Week 14 – Digital research methods

Joshua Rosenberg and Alex Lishinski

April 22, 2021

Welcome!

Welcome to *week 14* – our last class (and week of class)!

Record the meeting

Breakout rooms!

Starting with whomever likes classical music the most:

- What is one thing you are proud of accomplishing this semester related to using R/this class?
- What is an R– or data science–related goal that you have beyond this class?

A recap of last week

- The map() functions can help you to read in multiple files
- Cases with missing values can still be used

Homework highlights

```
library(readr)
covid_data <- c("https://raw.githubusercontent.com/making-data-science-count/s21-intro-to-data-sci-me
"https://raw.githubusercontent.com/making-data-science-count/s21-intro-to-data-sci-methods-in-ed/main
mydata<-map_df(covid_data,read_csv)
```

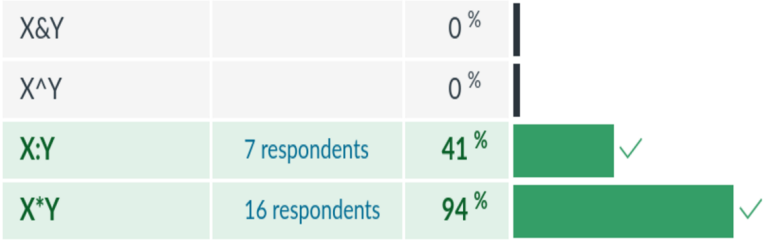
```
janssen <-mydata[1:378,c(1,2,3,8)]
moderna <-mydata[379:1449,4:7]
pfizer <-mydata[1450:2583,c(1,2,3,8)]

janssen <- janssen %>%
  rename(
    State = Jurisdiction,
    Week = `Week of Allocations`,
    First_Dose = `1st Dose Allocations`,
    Second_Dose = `2nd Dose Allocations`
  )
pfizer <- pfizer %>%
  rename(
    State = Jurisdiction,
    Week = `Week of Allocations`,
    First_Dose = `1st Dose Allocations`,
    Second_Dose = `2nd Dose Allocations`
  )
```

Exam 3

Great job!

Attempts: 17 out of 17

What is the syntax for adding an interaction to a regression model between variables X and Y?



35% answered correctly

Exam 3

Great job!

Attempts: 17 out of 17

Which of the following functions are used to extract the output from lm() models for further manipulation or better formatting?

broom::tidy()	14 respondents	**82 %**	✓
sjPlot::tab_model()	5 respondents	**29 %**	✓
broom::augment()	11 respondents	65 %	
broom::sweep()		0 %	

6% answered correctly

Topics for today

Record the meeting

- A. Introduction to digital research methods: Network and text analysis
- B. Looking ahead: where to next?
- C. Some (data–based) reflections

A. Digital research methods

Now is a good time to be studying social media (or exploring social media data as a hobby)

Facebook and Twitter have programs that allow accesss to researchers, including graduate students:

- Facebook and Instagram: https://research.fb.com/blog/2020/07/crowdtangle-opens-public-application-for-academics/
- Twitter: https://developer.twitter.com/en/solutions/academic-research

A. Digital research methods

There is a general *pattern* for how to access social media (and other!) data using R:

1. Apply for access to an "Application Programming Interface", or API, for a platform (or website)
2. Receive an API key automatically/within a few days
3. Use/enter the API key within R
4. Access data

A. Digital research methods

An example

AERA was recognized as quite unsuccessful this year; the organizers offered a full refund to all attendees due to technical difficulties

A parody account helped to bring levity to the situation: https://twitter.com/AeraHelpers

But, just how negative were posts about AERA on social media (Twitter)? We can use digital (or data science) research methods, those that would be hard or impossible to do without coding/using R.

```
library(academictwitteR)

all_ngsschat_tweets <- get_hashtag_tweets("#AERA21", "2010-01-01T00:00:00Z", "2020-12-31T00:00:00Z",
```

A. Digital research methods

There are a number of ways to "hide" your token so you don't have to repeatedly type it.

```
library(usethis)
edit_r_environ() # bearer_token = "xxx"

all_ngsschat_tweets <- get_hashtag_tweets("#AERA21", "2010-01-01T00:00:00Z", "2020-12-31T00:00:00Z",
```

