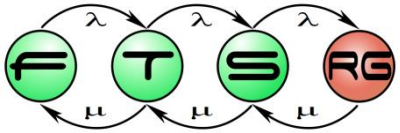
Vizuális adatanalízis

„Big Data” elemzési módszerek

Kocsis Imre, Salánki Ágnes, Gönczy Lászó
ikocsis, salanki@mit.bme.hu
2015.10.21.



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, …

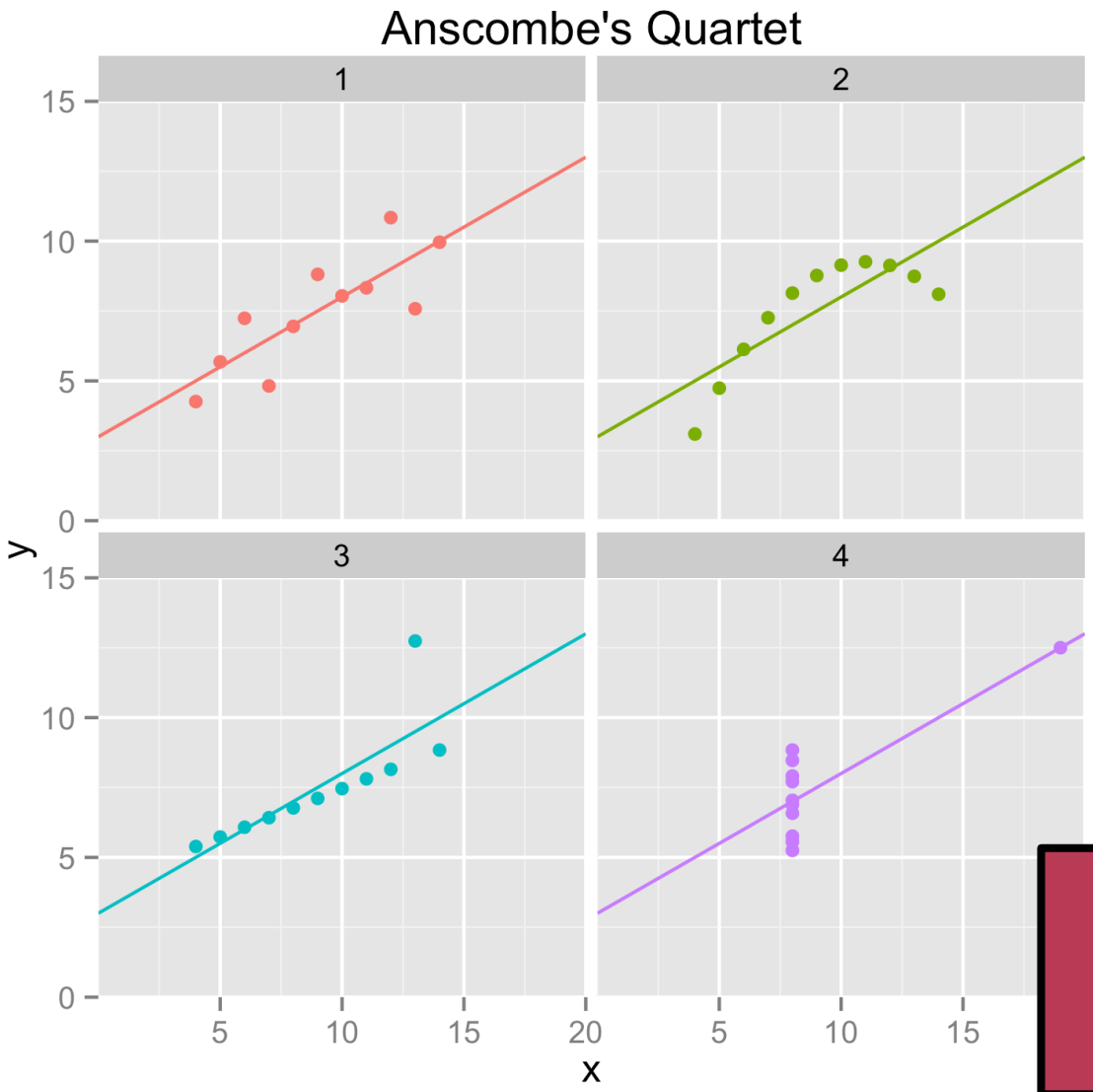
- Legismertebb vizionáriusa: John W. Tukey

[2] és [3] alapján

EDA

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

Anscombe négyese



Hibás feltételezések elkerülése… és intuíció:

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.”

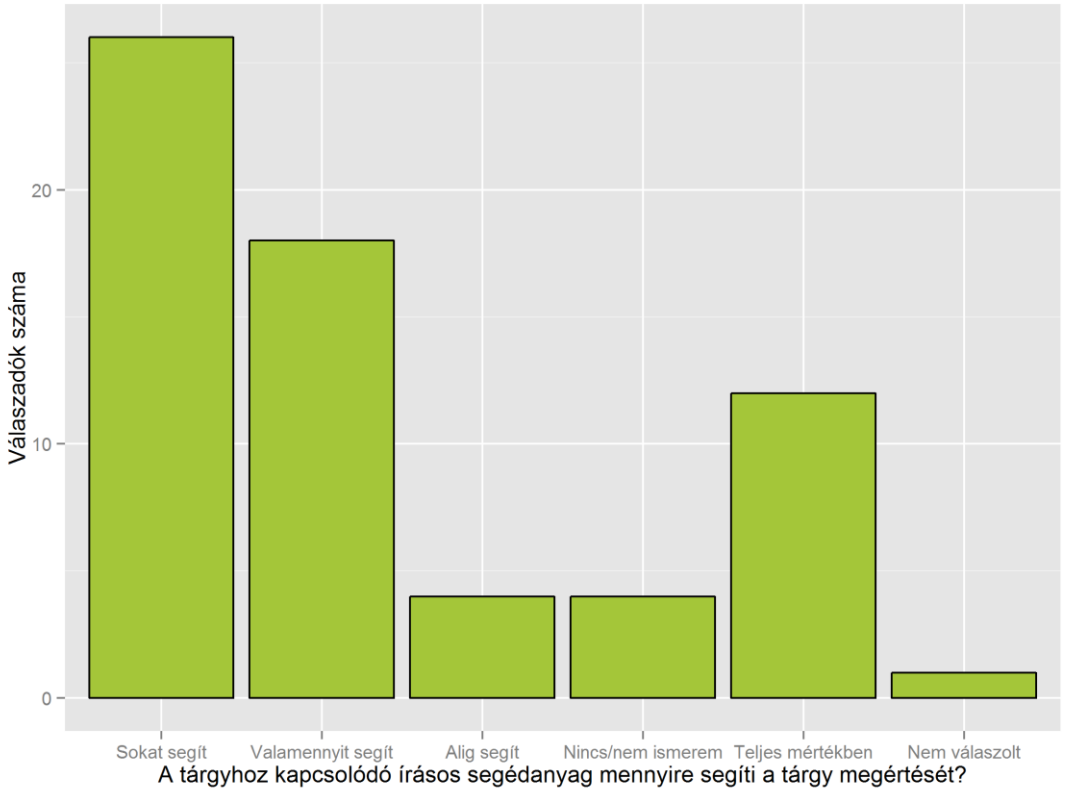
Prof. Alfred Inselberg

Forrás: [5] és [6]

Oszlopdiagram (bar chart)

- Megjelenített dimenziók száma: 1

- Ábrázolt összefügg.:
  - Diszkrét változó egyes értékeinek abszolút gyakorisága
- Adategység:
  - Oszlop – az oszlop mag az adott érték abs. gya tükrözi
- Tervezői döntés:
  - Csoportok kialakítása?
  - Értékkészlet darabolása?



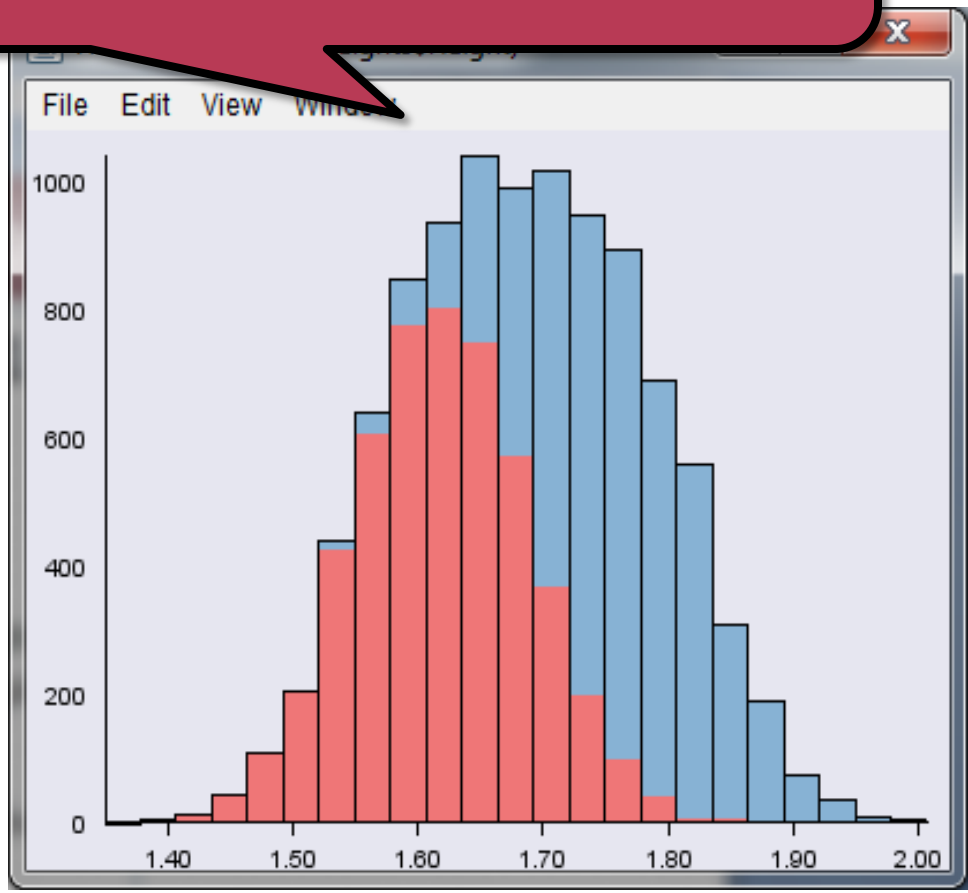
Hisztogram

- Megjelenített dim.k: 1
- Ábrázolt összefügg.:
  - folytonos változó eloszlása
- Adategység:
  - Oszlop – az oszlop magassága az adott érték abs. gyakoriságát tükrözi

Fontos percentilisek?

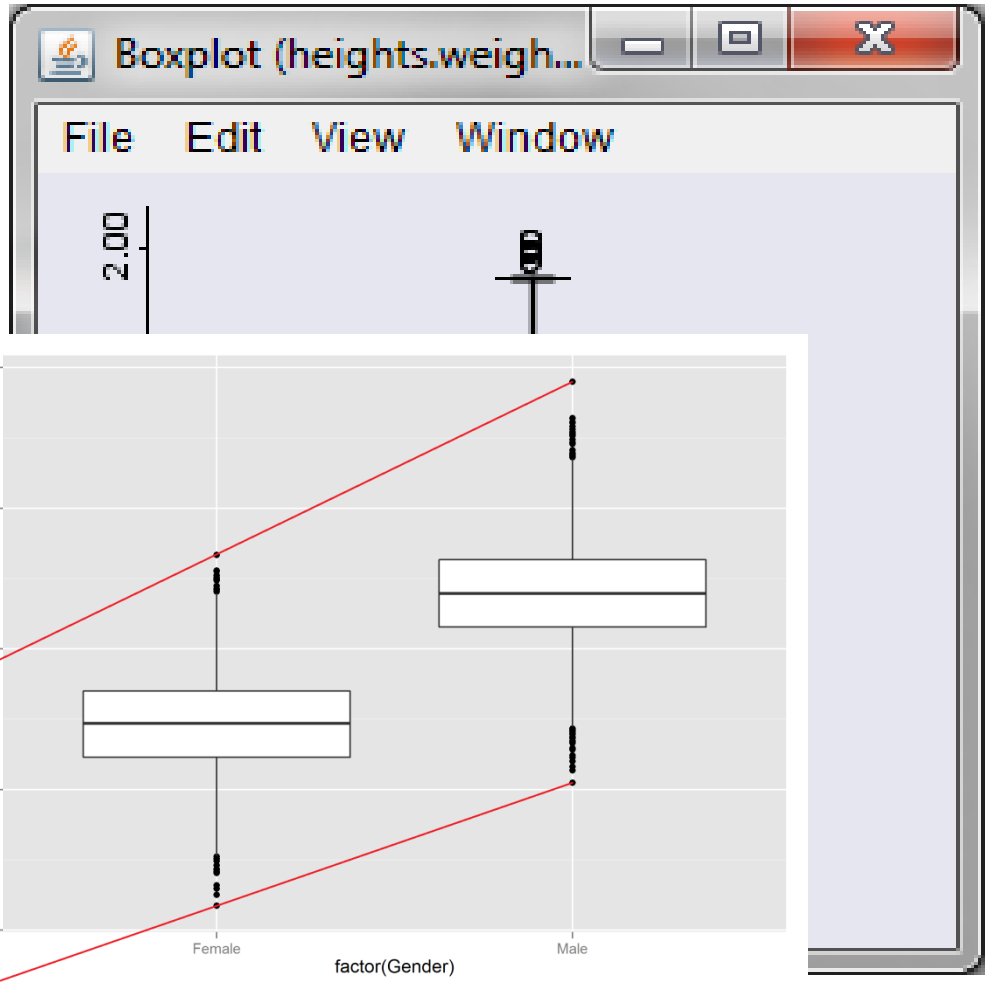
- Tervezői döntés:
  - Oszlopok szélessége?

Nők és férfiak magasságának eloszlása is szép haranggörbe

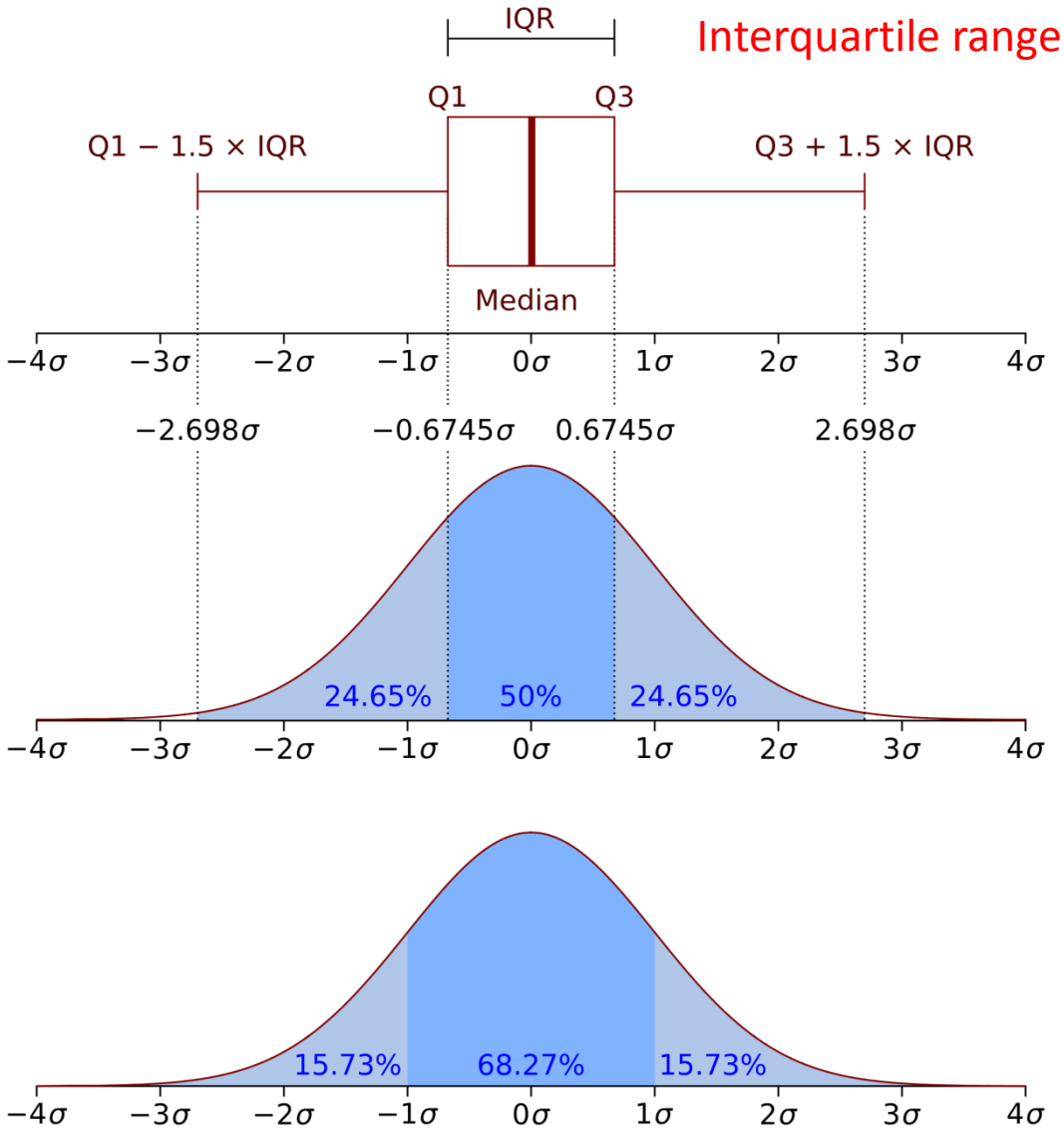


Doboz diagram (boxplot)

- Megjelenített dim.k: 1
- 5 értékkel jellemzésként
- Ábrázolt összefügg.:
  - folytonos változó fontos percentilis
- Adategység:
  - Doboz – szélei jelzik az alsó és felső kvartiliseket,
  - Középen a medián.
  - A minimum és a maximum általában még pontosan jelezve,
  - Outlierek már csak pöttyökkel.



Boxplot


IQR
Interquartile range
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%
15.73%
68.27%
15.73%

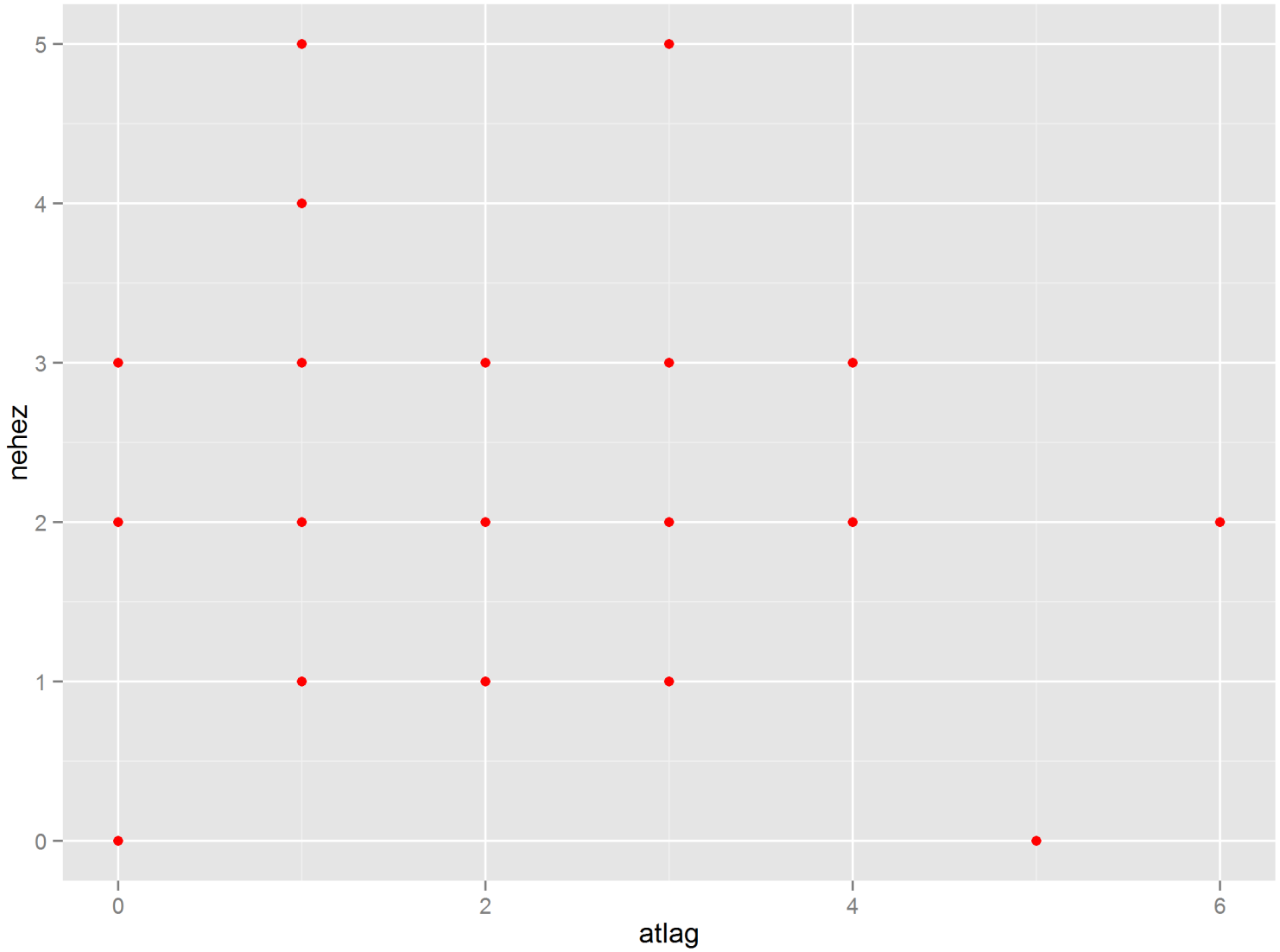
KÉT VÁLTOZÓ

Cél: tartományok, összefüggések keresése

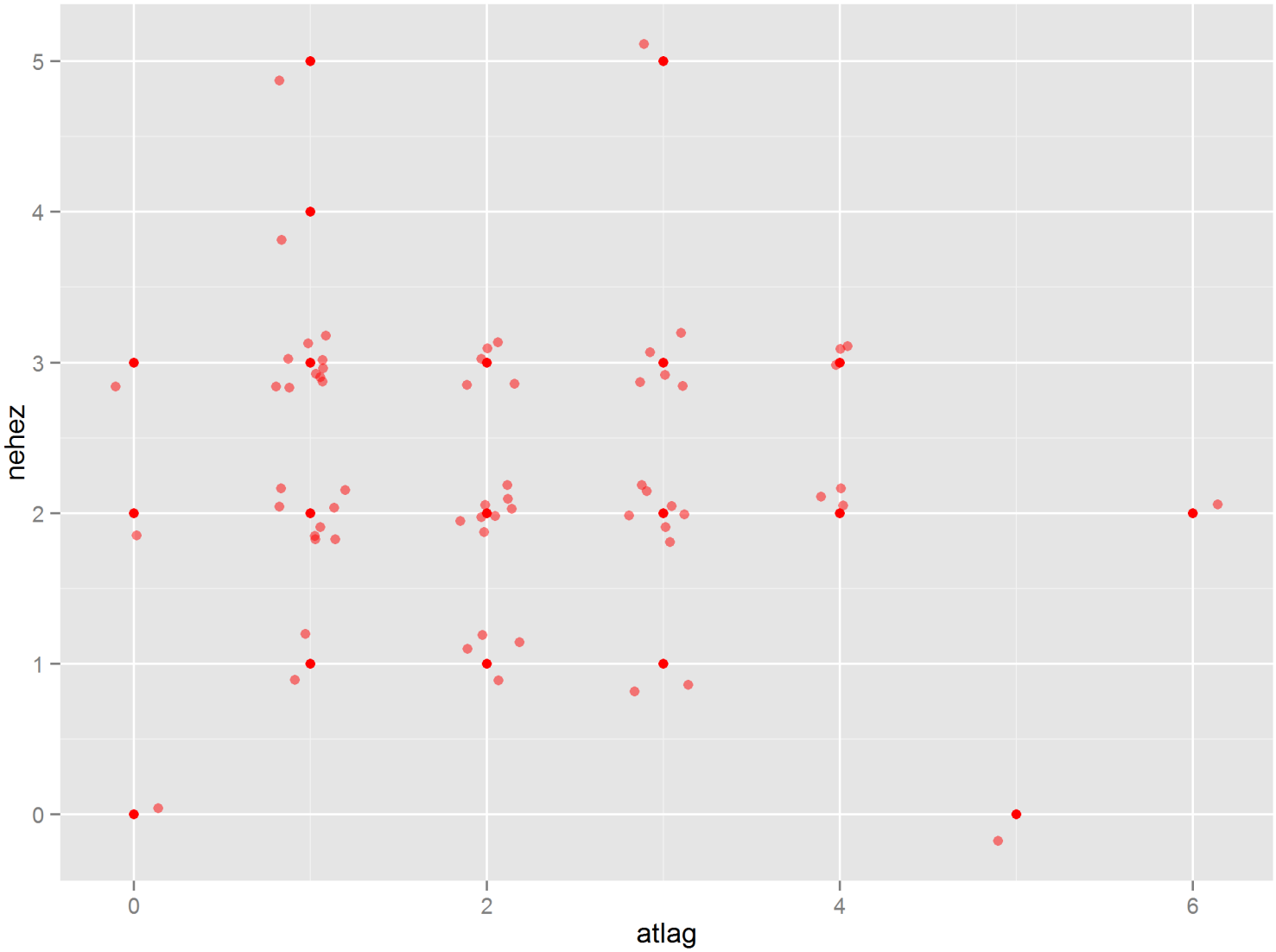
Pont – pont diagram (scatterplot)

- Megjelenített dim.k: 2
- Ábrázolt összefügg.:
  - 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?

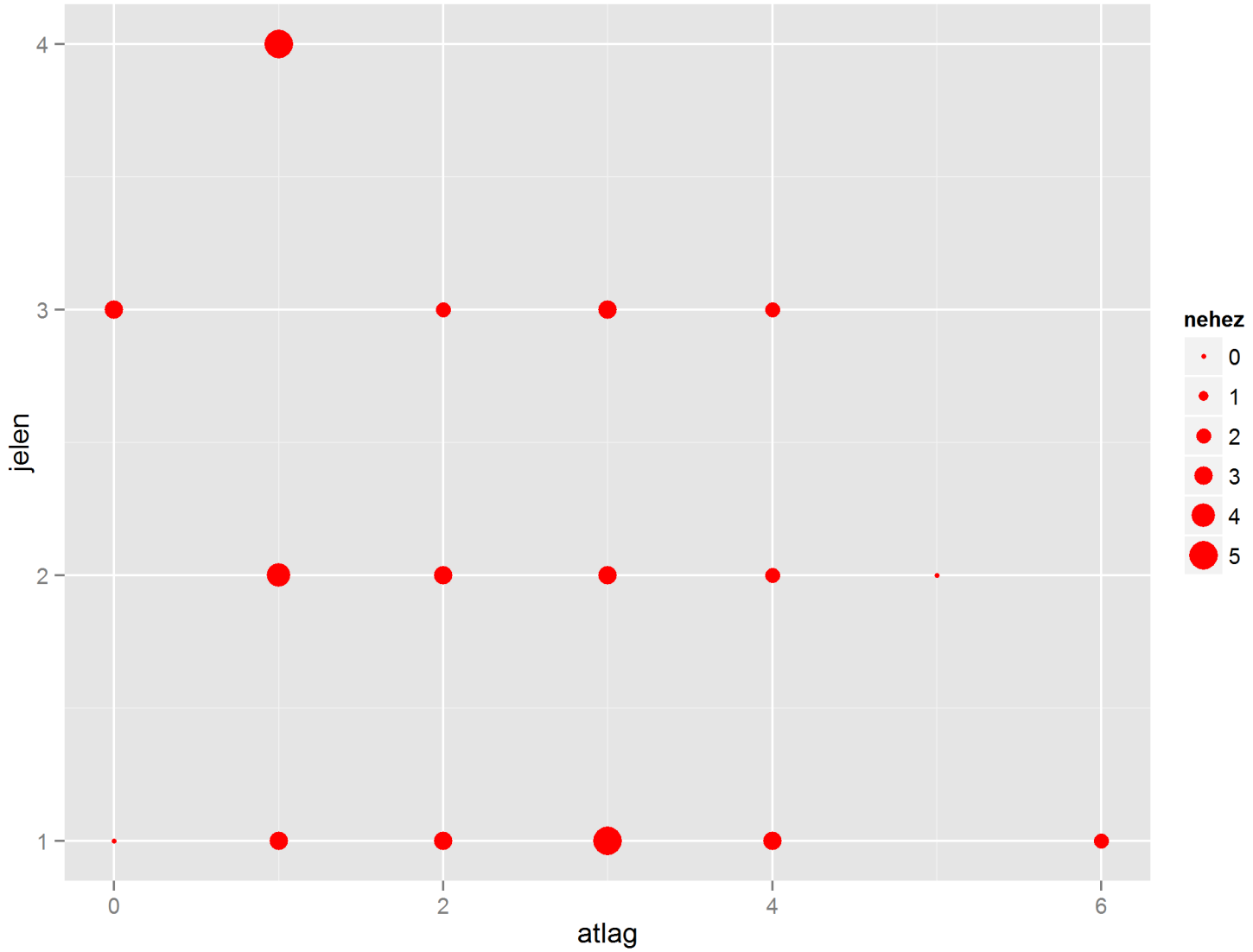
Hol volt, hol nem volt...



