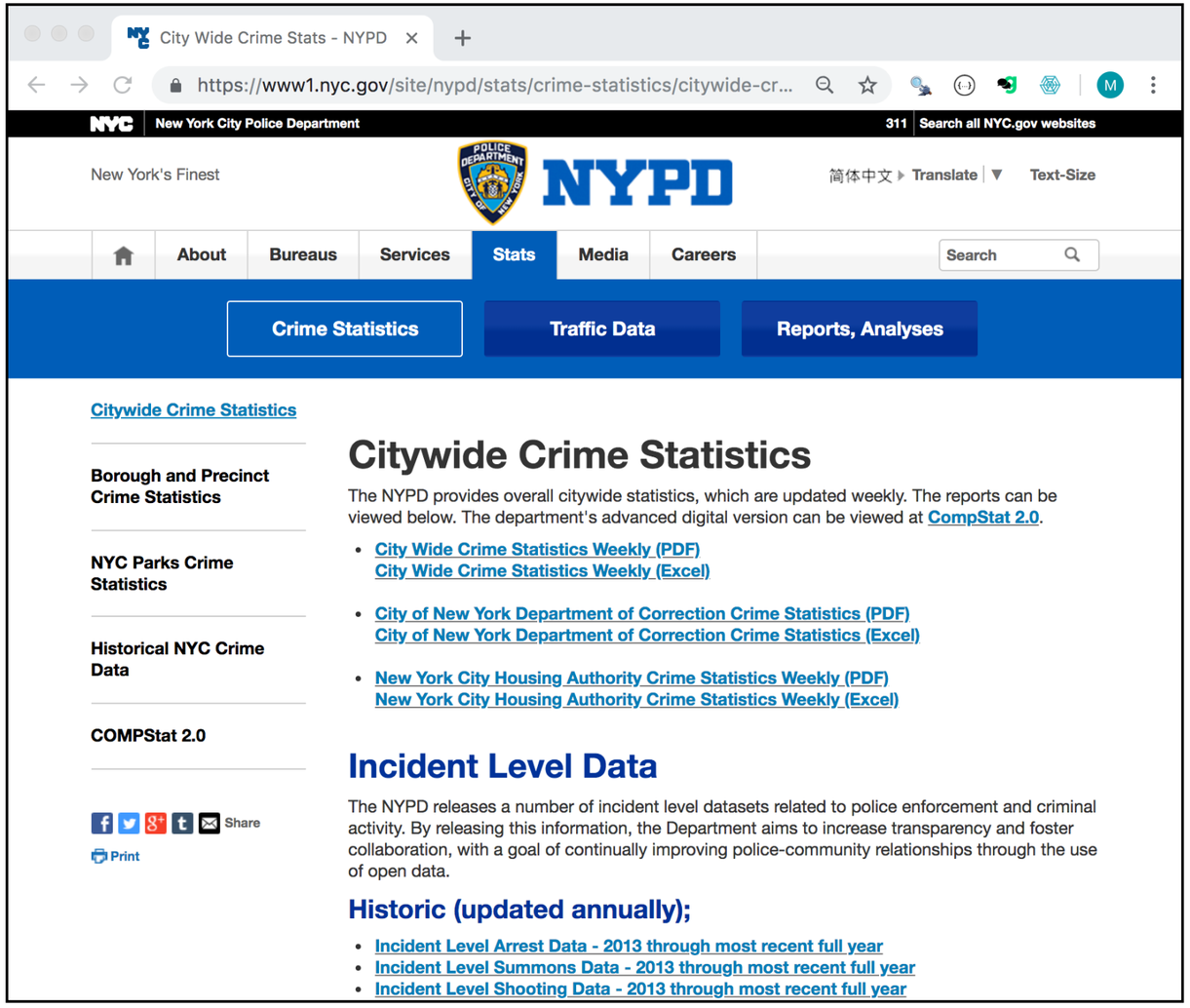
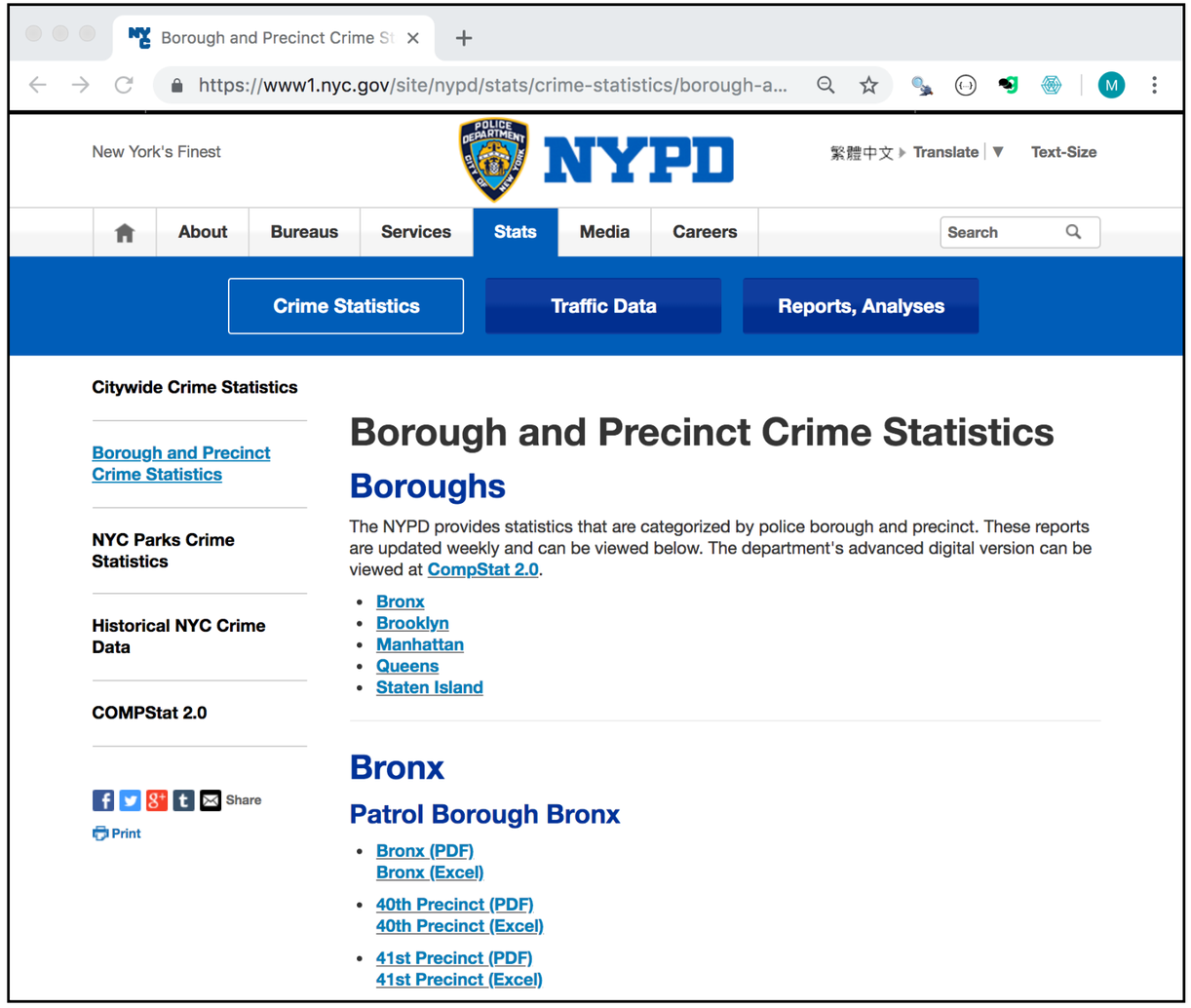
Scraping data from the web

With the move toward open data, organizations are increasingly making their data available online. But there is a good deal of variation in what, precisely, it means to publish data on the web.

Take data on crime or traffic incidents in NYC. The NYPD makes various PDF documents and Excel files for aggregate data, tabulated citywide, by borough or by precinct.

Incident-level data is rolled up to the nearest full quarter, although weekly updates are available via their CompStat 2.0 portal.


City Wide Crime Stats - NYPD
https://www1.nyc.gov/site/nypd/stats/crime-statistics/citywide-cr...
NYC New York City Police Department
311 Search all NYC.gov websites
New York's Finest
NYPD
简体中文 Translate Text-Size
About Bureaus Services Stats Media Careers
Search
Crime Statistics
Traffic Data
Reports, Analyses
Citywide Crime Statistics
Borough and Precinct Crime Statistics
NYC Parks Crime Statistics
Historical NYC Crime Data
COMPStat 2.0
Share
Print
Citywide Crime Statistics
The NYPD provides overall citywide statistics, which are updated weekly. The reports can be viewed below. The department's advanced digital version can be viewed at CompStat 2.0.
City Wide Crime Statistics Weekly (PDF)
City Wide Crime Statistics Weekly (Excel)
City of New York Department of Correction Crime Statistics (PDF)
City of New York Department of Correction Crime Statistics (Excel)
New York City Housing Authority Crime Statistics Weekly (PDF)
New York City Housing Authority Crime Statistics Weekly (Excel)
Incident Level Data
The NYPD releases a number of incident level datasets related to police enforcement and criminal activity. By releasing this information, the Department aims to increase transparency and foster collaboration, with a goal of continually improving police-community relationships through the use of open data.
Historic (updated annually);
Incident Level Arrest Data - 2013 through most recent full year
Incident Level Summons Data - 2013 through most recent full year
Incident Level Shooting Data - 2013 through most recent full year


Borough and Precinct Crime St
https://www1.nyc.gov/site/nypd/stats/crime-statistics/borough-a...
New York's Finest
NYPD
繁體中文 Translate Text-Size
About Bureaus Services Stats Media Careers
Search
Crime Statistics
Traffic Data
Reports, Analyses
Citywide Crime Statistics
Borough and Precinct Crime Statistics
NYC Parks Crime Statistics
Historical NYC Crime Data
COMPStat 2.0
Share
Print
Borough and Precinct Crime Statistics
Boroughs
The NYPD provides statistics that are categorized by police borough and precinct. These reports are updated weekly and can be viewed below. The department's advanced digital version can be viewed at CompStat 2.0.
Bronx
Brooklyn
Manhattan
Queens
Staten Island
Bronx
Patrol Borough Bronx
Bronx (PDF)
Bronx (Excel)
40th Precinct (PDF)
40th Precinct (Excel)
41st Precinct (PDF)
41st Precinct (Excel)

