Well-documented data analysis with R Notebooks and the tidyverse

Mike Trizna, Consortium for the Barcode of Life
Smithsonian Peer-led Bioinformatics Series
November 22, 2016

Prerequisites

- R Studio, Version 1.0 or greater
  - Download and install from rstudio.com. Current version is 1.0.44.
- Once R Studio is installed, go to RStudio console and enter:
  - install.packages("tidyverse")
  - install.packages("Lahman")

Talk Outline

- Who is Hadley Wickham, and what is the "tidyverse"?
- R Notebook introduction
- Background on Lahman baseball database
- The useful verbs of dplyr
- The pipes of magrittr
- Other odds and ends from tidyverse

Hadley Wickham

Hadley Wickham
hadley

Follow

Overview | Repositories 196 | Stars 110 | Followers 8k | Following 7

Pinned repositories

tidyverse/ggplot2
An implementation of the Grammar of Graphics in R
R ★ 2.1k 793

dplyr
Dplyr: A grammar of data manipulation
R ★ 1.5k 604

devtools
Tools to make an R developer's life easier
R ★ 1.4k 470

tidyverse/tidyverse
Easily install and load packages from the tidyverse
R ★ 226 31

Screenshot from https://github.com/hadley

The tidyverse

- Name is taken from a publication by Hadley Wickham in the Journal of Statistical Science: http://vita.had.co.nz/papers/tidy-data.pdf

There are three interrelated rules which make a dataset tidy:

1. Each variable must have its own column.
2. Each observation must have its own row.
3. Each value must have its own cell.

Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

Tidy Data

Hadley Wickham
RStudio

Abstract

A huge amount of effort is spent cleaning data to get it ready for analysis, but there has been little research on how to make data cleaning as easy and effective as possible. This paper tackles a small, but important, component of data cleaning: data tidying. Tidy datasets are easy to manipulate, model and visualise, and have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table. This framework makes it easy to tidy messy datasets because only a small set of tools are needed to deal with a wide range of un-tidy datasets. This structure also makes it easier to develop tidy tools for data analysis, tools that both input and output tidy datasets. The advantages of a consistent data structure and matching tools are demonstrated with a case study free from mundane data manipulation chores.

Keywords: data cleaning, data tidying, relational databases, R.

The tidyverse

for data visualization

for chaining commands

for data manipulation

for data tidying

for data import

Working with specific types of vectors:

- hms, for times.
- stringr, for strings.
- lubridate, for date/times.
- forcats, for factors.

Importing other types of data:

- DBI, for databases.
- haven, for SPSS, SAS and Stata files.
- httr, for web apis.
- jsonlite for JSON.
- readxl, for .xls and .xlsx files.
- rvest, for web scraping.
- xml2, for XML.

Images taken from https://www.rstudio.com/products/rpackages/, and package descriptions take from https://github.com/tidyverse/tidyverse

R Notebooks

Similar to Jupyter notebooks, an R Notebook is a combination of documentation in Markdown format and executable code blocks.

R Notebooks are great for:

- Iterating quickly on code, and seeing output immediately
- Full reproducible record of data analysis and/or transformation
- Export to several different formats that can be shared

Lahman Baseball Statistics Database

- First created by journalist Sean Lahman in 2012, and updated every year
- Contains complete batting and pitching statistics from 1871 to 2015, plus fielding statistics, standings, team stats, managerial records, post-season data, and more.
- Published as a set of 24 csv files at http://www.seanlahman.com/baseball-archive/statistics/.
- Lahman R package contains all of these tables in data frame format.

The "verbs" of dplyr

- filter()
- select()
- arrange()
- mutate()
- summarize()
- group_by()