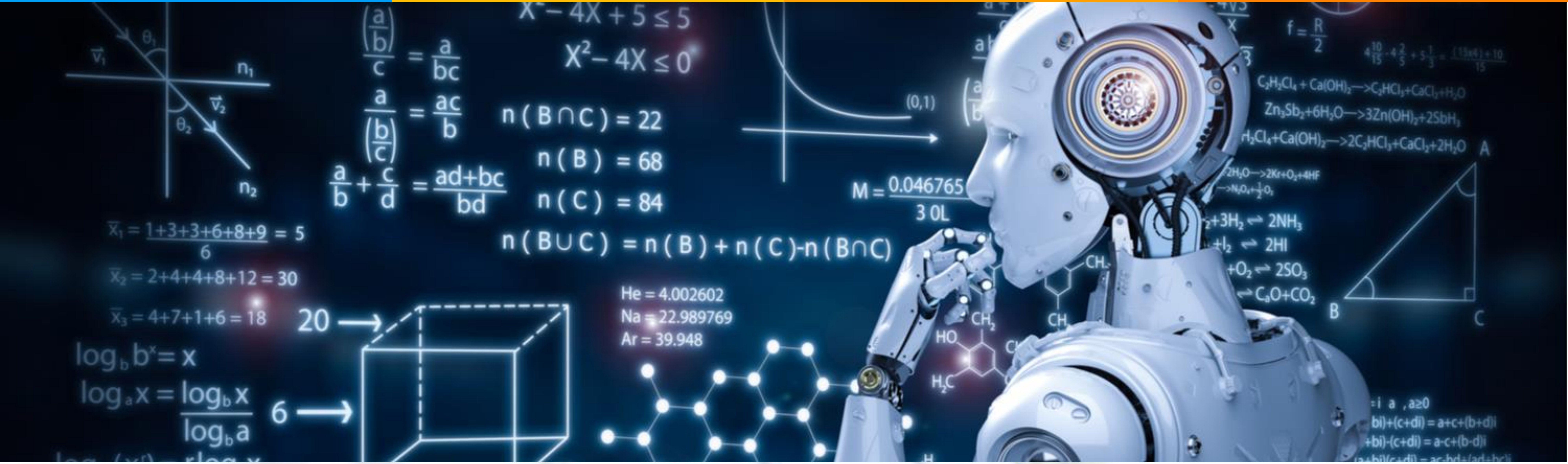
Duration: 60 min | Srikanth Shenoy | AML 5101 | M.E. AI & ML



$X^2 - 4X + 5 \le 5$
$X^2 - 4X \le 0$
$\frac{\left(\frac{a}{b}\right)}{c} = \frac{a}{bc}$
$\frac{a}{\left(\frac{b}{c}\right)} = \frac{ac}{b}$
$\frac{a}{b} + \frac{c}{d} = \frac{ad+bc}{bd}$
$n(B \cap C) = 22$
$n(B) = 68$
$n(C) = 84$
$n(B \cup C) = n(B) + n(C) - n(B \cap C)$
$\bar{x}_1 = \frac{1+3+3+6+8+9}{6} = 5$
$\bar{x}_2 = 2+4+4+8+12 = 30$
$\bar{x}_3 = 4+7+1+6 = 18$
$\log_b b^x = x$
$\log_a x = \frac{\log_b x}{\log_b a}$
20
6
(0,1)
$M = \frac{0.046765}{3\,0L}$
He = 4.002602
Na = 22.989769
Ar = 39.948
$f = \frac{R}{2}$
n_1
n_2
A
B
C

