Introduction to folders, files, and software

the stuff you'll use for data projects

by Martin Frigaard

Written: September 21 2021

Updated: November 30 2021

Created using the "λέξις" theme

Outline

What is data journalism?

- *Definitions & Examples*

Software

Open source & Proprietary

Code files

R Code, HTML & CSS

Data files

.CSV, .XML & .JSON

What is data journalism?

Wikipedia:

a journalistic process based on analyzing and filtering large data sets for the purpose of creating or elevating a news story

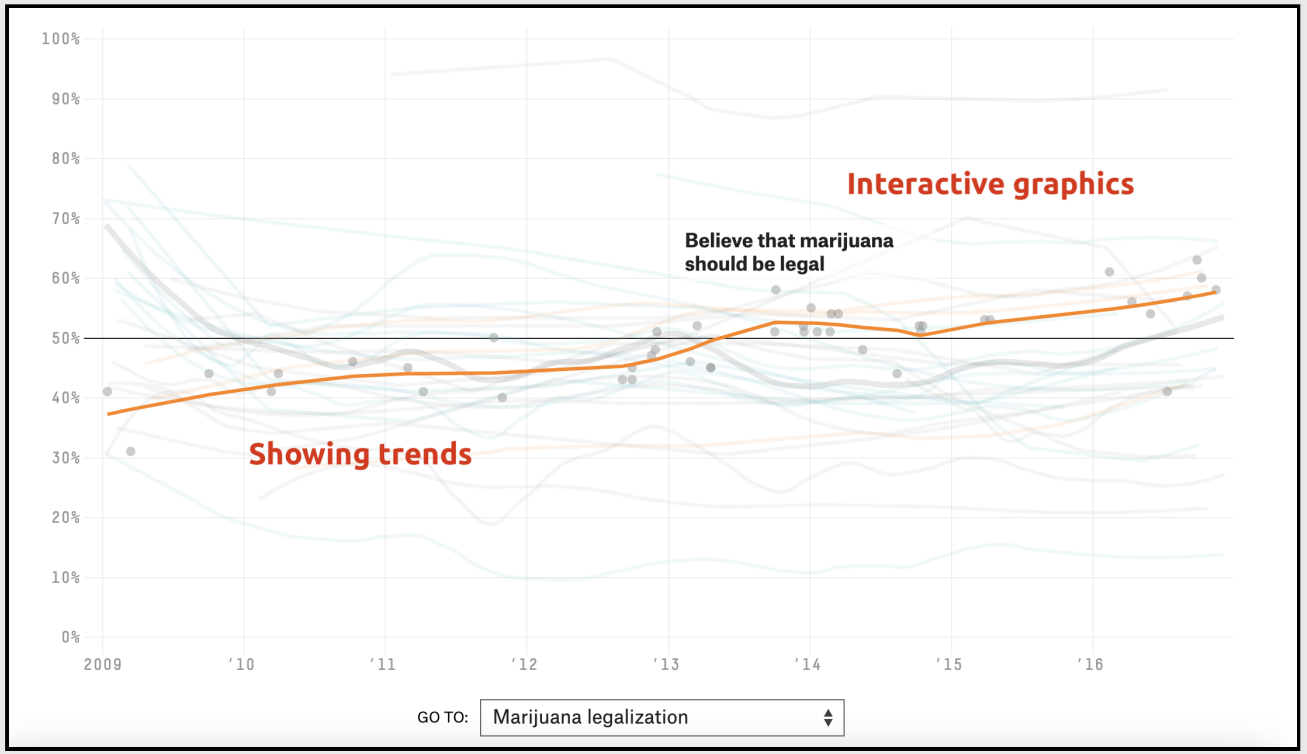
fivethirtyeight:

Data-driven news and analysis

the upshot (NYT):

