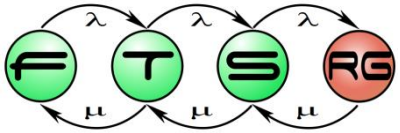
Leíró statisztika, EDA, vizualizáció

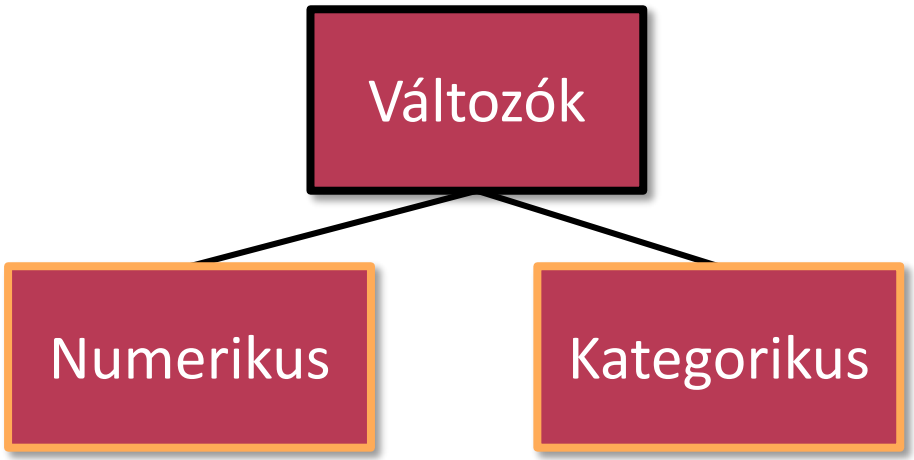
2017 ősz, 2./3. alkalom

Kocsis Imre, ikocsis@mit.bme.hu



Numerikus és kategorikus változók

- Numerikus (numerical)
  - az alapvető aritmetikai műveletek értelmesek



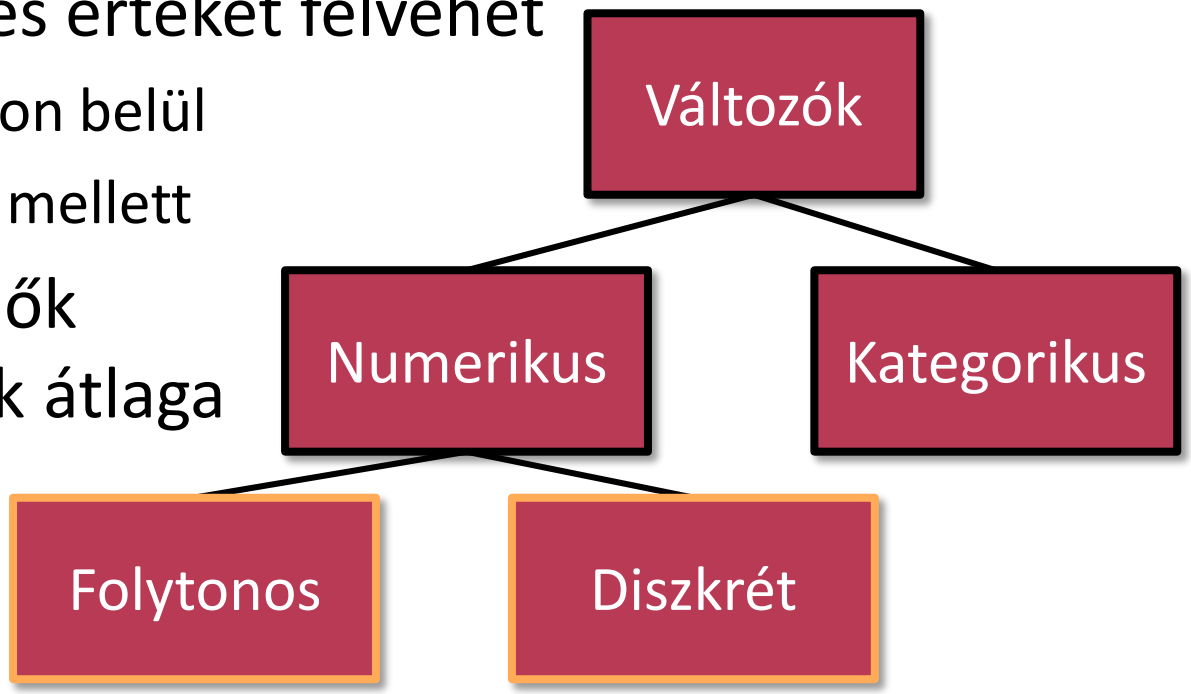
- Kategorikus (categorical)
  - Matematikai műveletek nem értelmezhetőek rajtuk, legfeljebb sorba rendezés

kor **átlaghőmérséklet**

telefonszám **nem**

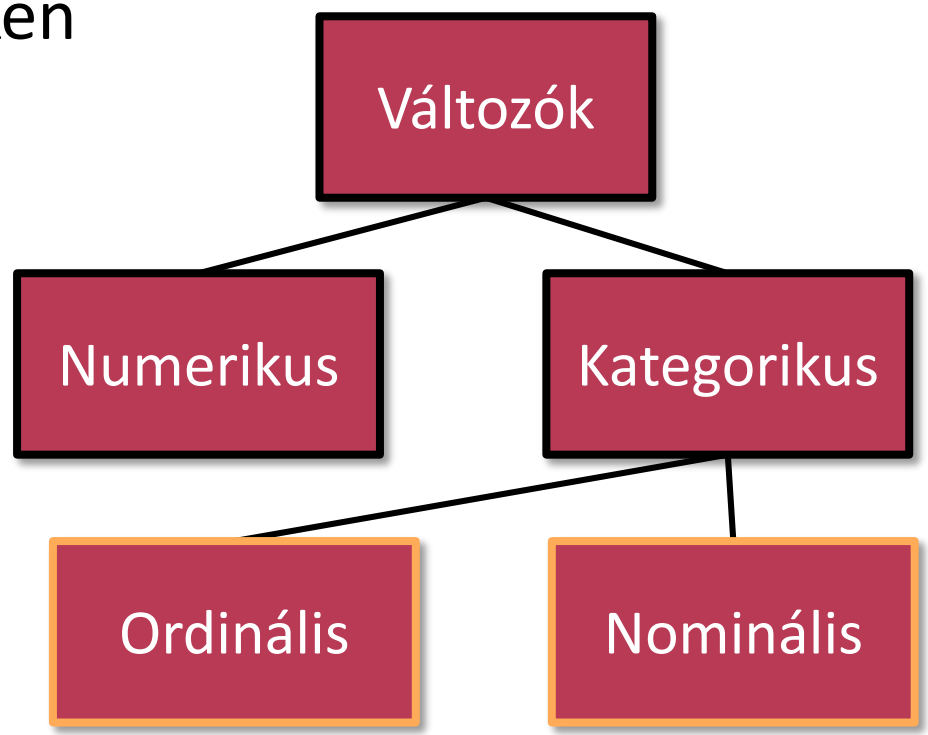
Numerikus változók

- Folytonos
  - Mért – tetszőleges értéket felvehet
    - adott tartományon belül
    - adott pontosság mellett
  - Pl. a teremben ülők ZH pontszámának átlaga
- Diszkrét
  - Számolt – véges sok értéket vehet fel adott tartományban
  - Pl. az előadáson ülők száma



Kategorikus változók

- Ordinális
  - Teljes rendezés az értékeken
  - Pl. szállodai csillagok
- Nominális



9. Ajánlanád-e a tárgyat másoknak?

- ○ Mindenkit rábeszélnék
- ○ Nyugodtan ajánlanám
- ○ Esetleg ajánlanám
- ○ Inkább lebeszélném róla
- ○ Feltétlenül lebeszélném
- ○ Nem kívánok válaszolni

Az “adatkeret” (data frame)

- **Táblázat sora** = megfigyelés/bejegyzés
- **Táblázat oszlopa** = tulajdonság

Név	Típus	Méret (kB)	Utolsó módosítás
Dokumentumok	könyvtár		2016.02.02
szerződés.pdf	fájl	569	2015.11.09
Képek	könyvtár		2016.02.02
logó.png	fájl	92	2015.03.06
alaprajz.jpg	fájl	1226	2016.02.02

- Adatelemzési eszközök (pl. R, Python): **data frame**
  - Egy sor egy mérés
  - Egyes oszlopoknak **típusai** vannak

Rövid demo: adatkeretek R-ben

Az “adatkeret”

- A legtöbb ált. célú adatelemző eszköz fókuszában
  - → kulcsrakész csomagok bemenete
- Denormalizált, ~~idegen kulcsok~~
  - “At rest” adaton ez nem baj
- SQL-től és CSV/TSV-től “egy lépésre”
  - + Excel…

- Egyéb nagy kategóriák
  - Strukturálatlan adat – pl. természetes szöveg
  - Szemi-strukturált adat – pl. XML
  - Más strukturált: gráfok, n-dim. mátrixok, …

“Széles” és “szűk”/”hosszú” adat

Person	Age	Weight
Bob	32	128
Alice	24	86
Steve	64	95

Person	Variable	Value
Bob	Age	32
Bob	Weight	128
Alice	Age	24
Alice	Weight	86
Steve	Age	64
Steve	Weight	95

Alapvető leíró statisztikák

Leíró statisztika

- Vizsgált adatok alapvető jellemzői
  - Kvantitatív
  - Erősen absztrahál, „összefoglal”

- Egyfajta ellentéte: következtető (*inferential*) stat.
  - Megfigyelt mintán túlmutató következtetések
  - Pl. populáció tulajdonságaira következtetés mintából

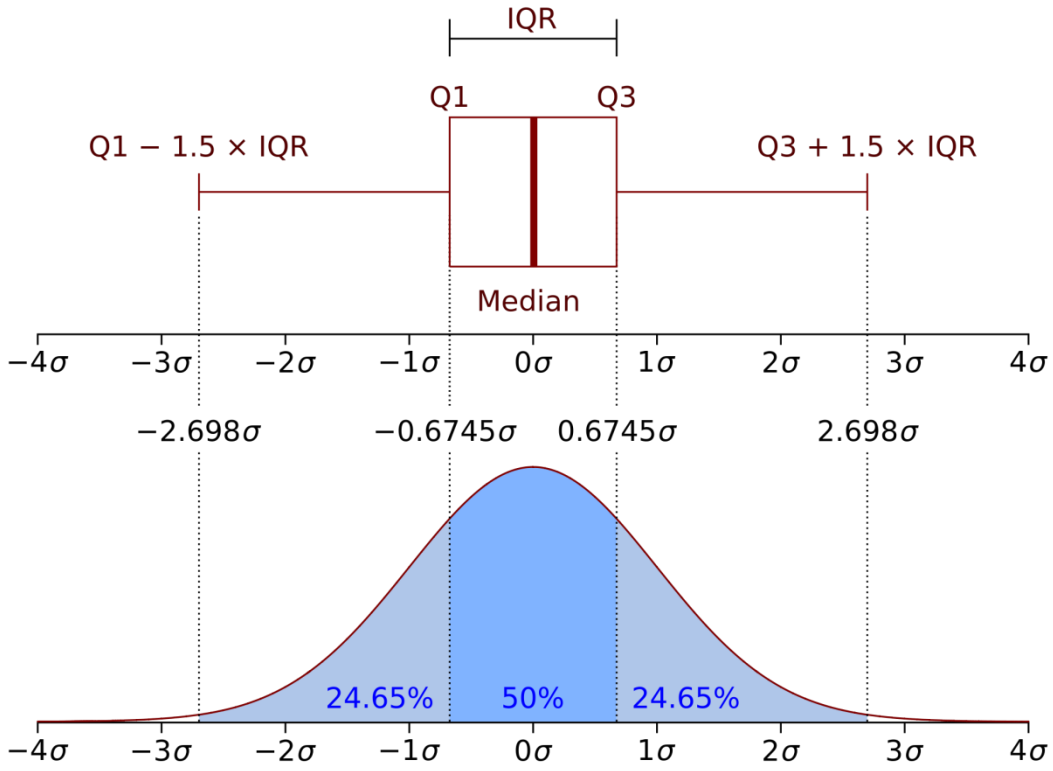
(Folytonos) megfigyelések jellemzése

- Átlag, medián, módusz

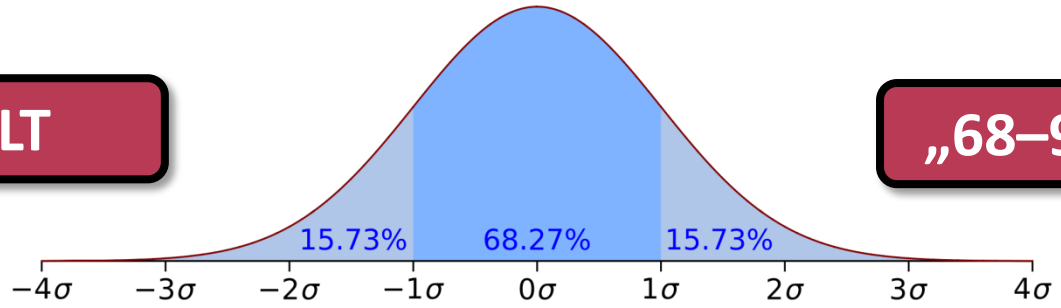
- Percentilis
  - Az n-edik percentilisnél az értékek $n\%$-a kisebb
- Kvartilis
  - Q1, Q3: 25. és 75. percentilis
  - Q2: medián

- Inter-quartile range (IQR)
  - Q3 – Q1

Kvartilisek szerepe



Ismétlés: CLT



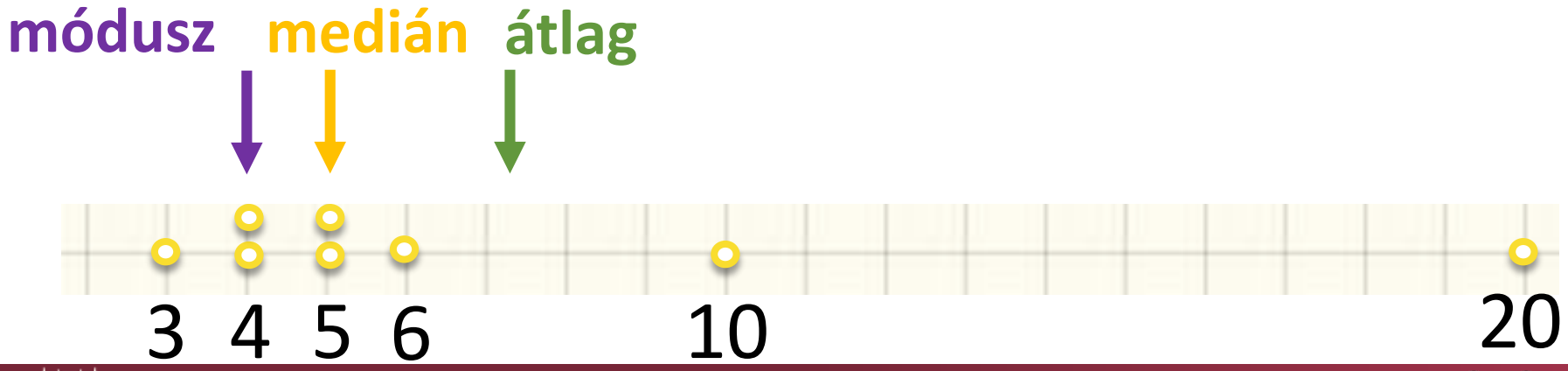
„68–95–99.7 rule”

Centrális tendencia és diszperzió

- Centrális jelleg jellemzői:
  - Átlag, medián, multimodalitás (illetve módus)
- „Diszperzió” jellemzői
  - Percentilisek, szórás(ok), variancia
- Melyik mennyire érzékeny a kiugró értékekre?
- Megj.: a mintaátlag vs. populáció-átlag jellegű kérdésekkel itt nem foglalkozunk
  - (Mi minek hogyan milyen becslője…)

(Folytonos) megfigyelések jellemzése

- A „központ” jellemzése
  - Átlag, **medián**, módusz
  - {3, 4, 4, 5, 5, 6, 10, 20}
    - Átlag: ~ 7.125
    - Medián: 5
    - Módusz: 4 és 5



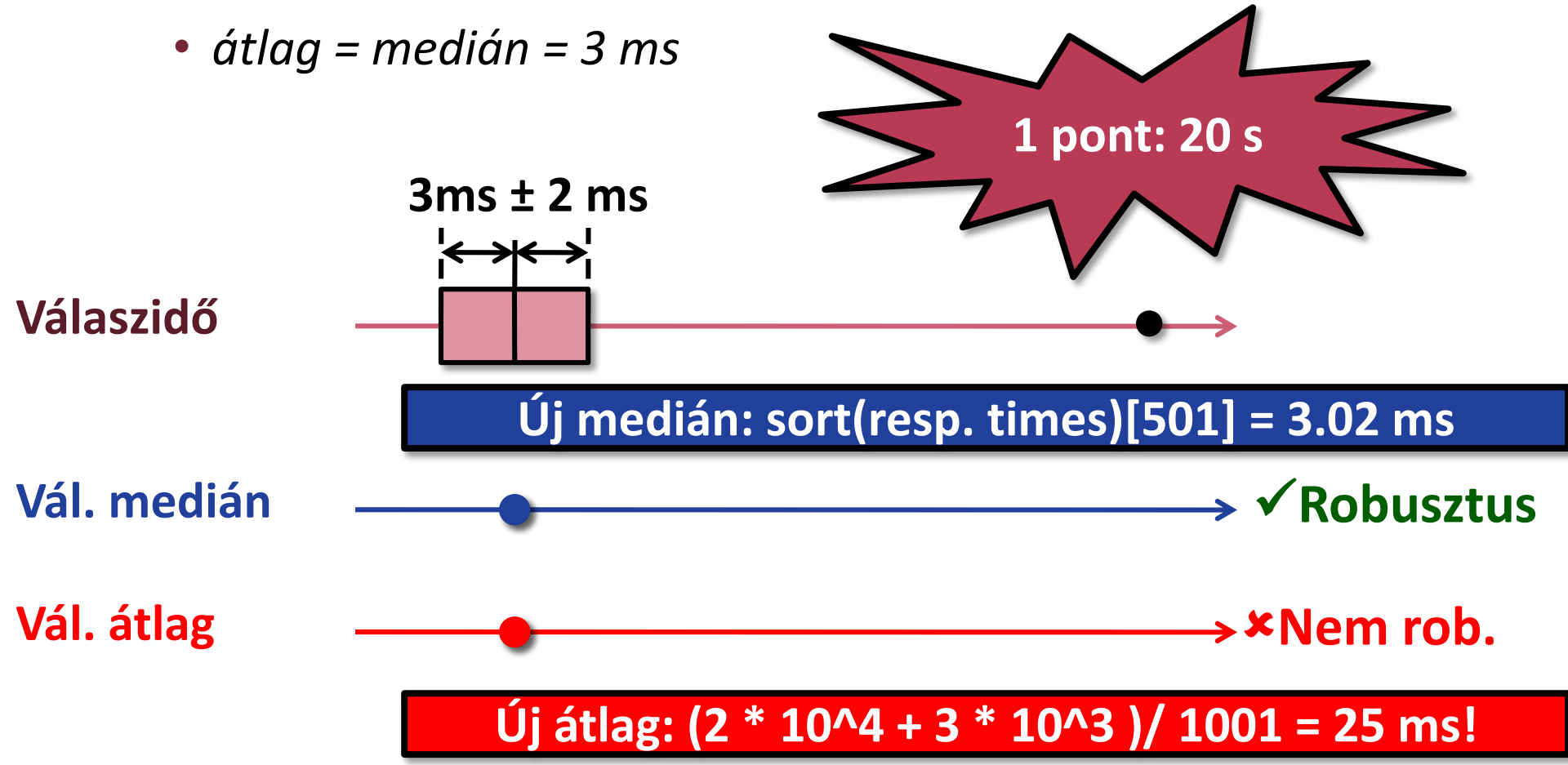
(Folytonos) megfigyelések jellemzése

Ha az értékeket növekvően sorba rendezzük, akkor a középső adat az adathalmaz **mediánja**. Ha nincs középső adat (páros számú érték esetén), akkor a **medián** a két középső érték átlaga (számtani közepe).

A **módusz** az adathalmazban legtöbbször előforduló érték. Ez nem feltétlenül egyértelmű, ilyenkor több móduszról beszélünk.

Robusztus mérőszámok

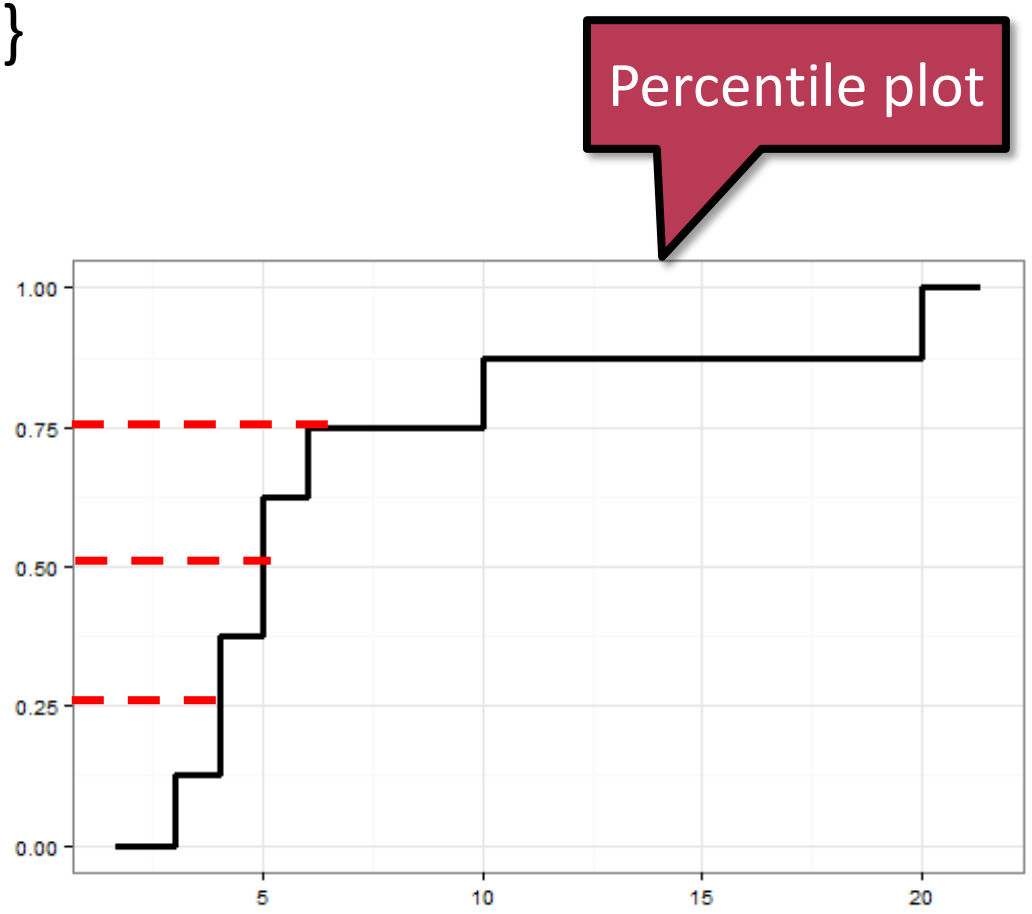
- Alaphalmaz
  - 1000 pont ~ U(1, 5) egyenletes eloszlás
    - *átlag = medián = 3 ms*



Terjedelem jellemzése: percentilisek

Az n-edik **percentilisnél** az értékek $n\%$-a kisebb.