NYPD CompStat 2.0

compstat.nypdonline.org/2e5c3f4b-85c1-4635-83c6-22b27fe7c75c/view/90

Update

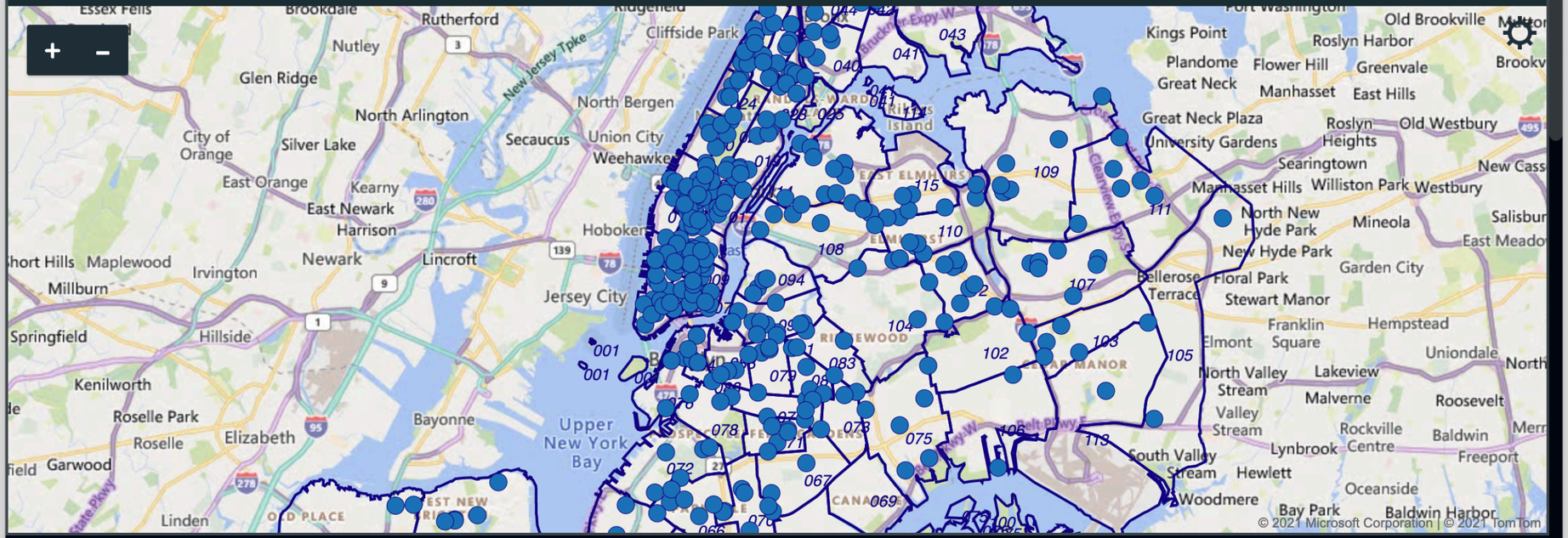
NYPD CompStat 2.0

As of 11/14/21

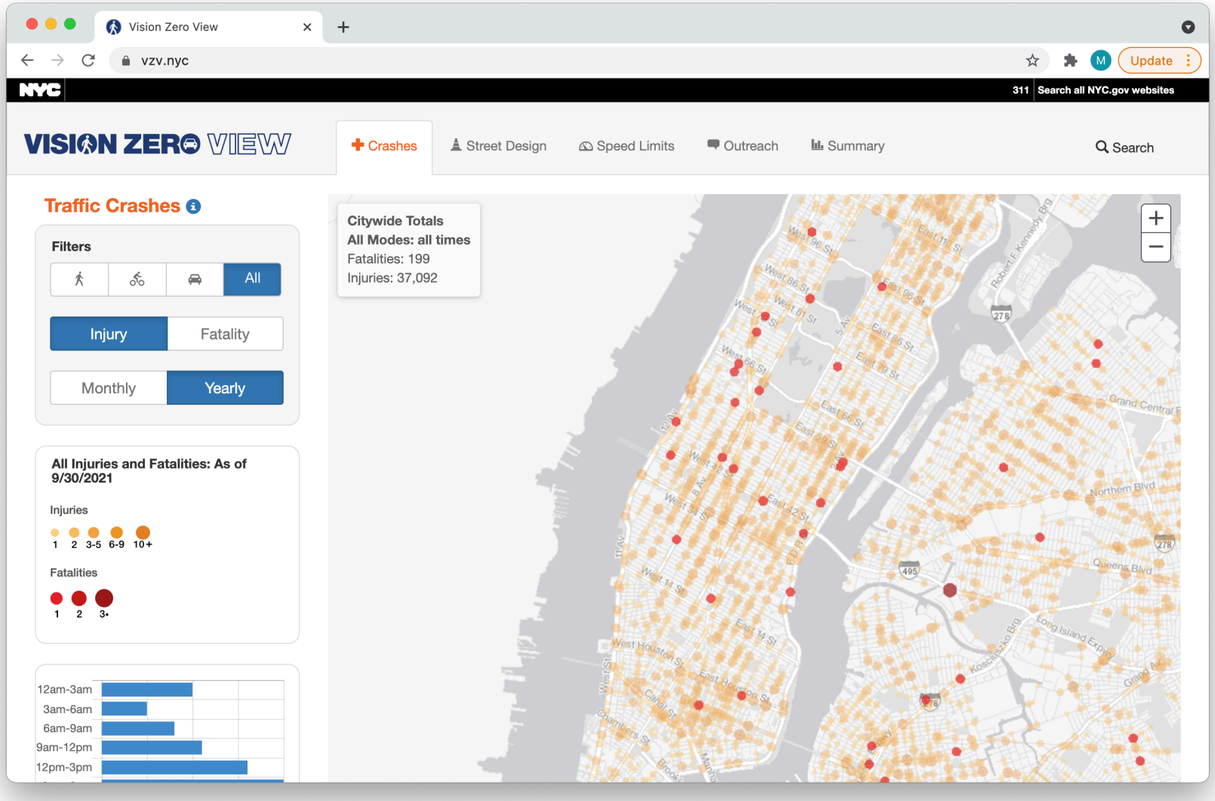
Citywide | Hate Crimes

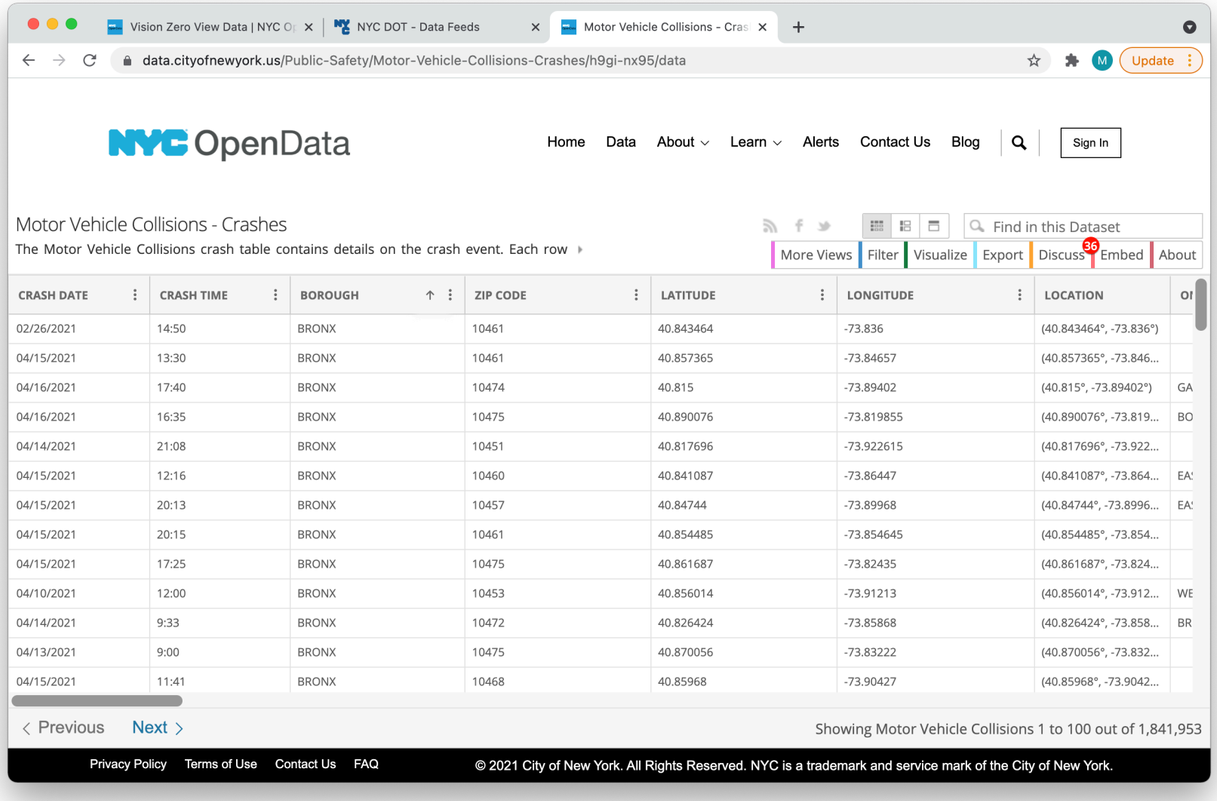
Period	2021	2020	% Change
Year to Date	475	238	99.6 %

Incident Map



Day of Week


Vision Zero View
vzv.nyc
Update
NYC
311 Search all NYC.gov websites
VISION ZERO VIEW
Crashes
Street Design
Speed Limits
Outreach
Summary
Search
Traffic Crashes
Filters
All
Injury
Fatality
Monthly
Yearly
All Injuries and Fatalities: As of 9/30/2021
Injuries
1 2 3-5 6-9 10+
Fatalities
1 2 3+
12am-3am
3am-6am
6am-9am
9am-12pm
12pm-3pm
Citywide Totals
All Modes: all times
Fatalities: 199
Injuries: 37,092


Vision Zero View Data | NYC O
NYC DOT - Data Feeds
Motor Vehicle Collisions - Cras
data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/data
Update
NYC OpenData
Home
Data
About
Learn
Alerts
Contact Us
Blog
Sign In
Motor Vehicle Collisions - Crashes
The Motor Vehicle Collisions crash table contains details on the crash event. Each row
Find in this Dataset
More Views
Filter
Visualize
Export
Discuss
Embed
About
CRASH DATE
CRASH TIME
BOROUGH
ZIP CODE
LATITUDE
LONGITUDE
LOCATION
02/26/2021 14:50 BRONX 10461 40.843464 -73.836 (40.843464°, -73.836°)
04/15/2021 13:30 BRONX 10461 40.857365 -73.84657 (40.857365°, -73.846...
04/16/2021 17:40 BRONX 10474 40.815 -73.89402 (40.815°, -73.89402°)
04/16/2021 16:35 BRONX 10475 40.890076 -73.819855 (40.890076°, -73.819...
04/14/2021 21:08 BRONX 10451 40.817696 -73.922615 (40.817696°, -73.922...
04/15/2021 12:16 BRONX 10460 40.841087 -73.86447 (40.841087°, -73.864...
04/15/2021 20:13 BRONX 10457 40.84744 -73.89968 (40.84744°, -73.8996...
04/15/2021 20:15 BRONX 10461 40.854485 -73.854645 (40.854485°, -73.854...
04/15/2021 17:25 BRONX 10475 40.861687 -73.82435 (40.861687°, -73.824...
04/10/2021 12:00 BRONX 10453 40.856014 -73.91213 (40.856014°, -73.912...
04/14/2021 9:33 BRONX 10472 40.826424 -73.85868 (40.826424°, -73.858...
04/13/2021 9:00 BRONX 10475 40.870056 -73.83222 (40.870056°, -73.832...
04/15/2021 11:41 BRONX 10468 40.85968 -73.90427 (40.85968°, -73.9042...
Previous
Next
Showing Motor Vehicle Collisions 1 to 100 out of 1,841,953
Privacy Policy
Terms of Use
Contact Us
FAQ
© 2021 City of New York. All Rights Reserved. NYC is a trademark and service mark of the City of New York.

NYC Open Data -

opendata.cityofnewyork.us

Update

NYC | OpenData | 311 | Search all NYC.gov websites

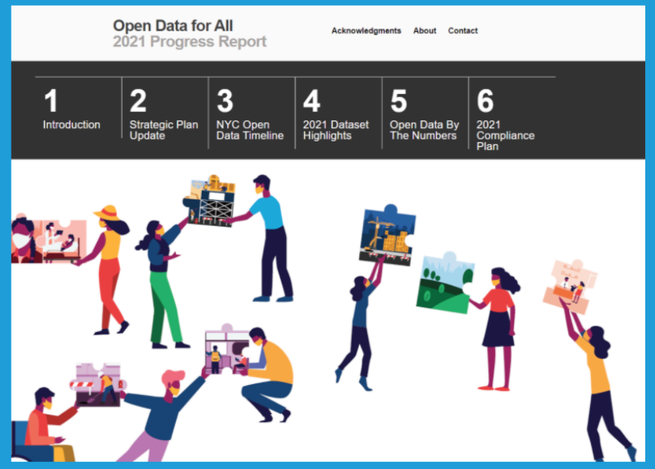
NYC OpenData

Home | Data | About | Learn | Contact Us | Blog | Sign In

Open Data for All New Yorkers

Open Data is free public data published by New York City agencies and other partners. **Share your work during Open Data Week 2021** or **sign up for the NYC Open Data mailing list** to learn about training opportunities and upcoming events.

