COESIONE ITALIA 21-27 EMILIA-ROMAGNA

Cofinanziato dall'Unione europea

REPUBBLICA ITALIANA

Regione Emilia-Romagna

DATA LAB

GUARDA AVANTI

Big Data, nuove competenze
per nuove professioni.

FORM INDUSTRIA EMILIA ROMAGNA

ifoa Sapere utile

REISS ROMOLI la passione della conoscenza

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

Università degli Studi di Ferrara

UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

UNIVERSITÀ DI PARMA

POLITECNICO MILANO 1863 POLO TERRITORIALE DI PIACENZA

UNIVERSITÀ CATTOLICA del Sacro Cuore

"Anticipare la crescita con le nuove competenze sui Big Data" Operazione Rif. PA 2023-19167/RER approvata con DGR n° 843 del 29 maggio 2023 e co-finanziata dal Fondo Sociale Europeo Plus 2021-2027 Regione Emilia-Romagna

ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

Operazione Rif. PA 2023-19167/RER/10/1, "ANTICIPARE LA CRESCITA CON LE NUOVE COMPETENZE SUI BIG DATA", approvata dalla Regione Emilia-Romagna con DGR n° 843 del 29/05/2023 e co-finanziata dal Fondo Sociale Europeo Plus 2021-2027

DATA LAB

CORRELAZIONE

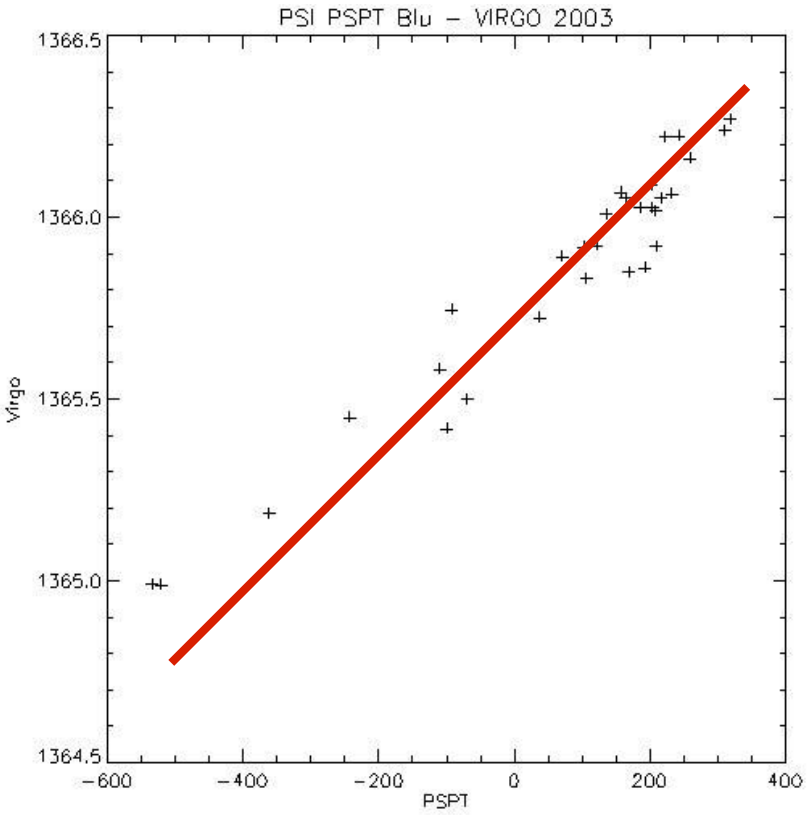
LA CORRELAZIONE LINEARE

La correlazione indica la tendenza che hanno due variabili (X e Y) a *variare insieme*, ovvero, a *covariare*. Ad esempio, si può supporre che vi sia una relazione tra il nostro salario ed i soldi spesi in viaggi, nel senso che all'aumentare dell'uno aumenta anche l'altro.

Quando si parla di correlazione bisogna prendere in considerazione due aspetti: *il tipo di relazione esistente* tra due variabili e *la forma della relazione*.

Per quanto riguarda il **tipo di relazione**, essa può essere *lineare* o *non lineare*

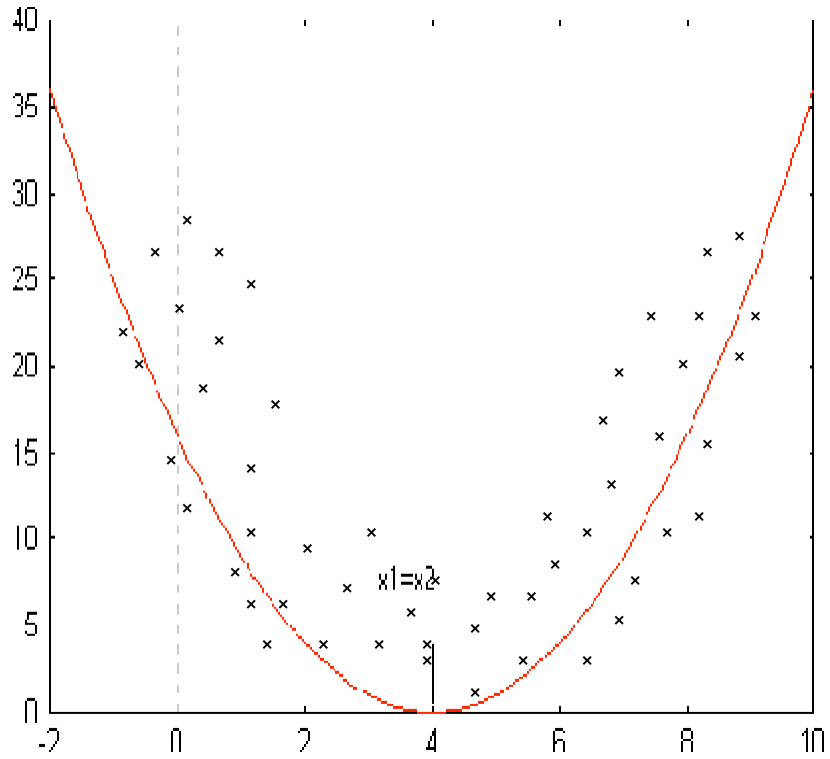
- La relazione è di tipo *lineare* se, rappresentata su assi cartesiane, si avvicina alla forma di una retta.



