ETC5512: Wild Caught Data

Week 11

Sports data and web scraping

Lecturer: *Dianne Cook*

Department of Econometrics and Business Statistics

ETC5512.Clayton-x@monash.edu

Image source: https://www.welcometocountry.org/digital-genocide-ash-bartys-race-repeatedly-removed/

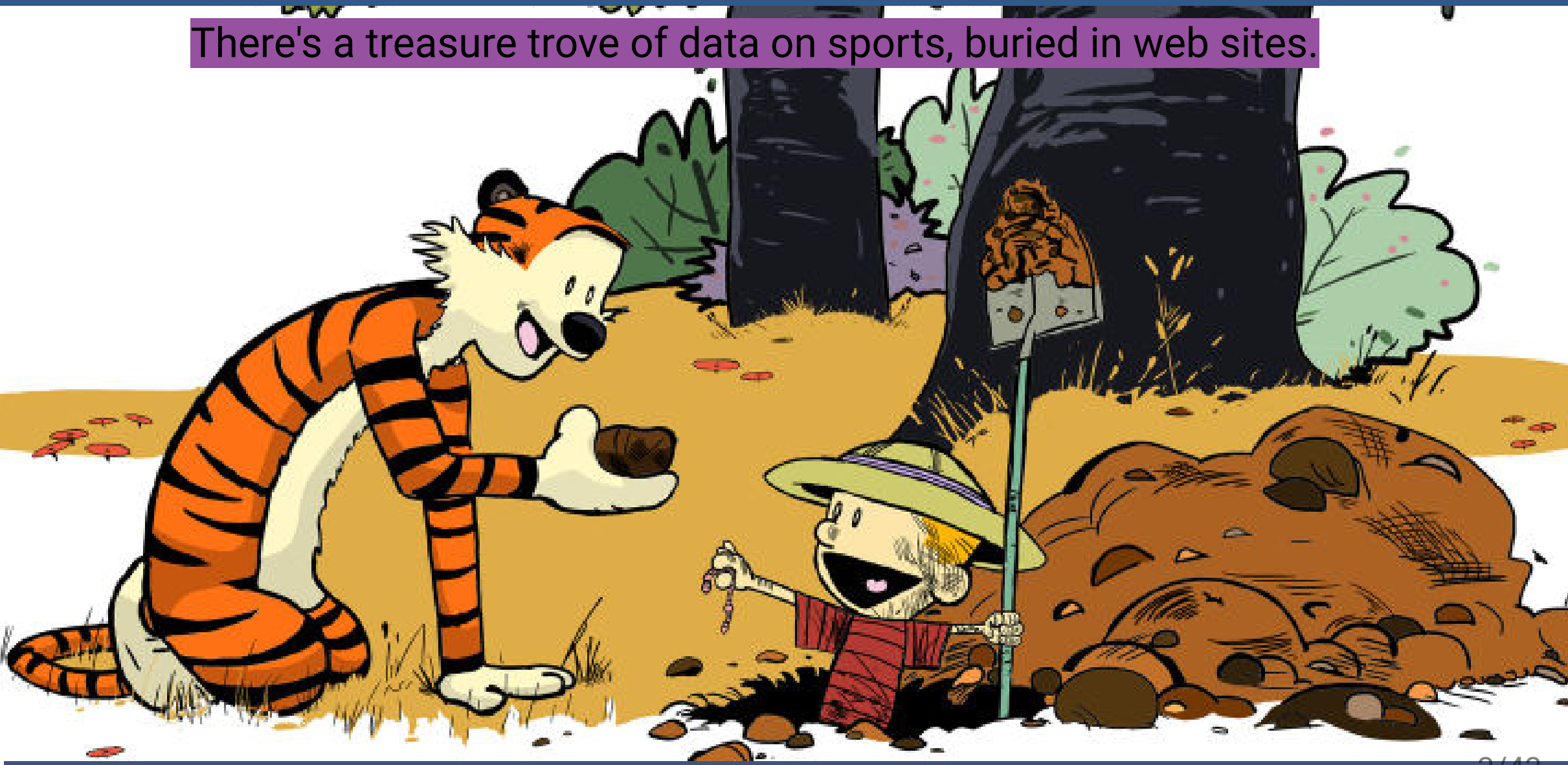
What about Barty!

- current ranking: 1
- singles titles: 8
- Prize Money: $17,594,569
- Win/Loss Singles: 252/94

