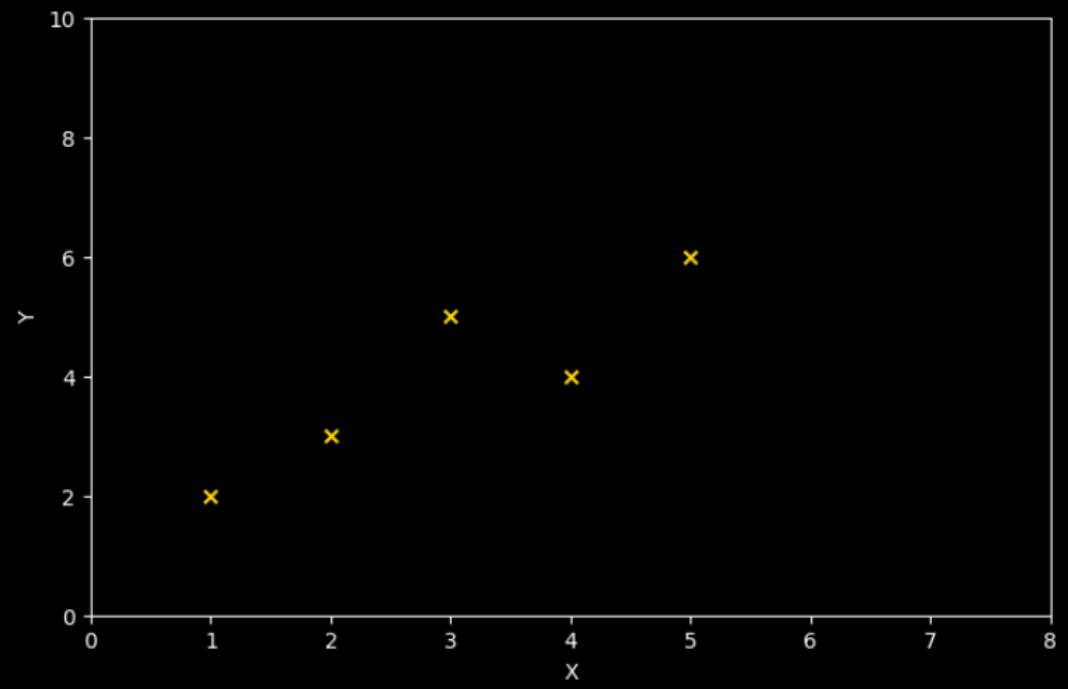
Covariance

Now let's plot both these separately. Suppose in 1st figure, X is in meters and Y is in Kg, then covariance = 22.5 m-kg. In 2nd figure, X is in hours and Y is in score, and covariance = 2.25 hours-score. Numerically they are different because of different units, but visually they look identical.

X (m)	Y (Kg)
1	20
2	30
3	50
4	40
5	60

X(hr)	Y (score)
1	2
2	3
3	5
4	4
5	6





Correlation

If we divide Cov(X,Y) by standard deviation (X) and standard deviation (Y), then we get **correlation.**

Correlation measures the strength of linear relationship between 2 variables.

$$r = \frac{\mathrm{Cov}(X,Y)}{s_X\, s_Y}$$

$$S_X = \sqrt{\frac{\sum(x_i - \bar{X})^2}{n-1}}$$

$$S_Y = \sqrt{\frac{\sum(y_i - \bar{Y})^2}{n-1}}$$

Side Note: Above is for **sample.** For population, n-1 → n

We use correlation over covariance because it's a **standardized** measure that is **easy to interpret and compare across different datasets**: Dataset1 could be between X in kg and Y in meter, in the Dataset2, X could be in hours and Y in score. We can still compare the datasets

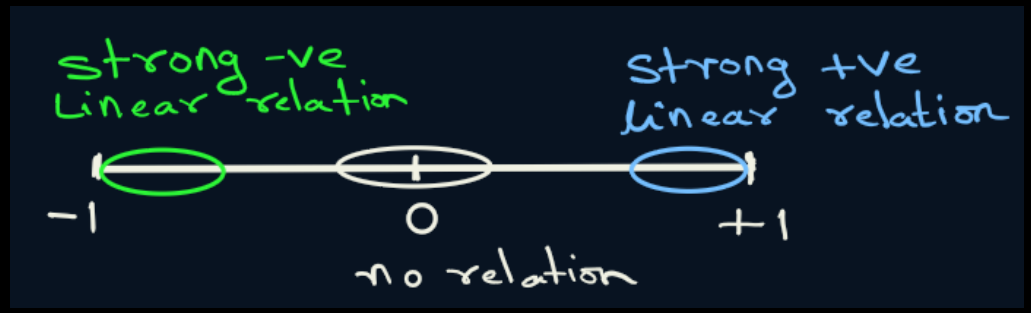
Correlation

Key properties of correlation

- It is unitless

$$r = \frac{\mathrm{Cov}(X, Y)}{s_X \, s_Y}$$

- Ranges from -1 to 1



- If the values of X and/or Y are converted to a different scale, the value of r will be unchanged. (Shown in next 2 examples)

Correlation

Example: Find correlation.

