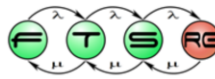


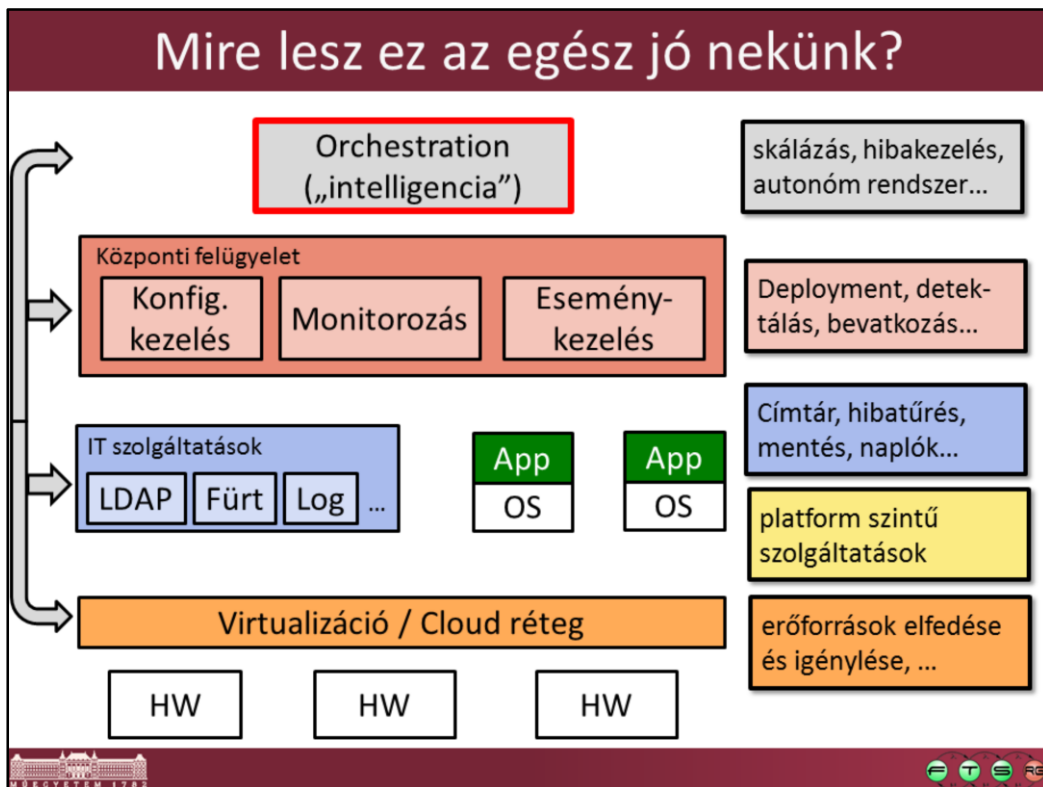
IT adatok vizuális elemzése

Salánki Ágnes, Kocsis Imre



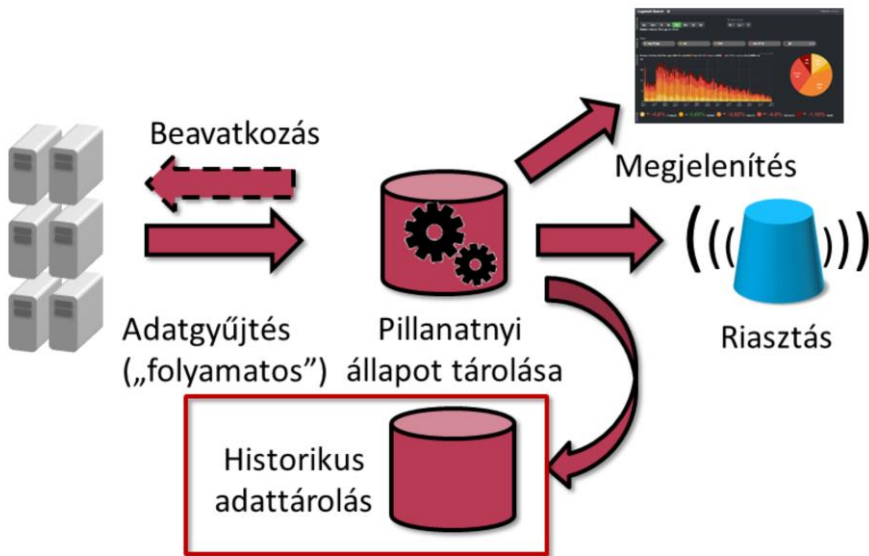
Utolsó módosítás: 2016.05.11.

Dr. Pataricza András (Rendszermodellezés) és Kocsis Imre (Big Data elemzési módszerek) idevágó fóliáit felhasználva.



Az alsó három szinttel foglalkoztak már a korábbi előadások. A vizuális analízis a monitorozott adatokat használja bemenetként, a rendszerbe általában a Központi felügyelet blokkba kapcsolódik vissza.

Mire lesz ez az egész jó nekünk?



A historikus adattárolás egyrészt szép dolog, másrészt meg: mit kezdünk azokkal az adatokkal, amiket eltárolunk? Szépen kielemezzük! 😊

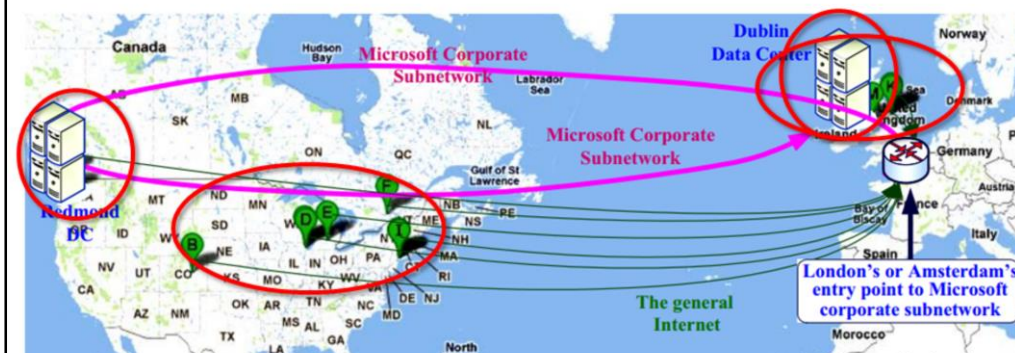
Mit kezdünk az adatokkal?





- Kapacitásstervezés, teljesítménymenedzsment → nem elég tudni a pillanatnyi értékeket feldolgozni, a hosszú távú tervezéshez pontosan kell tudni, melyik erőforrásból mekkora tartalékunk van
- Monitorozási szabályok → az adatelemzés segíthet annak kitalálásában, pontosan mely metrikákat érdemes kivezetni pl. real-time, vagy melyeket hogyan kell aggregálni
- Rendelkezésre állás növelés → a szűk keresztmetszetek néha nem mennek fejből
- Kísérlettervezés → meddig kell kiterhelni a rendszert, hogy minden működési módot lássunk?

Esettanulmány: cloud benchmarking

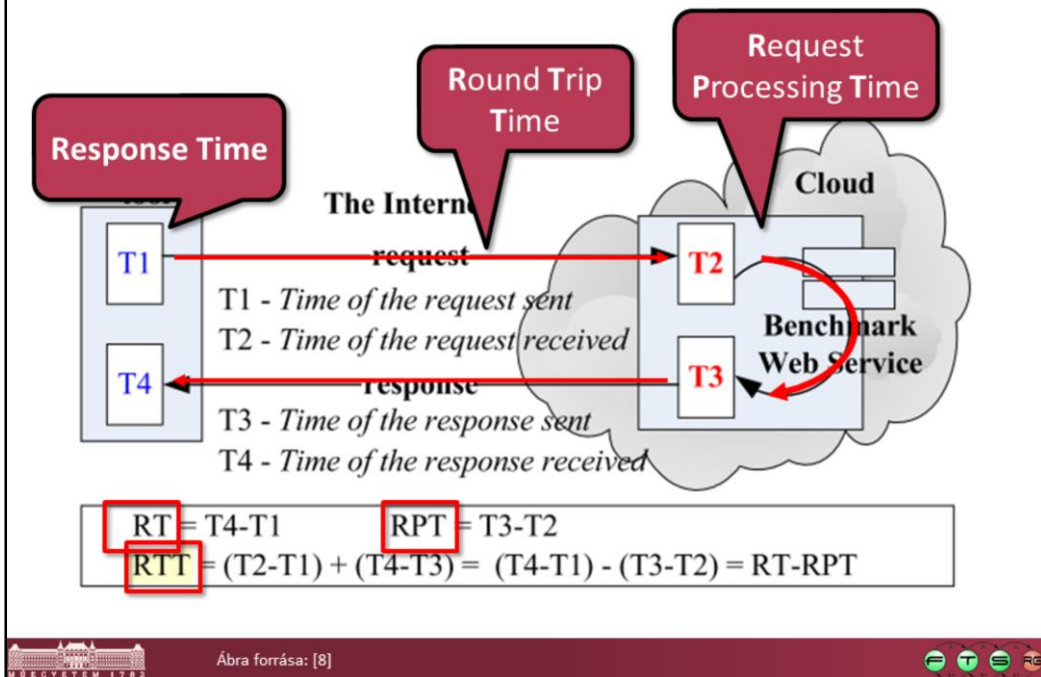


Ábra forrása: [6], [7]

Kísérlet eredeti célja: hogyan befolyásolják a kliens/szerver implementációk/helyszínek egy MS Azure-ban futó webszolgáltatás teljesítményét? Felhasználói szemszög: a QoS metrikánk ezúttal a válaszidő.

Az eredeti esettanulmányát lásd a [6] cikkben.

