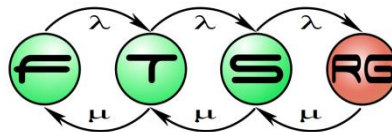


Vizuális adatelemzés

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 hétköznapijai

Analóg megjelenítés



Digitális megjelenítés



Analóg + koord. rei

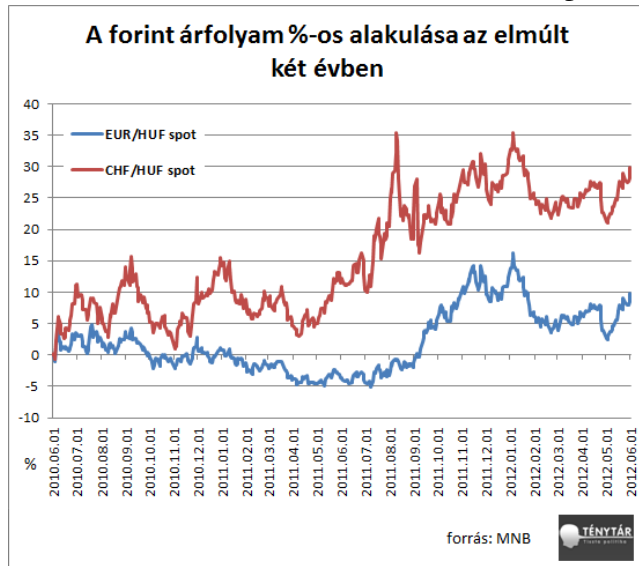


id megjelenítés



A vizualizáció hétköznapijai

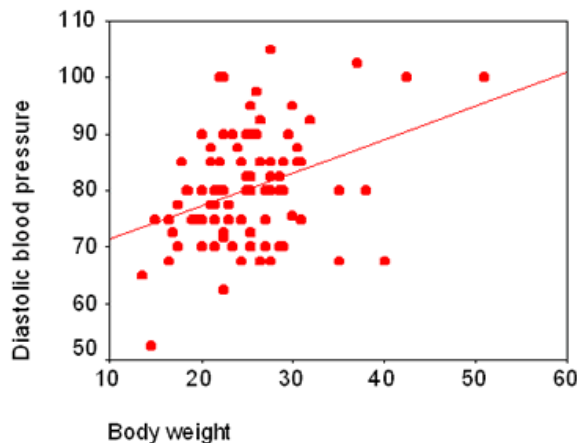
Trend analízis és előrejelzés



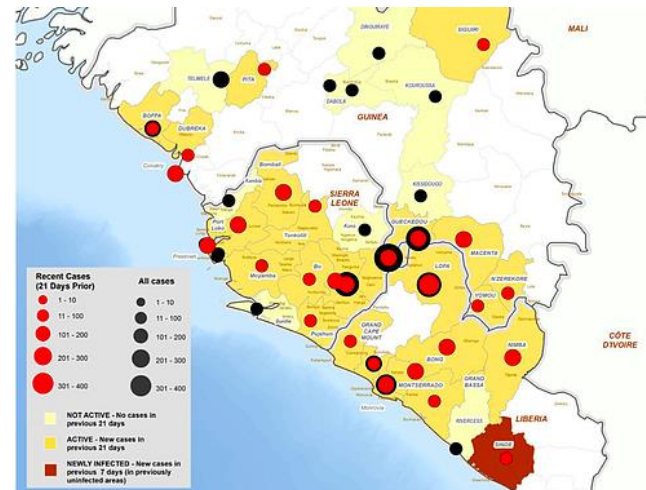
Idősor analízis



Korrelációanalízis

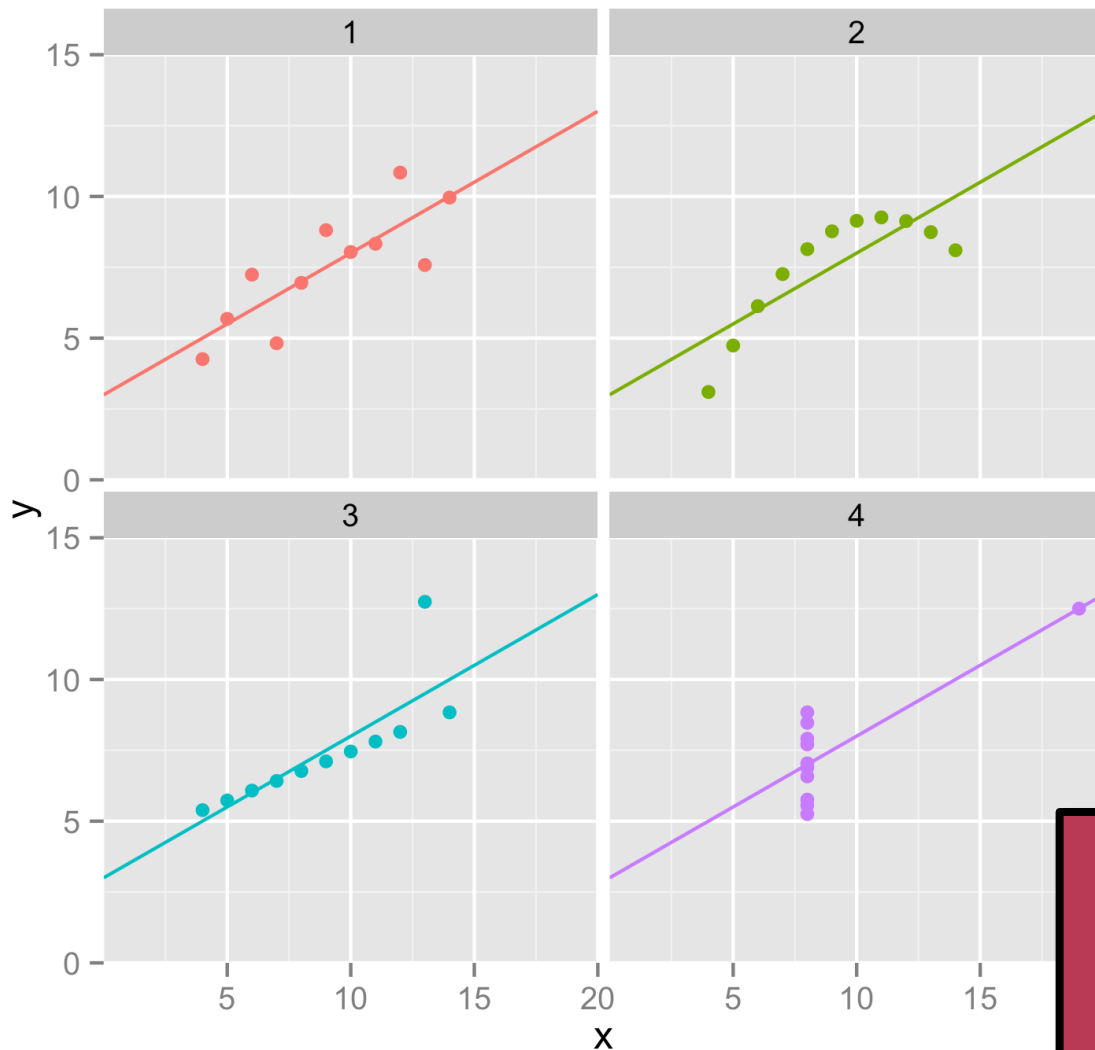


Térbeli analízis



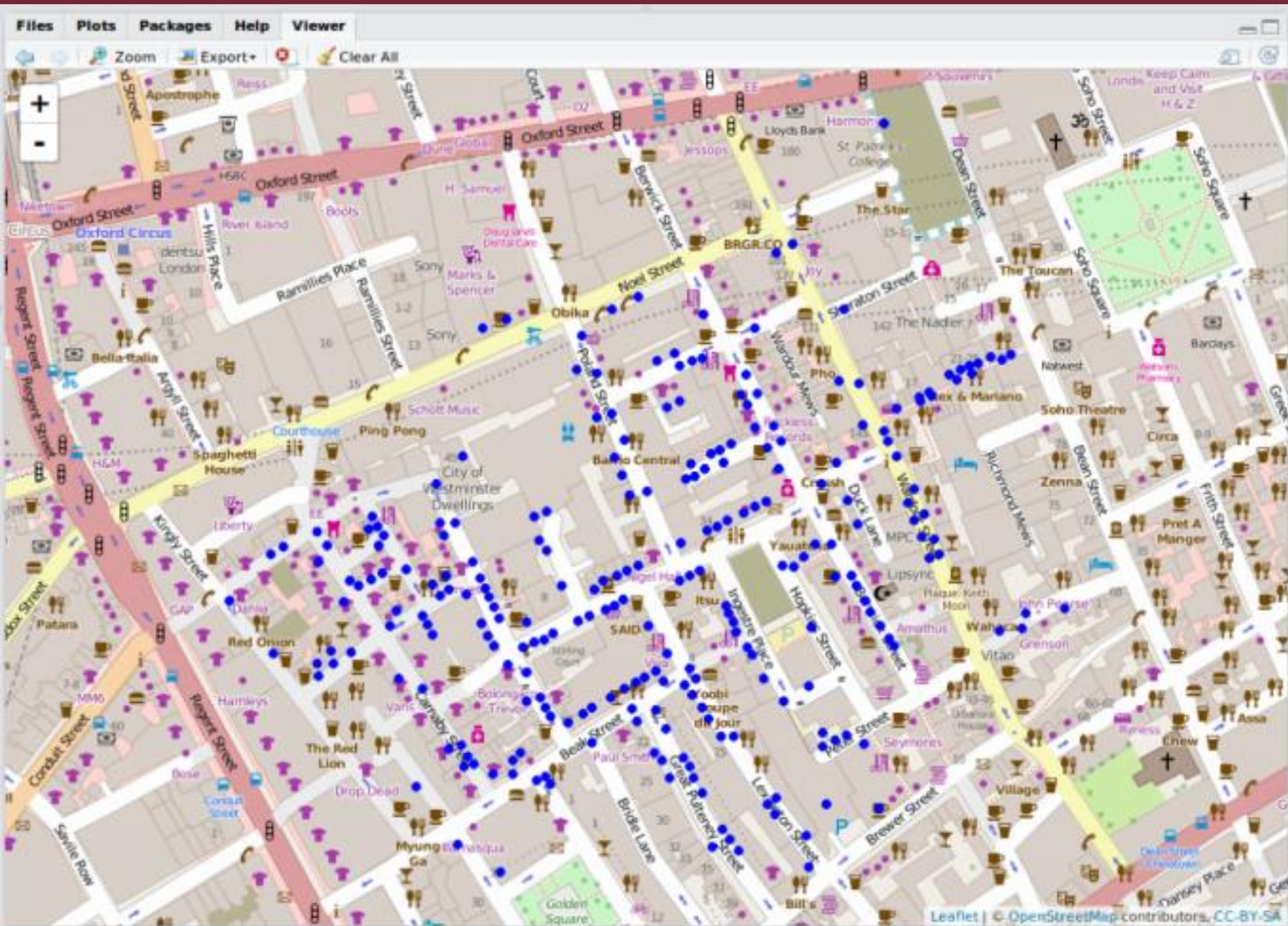
Számítások ellenőrzése

Anscombe's Quartet

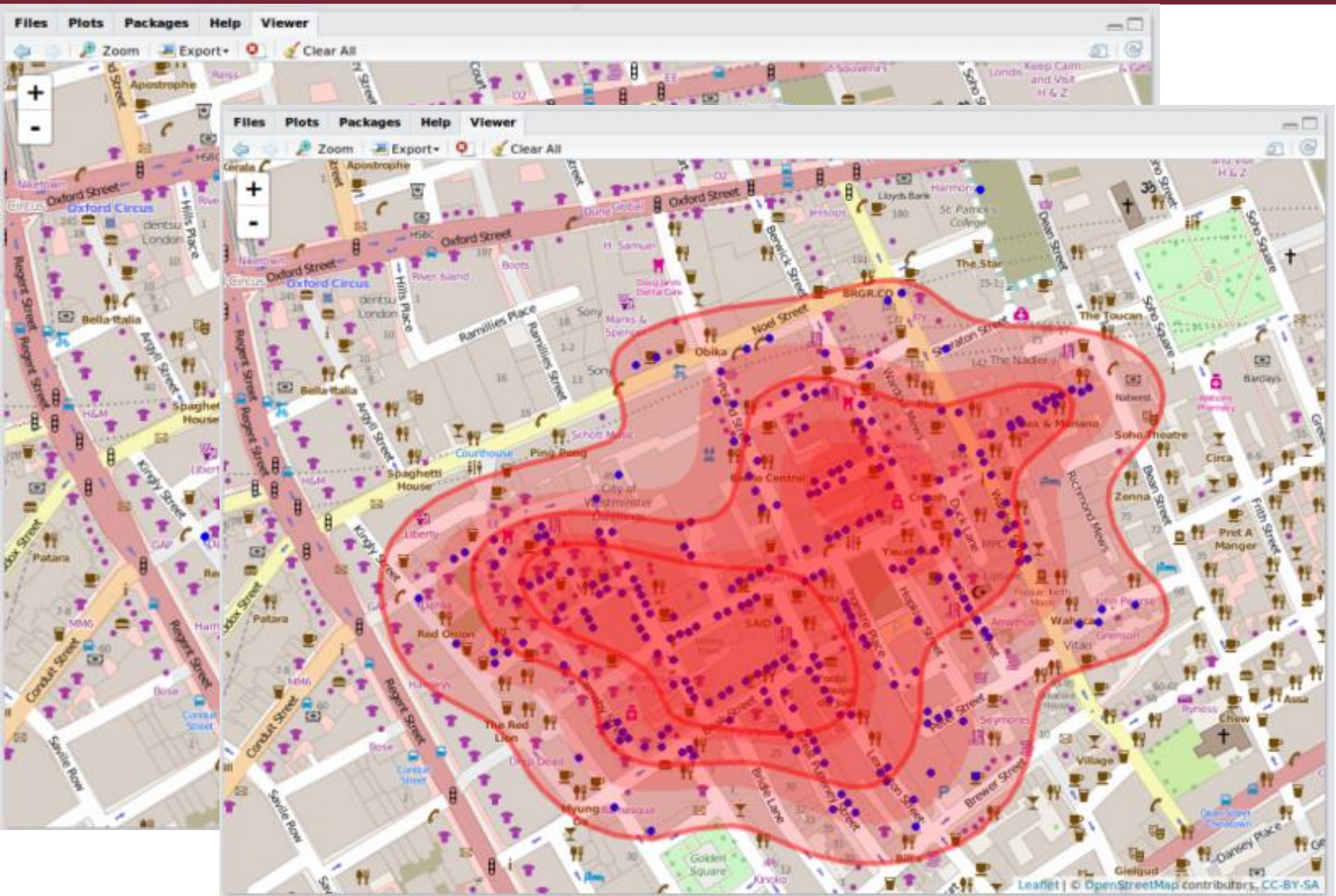


Hibás feltételezések
elkerülése... és intuíció

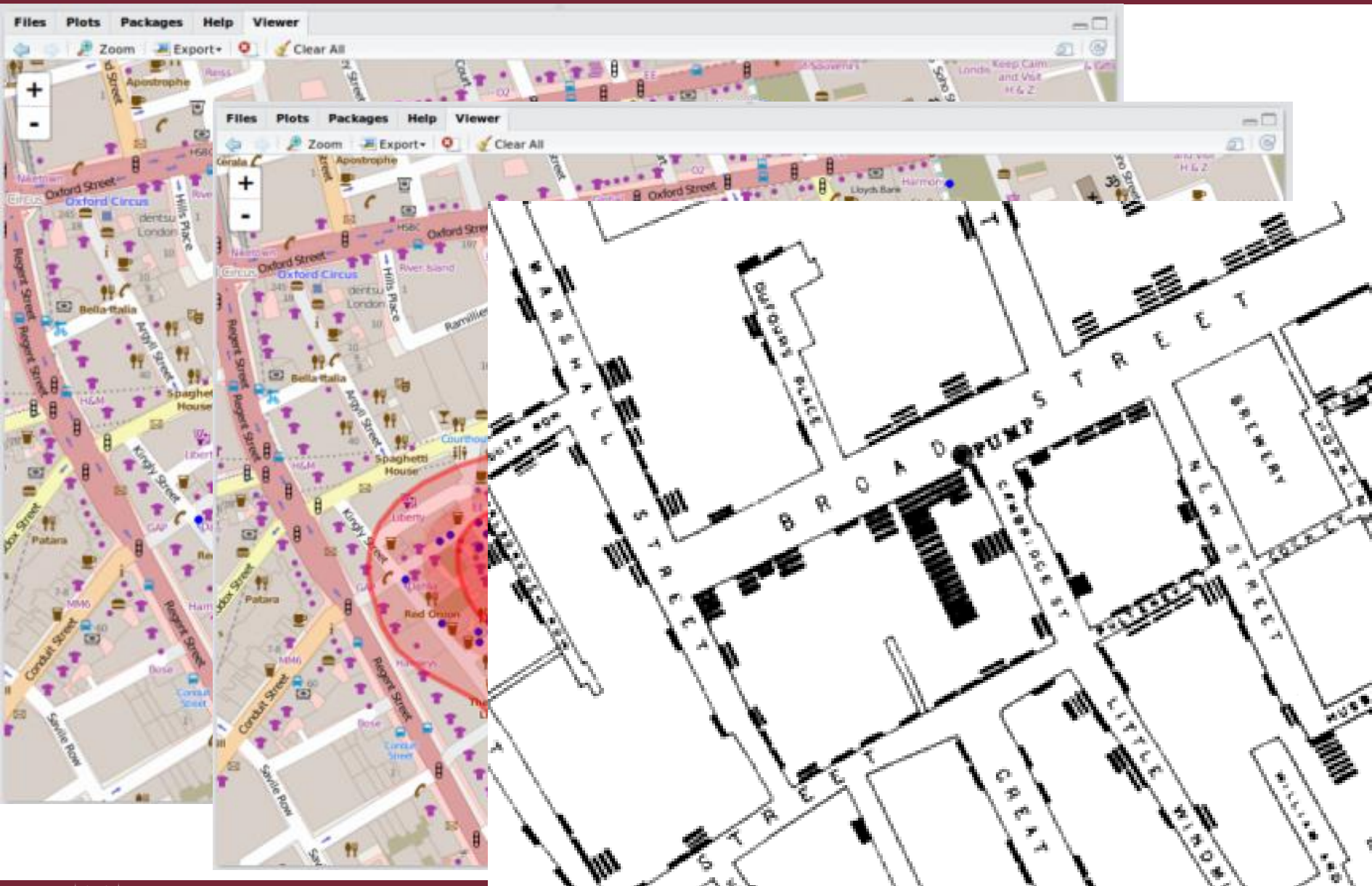
Összefüggések feltárása



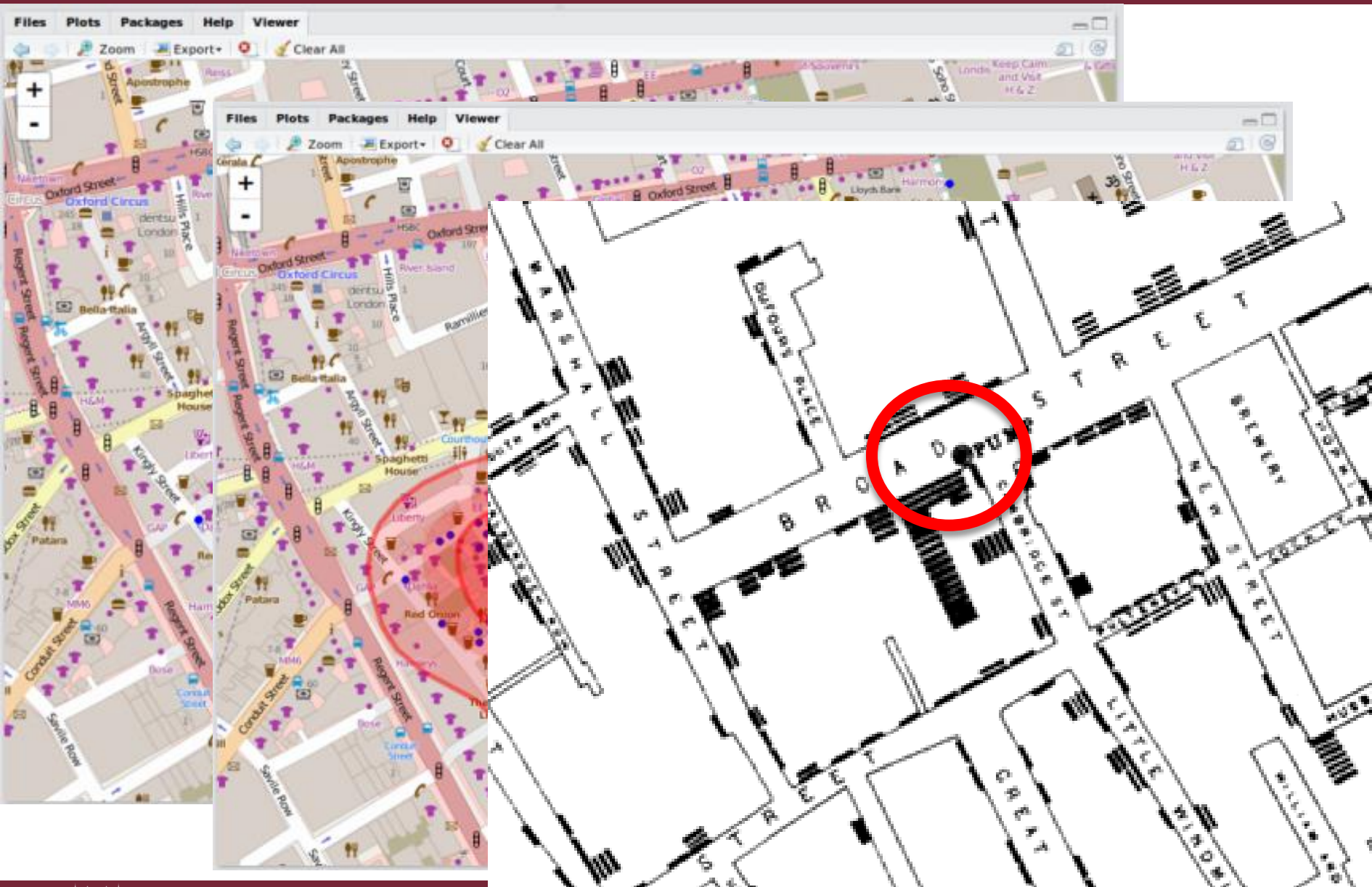
Összefüggések feltárása



Összefüggések feltárása



Összefüggések feltárása



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

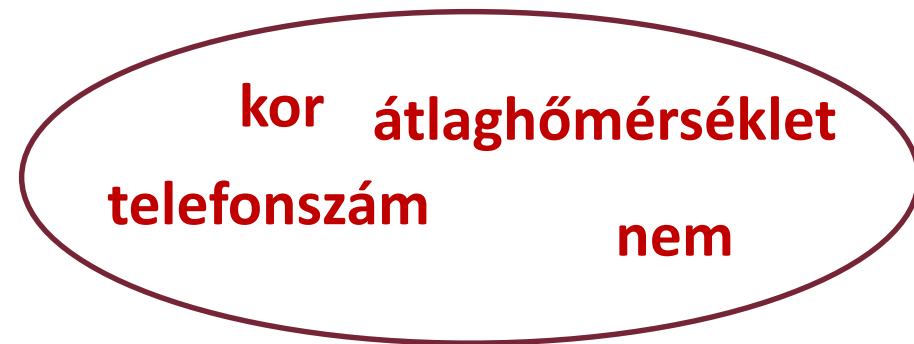
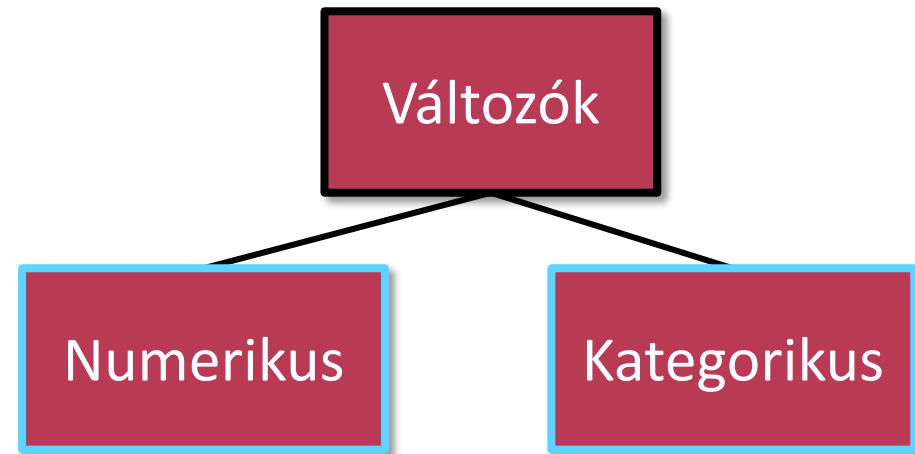
Numerikus és kategorikus változók

- Numerikus (numerical)

- az alapvető aritmetikai műveletek értelmesek

- Kategorikus (categorical)

- Matematikai műveletek nem értelmezhetőek rajtuk, legfeljebb sorba rendezés



Numerikus és kategorikus változók

- Numerikus (numerical)

- az alapvető aritmetikai

Ha egy változó értékei között az alapvető aritmetikai műveletek értelmezettek, akkor a változót **numerikusnak** tekintjük.

Minden más esetben – tehát ha megkülönböztetés, vagy legfeljebb sorba rendezés értelmes az értékek között – akkor a változót **kategorikusnak** tekintjük.

rajtuk, legfeljebb sorba
rendezés

telefonszám

nem