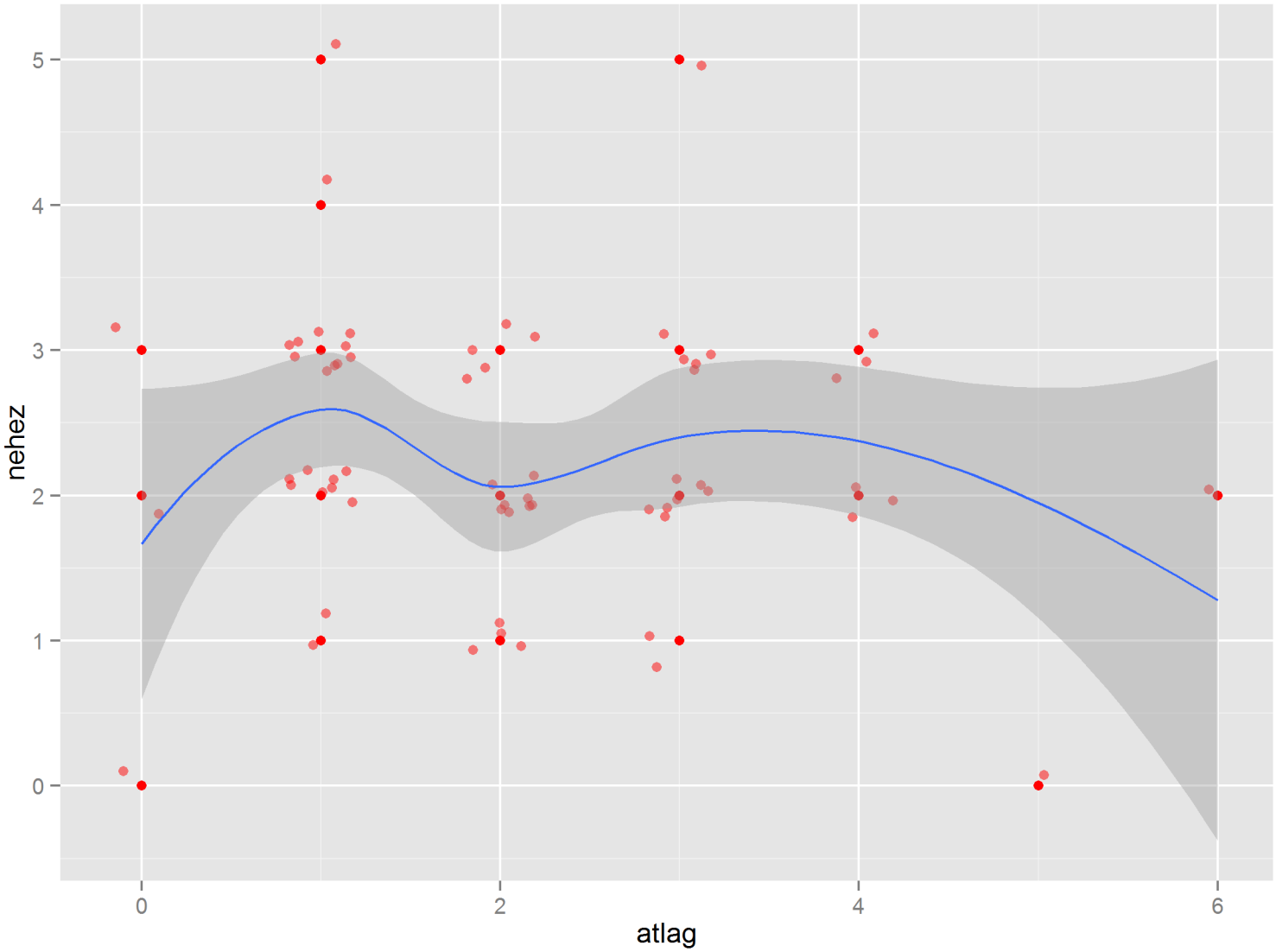
Szétszórjuk



A pontok....



És megpróbáljuk közelíteni...



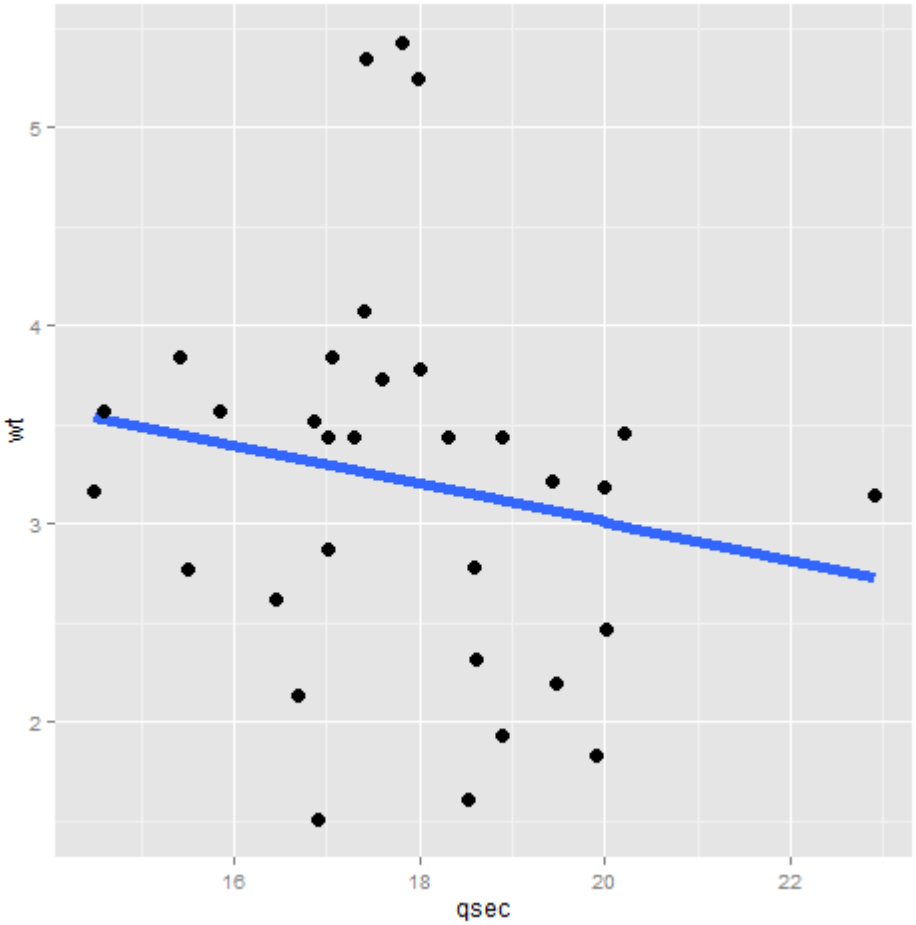
Simító görbe (*smoothing spline*) [11,12]

$$SS^*(h) = \left[\left(\sum_{i=1}^{n}(Y_i - \mu(x_i))^2\right) + \lambda \int_{x_{min}}^{x_{max}} \mu''(x)^2 dx\right]$$

- „*Penalized sum of squares*”
- Feladat: minimalizáló $\hat{\mu}(x)$ függvény megtalálása
- Első tag: maradvány-hibanégyzetösszeg
- Második tag: „*roughness penalty*”
  - Minél gyorsabban nő a meredekség, annál nagyobb
- Megoldása köbös (cubic) spline

Simító görbe (*smoothing spline*) [11,12]

- λ simító paraméter
  - Adat követése $(\sum_{i=1}^{n}(Y_i - \mu(x_i))^2)$
  - Simaság $\int_{x_1}^{x_n} \mu"(x)^2 dx$
  - λ =0 esetén interpolációs görbe
  - $\lambda \rightarrow \infty$ esetében lineáris regresszió (ill. *linear least squares estimate*)


wt
5
4
3
2
16
18
20
22
qsec

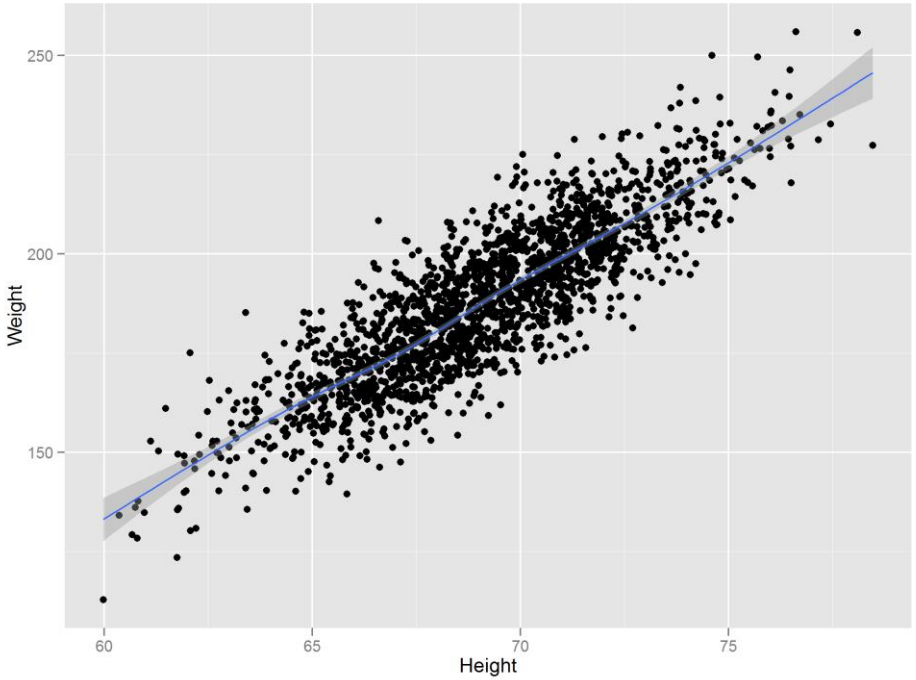
Regresszió [12]

- Cél:

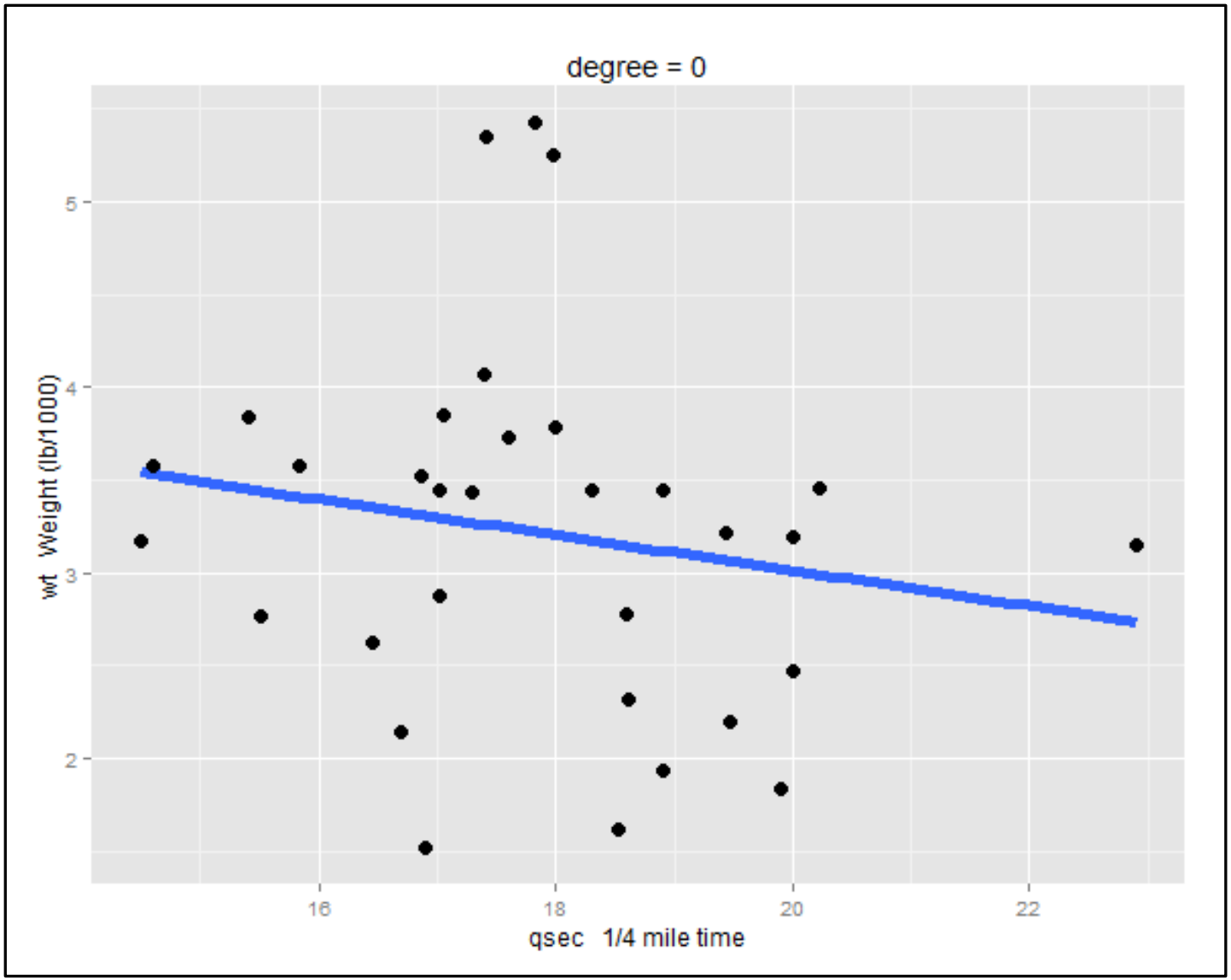
  megtalálni egy olyan f függvényt, amelynek inputja az attribútumok értéke, az outputja pedig a lehető legjobban közelíti (négyzetes hibaérték) a valóságot

- Példa:

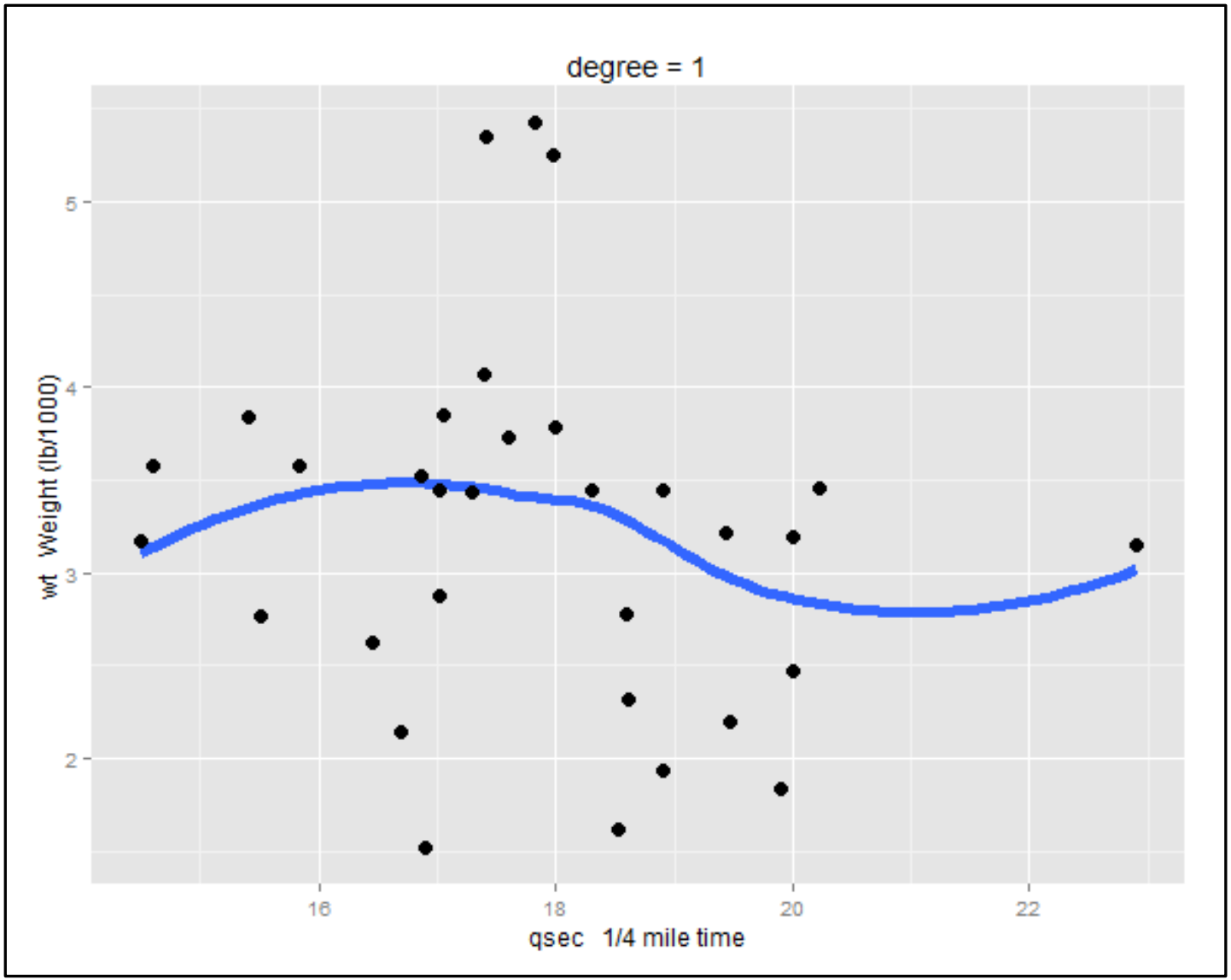
  - testtömeg/magasság együtte eloszlás valójában egyenesre illeszthető,
  - web forgalom jóslása



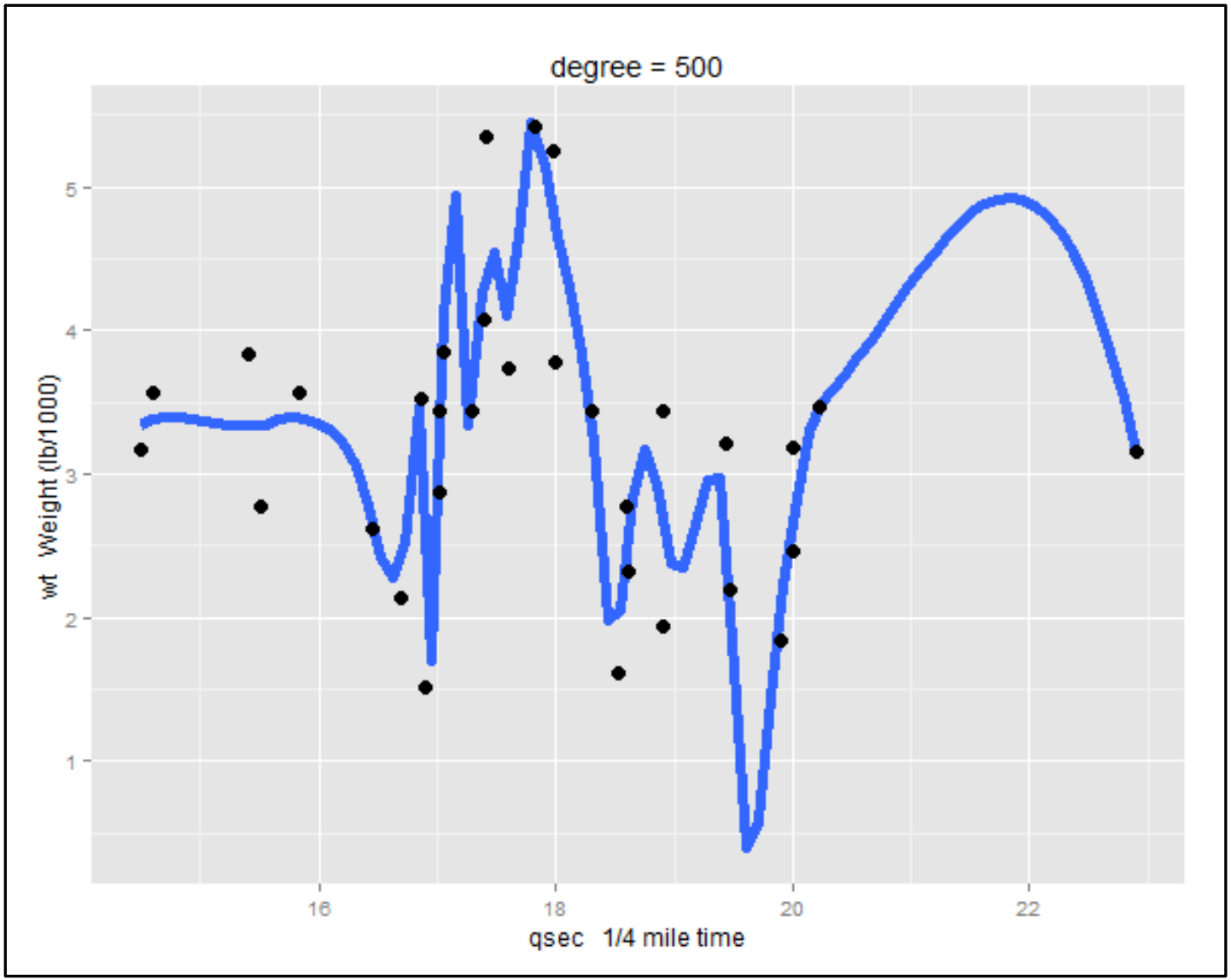
Lokális p-rendű LS polinomiális regresszió



Lokális p-rendű LS polinomiális regresszió

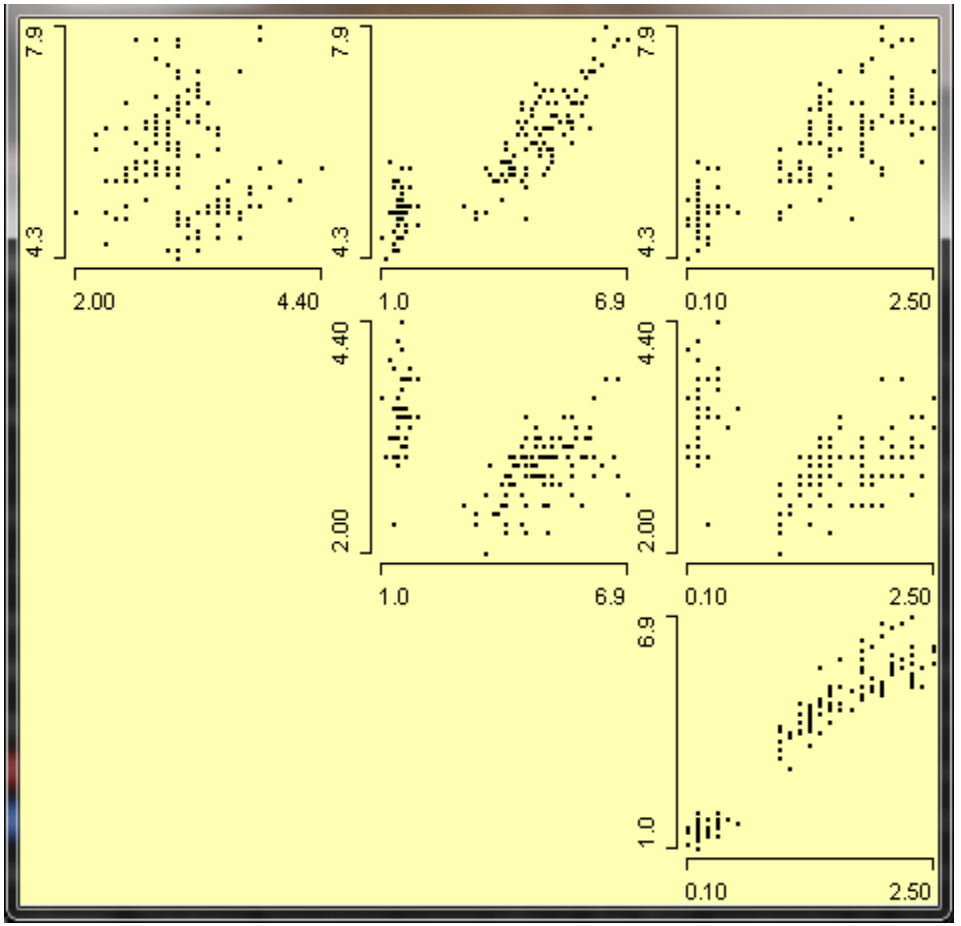


Lokális p-rendű LS polinomiális regresszió



Scatterplot mátrix

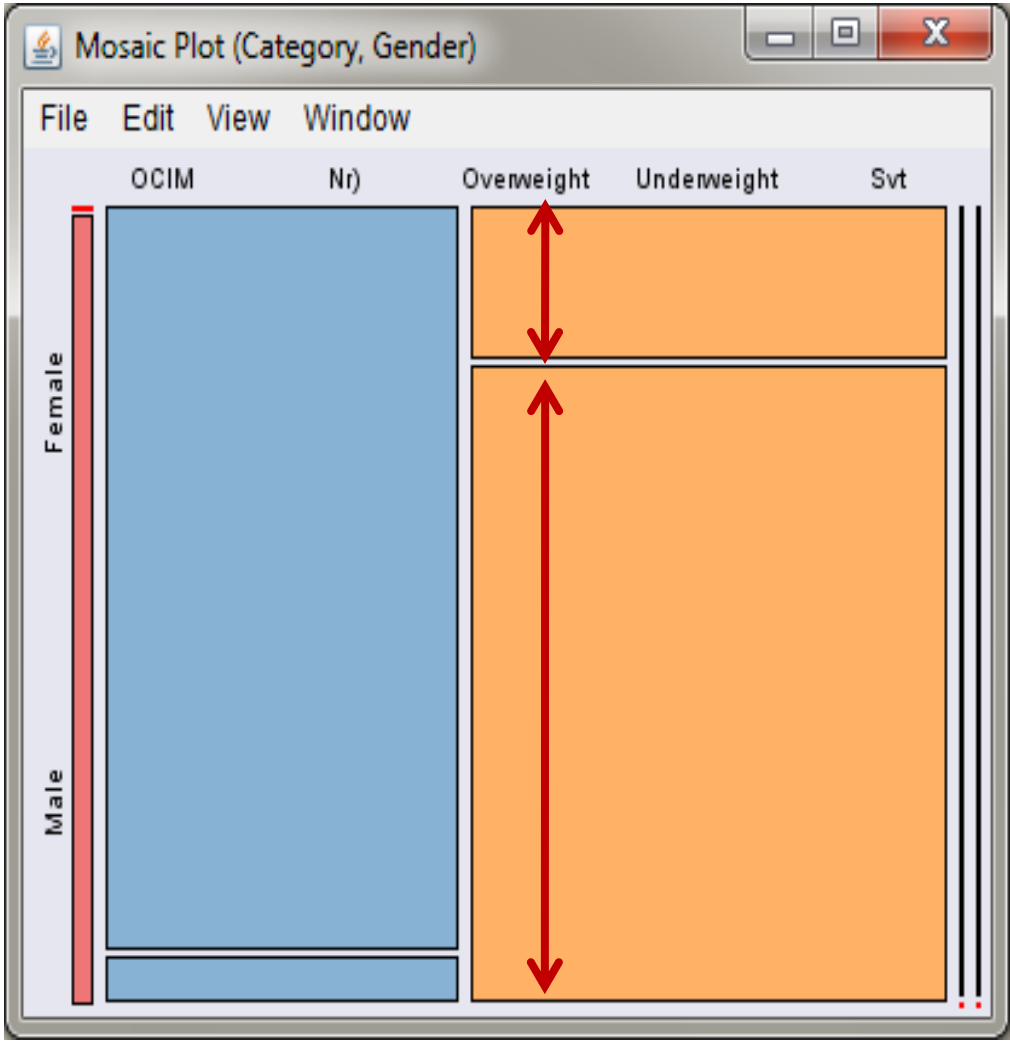
- Megjelenített dim.k: n
- Ábrázolt összefügg.:
  - A változópárok együttes eloszlása
- Adategység:
  - Scatterplot – minden diagram a neki megfelelő változók együttes eloszlását mutatja be



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

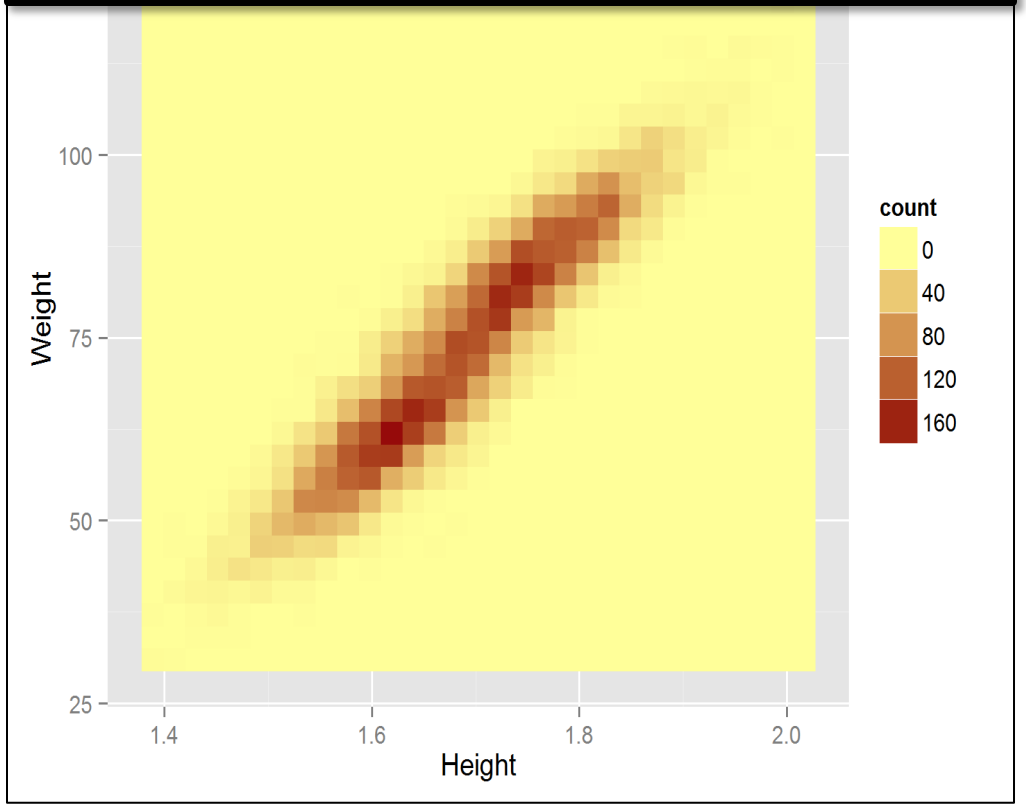
A túlsúlyosak nagy része férfi!



