Psychology of Data Visualization:
Course Overview

Michael Friendly
Psych 6135
http://euclid.psych.yorku.ca/www/psy6135/
@datavisFriendly

Introducing: me

I wear two hats, both reflected on my license plate:
Statistical graphics developer (categorical & multivariate data analysis)

History of data visualization: Les Chevaliers; Friendly & Wainer (2021)

Course Topics

• Varieties of information visualization
  ▪ Goals of visualization
  ▪ Survey of graphic forms
• History of information visualization
• Psychological models, theories and results
  ▪ What can people see, understand and remember from data displays?
  ▪ Perceptual aspects, cognitive aspects
• Software tools for information visualization
• Visualization in statistics: case studies
  ▪ Categorical data; High-D data; Dynamic and interactive methods
• Human factors research: how to tell what works
Your role

- Weekly readings – see the course web site for updates
- Discussion – no formal grade, but please contribute
- Discussion leader (20%)
  - Each week 1-2 of you will lead a brief discussion on one of the readings or sub-topics (~ 5 min.)
- Class presentation (40%)
  - In the last 2-3 weeks, each person will give a ~ 20 min presentation on a topic of research, application or software related to data visualization
- Research proposal (40%)
  - Prepare a brief research proposal on a data visualization topic

Books & Readings

- Colin Ware, *Information Visualization*, 3rd Ed.
  What perceptual science has to say about data visualization, from a bottom-up perspective
  Course notes at: [http://ccom.unh.edu/vislab/VisCourse/index.html](http://ccom.unh.edu/vislab/VisCourse/index.html)

- Alberto Cairo, *The Truthful Art*
  Information graphics from a communication perspective

- Steven Kosslyn, *Elements of Graph Design*
  A cognitive psychologist looks at graphs and presents some dos and don’ts

- Hadley Wickham, *ggplot2: Elegant graphics for data analysis*, 2nd Ed.
  Complete ggplot2 documentation: [http://docs.ggplot2.org/current/](http://docs.ggplot2.org/current/)

More books I like

- Tamara Munzner (2014), *Visualization Analysis & Design*
  An attractive new book combining computer science and design perspectives

- Howard Wainer (2005), *Graphic discovery: a trout in the milk and other visual adventures*
  A collection of essays on the history of graphics and other topics

- Manuel Lima, *The Book of Trees: Visualizing branches of knowledge*
  A visual delight; an entire history of tree-type diagrams

- Keiran Healy, *Data Visualization: A Practical Introduction*
  An accessible primer on how to create effective graphics from data using ggplot2
  Online: [http://socvis.co](http://socvis.co)

Tufte Stufte

Four books by Edward Tufte largely defined the landscape for data visualization and information design

Concepts introduced:
- chart junk,
- data-ink ratio,
- small multiples,
- substance takes precedence over visual design

Web site: [https://www.edwardtufte.com](https://www.edwardtufte.com)
**Blogs & Web resources**

My web site, [http://datavis.ca](http://datavis.ca). Contains the Milestone Project on the history of data vis, Data Visualization gallery, links to books, papers and courses.

Kaiser Fung, [http://junkcharts.typepad.com/](http://junkcharts.typepad.com/). Fung discusses a variety of data displays and discusses how they can be improved.

Nathan Yau’s blog, [http://flowingdata.com](http://flowingdata.com). A large number of blog posts illustrating data visualization methods with tutorials on how to do these with R and other software.

[http://visiphilia.org/](http://visiphilia.org/) Statistics Di Cook and Heike Hofmann from Iowa State University blog about data visualization topics, using R

Manuel Lima’s blog, [http://www.visualcomplexity.com/vc/blog/](http://www.visualcomplexity.com/vc/blog/) with hundreds of projects on all types of visualizations

---

**Data, pictures, models & stories**

Goal: Tell a credible story about some real data problem

- Measles vaccination
- Global warming
- ...

---

**DATA STORIES**


Annual awards celebrate excellence and beauty in data visualizations, infographics, interactives & information art. [https://www.informationisbeautifulawards.com](https://www.informationisbeautifulawards.com)

[https://www.r-bloggers.com/](https://www.r-bloggers.com/), A large collection of posts on R news and tutorials by over 750 R bloggers.


