September 5, 2017

CS 361: Probability & Statistics

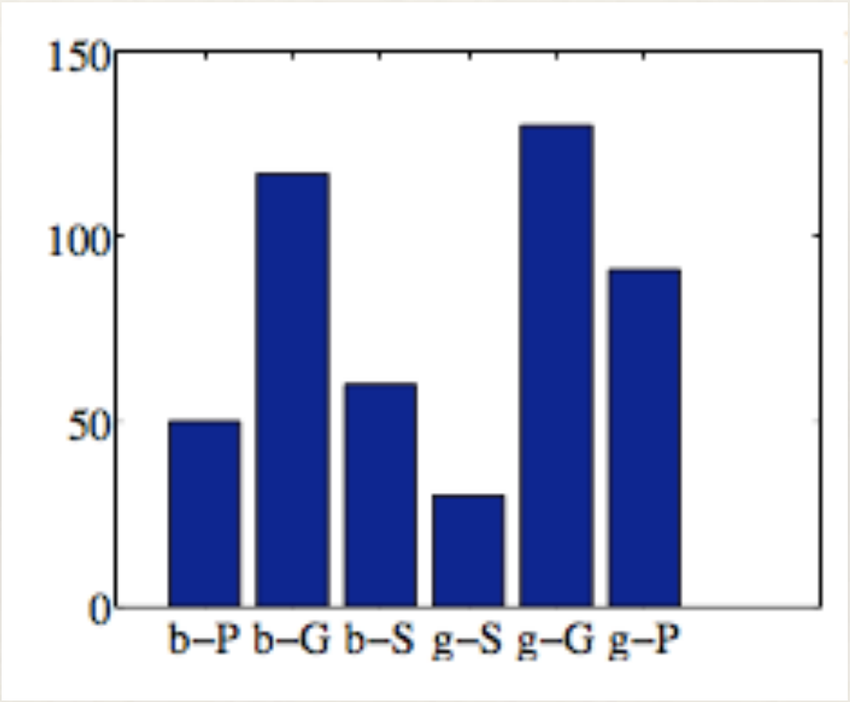
Relationships in data

Relationships

- Most of what we have covered so far has been about visualizing or describing a single dimension of a dataset
- We are often 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.

Plotting 2d categorical data

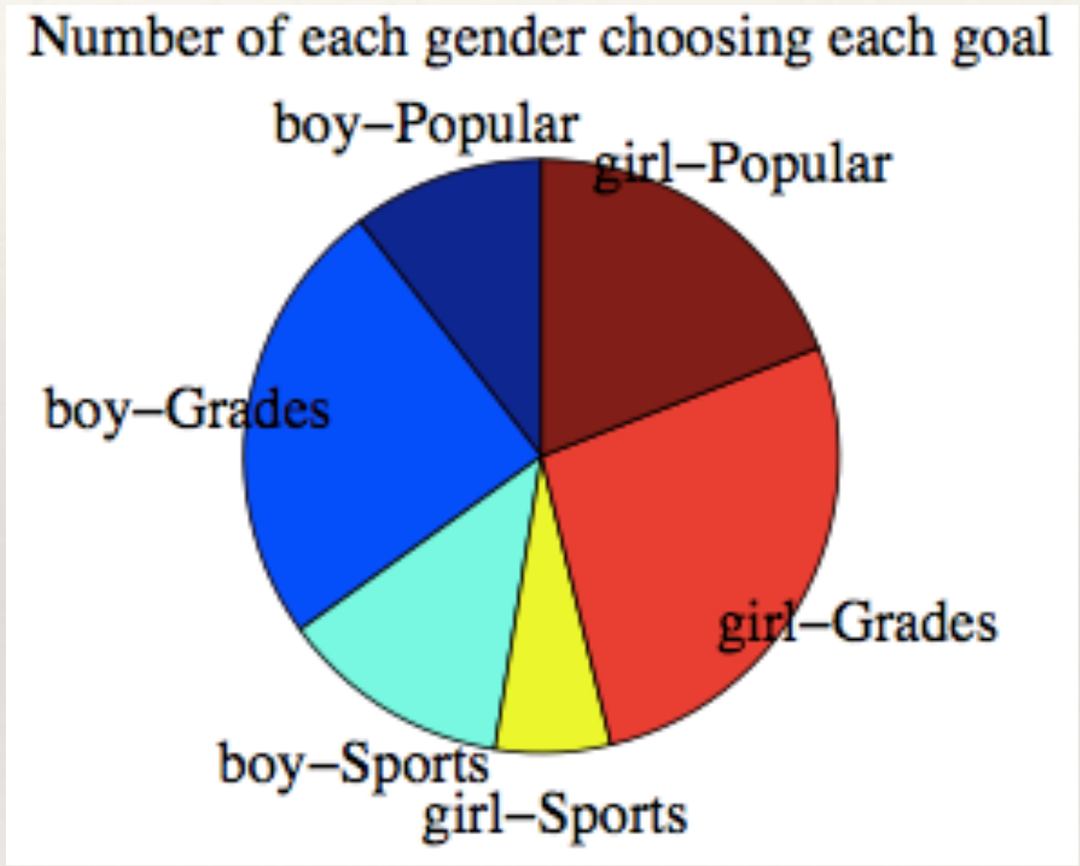
- We could try and collapse two categorical variables with *m* and *n* different values into a single variable with *mn* values and use a bar chart



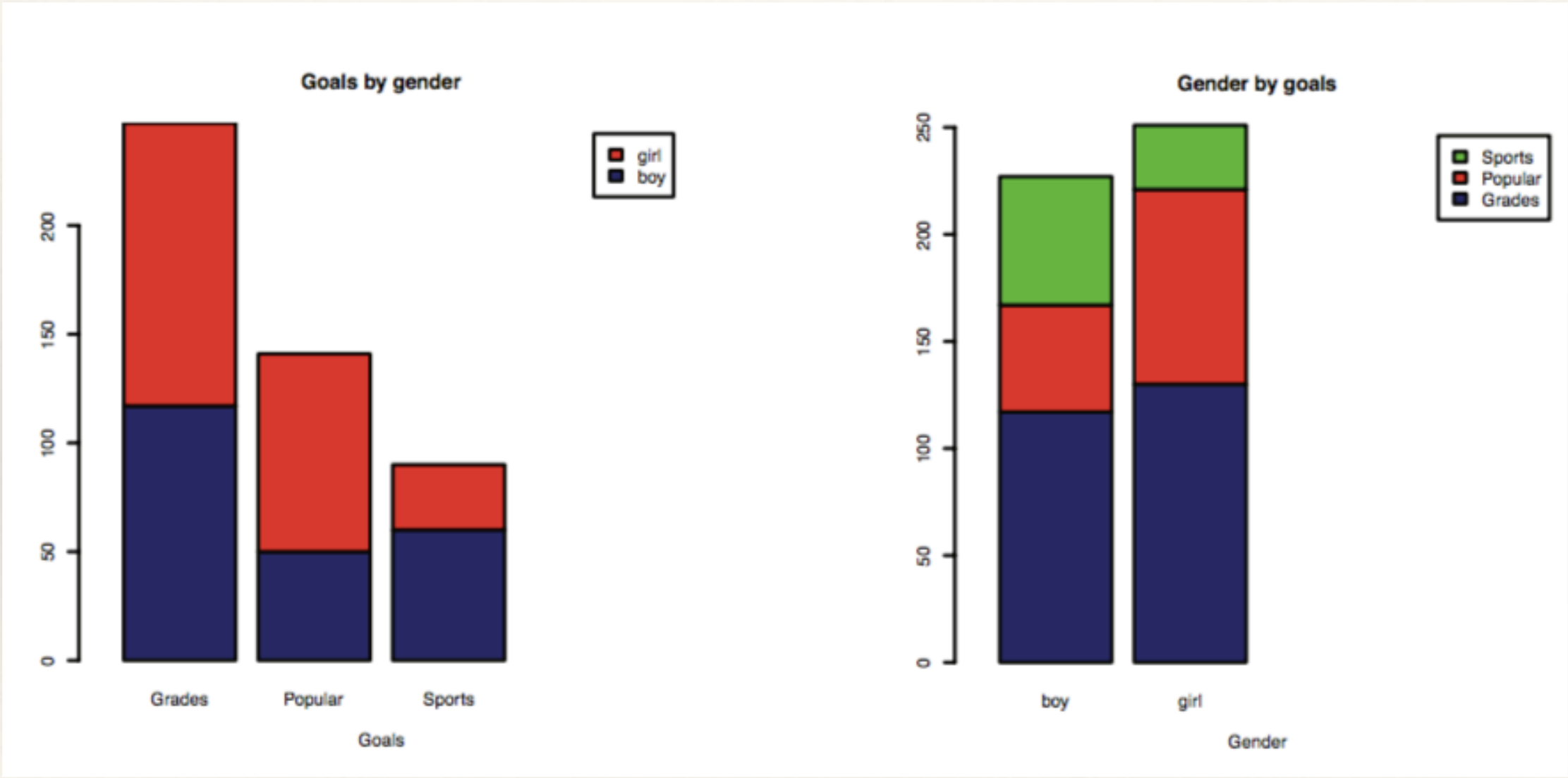
- What could go wrong?
- *mn* could be too large for the chart to be readable

