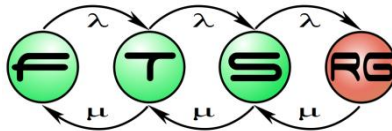


Vizuális adatanalízis



Exploratory data analysis (EDA)

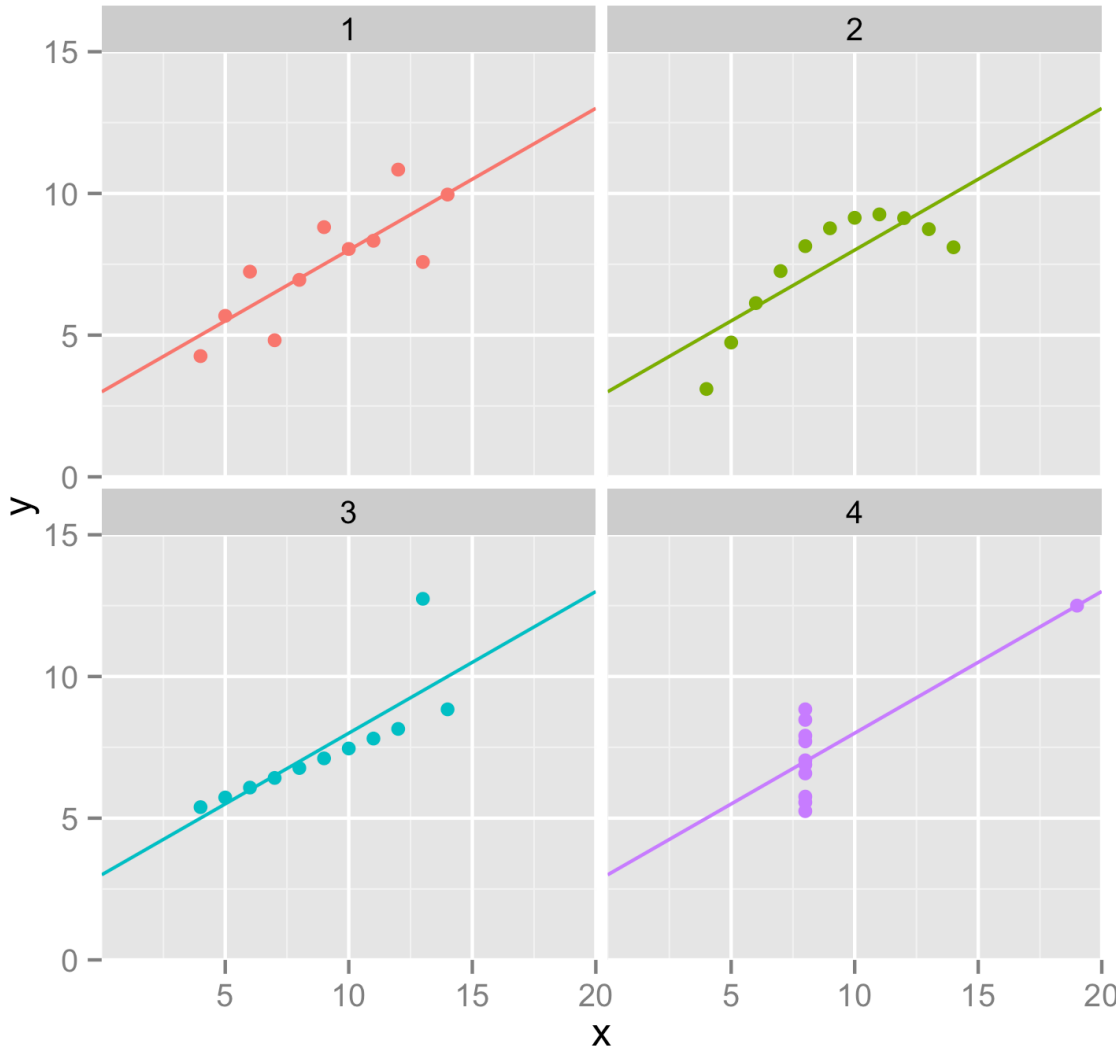
○ Cél

- Adatok megértése
 - Mi jó, mi nem?
 - Melyek a minőségileg eltérő csoportok?
 - Mik a fontos jellemzők?
- Jelenségek megsejtése
 - Korrelációkeresés (mi okoz mit?)
 - Minőségileg eltérő tartományok
- Precíz statisztikai módszerek kiválasztása

- Statisztikai analízis módszerek
 - Vizualizálás
 - Statisztikai nehéztűzéség nélkül
- Tukey, 60-as évek közepe
- Robusztus statisztika
 - Csökkentjen az érzékenység a mérési hibára
- Nemparametrikus statisztika
 - Ne kelljen az ismeretlen eloszlásra feltételezéseket tenni
- <http://www.visual-analytics.eu/>
- <http://www.rosuda.org/mondrian/>

Anscombe négyese

Anscombe's Quartet



■ Mindegyikre:

Átlag:

$$M[x] = 9$$

$$M[y] \sim 7.5$$

Szórás:

$$\sigma[x] = 11$$

$$\sigma[y] \sim 4.12$$

Korreláció:

$$C(x, y) \sim 0.816$$

Regressziós egyenes:

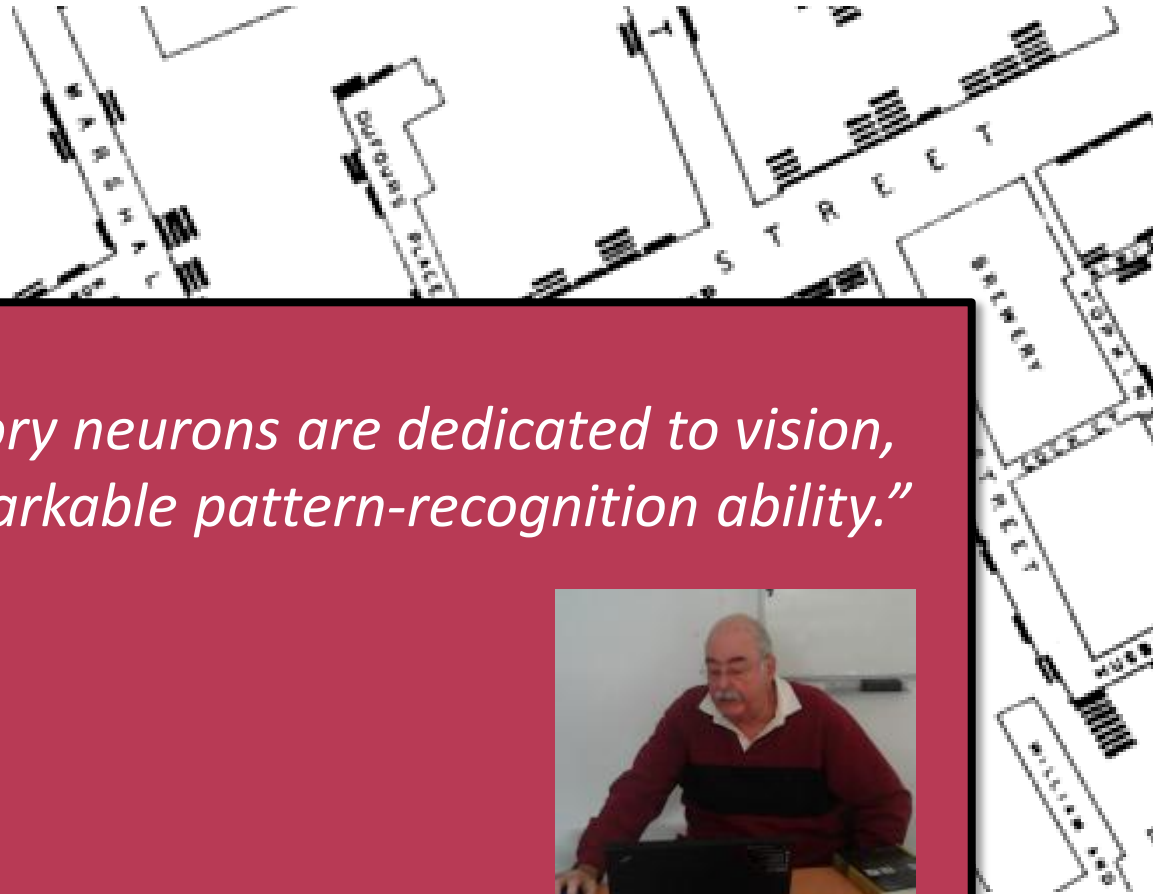
$$y \sim 3 + 0.5x$$

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

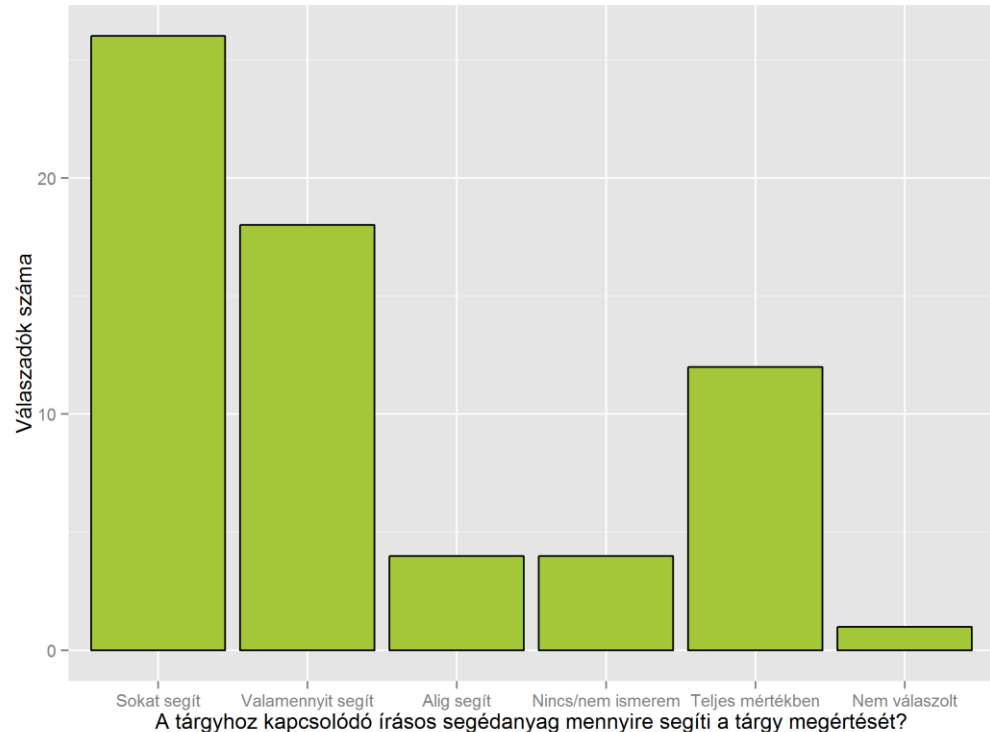


Forrás: [5] és [6]

EGYEDI VÁLTOZÓK

Oszlopdiagram (bar chart)

- Megjelenített dimenziók száma: 1
- Ábrázolt összefügg.:
 - Diszkrét változó egyes értékeinek abszolút gyakorisága
- Adategység:
 - Oszlop: magassága az adott érték abszolút gyakoriságát tükrözi
- Tervezői döntés:
 - Csoportok kialakítása?
 - Értékkészlet darabolása?



