Stats 195

Intro to R Programming

Haben Michael (haben.michael@stanford.edu)

Spring 2017

Outline

Course Logistics

Overview of R

Pros

Cons

- ▶ will honor prereqs (no programming assumed)
- ▶ website will contain scripts from class, hw, solutions etc.: https://haben-michael.github.io/stats195/
- ▶ see website for a list of similar courses
- ▶ will periodically break for about a couple minutes to let you try things in R, will take questions at that time
- ▶ auditors welcome as long as there are seats
- ▶ requirement for satisfactory grade: 3 hw assignments, each with a score of at least 60%
- ▶ solutions will be posted quickly so late hw not accepted except as required by university policy
- ▶ review the CS honor code on course website

Schedule (more details on coursework)

- ▶ first 2-3 classes: using R like a calculator, then using R as a programming language
- ▶ classes 4-6: fundamental applications: exploratory data analysis, some statistical tools, text processing
- ▶ 6/7-8: topics TBD (see coursework for previous years), may need to do dplyr and/or ggplot2 here

Outline

Course Logistics

Overview of R

Pros

Cons

A programming language oriented toward statistics and data analysis

- ▶ data types, preloaded routines, built-in graphics
- ▶ interactive/interpreted, a way to converse with the computer

- uses: data mining/analysis, linear and non-linear modeling, 2-d graphics, classical statistics, time series analysis, classification, many more.
- practitioners: statisticians and data analysts of all stripes in both academia and industry, finance, bioinformatics/genomics, many more.

History

- ▶ 80s/90s: S/S-Plus at Bell Labs (John Chambers)
- ▶ late 90s/2000s: R open-source implementation of S specification out of NZ (**R**obert Gentleman, **R**oss Ihaka)

Outline

Course Logistics

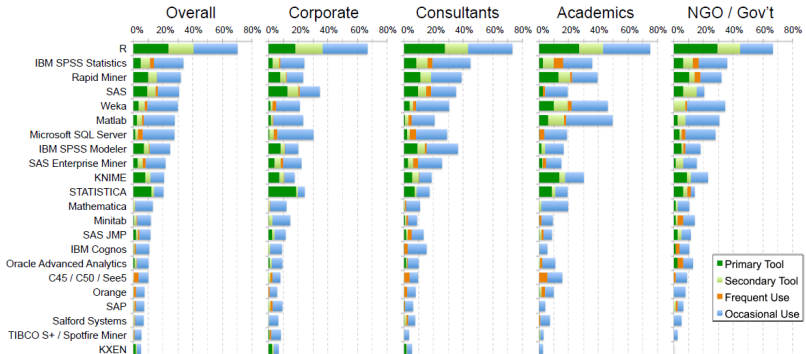Overview of R

Pros

Cons

popularity

Figure: Rexer Analytics Data Miner Survey 2013

free and open source

- ▶ programming languages commonly open source, but this has not been so in the area of statistical computing: competitors MATLAB, SAS, STATA, SPSS etc. are costly

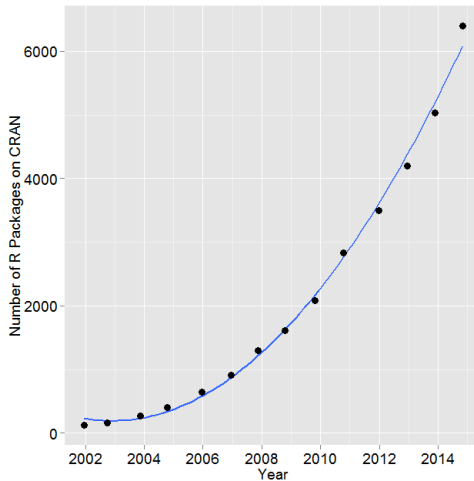large base of contributors

▶ many packages for many applications

Figure: (ref: r4stats.com)

▶ competitors also have large contributor bases although the applications themselves are costly, R cuts the vendor out of the picture

▶ good for stability/longevity of the language

▶ at least for statistics (also e.g., bioinformatics) R seems to be the first choice for the implementation of new research

good online support (compare stackexchange Q&As–almost entirely R–to SAS or MATLAB forums)
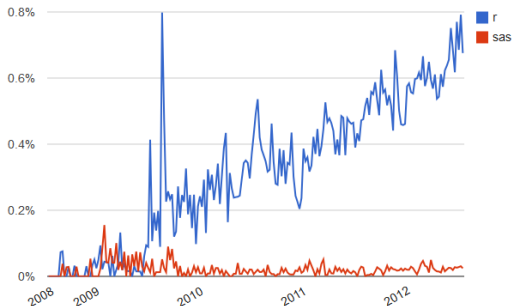

Figure: Number of R- and SAS-related posts to StackOverflow by week (ref: r4stats.com)

Outline

Course Logistics

Overview of R

Pros

Cons

Cons:

- ▶ learning curve
- ▶ slow with large data, memory limits, parallelism not built in
- ▶ developed by statisticians–quirks that can clash with common programming language customs, e.g., array indexing from 1, "." naming convention, strings are atomic types (but still much better than SAS/SPSS/etc. macro languages), no real language definition