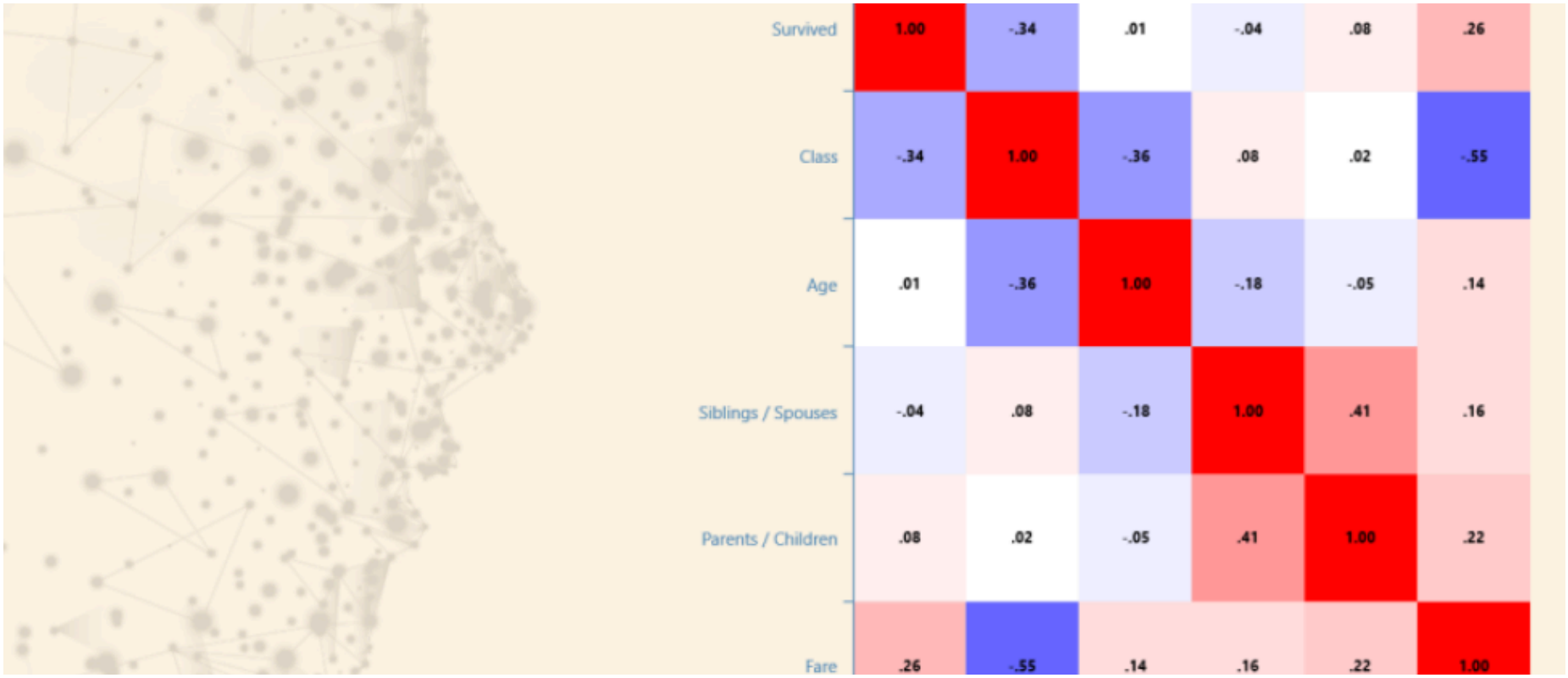


Correlation Analysis

Correlation Analysis

A STATISTICAL METHOD USED TO EVALUATE THE STRENGTH OF RELATIONSHIP BETWEEN TWO QUANTITATIVE VARIABLES.

A HIGH CORRELATION MEANS THAT TWO OR MORE VARIABLES HAVE A STRONG RELATIONSHIP WITH EACH OTHER

A WEAK CORRELATION MEANS THAT THE VARIABLES ARE HARDLY RELATED

IT IS CONNECTED TO THE LINEAR REGRESSION ANALYSIS THAT MODELS THE ASSOCIATION BETWEEN A DEPENDENT VARIABLE, CALLED RESPONSE, AND ONE OR MORE EXPLANATORY OR INDEPENDENT VARIABLES.