A. Digital research methods

Once we have the data, there are a number of things we can do. Let's collect data for a different hashtag; in fact, a collection of hashtags.

```
hashtags_to_search <- c(str_c("#AERA", 14:21), str_c("AERA20", 14:21)) %>%
  str_c(collapse = " OR ")

hashtags_to_search
```

```
## [1] "#AERA14 OR #AERA15 OR #AERA16 OR #AERA17 OR #AERA18 OR #AERA19 OR #AERA20 OR #AERA21 OR AERA2014 OR AE
```

```
all_aera_tweets <- c(str_c("#AERA", 14:21), str_c("AERA20", 14:21)) %>%
  map(get_hashtag_tweets, "2010-01-01T00:00:00Z", "2021-04-17T00:00:00Z", bearer_token = Sys.getenv("
```

A. Digital research methods

Big idea: We can treat text like other types of data

```
all_aera_tweets
```

```
## # A tibble: 178,868 x 90
##    user_id  status_id  created_at          screen_name text      source display_text_wi… reply_to_status… r
##    <chr>    <chr>      <dttm>              <chr>       <chr>     <chr>              <dbl> <chr>             <
##  1 2818093… 101683922… 2018-07-11 00:18:58 anna_phd    "@cheris… Tweet…               220 101683797700200… 5
##  2 2818093… 109963488… 2019-02-24 11:39:21 anna_phd    "Heads u… Twitt…               144 <NA>              <
##  3 2818093… 138105460… 2021-04-11 01:20:51 anna_phd    "Woo hoo… Twitt…               140 <NA>              <
##  4 2818093… 110562275… 2019-03-13 00:12:59 anna_phd    "Hi #lit… Tweet…               211 <NA>              <
##  5 2818093… 638022356… 2015-08-30 16:15:58 anna_phd    "reviewi… Tweet…               140 <NA>              <
##  6 2818093… 138075285… 2021-04-10 05:21:48 anna_phd    "Join us… Twitt…                90 <NA>              <
##  7 2818093… 111461120… 2019-04-06 19:29:55 anna_phd    "@bnquis… Twitt…               139 <NA>              <
##  8 2818093… 111338594… 2019-04-03 10:21:09 anna_phd    "Poster … Twitt…               140 <NA>              <
##  9 2818093… 452436584… 2014-04-05 13:24:22 anna_phd    "\"In an… Tweet…               120 <NA>              <
## 10 2818093… 452879279… 2014-04-06 18:43:28 anna_phd    "Thinkin… Tweet…               140 <NA>              <
## # … with 178,858 more rows, and 81 more variables: reply_to_screen_name <chr>, is_quote <lgl>, is_retweet <
## #   favorite_count <int>, retweet_count <int>, quote_count <int>, reply_count <int>, hashtags <list>, symbo
## #   urls_url <list>, urls_t.co <list>, urls_expanded_url <list>, media_url <list>, media_t.co <list>,
## #   media_expanded_url <list>, media_type <list>, ext_media_url <list>, ext_media_t.co <list>,
## #   ext_media_expanded_url <list>, ext_media_type <chr>, mentions_user_id <list>, mentions_screen_name <lis
## #   lang <chr>, quoted_status_id <chr>, quoted_text <chr>, quoted_created_at <dttm>, quoted_source <chr>,
## #   quoted_favorite_count <int>, quoted_retweet_count <int>, quoted_user_id <chr>, quoted_screen_name <chr>
## #   quoted_name <chr>, quoted_followers_count <int>, quoted_friends_count <int>, quoted_statuses_count <int
## #   quoted_location <chr>, quoted_description <chr>, quoted_verified <lgl>, retweet_status_id <chr>, retwee
## #   retweet_created_at <dttm>, retweet_source <chr>, retweet_favorite_count <int>, retweet_retweet_count <i
## #   retweet_user_id <chr>, retweet_screen_name <chr>, retweet_name <chr>, retweet_followers_count <int>,
## #   retweet_friends_count <int>, retweet_statuses_count <int>, retweet_location <chr>, retweet_description
## #   retweet_verified <lgl>, place_url <chr>, place_name <chr>, place_full_name <chr>, place_type <chr>, cou
## #   country_code <chr>, geo_coords <list>, coords_coords <list>, bbox_coords <list>, status_url <chr>, name
## #   location <chr>, description <chr>, url <chr>, protected <lgl>, followers_count <int>, friends_count <in
## #   listed_count <int>, statuses_count <int>, favourites_count <int>, account_created_at <dttm>, verified <
## #   profile_url <chr>, profile_expanded_url <chr>, account_lang <lgl>, profile_banner_url <chr>,
## #   profile_background_url <chr>, profile_image_url <chr>
```

