## Statisztika

### Politológus képzés

#### Daróczi Gergely

Politológia Tanszék

2012. május 8.



### Outline

- Mintaválasztás (ismétlés)
- A változók közötti kapcsolatról
- Korreláció
  - Elméleti háttér
  - Gyakorlat
  - A korrelációs együttható korlátairól
  - Gyakorlat
- Mereszttábla
  - Elméleti háttér
  - Simpson paradoxon
- 5 Standardizálás és dekompozíció
- 6 Grafikonok

#### A mintaválasztás

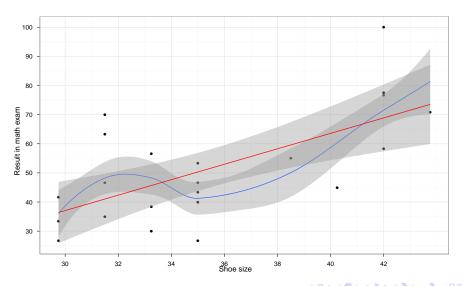
Valószínűségi vs. nem-valószínűségi mintavétel

	Nő	Férfi	Σ
Elméleti matematika Környezettudomány	10	10	20
Környezettudomány	40	10	50
Rendezvényszervező	10	20	30
Σ	60	40	100

Okos diákok nagy cipőben (példa)

Egy mini-kutatást végeztünk a diákok cipőmérete és matematika felkészültségéről. A következő eredményeket kaptuk:

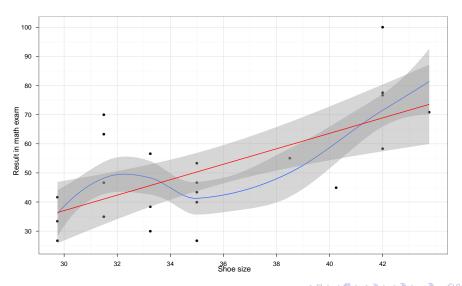
	Cipőméret	Matematika eredmény
1	29.75	26.67
2	29.75	33.33
3	29.75	41.67
4	31.50	35.00
5	31.50	46.67
6	31.50	63.33
7	31.50	70.00
8	33.25	30.00
9	33.25	38.33
10	33.25	56.67
11	35.00	26.67
12	35.00	40.00
13	35.00	43.33
14	35.00	46.67
15	35.00	53.33
16	38.50	55.00
17	40.25	45.00
18	42.00	58.33
19	42.00	76.67
20	42.00	77.50
21	42.00	100.00
22	43.75	70.83

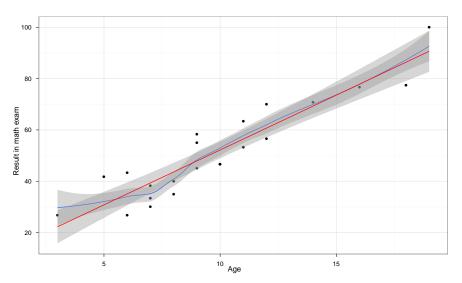


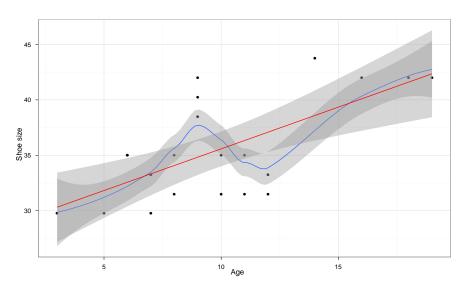
Okos diákok nagy cipőben (példa)

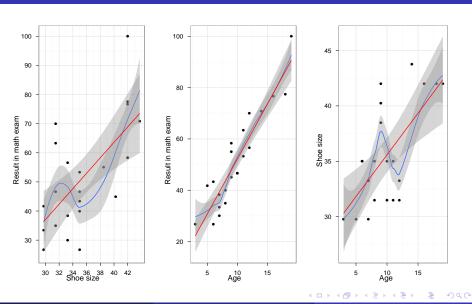
Egy mini-kutatást végeztünk a diákok cipőmérete, matematika felkészültségéről és **életkoráról**. A következő eredményeket kaptuk:

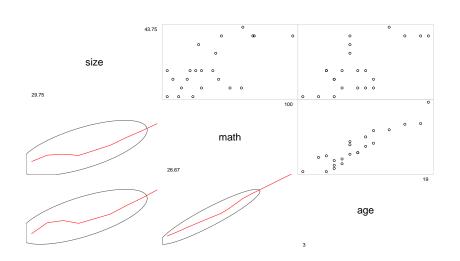
	Cipőméret	Matematika eredmény	Age
1	29.75	26.67	3
2	29.75	33.33	7
3	29.75	41.67	5
4	31.50	35.00	8
5	31.50	46.67	10
6	31.50	63.33	11
7	31.50	70.00	12
8	33.25	30.00	7
9	33.25	38.33	7
10	33.25	56.67	12
11	35.00	26.67	6
12	35.00	40.00	8
13	35.00	43.33	6
14	35.00	46.67	10
15	35.00	53.33	11
16	38.50	55.00	9
17	40.25	45.00	9
18	42.00	58.33	9
19	42.00	76.67	16
20	42.00	77.50	18
21	42.00	100.00	19
22	43.75	70.83	14

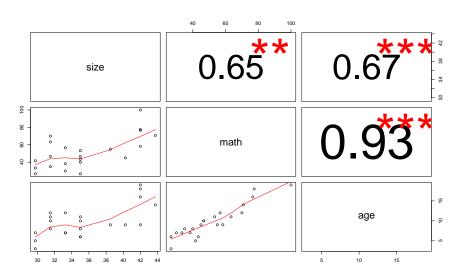












Okos diákok nagy cipőben (példa)

# Parciális korreláció:

$$r_{matek,cipo\cdot kor} = 0.11$$

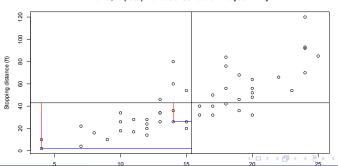
$$r_{matek,kor\cdot cipo} = 0.87$$

$$r_{cipo,kor\cdot matek} = 0.22$$

x és y változók együttes szórása:

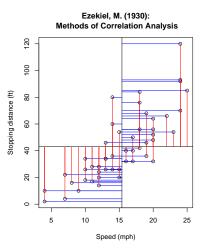
$$COV(xy) = \sum_{i=1}^{n} \frac{(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$
 $emlkeztet: \sigma = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n}}$ 

Ezekiel, M. (1930) Methods of Correlation Analysis. Wiley.

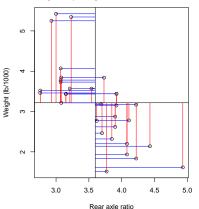


#### Elméleti háttér

#### Kovariancia



Henderson & Velleman (1981): Building multiple regression models interactively



$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$
1.0 0.8 0.4 0.0 -0.4 -0.8 -1.0
1.0 1.0 1.0 -1.0 -1.0 -1.0

$$\hat{r}_{XY \cdot \mathbf{Z}} = \frac{N \sum_{i=1}^{N} r_{X,i} r_{Y,i} - \sum_{i=1}^{N} r_{X,i} \sum_{i=1}^{N} r_{Y,i}}{\sqrt{N \sum_{i=1}^{N} r_{X,i}^2 - \left(\sum_{i=1}^{N} r_{X,i}\right)^2}} \sqrt{N \sum_{i=1}^{N} r_{Y,i}^2 - \left(\sum_{i=1}^{N} r_{Y,i}\right)^2}$$

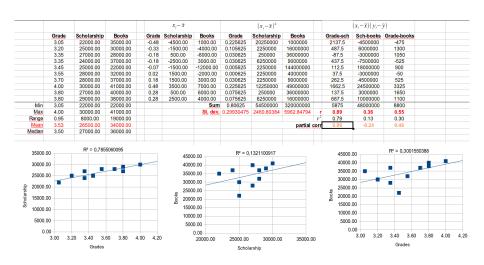
három változó esetén:

$$\hat{r}_{XY \cdot \mathbf{Z}} = \frac{r_{XY} - r_{X\mathbf{Z}}r_{Y\mathbf{Z}}}{\sqrt{(1 - r_{X\mathbf{Z}}^2)(1 - r_{Y\mathbf{Z}}^2)}}$$

- Mit takar a korreláció és parciális korreláció kifejezés?
- Határozza meg a korrelációs együtthatót az alábbi változó-párok esetében!
- Mennyiben különbözik a parciális korreláció értéke?

