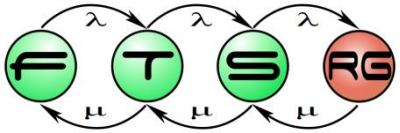
Nagyméretű adathalmazok vizualizációja

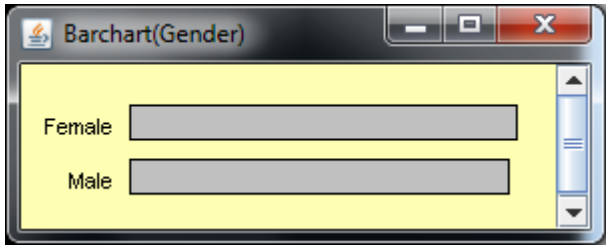
„Big Data” elemzési módszerek

Salánki Ágnes, Kocsis Imre
salanki, ikocsis@mit.bme.hu
2015.10.22.

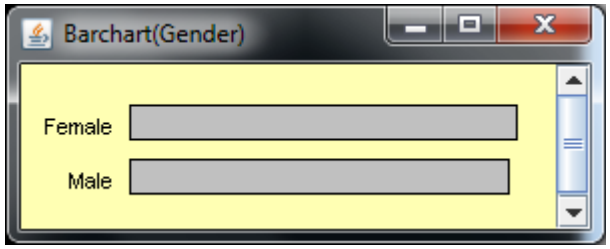


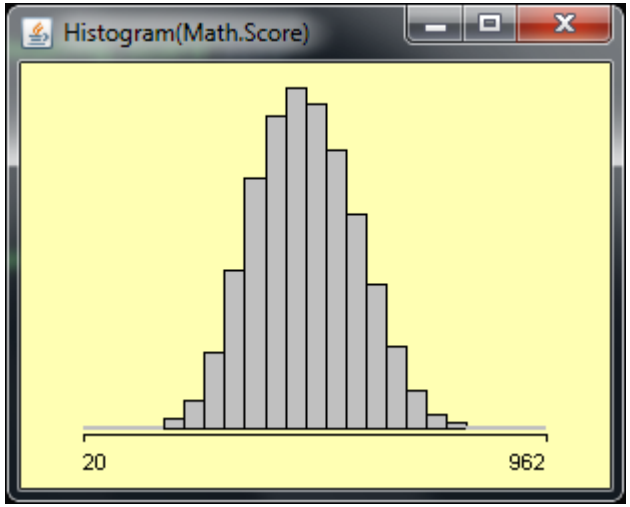
Alapvető vizualizációs eszközök

Alapvető vizualizációs eszközök

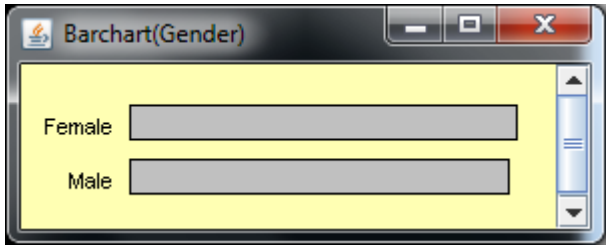


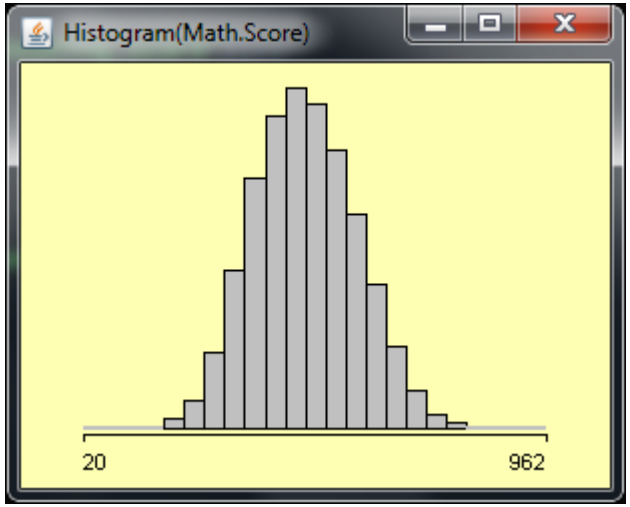
Alapvető vizualizációs eszközök

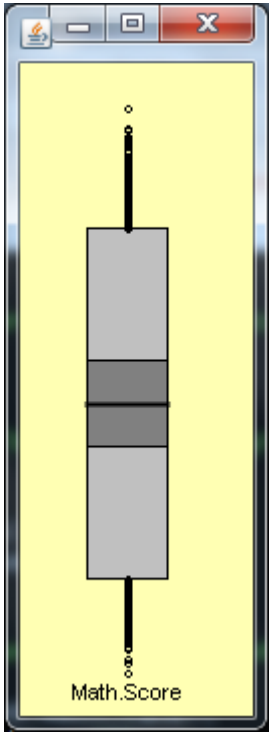




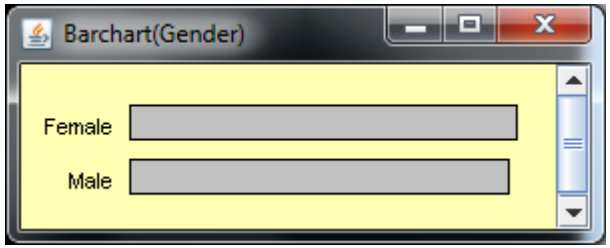
Alapvető vizualizációs eszközök

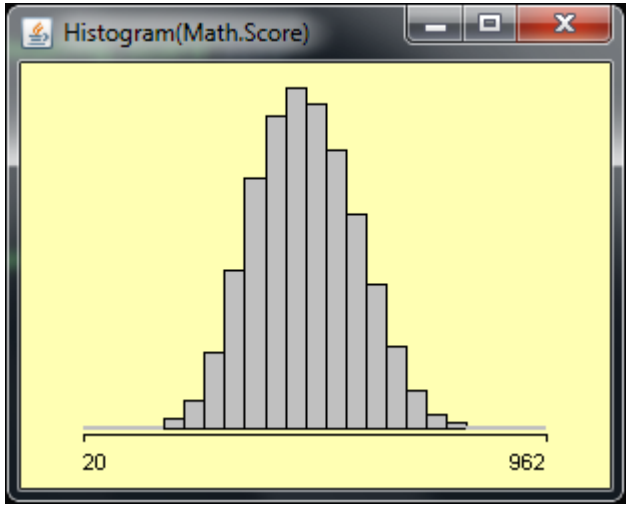


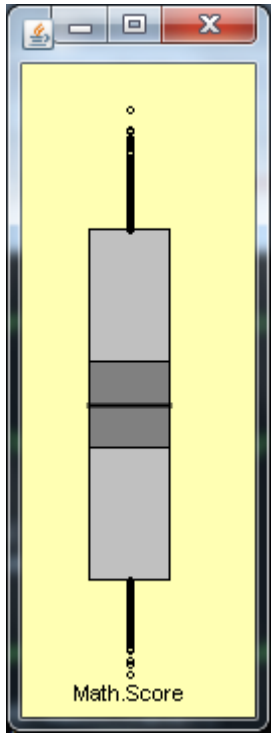


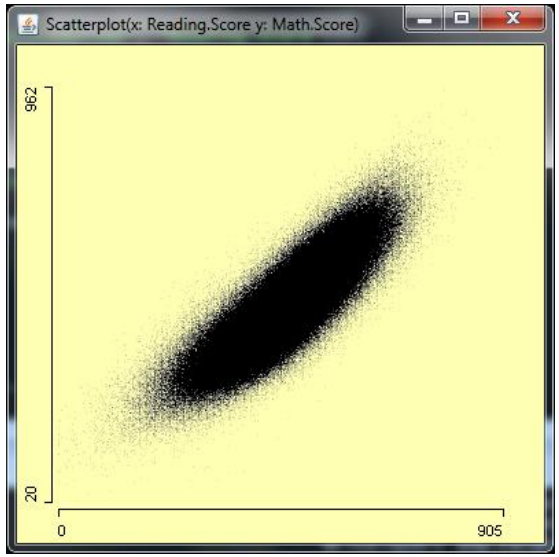


Alapvető vizualizációs eszközök

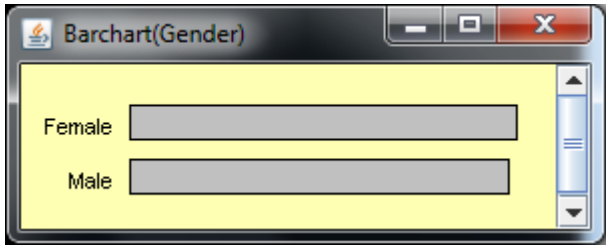


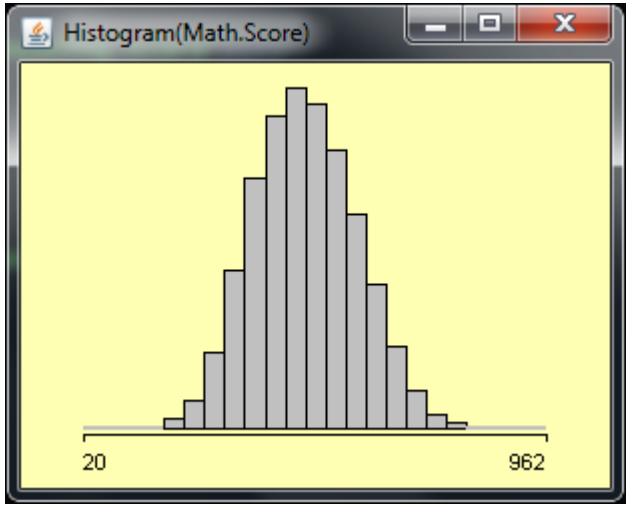


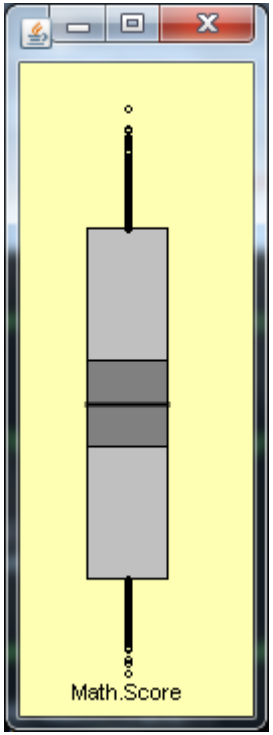


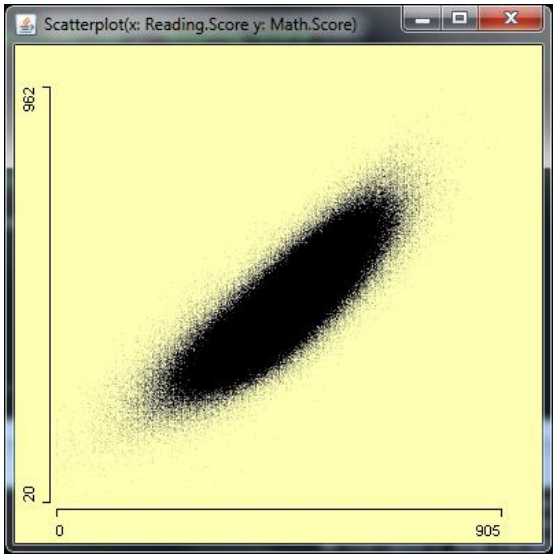


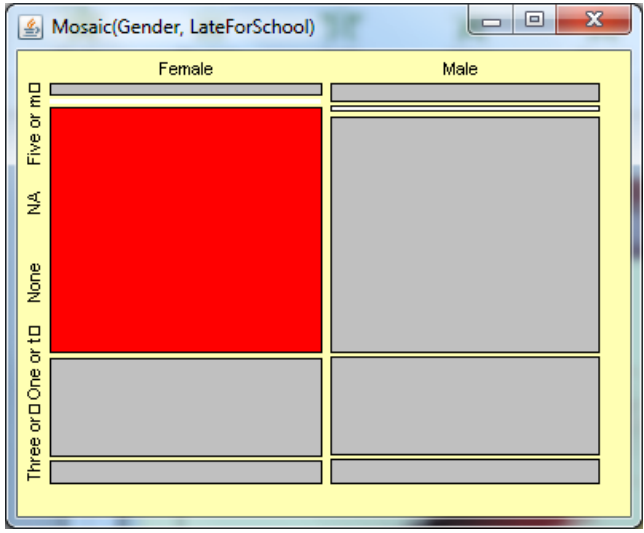
Alapvető vizualizációs eszközök



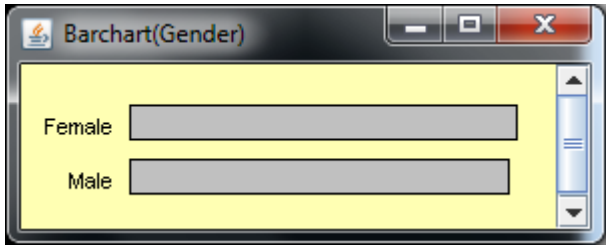


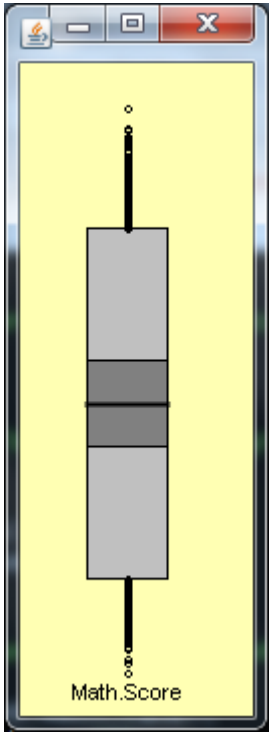


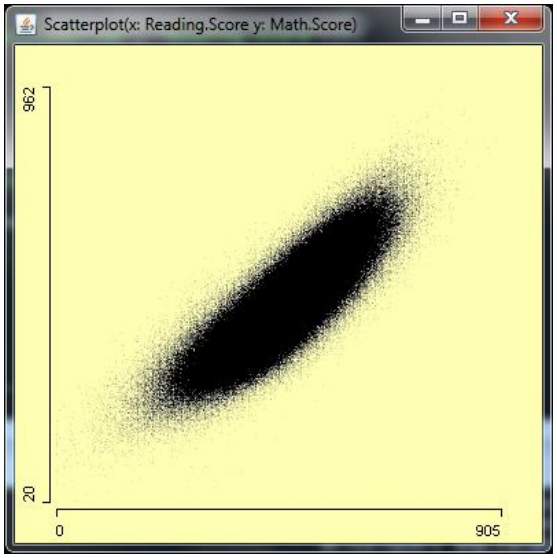


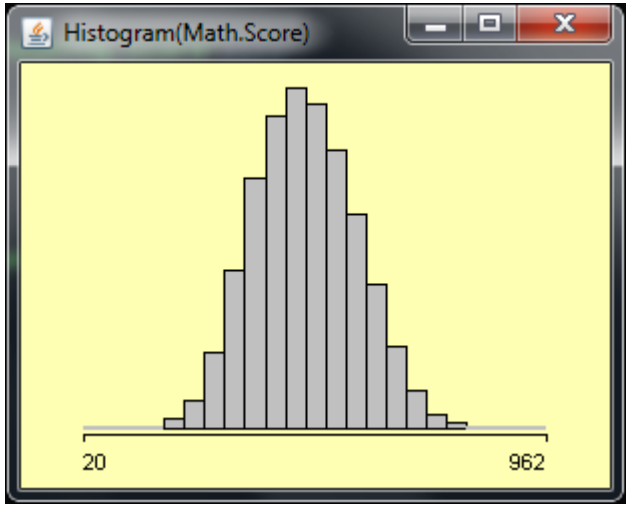


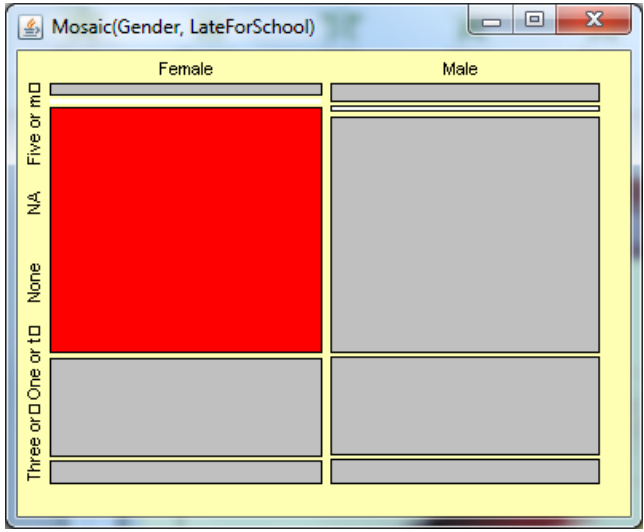
Alapvető vizualizációs eszközök

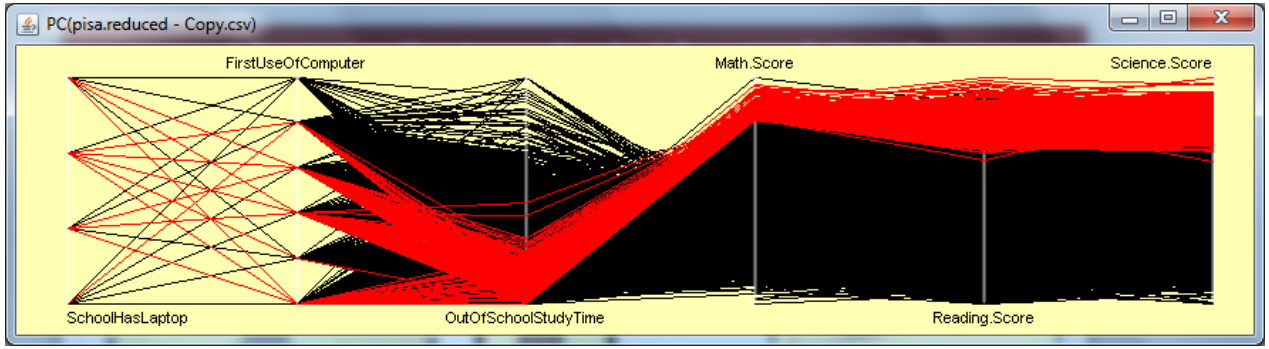












Nagyméretű adathalmazok vizualizációja

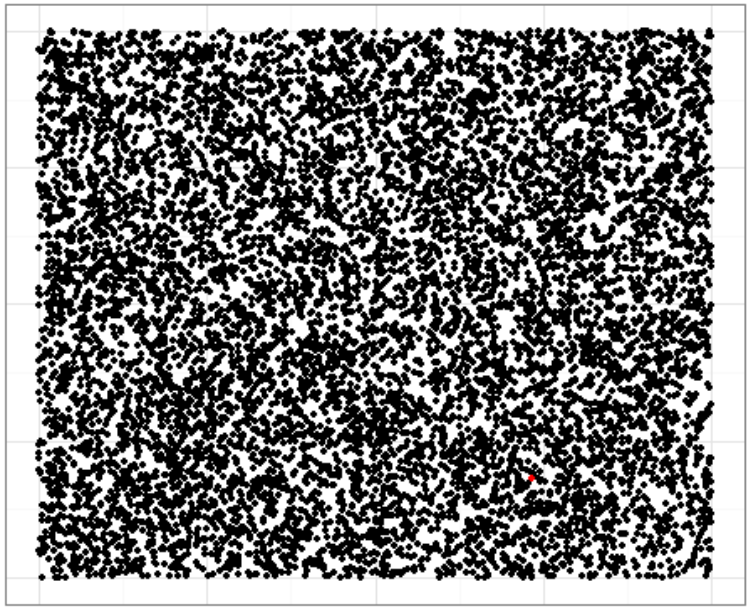
- Nem sokdimenziós,
  - Dimenziócsökkentő módszerek
    - PCA
    - MDS
    - stb.
- hanem sok megfigyelést tartalmazó adatok
  - Számolni úgyis sokáig kell
  - Megjelenítés?

Miért nehéz?

- „Visual bias of humans”
- Nagyon nagy számú objektumot általában nehéz megkülönböztetni
- Véges képernyőméret
- Ha úgyis csak $X \times Y$ pixelt rajzolunk ki, akkor felesleges kiszámolni mindent aztán meg kerekíteni

Miért nehéz?

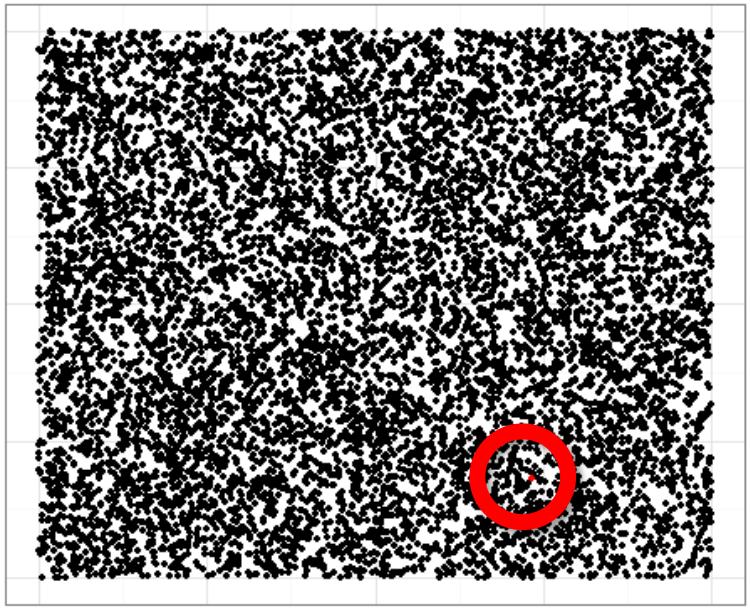
- „Visual bias of humans”
