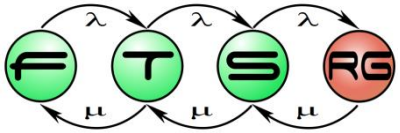
Nagy méretű adathalmazok vizualizációja

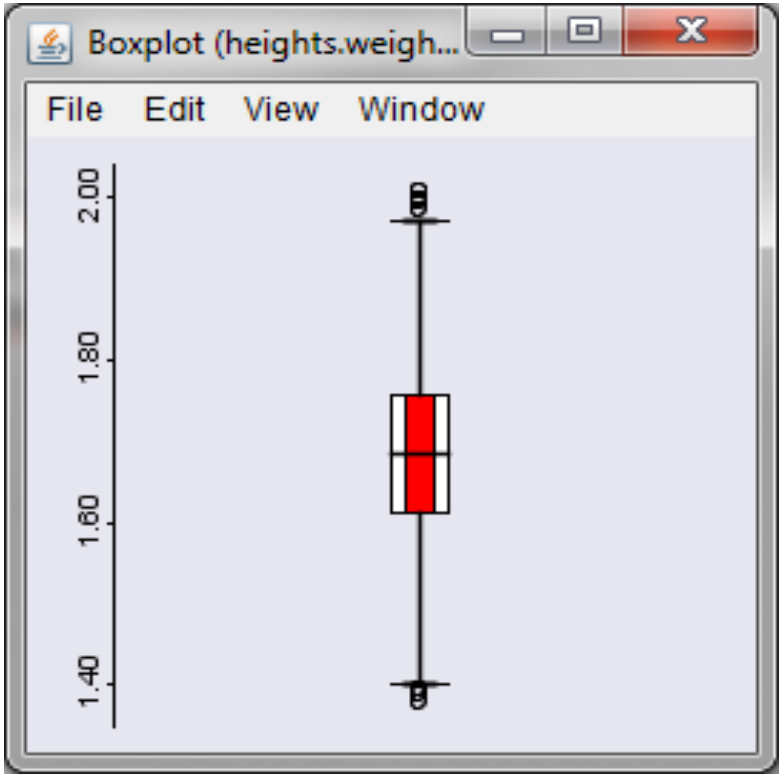
„Big Data” elemzési módszerek

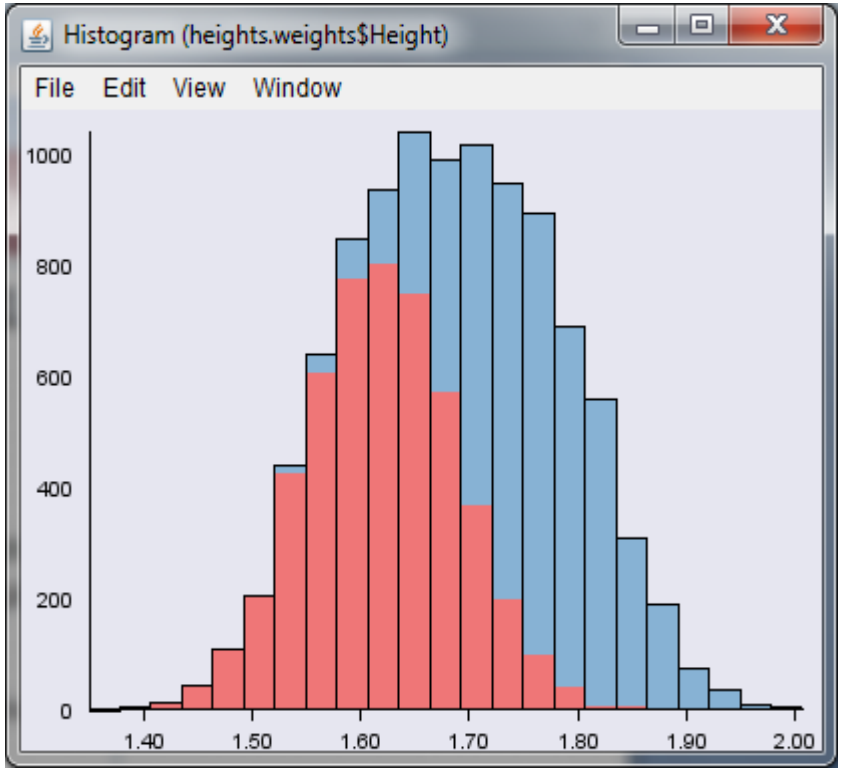
Kocsis Imre, Salánki Ágnes
ikocsis, salanki@mit.bme.hu
2014.10.15.



Aggregálunk

- n nagy → mesterségesen tömörítünk





Bin-summarize-smooth-visualize

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

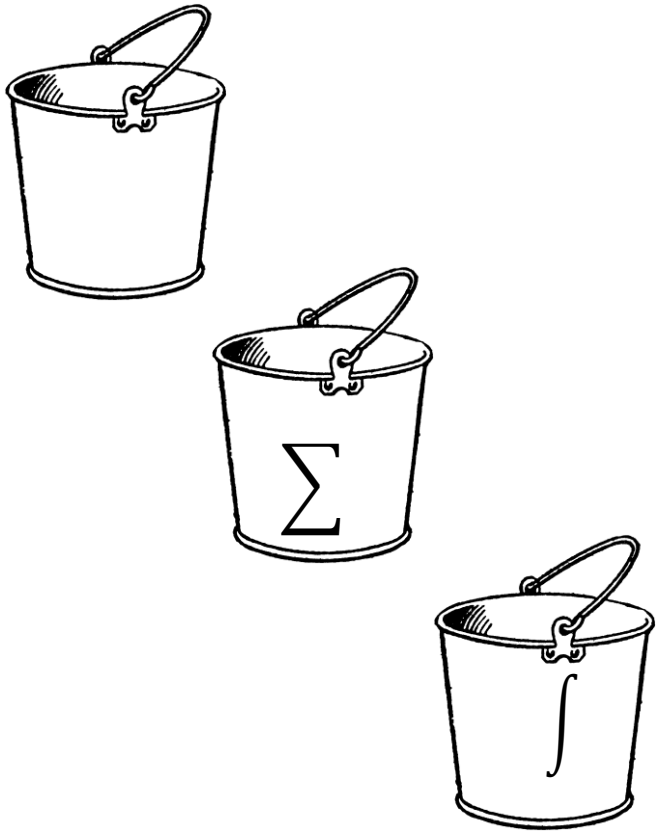
Bin-summarize-smooth-visualize

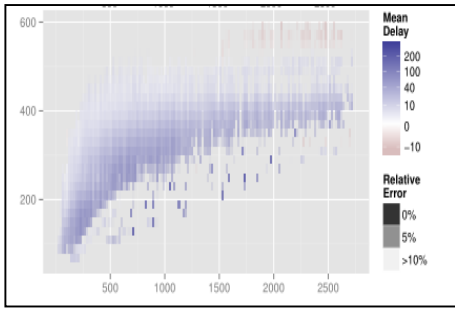
Bin

Summarize

Smooth

Visualize





„Condense”

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

Példa adatsor: flight data

- ASA Data Expo ’09
- http://stat-computing.org/dataexpo/2009/the-data.html

- Változók
  - Year, Month, DayOfMonth, DayOfWeek
  - DepTime, SchDepTime, ArrTime, SchArrTime
  - ArrDelay, DepDelay
  - Origin, Dest
  - Distance

„Bin”

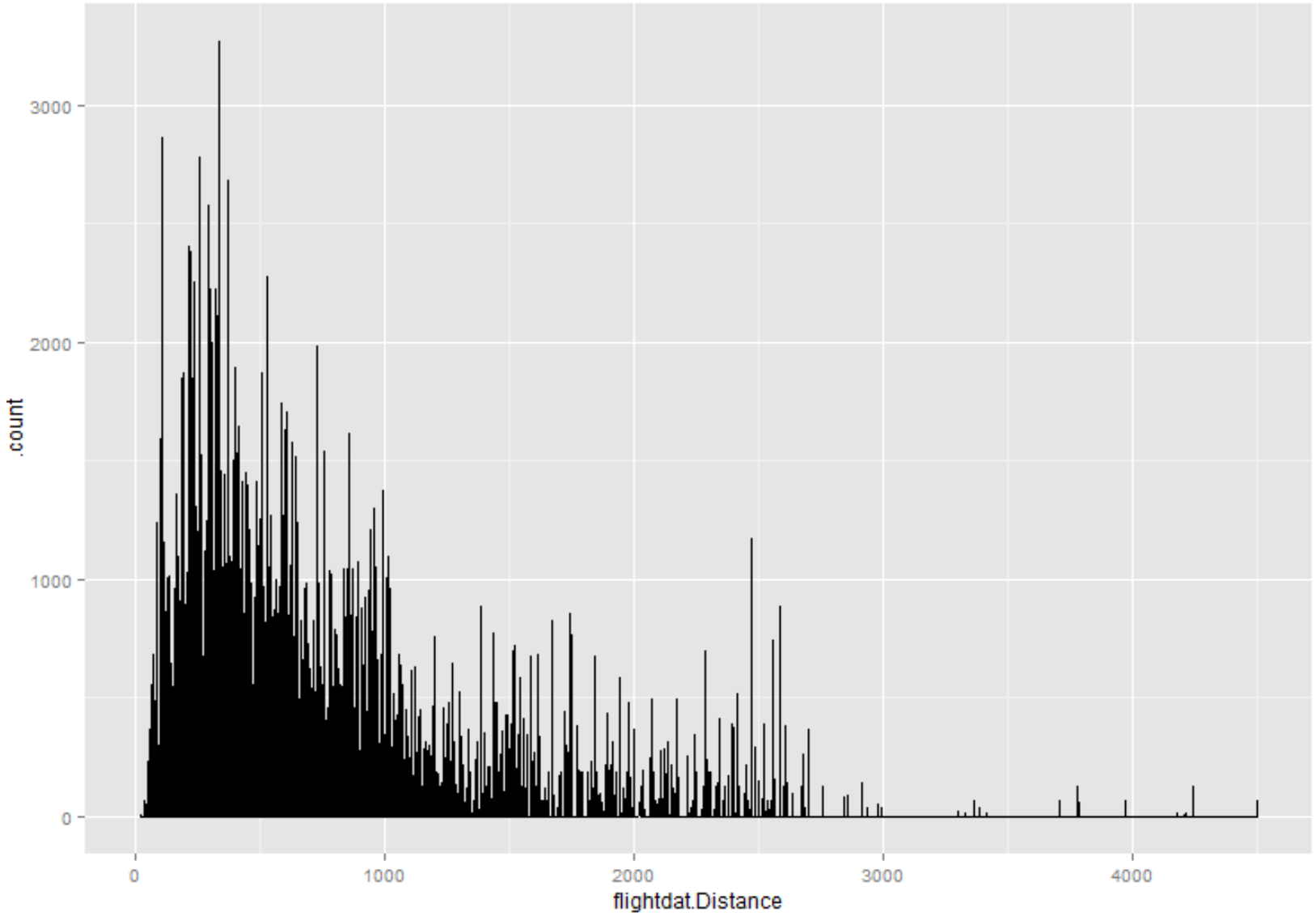
- Fix szélességű dobozok
- Egy dimenzióban: $\left\lfloor \frac{x - origin}{width} \right\rfloor + 1$
- Általánosítás több dimenzióban

