Practice3

Sebastian Montesinos

Due by midnight, Friday, March 4

Reminder: Practice assignments may be completed working with other individuals.

Reading

The associated reading for the week is Sections 6.4, 8.5-8.7, and 8.9-8.10.

Practicing Academic Integrity

If you worked with others or used resources outside of provided course material (anything besides our textbook, course materials in the repo, labs, R help menu) to complete this assignment, please acknowledge them below using a bulleted list.

I acknowledge the following individuals with whom I worked on this assignment:

Name(s) and corresponding problem(s)

-

I used the following sources to help complete this assignment:

Source(s) and corresponding problem(s)

-

1 - Scraping Tables

The text example showed how to scrape tables from a Wikipedia page. We also saw how to scrape a table from basketball-reference.com in our lecture notes. For this exercise, your task is to:

- scrape a table of your choosing from a different website (yes, it can be a different Wikipedia page),
- clean it up (i.e. understandable variable names, etc. in a display), and
- display a few rows of it in a nice table.

You must be sure that scraping the table is allowed. Your code should show appropriate documentation of your steps.

Solution:

```
url <- "https://en.wikipedia.org/wiki/List_of_highest_scores_in_figure_skating#Men"
#Checking if I can scrape
paths_allowed(url)
```

```
##  en.wikipedia.org
```

```
## [1] TRUE
```

```
#Reading in the tables
mens_records <- url %>%
  read_html() %>%
  html_elements("table")

#Scraping a particular table
men_total <- mens_records %>%
  purrr::pluck(9) %>%
  html_table()

#Displaying the first few rows of the table
men_total_display <- head(men_total, 3) %>%
  kable(caption = "Top 3 Best Total Scores in Men's Ice Skating",
        booktabs = TRUE)
men_total_display
```

Table 1: Top 3 Best Total Scores in Men's Ice Skating

Rank	Name	Nation	Score	Event
1	Nathan Chen	United States	335.30	2019–20 Grand Prix Final
2	Yuzuru Hanyu	Japan	322.59	2019 Skate Canada
3	Yuma Kagiyama	Japan	310.05	2022 Winter Olympics

2 - MDSR 8.6

Complete MDSR 8.6, which states: "A Slate article (http://tinyurl.com/slate-ethics) discussed whether race/ethnicity should be included in a predictive model for how long a homeless family would stay in homeless services. Discuss the ethical considerations involved in whether race/ethnicity should be included as a predictor in the model."

Solution:

Since these algorithms are being used in ways that will directly effect people's material lives, we need to be very careful in considering whether to include race. On the one hand, race is an additional piece of information that could make the algorithm more predictive. However, including race means that one could bias the model against historically discriminated against minorities. Since these groups have faced systemic discrimination in the past, it is likely that they have had more trouble staying with and accessing homeless services. Therefore, including race in the model will probably output that people from these groups are less likely to stay in help programs. The issue with this result is that the factors that underlie this result would be the result of patterns of injustice, and such algorithms are blind to this pattern.

Given the potential for bias in algorithms that try to make such determinations, ethical practices in data science demand that special care be given to model testing and transparency in this process. This means that special consideration needs to be given to producing reproducible analyses that are available for others to access and scrutinize for possible 'proxies' for race. This will ensure that data scientists live up to the principles of minimizing and recognizing bias in their work.

3 - Scraping Text with Weather Data

We want to get a tiny bit of practice with the web developer tools demo-ed in class for scraping in this exercise.

Go to the National Weather Service website and get a forecast page for a city of your choice (maybe your hometown, or Amherst, or a place you want to visit in the States, etc.).

> part a - Save the url of the page as `weatherurl`. Then, check that you are allowed to access the page for scraping.

Solution:

```
# National Weather Service for Santa Fe, NM forecast
weatherurl <- "https://forecast.weather.gov/MapClick.php?CityName=Santa+Fe&state=NM&site=ABQ&textField1=

#Checking Permissions
paths_allowed(weatherurl)
```

```
## [1] TRUE
```

In class, we accessed the table of current conditions and the extended forecast temperatures for the Amherst page via text. Above but near the table of current conditions is information about the local site the conditions are taken from. This includes the latitude, longitude, and elevation of the site.

> part b - Adjust the commands demonstrated in class (used to get the extended forecast temperature information) to get these 3 pieces of information off your chosen page. Print the information to the screen from the website.

```
latlong <- weatherurl %>%
  read_html() %>%
#Get the span.smallTxt elements on the page
  html_elements("span.smallTxt") %>%
#Convert into a text format
  html_text()

#Print text
print(latlong)
```

```
## [1] "Lat: 35.61°NLon: 106.1°WElev: 6260ft."
```