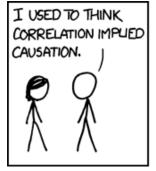
Érdemjegy (átlag)	Ösztöndíj (HUF)	Kiadás könyvekre (HUF)
3.05	22000	3500
3.2	25000	3000
3.35	27000	2800
3.35	24000	3700
3.45	25000	2200
3.55	28000	3200
3.7	28000	3700
45	30000	4100
3.8	27000	4000
3.8	29000	3800

## Megoldás

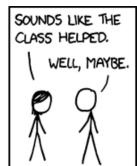


- Korreláció és linearitás
- Korreláció és kauzalitás
- Lazarsfeld paradigma

Correlation does not imply causation!







Forrás: http://xkcd.com/552

Correlation does not imply causation! - Elméleti háttér

Arisztotelész: logika, szillogizmus – if  $(A o B)\&(B o C) \Rightarrow A o C$ 

David Hume: szkepticizmus

- "only correlation can actually be perceived [not causality]"
- I. holnap vajon felkel a nap?
- I. "If I see a billiard ball moving towards another, on a smooth table, I can easily conceive to stop upon contact."

Popper: falszifikáció

Pearl, J. - *Causality: Models, Reasoning, and Inference*, Cambridge University Press, 2000

Lazarsfeld paradigma

Stouffer: The American Soldier

Soldiers in branches with higher promotion rates are happier than soldiers in branches with lower rates of promotion.

Lazarsfeld paradigma

Stouffer: The American Soldier

 $H_0$ : Soldiers in branches with higher promotion rates are happier than soldiers in branches with lower rates of promotion. **Ámde:** 

"Soldiers in branches with higher promotion rates were more pessimistic about their own chances of being promoted than soldiers in branches with lower rates of promotion."

Lazarsfeld paradigma

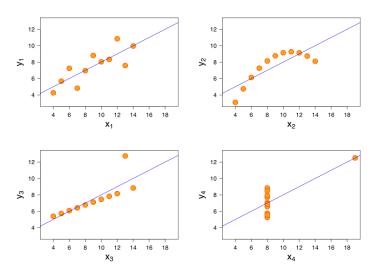
Stouffer: The American Soldier

 $H_0$ : Soldiers in branches with higher promotion rates are happier than soldiers in branches with lower rates of promotion. **Ámde:** 

"Soldiers in branches with higher promotion rates were more pessimistic about their own chances of being promoted than soldiers in branches with lower rates of promotion."

Kulcsszavak: referencia csoport, relatív depriváció

Linearitás

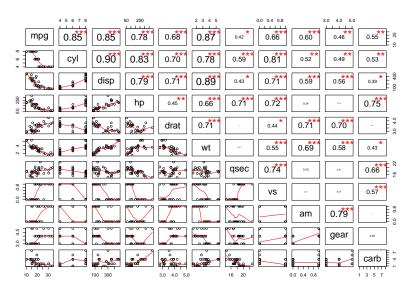


Forrás: Anscombe, F. J. (1973) Graphs in statistical analysis. American Statistician 27, 17-21 ...

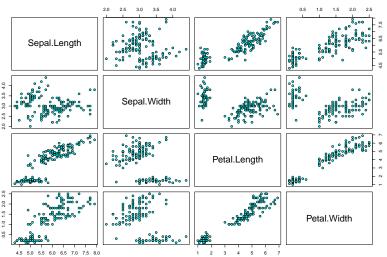
The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

- mpg: Miles/(US) gallon
- cyl: Number of cylinders
- disp: Displacement (cu.in.)
- hp: Gross horsepower
- drat: Rear axle ratio
- wt: Weight (lb/1000)
- qsec: 1/4 mile time
- vs: V/S
- am: Transmission (0 = automatic, 1 = manual)
- gear: Number of forward gears
- carb: Number of carburetors

Source: Henderson and Velleman (1981), Building multiple regression models interactively. Biometrics, 37, 391-411.



