January 24, 2018

CS 361: Probability & Statistics

Relationships in data

Standard coordinates

- If we have two quantities of interest in a dataset, we might like to plot their histograms and compare the two quantities, even if they have different units or measure entirely different things
- If we wanted to compare a histogram of internship earnings and GPA to see if there's something similar about the histograms, how to proceed?

Standard coordinates

A transformation of the dataset $\{x\}$

Create a new dataset $\{\hat{x}\}$ with standardized location — subtract the mean from each data item

And standardized scale — divide each data item by the standard deviation

$\{\hat{x}\}$ is dimensionless, has 0 mean, and unit standard deviation

Standard coordinates

- Recipe for computing standard coordinates from your dataset

Definition: 2.8 *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.

Standard normal data

- A wide variety of data, when standardized, will have a particular look and even fit a particular mathematical curve.

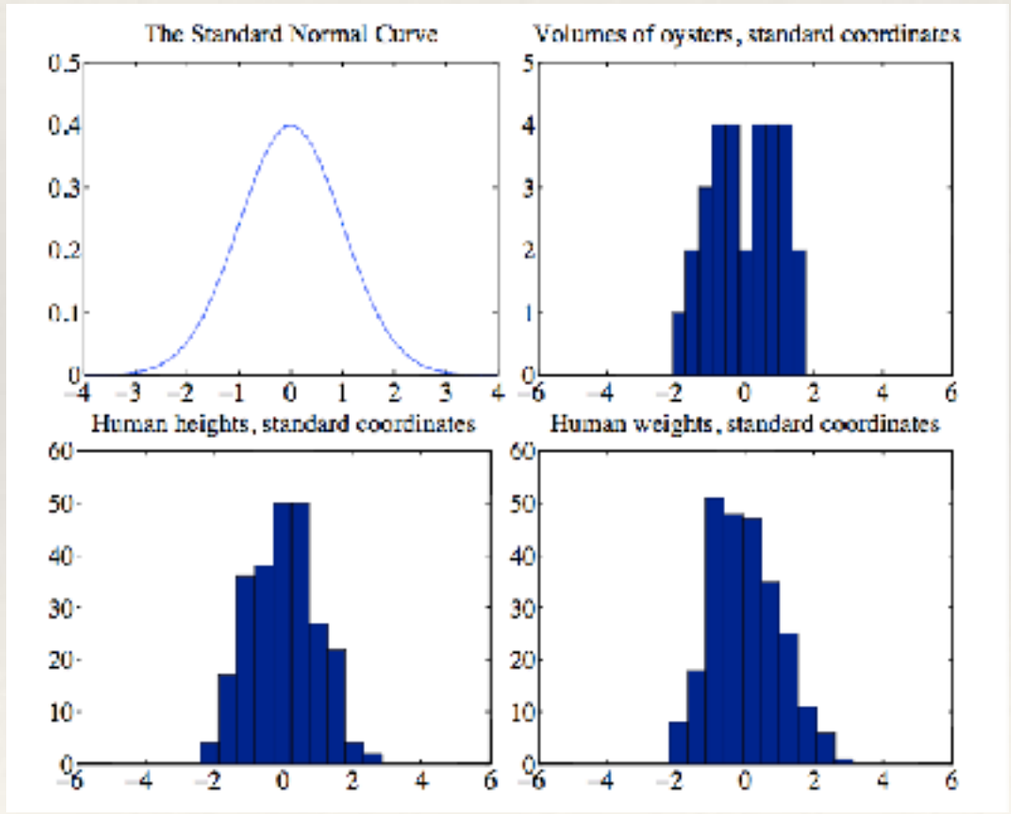
Definition: 2.9 *Standard normal data*

Data is standard **normal data** if, when we have a great deal of data, the histogram is a close approximation to the **standard normal curve**. This curve is given by

$$y(x) = \frac{1}{\sqrt{2\pi}} e^{(-x^2/2)}$$

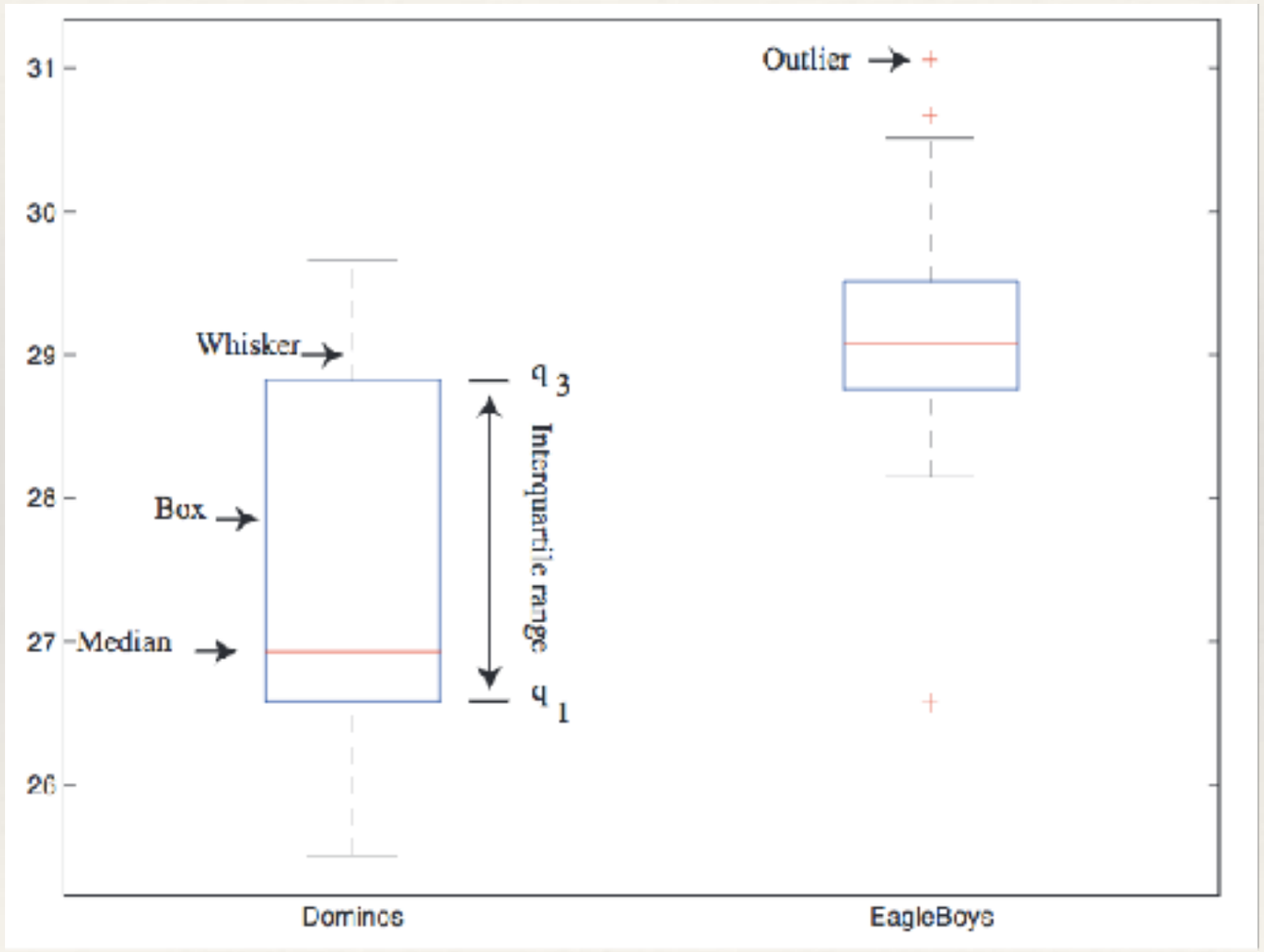
Definition: 2.10 *Normal data*

Data is **normal data** if, when we subtract the mean and divide by the standard deviation (i.e. compute standard coordinates), it becomes standard normal data.



Boxplots

- Another type of visualization



How to make a box plot

- Height of the box is from q1 to q3, width is whatever makes it look nice
- Identify the median
- Use a rule for outliers: bigger than q3 + 1.5(iqr) or smaller than q1 - 1.5(iqr) for example
- Whiskers extend to the largest data item which isn't an outlier and smallest data item which isn't an outlier
- Outlier data points indicated

Chapter 2: Relationships in data

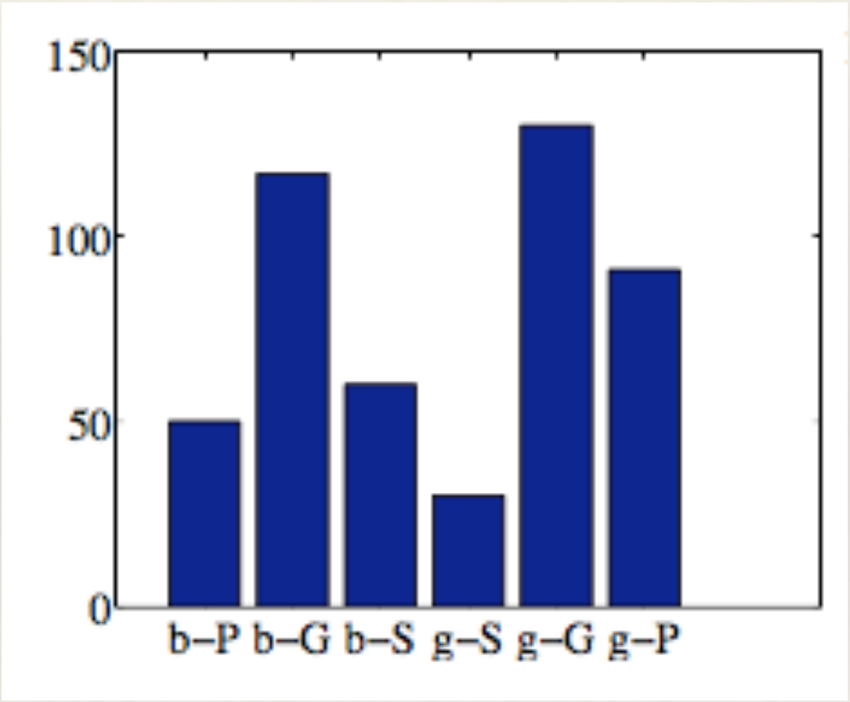
Relationships

- Most of what we have covered so far has been about visualizing or describing a single dimension of a dataset
- We may be interested in how two or more dimensions of a dataset relate to one another
- We might expect there to be a relationship among: temperature and latitude, height and weight, hours spent studying and GPA, etc.