Hőtérkép (heat map)

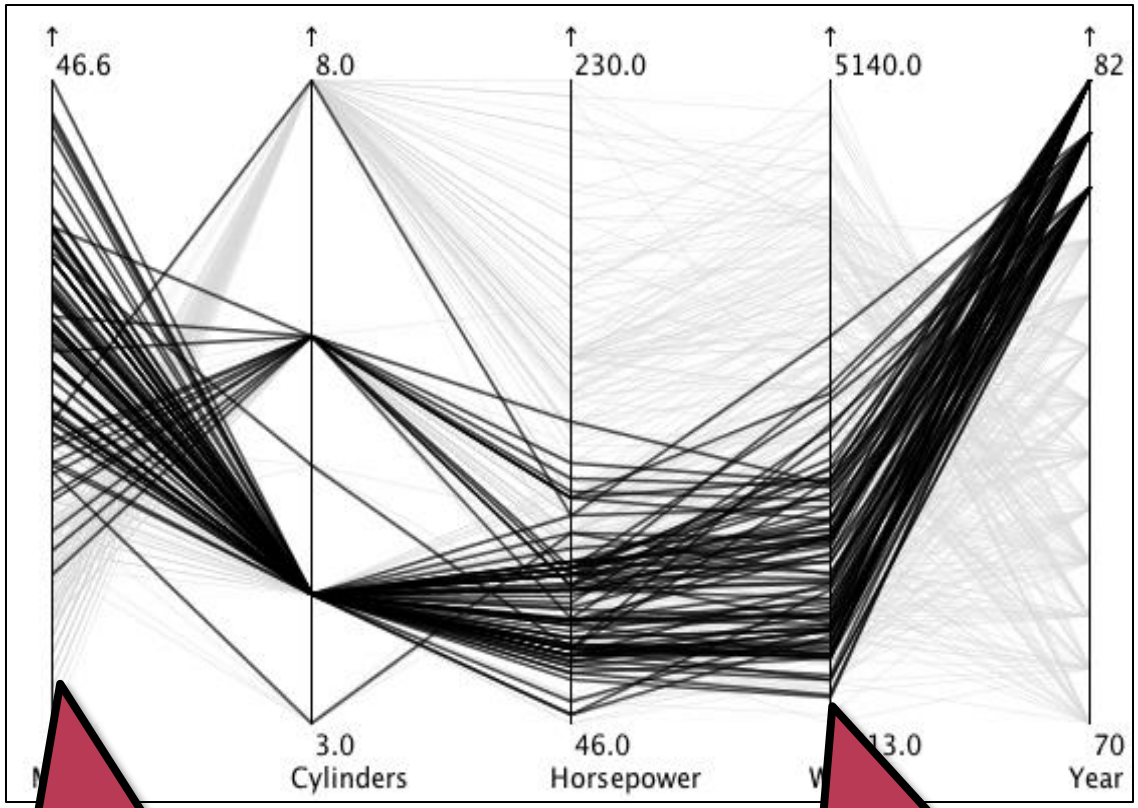
- Megjelenített dim.k: 3
- Ábrázolt összefügg.:
  - *sűrű* 3D struktúrák összefüggései
- Adategység:
  - tile – azonos „magasságú” összefüggő területrész
- Tervezői döntés:
  - tile-ok mérete?

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



Párhuzamos koordináták

- Megjelenített dim.k: n
- Ábrázolt összefügg.:
  - Rekordok/attribútumok hasonlósága
- Adategység:
  - Törött vonal – az egyes attribútumtengelyeken felvett értékek rendezett sorozata
- Korlátok:
  - Tengelyek (attribútumok) más mértékegysége/nagyságrendje stb torzíthat

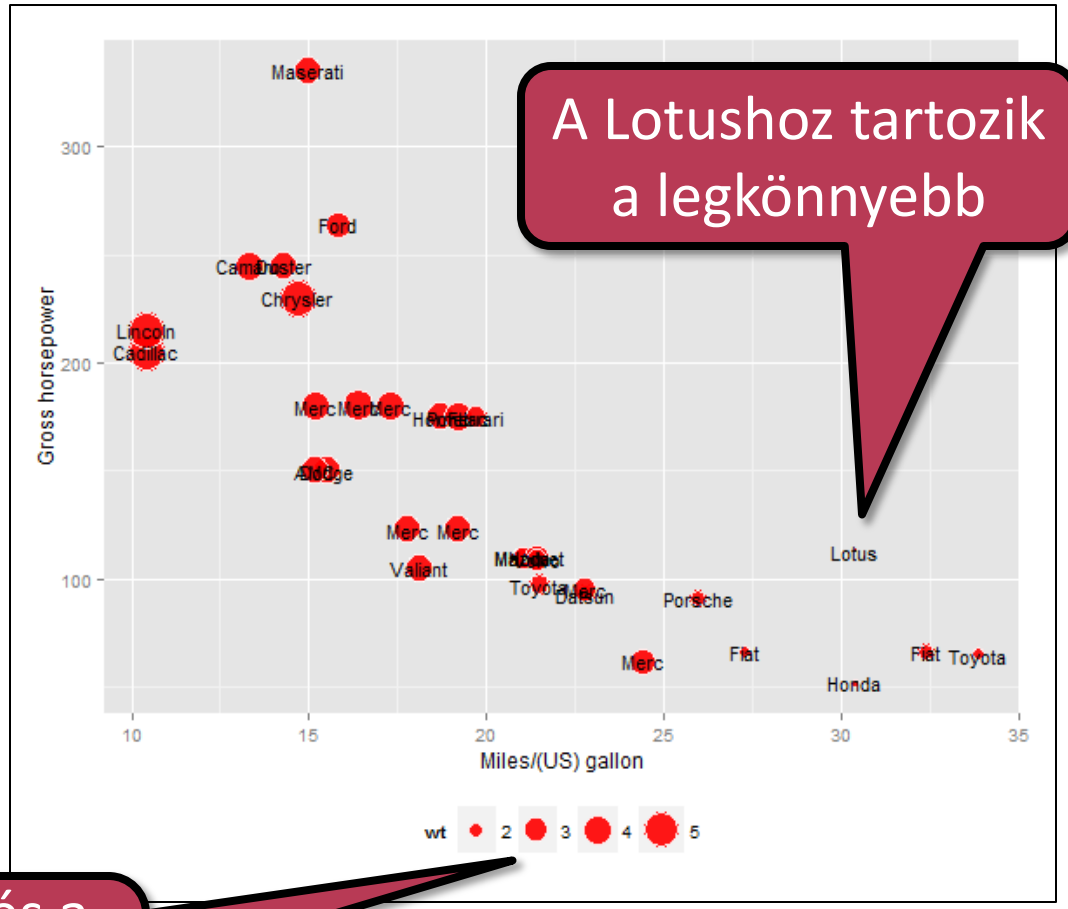


… és a fogyasztás is ☺

Az új autókban a tömeg kisebb…

Buborék diagram (bubble chart)

- Megjelenített dim.k: 3
- Ábrázolt összefügg.:
  - *ritka* 3D struktúrák összefüggései
- Adategység:
  - körlap – 3 attribútummal leírható: X és Y koordináta a középpontra + sugár
- Korlátok
  - overplotting torzíthat (ha a ritka struktúrában vannak sűrű részek)



A Lotushoz tartozik a legkönnyebb

Az X, Y pozíciót a fogyasztás és a teljesítmény adja, a kör sugara a tömeget mutatja

Interaktív statisztikai grafika

Vezetett adatbejárás – „data tour”

Lekérdezések

Kijelölés és csatolt kiemelés

Csatolt analízisek

Interakció az ábrákkal

Ábrák képzése – „plotolás”

[7] alapján

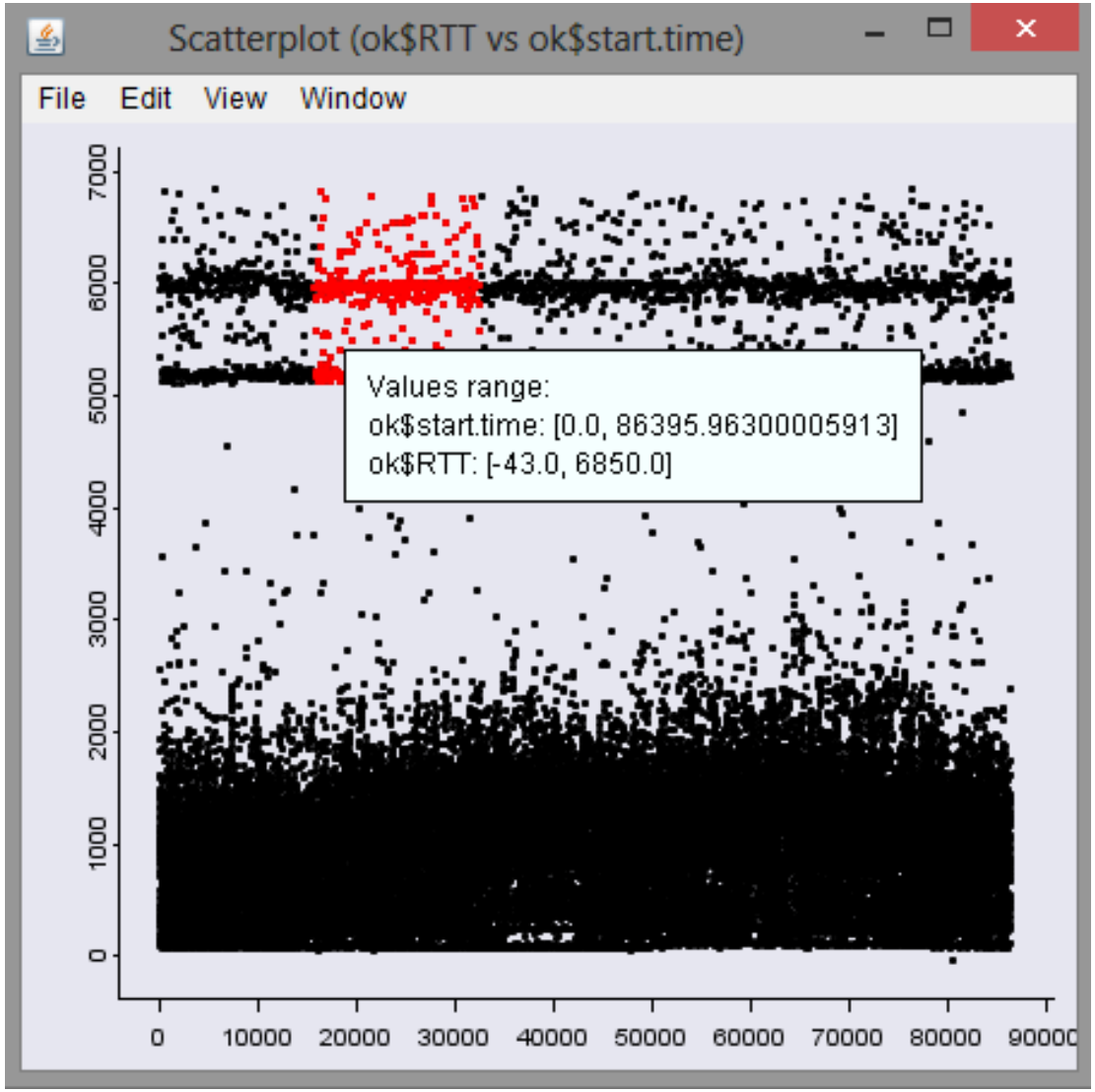
iPlots

- Interaktív statisztikai grafika R-ben
- 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~~

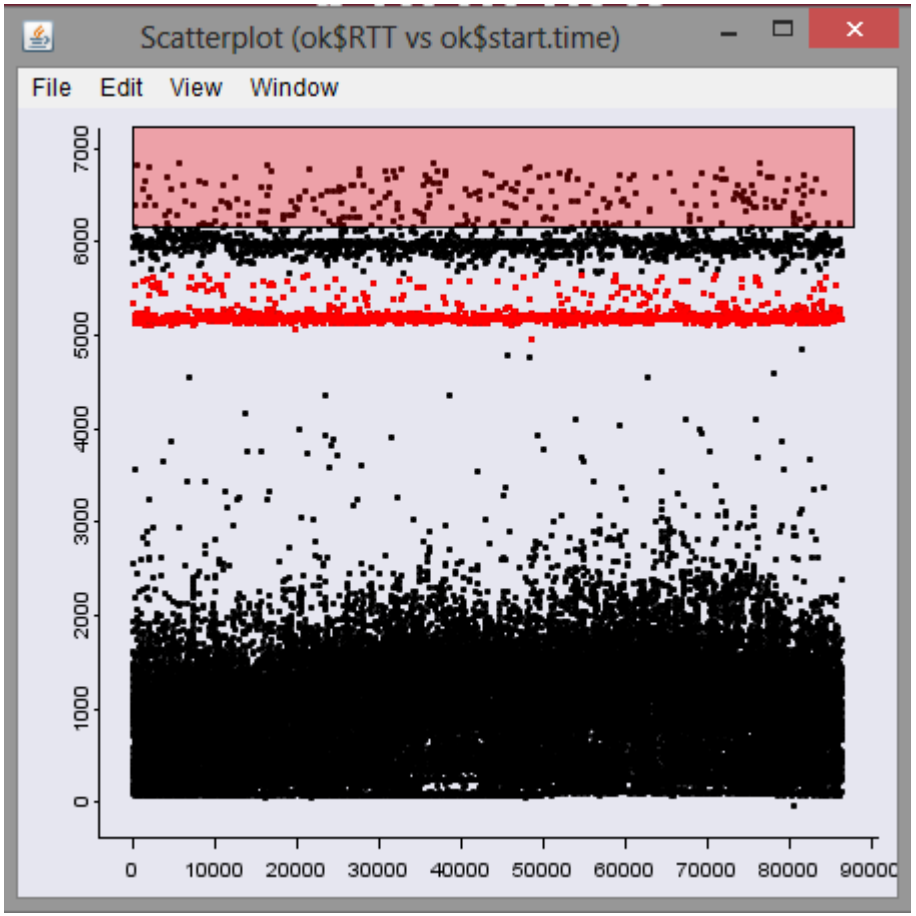


Kijelölés

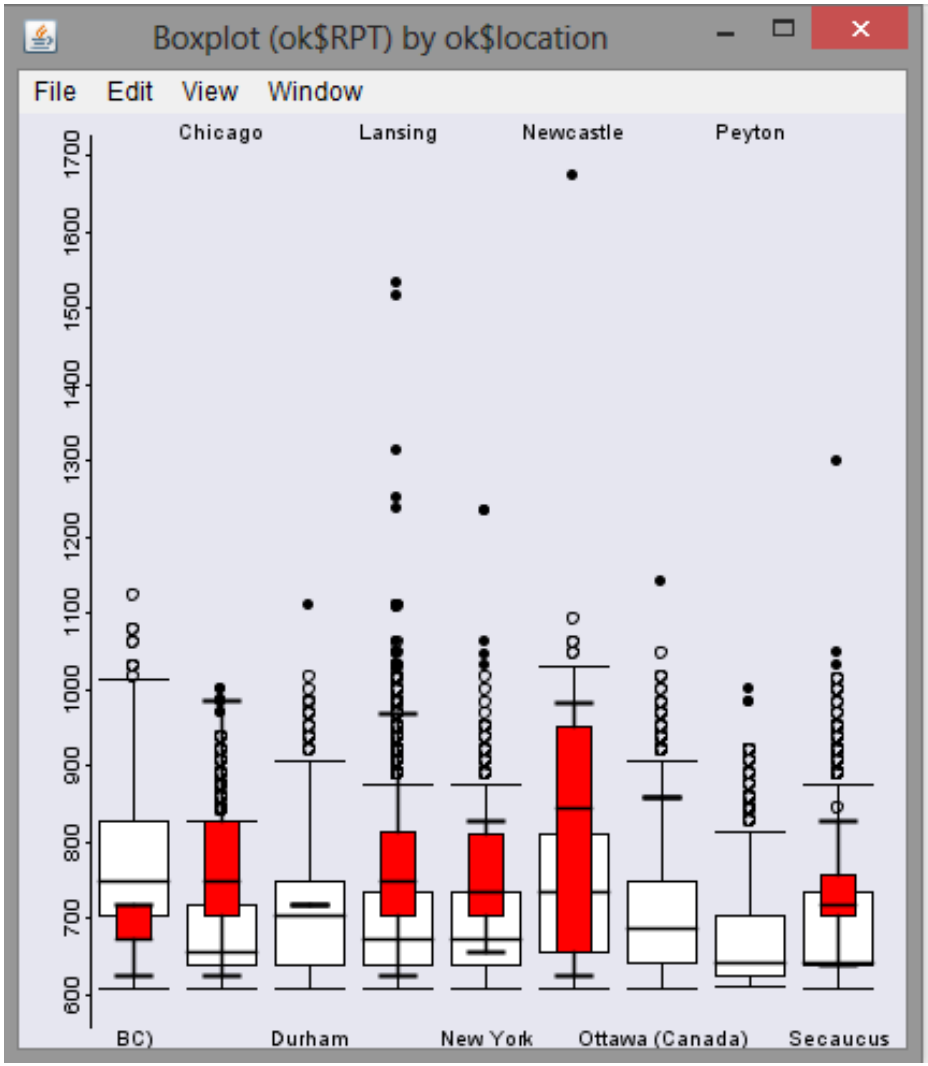
- SHIFT-CTRL: OR
- SHIFT: XOR

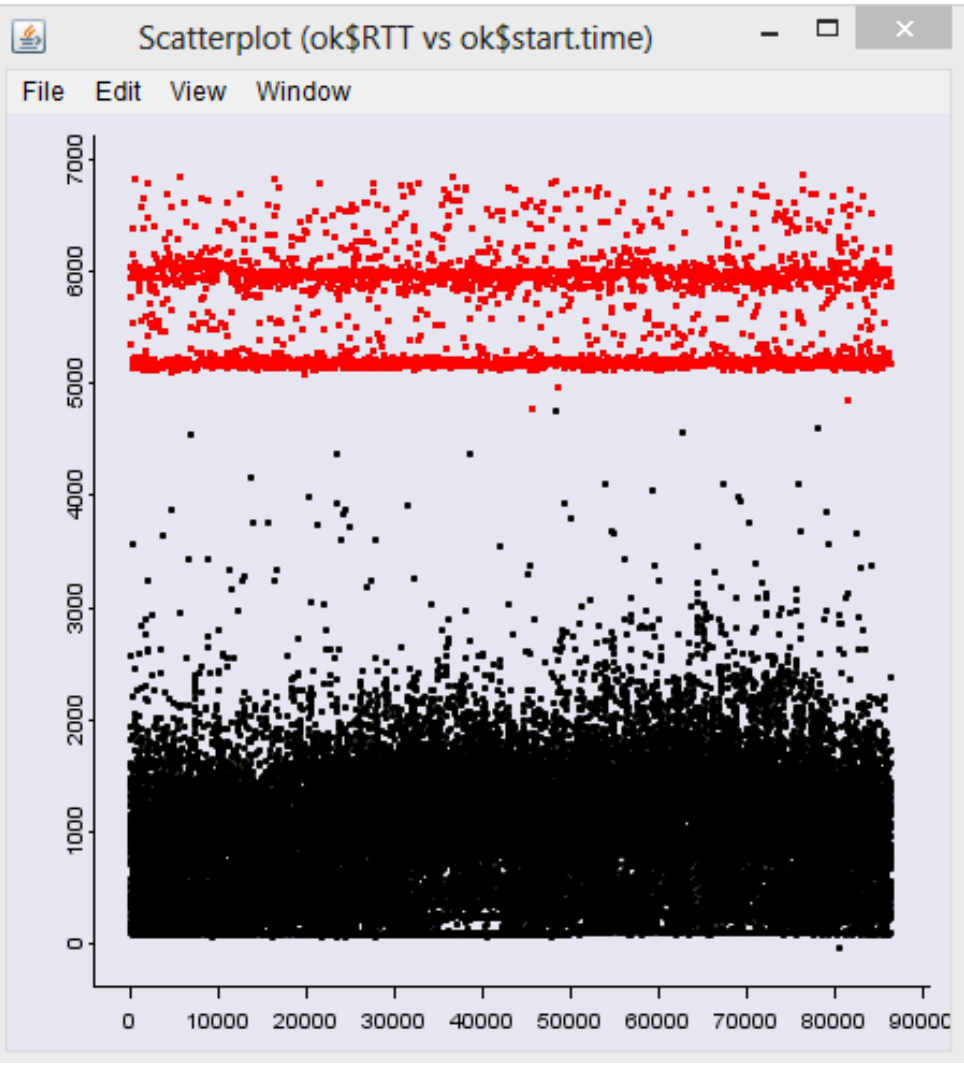
- Pointer, Drag-box, ~~Brush, Slicer, Lasso~~

- ~~Kijelölés-sorozatok~~

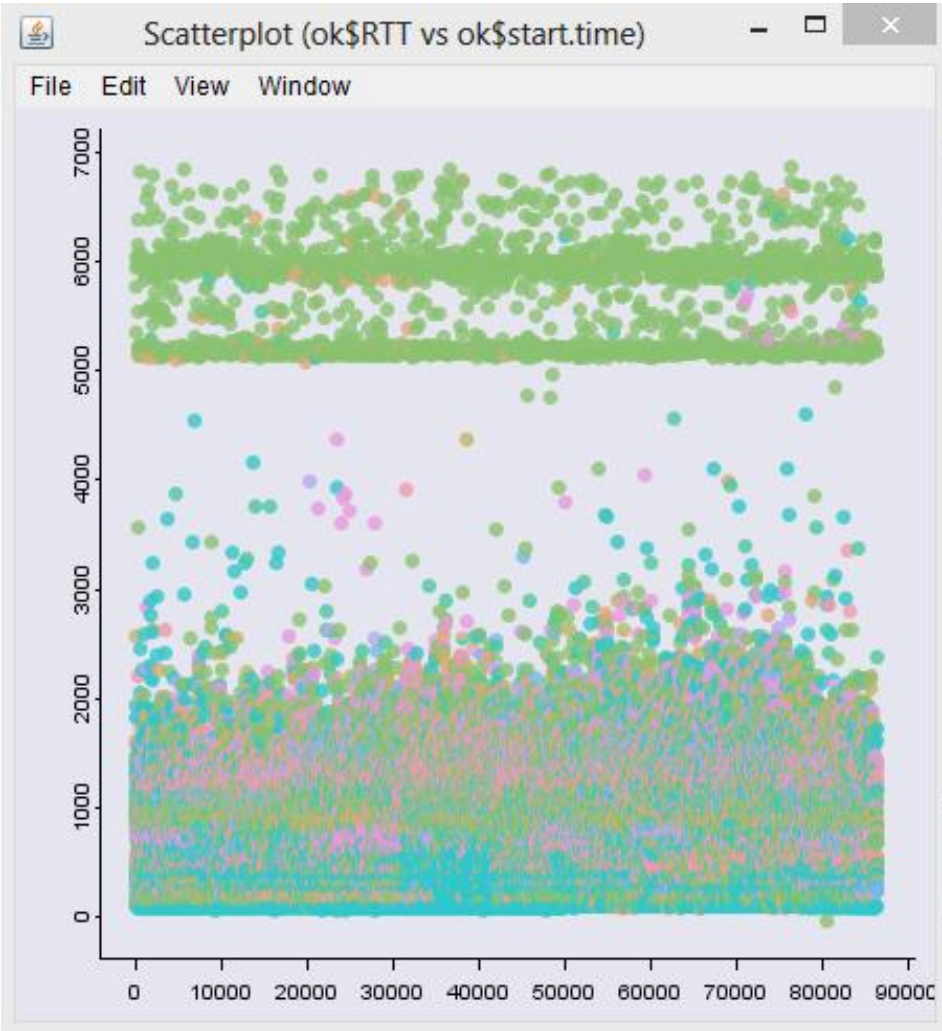


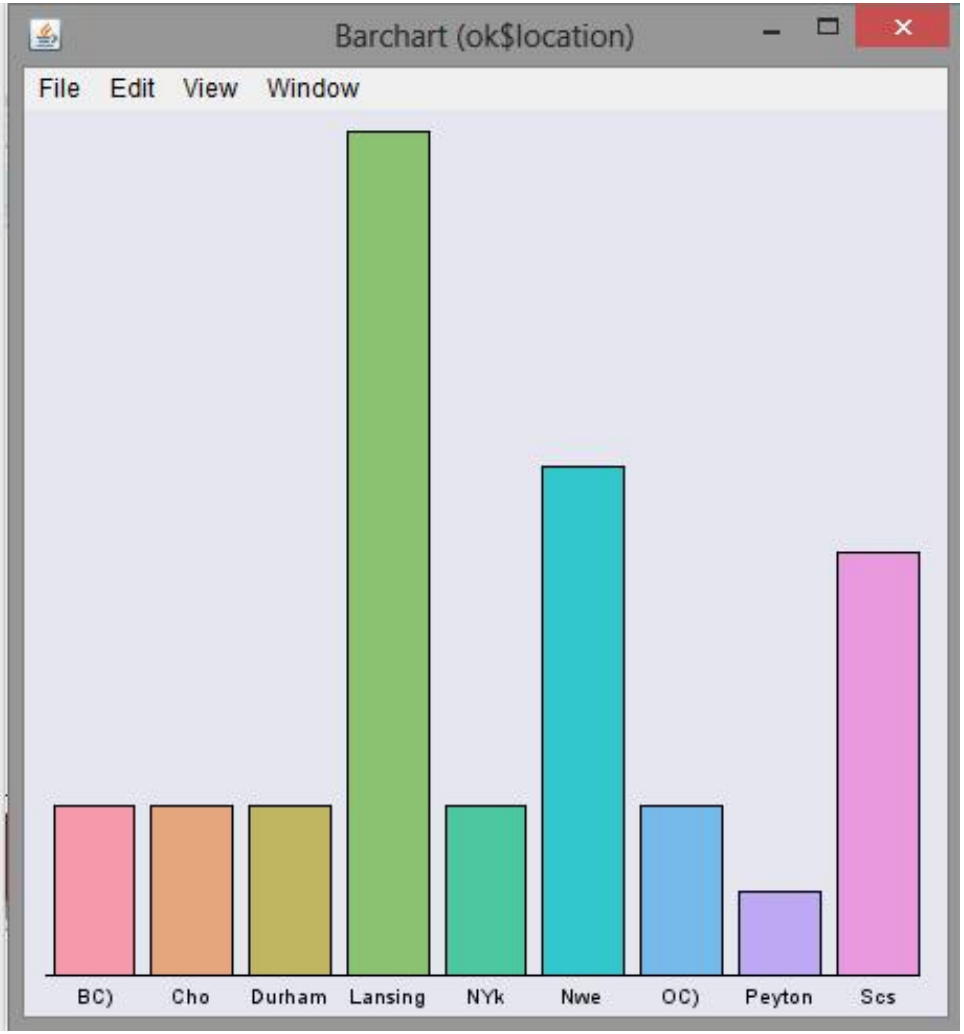
Csatolt kiemelés





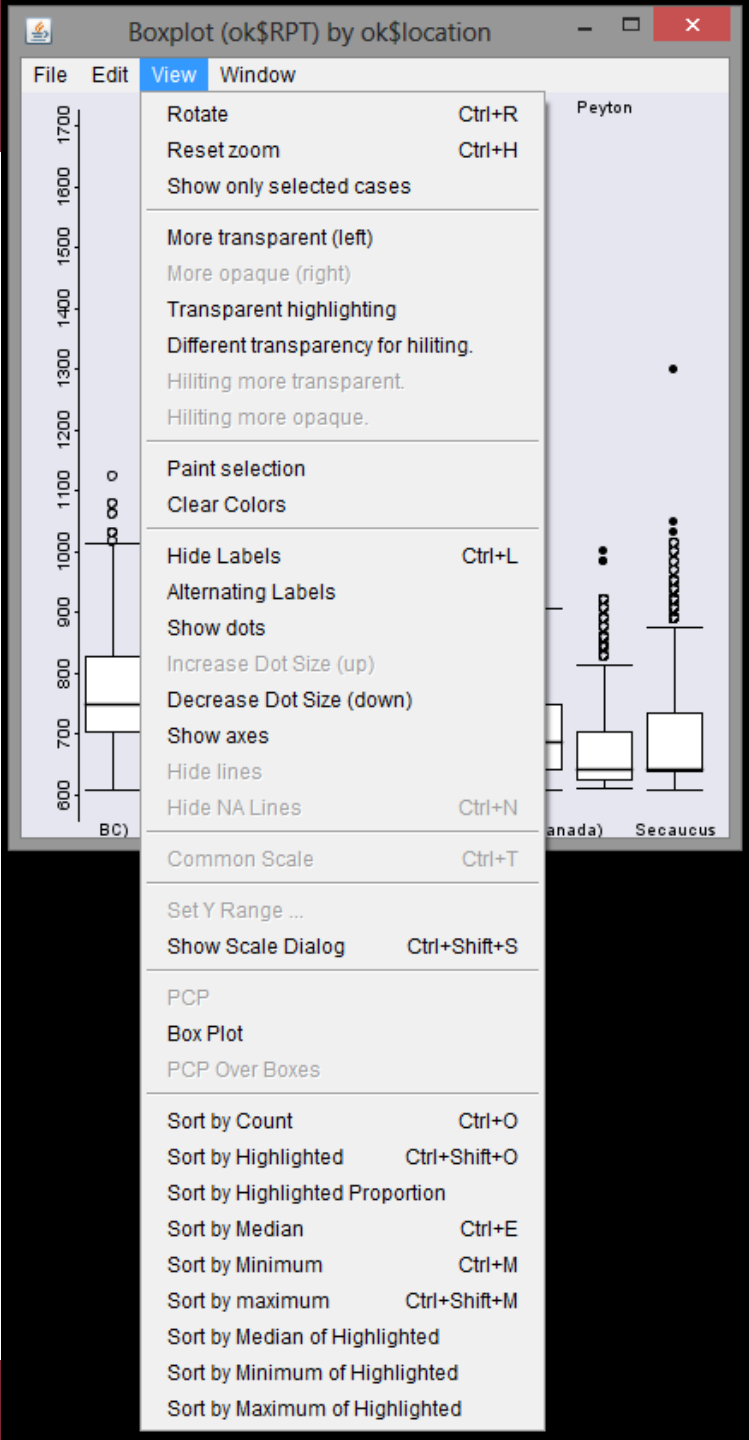
„Color brush”


