Transparent and Reproducible Research with R

Daniel Anderson[1] & Joshua Rosenberg[2]

[1] University of Oregon
[2] University of Tennessee, Knoxville

Reproducibility of research findings is critical to the validity of inferences from studies. If an independent evaluator with access to the study data is unable to reproduce the published findings *exactly*, the trustworthiness of the findings are called into question, as highlighted by several prominent examples (e.g., the Duke crisis; see Peng, 2015). Such issues are also essential in a time in which established findings are being called into question (i.e., concerns about the reproducibility of psychological science findings) because of choices made in the course of a research study.

In this training, we provide an overview of reproducibility and open science, and introduce participants to tools that increase the likelihood of reproducible and transparent workflows. We emphasize tools from the R software environment to weave text with analysis code (e.g., R Markdown), version control to document the entire history of a project, and platforms for sharing analysis workflows publicly.

In the first hour of this four-hour training, we introduce participants to the ideas motivating open and reproducible research in educational research. In the second and third hours, we discuss the basics of R Markdown and the various formats to which documents can be rendered. Finally, in the fourth hour, we provide a primer on version control using *GitHub*. Our target audience is early-career scholars as well as researchers at any stage looking for tools to help increase the likelihood of their work being reproducible. The format will include part lecture and part hands-on applied work.

Correspondence concerning this article should be addressed to Daniel Anderson, 5262 University of Oregon. E-mail: daniela@uoregon.edu

Course Faculty Course Format

List of instructors

Presenters	Affiliation	Email
Daniel Anderson	University of Oregon	daniela@uoregon.edu
Joshua Rosenberg	University of Tennessee, Knoxville	jrosenb8@utk.edu

Course format

This tutorial will be a **4-hour mini-course** on tools for conducting open and reproducible research with R. It will overview the idea and importance of reproducible research for educational researchers, cover R Markdown for weaving text with code, and discuss *git*/*GitHub* for version control and collaboration. Part of the benefit of R Markdown is that the syntax itself is relatively straightforward (learning to code with R is much more difficult than learning R Markdown), but is a powerful framework for presenting research findings in a variety of formats, including static web pages, slides, technical reports, and even APA-formatted journal articles. While the heart of our presentation will focus on helping participants to become familiar with and comfortable using R Markdown and version control toward the ultimate goal of more open, transparent, and reproducible analysis workflows, we also spend some time discussing these alternative output formats by way of motivating participants to continue learning beyond the mini-course. Both of the authors are experienced R users and have been using reproducible research principles in their applied work for a number of years.

Proposal narrative: AERA Professional Development Proposal

The purpose of this proposal is to provide participants with an introduction to tools for open, transparent, and reproducible analysis workflows. Carrying out educational research in reproducible frameworks represents a compromise between publishing a journal article alone, and full-scale replication (i.e., the "gold standard"). Some (e.g., R. D. Peng, 2011) have argued that reproducibility should be a *minimal* standard as a means combating the *replicability crisis* (see Hedges, 2018). We highlight *R Markdown*, a tool for integrating text with code, and *git/GitHub*, for documentation of the project history and to support collaboration. Through the use of these tools, quantitative analyses are more open and transparent, and the likelihood of reproducibility and trustworthiness is increased.

Prerequisite skills or knowledge needed for course participation

All participants should have an interest in conducting open and transparent analyses. This training will be most useful to those with experience planning, carrying out, and writing up the results of a research project. We will specifically discuss moving from existing frameworks to reproducible workflows using R, R Markdown, and *git/GitHub*.

All participants should have at least a basic familiarity with R and be comfortable with the idea of working in a scripting environment. Note that the presenters have partnered with DataCamp (https://www.datacamp.com) through the use of an academic account to provide a platform for less experienced users to get "up to speed" prior to the training. DataCamp is an online learning platform for R (and related data science technologies) that includes direct instruction and opportunities to practice and apply the learned skills, all within the online platform. Users with less experience with R will be asked to complete the Introduction to R, Working with the RStudio IDE: Part 1, and Introduction to the *tidyverse* online modules prior to the training. Following the training, we recommend all participants to re-visit the skills they have learned by completing the Reporting with R Markdown module.

Target course participants

Our target audience includes graduate students, emerging or early-career researchers, and continuing researchers interested in their own or their students'/trainees' work becoming more open, transparent, and reproducible. Because both an overview of ideas related to reproducibility as well as a number of quantitative and computational tools and approaches will be described, participants with less experience can benefit from developing a more conceptual understanding of open science and can turn to the tools later. This training is aimed at beginners who have little to no experience with R Markdown or version control, but who feel comfortable learning code and have at least a basic understanding of R.

Rationale

The basic premise of reproducible research is that quantitative analyses should be conducted and documented with sufficient clarity that independent researchers could reproduce all the results, exactly. Initially, this may sound relatively straightforward—of course research findings should be reproducible—and we may like to think that most research in education

adheres to these principles. Unfortunately, this is generally not the case. For example, in a large- scale review of growth models published from 2007-2012 across 47 education and psychology journals, Stevens J. J. and Tindal (2013) found that documenting even relatively routine procedures, such as how missing data were handled, was extraordinarily difficult. Indeed, in the vast majority of cases, this information was simply missing. These findings are congruent with Buckheit and Donoho (1995), who argue "an article about computational science in a scientific publication is **not** the scholarship itself, it is merely the **advertising** of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures" (p. 5, emphasis in the original). This may seem a somewhat extreme view, but it demonstrates that a journal article on its own is generally insufficient for the accumulation of scientific evidence. That is, it is difficult to build off the work of others, or verify study results, if you only have access to the published findings. We note that reproducibility does not imply "correctness", but rather a transparency in process: In fact, part of the reason reproducible research is essential is that it allows other researchers to verify the process and analysis, and ultimately, the validity of the inferences made from a study.

Conducting reproducible research. Considerable recent attention has been paid to open and reproducible research in science generally (Bartling & Friesike, 2014; National Academies of Sciences & Medicine, 2018), but also in educational research (Cook, Lloyd, Mellor, Nosek, & Therrien, 2018; McBee, Makel, Peters, & Matthews, 2017; Zee & Reich, 2018). What is often lacking, however, is a clear accounting of how to actually begin engaging in open and reproducible research. This training seeks to fill that gap by providing an initial introduction to tools to help make the process more tractable. In particular, we advocate for *conducting science publicly* through the publication of living code that is modified and updated as the project matures, along with *literate programming*, a concept introduced by Knuth (1984) that weaves substantive text from the manuscript with analysis code. Although literate programming has its own learning curve, leading to an initial dip in productivity, it can eventually lead to large gains in efficiency by all tables, figures, and in-text references to statistics (e.g., sample means) being produced through code and updated automatically. There is therefore no hand- entering of data/statistics into tables, which can be error-prone, and any tweaks to the model or data (including new data being added to the research) results in the entire document being updated automatically. The manuscript is therefore *dynamic* relative to the analysis (Xie, 2016). This process may sound complicated to implement, and it was even a few short years ago. Yet, the toolkit for producing dynamic, reproducible documents is rapidly expanding and is now far more accessible for the applied researcher. When this process is paired with a version control system such as *git*, and made publicly available through platforms such as *GitHub*, the project and process are far more open and transparent. Importantly, however, our training also discusses methods to ensure specific parts of the project (e.g., the raw data) are *not* available publicly.

Learning objectives

Participants in this training will (1) understand why reproducibility is an increasingly important consideration for educational researchers, (2) be introduced to tools, specifically R Markdown and *git*/*GitHub*, for carrying out open and reproducible research, and (3) understand some of the remaining challenges that remain for open and reproducible research.

Course content

Our course introduces participants to the fundamental tenants of open and reproducible analysis workflows, including literate programming, documenting the project history, and working from a public platform (*GitHub*). Given that the training is four hours, we aim only to introduce participants to these concepts. However, despite the session serving as a primer, we prioritize hands-on applied practice. In our experience, the initial step in getting started can often be the greatest hurdle. A preliminary schedule follows:

Hour 1: Introduction. The first hour will focus primarily on the substantive side of reproducible research–i.e., why are we all here? We will discuss the importance of reproducible research, covering infamous case-studies such as the Duke crisis (R. Peng, 2015) and others, but also introduce the ideas of literate programming and conducting science publicly. During the first hour, no specific code or tools will be covered and the focus will be on high-level conceptual understandings. The format will be primarily lecturing with slides, with brief breakout sessions (< 5 minutes) in groups to discuss the covered topics.

Hour 2: R Markdown I. The second hour will be more hands-on and applied, asking participants to follow-along with one instructor, while the other roams the room and helps participants who are having trouble. We will introduce the very basics of R Markdown, including delineating code chunks from plain text, creating headers at different levels, creating bulleted lists, and bolding/italicizing text. We will also cover different code chunk options, including hiding/showing the code evaluating/not evaluating the code chunk. By the end of Hour 1, all participants should have at least a basic R Markdown document rendered to HTML with each of the aforementioned features.

Hour 3: R Markdown II. In the third hour we discuss moving R Markdown documents to different formats, including PDF and Microsoft Word. We then discuss the *papaja* (Aust & Barth, 2018) R package for preparing APA formatted manuscripts using the same basic R Markdown features. We also discuss including references within an R Markdown document. Again, one instructor will lead a guided walk-through while the other circulates the room to assist participants. One possible complication with *papaja* is that it requires a *tex* distribution. We plan to address this by providing instructions prior to the training on installing the *tinytex* (Xie, 2018) package, which has all the required functionality but is a much smaller installation. By the end of Hour 3, all participants should have a basic APA formatted manuscript with at least one in-text citation and accompanying bibliography.

Hour 4: Use of git/GitHub. In the final hour of the training, we introduce participants to *GitHub*. The first 20 minutes of this section will be devoted to lecture, introducing participants to the basic commands (i.e., what a *repository* or *repo* is, as well as what it means to *stage*, *commit*, *pull*, *push*, and *clone*). In the last 40 minutes we guide participants through creating a new repository and pushing their existing project to that repository. We then walk them through the process of making changes, committing those changes, and pushing them to the repository. Finally, we walk participants through the basics of *GitHub* platform to view the history of a project and conduct work openly. We also discuss the use of a *.gitignore* file to ensure specific files do *not* get pushed to the repository. By the end of Hour 4, all participants should have created their first publicly viewable GitHub repository that documents the history of their reproducible project from the workshop from their initial commit.

References

Aust, F., & Barth, M. (2018). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Bartling, S., & Friesike, S. (2014). *Opening science: The evolving guide on how the internet is changing research, collaboration and scholarly publishing*. Springer.

Buckheit, J. B., & Donoho, D. L. (1995). Wavelab and reproducible research. In *Wavelets and statistics* (pp. 55–81). Springer.

Cook, B. G., Lloyd, J. W., Mellor, D., Nosek, B. A., & Therrien, W. (2018). Promoting open science to increase the trustworthiness of evidence in special education.

Hedges, L. V. (2018). Challenges in building usable knowledge in education. *Journal of Research on Educational Effectiveness*, *11*(1), 1–21.

Knuth, D. E. (1984). Literate programming. *The Computer Journal*, *27*(2), 97–111.

McBee, M., Makel, M., Peters, S. J., & Matthews, M. S. (2017). A manifesto for open science in giftedness research.

National Academies of Sciences, Engineering, & Medicine. (2018). *Open science by design: Realizing a vision for 21st century research*. Washington, DC: The National Academies Press. doi:10.17226/25116

Peng, R. (2015). The reproducibility crisis in science: A statistical counterattack. *Significance*, *12*(3), 30–32.

Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*(6060), 1226–1227.

Stevens J. J., Nese J. F. T., & Tindal, G. (2013). *Using longitudinal models to track student achievement: A literature synthesis.* Unpublished manuscript.

Xie, Y. (2016). *Dynamic documents with r and knitr*. Chapman; Hall/CRC.

Xie, Y. (2018). *Tinytex: Helper functions to install and maintain 'tex live', and compile 'latex' documents*. Retrieved from https://CRAN.R-project.org/package=tinytex

Zee, T. van der, & Reich, J. (2018). Open education science. *AERA Open*, *4*(3), 2332858418787466.

