Styles of data analysis

DAAG Chapter 2

Objectives

- Learn the common tools of Exploratory Data Analysis
  - Histograms, density plots, boxplots
  - Scatterplots and scatterplot matrices
  - Data summaries
- Learn about what to look for and what can go wrong
  - Outliers, skewness, clustering
  - Non-linearity, heteroscedasticity
- Be mindful of good statistical practice, overreaching, overfitting, …

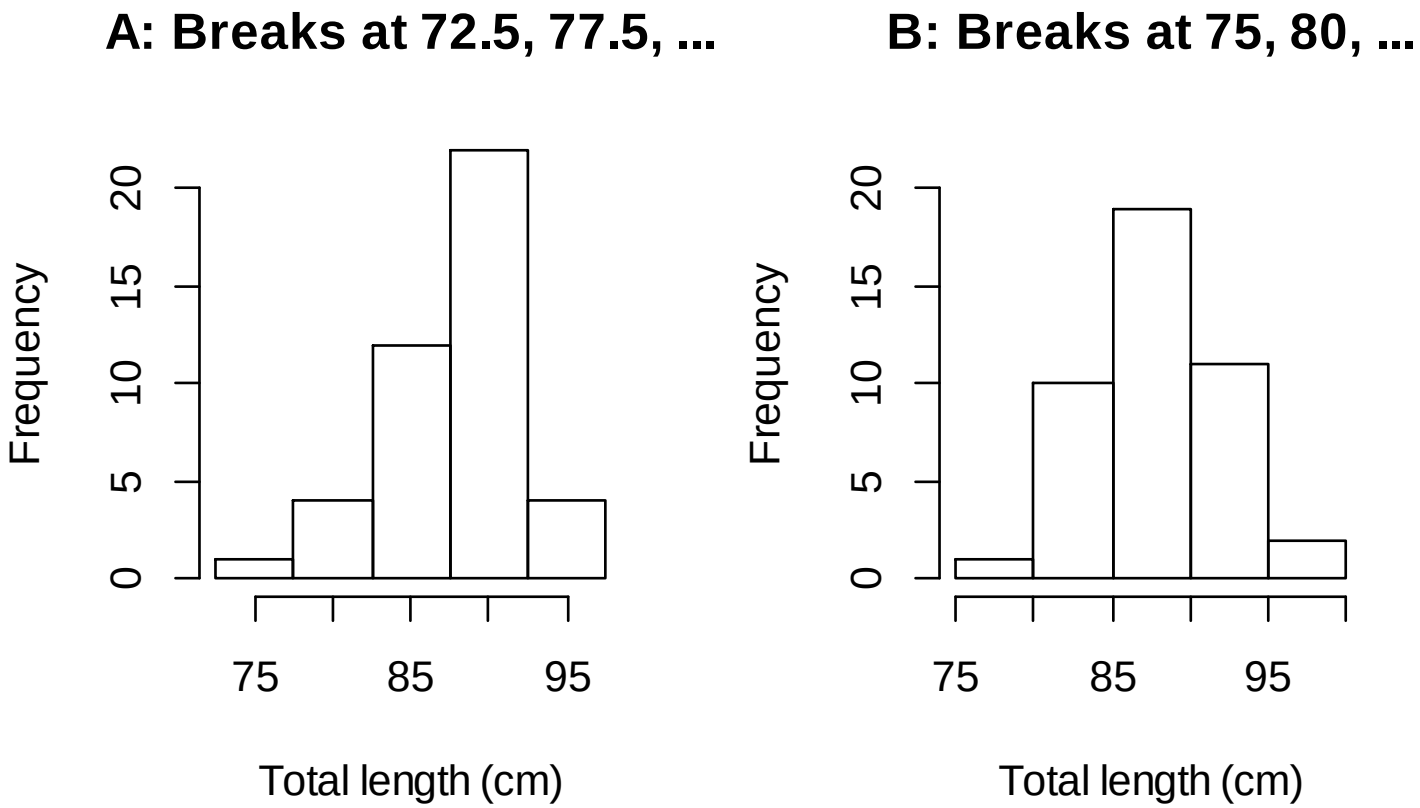
What is the first rule of data analysis?

Plot your data!