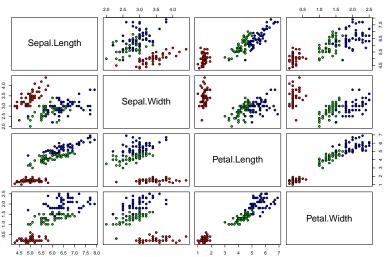
#### Edgar Anderson's Iris Data



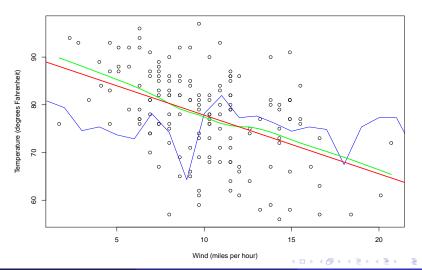
Anderson, Edgar (1935). The irises of the Gaspe Peninsula, Bulletin of the American Iris Society, 59, 2-5.

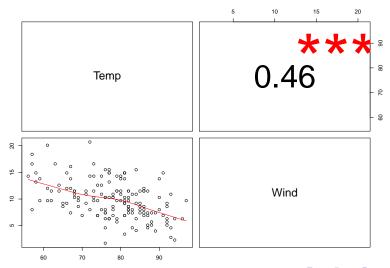
## Gyakorlat #2

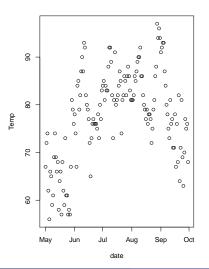
#### Edgar Anderson's Iris Data

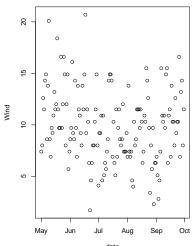


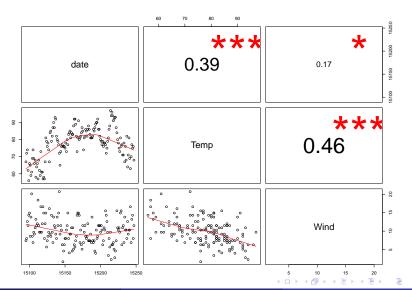
Anderson, Edgar (1935). The irises of the Gaspe Peninsula, Bulletin of the American Iris Society, 59, 2-5.











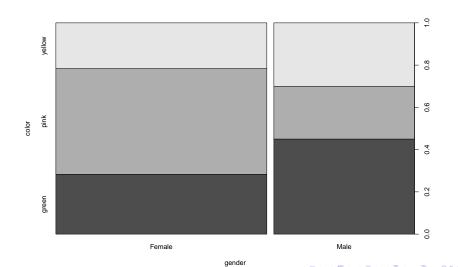
#### Kereszttábla

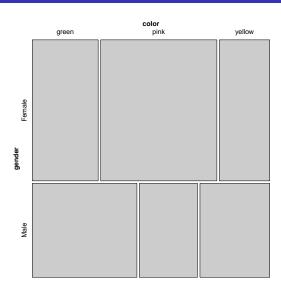
Alacsony mérési szintű (kvalitatív) változók

szín

### Kereszttábla

Alacsony mérési szintű (kvalitatív) változók





	zöld	piros	sárga
nők	17	30	13
férfiak	18	10	12

	zöld	piros	sárga	
nők	17	30	13	Marginals
férfiak	18	10	12	iviaryiriais
	Marginals			N

	zöld	piros	sárga	Σ
nők	17	30	13	60
férfiak	18	10	12	40
Σ	35	40	25	100

#### Százalékok

	zöld	piros	sárga	Σ
nők	17	30	13	60
férfiak	18	10	12	40
Σ	35	40	25	100

1. táblázat. Tapasztalt értékek

	zöld	piros	sárga	Σ
nők	17 %	30 %	13 %	60 %
férfiak	18 %	10 %	12 %	40 %
Σ	35 %	40 %	25 %	100 %

