STOR 320 Web Scraping

Lecture 9

Yao Li

Department of Statistics and Operations Research

UNC Chapel Hill

Motivation

- Relying on Downloadable CSV's Puts You at a Disadvantage
- Majority of Data Is Found Online
- Negative: Online Data is Unstructured in HTML Format
- Positive: Online Data is Often Updated, Relevant, & Untapped

Motivation

- Example 1: Population of states

Population of states, territories, divisions and regions

State/federal district/territory/ division/region	Rank	2019 population	Rank	2010 population	Rank	2000 population	Rank	2000-2010 change	Geographic sort
Massachusetts	15	6,892,503	14	6,547,629	13	6,349,097	43	3.1%	NEng
Connecticut	29	3,565,287	29	3,574,097	29	3,405,565	35	4.9%	NEng
New Hampshire	41	1,359,711	42	1,316,470	41	1,235,786	32	6.5%	NEng
Maine	42	1,344,212	41	1,328,361	40	1,274,923	39	4.2%	NEng
Rhode Island	44	1,059,361	43	1,052,567	43	1,048,319	49	0.4%	NEng
Vermont	49	623,989	49	625,741	49	608,827	44	2.8%	NEng
New England	**9**	**14,845,063**	**9**	**14,444,865**	**9**	**13,922,517**	**7**	**3.8%**	**NEast**
New York	4	19,453,561	3	19,378,102	3	18,976,457	46	2.1%	MAtl
Pennsylvania	5	12,801,989	6	12,702,379	6	12,281,054	41	3.4%	MAtl
New Jersey	11	8,882,190	11	8,791,894	9	8,414,350	37	4.5%	MAtl
Mid-Atlantic	**4**	**41,137,740**	**4**	**40,872,375**	**4**	**39,671,861**	**8**	**3.0%**	**NEast**
Northeast	**4**	**55,982,803**	**4**	**55,317,240**	**4**	**53,594,378**	**4**	**3.2%**	**USA**
Florida	3	21,477,737	4	18,801,310	4	15,982,378	8	17.6%	SAtl
Georgia	8	10,617,423	9	9,687,653	10	8,186,453	7	18.3%	SAtl
North Carolina	9	10,488,084	10	9,535,483	11	8,049,313	6	18.5%	SAtl
Virginia	12	8,535,519	12	8,001,024	12	7,078,515	16	13.0%	SAtl
Maryland	19	6,045,680	19	5,773,552	19	5,296,486	23	9.0%	SAtl
South Carolina	23	5,148,714	24	4,625,364	26	4,012,012	10	15.3%	SAtl
West Virginia	38	1,792,147	37	1,852,994	37	1,808,344	45	2.5%	SAtl

Motivation

- Example 2: Blood Pressure Chart

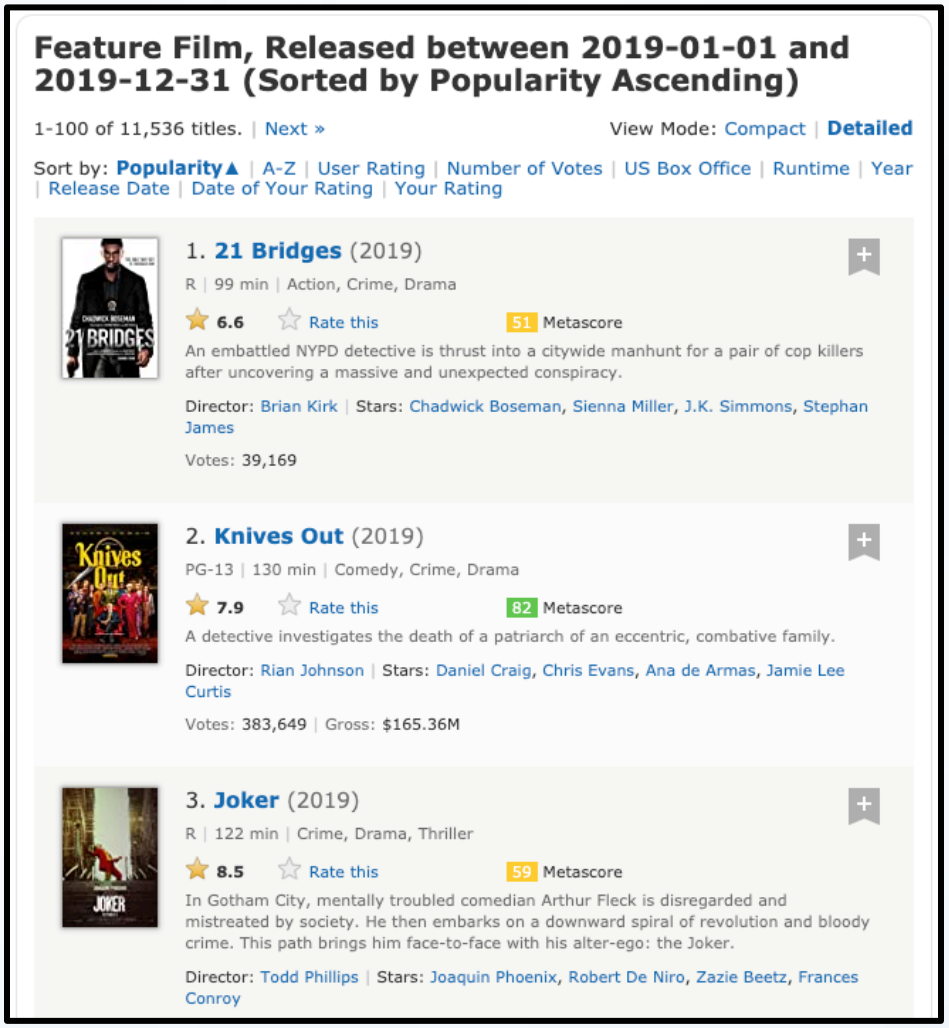
What Should Blood Pressure be According to Age?