- Percentilis
  - {3, 4, 4, 5, 5, 6, 10, 20}
    - 50. percentilis: 5
    - 25. percentilis: 4
    - 75. percentilis: 6
- Kvartilis
  - Q1: 25. percentilis
  - Q3: 75. percentilis
  - **Q2: medián**



Minta-variancia; minta kovariancia-mátrix

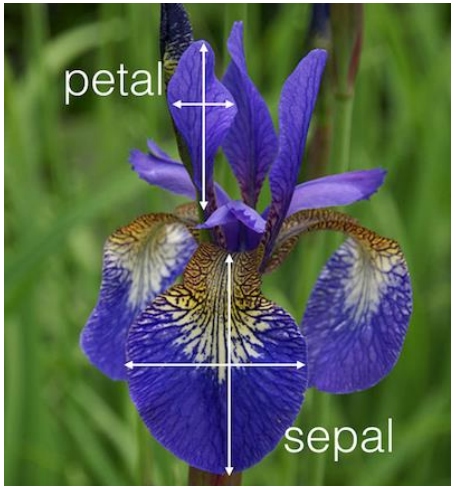
$$s^2_{N-1} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

$$cov(X,Y) = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{x})(Y_i - \bar{y})$$

Teljes populációra (nem mintával becslünk) N-1 helyett N.
A gyakorlatban legtöbbször kevéssé releváns kérdés.

Példa - felvezetés

- Fisher “Iris” adatkészlete
  - “The use of multiple measurements in taxonomic problems” (Fisher, 1936)
  - Cél: osztályozás folytonos jellemzők alapján
  - 50 minta, morfológiai jellemzők
  - 3 faj: setosa, versicolor, virginica



Példa - felvezetés

- A **csésze** (*kalyx*; virágképletbeli jele: **K**) a kétnemű virágtakarójú virágok külső takaróköre, a **csészelevelek** *(sepala)* összessége. A csésze a pártát övezi.

- A **párta** (*corolla*; virágképletbeli jele: **C**) a kétnemű virágtakarójú virágok belső takaróköre, a **sziromlevelek** *(petala)* összessége. A pártát a csésze övezi.

Demo

- R Studio
- iris adatkészlet
- Leíró statisztikák

Variancia, kovariancia: példa

```
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
>
```

Variancia, kovariancia: példa

```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
> 
```

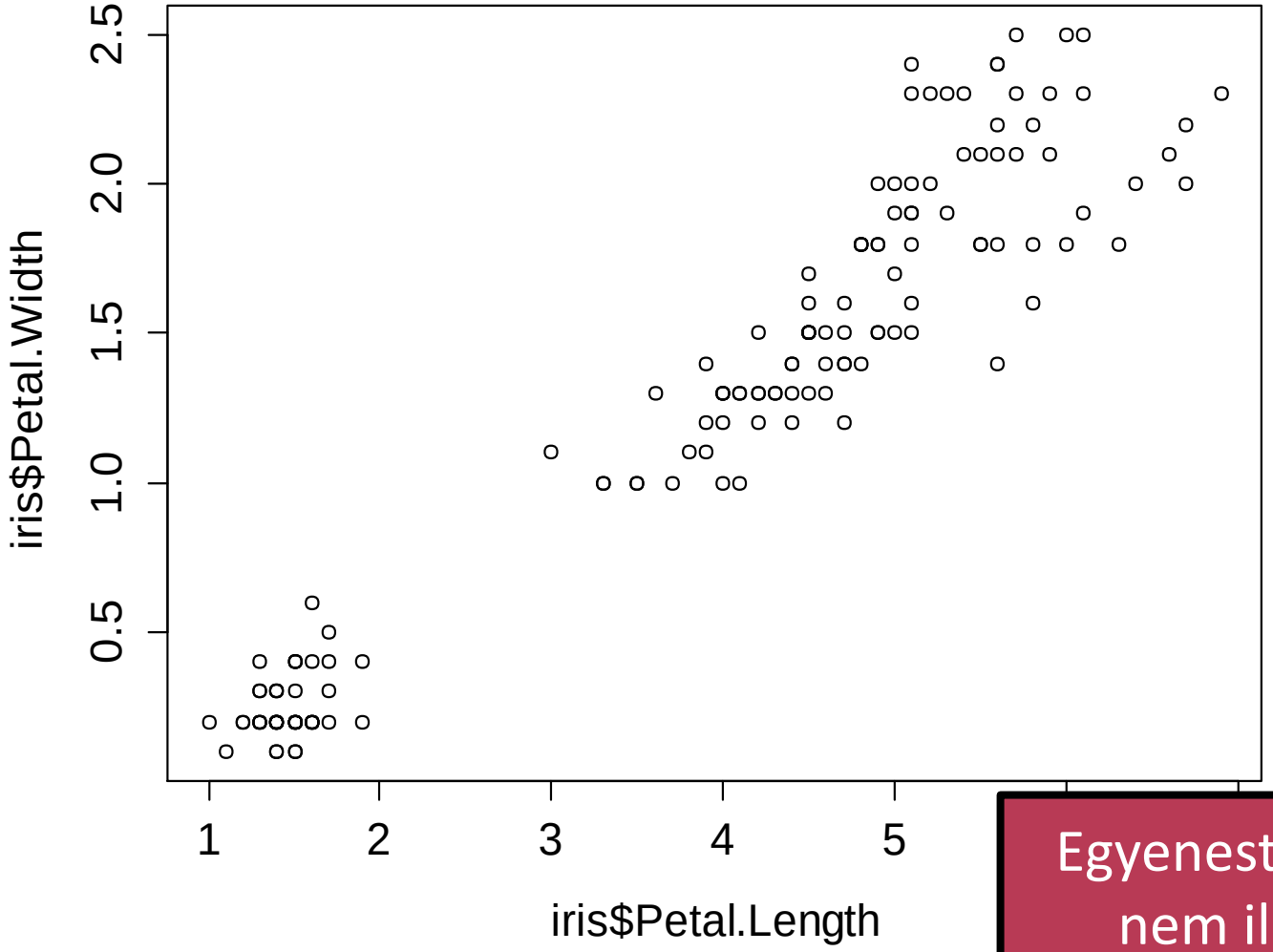
(Minta) variancia, kovariancia: példa

```
> cov(iris[,1:4])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    0.6856935  -0.0424340    1.2743154   0.5162707
Sepal.Width    -0.0424340   0.1899794   -0.3296564  -0.1216394
Petal.Length    1.2743154  -0.3296564    3.1162779   1.2956094
Petal.Width     0.5162707  -0.1216394    1.2956094   0.5810063
```

Normalizálás (szórások szorzatával): Pearson-féle lineáris korrelációs koefficiens

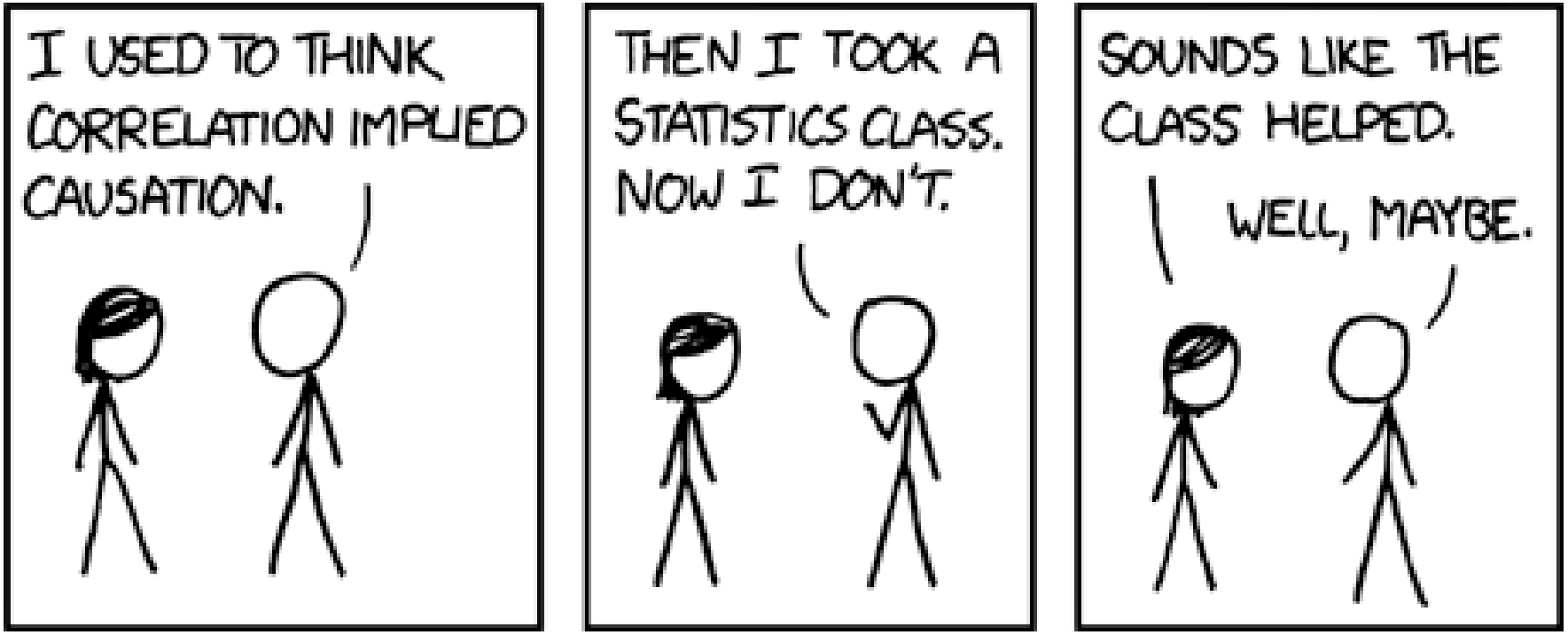
```
> cor(iris[,1:4])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
> 
```

Lineáris korrelációs koefficiens



Egyenest most még nem illesztünk

Correlation != Causation

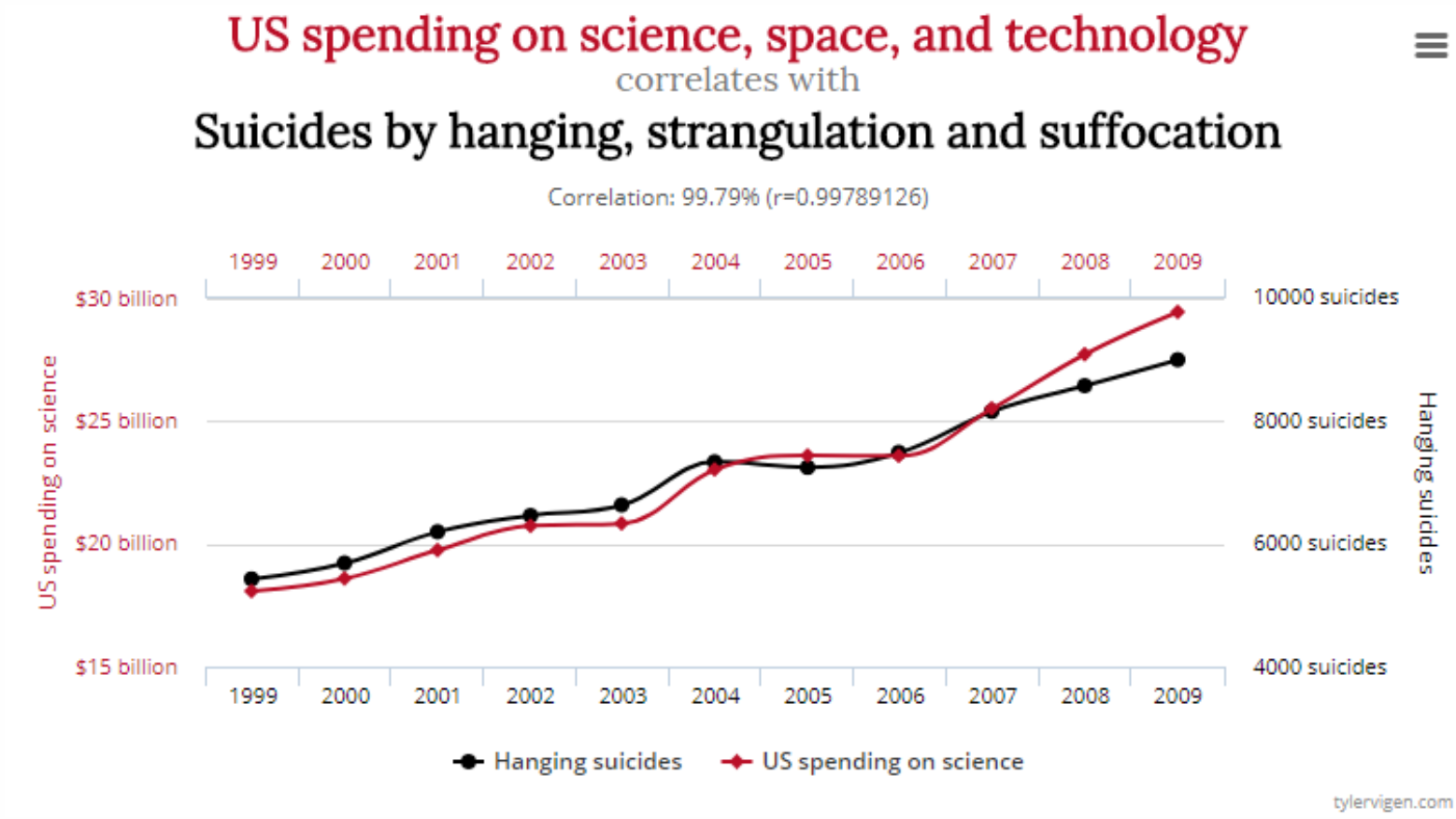


Forrás: https://xkcd.com/552/

! Az oksági kapcsolatok felderítése sem esélytelen, de jóval bonyolultabb, mint az irány nélküli, „leíró” korreláció. Lásd pl.: Pearl, Judea. *Causality*. Cambridge University Press, 2009.

On an even lighter(?) note

tylervigen.com/spurious-correlations

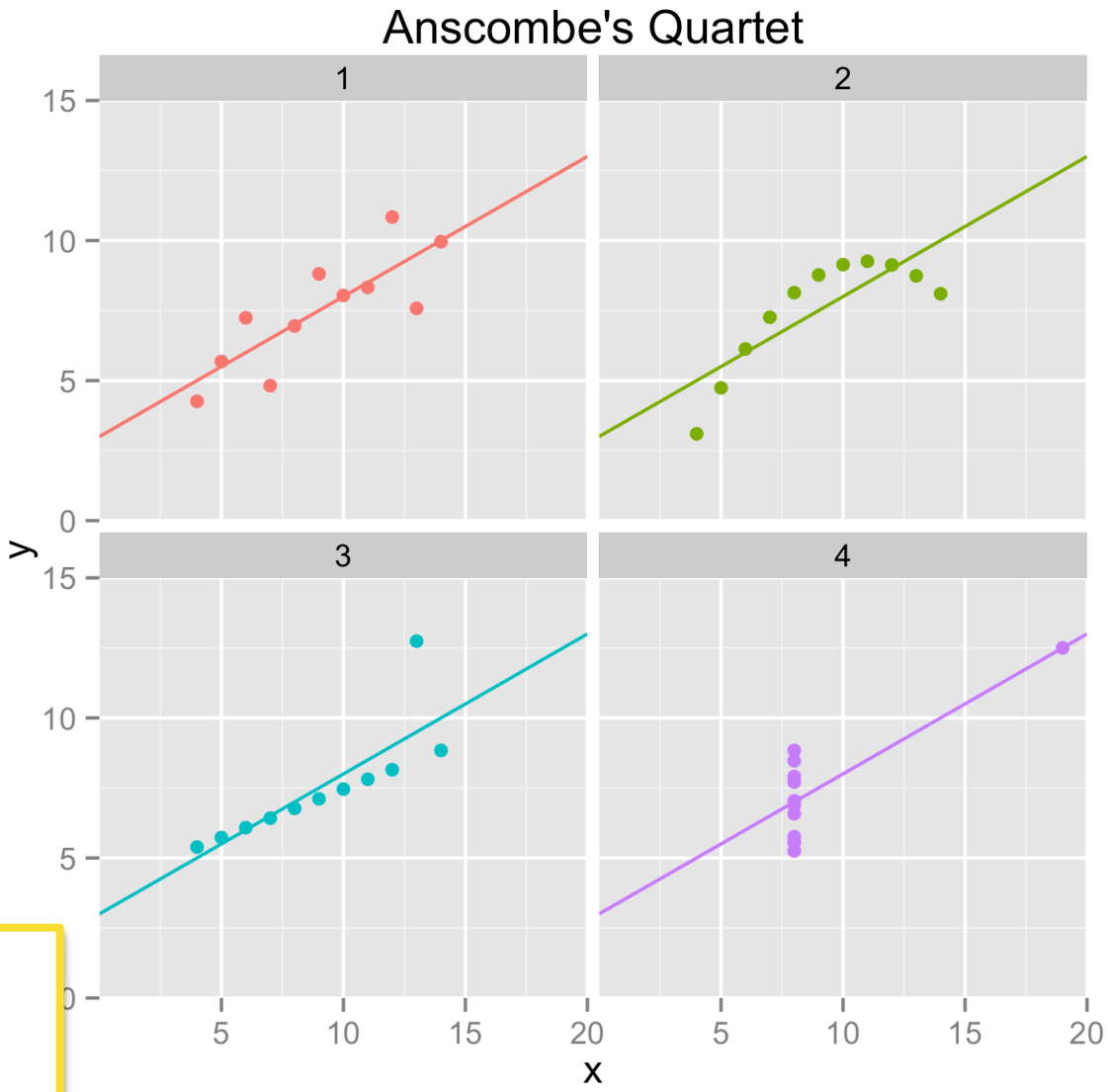


Exploratory Data Analysis

Demo

- R Studio
- Anscombe adatkeret
- Átlagok, mediánok, varianciák, kovarianciák, korrelációk

Anscombe négyese



Okok?
Tanulság?

Felderítő adatanalízis

- *Exploratory Data Analysis*: statisztikai tradíció,
  - mely koncepcionális
  - és számítási eszközökkel segíti
  - minták felismerését és ezen keresztül
  - hipotézisek felállítását és finomítását.

- Komplementere: *Confirmatory Data Analysis*
  - Hipotézistesztelés, modellválasztás, paraméterillesztés, …

- Legismertebb vizionáriusa: John W. Tukey

EDA

- Cél: adatok „megértése”
  - „detektívmunka”
  - erősen ad-hoc
- Fő eszköz: adatok „bejárása” grafikus reprezentációkkal
- Hipotézisek: iteratív folyamat
- Flexibilitás és pragmatizmus

Dr. John Snow és az 1854-es kolerajárvány

- A járvány nem „miazmikus”

„About half of our sensory neurons are dedicated to vision, endowing us with a remarkable pattern-recognition ability.”

Prof. Alfred Inselberg

Mindent a szemnek!

„Masszív” erőforrások

- 120.000.000 szenzor
- 10^{10} feldolgozó egység

A folyamat alapja az interakció

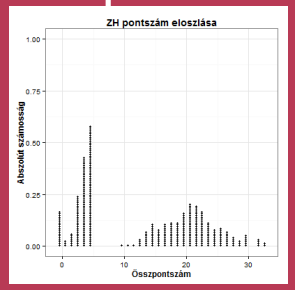
1. Adatvizualizáció
   – több ábra együttes vizsgálata
2. Vizuális kiértékelés
   – emberi kognitív képességek használata
3. Vizuális kiválasztás és manipuláció
4. Interpretáció, korreláció más modellekkel, kiértékelés

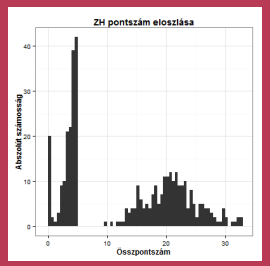
1 változó – eloszlásokra

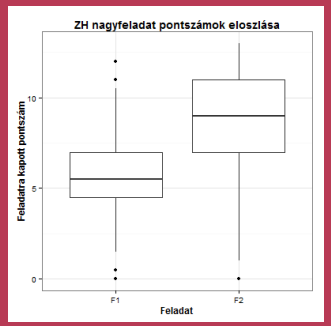
Változók

Numerikus

ZH pontszám: [13, 15, 2, …]

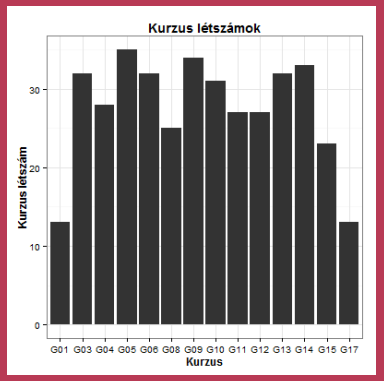






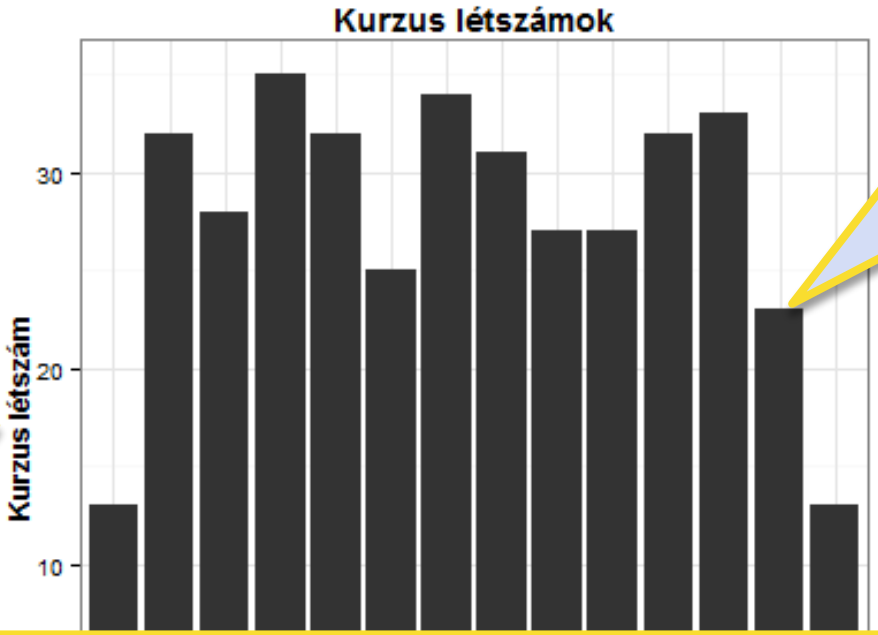
Kategorikus

Kurzus: [G01, G03, G15, G17, …]



Oszlopdiagram

- Bemenő változó: kurzus kód
- Kérdés: az egyes kurzusokra hányan járnak?