Visualizing relationships

Plotting 2d categorical data

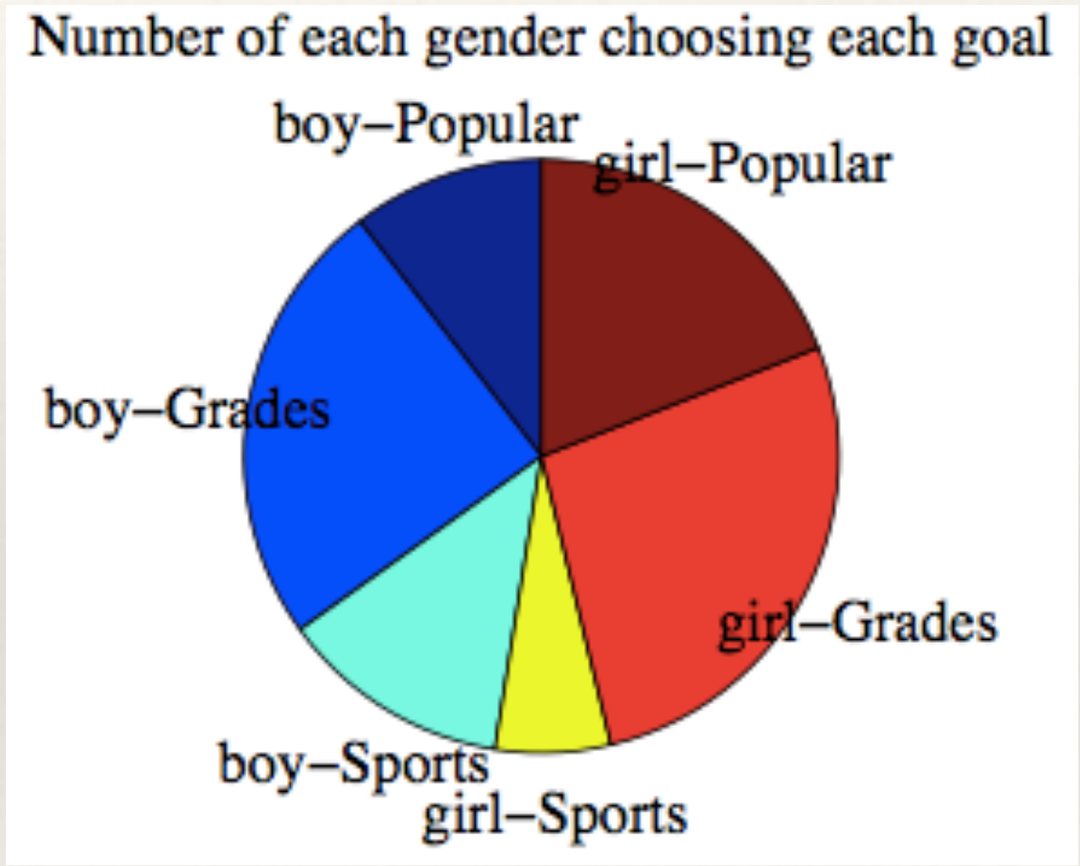
- We could try and collapse two categorical variables with *m* and *n* different values into a single variable with *mn* values and use a bar chart



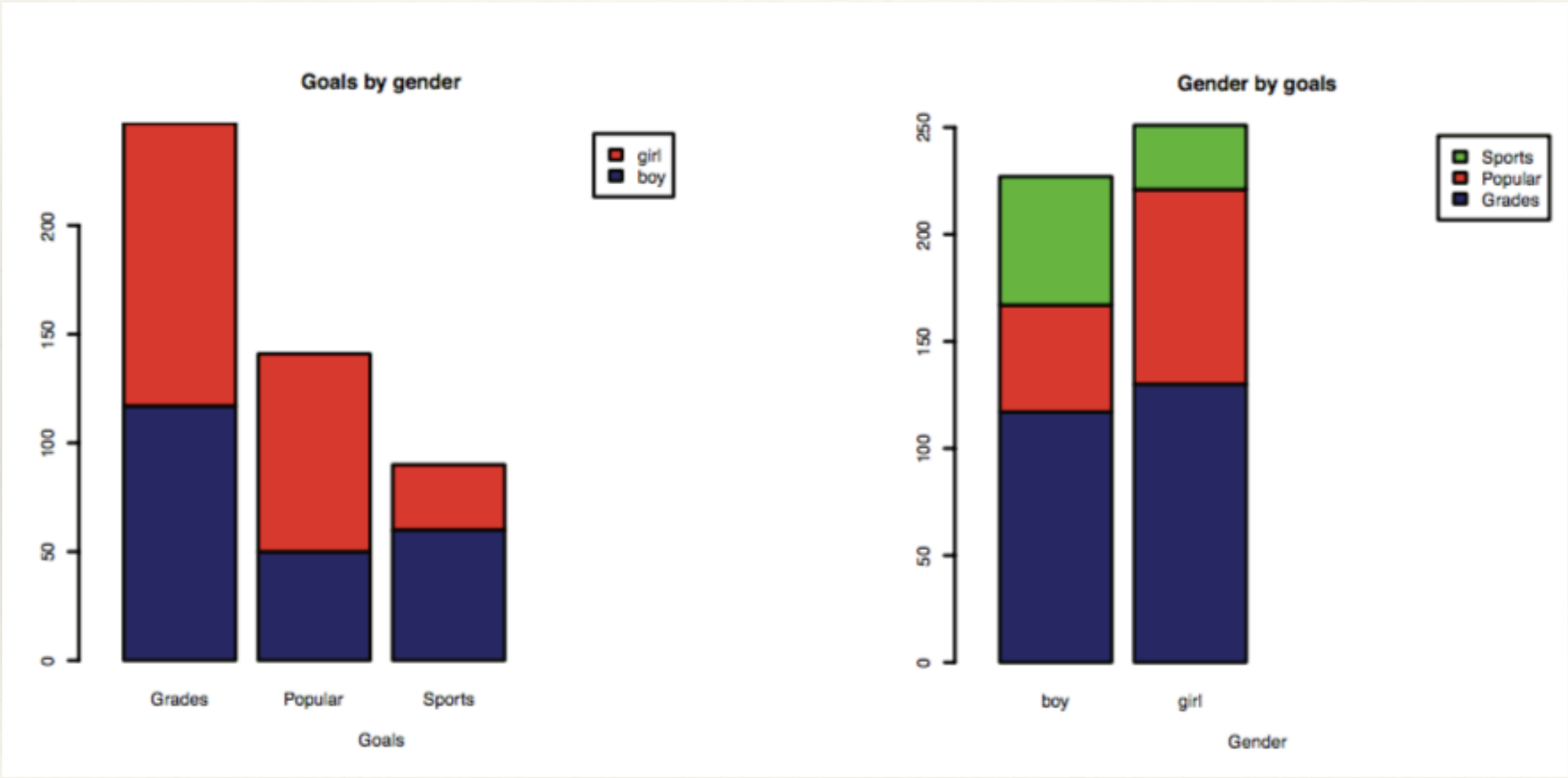
- What could go wrong?
- *mn* could be too large for the chart to be readable

Pie charts

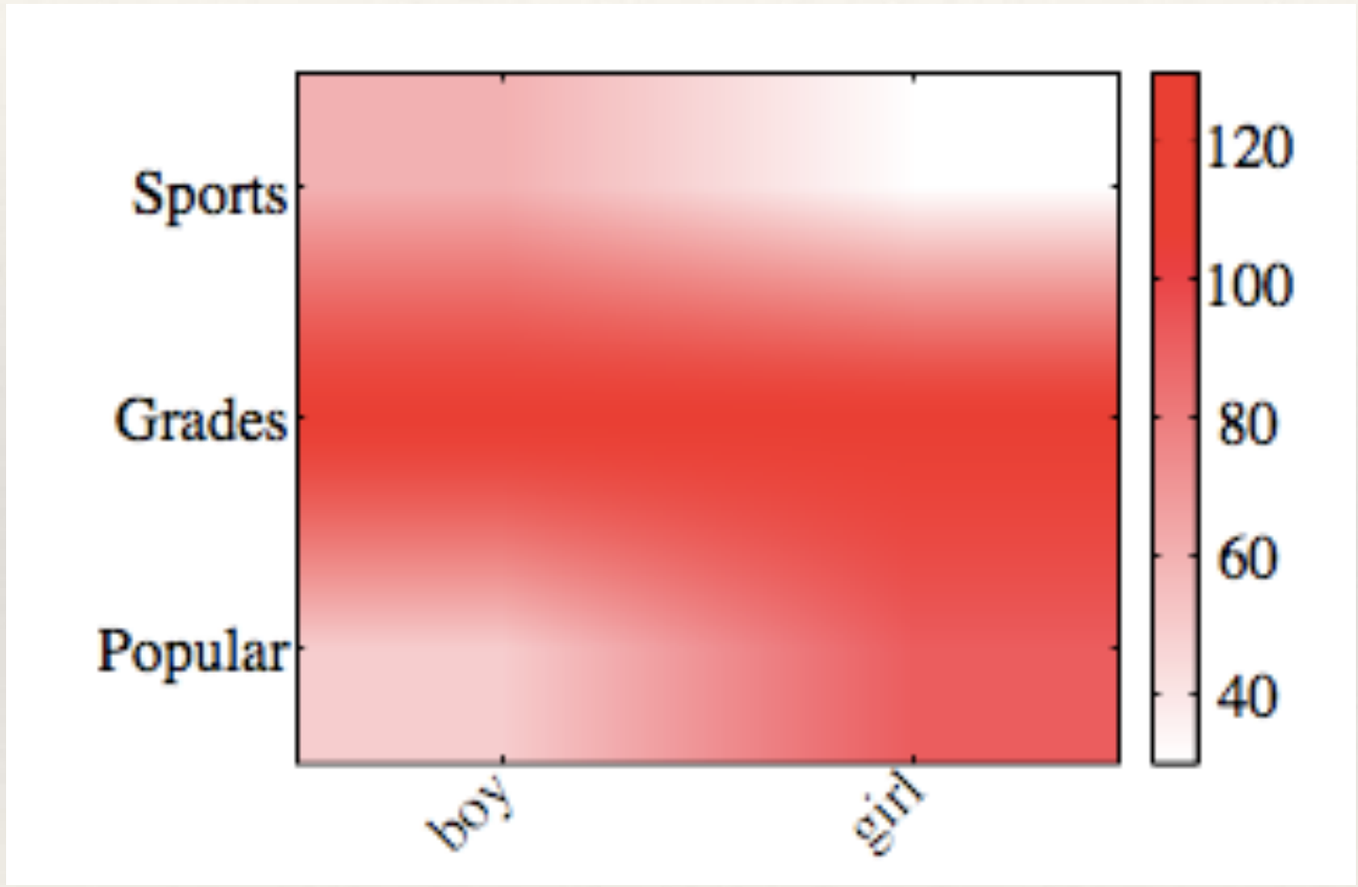
- An alternative is a pie chart
- Small differences may be hard to judge, though