2. táblázat. Teljes százalék

#### Sorszázalék

	zöld	piros	sárga	Σ
nők	17	30	13	60
férfiak	18	10	12	40
Σ	35	40	25	100

### 3. táblázat. Tapasztalt értékek

	zöld	piros	sárga	Σ
nők	28.3 %	50 %	21.7 %	100 %
férfiak	45 %	25 %	30 %	100 %
Σ	35 %	40 %	25 %	100 %

4. táblázat. Sorszázalék

#### Oszlopszázalék

	zöld	piros	sárga	Σ
nők	17	30	13	60
férfiak	18	10	12	40
Σ	35	40	25	100

5. táblázat. Tapasztalt értékek

	zöld	piros	sárga	Σ
nők	48.63 %	75 %	52 %	60 %
férfiak	51.4 %	25 %	48 %	40 %
Σ	100 %	100 %	100 %	100 %

6. táblázat. Oszlopszázalék

#### Várható érték

	zöld	piros	sárga	Σ
nők	17	30	13	60
férfiak	18	10	12	40
Σ	35	40	25	100

7. táblázat. Tapasztalt érték

	zöld	piros	sárga	Σ
nők	21	24	15	60
férfiak	14	16	10	40
Σ	35	40	25	100

8. táblázat. Várható érték

Khí-négyzet statisztika

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

#### where:

- $\chi^2$ : Pearson-féle teszt statisztika,
- O<sub>i</sub>: tapasztalt érték,
- *E<sub>i</sub>*: várható éréték,
- n: cellák száma.

H<sub>0</sub>: a tapasztalt és a várható érték megegyezik

Követelmények?



#### Khí-négyzet

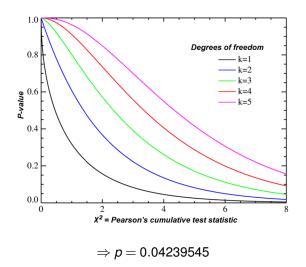
	zöld	piros	sárga	Σ
nők	$\frac{(17-21)^2}{21}$	$\frac{(30-24)^2}{24}$	$\frac{(13-15)^2}{15}$	-
férfiak	$\frac{(18-14)^2}{14}$	$\frac{(10-16)^2}{16}$	$\frac{(12-10)^2}{10}$	-
Σ	-	-	-	-

9. táblázat. Számított távolság a várt és tapasztalt értékek között

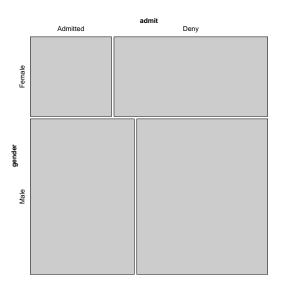
$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} = 6.321429$$

szabadságfok: 
$$(3-1)(2-1) = 2$$

#### Khí-négyzet



A Berkeley egyetem esete (Bickel et al.)



A Berkeley egyetem esete (Bickel et al.)

	Felvett	Elutasított	Σ
nők	1494	2827	4321
férfiak	3738	4704	8442
Σ	5232	7531	12763

10. táblázat. Tapasztalt értékek

A Berkeley egyetem esete (Bickel et al.)

	Felvett	Elutasított	Σ
nők	1494	2827	4321
férfiak	3738	4704	8442
Σ	5232	7531	12763

10. táblázat. Tapasztalt értékek

	felvett	elutasított	Σ
nők	34.6 %	65.4 %	100 %
férfiak	44.3 %	55.7 %	100 %
Σ	41 %	59 %	100 %

11. táblázat. Sorszázalék

A Berkeley egyetem esete (Bickel et al.)

	Felvett	Elutasított	Σ
nők	1494	2827	4321
férfiak	3738	4704	8442
Σ	5232	7531	12763

10. táblázat. Tapasztalt értékek

	felvett	elutasított	Σ
nők	34.6 %	65.4 %	100 %
férfiak	44.3 %	55.7 %	100 %
Σ	41 %	59 %	100 %

11. táblázat. Sorszázalék

 $\chi^2 = 110.8489$ ; d.f. = 1; p = 6.385628e - 26

A Berkeley egyetem esete (Bickel et al.)