---

**Two paths to enlightenment**

1. *Exploratory*:
   - **data**
   - **visualization**
   - **story**

2. *Model-based*:
   - **data**
   - **model**
   - **summary**
   - **story**
Now, tell the story!

Pictures and images in a wider context

Modes of communication, as composed of words (story), numbers (symbols) and pictures (images) in different proportions.

- Poetry = 60% words + 40% images
- Table = 10% words + 80% numbers + 10% images

Words, numbers and pictures

- Beauty: The 4th dimension
- Modes of communication also vary in beauty & aesthetic appeal

Roles of graphics in communication

- Graphs (& tables) are forms of communication:
  - What is the audience?
  - What is the message?

  **Analysis graphs:** design to see patterns, trends, aid the process of data description, interpretation

  **Presentation graphs:** design to attract attention, make a point, illustrate a conclusion

**Basic functions of data display**

- **Primary Use**
  - Perception
  - Detection
  - Comparison
  - Aesthetics
  - Rhetoric
  - Exposition

- **Presentation Goal**
  - to Demonstrate
  - to Inform
  - to Persuade

- **Design Principles**
Different graphs for different purposes

Presentation
Goal: the Wow! experience
Single image for a large audience
Tells a clear story!

Exploration
Goal: the Ah ha! Experience
Many images, for a narrow audience (you!), linked to analysis

Powerful graphs: Measels and vaccines

Visualizing the impact of health policy interventions

In 2015 Tynan DeBold & Dov Friedman in the Wall Street Journal show the effect of the introduction of vaccination programs in the US states on disease incidence, using color-coded heat maps for a variety of diseases

Measles was decimated!

The message hits you between the eyes!

Powerful graphs make comparison easy

In 2014, vaccination rates declined and measles re-emerged in those areas

Effective graphs can cure ignorance, but not stupidity.


Presentation graph: Nightingale (1857)

Data graph: Nightingale (1857)

The same, as a data graph, using time-series line plots
Many statisticians might prefer this today, but it doesn’t draw attention or interest as Flo’s original did.
Rhetorical graph: Welfare income and Homeless deaths after the “Common Sense Revolution”

Scott Sorli (2007)

Analysis graph: Deaths vs. Income

Scatterplot of deaths vs. income
- Loess smooth + CI band
- Labels: year
- Color: party in power

The message here is interesting, but it lacks the power and eloquence of the original graph

As well, the relationship of deaths to time & party is lost

Racial profiling: Analysis graph

- Toronto Star (2002) study of police actions on a charge of simple possession of marijuana
  - release with a summons (Form 9) vs. hold for bail (Show cause)
  - Evidence for racial bias?
- First graph: mosaic display
  - area ~ frequency
  - shading: ~ residual
    - Obs > Expected in blue
    - Obs < Expected in red

Racial profiling: The process

How to communicate these results most effectively?
- What is the message? What features are directly comprehensible to the audience?
Racial profiling: Presentation graphic

Together, we created this self-explaining infographic

Title gives the main conclusion

Legend gives a layman’s description of shading levels

Text description gives details

Bar width = charges Divided by % release

numbers shown in the cells

Why plot your data?

Graphs help us to see patterns, trends, anomalies and other features not otherwise easily apparent from numerical summaries.

Source: http://xkcd.com/523/

Why plot your data?

Three data sets with exactly the same bivariate summary statistics:

- Same correlations, linear regression lines, etc
- Indistinguishable from standard printed output
- Totally different interpretations!

Comparing groups: Analysis vs. Presentation graphs

Six different graphs for comparing groups in a one-way design

- which group means differ?
- equal variability?
- distribution shape?
- what do error bars mean?
- unusual observations?

Never use dynamite plots

Always explain what error bars mean

Consider tradeoff between summarization & exposure
Graphs of model coefficients are often clearer than tables:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient (Standard Error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>1.23 (0.24) <strong>P &lt; 0.05</strong></td>
</tr>
<tr>
<td>Chile</td>
<td>0.98 (0.36) <strong>P &lt; 0.01</strong></td>
</tr>
<tr>
<td>Colombia</td>
<td>1.86 (0.45) <strong>P &lt; 0.001</strong></td>
</tr>
<tr>
<td>Mexico</td>
<td>0.81 (0.27) <strong>P &lt; 0.05</strong></td>
</tr>
<tr>
<td>Venezuela</td>
<td>0.53 (0.20) <strong>P &lt; 0.01</strong></td>
</tr>
</tbody>
</table>

