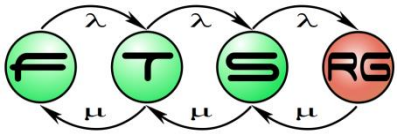
Vizuális adatelemzés

Rendszermodellezés 2016.

Budapest University of Technology and Economics
Fault Tolerant Systems Research Group



Mi is lesz?

- Miért vizualizálunk?
- Mit vizualizálunk?
- Hogyan vizualizálunk?
- Mire következtetünk?

Mi is lesz?

- Miért vizualizálunk?
- Mit vizualizálunk?
- Hogyan vizualizálunk?
- Mire következtetünk?

A vizualizáció hétköznapjai

Analóg megjelenítés

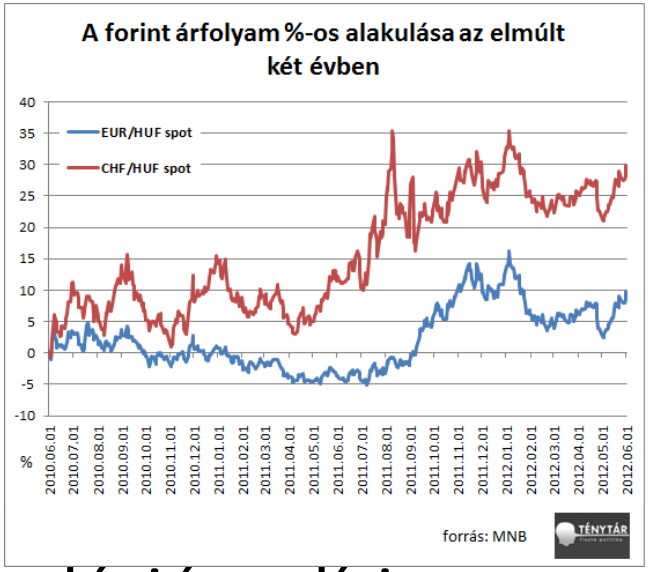
Digitális megjelenítés

Analóg + koord. rei

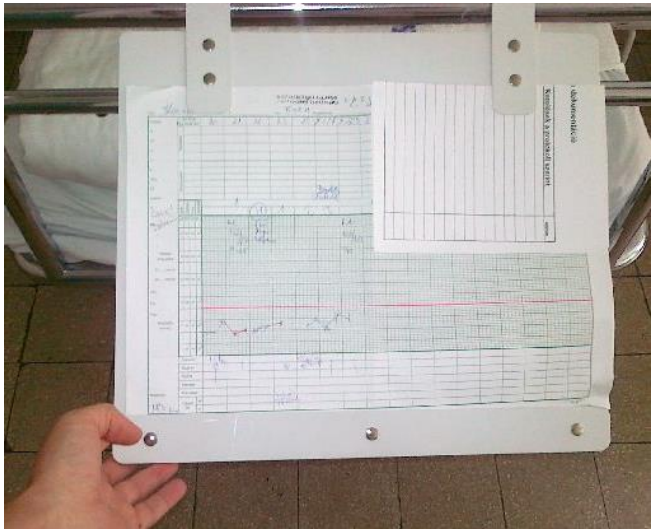
id megjelenítés

A vizualizáció hétköznapjai

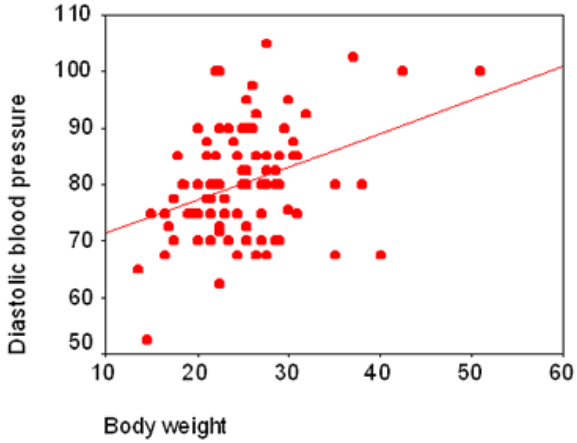
Trend analízis és előrejelzés



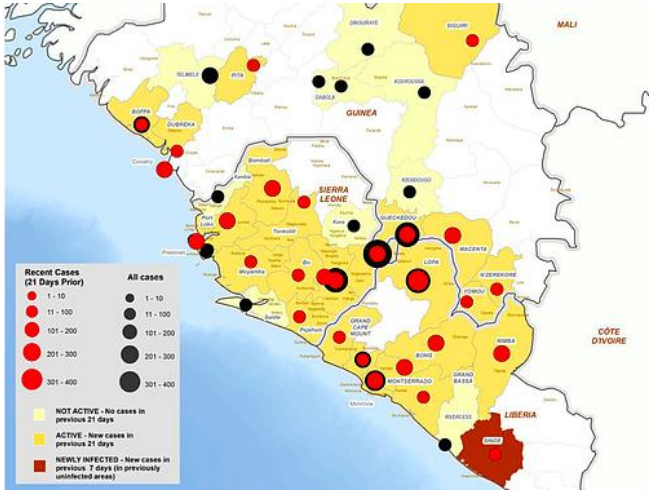
Idősor analízis



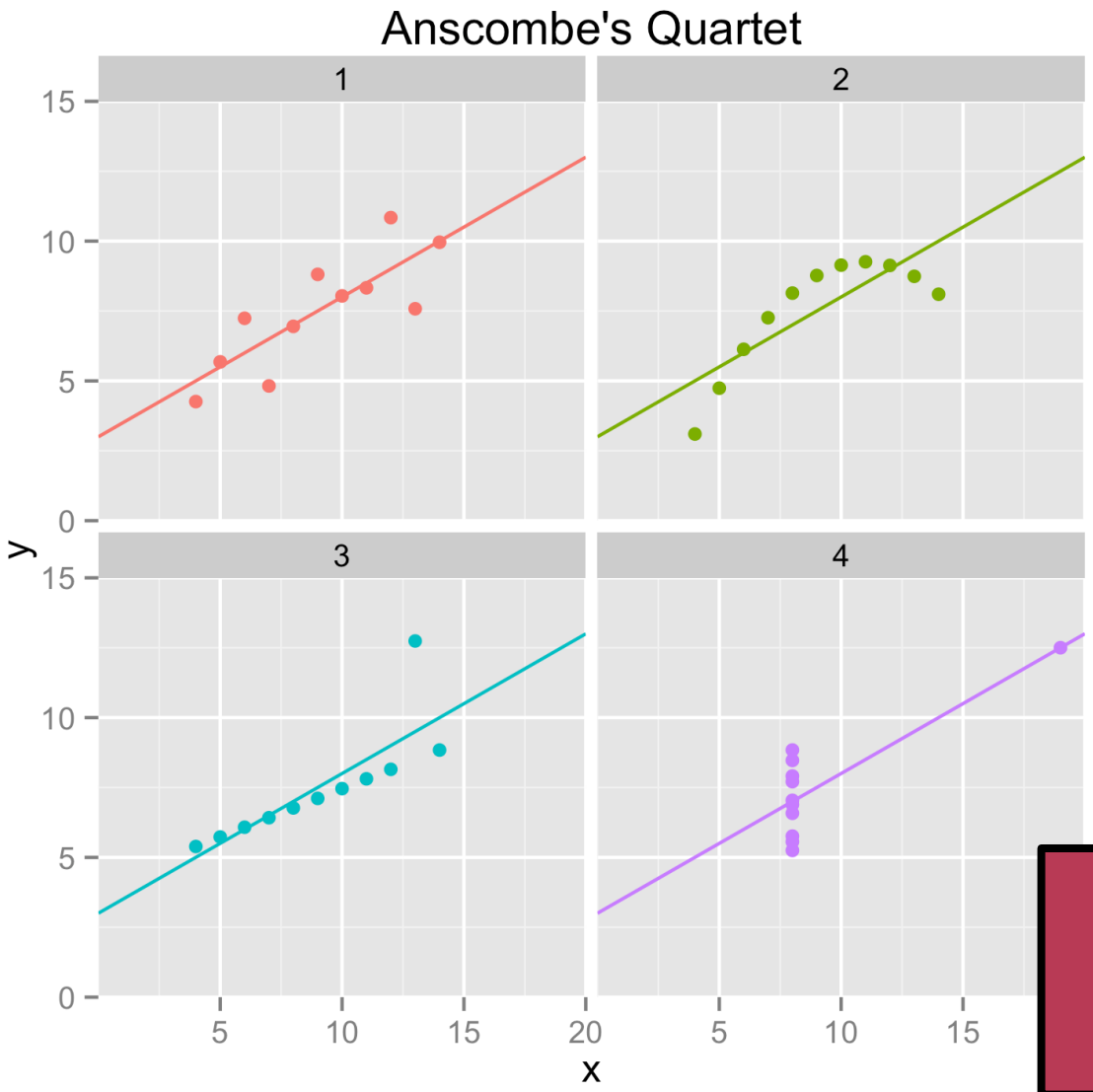
Korrelációanalízis



Térbeli analízis

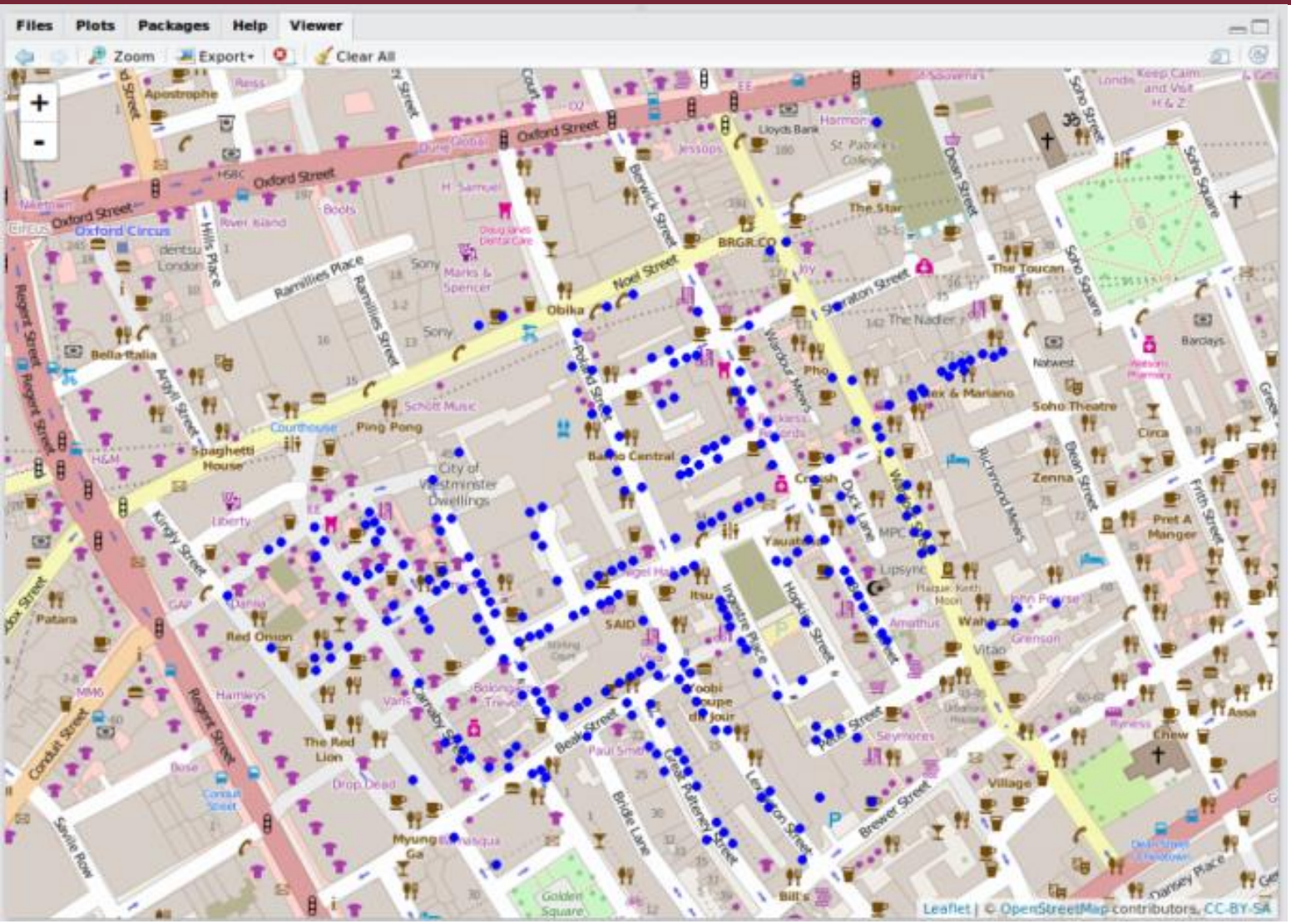


Számítások ellenőrzése



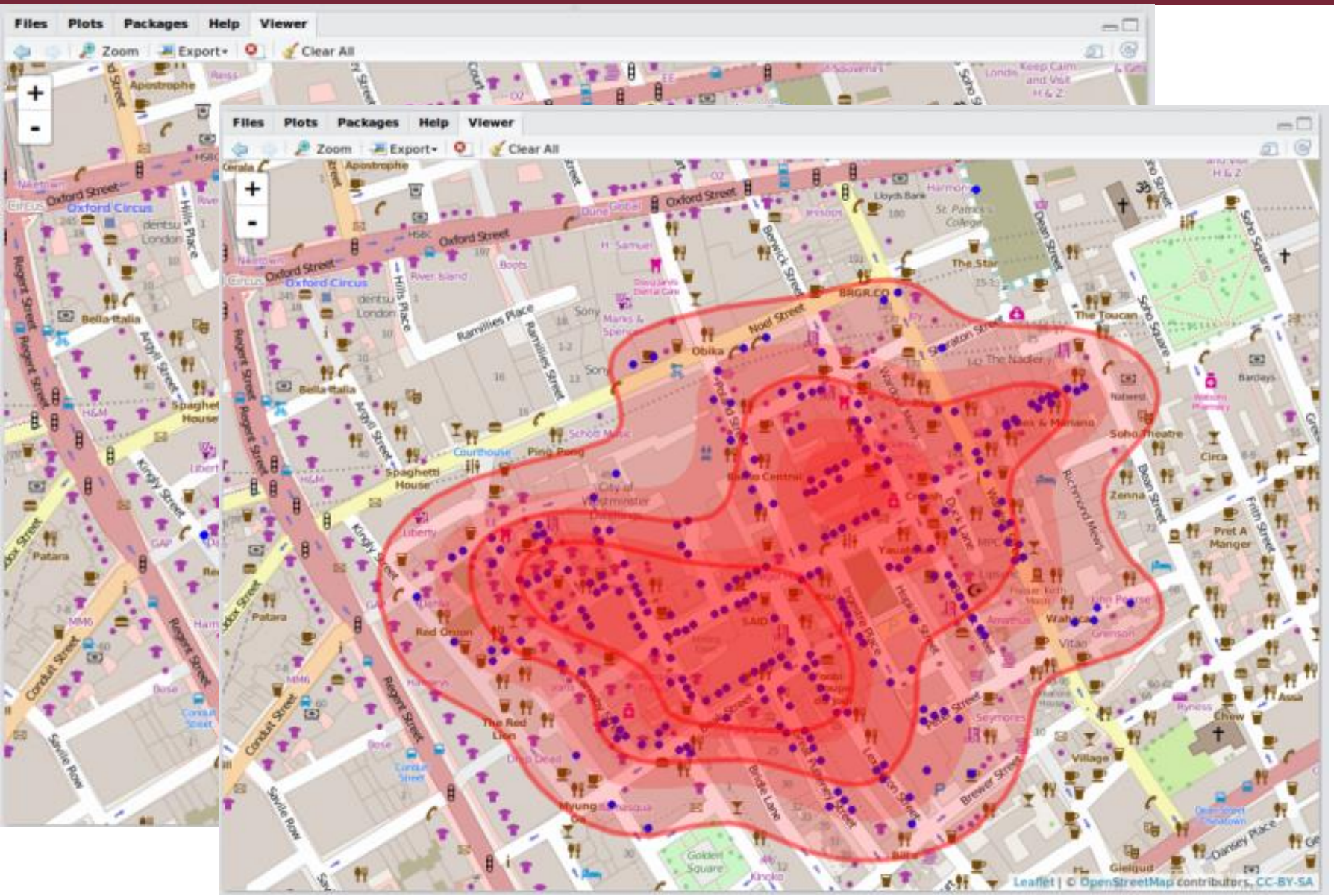
Hibás feltételezések elkerülése… és intuíció

Összefüggések feltárása



Forrás: https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak

Összefüggések feltárása

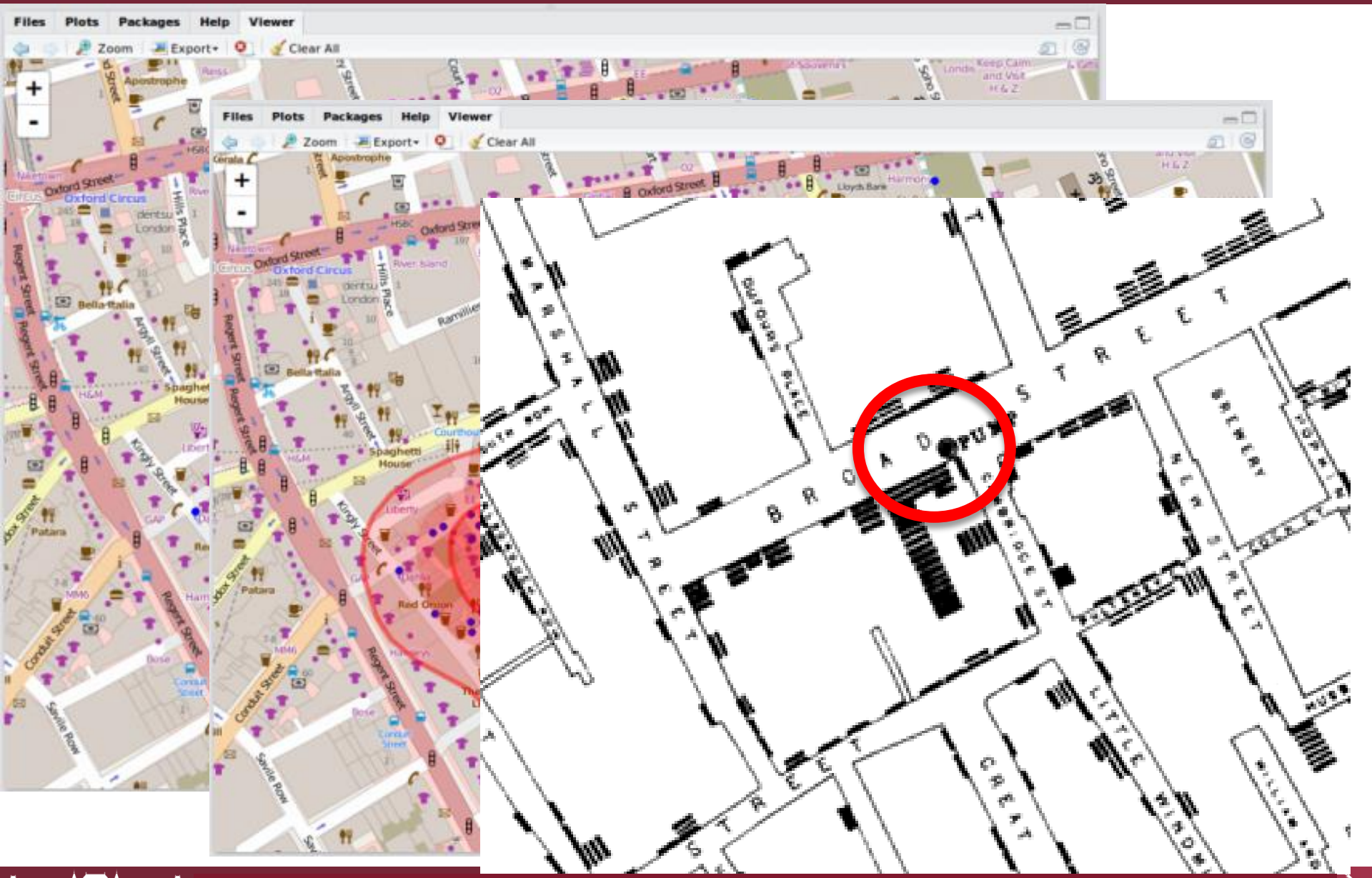


Forrás: https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak

Összefüggések feltárása

Forrás: https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak

Összefüggések feltárása



Forrás: https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak

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

Mi is lesz?

