UNIVERSITATEA DE MEDICINĂ, FARMACIE, ȘTIINȚE ȘI TEHNOLOGIE „GEORGE EMIL PALADE” DIN TÂRGU MUREȘ

PROBABILITĂȚI ȘI STATISTICĂ ÎN SISTEME MEDICALE

Cursul 7, 14-15 octombrie 2020

Analiza regresiei și corelației pentru date din sisteme medicale

PARTEA I - ANALIZA CORELATIEI

prof. univ. dr. habil Manuela Rozalia GABOR

PARTEA I
ANALIZA CORELATIEI

1. Generalități
2. Coeficient de corelație
3. Corelație parametrică
4. Corelație neparametrică
5. Reprezentarea grafică a corelației
6. Exemplu numeric
7. Coeficient de determinare
8. “Funny correlations”

1.GENERALITĂȚI

1. CORELAȚIE ȘI REGRESIE – ASPECTE GENERALE

Termenul CORELAȚIE

Este folosit pentru a sublinia existența unei anumite forme de asociere între două variabile studiate. De exemplu, în domeniul medical putem spune că am observat o "corelație" între zilele cu ceață și declanșarea crizelor de astm.

Pe de altă parte în domeniul biostatisticii, termenul de corelație este folosit pentru a reliefa existența unei asocieri între două variabile cantitative. În mod obișnuit, suntem tentați să presupunem că această asociere este "liniară", în sensul că una dintre variabile (să o notăm cu y) crește sau descrește într-o anumită măsură, "proporțional" cu creșterea sau descreșterea celeilalte variabile studiate (notată cu x).

Variabilă DEPEDENTĂ vs variabilă INDEPEDENTĂ

Variabila "**y**" va fi considerată "**variabila dependentă**", ce prezintă un anumit grad de asociere față de variabila "**x**", "**variabila independentă**".

REGRESIE

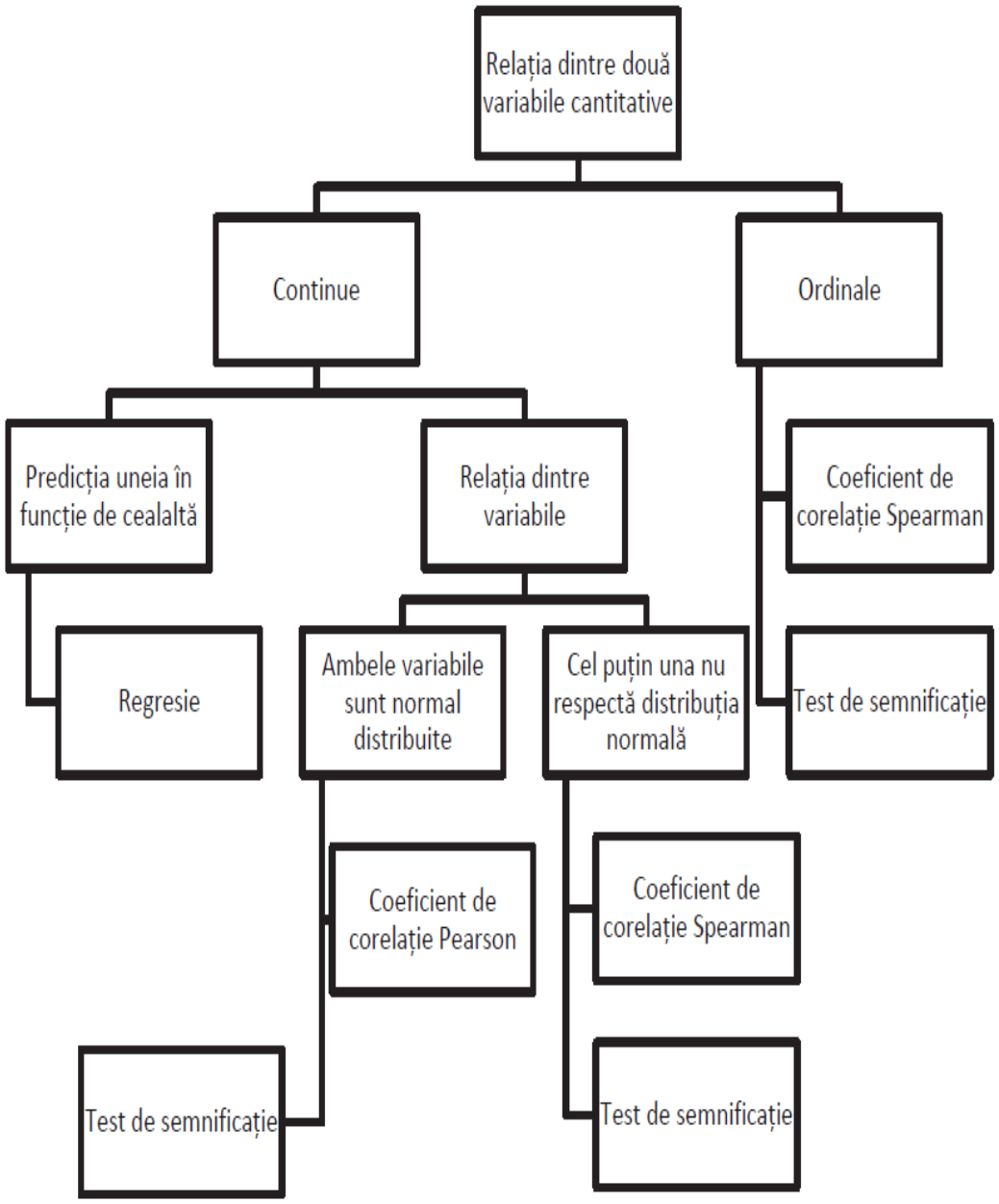
În astfel de circumstanțe este adesea folosit termenul de **regresie** (liniară), termen ce implică estimarea celei mai potrivite linii drepte care să reliefeze asocierea, așa cum veți vedea în următorul curs.

2. COEFICIENT DE CORELAȚIE

Gradul de asociere al variabilelor menționate anterior este măsurat cu ajutorul coeficientului de corelație, propus de Pearson și Bravais și care este o măsură a asocierii “liniare” a celor două variabile. Dacă însă de asocierea dintre variabile nu este liniară ci poate fi exprimată doar cu ajutorul unor curbe, aceasta înseamnă că sunt necesare alte măsurători ale corelației, folosind metode mai complexe.

Coeficientul de corelație este o măsură a asocierii între două variabile (variabila independentă și cea dependentă) ce poate lua valori cuprinse între -1….0….+1. Există două tipuri de corelație:

- Parametrică (Pearson)
- Neprametrică (Kendall , Spearman)



ALEGEREA CORECTĂ A METODEI DE CORELAȚIE (PARAMETRICĂ/NEPARAMETRICĂ)

3.COEFICIENT DE CORELAȚIE PARAMETRICĂ (PEARSON)

3. CORELAȚIA PARAMETRICĂ

Coeficientul lui Pearson, pentru variabile cantitative, numerice.

Vezi Aplicația rezolvată nr. 1 – Calculul coeficientului de corelație simplă (pe baza formulei)

- Coeficientul de corelație „r” este un număr calculat direct din datele observate și poate varia între –1 și +1. Formulele de calcul ale coeficientului de corelație „r” diferă ușor, în funcție de notațiile folosite de diverși autori.
- Dacă xi sunt valorile măsurate ale variabilei X (variabila independentă) și yi sunt valorile măsurate ale variabilei Y (variabila dependentă), atunci coeficientul de corelație se calculează astfel:

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

unde n= numărul perechilor de date, iar Sx, Sy – abaterile standard în cazul celor două variabile.

APLICAȚIA REZOLVATĂ NR. 1 – CALCULUL COEFICIENTULUI DE CORELAȚIE SIMPLĂ (PE BAZA FORMULEI)

Coeficientul lui Pearson, pentru variabile cantitative, numerice.

Tabelul 2. Calcule intermediare pentru coeficientul de corelație simplă

Anul	x_i	y_i	$x_i * y_i$	x_i^2	y_i^2
2007	2141,3	52,6	112632,38	4585165,69	2766,76
2008	2865,4	13,7	39255,98	8210517,16	187,69
2009	3694,0	78,9	291456,6	13645636,0	6225,21
2010	4670.9	172,1	803861,89	21817306,81	29618,41
2011	5586,2	167,6	936247,12	31205630,44	28089,76
2012	6438,4	274,5	1767340,8	41452994,56	75350,25
2013	7260,7	331,0	2403291,7	52717764,49	109561,0
Total	**Σx_i = 32656,9**	**Σy_i = 1090,4**	**$\Sigma x_i * y_i$ = 6354086,47**	**Σx_i^2 = 173635015**	**Σy_i^2 = 251799,1**

Înlocuind valorile în formula coeficientului de corelație simplă, rezultă:

$$R_{xy} = \frac{7 * 6354086,47 - 32656,9 * 1090,4}{\sqrt{[7 * 173635015 - 32656,9^2][7 * 251799,1 - 1090,4^2]}}$$

$$R_{xy} = \frac{869521,53}{\sqrt{148971988,4 * 573621,4}} = \frac{869521,53}{9244107} = 0,959478$$

MATRICE DE CORELAȚIE

(varianta "clasică)

Corelații negative /inverse

Corelații pozitive/ directe

	Mar-02	Mar-03	Sep-04	Sep-05	Mar-06	Sep-06	Mar-07	Sep-07	Sep-08	Sep-09	Sep-10	Sep-11	Sep-15
Mar-02	1	0.79906	0.19275	0.29824	-0.11094	-0.51331	-0.40007	-0.2006	-0.31702	-0.11459	-0.34535	-0.0612	-0.19537
Mar-03	0.79906	1	0.15338	0.21736	-0.17121	-0.52272	-0.46223	-0.34609	-0.28007	-0.17346	-0.37619	-0.06178	-0.28296
Sep-04	0.19275	0.15338	1	0.03591	0.08622	-0.21323	-0.21355	0.02134	0.00487	-0.02944	0.06525	0.05332	-0.25931
Sep-05	0.29824	0.21736	0.03591	1	0.33257	-0.1912	0.02597	0.00351	0.04272	-0.12306	-0.12325	0.44208	-0.09561
Mar-06	-0.11094	-0.17121	0.08622	0.33257	1	0.10096	0.23691	0.11148	0.11228	0.04507	-0.06023	0.26882	-0.08068
Sep-06	-0.51331	-0.52272	-0.21323	-0.1912	0.10096	1	0.52186	0.43776	0.26125	0.18527	0.2786	-0.0665	0.16822
Mar-07	-0.40007	-0.46223	-0.21355	0.02597	0.23691	0.52186	1	0.35157	0.15614	0.17502	0.32204	0.02287	0.37555
Sep-07	-0.2006	-0.34609	0.02134	0.00351	0.11148	0.43776	0.35157	1	0.20064	0.17582	0.3853	0.16359	0.21875
Sep-08	-0.31702	-0.28007	0.00487	0.04272	0.11228	0.26125	0.15614	0.20064	1	0.12242	0.1521	0.30017	0.0795
Sep-09	-0.11459	-0.17346	-0.02944	-0.12306	0.04507	0.18527	0.17502	0.17582	0.12242	1	0.07619	-0.01626	0.12753
Sep-10	-0.34535	-0.37619	0.06525	-0.12325	-0.06023	0.2786	0.32204	0.3853	0.1521	0.07619	1	0.13857	0.0632
Sep-11	-0.0612	-0.06178	0.05332	0.44208	0.26882	-0.0665	0.02287	0.16359	0.30017	-0.01626	0.13857	1	0.03666
Sep-15	-0.19537	-0.28296	-0.25931	-0.09561	-0.08068	0.16822	0.37555	0.21875	0.0795	0.12753	0.0632	0.03666	1

MATRICE DE CORELAȚIE

(output SPSS/PSPP)

Correlations

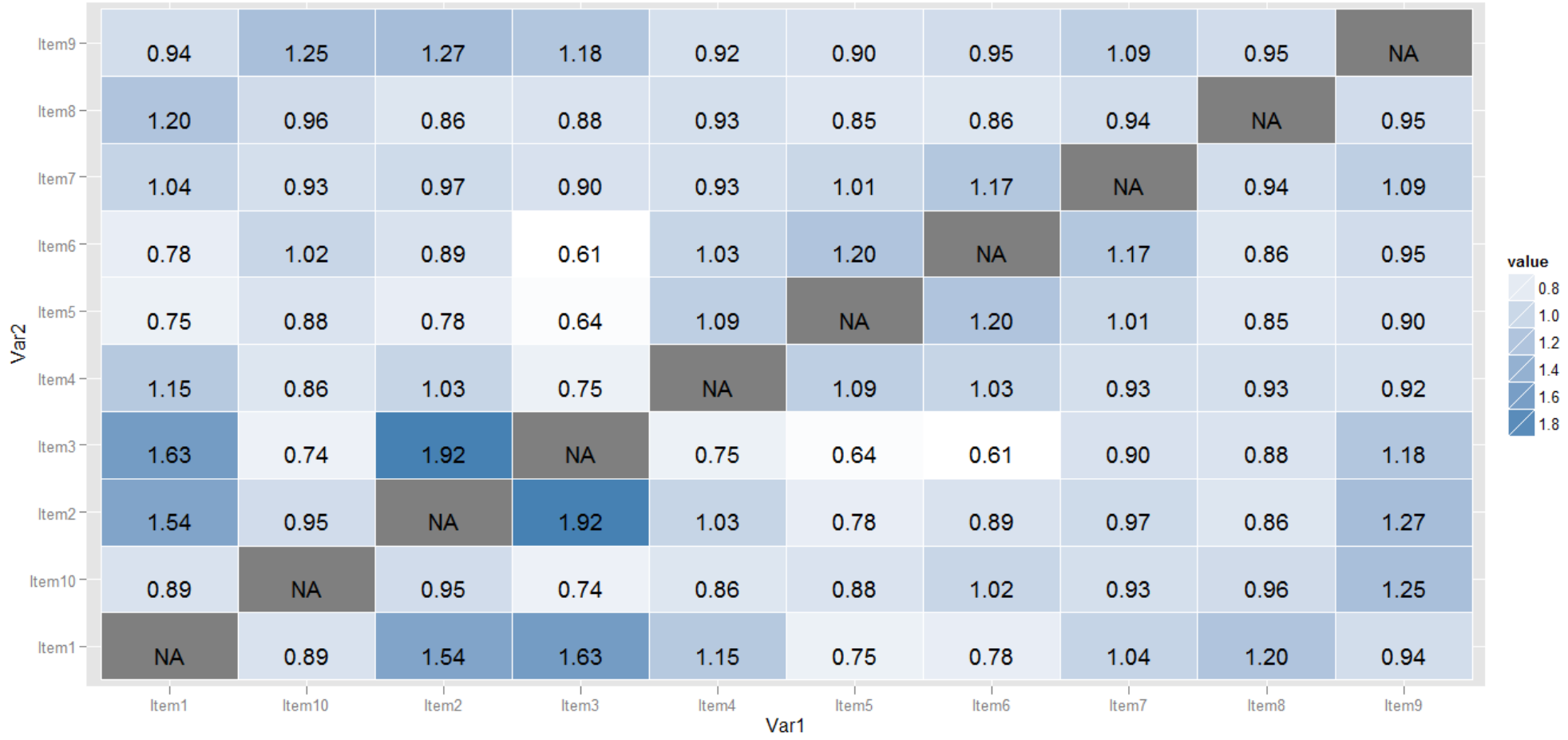
		reading score	writing score	math score	science score	female
reading score	Pearson Correlation[a]	1	.597**	.662**	.630**	-.053
	Sig. (2-tailed)[b]	.	.000	.000	.000	.455
	N[c]	200	200	200	200	200
writing score	Pearson Correlation	.597**	1	.617**	.570**	.256**
	Sig. (2-tailed)	.000	.	.000	.000	.000
	N	200	200	200	200	200
math score	Pearson Correlation	.662**	.617**	1	.631**	-.029
	Sig. (2-tailed)	.000	.000	.	.000	.680
	N	200	200	200	200	200
science score	Pearson Correlation	.630**	.570**	.631**	1	-.128
	Sig. (2-tailed)	.000	.000	.000	.	.071
	N	200	200	200	200	200
female	Pearson Correlation	-.053	.256**	-.029	-.128	1
	Sig. (2-tailed)	.455	.000	.680	.071	.
	N	200	200	200	200	200

**. Correlation is significant at the 0.01 level (2-tailed).

p- value> 0.05

MATRICE DE CORELAȚIE

(Excel)