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

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

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

„Bin”


3000
2000
1000
0
.count
0
1000
2000
3000
4000
flightdat.Distance

„Summarise”

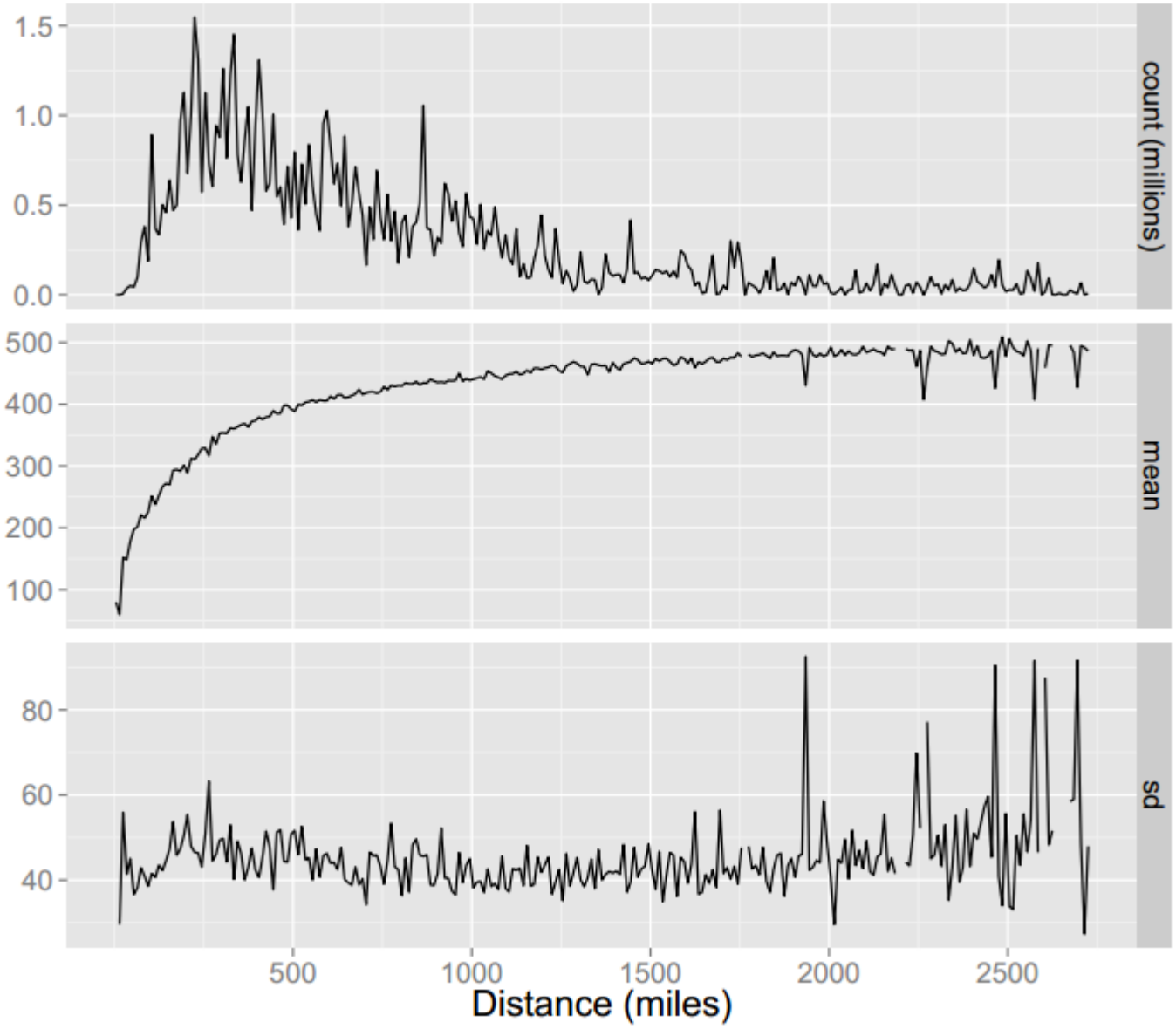
- Összefoglaló statisztikák típusai:
  - Disztributív
    - egyetlen, adott méretű köztestár
    - eredmények kombinálhatóak
    - pl. count, sum
  - Algebrai
    - disztributív statisztikák fix száma kell hozzá
    - Pl. átlag: count + sum
  - Holisztikus
    - bemenettel növekvő köztestár kell
   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

„Summarise”

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

1. Általában jól párhuzamosítható
2. Interaktív vizualizáció

„Summarize”

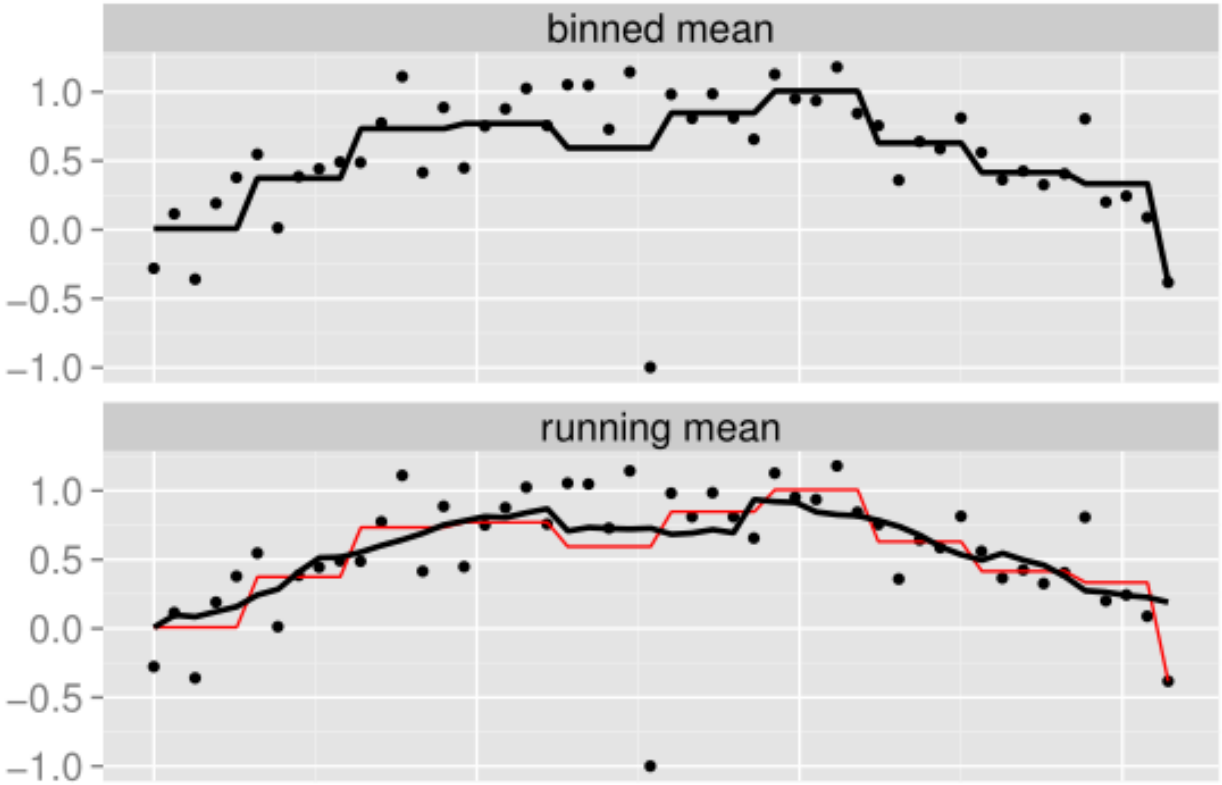

1.5
1.0
0.5
0.0
count (millions)
500
400
300
200
100
mean
80
60
40
sd
500
1000
1500
2000
2500
Distance (miles)

„Smooth”

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

„Smooth”

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



„Smooth”

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

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

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

- h: „sávszélesség”
  - Szomszédság mérete

- K itt: „triweight”