Miért vizualizálunk?

Mit vizualizálunk?

Hogyan vizualizálunk?

Mire következtetünk?

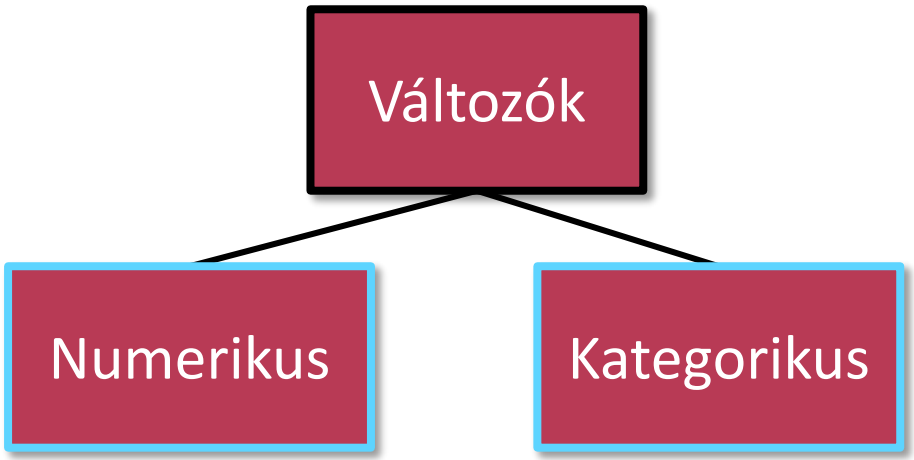
Emlékeztető: táblázatos ábrázolás

- **Táblázat sora** = modellelem
- **Táblázat oszlopa** = tulajdonság

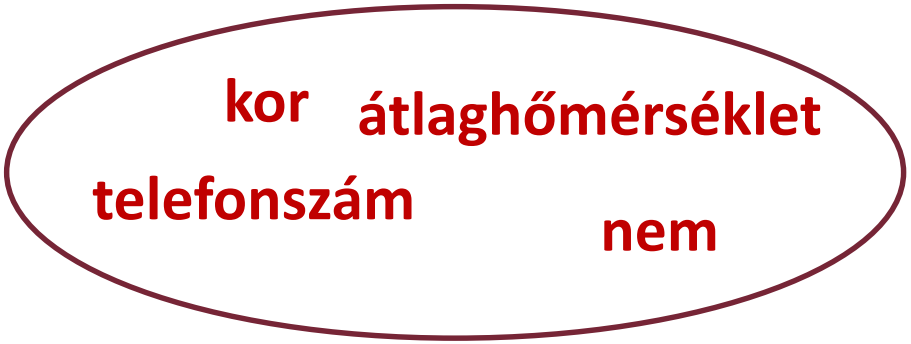
Név	Típus	Méret (kB)	Utolsó módosítás
Dokumentumok	könyvtár		2016.02.02
szerződés.pdf	fájl	569	2015.11.09
Képek	könyvtár		2016.02.02
logó.png	fájl	92	2015.03.06
alaprajz.jpg	fájl	1226	2016.02.02

Numerikus és kategorikus változók

- Numerikus (numerical)
  - az alapvető aritmetikai műveletek értelmesek

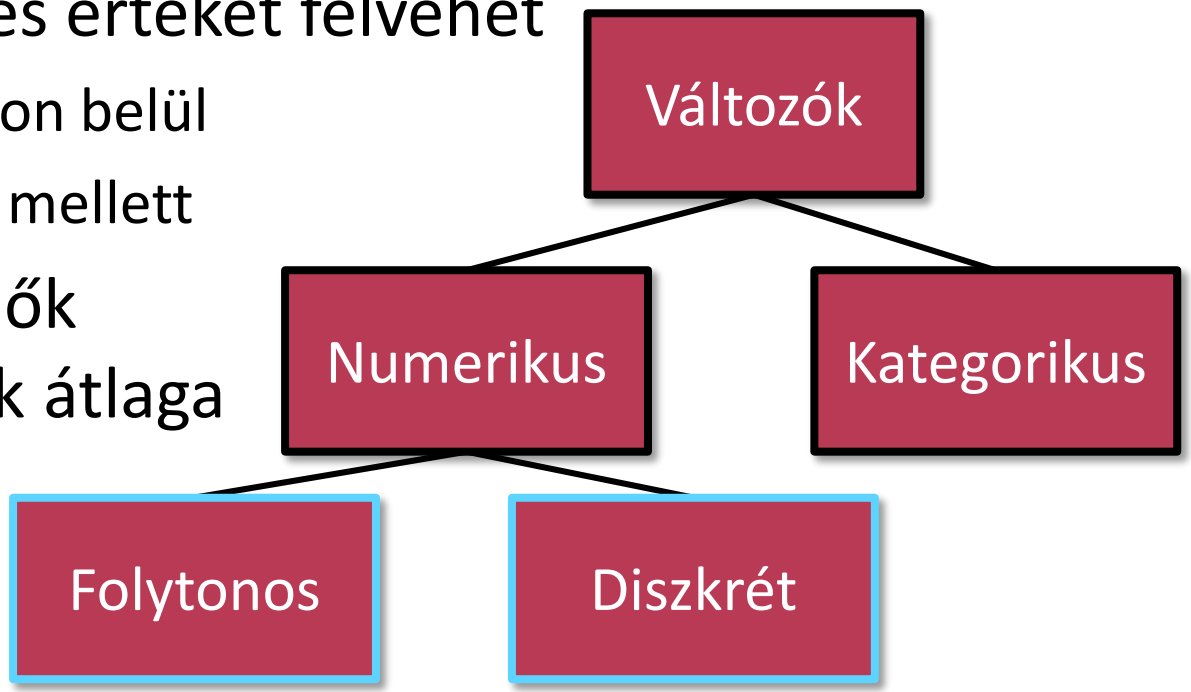


- Kategorikus (categorical)
  - Matematikai műveletek nem értelmezhetőek rajtuk, legfeljebb sorba rendezés



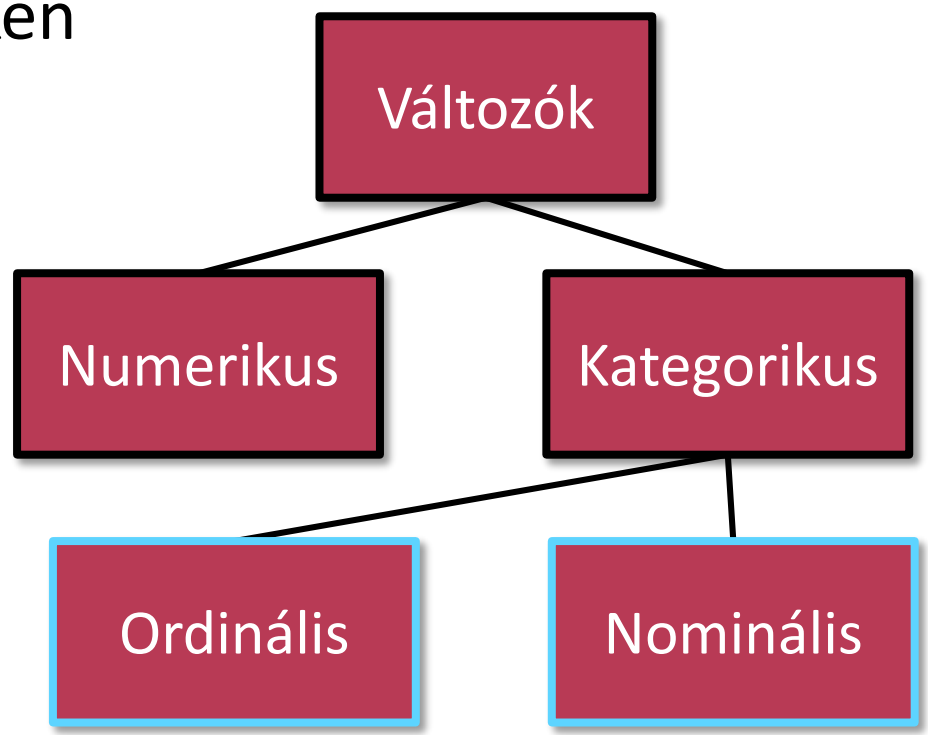
Numerikus változók

- 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

Mi is lesz?

Miért vizualizálunk?

Mit vizualizálunk?

Hogyan vizualizálunk?

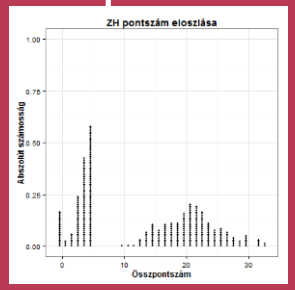
Mire következtetünk?

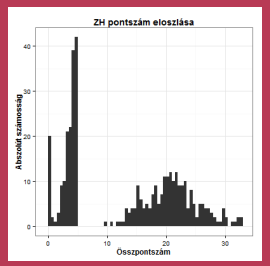
1 változó – eloszlásokra

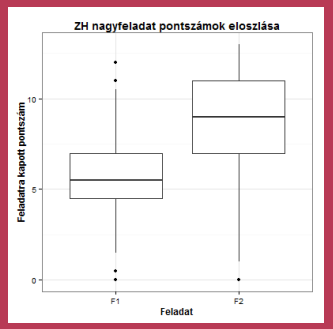
Változók

Numerikus

ZH pontszám: [13, 15, 2, …]

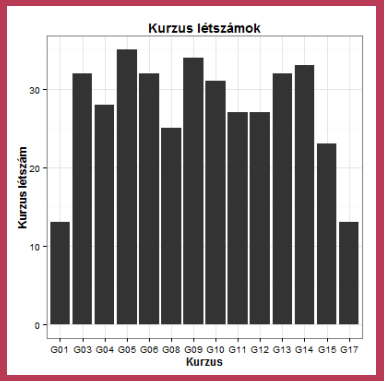






Kategorikus

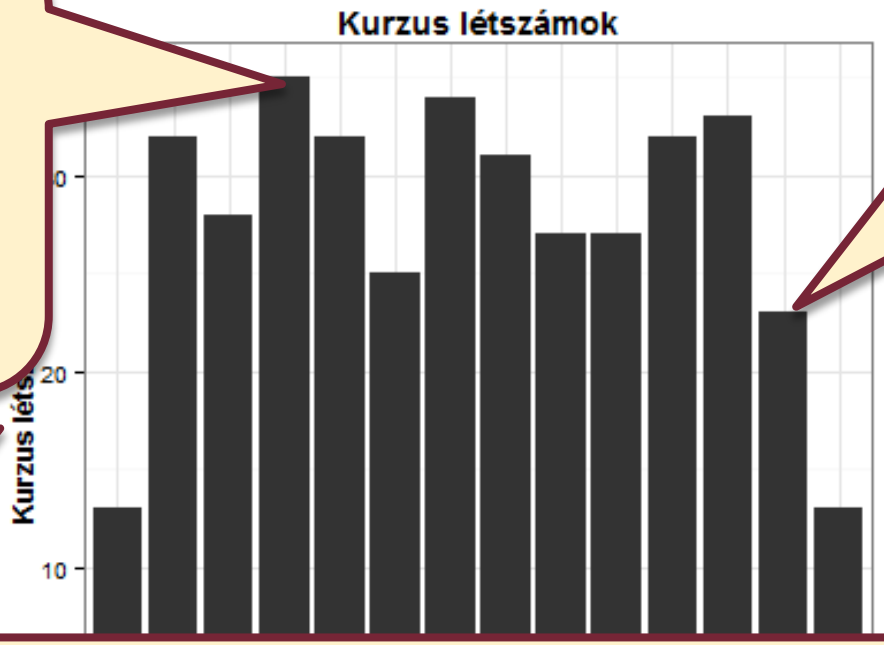
Kurzus: [G01, G03, G15, G17, …]



Oszlopdiagram

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

Vannak nagyon népszerű időpontok/ gyakvezek?



Oszlop-magasság: adott érték gyakorisága

abszolút gyakoriság!

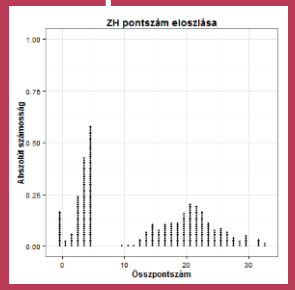
Tervezői döntés: értékkészlet darabolása
Pl.: kedd-csütörtök-péntek

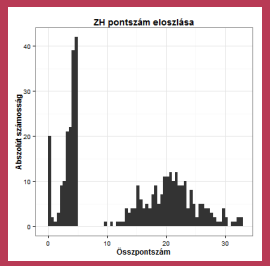
1 változó – eloszlásokra

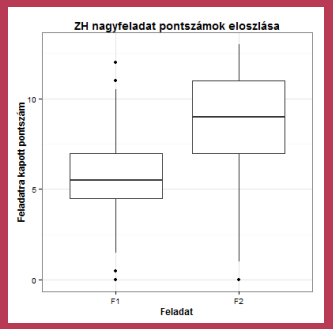
Változók

Numerikus

ZH pontszám: [13, 15, 2, …]

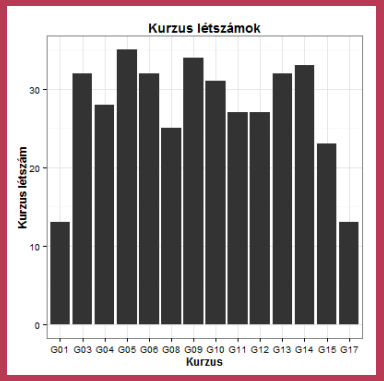






Kategorikus

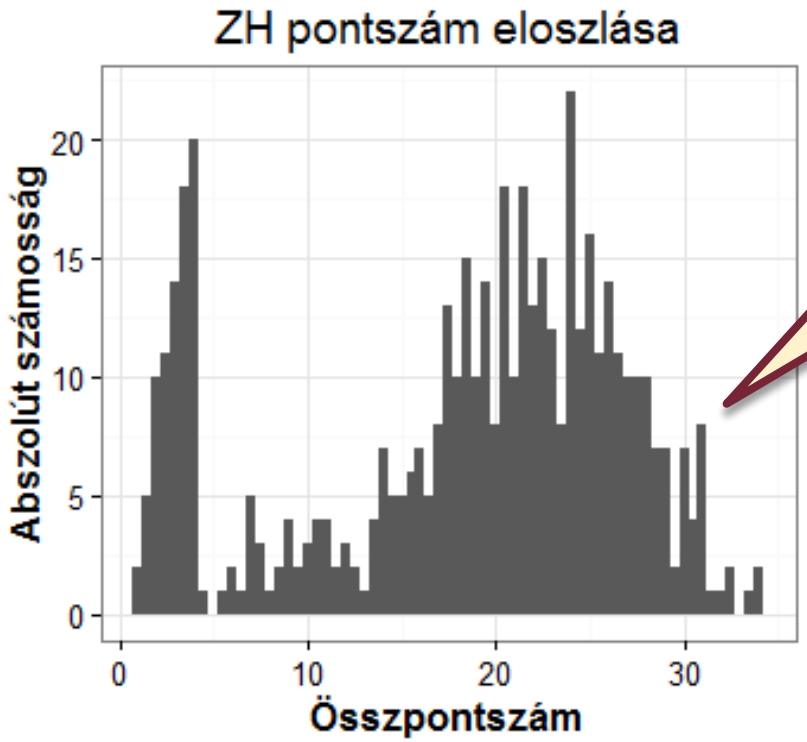
Kurzus: [G01, G03, G15, G17, …]



Hisztogram

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

abszolút gyakoriság!