Esettanulmány: cloud benchmarking



WSsDAT – percenként kiküld egy ilyen kérést a megfelelő szerverhez

A vizsgált webszolgáltatás: kérésre egy 50x50x50-es mátrixot rendezzünk ellenkező irányba, majd ennek az első 100KByte-nyi adatával térjünk vissza

Esettanulmány: cloud benchmarking

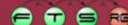
**44 608 rekord
manuálisan?**

analysis.R x d x										
Filter										
	ip	RT	RPT	RTT	location	client.type	DC	start.time	Country	
44574	New York 1	1428	641	787	New York	Java client	Redmond DC	85219.87	USA	
44575	New York 1	1269	625	644	New York	Java client	Redmond DC	85281.71	USA	
44576	New York 1	992	641	351	New York	Java client	Redmond DC	85343.76	USA	
44577	New York 1	1076	718	358	New York	Java client	Redmond DC	85405.53	USA	
44578	New York 1	985	719	266	New York	Java client	Redmond DC	85467.36	USA	
44579	New York 1	996	734	262	New York	Java client	Redmond DC	85529.27	USA	
44580	New York 1	900	640	260	New York	Java client	Redmond DC	85591.19	USA	
44581	New York 1	1917	640	1277	New York	Java client	Redmond DC	85653.09	USA	
44582	New York 1	2328	625	1703	New York	Java client	Redmond DC	85715.09	USA	

Lehetséges elemzési megközelítések

1. Leíró statisztika

Client's location		Client's IP address	Response Time (RT), ms				
			Min	Avg	Max	Std Dev	CV, %
Canada	Ottawa	216.151.172.x	1484	2007	3917	255	12.7
	Burnaby	64.151.226.x	1841	2329	287448	7648	328.4
UK	Newcastle	10.8.146.x	859	1577	4016	480	30.4
	Newcastle	10.8.151.x	890	1498	3438	423	28.2
	Durham	213.175.197.x	814	1922	47740	2132	110.9
USA	New York	69.72.183.x	1331	2018	7145	575	28.5
	Chicago	209.188.85.x	1133	2104	11729	875	41.6
	Peyton	64.64.0.x	1289	1961	4637	350	17.9
	Lansing	67.225.254.x	902	2051	4483	396	19.3
		67.225.254.x	1421	2024	3992	365	18.0
		67.227.193.x	1441	2068	19365	1004	48.5
		67.227.216.x	6123	6911	11520	366	5.3
		67.227.216.x	1420	2016	3770	344	17.1
	Secaucus	204.14.93.x	1262	1983	47056	1499	75.6
		208.87.24.x	1216	1991	46457	1271	63.8
		208.87.25.x	1212	2030	59475	2232	110.0
		64.20.37.x	1216	2275	132498	5135	225.7



CV -- coefficient of variation: szórás / átlag. Miért fontos: mert pl. a 2 ms-es szórás más jelent 4 ms-os átlag válaszdőnél és egy 4 s-os átlagos válaszdőnél.

Lehetséges elemzési megközelítések

1. *Leíró statisztika*

2. *Exploratory Data Analysis (EDA)*

- statisztikai tradíció,
- mely koncepcionális
- és számítási eszközökkel segíti
- minták felismerését és ezen keresztül
- hipotézisek felállítását és finomítását.

[1] és [2] alapján



- Statisztikai tradíció: ugyanolyan mint a matematikai statisztikából leadott Confirmatory Analysis (hipotézistesztesztelés, modellválasztás, paraméterillesztés – amit valószínűségszámításból tanultunk).
- Koncepcionális és számítási eszközök: nem használunk mély statisztikát, elemi számításokat és ábrákat annál inkább + az interpretáció sikeressége nagyban függ a szakértői tudástól.
- Minták felismerését és hipotézisek felállítását segíti: valahol tehát az adatbányászat és a statisztika között van.

Exploratory Data Analysis

- Cél: adatok „megértése”
 - „detektív munka”
 - erősen ad-hoc
- Fő eszköz: adatok „bejárása” grafikus reprezentációkkal
- Hipotézisteszteléssel: iteratív folyamat



Erősen ad-hoc: az adatok szisztematikus átvizsgálása, de nincs jól bevált recept, leginkább mélységi keresést végzünk, ha találunk valami érdekeset, akkor aztán megfogalmazunk egy hipotézist („Statisztikailag szignifikáns-e a válaszdő várható értékének különbsége Java és .Net kliens esetén?” vagy „Normál eloszlást követ-e az RPT?”), aztán a megfelelő teszttel leellenőrizzük. Valamit mond a teszt, aztán görgetünk vissza és folytatjuk a felderítést.

Mi nem kell hozzá? Mély statisztikai ismeret: se centrális határeloszlás tétel, se nagy számok törvénye, az majd a 2. fázisban

Mi kell hozzá? Szakértői tudás nélkül nem megy – ezért beszélünk róla egyáltalán IRF-ből.

Miről lesz szó?

- Adatelemzési alapfogalmak
- Alapvető diagramtípusok
- Interaktív EDA eszközök – elvárt funkcionalitás
- Esettanulmány: cloud benchmarking

Miről nem lesz szó?

- Adatbányászat
- Hipotézistesztesztelés
- Kísérlettervezés
 - Pl. Rendszermodellezés tárgyunk
- Számítógépes grafika
- Információvizualizáció
 - Pl. blogok: Junk charts [8], Flowing data [9]



Prezentáció vs. felderítés

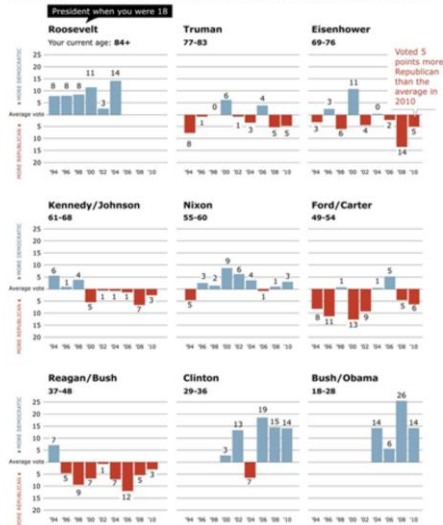
- **Prezentáció**
 - Statikus
 - Jó minőségű
 - Tömör
 - Sok annotáció: nagy közönség
- **ggplot2 csomag (R)**
Adobe Illustrator, Inkscape
- **Felderítő ábrázolás**
 - Interaktív
 - Gyors
 - Több különálló ábrát kapcsol össze
 - Néha tengelyfeliratok sem: az elemző az interpreter
- **Pl. Mondrian, iplots (R)**
Many Eyes, Tableau



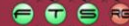
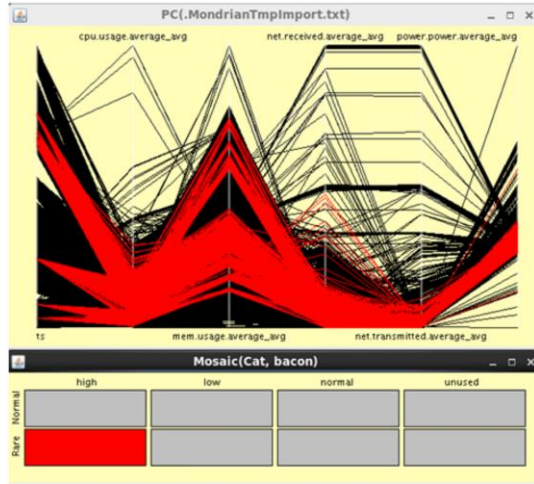
Prezentáció vs. felderítés

Presidential Legacies: How Those Who Came of Age Under Different Presidents Have Voted

How much more Democratic or Republican each group voted relative to the national average in each election



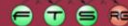
Source: Pew Research Center, *The Generation Gap and the 2012 Election* (p. 16)
Based on likely voters for Pew Research Center surveys conducted in the fall of each election year. Presidential vote for 1996, 2000, 2004, and 2008. Generic House vote for 1994, 1998, 2002, 2006, and 2010.



Adatmennyiség?

Hipotézismentes adatgyűjtés (mindent eltárolunk)

- Egyszerű instrumentáció (pl. HF3)
 - 8 metrika, egyperces mintavételezéssel egy hónapig
 - $8 \times 60 \times 24 \times 30 \approx 350\,000$ adatpont
- Windows 7 OS: perfmon
 - ≈ 100 körüli metrikaszám, egy másodperces mintavétel
 - Egy nap alatt $100 \times 3600 \times 24 \approx 860\,000$
- Tanszéki VCL
 - ≈ 70 metrika, hosztonként 20 mp-es mintavételezéssel
 - Egy hónap alatt $70 \times 10 \times 180 \times 24 \times 30 \approx 90M$



Hipotézismentes adatgyűjtés: mindent eltárolunk, amit tudunk (az adatelemzés kezdetén pl. fogalmunk sincs, mit nézzünk)

Mi a baj ekkora mennyiségnél? A táblázatokat

1. vagy nagyon sokáig kell böngészni, vagy
2. az aggregálás a világon mindent kisimít ekkora távon, a trend éppen látszik, a tranziensek pont nem és ez baj, mert minket általában az érdekel, hol és mikor történt valami hiba.

Miről lesz szó?

- **Adatelemzési alapfogalmak**
- Alapvető diagramtípusok
- Interaktív EDA eszközök – elvárt funkcionalitás
- Esettanulmány: cloud benchmarking



Rekordok és változók

Változók/Attribútumok

- Start.time
- Country
- Location
- IP
- Client.type
- DC
- RT, RPT, RTT

Rekord/megfigyelés

	ip	RT	RPT	RTT	location	client.type	DC	start.time	Country
44574	New York 1	1428	641	787	New York	Java client	Redmond DC	85219.87	USA
44575	New York 1	1269	625	644	New York	Java client	Redmond DC	85281.71	USA
44576	New York 1	992	641	351	New York	Java client	Redmond DC	85343.76	USA
44577	New York 1	1076	718	358	New York	Java client	Redmond DC	85405.53	USA
44578	New York 1	985	719	266	New York	Java client	Redmond DC	85467.36	USA
44579	New York 1	996	734	262	New York	Java client	Redmond DC	85529.27	USA
44580	New York 1	900	640	260	New York	Java client	Redmond DC	85591.19	USA
44581	New York 1	1917	640	1277	New York	Java client	Redmond DC	85653.09	USA
44582	New York 1	2328	625	1703	New York	Java client	Redmond DC	85715.09	USA

Rekord: általában egy egyértelműen azonosítható mérési regisztrátum

Változó: minden, amit az adott pillanatban tudunk a mérésről és a mért adatról

Hogyan érdemes az adatokat szervezni?

Tidy data

- Minden változó egy-egy oszlop
- Minden megfigyelés egy-egy sor
- Minden megfigyelési egység egy-egy tábla

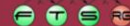
	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

„Messy” data

name	trt	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Wickham, Hadley . "Tidy Data" *Journal of Statistical Software*, Volume 59, Issue 10 (2014)

URL: <https://www.jstatsoft.org/article/view/v059i10>



Miért fontos ezt tudni? Mert érdemes már rögtön így naplózni az adatokat, hogy segítse a későbbi kiértékelést.

