Introduction and Course Motivation

PSS SUMMER SCHOOL

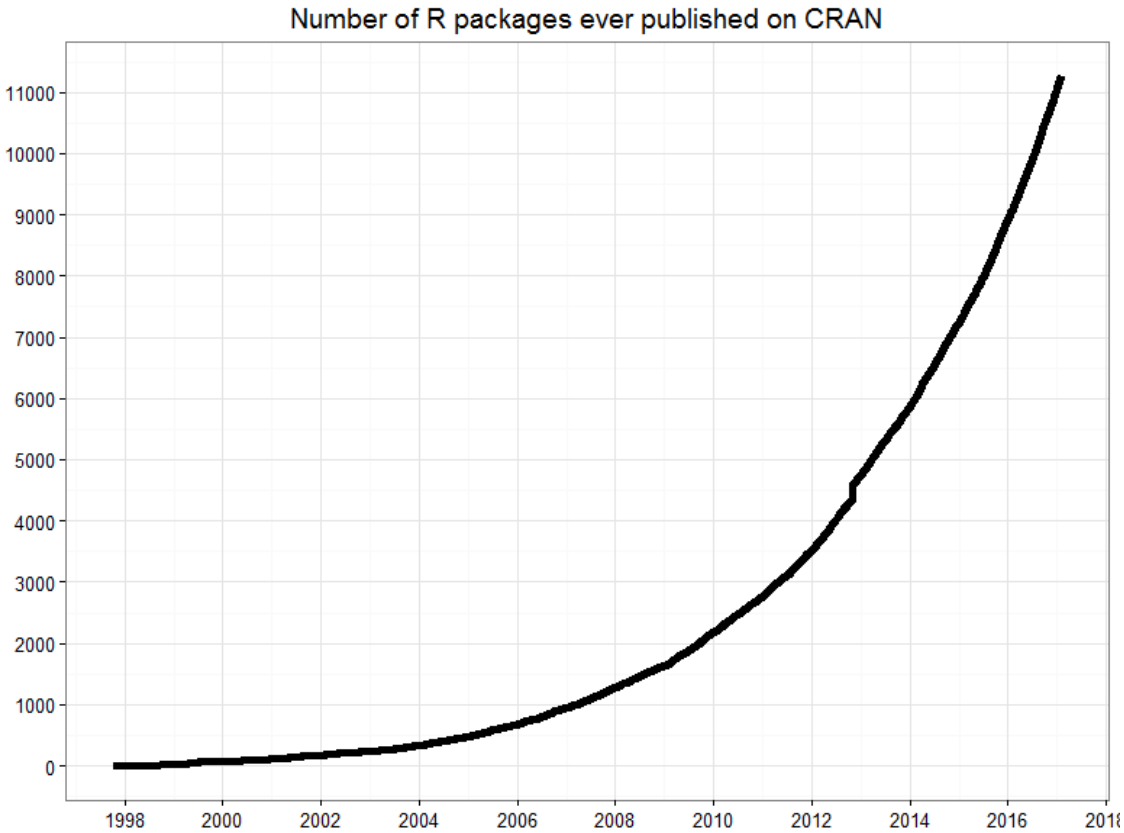
Anton Badev, Daniel Nikolic, Justin Skillman

June 21, 2017

What is R?

- Open source language first released in 1993
- Extensive library of statistical functions
- General purpose programming language*
- Adept at handling and processing *complex* data
- Dynamically typed & interpreted (vs statically typed & compiled)
- Estimated userbase of 250,000 to 2 million

Packages in R



Source: Smith, David. "CRAN now has 10,000 R packages. Here's how to find the ones you need.". RevolutionAnalystics, 2017. Web. 23 May. 2017.

R vs Spreadsheet Applications (Excel)

R

- Reusable/documented routines
- Statistics libraries for analysis
- Built-in debugging
- Easily interact with many types of databases
- Perform complex operations

Microsoft Excel

- User friendly
- Integrates with MS Office Suite
- Fast for small projects
- Standard software found on workstations

R vs SAS vs STATA vs MATLAB

R

- Flexible programming capabilities
- Extensive graphics packages
- Free open source software
- Larger possibility for errors or problems in 3rd party packages

STATA

- Simple scripting interface
- Standard statistical methods implemented
- Minimal keystrokes
- Only holds one dataset in memory

SAS

- Broad statistical capabilities
- Lower learning curve than general programming languages
- Widely used in academia
- Proprietary source code