DanielAnderson

Research Assistant Professor: University of Oregon

contact
5262 Universitiy of Oregon, Eugene, OR
daniela@uoregon.edu

website
www.datalorax.com.com

twitter
datalorax

github
datalorax

programming
dev: R **use:** R, CSS, Mplus

Interests

Data science education, `R` package development, open science, growth modeling, predictive modeling, achievement gaps

Education

PhD Educational Research Methodology — Unviversity of Oregon
Dissertation: Teacher and School Contributions to Student Growth

MS Educational Leadership — University of Oregon

BS Elementary Education — Utah State University

Teaching

2018-19	First three courses of a planned 5-course data science specialization.
Fall '17	Exploring Data with R
Spring '17	Exploring Data with R
Winter '17	A taste of R: 4-session mini-course for faculty development

Selected Publications

11 Total publications. See here for a full listing.

Fien, H., **Anderson, D**, Nelson, N. J., Kennedy, P., Baker, S. K., & Stoolmiller, M. (2018). Examining the impact and school-level predictors of impact variability of an 8th grade reading intervention on at-risk students' reading achievement. *Learning Disabilities Research & Practice*, *33*(1), 37–50.

Anderson, D, Kahn, J. D., & Tindal, G. (2017). Exploring the robustness of a unidimensional item response theory model with empirically multidimensional data. *Applied Measurement in Education*, *30*(3), 163–177.

Farley, D., **Anderson, D**, Irvin, P. S., & Tindal, G. (2017). Modeling reading growth in grades 3 to 5 with an alternate assessment. *Remedial and Special Education*, *38*(4), 195–206.

Anderson, D, Irvin, S., Alonzo, J., & Tindal, G. (2015). Gauging item alignment through online systems while controlling for rater effects. *Educational Measurement: Issues and Practice*, *34*(1), 22–33.

Anderson, D, Farley, D., & Tindal, G. (2015). Test design considerations for students with significant cognitive disabilities. *The Journal of Special Education*, *49*(1), 3–15.

Joshua M. Rosenberg

Assistant Professor, University of Tennessee, Knoxville
420 Claxton Complex, 1126 Volunteer Blvd., Knoxville, TN, 37996-3452
jrosenb8@utk.edu 865-974-5973 jmichaelrosenberg.com | *Updated:* July 23, 2018

Education

PhD, Educational Psychology & Educational Technology
Michigan State University 2018

Professional Experience

Assistant Professor of STEM Education 2018
Department of Theory and Practice in Teacher Education
University of Tennessee, Knoxville

Representative Publications (11 peer-reviewed journal articles in total)

Beymer, P. N., Rosenberg, J. M., Schmidt, J. A., & Naftzger, N. (2018). Examining relationships between choice, affect, and engagement in out-of-school time STEM programs. *Journal of Youth and Adolescence, 47*(6), 1178-1191. https://doi.org/10.1007/s10964-018-0814-9

Akcaoglu, M., Rosenberg, J. M., Ranellucci, J., & Schwarz, C. V. (2018). Outcomes from a self-generated utility value intervention on fifth and sixth-grade students' value and interest in science. *International Journal of Educational Research, 87*, 67-77. https://www.sciencedirect.com/science/article/pii/S0883035517308492

Schmidt, J. A., Rosenberg, J. M., & Beymer, P. (2018). A person-in-context approach to student engagement in science: Examining learning activities and choice. *Journal of Research in Science Teaching, 55*(1), 19-43. https://dx.doi.org/10.1002/tea.21409

Rosenberg, J. M., Greenhalgh, S. P., Koehler, M. J., Hamilton, E., & Akcaoglu, M. (2016). An investigation of State Educational Twitter Hashtags (SETHs) as affinity spaces. *E-Learning and Digital Media, 13*(1-2), 24-44. http://dx.doi.org/10.1177/2042753016672351

Select Invited Talks

Rosenberg, J. M. (March, 2016). An introduction to R for programming and statistical analysis in education. Georgia Southern University College of Education, Statesboro, GA.

Select R packages

tidyLPA: Interface to MCLUST to perform Latent Profile Analysis in R (w/ J. Schmidt, P. Beymer, & R. Steingut; v. 0.1.0). https://cran.r-project.org/web/packages/tidyLPA/index.html

konfound: R package to carry out sensitivity analysis (with R. Xu & K. Frank). https://github.com/jrosen48/konfound

Examples of Past Presentations on Related Content

Dynamic Documents: An Introduction

Day 2: Morning, Session 1

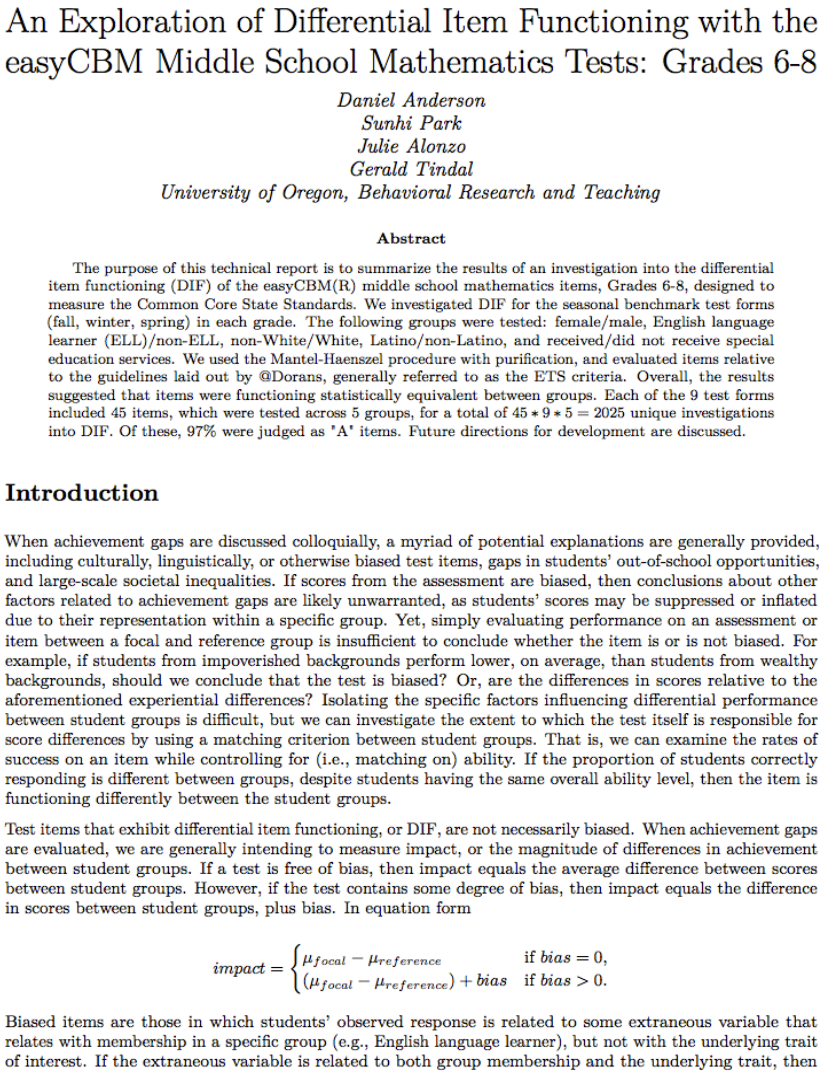
Daniel Anderson
R Training: Florida State University, June 22, 2016

A Brief Intro to dynamic documents

(focusing mostly on R Markdown)

What is R Markdown

- Simple language for converting R code/output into other formats, most notably HTML and PDF
- These slides were produced using a variant of Markdown



An Exploration of Differential Item Functioning with the easyCBM Middle School Mathematics Tests: Grades 6-8

Daniel Anderson
Sunhi Park
Julie Alonzo
Gerald Tindal
University of Oregon, Behavioral Research and Teaching

Abstract

The purpose of this technical report is to summarize the results of an investigation into the differential item functioning (DIF) of the easyCBM(R) middle school mathematics items, Grades 6-8, designed to measure the Common Core State Standards. We investigated DIF for the seasonal benchmark test forms (fall, winter, spring) in each grade. The following groups were tested: female/male, English language learner (ELL)/non-ELL, non-White/White, Latino/non-Latino, and received/did not receive special education services. We used the Mantel-Haenszel procedure with purification, and evaluated items relative to the guidelines laid out by @Dorans, generally referred to as the ETS criteria. Overall, the results suggested that items were functioning statistically equivalent between groups. Each of the 9 test forms included 45 items, which were tested across 5 groups, for a total of $45 * 9 * 5 = 2025$ unique investigations into DIF. Of these, 97% were judged as "A" items. Future directions for development are discussed.

Introduction

When achievement gaps are discussed colloquially, a myriad of potential explanations are generally provided, including culturally, linguistically, or otherwise biased test items, gaps in students' out-of-school opportunities, and large-scale societal inequalities. If scores from the assessment are biased, then conclusions about other factors related to achievement gaps are likely unwarranted, as students' scores may be suppressed or inflated due to their representation within a specific group. Yet, simply evaluating performance on an assessment or item between a focal and reference group is insufficient to conclude whether the item is or is not biased. For example, if students from impoverished backgrounds perform lower, on average, than students from wealthy backgrounds, should we conclude that the test is biased? Or, are the differences in scores relative to the aforementioned experiential differences? Isolating the specific factors influencing differential performance between student groups is difficult, but we can investigate the extent to which the test itself is responsible for score differences by using a matching criterion between student groups. That is, we can examine the rates of success on an item while controlling for (i.e., matching on) ability. If the proportion of students correctly responding is different between groups, despite students having the same overall ability level, then the item is functioning differently between the student groups.

Test items that exhibit differential item functioning, or DIF, are not necessarily biased. When achievement gaps are evaluated, we are generally intending to measure impact, or the magnitude of differences in achievement between student groups. If a test is free of bias, then impact equals the average difference between scores between student groups. However, if the test contains some degree of bias, then impact equals the difference in scores between student groups, plus bias. In equation form

$$impact = \begin{cases} \mu_{focal} - \mu_{reference} & \text{if } bias = 0, \\ (\mu_{focal} - \mu_{reference}) + bias & \text{if } bias > 0. \end{cases}$$

Biased items are those in which students' observed response is related to some extraneous variable that relates with membership in a specific group (e.g., English language learner), but not with the underlying trait of interest. If the extraneous variable is related to both group membership and the underlying trait, then

Why produce dynamic documents?

- Reproducible research principles
    - Increase transparency
- Can (eventually) be more efficient
- Simple for simple tasks (like homeworks)
    - Complexity increases as you ask more of it

Before we get too far...

Reproducible research

A couple caveats

- Much of what I'm going to be discussing is largely *not* how I have interacted with research to this point. Instead, it represents an ideal that I have only recently begun working towards.
- None of what I will talk about should be taken as a referendum on you or your current practices. However, I hope to convince you that you should be working toward the reproducible research ideal, and that, as a field, we should be moving toward reproducible research being the *minimal standard*.
- I will be focusing on reproducible research with R. Other options are available but, in my view, none are as clear, comprehensive, and easy to implement as the tools at your disposal through R.

What is reproducible research?

- **Replicability** is the gold standard for research. Ideally, most research would be verified through replication.
- Reproducibility represents a minimal standard, which itself can aid replication (tremendously), by conducting and documenting the research sufficiently that **an independent researcher could reproduce all the results from a study**, provided the data were available

Why should we care?

- Reproducibility as an ethical standard
  - More transparency
  - More potential for results to be verified (and errors found/corrected)
- If your work **is not** reproducible, it is usually not truly replicable.
- If your work **is** replicable, then others have a "recipe" for replication

Are journal articles research?

- Initially, we may think of journal articles as research, but really the research is everything that went into the article, not the article itself.
- Some (Buckheit & Donoho, 2015) conceive of the article as the "advertisement".
- If all we have is the advertisement, can we really fully understand the steps and decisions made during the research?
    - In large-scale data analysis, the answer is generally "no".

Tangential benefits

Striving toward reproducible research will:

- Make your own code more efficient/easily interpretable
    - Can help with collaboration on a project
- Reduce errors
- Increase efficiency by not having to redo tables and figures with each tweak to a model.

What does the process actually look like?

- Start with a basic text document (not Word, text)
- Use the text document to write your article
- Embed code within the text document that corresponds to your analysis. Note this is not just copying the code in. The code should be live and what you're working with while conducting your research.
- Render the document into a different format (pdf, html, etc.).
    - Select which code (if any) will be displayed
    - Build tables of results and plots to be produced
- Readers can then read the "advertisement", but if they are interested in reproducing your results (maybe because they disagree with you, or they think your results are weird and want to clearly see all the steps you took), they can access the text file that contains the computer code.
- The end result is a single product that has the advertisement and the research process embedded.