Pie charts

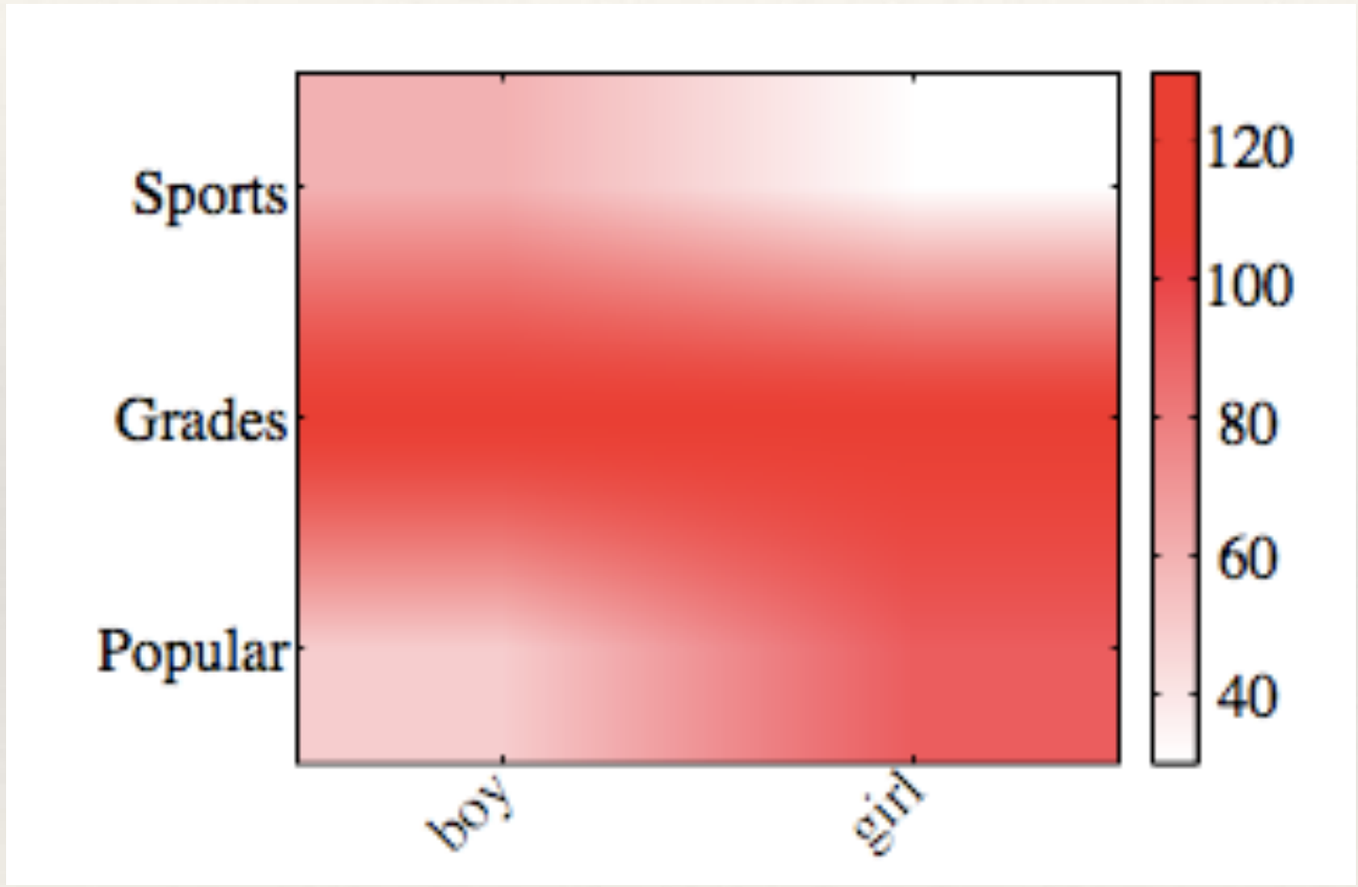
- An alternative is a pie chart
- Small differences may be hard to judge, though