Oszlop-magasság: adott érték gyakorisága

abszolút gyakoriság!

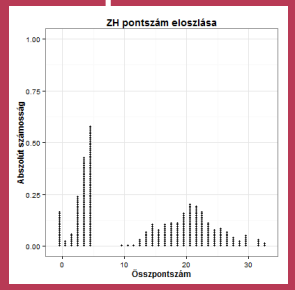
Tervezői döntés: értékkészlet darabolása
Pl.: kedd-csütörtök-péntek

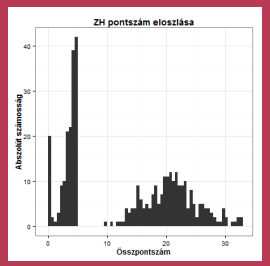
1 változó – eloszlásokra

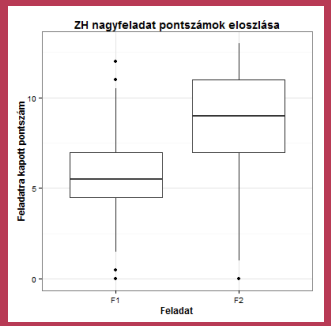
Változók

Numerikus

ZH pontszám: [13, 15, 2, …]

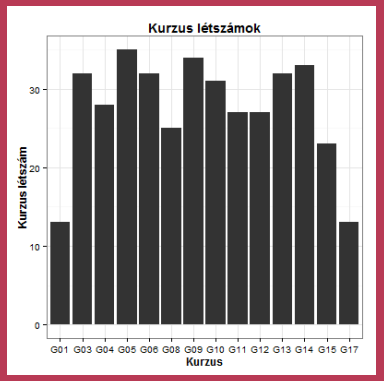






Kategorikus

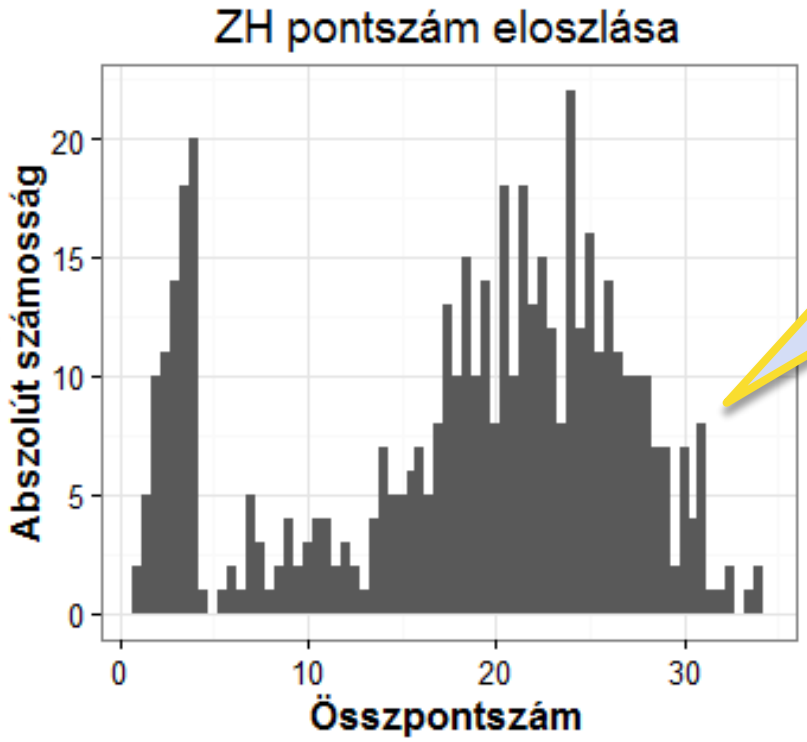
Kurzus: [G01, G03, G15, G17, …]



Hisztogram

- Bemenő változó: ZH összpontszám
- Kérdés: hogyan alakultak a ZH pontszámok?

abszolút gyakoriság!

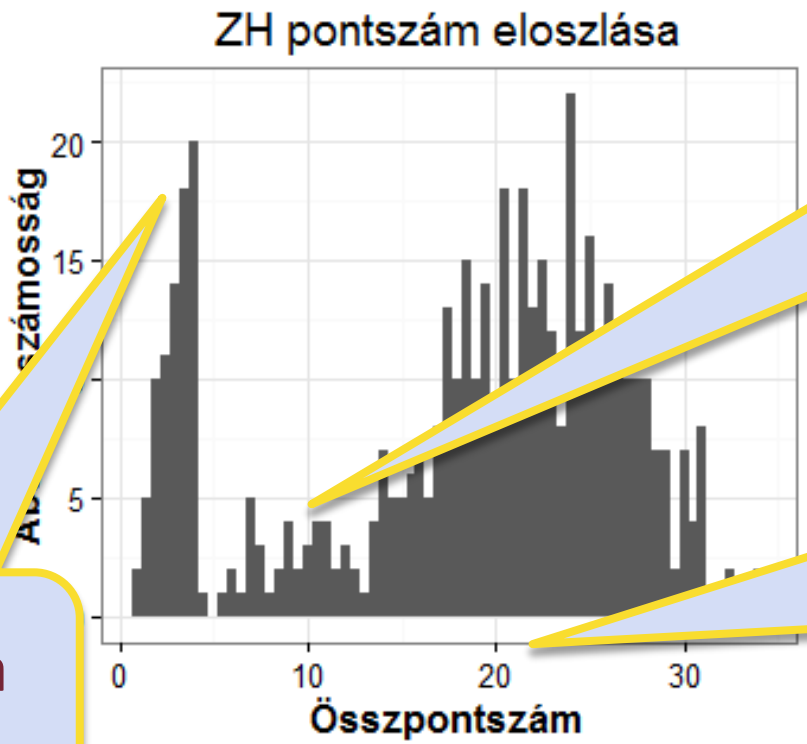


Oszlop-magasság: adott **intervallum** számossága

Tervezői döntés: mekkora legyen az intervallum hossza (bin size)?
Pl.: elég 1 pontos felbontással, vagy menjünk fél pontokig?

Hisztogram

- Bemenő változó: ZH összpontszám
- Kérdés: hogyan alakultak a ZH pontszámok?



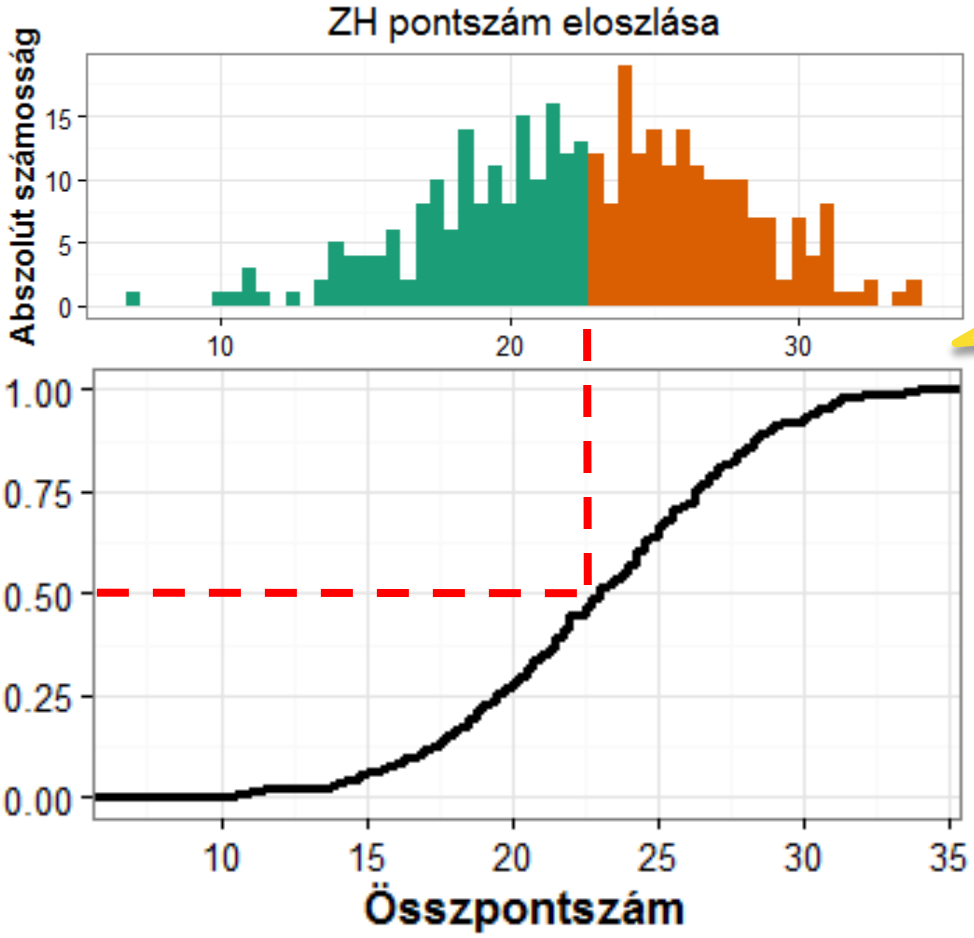
Akik átmentek a beugrón, valószínűleg át is mentek

Sokan voltak a határon

18 pont körül volt az átlag, 20 körül a medián

Ami nem látszik a hisztogramról

- Hol van az adatok „közepe”?

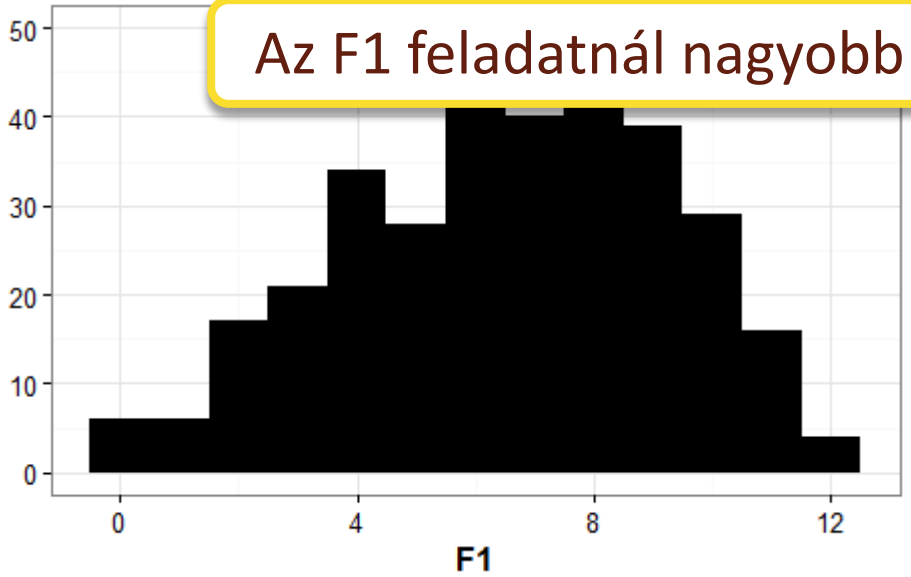


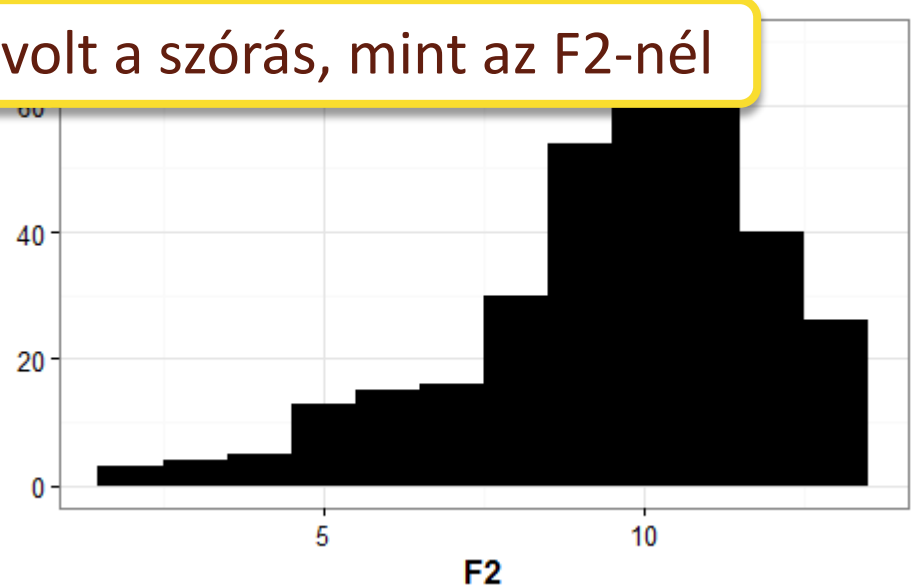
Az átmentek összpontszám mediánja 23

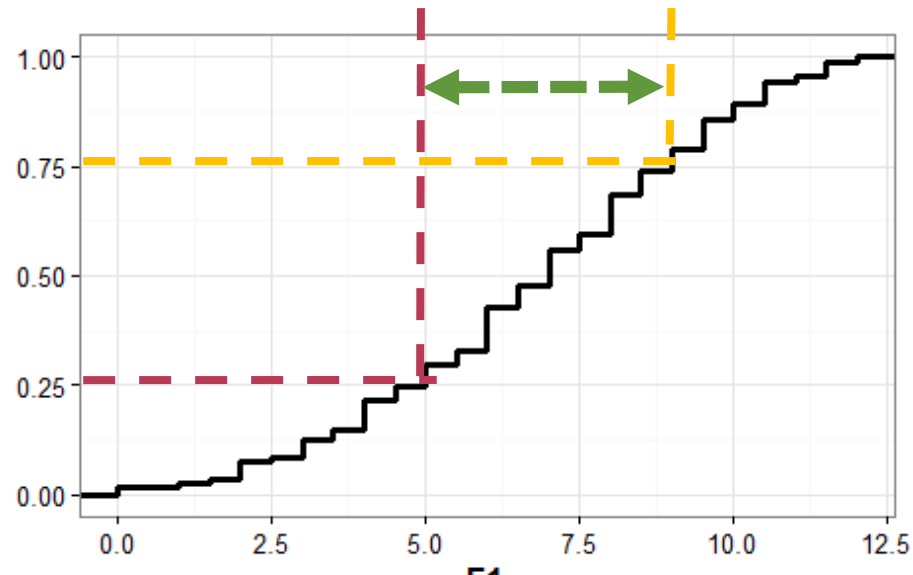
Empirikus CDF vs hisztogram

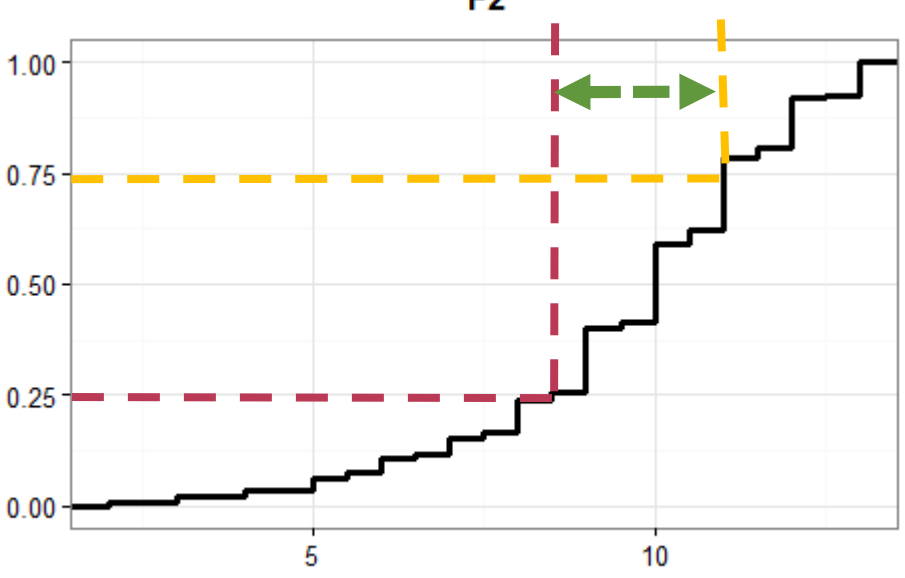
- Mennyire „szórtak” az adatok?

Az F1 feladatnál nagyobb volt a szórás, mint az F2-nél



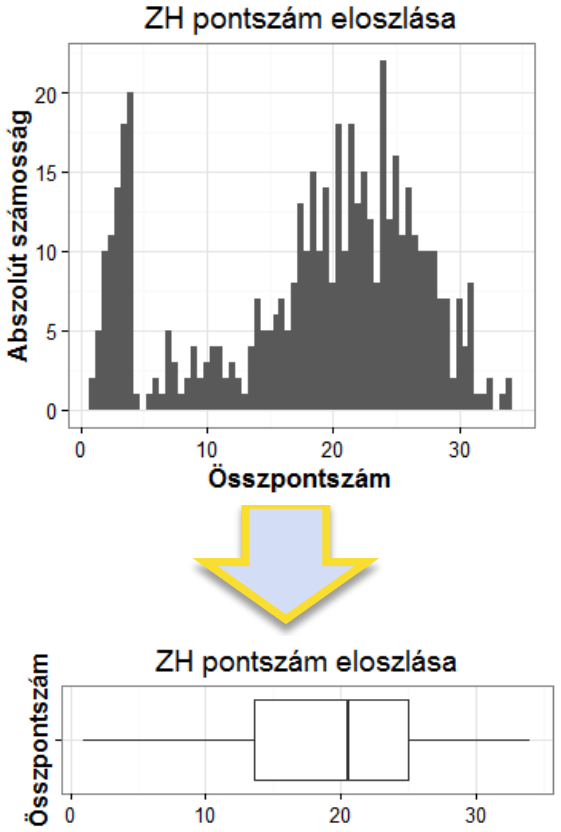






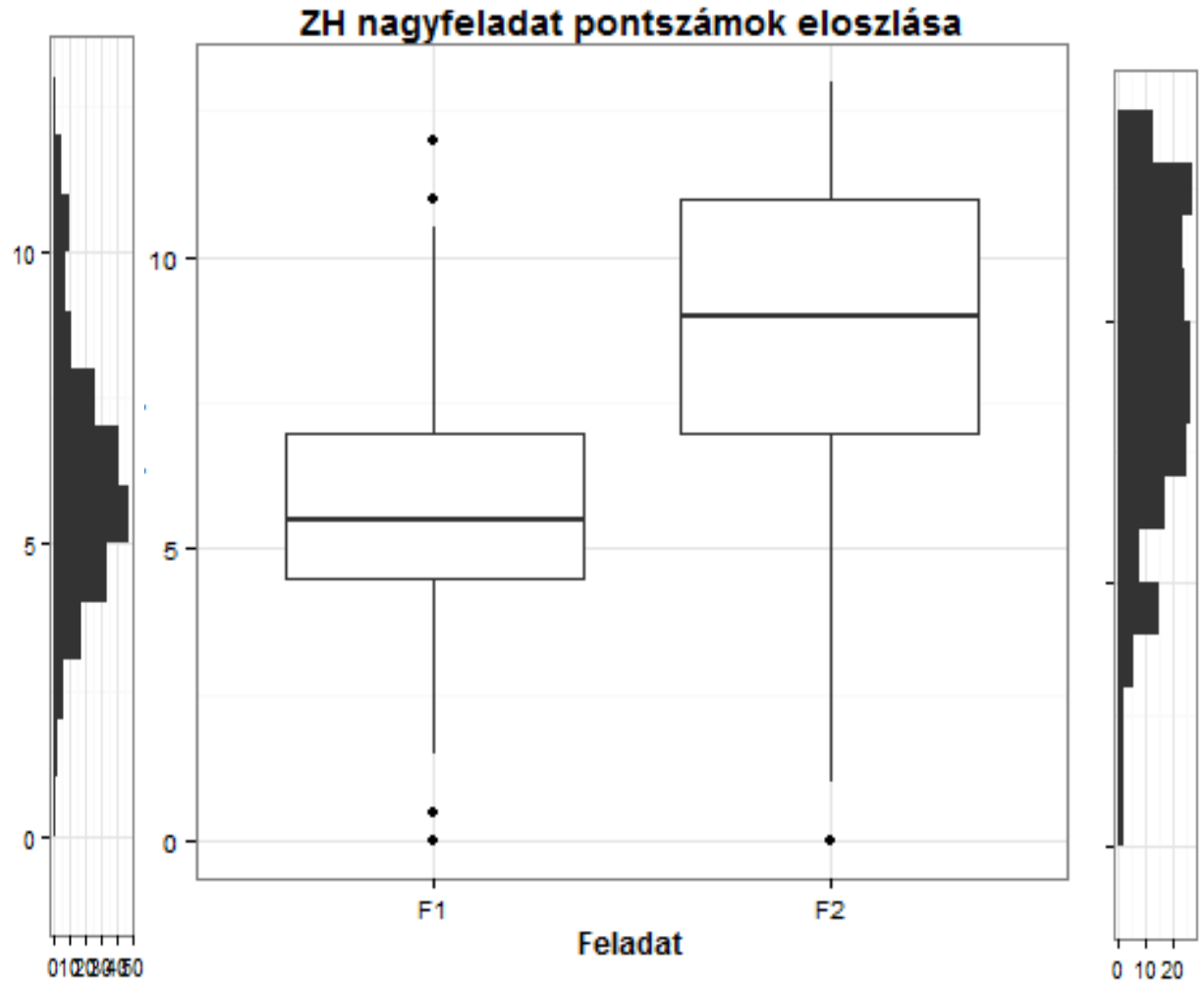
Boxplot

- Bemenő változó: ZH összpontszám
- Kérdés: hogyan alakultak a ZH pontszámok **úgy nagyjából**?