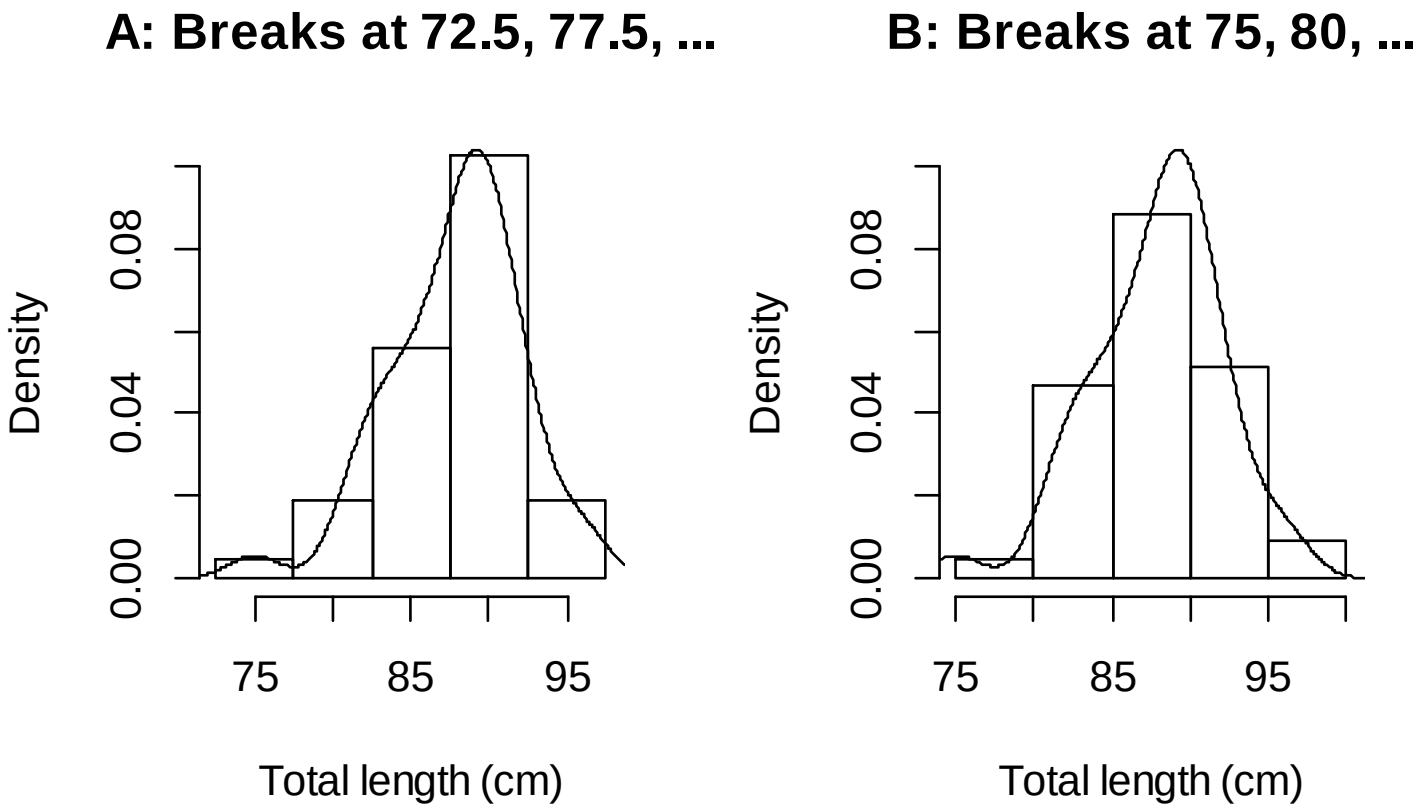
Exploratory data analysis

- Formalized by John Tukey
  - Guiding principle: let the data speak for themselves
- Why do EDA?
  - Suggest new ideas or understandings
  - Reveal problematic assumptions made before data collection
  - Check on assumptions to be made in subsequent analysis
  - Suggest future research questions or directions

