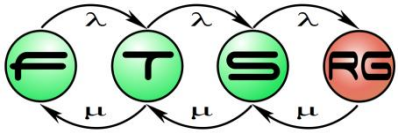
Vizuális adatanalízis



Exploratoratory 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 nehéztüzérség nélkül
- Tukey, 60-as évek közepe
- Robusztus statisztika
  - Csökkentjen az érzéekenység a mérési hibára
- Nemparametrikus statisztika
  - Ne kelljen az ismeretlen eloszlásra feltételezéseket tenni

- http://www.visual-analytics.eu/
- http://www.rosuda.org/mondrian/

EGYEDI VÁLTOZÓK

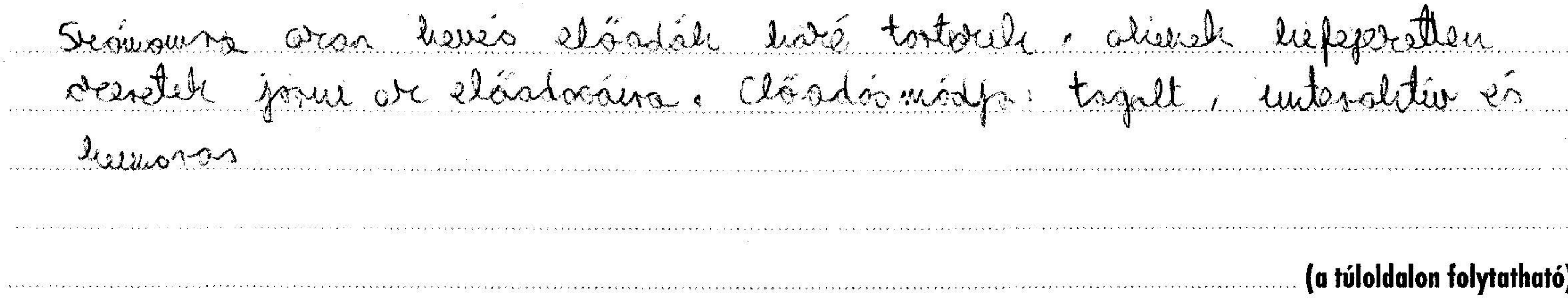
Egy kis példa: OHV

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	kerdoiv a	1.kerdes	2.kerdes	3.kerdes	4.kerdes	5.kerdes	6.kerdes	7.kerdes	8.kerdes	9.kerdes	10.kerdes	11.kerdes	12.kerdes	13.kerdes	14.kerdes	15.kerdes	16.kerdes
2	10015.tif	3	2	2	2	1	3	3	3	1	2	2	1	1	2	1	2
3	10013.tif	3	2	2	2	1	3	3	3	1	1	1	3	1	1	1	2

```
jelen,atlag,problema,gyak,segeda,előtan,nehez,tempo,ajánl,logikus,öf,figyele
1,1,3,2,2,1,3,3,3,3,2,3,1,1,2,3
1,1,3,6,3,6,2,2,4,4,3,3,1,1,1,3
3,3,2,2,2,3,2,2,2,2,2,2,1,1,1,2
1,3,3,6,4,3,2,3,3,2,1,2,1,1,1,2
1,1,2,2,2,1,3,3,2,1,2,2,1,1,1,2
3,0,1,2,2,2,2,2,2,2,1,2,1,2,1,1
1,2,5,6,3,6,1,1,3,3,3,2,1,3,2,3
1,2,3,6,3,6,1,3,2,1,1,1,1,1,1,2
```

CSV és rövid nevek

Nehéz értelmezni...

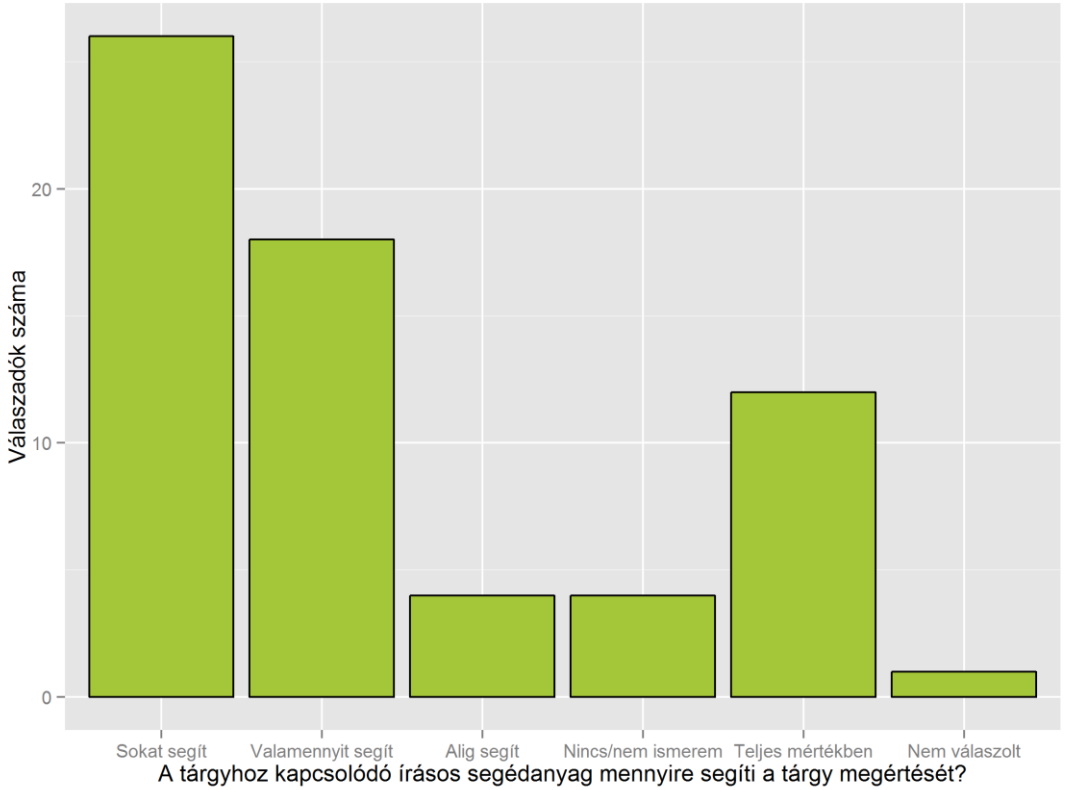


Számomra azon kevés előadók közé tartozik, akiknek kifejezetten érdemes jönni az előadására. Előadásmódja: tagolt, interaktív és humoros

(a túloldalon folytatható)

Oszlopdiagram (bar chart)

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

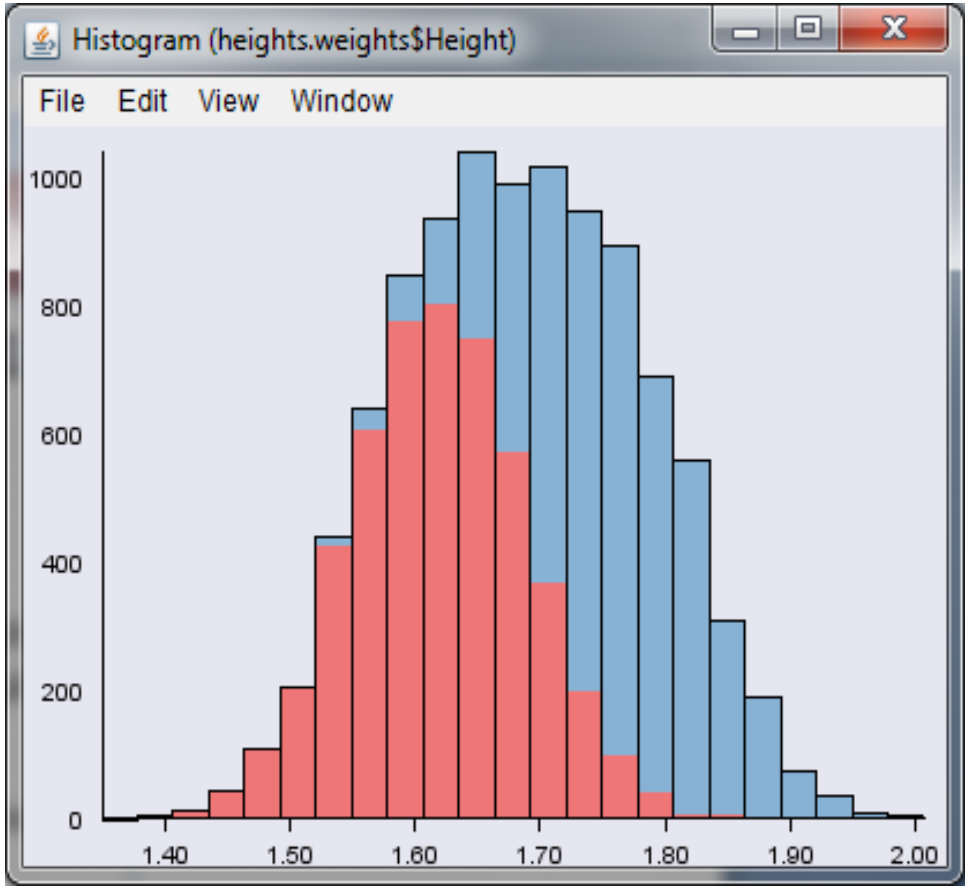


Hisztogram

- Megjelenített dimenziók száma: 1
- Ábrázolt összefügg.:
  - Folytonos változó eloszlása
- Adategység:
  - Oszlop – az oszlop magassága az adott érték abs. gyakoriságát tükrözi

Fontos percentilisek?

- Tervezői döntés:
  - Oszlopok szélessége?



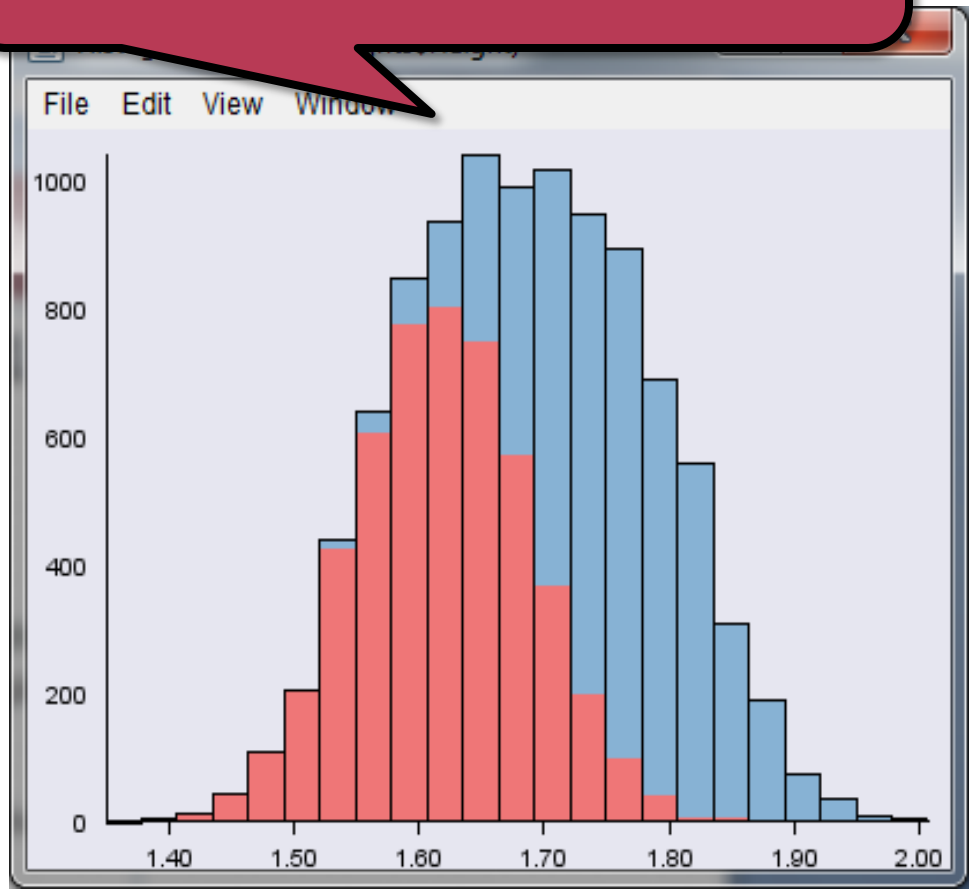
Hisztogram

- Megjelenített dimenzi
- Ábrázolt összefügg.:
  - Folytonos változó eloszlása
- Adategység:
  - Oszlop – az oszlop magassága az adott érték abs. gyakoriságát tükrözi

Fontos percentilisek?

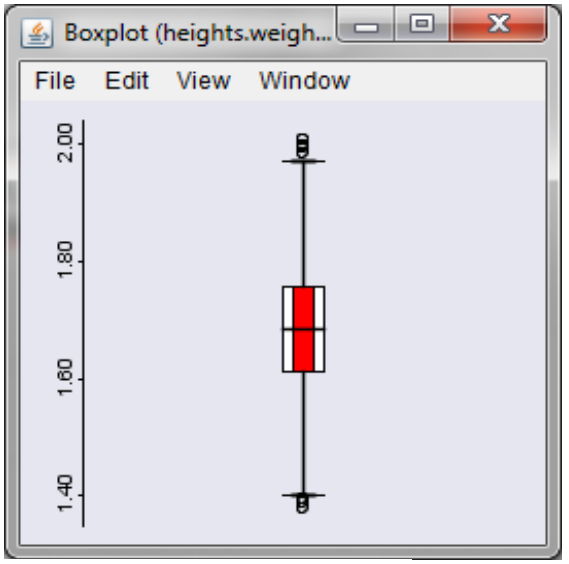
- Tervezői döntés:
  - Oszlopok szélessége?