In questo all'aumentare (o al diminuire) di X aumenta (o diminuisce) Y

Ad esempio, all'aumentare dell'altezza di una persona aumenta anche il suo peso.

- La relazione è di tipo *non lineare,* se rappresentata su assi cartesiane, ha un andamento curvilineo (parabola o iperbole).

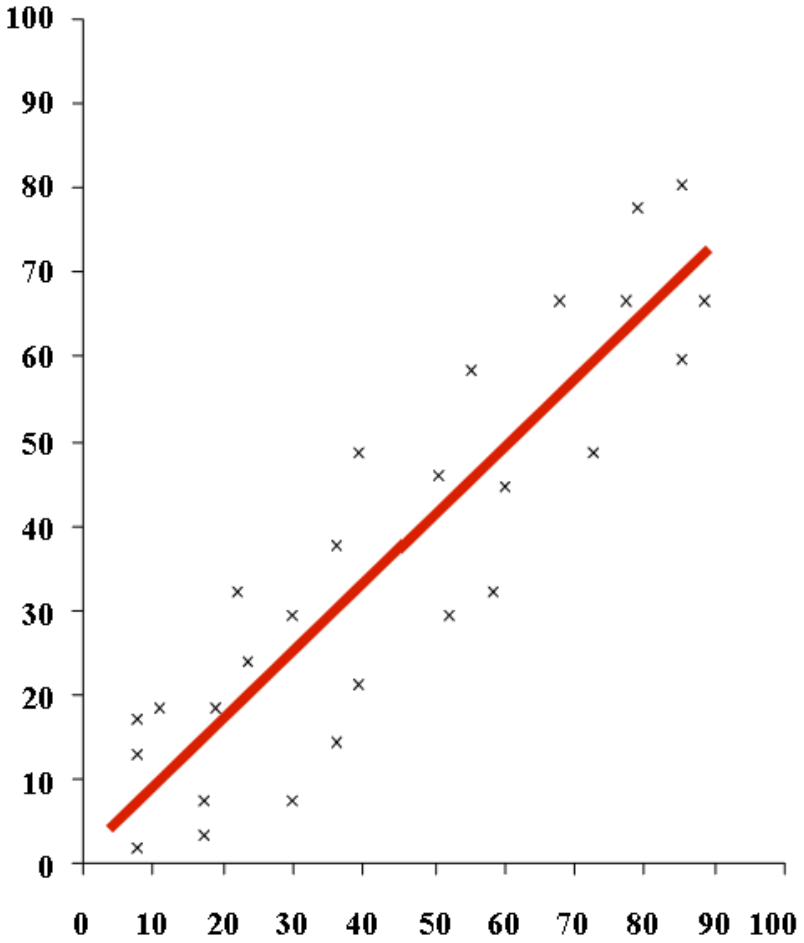


In questo caso a livelli bassi e alti di X corrispondono livelli bassi di Y; mentre a livelli intermedi di X corrispondono livelli alti di Y.

Ad esempio, il tempo impiegato per risolvere un problema è alto quando l'ansia è bassa o alta, è minimo quando l'ansia ha livelli medi

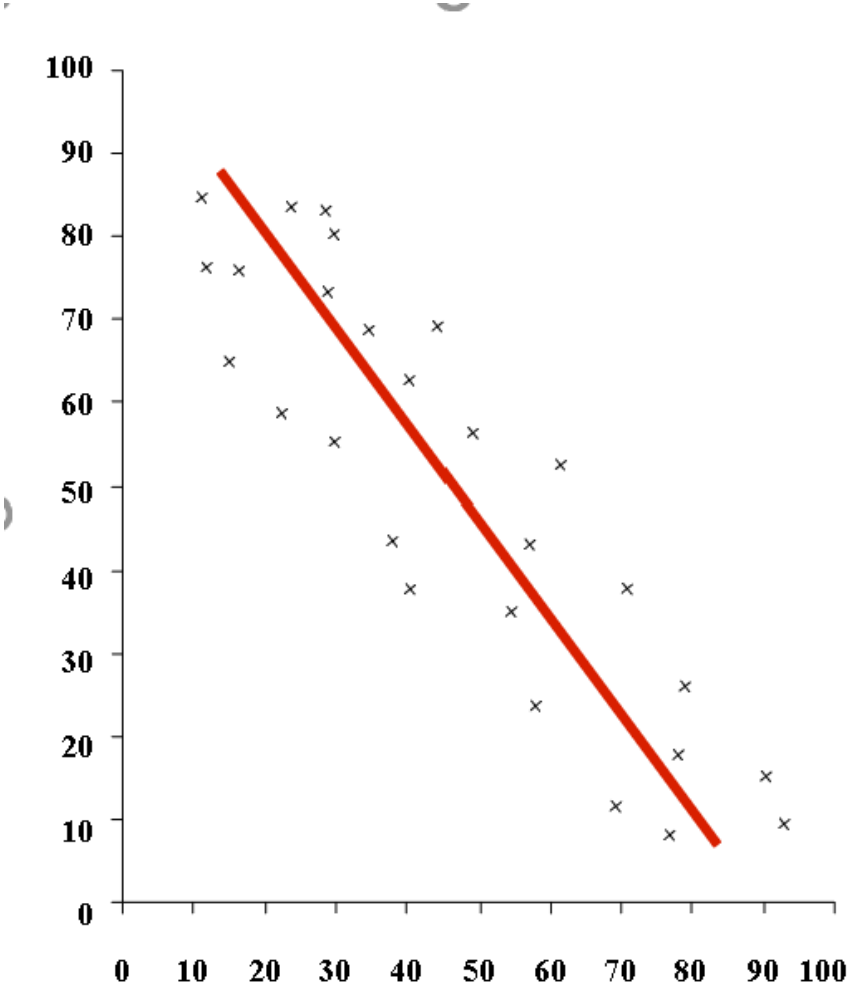
Per quanto riguarda la forma della relazione, si distinguono l'entità e la direzione.