Other reasons dynamic documents are useful

Outside of reproducibility, you may want to use R Markdown to:

- Produce slides
    - Just be careful, I have a horror story
- Keep track of your analysis (notes, essentially), even if you end up using something like Word
- Share code with others
- Quickly share results with others
- etc... ideas?

On to the mechanics

YAML Front Matter

Not explicitly necessary, but generally helpful

```
---
title: Example Markdown document
author: Daniel Anderson
date: "2015-09-17"
---
```

Example Markdown document

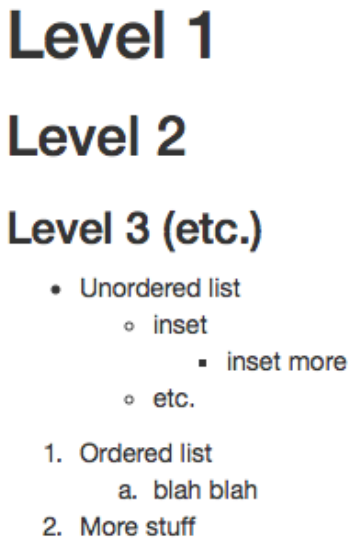
Daniel Anderson

2015-09-17

- Three dashes before and after the YAML fields
- Case sensitive
- Many other fields are possible.
    - For example, you may want to include an `output:` argument (`pdf_document`, `html_document`, `word_document`). Must be specified as it is rendered, if not supplied.

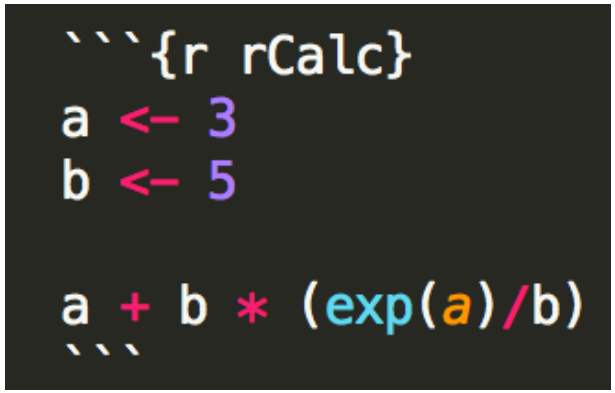
Headings and Lists

```
# Level 1
## Level 2
### Level 3 (etc.)


* Unordered list
  - inset
    + inset more
  - etc.

1. Ordered list
  a. blah blah
2. More stuff
```



Code chunks

Start a code chunk with ```{r chunkName, chunkOptions}, then produce some r code, then close the chunk with three additional back ticks ```.




```
a <- 3
b <- 5

a + b * (exp(a)/b)
```

```
## [1] 23.08554
```

A few select chunk options

OPTIONS	ARGUMENTS	DEFAULT	RESULT
eval	logical	TRUE	Evaluate the code?
echo	logical	TRUE	Show the code?
results	markup, asis, hold, hide	markup	Render the results
warning	logical	TRUE	Print warnings?
error	logical	TRUE	Preserve errors? (if FALSE, quit)
message	logical	TRUE	Print any messages?
include	logical	TRUE	Include any of the code or output or code?
tidy	logical	FALSE	Tidy code? (see formatR package)

(and a few more)

	OPTIONS	ARGUMENTS	DEFAULT	RESULT
9	cache	logical, 0:3	FALSE	Cache code chunks?
10	cache.comments	logical	NULL	Cache invalidated by comment changes?
11	dependson	char, num	NULL	Current chunk depend on prior cached chunks?
12	autodep	logical	FALSE	Should dependencies be determined automatically? (if TRUE, no need for dependson)
13	fig.height/fig.width	numeric	7, 7	Height and width of figure
14	fig.show	asis, hold, animate, hide	asis	How the figure should be displayed
15	interval	numeric	1	Interval (speed) When fig.show = 'animate'

For complete documentation, see http://yihui.name/knitr/options/

echo and eval

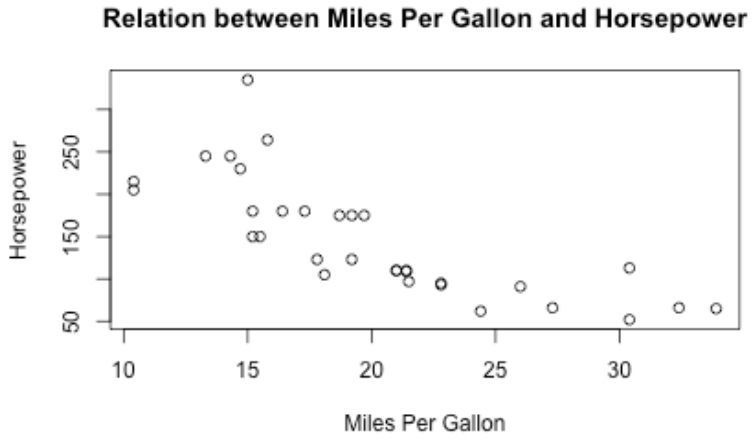
You can show code without evaluating it, using `eval = FALSE`.

````
```{r ex_rCalc2, eval = FALSE}
a + b * (exp(a)/b)
```
````

```
a + b * (exp(a)/b)
```

Alternatively, you can evaluate the code without displaying it, using `echo = FALSE`.

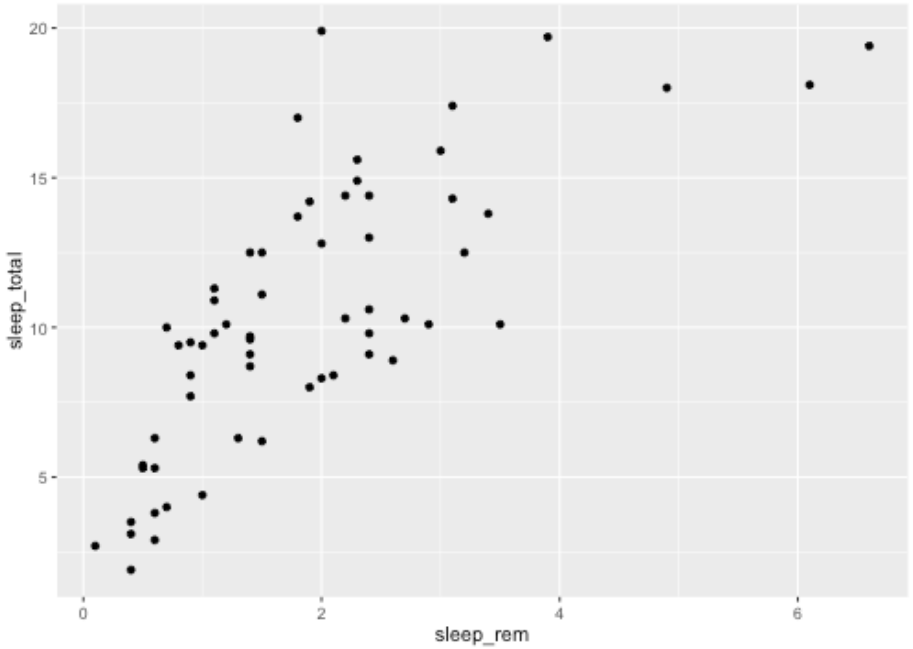
````
```{r plotExample, echo = FALSE, fig.width = 6, fig.height = 3.8}
data(mtcars)
with(mtcars, plot(mpg, hp,
  xlab = "Miles Per Gallon",
  ylab = "Horsepower",
  main = "Relation between Miles Per Gallon and Horsepower"))
```
````



warning

Warning = FALSE

```
ggplot(msleep,
  aes(sleep_rem, sleep_total)) +
  geom_point()
```

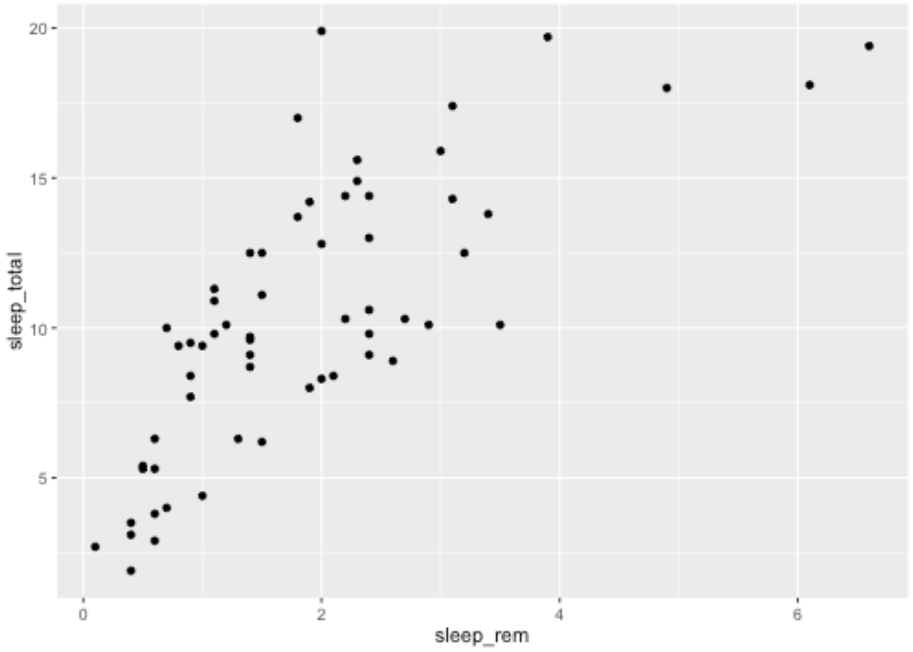


Warning is printed to the console when rendering.

Warning = TRUE

```
ggplot(msleep,
  aes(sleep_rem, sleep_total)) +
  geom_point()
```

```
## Warning: Removed 22 rows containing missing
```



Show errors

Default

```
ggplot(msleep,
  aes(sleep, sleep_total)) +
  geom_point()
```

```
## Don't know how to automatically pick scale for object of type data.frame. Defaulting to cor
```

```
## Error: Aesthetics must be either length 1 or the same as the data (83): x, y
```

If `error = FALSE`, the document won't render if it encounters an error.

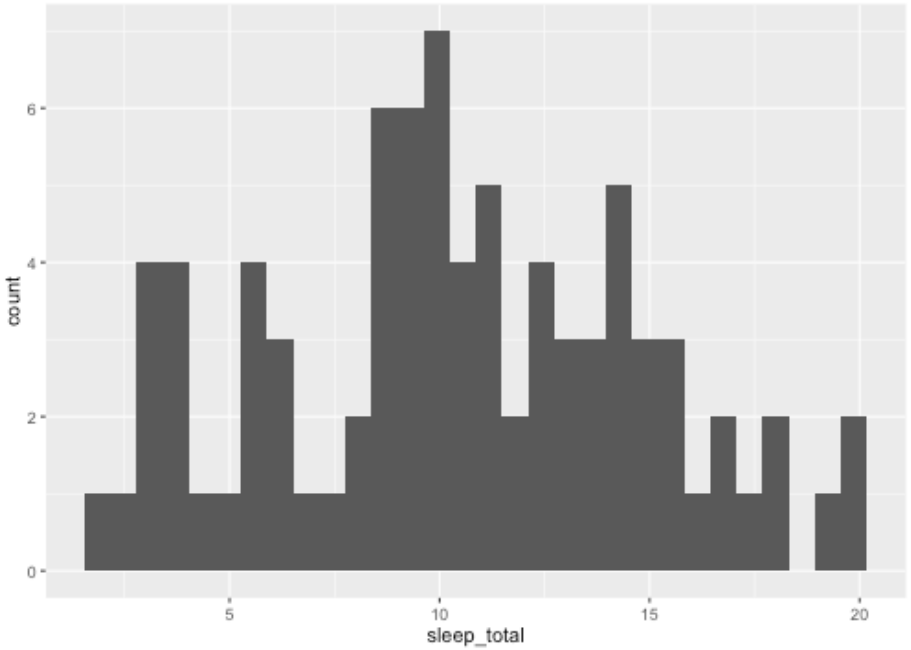
```
  |.....................................                | 67%
  |......................................               | 71%
  |.......................................              | 76%
  |........................................             | 81%
  |.........................................            | 86%
label: showErrors (with options)
List of 1
 $ error: logi FALSE

Quitting from lines 300-303 (dynamicDocuments.Rmd)
Error: Aesthetics must be either length 1 or the same as the data (83): x, y

> |
```