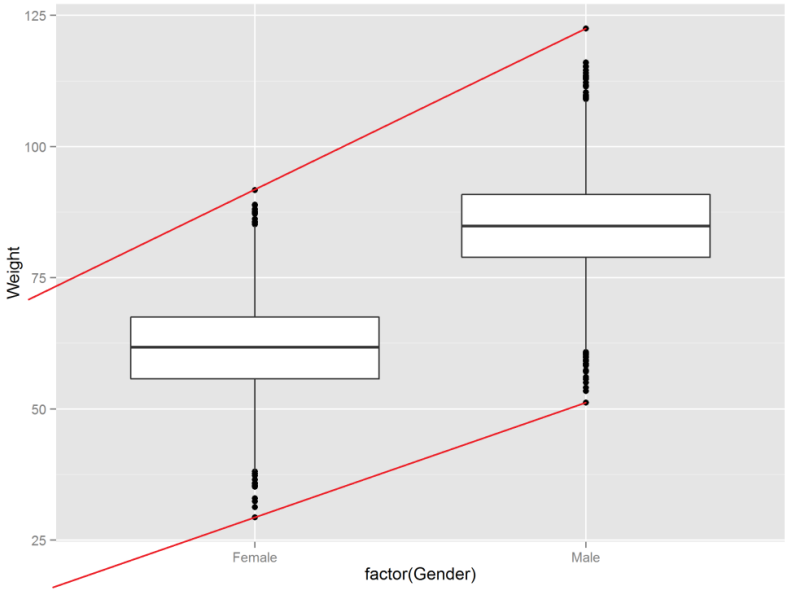
Nők és férfiak magasságának eloszlása is szép haranggörbe



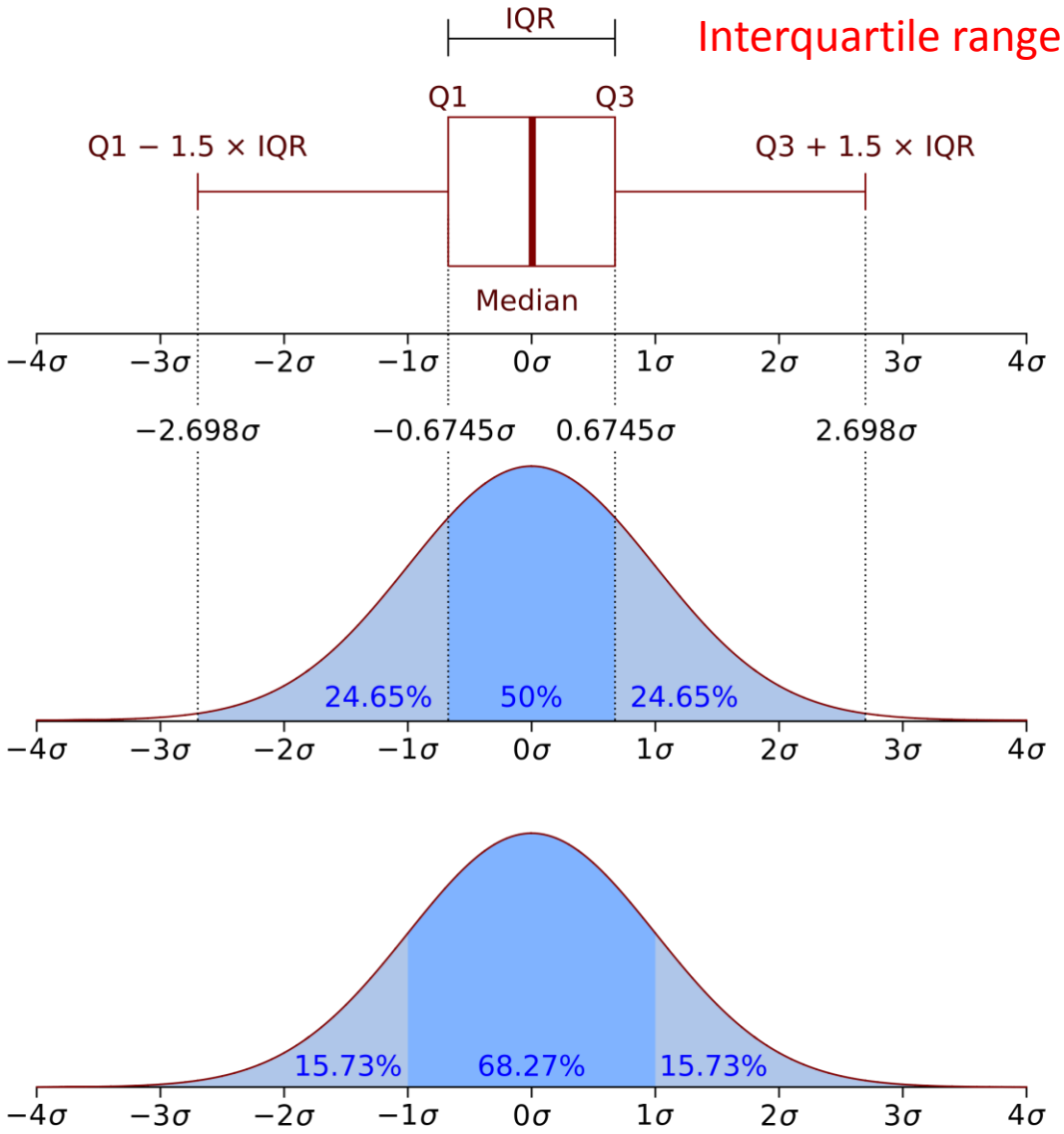
Doboz diagram (boxplot)

- Megjelenített dimenziók száma: 1
- 5 értékkel jellemzésként
- Ábrázolt összefügg.:
  - Folytonos változó fontos percentilisei
- 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

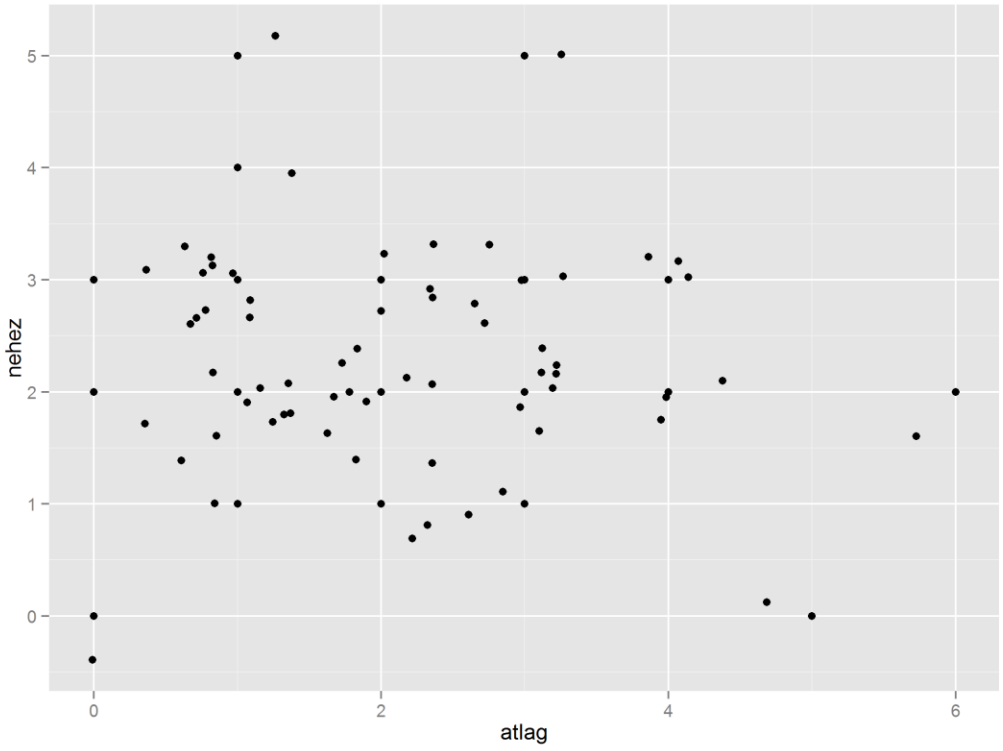

IQR
Interquartile range
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%
15.73%
68.27%
15.73%

KÉT VÁLTOZÓ

Cél: tartományok, összefüggések keresése

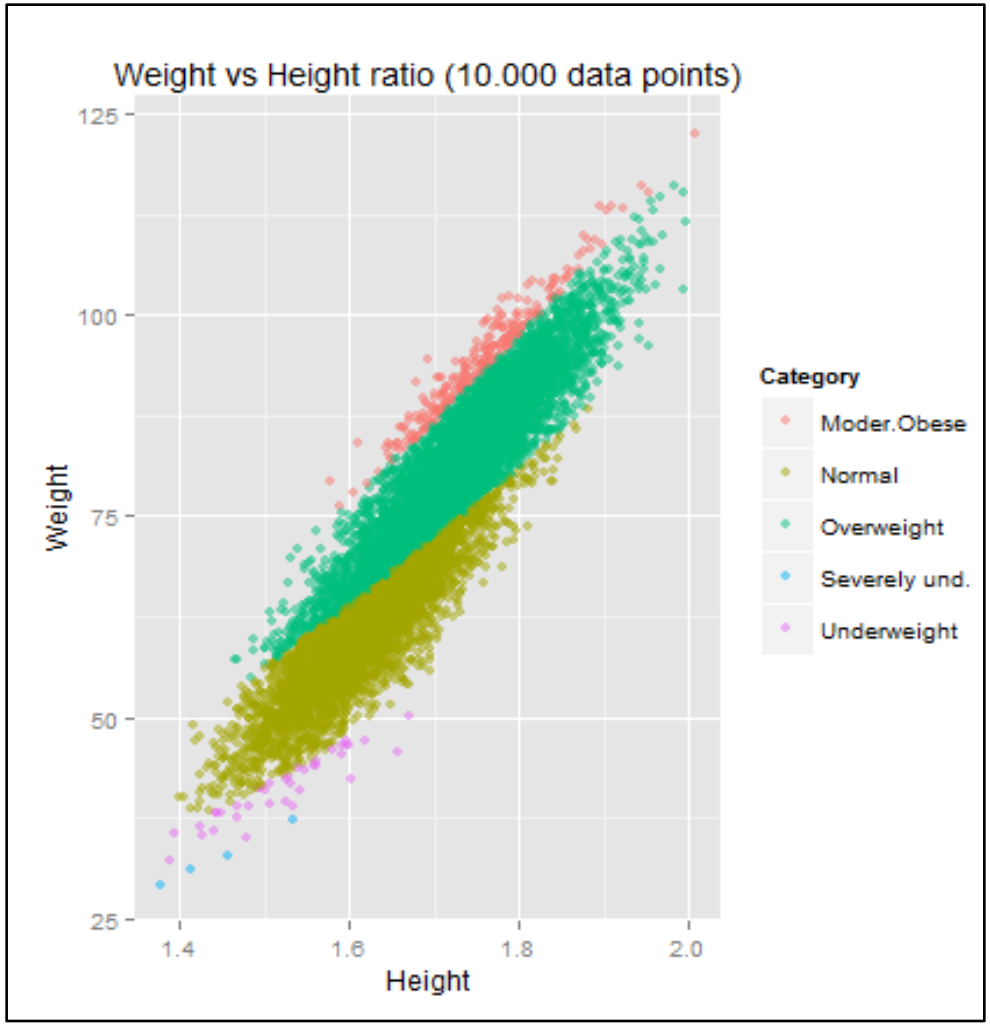
Pont – pont diagram (scatterplot)

- Megjelenített dimenziók száma: 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?

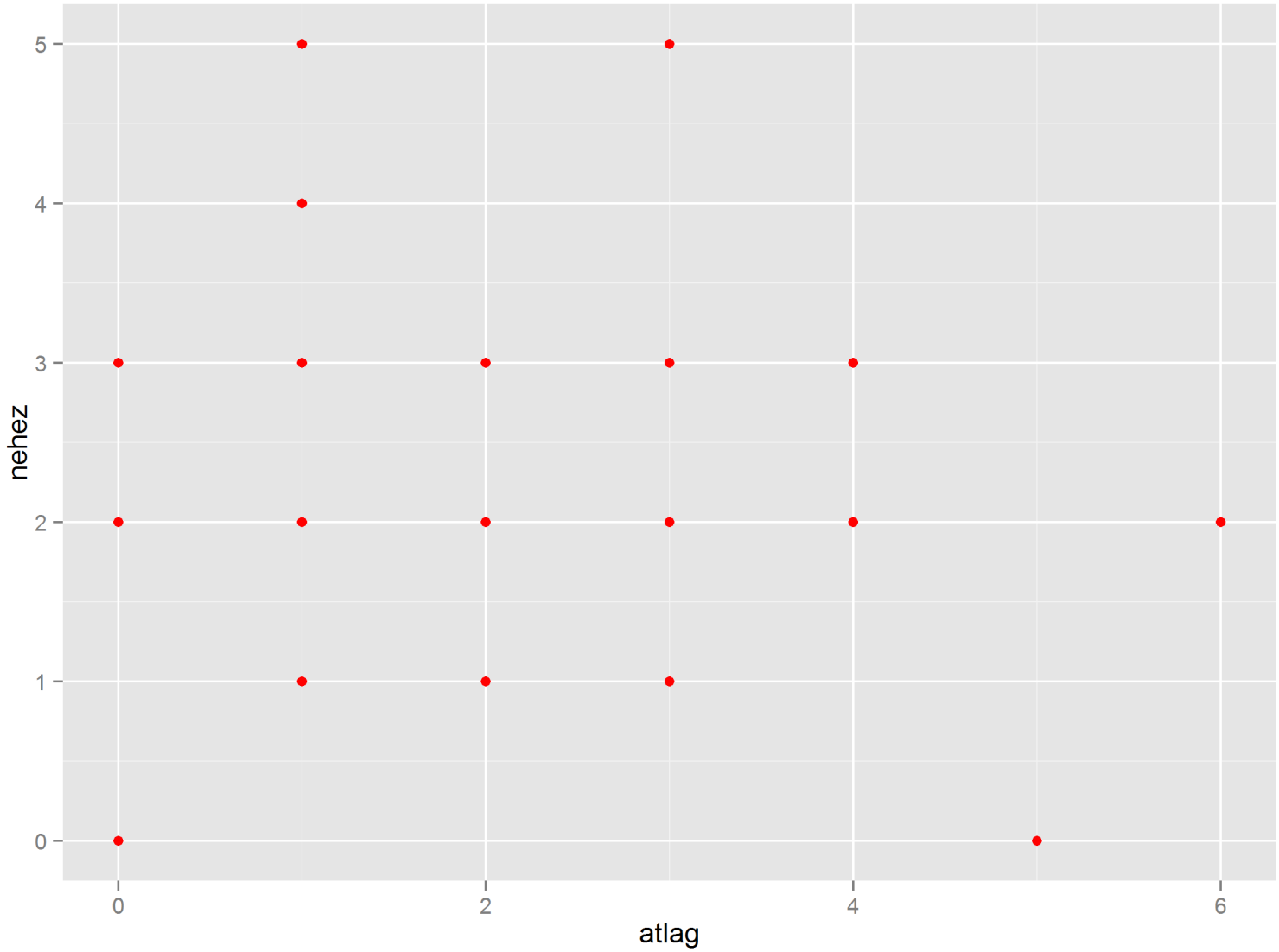


Pont – pont diagram (scatterplot)