R^2 = 0.15, Adjusted R^2 = 0.12
n = 500

Source: tables2graphs.com

Effective data display:

- **Make the data stand out**
  - Fill the data region (axes, ranges)
  - Use visually distinct symbols (shape, color) for different groups
  - Avoid chart junk, heavy grid lines that detract from the data

- **Facilitate comparison**
  - Emphasize the important comparisons visually
  - Side-by-side easier than in separate panels
  - “data” vs. a “standard” easier against a horizontal line
  - Show uncertainty where possible

- **Effect ordering**
  - For variables and unordered factors, arrange them according to the effects to be seen

Comparisons— Make visual comparisons easy
- Visual grouping— connect with lines, make key comparisons contiguous
- Baselines— compare data to model against a line, preferably horizontal
- Frequencies often better plotted on a square-root scale

Make comparisons direct
- Use points not bars
- Connect similar by lines
- Same panel rather than different panels
Showing uncertainty
- Standard plots of observed vs. predicted lack a basis for assessment of uncertainty
- Confidence envelopes indicate extent of deviation
- Identify “noteworthy” observations to track them down

Example: Normal QQ plots used to assess normality of data

Effect ordering
- Information presentation is always ordered
  - in time or sequence (a talk or written paper)
  - in space (table or graph)
  - Constraints of time & space are dominant—can conceal or reveal the important message
- Effect ordering for data display
  - Sort the data by the effects to be seen
  - Order the data to facilitate the task at hand
    - lookup – find a value
    - comparison – which is greater?
    - detection – find patterns, trends, anomalies

Effect order failure: the Challenger disaster
- Few events in history provide as compelling illustration of importance of appropriate ordering and display of information
  - On January 28, 1986, the space shuttle Challenger exploded on take-off.
  - The cause was later determined to be that rubber O-rings failed due to cold weather
- Tables and charts presented to NASA by Thiokol engineers showed data from prior launches ordered by time (launch number), rather than by temperature—the crucial factor.
- The engineers’ charts were also remarkable for information obfuscation: “erosion depth” (O-ring damage), “blow-by” (soot on O-rings), ...

Visual explanation: Physics
- NASA appointed members of the Rogers Commission to investigate the cause of the disaster
- the noted physicist Richard Feynman discovered the cause: at low temperature, O-rings became brittle and were subject to failure
- in his testimony, he demonstrated the effect by plunging a rubber O-ring into a cup of ice water
Visual explanation: Graphics

- Subsequent statistical analysis showed the relationship between launch temperature and O-ring failures
- As Tufte (1997) notes: the fatal flaw was in the ordering of the data.

The graph shown here is the result of a statistical model fit to the data
- The thick line shows the predicted value of failure vs. temperature
- The red dotted lines show uncertainty of the predicted values

A presentation version of the previous graph alters the scales and describes the story in text annotations

Graphic displays: Main effect ordering

- To see trends, patterns, anomalies: Sort unordered factors by means or medians

Data on barley yields
10 varieties x 6 sites x 2 years
3 way dot plot, sorted by main effect means
- Which site has the highest yield?
- Which variety is highest on average?
- Which site stands out in pattern over year?

Tabular displays: Main effect ordering

- Tables are often presented with rows/cols ordered alphabetically
  - good for lookup
  - bad for seeing patterns, trends, anomalies

