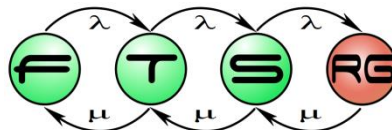


Interaktív és Big Data vizualizáció

2017 ősz, 4./5. alkalom

Kocsis Imre, ikocsis@mit.bme.hu

Budapest University of Technology and Economics
Fault Tolerant Systems Research Group



Interaktív statisztikai grafika

Vezetett adatbejárás – „data tour” / “projection pursuits”

Lekérdezések

Kijelölés és
csatolt
kiemelés

Csatolt
analízisek

Interakció az
ábrákkal

Ábrák képzése – „plotolás”

Interaktív statisztikai grafika

Nem tárgyaljuk, ipari eszközök nem implementálják
lásd pl. R tourr csomag

Vezetett adatbejárás – „data tour” / “projection pursuits”

Lekérdezések

Kijelölés és
csatolt kiemelés,
color brush

Csatolt
analízisek

Interakció az
ábrákkal

BI eszközökben helyette inkább:

- kijelölés és csatolt szűrés
- csatolt drill down / roll up

ás”

[1] alapján

Előző előadás

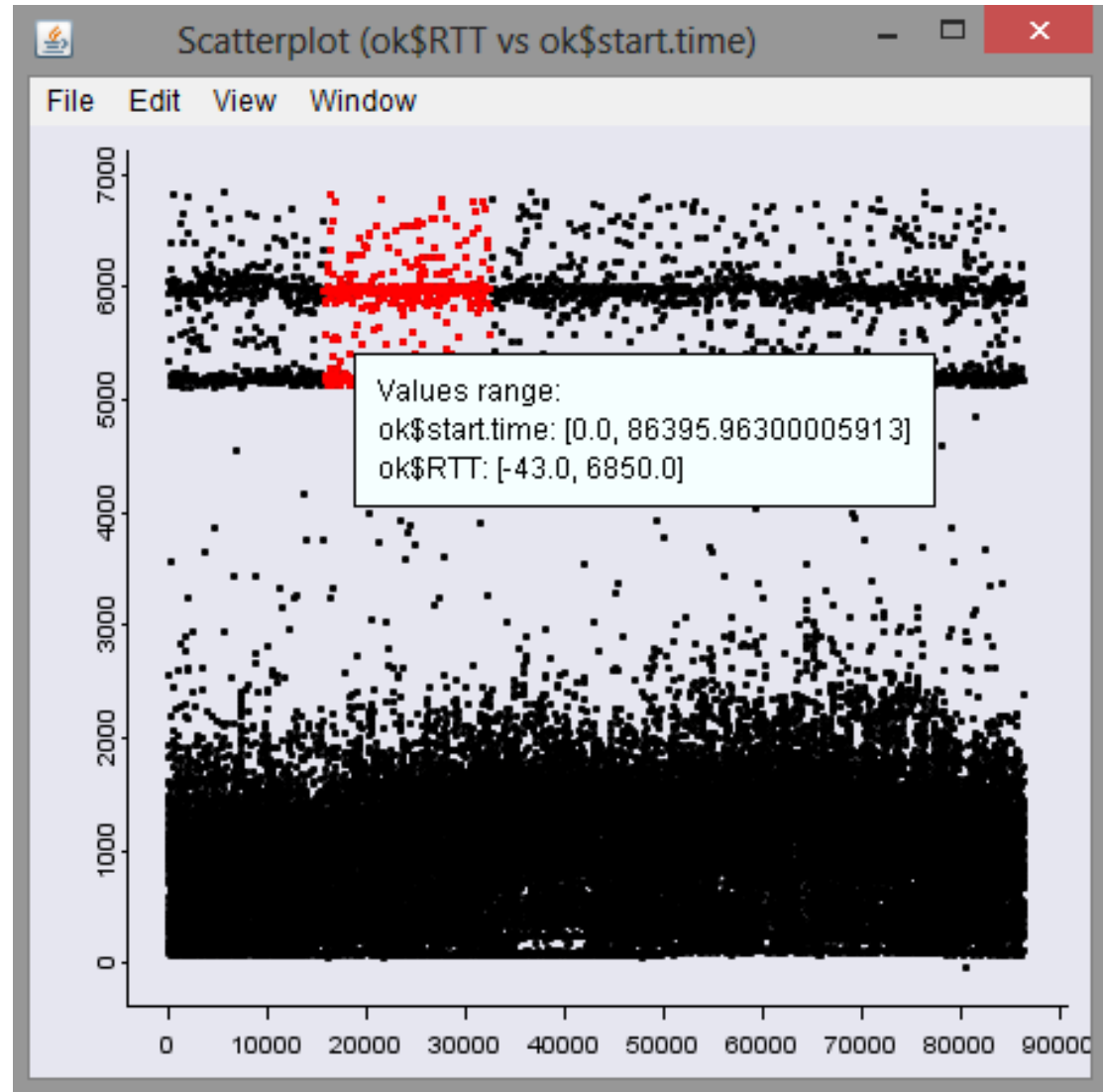
iPlots

- R csomag interaktív vizualizációhoz
- Oktatási példa – képességei elérhetőek / mostanában jelennek meg az ipari eszközökben
 - “Élesben” használata csak erősen megfontolva javasolt
 - “Industry strength”: notebook/webalkalmazás-ba ágyazva
 - R: shiny + ggplots2/D3.js/..., Python: pl. Jupyter + bokeh
- <http://stats.math.uni-augsburg.de/iplots/>
 - Mondrian, Rserve, rJava
- Interaktív...

Bar chart, Box plot, Hammock plot, Histogram, Map, Mosaic Plot, Parallel Coordinates Plot, Scatterplot