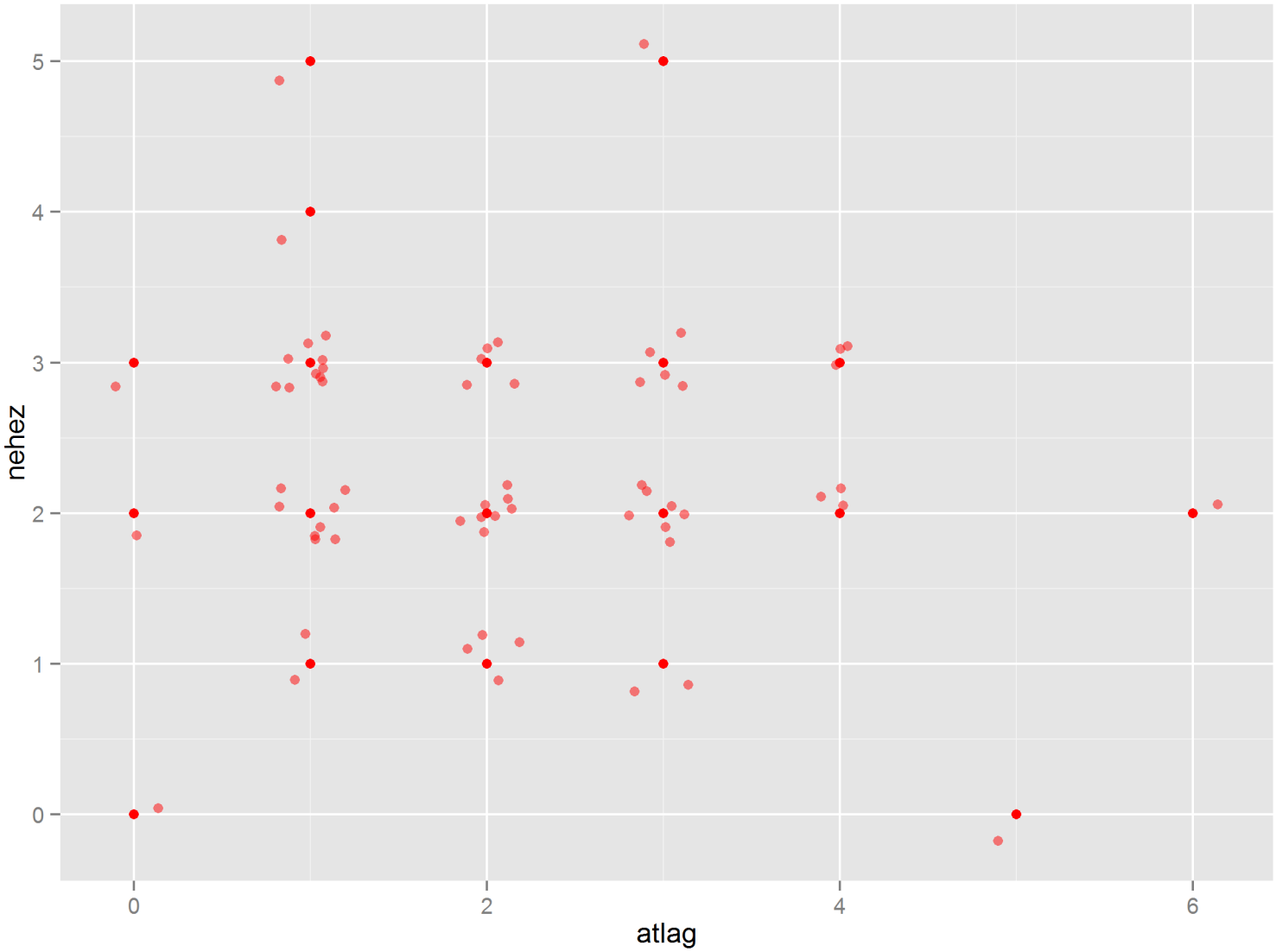
- Megjelenített dimenziók száma: 2
- Ábrázolt összefügg.:
  - Folytonos változók együttes eloszlása
- Adategység:
  - Pont – $X = x_i$, $Y = y_i$ előfordulás
- Korlát:
  - Ha az egyik változó értéke hiányzik → nem tudjuk felrajzolni
- Tervezői döntés
  - Overplotting?



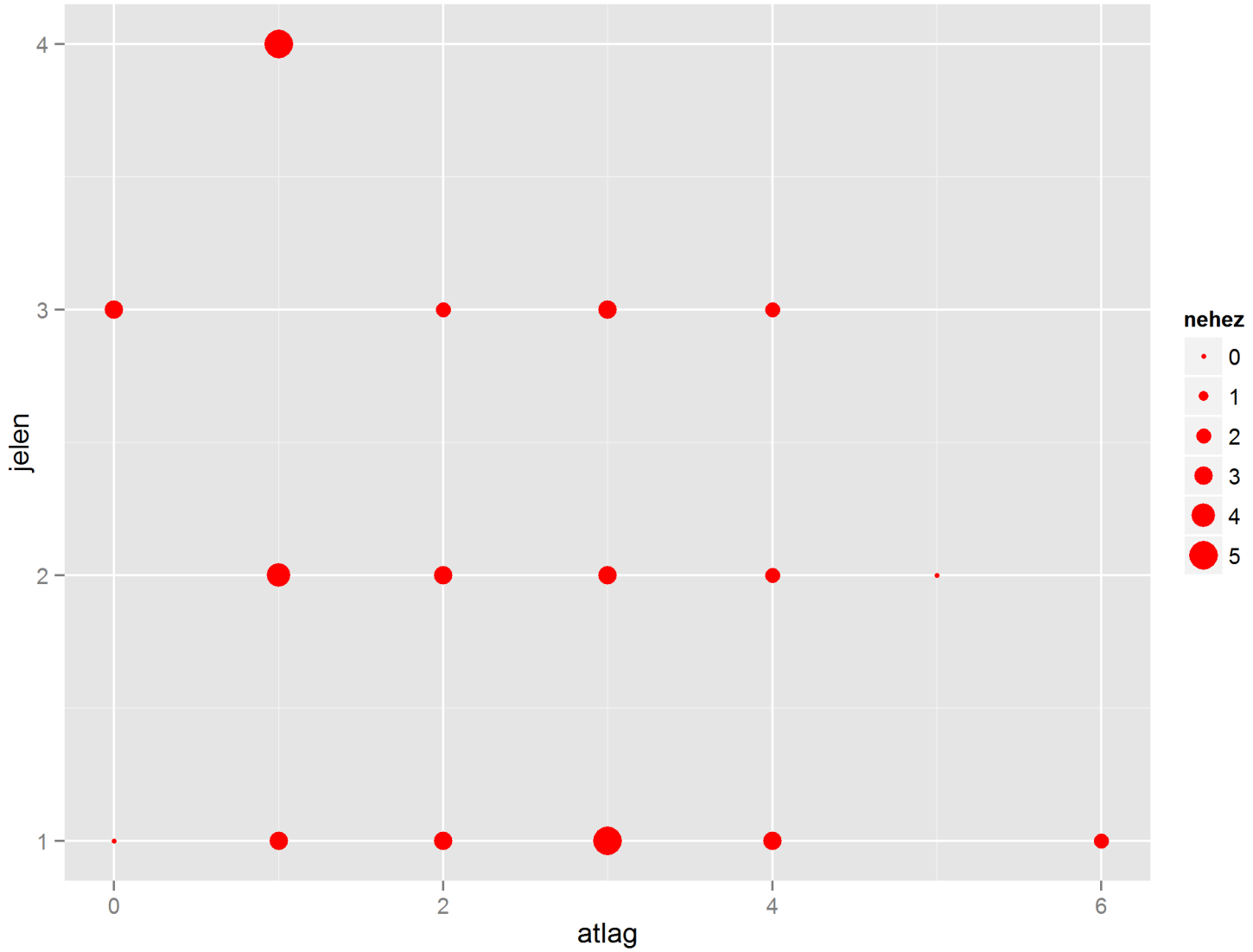
Hol volt, hol nem volt...



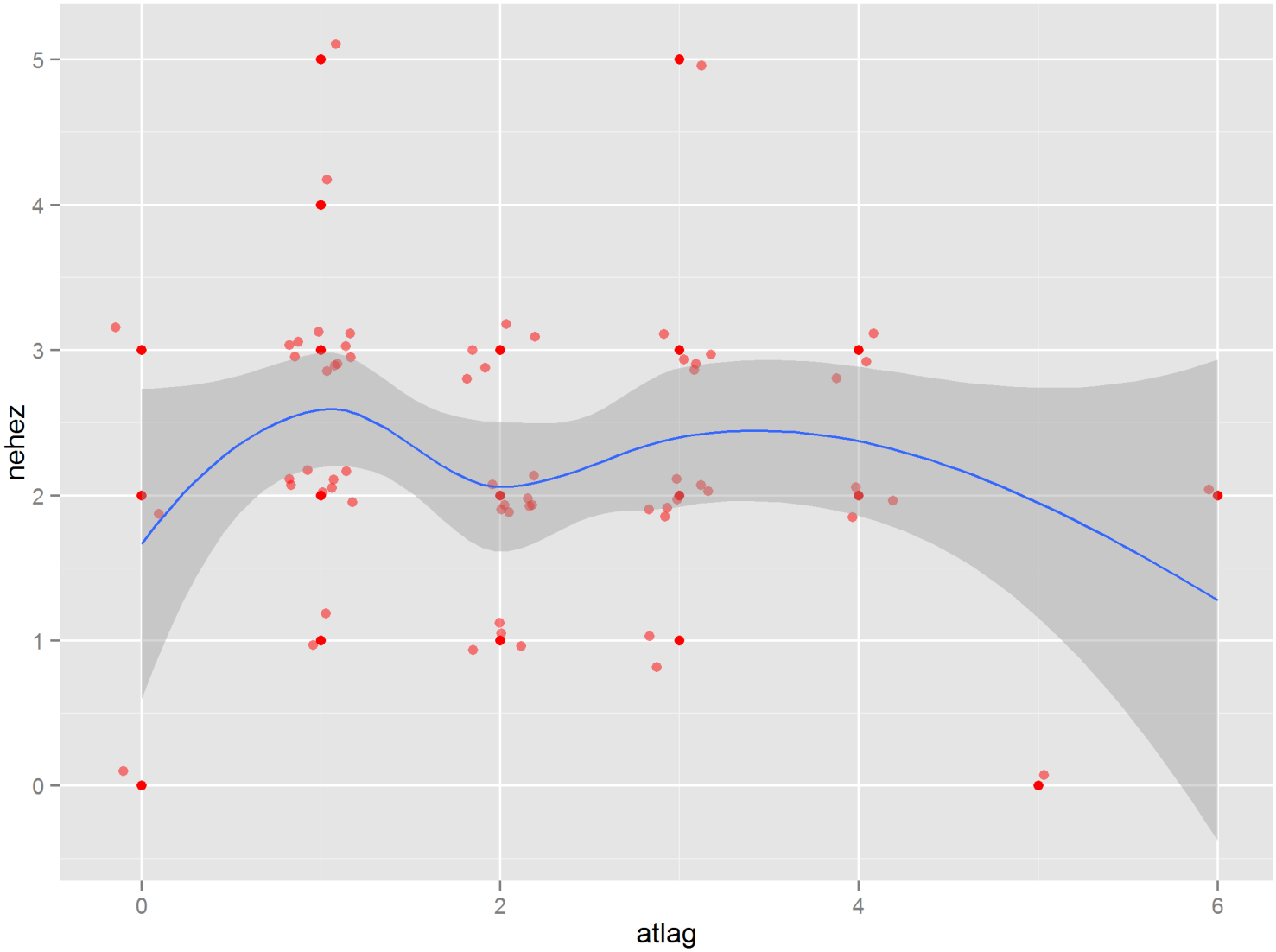
Szétszórjuk



A pontok....