MATRICE DE CORELAȚIE

Correlation Matrix 3 Year

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1															
2	NA														
3	0.54	NA													
4	0.52	NA	0.21												
5	0.59	NA	0.34	0.45											
6	0.50	NA	0.09	0.46	0.27										
7	0.21	NA	-0.16	0.69	0.28	0.32									
8	-0.01	NA	-0.15	0.15	0.11	0.13	0.35								
9	0.44	NA	0.14	0.48	0.64	0.52	0.37	-0.08							
10	0.58	NA	0.50	0.49	0.69	0.31	0.41	0.08	0.63						
11	0.62	NA	0.31	0.54	0.75	0.53	0.46	0.02	0.85	0.78					
12	-0.03	NA	-0.20	0.43	-0.06	-0.08	0.52	0.55	-0.10	0.04	-0.01				
13	0.23	NA	0.08	0.55	0.20	0.15	0.69	0.48	0.17	0.45	0.33	0.67			
14	0.57	NA	0.27	0.67	0.59	0.70	0.47	0.04	0.69	0.68	0.79	-0.02	0.32		
15	0.02	NA	-0.38	0.43	0.10	0.01	0.64	0.39	0.08	0.10	0.14	0.71	0.66	0.11	

Degree of Correlation

High	Moderate	None	Moderately Negative	Highly Negative
0.70 to 1.00	0.11 to 0.69	0.10 to -0.10	-0.11 to -0.69	-0.70 to -1.00

MATRICE DE CORELAȚIE

Correlation Matrix

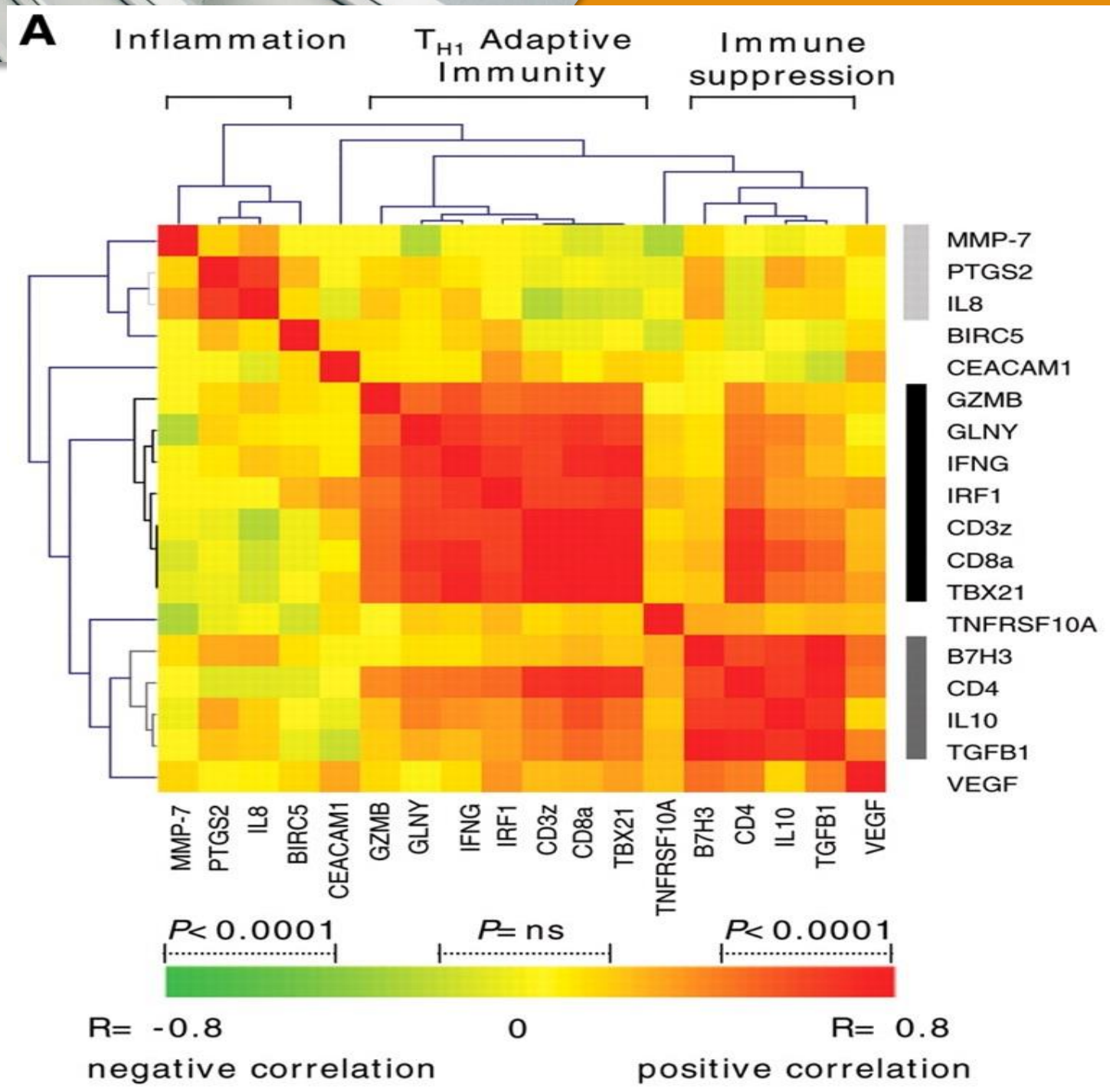
	S&P 500	AC World	World xUS	EAFE	Europe	EM	Japan	cash	US IG	high yield	glob IG	ST Gvt	EMD	REITs	LLs	gold	commodities	GSCI Energy	GSCI Base Metals	GSCI Prec Metals	GSCI Softs	duration	US 60/40	Small Caps
S&P 500	1.00	0.99	0.94	0.94	0.93	0.87	0.93	0.44	-0.15	0.90	-0.01	-0.39	0.54	0.73	0.73	0.11	0.21	0.77	0.83	-0.09	-0.07	-0.34	0.97	0.97
AC World	0.99	1.00	0.98	0.98	0.96	0.92	0.95	0.47	-0.12	0.90	0.05	-0.35	0.61	0.71	0.69	0.10	0.16	0.77	0.84	-0.07	-0.02	-0.33	0.98	0.98
World xUS	0.94	0.98	1.00	0.99	0.98	0.96	0.94	0.49	-0.07	0.87	0.13	-0.28	0.67	0.66	0.61	0.09	0.08	0.75	0.82	-0.05	0.04	-0.29	0.95	0.95
EAFE	0.94	0.98	0.99	1.00	0.99	0.92	0.95	0.50	-0.07	0.88	0.11	-0.29	0.65	0.64	0.64	0.09	0.12	0.78	0.79	-0.07	0.03	-0.29	0.95	0.95
Europe	0.93	0.96	0.98	0.99	1.00	0.90	0.91	0.46	-0.09	0.88	0.10	-0.27	0.64	0.58	0.65	0.11	0.15	0.82	0.76	-0.04	-0.02	-0.31	0.93	0.93
EM	0.87	0.92	0.96	0.92	0.90	1.00	0.87	0.43	-0.05	0.76	0.17	-0.24	0.66	0.61	0.44	0.07	-0.02	0.60	0.81	-0.01	0.08	-0.28	0.89	0.89
Japan	0.93	0.95	0.94	0.95	0.91	0.87	1.00	0.52	-0.03	0.84	0.08	-0.32	0.59	0.78	0.63	0.06	0.10	0.69	0.80	-0.13	0.09	-0.20	0.94	0.93
cash	0.44	0.47	0.49	0.50	0.46	0.43	0.52	1.00	0.38	0.47	0.51	0.21	0.67	0.37	0.27	0.10	-0.27	0.34	0.17	0.20	0.24	0.19	0.52	0.42
US IG	-0.15	-0.12	-0.07	-0.07	-0.09	-0.05	-0.03	0.38	1.00	-0.01	0.92	0.90	0.49	0.07	-0.40	0.35	-0.79	-0.20	-0.09	0.58	-0.31	0.95	0.07	-0.20
high yield	0.90	0.90	0.87	0.88	0.88	0.76	0.84	0.47	-0.01	1.00	0.14	-0.26	0.72	0.74	0.82	0.23	0.19	0.90	0.76	0.06	-0.14	-0.22	0.91	0.90
glob IG	-0.01	0.05	0.13	0.11	0.10	0.17	0.08	0.51	0.92	0.14	1.00	0.85	0.68	0.09	-0.34	0.52	-0.80	-0.05	0.04	0.75	-0.24	0.79	0.21	-0.04
ST Gvt	-0.39	-0.35	-0.28	-0.29	-0.27	-0.24	-0.32	0.21	0.90	-0.26	0.85	1.00	0.24	-0.20	-0.59	0.25	-0.81	-0.40	-0.36	0.58	-0.21	0.88	-0.20	-0.44
EMD	0.54	0.61	0.67	0.65	0.64	0.66	0.59	0.67	0.49	0.72	0.68	0.24	1.00	0.46	0.30	0.41	-0.32	0.58	0.57	0.44	-0.11	0.23	0.68	0.58
REITs	0.73	0.71	0.66	0.64	0.58	0.61	0.78	0.37	0.07	0.74	0.09	-0.20	0.46	1.00	0.64	0.04	-0.02	0.45	0.74	-0.11	0.13	-0.02	0.76	0.74
LLs	0.73	0.69	0.61	0.64	0.65	0.44	0.63	0.27	-0.40	0.82	-0.34	-0.59	0.30	0.64	1.00	-0.04	0.49	0.81	0.52	-0.27	-0.03	-0.51	0.64	0.76
gold	0.11	0.10	0.09	0.09	0.11	0.07	0.06	0.10	0.35	0.23	0.52	0.25	0.41	0.04	-0.04	1.00	-0.31	0.13	0.22	0.90	-0.52	0.31	0.19	0.10
commodities	0.21	0.16	0.08	0.12	0.15	-0.02	0.10	-0.27	-0.79	0.19	-0.80	-0.81	-0.32	-0.02	0.49	-0.31	1.00	0.49	0.11	-0.59	0.19	-0.78	0.02	0.18
GSCI Energy	0.77	0.77	0.75	0.78	0.82	0.60	0.69	0.34	-0.20	0.90	-0.05	-0.40	0.58	0.45	0.81	0.13	0.49	1.00	0.58	-0.08	-0.14	-0.40	0.73	0.76
GSCI Base Meta	0.83	0.84	0.82	0.79	0.76	0.81	0.80	0.17	-0.09	0.76	0.04	-0.36	0.57	0.74	0.52	0.22	0.11	0.58	1.00	-0.02	-0.04	-0.23	0.83	0.86
GSCI Prec Metal	-0.09	-0.07	-0.05	-0.07	-0.04	-0.01	-0.13	0.20	0.58	0.06	0.75	0.58	0.44	-0.11	-0.27	0.90	-0.59	-0.08	-0.02	1.00	-0.49	0.51	0.04	-0.11
GSCI Softs	-0.07	-0.02	0.04	0.03	-0.02	0.08	0.09	0.24	-0.31	-0.14	-0.24	-0.21	-0.11	0.13	-0.03	-0.52	0.19	-0.14	-0.04	-0.49	1.00	-0.32	-0.11	0.00
duration	-0.34	-0.33	-0.29	-0.29	-0.31	-0.28	-0.20	0.19	0.95	-0.22	0.79	0.88	0.23	-0.02	-0.51	0.31	-0.78	-0.40	-0.23	0.51	-0.32	1.00	-0.14	-0.39
US 60/40	0.97	0.98	0.95	0.95	0.93	0.89	0.94	0.52	0.07	0.91	0.21	-0.20	0.68	0.76	0.64	0.19	0.02	0.73	0.83	0.04	-0.11	-0.14	1.00	0.95
Small Caps	0.97	0.98	0.95	0.95	0.93	0.89	0.93	0.42	-0.20	0.90	-0.04	-0.44	0.58	0.74	0.76	0.10	0.18	0.76	0.86	-0.11	0.00	-0.39	0.95	1.00

Data source: FMRCo, Bloomberg, Haver Analytics, FactSet. Data as of 10/11/2019. Past performance is no guarantee of future results.

MATRICE DE CORELAȚIE

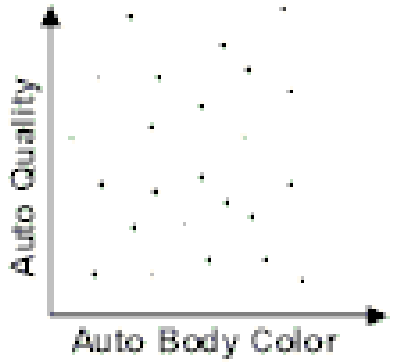
(**Excel**

Correlations matrix & heat map)

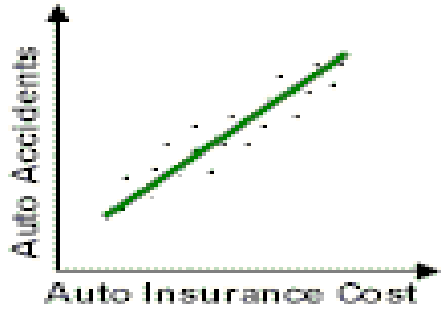


Putem întâlni următoarele situații:

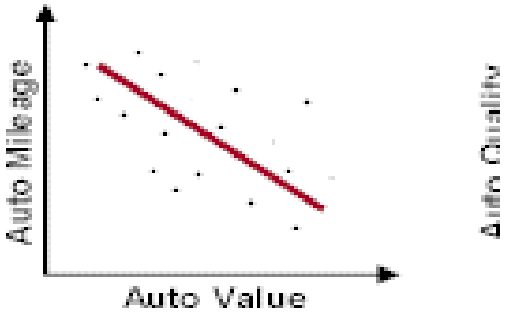
Zero



Positive



Negative



Dacă r = 0

- atunci înseamnă că **NU** avem nici o corelaţie între cele două variabile.
- De exemplu, **nu există nici legătură între presiunea sanguină şi numărul de fire de păr din cap**.

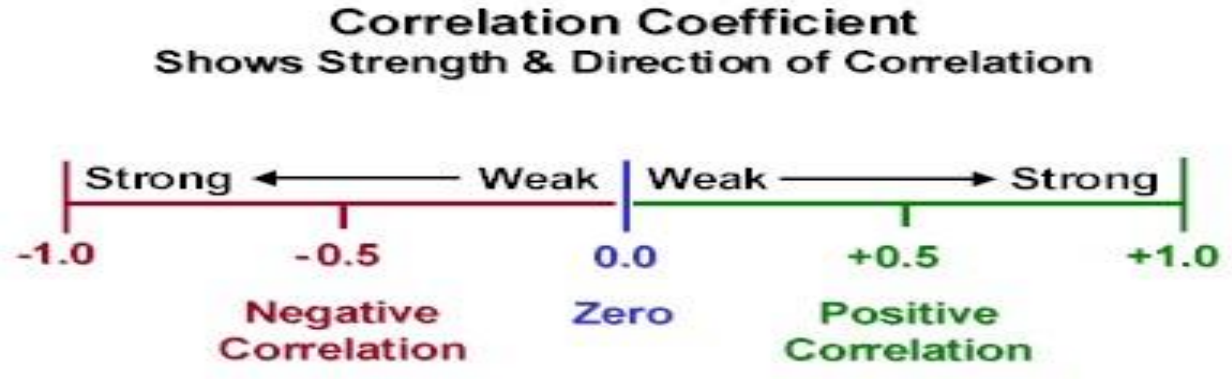
Dacă r = + 1

- înseamnă că avem o corelaţie **pozitivă/directă** perfectă, adică există o dependenţă directă între cele două variabile. O persoană care are o valoare mare la prima variabilă va avea o valoare mare şi la cea de a doua. De asemenea, valoarea unei variabile poate fi prevăzută exact pe baza valorii celei de a doua variabile.
- Un exemplu de acest tip este **corelaţia dintre vârsta unui copac şi numărul său de inele.**

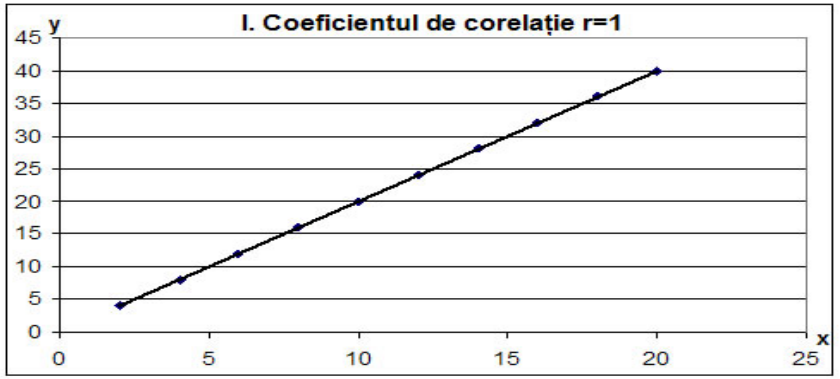
Dacă r = -1

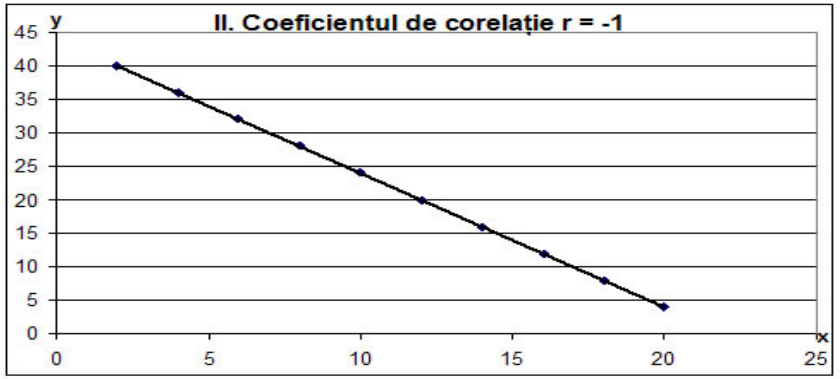
- atunci avem o dependenţă **inversă/negativă** perfectă. O valoare mare a unei variabile înseamnă o valoare mică a celeilalte variabile.