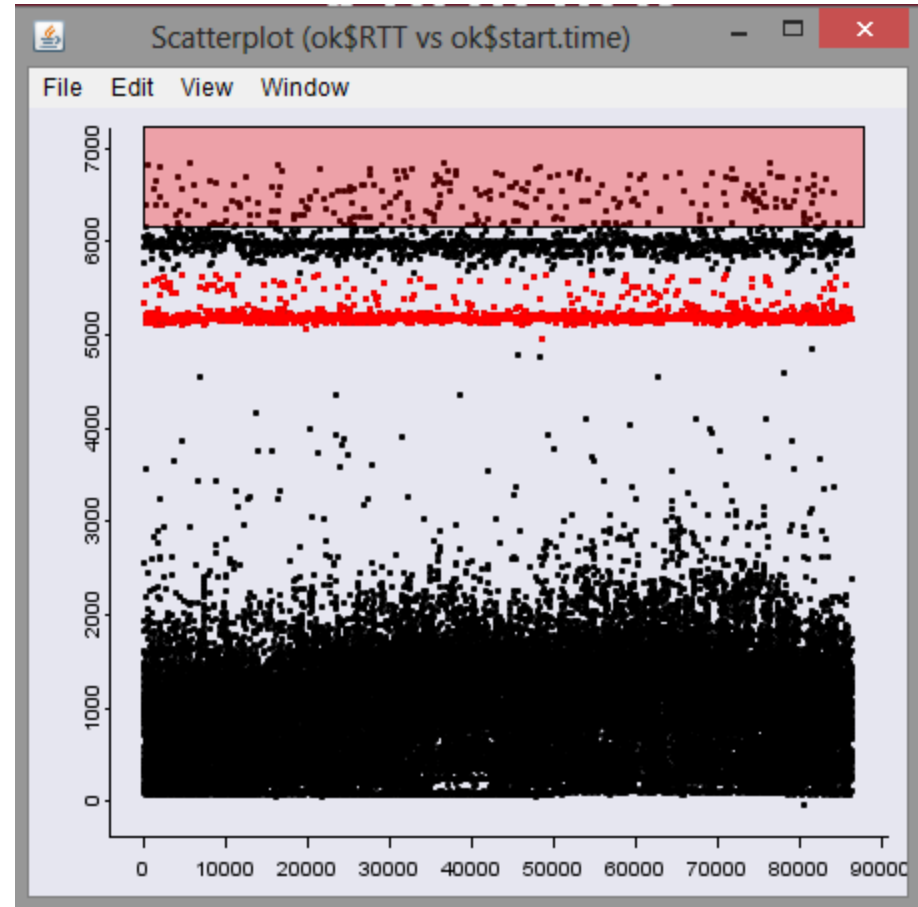
Lekérdezések

- „Query”
- iPlots: CTRL
- ~~Többszintű lekérdezés~~

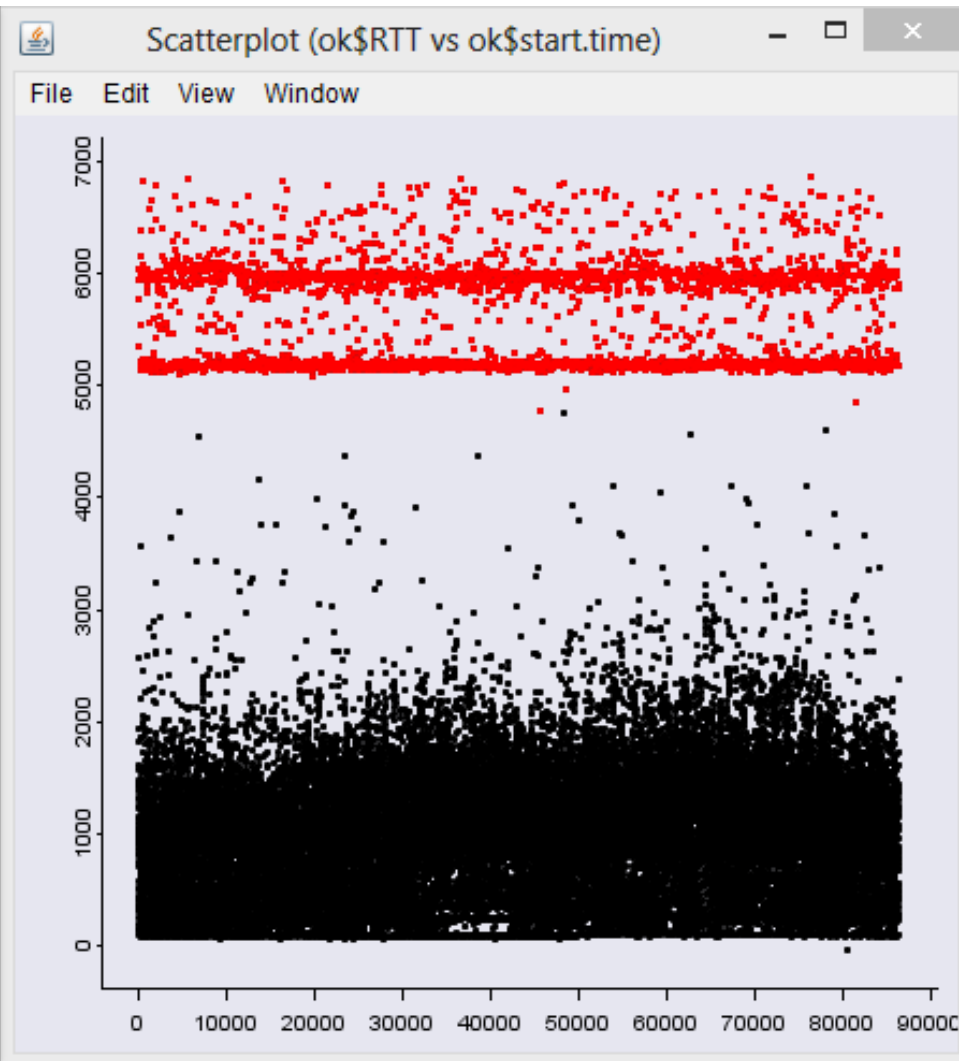
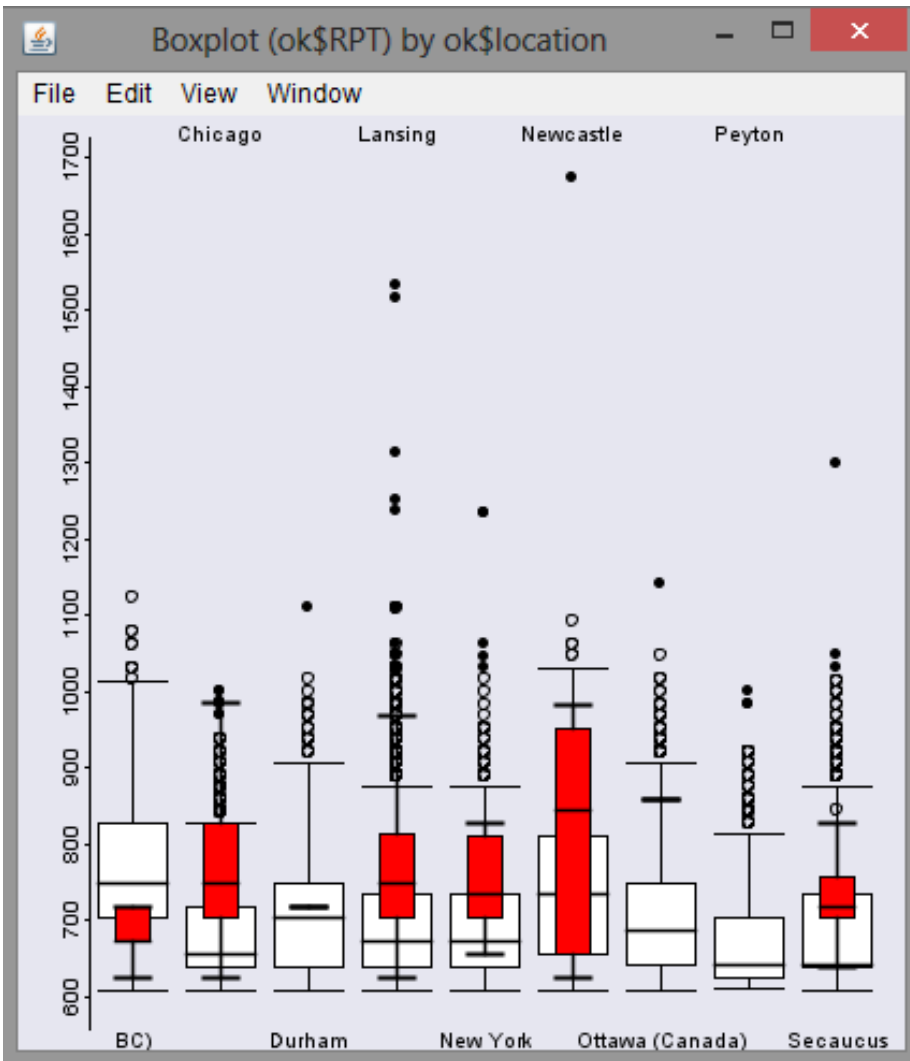


Kijelölés

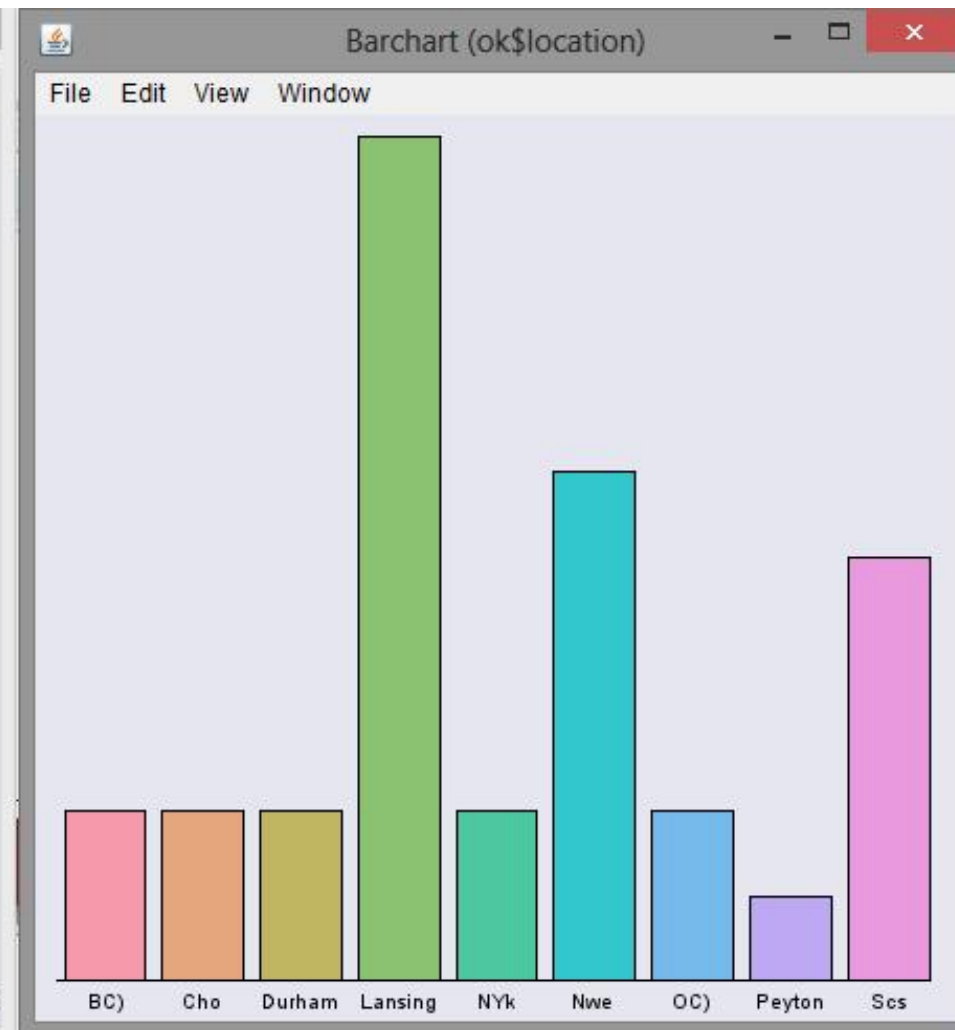
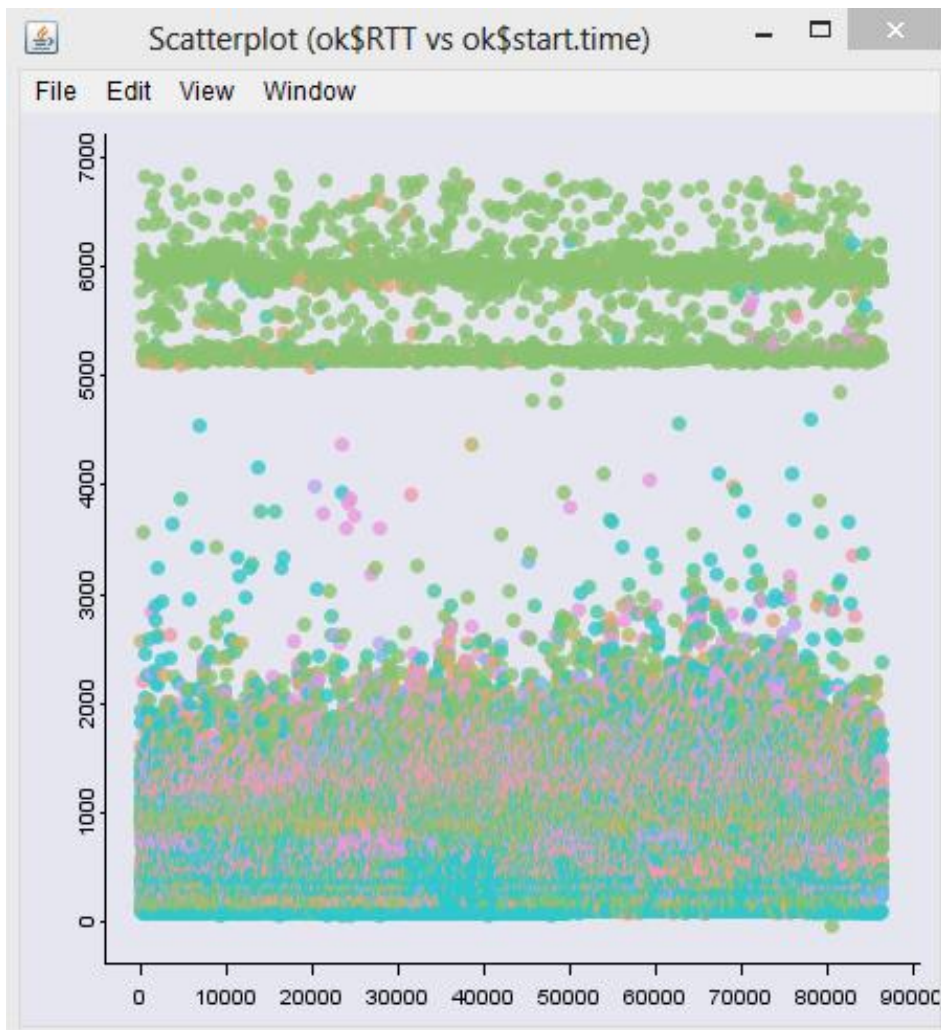
- SHIFT-CTRL: OR
- SHIFT: XOR
- Pointer, Drag-box, ~~Brush, Slicer, Lasso~~
- ~~Kijelölés-sorozatok~~



