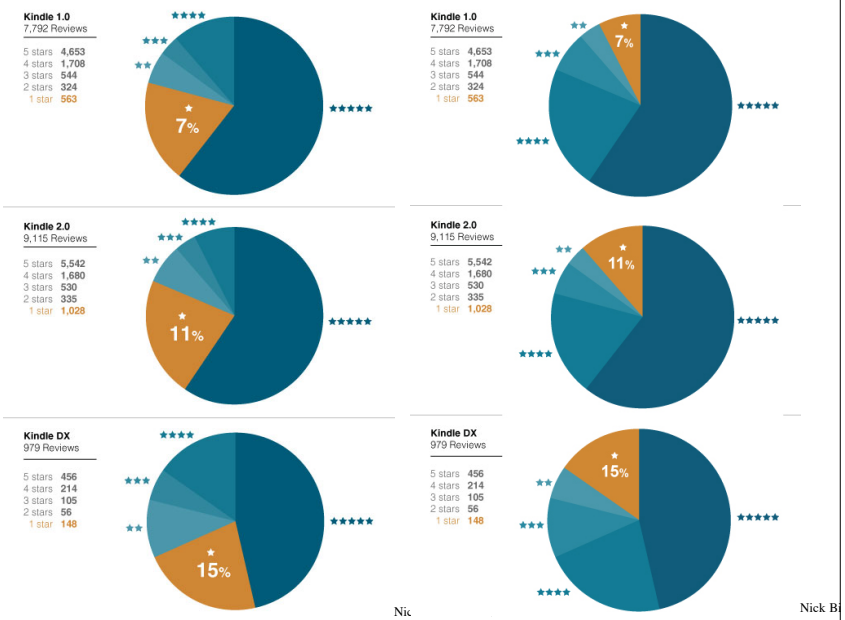
Graphical Data Analysis

Antony Unwin
University of Augsburg
unwin@math.uni-augsburg.de

Winter Term 2011/12



Times Customer feedback ratings for three versions of the Amazon Kindle.

Aims of the course

- At the end you should be able to
  - criticise published graphics constructively
  - carry out graphical data analyses
  - integrate graphical analysis and statistical modelling

Course Structure

- Review theories for graphics
- Examine graphics in practice
- Study interactive graphics
- Graphics and statistical models
- Format of course
  - Discuss a published graphic each lecture
  - Study a graphical data analysis each lecture
  - Evaluate blogs and websites for data visualization

Course information

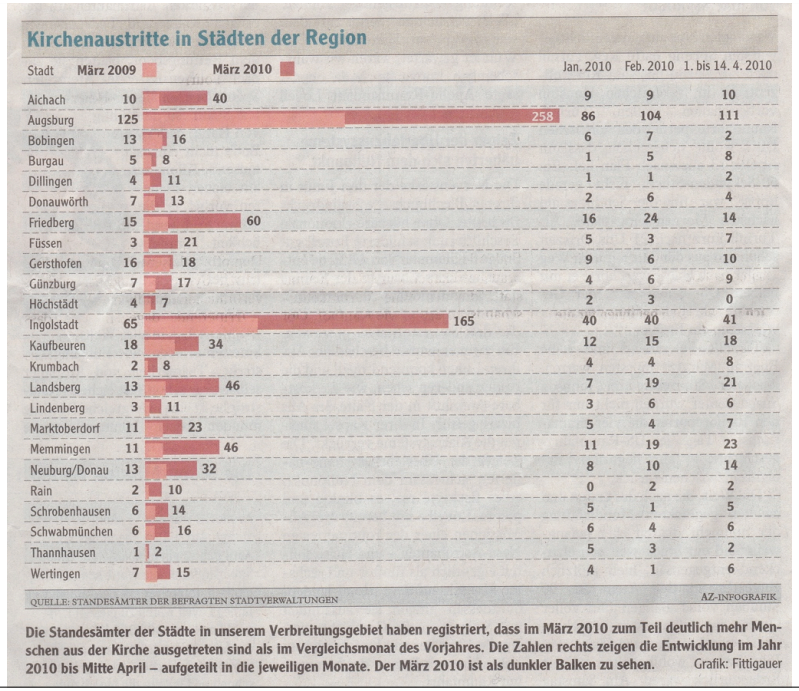
- Website: rosuda.org/lehre/WS1112-f/GDAw11.shtml
  - pdf's of slides
  - question sheets
  - datasets
  - code
  - general information
- Assessment
  - oral examination (based on individual project work)

Information Visualization

- All kinds of information
  - Scientific theories
  - Spatial layouts
  - Organisational structures
  - Social networks
  - Lines of argument
  - …
- This course is only concerned with data visualization

Why visualize data?

