

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-nativity-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>

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

CRAN - Package forcats

← → ↺

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

forcats: Tools for Working with Categorical Variables (Factors)

Helpers for reordering factor levels (including moving specified levels to front, ordering by first appearance, reversing, and randomly shuffling), and tools for modifying factor levels (including collapsing rare levels into other, 'anonymising', and manually 'recoding').

Version: 0.5.0

Depends: R (≥ 3.2)

Imports: [ellipsis](#), [magrittr](#), [rlang](#), [tibble](#)

Suggests: [covr](#), [ggplot2](#), [testthat](#), [readr](#), [knitr](#), [rmarkdown](#), [dplyr](#)

Published: 2020-03-01

Author: Hadley Wickham [aut, cre], RStudio [cph, fnd]

Maintainer: Hadley Wickham <hadley at rstudio.com>

BugReports: <https://github.com/tidyverse/forcats/issues>

License: [GPL-3](#)

URL: <http://forcats.tidyverse.org>, <https://github.com/tidyverse/forcats>

NeedsCompilation: no

Materials: [README NEWS](#)

CRAN checks: [forcats results](#)

Downloads:

Reference manual: [forcats.pdf](#)

Vignettes: [Introduction to forcats](#)

Package source: [forcats_0.5.0.tar.gz](#)

Windows binaries: r-devel: [forcats_0.5.0.zip](#), r-release: [forcats_0.5.0.zip](#), r-oldrel: [forcats_0.5.0.zip](#)

macOS binaries: r-release: [forcats_0.5.0.tgz](#), r-oldrel: [forcats_0.5.0.tgz](#)

Old sources: [forcats archive](#)

Reverse dependencies:

Reverse imports: [adventr](#), [animalcules](#), [animaltracker](#), [apyramid](#), [brinton](#), [bupaR](#), [calendr](#), [causact](#), [ccostr](#), [CGPfunctions](#), [cheese](#), [chorrds](#), [clustrngr](#), [cocktailApp](#), [codebook](#), [correlationfunnel](#), [dabestr](#), [datos](#), [DCPO](#), [describedata](#), [dfoliatR](#), [easyalluvial](#), [edeaR](#), [ENMTools](#), [epikit](#), [explore](#), [ezplot](#), [factorMerger](#), [finalfit](#), [fipe](#), [forestmangr](#), [gemtc](#), [genogeographer](#), [ggdag](#), [ggdist](#), [glancedata](#), [glmmfields](#), [glmSparseNet](#), [gtsummary](#), [haven](#), [HiLDA](#), [iNZightTools](#), [iNZightTS](#), [iotables](#), [kayadata](#), [lipidr](#), [matchmaker](#), [metabolic](#), [modeltime](#), [multinma](#), [naniar](#), [ngsReports](#), [noaastormevents](#), [OutliersO3](#), [pairwiseComparisons](#), [palaeoSig](#), [parcats](#), [partition](#), [plot3logit](#), [polAr](#), [pollster](#), [prcr](#), [processmapR](#), [RCT](#), [RSDA](#), [scipub](#), [SCnorm](#), [SEERaBomb](#), [simplecolors](#), [simplevis](#), [suddengains](#), [survivalAnalysis](#), [survParamSim](#), [testcorr](#), [TextForecast](#), [tidybayes](#), [tidycat](#), [tidycomm](#), [tidync](#), [tidytidbits](#), [tidyverse](#), [timetk](#), [Tplyr](#), [unheadr](#), [unpivotr](#)

Elements

Console

Sources

Network

Performance

Memory

Application

Sec

```

<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>...</head>
  <body>
    <h2>forcats: Tools for Working with Categorical Variables (Factors)</h2>
    <p>
      "Helpers for reordering factor levels (including
      moving specified levels to front, ordering by first appearance,
      reversing, and randomly shuffling), and tools for modifying factor
      levels (including collapsing rare levels into other, 'anonymising',
      and manually 'recoding')."
    </p>
    <table summary="Package forcats summary">...</table>
    ...
    <h4>Downloads:</h4> == $0
    <table summary="Package forcats downloads">
      <tbody>
        <tr>...</tr>
        <tr>...</tr>
        <tr>...</tr>
        <tr>...</tr>
        <tr>...</tr>
        <tr>...</tr>
      </tbody>
    </table>
    <h4>Reverse dependencies:</h4>
    <table summary="Package forcats reverse dependencies">
      <tbody>...</tbody>
    </table>
    <h4>Linking:</h4>
    <p>
      "Please use the canonical form
      "
      <a href="https://CRAN.R-project.org/package=forcats">...</a>
      "
      to link to this page."
    </p>
  </body>
</html>

```

<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

Version:

Depends:

Imports:

Suggests:

3.0.0)

Published:

Author:

RStudio [cph, fnd]

Maintainer:

X2

0.5.2

R (≥ 3.4)

cli, ellipsis, glue, lifecycle, magrittr, rlang ($\geq 1.0.0$), tibble, withr

covr, dplyr, ggplot2, knitr, readr, rmarkdown, testthat (\geq

2022-08-19

Hadley Wickham [aut, cre],

Hadley Wickham <hadley at rstudio.com>

X1

BugReports:

License:

URL:

<https://github.com/tidyverse/forcats>

NeedsCompilation:

Materials:

CRAN checks:

X2

<https://github.com/tidyverse/forcats/issues>

MIT + file LICENSE

<https://forcats.tidyverse.org/>,

no

README NEWS

forcats results

```
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] "0.5.2" "R (≥ 3.4)" "cli, ellipsis, glue, lifecycle, magrittr, rlang (≥ 1.0.0),  
tibble, withr" "covr, dplyr, ggplot2, knitr, readr, rmarkdown, testthat (≥\n3.0.0)" ...
```

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

The most recent version of **forcats** on CRAN is 0.5.2, published on 2022-08-19.

(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_element("#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_elements("h2")
```

id attribute

```
html_elements("#current_visitors")
```

class attribute

```
html_elements(".data")
```

Examples

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

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

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

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

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

```
## {html_node}
## <h2 id="current_visitors" class="data">
```

```
webdata %>% html_elements(".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_elements("h2") %>% html_text()
```

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

Where's the number?

```
webdata_d1 <- read_html("analytics.html")
webdata_d1 %>% html_elements("h2") %>% html_text()
```

```
## [1] "414,029"
```

```
webdata_d1 %>% html_elements(".data")
```

```
## {xml_nodeset (16)}
## [1] <h2 id="current_visitors" class="data">414,029</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= ...
```