Csatolt kiemelés

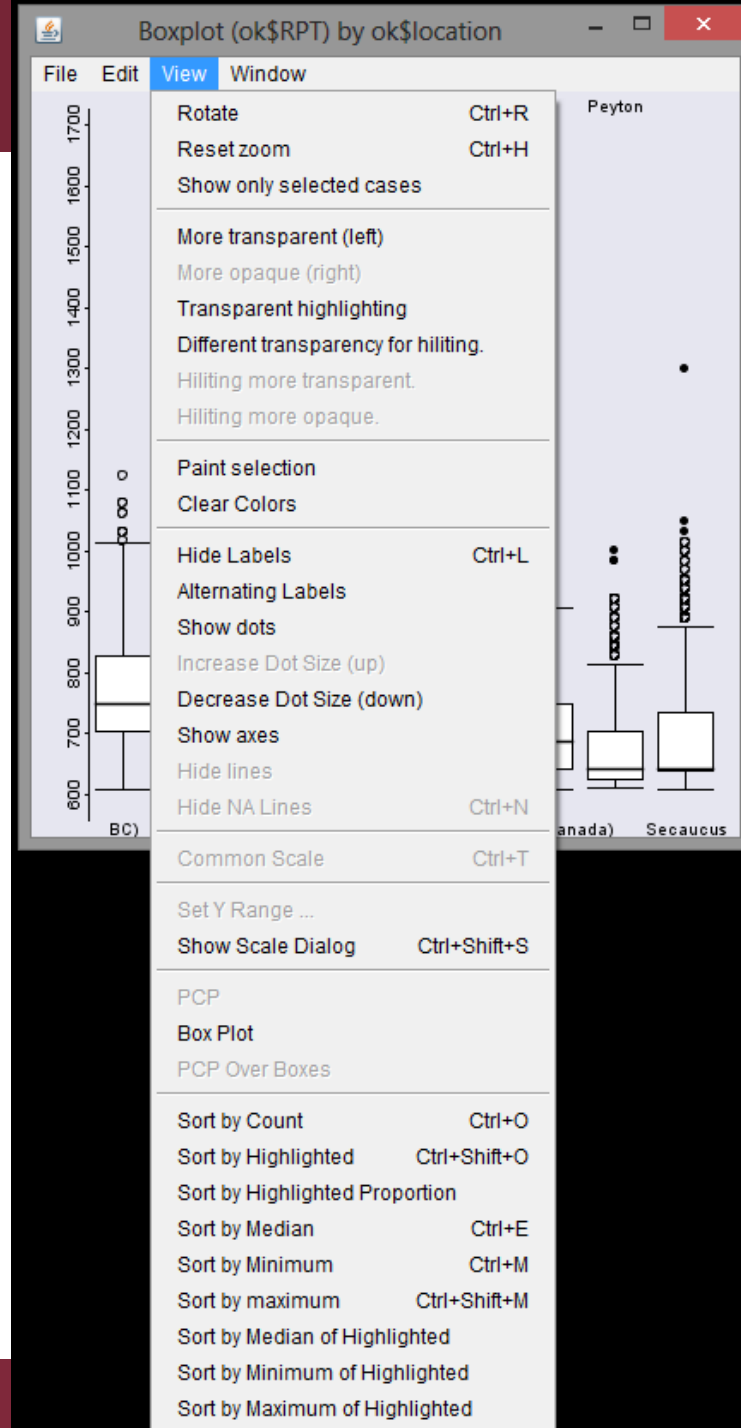


„Color brush”



Interakció az ábrákkal

- Billentyűkombinációk és menük
- Paraméterek (pl. hisztogram)
- Tengelyek megcserélése
- Skálázás
- Nagyítás (középső egérgomb)
- Áttetszőség ($\leftrightarrow$)



Vizuális elemzés: motivációs példa

- Számítási felhő teljesítménymérések
- Response Time = Request Processing Time + Round Trip Time

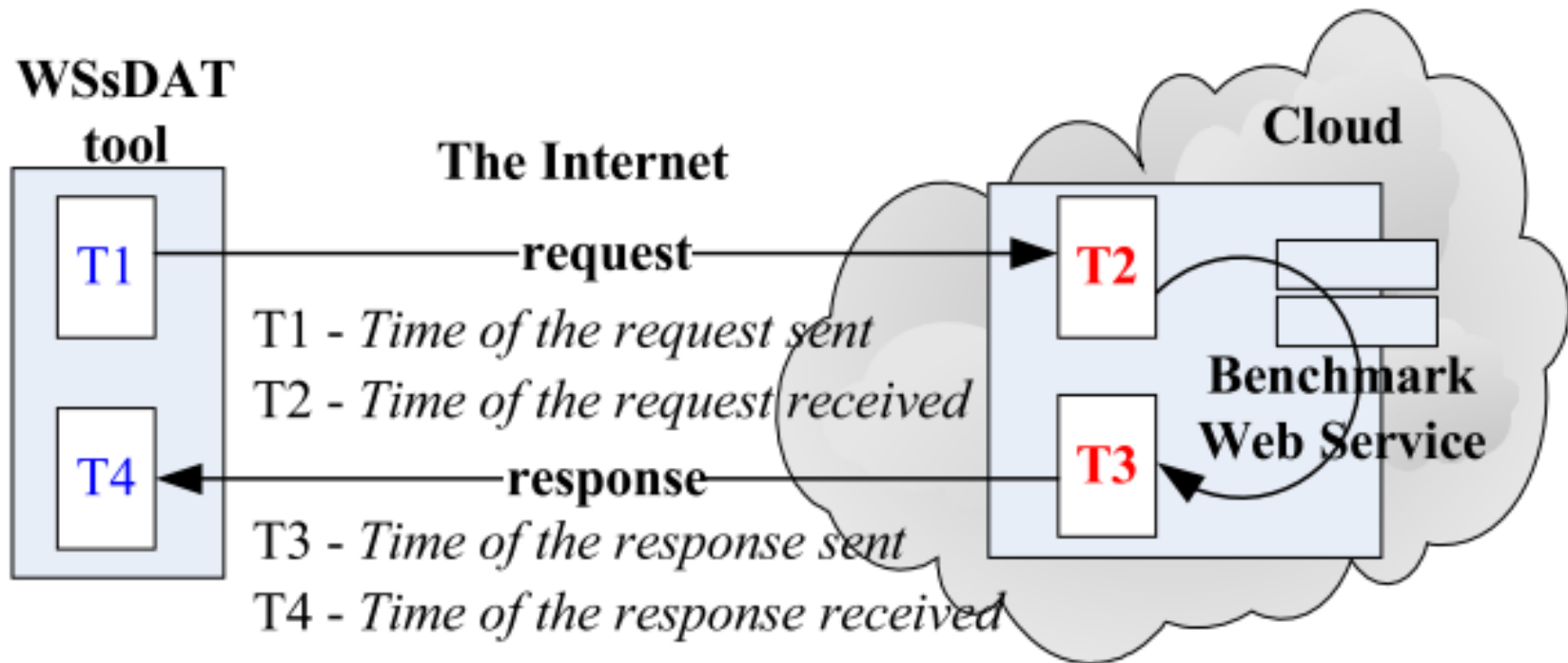
Timestamp	Client IP	Client location	Client type	Server location	RPT/RTT/RT (ms)
-----------	-----------	-----------------	-------------	-----------------	-----------------

Interaktív adatelemzés demo Mondrian-nal: www.theusrus.de

Kapcs. cikk:

https://inf.mit.bme.hu/sites/default/files/publications/2013_SERENE_Pataricza-ag-ar_v2.pdf