La **direzione** può essere: *positiva*, se all'aumentare di una variabile aumenta anche l'altra.



Ad esempio, all'aumentare dei matrimoni aumentano il numero di nascite

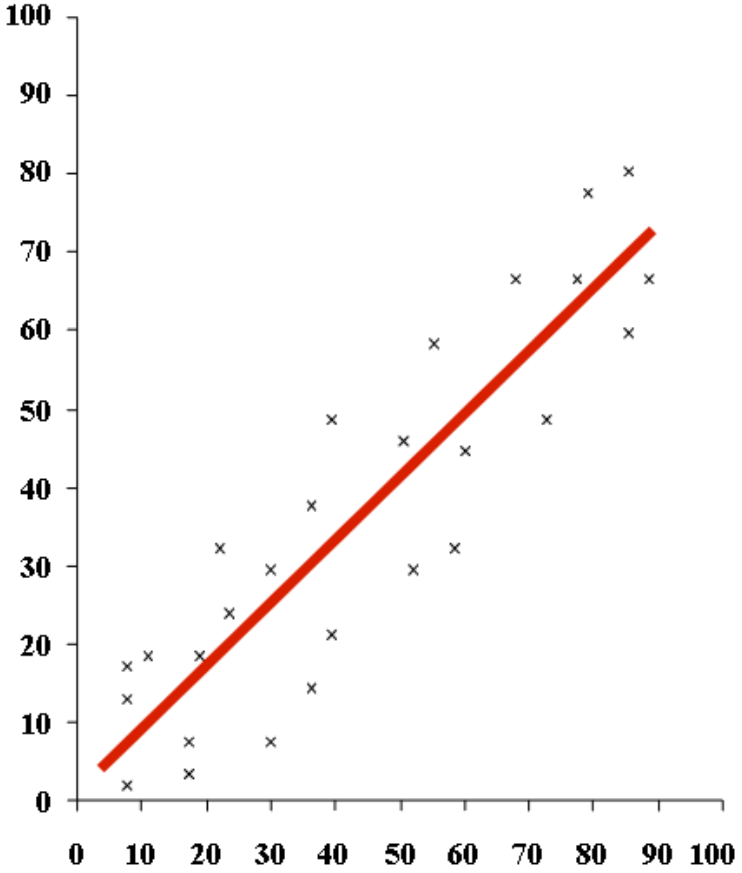
La direzione è negativa se all’aumentare di una variabile diminuisce l’atra.



Ad esempio all’aumentare dei divorzi diminuiscono il numero di nascite

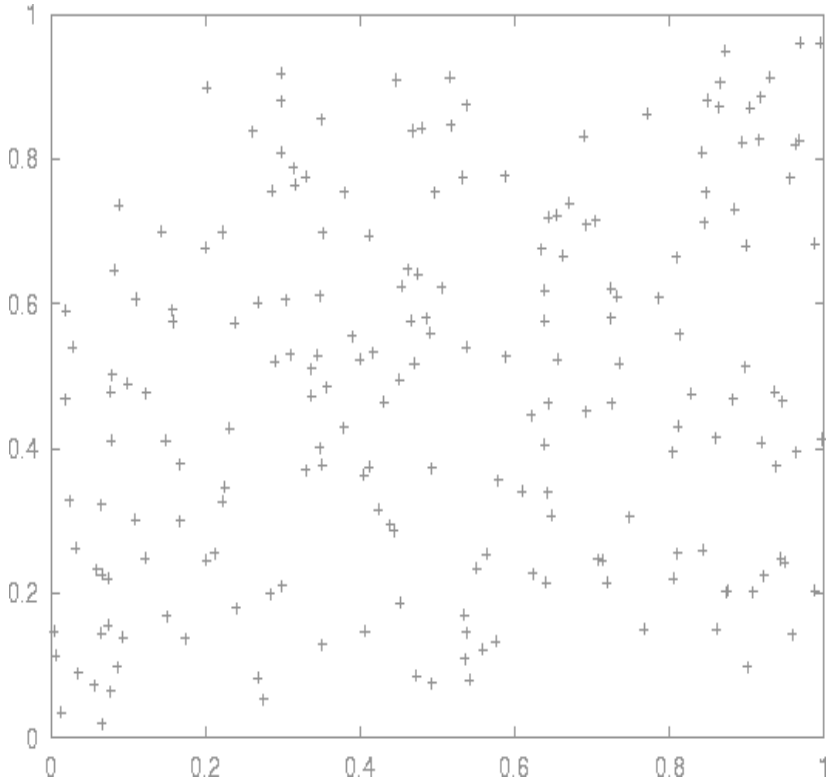
L’entità si riferisce alla forza della relazione esistente tra due variabili.