Scatterplot (ok$RTT vs ok$start.time)
File Edit View Window
7000
6000
5000
4000
3000
2000
1000
0
0 10000 20000 30000 40000 50000 60000 70000 80000 90000


Barchart (ok$location)
File Edit View Window
BC) Cho Durham Lansing NYk Nwe OC) Peyton Scs

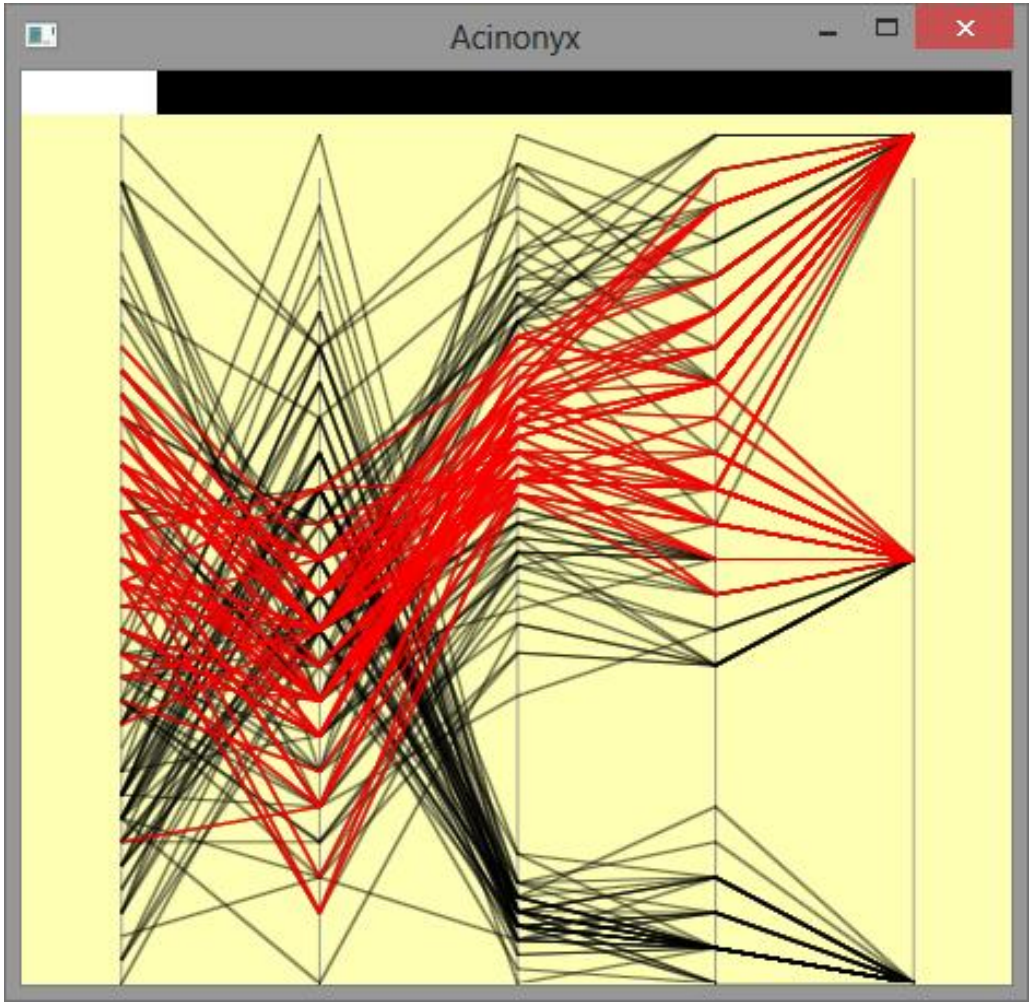
Interakció az ábrákkal

- Billentyűkombinációk és menük
- Paraméterek (pl. hisztogram)
- Tengelyek megcserélése
- Skálázás
- Nagyítás (középső egérgomb)
- Áttetszőség (←→)



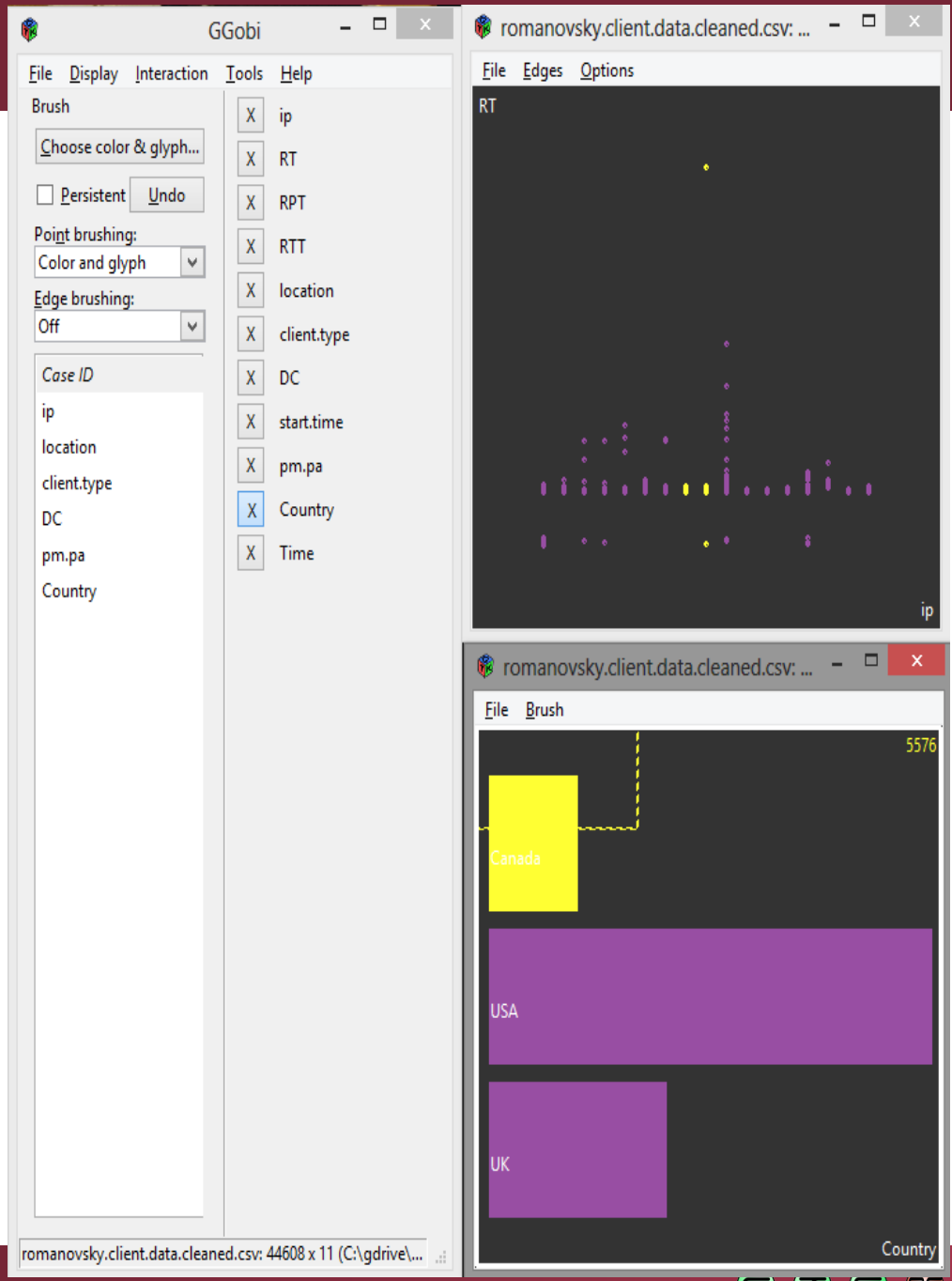
iPlots alternatívák: Acynonix

- „iPlots eXtreme”
- OpenGL gyorsítás
- Kiforrottság?

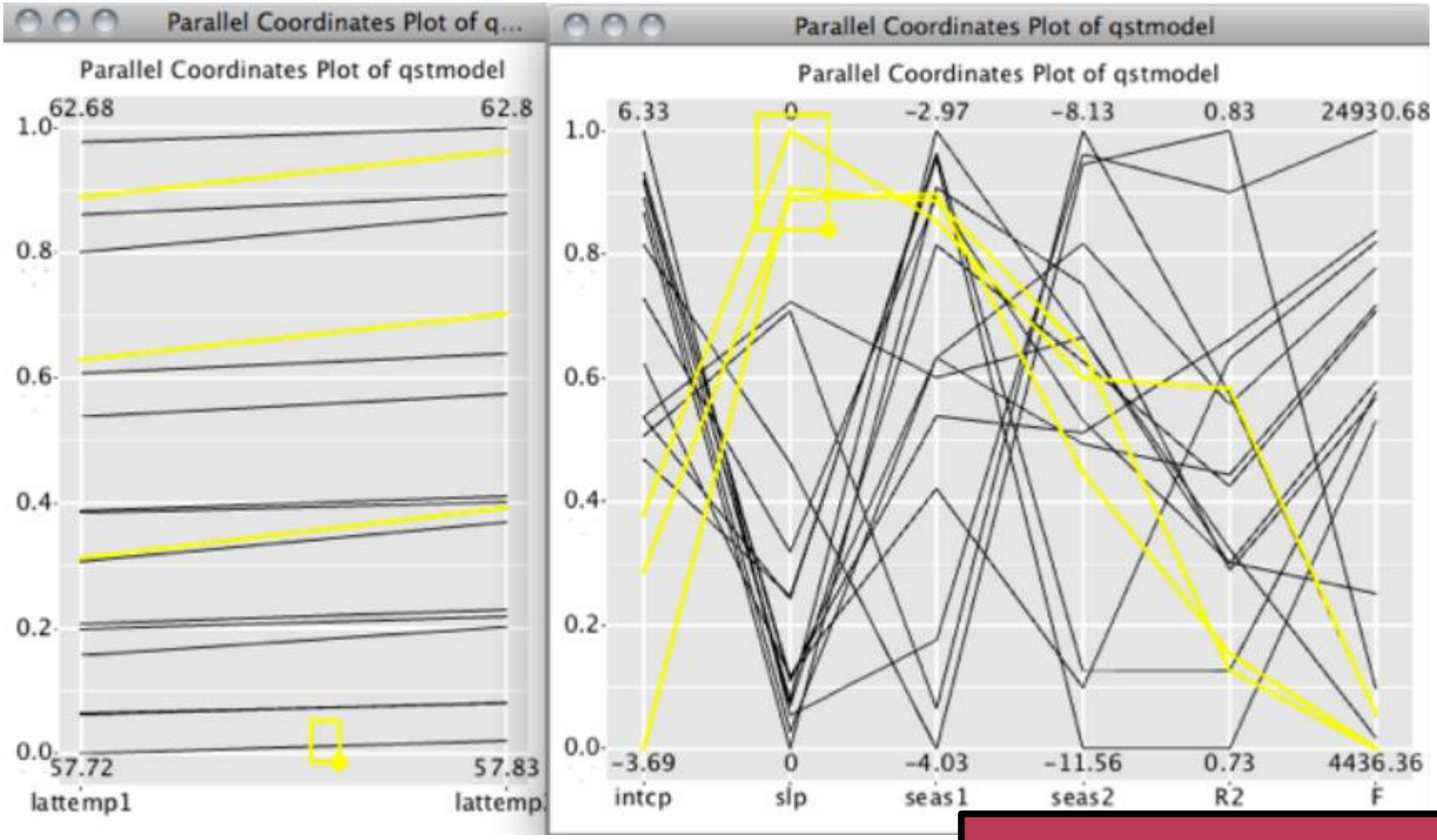


rggobi

- GGobi kötés
- Kiváló eszköz…
- … de nehézkes,
- GTK és C++,
- nincs aktív fejlesztés



cranvas



Qt; forever github…?

Forrás: [10], p 16

További alternatívák

- RStudio ggvis?
- RNavGraph?

- Ha nem kell komoly R kötés:
  - **Mondrian**, XmdvTool, Spotfire, Tableau, SAS JMP, Minitab, DataDesk, …

- Az R-be ágyazás előnyei:
  - Helyben az adat
  - Helyben a statisztika
  - **Helyben iteratív adatfinomítás**

Példa elemzési feladat

- Pataricza et al.: Empirical Assessment of Resilience
  - Az EDA-t a szolgáltatásbiztonság (*dependability*) elemzésében is kellene használnunk
  - [9]

- Itt:
  - Interaktív technikák szemléltetése
  - [9] munkafolyamatának néhány lépésén keresztül

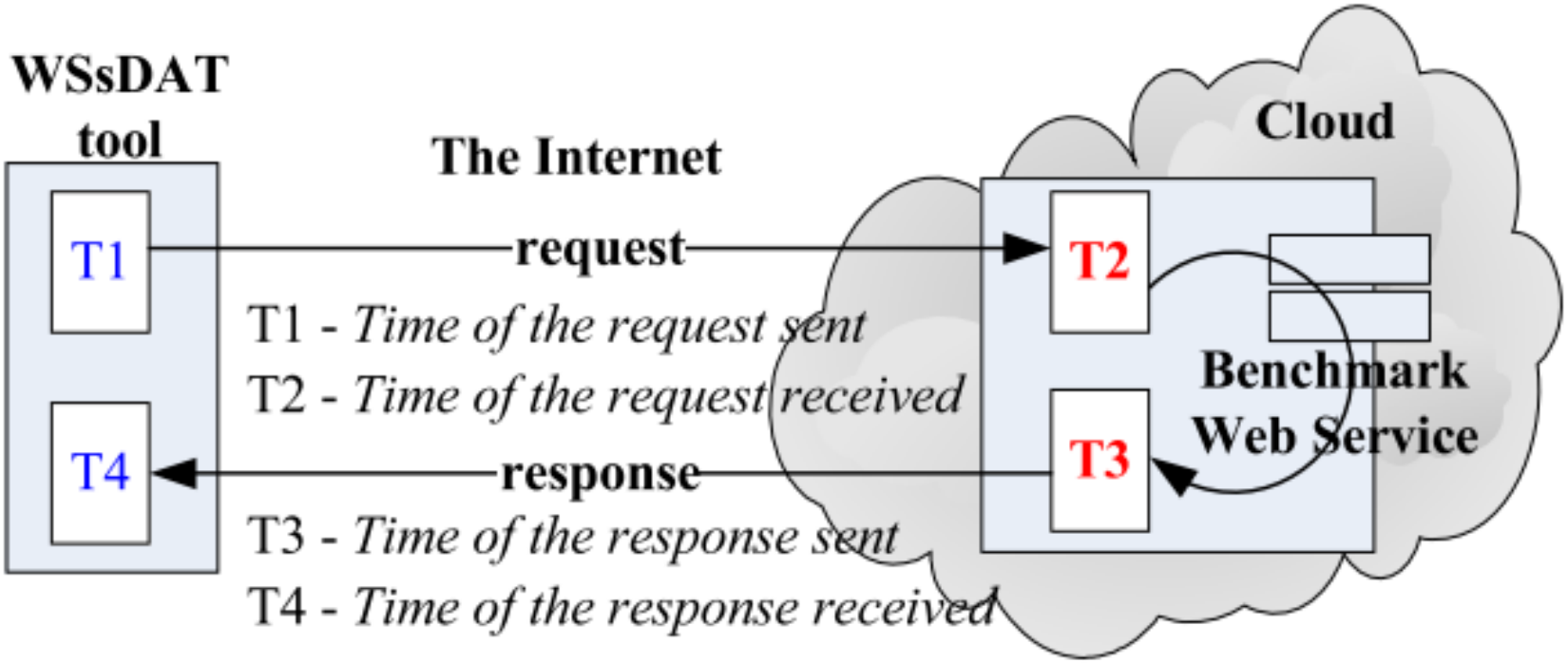
Példa adatkészlet

- Számítási felhő teljesítménymérések
  - Gorbenko et al. [8]

- Response Time = Request Processing Time + Round Trip Time

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

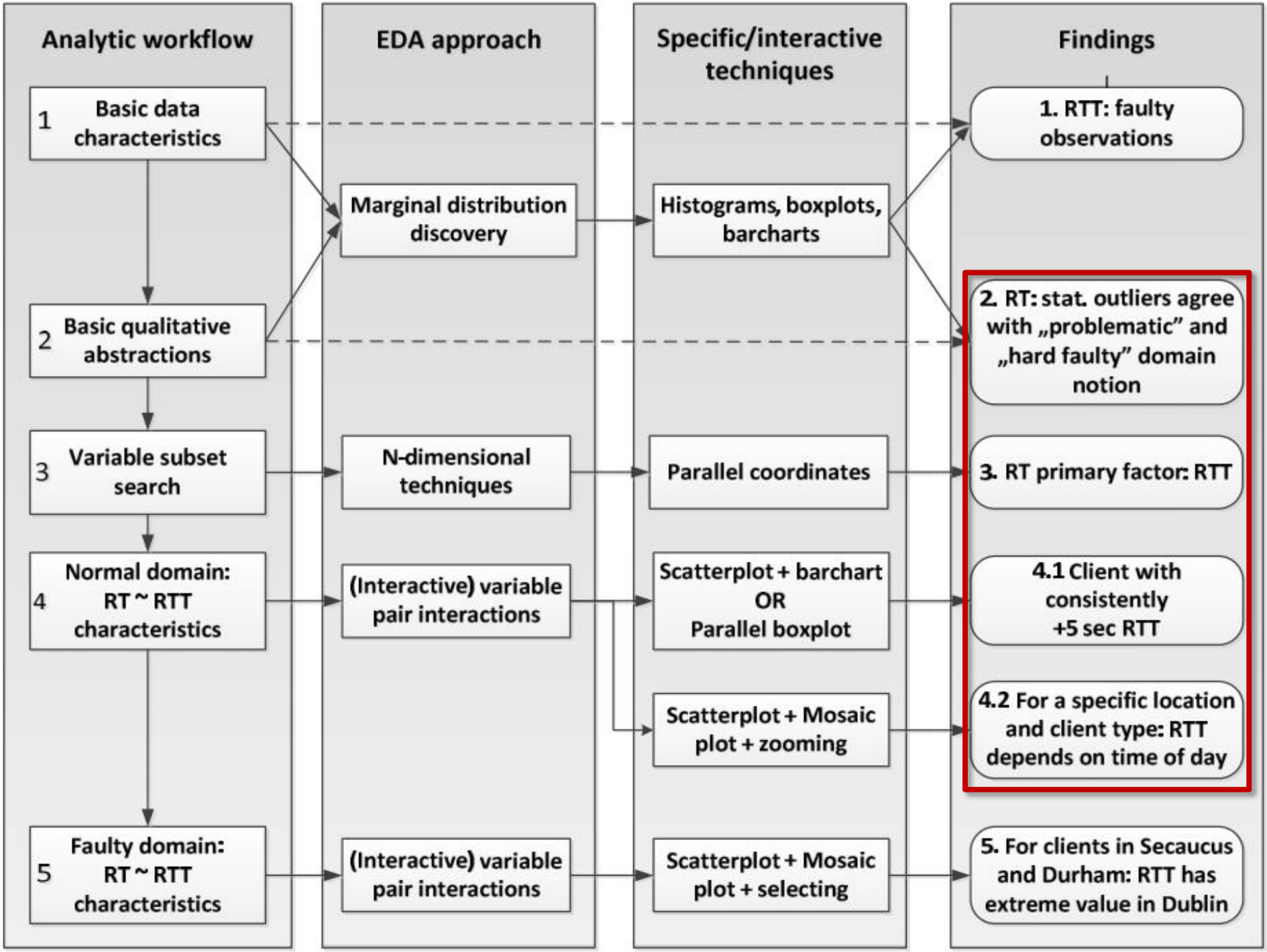
Példa adatkészlet



$$RT = T4-T1 \qquad RPT = T3-T2$$

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

Forrás: [8], p 186


Analytic workflow
EDA approach
Specific/interactive techniques
Findings
1 Basic data characteristics
2 Basic qualitative abstractions
3 Variable subset search
4 Normal domain: RT ~ RTT characteristics
5 Faulty domain: RT ~ RTT characteristics
Marginal distribution discovery
N-dimensional techniques
(Interactive) variable pair interactions
(Interactive) variable pair interactions
Histograms, boxplots, barcharts
Parallel coordinates
Scatterplot + barchart OR Parallel boxplot
Scatterplot + Mosaic plot + zooming
Scatterplot + Mosaic plot + selecting
1. RTT: faulty observations
2. RT: stat. outliers agree with „problematic" and „hard faulty" domain notion
3. RT primary factor: RTT
4.1 Client with consistently +5 sec RTT
4.2 For a specific location and client type: RTT depends on time of day
5. For clients in Secaucus and Durham: RTT has extreme value in Dublin

DEMO Adatkészlet

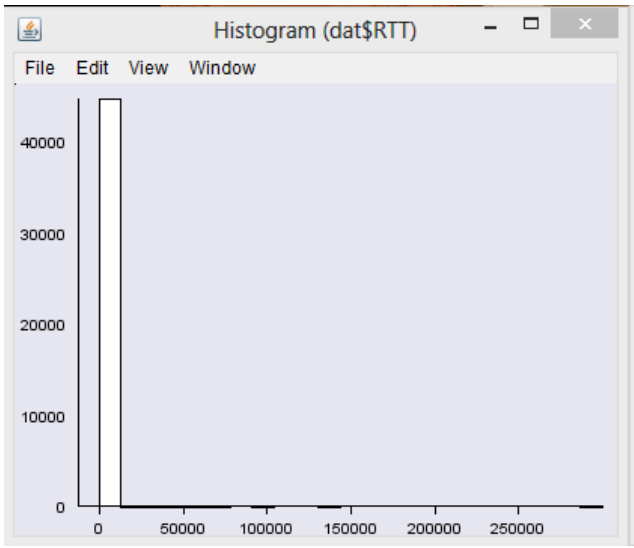
```
library('iplots')
dat <- read.table(myfilepath, sep=',', header=TRUE,
colClasses=c('factor', 'double', 'double', 'double', 'factor',
             'factor', 'factor', 'double', 'factor'))


dat$pm.pa <- NULL
dat$Time <- NULL
dat$start.time <- dat$start.time - min(dat$start.time, na.rm=TRUE)
dat <- dat[rowSums(is.na(dat)) == 0,]
```

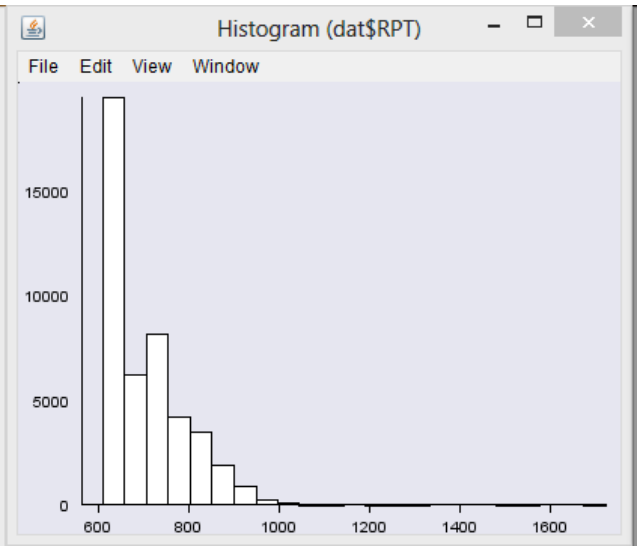
DEMO Adatkészlet

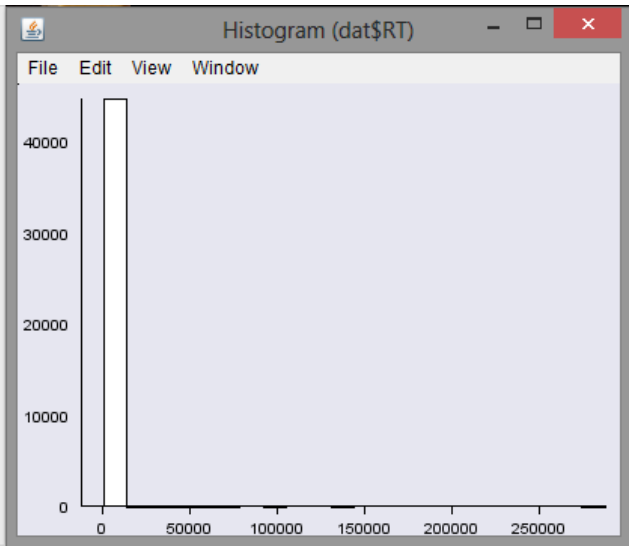
```
> summary(dat)
                    ip                RT                 RPT              RTT
[redacted] (laptop): 5570   Min.   :   613   Min.   : 609   Min.   :   -43
[redacted] (Desktop): 2788   1st Qu.:  1253   1st Qu.: 640   1st Qu.:   532
[redacted]          : 2788   Median :  1715   Median : 688   Median :  1015
[redacted]          : 2788   Mean   :  2011   Mean   : 709   Mean   :  1302
[redacted]          : 2788   3rd Qu.:  2088   3rd Qu.: 750   3rd Qu.:  1401
[redacted]          : 2788   Max.   :287448   Max.   :1672   Max.   :286526
(Other)             :25083
     location                  client.type                DC          start.time
Lansing  :13936   Java client     :40417   Dublin DC :26476   Min.   :    0
Newcastle: 8358   Microsoft client: 4176   Redmond DC:18117   1st Qu.:21548
Secaucus : 6966                                               Median :43185
Chicago  : 2788                                               Mean   :43178
Durham   : 2788                                               3rd Qu.:64798
New York : 2788                                               Max.   :86396
(Other)  : 6969
  Country
Canada: 5575
UK    :11146
USA   :27872
```

