Infovis and Statistical Graphics

Andrew Gelman and Antony Unwin
Columbia University and University of Augsburg

Empirical Legal Studies Conference, 5 Nov 2011
African Countries by GDP

Top countries by GDP in U.S. $ billions:

- **South Africa**: $285.4 b
- **Egypt**: $188.4 b
- **Nigeria**: $173 b
- **Algeria**: $140.6 b
- **Morocco**: $91.4 b
- **Angola**: $75.5 b
- **Libya**: $62.3 b
- **Kenya**: $39.6 b
- **Ethiopia**: $29.4 b
- **Ghana**: $28.5 b
- **Cameroon**: $26.2 b

GDP calculation:
Private consumption + gross investment + government spending + (exports − imports)
The informative (but boring) stat graphic
Different tools, different goals

Can we uncover the differences between the values and priorities of infovis and statistical graphics?
5 Best Data Visualization Projects of the Year
Honorable mention: Wordle
• Nathan Yau: “It's hard to say what exactly made Wordle so popular, but I [Yau] think it was a mix of randomness, aesthetics, and customization options

• Our view: Visualization as a fun puzzle
#3. Box office streamgraphs

Lee Byron
• Yau: “You can see Oscar contenders attract a smaller audience than the holiday and summer blockbusters and kind of slowly build an audience.”

• Our view:
  • Huh?
  • Better to have two graphs:
    (1) total sales over time, (2) trajectories for individual movies
  • Again, graph as puzzle
• Yau: “Discussion burst out across the Web . . . that I am convinced would not have come about if instead of a Streamgraph, they used say, a **stacked bar chart.**” [emphasis added]
“5 Best Data Visualizations”: our view

• Eye-catching graphics
• State-of-the art methods in stat and comp sci
• No attempt to achieve the traditional goals of statistical graphics (communication, discovery)
Winner of *Guardian* newspaper’s Visualization Contest

*Final Destination*
Density of fatal accidents 1942-2009

David McCandless
• Our view:
  – Display looks clean and efficient but isn’t!
  – Analogy to some modern architecture
Florence Nightingale’s coxcomb

http://www.Florence-Nightingale-Avenging-Angel.co.uk/Coxcomb.htm

Diagram of the Causes of Mortality in the Army in the East.

2. April 1855 to March 1856.

The areas of the blue, red, & black wedges are each measured from the centre as the common vertex.
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic Diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes.
The black line across the red triangle in Nov. 1854 marks the boundary of the deaths from all other causes during the month.
In October 1854, & April 1855, the black area coincides with the red, in January & February 1856, the blue coincides with the black.
The entire areas may be compared by following the blue, the red & the black lines enclosing them.
• Our view:
  
  – Excellent “infographic”—it’s attractive, grabby, thought-provoking
  
  – Graph as puzzle
  
  – Not a good “statistical graphic,” does not push to deeper understanding
  
  – “Clock plot” as dead end
Challenges in effectiveness research

Research: Why Chart Junk is More Useful than Plain Graphs

“Yep, it has been scientifically proven: the accuracy of people in describing charts with ‘chart junk’ is no worse than for plain charts, and the recall after a 2-3 week gap was actually significantly better. In addition, people overwhelmingly preferred ‘chart junk’ diagrams . . .”

But, before you go and slashdot this . . .
The “chartjunk” study is . . . junk!

- OK. Good chartjunk is better than crap chartjunk
Some practical tips

• Line plots and small multiples
• Avoid the graphical equivalent of the data dump
• Don’t try to cram everything into one plot
• Combine graphics with text
  – A picture plus 1000 words is worth more than two pictures or 2000 words
<table>
<thead>
<tr>
<th>Category</th>
<th>Sample Size</th>
<th>Percentage</th>
<th>Proportion Supporting Death Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>26953</td>
<td>(46%)</td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>31300</td>
<td>(54%)</td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>6516</td>
<td>(11%)</td>
<td></td>
</tr>
<tr>
<td>Non-black</td>
<td>51737</td>
<td>(89%)</td>
<td></td>
</tr>
<tr>
<td>18–29</td>
<td>12460</td>
<td>(21%)</td>
<td></td>
</tr>
<tr>
<td>30–44</td>
<td>18619</td>
<td>(32%)</td>
<td></td>
</tr>
<tr>
<td>45–64</td>
<td>17526</td>
<td>(30%)</td>
<td></td>
</tr>
<tr>
<td>65+</td>
<td>9648</td>
<td>(17%)</td>
<td></td>
</tr>
<tr>
<td>Less than High School</td>
<td>18211</td>
<td>(31%)</td>
<td></td>
</tr>
<tr>
<td>High School</td>
<td>25010</td>
<td>(43%)</td>
<td></td>
</tr>
<tr>
<td>Some College</td>
<td>5415</td>
<td>(9%)</td>
<td></td>
</tr>
<tr>
<td>College grad</td>
<td>7170</td>
<td>(12%)</td>
<td></td>
</tr>
<tr>
<td>Grad School</td>
<td>2447</td>
<td>(4%)</td>
<td></td>
</tr>
</tbody>
</table>
Graphing data and fitted models

MA (North); Mean yearly sample size = 43

OH (Midwest); Mean yearly sample size = 85
WHO SUPPORTS HEALTH CARE REFORM?

**AGE**
- 18 to 29
- 30 to 44
- 45 to 64
- 65+

**INCOME**
- Under $20,000
- $20,000 - $40,000
- $40,000 - $75,000
- Over $75,000

*Support for increased federal health care spending for the uninsured, based on the 2004 Annenberg survey.*
Last letters of boys’ names, 100 yrs ago

John, James, Edward, George, Henry, . . .
Last letters of boys’ names, 50 yrs ago

Michael, Thomas, Larry, Jeffrey, . . .
Last letters of boys’ names, now

Ethan (#8), John (18), Jonathan (19), Brandon (21), Christian (22), Dylan (23), also #25, 27, 28, 29, ...
The trend in last letters of boys’ names

- The long tail . . . and the paradox of freedom
Conclusion: Infovis vs. stat graphics

• Infovis:
  – Visual creativity, up-to-the-minute technology
  – Puzzles and the joy of recognition

• Statistical graphics:
  – Replication of standard forms
  – Discovery of the unexpected

• We can work together