Message

Some functions will return messages. You may want to suppress these.

message = FALSE

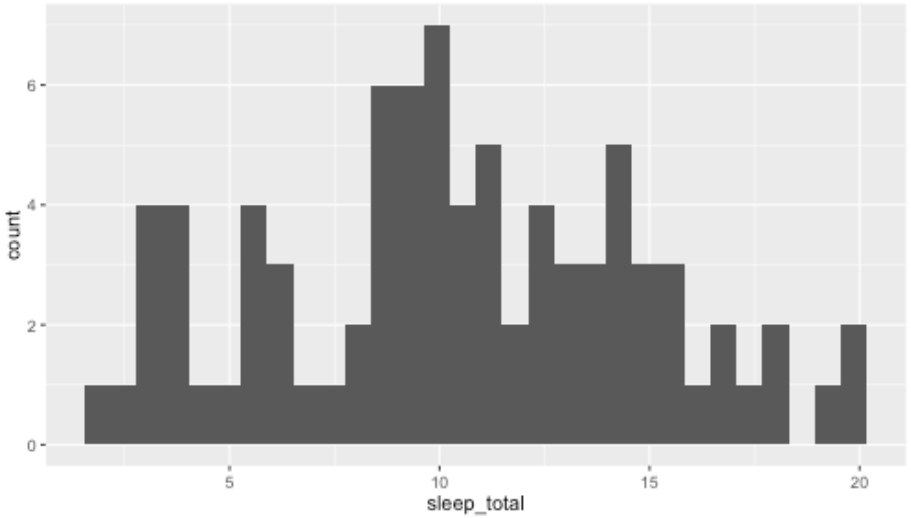
```
ggplot(msleep,
  aes(sleep_total)) +
  geom_histogram()
```



message = TRUE

```
ggplot(msleep,
  aes(sleep_total)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better v
```



tidy

Tidy = FALSE

```
matRow<-matrix(c(10,11,12,13,20,21,22,23,
  30,31,32,33),nrow=3,ncol=4,byrow=TRUE)

matRow<-matrix(c(10,11,12,13,
                 20,21,22,23,
                 30,31,32,33),
nrow=3,ncol=4,byrow=TRUE)
```

Tidy = TRUE

```
matRow <- matrix(c(10, 11, 12, 13, 20, 21, 22, 
    ncol = 4, byrow = TRUE)

matRow <- matrix(c(10, 11, 12, 13, 20, 21, 22, 
    ncol = 4, byrow = TRUE)
```

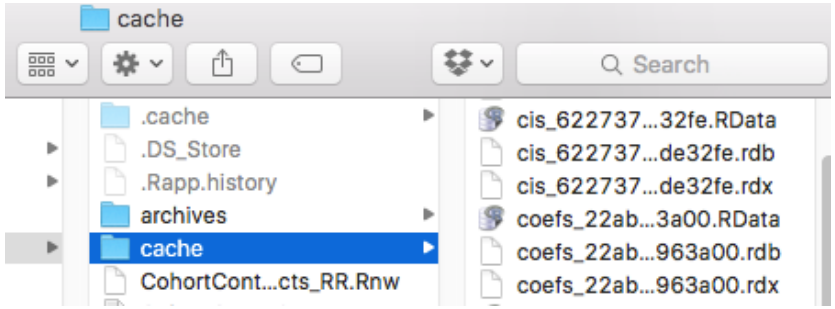
(It can only do so much, and sometimes ends up looking worse. Follow a style!)

cache and dependencies

Somewhat complicated

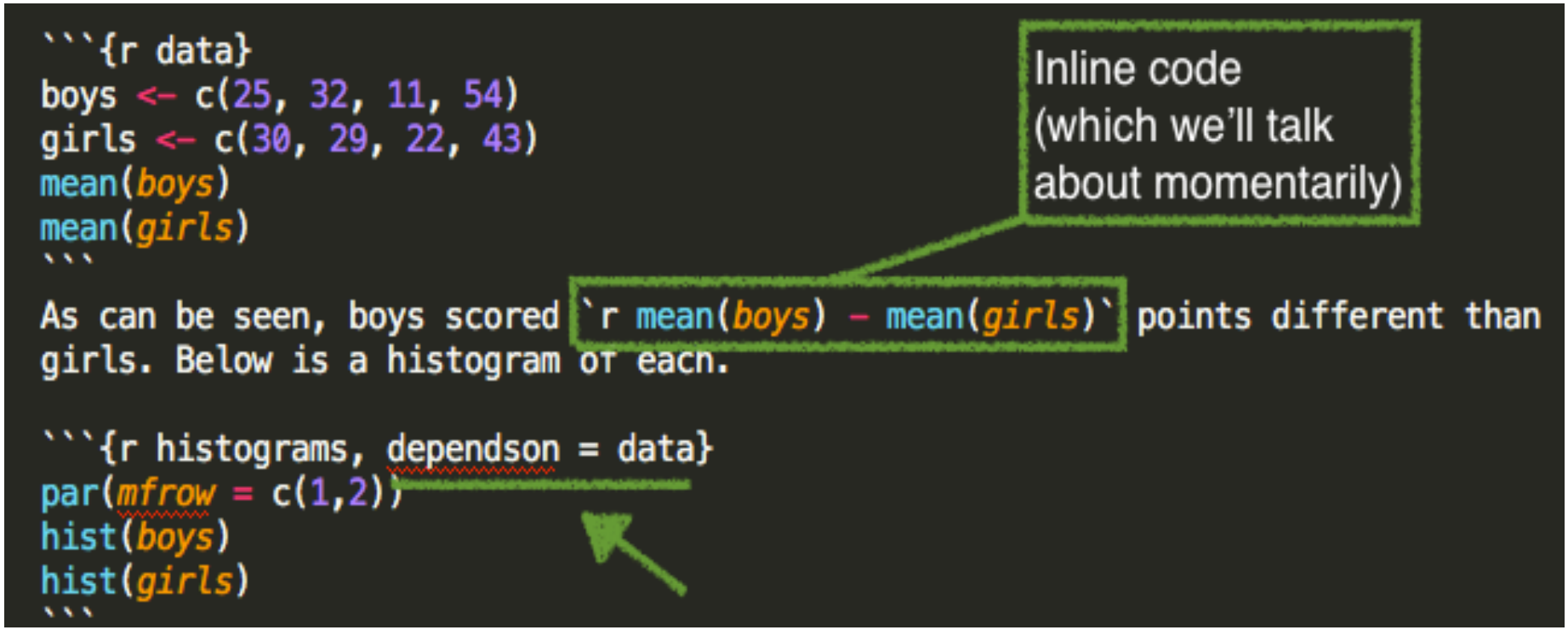
- When chunks take a long time to process, it is usually a good idea to *cache* them.
  - Create temporary files with the results of the chunk
  - These files are then called when the document is rendered, and need not be re-run, provided **nothing in the chunk has changed**.
  - Often, chunks may depend on the results of previous chunks. These are *dependencies*. All dependencies must then be updated, if a chunk with dependencies is updated.
  - You can declare dependencies manually or automatically

cache



- Cache files are named according to your chunk names (why it's important to name your chunks)
- Note that some packages that use *knitr* (i.e., *slidify*, which was used to produce these slides), will cache for you automatically. And the slidify cache is in a hidden folder (which can be really annoying)

Declaring dependencies manually



```
boys <- c(25, 32, 11, 54)
girls <- c(30, 29, 22, 43)
mean(boys)
```

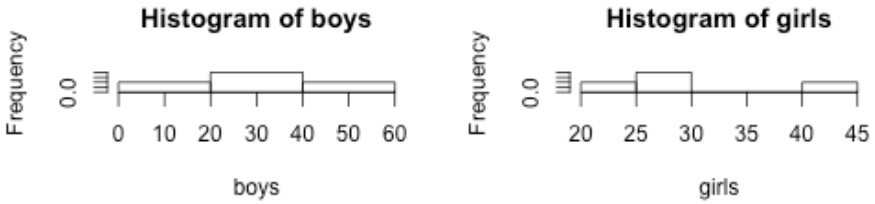
```
## [1] 30.5
```

```
mean(girls)
```

```
## [1] 31
```

As can be seen, boys scored -0.5 points different than girls. Below is a histogram of each.

```
par(mfrow = c(1,2))
hist(boys)
hist(girls)
```



Setting global options

In other words, change the default behavior

```
opts_chunk$set(options)
```

For example, you can setup all chunks to be cached, and for the dependencies to be automatically determined, with the following code:

```
opts_chunk$set(cache = TRUE, autodep = TRUE)
```

```
dep_auto()
```

Note that `dep_auto()` is a function that must be run on its own (which finds the dependencies).

In other cases you may wan to suppress all the code. For example, when preparing a report for somebody.

```
opts_chunk$set(echo = FALSE)
```

You can always override the defaults (global options) within a particular chunk, e.g.
```{r, chunkName, echo = TRUE}

```

include

````
```{r setup, include = FALSE}
library(knitr)

# Set global chunk options
opts_chunk$set(cache = TRUE, cache.comments = FALSE, autodep = T

dep_auto()
```
````

The `include` argument is used to evaluate code that is not included in the document at all. For example, when setting up your global options.

tables (very briefly)

- Packages that can produce tables for R markdown (in order from least to most flexible)
  - *knitr*
  - *pander*
  - *xtable*

Displaying tables

Change the `results` chunk option to "asis"

knitr::kable

For very simple tables, use `kable` from the *knitr* package

```
id <- rep(1:3, each = 2)
condition <- rep(c("A", "B"), 3)
score <- rnorm(6, 10, 3)
data <- data.frame(id, condition, score)

library(knitr)
kable(data)
```

ID	CONDITION	SCORE
1	A	8.941231
1	B	9.424647
2	A	9.292759
2	B	9.558533
3	A	6.450977
3	B	8.117230

pander

- Great for producing summary tables.
  - Must specify `style = "rmarkdown"`
- Doesn't seem to work well with *slidify* (not sure why).
- Hopefully we'll have time to look at this a bit with the example.

```
library(pander)
pander(lm(Sepal.Width ~ Species, data = iris),
    covariate.labels = c("Versicolor" ,  "Virginica" ),
    style = "rmarkdown")
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |:---|:---:|:---:|:---:|:---:| | **Versicolor** | -0.658 | 0.06794 | -9.685 | 1.832e-17 | | **Virginica** | -0.454 | 0.06794 | -6.683 | 4.539e-10 | | **(Intercept)** | 3.428 | 0.04804 | 71.36 | 5.708e-116 |

Table: Fitting linear model: Sepal.Width ~ Species

xtable

For `xtable`, you have to make sure you specify `results = "asis"`.

If you're in a markup environment (what we've been talking about), you have to also make sure you specify `type = "html"`.

