Data ingestation, data collection, and data analysis: key components in the statistics and data science analysis cycle

Mine Dogucu (University of California, Irvine)

Albert Y. Kim (Smith College)

CAUSE/JSDSE webinar series

Welcome from our host and moderator

Nicholas Horton (Amherst College)

Journal of Statistics and Data Science Education update

- Special issue published yesterday on “Computing in the Statistics and Data Science Curriculum”, Volume 29, supplementary (2021)
- Editorial, commentary by Nolan and Temple Lang, plus 13 original papers are included (including the ones we will be discussing today)
- As always, all open access, no author fees, per journal policy

https://www.tandfonline.com/toc/ujse21/29/sup1?nav=tocList

CAUSE/Journal of Statistics and Data Science Education webinar series

Upcoming webinars:

- "Teaching statistics online during the pandemic" (Tuesday, March 30th, 10:00am-11:30am ET)
- "What makes a good statistical question?" (Tuesday, April 27th, 4:00-4:45pm ET)
- Signup at https://www.causeweb.org/cause/webinars

Webinars are recorded and posted (with slides) at that same site

Consortium for the Advancement of Undergraduate Statistics Education

https://www.causeweb.org/cause

June 28 - July 1, 2021, registration only $25 (opens soon)

Consortium for the Advancement of Undergraduate Statistics Education

https://www.causeweb.org/cause

Undergraduate Statistics Project Competition
(Class Project and Research Project)
Undergraduate Statistics Research Conference
(eUSR)

More info at https://www.causeweb.org/usproc

Today's Breaking News

International Prize in Statistics Awarded to Nan Laird for Methods of Analyzing Data from Longitudinal Studies

