A Primer on Data Visualization Methods

David S. Pieczkiewicz, PhD
Clinical Assistant Professor
Institute for Health Informatics
University of Minnesota, USA

First International Conference on Research Methods for Standardized Terminologies
April 3, 2013
Identical Statistics...

<table>
<thead>
<tr>
<th>( x_1 )</th>
<th>( y_1 )</th>
<th>( x_2 )</th>
<th>( y_2 )</th>
<th>( x_3 )</th>
<th>( y_3 )</th>
<th>( x_4 )</th>
<th>( y_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.0</td>
<td>8.04</td>
<td>10.0</td>
<td>9.14</td>
<td>10.0</td>
<td>7.46</td>
<td>8.0</td>
<td>6.58</td>
</tr>
<tr>
<td>8.0</td>
<td>6.95</td>
<td>8.0</td>
<td>8.14</td>
<td>8.0</td>
<td>6.77</td>
<td>8.0</td>
<td>5.76</td>
</tr>
<tr>
<td>13.0</td>
<td>7.58</td>
<td>13.0</td>
<td>8.74</td>
<td>13.0</td>
<td>12.7</td>
<td>8.0</td>
<td>7.71</td>
</tr>
<tr>
<td>9.0</td>
<td>8.81</td>
<td>9.0</td>
<td>8.77</td>
<td>9.0</td>
<td>7.11</td>
<td>8.0</td>
<td>8.84</td>
</tr>
<tr>
<td>11.0</td>
<td>8.33</td>
<td>11.0</td>
<td>9.26</td>
<td>11.0</td>
<td>7.81</td>
<td>8.0</td>
<td>8.47</td>
</tr>
<tr>
<td>14.0</td>
<td>9.96</td>
<td>14.0</td>
<td>8.10</td>
<td>14.0</td>
<td>8.84</td>
<td>8.0</td>
<td>7.04</td>
</tr>
<tr>
<td>6.0</td>
<td>7.24</td>
<td>6.0</td>
<td>6.13</td>
<td>6.0</td>
<td>6.08</td>
<td>8.0</td>
<td>5.25</td>
</tr>
<tr>
<td>4.0</td>
<td>4.26</td>
<td>4.0</td>
<td>3.10</td>
<td>4.0</td>
<td>5.39</td>
<td>19.0</td>
<td>12.5</td>
</tr>
<tr>
<td>12.0</td>
<td>10.8</td>
<td>12.0</td>
<td>9.13</td>
<td>12.0</td>
<td>8.15</td>
<td>8.0</td>
<td>5.56</td>
</tr>
<tr>
<td>7.0</td>
<td>4.82</td>
<td>7.0</td>
<td>7.26</td>
<td>7.0</td>
<td>6.42</td>
<td>8.0</td>
<td>7.91</td>
</tr>
<tr>
<td>5.0</td>
<td>5.68</td>
<td>5.0</td>
<td>4.74</td>
<td>5.0</td>
<td>5.73</td>
<td>8.0</td>
<td>6.89</td>
</tr>
</tbody>
</table>

\[
\bar{x}_i \quad 9.0 \\
\sigma^2_{x_i} \quad 11.0 \\
\bar{y}_i \quad 7.5 \\
\sigma^2_{y_i} \quad 4.12 \\
\rho_{x_iy_i} \quad 0.866 \\
\text{LS linear fit} \quad y_i = 3 + 0.5x_i
\]
...Different Plots!

\[(x_1, y_1)\]

\[(x_2, y_2)\]

\[(x_3, y_3)\]

\[(x_4, y_4)\]
Why Visualize Data?

• We usually do it for one of two reasons:

  — To persuade

  — To explore and discover

The greatest value of a picture is when it forces us to notice what we never expected to see. – John W. Tukey, statistician (1915-2000)
Reading Visualizations

• *Perception*: The low-level activity of *sensing* the visual aspects of a display
• *Cognition*: The higher-level process of *interpreting* the display and translating it into meaning
• *The Challenge*: Using what we know about both to make visualizations better
# Common Perceptual Tasks