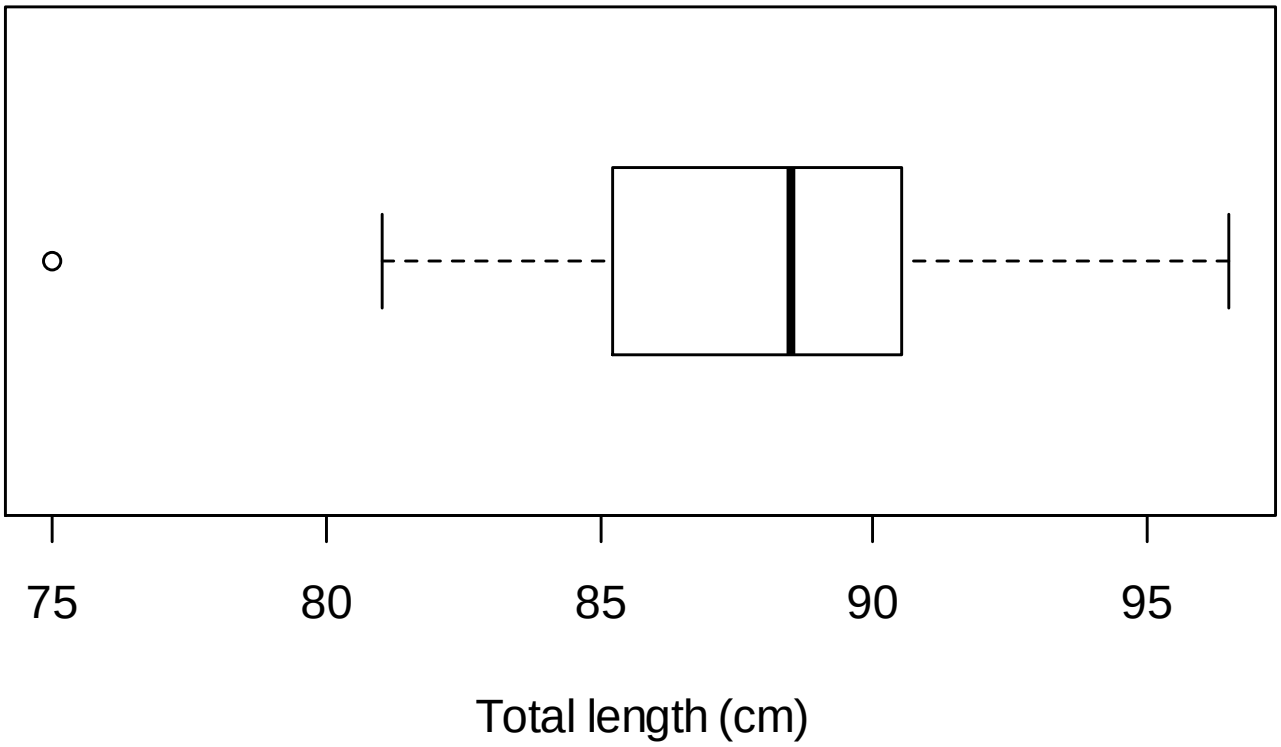
Plots for a single variable



Plots for a single variable



Plots for a single variable

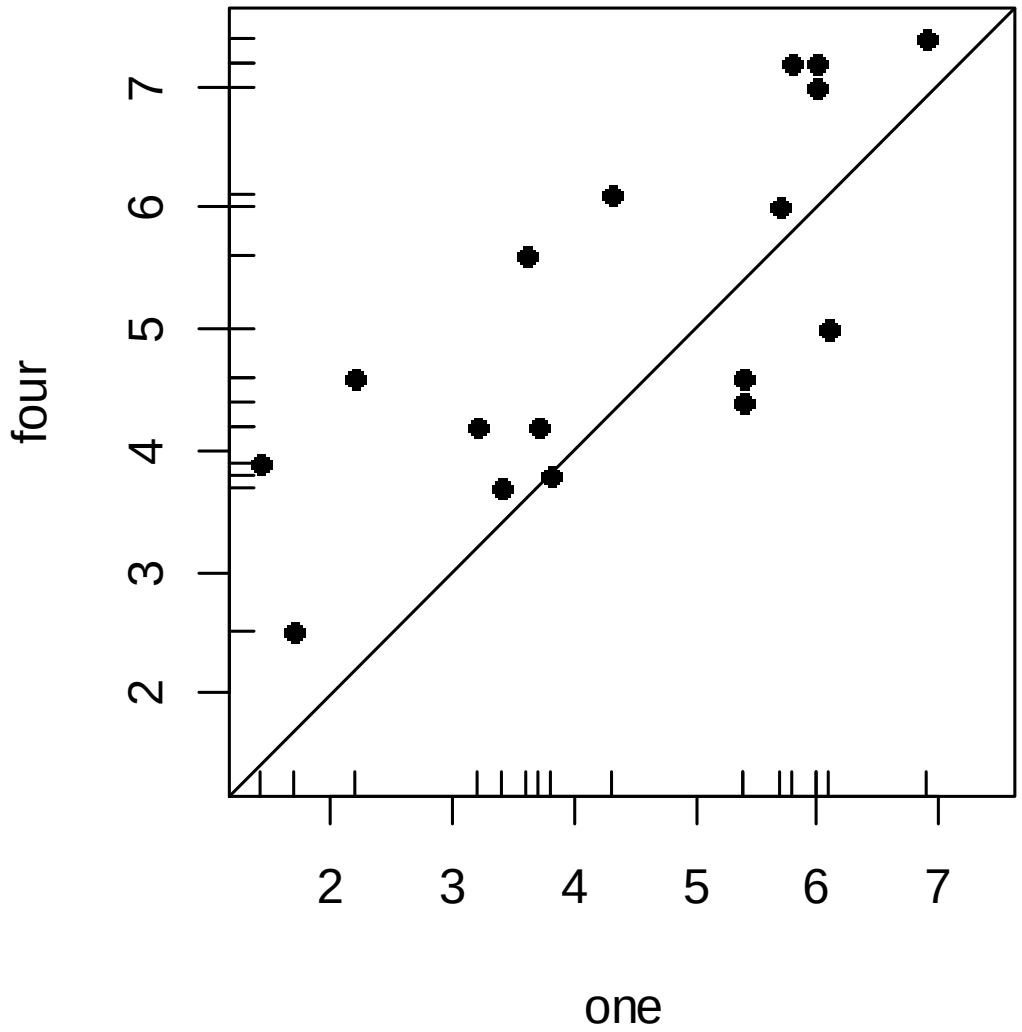


Plots for bivariate data

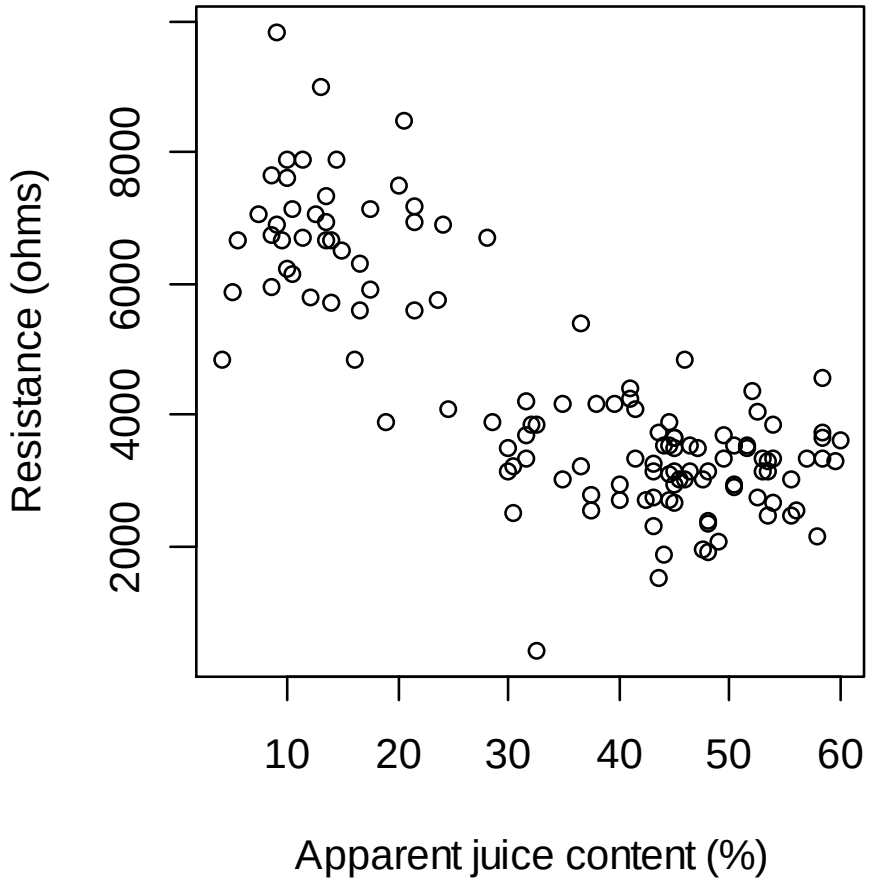
- Experiment with 17 tasters
  - Milk sample with 1 unit of sweetener
  - Milk sample with 4 units of sweetener
- Each person rated the sweetness of the two samples

Plots for bivariate data

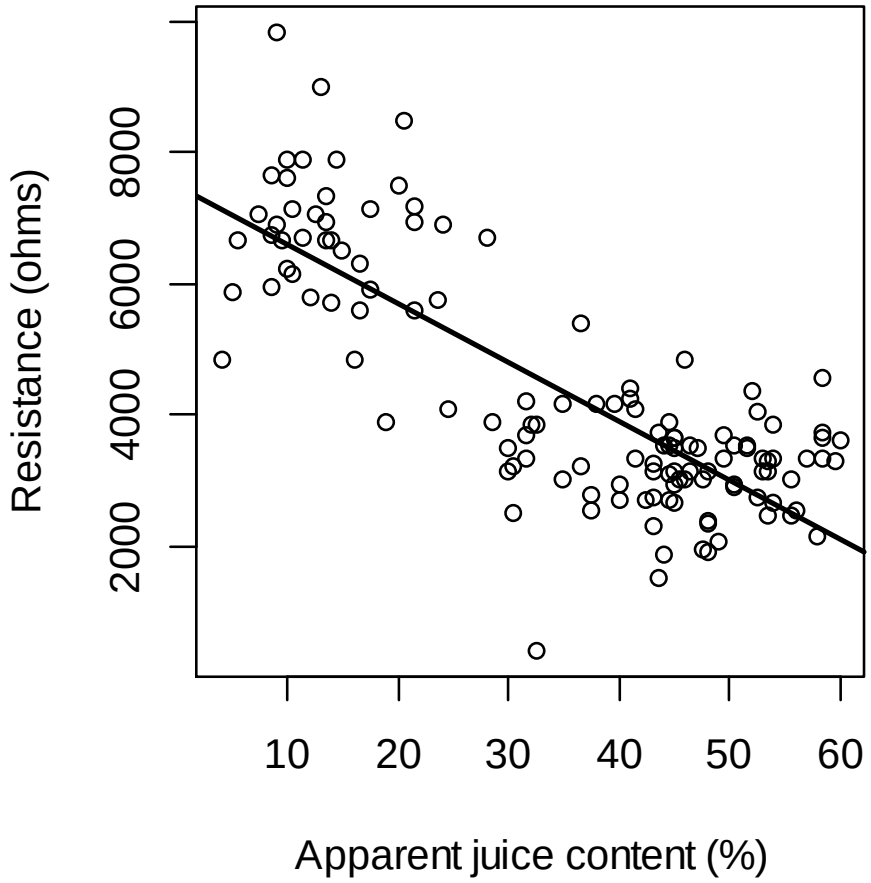
- 1:1 plot ratio
- Rug shows where points lie on the axis
- Most people think “four” is sweeter
- There is a positive relationship between ratings