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

„Smoot

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

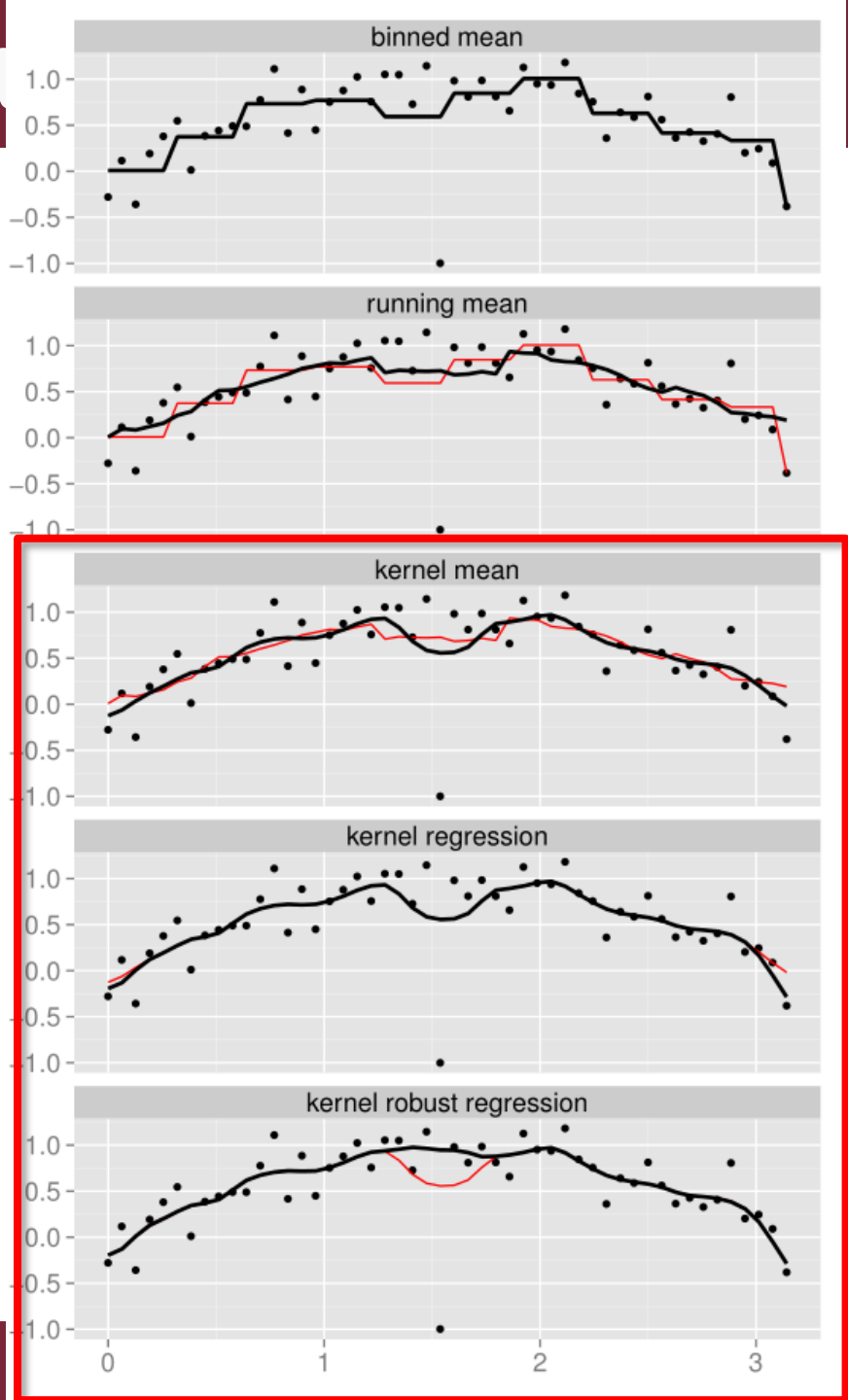
- j-edik bin közelítésénél az i-edik súlya:

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

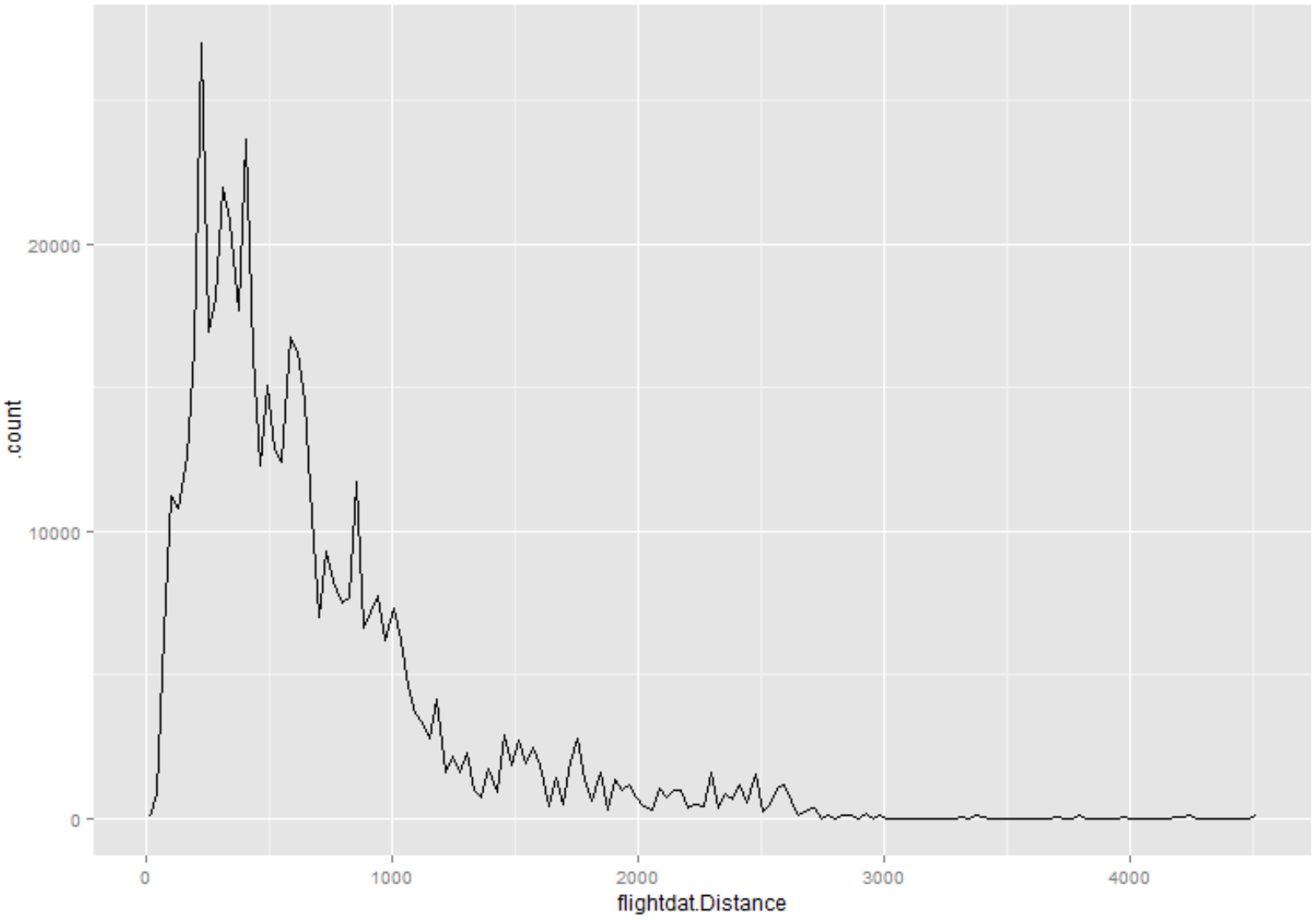
- h: „sávszélesség”
  - Szomszédság mérete

- K itt: „triweight”

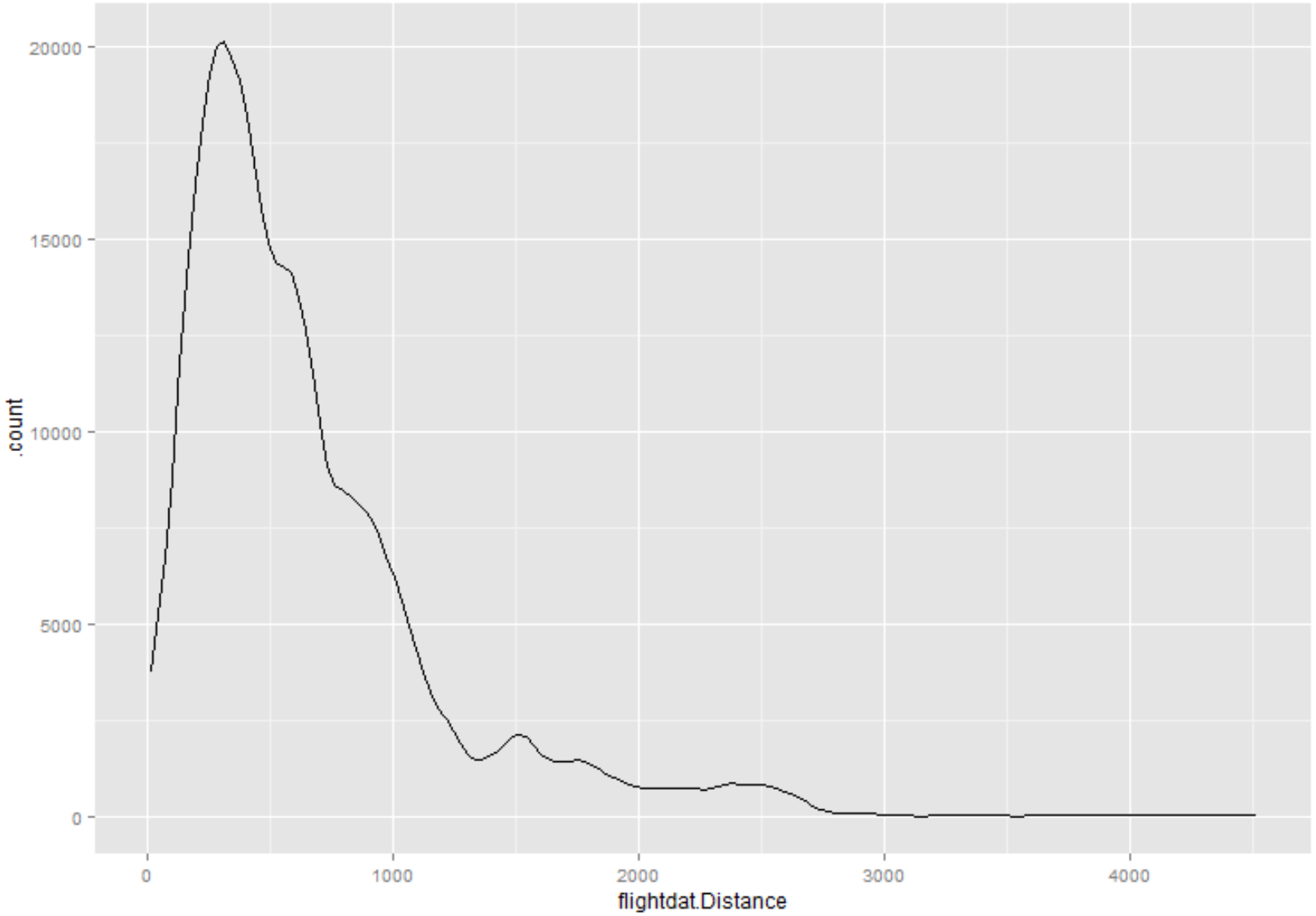
$$K(x) = (1 - |x|^3)^2 I_{|x|<1}$$



Automatikus sávszélesség választás?



Automatikus sávszélesség választás?



Automatikus sávszélesség választás?

