Visualization and descriptive statistics

D.A. Forsyth

What's going on here?

- Most important, most creative scientific question
- Getting answers
  - Make helpful pictures and look at them
  - Compute numbers in support of making pictures
- Data has types
  - Continuous
  - Discrete
    - Ordinal (can be ordered)
    - Categorical (no natural order, "cat" vs "hat")
- Different plots apply

Histograms

Gender	Goal	Gender	Goal
boy	Sports	girl	Sports
boy	Popular	girl	Grades
girl	Popular	boy	Popular
girl	Popular	boy	Popular
girl	Popular	boy	Popular
girl	Popular	girl	Grades
girl	Popular	girl	Sports
girl	Grades	girl	Popular
girl	Sports	girl	Grades
girl	Sports	girl	Sports

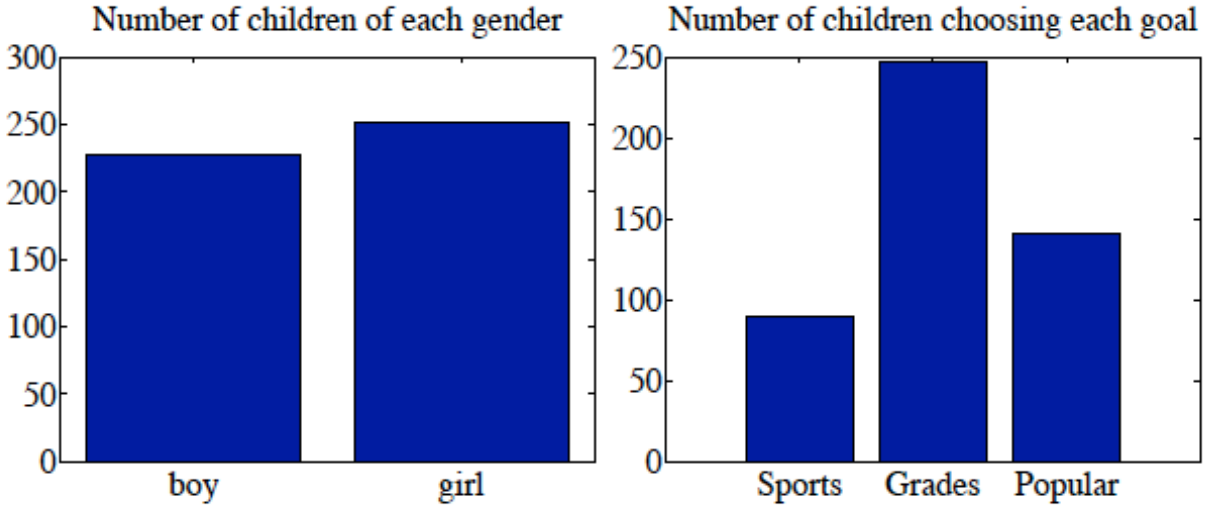
Categorical data

Ick!

Bar Charts

Gender	Goal	Gender	Goal
boy	Sports	girl	Sports
boy	Popular	girl	Grades
girl	Popular	boy	Popular
girl	Popular	boy	Popular
girl	Popular	boy	Popular
girl	Popular	girl	Grades
girl	Popular	girl	Sports
girl	Grades	girl	Popular
girl	Sports	girl	Grades
girl	Sports	girl	Sports

Categorical data - counts in category



Histograms

Index	net worth
1	100, 360
2	109, 770
3	96, 860
4	97, 860
5	108, 930
6	124, 330
7	101, 300
8	112, 710
9	106, 740
10	120, 170

Ick!

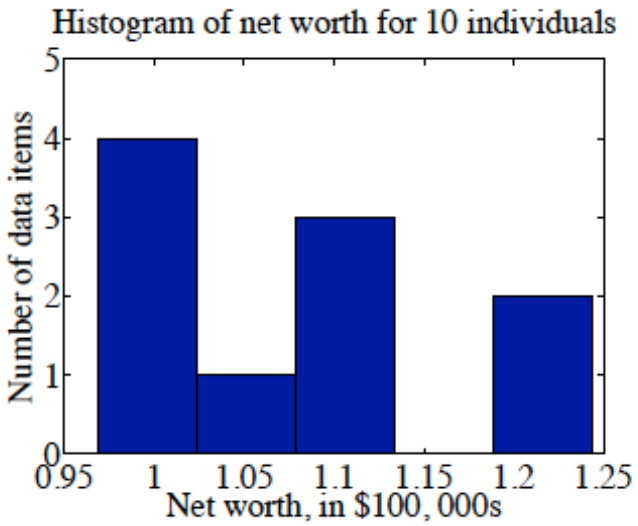
Index	Taste score	Index	Taste score
1	12.3	11	34.9
2	20.9	12	57.2
3	39	13	0.7
4	47.9	14	25.9
5	5.6	15	54.9
6	25.9	16	40.9
7	37.3	17	15.9
8	21.9	18	6.4
9	18.1	19	18
10	21	20	38.9

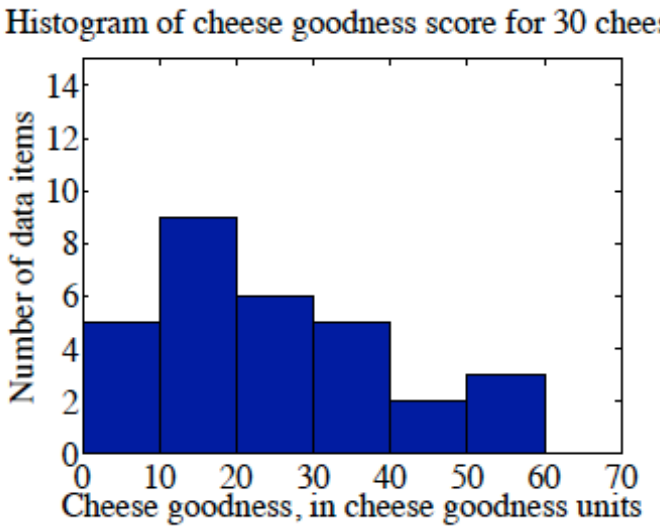
Continuous data

Histograms

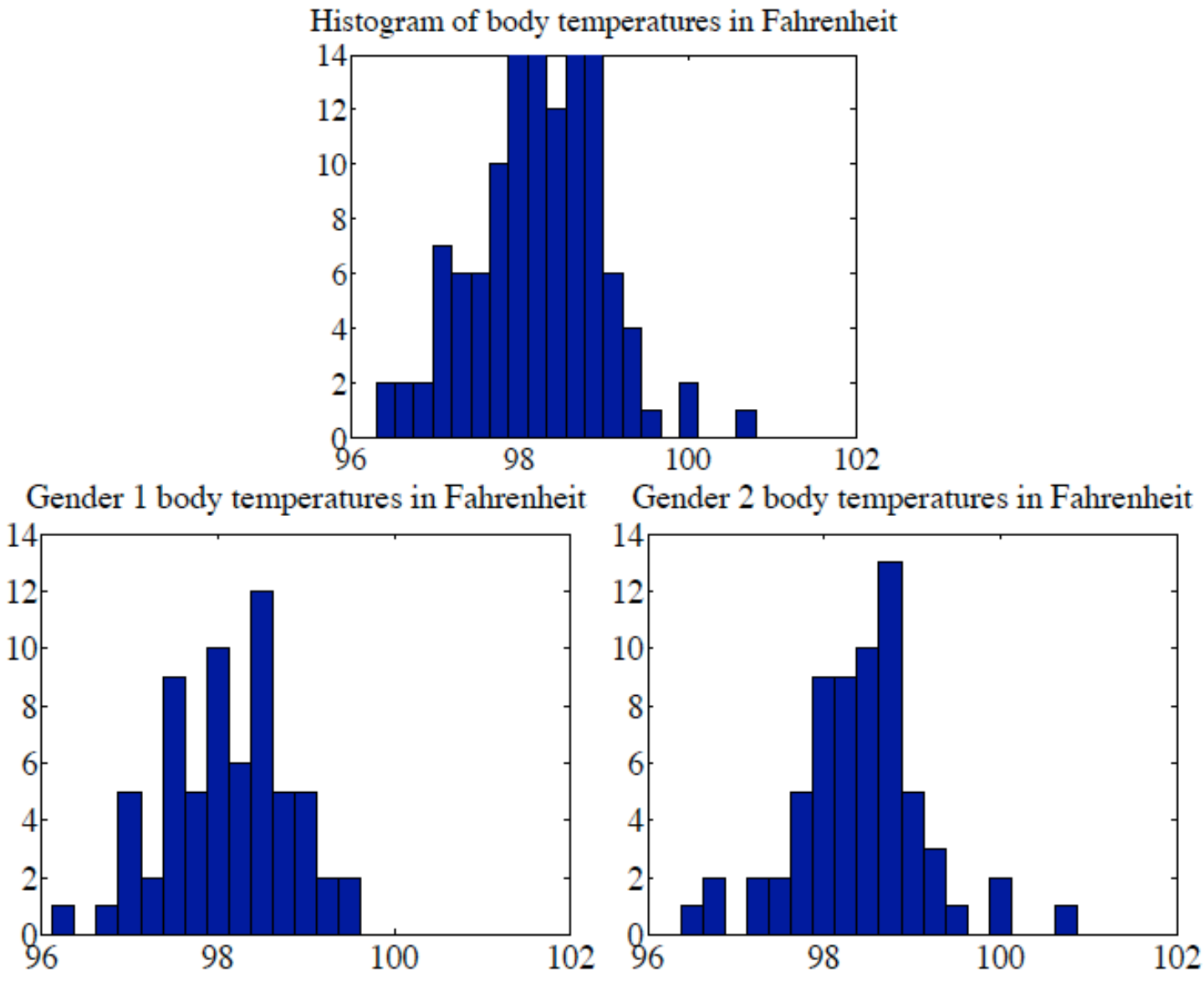
Index	net worth
1	100, 360
2	109, 770
3	96, 860
4	97, 860
5	108, 930
6	124, 330
7	101, 300
8	112, 710
9	106, 740
10	120, 170

Index	Taste score	Index	Taste score
1	12.3	11	34.9
2	20.9	12	57.2
3	39	13	0.7
4	47.9	14	25.9
5	5.6	15	54.9
6	25.9	16	40.9
7	37.3	17	15.9
8	21.9	18	6.4
9	18.1	19	18
10	21	20	38.9





Conditional Histograms



Data example

- Clicks, impressions and ages for NYT website
- https://github.com/oreillymedia/doing_data_science
- Question: Look at data - what's going on?

- Example R code on webpage

Why R?

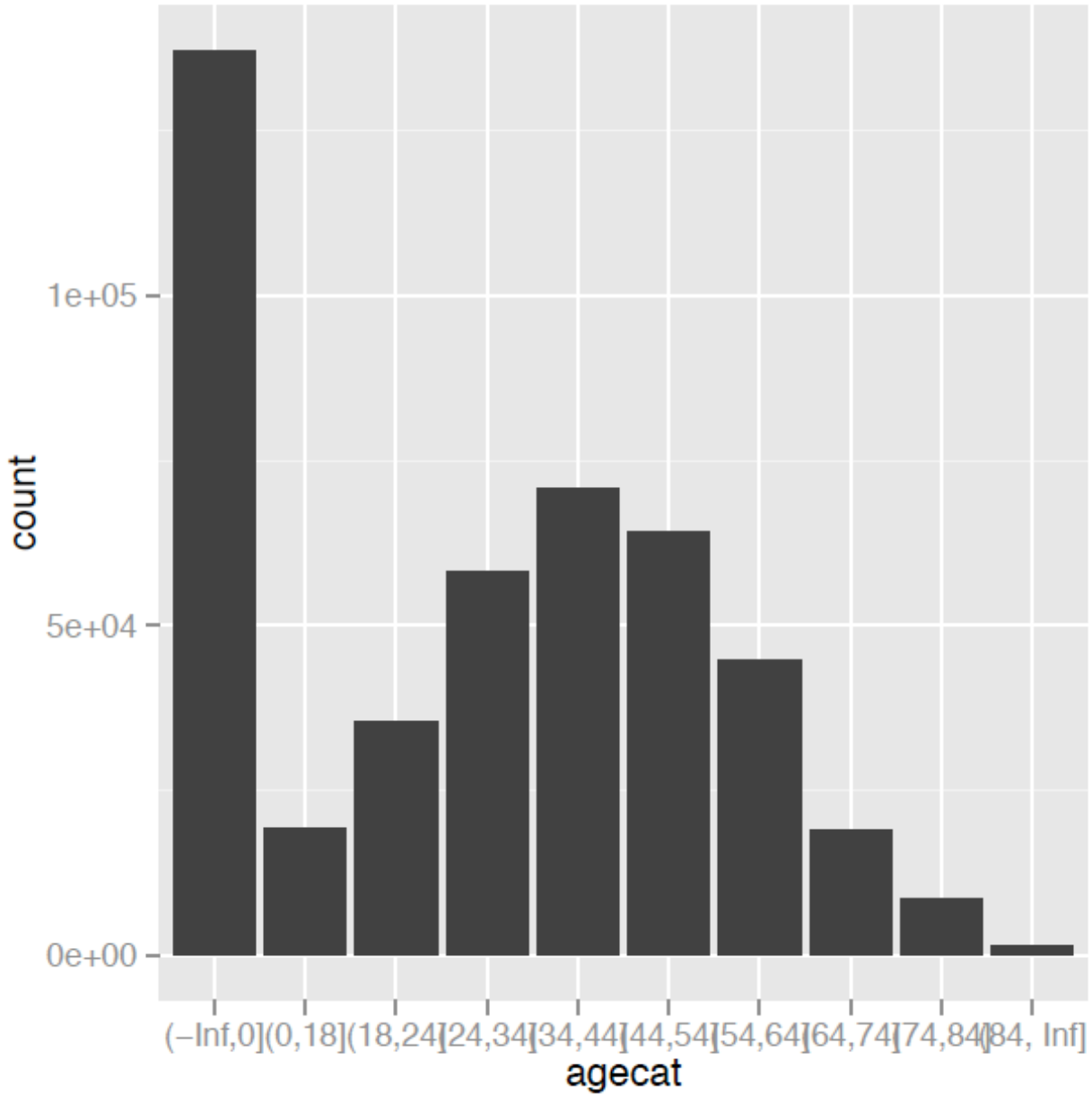
- It's free
- It's easy to get pictures up and going
  - from weirdly formatted datasets
- Many, many tools
  - most of the code I'll work with is downloaded/copied
  - that's the right strategy
  - work with tools *without* implementing them

Some R

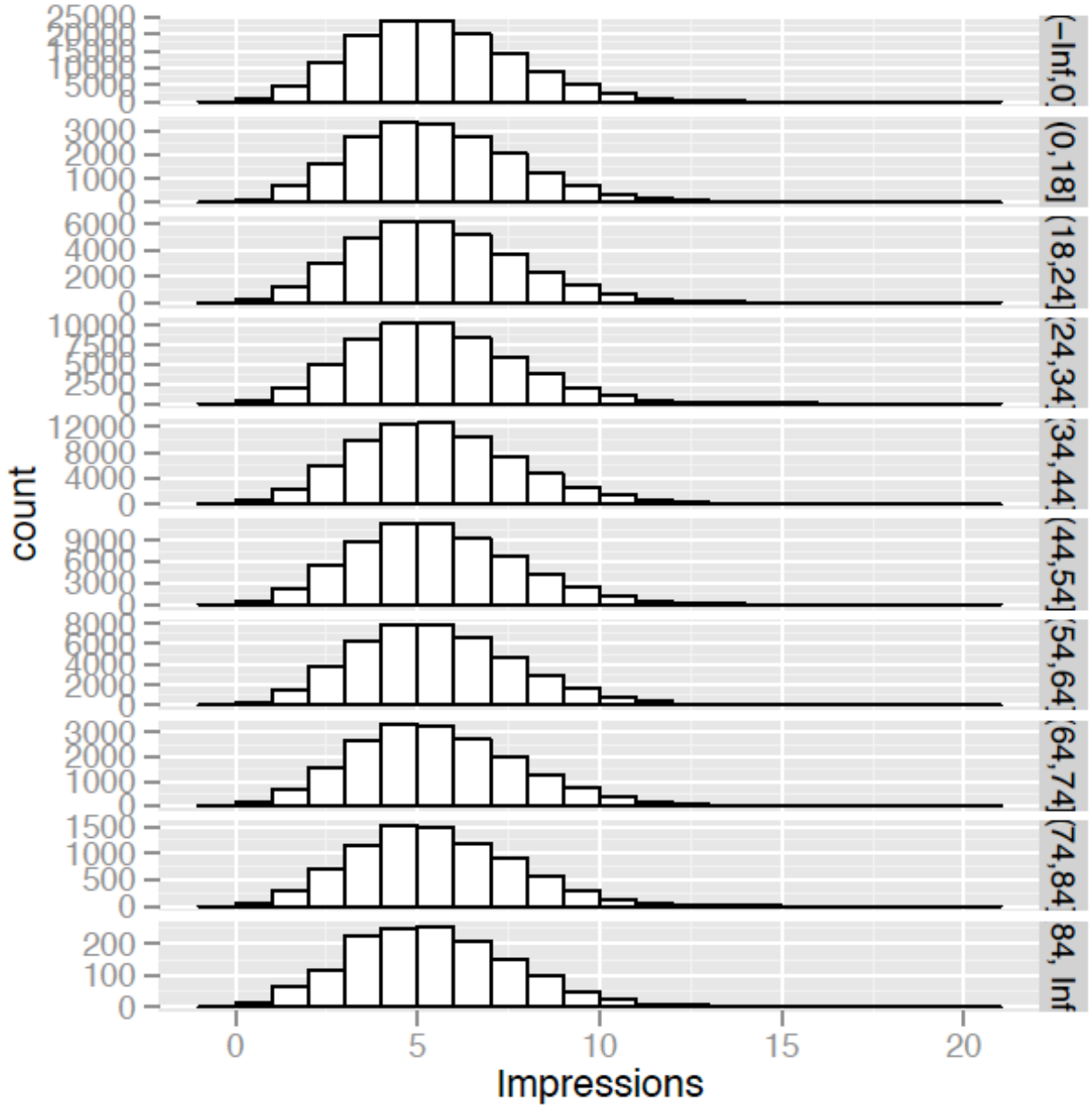
```
setwd('/users/daf/Current/courses/BigData/Examples')

data1<-read.csv('/users/daf/Current/courses/BigData/doing_data_science-master/dds_datasets/dds_ch2_nyt/nyt1.csv')

data1$agecat<-cut(data1$Age, c(-Inf, 0, 18, 24, 34, 44, 54, 64, 74, 84, Inf))
# This breaks the Age column into categories

data1$impcat<-cut(data1$Impressions, c(-Inf, 0, 1, 2, 3, 4, 5, Inf))
# This breaks the impression column into categories

summary(data1)
```

```
      Age            Gender       Impressions         Clicks          Signed_In          agecat          impcat
 Min.   :  0.00   Min.   :0.000   Min.   :  0.000   Min.   :0.00000   Min.   :0.0000   (-Inf,0]:137106   (-Inf,0]:  3066
 1st Qu.:  0.00   1st Qu.:0.000   1st Qu.:  3.000   1st Qu.:0.00000   1st Qu.:0.0000   (34,44] : 70860   (0,1]   : 15483
 Median : 31.00   Median :0.000   Median :  5.000   Median :0.00000   Median :1.0000   (44,54] : 64288   (1,2]   : 38433
 Mean   : 29.48   Mean   :0.367   Mean   :  5.007   Mean   :0.09259   Mean   :0.7009   (24,34] : 58174   (2,3]   : 64121
 3rd Qu.: 48.00   3rd Qu.:1.000   3rd Qu.:  6.000   3rd Qu.:0.00000   3rd Qu.:1.0000   (54,64] : 44738   (3,4]   : 80303
 Max.   :108.00   Max.   :1.000   Max.   :20.000   Max.   :4.00000   Max.   :1.0000   (18,24] : 35270   (4,5]   : 80477
                                                                                       (Other) : 48005   (5, Inf]:176558
```