Quanto più i punteggi sono raggruppati attorno ad una retta, tanto *più forte* è la relazione tra due variabili.



Ad esempio, quanto più elevata è la temperatura, tanto più si suda.

Se i punteggi sono dispersi in maniera uniforme, invece, tra le due variabili non esiste alcuna relazione.



Ad esempio, non esiste alcuna relazione tra la temperatura e la nostra età

Per esprimere la relazione esistente tra due variabili, si utilizza il **coefficiente di correlazione**

Tale coefficiente è standardizzato e può assumere valori che vanno da **–1.00** (correlazione perfetta negativa) e **+1.00** (correlazione perfetta positiva). Una correlazione uguale a **0** indica che tra le due variabili non vi è alcuna relazione.

Nota. La correlazione non include il concetto di causa-effetto, ma solo quello di rapporto tra variabili. La correlazione ci permette di affermare che tra due variabili c'è una *relazione sistematica*, ma non che una causa l'altra.

Esistono vari tipi di coefficienti di correlazione a seconda del tipo di scala della variabile.

- Per le scale a **intervalli** o **rapporti** equivalenti si usa il coefficiente ***r* di Pearson**.

- Per le scale **ordinali** si usano il coefficiente r_s **di Spearman** o il coefficiente ***tau* di Kendall**.

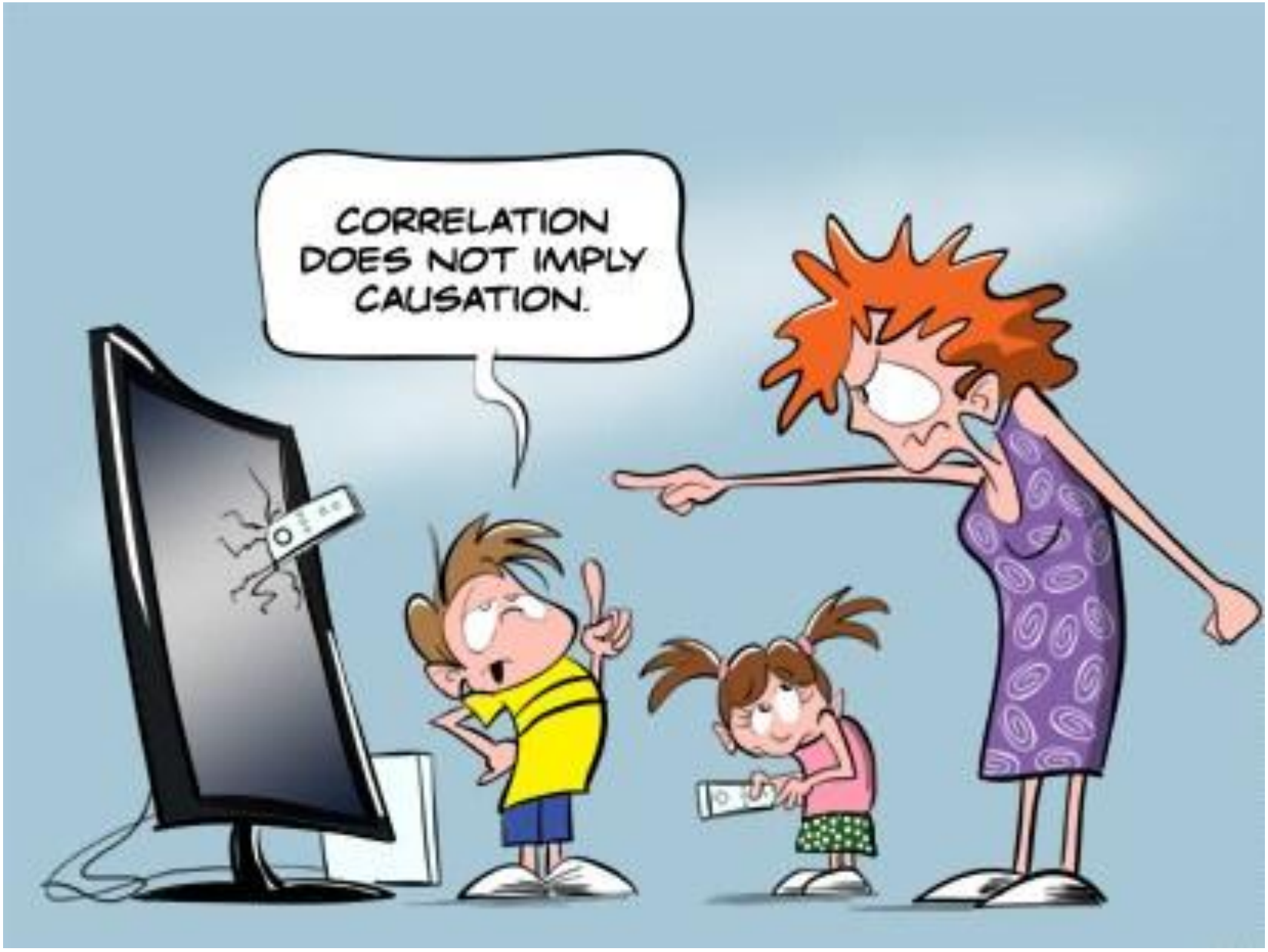
- Per le scale a **intervalli** o **rapporti** equivalenti si usa il coefficiente ***r* di Pearson**.

