

Scraping websites with R

Using rvest and the tidyverse

Hadley Wickham

Chief Scientist, Posit

March 2024



Getting data off a site

Easy

~~1. Official API~~ This afternoon

2. Unofficial API

3. Static HTML

4. Dynamic HTML

Hard

1. HTML structure

2. rvest basics

3. Extracting data

4. Pagination

5. Live HTML

6. Unofficial APIs

HTML structure

HTML is a tree

```
<!doctype html>
<html lang="en-US">
  <head>
    <title>Page title</title>
    <meta ... >
    <script ... ></script>
    ...
  </head>
  <body>
    ...
  </body>
</html>
```

The tree is made up of elements

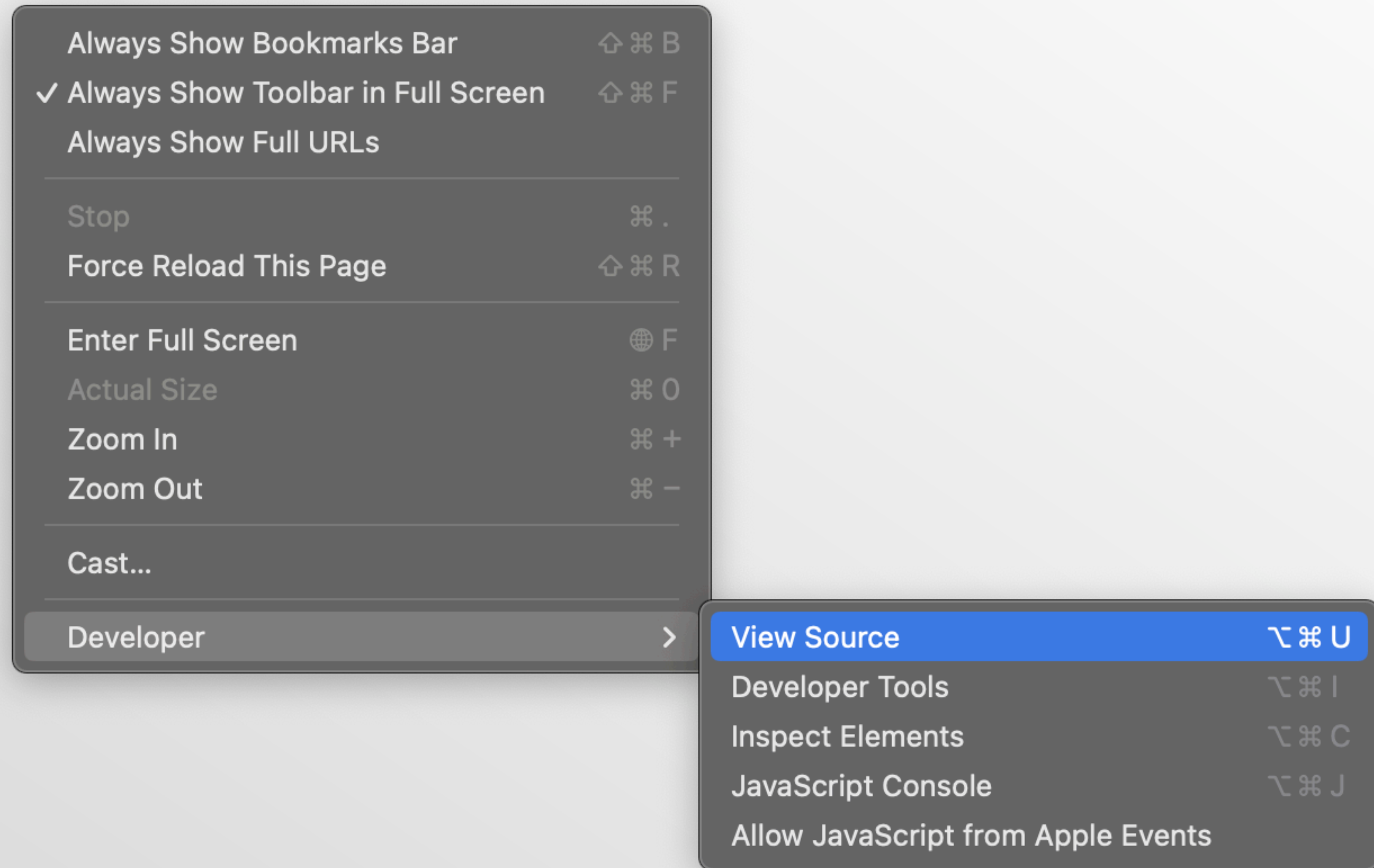
The diagram illustrates the structure of an HTML element using the example `<p class="nice">Hello world!</p>`. Brackets and labels identify the following parts:

- Opening tag**: A bracket above the `<p` portion of the code.
- Closing tag**: A bracket above the `</p>` portion of the code.
- An attribute and its value**: A bracket below the `class="nice"` portion of the code.
- Enclosed text content**: A bracket below the `Hello world!` text between the tags.

The stuff you see on a page comes from the body

```
<body>
  <h1>Top level heading</h1>
  <p>A paragraph containing text that is <b>bold</b> and
    an image: </p>
  <ul>
    <li>A bulleted list</li>
    <li>Bullet two</li>
  </ul>
  <table>
    <tr><th>A</th><th>B</th></tr>
    <tr><td>1</td><td>2</td></tr>
  </table>
</body>
```

Best way to see the tree of a real page is to use DevTools



Or right-click and choose inspect

Hoping to resolve the matter with a blockade of deadly battleships, the greedy Trade Federation then invaded the small planet of Naboo.

While the Confederacy debates this alarming chain of events, the Supreme Council has appointed two Jedi Knights, the guardians of peace and justice in the galaxy, to resolve the conflict....

Attack of the

Released: 2001

Director: George Lucas

Look Up "resolve"

Copy

Copy Link to Highlight

Search Google for "resolve"

Print...

Translate Selection to English

Open in Reading Mode **NEW**

Inspect

Speech >

Services >



Elements Console Sources Network Performance Memory Application Security

```
<!DOCTYPE html>
<!-- Generated by pkgdown: do not edit by hand -->
<html lang="en">
  ><head> ... </head>
  ▼<body> flex
    <div id="MathJax_Message" style="display: none;"></div>
    <a href="#container" class="visually-hidden-focusable">Skip to content</a>
    ▼<nav class="navbar fixed-top navbar-light navbar-expand-lg bg-none headroom headroom--top
      headroom--not-bottom"> flex
      ><div class="container"> ... </div> flex
      </nav>
      ><div class="container template-article" id="container"> ... </div> flex
      ><footer> ... </footer> flex
    </body>
  ...</html> == $0
```

Your turn

Go to <<https://rvest.tidyverse.org/articles/starwars.html>>, open the developer tools, and use the "elements" view to answer the following questions:

What element contains the film titles?

What element contains the name of the director? What attributes does this element have?

How many paragraphs does the “crawl” at the start of each movie have?

Where does the table of contents live relative to the film data?

