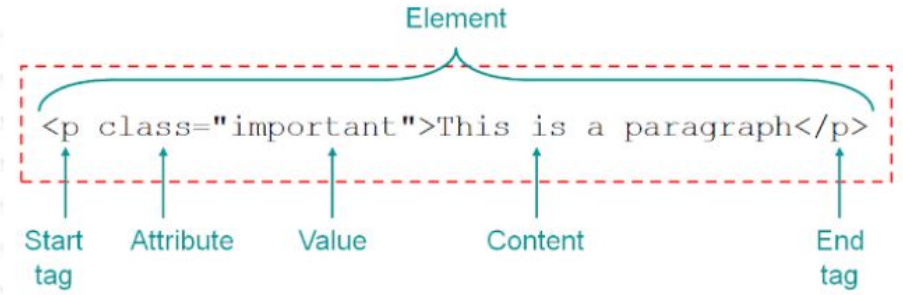
Scraping Web Data (with WNBA data)

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

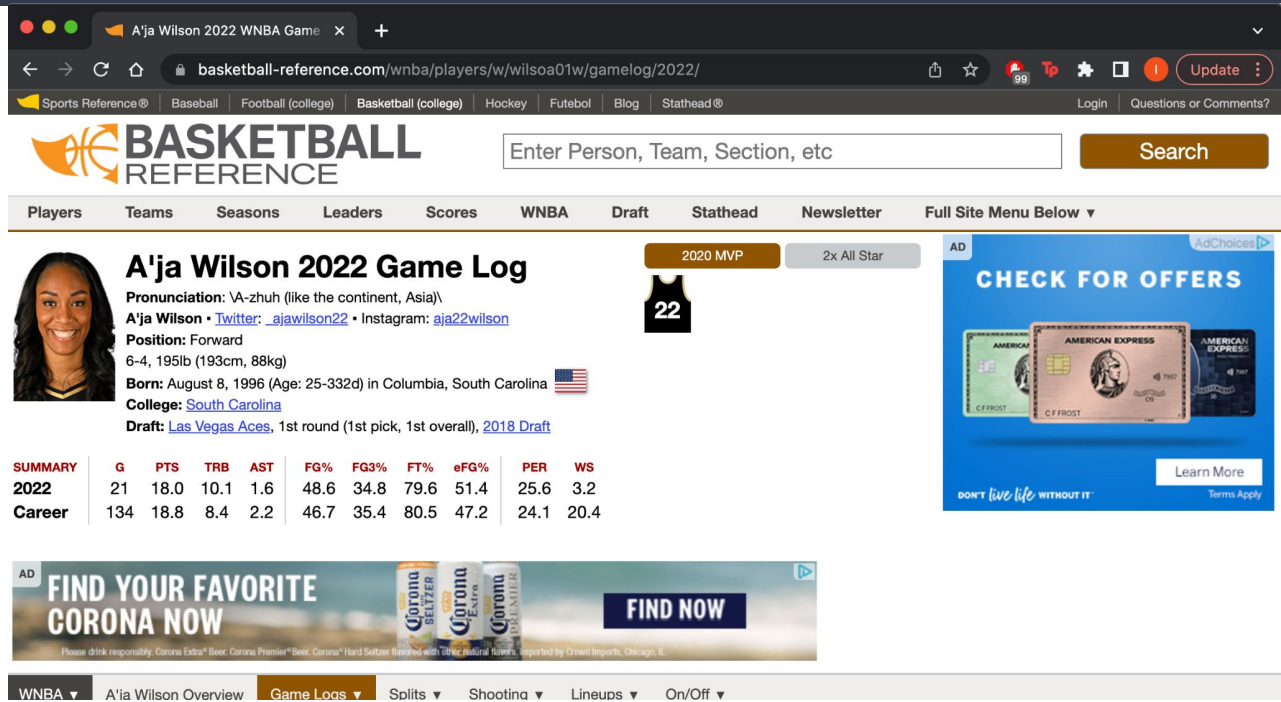
- Webpage with URL for R to access
  - First we want a datapage with a table to access data from

- 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

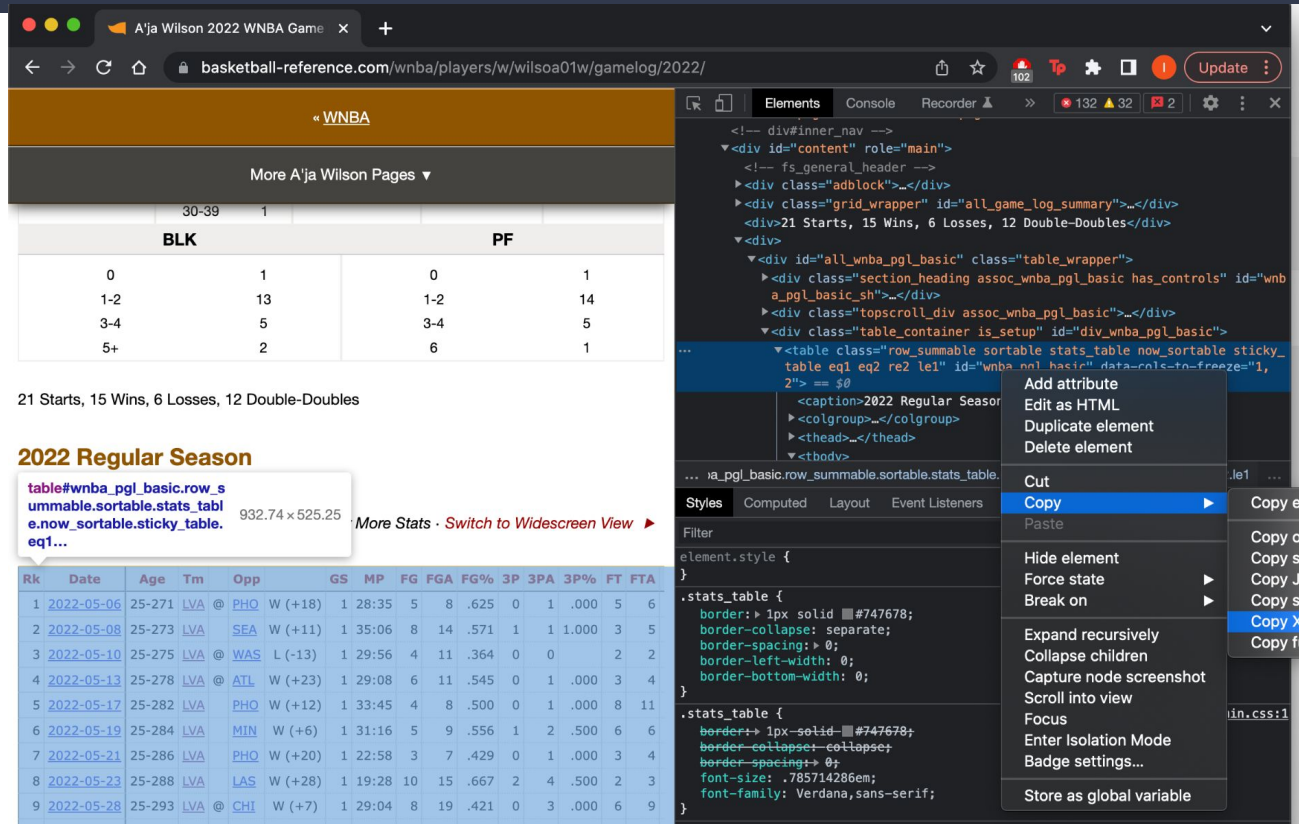
Webpage

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



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.



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