Változók: kontextus és viselkedési

- Kontextus
 - a mérési konfigurációt jellemzi
- Viselkedési
 - maga a mért érték



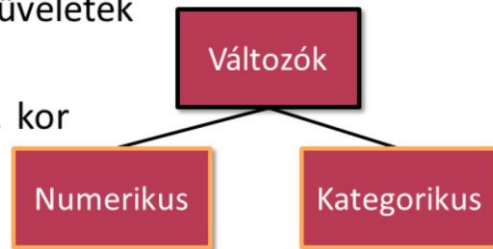
Mindig a kísérletből derül ki, hogy egy adott változó az kontextus vagy viselkedési.
Például:

- a) Hőmérsékletet mérünk MO-n: a hőmérséklet viselkedési
- b) Adott hőmérsékletre felmelegítjük a szerver szobát és számoljuk, hány gép adja meg magát: a hőmérséklet kontextus

Numerikus és kategorikus változók

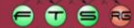
■ Numerikus (numerical)

- az alapvető aritmetikai műveletek értelmesek
- Pl. napi átlaghőmérséklet, kor



■ Kategorikus (categorical)

- Csak a megkülönböztetés miatt
- Pl. telefonszám, nem



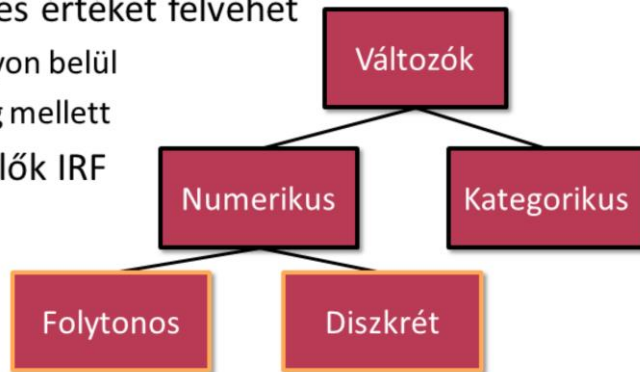
Nem a reprezentáció a lényeg, hanem az interpretáció!

Ha a nem az $\{1, 2\}$ halmazból veszi fel az értékét (és nem a $\{\text{nő}, \text{férfi}\}$ -ből), akkor is kategorikus lesz, mert minden művelet értelmetlen rá.

Numerikus változók

■ Folytonos

- Mért – tetszőleges értéket felvehet
 - adott tartományon belül
 - adott pontosság mellett
- Pl. a teremben ülők IRF jegyének átlaga



■ Diszkrét

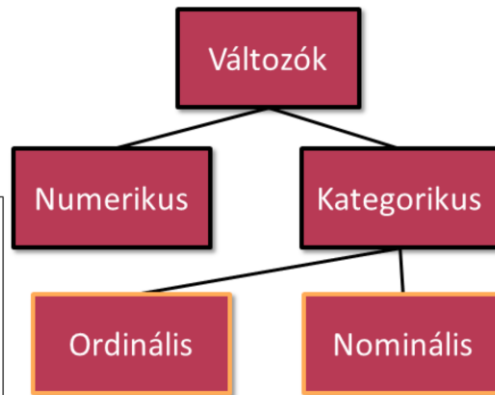
- Számolt – véges sok értéket vehet fel adott tartományban
- Pl. IRF előadáson ülők száma

Kategorikus változók

■ Nominális (nem értelmezhető a rendezés)

■ Ordinális (rendezett)

○ szintek között hierarchia



9. Ajánlanád-e a tárgyat másoknak?

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



Miért fontos, hogy tudjuk a típust? Mert tudnunk kell, milyen ábrázolásmód passzol hozzá.

Esettanulmány: változók típusai

- Start.time – numerikus, folytonos
- Country – kategorikus, nominális
- Location – kategorikus, nominális
- IP – kategorikus, nominális
- Client.type – kategorikus, nominális
- DC – kategorikus, nominális
- RT, RPT, RTT – numerikus, folytonos



Ha tudjuk a megfelelő ábrázolásmódot, onnan akár vissza is következtethetünk a típusra: pl. az RT vs. Start.time inkább scatterplotra menjen és nem mozaikplotra → inkább legyen numerikus, mint kategorikus.
A folytonos-diszkrét általában értelmezés kérdése.

Miről lesz szó?

- Adatelemzési alapfogalmak
- **Alapvető diagramtípusok**
- Interaktív EDA eszközök – elvárt funkcionalitás
- Esettanulmány: cloud benchmarking

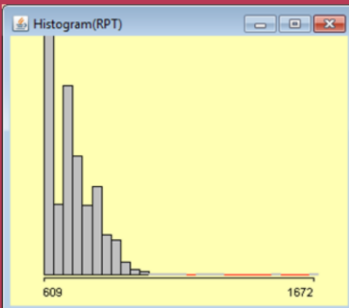


1 változó

Változók

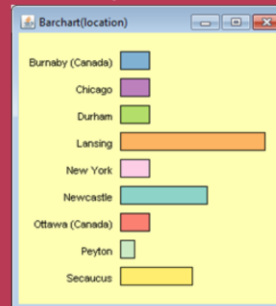
Numerikus

{RPT: 609, 613, 913, ...}



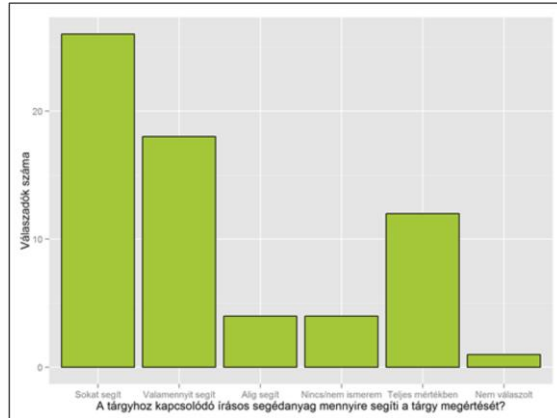
Kategorikus

{location: Peyton, Durham, ...}



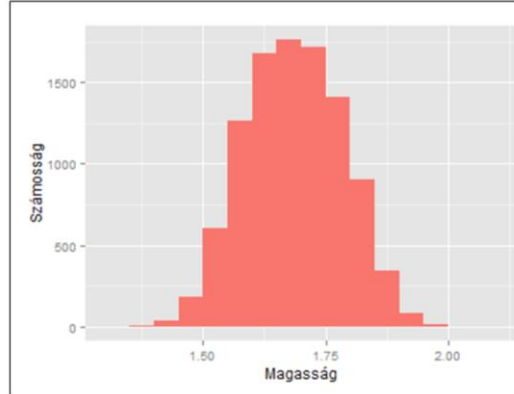
Oszlopdiagram (bar chart)

- Megjelenített dimenziók száma: 1
- Ábrázolt összefüggés:
 - Kategorikus változó egyes értékeinek abszolút gyakorisága
- Adategység:
 - Oszlop – magassága: adott érték gyakorisága
- Tervezői döntés:
 - Értékkészlet darabolása?



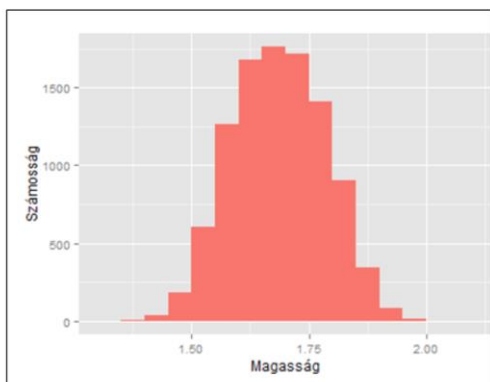
Hisztogram

- Megjelenített dimenziók száma: 1
- Értékkészlet intervallumokra bontása (bin)
- Ábrázolt összefüggés:
 - Folytonos változó intervallumainak számossága
- Adategység:
 - Oszlop – magassága: intervallum számossága
- Tervezői döntés:
 - Oszlopszélesség/kezdőpont?

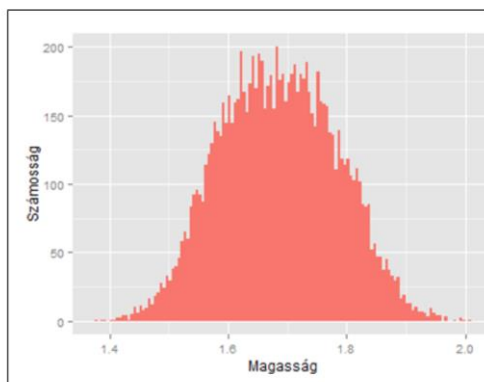


Intervallum mérete = oszlopszélesség (bin width)

Hisztogram: Oszlopszélesség hatása



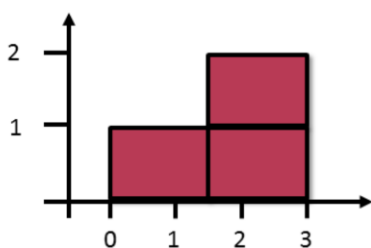
Oszlopszélesség: 0,1



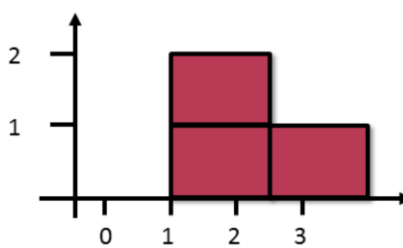
Oszlopszélesség: 0,005

Hisztogram: Kezdőpont hatása

- $\{1, 2, 3\}$ ábrázolása 1,5 oszlopszélességgel



Kezdőpont: 0



Kezdőpont: 1



$S := \{1.0, 2.0, 3.0\}$, az oszlopszélesség legyen 1.5.

Két különböző hisztogram is ér: $([0, 1.5), [1.5, 3.0])$ és $([1.0, 2.5), [2.5, 4.0]) \rightarrow$ ezek mindegyike érvényes, csak máshonnan indul!