Home-grown champion

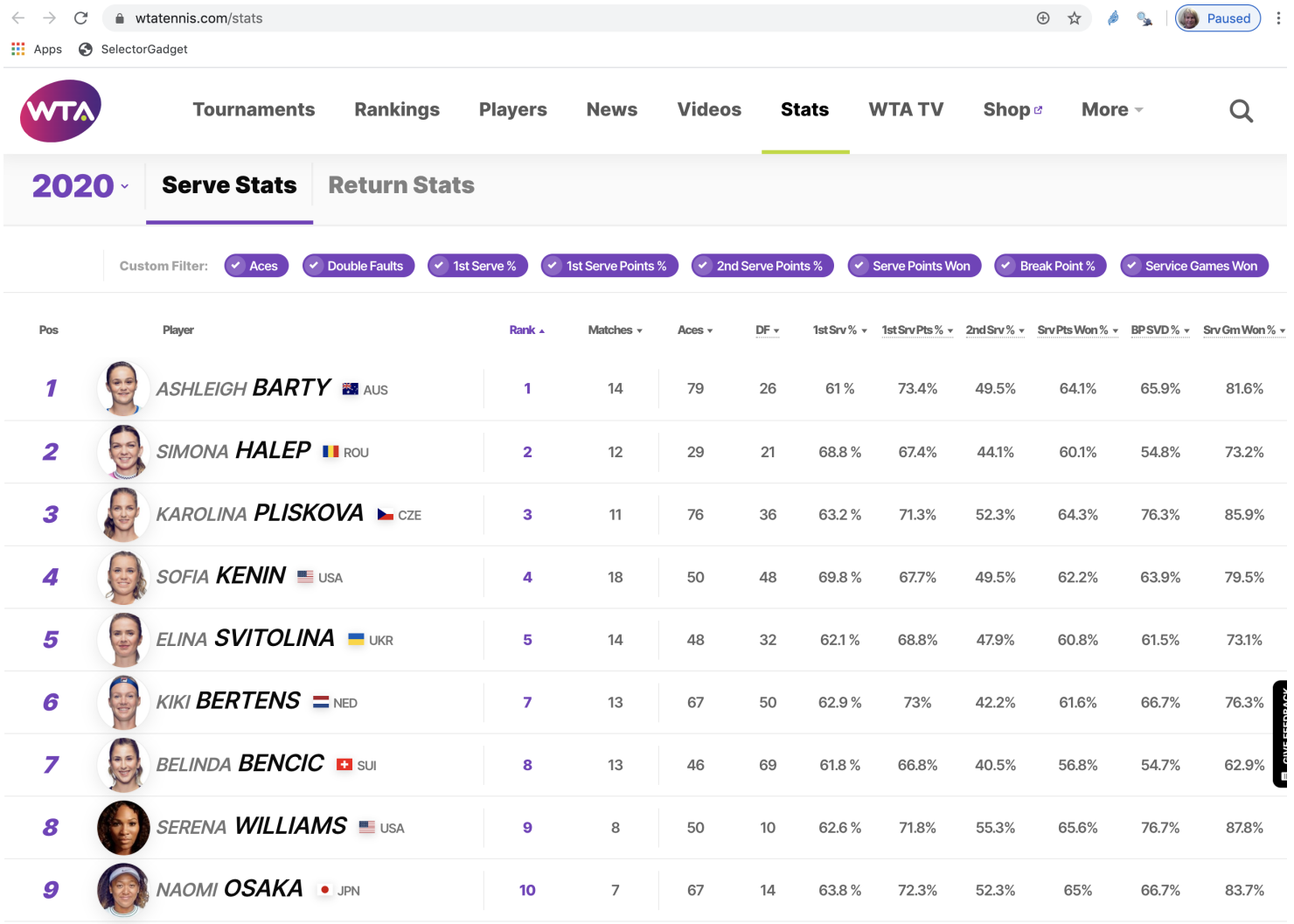
Sports data

There's a treasure trove of data on sports, buried in web sites.



Images source: https://boardgamegeek.com/

WTA



Pos	Player	Rank	Matches	Aces	DF	1st Srv %	1st Srv Pts %	2nd Srv %	Srv Pts Won %	BP SVD %	Srv Gm Won %
1	ASHLEIGH BARTY AUS	1	14	79	26	61 %	73.4%	49.5%	64.1%	65.9%	81.6%
2	SIMONA HALEP ROU	2	12	29	21	68.8 %	67.4%	44.1%	60.1%	54.8%	73.2%
3	KAROLINA PLISKOVA CZE	3	11	76	36	63.2 %	71.3%	52.3%	64.3%	76.3%	85.9%
4	SOFIA KENIN USA	4	18	50	48	69.8 %	67.7%	49.5%	62.2%	63.9%	79.5%
5	ELINA SVITOLINA UKR	5	14	48	32	62.1 %	68.8%	47.9%	60.8%	61.5%	73.1%
6	KIKI BERTENS NED	7	13	67	50	62.9 %	73%	42.2%	61.6%	66.7%	76.3%
7	BELINDA BENCIC SUI	8	13	46	69	61.8 %	66.8%	40.5%	56.8%	54.7%	62.9%
8	SERENA WILLIAMS USA	9	8	50	10	62.6 %	71.8%	55.3%	65.6%	76.7%	87.8%
9	NAOMI OSAKA JPN	10	7	67	14	63.8 %	72.3%	52.3%	65%	66.7%	83.7%

Source: https://www.wtatennis.com/stats

html source

```
<source srcset="https://photoresources.wtatennis.com/photo-resources/2019/10/08/f14eec26-4f99-4563-b904-
d94b7c70b7a1/vnoiRejq.jpg?width=56, https://photoresources.wtatennis.com/photo-resources/2019/10/08/f14eec26-
4f99-4563-b904-d94b7c70b7a1/vnoiRejq.jpg?width=112 2x" media="(max-width: 840px)">
<source srcset="https://photoresources.wtatennis.com/photo-resources/2019/10/08/f14eec26-4f99-4563-b904-
d94b7c70b7a1/vnoiRejq.jpg?width=56, https://photoresources.wtatennis.com/photo-resources/2019/10/08/f14eec26-
4f99-4563-b904-d94b7c70b7a1/vnoiRejq.jpg?width=112, https://photoresources.wtatennis.com/photo-resources/2019/10/
08/f14eec26-4f99-4563-b904-d94b7c70b7a1/vnoiRejq.jpg?width=168 2x" media="(min-width: 840px)">
<img class="object-fit-cover-picture__img js-faded-image fade-in-on-load is-loaded" src="https://
photoresources.wtatennis.com/photo-resources/2019/10/08/f14eec26-4f99-4563-b904-d94b7c70b7a1/vnoiRejq.jpg?
width=56" alt="Ashleigh Barty - Default Crop"> = $0
</picture>
</div>
▶ <div class="player-name">…</div>
</td>
▶ <td class="stats-list__cell stats-list__cell--rank stats-list__cell--fixed-width stats-list__cell--current">…</td>
▶ <td class="stats-list__cell stats-list__cell--matches stats-list__cell--fixed-width ">…</td>
<td class="stats-list__cell stats-list__cell--stat stats-list__column-serve " data-stat="Aces">79</td>
<td class="stats-list__cell stats-list__cell--stat stats-list__column-serve " data-stat="Double_Faults">26</td>
<td class="stats-list__cell stats-list__cell--stat stats-list__column-serve " data-stat="first_serve_percent">61 %</td>
<td class="stats-list__cell stats-list__cell--stat stats-list__column-serve " data-stat=
"first_serve_won_percent">73.4%</td>
<td class="stats-list__cell stats-list__cell--stat stats-list__column-serve " data-stat=
"second_serve_won_percent">49.5%</td>
<td class="stats-list__cell stats-list__cell--stat stats-list__column-serve " data-stat=
"service_points_won_percent">64.1%</td>
<td class="stats-list__cell stats-list__cell--stat stats-list__column-serve " data-stat=
"breakpoint_saved_percent">65.9%</td>
```

