UCLA
Technology Innovations in Statistics Education

Title

The fivethirtyeight R Package: "Tame Data" Principles for Introductory Statistics and Data Science Courses

Permalink

Journal

Authors

Kim, Albert Y.
Ismay, Chester
Chunn, Jennifer

Publication Date

2018

DOI

Supplemental Material

Copyright Information

Peer reviewed

The `fivethirtyeight` R Package: "Tame Data" Principles for Introductory Statistics and Data Science Courses

Introduction

At the time of release of the `fivethirtyeight` package, the authors were all teaching in undergraduate settings, in particular introductory statistics, data science, quantitative methods, and business analytics courses. These courses were taught in a wide variety of settings for students coming from a diverse array of backgrounds, with each student having their own set of prior experience, requirements, and needs (see Biographies). Despite these variations, we agreed that there is a basic set of pedagogical principles that any introductory data-centric course should follow for it to truly serve the 21st century student. In particular, we agreed that such courses should revolve around the use of data that satisfy what we term the *3 R's*. We want to use data that is

1. **Rich enough** to answer meaningful questions with;
2. **Real enough** to ensure that there is context;
3. **Realistic enough** to convey to students that data as it exists "in the wild" often needs wrangling/pre-processing;

It is against this backdrop that we present `fivethirtyeight`: an R package of data and code behind the stories and interactives at FiveThirtyEight. FiveThirtyEight.com is a data-driven journalism website founded by Nate Silver and owned by Disney/ESPN that reports on politics, economics, sports, and other current events. FiveThirtyEight has been forward-thinking in making the data used in many of their articles open and accessible on their GitHub repository page https://github.com/fivethirtyeight/data (GitHub is a web-based repository for collaboration on code and data). Our `fivethirtyeight` R package goes one step further by making this data and its corresponding documentation easily accessible, all with students in introductory statistics and data science courses that use R in mind. The homepage for the `fivethirtyeight` R package can be found at https://fivethirtyeight-r.netlify.com/.

Educational Landscape

The era of "Big Data" has led to renewed interest by the general public in the discipline of statistics. Given the increase in prominence of data-driven decision making in academia, industry, and government, there has been a corresponding increase in the number of statistics majors (American Statistical Association Undergraduate Guidelines Workgroup 2014). In parallel with this growth, college administrators are pushing many new data science/data analytics education initiatives; this demand for data-centric education is forecasted to continue

well into the future (Manyika et al. 2011). Correspondingly, there has also been a rise in the availability and accessibility of laptop computers, open source statistical and computational tools, open source learning materials, and open data licenses. This has led to students being increasingly freed from the physical constraint of computer labs centered around desktop machines using proprietary commercial software, allowing them to work in physical settings and environments that they are more comfortable in. Additionally, given the open source nature of many new pedagogical tools, instruction manuals are increasingly being replaced with crowd-sourced approaches to help and documentation, such as GitHub and StackOverflow, opening up a wide world of support.

Statistics educators must capitalize on this renewed interest in the field, the rise of data science/analytics, and the increased availability and accessibility of technological tools to convey the importance of many critical statistical concepts to a new generation of students and practitioners, including distributions, sampling error, uncertainty, and modeling (Cobb 2015; DeVeaux et al. 2016; Nolan and Lang 2010). In particular, one opportunity for experimentation with new pedagogical approaches, as espoused by the Guidelines for Assessment and Instruction in Statistics Education (GAISE), is to "integrate real data with a context and purpose" into the classroom (GAISE College Report ASA Revision Committee 2016). Such integration of real datasets in the classroom has long had many advocates (Gould 2010).

Two Conflicting Goals

One overarching goal we shared when designing courses is Cobb's principle of "minimizing prerequisites to research" for students, where "research" in this context does not necessarily imply research suitable for publication in academic journals, but rather simply "answering questions with data" (Cobb 2015). Cobb argues that students should be exposed to real and substantive data analysis as quickly as possible, even if this entails simplifying certain important topics at first and pushing aside others until later in the curriculum. Otherwise, Cobb argues we risk alienating a large number of students early on, thereby never allowing them to be exposed to the great potential of data analysis to answer questions on topics they may be curious about. This concept is further illustrated in what Grolemund and Wickham refer to as the "data/science pipeline" in Figure 1 (Grolemund and Wickham 2017). We feel that it is important to expose students to all these elements of this statistical/data science analysis cycle, from data importation right through to the communication of the results, and not just certain portions of the pipeline as is often the case in many introductory courses.

This thinking is in line with Perkins' philosophy of "making learning more meaningful by teaching the 'whole game' " instead of just isolated pieces of a discipline (Perkins 2010). So just as how aspiring softball/baseball players first learn to play a simplified version of the game called "teeball" instead of only learning how to swing a bat, we feel that introductory statistics and data science students should be exposed to the entirety of the data/science pipeline early, even if only in a simplified form.

Another overarching goal we shared when designing courses is emphasizing the use of "real world" data. Dr. Jenny Bryan at the University of British Columbia and RStudio has famously
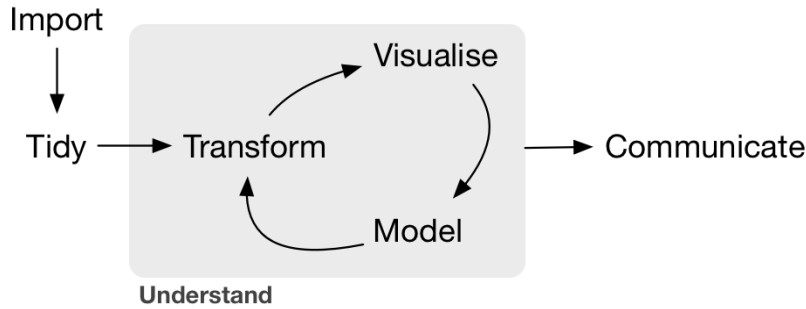
Figure 1: Grolemund and Wickham's "Data/Science Pipeline"

quoted that traditional data used in the classroom is like a "teddy bear," whereas real world data is more like a "grizzly bear with salmon blood dripping from its mouth" (Jedamski 2015). To shield students from the wrangling and pre-processing necessary to prepare data for analysis is to ultimately do students a disservice, as we would be blinding them to the reality that many statisticians and data scientists spend a substantial amount of their time performing such tasks (Lohr 2014).

These two goals, to 'minimize prerequisites to research' while also using real-world data, however, are in direct conflict with each other. On the one hand, one cannot expect novices to immediately tackle many raw datasets as they exist "in the wild," thereby placing barriers to research at a time where there is high risk of alienating them. On the other hand, to present students with datasets that are overly curated would betray the true nature of the work done by statisticians and data scientists. We argue instead that a careful and thoughtful balance between these two goals is essential when preparing datasets for use in introductory statistics and data science courses.