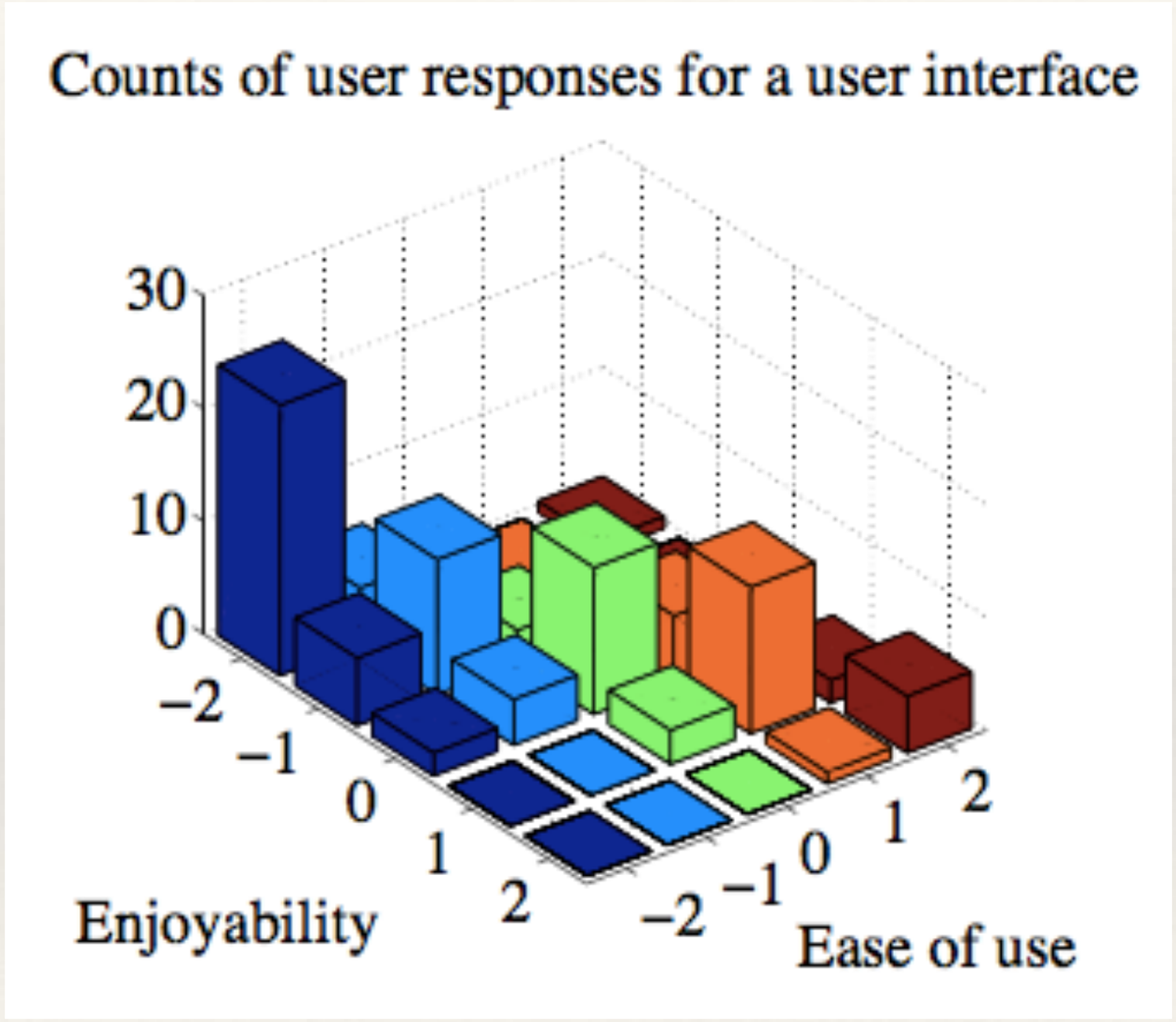
Stacked bar charts



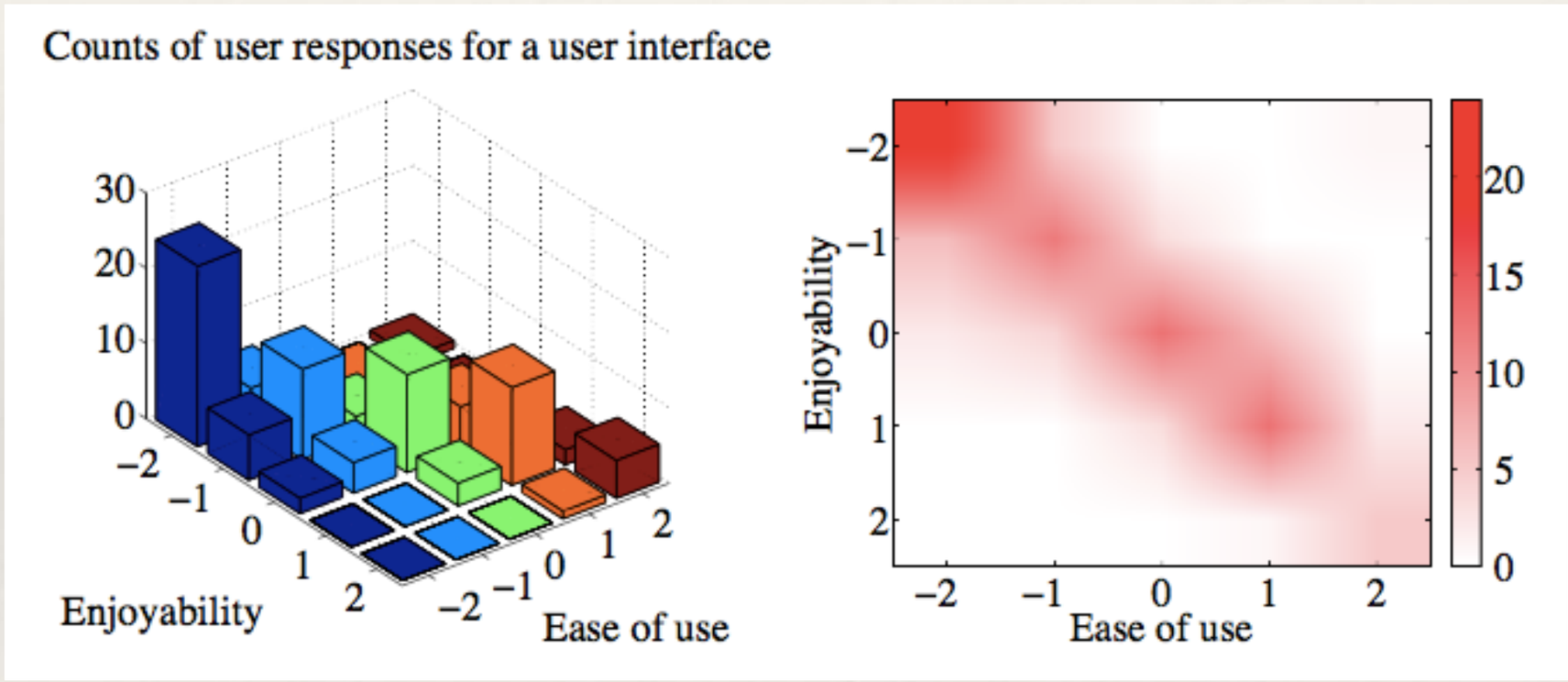
Heat Maps



3D bar charts

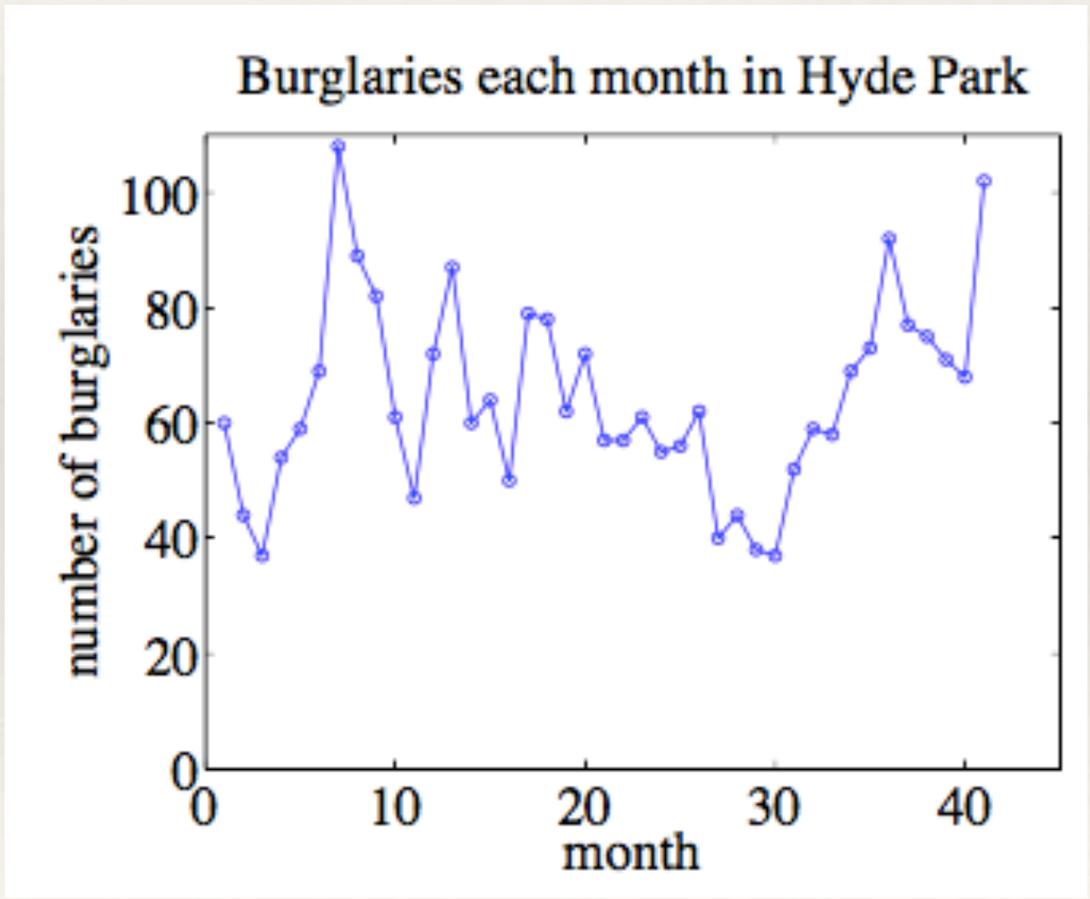


3D bar charts — occlusion

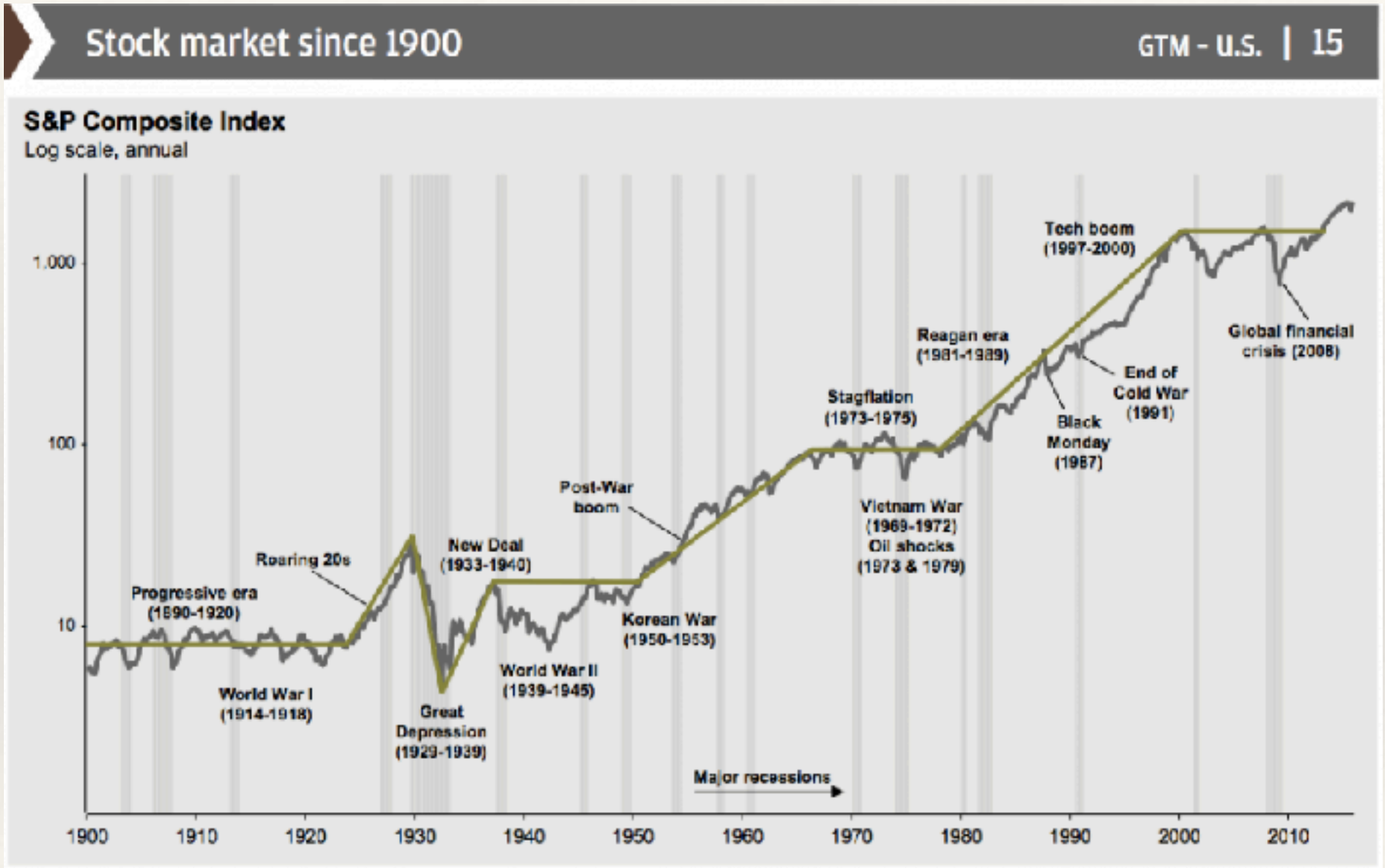


Time series data

- 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



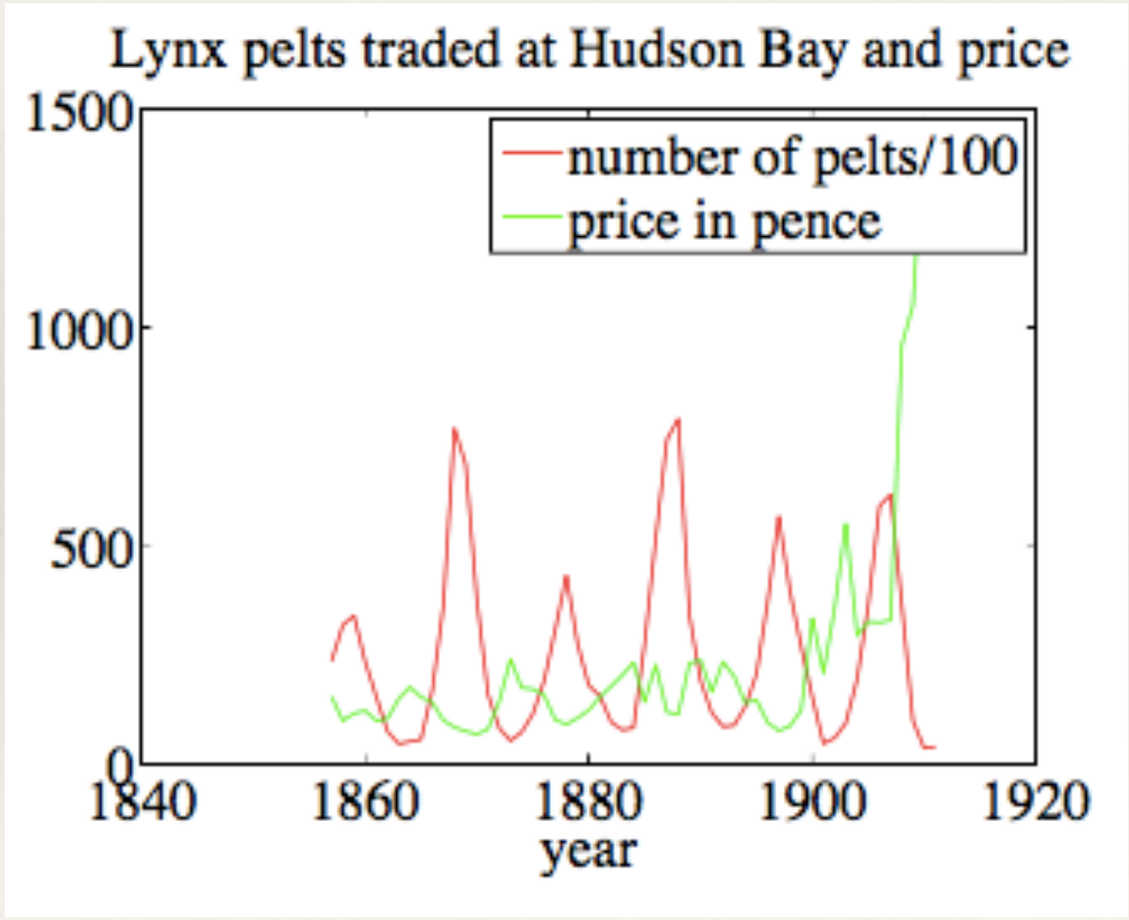
Time series



