432 Class 01

https://thomaselove.github.io/432-2023/

2023-01-17

Today's Agenda

1. Mechanics of the course
2. Why I write dates the way I do
3. Data organization in spreadsheets
4. Naming Things and Getting Organized
5. Switching from R Markdown to Quarto
6. Building and Validating an ANOVA Model

Section 1

Course Mechanics

Welcome to 432.

Everything is at https://thomaselove.github.io/432-2023

- Syllabus
- Calendar
  - with all deadlines, and links to class READMEs
- Course Notes
- Details on Assignments to come (+ next slide)
- R and Data
  - Updating / Installing R, RStudio, necessary R Packages
  - Review / Learn some R Basics (also see 431 web site)
- Sources
  - Books, Articles, YouTube series, etc.
- Links to Canvas, Piazza and Contact Us

Assignments

Every deliverable for the entire semester is listed in the Calendar, except for the Welcome to 432 Survey, which at least 30 of you've done, and if you haven't, visit https://bit.ly/432-2022-welcome-survey.

- Two projects
  - Project 1 (use publicly available data for linear & logistic models)
    1. Proposal due 2022-01-31 (data selection, cleaning, exploration)
    2. Final Materials due 2022-03-04 (analyses, discussion)
  - Project 2 (use almost any data you like and analyze it well)
- Two Quizzes (Quiz 1 due 2022-02-21, Quiz 2 due 2022-04-18)
  - Multiple choice and short answer, mostly, taken via a Google Form
- Six labs
  - Labs will be posted before our next class. Lab 01 is due Monday 2022-01-24 at 9 PM.
- Ten minute papers
  - First is due 2022-01-19. These actually take about 5 minutes each.

Syllabus and Instructions will provide more information on grading/feedback.

The Spring 2022 Teaching Assistants for 432 are:

- Stephanie Merlino Barr, PhD student in Clinical Translational Science
- Wyatt Bensken, PhD student in Epidemiology & Biostatistics
- Ali Elsharkawi, MS student in Clinical Research
- Shiying Liu, PhD student in Epidemiology & Biostatistics
- Marie Michenkova, MS student in Biomedical Health Informatics
- Julia Yang Payne, PhD student in Clinical & Translational Science
- Monika Strah, MS student in Epidemiology & Biostatistics
- Yanning Wu, PhD student in Epidemiology & Biostatistics

All return from working with students in 431 this past Fall, and I couldn't be more grateful for their energy and effort. Learn more about the TAs in Section 6 of the Syllabus.

Getting Help

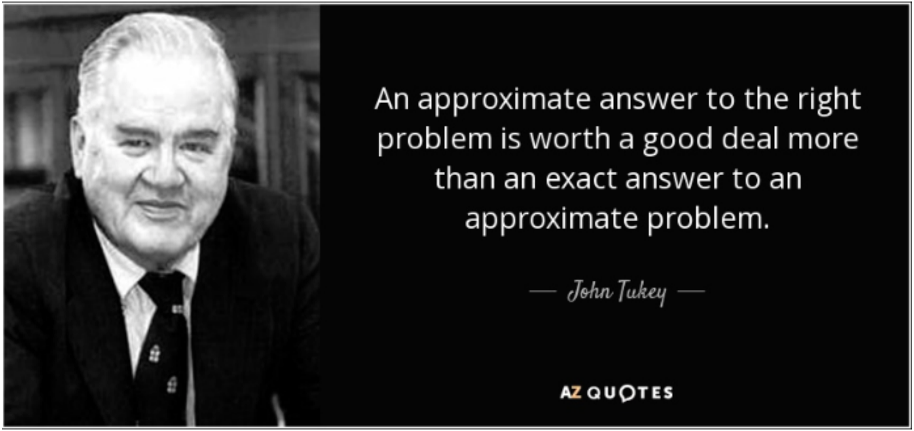
- Piazza is the location for discussion about the class. I follow it closely.
- We have 8 teaching assistants volunteering their time to help you.
- TAs will hold Office Hours beginning next Monday 2021-01-17 via Zoom, and the details will be available on Canvas (see Announcements) and our shared Google Drive.
- Dr. Love is available before and (especially) after class.
- Email Dr. Love directly only if you have a matter you need to discuss with him specifically. He's at `Thomas dot Love` at `case dot edu`.

We WELCOME your questions/comments/corrections/thoughts!