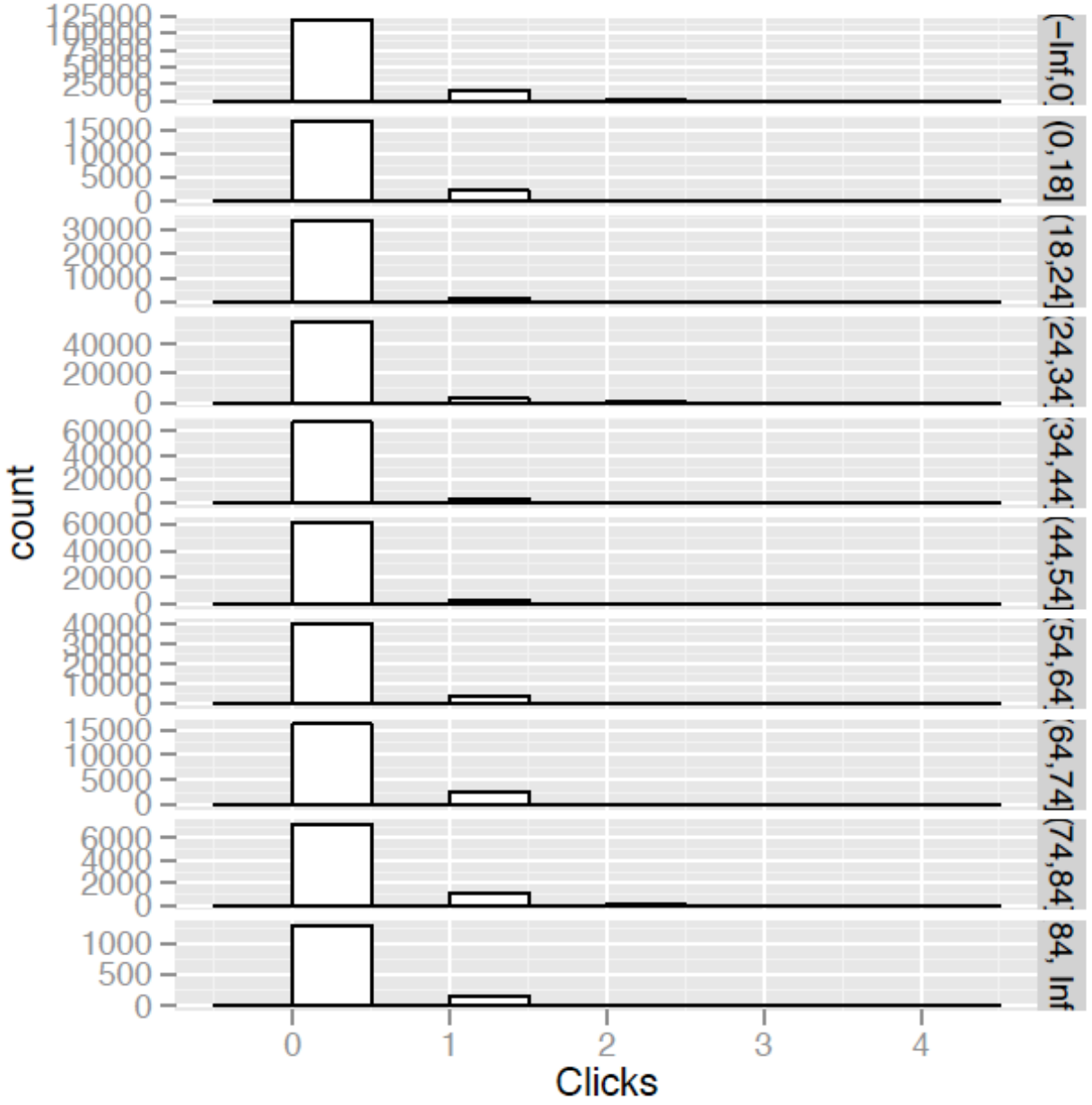
Users by age


1e+05
5e+04
0e+00
count
(−Inf,0]
(0,18]
(18,24]
24,34]
34,44]
44,54]
54,64]
64,74]
74,84]
84, Inf]
agecat

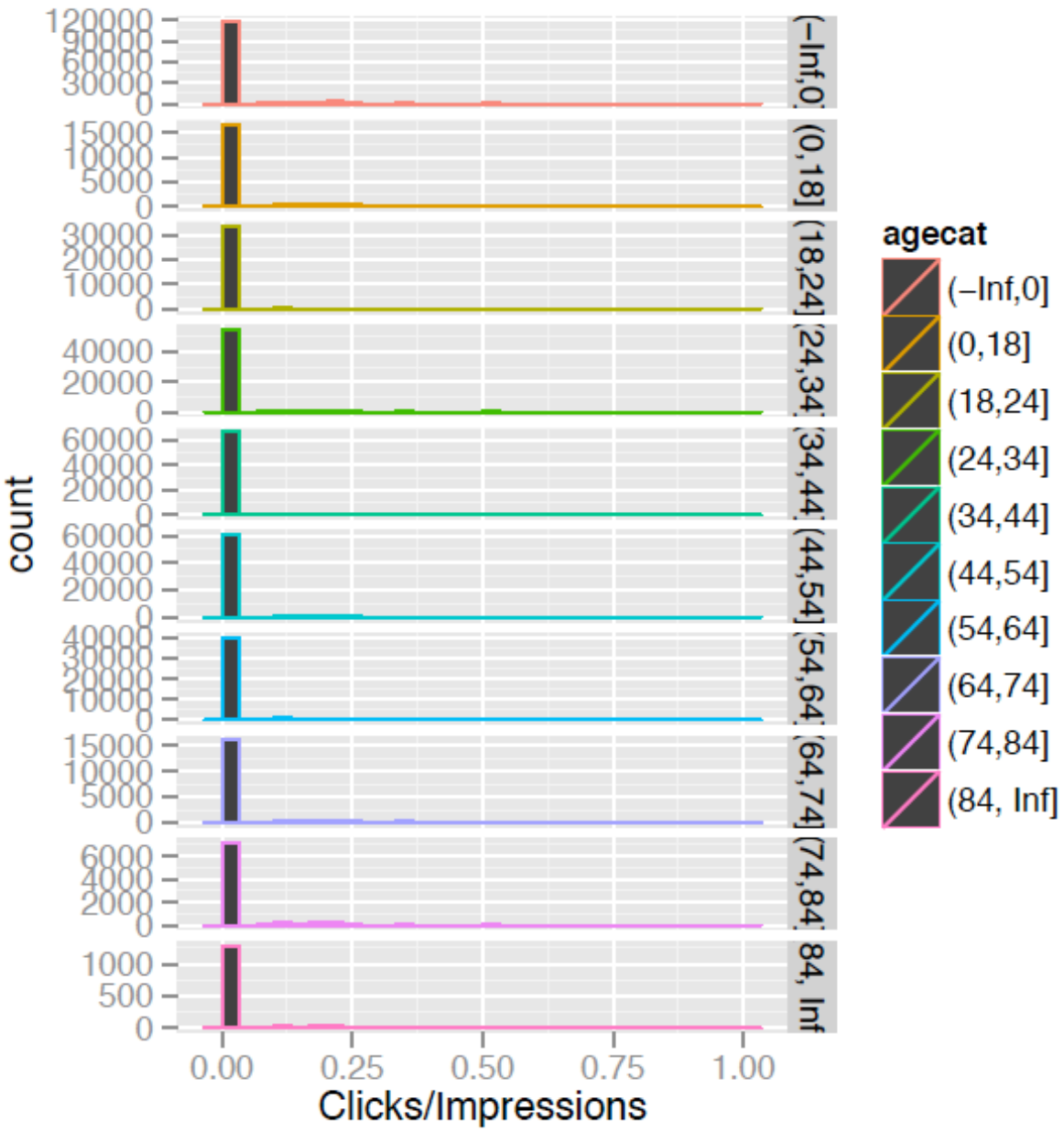
Impression histogram, faceted by age



Click histogram, faceted by age

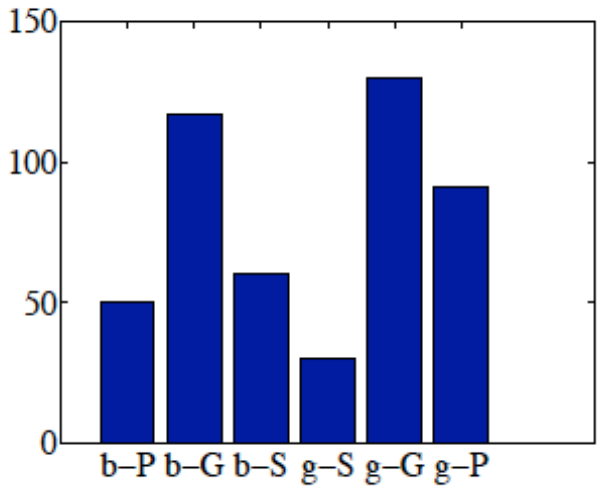


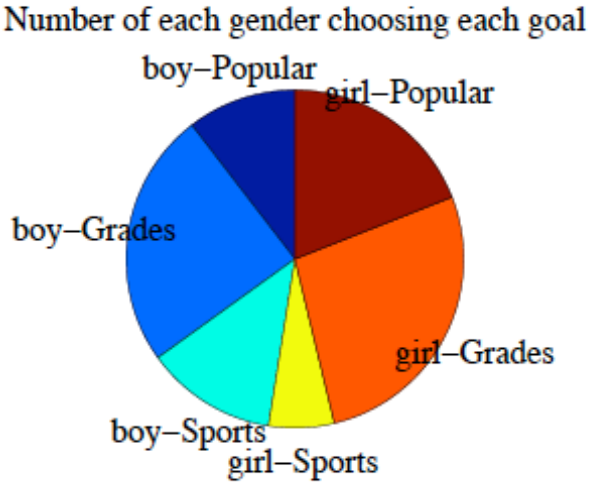
Click/Impression histogram, faceted by age



2D Data

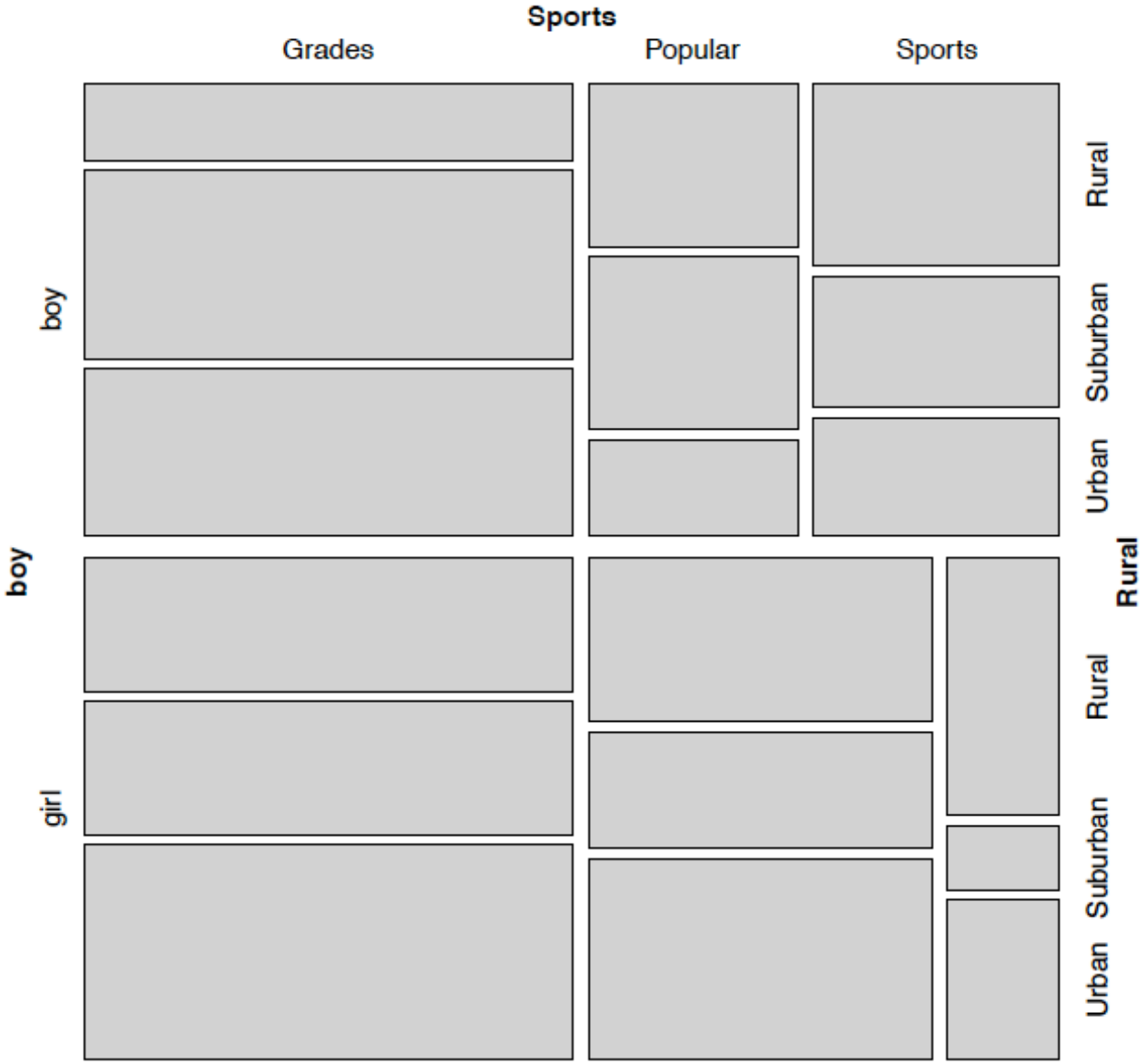
Categorical data





Pie charts are deprecated - it's hard to judge area by eye accurately

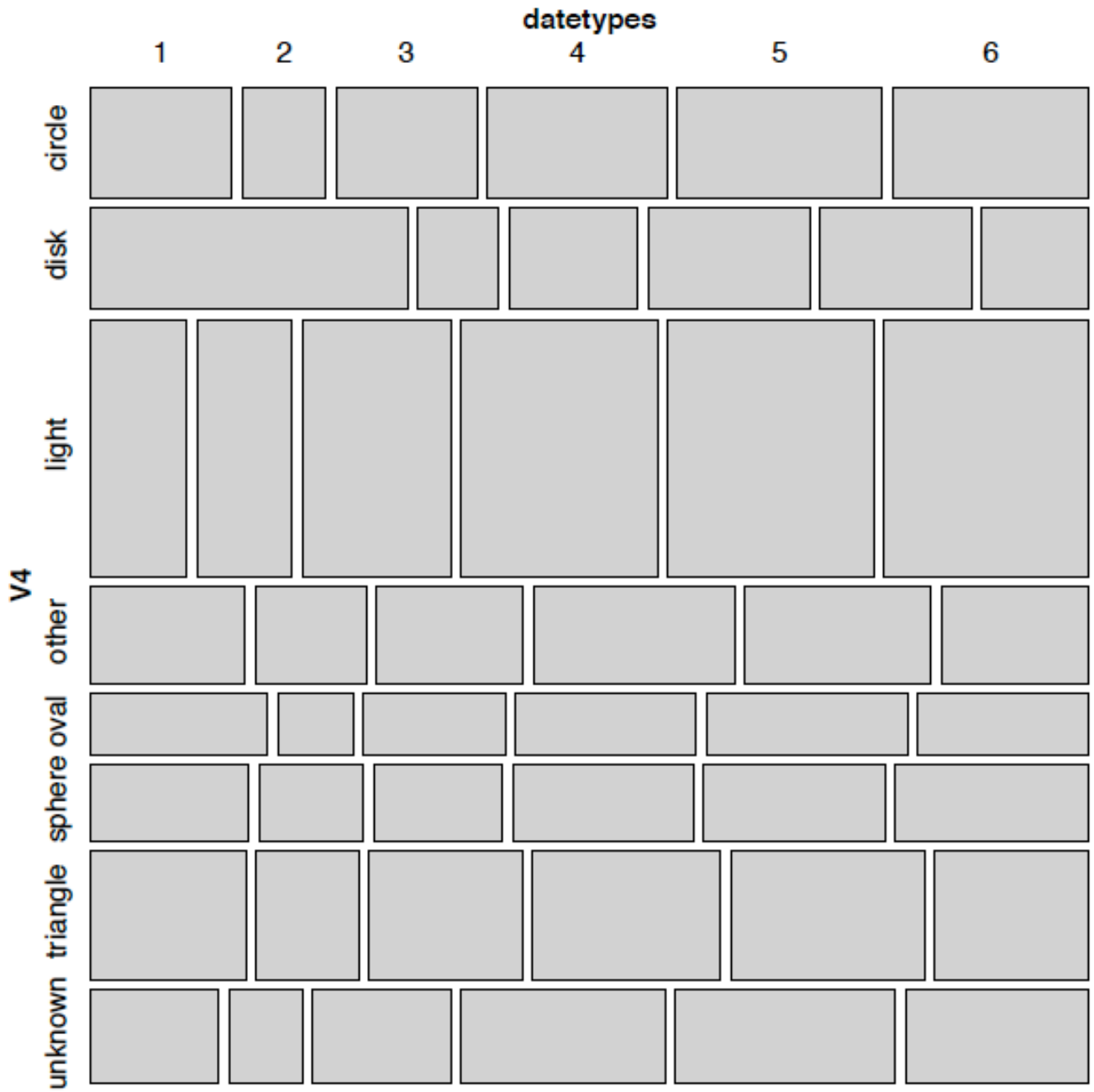
Mosaic Plots



The UFO data set

http://www.infochimps.com/datasets/60000-documented-ufo-sightings-with-text-descriptions-and-metada

- UFO sighting data
  - date of sighting; date of report; location; description; some free text
    - rather messy data
  - about 15 years of sightings (‘95 - ’08 with some others)
  - broke into 1000 day blocks
  - looked at most common shape descriptors
    - (' disk', ' light', ' circle', ' triangle', ' sphere', ' oval', ' other', ' unknown')
  - great example of categorical data
- R-code on website
  - not great code, but informative
    - building a map, merging datasets, reading datasets, mosaic plots
  - you should look at this