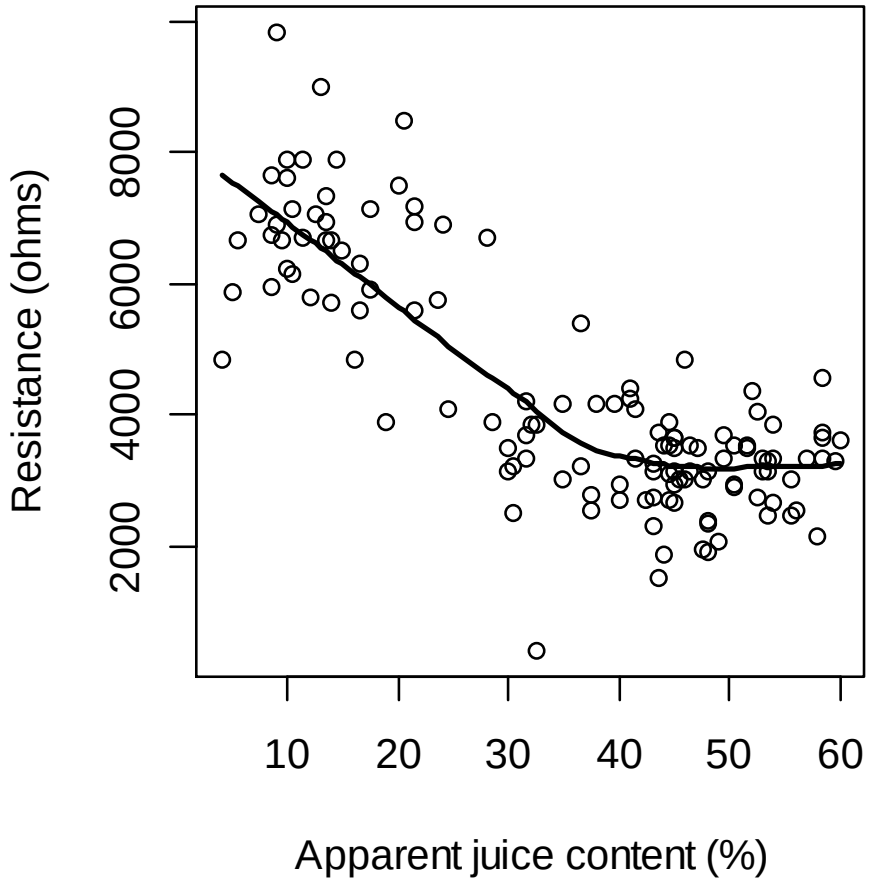
Plots for bivariate data



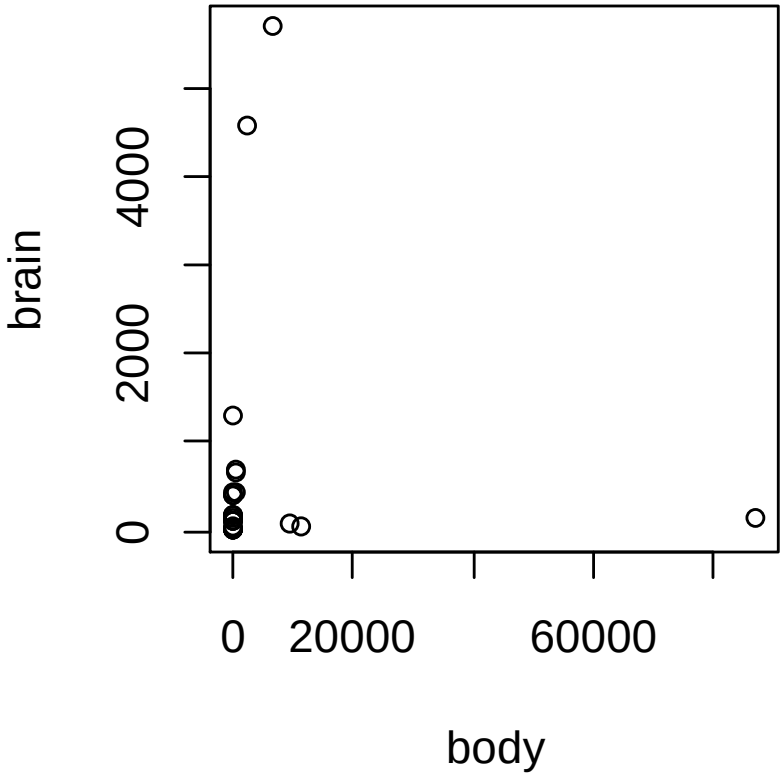
Plots for bivariate data

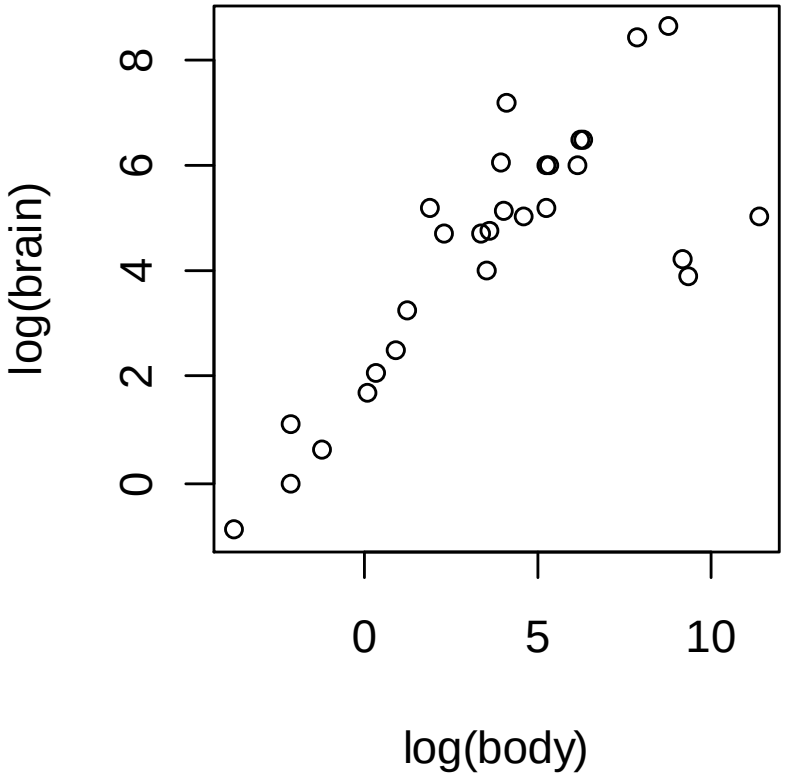


Plots for bivariate data



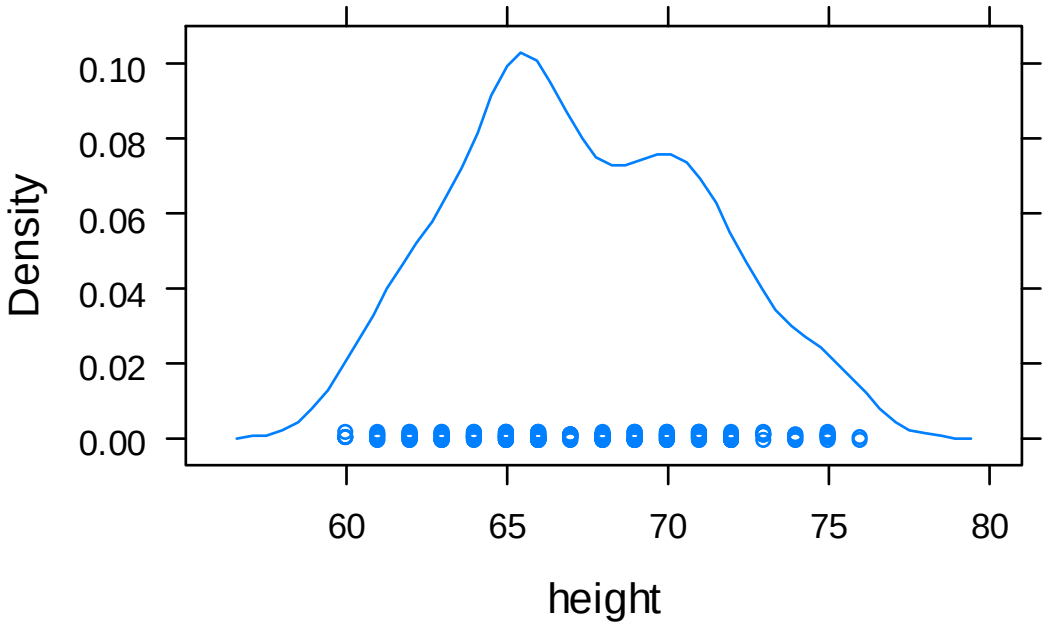