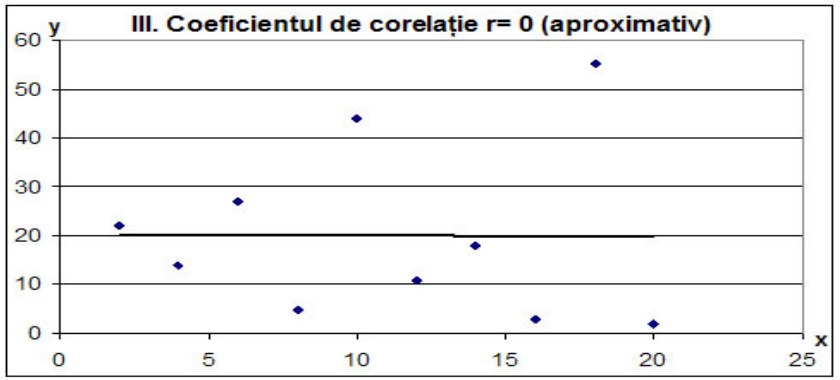
VALOAREA COEFICIENTULUI DE CORELAȚIE ȘI SMENIFICAȚIA LUI

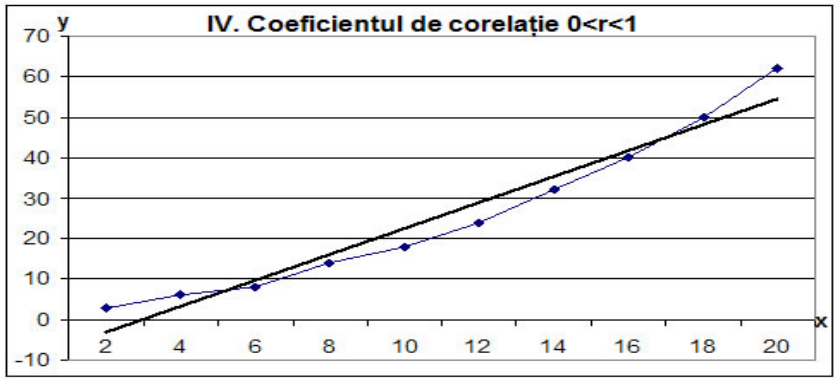


- Dacă coeficientul de corelaţie este între 0 şi +1 sau între –1 şi 0, atunci valoarea lui **r** ne dă tăria dependenţei celor două variabile.
- Aceste situaţii sunt prezentate în figurile și exemplele următoare:









Dacă dorim să realizăm neapărat o clasificare a intensității asocierii (corelației) între variabila independentă și cea dependentă, putem considera, în valori absolute, următoarele intervale:

Interval	Interpretare
$\|0 < r < 0,19\|$	Asociere/corelație de intensitate foarte slabă
$\|0,20 < r < 0,39\|$	Asociere/corelație de intensitate slabă
$\|0,40 < r < 0,59\|$	Asociere/corelație de intensitate medie (moderată)
$\|0,60 < r < 0,79\|$	Asociere/corelație de intensitate puternică
$\|0,80 < r < 1\|$	Asociere/corelație de intensitate ft. puternică

Trebuie ţinut însă seama de faptul că aceste limite din tabelsunt oarecum arbitrare, astfel că, trebuie să ţinem seama şi de contextul în care am desfăşurat experimentele, respectiv în care am făcut măsurătorile.

TESTUL DE SEMNIFICAȚIE PENTRU COEFICIENTUL DE CORELAȚIE PEARSON

Semnificaţia coeficientului de corelaţie Pearson poate fi evaluată dacă valoarea observată a apărut datorită întâmplării (dacă este semnificativ diferită de zero).

Interpretarea probabilității furnizate de acest test este că datele experimentale ne permit sau nu ne permit enunţarea existenţei unei relaţii între variabilele luate în calcul.

Din punct de vedere matematic există mai multe posibilități de teste de evaluare a semnificației coeficientului de corelație (**test F - Fisher, test t Student**) dar interpretarea acestor teste și rezultatele produse sunt de cele mai multe ori identice.

Valoarea r	p > 0,05	p < 0,05
-0,25 la 0,25	Nu are semnificatie statistică	Corelaţie slabă sau nulă
0,25 la 0,50 (-0,25 la -0,50)	Nu are semnificatie statistică	Grad de asociere acceptabil
0,5 la 0,75 (-0,5 la -0,75)	Nu are semnificatie statistică	Corelaţie moderată spre bună
> 0,75 (< -0,75)	Nu are semnificatie statistică	Foarte bună asociere sau corelaţie
r < -1; r > 1	Eroare	Eroare

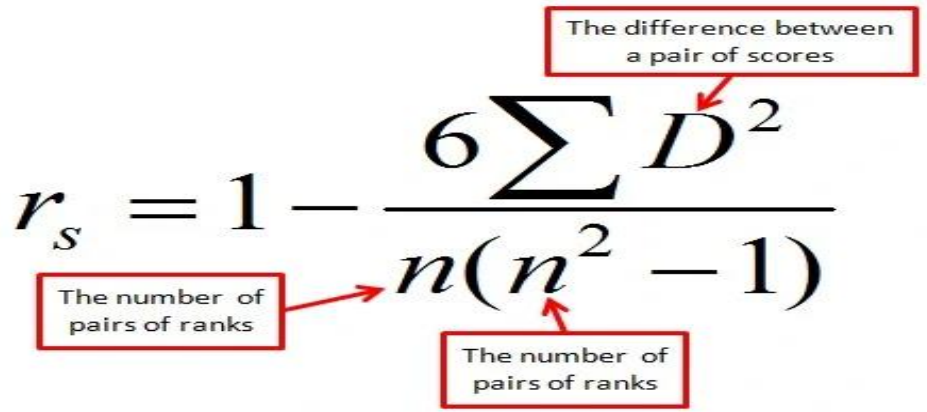
4.COEFICIENT DE CORELAȚIE NEPARAMETRICĂ (KENDALL / SPEARMAN)

CORELAREA RANGURILOR

Coeficientul de corelație Spearman

În măsura în care între diferitele grupuri rezultate din utilizarea variabilelor ca și elemente de grupare nu rezultă semnificație statistică (printr-o comparație de tip ANOVA) se poate utiliza coeficientul Spearman.

Coeficientul de corelaţie Spearman, notat rs, este analogul nonparametric al coeficientul de corelaţie Pearson, calculat pentru a fi utilizat în special cu date ordinale.



unde **n** este numărul de perechi de variabile

CORELAREA RANGURILOR

Coeficientul de corelație Kendall

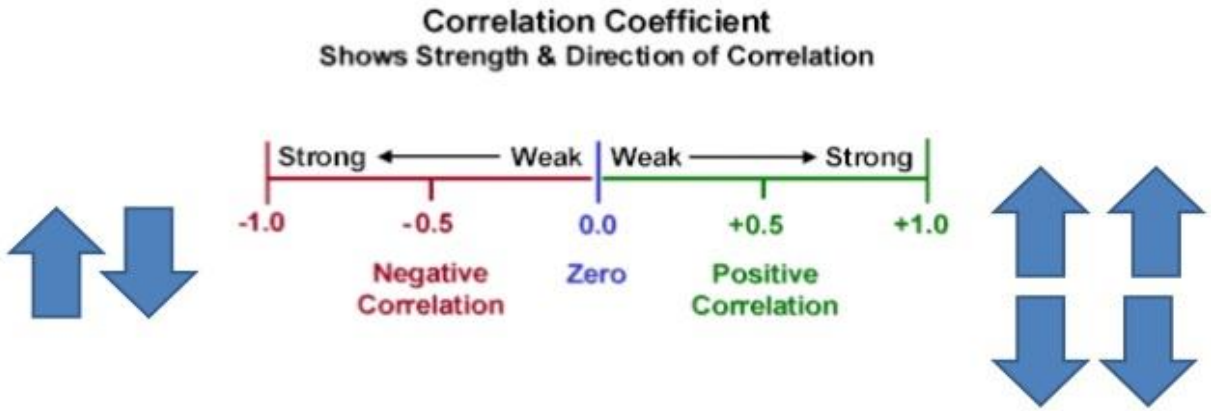
- Coeficientul Kendall:

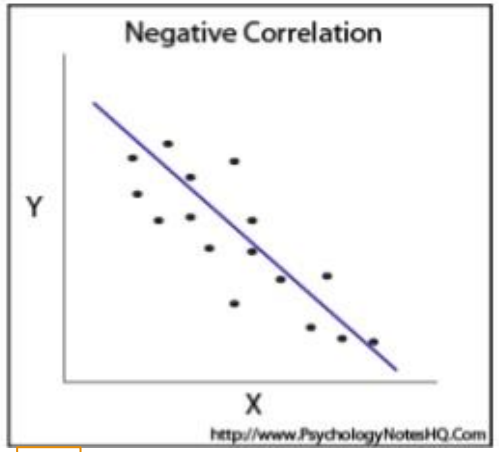
$$r_K = \frac{2(\sum P_i - \sum Q_i)}{n(n-1)}$$

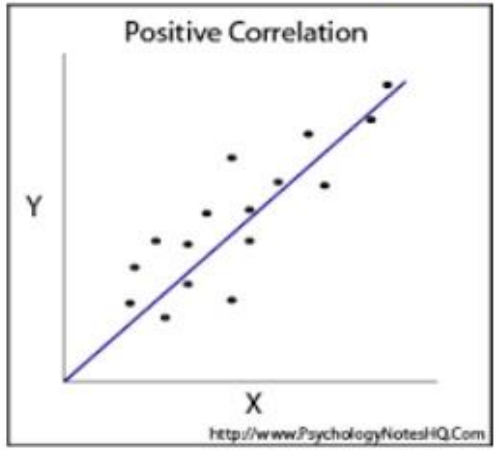
unde:
- P_i = *ranguri superioare*
- Q_i = *ranguri inferioare*
- n = *numărul observațiilor statistice*

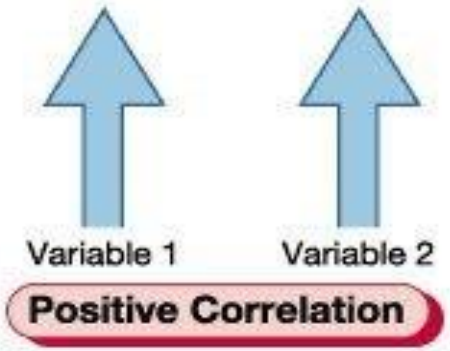
DIRECȚIA ȘI INTENSITATEA CORELAȚIEI

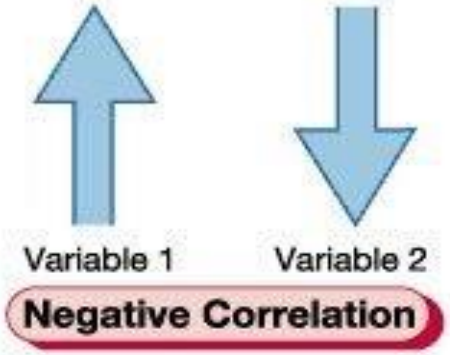
(aceleași principii ca și la corelația parametrică)

