WELCOME!

CSU, Chico

Data Science Initiative

Tidy Data Hands On Workshop

07 April 2017

Workshop Leader:
Dr. Rick Hubbard
Business Information Systems Dept

PARTICIPANT's GUIDE

datascience.csuchico.edu

CSU, Chico
Data Science Initiative
Tidy Data Workshop—Participant’s Guide – April 7, 2017

PURPOSE

Easy-to-follow, lightweight, practical, pragmatic techniques for “Wrangling” datasets into a form suitable for subsequent modeling and analysis*.

*NOTE: Please see the datascience.csuchico.edu website for up-to-date information regarding future Workshops, which will include sessions on topics such as: modeling, analysis, statistical methods, and the other topics relevant to Data Science, Business Analytics, and Research.

Datasets in such forms are often referred to as *“Tidy Data.”*

Develop/Enhance skills thru a Case Project.

HANDS-ON WORKSHOP—Presumptions

Ideally:

1. Generally speaking, you are technically-savvy (especially regarding paths, filenames, and navigating directories).
2. You are aware that—currently—the world’s leading Data Analytical & Computation Language is R…and the most widely used R-oriented IDE is RStudio.
3. Ideally, with respect to R, you know what the “<-“ operator means and how to access a data frame.
4. Ideally you have some familiarity with R and RStudio…or know that the person next do you does!
5. You have a relatively powerful laptop with the latest versions of R and RStudio loaded.
6. You have successfully logged into an available campus WiFi network!!
7. Also, you have loaded the R Package ‘tidyverse’.

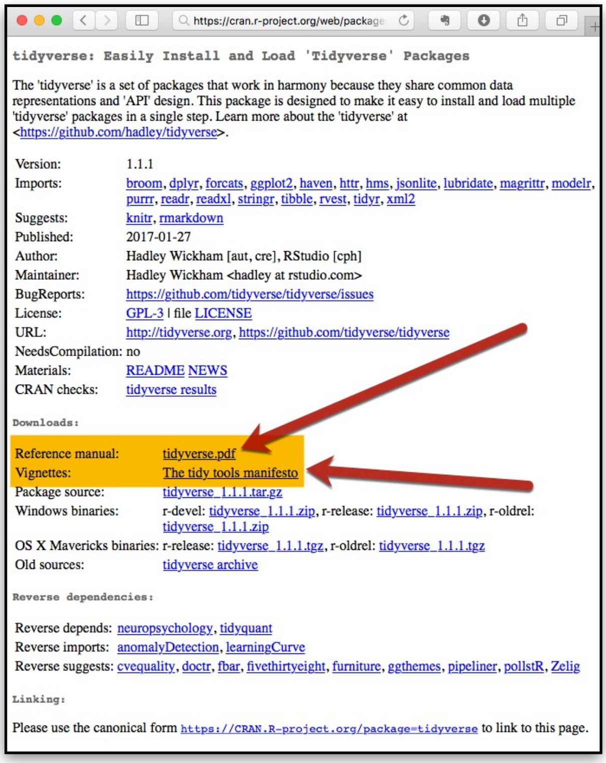
AGENDA

1. Purpose
2. Presumptions
3. Resources
4. An Ideal for Tidying Data
5. A Data Tidying Workflow
6. Activities
    6.1. Activation
    6.2. Workshop Scenario
    6.3. What's the Problem?
    6.4. Access MS-Excel
    6.5. Load Dataset
    6.6. How Untidy?
    6.7.Ideal Shape
    6.8. Fix Lexical Problems
    6.9. Fix Semantic Problems
    6.10. Make Relational

Disclaimer This is a lot to cover and it's unlikely we'll get through it all.

IMPORTANT NOTE: Contrary to the data-centric view of Data Scientists, your DSI-Committee is outright guessing as to appropriate levels-of-detail for these Workshops.

PLEASE LET US KNOW!
…too high-level?
…too detailed?
…Goldilocks?



tidyverse: Easily Install and Load 'Tidyverse' Packages

The 'tidyverse' is a set of packages that work in harmony because they share common data representations and 'API' design. This package is designed to make it easy to install and load multiple 'tidyverse' packages in a single step. Learn more about the 'tidyverse' at <https://github.com/hadley/tidyverse>.