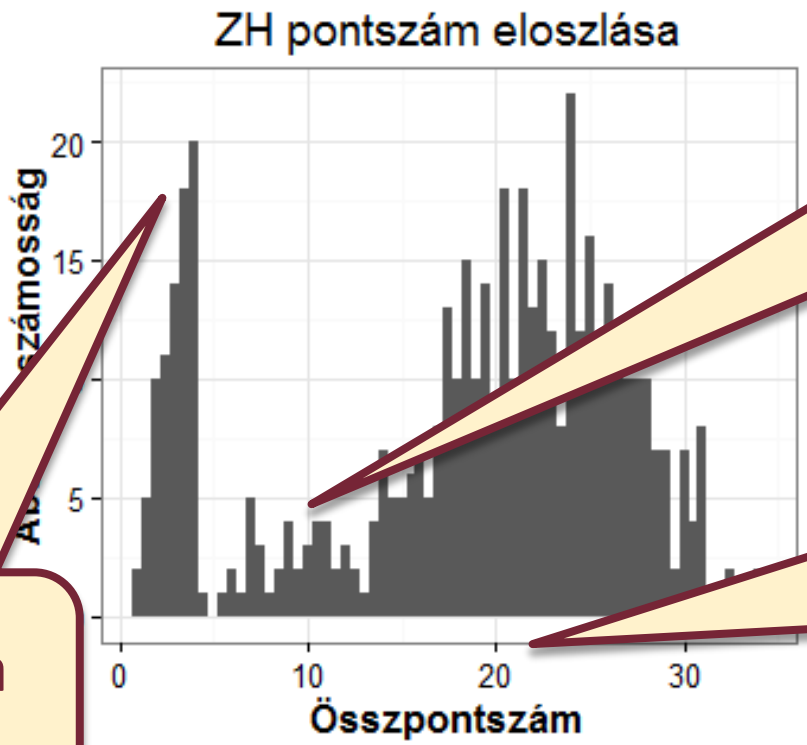


Oszlop-magasság: adott **intervallum** számossága

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

Hisztogram

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



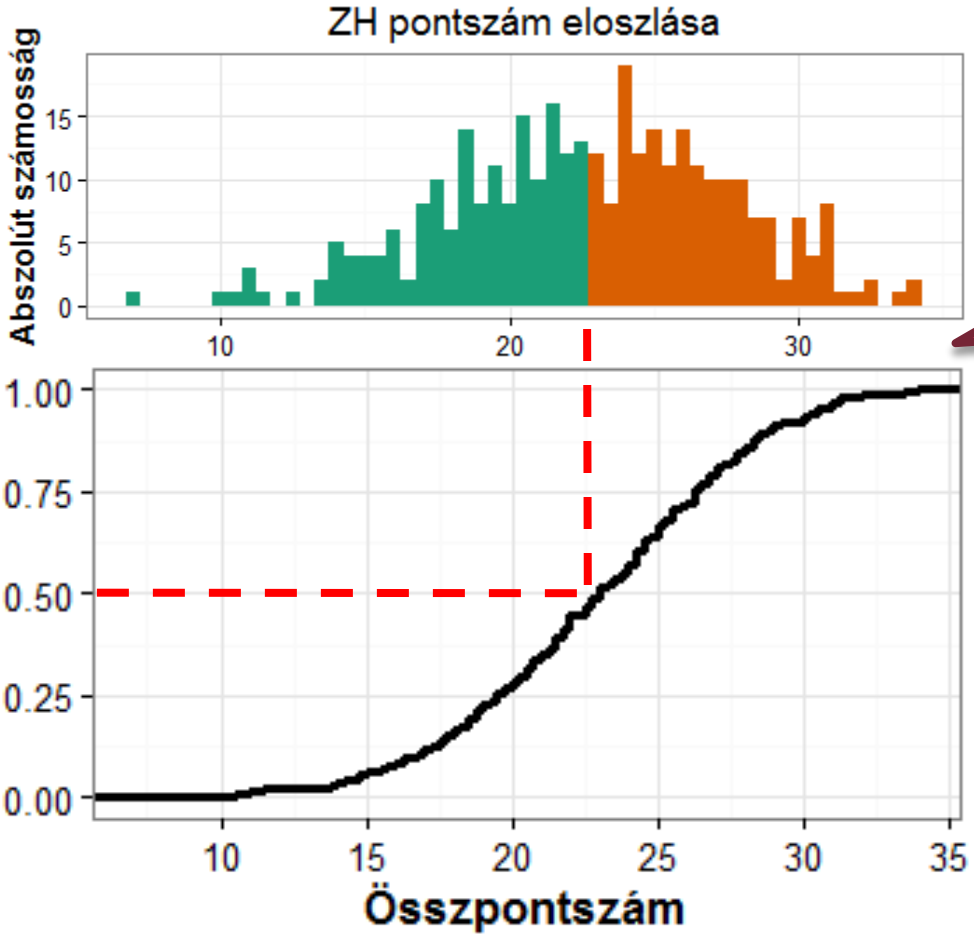
Akik átmentek a beugrón, valószínűleg át is mentek

Sokan voltak a határon

18 pont körül volt az átlag, 20 körül a medián

Egyszerű statisztikai jellemzés

- Hol van az adatok „közepe”?

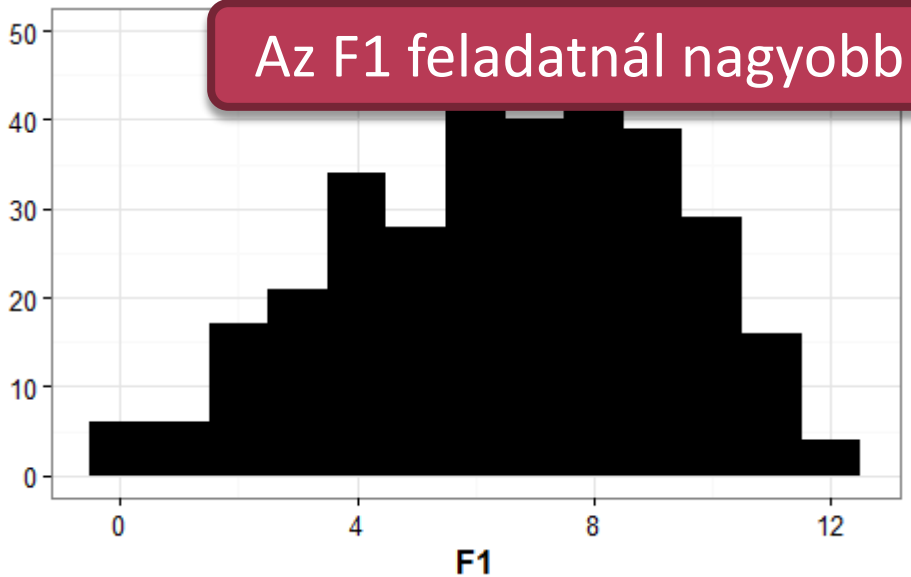


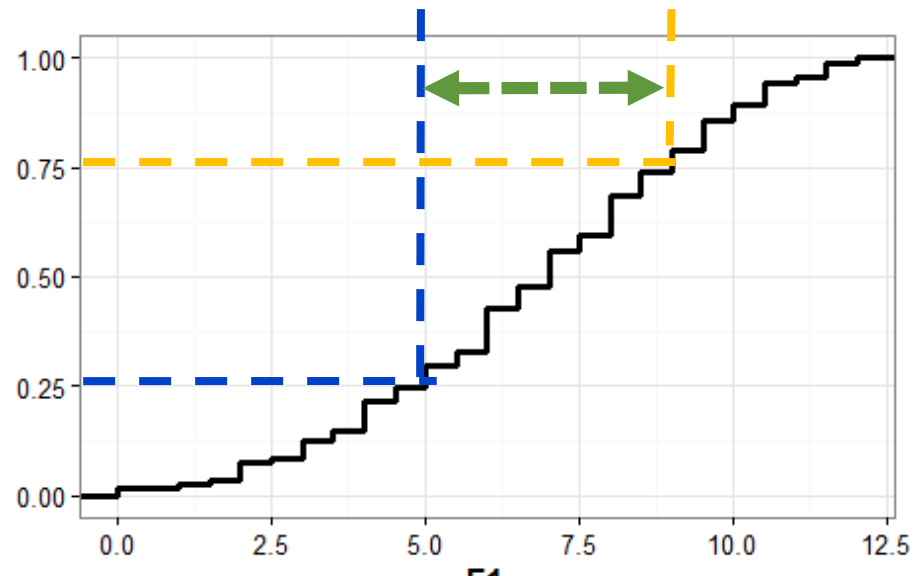
Az átmentek összpontszám ának mediánja 23

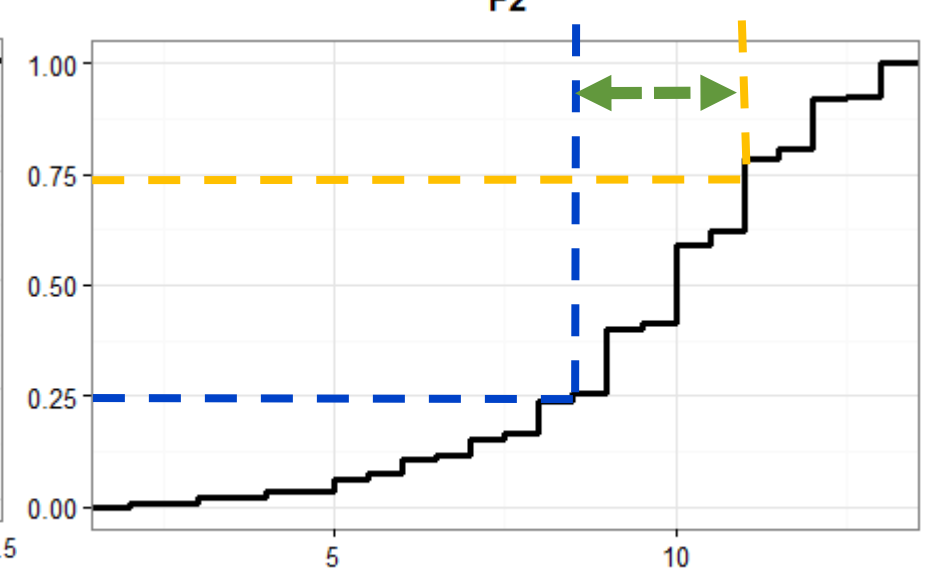
Egyszerű statisztikai jellemzés

- Mennyire „szórtak” az adatok?

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



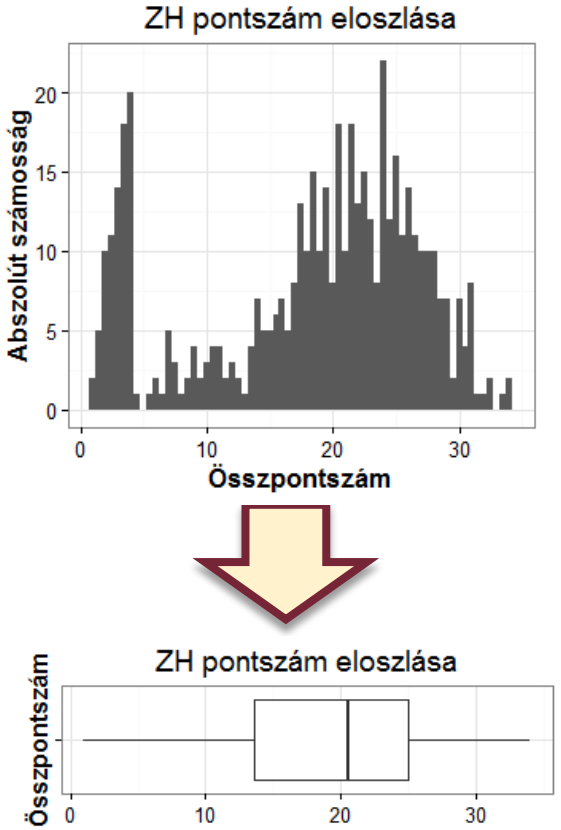




Boxplot

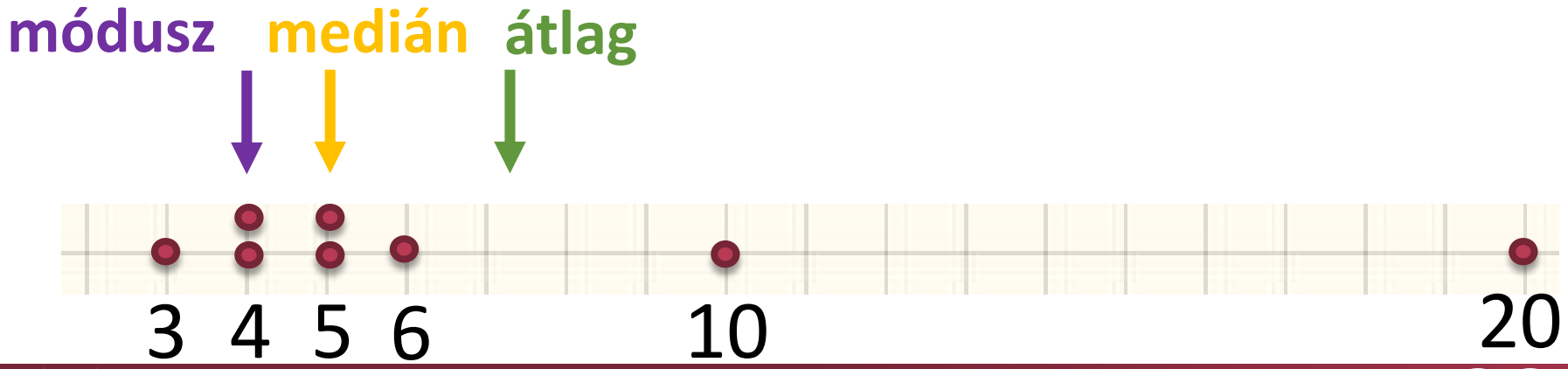
- Bemenő változó: ZH összpontszám
- Kérdés: hogyan alakultak a ZH pontszámok **úgy nagyjából**?

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



(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) mefigyelések jellemzése

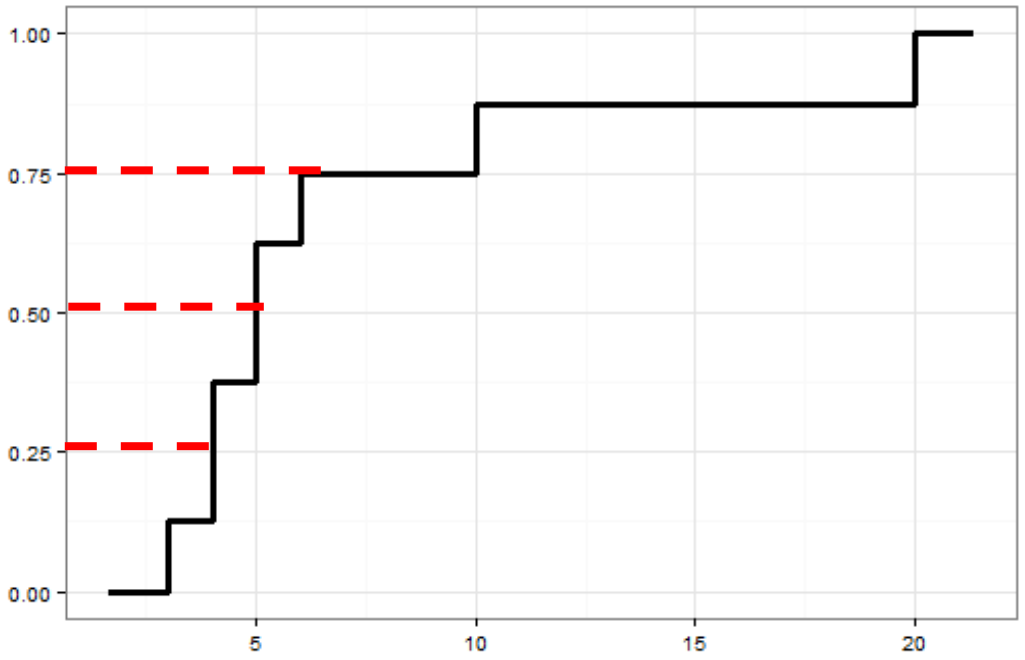
Ha az értékeket növekvően sorba rendezzük, akkor a középső adat az adathalmaz **mediánja**. Ha nincs középső adat (páros számú érték esetén), akkor a **medián** a két középső érték átlaga (számtani közepe).