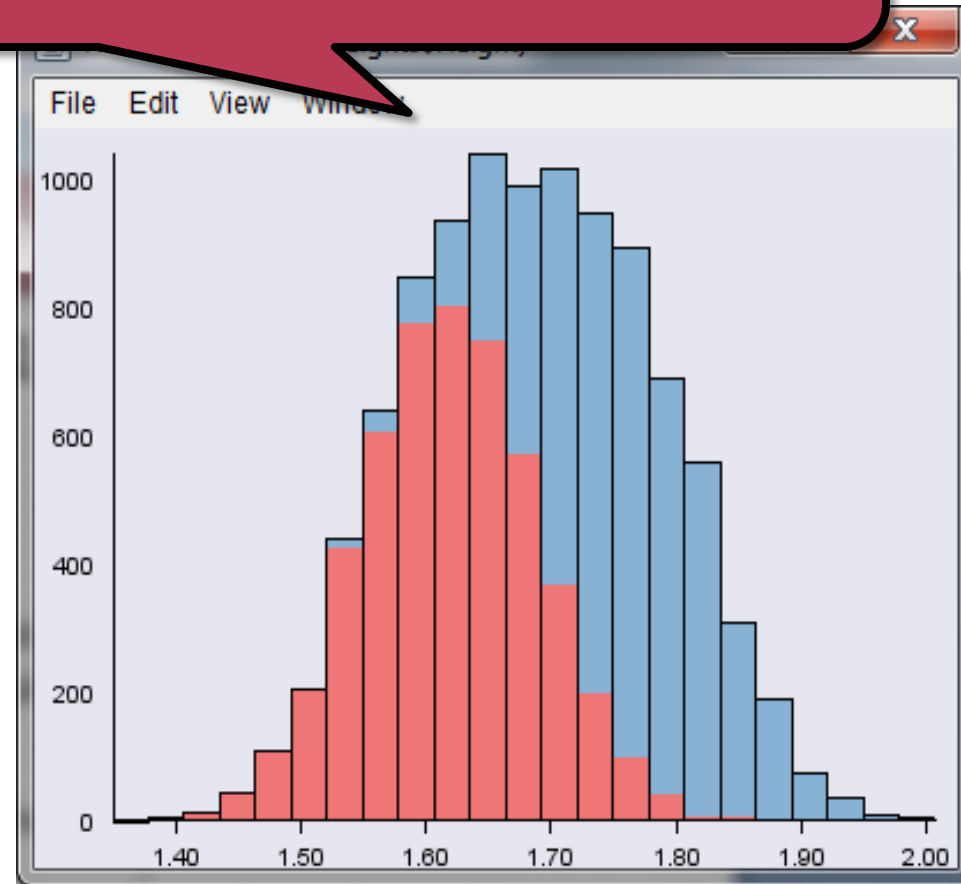
Hisztogram

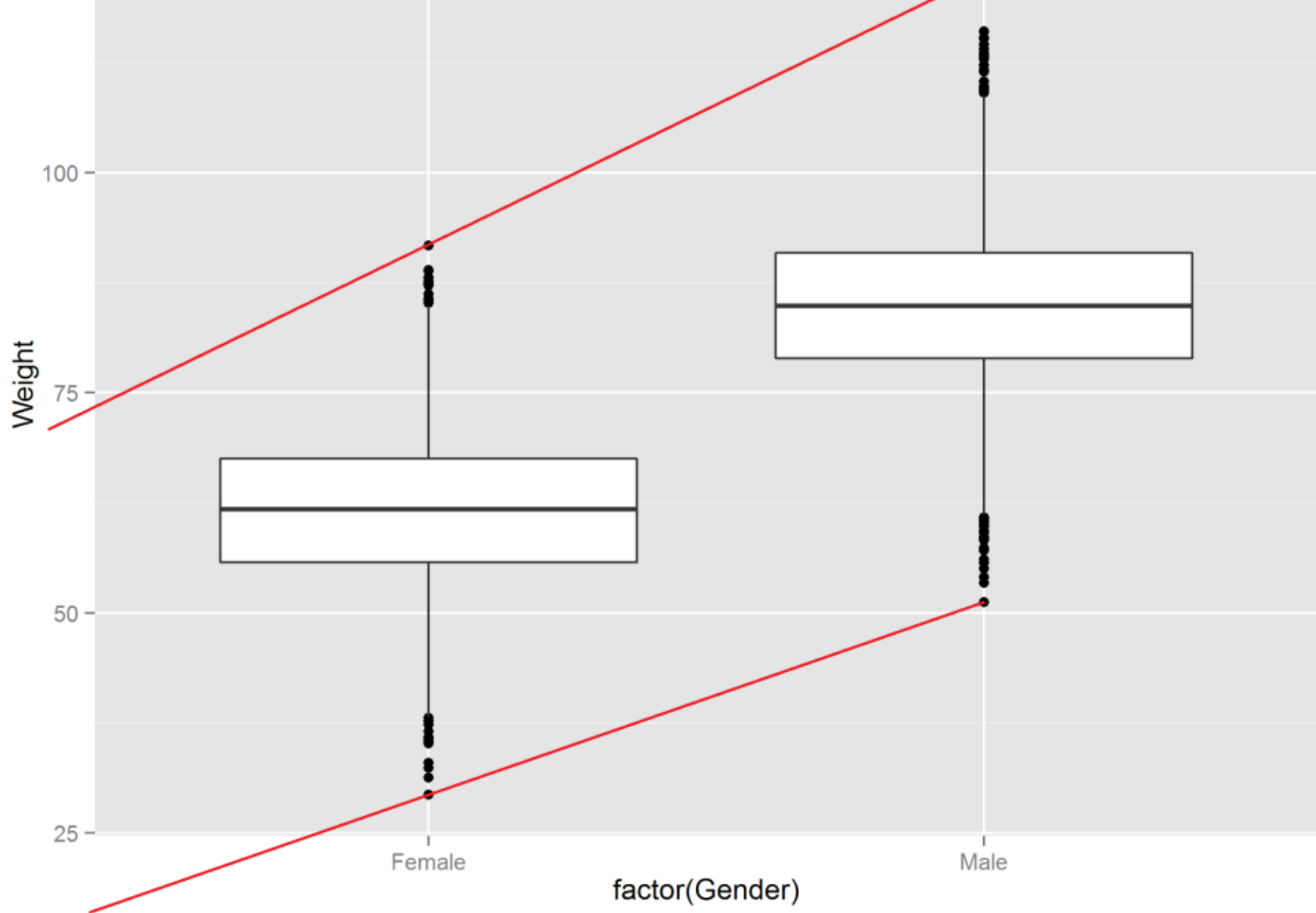
- Megjelenített dim.k: 1
- Ábrázolt összefügg.:
 - folytonos változó eloszlása
- Adategység:
 - Oszlop – az oszlop magassága az adott érték absz. gyakoriságát tükrözi

Fontos percentilisek?

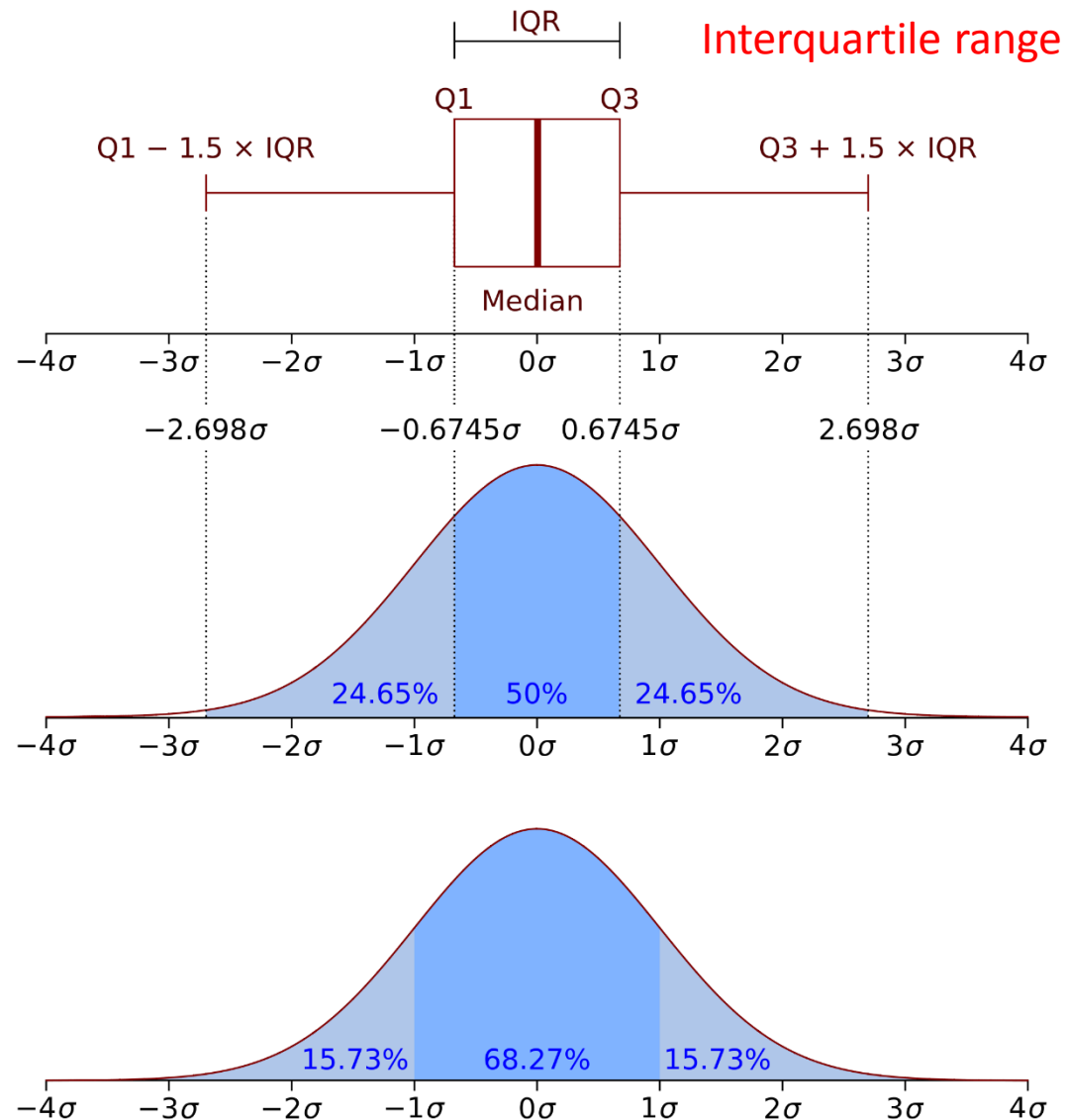
- Tervezői döntés:
 - Oszlopok szélessége?