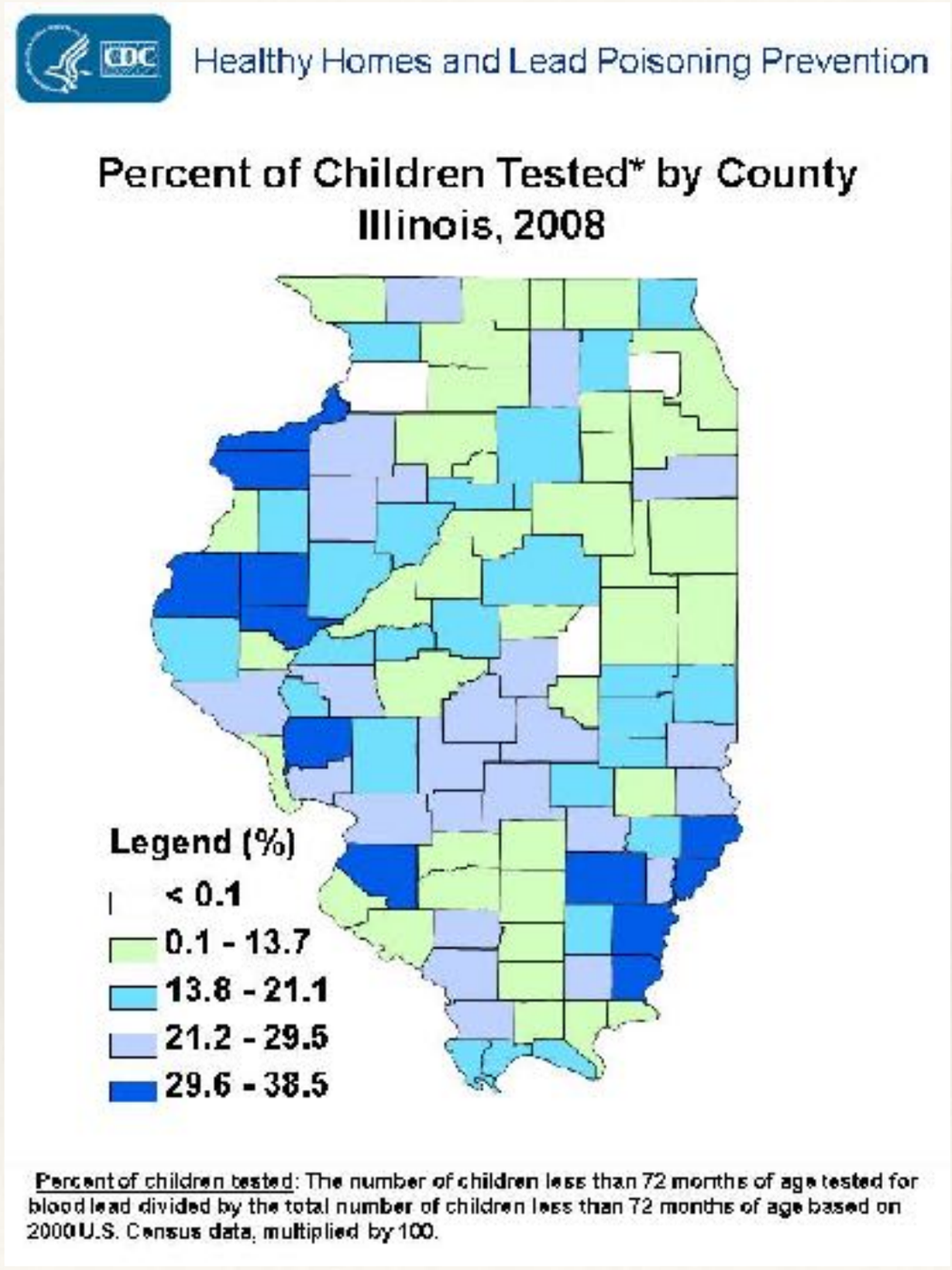
Source: FactSet, NBER, Robert Shiller, J.P. Morgan Asset Management.
Data shown in log scale to best illustrate long-term index patterns.

Relationships in time series

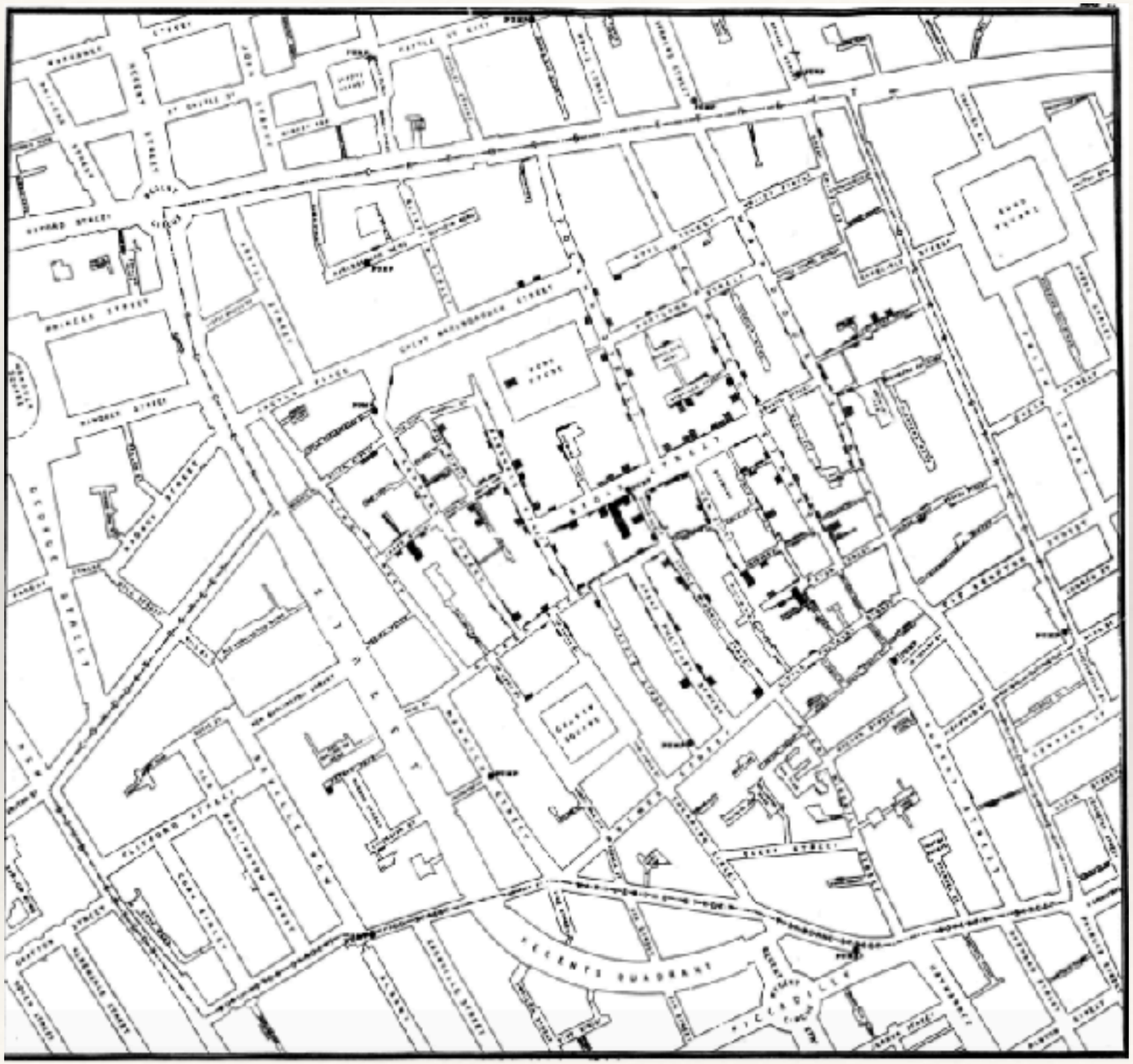
- We can visually inspect a relationship between two variables in a dataset with these plots
- Why might the number of pelts be periodic?
- The price?



