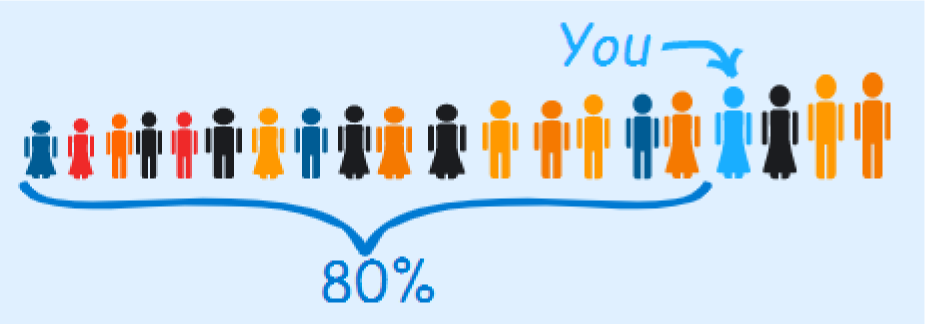
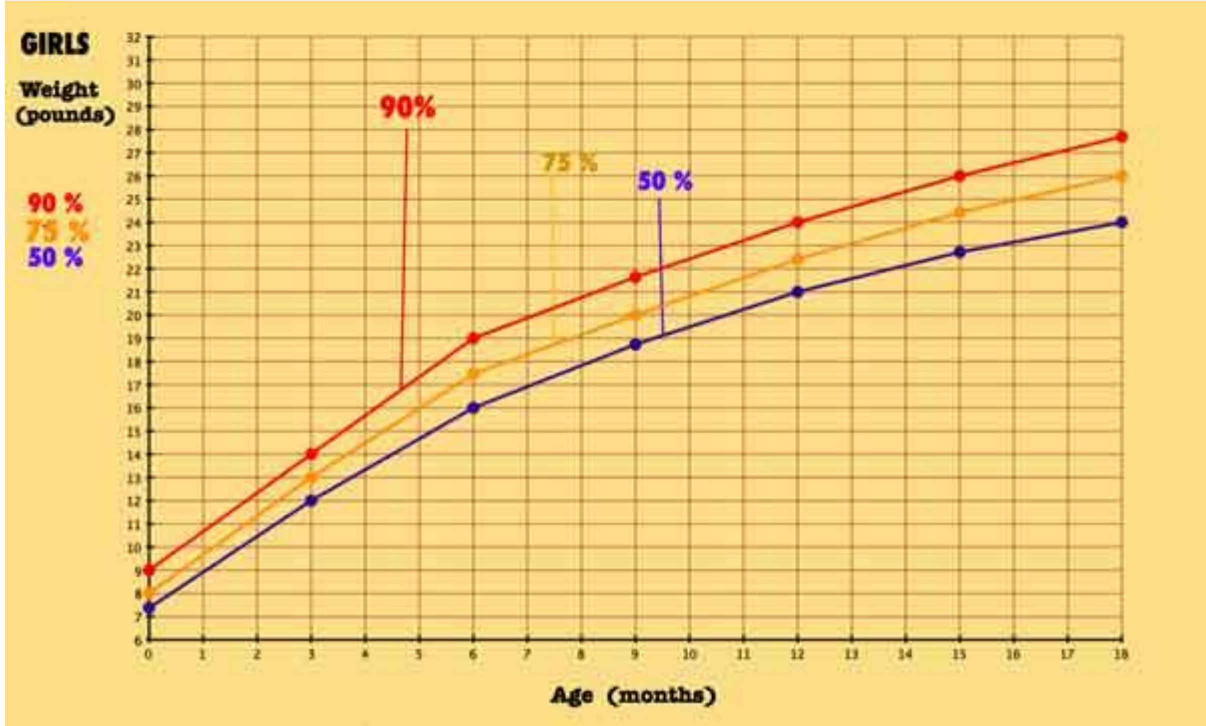
Percentiles





