Two principled approaches to data visualization

Stat Bytes
Jerzy Wieczorek
7/8/2015
How many 6s are there?
Preattentive processing

(Alberto Cairo, *The Functional Art*)
What could be improved here?

(Total Volume of Retail Trade Turnover by Formation Sources)

(National Statistical Service of the Republic of Armenia)
Total Volume of Retail Trade Turnover by Formation Sources

- Shops
- Agricultural products markets
- Other trade units
- Kiosks
- Consumer goods markets

Percent of total
What could be improved here?

(National Statistical Service of the Republic of Armenia)
Crude marriage and divorce rates of city Yerevan and RA marzes, 2011

- Vayots dзор
- Armavir
- Ararat
- Kotayk
- Yerevan
- Geharkunik
- Aragatson
- Shirak
- Tavush
- Lori
- Syunik

Per 1000 population
Some preattentive-processing principles

• Colors should be few, easily-distinguished, & meaningful
• Use direct labels, not legends
• Map data to an easy-to-compare perception (length > angle)
• Show comparisons directly
• Rank informatively
Useful features of a color scheme?
Useful features of a color scheme

• Different schemes for Qualitative vs Quantitative data
• Among Quantitative data, different schemes for Sequential vs Divergent data
• Colorblind friendly
• Print and photocopy friendly

http://colorbrewer2.org/
RColorBrewer package
R demo

• RColorBrewer
• directlabels


National Health and Nutrition Examination Survey (NHANES) 2011-2012 data subset: infants aged 0 – 6 months, complete-cases among 3 largest race/ethnic groups.

Variables: GENDER, MONTHS, RACETH, WEIGHT_KG, LENGTH_CM, HEAD_CM
What could be improved here?

(National Statistical Service of the Republic of Armenia)
Structure of Services, 2007

<table>
<thead>
<tr>
<th>Service</th>
<th>Percent of total volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communication</td>
<td>25</td>
</tr>
<tr>
<td>Transport</td>
<td>20</td>
</tr>
<tr>
<td>Financial activity</td>
<td>15</td>
</tr>
<tr>
<td>Other</td>
<td>5</td>
</tr>
<tr>
<td>Health</td>
<td>5</td>
</tr>
<tr>
<td>Education</td>
<td>5</td>
</tr>
<tr>
<td>Hotels and restaurants</td>
<td>2.5</td>
</tr>
</tbody>
</table>
The Grammar of Graphics, Wilkinson

“This system is capable of producing some hideous graphics. ... This system cannot produce a meaningless graphic, however.”

Not a list of plot types (like Excel), but a framework for describing what to plot.

Map each data variable to a perceptual variable or \texttt{aes} (position, length, color...), transform it with a \texttt{stat} (identity, bin, smooth, quantile...), plot it as a \texttt{geom} (point, line, bar...), and optionally \texttt{facet} into subplots by one or two data variables
GoG practice

Describe chart from slide 13 using Grammar of Graphics framework:

What data variables correspond to which \texttt{aes}?
What \texttt{stat} and \texttt{geom} are plotted?
Any \texttt{facet}?

Repeat for WHO and Mayo Clinic charts on next slides
WHO Child Growth Standards
Mayo Clinic Guide to Baby’s First Year

Month 1

Length (inches)

Month 6

Length (inches)
R demo

ggplot2

http://docs.ggplot2.org/

Mimic the WHO Child Growth Standards charts of
Length-for-age, by gender
http://www.who.int/childgrowth/standards/en/

and

Mayo Clinic Guide to your Baby’s First Year charts of
Length-for-weight, by month and gender
R demo

• `aes(x, y, color, linetype, size, shape, alpha)`
• `geom_point()`
• `geom_line(stat = "quantile")`
• `facet_grid(... ~ ...)`

What else can you plot with these commands and this data?

What else would you like to plot? [http://docs.ggplot2.org/](http://docs.ggplot2.org/)

Can you show all 6 variables at once? (Should you? 😊)