Tools You Will Definitely Use in this Class

- **Course Website** (see the bottom of this slide) especially the Calendar
    - Each class has a README (announcements, reminders, etc.) plus slides
- **R, RStudio and R Markdown** for, well, everything
- **Canvas** for access to Zoom meetings *and 432 recordings*, submission of most assignments
- **Google Drive via CWRU** for *recordings from 431*, forms (Minute Papers/Surveys/Quizzes) and receiving feedback on labs, projects, and Minute Papers
- **Piazza** is our discussion board. It's a moderated place to ask questions, answer questions of your colleagues, and get help fast. You don't have to pay to use it.
- **Zoom** for class sessions and for TA office hours

Some source materials are **password-protected**. What is the password?



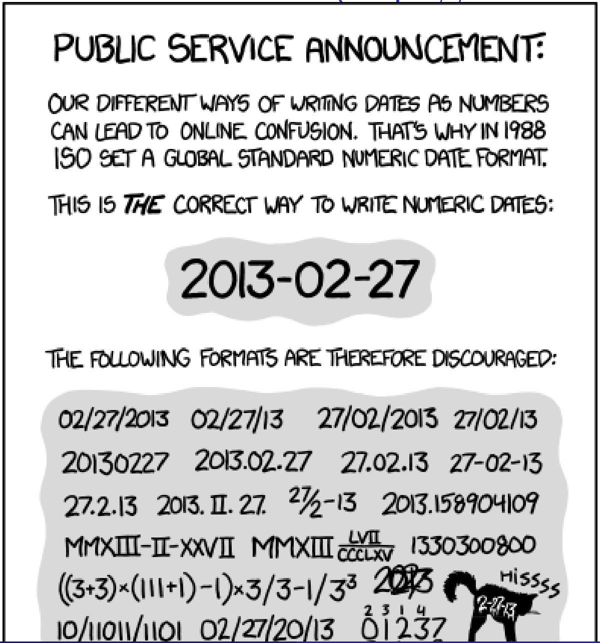
An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.

— John Tukey —

Section 2

Why I Write Dates The Way I Do

How To Write Dates (https://xkcd.com/1179/)


PUBLIC SERVICE ANNOUNCEMENT:
OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.
THIS IS *THE* CORRECT WAY TO WRITE NUMERIC DATES:
2013-02-27
THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:
02/27/2013 02/27/13 27/02/2013 27/02/13
20130227 2013.02.27 27.02.13 27-02-13
27.2.13 2013. II. 27. 27/2-13 2013.158904109
MMXIII-II-XXVII MMXIII LVII/CCCLXV 1330300800
((3+3)×(111+1)-1)×3/3-1/3³ 2013
HISSSSS
2-27-13
10/11011/1101 02/27/20/13 0 1 2 3 7
2 3 1 4

Section 3

Data Organization in Spreadsheets

Tidy Data (Wickham)

> *"A huge amount of effort is spent cleaning data to get it ready for analysis, but there has been little research on how to make data cleaning as easy and effective as possible....*

Tidy datasets are easy to manipulate, model and visualize, and have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table.

> *This framework makes it easy to tidy messy datasets because only a small set of tools are needed to deal with a wide range of un-tidy datasets. This structure also makes it easier to develop tidy tools for data analysis, tools that both input and output tidy datasets. The advantages of a consistent data structure and matching tools are demonstrated with a case study free from mundane data manipulation chores."*

https://www.jstatsoft.org/article/view/v059i10

“Data Tidying” presentation in *R for Data Science*

- Defines tidy data
- Demonstrates methods for tidying messy data in R

Read Sections

- 5 (Data transformation),
- 10 (Tibbles), 11 (Data import) and, especially, 12 (Tidy data)

https://r4ds.had.co.nz/

Data Organization in Spreadsheets (Broman & Woo)

- Create a data dictionary.
  - Jeff Leek has good thoughts on this in “How to Share Data with a Statistician” at https://github.com/jtleek/datasharing
  - Shannon Ellis and Jeff Leek’s preprint “How to Share data for Collaboration” touches on many of the same points at https://peerj.com/preprints/3139v5.pdf

We want:

1. The raw data.
2. A tidy data set.
3. A codebook describing each variable and its values in the tidy data set.
4. An explicit and exact recipe describing how you went from 1 to 2 and 3.

