Statistical Graphics for High-D data

Hadley Wickham

Deborah F. Swayne

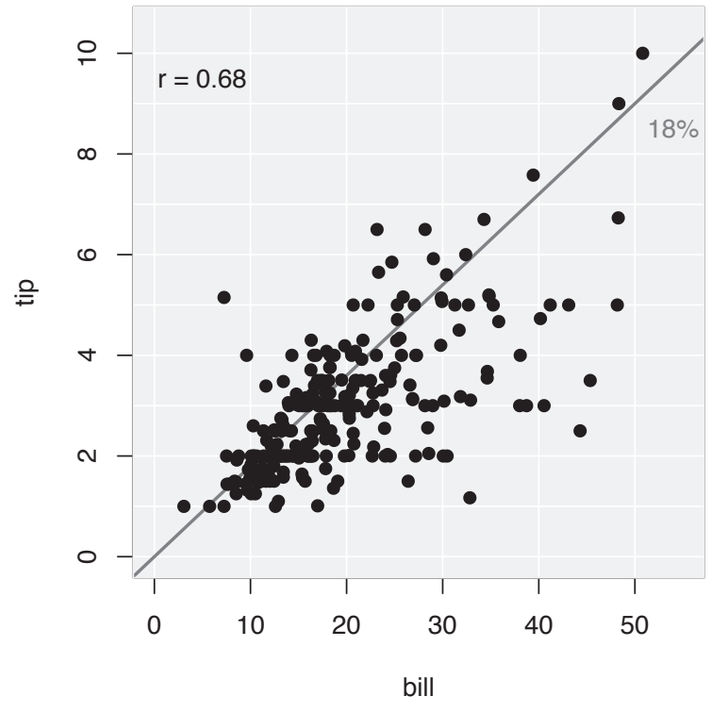
Dianne Cook

InfoVis 2007

Terminology

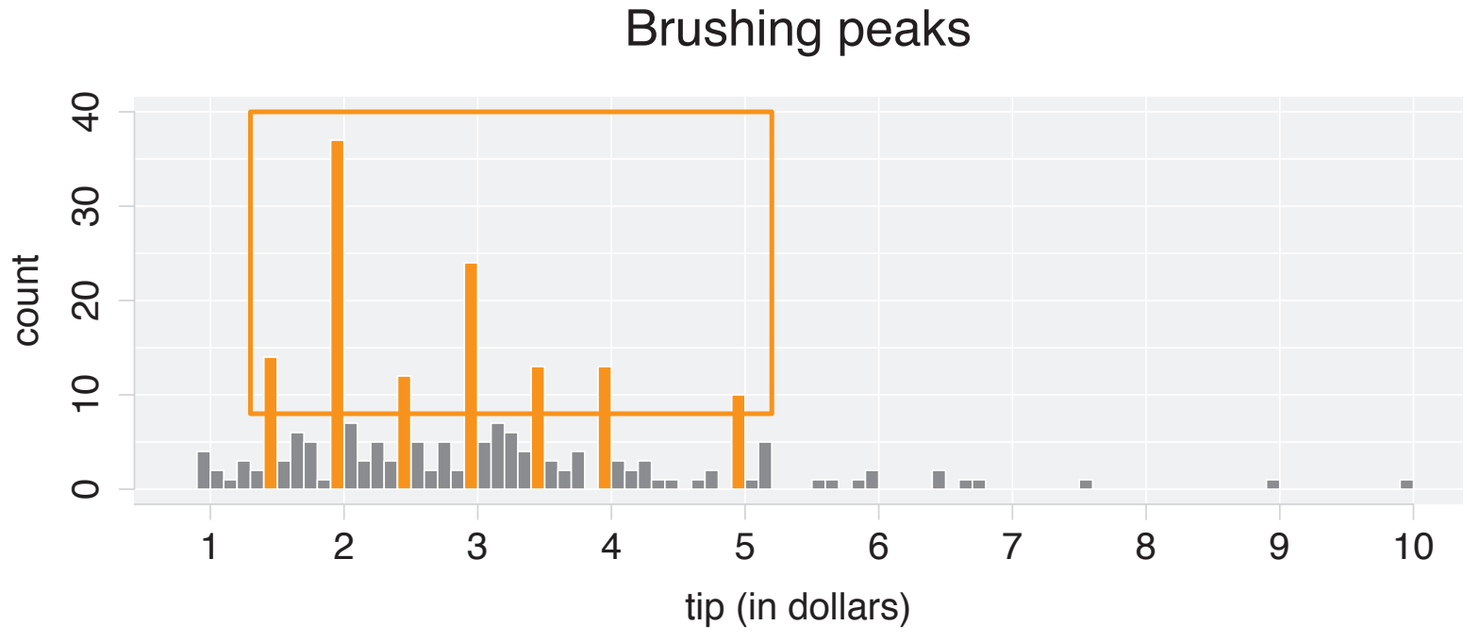
- *Data* refers to information that is structured in some schematic form such as a table.
- Data includes *attributes* or *variables*, eg number of hits on a web page, frequencies of words in text, weight, income per household.
- Data matrix: observations in rows, and variables in columns (GGobi: csv, or xml)