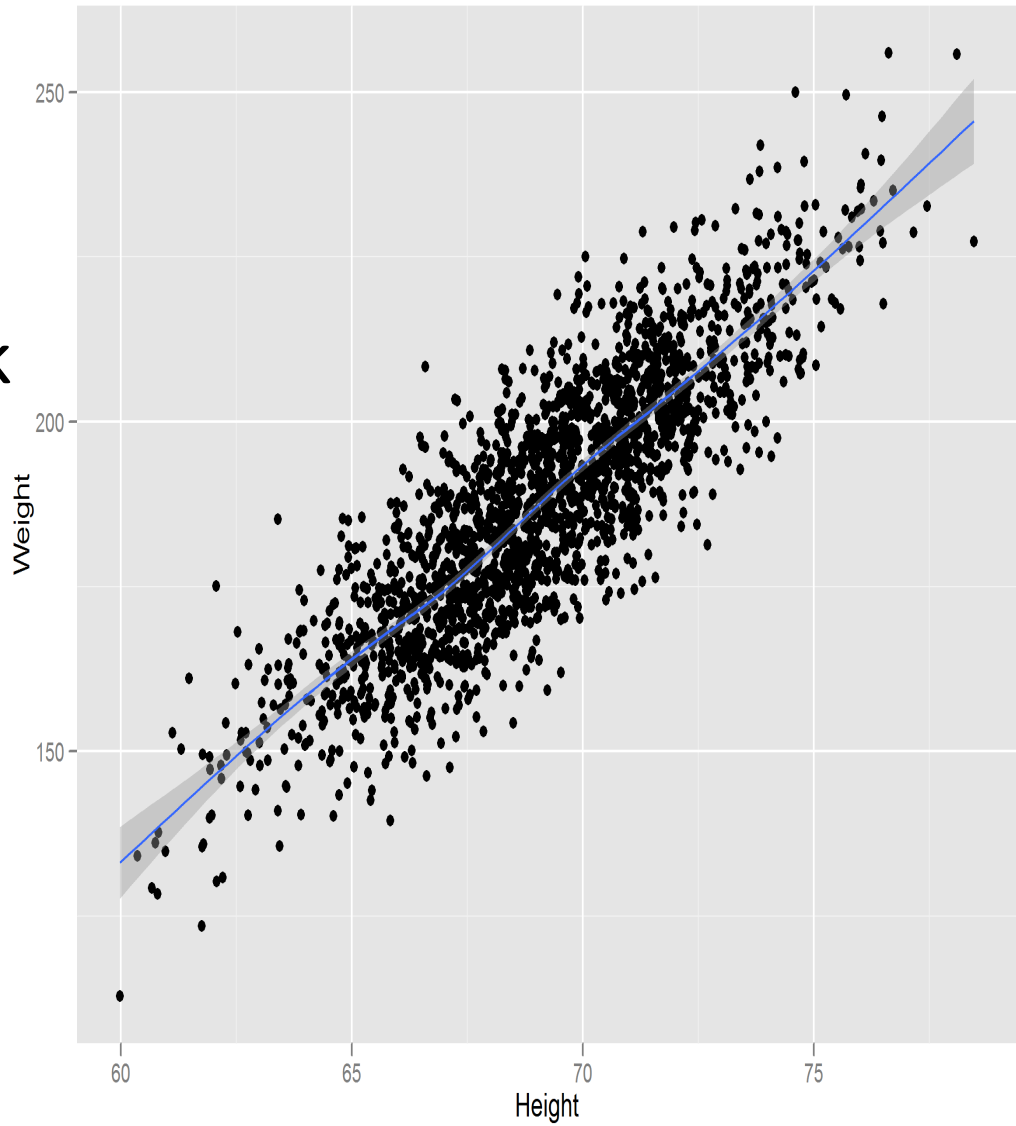
És megpróbáljuk közelíteni...



Regresszió

f függvény,

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



Regressziós módszerek

- Alapelv:

$$Y_t = f(\bullet) + \varepsilon_t$$

Véletlen változó

Közelítés

Hiba

Jósolt esemény

$$Y = f(X_1, X_2, \ldots, X_n)$$

Megfigyelhető változók

- Átlagos hiba (mean error)

Becsült érték

$$ME = \frac{\sum_{t=1}^{n} (Y_t - F_t)}{n}$$

Mért érték

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

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

- cél:

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

Levezetés (parc. deriválás)

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

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

$$\boxed{a=\overline{Y}-b\,\overline{X}}$$

Xi, Yi a mért értékpárok (pl. idő, terhelés)

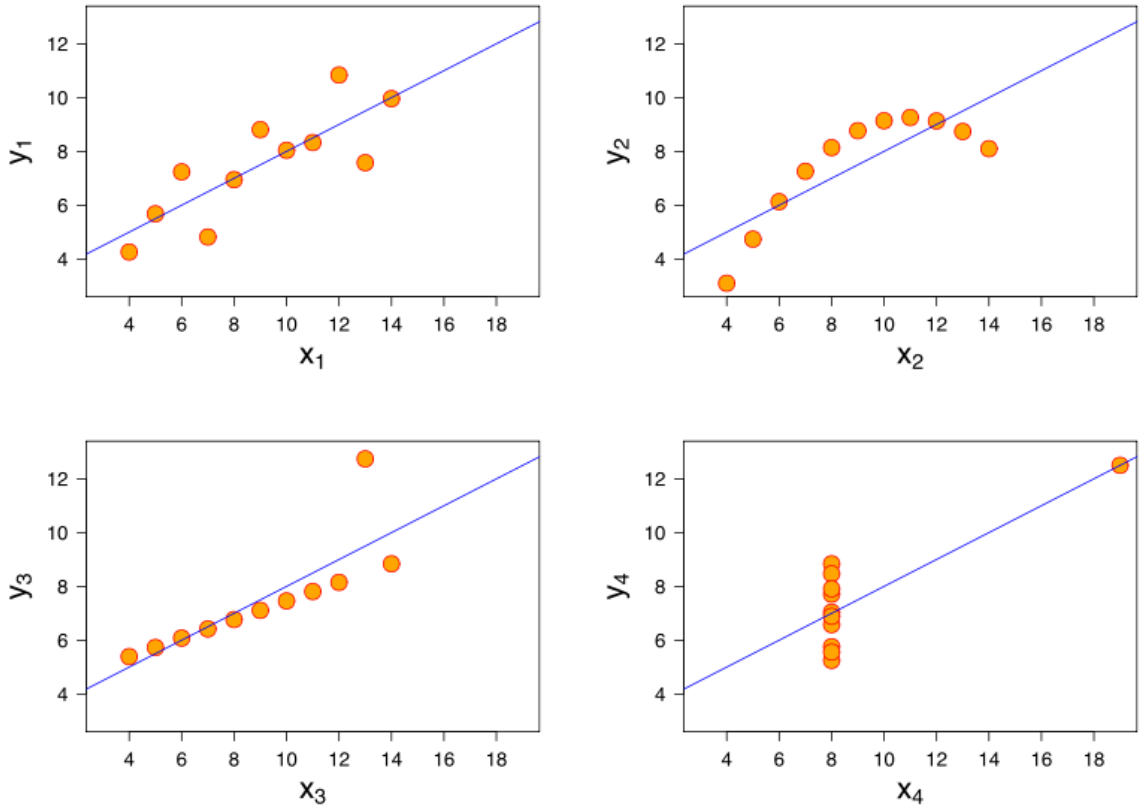
$$\frac{d\sum_{t=1}^{n}\left[Y_t-(a+bX_t)\right]^2}{db}=\sum_{t=1}^{n}X_t\left[Y_t-(a+bX_t)\right]=0$$

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

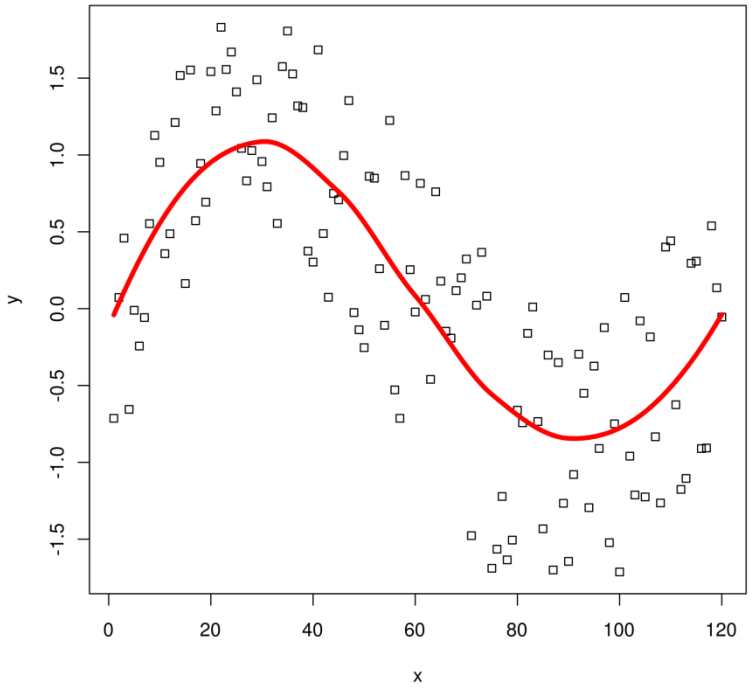
$$\boxed{b=\frac{n\sum_{t=1}^{n}X_tY_t-\left(\sum_{t=1}^{n}X_t\right)\left(\sum_{t=1}^{n}Y_t\right)}{n\sum_{t=1}^{n}X_t^2-\left(\sum_{t=1}^{n}X_t\right)^2}}$$

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

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

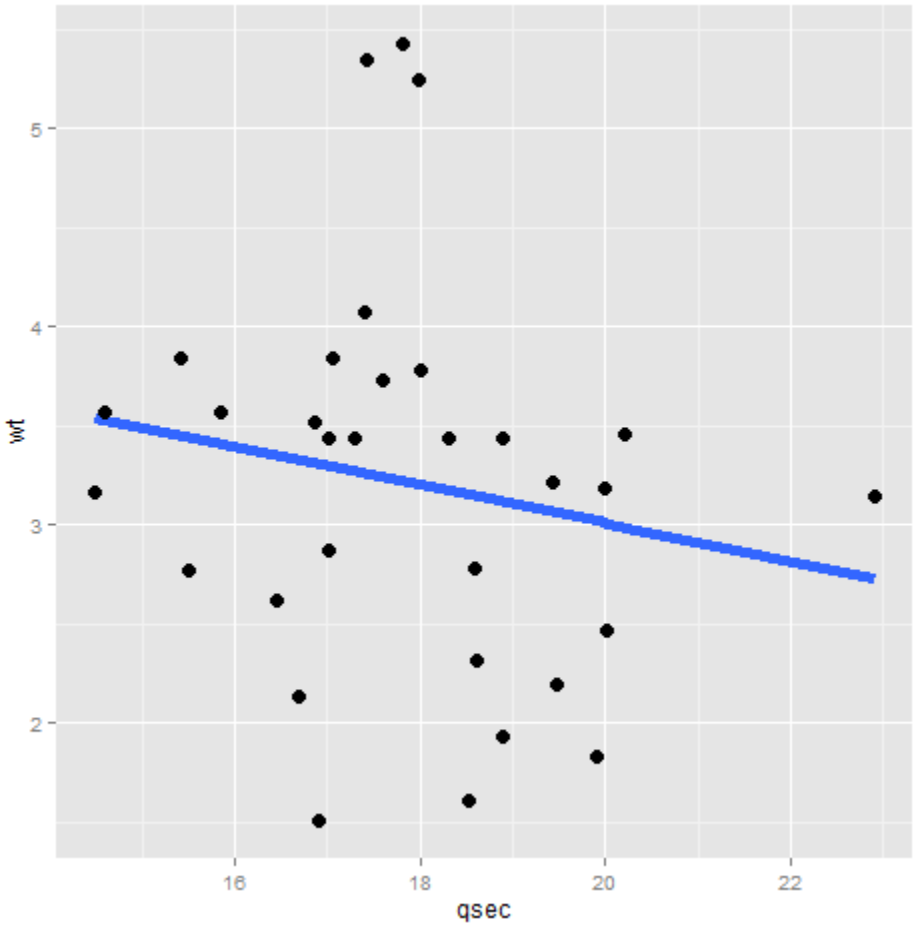
$$w(x_0) = T\left(\frac{|x - x_0|}{s}\right)$$

  - Nagy adatkészlet
  - Outlier érzékenység
  - Nem ad zárt alakot

Simító görbe

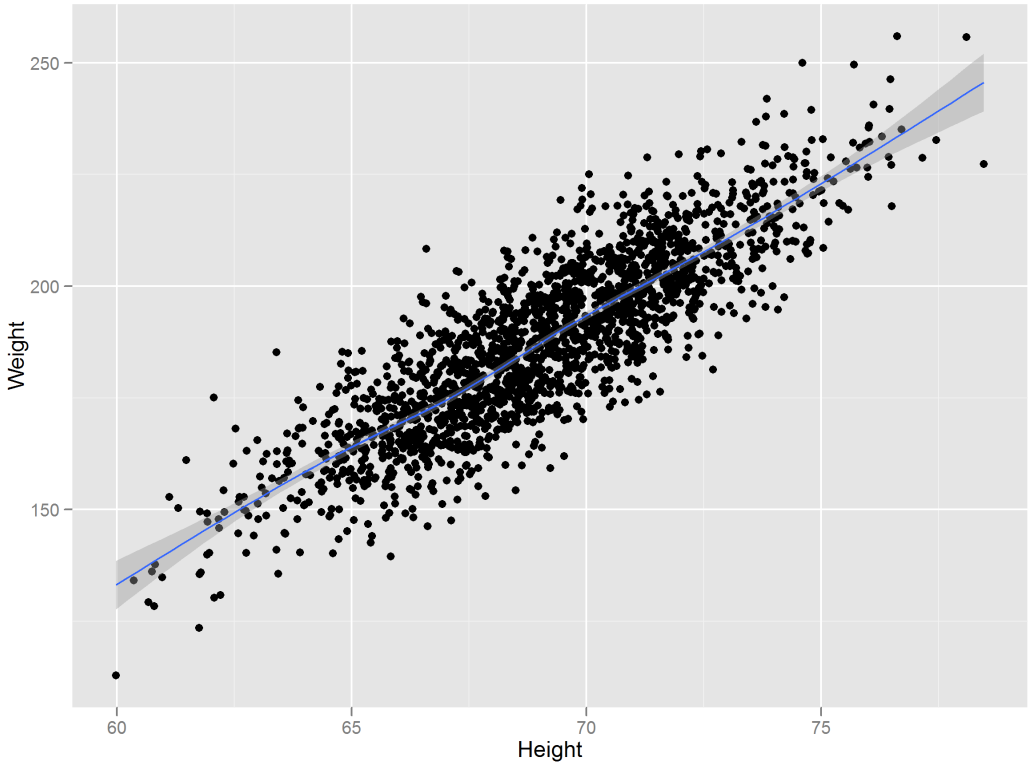
$$min\left[\left(\sum_{i=1}^{n}(Y_i - \hat{\mu}(x_i))^2\right) + \lambda \int_{x_1}^{x_n} \hat{\mu}''(x)^2 dx\right]$$

