Lecture 6: Data Acquisition

James Sears*

AFRE 891/991 SS 25

Michigan State University

*Parts of these slides are adapted from **“Data Science for Economists”** by Grant McDermott and **“Advanced Data Analytics”** by Nick Hagerty.

Table of Contents

Prologue

Prologue

Let's load in the packages that we'll need for today.

```
pacman::p_load(lubridate, rvest, stringr, tidyverse)
```

Data Acquisition

In order to wrangle, clean, or visualize data, we first need… data.

Where Do Data Come from?

There is a whole spectrum, from DIY to plug-and-chug.

1. Pre-cleaned datasets posted on secondary repositories

- General and journal replication repositories
  - e.g. **Harvard Dataverse**
- Community-based repositories
  - e.g. **Kaggle**
- Public GitHub repositories
  - e.g. **Johns Hopkins COVID-19 caseloads, deaths, and vaccination data**

Where Do Data Come from?

There is a whole spectrum, from DIY to plug-and-chug.

2. Open data libraries

- US Gov't: **Data.gov**
- Institutions, non-profits, and thinktanks
  - e.g. **World Bank Open Data**, **Pew Research**, **NBER Public Use Data**, **Economic Policy Institute**, **AEA Data Sources**
- Tech and other sites
  - e.g. **data.world**, **Our World in Data**, **Stanford Large Network Datasets**
- **Map of 2600+ Open Data Portals Worldwide**

Where Do Data Come from?

There is a whole spectrum, from DIY to plug-and-chug.

3. Websites of primary data providers

- Government agencies; some private companies and NGOs; scientific researchers.
    - i.e. **EPA + USGS Water Quality Portal**

4. Programmatic data access through Application Program Interfaces (APIs)

- We'll talk more about these next lecture

Where Do Data Come from?

There is a whole spectrum, from DIY to plug-and-chug.

5. Liberate previously inaccessible data

- Build relationships with people in government or the private sector.
- Find the right person, cold-email them and ask politely.
    - i.e. my master's thesis!
- File a Freedom of Information Act (FOIA) request (a last resort; very aggressive move).

Where Do Data Come from?

6. Compile data yourself

- Assemble systematic information from many disparate sources.
- E.g. historical archives, websites, PDF reports.

7. Collect your own primary data

- Run surveys or experiments.

Where to Look for Data?

There is **no "one-stop shop"**. Where to look entirely depends on your topic.

For economics research:

- **Search the literature:** Find papers related to your topic and check the Data section.
    - Good for learning the "standard" sources for common things (e.g., weather data).
- **Find your own data** that the literature hasn't used yet.
    - It's hard to find a novel use for an already widely-used dataset.
    - Cross-disciplinary arbitrage
- **Combine data in new ways:** most new projects will require joining data from 2+ sources.
    - E.g. state-level policy changes + household-level outcome data.

Where to Look for Data?

There is **no "one-stop shop"**. Where to look entirely depends on your topic.

A few useful starting points:

- **"How to Find Data: Tips for Finding Data"** (Davidson College Library).
- **"Data Sets for Quantitative Research"** (University of Missouri Libraries).
- **Google Dataset Search**, **AWS Data Exchange**

The majority of data sources described above have the data easily accessible once found.