Lecture 07 – Correlation & Autocorrelation

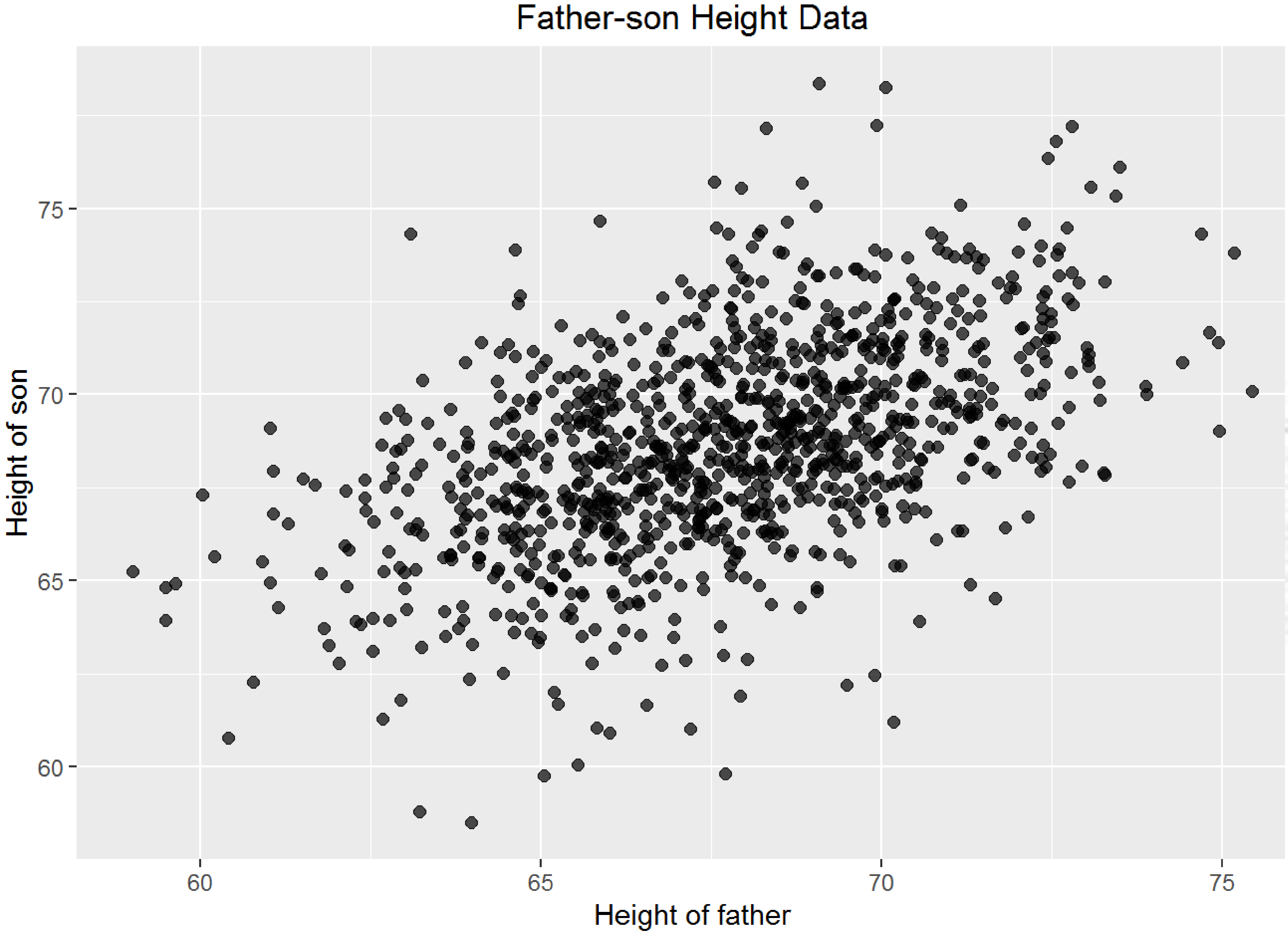
Recap

- RMS, SD, Chebyshev inequality

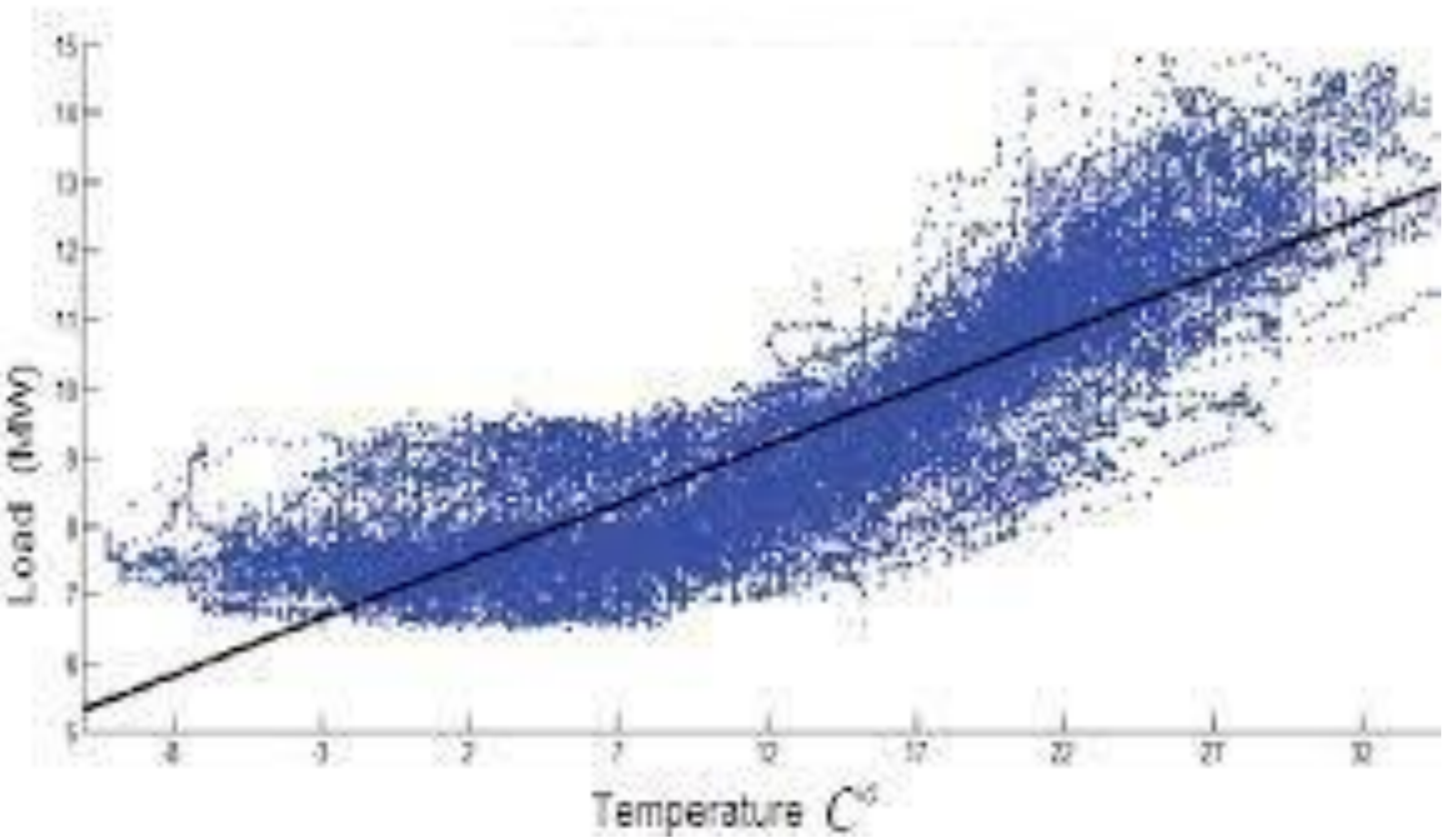
1. Correlation

Correlation

- Father-son heights

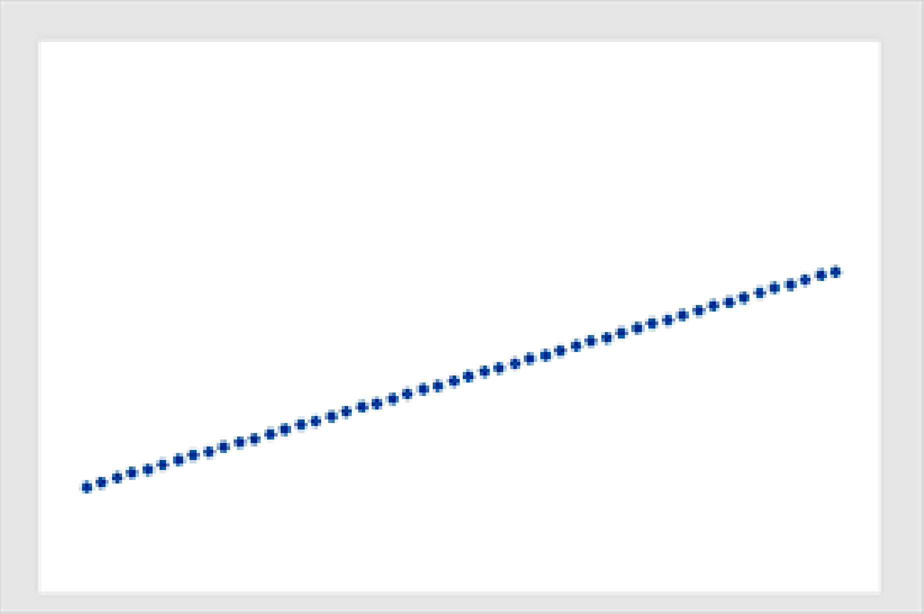
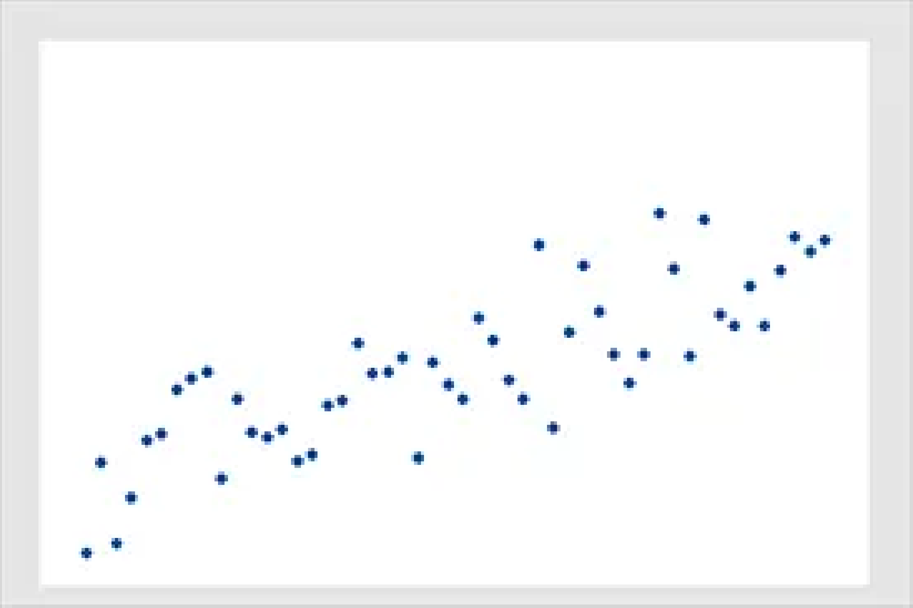
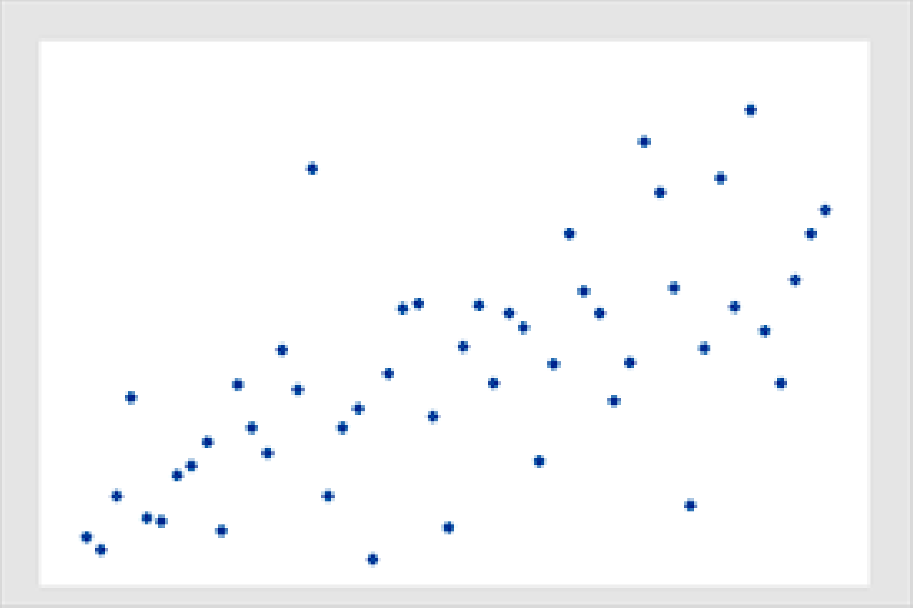
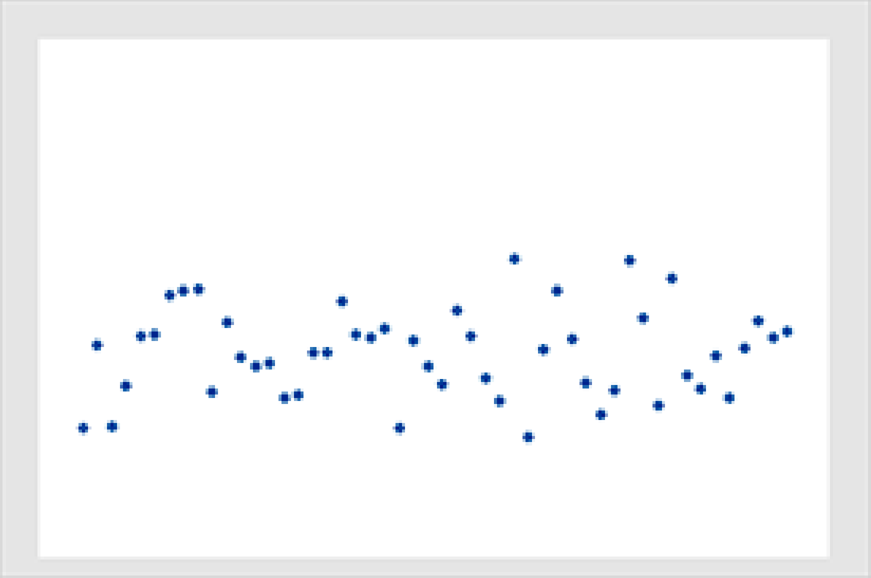