X	Y
1	20
2	30
3	50
4	40
5	60

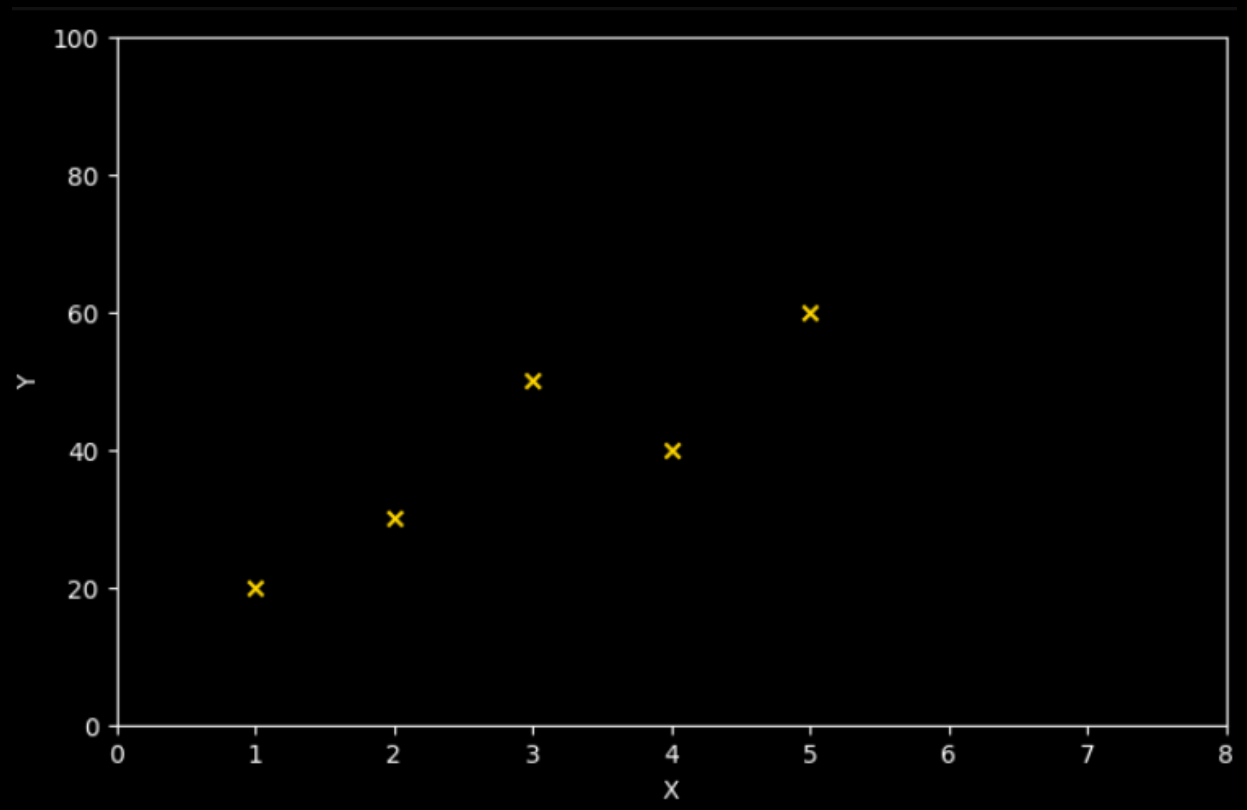
$$r = \frac{\mathrm{Cov}(X,Y)}{s_X\, s_Y}$$

$$S_X = \sqrt{\frac{\sum(x_i - \bar{X})^2}{n-1}}$$

We already found Cov(X,Y) = 22.5

The S_X = 1.58, S_Y = 15.8 →

$$r = \frac{22.5}{1.58 \times 15.8} = 0.9$$



Correlation

Example(Same data as before with Y scaled down by factor of 10): Find correlation.

X	Y
1	2
2	3
3	5
4	4
5	6

$$r = \frac{\mathrm{Cov}(X,Y)}{s_X\, s_Y}$$

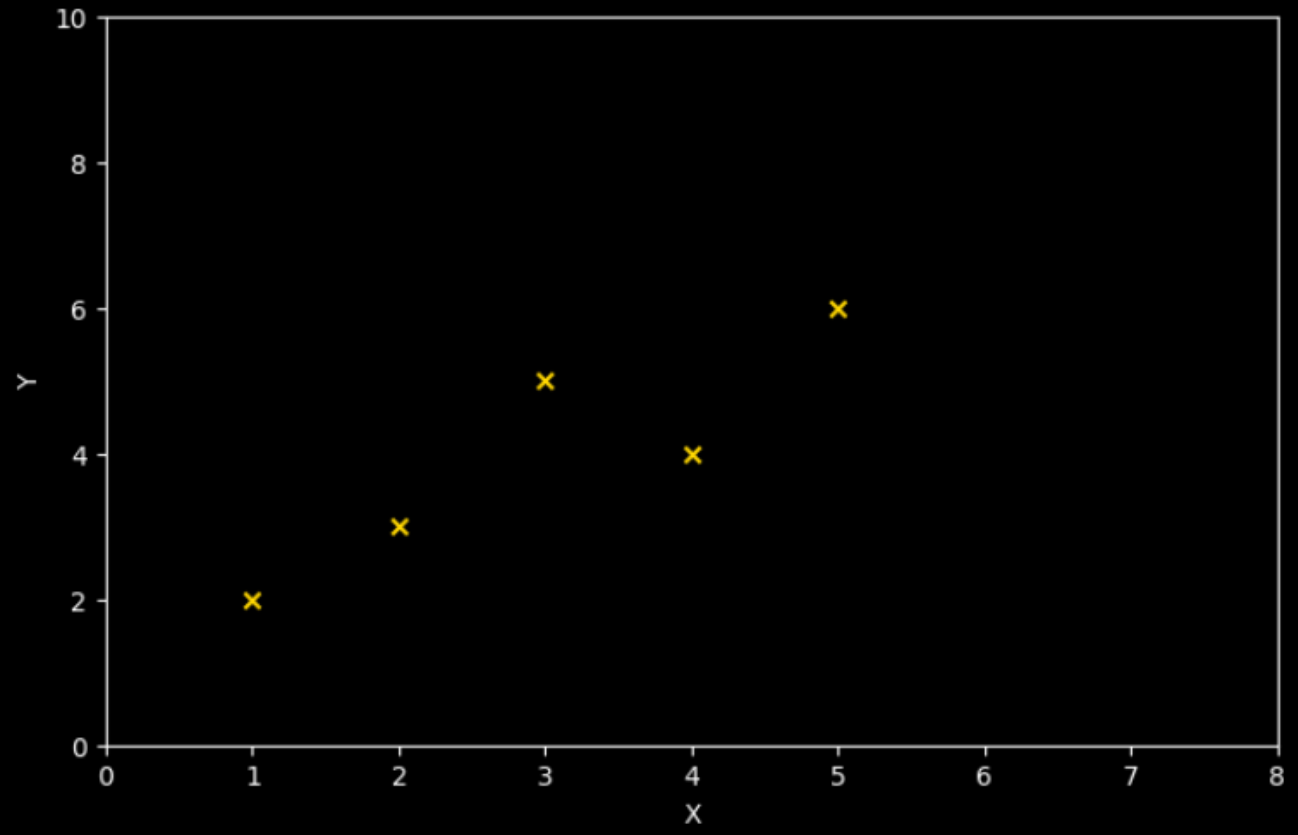
$$S_X = \sqrt{\frac{\sum(x_i - \bar{X})^2}{n-1}} \qquad S_Y = \sqrt{\frac{\sum(y_i - \bar{Y})^2}{n-1}}$$