Types of Correlation

- **Pearson's correlation**: This is the most common correlation method. It corresponds to the covariance of the two variables normalized (i.e., divided) by the product of their standard deviations. Assumes a linear relationship between variables

- **Spearman's rank correlation**: A non-parametric measure of correlation, the Spearman correlation between two variables is equal to the Pearson correlation between the rank scores of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not). The relationship may non-linear

- **Kendall's rank correlation**: In the normal case, the Kendall correlation is preferred to the Spearman correlation because of a smaller gross error sensitivity (GES) and a smaller asymptotic variance (AV), making it more robust and more efficient.

However, the interpretation of Kendall's tau is less direct compared to that of the Spearman's rho, in the sense that it quantifies the difference between the % of concordant and discordant pairs among all possible pairwise events.

Pearson Correlation

IT IS DEFINED AS THE QUALITY OF LEAST SQUARES FITTING TO THE ORIGINAL DATA.

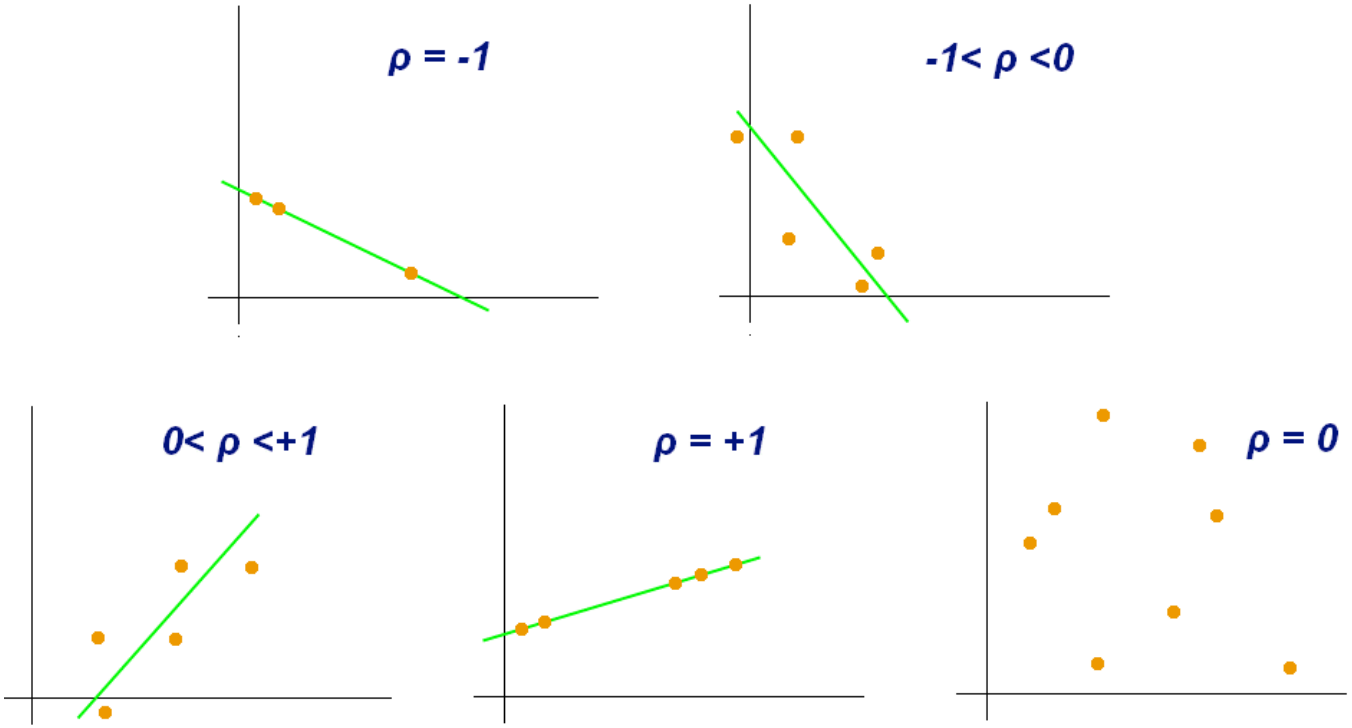
IT IS OBTAINED BY TAKING THE RATIO OF THE COVARIANCE OF THE TWO VARIABLES IN QUESTION OF OUR NUMERICAL DATASET, NORMALIZED TO THE SQUARE ROOT OF THEIR VARIANCES.

DIVIDES THE COVARIANCE OF THE TWO VARIABLES BY THE PRODUCT OF THEIR STANDARD DEVIATIONS.

https://machinelearningmastery.com/how-to-calculate-nonparametric-rank-correlation-in-python/

Pearson Correlation

- Pearson correlation coefficient (PCC) *r or* ρ is a measure of the linear correlation between two variables X and Y.
- It has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.



