Big Data: Data Wrangling Boot Camp
Web Crawling with R and CSS

Chuck Cartledge, PhD

25 February 2018

Table of contents (1 of 1)

What are we going to cover?

- Look to the future
- **Data wrangle** static Web pages from different sources
- Explore a few of the mysteries of CSS
- Understand how to download web pages

What is HTML?[3]

HTML is the standard markup language for creating Web pages.

- HTML stands for Hyper Text Markup Language
- HTML describes the structure of Web pages using markup
- HTML elements are the building blocks of HTML pages
- HTML elements are represented by tags
- HTML tags label pieces of content such as "heading", "paragraph", "table", and so on
- Browsers do not display the HTML tags, but use them to render the content of the page

A Simple HTML Document

Example

```
<!DOCTYPE html>
<html>
<head>
<title>Page Title</title>
</head>
<body>

<h1>My First Heading</h1>
<p>My first paragraph.</p>

</body>
</html>
```

Try it Yourself »

Image from [3].

HTML tags have attributes

Attributes help the browser to display tag related information "correctly."
Correctly can be dependent on which browser is being used.

All HTML Attributes

Attribute	Belongs to	Description
accept	<input>	Specifies the types of files that the server accepts (only for type="file")
accept-charset	<form>	Specifies the character encodings that are to be used for the form submission
accesskey	Global Attributes	Specifies a shortcut key to activate/focus an element
action	<form>	Specifies where to send the form-data when a form is submitted
align	Not supported in HTML 5.	Specifies the alignment according to surrounding elements. Use CSS instead
alt	<area>, <img>, <input>	Specifies an alternate text when the original element fails to display
async	<script>	Specifies that the script is executed

`https://www.w3schools.com/tags/ref_attributes.asp`

How to manage the "look" of a company's web pages?

HTML attributes can be used for "branding."[1]

- Website Content
- Overall Design and Layout
- Use of Innovation

Image from [2].

We're going to focus on design and layout.

Managing HTML attributes

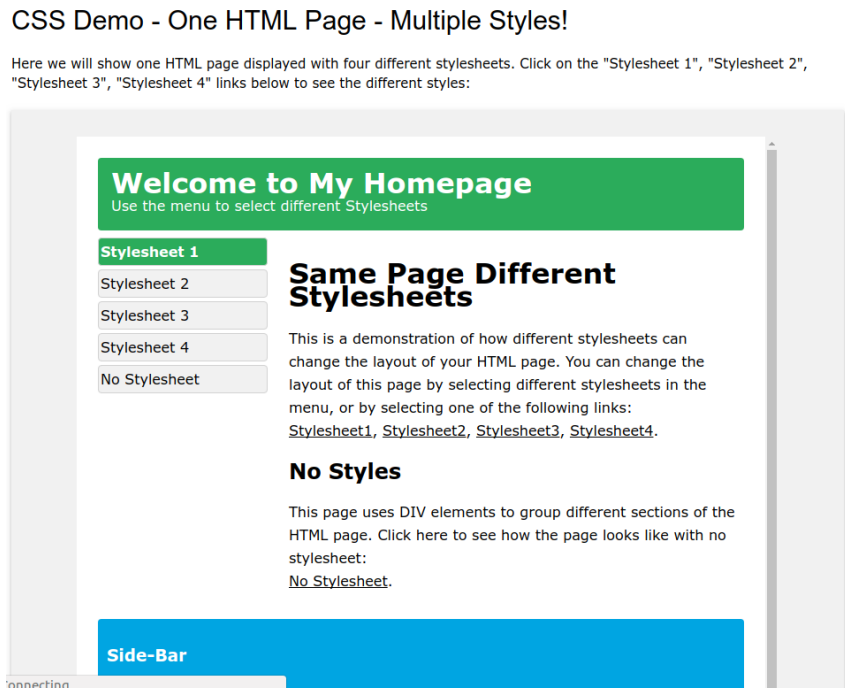
- HTML attributes on a few web pages can be managed by hand.
- HTML attributes on a few 10s of web pages might be managed by hand.
- HTML attributes on 1,000s of web pages can not be managed by hand.
- Cascading Style Sheets (CSS) are designed to work with HTML pages.
- HTML attributes can be written into the page, or can take values from a CSS.

The browser renders the page based on the HTML attribute values

Same HTML, different CSSs

To do:

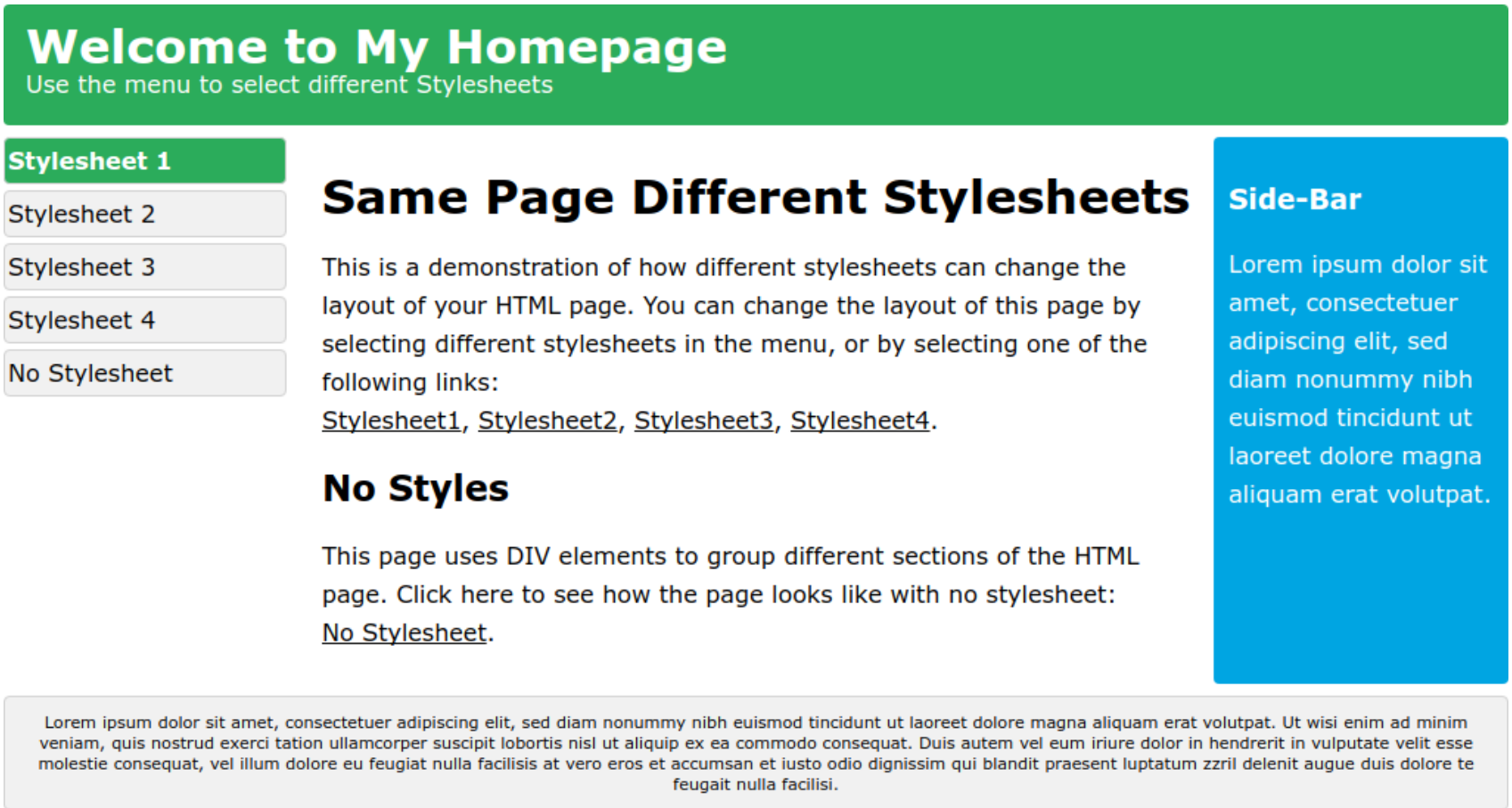
1. Load the page
2. Locate and press the "Stylesheet ..." button
3. Repeat as desired