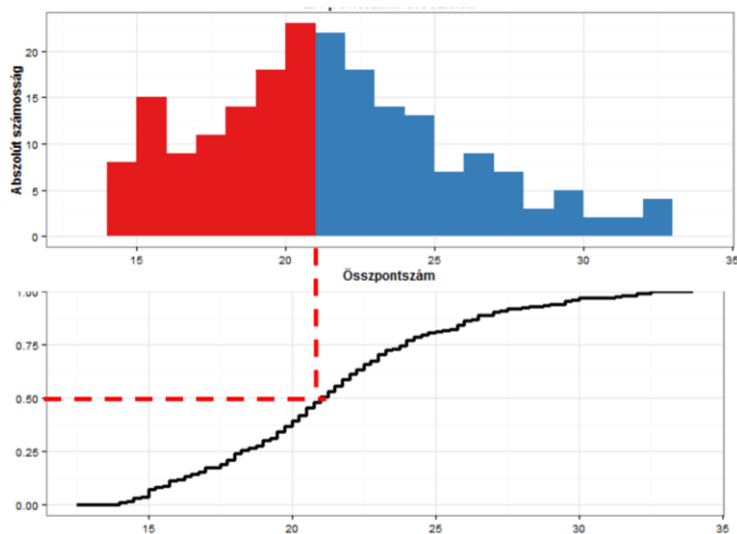
Hisztogramról leolvasható információ

- Néha elég tudni,
 - Hol van az adatok „közepe”?
 - Mennyire „szórtak” az adatok?
 - Vannak-e kilógó értékek?



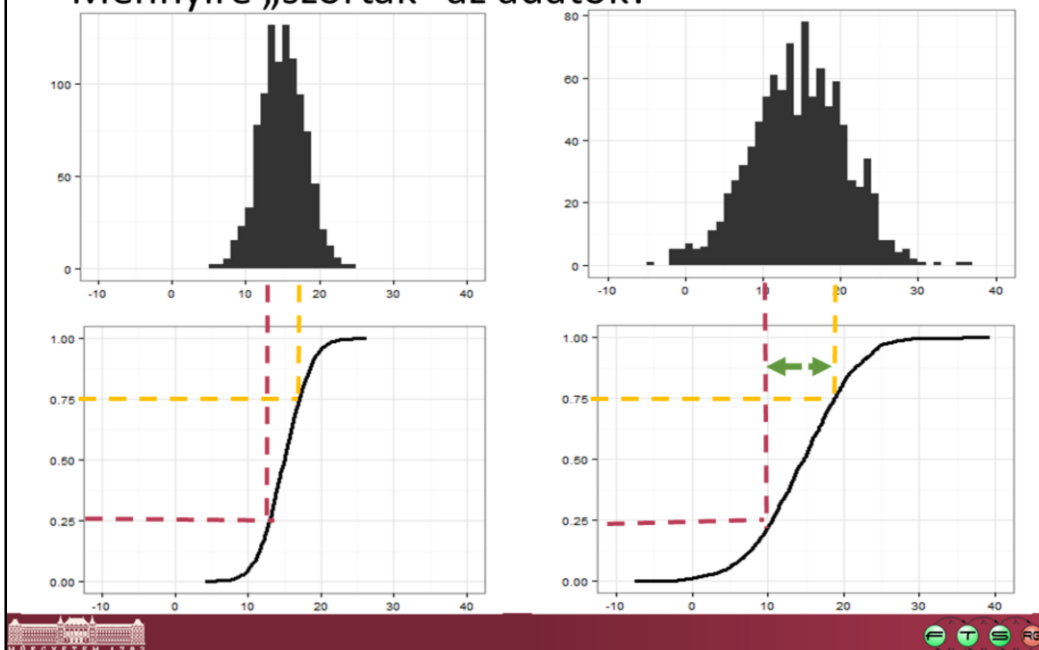
Egyszerű statisztikai jellemzés

- Hol van az adatok „közepe”?



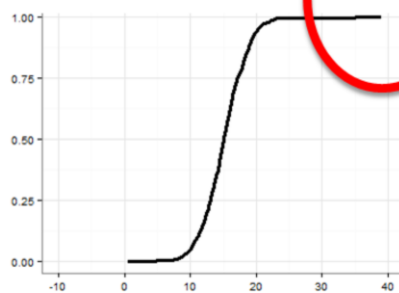
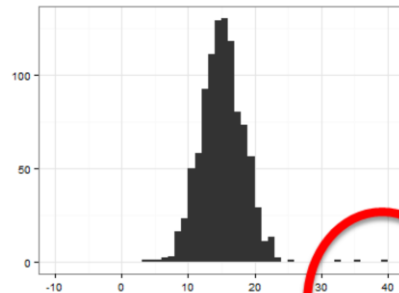
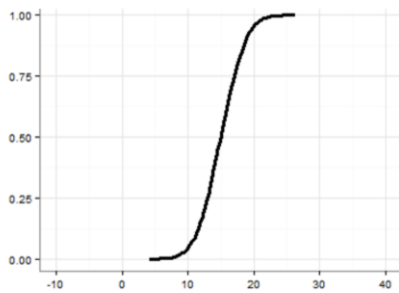
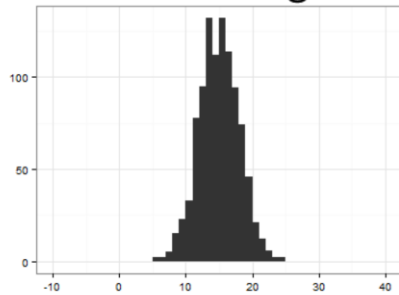
Egyszerű statisztikai jellemzés

■ Mennyire „szórtak” az adatok?



Egyszerű statisztikai jellemzés

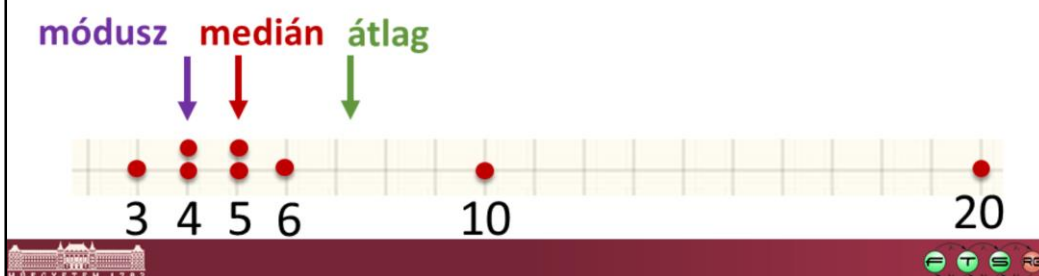
■ Vannak-e kilógók?



Kitérő: Egy kis leíró statisztika...

■ A „központ” jellemzése

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



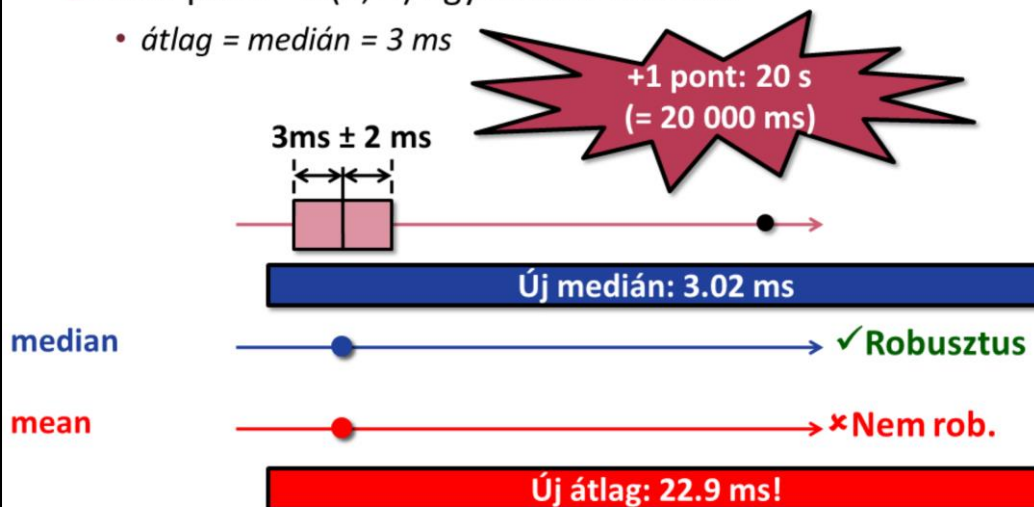
- Medián: Ha az értékeket növekvően sorba rendezzük, akkor a középső adat az adathalmaz mediánja. Ha nincs középső adat (páros számú érték esetén), akkor a medián a két középső érték átlaga (számtani közepe).
- Módusz: A módusz az adathalmazban leggyakrabban előforduló érték. Ez nem feltétlenül egyértelmű, ilyenkor több móduszból beszélünk.

Kitérő: Robusztus mérőszámok

■ Példa: alaphalmaz

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

- átlag = medián = 3 ms

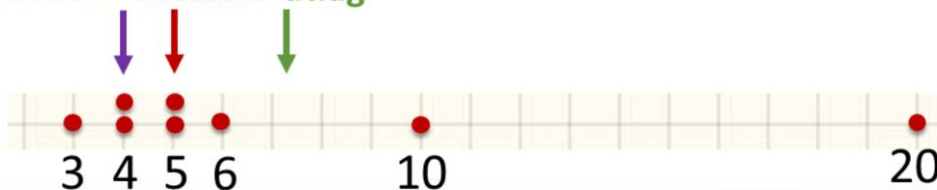


Ha az outlierekre vagyunk kíváncsiak, akkor persze számoljunk átlagot, de ha az adatsort jellemezni szeretnénk, akkor jobb robusztus statisztikákat használni.

Kitérő: Egy kis leíró statisztika...