Pearson Correlation

$$\rho_{X,Y} = \operatorname{corr}(X,Y)$$

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

$$= \frac{\operatorname{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

In terms of statistical moments:

$$\rho_{X,Y} = \frac{\operatorname{E}(XY) - \operatorname{E}(X)\operatorname{E}(Y)}{\sqrt{\operatorname{E}(X^2) - \operatorname{E}(X)^2} \cdot \sqrt{\operatorname{E}(Y^2) - \operatorname{E}(Y)^2}}$$

Pearson Correlation

Sample Correlation Coefficient:

Given a series of n measurements of the pair (X_i, Y_i) indexed by $i = 1, \ldots, n$, the *sample correlation coefficient* can be used to estimate the population Pearson correlation $\rho_{X,Y}$ between X and Y. The sample correlation coefficient is defined as

$$r_{xy} \stackrel{\text{def}}{=} \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

where $\bar{x}$ and $\bar{y}$ are the sample means of X and Y, and s_x and s_y are the corrected sample standard deviations of X and Y.

Assumptions in Pearson Correlation

- For the Pearson *r* correlation, both variables should be normally distributed (normally distributed variables have a bell-shaped curve).
- Other assumptions include linearity and homoscedasticity.
- Linearity assumes a straight-line relationship between each of the two variables
- homoscedasticity assumes that data is equally distributed about the regression line.

Spearman Correlation

- It is a nonparametric measure of rank correlation (statistical dependence between the rankings of two variables)
- It assesses how well the relationship between two variables can be described using a **monotonic** function
- The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the rank variables

For a sample of size *n*, the *n* raw scores X_i, Y_i are converted to ranks $\operatorname{rg} X_i, \operatorname{rg} Y_i$, and r_s is computed as

$$r_s = \rho_{\operatorname{rg}_X, \operatorname{rg}_Y} = \frac{\operatorname{cov}(\operatorname{rg}_X, \operatorname{rg}_Y)}{\sigma_{\operatorname{rg}_X} \sigma_{\operatorname{rg}_Y}},$$

where

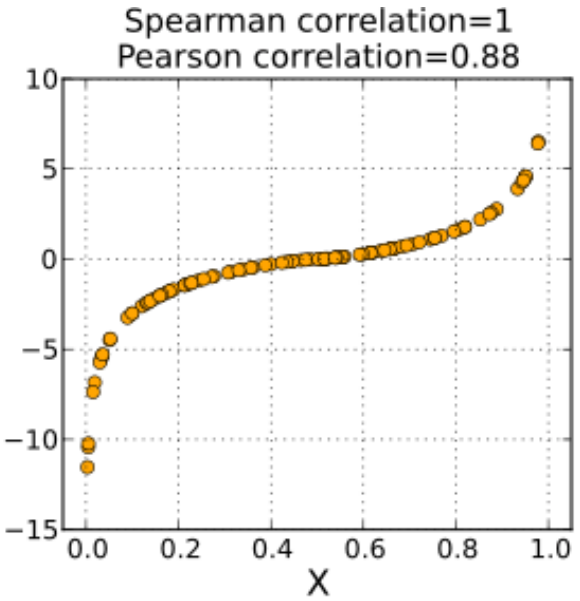
ρ denotes the usual Pearson correlation coefficient, but applied to the rank variables,

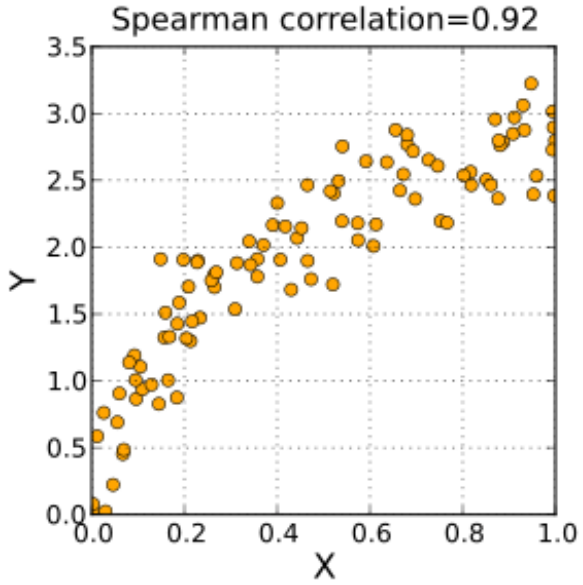
$\operatorname{cov}(\operatorname{rg}_X, \operatorname{rg}_Y)$ is the covariance of the rank variables,