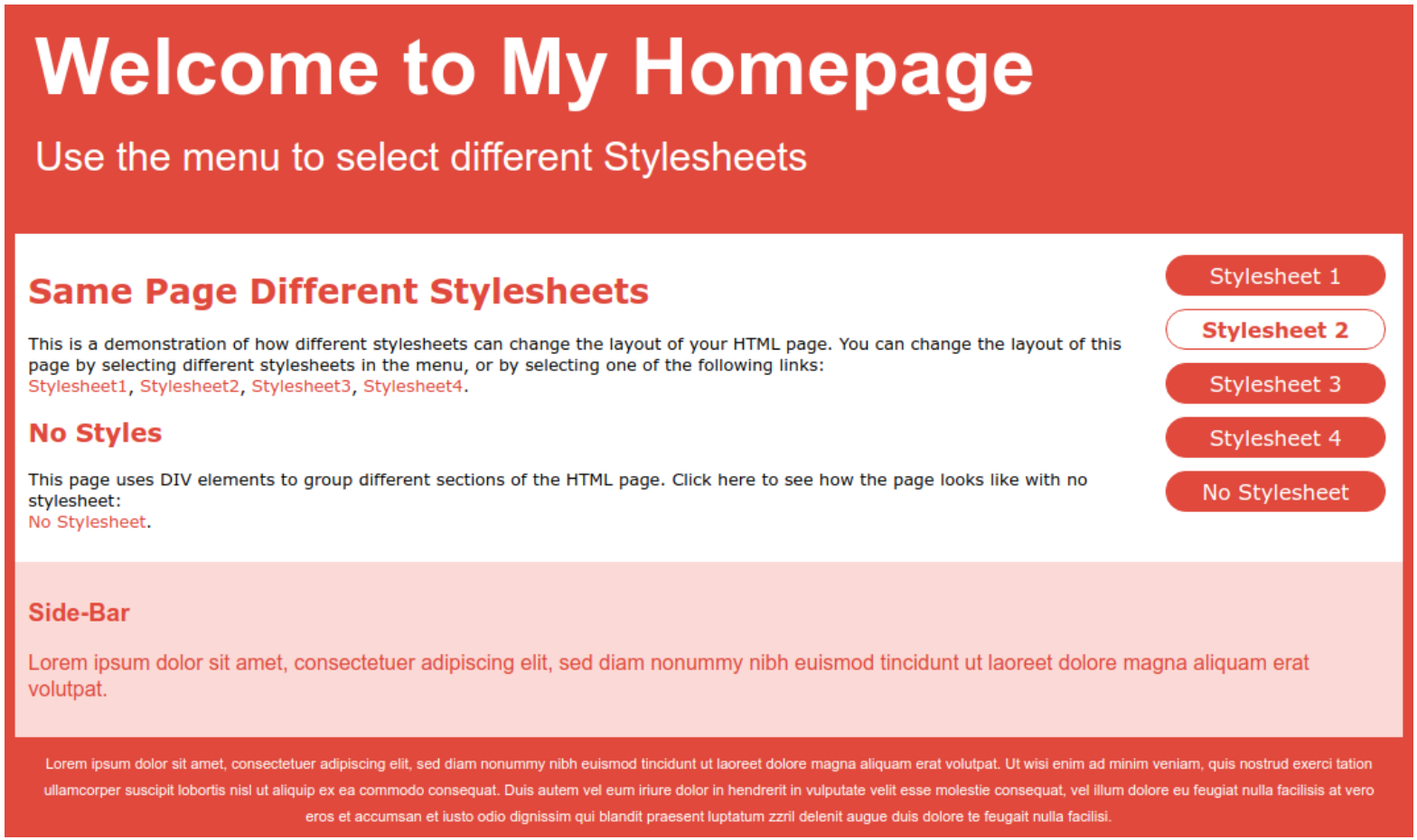
https://www.w3schools.com/css/css_intro.asp

The same HTML tags can be rendered differently based on the CSS.

Sample stylesheet 1.



Sample stylesheet 2.



Sample stylesheet 3.

Welcome to My Homepage

Use the menu to select different Stylesheets

Stylesheet 1 | Stylesheet 2 | Stylesheet 3 | Stylesheet 4 | No Stylesheet

Same Page Different Stylesheets

This is a demonstration of how different stylesheets can change the layout of your HTML page. You can change the layout of this page by selecting different stylesheets in the menu, or by selecting one of the following links:
Stylesheet1, Stylesheet2, Stylesheet3, Stylesheet4.

No Styles

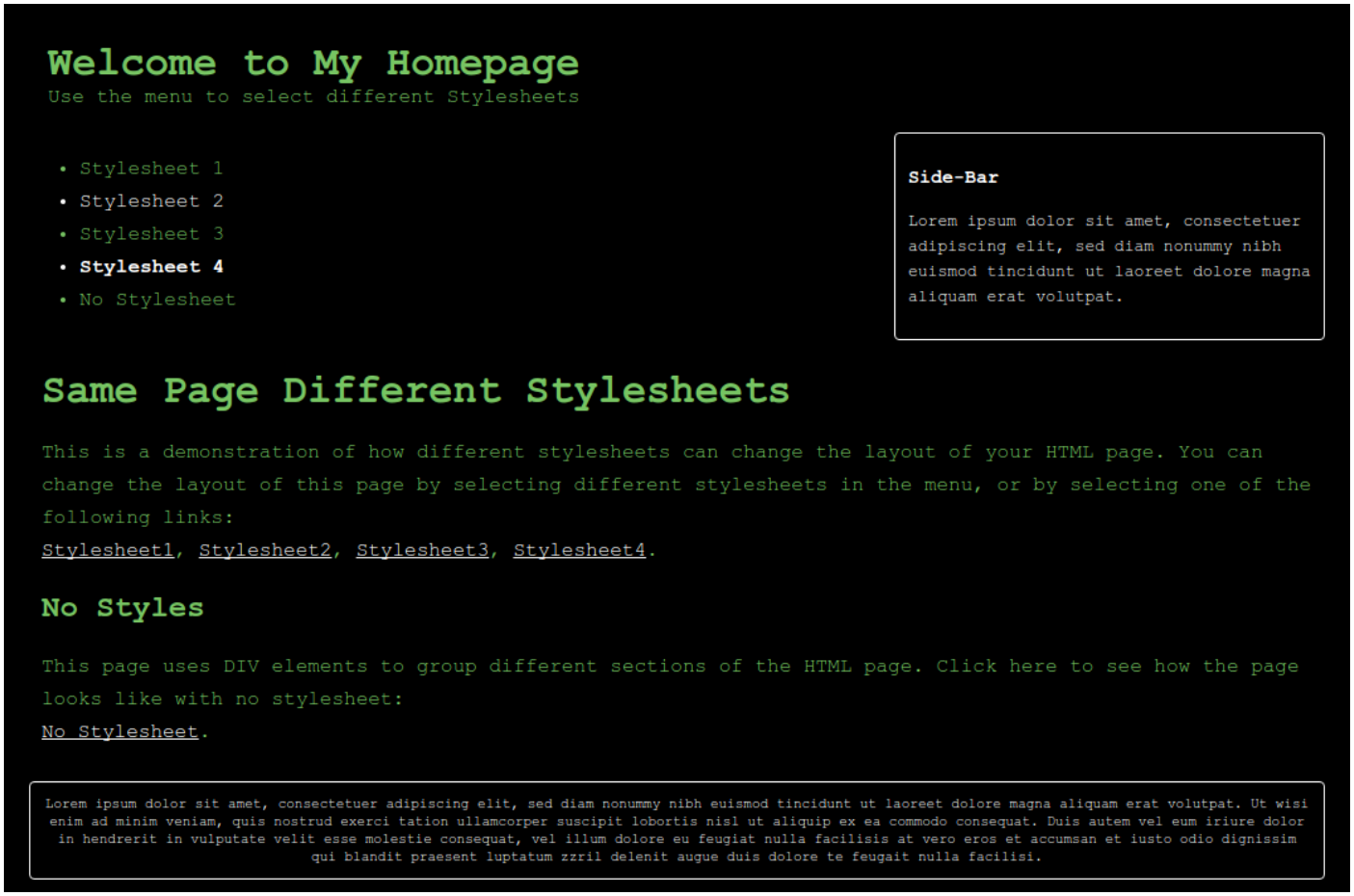
This page uses DIV elements to group different sections of the HTML page. Click here to see how the page looks like with no stylesheet:
No Stylesheet.

Side-Bar

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi.

Sample stylesheet 4.



Without a stylesheet.

Welcome to My Homepage

Use the menu to select different Stylesheets

- Stylesheet 1
- Stylesheet 2
- Stylesheet 3
- Stylesheet 4
- No Stylesheet

Same Page Different Stylesheets

This is a demonstration of how different stylesheets can change the layout of your HTML page. You can change the layout of this page by selecting different stylesheets in the menu, or by selecting one of the following links:
Stylesheet1, Stylesheet2, Stylesheet3, Stylesheet4.

No Styles

This page uses DIV elements to group different sections of the HTML page. Click here to see how the page looks like with no stylesheet:
No Stylesheet.

Side-Bar

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi.