Conclusion: UFO shapes haven't changed over time

Ordinal data

	-2	-1	0	1	2
-2	24	5	0	0	1
-1	6	12	3	0	0
0	2	4	13	6	0
1	0	0	3	13	2
2	0	0	0	1	5

TABLE 2.3: *I simulated data representing user evaluations of a user interface. Each cell in the table on the* **left** *contains the count of users rating "ease of use" (horizontal, on a scale of -2 -very bad- to 2 -very good) vs. "enjoyability" (vertical, same scale). Users who found the interface hard to use did not like using it either. While this data is categorical, it's also ordinal, so that the order of the cells is determined. It wouldn't make sense, for example, to reorder the columns of the table or the rows of the table.*

Ordinal data

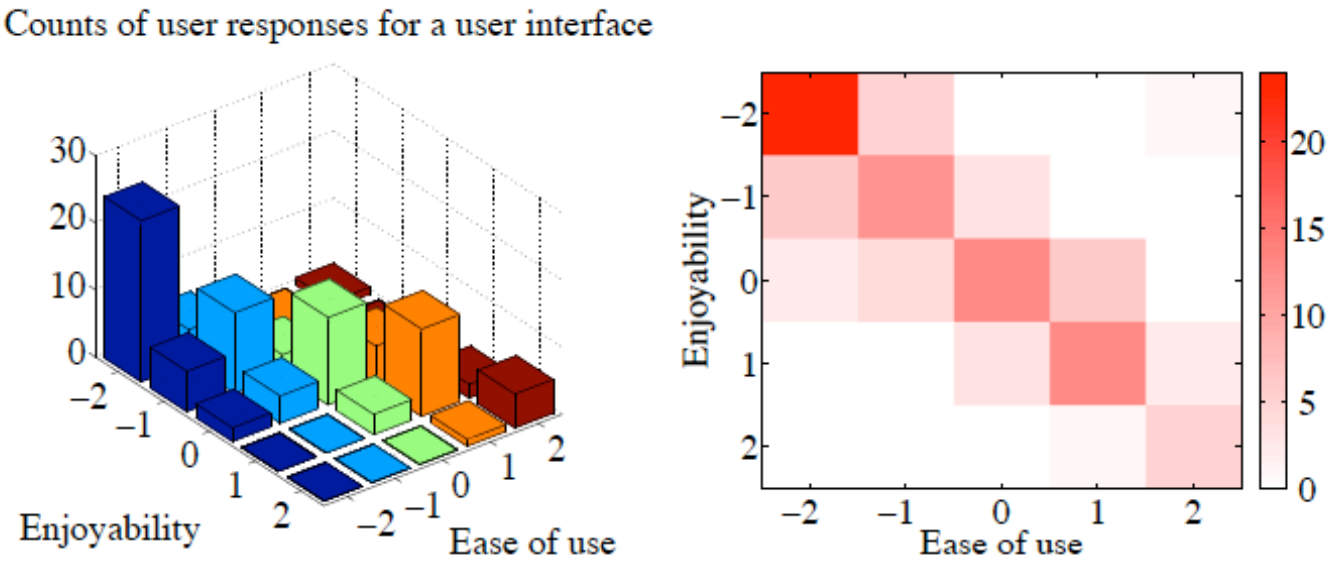
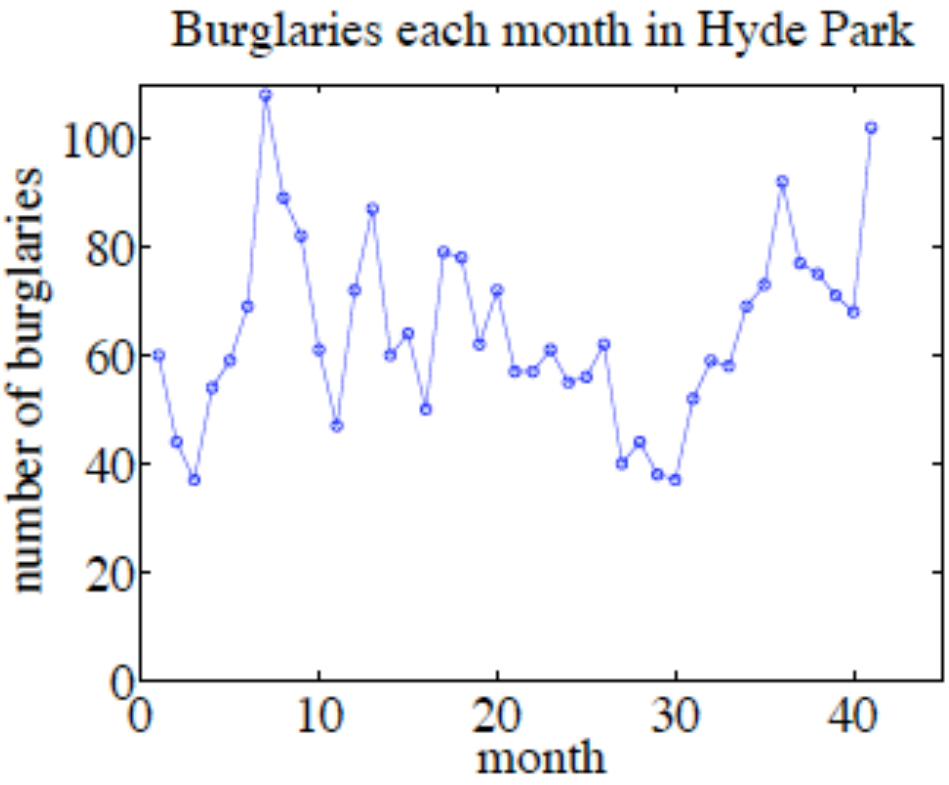
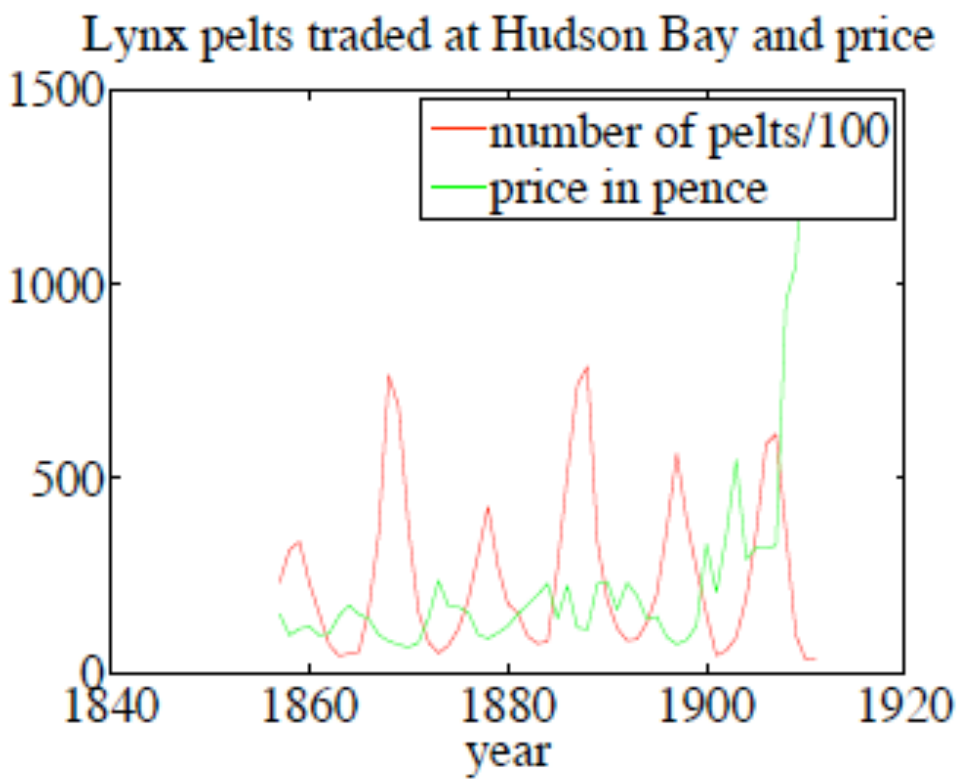


FIGURE 2.6: *On the* **left**, *a 3D bar chart of the data. The height of each bar is given by the number of users in each cell. This figure immediately reveals that users who found the interface hard to use did not like using it either. However, some of the bars at the back are hidden, so some structure might be hard to infer. On the* **right**, *a heat map of this data. Again, this figure immediately reveals that users who found the interface hard to use did not like using it either. It's more apparent that everyone disliked the interface, though, and it's clear that there is no important hidden structure.*

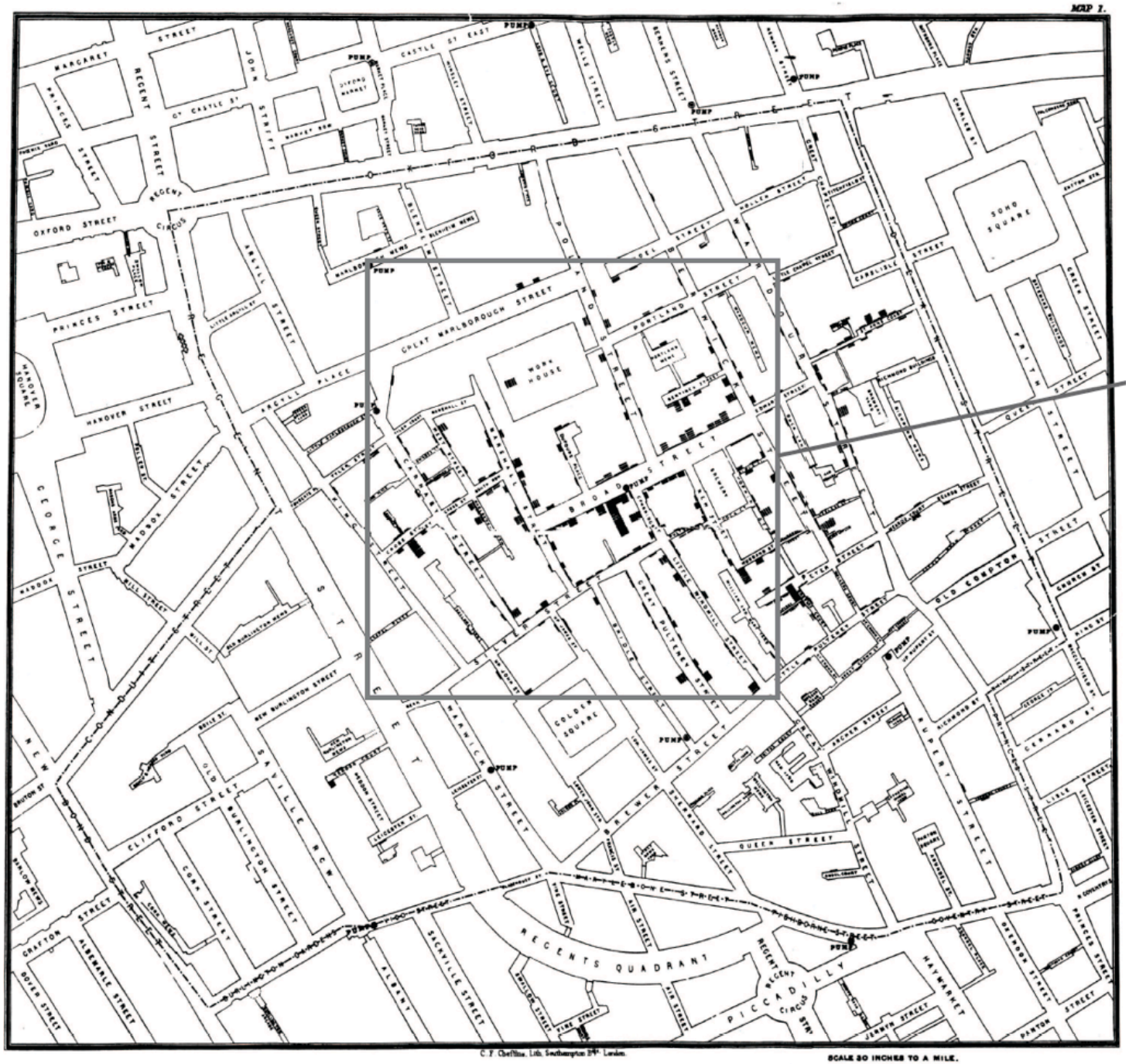
Series


Burglaries each month in Hyde Park
number of burglaries
0
20
40
60
80
100
0
10
20
30
40
month


Lynx pelts traded at Hudson Bay and price
number of pelts/100
price in pence
0
500
1000
1500
1840
1860
1880
1900
1920
year

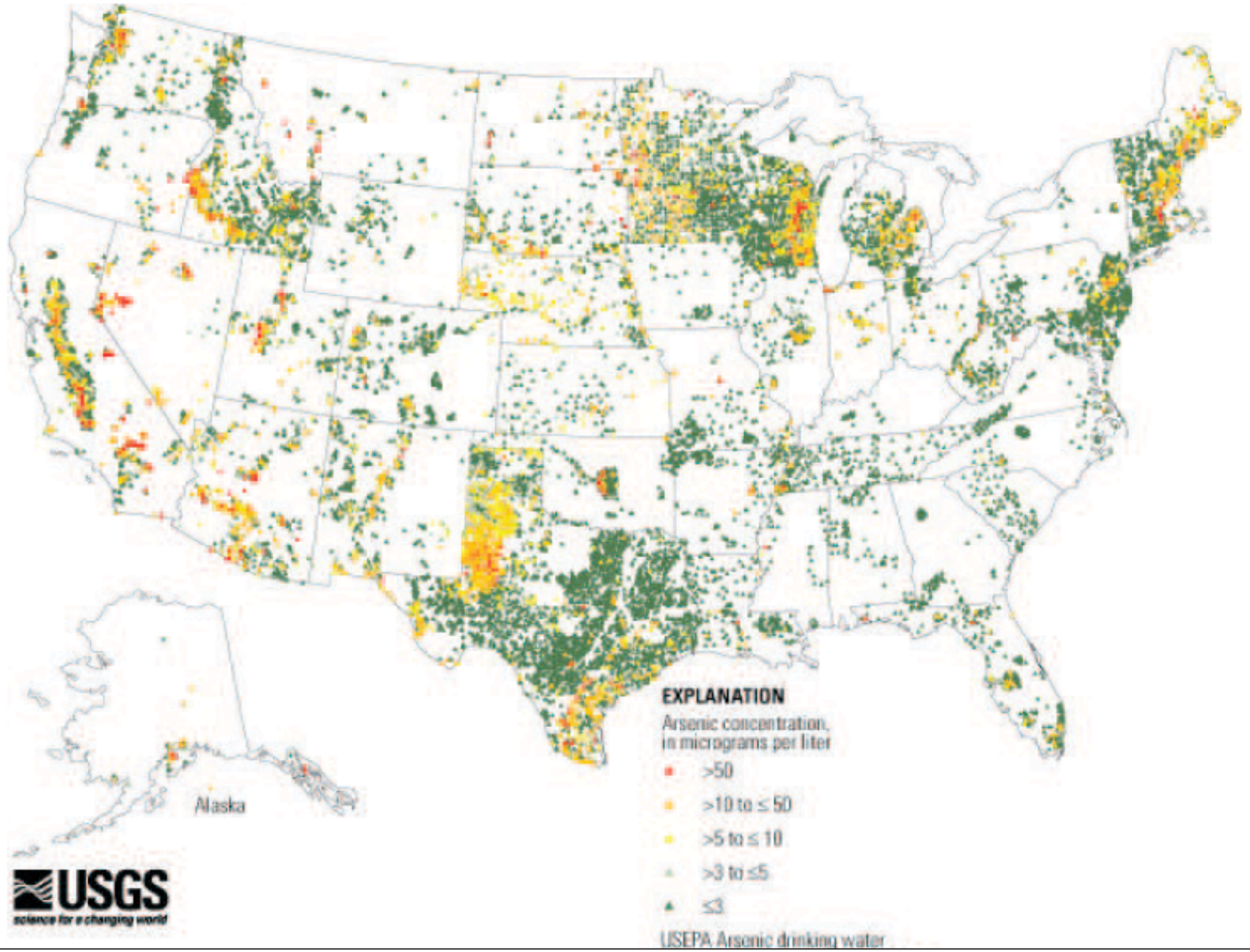
Scatter plots

- Plot a marker at a location where there is a datapoint
- Simplest case - geographic

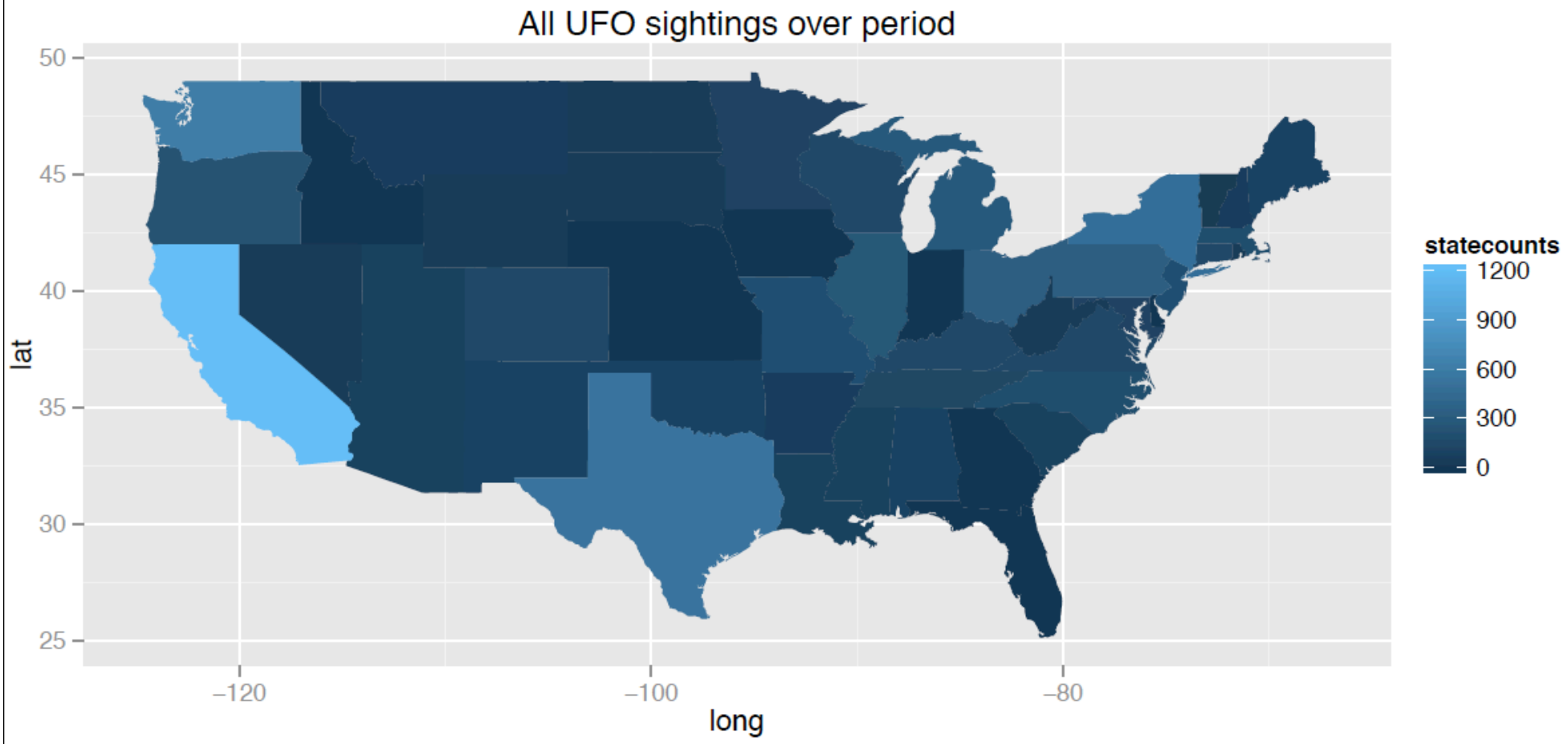

MAP 1.
REGENT STREET
OXFORD STREET
PRINCES STREET
HANOVER STREET
GEORGE STREET
MADDOX STREET
CLIFFORD STREET
SAVILLE ROW
NEW BURLINGTON STREET
CORK STREET
BURLINGTON STREET
BOYLE ST.
BRUTON ST.
CRAFTON STREET
ALBEMARLE STREET
BOYER STREET
ARGYLL PLACE
GREAT MARLBOROUGH STREET
LITTLE ARGYLL ST.
MARLBOROUGH MEWS
HANOVER SQUARE
WORK HOUSE
BROAD STREET
POLAND STREET
PORTLAND STREET
GREAT PULTENEY STREET
LOWER JAMES STREET
BEAK STREET
WARWICK STREET
GOLDEN SQUARE
QUEEN STREET
BREWER STREET
RUPERT STREET
OLD COMPTON STREET
SOHO SQUARE
CHARLES ST.
GREEK STREET
FRITH STREET
QUEEN STREET
RICHMOND ST.
LISLE STREET
LEICESTER STREET
PRINCES STREET
SECKVILLE STREET
ALBANY
REGENTS QUADRANT
REGENT CIRCUS
PICCADILLY
HAYMARKET
JERMYN STREET
PANTON STREET
CASTLE ST EAST
OXFORD MARKET
NEWMAN STREET
RATHBONE STREET
PUMP
C. F. Cheffins, Lith. Southampton Bldgs. London.
SCALE 30 INCHES TO A MILE.