As a means to this end we created the `fivethirtyeight` package, whose features and basic uses we present in Section 2. In Section 3 we present discussions on how we set the balance between these two goals and how this balance translated to the ultimate design of the package. These ideas are embodied in our proposed: *"Tame data" principles for introductory statistics and data science courses.*

The `fivethirtyeight` R Package

Many canonical datasets, such as `iris`, `UCBAdmissions`, Old `faithful`, and `anscombe`'s quartet have long been available for use in R. Other datasets, such as `mtcars` and `diamonds`, are used by many R packages as common examples to demonstrate the package's functionality. More interestingly, many newer and larger datasets are included in R packages devoted exclusively to their dissemination, such as `nycflights13`, `babynames`, `gapminder`, and `okcupiddata` (Bryan 2015; Horton et al. 2015; Kim and Escobedo-Land 2015; Wickham 2017a). Disseminating data via R packages has numerous advantages, in particular standardized web-based installation and relatively simple data-loading procedures.

Building off this precedent, we created the `fivethirtyeight` R package of data and code behind the stories and interactive visualizations at FiveThirtyEight.com. All with the goal of providing the novice student a "low barrier to entry" to exploring datasets that satisfy the 3 R's, we

1. **Made the data easy to load**: All data is loaded from an R package, instead of through the importing of comma separated values (CSV) files, Excel spreadsheets, and other file formats.
2. **Made the data easy to understand**: We provide ample documentation in the help file for each dataset, including:
   1. A thorough description of the observational unit and all variables.
   2. If documented, the data sources.
   3. A link to the original article on FiveThirtyEight. We felt that inclusion of these links was critical to provide information on the data's context, thereby emphasizing how "real" the data are.
3. **Made the data easy to explore**: We pre-process the data "just enough" to balance between the two conflicting goals from Section 1.2, thereby providing students access to data that is both "rich" and "realistic", but not so much so that we run counter to the goal of "minimizing prerequisites to research." We explicitly define what is meant by "just enough" in the Section 3: "Tame Data" Principles for Introductory Statistics and Data Science Courses.
4. **Made example data analyses easy to access and reproduce**: We include crowd-sourced data analysis examples in *vignettes* written in R Markdown, available at https://fivethirtyeight-r.netlify.com/articles/.

Basic Usage

We demonstrate the most basic use of the package: viewing the contents of a dataset and opening its help file. As an example, we view the `bechdel` dataset and its corresponding help file, which contains a link to the original FiveThirtyEight article "The Dollar-And-Cents Case Against Hollywood's Exclusion of Women" (Hickey 2014a), a codebook describing all 15 variables, and other information.

```r
library(fivethirtyeight)
head(bechdel)
?bechdel
# If using RStudio:
View(bechdel)
```

The current version of the `fivethirtyeight` package (v0.3.0.9000) contains 75 datasets, a detailed list of which can be viewed by running:

```r
vignette("fivethirtyeight", package="fivethirtyeight")
```

Example Usage

We now present a more thorough example usage of the package, suitable for assignment in an introductory statistics or data science course, centering around the `hate_crimes` dataset used in the FiveThirtyEight article "Higher Rates Of Hate Crimes Are Tied To Income Inequality." (Majumder 2017) In this article, the authors were interested in explaining the occurrence of hate crimes in the 50 US states and the District of Columbia based on demographic features of those states such as median income, education, and income inequality. In particular, they used multiple regression using both

1. `hate_crimes_per_100k_splc`: the number of hate crimes per 100,000 individuals between November 9-18, 2016 as reported by the Southern Poverty Law Center
2. `avg_hatecrimes_per_100k_fbi`: the average annual hate crime rate per 100,000 individuals between 2010 and 2015 as reported by the FBI

as outcome variables. An example exercise would be to have students:

1. Read the original FiveThirtyEight article and any other necessary supplementary materials for context.
2. Look over the help file for the dataset to know what variables they have access to.
3. Perform an exploratory data analysis of the `hate_crimes` data by visually inspecting the raw values, computing summary statistics, and creating visualizations.
4. Run regressions with either of the above outcome variables to investigate which state-level demographic features are associated with hate crimes.
5. Summarize their findings in written form.

We provide example solution code for this assignment using `hate_crimes_per_100k_splc` as the outcome variable and `gini_index` as the explanatory variable. The Gini index is a measure of income inequality where a value of 0 indicates perfect income equality and a value of 1 indicates perfect income inequality

First, have the students read the `hate_crimes` help file and look at the dataset's raw values (note here the output is suppressed):

```r
library(fivethirtyeight)
?hate_crimes
head(hate_crimes)
# If using RStudio:
View(hate_crimes)
```

then have students return summary statistics about both the outcome and explanatory variables:

```r
summary(hate_crimes$gini_index)
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>  0.4190  0.4400  0.4540  0.4538  0.4665  0.5320
summary(hate_crimes$hate_crimes_per_100k_splc)
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
```
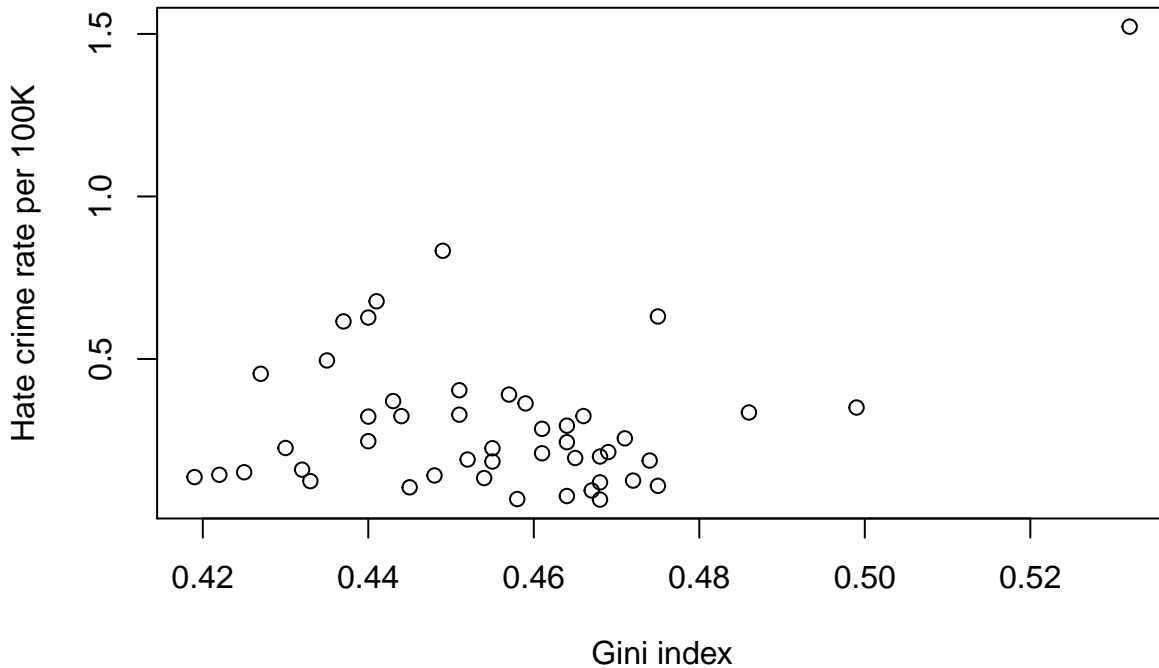
Figure 2: Relationship between hate crime incidence and income inequality.

```
#> 0.06745 0.14271 0.22620 0.30409 0.35694 1.52230        4
```

then have students produce a scatterplot visualizing the relationship between these two variables (see Figure 2):

```
plot(hate_crimes$gini_index, hate_crimes$hate_crimes_per_100k_splc,
    xlab = "Gini index", ylab = "Hate crime rate per 100K",
    title = "Hate Crimes per 100K Nov 9-18 2016 (SPLC)")
```

then have students perform the corresponding simple linear regression:

```
lm(hate_crimes_per_100k_splc ~ gini_index, data = hate_crimes)
```

which leads to a summary of results similar to the following:

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-1.527	0.783	-1.950	0.057	-3.105	0.05
gini_index	4.021	1.718	2.341	0.024	0.561	7.48

During the analysis of this particular dataset, many interesting pedagogical questions can be posed to students, including:

6

1. What is the Gini index? Students might not be familiar with this measure of income inequality, necessitating outside investigations (on Wikipedia for example).
2. What mechanisms generated the outcome variable? For example, how did the SPLC compute their counts as compared to the FBI? Are these counts based on *actual* numbers of hate crimes or *reported* numbers of hate crimes?
3. There were four states that did not have hate crimes measures (Hawaii, North Dakota, South Dakota, and Wyoming) yielding four `NA` missing values. What is the nature of this missingness? Is there something particular about these four states?
4. The plot indicates one clear outlier in the relationship between hate crimes and income inequality. Upon further investigating this outlier can be identified as the District of Columbia (DC). What is different about DC when compared to the 50 states? Are their populations different? Does DC's extremely small geographical area and urban nature answer this question?
5. The regression table suggests that there there is a positive association between hate crimes and income inequality. However, how much is this is driven by DC's outlier value? Furthermore, what does the traditional interpretation of a slope coefficient b_1 (for every increase of 1 unit in x, there is an associated increase of on average b_1 units of y) mean in this case when the Gini index itself is a value between 0 and 1? What might be a more informative interpretation of the slope coefficient?

Vignettes

The other component of the `fivethirtyeight` package is our use of package vignettes. Vignettes are included in many R packages as long-form guides and extended documentation to supplement existing help files (Wickham 2015). In this package, we use vignettes not for documentation purposes, but as example data analyses based on the datasets included in the package. They are written in R Markdown, a notebook-style file format that weaves together narrative text and code to produce elegantly formatted output (Allaire et al. 2017). Since the written text and source code are included in the same document, users can easily reproduce the analysis in its entirety.

In the article "The Dollar-And-Cents Case Against Hollywood's Exclusion of Women" (Hickey 2014a) corresponding to the earlier introduced `bechdel` dataset, the authors describe a rough metric for evaluating the representation of women in a particular movie. A movie is described as passing the "Bechdel Test" if it satisfies the following criteria:

1. There are at least two named female protagonists.
2. They talk to each.
3. They talk to each other about something other than a male protagonist.

As described in the `bechdel` help file, the variable `title` identifies the movie while the variable `clean_test` displays the Bechdel Test rating, a categorical variable with five levels: `ok` (passes test), `dubious`, `men` (female protagonists only talk about male protagonists), `no talk` (female protagonists don't talk to each other), and `nowomen` (fewer than two female protagonists). The vignette associated with this `bechdel` dataset is accessible by running:

```r
vignette("bechdel", package="fivethirtyeight")
```

A list of all user contributed vignettes, which are hosted on the package homepage at https://fivethirtyeight-r.netlify.com/articles/, can viewed by running:

```r
vignette("user_contributed_vignettes", package="fivethirtyeight")
```

Note however due to R package size restrictions, the version of the `fivethirtyeight` package on CRAN which is installed via the command `install.packages("fivethirtyeight")` only includes the vignette corresponding to the `bechdel` dataset. In order install all vignettes and their `.Rmd` R Markdown source code locally, the development version of the `fivethirtyeight` package must be installed by running:

```r
# If you haven't installed remotes package yet, do so via:
# install.packages("remotes")
remotes::install_github("rudeboybert/fivethirtyeight", build_vignettes = TRUE)
```

Package Source Code

As an additional resource for instructors, all the source code used to pre-process the raw data available on the FiveThirtyEight GitHub repository page at https://github.com/fivethirtyeight/data to for inclusion in the `fivethirtyeight` package is accessible on the package's GitHub repository page. This code can be used as data wrangling examples for more advanced students wanting to learn more advanced data wrangling topics such as strings/text data and dates:

- The raw data files can also be found in https://github.com/rudeboybert/fivethirtyeight/tree/master/data-raw.
- In the same directory, all `.R` files prefixed with `process_data_sets_` contain all the source code to pre-process the raw data. At the time of publication of this article, there were three such files (`process_data_sets_albert.R`, `process_data_sets_chester.R`, and `process_data_sets_jen.R`); in the future we anticipate having more.

"Tame Data" Principles for Introductory Statistics and Data Science Courses

In the spirit of balancing between the conflicting goals stated in Section 1.2, we propose the following *"Tame data" principles for introductory statistics and data science courses*. They are an explicit articulation of the guidelines we followed during the design of the `fivethirtyeight` R package, in particular the pre-processing we performed on the raw data available on the FiveThirtyEight GitHub repository page. The implementation of these principles removes some of the biggest barriers faced by novice students when first constructing basic data visualizations and performing rudimentary data wrangling, thereby

"minimizing prerequisites for research." However, their implementation does not simplify the datasets to the point that they no longer present students with a realistic view of data as it exists "in the wild."

1. **Naming conventions for data frame and variable names**:
   1. Whenever possible, all names should be no more than 20 characters long. Exceptions to this rule exist when shortening the names to less than 20 characters would lead to a loss of information.
   2. Use only lower case characters and replace all spaces with underscores. This format is known as `snake_case` and is an alternative to `camelCase`, where successive words are delineated with upper case characters.
   3. In the case of variable (column) names within a data frame, use underscores instead of spaces.
2. **Variables identifying observational units**:
   1. Any variables uniquely identifying each observational unit should be in the left-hand columns.
3. **Dates**:
   1. If only a `year` variable exists, then it should be represented as a numerical variable.
   2. If there are `year` and `month` variables, then convert them to `Date` objects as `year-month-01`. In other words, associate all observations from the same month to have a `day` of `01` so that a correct `Date` object can be assigned.
   3. If there are `year`, `month`, and `day` variables, then convert them to `Date` objects as `year-month-day`.
4. **Ordered Factors, Factors, Characters, and Logicals**:
   1. Ordinal categorical variables are represented as `ordered` factors.
   2. Categorical variables with a fixed and known set of levels are represented as regular `factor`s.
   3. Categorical variables whose possible levels are either unknown or of a very large number are represented as `character`s.
   4. Any "yes/no" character encoding of binary variables is converted to `TRUE/FALSE` logical variables.
5. **Tidy data format**:
   1. Whenever possible, save all data frames in "tidy" data format as defined by Wickham (2014):
      a) Each variable forms a column.
      b) Each observation forms a row.
      c) Each type of observational unit forms a table.
   2. If converting the raw data to "tidy" data format alters the dataset too much, then make the code to convert to tidy format easily accessible.