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

módusz medián átlag



Kitérő: Egy kis leíró statisztika

■ Percentilis

- Az n -edik percentilisének az adatok $n\%$ -a kisebb

- {3, 4, 4, 5, 5, 6, 10, 20}

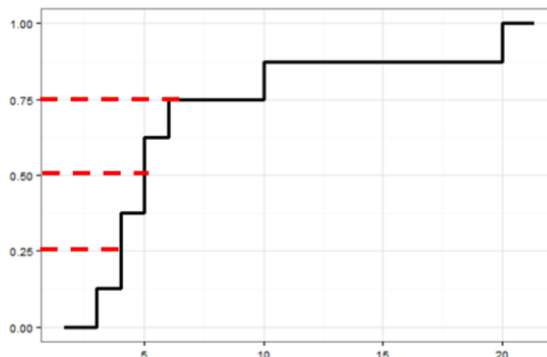
- 50. percentilis: 5
- 25. percentilis: 4
- 75. percentilis: 6

■ Kvartilis

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

■ Inter-quartile range (IQR)

- $Q3 - Q1$



Doboz diagram (boxplot)

- Megjelenített dimenziók száma: 1

- Ábrázolt összefüggés:

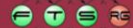
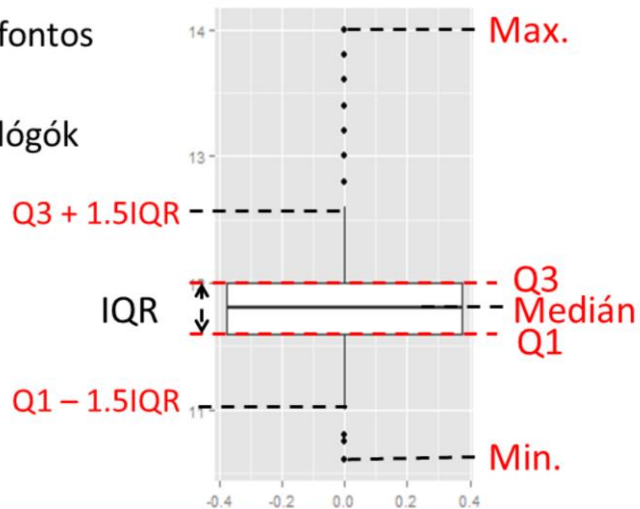
- Folytonos változó fontos percentilisei
- 5 fontos érték + kilógók

- Adategység:

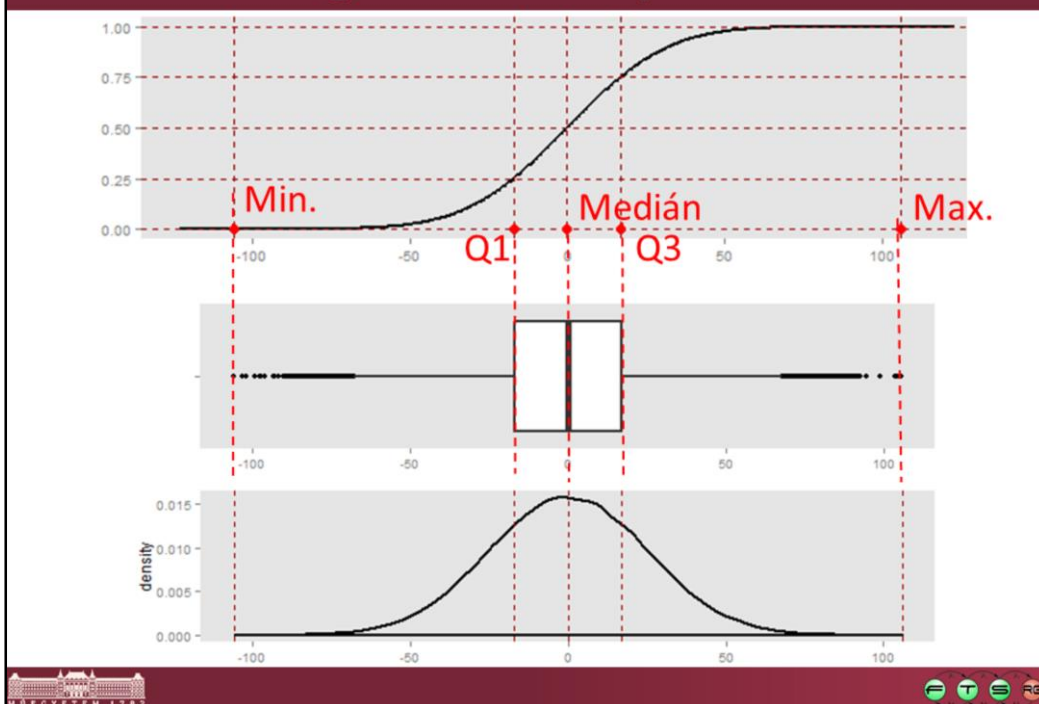
- Doboz

- Tervezői döntés:

- Outlierek?

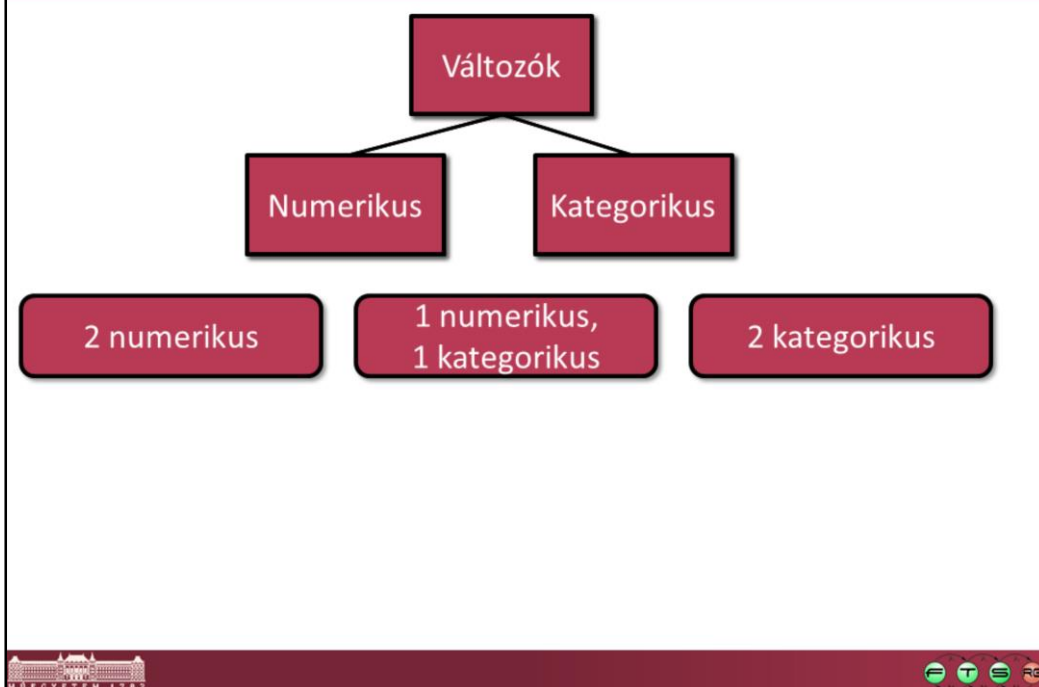


Doboz diagram: fontos percentilisek?

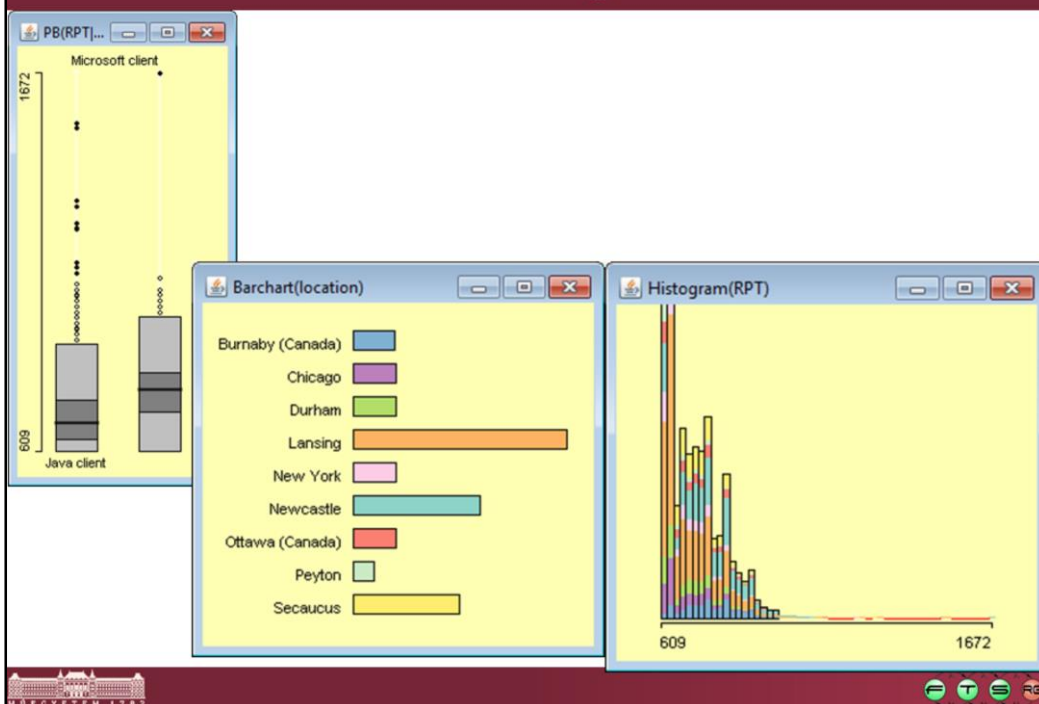


Felső: eloszlásfüggvény, alsó: sűrűségfüggvény

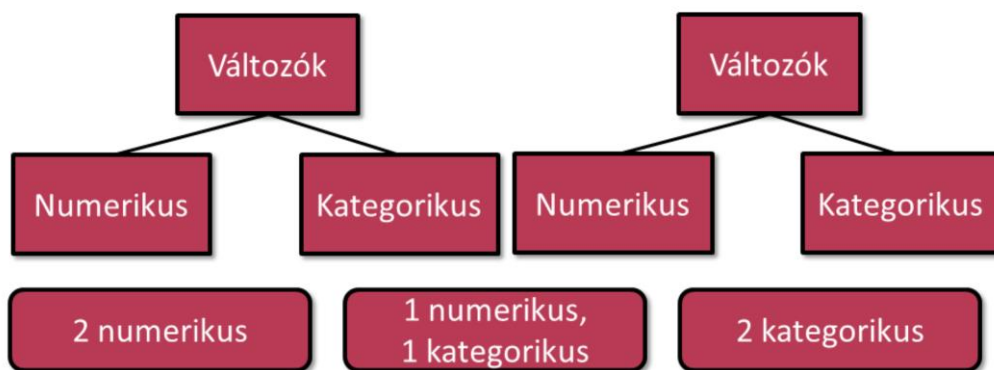
2 változó kapcsolata



Numerikus kategóriánként



2 változó kapcsolata



Pont – pont diagram (scatterplot)