```
library(xtable)
mat <- round(matrix(c(0.9, 0.89, 200, 0.045, 2.0), c(1, 5)), 4)
rownames(mat) <- "$y_{t-1}$"
colnames(mat) <- c("$R^2$", "$\\bar{x}$", "F-stat", "S.E.E", "DW")
mat <- xtable(mat)
print(mat,
  sanitize.text.function = function(x) {x},
  type = "html")
```

	R^2	$\bar{x}$	F-STAT	S.E.E	DW
y_{t-1}	0.90	0.89	200.00	0.04	2.00

Same example, but without specifying `results = "asis"`

```
print(mat,
  sanitize.text.function = function(x) {x},
  type = "html")
```

```
## <!-- html table generated in R 3.3.0 by xtable 1.8-2 package -->
## <!-- Fri Jun 17 20:37:33 2016 -->
## <table border=1>
## <tr> <th>  </th> <th> $R^2$ </th> <th> $\bar{x}$ </th> <th> F-stat </th> <th> S.E.E </th> <
##   <tr> <td align="right"> $y_{t-1}$ </td> <td align="right"> 0.90 </td> <td align="right">
##    </table>
```

Other options for results: `hold`

```
m1 <- lm(mpg ~ ., data = mtcars)
coef(m1)
coef(summary(m1))[, "Std. Error"]
arm::display(m1)
```

```
## (Intercept)         cyl        disp          hp        drat          wt
## 12.30337416 -0.11144048  0.01333524 -0.02148212  0.78711097 -3.71530393
##        qsec          vs          am        gear        carb
##  0.82104075  0.31776281  2.52022689  0.65541302 -0.19941925
## (Intercept)         cyl        disp          hp        drat          wt
## 18.71788443  1.04502336  0.01785750  0.02176858  1.63537307  1.89441430
##        qsec          vs          am        gear        carb
##  0.73084480  2.10450861  2.05665055  1.49325996  0.82875250
## lm(formula = mpg ~ ., data = mtcars)
##             coef.est coef.se
## (Intercept) 12.30    18.72
## cyl         -0.11     1.05
## disp         0.01     0.02
## hp          -0.02     0.02
## drat         0.79     1.64
## wt          -3.72     1.89
## qsec         0.82     0.73
## vs           0.32     2.10
```

Same chunk, no hold

```
m1 <- lm(mpg ~ ., data = mtcars)
coef(m1)
```

```
## (Intercept)         cyl        disp          hp        drat          wt
## 12.30337416 -0.11144048  0.01333524 -0.02148212  0.78711097 -3.71530393
##        qsec          vs          am        gear        carb
##  0.82104075  0.31776281  2.52022689  0.65541302 -0.19941925
```

```
coef(summary(m1))[, "Std. Error"]
```

```
## (Intercept)         cyl        disp          hp        drat          wt
## 18.71788443  1.04502336  0.01785750  0.02176858  1.63537307  1.89441430
##        qsec          vs          am        gear        carb
##  0.73084480  2.10450861  2.05665055  1.49325996  0.82875250
```

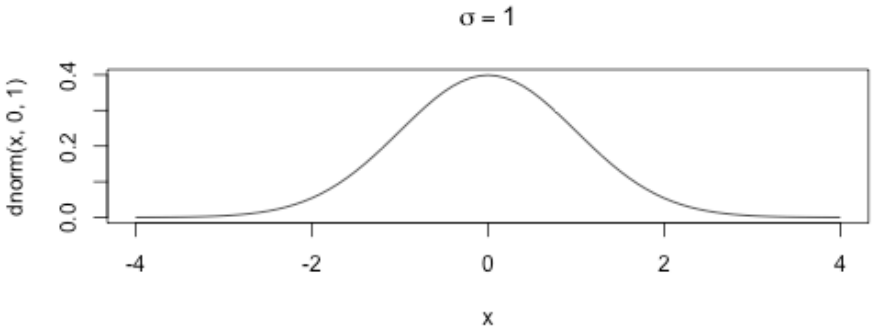
```
arm::display(m1)
```

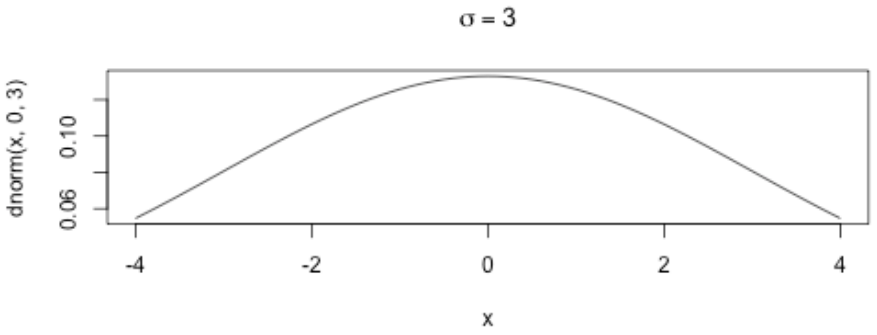
```
## lm(formula = mpg ~ ., data = mtcars)
##             coef.est coef.se
## (Intercept) 12.30    18.72
## cyl         -0.11     1.05
```

Hold figures

`fig.show = "hold"`

```
x <- seq(-4, 4, 0.1)
plot(x, dnorm(x, 0, 1), type = "l", main = expression(sigma == 1))
plot(x, dnorm(x, 0, 3), type = "l", main = expression(sigma == 3))
```





Inline code

A single back tick followed by `r` prooduces inline code to be evaluated.

```
This is an example of inline code, where I want to refer to the sum of `a` and
  `b`, which is `r a + b`.
```

This is an example of inline code, where I want to refer to the sum of `a` and `b`, which is 8.

This is *extremely* useful in writing reports. Never have to update any numbers in text, regardless of changes to your models or data (if you are careful about it).

Citations (quickly)

To include references in your paper, you must:

- Create an external .bib file using LaTeX formatting (we'll get to this)
- Include `bibliography: nameOfYourBibFile.bib` in your YAML front matter.
- Refer to the citations in text using `@`

Creating a .bib doc

The persistence of school-level value-added

DC Briggs, JP Weeks - Journal of Educational and Behavioral ..., 2011 - jeb.sagepub.com

Abstract Using longitudinal data for an entire state from 2004 to 2008, this article describes the results from an empirical investigation of the persistence of value-added school effects on student achievement in reading and math. It shows that when schools are the principal ...

Cited by 20 Related articles All 8 versions Web of Science: 4 Cite Save More

- Cite

Copy and paste a formatted citation or use one of the links to import into a bibliography manager.

MLA Briggs, Derek C., and Jonathan P. Weeks. "The persistence of school-lev value-added." *Journal of Educational and Behavioral Statistics* 36.5 (2011 616-637.

APA Briggs, D. C., & Weeks, J. P. (2011). The persistence of school-level valu added. *Journal of Educational and Behavioral Statistics*, *36*(5), 616-637.

Chicago Briggs, Derek C., and Jonathan P. Weeks. "The persistence of school-lev value-added." *Journal of Educational and Behavioral Statistics* 36, no. 5 (2011): 616-637.

Harvard Briggs, D.C. and Weeks, J.P., 2011. The persistence of school-level value-added. *Journal of Educational and Behavioral Statistics*, *36*(5), pp.616-637.

Vancouver Briggs DC, Weeks JP. The persistence of school-level value-added. Journal of Educational and Behavioral Statistics. 2011 Oct 1;36(5):616-37.

BibTeX EndNote RefMan RefWorks

```
@article{briggs2011persistence,
  title={The persistence of school-level value-added},
  author={Briggs, Derek C and Weeks, Jonathan P},
  journal={Journal of Educational and Behavioral Statistics},
  volume={36},
  number={5},
  pages={616--637},
  year={2011},
  publisher={SAGE Publications}
}
```

```
@article{Briggs11,
  title={The persistence of school-level value-added},
  author={Briggs, Derek C and Weeks, Jonathan P},
  journal={Journal of Educational and Behavioral Statistics}
  volume={36},
  number={5},
  pages={616--637},
  year={2011},
  publisher={SAGE Publications}
}
```

Tag for in-text referencing

In text citations

CITATION STYLE	OUTPUT
@Briggs11	Briggs and Weeks (2011)
[see @Baldwin2014; @Caruso2000]	(see Baldwin et al. 2014; Caruso 2000)
[@Linn02, p. 9]	(Linn and Haug 2002, 9)
[-@Goldhaber08]	(2008)

Note this is not APA. However, references are included automatically at the end of the document. Include `# References` as the last line of your document to give it a title.

References

References

Baldwin, Scott A, Zac E Imel, Scott R Braithwaite, and David C Atkins. 2014. "Analyzing Multiple Outcomes in Clinical Research Using Multivariate Multilevel Models." *Journal of Consulting and Clinical Psychology* 82 (5). American Psychological Association: 920.

Briggs, Derek C, and Jonathan P Weeks. 2011. "The Persistence of School-Level Value-Added." *Journal of Educational and Behavioral Statistics* 36 (5). SAGE Publications: 616–37.

Caruso, John C. 2000. "Reliability Generalization of the NEO Personality Scales." *Educational and Psychological Measurement* 60 (2). Sage Publications: 236–54.

Goldhaber, D., and M. Hansen. 2008. "Is It Just a Bad Class? Assessing the Stability of Measured Teacher Performance. CPRE Working Paper No. 2008-5, University of Washington." Report.

Linn, R. L., and C. Haug. 2002. "Stability of School-Building Accountability Scores and Gains." Journal Article. *Educational Evaluation and Policy Analysis* 24: 29–36. doi:10.3102/01623737024001029.

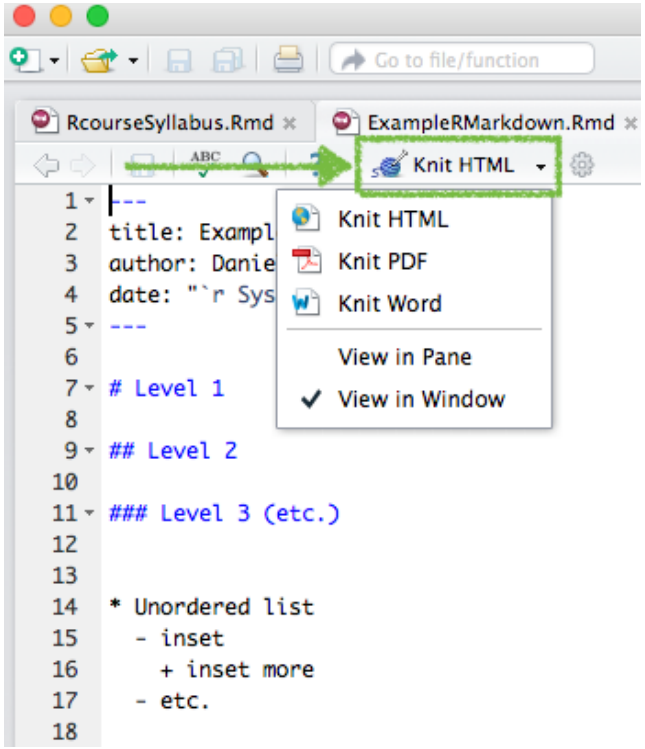
Rendering the document

Through a text editor (e.g., SublimeText)

```
install.packages("rmarkdown")
library(rmarkdown)
setwd("dir/to/Rmd/doc")
render("ExampleRMarkdown.Rmd",
  "html_document")
```

Note that the document type need not be specified if `output:` is supplied in the YAML front matter.

Through RStudio



Final Product!

ExampleRMarkdown.Rmd — UNREGISTERED

ExampleRMarkdown.Rmd

```
---
title: Example Markdown document
author: Daniel Anderson
date: "`r Sys.Date()`"
output: html_document
---

```{r setup, include = FALSE}
library(knitr)

# Set global chunk options
opts_chunk$set(cache = TRUE, cache.comments = FALSE, autodep = TRUE)

# Determine caching dependencies automatically
dep_auto()
```

Level 1

Level 2

Level 3 (etc.)

* Unordered list
  - inset
    + inset more
  - etc.

1. Ordered list
   a. blah blah
2. More stuff

```{r ex_rCalc1}
a <- 3
b <- 5

a + b * (exp(a)/b)
```

This is an example of inline code, where I want to refer to the sum of `a` and
  `b`, which is `r a + b`.
```

117 Words, git branch: master, index: 1?, working: 1?, Line 1, Column 1 | 12 misspelled words | Tab Size: 4 | R Markdown

file:///Users/Daniel/Dropbox/Teaching/Rcourse/Week1/ExampleR

irst Row Apple Gmail MidMathDev UO Mail Facebook espn Duckweb Blackboard Wells Farg

r - How to add whitespace to an RMarkdown docume... Exampl

Example Markdown document

Daniel Anderson

2015-09-17

Level 1

Level 2

Level 3 (etc.)

- Unordered list
- inset
  - inset more
- etc.

1. Ordered list

a. blah blah

2. More stuff

```
a <- 3
b <- 5

a + b * (exp(a)/b)
```

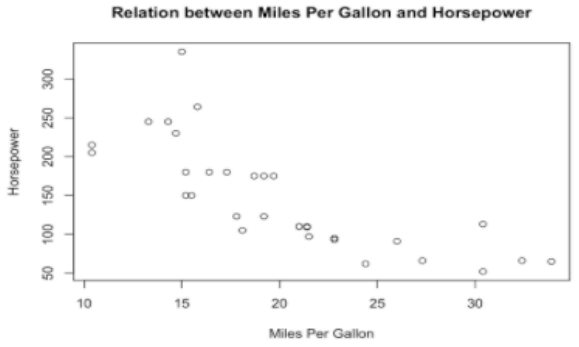
```
## [1] 23.08554
```

This is an example of inline code, where I want to refer to the sum of `a` and `b`, which is 8.

You can show code without evaluating it, using `eval = FALSE`.

