Web Scraping: A Primer

July 20th, 2023

Scraping HTML Tables

```
nhl_url <- "https://www.hockey-reference.com/leaders/games_played_career.html"
```

```
nhl_tbl <- nhl_url |>
  read_html() |>
  html_element(css = "#stats_career_NHL") |>
  html_table()
# found the css element by going through inspecting the html for the table
# and copying the selector
```

Click to go to Stringr cheat sheet

```
nhl_tbl |>
  mutate(HOF = if_else(str_detect(Player, "\\*"), 1, 0),
         Player = str_remove(Player, "\\*"),
         Rank = as.numeric(str_remove(Rank, "\\."))) |>
  fill(Rank)
```

```
## # A tibble: 250 x 5
##     Rank Player          Years      GP   HOF
##    <dbl> <chr>           <chr>   <int> <dbl>
##  1     1 Patrick Marleau 1997-21  1779     0
##  2     2 Gordie Howe     1946-80  1767     1
##  3     3 Mark Messier    1979-04  1756     1
##  4     4 Jaromír Jágr    1990-18  1733     0
##  5     5 Ron Francis     1981-04  1731     1
##  6     6 Joe Thornton    1997-22  1714     0
##  7     7 Zdeno Chára     1997-22  1680     0
##  8     8 Mark Recchi     1988-11  1652     1
##  9     9 Chris Chelios   1983-10  1651     1
## 10    10 Dave Andreychuk 1982-06  1639     1
## # i 240 more rows
```

```
fb_url <- "https://fbref.com/en/comps/183/2017-2018/2017-2018-Frauen-Bundesliga-Stats"

fb_url |>
  read_html() |>
  html_element(css = "#results2017-20181831_overall") |>
  html_table()
```

```
## # A tibble: 12 x 15
##       Rk Squad            MP     W     D     L    GF    GA    GD   Pts `Pts/MP`
##    <int> <chr>         <int> <int> <int> <int> <int> <int> <int> <int>    <dbl>
##  1     1 Wolfsburg        22    18     2     2    56     8    48    56     2.55
##  2     2 Bayern Munich    22    17     2     3    62    15    47    53     2.41
##  3     3 Freiburg         22    15     3     4    50    15    35    48     2.18
```

```
##  4     4 Turbine Potsd~    22    13     6     3    50    21    29    45     2.05
##  5     5 Essen             22    12     3     7    43    30    13    39     1.77
##  6     6 FFC Frankfurt     22    10     1    11    29    25     4    31     1.41
##  7     7 Sand              22     9     3    10    32    34    -2    30     1.36
##  8     8 Hoffenheim        22     8     1    13    22    32   -10    25     1.14
##  9     9 MSV Duisburg      22     6     0    16    16    33   -17    18     0.82
## 10    10 Werder Bremen     22     3     5    14    26    59   -33    14     0.64
## 11    11 Köln              22     3     2    17     8    78   -70    11     0.5
## 12    12 USV Jena          22     2     4    16    12    56   -44    10     0.45
## # i 4 more variables: Attendance <chr>, `Top Team Scorer` <chr>,
## #   Goalkeeper <chr>, Notes <chr>
```

Scraping Images

```
fb_url <- "https://fbref.com/en/comps/183/2017-2018/2017-2018-Frauen-Bundesliga-Stats"

fb_node <- fb_url |>
  read_html() |>
  html_element(css = "#results2017-20181831_overall")
```

```
fb_imgs <- fb_node |>
  html_elements("img") |>
  html_attr("src")
```