- Temperature-Electric bill

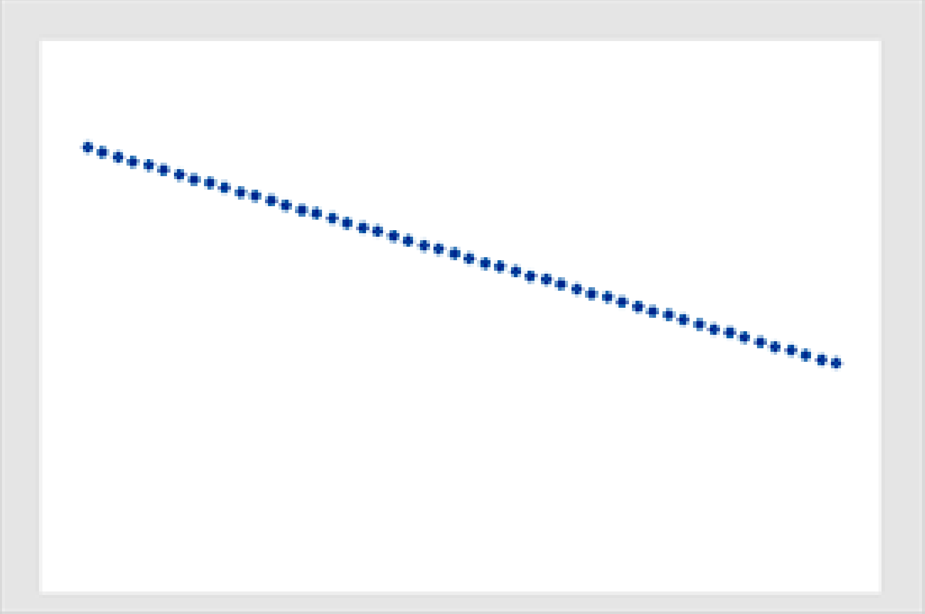
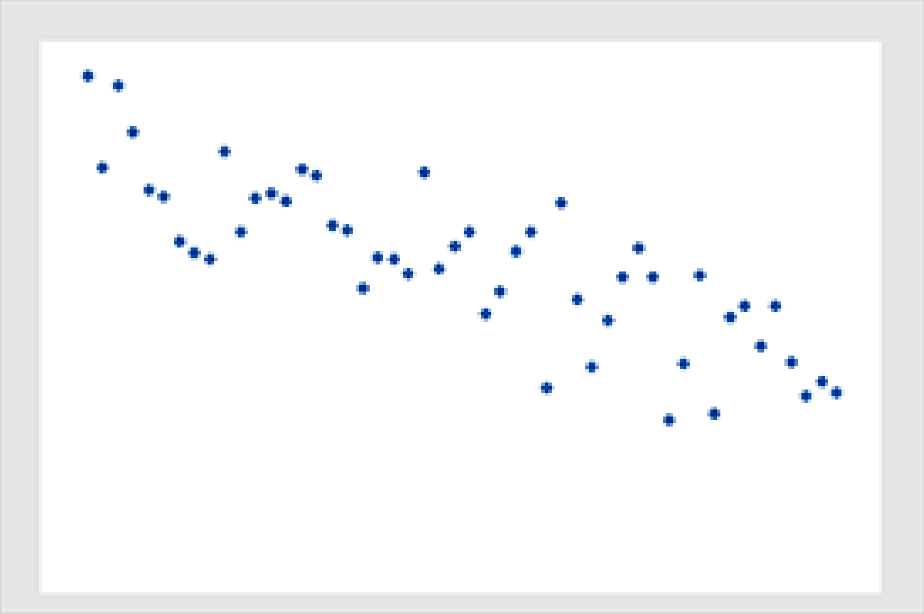
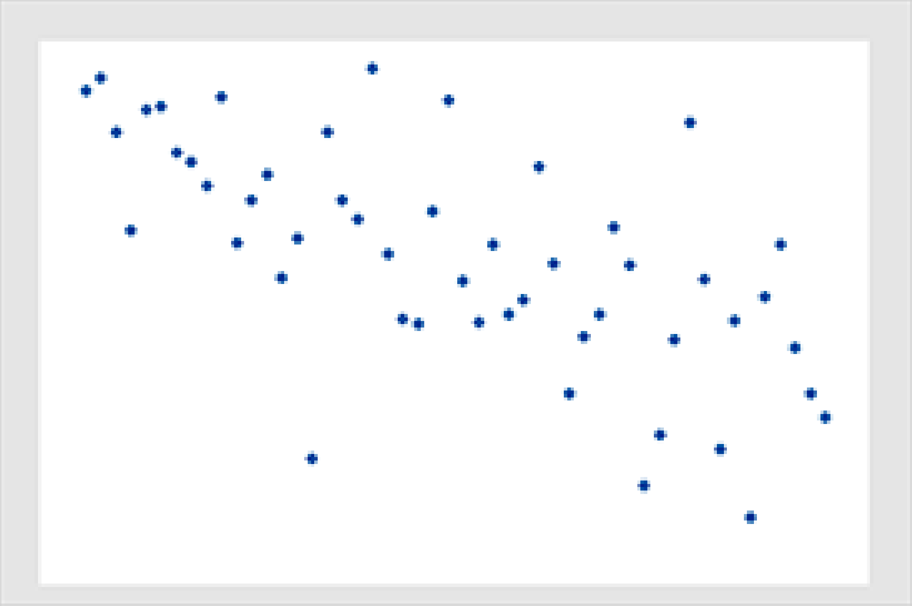


Correlation strength & coefficients

- Very Strong Strong Moderate None

- 1 0.8 0.6 0
- -1 -0.8 -0.6

Correlation coefficient

Simple form

$$\rho = \frac{(a-\bar{a})^T(b-\bar{b})}{\|a-\bar{a}\|\|b-\bar{b}\|}$$

Denominator for normalizing between -1 and 1

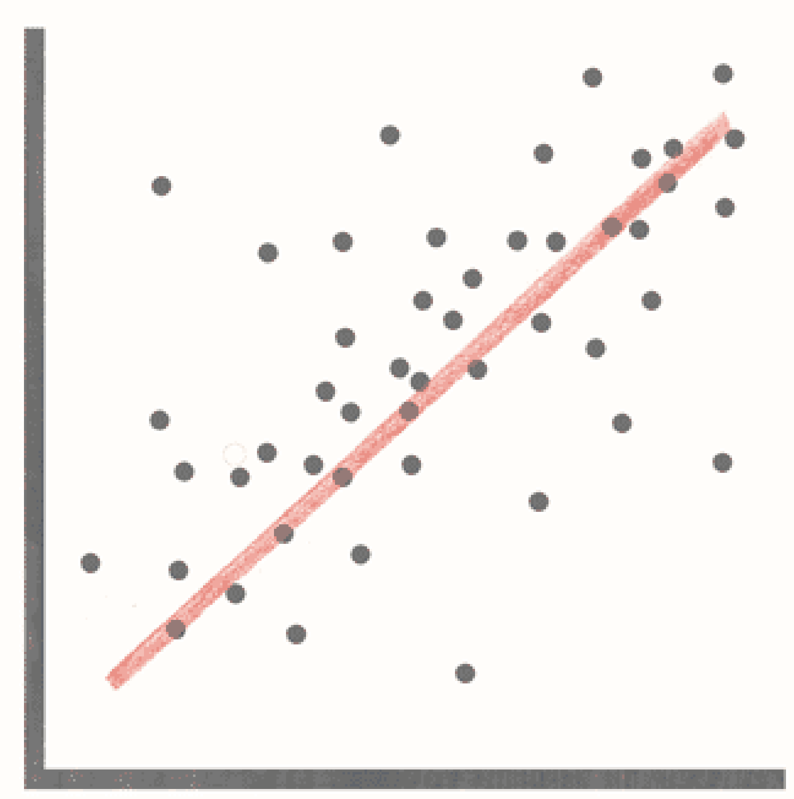
Nasty form

$$\rho = \frac{(a-\frac{1}{n}\mathbf{1}^T a\mathbf{1})^T(b-\frac{1}{n}\mathbf{1}^T b\mathbf{1})}{\|a-\frac{1}{n}\mathbf{1}^T a\mathbf{1}\|\|b-\frac{1}{n}\mathbf{1}^T b\mathbf{1}\|}$$