A. Digital research methods

Big idea: We can treat text like other types of data

```
library(tidytext)

all_aera_tweets_words <- all_aera_tweets %>%
  select(created_at, text) %>%
  unnest_tokens(word, text)

all_aera_tweets_words
```

```
## # A tibble: 4,257,813 x 2
##    created_at          word
##    <dttm>              <chr>
##  1 2018-07-11 00:18:58 cherisemcb
##  2 2018-07-11 00:18:58 aeraticl
##  3 2018-07-11 00:18:58 a2
##  4 2018-07-11 00:18:58 i
##  5 2018-07-11 00:18:58 too
##  6 2018-07-11 00:18:58 am
##  7 2018-07-11 00:18:58 interested
##  8 2018-07-11 00:18:58 in
##  9 2018-07-11 00:18:58 broadening
## 10 2018-07-11 00:18:58 my
## # … with 4,257,803 more rows
```

A. Digital research methods

Sentiment analysis is as straightforward as *joining* a pre-identified collection of positive and negative words – a dictionary

```
all_aera_tweets_words %>%
  inner_join(get_sentiments("bing"))
```

```
## # A tibble: 177,791 x 3
##    created_at          word     sentiment
##    <dttm>              <chr>    <chr>
##  1 2018-07-11 00:18:58 work     positive
##  2 2018-07-11 00:18:58 work     positive
##  3 2018-07-11 00:18:58 love     positive
##  4 2018-07-11 00:18:58 fresh    positive
##  5 2021-04-11 01:20:51 woo      positive
##  6 2021-04-11 01:20:51 winning  positive
##  7 2021-04-11 01:20:51 best     positive
##  8 2021-04-11 01:20:51 award    positive
##  9 2019-03-13 00:12:59 thrilled positive
## 10 2019-03-13 00:12:59 bad      negative
## # … with 177,781 more rows
```

A. Digital research methods

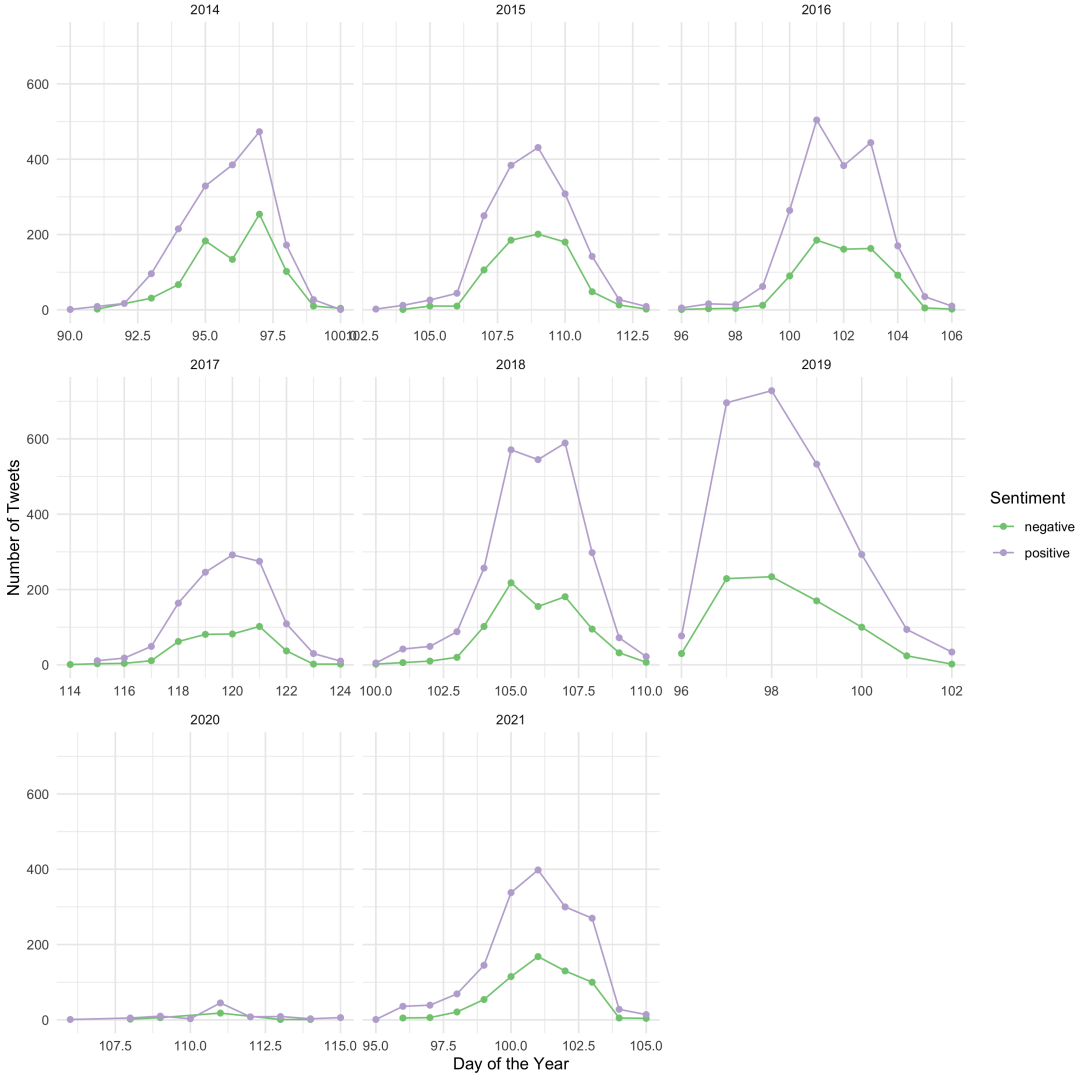
A different dictionary; from these, we can summarize and visualize the sentiment scores

```
all_aera_tweets_words %>%
  inner_join(get_sentiments("afinn"))
```

```
## # A tibble: 186,189 x 3
##    created_at          word       value
##    <dttm>              <chr>      <dbl>
##  1 2018-07-11 00:18:58 interested     2
##  2 2018-07-11 00:18:58 love           3
##  3 2018-07-11 00:18:58 fresh          1
##  4 2019-02-24 11:39:21 hope           2
##  5 2021-04-11 01:20:51 woo            3
##  6 2021-04-11 01:20:51 congrats       2
##  7 2021-04-11 01:20:51 winning        4
##  8 2021-04-11 01:20:51 best           3
##  9 2021-04-11 01:20:51 award          3
## 10 2019-03-13 00:12:59 thrilled       5
## # … with 186,179 more rows
```