Static vs dynamic

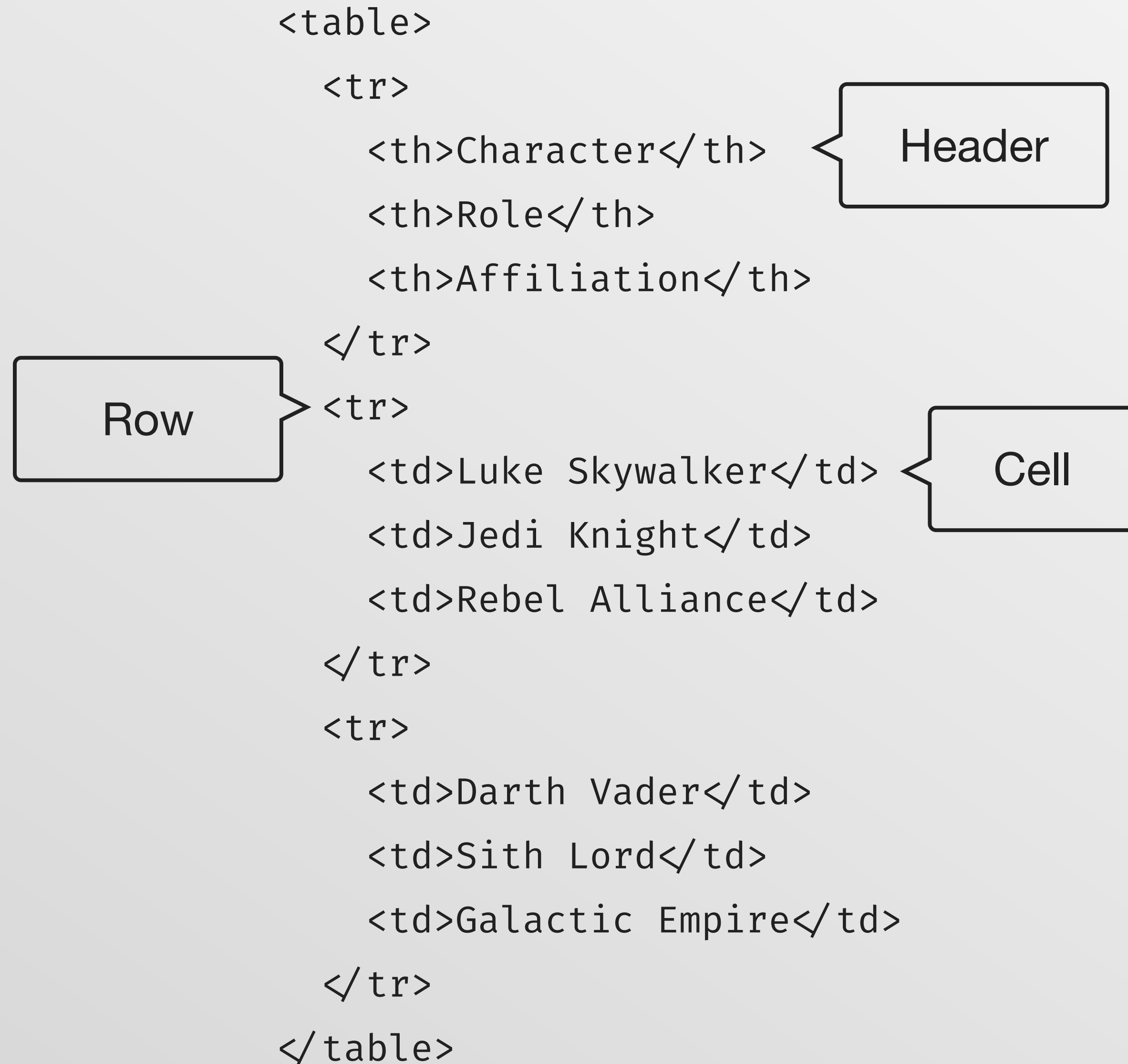
The HTML displayed in the source pane is usually generated from a HTML file that you can find in the sources pane.

Sometimes, however, the HTML is generated **dynamically** with Javascript. We'll come back to this later, but I wanted to illustrate the difference.

Let's look at <<https://rvest.tidyverse.org/dev/articles/starwars-dynamic.html>>

invest basics

Easiest to get data from HTML if it's already in a table



Can read with html_table()

```
library(rvest)
html ← minimal_html("<table>
  <tr>
    <th>Character</th>
    <th>Role</th>
    <th>Affiliation</th>
  </tr>
  <tr>
    <td>Luke Skywalker</td>
    <td>Jedi Knight</td>
    <td>Rebel Alliance</td>
  </tr>
</table>")
html_table(html)
```

Let's look at a more realistic example

```
# I want to get the sound track for Star Wars movie
```

```
url ← "https://en.wikipedia.org/wiki/Star\_Wars\_\(soundtrack\)"
```

```
html ← read_html(url)
```

```
html ▷
```

```
  html_table() ▷
```

```
  _[5:8]
```


Can we do better than asking for tables 5 through 8?

```
# I want to get the sound track for Star Wars movie
```

```
url ← "https://en.wikipedia.org/wiki/Star\_Wars\_\(soundtrack\)"
```

```
html ← read_html(url)
```

```
html ▷
```

```
  html_table() ▷
```

```
  _[5:8]
```

Your turn

Open <[https://en.wikipedia.org/wiki/Star Wars \(soundtrack\)](https://en.wikipedia.org/wiki/Star_Wars_soundtrack)>

Using the developer tools, can you find something in the structure of the HTML that uniquely identifies these tables?

We can use `html_elements()` with CSS selectors

```
url ← "https://en.wikipedia.org/wiki/Star_Wars_(soundtrack)"
```

```
html ← read_html(url)
```

```
html ▷
```

```
  html_elements(".tracklist") ▷
```

```
  html_table()
```

```
# .tracklist means all elements with class tracklist
```

CSS selectors

CSS = cascading style sheets

Primary purpose is to separate the visual appearance (style) from its underlying semantics.

Used to say (e.g.) “make this box blue” or “make all links green”.

It’s a domain specific language for selecting elements in the HTML tree. We’ll use it to identify the HTML elements that contain the data we care about.

Most important selectors

- `.brown` = all elements with class "brown"
- `#abc` = single element with id "abc"
- `p` = all paragraphs
- `p.important` = all paragraphs with "important" class
- `p b` = all bold elements that are descendants of a paragraph
- `p > b` = all bold elements that are children of a paragraph

Your turn

Practice your CSS selectors on <<https://flukeout.github.io/>>

Extracting data

1. Find the “rows” with `html_elements()`
2. Find the “columns” with `html_element()`
3. Extract the data with `html_text2()` or `html_attr()`
4. Make a tibble
5. Clean it up

Your turn

- Head back to <https://rvest.tidyverse.org/articles/starwars.html>
- What are the rows? How can you identify them with a css selector?
- What are the columns? How can you identify them with a css selector?