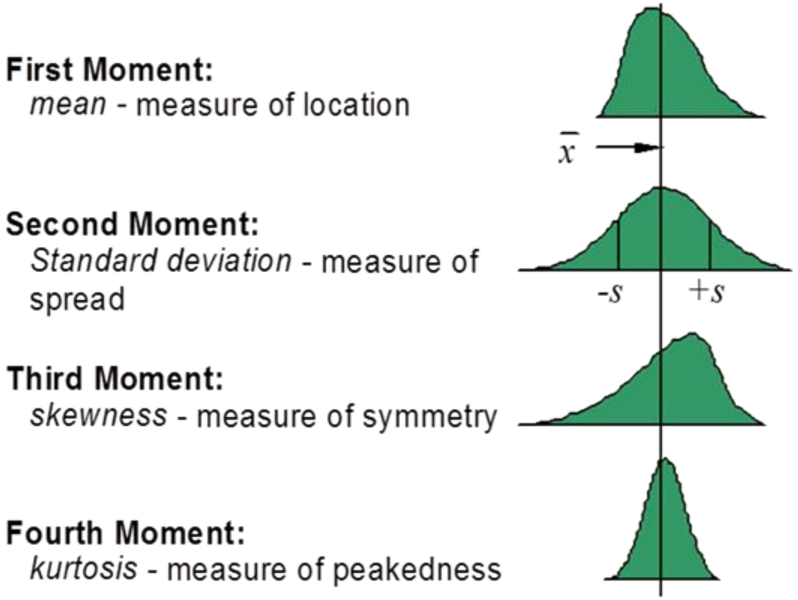
7. Percentiles.ipynb

Moments

- Quantitative measures of the **shape** of a probability density function
- Mathematically:

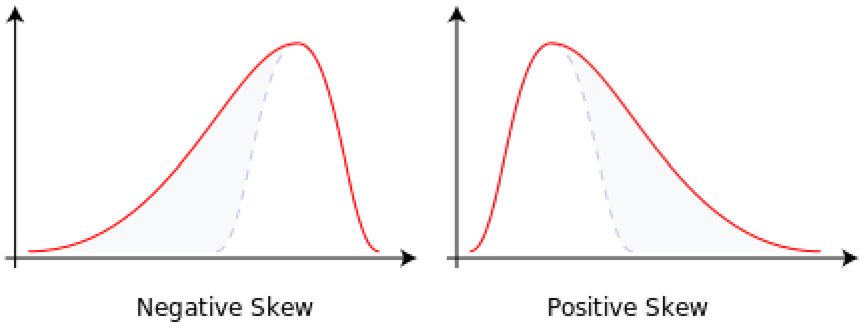
$$\mu_n = \int_{-\infty}^{\infty}(x-c)^n f(x)dx \text{ (for moment } n \text{ around value } c\text{)}$$

(Variance)