Search Open Data for things like 311, Buildings, Crime



Learn about the next decade of NYC Open Data, and read our 2021 Report

Translate »

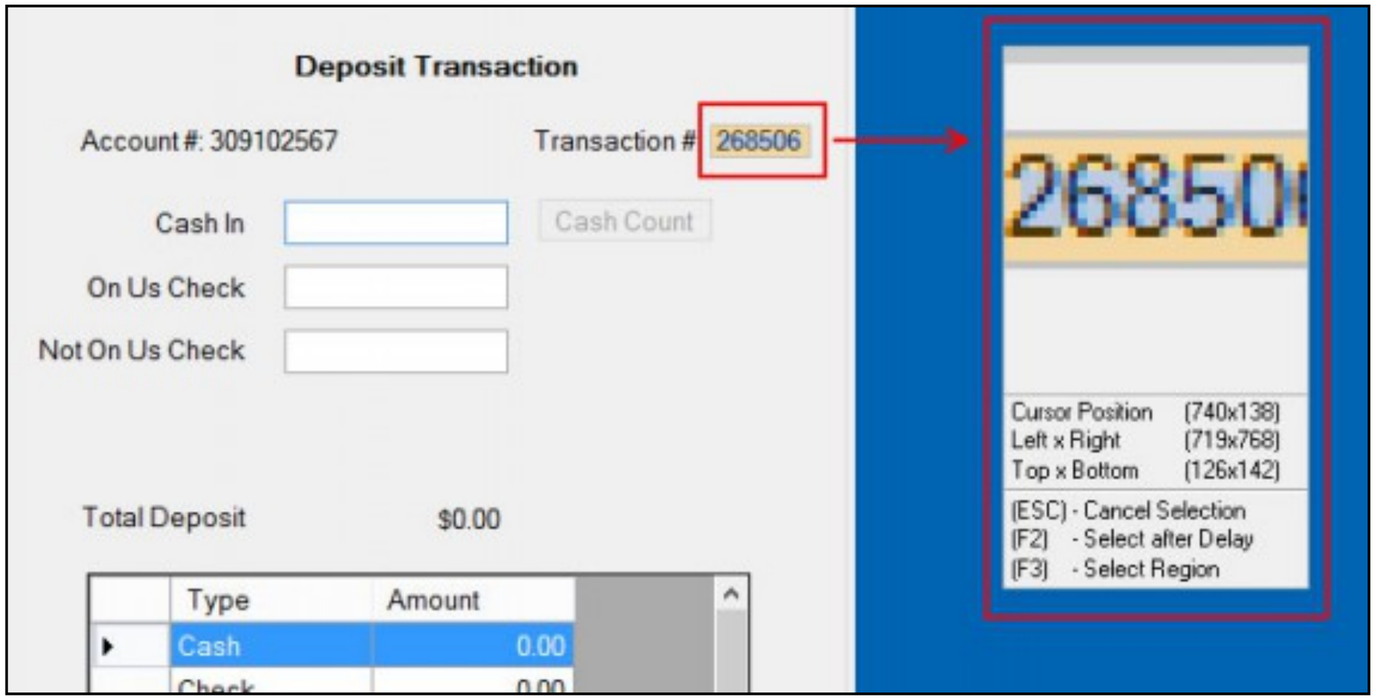
Scraping data from the web

Whenever possible, we would like to find data published for us to use in a sensible format. With luck, data are in an easily accessible portal or, if we ask nicely, someone will email us a copy of the data we are after (email or Dropbox or SecureDrop or whatever seems appropriate given sensitivities around the data).

That is, the verb “scrape” implies we expend some amount of effort, or apply a degree of force. All things being equal, we’d prefer not to work that hard if we don’t have to. **So the first rule of scraping is that we don’t want to do it** — we should always first make sure data are not available through simpler means.

Scraping data from the web

So what do we mean by “scraping” or more generally “data scraping”? Back in the day, we might read data from a program output that was being shown on a screen. A “screen scraper” would read data from the user interface, perhaps simulating user input to collect a multiple fields of information.



Scraping data from the web

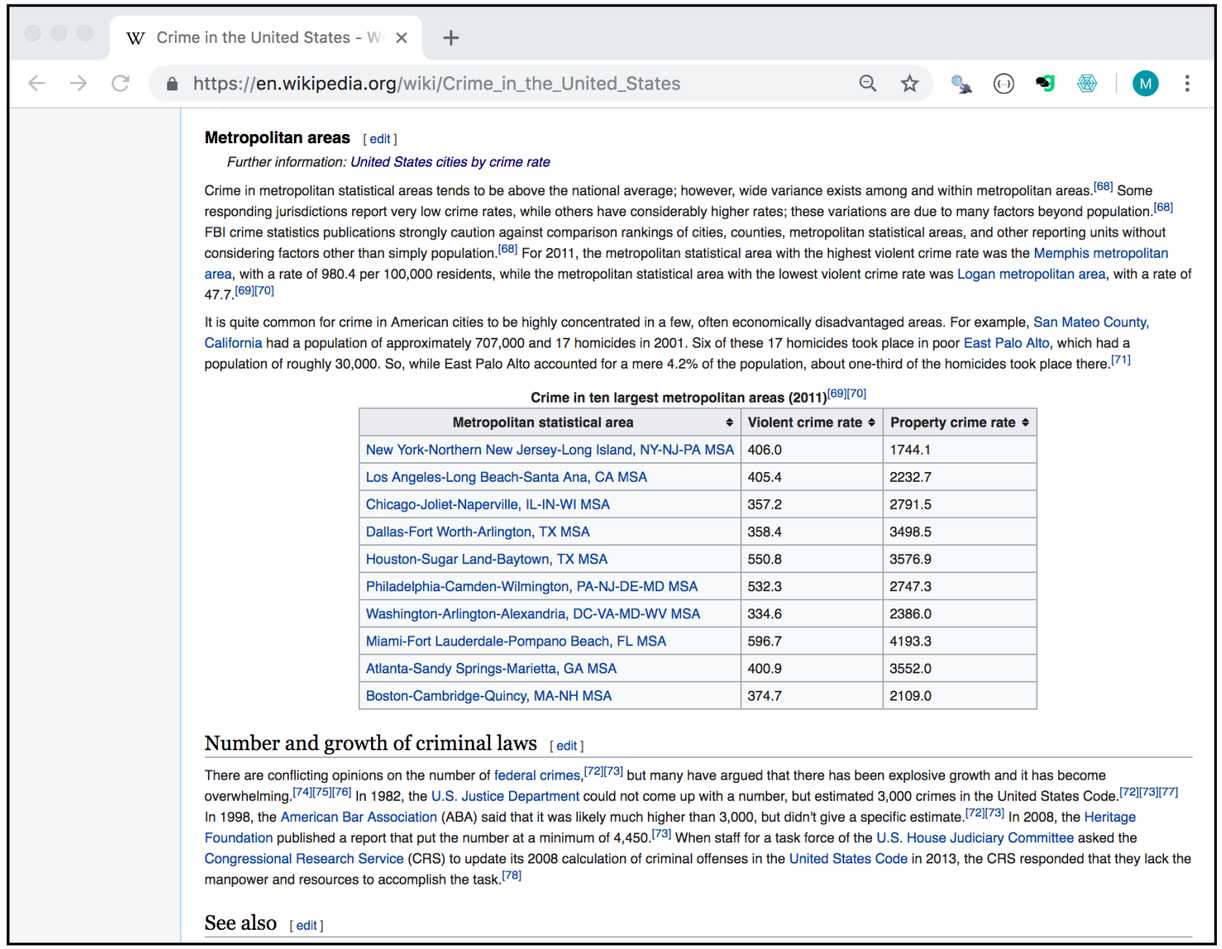
In a similar way, “web scraping” involves taking data that was formatted for one purpose and translating it for another. Let’s take a simple example. Look at the page on crime statistics by city on Wikipedia.

You might be interested in the table that appears toward the bottom of the article describing the crime rates for the largest metropolitan areas in the US.

Scraping data from the web

In this format, on the web page, there is a limit to what you can do with these data aside from simply reading them. And that's basically the point of a web page — information is formatted as a "document" that is meant to be read by a human.

But what if we wanted to compute an average rate for these areas or wanted to compare the violent and property crime rates? OK silly actions for these data but the point is that on the web page they are somewhat "inert".



Metropolitan areas [edit]

Further information: United States cities by crime rate

Crime in metropolitan statistical areas tends to be above the national average; however, wide variance exists among and within metropolitan areas.[68] Some responding jurisdictions report very low crime rates, while others have considerably higher rates; these variations are due to many factors beyond population.[68] FBI crime statistics publications strongly caution against comparison rankings of cities, counties, metropolitan statistical areas, and other reporting units without considering factors other than simply population.[68] For 2011, the metropolitan statistical area with the highest violent crime rate was the Memphis metropolitan area, with a rate of 980.4 per 100,000 residents, while the metropolitan statistical area with the lowest violent crime rate was Logan metropolitan area, with a rate of 47.7.[69][70]

It is quite common for crime in American cities to be highly concentrated in a few, often economically disadvantaged areas. For example, San Mateo County, California had a population of approximately 707,000 and 17 homicides in 2001. Six of these 17 homicides took place in poor East Palo Alto, which had a population of roughly 30,000. So, while East Palo Alto accounted for a mere 4.2% of the population, about one-third of the homicides took place there.[71]

Crime in ten largest metropolitan areas (2011)[69][70]

Metropolitan statistical area	Violent crime rate	Property crime rate
New York-Northern New Jersey-Long Island, NY-NJ-PA MSA	406.0	1744.1
Los Angeles-Long Beach-Santa Ana, CA MSA	405.4	2232.7
Chicago-Joliet-Naperville, IL-IN-WI MSA	357.2	2791.5
Dallas-Fort Worth-Arlington, TX MSA	358.4	3498.5
Houston-Sugar Land-Baytown, TX MSA	550.8	3576.9
Philadelphia-Camden-Wilmington, PA-NJ-DE-MD MSA	532.3	2747.3
Washington-Arlington-Alexandria, DC-VA-MD-WV MSA	334.6	2386.0
Miami-Fort Lauderdale-Pompano Beach, FL MSA	596.7	4193.3
Atlanta-Sandy Springs-Marietta, GA MSA	400.9	3552.0
Boston-Cambridge-Quincy, MA-NH MSA	374.7	2109.0

Number and growth of criminal laws [edit]

There are conflicting opinions on the number of federal crimes,[72][73] but many have argued that there has been explosive growth and it has become overwhelming.[74][75][76] In 1982, the U.S. Justice Department could not come up with a number, but estimated 3,000 crimes in the United States Code.[72][73][77] In 1998, the American Bar Association (ABA) said that it was likely much higher than 3,000, but didn't give a specific estimate.[72][73] In 2008, the Heritage Foundation published a report that put the number at a minimum of 4,450.[73] When staff for a task force of the U.S. House Judiciary Committee asked the Congressional Research Service (CRS) to update its 2008 calculation of criminal offenses in the United States Code in 2013, the CRS responded that they lack the manpower and resources to accomplish the task.[78]

See also [edit]

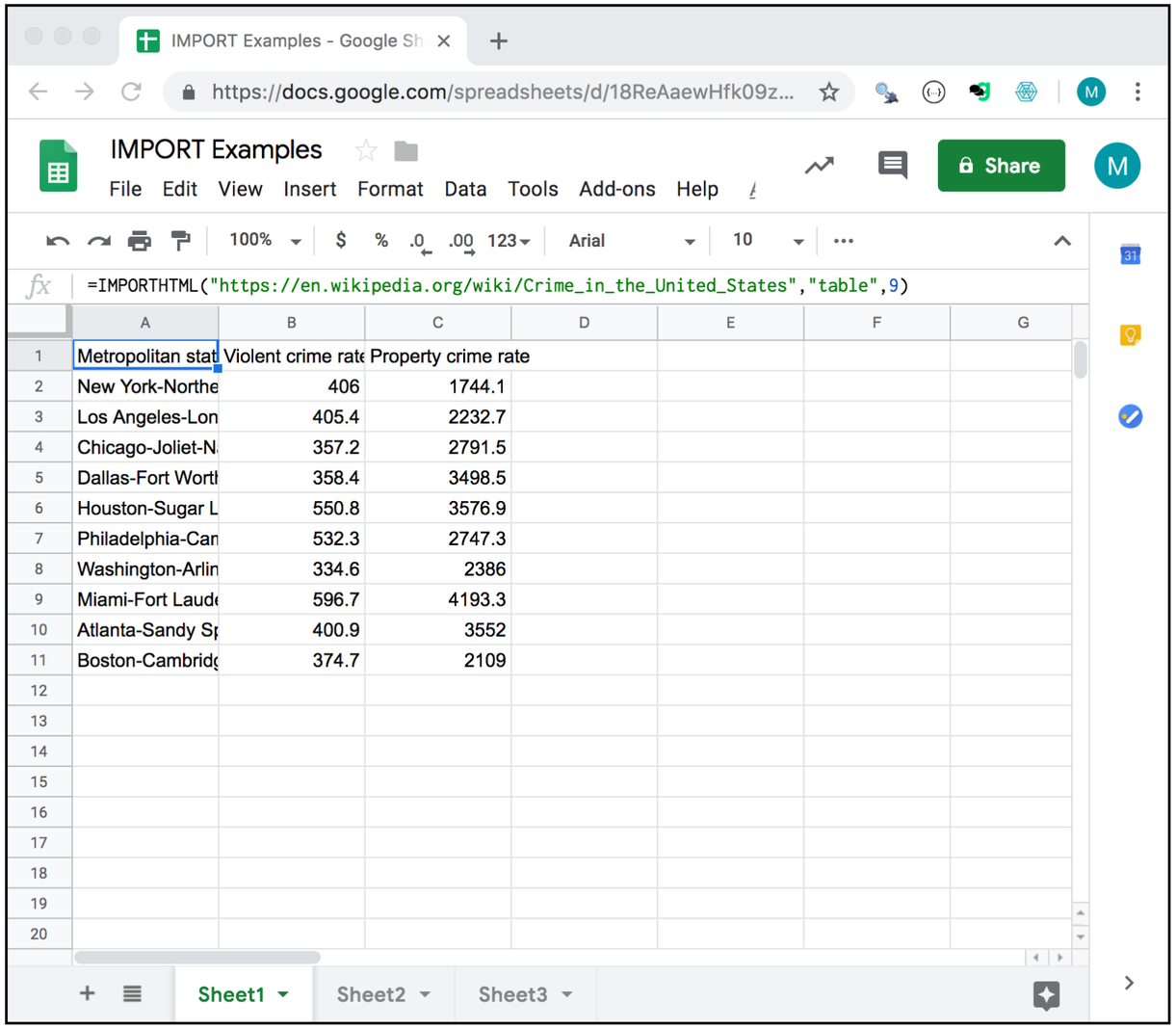
Scraping data from the web

Here we open a new sheet and in the `A1` cell we use the `IMPORTHTML` function. Specifically, you call it using the command

```
=IMPORTHTML( url, item, number)
```

where `url` is the address of the web page we want to scrape a table from, item specifies either a `"table"` or a `"list"` (more on the latter shortly), and `number` specifies which table on the page we want (where we start counting from 1)