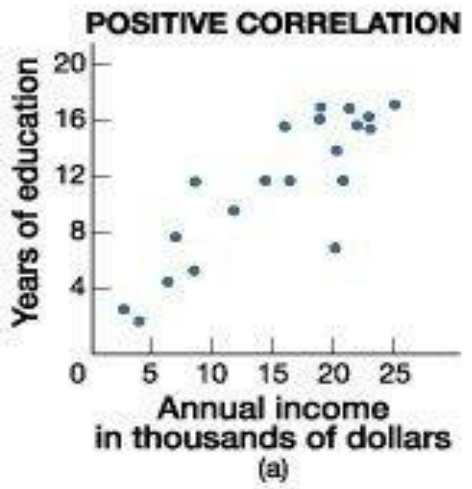


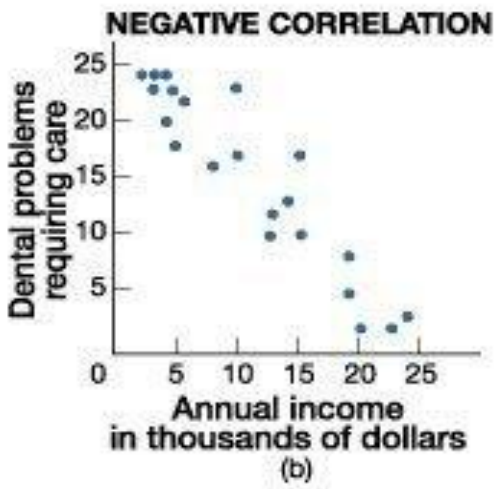










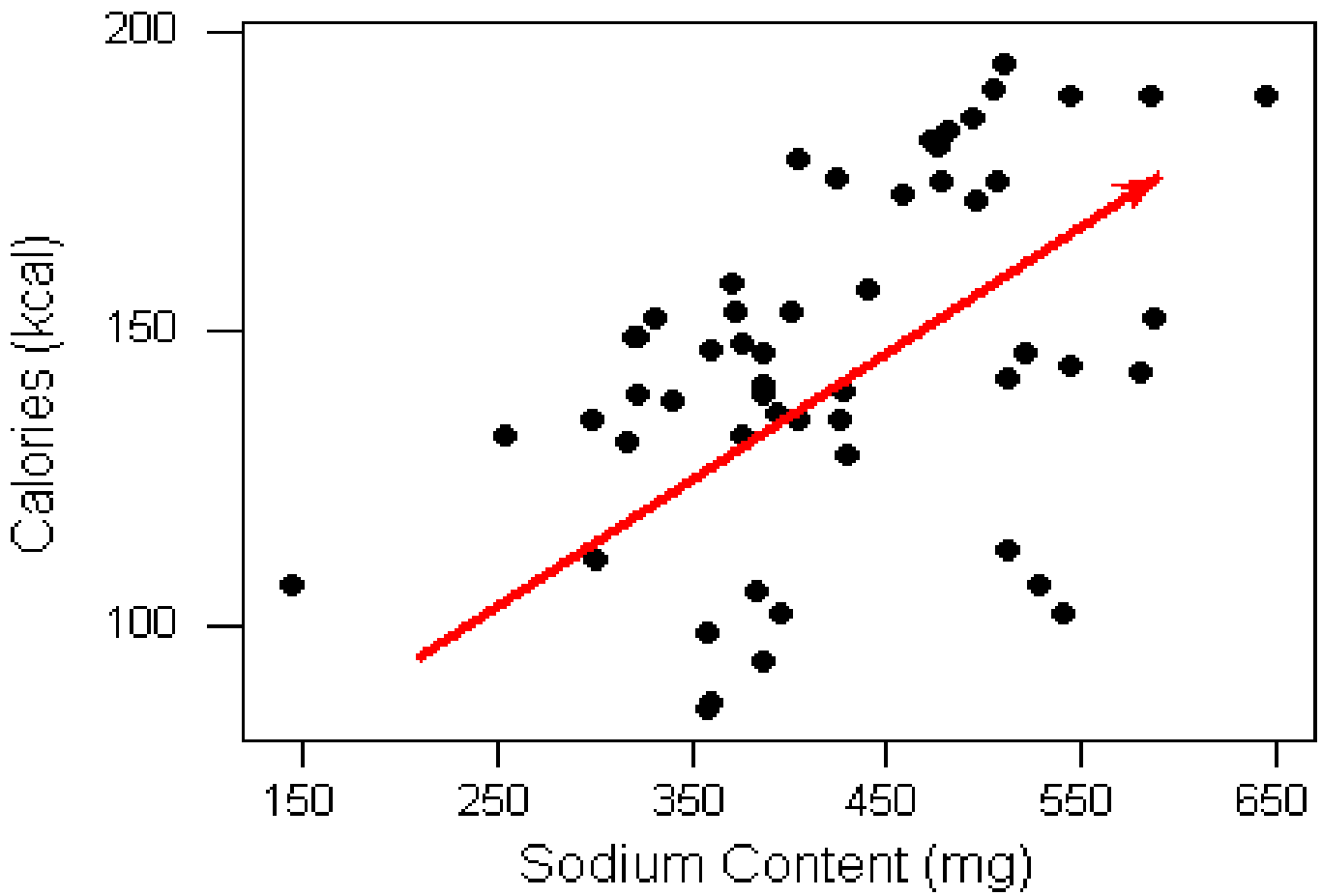


5. REPREZENTAREA GRAFICĂ A DATELOR ÎN CAZUL ANALIZEI CORELAȚIEI/REGRESIEI

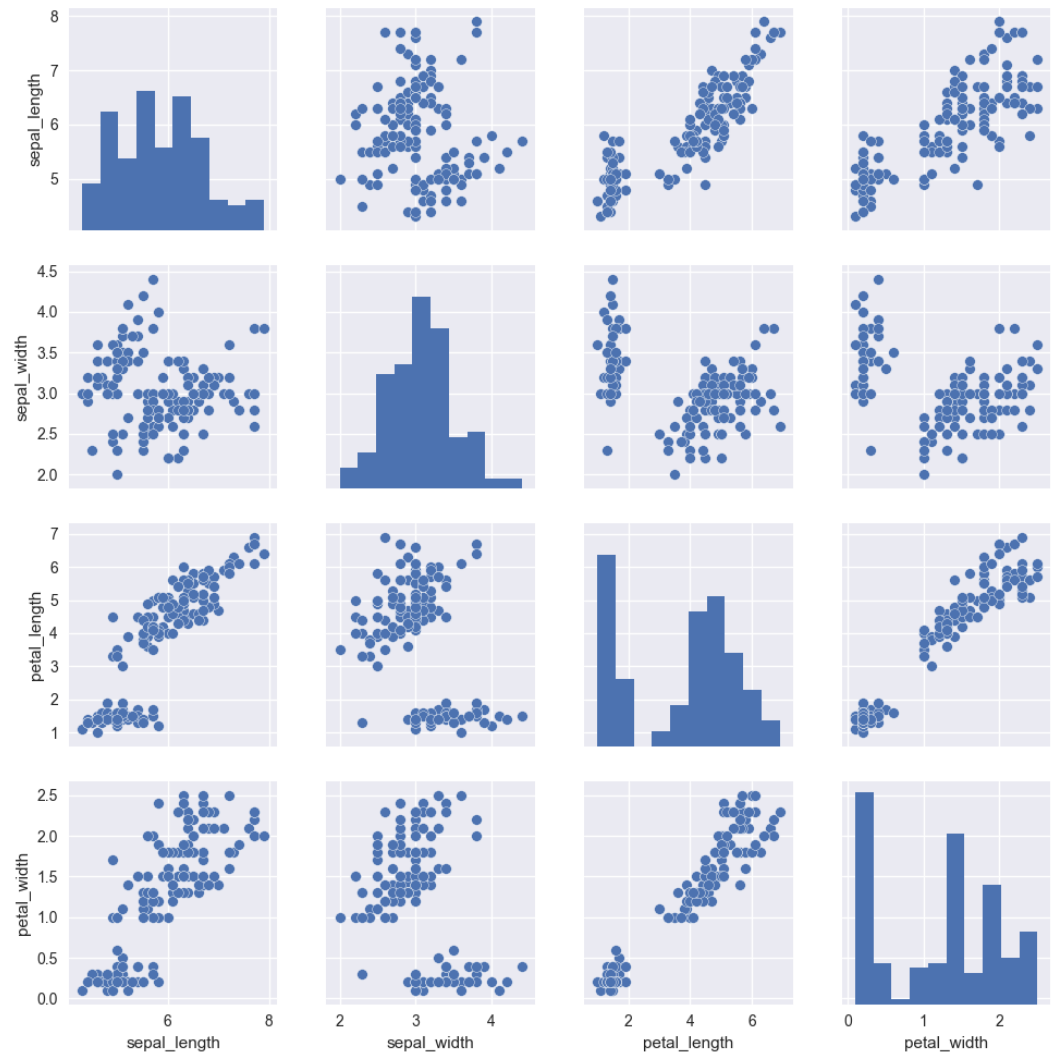
În momentul în are un cercetător a colectat două serii de observaţii (măsurători) şi doreşte să vadă dacă există o asociere între ele, primul lucru care trebuie făcut este reprezentarea lor grafică, sub forma unei aşa-numite “diagrame de împrăştiere” (scatter diagram - diagramă de împrăştiere a rezultatelor, într-o traducere aproximativă a termenului).

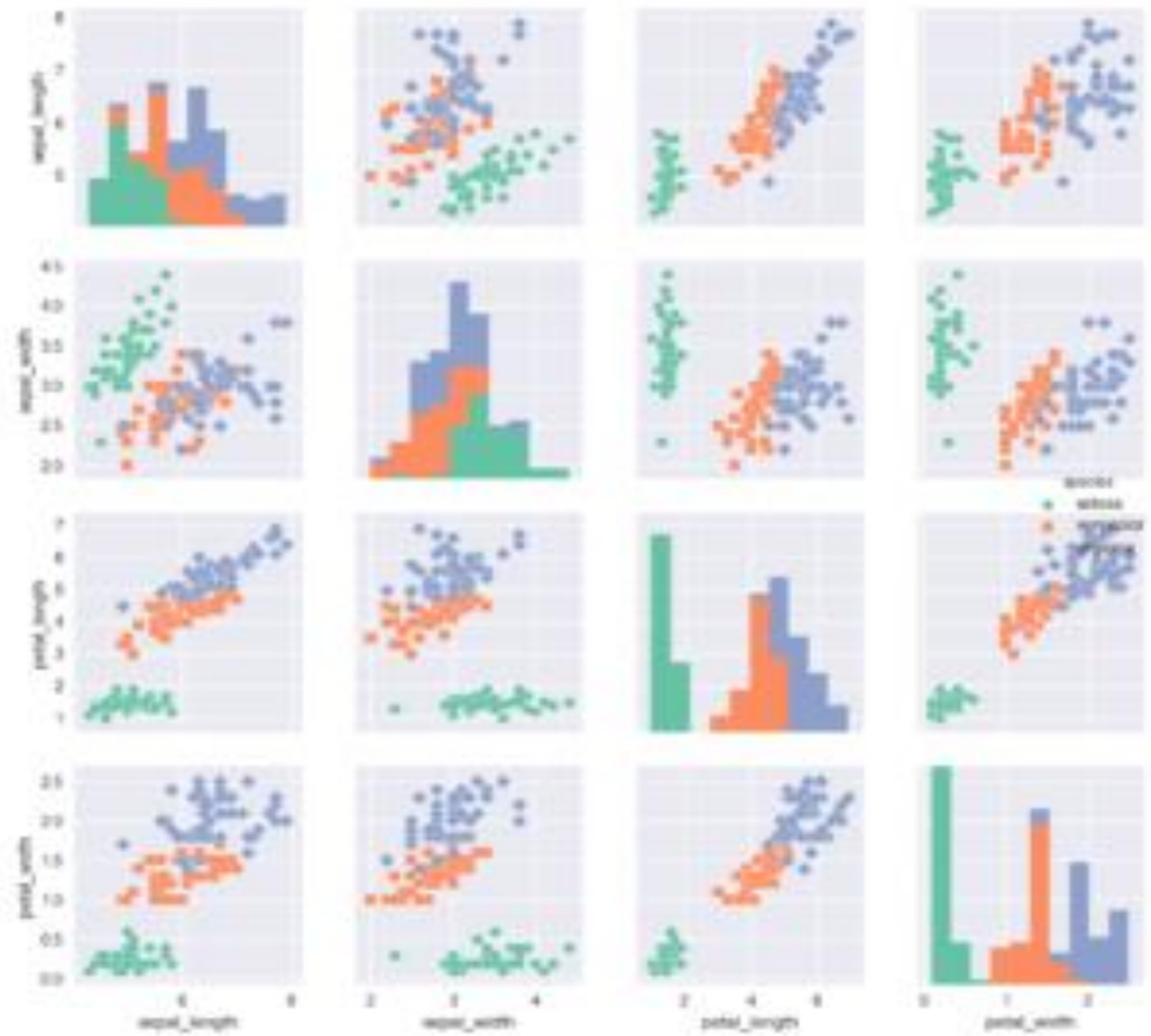
Majoritatea programelor de calcul tabelar (cum este MS Excel) oferă posibilitatea realizării unei astfel de diagrame, similară celei din figura alăturată:

Acest tip de diagramă foloseşte cele două axe de coordonate pentru a reprezentata cele două seturi de măsurători: pe axa X se află măsurătorile legate de variabila independentă, iar pe axa Y măsurătorile efectuate în cazul variabilei dependente