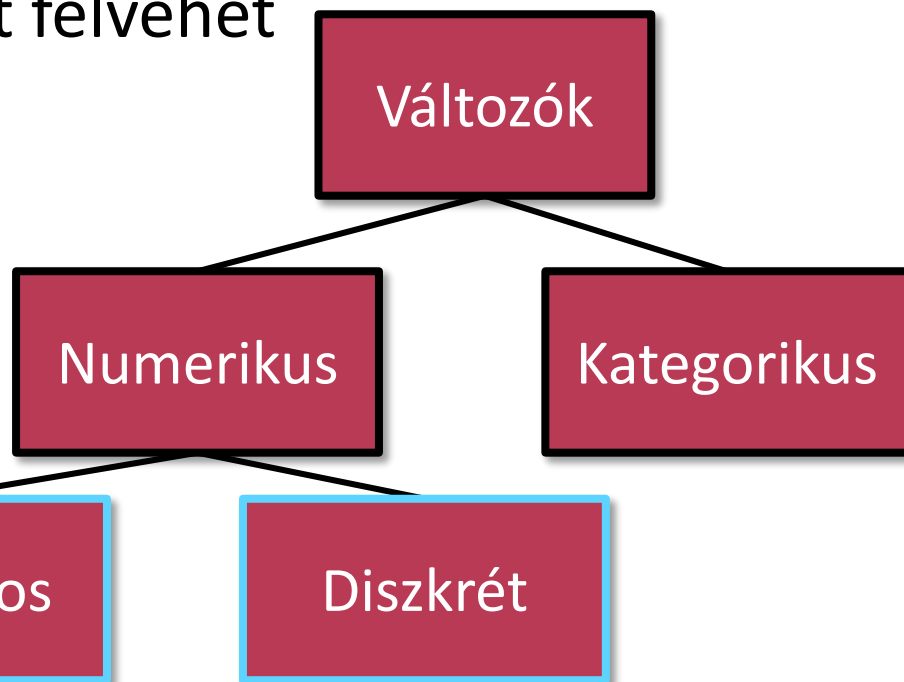
Numerikus változók

■ 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

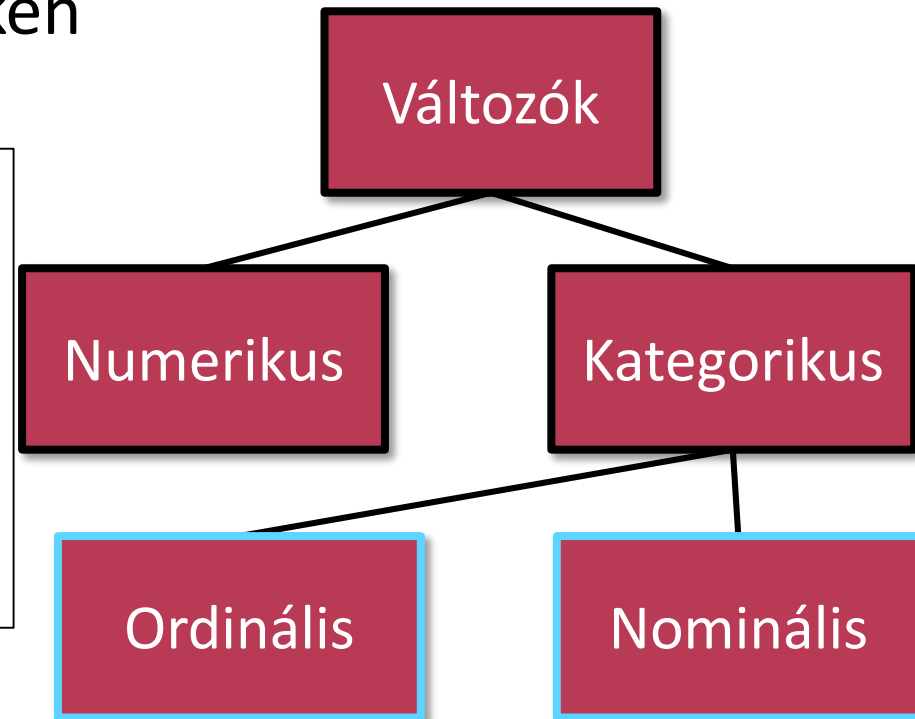
- Teljes rendezés az értékeken
- Pl. szállodai csillagok

9. Ajánlanád-e a tárgyat másoknak?

- ☐ Mindenkit rábeszelnék
- ☐ Nyugodtan ajánlanám
- ☐ Esetleg ajánlanám
- ☐ Inkább lebeszelném róla
- ☐ Feltétlenül lebeszelném
- ☐ Nem kívánok válaszolni

■ Nominális

- Pl. férfi, nő



Mi is lesz?

Miért vizualizálunk?



Mit vizualizálunk?



Hogyan vizualizálunk?



Mire következtetünk?

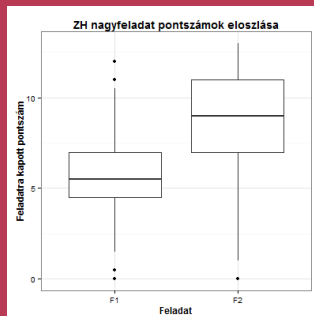
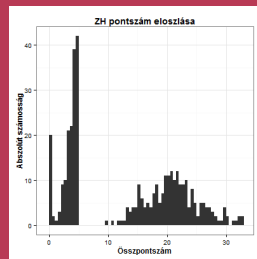
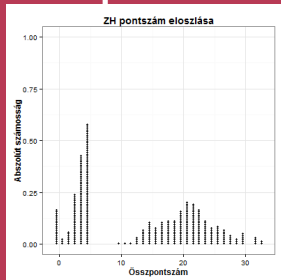
1 változó – eloszlásokra

Változók

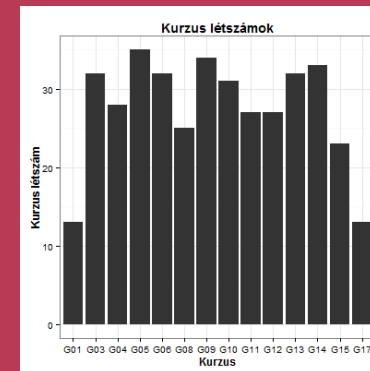
Numerikus

Kategorikus

ZH pontszám: [13, 15, 2, ...]



Kurzus: [G01, G03, G15, G17, ...]

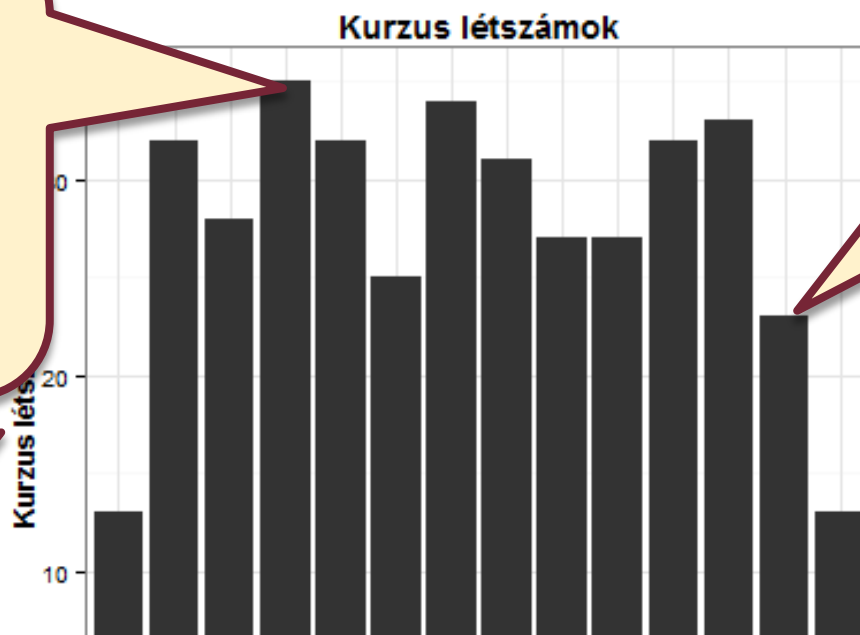


Oszlopdiagram

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

Vannak nagyon népszerű időpontok/gyakvezek?

abszolút gyakoriság!



Oszlop-magasság: adott érték gyakorisága

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

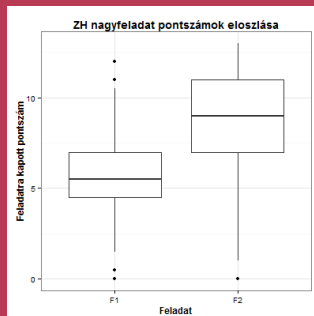
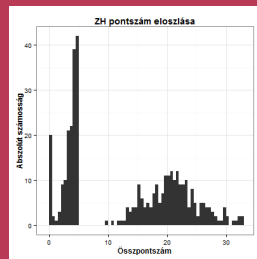
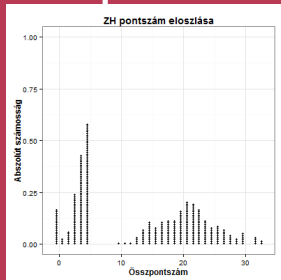
1 változó – eloszlásokra

Változók

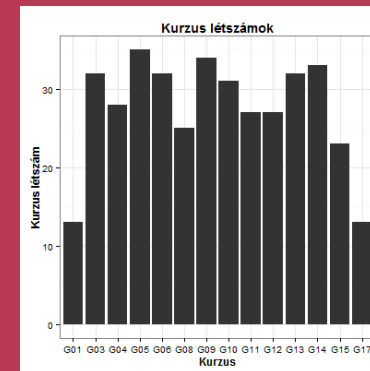
Numerikus

Kategorikus

ZH pontszám: [13, 15, 2, ...]



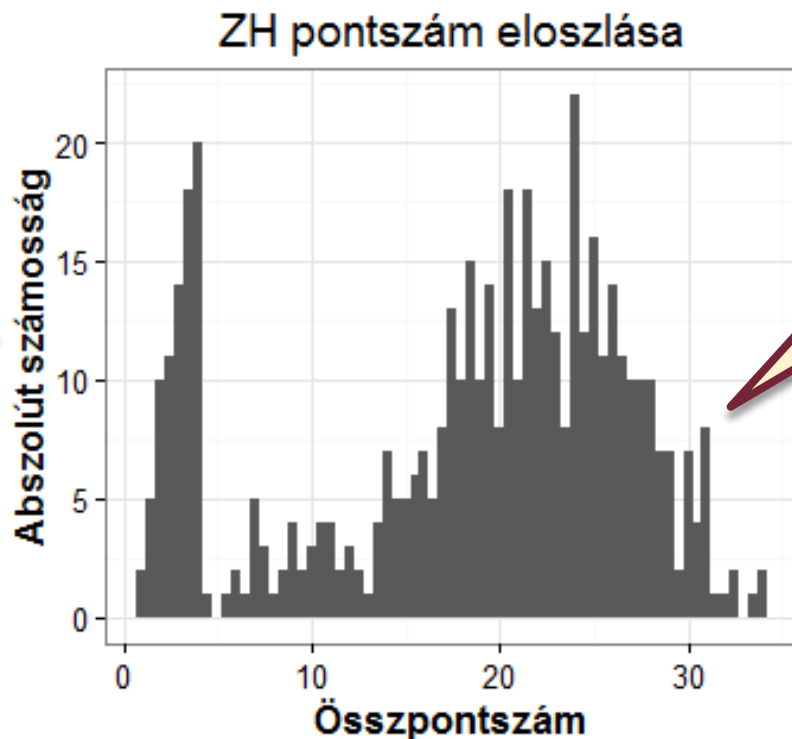
Kurzus: [G01, G03, G15, G17, ...]



Hisztogram

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

abszolút
gyakoriság!

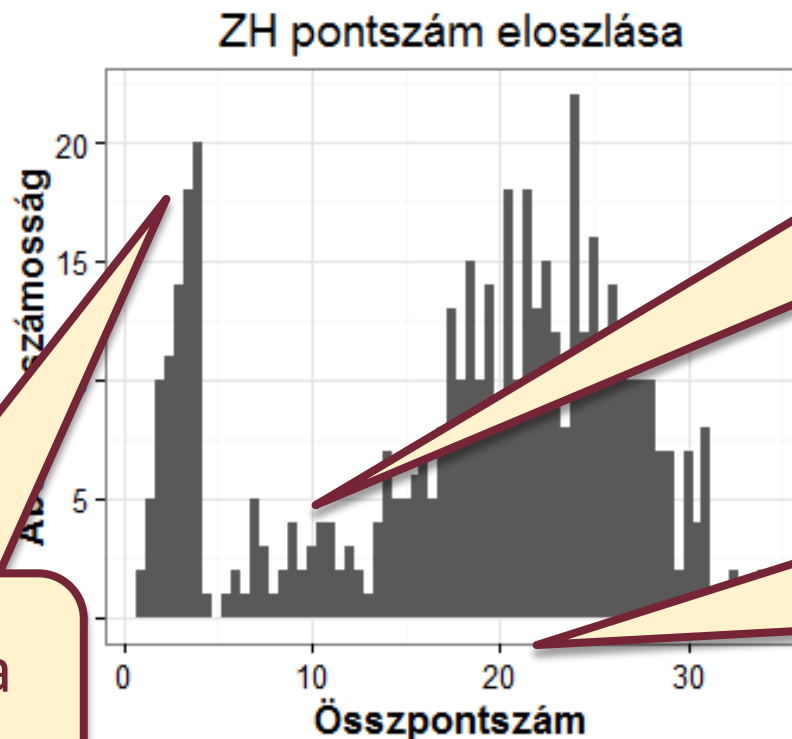


Oszlop-
magasság:
adott
intervallum
számossága

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?