Terminology

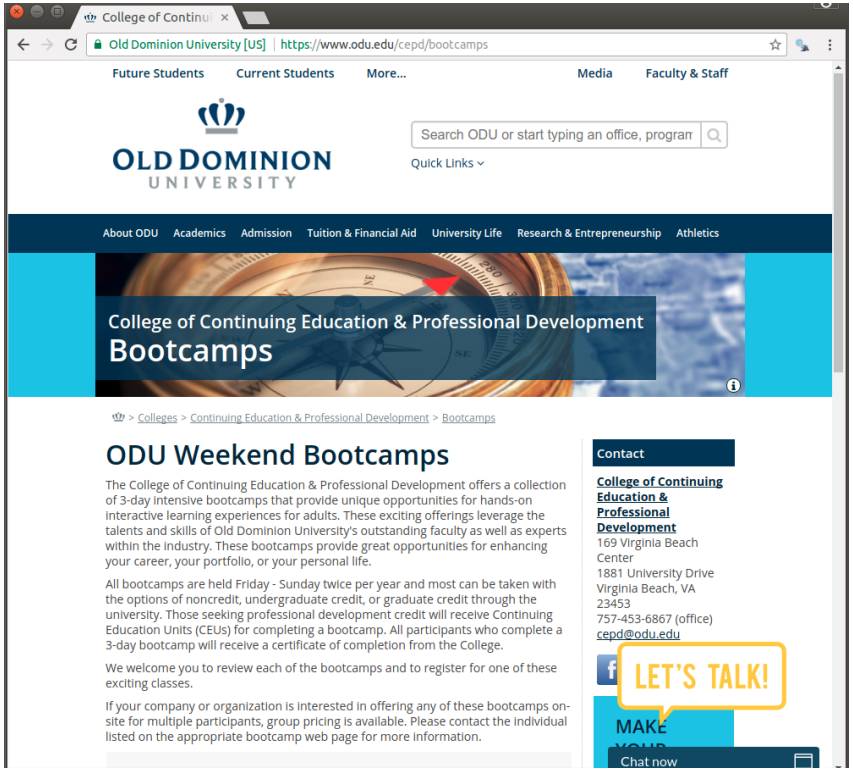
- HTML uses tags and attributes. These are sometimes called "elements."
- CSS uses selectors and declarations.
- CSS declarations have two fields: property and value.
- CSS selectors are used to "find" (or select) HTML elements based on their name, id, class, attribute, and more[1].

Amazon reports over 6,000 CSS books.

[1] `https://www.w3schools.com/css/css_syntax.asp`

Start somewhere easy:

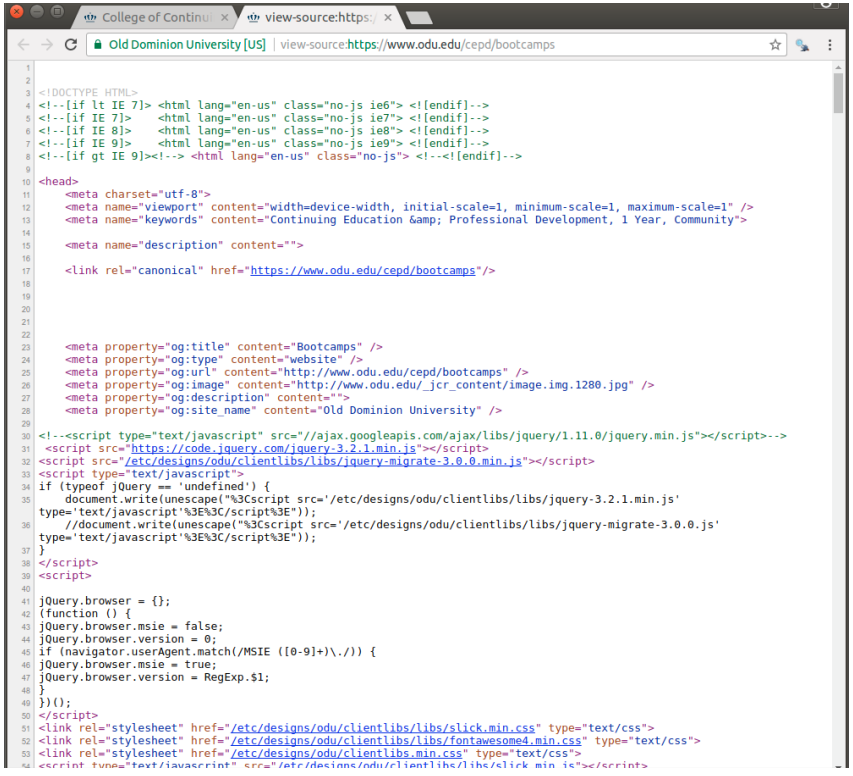
Bring up the ODU College of Continuing Education and Professional Development boot camp web page.



`https://www.odu.edu/cepd/bootcamps`

Look at the HTML tags (elements)

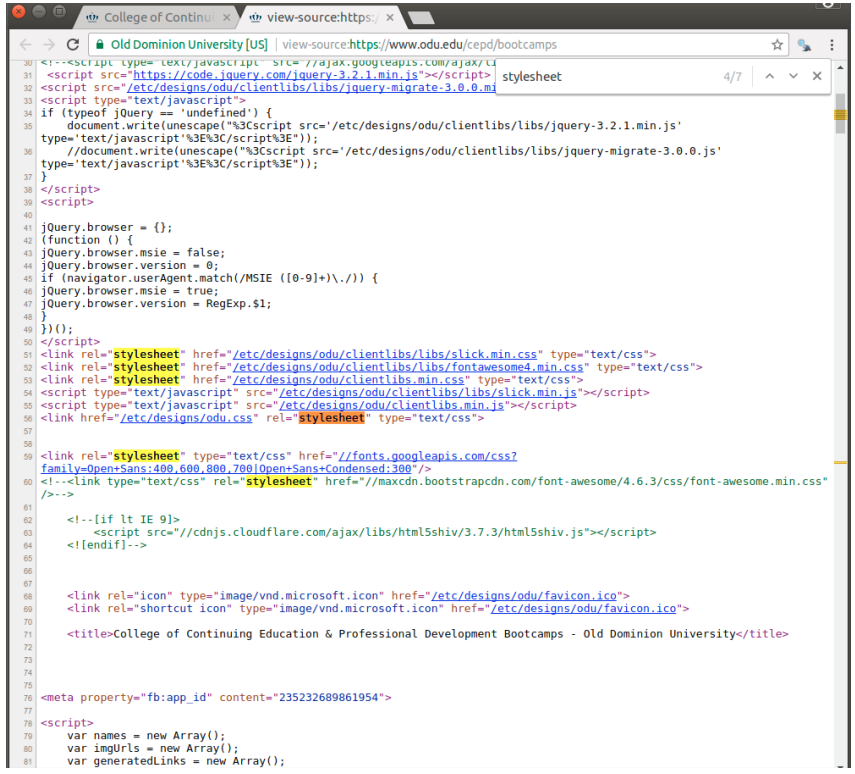
Right click on the page and select "View page source"



`https://www.odu.edu/cepd/bootcamps`

Find the style sheet references

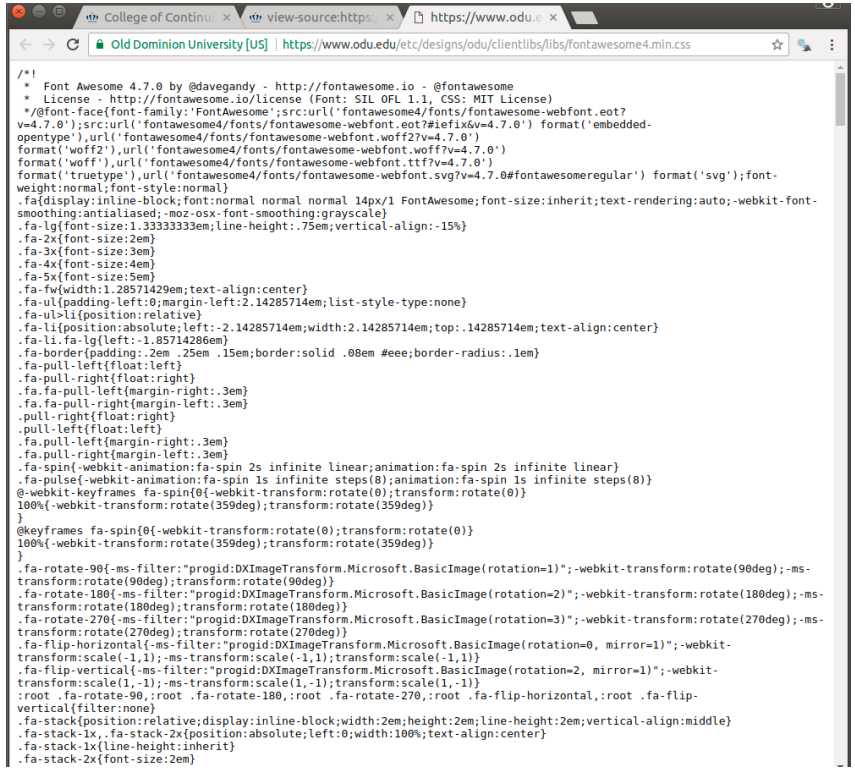
Execute a CTRL-F to locate the string: stylesheet



https://www.odu.edu/cepd/bootcamps

Looking at a CSS

Click on one of the links near the string “stylesheet”



