Topics in Data Science = Өгөгдлийн шинжлэх ухаан

Rutgers University has made this article freely available. Please share how this access benefits you.
Your story matters. [https://rucore.libraries.rutgers.edu/rutgers-lib/52378/story/]

Citation to Publisher
Version: No citation available.


Terms of Use: Copyright for scholarly resources published in RUcore is retained by the copyright holder. By virtue of its appearance in this open access medium, you are free to use this resource, with proper attribution, in educational and other non-commercial settings. Other uses, such as reproduction or republication, may require the permission of the copyright holder.

Article begins on next page
Data Visualization

Ryan Womack

Data Librarian, Rutgers University, https://ryanwomack.com

National Statistics Office of Mongolia, Ulaanbaatar, Mongolia, May 8, 2017

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.
About Me

- **Data Librarian**
- **Masters in Economics, Library Science, and Statistics**
  - working in the intersection of these fields
  - help people find and use data
  - manage data for the Libraries
- interaction with advanced scholarship and research
About Rutgers

- **Rutgers**, The State University of New Jersey
- New Jersey, 8.8 million people, between New York and Philadelphia
- Rutgers founded in 1766, over 250 years old
- Carnegie Classification: Research - Very High Research Activity, the highest classification (R1)
- more than 100 major programs of study
- over 68,000 students (7,500 international from over 120 countries)
- Top 100 university in the world (Times Higher Education, Shanghai Ranking, CWUR, and others)
About Rutgers
About IASSIST

IASSIST is the International Association for Social Science Information Services and Technology

A community of over 300 data professionals working in information technology, libraries, data services, research & higher education, government, non-profit and private research sector

IASSIST seeks to

- Foster and promote a network of excellence for data service delivery
- Advance infrastructure in the social sciences and beyond
- Provide opportunities for collegial exchange of sound professional practices

Annual conference is a forum for presentation and networking (US, Canada, Europe, next Asia?)

IASSIST is seeking increased membership from Asia
Why Data Visualization?

Data visualization can:

- provide clear understanding of patterns in data
- detect hidden structures in data
- condense information
Anscombe’s Quartet

For example, see Anscombe’s quartet (image source: http://commons.wikimedia.org/wiki/File:Anscombe%27s_quartet_3.svg):

![Graphs of Anscombe's Quartet](http://commons.wikimedia.org/wiki/File:Anscombe%27s_quartet_3.svg)
Astronomical observations, charts, and maps led in graphical innovation prior to 1800. See also Classic Data Visualizations.

William Playfair is the pioneer of the line chart, bar chart, time series plots, and pie chart.

- Playfair, W. (1786). *Commercial and Political Atlas: Representing, by Copper-Plate Charts, the Progress of the Commerce, Revenues, Expenditure, and Debts of England, during the Whole of the Eighteenth Century*,


Playfair Examples
Playfair Examples

CHART of EXPORTS and IMPORTS to and from the EAST INDIES
From the Year 1700 to 1780 by W. Playfair

The Bottom Line is Divided into Years the Right hand Line into HUNDRED THOUSAND POUNDS
Minard

Charles Joseph Minard was the next influential data graphic creator after Playfair.

- Minard’s flow map of Napoleon’s Russian campaign is celebrated by Tufte and others as one of the greatest information graphics.
- It embodies an ideal of highly compressed informative elements, presented with style.
- Six variables: size, location in 2 dimensions, the direction of the army, temperature, date [and group]
- However, this is a one-off design that crosses into Infographics, but it can be reproduced in R and other software.
Minard Examples
Minard Examples
In the 20th century, statisticians such as Ronald Fisher and John Tukey continued to advance graphical methods for the analysis of data.

Fisher emphasized plotting the data to understand relationships.

Tukey’s *Exploratory Data Analysis* emphasized the use of graphics to understand the data during analysis, rather than the final presentation to an outside audience.

Tukey created the box and whiskers plot and the stem and leaf plot.
From Barchart to Dot Plot

- The Cleveland dot plot
  - use to compare labeled quantities, ordered lists
Barchart v. Dot Plot
Visualizing Distributions of Data

- **Box and Whiskers Plot**
  - illustrate quantiles and outliers. There is also a Tufte version.

- **Violin plot**
  - Blends density information with box and whiskers style (in an artistic manner)
Box Plot v. Violin Plot
Visualizing Categorical Data

- Beyond the pie chart
- The *mosaic plot* allows multiple categories to be displayed on the same graph, but can be complicated to interpret.
- The *spineplot* is a variant of the mosaic plot, plotting proportions in 2 dimensions.
Pie Chart v. Mosaic Plot
Maps and Glyphs

Maps are obviously an important and widespread way of presenting data.

- We examine a few examples of choropleth maps, in which shading indicates data levels.
- See also Interactive Maps in R and 5 kinds of Interactive maps in Plot.ly for further exploration.