By advocating for the shielding of novices from such pre-processing, we are not arguing for their diminished importance in a data scientist/statistician's toolkit. Rather we argue that these are intermediate topics that should be left to later in the curriculum, either later in the same course or in a later course altogether. We have found in our experience that students are better motivated to learn such intermediate topics only after they have had a chance to develop basic data intuition and perform elementary data wrangling, visualization,

modeling, and analysis. Furthermore, we note that while some of these principles are statistical language/software independent, many are specific to R and thus they should be viewed in the context of a course syllabus centered around the use of R. We present examples of four of these five principles in application and demonstrate the advantages they yield for novice R users, using both base R and `tidyverse`-centric approaches (Wickham 2017b).

Naming Conventions

Whereas the first three subpoints relating to data frame and variable naming are more cosmetic in nature, the fourth principle relates to an R-specific issue faced by novices. Variable names that include spaces require different treatment than those that do not, specifically the use of tick marks when referring to them. While this is a topic R users eventually have to learn, we argue that it is not an immediate priority.

As an example of the importance of preprocessing variable names, consider the data corresponding to the FiveThirtyEight article "41 Percent Of Fliers Think You're Rude If You Recline Your Seat" (Hickey 2014b) where survey respondents were asked, among other things, if 1) they considered it rude to bring a baby on a flight and 2) whether or not they had children under the age of 18. The raw data corresponding to this article is saved in CSV format on FiveThirtyEight's GitHub repository page https://github.com/fivethirtyeight and can be accessed via the shortened bit.ly link http://bit.ly/2vg8gTf. We load and save this in a data frame `flying_raw` and look at the first 5 variable names:

```
library(tidyverse)
flying_raw <- read_csv("http://bit.ly/2vg8gTf")
colnames(flying_raw)[1:5]
#> [1] "RespondentID"
#> [2] "How often do you travel by plane?"
#> [3] "Do you ever recline your seat when you fly?"
#> [4] "How tall are you?"
#> [5] "Do you have any children under 18?"
```

We contrast this to the corresponding `flying` data frame in the `fivethirtyeight` package:

```
library(fivethirtyeight)
colnames(flying)[1:5]
#> [1] "respondent_id"     "gender"              "age"
#> [4] "height"            "children_under_18"
```

We see that in the latter case the variable names are much shorter and cleaner. This simplification leads to much less cumbersome code to wrangle and visualize this data. For example, consider the following two `ggplot()` commands to generate the barplot in Figure 3 to visualize the relationship between the two categorical variables of interest: using the raw data necessitates tick marks to access the variables, whereas using the latter data doesn't.
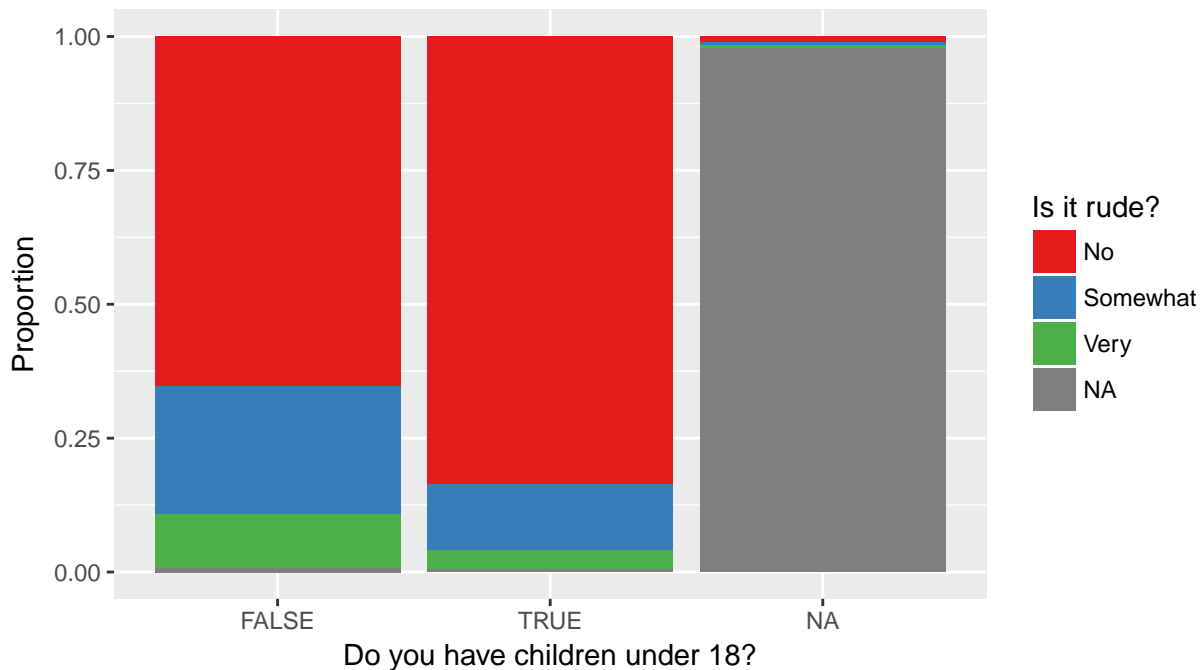
Figure 3: Attitudes about bringing babies on a flight.

```r
# Using raw data:
ggplot(flying_raw,
       aes(x = `Do you have any children under 18?`,
           fill = `In general, is itrude to bring a baby on a plane?`)) +
  geom_bar(position = "fill") +
  labs(x = "Children under 18?", y = "Proportion", fill = "Is it rude?")

# Using fivethirtyeight package data:
ggplot(flying, aes(x = children_under_18, fill = baby)) +
  geom_bar(position = "fill") +
  labs(x = "Children under 18?", y = "Proportion", fill = "Is it rude?")
```

Dates

Although recent advances such as the `lubridate` package have made wrangling dates much easier, performing such tasks can still be very challenging for students who have not had experience (Grolemund and Wickham 2011). In datasets where only a numerical variable indicating the year exists, it can be argued no pre-processing is necessary. However, when a month and/or day variables exist along with a year variable, we argue that pre-processing should be done. Specifically, they should be combined and converted to `Date` objects. This allows for easy creation of time series plots with well formatted x-axes and for performing of basic date arithmetic.