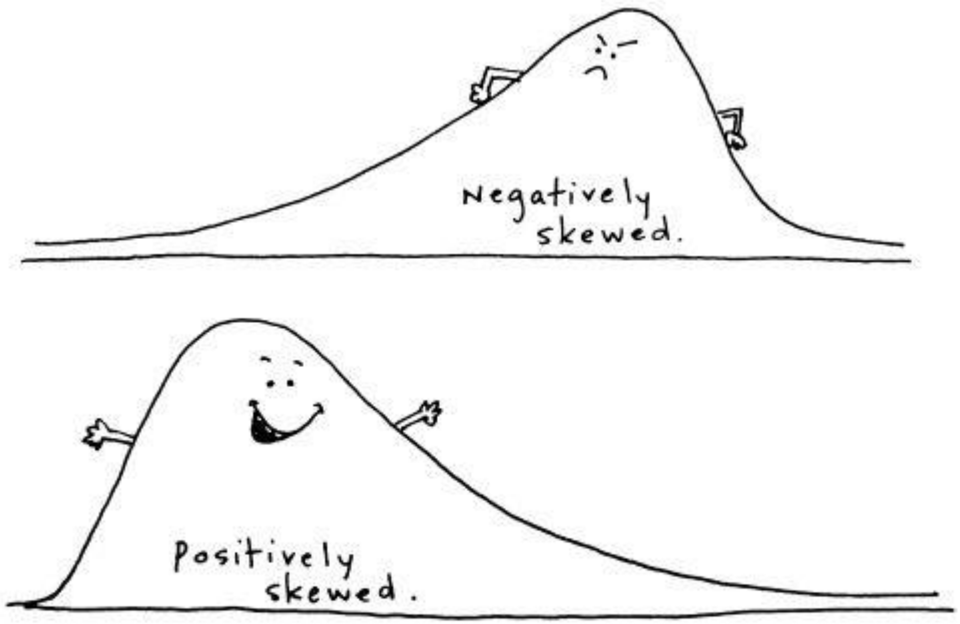


Univariate Analysis - Numerical			
Statistics	**Visualization**	**Equation**	**Description**
Count	Histogram	N	The number of values (observations) of the variable.
Minimum	Box Plot	Min	The smallest value of the variable.
Maximum	Box Plot	Max	The largest value of the variable.
Mean	Box Plot	$\bar{X} = \frac{\sum X}{N}$	The sum of the values divided by the count.
Median	Box Plot	$\tilde{X}$	The middle value. Below and above median lies an equal number of values.
Mode	Histogram		The most frequent value. There can be more than one mode.
Quantile	Box Plot	Q_k	A set of 'cut points' that divide a set of data into groups containing equal numbers of values (Quartile, Quintile, Percentile, ...).
Range	Box Plot	$Max\text{-}Min$	The difference between maximum and minimum.
Variance	Histogram	$S^2 = \frac{\sum(X-\bar{X})^2}{N-1}$	A measure of data dispersion.
Standard Deviation	Histogram	$S = \sqrt{S^2}$	The square root of variance.
Coefficient of Deviation	Histogram	$CV = \frac{S}{\bar{X}} \times 100\%$	A measure of data dispersion divided by mean.
Skewness	Histogram	$\frac{N}{(N-1)(N-2)}\sum\left(\frac{X-\bar{X}}{S}\right)^3$	A measure of symmetry or asymmetry in the distribution of data.
Kurtosis	Histogram	$\left[\frac{N(N+1)}{(N-1)(N-2)(N-3)}\sum\left(\frac{X-\bar{X}}{S}\right)^4\right] - \frac{3(N-1)^2}{(N-2)(N-3)}$	A measure of whether the data are peaked or flat relative to a normal distribution.

Skew

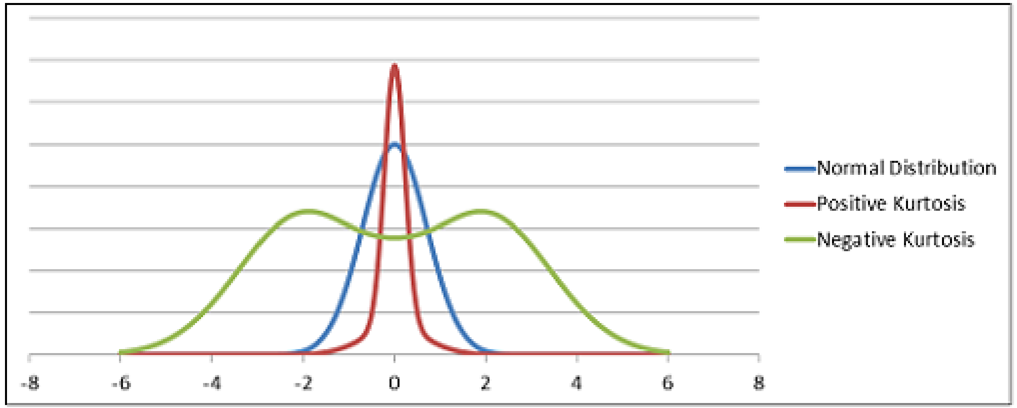


Negative Skew
Positive Skew



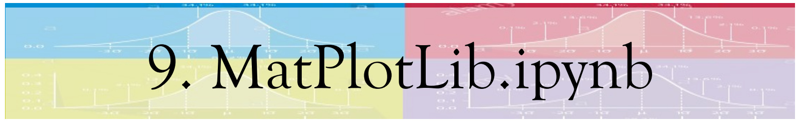
Negatively skewed.
Positively skewed.