Analytical journalism in words and graphics

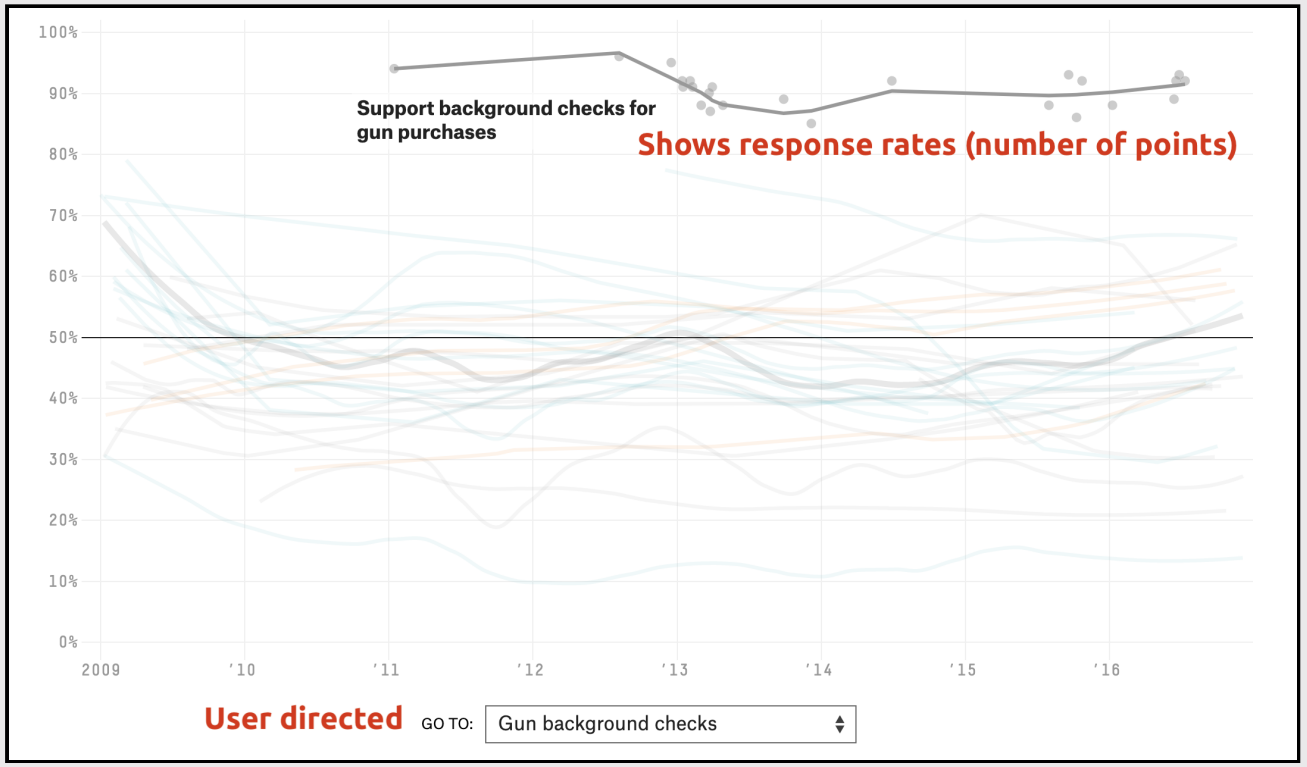
Example 1: FiveThirtyEight's *"How America's Thinking Changed Under Obama: Public opinion on 32 big issues over the past eight years"*