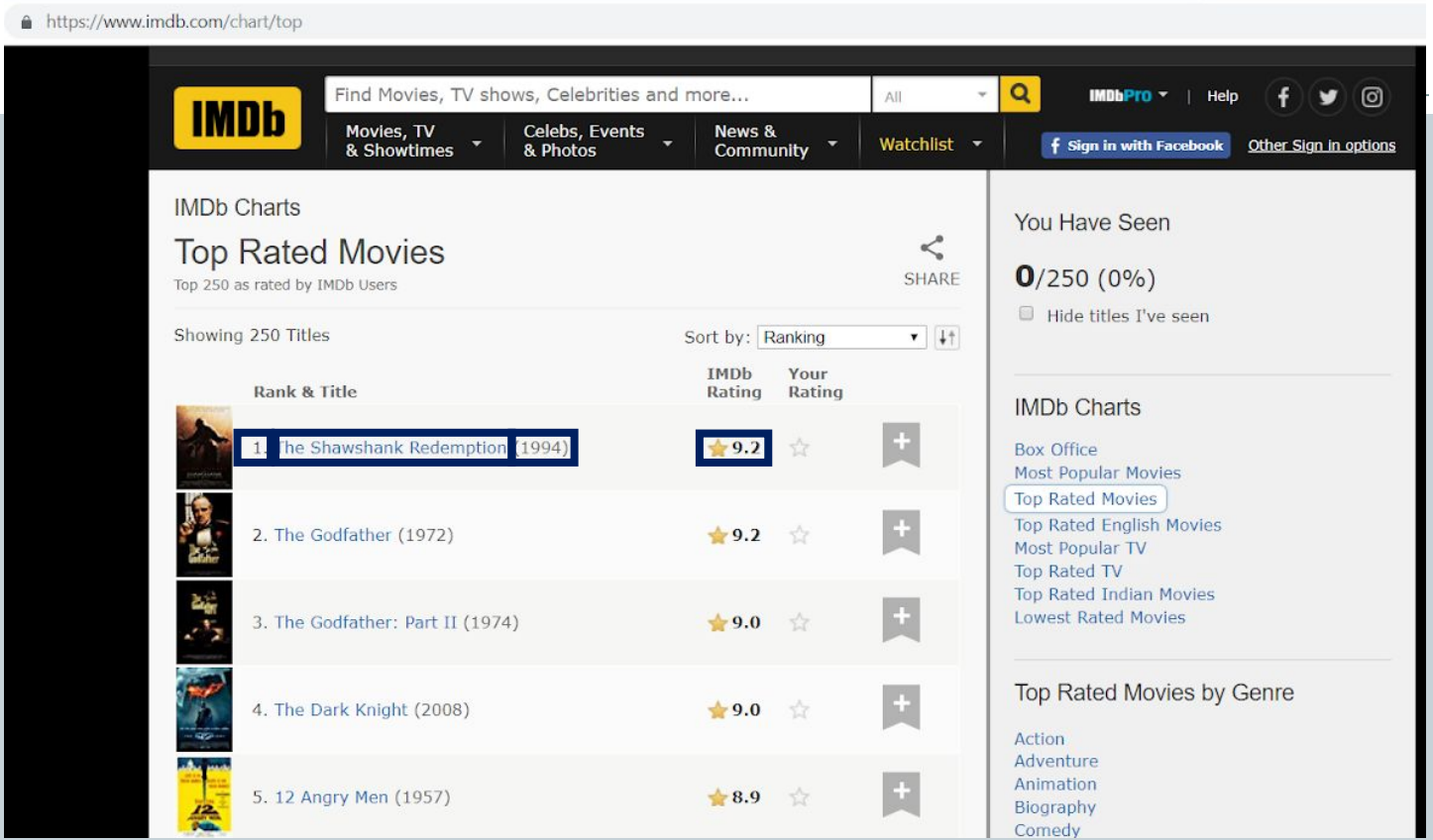
https://www.statprize.org

Web Scraping in the Statistics and Data Science Curriculum

Mine Dogucu
Department of Statistics
University of California/Irvine

Dogucu and Cetinkaya-Rundel (2021): “Web Scraping in the Statistics and Data Science Curriculum: Challenges and Opportunities” (https://github.com/mdogucu/web-scrape)


Web
Scraping
Protocol
CSV


https://www.imdb.com/chart/top
IMDb
Find Movies, TV shows, Celebrities and more...
All
IMDbPro
Help
Movies, TV & Showtimes
Celebs, Events & Photos
News & Community
Watchlist
Sign in with Facebook
Other Sign in options
IMDb Charts
Top Rated Movies
Top 250 as rated by IMDb Users
SHARE
Showing 250 Titles
Sort by: Ranking
Rank & Title
IMDb Rating
Your Rating
1. The Shawshank Redemption (1994) 9.2
2. The Godfather (1972) 9.2
3. The Godfather: Part II (1974) 9.0
4. The Dark Knight (2008) 9.0
5. 12 Angry Men (1957) 8.9
You Have Seen
0/250 (0%)
Hide titles I've seen
IMDb Charts
Box Office
Most Popular Movies
Top Rated Movies
Top Rated English Movies
Most Popular TV
Top Rated TV
Top Rated Indian Movies
Lowest Rated Movies
Top Rated Movies by Genre
Action
Adventure
Animation
Biography
Comedy

Web Scraping

	title	year	rating
1	The Shawshank Redemption	1994	9.2
2	The Godfather	1972	9.2
3	The Godfather: Part II	1974	9.0
4	The Dark Knight	2008	9.0
5	12 Angry Men	1957	8.9
6	Schindler's List	1993	8.9
7	The Lord of the Rings: The Return of the King	2003	8.9
8	Pulp Fiction	1994	8.9
9	The Good, the Bad and the Ugly	1966	8.8
10	Fight Club	1999	8.8
11	The Lord of the Rings: The Fellowship of the Ring	2001	8.8

The Journal of Statistics Education was established in 1992. *You are currently reading a manuscript from this journal.*

The American Statistician was established in 1947.

Figure 2. HTML document with CSS example.