Solution

starwars.R

Your turn

- Go to <https://quotes.toscrape.com/>
- Identify the rows and columns (including the URL to the author page), and the selectors that will identify them.
- (What might you want to do with the tags?)
- Scrape into a tibble

Solution

quotes.R

html_elements() vs html_element()

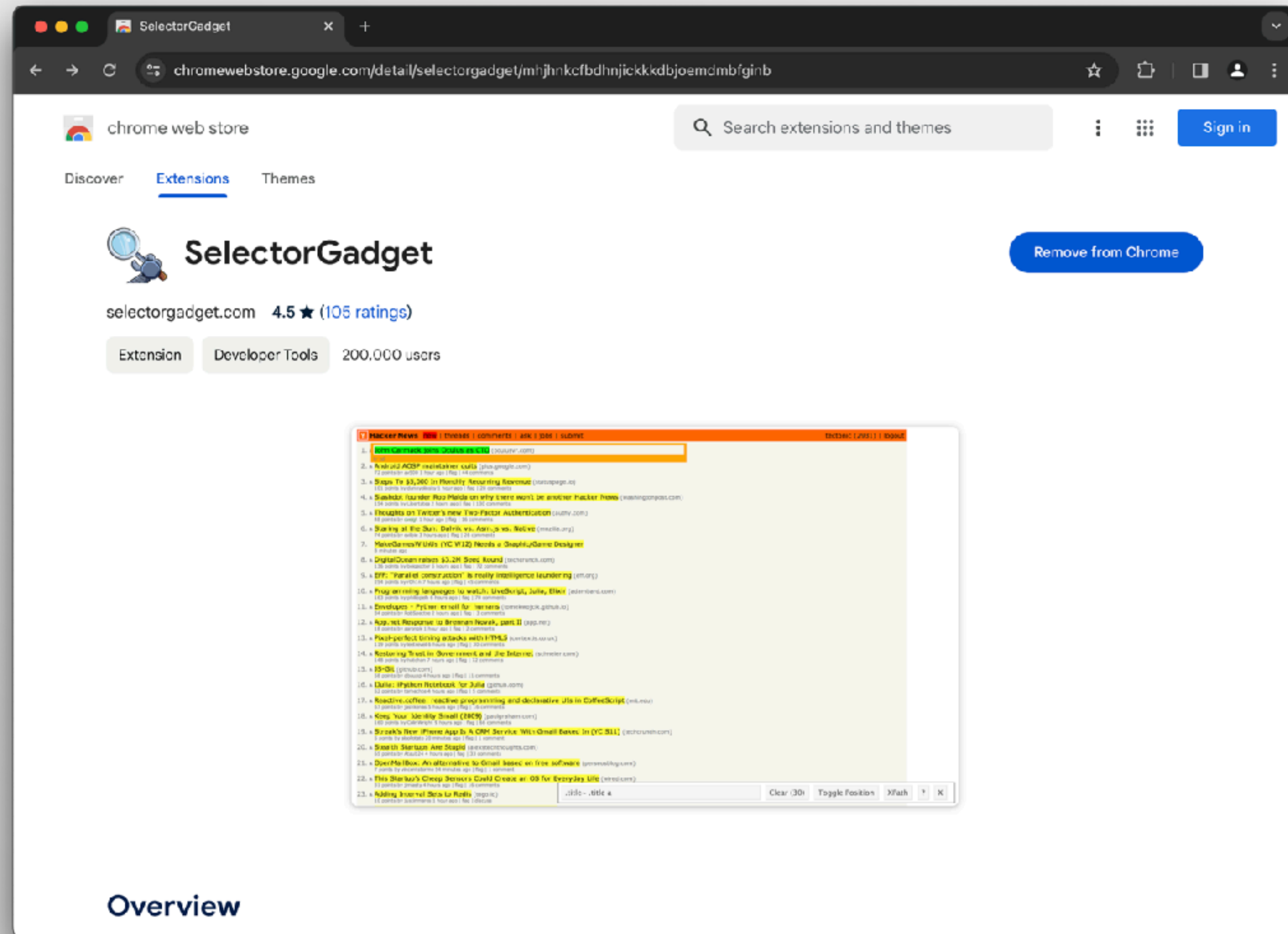
html_elements()	html_element()
n -> m	n -> n
length(0)	NA
Find rows	Find column in each row

Ways to find the selector

1. Directly inspecting the HTML
2. In DevTools, right-click & choose “Copy selector” (then simplify)
3. SelectorGadget

Google for selector gadget

https://chrome.google.com/webstore/detail/selectorgadget



More practice

Scrape all the books off <<http://books.toscrape.com/>>

Make sure to capture the (full) title, the rating, the path to the cover image and the price.

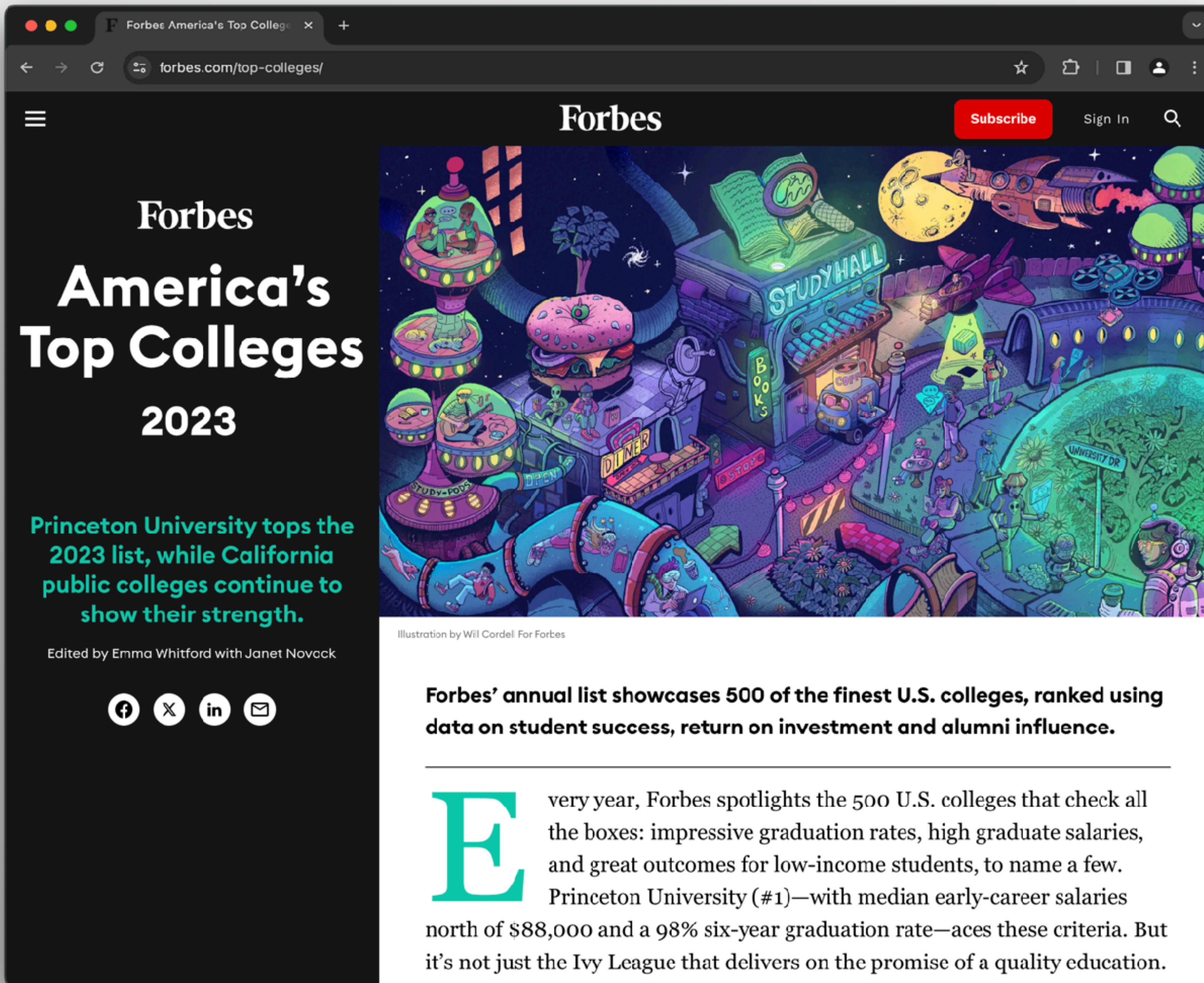
(Don't worry about the pagination yet)