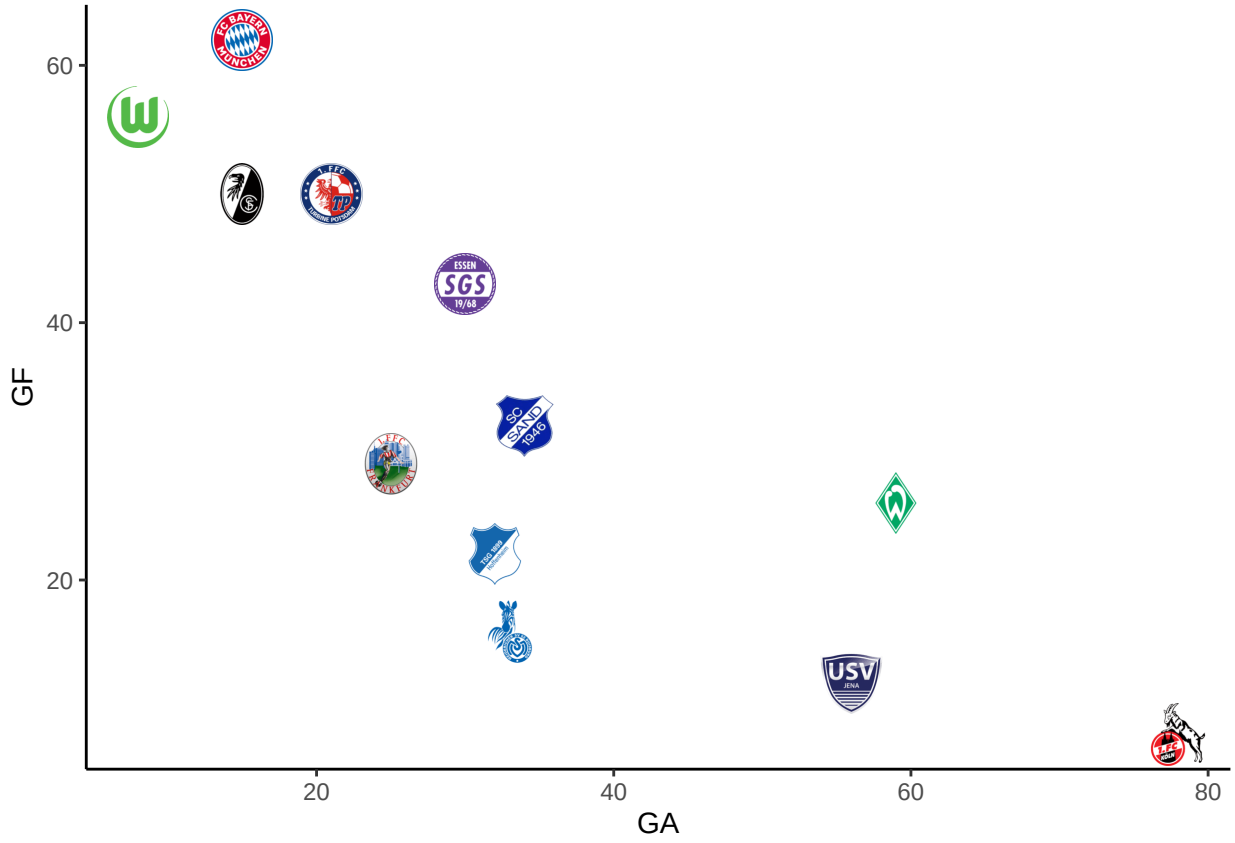
```
fb_links <- fb_node |>
  html_elements("a") |>
  html_attr("href") |>
  str_subset("squads")
```

```
fb_tbl <- fb_node |>
  html_table() |>
  mutate(img = fb_imgs,
         link = fb_links)

head(fb_tbl)
```

```
## # A tibble: 6 x 17
##      Rk Squad              MP     W     D     L    GF    GA    GD   Pts `Pts/MP`
##   <int> <chr>           <int> <int> <int> <int> <int> <int> <int> <int>    <dbl>
## 1     1 Wolfsburg          22    18     2     2    56     8    48    56     2.55
## 2     2 Bayern Munich      22    17     2     3    62    15    47    53     2.41
## 3     3 Freiburg           22    15     3     4    50    15    35    48     2.18
## 4     4 Turbine Potsdam    22    13     6     3    50    21    29    45     2.05
## 5     5 Essen              22    12     3     7    43    30    13    39     1.77
## 6     6 FFC Frankfurt      22    10     1    11    29    25     4    31     1.41
## # i 6 more variables: Attendance <chr>, `Top Team Scorer` <chr>,
## #   Goalkeeper <chr>, Notes <chr>, img <chr>, link <chr>
```

```
fb_tbl |>
  mutate(img = str_remove(img, "mini.")) |>
  ggplot(aes(GA, GF)) +
  geom_image(aes(image = img), size = 0.08, asp = 1) +
  theme_classic()
```