- Nagyon nagy számú objektumot általában nehéz megkülönböztetni



- Véges képernyőméret
- Ha úgyis csak $X \times Y$ pixelt rajzolunk ki, akkor felesleges kiszámolni mindent aztán meg kerekíteni

Miért nehéz?

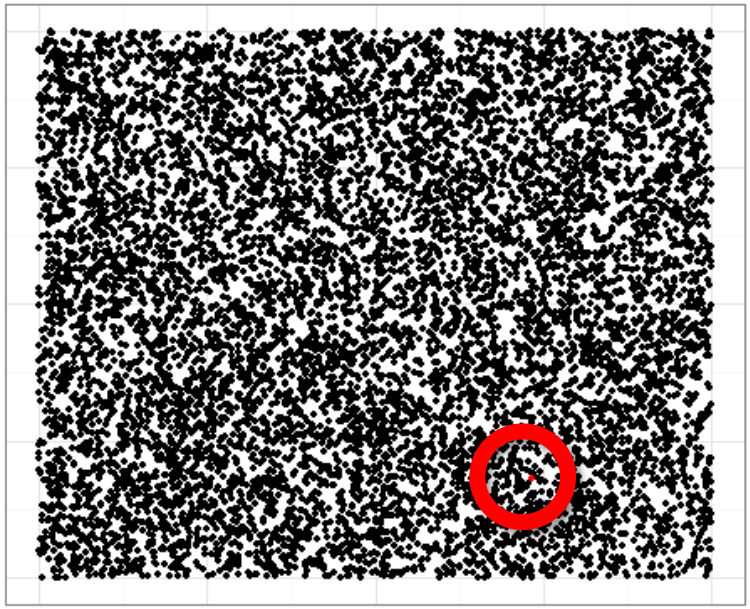
- „Visual bias of humans”
- Nagyon nagy számú objektumot általában nehéz megkülönböztetni



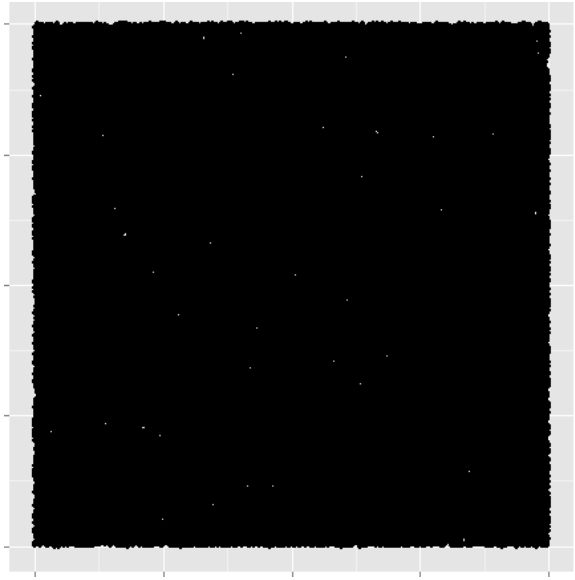
- Véges képernyőméret
- Ha úgyis csak $X \times Y$ pixelt rajzolunk ki, akkor felesleges kiszámolni mindent aztán meg kerekíteni

Miért nehéz?

- „Visual bias of humans”
- Nagyon nagy számú objektumot általában nehéz megkülönböztetni



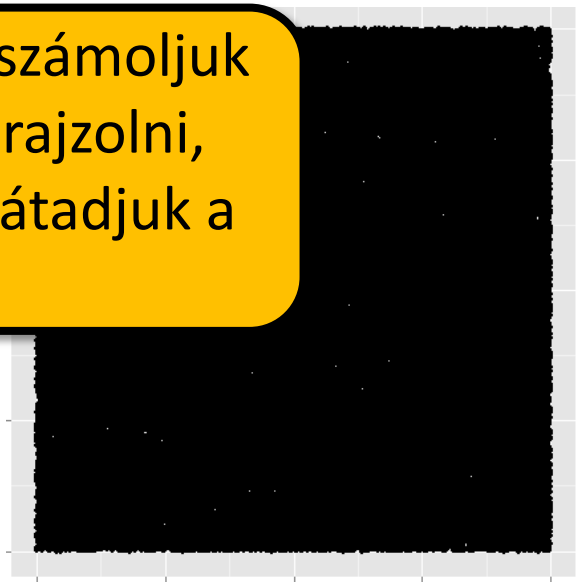
- Véges képernyőméret
- Ha úgyis csak $X \times Y$ pixelt rajzolunk ki, akkor felesleges kiszámolni mindent aztán meg kerekíteni



Miért nehéz?

- „Visual bias of humans”
- Nagyon nagy számú objektumot általában nehéz megkülönböztetni
- Véges képernyőméret
- Ha úgyis csak $X \times Y$ pixelt rajzolunk ki, akkor felesleges kiszámolni mindent aztán meg kerekíteni

A klasszikus megoldás: előbb számoljuk ki *pontosan*, mit kell majd kirajzolni, majd a kirajzolófüggvénynek átadjuk a konkrétumokat



Esettanulmány

- PISA 2012

  > colnames(pisa2012)

  [1] "CNT" "OECD" "**BirthMonth**" "**BirthYear**"

  [5] "Gender" "LateForSchool"

  "Possessions.Computer" "EnjoysMath„

  [9] "ParentsLikeMath" "PlaysChess"

  "ComputerProgramming" "SchoolHasLaptop„

  [13] "**FirstUseOfComputer**" "EdCodeMother"

  "EdCodeFather" "**OutOfSchoolStudyTime**„

  [17] "**Math.Score**" "**Reading.Score**"

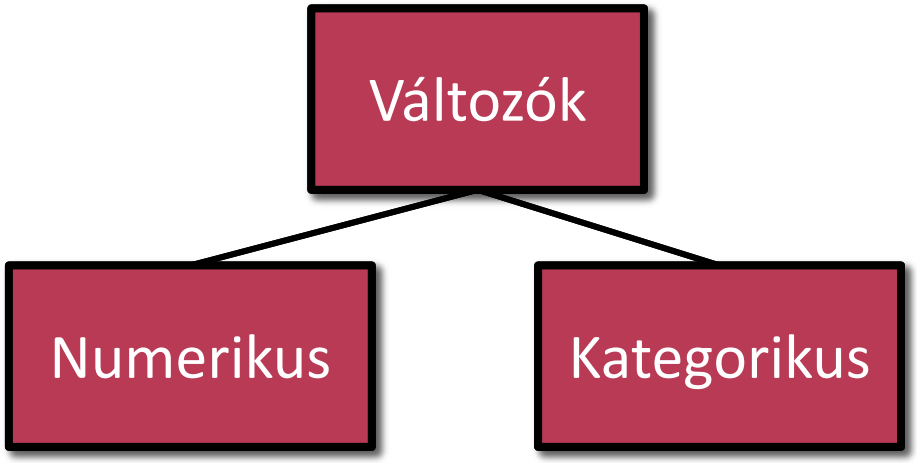
  "**Science.Score**"

Bigdata vizualizáció R alapokon

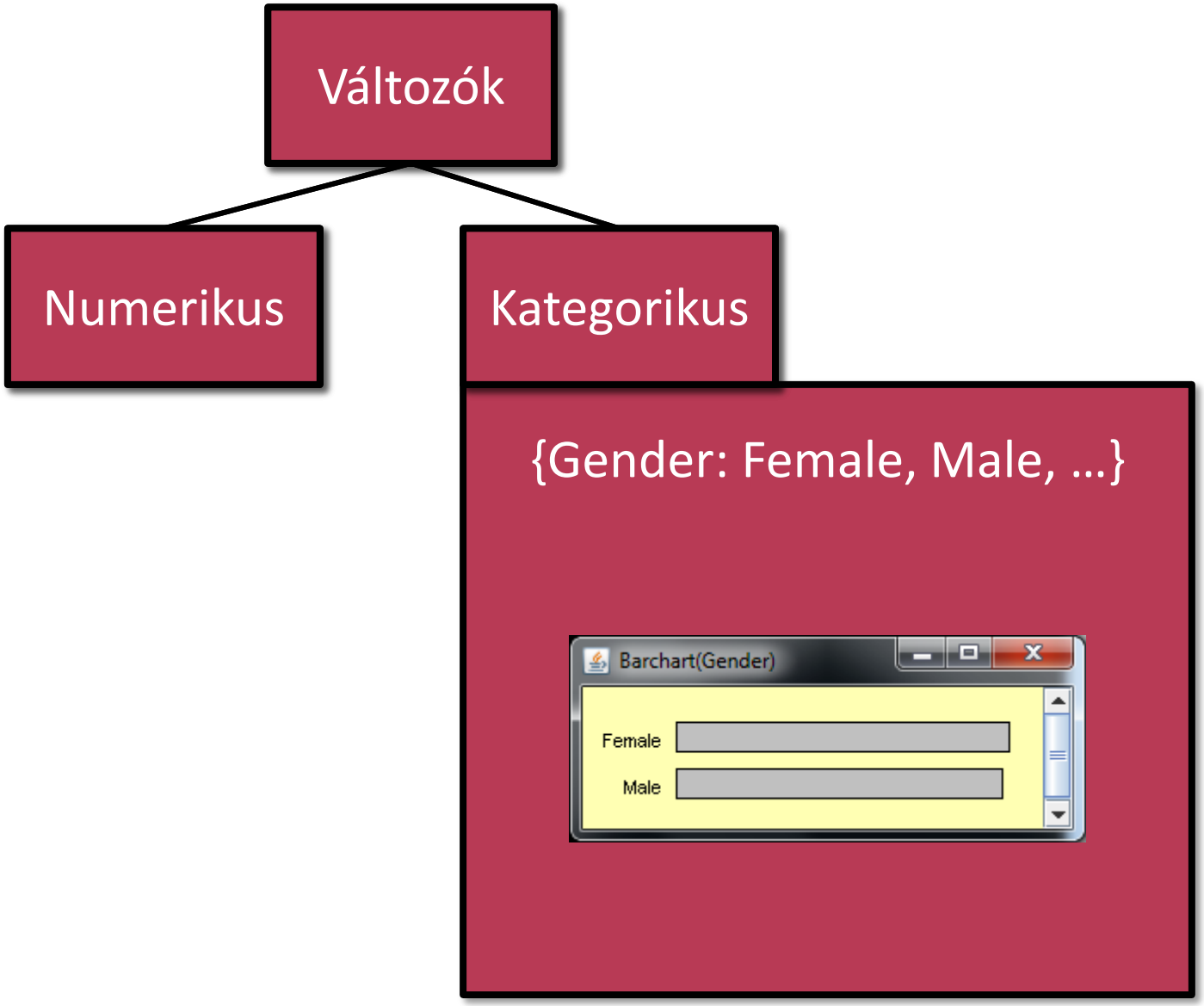
- bigvis
- tabplot
- trelliscope

BIGVIS

1 változó



1 változó

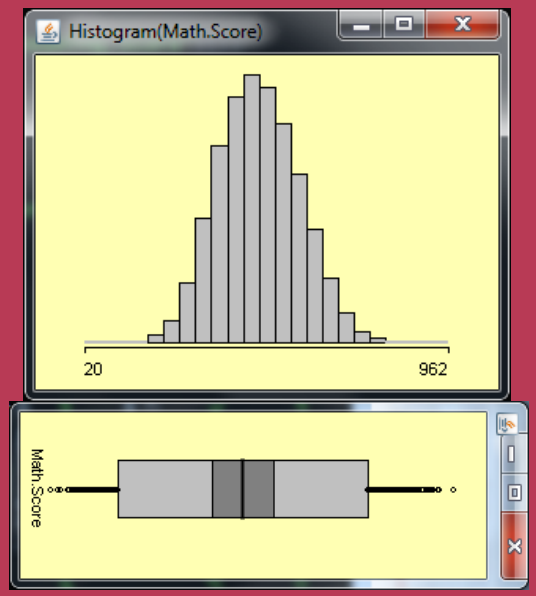

Változók
Numerikus
Kategorikus
{Gender: Female, Male, …}
Barchart(Gender)
Female
Male

