Data Literacy for All, with R

Ryan Womack

May 29, 2018

First presented at IASSIST 2017, Lawrence, Kansas updated for IASSIST 2018, Montréal, Québec

May 29, 2018

Ryan Womack, Data Librarian

Rutgers University, rwomack@rutgers.edu

https://ryanwomack.com

Outline

1. What is data literacy?
2. Packaging Data
3. Packaging Functions
4. Designing functions for data literacy and data exploration
5. Interactivity - ggvis and shiny
6. Interactivity - population pyramid
7. Building Instruction around Interactive Functions - the Rutgers Future Scholars
8. Group Discussion

What is data literacy?

see associated slides, Data_Literacy_for_All_Womack.pdf

Statistical Literacy

Statistical literacy involves the following elements: - Literacy (reading, but also including handling graphs, charts, and tables, and other forms of textual evidence) - Statistical knowledge - Mathematical knowledge - Context - Critical Skills

Data Literacy

Data literacy requires statistical literacy, but with more emphasis on data wrangling and data exploration.

Prado and Manzi –> "access, interpret, critically assess, manage, handle and ethically use data"

That is a lot. For education, often we need to focus on the core of statistical literacy in a gentle way before getting into the mechanical details of our software tools.

As data professionals, we can do some of the work for our audiences to mask the complexity of the tools and highlight the data itself.

Packages

Some guides:

RStudio and RMarkdown http://rmarkdown.rstudio.com/

Build Project https://support.rstudio.com/hc/en-us/articles/200526207-Using-Projects and https://support.rstudio.com/hc/en-us/articles/200486508-Building-Testing-and-Distributing-Packages https://bookdown.org/rdpeng/RProgDA/building-r-packages.html

Packages - what are they

The structure of a package

How RStudio helps the process

From the command line: use R CMD build packagename R CMD INSTALL packagename

A simple way to get started: https://hilaryparker.com/2014/04/29/writing-an-r-package-from-scratch/

Hadley Wickham's R Packages (Recommended) http://r-pkgs.had.co.nz/

Official CRAN documentation https://cran.r-project.org/doc/manuals/R-exts.html

Packaging Data

We will prepare and extract some data with WDI.R

Any data can be stored as an R data file and bundled in a package.

The "devtools" package has a useful use_data() command.

Note the drat package (https://journal.r-project.org/archive/2017/RJ-2017-026/index.html) allows for access to larger datasets. CRAN has a 5 MB limit.

Packaging Functions

We will walk through:

- Writing a function
- Writing documentation
- Saving a function
- Checking and testing...
- Building and sharing a package

How to distribute your package

We can distribute packages several ways

Local, e.g. library(mypkg, lib.loc = "f:/R-packages")

RForge - like a Github for R

Github quick guide http://rogerdudler.github.io/git-guide/ – git init – git add* – git commit -m "message" – git remote set-url origin https://github.com/ryandata/test11111.git – git push origin master

install_github("ryandata/test11111")

Designing functions for data literacy and data exploration

We want to build functions that the end user can quickly apply to their own data exploration needs.

One package designed for this is mosaic

Interactivity - ggvis and shiny

ggvis provides a lightweight way to introduce some dynamic, interactive elements to your plots

shiny provides a suite of tools to design customized interactive web-accessible data sites, while retaining R for data analysis.

See the shinyapp.R file.

Interactivity - population pyramid

This example illustrates the use of interactivity to uncover population patterns.

See http://www.arilamstein.com/blog/2016/06/06/idbr-access-us-census-bureau-international-data-base-r/

https://walkerke.github.io/2014/06/rcharts-pyramids/

Some of my other R materials may be found at http://libguides.rutgers.edu/data_R

Building Instruction around Interactive Functions - the Rutgers Future Scholars

Materials for an introduction to data for Rutgers Future Scholars can be found at

http://libguides.rutgers.edu/rutgersfuturescholars

The pyramid.R file creates a set of functions using the Census International Data Base API. These functions can be used to analyze population growth for any country and dates desired.

Working with **real** and **live** data in a simple way motivates rapid growth in data literacy.

We can build customized kits like this for many purposes, such as international data. The ropensci project is a leading example of the explosion of packages being created for open data and other innovative applications. Data literacy should be a part of this trend.

Some of my other R materials may be found at http://libguides.rutgers.edu/data_R

Open Discussion

- What do you see as areas that would benefit from this guided approach to data literacy?
- Are there particular sources you would use?
- Specific techniques you would focus use?
- Which audiences would you design for?
- What are the challenges you would face?
- What is the role for data professionals?

Live PollEverywhere at https://pollev.com/ryanwomack427

Lagniappe - Distributing files via PirateBox

This is not really that difficult if you follow the instructions here: https://piratebox.cc/openwrt:diy#post-installation

with a little customization help for your SSID and Home Page here: https://www.youtube.com/watch?v=vc6od_2mess https://www.youtube.com/watch?v=asCC12QAHr0#t=19.562452

Keep exploring!