Nők és férfiak magasságának eloszlása is szép haranggörbe

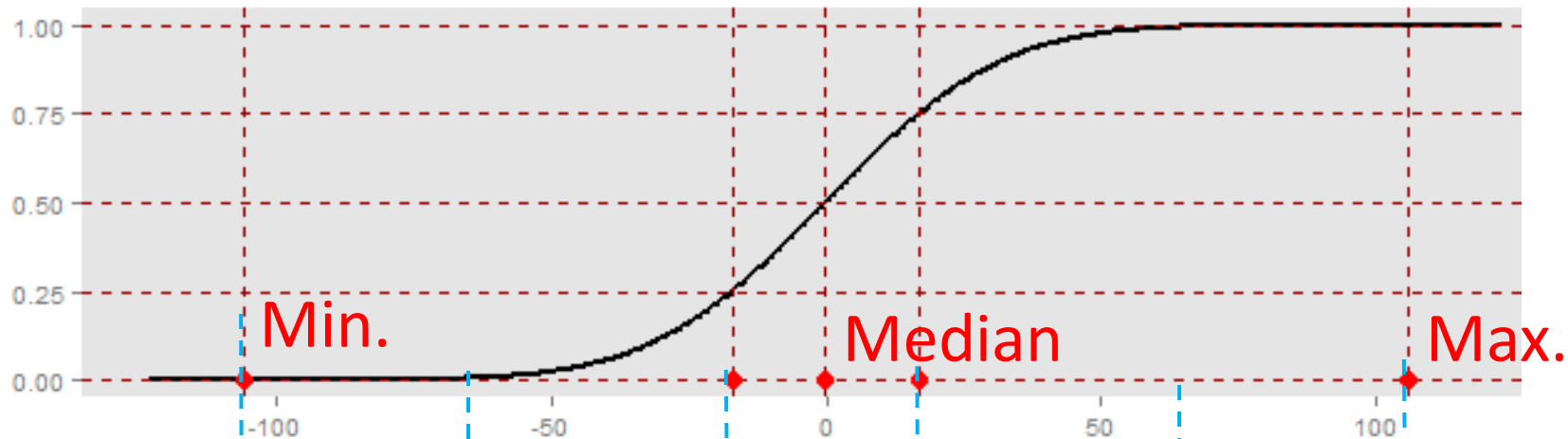




Boxplot

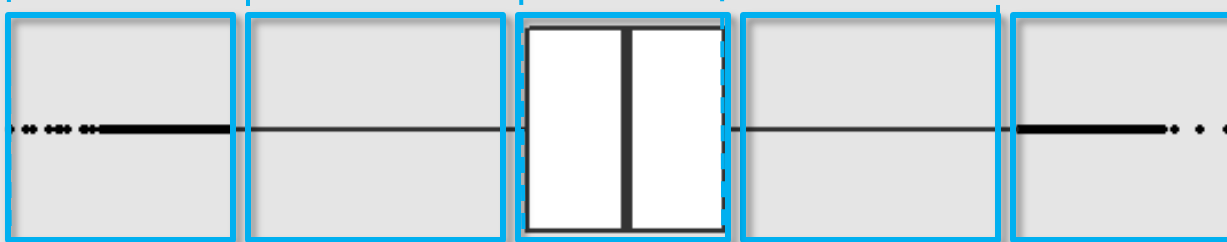


Kvalitatív gondolkozás



Extr. small Small Normal Large Extr. large

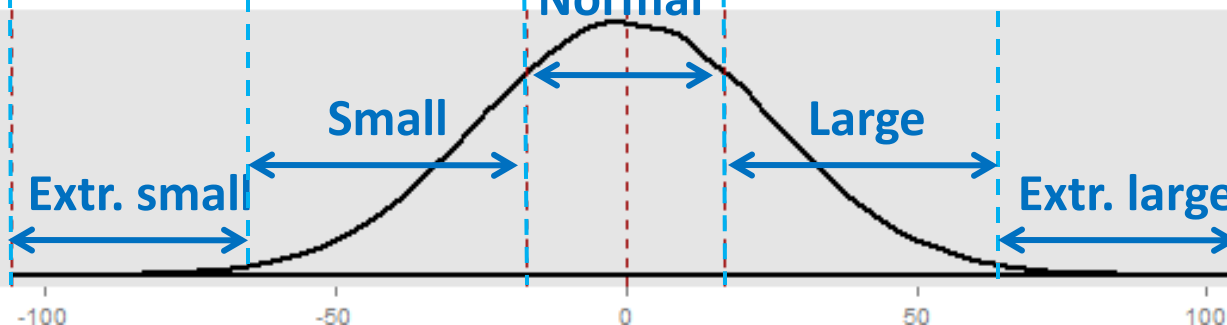
Kvalitatív
tartomány



$Q1 - 1.5 IQR$

Normal

Kvantitatív
tartomány

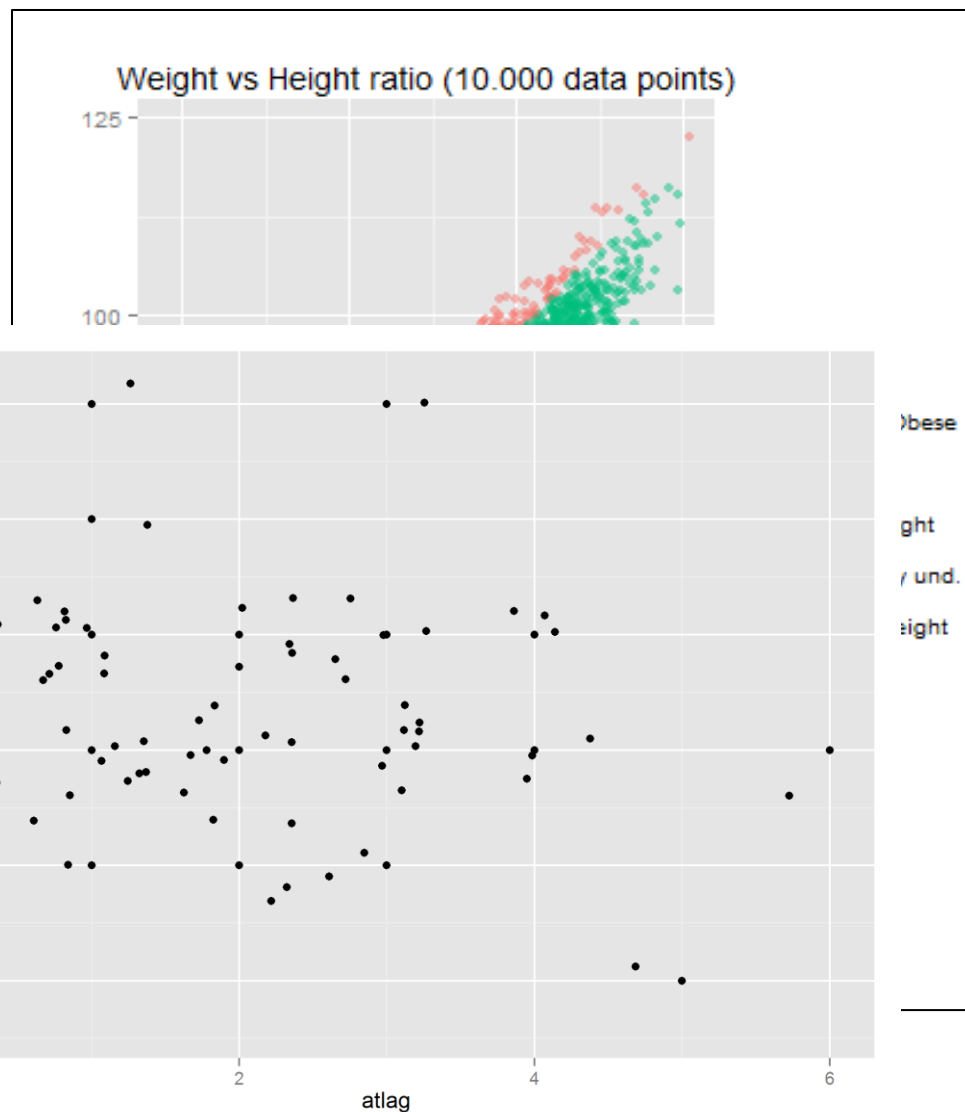


KÉT VÁLTOZÓ

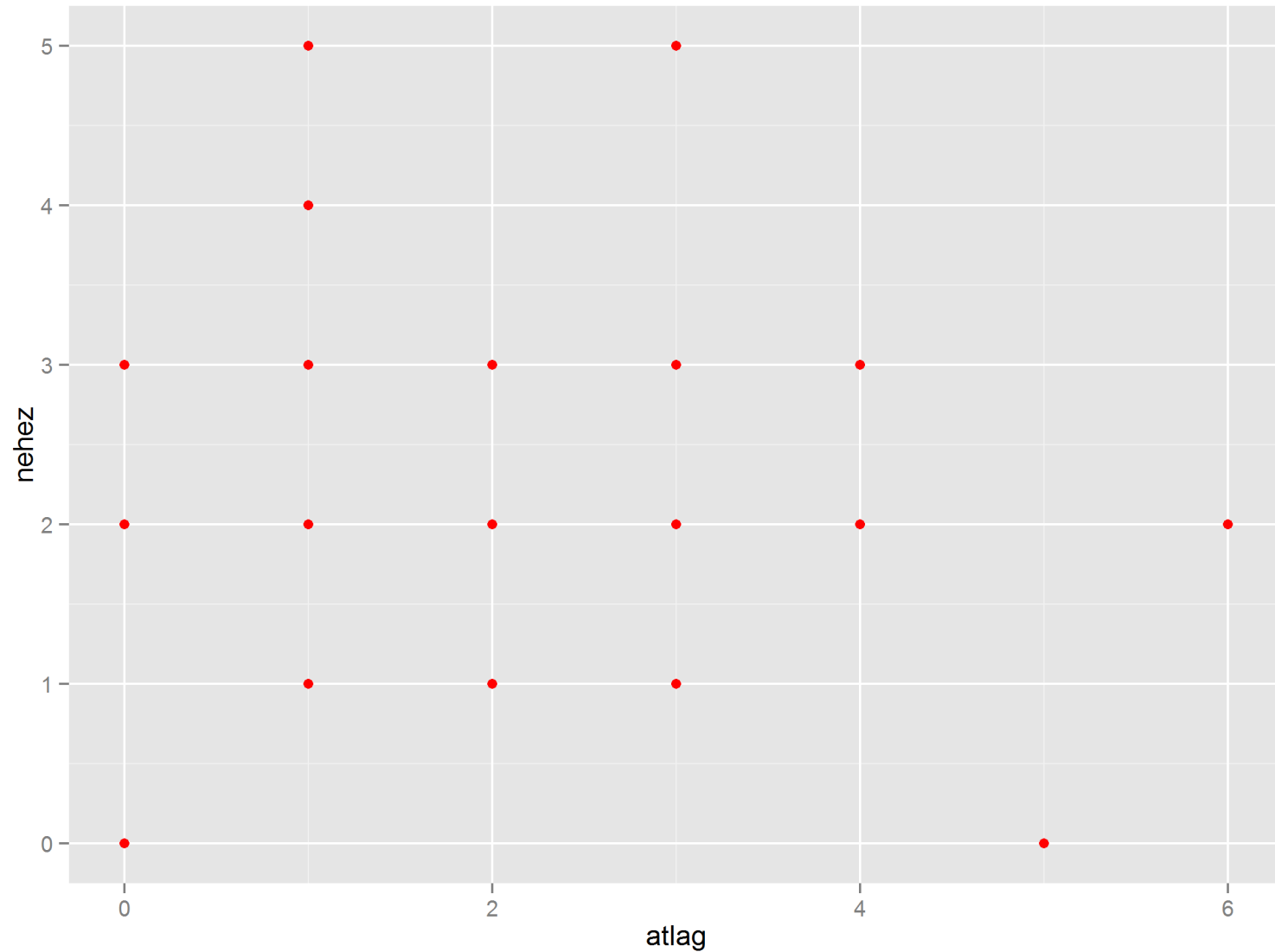
Cél: tartományok, összefüggések keresése

Pont – pont diagram (scatterplot)

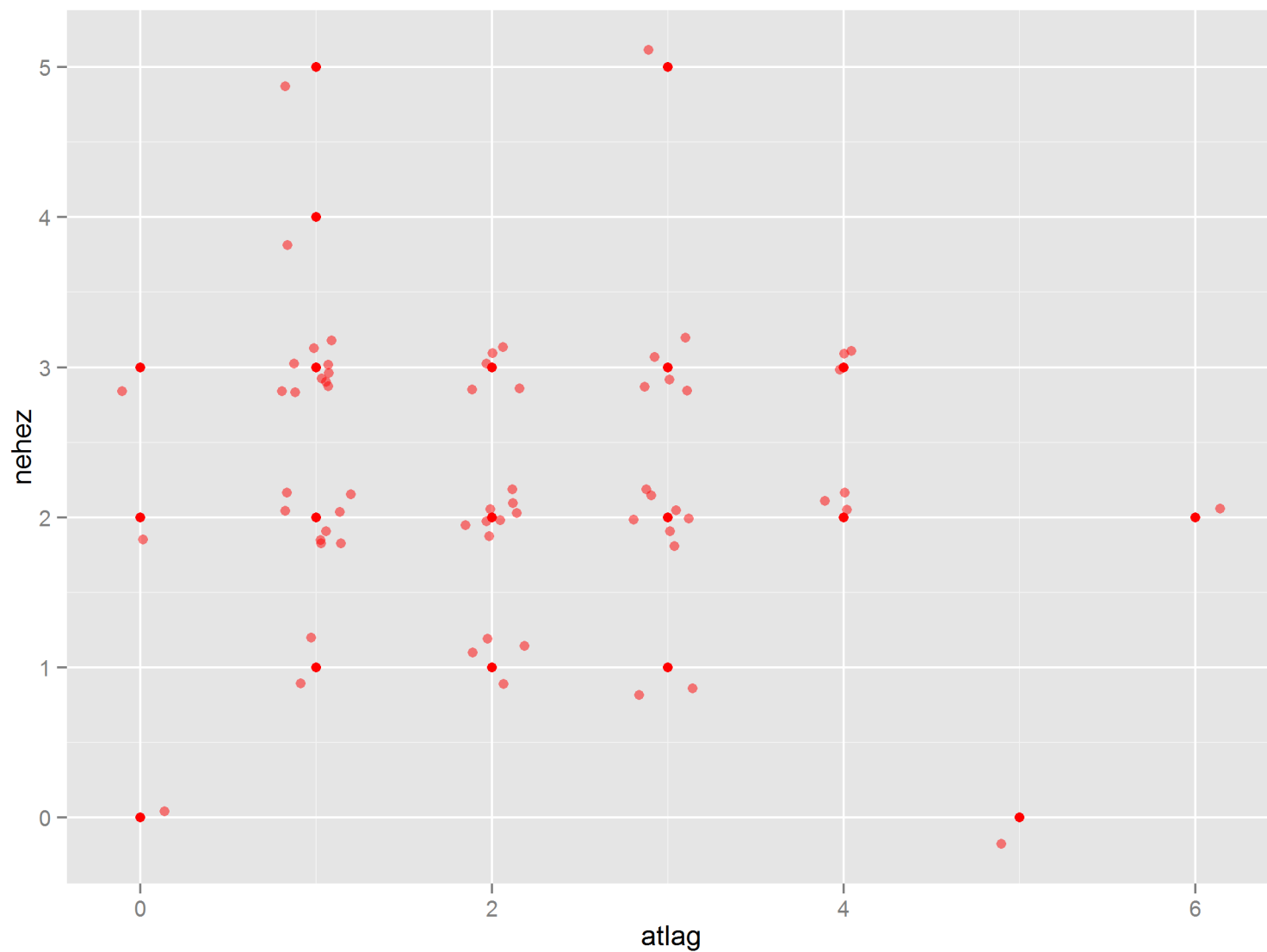
- Megjelenített dim.k: 2
- Ábrázolt összefügg.:
 - folytonos változók együttes eloszlása
- Adategység:
 - pont – $X = x_i, Y = Y_i$
előfordulás
- Korlát:
 - ha az egyik változó értéke hiányzik →
nem tudjuk felrajzolni
- Tervezői döntés:
 - Overplotting?