Tale coefficiente è calcolato come rapporto tra la covarianza delle due variabili e il prodotto delle loro deviazioni standard

$$-1 \leq \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_x)^2}\sqrt{\sum_{i=1}^{n}(y_i - \mu_y)^2}} \leq +1$$

Può assumere valori che vanno da **–1.00** (tra le due variabili vi è una correlazione perfetta negativa) e **+1.00** (tra le due variabili vi è una correlazione perfetta positiva). Una correlazione uguale a 0 indica che tra le due variabili non vi è alcuna relazione.

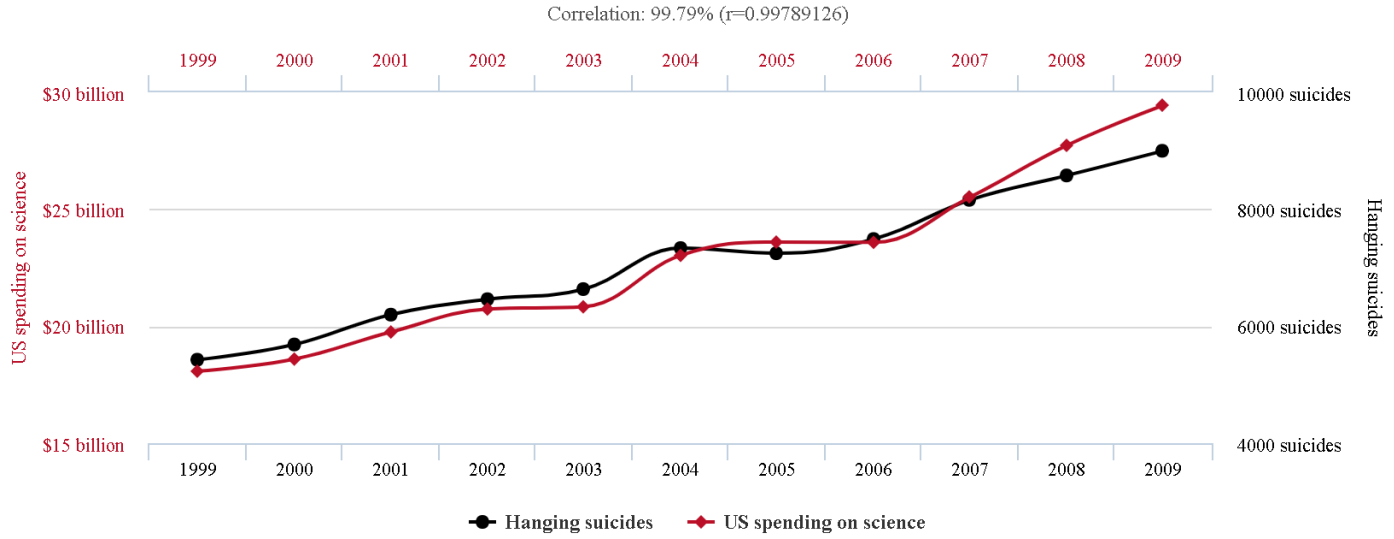
	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	Price
CRIM	1.000000	-0.200469	0.406583	-0.055892	0.420972	-0.219247	0.352734	-0.379670	0.625505	0.582764	0.289946	-0.385064	0.455621	-0.388305
ZN	-0.200469	1.000000	-0.533828	-0.042697	-0.516604	0.311991	-0.569537	0.664408	-0.311948	-0.314563	-0.391679	0.175520	-0.412995	0.360445
INDUS	0.406583	-0.533828	1.000000	0.062938	0.763651	-0.391676	0.644779	-0.708027	0.595129	0.720760	0.383248	-0.356977	0.603800	-0.483725
CHAS	-0.055892	-0.042697	0.062938	1.000000	0.091203	0.091251	0.086518	-0.099176	-0.007368	-0.035587	-0.121515	0.048788	-0.053929	0.175260
NOX	0.420972	-0.516604	0.763651	0.091203	1.000000	-0.302188	0.731470	-0.769230	0.611441	0.668023	0.188933	-0.380051	0.590879	-0.427321
RM	-0.219247	0.311991	-0.391676	0.091251	-0.302188	1.000000	-0.240265	0.205246	-0.209847	-0.292048	-0.355501	0.128069	-0.613808	0.695360
AGE	0.352734	-0.569537	0.644779	0.086518	0.731470	-0.240265	1.000000	-0.747881	0.456022	0.506456	0.261515	-0.273534	0.602339	-0.376955
DIS	-0.379670	0.664408	-0.708027	-0.099176	-0.769230	0.205246	-0.747881	1.000000	-0.494588	-0.534432	-0.232471	0.291512	-0.496996	0.249929
RAD	0.625505	-0.311948	0.595129	-0.007368	0.611441	-0.209847	0.456022	-0.494588	1.000000	0.910228	0.464741	-0.444413	0.488676	-0.381626
TAX	0.582764	-0.314563	0.720760	-0.035587	0.668023	-0.292048	0.506456	-0.534432	0.910228	1.000000	0.460853	-0.441808	0.543993	-0.468536
PTRATIO	0.289946	-0.391679	0.383248	-0.121515	0.188933	-0.355501	0.261515	-0.232471	0.464741	0.460853	1.000000	-0.177383	0.374044	-0.507787
B	-0.385064	0.175520	-0.356977	0.048788	-0.380051	0.128069	-0.273534	0.291512	-0.444413	-0.441808	-0.177383	1.000000	-0.366087	0.333461
LSTAT	0.455621	-0.412995	0.603800	-0.053929	0.590879	-0.613808	0.602339	-0.496996	0.488676	0.543993	0.374044	-0.366087	1.000000	-0.737663
Price	-0.388305	0.360445	-0.483725	0.175260	-0.427321	0.695360	-0.376955	0.249929	-0.381626	-0.468536	-0.507787	0.333461	-0.737663	1.000000