DEMO RT, RPT, RTT vizsgálata

```
> ihist(dat$RT)
ID:1 Name: "Histogram (dat$RT)"
> ihist(dat$RPT)
ID:2 Name: "Histogram (dat$RPT)"
> ihist(dat$RTT)
ID:3 Name: "Histogram (dat$RTT)"
>
```



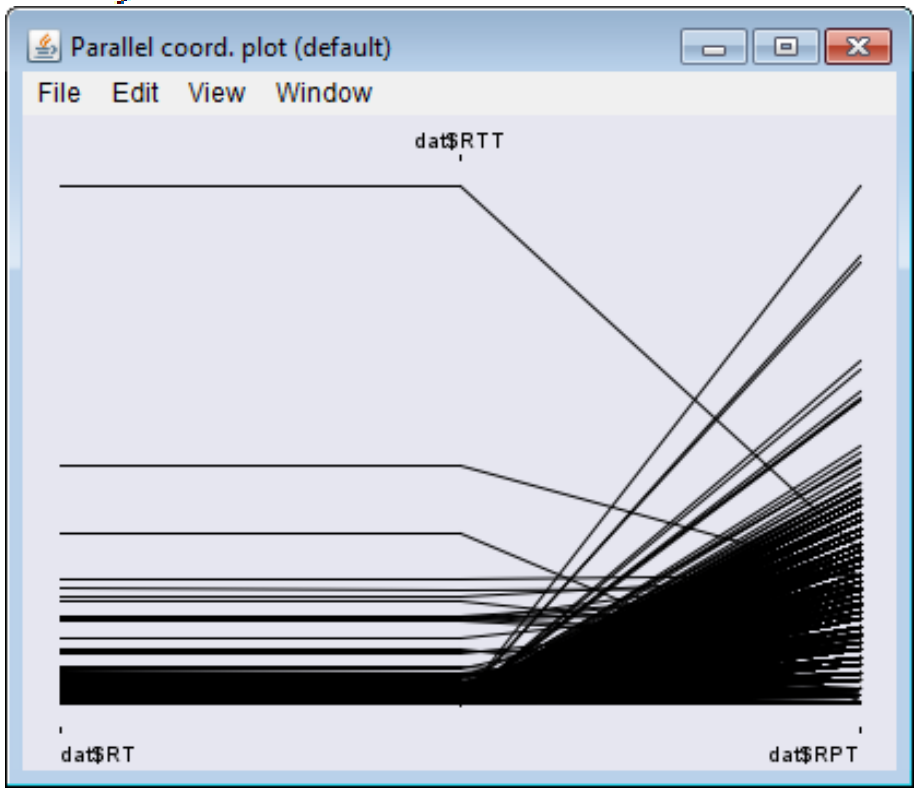




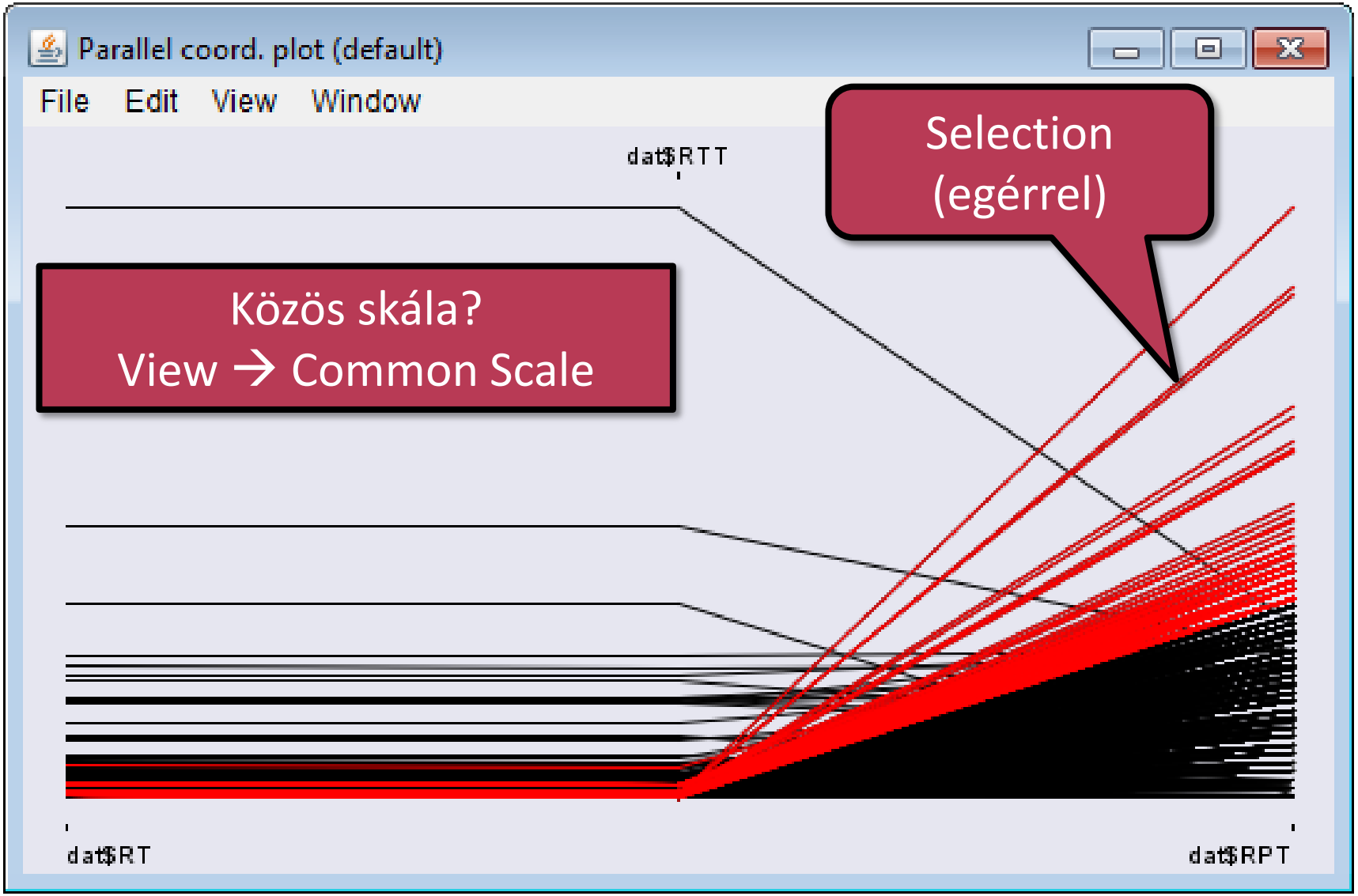
Kapcsolatok?

DEMO RT, RPT, RTT vizsgálata

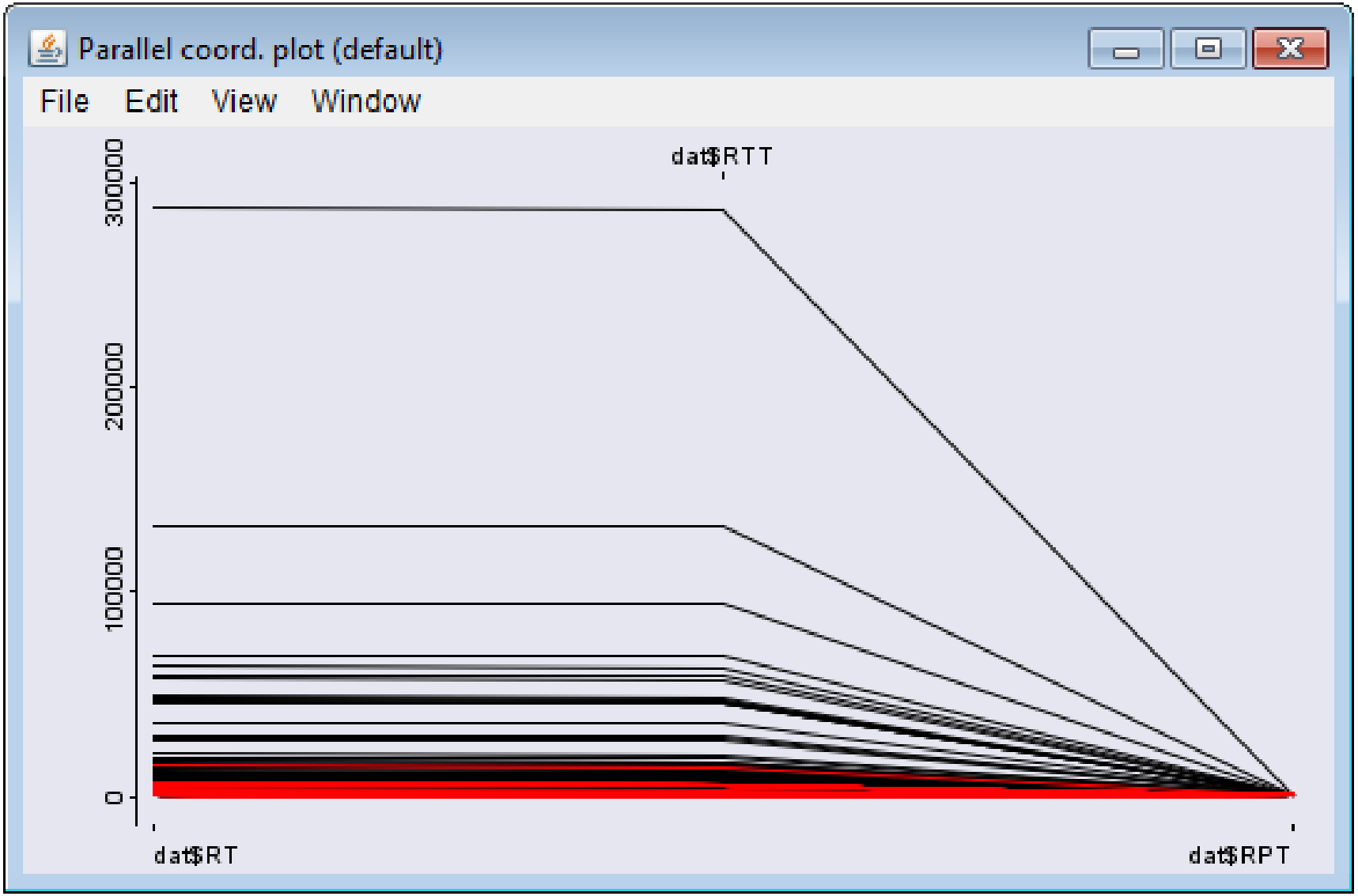
```
> ihist(dat$RT)
ID:1 Name: "Histogram (dat$RT)"
> ihist(dat$RPT)
ID:2 Name: "Histogram (dat$RPT)"
> ihist(dat$RTT)
ID:3 Name: "Histogram (dat$RTT)"
> ipcp(dat$RT, dat$RTT, dat$RPT)
ID:4 Name: "Parallel coord. plot (default)"
>
```



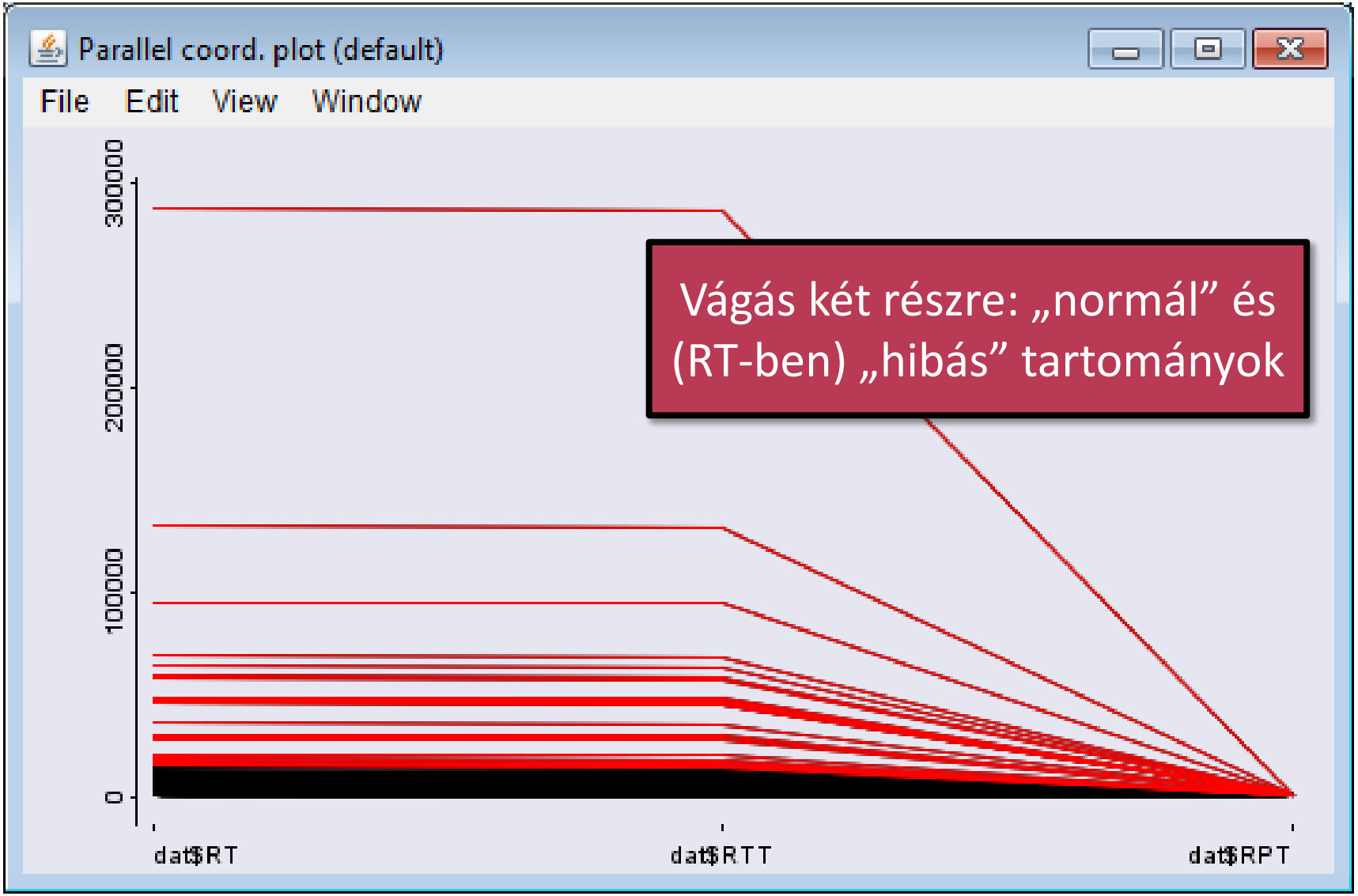
DEMO RT, RPT, RTT vizsgálata



DEMO „Common scale” után

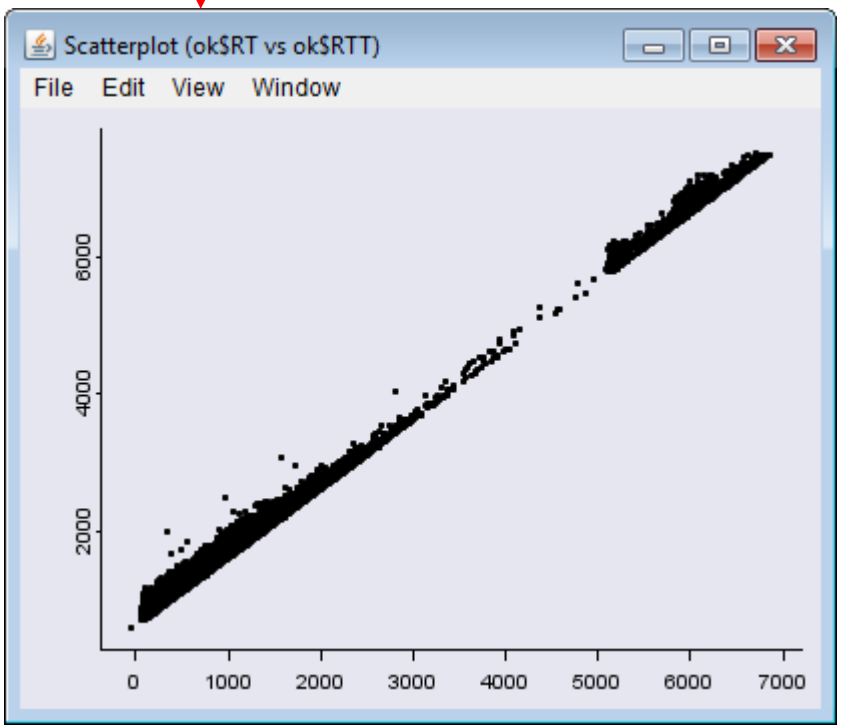


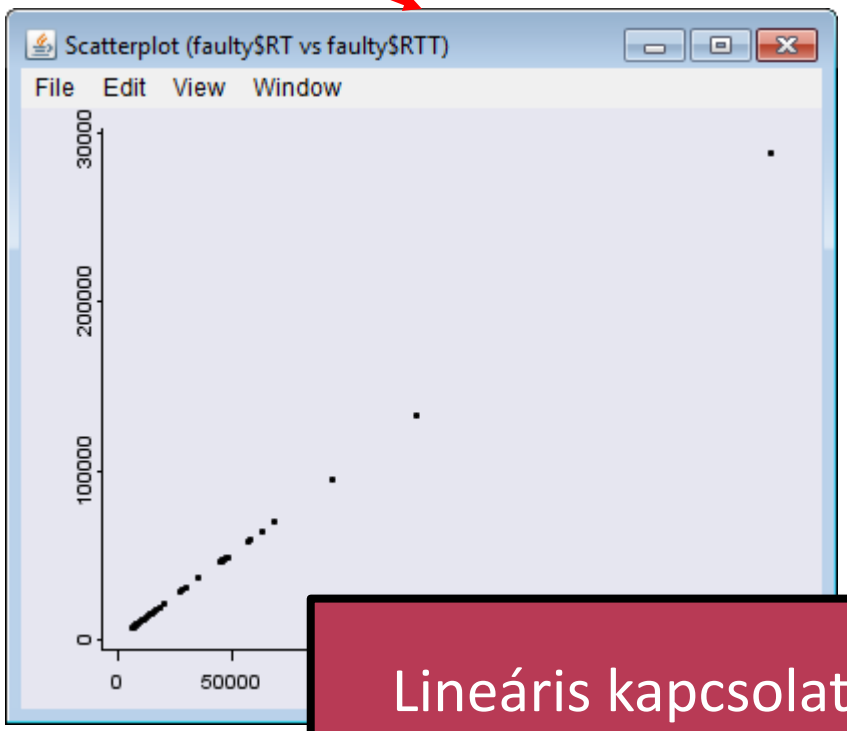
DEMO RT ~ RTT?


Parallel coord. plot (default)
File
Edit
View
Window
300000
200000
100000
0
dat$RT
dat$RTT
dat$RPT
Vágás két részre: „normál” és (RT-ben) „hibás” tartományok

DEMO Vágás

```
ID:1 Name: "Parallel coord. plot (default)"
> ok <- dat[dat$RT < 7500,]
> faulty <- dat[dat$RT >= 7500,]
> iplot(faulty$RTT, faulty$RT)
iSet and data length differ. Please observe the dialog box (it may be hidden by the R window).
ID:1 Name: "Scatterplot (faulty$RT vs faulty$RTT)"
> iplot(ok$RTT, ok$RT)
iSet and data length differ. Please observe the dialog box (it may be hidden by the R window).
ID:1 Name: "Scatterplot (ok$RT vs ok$RTT)"
>
```



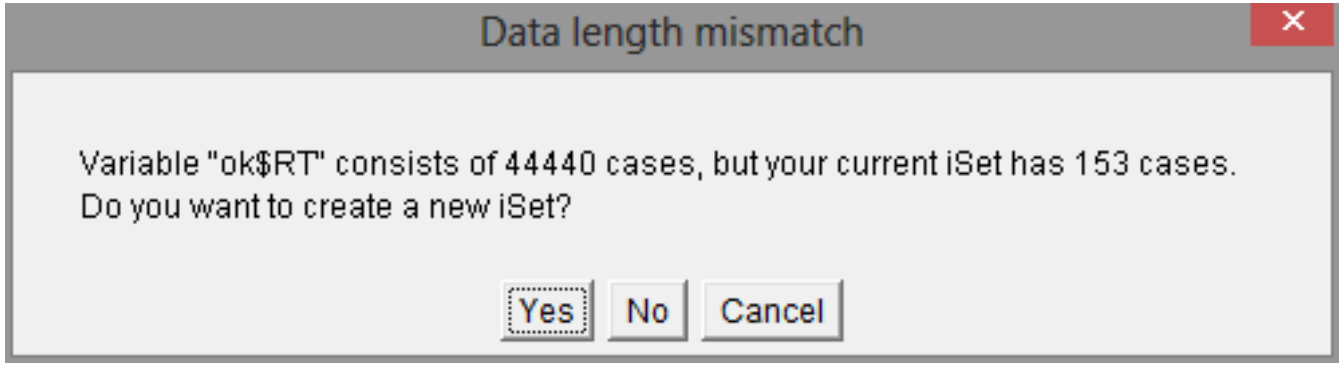


Lineáris kapcsolat?

DEMO Vágás

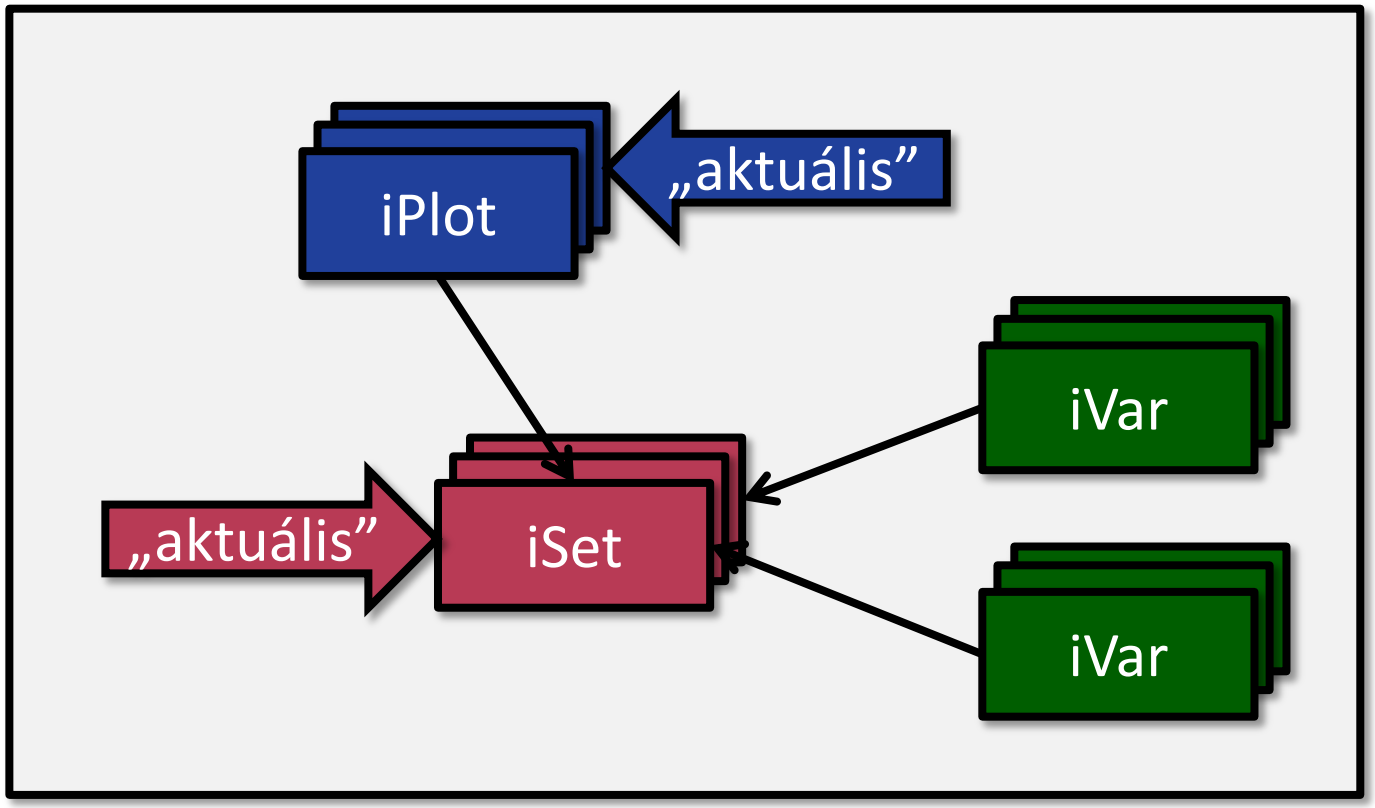
```
ID:1 Name: "Parallel coord. plot (default)"
> ok <- dat[dat$RT < 7500,]
> faulty <- dat[dat$RT >= 7500,]
> iplot(faulty$RTT, faulty$RT)
iSet and data length differ. Please observe the dialog box (it may be hidden by the R window).
ID:1 Name: "Scatterplot (faulty$RT vs faulty$RTT)"
> iplot(ok$RTT, ok$RT)
iSet and data length differ. Please observe the dialog box (it may be hidden by the R window).
ID:1 Name: "Scatterplot (ok$RT vs ok$RTT)"
>
```

???