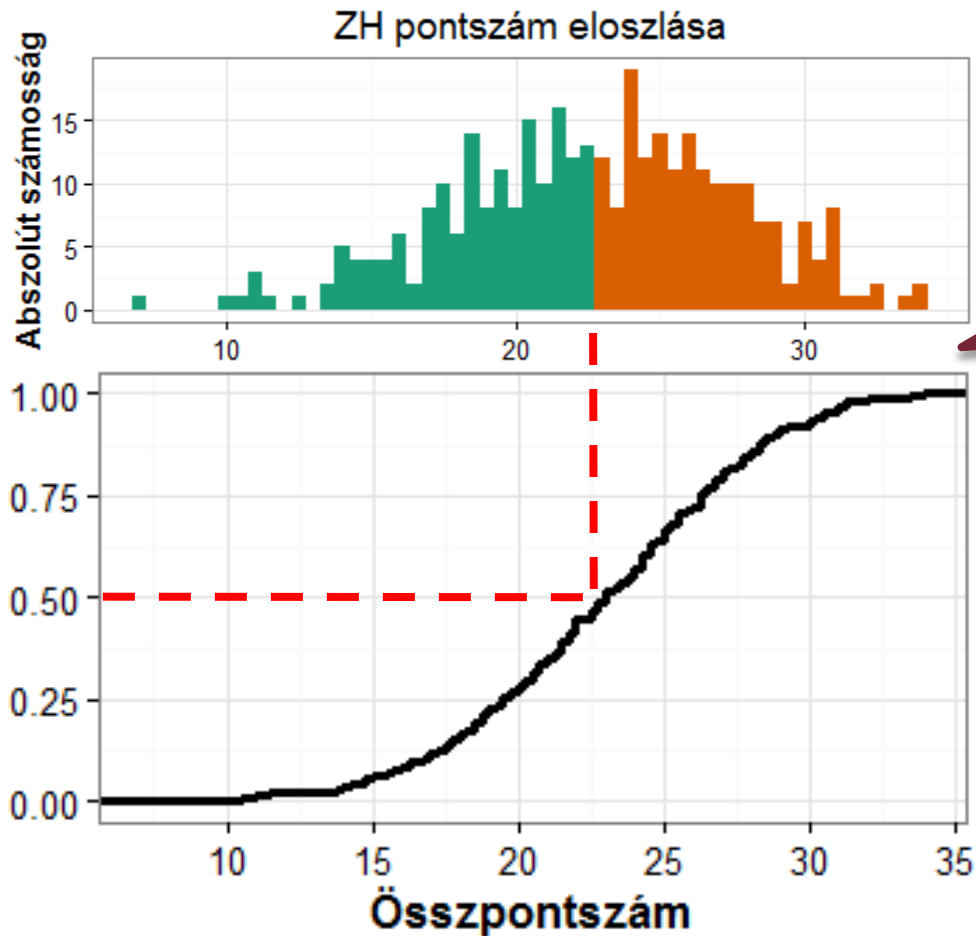
Sokan voltak a határon

Akik átmentek a beugrón, valószínűleg át is mentek

18 pont körül volt az átlag, 20 körül a medián

Egyszerű statisztikai jellemzés

- Hol van az adatok „közepe”?

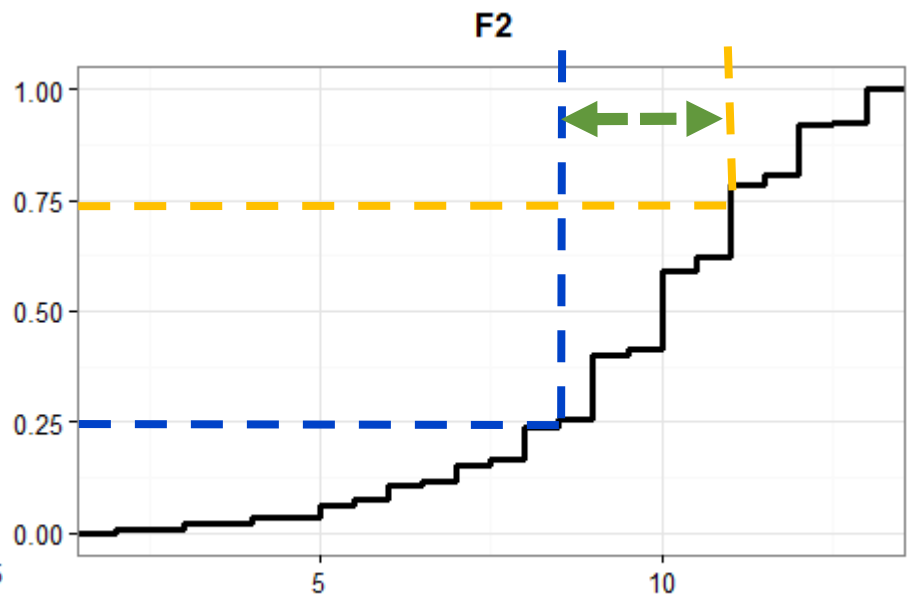
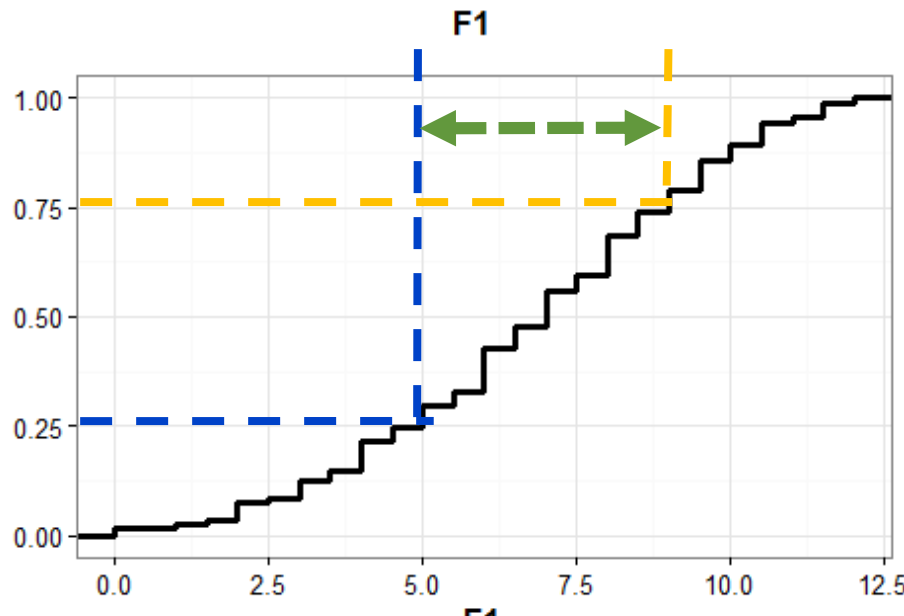
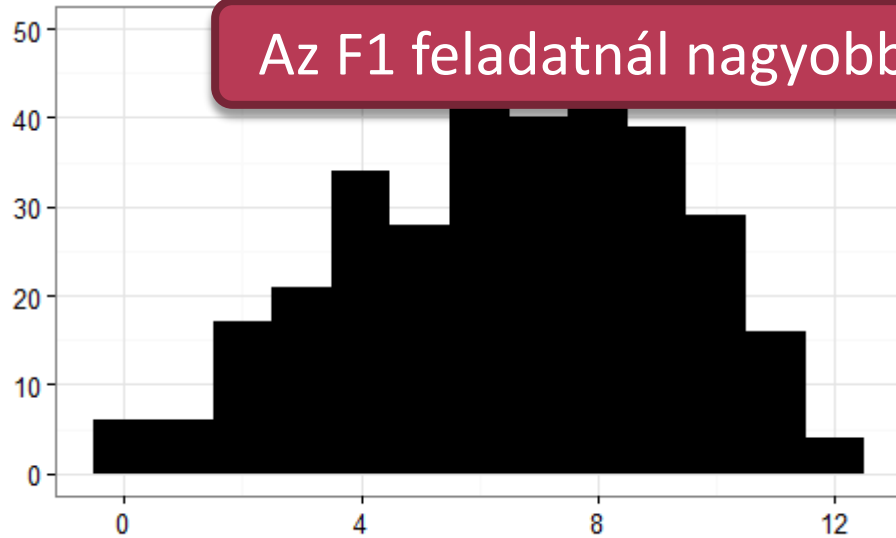


Az átmentek
összpontszám
ának mediánja
23

Egyszerű statisztikai jellemzés

■ Mennyire „szórtak” az adatok?

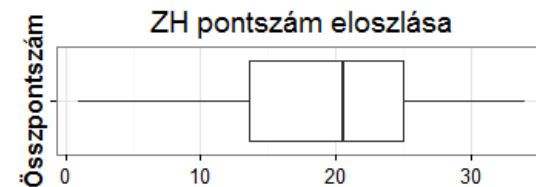
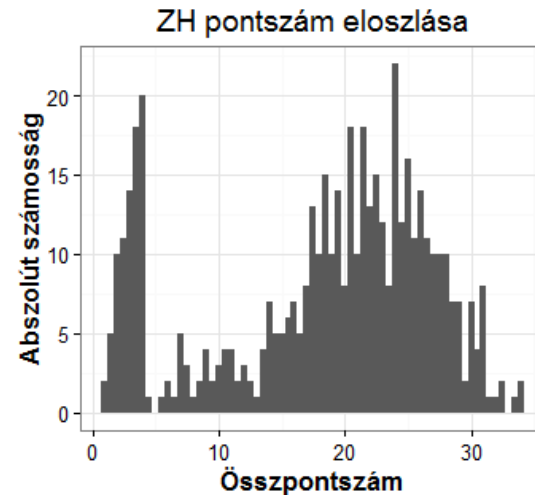
Az F1 feladatnál nagyobb volt a szórás, mint az F2-nél



Boxplot

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

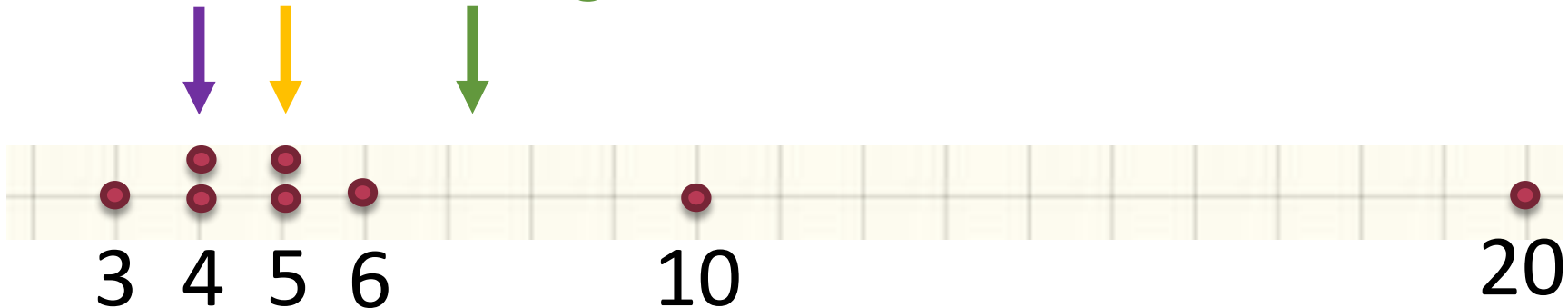
Egyfajta absztrakció itt is:
legyenek intervallumok,
felesleges minden pontot
kirajzolni



(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

módusz medián átlag



(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 módusról beszélünk.

Terjedelem jellemzése: percentilisek

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

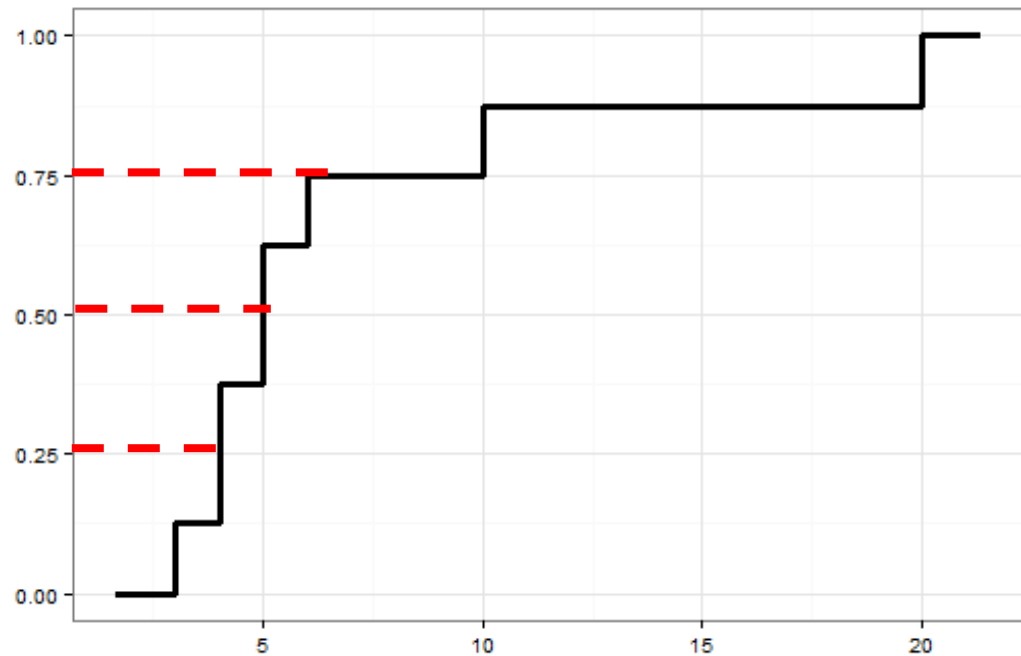
■ Percentilis

○ {3, 4, 4, 5, 5, 6, 10, 20}

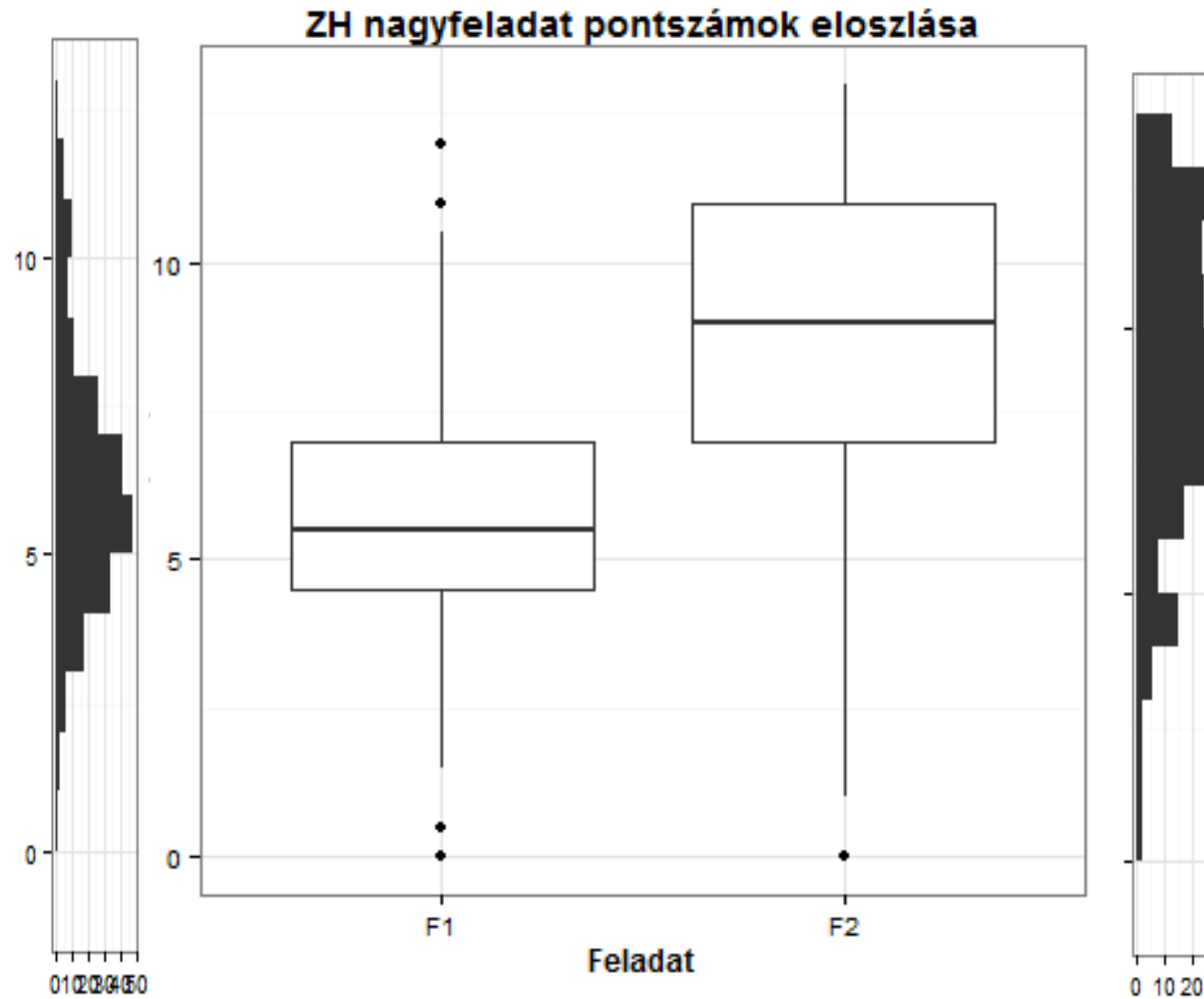
- 50. percentilis: 5
- 25. percentilis: 4
- 75. percentilis: 6

■ Kvartilis

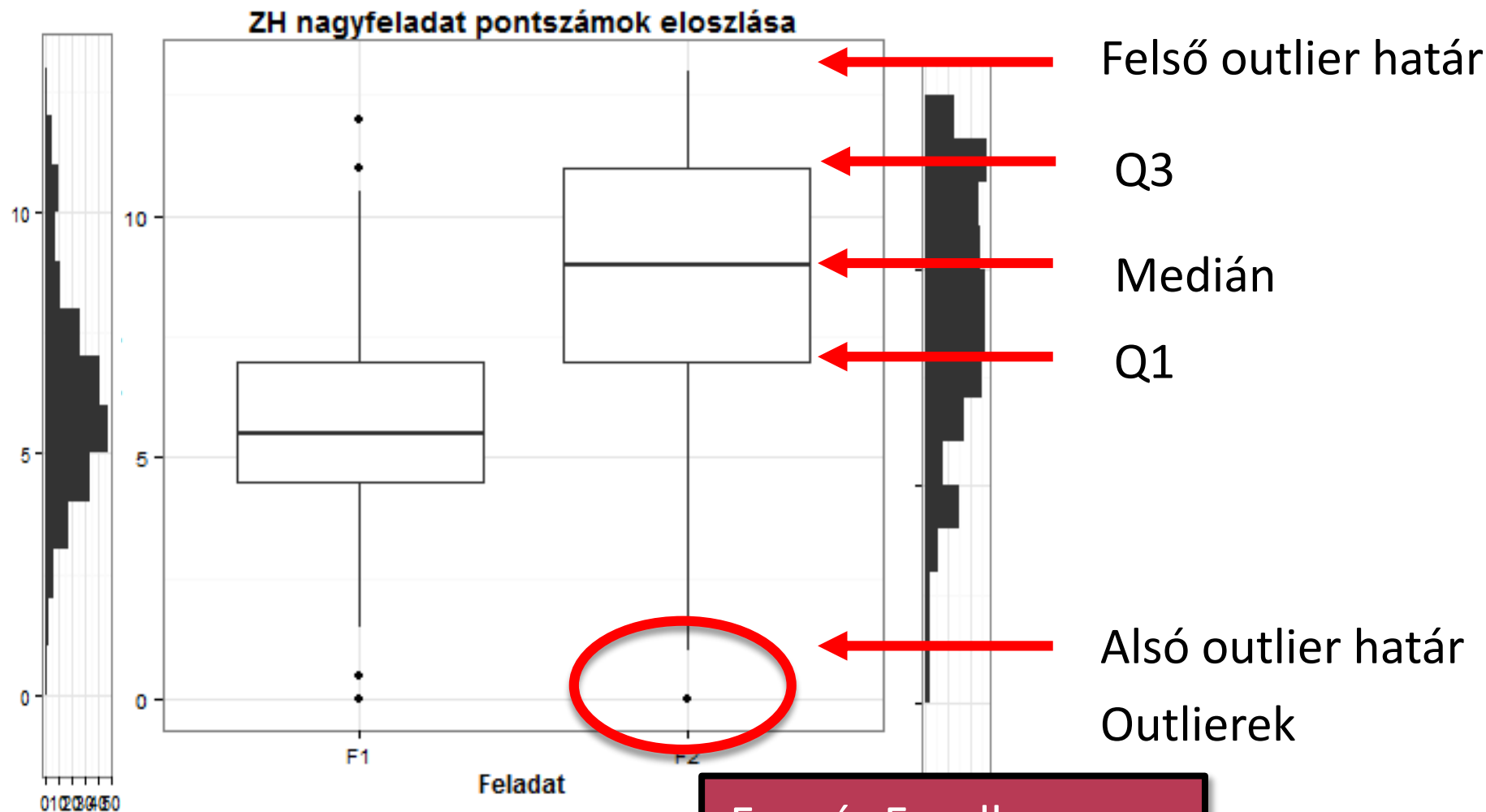
- Q1: 25. percentilis
- Q3: 75. percentilis
- **Q2: medián**