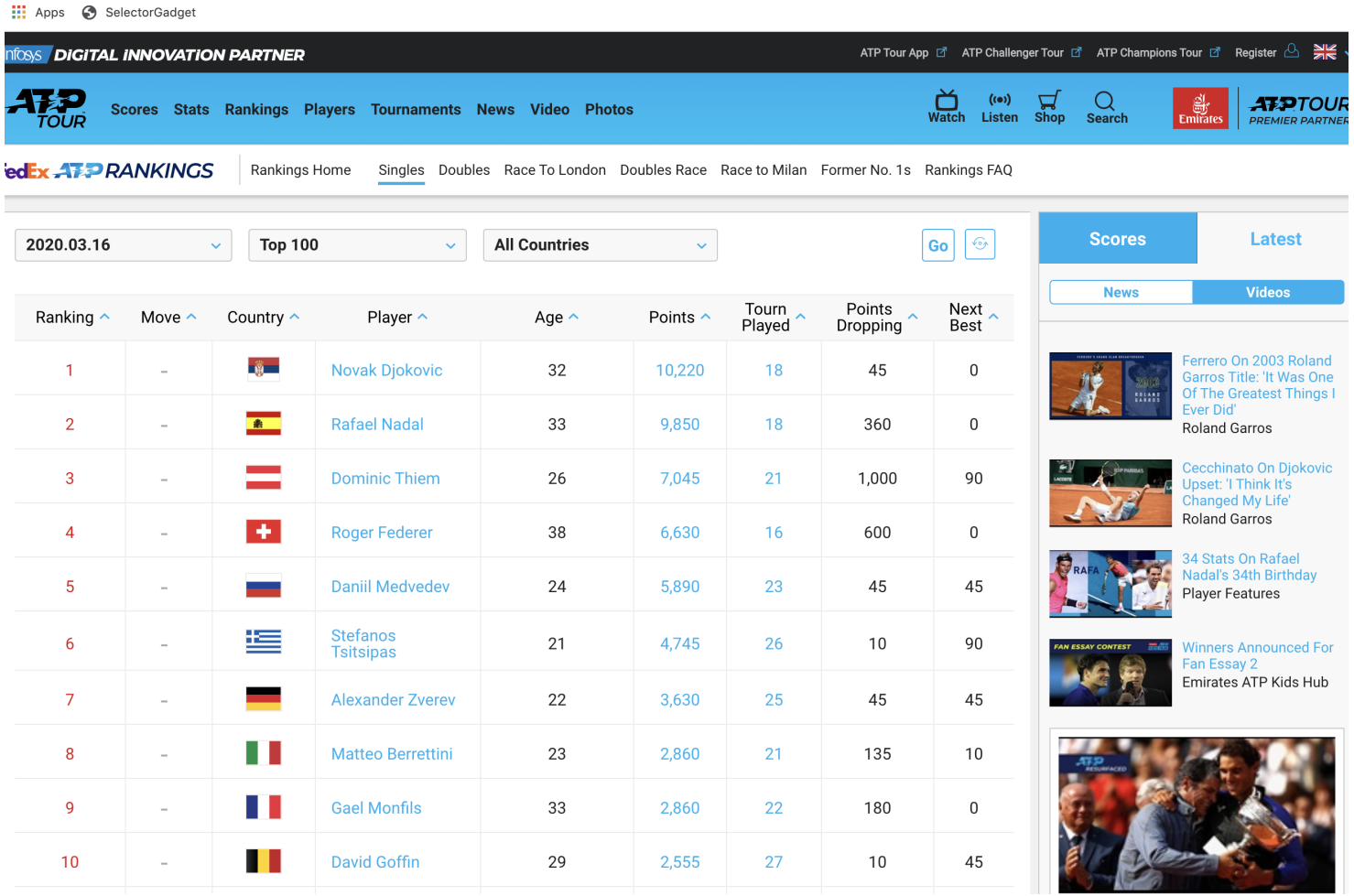
`html` is text, verbose, but nicely organised by tags. Web scraping allows harvesting data provided in web pages, by extracting the data components of the text.

```
## # A tibble: 22 x 4
##    Player                                          Rank Matches  Aces
##    <chr>                                          <int>   <int> <int>
##  1 ASHLEIGH A. BARTY  ASHLEIGH BARTY    AUS          1      14    79
##  2 KAROLINA K. PLISKOVA  KAROLINA PLISKOVA    CZE     3      11    76
##  3 SOFIA S. KENIN  SOFIA KENIN    USA                 4      18    50
##  4 ELINA E. SVITOLINA  ELINA SVITOLINA    UKR         5      14    48
##  5 KIKI K. BERTENS  KIKI BERTENS    NED               7      13    67
##  6 BELINDA B. BENCIC  BELINDA BENCIC    SUI           8      13    46
##  7 SERENA S. WILLIAMS  SERENA WILLIAMS    USA         9       8    50
##  8 NAOMI N. OSAKA  NAOMI OSAKA    JPN                10       7    67
##  9 ARYNA A. SABALENKA  ARYNA SABALENKA    BLR        11      15    64
## 10 PETRA P. KVITOVA  PETRA KVITOVA    CZE            12      15    77
## 11 MADISON M. KEYS  MADISON KEYS    USA              13       8    46
## 12 GARBIÑE G. MUGURUZA  GARBIÑE MUGURUZA    ESP      16      20   122
## 13 ELENA E. RYBAKINA  ELENA RYBAKINA    KAZ          17      25   146
## 14 MARIA M. SAKKARI  MARIA SAKKARI    GRE            20      15    60
## 15 ELISE E. MERTENS  ELISE MERTENS    BEL            23      14    46
## # … with 7 more rows
```

That took me about a half day to work out.

- The WTA (women's tennis) web site is difficult to scrape because the table content is dynamic. There are numerous javascripts which extract and load the data.
- The trick for a page like this is to save a local copy of the web page, and read it into R from this. Directly reading from the URL gets empty objects.
- It's not easy to tell that a page is dynamic, and its hard to determine if its just stupid me. Need more practice.
- ATP (men's tennis site) is much easier - its just tables, even though the reader can choose to display different tables in the page. This format is easier to automate.

ATP



Ranking	Move	Country	Player	Age	Points	Tourn Played	Points Dropping	Next Best
1	-		Novak Djokovic	32	10,220	18	45	0
2	-		Rafael Nadal	33	9,850	18	360	0
3	-		Dominic Thiem	26	7,045	21	1,000	90
4	-		Roger Federer	38	6,630	16	600	0
5	-		Daniil Medvedev	24	5,890	23	45	45
6	-		Stefanos Tsitsipas	21	4,745	26	10	90
7	-		Alexander Zverev	22	3,630	25	45	45
8	-		Matteo Berrettini	23	2,860	21	135	10
9	-		Gael Monfils	33	2,860	22	180	0
10	-		David Goffin	29	2,555	27	10	45