CORRELATION
DOES NOT IMPLY
CAUSATION.

US spending on science, space, and technology

correlates with

Suicides by hanging, strangulation and suffocation

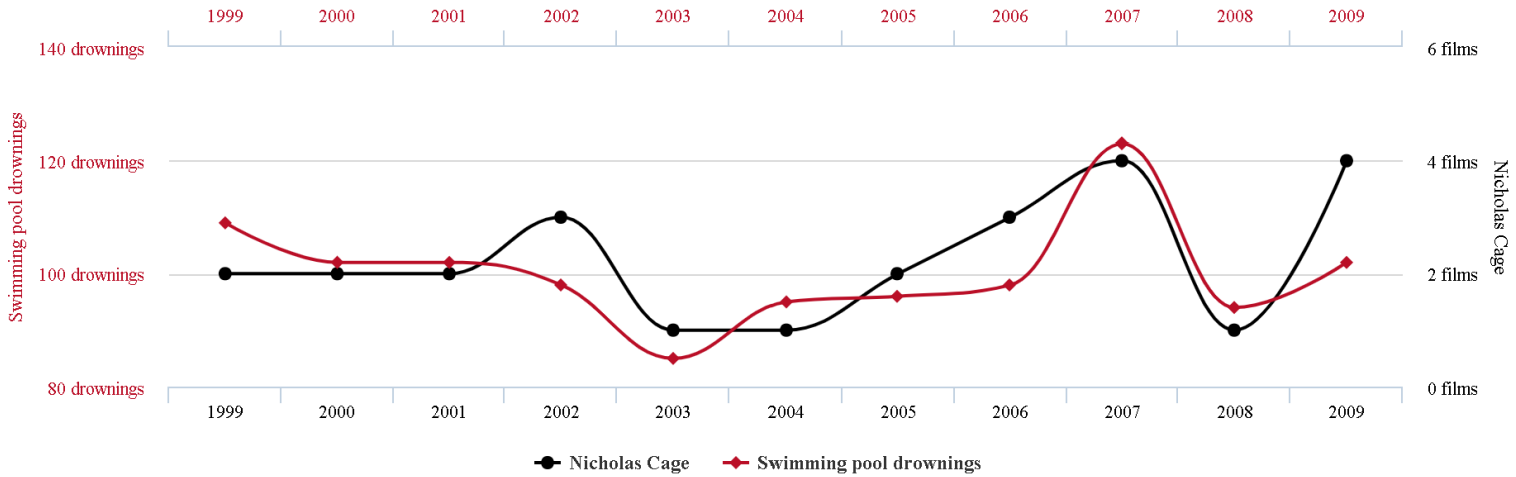


tylervigen.com

Number of people who drowned by falling into a pool

correlates with

Films Nicolas Cage appeared in

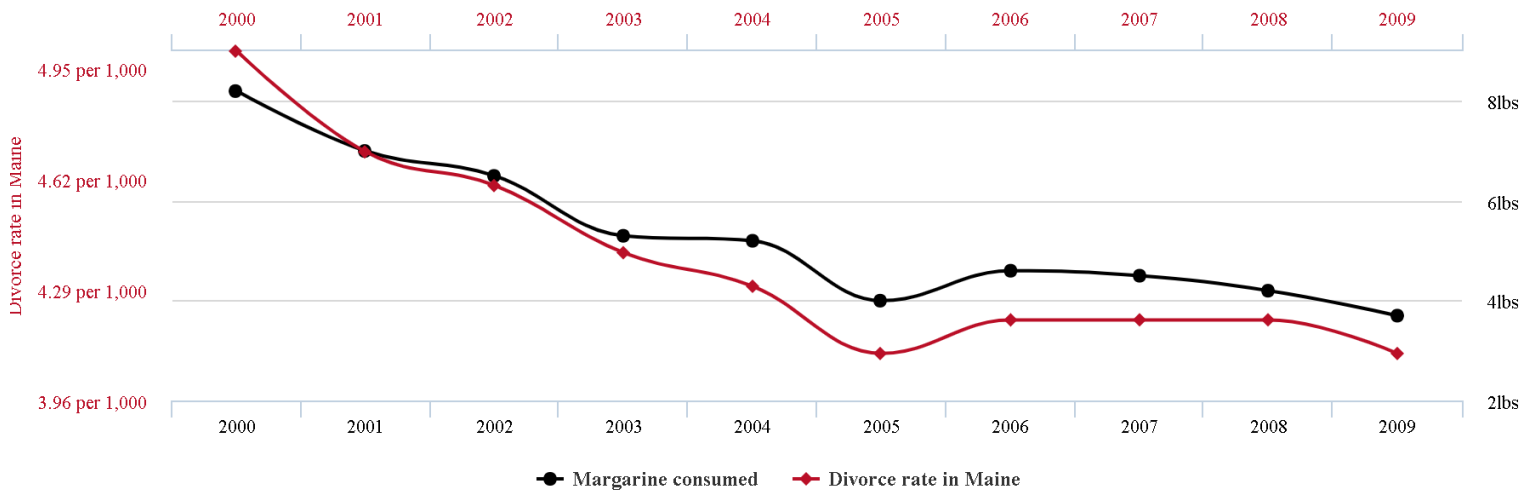


Divorce rate in Maine

correlates with

Per capita consumption of margarine

Correlation: 99.26% (r=0.992558)

