Introduction to R for data analysis

Research Computing Center and the Dept. of Human Genetics University of Chicago

Peter Carbonetto
pcarbo@uchicago.edu

August 1, 2017

http://tinyurl.com/y7bqsl2x

Outline of workshop

1. Initial setup (10 min).
2. Introduction (5 min).
3. An exploratory data analysis in R:
    a. Set up your R environment (15 min).
    b. Import & inspect the Divvy station data (20 min).
    c. Import & inspect the Divvy trip data (20 min).
    d. Create plots using the station & trip data (20 min).
    e. Analyze U of C biking trends from Divvy data (20 min).
4. Recap (5 min).

Getting started…

- Initial setup:
  - Wireless Internet
  - Power outlets
  - RCC cluster access
  - Install R and/or RStudio
  - Download git repository
  - Download data files
  - *Windows difficulties*

- Introduce yourself to your neighbors.
- Breaks.
- Ask me questions
  - *Keyboard shortcuts*
- Pace & experience levels.

Aims of workshop

1. Understand why R has become important for many areas of research.
2. Set up your laptop or RCC cluster account to do interactive programming in R.
3. Learn how to install & use R packages.
4. Learn basic elements of R data analysis by example.
5. Work with “R Markdown” documents.

What this workshop does ***not*** cover

- How to write R code.
- Fundamentals of statistical analysis.
- Syntax and grammar of the R programming language.
- How to submit R computing jobs on the RCC cluster.
- High-performance computing in R (“Big Data”).

The Software Carpentry approach

1. Learning through “live coding.”
   - Especially learning from our mistakes!
2. Hands on—*using your own computer*.
3. Lateral knowledge transfer.
4. Collaborative note-taking (e.g., Etherpad).

Key features of R

1. R is based on the statistical programming language **S**.
2. R is **open source** (GPL).
3. R is a **programming environment**.
4. RStudio provides a **free IDE** = integrated development interface.
5. R is **community-driven**.

R is community-driven

10,048 active packages

5,871 package maintainers

188 updates last week

6,836,151 downloads last week

Rcpp – 0.12.9
23 days ago by Dirk Eddelbuettel
Seamless R and C++ Integration

ggplot2 – 2.2.1
a month ago by Hadley Wickham
Create Elegant Data Visualisations Using the Grammar of Graphics

digest – 0.6.12
10 days ago by Dirk Eddelbuettel
Create Compact Hash Digests of R Objects

tibble – 1.2
5 months ago by Kirill Müller
Simple Data Frames

lazyeval – 0.2.0
8 months ago by Hadley Wickham
Lazy (Non-Standard) Evaluation

assertthat – 0.1
3 years ago by 'Hadley Wickham'
Easy pre and post assertions.

BH – 1.62.0-1
3 months ago by Dirk Eddelbuettel
Boost C++ Header Files

R6 – 2.2.0
4 months ago by Winston Chang
Classes with Reference Semantics

magrittr – 1.5
2 years ago by Stefan Milton Bache
A Forward-Pipe Operator for R

plyr – 1.8.4
8 months ago by Hadley Wickham
Tools for Splitting, Applying and Combining Data

jsonlite – 1.2
a month ago by Jeroen Ooms
A Robust, High Performance JSON Parser and Generator for R

stringr – 1.1.0
6 months ago by Hadley Wickham
Simple, Consistent Wrappers for Common String Operations

curl – 2.3
2 months ago by Jeroen Ooms
A Modern and Flexible Web Client for R

stringi – 1.1.2
4 months ago by Marek Gagolewski
Character String Processing Facilities

scales – 0.4.1
3 months ago by Hadley Wickham
Scale Functions for Visualization

reshape2 – 1.4.2
3 months ago by Hadley Wickham
Flexibly Reshape Data: A Reboot of the Reshape Package