Kurtosis



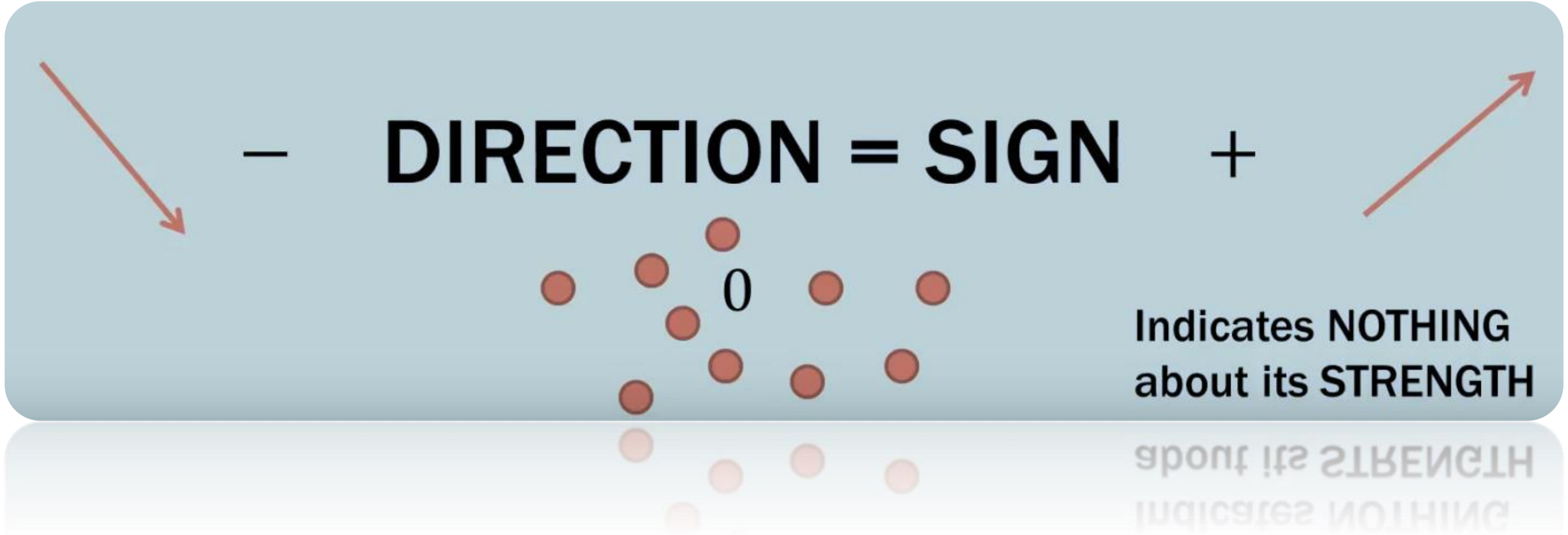
Normal Distribution
Positive Kurtosis
Negative Kurtosis
-8
-6
-4
-2
0
2
4
6
8

MatPlotLib - graphs



Covariance

- A descriptive measure of the **linear association** between 2 variables:
  - A positive value indicates a direct or **increasing linear** relationship
  - A negative value indicates a **decreasing** relationship



https://www.youtube.com/watch?v=xGbpuFNR1ME&index=7&list=PLAshlHpA2Iwc10-3HIioqUtqG0Fc4MNpp