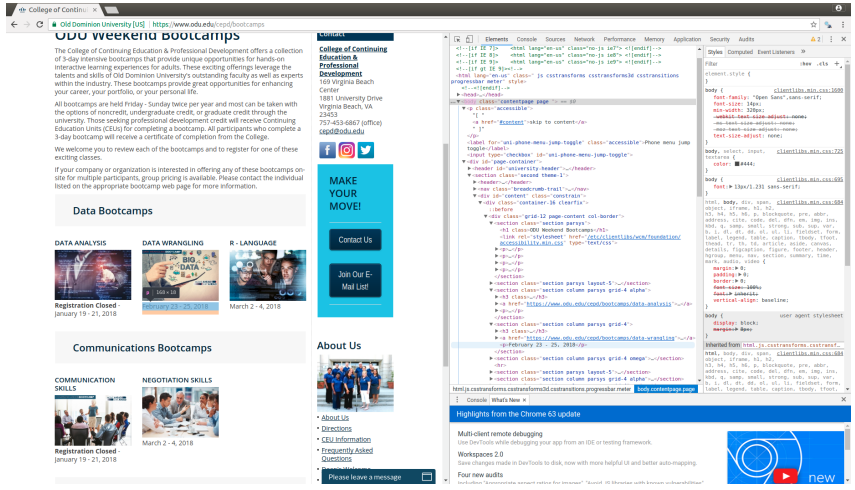
`https://www.odu.edu/etc/designs/odu/clientlibs/libs/fontawesome4.min.css`

CSS are meant for computer consumption, not humans.

We'll be picking apart a page to find the CSS selectors

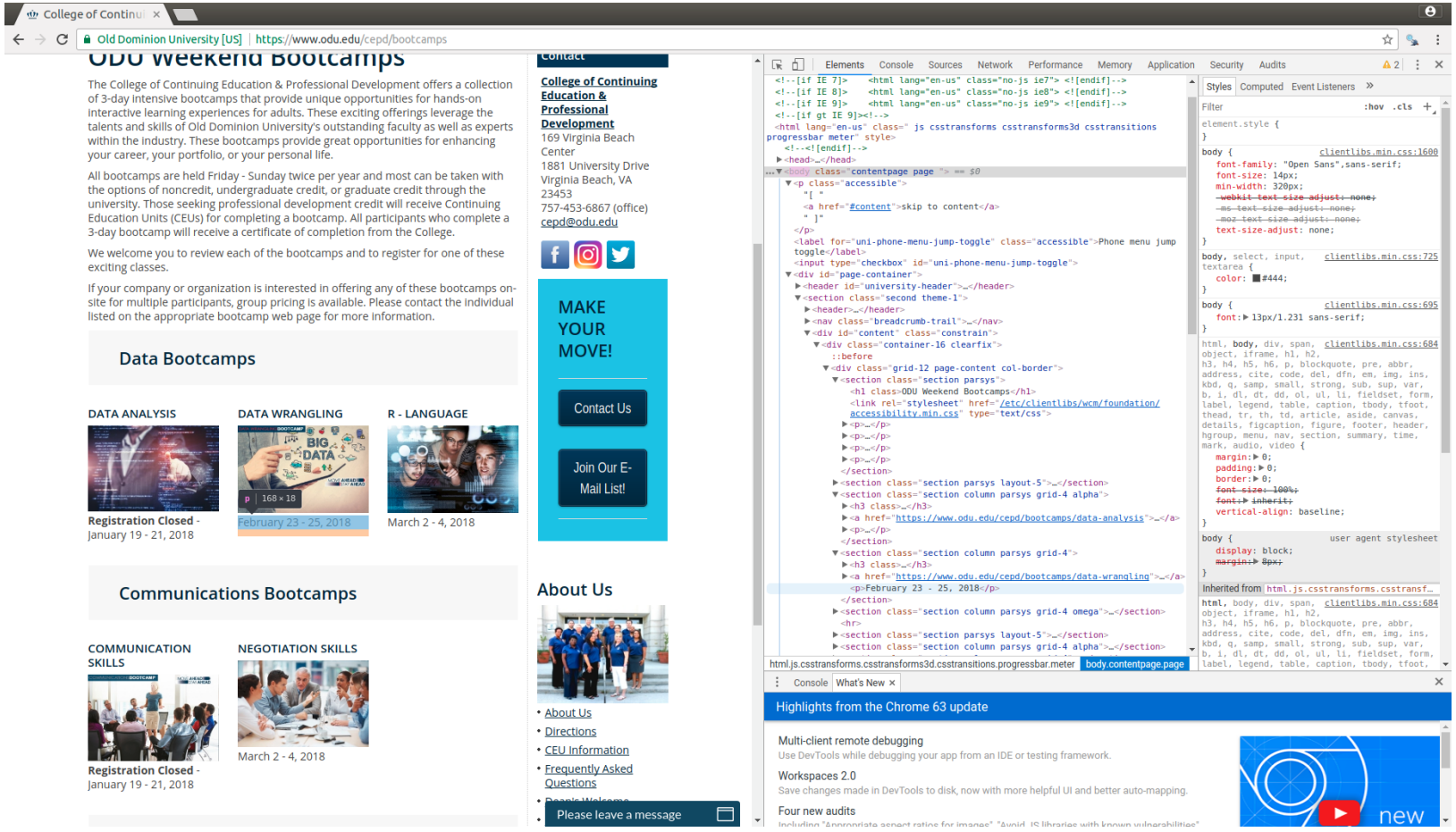
A few steps using Google Chrome.

1. Open this url: `https://www.odu.edu/cepd/bootcamps`
2. Press F12 (the debug key)
3. Inside the HTML source, drop down and expand until our class date is selected on the left



This is doable, but there are easier ways.

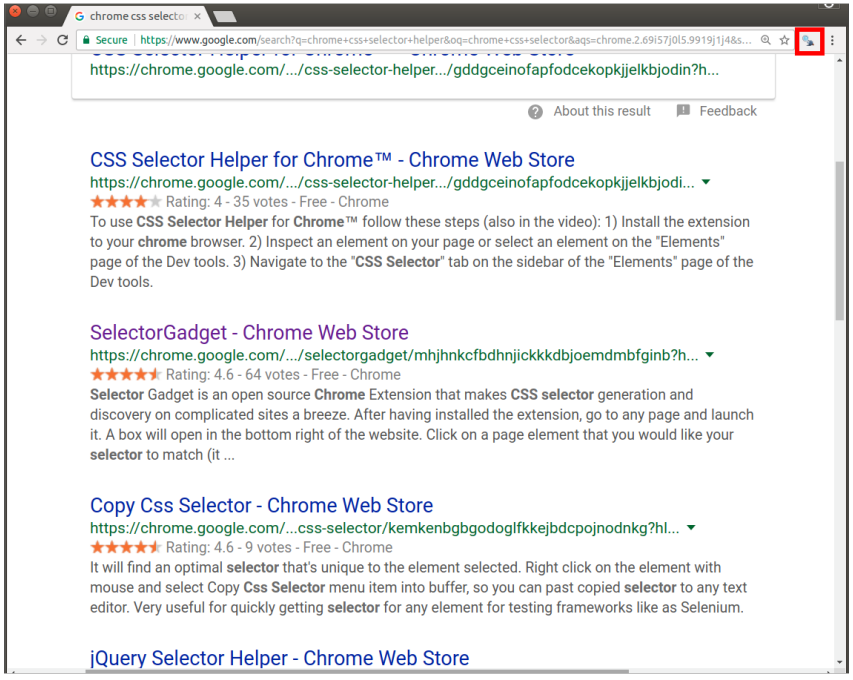
Same image.



This is doable, but there are easier ways.

Chrome has a free plug-in: SelectorGadget

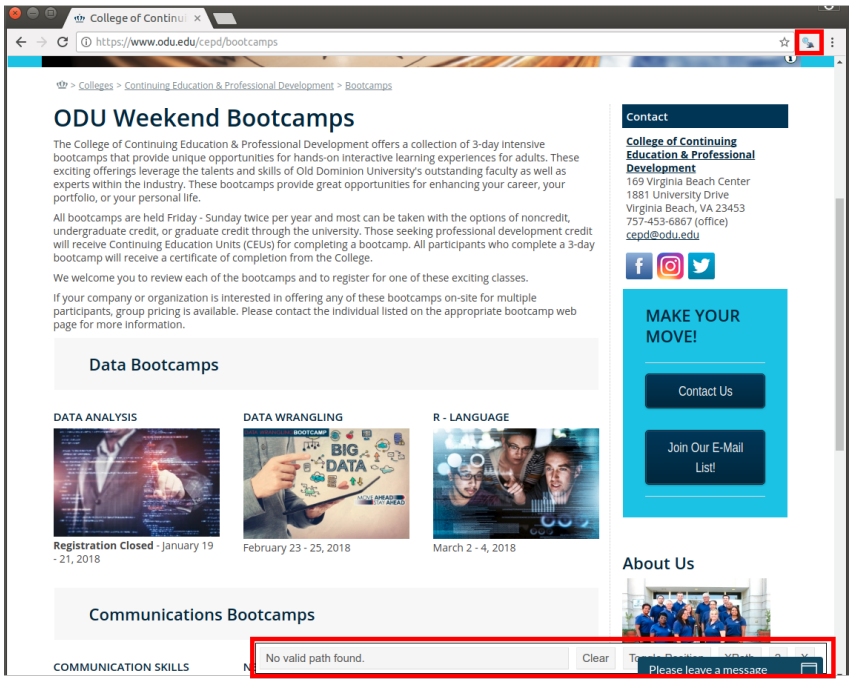
- The plug-in is free.
- A new icon is added to the browser



See the red rectangle in the image.