Table 1: Average Barley Yields (rounded), Means by Site and Variety

<table>
<thead>
<tr>
<th>Variety</th>
<th>Crookston</th>
<th>Duluth</th>
<th>Grand Rapids</th>
<th>Morris</th>
<th>University Farm</th>
<th>Waseca</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glabron</td>
<td>32</td>
<td>28</td>
<td>22</td>
<td>32</td>
<td>40</td>
<td>46</td>
<td>33.3</td>
</tr>
<tr>
<td>Manchuria</td>
<td>36</td>
<td>26</td>
<td>28</td>
<td>31</td>
<td>27</td>
<td>41</td>
<td>31.5</td>
</tr>
<tr>
<td>No. 45*</td>
<td>40</td>
<td>28</td>
<td>26</td>
<td>36</td>
<td>35</td>
<td>50</td>
<td>35.8</td>
</tr>
<tr>
<td>No. 462</td>
<td>40</td>
<td>25</td>
<td>22</td>
<td>39</td>
<td>31</td>
<td>55</td>
<td>35.4</td>
</tr>
<tr>
<td>No. 475</td>
<td>38</td>
<td>30</td>
<td>17</td>
<td>33</td>
<td>27</td>
<td>44</td>
<td>31.8</td>
</tr>
<tr>
<td>Pearlband</td>
<td>33</td>
<td>32</td>
<td>31</td>
<td>37</td>
<td>30</td>
<td>42</td>
<td>34.2</td>
</tr>
<tr>
<td>Stensonic</td>
<td>31</td>
<td>24</td>
<td>23</td>
<td>30</td>
<td>31</td>
<td>43</td>
<td>30.4</td>
</tr>
<tr>
<td>Trobi</td>
<td>44</td>
<td>32</td>
<td>25</td>
<td>45</td>
<td>33</td>
<td>57</td>
<td>39.9</td>
</tr>
<tr>
<td>Veler</td>
<td>37</td>
<td>24</td>
<td>28</td>
<td>32</td>
<td>33</td>
<td>44</td>
<td>33.1</td>
</tr>
<tr>
<td>Wisconsin No. 38</td>
<td>43</td>
<td>30</td>
<td>28</td>
<td>39</td>
<td>39</td>
<td>58</td>
<td>39.4</td>
</tr>
</tbody>
</table>

Mean: 37.4 28.0 24.9 35.4 32.7 48.1 34.4
Tabular displays: Main effect ordering

- Better: sort rows/cols by means/medians
- Shade cells according to residual from additive model

Yield difference, $\Delta y_{ij} = 1931 - 1932$ by Variety & Site

Ordered: by row and column means; shaded: by value ($|\Delta y_{ij}| > (2.3) \times \sigma (\Delta y_{ij})$)

What features stand out?

Table 2: Average Barley Yields, sorted by Mean, shaded by residual from the model $Yield = Variety + Site$

<table>
<thead>
<tr>
<th>Variety</th>
<th>Grand Rapids</th>
<th>Duluth</th>
<th>University Farm</th>
<th>Morris</th>
<th>Crookston</th>
<th>Waseca</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Svansota</td>
<td>23</td>
<td>24</td>
<td>31</td>
<td>30</td>
<td>31</td>
<td>43</td>
<td>30.4</td>
</tr>
<tr>
<td>Manchuria</td>
<td>28</td>
<td>26</td>
<td>27</td>
<td>31</td>
<td>36</td>
<td>41</td>
<td>31.5</td>
</tr>
<tr>
<td>No. 475</td>
<td>17</td>
<td>30</td>
<td>27</td>
<td>33</td>
<td>38</td>
<td>44</td>
<td>31.8</td>
</tr>
<tr>
<td>Velvet</td>
<td>28</td>
<td>24</td>
<td>33</td>
<td>32</td>
<td>37</td>
<td>44</td>
<td>33.1</td>
</tr>
<tr>
<td>Glabron</td>
<td>22</td>
<td>28</td>
<td>40</td>
<td>32</td>
<td>32</td>
<td>46</td>
<td>33.3</td>
</tr>
<tr>
<td>Feastland</td>
<td>31</td>
<td>32</td>
<td>30</td>
<td>37</td>
<td>33</td>
<td>42</td>
<td>34.2</td>
</tr>
<tr>
<td>No. 462</td>
<td>22</td>
<td>25</td>
<td>31</td>
<td>39</td>
<td>40</td>
<td>55</td>
<td>35.4</td>
</tr>
<tr>
<td>No. 457</td>
<td>26</td>
<td>28</td>
<td>35</td>
<td>36</td>
<td>40</td>
<td>50</td>
<td>35.6</td>
</tr>
<tr>
<td>Wisconsin No. 38</td>
<td>28</td>
<td>30</td>
<td>39</td>
<td>38</td>
<td>43</td>
<td>57</td>
<td>39.4</td>
</tr>
<tr>
<td>Trebi</td>
<td>25</td>
<td>32</td>
<td>33</td>
<td>45</td>
<td>44</td>
<td>57</td>
<td>39.4</td>
</tr>
<tr>
<td>Mean</td>
<td>24.9</td>
<td>28.0</td>
<td>32.7</td>
<td>35.4</td>
<td>37.4</td>
<td>48.1</td>
<td>34.4</td>
</tr>
</tbody>
</table>