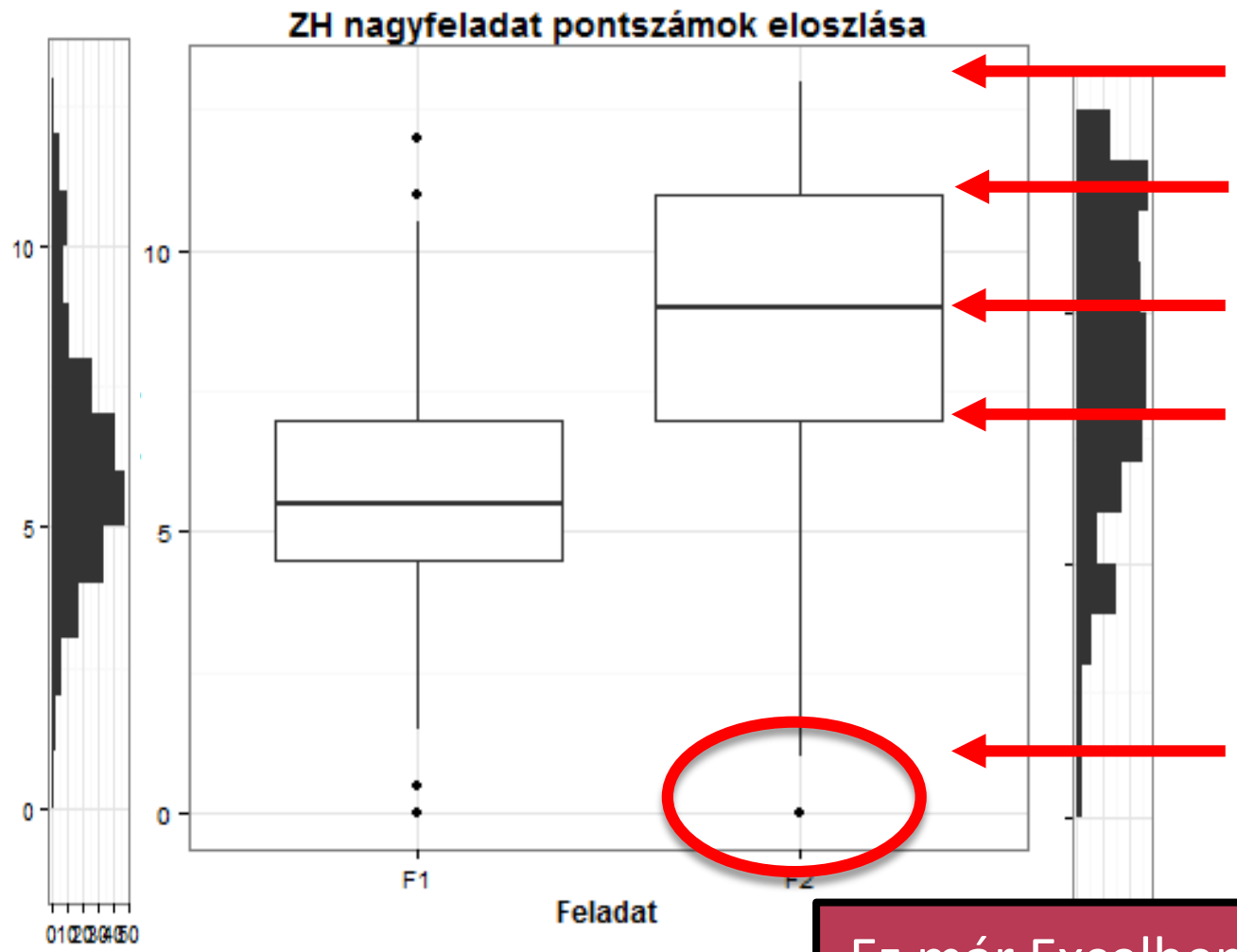


Egyfajta absztrakció itt is: legyenek intervallumok, felesleges minden pontot kirajzolni

Boxplot (Box and whisker plot)



Boxplot (Box and whisker plot)



Felső outlier határ

Q3

Medián

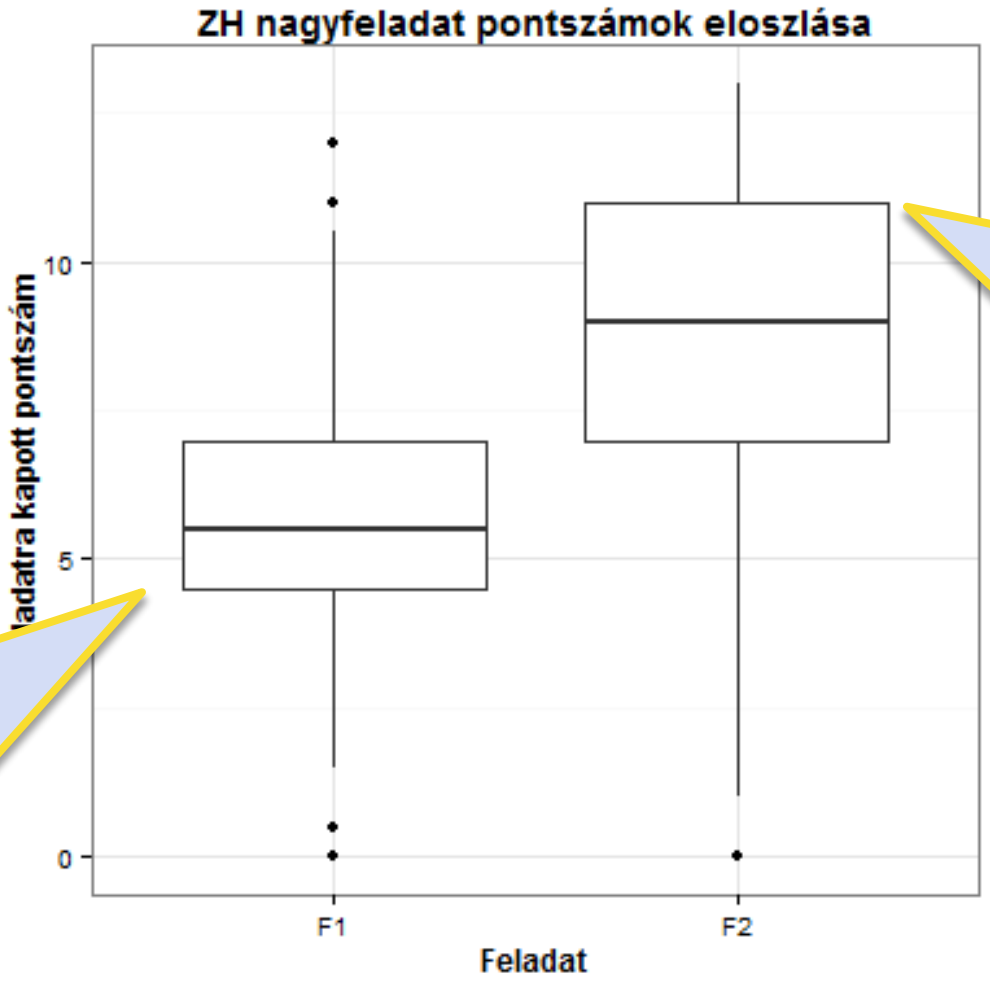
Q1

Alsó outlier határ

Outlierek

Ez már Excelben nem könnyű…

Boxplot (Box and whisker plot)

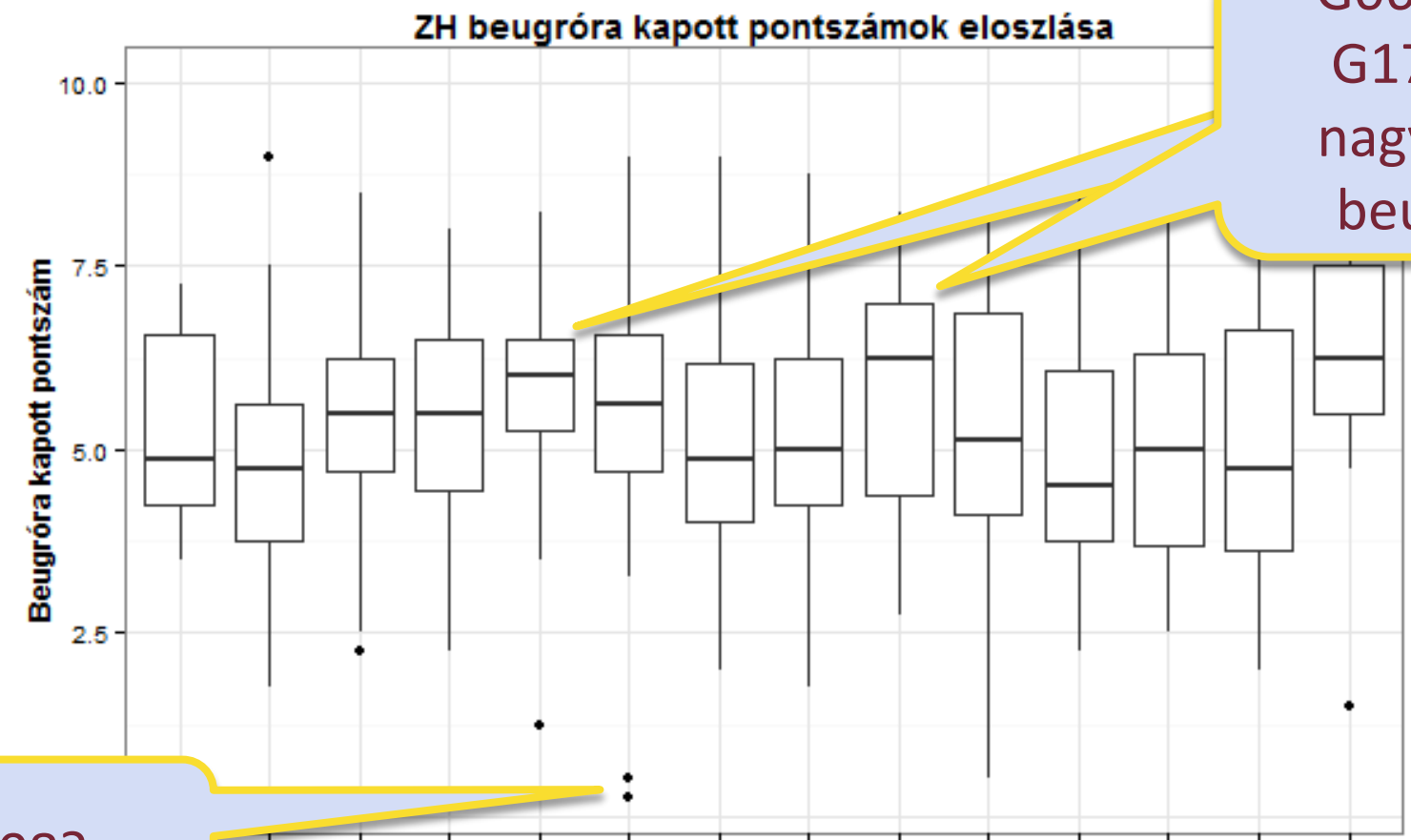


F2-re általában több pontot kaptak, mint F1-re

Az F1 pontszámok 50%-a 4.5 és 7.5 között volt

Boxplot (Box and whisker plot)

- Melyik csoportban hogyan sikerültek a beugrók?

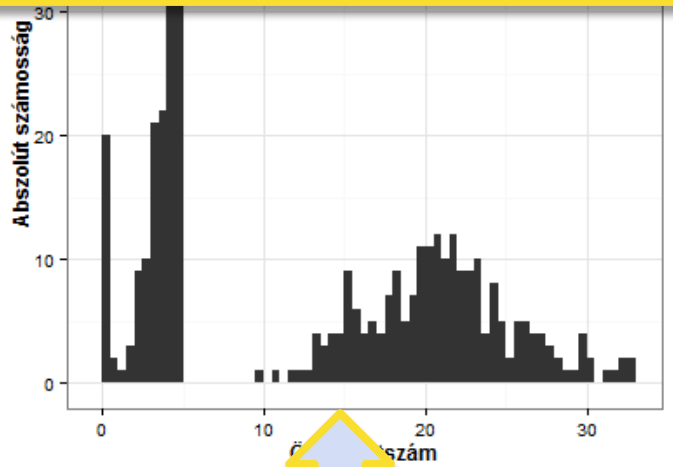


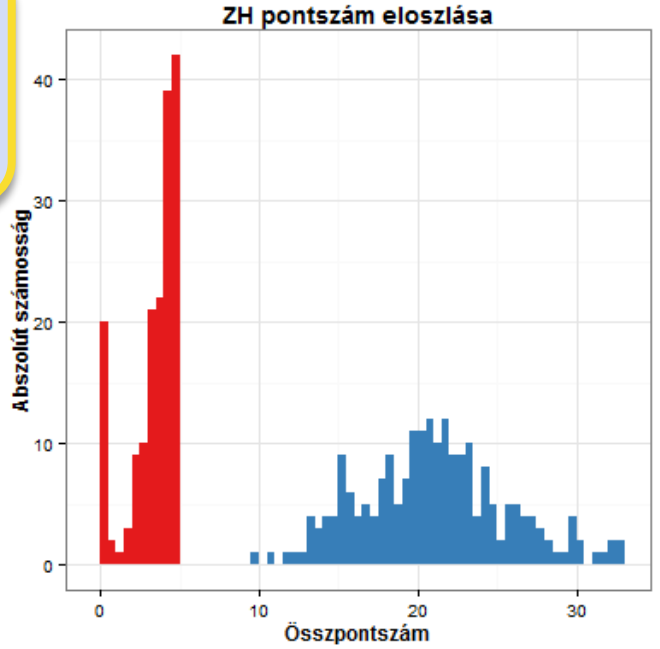
G06, G11, G17-ben nagyon jó beugrók

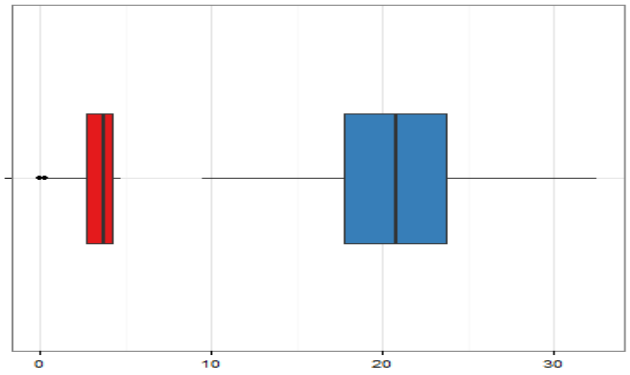
G08?

Boxplot (Box and whisker plot)

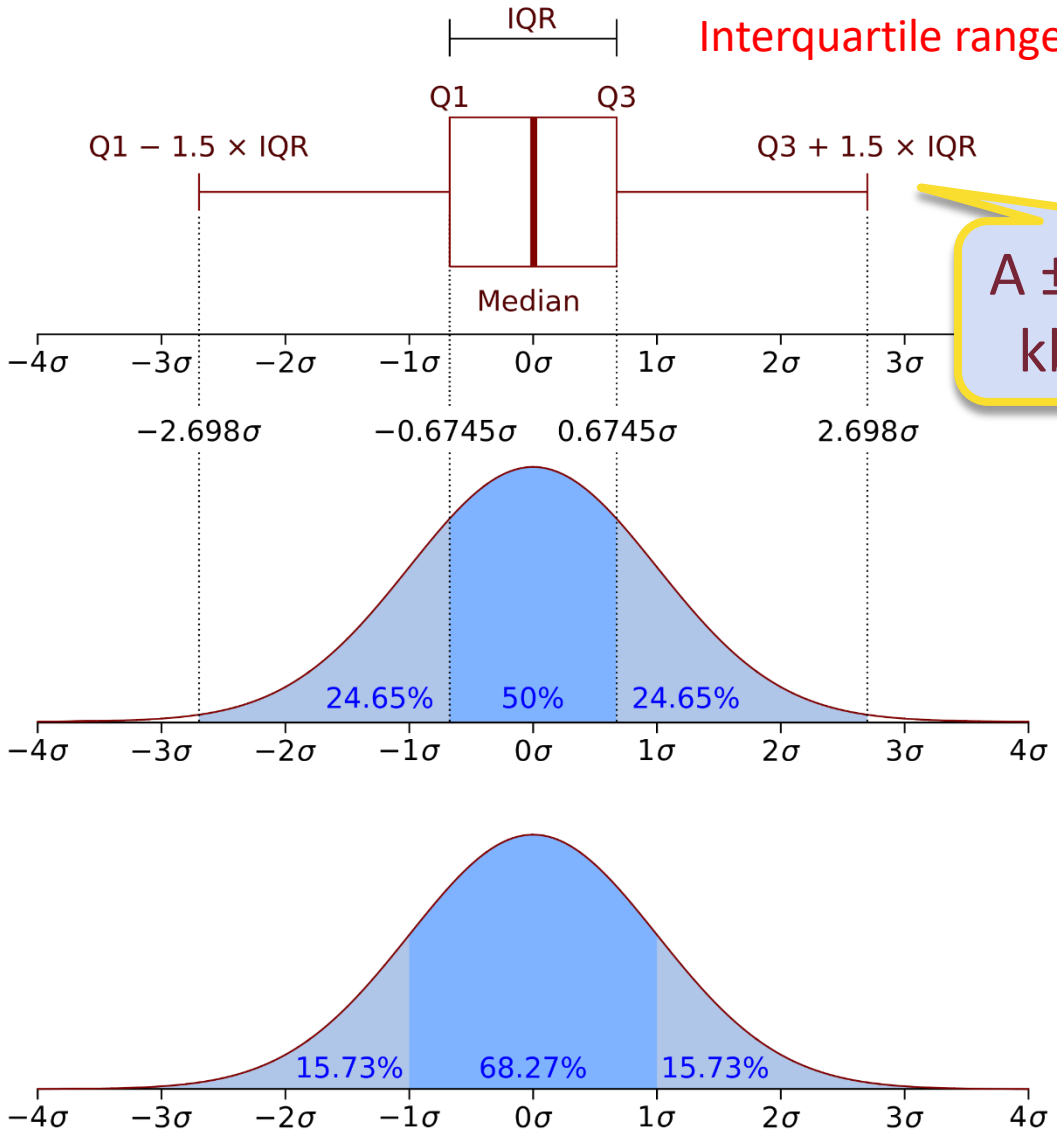
Absztrakció: a boxplottal fontos információt is veszíthetünk!