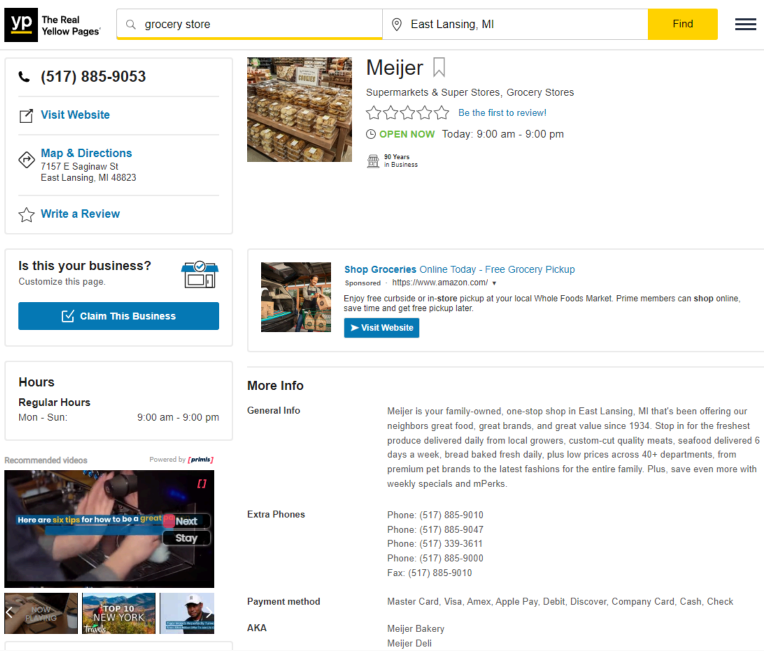
If the data aren't already machine readable, then we can take advantage of...

Intro to Web Scraping

Web Scraping

Web scraping is the process of **extracting semi-structured web data** and **converting into a structured dataset**

- Useful when information is already online but not available in a handy format.



(517) 885-9053

Map & Directions
7157 E Saginaw St
East Lansing, MI 48823

Meijer

Supermarkets & Super Stores, Grocery Stores

☆☆☆☆☆ Be the first to review!

OPEN NOW Today: 9:00 am - 9:00 pm

90 Years in Business

Is this your business?
Customize this page.

Shop Groceries Online Today - Free Grocery Pickup
Sponsored · https://www.amazon.com/ ▾
Enjoy free curbside or in-**store** pickup at your local Whole Foods Market. Prime members can **shop** online, save time and get free pickup later.

Hours
Regular Hours
Mon - Sun: 9:00 am - 9:00 pm

More Info

General Info	Meijer is your family-owned, one-stop shop in East Lansing, MI that's been offering our neighbors great food, great brands, and great value since 1934. Stop in for the freshest produce delivered daily from local growers, custom-cut quality meats, seafood delivered 6 days a week, bread baked fresh daily, plus low prices across 40+ departments, from premium pet brands to the latest fashions for the entire family. Plus, save even more with weekly specials and mPerks.
Extra Phones	Phone: (517) 885-9010 Phone: (517) 885-9047 Phone: (517) 339-3611 Phone: (517) 885-9000 Fax: (517) 885-9010
Payment method	Master Card, Visa, Amex, Apple Pay, Debit, Discover, Company Card, Cash, Check
AKA	Meijer Bakery Meijer Deli

Structure of Webpages

Webpages are largely made out of **two types of files** that we have to parse:

HTML

- **H**yper**T**ext **M**arkup **L**anguage is a **markup language**, Like Markdown
- It specifies the **structure** of a webpage.

CSS

- **C**ascading **S**tyle **S**heets is a language for **formatting the appearance of a webpage**
    - CSS **properties** specify **how**to format: what font, what color, how wide
    - CSS **selectors** specify **what** to format: which structural elements get what rule

How Websites Render Content

It's also worth realising that there are **two ways** that web content gets rendered in a browser:

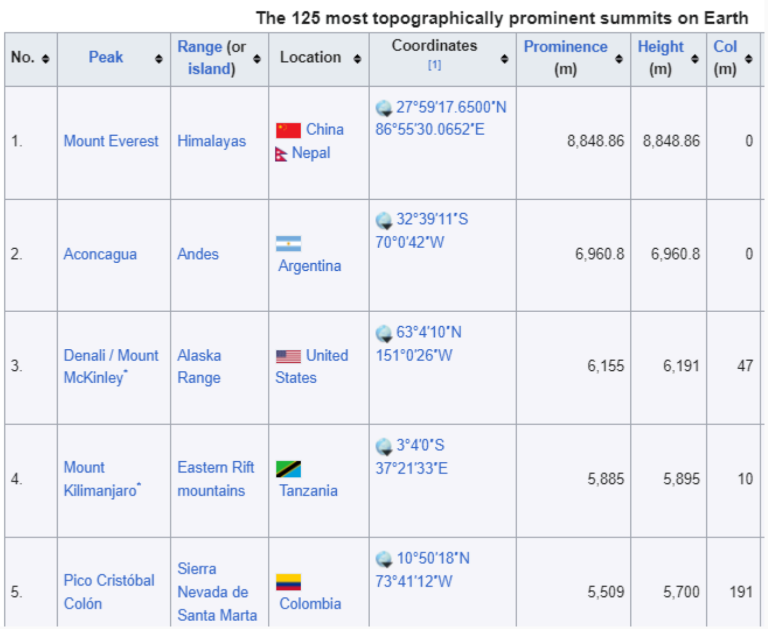
1. **Server-side (back-end)**
2. **Client-side (front-end)**