How to use the SelectorGadget? (1 of 4)

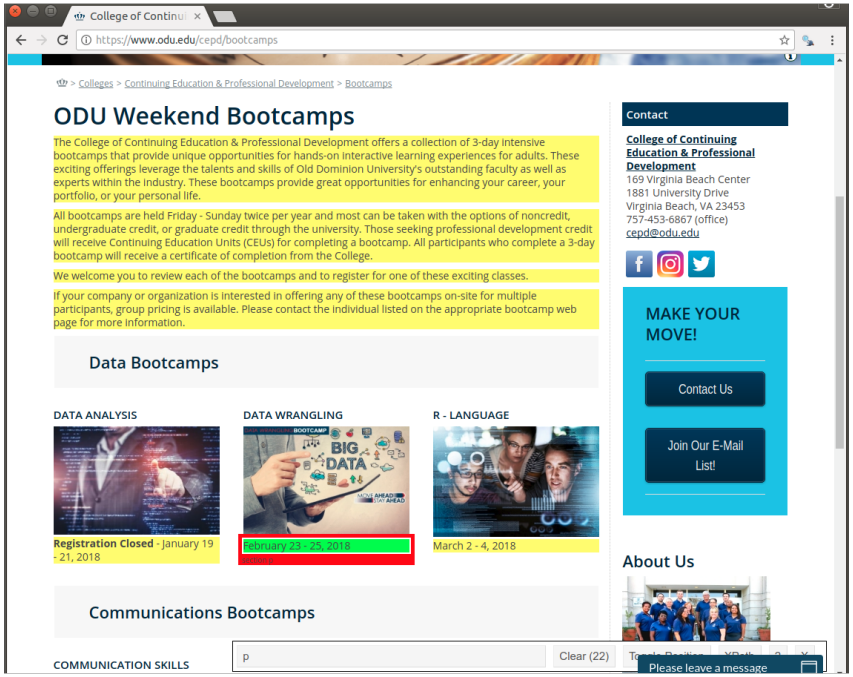
1. Load the page of interest.
2. Click the icon (small square at top of page).
3. See the pop-up windows (wide rectangle at bottom of page)



Press the gadget icon a second time to turn it off.

How to use the SelectorGadget? (2 of 4)

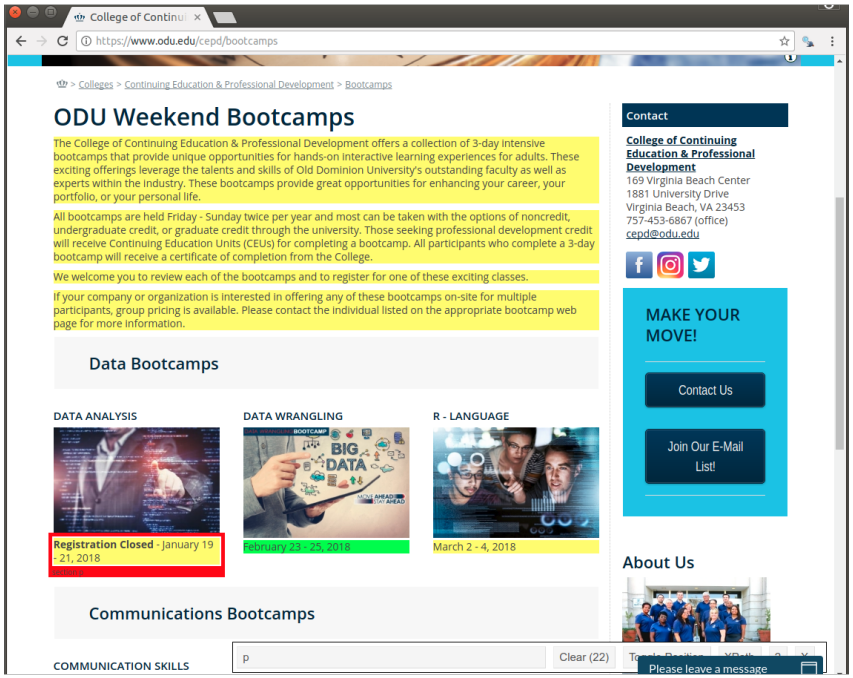
1. Click on the "item" of interest. The item will turn green. The bottom pop-up will have the CSS selector that identifies what was clicked.



All other HTML elements what match the CSS selector will turn yellow.

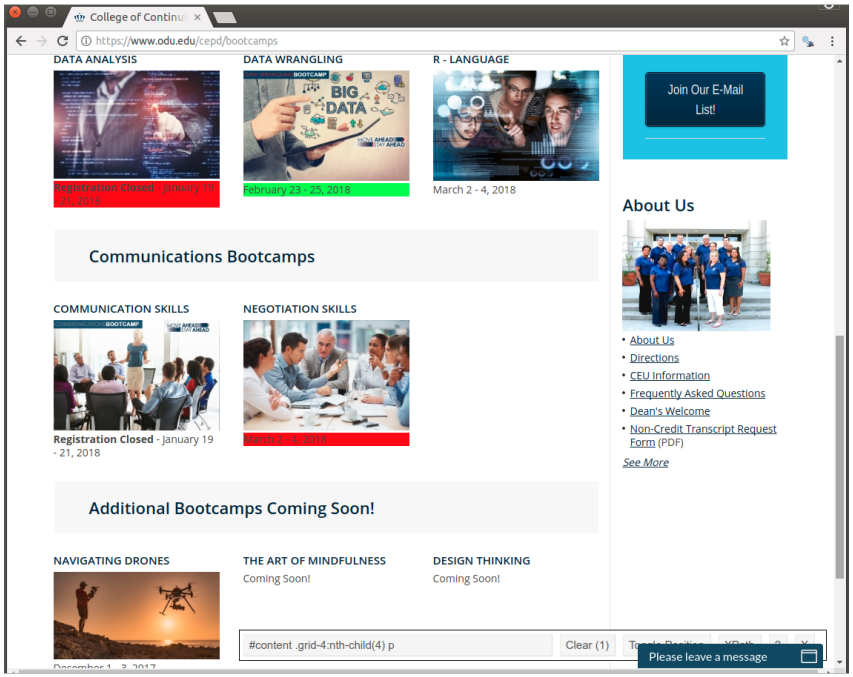
How to use the SelectorGadget? (3 of 4)

1. Click on the "item" **NOT** of interest. The item will be outlined in red.
2. Repeat the above step until only the "item" of interest is highlighted. This may take several iterations, and scrolling up and down the page to ensure only the items of interest are highlighted.



How to use the SelectorGadget? (4 of 4)

Once only the "items" of interest are green:
The gadget contains the CSS selector. In this case, it is:
#content .grid-4:nth-child(4) p
Where spaces, and case are important.



The selector is what we are after.

Getting a field of data.

Copy and paste these lines into RStudio:

```
library(xml2)
library(rvest)

url <- "https://www.odu.edu/cepd/bootcamps"
selector <- "#content .grid-4:nth-child(4) p"

webPage <- read_html(url)

data <- html_text(html_nodes(webPage, selector))

data
```

The result is: `February 23 - 25, 2018`

An interesting question.

Our numerically inclined supervisor has an interest in politics. Specifically, can bi-partisan support be quantified?

How can we use R, CSS selectors, and a little math to answer the question?

Where/how to begin?

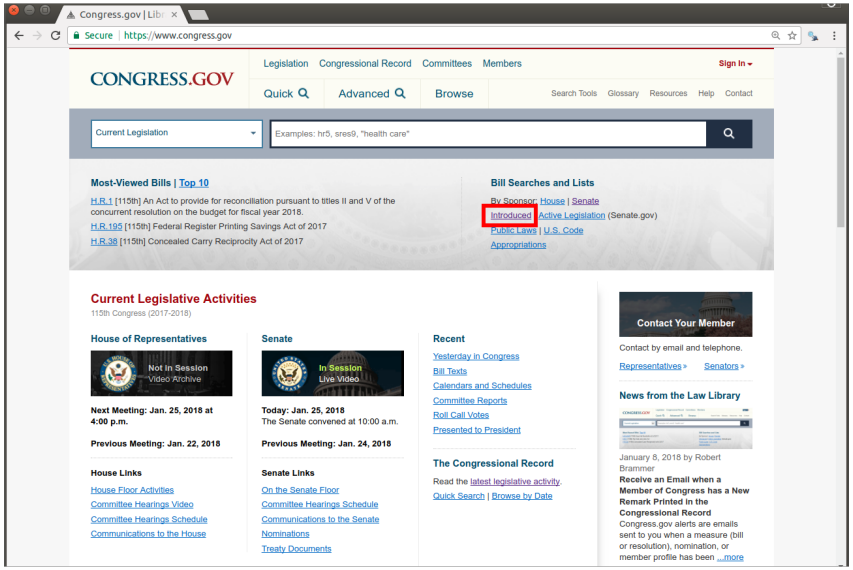
We'll need to do the following:

1. Get a list of all the Senate bills
2. Get each bill's sponsor (there will only be one)
3. Get each bill's cosponsor (there may be none, or a bunch)
4. Get each member's party (there are three)
5. See how often members of different parties are on the same bill

All of this leading to a data structure, indexed by bill number, where each bill has a list containing the sponsor and any cosponsors in that order.

