Coding Grace - Workshop 2016

Tidy Data with R

Tidy datasets are all alike but every messy dataset is messy in its own way. Hadley Wickham

Data science, at its heart, is a computer programming exercise. Data scientists use computers to store, transform, visualize, and model their data. As a result, every data science project begins with the same task: you must prepare your data to use it with a computer.

- **dplyr** - data manipulation
- **magrittr** - pipe operator

Tidy Data

Abstract from **dplyr** talk

- To make the most of dplyr, Hadley Wickham recommends that you familiarise yourself with the **principles of tidy data**.
- This will help you get your data into a form that works well with **dplyr**, **ggplot2** and `R`'s many modelling functions.

Tidy Data With R

- **tidyr** is a reframing of reshape2 designed to accompany the tidy data framework, and to work hand-in-hand with **magrittr** and **dplyr** to build a solid pipeline for data analysis.
  (from Hadley Wickham's abstract)

Principles of Tidy Data

- Tidy data was popularized by Hadley Wickham, and it serves as the basis for many R packages and functions.
- You can learn more about tidy data by reading **Tidy Data** a paper written by Hadley Wickham and published in the Journal of Statistical Software.
- Tidy Data is available online at `www.jstatsoft.org/v59/i10/paper`.

Tidy Data

Three Principles from Hadley Wickham's paper

1. Each variable forms a column,
2. Each observation forms a row,
3. Each table/file stores data about one kind of observation.

Remark:
The paper "**Tidy Data**" by Hadley Wickham (RStudio) can be downloaded from

`http://vita.had.co.nz/papers/tidy-data.pdf`

Tidy Data with R

R follows a set of conventions that makes one layout of tabular data much easier to work with than others.

- **Each column is a variable:** Each variable in the data set is placed in its **own** column
- **Each row is an observation:** Each observation is placed in its **own** row
- Each value is placed in its own cell

Data that satisfies these rules is known as **tidy data**.

Tidy Data with R

Dataframes (if you are not familiar)

- A data frame is a list of vectors that R displays as a table.
- When your data is tidy, the values of each variable fall in their own column vector.

```
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
>
```

tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions

An evolution of 'reshape2'. It's designed specifically for data tidying (not general reshaping or aggregating) and works well with 'dplyr' data pipelines.

Version:	0.3.1
Depends:	R (≥ 3.1.0)
Imports:	dplyr (≥ 0.4), stringi, lazyeval, magrittr, Rcpp
LinkingTo:	Rcpp
Suggests:	knitr, testthat, data.table, covr
Published:	2015-09-10
Author:	Hadley Wickham [aut, cre], RStudio [cph]
Maintainer:	Hadley Wickham <hadley at rstudio.com>
BugReports:	https://github.com/hadley/tidyr/issues
License:	MIT + file LICENSE
URL:	https://github.com/hadley/tidyr
NeedsCompilation:	yes
Materials:	README
CRAN checks:	tidyr results

tidyR

Abstract for tidyr

- tidyr is new package that makes it easy to "tidy your data.
- Tidy data is data thats easy to work with: its easy to munge (with **dplyr**), visualise (with **ggplot2** or **ggvis**) and model (with R's hundreds of modelling packages).

tidyr's verbs

- **tidyr** provides four main functions for tidying your messy data: `gather()`, `separate()`, `spread()` and `unite()`.

```
gather()
spread()
separate()
unite()
```