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

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

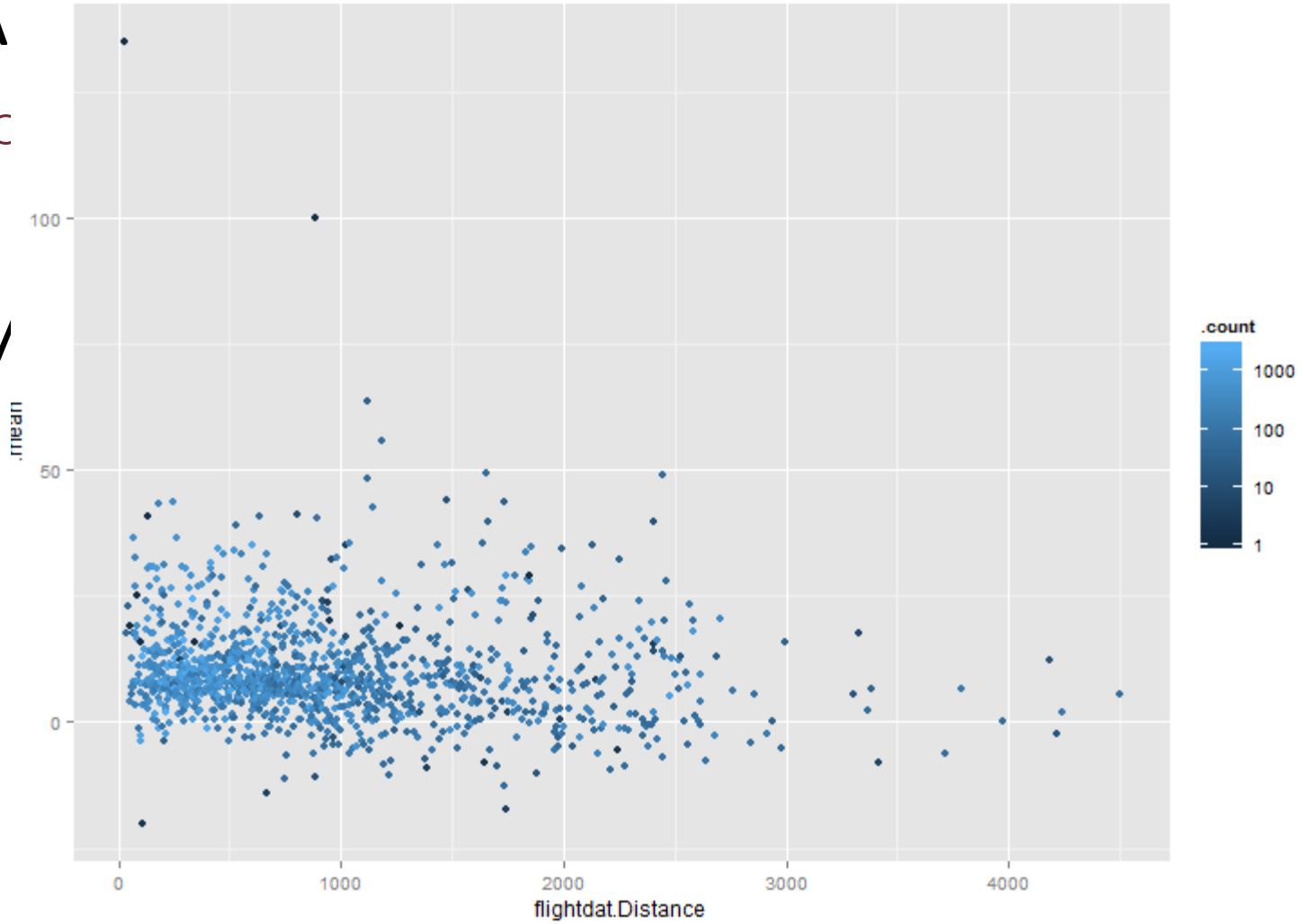
Két változó?

- Az egyik bin, a másik statisztika alapja
  - mean, median, std. dev.

- Mindkettő bin alapja, statisztika: „count”

Két változó?

- A
  - C
- M




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

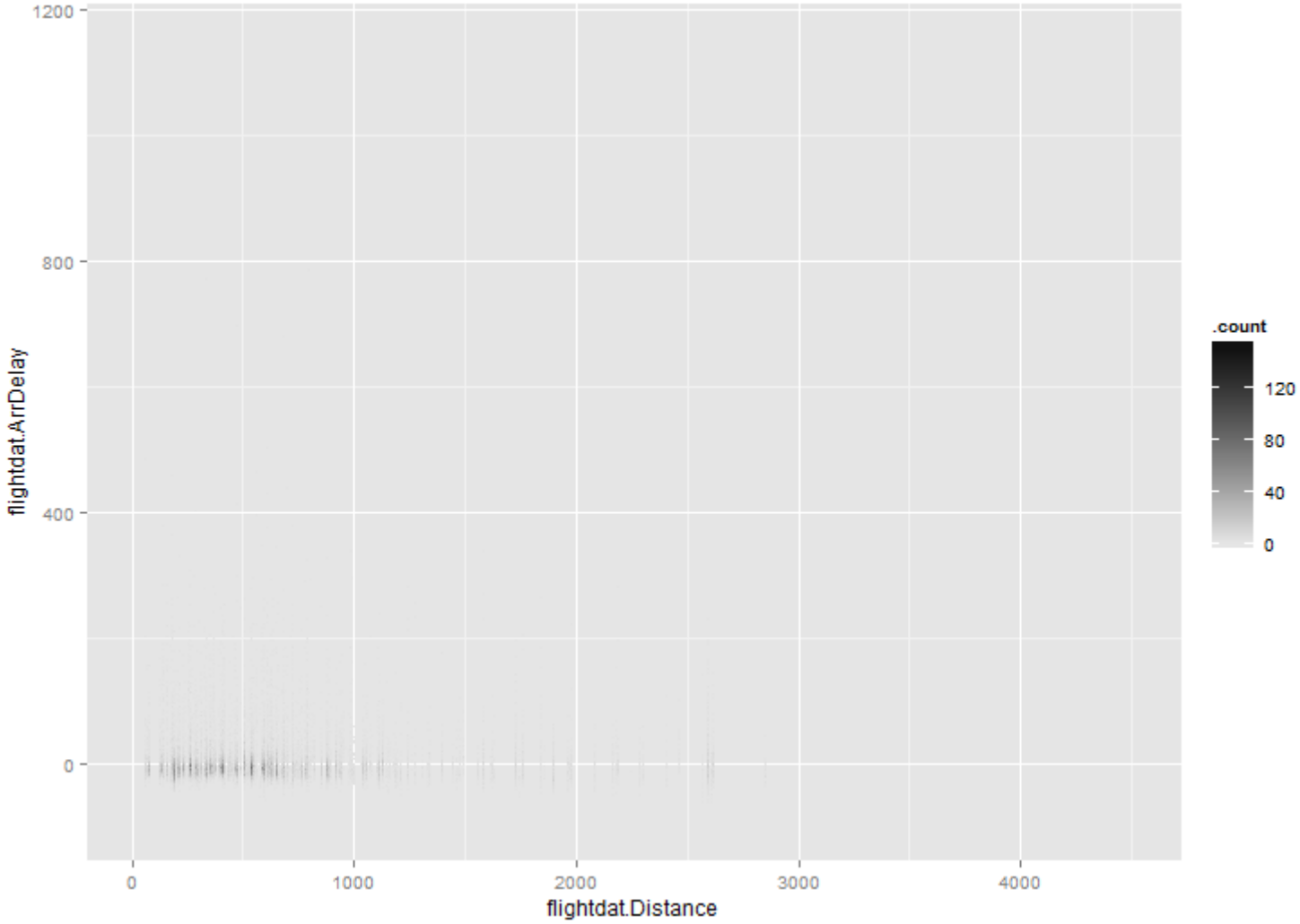
Két változó?

- Az egyik bin, a másik statisztika alapja
  - mean, median, std. dev.

- Mindkettő bin alapja, statisztika: „count”

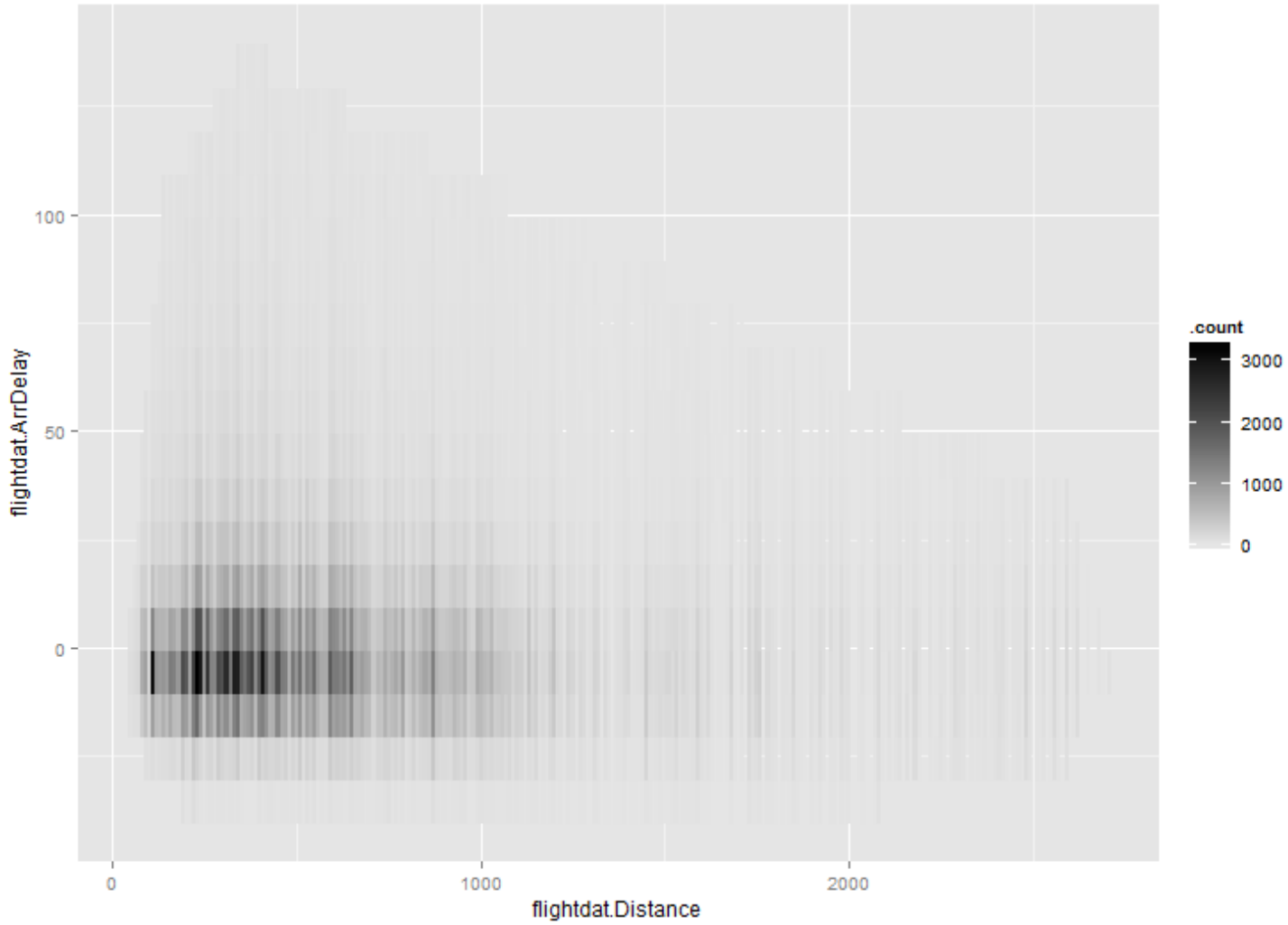
Két változó?