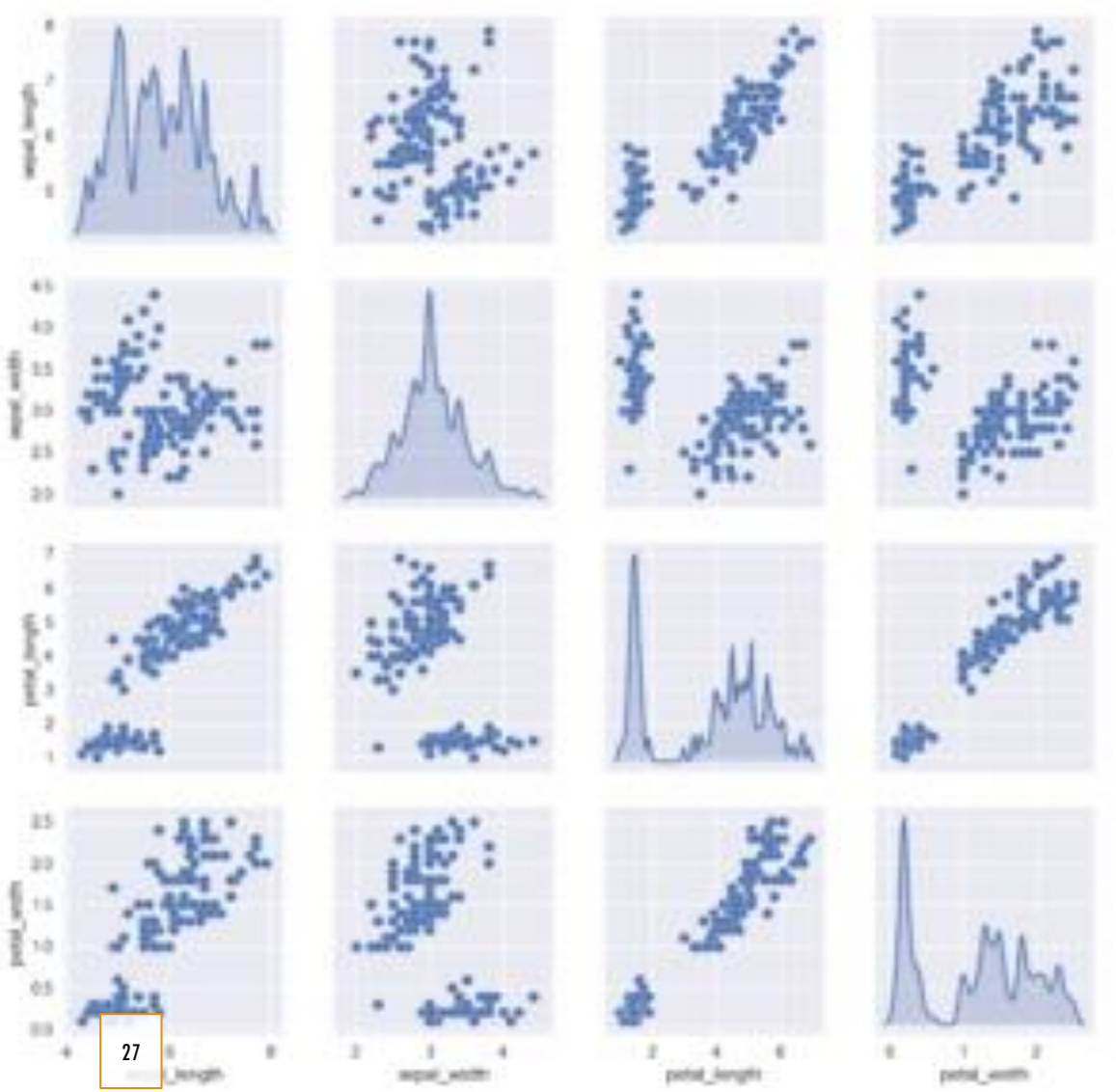


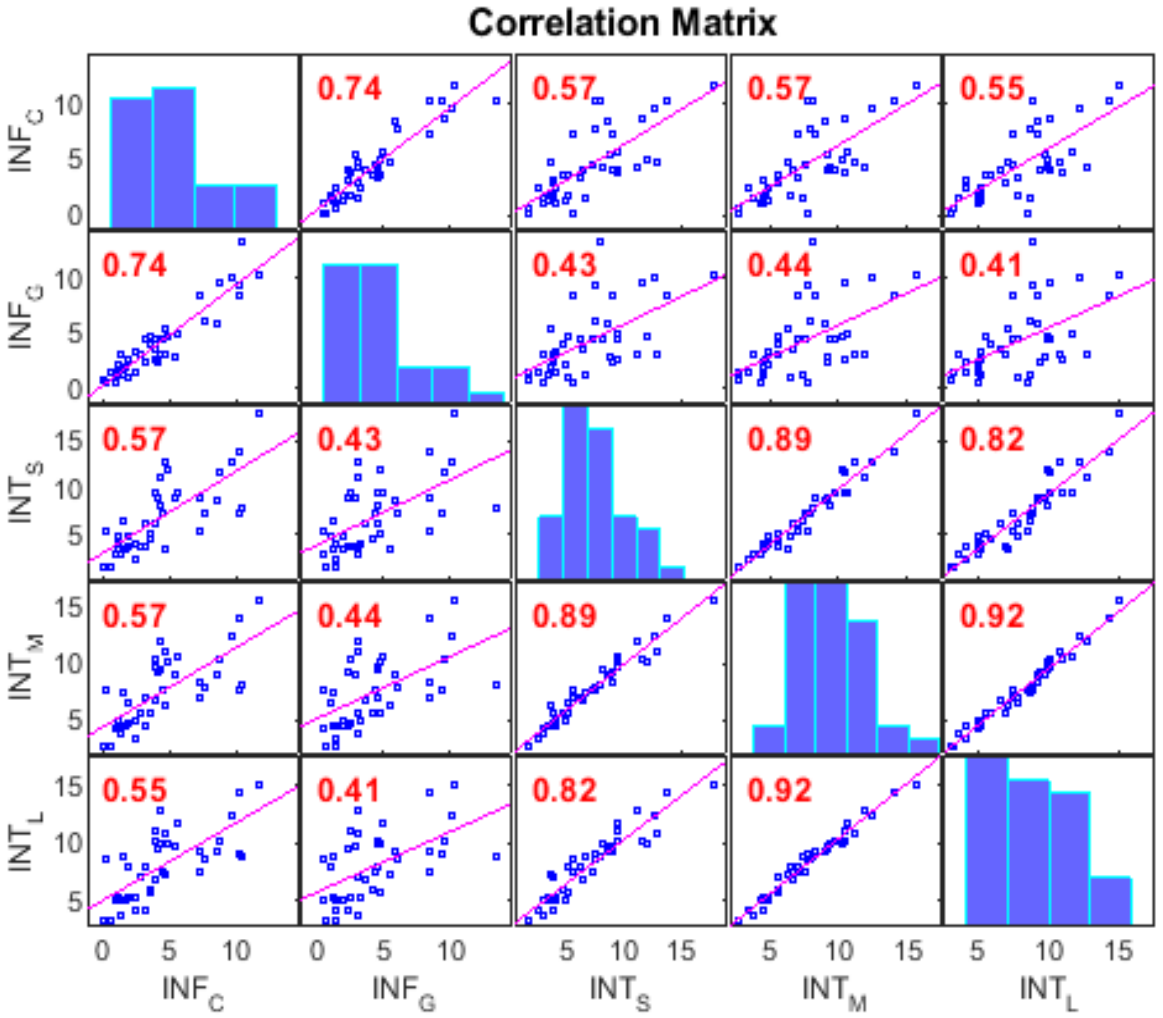
ALTE TIPURI DE REPREZENTĂRI GRAFICE PENTRU CORELAȚII



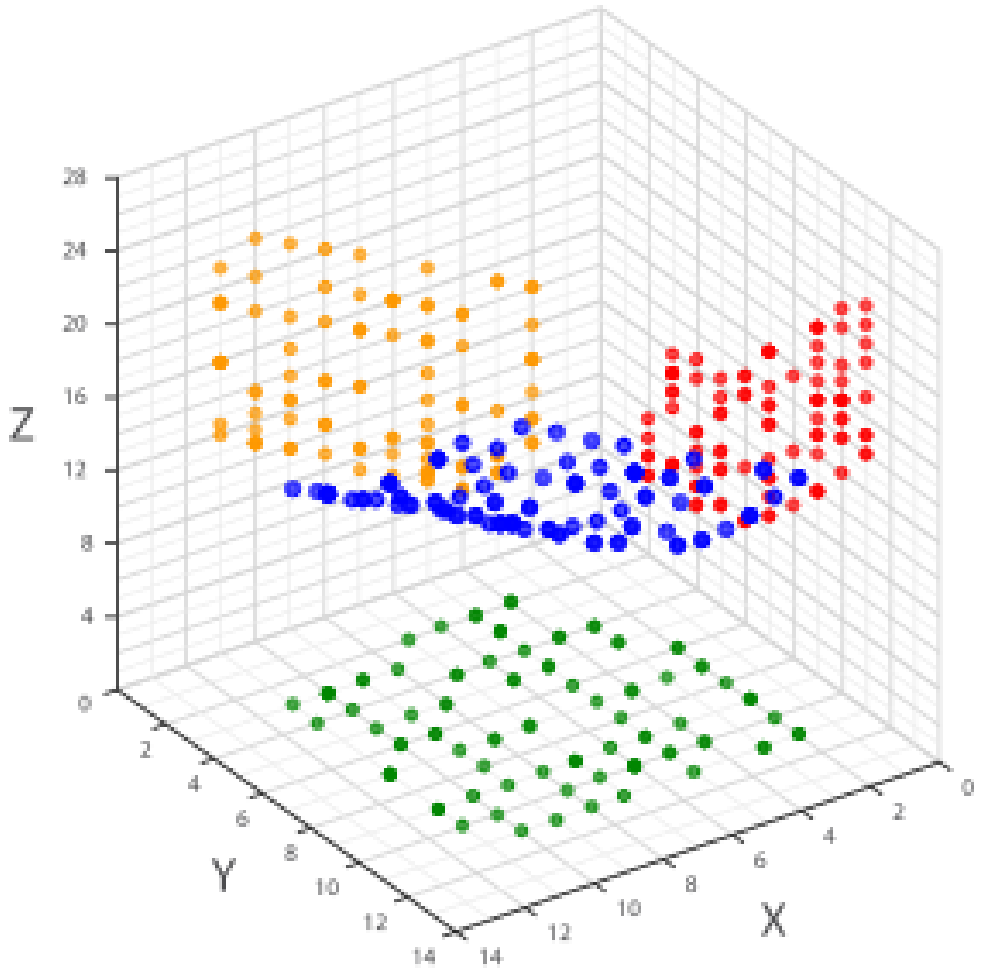


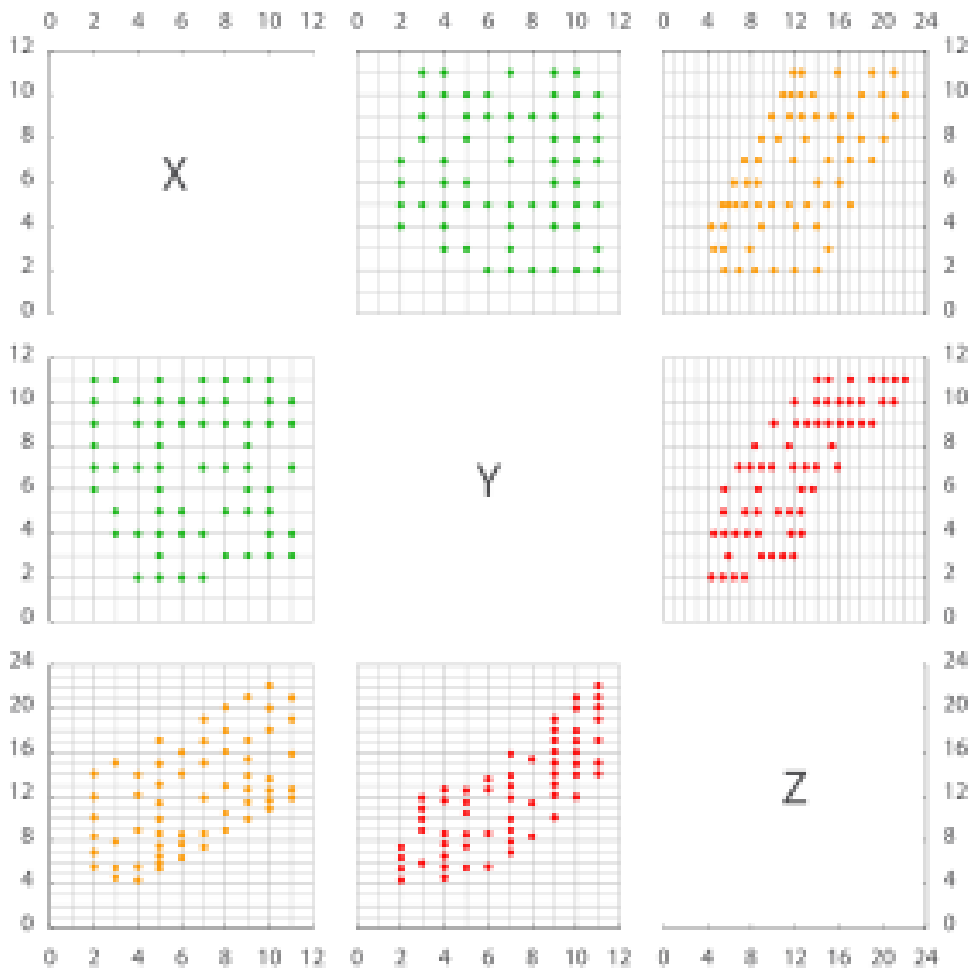
ALTE TIPURI DE REPREZENTĂRI GRAFICE PENTRU CORELAȚII



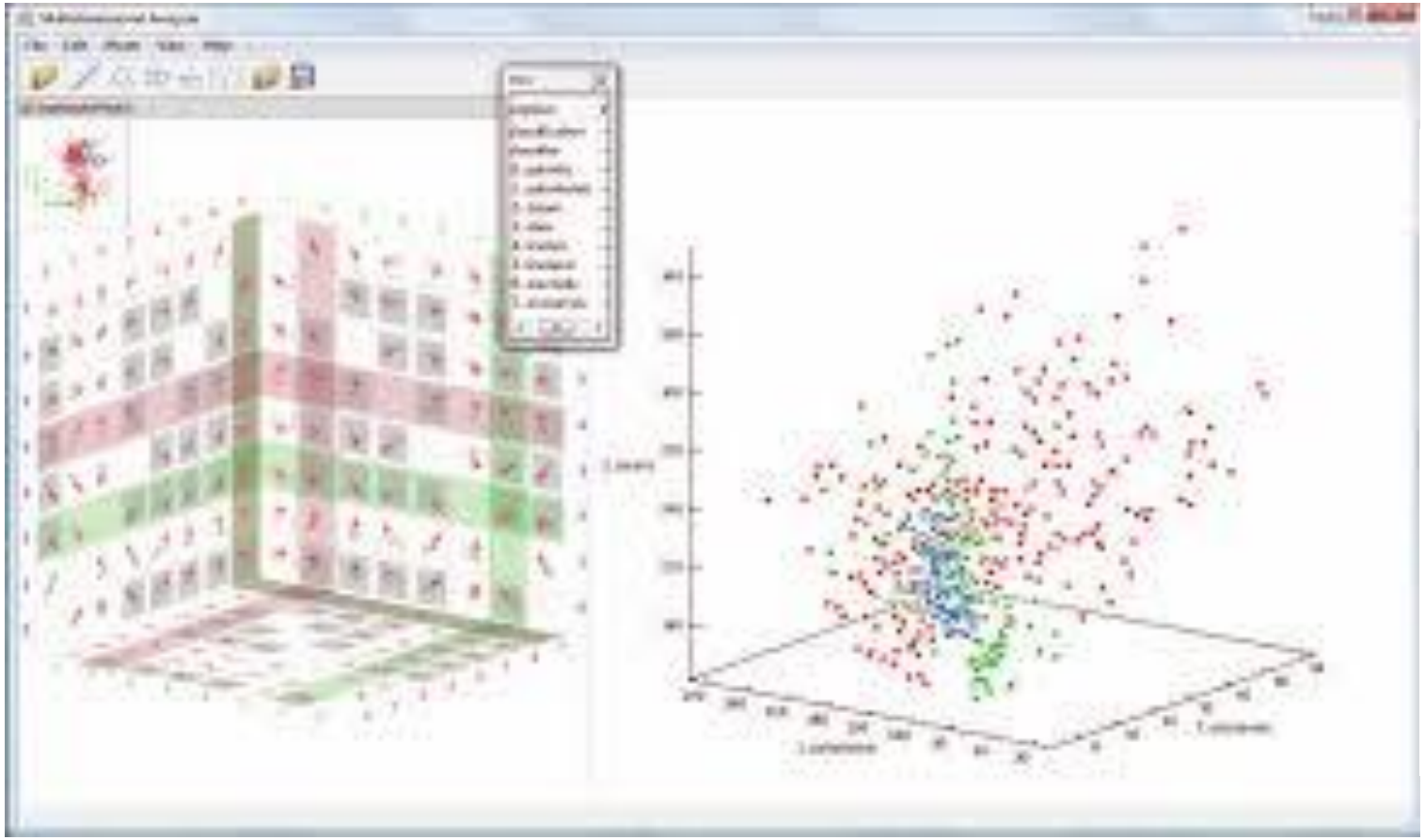


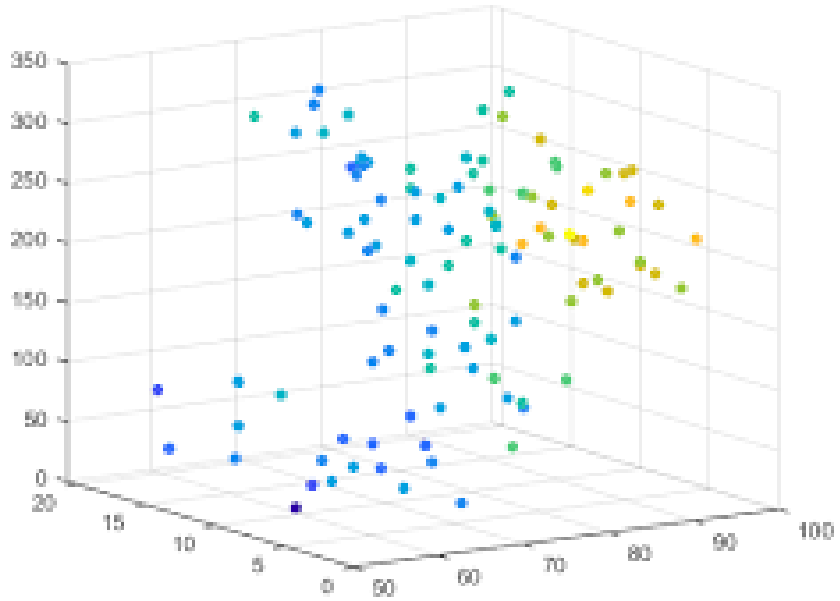
ALTE TIPURI DE REPREZENTĂRI GRAFICE PENTRU CORELAȚII



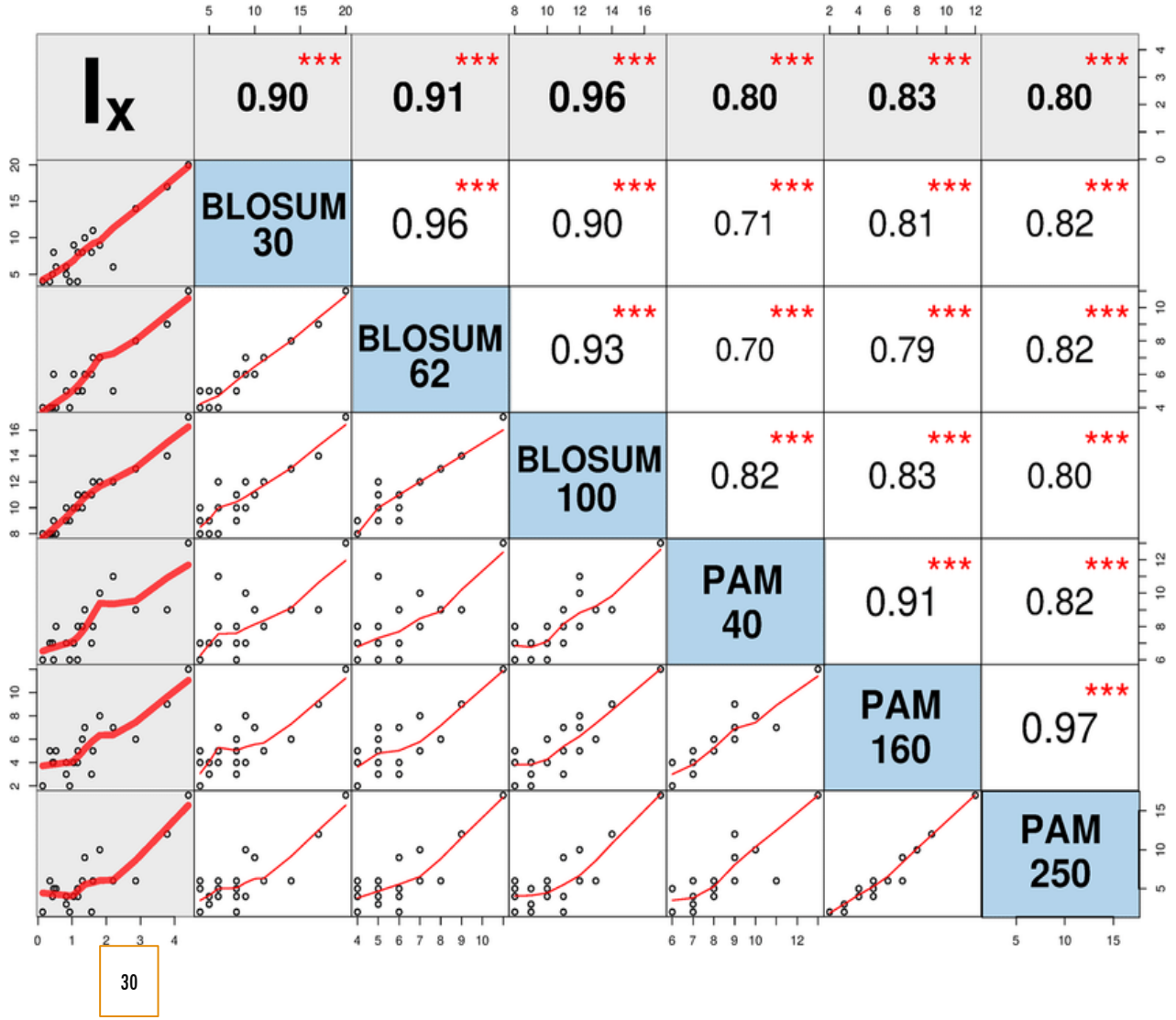


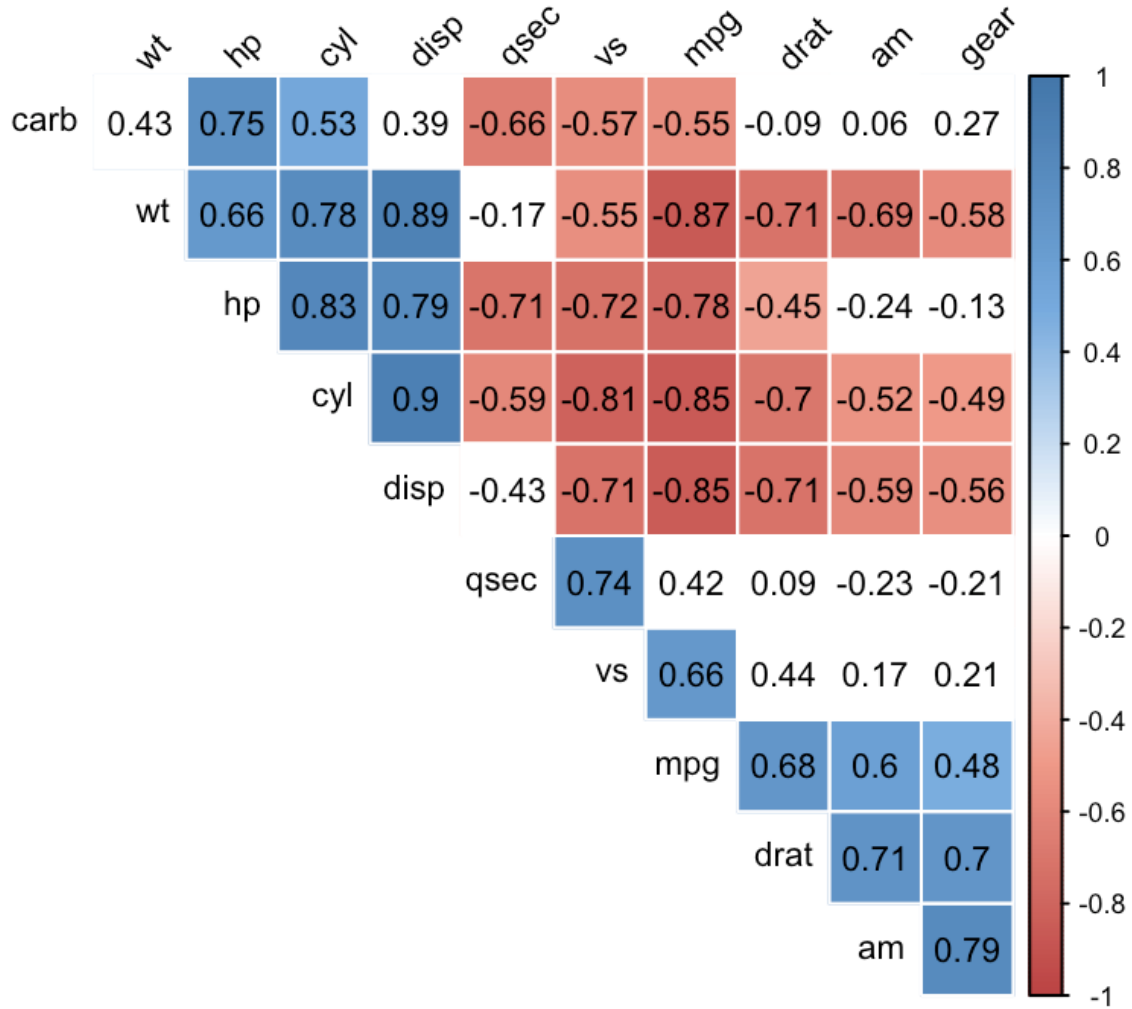
ALTE TIPURI DE REPREZENTĂRI GRAFICE PENTRU CORELAȚII



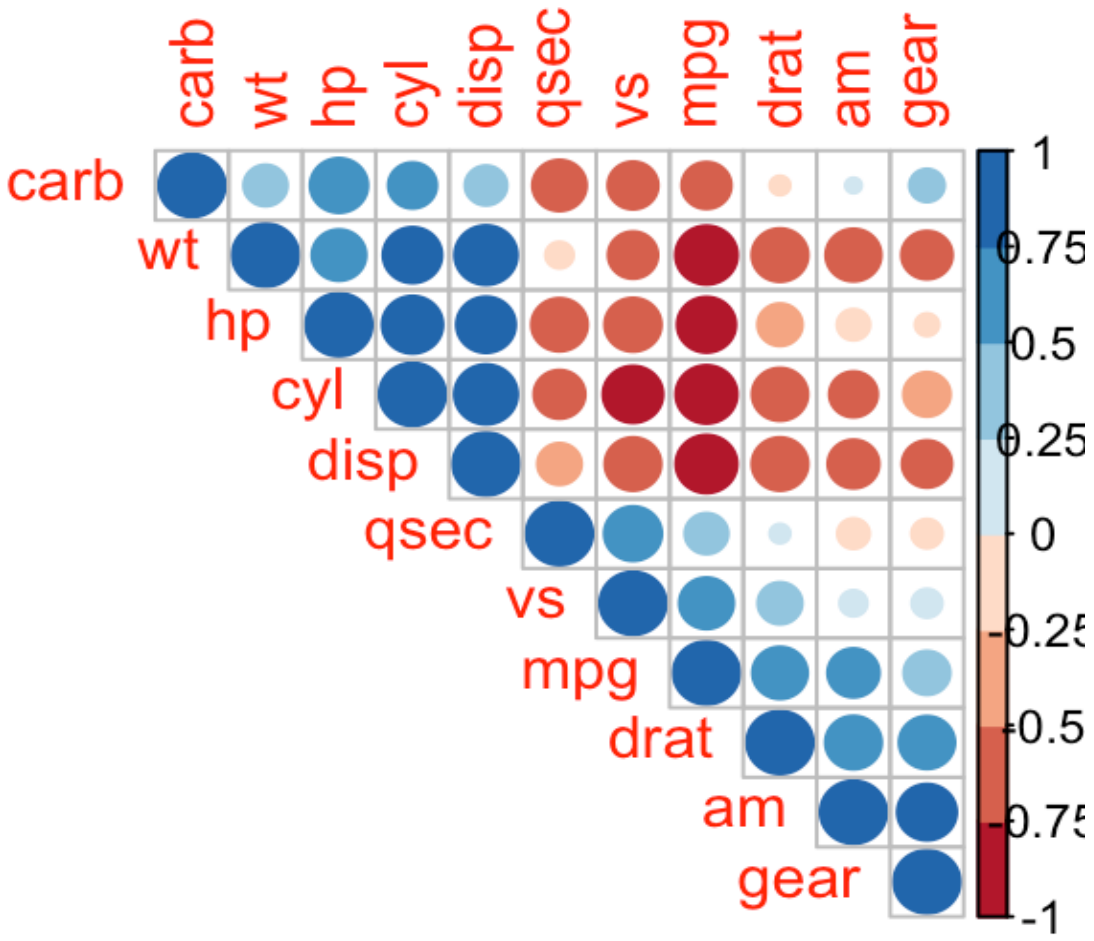


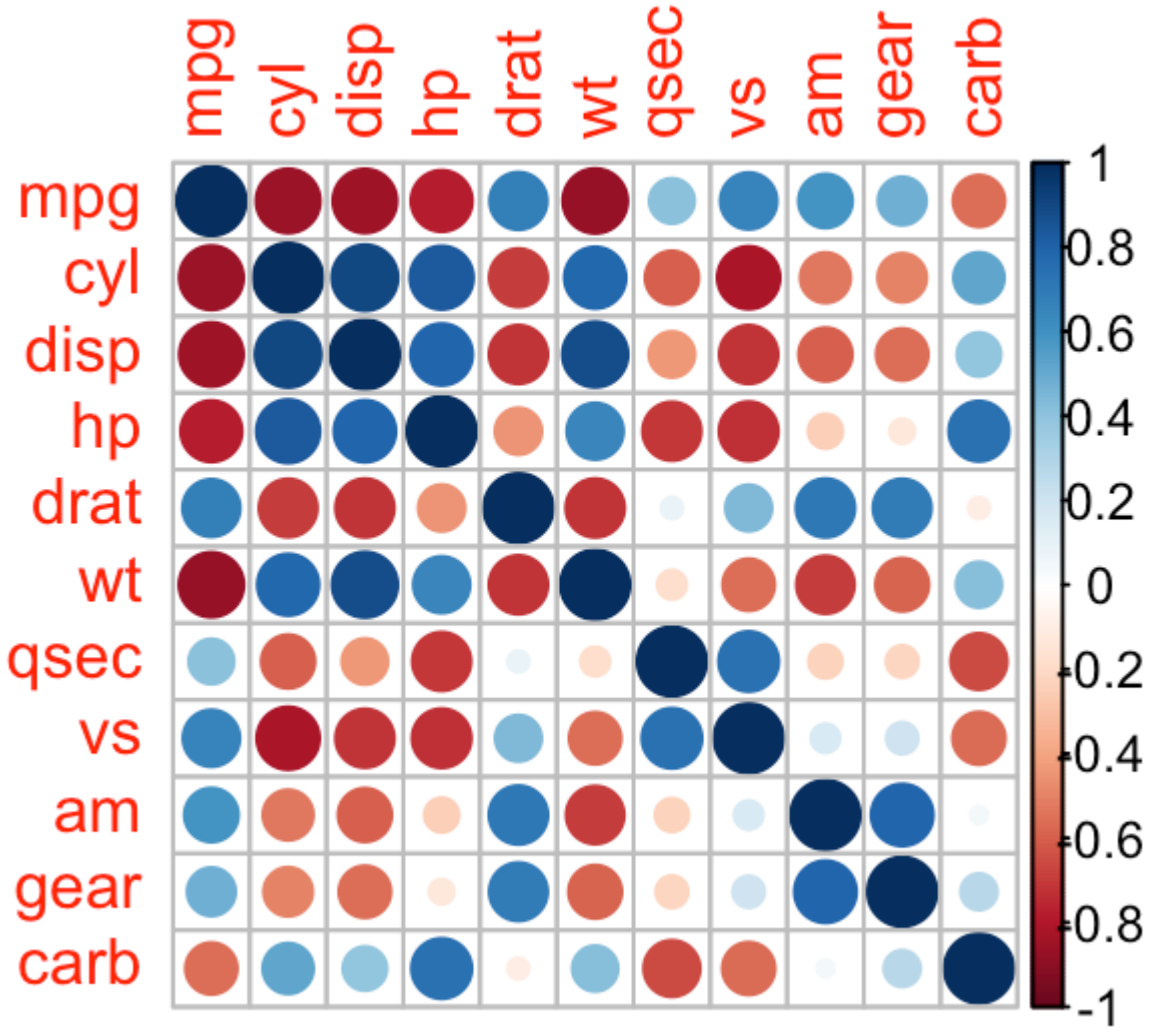
ALTE TIPURI DE REPREZENTĂRI GRAFICE PENTRU CORELAȚII



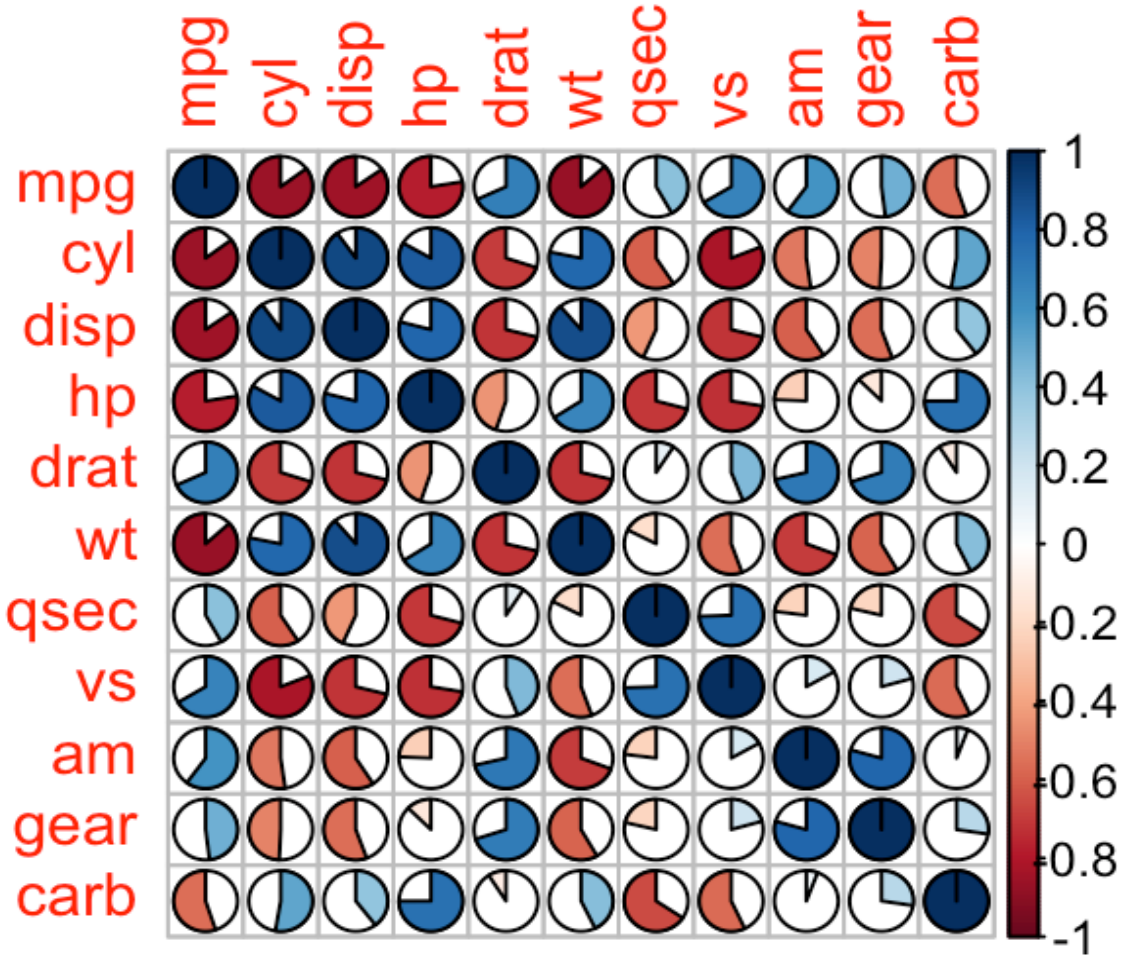


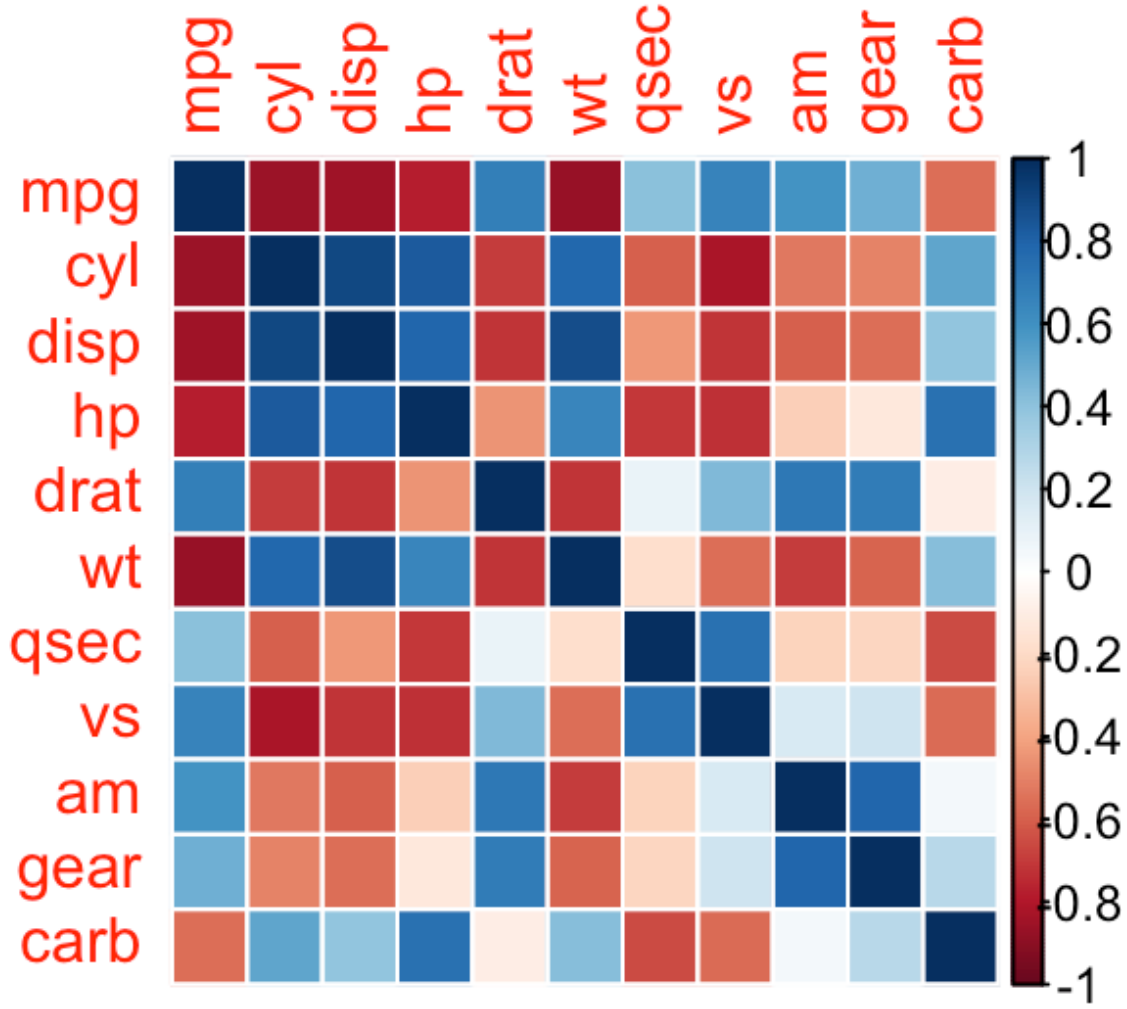
ALTE TIPURI DE REPREZENTĂRI GRAFICE PENTRU CORELAȚII





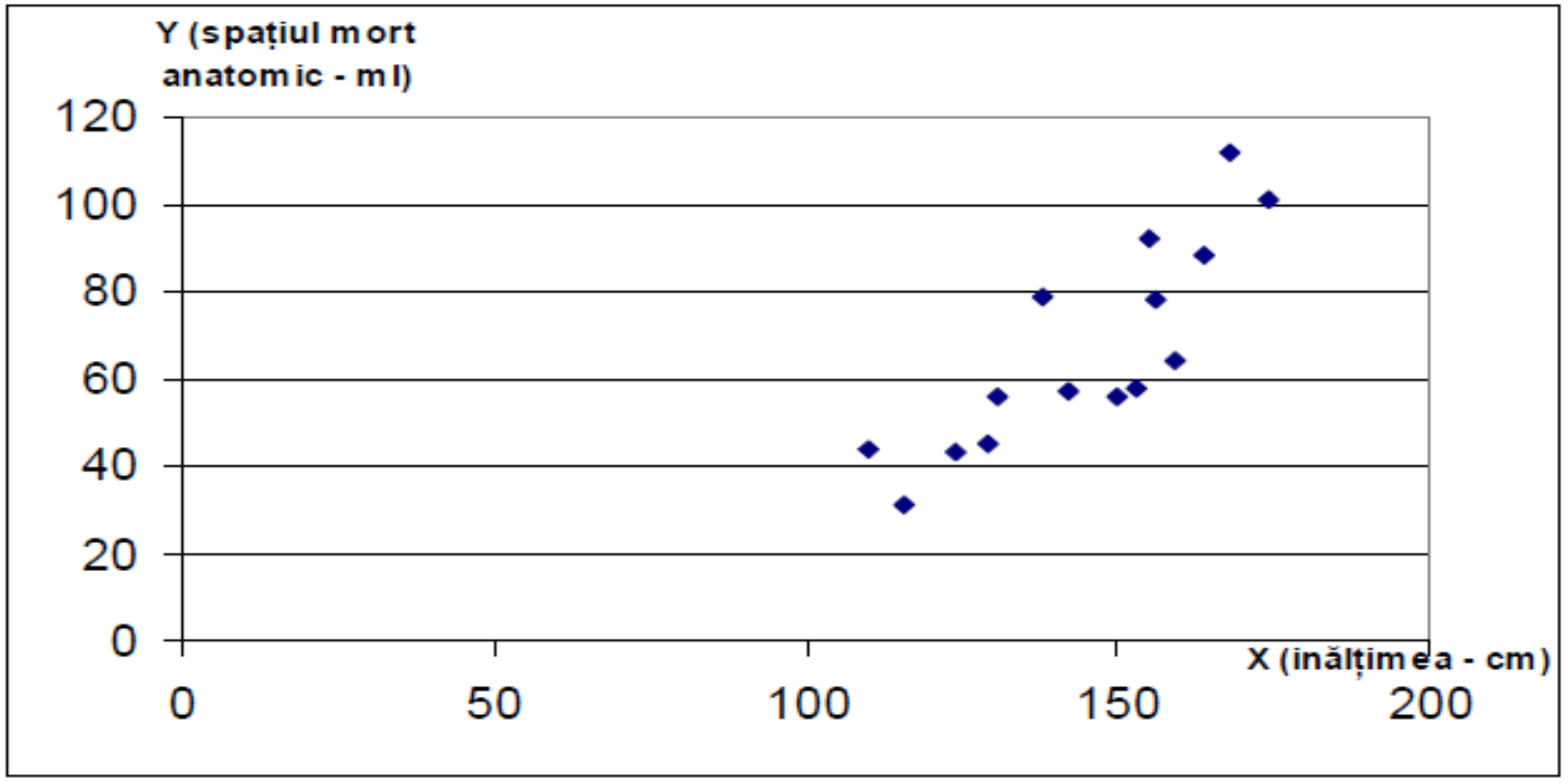
ALTE TIPURI DE REPREZENTĂRI GRAFICE PENTRU CORELAȚII





6. EXEMPLU NUMERIC

Nr.crt. subiect	**Înălțimea (cm)** – variabila independentă	**Spațiul pulmonar mort anatomic** – variabila dependentă
1	110	44
2	116	31
3	124	43
4	129	45
5	131	56
6	138	79
7	142	57
8	150	56
9	153	58
10	155	92
11	156	78
12	159	64
13	164	88
14	168	112
15	174	101
Statistică descriptivă (n=15)	$\bar{x} = 144{,}60$ $SD_x = 19{,}37$	$\bar{y} = 66{,}93$ $SD_y = 23{,}65$