Scraping Text

```
wimbledon_url <- "https://en.wikipedia.org/wiki/2009_Wimbledon_Championships_-_Women%27s_singles"
```

```
wimbledon_info <- wimbledon_url |>
  read_html() |>
  html_element(css = "#mw-content-text > div.mw-parser-output > div:nth-child(13)") |>
  html_text2() |>
  str_split_1("\\n")

wimbledon_info
```

```
##  [1] "01. Dinara Safina (semifinals)"
##  [2] "02. Serena Williams (champion)"
##  [3] "03. Venus Williams (final)"
##  [4] "04. Elena Dementieva (semifinals)"
##  [5] "05. Svetlana Kuznetsova (third round)"
##  [6] "06. Jelena Janković (third round)"
##  [7] "07. Vera Zvonareva (third round, withdrew due to an ankle injury)"
##  [8] "08. Victoria Azarenka (quarterfinals)"
##  [9] "09. Caroline Wozniacki (fourth round)"
```

```
## [10] "10. Nadia Petrova (fourth round)"
## [11] "11. Agnieszka Radwańska (quarterfinals)"
## [12] "12. Marion Bartoli (third round)"
## [13] "13. Ana Ivanovic (fourth round, retired due to a thigh injury)"
## [14] "14. Dominika Cibulková (third round)"
## [15] "15. Flavia Pennetta (third round)"
## [16] "16. Zheng Jie (second round)"
## [17] "17. Amélie Mauresmo (fourth round)"
## [18] "18. Samantha Stosur (third round)"
## [19] "19. Li Na (third round)"
## [20] "20. Anabel Medina Garrigues (third round)"
## [21] "21. Patty Schnyder (first round)"
## [22] "22. Alizé Cornet (first round)"
## [23] "23. Aleksandra Wozniak (first round)"
## [24] "24. Maria Sharapova (second round)"
## [25] "25. Kaia Kanepi (first round)"
## [26] "26. Virginie Razzano (fourth round)"
## [27] "27. Alisa Kleybanova (second round)"
## [28] "28. Sorana Cîrstea (third round)"
## [29] "29. Sybille Bammer (first round)"
## [30] "30. Ágnes Szávay (first round)"
## [31] "31. Anastasia Pavlyuchenkova (second round)"
## [32] "32. Anna Chakvetadze (first round)"
```

APIs

```
f1_api <- "http://ergast.com/api/f1/constructorStandings/1/constructors.json"
f1_response <- f1_api |>
  GET()
f1_response
```

```
## Response [http://ergast.com/api/f1/constructorStandings/1/constructors.json]
##   Date: 2023-07-20 14:59
##   Status: 200
##   Content-Type: application/json; charset=utf-8
##   Size: 2.4 kB
```

```
f1_content <- f1_response |>
  content()
glimpse(f1_content)
```

```
## List of 1
##  $ MRData:List of 7
##   ..$ xmlns           : chr "http://ergast.com/mrd/1.5"
##   ..$ series          : chr "f1"
##   ..$ url             : chr "http://ergast.com/api/f1/constructorstandings/1/constructors.json"
##   ..$ limit           : chr "30"
##   ..$ offset          : chr "0"
##   ..$ total           : chr "17"
##   ..$ ConstructorTable:List of 2
##   .. ..$ constructorStandings: chr "1"
##   .. ..$ Constructors        :List of 17
```

```
f1_constructor_list <- f1_content |>
  pluck("MRData") |>
```

```
  pluck("ConstructorTable") |>
  pluck("Constructors")

f1_constructor_list[[1]]
```

```
## $constructorId
## [1] "benetton"
##
## $url
## [1] "http://en.wikipedia.org/wiki/Benetton_Formula"
##
## $name
## [1] "Benetton"
##
## $nationality
## [1] "Italian"
```

```
f1_constructor_tbl <- f1_constructor_list |>
  as_tibble_col(column_name = "info") |> # convert list to tibble
  unnest_wider(info) # unnest a list-column into columns
f1_constructor_tbl
```

```
## # A tibble: 17 x 4
##    constructorId url                                      name  nationality
##    <chr>         <chr>                                    <chr> <chr>
##  1 benetton      http://en.wikipedia.org/wiki/Benetton_Formula Bene~ Italian
##  2 brabham-repco http://en.wikipedia.org/wiki/Brabham     Brab~ British
##  3 brawn         http://en.wikipedia.org/wiki/Brawn_GP    Brawn British
##  4 brm           http://en.wikipedia.org/wiki/BRM         BRM   British
##  5 cooper-climax http://en.wikipedia.org/wiki/Cooper_Car_Comp~ Coop~ British
##  6 ferrari       http://en.wikipedia.org/wiki/Scuderia_Ferrari Ferr~ Italian
##  7 lotus-climax  http://en.wikipedia.org/wiki/Team_Lotus   Lotu~ British
##  8 lotus-ford    http://en.wikipedia.org/wiki/Team_Lotus   Lotu~ British
##  9 matra-ford    http://en.wikipedia.org/wiki/Matra       Matr~ French
## 10 mclaren       http://en.wikipedia.org/wiki/McLaren     McLa~ British
## 11 mercedes      http://en.wikipedia.org/wiki/Mercedes-Benz_i~ Merc~ German
## 12 red_bull      http://en.wikipedia.org/wiki/Red_Bull_Racing  Red ~ Austrian
## 13 renault       http://en.wikipedia.org/wiki/Renault_in_Form~ Rena~ French
## 14 team_lotus    http://en.wikipedia.org/wiki/Team_Lotus   Team~ British
## 15 tyrrell       http://en.wikipedia.org/wiki/Tyrrell_Racing   Tyrr~ British
## 16 vanwall       http://en.wikipedia.org/wiki/Vanwall      Vanw~ British
## 17 williams      http://en.wikipedia.org/wiki/Williams_Grand_~ Will~ British
```