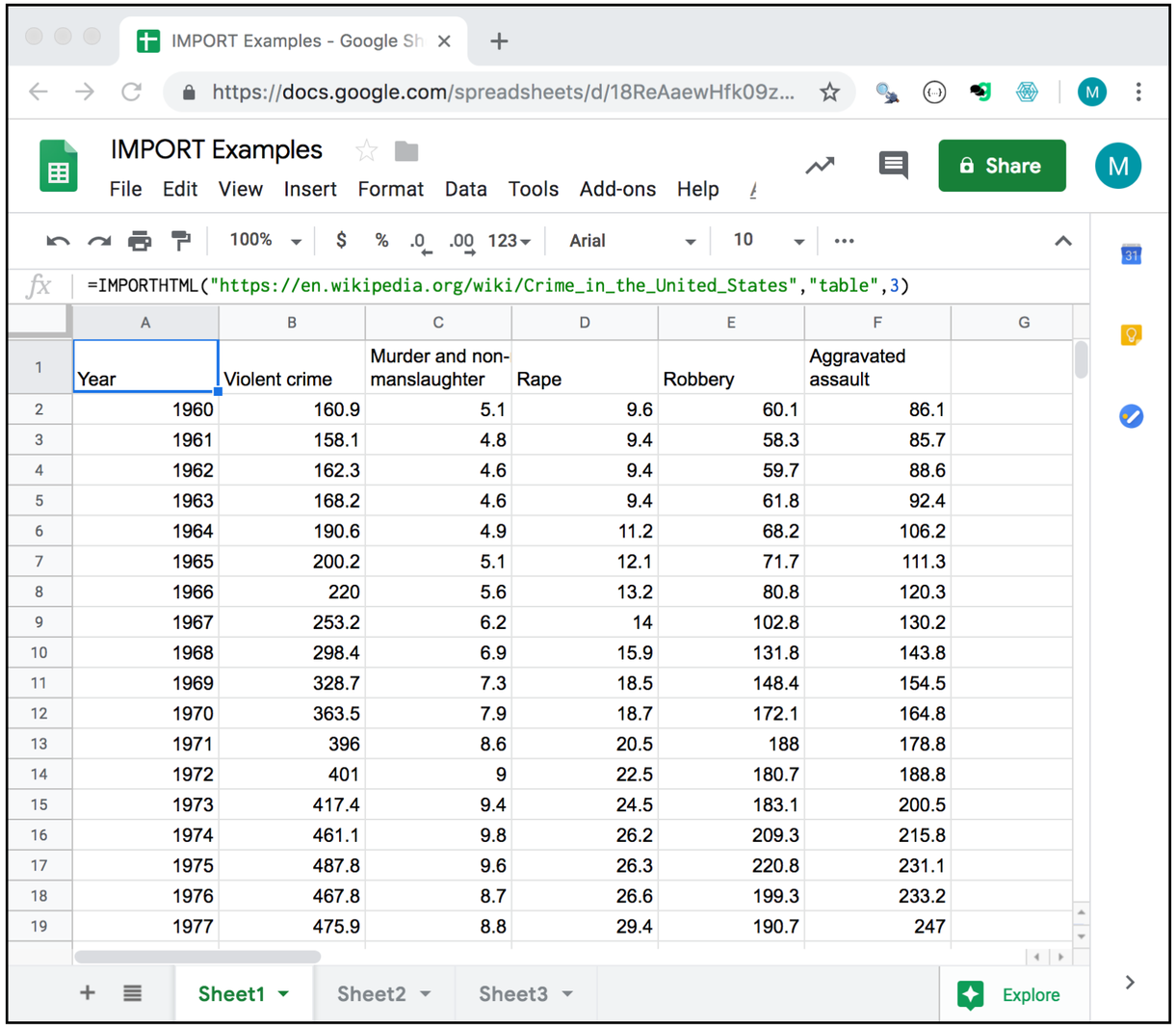
The table with crime rates for large metropolitan areas can be scraped into Google Sheets using the command below in cell `A1` of a blank table.



	A	B	C
1	Metropolitan stat	Violent crime rate	Property crime rate
2	New York-Northe	406	1744.1
3	Los Angeles-Lon	405.4	2232.7
4	Chicago-Joliet-N	357.2	2791.5
5	Dallas-Fort Worth	358.4	3498.5
6	Houston-Sugar L	550.8	3576.9
7	Philadelphia-Can	532.3	2747.3
8	Washington-Arlin	334.6	2386
9	Miami-Fort Laude	596.7	4193.3
10	Atlanta-Sandy Sp	400.9	3552
11	Boston-Cambridg	374.7	2109

```
=IMPORTHTML("https://en.wikipedia.org/wiki/Crime_in_the_United_States","table",8)
```

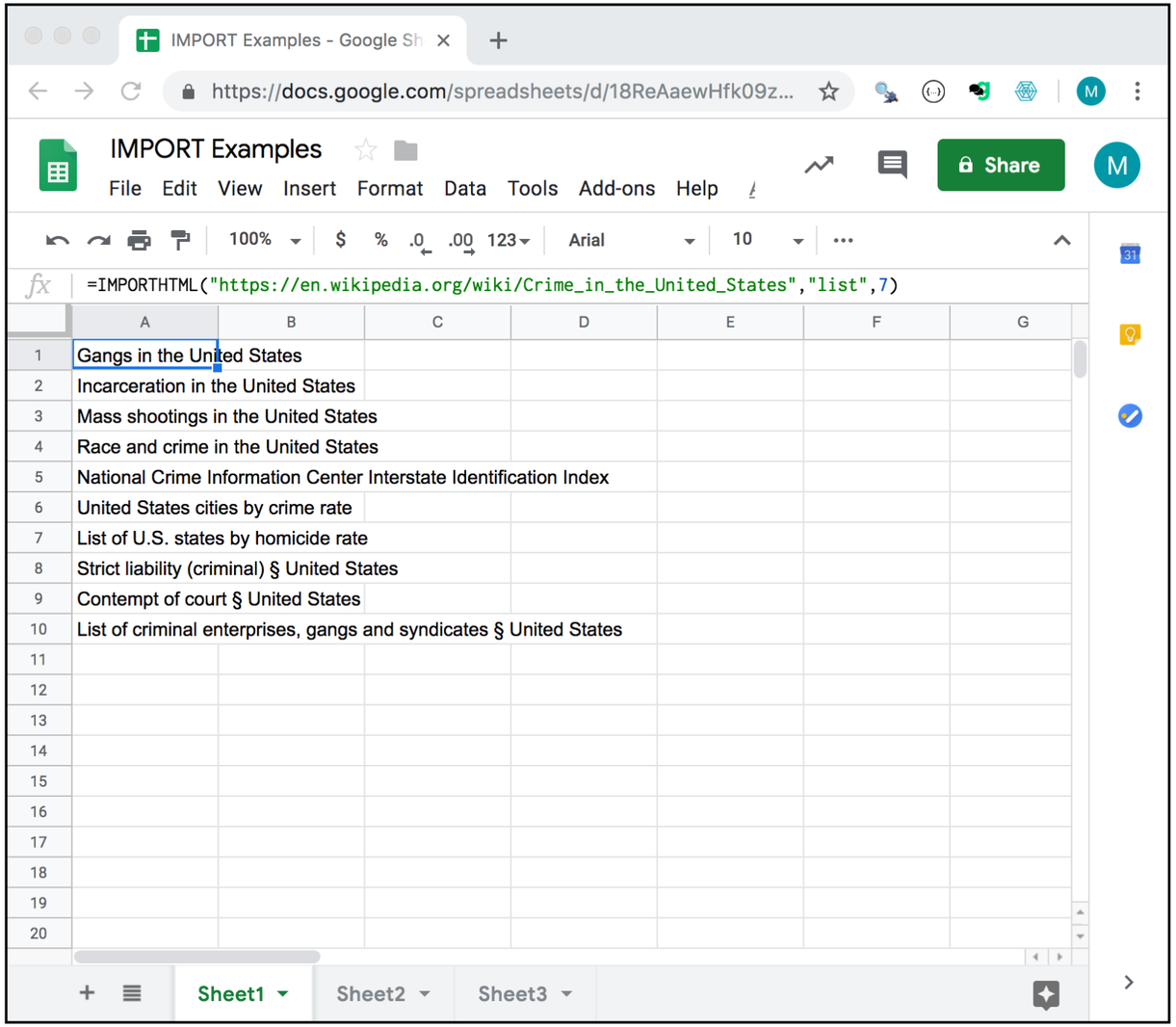
```
=IMPORTHTML("https://en.wikipedia.org/wiki/Crime_in_the_United_States","table",2)
```



	A	B	C	D	E	F	G
1	Year	Violent crime	Murder and non-manslaughter	Rape	Robbery	Aggravated assault	
2	1960	160.9	5.1	9.6	60.1	86.1	
3	1961	158.1	4.8	9.4	58.3	85.7	
4	1962	162.3	4.6	9.4	59.7	88.6	
5	1963	168.2	4.6	9.4	61.8	92.4	
6	1964	190.6	4.9	11.2	68.2	106.2	
7	1965	200.2	5.1	12.1	71.7	111.3	
8	1966	220	5.6	13.2	80.8	120.3	
9	1967	253.2	6.2	14	102.8	130.2	
10	1968	298.4	6.9	15.9	131.8	143.8	
11	1969	328.7	7.3	18.5	148.4	154.5	
12	1970	363.5	7.9	18.7	172.1	164.8	
13	1971	396	8.6	20.5	188	178.8	
14	1972	401	9	22.5	180.7	188.8	
15	1973	417.4	9.4	24.5	183.1	200.5	
16	1974	461.1	9.8	26.2	209.3	215.8	
17	1975	487.8	9.6	26.3	220.8	231.1	
18	1976	467.8	8.7	26.6	199.3	233.2	
19	1977	475.9	8.8	29.4	190.7	247	

Note that table 2 is not visible on the page, but has to be expanded.

```
=IMPORTHTML("https://en.wikipedia.org/wiki/Crime_in_the_United_States","list",11)
```



Specifying “list” instead of “table” gets you access to the lists on the page. Here is the “See Also” list.

Scraping data from the web

Now, the Wikipedia page being displayed in your browser is written in HTML (the Hypertext Markup Language). This is a format that allows you to describe “documents.”

HTML is rich in “tags” that format information for display in a browser. There are tags to indicate titles (headers), paragraphs, lists, tables, and links. You can find a complete list here.