Source: https://www.atptour.com/en/rankings/singles

```
library(rvest)
library(tidyverse)
url_atp <- "https://www.atptour.com/en/rankings/s
atp_html <- read_html(url_atp)
atp_rankings <- html_node(atp_html, "table") %>%
  html_table(fill=TRUE)
```

```
## # A tibble: 100 x 6
##    Player              Age Points `Tourn Played` `Points Dropping` `Next Best`
##    <chr>             <int> <chr>           <int> <chr>                    <int>
##  1 Novak Djokovic       32 10,220             18 45                           0
##  2 Rafael Nadal         33 9,850              18 360                          0
##  3 Dominic Thiem        26 7,045              21 1,000                       90
##  4 Roger Federer        38 6,630              16 600                          0
##  5 Daniil Medvedev      24 5,890              23 45                          45
##  6 Stefanos Tsitsipas   21 4,745              26 10                          90
##  7 Alexander Zverev     22 3,630              25 45                          45
##  8 Matteo Berrettini    23 2,860              21 135                         10
##  9 Gael Monfils         33 2,860              22 180                          0
## 10 David Goffin         29 2,555              27 10                          45
## # … with 90 more rows
```

Data will require more processing

Notice the format of variables:

- Points is interpreted as a character
- `Points dropping` is also a character

the "," in the field isn't read as a separator in a number. These columns will need to be converted to numeric, after stripping out the "," with a text substitution.

case study

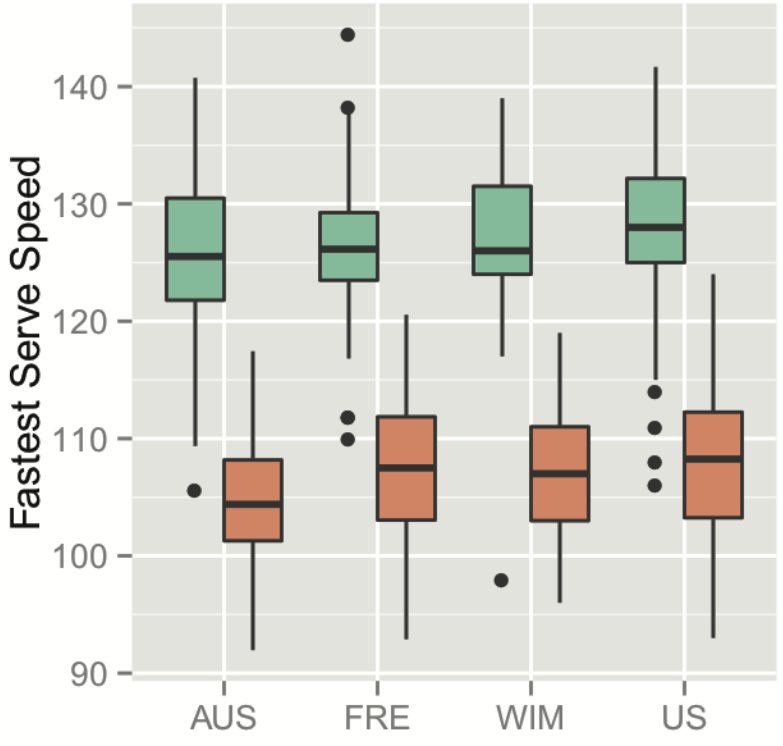
In tennis, do smashes win matches?

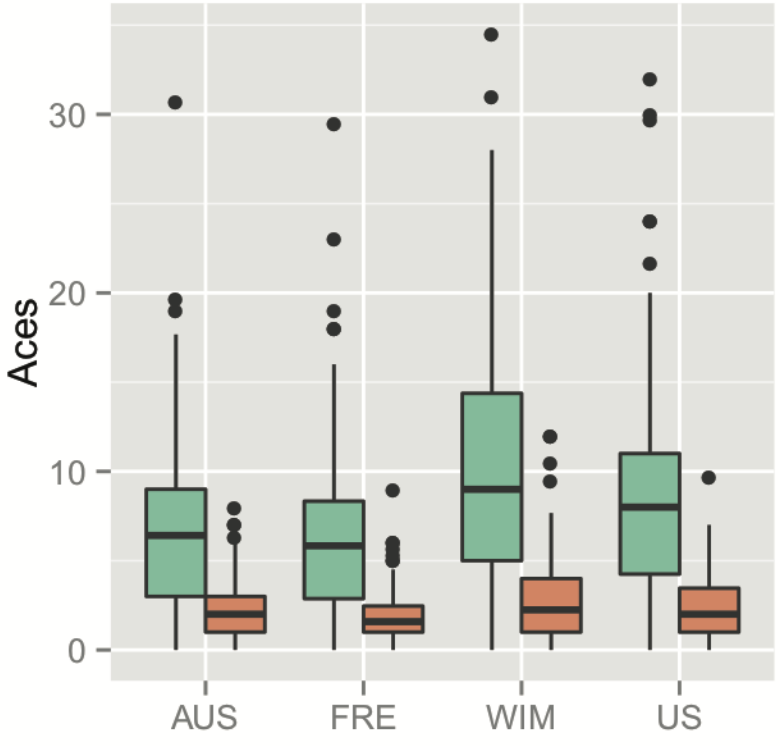
Smashes win matches analysis

Data for women and men's singles matches was scraped from the 2012 Grand Slam web sites. Statistics for each match were recorded:

- Aces
- Fastest serve speed
- Winners
- Unforced errors
- Return games won
- First serve %
- Second serve %
- Receiving points win