```
a + b * (exp(a)/b)
```

Alternatively, you can evaluate the code without displaying it



A few complications

If you use *RStudio*, you should be able to render HTML output automatically with the `knit2html` button.

However, if you use a text editor (like I do), then you'll need to also install *pandoc* (http://pandoc.org).

PDF output

Regardless of whether you use RStudio or not, you will also need to install a TeX distribution.

- Macs: MacTeX (http://tug.org/mactex/)

- Windows: MikTeX (http://miktex.org)

Summarizing

- R Markdown is relatively simple and easy to learn.
- Tables are probably the most difficult piece.
- Lots of options to get it to do what you want.
- Great for sharing and documenting your work.

but...

- The more you ask from it, the more difficult it will become.
- At a certain point, you may need more flexibility.

Writing a paper in apa style

- Stick with R Markdown and try the *papaja* package (https://github.com/crsh/papaja)
    - I have little to no experience with it

```
install.packages("devtools")
library(devtools)
install_github("crsh/papaja")
```

- Go with the more advanced *.RNW* format (versus *.Rmd*)
    - Essentially you build your paper with LaTeX, embedding code through the *knitr* package

```
mobile <- function(d) {
  moved <- sapply(
        split(mth[ ,c("ScID_Pred", "ScID_Out")], mth$combo),
          function(x) table(x$ScID_Pred != x$ScID_Out)
          )
  round((moved / colSums(moved)) * 100)
}

## Explore variability by cohort
# sapply(split(mth, mth$cohort), mobile)
# sapply(split(rdg, rdg$cohort), mobile)

mobileM <- mobile(mth)
mobileR <- mobile(rdg)

@
```

```
Operational statewide accountability data from one state located in the Pacific
  Northwest were used, collected across the 2007-08 to 2011-12 school years.
  Three complete cohorts of students were matched longitudinally across Grades
  3-5. These cohorts are referred to throughout as the `08, `09, and `10
  Cohorts, corresponding to the year in which each cohort completed third
  grade. Approximately \Sexpr{cohSize} students were represented in each cohort
  across the \Sexpr{nSch} schools in the sample. The mean sample size per
  school was approximately \Sexpr{nByJ_Mean} with a standard deviation of
  \Sexpr{nByJ_SD}. Sample means and standard deviations are displayed by cohort
  and for the overall sample in Table \ref{tab:demos}. Overall, approximately
  \Sexpr{percDems["nonWhite"]}\% of the sample was non-White,
  \Sexpr{percDems["SWD"]}\% had a documented disability, and
  \Sexpr{percDems["FRL"]}\% were eligible for free or reduced price lunch.
  Approximately \Sexpr{mobileM[2, 1]}\% of the sample moved schools between
  Grades 3 and 4, while \Sexpr{mobileM[2, 3]}\% moved between Grades 4 and 5,
  and \Sexpr{mobileM[2, 2]}\% moved between Grades 3 and 5.
```

```
<<demos>>=
dems[[1]]$Subject <- rep("Math", nrow(dems[[1]]))
dems[[2]]$Subject <- rep("Rdg", nrow(dems[[2]]))

desc <- function(dem, sub) {
  mns <- tapply(dem$RIT, list(dem$cohort, dem$Grade),
        mean, na.rm = TRUE)
  overall_mn <- tapply(dem$RIT, dem$Grade, mean, na.rm = TRUE)
```

2. What proportion of the variance in students' scores is attributable to *school*, *cohort*, or *content* facets?

3. How does the number of cohorts modeled impact the reliability of school effect estimates?

Method

Sample and Data

Operational statewide accountability data from one state located in the Pacific Northwest were used, collected across the 2007-08 to 2011-12 school years. Three complete cohorts of students were matched longitudinally across Grades 3-5. These cohorts are referred to throughout as the '08, '09, and '10 Cohorts, corresponding to the year in which each cohort completed third grade. Approximately 27,000 students were represented in each cohort across the 727 schools in the sample. The mean sample size per school was approximately 122 with a standard deviation of 95. Sample means and standard deviations are displayed by cohort and for the overall sample in Table 1. Overall, approximately 35% of the sample was non-White, 12% had a documented disability, and 50% were eligible for free or reduced price lunch. Approximately 10% of the sample moved schools between Grades 3 and 4, while 9% moved between Grades 4 and 5, and 17% moved between Grades 3 and 5.

Measures

Final remarks on R Markdown

- Make sure to look at the documentation
    - http://RMarkdown.rstudio.com
    - http://RMarkdown.rstudio.com/authoring_basics.html
    - http://RMarkdown.rstudio.com/authoring_rcodechunks.html
- The more you ask from it, the more complicated it becomes.
- Challenges
    - Word is the industry standard (frustratingly so, to me)
        - Word output is less than ideal
    - Can be difficult when collaborating with others
    - Some journal articles *require* papers submitted in Word
        - Potentially get a pdf to word converter, but still less than ideal
    - Advanced features have a relatively steep learning curve

Take home message

- It's a fairly big challenge to start to write *papers* using this method
- Fairly straightforward as a method to produce reports/keep track of your analysis
- Start small and work your way up; don't get discouraged too easily

I'm still actively learning this whole process. I recommend Yihui's book, it's quite good.

Let's practice!
(if we have time)

Dr. Rosenberg Presentation

Outline

1. Background
2. Wrangling, Plotting, and Modeling
3. Essential Functionality
4. Advanced Functionality
5. Additional Resources

Part I: Background

Why use R: Accessibility

-
  - A script documents all your work, from data access to reporting, and can instantly be re-run at any time
-
  - As an open-source project, you can use R free of charge: no worries about subscription fees, license managers, or user limits.
-
  - All of the standard data analysis tools are built right into the R language (and many others are available via “packages”)
-
  - One of the design principles of R was that visualization of data through charts and graphs is an essential part of the data analysis process, so it has excellent tools for creating graphics

Why use R: Community

-
  - Leading academics and researches from around the world use R to develop the latest methods in statistics, machine learning, and predictive modeling
-
  - There's a wealth of community resources for R available on the Web, for help in just about every domain
-
  - Available Linux, Mac, and Windows
-
  - R users come from myriad academic departments and industries

Part II: Wrangling, Plotting, and Modeling

Focus on data frames (and tidyverse)

```r
# install.packages("tidyverse")
library(tidyverse)

df <- nycflights13::flights
```

Explore your data: Descriptive statistics

```
df_ss <- select(df, dep_delay, arr_delay, air_time, distance,
carrier)
psych::describe(df_ss)
```

```
          vars      n    mean     sd median trimmed    mad min  max
range
dep_delay    1 328521   12.64  40.21     -2    3.32   5.93 -43 1301
1344
arr_delay    2 327346    6.90  44.63     -5   -1.03  20.76 -86 1272
1358
air_time     3 327346  150.69  93.69    129  140.03  75.61  20  695
675
distance     4 336776 1039.91 733.23    872  955.27 569.32  17 4983
4966
carrier*     5 336776    9.00   0.00      9    9.00   0.00   9    9
0
          skew kurtosis   se
dep_delay 4.80    43.95 0.07
arr_delay 3.72    29.23 0.08
air_time  1.07     0.86 0.16
distance  1.13     1.19 1.26
carrier*   NaN      NaN 0.00
```

What's going on here?

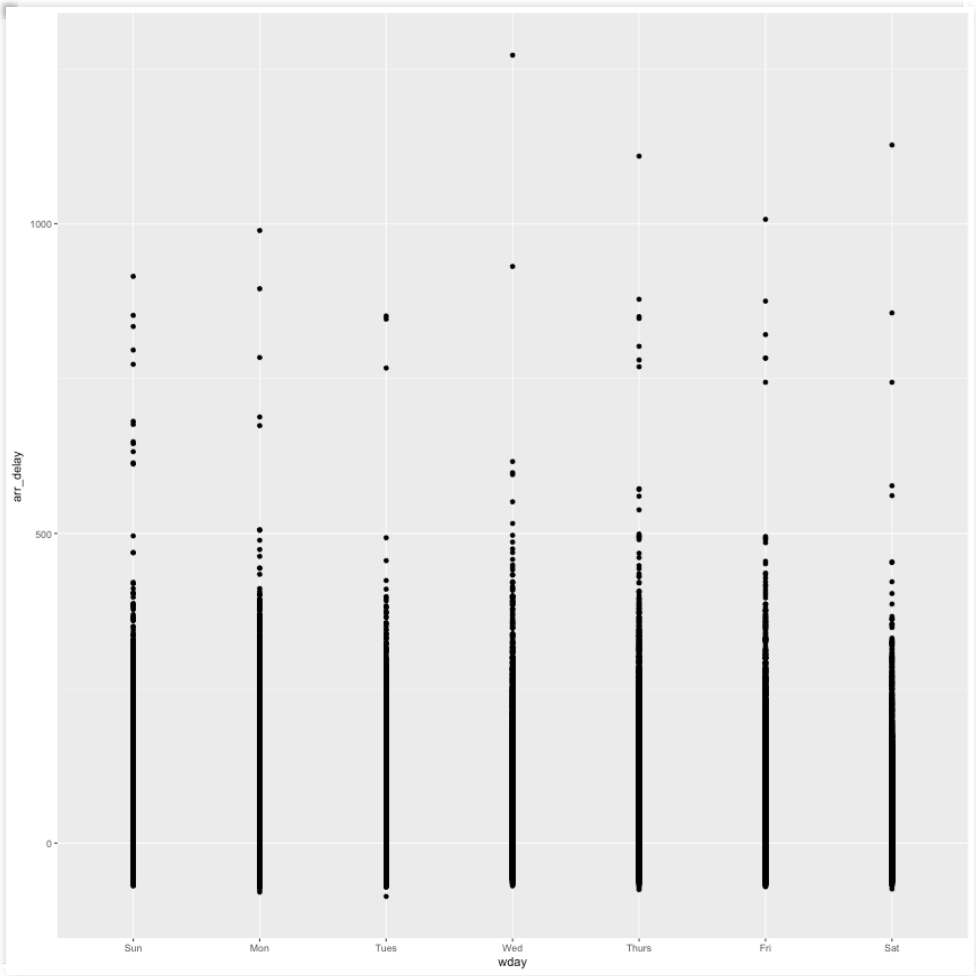
```
df %>%
    select(dep_delay, arr_delay, air_time, distance, carrier) %>%
    group_by(carrier) %>%
    summarize(dep_delay_mean = mean(dep_delay, na.rm = T))
```

```
df %>%
    select(dep_delay, arr_delay, air_time, distance, carrier) %>%
    group_by(carrier) %>%
    summarize(dep_delay_mean = mean(dep_delay, na.rm = T))
```

```
# A tibble: 16 × 2
   carrier dep_delay_mean
     <chr>          <dbl>
1       9E      16.725769
2       AA       8.586016
3       AS       5.804775
4       B6      13.022522
5       DL       9.264505
6       EV      19.955390
7       F9      20.215543
8       FL      18.726075
9       HA       4.900585
10      MQ      10.552041
11      OO      12.586207
12      UA      12.106073
13      US       3.782418
14      VX      12.869421
15      WN      17.711744
16      YV      18.996330
```

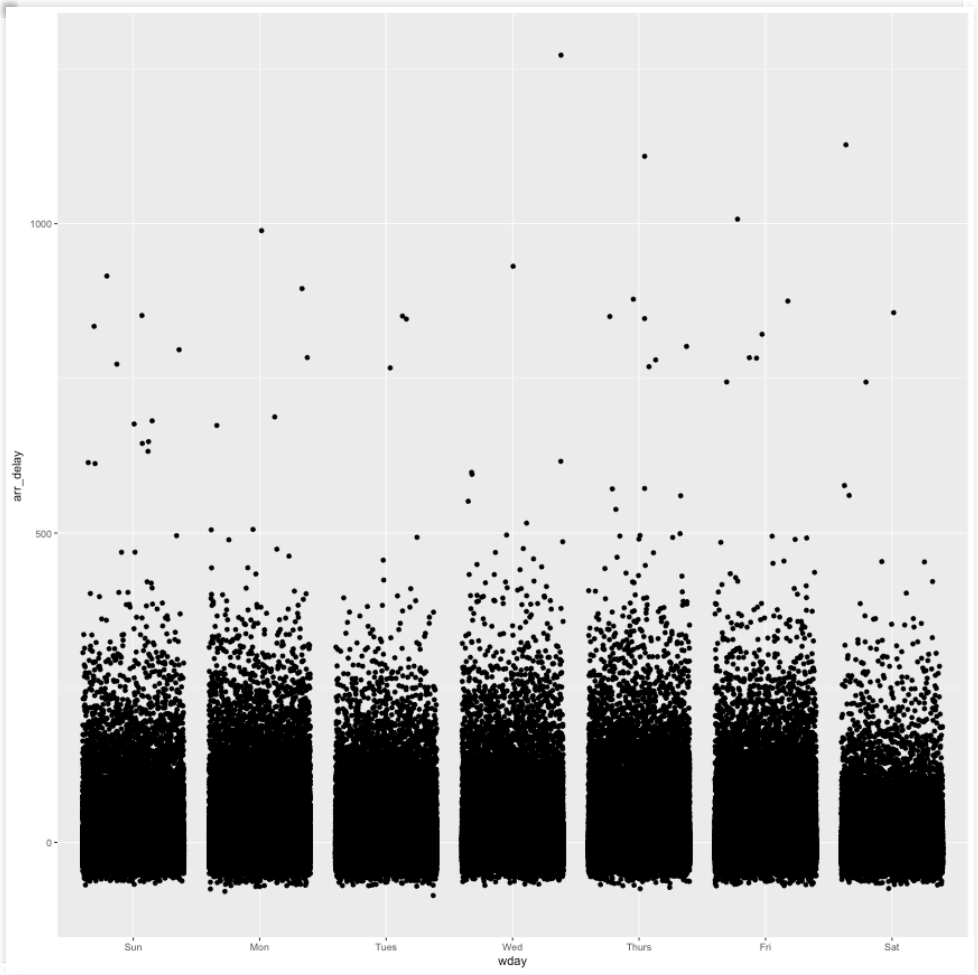
Explore your data: Plotting Distributions