Version:	1.1.1
Imports:	broom, dplyr, forcats, ggplot2, haven, httr, hms, jsonlite, lubridate, magrittr, modelr, purrr, readr, readxl, stringr, tibble, rvest, tidyr, xml2
Suggests:	knitr, rmarkdown
Published:	2017-01-27
Author:	Hadley Wickham [aut, cre], RStudio [cph]
Maintainer:	Hadley Wickham <hadley at rstudio.com>
BugReports:	https://github.com/tidyverse/tidyverse/issues
License:	GPL-3 | file LICENSE
URL:	http://tidyverse.org, https://github.com/tidyverse/tidyverse
NeedsCompilation:	no
Materials:	README NEWS
CRAN checks:	tidyverse results

Downloads:

Reference manual:	tidyverse.pdf
Vignettes:	The tidy tools manifesto
Package source:	tidyverse_1.1.1.tar.gz
Windows binaries:	r-devel: tidyverse_1.1.1.zip, r-release: tidyverse_1.1.1.zip, r-oldrel: tidyverse_1.1.1.zip
OS X Mavericks binaries:	r-release: tidyverse_1.1.1.tgz, r-oldrel: tidyverse_1.1.1.tgz
Old sources:	tidyverse archive

Reverse dependencies:

Reverse depends:	neuropsychology, tidyquant
Reverse imports:	anomalyDetection, learningCurve
Reverse suggests:	cvequality, doctr, fbar, fivethirtyeight, furniture, ggthemes, pipeliner, pollstR, Zelig

Linking:

Please use the canonical form https://CRAN.R-project.org/package=tidyverse to link to this page.

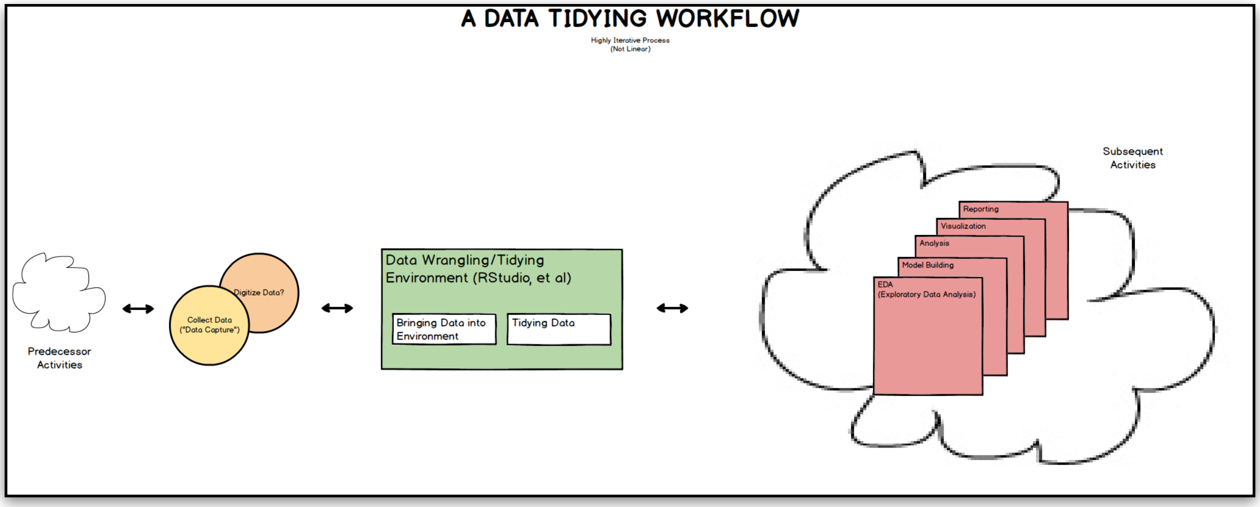
RESOURCES

Tidyverse on CRAN: https://cran.r-project.org/package=tidyverse (see, screenshot to the right).

Cheatsheets from RStudio

- **Data Importing**: https://raw.githubusercontent.com/rstudio/cheatsheets/master/source/pdfs/data-import-cheatsheet.pdf
- **Data Transformation**: https://raw.githubusercontent.com/rstudio/cheatsheets/master/source/pdfs/data-transformation-cheatsheet.pdf
- **RStudio IDE**: https://www.rstudio.com/wp-content/uploads/2016/01/rstudio-IDE-cheatsheet.pdf
- **RegEx**: https://www.rstudio.com/wp-content/uploads/2016/09/RegExCheatsheet.pdf