Interactive graphics

Color highlights trend-lines

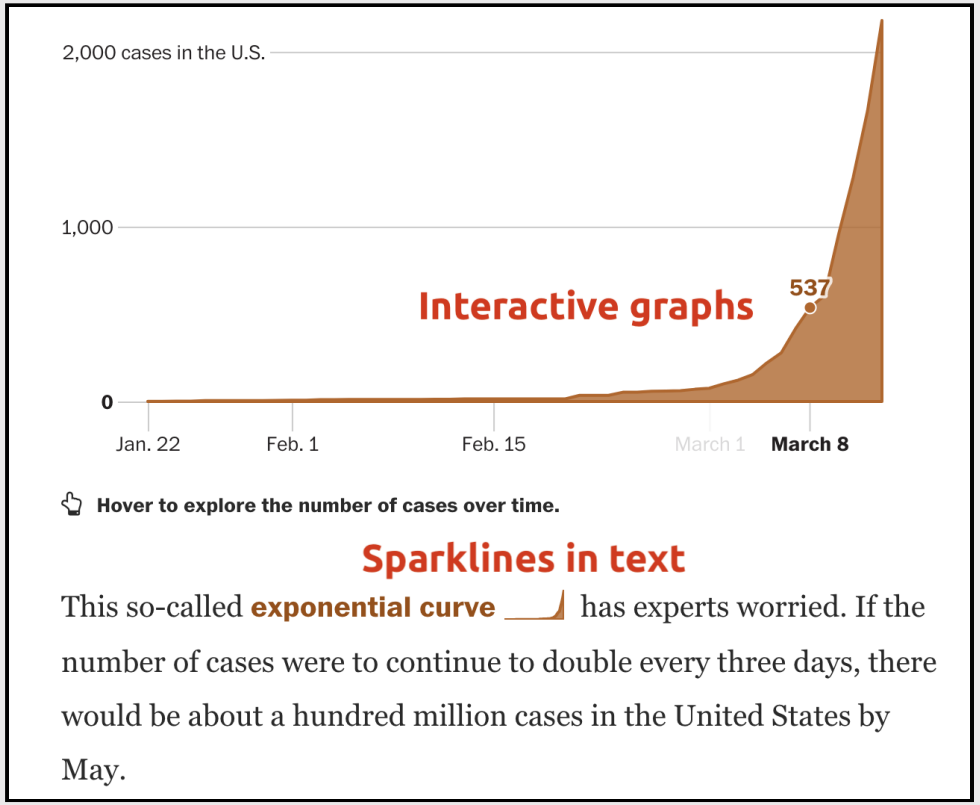
Example 1: FiveThirtyEight's *"How America's Thinking Changed Under Obama: Public opinion on 32 big issues over the past eight years"*



Top plot shows changes over time

Selector on bottom allows user to change graph

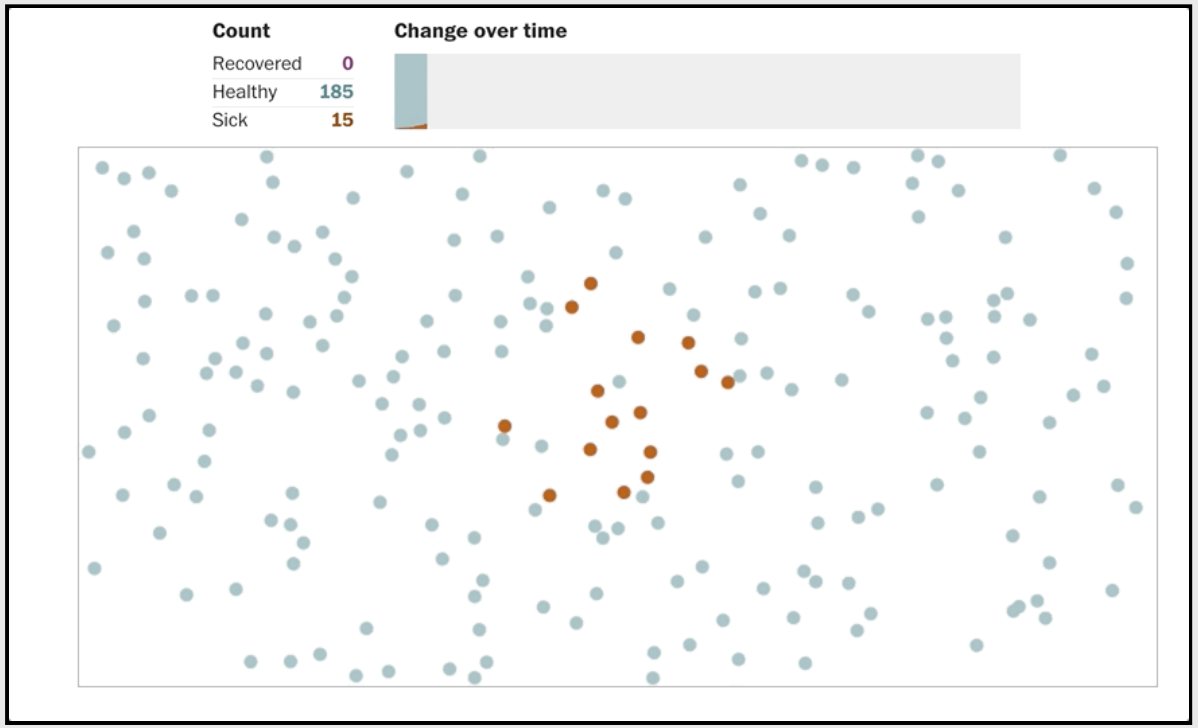
Example 2: Washington Post's *Why outbreaks like coronavirus spread exponentially, and how to "flatten the curve"*