PUMP
GREAT MARLBOROUGH STREET
WORK HOUSE
BROAD STREET
PUMP
GREAT PULTENEY STREET

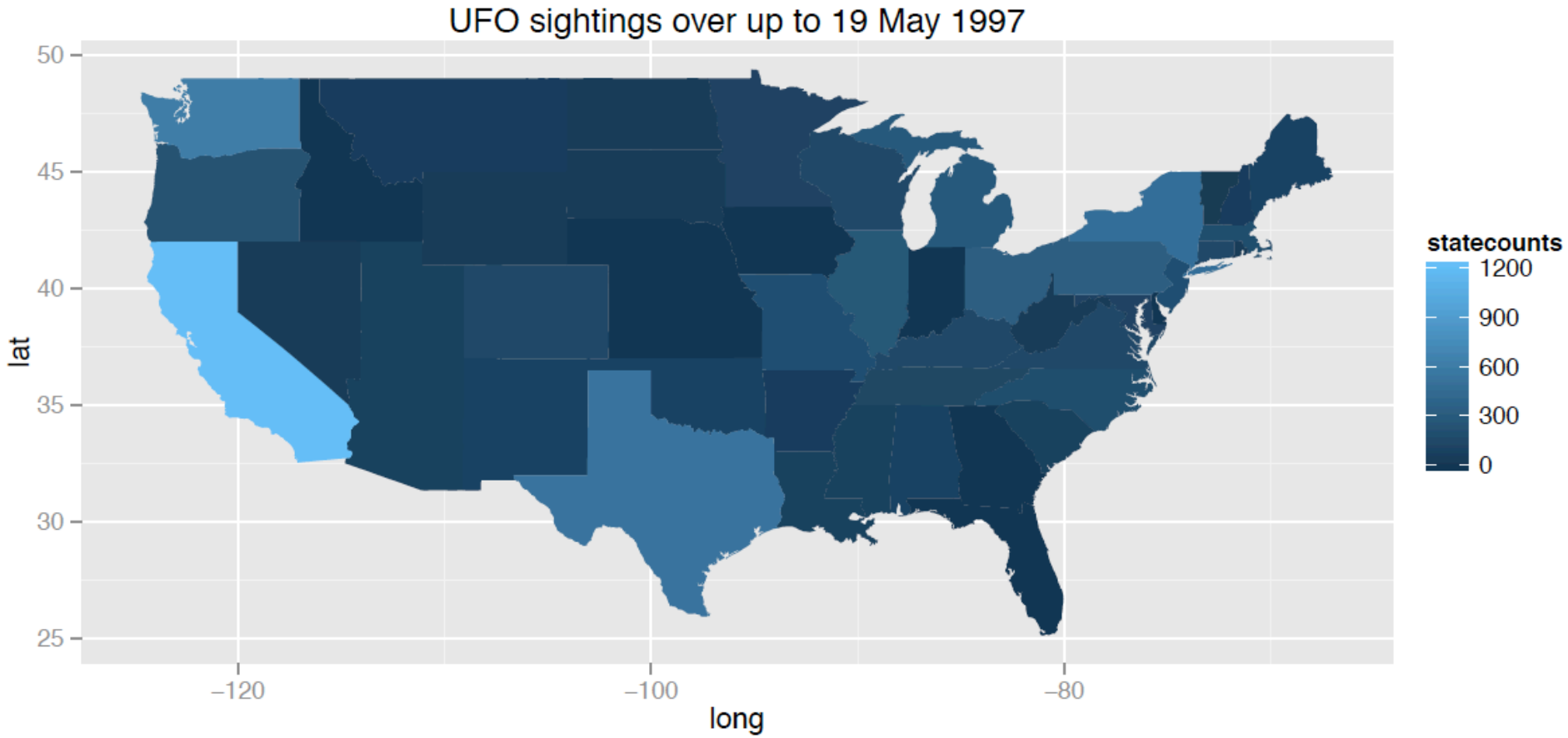
Arsenic in well water



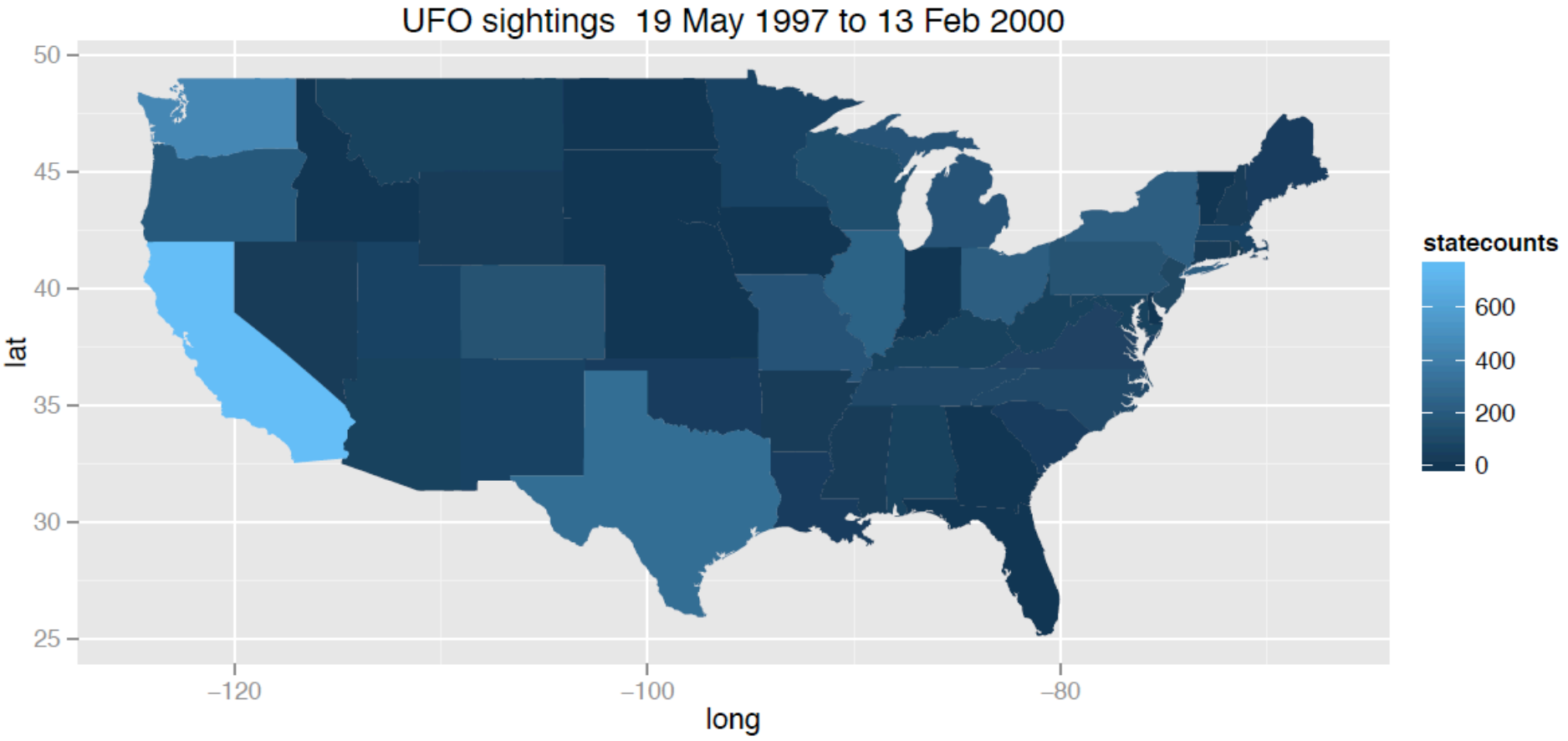
UFO sightings by state



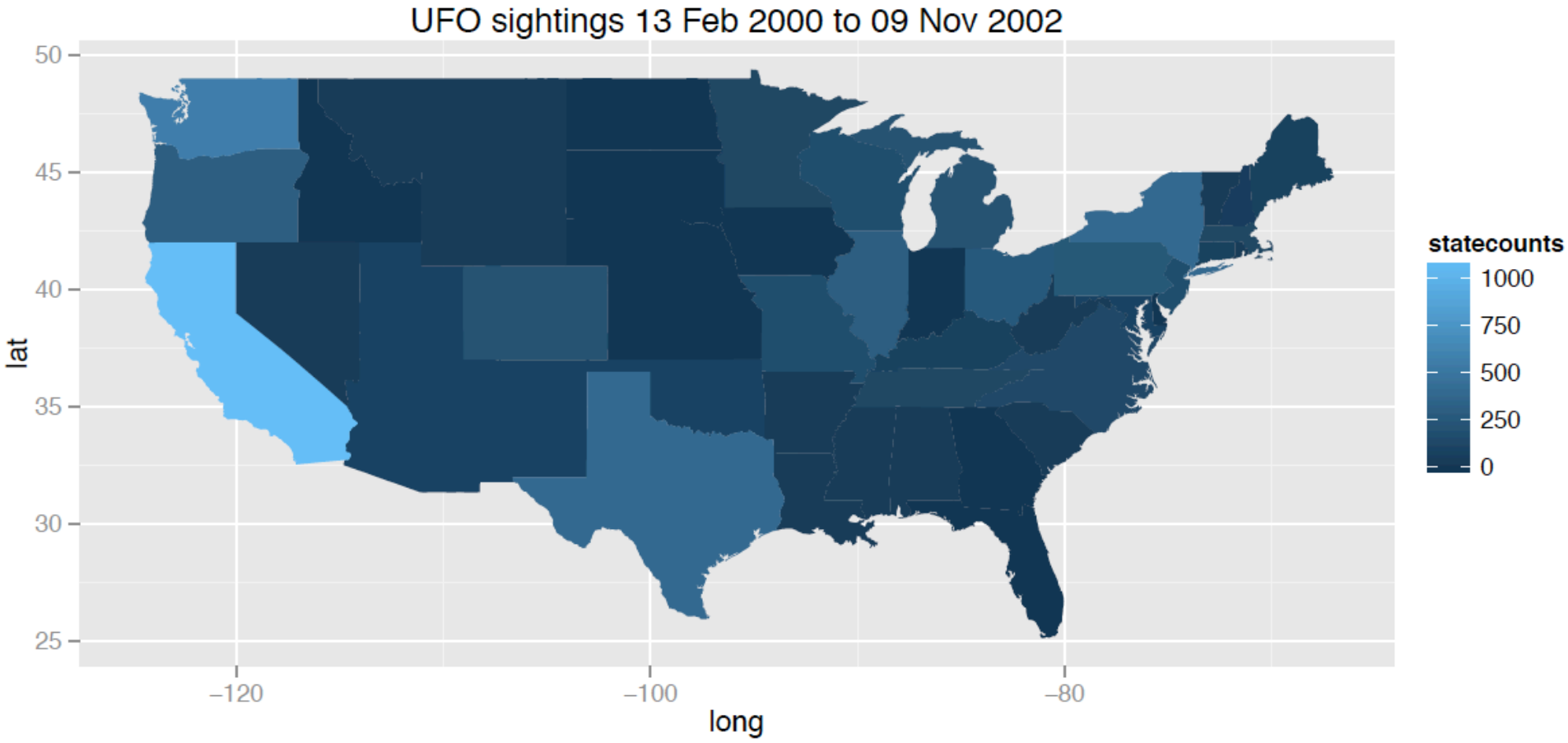
UFO's by interval



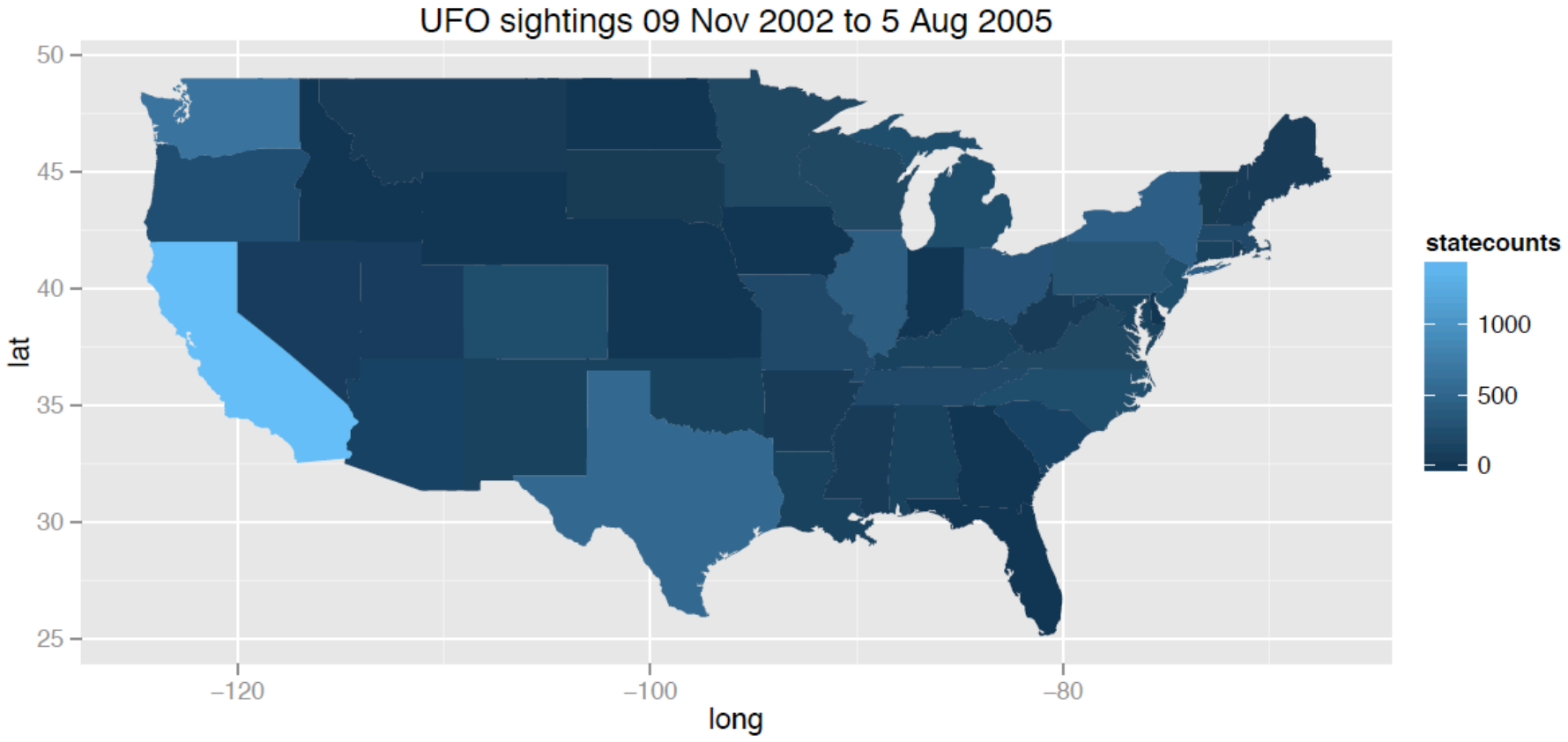
UFO's by interval



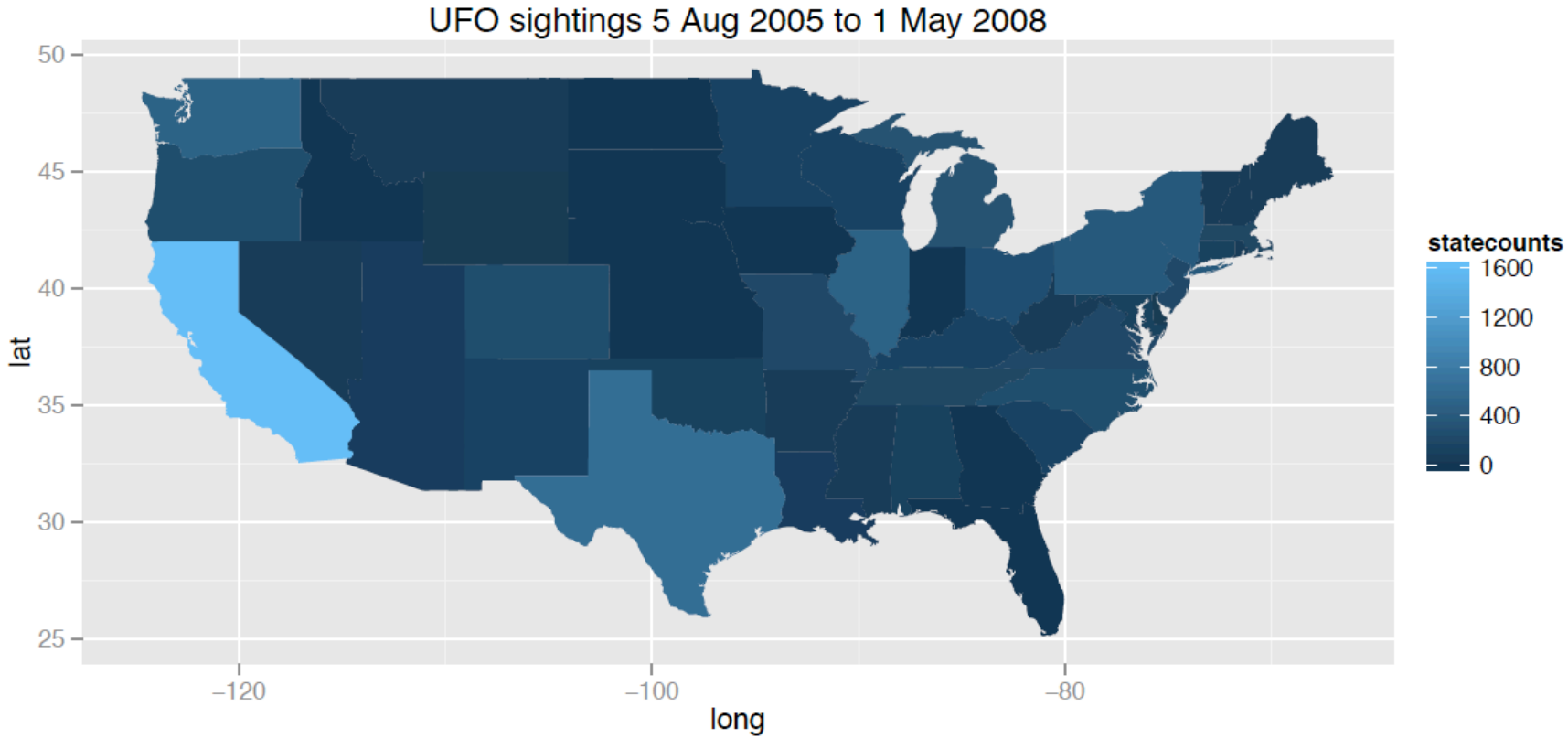
UFO's by interval



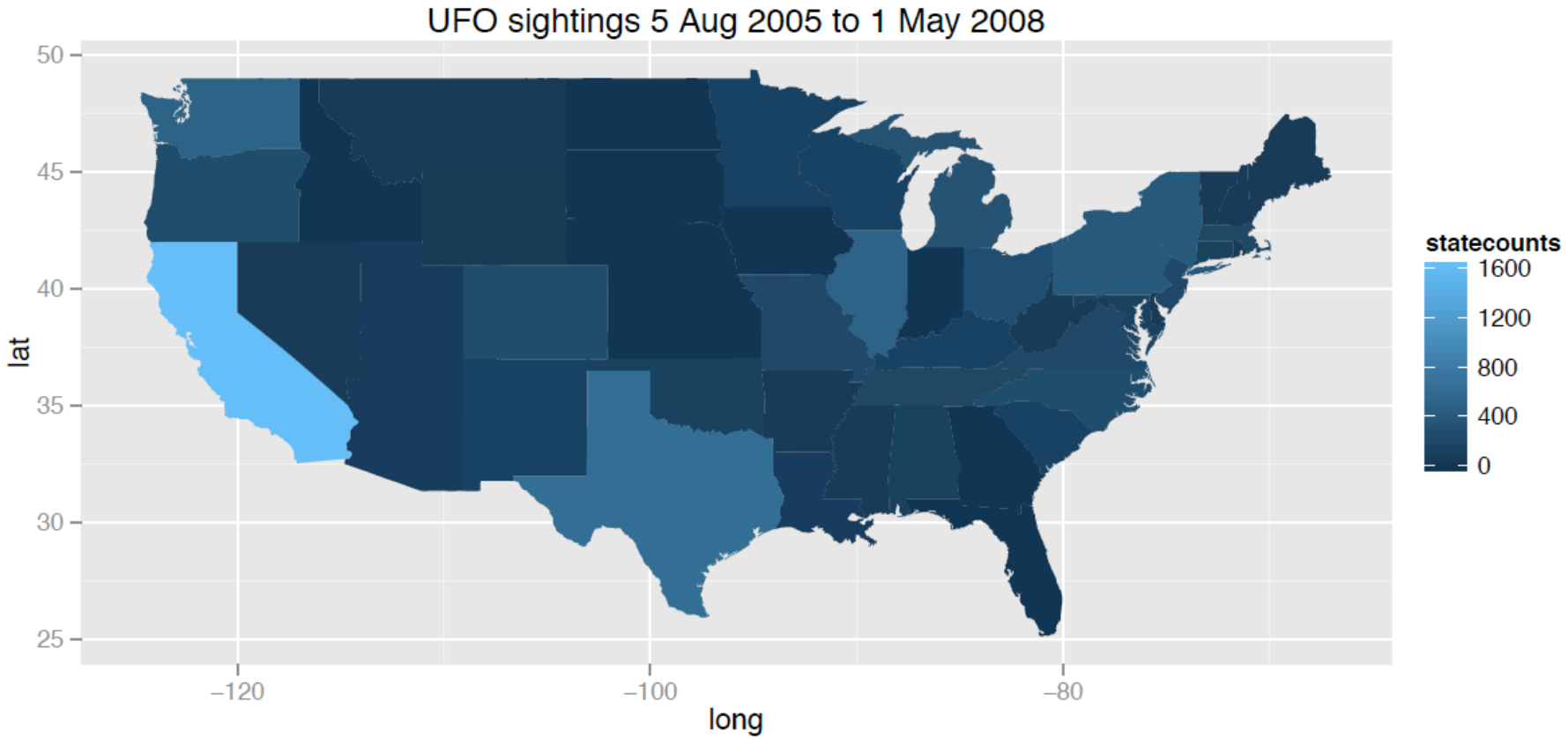
UFO's by interval



UFO's by interval



UFO's by interval

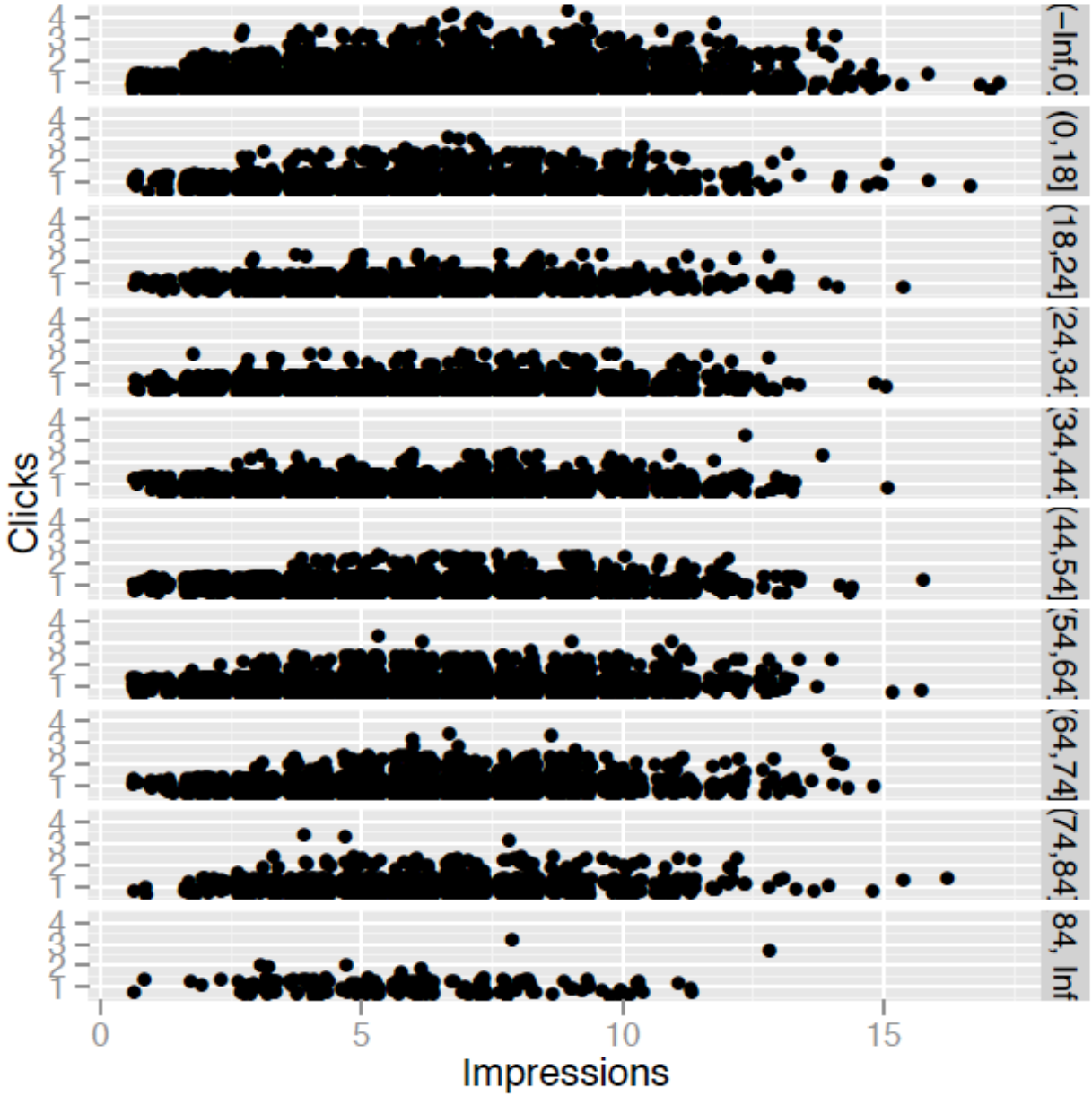