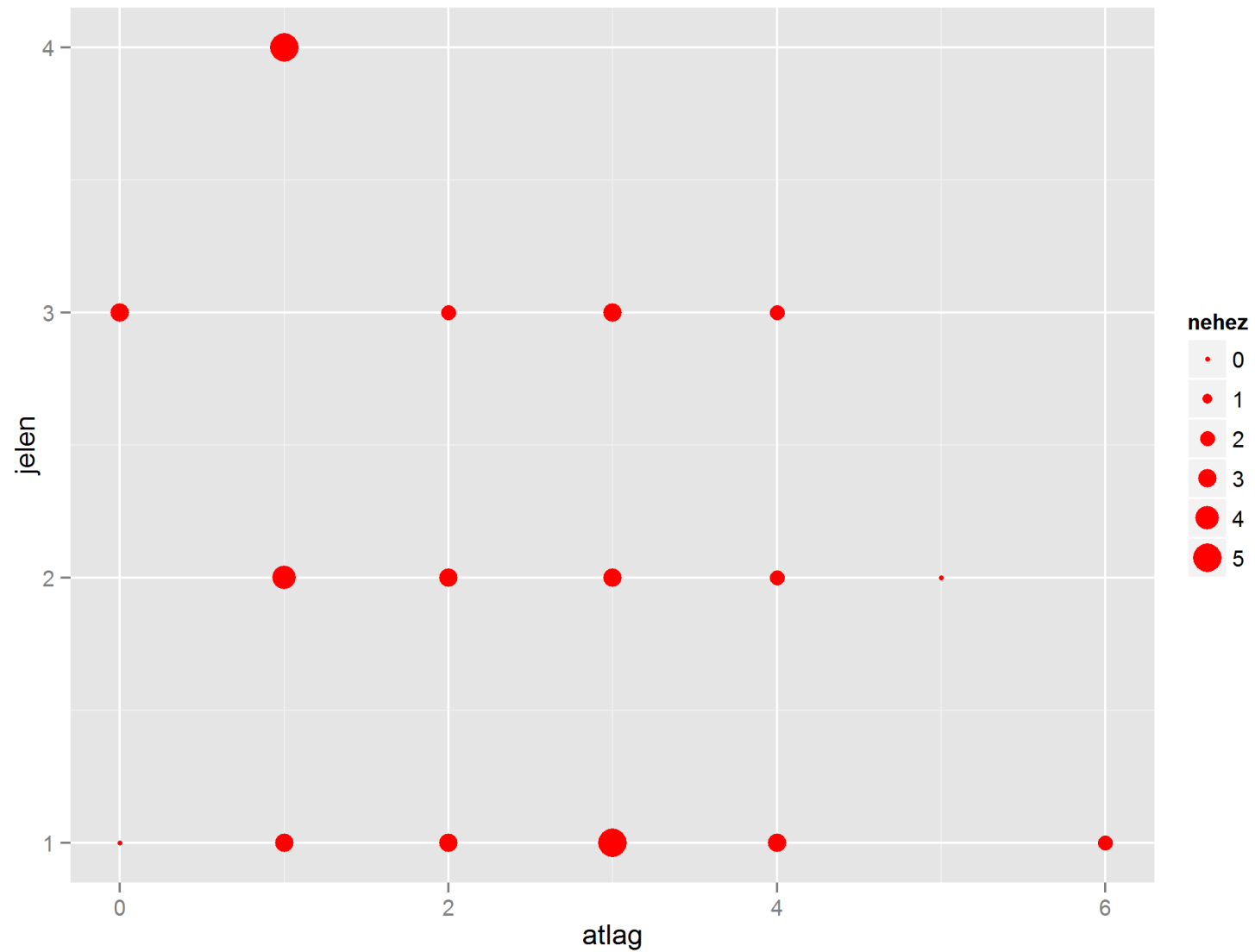
Hol volt, hol nem volt...



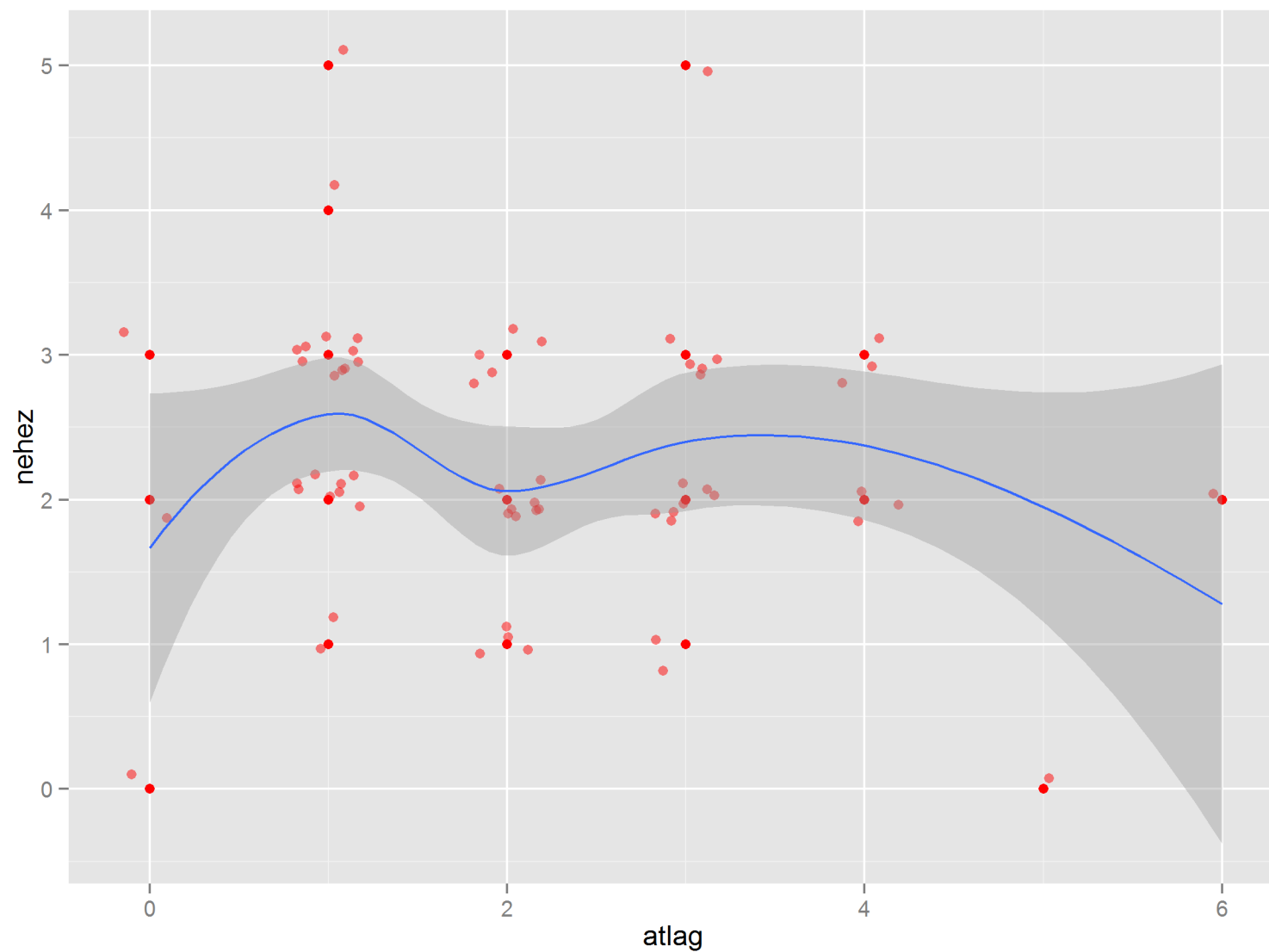
Szétszórjuk



A pontok....



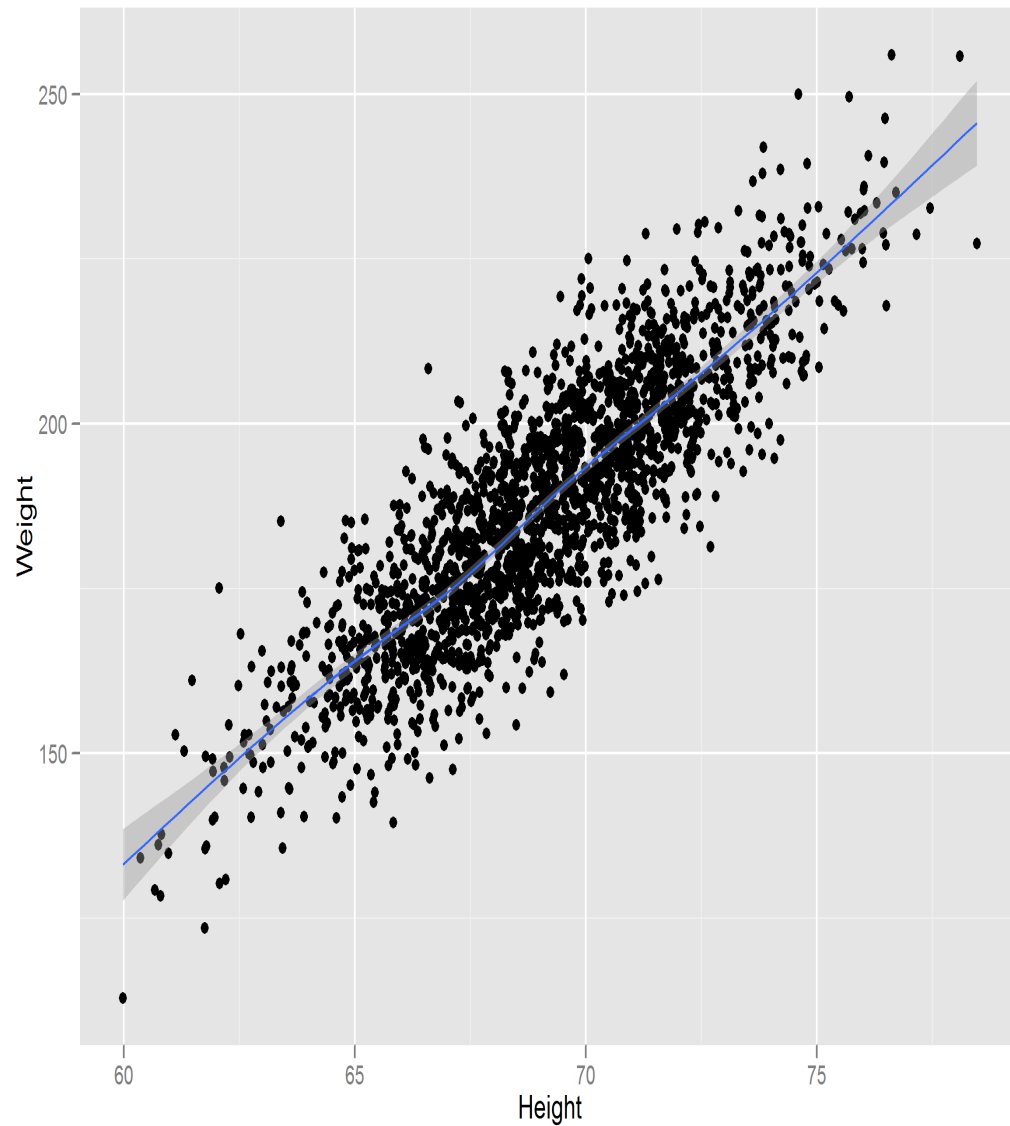
És megpróbáljuk közelíteni...



Regresszió

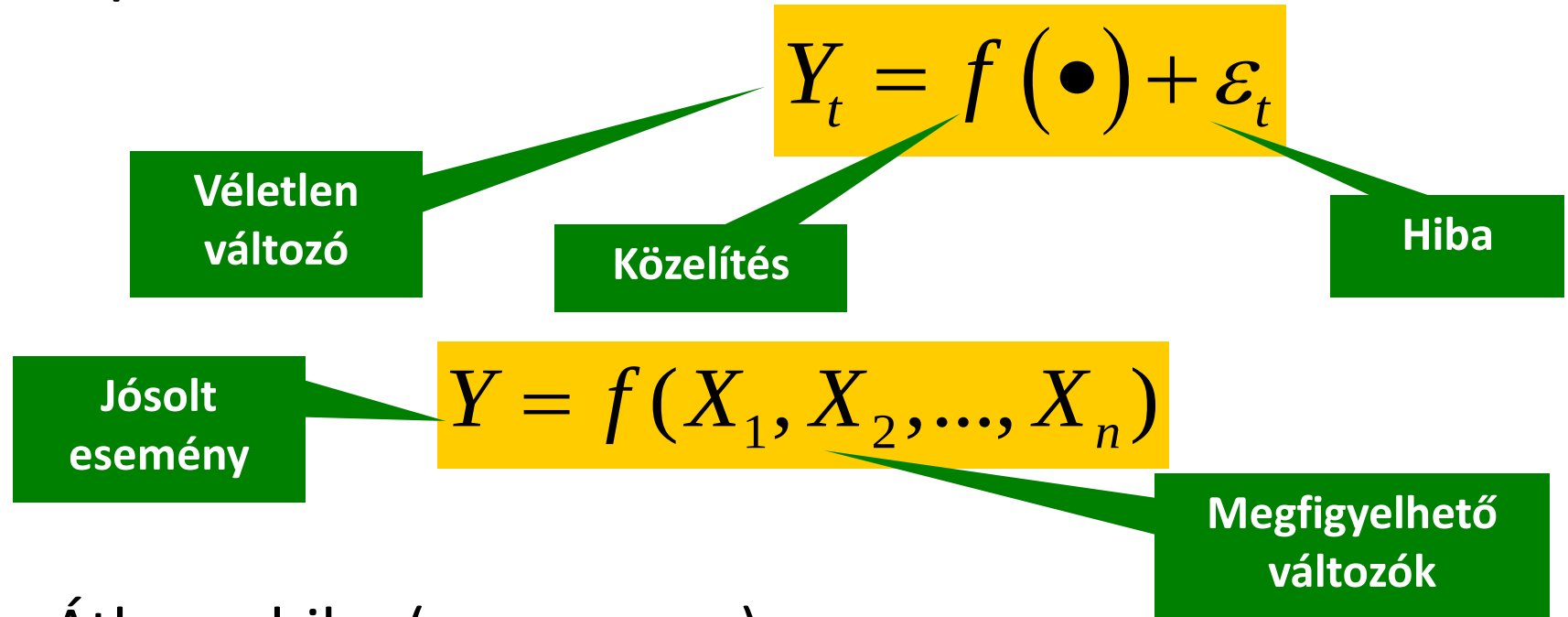
f függvény,

- bemenet:
az attribútumok értéke
- kimenet: megfigyelések
legjobb közelítése
- „ökölszabály”
- Példa:
testtömeg/magasság
együttes eloszlás
valójában egyenesre
illeszthető,