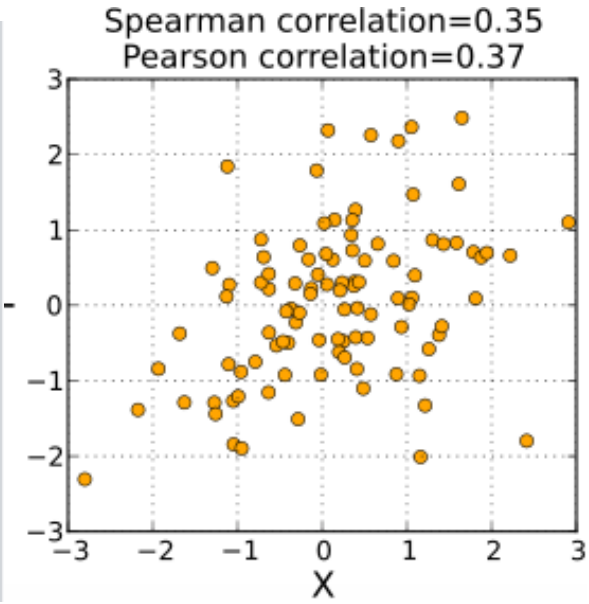
$\sigma_{\operatorname{rg}_X}$ and $\sigma_{\operatorname{rg}_Y}$ are the standard deviations of the rank variables.

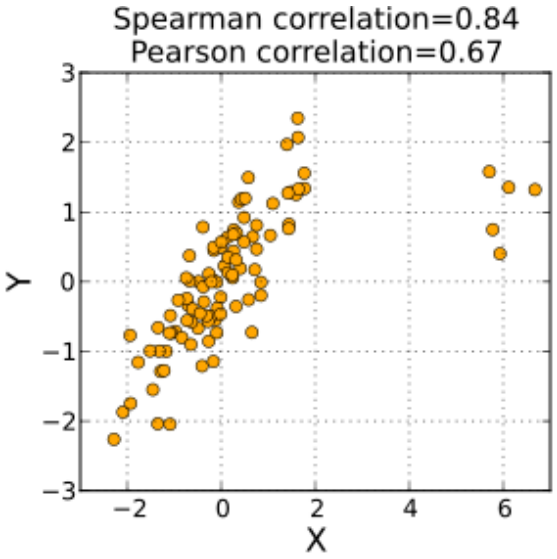
https://machinelearningmastery.com/how-to-calculate-nonparametric-rank-correlation-in-python/

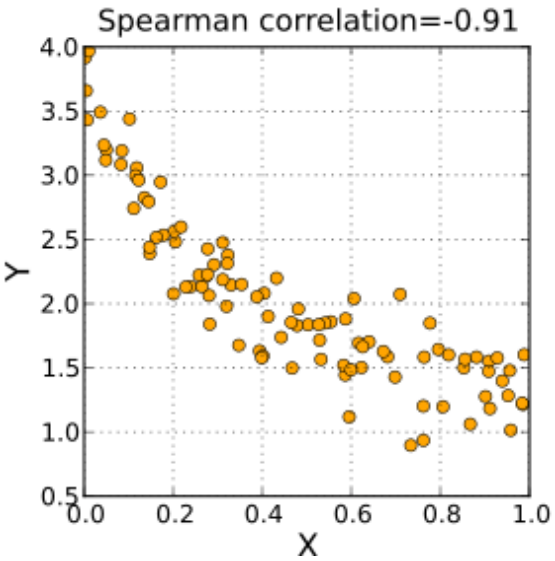
Spearman Correlation











Spearman Correlation

IQ, X_i	Hours of TV per week, Y_i
106	7
86	0
100	27
101	50
99	28
103	29
97	20
113	12
112	6
110	17

IQ, X_i	Hours of TV per week, Y_i	rank x_i	rank y_i	d_i	d_i^2
86	0	1	1	0	0
97	20	2	6	−4	16
99	28	3	8	−5	25
100	27	4	7	−3	9
101	50	5	10	−5	25
103	29	6	9	−3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