- 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 $\int_{x_1}^{x_n} \hat{\mu}''(x)^2 dx$
  - λ =0 esetén interpolációs görbe
  - $\lambda \rightarrow \infty$ esetében lineáris regresszió

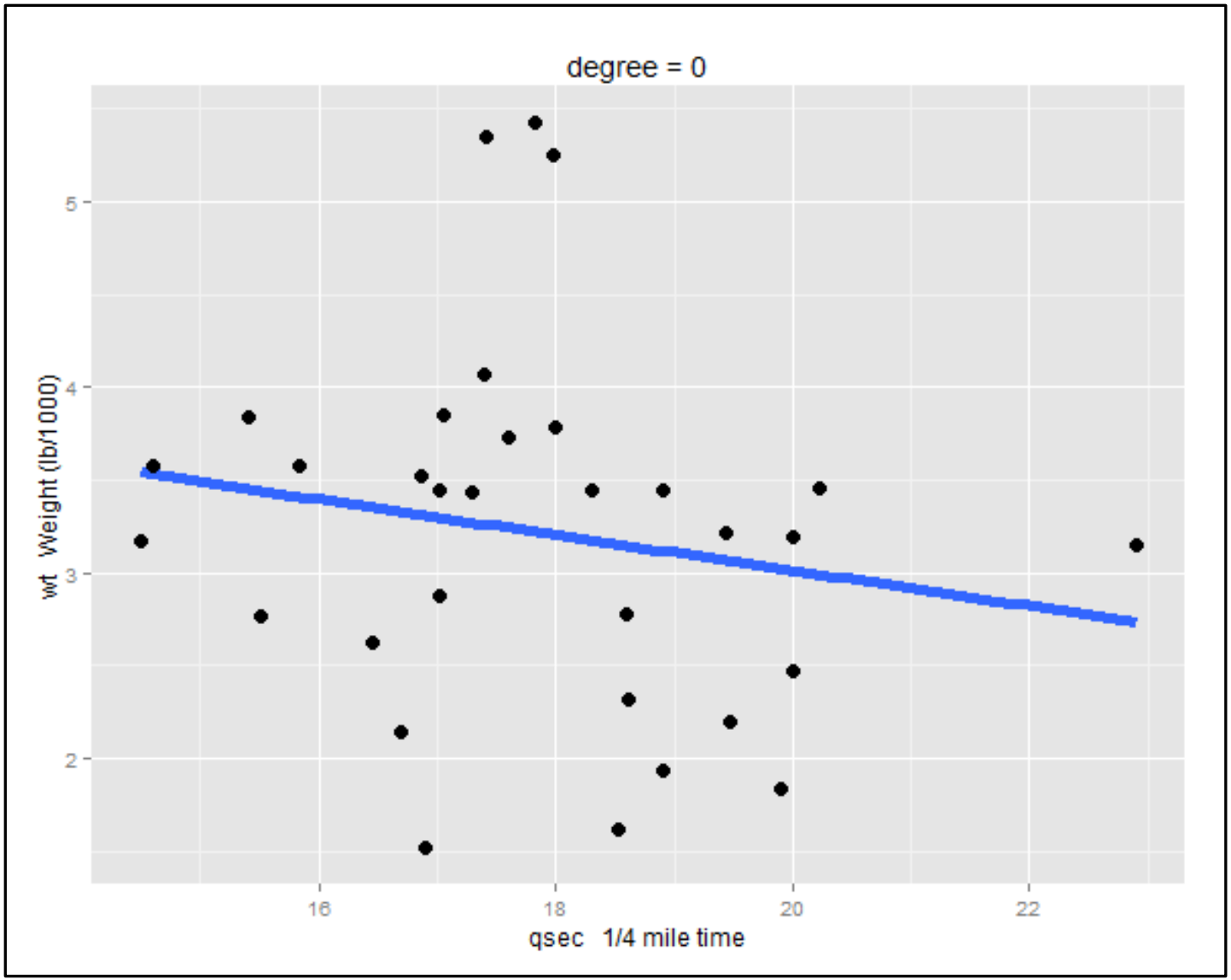

wt
5
4
3
2
16
18
20
22
qsec

Regresszió

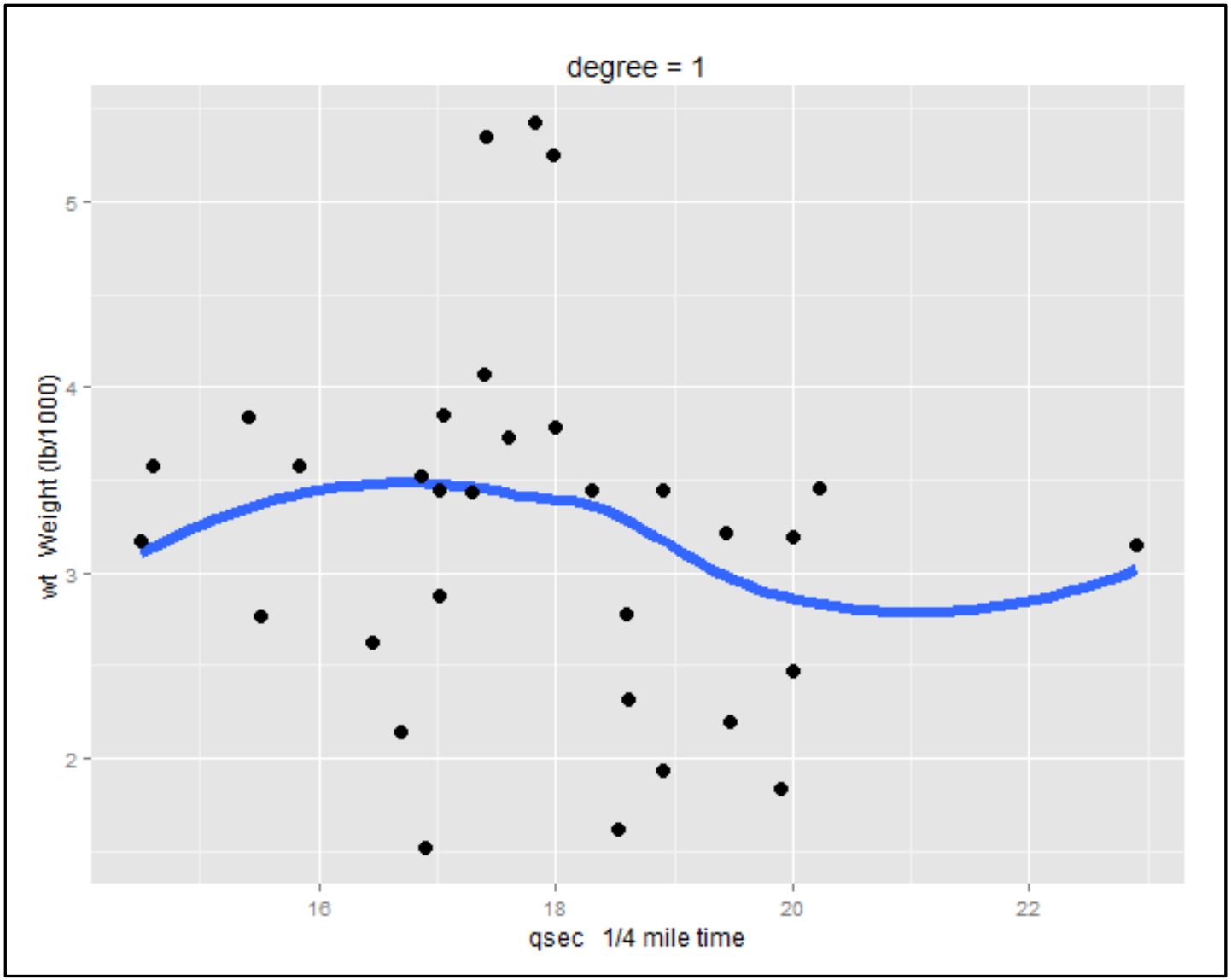
- Cél:
  - megtalálni egy olyan f függvényt, amelynek inputja az attribútumok értéke, az outputja pedig a lehető legjobban közelíti (négyzetes hibaérték) a valóságot
- Példa:
  - testtömeg/magasság együttes eloszlás valójában egyenesre illeszthető,
  - web forgalom jóslása