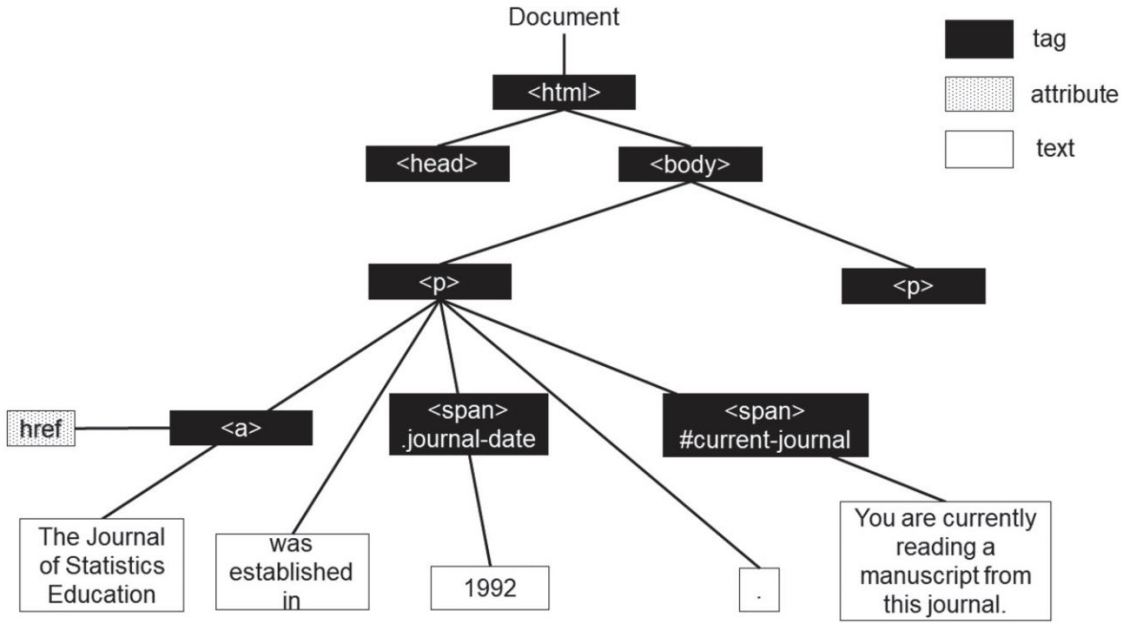


Figure 3. Partial HTML tree of HTML and CSS example.