$$r = \frac{150605 - (15 * 144{,}60 * 66{,}93)}{14 * 19{,}37 * 23{,}65} = \frac{5426{,}6}{6412{,}06} = 0{,}846$$

Coeficientul de corelaţie pare să indice o corelaţie pozitivă puternică între mărimea spaţiului mort anatomic şi înălţimea copiilor.

Dar în interpretarea corelaţiei este important să ne amintim că existenţa unei corelaţii între două variabile nu implică în mod necesar cauzalitatea, aceasta se poate datora unor cauze comune. Prin urmare trebuie avut grijă la interpretarea acestor coeficienţi de corelaţie.

7. COEFICIENTUL DE DETERMINARE

O parte a variaţiilor valorilor măsurate în cazul *variabile dependente* (exprimate cu ajutorul *varianţei*, mărime calculată în cadrul analizei statistice descriptive) se pot datora într-adevăr *existenţei unei (co)relaţii cu variabila independentă*, pe când o altă parte se *datorează unor cauze nedeterminate* (adesea aleatorii).

De aceea avem nevoie de o mărime care să cuantifice cât din această varianţă a variabilei dependente se datorează influenţei variabilei independente.

Această mărime se numeşte **coeficient de determinare** şi este egal cu **r2**. Pentru exemplul studiat anterior, **r2 = 0,716**, astfel că putem afirma faptul că aproximativ 72% din variaţia existentă între volumul spaţiului mort anatomic la lotul de copii studiat se datorează variaţiilor înălţimii acestora.