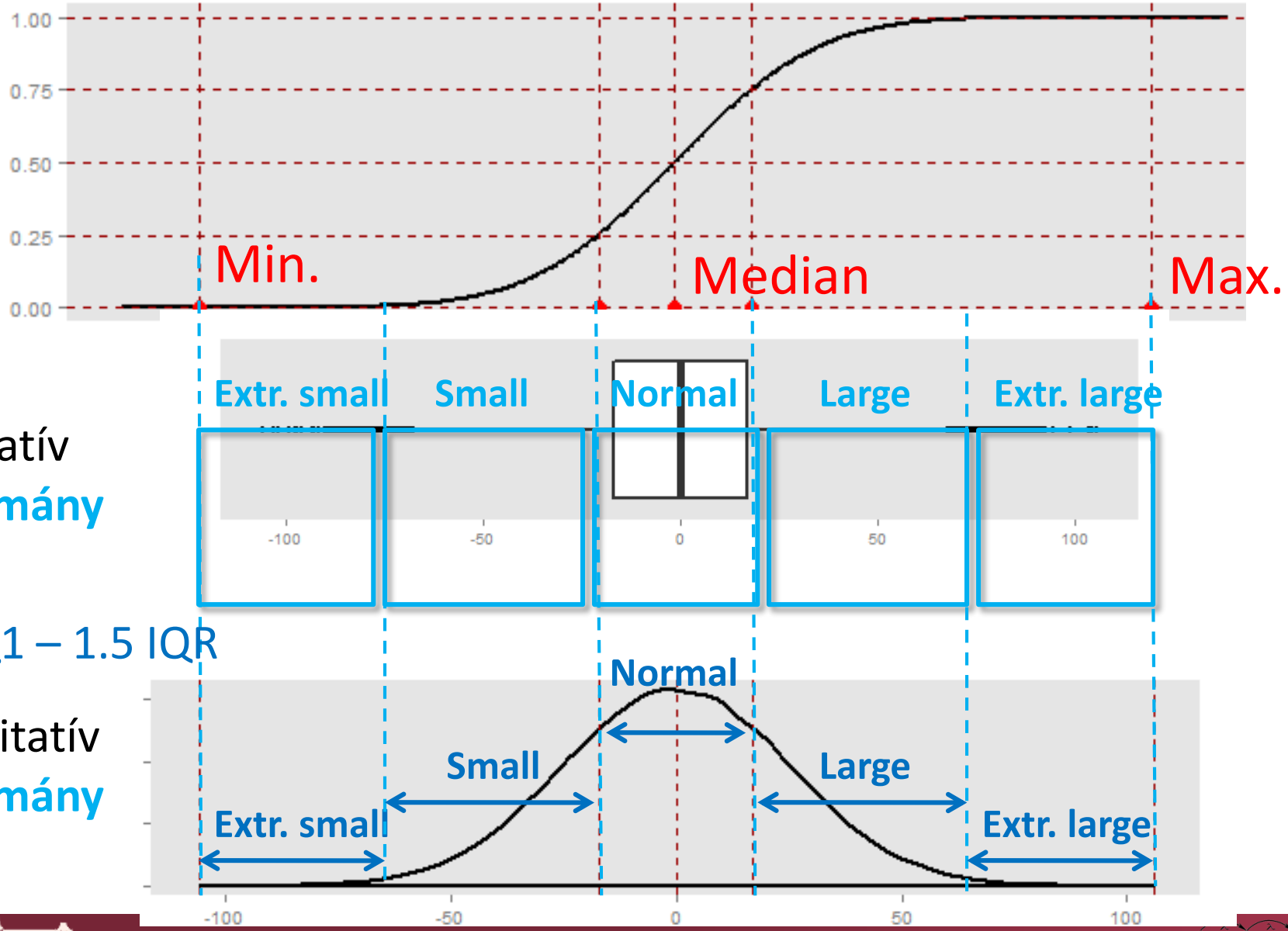


Boxplot (Box and whisker plot)



A ±1.5 × IQR-hez kb. 3σ tartozik

Boxplot: kvalitatív jellemzés

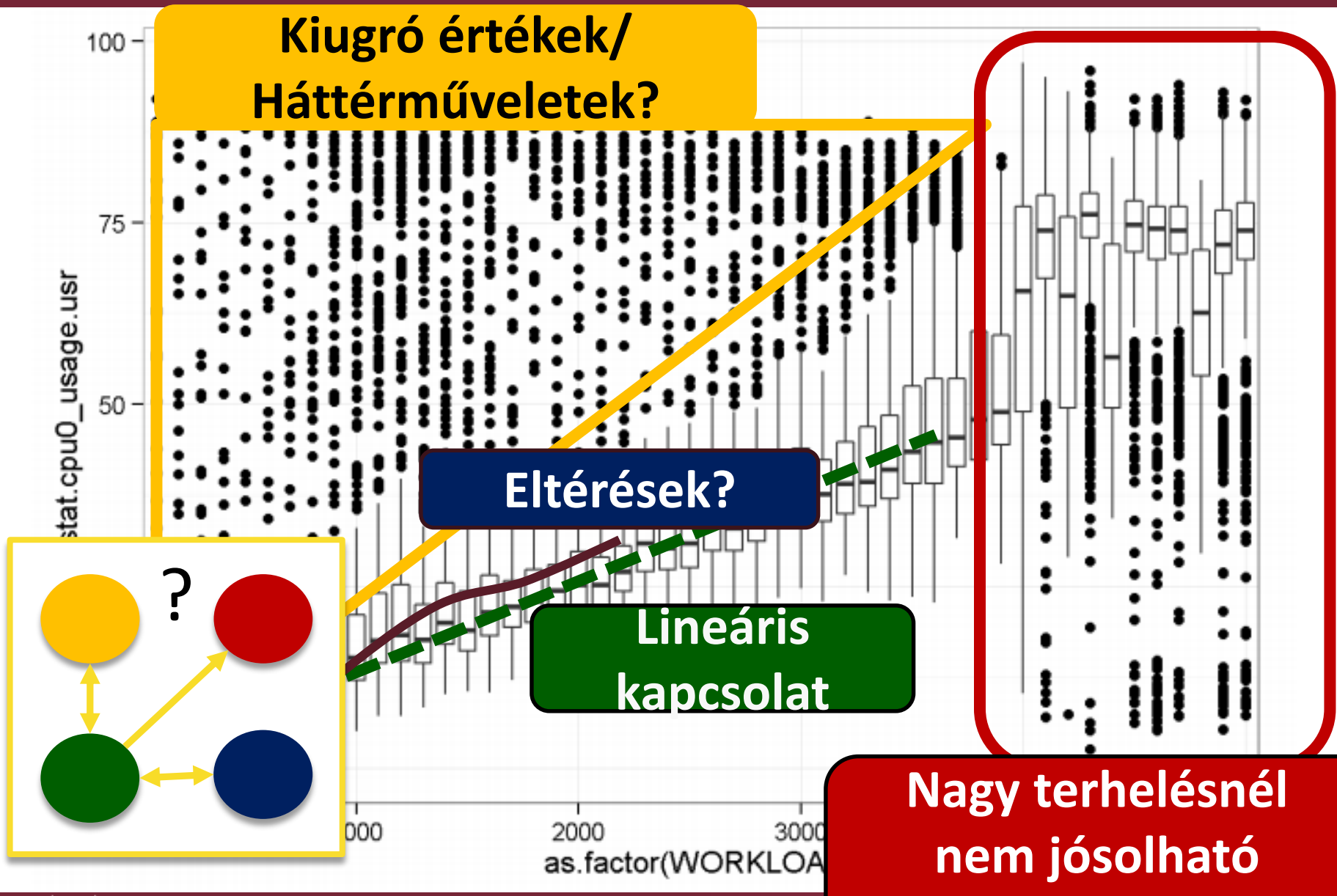


Kvalitatív **tartomány**

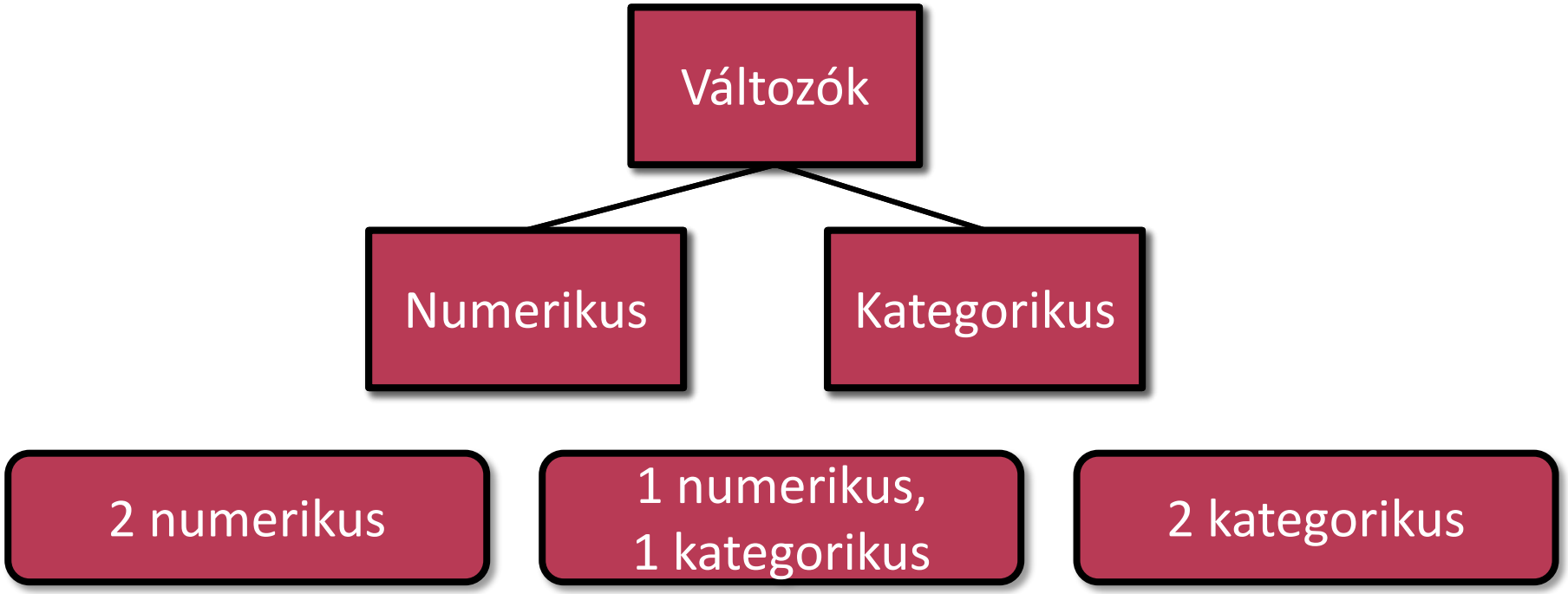
Q1 – 1.5 IQR

Kvantitatív **tartomány**

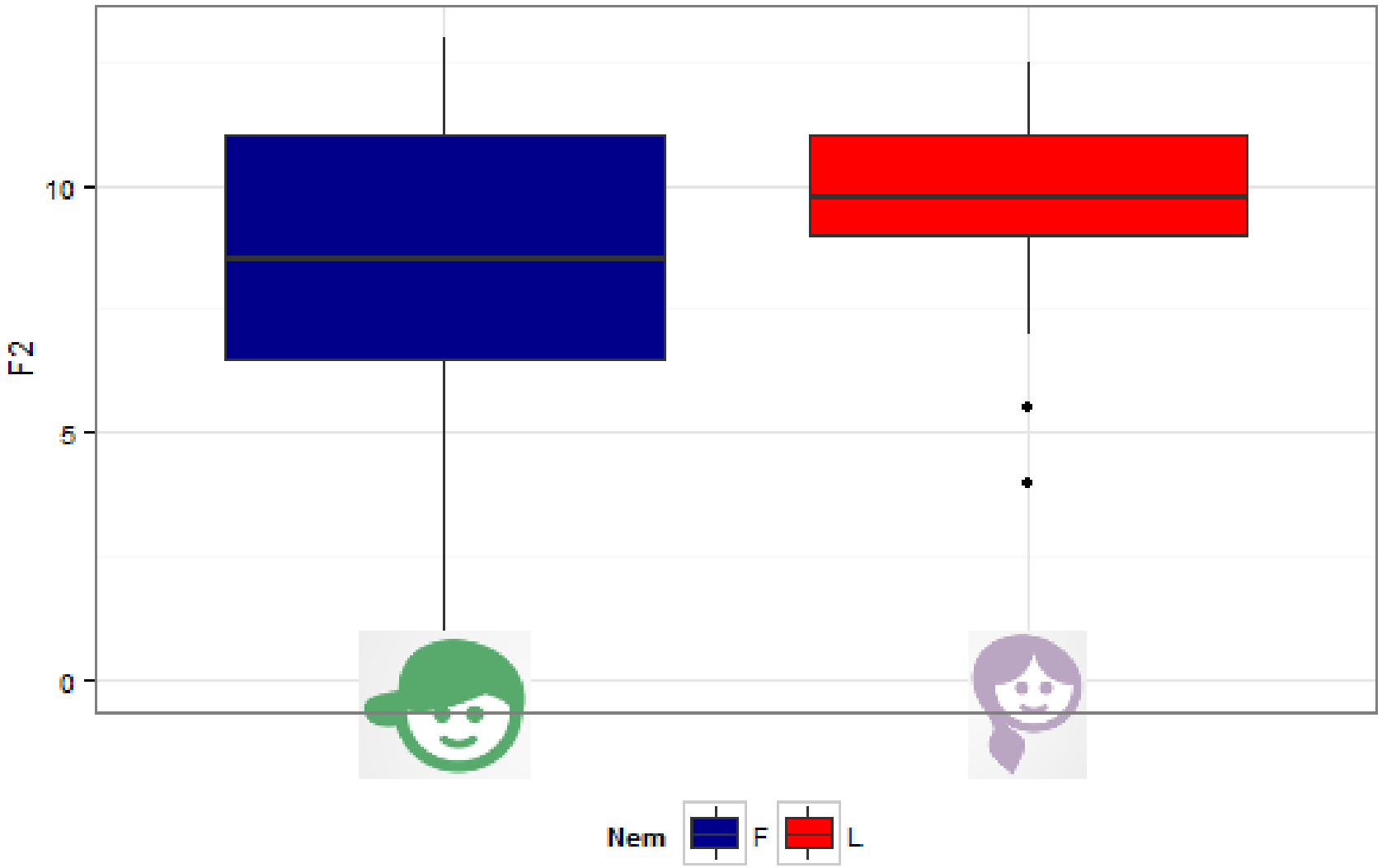
Példa: terhelés vs. kihasználtság



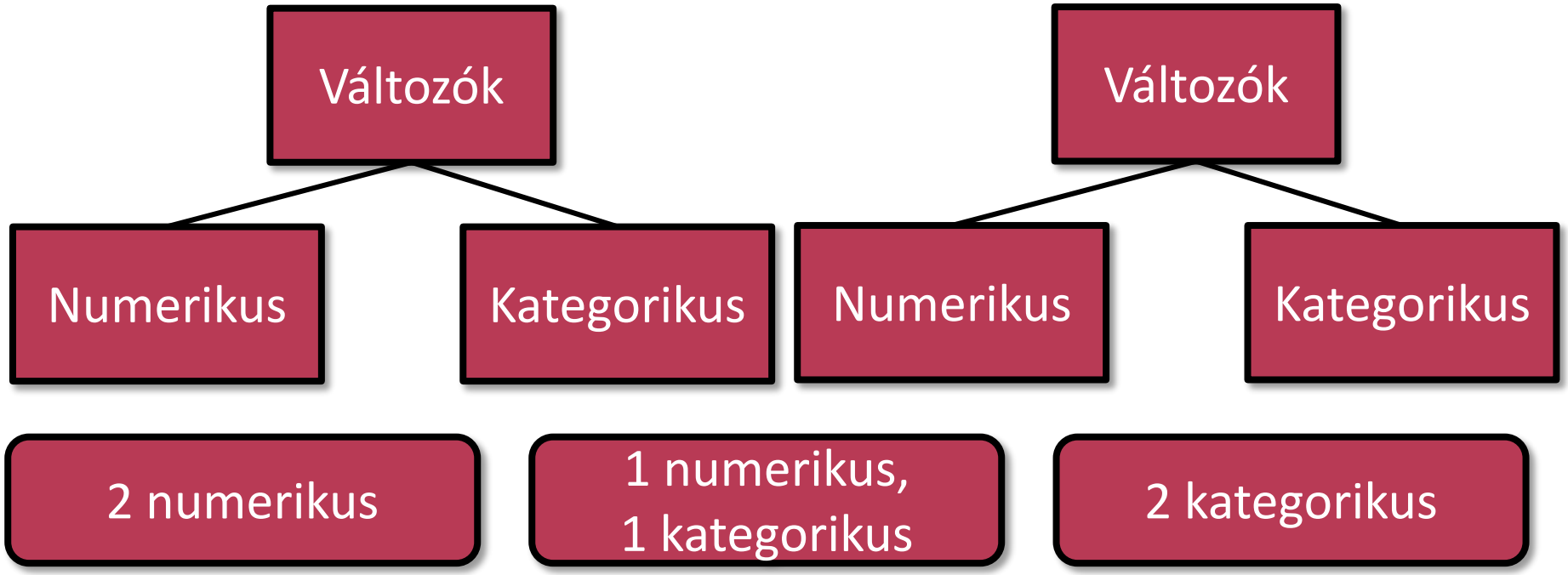
2 változó kapcsolata



Numerikus kategóriánként



2 változó kapcsolata

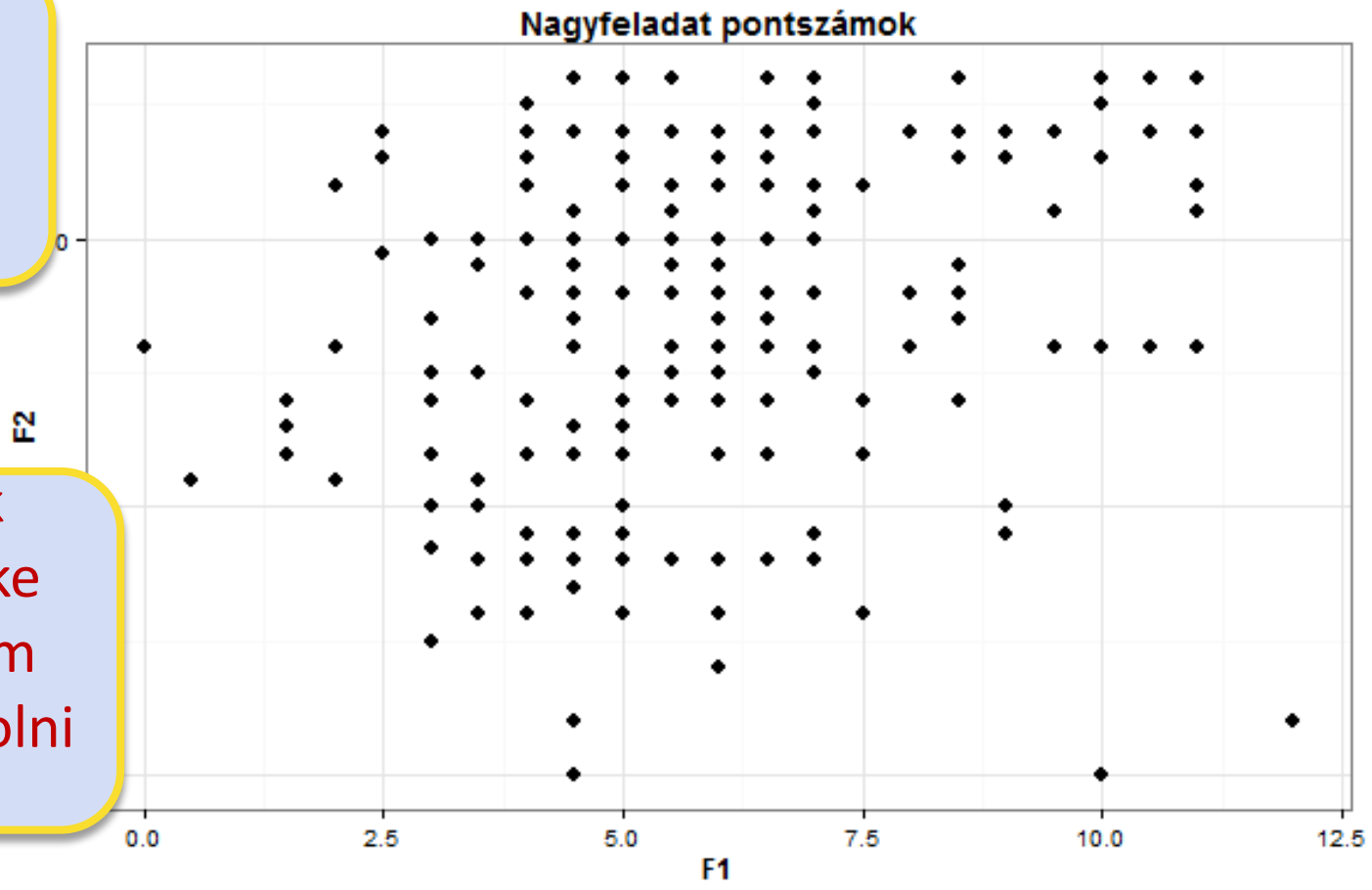


Pont – pont diagram (scatterplot)

- Bemenő változó: nagyfeladatokra kapott pontok
- Kérdés: hogyan viszonyulnak egymáshoz?

Együttesen előforduló pontpárokat vizualizálunk

Ha az egyik változó értéke hiányzik, nem tudjuk felrajzolni ☹



Pont – pont diagram (scatterplot)

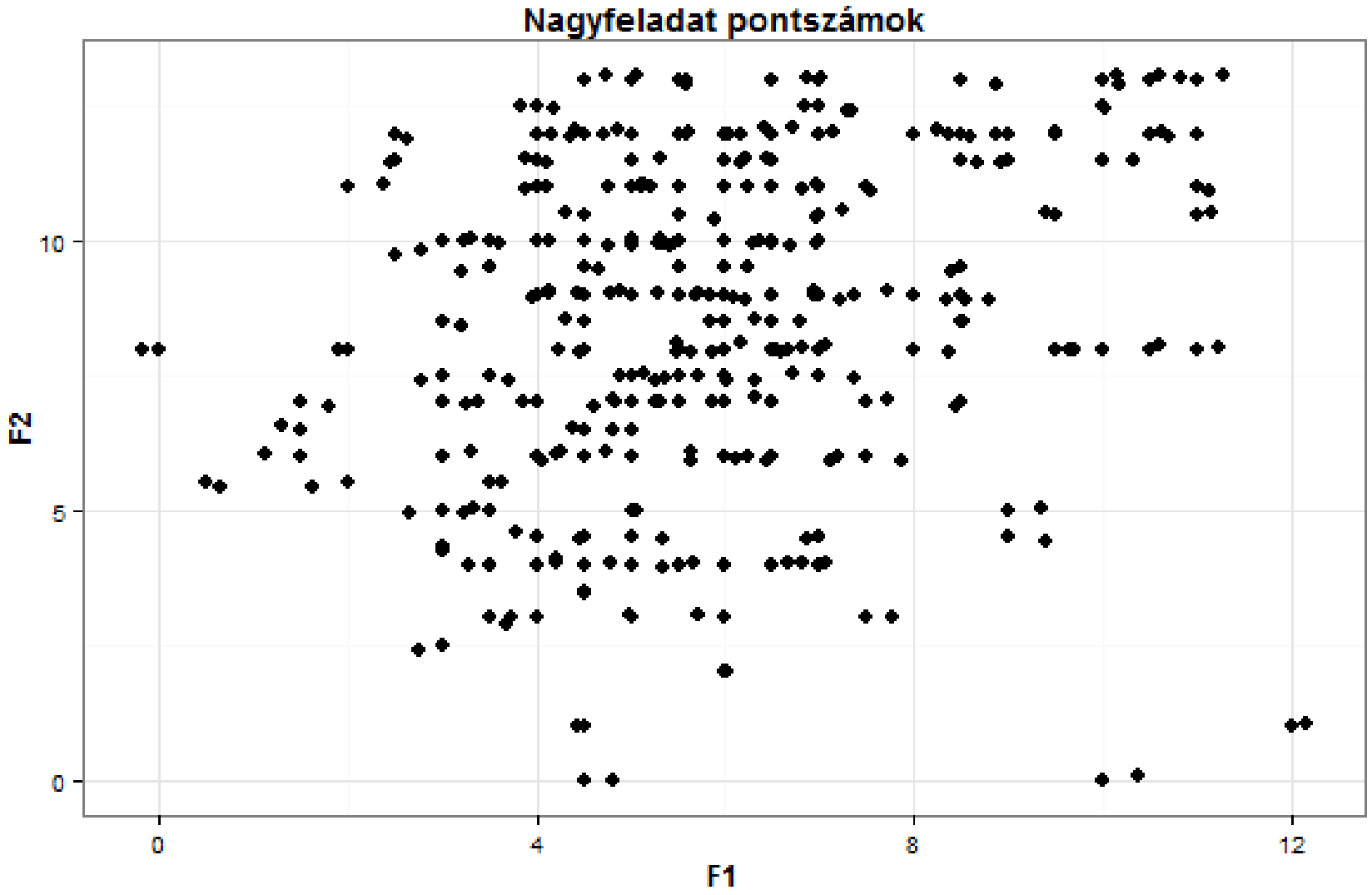
- Bemenő változó: nagyfeladatokra kapott pontok
- Kérdés: hogyan viszonyulnak egymáshoz?