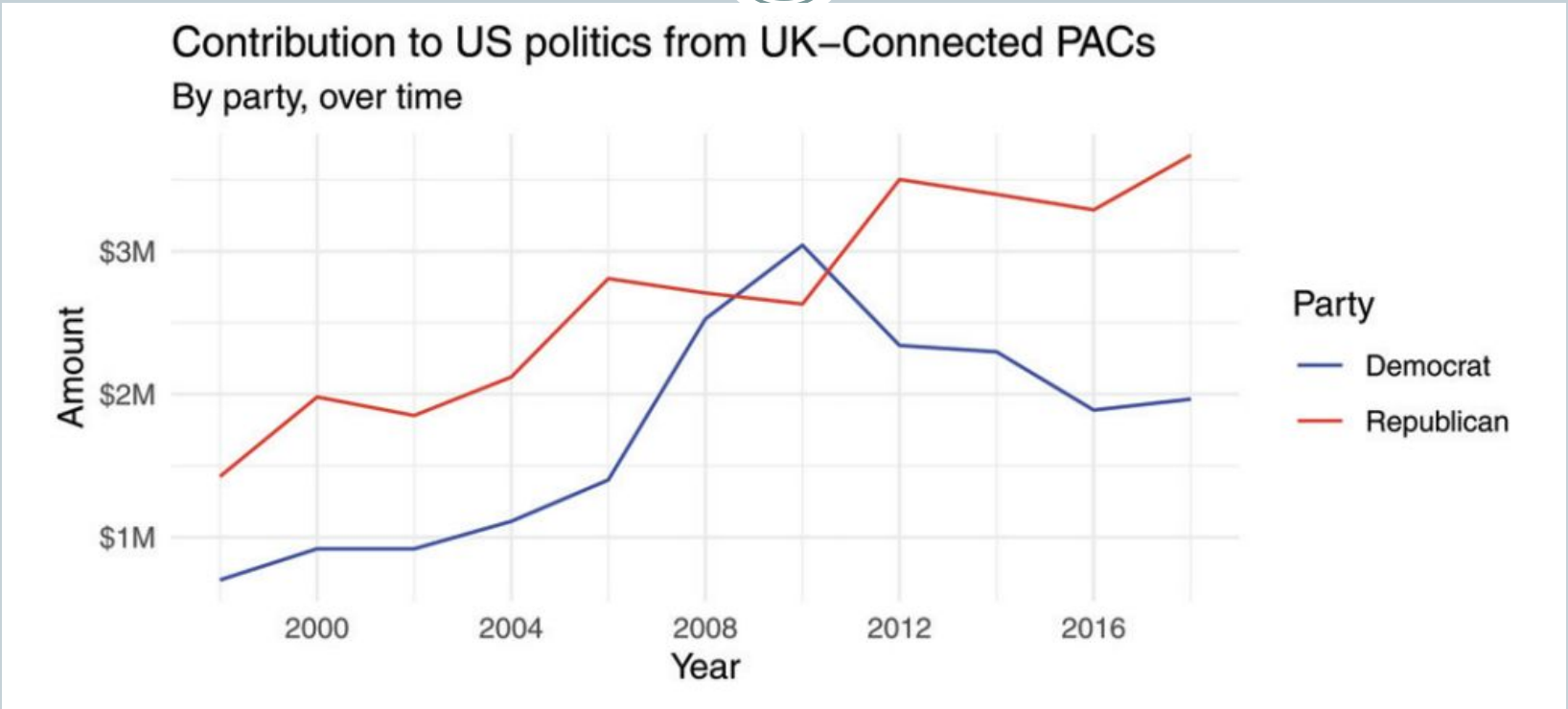
Selector Gadget

PAC Name (Affiliate)	Country of Origin/Parent Company	Total	Dems	Repubs
7-Eleven	Japan/Seven & I Holdings	$1,000	$0	$1,000
ABB Group (ABB Group)	Switzerland/Asea Brown Boveri	$8,000	$3,500	$4,500
Accenture (Accenture)	Ireland/Accenture plc	$82,000	$49,000	$33,000
Air Liquide America	France/L'Air Liquide SA	$14,000	$5,000	$9,000
Airbus Group	Netherlands/Airbus Group	$159,000	$66,000	$93,000
Alkermes Inc	Ireland/Alkermes Plc	$77,250	$25,750	$51,500
Allergan PLC (Allergan PLC)	Ireland/Allergan PLC	$111,000	$6,000	$105,000
Allianz of America (Allianz)	Germany/Allianz AG Holding	$46,500	$19,350	$27,150
Anheuser-Busch (Anheuser-Busch InBev)	Belgium/Anheuser-Busch InBev	$252,000	$127,000	$125,000
AON Corp (AON plc)	UK/AON PLC	$45,000	$17,500	$27,500
APL Maritime (CMA CGM)	France/CMA CGM SA	$15,000	$8,500	$6,500

Figure 4. Identifying the CSS selector for the table of contributions using the SelectorGadget. Retrieved on April 30, 2020.

Classroom Examples

Level 1: Scraping a Table From a Single Website
Level 2: Writing Functions
Level 3: Iteration
Level 4: Data Cleaning and Visualization


Contribution to US politics from UK–Connected PACs
By party, over time
Amount
$3M
$2M
$1M
2000
2004
2008
2012
2016
Year
Party
Democrat
Republican

Challenges

- Reproducibility
- Missing nodes
- Lack of Control over Connectivity
- Lack of Control over Content

Opportunities

- Current, rich, and interesting data
- Different size, shape, and format of data
- Blending of computing and statistics
- Discussion of ethics

Questions to Ask Before Teaching Web Scraping

- Is the data from human subjects? If yes, is it ethical to scrape the data?
- Does the website provide an API?
- Does the website allow web scraping?
- Are the data provided in an HTML table?
- Are the CSS selectors easy to select with SelectorGadget?
- Is there non-numeric data? If yes, how easy is it to manipulate it?
- Would the process of scraping involve iteration over multiple pages? If yes, how much data are you planning to scrape, all or a sample?

Summary

Goal	Section
You want to learn the basic idea behind HTML & CSS	Section 2.1 (HTML & CSS)
You want to learn web scraping using R	Sections 2 (Technical Tools) and 3 (Classroom Examples).
You want to know why one should learn web scraping and/or teach web scraping	Section 1 (Introduction) and Section 5
You are trying to learn web scraping but running into problems	Section 4 (Challenges)
You want to teach web scraping and need an example	Section 3 (Classroom Examples)
You want to teach web scraping or already are teaching web scraping and want to consider the bigger picture.	Section 4 (Challenges) and Section 5 (Opportunities)

