Web scraping in R using the rvest package

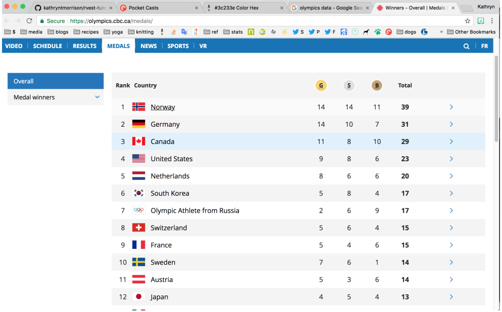
R Ladies Meetup

Kathryn Morrison, Co-Founder
Precision Analytics

March 15, 2018

What is web scraping and why is it useful?

- Say you want to download some data off a website to do some analysis:



Rank	Country	G	S	B	Total	
1	Norway	14	14	11	**39**	›
2	Germany	14	10	7	**31**	›
3	Canada	11	8	10	**29**	›
4	United States	9	8	6	**23**	›
5	Netherlands	8	6	6	**20**	›
6	South Korea	5	8	4	**17**	›
7	Olympic Athlete from Russia	2	6	9	**17**	›
8	Switzerland	5	6	4	**15**	›
9	France	5	4	6	**15**	›
10	Sweden	7	6	1	**14**	›
11	Austria	5	3	6	**14**	›
12	Japan	4	5	4	**13**	›

- You could manually copy and paste the data into a spreadsheet, but it would take a long time.
- Web scraping is a faster alternative to this

What is web scraping and why is it useful?

- Web scraping is basically automating the collection/download of unstructured data from online sources into a format you can use, e.g., plain text, csv, JSON, etc
- You can use many languages (e.g., python, java, *R*!)

Some popular approaches to web scraping:

1. Text pattern matching with regular expressions
2. Using an API
3. Parsing a website (what we'll do)

- I'll go through a simple parsing example
- If you're using an API or something more complex, you may need more technical skills/tools than rvest

Scraping in rvest

- Knowing some basics of HTML and CSS is helpful (but not required for the basics)
- You can use the SelectorGadget chrome extension to find the necessary tags

The rvest package was inspired by ‘Beautiful Soup’ in python

Rvest was created by Hadley Wickham (of course)

We’ll go through a really simple example