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

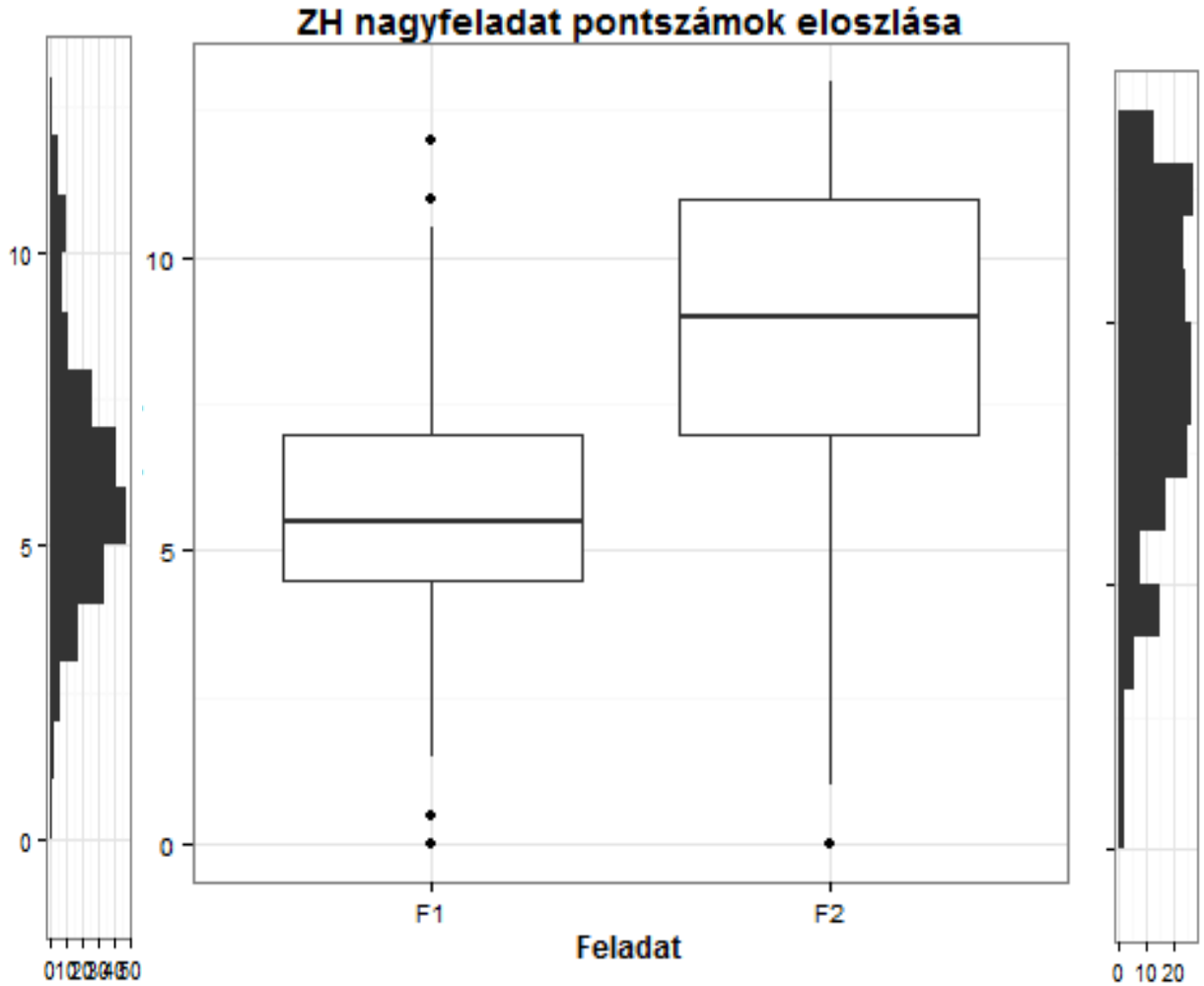
Terjedelem jellemzése: percentilisek

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

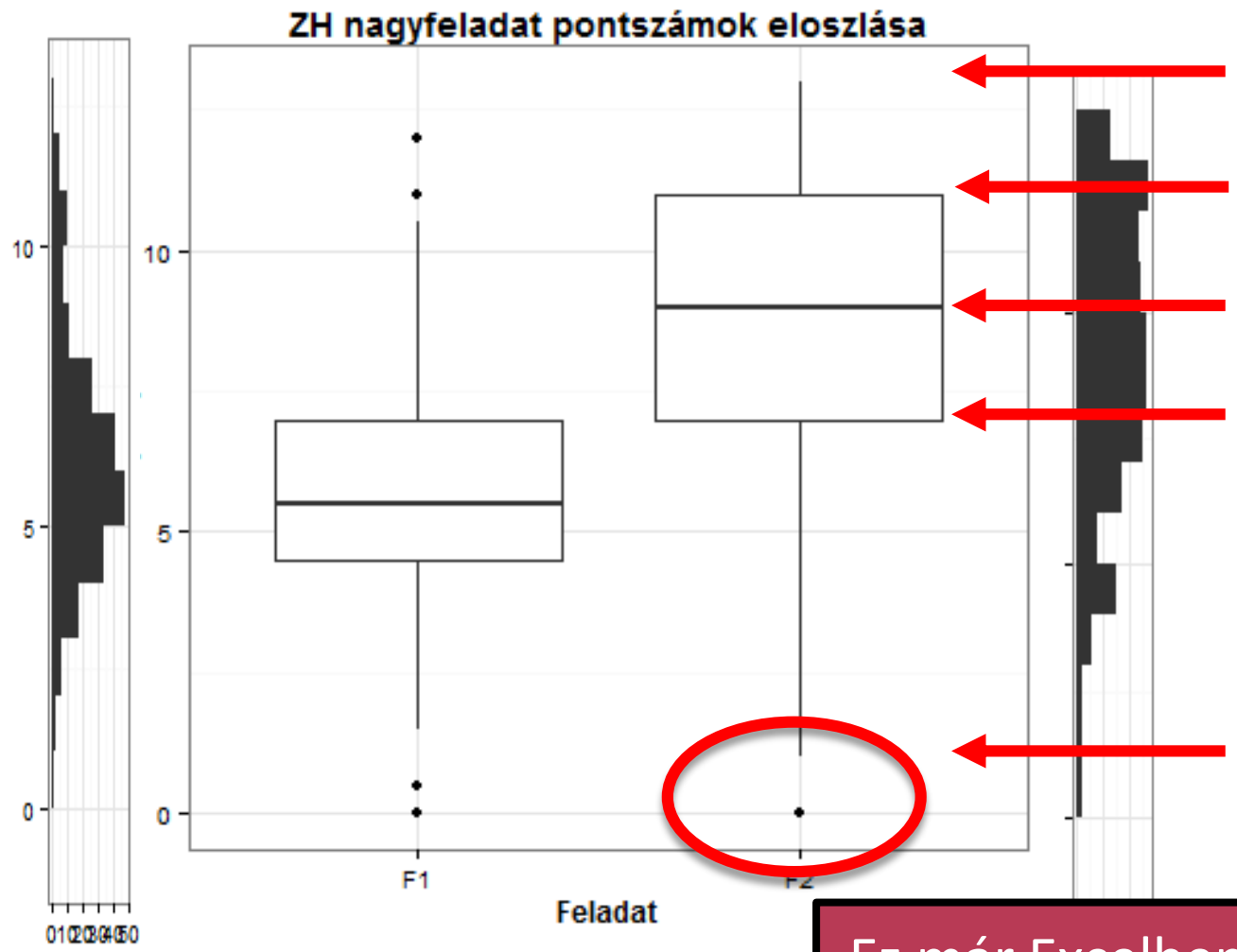
- Percentilis
  - {3, 4, 4, 5, 5, 6, 10, 20}
    - 50. percentilis: 5
    - 25. percentilis: 4
    - 75. percentilis: 6
- Kvartilis
  - Q1: 25. percentilis
  - Q3: 75. percentilis
  - **Q2: medián**



Boxplot (Box and whisker plot)



Boxplot (Box and whisker plot)



Felső outlier határ

Q3

Medián

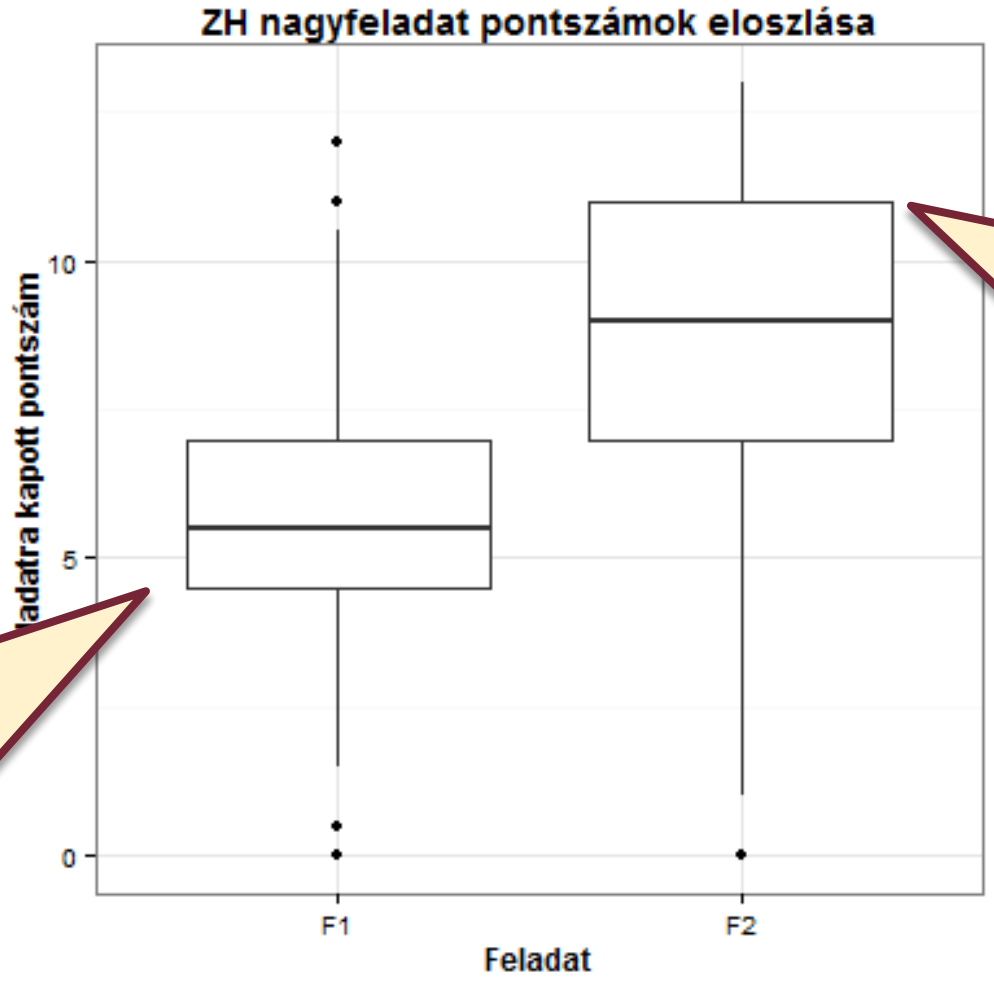
Q1

Alsó outlier határ

Outlierek

Ez már Excelben nem könnyű…

Boxplot (Box and whisker plot)

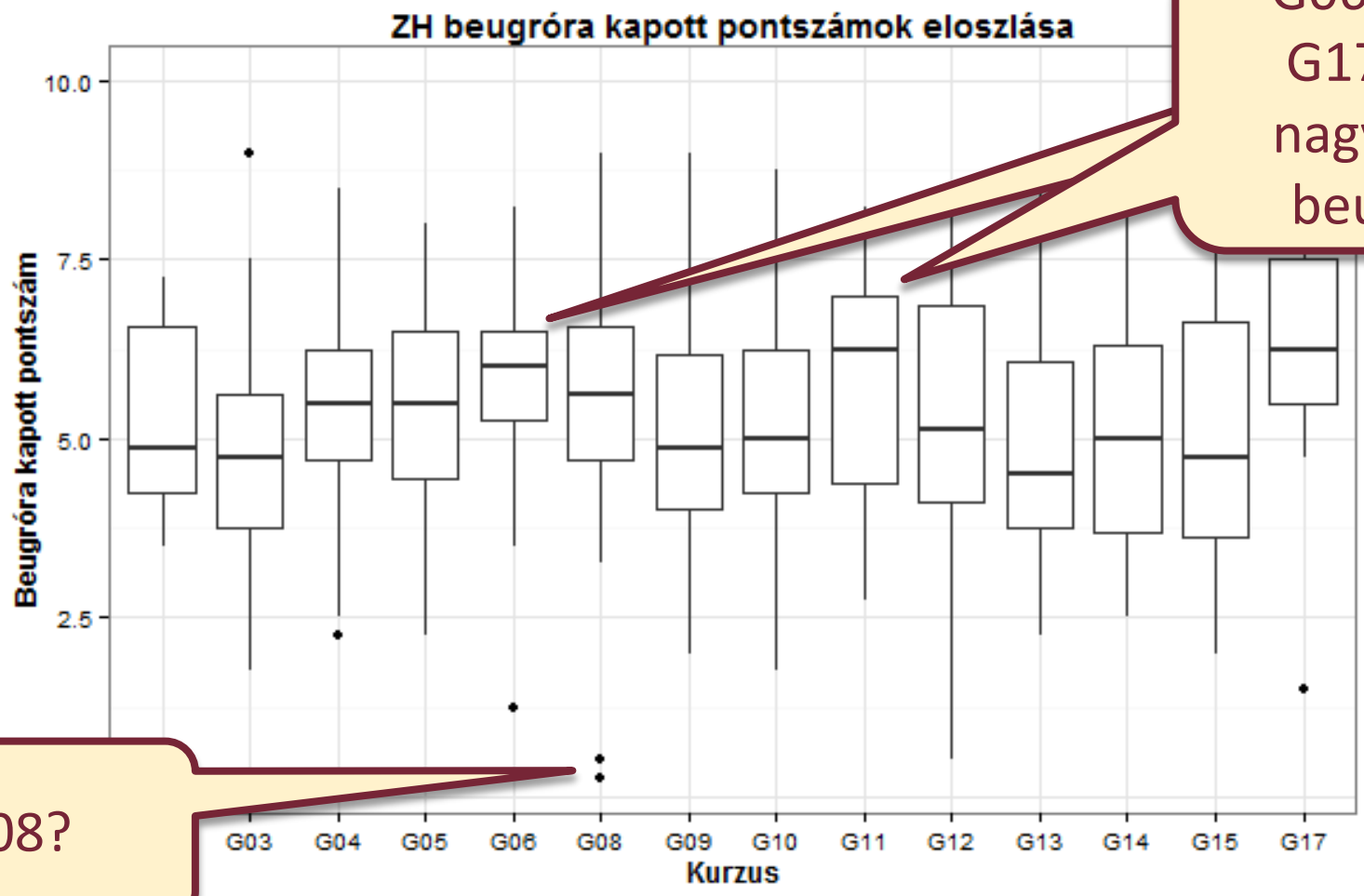


F2-re általában több pontot kaptak, mint F1-re

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

Boxplot (Box and whisker plot)

- Melyik csoportban hogyan sikerültek a beugrók?

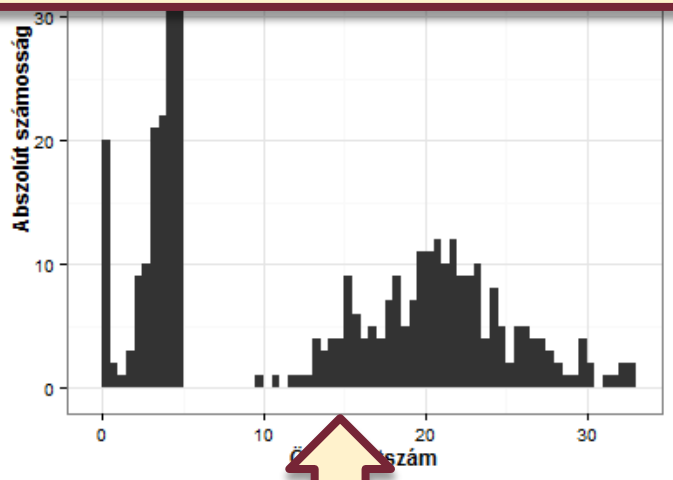


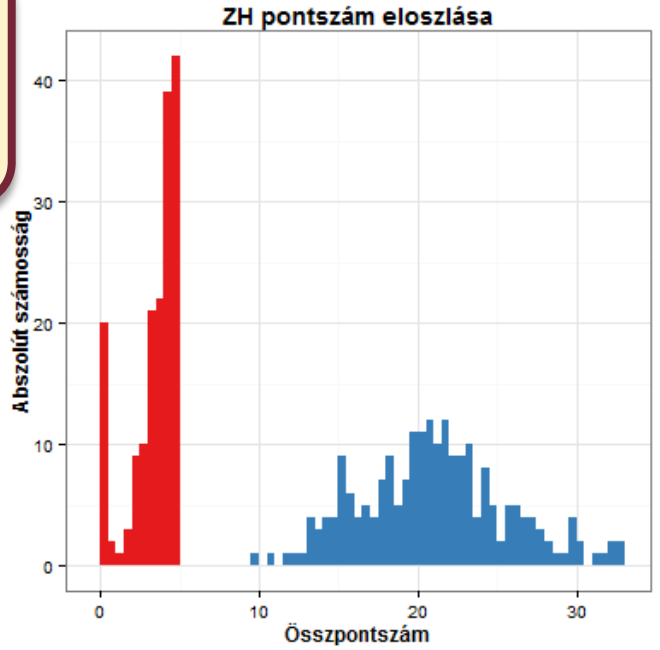
G06, G11, G17-ben nagyon jó beugrók

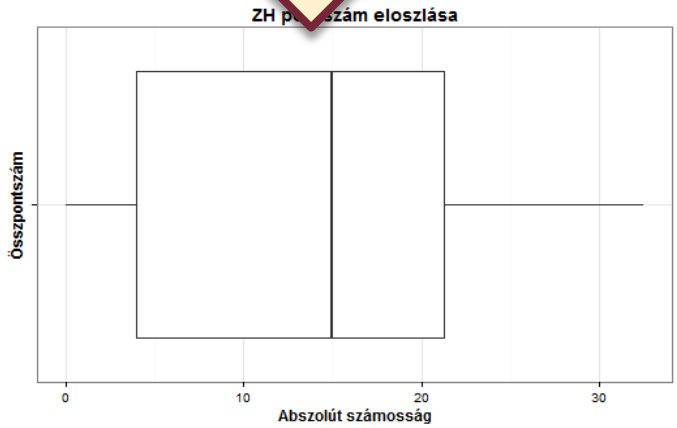
G08?

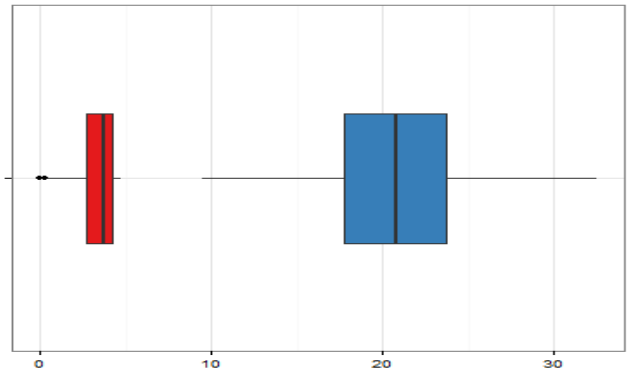
Boxplot (Box and whisker plot)

Absztrakció: a boxplottal fontos információt is veszíthetünk!