We already found Cov(X, Y) = 2.25

The s_X = 1.58, s_Y = 1.58 →

$$r = \frac{2.25}{1.58 \times 1.58} = 0.9$$

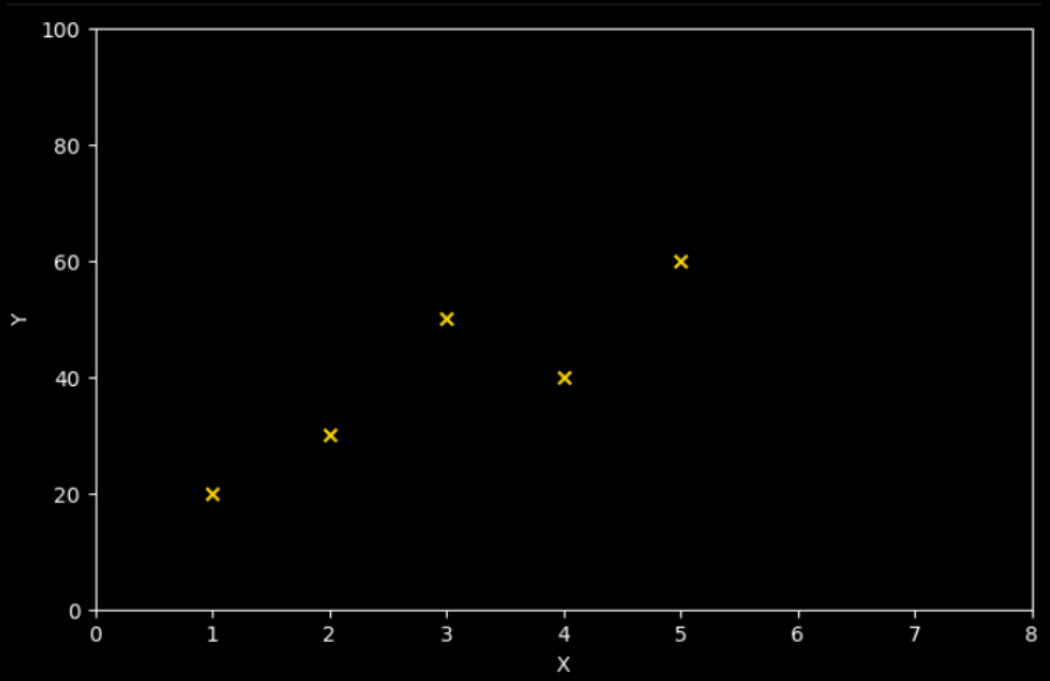
The correlation is same as before.
Correlation is unaffected by scaling

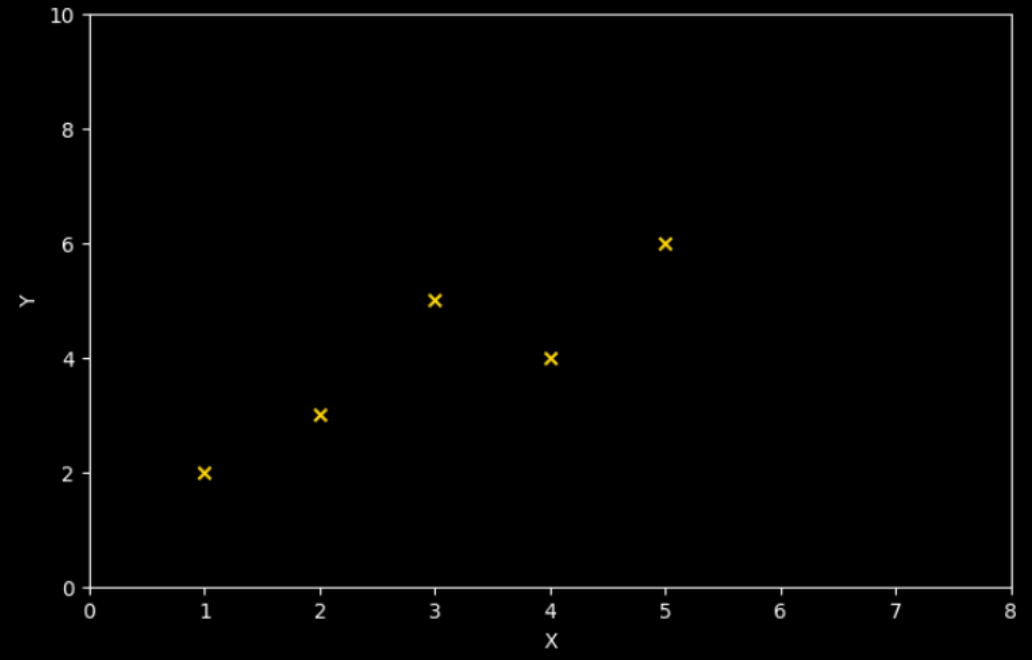


Now let’s plot both these separately. The covariance is different: +22.5 -vs- +2.25.
The correlation is same: 0.9

X	Y
1	20
2	30
3	50
4	40
5	60

X	Y
1	2
2	3
3	5
4	4
5	6

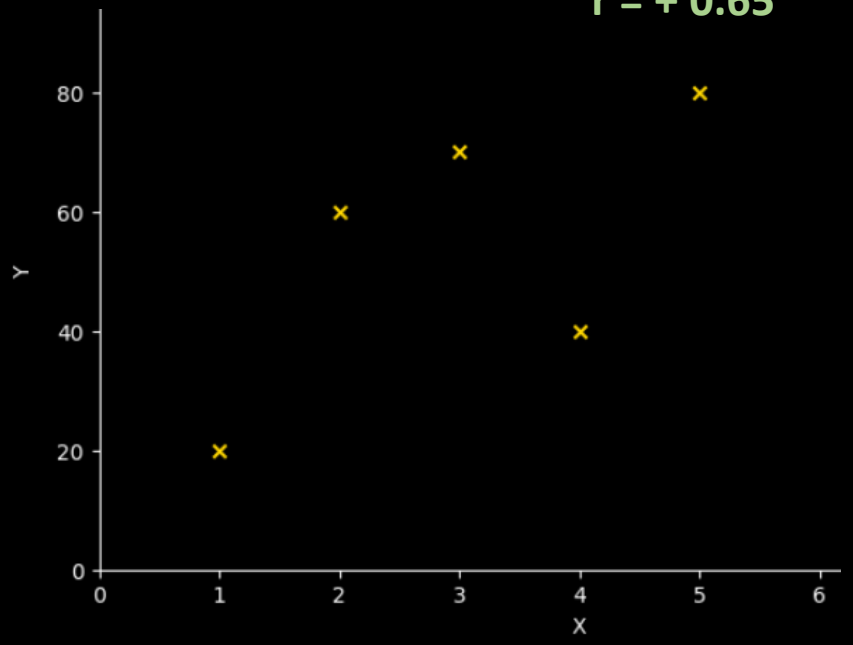




Problem: Which of the following has strong linear relationship between X and Y ? How would you measure it?
Ans: Find correlation and that would tell you how strong the linear relation is between the 2 variables.
Correlation is maximum for figure with r = +1