Plots for bivariate data



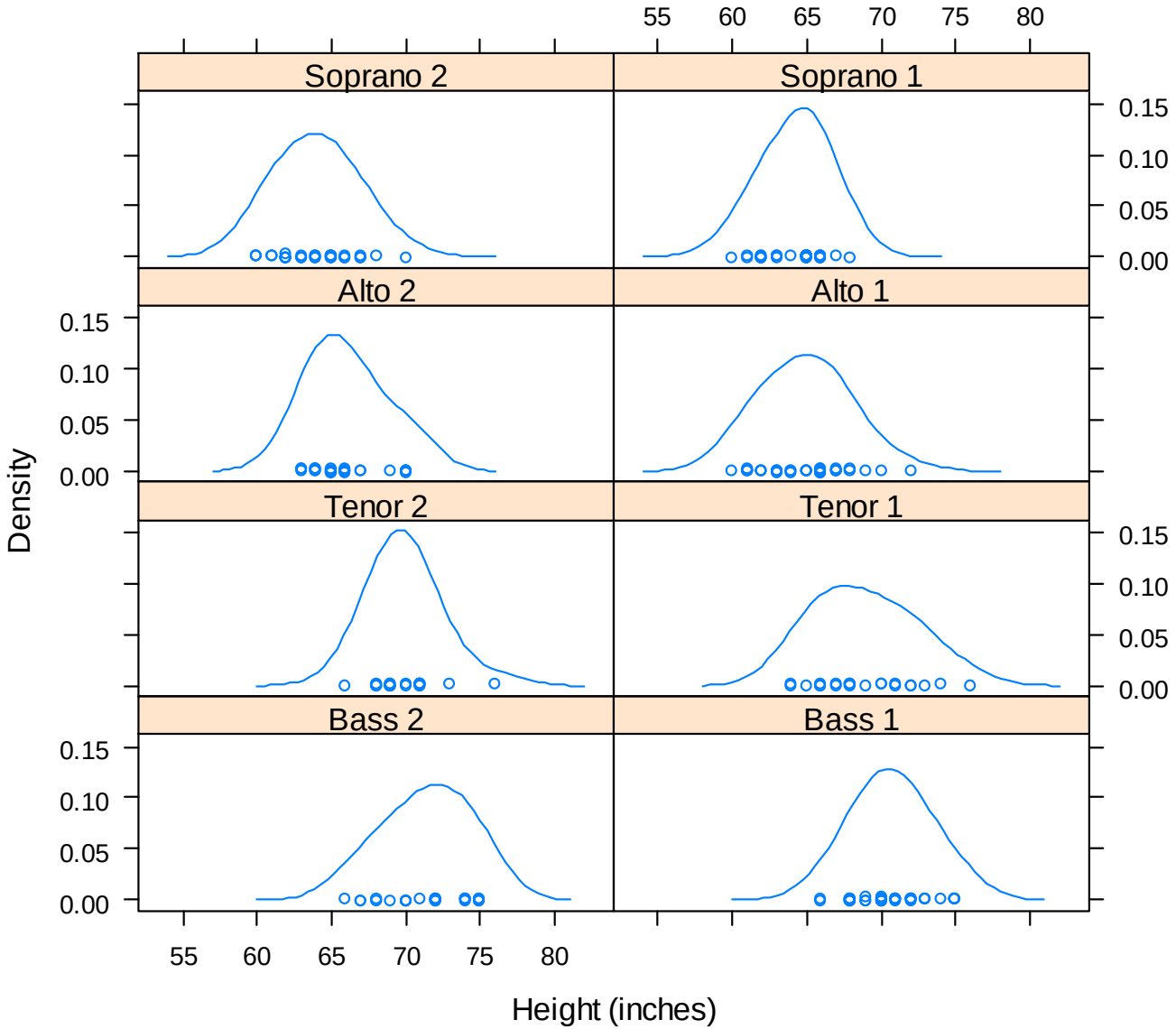


Clustering

- Heights in inches of the singers in the New York Choral Society in 1979

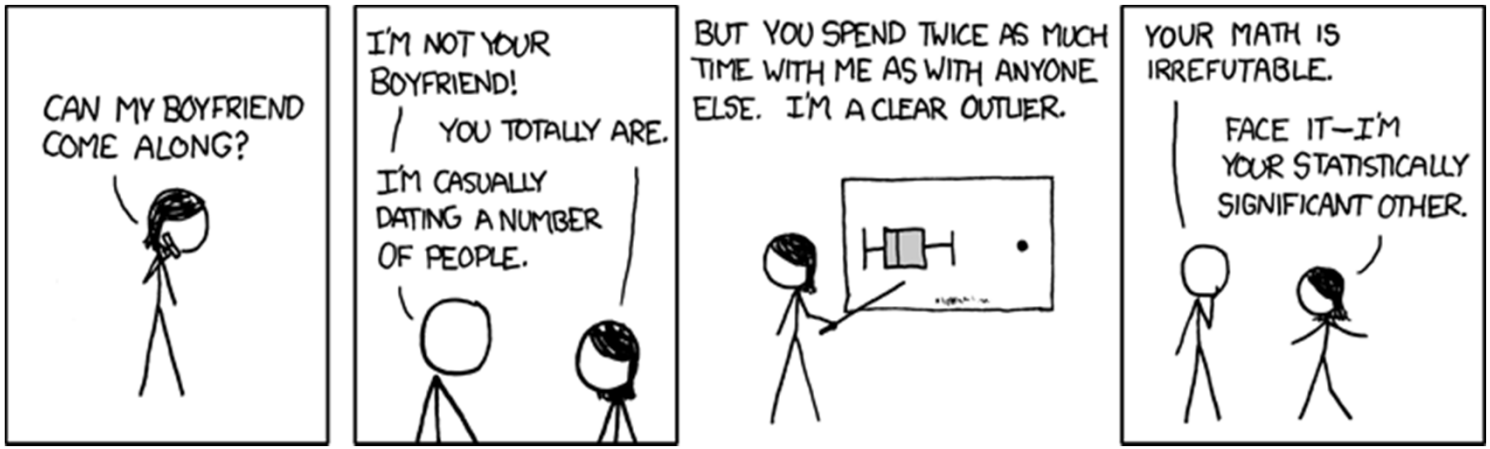


Clustering


Soprano 2
Soprano 1
Alto 2
Alto 1
Tenor 2