Covariance formula

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Sample Covariance

$$\sigma_{xy} = \frac{\Sigma(x_i - \mu_x)(y_i - \mu_y)}{N}$$

Population Covariance

Workers

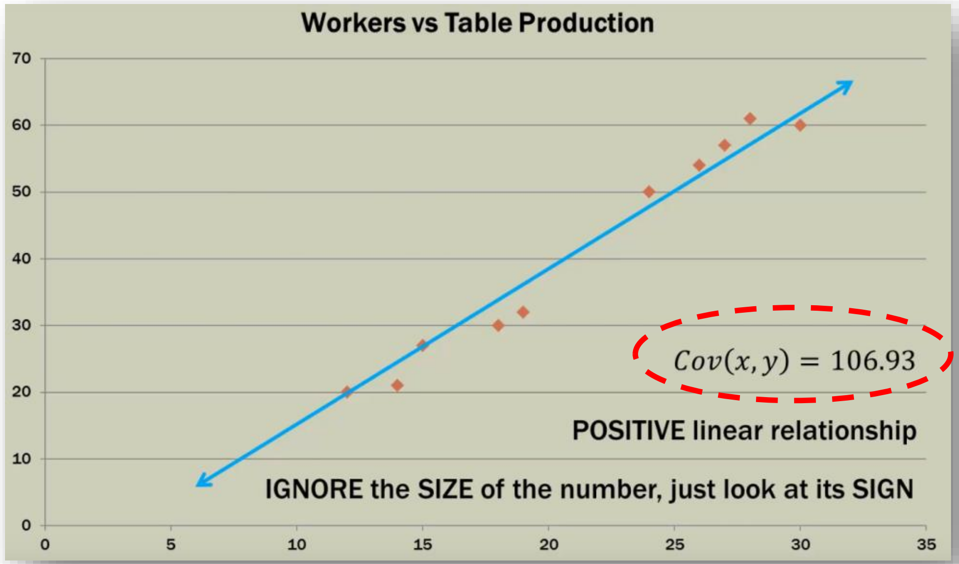
Table Productions

x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
12	20	-9.3	-21.2	197.16
30	60	8.7	18.8	163.56
15	27	-6.3	-14.2	89.46
24	50	2.7	8.8	23.76
14	21	-7.3	-20.2	147.46
18	30	-3.3	-11.2	36.96
28	61	6.7	19.8	132.66
26	54	4.7	12.8	60.16
19	32	-2.3	-9.2	21.16
27	57	5.7	15.8	90.06
$\bar{x} = 21.3$	$\bar{y} = 41.2$			$\Sigma = 962.4$