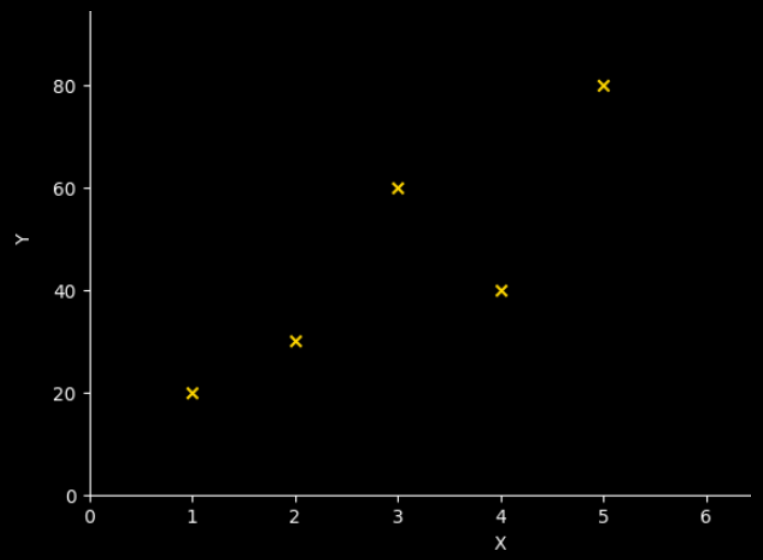
X	Y
1	20
2	60
3	70
4	40
5	80

r = + 0.65



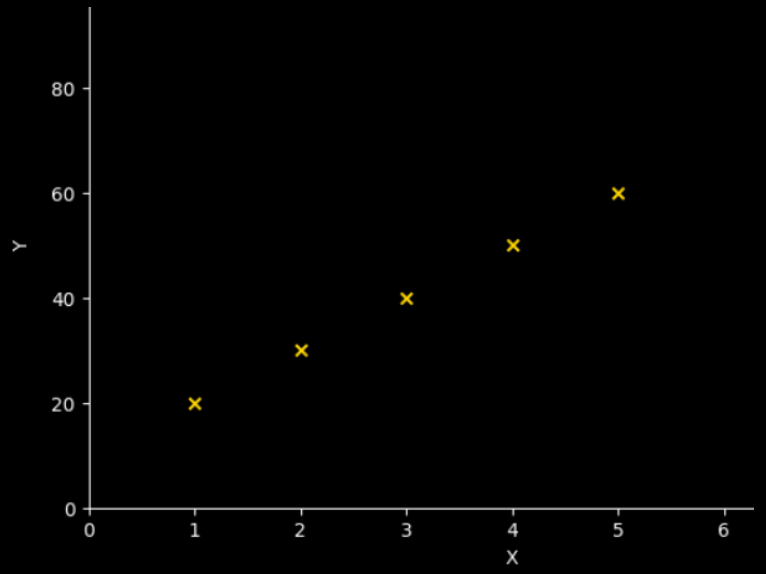
X	Y
1	20
2	30
3	60
4	40
5	80

r = + 0.85



X	Y
1	20
2	30
3	40
4	50
5	60

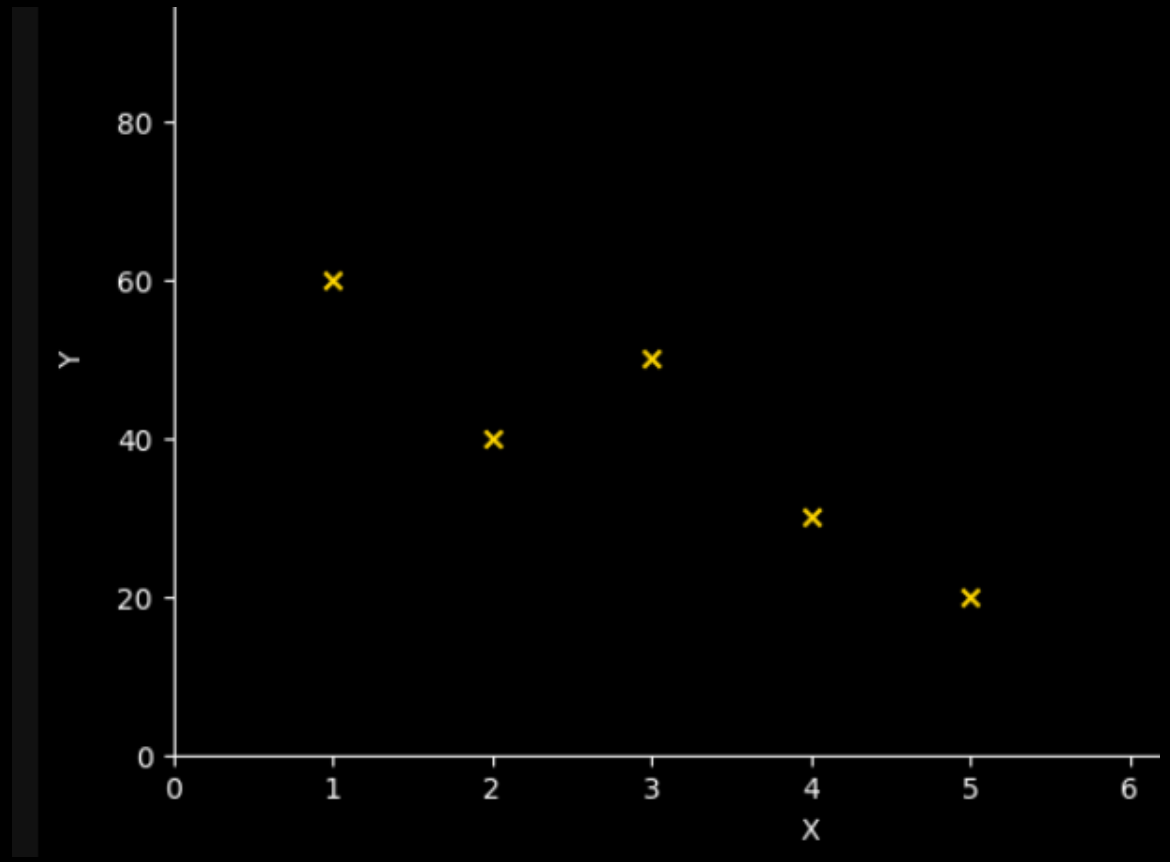
r = + 1



Example: X and Y have a strong negative correlation.