Boxplot (Box and whisker plot)

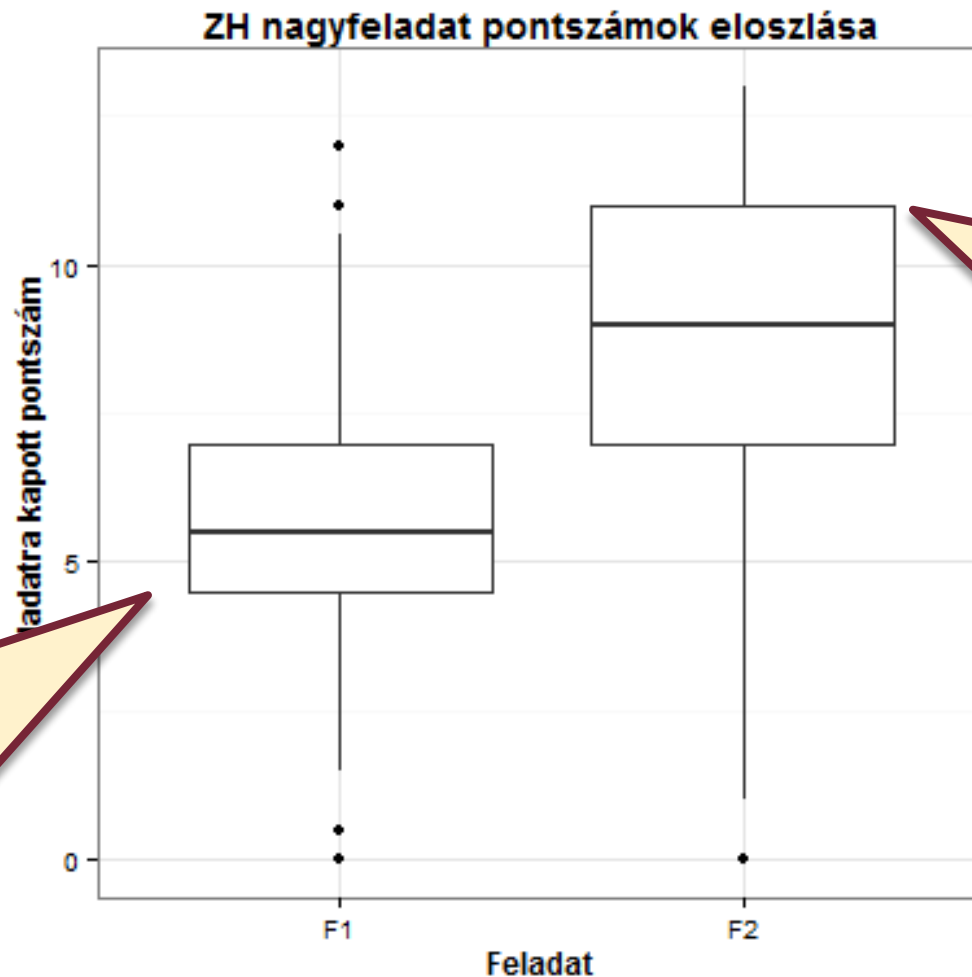


Boxplot (Box and whisker plot)



Ez már Excelben nem
könnyű...

Boxplot (Box and whisker plot)

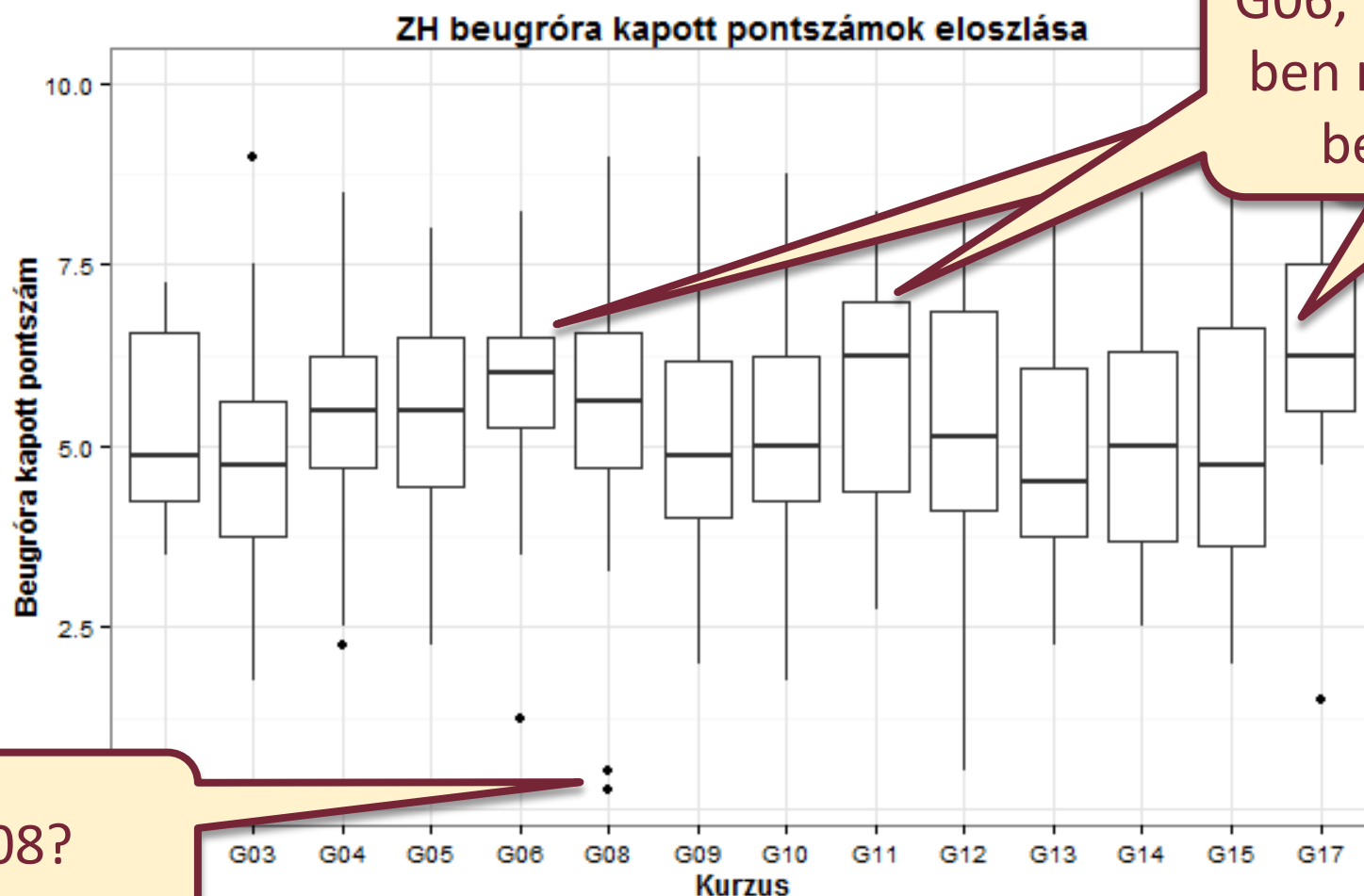


Az F1
pontszámok
50%-a
4.5 és 7.5
között volt

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

Boxplot (Box and whisker plot)

- Melyik csoportban hogyan sikerültek a beugrók?

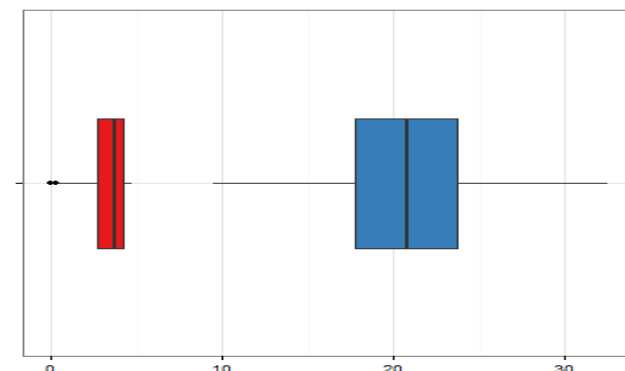
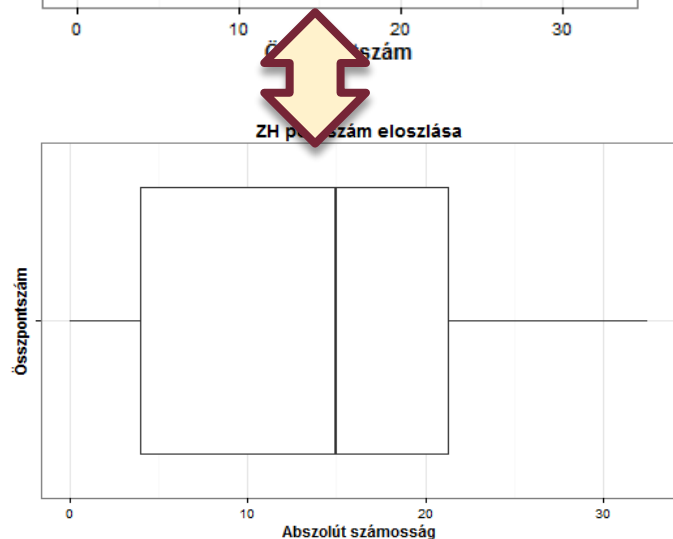
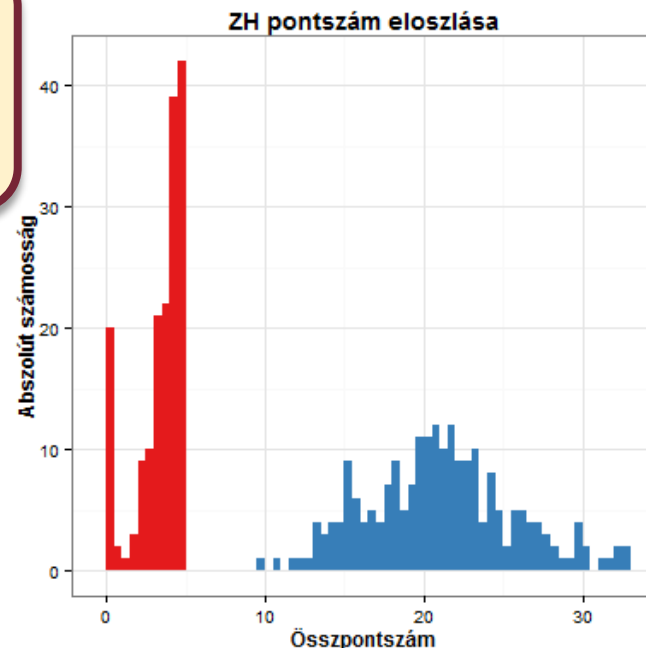
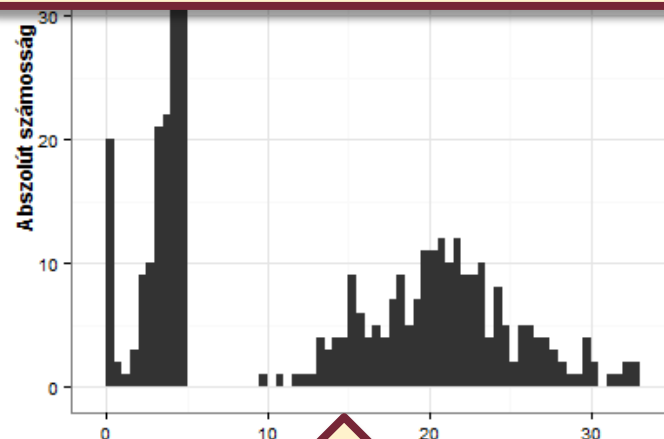


G06, G11, G17-
ben nagyon jó
beugrók

G08?

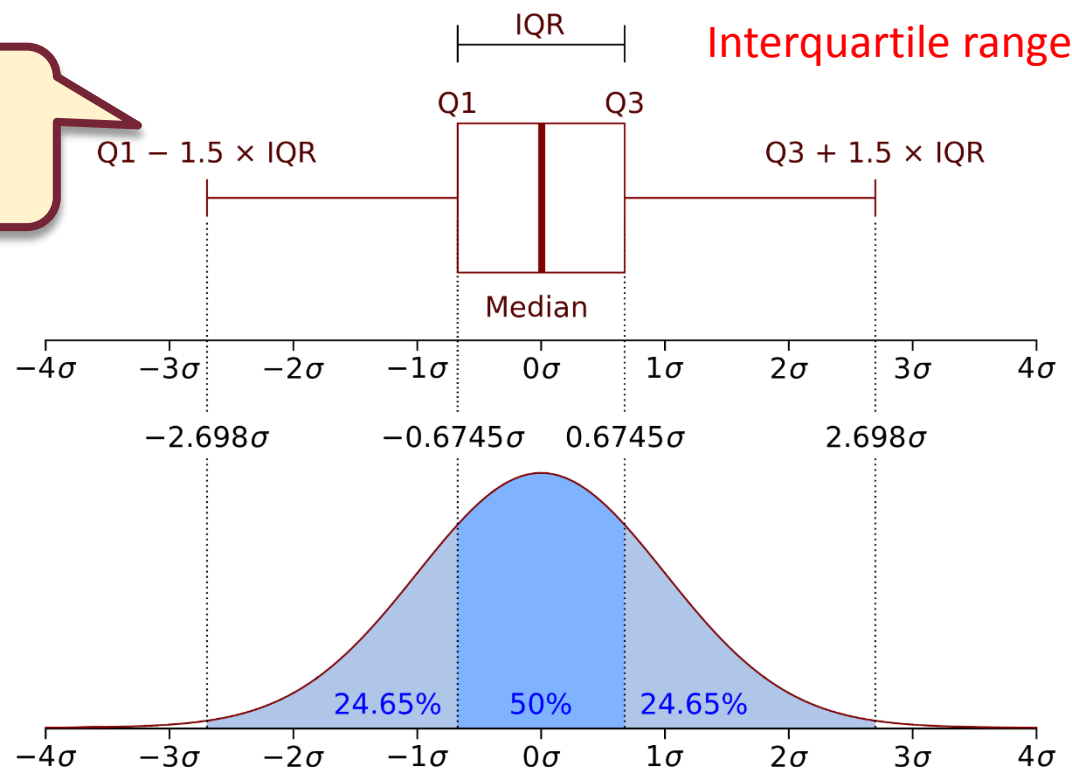
Boxplot (Box and whisker plot)

Absztrakció: a boxplottal fontos információt is veszíthetünk!



Boxplot (Box and whisker plot)

A ± 1.5 csak
konvenció



Ez majd valószínűesszámításból érthetőbb lesz:
“normális eloszlásnál ez kb. a $\pm 3\sigma$ -nak felel meg”

Miért medián, miért nem átlag?

■ Alaphalmaz

○ 1000 pont $\sim U(1, 5)$ egyenletes eloszlás

• *átlag = medián = 3 ms*

1 pont: 20 s
(= 20 000)

3ms $\pm$ 2 ms



Válaszidő

Új medián: `sort(resp. times)[501] = 3.02 ms`

Vál. medián

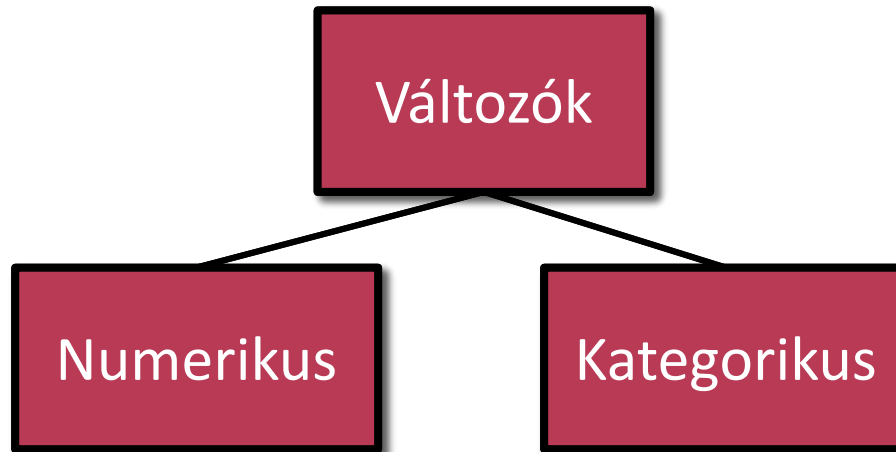
→ ✓ Robusztus

Vál. átlag

→ ✗ Nem rob.