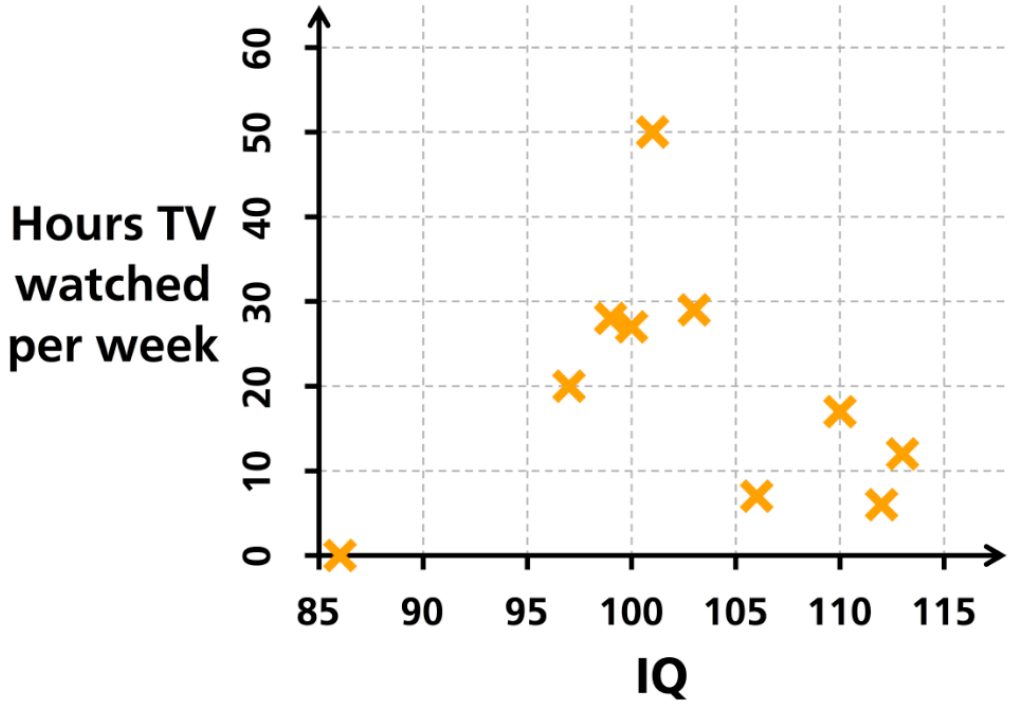
Firstly, evaluate d_i^2. To do so use the following steps, reflected in the table below.

1. Sort the data by the first column (X_i). Create a new column x_i and assign it the ranked values 1, 2, 3, ..., *n*.
2. Next, sort the data by the second column (Y_i). Create a fourth column y_i and similarly assign it the ranked values 1, 2, 3, ..., *n*.
3. Create a fifth column d_i to hold the differences between the two rank columns (x_i and y_i).
4. Create one final column d_i^2 to hold the value of column d_i squared.

$\rho = -29/165 = -0.175757575$

Spearman Correlation

- That the value is close to zero shows that the correlation between IQ and hours spent watching TV is very low
- The negative value suggests that the longer the time spent watching television the lower the IQ.



Kendall Correlation

- The coefficient is often referred to by the lowercase Greek letter tau (t).
- The test may be called Kendall's tau.
- It calculates a normalized score for the number of matching or concordant rankings between the two samples. As such, the test is also referred to as Kendall's concordance test.
- The test takes the two data samples as arguments and returns the correlation coefficient and the p-value.
- As a statistical hypothesis test, the method assumes (H0) that there is no association between the two samples.

https://machinelearningmastery.com/how-to-calculate-nonparametric-rank-correlation-in-python/

Correlation in Python

- Download Python script correlation1.py

- https://machinelearningmastery.com/how-to-use-correlation-to-understand-the-relationship-between-variables/

Correlation in Python

- Download Python script correlation2.py

- https://towardsdatascience.com/better-heatmaps-and-correlation-matrix-plots-in-python-41445d0f2bec

17 Normality Tests with Python

Download Python script
normality-tests.py

https://machinelearningmastery.com/statistical-hypothesis-tests-in-python-cheat-sheet/