interactive graphs

sparklines in text

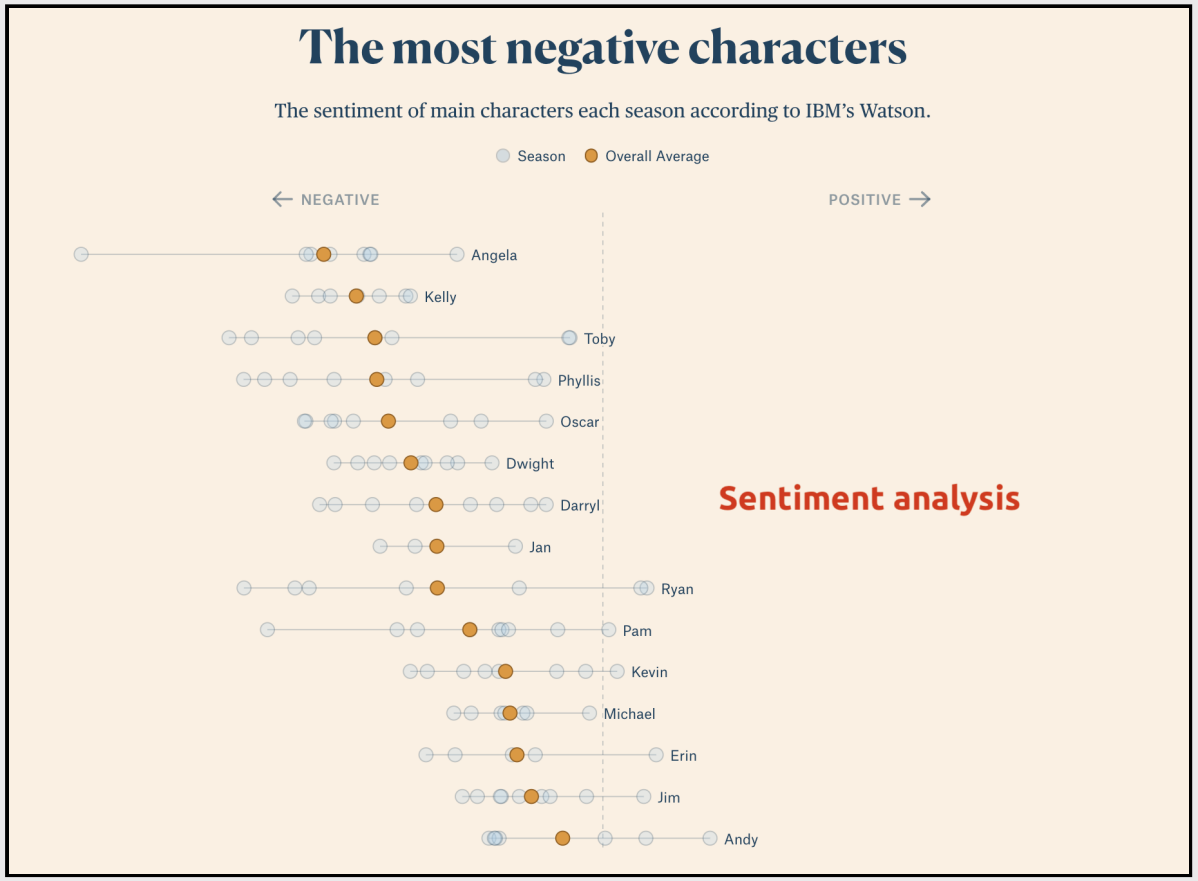
Example 2: Washington Post's *Why outbreaks like coronavirus spread exponentially, and how to "flatten the curve"*