dplyr – 0.5.0
7 months ago by Hadley Wickham
A Grammar of Data Manipulation

data.table – 1.10.4
5 days ago by Matt Dowle
Extension of `data.frame`

colorspace – 1.3-2
2 months ago by Achim Zeileis
Color Space Manipulation

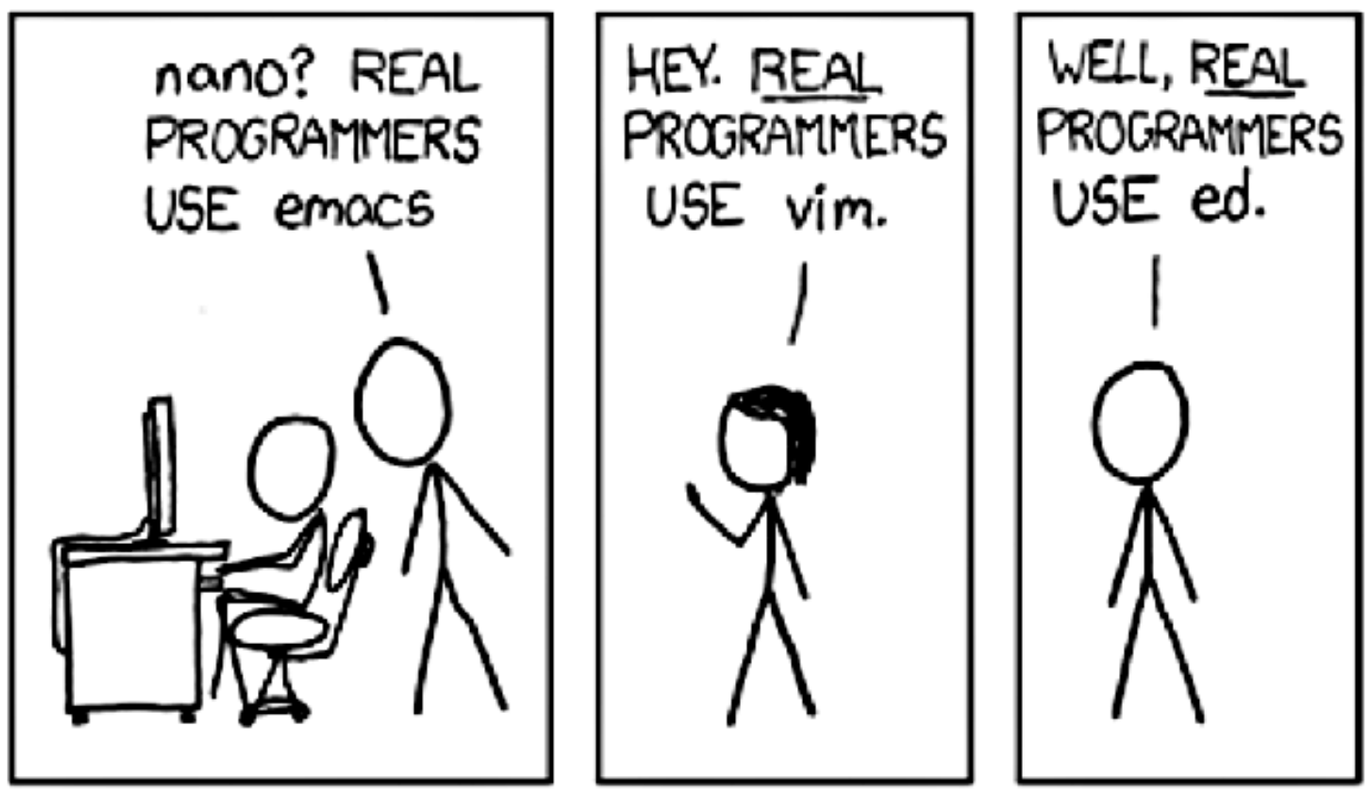
RColorBrewer – 1.1-2
2 years ago by Erich Neuwirth
ColorBrewer Palettes

source: www.r-pkg.org (accessed Feb. 6, 2017)

Options for setting up your *interactive* R data analysis environment

1. On your laptop:
   - R + text editor + X Window System
   - IDE #1: RStudio Desktop
   - IDE #2: Jupyter notebook + R kernel
2. On the RCC cluster:
   - R + text editor + ThinLinc
   - IDE #1: RStudio Desktop + ThinLinc
   - IDE #2: RStudio Server (*limited availability*)
   - IDE #3: Jupyter notebook with R kernel
3. *Non-interactive* data analysis:
   `R CMD BATCH my_analysis.R.`

There is no best tool—use *whatever works for you*.



https://xkcd.com/378/

Some general advice

1. Use packages—don't reinvent the wheel.
2. `help(cool_function)` & stackoverflow.com.
3. Use **midway2**, not midway1.
4. Email help@rcc.uchicago.edu R help on the RCC cluster.
5. Learn to avoid loops as much as possible; e.g., use `apply()`, `lapply()`, `tapply()`, `do.call()`.
6. The "defaults" in R are often not what you want—check the function inputs carefully.
7. Use R Markdown or Jupyter notebooks to document your analyses.
8. Document your setup—start with `sessionInfo()`.
9. See "Great resources for R."

After the workshop

I'm happy to talk individually about using R for your research project.

You will receive an email requesting feedback on this workshop. **Please complete this survey!**