Data Organization in Spreadsheets: **Be Consistent**

- Consistent codes for categorical variables.
  - Either "M" or "Male" but not both at the same time.
  - Make it clear enough to reduce dependence on a codebook.
  - No spaces or special characters other than _ in category names.
- Consistent fixed codes for missing values.
  - NA is the most convenient R choice.
- Consistent variable names
  - In R, I'll use `clean_names` from the `janitor` package to turn everything into snake_case.
  - In R, start your variable names with letters. No spaces, no special characters other than _.
- Consistent subject / record identifiers
  - And if you're building a .csv in Excel, don't use ID as the name of that identifier.
- Consistent data layouts across multiple files.

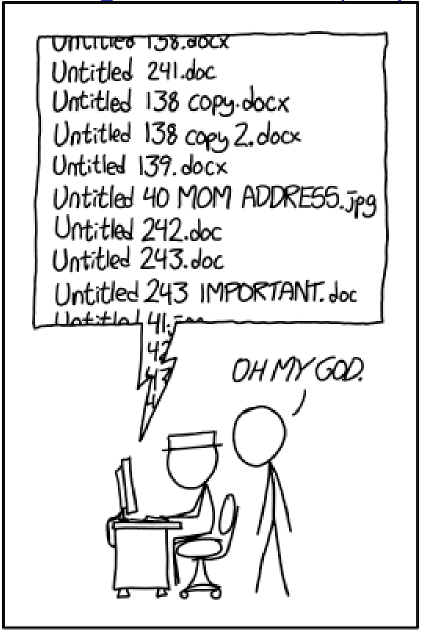
What Goes in a Cell?

- Make your data a rectangle.
  - Each row represents a record (sometimes a subject).
  - Each column represents a variable.
  - First column is a unique identifier for each record.
- No empty cells.
- One Thing in each cell.
- No calculations in the raw data
- No font colors
- No highlighting

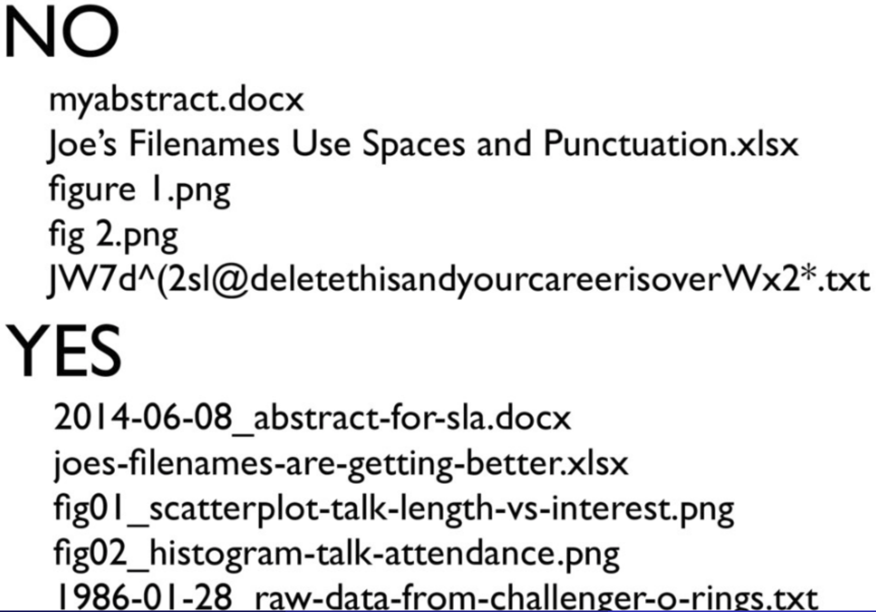
Section 4

Naming Things and Getting Organized

Naming Files is Hard (https://xkcd.com/1459/)



How To Name Files



NO

myabstract.docx
Joe’s Filenames Use Spaces and Punctuation.xlsx
figure 1.png
fig 2.png
JW7d^(2sl@deletethisandyourcareerisoverWx2*.txt

YES

2014-06-08_abstract-for-sla.docx
joes-filenames-are-getting-better.xlsx
fig01_scatterplot-talk-length-vs-interest.png
fig02_histogram-talk-attendance.png
1986-01-28_raw-data-from-challenger-o-rings.txt

Data Organization in Spreadsheets: Use consistent, strong file names.