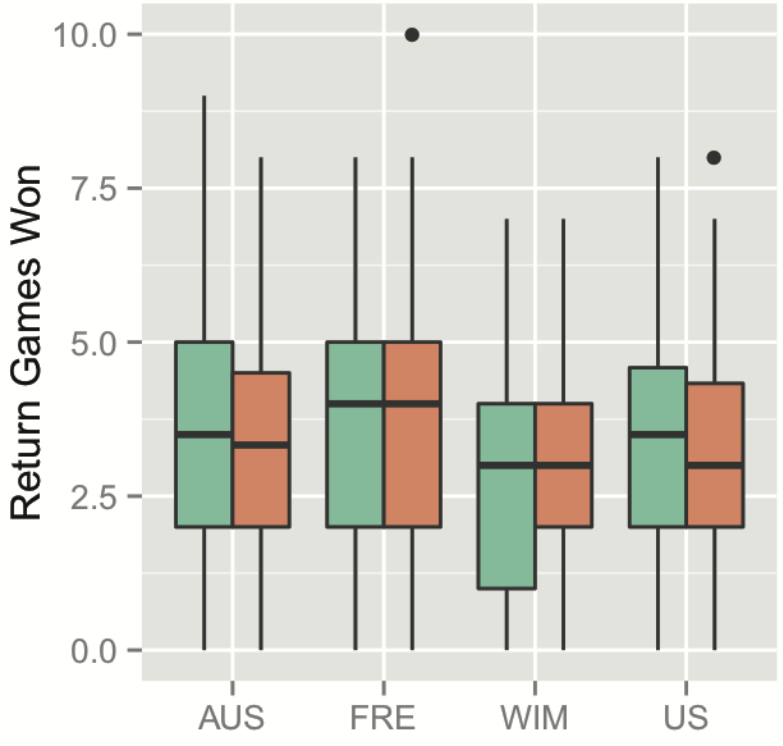
Smashes win matches analysis

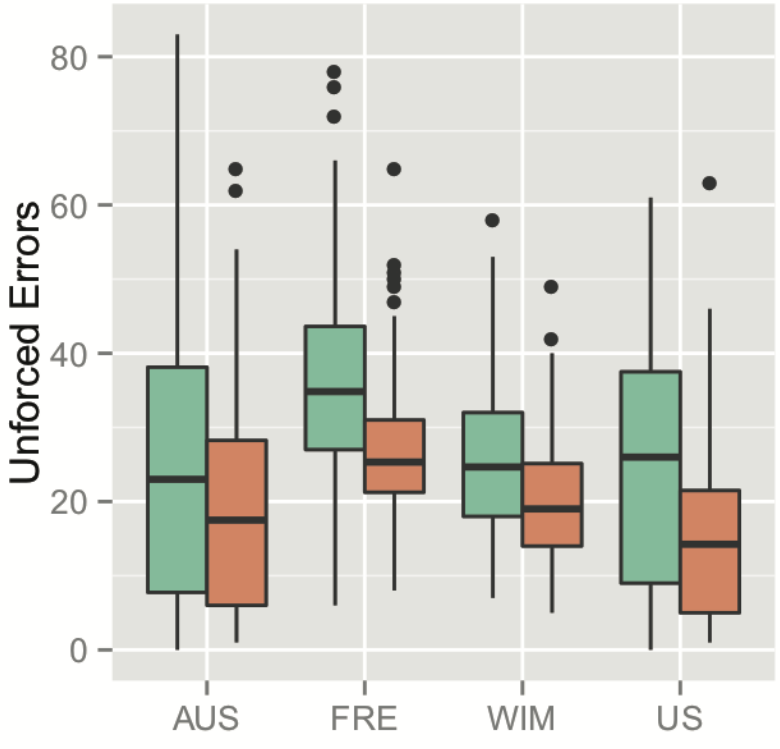




Fastest serve speed, and number of aces, in a match, by Grand Slam, and comparing men and women.

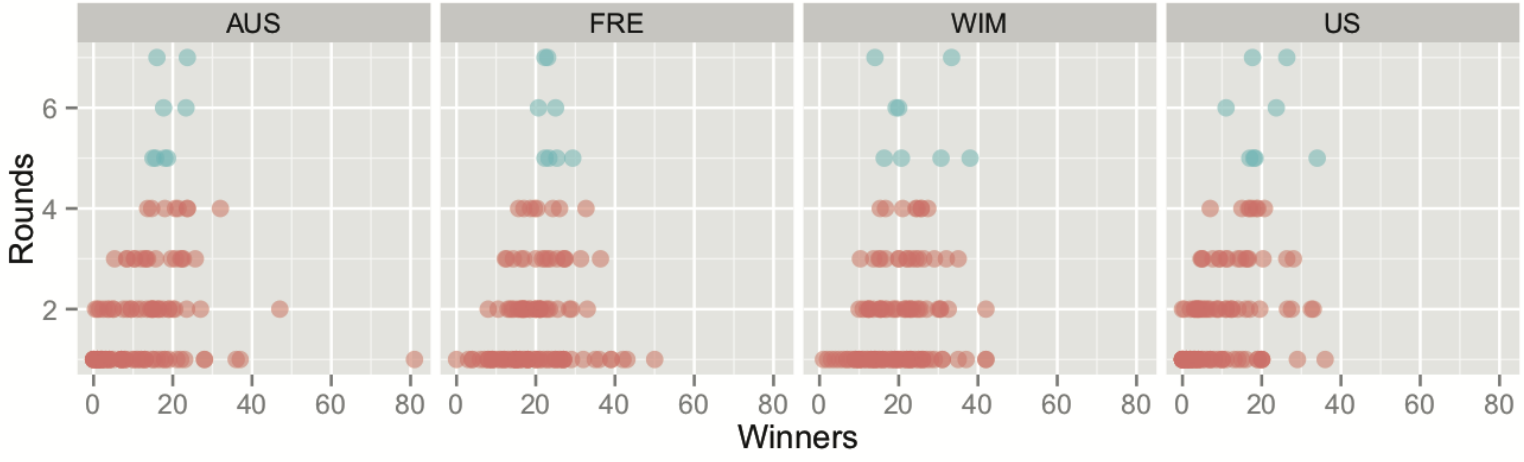
Smashes win matches analysis

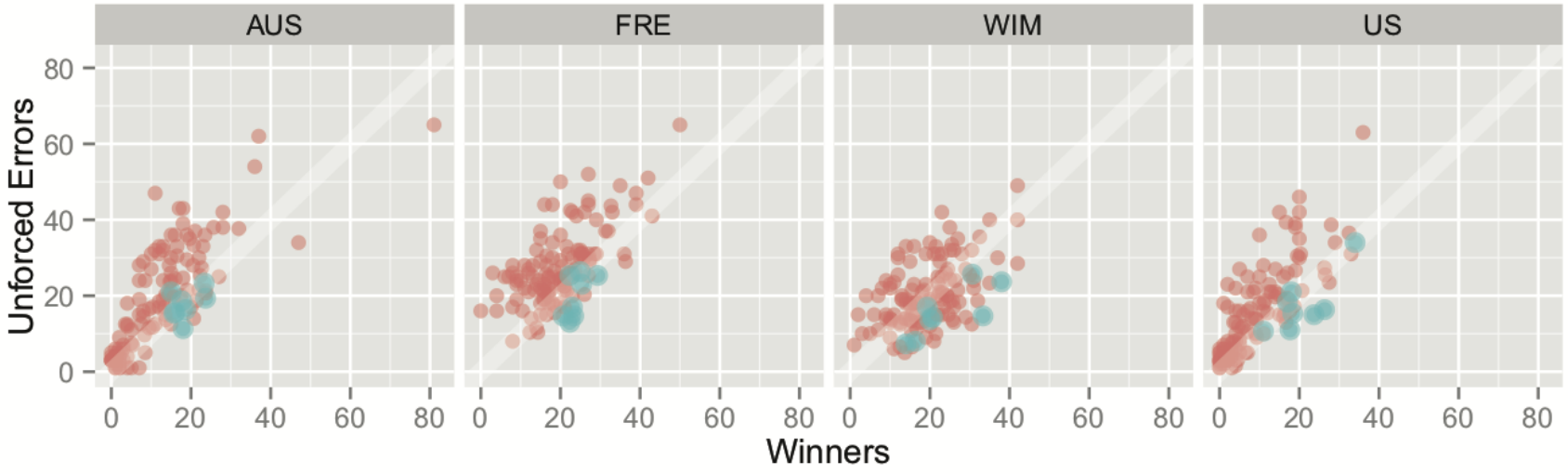




Return games won, and number of unforced errors.