A. One way of representing the results

Using the bing dictionary; see academic-twitter-example.R



A. Digital research methods

This analysis invites a number of questions:

- Was AERA really not that much more negative than in years past? Or, is there sone difference?
- Is there an effect of the dictionary selected; is the "bing" dictionary not detecting the *type* of negativity in these short, social media–based posts?
- Were negative posts this year interacted with more than positive posts?
- Could a qualitative analysis reveal nuances between positive and negative labeled posts?
- How does AERA compare to other conferences, like LAK, that appeared to have been better received?

A. Digital research methods

Some completed and ongoing projects Josh has (and others in class have!) worked on:

Rosenberg, J. M., Reid, J., Dyer, E., Koehler, M. J., Fischer, C., & McKenna, T. J. (2020). Idle chatter or compelling conversation? The potential of the social media–based #NGSSchat network as a support for science education reform efforts. Journal of Research in Science Teaching, 57(9), 1322–1355. https://onlinelibrary.wiley.com/doi/10.1002/tea.21660

Greenhalgh, S. P., Rosenberg, J. M., Koehler, M. J., Akcaoglu, M., & Staudt Willet, K. B. (2020). Identifying multiple learning spaces within a single teacher–focused Twitter hashtag. Computers & Education, 148(4). https://doi.org/10.1016/j.compedu.2020.103809

Greenhalgh, S. P., Staudt Willet, K. B., Rosenberg, J. M., & Koehler, M. J. (2018). Tweet, and we shall find: Using digital methods to locate participants in educational hashtags. TechTrends, 62(5), 501–508. https://doi.org/10.1007/s11528–018–0313–6

Kimmons, R., Rosenberg, J., & Allman, B. (2021). Trends in Educational Technology: What Facebook, Twitter, and Scopus Can Tell us about Current Research and Practice. TechTrends, 1–12. https://link.springer.com/article/10.1007/s11528–021–00589–6

Understanding public sentiment about educational reforms: The Next Generation Science Standards on Twitter. https://osf.io/xymsd/

"We are trying to communicate the best we can": Districts' communication on Twitter during the COVID–19 pandemic. https://osf.io/qpu8v/

Is student privacy "quick and easy"? Investigating student images and names on K–12 educational institution's Facebook postings. https://osf.io/5tpb9/

B. Looking ahead

There are so many things you can do!

