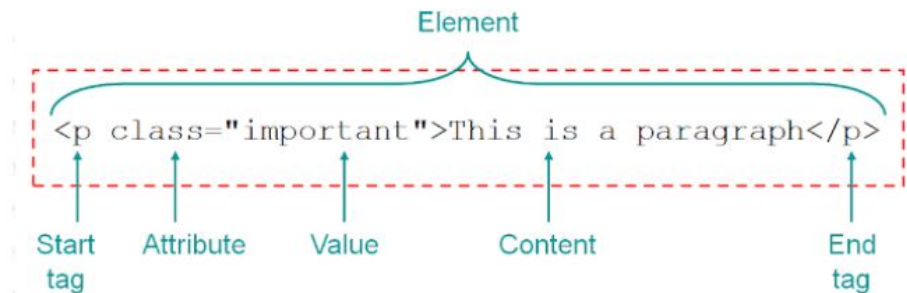


Scraping Web Data (with WNBA data)

Introduction to Scraping

- Webpages are structured using HTML
- HTML consists of various pieces to help better display things, but the content is what we're after
- Using a package called rvest we can allow R to parse through webpages and extract certain HTML elements



Things we need

- Webpage with URL for R to access
 - First we want a datapage with a table to access data from
- HTML xpath
 - We then want the HTML path to the table so that R can extract the data when it goes to the webpage
- Rvest package
 - Finally we need the R package 'rvest' installed to make the data usable in R

Webpage

We'll use the basketball reference page for WNBA star A'ja Wilson to get her 2022 game log statistics

The screenshot shows the Basketball Reference website for A'ja Wilson's 2022 game log. The page includes a navigation bar with links to various sports and sections, a search bar, and a main content area with a player profile and a game log table. There are also advertisements for American Express and Corona beer.

Basketball Reference

Enter Person, Team, Section, etc **Search**

Players Teams Seasons Leaders Scores WNBA Draft Stathead Newsletter Full Site Menu Below

A'ja Wilson 2022 Game Log

Pronunciation: \A-zhuh (like the continent, Asia)\
A'ja Wilson • [Twitter: @ajawilson22](#) • [Instagram: aja22wilson](#)
Position: Forward
6-4, 195lb (193cm, 88kg)
Born: August 8, 1996 (Age: 25-332d) in Columbia, South Carolina
College: [South Carolina](#)
Draft: [Las Vegas Aces](#), 1st round (1st pick, 1st overall), [2018 Draft](#)

2020 MVP **2x All Star**

SUMMARY

	G	PTS	TRB	AST	FG%	FG3%	FT%	eFG%	PER	WS
2022	21	18.0	10.1	1.6	48.6	34.8	79.6	51.4	25.6	3.2
Career	134	18.8	8.4	2.2	46.7	35.4	80.5	47.2	24.1	20.4

AD **FIND YOUR FAVORITE CORONA NOW** **FIND NOW**

AD **CHECK FOR OFFERS** **AMERICAN EXPRESS** **Learn More**

WNBA A'ja Wilson Overview **Game Logs** Splits Shooting Lineups On/Off

X Path

Next we want to left click the table of data, select 'Inspect Element' then when the table is highlighted we want to left click the highlighted HTML code on the side and select Copy then Copy XPath.

A'ja Wilson 2022 WNBA Game Log

« WNBA

More A'ja Wilson Pages ▼

Season	Games	Points	Rebounds	Assists	Steals	Blocks	Minutes
30-39	1						
BLK							
0	1	0	1				
1-2	13	1-2	14				
3-4	5	3-4	5				
5+	2	6	1				

21 Starts, 15 Wins, 6 Losses, 12 Double-Doubles

2022 Regular Season

table#wnba_pgl_basic.row_s
ummable.sortable.stats_tabl
e.now_sortable.sticky_table.
eq1... 932.74 x 525.25 More Stats · Switch to Widescreen View ▶

Rk	Date	Age	Tm	Opp	GS	MP	FG	FGA	FG%	3P	3PA	3P%	FT	FTA	
1	2022-05-06	25-271	LVA	@ PHO	W (+18)	1	28:35	5	8	.625	0	1	.000	5	6
2	2022-05-08	25-273	LVA	SEA	W (+11)	1	35:06	8	14	.571	1	1	1.000	3	5
3	2022-05-10	25-275	LVA	@ WAS	L (-13)	1	29:56	4	11	.364	0	0		2	2
4	2022-05-13	25-278	LVA	@ ATL	W (+23)	1	29:08	6	11	.545	0	1	.000	3	4
5	2022-05-17	25-282	LVA	PHO	W (+12)	1	33:45	4	8	.500	0	1	.000	8	11
6	2022-05-19	25-284	LVA	MIN	W (+6)	1	31:16	5	9	.556	1	2	.500	6	6
7	2022-05-21	25-286	LVA	PHO	W (+20)	1	22:58	3	7	.429	0	1	.000	3	4
8	2022-05-23	25-288	LVA	LAS	W (+28)	1	19:28	10	15	.667	2	4	.500	2	3
9	2022-05-28	25-293	LVA	@ CHI	W (+7)	1	29:04	8	19	.421	0	3	.000	6	9

Getting the data in R

Finally we can put the pieces together in R. We'll copy the webpage URL into R as a string. We then use the `rvest::html_elements()` function to extract the html elements from the chosen xpath. The xpath that we copied earlier can be pasted as the highlighted input into the function. We then use the `rvest::html_table()` function to turn the elements into a readable table and extract the table to get usable data

```
```{r}
wilson <- 'https://www.basketball-reference.com/wnba/players/w/wilsoa01w/gamelog/2022/'
wil_doc <- rvest::read_html(wilson)

wil_doc %>%
 rvest::html_elements(., xpath = "//*[@id = 'div_wnba_pgl_basic']") %>%
 rvest::html_table() -> wil
wil <- wil[[1]]
head(wil)

...

```

Rk <chr>	Date <chr>	Age <chr>	Tm <chr>	<chr>	Opp <chr>	<chr>
1	2022-05-06	25-271	LVA	@	PHO	W (+18)
2	2022-05-08	25-273	LVA		SEA	W (+11)
3	2022-05-10	25-275	LVA	@	WAS	L (-13)
4	2022-05-13	25-278	LVA	@	ATL	W (+23)
5	2022-05-17	25-282	LVA		PHO	W (+12)
6	2022-05-19	25-284	LVA		MIN	W (+6)

6 rows | 1-10 of 28 columns