As an example of the importance of preprocessing dates, consider the data corresponding to the FiveThirtyEight article "Some People Are Too Superstitious To Have A Baby On Friday The 13th" (Bialik 2016) of the number of daily births in the United States between 1994 and 2003. The raw data corresponding to this article is saved in CSV format on FiveThirtyEight's GitHub repository page https://github.com/fivethirtyeight/ and can be accessed via the shortened bit.ly link http://bit.ly/2vgRFiw. We load this data, filter for only those rows corresponding to 1999 births, and save this in a data frame `US_births_1999_raw`. The raw data is saved in a format that makes it difficult for novices to create a time series plot. Furthermore, people do not typically think of the day of the week (Sunday, Monday, etc) in terms of a numerical value between 1 and 7.

```
library(tidyverse)
US_births_1999_2003_raw <- read_csv("http://bit.ly/2vgRFiw")
US_births_1999_raw <- US_births_1999_2003_raw[US_births_1999_2003_raw$year == 1999, ]
head(US_births_1999_raw)
#> # A tibble: 6 x 5
#>    year month date_of_month day_of_week births
#>   <int> <int>         <int>       <int>  <int>
#> 1  1999     1             1           5   8163
#> 2  1999     1             2           6   7637
#> 3  1999     1             3           7   7416
#> 4  1999     1             4           1  10396
#> 5  1999     1             5           2  12004
#> 6  1999     1             6           3  11718
```

In contrast, when using the pre-processed `US_births_1994_2003` data frame from the `fivethirtyeight` package we observe that there is a variable `date`, which can be treated as a numerical variable. This allows for the `date` variable to be plotted with informative tick marks on the x-axis as in Figure 4. Furthermore, the day of the week is indicated with more informative text rather that values between 1 and 7.

```
library(fivethirtyeight)
US_births_1999 <- US_births_1994_2003[US_births_1994_2003$year == 1999, ]
head(US_births_1999)
#> # A tibble: 6 x 6
#>    year month date_of_month       date day_of_week births
#>   <int> <int>         <int>     <date>        <ord>  <int>
#> 1  1999     1             1 1999-01-01          Fri   8163
#> 2  1999     1             2 1999-01-02          Sat   7637
#> 3  1999     1             3 1999-01-03          Sun   7416
#> 4  1999     1             4 1999-01-04          Mon  10396
#> 5  1999     1             5 1999-01-05         Tues  12004
#> 6  1999     1             6 1999-01-06          Wed  11718
plot(US_births_1999$date, US_births_1999$births, type = "l",
     xlab = "Date", ylab = "# of births")
```
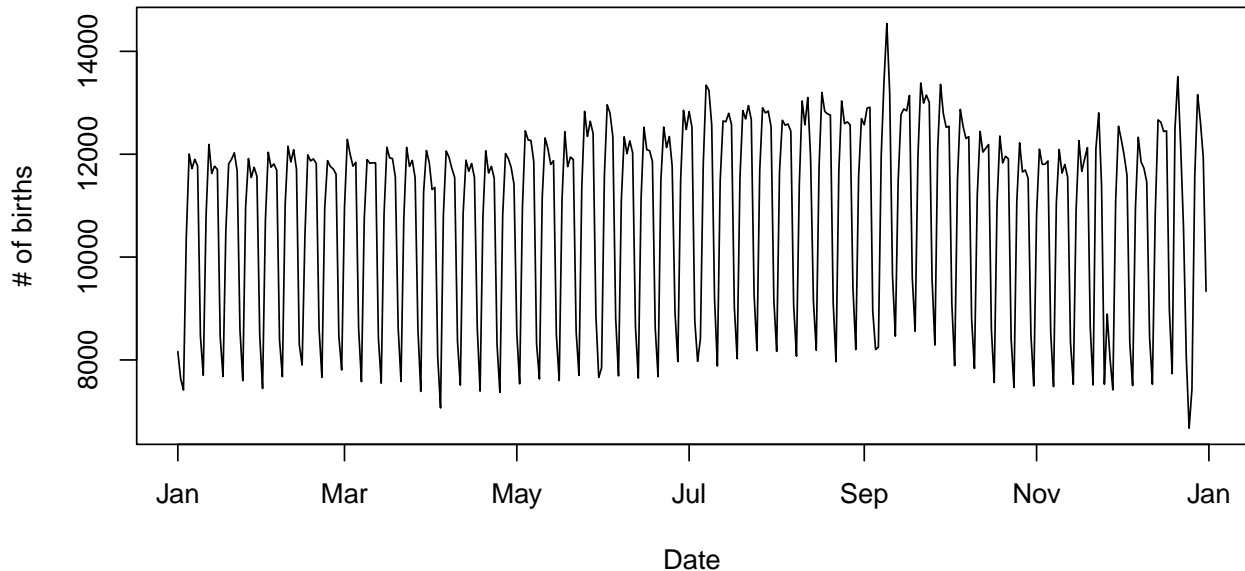
Figure 4: Number of US births in 1999.

One exercise we can assign to students is for them to investigate the anomalous spike in the number of births that occured roughly a month before October 1st, 1999:

```
head(US_births_1999[which.max(US_births_1999$births), ])
#> # A tibble: 1 x 6
#>    year month date_of_month       date day_of_week births
#>   <int> <int>        <int>      <date>       <ord>  <int>
#> 1  1999     9            9 1999-09-09       Thurs  14540
```

Much as the FiveThirtyEight article suggested that there are relatively fewer births on Fridays that fall on the 13th because many parents avoid inducing labor on that date for likely superstitious reasons, a plausible explanation for the spike of births on 9/9/99 is that parents were choosing to deliberately induce labor on that date given its aesthetic appeal.

Factors vs Characters

In our experience, among the earliest questions students pose when creating data visualizations involving a categorical variable (such as barplots and boxplots) is "How do I reorder the bars/boxes?" This is because the default alphabetical ordering of a categorical variable is often not intuitive. While there have been many recent advances in dealing with factor variables in R such as the `forcats` package (McNamara and Horton 2017; Wickham 2017c), we argue that the reordering of categorical variables is still sufficiently too complex for novices that it would run counter to Cobb's philosophy of "minimizing prerequisites to research." Rather, we suggest that this topic should only be addressed after students have developed a basic familiarity with R. Therefore when pre-processing the datasets for this package, we represented ordinal categorical variables as `ordered` factors with the appropriate levels,

categorical variables with a fixed and known set of levels as regular `factor`s, and categorical variables whose possible levels are either unknown or of a very large number as `character`s.

As an example of the importance of preprocessing categorical variables, consider again the data corresponding to the FiveThirtyEight article "The Dollar-And-Cents Case Against Hollywood's Exclusion of Women" (Hickey 2014a) of Bechdel Test scores of gender representation for various movies. This article includes Figure 5, showing the breakdown of Bechdel Test results for movies in 5-year time periods. A key observation is that within each bar corresponding to a 5-year time period, the vertical ordering of the blue and red segments matches the hierarchical nature of the five possible Bechdel Test outcomes. For example, the question of whether female protagonists talk about something other than male protagonists is only relevant if the movie has more than two women and the women speak to each other.
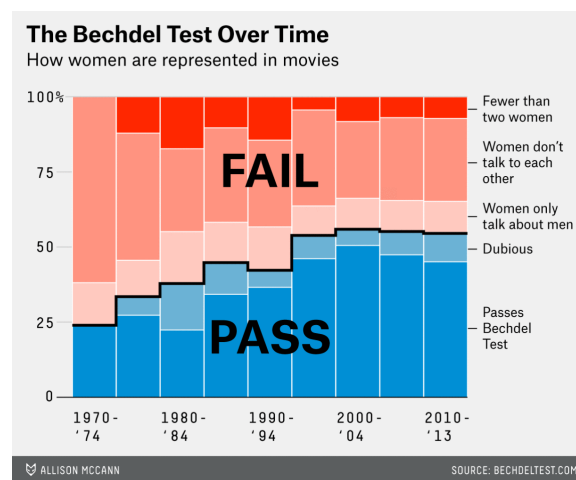