Some broad suggestions, first:

- It's normal to feel overwhelmed while learning skills like R and data science; adopt a long–term, growth approach (or mindset) to learning R
- Surround yourself with resources
- Ask for help!
    - Peers and us (we're part of your R network now!)
- Share with others
    - Peers and us!
- Welcome others

B. Looking ahead

We also have some specific suggestions:

- Ask and receive help via the RStudio Community: https://community.rstudio.com/
- create a blog with blogdown: https://bookdown.org/yihui/blogdown/
- Twitter and #rstats
- R Weekly
- R for Data Science: https://r4ds.had.co.nz/
- Twitter users:
    - @ivelasq
    - @kierisi
    - @ry_estrellado
    - @efreer20
    - @hadleywickham
- Other resources: https://datascienceineducation.com/c18.html

B. Looking ahead

- Research presentations and papers that you lead or co–author
- Workshops/training opportunities at UTK and beyond
- Job opportunities (Skill with R is in–demand at present)
- Other courses at UTK
    - . . . including those offered within TPTE (Josh and Alex's department; consider the graduate certificate)

C. Some reflections

- Teaching this class has been an amazing experience
- We are really thankful for your ambition, persistence, and positivity
- Kudos to you all

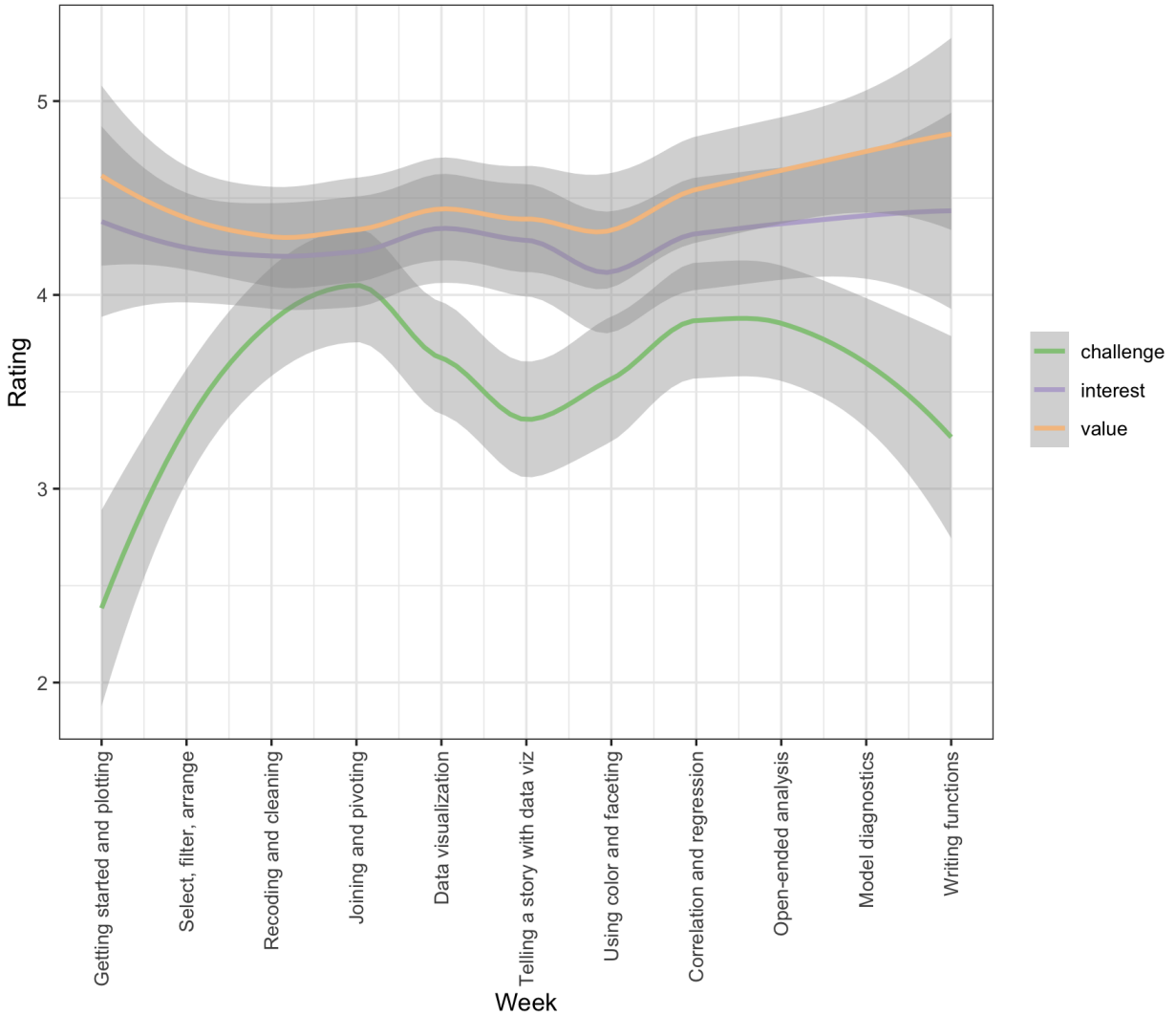
C. Some reflections

We are learning, too; we learned a lot:

- Assigning base groups!
- Make the class *longer*
- You can do ambitious work early on; emphasize early
- Keep homework relatively narrow, focused, and self–contained
- Emphasize projects/authentic contexts earlier and more next time
- Small technical issues can present headaches; use R Studio Cloud (early)?

C. Some reflections

See exit-ticket-analysis.R



C. Some reflections

Take–aways from the homework reflections:

- Interest and value remained relatively high for most of you (with substantial between–individual and some between–assignment variation)
- Challenge was moderate throughout, with some peaks that we may wish to avoid (or provide support/instruction around)
- Your comments when submitting the homework were great as a barometer for us
- The same is true for your end–of–class check–outs
- Data science is hard; teaching it is fun and hard in some ways, too; thanks for sticking it through and for remaining upbeat

C. Some reflections

Other highlights

- Kudos to everyone dealing with the pandemic; it has been a challenging time – at best
- Special kudos to our mathematics teachers
- Special kudos to anyone who was caring for oneself or others this semester
- Let's move forward knowing that we can get through this challenging time, hopefully stronger in some ways

Data Science Certificate