1 változó

Változók

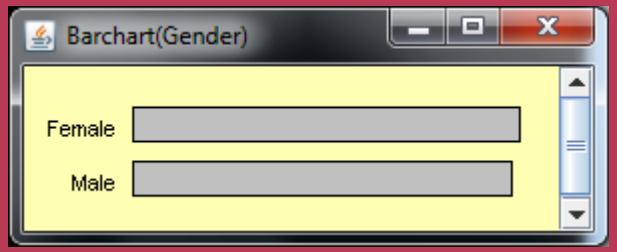
Numerikus

{Math.Score: 609, 613, … }



Kategorikus

{Gender: Female, Male, …}



Jellemző mérőszámok egyváltozós esetben

- Oszlopdiagram, hisztogram
  - abszolút számosság
- Boxplot
  - Kvartilisek + outlierek

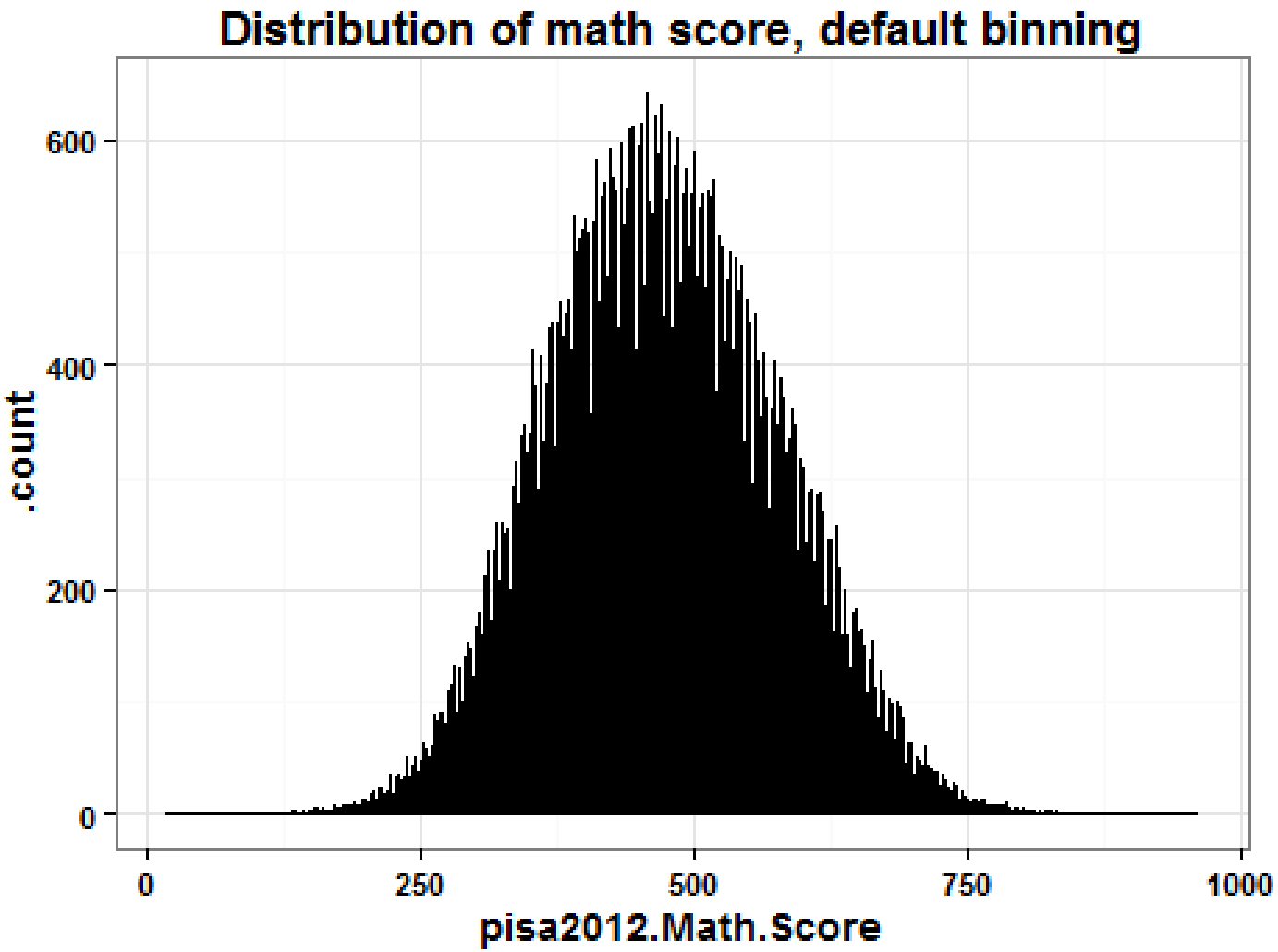
- Mit mennyire nehéz kiszámolni?
  - Disztributív
  - Algebrai
  - holisztikus

Leíró statisztikák

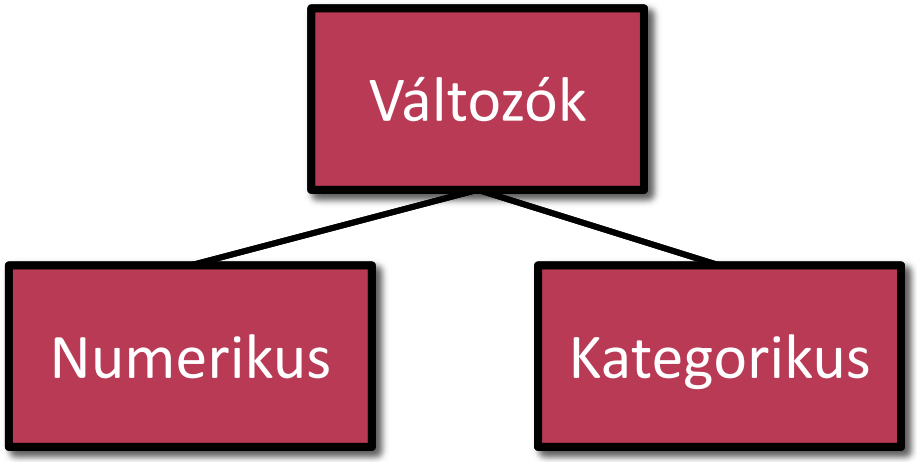
- Összefoglaló statisztikák típusai:
  - Disztributív
    - egyetlen, adott méretű köztestár
    - eredmények kombinálhatóak
    - pl. count, sum
  - Algebrai
    - disztributív statisztikák fix száma kell hozzá
    - Pl. átlag: count + sum
  - Holisztikus
    - bemenettel növekvő köztestár kell
   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

1. Általában jól párhuzamosítható
2. Interaktív vizualizációs technikákat támogatja (lásd később „iterative refinement”)

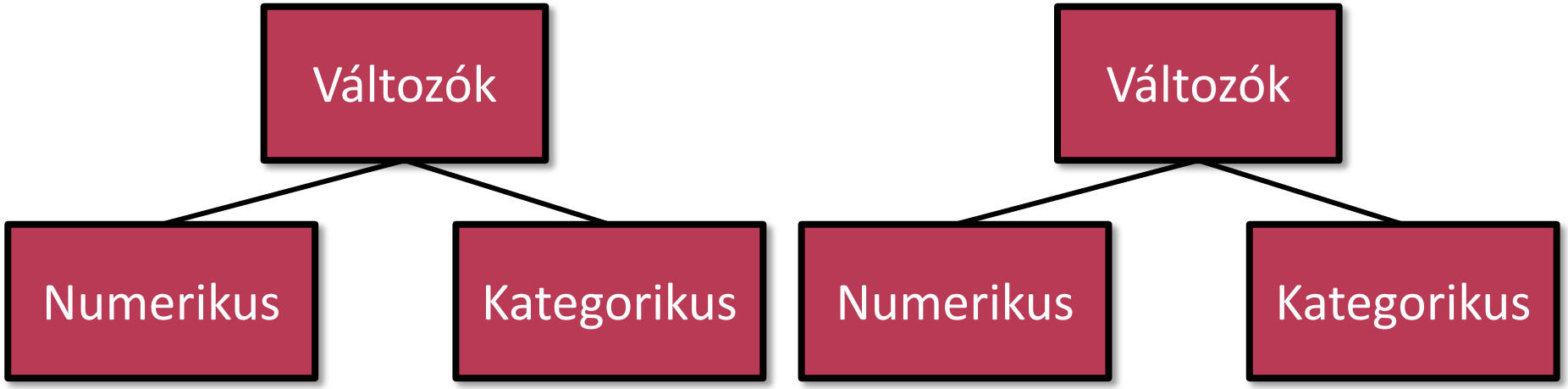
„Bin”


Distribution of math score, default binning
.count
600
400
200
0
0
250
500
750
1000
pisa2012.Math.Score

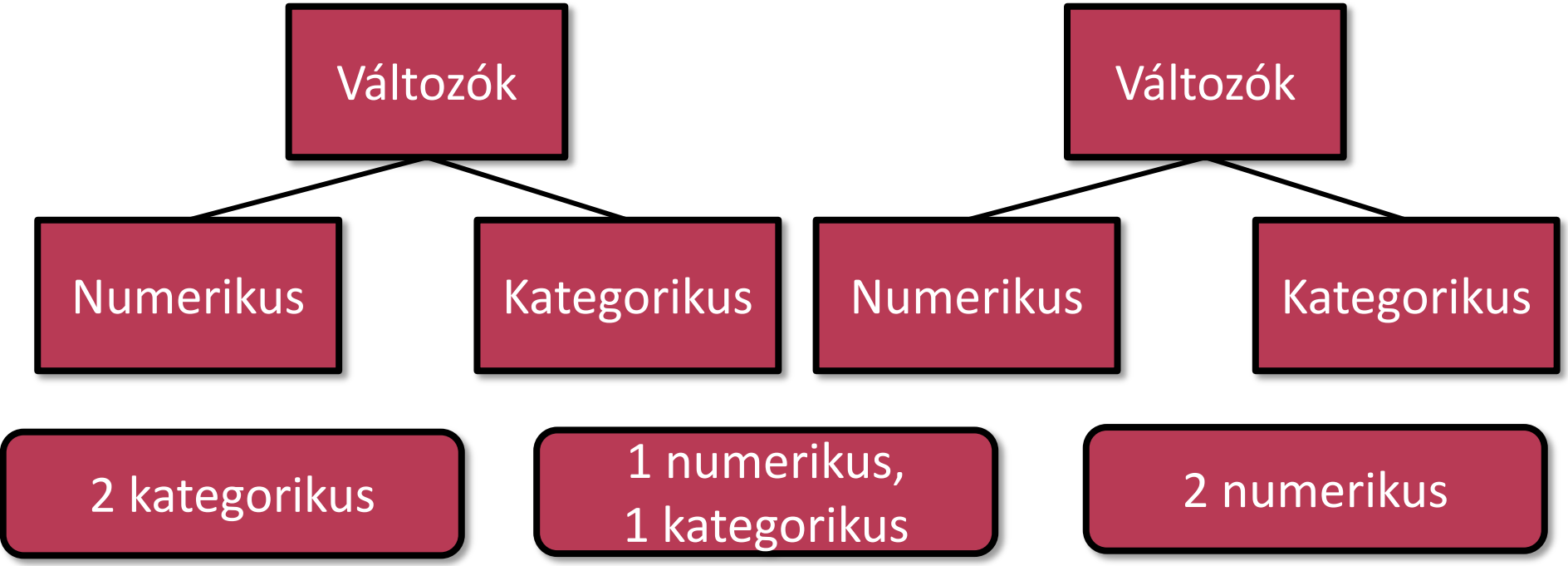
2 változó kapcsolata



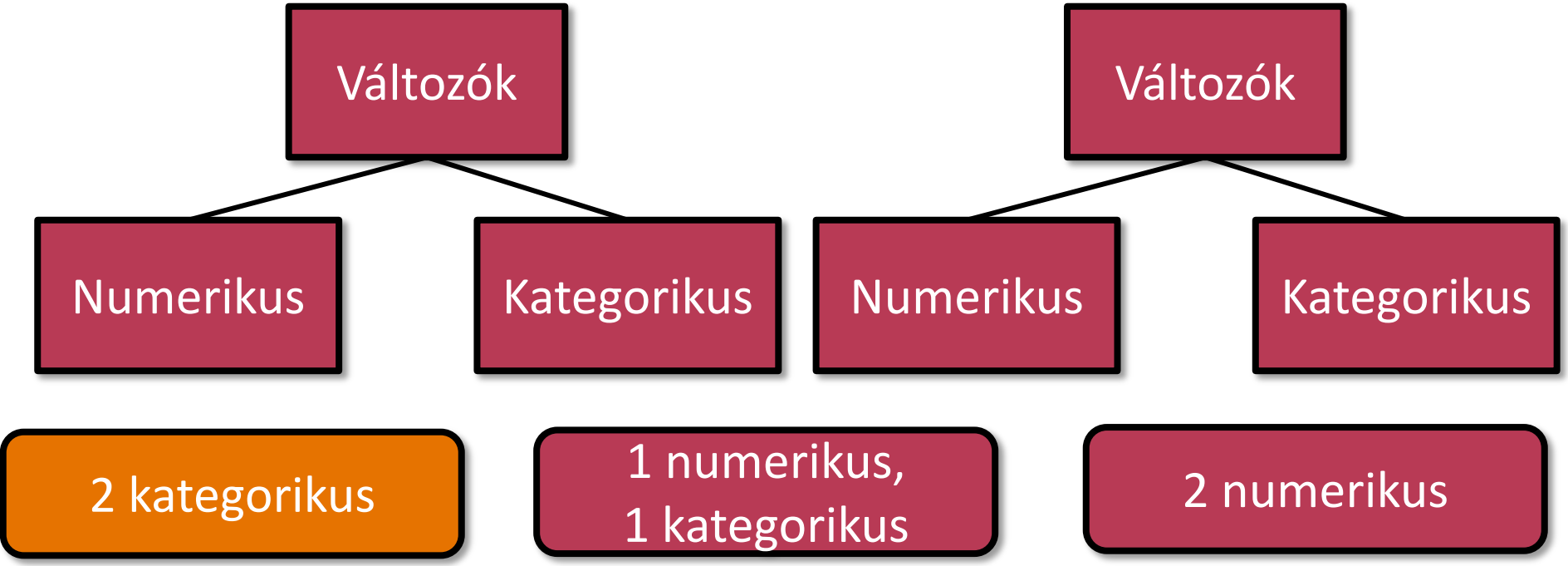
2 változó kapcsolata



2 változó kapcsolata



2 változó kapcsolata

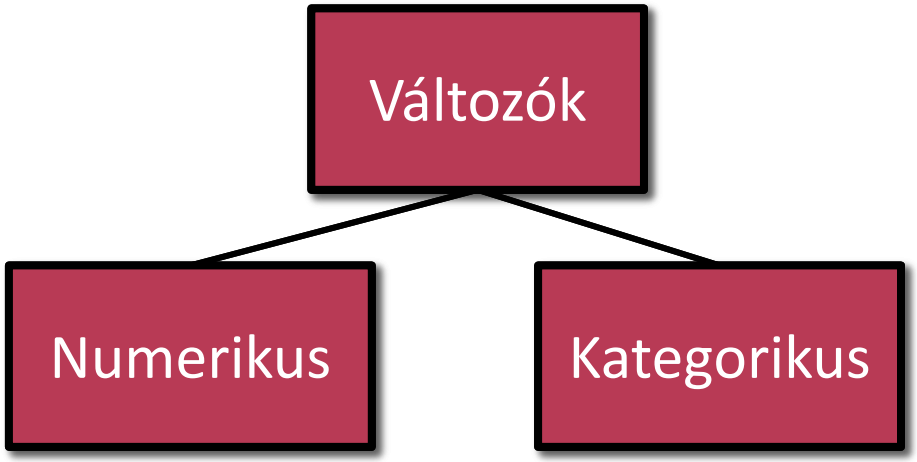


2 kategorikus

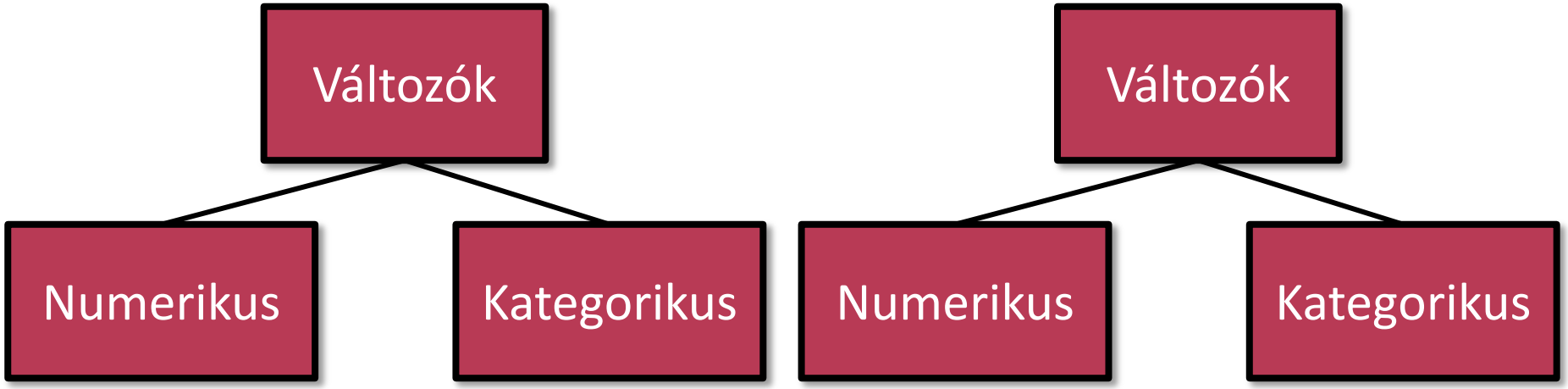
- 2 count
- Sajnos nem megy egymástól függetlenül ☹
  - $p(X = x_1, Y = y_1) \neq p(X = x_1) \times p(Y = y_1)$

