



Web Scraping I

Motivation for Web Scraping



- 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 for Web Scraping

- Example 1: ESPN NHL Stats

NHL Player Points Statistics - 2017-18

Statistics: Points | Shooting | Goaltending | Defensive | Time On Ice | Faceoffs | Major Penalties | Minor Penalties

Season: 2017-18 Regular Season

League: NHL

Splits: Total

Positions: All

Points Leaders - All Players

													PP		SH	
RK	PLAYER	TEAM	GP	G	A	PTS	+/-	PIM	PTS/G	SOG	PCT	GWG	G	A	G	A
1	Connor McDavid, C	EDM	82	41	67	108	20	26	1.32	274	15.0	7	5	15	1	3
2	Claude Giroux, LW	PHI	82	34	68	102	28	20	1.24	193	17.6	1	9	27	0	0
3	Nikita Kucherov, RW	TB	80	39	61	100	15	42	1.25	279	14.0	7	8	28	0	0
4	Evgeni Malkin, C	PIT	78	42	56	98	16	87	1.26	239	17.6	7	14	24	0	0
5	Nathan MacKinnon, C	COL	74	39	58	97	11	55	1.31	284	13.7	12	12	20	0	1
6	Taylor Hall, LW	NJ	76	39	54	93	14	34	1.22	278	14.0	7	13	24	1	0
7	Anze Kopitar, C	LA	82	35	57	92	21	20	1.12	200	17.5	6	7	20	0	2
	Phil Kessel, RW	PIT	82	34	58	92	-4	36	1.12	261	13.0	6	12	30	0	0
9	Blake Wheeler, RW	WPG	81	23	68	91	13	52	1.12	246	9.4	2	6	34	0	2
10	Sidney Crosby, C	PIT	82	29	60	89	0	46	1.09	247	11.7	6	9	29	0	0



Motivation for Web Scraping



- 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 for Web Scraping



- Example 3: AP Top 50 Stories

AP Top News

50 stories

20 mins ago

'Deliberate act of compassion' a reaction to Vegas shooting



LAS VEGAS (AP) — As a cloud-streaked orange sunset glowed over Las Vegas, officials, victims' families and survivors of year's mass shooting at a country music festival marked the first anniversary of the tragedy by placing roses on a tribute wall and dedicating a memorial garden Wednesday...

[Shootings](#) [Las Vegas mass shooting](#) [North America](#) [Las Vegas](#) [Brian Sandoval](#) [U.S. News](#)

2 hours ago

White House gives FBI freer rein in Kavanaugh investigation



WASHINGTON (AP) — The White House has given the FBI clearance to interview anyone it wants to by Friday in its investigation of sexual misconduct allegations against Supreme Court nominee Brett Kavanaugh.

The new guidance, described to The Associated Press by a person familiar with it, was...

[Sexual misconduct](#) [Supreme courts](#) [Kavanaugh nomination](#) [Politics](#) [North America](#) [U.S. Supreme Court](#) [Courts](#) [Christine Blasey Ford](#)

Web Scraping Defined



- Process of Converting Currently Unstructured Data on Web to Structured Data in R
- Ideas:
 - ESPN Table to CSV
 - Blood Pressure Chart to Tibble
 - Top News Stories to List in R
- Absolutely Crucial Skill for Modern Data Scientists