Figure 5: Original Bechdel barplot in FiveThirtyEight Article

We now reconstruct this graphic using the data provided by FiveThirtyEight using both the raw data saved in CSV format on FiveThirtyEight's GitHub repository page https://github.com/fivethirtyeight (accessible via the shortened bit.ly link http://bit.ly/2uD3ls6) and using the pre-processed version in the `fivethirtyeight` R package. In both cases, we discretize the `year` variable into 5-year bins using the vector `year_bins` and plot a stacked barplot of proportions.

```r
year_bins <- c("'70-'74", "'75-'79", "'80-'84", "'85-'89", "'90-'94",
               "'95-'99", "'00-'04", "'05-'09", "'10-'13")

# Using raw data:
library(tidyverse)
bechdel_raw <- read_csv("http://bit.ly/2uD3ls6") %>%
  mutate(five_year = cut(year, breaks = seq(1969, 2014, 5), labels = year_bins))

ggplot(bechdel_raw, aes(x = five_year, fill = clean_test)) +
  geom_bar(position = "fill", color = "black") +
  labs(x = "Year", y = "Proportion", fill = "Bechdel Test") +
```

```
  scale_fill_brewer(palette = "YlGnBu")

# Using fivethirtyeight package data:
library(fivethirtyeight)
bechdel <- bechdel %>%
  mutate(five_year = cut(year, breaks = seq(1969, 2014, 5), labels = year_bins))

ggplot(bechdel, aes(x = five_year, fill = clean_test)) +
  geom_bar(position = "fill", color = "black") +
  labs(x = "Year", y = "Proportion", fill = "Bechdel Test") +
  scale_fill_brewer(palette = "YlGnBu")
```

We observe in the top plot of Figure 6 based on the raw data has the uninformative default alphabetical ordering of the Bechdel Test outcomes, whereas the bottom plot in Figure 6 based on the pre-processed `bechdel` data from the `fivethirtyeight` package has the correct hierarchical ordering of the outcomes.

Tidy Data Format

Whenever possible, we saved data frames in "tidy" data format (Wickham 2014). Wickham describes a dataset/data frame as being in tidy format if it satisfies the following criteria:

a) Each variable must have its own column.
b) Each observation must have its own row.
c) Each type of observational unit forms a table.

This format is also known as "long/narrow" format, as opposed to "wide" format. While the information contained in a dataset is identical irrespective of it being in "tidy" format or not, the chief benefit of these format guidelines is that they act as a set of standards for the input and output format of many different functions in many different R packages, in particular those in the `tidyverse` suite of R packages for data science. These guidelines are espoused in the first principle of the "tidy tools manifesto" to "reuse existing data structures" (Wickham 2017d). By following this formatting guideline, all datasets in the `fivethirtyeight` package fit seamlessly into the `tidyverse` ecosystem of R packages, in particular the `ggplot2` package for data visualization and the `dplyr` package for data wrangling.

Converting datasets to "tidy" format in R necessitates understanding of the `gather()` and `spread()` commands of the `tidyr` package. We argue that the topic of converting between wide data format and tidy/long format is best left to an intermediate stage of learning, given the relative complexity of the necessary commands. Therefore we pre-processed all data in the package to follow the tidy data format. For datasets where we felt converting the data to tidy format would alter them too much from their original form, we include the `gather()` code necessary to convert them from wide data format to tidy format in the `@examples` section of the corresponding help files.
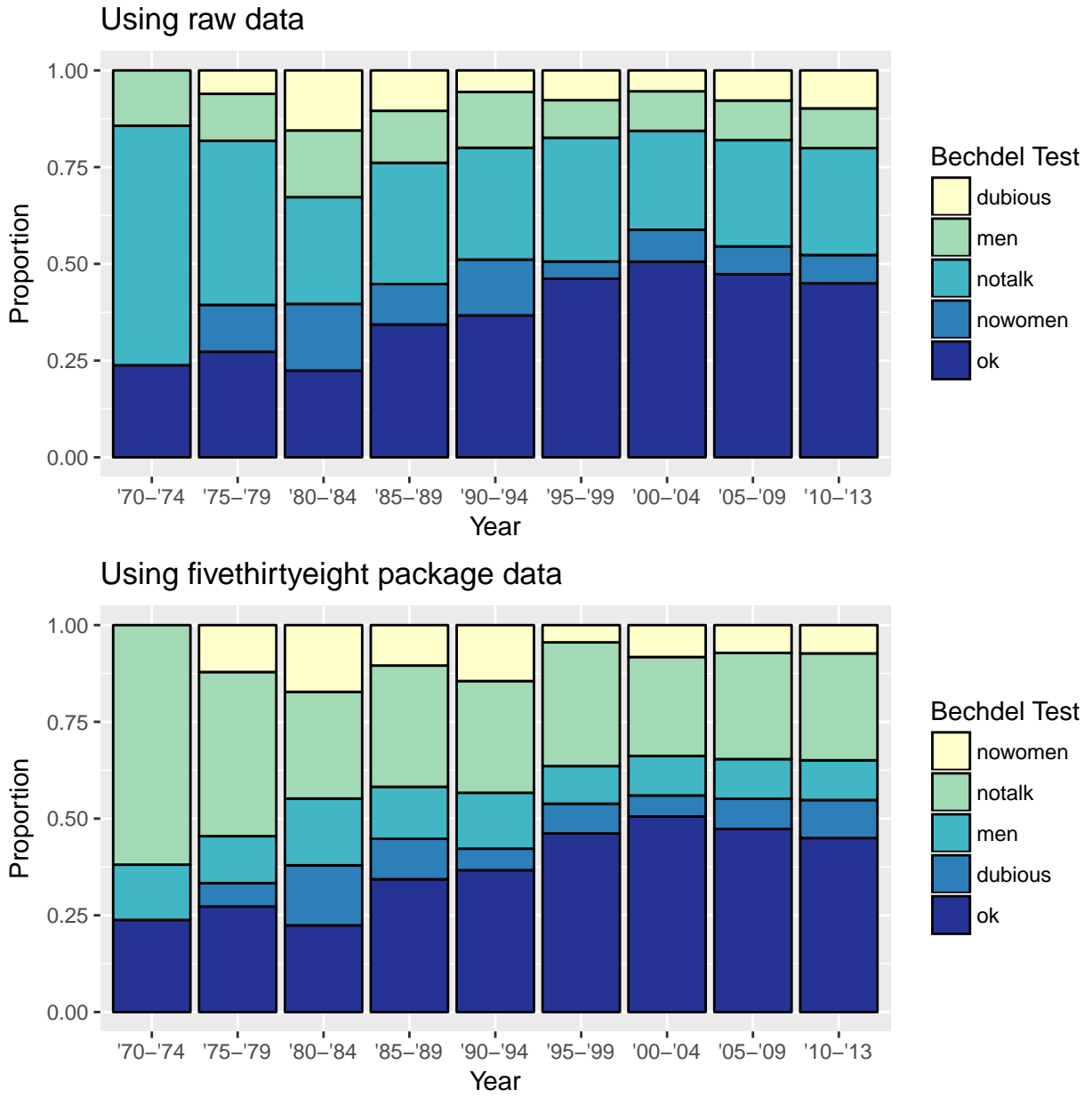
Figure 6: Barcharts of Bechdel Test results across time.