Approx. BP According to Age Chart

Age	Low		Normal		Elevated		Stage 1 Hypertension		Stage 2 Hypertension	
	S	D	S	D	S	D	S	D	S	D
17-19	< 90	< 60	< 120	< 80	120-129	< 80	130-139	80-89	140+	90+
20-24	< 90	< 60	< 120	< 80	120-129	< 80	130-139	80-89	140+	90+
25-29	< 90	< 60	< 120	< 80	120-129	< 80	130-139	80-89	140+	90+
30-34	< 90	< 60	< 120	< 80	120-129	< 80	130-139	80-89	140+	90+
35-39	< 90	< 60	< 120	< 80	120-129	< 80	130-139	80-89	140+	90+
40-44	< 90	< 60	< 120	< 80	120-129	< 80	130-139	80-89	140+	90+
45-49	< 90	< 60	< 120	< 80	120-129	< 80	130-139	80-89	140+	90+
50-54	< 90	< 60	< 120	< 80	120-129	< 80	130-139	80-89	140+	90+
55-59	< 90	< 60	< 120	< 80	120-129	< 80	130-139	80-89	140+	90+
60+	< 90	< 60	120	< 80	120-129	< 80	130-139	80-89	140+	90+

Motivation

- Example 3: Movie List



Feature Film, Released between 2019-01-01 and 2019-12-31 (Sorted by Popularity Ascending)

1-100 of 11,536 titles. | Next » View Mode: Compact | **Detailed**

Sort by: **Popularity▲** | A-Z | User Rating | Number of Votes | US Box Office | Runtime | Year | Release Date | Date of Your Rating | Your Rating

1. **21 Bridges** (2019)

R | 99 min | Action, Crime, Drama

6.6 Rate this 51 Metascore

An embattled NYPD detective is thrust into a citywide manhunt for a pair of cop killers after uncovering a massive and unexpected conspiracy.

Director: Brian Kirk | Stars: Chadwick Boseman, Sienna Miller, J.K. Simmons, Stephan James

Votes: 39,169

2. **Knives Out** (2019)

PG-13 | 130 min | Comedy, Crime, Drama

7.9 Rate this 82 Metascore

A detective investigates the death of a patriarch of an eccentric, combative family.

Director: Rian Johnson | Stars: Daniel Craig, Chris Evans, Ana de Armas, Jamie Lee Curtis

Votes: 383,649 | Gross: $165.36M

3. **Joker** (2019)

R | 122 min | Crime, Drama, Thriller

8.5 Rate this 59 Metascore

In Gotham City, mentally troubled comedian Arthur Fleck is disregarded and mistreated by society. He then embarks on a downward spiral of revolution and bloody crime. This path brings him face-to-face with his alter-ego: the Joker.