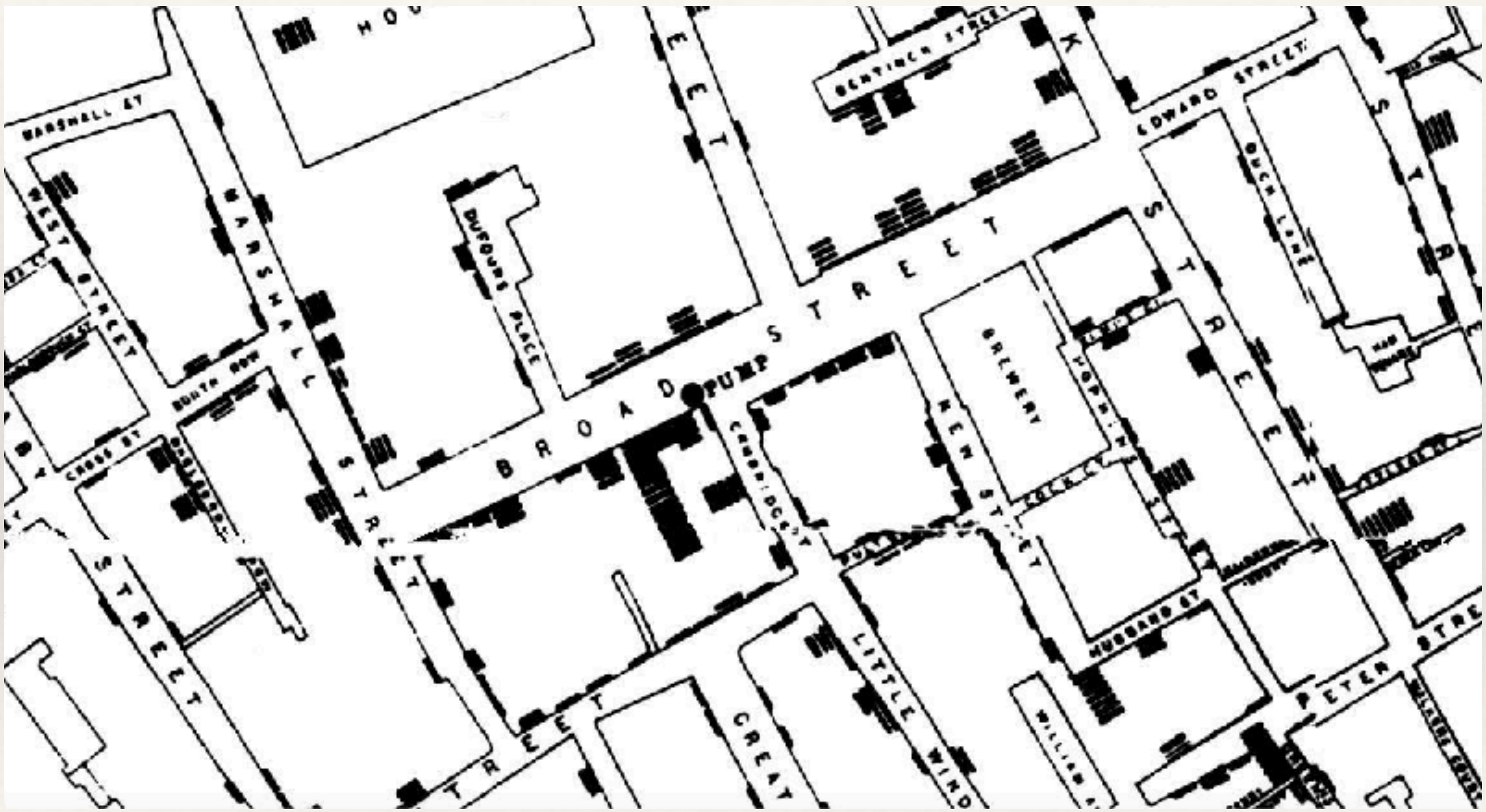
Plotting spatial data



Spatial data and Cholera

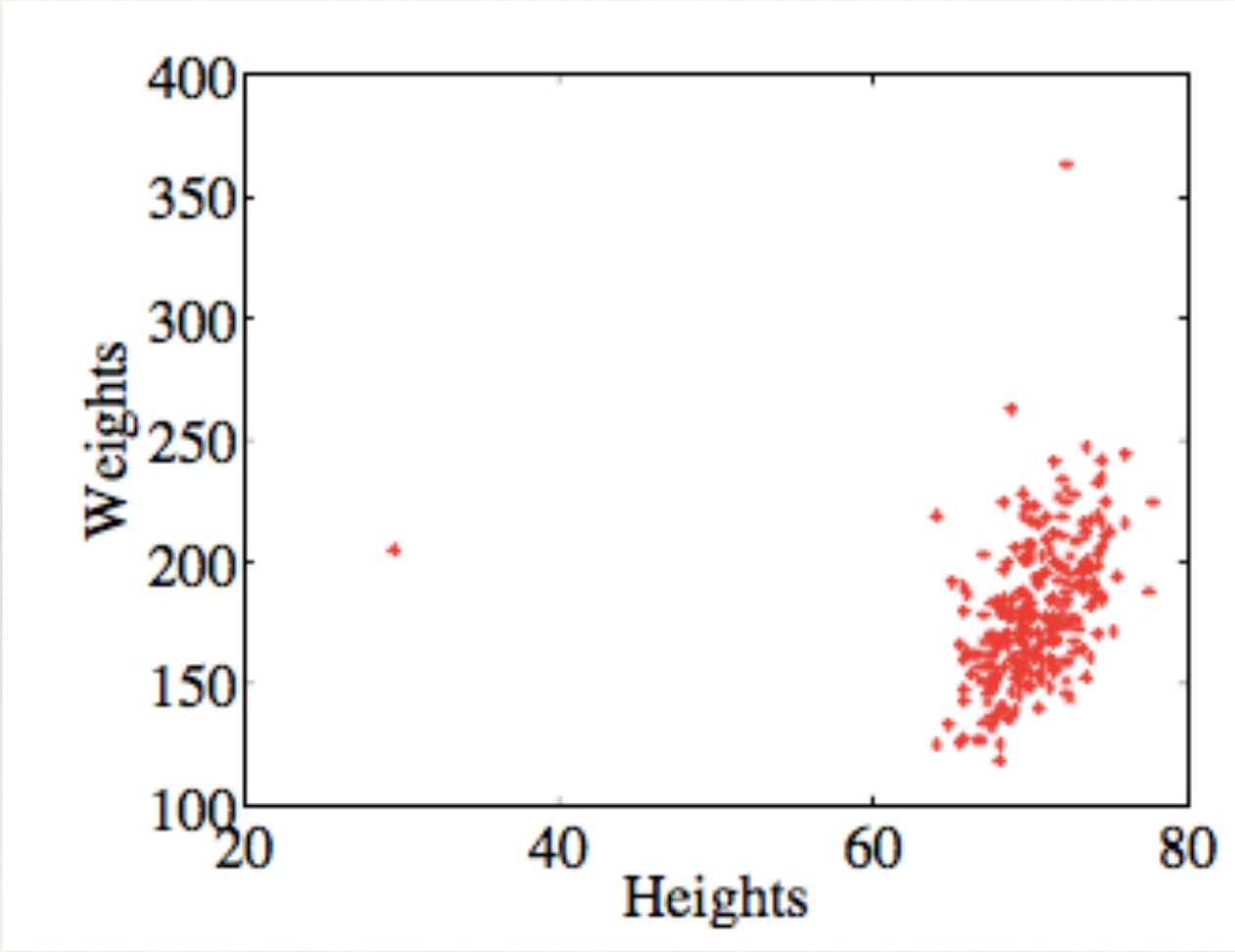


Spatial data



Visualizing two variables: scatterplots

- One numerical variable on each axis

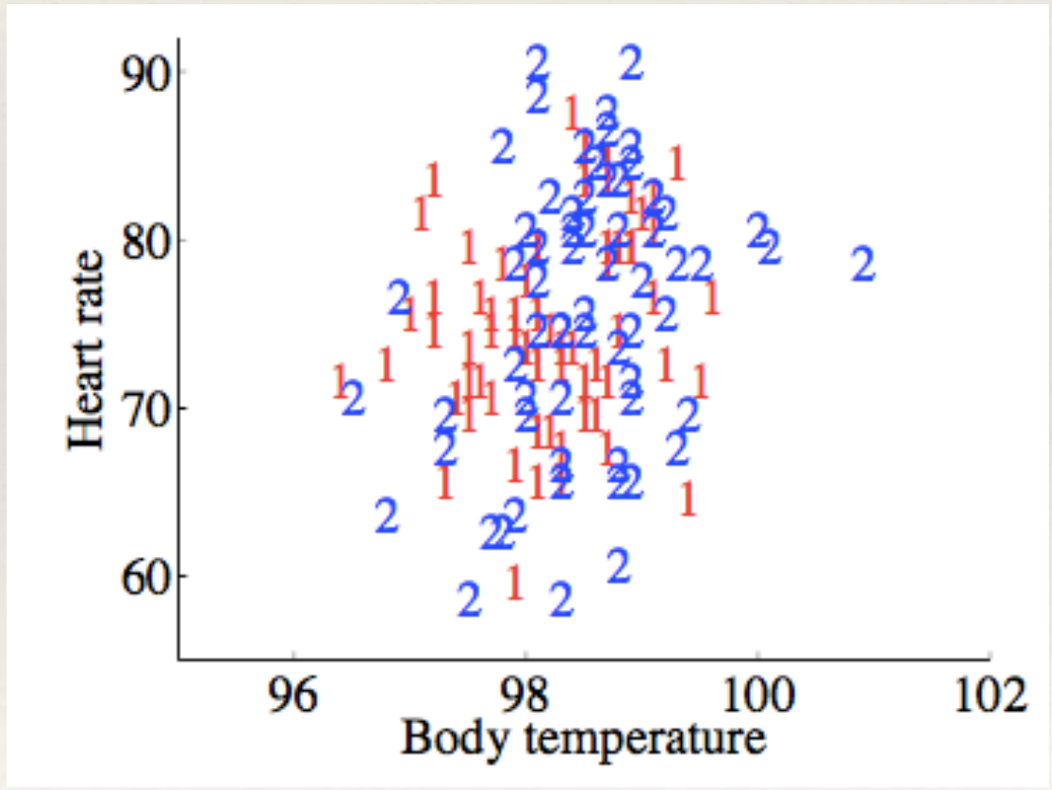


Scatterplots

- 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?)