You can read **here** for more details (including example scripts), but for our purposes the essential features are as follows...

Server-Side Content Rendering

The scripts that build a **server-side** website aren't run on our computer

- Rather, the script that "builds" the site is **run on a host server**
    - All information is directly embedded in the webpage's HTML
- e.g. Wikipedia tables are already populated with all the info we see in our browser



The 125 most topographically prominent summits on Earth

No.	Peak	Range (or island)	Location	Coordinates [1]	Prominence (m)	Height (m)	Col (m)
1.	Mount Everest	Himalayas	China Nepal	27°59′17.6500″N 86°55′30.0652″E	8,848.86	8,848.86	0
2.	Aconcagua	Andes	Argentina	32°39′11″S 70°0′42″W	6,960.8	6,960.8	0
3.	Denali / Mount McKinley*	Alaska Range	United States	63°4′10″N 151°0′26″W	6,155	6,191	47
4.	Mount Kilimanjaro*	Eastern Rift mountains	Tanzania	3°4′0″S 37°21′33″E	5,885	5,895	10
5.	Pico Cristóbal Colón	Sierra Nevada de Santa Marta	Colombia	10°50′18″N 73°41′12″W	5,509	5,700	191

Server-Side Content Rendering

The scripts that build a **server-side** website aren't run on our computer

- Rather, the script that "builds" the site is **run on a host server**
    - All information is .hi-blue[directly embedded] in the webpage's HTML
- e.g. Wikipedia tables are already populated with all the info we see in our browser

- **Webscraping process:** finding the correct selectors (CSS or Xpath), iterating through (dynamic) webpages (e.g. "Next page" and "Show more" tabs)
- **Key concepts:** CSS, Xpath, HTML

