# Getting Data

# Considerations for deciding how to get data

- reproducibility of workflow
- frequency with which data is updated
- available formats (may not be identical)
- time to process data

# Web scraping

- Web scraping is a last resort, other methods are generally preferable if available
- Better to find an API, use **httr** package
- Even better, find an R package
  ex. https://cran.r-project.org/web/packages/atus/index.html

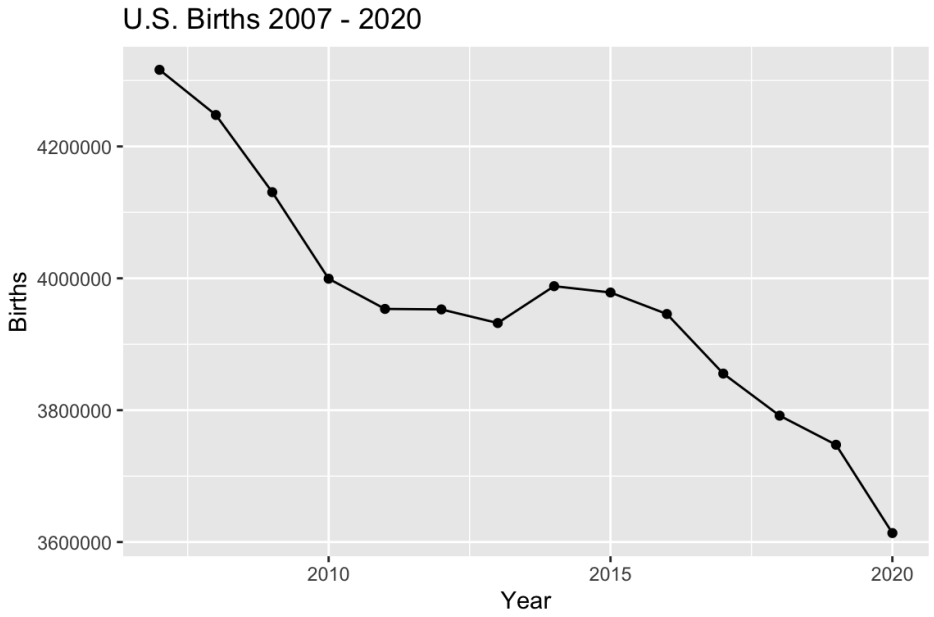
# Case study: CDC birth data

Options:

1. `.txt` file from CDC https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm
2. `.csv` file from NBER https://www.nber.org/research/data/vital-statistics-natality-birth-data (2.46GB unzipped, 200MB zipped)
3. CDC Wonder API web interface https://wonder.cdc.gov/
4. CDC Wonder API https://github.com/socdataR/wonderapi

# CDC birth data API option

```
library(tidyverse)
library(wonderapi)
natdata <- getData(TRUE, "Natality for 2007 - 2020")
ggplot(natdata, aes(Year, Births)) +
  geom_line() +
  geom_point() +
  ggtitle("U.S. Births 2007 - 2020")
```