X	Y
1	60
2	40
3	50
4	30
5	20

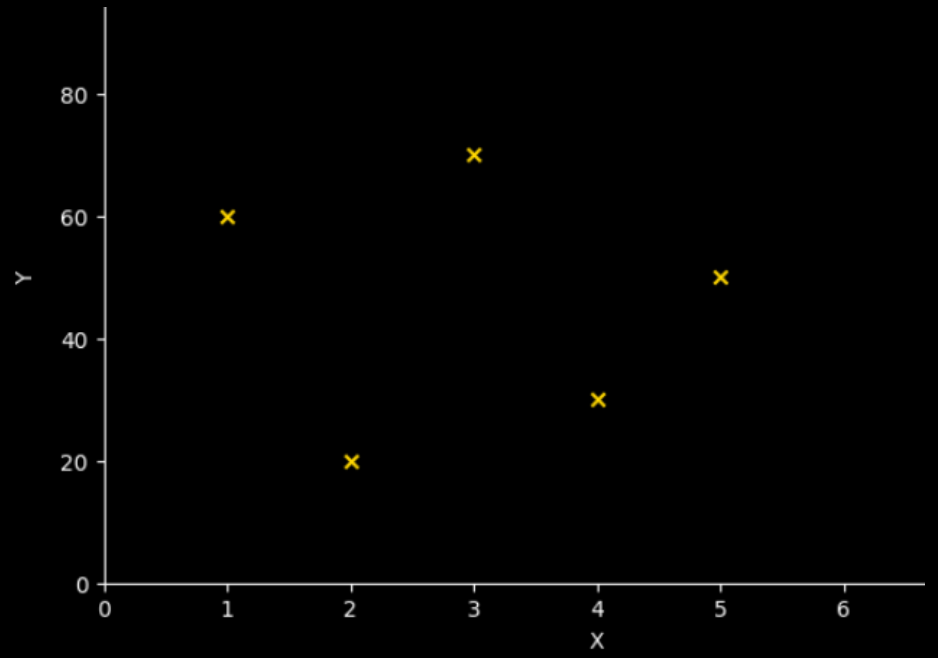
r = - 0.9

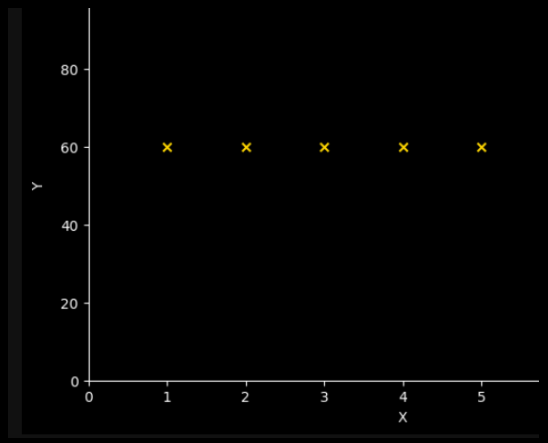


Example: X and Y have almost 0 correlation

X	Y
1	60
2	20
3	70
4	30
5	50

r = - 0.07





Recap

Cov (X,Y) > 0: X and Y move in same direction

Cov (X,Y) < 0: X and Y move in opposite direction

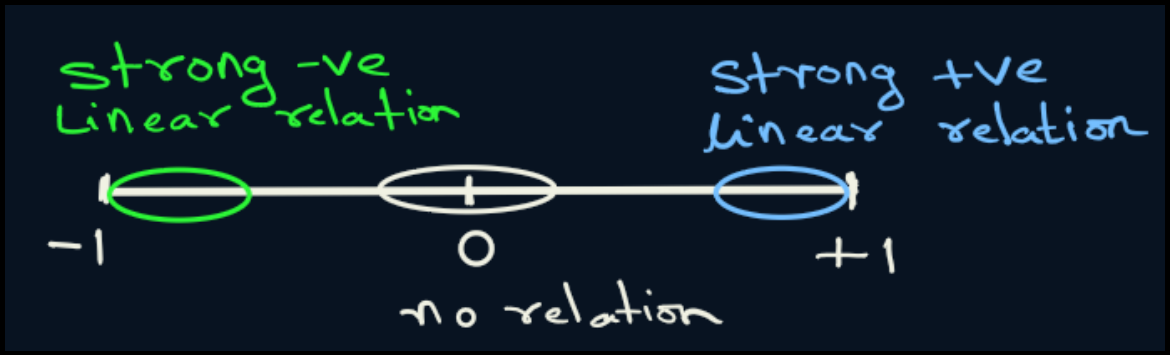
Cov (X,Y) = 0: X and Y are independent

Major flaw of covariance:

- It does not tell you the strength of linear relationship between X and Y.

So, we need correlation.

Correlation ranges from -1 to 1:

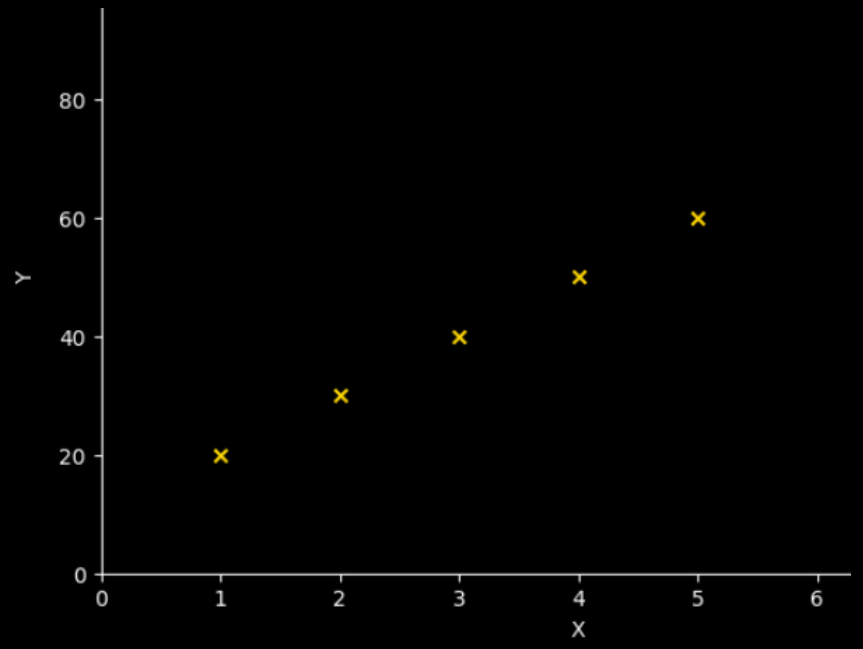


Outliers

The value of r is **sensitive to outliers** and can change dramatically if they are present in the data.

