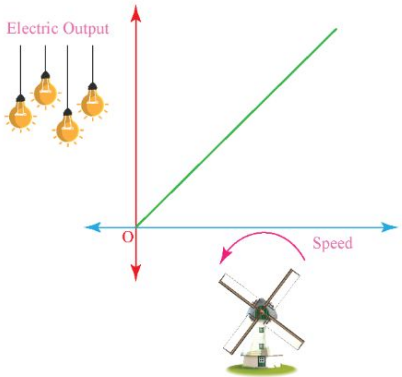
Data Curation Techniques

Siddharth R

Pearson Correlation Coefficient

- To measure how strong a relationship is between two variables.
- The Pearson Correlation Coefficient (PCC), often denoted as r, is a statistical measure that quantifies the **linear** relationship between two variables.
- It tells us both the **strength** and **direction** of the association between them.



For example:

- Does the amount of time spent studying (X) affect exam scores (Y)?
- Is there a connection between height (X) and weight (Y)?

Assumptions: Linearity, Continuous data, Without outlier, Normally distributed

Pearson Correlation Coefficient

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

n : sample size

x_i, y_i : individual sample points indexed with i

$\bar{x}, \bar{y}$: sample mean

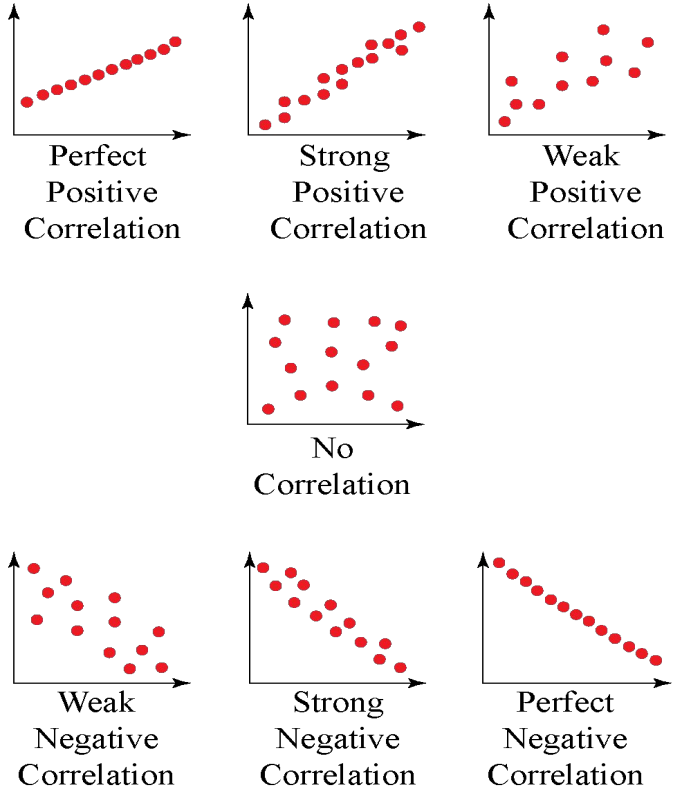
Pearson Correlation Coefficient

Interpretation

The value of r always lies between -1 and +1:

- r = +1 → Perfect positive correlation: As X increases, Y always increases proportionally.
- r = -1→ Perfect negative correlation: As X increases, Y always decreases proportionally.
- r = 0 → No linear correlation: Changes in X do not predict changes in Y.

Pearson Correlation Coefficient



Sample Correlation Strength

Correlation Coefficient Size (r)	Correlation Strength
.91 to 1.00 or -.91 to -1.00	Very Strong
.71 to .90 or -.71 to -.90	Strong
.51 to .70 or -.51 to -.70	Medium
.31 to .50 or -.31 to -.50	Weak
.01 to .30 or -.01 to -.30	Very Weak
.00	No Correlation

Summary

- Variance vs Covariance vs Correlation
- Variance: Understanding data spread — like how diverse students' test scores are.
- Covariance: Checking direction of relationship — like whether hours studied and exam scores tend to move together.
- Correlation: Assessing strength of relationship — like how strongly connected savings rate and investment returns are.
- Variance and Covariance is scale dependent where correlation is not

Thank You !!!