Director: Todd Phillips | Stars: Joaquin Phoenix, Robert De Niro, Zazie Beetz, Frances Conroy

Web Scraping Definition

- Process of Converting Currently Unstructured Data on Web to Structured Data in R

- Ideas:
  - Population Table to CSV
  - Blood Pressure Chart to Tibble
  - Movies to List in R

- Absolutely Crucial Skill for Modern Data Scientists

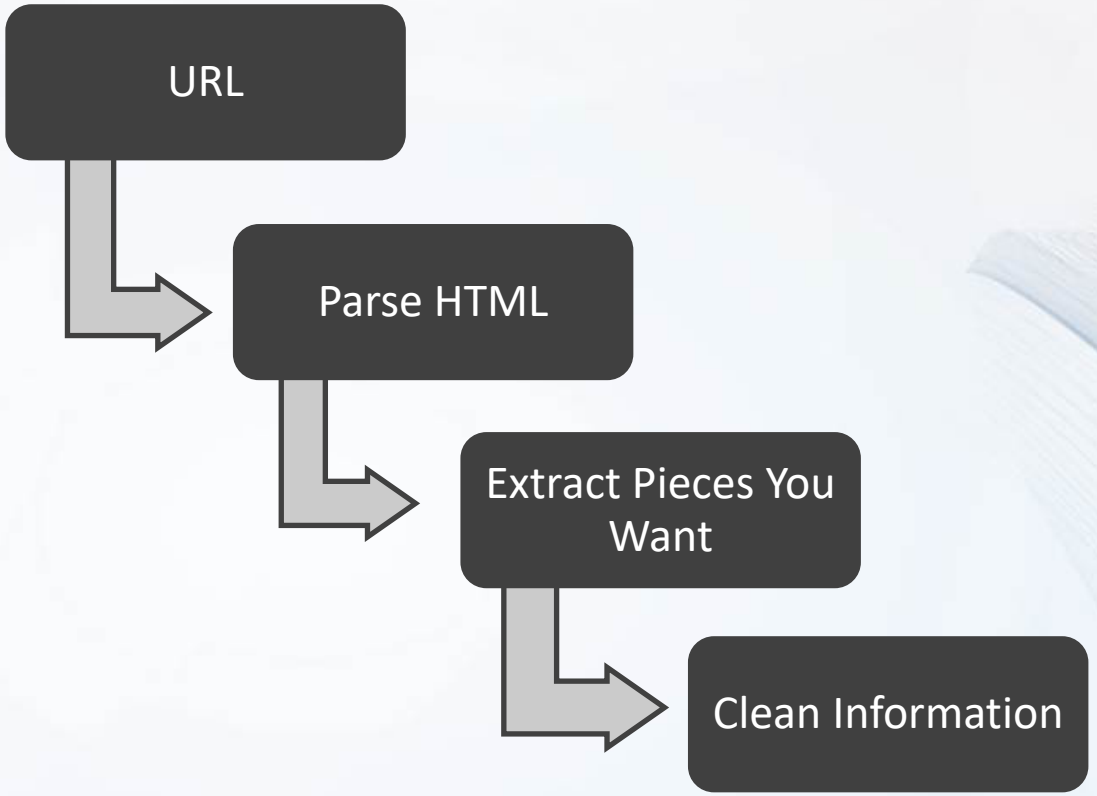
Web Scraping in R

- The rvest package

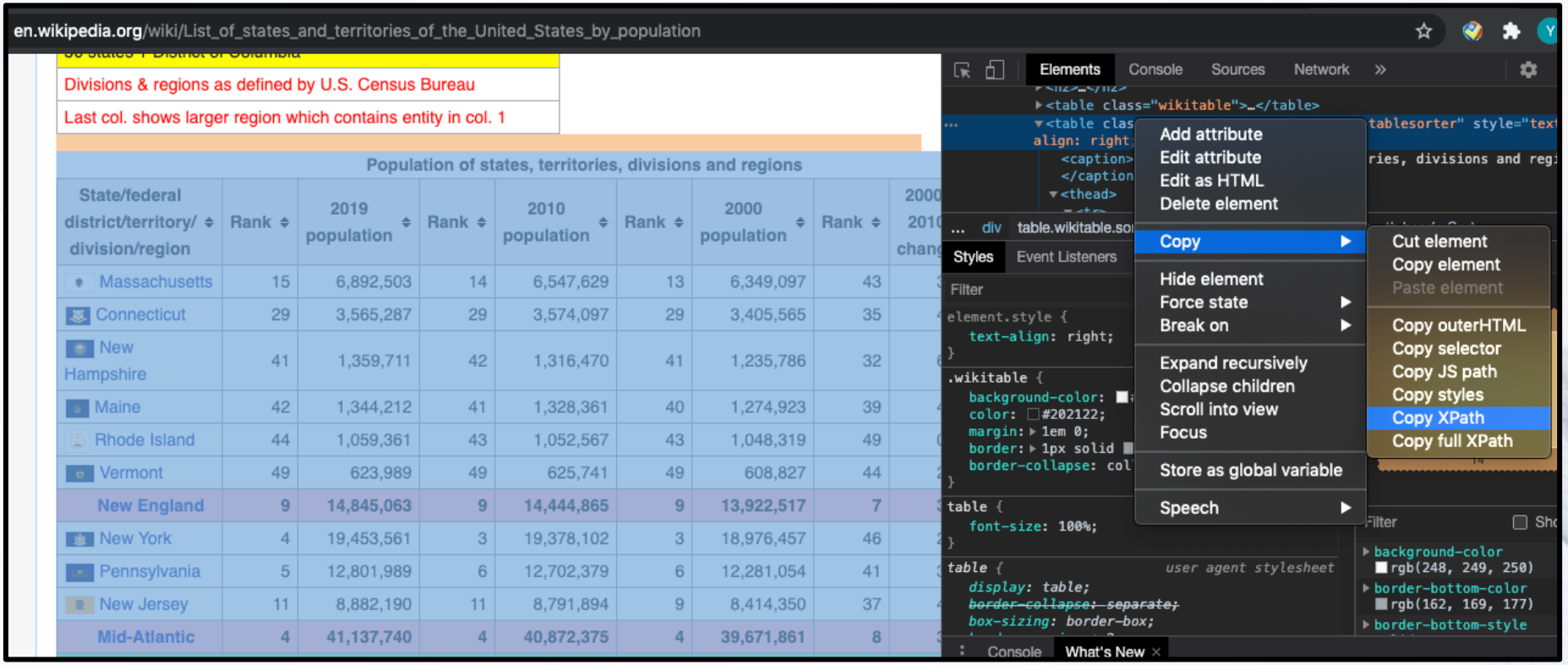
```
> library(rvest)
```

  - Written by Hadley Wickham

  - General Process:



Example 1



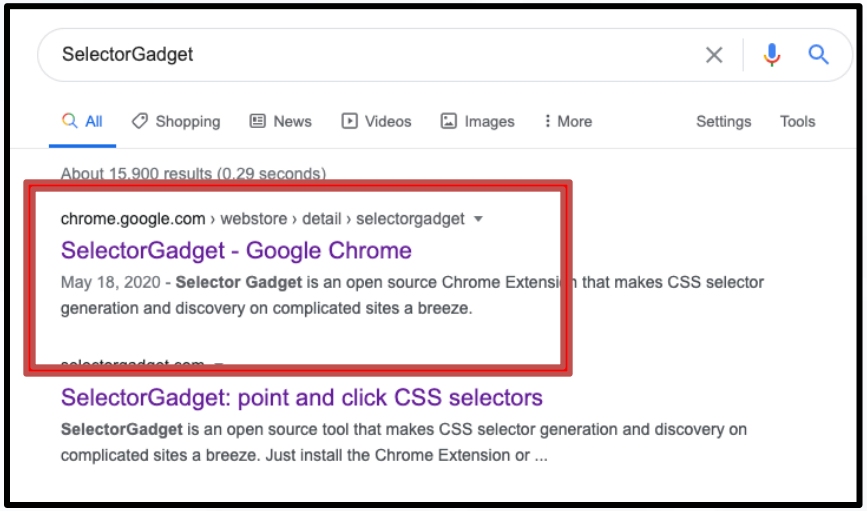
- Right click the table → inspect
- Select element, right click → copy XPath