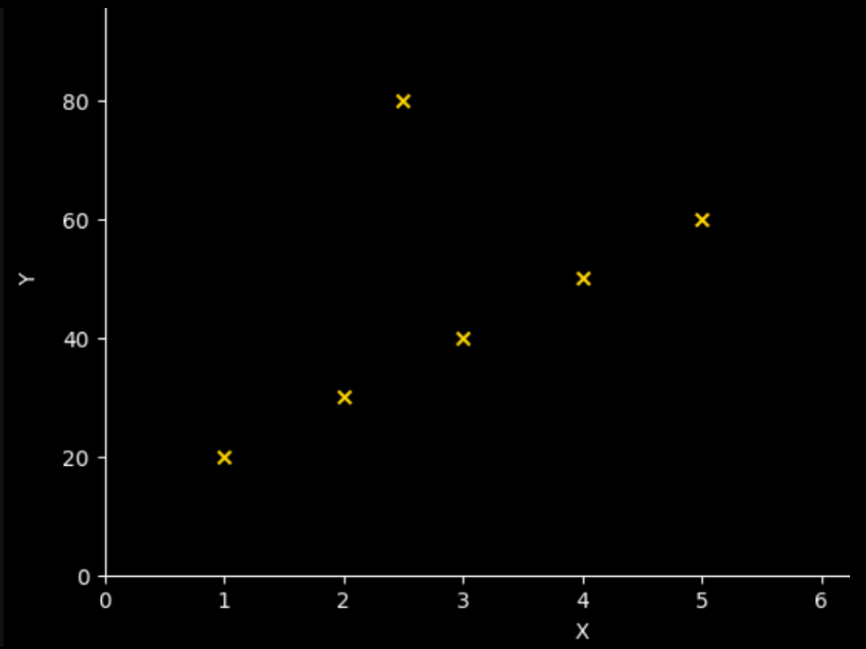
X	Y
1	20
2	30
3	40
4	50
5	60

r = + 1



X	Y
1	20
2	30
2.5	**80**
3	100
4	50
5	60

r = + 0.53



Covariance and Correlation Matrix

$$\Sigma = \begin{bmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_p) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) & \cdots & \mathrm{Cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_p, X_1) & \mathrm{Cov}(X_p, X_2) & \cdots & \mathrm{Cov}(X_p, X_p) \end{bmatrix}$$

$$R = \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1p} \\ r_{21} & 1 & r_{23} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ r_{p1} & r_{p2} & r_{p3} & \cdots & 1 \end{bmatrix}$$

$$r_{ij} = \frac{\mathrm{Cov}(X_i, X_j)}{\sigma_{X_i} \sigma_{X_j}}$$

Covariance and Correlation Matrix

Printing covariance and correlation between columns x1, x2, x3 and x4 using pandas.

4 variable:

```
import pandas as pd

x1 = [1, 2, 3, 4, 5]
x2 = [2, 3, 5, 4, 6]
x3 = [6, 4, 5, 3, 2]
x4 = [6, 2, 7, 3, 5]

df = pd.DataFrame({'x1': x1,
                   'x2': x2,
                   'x3': x3,
                   'x4': x4}
)

print("\nCovariance:\n", df.cov())
print("\nCorrelation:\n", round(df.corr(),2))
```

```
Covariance:
       x1    x2    x3    x4
x1  2.50  2.25 -2.25 -0.25
x2  2.25  2.50 -1.75  0.75
x3 -2.25 -1.75  2.50  1.50
x4 -0.25  0.75  1.50  4.30

Correlation:
       x1    x2    x3    x4
x1  1.00  0.90 -0.90 -0.08
x2  0.90  1.00 -0.70  0.23
x3 -0.90 -0.70  1.00  0.46
x4 -0.08  0.23  0.46  1.00
```

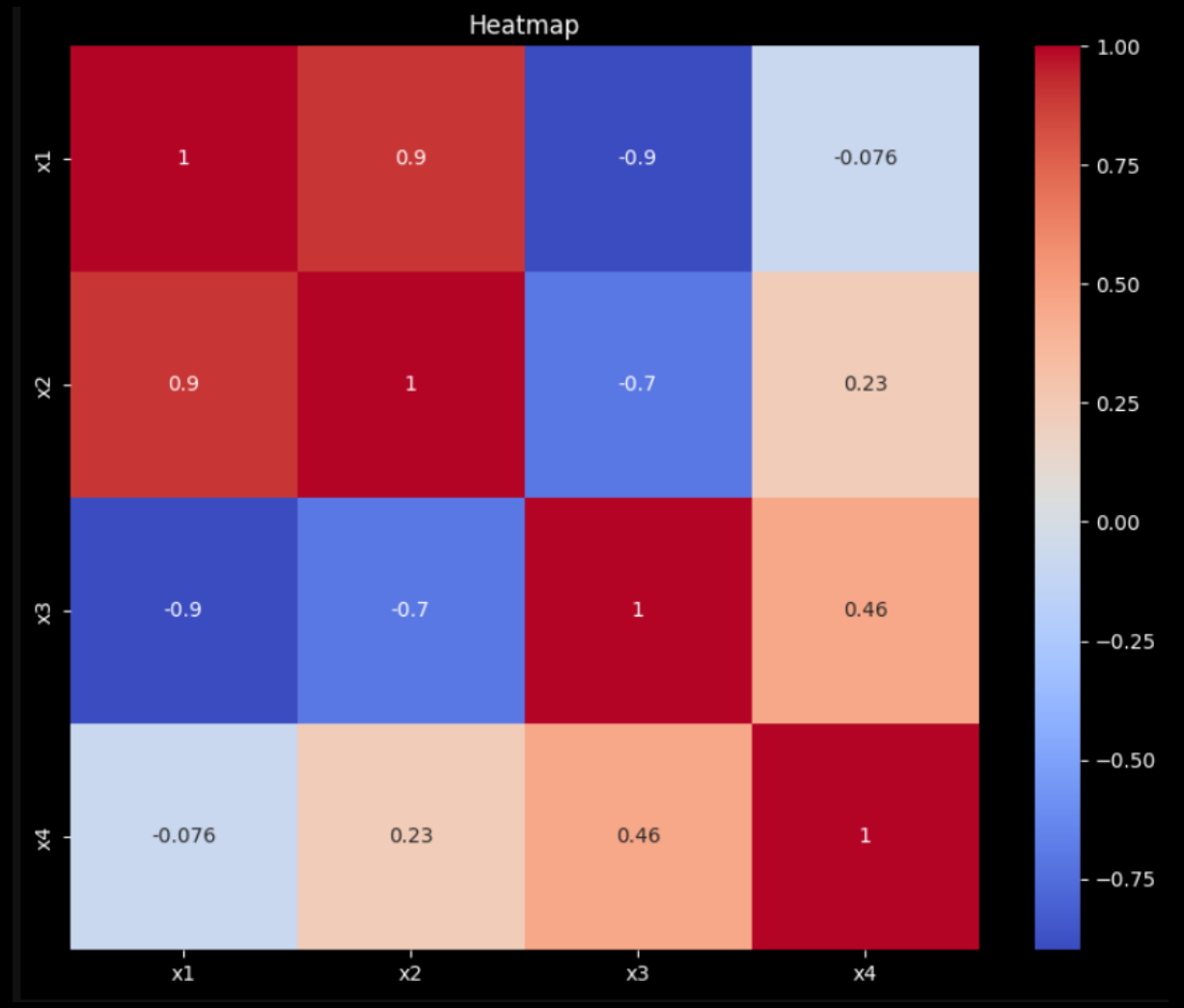
Printing covariance and correlation between columns x1, x2, x3 and x4 using pandas.

```
import matplotlib.pyplot as plt
import pandas as pd

x1 = [1, 2, 3, 4, 5]
x2 = [2, 3, 5, 4, 6]
x3 = [6, 4, 5, 3, 2]
x4 = [6, 2, 7, 3, 5]

df = pd.DataFrame({'x1': x1,
                   'x2': x2,
                   'x3': x3,
                   'x4': x4}
)

plt.figure(figsize=(10, 8))
sns.heatmap( df.corr(), annot=True, cmap="coolwarm")
plt.title("Heatmap")
plt.show()
```