- Előbb az egyik szerint számolunk, aztán kategóriánként a másik szerint

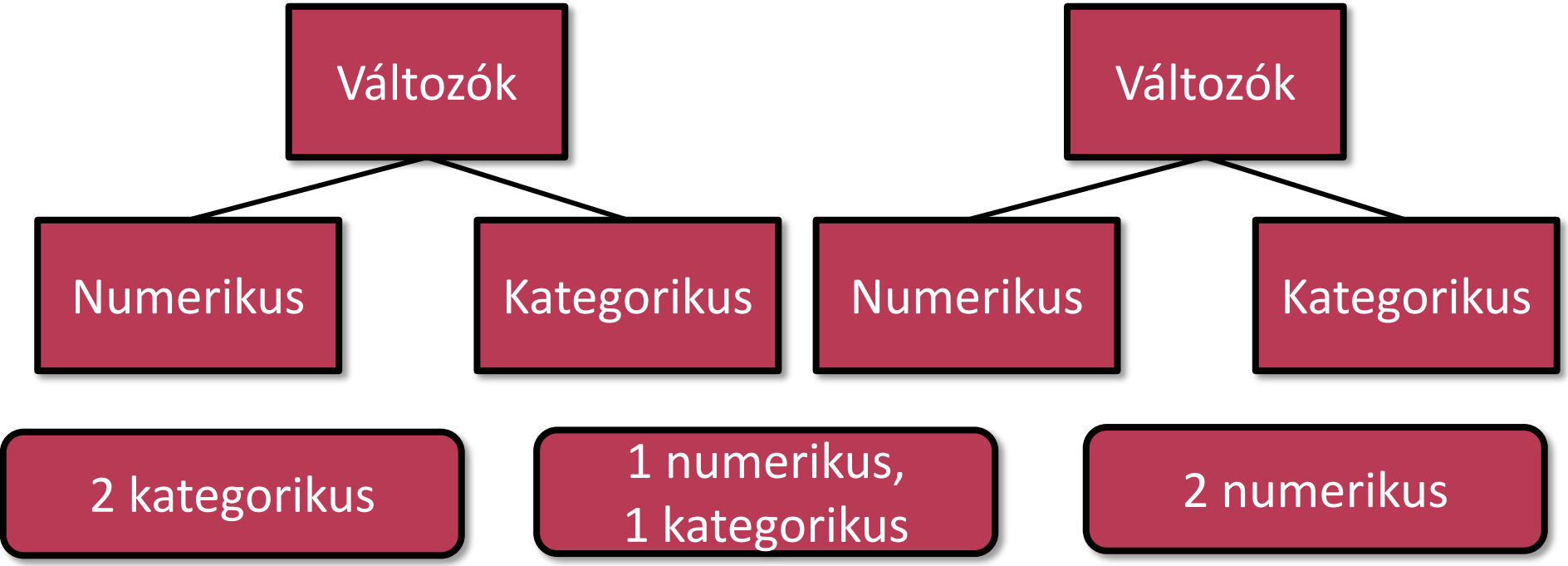
2 változó kapcsolata



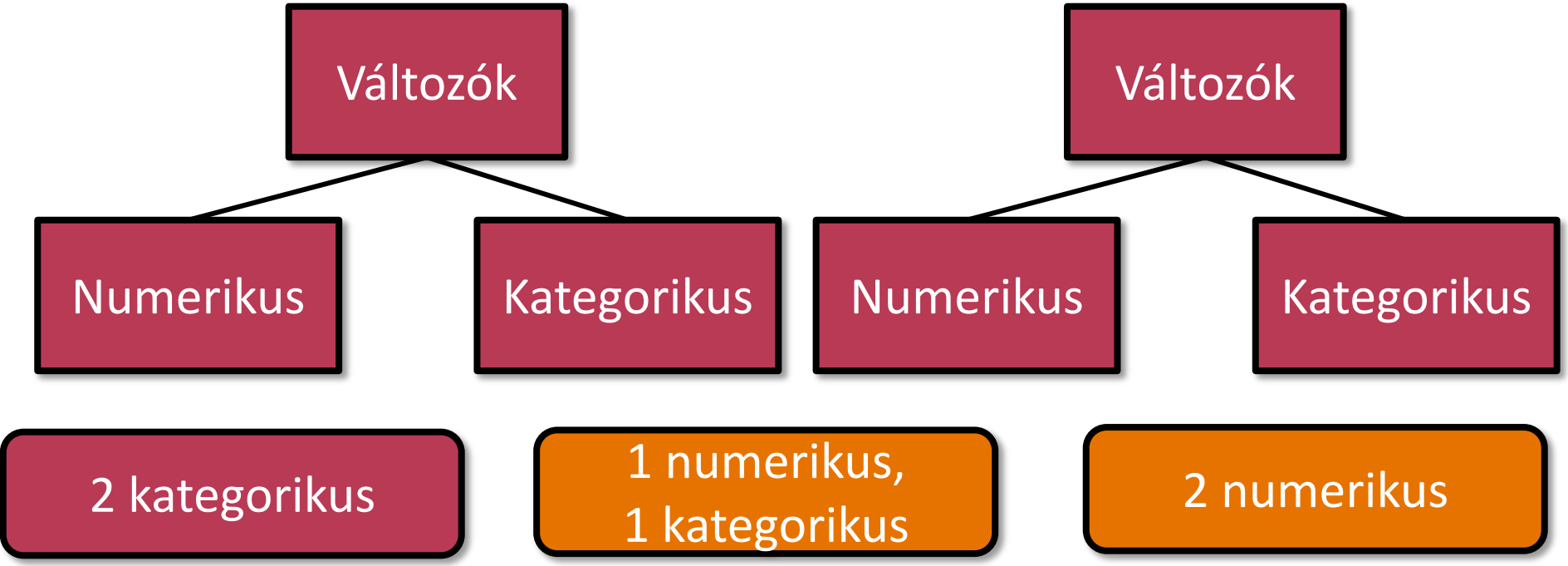
2 változó kapcsolata



2 változó kapcsolata



2 változó kapcsolata



„Bin”

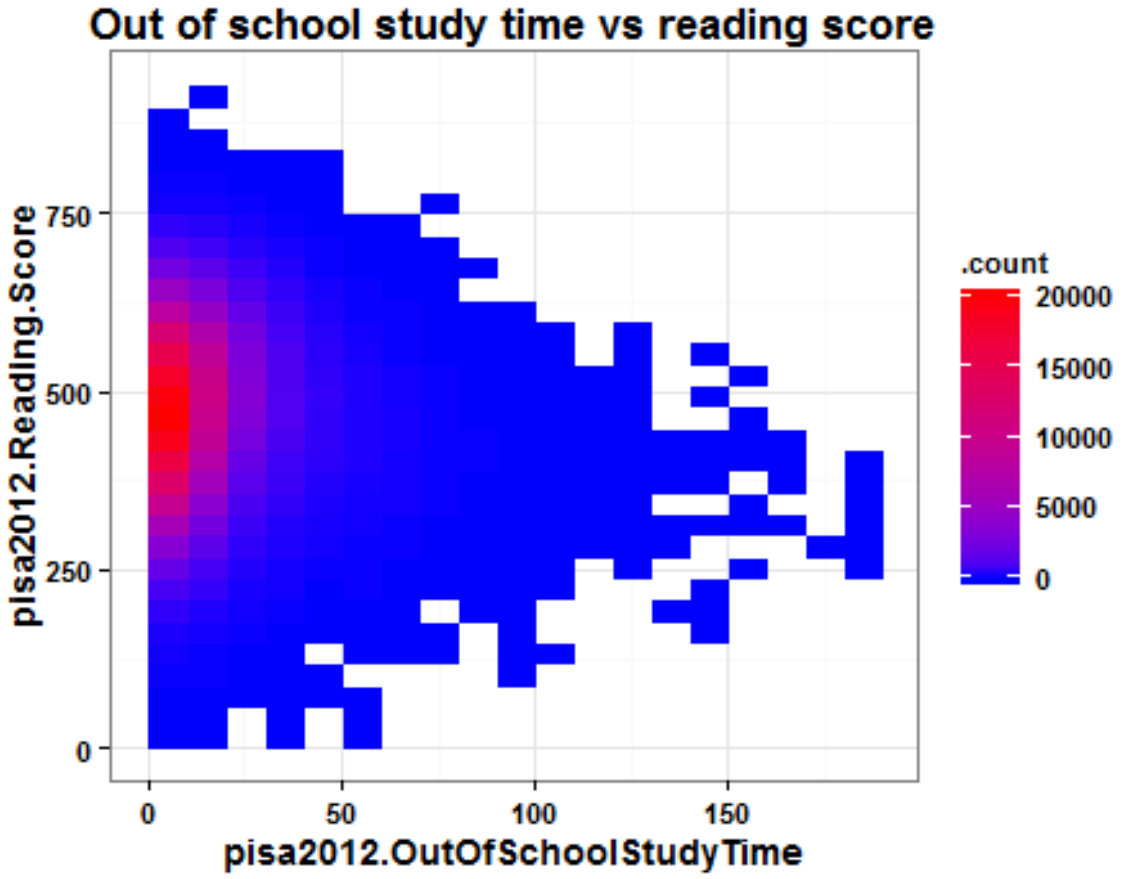
- Fix szélességű dobozok
- Egy dimenzióban: $\left\lfloor \frac{x - origin}{width} \right\rfloor + 1$
- Általánosítás több dimenzióban

$$= x_1 + x_2 \cdot n_1 + x_3 \cdot n_1 \cdot n_2 + \cdots + x_m \prod_{i=1}^{m-1} n_i$$

$$= x_1 + n_1 \cdot (x_2 + n_2 \cdot (x_3 + \cdots (x_m))$$

Numerikus is

- Mindkettő „bin”, a statisztika alapja count



Numerikus is

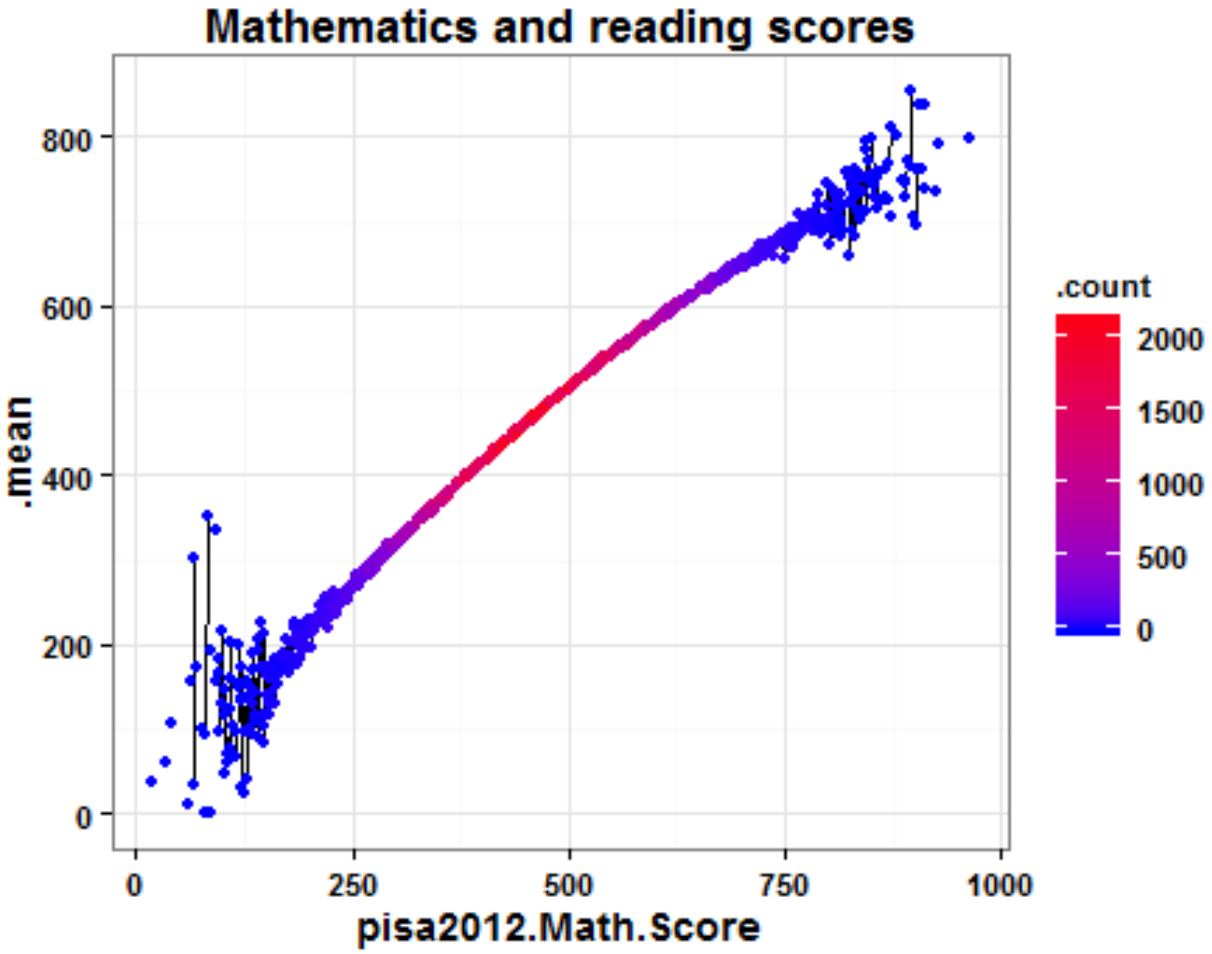
- Mindkettő „bin”, a statisztika alapja count

- Az egyiknek csak valamelyik leíró statisztikáját (pl. átlag) számoljuk/vizualizáljuk