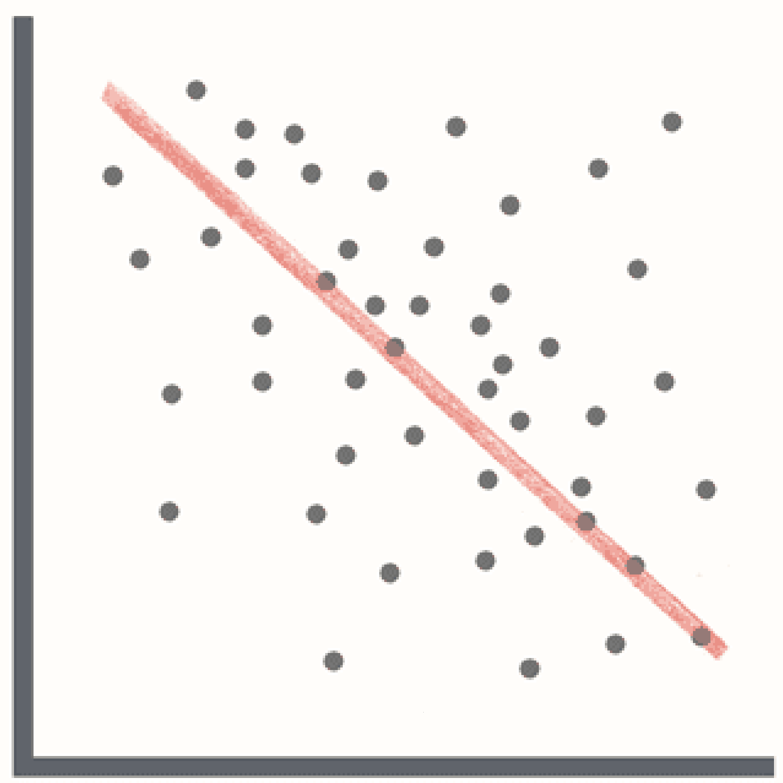
- Why not $\rho = \frac{a^T b}{\|a\|\|b\|}$