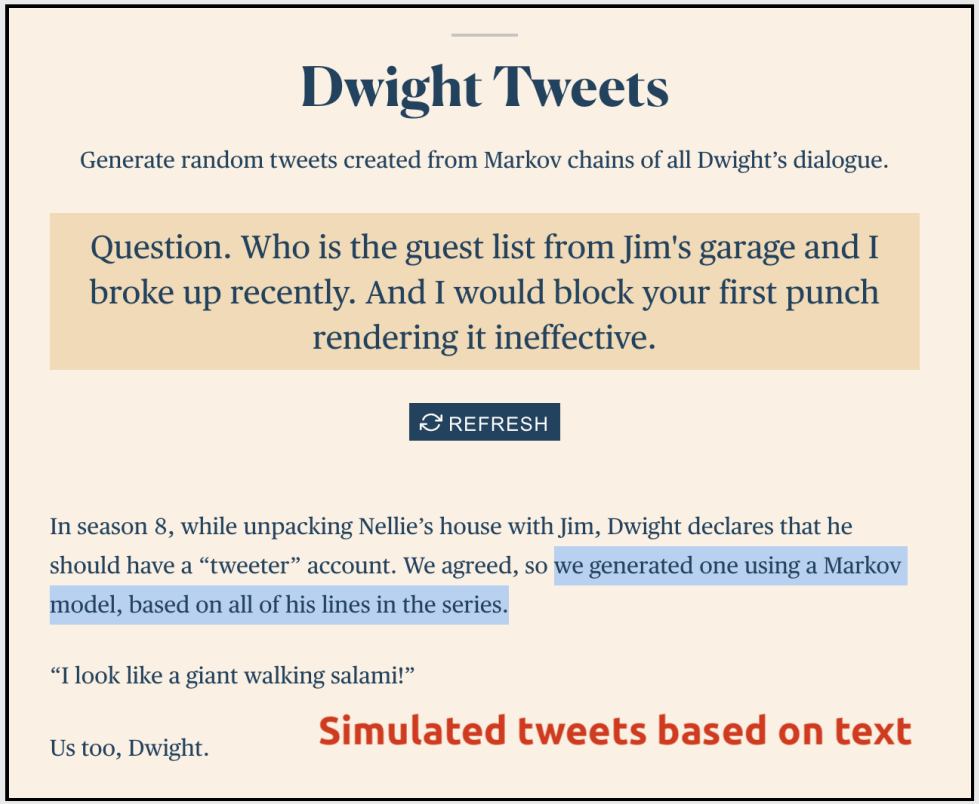


Animated area chart

Scatter-plot changes color in relation to area chart

Example 3: The Pudding's *'The Office' Dialogue in Five Charts: A breakdown of how every character contributed to the show.*





Code for Data Journalists

Code file formats and extensions

We can determine the language of code by it's file extension

file.R = R code file (or script)

file.html = HTML code file (or webpage)

file.css = CSS code file (or stylesheet)

Code file formats and extensions

Text editors (*like Notepad or TextEdit*) can be used to view code files

```
'Below is the plot.R code file in TextEdit'
```

plot.R

```
# load package
library(ggplot2)

# import data
diamonds <- ggplot2::diamonds

# build plot
ggplot(data = diamonds, mapping = aes(x = carat, y = price)) +
  geom_point(aes(color = cut))
```

Code for Data Journalists

R

https://github.com/mjfrigaard/csuc-data-journalism

Why write code?

R is a *language*, which means it gives us the ability to express our ideas with precision

R code is text, so we can use copy + paste and Google (especially for errors!)

```
library(ggplot2)
mtcars %>% ggplot(aes(mpg, disp)) %>%
  geom_point() +
  geom_smooth()
```

```
  Error: `mapping` must be created by
`aes()`
  Did you use %>% instead of +?
```

copy + paste error message and Google it!