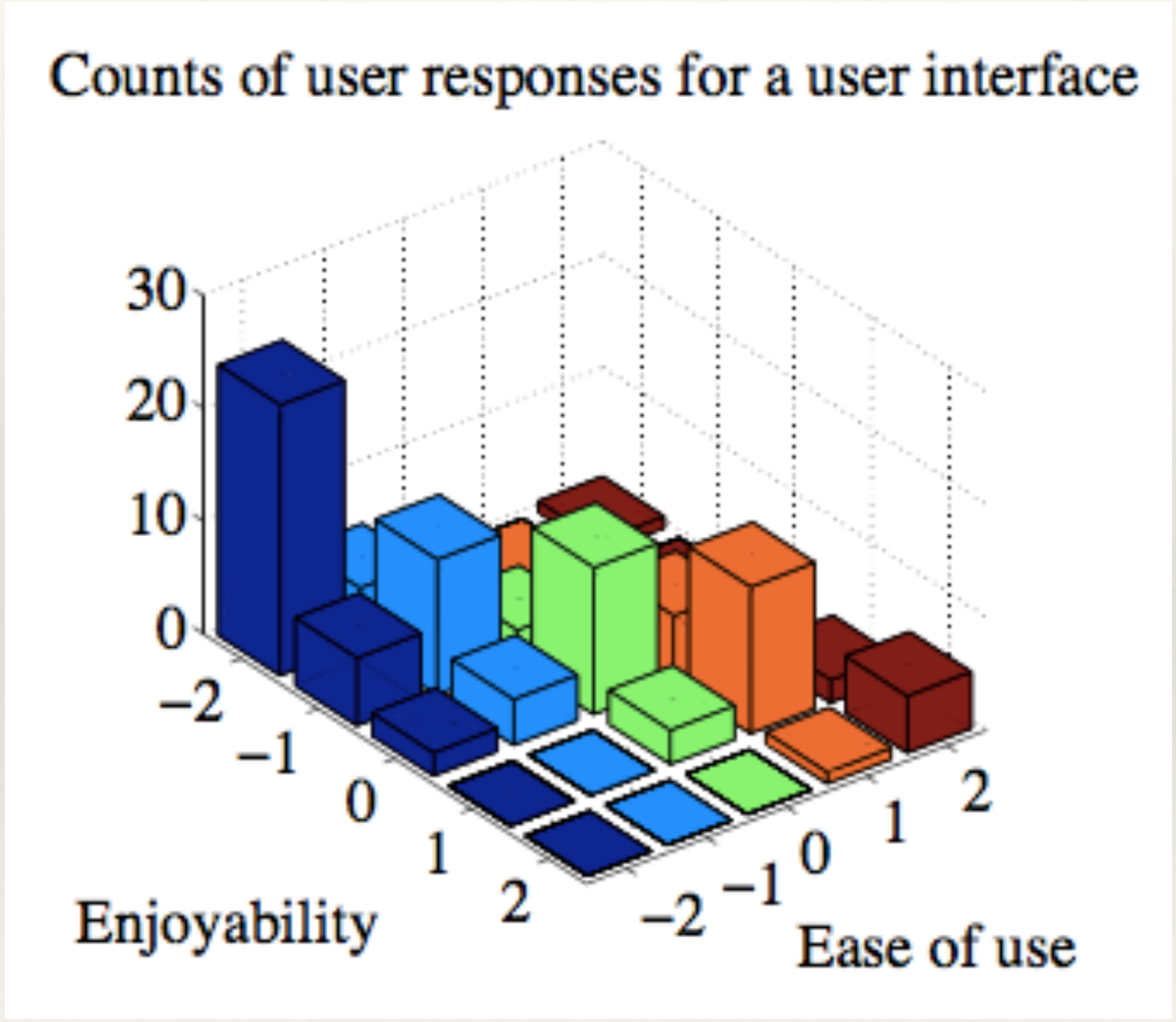
Stacked bar charts



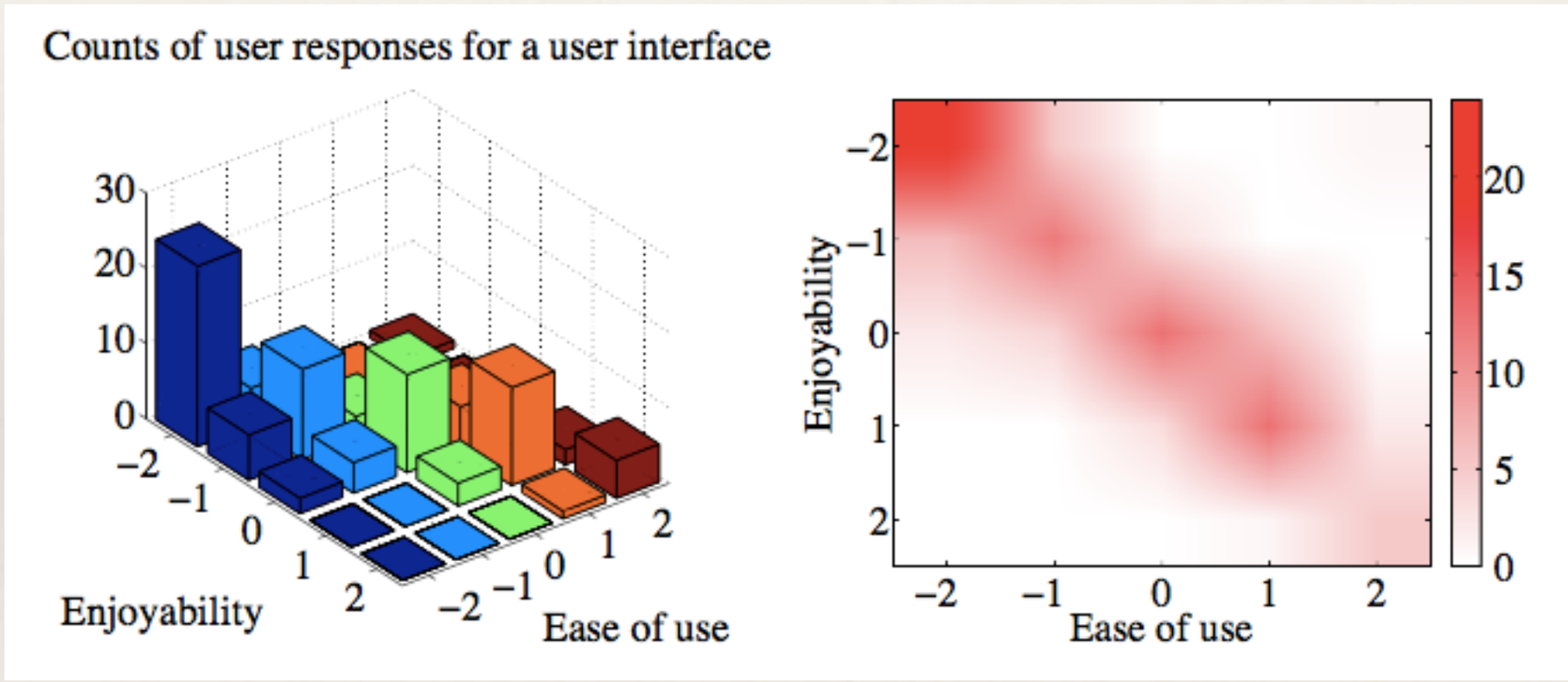
Heat Maps



3D bar charts



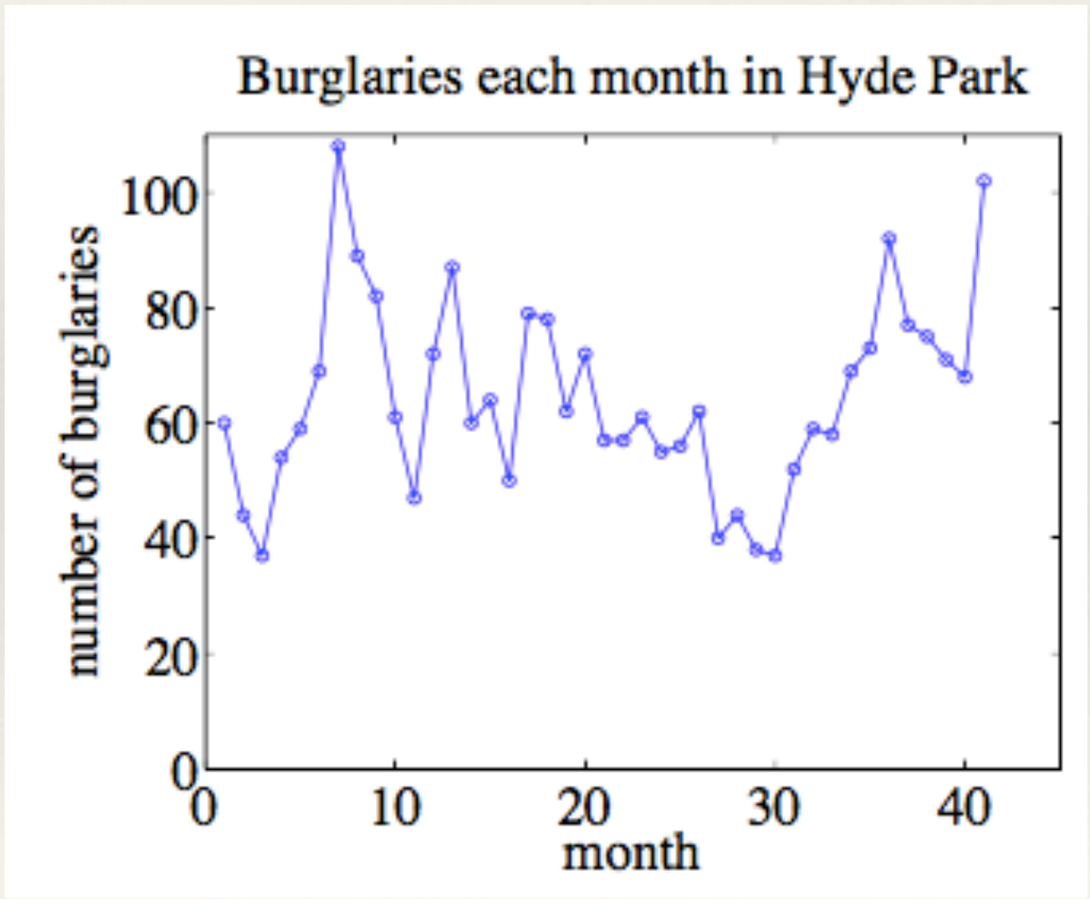
3D bar charts — occlusion



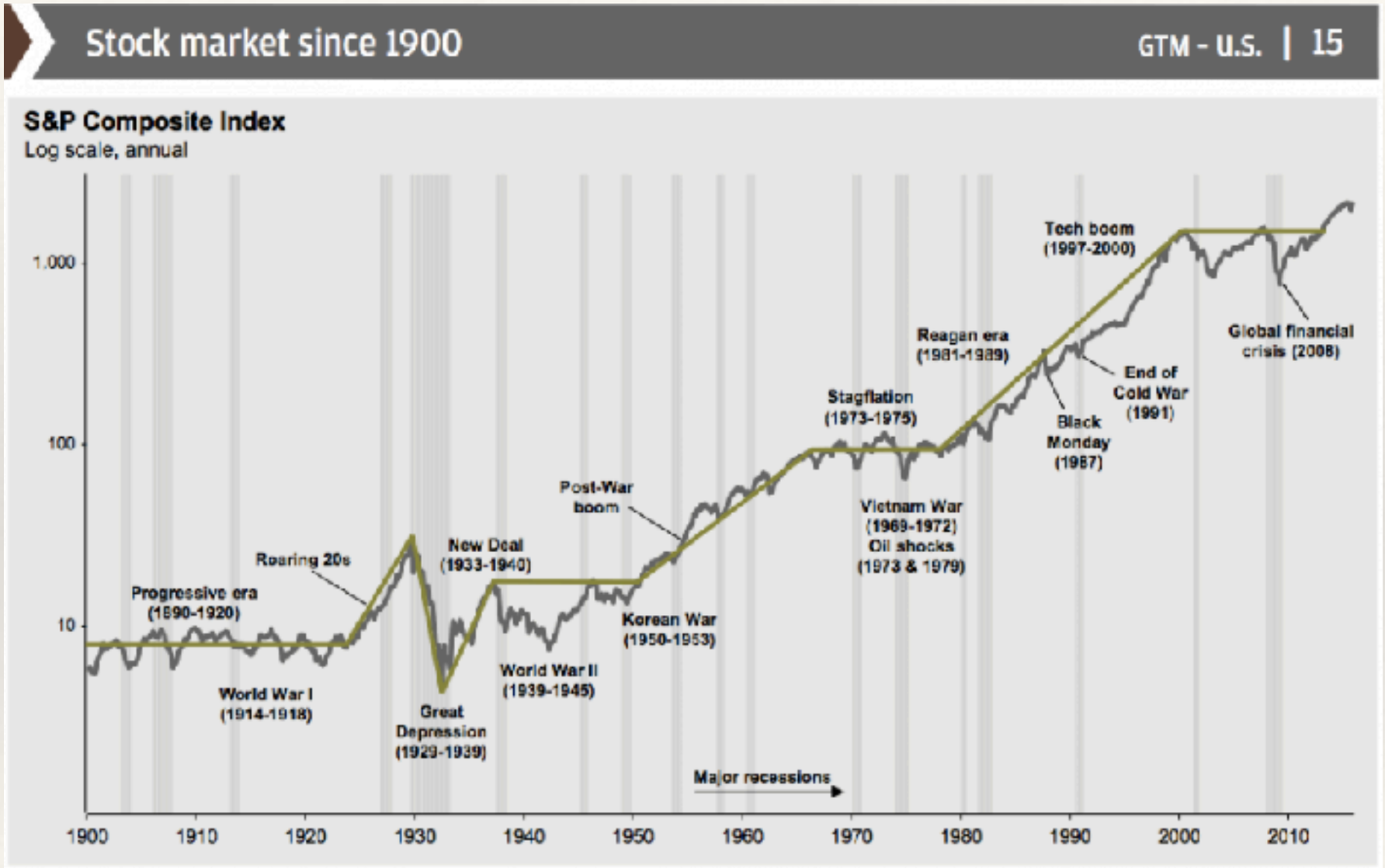
Time series and spatial data

Time series data

- Sometimes there is something in the data that indicates an ordering in time for the data (month, day, year, etc)
- Plot the points Connect with lines
- Trends: upwards, downwards, periodic?
- Ask yourself what's happening before and after the cutoffs chosen