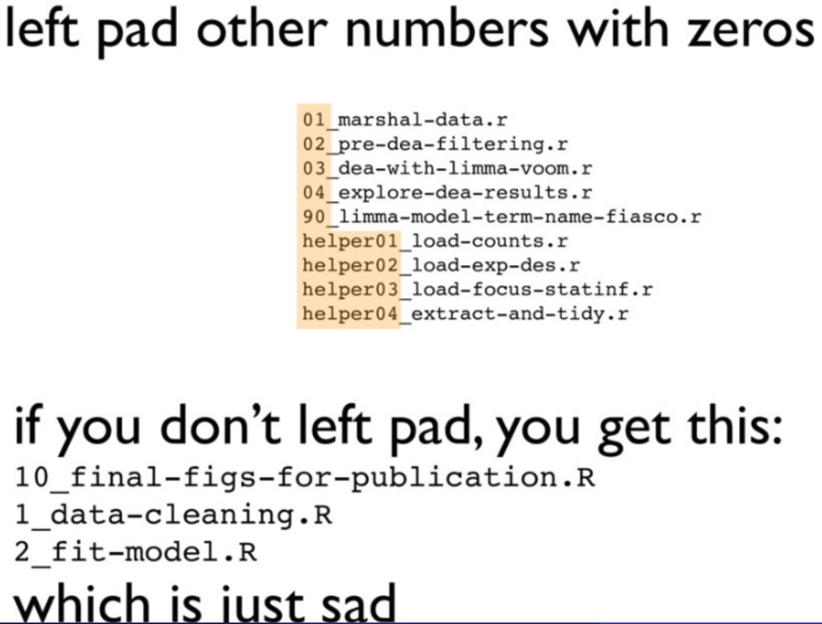
Jenny Bryan's advice on "Naming Things" hold up well. There's a full presentation at SpeakerDeck.

Good file names:

- are machine readable (easy to search, easy to extract info from names)
- are human readable (name contains content information, so it's easy to figure out what something is based on its name)
- play well with default ordering (something numeric first, left padded with zeros as needed, use ISO 8601 standard for dates)

Avoid: spaces, punctuation, accented characters, case sensitivity

from Jenny Bryan's "Naming Things" slides...



left pad other numbers with zeros

```
01_marshal-data.r
02_pre-dea-filtering.r
03_dea-with-limma-voom.r
04_explore-dea-results.r
90_limma-model-term-name-fiasco.r
helper01_load-counts.r
helper02_load-exp-des.r
helper03_load-focus-statinf.r
helper04_extract-and-tidy.r
```

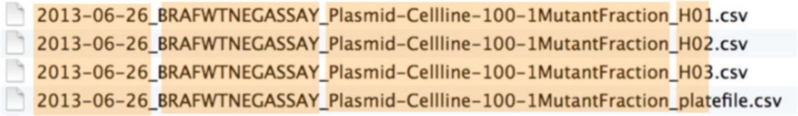
if you don't left pad, you get this:

```
10_final-figs-for-publication.R
1_data-cleaning.R
2_fit-model.R
```

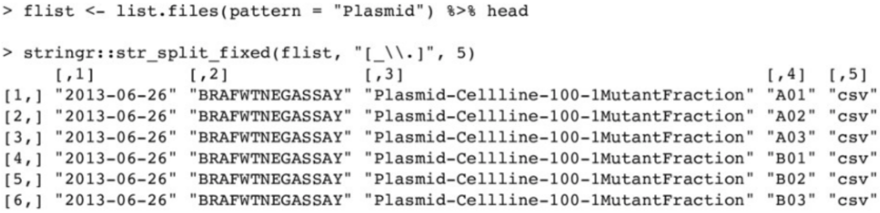
which is just sad

Jenny Bryan: Deliberate Use of Delimiters

Deliberately use delimiters to make things easy to compute on and make it easy to recover meta-data from the filenames.



2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_platefile.csv

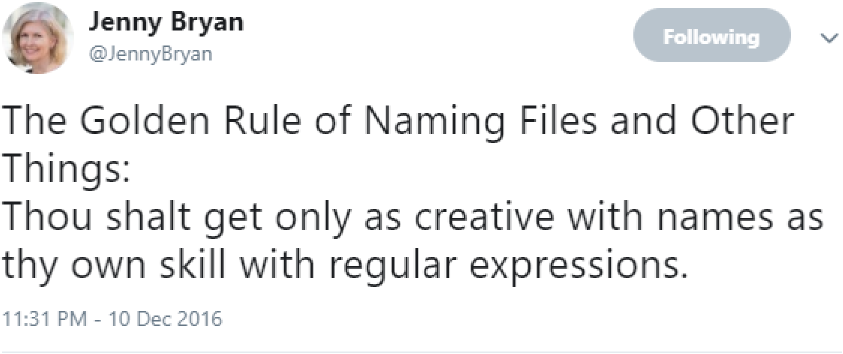


```
> flist <- list.files(pattern = "Plasmid") %>% head

> stringr::str_split_fixed(flist, "[_\\.]", 5)
     [,1]         [,2]            [,3]                                    [,4]  [,5]
[1,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "A01" "csv"
[2,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "A02" "csv"
[3,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "A03" "csv"
[4,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "B01" "csv"
[5,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "B02" "csv"
[6,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "B03" "csv"
```

"_" underscore used to delimit units of meta-data I want later

Don’t get too cute.