Table 3: Yield Differences, 1931-1932, sorted by mean difference, and shaded by value

<table>
<thead>
<tr>
<th>Variety</th>
<th>Morris</th>
<th>Duluth</th>
<th>University Farm</th>
<th>Grand Rapids</th>
<th>Waseca</th>
<th>Crookston</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. 475</td>
<td>-22</td>
<td>-6</td>
<td>-5</td>
<td>4</td>
<td>6</td>
<td>12</td>
<td>0.1</td>
</tr>
<tr>
<td>Wisconsin No. 38</td>
<td>-18</td>
<td>2</td>
<td>1</td>
<td>14</td>
<td>4</td>
<td>14</td>
<td>2.4</td>
</tr>
<tr>
<td>Velvet</td>
<td>-13</td>
<td>4</td>
<td>4</td>
<td>-9</td>
<td>13</td>
<td>9</td>
<td>2.9</td>
</tr>
<tr>
<td>Manchuria</td>
<td>-7</td>
<td>6</td>
<td>0</td>
<td>11</td>
<td>15</td>
<td>7</td>
<td>5.5</td>
</tr>
<tr>
<td>Trebi</td>
<td>-3</td>
<td>3</td>
<td>7</td>
<td>9</td>
<td>15</td>
<td>5</td>
<td>6.1</td>
</tr>
<tr>
<td>Svansota</td>
<td>-9</td>
<td>3</td>
<td>6</td>
<td>13</td>
<td>9</td>
<td>20</td>
<td>7.3</td>
</tr>
<tr>
<td>No. 462</td>
<td>-17</td>
<td>6</td>
<td>11</td>
<td>5</td>
<td>21</td>
<td>18</td>
<td>7.4</td>
</tr>
<tr>
<td>Glabron</td>
<td>-6</td>
<td>4</td>
<td>6</td>
<td>15</td>
<td>17</td>
<td>12</td>
<td>8.0</td>
</tr>
<tr>
<td>No. 457</td>
<td>-15</td>
<td>11</td>
<td>17</td>
<td>13</td>
<td>16</td>
<td>11</td>
<td>8.8</td>
</tr>
<tr>
<td>Mean</td>
<td>-12.2</td>
<td>4.6</td>
<td>6.3</td>
<td>8.2</td>
<td>12.5</td>
<td>12.5</td>
<td>5.3</td>
</tr>
</tbody>
</table>

Graphical display: Two-way tables

Tukey two-way plot of average barley yield

If there is no interaction,

$Y_{ij} = \mu + \alpha_{site} + \beta_{variety}$

Site & variety effects sorted automatically
Effects are spaced by fitted values

More variation among sites than varieties
Waseca best, by a wide margin

Multivariate data: correlation ordering

- Arrange **variables** so that:
  - Similar variables are contiguous
  - Ordered to show patterns of relations

- Arrange **observations** so that:
  - Similar variables are contiguous
  - Ordered to show patterns of relations
Correlation matrices

```r
> cor(bb)
     Assists Atbat Errors Hits Homer logSal Putouts RBI Runs Walks Years
Assists  1.0000 0.3421 0.7035 0.3040 -0.1616 0.0500 -0.0434 0.0629 0.179 0.1025 -0.0851
Atbat  0.3421 1.0000 0.3256 0.9640 0.5510 0.4149 0.3096 0.7360 0.300 0.6244 0.1027
Errors 0.7035 0.3256 1.0000 0.2799 -0.0097 0.2053 0.1502 0.193 0.0189 -0.1565
Hits  0.3040 0.9640 0.2798 1.0000 0.5303 0.4466 0.2979 0.7885 0.411 0.5873 0.0184
Homer -0.1616 0.5531 -0.0097 0.5306 1.0000 0.3398 0.2509 0.8491 0.631 0.4405 0.1135
LogSal 0.0500 0.4149 -0.0097 0.5306 0.3398 1.0000 0.2245 0.4464 0.426 0.4324 0.5374
Putouts -0.0434 0.3096 0.2053 0.2979 0.2509 0.2245 1.0000 0.3121 0.271 0.2809 -0.0200
RBI  0.1025 0.6244 0.2053 0.7885 0.8491 0.4464 0.4461 1.0000 0.779 0.5685 0.1237
Runs 0.179 0.1025 0.1502 0.7885 0.8491 0.4464 0.2809 1.0000 0.779 0.5685 -0.0200
Walks 0.1025 0.6244 0.193 0.5873 0.5873 0.426 0.2809 0.779 1.0000 0.697 0.1348
Years -0.0851 0.0187 -0.1565 0.0184 0.0184 0.5374 -0.0200 0.271 0.0184 1.0000 0.1348
```