Új átlag: $\frac{(3 \times 10^3 + 2 \times 10^4)}{1001} = 23 \text{ ms!}$

2 változó kapcsolata

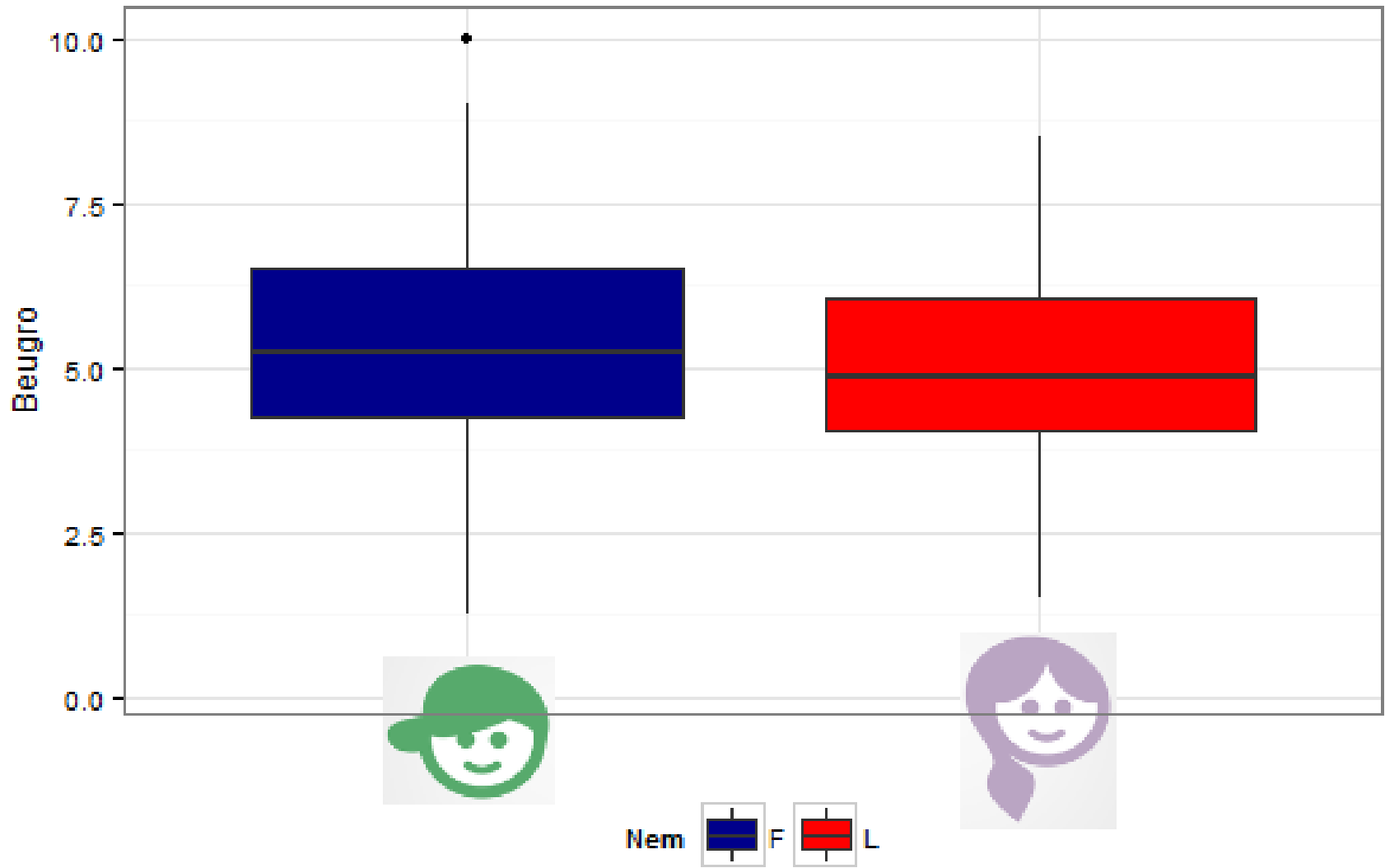


2 numerikus

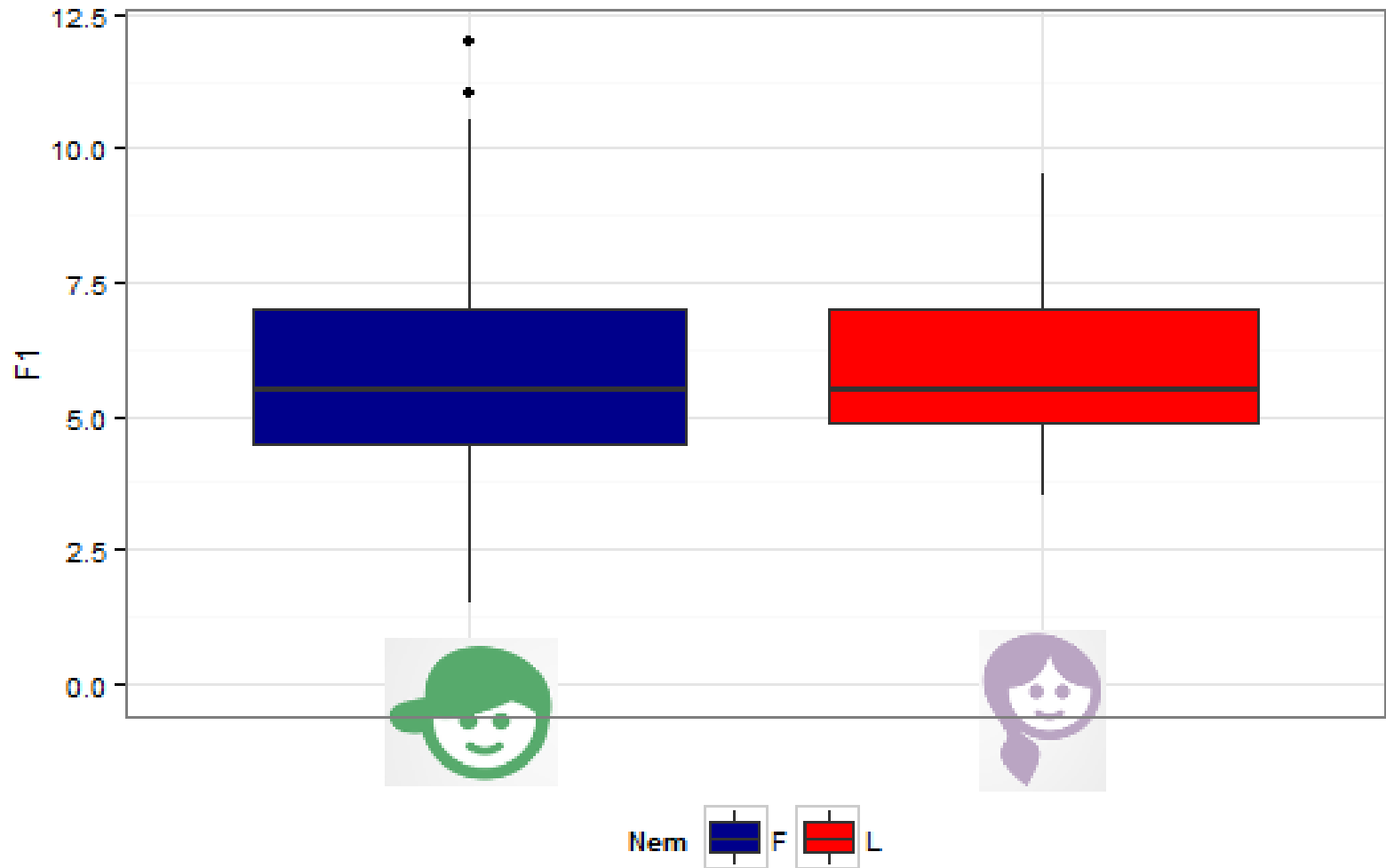
1 numerikus,
1 kategorikus

2 kategorikus

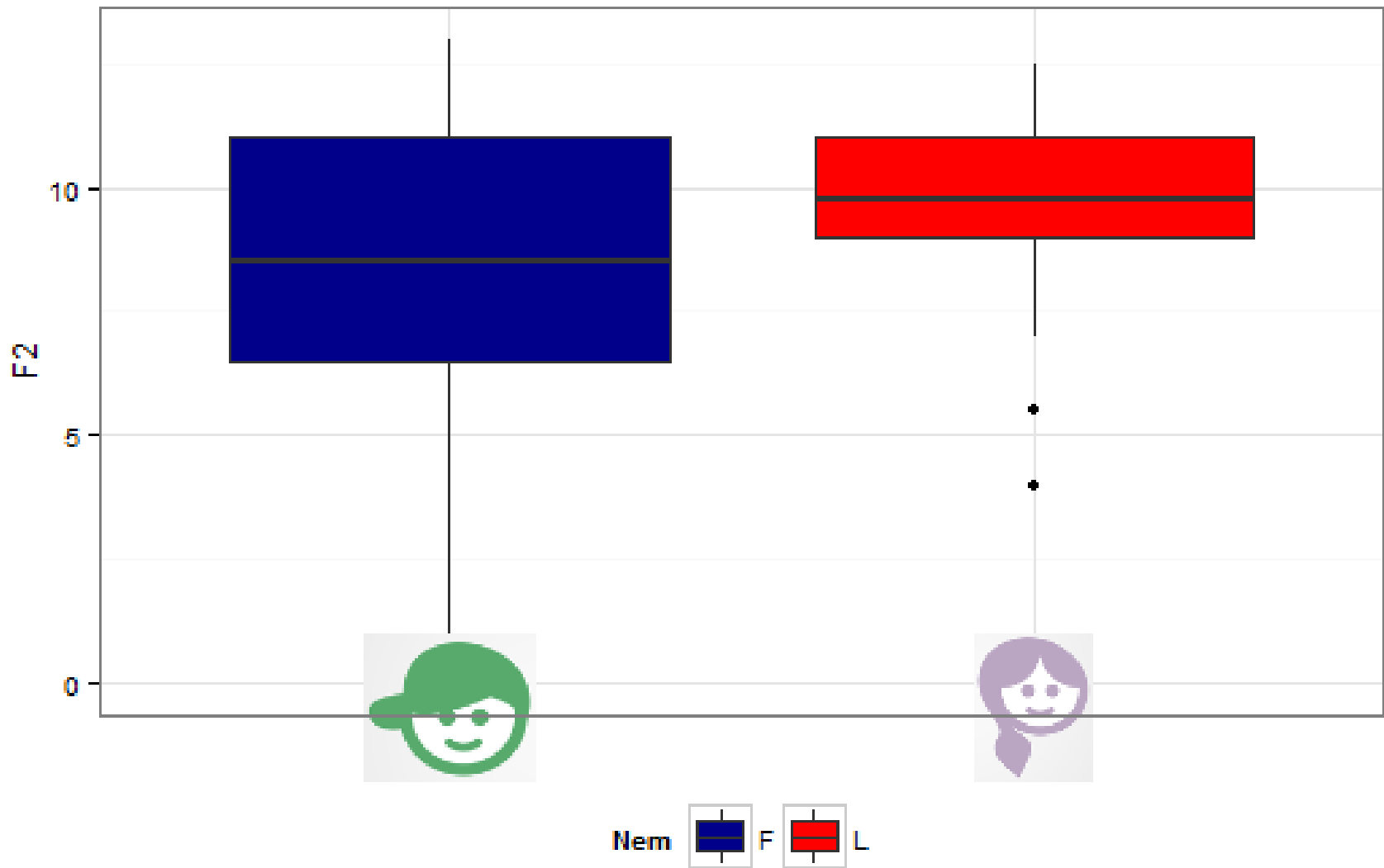
Numerikus kategóriánként



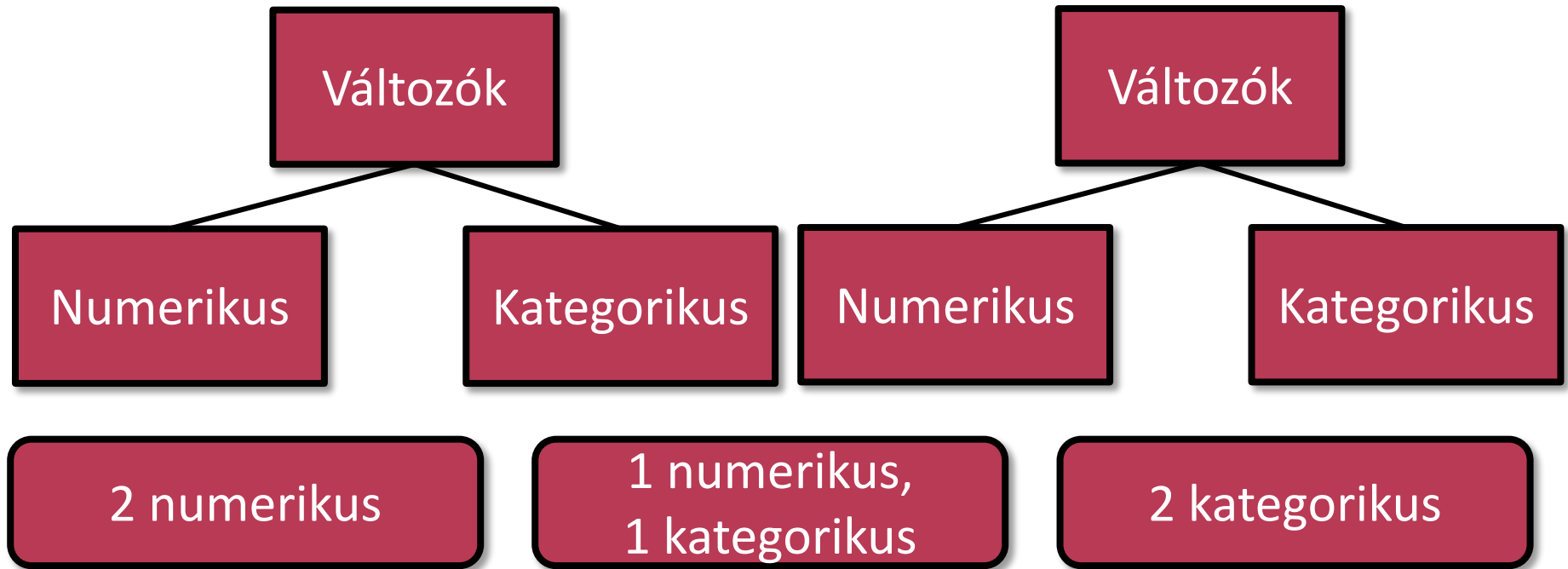
Numerikus kategóriánként



Numerikus kategóriánként



2 változó kapcsolata

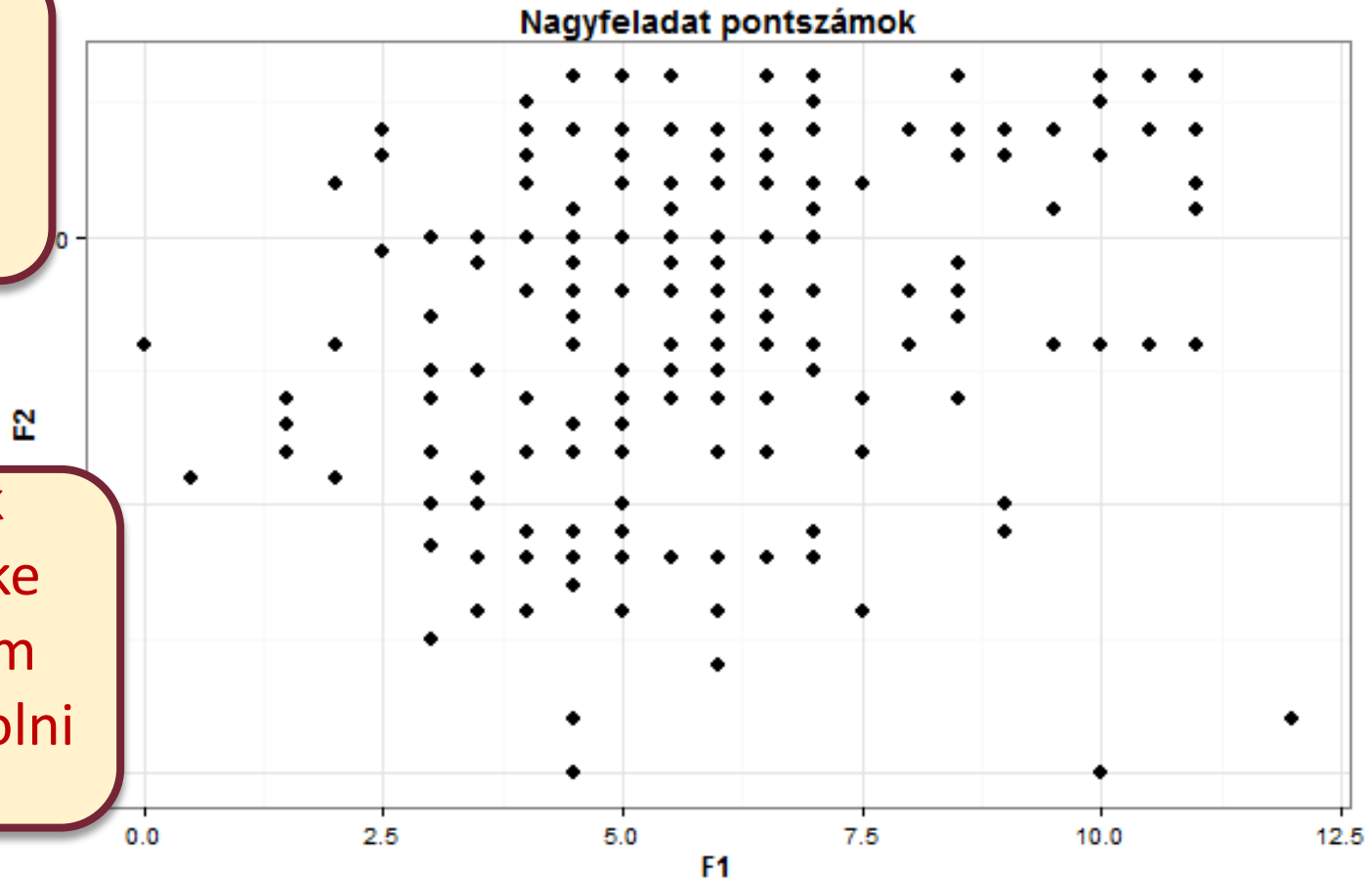


Pont – pont diagram (scatterplot)

- 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

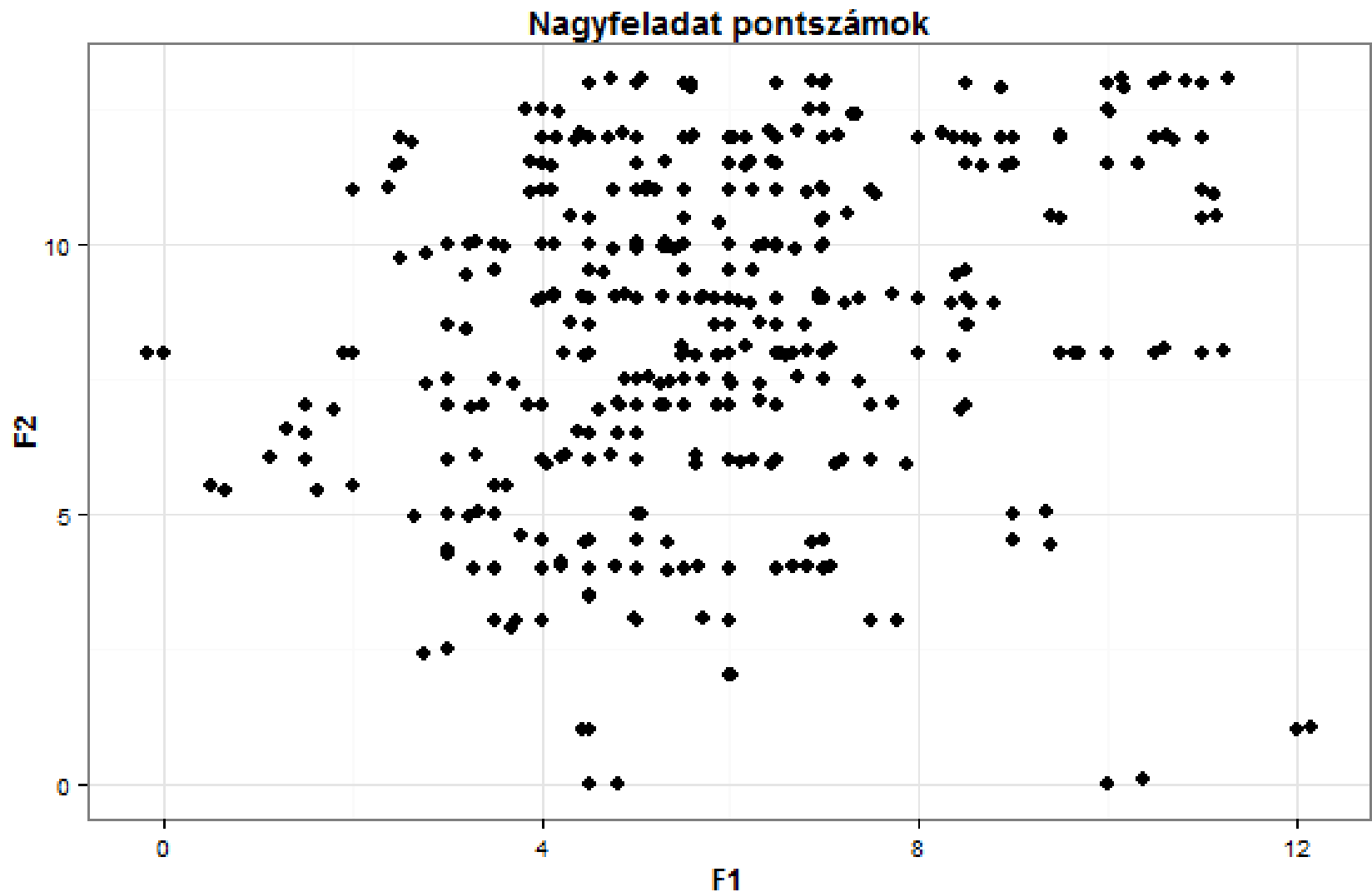


Hogyan kezeljük a takarásokat?