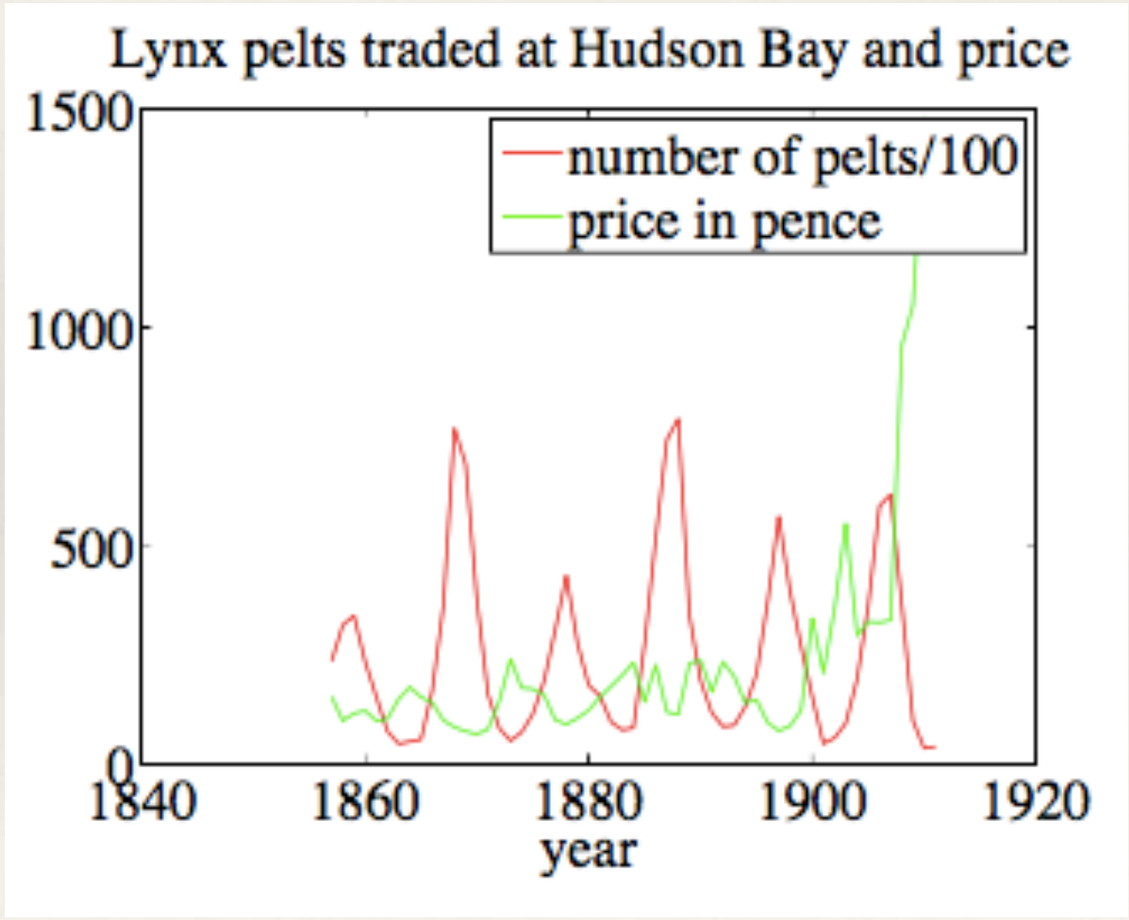
Time series



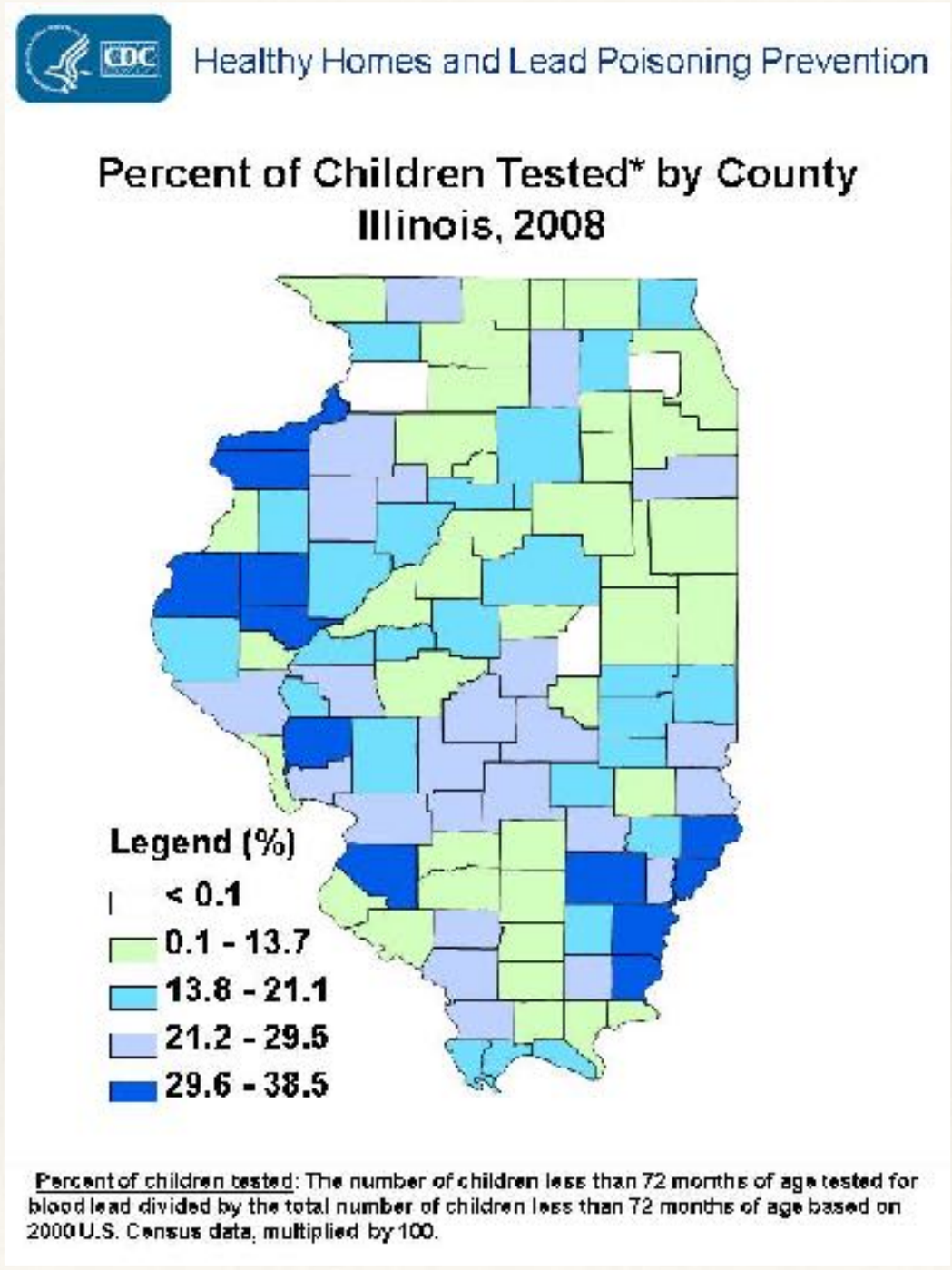
Source: FactSet, NBER, Robert Shiller, J.P. Morgan Asset Management.
Data shown in log scale to best illustrate long-term index patterns.

Relationships in time series

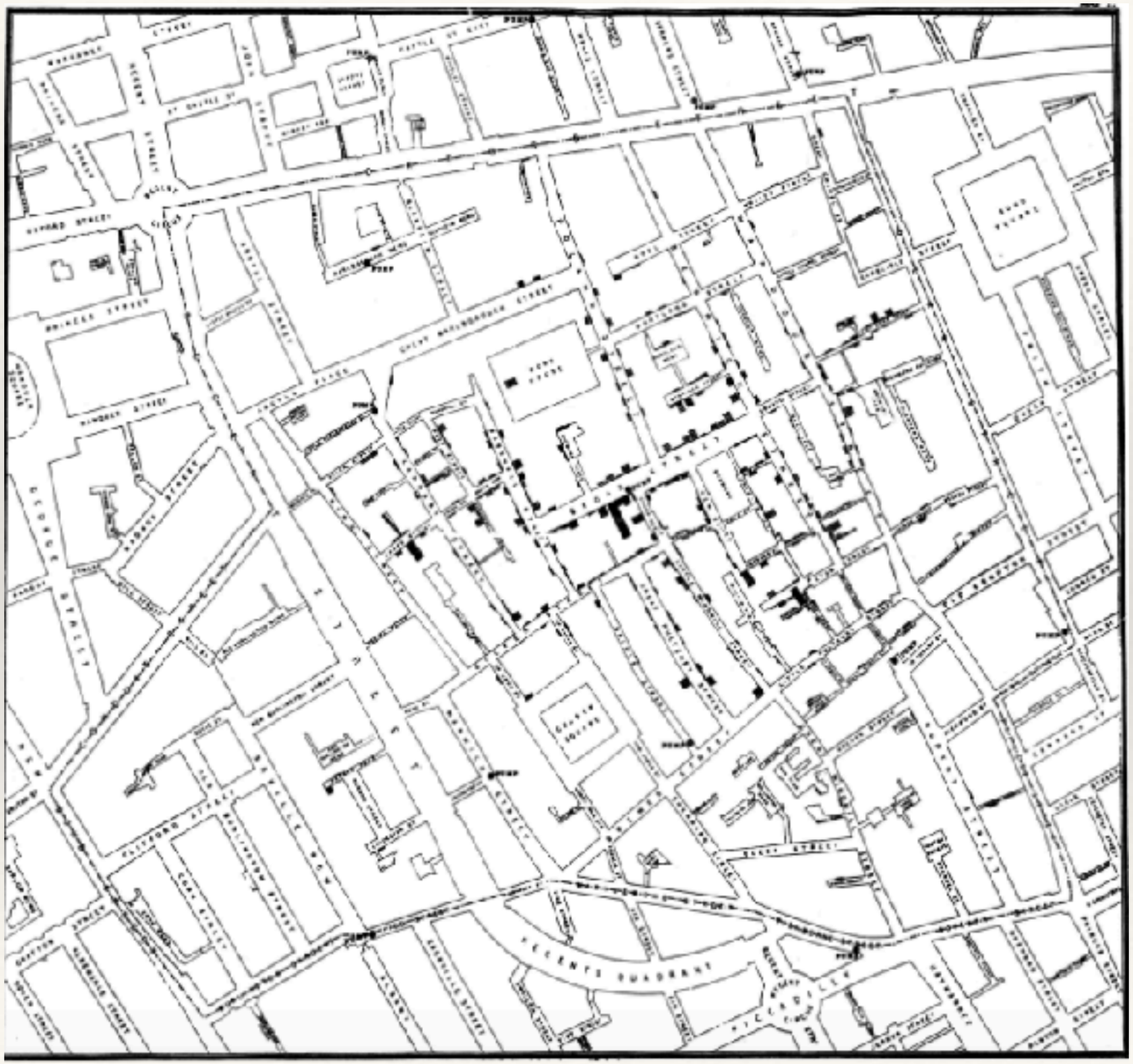
- We can visually inspect a relationship between two variables in a dataset with these plots
- Why might the number of pelts be periodic?
- The price?