- Az egyik bin, a másik statisztika alapja
  -
- M



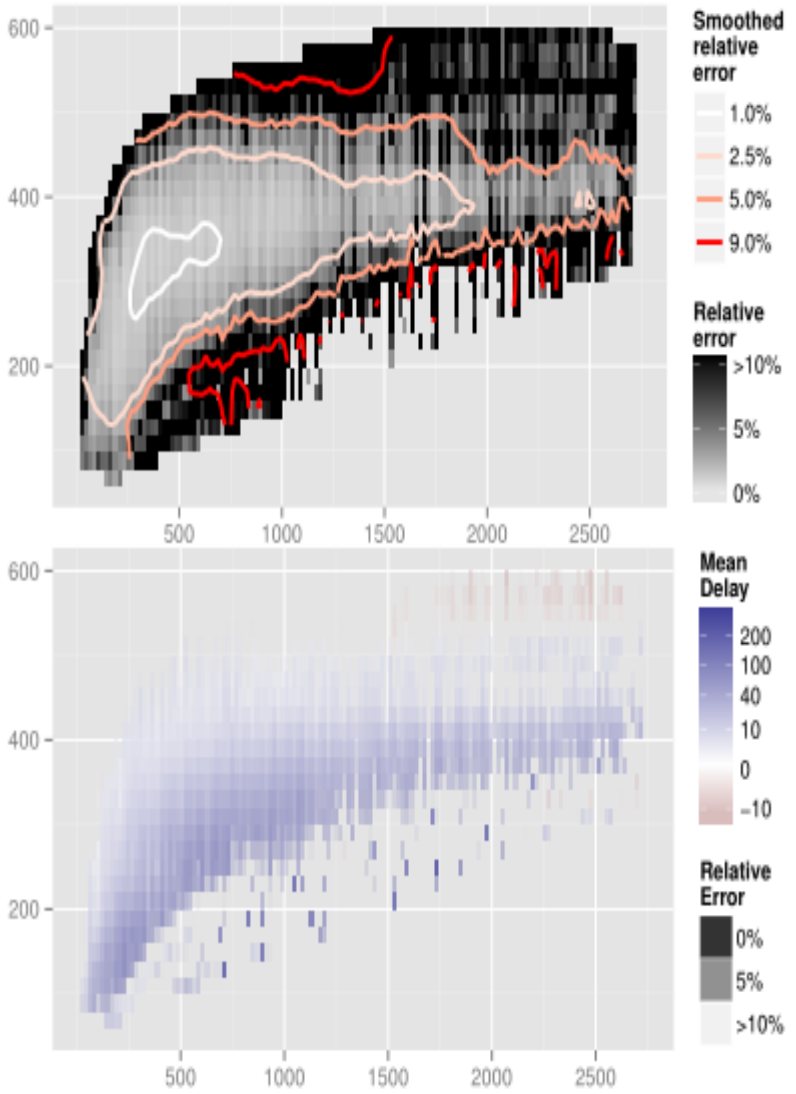
Két változó?

- Az egyik bin, a másik statisztika alapja
  -
- M



Vizualizáció

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



Ábra forrása: [1]

Hivatkozások

- [1] H. Wickham: Bin-summarize-smooth: A framework for visualizing large data. http://vita.had.co.nz/papers/bigvis.pdf (*A cikk az IEEE Transactions on Visualization and Computer Graphics folyóiratban fog megjelenni.*)
- [2] Bigvis-t bemutató meetup oldala: http://www.meetup.com/nyhackr/events/112271042/

Előkészületek

```
flightdat <- read.table('C:/Users/ikocsis/Desktop/2008_sm.csv', header=TRUE, sep=',')
save(flightdat, file="C:/Users/ikocsis/Desktop/flightdat.RData")

install.packages('devtools')
install_github('bigvis', 'hadley')
library('bigvis')
library('ggplot2')

colnames(flightdat)
summary(flightdat)
```

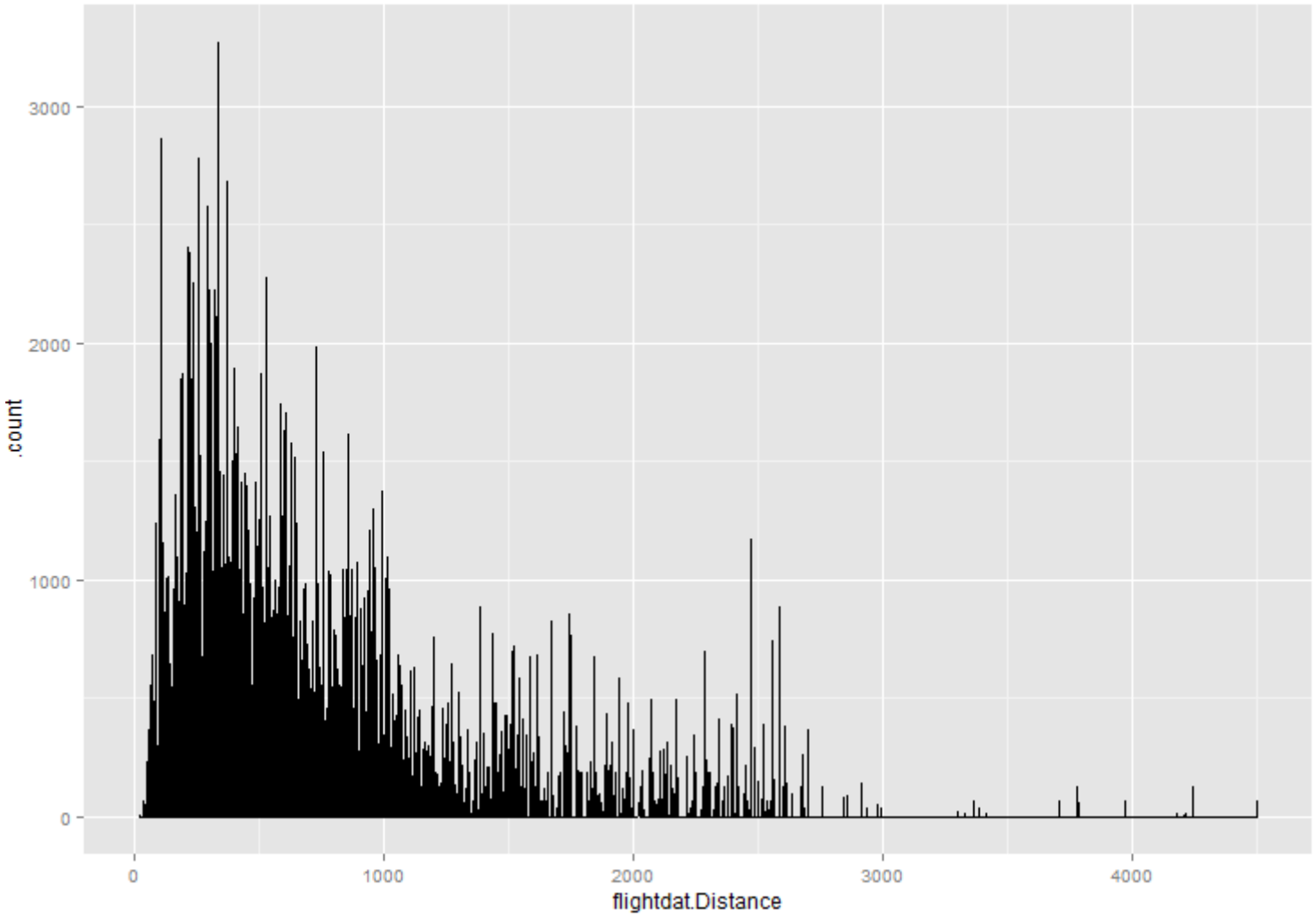
Előkészületek

```
> summary(flightdat)
      Year          Month          DayofMonth       DayOfWeek        DepTime       CRSDepTime       ArrTime       CRSArrTime    UniqueCarrier
 Min.   :2008   Min.   :1   Min.   : 1.00   Min.   :1.000   Min.   :   1   Min.   :  10   Min.   :   1   Min.   :   1   WN      :101396
 1st Qu.:2008   1st Qu.:1   1st Qu.: 8.00   1st Qu.:2.000   1st Qu.: 938   1st Qu.: 932   1st Qu.:1116   1st Qu.:1120   OO      : 48992
 Median :2008   Median :1   Median :16.00   Median :4.000   Median :1335   Median :1325   Median :1520   Median :1520   MQ      : 43454
 Mean   :2008   Mean   :1   Mean   :16.22   Mean   :3.842   Mean   :1344   Mean   :1332   Mean   :1491   Mean   :1498   US      : 39226
 3rd Qu.:2008   3rd Qu.:1   3rd Qu.:24.00   3rd Qu.:5.000   3rd Qu.:1731   3rd Qu.:1715   3rd Qu.:1910   3rd Qu.:1905   UA      : 38026
 Max.   :2008   Max.   :1   Max.   :31.00   Max.   :7.000   Max.   :2400   Max.   :2359   Max.   :2400   Max.   :2400   XE      : 35058
                                                            NA's   :15292                 NA's   :16367                 (Other):193847
   FlightNum        TailNum      ActualElapsedTime CRSElapsedTime      AirTime         ArrDelay          DepDelay          Origin
 Min.   :   1          :  8607   Min.   : 16.0     Min.   :   1.0   Min.   :  0.00   Min.   : -91.00   Min.   : -92.00   ATL     : 27370
 1st Qu.: 687   N484HA :   398   1st Qu.: 75.0     1st Qu.:  76.0   1st Qu.: 54.00   1st Qu.:  -9.00   1st Qu.:  -4.00   ORD     : 25235
 Median :1828   N480HA :   391   Median :104.0     Median : 105.0   Median : 80.00   Median :   0.00   Median :   0.00   DEN     : 18341
 Mean   :2483   N487HA :   379   Mean   :120.7     Mean   : 121.4   Mean   : 97.48   Mean   :  10.55   Mean   :  11.81   PHX     : 16413
 3rd Qu.:4007   N475HA :   374   3rd Qu.:146.0     3rd Qu.: 146.0   3rd Qu.:121.00   3rd Qu.:  15.00   3rd Qu.:  11.00   LAX     : 16033
 Max.   :7829   N481HA :   370   Max.   :693.0     Max.   :1435.0   Max.   :590.00   Max.   :1147.00   Max.   :1172.00   DFW     : 15366
                (Other):489480   NA's   :16367     NA's   :106      NA's   :16367    NA's   :16367     NA's   :15292     (Other):381241
      Dest          Distance         TaxiIn          TaxiOut         Cancelled      CancellationCode    Diverted       CarrierDelay
 ATL    : 27348   Min.   :  24.0   Min.   :  0.000   Min.   :  0.00   Min.   :0.00000    :484707   Min.   :0.00000   Min.   :   0.0
 ORD    : 25267   1st Qu.: 306.0   1st Qu.:  4.000   1st Qu.: 10.00   1st Qu.:0.00000   A:  5536   1st Qu.:0.00000   1st Qu.:   0.0
 DEN    : 18435   Median : 516.0   Median :  6.000   Median : 14.00   Median :0.00000   B:  6114   Median :0.00000   Median :   0.0
 PHX    : 16371   Mean   : 661.2   Mean   :  6.793   Mean   : 16.42   Mean   :0.03058   C:  3642   Mean   :0.00215   Mean   :  16.3
 LAX    : 16266   3rd Qu.: 853.0   3rd Qu.:  8.000   3rd Qu.: 19.00   3rd Qu.:0.00000              3rd Qu.:0.00000   3rd Qu.:  17.0
 DFW    : 15104   Max.   :4502.0   Max.   :213.000   Max.   :383.00   Max.   :1.00000              Max.   :1.00000   Max.   :1120.0
 (Other):381208                     NA's   :16367     NA's   :15292                                                    NA's   :376040
  WeatherDelay       NASDelay       SecurityDelay     LateAircraftDelay
 Min.   :   0.0   Min.   :  0.0   Min.   :  0.0   Min.   :  0.0
 1st Qu.:   0.0   1st Qu.:  0.0   1st Qu.:  0.0   1st Qu.:  0.0
 Median :   0.0   Median :  3.0   Median :  0.0   Median :  0.0
 Mean   :   3.1   Mean   : 15.7   Mean   :  0.1   Mean   : 21.5
 3rd Qu.:   0.0   3rd Qu.: 19.0   3rd Qu.:  0.0   3rd Qu.: 27.0
 Max.   :1049.0   Max.   :896.0   Max.   :136.0   Max.   :897.0
 NA's   :376040   NA's   :376040  NA's   :376040  NA's   :376040
```