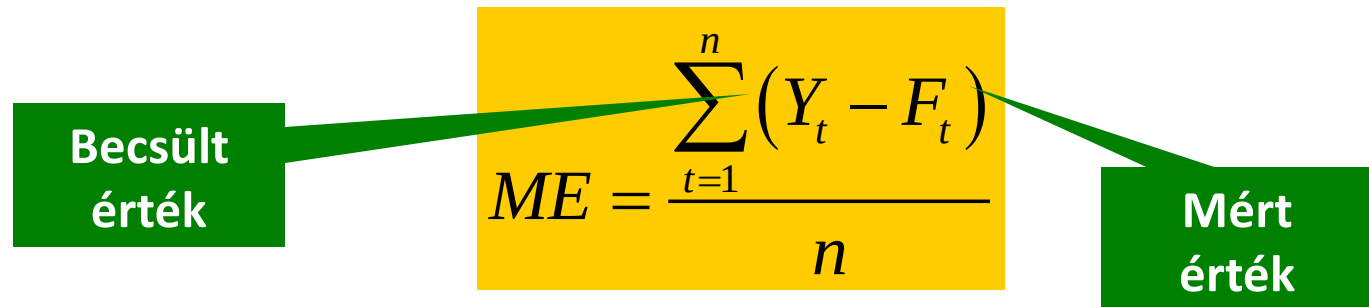


Regressziós módszerek

■ Alapelv:



• Átlagos hiba (mean error)



Lineáris regresszió

- Egyszerű lin. függvény illesztése az adatokra
 - nem vár alapvető változást a rendszer viselkedésében

$$Y = a + bX$$

- Legkisebb négyzetek módszere
 - keressük azokat az a, b paramétereket, amelyekre

$$SSE = \sum_{t=1}^n \varepsilon_t^2 = \sum_{t=1}^n (Y_t - F_t)^2 \quad \text{minimális (Sum of Squared Errors)}$$

- cél:

$$\sum_{t=1}^n (Y_t - F_t)^2 = \sum_{t=1}^n [Y_t - (a + bX_t)]^2$$

Levezetés (parc. deriválás)

$$\frac{d \sum_{t=1}^n [Y_t - (a + bX_t)]^2}{da} = \sum_{t=1}^n (-2) [Y_t - (a + bX_t)] = 0$$

$$na = \sum_{t=1}^n (Y_t - bX_t)$$

$$a = \bar{Y} - b\bar{X}$$

**X_i, Y_i a mért értékpárok
(pl. idő, terhelés)**

$$\frac{d \sum_{t=1}^n [Y_t - (a + bX_t)]^2}{db} = \sum_{t=1}^n X_t [Y_t - (a + bX_t)] = 0$$

$$\sum_{t=1}^n X_t \left[Y_t - \frac{1}{n} \sum_{t=1}^n (Y_t - bX_t) - bX_t \right] = \sum_{t=1}^n X_t Y_t - \frac{1}{n} \left(\sum_{t=1}^n X_t \right) \left(\sum_{t=1}^n Y_t \right) + \frac{1}{n} b \left(\sum_{t=1}^n X_t \right) \left(\sum_{t=1}^n X_t \right) - b \sum_{t=1}^n X_t^2 = 0$$

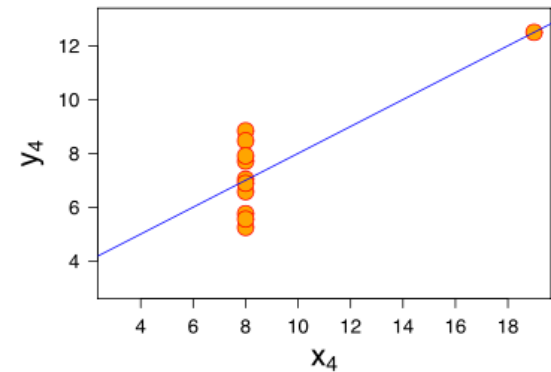
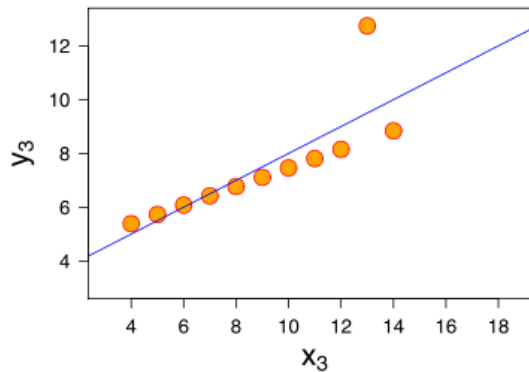
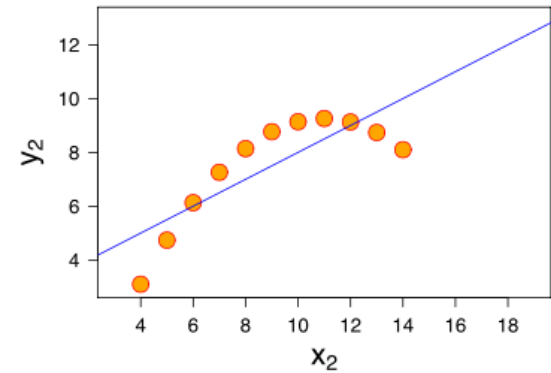
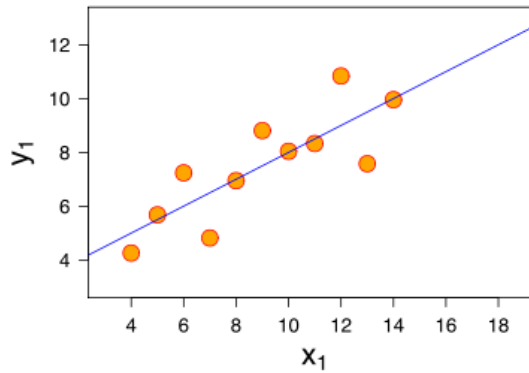
$$b = \frac{n \sum_{t=1}^n X_t Y_t - \left(\sum_{t=1}^n X_t \right) \left(\sum_{t=1}^n Y_t \right)}{n \sum_{t=1}^n X_t^2 - \left(\sum_{t=1}^n X_t \right)^2}$$

Lineáris regresszió

- Legjobban illeszkedő egyenes
- $\min(\sum_{i=1}^n |Y_i - \hat{\mu}(x_i)|^2)$, ahol $\hat{\mu}(x) = ax + b$

- **DE:**
Anscombe's quartet

- Minőségileg különböző adatok
- Azonos regressziós egyenes



Loess görbe (Locally weighted polynomial regression)

- Pont környezetében polinomiális közelítések összefűzve

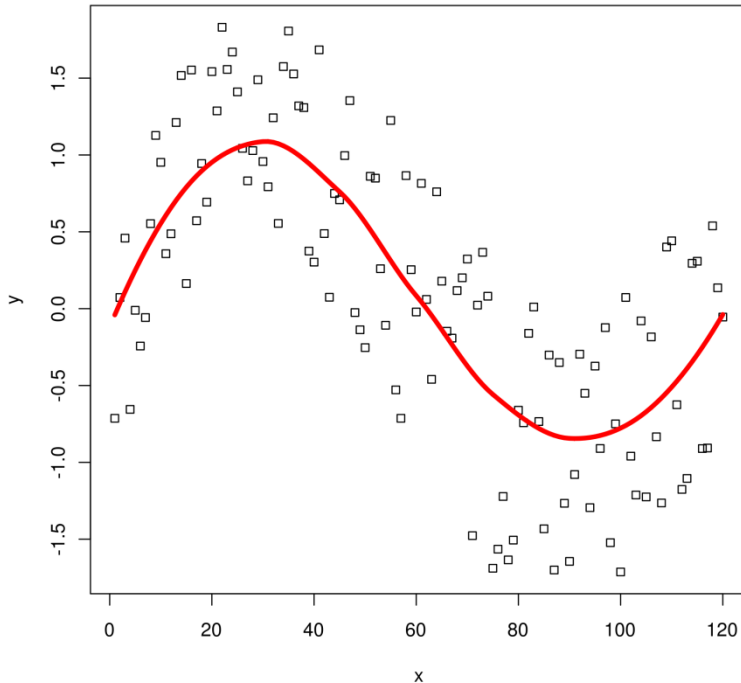
- Tipikusan 1 vagy 2 fokú
- Környezet

- Fix intervallum (span)
- Fix darabszám

$$T(u) = \begin{cases} ((1 - |u|)^3)^3 & \text{for } |u| < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$w(x_0) = T\left(\frac{|x - x_0|}{s}\right)$$

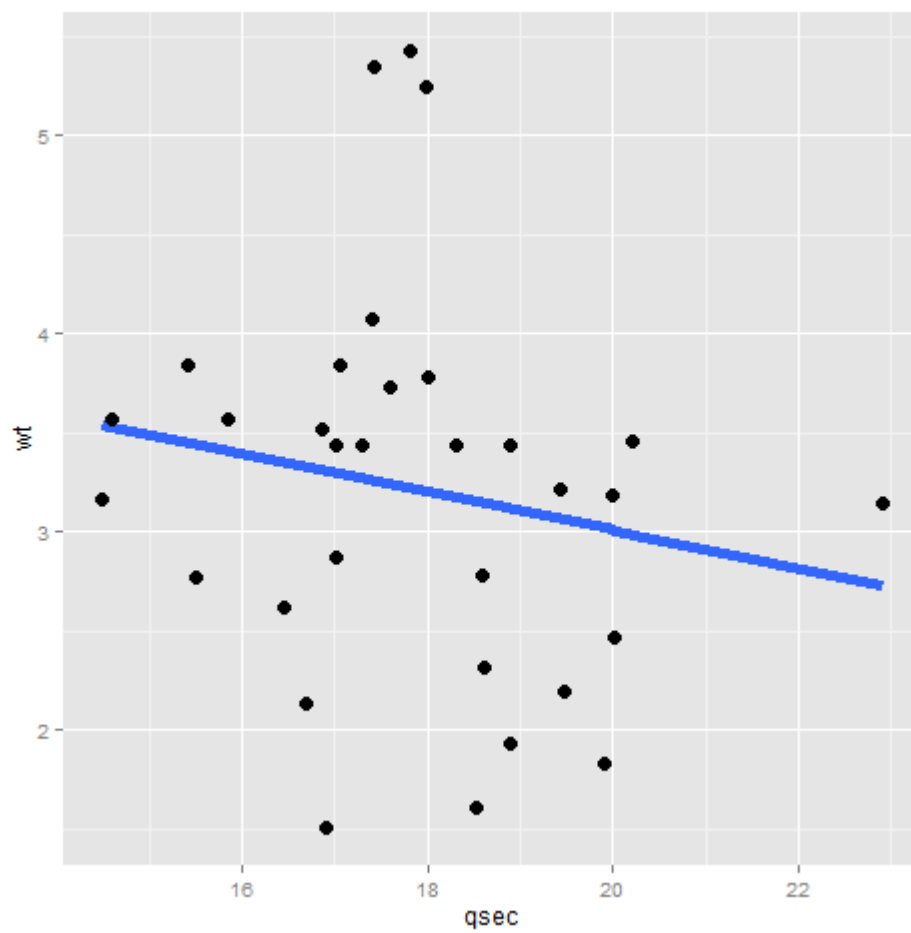
- Nagy adatkészlet
- Outlier érzékenység
- Nem ad zárt alakot



Simító görbe

$$\min \left[\left(\sum_{i=1}^n (Y_i - \hat{\mu}(x_i))^2 \right) + \lambda \int_{x_1}^{x_n} \hat{\mu}''(x)^2 dx \right]$$

- Egy adott osztályból legjobban közelítő $\hat{\mu}$ függvény
- λ simító paraméter
 - Adat követése $(\sum_{i=1}^n (Y_i - \hat{\mu}(x_i))^2)$
 - Simaság $\int_{x_1}^{x_n} \hat{\mu}''(x)^2 dx$
 - $\lambda = 0$ esetén interpolációs görbe
 - $\lambda \rightarrow \infty$ esetében lineáris regresszió