Tenor 1
Bass 2
Bass 1
Density
Height (inches)
55
60
65
70
75
80
0.00
0.05
0.10
0.15

Outliers

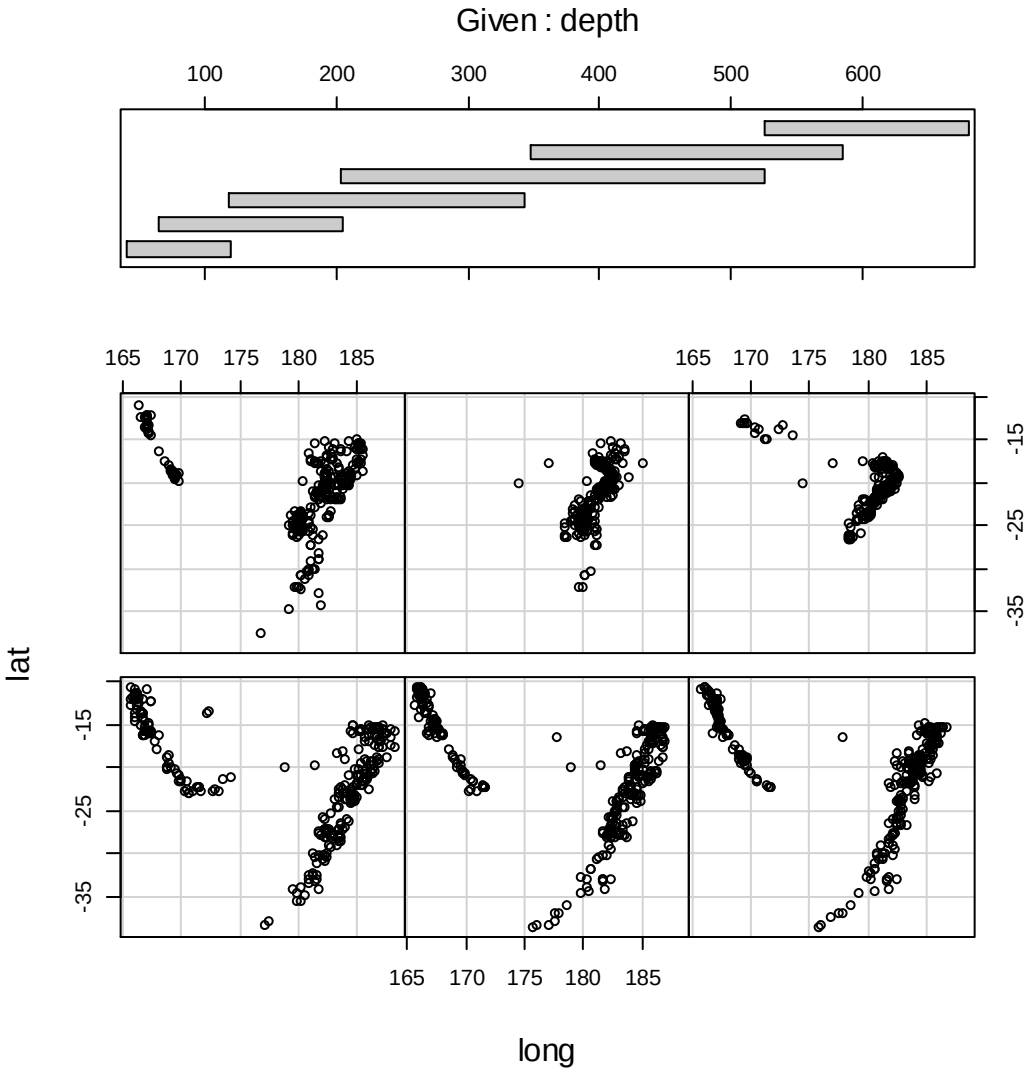
- Require special treatment
- Could be highly influential in subsequent modeling
- Could suggest new understanding