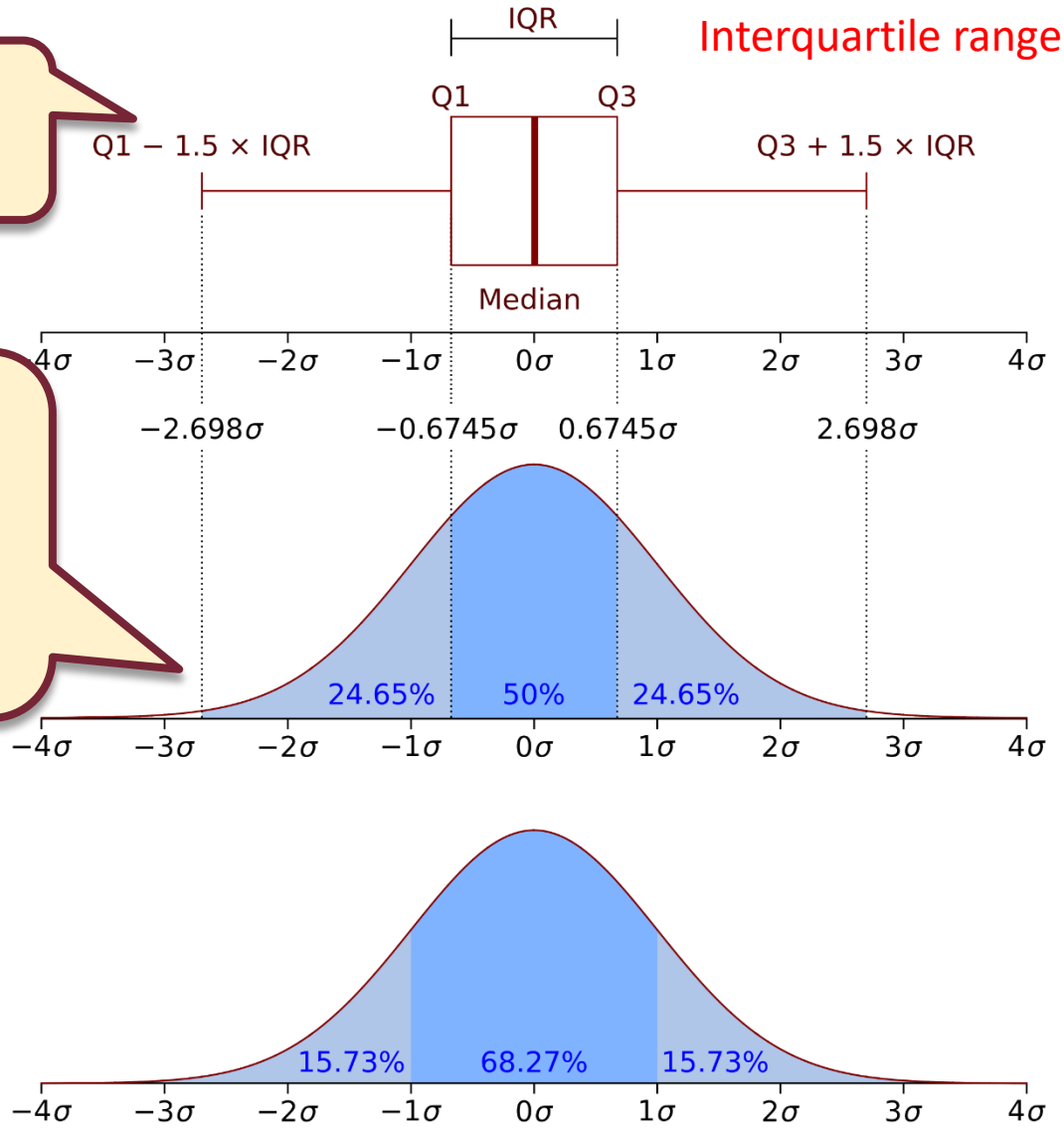




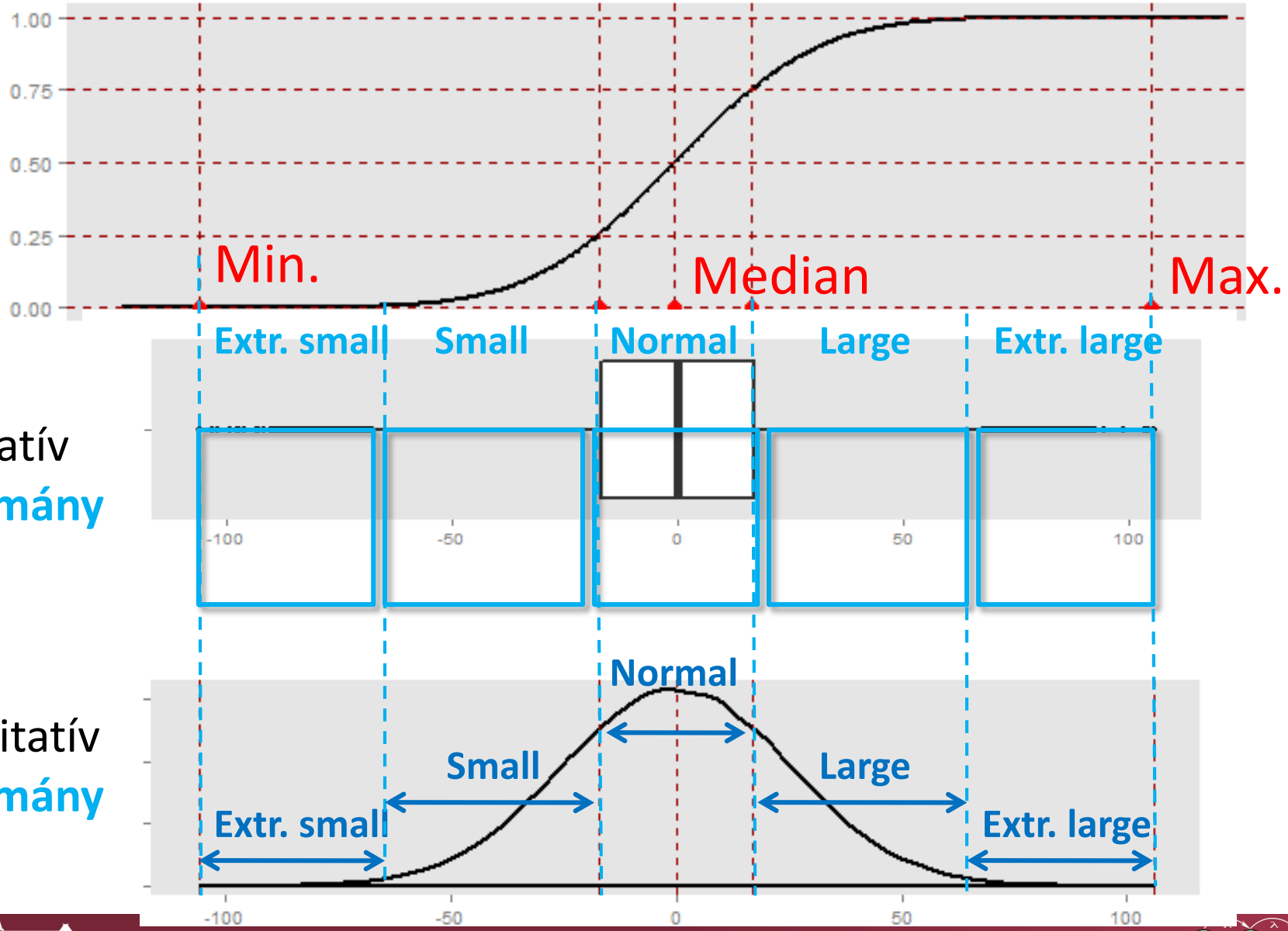
Boxplot (Box and whisker plot)

A ±1.5 csak konvenció

Normális eloszlásnál ez kb. a ± 3σ-nak felel meg



Boxplot: kvalitatív jellemzés

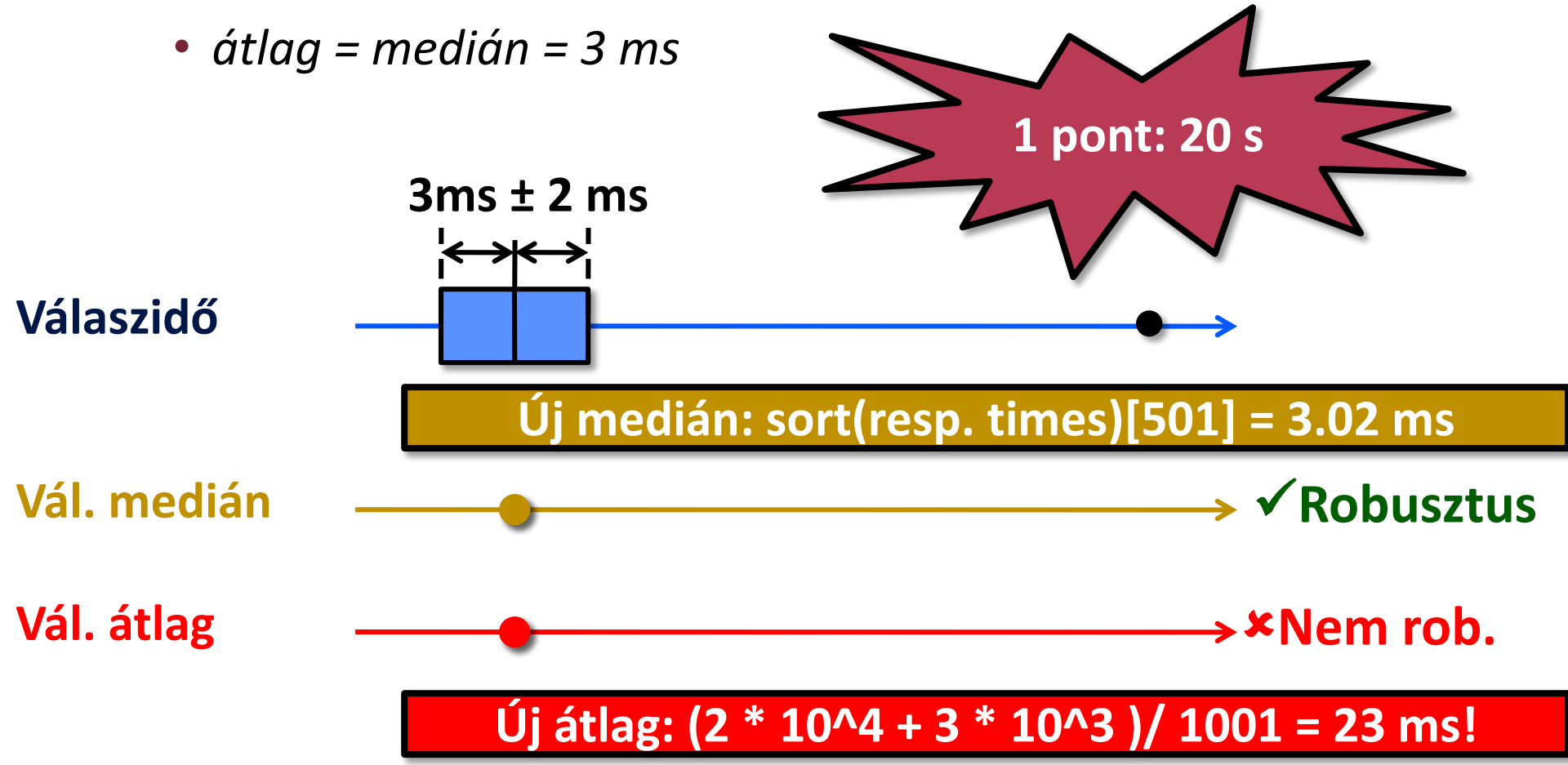


Kvalitatív **tartomány**

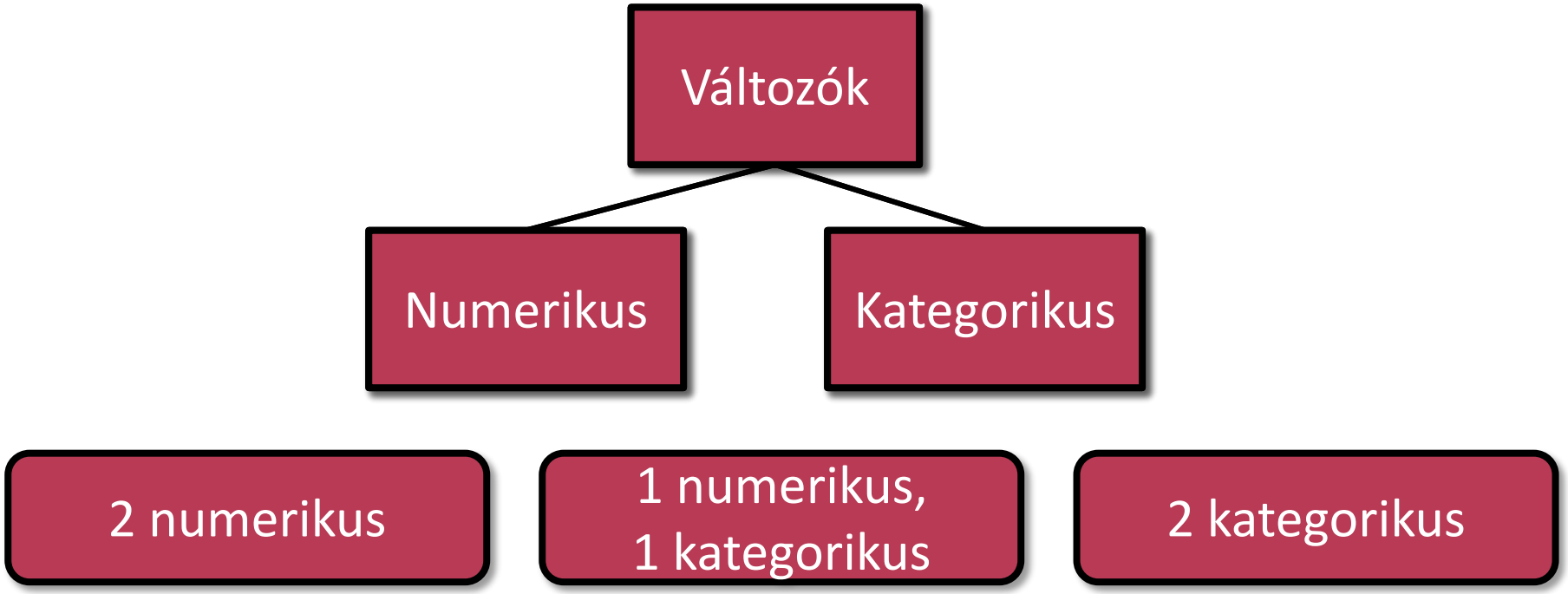
Kvantitatív **tartomány**

Miért medián, miért nem átlag?

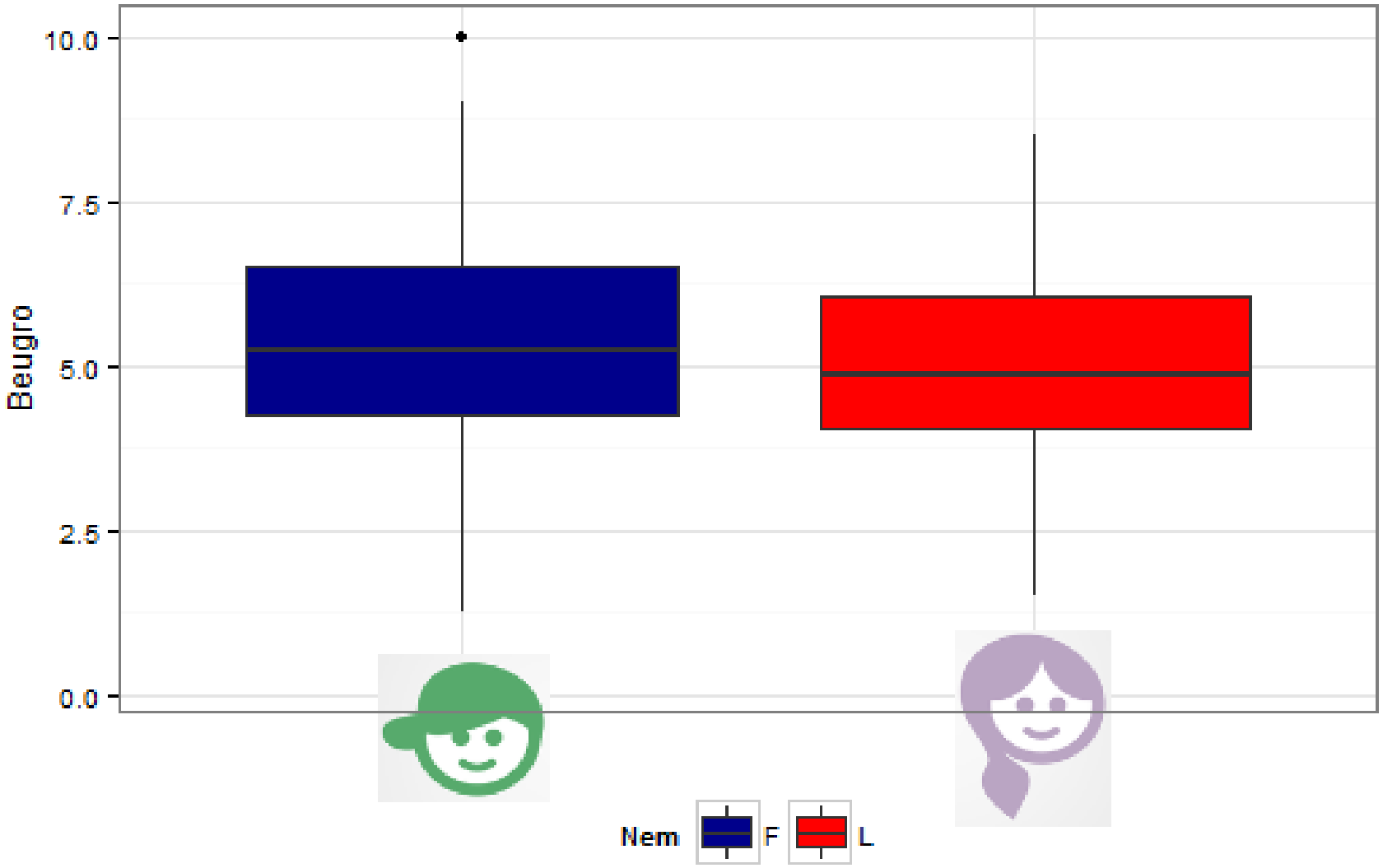
- Alaphalmaz
  - 1000 pont ~ U(1, 5) egyenletes eloszlás
    - *átlag = medián = 3 ms*