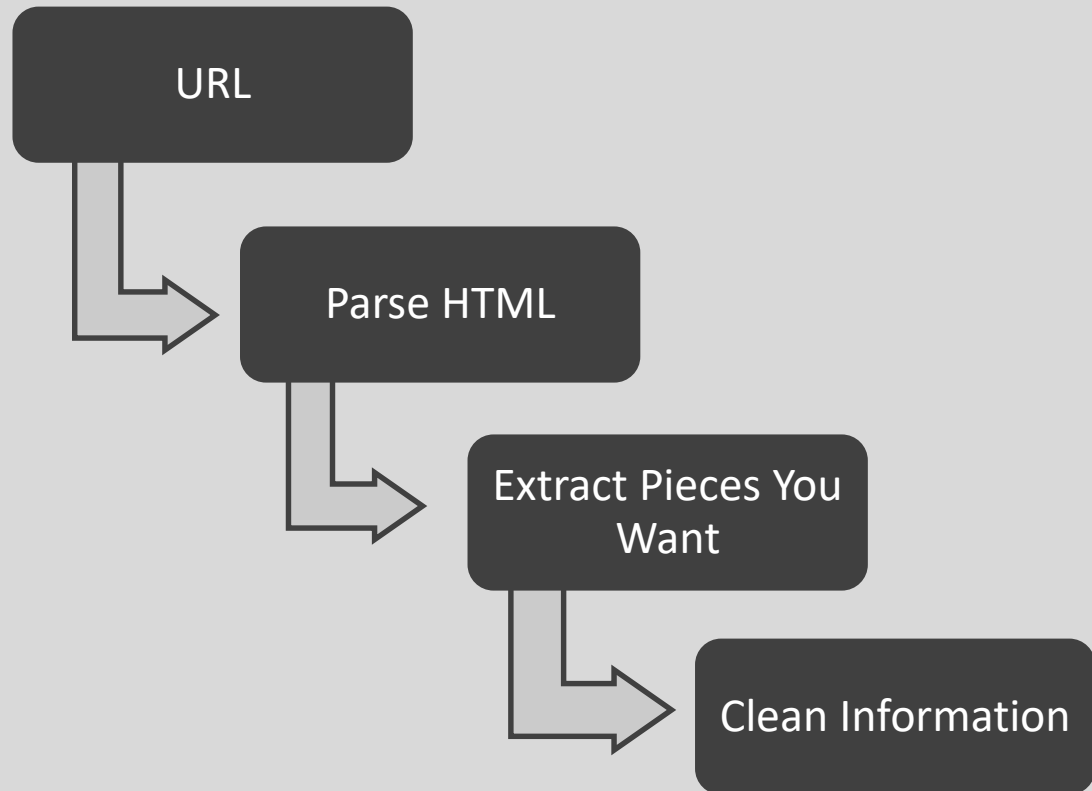
Web Scraping in R



- The rvest package

```
> library(rvest)
```

- Written by Hadley Wickham
- General Process:



Tutorial 7 Introduction



- Step 1: Open Tutorial 7
- Step 2: Ensure You Have the Following R Packages Installed
 - tidyverse
 - rvest (Requires Internet)
 - noncensus
- 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							
State ↕	City ↕	Population ↕	Violent Crime				
			Total ↕	Murder and Nonnegligent Manslaughter ↕	Rape ↕	Robbery ↕	Aggravated Assault ↕
New Mexico	Albuquerque	559,721	965.8	7.7	72.2	301.2	584.8
California	Anaheim	349,471	363.7	5.2	36.9	125.6	196.0
Alaska	Anchorage	301,239	1070.9	8.6	171.6	206.1	684.5
Texas	Arlington	387,565	502.1	2.1	53.7	136.5	309.9
Georgia	Atlanta	464,710	1119.6	20.2	36.6	429.3	633.5
Colorado	Aurora	360,237	460.8	6.7	97.7	124.1	232.3
Texas	Austin	938,728	372.5	2.5	51.9	99.0	219.2
California	Bakersfield	373,887	484.1	5.9	19.0	175.2	284.0
Maryland	Baltimore	621,252	1535.9	57.8	46.2	694.2	740.1
Massachusetts	Boston	665,258	706.8	5.7	36.1	233.1	431.9
New York	Buffalo	258,096	1118.6	15.9	67.0	400.2	635.4

➡ 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()`
 - Additional Problem in Population Variable
- 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

Aaronsburg / Coordinates

40.8998° N, 77.4533° W



- 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	<u>New Jersey</u>	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 Should All Be Saved to CSV's
- We Will Begin the Last Half of This Tutorial By Loading in this Information
- 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

Closing



Disperse
and Make
Reasonable
Decisions