

Reading Set 4

Dhyey Mavani

Due by 10pm ET on Monday

Reading Set Information

A more thorough reading and light practice of the textbook reading prior to class allows us to jump into things more quickly in class and dive deeper into topics. As you actively read the textbook, you will work through the Reading Sets to help you engage with the new concepts and skills, often by replicating on your own the examples covered in the book.

These should be completed on your own without help from your peers. While most of our work in this class will be collaborative, it is important each individual completes the active readings. The problems should be straightforward based on the textbook readings, but if you have any questions, feel free to ask me!

GitHub Workflow

1. Before editing this file, verify you are working on the copy saved in *your* repo for the course (check the filepath and the project name in the top right corner).
2. Before editing this file, make an initial commit of the file to your repo to add your copy of the problem set.
3. Change your name at the top of the file and get started!
4. You should *save*, *knit*, and *commit* the .Rmd file each time you've finished a question, if not more often.
5. You should occasionally *push* the updated version of the .Rmd file back onto GitHub. When you are ready to push, you can click on the Git pane and then click **Push**. You can also do this after each commit in RStudio by clicking **Push** in the top right of the *Commit* pop-up window.
6. When you think you are done with the assignment, save the pdf as "*Name_thisfilename_date.pdf*" (it's okay to leave out the date if you don't need it) before committing and pushing (this is generally good practice but also helps me in those times where I need to download all student homework files).

Gradescope Upload

For each question (e.g., 3.1), allocate all pages associated with the specific question. If your work for a question runs onto a page that you did not select, you may not get credit for the work. If you do not allocate *any* pages when you upload your pdf, you may get a zero for the assignment.

You can resubmit your work as many times as you want before the deadline, so you should not wait until the last minute to submit some version of your work. Unexpected delays/crises that occur on the day the assignment is due do not warrant extensions (please submit whatever you have done to receive partial credit).

Problem 1 Web scraping In Section 6.4.1.2, the **rvest** package is used to scrape a Wikipedia page. BUT **WAIT!** While we may have the technical ability to scrape a webpage, that doesn't necessarily mean we are *allowed* to scrape it. **ETHICS ALERT!** *Before scraping a web page, you should always check whether doing so is allowed.* If you're unsure of the permissions for a particular domain, you can use the handy `paths_allowed()` function within the **robotstxt** package.

- 1.1 Check the permissions for the Wikipedia page using the code below. If the code returns "TRUE", then that indicates a bot has permission to access the page. Do you (via R) have permission to access the page?

```
library(robotstxt)
# Define url since we will use it again
url <- "https://en.wikipedia.org/wiki/Mile_run_world_record_progression"

# Check bot permissions
paths_allowed(url)
```

```
[1] TRUE
```

- 1.2 Now, use the code chunk below to follow along with the code in Section 6.4.1.2 to scrape the tables from the Wikipedia page on *Mile run world record progression*. Use `length(tables)` to identify how many tables are in the object you created called `tables`. How many tables are there?

```
library(rvest)
tables <- url %>%
  read_html() %>%
  html_nodes("table")

length(tables)
```

```
[1] 12
```

- 1.3 Next, look at the [Wikipedia page](#). We want to work with the table toward the bottom titled "Women Indoor IAAF era" shows four records: one for Mary Decker, two for Doina Melinte, and one for Genzebe Dibaba. From your `tables` object created in 1.2, create a dataframe called `women_indoor` that includes this "Women Indoor IAAF era" table data. *Hint:* You can use the same code as used in the textbook to create the `amateur` and `records` tables, except you'll need to update the table number that's plucked.

```
women_indoor <- tables %>%
  purrr::pluck(10) %>%
  html_table()
```

- 1.4 Use `kable()` to display the table from 1.3. Who holds the indoor one-mile world record for IAAF women, and what was her time?