Pagination

Demo

pagination.R

Live HTML



forbes.com/top-colleges/

Forbes

SubscribeSign In

OVERALL

STATE

PUBLIC/PRIVATE

SCHOOL SIZE

CAMPUS SETTING

SEARCH

RANK	NAME	STATE	TYPE	AV. GRANT AID	AV. DEBT	MEDIAN 10-YEAR SALARY	FINANCIAL GRADE
	Princeton University	NJ	Private not-for-profit	\$47,136	\$7,216	\$177,300	A+
	Yale University	CT	Private not-for-profit	\$58,715	\$4,968	\$163,900	A+
	Stanford University	CA	Private not-for-profit	\$56,211	\$8,868	\$173,800	A+
	Massachusetts Institute of Technology	MA	Private not-for-profit	\$32,562	\$7,235	\$182,800	A+
	University of California, Berkeley	CA	Public	\$21,406	\$7,202	\$161,300	
	Columbia University	NY	Private not-for-profit	\$57,726	\$13,338	\$150,900	A+
	University of California, Los Angeles	CA	Public	\$17,592	\$5,965	\$140,300	
	University of Pennsylvania	PA	Private not-for-profit	\$50,778	\$10,510	\$165,700	A+
	Harvard	MA	Private not-for-	\$59,840	\$8,700	\$169,500	A+

Top Colleges 2023: Why California's Public Universities Are So Good

It's no surprise that Princeton and Harvard made the top 25 colleges, but despite California's woes, four of its public universities did so too.

READ MORE >

Your turn

Where does the data live? What defines the rows and the columns? (Hint: it looks like a table, but it's not)

What happens if you try to read this data into R?