Kapcsolat az R-rel

- Valójában Java-t használunk (+ df-másolás)
- Objektumok: Változóba regisztrálás, „léptetés”, listázás, módosítás



EDA több adatkészlet felett

- Quick & dirty EDA egy adatkeretre: nem kell foglalkoznunk iSet/iVar/iPlot-okkal

- iSet létrehozása: iset
  - Az iVar-ok a szelekciós operátorokkal elérhetők
- i{plot|bar|pcp|…}: az „aktuális” iSet-en
- Aktuális iSet „átállítása”: iset.set

- Kijelölés: iSet-en értelmezett!
- A végigvezetett demo-ban nincs ezekre szükség
  - Bár nem „szép” megoldás feleslegesen új iSet-eket létrehozni…

DEMO Több iSet explicit kezelése

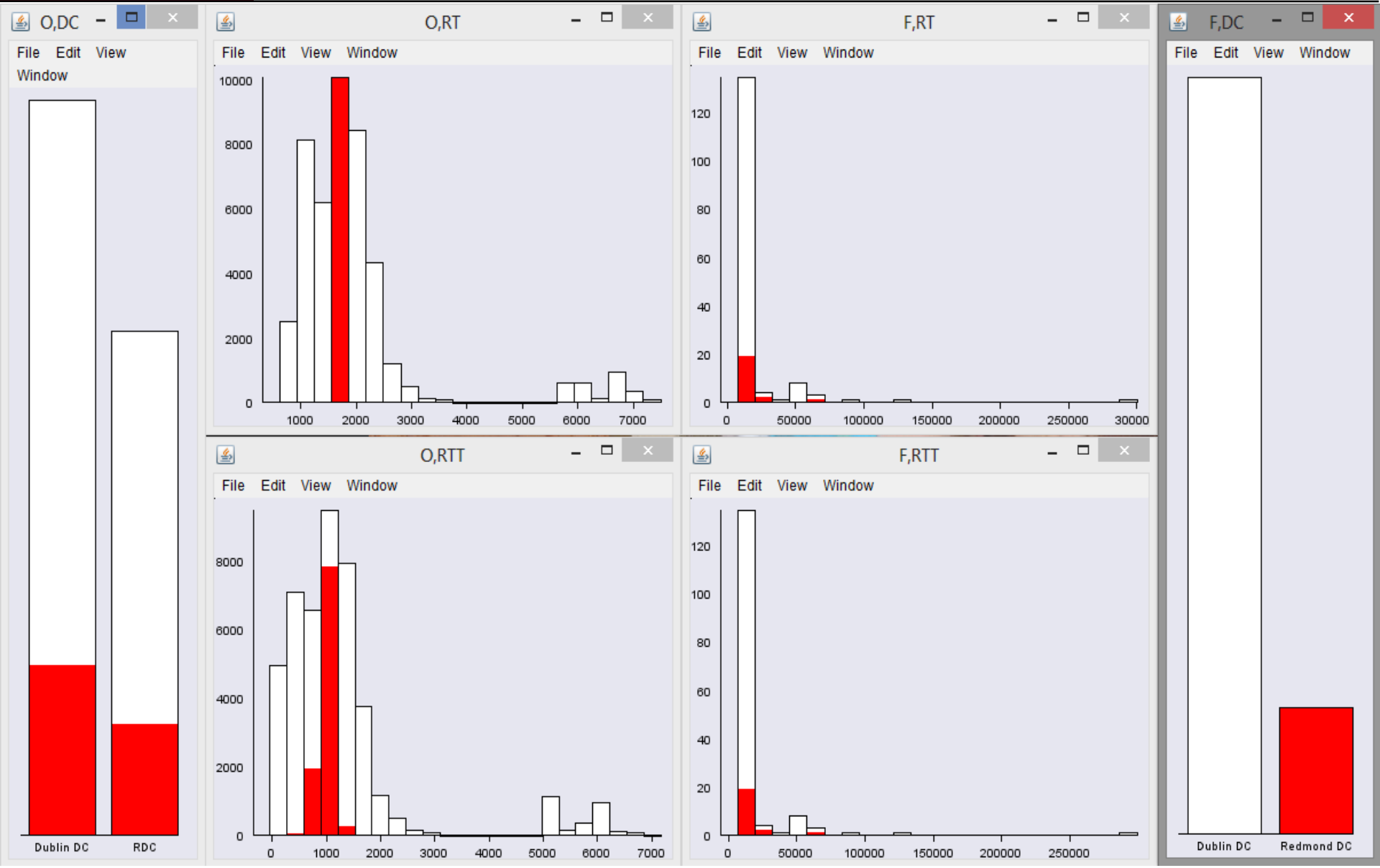
```
fts <- iset.new("faultyset", faulty)
ihist(fts$RT, title="F,RT")

oks <- iset.new("okset", ok)
ihist(oks$RT, title="O,RT")

iset.set("faultyset")
ihist(fts$RTT, title="F,RTT")

iset.set("okset")
ihist(oks$RTT, title="O,RTT")
ibar(oks$DC, title="O,DC")

iset.set("faultyset")
ibar(fts$DC, title="F,DC")
```

iSet, mint objektum

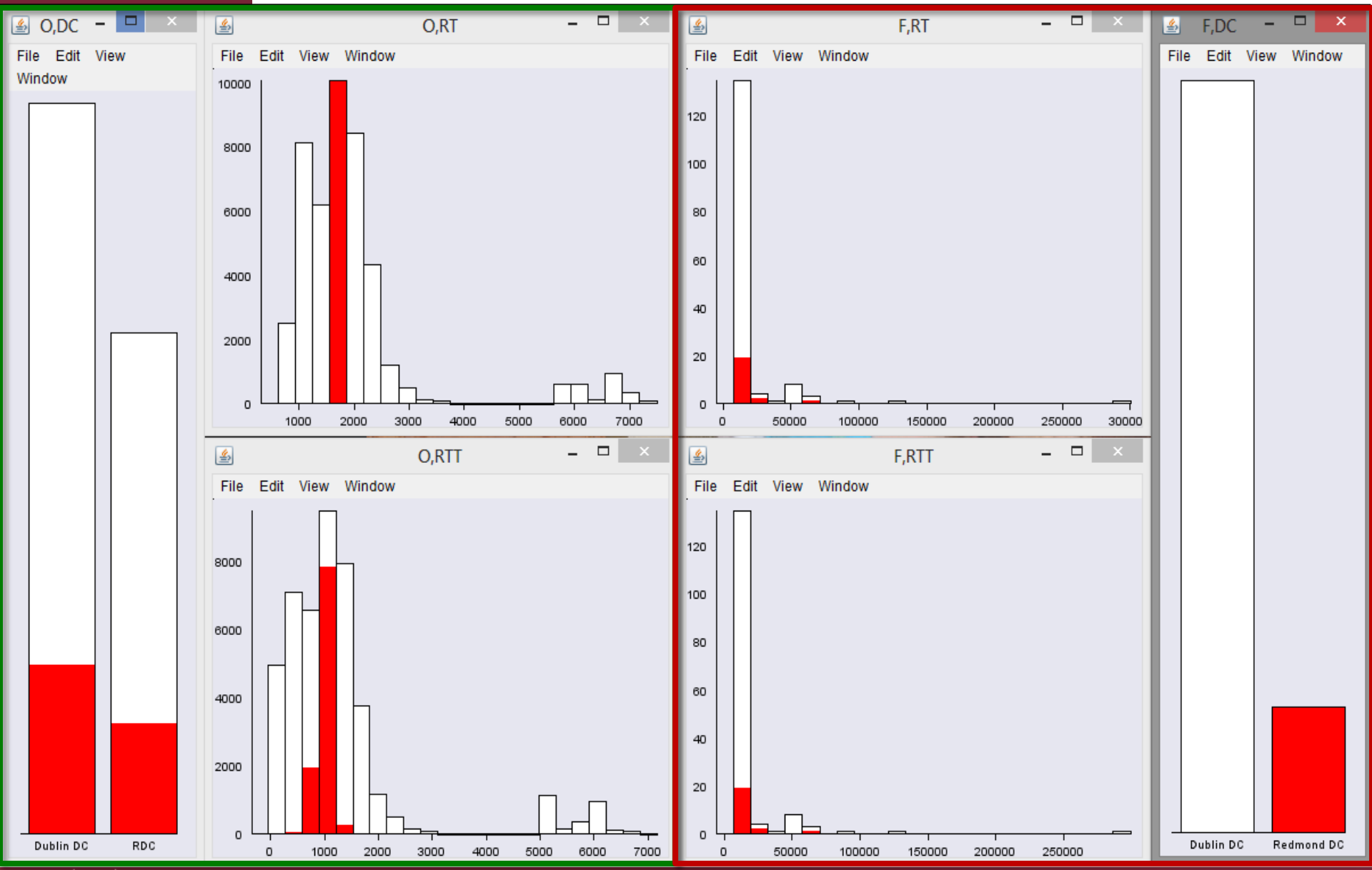
iSet-változó megjelenítése

Aktuális iSet átállítása

DEMO Több iSet explicit kezelése



DEMO Több iSet explicit kezelése



DEMO Több iSet explicit kezelése

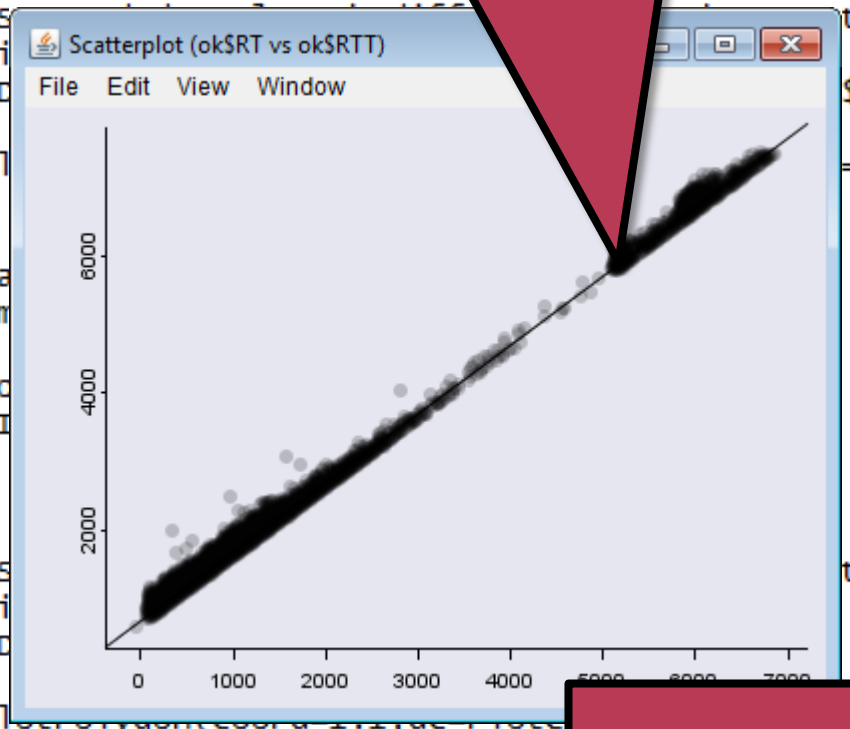
```
> iset.set(iset.next())
[1] "okset"
> iset.list()
faultyset okset
        2     3
> iplot.list()
[[1]]
ID:1 Name: "Histogram (RT)"
[[2]]
ID:2 Name: "Histogram (RTT)"
[[3]]
ID:3 Name: "Barchart (DC)"
```

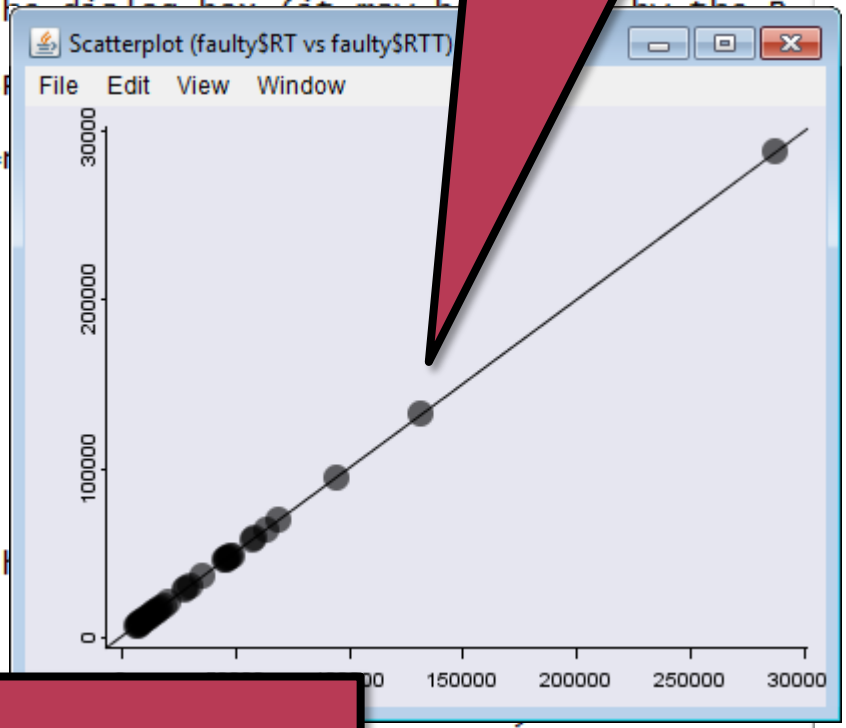
Az aktuális iSet-re

DEMO Visszatérve a példára…

Módosított átlátszóság
View → More transparent
(vagy ←→)

Nagyobb pontméret
View → Larger points
(vagy ↑↓)

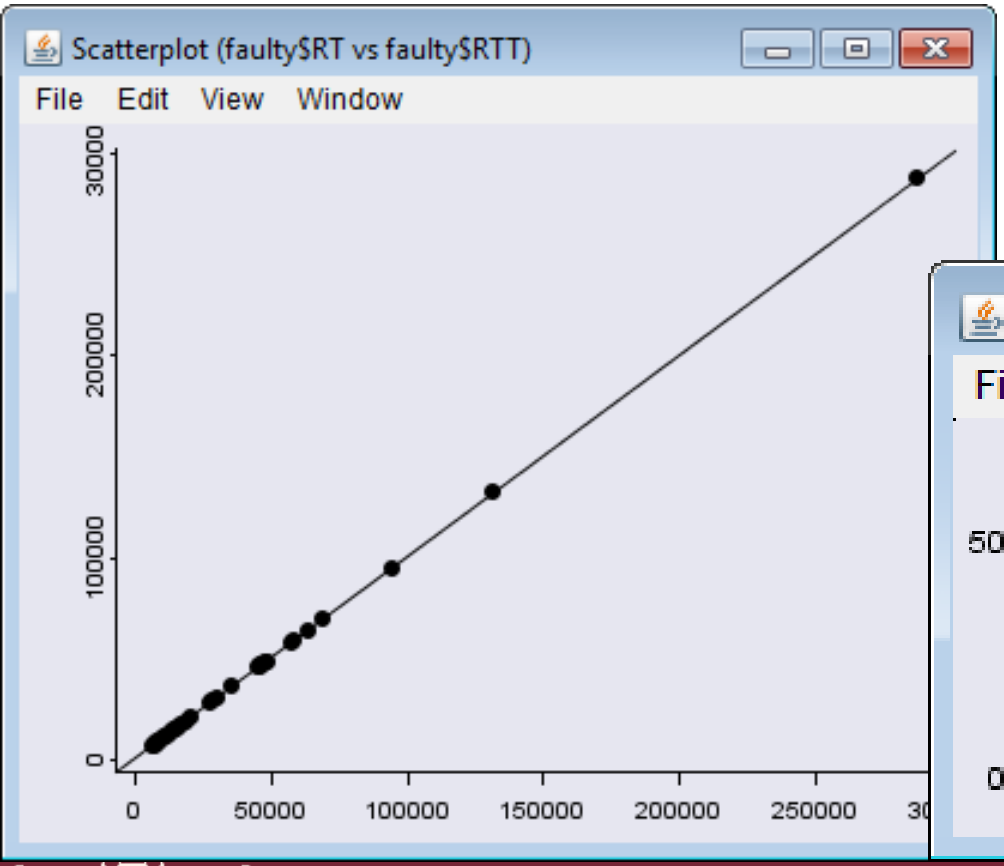


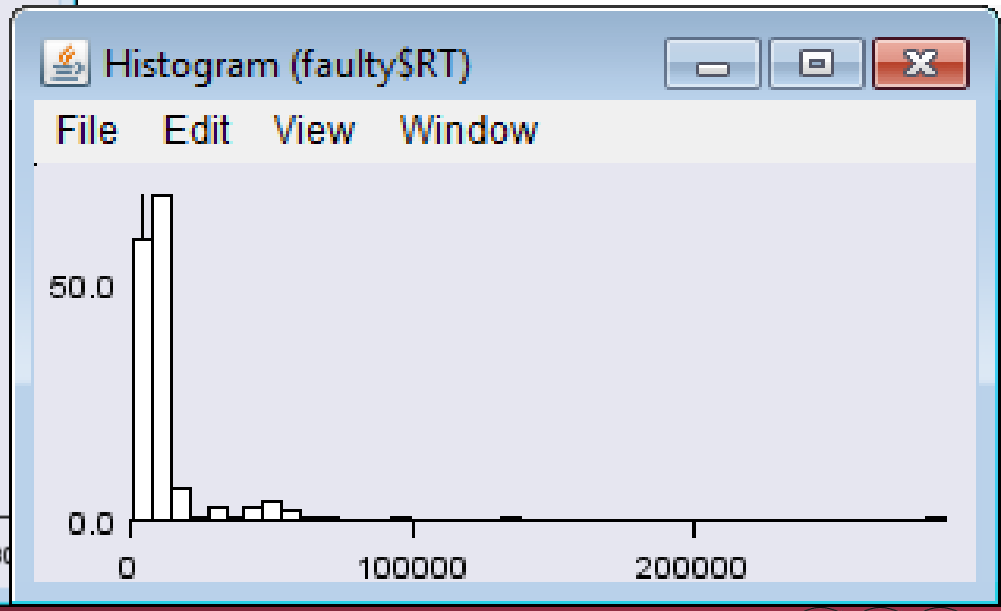


Lineáris kapcsolat!

DEMO RT vs. RTT – „kilógó” esetek

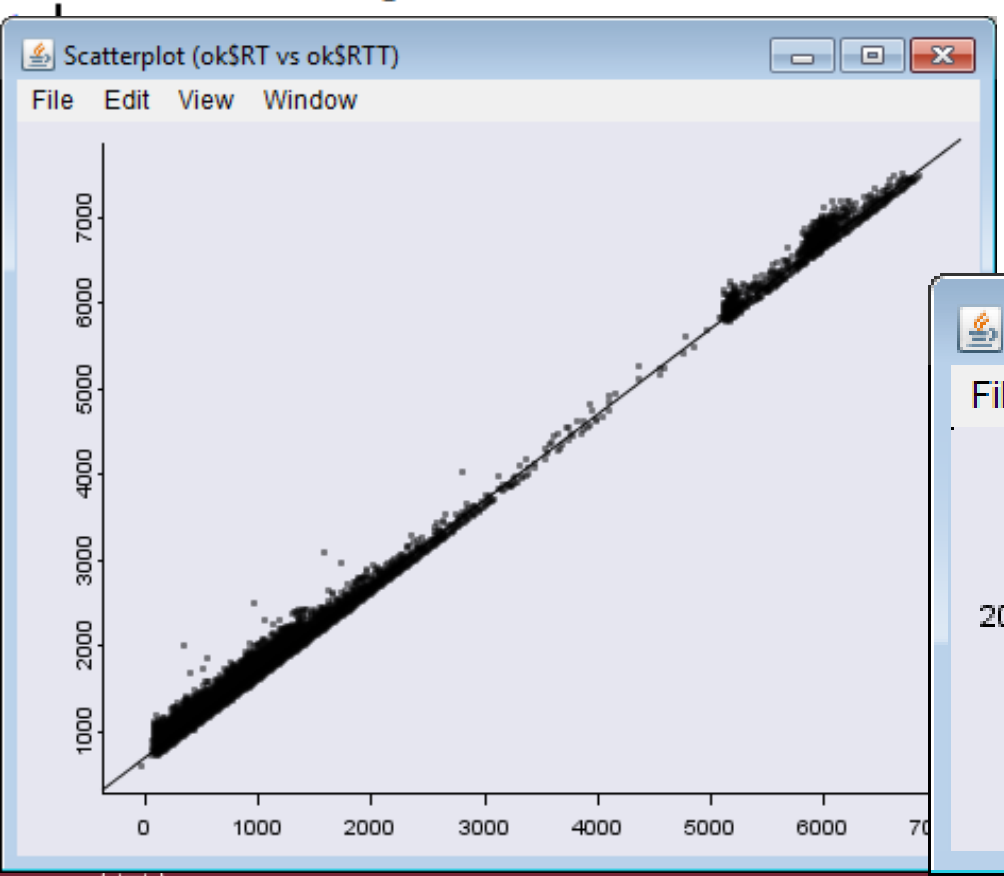
```
ID:1 Name: "Scatterplot (ok$RT vs ok$RTT)"
> iabline(708, 1)
PlotPolygon(coord=1:1,dc=PlotColor(black),fc=none,points=2,visible=true)
> ihist(faulty$RT)
iSet and data length differ. Please observe the dialog box (it may be hidden by the R
window).
ID:1 Name: "Histogram (faulty$RT)"
```



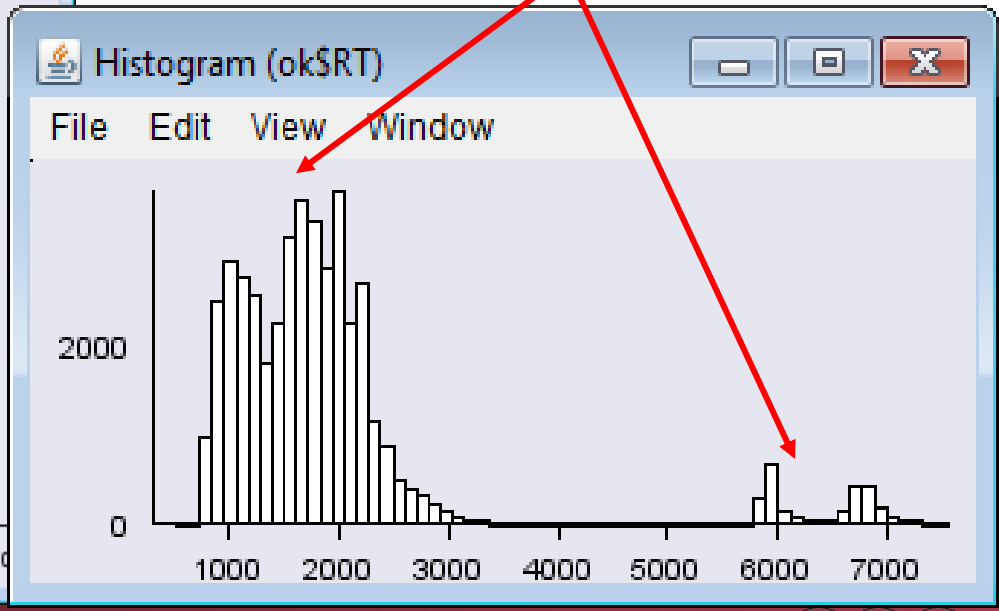


DEMO RT vs. RTT – „normál” esetek

```
> ihist(faulty$RT)
iset and data length differ. Please observe the dialog box (it may be hidden by the R
window).
ID:1 Name: "Histogram (faulty$RT)"
> ihist(ok$RT)
iset and data length differ. Please observe the dialog box (it may be hidden by the R
window).
ID:1 Name: "Histogram (ok$RT)"
```



Két diszjunkt tartomány?

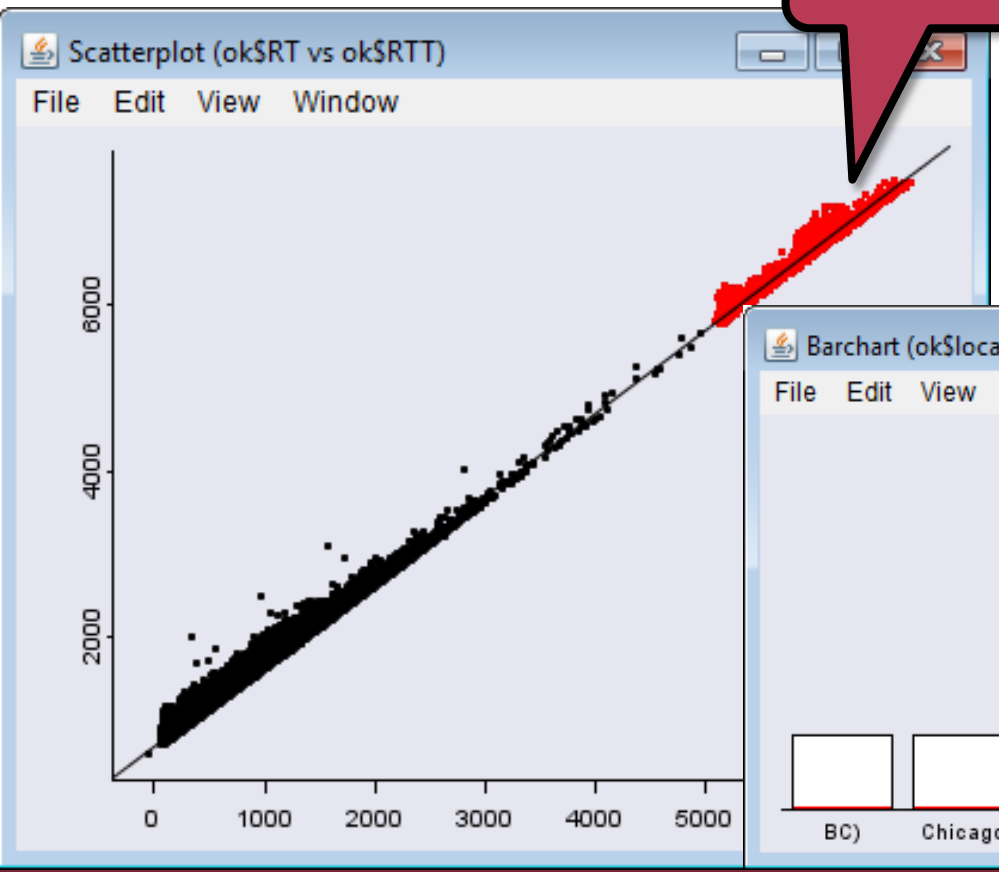


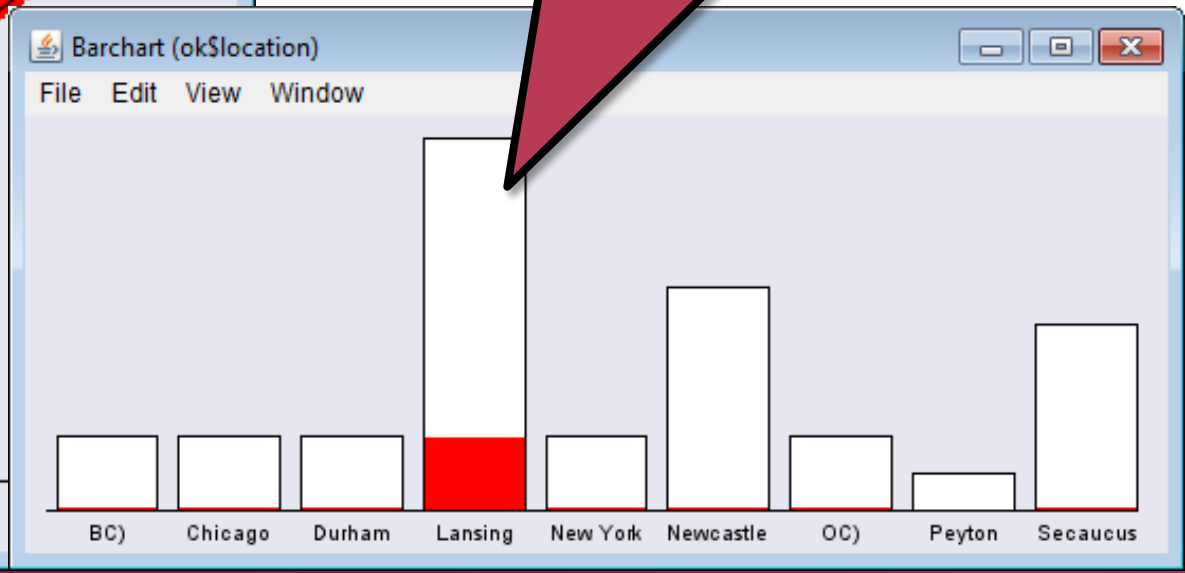
DEMO Gyanús kliens felderítése

```
> ihist(ok$RT)
iSet and data length differ. Please observe the dialog box (it may be hidden by the R
window).
ID:1 Name: "Histogram (ok$RT)"
> ibar(ok$location)
ID:2 Name: "Barchart (ok$location)"
>
```