Other correlation exercises in Python with House Prices

- https://www.geeksforgeeks.org/exploring-correlation-in-python/

More Correlation exercises in Python

https://www.reneshbedre.com/blog/correlation-analysis.html

Uses: from bioinfokit import analys, visuz

Correlation and Independence

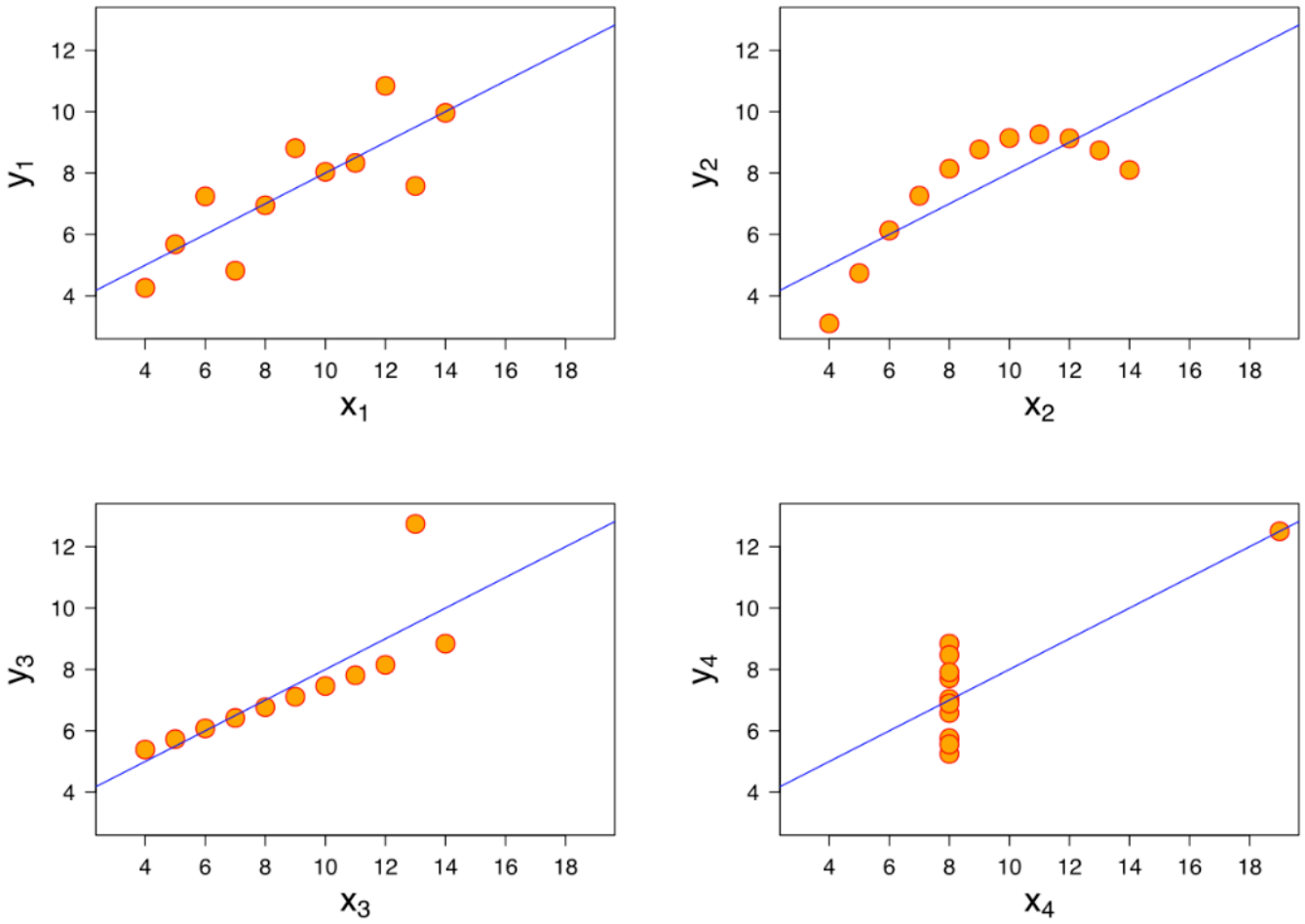
- If the variables are independent, Pearson's correlation coefficient is 0, but the converse is not true because the correlation coefficient detects only linear dependencies between two variables

$$X, Y \text{ independent} \quad \Rightarrow \quad \rho_{X,Y} = 0 \quad (X, Y \text{ uncorrelated})$$
$$\rho_{X,Y} = 0 \quad (X, Y \text{ uncorrelated}) \quad \not\Rightarrow \quad X, Y \text{ independent}$$

Correlation and Causation

- Correlation does not imply causation
- If X and Y are correlated there may be a third unobserved variable Z that causes X and Y
- If X causes Y then X and Y may be correlated

Caution with Correlation



Hands-on Exercise in R

- http://www.sthda.com/english/wiki/correlation-analyses-in-r
- https://statsandr.com/blog/correlation-coefficient-and-correlation-test-in-r/