Higher Round number indicates player made it further in the tournament. Statistics for women's singles matches.



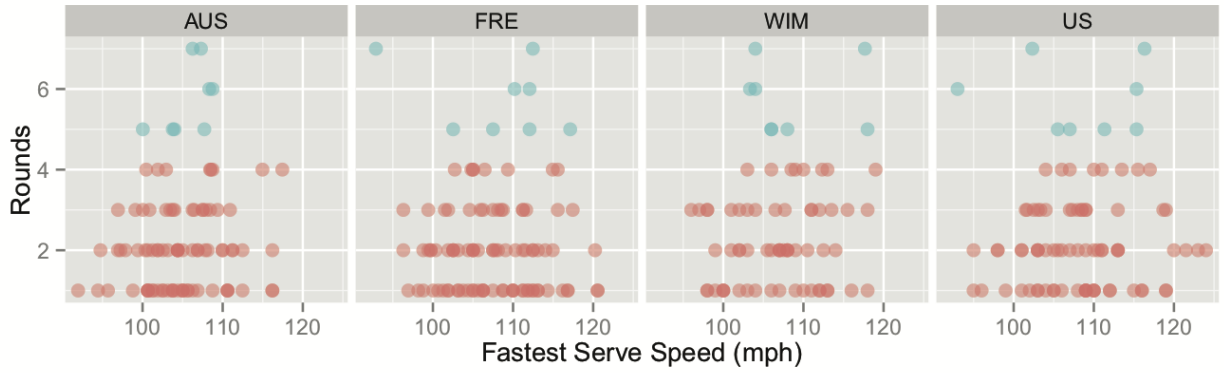


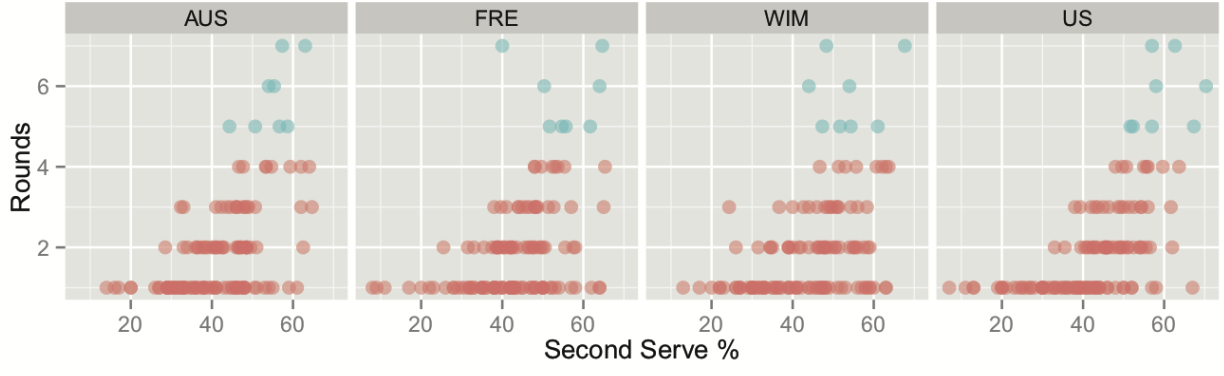
Generally, you want to have more winners than unforced errors. Too few winners might indicate that you are not working hard enough, to be able to win the match, and too many might indicate having to work too hard, so too risky, to win the match.

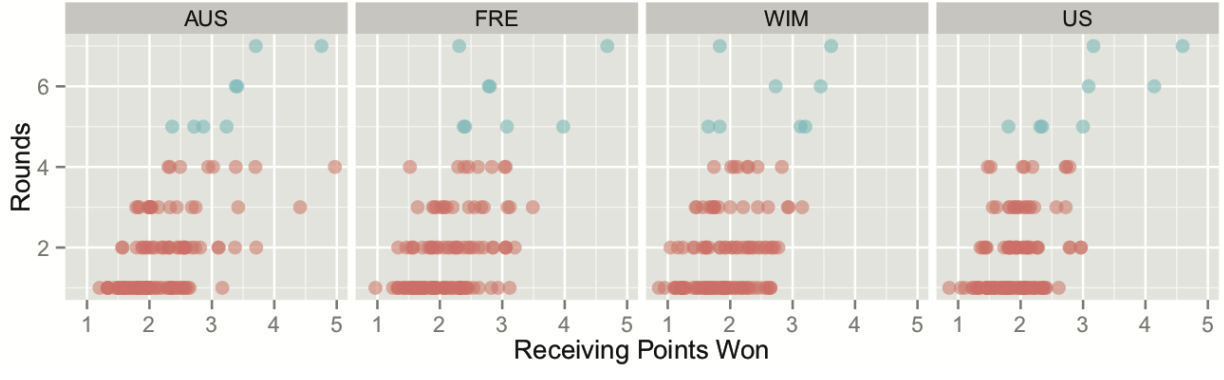
Similar for men's matches.

How important is serving?

Turns out, first serve not so much, as long as its reasonably good. The second serve % is a big indicator of progressing through a tournament, along with receiving points won.







Rules for tournament progression

Table 1. The most important statistics and values to match for men and women aiming to make the quarter-finals

Statistic	*Men's rule*	*Women's rule*
Return points won	> 39.0	> 45.6
First serve winning %	> 74.3	> 61.8
Second serve winning %	> 52.9	> 48.8
Unforced errors	between 22.3, 35.2	between 12.3, 26.5

Odds to win

Knowing the statistics of players in the first two rounds, gives pretty reliable odds of predicting the quarter finalists.

Table 2. Example odds for the women's and men's 2013 Australian Open

Player	*Odds*	*Player*	*Odds*
Serena Williams	1.1	Novak Djokovic	1.1
Maria Sharapova	1.2	Andy Murray	1.1
Sloane Stephens	1.2	Roger Federer	1.2
Caroline Wozniacki	1.2	Marin Cilic	1.3
Venus Williams	1.3	Juan Martin Del Potro	1.3
Victoria Azarenka	1.3	Jo-Wilfried Tsonga	1.4
Svetlana Kuznetsova	1.4	Tomas Berdych	1.7
Jamie Hampton	1.4	Stanislas Wawrinka	1.7

Women: S. Williams, Sharapova, Stephens, Azarenka, Kuznetsova, Radwańska, Na, Makarova

Men: Djokovic, Murray, Federer, Tsonga, Berdych, Chardy, Almagro, Ferrer

Smashes win matches analysis

Smashes are important, but only up to a point! The players who are successful are those who force the pace of the game with smashes, but who do not overdo it. Defensive play is probably more important: being able to win points on the opponent’s serve, and winning points on one’s own second serve, correlates best with getting through to the quarter-finals and the big money prizes.