After all, dot product measures similarity

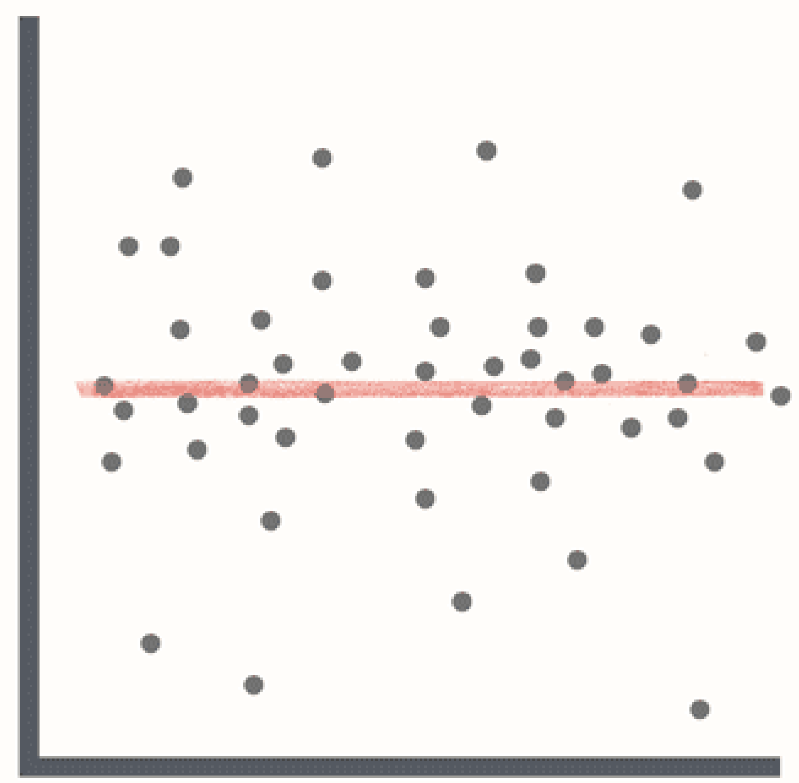
Intuition behind mean centering



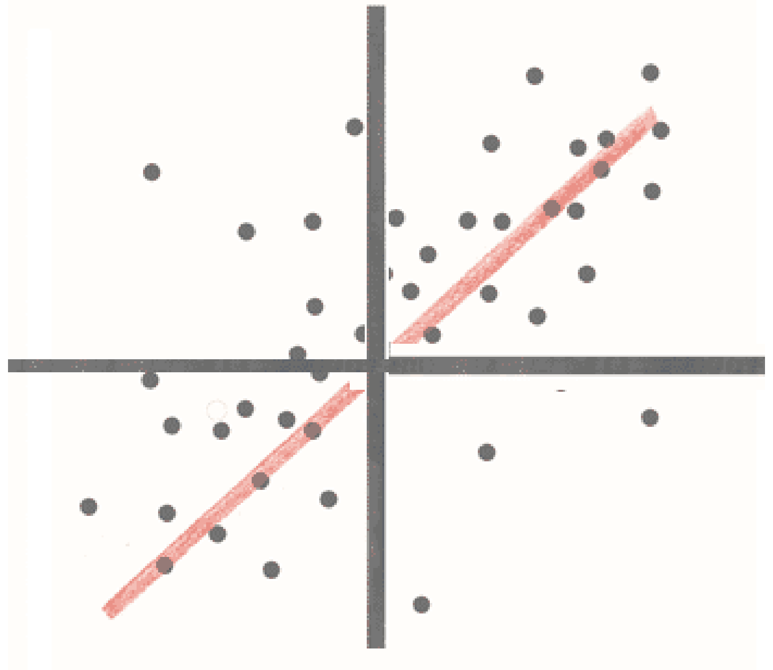
Positive Correlation



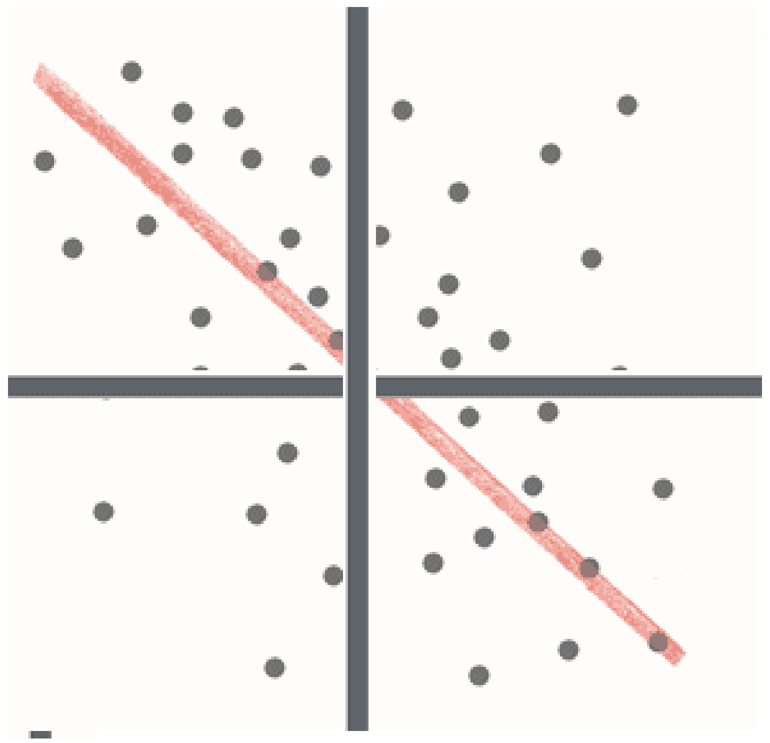
Negative Correlation

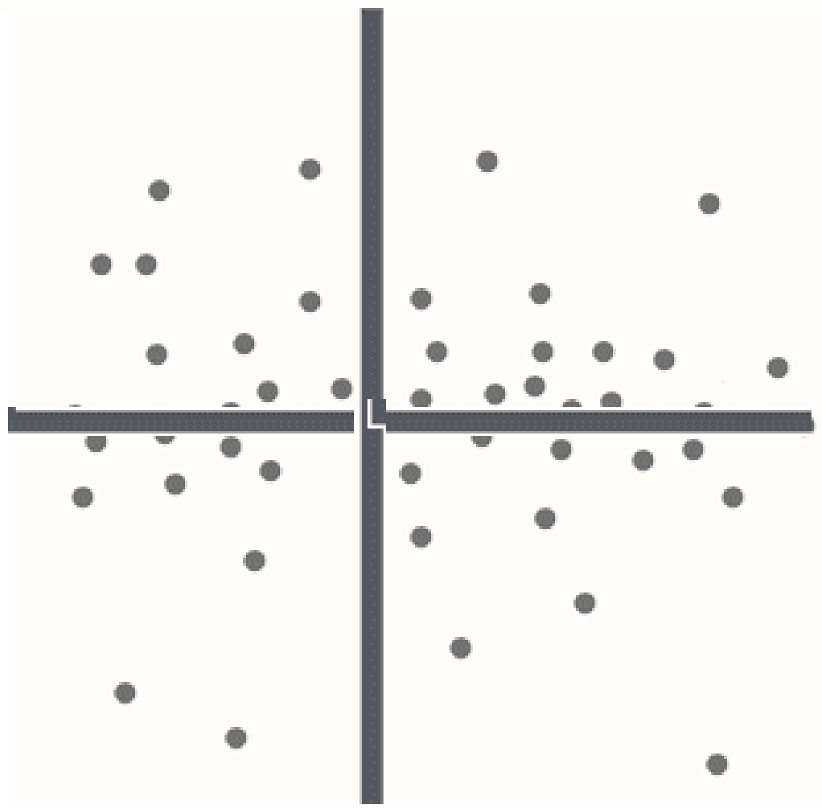


No Correlation



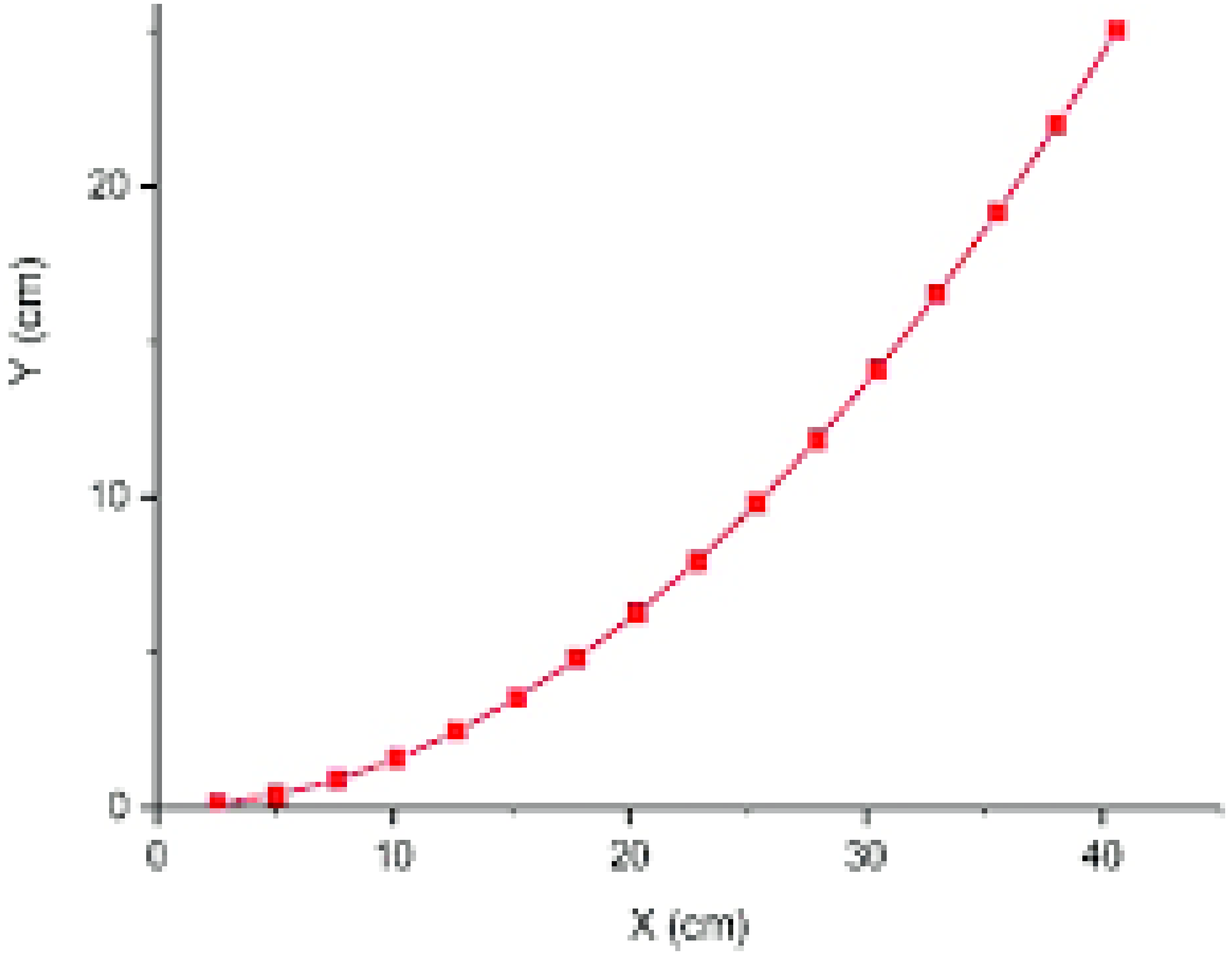
Positive Correlation

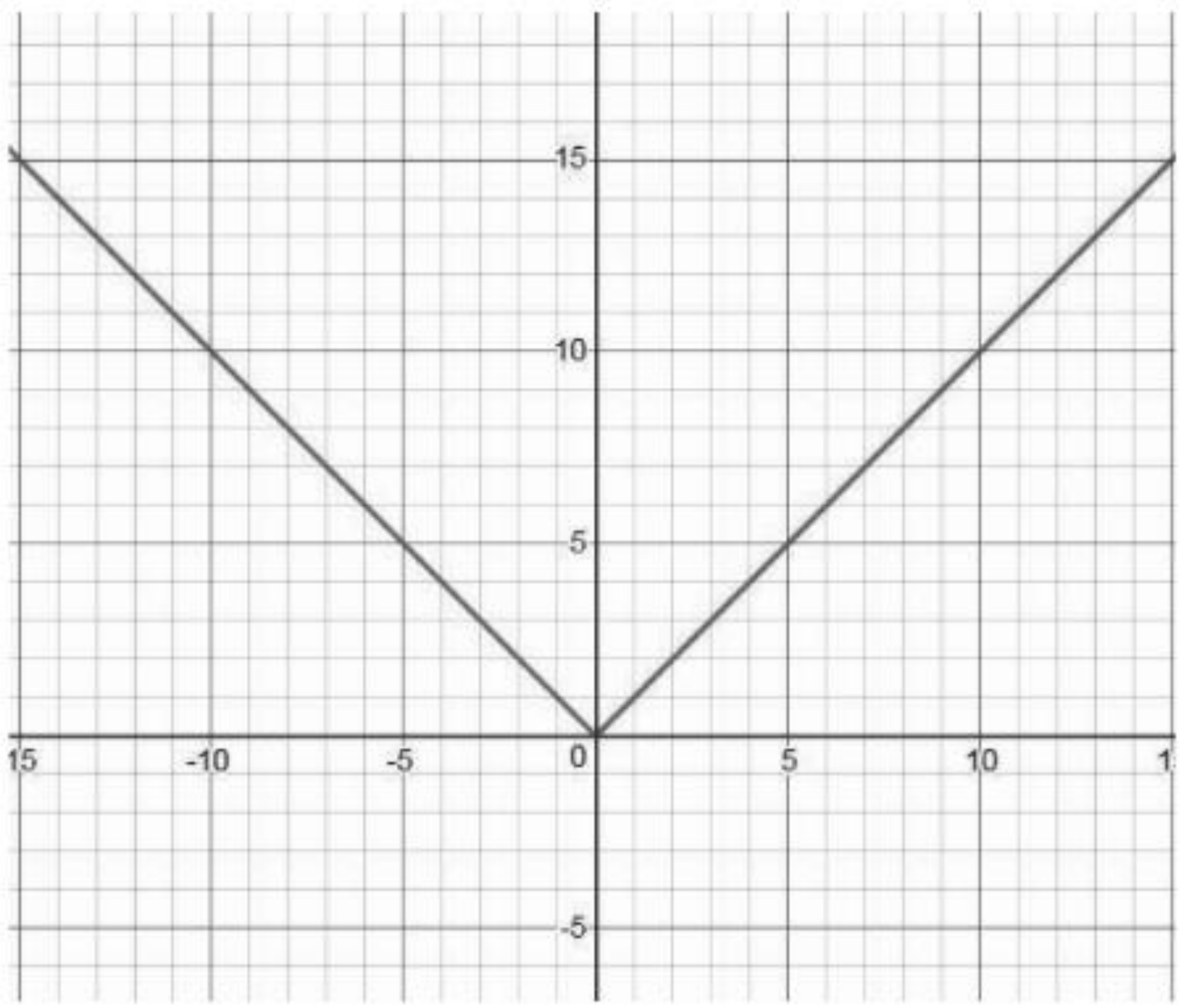




What is the correlation coefficient ?