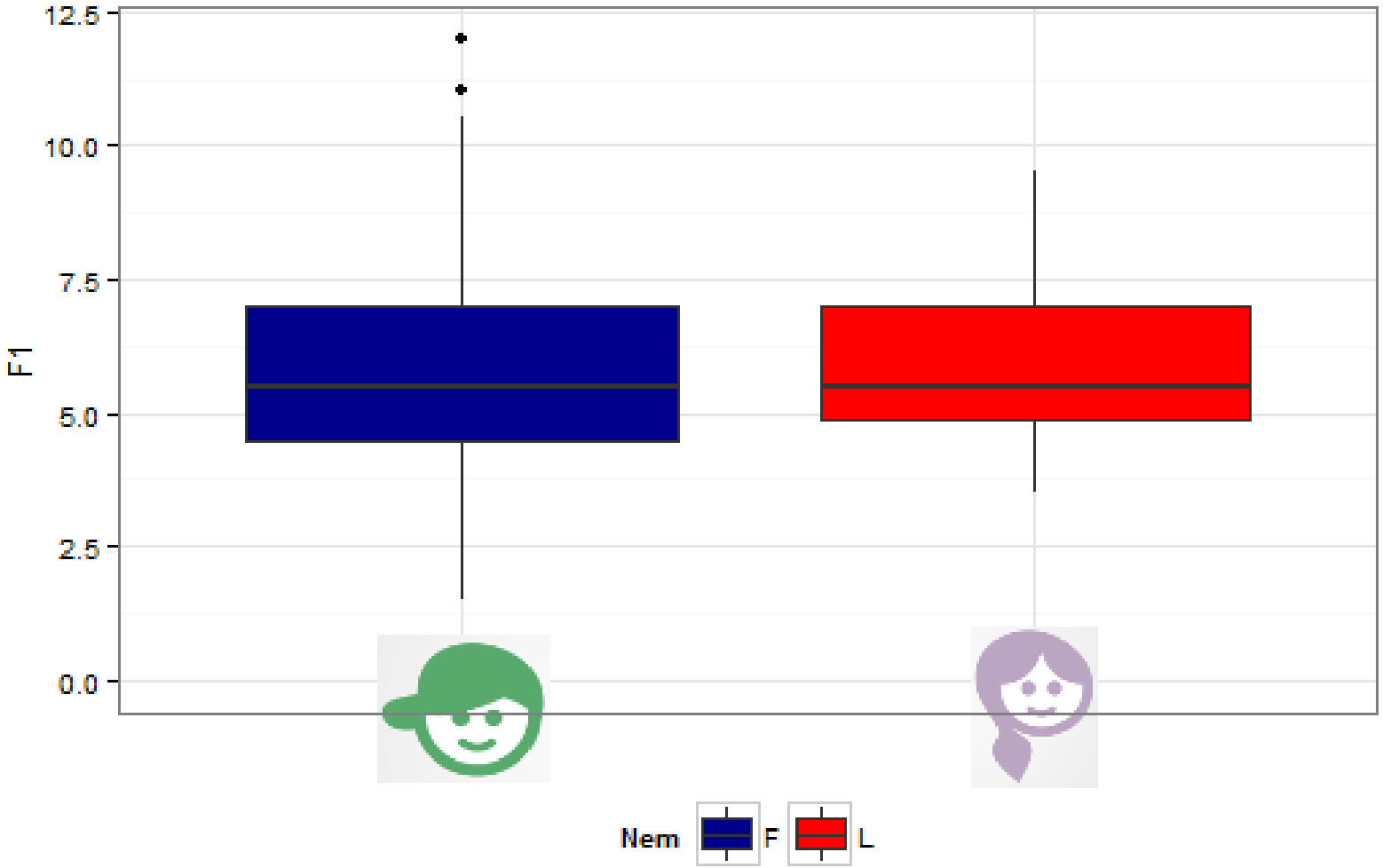
2 változó kapcsolata



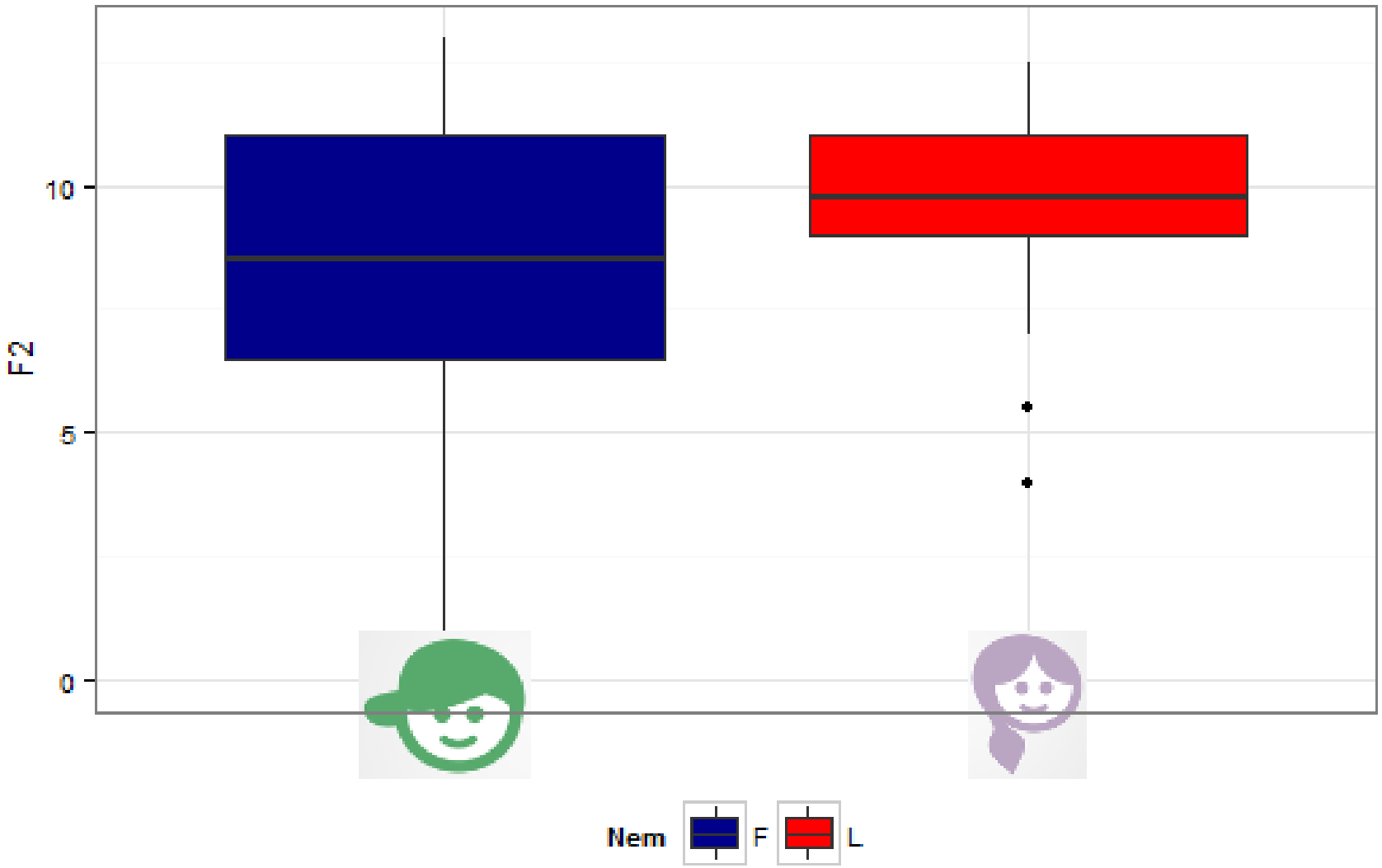
Numerikus kategóriánként



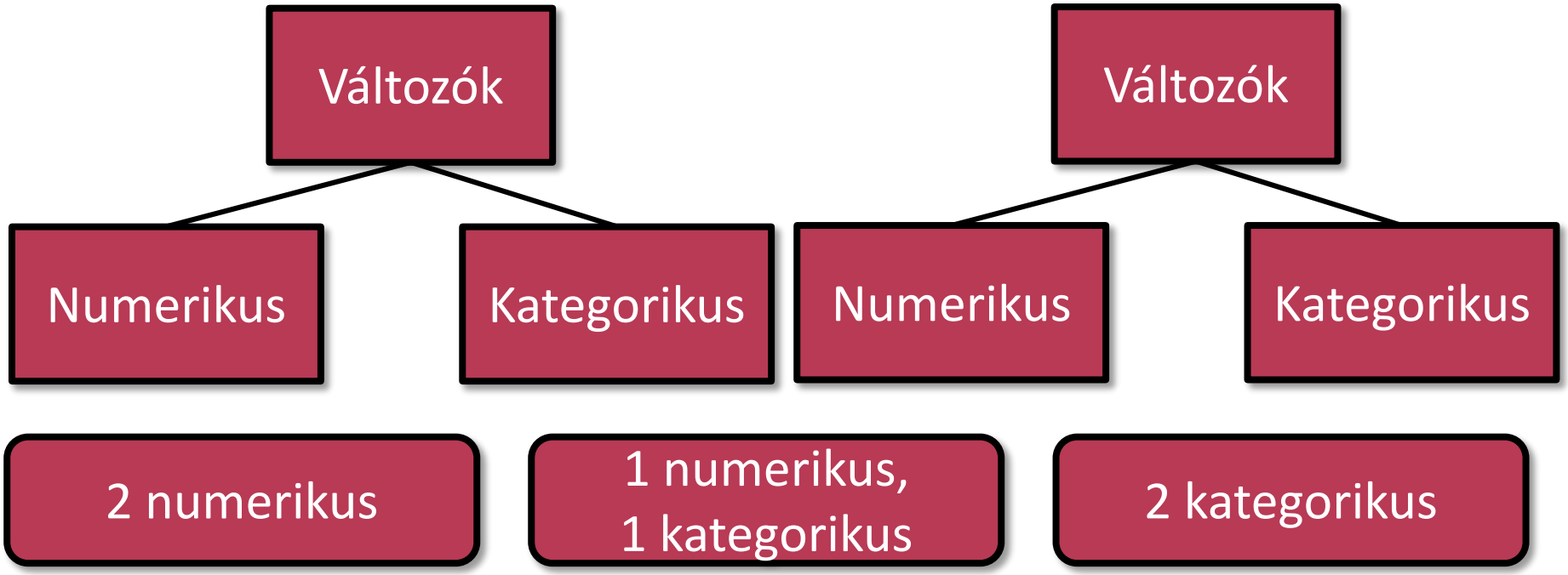
Numerikus kategóriánként



Numerikus kategóriánként



2 változó kapcsolata

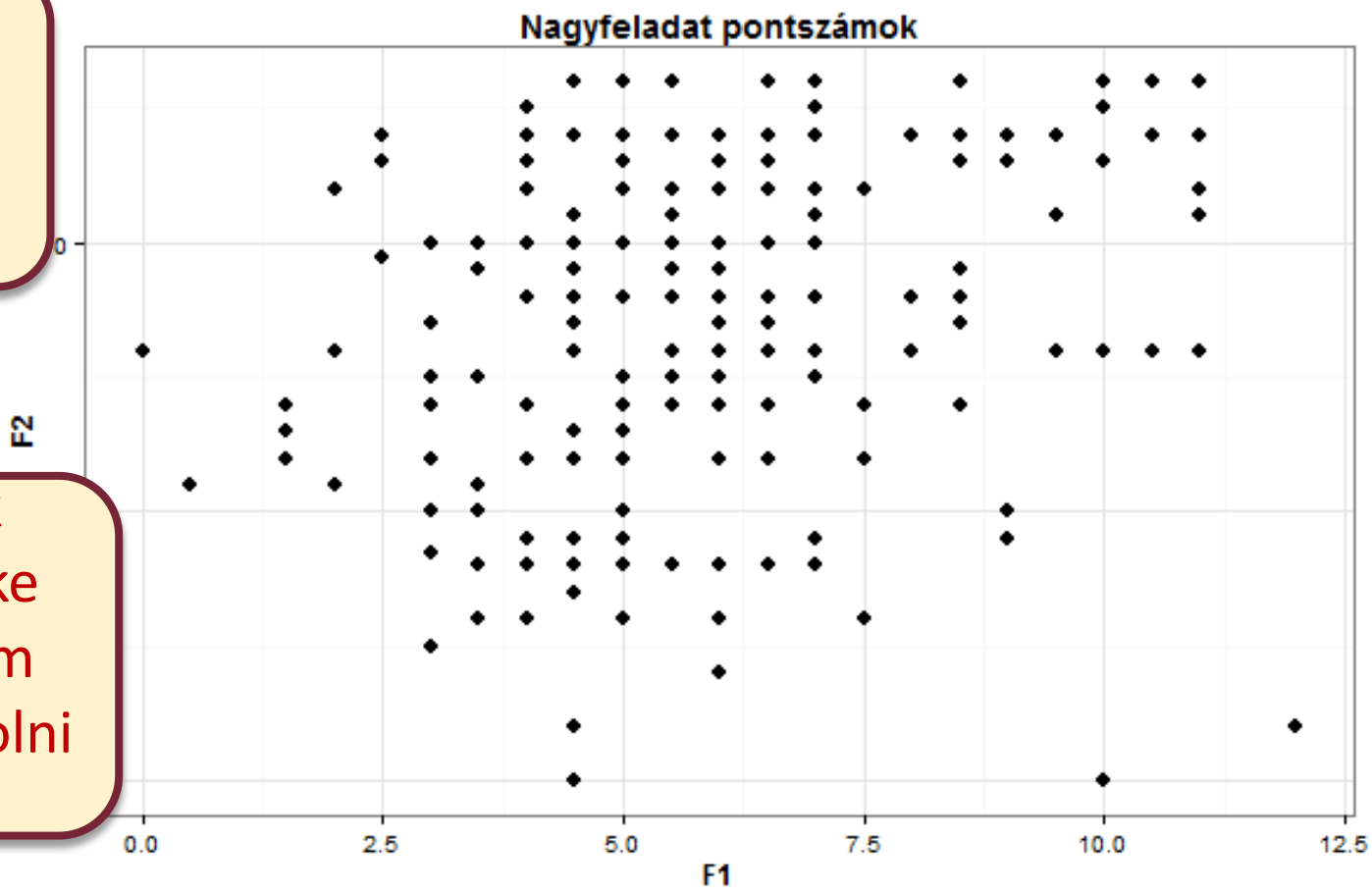


Pont – pont diagram (scatterplot)

- 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)

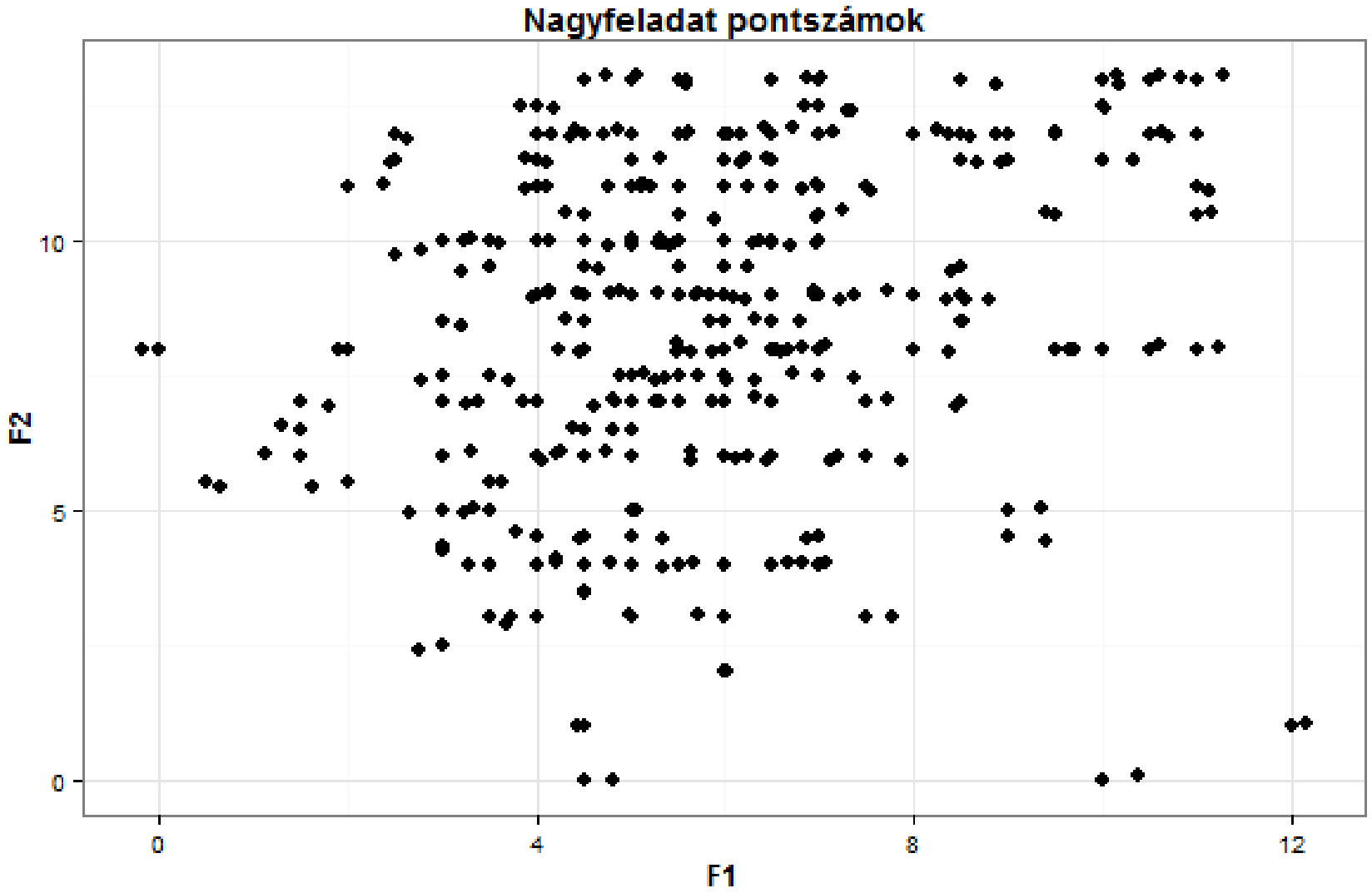
- Bemenő változó: nagyfeladatokra kapott pontok
- Kérdés: hogyan viszonyulnak egymáshoz?