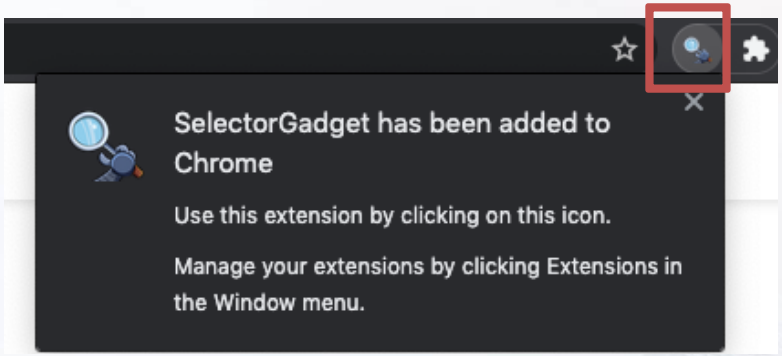
Example 1: code

```r
url <- "https://en.wikipedia.org/wiki/List_of_states_and_territories_of_the_United_States_by_population"

population <- url %>%
  read_html() %>%
  html_nodes(xpath='//*[@id="mw-content-text"]/div[1]/table[3]') %>%
  html_table() %>%
  .[[1]]

colnames(population)=c("State","Rank_2019","Pop_2019","Pop_2010","Rank_2010","Pop_2000","Rank_2000","Change","Geographic_sort")

head(population)
```

	State <chr>	Rank_2019 <chr>	Pop_2019 <chr>	Pop_2010 <chr>	Rank_2010 <chr>	Pop_2000 <chr>	Rank_2000 <chr>	Change <chr>
1	Massachusetts	15	6,892,503	14	6,547,629	13	6,349,097	43
2	Connecticut	29	3,565,287	29	3,574,097	29	3,405,565	35
3	New Hampshire	41	1,359,711	42	1,316,470	41	1,235,786	32
4	Maine	42	1,344,212	41	1,328,361	40	1,274,923	39
5	Rhode Island	44	1,059,361	43	1,052,567	43	1,048,319	49
6	Vermont	49	623,989	49	625,741	49	608,827	44