Nem biztos, hogy akinek megy az F1, megy az F2 is

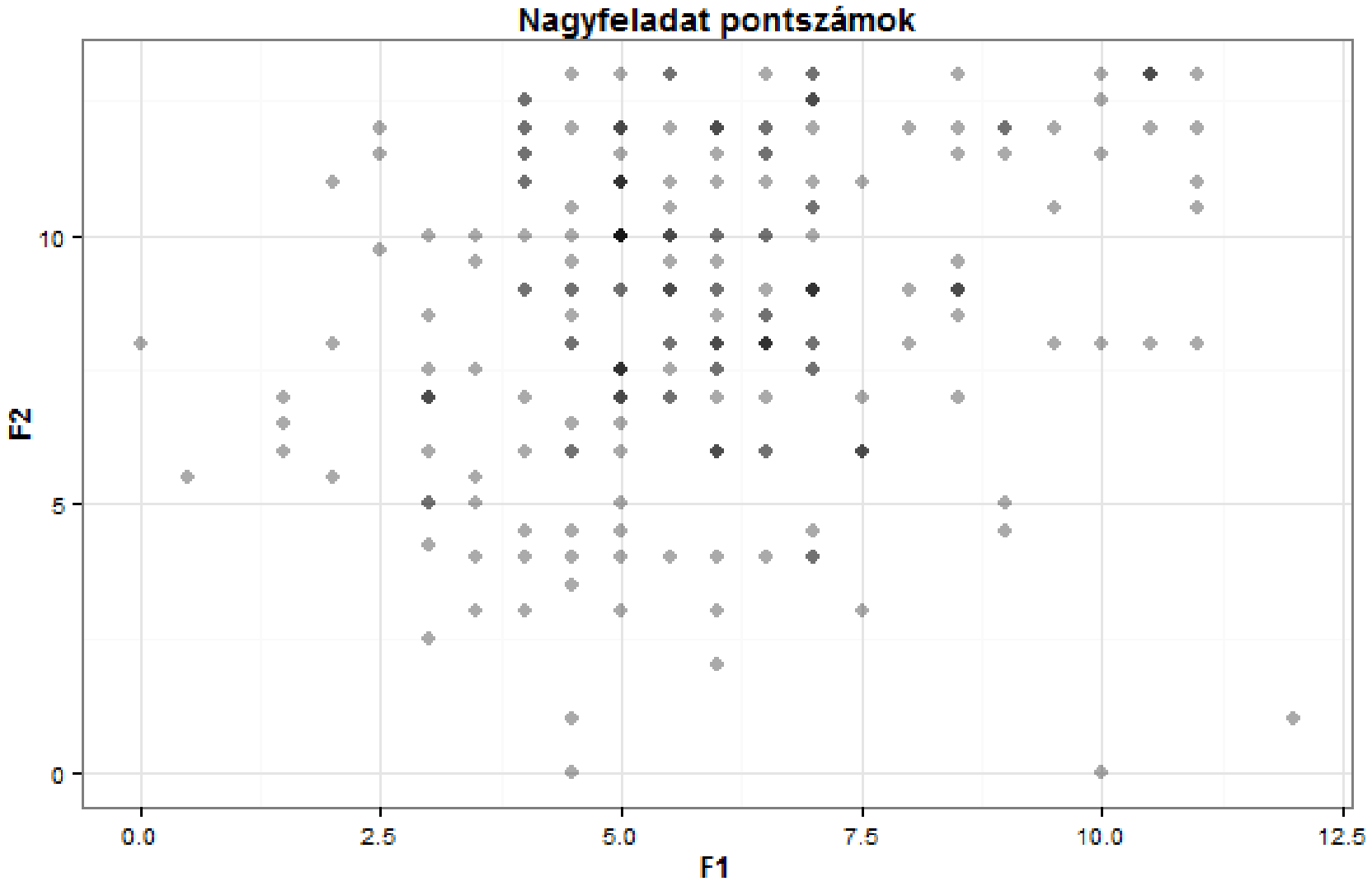
Hogyan kezeljük a takarásokat?

Overplotting

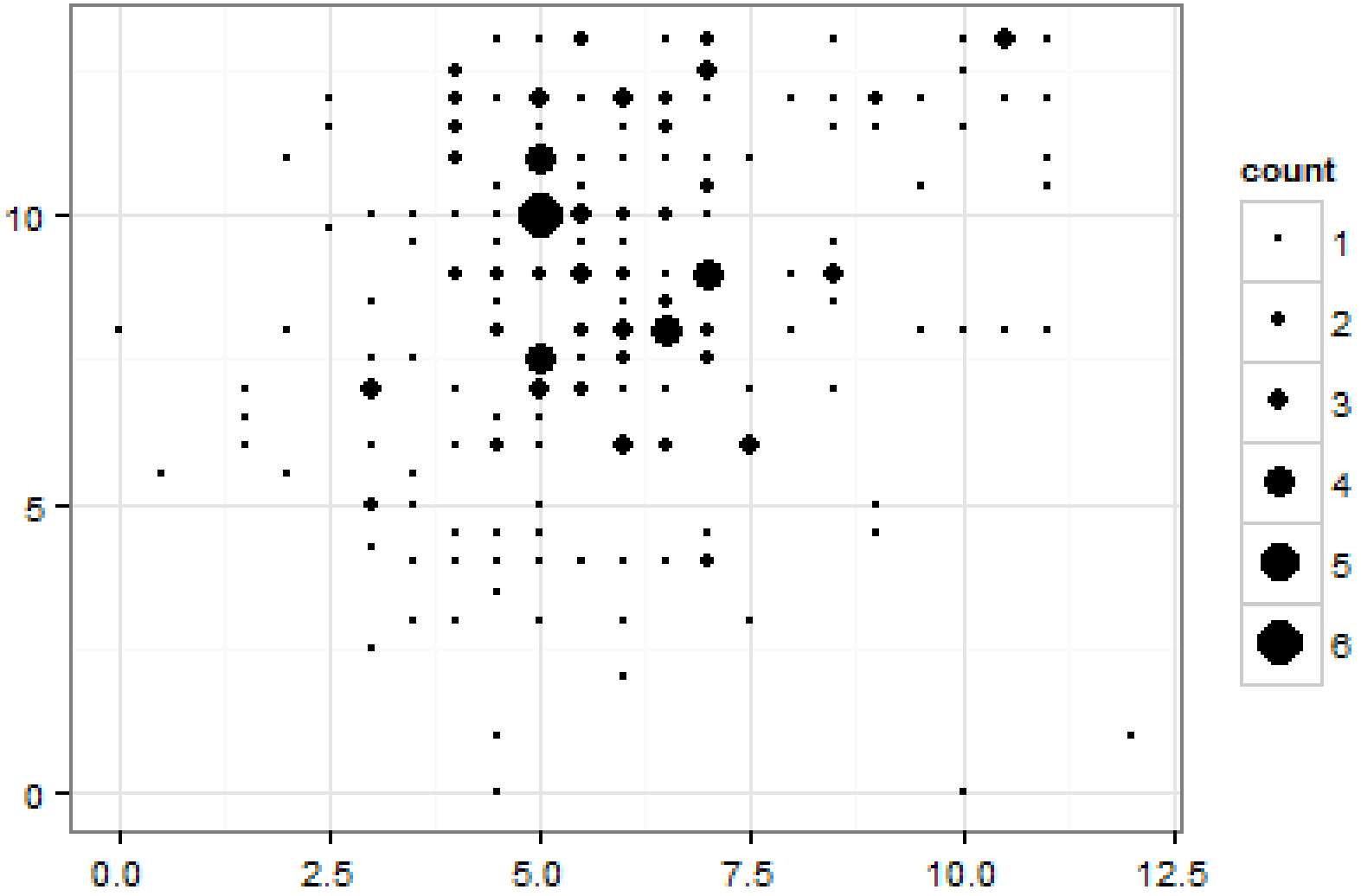
Overplotting megoldások 1: jitter



Overplotting megoldások 2: átlátszóság



Overplotting megoldások 3: méret



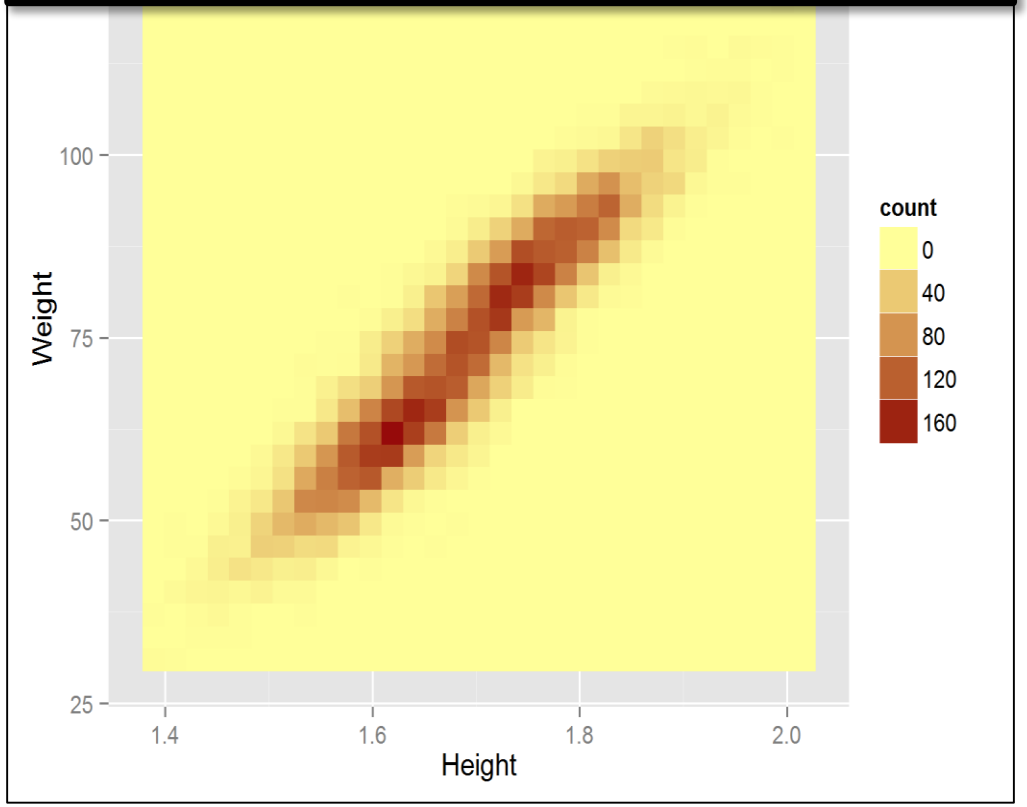
>=3 változó

- A grafikai objektumok attribútumait változtatom
  - Szín
  - Méret
  - Textúra
  - Hely – ez triviálisnak tűnik, de a treemapnél van jelentősége
- Pl. heatmap, treemap

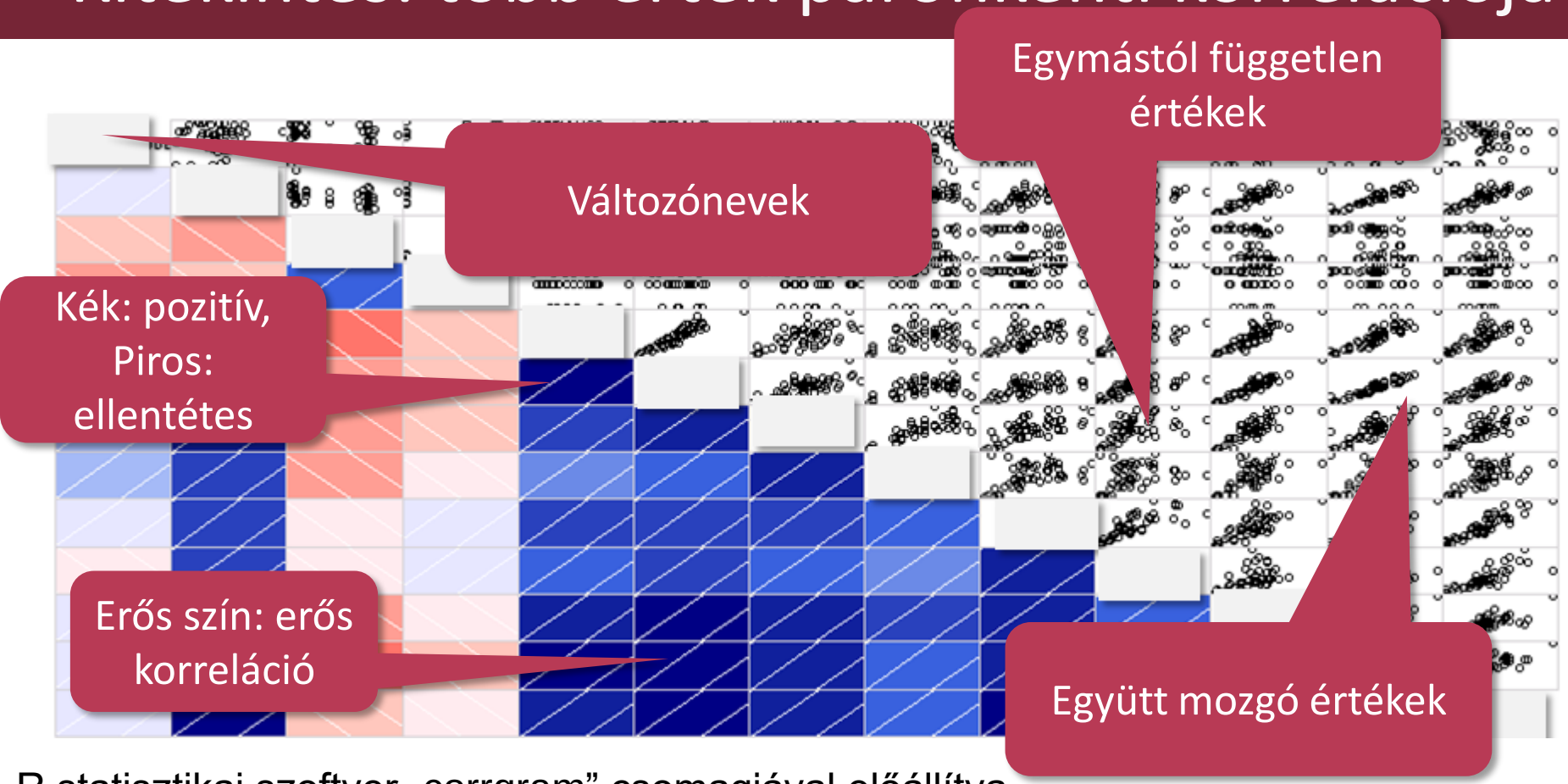
Hőtérkép (heat map)

- Megjelenített dim.k: 3
- Ábrázolt összefügg.:
  - *sűrű* 3D struktúrák összefüggései
- Adategység:
  - tile – azonos „magasságú” összefüggő területrész
- Tervezői döntés:
  - tile-ok mérete?

Színekkel kommunikál:
Pl. nincs senki, aki kétméteres lenne és 25 kiló, de sok 1.60-as van 60 kiló környékén

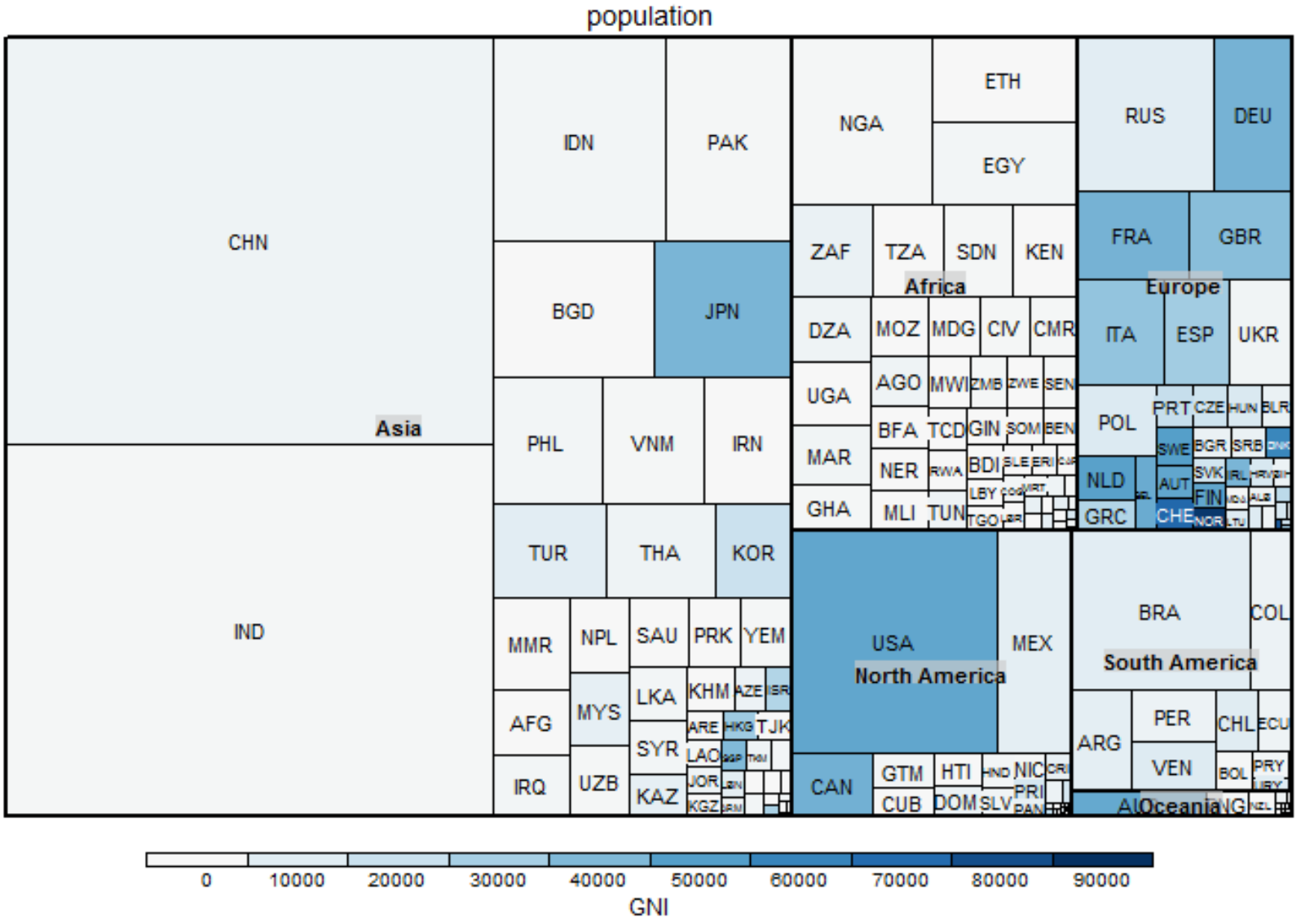


Kitekintés: több érték páronkénti korrelációja

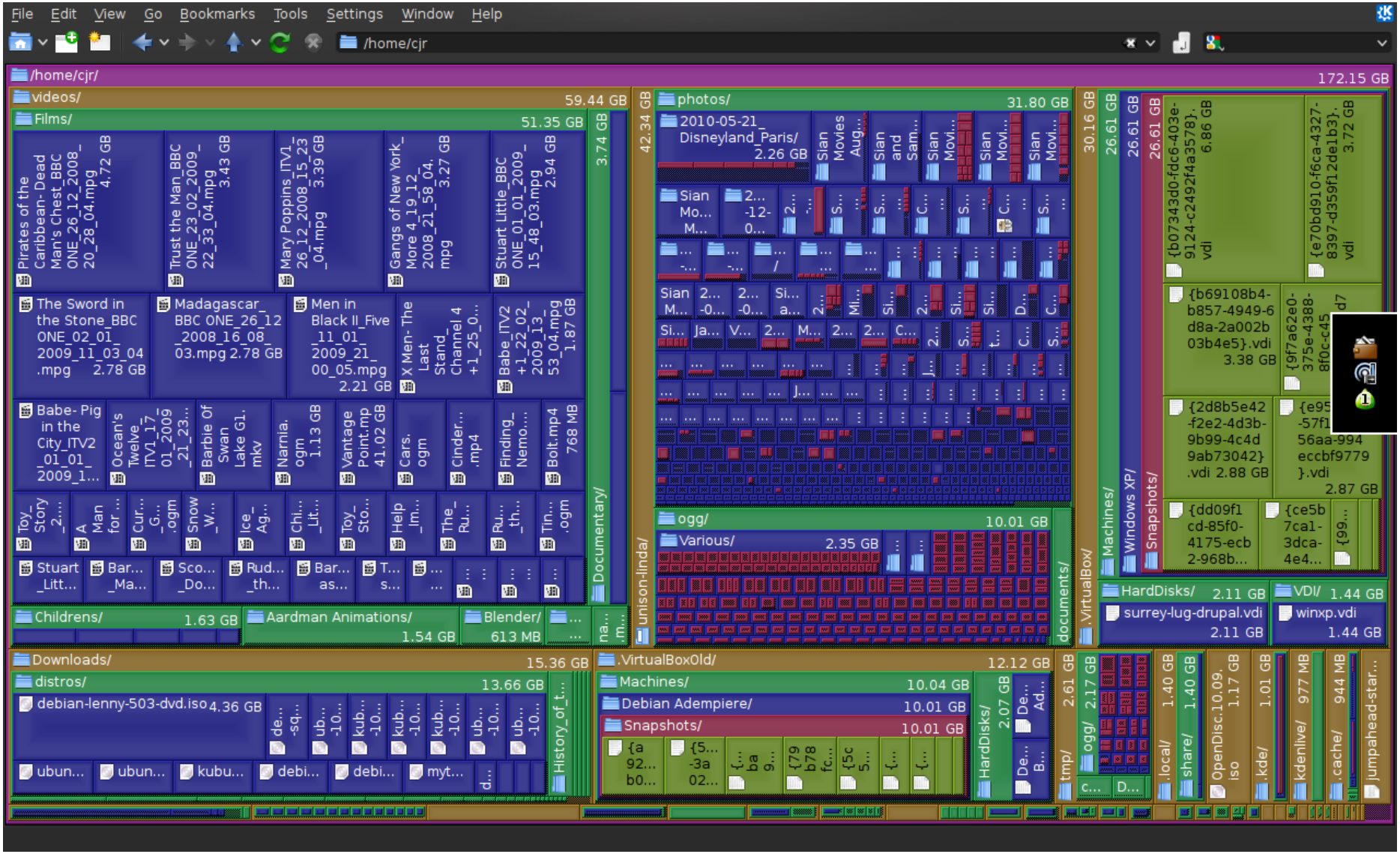


R statisztikai szoftver „corrgram” csomagjával előállítva.
Korreláció (ld. Valószínűségszámítás):
két érték közti lineáris kapcsolat erőssége és iránya
Átló felett: **scatterplot mátrix**
Cél: együtt mozgó értékek kiszűrése, **kiugró értékek (outlierek)** azonosítása.
→Mik a terhelés/előrejelzés szempontjából lényeges változók?