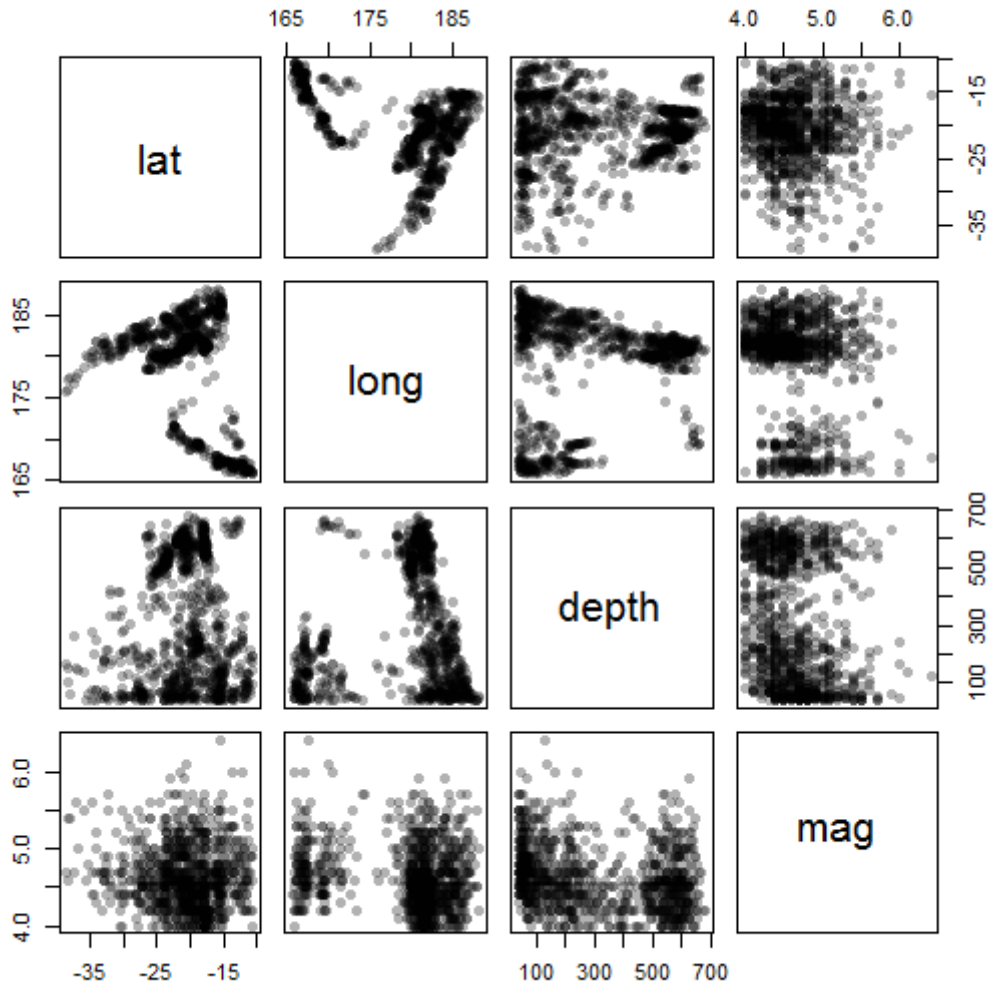
xkcd.com

Conditioning plots

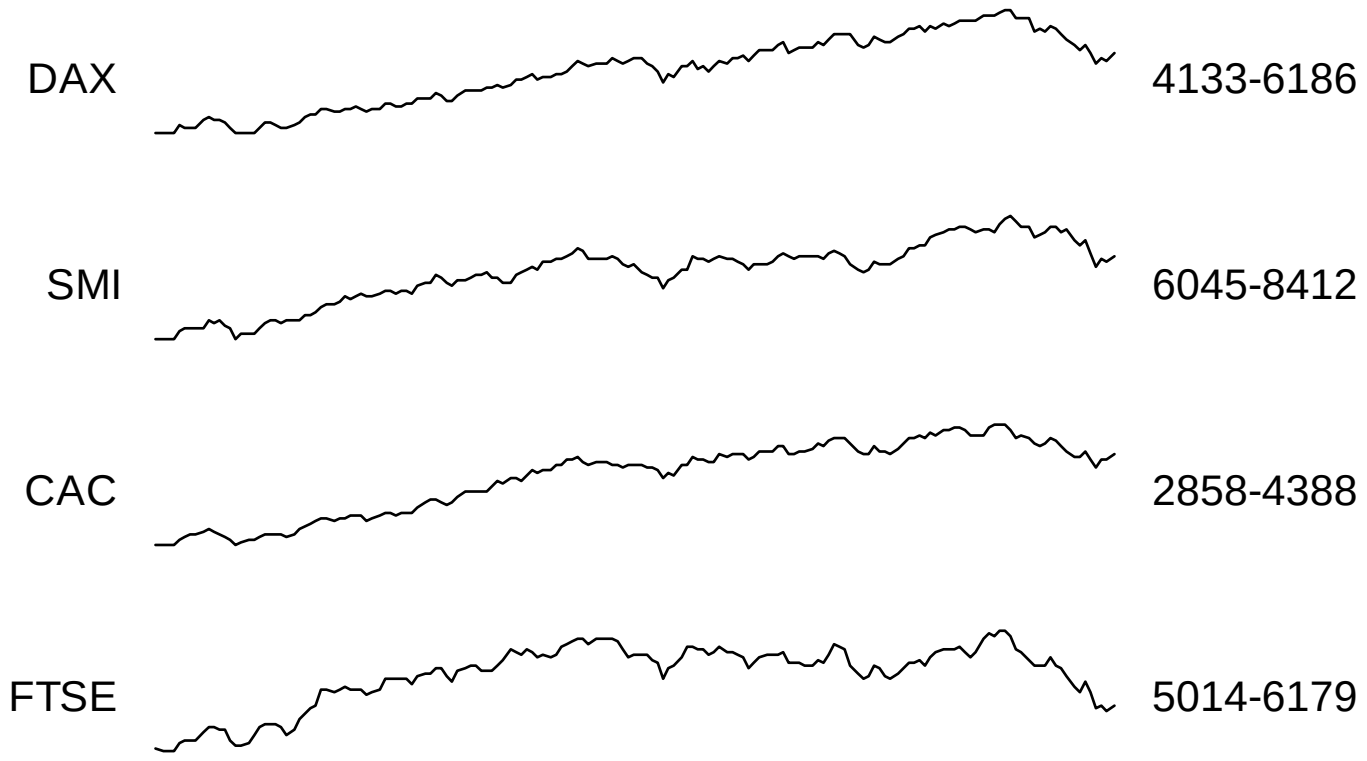
- Earthquake data from a location near Fiji
- Depth in km
- Data since 1964