For example, say we want to create a barplot comparing consumption of beer, spirits, and wine between the United States and France using the `drinks` dataset corresponding to the article "Dear Mona Followup: Where Do People Drink The Most Beer, Wine And Spirits?" (Chalabi 2014) The data is saved in "wide" format and thus cannot be used in the `ggplot()` function.

```
library(tidyverse)
library(fivethirtyeight)
drinks %>%
  filter(country %in% c("USA", "France"))
#> # A tibble: 2 x 5
#>   country beer_servings spirit_servings wine_servings
#>     <chr>         <int>           <int>         <int>
#> 1  France           127             151           370
#> 2     USA           249             158            84
#> # ... with 1 more variables: total_litres_of_pure_alcohol <dbl>
```

However, the help file for this dataset, accessible by typing `?drinks` in the console, provides the `gather()` code necessary to convert this data into "tidy" format:

```
drinks_tidy_US_FR <- drinks %>%
  filter(country %in% c("USA", "France")) %>%
  gather(type, servings, -c(country, total_litres_of_pure_alcohol))
drinks_tidy_US_FR
#> # A tibble: 6 x 4
#>   country total_litres_of_pure_alcohol            type servings
#>     <chr>                        <dbl>           <chr>    <int>
#> 1  France                         11.8   beer_servings      127
#> 2     USA                          8.7   beer_servings      249
#> 3  France                         11.8 spirit_servings      151
#> 4     USA                          8.7 spirit_servings      158
#> 5  France                         11.8   wine_servings      370
#> 6     USA                          8.7   wine_servings       84
```

This formatting of the data now allows itself to be used as input to the `ggplot()` function to create an appropriate barplot in Figure 7. Note in this case since the number of servings is pretabulated in the variable `servings`, which in turn is mapped to the y-axis, we use `geom_col()` instead of `geom_bar()` (`geom_col()` is equivalent to `geom_bar(stat = "identity")`).

```
ggplot(drinks_tidy_US_FR, aes(x=type, y=servings, fill=country)) +
  geom_col(position = "dodge") +
  labs(x = "Alcohol type", y = "Average number of servings")
```
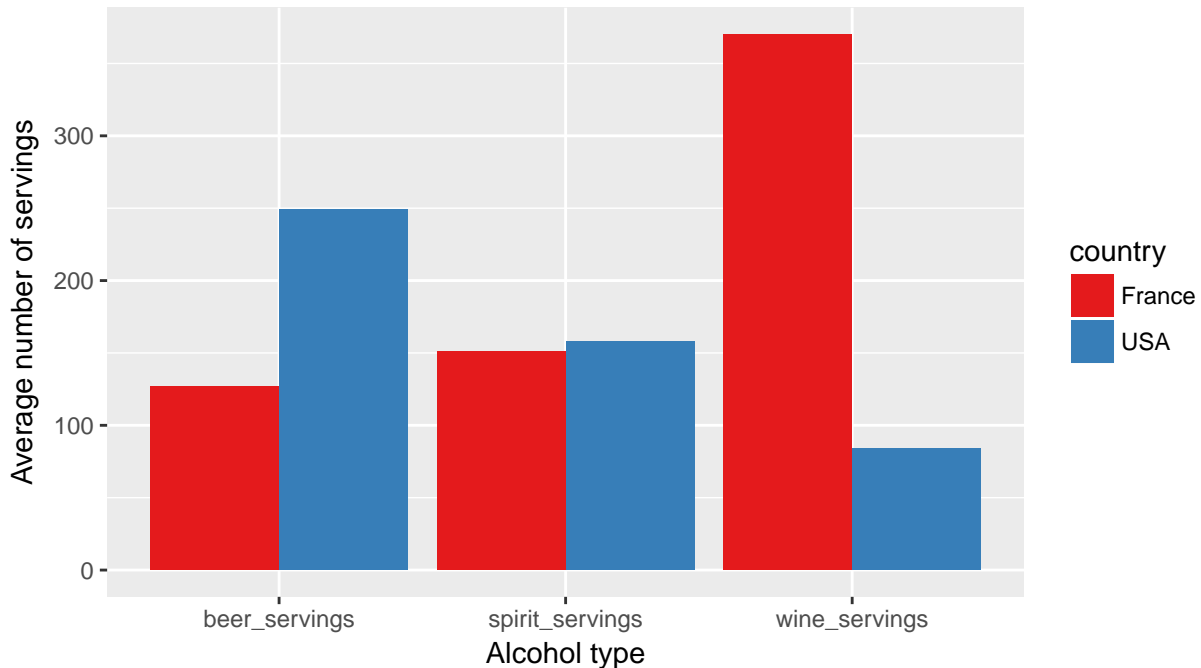
Figure 7: USA vs France alcohol consumption.

Conclusion

The Future

One immediate application of this package for instructors is to increase the data-centricism of their courses with minimal disruption to the syllabus. The data are easy to load via an R package and the help files contain contextual information, including a description of all variables and a link to the original article. For courses with a data science component, the datasets can provide students opportunities to "flex their data science muscles," in particular those relating to data visualization and wrangling. For example, Ismay had their SOC 301 Social Statistics students at Pacific University use these datasets for their group projects, which can be accessed at https://ismayc.github.io/soc301_s2017/. Finally, as stated in Section 2.3, for courses where data wrangling is a learning goal in and of itself, the source code converting the raw data to tame data format can be used as examples for students to learn more sophisticated data wrangling tools, such as those needed to wrangle dates and strings/text data.

Additionally, this package is being used in other statistics and data science pedagogical initiatives. Ismay and Kim have incorporated this package into "ModernDive: An Introduction to Statistical and Data Sciences via R" (Ismay and Kim 2017a), an open source, fully reproducible electronic textbook available at http://www.moderndive.com. Ismay is creating a companion DataCamp course "Effective Data Storytelling Using the `tidyverse`" (Ismay and Kim 2017b); DataCamp is an online interactive environment for learning data science currently via R and Python.

Biographies

Albert Y. Kim was an assistant professor of statistics at Middlebury College in Middlebury, Vermont, where he was a statistician in the Department of Mathematics. He taught an introductory statistics class that acts as a service class for disciplines needing their students be adept at statistical thinking. He is currently a Lecturer of Statistics at Amherst College, and will be joining Smith College as an Assistant Professor of Statistical and Data Sciences in the fall of 2018.

Chester Ismay was formerly a Data Science/Statistics Consultant and Visiting Assistant Professor of Mathematics at Reed College. He also taught Social Statistics for the Sociology Department at Pacific University. Prior to moving to the Portland, Oregon area, he taught introductory statistics, data analysis, probability, mathematical statistics, and introduction to computer science as the lone statistician at Ripon College. His focus has been and remains to be improving the communication of statistics and data science via clear visualization and careful, deliberate analysis. He is currently a Data Science Curriculum Lead at DataCamp.

