



# *Web Scraping II*

## Recap of Web Scraping I



- 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 (Requires Internet)
- 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

| Rank | Name         | State        | 2019<br>Population<br>▼ | 2016<br>Population | 2010<br>Census | Change | 2019<br>Density ≡      |
|------|--------------|--------------|-------------------------|--------------------|----------------|--------|------------------------|
| 1    | New York     | New York     | 8,601,186               | 8,537,673          | 8,175,133      | 0.25%  | 11,056/km <sup>2</sup> |
| 2    | Los Angeles  | California   | 4,057,841               | 3,976,322          | 3,792,621      | 0.67%  | 3,343/km <sup>2</sup>  |
| 3    | Chicago      | Illinois     | 2,679,044               | 2,704,958          | 2,695,598      | -0.32% | 4,550/km <sup>2</sup>  |
| 4    | Houston      | Texas        | 2,359,480               | 2,303,482          | 2,099,451      | 0.80%  | 1,431/km <sup>2</sup>  |
| 5    | Phoenix      | Arizona      | 1,711,356               | 1,615,017          | 1,445,632      | 1.91%  | 1,276/km <sup>2</sup>  |
| 6    | Philadelphia | Pennsylvania | 1,576,596               | 1,567,872          | 1,526,006      | 0.18%  | 4,537/km <sup>2</sup>  |

- 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

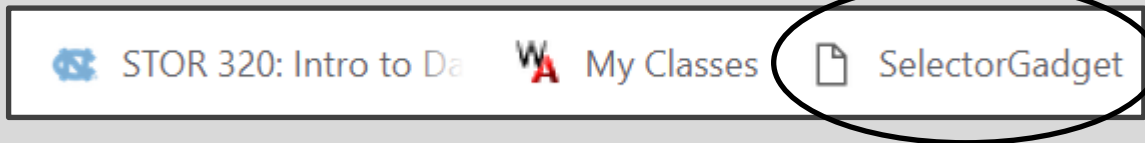


- Step 3: Run Chunk 1
  - What is required to convert the Percentage Change to a numeric variable?
  - What is required to convert the 2019 Density to a numeric variable?
- Step 4: Run Chunk 2
  - Notice: `/.*`
- Step 5: No-Knitter

## 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       |
| 2. Maine         |
| 3. Minnesota     |
| 4. Utah          |
| 5. New Hampshire |
| 6. Connecticut   |
| 7. Rhode Island  |
| 8. Hawaii        |
| 9. Massachusetts |
| 10. Washington   |

|                    |
|--------------------|
| 1. Mississippi     |
| 2. Louisiana       |
| 3. Oklahoma        |
| 4. Texas           |
| 5. Florida         |
| 6. Arkansas        |
| 7. Alabama         |
| 8. Missouri        |
| 9. Alaska          |
| 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

STOR 320: Intro to Da  My Classes  SelectorGadget

- Locate This Box


No valid path found.



## Part 2: Inclusion of Expert Opinion



- Step 4: Continued
  - Find Content You Want



US STATES

1. Vermont
2. Maine
3. Minnesota
4. Utah
5. New Hampshire
6. Connecticut
7. Rhode Island
8. Hawaii
9. Massachusetts
10. Washington

Hover Over Text We Want

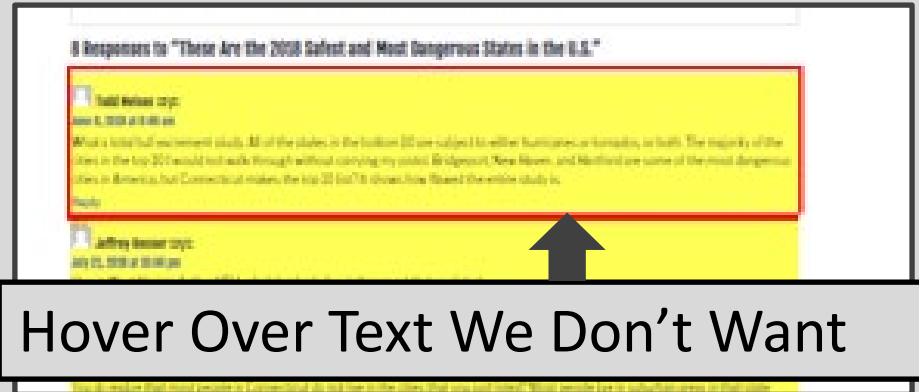
- Point and Click to Select Info
- Info We Want is Highlighted
- Info We Don't Want, As Well

[illegible]

## Part 2: Inclusion of Expert Opinion



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



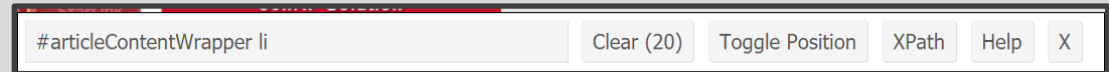
- Point and Click to Deselect
- Locate This Box



## 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

Closing



Disperse  
and Make  
Reasonable  
Decisions