Válasszunk egy változót…

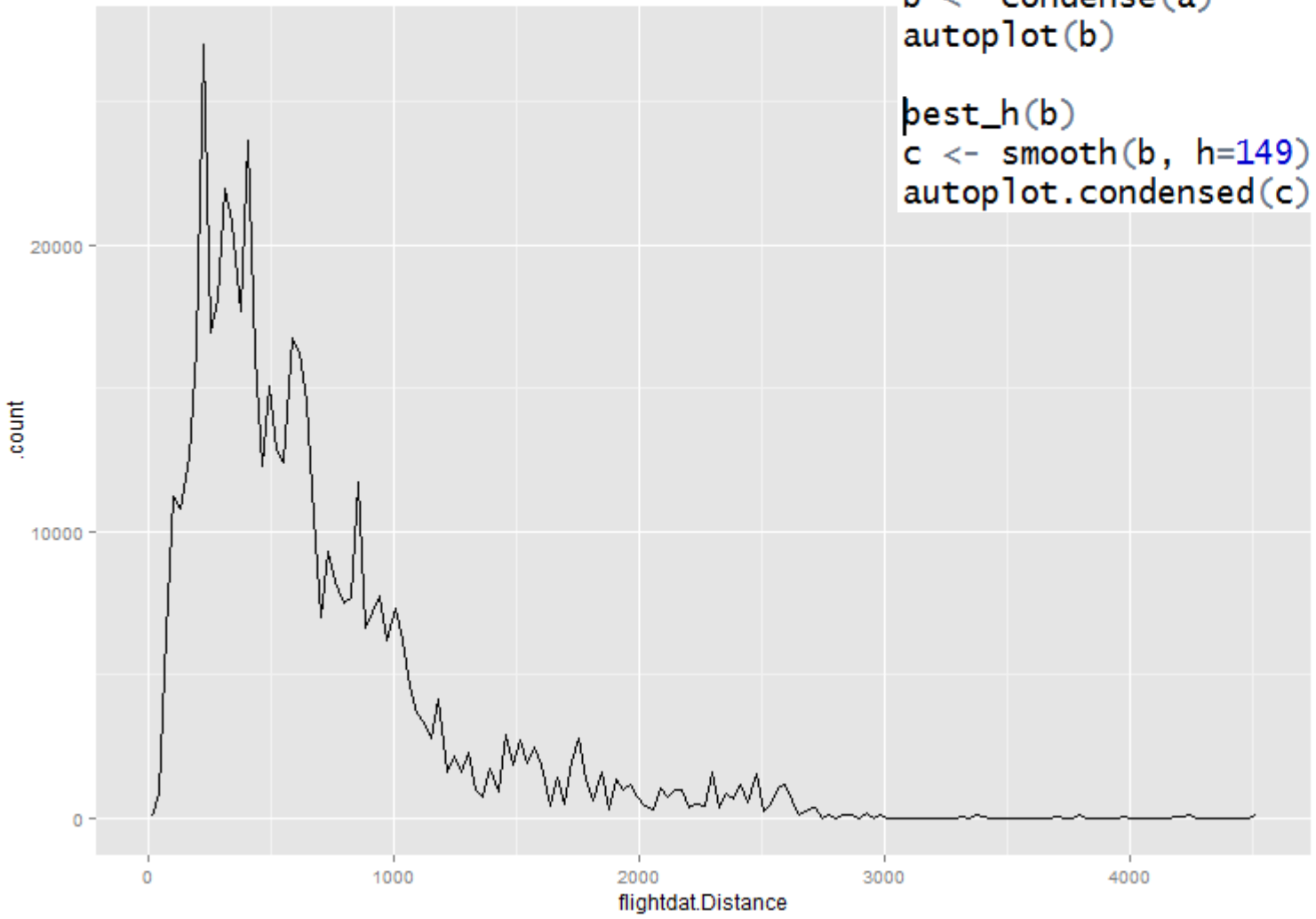
```
> a <- bin(flightdat$Distance)
> a
Binned [499999]. Width: 0.5 Origin: 23.5
> b <- condense(a)
Summarising with count
> head(b)
  flightdat.Distance .count
1                 NA      0
2              23.75      0
3              24.25      1
4              24.75      0
5              25.25      0
6              25.75      0
> autoplot.condensed(b)
>
```

Válasszunk egy változót…



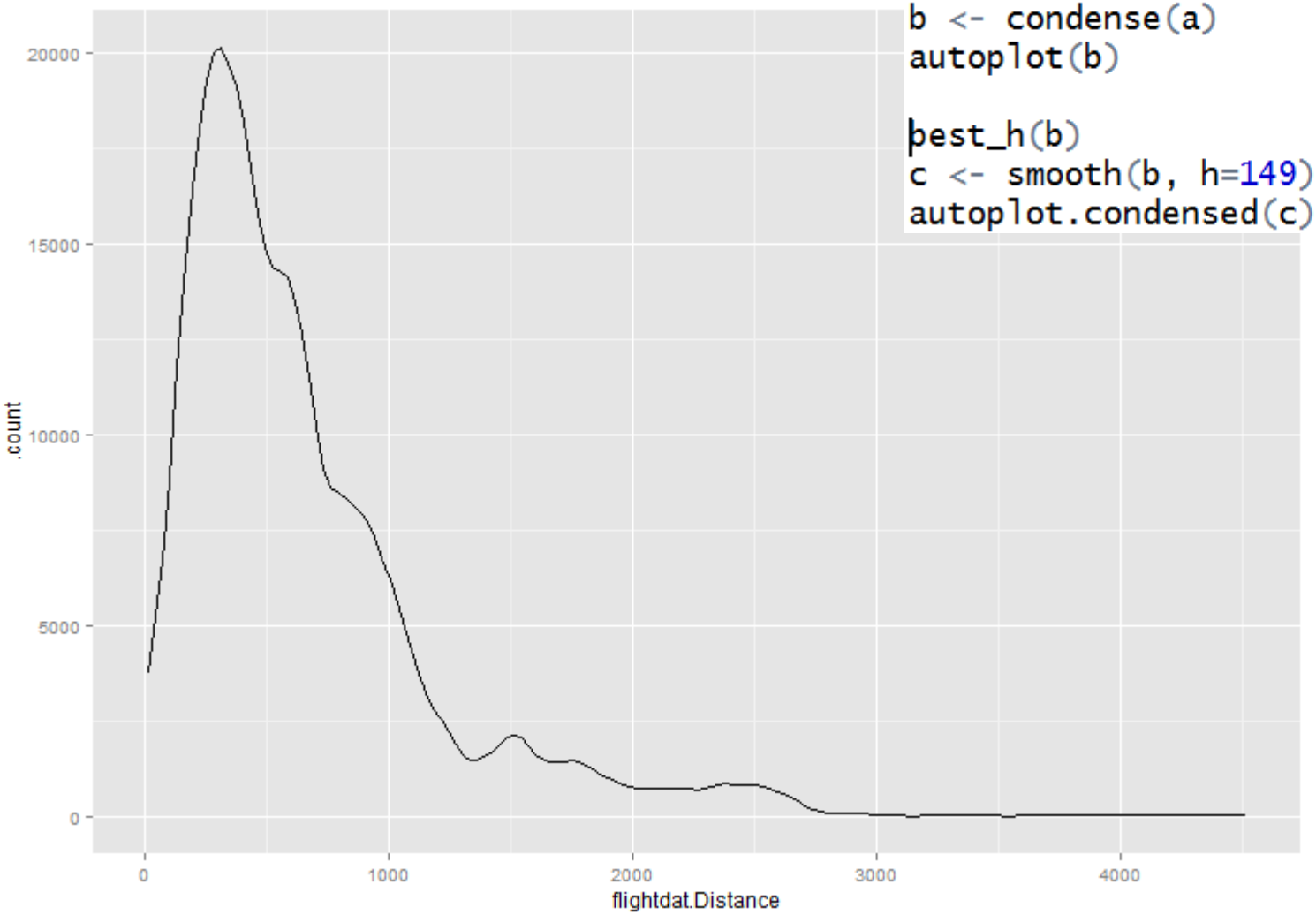
Binning; simítás

```
a <- bin(flightdat$Distance, width=30)
b <- condense(a)
autoplot(b)

best_h(b)
c <- smooth(b, h=149)
autoplot.condensed(c)
```