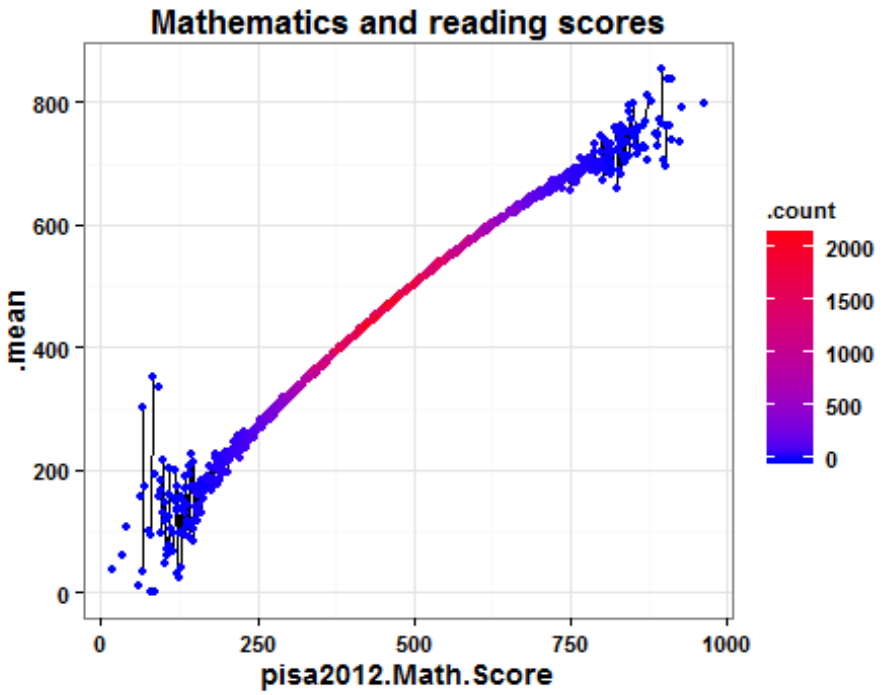
Numerikus is

- Mind

- Az eg át (pl. átlag

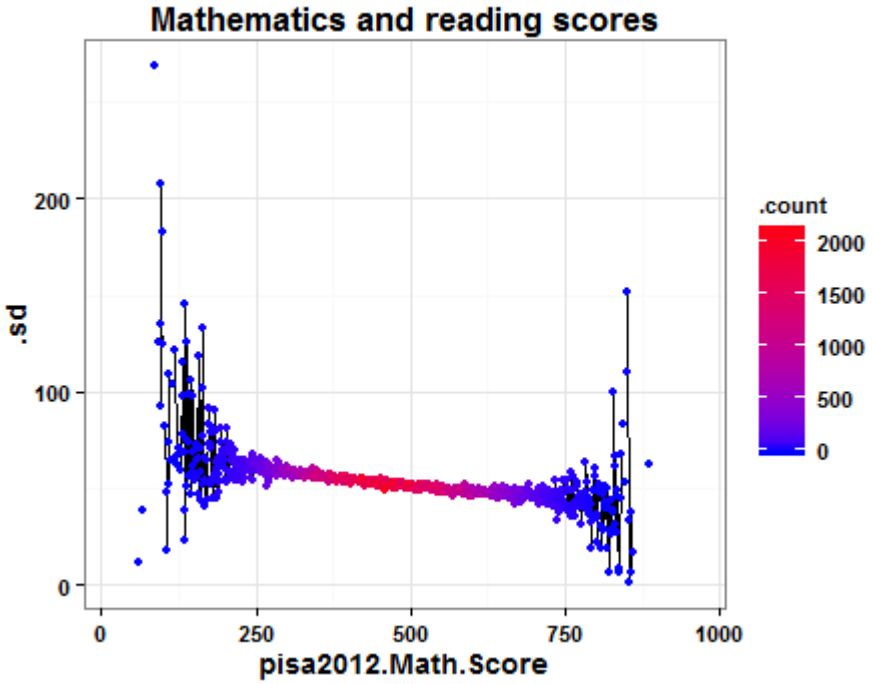


Numerikus is



ztika alapja count

lyik leíró statisztikáját (pl.
áljuk



$n > 2$ változó

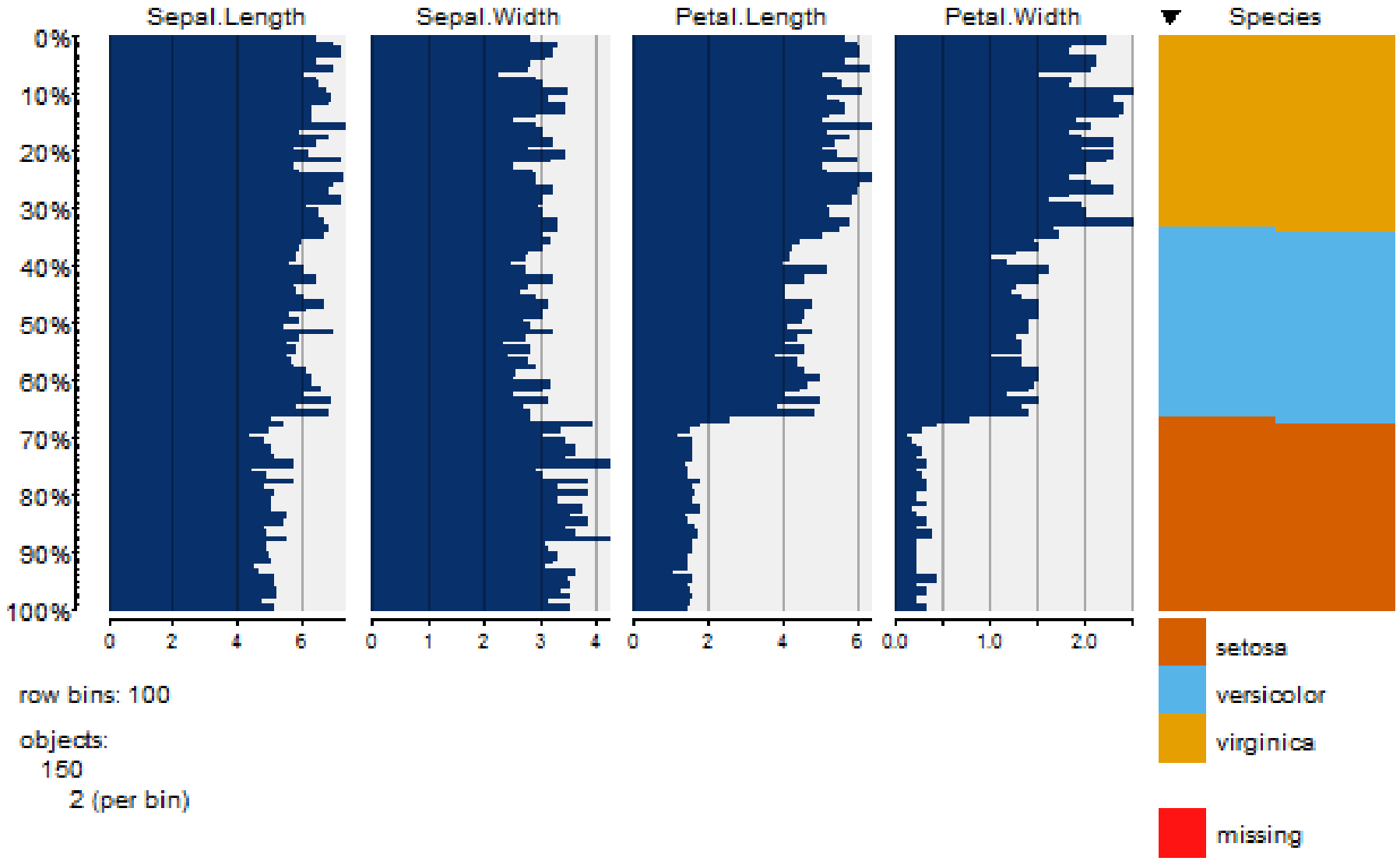
- Általánosítás: darabolunk a változók szerint, majd ugyanaz, mint fent

- Általános elv: **small multiples**
  - Ugyanazt rajzoljuk ki kategóriák szerint
  - Pl., matematika-olvasás pontszám scatterplotja országonként
  - R specifikusan: facet in ggplot2, trellis in lattice

- Inkább kategorikus, nagyon felderítés: tabplot
- Inkább numerikus, inkább részletek : Trelliscope

TABPLOT

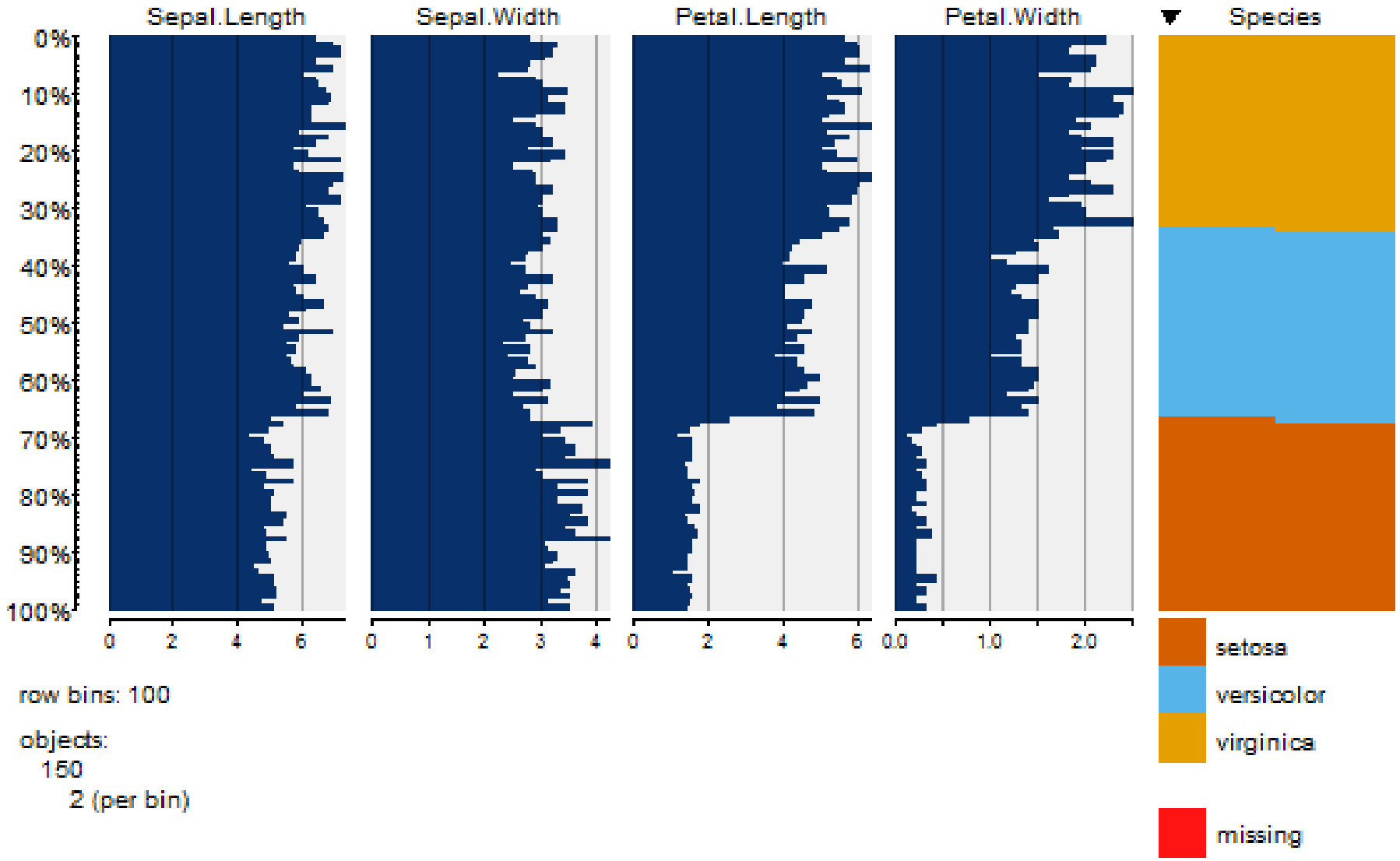
tabplot


Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
Species
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
0
2
4
6
0
1
2
3
4
0
2
4
6
0.0
1.0
2.0
setosa
versicolor
virginica
missing
row bins: 100
objects:
150
2 (per bin)

tableplot

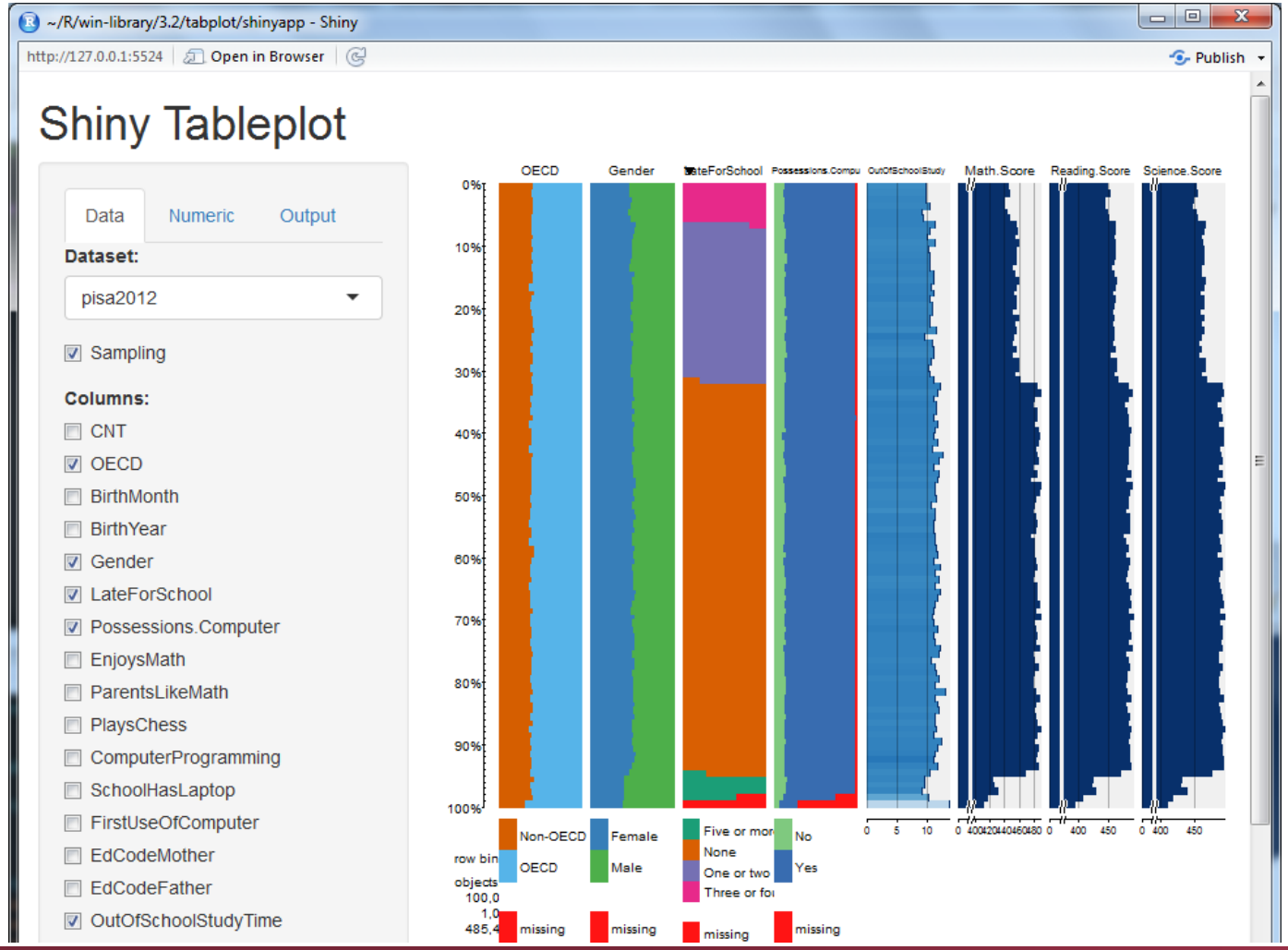
- Szűrünk adott változó lehetséges értékei alapján
  - Kategorikus: értékkészlet elemei
  - Numerikus: percentilisekkel
- A vetített változók jellemző eloszlásait vizualizáljuk
  - Kategorikus: stacked barchart gyakoriság alapján
  - Numerikus: átlaggal

tabplot


Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
Species
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
0
2
4
6
0
1
2
3
4
0
2
4
6
0.0
1.0
2.0
setosa
versicolor
virginica
missing
row bins: 100
objects:
150
2 (per bin)

tabplot

- PISA 2012

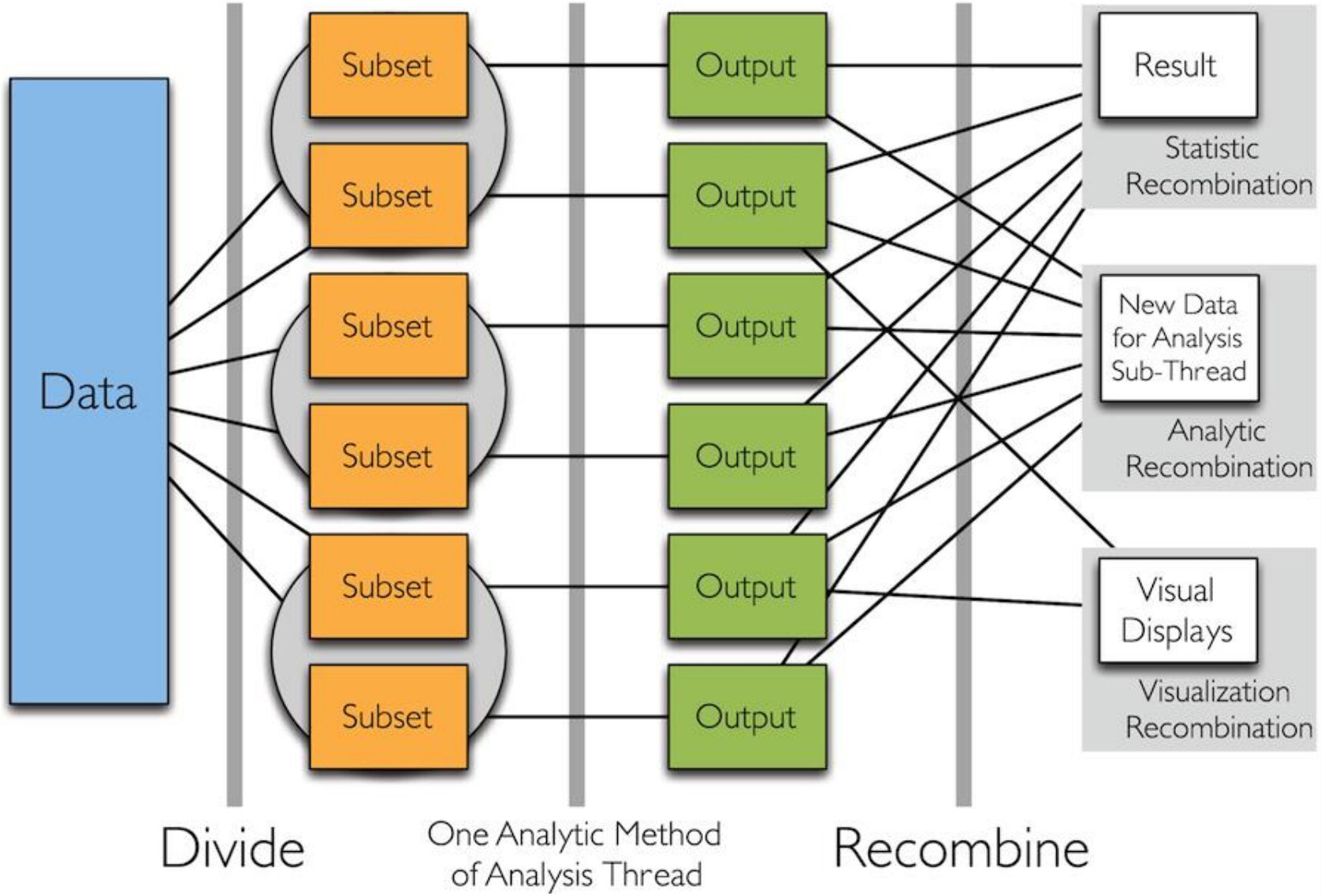


TRELLISCOPE

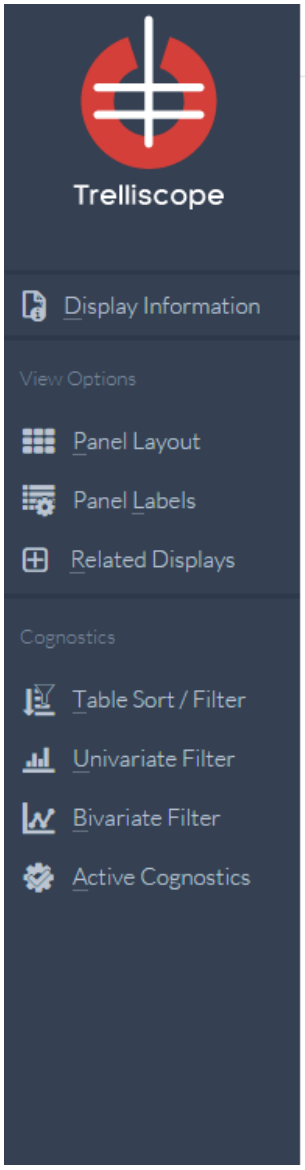
Trelliscope

- Small multiples vizualizáció
  - A rendező attribútumlista többelemes is lehet (a tabplotban egy változó szerint rendezünk csak)
- Bemeneti adatszerkezet: ddf (distributed data frame)
  - Memória, lokális diszk
  - HDFS
- Divide & recombine elv még a rajzolás előtt