Példa adatkészlet



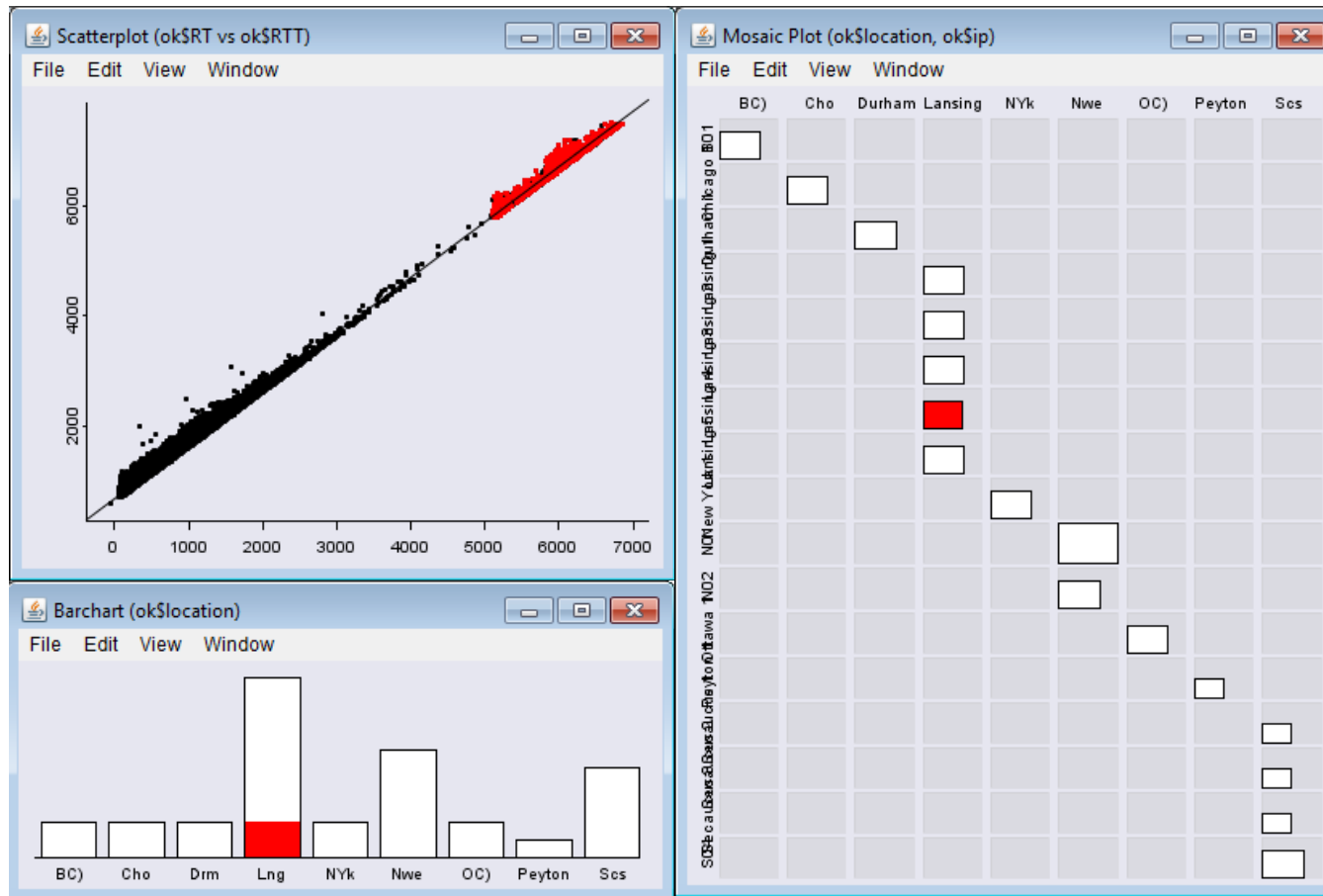
$$RT = T4 - T1$$

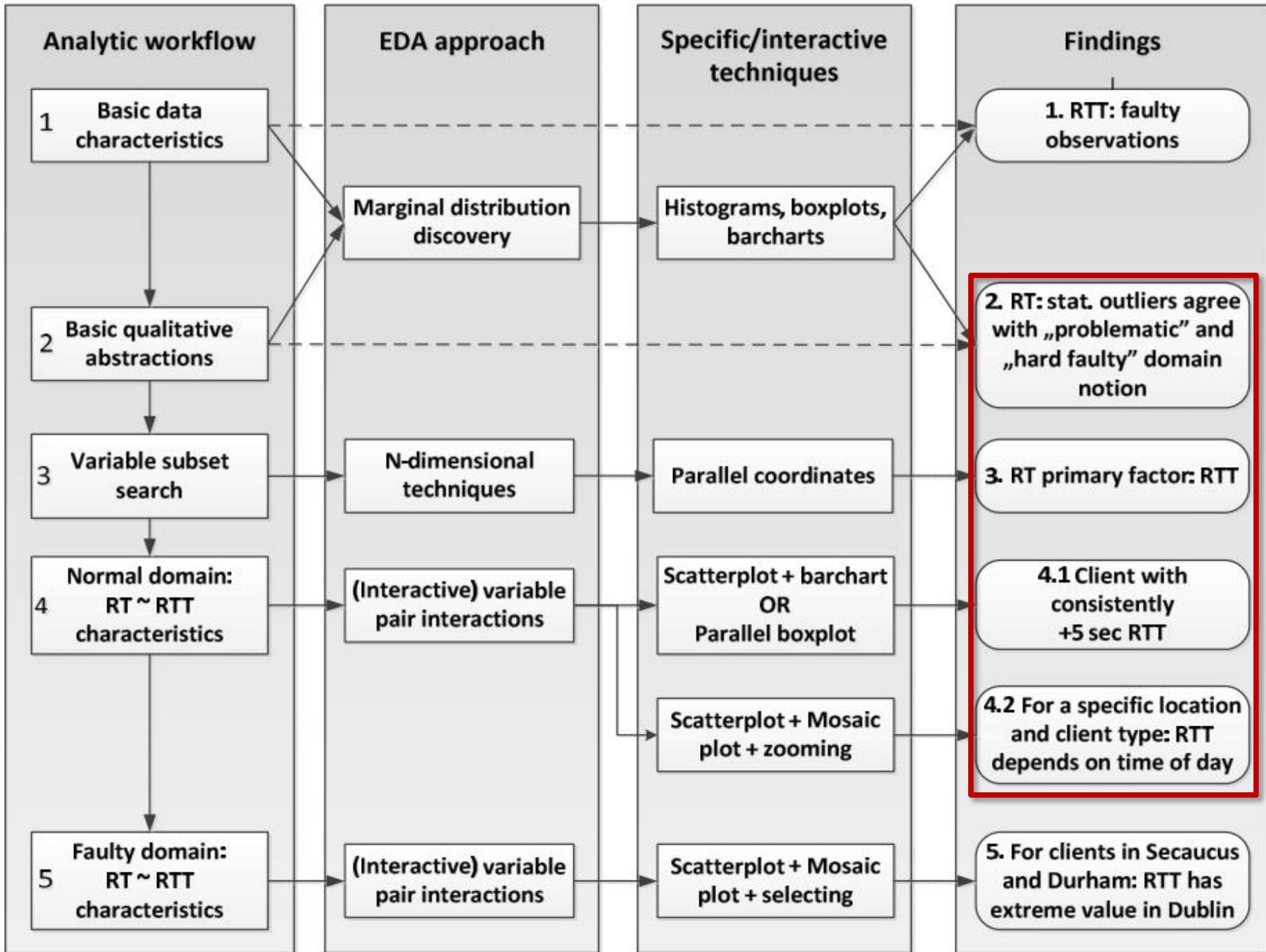
$$RPT = T3 - T2$$

$$RTT = (T2 - T1) + (T4 - T3) = (T4 - T1) - (T3 - T2) = RT - RPT$$

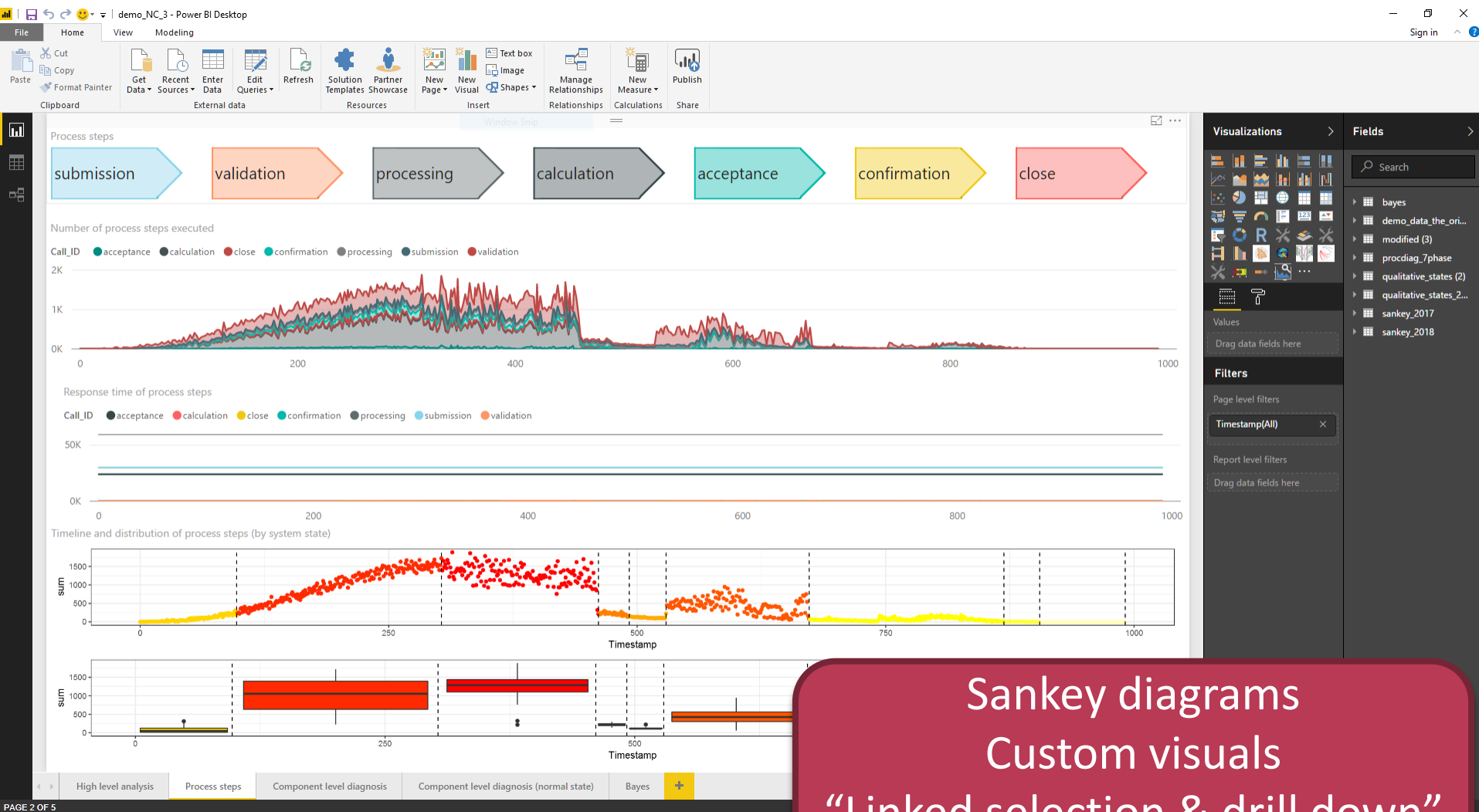
Demo

■ R + iPlots





Demo: vizuális elemzés PowerBI-ban

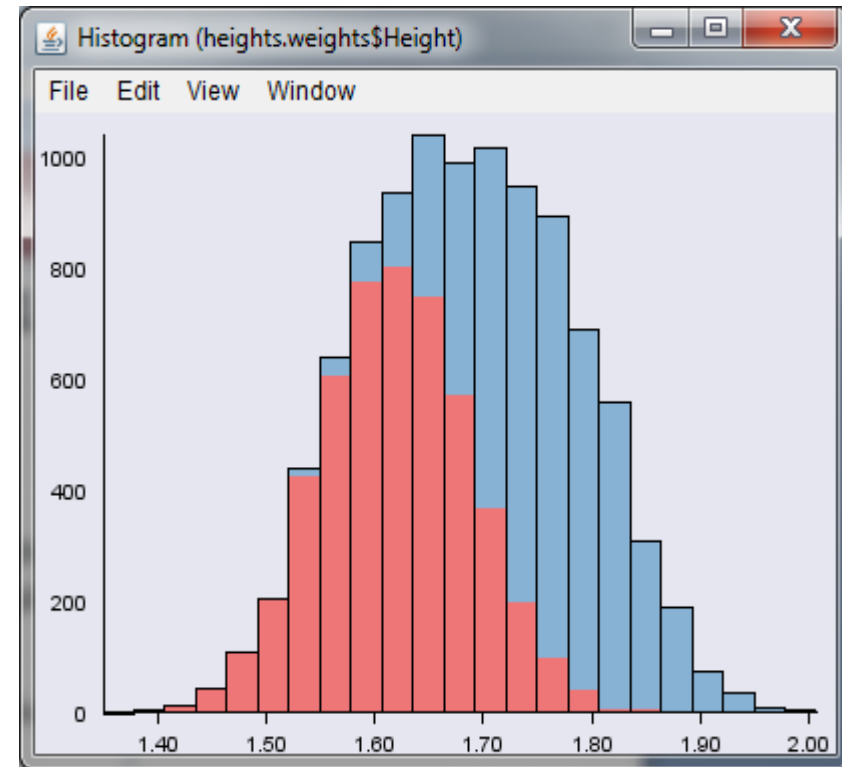
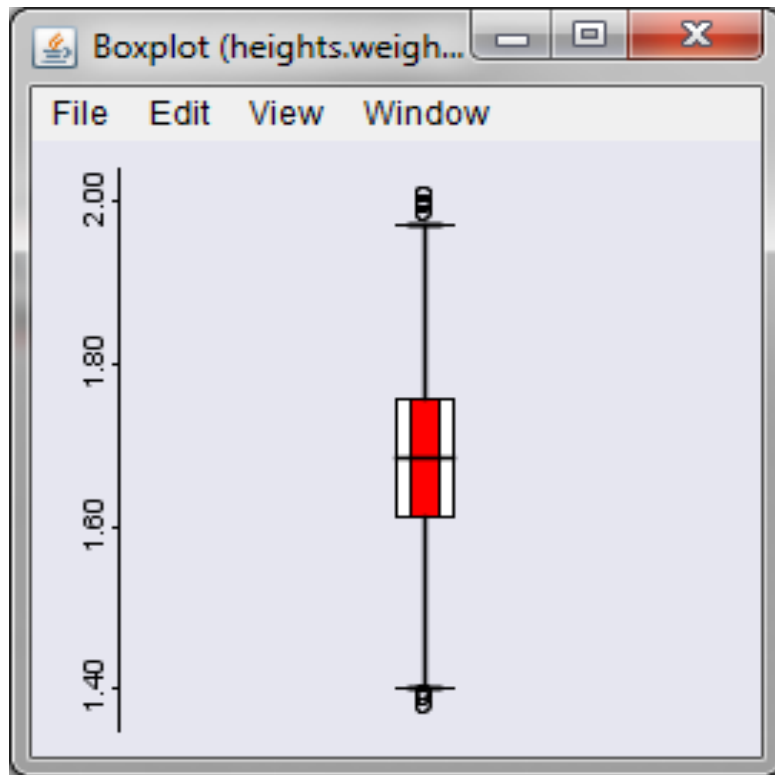


Sankey diagrams
Custom visuals
"Linked selection & drill down"
"Big Data képes": DirectQuery!