Add SelectorGadget



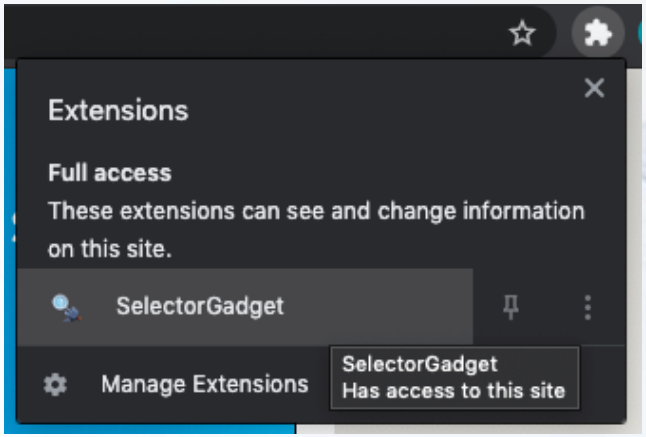


Example 2

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

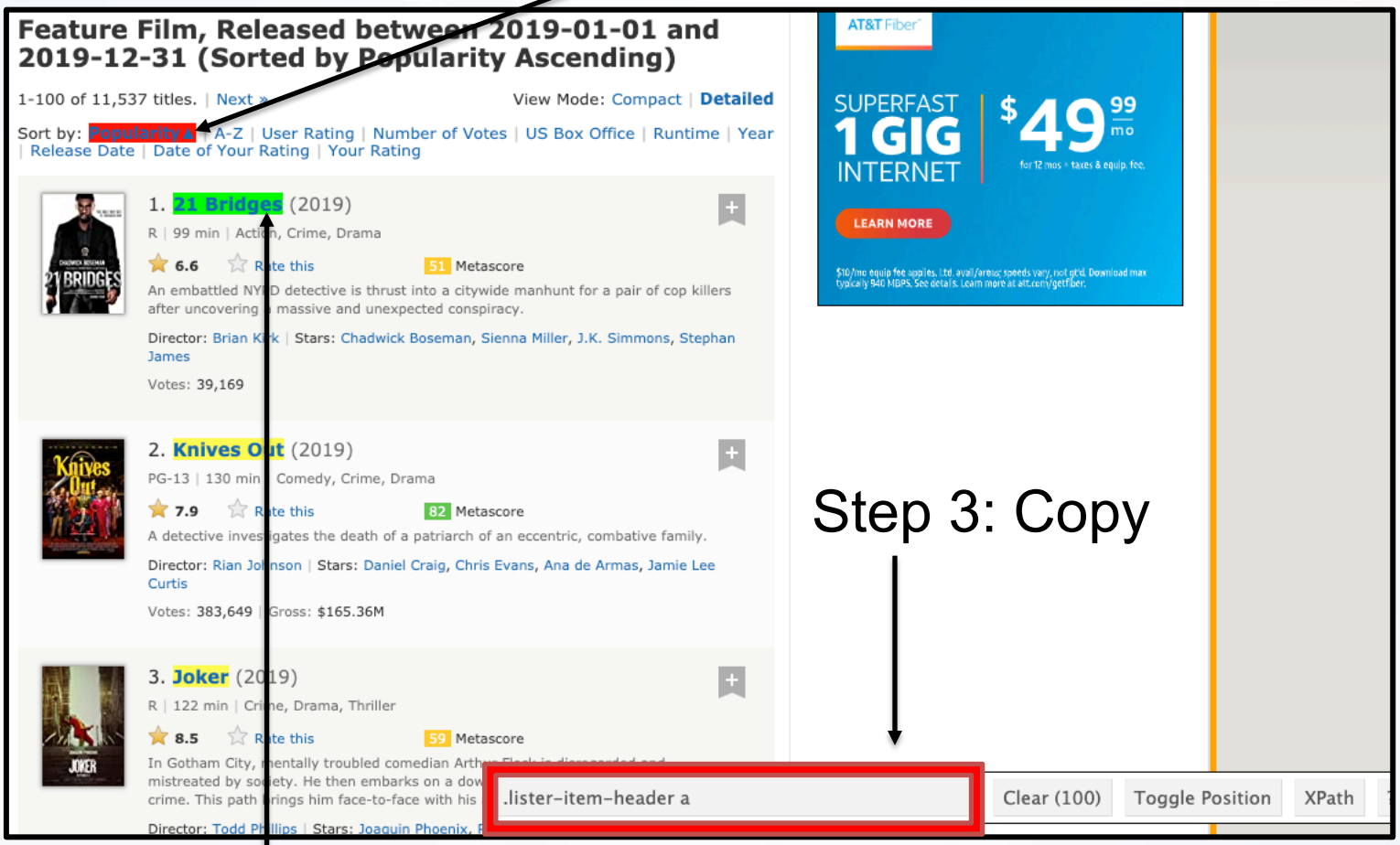
https://www.imdb.com/search/title/?count=100&release_date=2019,2019&title_type=feature

  - Choose SelectorGadget



  - Locate This Box on the bottom

Example 2



Step 2: Deselect

Step 3: Copy

Step1: Select things you want to scrap

Example 2: code

```r
url <- "https://www.imdb.com/search/title/?count=100&release_date=2019,2019&title_type=feature"

movies <- url %>%
  read_html() %>%
  html_nodes('.lister-item-header a') %>%
  html_text()
```

```
 [1] "21 Bridges"
 [2] "Knives Out"
 [3] "Joker"
 [4] "Avengers: Endgame"
 [5] "I See You"
 [6] "Once Upon a Time... in Hollywood"
 [7] "The Peanut Butter Falcon"
 [8] "Parasite"
 [9] "Ready or Not"
[10] "Jojo Rabbit"
[11] "The Gentlemen"
[12] "Doctor Sleep"
[13] "1BR"
[14] "The Lighthouse"
[15] "Jumanji: The Next Level"
[16] "Bombshell"
[17] "1917"
[18] "Get Duked!"
```

Tutorial 6

- Step 1: Open Tutorial
- Step 2: Ensure You Have the Following R Packages Installed
  - tidyverse
  - rvest
  - devtools
  - noncensus (Install from Github)
- Step 3: Knit and Run
- Step 4: Read the Introduction

Part 1: Violent Crimes in US Cities

- Step 1: Wikipedia Violent Crimes
- Step 2: Locate the Table

Crime rates per 100,000 people per year

State	City	Population	Violent crime				
			Total	Murder and Nonnegligent manslaughter	Rape[1]	Robbery	Aggravated assault
Alabama	Mobile[3]	248431	740.25	20.13	57.16	177.11	485.85
Alaska	Anchorage	296188	1203.29	9.12	132.01	262.67	799.49
Arizona	Chandler	249355	259.47	2.01	52.13	56.95	148.38
Arizona	Gilbert	242090	85.51	2.07	16.11	21.07	46.26
Arizona	Glendale	249273	488.22	4.81	38.91	192.96	251.53
Arizona	Mesa	492268	415.83	4.67	51.19	92.23	267.74
Arizona	Phoenix	1644177	760.93	9.55	69.46	200.28	481.64
Arizona	Scottsdale	251840	157.24	1.99	40.90	39.71	74.65
Arizona	Tucson	532323	801.77	8.64	93.55	268.82	430.75
California	Anaheim	353400	354.56	2.83	32.54	135.82	183.36
California	Bakersfield	381154	479.33	10.76	24.14	197.56	246.88

Goal: Read Table Into R

Part 1: Violent Crimes in US Cities

- Step 3: What Do You Expect to Be a Problem in the Data?