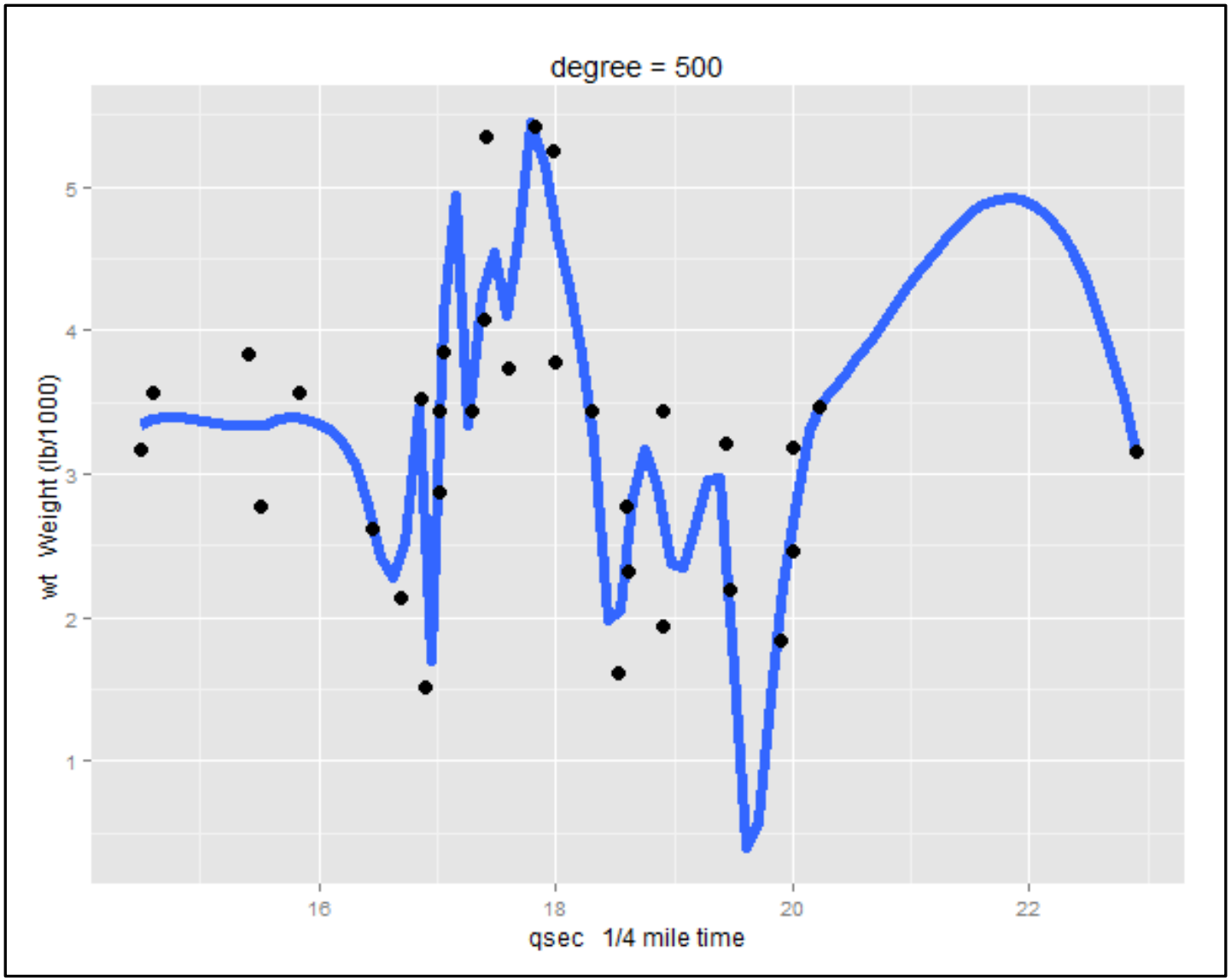


$\lambda = 0$



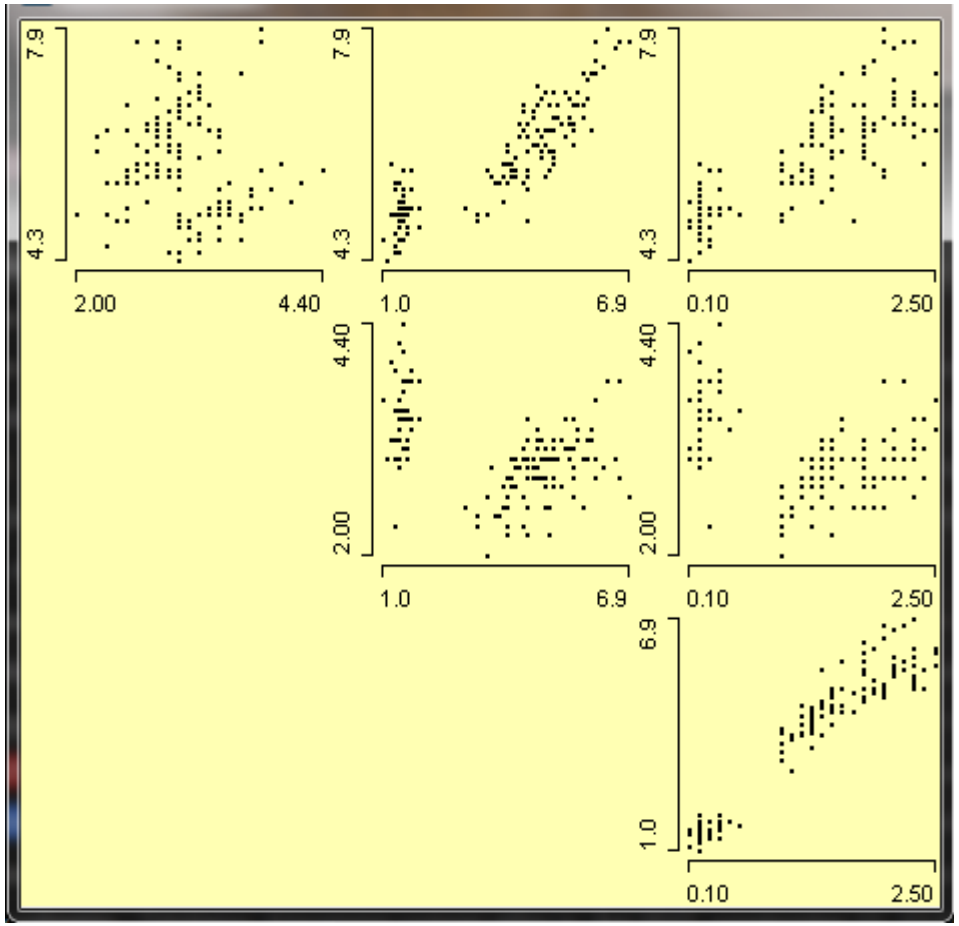
$\lambda = 1$




degree = 500
wt Weight (lb/1000)
5
4
3
2
1
16
18
20
22
qsec 1/4 mile time

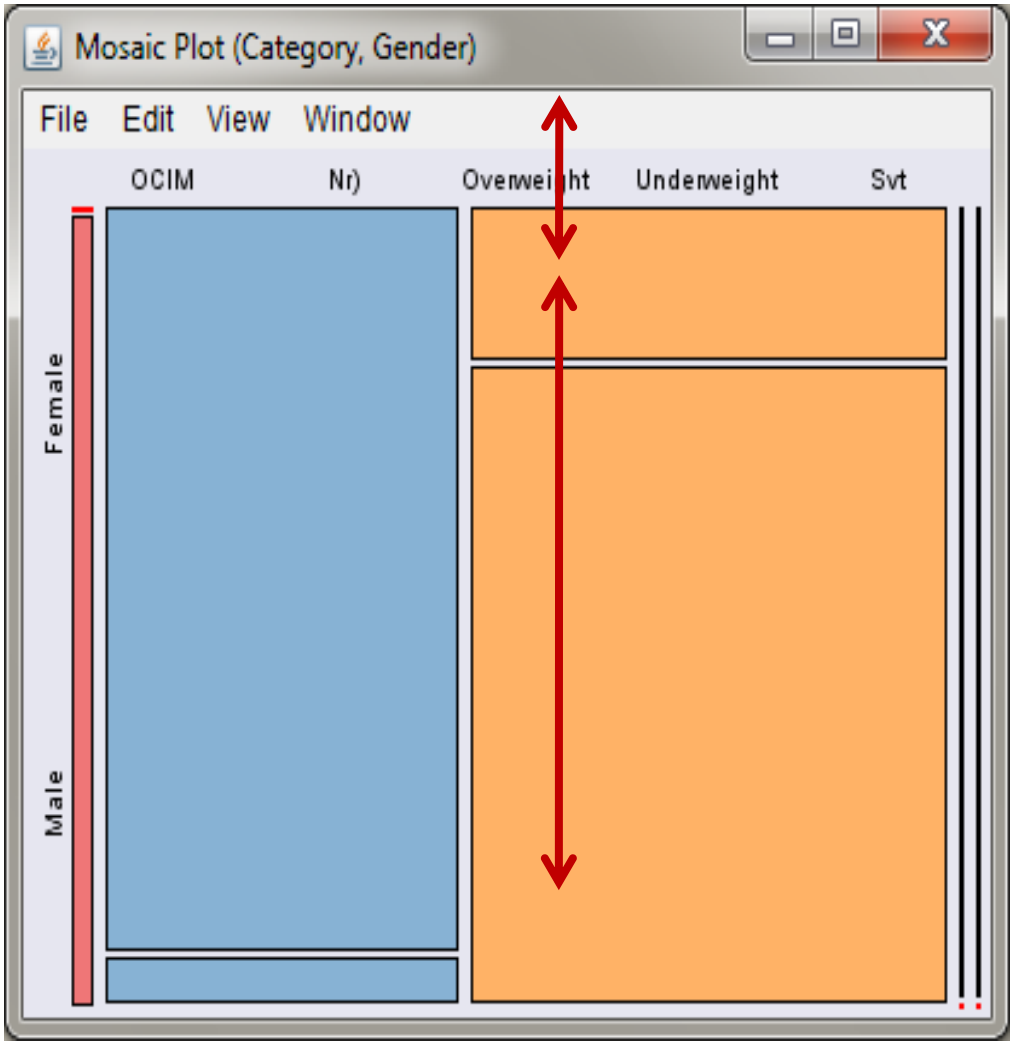
Scatterplot mátrix

- Megjelenített dimenziók száma: n
- Ábrázolt összefügg.:
  - A változó párok együttes eloszlása
- Adategység:
  - Scatterplot – minden diagram a neki megfelelő változók együttes eloszlását mutatja be



Mozaik diagram (mosaic plot)

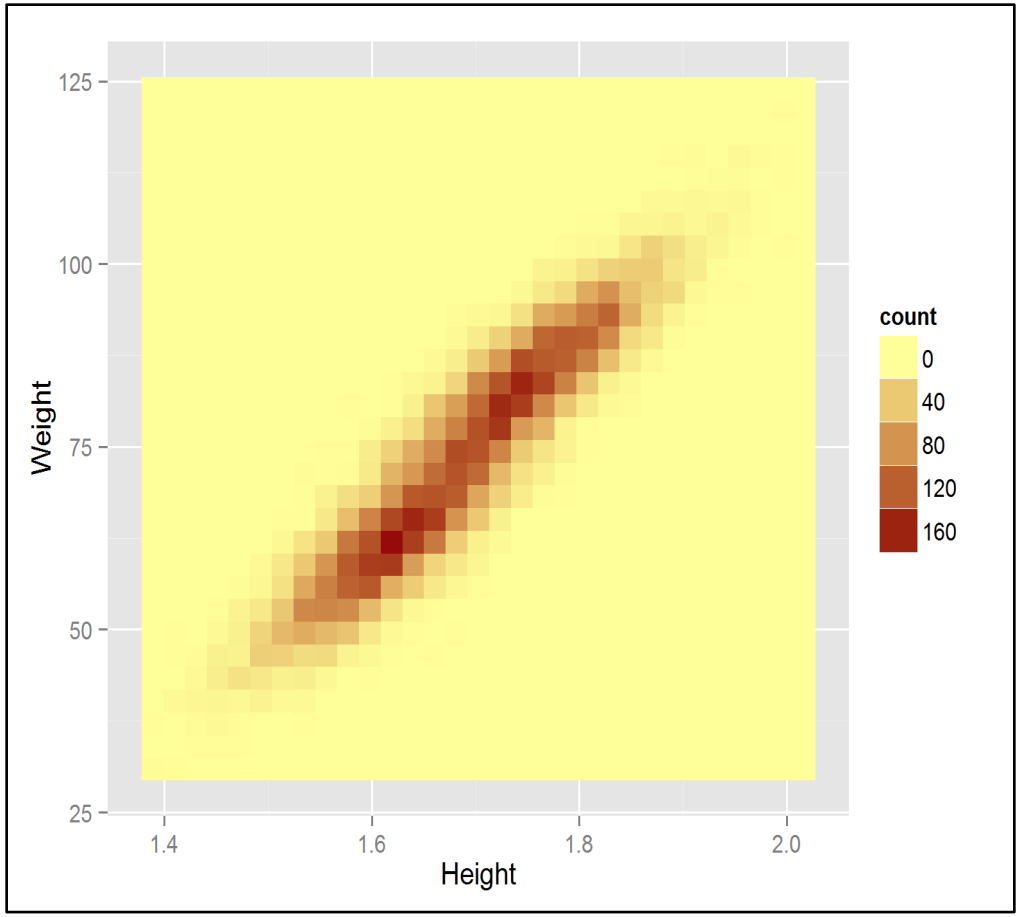
- Megjelenített dimenziók száma: 2
- Ábrázolt összefügg.:
  - Két diszkrét változó együttes eloszlása
- Adategység:
  - Téglalap – a téglalap *területe* arányos az $(X = x_i, Y = y_i)$ értékpárok gyakoriságával
- Korlát:
  - Sorfolytonos olvasása nehézkes



A túlsúlyosak nagy része férfi!

Hőtérkép (heat map)

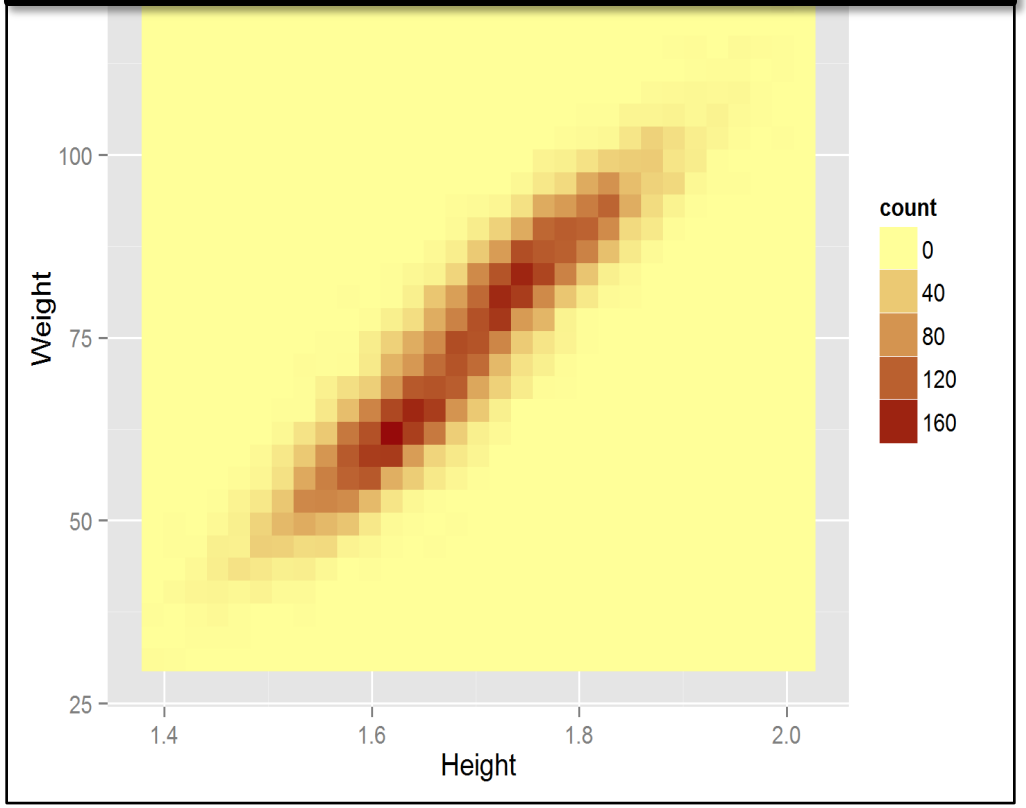
- Megjelenített dimenziók száma: 3
- Ábrázolt összefügg.:
  - *Sűrű* 3D struktúrák összefüggései
- Adategység:
  - Tile – azonos „magasságú” összefüggő területrész
- Tervezői döntés:
  - Tile-ok mérete?