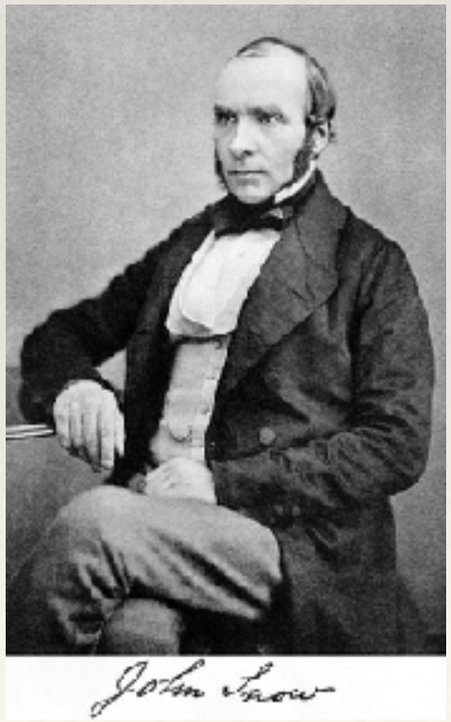


Plotting spatial data



Spatial data and Cholera



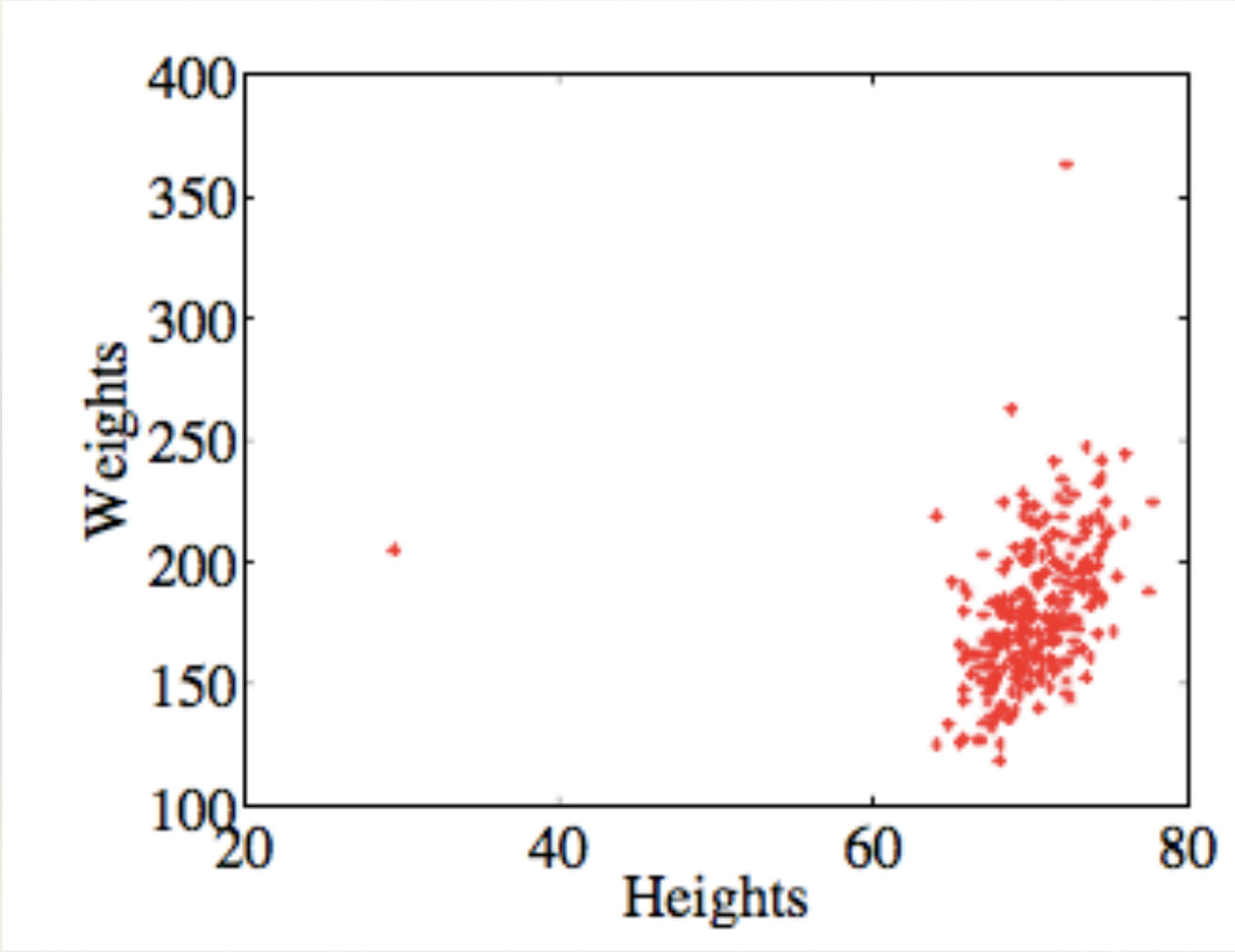


Spatial data

Scatterplots

Visualizing two variables: scatterplots

- 2D plot with one numerical variable on each axis

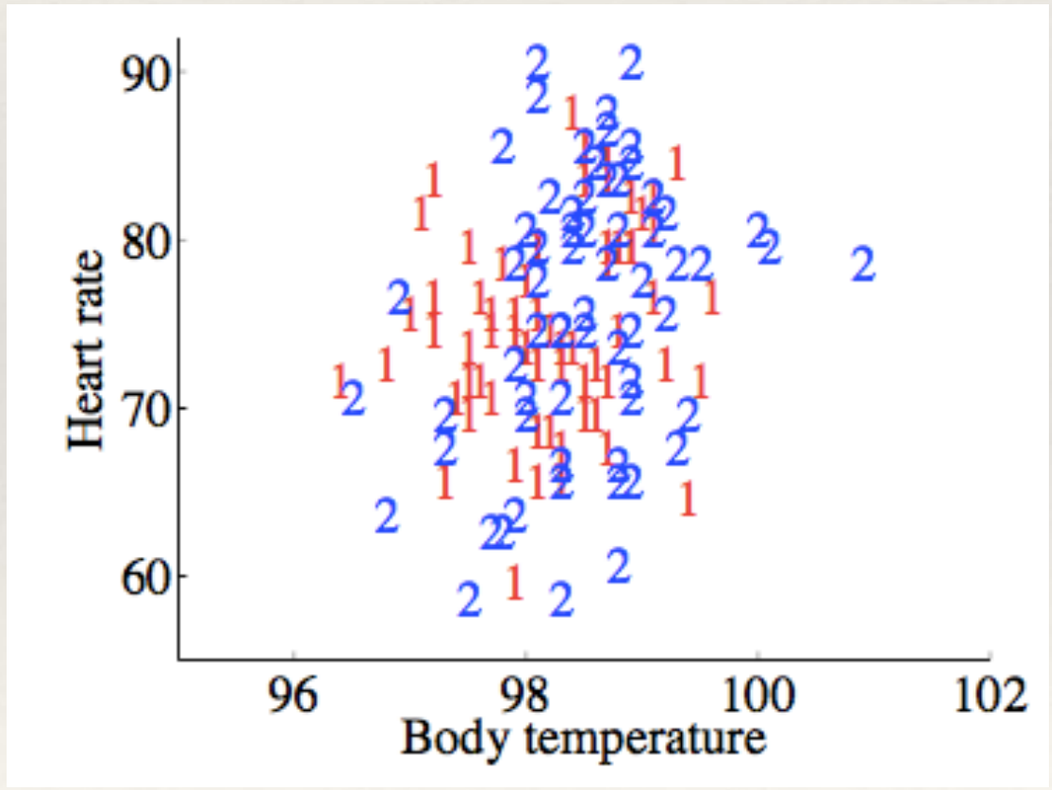


Scatterplots

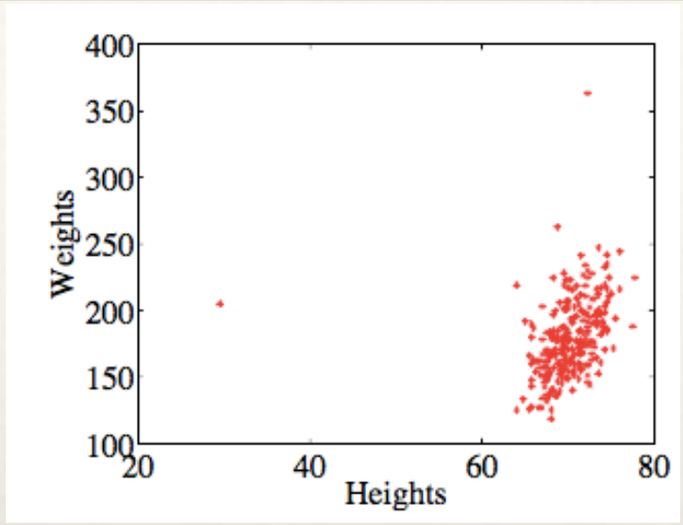
- Choose 2 of the d variables in your dataset that you're interested in investigating for some relationship
- Call one of the variables x and one y
- Creating a new dataset $\{\mathbf{x_i}\} = \{(x_i, y_i)\}$
- Then plot a mark on a graph for each data item at the (x,y) coordinate given by the 2 variables you've chosen to look at
- It doesn't really matter which is x and y (what if we flipped them?)

Including more information

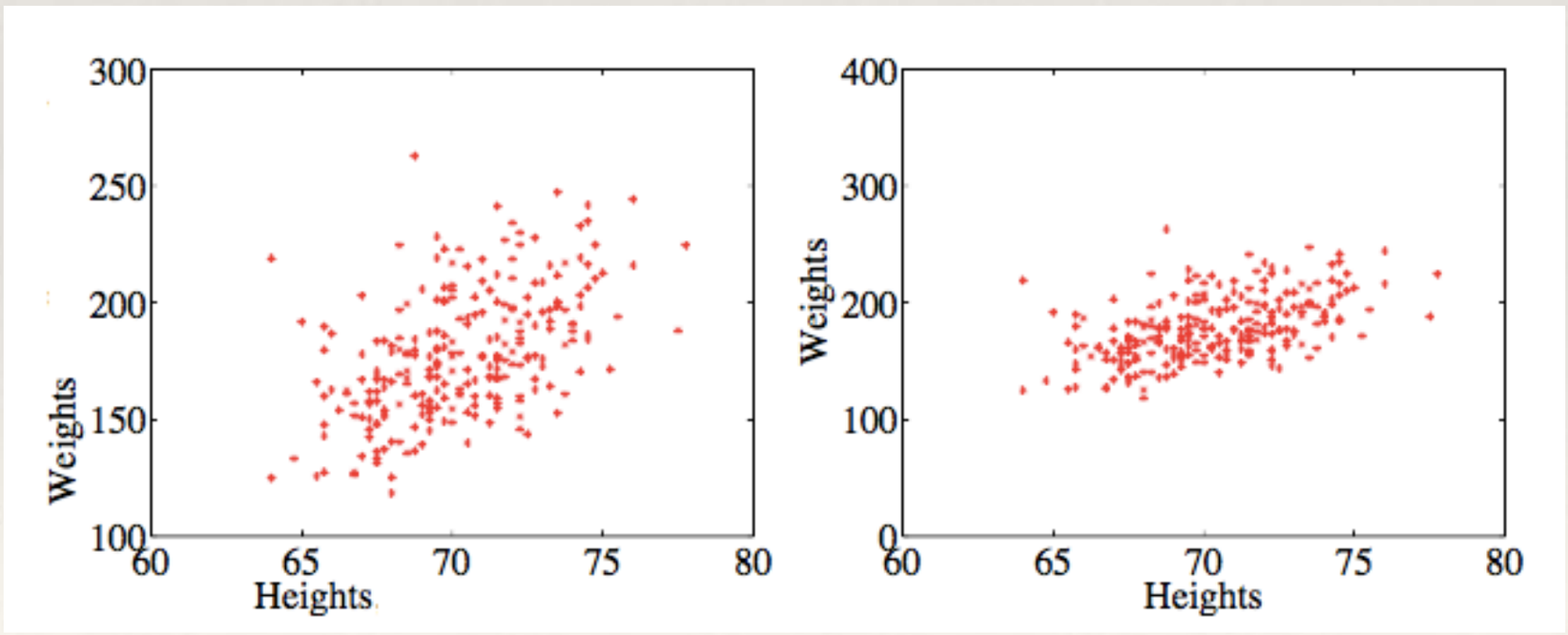
- It's possible to use point size, point color, or different types of points (x's and o's for instance) to indicate the values of other variables in the plot
  - Any difference between sex 1 and 2?
  - Relationship between temp and HR?