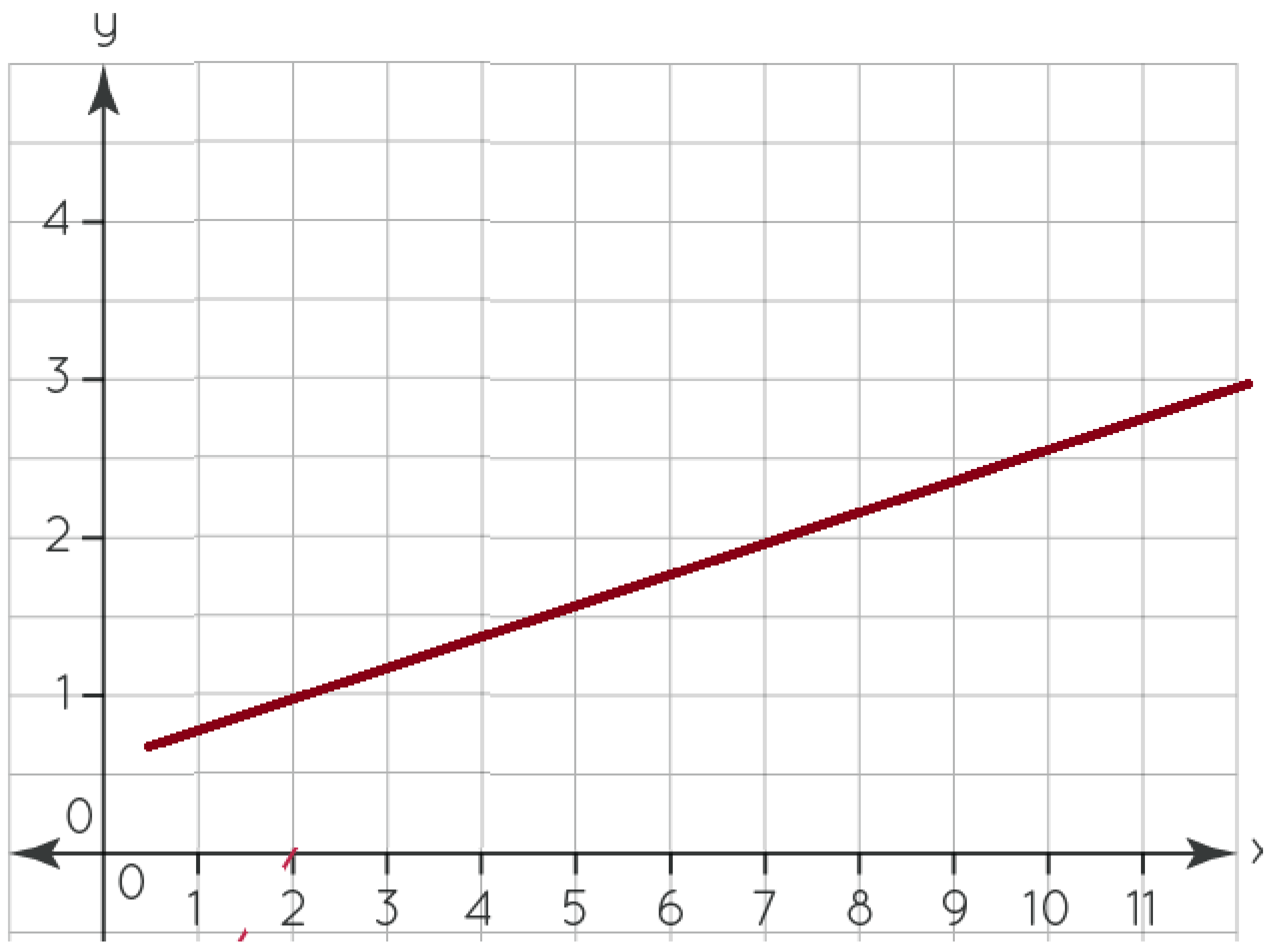
- Imagine two datasets with points **perfectly** along the curve
- What is the correl coeff?

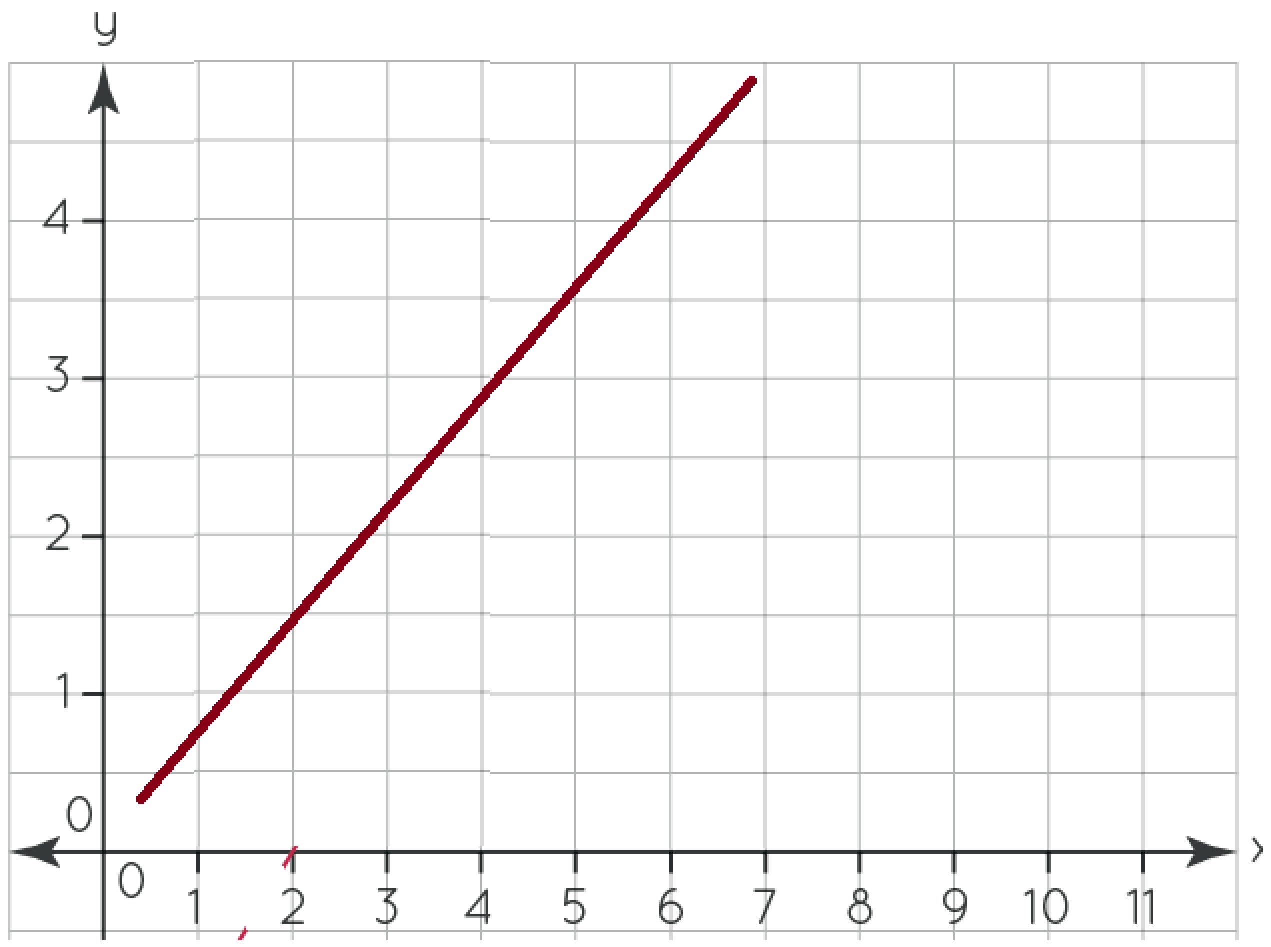




Which has more correlation coefficient ?

- Imagine two datasets where points are **perfectly** along these lines
- Which has more correlation coefficient?





Correlation coefficient other forms

In Linear Algebra

$$\rho = \frac{(a-\bar{a})^T(b-\bar{b})}{\|a-\bar{a}\|\|b-\bar{b}\|}$$

Cov(a,a) = Var(a)

$$\rho = \frac{u^T v}{n} \quad \Longleftarrow$$

$$where \quad u = \frac{a-\bar{a}}{\sigma_a} \quad v = \frac{b-\bar{b}}{\sigma_b}$$

In Statistics

$$Cov(a,b) = \frac{\Sigma_{i=1}^{n}(a-\bar{a})(b-\bar{b})}{n}$$

$$\rho = Correl(a,b) = \frac{Cov(a,b)}{\sigma_a \sigma_b}$$

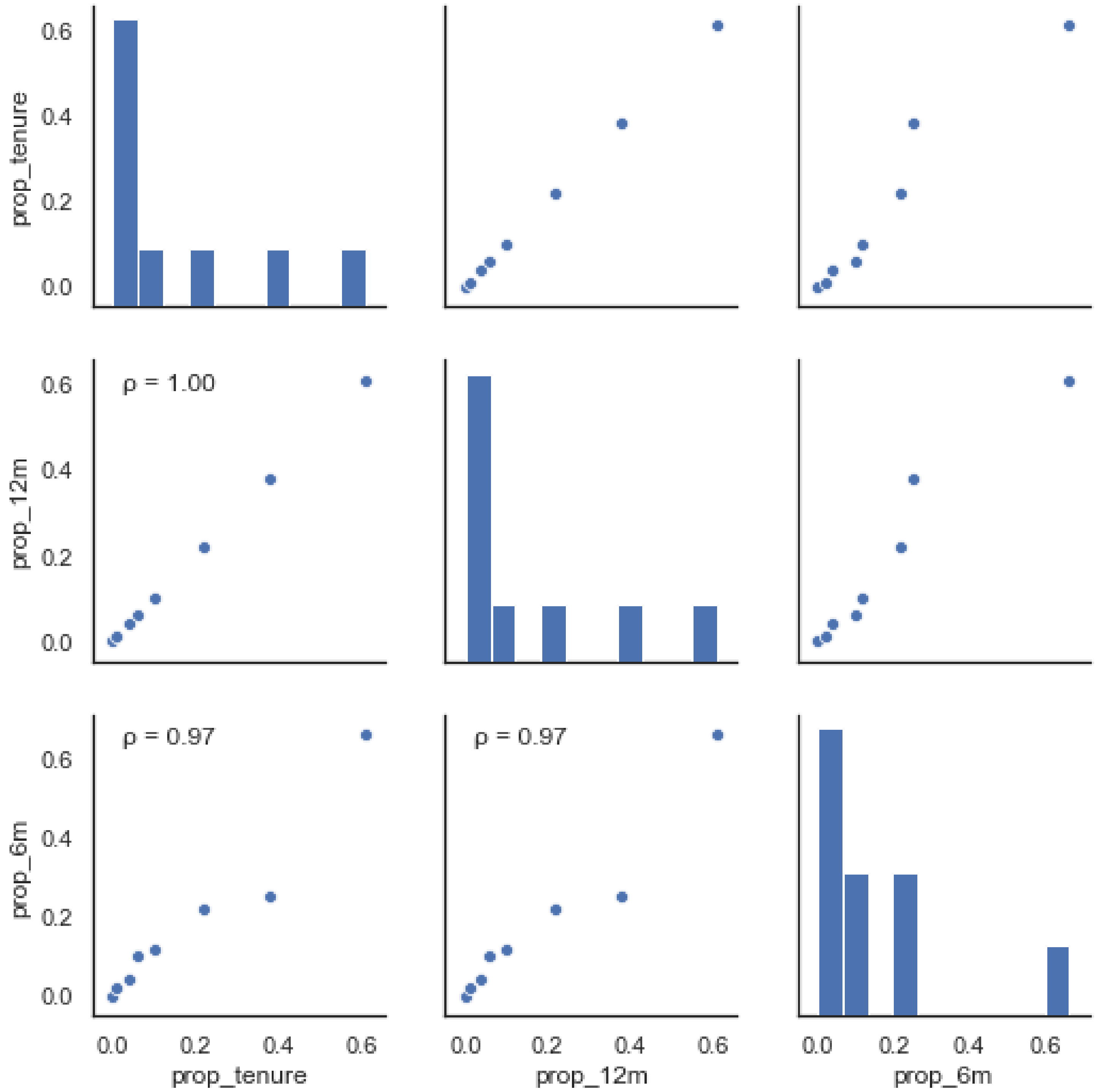
$$\rho = \frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{a_i-\bar{a}}{\sigma_a}\right)\left(\frac{b_i-\bar{b}}{\sigma_b}\right)$$