Hőtérkép (heat map)

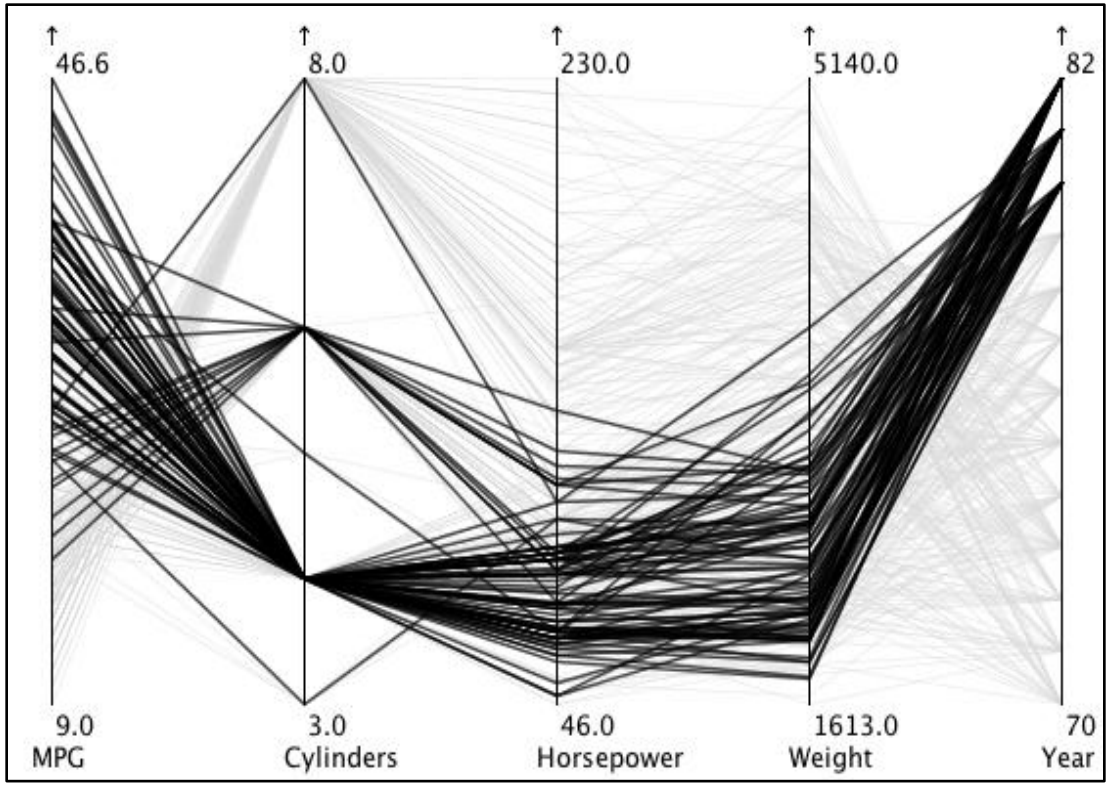
- Megjelenített dimenzi
- Ábrázolt összefügg.:
  - *Sűrű* 3D struktúrák összefüggései
- Adategység:
  - Tile – azonos „magasságú” összefüggő területrész
- Tervezői döntés:
  - Tile-ok mérete?

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



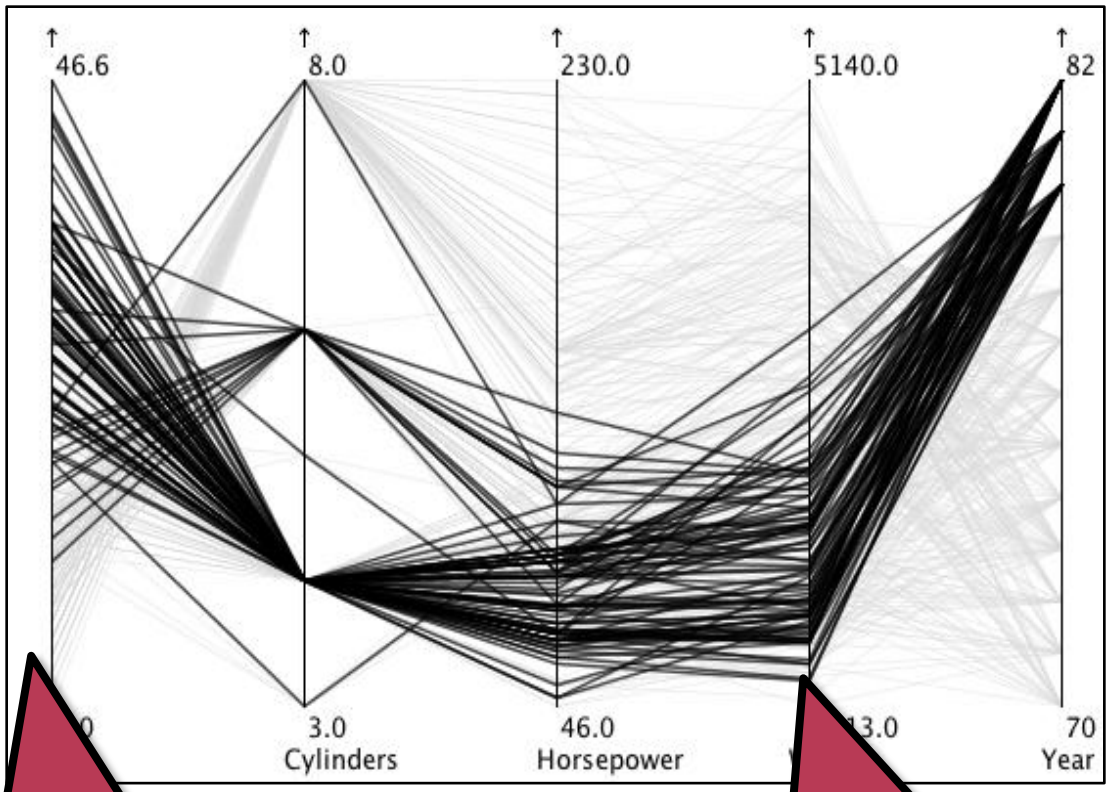
Párhuzamos koordináták

- Megjelenített dimenziók száma: n
- Ábrázolt összefügg.:
  - Rekordok/attrib.-ok hasonlósága
- Adategység:
  - Törött vonal – az egyenes attrib.-tengelyeken felvett értékek rendezett sorozata
- Korlátok:
  - Tengelyek (attrib.-ok) más mértékegysége/nagyságrendje stb. torzíthat



Párhuzamos koordináták

- Megjelenített dimenziók száma: n
- Ábrázolt összefügg.:
  - Rekordok/attrib.-ok hasonlósága
- Adategység:
  - Törött vonal – az egyenes attrib.-tengelyeken felvett értékek rendezett sorozata
- Korlátok:
  - Tengelyek (attrib.-ok) más m

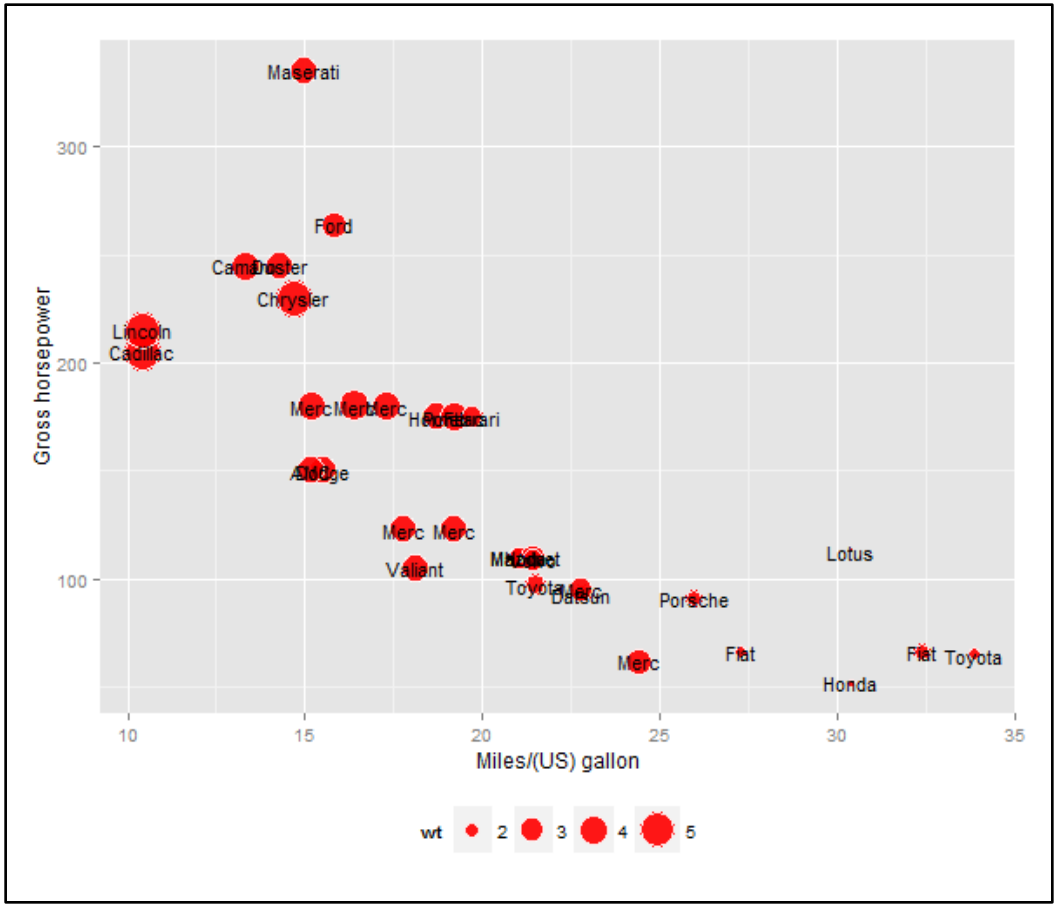


… de a fogyasztás nagyobb

Az új autókban a tömeg kisebb…

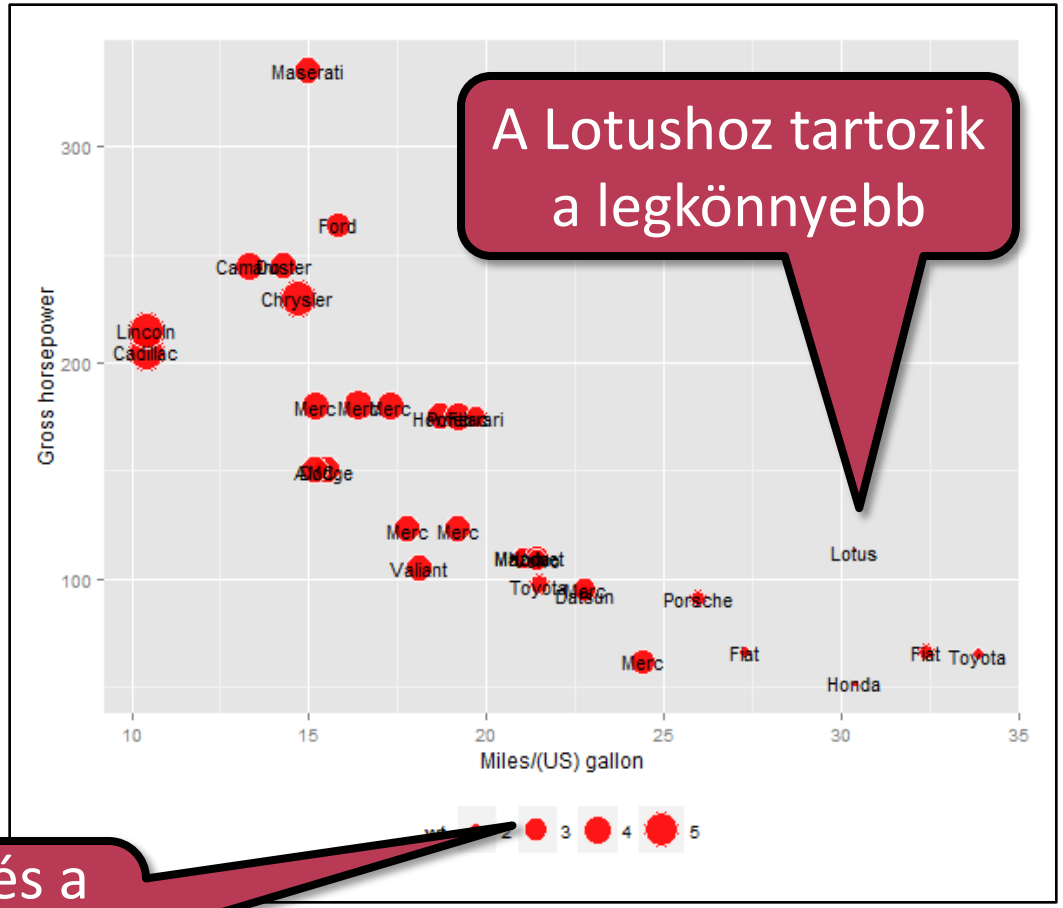
Buborék diagram (bubble chart)

- Megjelenített dimenziók száma: 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)



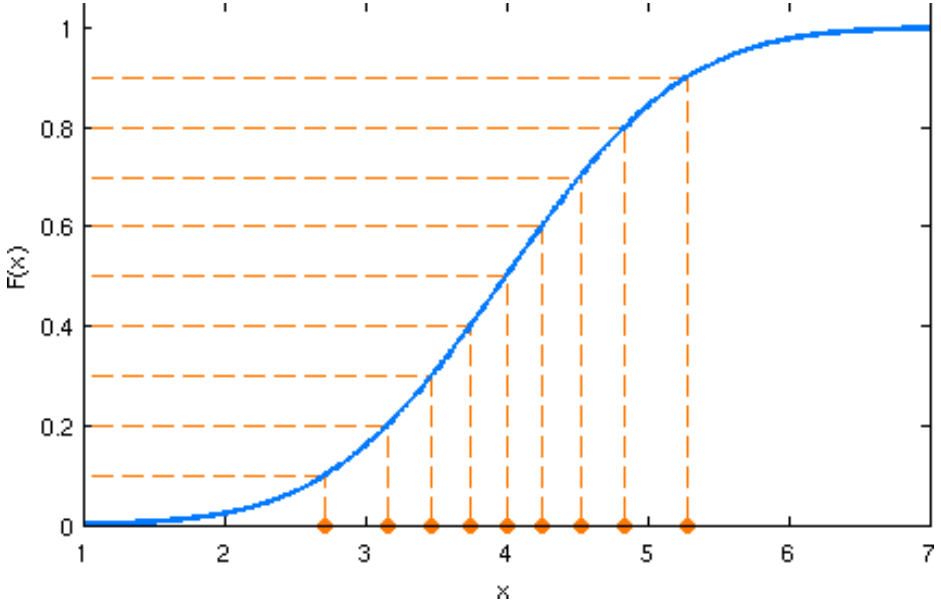
Buborék diagram (bubble chart)