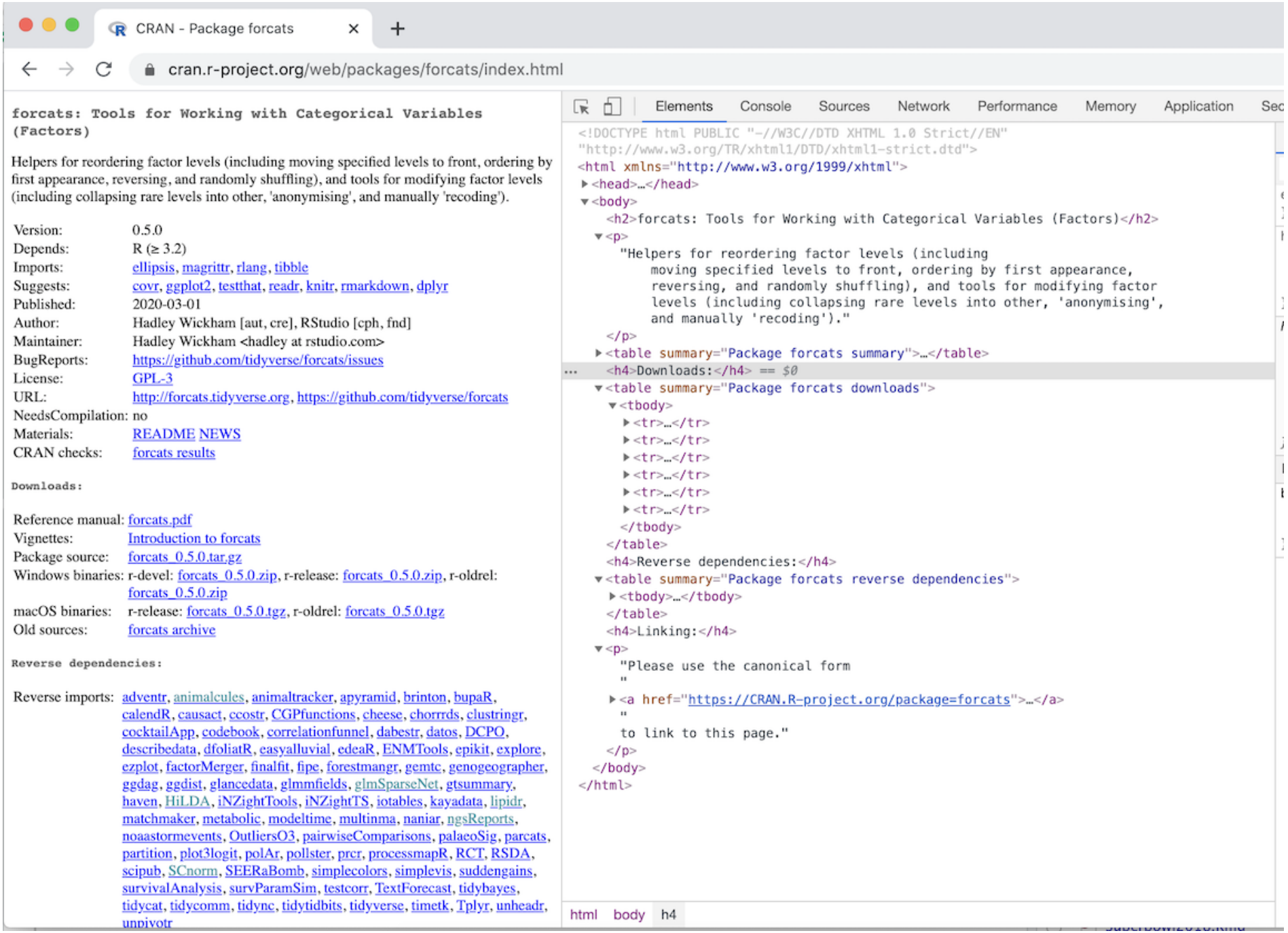
# Web scraping, what not to do

- Scrape all Southwest Airlines data and send consumers notifications if their ticket prices decreased after purchase
- Buy a International Council of Shopping Centers membership, agree to terms of membership, then scrape the entire proprietary membership directory and contact members
- Scrape data that is for sale

# Web scraping, what you should do

- think and investigate legal issues
- think about ethical questions
- limit bandwidth use
- scrape only what you need

# Structure of an HTML page




https://cran.r-project.org/web/packages/forcats/index.html

# rvest package

```
library(tidyverse)
library(rvest)
library(robotstxt)

paths_allowed("https://cran.r-project.org/web/packages/forcats/index.html")
```

```
## [1] TRUE
```

```
paths_allowed("https://cran.r-project.org/web/packages/forcats/DESCRIPTION")
```

```
## [1] FALSE
```

