Web Scraping with `rvest`

Math 241, Week 5

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

Grab Tables from the Web

Let's grab the **Portland area sports teams** table on Portland's Wikipedia page.

```
#Store url
url <- "https://en.wikipedia.org/wiki/Portland,_Oregon"

## Scrape html and store table

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

#Grab the specific table
champ_table <- html_table(tables[[8]], fill = TRUE)
champ_table
```

```
## # A tibble: 6 x 6
##   Club                  Sport      'Current League' Championships Venue Founded
##   <chr>                 <chr>      <chr>            <chr>         <chr>   <int>
## 1 Portland Trail Blazers Basketball NBA              1 (1977)      Moda~    1970
## 2 Portland Winterhawks  Hockey     WHL              3 (1981-82, ~ Vete~    1976
## 3 Portland Timbers      Soccer     MLS              1 (2015)      Prov~    2009
## 4 Portland Thorns FC    Soccer     NWSL             3 (2013, 201~ Prov~    2012
## 5 Hillsboro Hops        Baseball   Northwest League 3 (2014, 201~ Ron ~    2013
## 6 Portland Timbers 2    Soccer     MLS Next Pro     0             Hill~    2014
```

Population data over time

Using the same approach, grab the table of Portland's population over time (table 4), and create a simple visualization of the total population count over time.

Another Example

- Although we saw that we can use `datapasta` to grab these data, let's scrape the NYTimes.com's College Access Index table.

```r
# 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 = "#opinion-pell-table > div > div.g-item.g-sortable-table > table") %>%
  html_table()
```

Create a visualization

Choose 3 variables and create a heuristic visualization using this dataset.