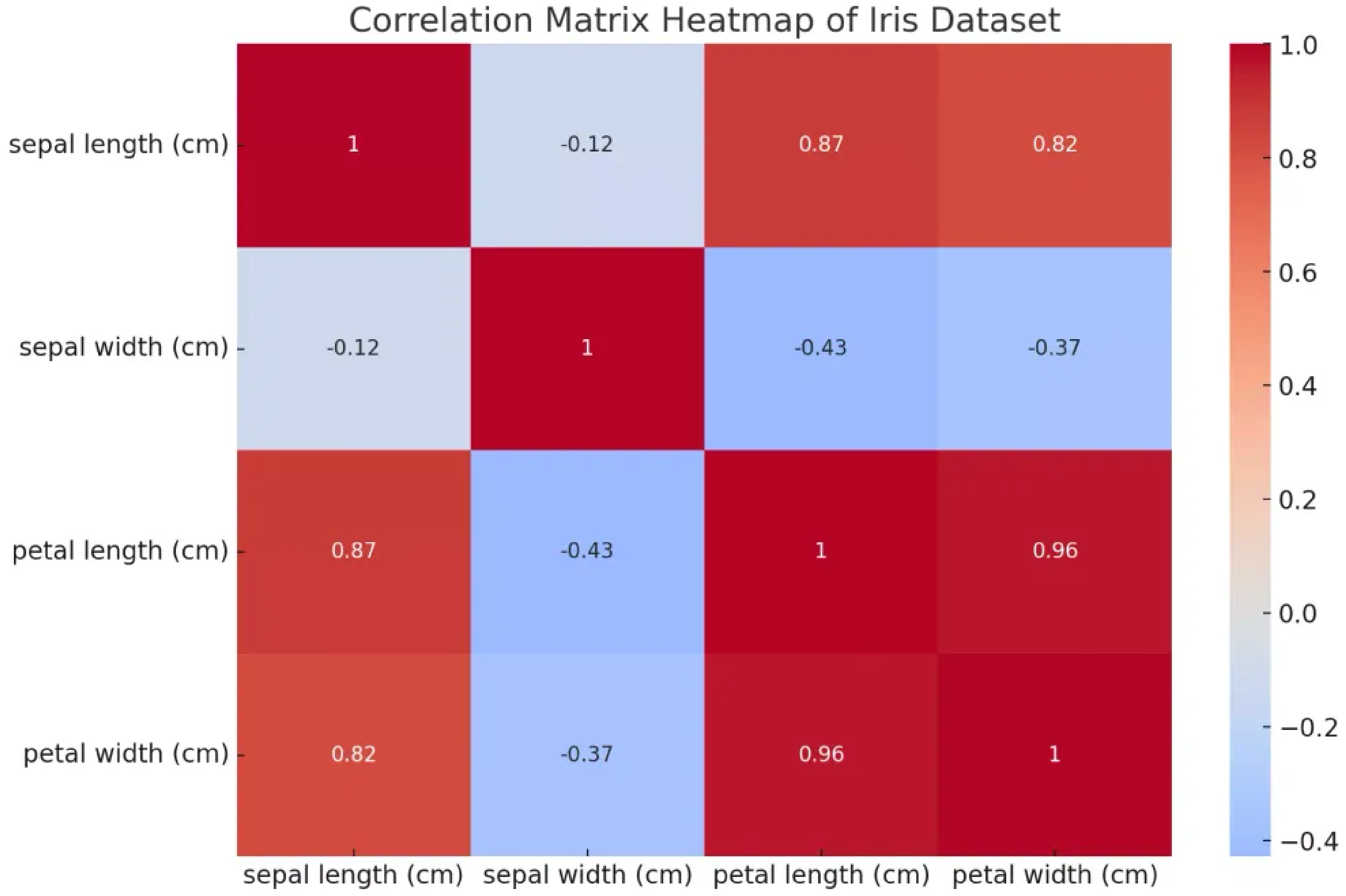
Covariance & Correlation in numpy

- Covariance: np.cov()
  - Returns covariance matrix
- Correlation Coefficient: np.corrcoef(a, b)
  - Two 1-D vectors passed
  - Supports only Pearson correlation coefficient
- There are two more – Spearman and Kendall
  - When are they used? - Reading assignment
- What is the relation to np.correlate()?
- Pandas corr()

Visualizing correlation

- Seaborn pair plots and heatmaps





Correlation is not causation

Correlation is not causation

- Ice cream sale increases as summer heat increases
- Shark attack increases as summer heat increases
- Ice cream sale is highly correlated to shark attacks
- Wrong to conclude ice cream sale caused shark attack

$X^2 - 4X + 5 \le 5$
$X^2 - 4X \le 0$
$\frac{\left(\frac{a}{b}\right)}{c} = \frac{a}{bc}$
$\frac{a}{\left(\frac{b}{c}\right)} = \frac{ac}{b}$
$\frac{a}{b} + \frac{c}{d} = \frac{ad+bc}{bd}$
$n(B \cap C) = 22$
$n(B) = 68$
$n(C) = 84$
$n(B \cup C) = n(B) + n(C) - n(B \cap C)$
$\bar{x}_1 = \frac{1+3+3+6+8+9}{6} = 5$
$\bar{x}_2 = 2+4+4+8+12 = 30$
$\bar{x}_3 = 4+7+1+6 = 18$
$\log_b b^x = x$
$\log_a x = \frac{\log_b x}{\log_b a}$
20 →
6 →
He = 4.002602
Na = 22.989769
Ar = 39.948
$M = \frac{0.046765}{3\ 0L}$
(0,1)
$f = \frac{R}{2}$
$Zn_3Sb_2 + 6H_2O \longrightarrow 3Zn(OH)_2 + 2SbH_3$
$2NH_3$
2HI
$2SO_3$
A
B
C
HO
CH_3
H_3C

2. Correlation between time series signals

Auto correlation

- Correlation of data with itself
- Auto-correlation is always 1
- What if it is time lagged version of itself?