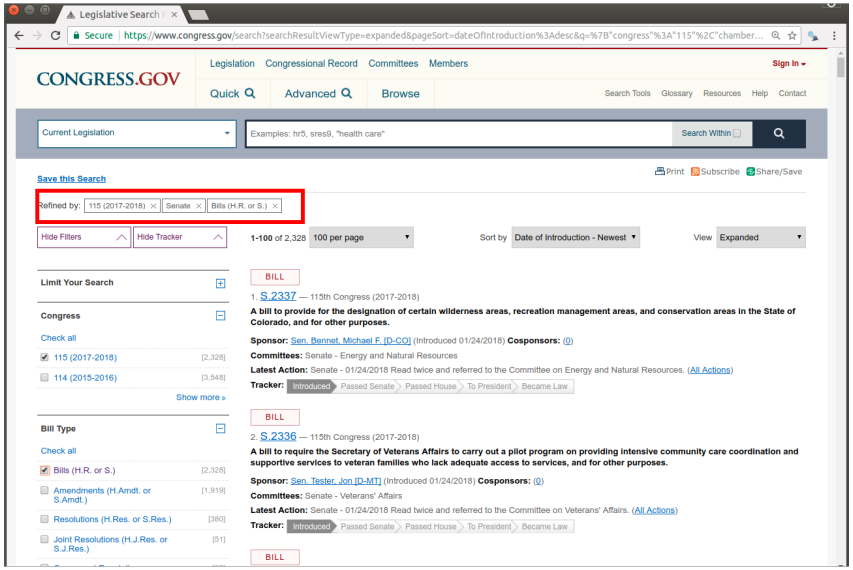
Where/how to get a list of Senate bills?

This is the first step in our search:
`https://www.congress.gov/`
and pressing the “Introduced” link.



Limiting the search to Senate bills.

Using the filters on the left, limit the search to the 115th Congress, originated in the Senate, and is a Bill.
There are 2,328 bills introduced.
Far too many to fit on a screen.



Looking at the page:

Some details about the page:

- The maximum number of results per page is 250.
- There are navigation controls at the bottom of the page.
- The only apparent difference between pages, is the value of the last argument.
- Changing the value of "page" to something extreme results in the last valid page

Sample URLs

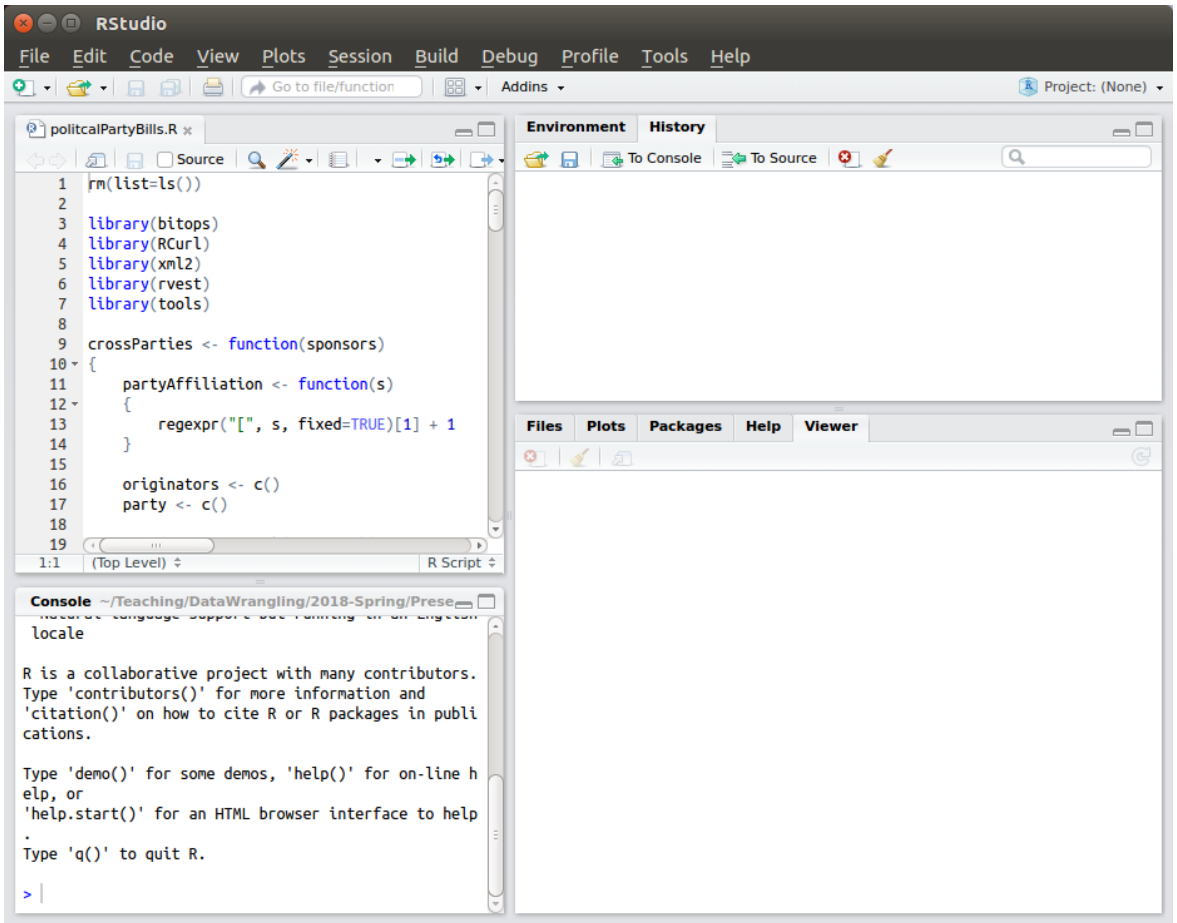
URLs to get pages 3 and 5.
Lines broken for readability

```
https://www.congress.gov/search?searchResultViewType=expanded&
       pageSort=dateOfIntroduction%3Adesc&q=%7B%22congress%22%3A%22115%22
       %2C%22chamber%22%3A%22Senate%22%2C%22type%22%3A%22bills%22%7D&page=3
https://www.congress.gov/search?searchResultViewType=expanded&
       pageSort=dateOfIntroduction%3Adesc&q=%7B%22congress%22%3A%22115%22
       %2C%22chamber%22%3A%22Senate%22%2C%22type%22%3A%22bills%22%7D&page=5
```

So:

1. We can use SelectorGadget to identify the Bill number
2. We can loop across all pages, starting at 1 and continuing until we get the same page twice
3. Collect all the Bill numbers as we go along

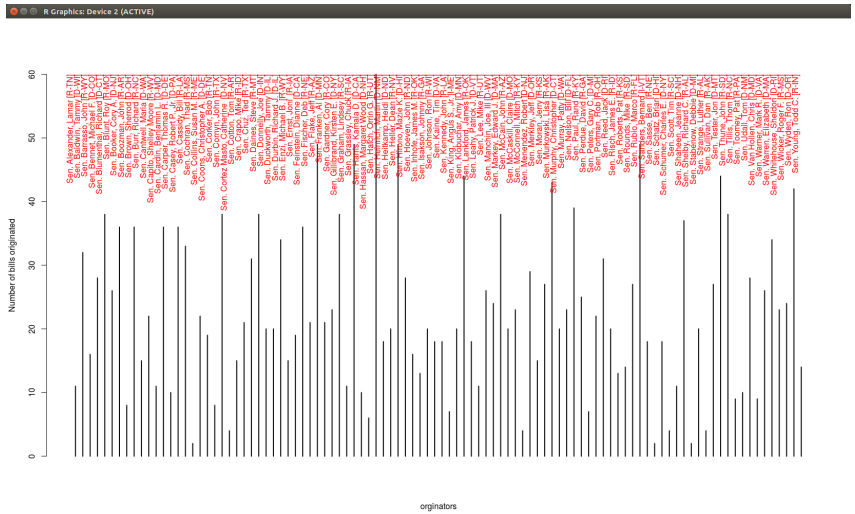
Lets open the attached file, and have some fun.