Statistical Thinking

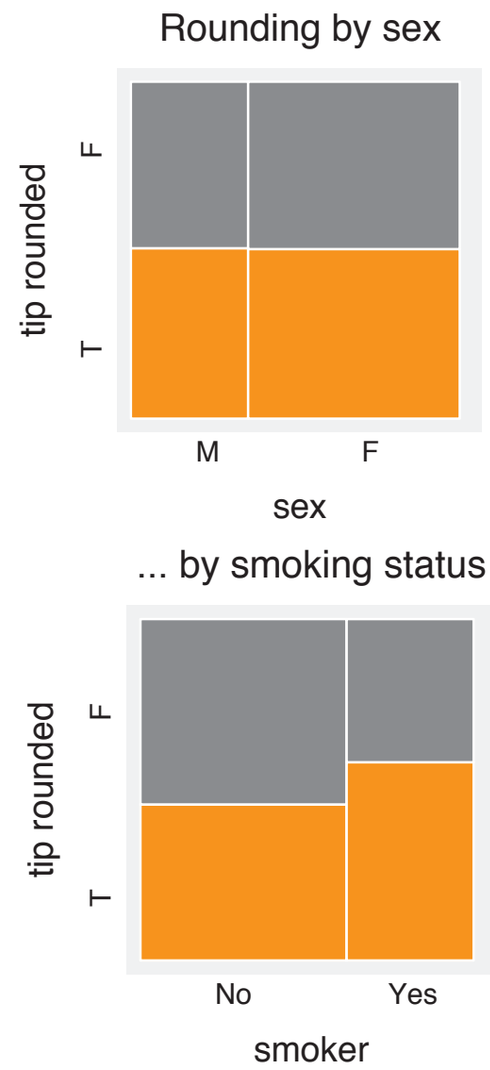


- Data visualization is an integral part of data analysis, to support and enrich exploration, modeling and inference.
- Concerned with variability in observations and errors in measurements, dealing with uncertainty.

Themes



- Start with low-D views, then get more complicated
- Many views, enhanced by interaction: linking between plots, and motion.
- Use similar methods for exploration and diagnostics. Display data and models together
- Using visualisation and analytic software together



Audience

- Not for “every one” - focus on statisticians and scientists
- Expect that users know (or want to learn) some statistics
- Research platform
- GUI, API, and CLI

GGobi, R and rggobi

- GGobi: http://www.ggobi.org
- R: language and environment for statistics, www.R-project.org
- rggobi provides scripted interface to GGobi from R
- R packages for specific visualisations: classifly, clusterfly
- R + rggobi code allows reproducibility

Quick history

- Dataviewer: Buja, Hurley, McDonald. 1986-
  Symbolics lisp machine
- XGobi: Swayne, Cook & Buja, 1991-.
  C & X Window System
- GGobi: Swayne, Cook, Buja, Temple Lang,
  Hofmann, Lawrence, Wickham. 1998-
  C & Gtk

http://stat-graphics.org/movies

Interactive and dynamic graphics for data analysis: with R and GGobi

www.ggobi.org/book

slides, R code, movies, data

Timeline

20 mins	Toolbox
30 mins	Missing values
45 mins	Supervised Classification
Break	
45 mins	Unsupervised Classification
30 mins	Inference