improveR

Contents

1 Introduction

1.1 Blurb

This short course covers the core skills required for a budding R user to develop a strong foundation for data analysis in the RStudio environment. Within the framework of a reproducible research workflow we will cover importing and cleaning data, efficient coding practices, writing your own functions and using the powerful `dplyr` data manipulation tools.

1.2 Key Topics

- Reproducible Research
- R Studio and project management
- Importing and cleaning data
- Good coding practices in R
- standard control structures
- Vectorisation and `apply` functions
- Writing your own funcitons
- Data manipulation with `dplyr`
- Piping/chaining commands

1.3 Course information

Intended audience Anyone interested in quantitative data analysis using open source tools.

Prior knowledge Knowledge of R (as covered in R: An introduction).

Resources Course handbook

Software RStudio & R 3.1.2

Format Presentation with practical exercises

Where next? R:

2 Reproducible Research [presentation only]

2.1 Why?

- Reinhart Rogoff Excel spreadsheet

3 Set-up [presentation and practical]

3.1 RStudio

3.2 Project management

3.3 Literate programming

3.3.1 Consistent coding style e.g.:

- Google style guide
- Hadley Wickham's style guide

3.3.2 Commenting

3.4 * Bonus section: github

3.5 PRACTICAL: new R project

- personalise RStudio settings (don't save .Rdata etc)
- new project folder with subfolders (data, figures, scripts)
- new Rproject

4 Workflow

4.1 Importing data

The original data should be read-only!!

- url
- unzip
- (colClasses)

4.2 Data tidying

- gather/ spread

4.3 PRACTICAL: Import and clean some data

- download and import data
- do some `tidyr` stuff with it
- think about commenting and file structure!

5 Efficient Coding

5.1 Standard control structures

5.1.1 Conditional execution

5.1.2 Looping

5.1.3 PRACTICAL

5.2 Vecotrisation and`apply` family of funcitons

5.2.1 PRACTICAL

benchmarking apply vs for loops

5.3 Writing your own functions

5.3.1 objects, types, environments

5.3.2 passing arguments

5.3.3 PRACTICAL

5.4 Data manipulation with `dplyr`

5.4.1 Subsetting

- `filter`
- `sample`
- `slice`
- `distinct`
- `select`

5.4.2 Grouping

- `group_b`

5.4.3 Summarizing

5.4.4 Making new variables

- `mutate`

5.4.5 Piping/chaining daisies

5.4.6 PRACTICAL

5.5 FINAL PRACTICAL

something along the lines of:

- Fun1: a function to be called in summarize or mutate (e.g. z-score)
- Fun2: a chain (that calls Fun1), and then filters the table in some way
- Fun3: a plotting function that takes the result of Fun2