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

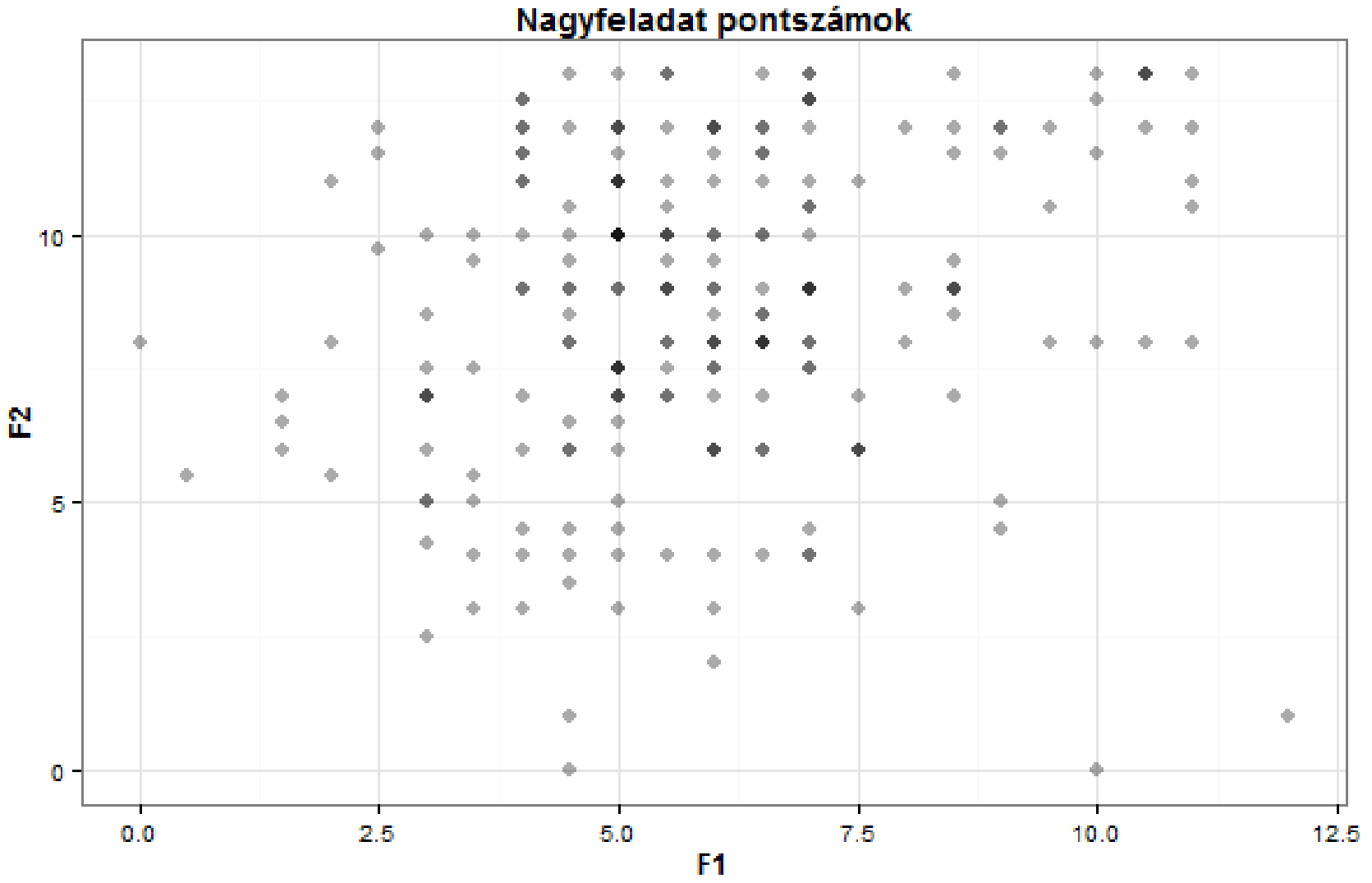
Hogyan kezeljük a takarásokat?

Overplotting

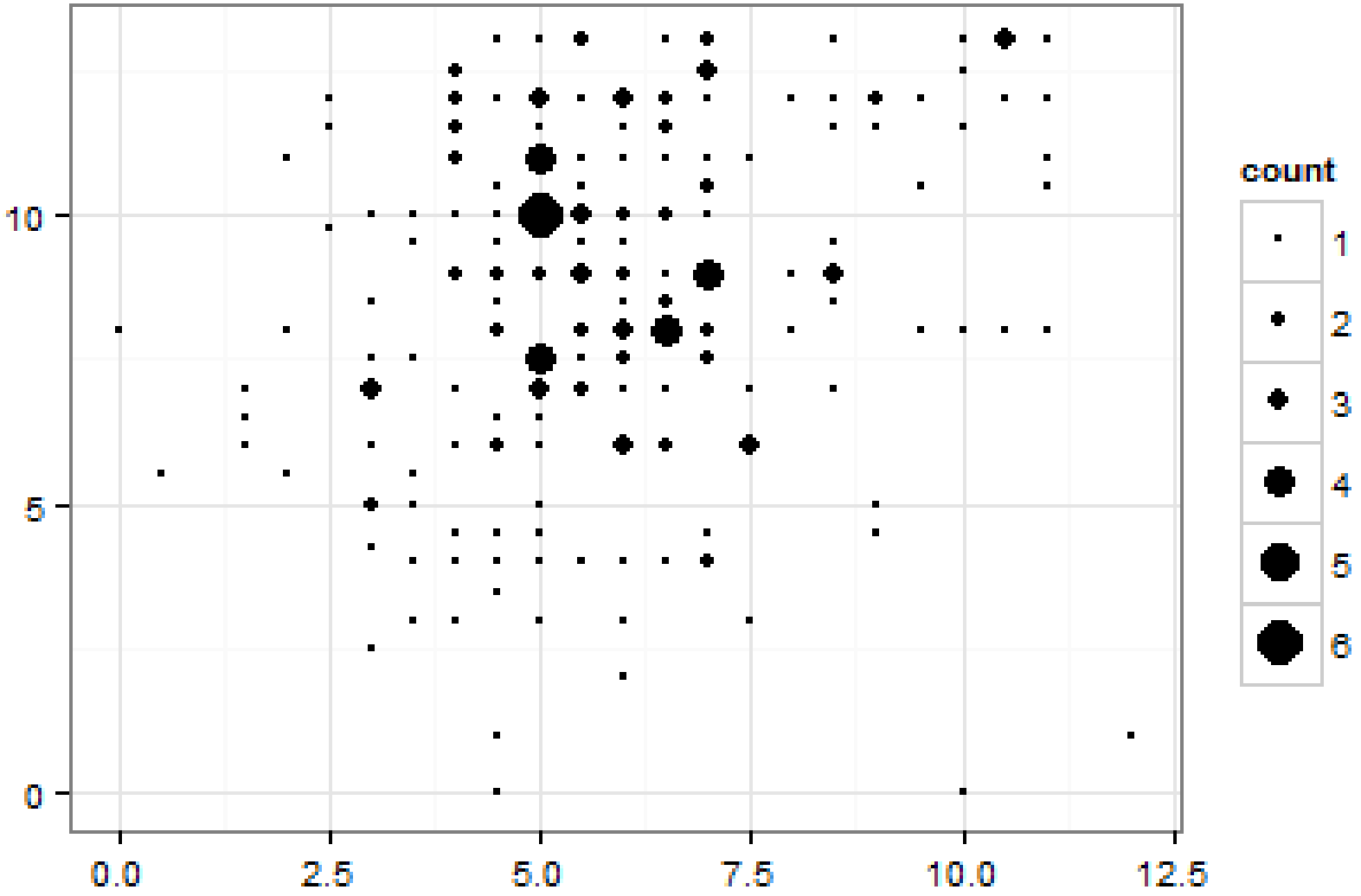
Overplotting megoldások 1: jitter



Overplotting megoldások 2: átlátszóság



Overplotting megoldások 3: méret

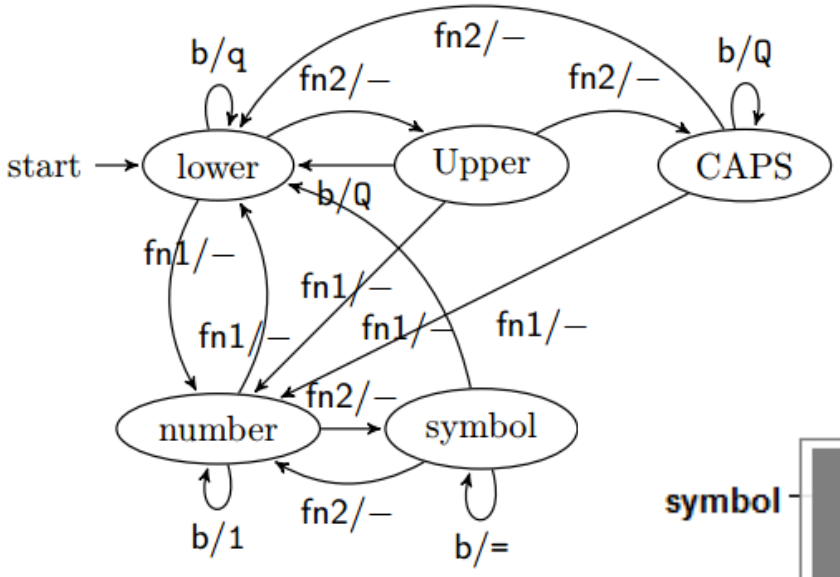


SOK VÁLTOZÓ

>=3 változó

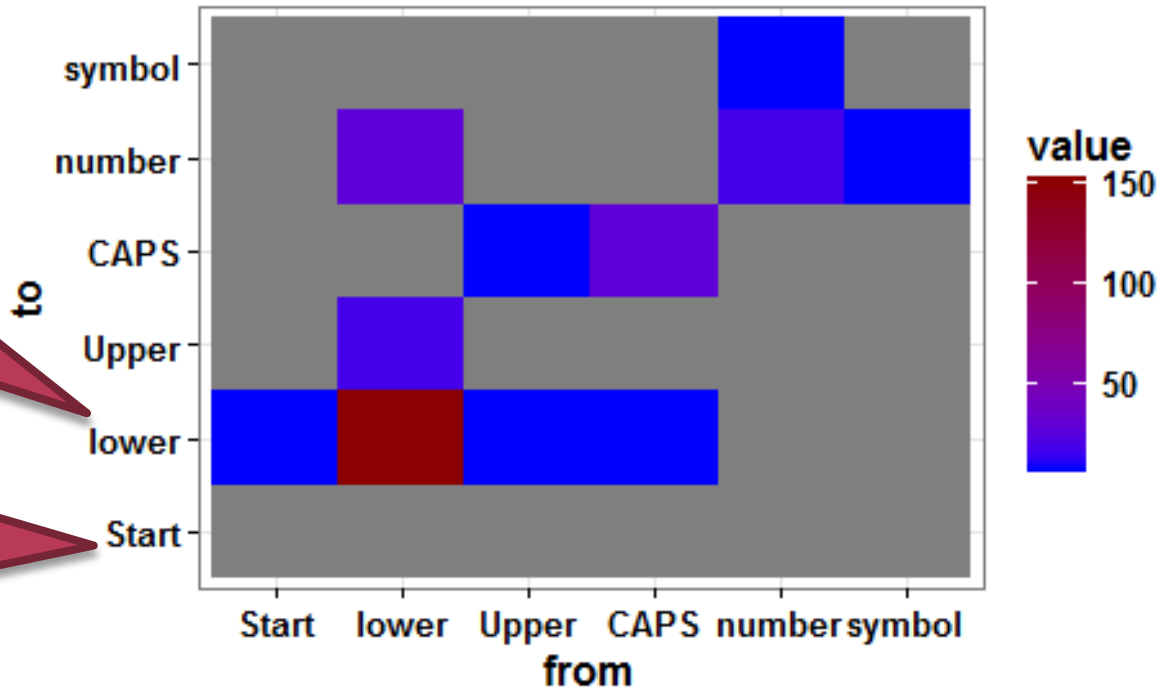
- A grafikai objektumok attribútumait változtatom
  - Szín
  - Méret
  - Textúra
  - Hely – ez triviálisnak tűnik, de a treemapnél van jelentősége
- Pl. heatmap, treemap