R Code - grammar & syntax

A code's *syntax* defines the rules for it's grammar and punctuation

```
add_stuff <- function(a, b) {
    c <- sum(a, b)
    return(c)
}
```

a plus b equals c

The characters and words have specific meanings (just like in English)

```
add_stuff(2, 2)
```

```
[1] 4
```

2 plus 2 equals 4

R Code - grammar & syntax

Characters and words have to be written in a particular order for R code work

```
add_stuff <- function(a, b) {
    c <- sum(a, b)
    return(z)
}
```

```
add_stuff(2, 2)
```

```
  Error in add_stuff(2, 2): object 'z' not found
```

a plus b equals z?

2 plus 2 equals ???

`R Code` - the basics

`objects` are like *nouns*

`functions` are like *verbs*

```
verb(noun)
```

is like...

```
function(object)
```

R Code - A quick example

```
qplot(data = diamonds,
          x = carat,
          y = price,
      color = cut,
       geom = "point")
```

The `qplot()` function **does things** to the objects (*the `carat` and `price` columns from the `diamonds` dataset*)

R Code - A quick example

```
qplot(data = diamonds,
          x = carat,
          y = price,
      color = cut,
       geom = "point")
```

We're telling R we want to use the `diamonds` data to plot the `carat` column on the `x` axis, and the `price` column on the `y`

R Code - A quick example

```
qplot(data = diamonds,
          x = carat,
          y = price,
      color = cut,
       geom = "point")
```

We want the graph to have 'points' (or dots), and we want these points colored by the `cut` column (we use the `"point"` geometric object for points).

R Code - A quick example

R Code | Plot

```
qplot(data = diamonds,
         x = carat, y = price, color = cut, geom = "point")
```

Code for Data Journalists

HTML

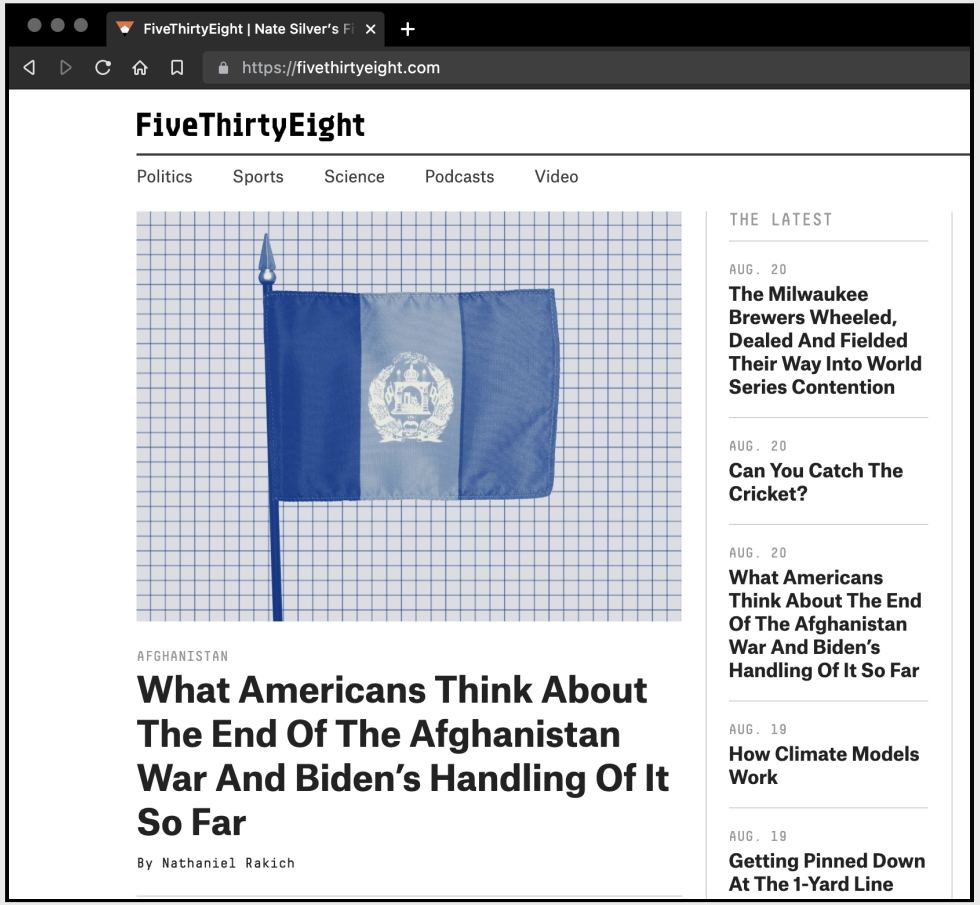
HTML

HTML stands for 'HyperText Markup Language' and is a computer language used to create web pages