MATLAB

- Optimized routines
- Extensive documentation and support from MathWorks
- Prohibitively expensive
- Proprietary source code

R vs Python

R

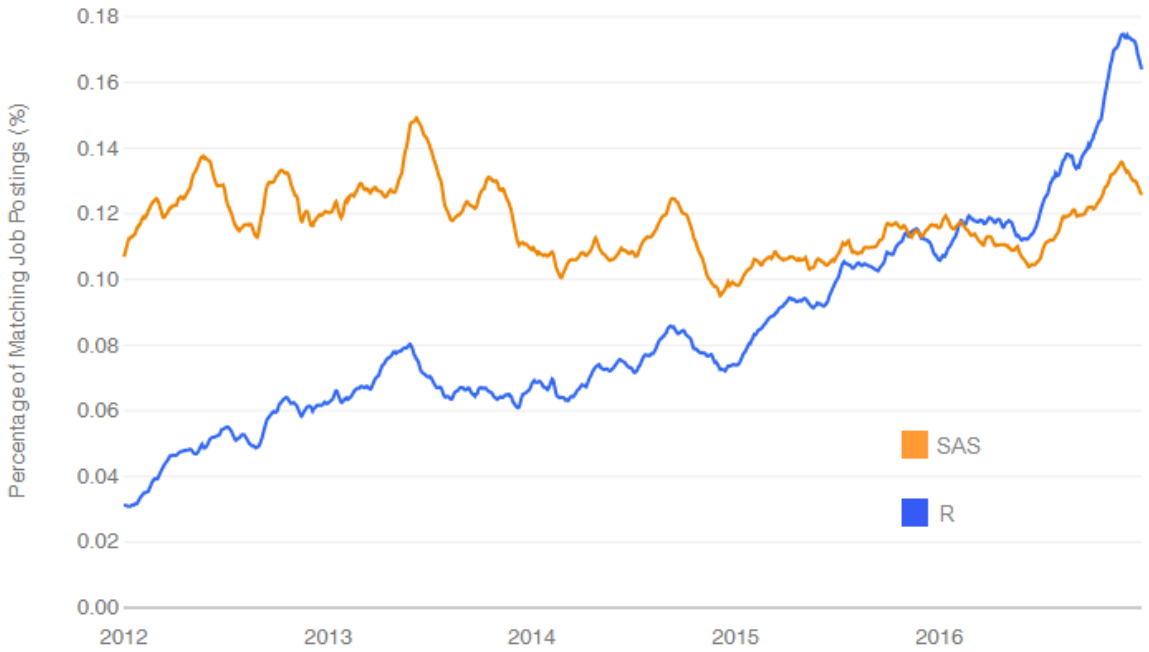
- Intended primarily for statistical data analysis
- Higher level
- More extensive libraries of cutting edge statistical methods
- More developed graphics capabilities
- Open source

Python

- Intended for general purpose programming
- Slightly lower level (still interpreted)
- Greatly improved (recently) tools for panel dataset analysis: *pandas* & *numPy*
- More developed machine learning tools (*scikit-learn*)
- Open source

Note: These are all points of, sometimes fervent, debate among each community

Software Popularity



Source: Indeed.com

What is LaTeX?

- LaTeX is a high-quality typesetting system designed for producing scientific documents
- Extensive capabilities for handling technical content including: mathematical formulas, diagrams, tables, graphs, and bibliography
- LaTeX refers to a set of encoding and tagging conventions, and is implemented in the TeX macro language
- This is the software you will use to write-up results and analysis for the assignments

Course Goals

- Mastery of basic programming concepts
- Learn to approach problems algorithmically
- Learn the tools to perform data analysis in R
- Implement a complete data analysis project

Course Syllabus

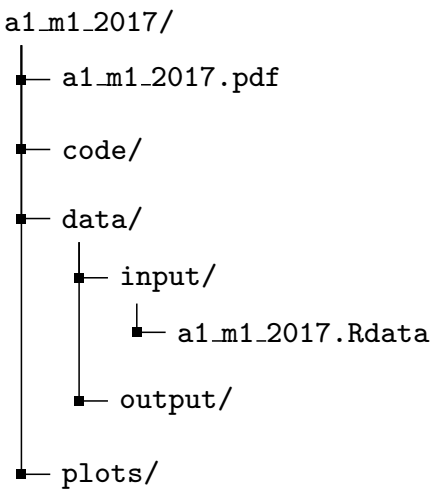
Date	Topic	Subtopics
6/21	Primitives	Primitive Data Types and Operations, Control Flow, Scope, R "vocabulary"
6/22	Functions	Functional Programming (the *apply* family), Functional Operators, Vectorization
6/23	Debugging	Condition Handling, *traceback* and *browser*, RStudio's Error Inspector
6/26	Data Manipulation	Methods for Fast Subsetting, Transformations, *reshape2*, *dplyr*
6/28	Visualization	Plotting with *ggplot2*: Plot Types and Formatting, Tabulation
6/30	Performance	Parallelization, Memory-usage, Code optimization
7/5	Final Exam	Group Presentations