Selection

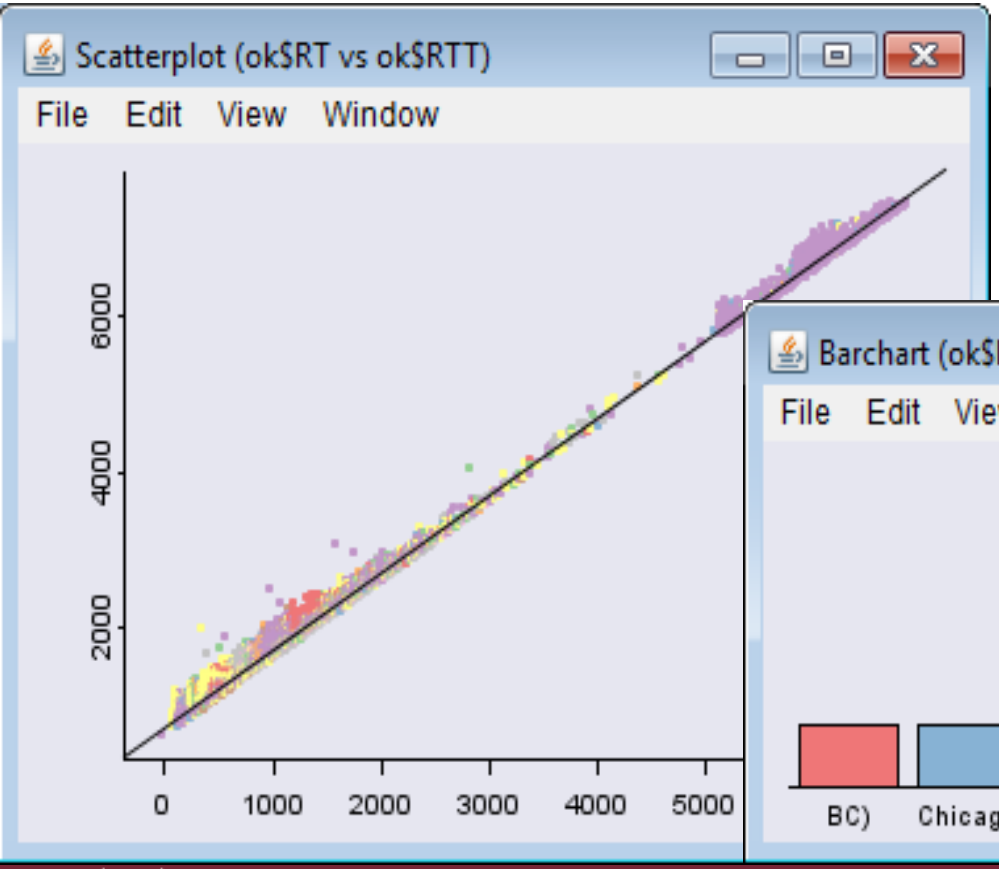
Linked Highlighting

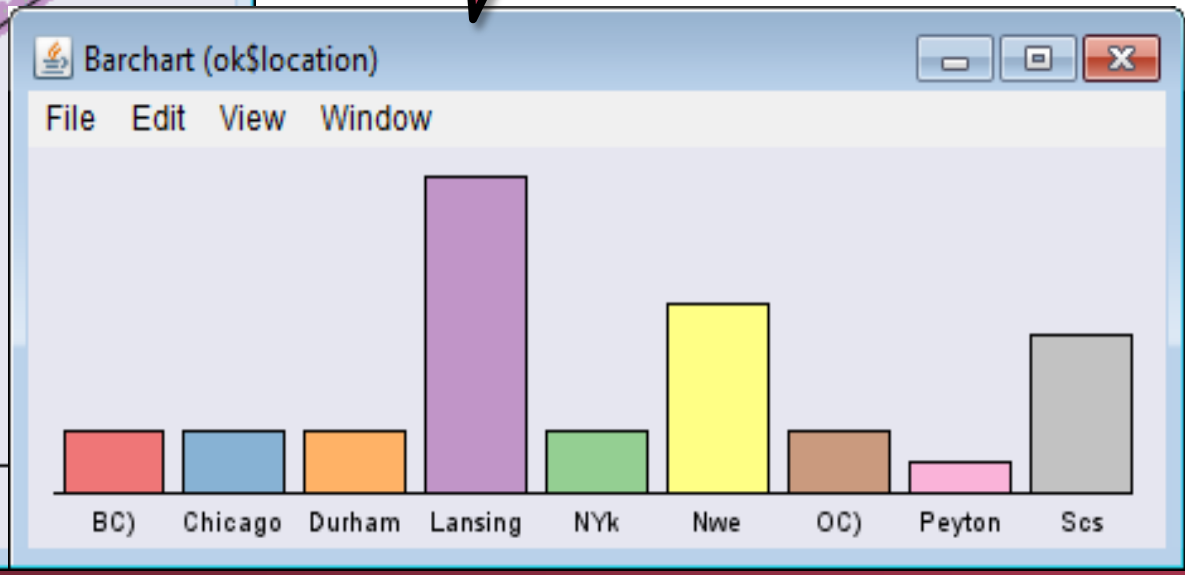




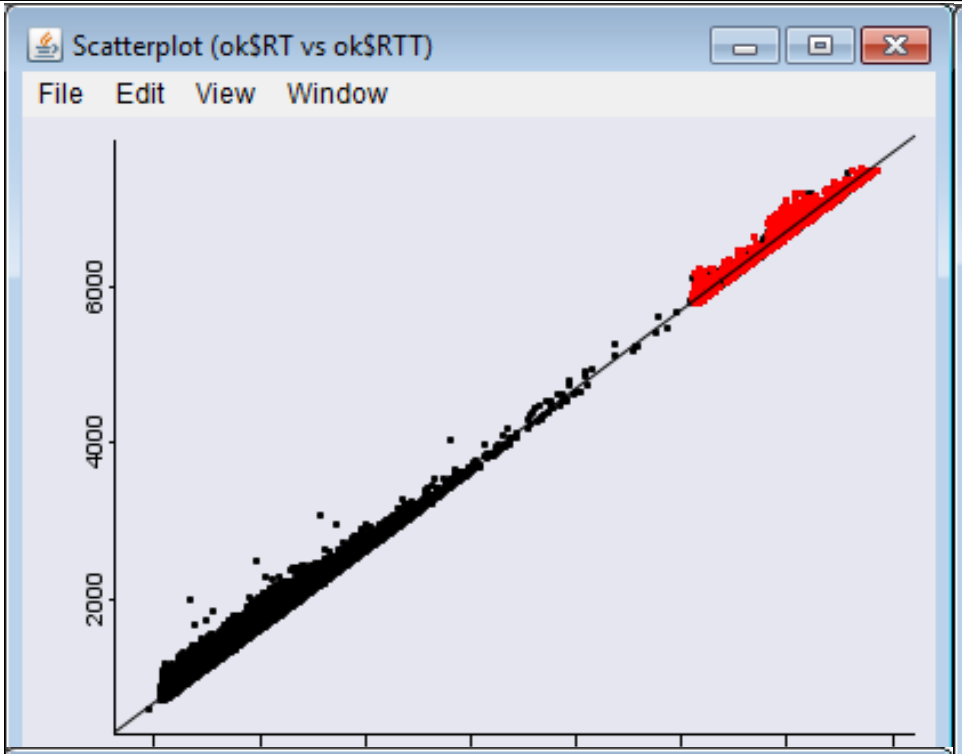
DEMO Gyanús kliens: csak Lansing

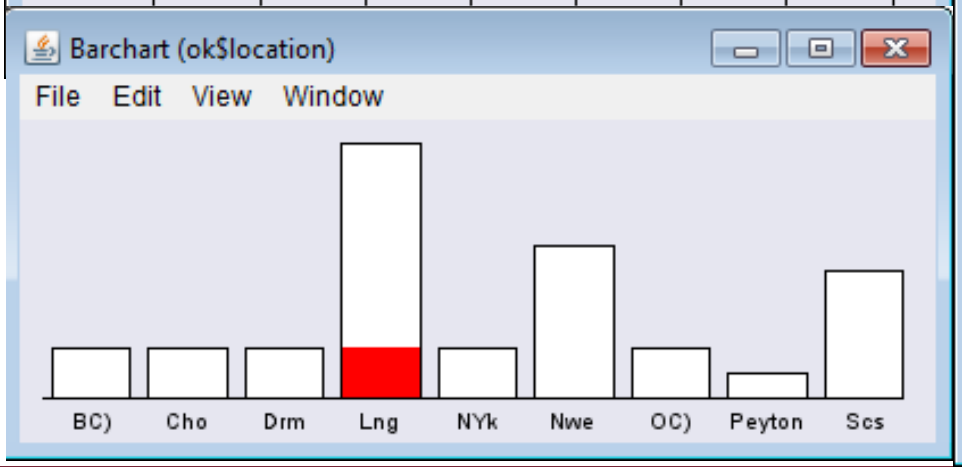
Color Brush:
View → Set Colors





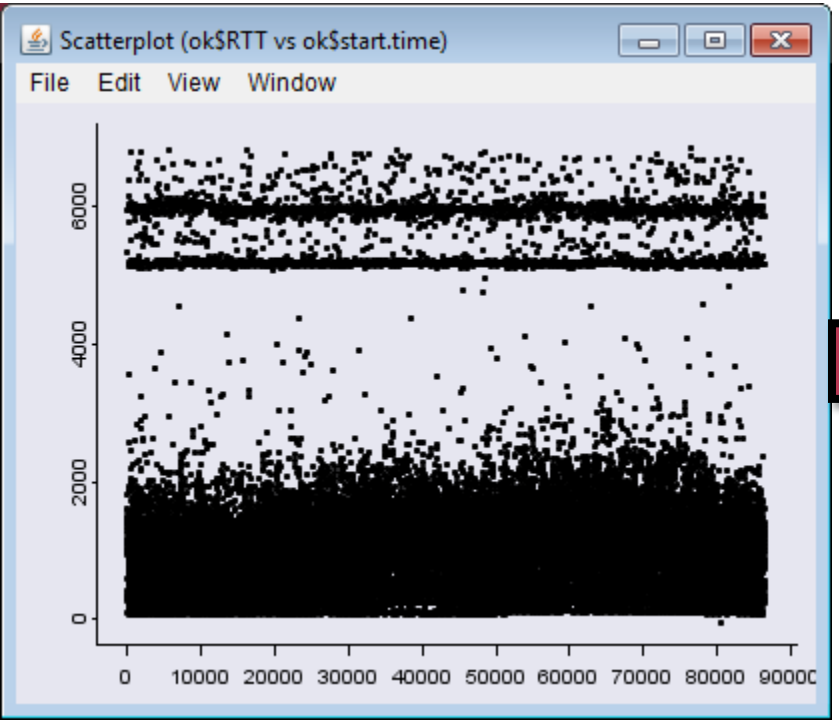
DEMO Gyanús kliens: csak Lansing 4!



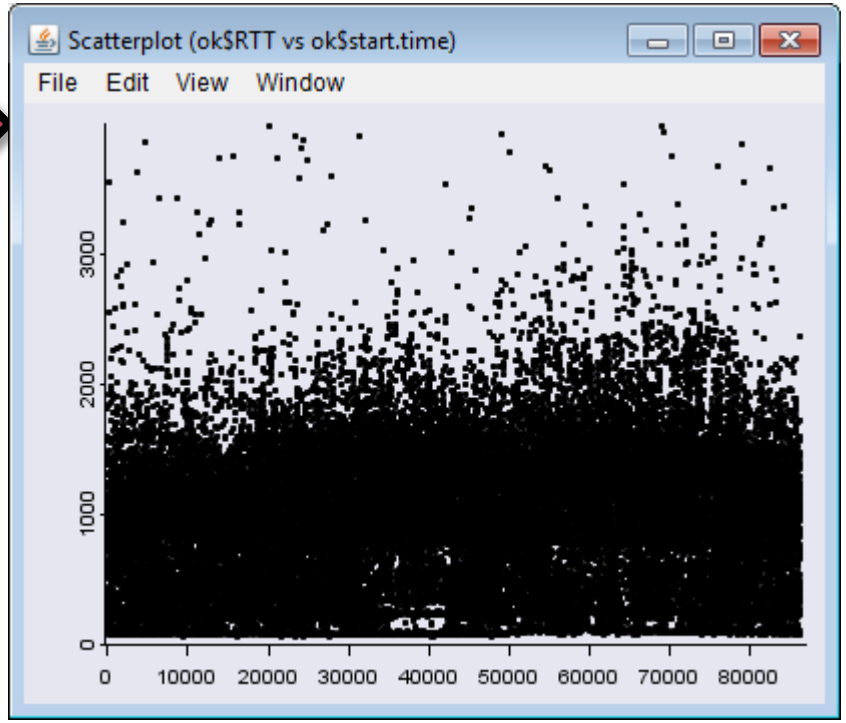


DEMO Időfüggő hálózati viselkedés

```
> ibar(ok$location)
ID:2 Name: "Barchart (ok$location)"
> iplot(ok$start.time, ok$RTT)
ID:3 Name: "Scatterplot (ok$RTT vs ok$start.time)"
```

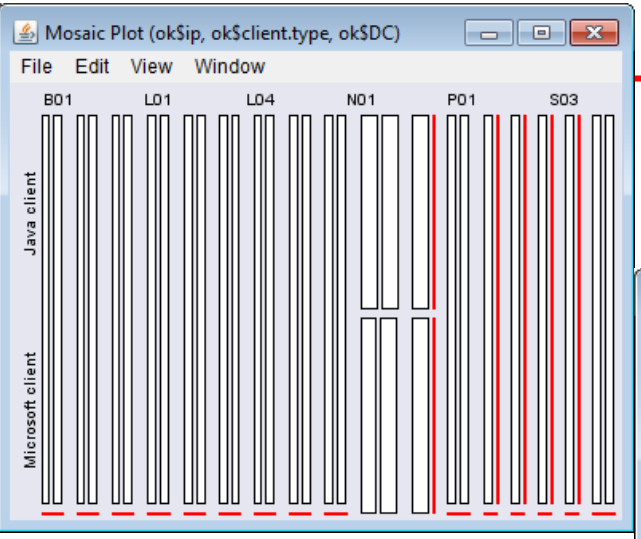


Zoom

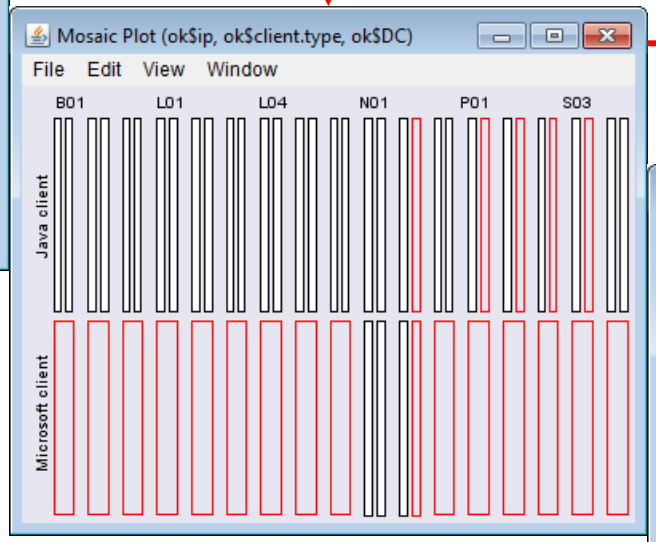


DEMO Időfüggő hálózati viselkedés

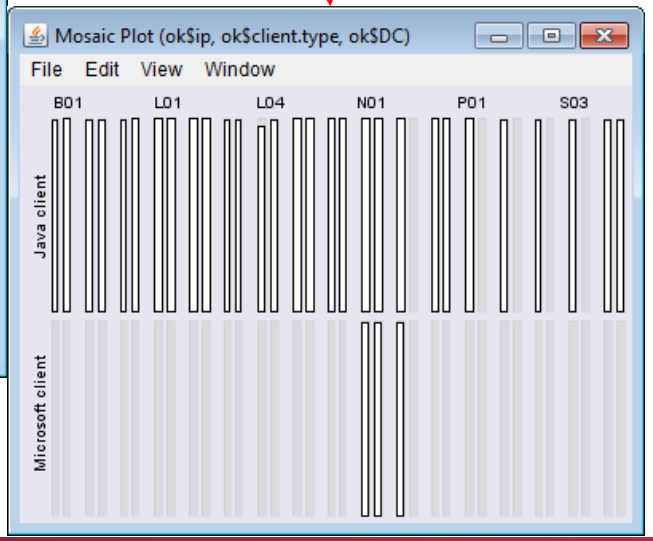
```
> iplot(ok$start.time, ok$RTT)
ID:3 Name: "Scatterplot (ok$RTT vs ok$start.time)"
> imosaic(ok$ip, ok$client.type, ok$DC)
ID:4 Name: "Mosaic plot (ok$ip, ok$client.type, ok$DC)"
>
```



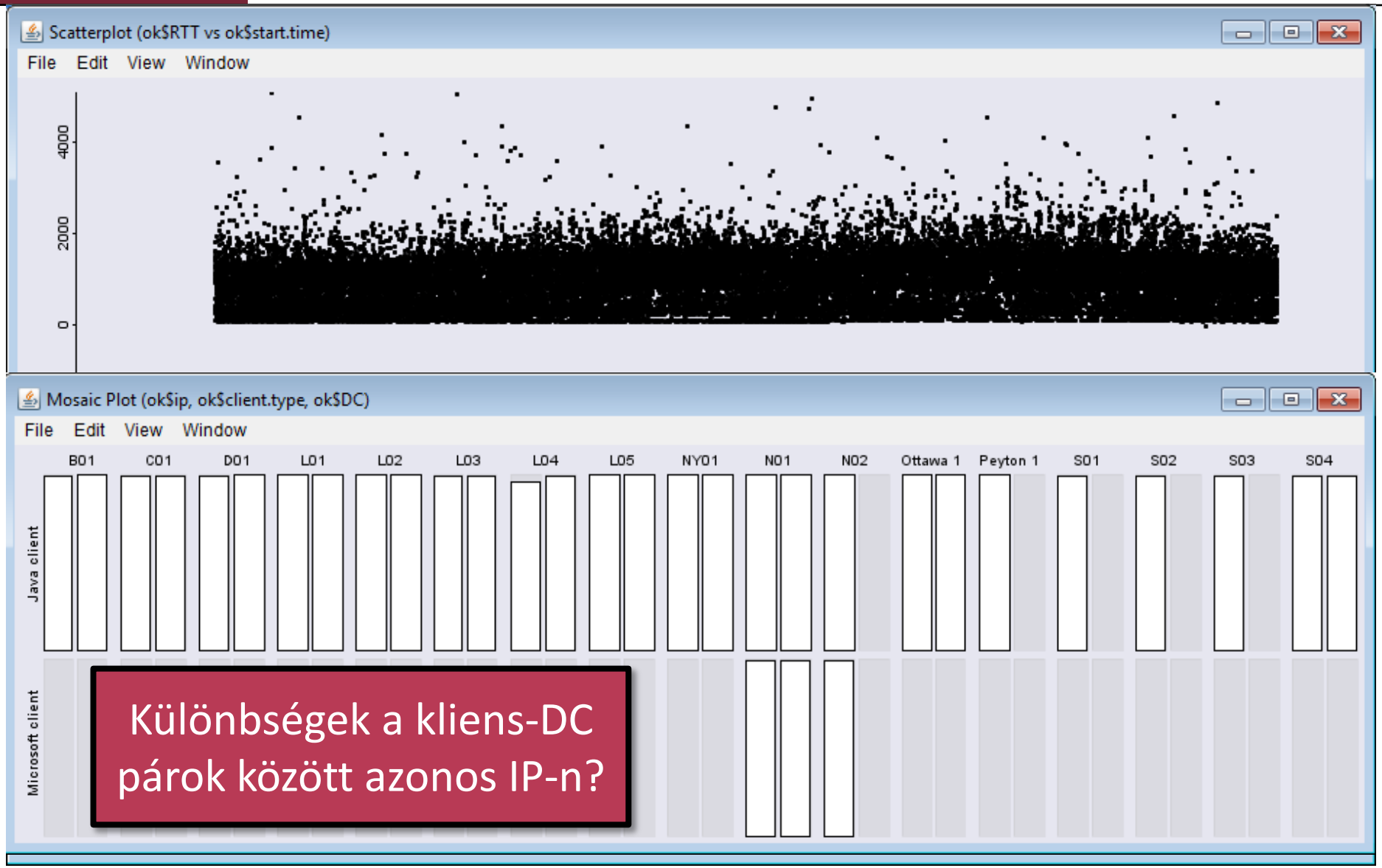
Azonos csempeméret:
View → Same bin size