HTML code can be run by opening the file containing the code with any web browser (Chrome, Safari, Firefox, etc.)

HTML5 is the current standard

HTML

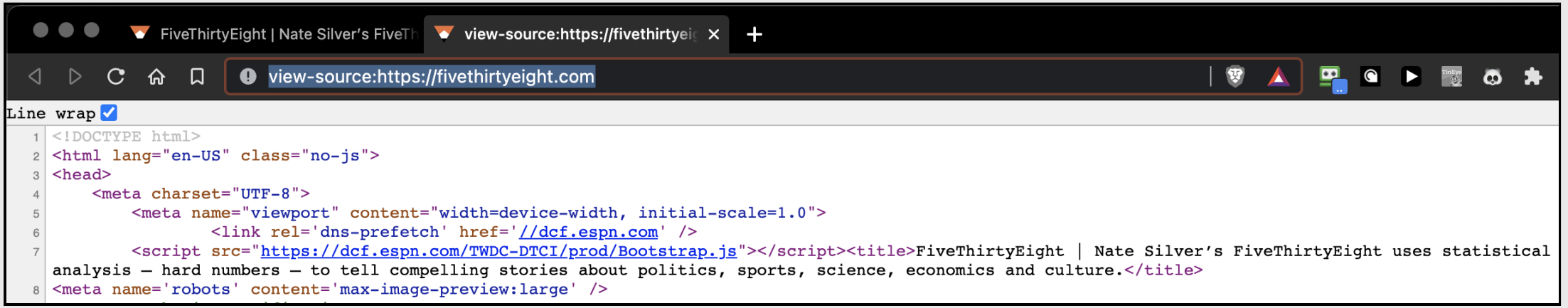


Head over to the fivethirtyeight landing page

Right click on the page and click 'view source'

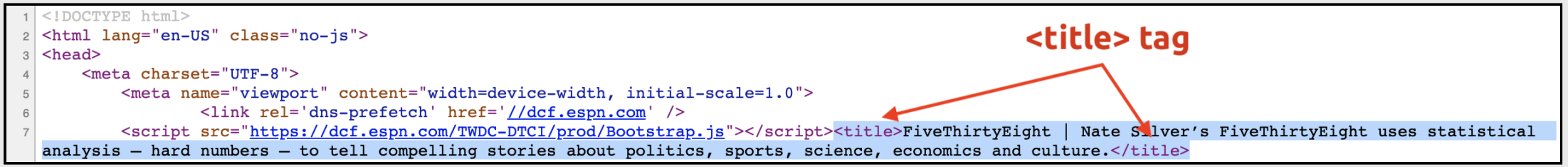
HTML

We're looking at the HTML code used to build the fivethirtyeight website (note the `<!DOCTYPE html>` at the top of the page)



HTML: structure

HTML consists of `elements` and `tags`



Elements have a start tag, followed by the element content, followed by an end tag

HTML - elements and tags

Start tag:

```
<title>
```

Content:

```
<title>FiveThirtyEight | Nate Silver’s FiveThirtyEight uses statisti...
```

End tag:

```
<title>FiveThirtyEight | Nate Silver’s FiveThirtyEight uses...</title>
```

HTML - attributes

Attributes appear in the start tag `<tag>`

Inside tags are `attribute(s)="attribute value(s)"`

End with `</tag>`

```
<img src="image.png">
```

The code above shows the start tag for an `img` element, with an attribute called `src` with a value `"image.png"`

