STOR 320 Web Scraping

Lecture 16

Yao Li
Department of Statistics and Operations Research
UNC Chapel Hill

Intermission

- Final 3 Data Frames From Last Tutorial Should All Be Saved to CSV's on PC
  - FINAL_VIOLENT.CSV
  - FINAL_ZIP.CSV
  - FINAL_STATE_ABBREV.CSV
- Think About What Other City Information Could Potentially Be a Factor in Violent Crimes
- Think About What Other City Information Could Potentially Be Influenced by the Prevalence of Violent Crimes

Tutorial 8 Introduction

- Step 1: Open Tutorial 8
- Step 2: Ensure You Have the Following R Packages Installed
  - tidyverse
  - rvest
- Step 3: Switch Knitter
- Step 4: Read the Introduction

Part 1: Connection to Population Change and Density

- Step 1: Select the Link and Observe the Following Table

2019 rank	City	State[c]	2019 estimate	2010 Census	Change	2016 land area		2016 population density		Location
1	*New York*[d]	New York	8,336,817	8,175,133	+1.98%	301.5 sq mi	780.9 km^2	28,317/sq mi	10,933/km^2	40.6635°N 73.9387°W
2	*Los Angeles*	California	3,979,576	3,792,621	+4.93%	468.7 sq mi	1,213.9 km^2	8,484/sq mi	3,276/km^2	34.0194°N 118.4108°W
3	*Chicago*	Illinois	2,693,976	2,695,598	−0.06%	227.3 sq mi	588.7 km^2	11,900/sq mi	4,600/km^2	41.8376°N 87.6818°W
4	*Houston*[3]	Texas	2,320,268	2,100,263	+10.48%	637.5 sq mi	1,651.1 km^2	3,613/sq mi	1,395/km^2	29.7866°N 95.3909°W
5	***Phoenix***	Arizona	1,680,992	1,445,632	+16.28%	517.6 sq mi	1,340.6 km^2	3,120/sq mi	1,200/km^2	33.5722°N 112.0901°W
6	*Philadelphia*[e]	Pennsylvania	1,584,064	1,526,006	+3.80%	134.2 sq mi	347.6 km^2	11,683/sq mi	4,511/km^2	40.0094°N 75.1333°W
7	San Antonio	Texas	1,547,253	1,327,407	+16.56%	461.0 sq mi	1,194.0 km^2	3,238/sq mi	1,250/km^2	29.4724°N 98.5251°W
8	San Diego	California	1,423,851	1,307,402	+8.91%	325.2 sq mi	842.3 km^2	4,325/sq mi	1,670/km^2	32.8153°N 117.1350°W
9	Dallas	Texas	1,343,573	1,197,816	+12.17%	340.9 sq mi	882.9 km^2	3,866/sq mi	1,493/km^2	32.7933°N 96.7665°W

- Step 2: Questions?
  - What is the Connection to Violent Crimes?
  - How is this Useful When Related to Violent Crimes?

Part 1: Connection to Population Change and Density

- Step 3: Run Chunk 1
  - What is required to convert the Pop_2019 to a numeric variable?
  - What is required to convert the Land to a numeric variable?
  - What is required to convert the Density to a numeric variable?
- Step 4: Run Chunk 2
  - Notice: “,|km2”,”,|/km2”

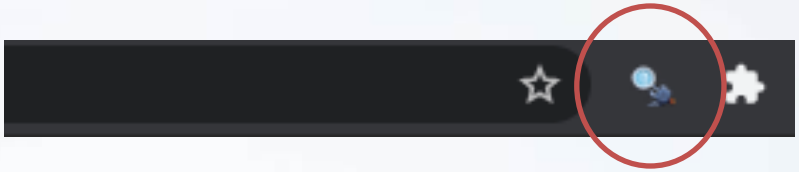
Part 1: Connection to Population Change and Density

- Step 5: Run Chunk 3
  - How to create a variable representing population change from 2016 to 2019?
  - How to create a variable representing population density in 2019?
  - How to clean the city name column?

Part 2: Inclusion of Expert Opinion

- Step 1: Selector Gadget Website

  - Open Source
  - Chrome Extension Exists
  - Easy: Drag Link to Bookmark Bar as Webpage Explains



- Step 2: Observe the Article on 2018's Safest and Most Dangerous States

  - What info could be of use?
  - Do you agree identification?