	Jelentkezők	Felvettek száma
férfiak	8442	44%
nők	4321	35%

	férfiak	(	nők	
szak	jelentkezők	felvett	jelentkezők	felvett
A	825	62%	108	82%
В	560	63%	25	68%
С	325	37%	593	34%
D	417	33%	375	35%
Е	191	28%	393	24%
F	272	6%	341	7%

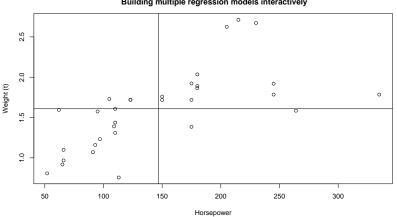
Baseball ütések

	1995		199	1996		Combined	
	Runs/Outs	%	Runs/Outs	%	Runs/Outs	%	
Derek Jeter	12/48	25 %	183/582	31.4 %	195/630	31 %	
David Justice	104/411	25.3 %	45/140	<b>32.1</b> %	149/551	27 %	

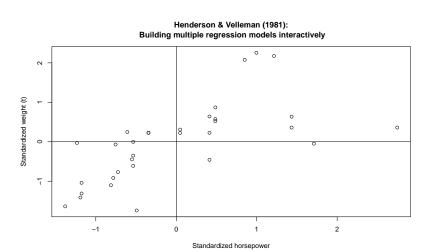
Melyikük a jobb játékos?

Egy egyszerű példa





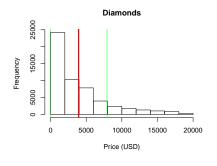
Egy egyszerű példa

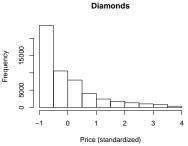


#### Elméleti háttér

Egy standardizált változó (z-values, z-scores, normal scores, standardized variables) azt mutatja, hogy hány szórásnyira esik az adott érték az átlagtól:

$$z=\frac{x-\mu}{\sigma}$$





Dekompozíció

Population and Deaths by Age in 1970 for White Females in Miami, Alaska, and the U.S.

		Miami			Alaska			U.S.	
Age	Pop.	Deaths	Rate*	Pop.	Deaths	Rate*	Pop.+	Deaths <sup>+</sup>	Rate*
< 15	114,350	136	1.19	37,164	59	1.59	23,961	32	1.34
15-24	80,259	57	0.71	20,036	18	0.90	15,420	9	0.58
25-44	133,440	208	1.56	32,693	37	1.13	21,353	30	1.40
45-64	142,670	1,016	7.12	14,947	90	6.02	19,609	140	7.14
65+	92,168	3,605	39.11	2,077	81	39.00	10,685	529	49.51
	562,887	5,022		106,917	285		91,028	740	
Crude death rate*			8.92			2.67			8.13