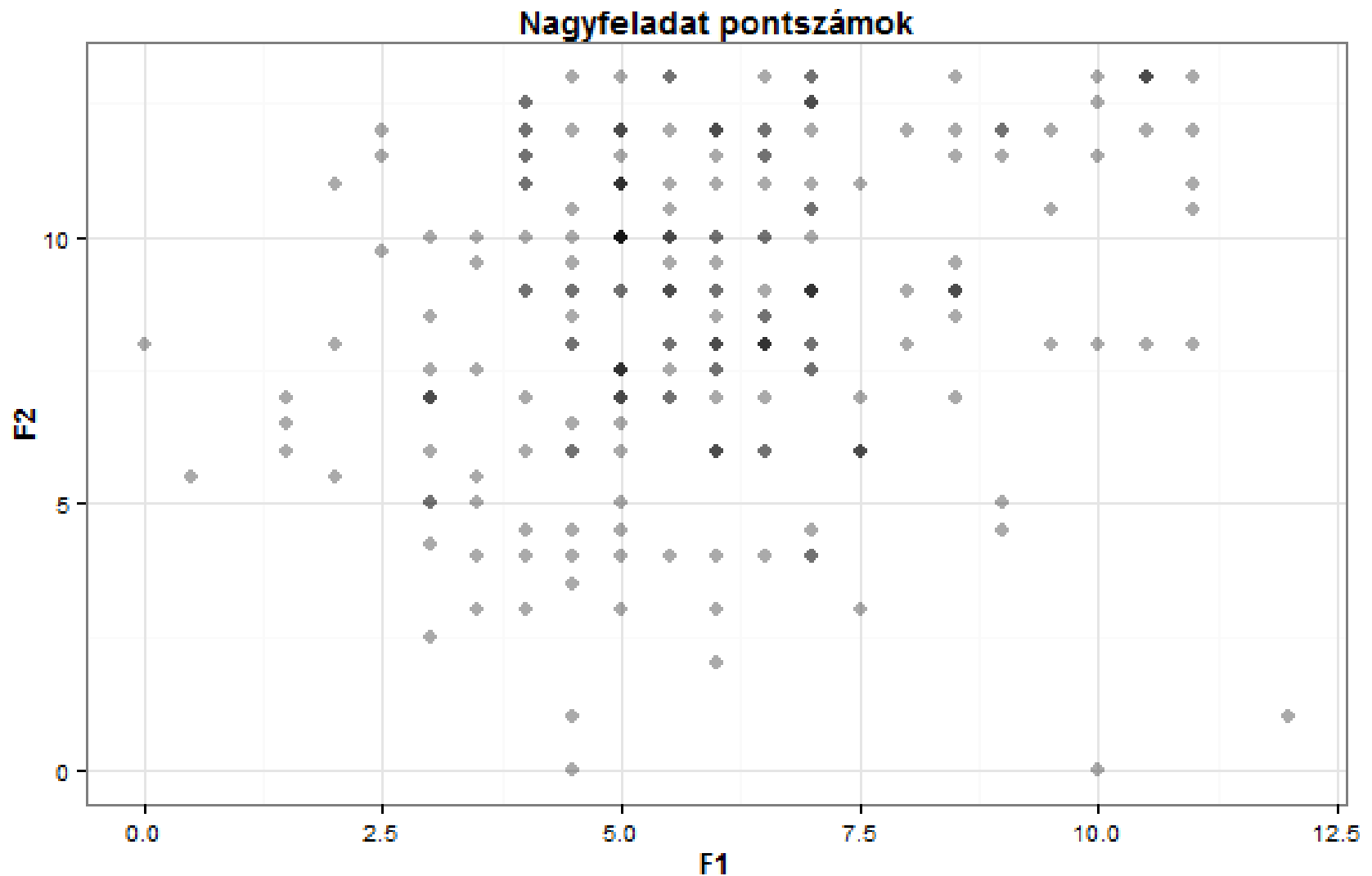
Overplotting



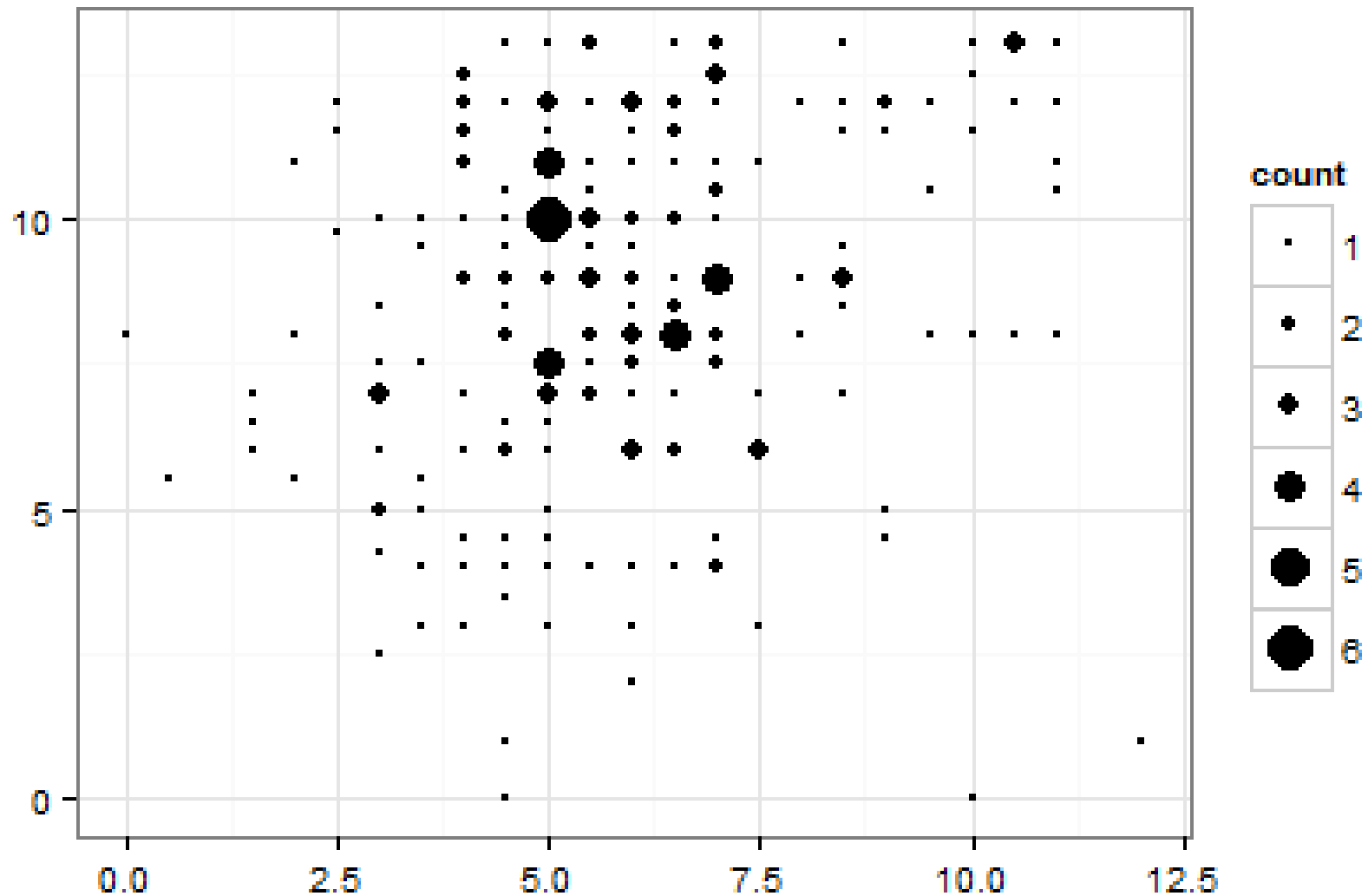
Overplotting megoldások 1: jitter



Overplotting megoldások 2: átlátszóság



Overplotting megoldások 3: méret

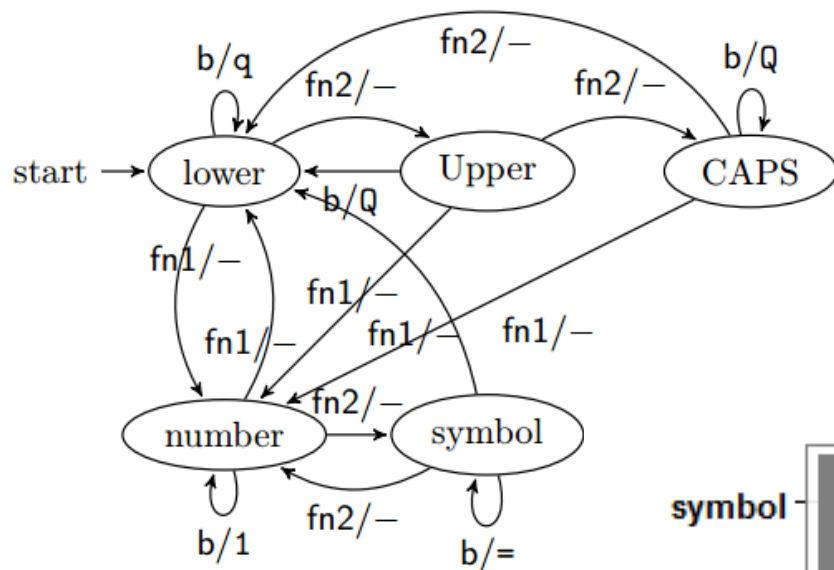


SOK VÁLTOZÓ

≥ 3 változó

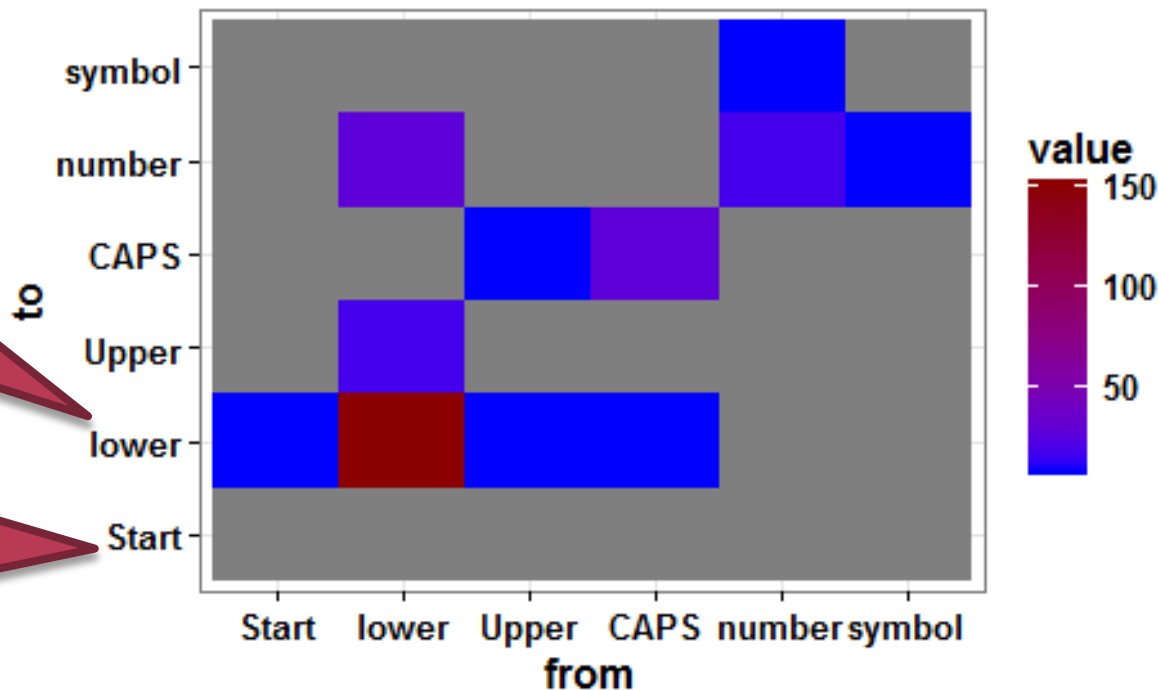
- A grafikai objektumok attribútumait változtatom
 - Szín
 - Méret
 - Textúra
 - Hely – ez triviálisnak tűnik, de a treemapnél van jelentősége
- Pl. heatmap, treemap