- Looking for global trends
  - overall structure
- Looking for local features
  - data quality
  - groups or clusters
  - outliers, tail distributions and extremes
  - patterns of all kinds



Kirchenaustritte in Städten der Region

Stadt	März 2009	März 2010	Jan. 2010	Feb. 2010	1. bis 14. 4. 2010
Aichach	10	40	9	9	10
Augsburg	125	258	86	104	111
Bobingen	13	16	6	7	2
Burgau	5	8	1	5	8
Dillingen	4	11	1	1	2
Donauwörth	7	13	2	6	4
Friedberg	15	60	16	24	14
Füssen	3	21	5	3	7
Gersthofen	16	18	9	6	10
Günzburg	7	17	4	6	9
Höchstädt	4	7	2	3	0
Ingolstadt	65	165	40	40	41
Kaufbeuren	18	34	12	15	18
Krumbach	2	8	4	4	8
Landsberg	13	46	7	19	21
Lindenberg	3	11	3	6	6
Marktoberdorf	11	23	5	4	7
Memmingen	11	46	11	19	23
Neuburg/Donau	13	32	8	10	14
Rain	2	10	0	2	2
Schrobenhausen	6	14	5	1	5
Schwabmünchen	6	16	6	4	6
Thannhausen	1	2	5	3	2
Wertingen	7	15	4	1	6

QUELLE: STANDESÄMTER DER BEFRAGTEN STADTVERWALTUNGEN AZ-INFOGRAFIK

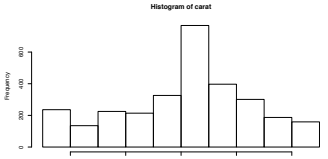
Die Standesämter der Städte in unserem Verbreitungsgebiet haben registriert, dass im März 2010 zum Teil deutlich mehr Menschen aus der Kirche ausgetreten sind als im Vergleichsmonat des Vorjahres. Die Zahlen rechts zeigen die Entwicklung im Jahr 2010 bis Mitte April – aufgeteilt in die jeweiligen Monate. Der März 2010 ist als dunkler Balken zu sehen. Grafik: Fittigauer

Bluenile Diamonds

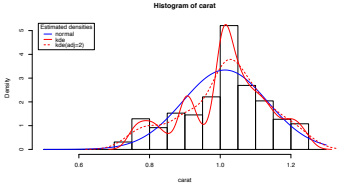
- Data from www.bluenile.com: 2948 diamonds
  - rank, pct
  - type (three)
  - carat
  - color (seven)
  - clarity (seven)
  - cut (five)
  - gcal (two)
  - price
  - cbody (two)

Example (1) Bluenile6

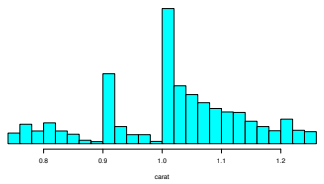
```
hist(carat)
```




```
#Add a density estimator
lines(density(carat))
#or
…,adjust=2))
```




```
#truehist is a density
#uses better default for no of bins
truehist(carat)
hist(carat,breaks="Scott")
```



Example (1) Comments

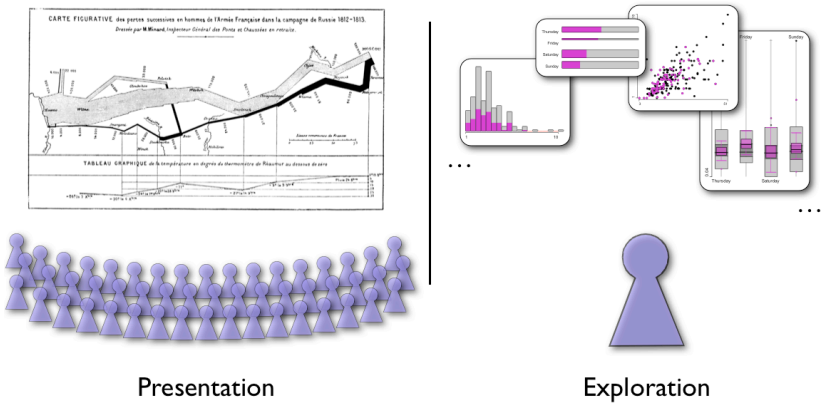
- Histograms are tricky (cf. lower/upper boundaries)
- If you know what you want to show, it is (almost) always possible to find a good graphic (Presentation)
- How do you find features? (Exploration)
  - try other binwidth methods or choose breaks
  - interactively vary binwidth and anchorpoint
  - use density estimates and models
  - use other graphics
  - use metadata and domain information