Devide and recombine

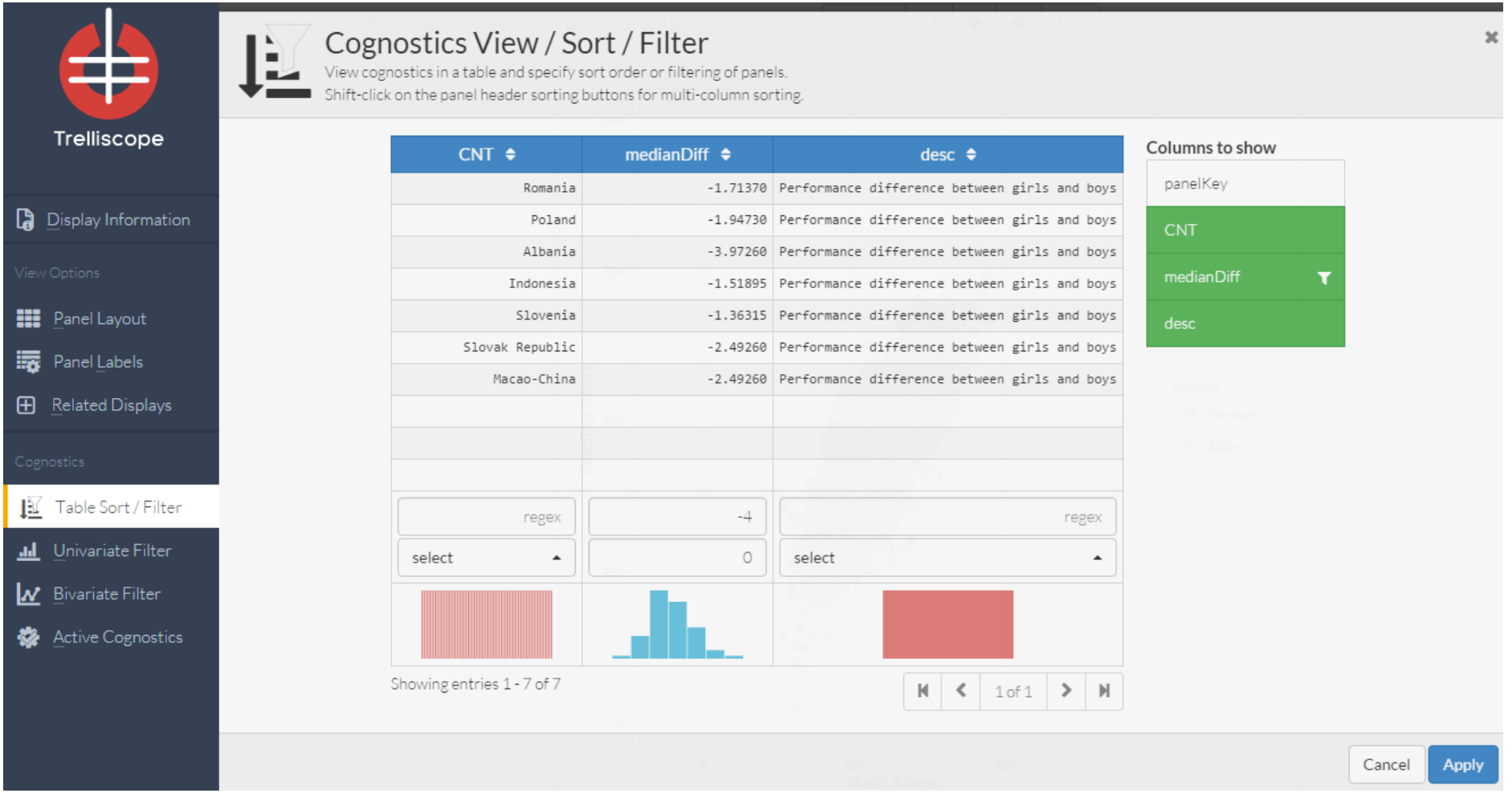


Trelliscope



Cognostics

- Tetszőleges függvényben megfogalmazható leírója az ábrának/adatnak
- Később ezekre lehet keresni vagy ezek alapján sorrendezni

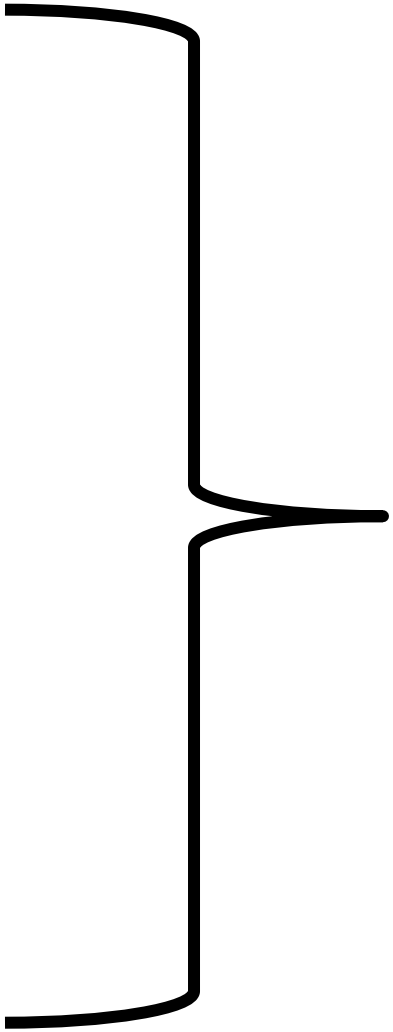

Trelliscope
Display Information
View Options
Panel Layout
Panel Labels
Related Displays
Cognostics
Table Sort / Filter
Univariate Filter
Bivariate Filter
Active Cognostics
Cognostics View / Sort / Filter
View cognostics in a table and specify sort order or filtering of panels.
Shift-click on the panel header sorting buttons for multi-column sorting.
CNT
medianDiff
desc
Romania -1.71370 Performance difference between girls and boys
Poland -1.94730 Performance difference between girls and boys
Albania -3.97260 Performance difference between girls and boys
Indonesia -1.51895 Performance difference between girls and boys
Slovenia -1.36315 Performance difference between girls and boys
Slovak Republic -2.49260 Performance difference between girls and boys
Macao-China -2.49260 Performance difference between girls and boys
regex
-4
regex
select
0
select
Showing entries 1 - 7 of 7
1 of 1
Columns to show
panelKey
CNT
medianDiff
desc
Cancel
Apply

Bigdata vizualizáció R alapokon: ami közös

- bigvis
- tabplot
- trelliscope

Bigdata vizualizáció R alapokon: ami közös

- bigvis
- tabplot
- trelliscope



Előbb számolunk, aztán rajzolunk

Bigdata vizualizáció R alapokon: a tradeoff

- bigvis
  - Interaktív változat
  - Adaptív binwidth
- tabplot
  - Többszörös rendezés
- Trelliscope
  - Interaktív változat →
    legalább select, zoom szinten koordinátánként

Bigdata vizualizáció R alapokon: a tradeoff

- bigvis
  - Interaktív változat
  - Adaptív binwidth
- tabplot
  - Többszörös rendezés
- Trelliscope
  - Interaktív változat →
    legalább select, zoom szinten koordinátánként

Outlierek? ☹

Bigdata vizualizáció R alapokon: különbségek

R csomag	Preferált vizualizáció	Paraméterez-hetőség	Kód mennyisége	Input
ggvis	1D, 2D	5 beépített leíró statisztika, ezek belső paraméterei	kevés	data.frame
tabplot	sokdimenziós, leginkább sok diszkrét	kb. semmi	kb. semmi	data.frame, ff
trelliscope	sokdimenziós	kb. minden	sok	ddf (memória, lokális diszk, hdfs)

HALADÓ TIPPEK-TRÜKKÖK

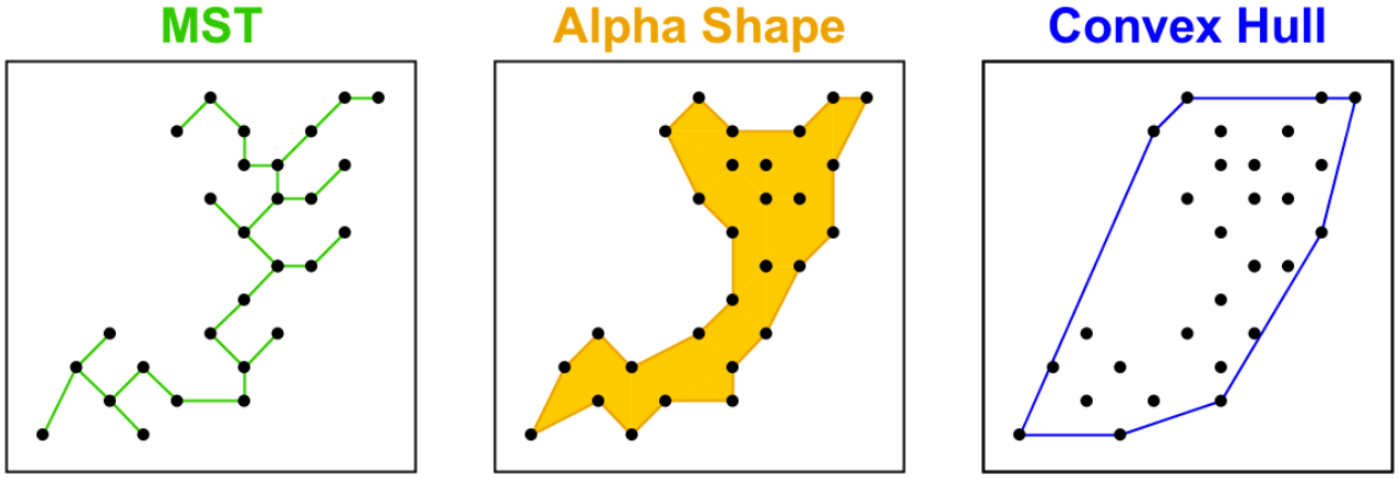
Milyen elven működnek a fenti eszközök?

MI TÖRTÉNIK, HA NEM TUDJUK ELŐRE, MIT ÉRDEMES MEGNÉZNI?

Scagnostic measures

Scagnostic measures

- Megpróbáljuk kitalálni a 2D scatterplotból az összefüggést, amit ábrázol
- 3 vizualizációs forma
  - Konvex burok (convex hull)
  - Alfa-burok (alpha hull)
  - Minimális feszítőfa (MST)

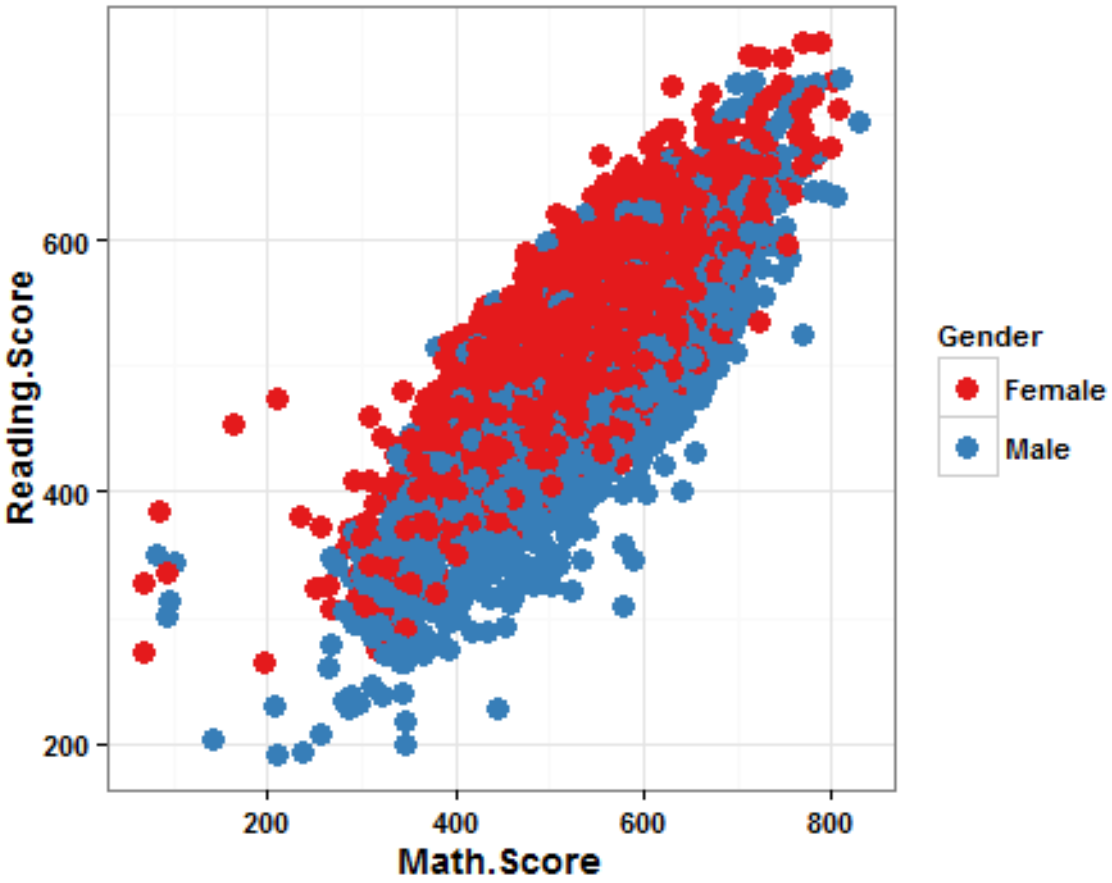


Scagnostic measures

- Outlying
- Skewed
- Clumpy
- Sparse
- Striated
- Convex
- Skinny
- Stringy
- Monotonic