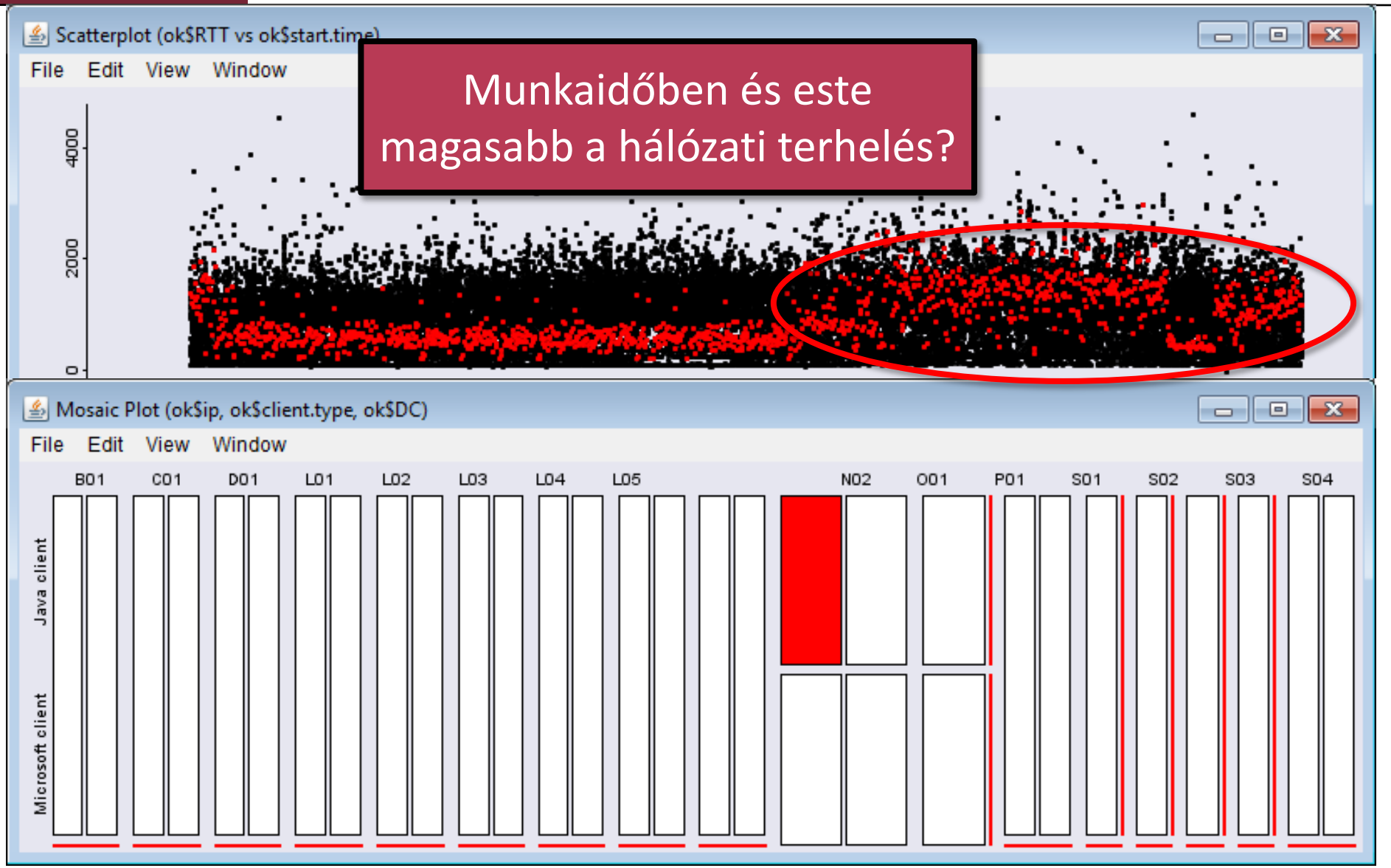
Flukt.-diagram:
View → Fluctuation



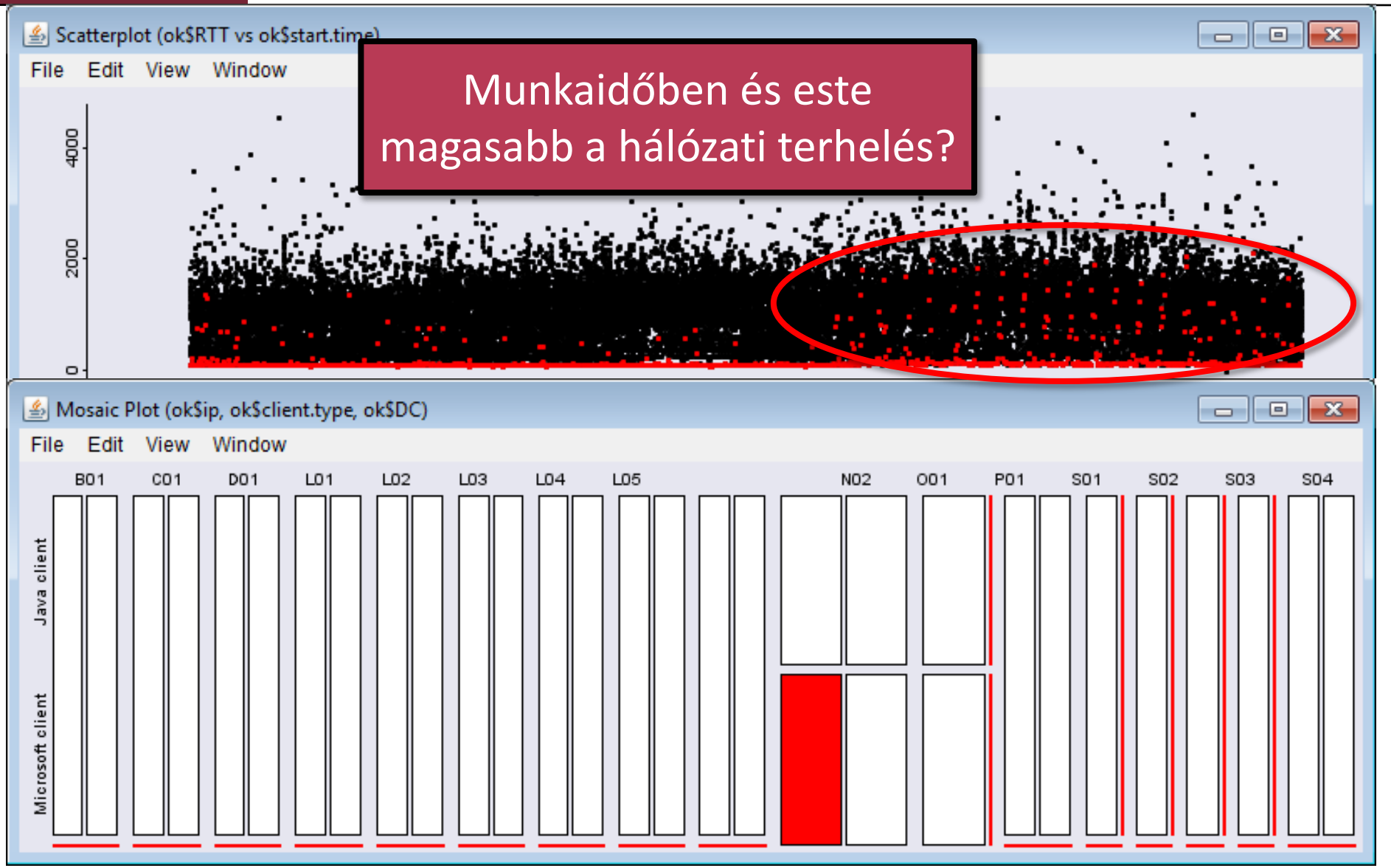
DEMO Időfüggő hálózati viselkedés



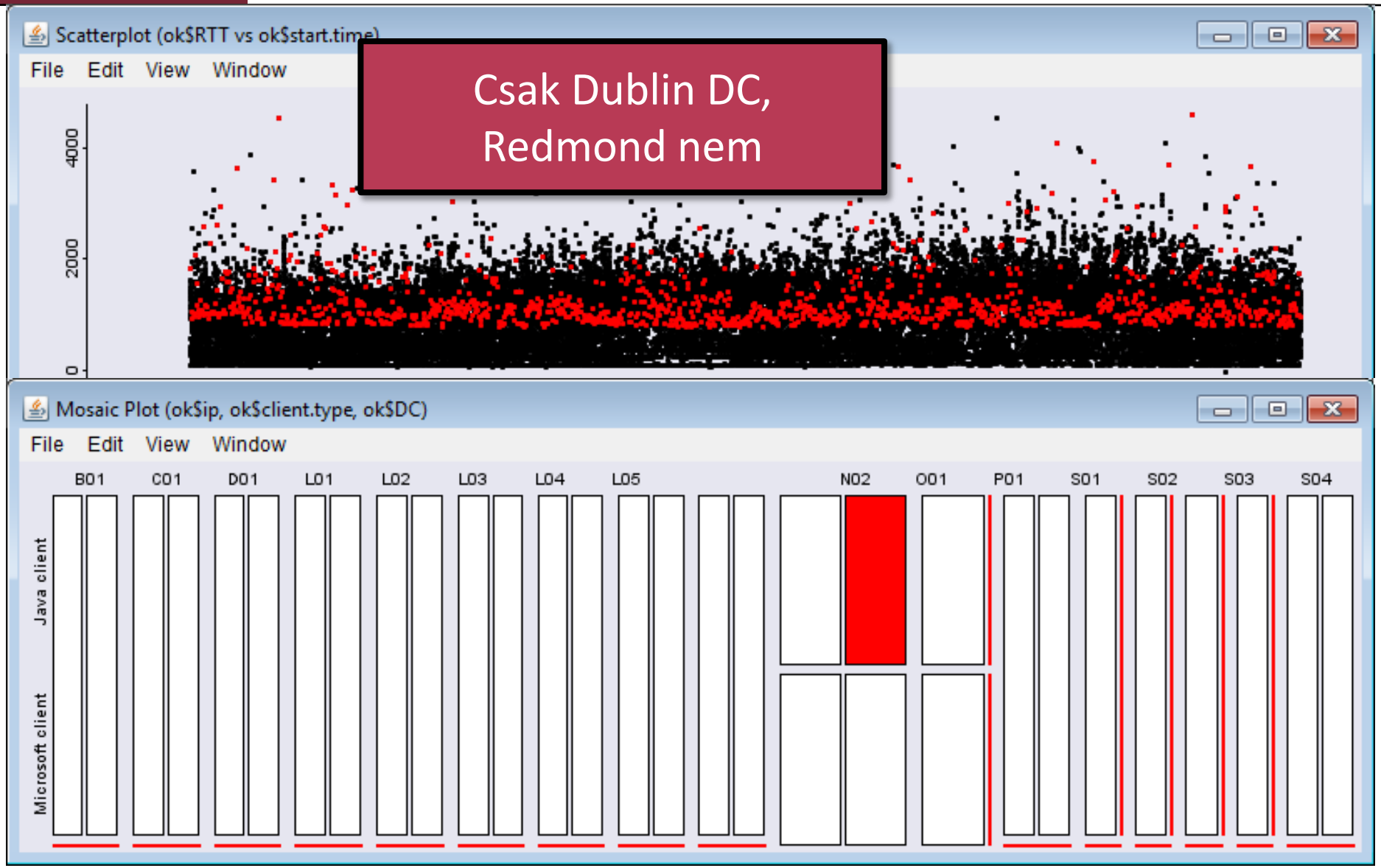
DEMO Időfüggő hálózati viselkedés



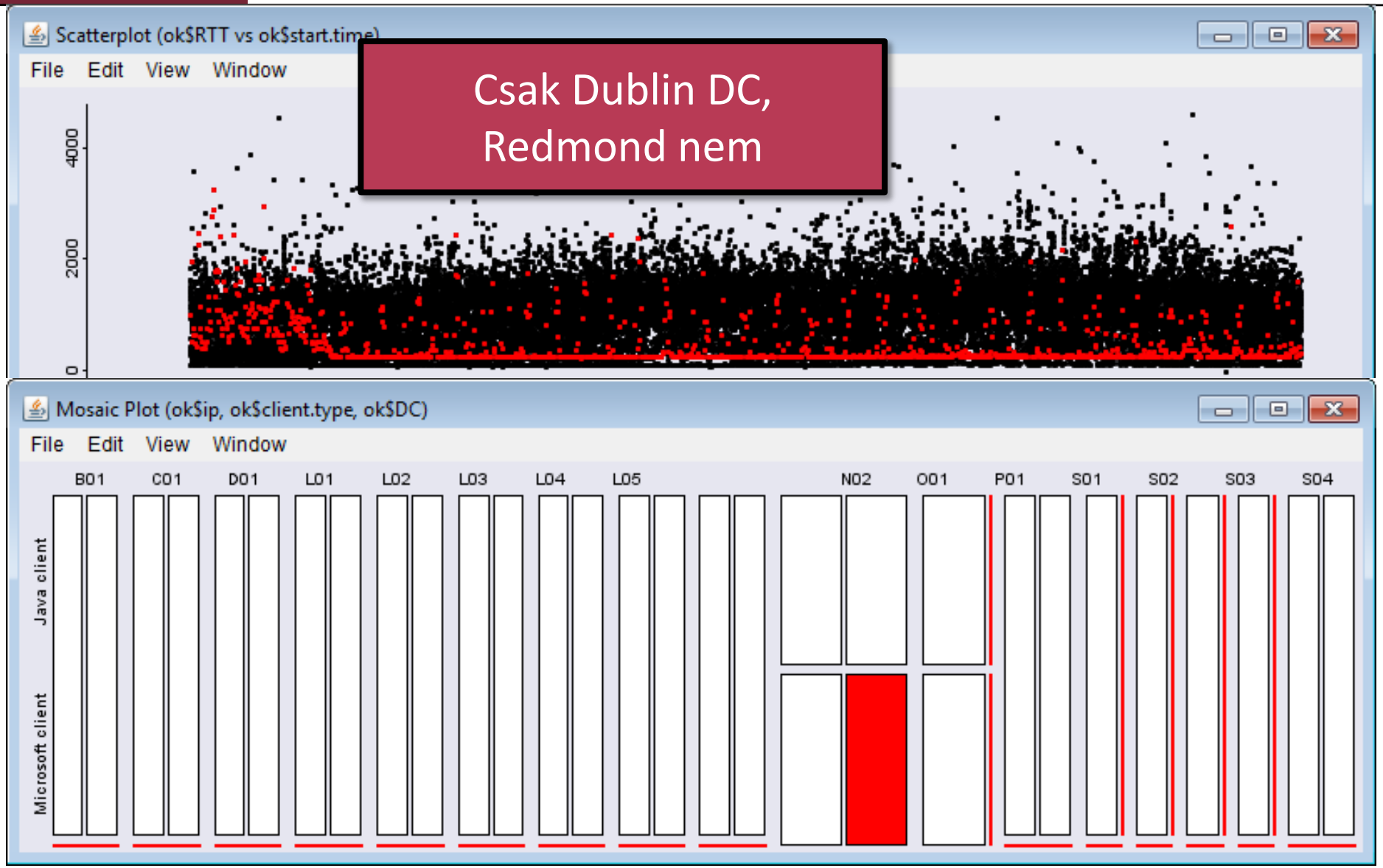
DEMO Időfüggő hálózati viselkedés



DEMO Időfüggő hálózati viselkedés



DEMO Időfüggő hálózati viselkedés

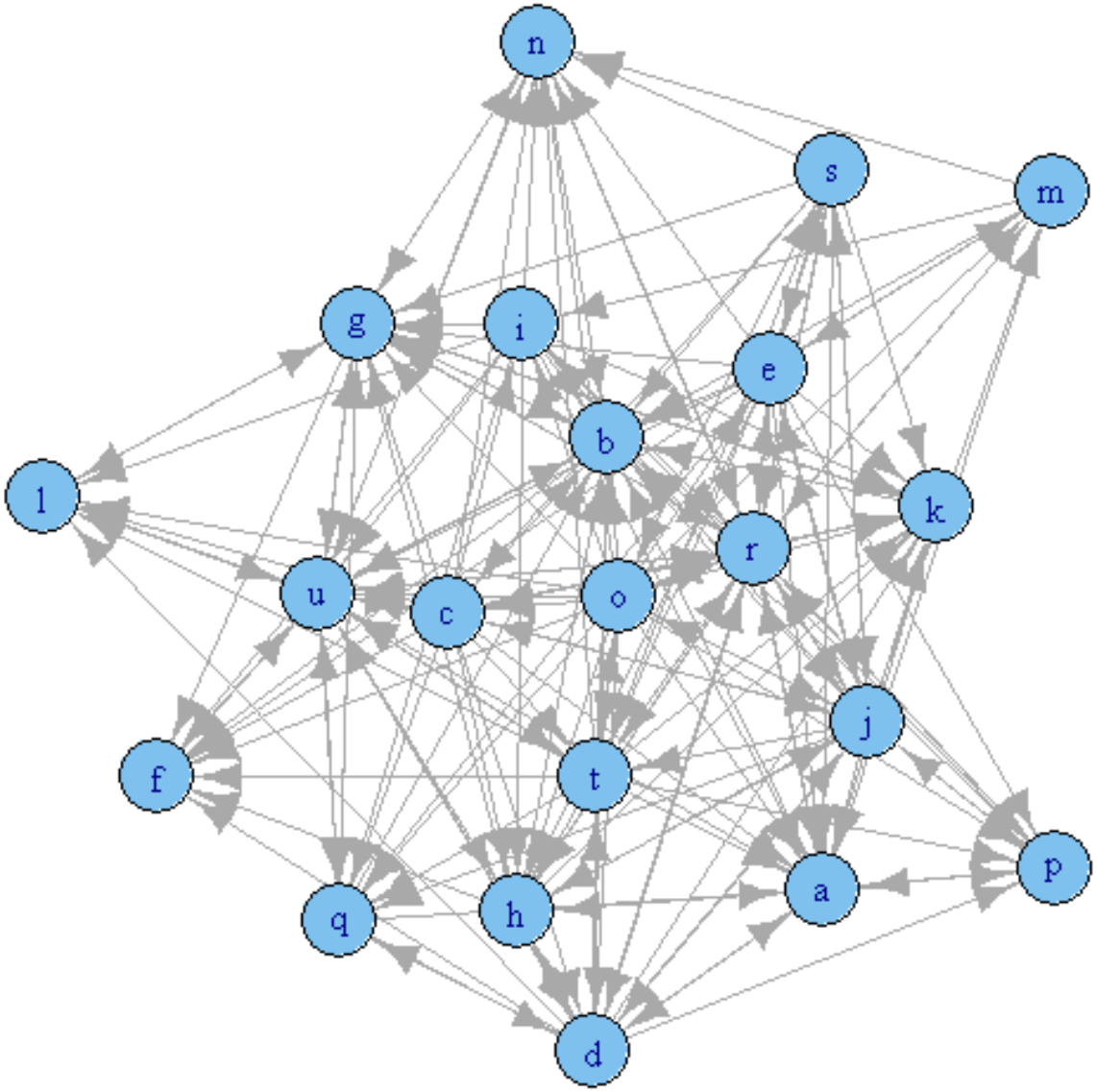


Fájó pontok

- Legalább Biggish Data?!?
  - OpenGL/DirectX
  - Statisztikai előfeldolgozás az adatokhoz közel?
- „Recordable EDA” =/= „reproducible research”
- rapporter.net, knitr, sweave, …:
  - A végeredmény
  - Folyamat kézi visszakövetése és átemelése

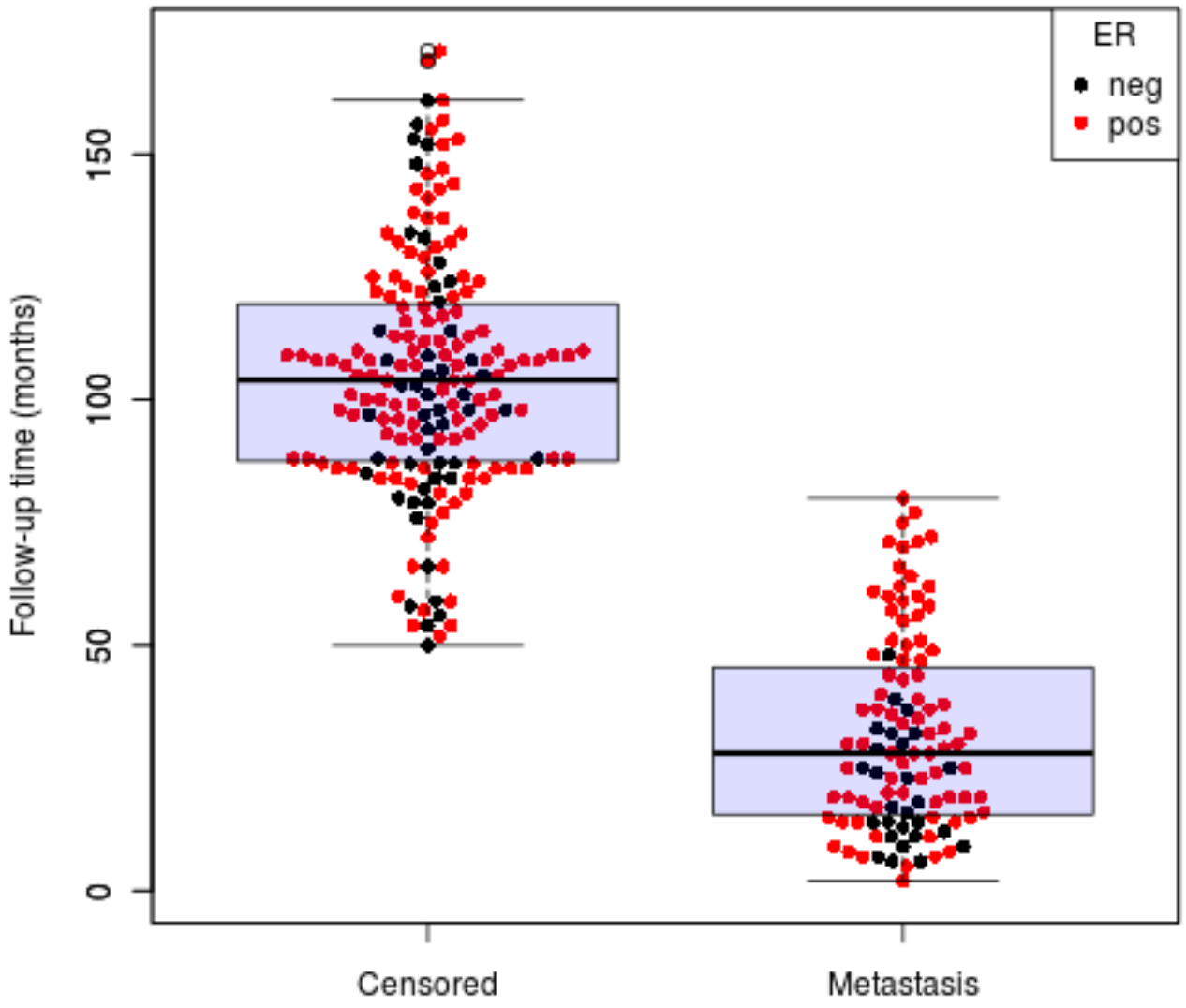
MEGJELENÍTŐ VIZUALIZÁCIÓ

Gráfok



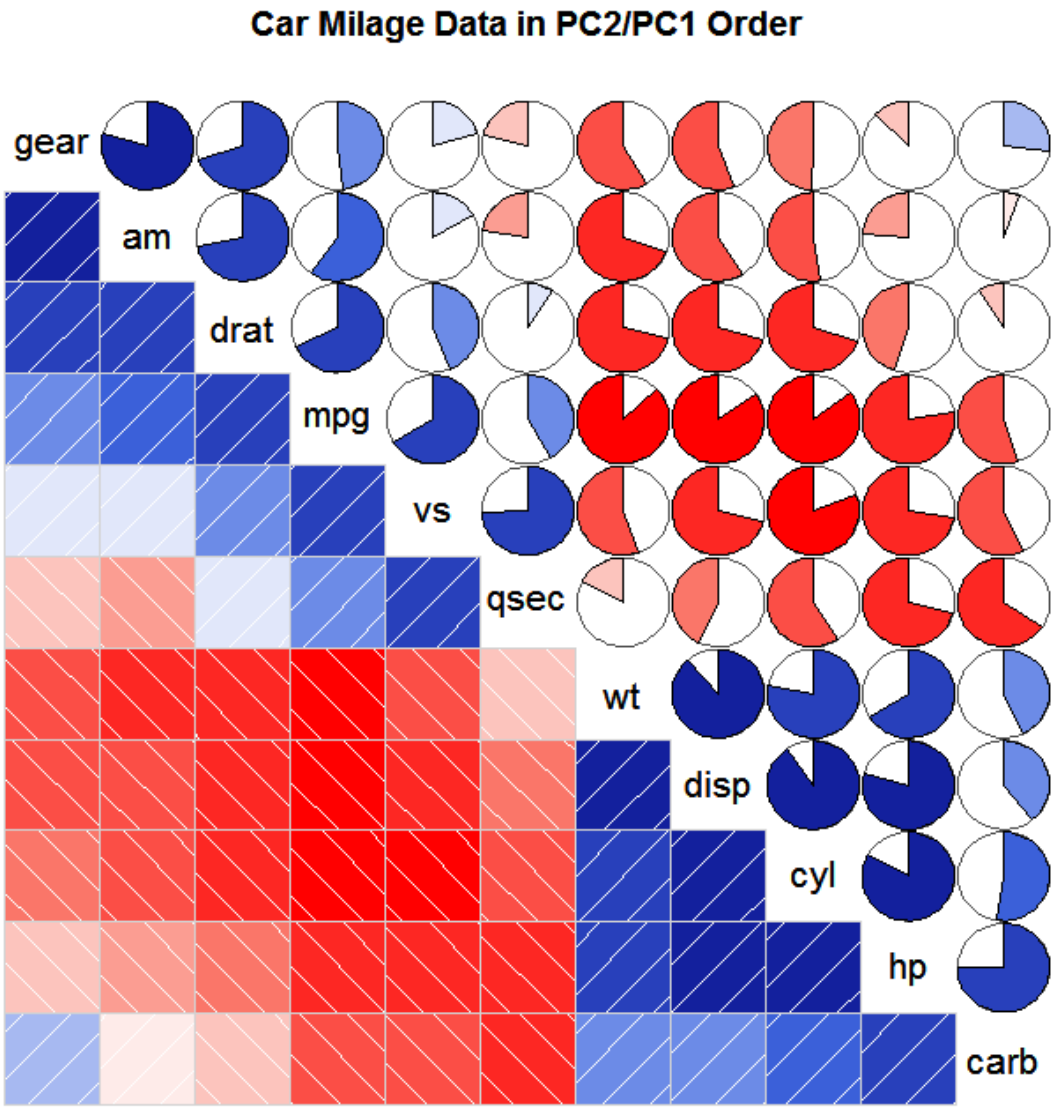
Rgraphviz

Beeswarm



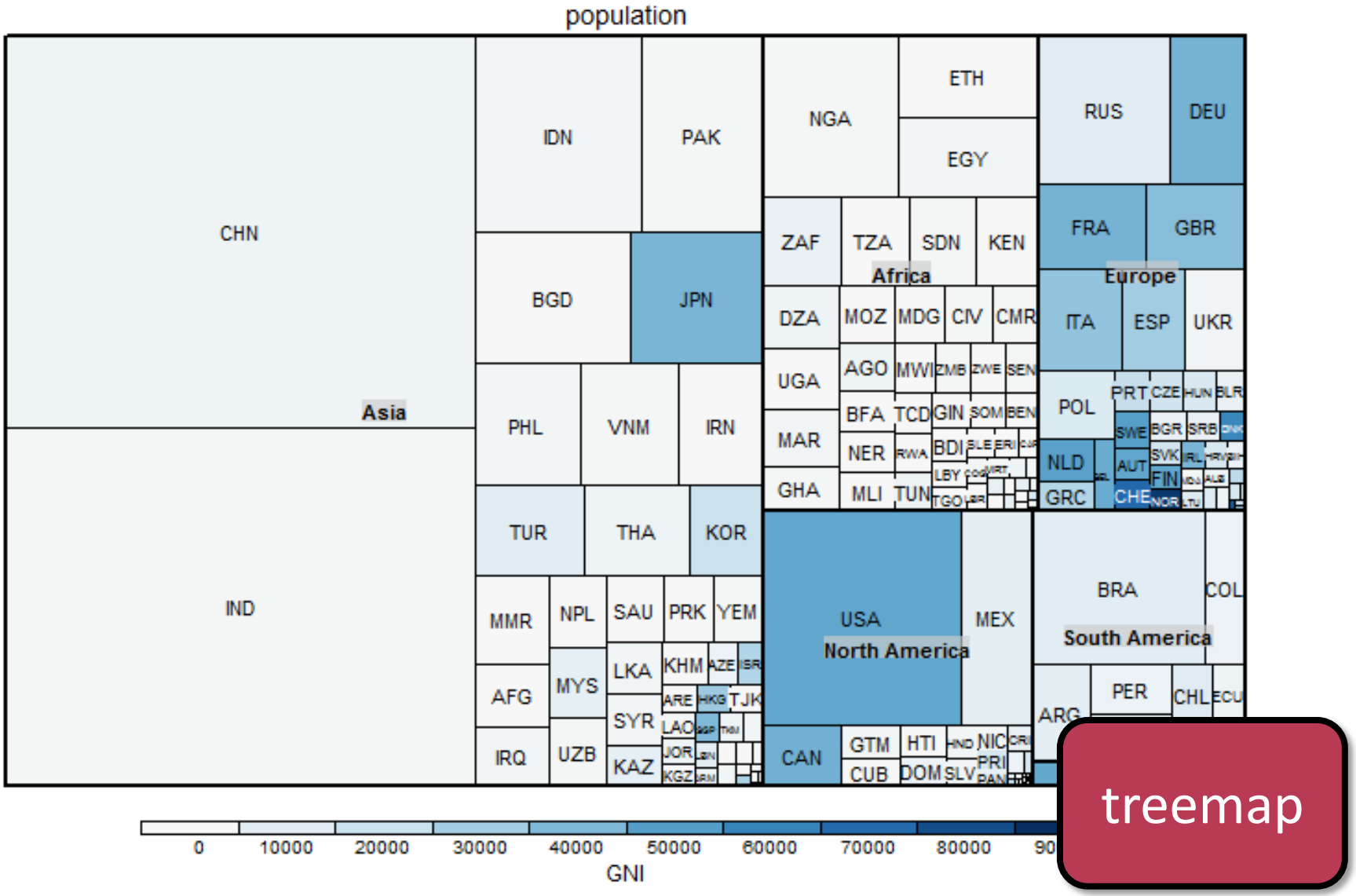
beeswarm

Korrelogram



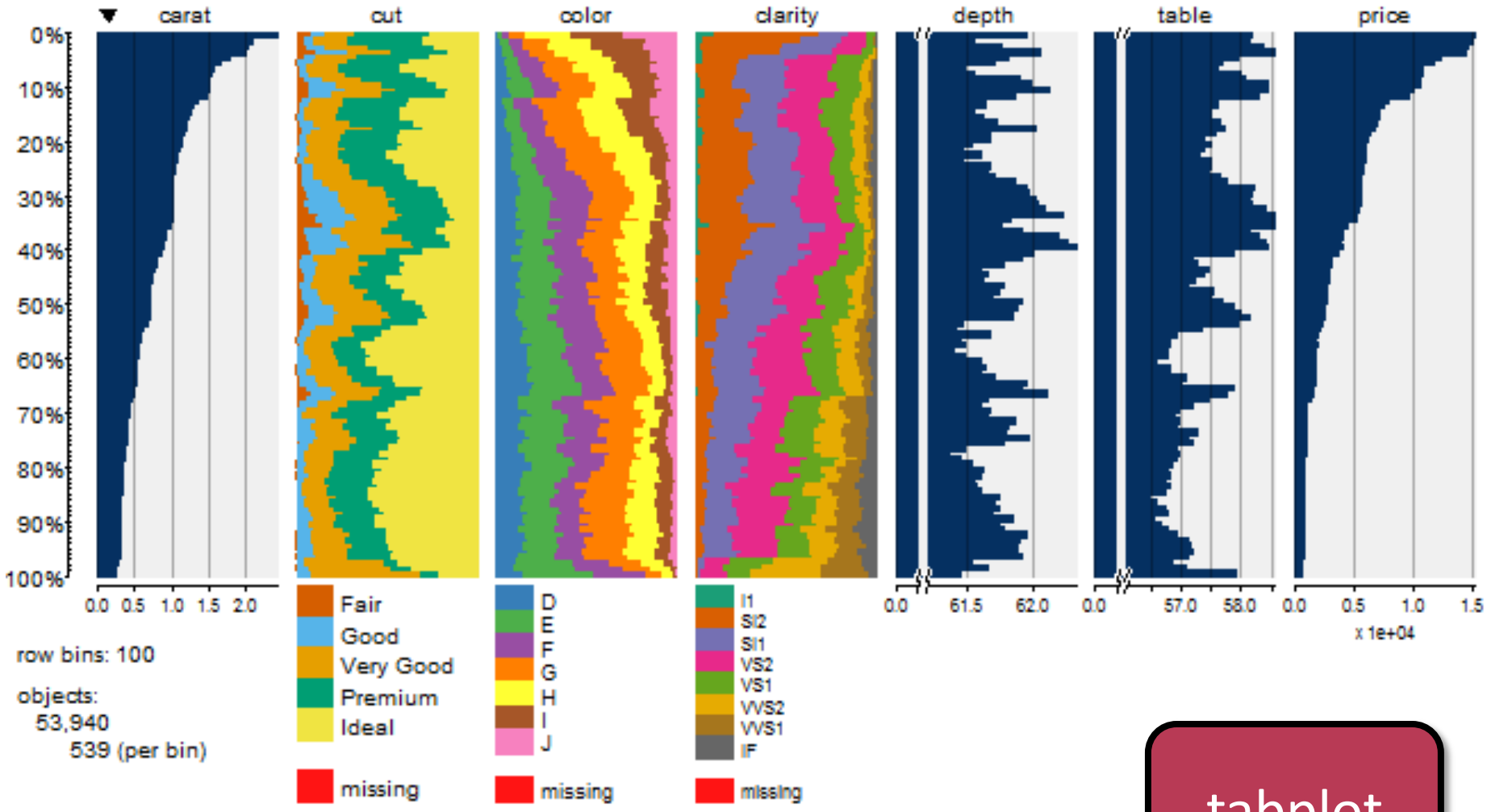
corrgram

Treemap



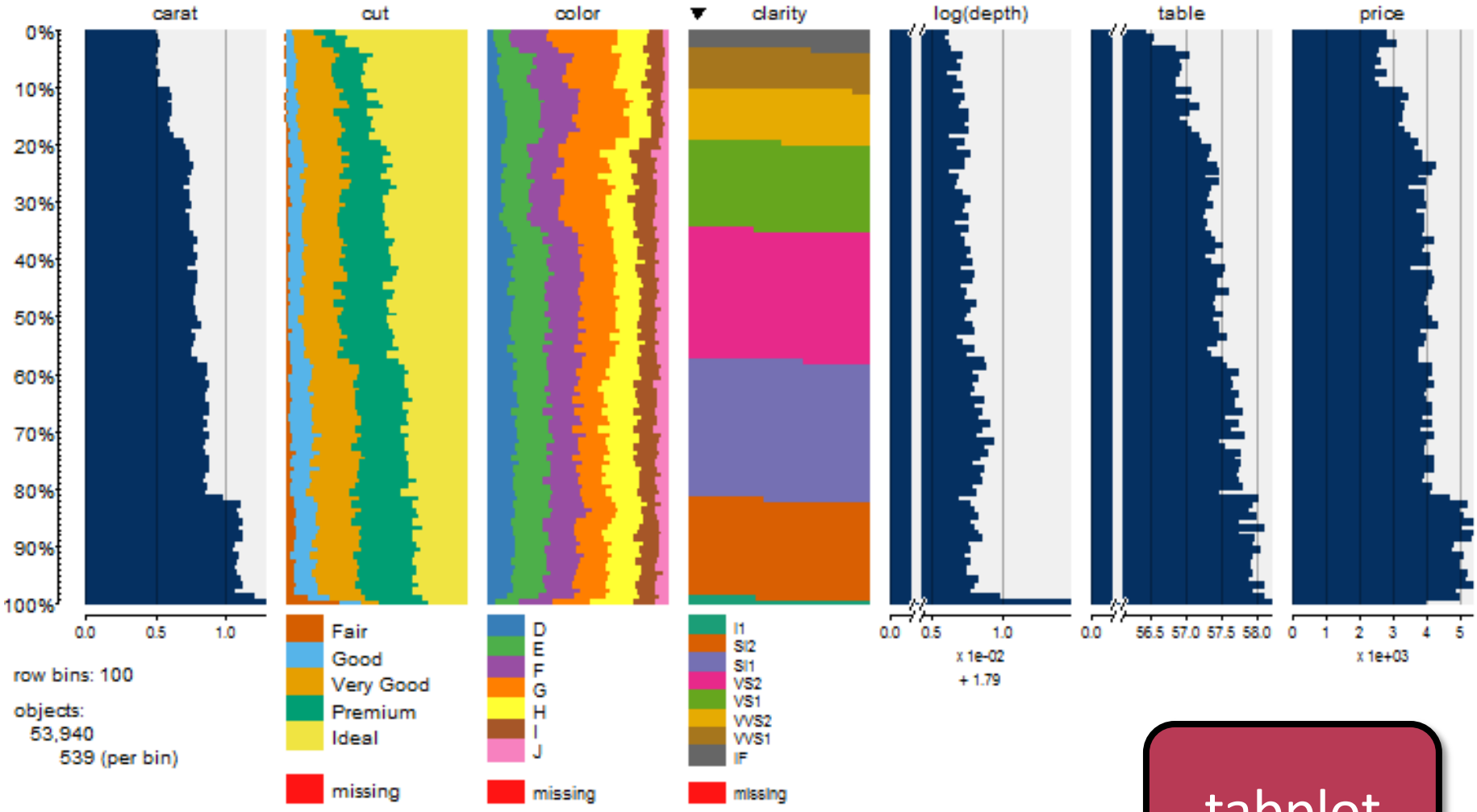
treemap

Tableplot



tabplot

Tableplot



tabplot

Hivatkozások

- [1] http://xkcd.com/435/
- [2] Behrens, J.T.: Principles and procedures of exploratory data analysis. Psychological Methods 2, 131–160 (1997)
- [3] Tukey, J.: We need both exploratory and confirmatory. The American Statistician 34, 23–25 (1980)
- [4] Anscombe, F. J.: Graphs in Statistical Analysis. American Statistician 27 (1): 17–21 (1973)
- [5] McLeod, K.S.: Our sense of Snow: the myth of John Snow in medical geography. Social Science and Medicine 50 (7-8): 923-935 (2000)
- [6] Inselberg, A.: Parallel Coordinates: Visual Multidimensional Geometry and its Applications. Springer Science+Business Media, New York (2009)
- [7] Theus, M., Urbanek, S.: Interactive graphics for data analysis: principles and examples. CRC Press (2011)
- [8] 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)
- [9] Pataricza, András, et al.: Empirical Assessment of Resilience. Software Engineering for Resilient Systems. 1-16. (2013)
- [10] http://streaming.stat.iastate.edu/~dicook/Reykjavik/vis/notes/6-interactive.pdf
- [11] http://en.wikipedia.org/wiki/Spline_interpolation
- [12] http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-nonparametric-regression.pdf