- Megjelenített dimenziók: 2

- Ábrázolt összefüggés:

- folytonos változók együttes eloszlása

- Adategység:

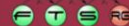
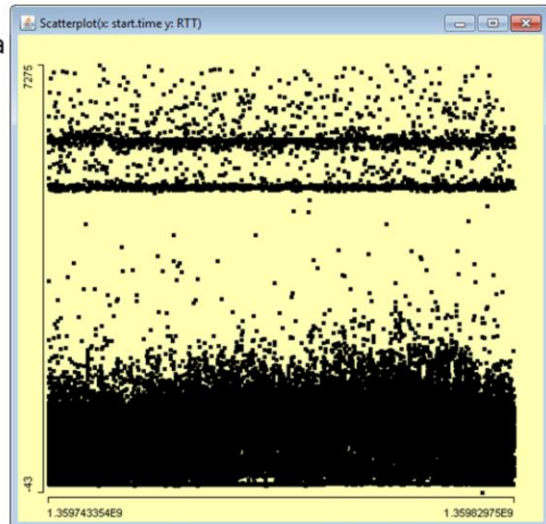
- pont – $X = x_i, Y = Y_i$
előfordulás

- Korlát:

- ha az egyik változó értéke
hiányzik, nem tudjuk felrajzolni

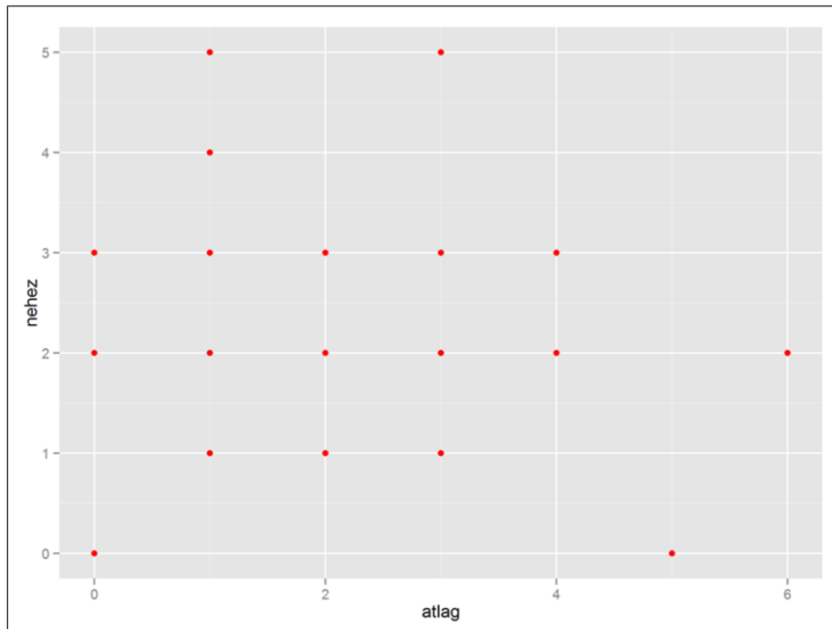
- Tervezői döntés:

- Overplotting?



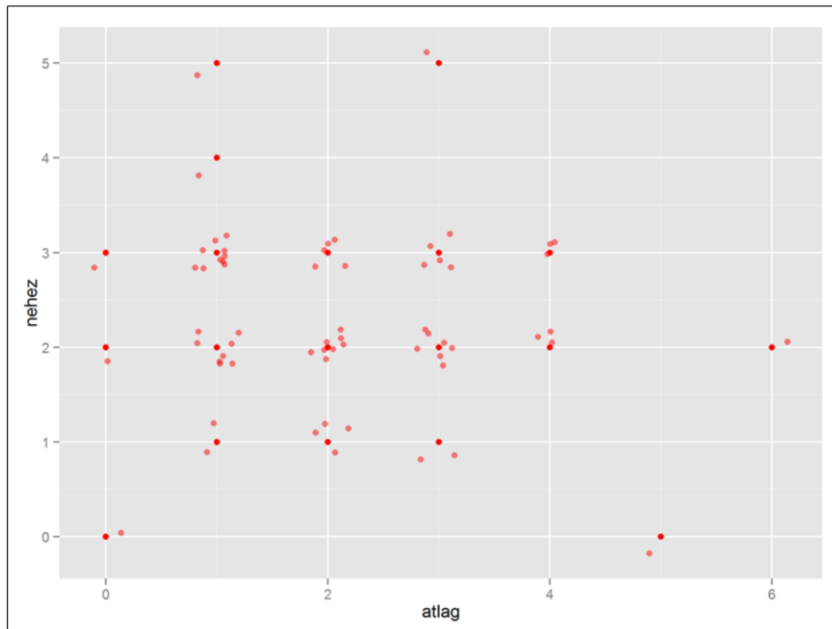
Overplotting: mi van akkor, ha több ugyanolyan (x,y) értékű pontom van?

Overplotting megoldások 1: jitter

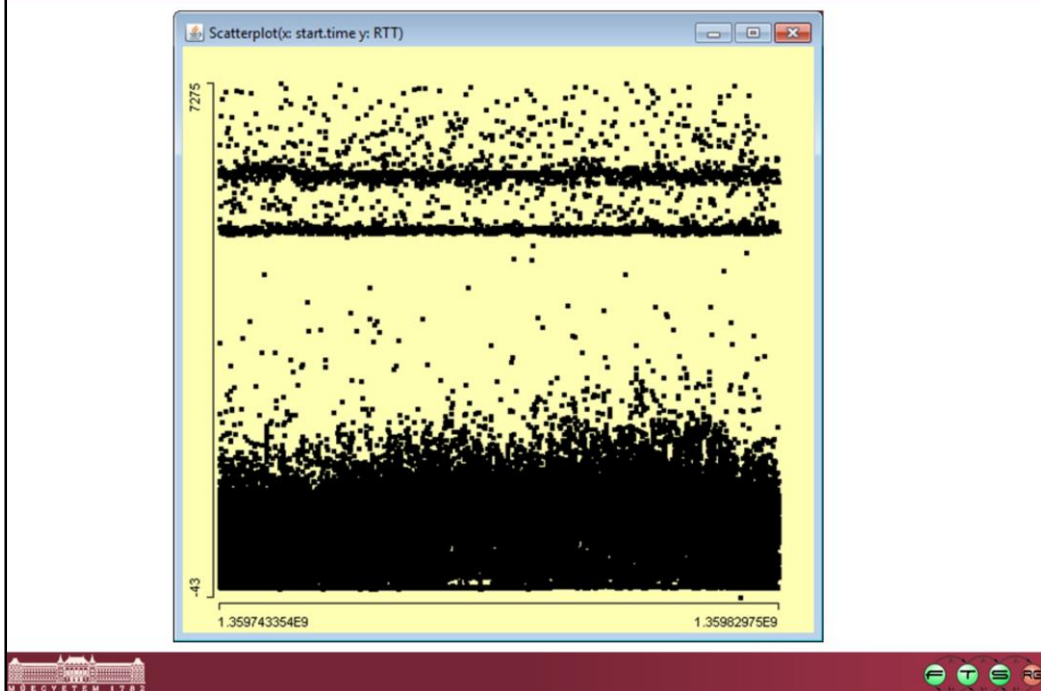


Két diszkrét numerikusnál – ritka, szét kell szedni

Overplotting megoldások 1: jitter

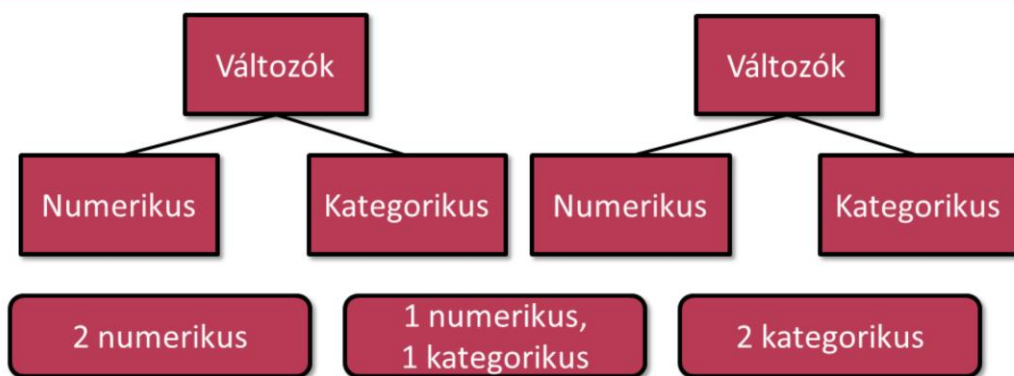


Overplotting megoldások 2: átlátszóság



Két folytonosnál: túl sűrű, szedjük szét!

2 változó kapcsolata



Mozaik diagram (mosaic plot)

- Megjelenített dimenziók száma: 2

A túlsúlyosak nagy része férfi!

- Ábrázolt összefüggés:

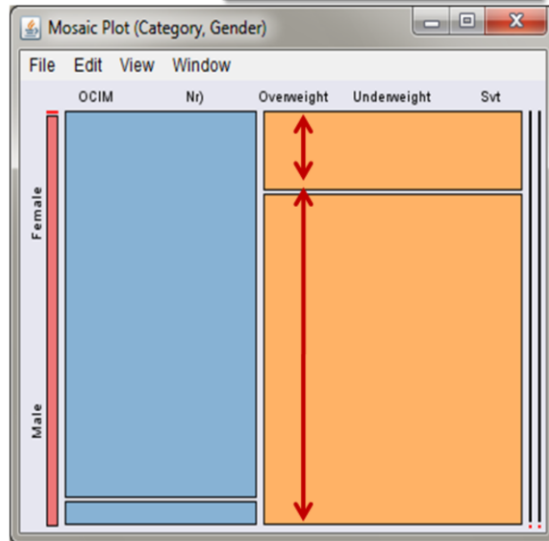
- 2 diszkrét változó e. e.