At the right, I’ve put a (slightly simplified) version of the table of crime rates on the Wikipedia page. You can see the content in your browser by “right clicking” on the table and choosing “inspect” — it will open the “Developer Tools” pane in Chrome and let you see the HTML behind this element.

```
<table>
   <tbody>
      <tr>
         <th>Metropolitan statistical area</th>
         <th>Violent crime rate</th>
         <th>Property crime rate</th>
      </tr>
      <tr>
         <td>New York-Northern New Jersey-Long Island</td>
         <td>406.0</td>
         <td>1744.1</td>
      </tr>
      <tr>
         <td>Los Angeles-Long Beach-Santa Ana, CA MSA</td>
         <td>405.4</td>
         <td>2232.7</td>
      </tr>
      <tr>
         <td>Chicago-Joliet-Naperville, IL-IN-WI MSA</td>
         <td>357.2</td>
         <td>2791.5</td>
      </tr>
      <tr>
         <td>Dallas-Fort Worth-Arlington, TX MSA</td>
         <td>358.4</td>
         <td>3498.5</td>
      </tr>
      ...
   </tbody>
</table>
```

Scraping data from the web

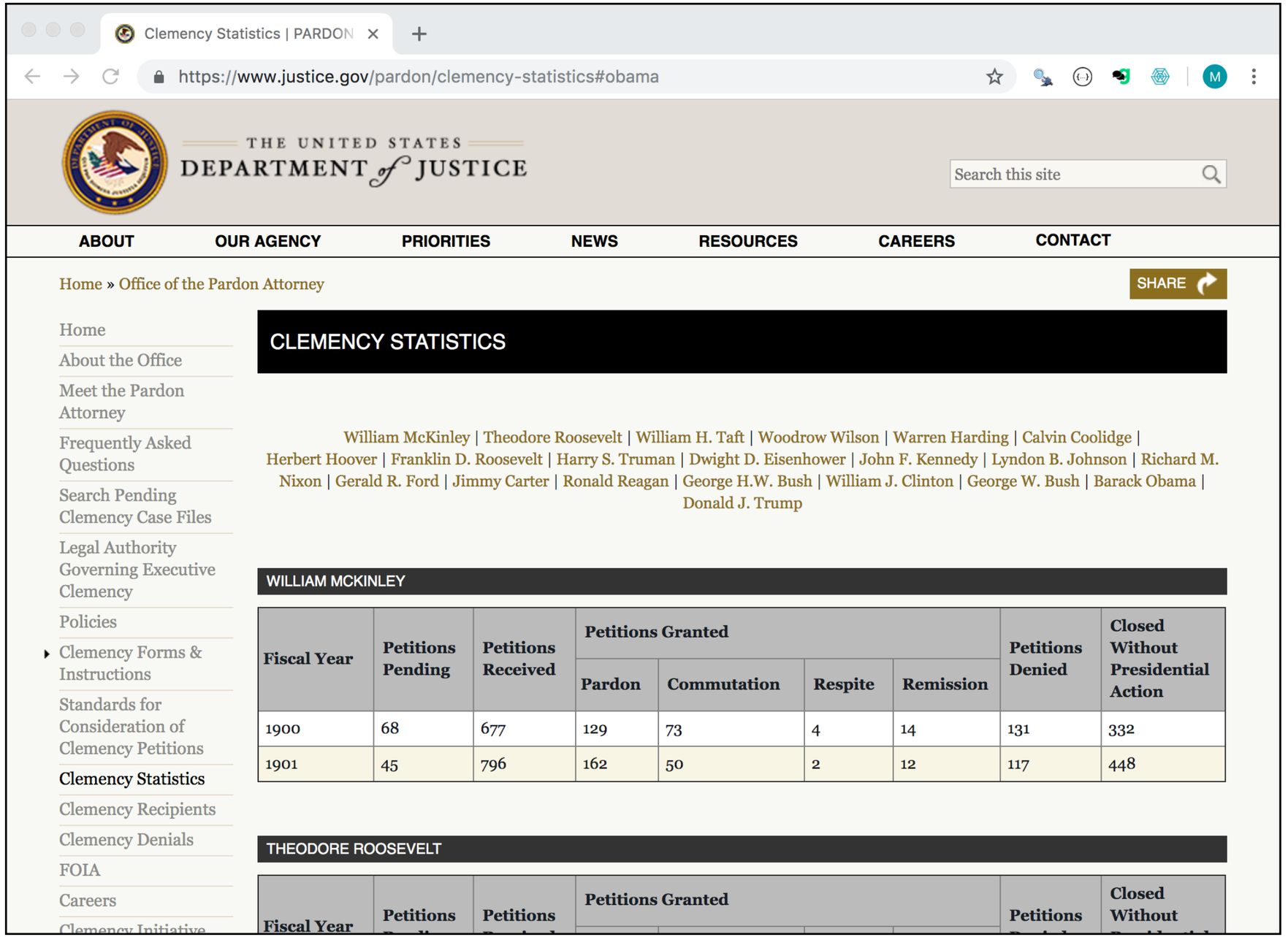
This should give you a good sense of the difference between data stored (or marked up) for a web page and data in a Google Sheet (well, any spreadsheet).

In general, the act of web scraping takes data in one perhaps unusable format and translates it into something we can compute with.

In the case of lists or tables, this translation is fairly direct (assuming people have entered data in nicely). But often, the elements we want to pull are not so direct, making scraping real work.

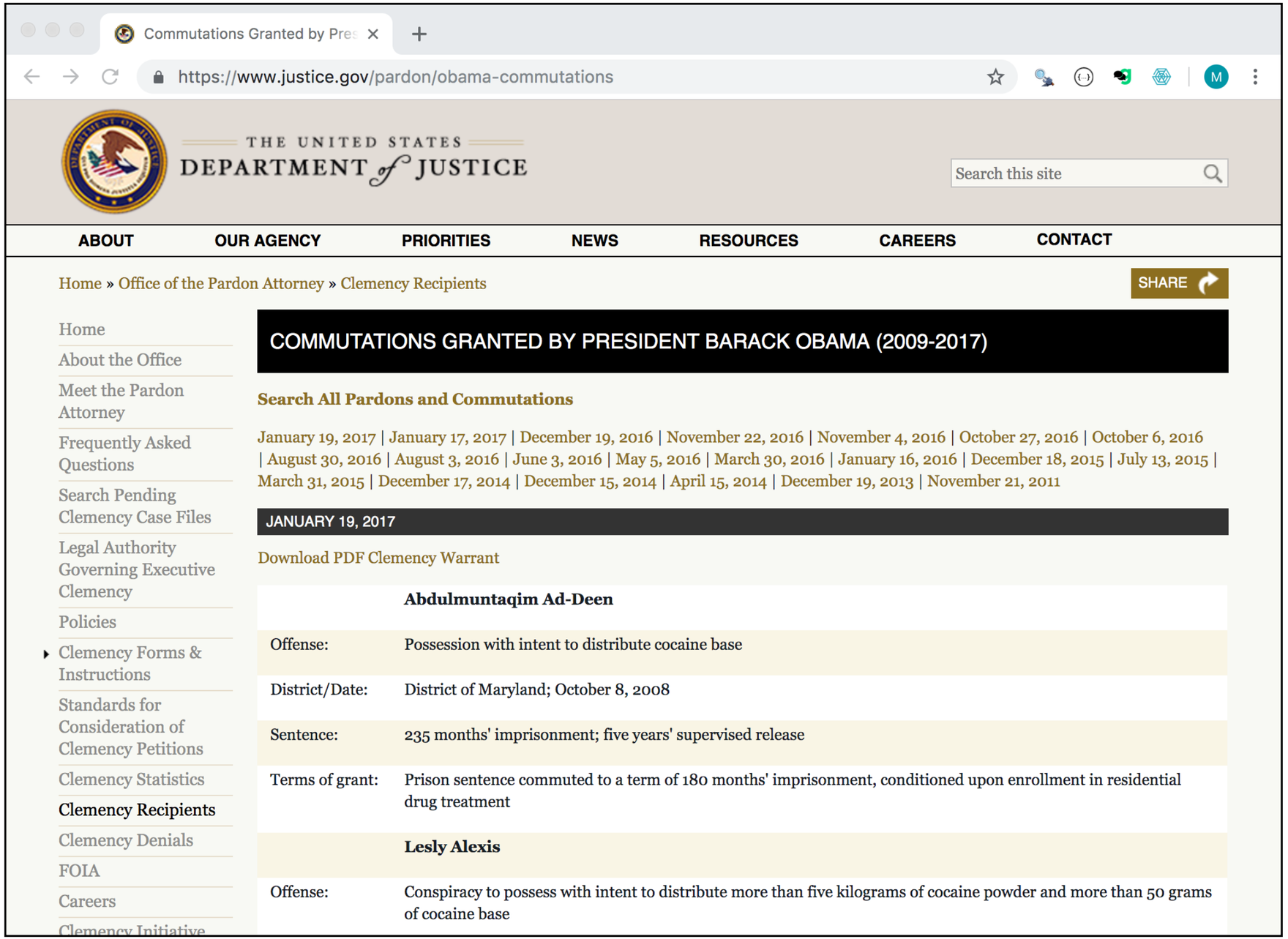
```
<ul>
  <li>
    <a href="/wiki/Gangs in the United States" title="Gangs in the
    United States">Gangs in the United States</a>
  </li>
  <li>
    <a href="/wiki/Incarceration in the United States"
    title="Incarceration in the United States">Incarceration in the
    United States</a>
  </li>
  <li>
    <a href="/wiki/Mass shootings in the United States" title="Mass
    shootings in the United States">Mass shootings in the United
    States</a>
  </li>
  <li>
    <a href="/wiki/Race and crime in the United States" title="Race
    and crim e in the United States">Race and crime in the United
    States</a>
  </li>
  …

  <li>
    <a href="/wiki/Contempt of court#United States" title="Contempt of
    court">Contempt of court § United States</a>
  </li>
  <li>
    <a href="/wiki/
    List of criminal enterprises, gangs and syndicates#United States"
    title="List of criminal enterprises, gangs and syndicates">List of
    criminal enterprises, gangs and syndicates § United States</
    a>
  </li>
</ul>
```



Another example

Here are a series of tables with clemency statistics for presidents going back to McKinley.



Another example

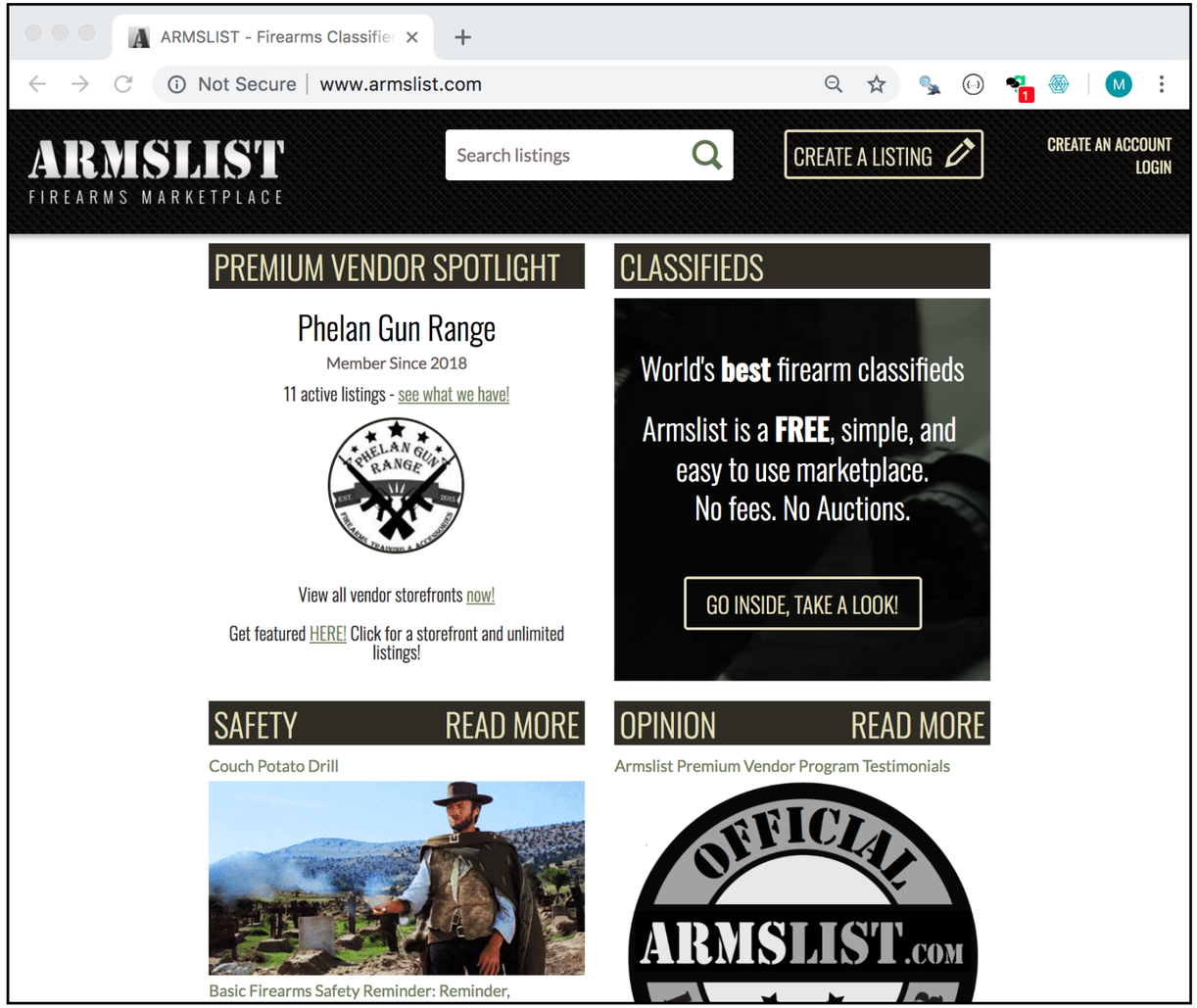
Who received a pardon or commutation is also available in a table, but a really awkward one… why?

Scraping data from the web

Here's another example, keeping with the very cheerful crime theme we have going. This is a site called ARMSLIST and was called out in the NYTimes for selling guns across state lines.