Legality of scraping

- Is web scraping legal? Yes, unless you use it unethically.
  - Search engines started as web scrapers, and it boost the visibility of the page, increasing the positive sentiment towards scraping.
- Copyright infringement: if the data is copyright protected, you can't upload it to your own site, or use it for commercial purposes
- Violation of the Computer Fraud and Abuse Act: unauthorised access, eg jerk.com
- Trespass to Chattel: Don't make so many requests that you slow the web site's performance

Source: https://www.import.io/post/6-misunderstandings-about-web-scraping/

ATP terms and conditions

Terms & Conditions

3. COPYRIGHTS AND TRADEMARKS

A. Ownership

ATP owns or has the right to use all of the data, information, text, images, streaming media, video, sounds, icons, scores, rankings, statistics, and other content contained on this Website (the "Content"), and the copyrights and other intellectual property rights therein, unless otherwise noted. You may print one copy of the Content of this Website for your own personal, non-commercial use, but you may not make more than one copy of such Content, modify it in any way, distribute or transmit it to any other person or company, frame or otherwise display any of the Content of this Website on your own or any other Website, or make any other use of it. Such copying, modification, distribution, transmission, display, or use is a breach of this Agreement and infringes ATP's copyrights, copyrights licensed to ATP, rights of privacy and publicity of ATP members and others, trademark rights, and/or other rights owned or licensed by ATP.

Trademarks and service marks owned by ATP include but are not limited to: ATP, the Tennis Player Design, ATP TOUR, ATPTour.com, Nitto ATP Finals, ATP Tour Masters 1000, ATP Tour 500, ATP Tour 250, ATP Challenger Tour, and Challenger Tour Shield Designs. The following trademarks or service marks appear on the Website with the permission of their respective owners: Peugeot, Rolex, FedEx, Infosys, Nitto, Emirates Airlines, Tennis Warehouse, Penn/Head, Moët & Chandon, Tecnifibre, Skins, ProSeries, Maui Jim, Daylong, Garanti Koza, Lacoste, Nature Valley, ATP Tour events and others. You may not use any such marks in any way.

What do you think are reasonable uses of scraping the ATP data?

- ⚽ Pull the statistics of your next opponent to find their strengths and weaknesses
- ⚽ Develop odds for a gambling enterprise
- ⚽ Examine the statistics of a player prior to an injury to determine if it might be preventable
- ⚽ Develop a player ranking to build the draw for a tournament

Keep in mind

- Web scraping doesn't work forever
  - Web sites change, and code needs to be rewritten
- A web site can be made to be almost scrape-proof, but technically if its visible is scrapable
- Its more than just coding. Its pretty time-intensive to build a scraper, and then the data extracted needs to be wrangled into shape

Source: https://www.import.io/post/6-misunderstandings-about-web-scraping/

Be polite!

```
library(polite)
tennis_bow <- bow(
  url = "https://www.atptour.com/en",  # base URL
  user_agent = "Wild-caught Data <https://wcd.numbat.space>",  # identify
  force = TRUE
)
tennis_bow

## <polite session> https://www.atptour.com/en
##     User-agent: Wild-caught Data <https://wcd.numbat.space>
##     robots.txt: 20 rules are defined for 0 bots
##    Crawl delay: 5 sec
##   The path is scrapable for this user-agent
```

Hypertext Markup Language HTML Introduction

```
<!DOCTYPE html>
<html>
<body>
    This is my first web page
</body>
</html>
```

Common tags

- *html*: opening tag declaring everything between is html format
- *body*: main content appearing in the page
- *title*: title in browser border
- *table*, *tr*, *td*: table start, row and column starts
- *img*: insert an image here
- *a*: insert a link, could be external, or internal
- *h1*, *h2*, *h3*: headings in the document

Attributes

```
<h1 style="color:Tomato;"> Bilby </h1>

<img width="512" src="https://commons.wikimedia.org/wiki/image.jpg">

<a title="my image" href="https://commons.wikimedia.org/wiki/image.jpg">
<img width="512" src="https://commons.wikimedia.org/wiki/image.jpg"></a>
```

Elements

- ⚽ Bilby
- ⚽ The image is the element in the third example

Beginners guide to html

HTML files are nothing more than simple text files, so to start writing in HTML, you need nothing more than a simple text editor.

Notepad is a common text editor on Windows-based computers (usually found under the Programs > Accessories menu) and Mac OSX computers come bundled with TextEdit but any program that lets you fiddle with text will do.

Type this in to your text editor:

```
This is my first web page
```

Now create a folder called "html" wherever you like to save files on your computer and save the file as "myfirstpage.html".

! Be careful. It is important that the extension ".html" is specified - some text editors, such as Notepad, will automatically save it as ".txt" otherwise.

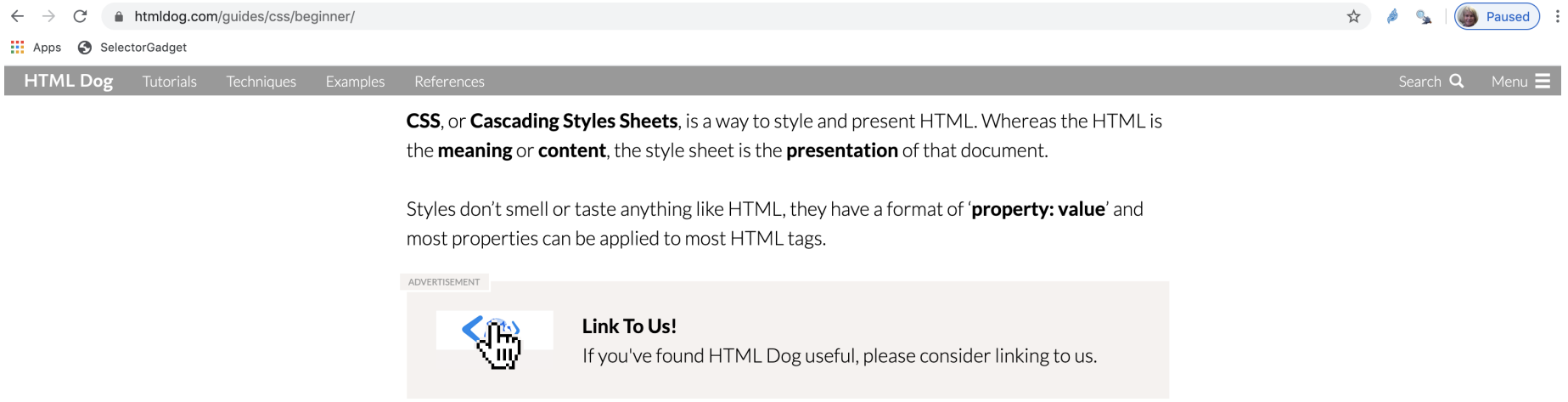
You also need to ensure that your file is being saved as **plain text**. TextEdit, for example, will start new files as "Rich text", containing lots of formatting extras, by default. In such cases, go into the preferences and make sure you check the "Plain text" format option **before creating** a new file.