'BIG DATA' VIZUALIZÁCIÓ

Aggregálunk

- n nagy $\rightarrow$ mesterségesen tömörítünk



Bin-summarize-smooth-visualize

- A képernyő pixelszáma erősen véges
- Az előfeldogozást „le kell csatolni” a megjelenítésről
- Lehetővé teszi a
 - Párhuzamosítást
 - Out-of-memory adatok megjelenítését
- A fontosabb 1d és 2d statisztikai eszközök
- *Alapvető forrás: [1] „H.Wickham: Bin-summarize-smooth: A framework for visualizing large data”*

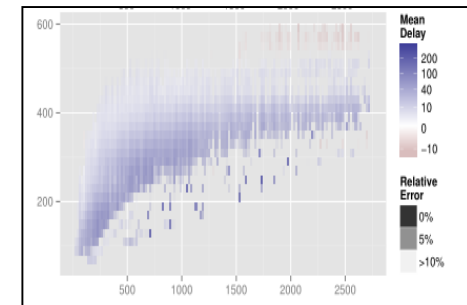
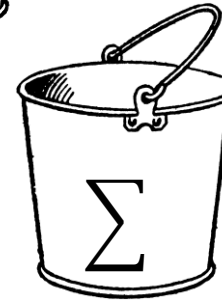
Bin-summarize-smooth-visualize

Bin

Summarize

Smooth

Visualize



„Condense”

- Bin + summary: nagy adat $\rightarrow$ „dobozolt” összefoglalók
- „dobozolás” (*binning*), majd néhány leíró statisztika bin-hez rendelése
- Binning: injektív leképezés
- Adatbázisban is végezhető

Példa adatsor: flight data

- ASA Data Expo '09
- <http://stat-computing.org/dataexpo/2009/the-data.html>
- Változók
 - Year, Month, DayOfMonth, DayOfWeek
 - DepTime, SchDepTime, ArrTime, SchArrTime
 - ArrDelay, DepDelay
 - Origin, Dest
 - Distance

„Bin”

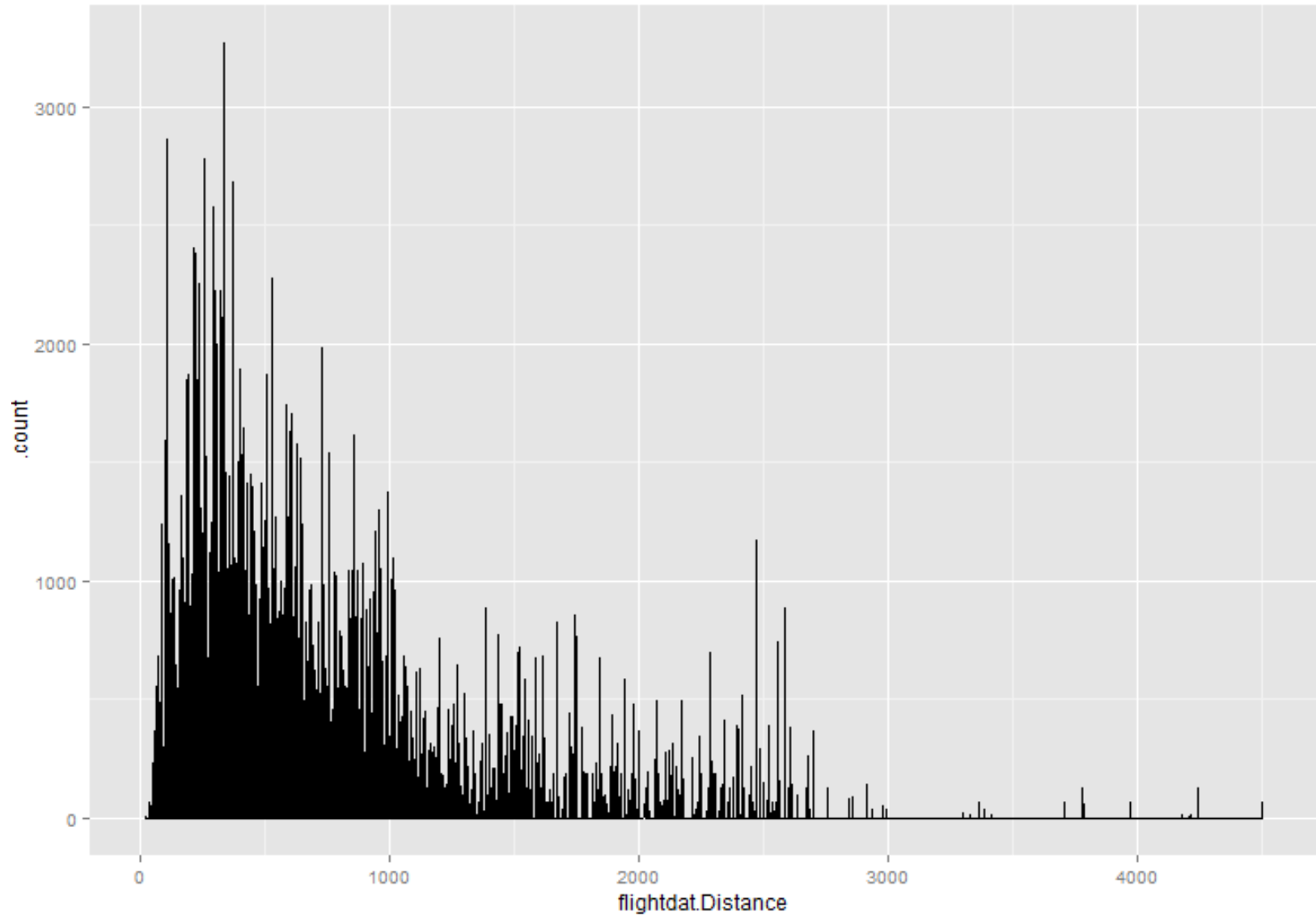
- Fix szélességű dobozok
- Egy dimenzióban: $\left\lfloor \frac{x - origin}{width} \right\rfloor + 1$

- Általánosítás több dimenzióban

$$\begin{aligned} &= x_1 + x_2 \cdot n_1 + x_3 \cdot n_1 \cdot n_2 + \cdots + x_m \prod_{i=1}^{m-1} n_i \\ &= x_1 + \cdot (x_2 + n_2 \cdot (x_3 + \cdots (x_m))) \end{aligned}$$

- Ritka adatok: jobb lenne a „nagyobb” szélesség
 - Pl. a variancia csökkentésére
 - Nehéz probléma → Inkább simítás

„Bin“



„Summarise”

- Összefoglaló statisztikák típusai:
 - Disztributív
 - egyetlen, adott méretű köztestár
 - eredmények kombinálhatóak
 - pl. count, sum
 - Algebrai
 - disztributív statisztikák fix száma kell hozzá
 - Pl. átlag: count + sum
 - Holisztikus
 - bemenettel növekvő köztestár kell
 - Pl. medián

„Summarise”

■ Összefoglaló statisztikák típusai:

○ Disztributív

- egyetlen, adott méretű köztestár
- eredmények kombinálhatóak
- pl. count, sum

○ Algebrai

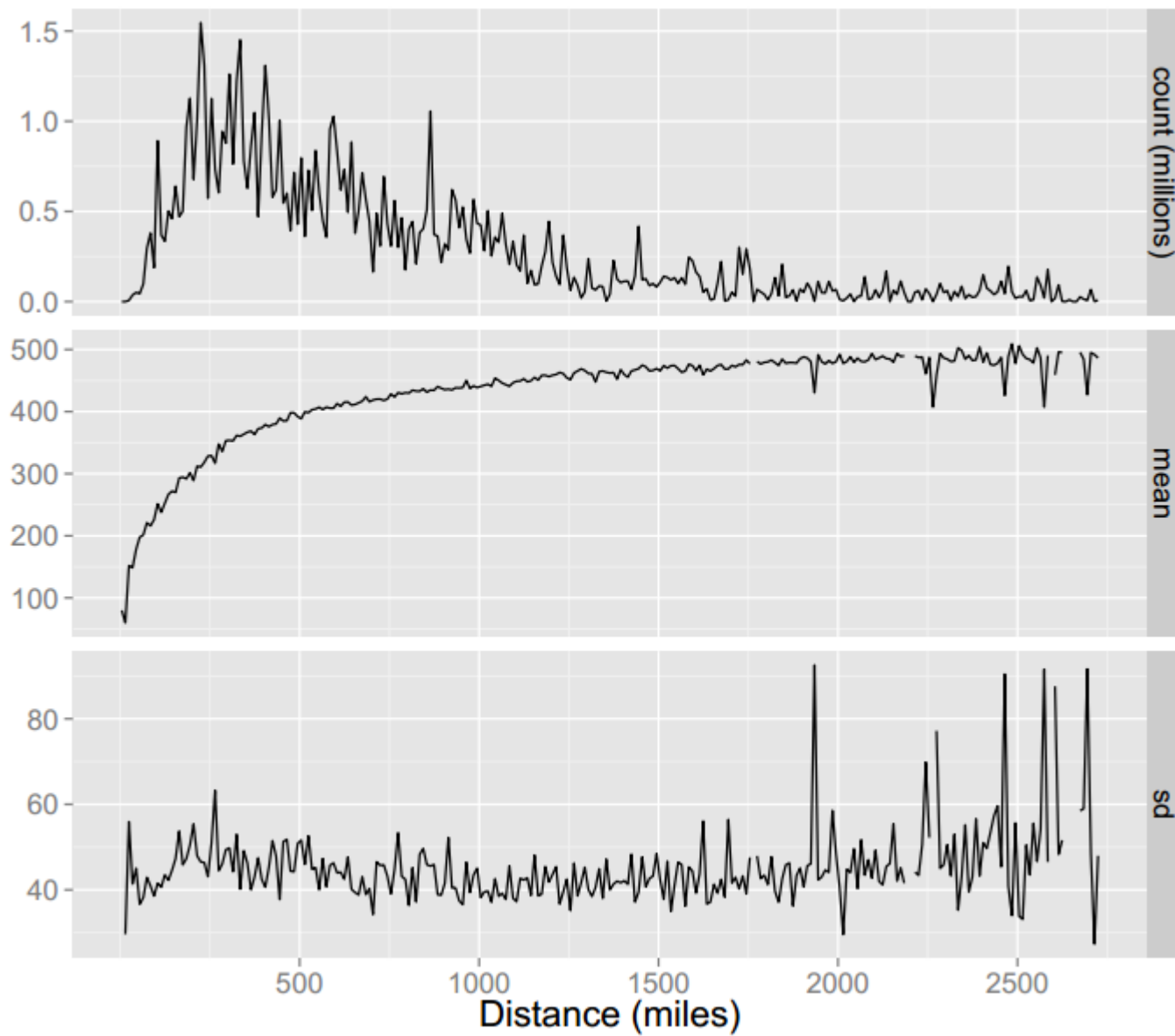
- disztributív statisztikák fix száma kell hozzá
- Pl. átlag: count + sum