https://www.datapedagogy.com/posts/2020-07-15-web-scrape

Playing the whole game: data collection and analysis with Google calendar

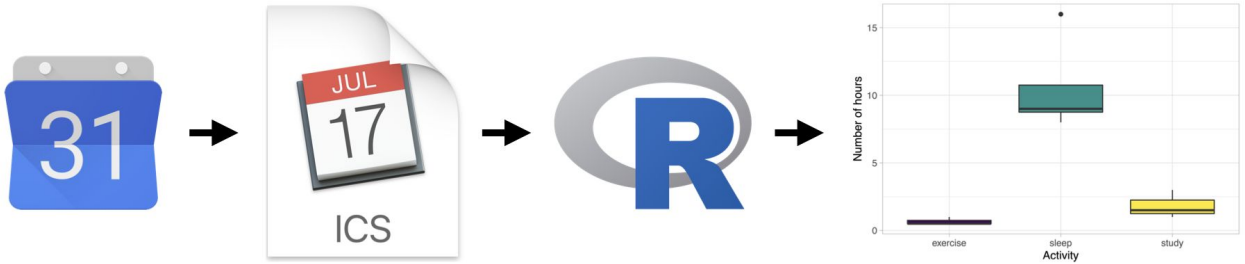
Albert Young-Sun Kim
Statistical and Data Sciences
Smith College

Kim and Hardin (2021) "'Playing the Whole Game': A Data Collection and Analysis Exercise With Google Calendar"

Like to paper & instructor resources:
(https://smithcollege-sds.github.io/sds-www/JSE_calendar.html)

Graphical Abstract

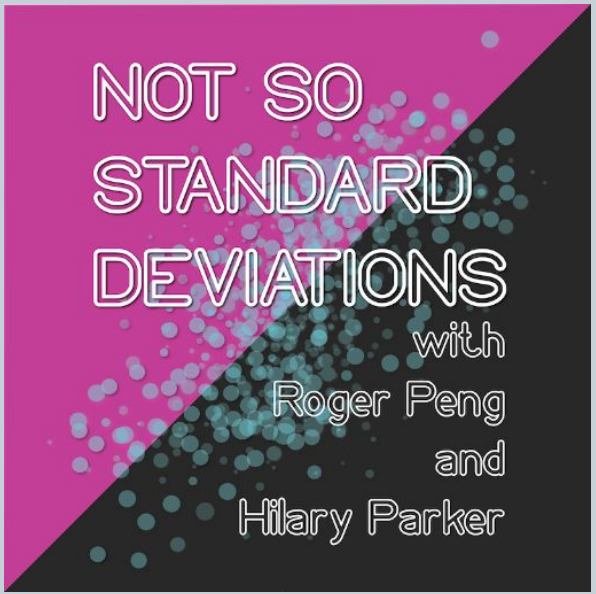
"Playing the whole game":
A data collection & analysis exercise with Google Calendar



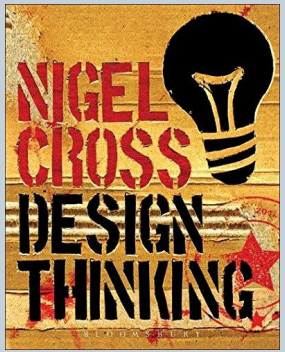
1. Log activities in Google Calendar
2. Export to `.ics` file format
3. Import to R using `ical` package
4. Analyze

Iterate as needed!