# Tables

```
forcats_data <- read_html("https://cran.r-project.org/web/packages/forcats/index.html") %>%
  html_table()

length(forcats_data)
```

```
## [1] 4
```

```
forcats_data[[1]]
```

| X1 | X2 |
|---|---|
| Version: | 0.5.1 |
| Depends: | R (≥ 3.2) |
| Imports: | ellipsis, magrittr, rlang, tibble |
| Suggests: | covr, dplyr, ggplot2, knitr, readr, rmarkdown, testthat |
| Published: | 2021-01-27 |
| Author: | Hadley Wickham [aut, cre], |
| RStudio [cph, fnd] | |
| Maintainer: | Hadley Wickham <hadley at rstudio.com> |
| BugReports: | https://github.com/tidyverse/forcats/issues |
| License: | MIT + file LICENSE |
| URL: | https://forcats.tidyverse.org, |
| https://github.com/tidyverse/forcats | |

| X1 | X2 |
| --- | --- |
| NeedsCompilation: | no |
| Materials: | README NEWS |
| CRAN checks: | forcats results |

```r
mytable <- forcats_data[[1]]
str(mytable)
```

```
## tibble [13 × 2] (S3: tbl_df/tbl/data.frame)
##  $ X1: chr [1:13] "Version:" "Depends:" "Imports:" "Suggests:" ...
##  $ X2: chr [1:13] "0.5.1" "R (≥ 3.2)" "ellipsis, magrittr, rlang, tibble" "covr, dplyr, ggplot2, knitr, readr,
rmarkdown, testthat" ...
```

```r
version <- mytable %>% filter(X1 == "Version:") %>% pull(X2)
date <- mytable %>% filter(X1 == "Published:") %>% pull(X2)
```

The most recent version of **forcats** on CRAN is 0.5.1, published on 2021-01-27.

(Use inline rmarkdown syntax to include values of variables within text sections.)

# Data not in table form

https://www.beckershospitalreview.com/public-health/states-ranked-by-percentage-of-covid-19-vaccines-administered.html

```
vaccine <- read_html("https://www.beckershospitalreview.com/public-health/states-ranked-by-percentage-of-covid-19-vaccines-administered.html")
```

```
vaccine |> html_node("#inner-article-content")
```

```
## {html_node}
## <div id="inner-article-content">
##  [1] <p>Wisconsin has administered the highest percentage of COVID-19 vaccine ...
##  [2] <script type="text/javascript">doNotShowRelatedArticles = 1;</script>
##  [3] <p>The <a href="https://covid.cdc.gov/covid-data-tracker/#vaccinations"  ...
##  [4] <p>As of 6 a.m. ET Nov. 29, a total of 570,662,725 vaccine doses had bee ...
##  [5] <p>Below are the states and Washington, D.C., ranked by the percentage o ...
##  [6] <p>1. <strong>Wisconsin</strong><br>Doses distributed to state: 9,222,53 ...
##  [7] <p>2. <strong>Connecticut</strong><br>Doses distributed to state: 6,789, ...
##  [8] <p>3. <strong>Massachusetts</strong><br>Doses distributed to state: 13,3 ...
##  [9] <p>4. <strong>New Mexico</strong><br>Doses distributed to state: 3,599,3 ...
## [10] <p>5. <strong>Vermont</strong><br>Doses distributed to state: 1,295,970< ...
## [11] <p>6. <strong>Rhode Island</strong><br>Doses distributed to state: 2,020 ...
## [12] <p>7. <strong>Colorado</strong><br>Doses distributed to state: 10,087,26 ...
## [13] <p>8. <strong>California</strong><br>Doses distributed to state: 70,222, ...
## [14] <p>9. <strong>New York State</strong><br>Doses distributed to state: 35, ...
## [15] <p>10. <strong>Virginia</strong><br>Doses distributed to state: 15,561,3 ...
## [16] <p>11. <strong>Maine</strong><br>Doses distributed to state: 2,642,860<b ...
## [17] <p>12. <strong>Illinois</strong><br>Doses distributed to state: 21,451,2 ...
## [18] <p>13. <strong>Minnesota</strong><br>Doses distributed to state: 9,786,0 ...
## [19] <p>14. <strong>Nevada</strong><br>Doses distributed to state: 4,757,360< ...
## [20] <p>15. <strong>Arizona</strong><br>Doses distributed to state: 11,680,64 ...
## ...
```