- Step 4: Run Chunk 1

  - Is This What You Expected?
  - What New Problems Arise?

- Step 5: Run Chunk 2

  - Select Wanted Information
  - Remove 1st Row – Subgroups
  - Rename Variables

Part 1: Violent Crimes in US Cities

- Step 6: Run Chunk 3
  - Converting Variable Types
    - as.numeric()
    - as.character()
    - as.date()
    - as.integer()
  - All Numeric Variables are Character Because of First Row
- Step 7: Run Chunk 4
  - City Variable Has Problems
  - Why Do We Care?

Part 1: Violent Crimes in US Cities

- Step 8: Run Chunk 5
  - String Functions Used
    - str_replace_all()
    - str_replace()
  - Conditional Mutation
    - ifelse()
- Step 9: Base Knit

Part 2: Geographical Locations of US Cities

- Step 1: What Additional Information Would We Need to Plot Crime Information on a Map?

- Step 2: Run Chunk 1

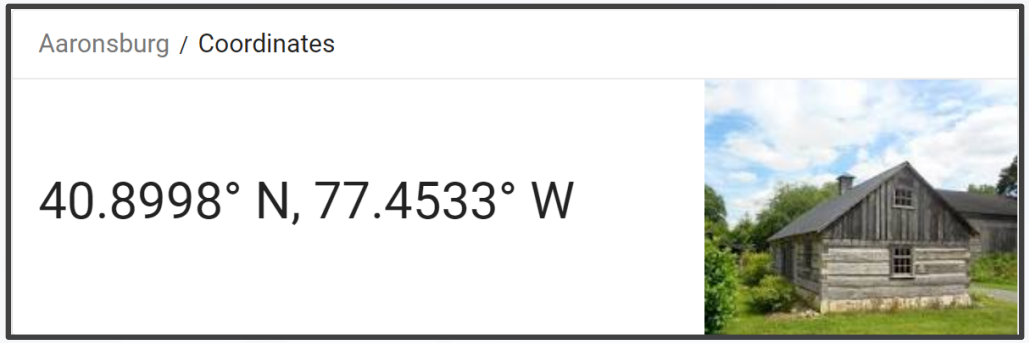
  - What Info is Important?
  - What Do You Notice About the City Variable?

- Step 3: Run Chunk 2

  - Goal: Find the Average Latitude and Longitude for Each City and State

Part 2: Geographical Locations of US Cities

- Step 4: Run Chunk 3
  - Examine the Output
  - Notice Aaronsburg, PA



  - Are We Ready to Merge?
    - #No
    - #WhyNot
- Step 5: Pinch Knit

Part 3: Linking State Names to State Abbreviations

- Step 1: Select Website Link

- Step 2: Examine the Table

Name	Abbreviation	Name	Abbreviation
Alabama	AL	Montana	MT
Alaska	AK	Nebraska	NE
Arizona	AZ	Nevada	NV
Arkansas	AR	New Hampshire	NH
California	CA	New Jersey	NJ
Colorado	CO	New Mexico	NM
Connecticut	CT	New York	NY

- Step 3: What is the Issue with the Way this Information is Presented and How Does this Pose a Threat to Our Existence?

Part 3: Linking State Names to State Abbreviations

- Step 4: Run Chunk 1
  - Did You Get What You Expected?
  - How Should We Fix This Data?

- Step 5: Run Chunk 2
  - Stacking Datasets
    - Horizontally `> cbind(x,y)`
    - Vertically `> rbind(x,y)`

- Step 6: Knitting Streak

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 7 Introduction

- Step 1: Open Tutorial 7
- 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