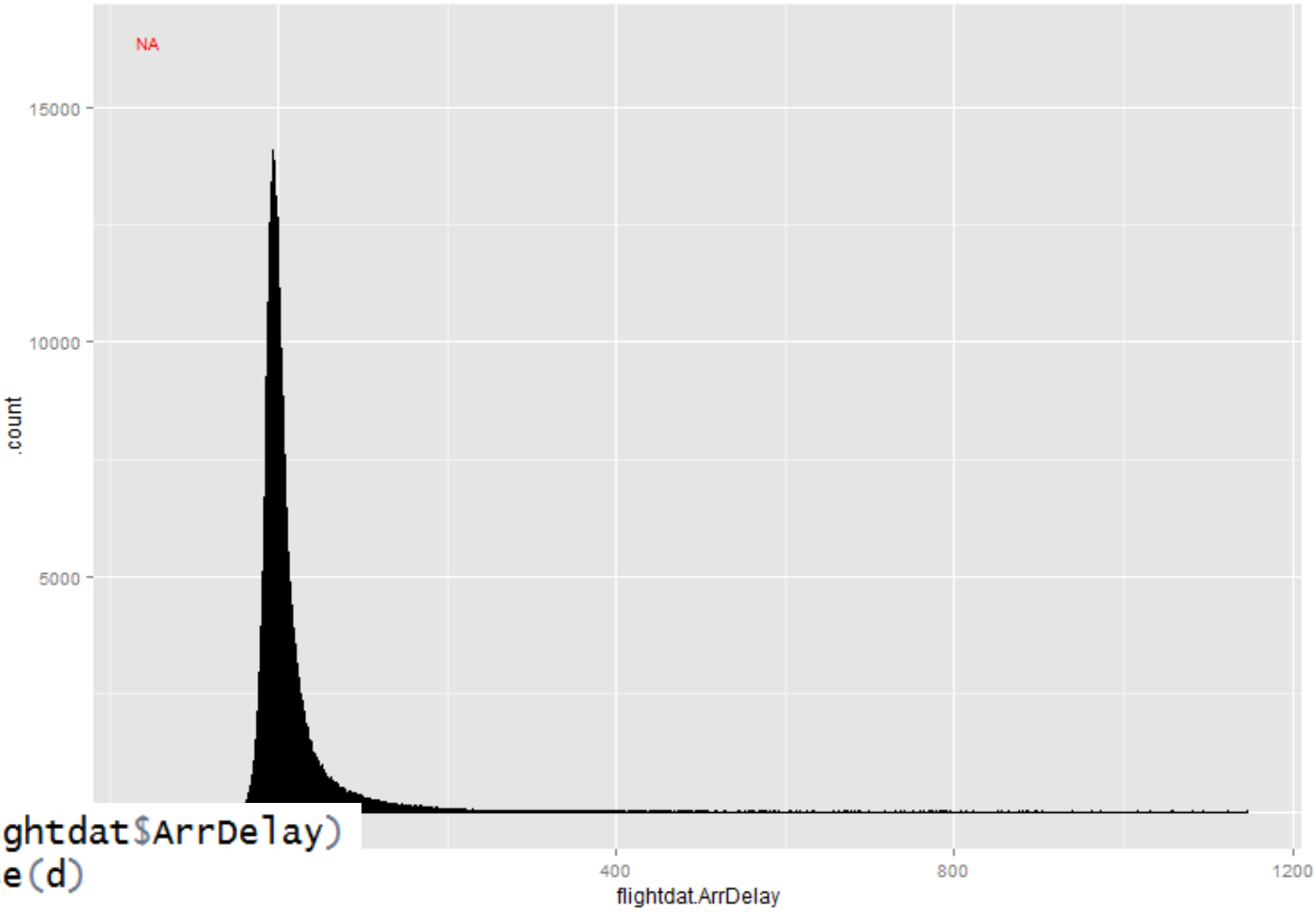


Binning; simítás

```
a <- bin(flightdat$Distance, width=30)
b <- condense(a)
autoplot(b)

best_h(b)
c <- smooth(b, h=149)
autoplot.condensed(c)
```

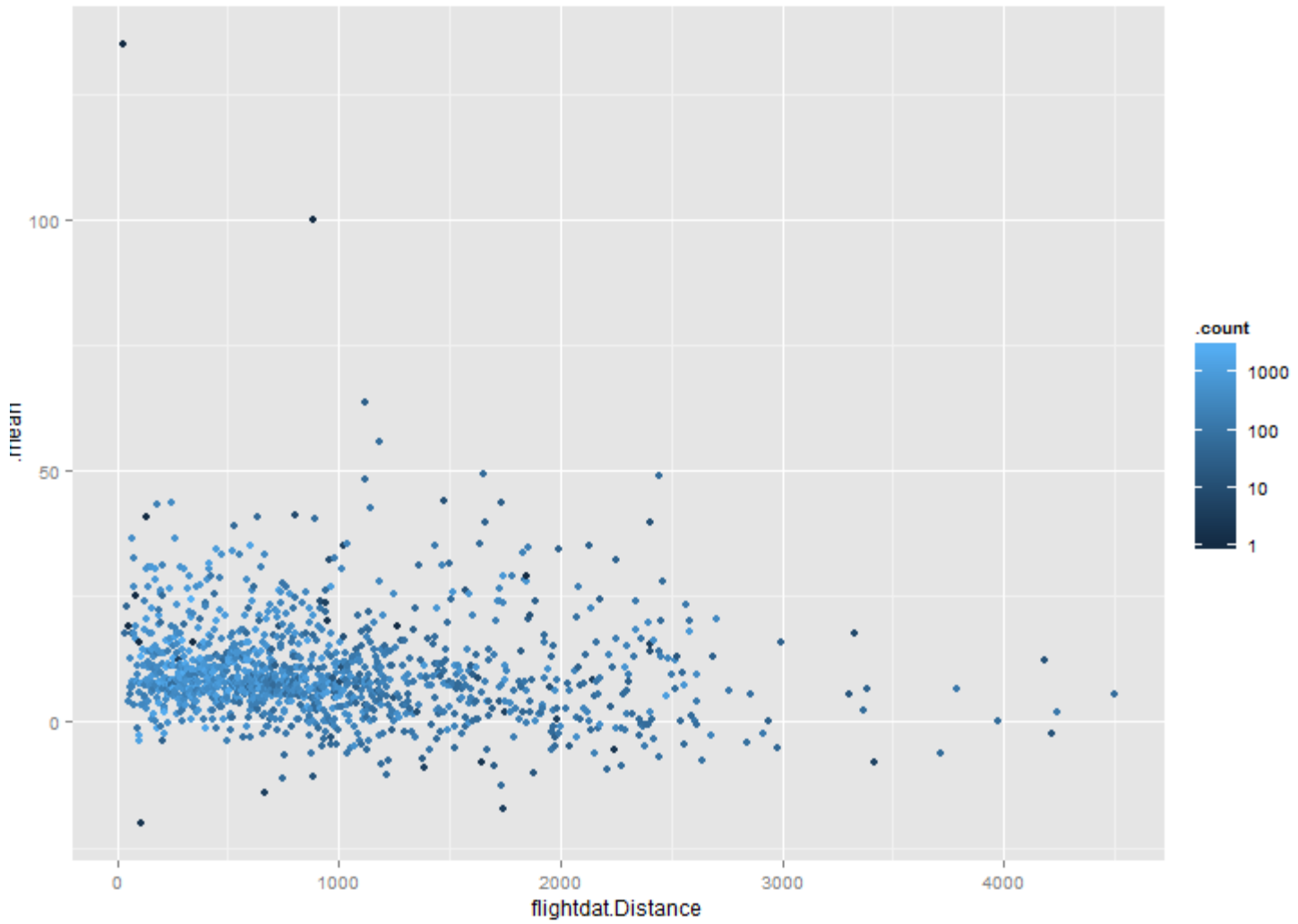


Másik változó




```
d <- bin(flightdat$ArrDelay)
e <- condense(d)
autoplot(e)
```