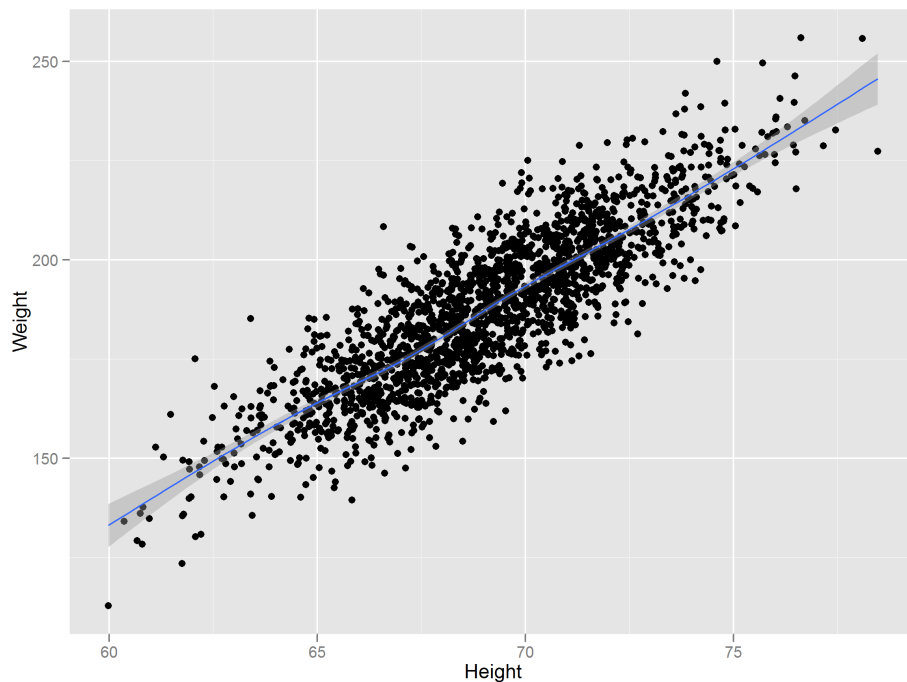
Regresszió

- Cél:

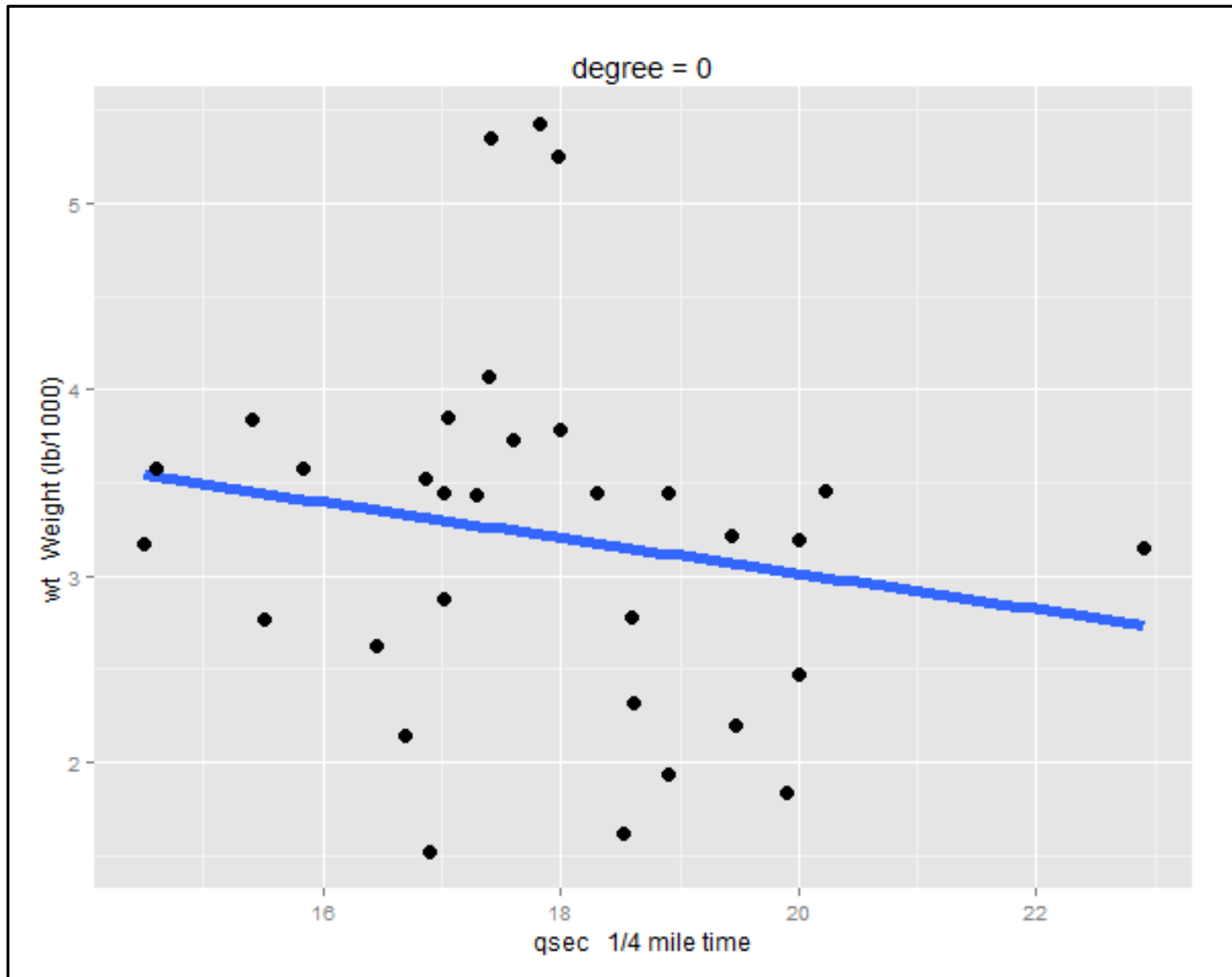
megtalálni egy olyan f függvényt, amelynek inputja az attribútumok értéke, az outputja pedig a lehető legjobban közelíti (négyzetes hibaérték) a valóságot

- Példa:

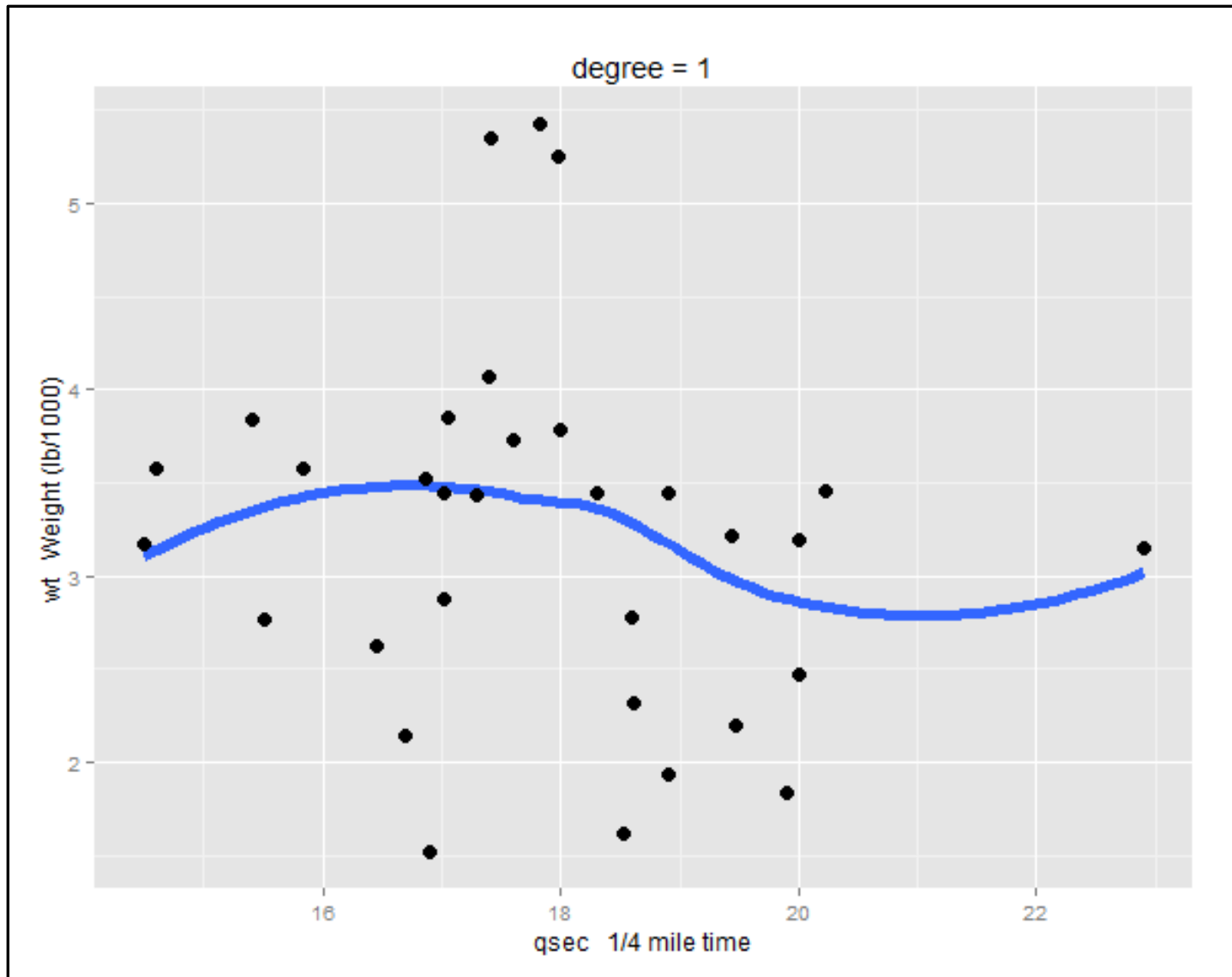
- testtömeg/magasság együttes eloszlás valójában egyenesre illeszthető,
- web forgalom jóslása