Interesting analogy

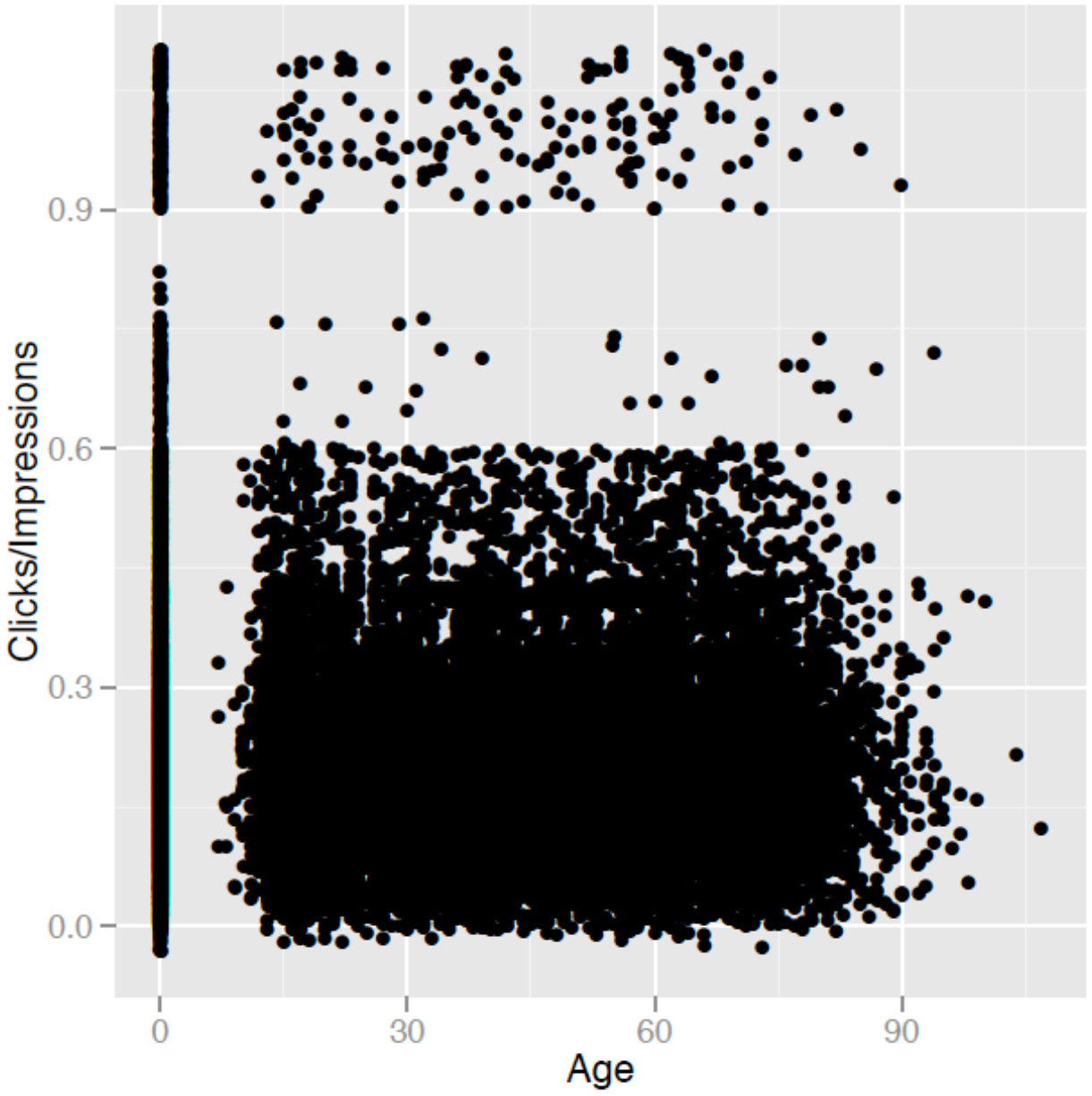
- Blackett's reasoning about submarine sightings in WWII
  - can estimate probability of sightings
  - lead to significantly improved sighting rates, aircraft painting and lighting strategies (see Korner, "The pleasures of counting" or good histories)

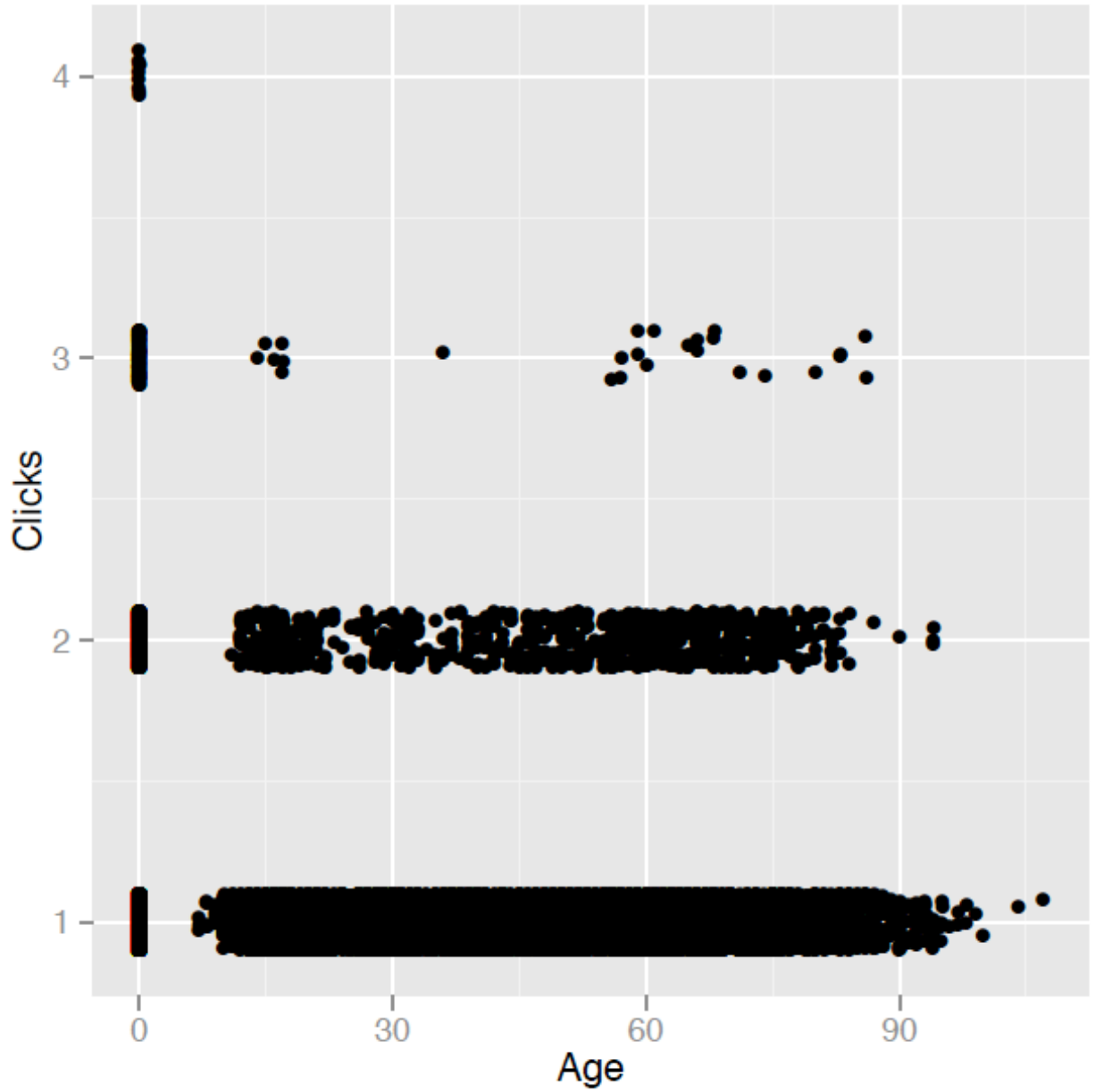
NYT data - remarks

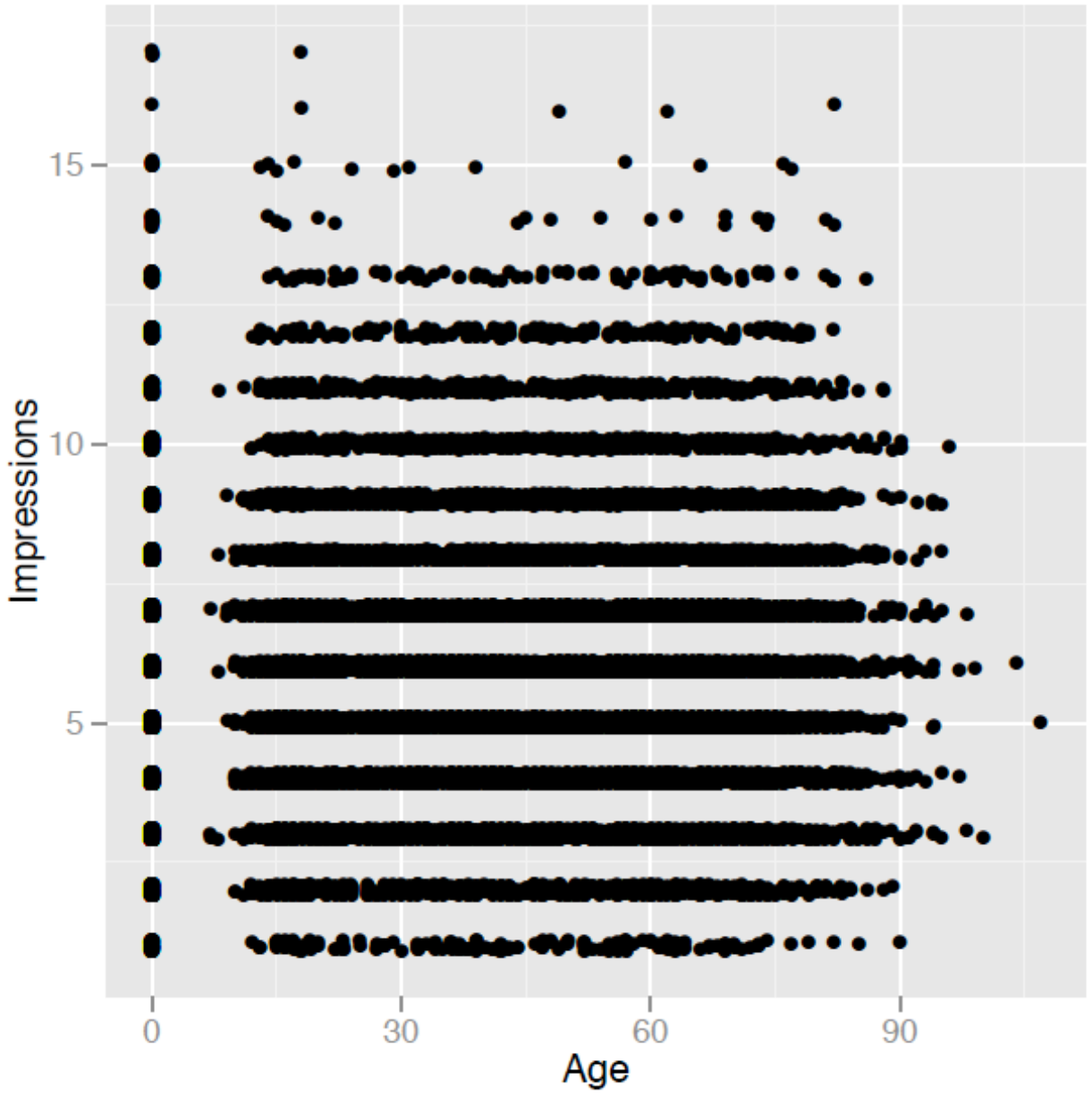
- Many data points lying on top of each other
  - scatter plot can be deceptive
  - jitter the points (move by a small random amount)