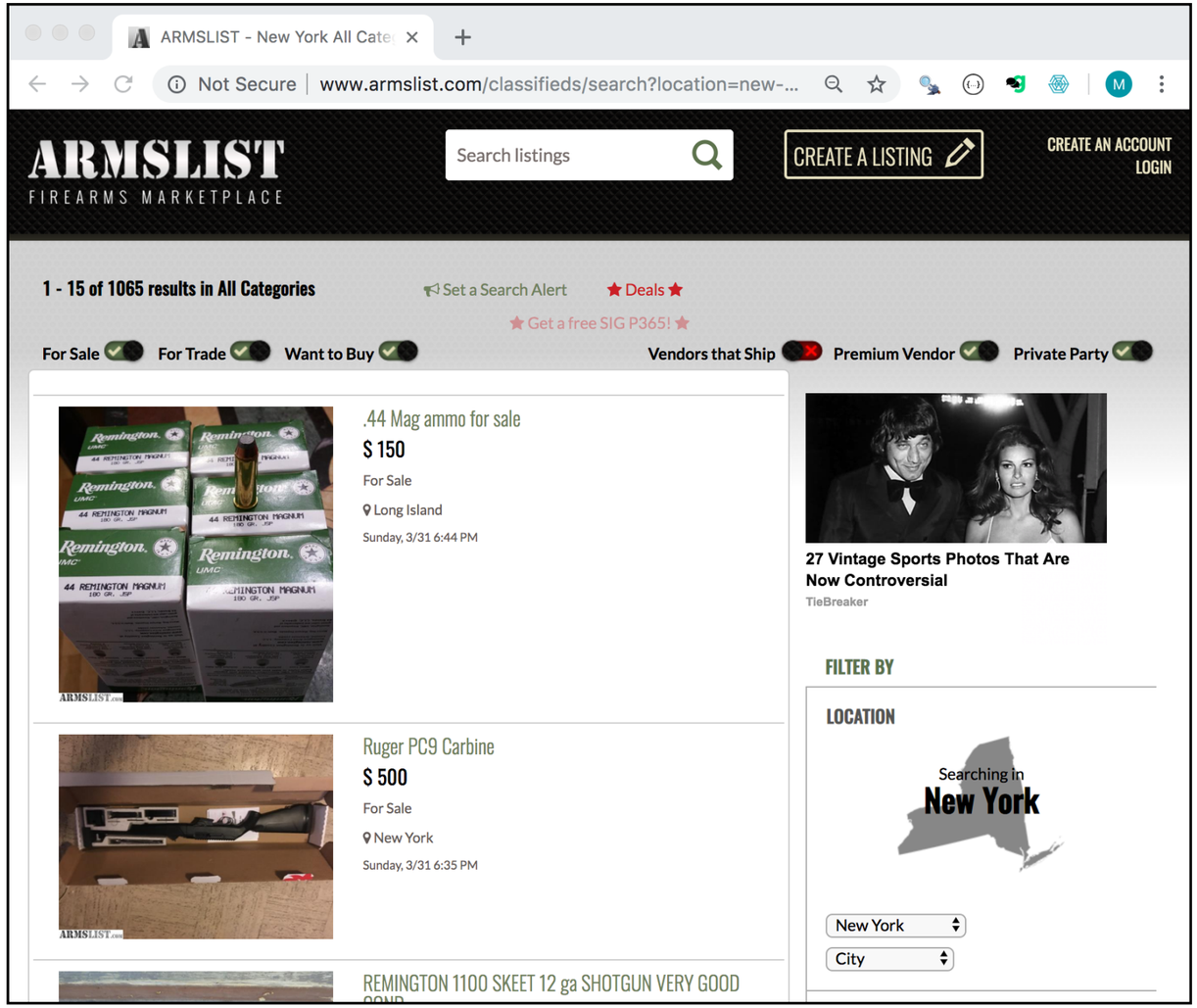
Select their "Go Inside, Take a Look" link under the "Classifieds" heading. Then select "New York" state (or any state you want really).

Here's what you get…



Scraping data from the web

You now get a long list of elements that each represent a firearm or ammunition for sale. But if you right click to “Inspect” the elements, you find that the data are not contained in a table.



Inspecting a web page

Instead, aspects of each item are divided into different paragraphs, different headings and then a series of nested `<div>` tags (these define divisions or sections in an HTML page).

Here's what you get from inspecting the elements. I've edited it down a little so that the main structure is clearer. Have a look at this and see different tags that have been assembled to present a single item for sale.

This is now where things get interesting. When we don't have a container like `<table>` or `<ul>` to tell us where items are on the page, how are we going to find them?

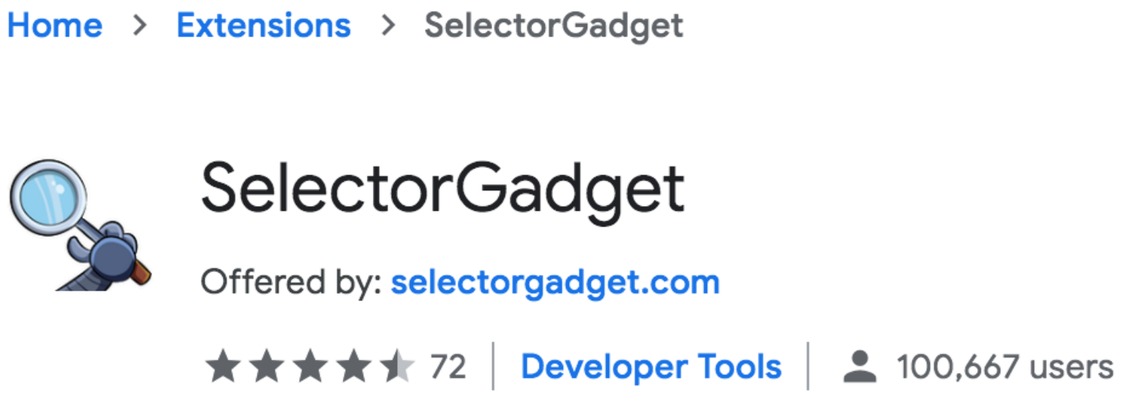
```
<div class="container-fluid">
  <div class="row">
    <div class="col-md-5">
      <a href="44-mag-ammo-for-sale">
        <img src="https://s3.amazon.com/4mag_ammo_for_sale_640.jpg"/>
      </a>
    </div>
    <div class="col-md-7">
      <h4>
        <a href="44-mag-ammo-for-sale">.44 Mag ammo for sale</a>
      </h4>
      <h4> $ 150 </h4>
      <p> <small>For Sale</small> </p>
      <p> <small>Long Island</small> </p>
      <p> <small>Sunday, 3/31 6:44 PM</small> </p>
    </div>
  </div>
</div>
```

Inspecting a web page

The Chrome extension SelectorGadget helps you identify patterns in how a web page is authored to help you identify the information you'd like to extract.

Whereas all we needed with our Google Sheets scraping tool was to know which table or list we wanted, for less structured data, we need to know what tags contain the data we're after.

Let's install the extension and see what it does with the ARMSLIST page. Let's find, say, the prices of the munitions.



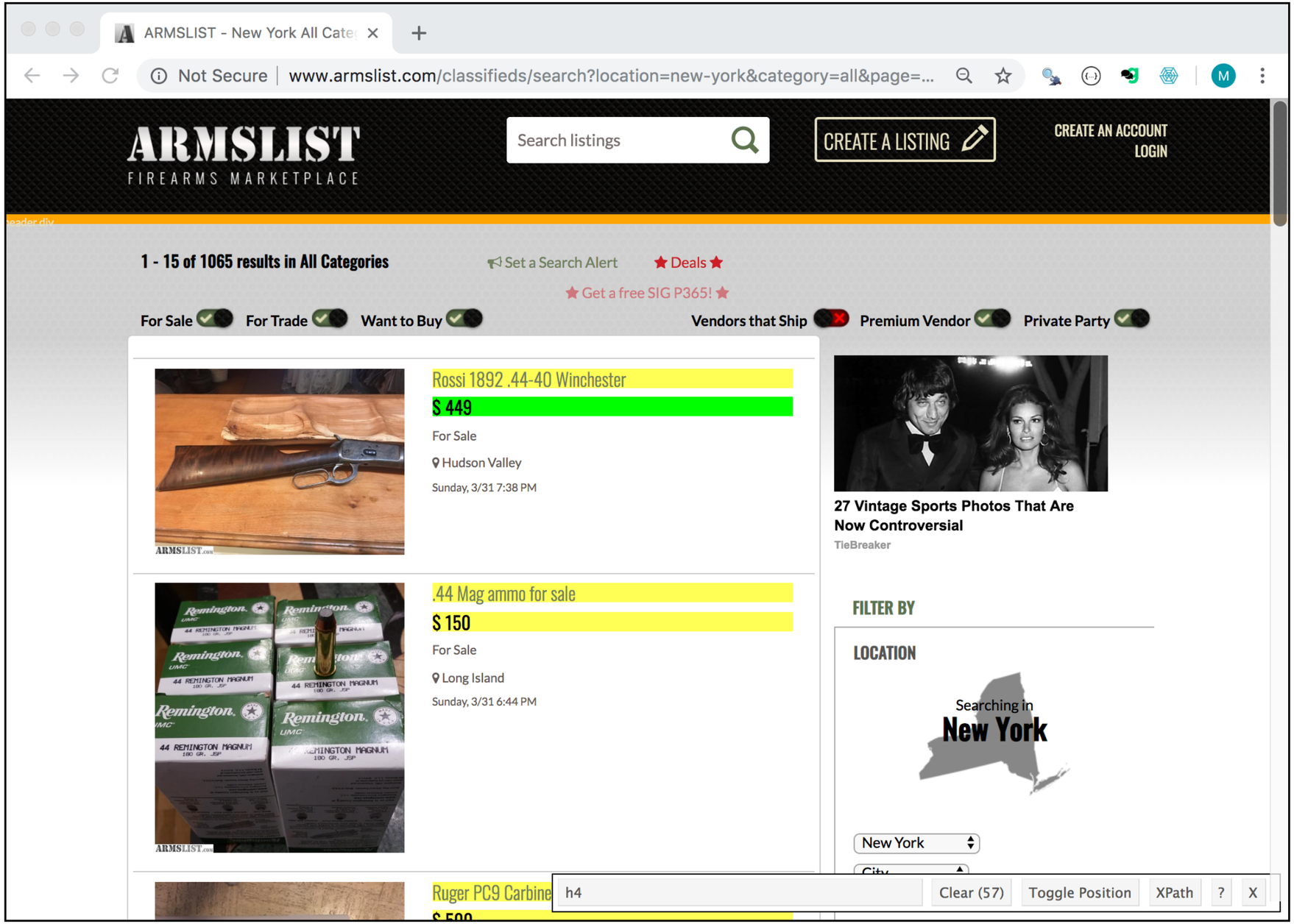
Inspecting a web page

After you start SelectorGadget, you will see a grey bar at the bottom of your screen. It is going to compile a “description” of where the data you’ve highlighted can be found.

So, if we click on $449 (the first price in the list of arms for New York State when I ran this) you will get a green line overlaid on the price and then a number of other things will turn yellow.

The bar at the bottom indicates that the price is included in a header tag `<h4>` (which we can verify from the snippet of the web page I included 2 slides back).

But other data are in `<h4>` tags also, including the name of the time for sale, as indicated by its yellow overlay.