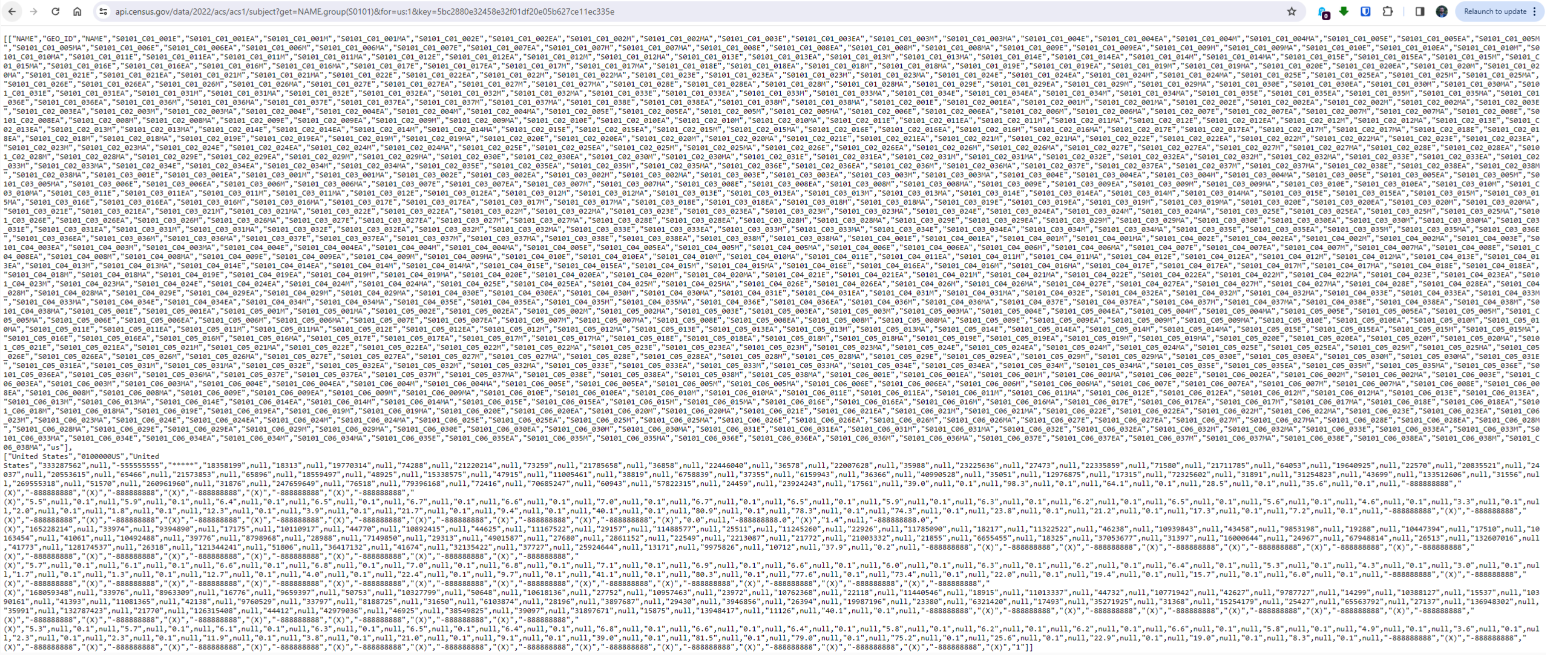
Client-Side Content Rendering

The scripts that build a **client-side** website aren't run on our computer

- Website contains an empty template of HTML and CSS
  - May contain a "skeleton" table without any values
- When we visit the page URL, our browser sends a **request** to the host server
- If everything is okay (e.g. our request is valid), then the server sends a **response** script, which our browser executes and uses to populate the HTML template with the specific information that we want.

Client-Side Content Rendering

- If everything is okay (e.g. our request is valid), then the server sends a **response** script, which our browser executes and uses to populate the HTML template with the specific information that we want.