$$Cov(x,y) = s_{xy} = \frac{962.4}{n-1}$$

$$\frac{962.4}{9}$$

$$Cov(x,y) = 106.93$$

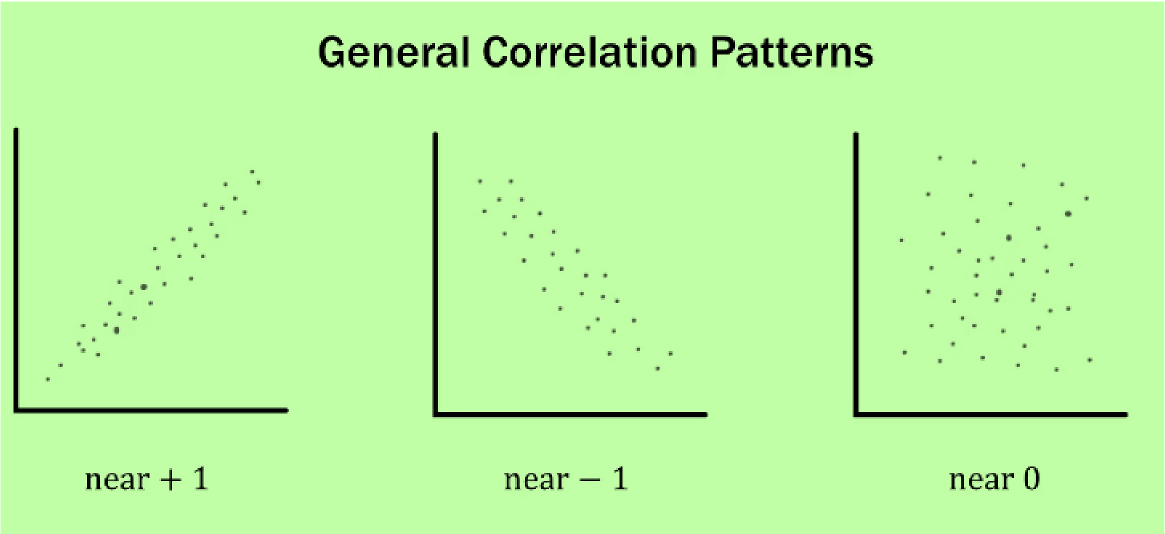


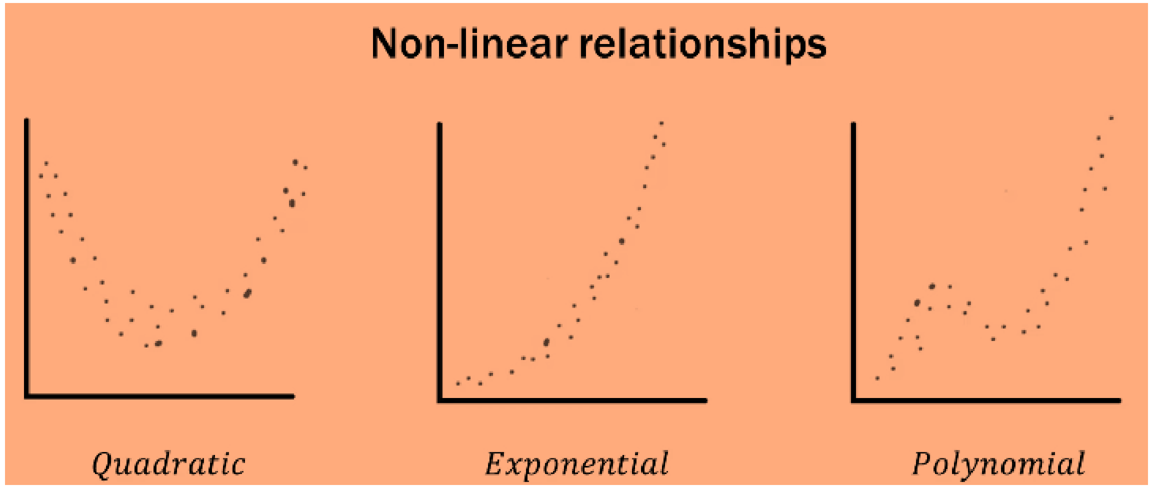
Correlation vs Covariance

- **Covariance** provides the **DIRECTION** (positive, negative, near zero) of the linear relationship between 2 variables
- **Correlation** provides both **DIRECTION** and **STRENGTH**
- Covariance result has no upper/lower bounds and its size is dependent on the scale of the variables
- Correlation is always between -**1** and **+1**

- **Correlation** is NOT **Causation**

Correlation





Correlation Formula

- r is called a (Pearson) correlation coefficient

$$r = \frac{Covariance\ (x, y)}{Standard\ Deviation(x) \times Standard\ Deviation\ (y)}$$

Can be written as:

$$r = \frac{Cov(x, y)}{s_x s_y}$$

Correlation Example

x	y	$(x_i - \bar{x})(y_i - \bar{y})$
12	20	197.16
30	60	163.56
15	27	89.46
24	50	23.76
14	21	147.46
18	30	36.96
28	61	132.66
26	54	60.16
19	32	21.16
27	57	90.06
$\bar{x} = 21.3$	$\bar{y} = 41.2$	$\Sigma = 962.4$