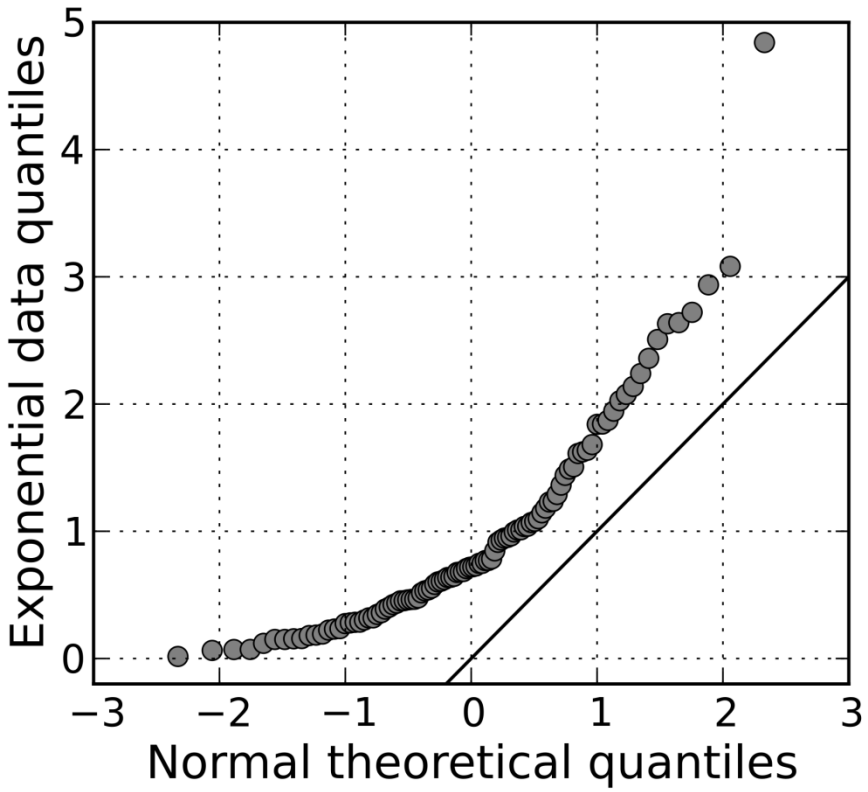
- Megjelenített dimenziók száma: 3
- Ábrázolt összefügg.:
  - *Ritka* 3D struktúrák összefüggései
- Adategység:
  - Körlap – 3 attribútummal leírható: X és Y koordináta a középpontra + sugár
- Korlátok:
  - Overplotting torzíthat



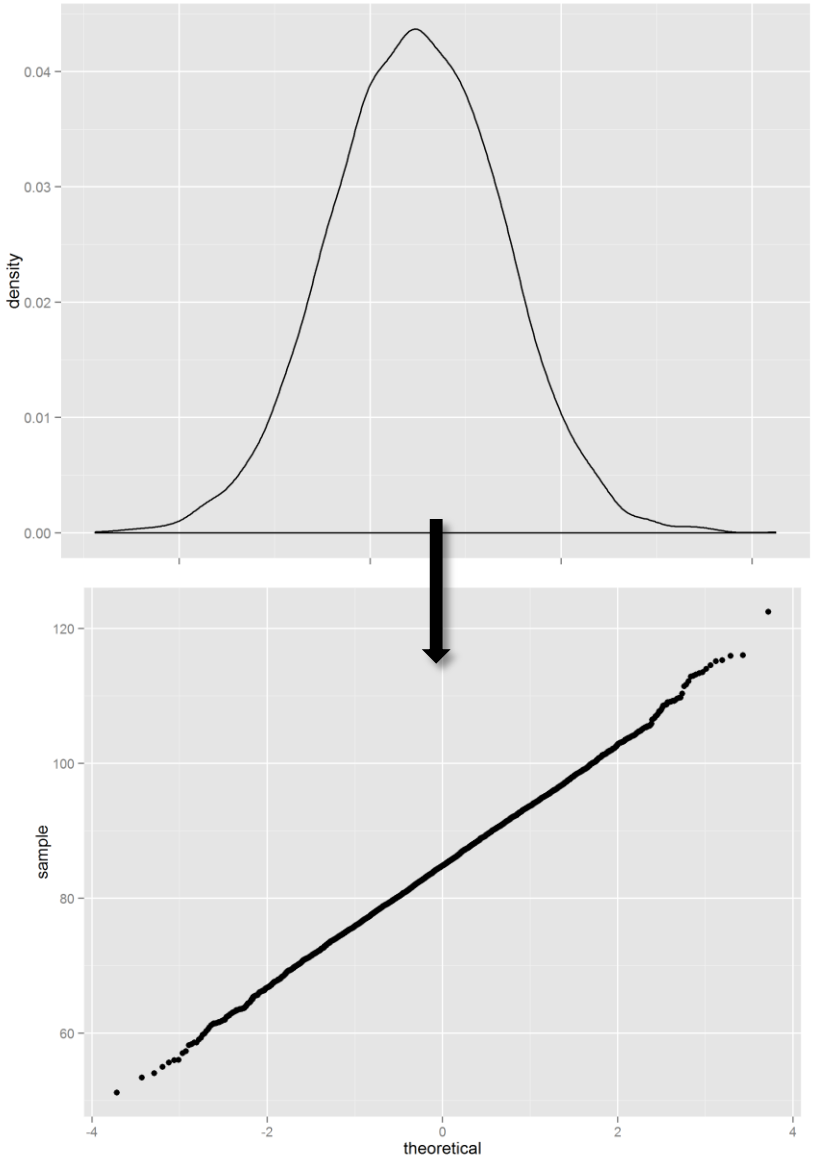
Az X, Y pozíciót a fogyasztás és a teljesítmény adja, a kör sugara a tömeget mutatja

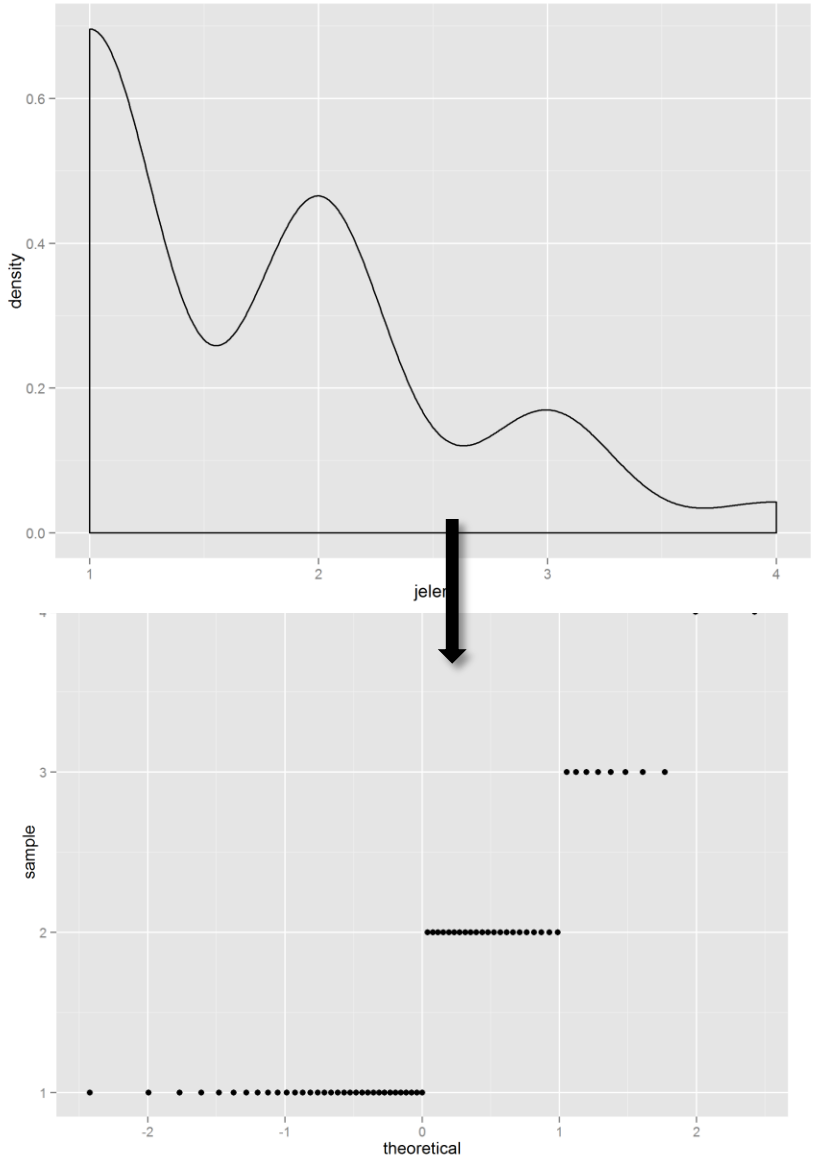
qqplot


F(x)
x
0
0.2
0.4
0.6
0.8
1
1
2
3
4
5
6
7


Exponential data quantiles
Normal theoretical quantiles
0
1
2
3
4
5
−3
−2
−1
0
1
2
3

qqplot


density
0.04
0.03
0.02
0.01
0.00
sample
120
100
80
60
-4
-2
0
2
4
theoretical

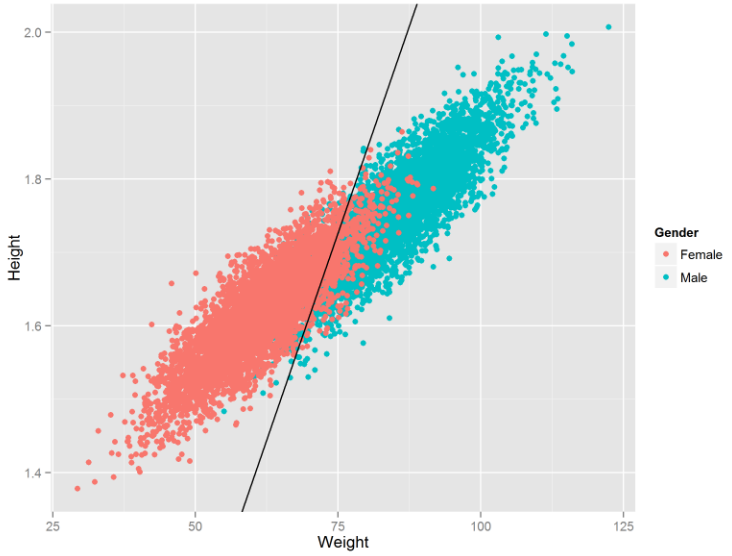

density
0.6
0.4
0.2
0.0
1
2
3
4
sample
3
2
1
-2
-1
0
1
2
theoretical

Data transformation: Box-Cox

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

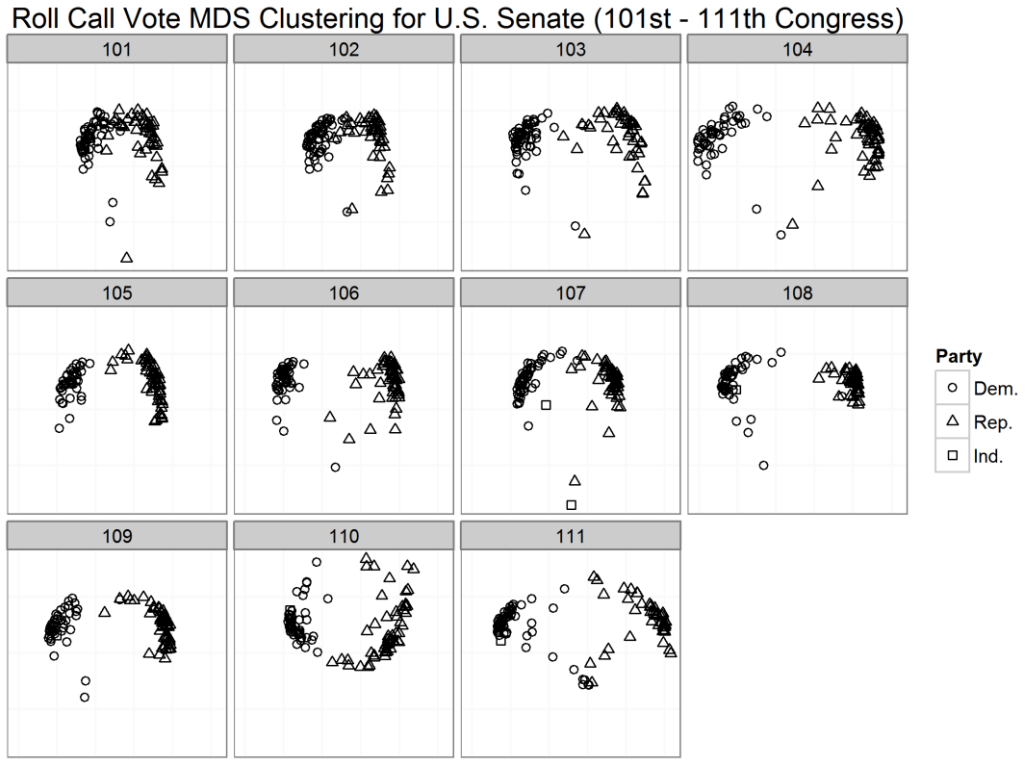
Osztályozás

- Cél:
  - Adott teszthalmaz meglétét feltételezve egy újonnan érkező elemről eldönteni, melyik osztályba tartozik (az osztályok száma és a tesztrekordok osztályba tartozása rögzített)
- Példa:
  - SPAM detektálás tartalom alapján,
  - testtömeg/magasság alapján alany nemének meghatározása



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?