HTML - tags to know

```html
<!-- html comments (not read/displayed by browser) -->
<!DOCTYPE html> <!-- document type declaration -->
<html lang="en-US"> <!-- describes the web page -->
  <head> <!-- header of the HTML document -->
    <title></title> <!-- title of the HTML document -->
  </head>
  <body> <!-- visible page content -->
    <h1> <!-- level 1 header (others include h2-h6) -->
    <!-- href is url, followed by displayed text -->
      <a
      href="https://www.website.com">link
      </a>
      <!-- src the path to an image file -->
        <img src="example.com/example.jpg">
    <h1/>
  </body>
</html>
```

starts with `<!DOCTYPE html>`

most of the code is in the `<body>`

website: `<a href="url>text</a>`

image `<img src="path/to/file.jpg">`

Code for Data Journalists

CSS

CSS

CSS stands for 'Cascading Style Sheets'

CSS is used for describing the layout, colors, and fonts of a HTML document

CSS - structure

```
<style>
    h1 {
        color: blue;
    }
</style>
```

1. `<style>` start tag
2. a selector (`h1`)
3. an open bracket (`{`)
4. property name (`color`)
5. colon (`:`)
6. property value (`blue`)
7. semi-colon (`;`)
8. a closed bracket (`}`)
9. `</style>` end tag

CSS - use

CSS is most useful when included in an external CSS file (i.e., my_style_sheet.css)

```
<link href="my_style_sheet.css" rel="stylesheet"
      type="text/css">
```

We can then reference the my_style_sheet.css style sheet using the <link> tag

Data file formats

Data file types

Data comes in a variety of formats, but this course will focus on 'plain text formats'

Text editors can read and write plain text files

Plain text files are portable across different computer operating systems

CSV - 'comma-separated values' files

The first line is a "header" and contains column (or variable) names in each of the fields (using letters, digits, and underscores)

Each following line represents a new row (or observation)

Any field *may be* quoted, but fields with embedded commas or double-quote characters *must be* quoted

CSV - 'comma-separated values' files

How .csv files look in text editors:

```
name, city, state
Sally, Commack, NY
Fred, Fort Dodge, IA
Deb, Phillipsburg, NJ
```

How .csv files look in a spreadsheet:

name	city	state
Sally	Commack	NY
Fred	Fort Dodge	IA
Deb	Phillipsburg	NJ

XML - Extensible Markup Language

`XML` consists of of XML elements with a start tag and an end tag (with plain text content or other XML elements in-between)

The start tag may include attributes of the form `attribute="value"` (case-sensitive)

All attribute values must be enclosed within double-quotes

XML - structure

```
<?xml version="2.0"?>
<heights>
<filename>ht.txt</filename>
<case date="24-JAN-2019"
      height="78.9"/>
</heights>
```

root element = <heights>
start tag = <filename>
content = ht.txt
end tag = </filename>
element name = case
attribute name = date
attribute value = "24-JAN-2019"
attribute name = height
attribute value = "78.9"

The root element is the `heights` element with `filename` and `case` elements nested within the `heights`

JSON - JavaScript Object Notation

JSON is a lightweight data storage format similar in structure to XML but different syntax/format

Common format for data from application programming interfaces (APIs)

JSON - structure

Data are stored as:

- **Numbers (double)**
- **Strings (double quoted)**
- **Boolean (true or false)**
- **Array (ordered, comma separated enclosed in square brackets `[ ]`)**
- **Object (unorderd, comma-separated collection of key:value pairs in curley brackets `{}`)**

JSON - structure

CSV vs. JSON formats

```
name, city, state
Sally, Commack, NY
Fred, Fort Dodge, IA
Deb, Phillipsburg, NJ
```

```
[
  {
    "name": "Sally",
    "city": "Commack",
    "state": "NY"
  },
  {
    "name": "Fred",
    "city": "Fort Dodge",
    "state": "IA"
  },
  {
    "name": "Deb",
    "city": "Phillipsburg",
    "state": "NJ"
  }
]
```

Recap

Data journalists use programming languages as a tool to process, store, and display data

Code is the preferred technology because it's a language and allows us to be precise and expressive

Plain text data file formats are simple, lowest-common-denominator storage formats

Data in a plain text formats are usually arranged in rows, with several values on each row