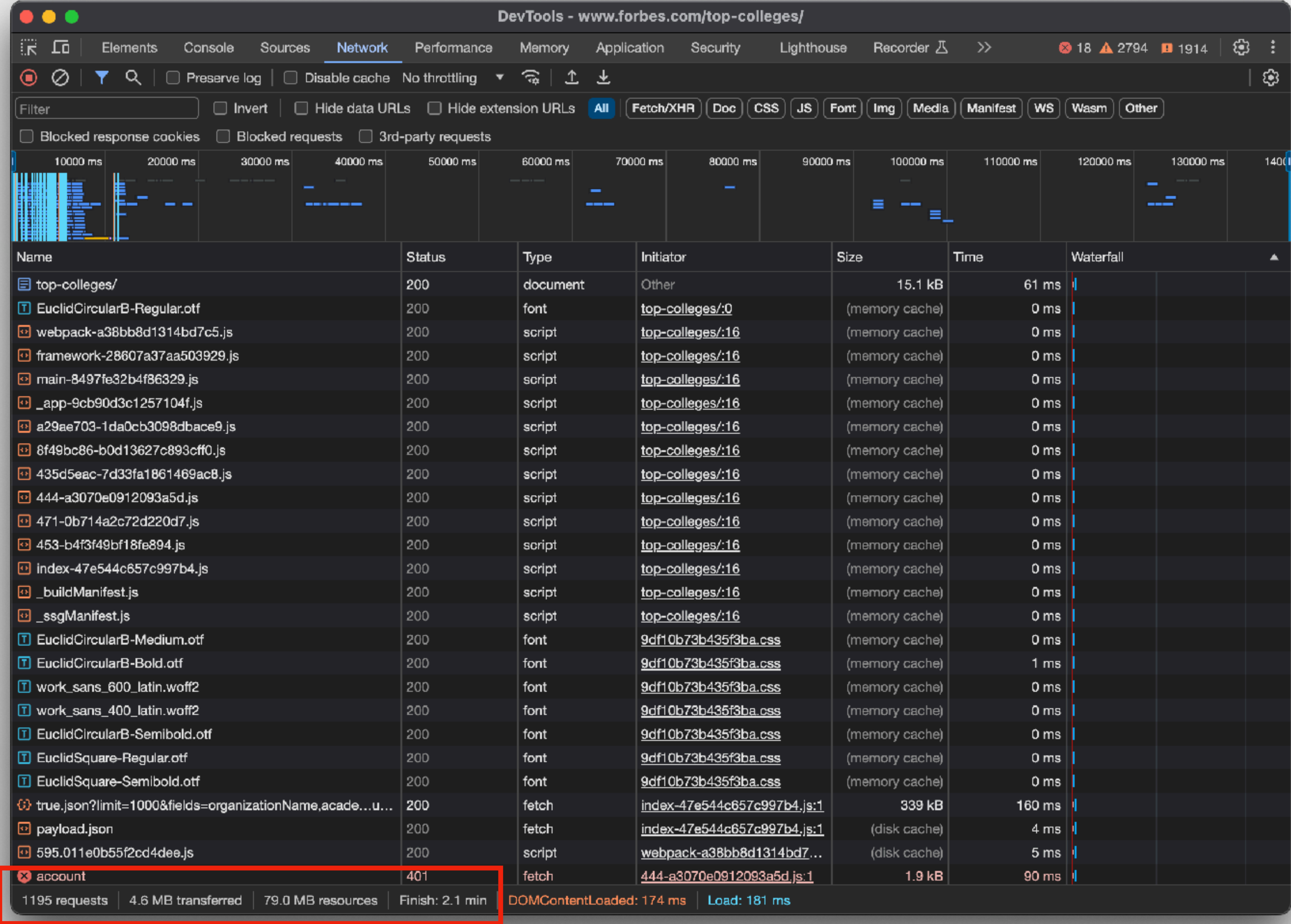
Solution

`forbes-live.R`

Unofficial API

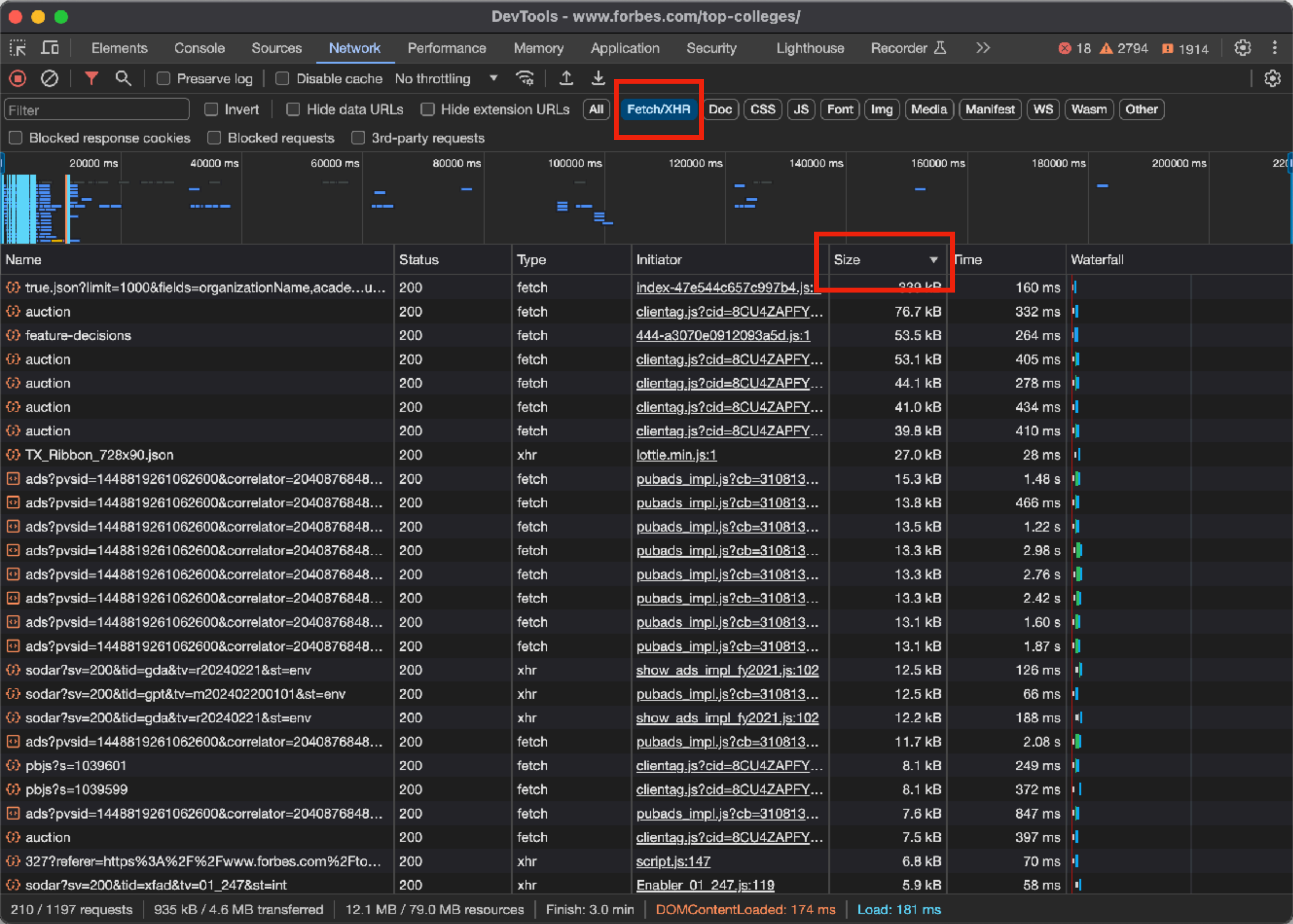
Unofficial API

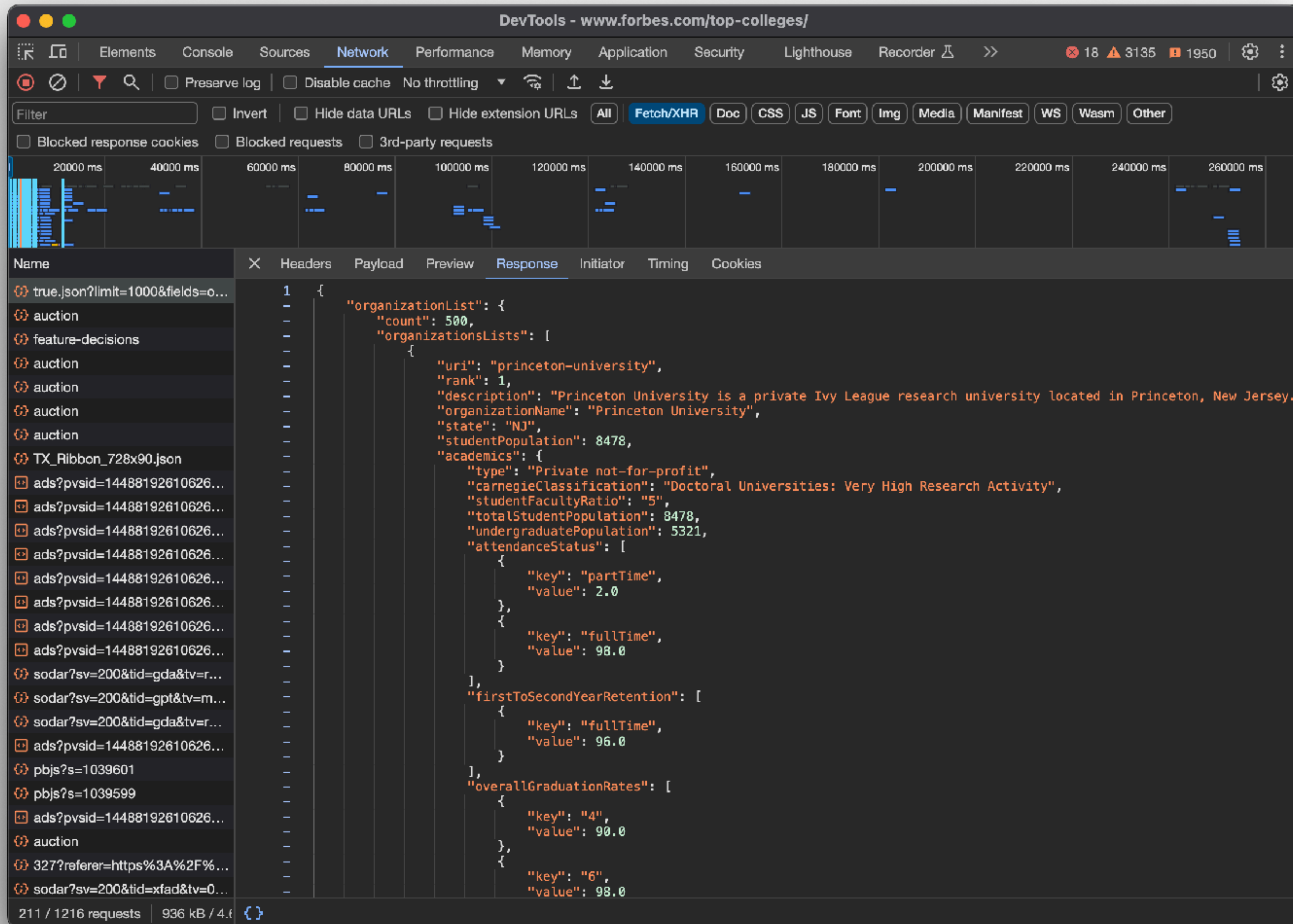
- Most websites that dynamically generate HTML do so from a JSON file.
- If you can find that JSON file you can work with it directly, making your life much easier.
- Find it with network pane in the browser developer tools.
- It's often obvious, but even when not it's worth spending 30 minutes on because it'll easily save that much time.
- (I think this Forbes site is an outlier; most of the time it will be a bit easier.)



Two useful heuristics to start with

- Filter to “Fetch/XHR”
- Sort by size (decreasing)





DevTools - www.forbes.com/top-colleges/

Elements Console Sources **Network** Performance Memory Application Security Lighthouse Recorder >> 18 3135 1950

Filter ☐ Preserve log ☐ Disable cache No throttling ☐ Blocked response cookies ☐ Blocked requests ☐ 3rd-party requests

☐ Invert ☐ Hide data URLs ☐ Hide extension URLs All Fetch/XHR Doc CSS JS Font Img Media Manifest WS Wasm Other

20000 ms 40000 ms 60000 ms 80000 ms 100000 ms 120000 ms 140000 ms 160000 ms 180000 ms 200000 ms 220000 ms 240000 ms 260000 ms 280000 ms 300000 ms 320000 ms 340000 ms

Name X Headers Payload Preview **Response** Initiator Timing Cookies

true.json?limit=1000&fields=o...
auction
feature-decisions
auction
auction
auction
auction
TX_Ribbon_728x90.json
ads?pvsid=14488192610626...
ads?pvsid=14488192610626...
ads?pvsid=14488192610626...
ads?pvsid=14488192610626...
ads?pvsid=14488192610626...
ads?pvsid=14488192610626...
ads?pvsid=14488192610626...
ads?pvsid=14488192610626...
sodar?sv=200&tid=gda&tv=r...
sodar?sv=200&tid=gpt&tv=m...
sodar?sv=200&tid=gda&tv=r...
ads?pvsid=14488192610626...
pbjs?s=1039601
pbjs?s=1039599
ads?pvsid=14488192610626...
auction
327?referer=https%3A%2F%...
sodar?sv=200&tid=xfad&tv=0...