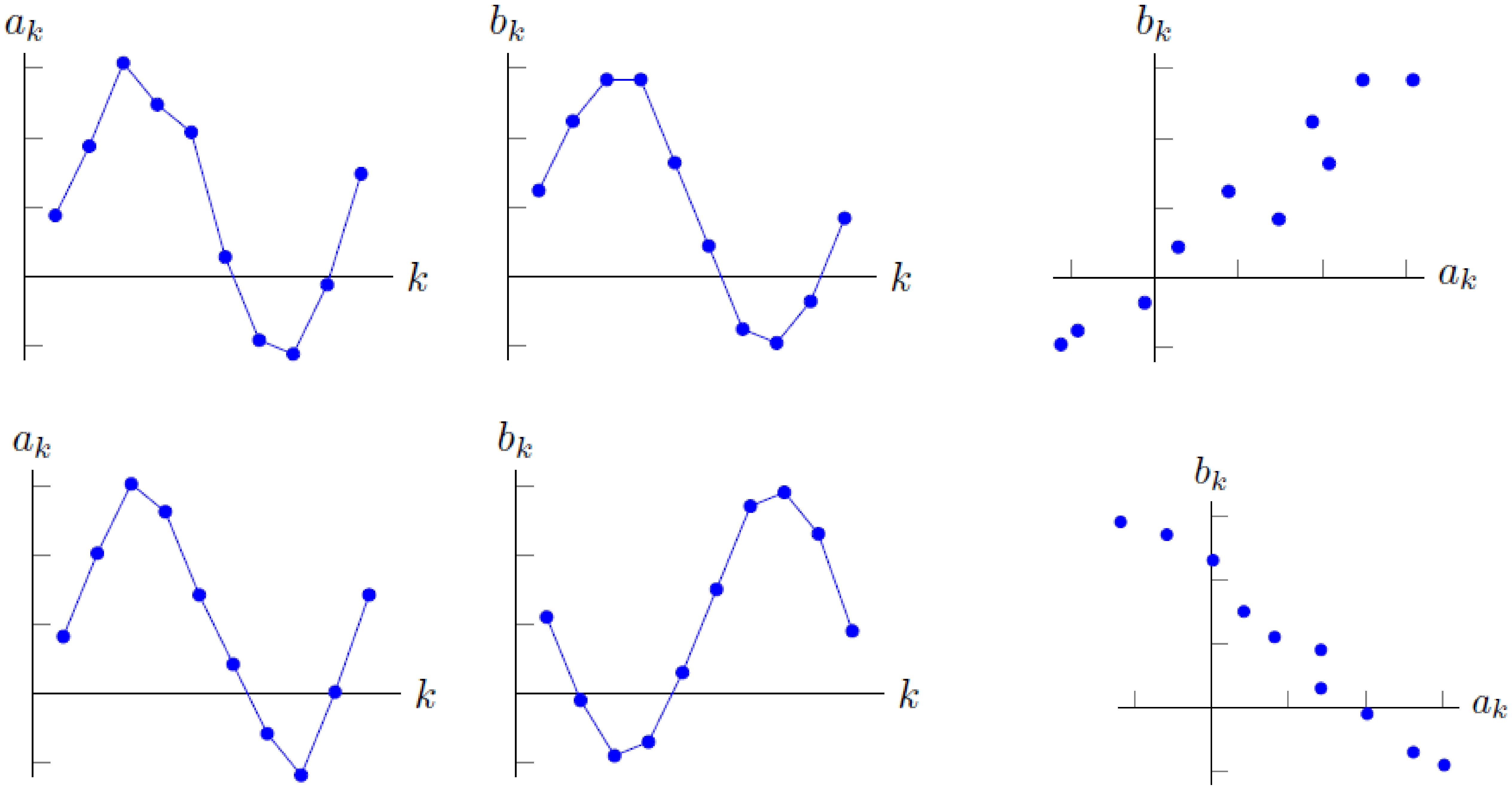
Auto correlation

- Correlation of time series signal with a lagged version of itself

$$r_k = \frac{\sum_{t=k+1}^{N}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{N}(y_t - \bar{y})^2}$$

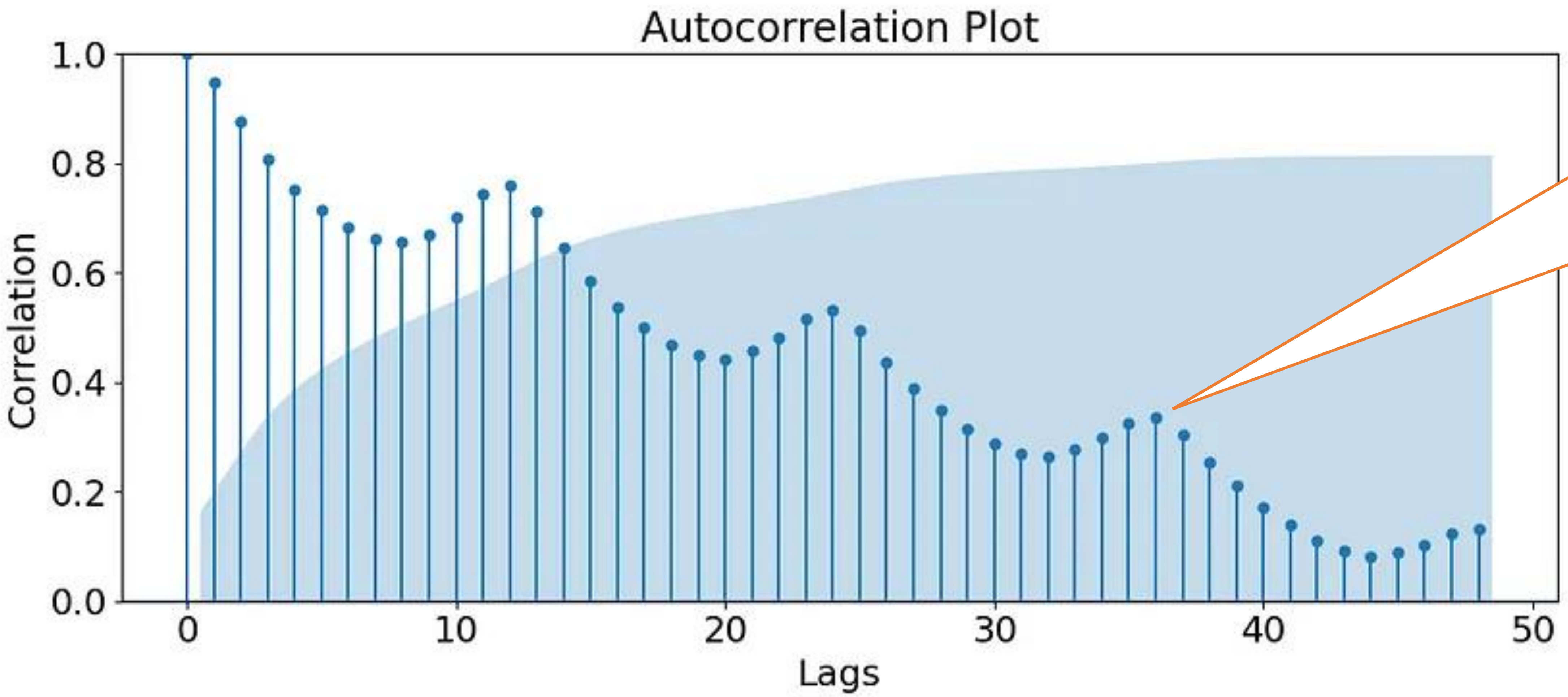
Auto correlation

- Correlation of time series signal with a lagged version of itself is often useful



Auto correlation

- Air pollution timeseries auto correlation



Why is autocorrelation decreasing even for 12 month cycle with the passage of time?

- Notice the cyclical pattern in multiple of 12

Correlation between time series

- Stock market correlation between stocks of same sector
- If they are not showing strong correlation, then something is wrong or some new opportunity
- Always analyze correlation with volatility (coefficient of standard deviation over a period)

$X^2-4X+5\leq5$
$X^2-4X\leq0$
$n(B\cap C)=22$
$n(B)=68$
$n(C)=84$
$n(B\cup C)=n(B)+n(C)-n(B\cap C)$
$\frac{a}{b}+\frac{c}{d}=\frac{ad+bc}{bd}$
$\bar{x}_1=\frac{1+3+3+6+8+9}{6}=5$
$\bar{x}_2=2+4+4+8+12=30$
$\bar{x}_3=4+7+1+6=18$
$\log_b b^x=x$
$\log_a x=\frac{\log_b x}{\log_b a}$
20 →
6 →
He = 4.002602
Na = 22.989769
Ar = 39.948
$M=\frac{0.046765}{3.0L}$
(0,1)
$f=\frac{R}{2}$
A
B
C

2. Standard deviation of sum of vectors

Standard deviation of the sum

- a and b are two equal sized vectors
- In statistical terms a and b are realizations of A and B

$$std(a+b) = \sqrt{std(a)^2 + std(b)^2 + 2\rho std(a) std(b)}$$

- Special cases:
  - Correlation coefficient = 1, -1 and 0

Standard deviation of the sum

$$std(a+b) = \sqrt{std(a)^2 + std(b)^2 + 2\rho std(a)std(b)}$$

- Correlation coefficient = 1

$$std(a+b) = std(a) + std(b)$$

- Correlation coefficient = -1

$$std(a+b) = std(a) - std(b)$$

- Correlation coefficient = 0 (Uncorrelated features in ML)

$$std(a+b) = \sqrt{std(a)^2 + std(b)^2}$$

Hedging investments

- Invest in two assets with same return (μ) & risk(σ)
- Asset returns over 5 year period in a and b vectors
- Hedged investment

$$c = \frac{a+b}{2} \qquad avg(c) = avg\left(\frac{a+b}{2}\right) = \mu \qquad \sigma_{\frac{a}{2}} = \sigma_{\frac{b}{2}} = \frac{\sigma}{2}$$

$$std(c) = \sqrt{std(\frac{a}{2})^2 + std(\frac{b}{2})^2 + 2\rho std(a)std(b)}$$

$$= \frac{\sqrt{2\sigma^2 + 2\rho\sigma^2}}{2} \qquad = \frac{\sigma}{\sqrt{2}}\sqrt{1+\rho}$$

Two special cases:
Rho = 0 and 1

Brief plan of what is next

- Remaining topics between chapter 1 & 3:
  - Linear Combinations

QUESTIONS

Literature Survey