Global and local

- Global properties belong to the whole dataset
  - sampling works (n large), models work
- Local properties belong to a subset
  - Finding local properties involves selecting and conditioning — and always comparing
  - Global models may not work well
  - In large-d datasets local applies to subsets of variables, associated by type or relevance

Presentation v. Exploration

- Presentation graphics usually involve only one graphic for viewing by a huge number of people
- Exploratory graphics usually involve a huge number of graphics for viewing by only one person
- Presentation graphics convey known information
- Exploratory graphics are used to find information



Books

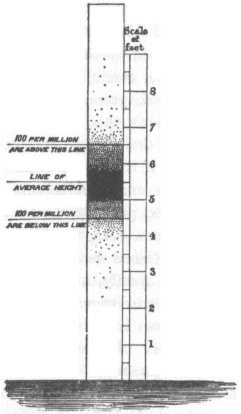
- “Grammar of Graphics” L. Wilkinson
- "ggplot2" H. Wickham
- "Interactive Graphics for Data Analysis" M. Theus, S. Urbanek
- "Graphics of Large Datasets" A. Unwin, M. Theus, H. Hofmann
- "Handbook of Data Visualization" (eds. Chen, Härdle, Unwin)
- Books by Edward Tufte

Software

- R
  - grid
  - ggplot2
  - lattice
  - vcd
  - iplots (Acinonyx)
  - …
- Mondrian
- SEURAT, GAUGUIN, …

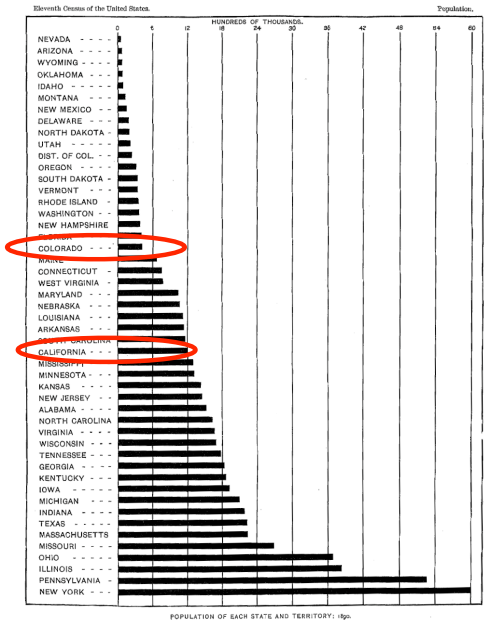
Data Visualization is not new

Galton's plot of the (hypothetical) heights of a million men

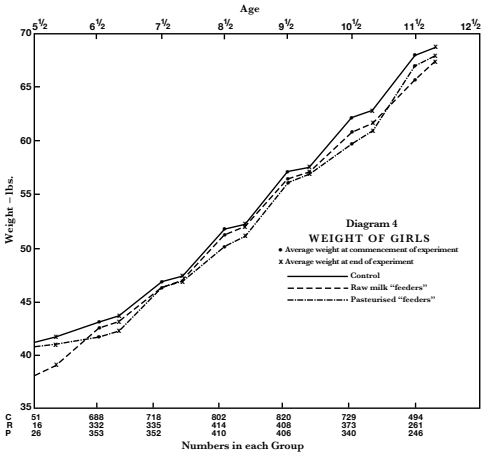


US Census 1890

63 million Hollerith cards



Student and Milk



The Lanarkshire Milk Experiment
Feb - June 1930

20000 Children
5000 got raw milk
5000 pasteurised
10000 got none

(Some) Websites (1)

- Gallery of Data Visualization
  - www.math.yorku.ca/SCS/Gallery/
- Statistical Modeling, Causal Inference, and Social Science
  - www.stat.columbia.edu/~gelman/blog/
- UK Local Government (public)
  - www.improving-visualisation.org
- Tableausoftware (commercial)
  - www.tableausoftware.com

(Some) Websites (2)

- Many Eyes
  - manyeyes.alphaworks.ibm.com/manyeyes/
- New York Times graphics
  - www.smallmeans.com/new-york-times-infographics/
- Junk Charts
  - junkcharts.typepad.com/
- Flowing Data
  - flowingdata.com
- Ask ET (Edward Tufte)
  - www.edwardtufte.com

(Some) Websites (3)

- Martin Theus Blog
  - www.theusRus.de/blog
- Data Applied (commercial)
  - www.data-applied.com/Web/Products/Overview.aspx
- Guardian newspaper
  - www.guardian.co.uk/data-store
- Name voyager and name mapper (some entertainment)
  - www.babynamewizard.com