Clicks
Impressions
0
5
10
15
(−Inf,0]
(0,18]
(18,24]
(24,34]
(34,44]
(44,54]
(54,64]
(64,74]
(74,84]
(84, Inf]

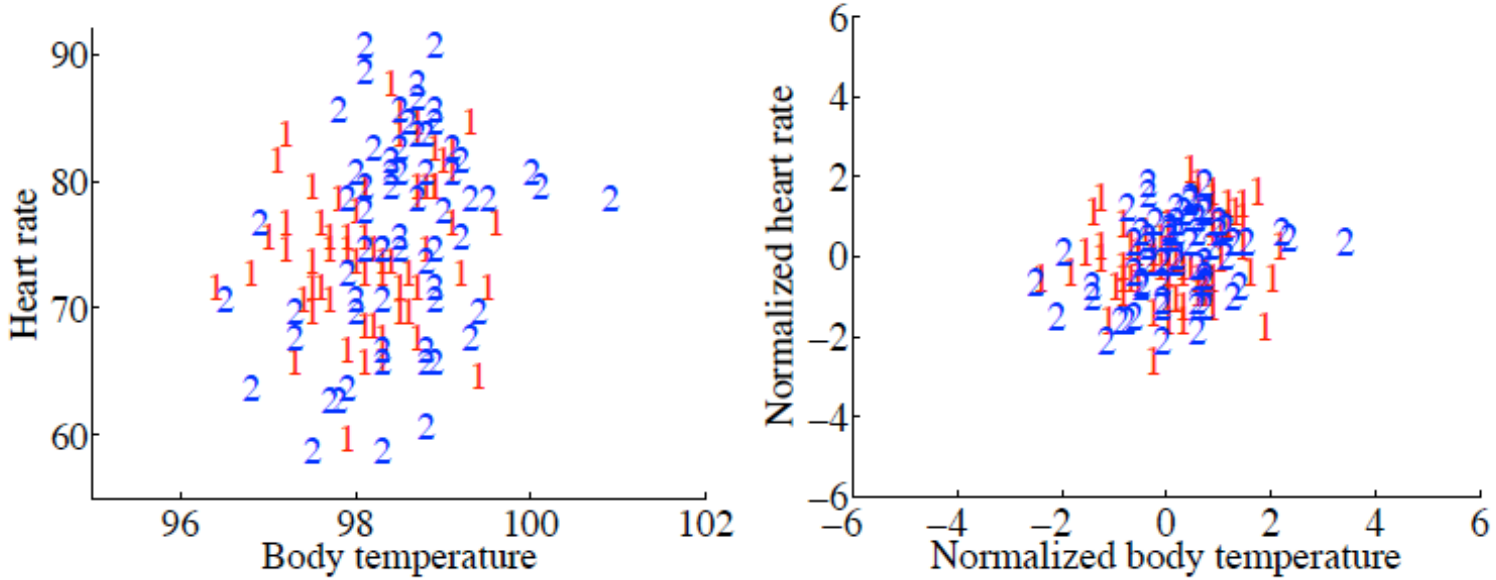
Age	Gender	Impressions	Clicks	Signed_In	agecat	impcat
Min. : 0.00	Min. :0.000	Min. : 0.000	Min. :0.00000	Min. :0.0000	(-Inf,0]:137106	(-Inf,0]: 3066
1st Qu.: 0.00	1st Qu.:0.000	1st Qu.: 3.000	1st Qu.:0.00000	1st Qu.:0.0000	(34,44] : 70860	(0,1] : 15483
Median : 31.00	Median :0.000	Median : 5.000	Median :0.00000	Median :1.0000	(44,54] : 64288	(1,2] : 38433
Mean : 29.48	Mean :0.367	Mean : 5.007	Mean :0.09259	Mean :0.7009	(24,34] : 58174	(2,3] : 64121
3rd Qu.: 48.00	3rd Qu.:1.000	3rd Qu.: 6.000	3rd Qu.:0.00000	3rd Qu.:1.0000	(54,64] : 44738	(3,4] : 80303
Max. :108.00	Max. :1.000	Max. :20.000	Max. :4.00000	Max. :1.0000	(18,24] : 35270	(4,5] : 80477
					(Other) : 48005	(5, Inf]:176558


Clicks/Impressions
0.9
0.6
0.3
0.0
0
30
60
90
Age

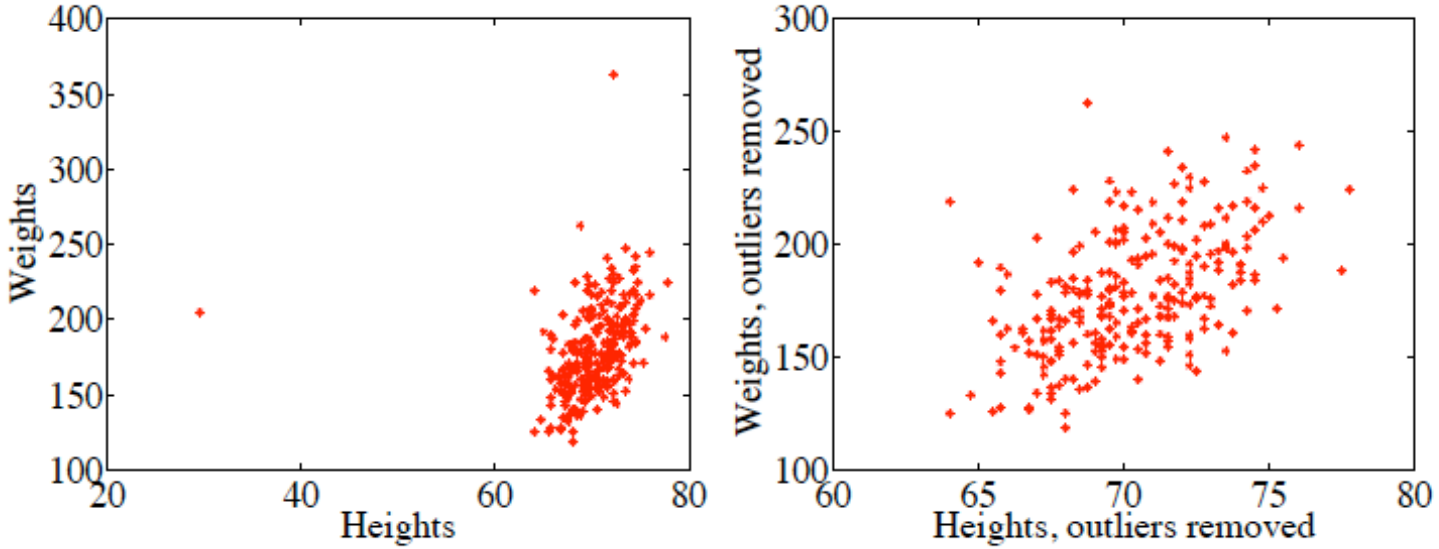

4
3
2
1
Clicks
0
30
60
90
Age


Impressions
15
10
5
0
30
60
90
Age

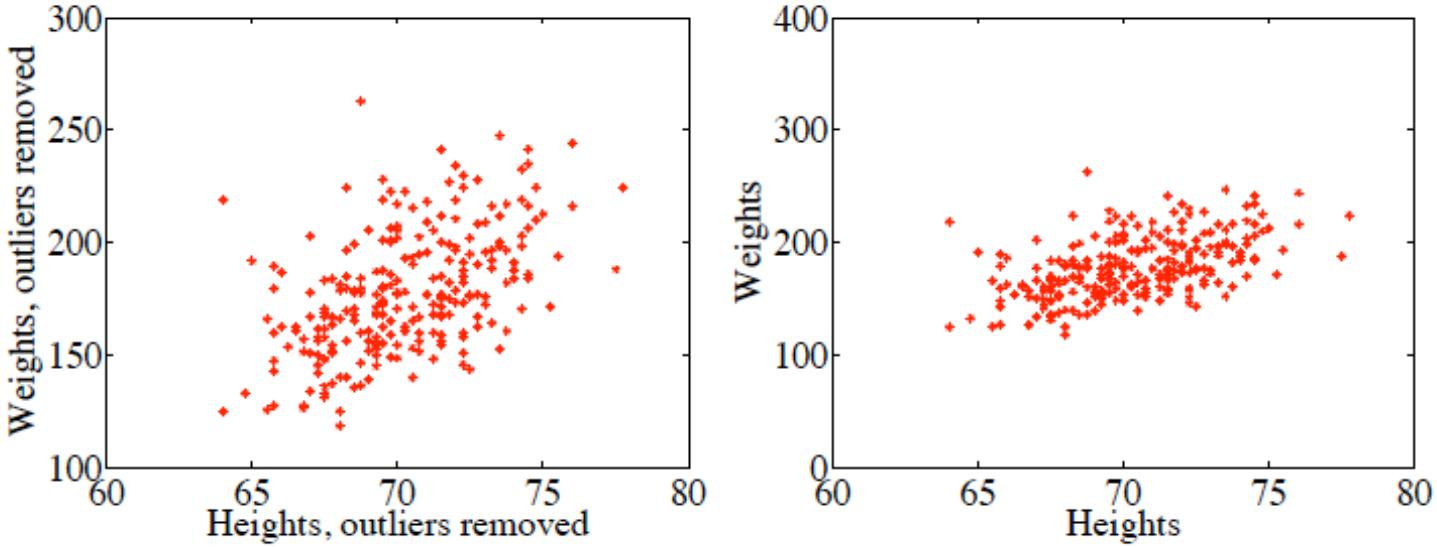
Scale is an issue



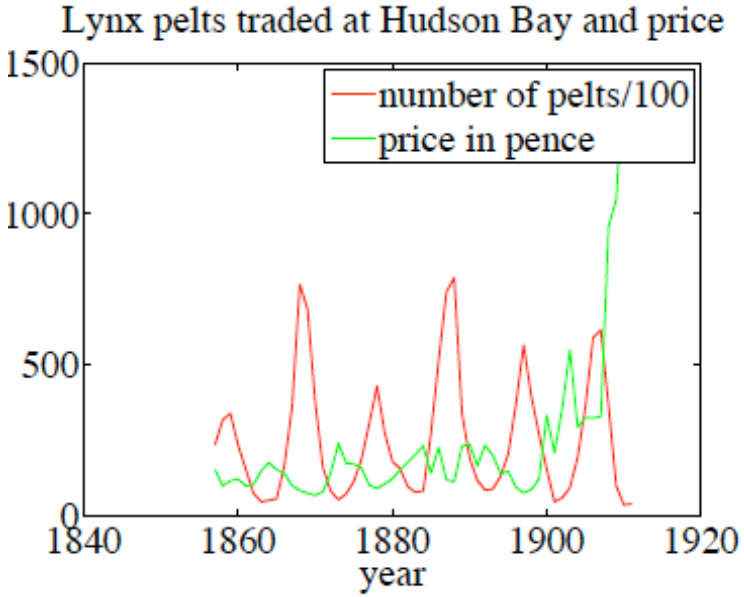
Outliers can set scale

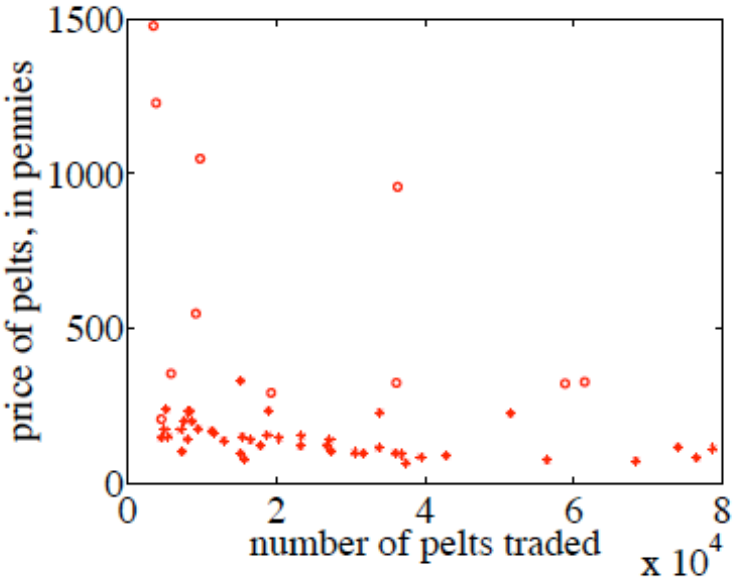


But scale is really a problem



Lynx pelts


Lynx pelts traded at Hudson Bay and price
number of pelts/100
price in pence
1500
1000
500
0
1840
1860
1880
1900
1920
year


price of pelts, in pennies
1500
1000
500
0
0
2
4
6
8
number of pelts traded
x 10^4

Data example

- Housing sales in NYC boroughs
- https://github.com/oreillymedia/doing_data_science
- Question: Look at real estate sales - what's going on?

Summary Statistics - mean

Definition: 3.1 *Mean*

Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. Their mean is

$$\text{mean}\left(\{x\}\right) = \frac{1}{N}\sum_{i=1}^{i=N} x_i.$$

The average

The best estimate of the value of a new datapoint in the absence of any other information about it

Summary statistics - Standard deviation

Definition: 3.2 *Standard deviation*

Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. The standard deviation of this dataset is is:

$$\text{std}(x_i) = \sqrt{\frac{1}{N}\sum_{i-1}^{i=N}(x_i - \text{mean}(\{x\}))^2} = \sqrt{\text{mean}(\{(x_i - \text{mean}(\{x\}))^2\})}.$$

Think of this as a scale

Average distance from mean

Important math properties in notes

Standard deviation

Proposition: *Assume we have a dataset* $\{x\}$ *of* N *data items,* $x_1, \ldots, x_N$. *Assume the standard deviation of this dataset is* $\mathsf{std}(x) = \sigma$. *Then there are at most* $\frac{1}{k^2}$ *data points lying* k *or more standard deviations away from the mean.*

= there are not many points many standard deviations away from the mean

Proposition: $(\mathsf{std}(x))^2 \leq \max_i (x_i - \mathsf{mean}(\{x\}))^2$.

= there is at least one point at least one standard deviation away from the mean

Standard coordinates

Definition: 3.8 *Standard coordinates*

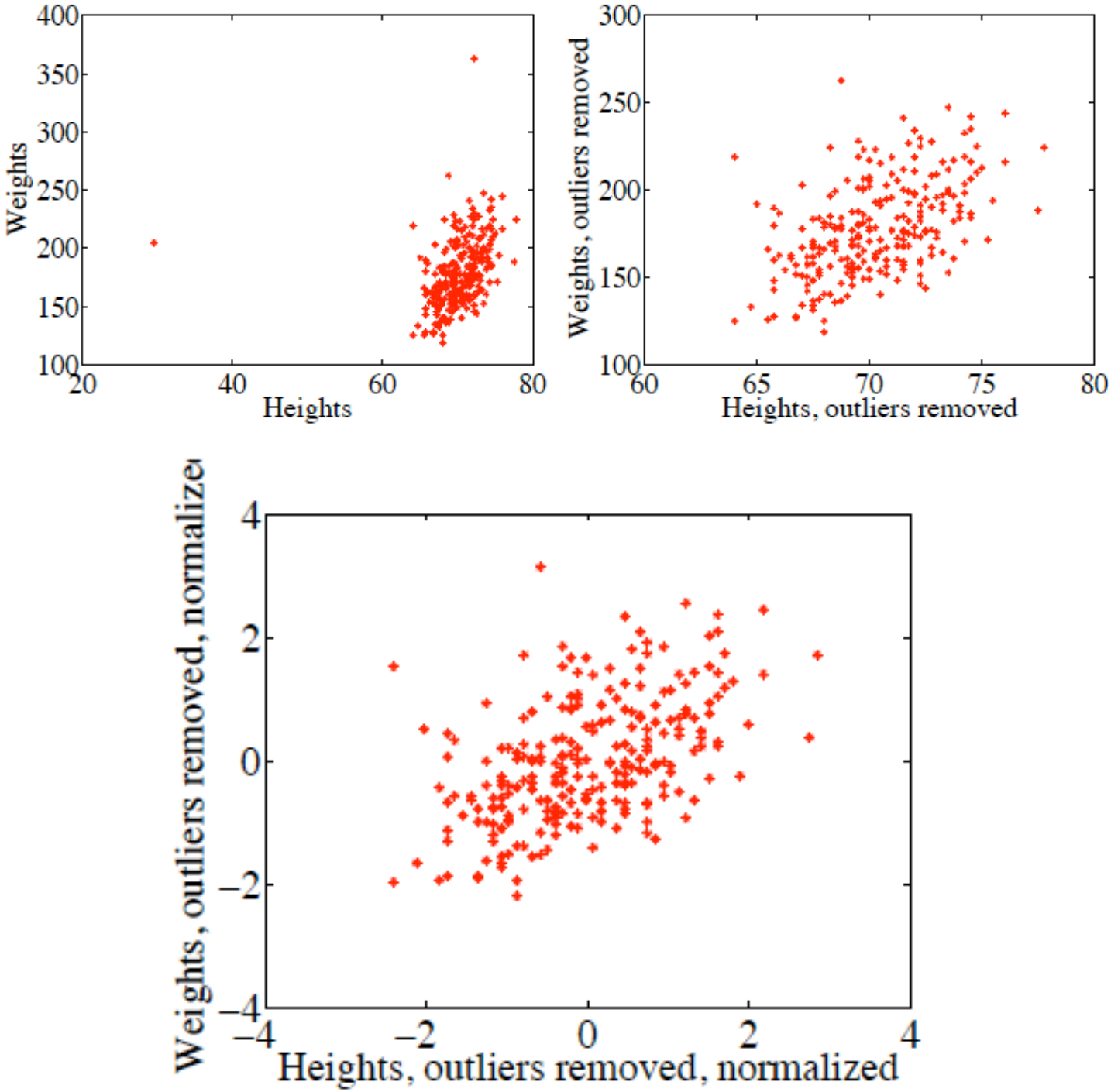
Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. We represent these data items in standard coordinates by computing