- Adategység:

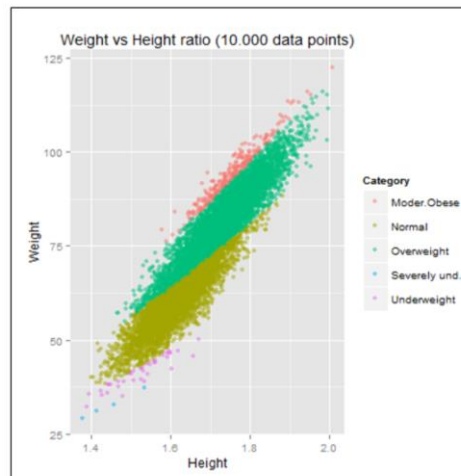
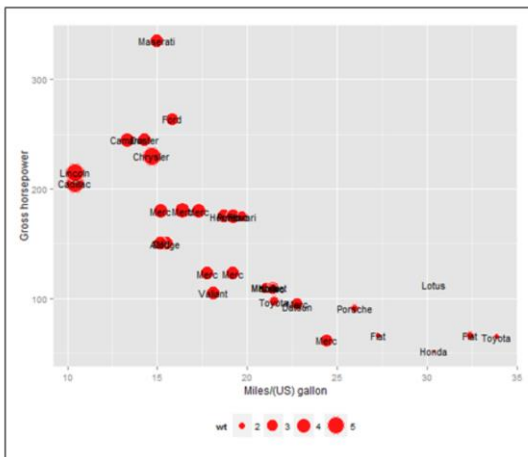
- Téglalap –
területe arányos az
($X = x_i, Y = y_i$)
értékpárok
gyakoriságával

- Korlát:

- Sorfolytonos olvasás?

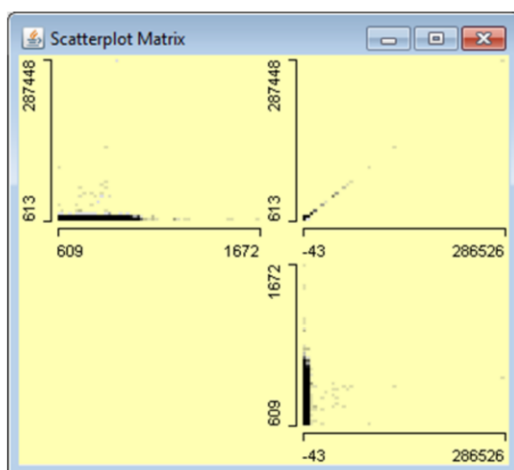


≥ 3 változó – mesterséges dimenziók

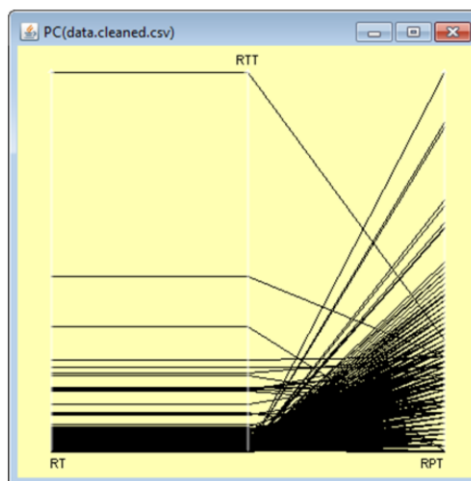


Alak (telített pötty vagy sima, csillag, négyzet stb.) + szín + méret – 3 extra dimenzió, szerencsés esetben akár 5 dimenziót is ábrázolhatunk egyszerre

≥ 3 változó – általánosítás



Egyenes arányosság az RT és az RTT között



Ábra forrása: [4], [5]

Scatterplot matrix: rettenetesen rosszul skálázódik ☹, gyorsan kiszűrhetők az egymással asszociáló változópárok ☺

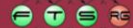
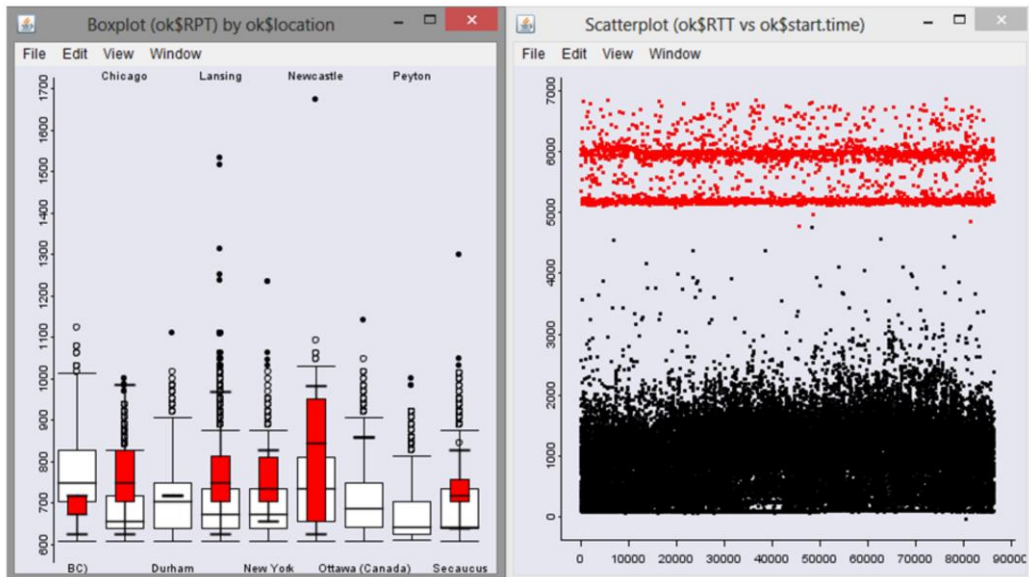
Párhuzamos koordináták: a tengelyek különböző nagyságrendje torzíthat az összefüggésen ☹, aki érti a geometriáját, az könnyen/gyorsan következtet (Pl. egy pontban metszik a tengelyek egymást → 1-es vagy -1-es korreláció látszik) ☺ → Optimális változósorrendet találni? ☹

Miről lesz szó?

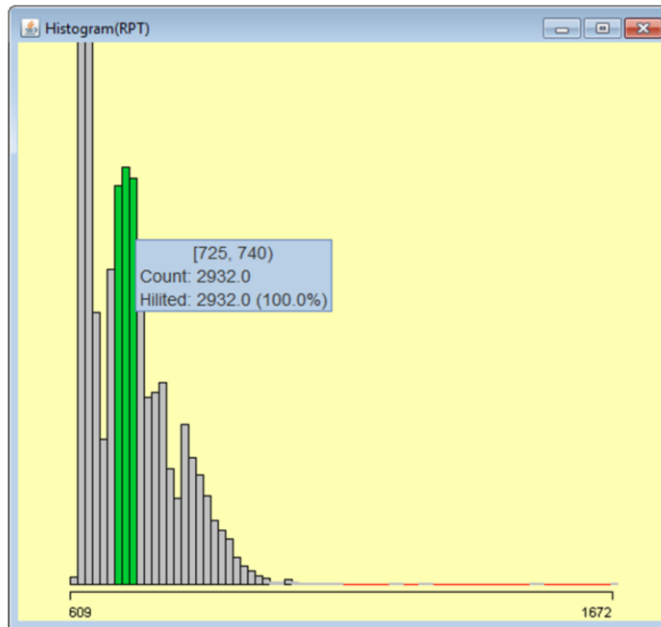
- Adatelemzési alapfogalmak
- Alapvető diagramtípusok
- **Interaktív EDA eszközök – elvárt funkcionalitás**
- Esettanulmány: cloud benchmarking



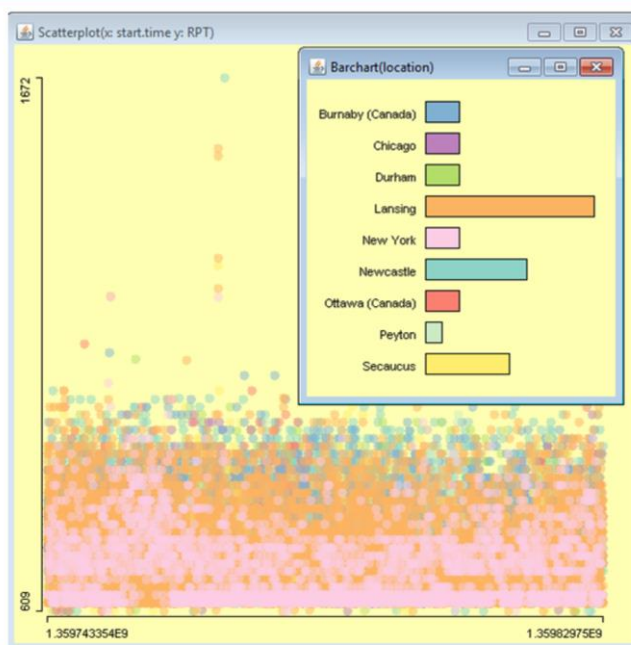
Adatkötés



Lekérdezések



Színezés/átlátszóság



Miről lesz szó?

- Adatelemzési alapfogalmak
- Alapvető diagramtípusok
- Interaktív EDA eszközök – elvárt funkcionalitás
- **Esettanulmány: cloud benchmarking**



Mit látunk az adatokból?