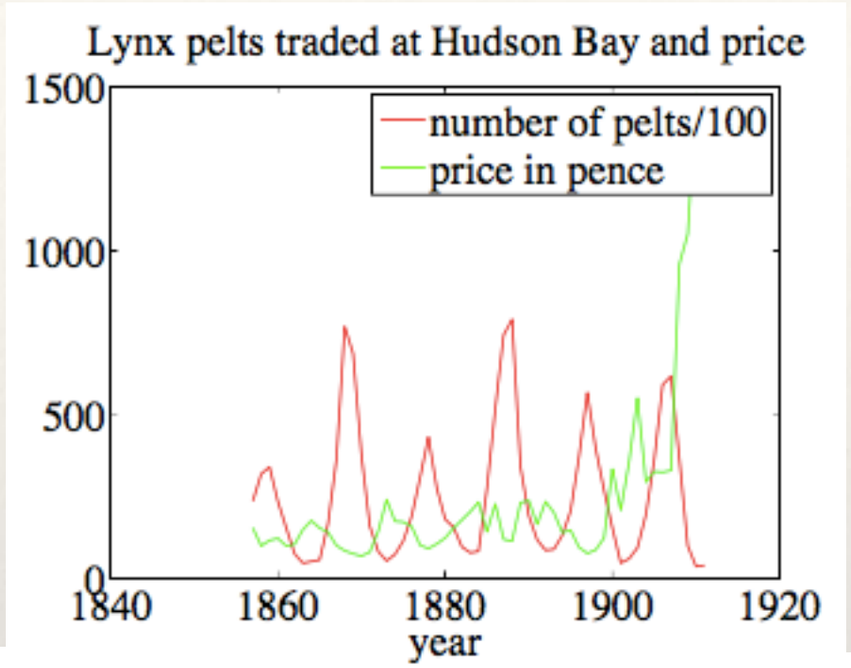
Scale matters



The same dataset, outliers removed, two different axis scales

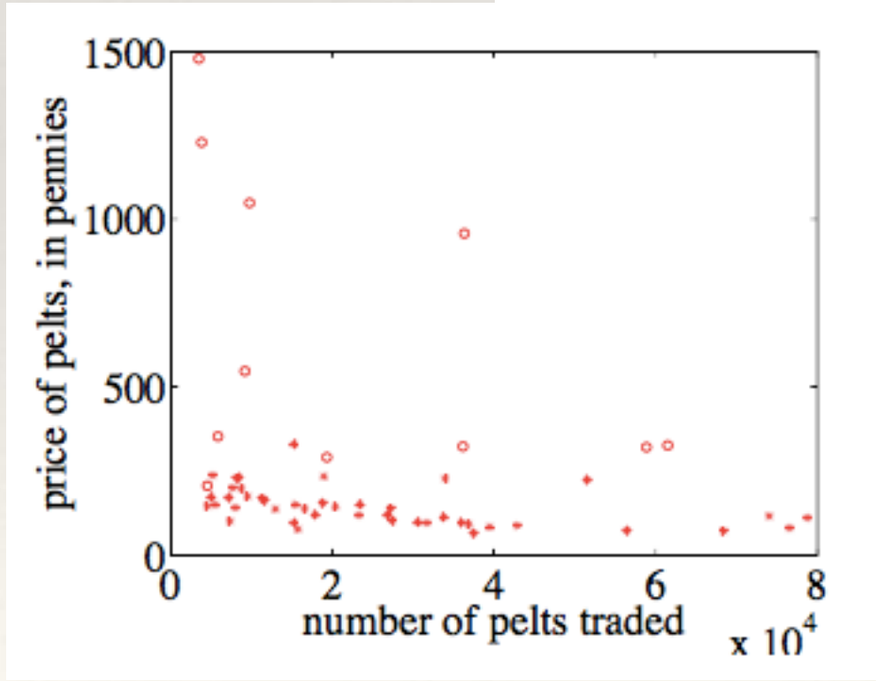


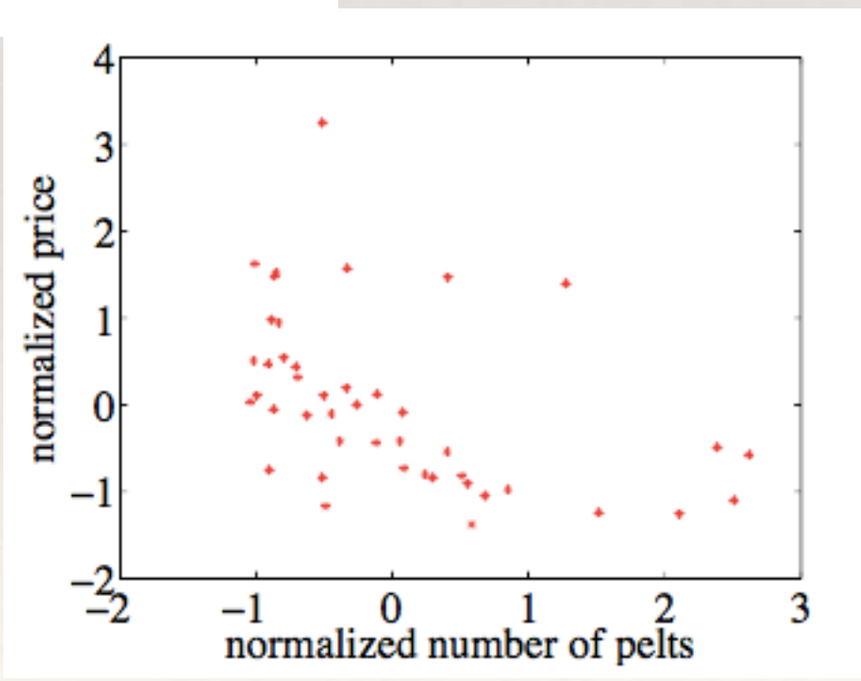
Normalization



The top figure shows the time series data, the bottom two are the same data on a scatter plot

In the bottom left, it's hard to see the law of supply and demand. Normalization reveals that this is a scaling artifact





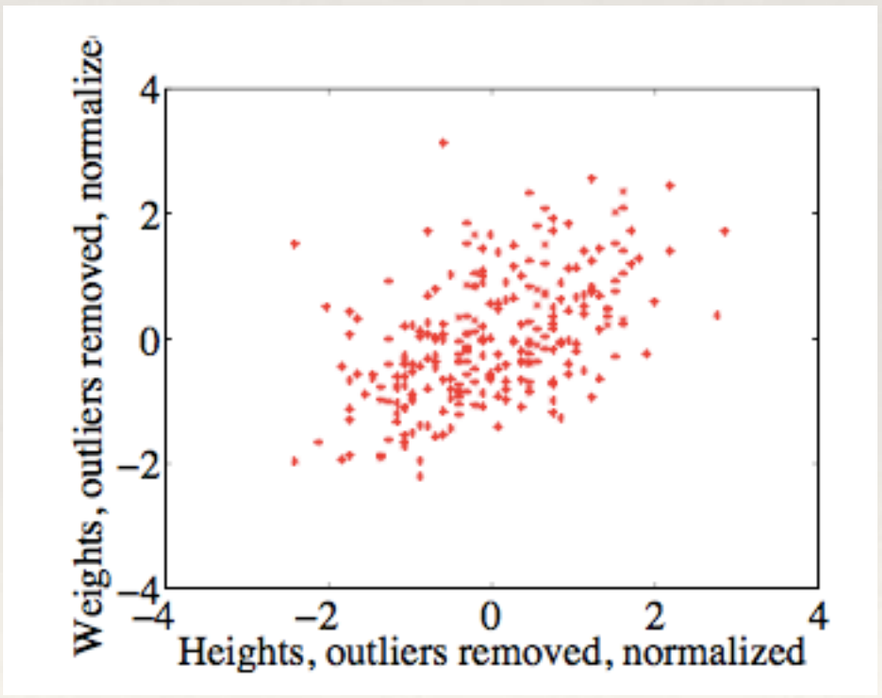
Scatterplots summary

- A good first choice when dealing with 2 numerical dimensions of data
- Scale matters, so it's a good idea to use standard coordinates
- May want to remove outliers or unusual data