Assignment Submission

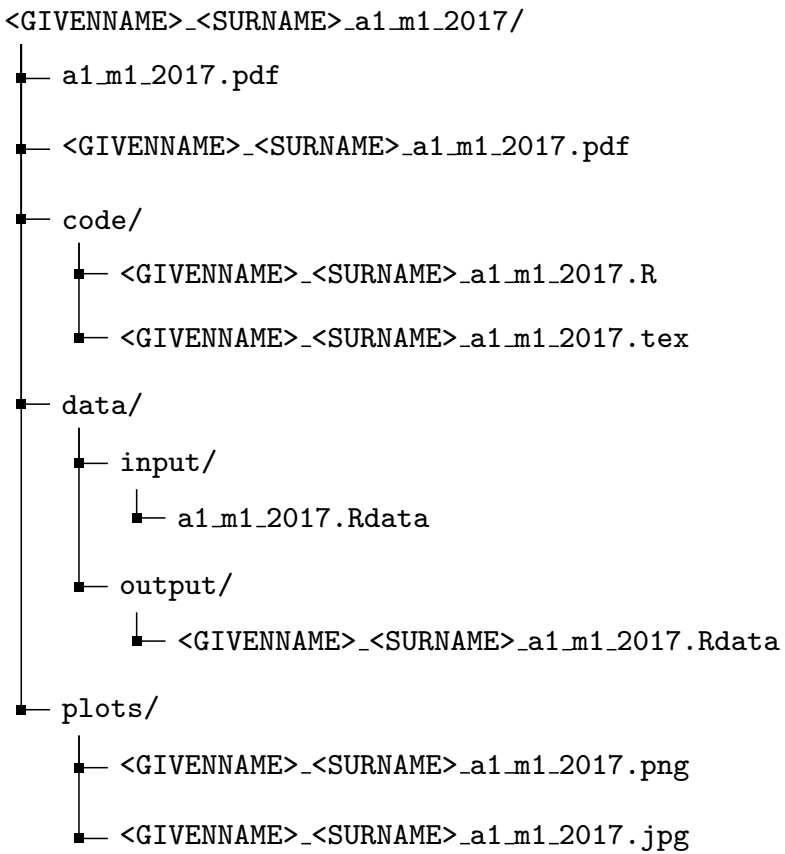
- I have emailed you a zip directory that contains everything you will need to complete the assignment
  - Assignment document
  - Directory structure
  - Data to be used
  - Lecture slides
- *DO NOT* change the directory structure
- To submit your assignment, email the zipped folder to *pss.ss.hw.submit@sp.frb.gov*

Example of Directory Setup

(zip to students)



(zip from students)



Example of Directory Structure Code

- To make code more portable follow the setup below
- We should only need to change the home directory line

```
# Set directory structure
home_dir <- "Documents/R/summer_school/assignment_1/"

code_dir <- paste0(home_dir,"code/")
data_in_dir <- paste0(home_dir,"data/input/")
data_out_dir <- paste0(home_dir,"data/output/")
plot_dir <- paste0(home_dir,"plots/")
```

Code Etiquette

- Meaningful names for everything
  - Bad: `foobar(x,y)`
  - Good: `compute_distance(x,y)`
- Comment your code, why is better than what
  - Bad: `x + y #Adds x and y`
  - Good: `x + y #To be used in compare_results function`
- Delineate sections of code using spacing or comment lines
- Consistency is most important

Help Resources

- Using Google *effectively* is a skill
- Distilling problems down to smaller and smaller parts is better
- Understand exactly where the error occurs using R's output
- Bad search: `make a word into two words`
- Good search: `R split text strings`
- `<programming language> <verb> <specific keywords>`
- Don't be afraid to go through multiple pages of Google results
- Refine your search as you begin to understand terminology better

Help Resources

- Websites to look for
  - stackoverflow.com
  - inside-r.org
  - r-bloggers.com
- Base R help function `help(<your command>)`
- Copy/Paste error into Google
- Office hours