TABLE I. MS AZURE END-TO-END RESPONSE TIME

Client's location		Client's IP address	Response Time (RT), ms				
			Min	Avg	Max	Std Dev	CV, %
Canada	Ottawa	216.151.172.x	1484	2007	3917	255	12.7
	Burnaby	64.151.226.x	1841	2329	287448	7648	328.4
UK	Newcastle	10.8.146.x	859	1577	4016	480	30.4
	Newcastle	10.8.151.x	890	1498	3438	423	28.2
	Durham	213.175.197.x	814	1922	47740	2132	110.9
USA	New York	69.72.183.x	1331	2018	7145	575	28.5
	Chicago	209.188.85.x	1133	2104	11729	875	41.6
	Peyton	64.64.0.x	1289	1961	4637	350	17.9
	Lansing	67.225.254.x	902	2051	4483	396	19.3
		67.225.254.x	1421	2024	3992	365	18.0
		67.227.193.x	1441	2068	19365	1004	48.5
		67.227.216.x	6123	6911	11520	366	5.3
		67.227.216.x	1420	2016	3770	344	17.1
	Secaucus	204.14.93.x	1262	1983	47056	1499	75.6
		208.87.24.x	1216	1991	46457	1271	63.8
		208.87.25.x	1212	2030	59475	2232	110.0
		64.20.37.x	1216	2275	132498	5135	225.7



Cloud benchmarking

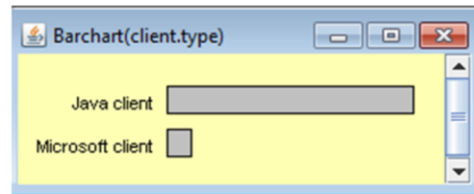
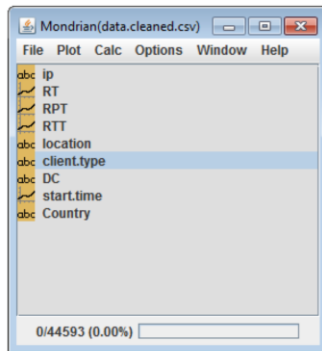
- Alapvető RT-RTT összefüggések
- Kísérlettervezési hiányosságok
- Konfiguráció hibák
- Térbeli/időbeli/kliensbeli függőségek



- RT-RTT scatterplot és a többi változat
- Hely-kliens típus párok ábrázolása, hiányoznak kombinációk
- Lansing-en belül az egyik kiugrik; boxploton megjelenik
- Datacenter összehasonlítása / munkaidőben nagyobb a késleltetés Newcastle-ben / java és .net kliens

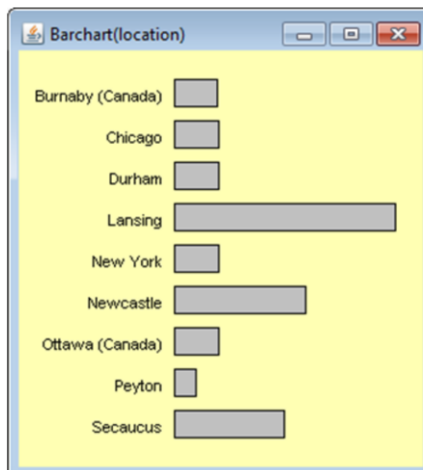
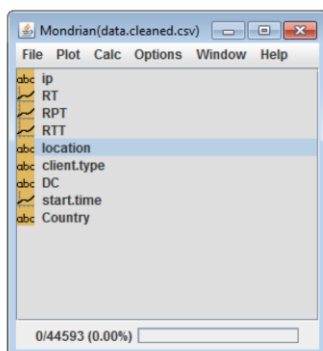
Oszlopdiagram (bar chart)

- Bemenő változó: kliens típus
- Kérdés: melyik klienssel mennyit mértünk?

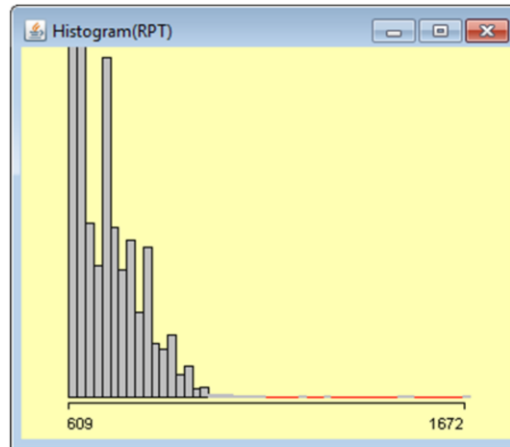
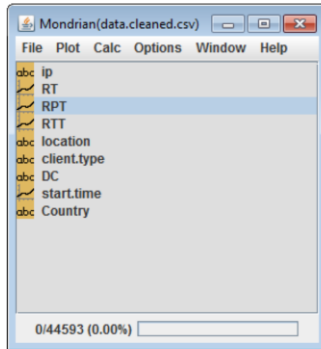


Oszlopdiagram (bar chart)

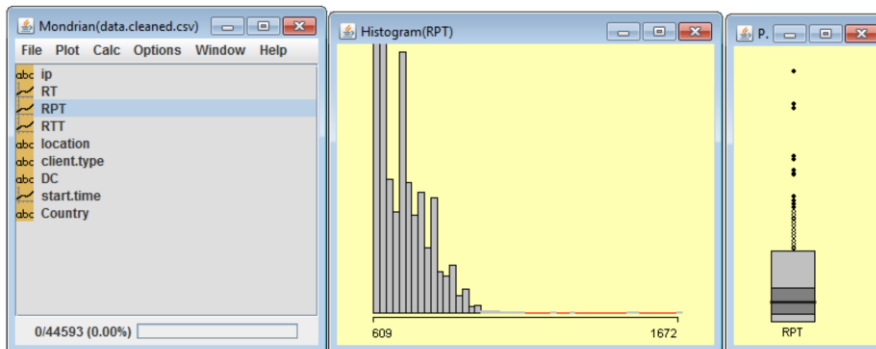
- Bemenő változó: mérési helyek
- Kérdés: milyen helyekről mértünk? Mennyit?



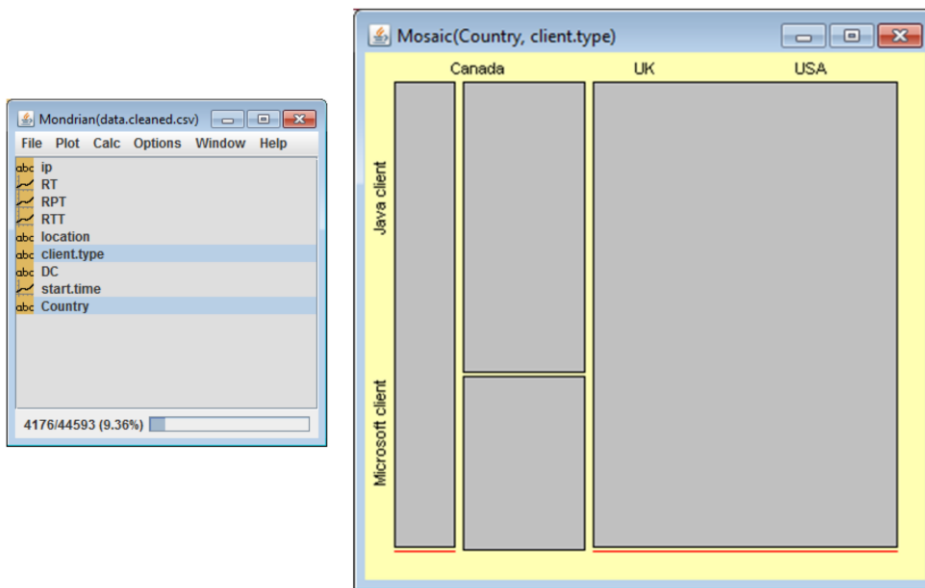
Hisztogram



Doboz diagram (boxplot)

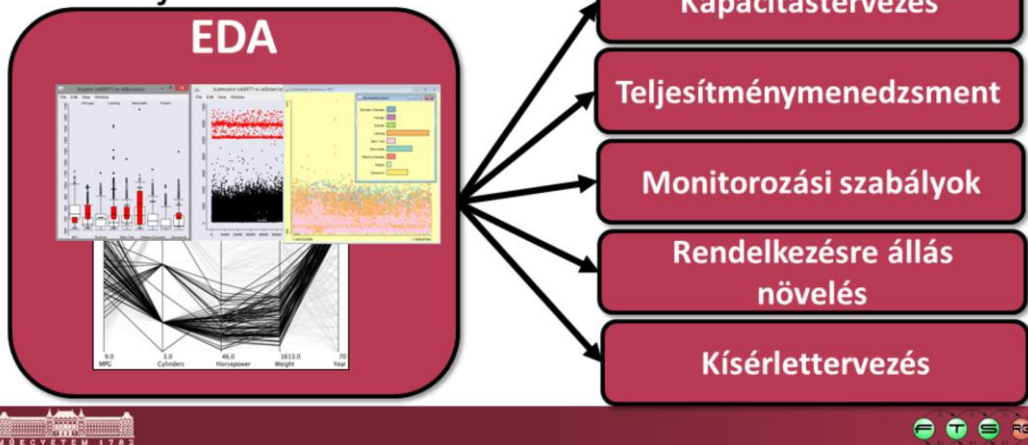


Mozaik diagram (mosaic plot)



Összefoglalás

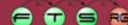
- Miért jó?
 - Összehasonlítás
 - Tetszőleges mélység
- Mire jó?



Szűk keresztmetszetek keresése, legyen az szolgáltatásbiztonság vagy teljesítménymenedzsment : időbeli/térbeli függés fontos lehet.
Kapacitástervezés, monitorozás: RTT befolyásol inkább, mint az RPT, akkor monitorozzuk/javítsuk ezt a részét a dolgoknak, felesleges jobb processzort venni, inkább a hálózatunk legyen normálisan bekonfigurálva

További információ

- I. Kocsis, „Vizuális analízis”. Intelligens adatelemzés, Typotex, 2014.
<http://docs.inf.mit.bme.hu/remo-jegyzet/felderito-adatelemzes-konyvfejezet.pdf>
- H. Hartmann, „Statistics for Engineers”, acm queue, 14:1, 2006.
<http://queue.acm.org/detail.cfm?id=2903468>



Hivatkozások

- [1] Behrens, J.T.: Principles and procedures of exploratory data analysis. *Psychological Methods* 2, 131–160 (1997)
- [2] Tukey, J.: We need both exploratory and confirmatory. *The American Statistician* 34, 23–25 (1980)
- [3] Yau, Nathan. *Visualize this: the FlowingData guide to design, visualization, and statistics*. John Wiley & Sons, 2011.
- [4] Inselberg, A.: *Parallel Coordinates: Visual Multidimensional Geometry and its Applications*. Springer Science+Business Media, New York (2009)
- [5] Theus, M., Urbanek, S.: *Interactive graphics for data analysis: principles and examples*. CRC Press (2011)
- [6] Gorbenko, A., Kharchenko, V., Mamutov, S., Tarasyuk, O., Romanovsky, A.: Exploring Uncertainty of Delays as a Factor in End-to-End Cloud Response Time. In: *2012 Ninth European Dependable Computing Conference*, pp. 185–190. IEEE (2012)
- [7] Pataricza, András, et al.: Empirical Assessment of Resilience. *Software Engineering for Resilient Systems*. 1-16. (2013)
- [8] Funk, Kaiser: Junk Charts blog, URL: <http://junkcharts.typepad.com/>
- [9] Yau, Nathan: FlowingData blog, URL: <http://flowingdata.com/>