<sup>\*</sup> Deaths per 1,000 population

<sup>+</sup> in thousands

#### Csoportosított oszlopdiagram

#### Category wise Sales for 1996



#### Rétegzett oszlopdiagram

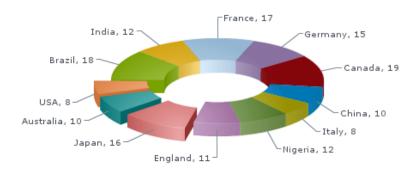




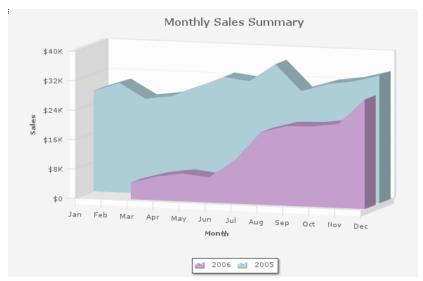
#### Vonaldiagram



#### Industrial Growth Rate (Country)



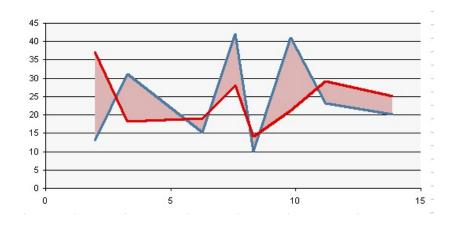
#### Terület



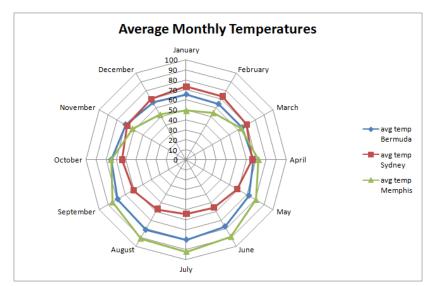
#### Összetett diagram



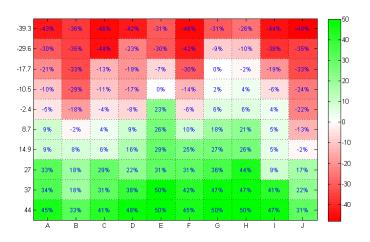
### Összetett diagram



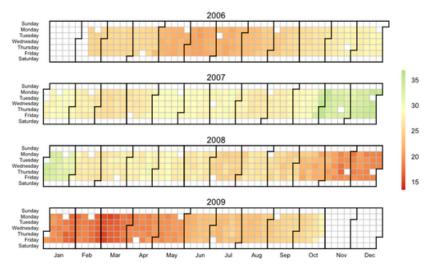
#### Poláris diagram



#### Heatmap



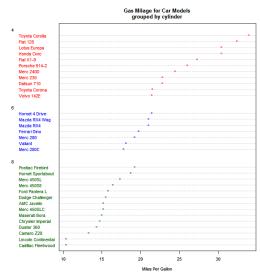
#### Calendar Heat Map of MSFT Adjusted Close



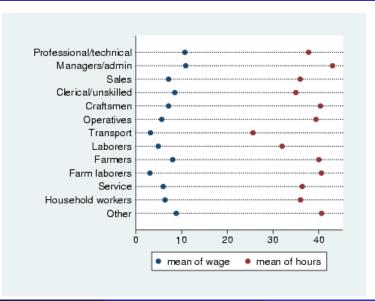
#### Waterfall

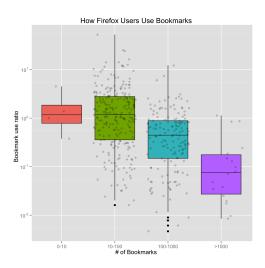


#### Dot plot

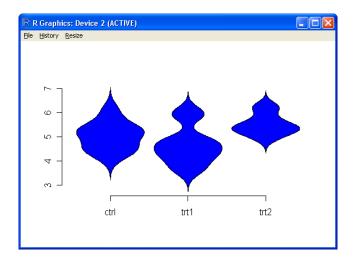


#### Dot plot

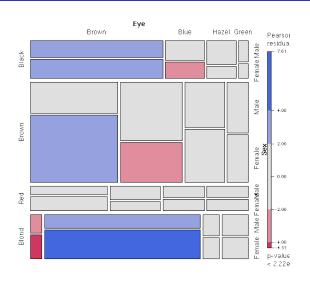




#### Violin plot



#### Mosaic chart

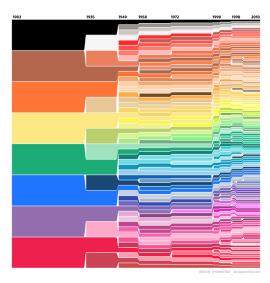


#### Szófelhő



60 / 63

"Crayola Color Chart, 1903-2010"



#### Érdekes honlapok

- http://www.visual-literacy.org/periodic\_table/periodic\_table.html
- http://www.edwardtufte.com/tufte/
- http://www.perceptualedge.com/
- http://www.visualcomplexity.com/vc/
- http://flowingdata.com/
- http://infosthetics.com/
- http://chartsgraphs.wordpress.com/
- http://www.informationisbeautiful.net/
- http://chartporn.org/

# Köszönöm a figyelmet!

Daróczi Gergely daroczi.gergely@btk.ppke.hu