Part 2: Inclusion of Expert Opinion

- Step 3: Information of Interest
  - Safe vs Dangerous

1. Vermont	1. Mississippi
2. Maine	2. Louisiana
3. Minnesota	3. Oklahoma
4. Utah	4. Texas
5. New Hampshire	5. Florida
6. Connecticut	6. Arkansas
7. Rhode Island	7. Alabama
8. Hawaii	8. Missouri
9. Massachusetts	9. Alaska
10. Washington	10. South Carolina

  - Goal: Scrape this Information into Vectors in R to Create a Table

Part 2: Inclusion of Expert Opinion

- Step 4: Identifying CSS Selector
  - Go to Web Page

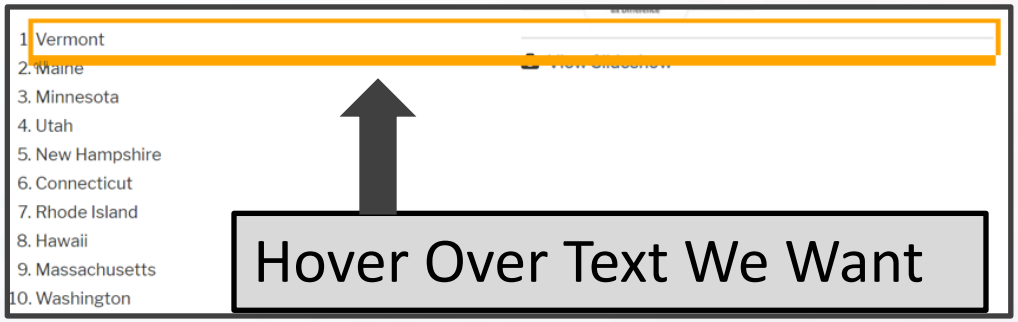
https://www.securitysales.com/fire-intrusion/2018-safest-most-dangerous-states-us/

  - Choose SelectorGadget in Bookmark Tab

  - Locate This Box

Part 2: Inclusion of Expert Opinion

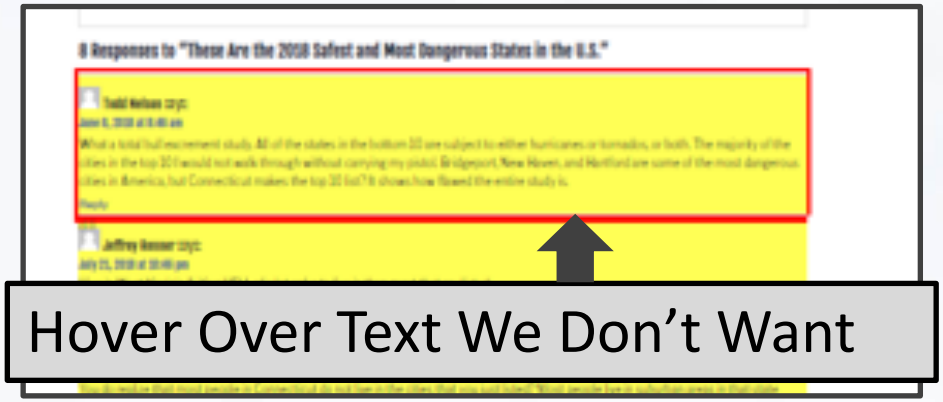
- Step 4: Continued
  - Find Content You Want



  - Point and Click to Select Info
  - Info We Want is Highlighted
  - Info We Don't Want, As Well

Part 2: Inclusion of Expert Opinion

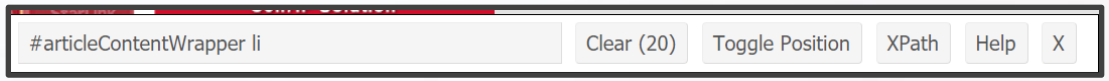
- Step 4: Continued
  - Find Content You Don't Want



  - Point and Click to Deselect
  - Locate This Box

Part 2: Inclusion of Expert Opinion

- Step 4: Continued
  - Locate This Box



  - Copy CSS Selector: “#articleContentWrapper li”

- Step 5: Run Chunk 1

```
SAFE_VS_DANGEROUS = URL.SAFE_VS_DANGEROUS %>%
                    read_html() %>%
                    html_nodes(css="#articleContentWrapper li") %>%
                    html_text()
```

- Step 6: Run Chunk 2
  - What About the Other States?
- Step 7: Walk-off Knit