**Glyphs** present iconic representations of data elements.

- **Weather maps** often use glyphs.
- A more dynamic example is here.
- As an R example, consider Chernoff faces and the ap1pack package. Also, Smiley faces [and many more graph variants in this chapter].
Choropleth Map v. Chernoff Faces
Interactive Data Viz - Principles

- Why aren’t all of our graphs interactive?
- *Brushing* is used to select data points and track them through various analyses.
- Drilling down, zooming, and subsetting are also interactive techniques.
- Data displays can be linked so that a selection in one panel modifies the output displayed in another panel.
- Interactivity is especially useful for data exploration, studying multidimensional relationships.
Interactive Data in Practice

There are many R packages that allow for interactive data work in a graphical user interface, including:

- **playwith** - versatile package that works with any graphics function. Graphics can be explored, edited, and exported.
  
  - requires separate installation of GTK+ on your computer [method varies by OS]
googleVis

In many contexts, visualizing the relationships between data elements is made easier by viewing related data interactively.

- Making this easy are googleVis and other “Vis” packages, e.g. bdvis for biodiversity or rainfreq.
- A Library example - comparing selected ARL Statistics for public CIC universities
Interactive Data on the Web - Rcharts

- **Rcharts** is a package that uses javascript to create interactive visualizations.
- Lattice-style commands are used.
- The package can output javascript for use in an HTML page.
- Some commands depend on supplemental javascript libraries that must be installed, such as NVD3
- Can embed in documents too, with slidify
The shiny package is developed by the Rstudio folks

You can learn shiny in half a day via the online tutorial

More custom control of the design is possible with shiny, in comparison to other do-it-all packages

Graphics use familiar R syntax (including ggplot2), with wrappers to implement web functionality

Every shiny app has the same structure: two R scripts saved together in a directory [ui and server files]

You must install the shiny server to deliver pages via the web
Interactive Data on the Web - shiny, cont.

- There are samples built into the shiny package.
- You can see more in the shiny gallery
- Rcharts works with shiny too.
The `ggvis` package is ALSO developed by the Rstudio folks

Think `ggplot` meets `shiny`

Similar syntax to `ggplot`

Some ability to add interactive controls

Can embed in `shiny` for web access
Other (non-R) options for Web visualization

- D3.js, free at http://d3js.org/
- Inkscape, free at https://inkscape.org/
- Tableau, free 1-year student license at http://www.tableau.com/academic/students
- Plot.ly environment at http://plot.ly
- Datavisualizationforall - free online book with methods
Population pyramids are one example where \textit{interactivity} + \textit{animation} = \textit{insight}.

- Populationpyramid.net - for all countries, basic animation
- The German Population Pyramid from Destatis is even more interactive
- Doing it in R is possible with these instructions (Part 1) and (Part 2)
- The \texttt{ggvis} package is ALSO developed by the Rstudio folks
Recent Developments and Tools

- Using R to Gain Insights into the Emotional Journeys in War and Peace
- Hourly Heatmap
- ropensci - for open science
- Digital Panopticon - some tools
Big data presents special issues for data visualization

- While many techniques and graphics are the same, exploration and plotting must be optimized for the size of the data set
- Representation of the complexity of the data may require special techniques
  - hexbin
  - bigvis
BIGVIS was an experimental package by Hadley Wickham to deal with the issues of Big Data

- There is a Preprint and R Meetup presentation by Hadley Wickham
- Complete code is available at https://github.com/hadley/bigvis-infovis
- Target: process 100 million observations in under 5 seconds.
- Fundamental principle: No need for more data points than there are pixels on the screen.
- “ggstat” package has been mentioned as a future project that will incorporate these ideas.
BigVis steps

- Condense (bin, condense)
- Smooth (smooth, best_h, peel)
- Visualize (autoplot plus standard methods)
**Trelliscope (DeltaRho, was Tessera)**

**DeltaRho** is developed by Purdue, Pacific Northwest National Laboratory, and Mozilla. Launched in November 2014, this project holds a lot of promise.

- Running in the R environment, Tessera provides its own commands that execute across a cluster, easing the burden of analysis in this environment.

- The **datadr** package “divides and recombines” in a manner similar to MapReduce, providing a simplified interface to Hadoop.

- Tessera has its own visualization interface, **Trelliscope**, that can handle views across many variables and observations. Described in this [paper](#).

- Try the **quickstart** to begin

- Live demo is [here](#).
Keep Exploring

Data Visualization represents a nearly infinite world of possibility for exploration:

- plunge into programming
- deep dives into data
- indulge in interactivity
- ...have fun and keep learning! [e.g., R-bloggers.com]
Data Visualization => Future

- Tools and packages are making complex and interactive visualization easier
- Evolving standards and expectations will make data visualization the norm
- Understanding data visualization is a component of data/information literacy
- Questions? Discussion?
Асуулт байна уу?
Маш их баярлалaa!
Намайг дагаарай...

- https://youtube.com/librarianwomack
- https://www.linkedin.com/in/ryanwomack
- https://twitter.com/ryandata
- https://ryandata.wordpress.com