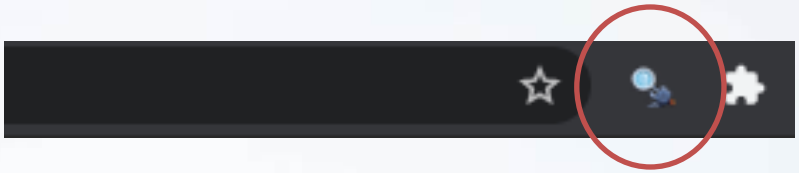
- Step 3: Run Chunk 1
    - What is required to convert the Pop_2019 to a numeric variable?
    - What is required to convert the Land to a numeric variable?
    - What is required to convert the Density to a numeric variable?
- Step 4: Run Chunk 2
    - Notice: “,|km2”,”,|/km2”

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

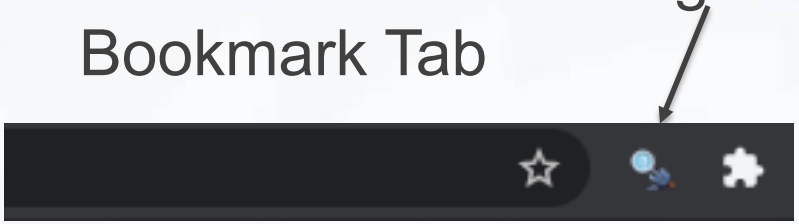
1. Vermont	1. Mississippi
2. Maine	2. Louisiana
3. Minnesota	3. Oklahoma
4. Utah	4. Texas
5. New Hampshire	5. Florida
6. Connecticut	6. Arkansas
7. Rhode Island	7. Alabama
8. Hawaii	8. Missouri
9. Massachusetts	9. Alaska
10. Washington	10. South Carolina

  - 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

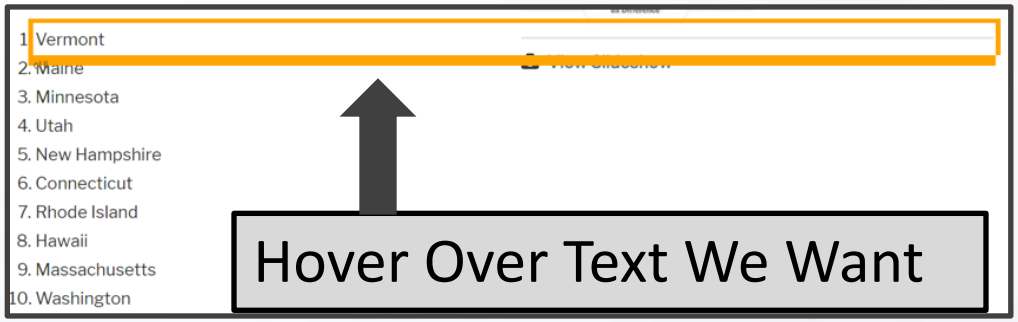
  - Choose SelectorGadget in Bookmark Tab



  - Locate This Box

Part 2: Inclusion of Expert Opinion

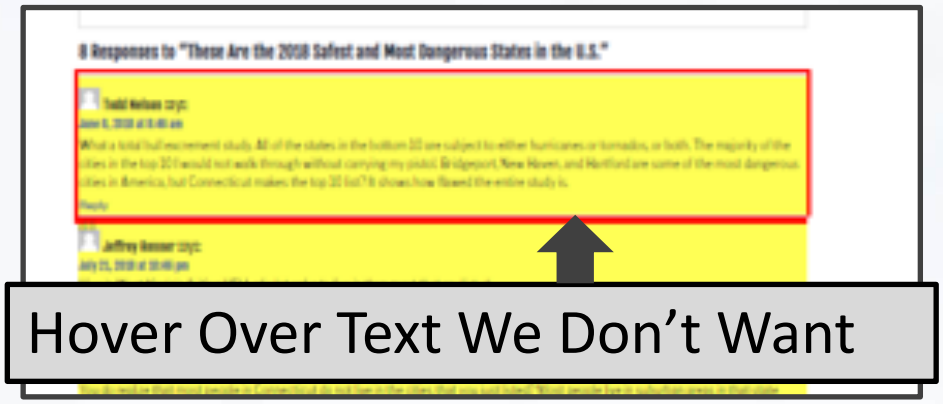
- Step 4: Continued
  - Find Content You Want



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

Part 2: Inclusion of Expert Opinion

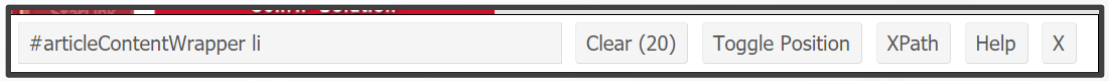
- Step 4: Continued
  - Find Content You Don't Want



  - 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