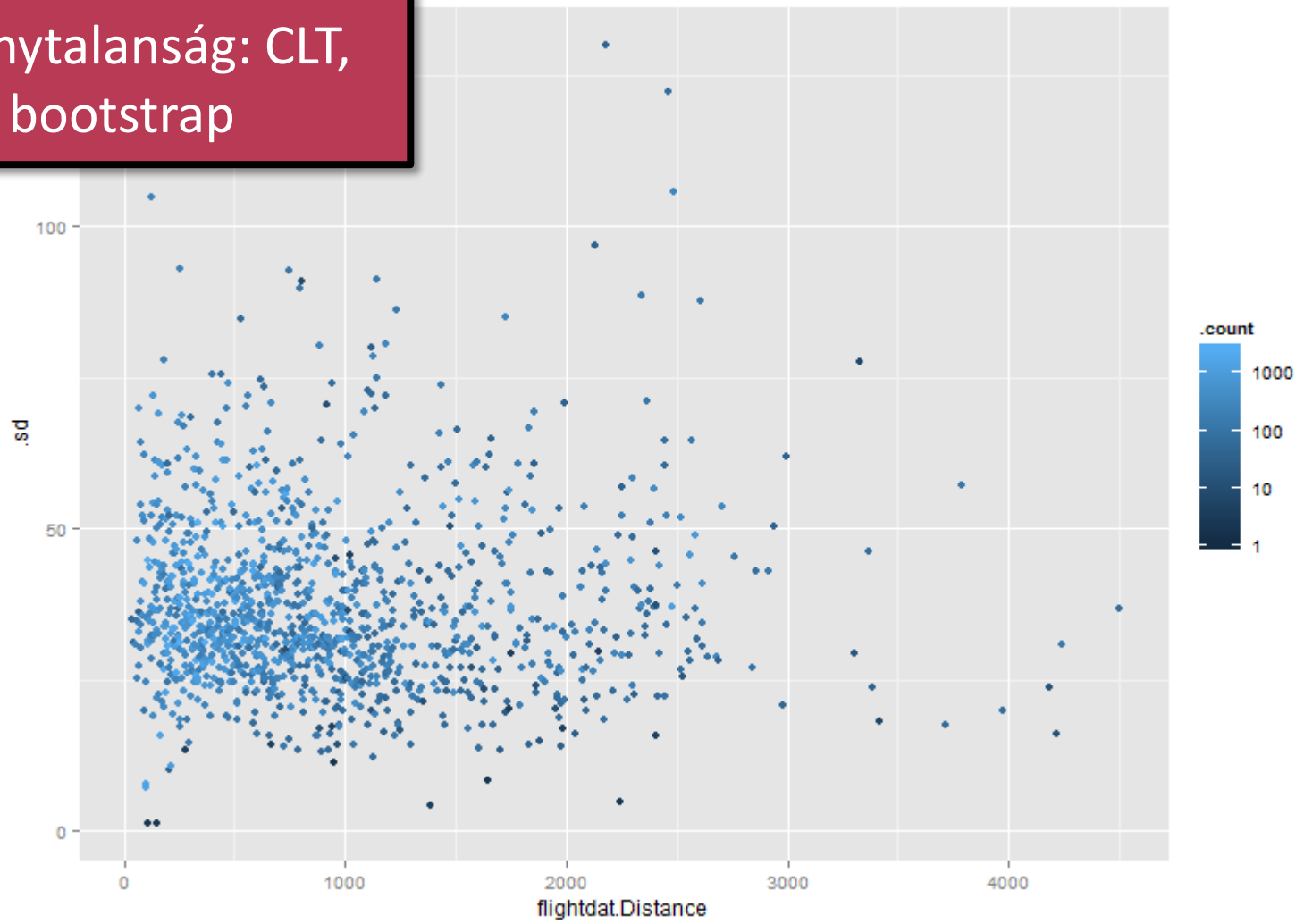
Két változó




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

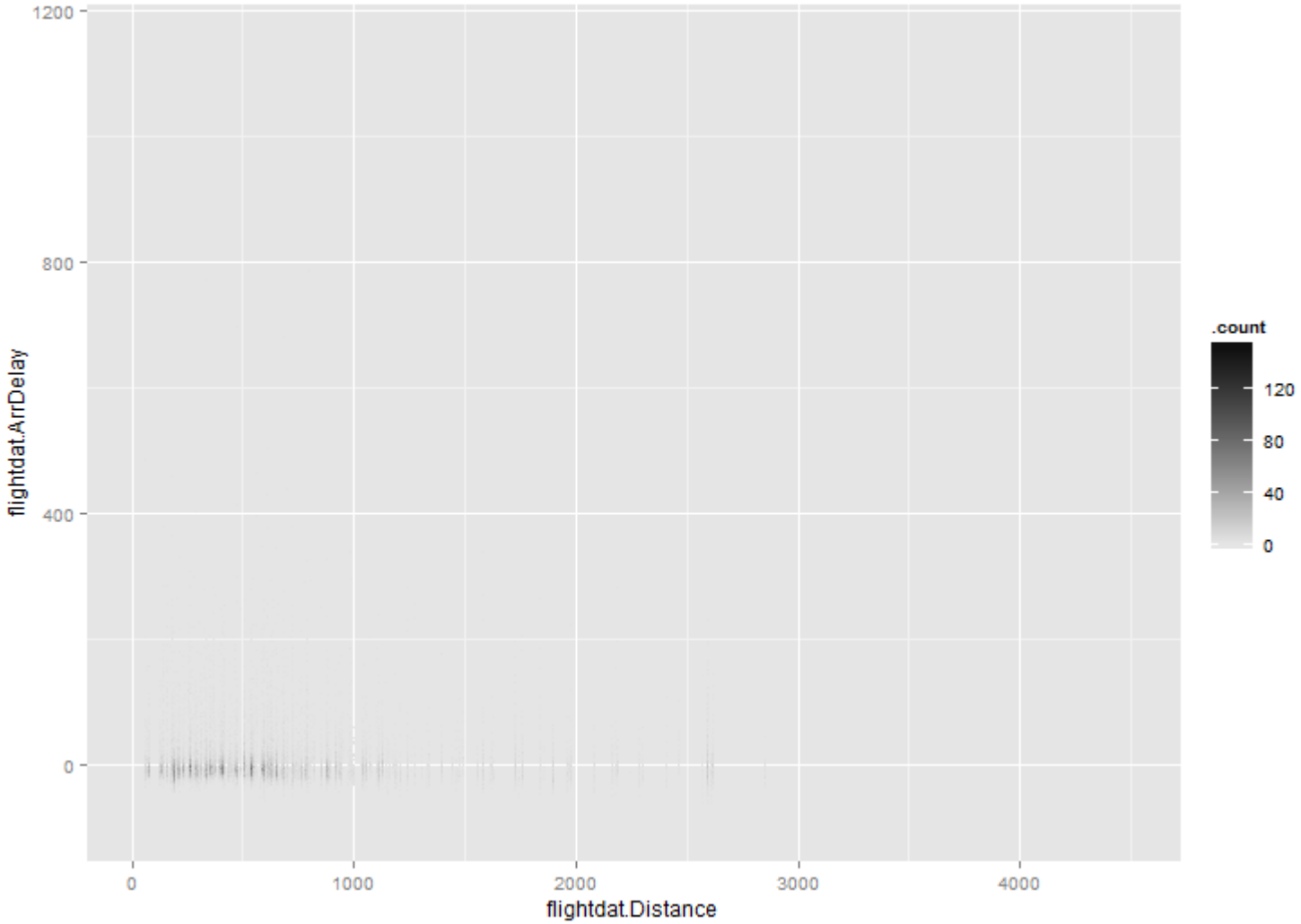
Két változó

Bizonytalanság: CLT, bootstrap



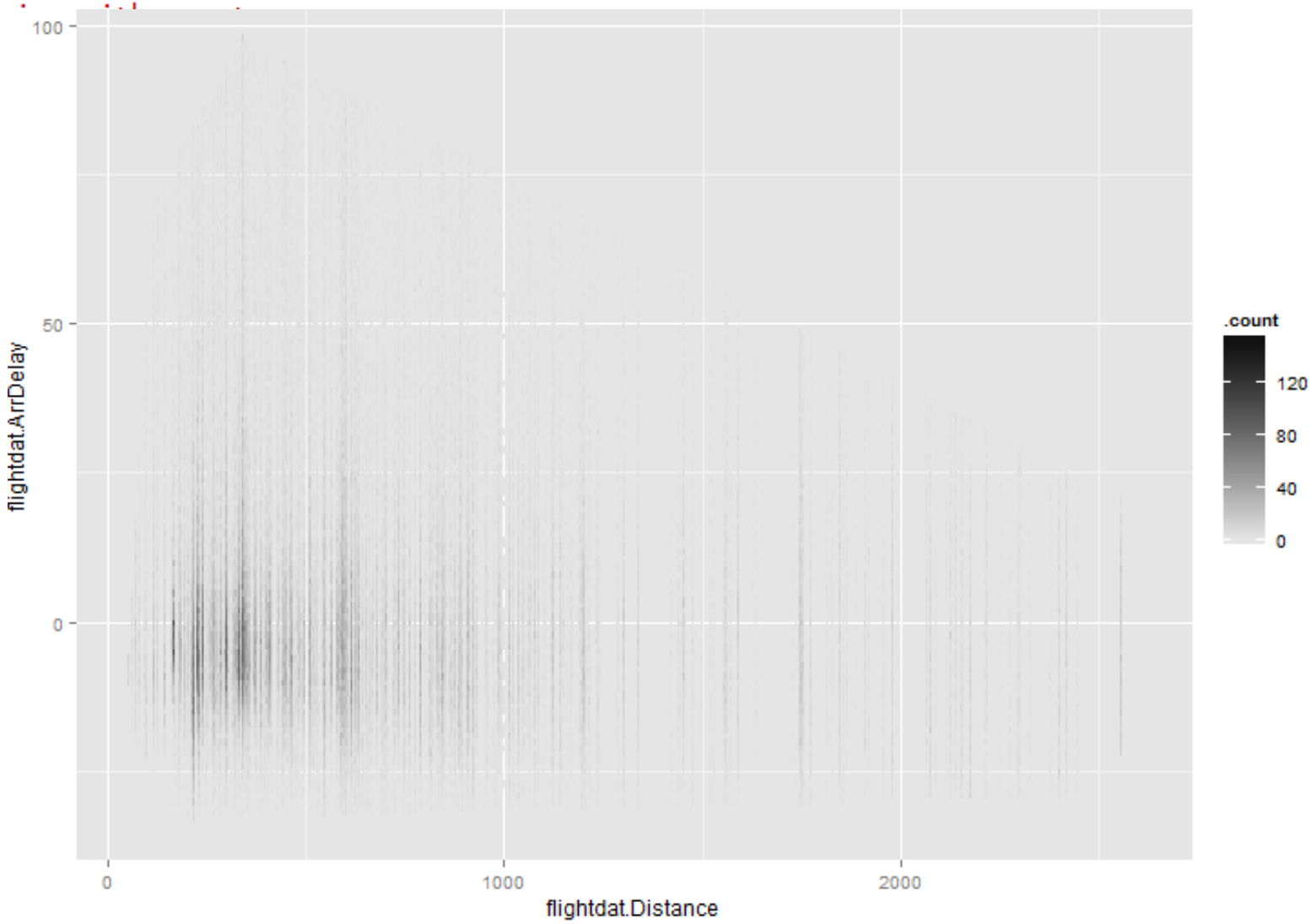

```
> autoplot(condense(bin(flightdat$Distance), z=flightdat$ArrDelay, summary="sd"))
```

...


1200
800
400
0
flightdat.ArrDelay
0
1000
2000
3000
4000
flightdat.Distance
.count
120
80
40
0

„Hámozás”

```
> w <- peel(condense(bin(flightdat$Distance), bin(flightdat$ArrDelay)), keep=0.5, central=TRUE)
```



… illetve kézivezérlés

```
> autoplot(peel(condense(bin(flightdat$Distance, width=10), bin(flightdat$ArrDelay, 10)), keep=0.9))
```