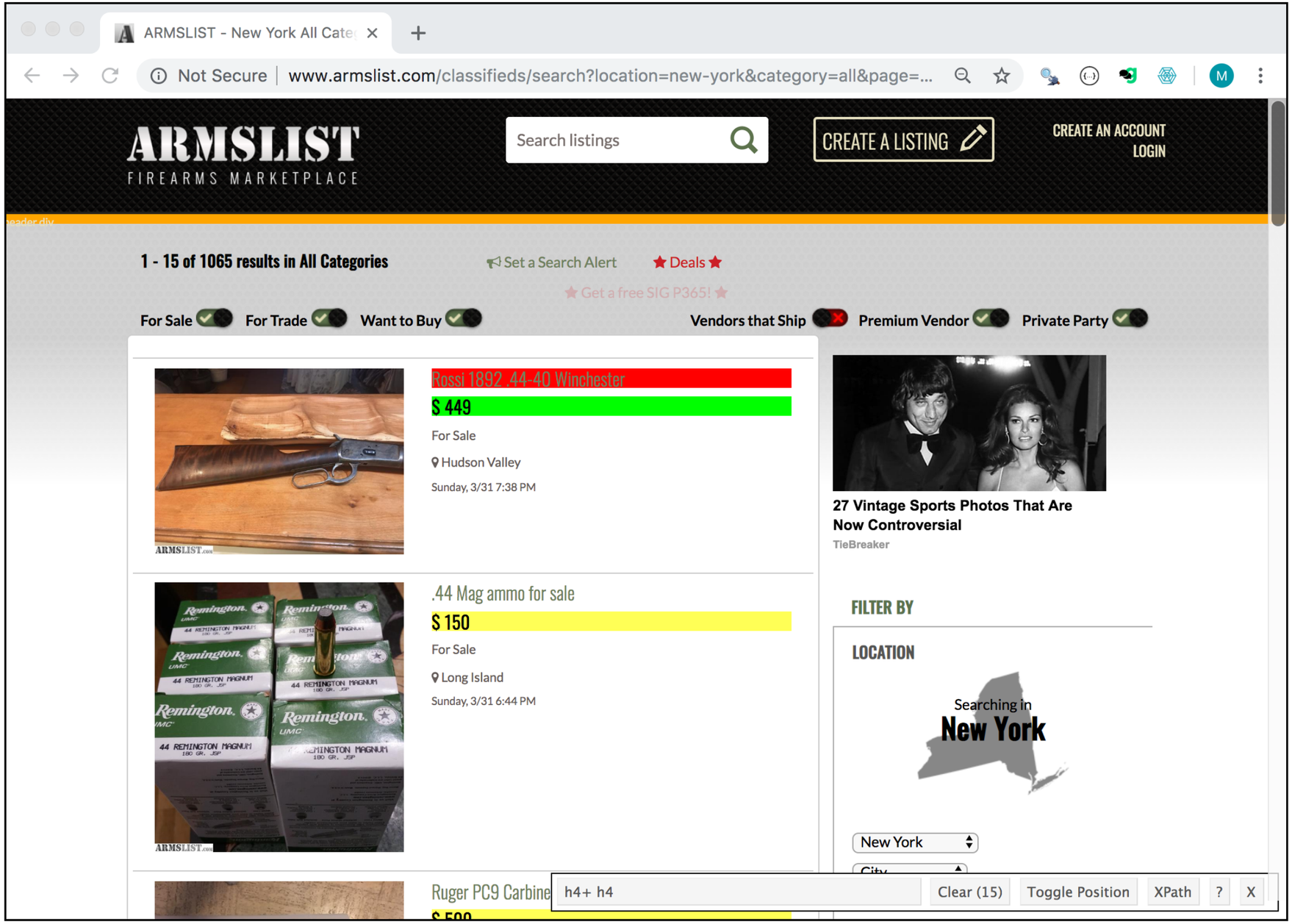
Inspecting a web page

To remove the titles from our selection, we can click on one and it will turn red (see the image on the right). The subsequent names are now no longer yellow as well.

Ah but the bar is now a bit more complex. It reads h4+h4. More on that shortly.

But it looks like the only elements in yellow are the prices we are looking for! This has given us a way to target the data we want.

That is, once we understand what the descriptor coming from SelectorGadget actually means…



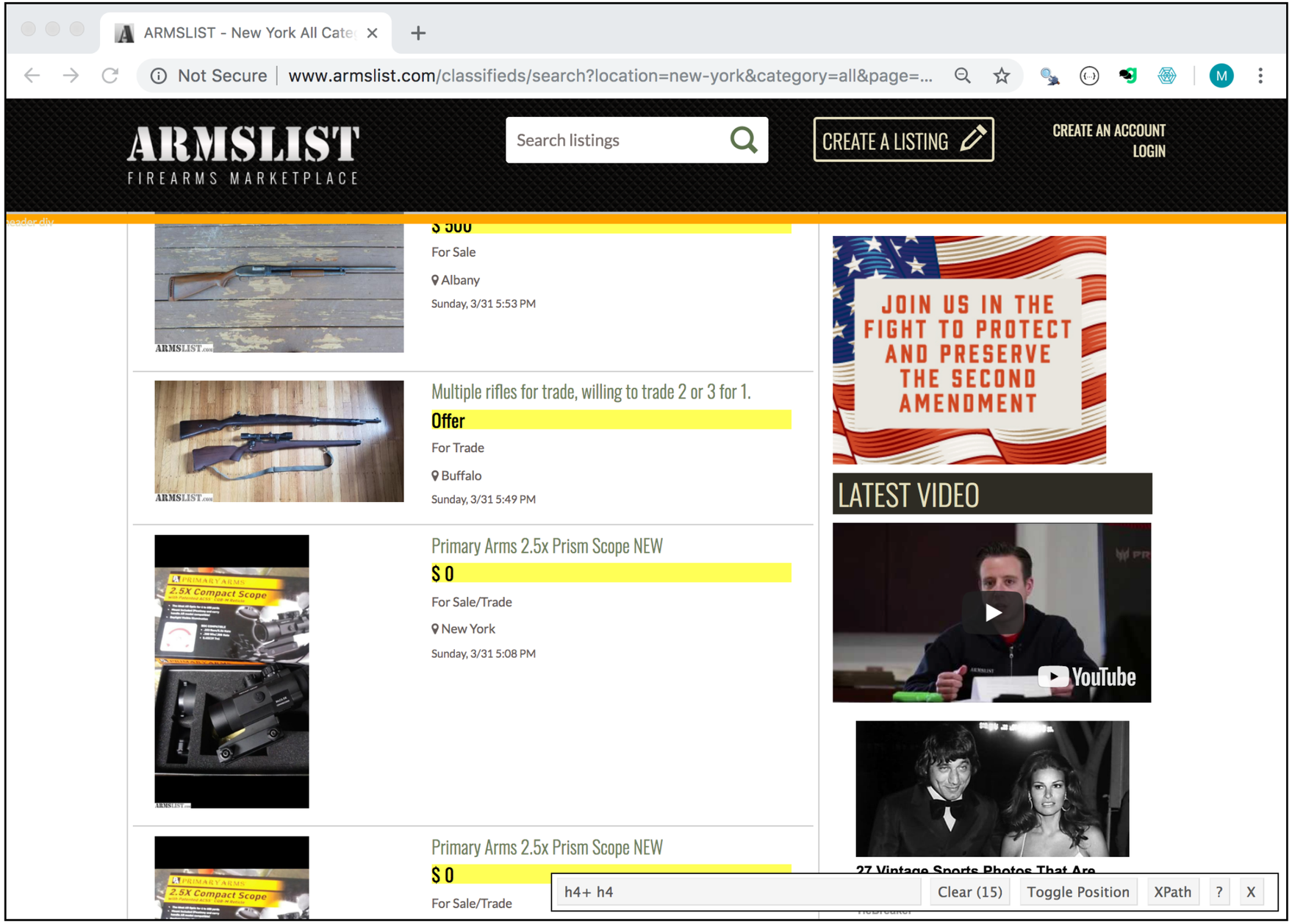
Inspecting a web page

To remove the titles from our selection, we can click on one and it will turn red (see the image on the right). The subsequent names are now no longer yellow as well.

Ah but the bar is now a bit more complex. It reads h4+h4. More on that shortly.

But it looks like the only elements in yellow are the prices we are looking for! This has given us a way to target the data we want.

That is, once we understand what the descriptor coming from SelectorGadget actually means…



Inspecting a web page

So what do we have? SelectorGadget has 15 matches (the number of items for sale on this page) and the pattern it has identified is `h4+h4`.

If you look at the snippet of HTML 5 slides back, you see that the price is in an `<h4>` tag that is followed by an `<h4>` tag containing the listing's name.

The expression `h4+h4` is called a "CSS selector" and this particular selector identifies the content in "an `h4` tag that follows an `h4` tag."

OK so, what is CSS? And what is a CSS selector?

CSS

CSS stands for “Cascading Style Sheets” and is a language that describes the “style” of an HTML document. With it, we can say that the background of the body of our document should be light blue, that <h1> headers should be white, with the text centered, our the text in a <p> paragraph should be rendered in Verdana with size 20 pixels.

CSS selectors are then used to find elements to apply a style to. The examples on the right are just single tags (the body, a header, a paragraph), but things can get arbitrarily more complex.

You can read about CSS and selectors here. And there’s an interesting illustration of the different CSS constructions here.

```
body {
  background-color: lightblue;
}

h1 {
  color: white;
  text-align: center;
}

p {
  font-family: verdana;
  font-size: 20px;
}
```

Click a selector:

#Lastname
.intro, #Lastname
h1
h1, p
div p
div > p
ul + p
ul ~ table
*
[id]
[id=my-Address]
[id$=ess]
[id|=my]
[id^=L]
[title~=beautiful]
[id*=s]
:checked
:disabled
:enabled
:empty
:focus
p:first-child
p::first-letter
p::first-line
p:first-of-type
h1:hover
input:in-range
input:out-of-range
input:invalid
input:valid
p:lang(it)
p:last-child
p:last-of-type

Selector:

ul + p

The element that are next to each <ul> elements.

Result:

<h1> **Welcome to My Homepage** </h1>

My name is Donald Duck.

I live in Duckburg

I have many friends:

<ul id="Listfriends>

- Goofy
- Mickey
- Daisy
- Pluto

All my friends are great!
But I really like Daisy!!

Ciao bella

<h3> **We are all animals!** </h3>

  My latest discoveries have led me to believe that we are all animals:

<table>

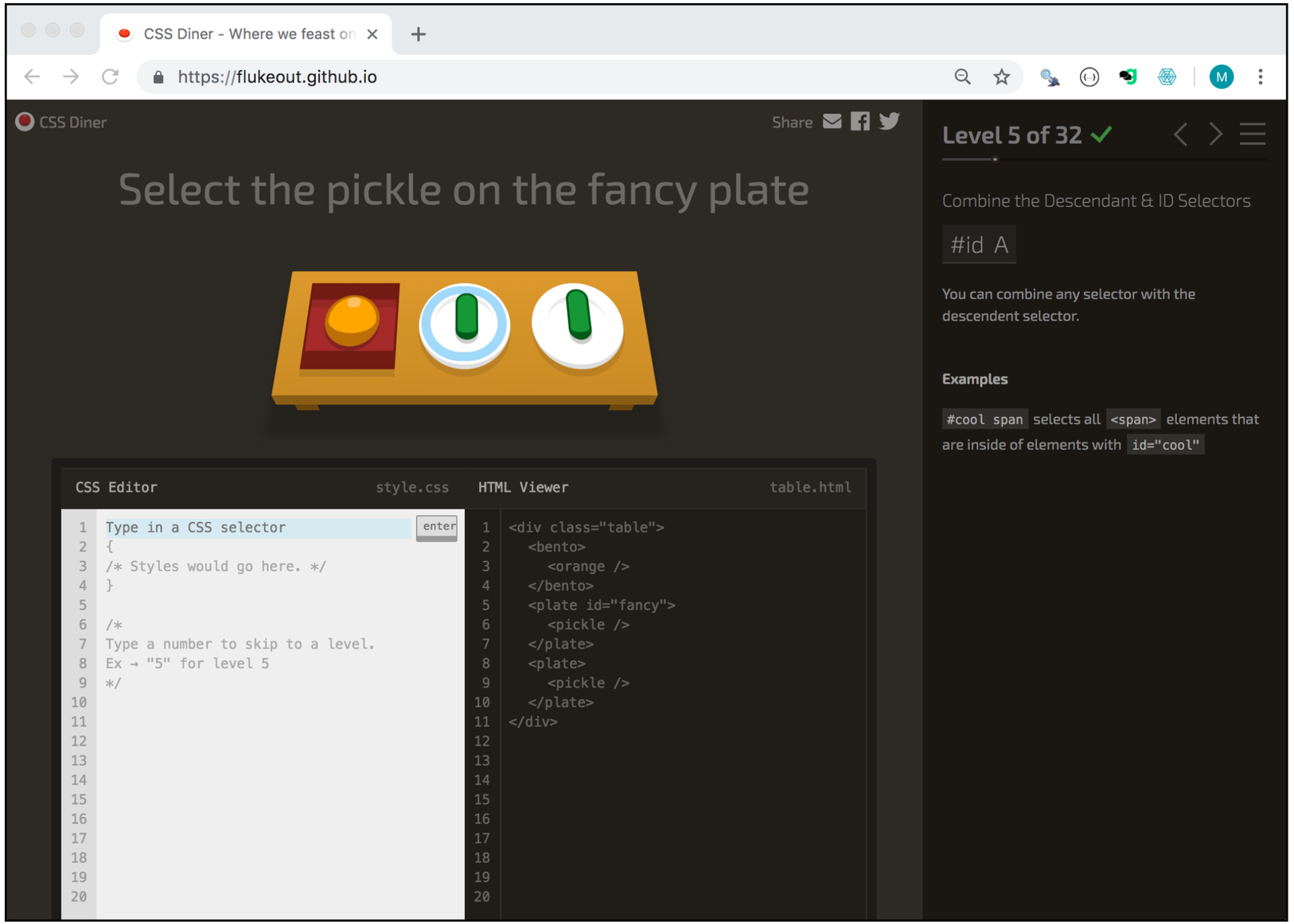
Name	Type of Animal

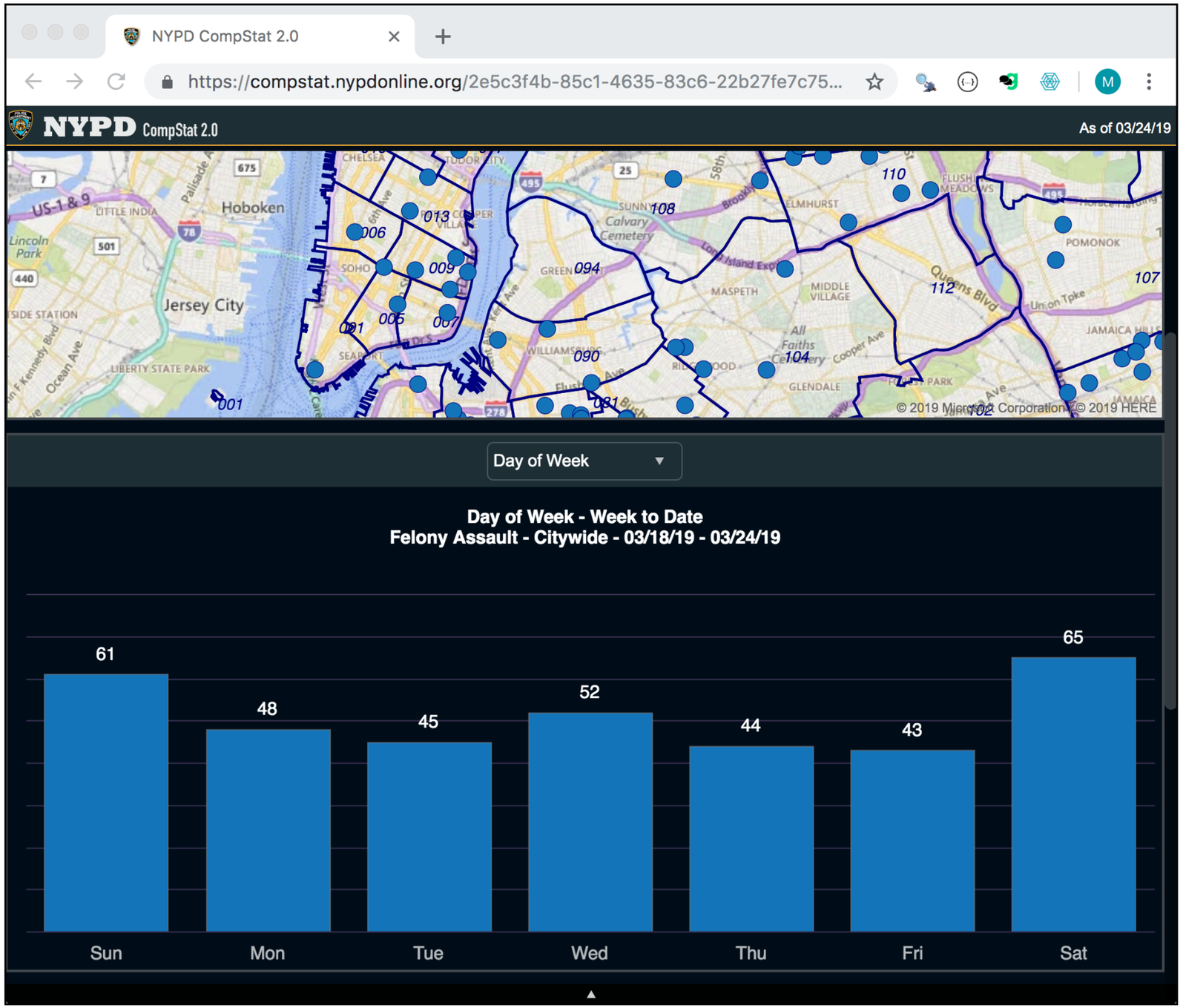
CSS

As is the case with anything on the web, there are far more entertaining ways to learn about CSS selectors.

The CSS Diner is a fun (well, that depends on your definition of “fun” I guess) game that will teach you pretty painlessly what each of the selector expressions does.

Again, the “selector” in CSS selector is meant to identify content to apply a style to (color, linewidth, size). **Notice that we are using them to find the data we want to extract!**

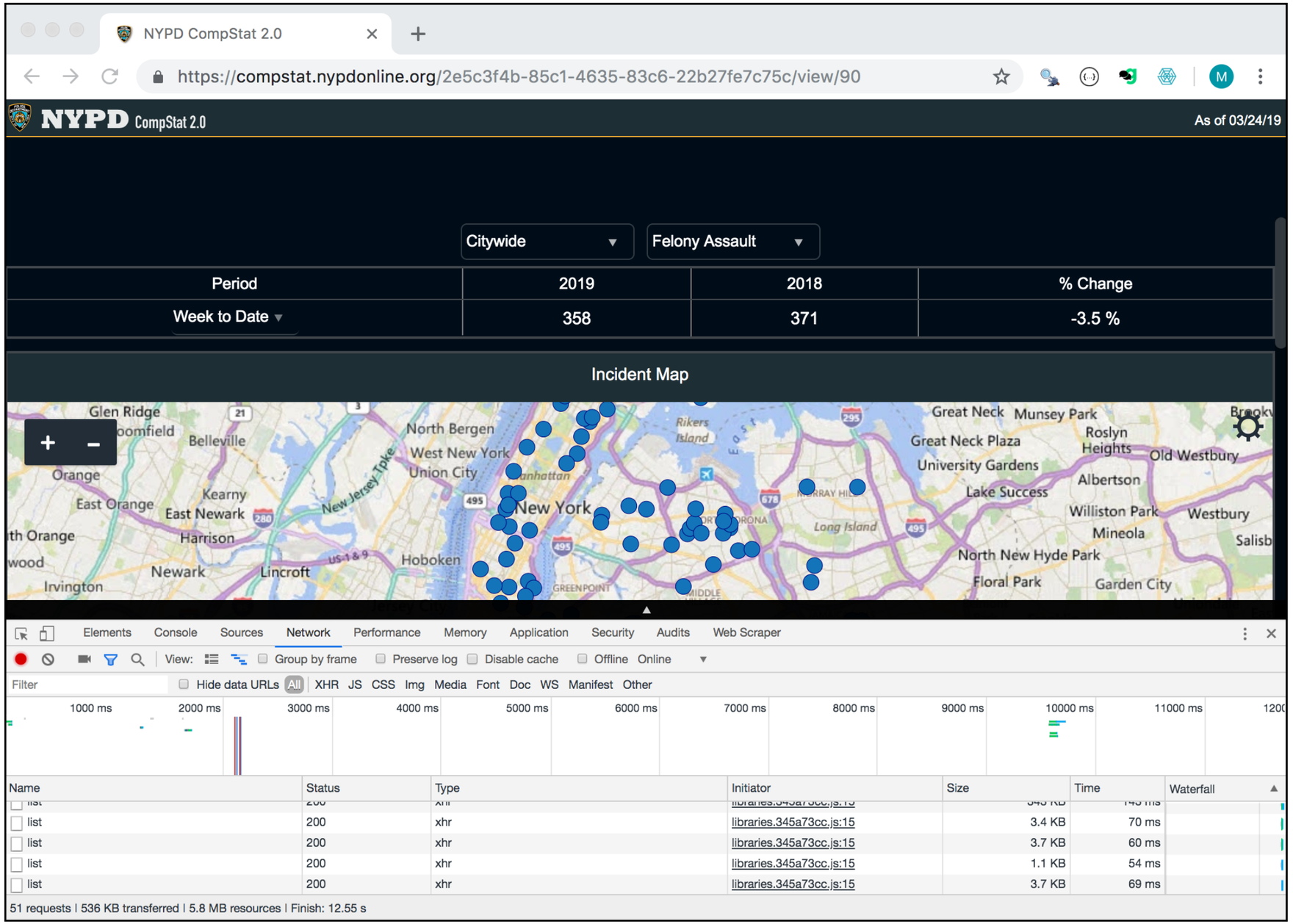




The Developer Tools

The developer tools are pretty powerful on their own. You can, for example, watch the “network” activity as CompStat 2.0 calls back to its server for data as you make selections.

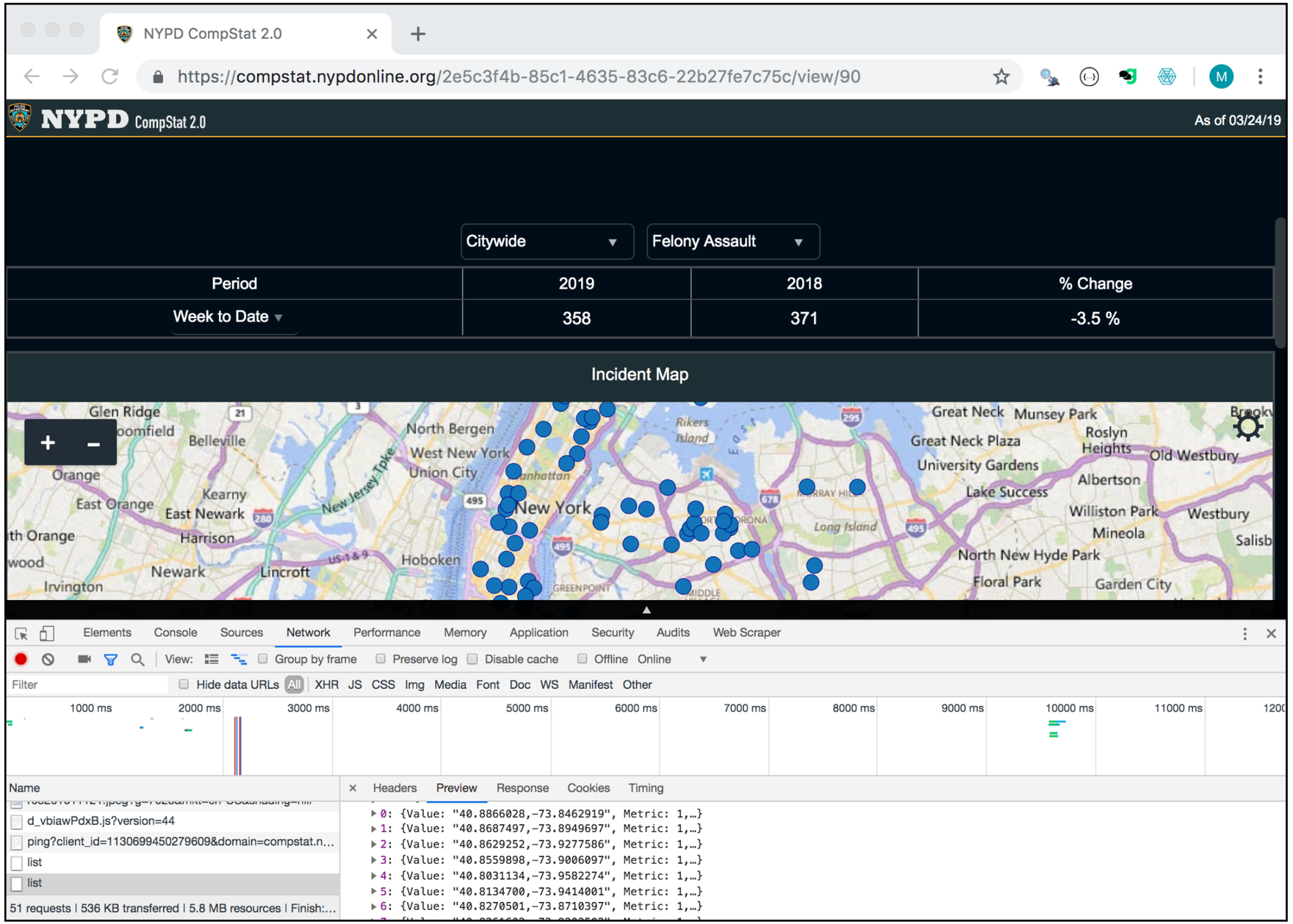
Whenever you see an interface like this, you should wonder where the data are coming from and if you can grab them. It’s always much easier than literally scraping.



The Developer Tools

The developer tools are pretty powerful on their own. You can, for example, watch the "network" activity as CompStat 2.0 calls back to its server for data as you make selections.

Whenever you see an interface like this, you should wonder where the data are coming from and if you can grab them. It's always much easier than literally scraping.



The Developer Tools

The developer tools are pretty powerful on their own. You can, for example, watch the “network” activity as CompStat 2.0 calls back to its server for data as you make selections.

Whenever you see an interface like this, you should wonder where the data are coming from and if you can grab them. It’s always much easier than literally scraping.