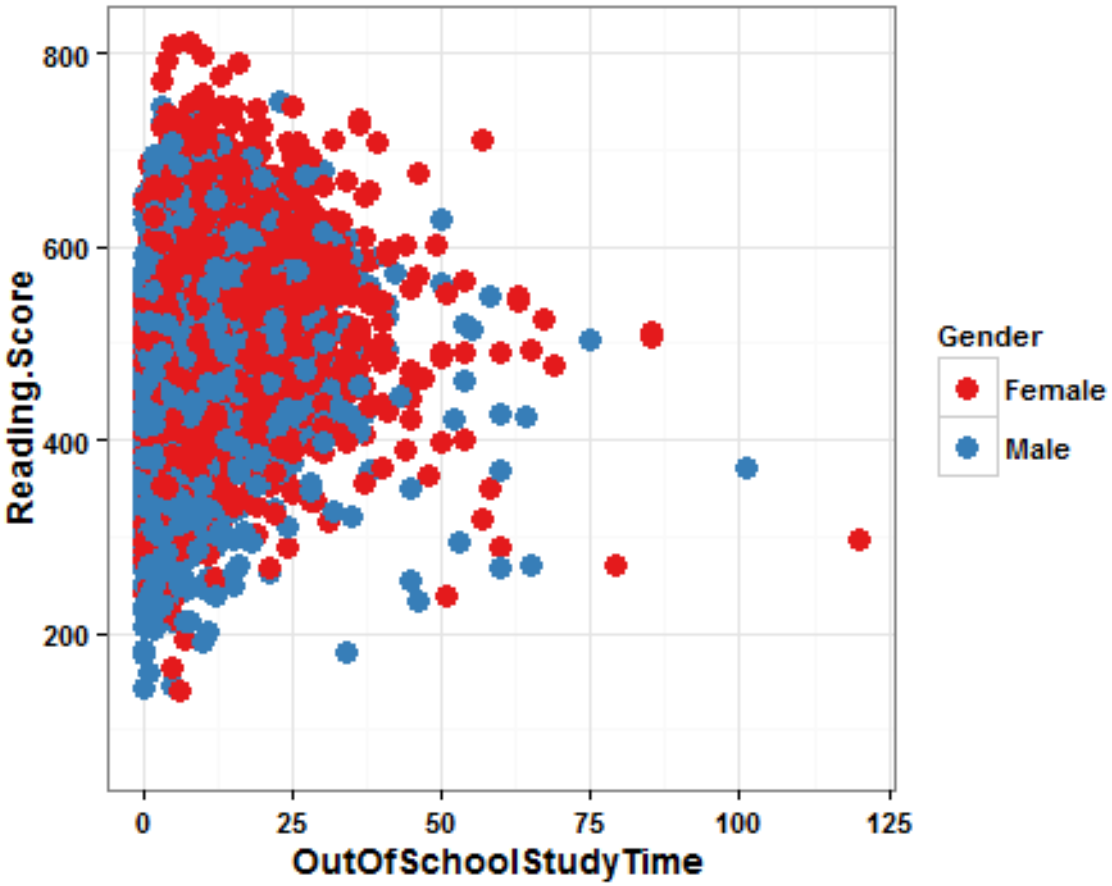
Scagnostic measures

- Outlying 0.17
- Skewed 0.68
- Clumpy 0.01
- Sparse 0.02
- Striated 0.02
- Convex 0.43
- Skinny 0.44
- Stringy 0.33
- Monotonic 0.67



Scagnostic measures

- Outlying 0.26
- Skewed 0.79
- Clumpy 0.01
- Sparse 0.02
- Striated 0.02
- Convex 0.47
- Skinny 0.42
- Stringy 0.33
- Monotonic 0.04

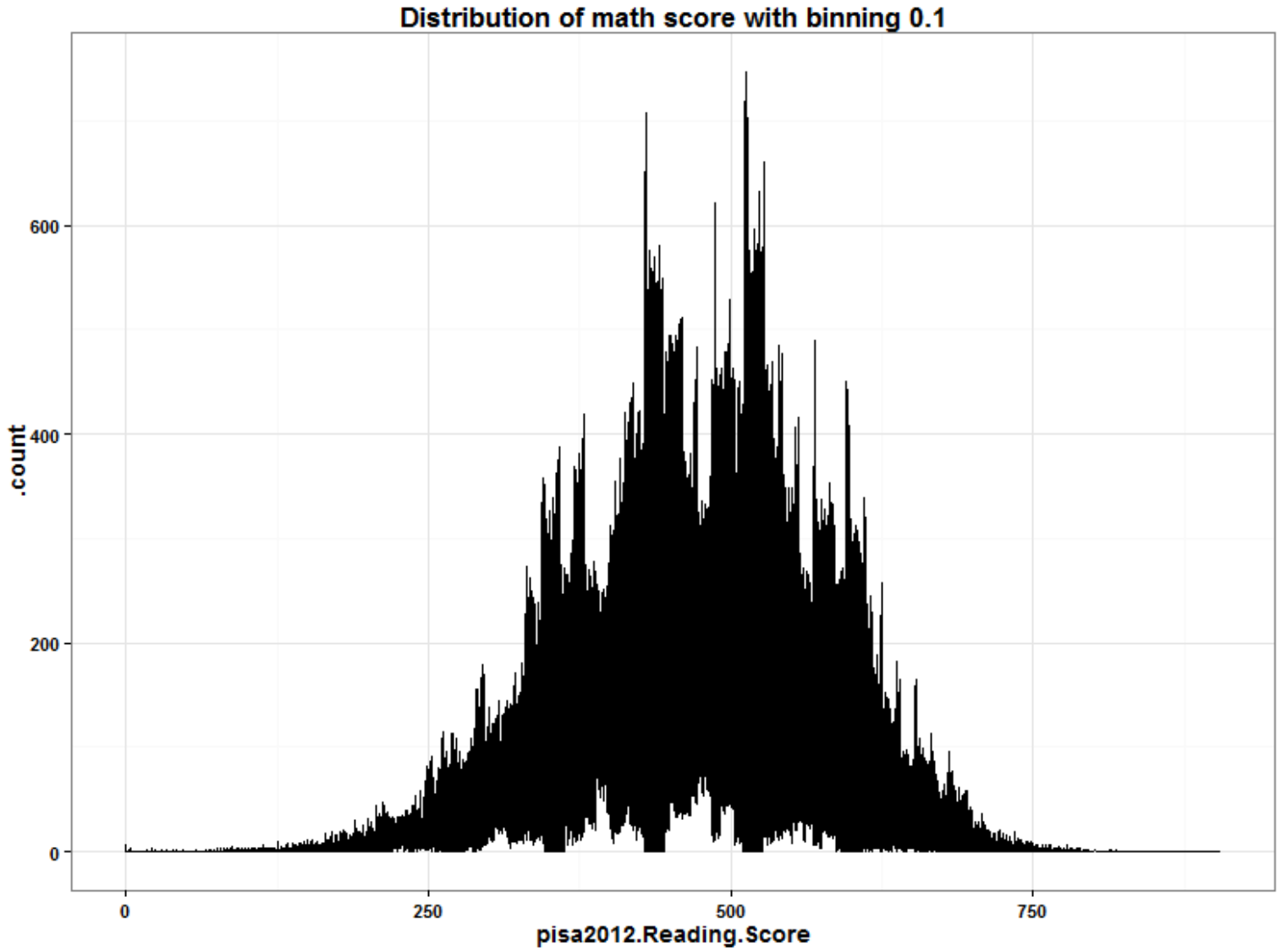


Bigvis: simítás

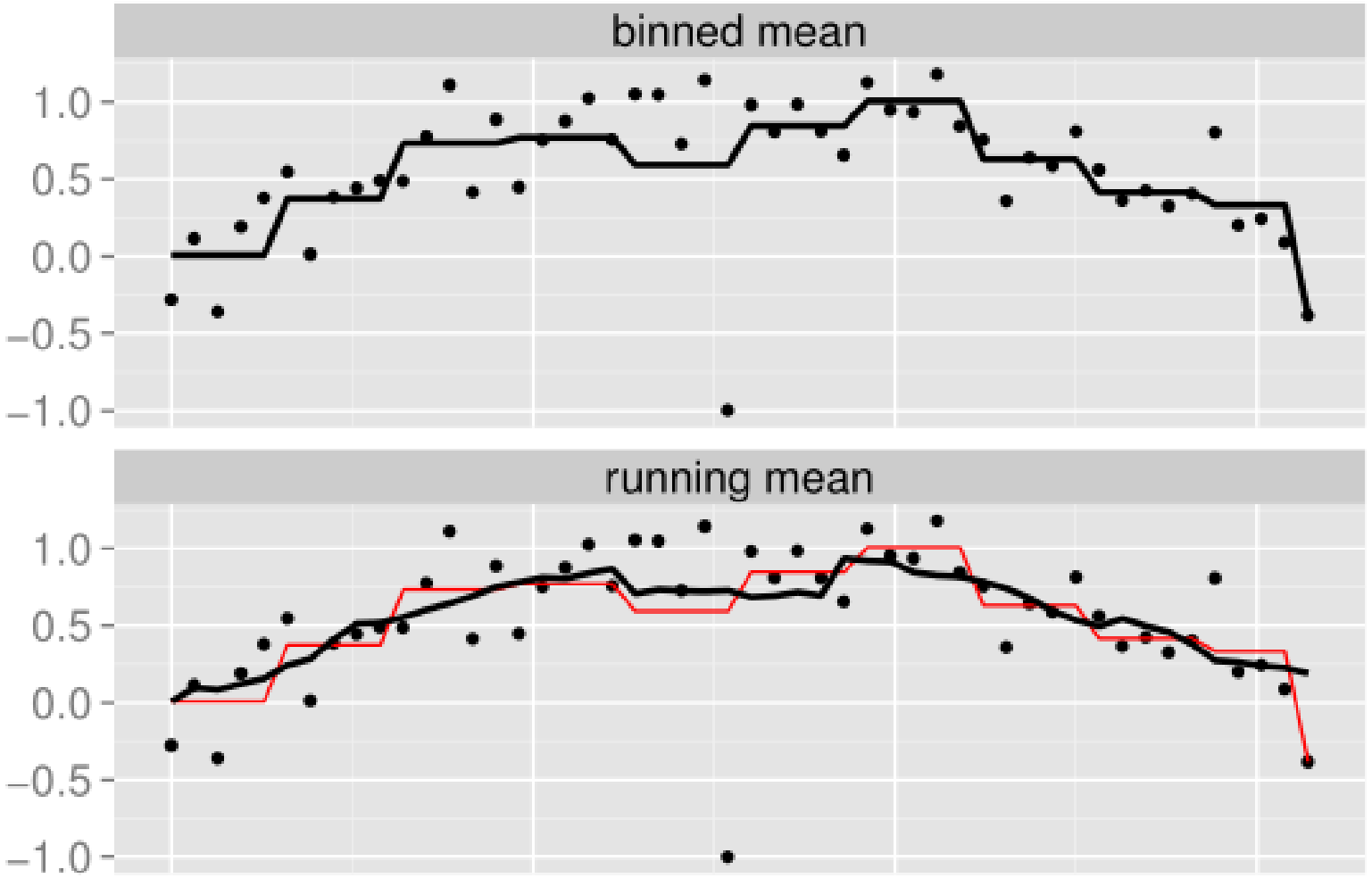
- Hogyan simítsuk el a zajt és hangsúlyozzuk a trendet?
- Hogyan szűrjük ki az outliereket?

- Általános ökölszabály: inkább legyen gyors mint robusztus (lásd még fent a „human bias” részt)

Bigvis: simítás



„Smooth”



„Smooth”

- Kernel módszerek:
  - nemcsak szomszédok,
  - de súlyozás is

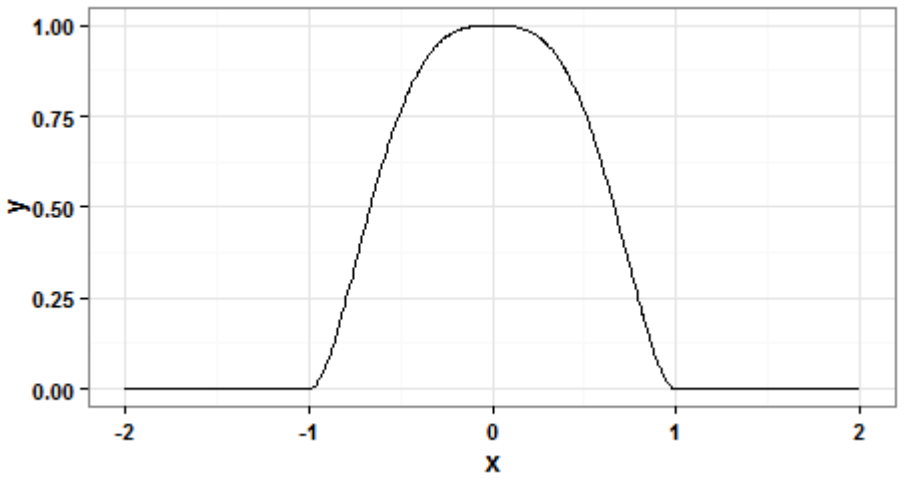
- j-edik bin közelítésénél az i-edik súlya:

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

- h: „sávszélesség”
  - Szomszédság mérete

- K itt: „triweight”

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



„Smoot

- Kernel módszerek:
  - nemcsak szomszédok,
  - de súlyozás is

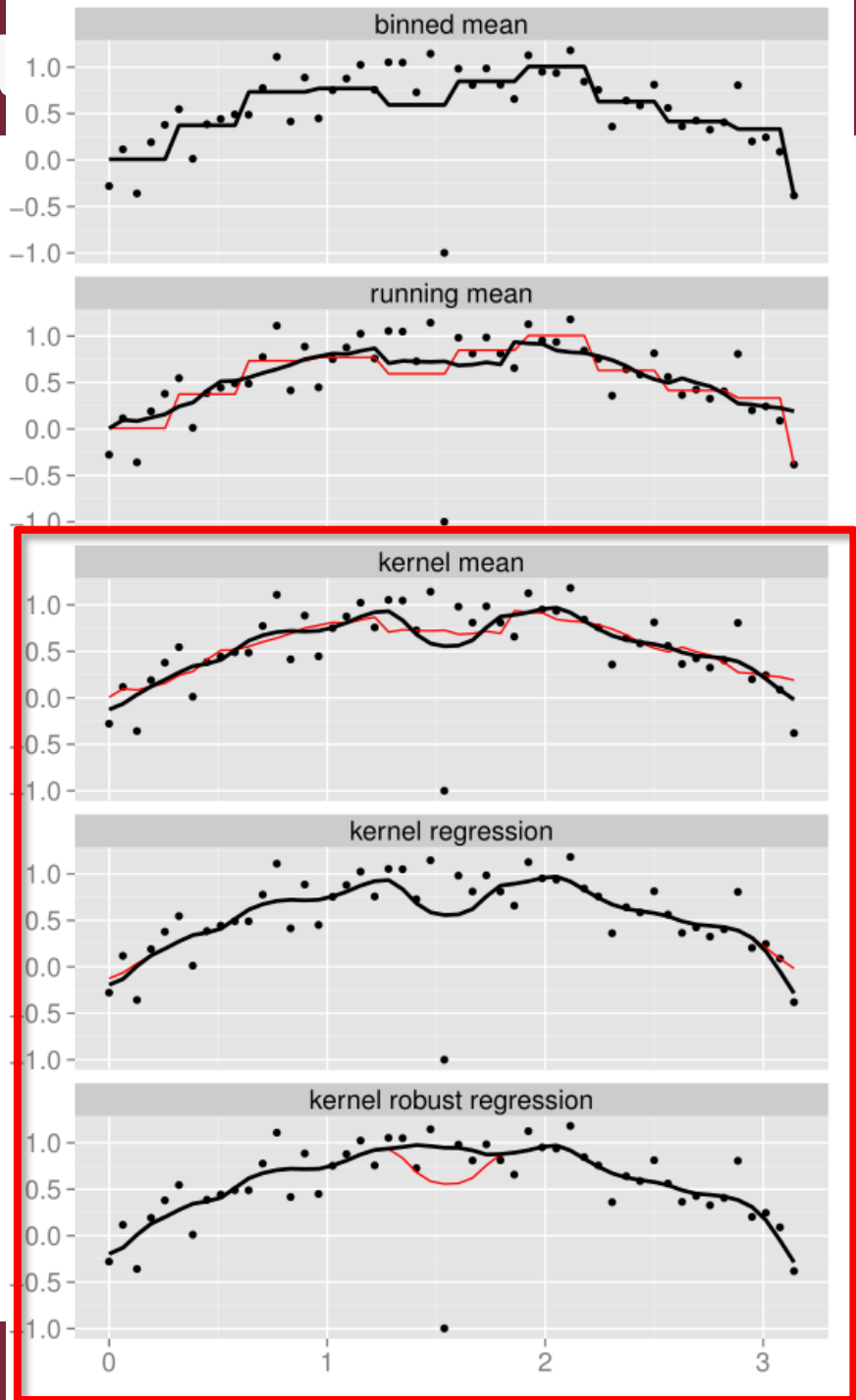
- j-edik bin közelítésénél az i-edik súlya:

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

- h: „sávszélesség”
  - Szomszédság mérete

- K itt: „triweight”