Heatmap: lefutási statisztikák

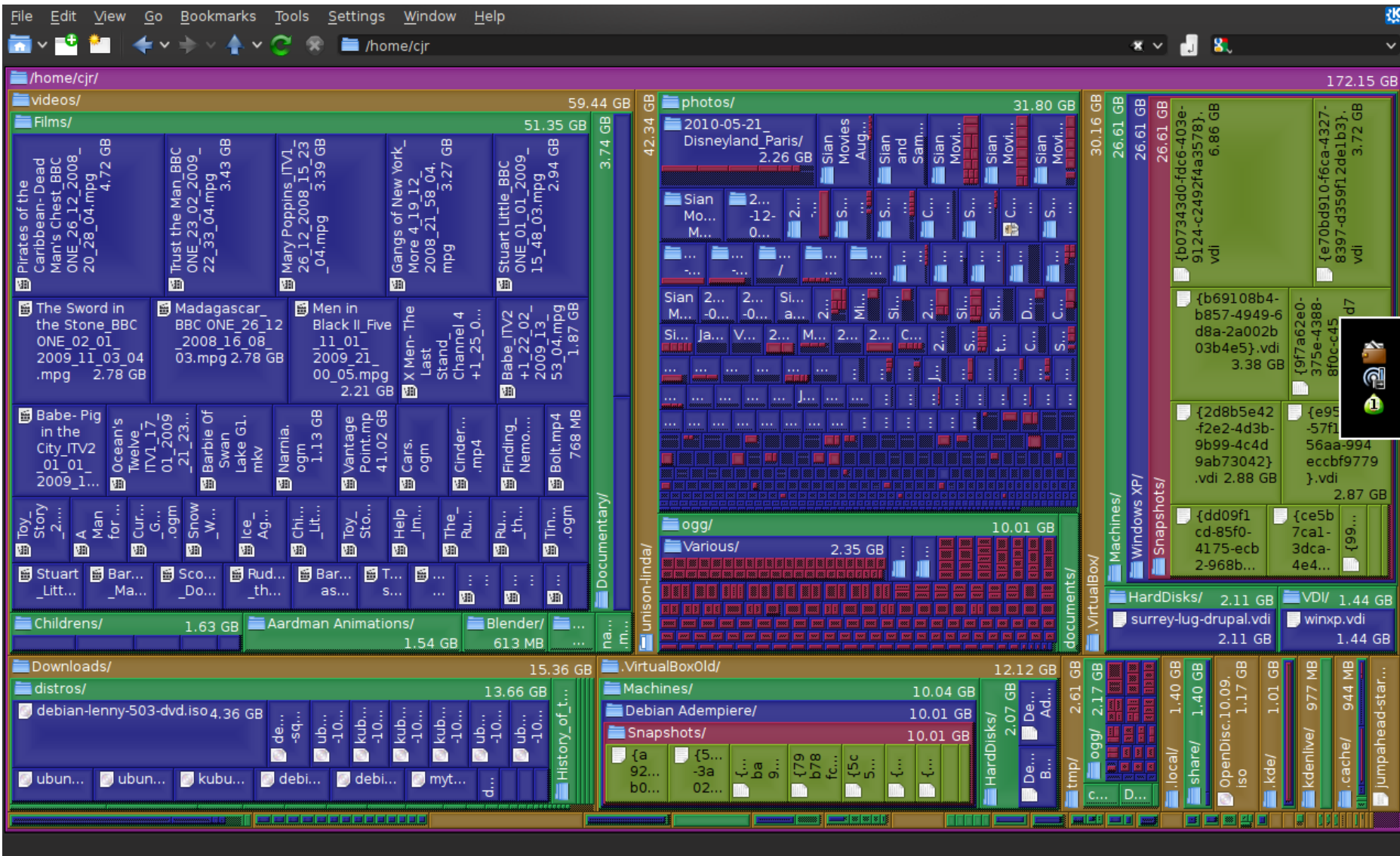


Inkább csak sima
szöveget írunk

A Startban mindig
csak kezdünk, oda
nem jutunk vissza

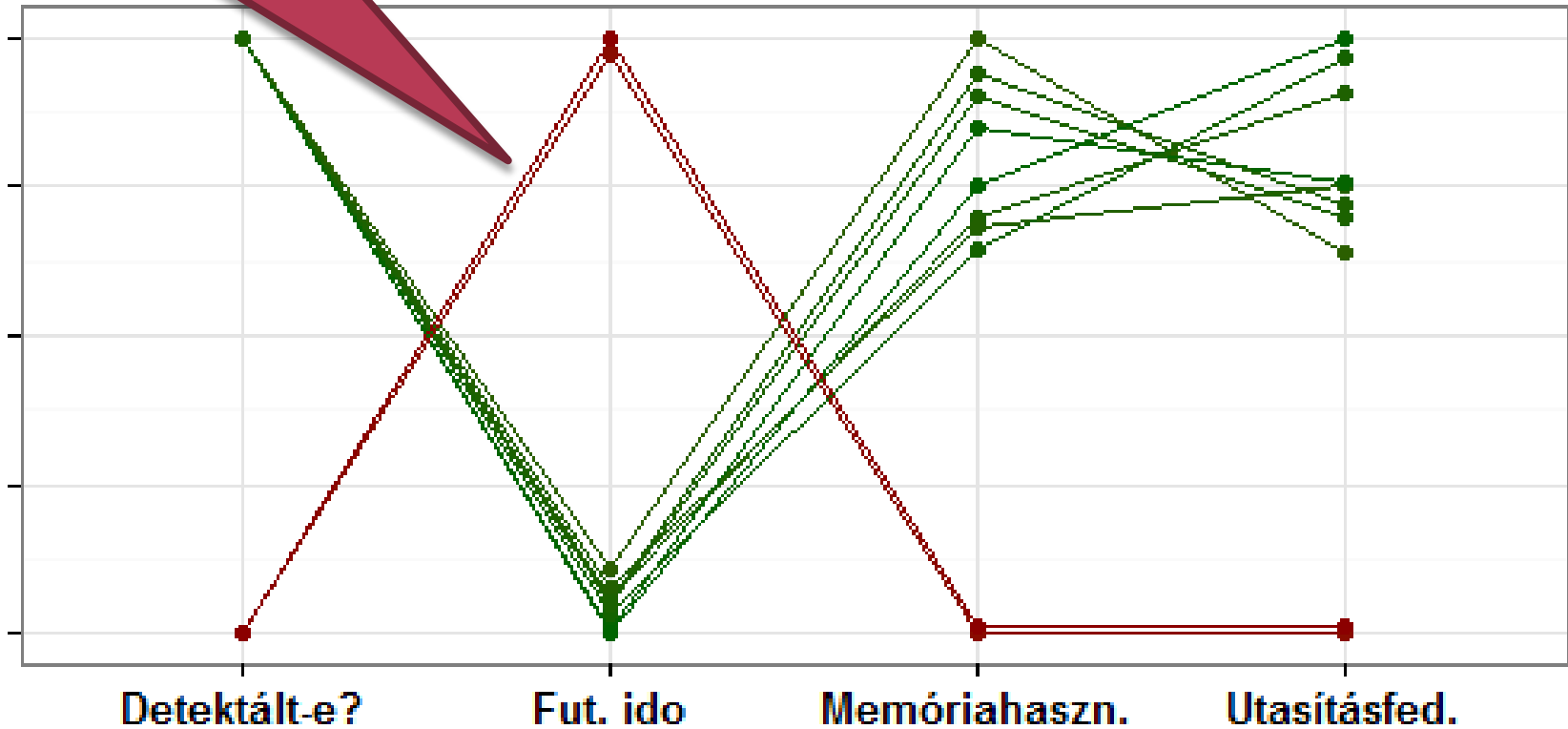


Treemap: állományrendszer



Párhuzamos koordináták: tesztesetek elemzése

1 teszteset 1 törött vonal

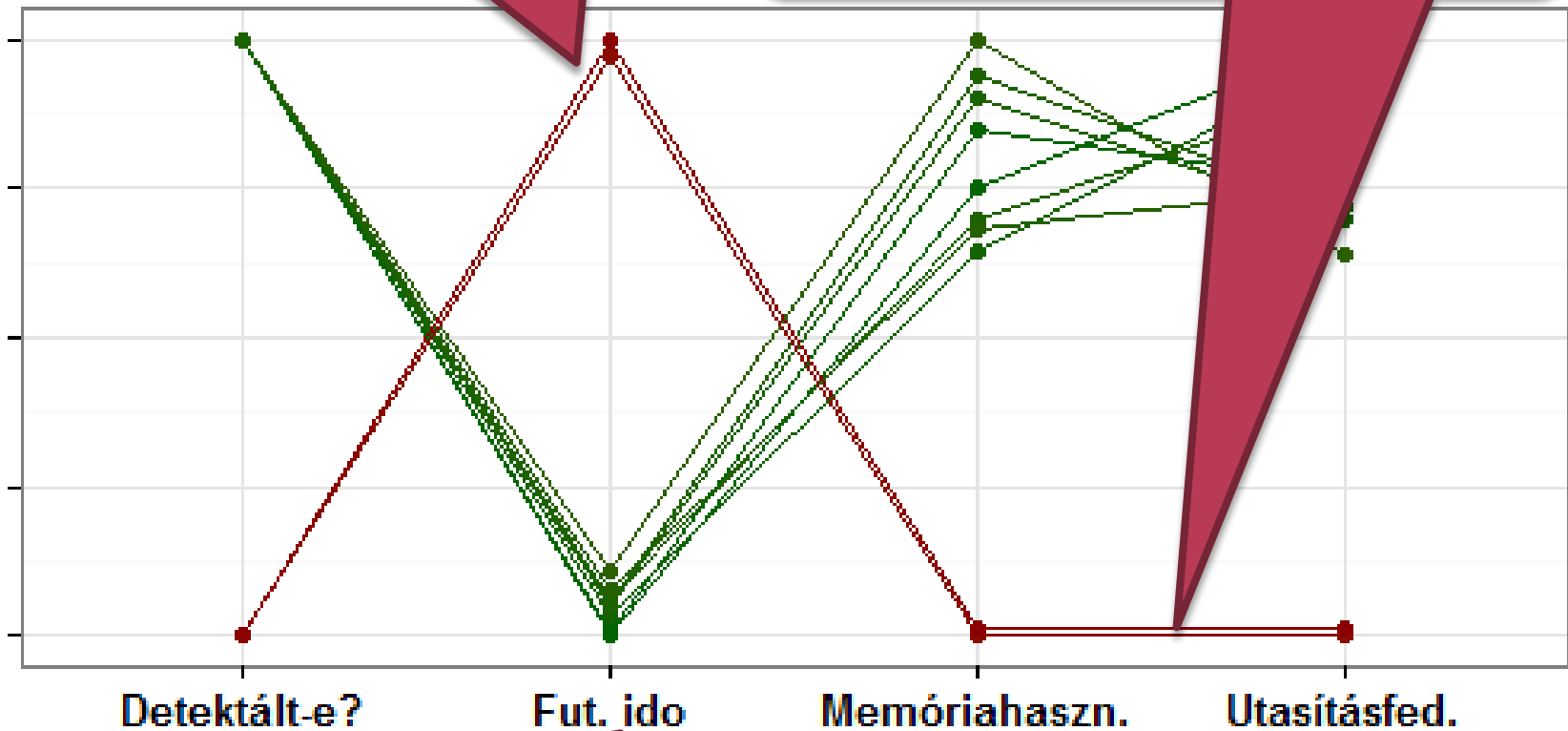


A változók az x tengelyen jelennek meg

Párhuzamos koordináták: tesztesetek elemzése

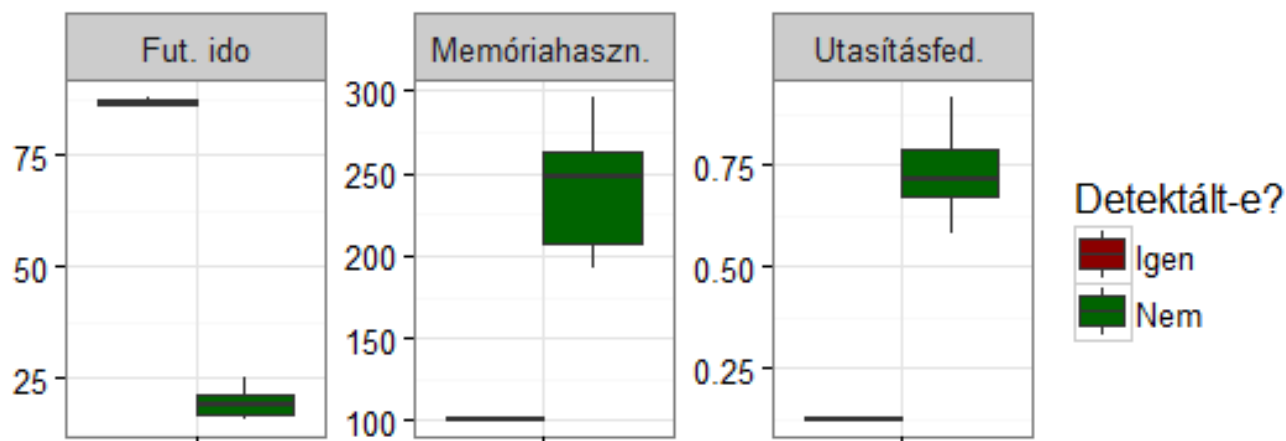
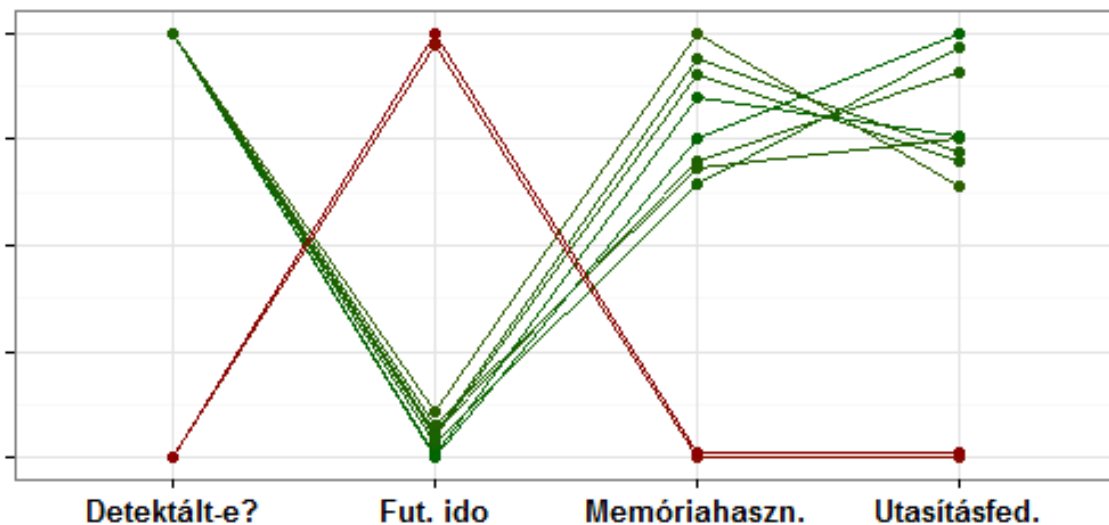
Timeout?

A hibát detektálók az érdemi számításig valószínűleg el sem jutnak



A futási idő és a memóriahasználat valószínűleg pozitív kapcsolatban állnak

Párhuzamos koordináták: viz. alternatívák



Benchmarking - Definíció

■ Wikipedia

*„In **computing**, a benchmark is the **act of running** a computer program, a set of programs, or other operations, in order to **assess the relative performance** of an object, normally by running a number of **standard tests** and trials against it.”*

A benchmarkolás

- egy **program** (programok, vagy más műveletek) **futtatása**
- **szabványos tesztekkel** vagy bemenetekkel
- egy objektum **relatív teljesítményének** felmérése érdekében

Benchmarking motiváció

- Fejlesztés alatt lévő rendszer
- Teljesítmény felmérése
 - Modell alapján becsült értékek
 - Megvalósított rendszer értékei „éles” helyzetben
- Tervezői és menedzsment döntések
 - Melyik részébe fektessünk több energiát ($\approx$ pénzt)?
- Mi alapján döntsünk?
 - Jelenlegi erősségek és gyengeségek
- Mi számít erősségnek vagy gyengeségnek?
 - Mire képesek a hasonló rendszerek, versenytársak?

Elvárások

- Ismételhetőség (Repeatability)
 - Egy elemen többször megismételt mérés/művelet eredményeinek változékonysága (szóródási faktor)
- Reprodukálhatóság (Reproducibility)
 - A mérési rendszer változékonysága, amelyet a műveletek viselkedéseinek különbsége okoz
- Általánosított felhasználói eset
 - Átlag felhasználó számára értelmezhető legyen az eredmény

- OLTP benchmark
 - Online Transaction Processing
 - Szoftverek és hardverek OLTP teljesítményét méri
- Komplex benchmark
 - Több tranzakció típus
 - Összetett adatbázis séma
 - Széles skálán mozgó adathalmaz méret
- Mért metrikák
 - Áteresztőképesség: transactions per minute (*tpmC*)
 - Ár-teljesítmény arány ($\approx$ hatékonyság): \$/tpmC

■ Mintaadatbázis

- Nagykereskedelmi beszállító cég rendelései
- 9 különböző tábla

■ 5 féle tranzakció típus

- Rendelés, fizetés, szállítás, rendelés státusza, raktár állapota
- Frissítés, beszúrás, törlés, megszakítás

■ ACID tranzakciók

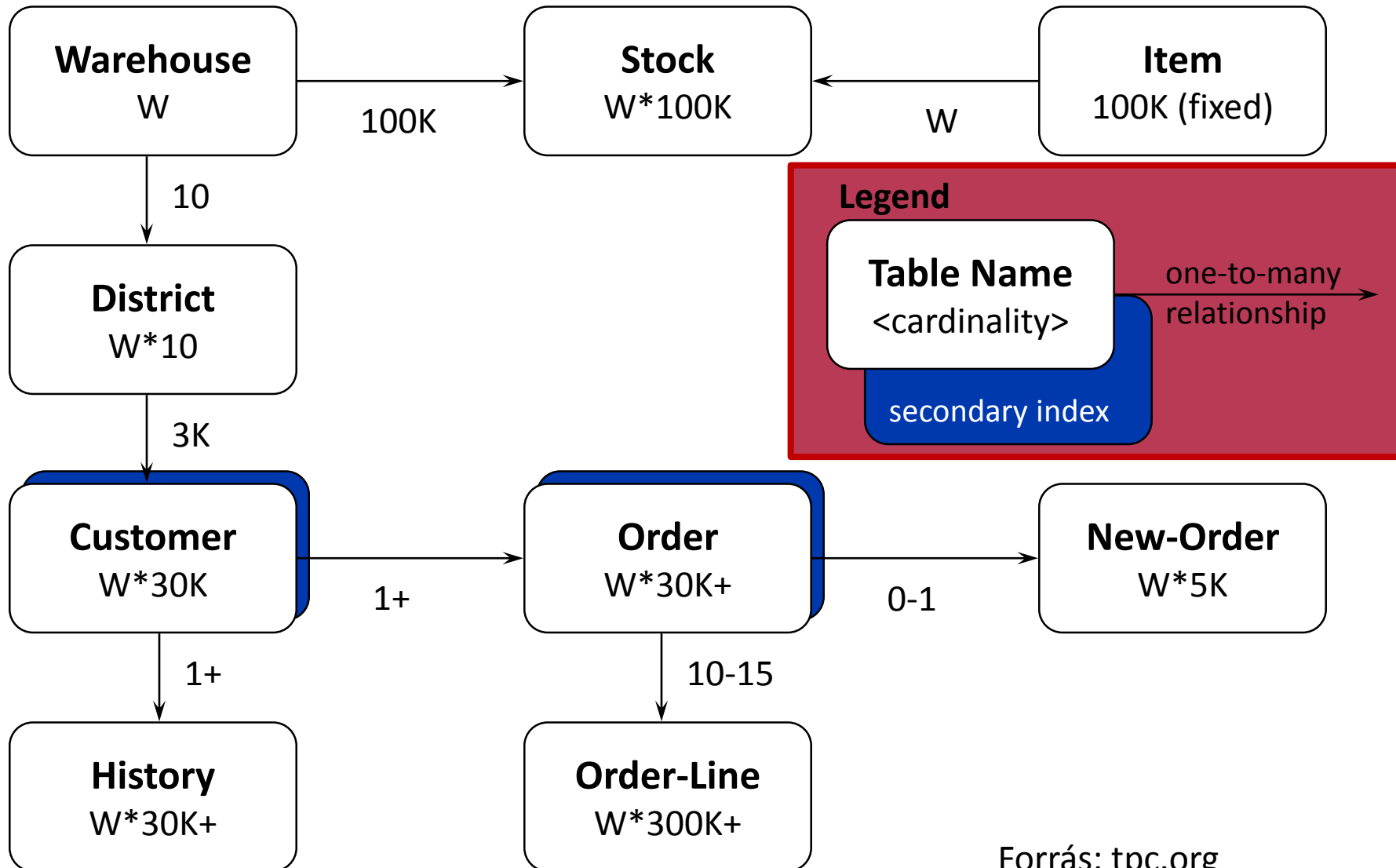
- Atomicity, Consistency, Isolation, Durability

■ Változatos felhasználói kérések szimulációja

- Tranzakció típusok kezdeményezése adott valószínűséggel

■ Időkorlát az egyes tranzakció típusokra

TPC-C adatbázis séma



Forrás: tpc.org

KIÉRTÉKELÉS

Adattisztítás

- Pénznem egységesítés (EUR, USD, AUD, JPY)
- Tizedesjegy, tízes elválasztó ellenőrzés

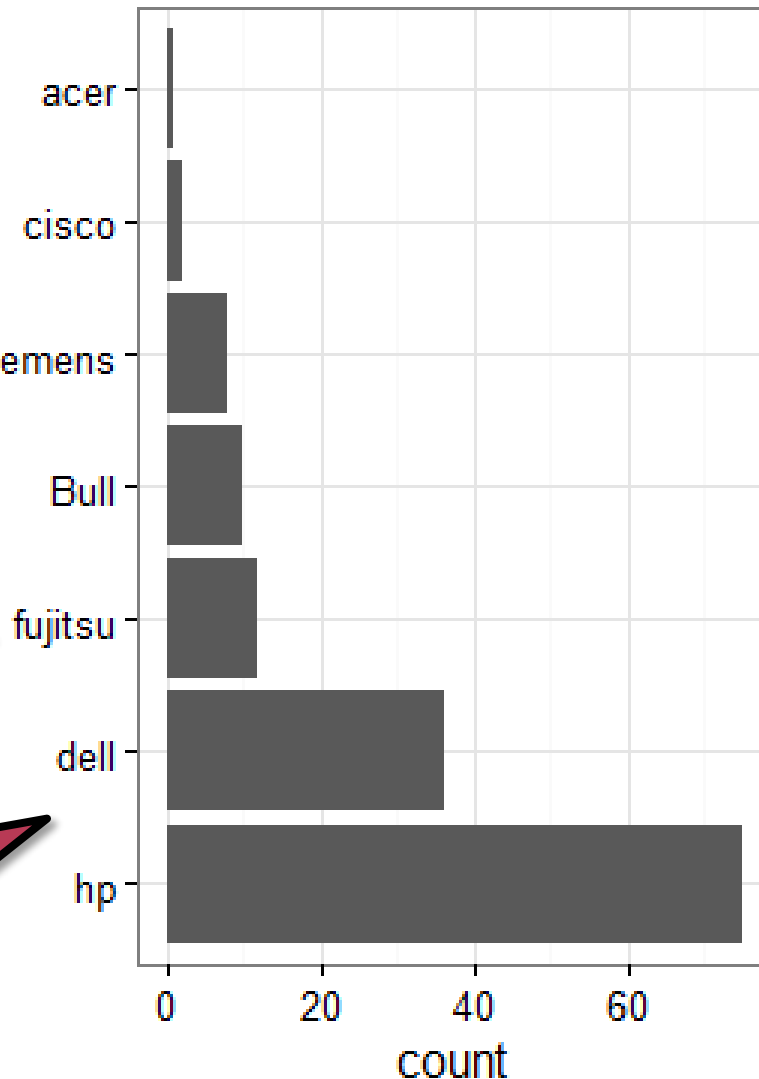
Hardware Vendor	System	tpmC	Price/tpmC	Watts/KtpmC	System Availability	Database	Operating System	TP Monitor	Date Submitted
	Altos R710 12GB/3.6GHz/2P C/S with 4 Altos R710	66,543	12.42 AUD	NR	12/01/05	Microsoft SQL Server 2000 Enterprise Edition SP3	Microsoft Windows Server 2003 Enterprise Edition	Microsoft COM+	01/12/06
	Bull Escala PL6460R	6,085,166	2.81 USD	NR	12/15/08	IBM DB2 9.5	IBM AIX 5L V5.3	Microsoft COM+	06/15/08
	Bull Escala PL860R	629,159	2.49 USD	NR	06/11/08	IBM DB2 9.5 Enterprise Edition	IBM AIX 5L V5.3	Microsoft COM+	06/11/08

Milyen rendszereket vizsgálunk?

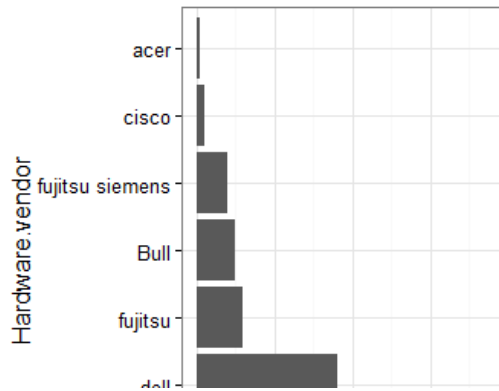
Egységesítsünk?