Source: https://htmldog.com/guides/html/beginner/

Intro to Cascading Style Sheets (css)

- a way to style and present HTML.
- to understand parts of the html, requires knowledge of the styling components too

Intro to css



Fun interactive way to learn about css selectors at http://flukeout.github.io/.

Learning to scrape with rvest

```
library(rvest)
library(tidyverse)
url_atp <- "https://www.atptour.com/en/rankings/s
atp_html <- read_html(url_atp)
atp_rankings <- html_node(atp_html, "table") %>%
  html_table(fill=TRUE)
```

Different selector

```
lego_movie <- read_html("http://www.imdb.com/title/tt1490017/")

rating <- lego_movie %>%
  html_nodes("strong span") %>%
  html_text() %>%
  as.numeric()
rating

## [1] 7.7
```

```
cast <- lego_movie %>%
  html_nodes("#titleCast .primary_photo img") %>%
  html_attr("alt")
cast
```

```
##  [1] "Will Arnett"     "Elizabeth Banks" "Craig Berry"     "Alison Brie"
##  [5] "David Burrows"   "Anthony Daniels" "Charlie Day"     "Amanda Farino
##  [9] "Keith Ferguson"  "Will Ferrell"    "Will Forte"      "Dave Franco"
## [13] "Morgan Freeman"  "Todd Hansen"     "Jonah Hill"
```

When pages make it difficult

```
url <- "https://www.wtatennis.com/stats"
wta_html <- read_html(url)
wta_rankings <- html_node(wta_html, "table")
wta_rankings

## {xml_missing}
## <NA>
```

Download a copy first

```
wta_html <- read_html("wta_rankings2.htm")
wta_rankings <- html_node(wta_html, "table") %>% html_table(fill=TRUE)
wta_rankings <- wta_rankings %>%
  janitor::remove_empty() %>%
  as_tibble()
wta_rankings
```

```
## # A tibble: 207 x 17
##      Pos Player  Rank Matches  Aces `DF  Double Fau… `1st Srv %`
##    <int> <chr>  <int>   <int> <int>             <int> <chr>
##  1     1 ASHLE…     1      14    79                26 61 %
##  2     2 SIMON…     2      12    29                21 68.8 %
##  3     3 KAROL…     3      11    76                36 63.2 %
##  4     4 SOFIA…     4      18    50                48 69.8 %
##  5     5 ELINA…     5      14    48                32 62.1 %
##  6     6 KIKI …     7      13    67                50 62.9 %
##  7     7 BELIN…     8      13    46                69 61.8 %
##  8     8 SEREN…     9       8    50                10 62.6 %
##  9     9 NAOMI…    10       7    67                14 63.8 %
```

Sports statistics scraping packages

- Tennis: deuce package (https://github.com/skoval/deuce)
- Cricket: cricketdata (https://github.com/ropenscilabs/cricketdata)
- AFL: fitzRoy (https://jimmyday12.github.io/fitzRoy/)
- baseball: Lahman, pitchRx
- basketball: ballr
- soccer: https://github.com/statsbomb/open-data, https://github.com/JoGall/soccermatics

deuce

```
# remotes::install_github("skoval/deuce")
library(deuce)
```

- Scrapes data from http://www.atpworldtour.com/, https://www.flashscore.com/tennis.
- Developed by a Tennis Australia data scientist Stephanie Kovalchik.

NOTE: Some components broken.

cricketdata

```
# remotes::install_github("ropenscilabs/cricketdat
library(cricketdata)
```

- Scrapes data from https://docs.ropensci.org/cricketdata/
- Developed by Rob Hyndman, Timothy Hyndman, Charles Gray, Sayani Gupta
- Interesting approach to getting the URLs for the data pages

cricketdata

https://stats.espncricinfo.com/ci/engine/stats/index.html?class=10;team=289;template=results;type=batting

```
auswt20 <- fetch_cricinfo("T20", "Women", country="Aust")
auswt20
```

```
## # A tibble: 53 x 17
##    Player Start   End Matches Innings NotOuts  Runs HighScore HighScoreNo
##    <chr>  <int> <int>   <int>   <int>   <int> <int>     <dbl> <lgl>
##  1 MM La… 2010  2020     104      98      21  2788       133 TRUE
##  2 AJ He… 2010  2020     112      97      16  2060       148 TRUE
##  3 BL Mo… 2016  2020      52      49      11  1452       117 TRUE
##  4 EJ Vi… 2009  2018      62      58      10  1369        90 TRUE
##  5 AJ Bl… 2005  2017      95      81      19  1314        61 FALSE
##  6 EA Pe… 2008  2020     120      72      29  1218        60 TRUE
##  7 JE Du… 2009  2015      64      55      10   941        68 TRUE
##  8 LJ Po… 2006  2012      40      40       2   784        61 FALSE
##  9 S Nit… 2005  2011      36      35       2   776        56 FALSE
## 10 LC St… 2005  2013      54      50      14   769        52 FALSE
## # … with 43 more rows, and 8 more variables: Average <dbl>, BallsFaced <
```

fitzRoy

```
# From CRAN
# install.packages("fitzRoy")
# or from GitHub
# remotes::install_github("jimmyday12/fitzRoy")
library(fitzRoy)
aflw <- get_aflw_match_data(start_year = 2020)
```

- ⚽ Gets statistics from https://womens.afl and https://afltables.com
- ⚽ Developed by James Day, Robert Nguyen, Matthew Erbs, Oscar Lane, Jason Zivkovic
- ⚽ Combination of scraping for men's data, and reading protected JSON data for women's by requesting a permission token

Working with data in elite sport

Dr Jacquie Tran | @jacquietran | 15 May 2019