# Reading Set 4

Kauila Tengan

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?

Yes, after using the `paths_allowed()` function in the **robotstxt** package, the bot returned "TRUE."

```
# 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?

There are 12 tables in the object `tables` we created.

```
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 `pluck`ed.

```
women_indoor <- tables %>%
  purrr::pluck(10) %>% #retrieves this specific data table from the wiki page
  html_table() #formats into the html wikipedia table form?

women_indoor
```

```
# A tibble: 4 x 5
  Time    Athlete        Nationality   Date              Venue
  <chr>   <chr>          <chr>         <chr>             <chr>
1 4:20.5  Mary Decker    United States February 19, 1982 San Diego  United Stat~
2 4:18.86 Doina Melinte  Romania       February 13, 1988 East Rutherford  Unite~
3 4:17.14 Doina Melinte  Romania       February 9, 1990  East Rutherford  Unite~
4 4:13.31 Genzebe Dibaba Ethiopia      February 17, 2016 Stockholm  Sweden
```

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?

According to our data table, the indoor women's one-mile world record is held by Genzebe Dibaba with a time of 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, with a time of 4:12.33

```
women_indoor2 <- women_indoor %>%
  select(-Venue) %>%  #remove Venue variable
  mutate(Type = "Indoors") #Assign indoor info for variable "Type"

women_outdoor <- tables %>%
  purrr::pluck(8) %>% #retrieves this specific data table from the wiki page
  html_table() %>% #formats into the html wikipedia table form?
  select(-Auto, -Venue) %>% #remove Venue variable
  mutate(Type = "Outdoors") #add variable "Type" for outdoors

women_records <- bind_rows(women_indoor2, women_outdoor) %>% #combine tables
  arrange(Time) # arrange time in ascending order

women_records %>%
  kable(longtable = TRUE) #make it aesthetically more pleasing
```

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

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 big thing that stuck out to me was the section about reproducibility. This summer I was tasked with updating 10 years worth of data in an excel spreadsheet for a journal. I had to update and recreate each of the figures, graphs, and tables. The last set of authors kept no trace of their work so I essentially had to do everything on my own with no experience in excel. With that said, I had difficulty conceptualizing how the next editors would be able to reproduce my work. This is a valuable lesson. Another major takeaway was spending a lot of time considering the ethical repercussions of my data and its analyses. On one hand the accuracy of it. Instead of producing something to fulfill a grant requirement that necessitates extensive wrangling to the point where it is no longer reproducible, that is unethical. Thirdly, even if the data comes out perfectly and statistically accurate, that does not necessarily correlate with being ethical. If my data, without much wrangling details a clear message that could disenfranchise groups, amplifying inequality, or impede on someone's privacy (like the Harvard study example), that is again not ethical.