Jennifer Chunn was formerly a full-time lecturer at Seattle University in the Albers School of Business and Economics. She taught required introductory statistics and regression courses to Business undergraduate majors, as well as a senior-level Applied Econometrics elective. In addition, she incorporated examples from the fivethirtyeight package into her Data Visualization and Statistical Modeling course in the Executive MBA program. She is currently a Manager of Data Science and Analytics at Nordstrom, Inc.

Acknowledgements

References

Allaire, J. J., Cheng, J., Xie, Y., McPherson, J., Chang, W., Allen, J., Wickham, H., Atkins, A., Hyndman, R., and Arslan, R. (2017), "rmarkdown: Dynamic Documents for R." R package version 1.6. *https://CRAN.R-project.org/package=rmarkdown.*

American Statistical Association Undergraduate Guidelines Workgroup (2014), *2014 Curriculum Guidelines for Undergraduate Programs in Statistical Science*, Alexandria, VA: American Statistical Association. Available at http://www.amstat.org/asa/files/pdfs/EDU-guidelines2014-11-15.pdf.

Bialik, C. (2016), "Some People Are Too Superstitious To Have A Baby On Friday The 13th," *FiveThirtyEight* [online]. Available at https://fivethirtyeight.com/features/some-people-are-too-superstitious-to-have-a-baby-on-friday-the-13th/.

Bryan, J. (2015), "gapminder: Data from Gapminder." R package version 0.3.0. https://CRAN.R-project.org/package=gapminder.

Chalabi, M. (2014), "Dear Mona Followup: Where Do People Drink The Most Beer, Wine And Spirits?," *FiveThirtyEight* [online]. Available at https://fivethirtyeight.com/features/some-people-are-too-superstitious-to-have-a-baby-on-friday-the-13th/.

Cobb, G. (2015), "Mere Renovation is Too Little Too Late: We Need to Rethink our Undergraduate Curriculum from the Ground Up," *The American Statistician* 69, 266-282.

DeVeaux, R., Agarwal, M., Averett, M., Baumer, B., Bray, A., Bressoud, T., Bryant, L., Cheng, L., Francis, A., Gould, R., Kim, A. Y., Kretchmar, M., Lu, Q., Moskol, A., Deborah Nolan, Pelayo, R., Raleigh, S., Sethi, R. J., Sondjaja, M., Tiruviluamala, N., Uhlig, P., Washington, T., Wesley, C., White, D., and Ye, P. (2016), "Curriculum Guidelines for Undergraduate Programs in Data Science," *The Annual Review of Statistics and Its Application* 4, 15-30.

GAISE College Report ASA Revision Committee (2016), *Guidelines for Assessment and Instruction in Statistics Education (GAISE) in Statistics Education (GAISE) College Report College Report 2016*, Alexandria, VA: American Statistical Association. Available at http://www.amstat.org/education/gaise, last accessed August 7, 2017.

Gould, R. (2010), "Statistics and the Modern Student," *International Statistics Review* 78, 297–315.

Grolemund, G., and Wickham, H. (2011), "Dates and Times Made Easy with lubridate," *Journal of Statistical Software* 40, 1–25. Available at: http://www.jstatsoft.org/v40/i03/.

Grolemund, G., and Wickham, H. (2017), *R for Data Science*, O'Reilly Media. Available at: http://r4ds.had.co.nz/.

Guzdial, M., and Tew, A. E. (2006), "Imagineering Inauthentic Legitimate Peripheral Participation: An Instructional Design Approach for Motivating Computing Education," in *Proceedings of the Second International Workshop on Computing Education Research*, ICER '06, New York, NY, USA: ACM, pp. 51–58. DOI: 10.1145/1151588.1151597.

Hickey, W. (2014a), "41 Percent Of Fliers Think You're Rude If You Recline Your Seat," *FiveThirtyEight* [online]. Available at: https://fivethirtyeight.com/datalab/airplane-etiquette-recline-seat/.

Hickey, W. (2014b), "The Dollar-And-Cents Case Against Hollywood's Exclusion of Women," *FiveThirtyEight* [online]. Available at: https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/.

Horton, N. J., Baumer, B., and Wickham, H. (2015), "Setting the stage for data science: integration of data management skills in introductory and second courses in statistics," *CHANCE,* 28, 40–50.

Ismay, C., and Kim, A. Y. (2017a), *An Introduction to Statistical and Data Sciences via R* [online]. Available at: http://www.moderndive.com/.

Ismay, C., and Kim, A. Y. (2017b), "Effective Data Storytelling Using the `tidyverse`," *DataCamp* [online]. Available at: https://www.datacamp.com/courses/effective-data-storytelling-using-the-tidyverse.

Jedamski, D. [d_jedamski]. (2015), .@JennyBryan describes data in school as "teddy bear", real world data more like "grizzly bear w/ salmon blood dripping from mouth" #JSM2015 [Tweet]. Retrieved from https://twitter.com/d_jedamski/status/631150332941332480.

Kim, A. Y., and Escobedo-Land, A. (2015), "OkCupid Profile Data for Introductory Statistics and Data Science Courses," *Journal of Statistics Education* 23, 97–107.

Lave, J., and Wenger, E. (1991), *Situated Learning: Legitimate Peripheral Participation*, Cambridge University Press.

Lohr, S. (2014), "For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights." *New York Times* [online]. Available at: https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html.

Majumder, M. (2017), "Higher Rates Of Hate Crimes Are Tied To Income Inequality," *FiveThirtyEight* [online]. Available at https://fivethirtyeight.com/features/higher-rates-of-hate-crimes-are-tied-to-income-inequality/.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. H. (2011). "Big data: The next frontier for innovation, competition, and productivity," *McKinsey Global Institute* [online]. Available at: http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation.

McNamara, A., and Horton, N. J. (2017), "Wrangling categorical data in R." *The American Statistician*. DOI: 10.1080/00031305.2017.1356375.

Nolan, D., and Lang, D. T. (2010), "Computing in the Statistics Curricula," *The American Statistician,* 64, 97–107.

Perkins, D. (2010), *Making Learning Whole: How Seven Principles of Teaching Can Transform Education*, Jossey-Bass.

Wickham, H. (2014), "Tidy Data," *Journal of Statistical Software* 59, 1-23. DOI: 10.18637/jss.v059.i10.

Wickham, H. (2015), *R Packages: Organize, Test, Document, and Share Your Code*. O'Reilly. Available at: http://r-pkgs.had.co.nz/.

Wickham, H. (2017a), "babynames: US Baby Names 1880-2015." R package version 0.3.0.

https://CRAN.R-project.org/package=babynames.

Wickham, H. (2017b), "forcats: Tools for Working with Categorical Variables (Factors)." R package version 0.2.0. https://CRAN.R-project.org/package=forcats.

Wickham, H. (2017c), *The tidy tools manifesto* [online]. Retrieved from https://cran.r-project.org/web/packages/tidyverse/vignettes/manifesto.html.

Wickham, H. (2017d), "tidyverse: Easily Install and Load 'Tidyverse' Packages." R package version 1.1.1. https://CRAN.R-project.org/package=tidyverse.