212 / 1218 requests | 936 kB / 4.6

Open in Sources panel
Open in new tab
Clear browser cache
Clear browser cookies
Copy
Block request URL
Block request domain
Sort By
Header Options
Override headers
Override content
Show all overrides
Save all as HAR with content

Copy URL
Copy as cURL
Copy as PowerShell
Copy as fetch
Copy as fetch (Node.js)
Copy response
Copy stack trace
Copy all URLs
Copy all as cURL
Copy all as PowerShell
Copy all as fetch
Copy all as fetch (Node.js)
Copy all as HAR

This gives a giant curl call

```
curl 'https://www.forbes.com/forbesapi/org/top-colleges/2023/position/true.json?
limit=1000&fields=organizationName,academics,state,financialAid,rank,medianBaseSalary,campusSetting,studentPopulation,squareImage,uri,description,grade' \
-X 'GET' \
-H 'Accept: */*' \
-H 'Sec-Fetch-Site: same-origin' \
-H 'Cookie: _ga_DLD85VJ5QY=GS1.1.1706542437.3.1.1706542642.60.0.0; VW0=27.700;
_ketch_consent_v1_=eyJiZWdhdmFvcmFsX2FkdGVydGlzaW5nIjpb7InN0YXR1cyI6ImdyYW50ZWQiLCJjYW5vbmljYWxQdXJwb3NlcyI6WyJiZWdhdmFvcmFsX2FkdGVydGlzaW5nIl19LCJhbmFseXRpY3MiO3RhZHVzIjoiZ3JhbnRlZCI6ImNhbm9uaWNhb
FB1cnBvc2VzIjpbImFuYWx5dGljcyJdfSwiZnVuY3Rpb25hbCI6eyJzdGF0dXMiOiJncmFudGVkIiwieWY2Fub25pY2F5UHVycG9zZXMiOlsicHJvZF9lbmhhbmNlbWVudCI6ImNhbm9uaWNhbFBIcnBvc2VzIjpbImVzc2VudGlhbF9zZXJ2aWNlcyJdfX0%3D;
_swb_consent_=eyJlbnZpcm9ubWVudENvZGUiOiJwcm9kdWN0aW9uIiwiaWRlbnRpdGllcyI6eyJfZ29vZ2xlQW5hbHl0aWNzQ2xpZW50SUQiOiJHQTUuMi4zNTcxMjI4NjguMTcwNTI10TE4OCIsInN3Yl93ZWJzaXRlX3NtYXJ0X3RhZyI6IjU3YzY4NDNhLWQ4ZT
ktNDBjMy05YmMxLTJhYWNkZWU3ZDE2OSJ9LCJqdXJpc2RpbY3Rpb25Db2RlIjoidXNfZ2VuZXJhbCI6ImNhbm9uaWNzQ2xpZW50SUQiOiJHQTUuMi4zNTcxMjI4NjguMTcwNTI10TE4OCIsInN3Yl93ZWJzaXRlX3NtYXJ0X3RhZyI6IjU3YzY4NDNhLWQ4ZT
UiOiJkaXNjbG9zdXJlIn0sImJlaGF2aW9yYWx5fWY2ZXJ0aXNpbmciO3RhZHVzIjoiZ3JhbnRlZCI6ImNhbm9uaWNhbFBIcnBvc2VzIjpbImVzc2VudGlhbF9zZXJ2aWNlcyJdfX0%3D;
J9LCJyZXF1aXJlZCI6eyJhbmFseXRpY3MiO3RhZHVzIjoiZ3JhbnRlZCI6ImNhbm9uaWNhbFBIcnBvc2VzIjpbImVzc2VudGlhbF9zZXJ2aWNlcyJdfX0%3D;
_gcl_au=1.1.1366914331.1705344922; _gid=GA1.2.412386523.1706542438; us_privacy=1---; usprivacy=1---; AWSALB=tnC2CAJGdOK0IsOKzBarRIkeT8r4YS5/
wR6+n+A2VQX2GfXa3lgP2KrEv6bGPRZyGIhGFvSCrNCftxLR8EXyMI3eDVjI5kBMlcIv9BthsZsyM9Vp0eTKR5yeR/H; AWSALBCORS=tnC2CAJGdOK0IsOKzBarRIkeT8r4YS5/
wR6+n+A2VQX2GfXa3lgP2KrEv6bGPRZyGIhGFvSCrNCftxLR8EXyMI3eDVjI5kBMlcIv9BthsZsyM9Vp0eTKR5yeR/H; BCSessionID=2ae07b14-70be-40b1-9e90-ed5d91b1be4f; ki_r=;
ki_t=1706121218201%3B1706542438105%3B1706542587089%3B2%3B12; client_id=14d486f22e5f65b3107348cd0ffeaa50923; lux_uid=170654243670875530; AMP_TOKEN=%24NOT_FOUND; _swb=57c6843a-
d8e9-40c3-9bc1-8cacdee7d169; rbzid=CXwv4IPK0a5WMSn9vGfFx/FfHDJXwSL2pHJ/swUFnuvHRpgd1qGWgsSHDvkfxbe6A3W7IDkaJbmIeiPvd/wC7NLDIMS6nZ4N6B7HWA1lvWFqWciQ/+GZ1Bm7YCvrGGhX059ttvv6mZYAXbx6MHiL6+/
BVKFl8m2Z5gx0M2r0M2j9D/QiW7bH4TR8S/oJmWPQi7TJT5F+3SMGf6SjtfG64f74hLDwf+zEy5y95JETCHlw70g8eg5Q05AALhKK+rhPV4z7F29XqrM2aGXgKHz//+Q==; rbzsessionid=da3b925ab2cf238acdfa18f6e699045b;
_ga_HY3LZWHH6W=GS1.1.1705344921.1.1.1705345703.0.0.0; _uetvid=a2f10940b3d711ee8c1881f6a3138e42; amp_9c5697=N1292876525 ... 1hk77ksom.1hk77kvfo.2.2.4; notice_behavior=implied,us; fadve2etidvcnt=2;
_clck=10sfr1j%7C2%7Cfif%7C0%7C1474; fadve2etid=N1292876525; fadvfpuid=FA77ce891e24882d9a03e9d2bc5bf16cf3; _ga_0Y2Y7WWQP1=GS1.1.1705259187.1.1.1705259215.0.0.0;
_ga_JFZ3B3QM86=GS1.1.1705259187.1.1.1705259215.0.0.0; cmapi_cookie_privacy=permit 1,2,3; fadvuke2etid=N255829421; blaize_session=d72df655-7d08-4673-bbd3-beadbe316b40;
blaize_tracking_id=418dd2e2-817f-48c7-9792-e36d6ce05bf9' \
-H 'Sec-Fetch-Dest: empty' \
-H 'Accept-Language: en-GB,en-US;q=0.9,en;q=0.8' \
-H 'Sec-Fetch-Mode: cors' \
-H 'Host: www.forbes.com' \
-H 'User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/17.2.1 Safari/605.1.15' \
-H 'Referer: https://www.forbes.com/top-colleges/' \
-H 'Accept-Encoding: gzip, deflate, br' \
```