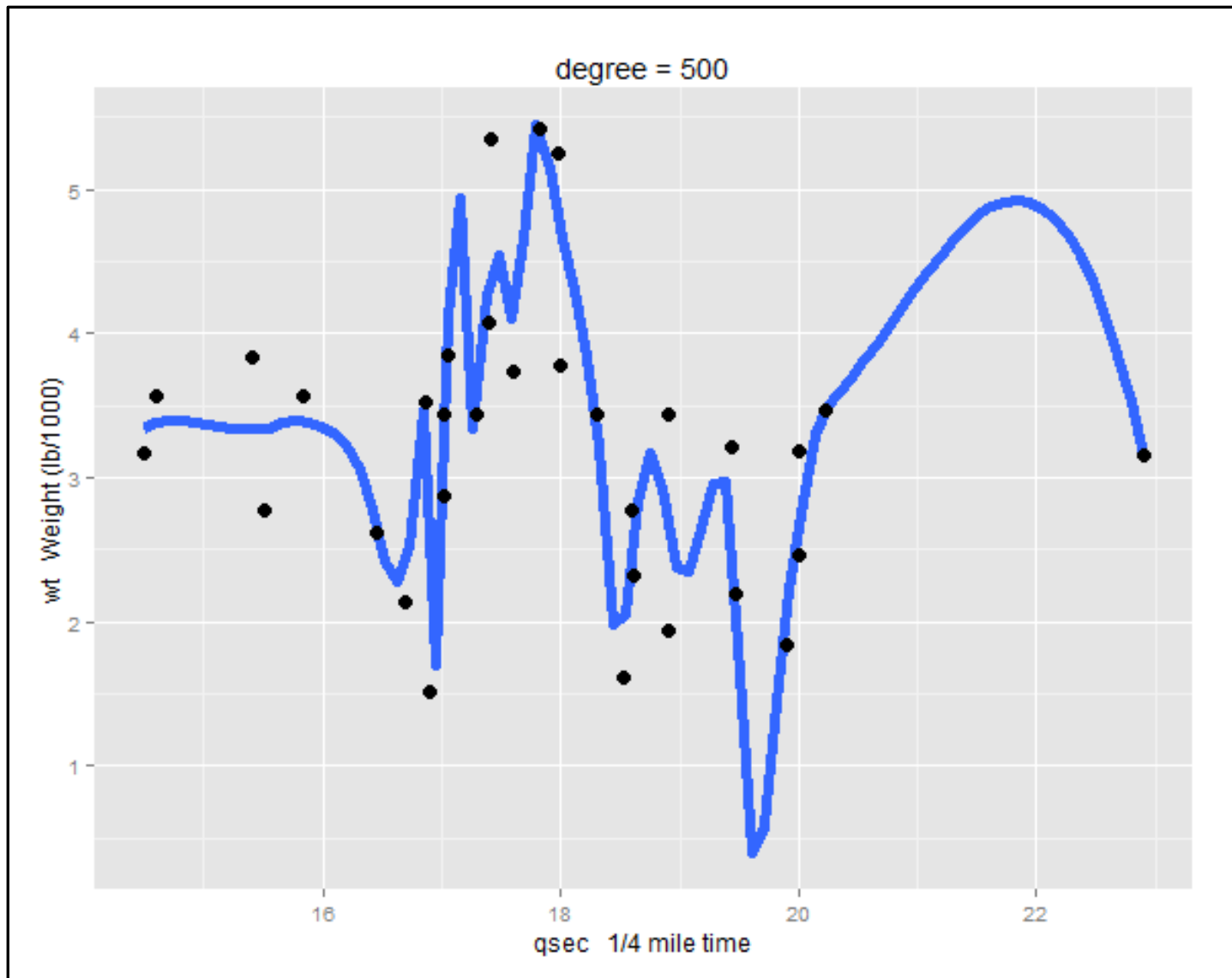


$$\lambda = 0$$



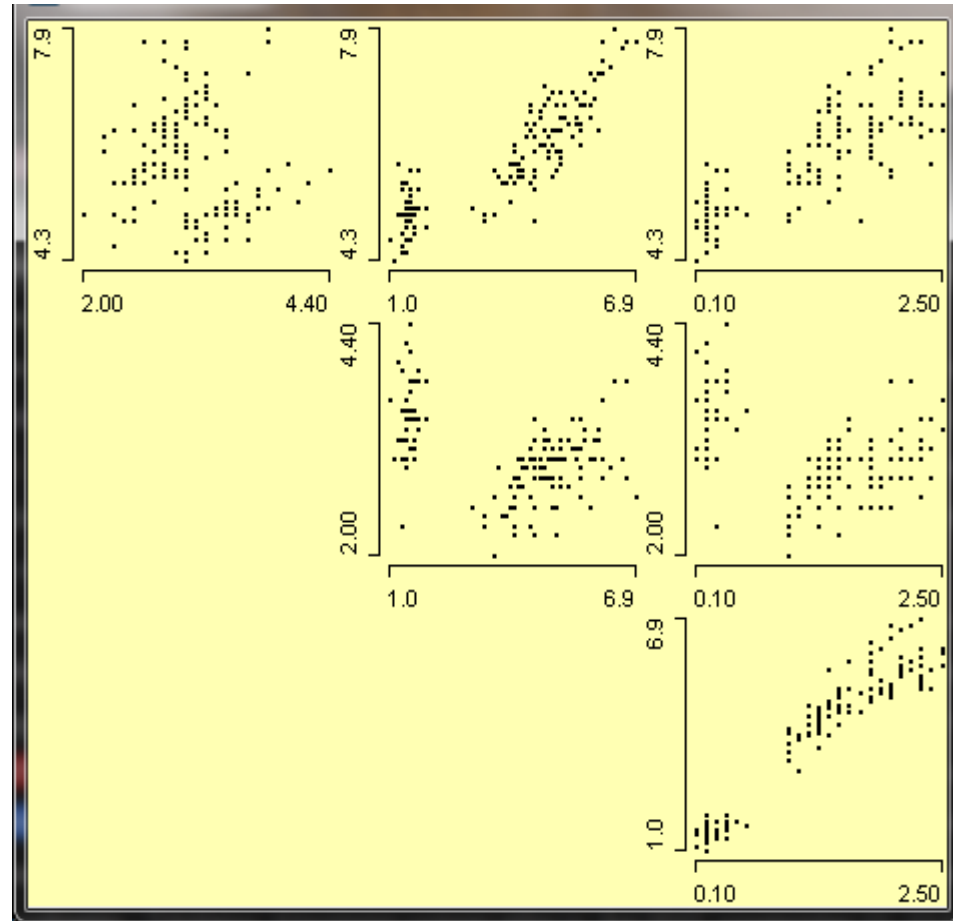
$$\lambda = 1$$





Scatterplot mátrix

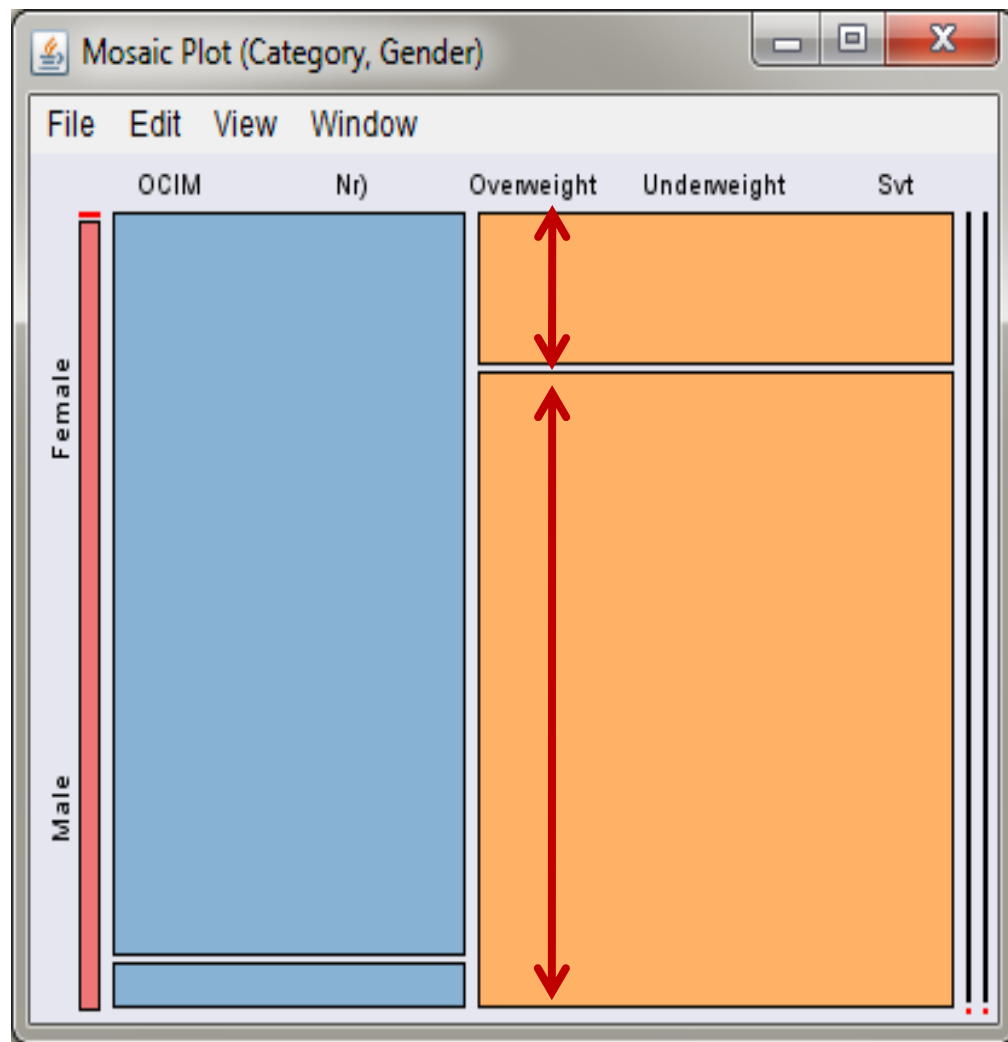
- Megjelenített dim.k: n
- Ábrázolt összefügg.:
 - A változópárok együttes eloszlása
- Adategység:
 - Scatterplot – minden diagram a neki megfelelő változók együttes eloszlását mutatja be



Mozaik diagram (mosaic plot)

- Megjelenített dim.k: 2
- Ábrázolt összefügg.:
 - két diszkrét változó együttes eloszlása
- Adategység:
 - Téglalap – a téglalap *területe* arányos az $(X = x_i, Y = y_i)$ értékpárok gyakoriságával
- Korlát:
 - Sorfolytonos olvasása nehézkes

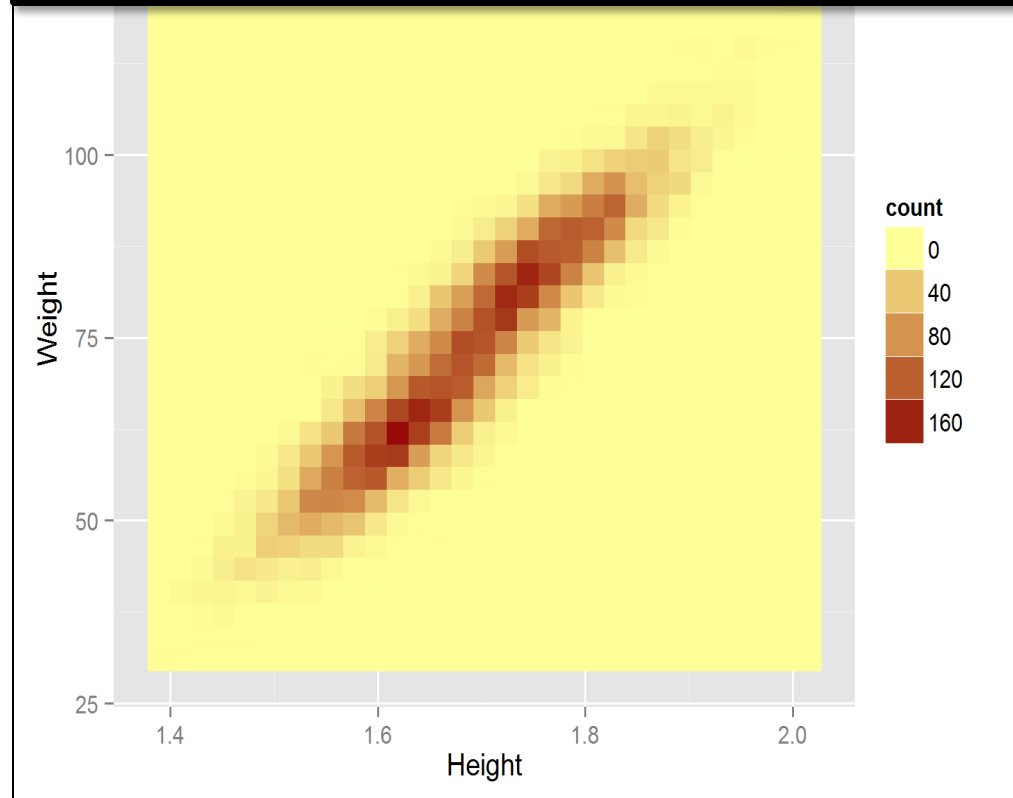
A túlsúlyosak nagy része férfi!



Hő térkép (heat map)

- Megjelenített dim.k: 3
- Ábrázolt összefügg.:
 - sűrű 3D struktúrák összefüggései
- Adategység:
 - tile – azonos „magasságú” összefüggő terület rész
- 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.:

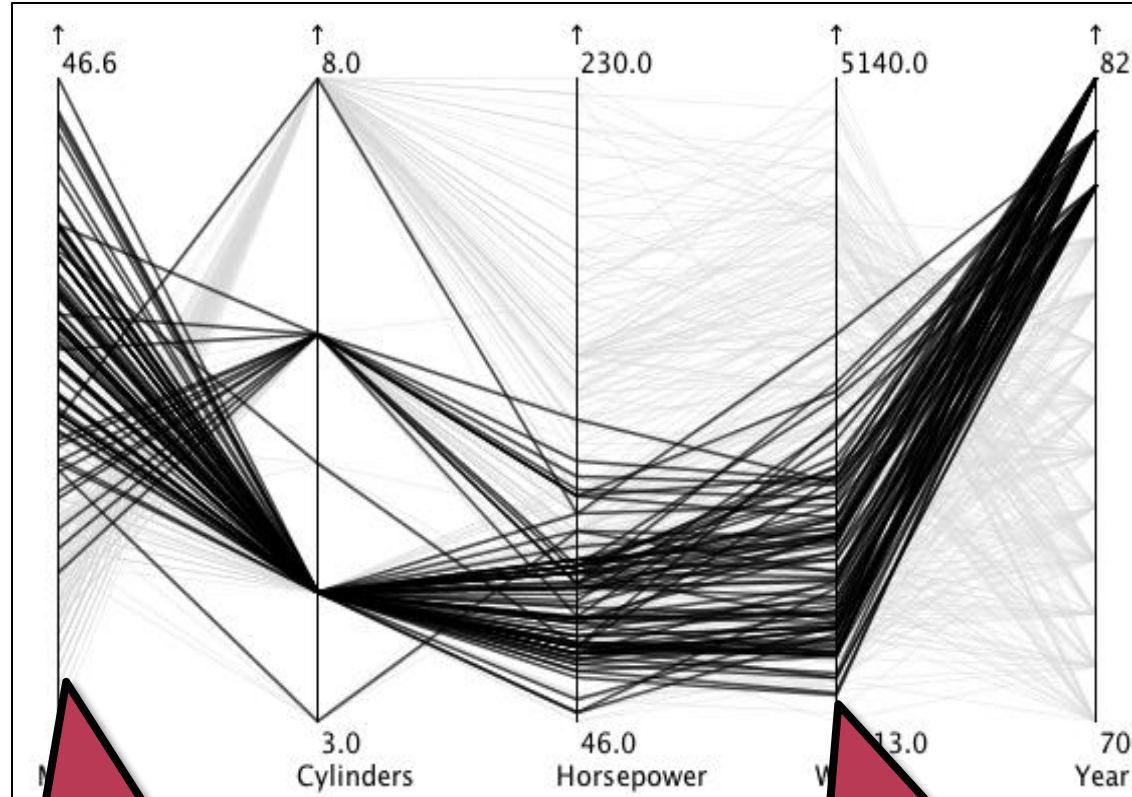
- Rekordok/attribútumok hasonlósága

- Adategység:

- Törött vonal – az egyes attribútumtengelyeken felvett értékek rendezett sorozata

- Korlátok:

- Tengelyek (attribútumok) más mértékegysége/nagyságrendje stb torzíthat

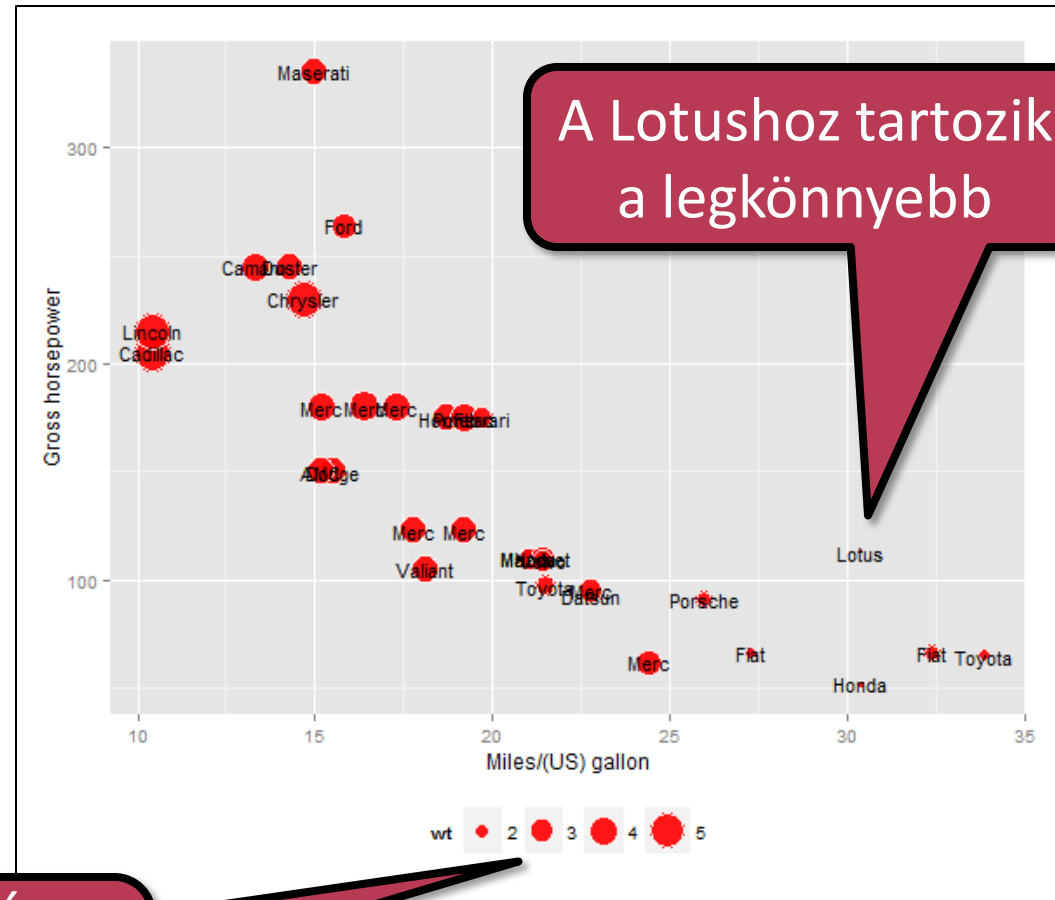


... de a fogyasztás
nagyobb

Az új autókban a tömeg
kisebb...

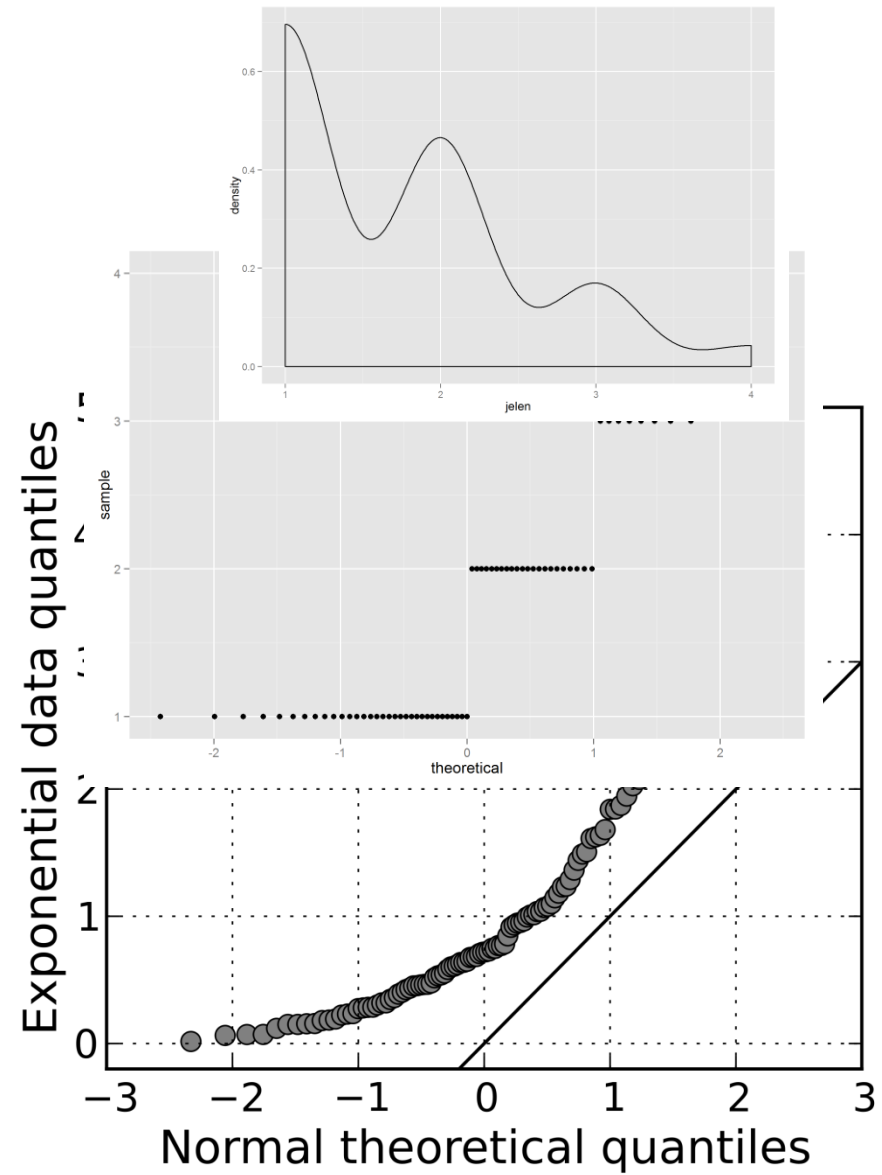
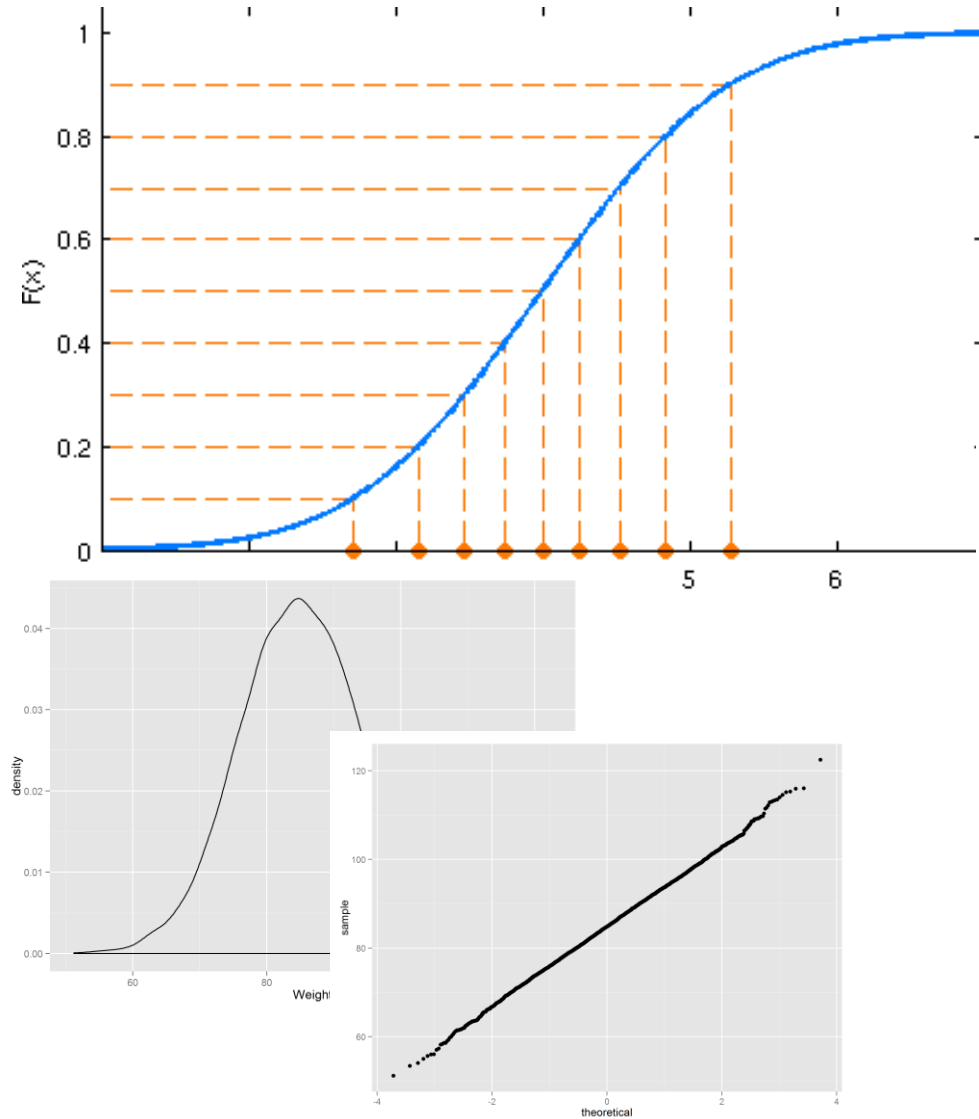
Buborék diagram (bubble chart)

- 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)



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

qqplot



Data transformation: Box-Cox

$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log(y_i), & \text{if } \lambda = 0. \end{cases}$$

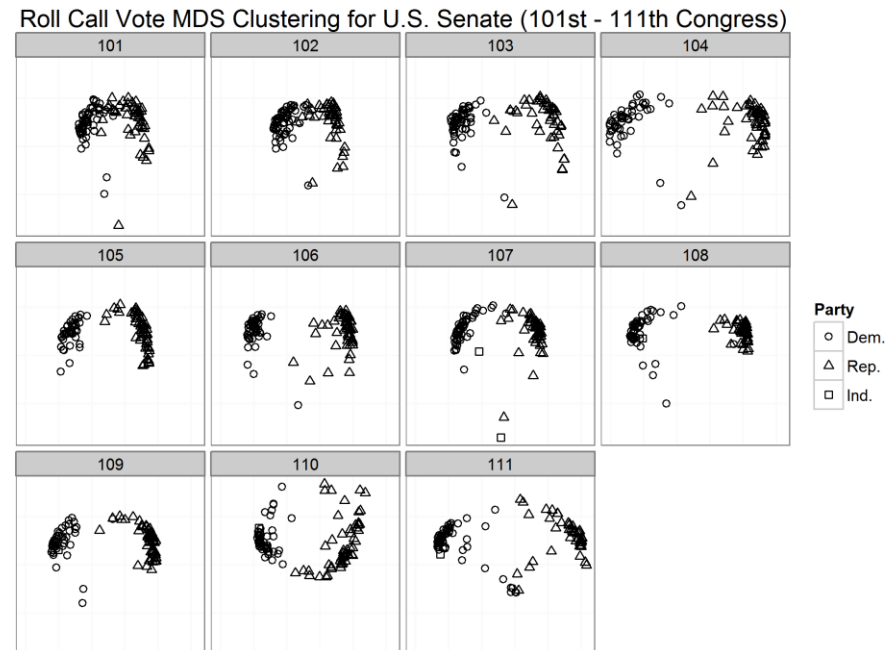
Klaszterezés

- Cél:

egy halmaz elemeit csoportokba sorolni úgy, hogy az egymáshoz "közel" lévő elemek egymáshoz "hasonlóak"

- Példa:

- ajánló rendszerek R csomagokról,
- szenátusi tagok klaszterezése szavazatuk alapján



PCA

- Cél:
megtalálni a rekordot legjobban jellemző faktorokat
- Példa:
 - Tőzsdei árfolyamok közül melyek határozzák meg legjobban a BUX index alakulását?