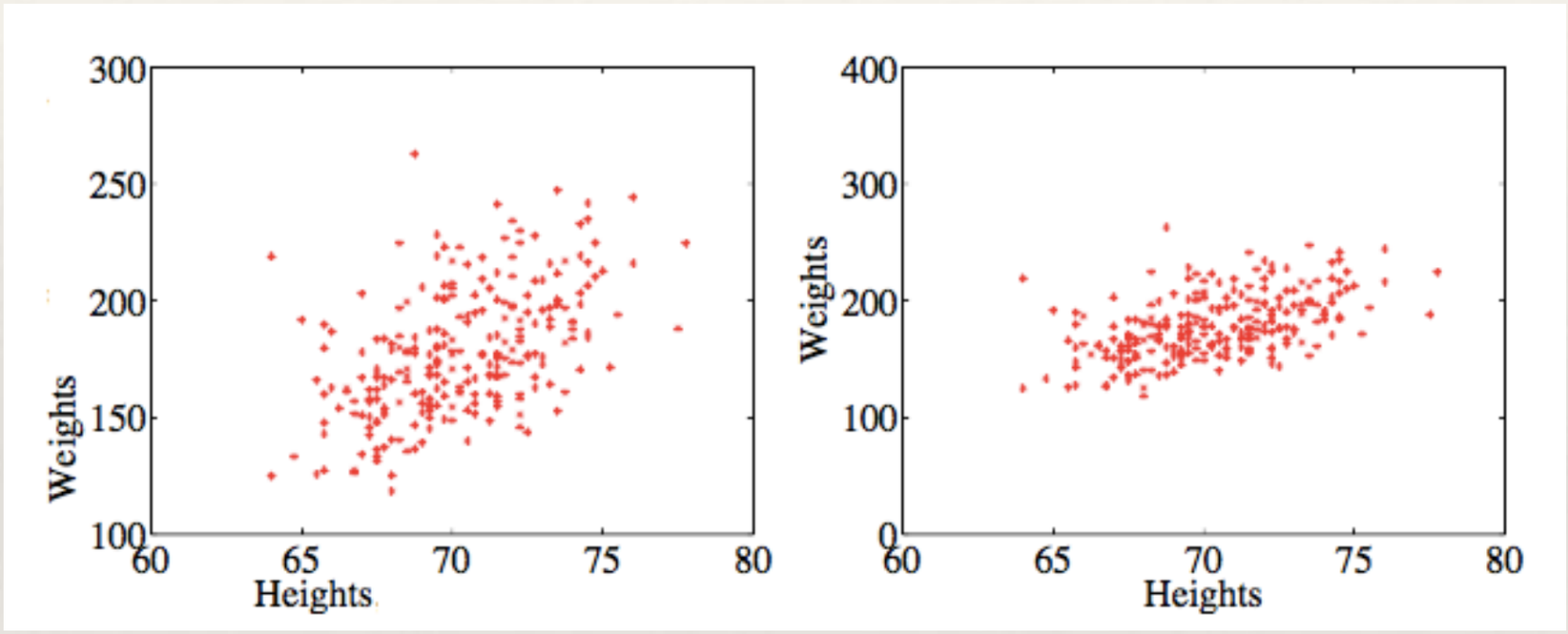
Including more information

- 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

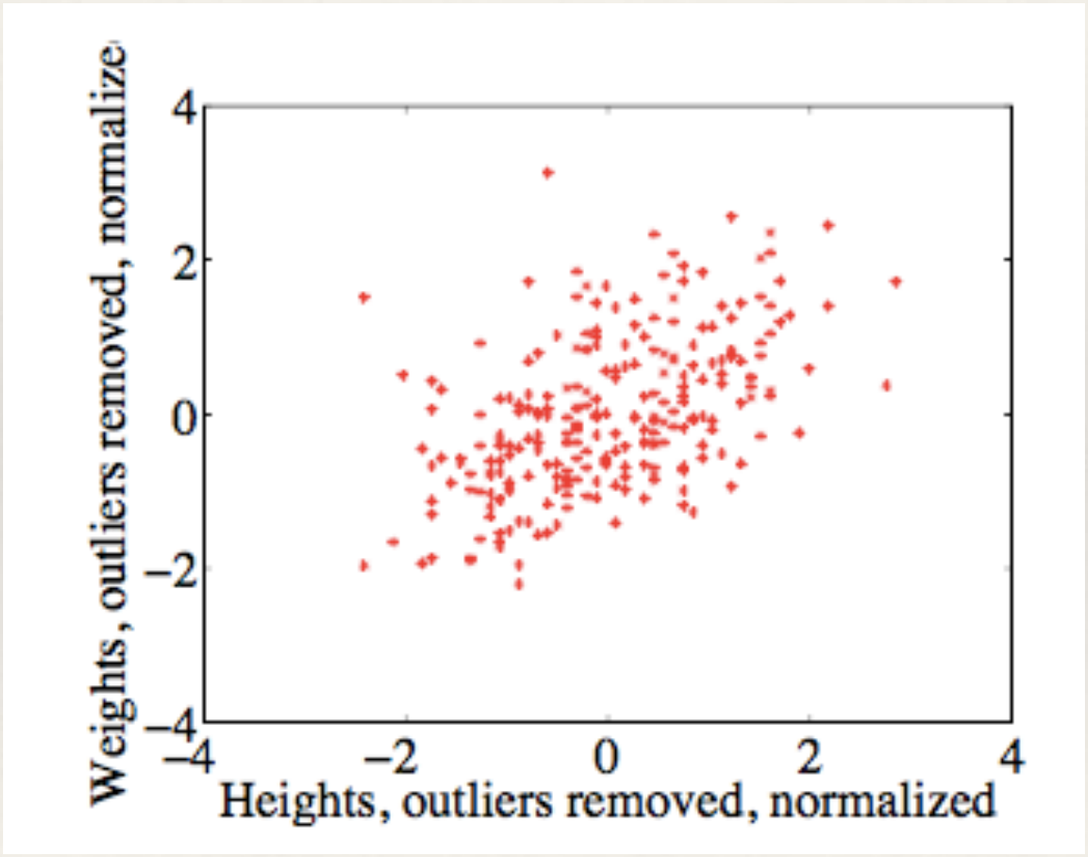
- The same dataset



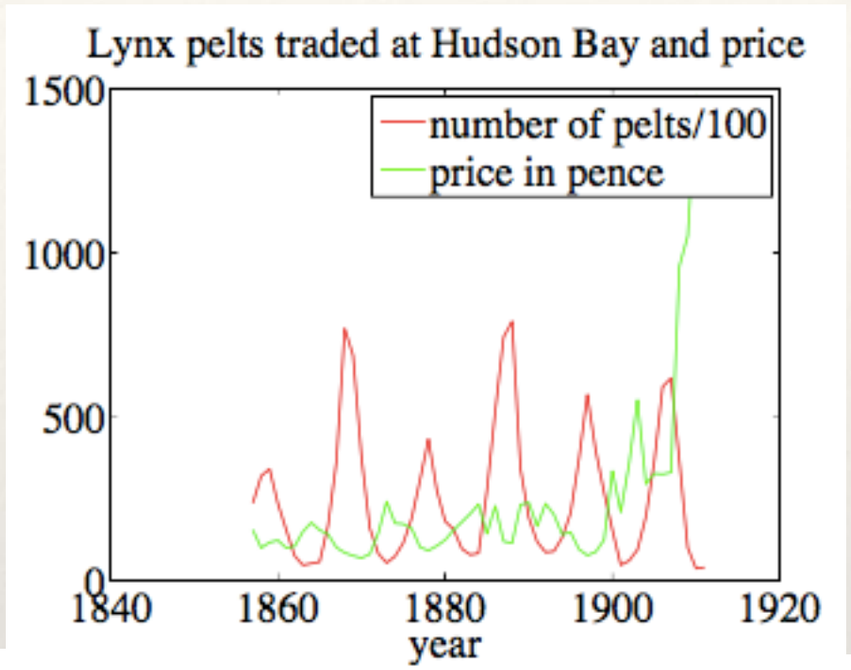
- Zooming out to capture outliers can make data look concentrated