And you can translate to R with `httr2::curl_translate()`

```
request("https://www.forbes.com/forbesapi/org/top-colleges/2023/position/true.json") ➤

req_method("GET") ➤

req_url_query(
  limit = "1000",
  fields = "organizationName,academics,state,financialAid,rank,medianBaseSalary,campusSetting,studentPopulation,squareImage,uri,description,grade",
) ➤

req_headers(
  Accept = "*/*",
  Cookie = "_ga_DLD85VJ5QY=GS1.1.1706542437.3.1.1706542642.60.0.0; VW0=27.700;
_ketch_consent_v1_=eyJiZWhhdmFvcmFsX2FkdVYdGlzaW5nIjpb7InN0YXR1cyI6ImdyYW50ZWQiLCJjYW5vbmljYWxQdXJwb3NlcyI6WyJiZWhhdmFvcmFsX2FkdVYdGlzaW5nIl19LCJhbmFseXRpY3MiOnsic3RhdHVzIjoZ3JhbnRLZCIsImNhbm9uaWNhb
FB1cnBvc2VzIjpbImFuYWx5dGljcyJdfSwiZnVuY3Rpb25hbCI6eyJzdGF0dXMiOiJncmFudGVkIiwieY2Fub25pY2F2SUHVycG9zZXMiOlsicHJvZF9lbmhhbmNlbWVudCIsInBlcnNvbFsaXphdGlviJdfSwicmVxdWlyZWQiOnsic3RhdHVzIjoZ3JhbnRLZCIsI
mNhbm9uaWNhbFB1cnBvc2VzIjpbImVzc2VudGlhbF9zZXJ2aWNlcyJdfX0%3D;
_swb_consent_=eyJlbnZpcm9ubWVudENvZGUiOiJwcm9kdWN0aW9uIiwiaWRlbnRpdGllcyI6eyJfZ29vZ2xlQW5hbHl0aWNzQ2xpZW50SUQiOiJHQTUuMi4zNTcxMjI4NjguMTcwNTI10TE4OCIsInN3Yl93ZWJzaXRlX3NtYXJ0X3RhZyI6IjU3YzY4NDNhLWQ4ZT
ktNDBjMy05YmMxLTJhYWNkZWU3ZDE2OSJ9LCJqdXJpc2RpbY3Rpb25Db2RlIjoZ3JhbnRLZCIsImNhbm9uaWNhbFB1cnBvc2VzIjpbImFuYWx5dGljcyJdfSwiZnVuY3Rpb25hbCI6eyJzdGF0dXMiOiJncmFudGVkIiwieY2Fub25pY2F2SUHVycG9zZXMiOlsicHJvZF9lbmhhbmNlbWVudCIsInBlcnNvbFsaXphdGlviJdfSwicmVxdWlyZWQiOnsic3RhdHVzIjoZ3JhbnRLZCIsI
mNhbm9uaWNhbFB1cnBvc2VzIjpbImVzc2VudGlhbF9zZXJ2aWNlcyJdfX0%3D;
_ketch_consent_v1_=eyJiZWhhdmFvcmFsX2FkdVYdGlzaW5nIjpb7InN0YXR1cyI6ImdyYW50ZWQiLCJjYW5vbmljYWxQdXJwb3NlcyI6WyJiZWhhdmFvcmFsX2FkdVYdGlzaW5nIl19LCJhbmFseXRpY3MiOnsic3RhdHVzIjoZ3JhbnRLZCIsImNhbm9uaWNhb
FB1cnBvc2VzIjpbImFuYWx5dGljcyJdfSwiZnVuY3Rpb25hbCI6eyJzdGF0dXMiOiJncmFudGVkIiwieY2Fub25pY2F2SUHVycG9zZXMiOlsicHJvZF9lbmhhbmNlbWVudCIsInBlcnNvbFsaXphdGlviJdfSwicmVxdWlyZWQiOnsic3RhdHVzIjoZ3JhbnRLZCIsI
mNhbm9uaWNhbFB1cnBvc2VzIjpbImVzc2VudGlhbF9zZXJ2aWNlcyJdfX0%3D;
_swb_consent_=eyJlbnZpcm9ubWVudENvZGUiOiJwcm9kdWN0aW9uIiwiaWRlbnRpdGllcyI6eyJfZ29vZ2xlQW5hbHl0aWNzQ2xpZW50SUQiOiJHQTUuMi4zNTcxMjI4NjguMTcwNTI10TE4OCIsInN3Yl93ZWJzaXRlX3NtYXJ0X3RhZyI6IjU3YzY4NDNhLWQ4ZT
ktNDBjMy05YmMxLTJhYWNkZWU3ZDE2OSJ9LCJqdXJpc2RpbY3Rpb25Db2RlIjoZ3JhbnRLZCIsImNhbm9uaWNhbFB1cnBvc2VzIjpbImFuYWx5dGljcyJdfSwiZnVuY3Rpb25hbCI6eyJzdGF0dXMiOiJncmFudGVkIiwieY2Fub25pY2F2SUHVycG9zZXMiOlsicHJvZF9lbmhhbmNlbWVudCIsInBlcnNvbFsaXphdGlviJdfSwicmVxdWlyZWQiOnsic3RhdHVzIjoZ3JhbnRLZCIsI
mNhbm9uaWNhbFB1cnBvc2VzIjpbImVzc2VudGlhbF9zZXJ2aWNlcyJdfX0%3D;
_ketch_consent_v1_=eyJiZWhhdmFvcmFsX2FkdVYdGlzaW5nIjpb7InN0YXR1cyI6ImdyYW50ZWQiLCJjYW5vbmljYWxQdXJwb3NlcyI6WyJiZWhhdmFvcmFsX2FkdVYdGlzaW5nIl19LCJhbmFseXRpY3MiOnsic3RhdHVzIjoZ3JhbnRLZCIsImNhbm9uaWNhb
FB1cnBvc2VzIjpbImFuYWx5dGljcyJdfSwiZnVuY3Rpb25hbCI6eyJzdGF0dXMiOiJncmFudGVkIiwieY2Fub25pY2F2SUHVycG9zZXMiOlsicHJvZF9lbmhhbmNlbWVudCIsInBlcnNvbFsaXphdGlviJdfSwicmVxdWlyZWQiOnsic3RhdHVzIjoZ3JhbnRLZCIsI
mNhbm9uaWNhbFB1cnBvc2VzIjpbImVzc2VudGlhbF9zZXJ2aWNlcyJdfX0%3D;
_swb_consent_=eyJlbnZpcm9ubWVudENvZGUiOiJwcm9kdWN0aW9uIiwiaWRlbnRpdGllcyI6eyJfZ29vZ2xlQW5hbHl0aWNzQ2xpZW50SUQiOiJHQTUuMi4zNTcxMjI4NjguMTcwNTI10TE4OCIsInN3Yl93ZWJzaXRlX3NtYXJ0X3RhZyI6IjU3YzY4NDNhLWQ4ZT