$$\hat{x}_i = \frac{(x_i - \text{mean}(\{x\}))}{\text{std}(x)}.$$

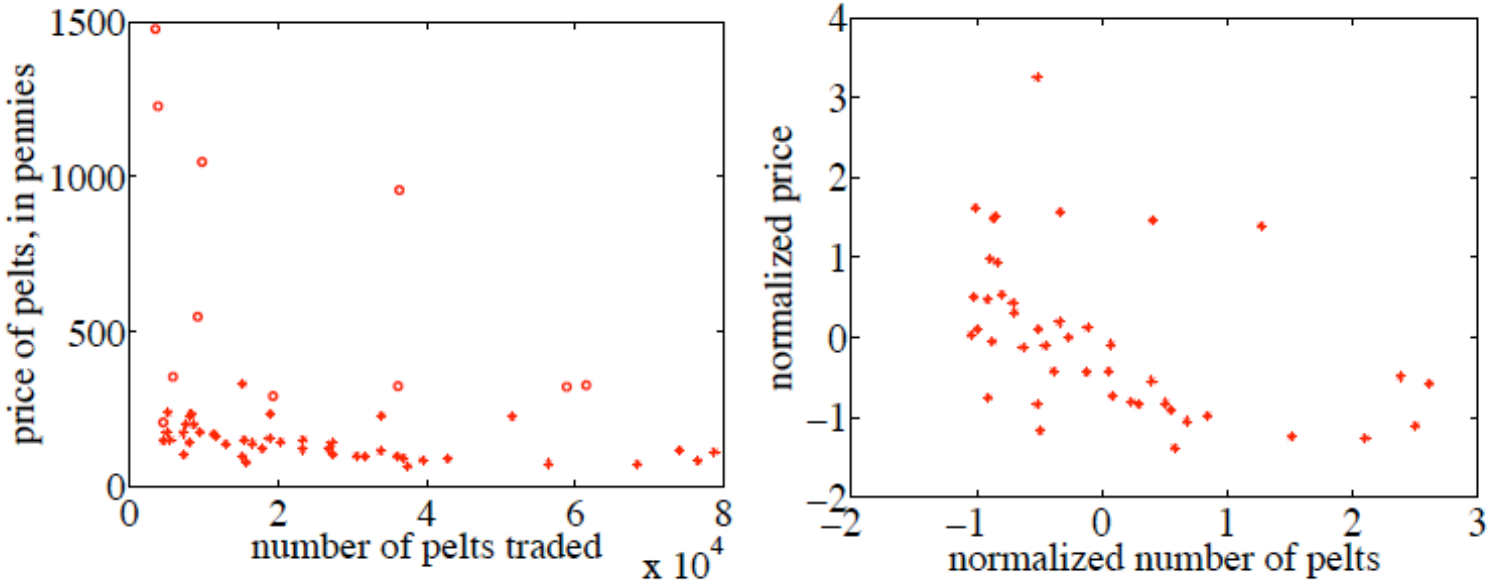
We write $\{\hat{x}\}$ for a dataset that happens to be in standard coordinates.

Suppressing scale effects

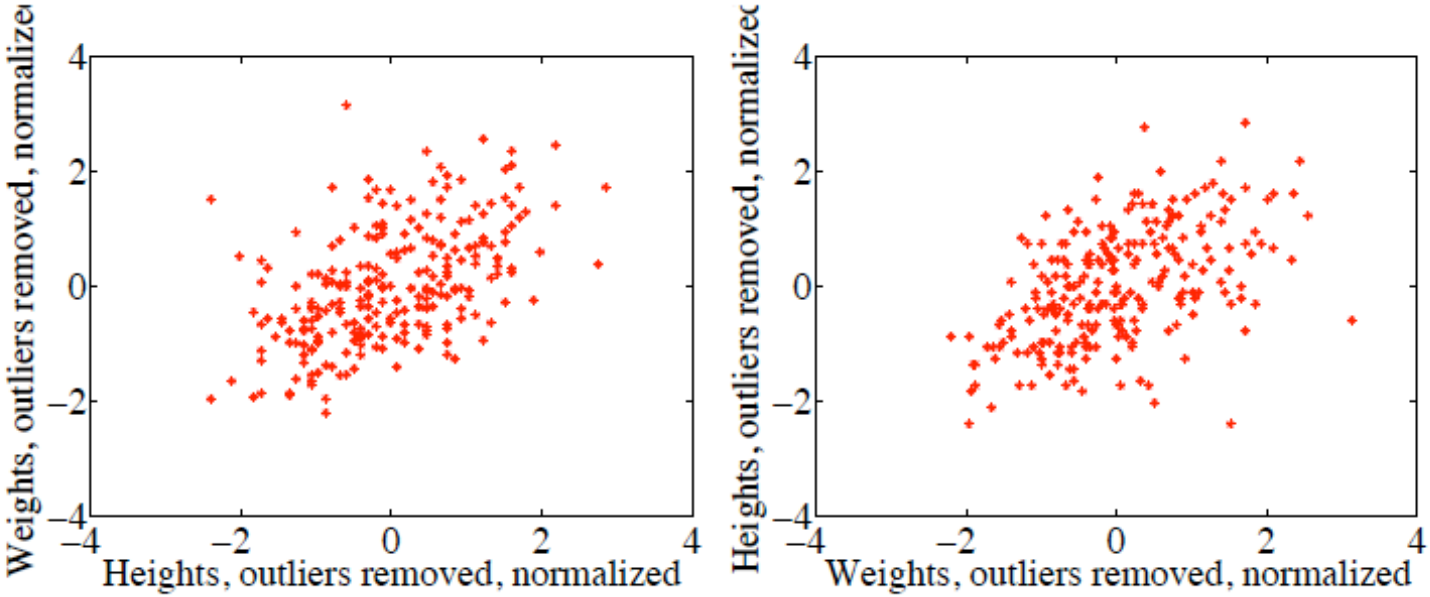
- Do scatter plots in standard coordinates for x, y

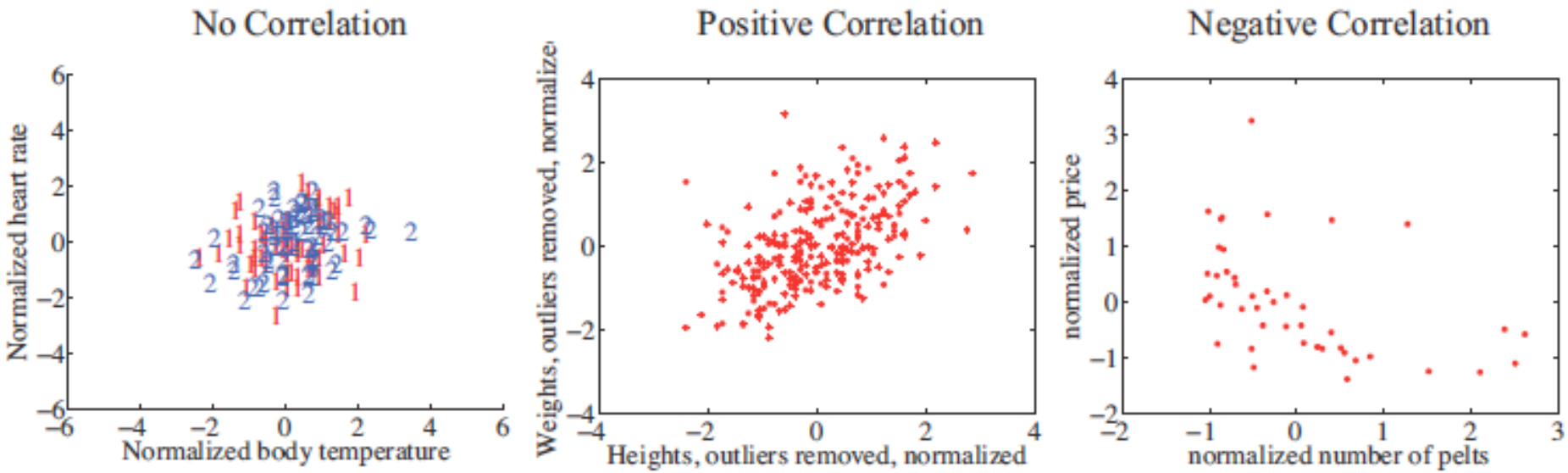

Weights
Heights
Weights, outliers removed
Heights, outliers removed
Weights, outliers removed, normalized
Heights, outliers removed, normalized

Lynx, normalized

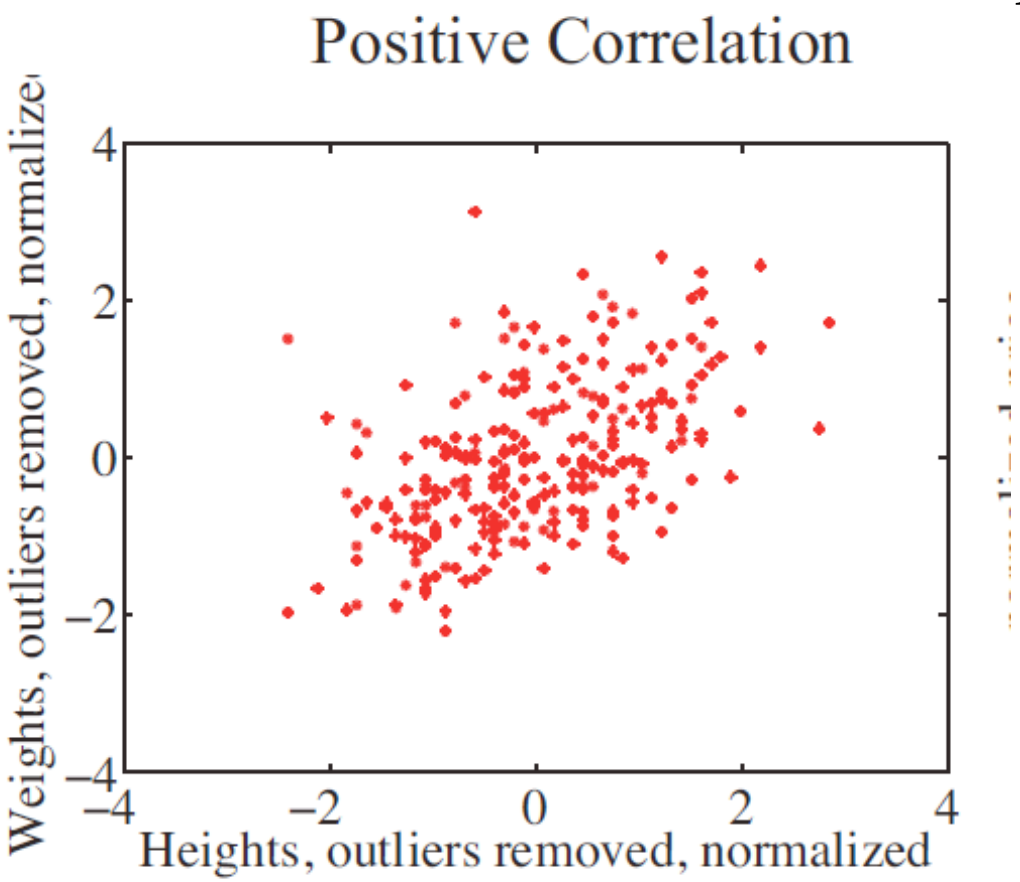


x, y don't really matter



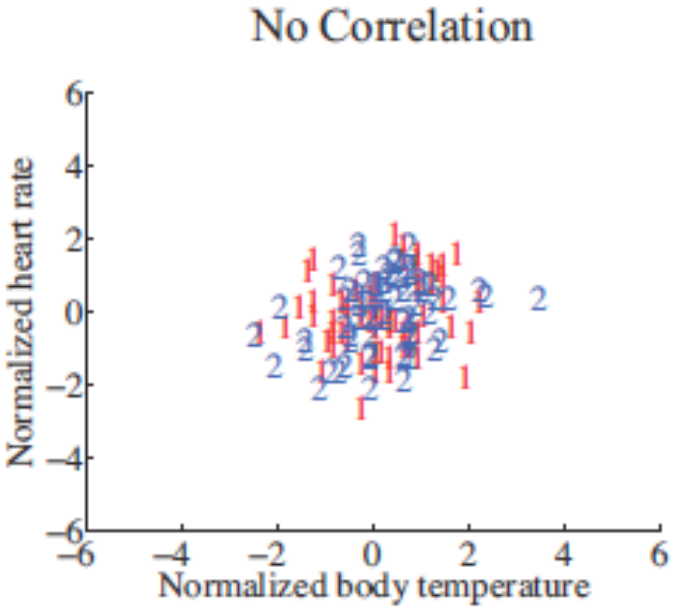

No Correlation
Normalized heart rate
Normalized body temperature
Positive Correlation
Weights, outliers removed, normalize
Heights, outliers removed, normalized
Negative Correlation
normalized price
normalized number of pelts

Positive Correlation

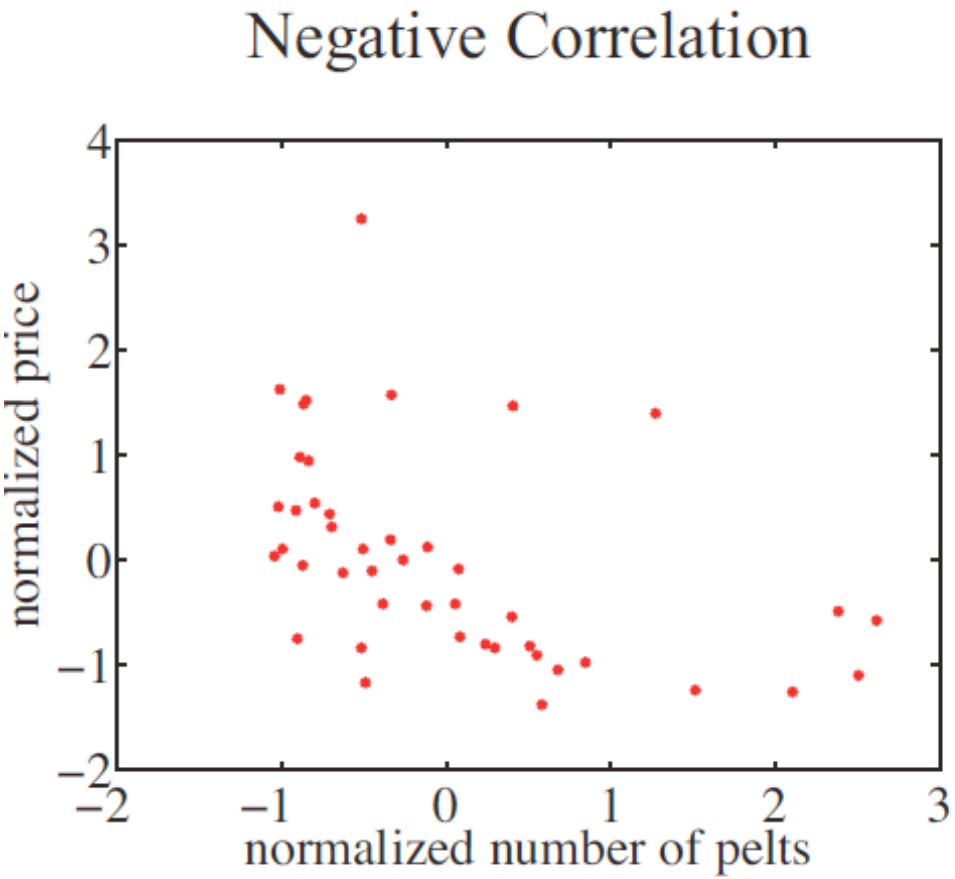


Positive correlation occurs when larger $\hat{x}$ values tend to appear with larger $\hat{y}$ values. This means that data points with with small (i.e. negative with large magnitude) $\hat{x}$ values must have small $\hat{y}$ values, otherwise the mean of $\hat{x}$ (resp. $\hat{y}$) would be too big. In turn, this means that the scatter plot should look like a "smear" of data from the bottom left of the graph to the top right. The smear might be broad or narrow, depending on some details we'll discuss below. Figure 3.11 shows

Zero Correlation



Zero correlation occurs when there is no relationship. This produces a characteristic shape in a scatter plot, but it takes a moment to understand why. If there really is no relationship, then knowing $\hat{x}$ will tell you nothing about $\hat{y}$. All we know is that $\mathsf{mean}(\{\hat{y}\}) = 0$, and $\mathsf{var}(\{\hat{y}\}) = 1$. Our value of $\hat{y}$ should have this mean and this variance, but it doesn't depend on $\hat{x}$ in any way. This is enough information to predict what the plot will look like. We know that $\mathsf{mean}(\{\hat{x}\}) = 0$ and $\mathsf{var}(\{\hat{x}\}) = 1$; so there will be many data points with $\hat{x}$ value close to zero, and few with a much larger or much smaller $\hat{x}$ value. The same applies to $\hat{y}$. Now consider the data points in a strip of $\hat{x}$ values. If this strip is far away from the origin, there will be few data points in the strip, because there aren't many big $\hat{x}$ values. If there is no relationship, we don't expect to see large or small $\hat{y}$ values in this strip, because there are few data points in the strip and because large or small $\hat{y}$ values are uncommon — we see them only if there are many data points,