Correlation

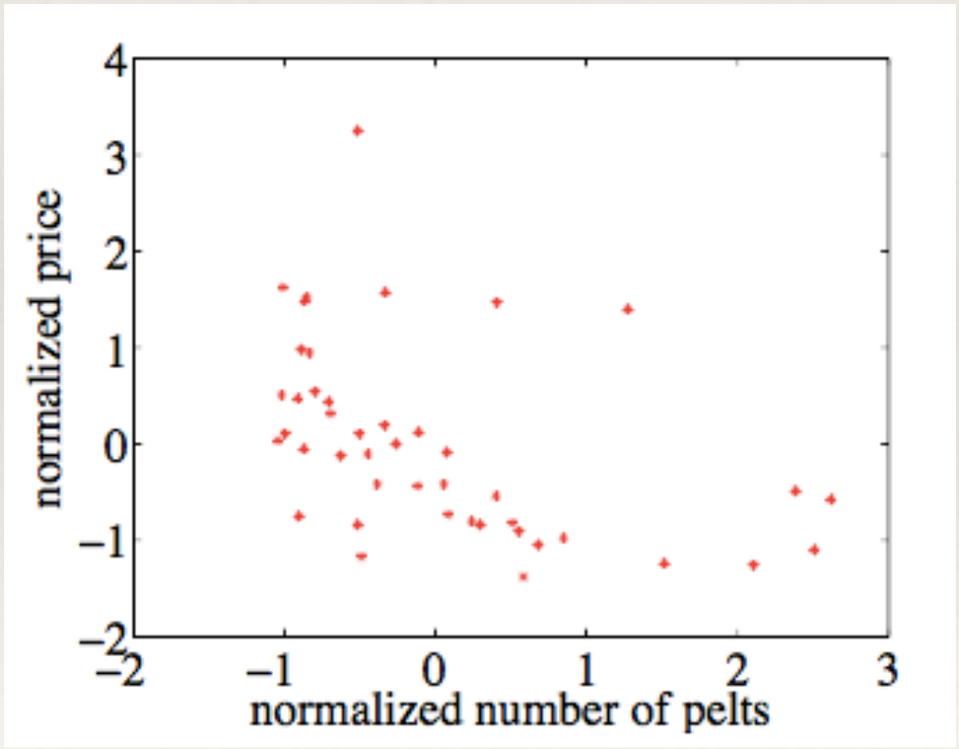
Correlation

- Broadly, if x changes, what does y do?
- If a small x and small y (respectively large x and large y) tend to occur together we say there is **positive correlation** between x and y



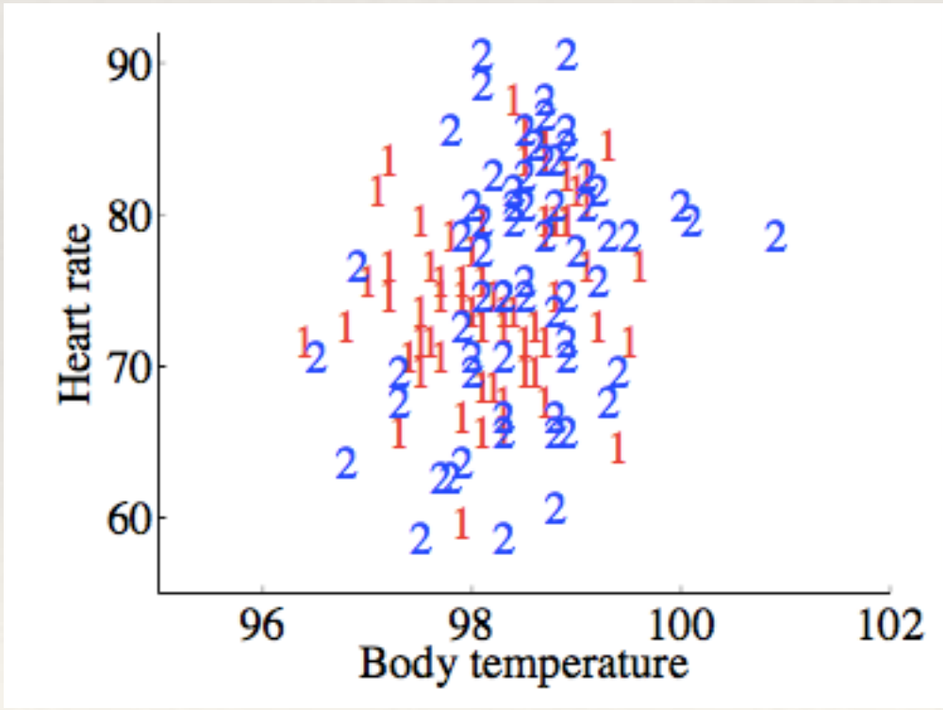
Correlation

- If small values of x tend to occur with large values of y and large values of x tend to occur with small values of y we say that x and y are **negatively correlated**



Correlation

- When there is no tendency for x and y to be either large or small together, we say there is **zero correlation**
- Our data will be more of a blob



Examples

Lines of code in a codebase and number of bugs?

Body temperature and height?

GPA and hours spent playing video games?

Earnings and happiness?

Correlation coefficient

Suppose we have N data items that are each 2-vectors

$(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$

Normalize the data

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

The correlation coefficient of x and y is the mean of the product $\hat{x}_i \hat{y}_i$, i.e.

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

Properties of correlation

- Correlation coefficient is symmetric

$$\text{corr}\left(\{(x,y)\}\right) = \text{corr}\left(\{(y,x)\}\right)$$

- Not changed by translating data

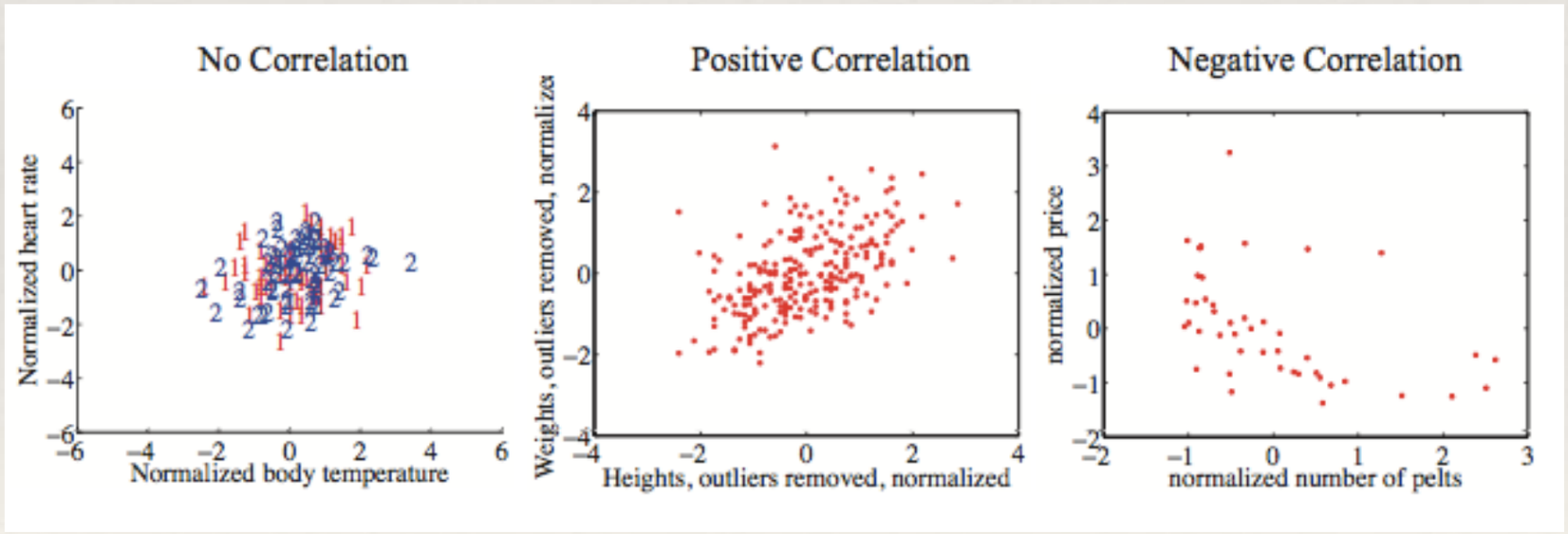
- Scaling may change the sign

$$\text{corr}(\{(ax+b, cy+d)\}) = \text{sign}(ac)\text{corr}(\{(x,y)\})$$

Properties of correlation

- How do these two conceptualizations line up?

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



Properties of correlation

The largest possible correlation is 1 and happens when $\hat{x} = \hat{y}$

The smallest possible correlation is -1 and happens when $\hat{x} = -\hat{y}$

Proving correlation bounds

Proposition: $\text{corr}(\{(x,y)\}) \leq 1$

First note that the correlation can be written as a dot product of two vectors

Let $\mathbf{x} = \frac{1}{\sqrt{N}}[\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_N]$ and $\mathbf{y} = \frac{1}{\sqrt{N}}[\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N]$

We have $\mathbf{x} \bullet \mathbf{y} = \frac{\sum_i \hat{x}_i \hat{y}_i}{N}$ or $\mathbf{x} \bullet \mathbf{y} = \text{corr}(\{(x,y)\})$

Either $\mathbf{x} \bullet \mathbf{y} \leq \mathbf{x} \bullet \mathbf{x}$ or $\mathbf{x} \bullet \mathbf{y} \leq \mathbf{y} \bullet \mathbf{y}$

But $\mathbf{x} \bullet \mathbf{x} = \frac{\sum_i \hat{x}_i^2}{N}$ which is $\text{std}(\{\hat{x}\})^2$

Recall

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

Proof

Since $\{\hat{x}\}$ is our standardized dataset, we have $\text{std}(\{\hat{x}\})^2 = 1$

Similar reasoning applies for y

So we know that $\mathbf{x} \bullet \mathbf{y} \leq 1$

And since $\mathbf{x} \bullet \mathbf{y} = \text{corr}(\{(x, y)\})$

We've shown that $\text{corr}(\{(x, y)\}) \leq 1$

Using correlation to predict

- One useful task is to take what we know about the data we have and make predictions about data we don't yet have or measurements we have that are incomplete
- Example: we might like to go into the fur pelt business and have a bunch of historical data on supply and prices. We know the price today and would like to guess as to the total supply
- That is we have a bunch of pairs *(x,y)* for prices and supply. But our state of knowledge today might be *(x_0, ???)*
- Correlation will be useful for this task

Prediction

- We want a predictor that we can apply to any x
- We want it to behave well on our existing data
- We can choose the predictor by considering the error the predictor will have

Prediction

Since it's possible to convert to and from standard coordinates and we know standard coordinates have nice properties like 0 mean and 1 standard deviation, we will first convert

We will write $\hat{y}_i^p$ to indicate our predicted value of $\hat{y}_i$ for the point $\hat{x}_i$