# Troubleshooting

- rvest makes it easy to identify nodes and parse text
- but… it doesn't work with all dynamically created content
- workaround: download page as "Webpage, complete" manually
- Or: use RSelenium

# Example

https://analytics.usa.gov/

```
<h2 id="current_visitors" class="data">319,942</h2>
```

`h2` tag

```
html_nodes("h2")
```

`id` attribute

```
html_nodes("#current_visitors")
```

`class` attribute

```
html_nodes(".data")
```

# Examples

```
library(robotstxt)
paths_allowed("https://analytics.usa.gov/")
```

```
## [1] TRUE
```

```
webdata <- read_html("https://analytics.usa.gov/")
webdata %>% html_nodes("h2")
```

```
## {xml_nodeset (1)}
## [1] <h2 id="current_visitors" class="data">...</h2>
```

```
webdata %>% html_nodes("#current_visitors")
```

```
## {xml_nodeset (1)}
## [1] <h2 id="current_visitors" class="data">...</h2>
```

```
webdata %>% html_nodes(".data")
```

```
## {xml_nodeset (16)}
##  [1] <h2 id="current_visitors" class="data">...</h2>
##  [2] <svg class="data time-series"></svg>
##  [3] <span id="total_visitors" class="data">...</span>
##  [4] <div class="data bar-chart">\n            </div>
##  [5] <div class="data bar-chart">\n            </div>
##  [6] <div class="data bar-chart">\n            </div>
##  [7] <div class="data bar-chart">\n            </div>
##  [8] <div class="data bar-chart">\n            </div>
##  [9] <div class="data bar-chart">\n            </div>
## [10] <div class="data bar-chart">\n            </div>
## [11] <div class="data bar-chart">\n            </div>
## [12] <div class="data bar-chart">\n            </div>
## [13] <div class="data bar-chart">\n            </div>
## [14] <div class="data bar-chart">\n            </div>
## [15] <div class="data bar-chart">\n            </div>
## [16] <div class="data bar-chart">\n            </div>
```

```
webdata %>% html_nodes("h2") %>% html_text()
```

```
## [1] "..."
```

Where's the number?

```
webdata_dl <- read_html("analytics.html")
webdata_dl %>% html_nodes("h2") %>% html_text()
```

```
## [1] "437,403"
```

```
webdata_dl %>% html_nodes(".data")
```

```
## {xml_nodeset (16)}
##  [1] <h2 id="current_visitors" class="data">437,403</h2>
##  [2] <svg class="data time-series" viewbox="0 0 700 150"><g class="axis y0" t ...
##  [3] <span id="total_visitors" class="data">5.60 billion</span>
##  [4] <div class="data bar-chart">\n            <div class="bin">\n<div class= ...
##  [5] <div class="data bar-chart">\n            <div class="bin">\n<div class= ...
##  [6] <div class="data bar-chart">\n            <div class="bin">\n<div class= ...
##  [7] <div class="data bar-chart">\n            <div class="bin">\n<div class= ...
##  [8] <div class="data bar-chart">\n            <div class="bin" data-share="2 ...
##  [9] <div class="data bar-chart">\n            <div class="bin">\n<div class= ...
## [10] <div class="data bar-chart">\n            <div class="bin">\n<div class= ...
## [11] <div class="data bar-chart">\n            <div class="bin" data-share="8 ...
## [12] <div class="data bar-chart">\n            <div class="bin" data-share="1 ...
## [13] <div class="data bar-chart">\n            <div class="bin">\n<div class= ...
## [14] <div class="data bar-chart">\n            <div class="bin">\n<div class= ...
## [15] <div class="data bar-chart">\n            <div class="bin">\n<div class= ...
## [16] <div class="data bar-chart">\n            <div class="bin">\n<div class= ...
```