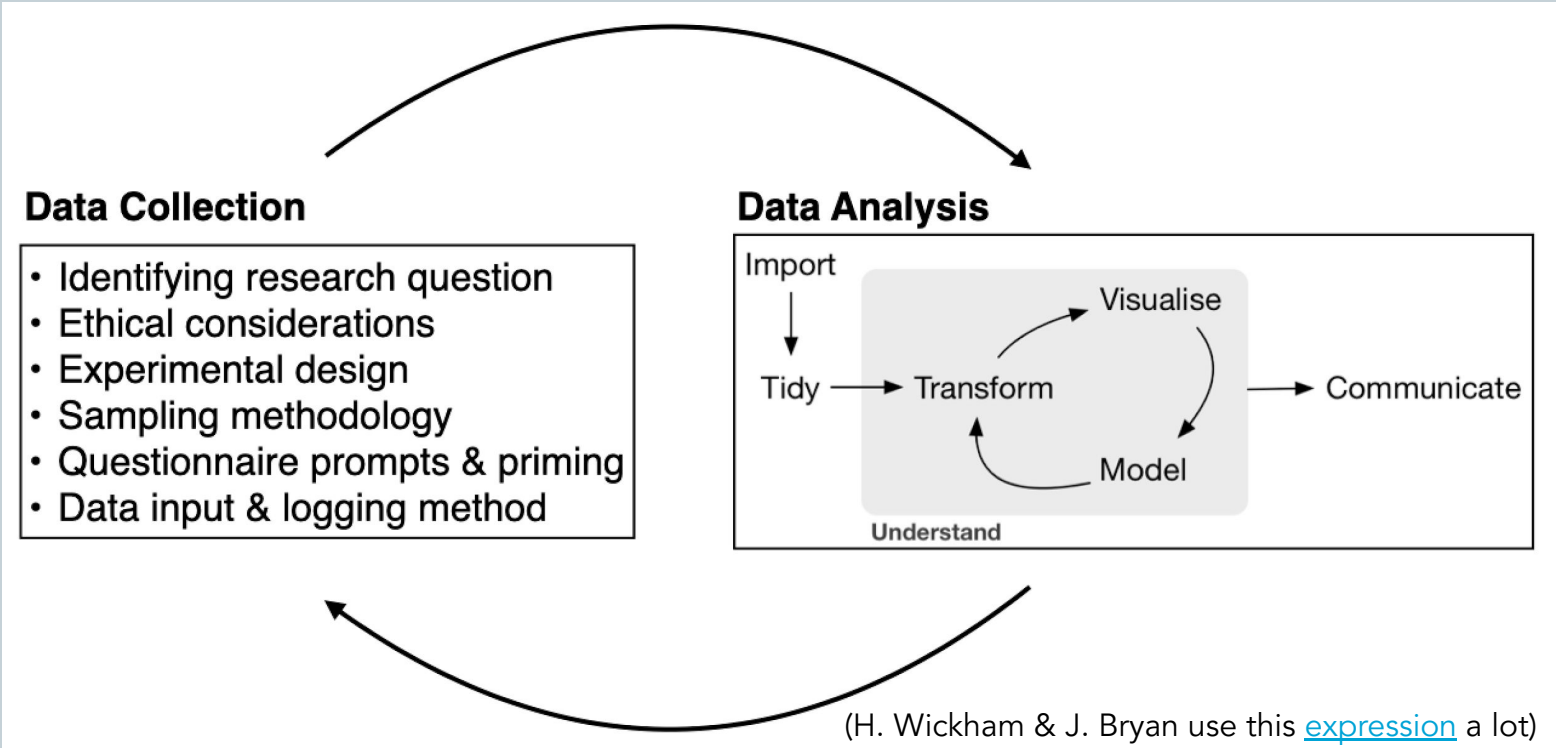
Jo Hardin's moment of 



- Ep 71 Compromised Shoe Situation
- Hosts collected data to understand what factors impact their commute time
- Part of first data science "design challenge"