EXTRA

Covariance-correlation

Suppose that you have samples of 2 data: X and Y. Then formula for Covariance and correlation is

$$\mathrm{Cov}(X,Y) = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})$$

$$r = \frac{\mathrm{Cov}(X,Y)}{s_X\, s_Y}$$

$$S_X = \sqrt{\frac{\sum(x_i - \bar{X})^2}{n-1}} \qquad S_Y = \sqrt{\frac{\sum(y_i - \bar{Y})^2}{n-1}}$$

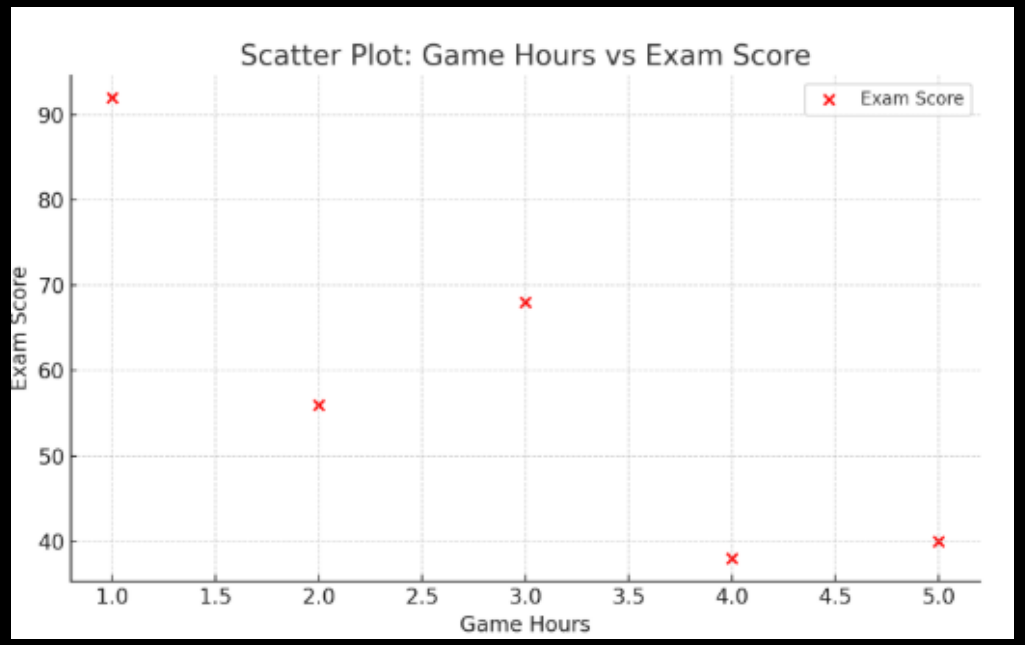
$$r_{XY} = \frac{\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{n}(x_k - \bar{x})^2 \sum_{k=1}^{n}(y_k - \bar{y})^2}}$$

Practice: Following is sample taken from a population. What is correlation coefficient between gamehours and ExamScore ?