Treemap

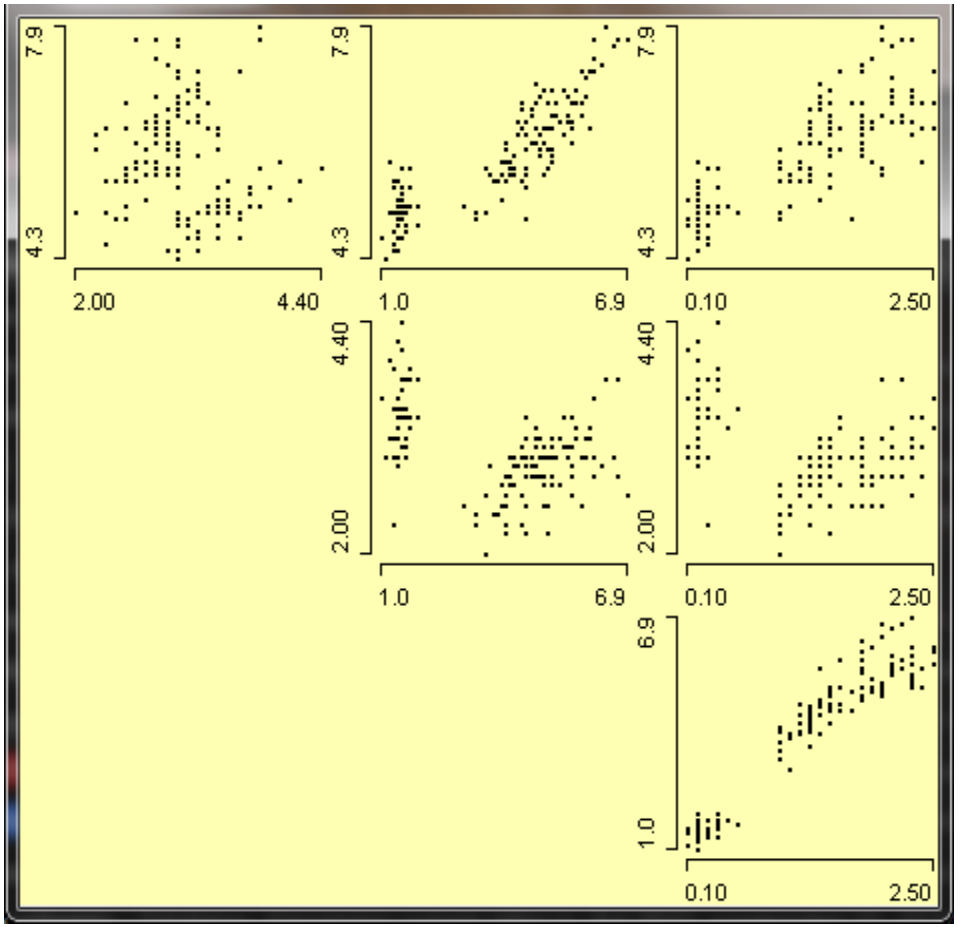

population
CHN
Asia
IND
IDN
PAK
BGD
JPN
PHL
VNM
IRN
TUR
THA
KOR
MMR
NPL
SAU
PRK
YEM
AFG
MYS
LKA
KHM
AZE
ISR
ARE
HKG
TJK
SYR
LAO
IRQ
UZB
KAZ
JOR
KGZ
NGA
ETH
EGY
ZAF
TZA
SDN
KEN
Africa
DZA
MOZ
MDG
CIV
CMR
UGA
AGO
MWI
ZMB
ZWE
SEN
MAR
BFA
TCD
GIN
SOM
BEN
NER
RWA
BDI
GHA
MLI
TUN
LBY
TGO
USA
North America
MEX
CAN
GTM
HTI
HND
NIC
CUB
DOM
SLV
PRI
RUS
DEU
FRA
GBR
Europe
ITA
ESP
UKR
POL
PRT
CZE
HUN
BLR
SWE
BGR
SRB
NLD
AUT
SVK
FIN
GRC
CHE
NOR
BRA
COL
South America
ARG
PER
CHL
VEN
BOL
Oceania
0
10000
20000
30000
40000
50000
60000
70000
80000
90000
GNI

Treemap: állományrendszer



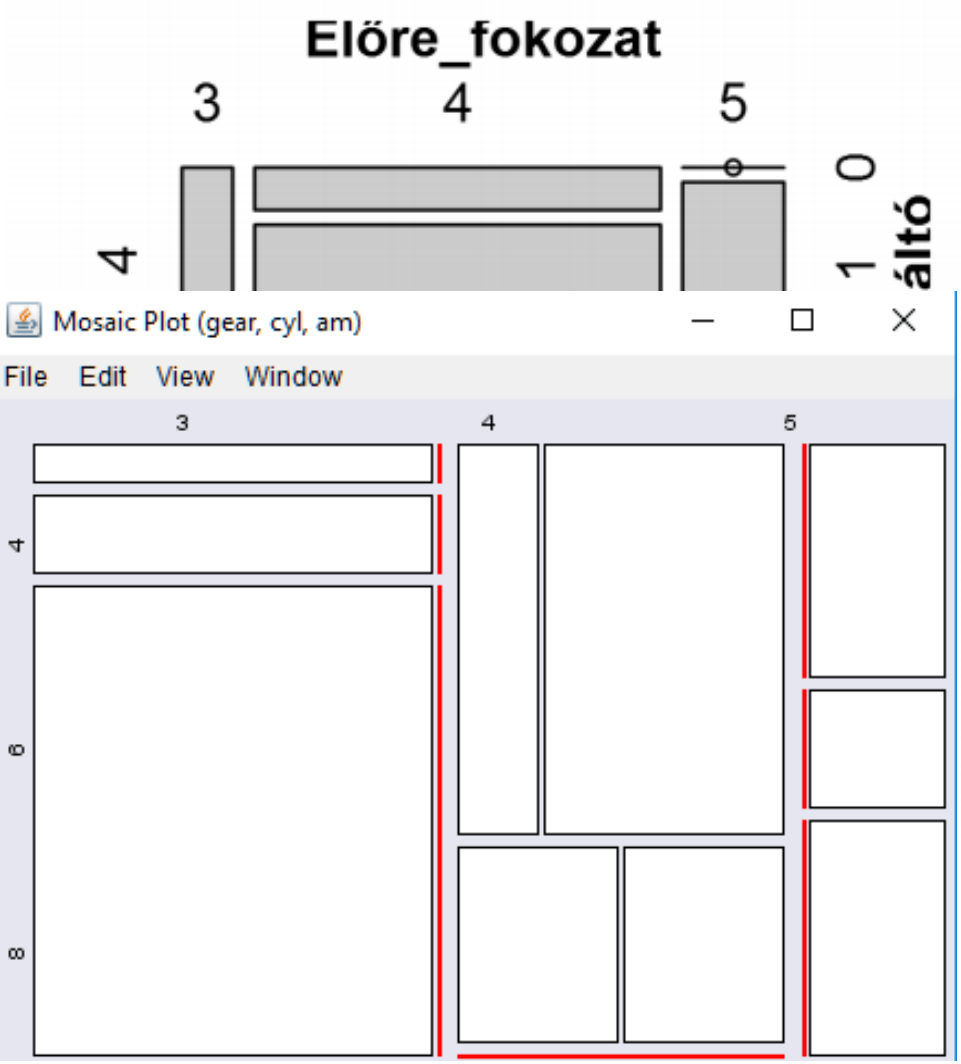
Scatterplot mátrix

- Megjelenített dim.k: n
- Ábrázolt összefügg.:
  - A változópárok együttes eloszlása
- Adategység:
  - Scatterplot – minden diagram a neki megfelelő változók együttes eloszlását mutatja be

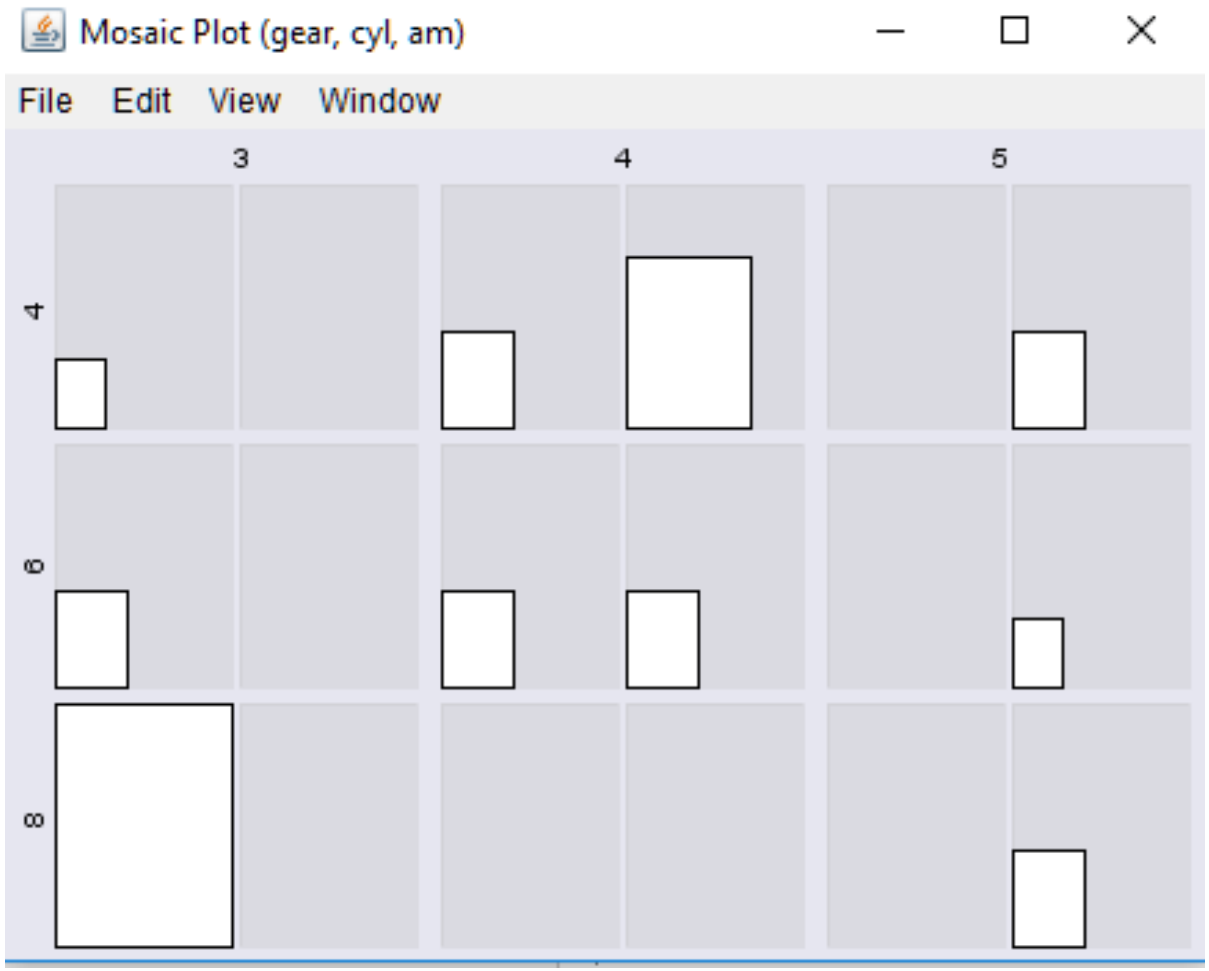


Mozaik diagram (mosaic plot)

- Megjelenített dim.k: 2
- Ábrázolt összefügg.:
  - két diszkrét változó együttes eloszlása
- Adategység:
  - Téglalap – a téglalap *területe* arányos az (X = xi, Y = yi) értékpárok gyakoriságával
- Korlát:
  - Sorfolytonos olvasása nehézkes

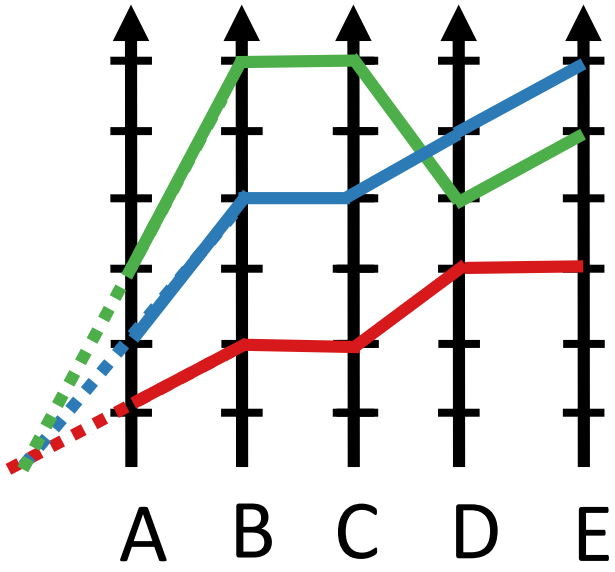


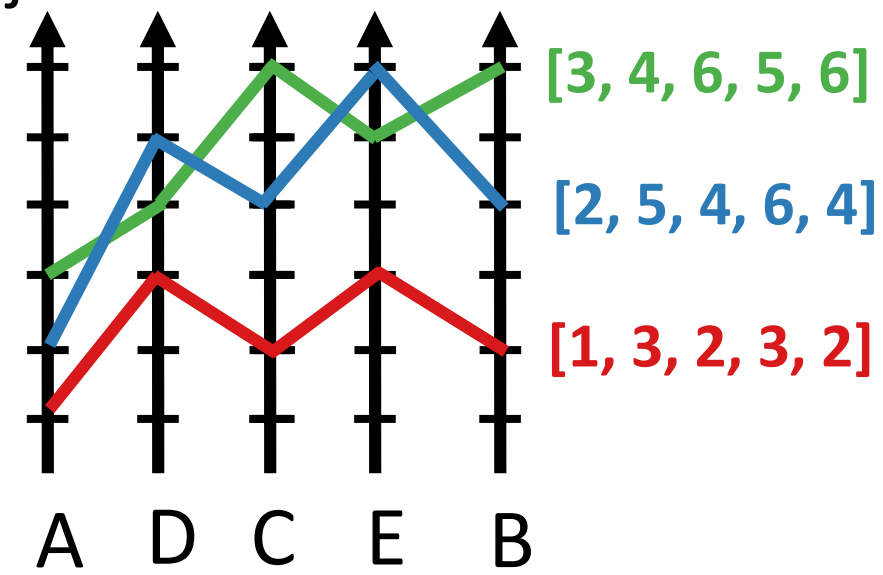
Változat: fluktuációs diagram



Párhuzamos koordináták

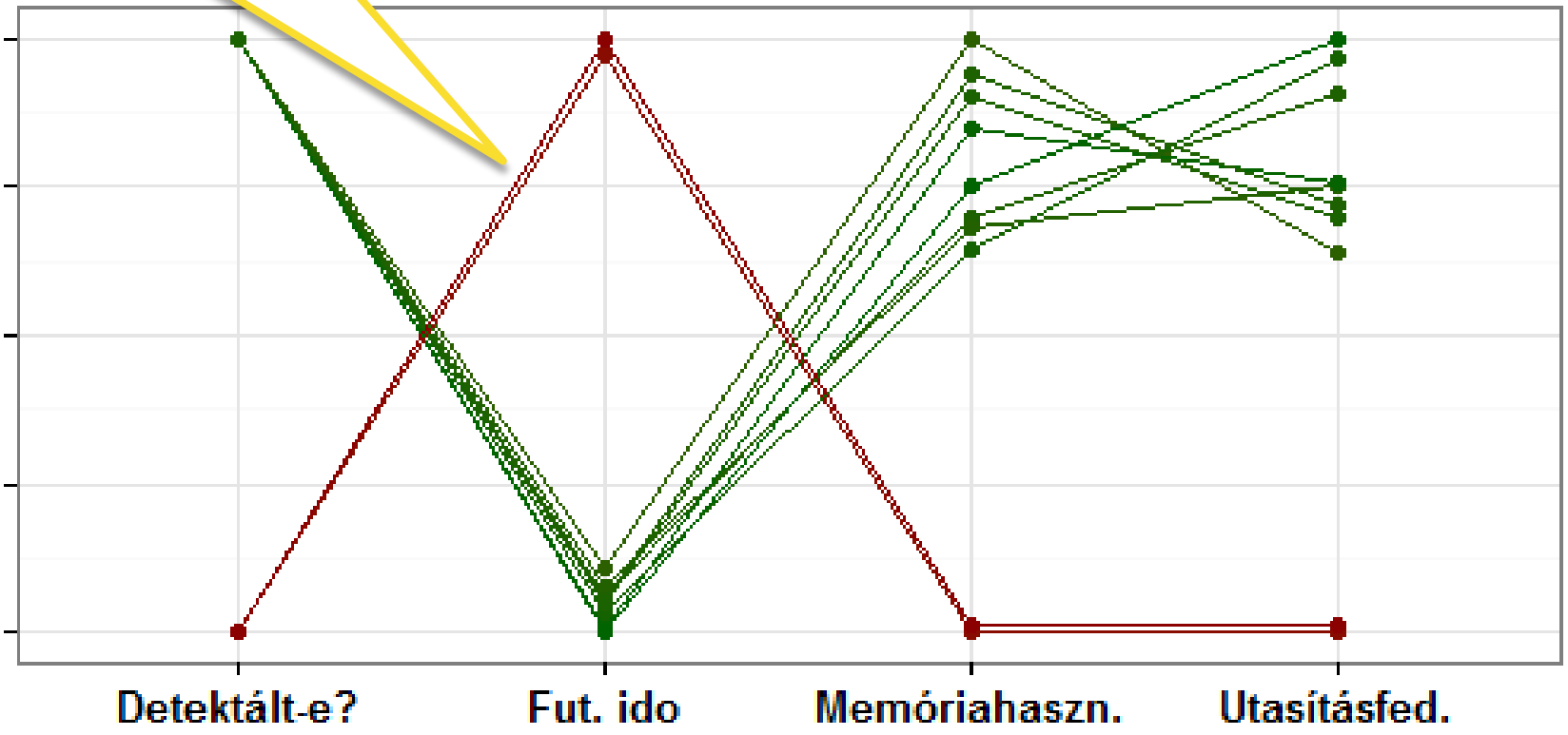
- Tengelyek: dimenziók/koordináták
  - tetszőleges számú
  - tetszőleges skála
- Egy vonal egy mérés (darabszám?)
- Kompakt és skálázható
- Koordináta sorrend befolyásolja a kiértékelést