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



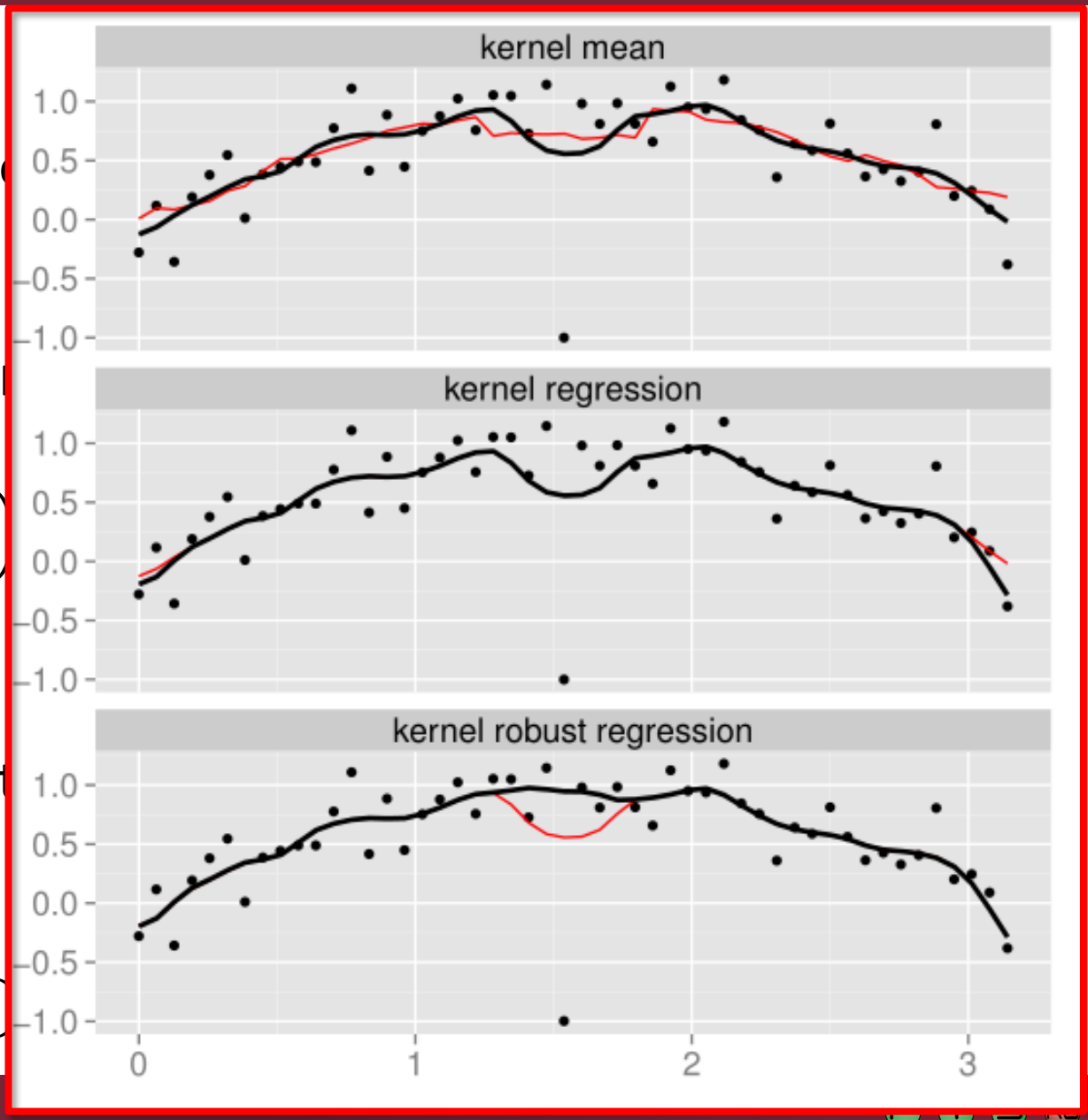
„Smooth”

- Kernel módszerek:
  - nemcsak szomszéd
  - de súlyozás is
- j-edik bin közelítésé súlya:

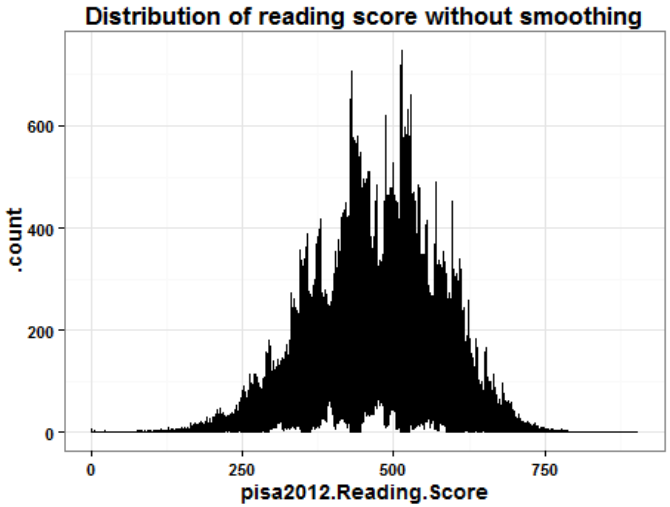
$$k_i = K\left(\frac{x_j - x_i}{h}\right)$$

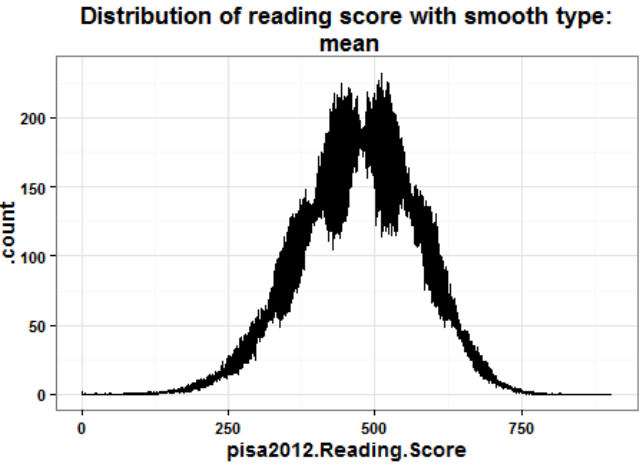
- h: „sávszélesség”
  - Szomszédság méret
- K itt: „triweight”

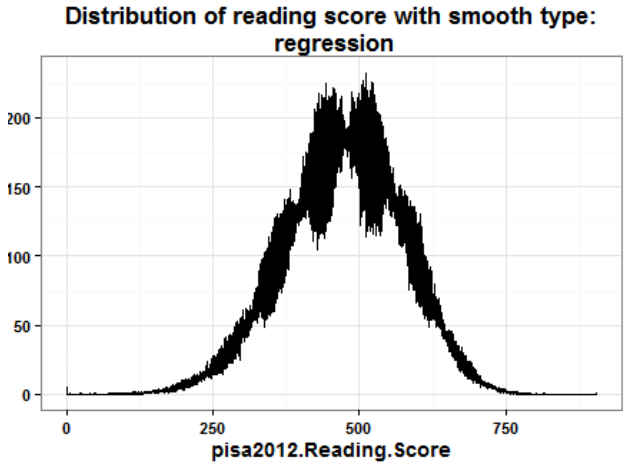
$$K(x) = (1 - |x|^3)$$

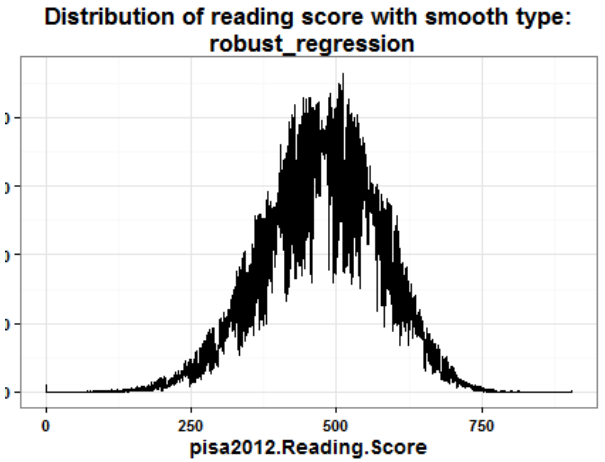


Bigvis: simítás, binwdith = 0.2

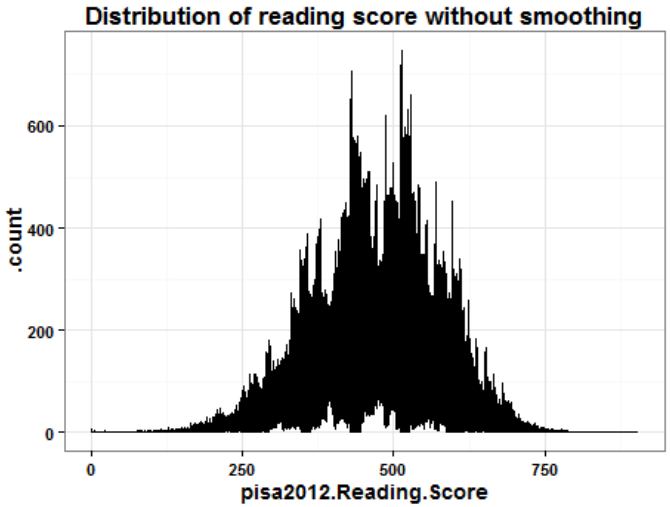


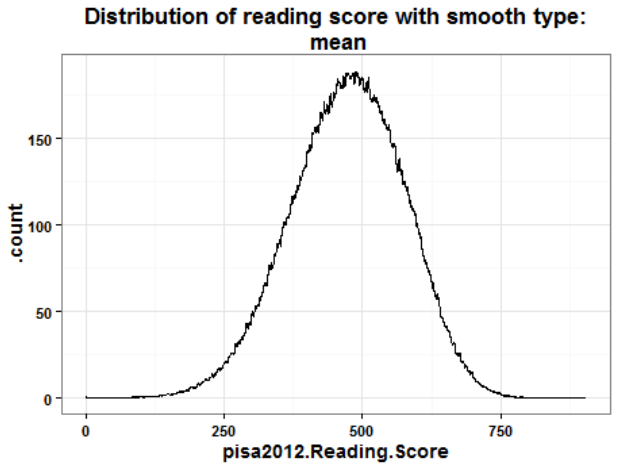


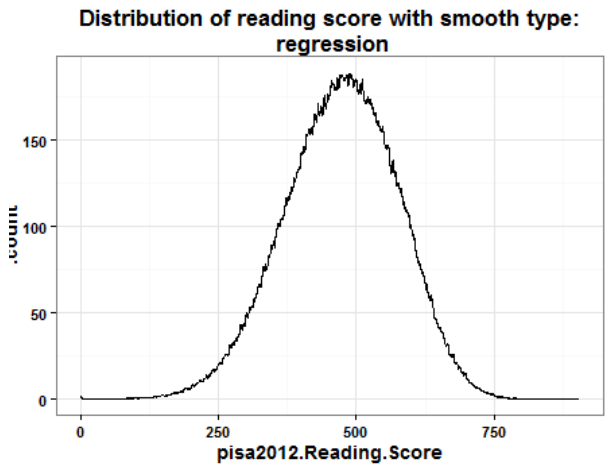


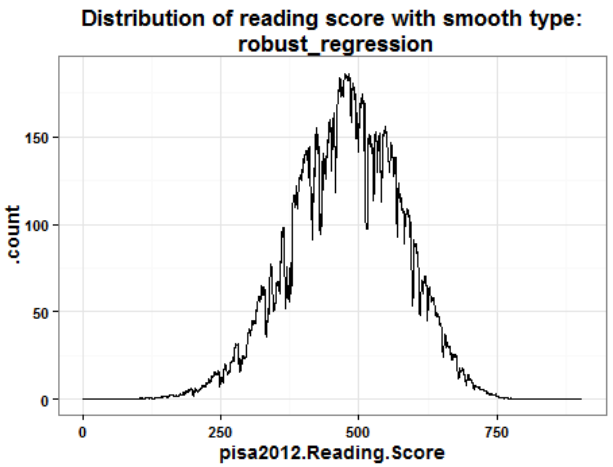


Bigvis: simítás, binwdith = 2









Bigvis: simítási sávszélesség

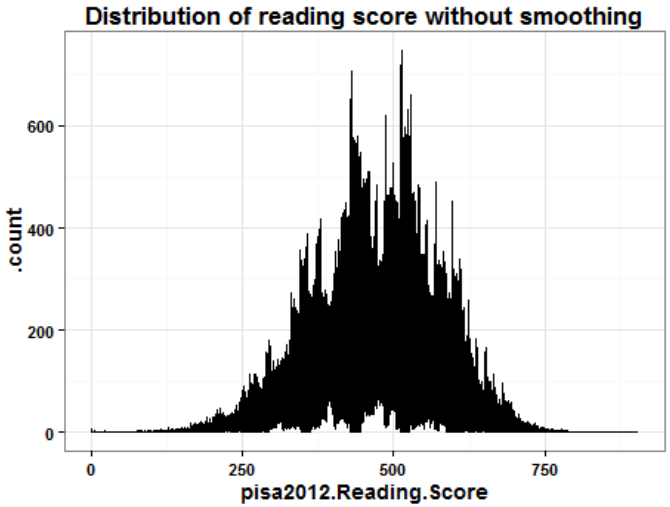
- Mekkora az ideális sávszélesség?

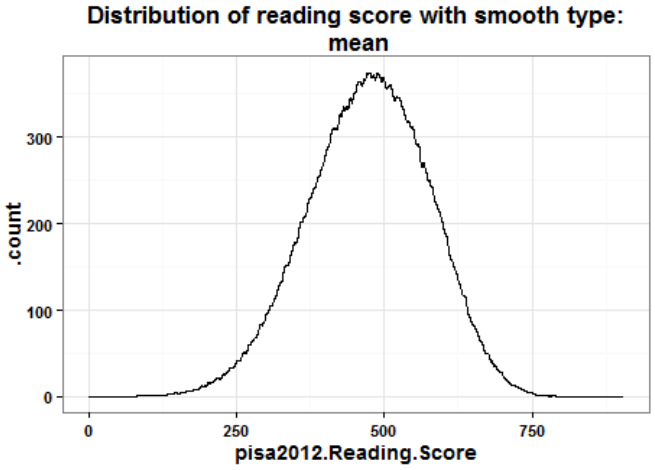
Automatikus sávszélesség választás?

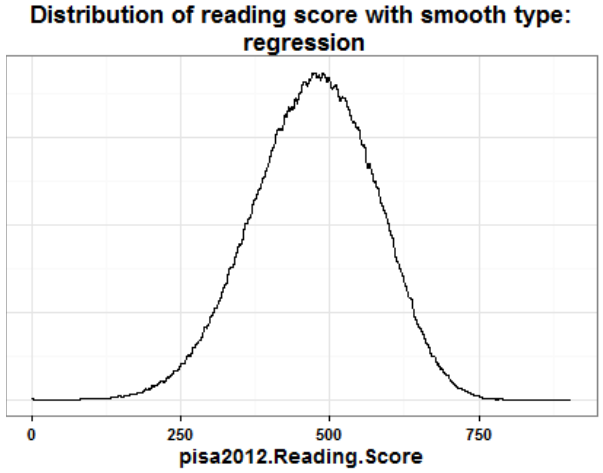
- Pl. „leave-one-out cross-validation” (LOOCV)

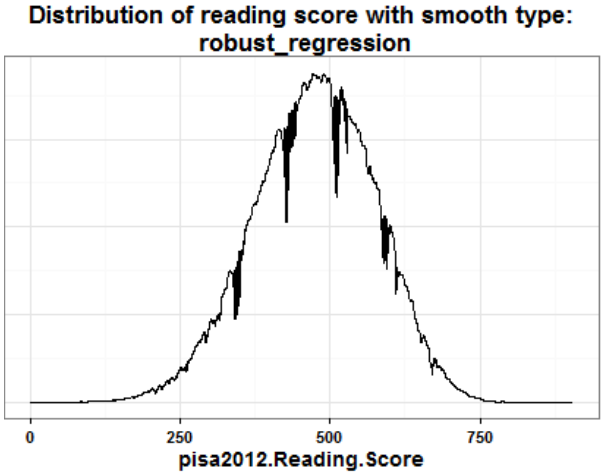
- aktuális statisztika és a nélküle simított összeh.
  - root mean squared error
  - $rmse = \sqrt{(y_i - \hat{y}_i)^2/n}$
  - keressük a minimumhoz tartozó h-t

Bigvis: simítás, binwdith = 3.5









Tableplot: zoom

- „Iterative refinement of computational results”

- Zoom esetén a binek és a statisztikák újraszámolódnak
- Támogatja a „top-down” elemzést
  - Előbb megnézzük a kategóriákat egymáshoz képest, aztán belezoomolunk

Nagyméretű adathalmazok vizualizációja

- „Perception-based precision”: elég a közelítés, a double pontosság se kell mindig
- „Perception-based convergence”: elég az első néhány iterációt vizsgálni
- „Screen-wise precision”: elég max. 3M különböző értéket számolni
- „Iterative refinement”: először elég az aggregátum, aztán majd a részletek

Hivatkozások

- Kódok: https://github.com/FTSRG/BigDANTE
- PISA 2012 adatsor: http://pisa2012.acer.edu.au/downloads.php

- Scagnostic measures: http://www2.cs.uic.edu/~tdang/file/pacificVisPoster2013.pdf

Hivatkozások

- Trelliscope: http://tessera.io/docs-trelliscope/
- Tabplot: http://www.jds-online.com/file_download/379/jds-1108.pdf
- Bigvis: http://vita.had.co.nz/papers/bigvis.pdf