○ Holisztikus

- bemenettel növekvő köztestár kell
- Pl. medián

1. Általában jól párhuzamosítható
2. Interaktív vizualizáció

„Summarize”

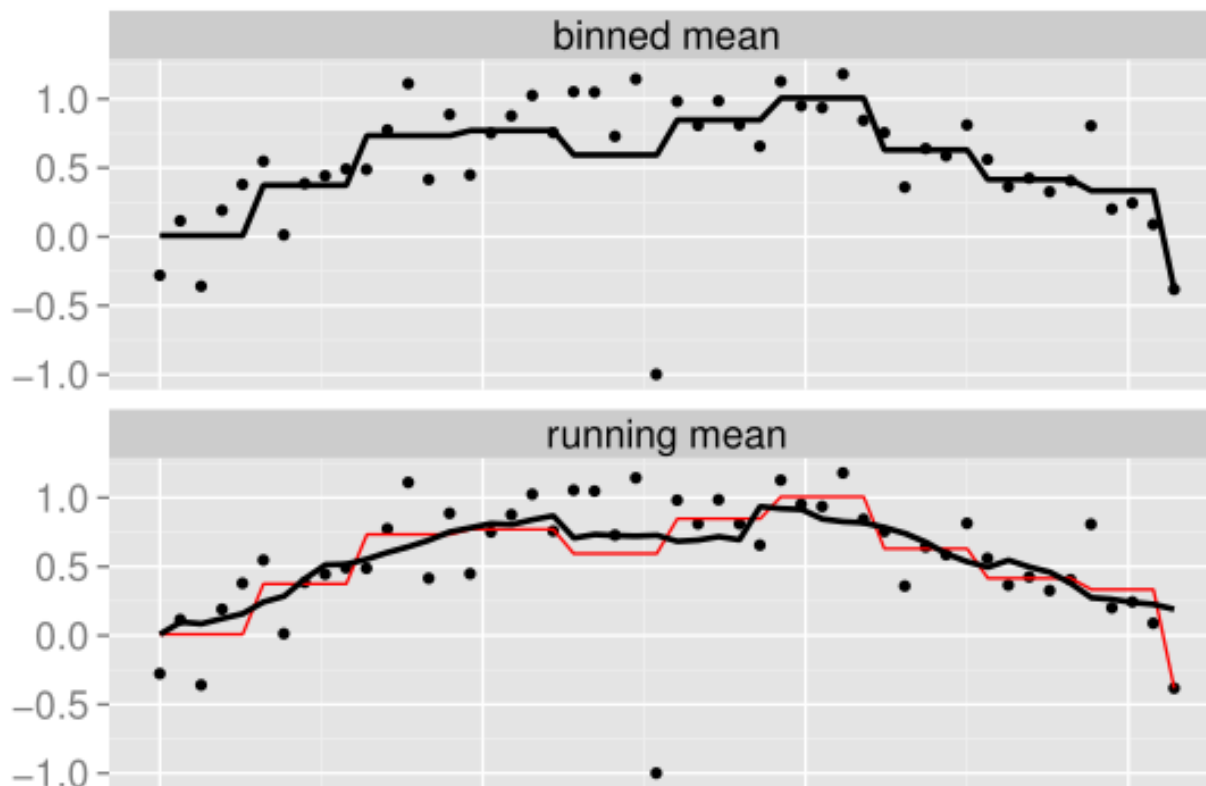


„Smooth”

- Túl kicsi a szélesség
- Inkább legyen gyors, mint robusztus

„Smooth”

- Túl kicsi a szélesség
- Inkább legyen gyors, mint robusztus



„Smooth”

- Kernel módszerek:
 - nemcsak szomszédok,
 - de súlyozás is
- j-edik bin közelítésénél az i-edik súlya:

$$k_i = K\left(\frac{x_j - x_i}{h}\right)$$

- h: „sávszélesség”
 - Szomszédság mérete

- K itt: „triweight”

$$K(x) = (1 - |x|^3)^2 I_{|x| < 1}$$

„Smoo

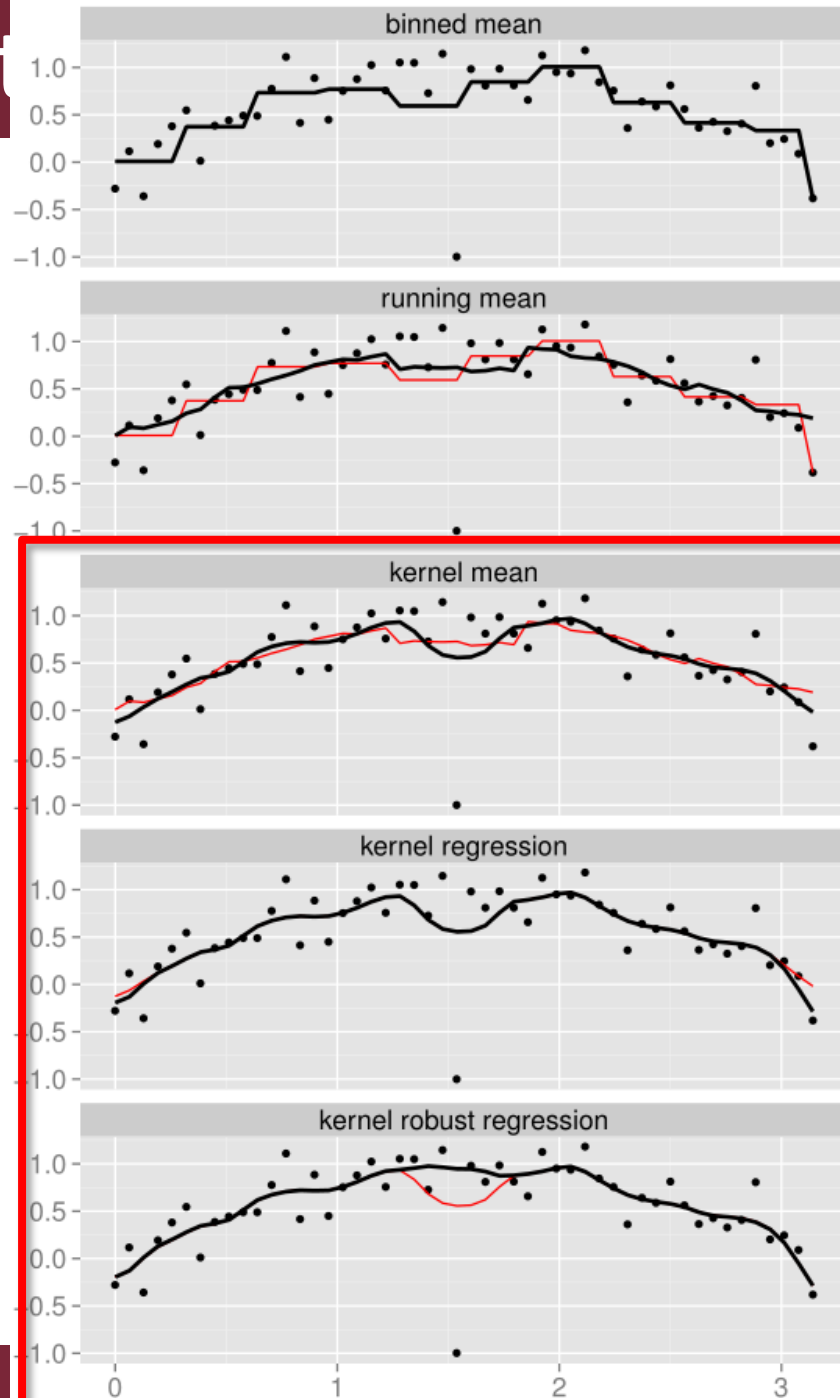
- Kernel módszerek:
 - nemcsak szomszédok,
 - de súlyozás is
- j-edik bin közelítésénél az i-edik súlya:

$$k_i = K\left(\frac{x_j - x_i}{h}\right)$$

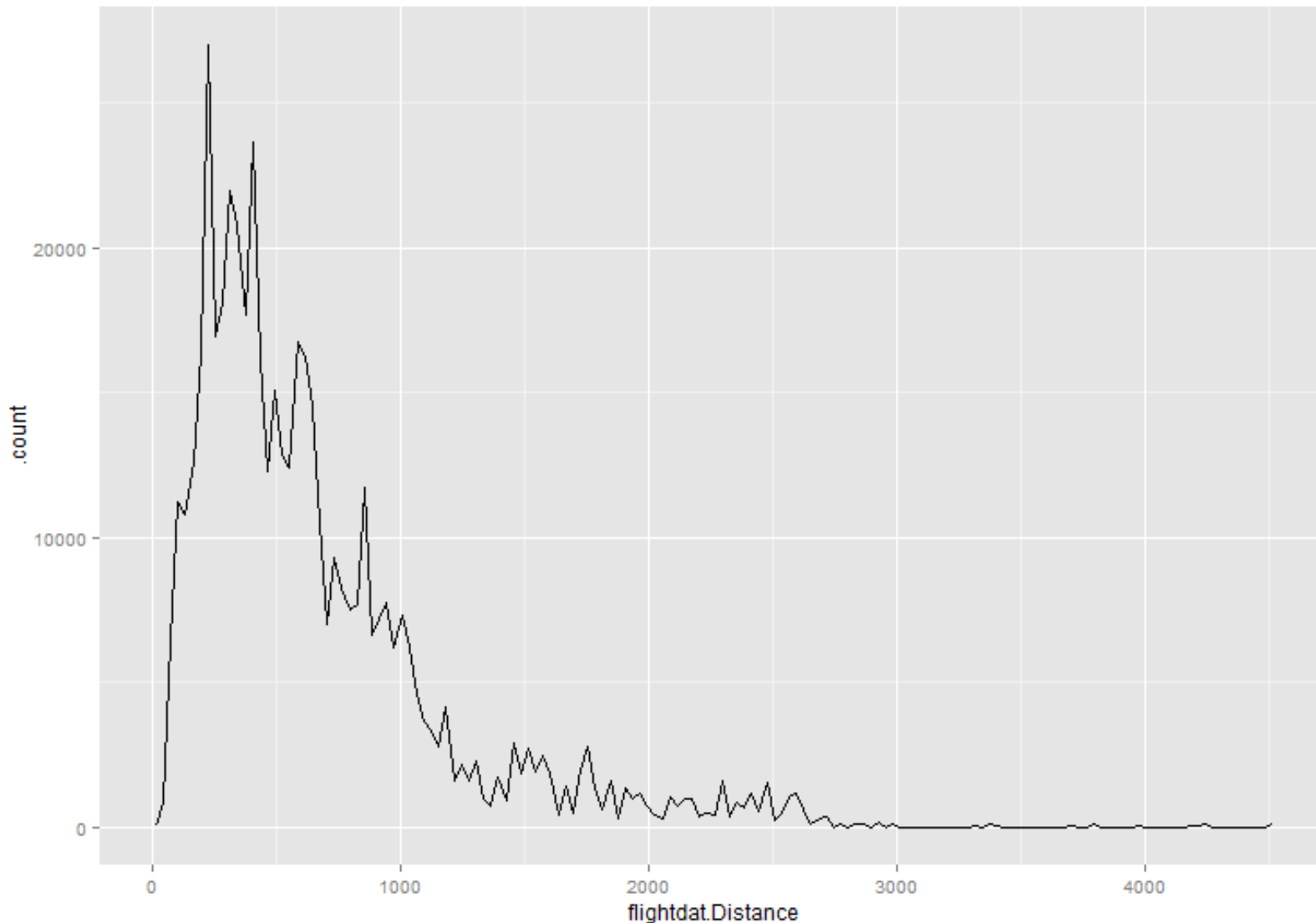
- h: „sávszélesség”
 - Szomszédság mérete

- K itt: „triweight”

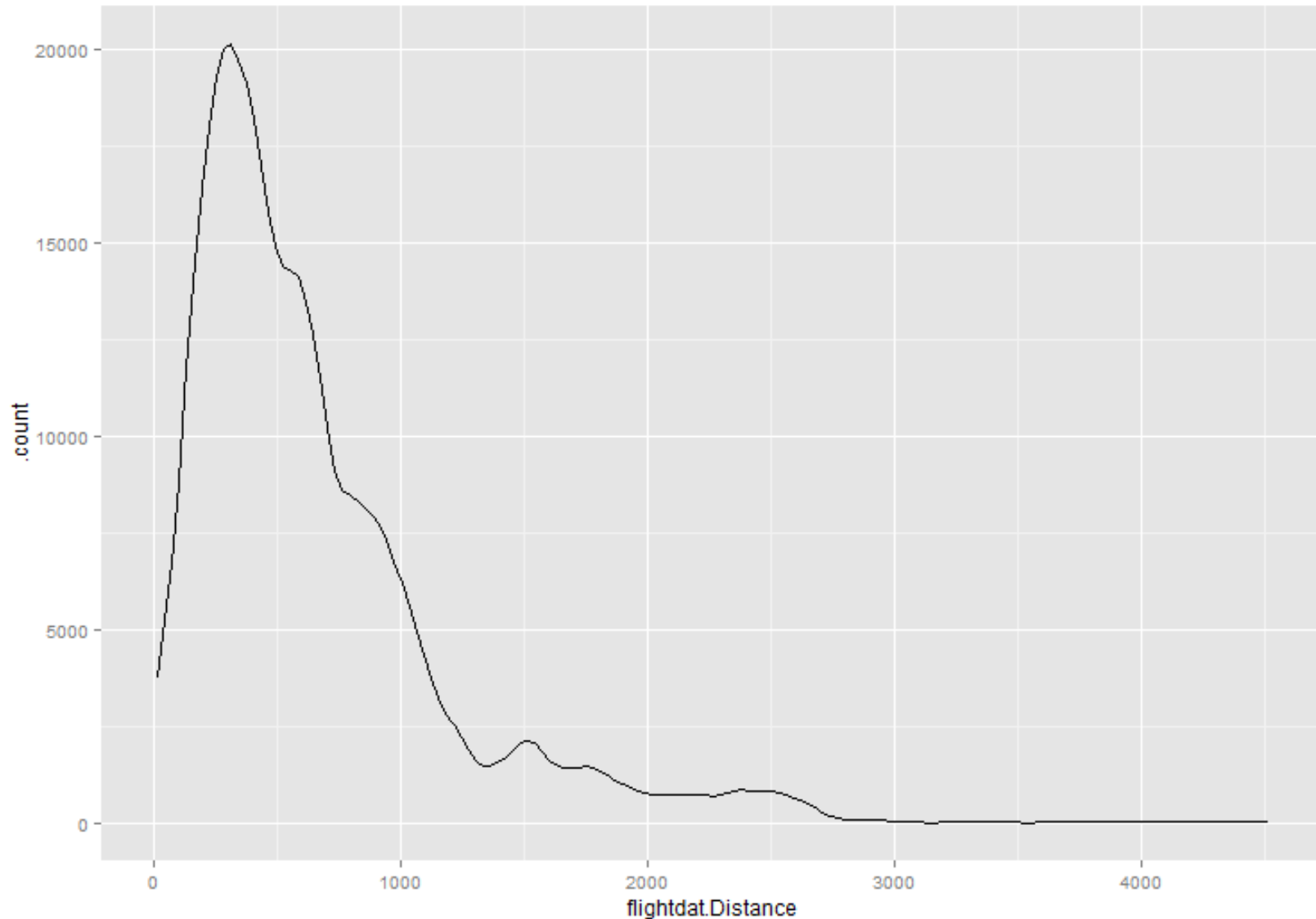
$$K(x) = (1 - |x|^3)^2 I_{|x| < 1}$$



Automatikus sáv szélesség választás?



Automatikus sáv szélesség választás?



Automatikus sávszélesség választás?

- Pl. „leave-one-out cross-validation” (LOOCV)
- aktuális statisztika és a simított összeg.
 - root mean squared error
 - $rmse = \sqrt{(y_i - \hat{y}_i)^2 / n}$
 - keressük a minimumhoz tartozó h -t

Két változó?

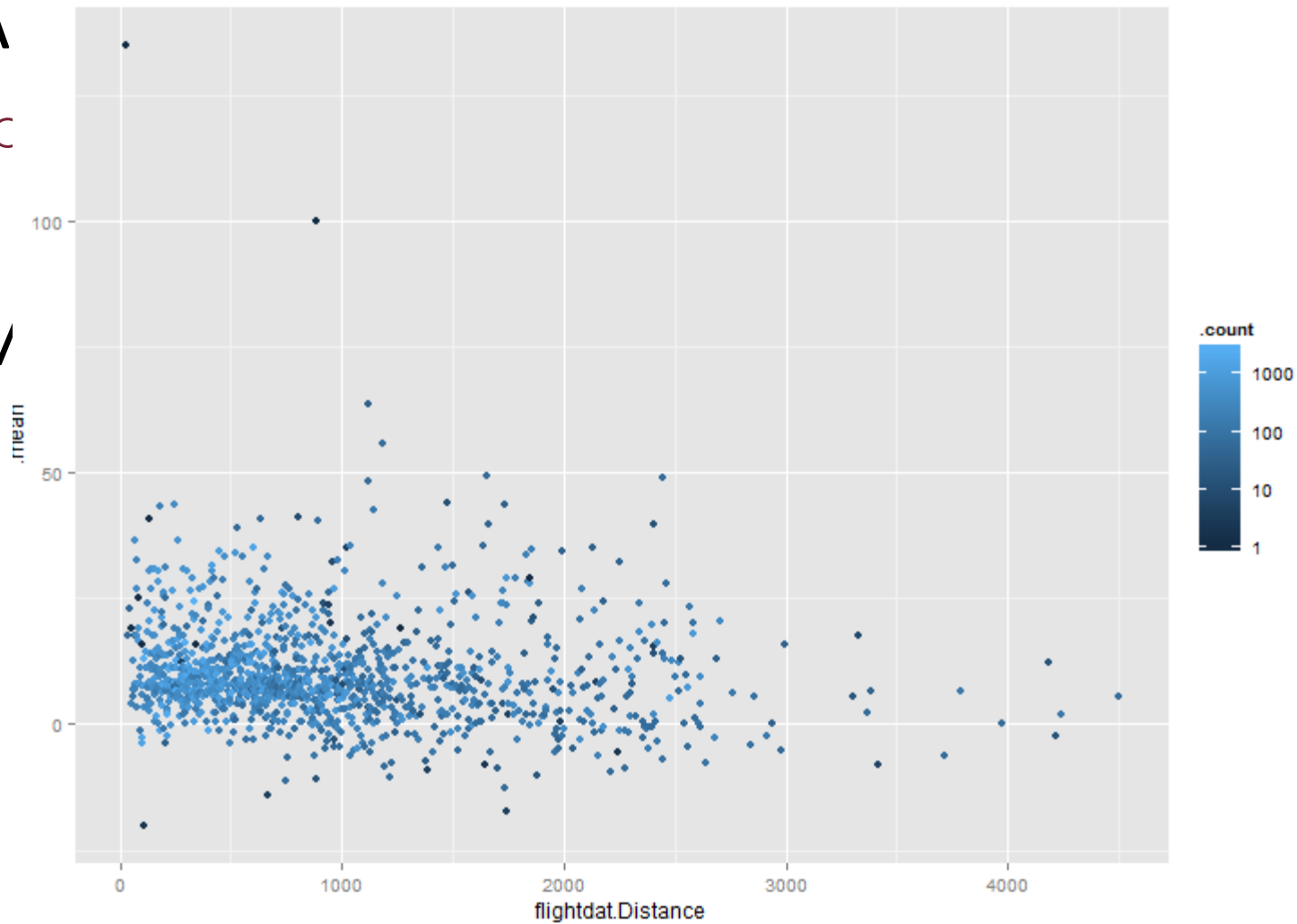
- Az egyik bin, a másik statisztika alapja
 - mean, median, std. dev.
- Mindkettő bin alapja, statisztika: „count”

Két változó?

■ A

C

■ N



```
autoplot(condense(bin(flightdat$Distance), z=flightdat$ArrDelay, summary="mean"))
```

Két változó?