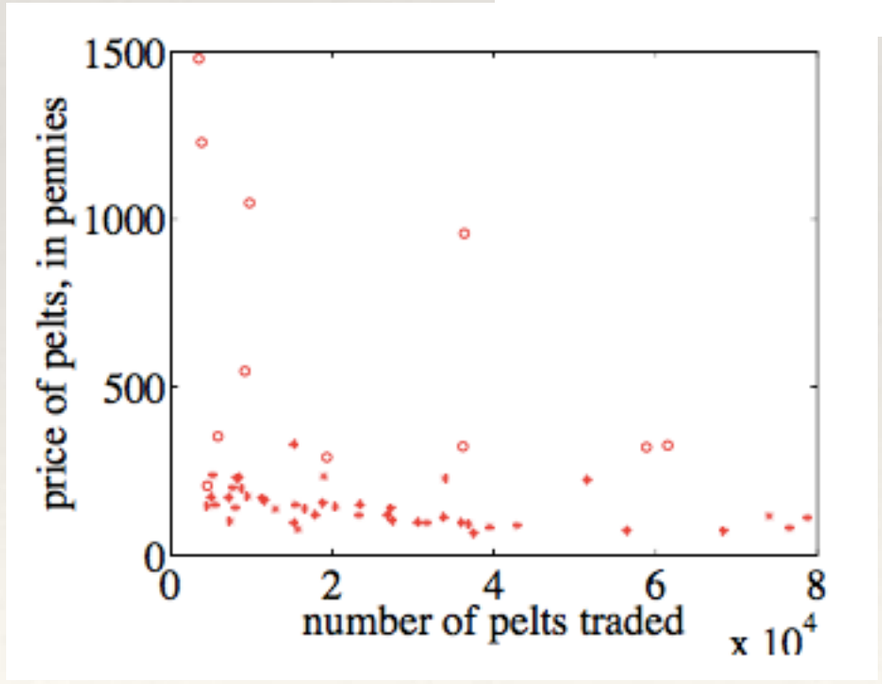
Normalization

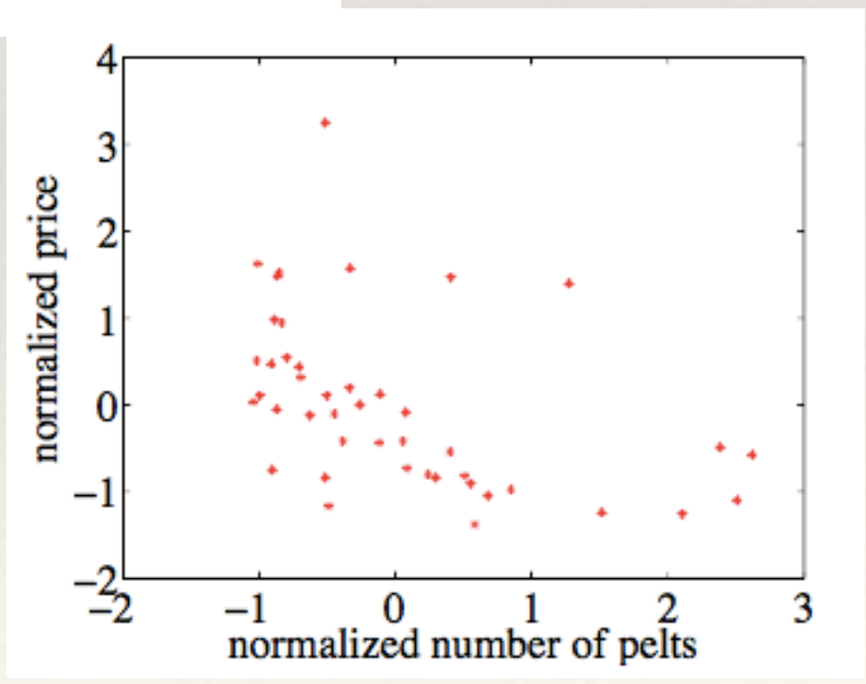
- We often normalize our data (put in standard coordinates) and remove outliers before plotting