Negative correlation

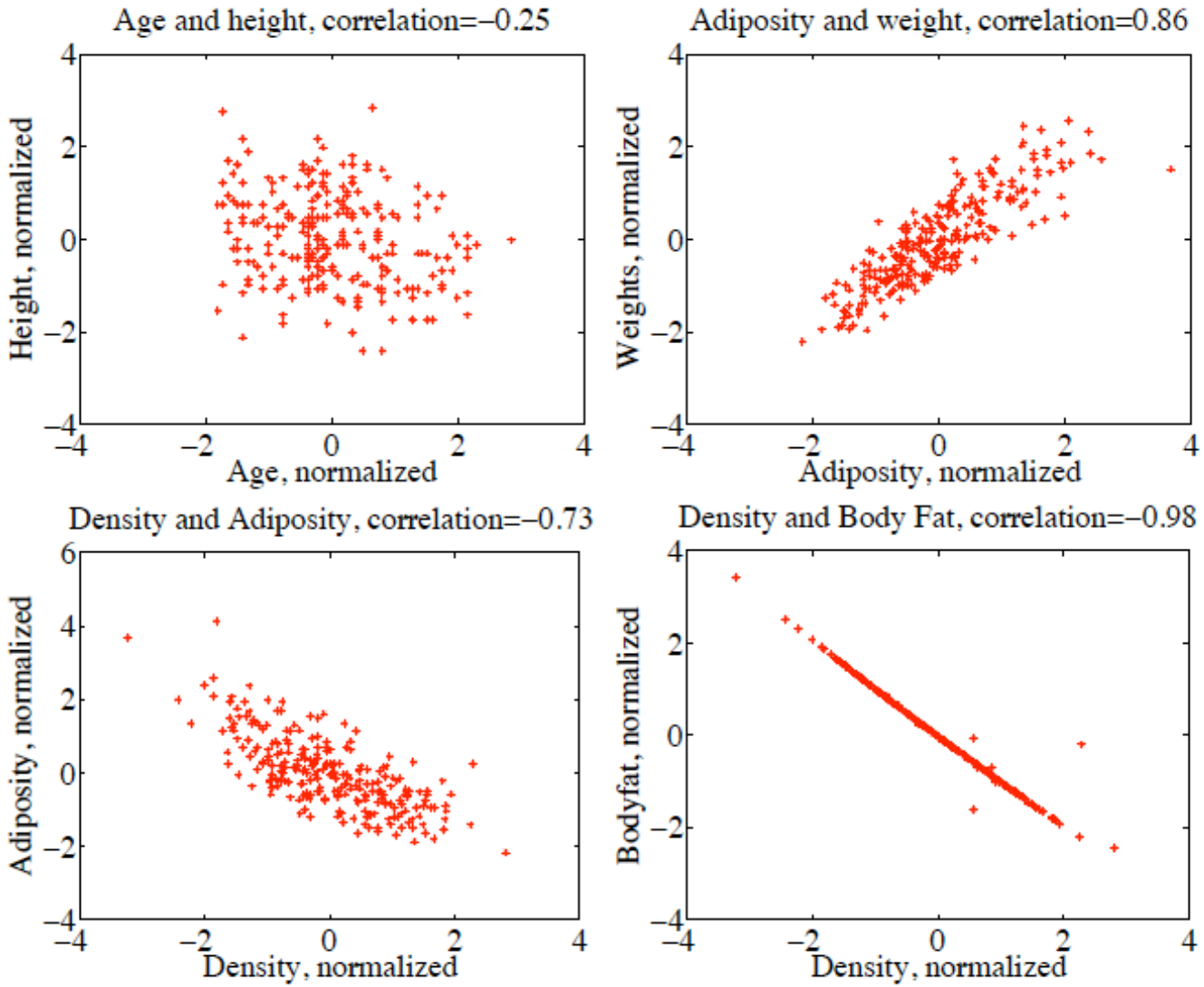
Negative correlation occurs when larger $\hat{x}$ values tend to appear with smaller $\hat{y}$ values. This means that data points with with small $\hat{x}$ values must have large $\hat{y}$ values, otherwise the mean of $\hat{x}$ (resp. $\hat{y}$) would be too big. In turn, this means that the scatter plot should look like a "smear" of data from the top left of the graph to the bottom right. The smear might be broad or narrow, depending on some details we'll discuss below. Figure 3.13 shows a normalized scatter plot of

The Correlation Coefficient

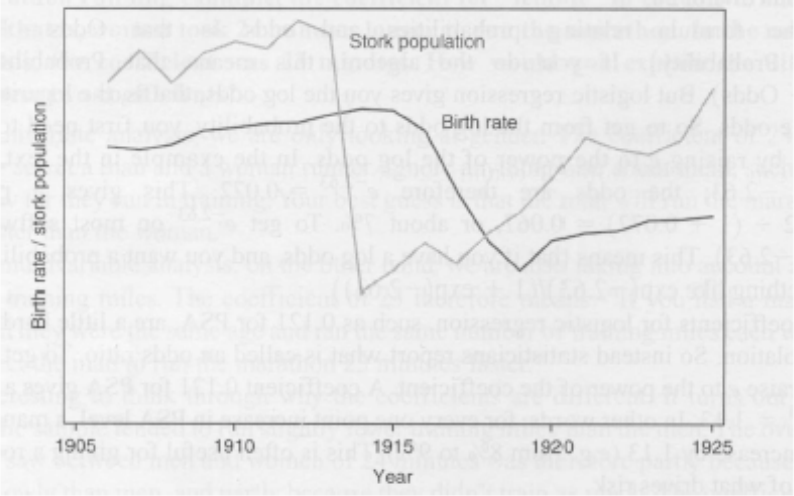
Definition: 3.11 *Correlation coefficient*

Assume we have N data items which are 2-vectors $(x_1, y_1), \ldots, (x_N, y_N)$, where $N > 1$. These could be obtained, for example, by extracting components from larger vectors. We compute the correlation coefficient by first normalizing the x and y coordinates to obtain $\hat{x}_i = \frac{(x_i - \mathsf{mean}(\{x\}))}{\mathsf{std}(x)}$, $\hat{y}_i = \frac{(y_i - \mathsf{mean}(\{y\}))}{\mathsf{std}(y)}$. The correlation coefficient is the mean value of $\hat{x}\hat{y}$, and can be computed as:

$$\mathsf{corr}\left(\{(x, y)\}\right) = \frac{\sum_i \hat{x}_i \hat{y}_i}{N}$$

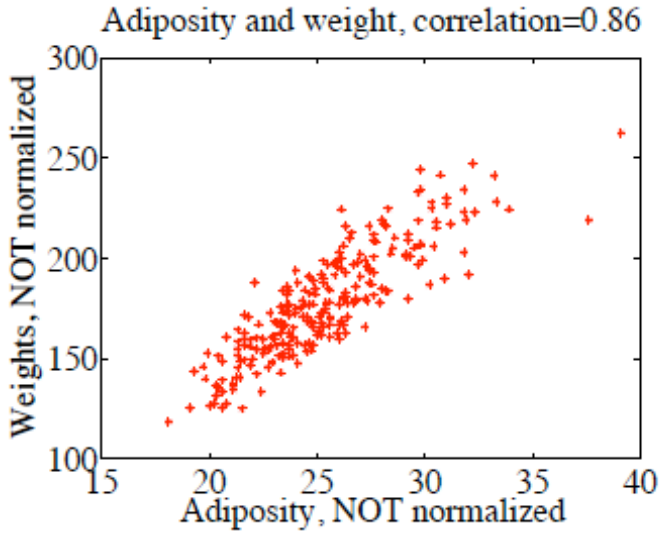

Age and height, correlation=–0.25
Height, normalized
Age, normalized
Adiposity and weight, correlation=0.86
Weights, normalized
Adiposity, normalized
Density and Adiposity, correlation=–0.73
Adiposity, normalized
Density, normalized
Density and Body Fat, correlation=–0.98
Bodyfat, normalized
Density, normalized

Correlation isn't causality



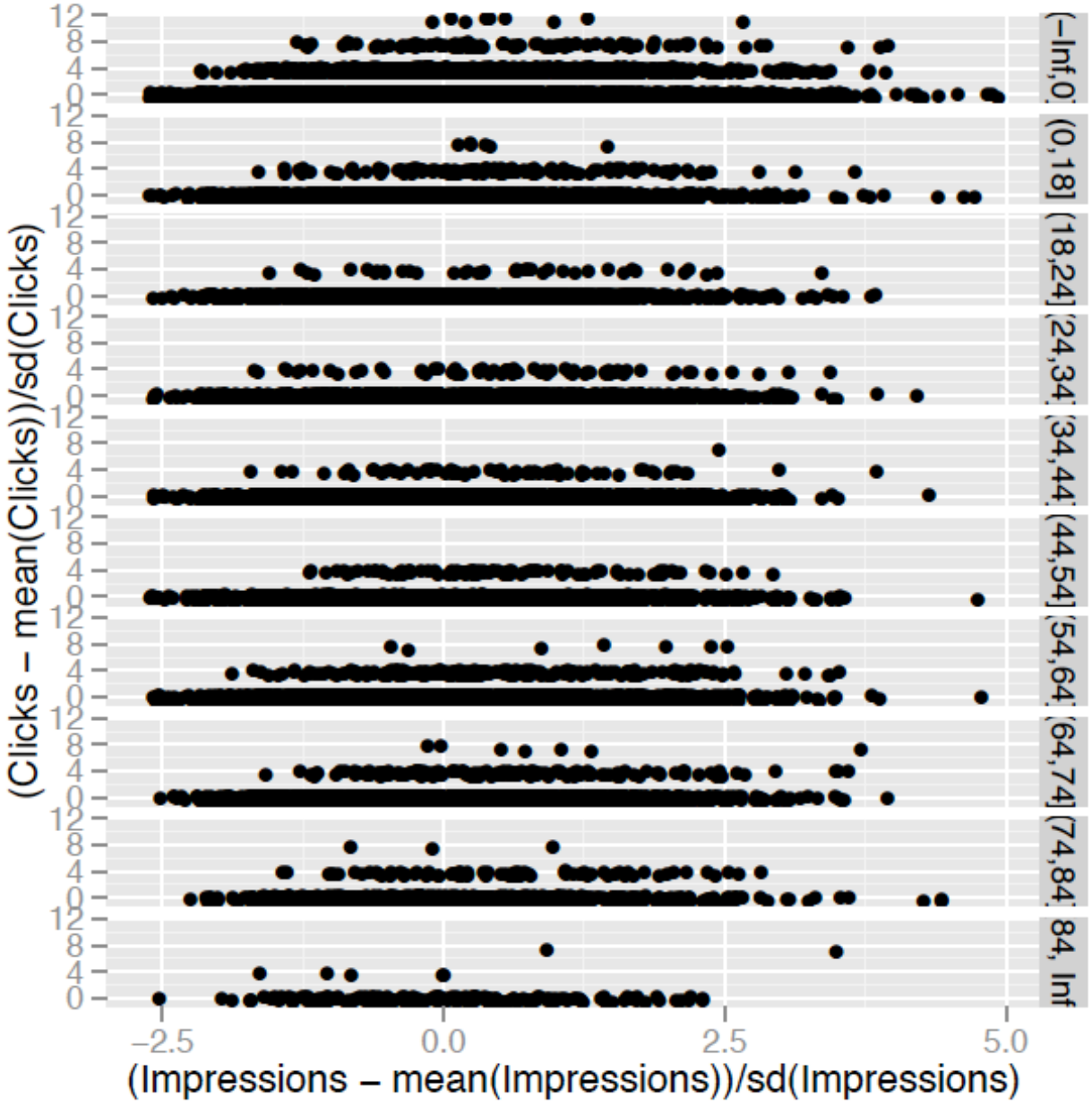
and foot size is positively correlated with reading ability, etc.

but can be used to predict

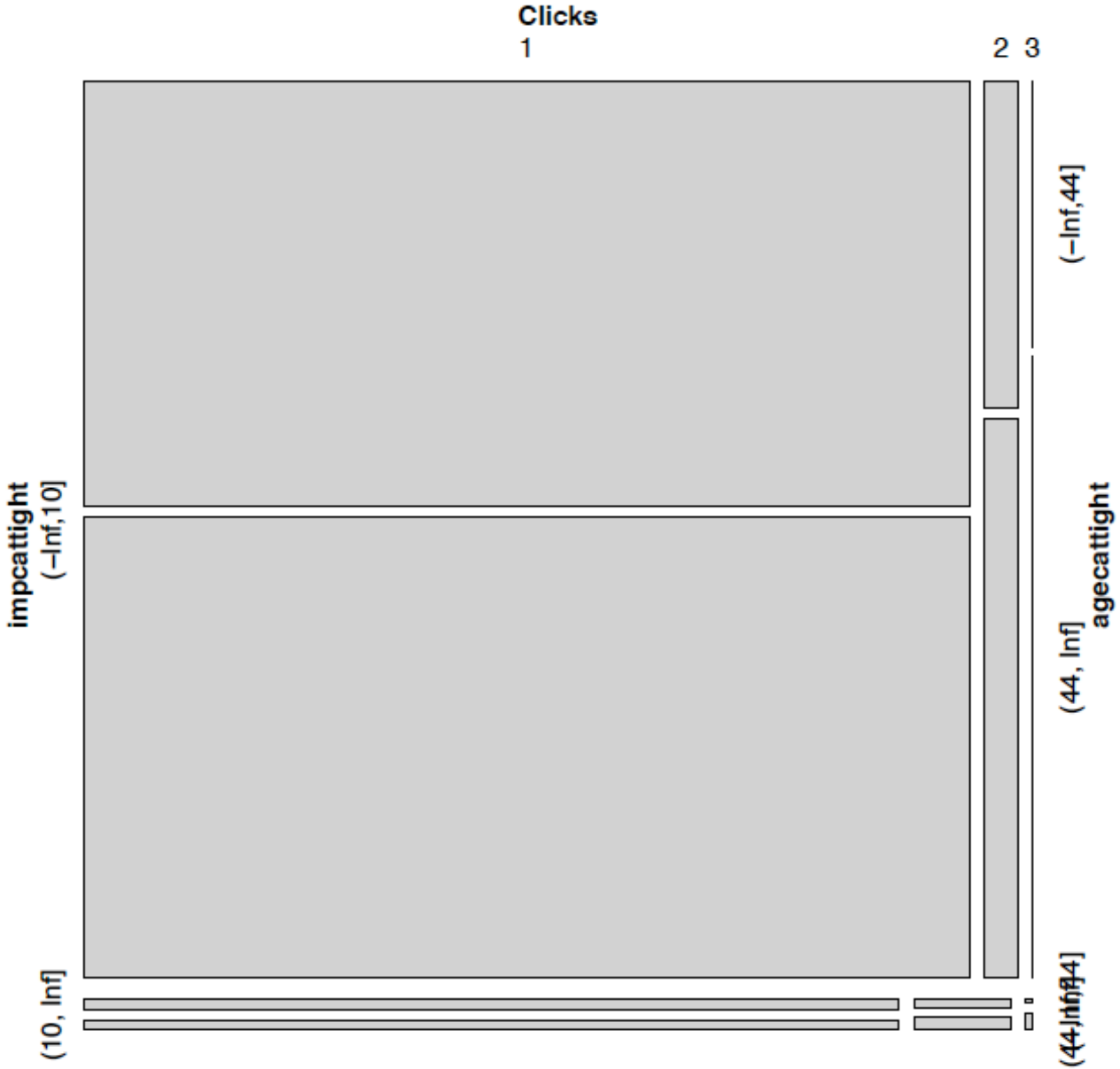


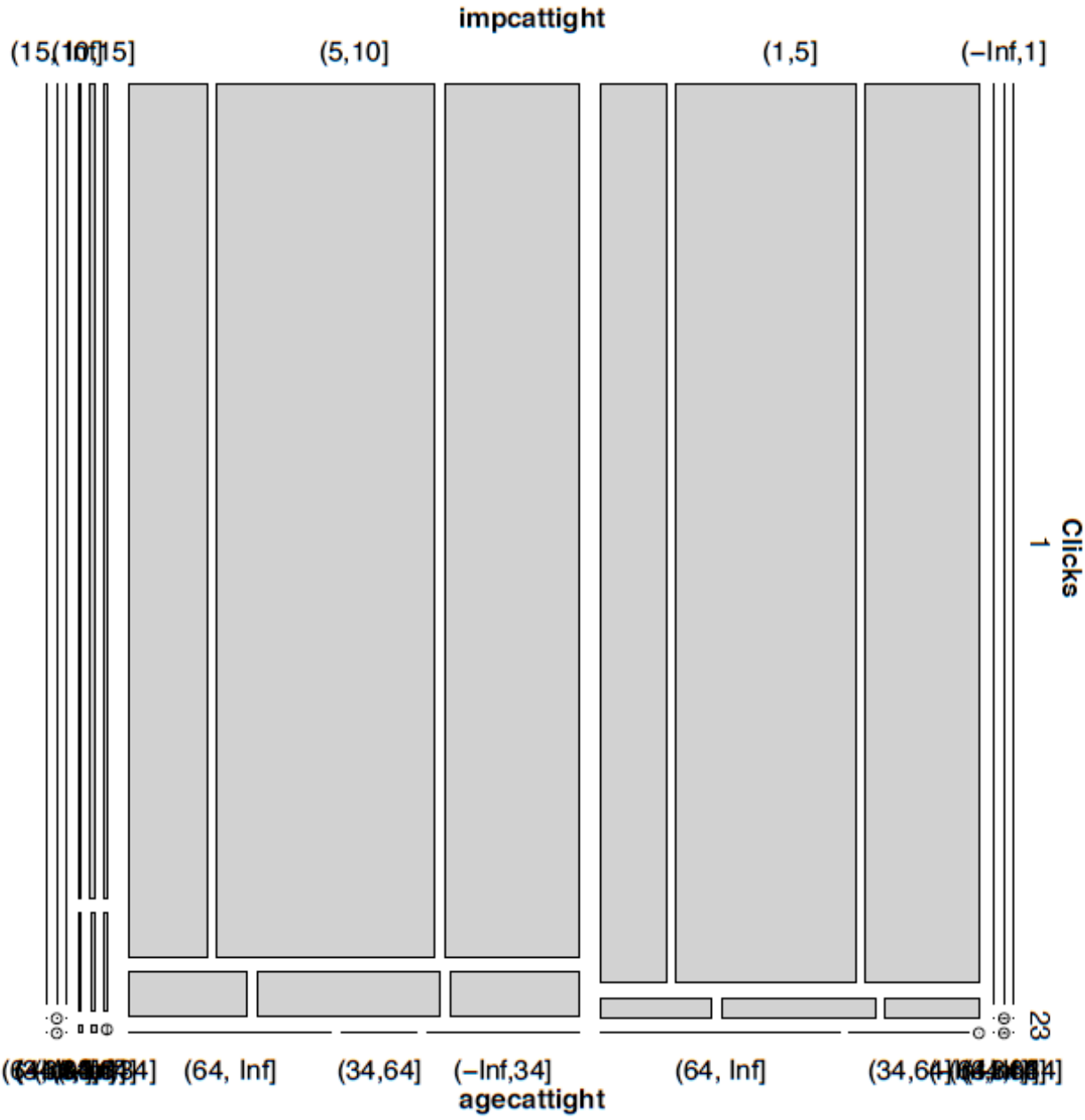
NYT normalized

- What's going wrong here?


(Clicks – mean(Clicks))/sd(Clicks)
(Impressions – mean(Impressions))/sd(Impressions)
(−Inf,0]
(0,18]
(18,24]
(24,34]
(34,44]
(44,54]
(54,64]
(64,74]
(74,84]
(84, Inf]
12
8
4
0
−2.5
0.0
2.5
5.0

A Mosaic Plot




impcattight
(15,[illegible]
(5,10]
(1,5]
(–Inf,1]
Clicks
1
23
(64, Inf]
(34,64]
(–Inf,34]
(64, Inf]
agecattight