> round(100*cor(bb))

```r
     Assists Atbat Errors Hits Homer logSal Putouts RBI Runs Walks Years
Assists     100    34     70   30  -16    5   -4  6  18  10  -9
Atbat 34   100     33    96    56   41   31  80  90  62   1
Errors 70  33  100     28    10  55  30  79  91  53   2
Hits -16  56  -1  53  100  34  25  85  63  44  11
Homer  5   41   -2  45  34  100  22  44  43  43  54
LogSal  5   41   -2  45  34  100  22  44  43  43  54
Putouts -4  51   8  50  25  22  100  31  27  28  -2
RBI  6  80  15  79  85  44  31 100  78  57  13
Runs 18  90  19  91  63  43  27  78 100  70  -1
Walks 10  62   8  53  44  43  28  57  70 100  13
Years -9   1   -6  2  11  54  -2 13  -1  13 100
```

Baseball data: Batting, fielding and (log) Salary
Nobody wants to see all those decimals

If you are going to present the numbers, round a lot

```r
> round(150*cor(bb))
     Assists Atbat Errors Hits Homer logSal Putouts RBI Runs Walks Years
Assists     100    34     70   30  -16    5   -4  6  18  10  -9
Atbat 34   100     33    96    56   41   31  80  90  62   1
Errors 70  33  100     28    10  55  30  79  91  53   2
Hits -16  56  -1  53  100  34  25  85  63  44  11
Homer  5   41   -2  45  34  100  22  44  43  43  54
LogSal  5   41   -2  45  34  100  22  44  43  43  54
Putouts -4  51   8  50  25  22  100  31  27  28  -2
RBI  6  80  15  79  85  44  31 100  78  57  13
Runs 18  90  19  91  63  43  27  78 100  70  -1
Walks 10  62   8  53  44  43  28  57  70 100  13
Years -9   1   -6  2  11  54  -2 13  -1  13 100
```

Correlation ordering: corrgrams

**Baseball data**
This is a corrgram display of the correlations among the baseball statistics, with the variables ordered alphabetically

The same display, with the variables sorted according to the angles between vectors in the PCA
Not that dramatic, but it isolates the positive & negative correlations
Graphs: Good/Bad, Excellent/Evil

- Like good writing, good graphical displays of data communicate ideas with:
  - clarity,
  - precision, and
  - efficiency—avoids graphic clutter
- Even better: excellent graphs make the message obvious
- Like poor writing, bad graphical displays:
  - distort or obscure the data,
  - make it harder to understand or compare, or
  - thwart the communicative effect the graph should convey.
- Even worse: evil graphs distort, or mislead.

Bad graphs are easy in Excel

Friends don’t let friends use Excel for data visualization or statistics

How many things are wrong with this graph?

Pie charts are easy to abuse

What’s wrong with this picture?

On the other hand, pie charts are a great source of merriment for people interested in graphics

But, can be used to great effect

This graphic uses pie charts to show the transport of different kinds of goods to the ports of Paris and the principal maritime ports
- the size of each pie reflects total
- the sectors reflect relative %
- location places them in context

*Album de Statistique Graphique, 1885, plate 17.*
3D pie charts are usually evil

What was the intent of the designer of this graphic? Which category led to the greatest total deaths? What was the proportion of deaths due to strokes? Did more people die from strokes vs. accidents?

Double Y-axis: Really evil graphs

After pie charts, double Y-axis graphs have caused more trouble than almost any other

OMG, autism has been increasing directly with sales of organic food!

But, can be used to great effect

William Playfair invented the pie chart, line chart and bar chart. In this figure, he shows 3 parallel time series over a 250-year period, 1560–1810

• weekly wages of a good mechanic
• price of wheat
• reigning monarch

Goal: show that workers were better off most recently (1810) than in the past
Or, another graph would have been better

A modern re-vision plots the ratio of price of wheat to wages directly

---

Even more evil: No scales, no data

Rep. Jason Chaffetz, R-Utah, sparred with Planned Parenthood president Cecile Richards during a high-profile hearing on Sept. 29, 2015 and presented this graph.