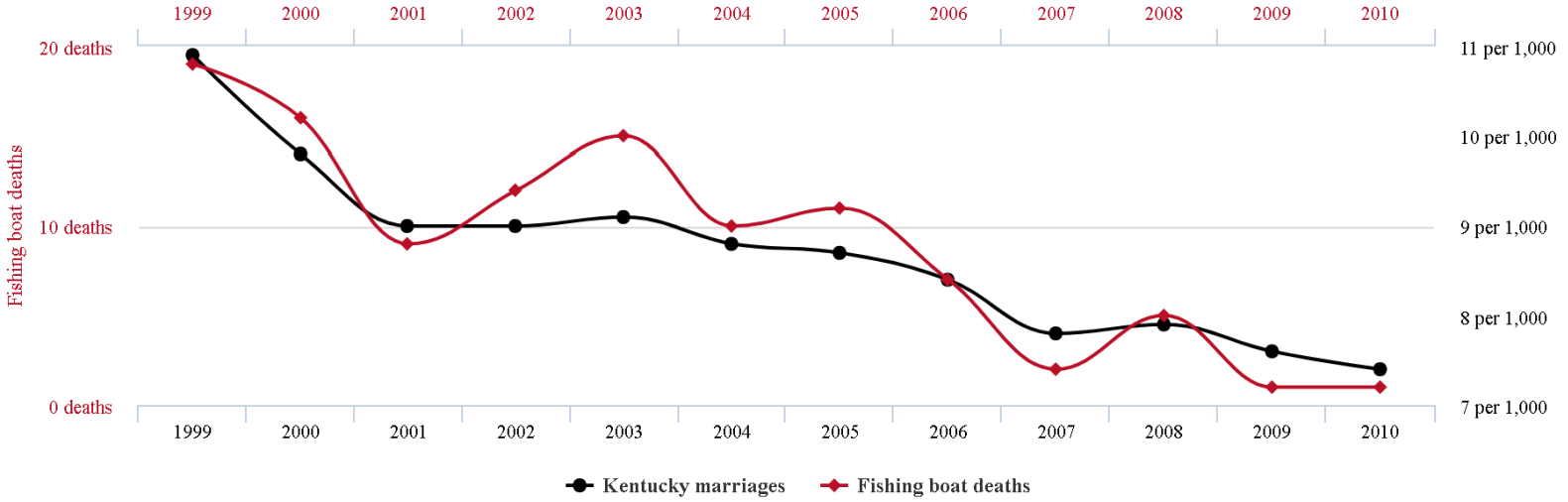


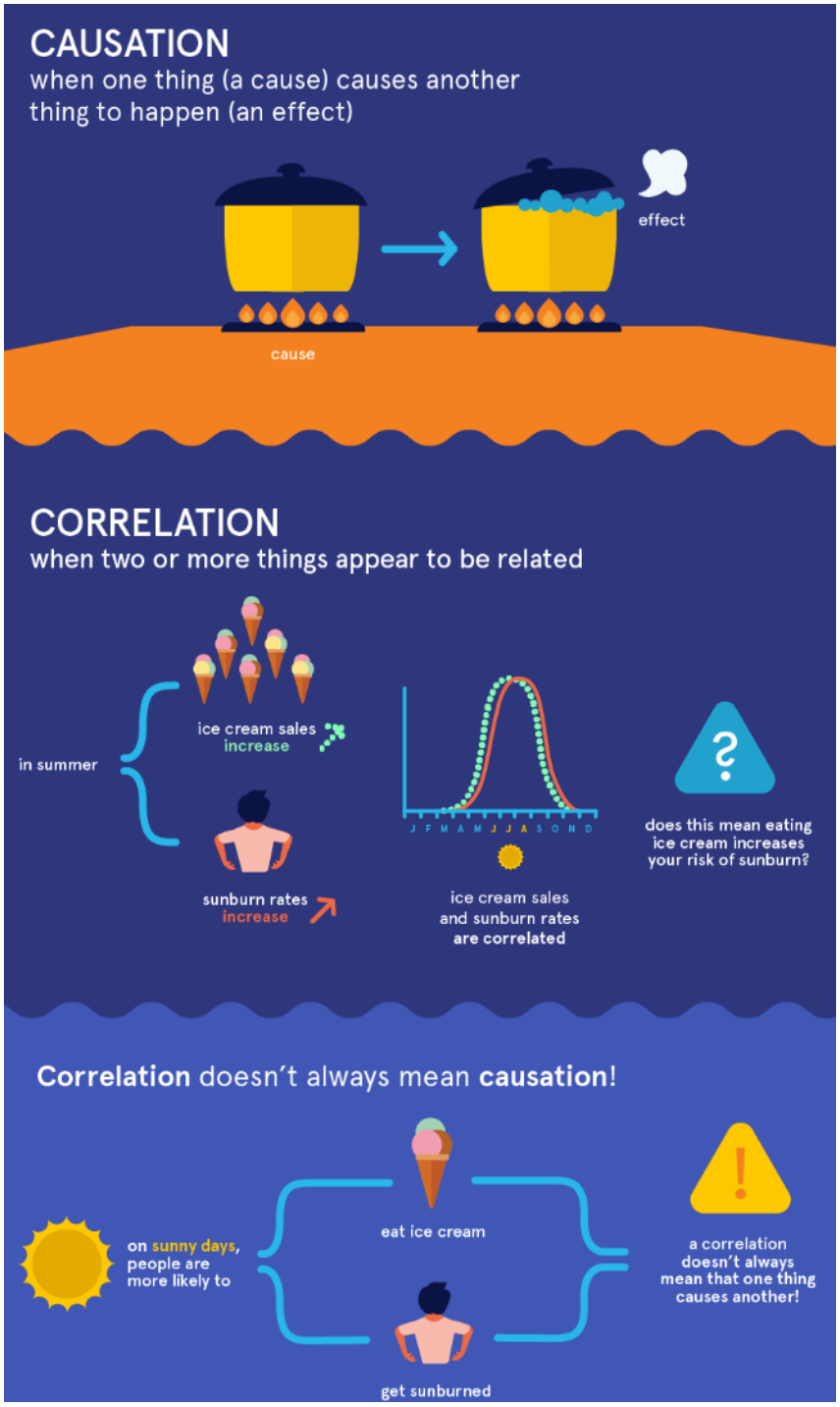
People who drowned after falling out of a fishing boat