Genzebe Dibaba from Ethiopia holds the indoor one-mile world record for IAAF women, and her time was 4:13.31

```
women_indoor %>%
  kable(booktabs = TRUE)
```

Time	Athlete	Nationality	Date	Venue
4:20.5	Mary Decker	United States	February 19, 1982	San Diego United States
4:18.86	Doina Melinte	Romania	February 13, 1988	East Rutherford United States
4:17.14	Doina Melinte	Romania	February 9, 1990	East Rutherford United States
4:13.31	Genzebe Dibaba	Ethiopia	February 17, 2016	Stockholm Sweden

- 1.5 Create a dataframe called `women_outdoor` that contains the table for “Women’s IAAF era” (starting with Anne Smith’s record and ending with Sifan Hassan’s record). Combine `women_indoor` and `women_outdoor` into one dataframe called `women_records` using the `bind_rows()` function. Include a variable called `Type` in this new dataframe to indicate whether a particular observation corresponds to an indoor record or an outdoor record (*hint*: create `Type` separately in each dataframe before combining). Finally, arrange `women_records` by ascending time, drop the `Venue` variable, and display the table using `kable()`. Who holds the fastest record, and was it from an indoor or outdoor event?

Sifan Hassan from Netherlands holds the fastest record outdoor.

```
women_outdoor <- tables %>%
  purrr::pluck(8) %>%
  html_table() %>%
  select(-Auto) # remove unwanted column

women_outdoor <- women_outdoor %>%
  mutate("Type" = "outdoor")

women_indoor <- women_indoor %>%
  mutate("Type" = "indoor")

women_records <- bind_rows(women_outdoor, women_indoor)

women_records <- women_records %>%
  arrange(Time)
```

```
women_records$Venue <- NULL
```

```
women_records %>%  
  kable(longtable = TRUE)
```

Time	Athlete	Nationality	Date	Type
4:12.33	Sifan Hassan	Netherlands	12 July 2019	outdoor
4:12.56	Svetlana Masterkova	Russia	14 August 1996[9]	outdoor
4:13.31	Genzebe Dibaba	Ethiopia	February 17, 2016	indoor
4:15.61	Paula Ivan	Romania	10 July 1989[9]	outdoor
4:16.71	Mary Decker-Slaney	United States	21 August 1985[9]	outdoor
4:17.14	Doina Melinte	Romania	February 9, 1990	indoor
4:17.44	Maricica Puică	Romania	9 September 1982[9]	outdoor
4:18.08	Mary Decker-Tabb	United States	9 July 1982[9]	outdoor
4:18.86	Doina Melinte	Romania	February 13, 1988	indoor
4:20.5	Mary Decker	United States	February 19, 1982	indoor
4:20.89	Lyudmila Veselkova	Soviet Union	12 September 1981[9]	outdoor
4:21.7	Mary Decker	United States	26 January 1980[9]	outdoor
4:22.1	Natalia Mărașescu	Romania	27 January 1979[9]	outdoor
4:23.8	Natalia Mărașescu	Romania	21 May 1977[9]	outdoor
4:29.5	Paola Pigni	Italy	8 August 1973[9]	outdoor
4:35.3	Ellen Tittel	West Germany	20 August 1971[9]	outdoor
4:36.8	Maria Gommers	Netherlands	14 June 1969[9]	outdoor
4:37.0	Anne Smith	United Kingdom	3 June 1967[9]	outdoor

Problem 2 As we wrap up the chapter on ethics, what are three major takeaways from Chapter 8 that had an impact on how you think about approaching your work as a budding data scientist?

The first takeaway for me is to protect the privacy and security of individuals represented in our data because we as data analysts or data scientists don't have the legal right to own the data unless and until the individual who generated the data signs an agreement with you. So, as a budding data scientist, I realize the responsibility to maintain the confidentiality of data while making sure that I don't alter it for any kind of personal or professional incentive as that would be illegal and against the data science ethics.

The second takeaway for me is to recognize and mitigate bias in ourselves and in the data we use because we all have limitations with human body and mind in terms of accuracy, so we might not be right and most efficient every time. For this collaboration to be possible, we should create reproducible and extensible work. Furthermore, be open to changing our methods and conclusions in response to new knowledge. This comes with respecting and inviting fair criticism while promoting the identification and open discussion of errors, risks, and unintended consequences of our work.

My third and final takeaway for data science ethics is to present our work in ways that empower others to make better-informed decisions. We should not hide information or modify the portrayal of data in accordance with our personal incentives because that will more often enter a negative bias than it will enter a positive bias.