```
df$wday <- lubridate::wday(df$time_hour, label = T)

ggplot(df, aes(x = wday, y = arr_delay)) +
    geom_point()
```



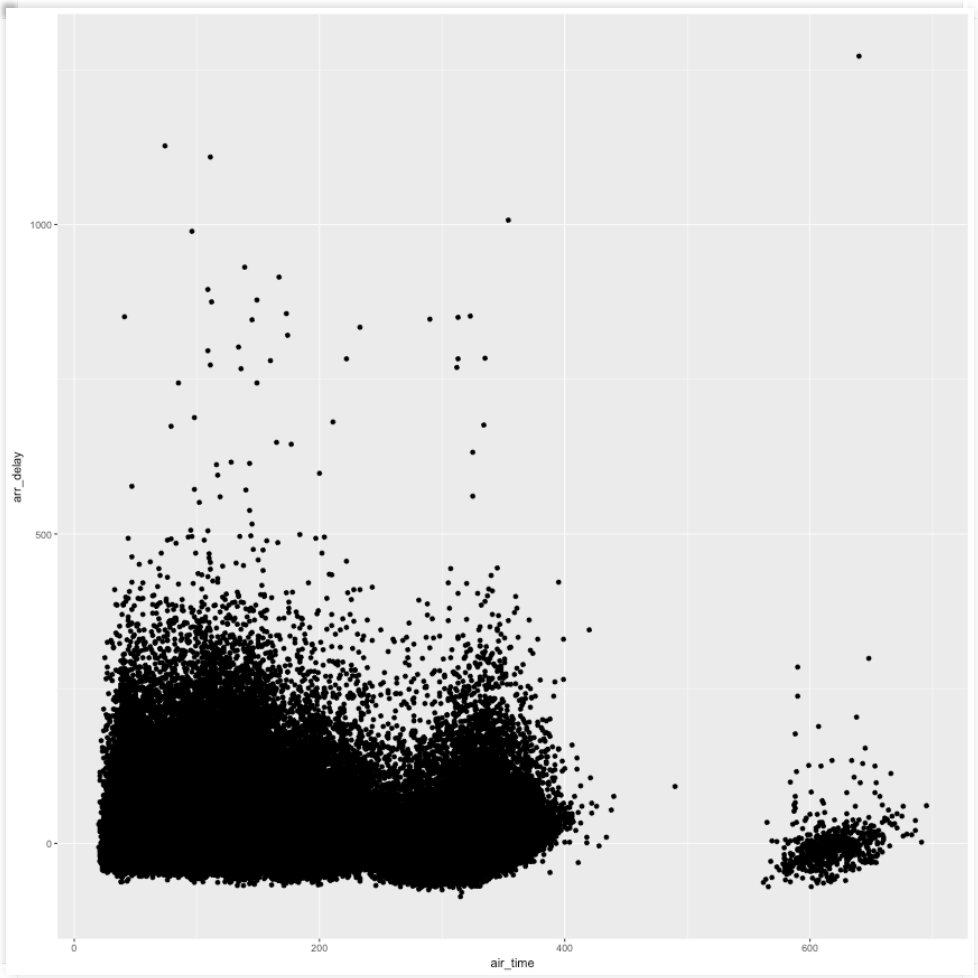
Explore your data: Plotting Distributions (With Some Random Noise)

```
library(ggplot2)
ggplot(df, aes(x = wday, y = arr_delay)) +
    geom_jitter()
```



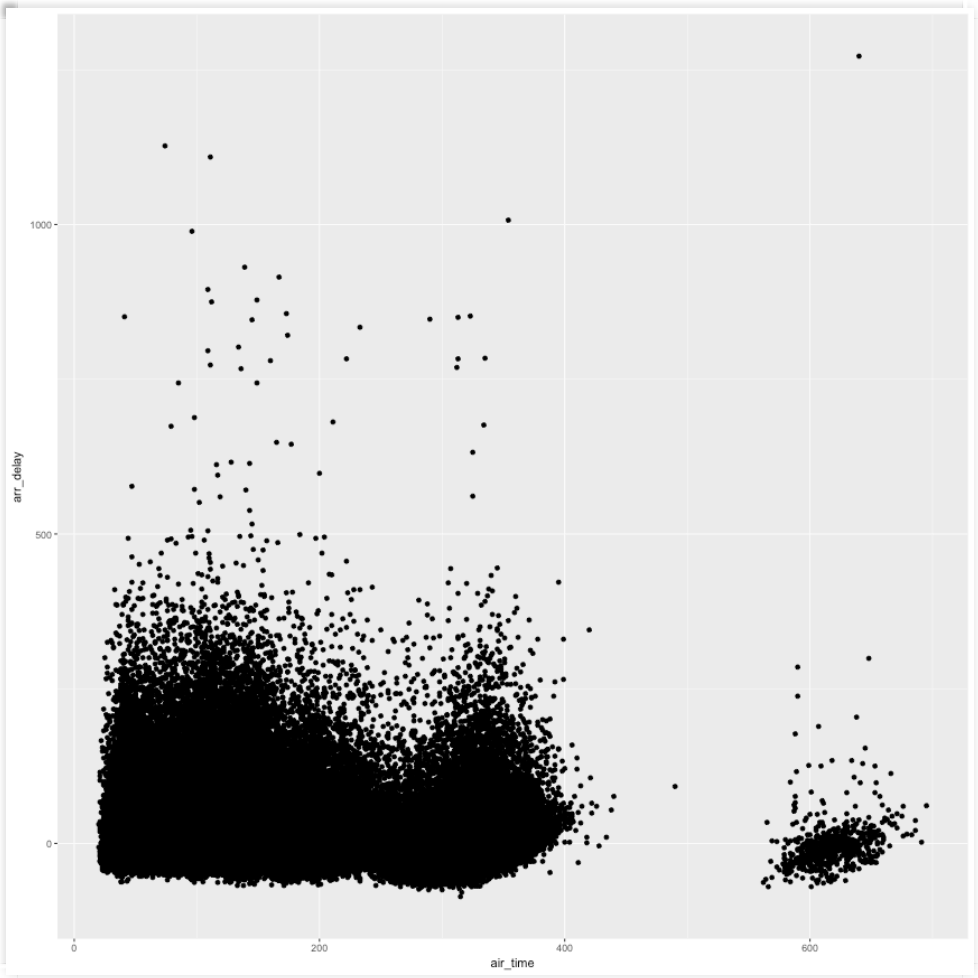
Explore your data: Plotting relationships

```
ggplot(df, aes(x = air_time, y = arr_delay)) +
    geom_point()
```



Explore your data: Plotting relationships

```
ggplot(df, aes(x = air_time, y = arr_delay)) +
    geom_point()
```



Explore your data: Plotting relationships

```
ggplot(df_ss, aes(x = air_time, y = arr_delay, color = carrier)) +
    geom_point()
```

Explore your data: Manipulate data (for plots)

```
to_plot <- df %>%
    group_by(carrier) %>%
    summarize(dep_delay_mean = mean(dep_delay, na.rm = T),
              n = n()) %>%
    filter(n > 10000) %>%
    arrange(desc(dep_delay_mean))
```

- EV: Express Jet
- WN: Southwest Airlines
- AA: American Airlines
- US: US Airways

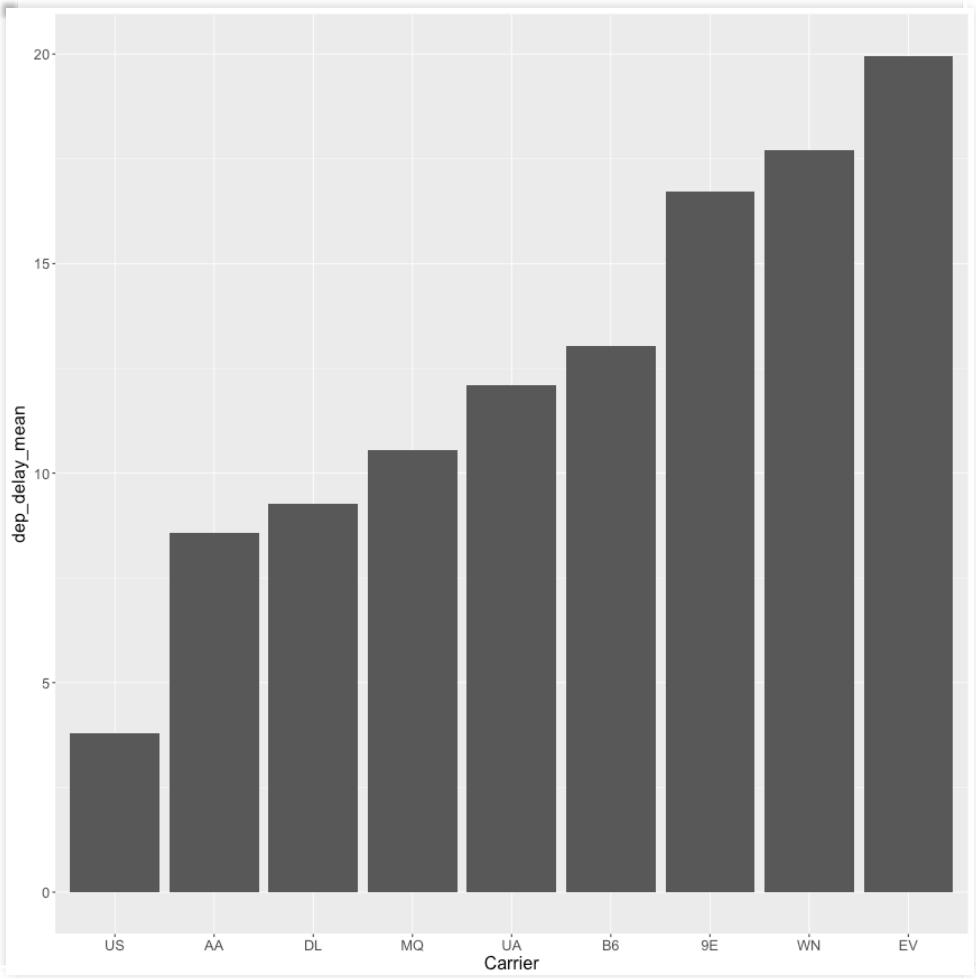
Explore your data: Manipulate data (for plots)

```
to_plot
```

```
# A tibble: 9 × 3
  carrier dep_delay_mean     n
    <chr>          <dbl> <int>
1      EV      19.955390 54173
2      WN      17.711744 12275
3      9E      16.725769 18460
4      B6      13.022522 54635
5      UA      12.106073 58665
6      MQ      10.552041 26397
7      DL       9.264505 48110
8      AA       8.586016 32729
9      US       3.782418 20536
```

Explore your data: Plotting Means