$$Cov(x,y) = s_{xy} = \frac{962.4}{n-1}$$

$$\frac{962.4}{9}$$

$$Cov(x,y) = 106.93$$

$s_x = 6.48$ $\quad s_y = 16.69$

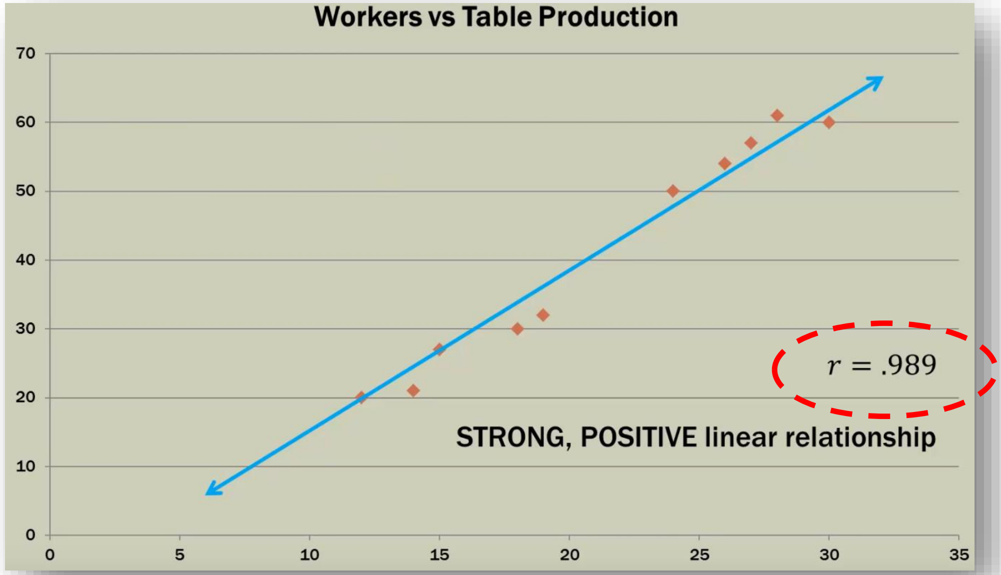
$$r = \frac{Cov(x,y)}{s_x s_y}$$

$$r = \frac{s_{xy}}{s_x s_y}$$

$$r = \frac{106.93}{6.48 \times 16.69}$$

$$r = \frac{106.93}{108.15}$$

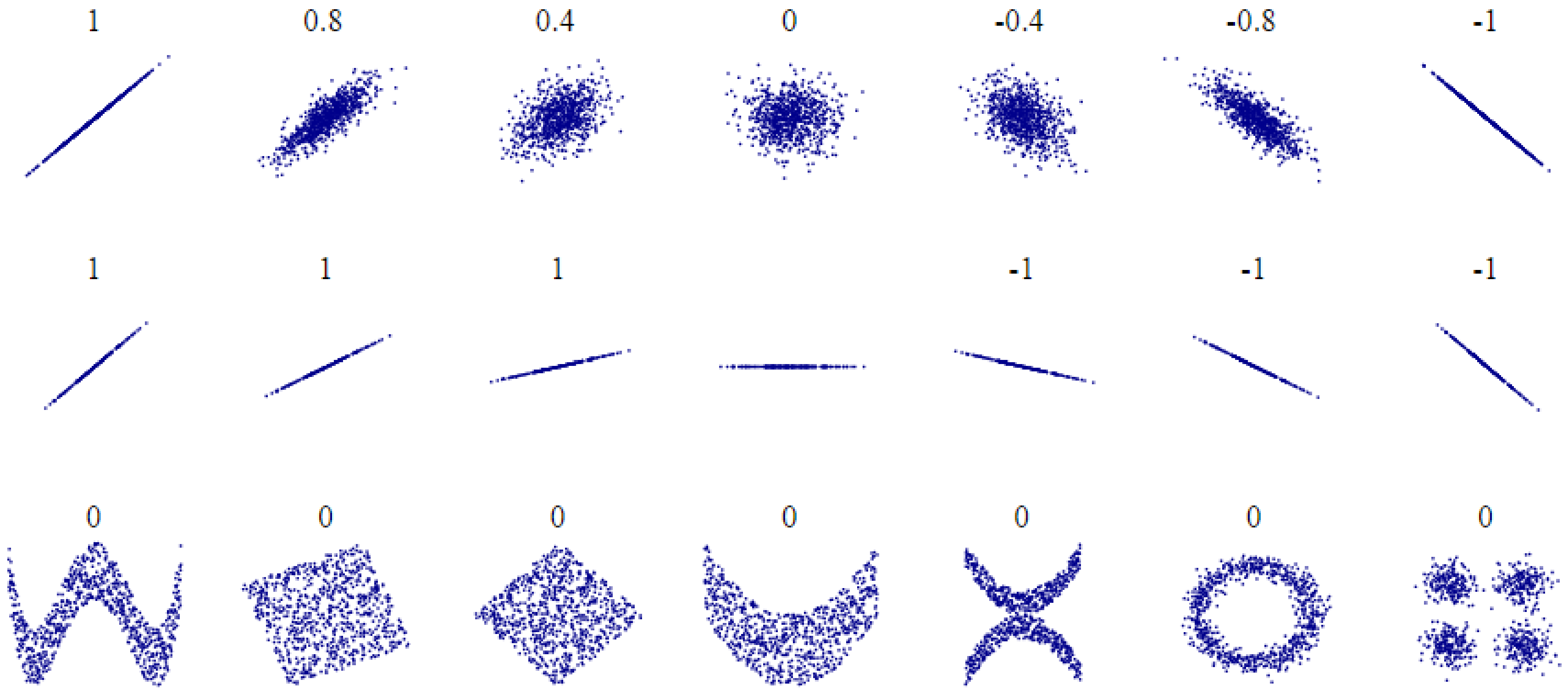
$$r = .989$$