An IDEAL for TIDYING DATA



1. Maximum Information in a Minimum of Space & Time.
2. Every type of observation has its own table. This implies use of a "Relational Data Model" (and "Data Normalization").
3. [Very Good Practice] Every observation is uniquely, and unambiguously, identified. Unique identification also extends to "reference data" (*e.g.*, descriptive details of categorical/factor variables, descriptive details of subjects, and the like).
4. Every attribute/variable occupies a unique column within a unique table. This means, all values are "atomic." (NOTE: This does not imply all attributes are exclusive to a specific table…which would render use of relational approaches impossible.)
5. Every observation occupies one, and only one, row in a unique table.
6. It is possible for "Units-of-Analysis" to span multiple tables (further evidence of the use of a Relational Data Model).
7. Units-of-Analysis should (must!?) reflect applicable semantics of the subject domain.

8. "Shape" of "Units-of-Analysis" should (must!?) enables subsequent modeling, analysis, reporting and visualization.
9. In Tidy Data code: **Avoid virtuosity*!** Adopt (so-called) "Literate Programming," wherein there are many small—readily readable—steps (instructions).

*A euphemism for *"…writing unreadable code."*

A DATA TIDYING WORKFLOW

(Although depicted linearly for readability, please be mindful that the underlying processes are extremely iterative.)

ACTIVITY 1. ACTIVATION

1. Boot laptop
2. Grab a cookie
3. Create a folder work this Workshop`{root}/DSI-Tidy-Data`"
4. Download:
    4.1. This *Workshop Participant Guide*
    4.2. Workshop Case Dataset (`DSI-Tidy-Data-Case-v20170407.xlsx`) into `{root}/DSI-Tidy-Data`
    4.3. Tidyverse Manual PDF (CRAN)
    4.4. Tidyverse Vignetts (CRAN)
    4.5. Four Cheatsheets (RStudio) (see Resources)
5. Invoke RStudio

ACTIVITY 2: WORKSHOP SCENARIO

To equip today's participants with a common array of Data Tidying skills… today's Hands-on Workshop will be based on a common Case Project; specifically a scenario wherein you are the Principal Investigator of an inquiry regarding the relationship of gravity experiments* and dietary practices of the first true scientists:

Cats

Why cats?

Because, what Data Scientists don't like cats?

***Double-Secret Bonus Points:** *Who is familiar with Feline Gravity Experiments?*

ACTIVITY 3: What's the Problem?

Before data can be modeled, analyzed or interpreted...
...data must be accessible in meaning-laden forms.

Annually, many Analysts around the globe report expending up to 80% of the effort and duration of an investigation in "wrangling" and "tidying" data into a "shape" that enables thoughtful, economical analysis.

ACTIVITY 4. ACCESS MS-EXCEL

As needed,

```
install.packages("xlsx")
library(xlsx)
```

Be mindful of directory! Replace `{root}` with the root for your computer:

```
print(getwd())
file_path <- "{root}/DSI-Tidy-Data/DSI-Tidy-Data-Case-v20170407.xlsx"
print(file_path)
```

ACTIVITY 5. LOAD DATASET

```
dt.cats <- read.xlsx(file_path, sheetName = 'Results')
```

How to observe loaded dataset?

ACTIVITY 6. DETERMINE—HOW UNTIDY?

To illustrate various problems and solutions the [contrived] data in our Workshop Scenario were collected with little consideration for how the data would be consumed, modeled or analyzed.

Examine loaded data…what problems—of a type that would probably impede modeling and analysis—do you detect?

In the allotted time, please list all problems you observer in this Google Doc:

https://docs.google.com/a/mail.csuchico.edu/document/d/16ODvzDGmT9DmsAERBwKbsyoncjkOwwxbCC24IHVttLQ/edit?usp=sharing

ACTIVITY 7. IDEAL "SHAPE"?

With respect to the Case Study investigation, please take a moment to determine what you believe to be the Ideal Shape for the dataset.

Your solution may (probably will) vary from everyone else's (including the Workshop leader).

If you want to record your concept of the "Tidy Data Shape" in the Google Doc, please feel free to do so.

ACTIVITY 8. DATA WRANGLING ONE: Fix Lexical Problems

Lexical (*e.g.*, numbers, dates, unit)

Convert oz to grams

```
dt.cats$Moist_grams <- (dt.cats$Moist_oz) * 28.4
```

Avoid storing derived data (*e.g.*, number of days since "X")

ACTIVITY 9. DATA WRANGLING TWO: Fix Semantic Errors

ACTIVITY 10. DATA WRANGLING THREE: Make Relational