Practic **coeficientul de determinare r2** este extrem de util deoarece este o măsură a procentului variaţiei ce poate fi "explicată" din totalul variaţiei observate

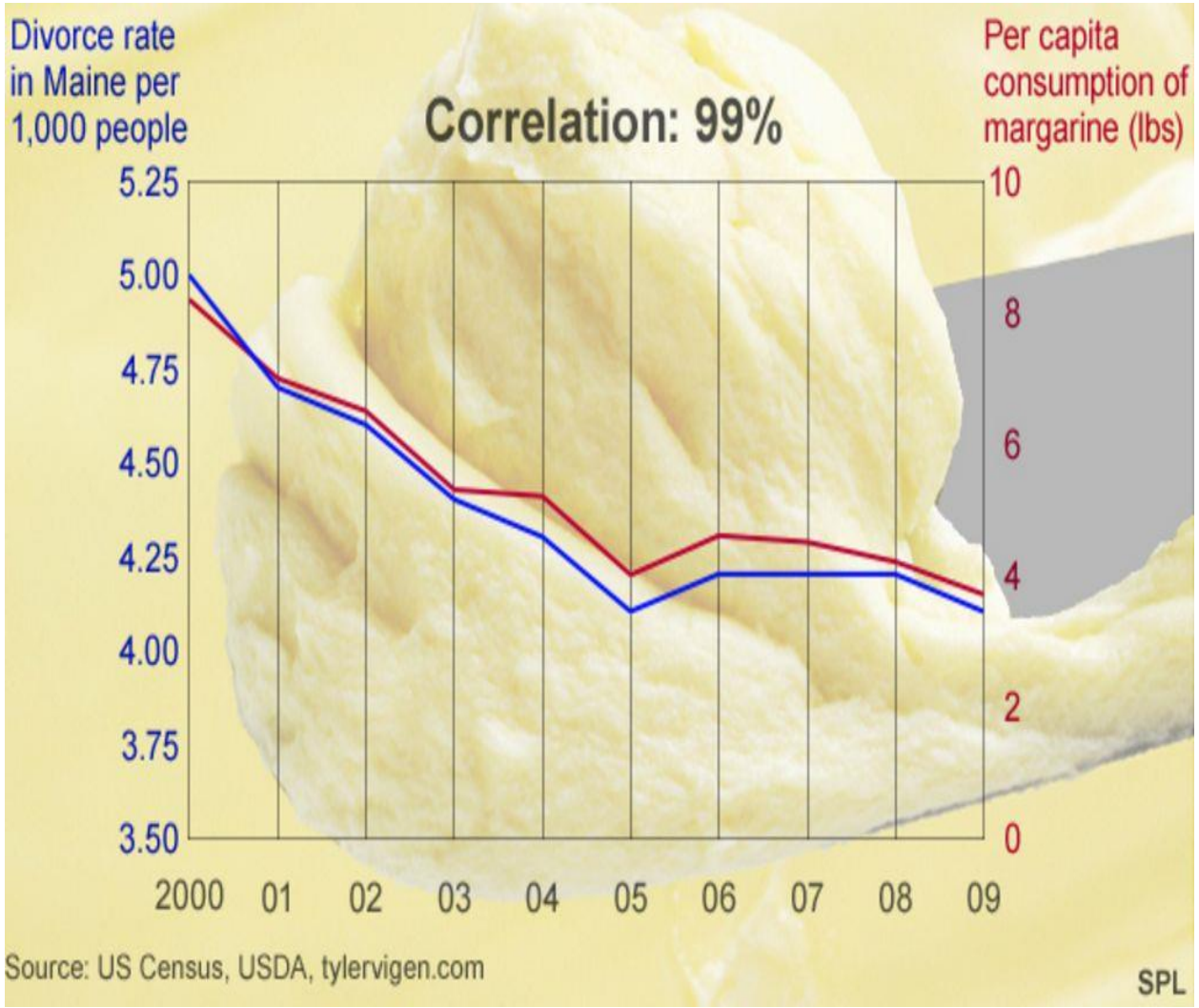
Coeficientul de determinare poate avea valori cuprinse între 0 şi 1 (0<r2<1).

8. “FUNNY CORRELATIONS”

NOTĂ: Alegerea variabilelor dependente și independente trebuie făcută cu precauție, deoarece putem să greșim ușor datorită unor factori de confuzie (De exemplu o a treia variabilă care le poate influența pe amândouă).

Astfel, este potrivit să presupunem că înălțimea unui lot de copii (variabila dependentă) este corelată pozitiv ($r>0$) cu vârsta acestora (variabila independentă).
Pe de altă parte, am putea constata existența unei corelații Negative ($r<0$) între numărul de cazuri de infarct miocardic (variabila "Dependentă") și consumul de înghețată (variabila "independentă"), când, De fapt, ambele variabile sunt influențate de o a treia, temperatura mediului Înconjurător, fără a avea o legătură directă una cu cealaltă. Numărul de Cazuri de infarct miocardic este corelat negativ, iar consumul de înghețată Corelat pozitiv cu creșterea temperaturii mediului înconjurător.

FUNNY CORRELATIONS



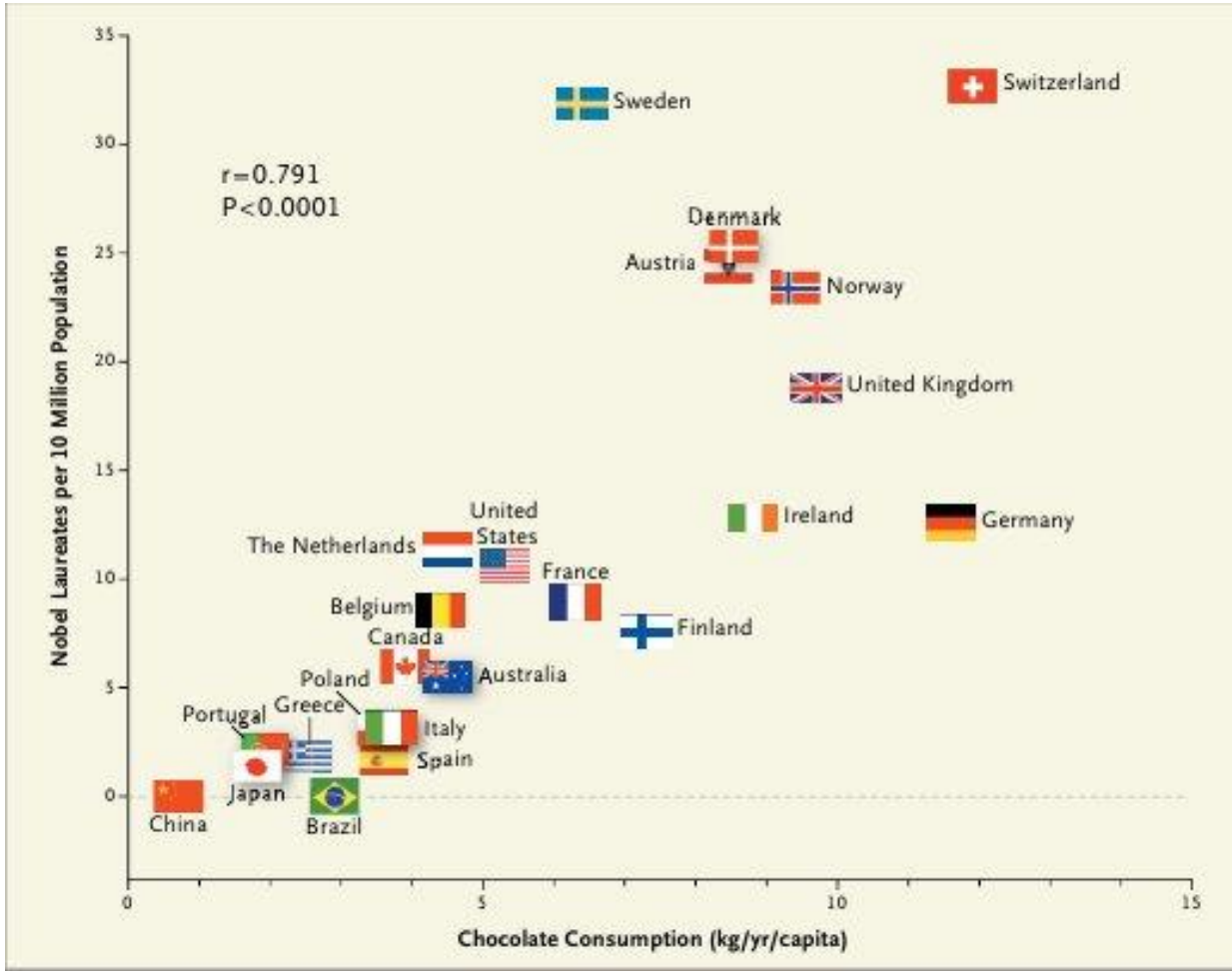
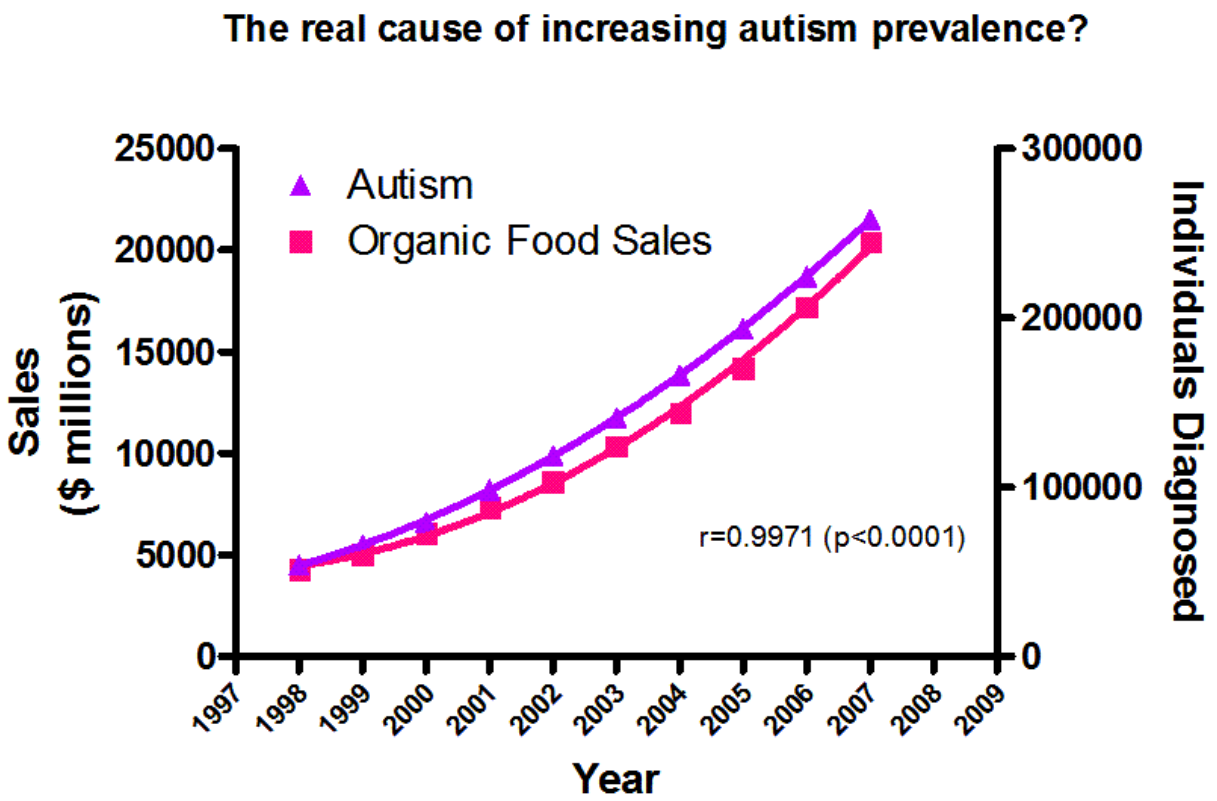
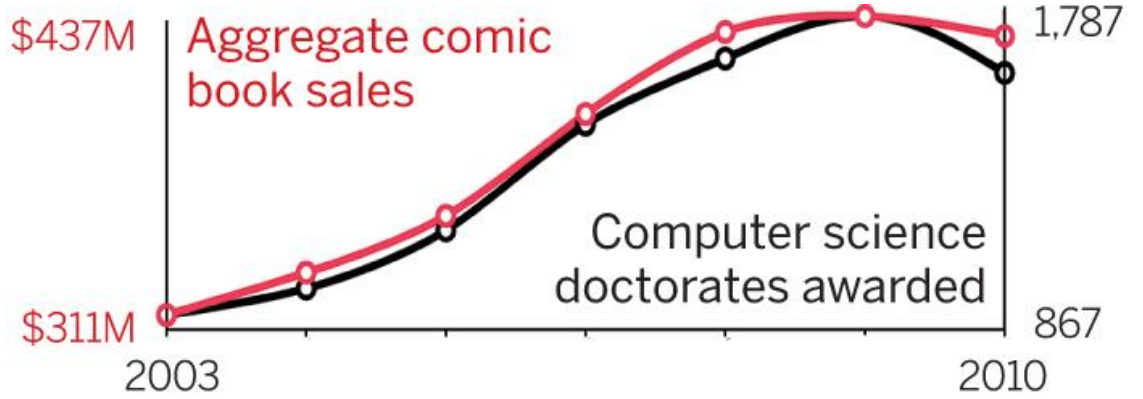