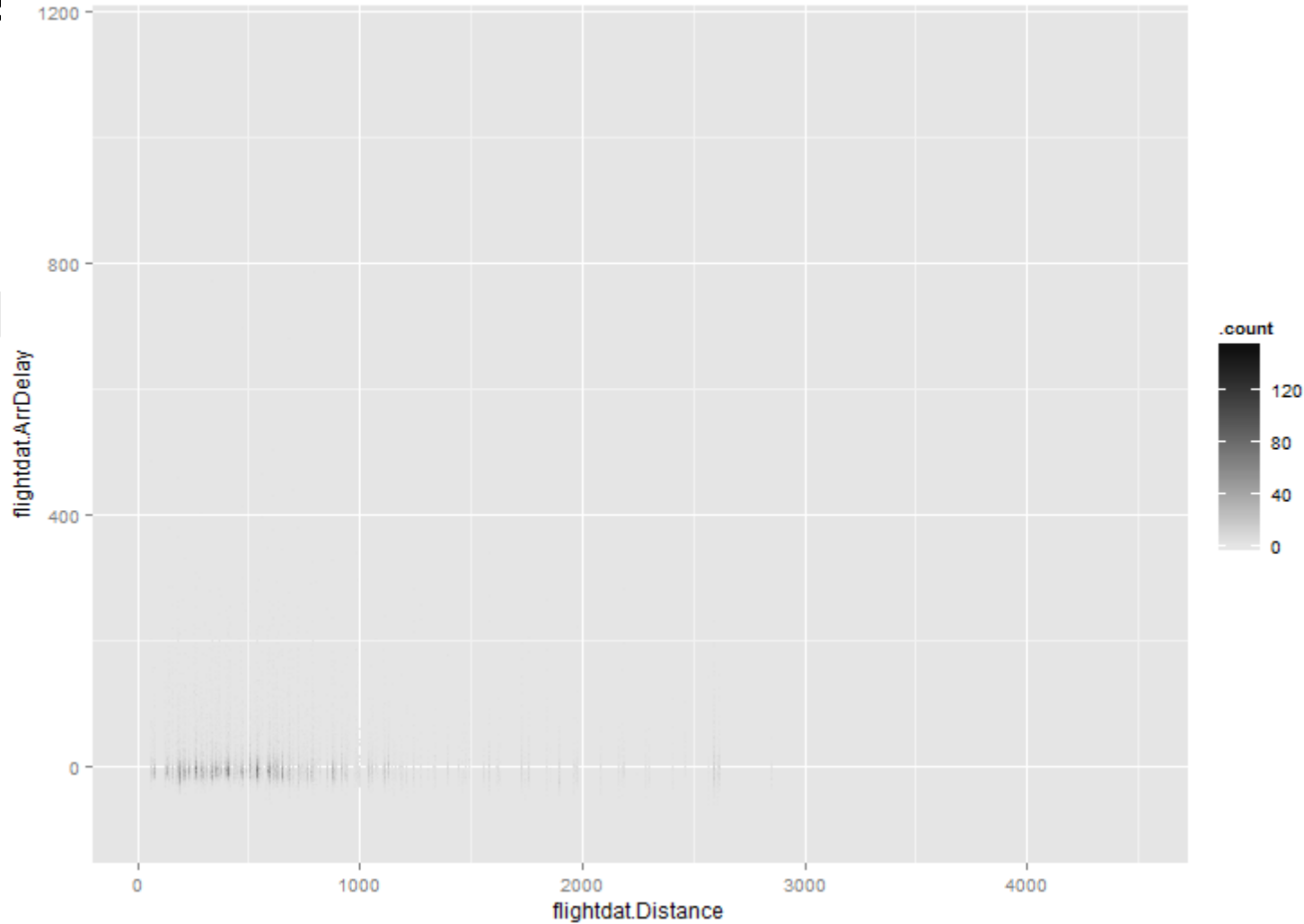
- Az egyik bin, a másik statisztika alapja
 - mean, median, std. dev.
- Mindkettő bin alapja, statisztika: „count”

Két változó?

- Az egyik bin a másik statisztika eleme



- M

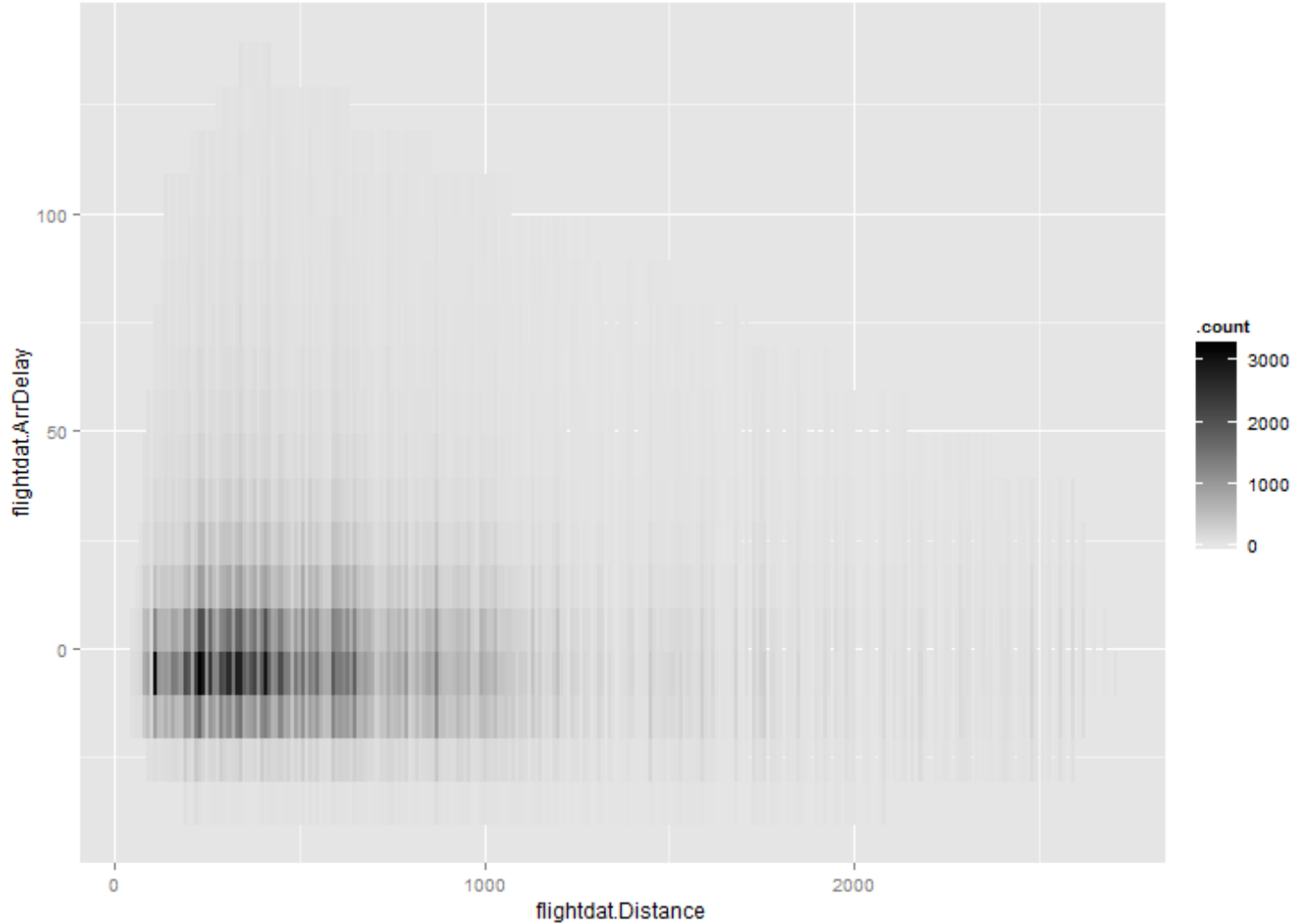


Két változó?

- Az egyik bin a másik statisztika eleme

○

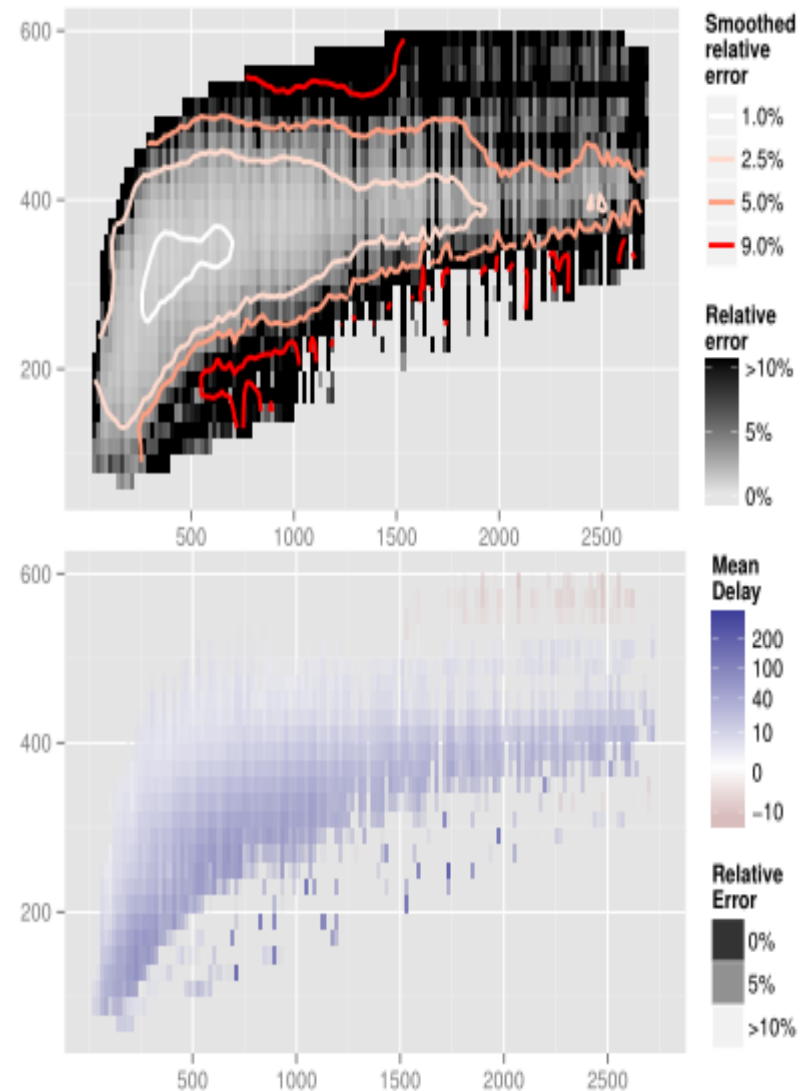
- M



Vizualizáció

- (2,1)-d plot:
heatmap/tile plot,
contour plot
- (n,m)-d plot:
 - „small multiples”
(faceting)
 - Interakció

Ábra forrása: [1]



Datashader

- Datashader:
 - <https://github.com/bokeh/datashader>
 - Erősen javasolt conda csomagkezelő segítségével telepíteni
- Futtatott plotting_pitfalls, nyc_taxi és census példák:
 - <https://github.com/bokeh/datashader/tree/master/examples>
- A megnézett példák (statikusan) itt:
 - <https://anaconda.org/jbednar/notebooks>