How a bill originator histogram may look

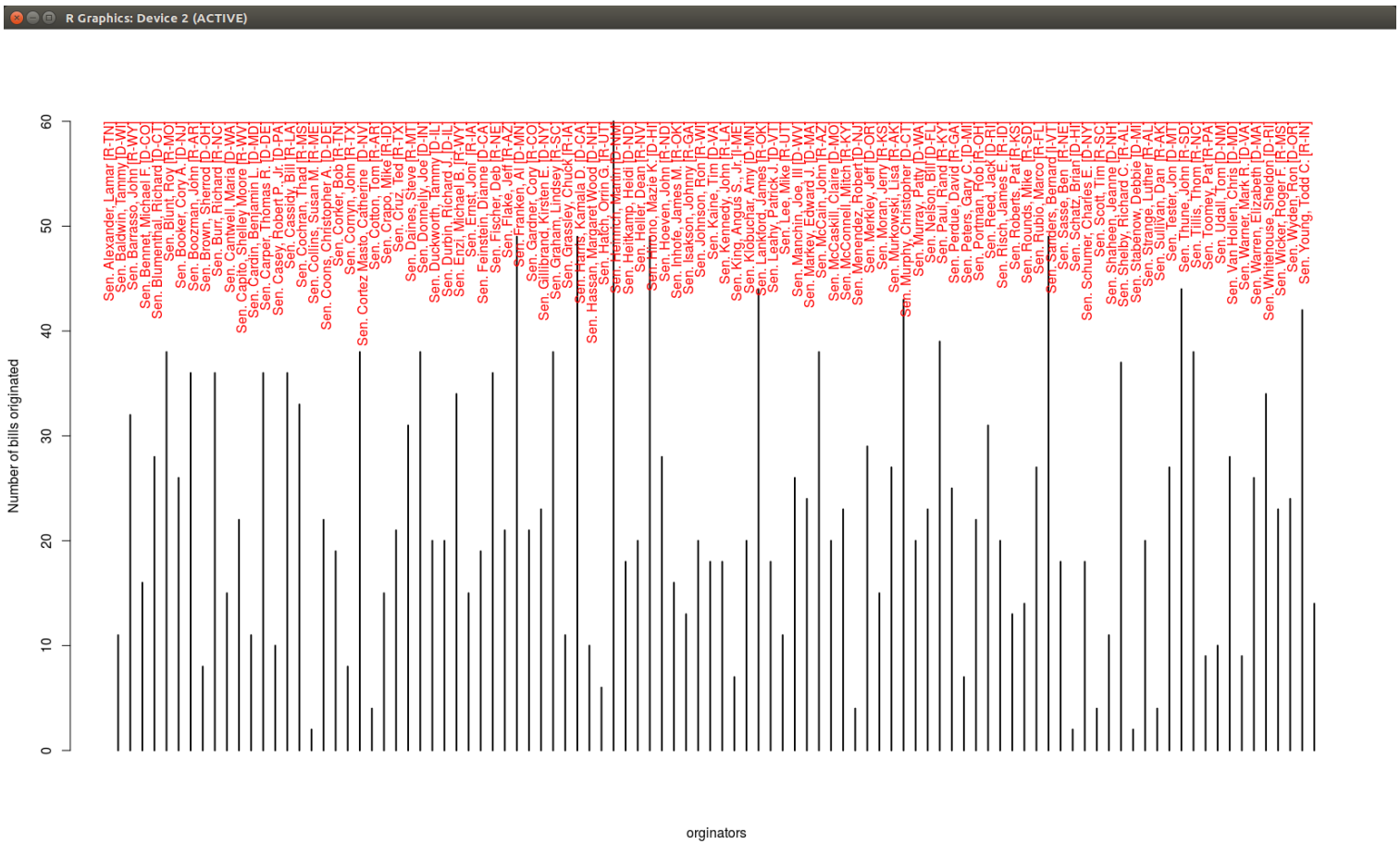
Things of interest:

- Orrin Hatch's office is really, really busy.
- Cochran's and Schalz's office, not so much. (Makes you wonder.)



Data is time sensitive, and deserves greater examination.

Same image.



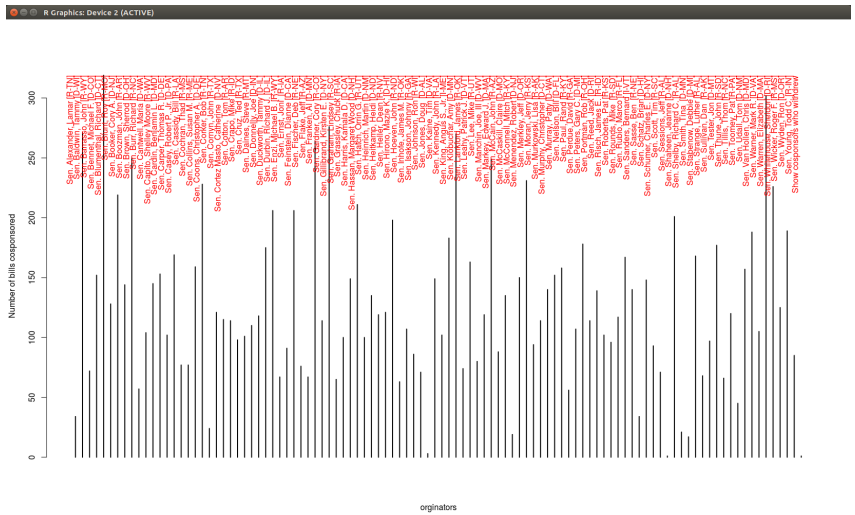
Data is time sensitive, and deserves greater examination.

Senate bill cosponsors.

Thins of interest:

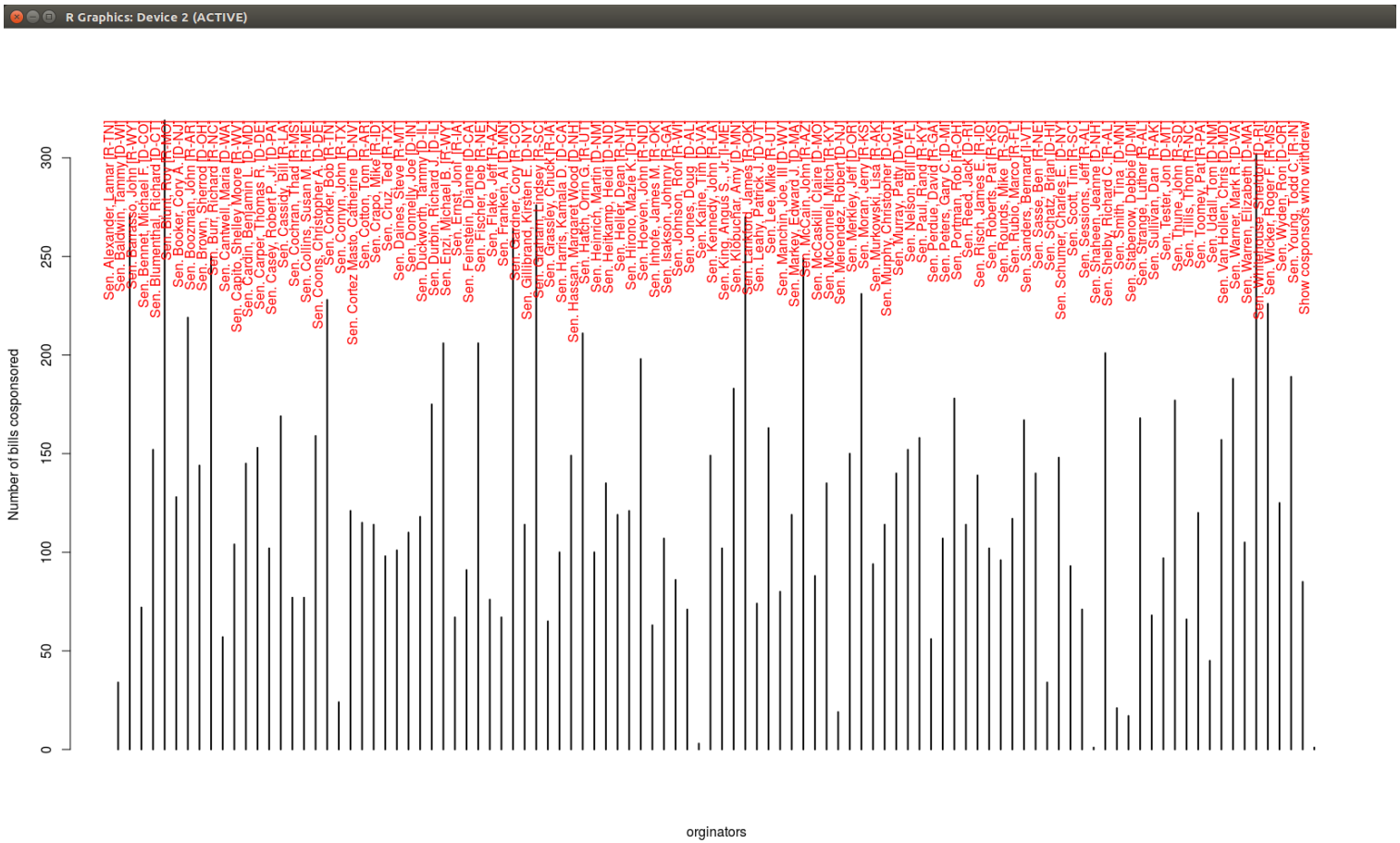
- There are some senators who are very popular.
- There are others, not so much.