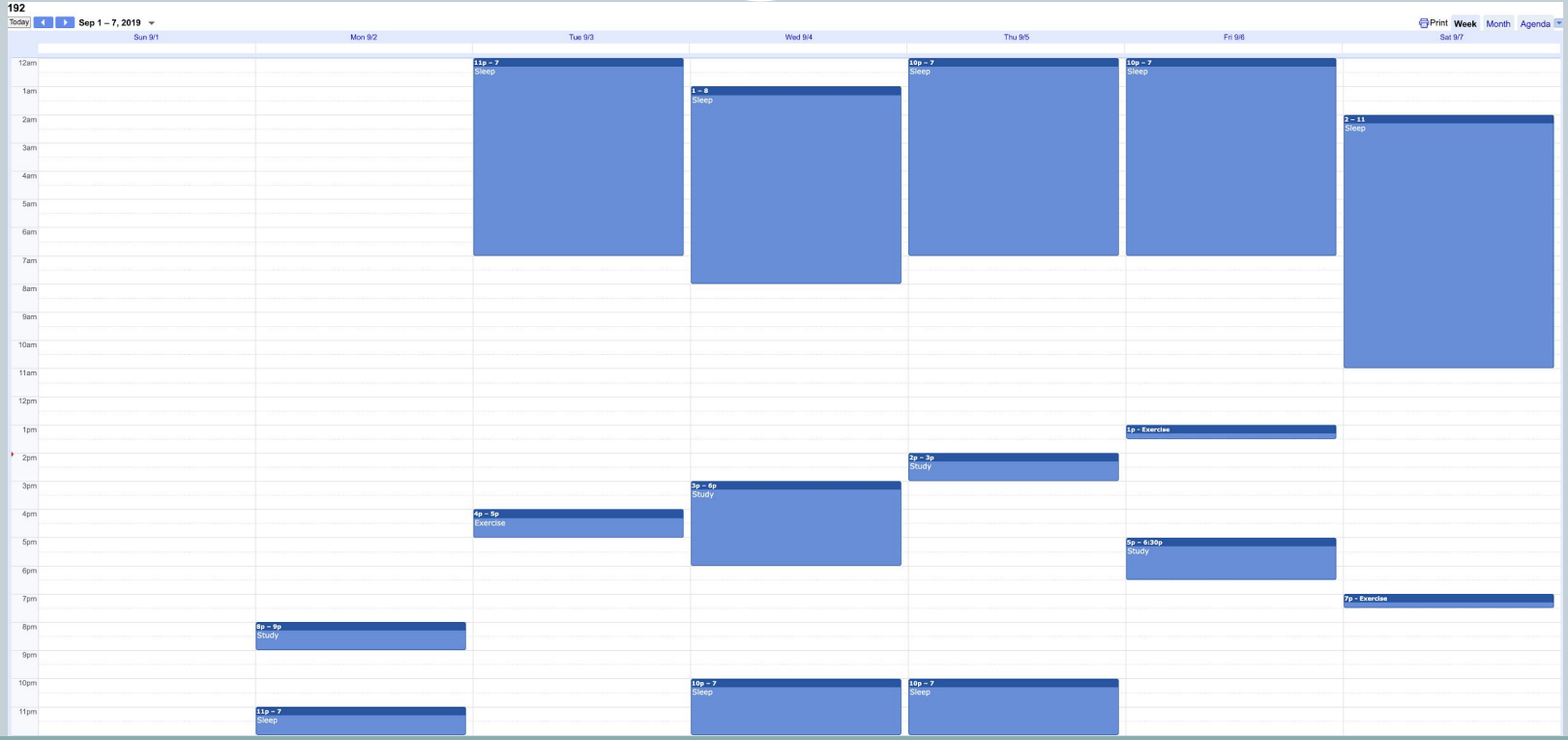


"Playing the whole game…"



(H. Wickham & J. Bryan use this expression a lot)