Scatterplot matrix



Sparklines



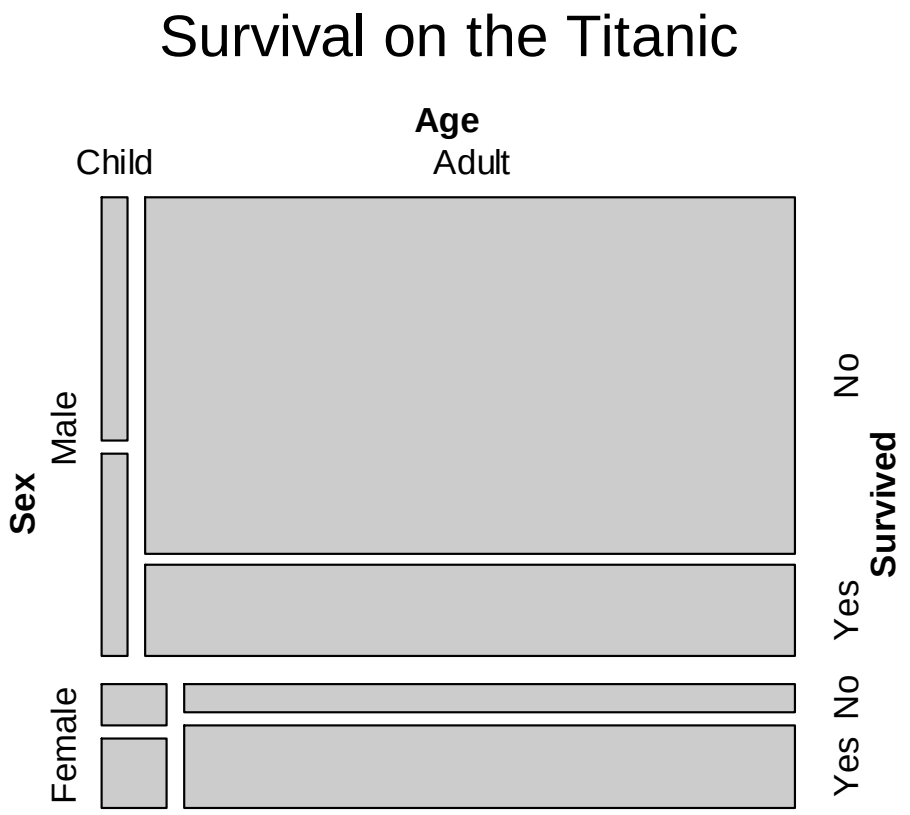
EU daily closing price indices: 1998

(Sparklines R code)

```
EU <- window( EuStockMarkets, start = 1998 )
par( mfcol = c(4,1), mar = c(1,5,1,8)+0.1, oma = c(2,0,0,0) )
for( i in 1:4 ){
  plot( EU[,i], axes = FALSE, xlab = "", ylab = "" )
  rr <- range( EU[,i] )
  mtext( paste( round(rr), collapse="-" ), 4, las = 1 )
  mtext( colnames(EU)[i], 2, las = 1 )
}
mtext("EU daily closing price indices: 1998",1,outer=TRUE, line=0)
```

Summary statistics

- Central tendency: mean, median, mode, …
- Dispersion: standard deviation, IQR, range, …
- Counts by group or category



The data analysis process

- Moving from EDA into more directed data analysis, we begin to ask questions of the data
  - Questions motivated by scientific understanding
    - Testing hypotheses
    - Mechanism is important
  - Questions motivated by a goal to predict
    - Prediction performance is important
    - Mechanism is not necessarily important

Observational vs Experimental Data

- Experimental data are the gold standard
  - Randomization allows isolation of effects
  - Caution about generalizing results
- Observational data are abundant
  - Experiments are not always possible
  - Features and relationships are difficult or impossible to isolate

Data from surveys

- Are we measuring what we think we are measuring?
  - Large field of research
  - Are we measuring the population of interest?
  - Non-response issues
  - Does the question measure what we are interested in?
- e.g. Would like to know whether people support handgun ownership.
  - Poll people leaving a sporting goods store.
  - Ask: “Have you considered handgun ownership for self defense?”

Planning ahead

- The best time to plan data analysis is **before** the data are collected
  - Preliminary data or data from another study can be used to design the analysis and experiment/survey
- The reality is that we are often asked to do data analysis after the fact
  - Although EDA can be useful, it is important to ask directed questions of the data to avoid fishing expeditions
  - Sometimes, it is not possible to answer a given question using a given dataset without resorting to *unreasonable* assumptions

Stat 862 students

- Reminder to see me this week about project alternative
- “Proposal” due date is Monday October 6