Web Scraping with `rvest`

Stat 108, Week 12

```
#Load web scraping library
library(rvest)
```

Understanding HTML Tables

- Let's scrape the tables in "pdxTreesHTMLTable.html".
  - It is in both the shared folder on the Server and the `materials` repo on GitHub.
- Now let's scrape the tables with `rvest`.

```
pdx_tables <- read_html("https://raw.githubusercontent.com/harvard-stat108s23/materials/main/handouts/p
  html_nodes("table") %>%
  html_table()
```

- Examine the structure.

```
pdx_tables
class(pdx_tables)
class(pdx_tables[2])
class(pdx_tables[[2]])

TREES <- pdx_tables[[2]]
TREES
```

Grab Tables from the Web

Let's grab the **Current teams** table on Boston sport teams' Wikipedia page.

```
#Store url
url <- "https://en.wikipedia.org/wiki/Sports_in_Boston"

## Scrape html and store table

#Option 1: Grab all the tables and then navigate to the one you wanted.
tables <- url %>%
  read_html() %>%
  html_nodes(css = "table")

#Grab the specific table
current_teams <- html_table(tables[[1]], fill = TRUE)
current_teams
```

```
## # A tibble: 5 x 6
##   Club                League Sport   `Venue (capacity)` Founded Championships
##   <chr>               <chr>  <chr>   <chr>                <int> <chr>
## 1 Boston Red Sox      MLB    Baseba~ Fenway Park (37,5~    1901 9 World Seri~
## 2 Boston Bruins       NHL    Ice Ho~ TD Garden (17,565)    1924 6 Stanley Cu~
```

```
## 3 Boston Celtics          NBA    Basket~ TD Garden (17,565)    1946 17 NBA titles
## 4 New England Patriots    NFL    Footba~ Gillette Stadium ~    1960 6 Super Bowls
## 5 New England Revolution MLS     Soccer  Gillette Stadium ~    1995 0 MLS Cups; ~
```

```
#Option 2: Use the specific css
#NOTICE: nodes is now node!
current_teams2 <- url %>%
  read_html() %>%
  html_node(css = "table.wikitable:nth-child(15)") %>%
  html_table()
```

Another Example

Let's scrape the NYTimes.com's College Access Index table.

```
# Store url
url <- "https://www.nytimes.com/interactive/2017/05/25/sunday-review/opinion-pell-table.html"

## Scrape html and store table

# Grab the table
tables <- url %>%
  read_html() %>%
  html_nodes(css = "table")

#Grab the specific table
college_access_table <- html_table(tables[[1]], fill = TRUE)

#Option 2: Use the specific css
college_access_table2 <- url %>%
  read_html() %>%
  html_node(css = ".table") %>%
  html_table()
```

Easier route: Let's see how to use `datapasta` to grab the data for this example!