The starting point



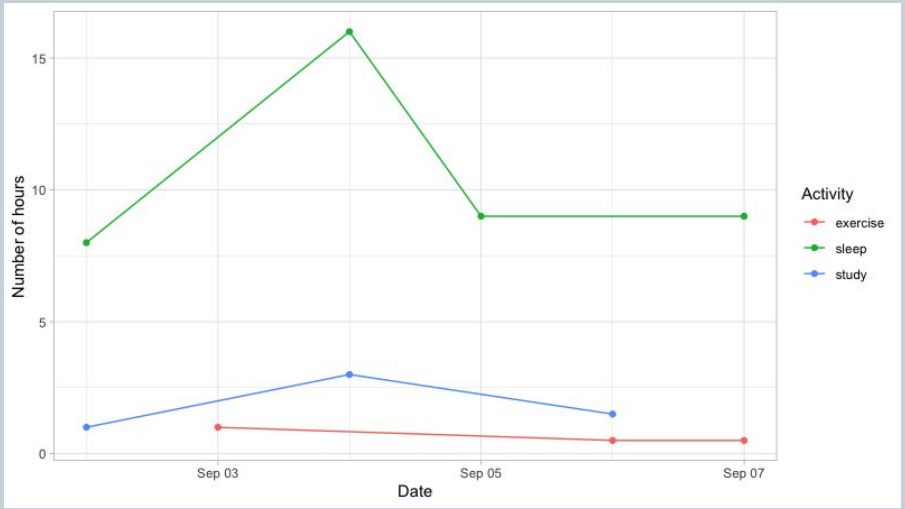
The code

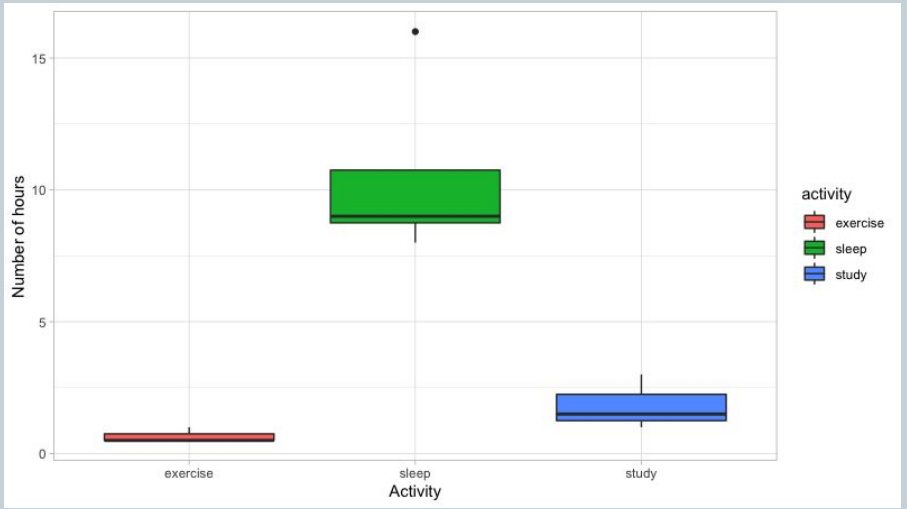
```r
library(ggplot2)
library(dplyr)
library(lubridate)
library(ical)

calendar_data <- "192.ics" %>%
  # Use ical package to import into R:
  ical_parse_df() %>%
  # Convert to "tibble" data frame format:
  as_tibble() %>%
  # Use lubridate packge to wrangle dates and times:
  mutate(
    start_datetime = with_tz(start, tzone = "America/New_York"),
    end_datetime = with_tz(end, tzone = "America/New_York"),
    duration = end_datetime - start_datetime,
    date = floor_date(start_datetime, unit = "day")
  ) %>%
  # Convert calendar entry to all lowercase and rename:
  mutate(activity = tolower(summary)) %>%
  # Compute total duration of time for each day & activity:
  group_by(date, activity) %>%
  summarize(duration = sum(duration)) %>%
  # Convert duration to numerical variable and set units. Here hours:
  mutate(
    duration = as.numeric(duration),
    hours = duration / 60
  ) %>%
  # Filter out only rows where date is later than 2019-09-01:
  filter(date > "2019-09-01")
```

The (post-processed) data

Date	Activity	Duration	Hours
2019-09-02	Sleep	480	8.0
2019-09-02	Study	60	1.0
2019-09-03	Exercise	60	1.0
2019-09-04	Sleep	960	16.0
2019-09-04	Study	180	3.0
2019-09-05	Sleep	540	9.0
2019-09-06	Exercise	30	0.5
2019-09-06	Study	90	1.5
2019-09-07	Exercise	30	0.5
2019-09-07	Sleep	540	9.0

The deliverable: Analysis of time spent





The reflection piece prompts

- Reflect on the process of playing the whole game
- As someone who provides data:
  What expectations do you have when you give your data?
- As someone who analyzes other's data:
  What legal and ethical responsibilities do you have?

Quotes on technical difficulties

- "One of the technical issues we ran into, and probably the defining experience of this project, was the difficulty in creating consistent error free information to export to our partner. It's ridiculous how small manual entry errors, like whether I used spaces (in the calendar entries), made such a difference"
- "In our group, one member accidentally shared their entire calendar data with the other member when only a small portion of this data was needed for analysis. The other member realized the first member's mistake, deleted the data, and showed the first member how to separate the events she wanted to be analyzed from the rest of her calendar."

(Quotes edited here for brevity)

Quotes on ethical considerations

- "Additionally, the individual should have the right to ask about the research project that the data is going to and how their data may be used."
- "The question was straightforward but it also made us uneasy as we had to consider both what we were comfortable sharing and how to handle our partner's data in an ethical manner… This in turn motivated me to handle her data with care and sensitivity."
- "Information that puts our security or identity at risk should be left out. However, if too little information is shared, it may affect the ability to accurately represent any phenomena occurring within the data."

Some real 💎💎💎

- "We had to decide whether or not Friday would be considered a weekday or a weekend."
- "Since then I have been more cautious about granting apps access to my information."
- "Figure 1 made me realize that I should also spend more time during the weekdays sleeping and hanging out with friends to take care of myself."
- From Kat Correira's STAT 231 course, how one student spent their leisure time during the Spring 2020 pandemic semester

Questions

1. What's been most challenging to you as you developed these activities (Mine, then Albert)?
2. How have you approached assessment of the learning outcomes for these activities (Albert, then Mine)?
3. Do you have any advice for instructors planning to incorporate these activities into their courses (Mine, then Albert)?

Charities

1. Stop AAPI Hate: An organization ensuring that hate crimes on the Asian & Pacific Islander community do not go under-reported
2. Red Canary Song: A grassroots collective of Asian & migrant sex workers, organizing transnationally
3. Data for Black Lives: A movement of scientists & activists. Data as protest. Data as accountability. Data as collective action.