Polite Package

Polite ensures that you're respecting the `robots.txt` and not submitting too many requests

```
wimbledon_url <- "https://en.wikipedia.org/wiki/2009_Wimbledon_Championships_-_Women's_singles"
session <- wimbledon_url |>
  bow()
session
```

```
## <polite session> https://en.wikipedia.org/wiki/2009_Wimbledon_Championships_-_Women's_singles
##     User-agent: polite R package
##     robots.txt: 456 rules are defined for 33 bots
##    Crawl delay: 5 sec
##   The path is scrapable for this user-agent
```

```
# scrape() essentially replaces read_html() seen earlier
session |>
  scrape() |>
  html_element("#mw-content-text > div.mw-parser-output > div:nth-child(13)") |>
  html_text2() |>
  str_split_1("\\n")
```

```
##  [1] "01. Dinara Safina (semifinals)"
##  [2] "02. Serena Williams (champion)"
##  [3] "03. Venus Williams (final)"
##  [4] "04. Elena Dementieva (semifinals)"
##  [5] "05. Svetlana Kuznetsova (third round)"
##  [6] "06. Jelena Janković (third round)"
##  [7] "07. Vera Zvonareva (third round, withdrew due to an ankle injury)"
##  [8] "08. Victoria Azarenka (quarterfinals)"
##  [9] "09. Caroline Wozniacki (fourth round)"
## [10] "10. Nadia Petrova (fourth round)"
## [11] "11. Agnieszka Radwańska (quarterfinals)"
## [12] "12. Marion Bartoli (third round)"
## [13] "13. Ana Ivanovic (fourth round, retired due to a thigh injury)"
## [14] "14. Dominika Cibulková (third round)"
## [15] "15. Flavia Pennetta (third round)"
## [16] "16. Zheng Jie (second round)"
## [17] "17. Amélie Mauresmo (fourth round)"
## [18] "18. Samantha Stosur (third round)"
## [19] "19. Li Na (third round)"
## [20] "20. Anabel Medina Garrigues (third round)"
## [21] "21. Patty Schnyder (first round)"
## [22] "22. Alizé Cornet (first round)"
## [23] "23. Aleksandra Wozniak (first round)"
## [24] "24. Maria Sharapova (second round)"
## [25] "25. Kaia Kanepi (first round)"
## [26] "26. Virginie Razzano (fourth round)"
## [27] "27. Alisa Kleybanova (second round)"
## [28] "28. Sorana Cîrstea (third round)"
## [29] "29. Sybille Bammer (first round)"
## [30] "30. Ágnes Szávay (first round)"
## [31] "31. Anastasia Pavlyuchenkova (second round)"
## [32] "32. Anna Chakvetadze (first round)"
```