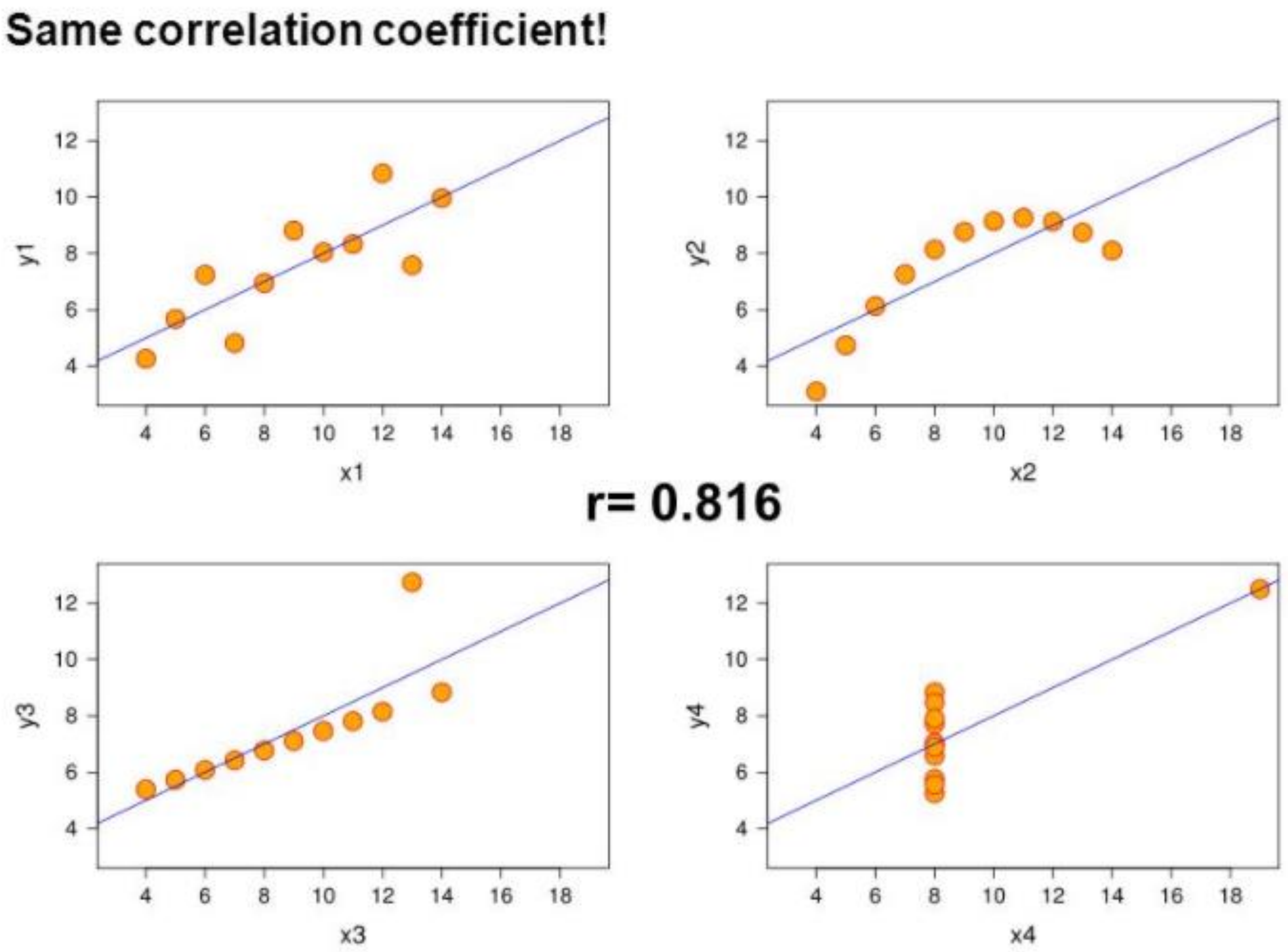
If $|r| \geq \dfrac{2}{\sqrt{n}}$, then a relationship exists

So for our problem: $|r| \geq \dfrac{2}{\sqrt{10}} = .632$ is the rule of thumb threshold

Correlation Examples



Correlation Coefficient can be tricky…



Correlation is NOT causation/causality!

- https://www.youtube.com/watch?v=Cl8zetzDBfM

END