Heatmap: lefutási statisztikák

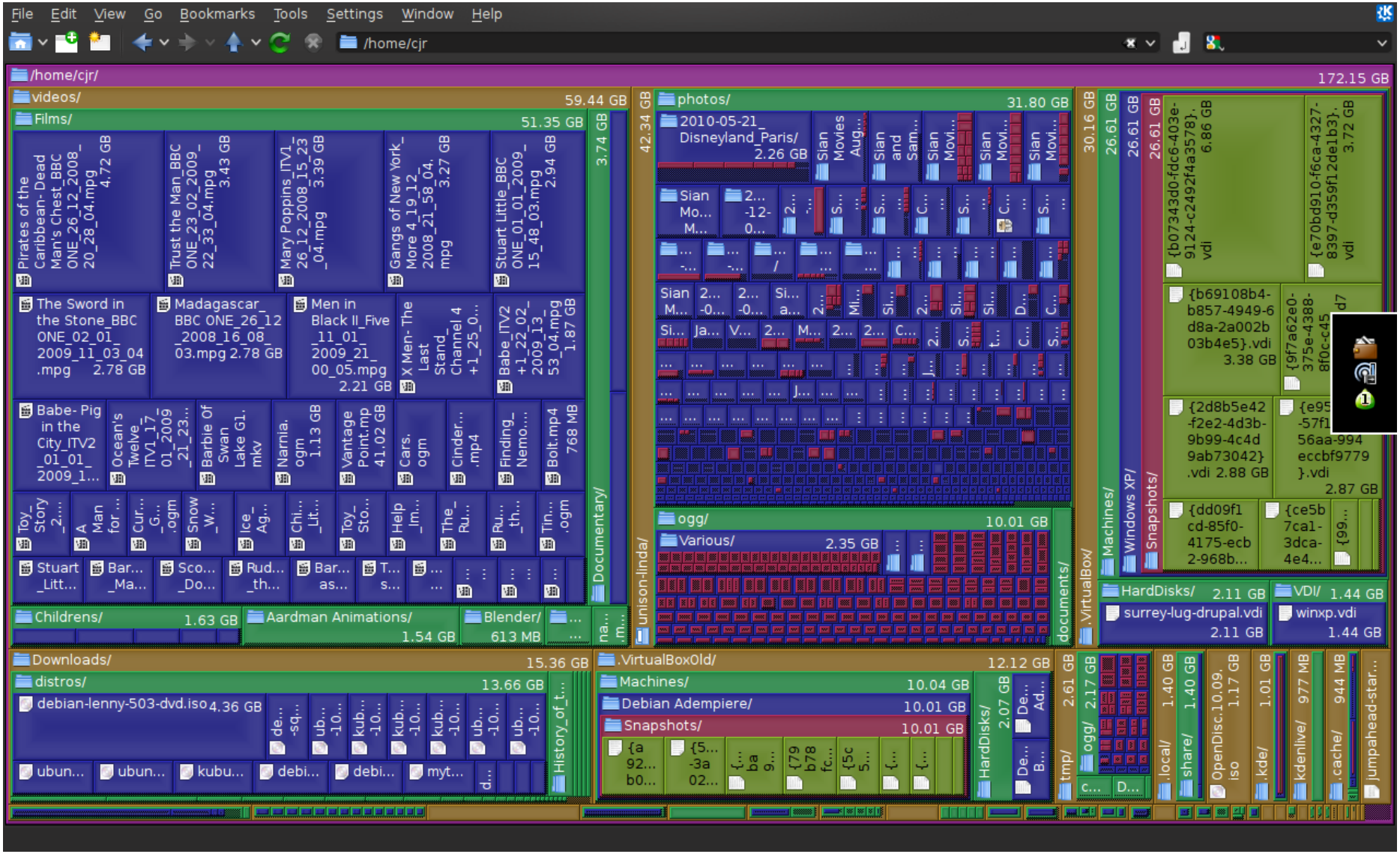


Inkább csak sima szöveget írunk

A Startban mindig csak kezdünk, oda nem jutunk vissza

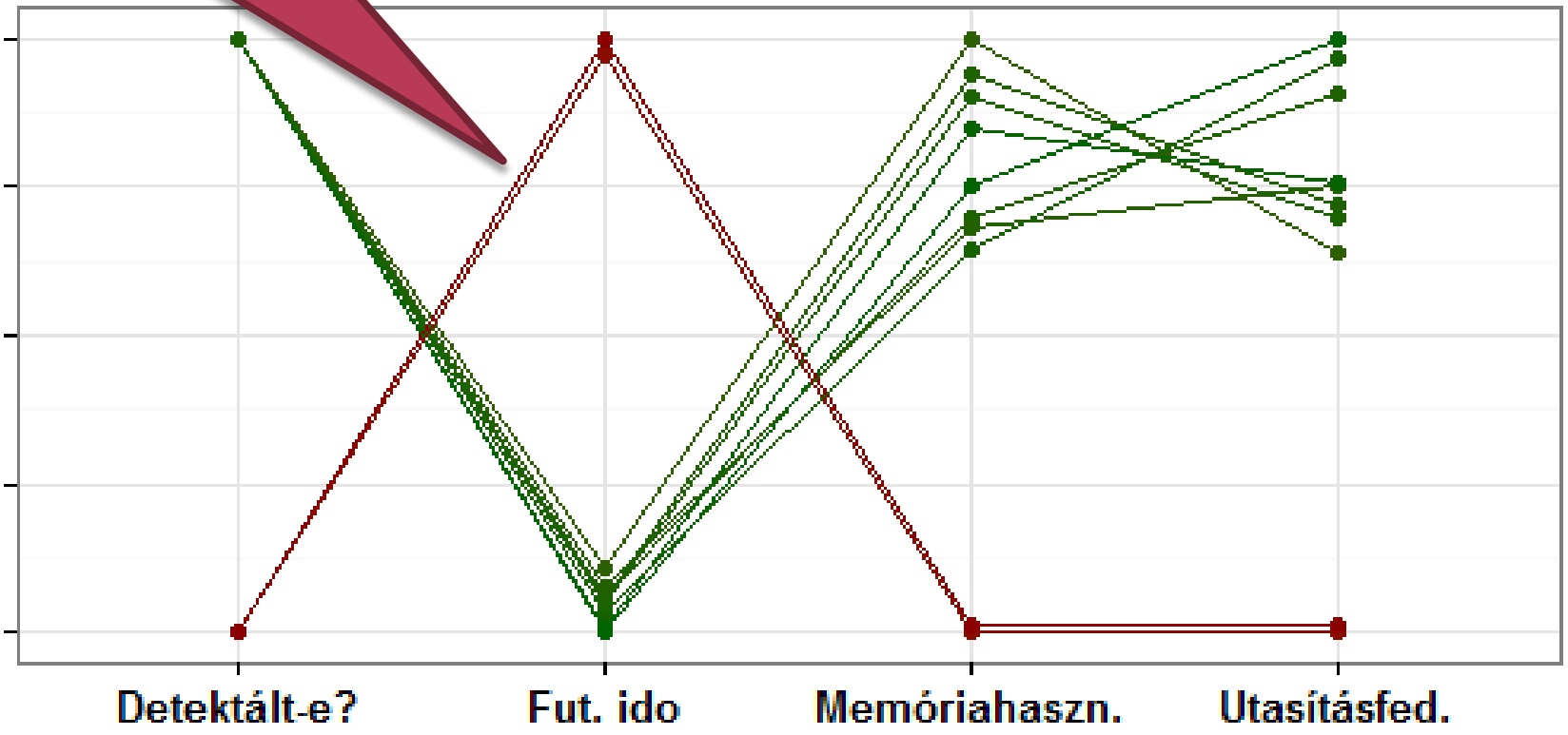


Treemap: állományrendszer



Párhuzamos koordináták: tesztesetek elemzése

1 teszteset 1 törött vonal

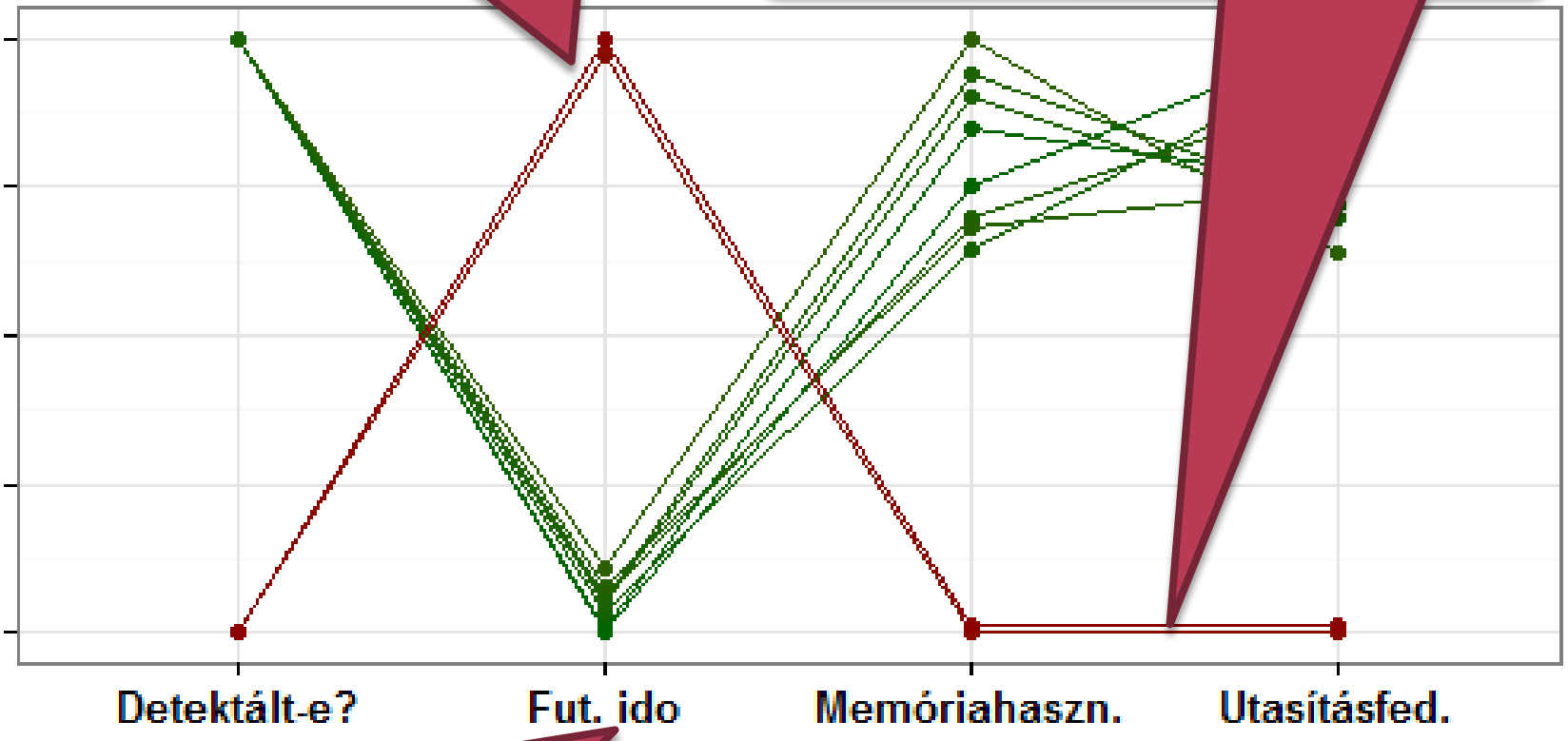


A változók az x tengelyen jelennek meg

Párhuzamos koordináták: tesztesetek elemzése

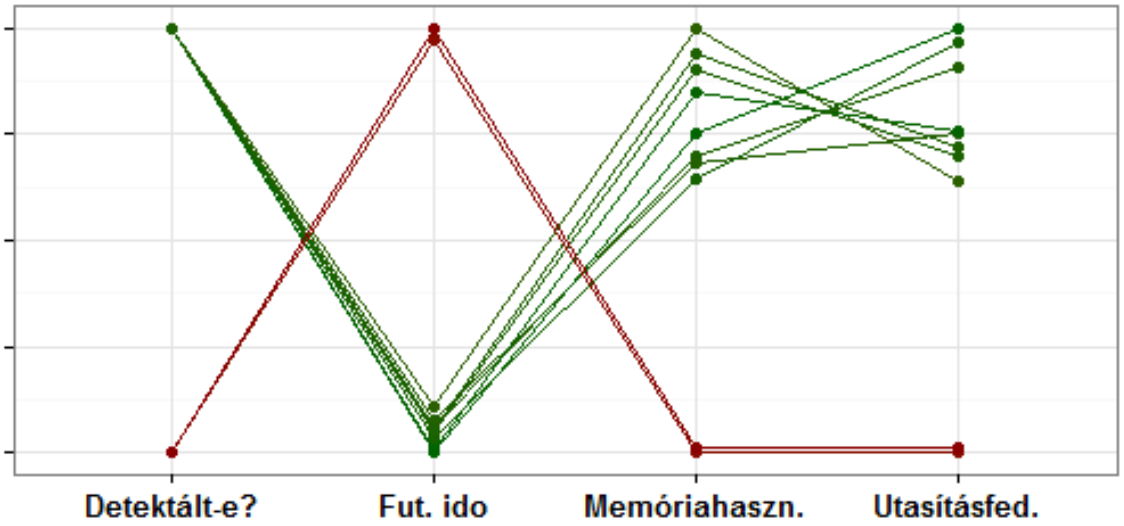
Timeout?

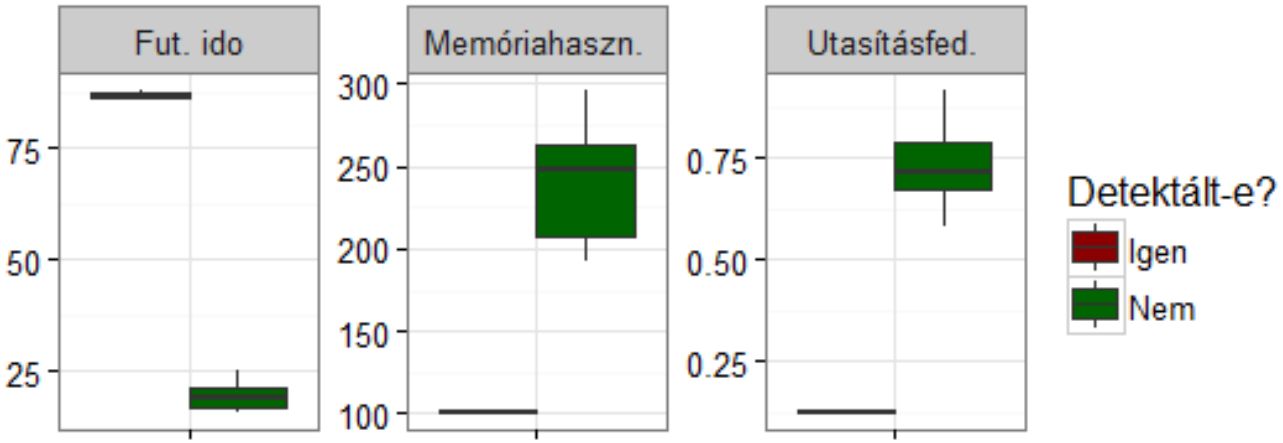
A hibát detektálók az érdemi számításig valószínűleg el sem jutnak



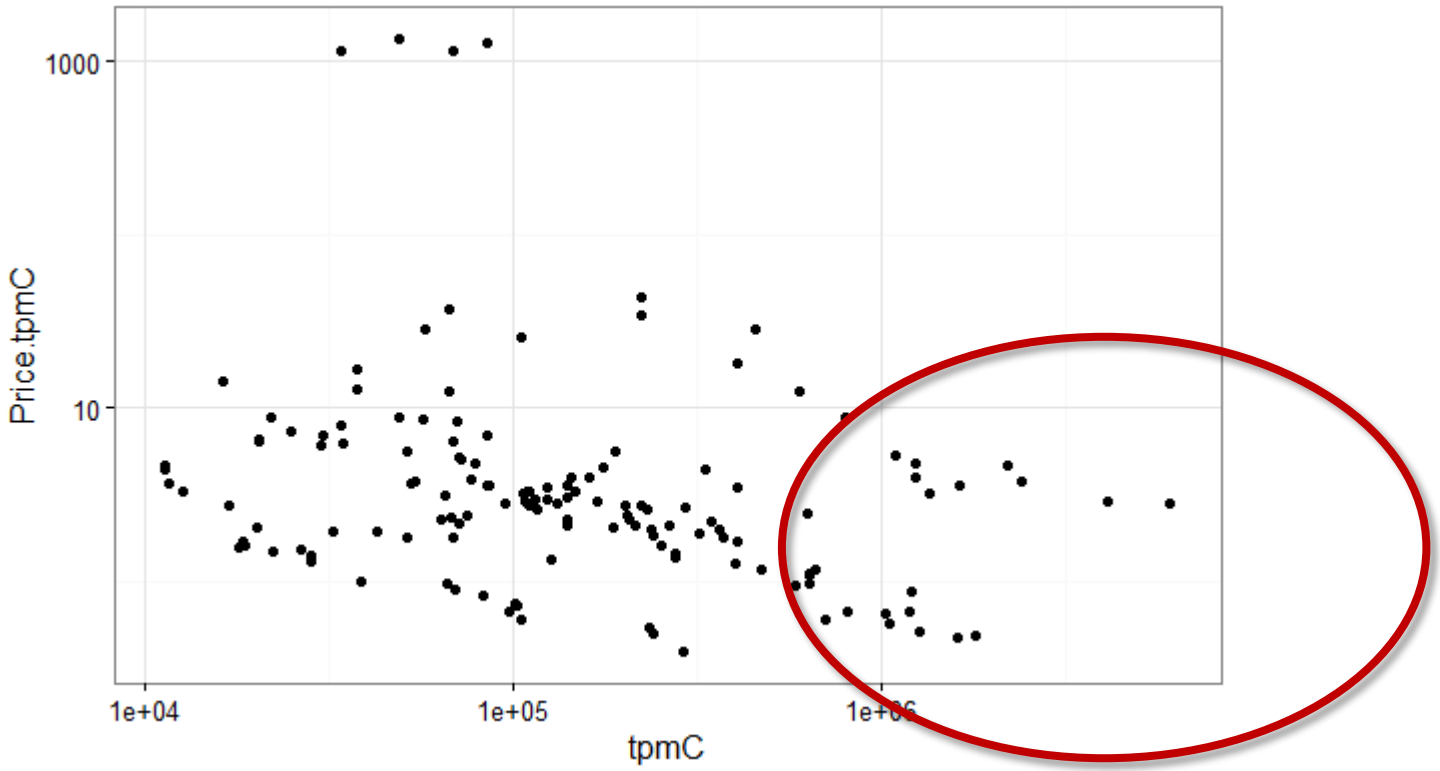
A futási idő és a memóriahasználat valószínűleg pozitív kapcsolatban állnak

Párhuzamos koordináták: viz. alternatívák



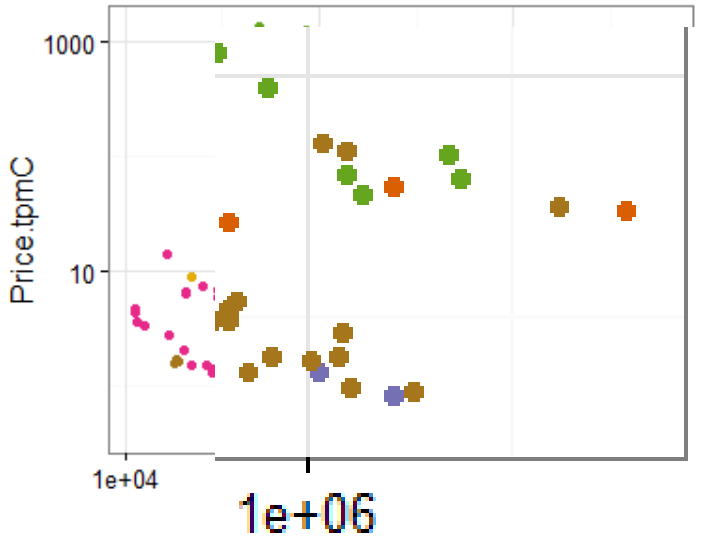


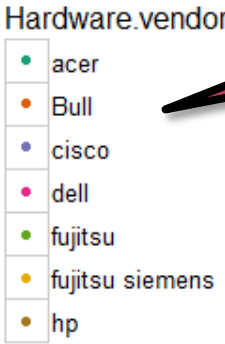
Benchmark eredmények



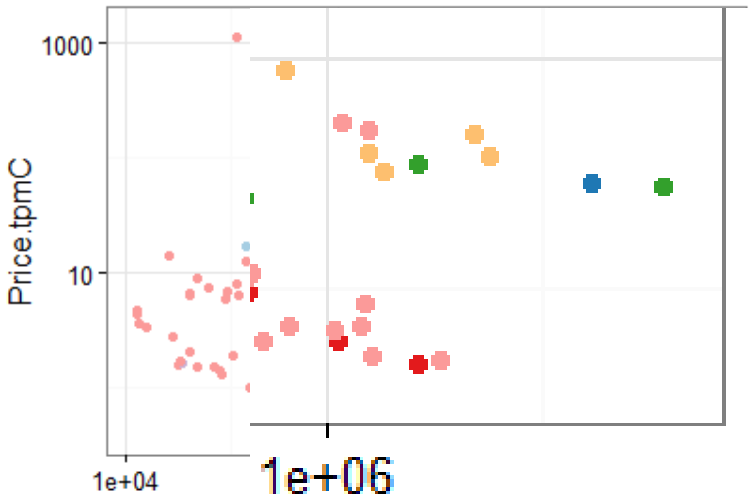
Logaritmikus skála?

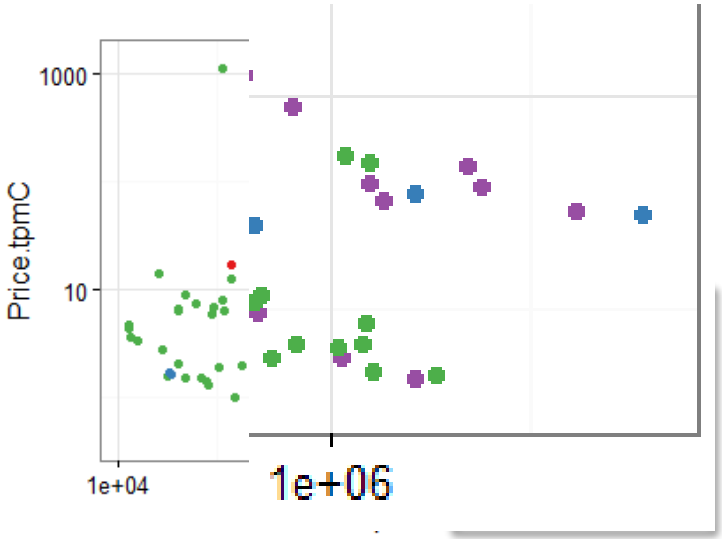
Benchmark eredmények

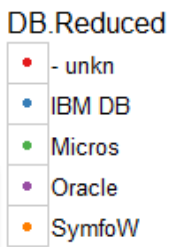




Az élmezőny elég változatos







Nincs legjobb OS és DB konfiguráció sem

Időbeli változások?

A legtöbb gyártónál
10 év alatt
egy nagyságrenddel csökkent a
fajlagos ár