Hardware vendor

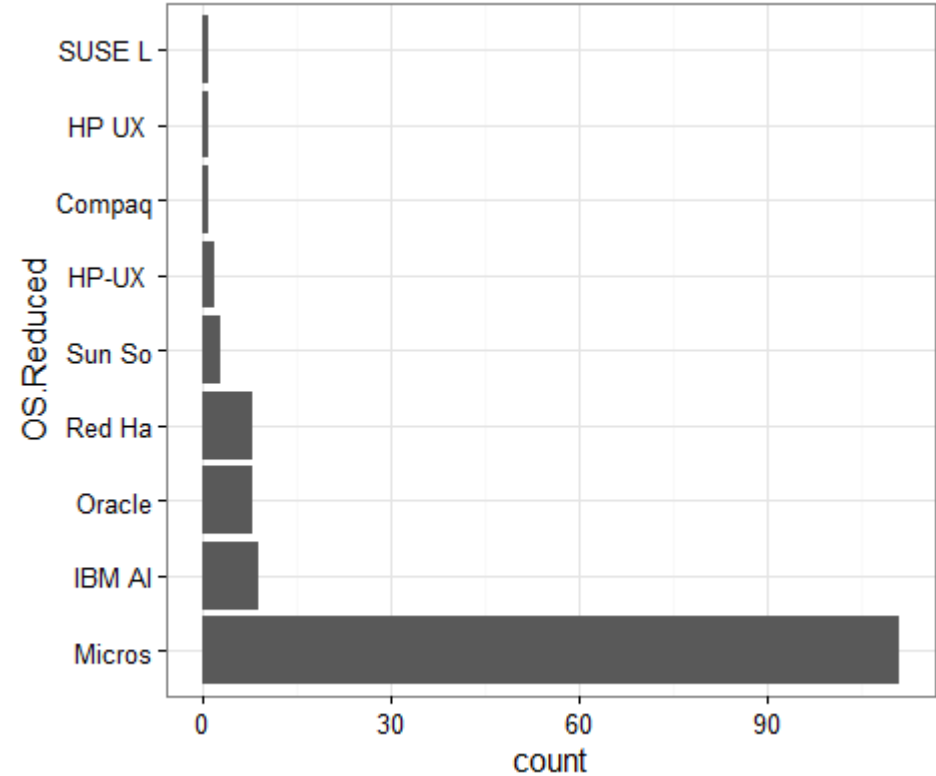
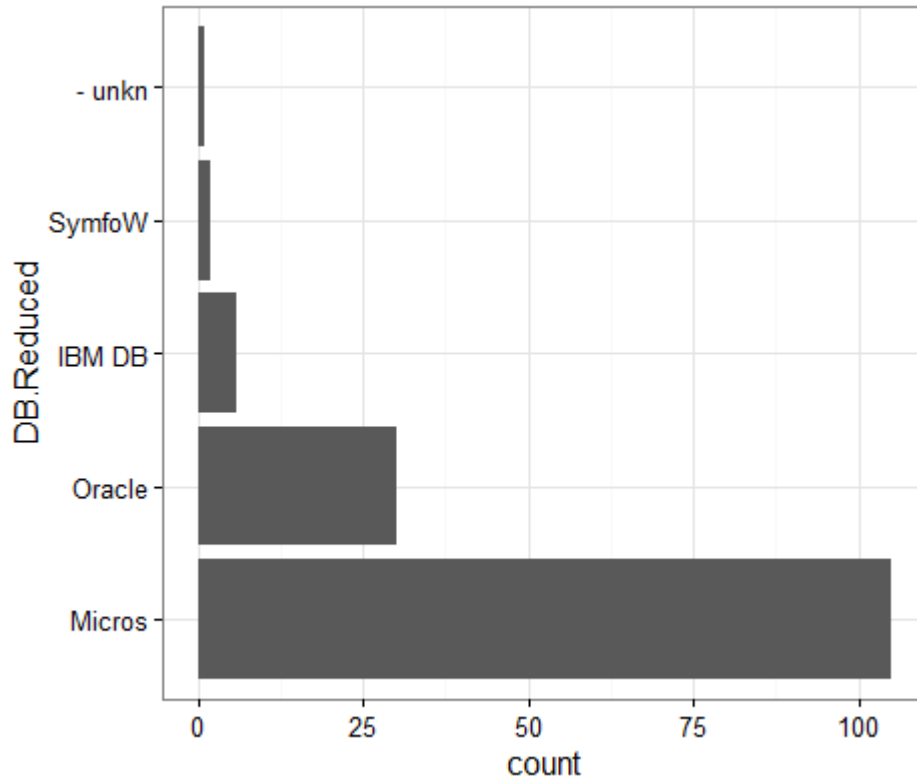
A két leggyakoribb gyártó
az esetek 77%-ában
szerepel



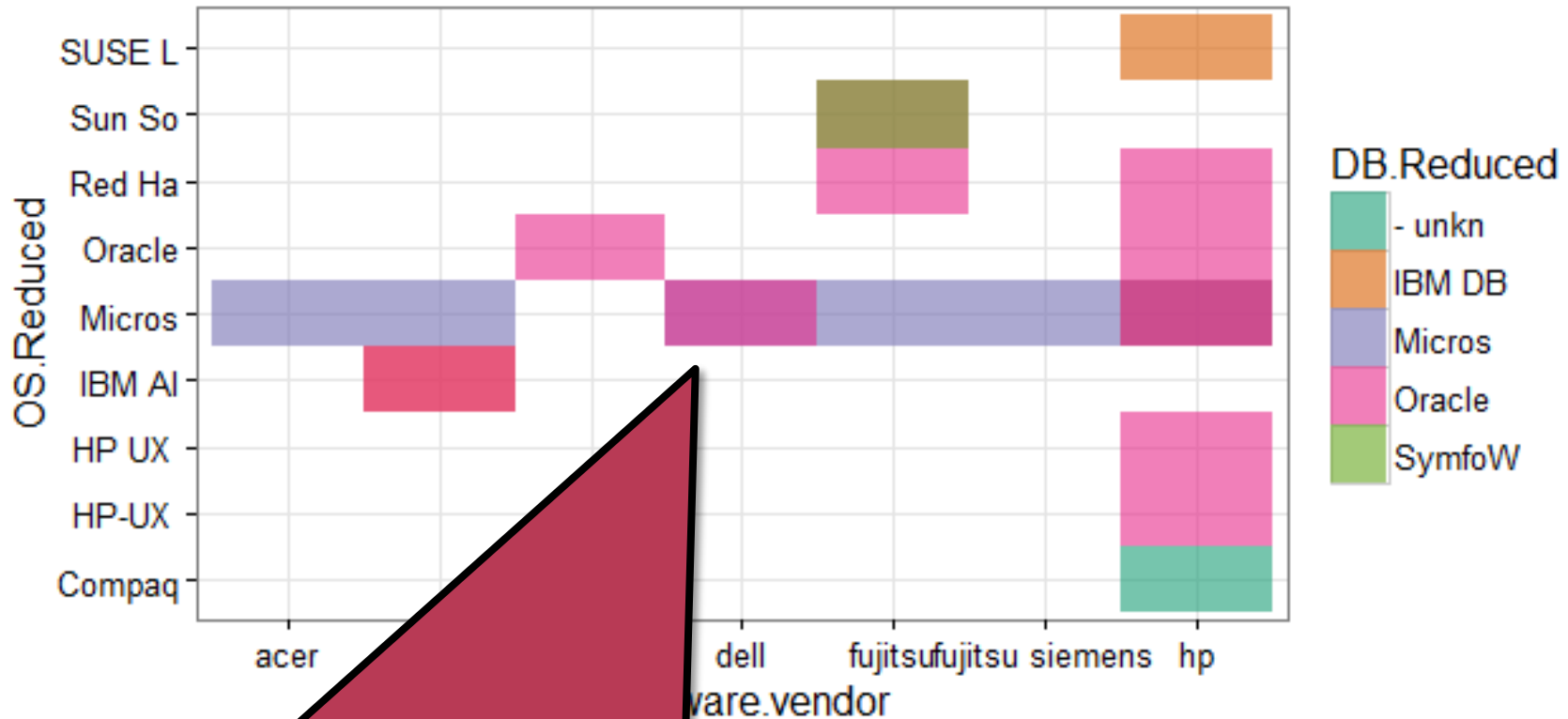
Milyen rendszereket vizsgálunk?



Típus szerinti
összevonás?

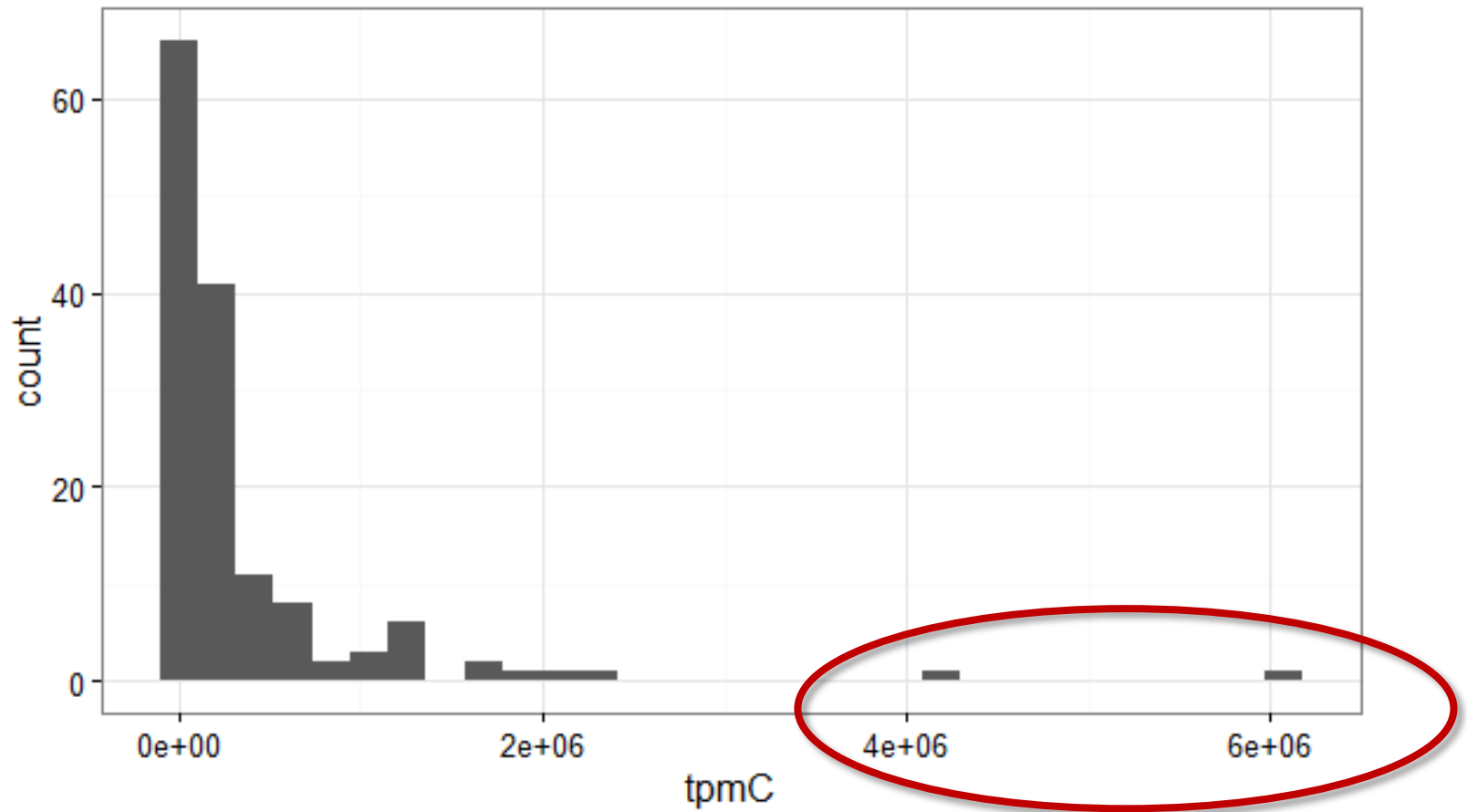


Milyen rendszereket vizsgálunk?



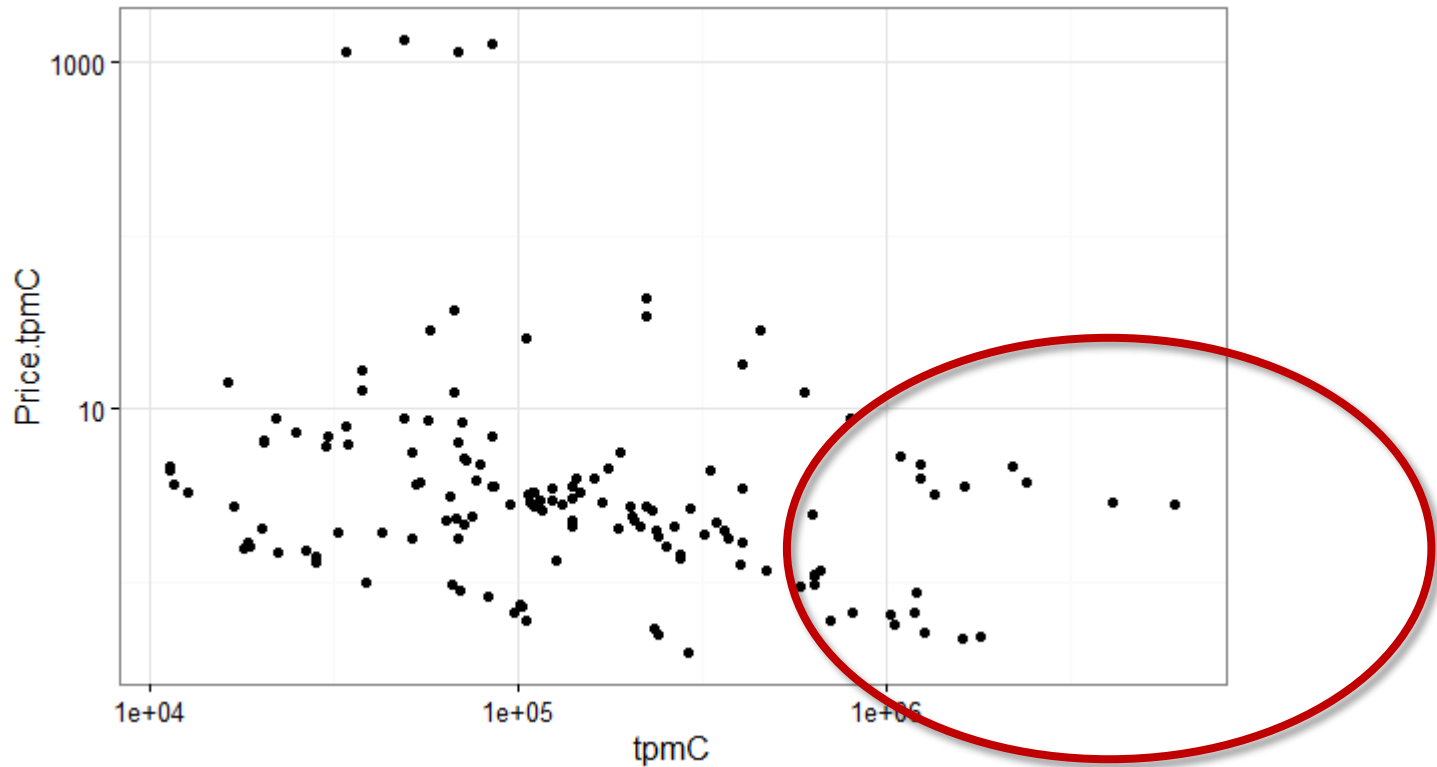
Általában többféle adatbázis többféle operációs rendszeren fut

Benchmark eredmények



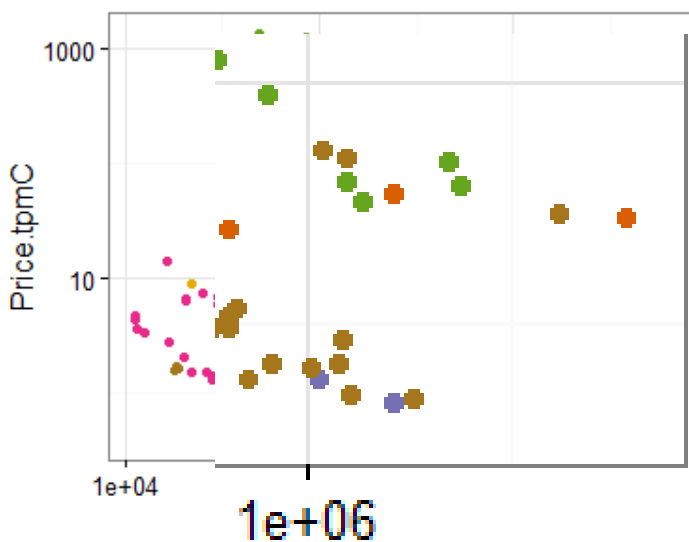
Milyen az ár/érték arányuk?

Benchmark eredmények



Logaritmikus skála?

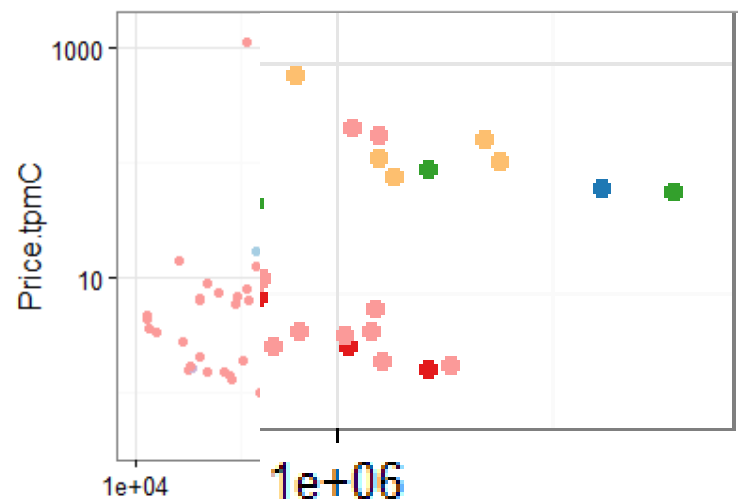
Benchmark eredmények



Hardware.vendor

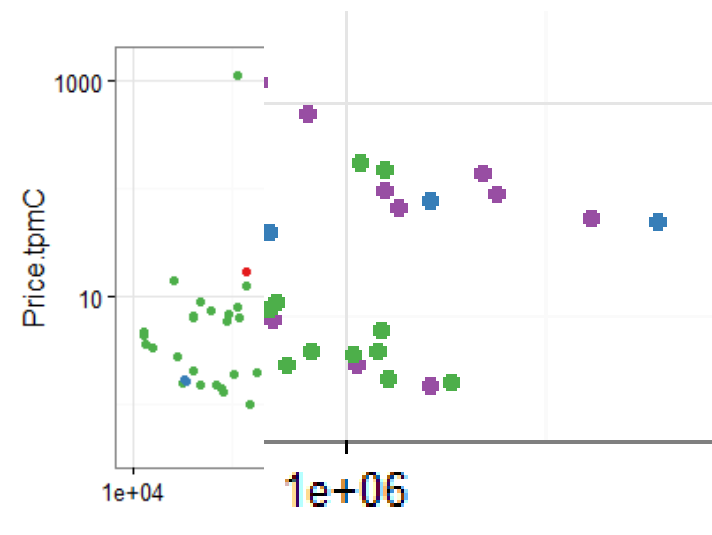
- acer
- Bull
- cisco
- dell
- fujitsu
- fujitsu siemens
- hp

Az élmezőny elég változatos



OS.Reduced

- Compaq
- HP-UX
- HP UX
- IBM AI
- Micros
- Oracle
- Red Ha
- Sun So
- SUSE L



DB.Reduced

- unkn
- IBM DB
- Micros
- Oracle
- SymfoW

Nincs legjobb OS és DB konfiguráció sem

Időbeli változások?

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