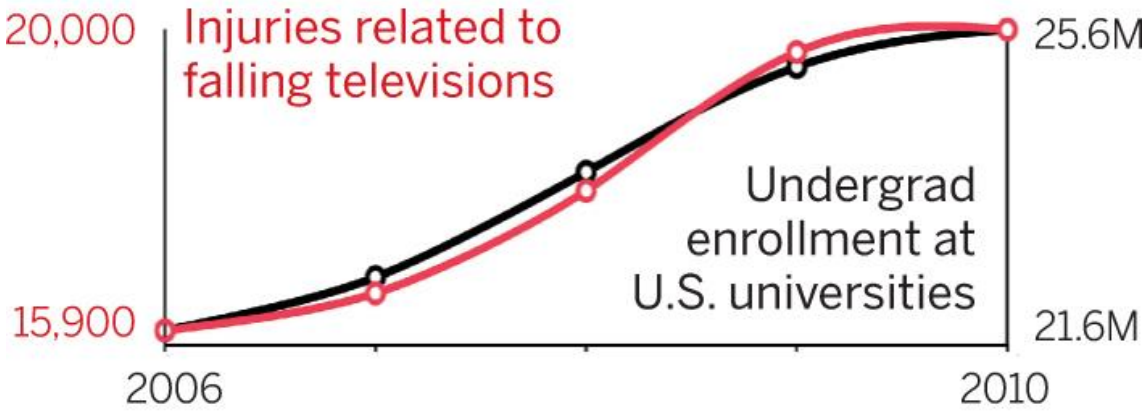
Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

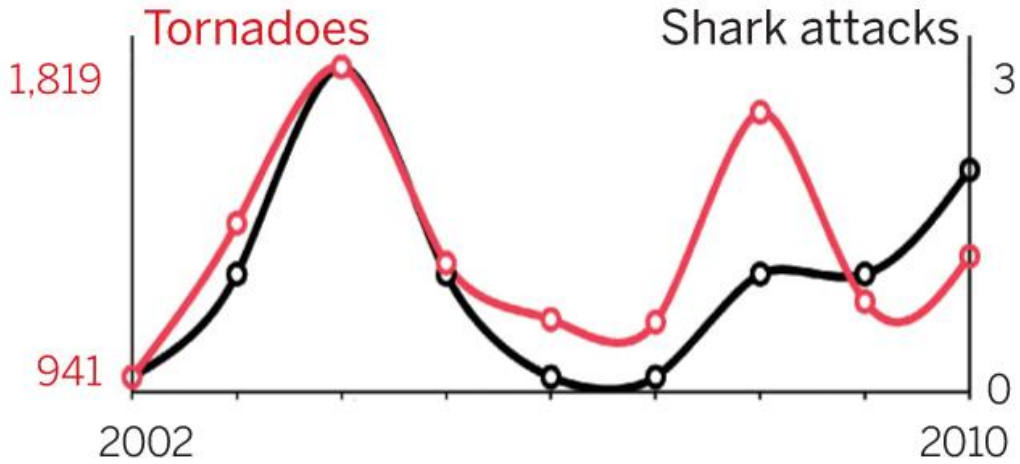
FUNNY CORRELATIONS



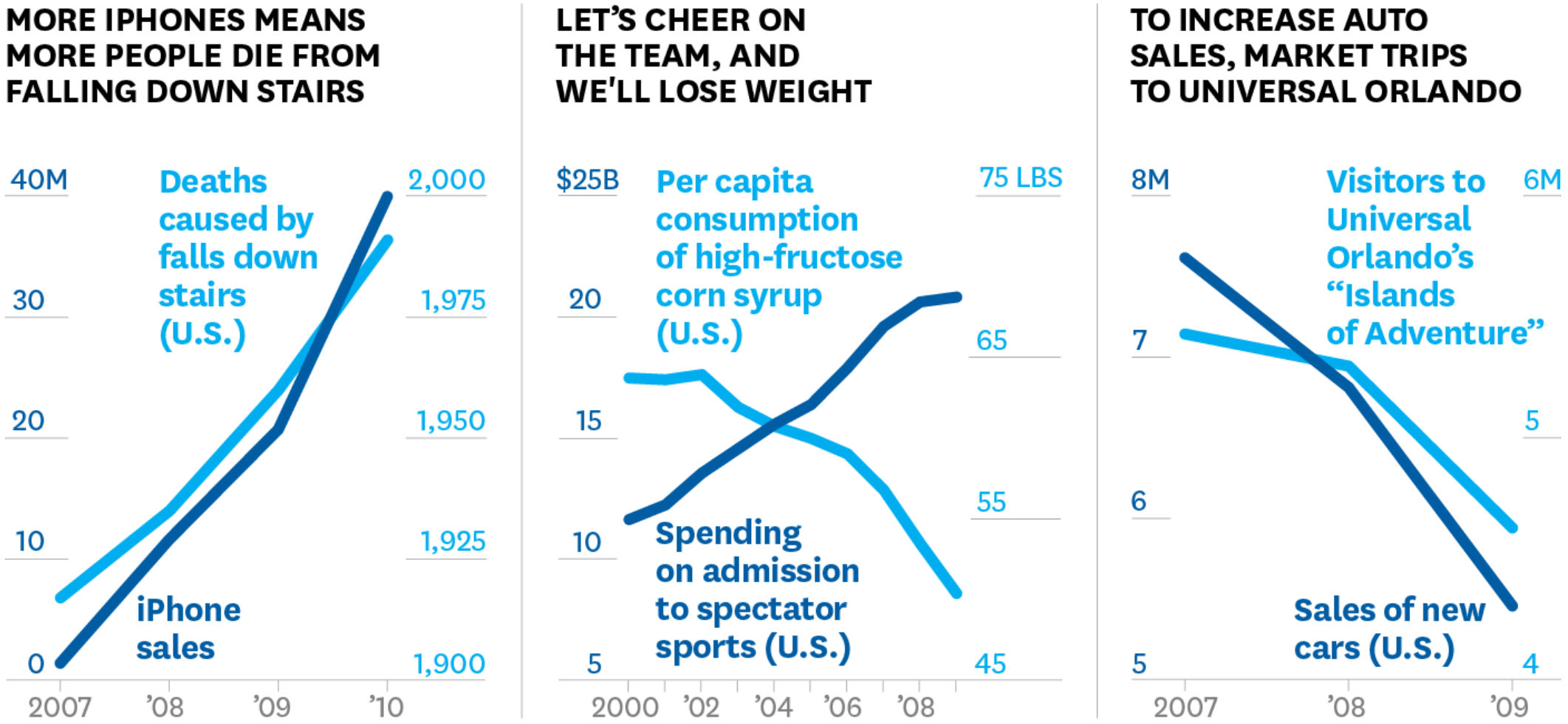
Sources: Organic Trade Association, 2011 Organic Industry Survey; U.S. Department of Education, Office of Special Education Programs, Data Analysis System (DANS), OMB# 1820-0043: "Children with Disabilities Receiving Special Education Under Part B of the Individuals with Disabilities Education Act





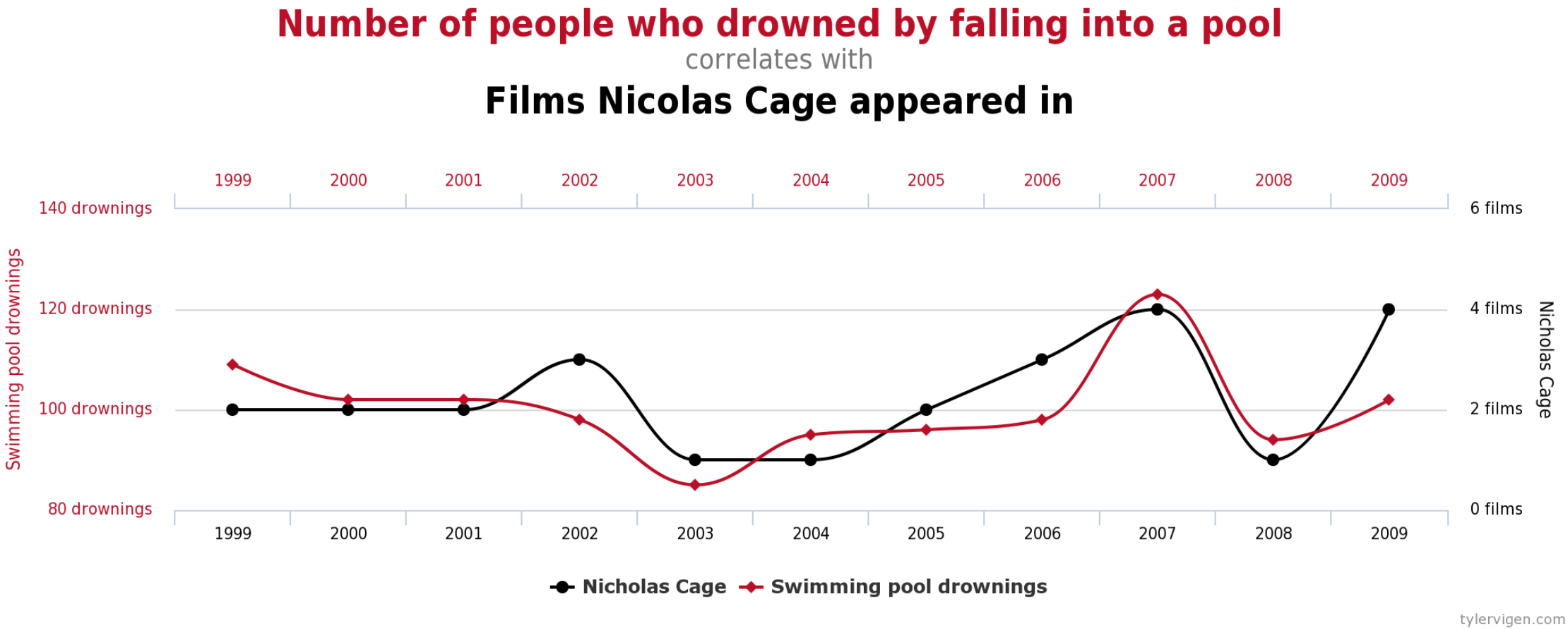


FUNNY CORRELATIONS

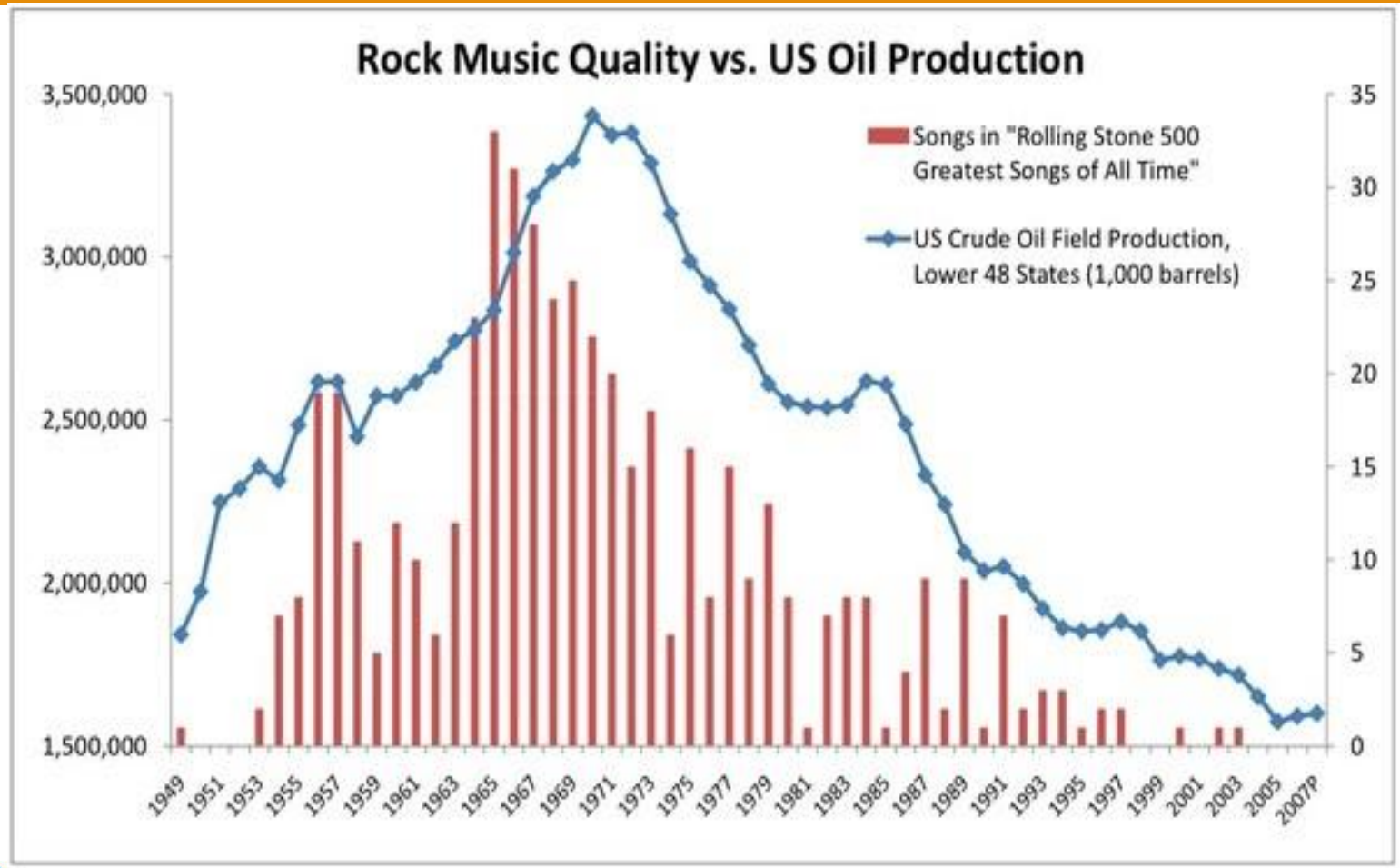


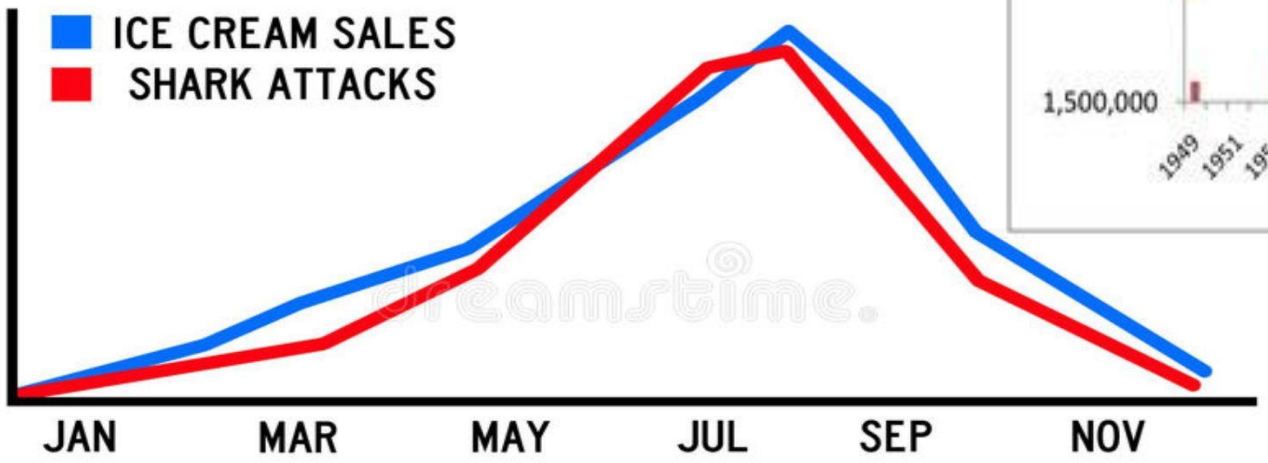
SOURCE TYLERVIGEN.COM
FROM "BEWARE SPURIOUS CORRELATIONS," JUNE 2015

FUNNY CORRELATIONS



FUNNY CORRELATIONS





Both ice cream sales and shark attacks increase when the weather is hot and sunny, but they are not caused by each other (they are caused by good weather, with lots of people at the beach, both eating ice cream and having a swim in the sea)