```
ggplot(to_plot, aes(x = reorder(carrier, dep_delay_mean), y =
dep_delay_mean)) +
    geom_col() +
    theme(text = element_text(size = 16)) +
    xlab("Carrier")
```



Model your data: Linear models

```
m1 <- lm(arr_delay ~ air_time, data = df)
arm::display(m1)
```

```
lm(formula = arr_delay ~ air_time, data = df)
            coef.est coef.se
(Intercept)  9.43     0.15
air_time    -0.02     0.00
---
n = 327346, k = 2
residual sd = 44.61, R-Squared = 0.00
```

Model your data: Linear models

```
m2 <- lm(arr_delay ~ air_time + distance, data = df)
arm::display(m2)
```

```
lm(formula = arr_delay ~ air_time + distance, data = df)
            coef.est coef.se
(Intercept) -1.46     0.17
air_time     0.67     0.01
distance    -0.09     0.00
---
n = 327346, k = 3
residual sd = 43.73, R-Squared = 0.04
```

Model your data: Linear models

```
m3 <- lm(arr_delay ~ air_time*distance, data = df)
arm::display(m3)
```

```
lm(formula = arr_delay ~ air_time * distance, data = df)
                  coef.est coef.se
(Intercept)       -0.07     0.24
air_time           0.66     0.01
distance          -0.09     0.00
air_time:distance  0.00     0.00
---
n = 327346, k = 4
residual sd = 43.72, R-Squared = 0.04
```

Model your data: Linear models

```
m3 <- lm(arr_delay ~ air_time*distance + carrier, data = df)
arm::display(m3)
```

Part III: Essential Functionality

Vectors

```
my_vector <- c(1:10)
```

```
my_vector
```

```
 [1]  1  2  3  4  5  6  7  8  9 10
```

```
mean(my_vector)
```

```
[1] 5.5
```

Base R functions

```
- ? # this is to find out what a function does
- str() # this is to find out the 'structure' of anything
- View() # this allows you to view a data frame (think spreadsheet)
or matrix
- class() # this tells you what kind of object this is
```

```
my_data[1, ] # just the first row of data frame
my_data[, 1] # just the first column of data frame
head(my_data) # first six rows of data frame
tail(my_data) # last six rows of data frame
```

Loading data (CSV)

```
setwd("~/documents") # this sets the working directory
my_data <- readr::read_csv("r_introduction_data.csv") # loads a CSV
and saves it to `my_data`
my_data
```

```
# A tibble: 212 × 57
     Int    StudentID Grade   Age Gender ClassTeacher SciTeacher
PrevIQWST
   <int>        <chr> <int> <int>  <int>        <int>      <int>
<int>
1      1 A.J. Miranda     1    11      0            5          2
NA
2      1      Abby B.     0    10      1            0          0
0
3      0       Abby E     0    10      0            1          0
0
4      0       Abi N.     1    11      1            4          2
NA
5      1   Abigail D.     1    11      1            7          3
NA
6      0       Adam F     1    12      0            5          2
NA
7     NA      Adam L.     1    11      0            4          2
NA
8      0      Adam T.     0     9      0            1          0
0
9      1   Addison D.     1    11      1            6          3
NA
10     1    Adelle S.     1    11      1            4          2
NA
```

```
# ... with 202 more rows, and 49 more variables: PreEff1 <int>,
#   PreEff2 <int>, PreEff3 <int>, PreEff4 <int>, PreEff5 <int>,
#   PreInt1 <int>, PreInt2Rev <int>, PreInt3 <int>, PreInt4 <int>,
#   PreInt5 <int>, PreVal1 <int>, PreVal2 <int>, PreVal3 <int>,
#   PreVal4 <int>, PreVal5 <int>, PreVal6 <int>, PreVal7 <int>,
#   PreVal8 <int>, PreEff_Ave <dbl>, PreInt_Ave <dbl>, PreVal_Ave
<dbl>,
#   PostEff1 <int>, PostEff2 <int>, PostEff3 <int>, PostEff4 <int>,
#   PostEff5 <int>, PostInt1 <int>, PostInt2 <int>, PostInt3 <int>,
#   PostInt4 <int>, PostInt5 <int>, PostVal1 <int>, PostVal2 <int>,
#   PostVal3 <int>, PostVal4 <int>, PostVal5 <int>, PostVal6 <int>,
#   PostVal7 <int>, PostVal8 <int>, PostMaintInt1 <int>,
#   PostMaintInt2 <int>, PostMaintInt3 <int>, PostMaintInt4 <int>,
#   PostEff_Ave <dbl>, PostInt_Ave <dbl>, PostVal_Ave <dbl>,
#   PostMaintInt_Ave <dbl>, PreAchievement <int>, PostAchievement
<int>
```

Loading data (Tab-delimited, Excel, and SPSS)

```
read.delim("filename.txt") # Tab-delimited
readxl::read_excel("filename.xlsx") # Excel
haven::read_sav("filename.sav") # SPSS
```

Calculating summary statistics

```
my_data %>% count(SciTeacher) # counts frequencies and creates a
table
```

```
# A tibble: 4 × 2
  SciTeacher     n
       <int> <int>
1          0    53
2          1    55
3          2    53
4          3    51
```

```
my_data_ss <- select(my_data, contains("_Ave")) # this selects any
variables containing "Ave"
summary(my_data_ss) # creates summary statistics for continuous
variables
```

```
   PreEff_Ave      PreInt_Ave      PreVal_Ave     PostEff_Ave
 Min.   :2.200   Min.   :1.000   Min.   :1.875   Min.   :1.200
 1st Qu.:5.000   1st Qu.:4.800   1st Qu.:5.125   1st Qu.:4.800
 Median :5.600   Median :6.200   Median :6.125   Median :5.600
 Mean   :5.466   Mean   :5.739   Mean   :5.780   Mean   :5.354
 3rd Qu.:6.200   3rd Qu.:6.800   3rd Qu.:6.625   3rd Qu.:6.200
 Max.   :7.000   Max.   :7.000   Max.   :7.000   Max.   :7.000
                                                 NA's   :16
  PostInt_Ave     PostVal_Ave    PostMaintInt_Ave
 Min.   :1.000   Min.   :1.500   Min.   :1.00
 1st Qu.:4.400   1st Qu.:4.875   1st Qu.:3.50
 Median :5.800   Median :6.250   Median :5.00
 Mean   :5.394   Mean   :5.691   Mean   :4.72
 3rd Qu.:6.600   3rd Qu.:6.750   3rd Qu.:6.00
 Max.   :7.000   Max.   :7.000   Max.   :7.00
```

```
NA's   :16      NA's   :16      NA's   :18
```

dplyr for data manipulation

```
dplyr::select(my_data, PreAchievement, PostAchievement) # Select
only certain columns

dplyr::filter(my_data, PreAchievement >= 3) # select only certain
rows

dplyr::arrange(my_data, PostAchievement) # arrange data by a
variable
```

```
my_data %>%
    filter(PreAchievement >= 3) %>%
    group_by(SciTeacher) %>%
    summarize(SciTeacher_mean = mean(SciTeacher))
```

tidyr for reshaping and tidying data

```
stocks <- data_frame(
  time = as.Date('2009-01-01') + 0:9,
  X = rnorm(10, 0, 1),
  Y = rnorm(10, 0, 2),
  Z = rnorm(10, 0, 4)
)

stocks
```

```
# A tibble: 10 × 4
         time           X          Y          Z
       <date>       <dbl>      <dbl>      <dbl>
1  2009-01-01  0.55486073  2.2439625  0.6818610
2  2009-01-02 -1.20030507 -3.1660577  2.3646694
3  2009-01-03 -0.44413959  1.7341190  0.6827402
4  2009-01-04 -0.18554758  1.7967152 -0.3609197
5  2009-01-05  0.96336125  1.3014036  6.3654195
6  2009-01-06 -0.69798296  2.2115162 -4.6764298
7  2009-01-07 -0.50668007  0.7689348  1.0283452
8  2009-01-08  1.96965999 -4.6243631  4.8479906
9  2009-01-09  0.01062281  3.0295156  2.9525947
10 2009-01-10  0.99066158 -0.0639218 -3.2256691
```

gather() for reshaping from "wide" to "long" format

```
gather(stocks, stock, price, -time)
```

```
# A tibble: 30 × 3
         time stock       price
       <date> <chr>       <dbl>
1  2009-01-01     X  0.55486073
2  2009-01-02     X -1.20030507
3  2009-01-03     X -0.44413959
4  2009-01-04     X -0.18554758
5  2009-01-05     X  0.96336125
6  2009-01-06     X -0.69798296
7  2009-01-07     X -0.50668007
8  2009-01-08     X  1.96965999
9  2009-01-09     X  0.01062281
10 2009-01-10     X  0.99066158
# ... with 20 more rows
```

spread() for reshaping from "long"" to "wide" format

```
stocks_long <- gather(stocks, stock, price, -time)
spread(stocks_long, stock, price)
```

```
# A tibble: 10 × 4
         time           X          Y          Z
*      <date>       <dbl>      <dbl>      <dbl>
1  2009-01-01  0.55486073  2.2439625  0.6818610
2  2009-01-02 -1.20030507 -3.1660577  2.3646694
3  2009-01-03 -0.44413959  1.7341190  0.6827402
4  2009-01-04 -0.18554758  1.7967152 -0.3609197
5  2009-01-05  0.96336125  1.3014036  6.3654195
6  2009-01-06 -0.69798296  2.2115162 -4.6764298
7  2009-01-07 -0.50668007  0.7689348  1.0283452
8  2009-01-08  1.96965999 -4.6243631  4.8479906
9  2009-01-09  0.01062281  3.0295156  2.9525947
10 2009-01-10  0.99066158 -0.0639218 -3.2256691
```

Part IV: Advanced functionality

Packages

- Linear mixed models modeling: `lme4`, `nlme`
- Latent variable modeling: `lavaan`, `OpenMx`
- Social Network Analysis: `igraph`, `statnet`
- Text analysis: `quanteda`, `tidytext`

Linear mixed effects (multi-level) models

```
library(lme4)

model_1 <- lmer(engagement ~ challenge + percomp + (1 |
program_ID), data = df)

summary(model1)
```

Structural equation modeling

```
library(lavaan)

model <- '
  # measurement model
    ind60 =~ x1 + x2 + x3
    dem60 =~ y1 + y2 + y3 + y4
    dem65 =~ y5 + y6 + y7 + y8
  # regressions
    dem60 ~ ind60
    dem65 ~ ind60 + dem60
  # residual correlations
    y1 ~~ y5
    y2 ~~ y4 + y6
    y3 ~~ y7
    y4 ~~ y8
    y6 ~~ y8
'

fit <- sem(model, data = PoliticalDemocracy)
summary(fit, standardized = TRUE, fit.measures = T)
```

Social network analysis

```
library(igraph)
g <- graph.edgelist(edgematrix) # loads two column matrix with ties
V(g)$size = log(degree(g)) # changes size of vertices
V(g)$label <- NA # removes vertex names
V(g)$color[year_1_cohort_bool] <- "blue" # changes vertex color
based on logical index
l <- layout.fruchterman.reingold(g) # selects layout
E(g)$weight <- 1 # weights each edge equal to 1
g <- simplify(g, edge.attr.comb = list(weight = "sum")) # sums
edgeweights for equivalent ties

plot(g, layout = l,
     edge.width = E(g)weight)
```

Text analysis

```
library(quanteda)
my_corpus <- corpus(inaugTexts)
summary(my_corpus, n = 3)

my_dfm <- dfm(my_corpus, ignoredFeatures = stopwords("english"))
```

What are some other things we can do?

```
library(matchit) # propensity score matching
library(mgcv) # generalized additive models
library(modelr) # helper functions for modeling
library(rvest) # web scraping
library(caret) # machine learning framework
```

Part V: Additional resources

Rmarkdown & knitr

-
-
-
-

Shiny

-
- http://shinyapps.io
-
- Example: SETHs

Packages

-
-
-
- Example: prcr

Additional resources

-
  - Quick-R
  - R Studio Cheat Sheets
  - Stack Overflow
  - #rstats
  - RBloggers
-
  - Gelman & Hill (2006)
  - Grolemund & Wickham (2014)
  - Wickham & Grolemund (2017)

Thank you!

- Email: **jrosen@msu.edu**
- Web: **http://jmichaelrosenberg.com**
- Twitter: @jrosenberg6432