"Tame" data principles and the fivethirtyeight R package

Albert Y. Kim - Amherst College -> Smith College (July 2018)

Tuesday June 12, 2018

Today's focus

What data to use in introductory statistics and data science courses?

Ideally data that's:

1. **Rich** enough to answer meaningful questions with
2. **Real** enough to ensure that there is context
3. **Realistic** enough to convey to the reality of much of the world's data

One goal

On the one hand, Cobb (2015) argues that we should

1. "Teach through research"
2. **"Minimize prerequisites to research"**

Another goal

On the other hand, from New York Times:

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

Monica Rogati, Jawbone's vice president for data science, with Brian Wilt, a senior data scientist.
Peter DaSilva for The New York Times

By Steve Lohr

Aug. 17, 2014

Analogy for second goal

sandy griffith
@sgrifter

love @JennyBryan's analogy of classroom data as teddybears & real data like a grizzly bear with salmon blood dripping out its mouth #jsm2015

1:10 PM - Aug 11, 2015

28 See sandy griffith's other Tweets

Two conflicting goals

- On the one hand: Minimize prerequisites to research
- On the other: Do not betray reality of data as it exists in much of the world

Back to analogy

In other words, a balancing act is required between:

Data with no prerequisites needed

Data as it exists "in the wild"

Data "taming"

Data "taming" sets out to balance:

- On the one hand: Performing enough pre-processing so that data is accessible to R novices
- On the other: Not performing so much pre-processing as to betray the reality of data as it exists "in the wild"

"Tame" data principles

We propose the following "tame" data principles to remove biggest hurdles R novices face:

1. Clean variable names
2. Identification variables in left-hand columns
3. Clean dates
4. Logically ordered categorical variables
5. Consistent "tidy" format

fivethirtyeight package

In the `fivethirtyeight` R package, Chester Ismay, Jennifer Chunn, and I:

- Take FiveThirtyEight's raw article data from GitHub
- **Pre-process the raw data so that it follows "tame" data principles**
- Make the tame data, documentation, and original article easily accessible via an R package

Examples

Following examples involve code, so I suggest you follow in HTML version of slides:

1. In your browser, go to bit.ly/causeweb_tame
2. In the left-hand menu, click on "Principle 1: Clean variable names"

Principle 1: Clean variable names

a) Comparing raw and tamed data

- Original article: 41 Percent Of Fliers Think You're Rude If You Recline Your Seat
- Raw CSV data: `flying-etiquette.csv`

```
library(readr)
library(fivethirtyeight)

# Raw data: variable names are unwieldy & have spaces
flying_raw <- read_csv("https://raw.githubusercontent.com/fivethirtyeight/data/master/flyin
colnames(flying_raw)[c(5, 19)]
```

```
## [1] "Do you have any children under 18?"
## [2] "In general, is itrude to bring a baby on a plane?"
```

```
# Tamed data: corresponding variable names are cleaner
colnames(flying)[c(5, 18)]
```

```
## [1] "children_under_18" "baby"
```

Principle 2: ID variables

More organizational. Any identification variables that uniquely identify the observations/rows should be place in the left-hand columns since they are of highest prominence. Such variables are used to key joins/merging of datasets.

- Original articles:
  1. Straight Outta Compton' Is The Rare Biopic Not About White Dudes
  2. A Statistical Analysis of the Work of Bob Ross
- Raw CSV data:
  1. `biopics.csv`
  2. `elements-by-episode.csv`

```
library(fivethirtyeight)

# Both title and imdb site tag uniquely identify movies. Show only 8 first
# columns and 3 first rows of dataset:
biopics[1:3, 1:8]
```

```
## # A tibble: 3 x 8
##   title   site   country year_release box_office director number_of_subje…
```

Principle 3: Dates

a) Comparing raw and tamed data

- Original article: Some People Are Too Superstitious To Have A Baby On Friday The 13th
- Raw CSV data: `US_births_1994-2003_CDC_NCHS.csv`

```
library(readr)
library(dplyr)
library(fivethirtyeight)

# Raw data: year, month, day are separate variables
US_births_1994_2003_raw <- read_csv("https://raw.githubusercontent.com/fivethirtyeight/data
head(US_births_1994_2003_raw)
```

```
## # A tibble: 6 x 5
##    year month date_of_month day_of_week births
##   <int> <int>         <int>       <int>  <int>
## 1  1994     1             1           6   8096
## 2  1994     1             2           7   7772
## 3  1994     1             3           1  10142
## 4  1994     1             4           2  11248
```