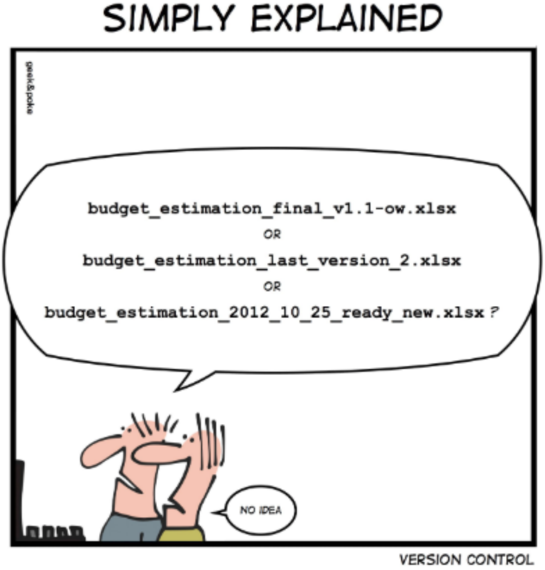


Jenny Bryan
@JennyBryan

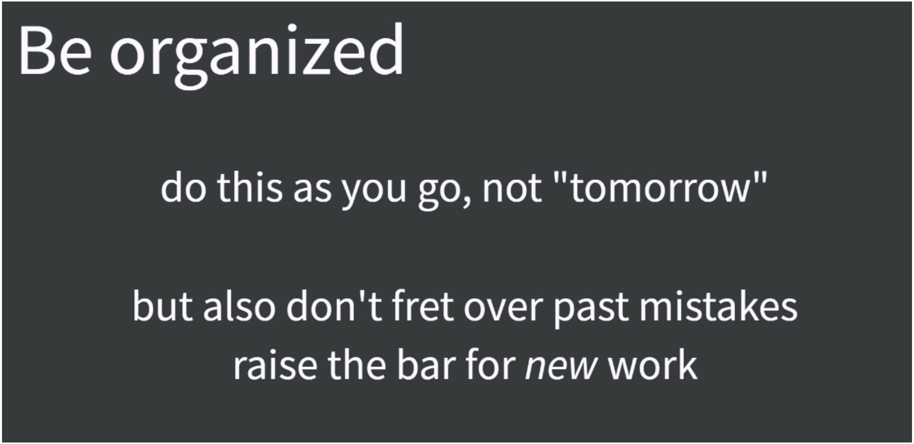
The Golden Rule of Naming Files and Other Things:
Thou shalt get only as creative with names as thy own skill with regular expressions.

11:31 PM - 10 Dec 2016

Goal: Avoid this...


SIMPLY EXPLAINED
geek&poke
budget_estimation_final_v1.1-ow.xlsx
OR
budget_estimation_last_version_2.xlsx
OR
budget_estimation_2012_10_25_ready_new.xlsx ?
NO IDEA
VERSION CONTROL

Get organized



Be organized

do this as you go, not "tomorrow"

but also don't fret over past mistakes
raise the bar for *new* work

Don’t spend a lot of time bemoaning or cleaning up past ills. Strive to improve this sort of thing going forward.

“Good Enough Practices in Scientific Computing”

1. Save the raw data.
2. Ensure that raw data is backed up in more than one location.
3. Create the data you wish to see in the world (the data you wish you had received.)
4. Create analysis-friendly, tidy data.
5. Record all of the steps used to process data.
6. Anticipate the need for multiple tables, and use a unique identifier for every record.

http://bit.ly/good-enuff

Lots of great advice here on software, collaboration and project organization.

Section 5

Switching from R Markdown to Quarto

details to come.

etc.

Section 6

Building and Validating an ANOVA Model

details to come.

etc.

Section 7

What Should I Be Working On?

For Next Time...

etc.