"In pink, that's the reduction in the breast exams, and the red is the increase in the abortions. That's what's going on in your organization."

Created by an anti-abortion group it is a deliberate attempt to mislead.

Can you see why?


---

Corrected graph

This graph shows the actual data from the Planned Parenthood reports used by Americans United for Life

The number of abortions was relatively steady.

Some services like pap smears, dropped due to changing medical standards about who should be screened and how often.

What are a few improvements that could be made to this graph?

---

Corrected graph, in context

Showing a wider range of PP activities puts these data in context

PP activities were far higher for contraception and STD testing
A study by Abigail Friendly (2017) wanted to show the use of benefits afforded to Toronto developers for their contributions of different types over time.

Color background scale from light to dark highlights the largest values.

Most frequent benefits appear at the top.

Can see overall trends and anomalies.

What happened in 2014-2016?


This graph reports the results of a survey by Sherman Kent for the CIA with the question:

What [probability/number] would you assign to the phrase “[phrase]”?

The goal was to contribute to an understanding of how intelligence analysts use these terms.

Why can this be considered a graphical failure?

This graph shows the same data, as both dotplots & boxplots.

We can see a lot more:

- “about even” has very low variability
- the last 3 categories are listed out of order
- the extreme outliers stand out
- skewness is – for high probability, + for low probability

Technical notes:

- software: ggplot2
- design: faint grid lines
- color: points use transparent color & jittering; outliers also shown in black

From: https://github.com/zonination/perceptions

This graph uses “ridgeline” plots to show the same data.

Each one is a small version of a density plot showing a smoothed version of the distribution.

Stacking them in this way allows center, variability, shape and other features to be readily compared.
Suzana Herculano-Houzel has a new method for determining counts of cortical neurons across different species. How to present this effectively?

Goal: compare mammal species brain size and cortical neuron count

Neuron count is shown both as numbers and bars

What do you think?

How could this be made better?

As a scatterplot

A scatterplot makes clear how humans differ from other species

• Using scaled images as point symbols also conveys brain size
• Primates are distinguished from non-primates by text color

This is arguably a more effective display.

As a scatterplot – log scale

Perhaps even better is to make the plot using log scales for both axes

The relationship is now approx. linear
Why graphs matter: Climate change

In the movie, *An Inconvenient Truth* (2006), Al Gore used the now-famous “hockey stick” graph to show that human activities had greatly increased the degree of global warming over the recent past.

The goal was to raise public awareness and call for action to curb environmental effects: CO₂ emissions as the main agent.

Climate change: Original graph

Sir John Houghton presents the original Northern Hemisphere hockey stick graph to the Intergovernmental Panel on Climate Change (IPCC) in 2005. It is based on an analysis by Mann, Bradley & Hughes (1990), with a smoothed curve and uncertainty intervals.

Climate change: data sources

The MBH (1999) paper had used a wide variety of data sources. They were combined using a novel statistical technique, the first eigenvector-based climate field reconstruction (CFR).

Climate scientists understood this; the sceptics did not.

Countering climate change

Taking a longer view, and adding a lot of extraneous historical details, climate sceptics were easily able to mount alternative explanations.
Time scale

Perhaps one fault with the original graphs was trying to show noisy data, from many sources, over too wide a time span.

Climate change: Infographic

A politically incorrect graphic shows very clearly the effect of global warming on panty size


Climate change: other explanations

This infographic attempts to relate global warming to the decrease in pirates

Aside from the substance, how many things are wrong about this graphic?

Simple explanation:
Lack of pirates causes global warming!

Conclusion:
To stop global warming, become a pirate!


Climate change: animation

This animation shows a rotating globe indicating local effects of global warming (red is warmer)

The graph below shows the global average temperature changes from historical averages

Video: https://youtu.be/xhqEkyjDBho
Summary

• Graphs as a form of communication
  ▪ Data (numbers), words, images → Stories
• Analysis graphs vs. presentation graphs
• Some principles of effective data display
  ▪ Make the data stand out
  ▪ Facilitate comparisons
  ▪ Effect ordering