Párhuzamos koordináták: tesztesetek elemzése

1 teszteset 1 törött vonal

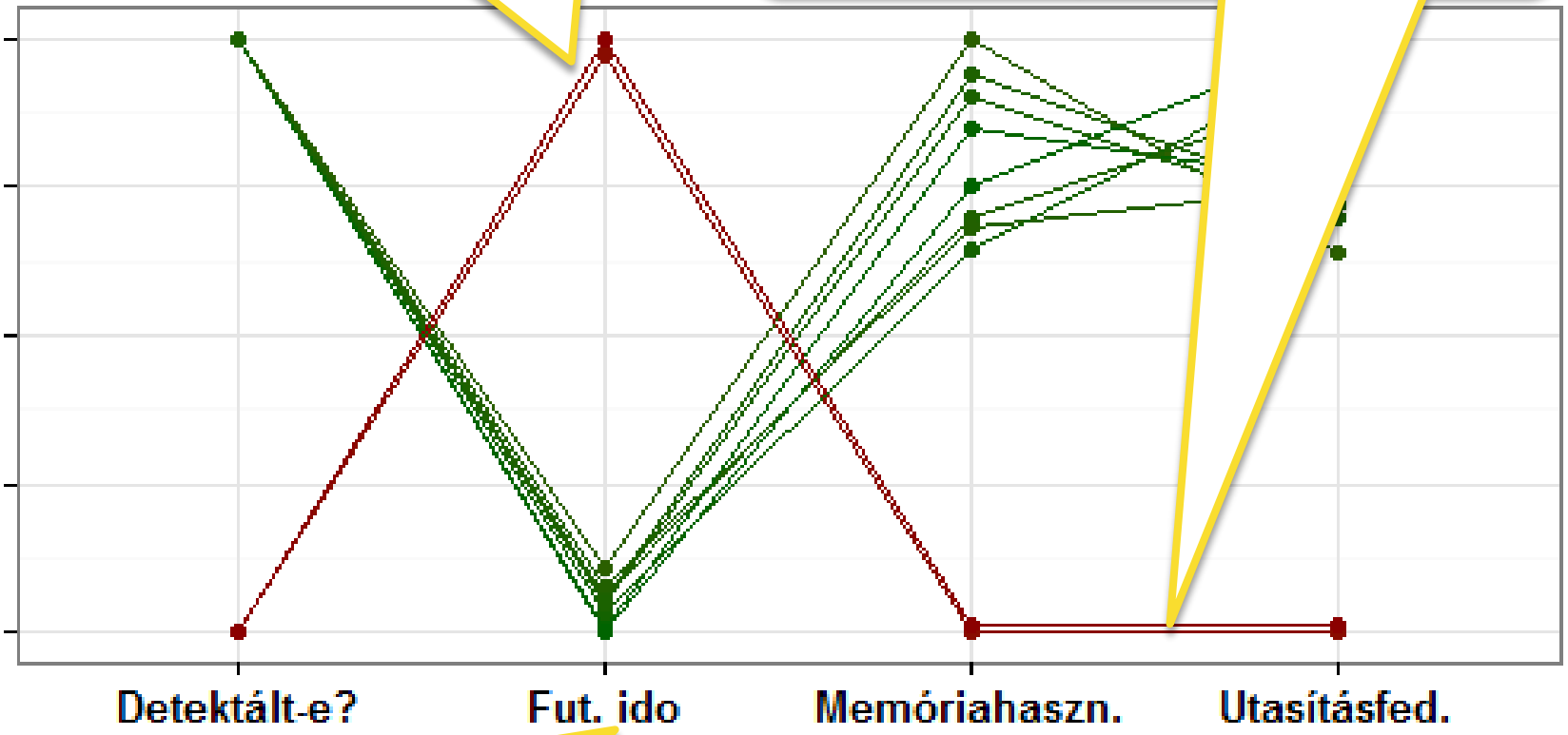


A változók az x tengelyen jelennek meg

Párhuzamos koordináták: tesztesetek elemzése

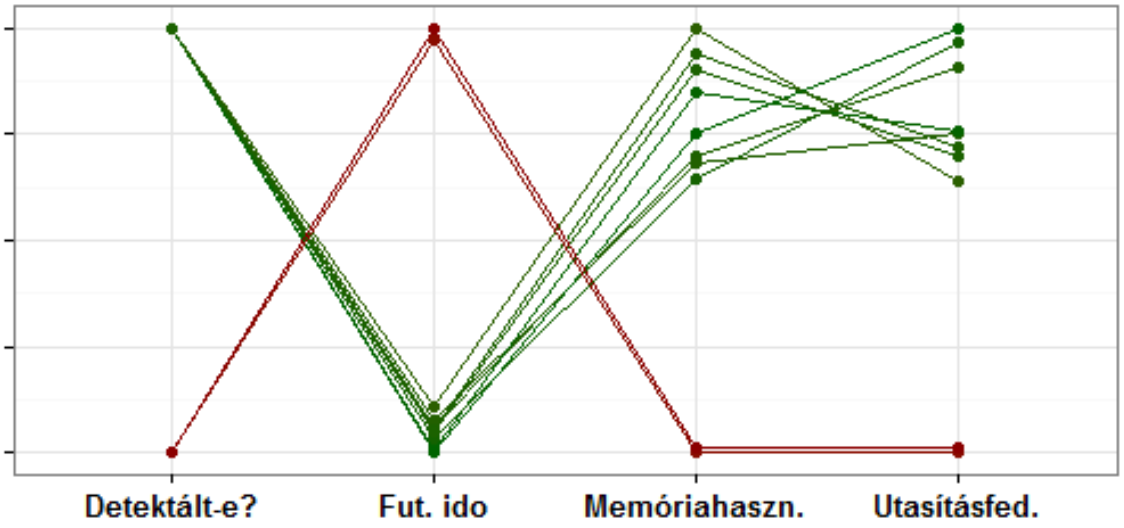
Timeout?

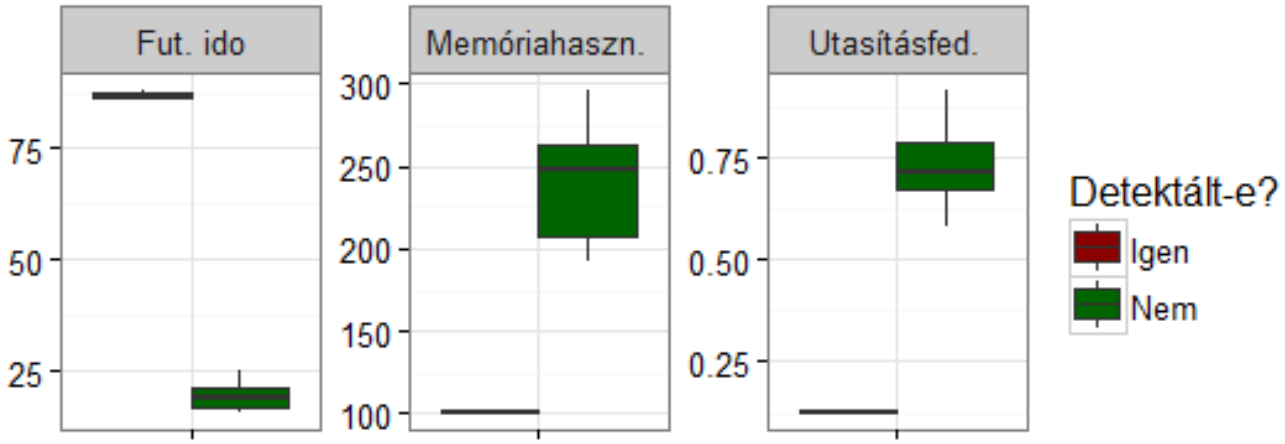
A hibát detektálók az érdemi számításig valószínűleg el sem jutnak



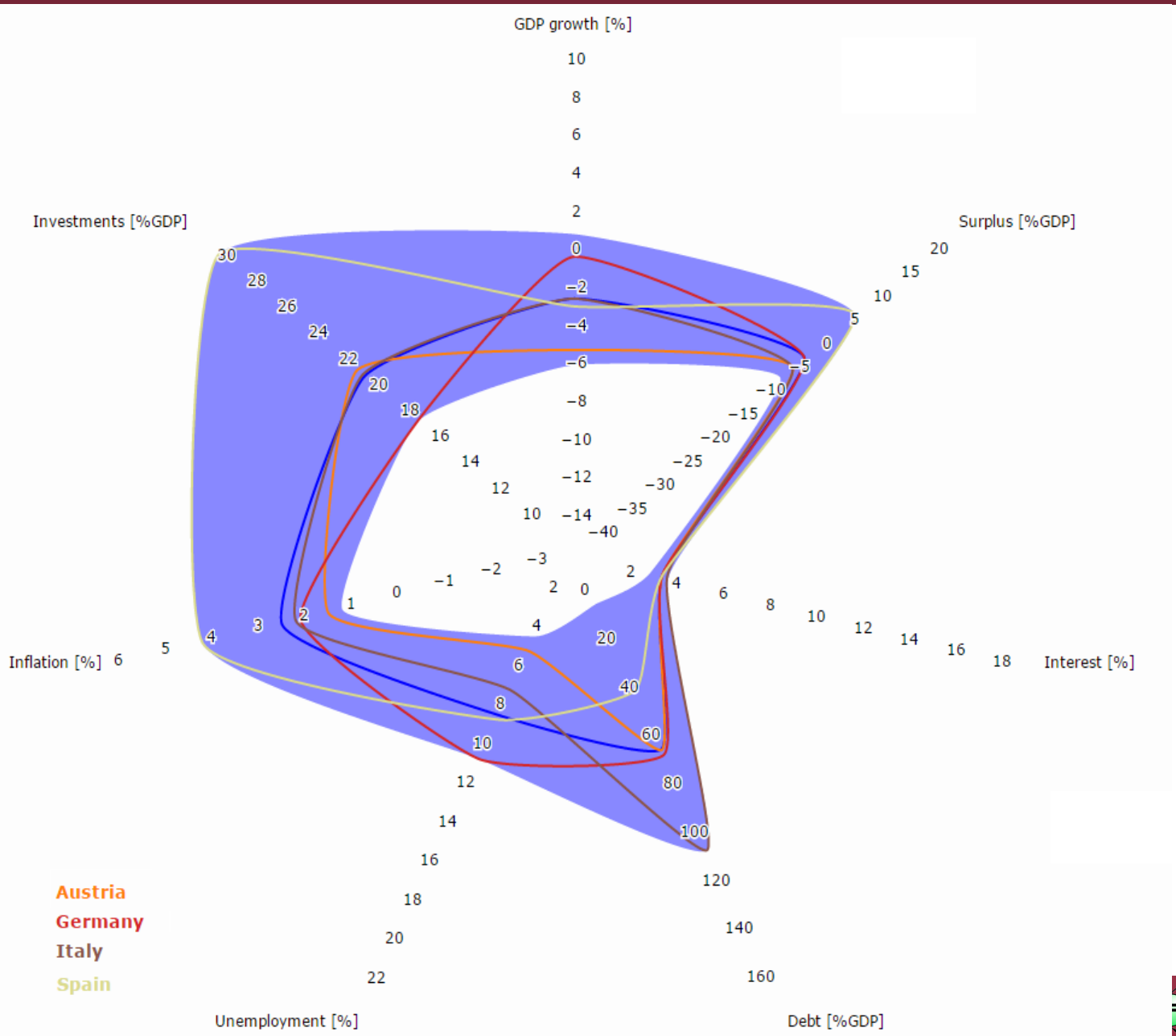
A futási idő és a memóriahasználat valószínűleg pozitív kapcsolatban állnak (sikeres teszteknél)

Párhuzamos koordináták: viz. alternatívák

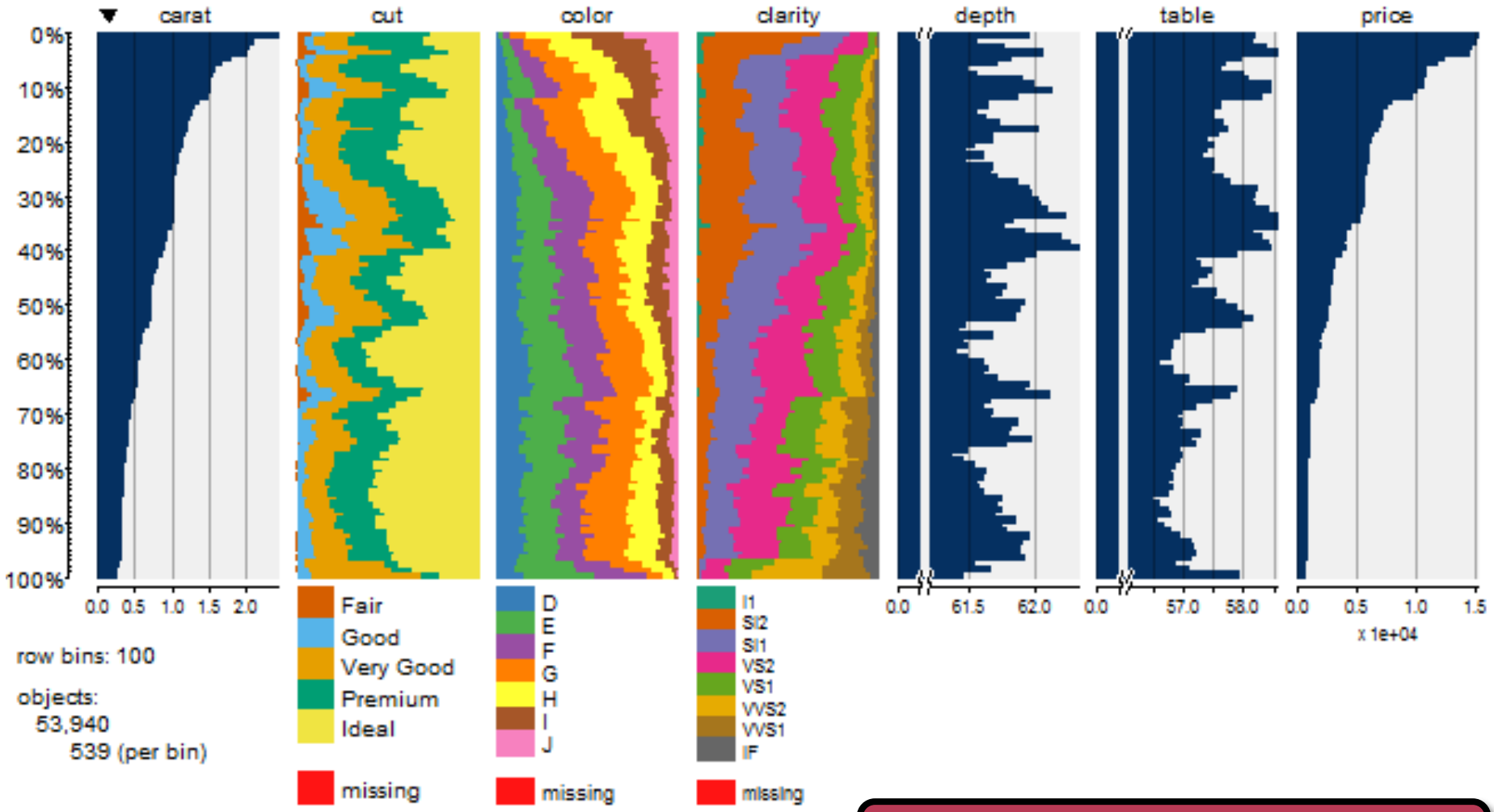




Radar chart: egy párhuzamos koord. kiterjesztés



Tableplot



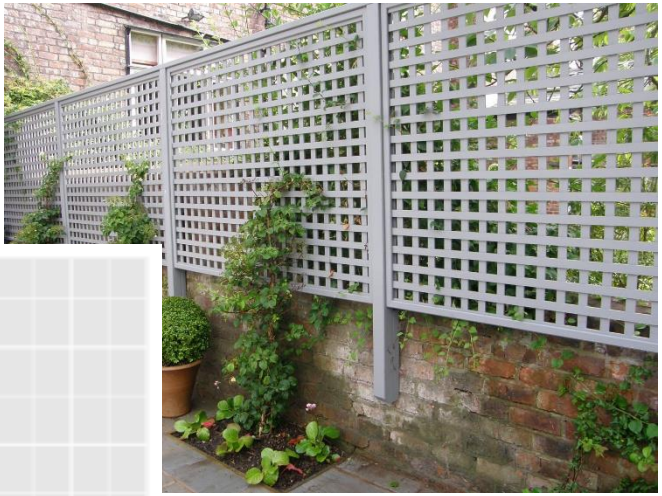
Small demo: itableplot

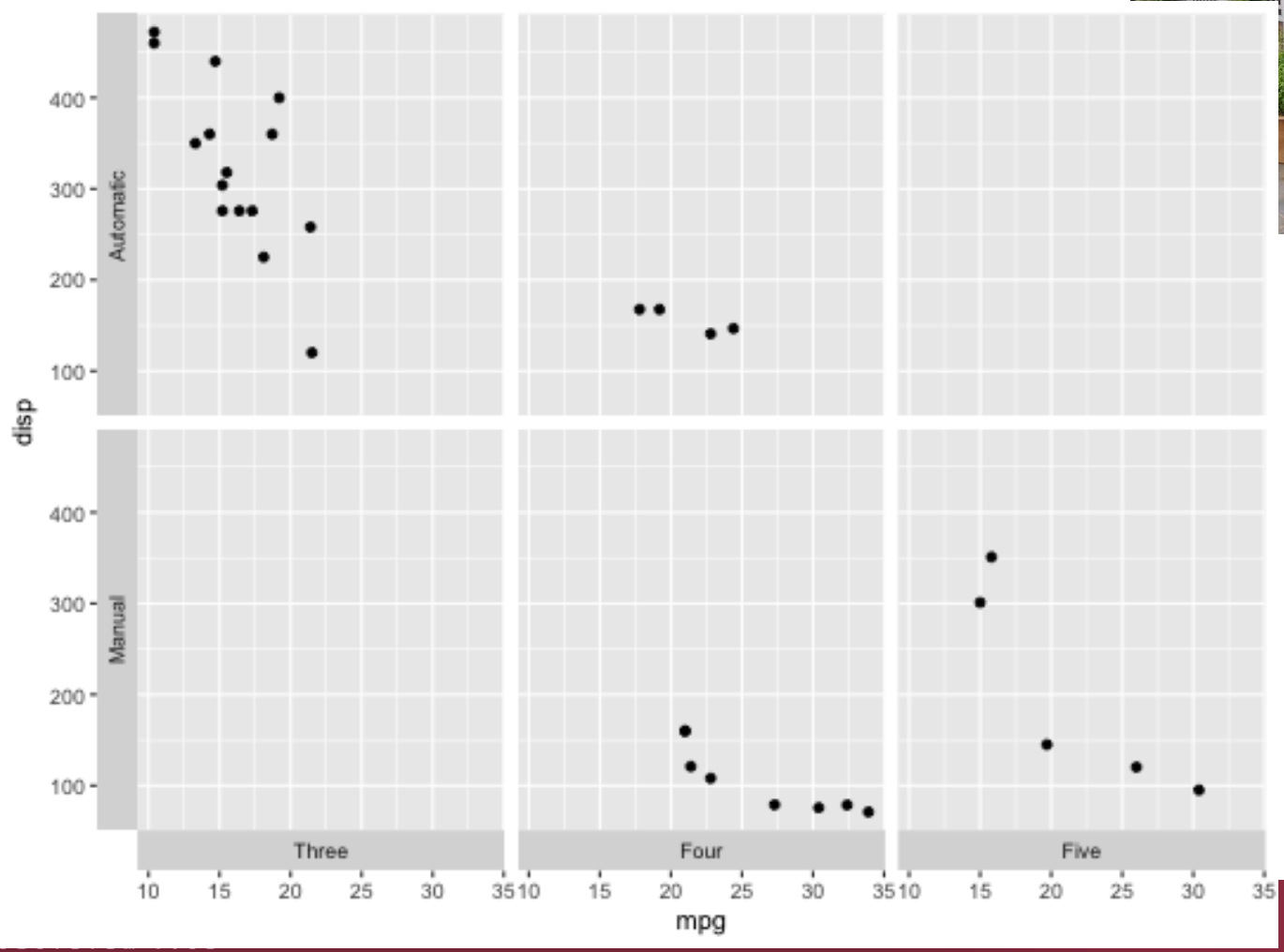
Demo

- R, tabplot package
- Attach data frame, call itableplot()
- iris, diamonds (from ggplot2)

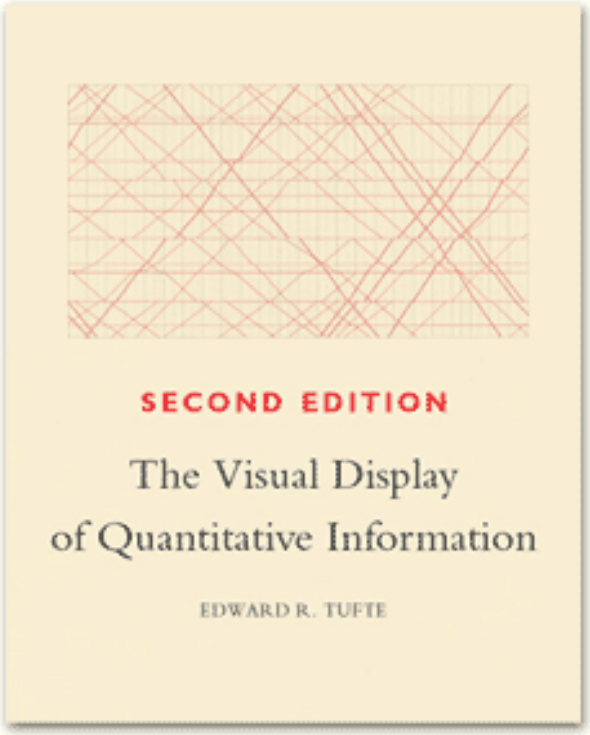
Trellis plot/Crossplot/„facets”/„small multiples”

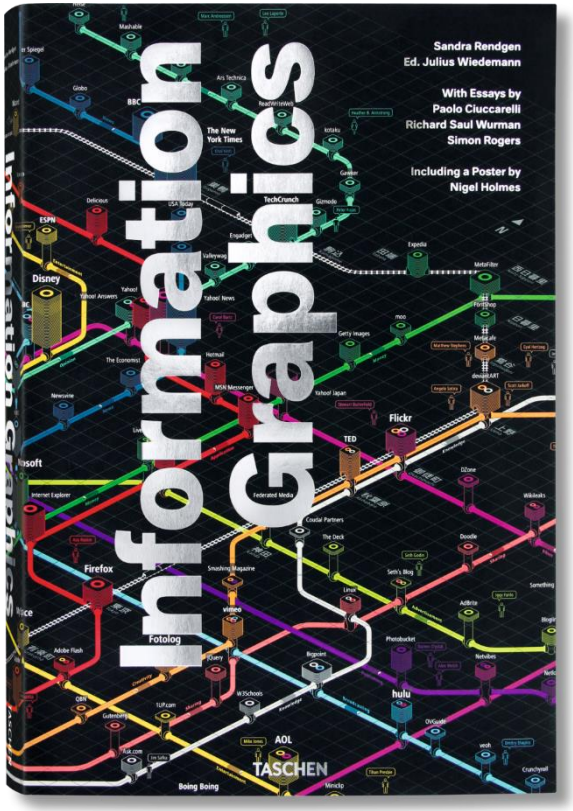
- Egy vagy több változó értékére kondícionálás
- Panelek

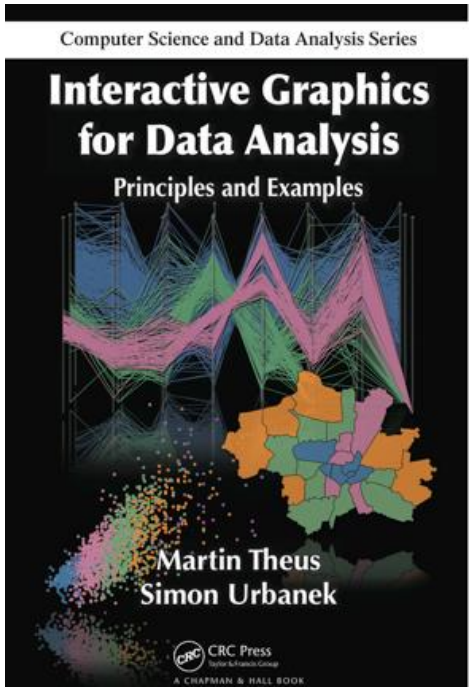




Javasolt olvasmányok







http://library.duke.edu/data/data-visualization