Normalization





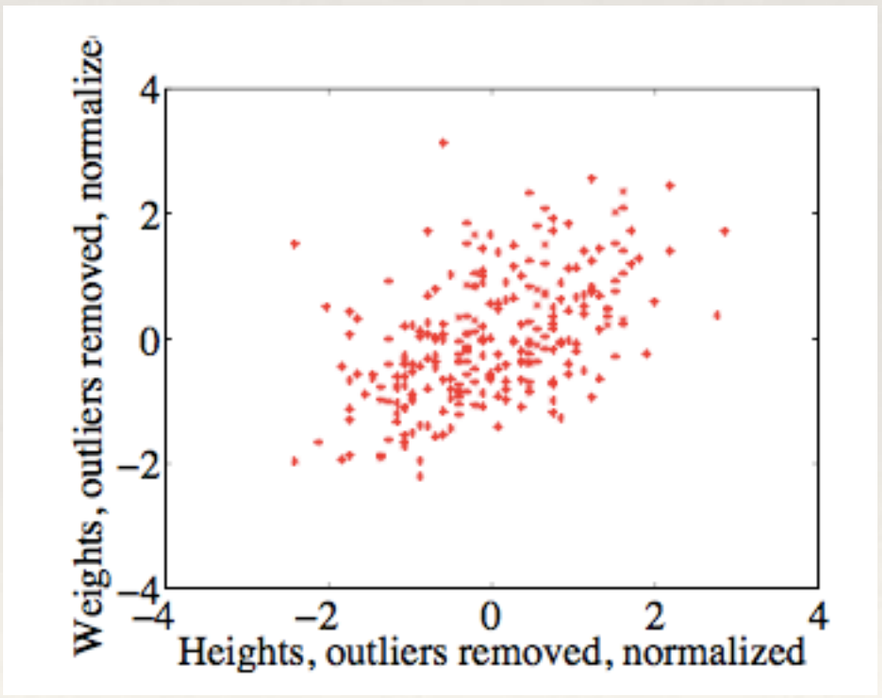


Scatterplots summary

- Should be the 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

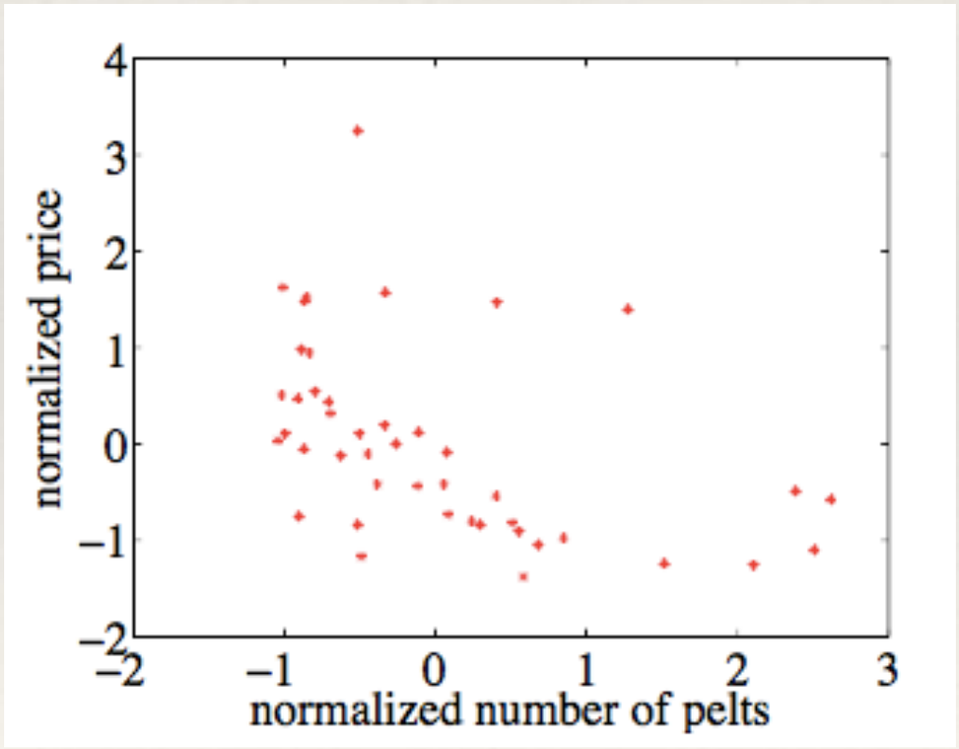
Correlation

- 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



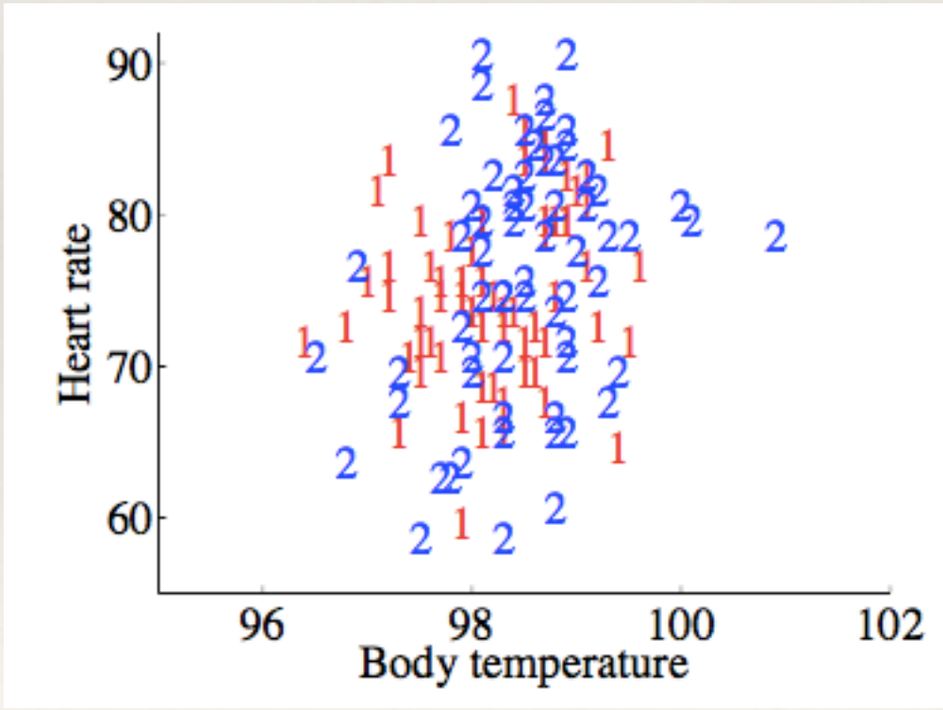
Correlation

- 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



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

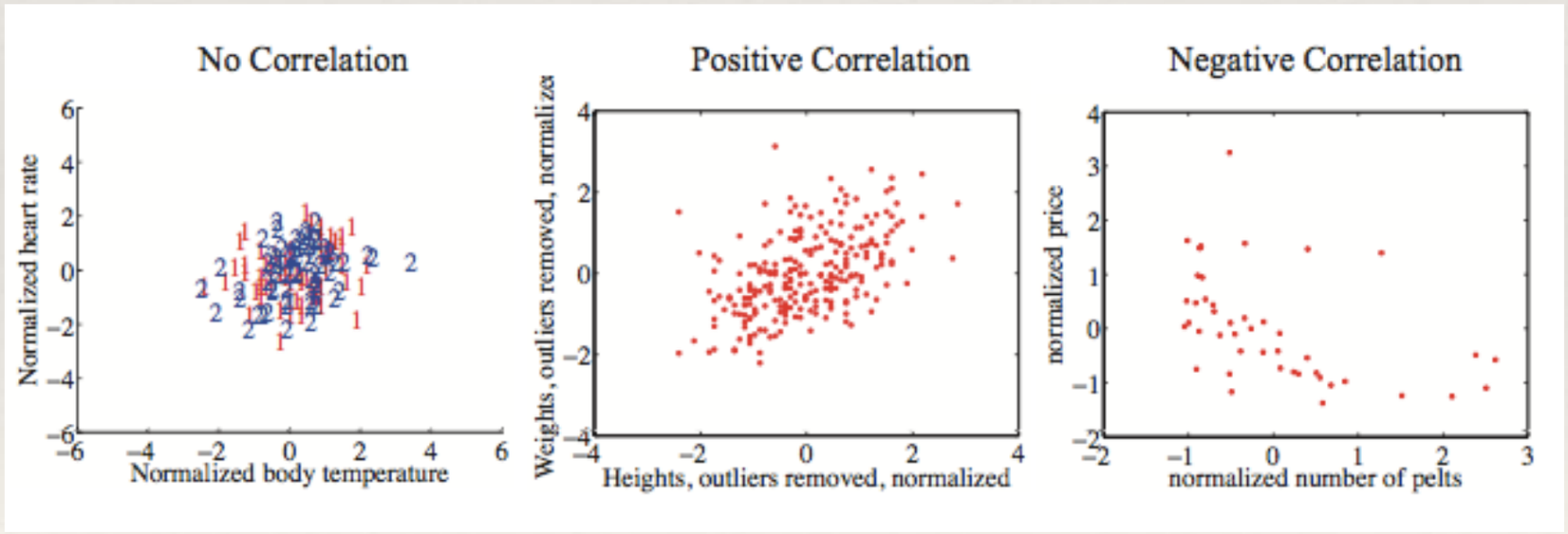
- Scaling may change the sign

$$\text{corr}\left(\{(ax+b,cx+d)\}\right) = \text{sign}(ab)\text{corr}\left(\{(x,y)\}\right)$$

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: $\mathrm{corr}(\{(x,y)\}) \leq 1$
- 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 $\mathbf{x} \bullet \mathbf{y} = \mathrm{corr}(\{(x,y)\})$
- 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 $\mathrm{std}(\{\hat{x}\})^2$

Recall

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

Proof

- Since $\{\hat{x}\}$ is our standardized dataset, we have $\text{std}(\{\hat{x}\})^2 = 1$
- Similar reasoning applies for y
- So we have that $\mathbf{x} \bullet \mathbf{y} \leq 1$
- And since $\mathbf{x} \bullet \mathbf{y} = \text{corr}(\{(x, y)\})$
- We've shown $\text{corr}(\{(x, y)\}) \leq 1$

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 $\hat{x}_i$

Prediction

- If we knew $\hat{y}_i$ we could define our prediction error as $u_i = \hat{y}_i - \hat{y}_i^p$
- *{u}* then is a dataset and we can perhaps use its mean and variance
- We will look at a simple predictor: a linear predictor
- So our prediction function will have the form

$$\hat{y}_i^p = a\hat{x}_i + b$$

Prediction

- The mean of our errors should be 0 or else we could reduce our prediction error by subtracting a constant
- Let's use this assumption to get a value for b in

$$\hat{y}_i^p = a\hat{x}_i + b$$

- Recalling that $u_i = \hat{y}_i - \hat{y}_i^p$ we have

$$\text{mean}(\{u\}) = \text{mean}(\{\hat{y} - \hat{y}^p\})$$

- Which we rewrite as

$$\text{mean}(\{u\}) = \text{mean}(\{\hat{y}\}) - \text{mean}(\{\hat{y}^p\})$$

Prediction

- Since $\hat{y}_i^p = a\hat{x}_i + b$ we rewrite

$$\text{mean}(\{u\}) = \text{mean}(\{\hat{y}\}) - \text{mean}(\{\hat{y}^p\})$$

- As

$$\text{mean}(\{u\}) = \text{mean}(\{\hat{y}\}) - \text{mean}(\{a\hat{x} + b\})$$

- Using what we know about the mean we rewrite

$$\text{mean}(\{u\}) = \text{mean}(\{\hat{y}\}) - a\,\text{mean}(\{\hat{x}\}) + b$$

Prediction

- Since we are in standard normal coordinates, we can simplify

$$\text{mean}(\{u\}) = \text{mean}(\{\hat{y}\}) - a\,\text{mean}(\{\hat{x}\}) + b$$

- To $\text{mean}(\{u\}) = 0 - a0 + b$
- Recall that we said we wanted the mean of our error to be 0, so we can solve for b above and get $b=0$

Prediction

- So now we have a predictor with the form $\hat{y}^p = a\hat{x}$
- Ideally our error would have 0 mean and 0 standard deviation which means we always predict exactly correctly
- Let's try and choose an a that minimizes standard deviation
- But since we want to keep our math simple, let's equivalently find an a that minimizes the variance of the error
- In order to have a shorthand available to us, let's write r for the correlation of x and y

Prediction

- We want to minimize $\mathrm{var}(\{u\}) = \mathrm{var}(\{\hat{y} - \hat{y}^p\})$
- Which we rewrite as $\mathrm{var}(\{u\}) = \mathrm{var}(\{\hat{y} - a\hat{x}\})$
- Remember that we can write the variance as the mean of some quantity: the squared distances from the mean of the data (in this case mean(u) = 0)

$$\mathrm{var}(\{u\}) = \mathrm{mean}(\{(\hat{y} - a\hat{x} - 0)^2\})$$

Prediction

- Expanding $\text{var}(\{u\}) = \text{mean}(\{(\hat{y} - a\hat{x})^2\})$ we get

$$\text{var}(\{u\}) = \text{mean}(\{(\hat{y})^2 - 2a\hat{x}\hat{y} - a^2(\hat{x})^2\})$$

- Which we rewrite as

$$\text{mean}(\{(\hat{y})^2\}) - 2a\,\text{mean}(\{\hat{x}\hat{y}\}) + a^2\,\text{mean}(\{(\hat{x})^2\})$$

- Each of these terms has something we recognize

$$\text{mean}(\{(\hat{y})^2\}) = (\text{std}(\{\hat{y}\}))^2 \qquad \text{mean}(\{\hat{x}\hat{y}\}) = \text{corr}(\{(x, y)\})$$

$$\text{mean}(\{(\hat{x})^2\}) = (\text{std}(\{\hat{x}\}))^2 \qquad \text{mean}(\{\hat{x}\hat{y}\}) = r$$

Prediction

- So we can simplify

$$\text{mean}(\{(\hat{y})^2\}) - 2a\,\text{mean}(\{\hat{x}\hat{y}\}) + a^2\,\text{mean}(\{(\hat{x})^2\})$$

- As $\text{var}(\{u\}) = 1 - 2ar + a^2$

- Since we are searching for an a that minimizes variance, we take the derivative and set equal to 0 and get

$$2r + 2a = 0$$

- So we will use $a=r$ in our predictor

Summary of what we proved

- We wanted a way of predicting y from x
- We chose to think in standard coordinates and to use a linear predictor of the form $\hat{y}_i^p = a\hat{x}_i + b$
- Assuming the mean of the error was 0 gave us $b=0$
- Minimizing the variance of the error gave us $a=r$
- So our final predictor is

$$\hat{y}_i^p = r\hat{x}_i$$

Prediction

- So here is our process for predicting y_0 from x_0

- Transform the data set into standard coordinates, to get

$$\hat{x}_i = \frac{1}{\text{std}(x)}(x_i - \text{mean}(\{x\}))$$

$$\hat{y}_i = \frac{1}{\text{std}(y)}(y_i - \text{mean}(\{y\}))$$

$$\hat{x}_0 = \frac{1}{\text{std}(x)}(x_0 - \text{mean}(\{x\})).$$

- Compute the correlation

$$r = \text{corr}(\{(x, y)\}) = \text{mean}(\{\hat{x}\hat{y}\}).$$

- Predict $\hat{y}_0 = r\hat{x}_0$.
- Transform this prediction into the original coordinate system, to get

$$y_0 = \text{std}(y) r \hat{x}_0 + \text{mean}(\{y\})$$