Principle 4: Categorical variables

a) Comparing raw and tamed data

- Original article: The Dollar-And-Cents Case Against Hollywood's Exclusion of Women
- Raw CSV data: `movies.csv`

```
library(readr)
library(ggplot2)
library(fivethirtyeight)
bechdel_raw <- read_csv("https://raw.githubusercontent.com/rudeboybert/fivethirtyeight/mast

# Raw data: categorical variable clean_test is saved as characters/strings
bechdel_raw$clean_test[1:5]
```

```
## [1] "notalk" "ok"     "notalk" "notalk" "men"
```

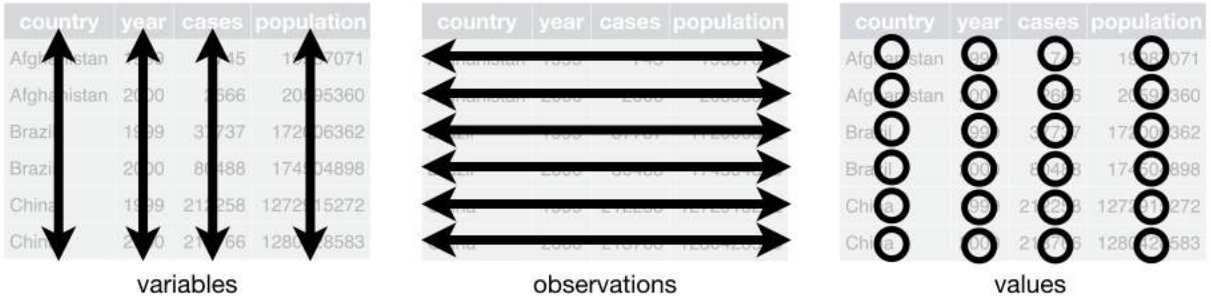
```
# Tamed data: clean_test is saved as factor
bechdel$clean_test[1:5]
```

```
## [1] notalk ok     notalk notalk men
## Levels: nowomen < notalk < men < dubious < ok
```

Principle 5: "Tidy" data format

"Tidy" data format is narrow/long format, as opposed to wide. This format is chosen for input/output data frame standardization across many R packages in the tidyverse: `ggplot2`, `dplyr`, etc. 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.



a) Comparing raw and tamed data

- Original article: Dear Mona Followup: Where Do People Drink The Most Beer, Wine And Spirits?

Advanced example

a) Comparing raw and tamed data

- Original article: The Last 10 Weeks Of 2016 Campaign Stops In One Handy Gif
- Raw CSV data were in two separate CSVs
  - `clinton.csv`
  - `trump.csv`

In the tamed `pres_2016_trail` data frame we:

1. Ensured `lat` and `lng` were in numerical format, not in degree/minute/second, North/South, and East/West format (A variation on Principle 3: Dates)
2. Combined both CSV's into one and added variable `candidate` (Principle 5: Tidy data format)

```
library(dplyr)
library(fivethirtyeight)

# Tamed data:
```

Comments

- Analogy I heard that I like: `fivethirtyeight` is like a data petting zoo
- No "universal" balance of two goals: it will vary depending on your students' experience, requirements, and needs
- Tame data principles and `fivethirtyeight` can be used in other contexts: 1) intermediate-level data science courses and 2) advanced projects

Used in data science courses

1. Recruited STAT231 Data Science students to "tame" datasets STAT135 Intro students found for their final projects
2. Available on GitHub: data wrangling source code by package authors to convert 538 raw CSV data to "tamed" format
   `process_data_sets_albert.R`, `process_data_sets_chester.R`, `process_data_sets_jen.R`

Used for advanced projects

- `fivethirtyeight` package is in maintenance mode: no new development, only need to add new datasets
- Get student interns to do it instead!
- Internship model of learning/development: learning R package construction, GitHub, communication and project management skills, etc. RStudio's 2018 `broom` package summer internship follows a similar model.
- **Undergraduate student** written data wrangling source code to convert 538 raw CSV data to "tamed" format `process_data_sets_maggie.R`, `process_data_sets_meredith.R`

Other resources

- Complete TISE article (HTML, PDF)
- Package homepage including list of all datasets
- Link to this presentation bit.ly/causeweb_tame