ktNDBjMy05YmMxLTJhYWNkZWU3ZDE2OSJ9LCJqdXJpc2RpbY3Rpb25Db2RlIjoZ3JhbnRLZCIsImNhbm9uaWNhbFB1cnBvc2VzIjpbImFuYWx5dGljcyJdfSwiZnVuY3Rpb25hbCI6eyJzdGF0dXMiOiJncmFudGVkIiwieY2Fub25pY2F2SUHVycG9zZXMiOlsicHJvZF9lbmhhbmNlbWVudCIsInBlcnNvbFsaXphdGlviJdfSwicmVxdWlyZWQiOnsic3RhdHVzIjoZ3JhbnRLZCIsI
mNhbm9uaWNhbFB1cnBvc2VzIjpbImVzc2VudGlhbF9zZXJ2aWNlcyJdfX0%3D;
_ketch_consent_v1_=eyJiZWhhdmFvcmFsX2FkdVYdGlzaW5nIjpb7InN0YXR1cyI6ImdyYW50ZWQiLCJjYW5vbmljYWxQdXJwb3NlcyI6WyJiZWhhdmFvcmFsX2FkdVYdGlzaW5nIl19LCJhbmFseXRpY3MiOnsic3RhdHVzIjoZ3JhbnRLZCIsImNhbm9uaWNhb
FB1cnBvc2VzIjpbImFuYWx5dGljcyJdfSwiZnVuY3Rpb25hbCI6eyJzdGF0dXMiOiJncmFudGVkIiwieY2Fub25pY2F2SUHVycG9zZXMiOlsicHJvZF9lbmhhbmNlbWVudCIsInBlcnNvbFsaXphdGlviJdfSwicmVxdWlyZWQiOnsic3RhdHVzIjoZ3JhbnRLZCIsI
mNhbm9uaWNhbFB1cnBvc2VzIjpbImVzc2VudGlhbF9zZXJ2aWNlcyJdfX0%3D;
_swb_consent_=eyJlbnZpcm9ubWVudENvZGUiOiJwcm9kdWN0aW9uIiwiaWRlbnRpdGllcyI6eyJfZ29vZ2xlQW5hbHl0aWNzQ2xpZW50SUQiOiJHQTUuMi4zNTcxMjI4NjguMTcwNTI10TE4OCIsInN3Yl93ZWJzaXRlX3NtYXJ0X3RhZyI6IjU3YzY4NDNhLWQ4ZT
ktNDBjMy05YmMxLTJhYWNkZWU3ZDE2OSJ9LCJqdXJpc2RpbY3Rpb25Db2RlIjoZ3JhbnRLZCIsImNhbm9uaWNhbFB1cnBvc2VzIjpbImFuYWx5dGljcyJdfSwiZnVuY3Rpb25hbCI6eyJzdGF0dXMiOiJncmFudGVkIiwieY2Fub25pY2F2SUHVycG9zZXMiOlsicHJvZF9lbmhhbmNlbWVudCIsInBlcnNvbFsaXphdGlviJdfSwicmVxdWlyZWQiOnsic3RhdHVzIjoZ3JhbnRLZCIsI
mNhbm9uaWNhbFB1cnBvc2VzIjpbImVzc2VudGlhbF9zZXJ2aWNlcyJdfX0%3D;
_ketch_consent_v1_=eyJiZWhhdmFvcmFsX2FkdVYdGlzaW5nIjpb7InN0YXR1cyI6ImdyYW50ZWQiLCJjYW5vbmljYWxQdXJwb3NlcyI6WyJiZWhhdmFvcmFsX2FkdVYdGlzaW5nIl19LCJhbmFseXRpY3MiOnsic3RhdHVzIjoZ3JhbnRLZCIsImNhbm9uaWNhb
FB1cnBvc2VzIjpbImFuYWx5dGljcyJdfSwiZnVuY3Rpb25hbCI6eyJzdGF0dXMiOiJncmFudGVkIiwieY2Fub25pY2F2SUHVycG9zZXMiOlsicHJvZF9lbmhhbmNlbWVudCIsInBlcnNvbFsaXphdGlviJdfSwicmVxdWlyZWQiOnsic3RhdHVzIjoZ3JhbnRLZCIsI
mNhbm9uaWNhbFB1cnBvc2VzIjpbImVzc2VudGlhbF9zZXJ2aWNlcyJdfX0%3D;
_swb_consent_=eyJlbnZpcm9ubWVudENvZGUiOiJwcm9kdWN0aW9uIiwiaWRlbnRpdGllcyI6eyJfZ29vZ2xlQW5hbHl0aWNzQ2xpZW50SUQiOiJHQTUuMi4zNTcxMjI4NjguMTcwNTI10TE4OCIsInN3Yl93ZWJzaXRlX3NtYXJ0X3RhZyI6IjU3YzY4NDNhLWQ4ZT
ktNDBjMy05YmMxLTJhYWNkZWU3ZDE2OSJ9LCJqdXJpc2RpbY3Rpb25Db2RlIjoZ3JhbnRLZCIsImNhbm9uaWNhbFB1cnBvc2VzIjpbImFuYWx5dGljcyJdfSwiZnVuY3Rpb25hbCI6eyJzdGF0dXMiOiJncmFudGVkIiwieY2Fub25pY2F2SUHVycG9zZXMiOlsicHJvZF9lbmhhbmNlbWVudCIsInBlcnNvbFsaXphdGlviJdfSwicmVxdWlyZWQiOnsic3RhdHVzIjoZ3JhbnRLZCIsI
mNhbm9uaWNhbFB1cnBvc2VzIjpbImVzc2VudGlhbF9zZXJ2aWNlcyJdfX0%3D;
_ketch_consent_v1_=eyJiZWhhdmFvcmFsX2FkdVYdGlzaW5nIjpb7InN0YXR1cyI6ImdyYW50ZWQiLCJjYW5vbmljYWxQdXJwb3NlcyI6WyJiZWhhdmFvcmFsX2FkdVYdGlzaW5nIl19LCJhbmFseXRpY3MiOnsic3RhdHVzIjoZ3JhbnRLZCIsImNhbm9uaWNhb
FB1cnBvc2VzIjpbImFuYWx5dGljcyJdfSwiZnVuY3Rpb25hbCI6eyJzdGF0dXMiOiJncmFudGVkIiwieY2Fub25pY2F2SUHVycG9zZXMiOlsicHJvZF9lbmhhbmNlbWVudCIsInBlcnNvbFsaXphdGlviJdfSwicmVxdWlyZWQiOnsic3RhdHVzIjoZ3JhbnRLZCIsI
mNhbm9uaWNhbFB1cnBvc2VzIjpbImVzc2VudGlhbF9zZXJ2aWNlcyJdfX0%3D;
_swb_consent_=eyJlbnZpcm9ubWVudENvZGUiOiJwcm9kdWN0aW9uIiwiaWRlbnRpdGllcyI6eyJfZ29vZ2xlQW5hbHl0aWNzQ2xpZW50SUQiOiJHQTUuMi4zNTcxMjI4NjguMTcwNTI10TE4OCIsInN3Yl93
```


Demo

`forbes-api.R`

If the data is stored as JSON in the HTML

This is a pain, but is fortunately relatively rare.

I have had some luck in the past with using the V8 R package to run the javascript code and then extract the JSON object back into R.