<table>
<thead>
<tr>
<th>Task</th>
<th>Illustration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position along aligned scales</td>
<td><img src="image1" alt="Illustration" /></td>
</tr>
<tr>
<td>Position along unaligned scales</td>
<td><img src="image2" alt="Illustration" /></td>
</tr>
<tr>
<td>Length</td>
<td><img src="image3" alt="Illustration" /></td>
</tr>
<tr>
<td>Angle and slope</td>
<td><img src="image4" alt="Illustration" /></td>
</tr>
<tr>
<td>Area</td>
<td><img src="image5" alt="Illustration" /></td>
</tr>
<tr>
<td>Volume</td>
<td><img src="image6" alt="Illustration" /></td>
</tr>
<tr>
<td>Color and saturation</td>
<td><img src="image7" alt="Illustration" /></td>
</tr>
</tbody>
</table>
Common Cognitive Tasks

• The parallel idea that there are basic, atomic tasks that are performed with visualizations:

  – Value extraction
  – Proportion extraction
  – Value comparison
  – Trend detection

• Visualizations have different degrees of cognitive burden for different tasks
# Two Views of Data

<table>
<thead>
<tr>
<th>Date</th>
<th>FEV₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>5/14/2000</td>
<td>0.963</td>
</tr>
<tr>
<td>5/17/2000</td>
<td>1.108</td>
</tr>
<tr>
<td>5/18/2000</td>
<td>1.026</td>
</tr>
<tr>
<td>5/19/2000</td>
<td>1.024</td>
</tr>
<tr>
<td>5/20/2000</td>
<td>1.005</td>
</tr>
<tr>
<td>5/21/2000</td>
<td>1.038</td>
</tr>
<tr>
<td>5/22/2000</td>
<td>0.999</td>
</tr>
<tr>
<td>5/23/2000</td>
<td>1.049</td>
</tr>
<tr>
<td>5/24/2000</td>
<td>1.068</td>
</tr>
<tr>
<td>5/25/2000</td>
<td>1.084</td>
</tr>
<tr>
<td>5/26/2000</td>
<td>1.024</td>
</tr>
<tr>
<td>5/27/2000</td>
<td>1.010</td>
</tr>
<tr>
<td>5/28/2000</td>
<td>1.023</td>
</tr>
</tbody>
</table>
There is No One-Best-Way!

- The “best” visualization format for a situation is highly task-dependent
- Some tasks may require combinations
- Many published guidelines, with different aims
  - Persuasive graphics
  - Statistical graphics
- There is no general theory of visualization
Suggested Practices

Value extraction
- Numeric tables

Proportions
- Pie charts, stacked bar charts

Value comparison
- Bar charts, line graphs, scatterplots

Trend detection
- Line graphs

Use the design which minimizes cognitive burden for the task at hand
Three Questions to Ask

• These should be asked when setting out to design a visualization:
  – Who is the intended *audience*?
  – What is the *goal* of the visualization? (exploration, education, persuasion)
  – What are the *data* composed of, statistically? (continuous, categorical, time series, *etc.*)
Omaha System Data

• Is categorical (problems, interventions, outcomes), and hierarchical
• Can be longitudinal as well as cross-sectional
• Can characterize populations as well as individuals
• *Initial recommendations*: Line graphs or pie charts, using color-coding for categories
• Can take “classic” or “modern” approaches to visualizing this data
Longitudinal Omaha Data

• *Classic approach*: line graphs or pie charts showing levels and proportions of problems, interventions, *etc*.

• *Modern approach*: “Stream graphs” that can layer many categories at once:
Categorical Omaha Data

- **Classic approach**: Tables or bar charts to show levels of knowledge-behavior-surveillance within a patient.
- **Modern approach**: “Sunburst graphs”, which can show characteristic shapes for different constellations of patient issues.
Creating Visualizations

• You do *not* have to be an artist!
• Basic applications like Microsoft Excel can do a lot, if you are willing to tweak default settings
• Other software exists to create sophisticated visualizations
• *Free* software also exists, but can require more effort to learn
Seattle Area Sales Activity

The line shows the median home sales price for the Seattle area. The individual sales are shown as circles, color coded by zip code. Highlighting zip code 98106 confirms this includes the more expensive parts of town.
ggplot2 (for R)
D3.js is a JavaScript library for manipulating documents based on data. D3 helps you bring data to life using HTML, SVG and CSS. D3’s emphasis on web standards gives you the full capabilities of modern browsers without tying yourself to a proprietary framework, combining powerful visualization components and a data-driven approach to DOM manipulation.

Download the latest version here:

- d3.v3.zip


Thank You!

piecz@umn.edu