It would be interesting to "normalize" sponsorship by length of time in the congress.



Data is time sensitive, and deserves greater examination.

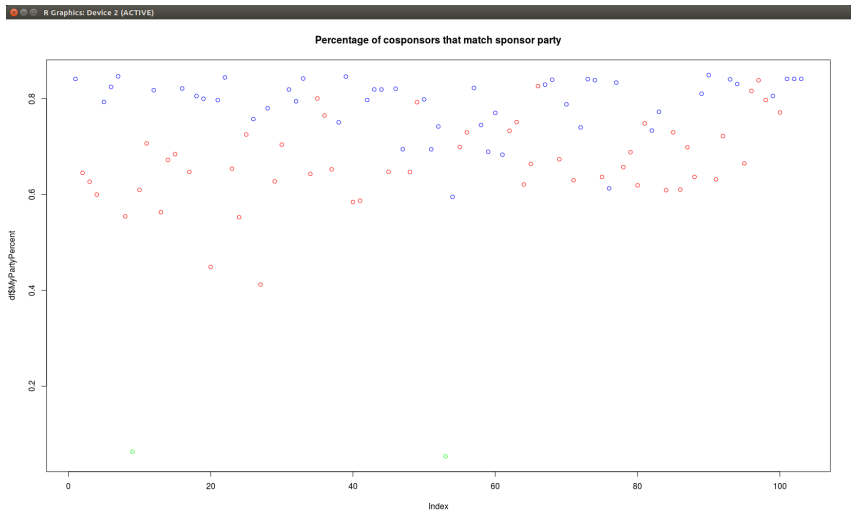
Same image.



Data is time sensitive, and deserves greater examination.

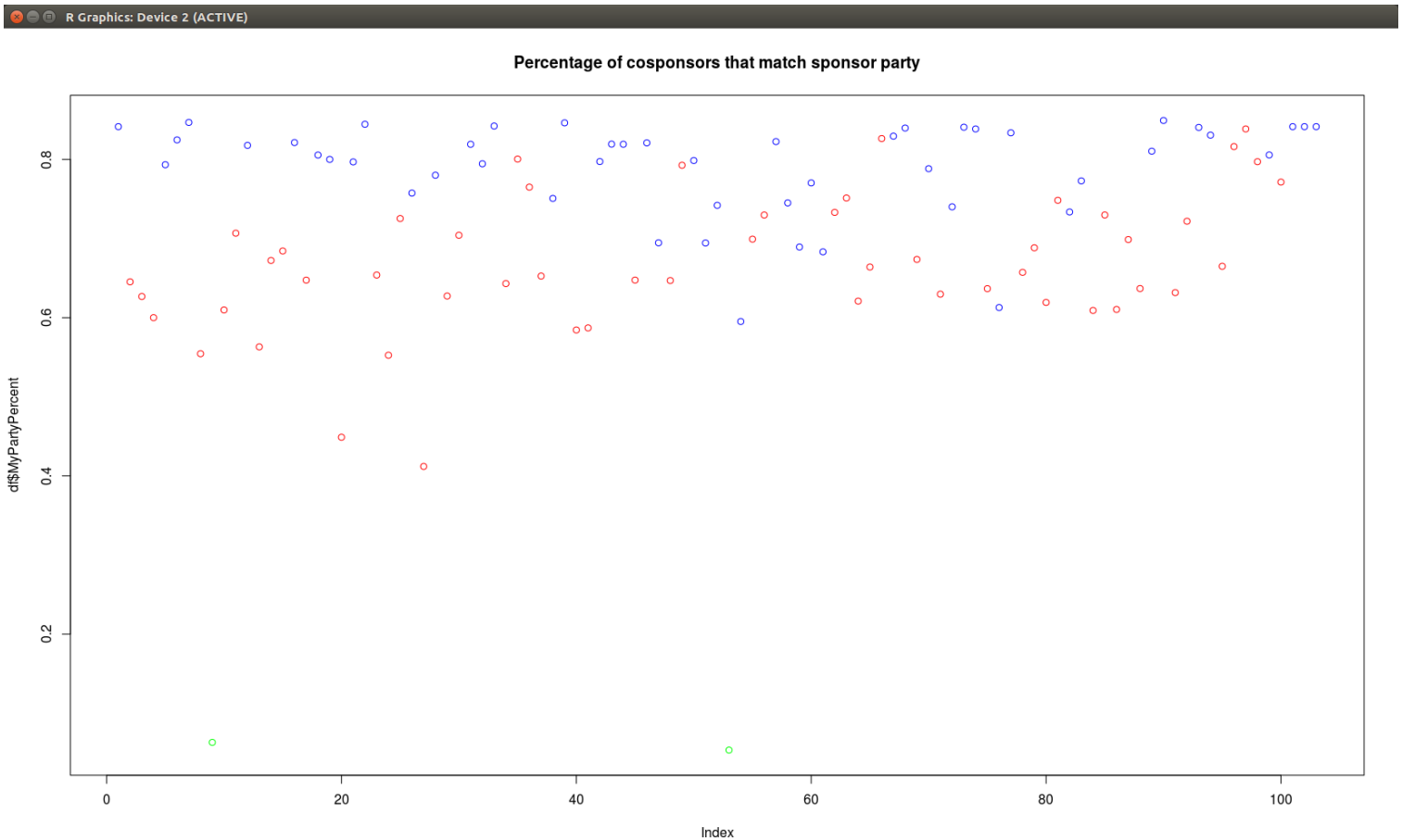
Given my party, what percentage of cosponsors match my party?

There is almost a division.
Republicans are red.
Democrats are blue.
Independents are green.



By and large, people stay within their party.

Same image.



By and large, people stay within their party.

Now that we have looked at the data, what other questions could be asked and answered?

In no particular order:

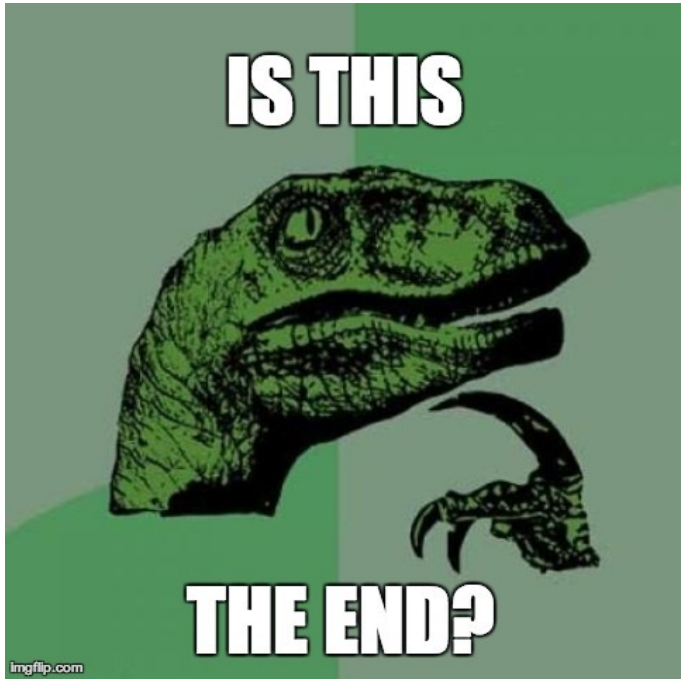
1. Who are the most influential senators? Where influential is measured by how many bills they are directly, or indirectly associated with.
2. Are senators from the same region more likely to work together than those not from the same region? (Think about senators around the Chesapeake Bay.)
3. What are individual senators really interested in? (Think about looking at the text of their sponsored and cosponsored bills to understand what they put their names to.)
4. Should be lots of others.

Q & A time.

Q: How many bureaucrats does it take to screw in a light bulb?
A: Two. One to assure everyone that everything possible is being done while the other screws the bulb into the water faucet.

What have we covered?

- Explored a little bit of cascading style sheets (CSS)
- Used R to download web pages
- Used R to extract data based on CSS selectors



Next: Exploring the wild and woolly Web world.

References (1 of 1)

[1] Corey Smith, HOW TO USE YOUR WEBSITE AS A BRANDING TOOL, https://www.tributemedia.com/blog/using-your-website-branding-tool, 2013.

[2] OBsurvey Staff, Most effective distribution methods to get answers, http://obsurvey.com/blog/most-effective-distribution-methods-to-get-answers/, 2014.

[3] W3 Staff, What is HTML?, https://www.w3schools.com/html/html_intro.asp, 2017.

Files of interest

1. Political parties