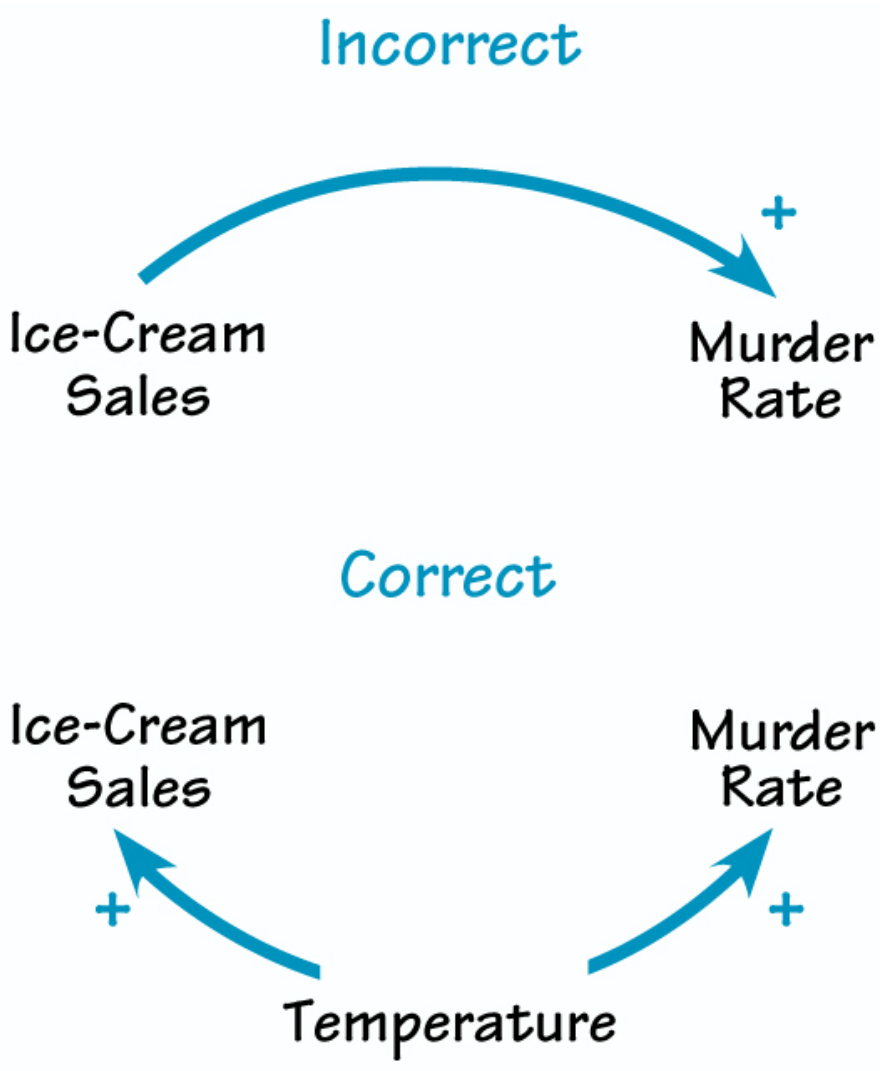
correlates with

Marriage rate in Kentucky

Correlation: 95.24% (r=0.952407)




CAUSATION
when one thing (a cause) causes another thing to happen (an effect)
effect
cause
CORRELATION
when two or more things appear to be related
in summer
ice cream sales increase
sunburn rates increase
J F M A M J J A S O N D
ice cream sales and sunburn rates are correlated
does this mean eating ice cream increases your risk of sunburn?
Correlation doesn't always mean causation!
on sunny days, people are more likely to
eat ice cream
get sunburned
a correlation doesn't always mean that one thing causes another!


Incorrect
Ice-Cream Sales
+
Murder Rate
Correct
Ice-Cream Sales
Murder Rate
+
+
Temperature