A bit more detail on our plans:

- We'll start Fall, 2021 (Spring, 2022 would be the next relevant class for you; certificate will likely be approved by this time)
- Can complete within two years
- Will allow for one eligible course to be replaced with another computationally–focused/data science class that you have taken or will taken; *so, you will need to only take three of the four classes to complete the certificate*

Course	Semester (Fall 1 = Fall, 2021; Spring 1 = Spring, 2022)	Possible Instructors
Introduction to Data Science in Education Using R	Fall 1, Fall 2 (offered every Fall)	Rosenberg, Lishinski
Studying Digital Learning Environments and Networks	Spring 1	Rosenberg, Lishinski
Predictive Modeling and Machine Learning	Fall 2	Rosenberg
Data Visualization	Spring 2	Schmidt; Velasquez

https://docs.google.com/document/d/1shNL9fY–vhBKK–zJw_V_iVhAKDtZ1Wqen4maJFnswQ4/edit

Logistics

This week

- Homework 14: Available tomorrow by noon; **Due by Thursday, 4/29**
- Reading:
    - Big Magic With R: Creating Learning Beyond Fear: https://alison.rbind.io/talk/2018-cascadia-bigmagic/

Final project recordings/presentations

- Strict deadline of 11:59 pm ET on 5/5
- We recommend using Zoom and recording to the Cloud – then sharing the link in the #final–projects channel.
- Also, *you will share your recording with the members of your base group*
- We'll ask you to view and comment on your base group member's presentations by Sunday, May 9 (just over four days); you can comment on anyone else's, too!
- Office hours:
    - Josh: next two Tuesdays (4/27 and 5/4), 4–5 pm
    - Alex: Wednesdays (4/28), 12–2 pm

Schedule

- The revised data ethics statement is due **April 29**
- The product for your final project and a recording that will serve as your presentation of it are both due by the end of the day on **May 5**
- The curating a resource assignment is due by the same date/time as the final project, on **May 5**

That's it!

Requests for input/suggestions

1. We value your *honest input* (TN Voice: https://utk.campuslabs.com/eval-home/ (4/29 deadline)
2. We'd also love your suggestions for how to improve the class in a very short end-of-class survey: https://utk.co1.qualtrics.com/jfe/form/SV_4G8IZYrfoGMuM7P (5/11 deadline)
3. We may ask you after grades are entered for you to consent for us to use your work for data science education research, with specifics only shared with your permission; possible incentive to participate; see email after 5/11

Wrapping up

In your base group's Slack channel:

- What is one thing you learned *this semester*?
- What is something you want to learn more about?
- Share your feelings in GIF form!

Please stay in touch!