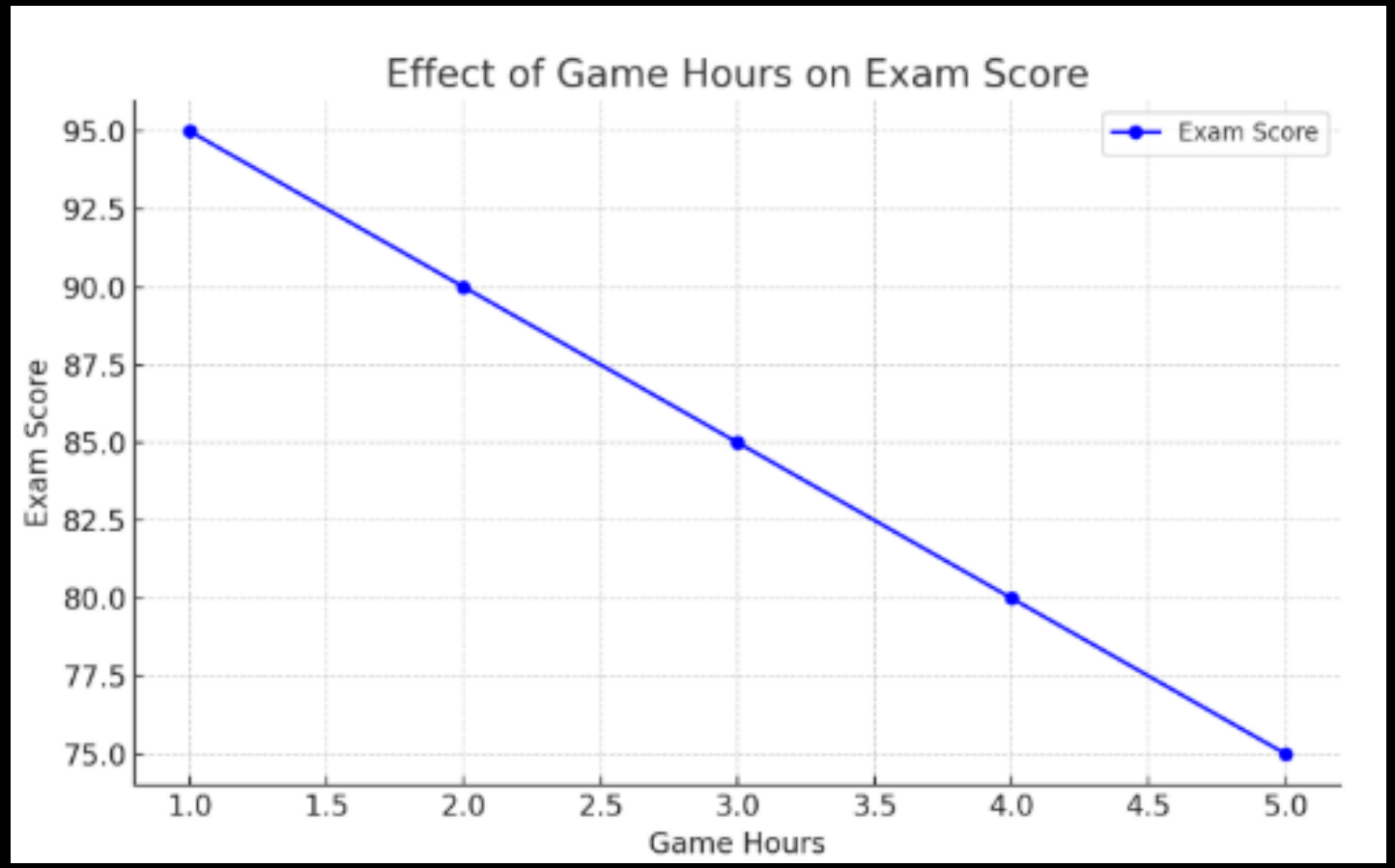
Student	GameHrs (X)	ExamScore (Y)
A	1	92
B	2	56
C	3	68
D	4	38
E	5	40

Performing the same calculation as above, we that the mean of each is 3 and 58.8,
Cov (X,Y) = -24.4

Correlation coefficient = -0.868



If we plot Gamehours versus Scores, we see following perfect linear plot. This is typical of features that have correlation coeff = 1



Covariance + correlation (p1)

Covariance measures how two variables change together—whether they increase or decrease in tandem

Population Covariance Formula

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

Sample Covariance Formula

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

Where:

- n = number of data points
- X_i, Y_i = the i-th observation of variables X and Y
- μ_X or $\bar{X}$ = mean of X
- μ_Y or $\bar{Y}$ = mean of Y

- The covariance matrix tells us **how features vary together**. For two features X and Y:

$$\text{Cov}(X,Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

- In matrix form:

$$\textbf{Covariance matrix} = \mathbf{C} = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1,X_2) & \dots \\ \text{Cov}(X_2,X_1) & \text{Var}(X_2) & \dots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

What does this matrix represent?

1. The **diagonal elements = variances of individual features** → how much that feature spreads out.
2. The **off-diagonal elements = covariances** → how two features change together:

Negative covariance: when one increases, the other decreases….X1 = age and X2 = number of pushups

Zero covariance: no linear relationship…..X1 = your age and X2 = neighbors income

Positive covariance: they increase together…. X1 = your age, X2 = income

For example X1 could be age, X2 could income and we may see that Cov(X1,X2) be +ve: as age goes up, income goes up to.

You will also see following formula in literature:

Covariance + correlation EXAMPLE (jupyter n/b)

Example: Suppose we're looking at:
GameHrs: The number of hours a student plays video games per day
ExamScore: The student's exam score out of 100.
Below is a **sample** taken from a population. Find covariance and correlation coeff. between the 2 features.

Student	GameHrs (X)	ExamScore (Y)
A	1	95
B	2	90
C	3	85
D	4	80
E	5	75

We expect that as GameHours (video games) increase, ExamScore decreases → negative covariance. Let's see if can deduce this mathematically.

Here we would use sample covariance formula.

Compute means

$$\bar{X} = (1+2+3+4+5)/5 = 3$$

$$\bar{Y} = (95+90+85+80+75)/5 = 85$$

$$cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

$$= \frac{1}{4}[(x_1-\bar{x})(y_1-\bar{y}) + (x_2-\bar{x})(y_2-\bar{y}) + (x_3-\bar{x})(y_3-\bar{y}) + (x_4-\bar{x})(y_4-\bar{y})]$$

$$= \frac{1}{4}[(1-3)(95-85) + (2-3)(90-85) + (3-3)(85-85) + (4-3)(80-85)]$$

$$= \frac{1}{4}[-50] = -12.5$$

GameHrs(X)	ExamScore(Y)	(X - 3)	(Y - 85)	Product
1	95	-2	10	-20
2	90	-1	5	-5
3	85	0	0	0
4	80	1	-5	-5
5	75	2	-10	-20

So the cova. is -12.5 hours - score point

Cont....

...cont Covariance + correlation EXAMPLE (jupyter n/b)

Why do we need correlation coefficient (r) ρ?

Covariance has units:

- Example: if X = meters, Y = seconds → cov(X,Y) is in meter·second
- This makes it hard to compare across different datasets that may have different units

The formula for correlation coefficient is

$$\text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}}$$

$$\text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Key point on **r**:
It is dimensionless (no units) and ∈ [-1, 1]
+1 → perfect positive linear relationship
0 → no linear relationship
-1 → perfect negative linear relationship

Continuing with previous example,

Std dev of X

$$\sigma_X = \sqrt{\frac{(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2}{5-1}}$$

$$= \sqrt{\frac{4+1+0+1+4}{4}} = \sqrt{\frac{10}{4}} = \sqrt{2.5} \approx 1.58$$

Std dev of Y

$$\sigma_Y = \sqrt{\frac{(95-85)^2 + (90-85)^2 + (85-85)^2 + (80-85)^2 + (75-85)^2}{4}}$$

$$= \sqrt{\frac{100+25+0+25+100}{4}} = \sqrt{\frac{250}{4}} = \sqrt{62.5} \approx 7.91$$

We already calculated:

$$\text{cov}(X,Y) = -12.5$$

Correlation coefficient

$$r = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{-12.5}{1.58 \times 7.91} = \frac{-12.5}{12.5} = -1.0$$