api.census.gov/data/2022/acs/acs1/subject?get=NAME,group(S0101)&for=us:1&key=5bc2880e32458e32f01df20e05b627ce11ec335e

```
[["NAME","GEO_ID","NAME","S0101_C01_001E","S0101_C01_001EA","S0101_C01_001M","S0101_C01_001MA","S0101_C01_002E","S0101_C01_002EA","S0101_C01_002M","S0101_C01_002MA","S0101_C01_003E","S0101_C01_003EA","S0101_C01_003M","S0101_C01_003MA","S0101_C01_004E","S0101_C01_004EA","S0101_C01_004M","S0101_C01_004MA","S0101_C01_005E","S0101_C01_005EA","S0101_C01_005M
","S0101_C01_005MA","S0101_C01_006E","S0101_C01_006EA","S0101_C01_006M","S0101_C01_006MA","S0101_C01_007E","S0101_C01_007EA","S0101_C01_007M","S0101_C01_007MA","S0101_C01_008E","S0101_C01_008EA","S0101_C01_008M","S0101_C01_008MA","S0101_C01_009E","S0101_C01_009EA","S0101_C01_009M","S0101_C01_009MA","S0101_C01_010E","S0101_C01_010EA","S0101_C01_010M","S0
101_C01_010MA","S0101_C01_011E","S0101_C01_011EA","S0101_C01_011M","S0101_C01_011MA","S0101_C01_012E","S0101_C01_012EA","S0101_C01_012M","S0101_C01_012MA","S0101_C01_013E","S0101_C01_013EA","S0101_C01_013M","S0101_C01_013MA","S0101_C01_014E","S0101_C01_014EA","S0101_C01_014M","S0101_C01_014MA","S0101_C01_015E","S0101_C01_015EA","S0101_C01_015M","S0101_C
01_015MA","S0101_C01_016E","S0101_C01_016EA","S0101_C01_016M","S0101_C01_016MA","S0101_C01_017E","S0101_C01_017EA","S0101_C01_017M","S0101_C01_017MA","S0101_C01_018E","S0101_C01_018EA","S0101_C01_018M","S0101_C01_018MA","S0101_C01_019E","S0101_C01_019EA","S0101_C01_019M","S0101_C01_019MA","S0101_C01_020E","S0101_C01_020EA","S0101_C01_020M","S0101_C01_02
0MA","S0101_C01_021E","S0101_C01_021EA","S0101_C01_021M","S0101_C01_021MA","S0101_C01_022E","S0101_C01_022EA","S0101_C01_022M","S0101_C01_022MA","S0101_C01_023E","S0101_C01_023EA","S0101_C01_023M","S0101_C01_023MA","S0101_C01_024E","S0101_C01_024EA","S0101_C01_024M","S0101_C01_024MA","S0101_C01_025E","S0101_C01_025EA","S0101_C01_025M","S0101_C01_025MA",
"S0101_C01_026E","S0101_C01_026EA","S0101_C01_026M","S0101_C01_026MA","S0101_C01_027E","S0101_C01_027EA","S0101_C01_027M","S0101_C01_027MA","S0101_C01_028E","S0101_C01_028EA","S0101_C01_028M","S0101_C01_028MA","S0101_C01_029E","S0101_C01_029EA","S0101_C01_029M","S0101_C01_029MA","S0101_C01_030E","S0101_C01_030EA","S0101_C01_030M","S0101_C01_030MA","S010
1_C01_031E","S0101_C01_031EA","S0101_C01_031M","S0101_C01_031MA","S0101_C01_032E","S0101_C01_032EA","S0101_C01_032M","S0101_C01_032MA","S0101_C01_033E","S0101_C01_033EA","S0101_C01_033M","S0101_C01_033MA","S0101_C01_034E","S0101_C01_034EA","S0101_C01_034M","S0101_C01_034MA","S0101_C01_035E","S0101_C01_035EA","S0101_C01_035M","S0101_C01_035MA","S0101_C01
_036E","S0101_C01_036EA","S0101_C01_036M","S0101_C01_036MA","S0101_C01_037E","S0101_C01_037EA","S0101_C01_037M","S0101_C01_037MA","S0101_C01_038E","S0101_C01_038EA","S0101_C01_038M","S0101_C01_038MA","S0101_C02_001E","S0101_C02_001EA","S0101_C02_001M","S0101_C02_001MA","S0101_C02_002E","S0101_C02_002EA","S0101_C02_002M","S0101_C02_002MA","S0101_C02_003E
","S0101_C02_003EA","S0101_C02_003M","S0101_C02_003MA","S0101_C02_004E","S0101_C02_004EA","S0101_C02_004M","S0101_C02_004MA","S0101_C02_005E","S0101_C02_005EA","S0101_C02_005M","S0101_C02_005MA","S0101_C02_006E","S0101_C02_006EA","S0101_C02_006M","S0101_C02_006MA","S0101_C02_007E","S0101_C02_007EA","S0101_C02_007M","S0101_C02_007MA","S0101_C02_008E","S0
101_C02_008EA","S0101_C02_008M","S0101_C02_008MA","S0101_C02_009E","S0101_C02_009EA","S0101_C02_009M","S0101_C02_009MA","S0101_C02_010E","S0101_C02_010EA","S0101_C02_010M","S0101_C02_010MA","S0101_C02_011E","S0101_C02_011EA","S0101_C02_011M","S0101_C02_011MA","S0101_C02_012E","S0101_C02_012EA","S0101_C02_012M","S0101_C02_012MA","S0101_C02_013E","S0101_C
02_013EA","S0101_C02_013M","S0101_C02_013MA","S0101_C02_014E","S0101_C02_014EA","S0101_C02_014M","S0101_C02_014MA","S0101_C02_015E","S0101_C02_015EA","S0101_C02_015M","S0101_C02_015MA","S0101_C02_016E","S0101_C02_016EA","S0101_C02_016M","S0101_C02_016MA","S0101_C02_017E","S0101_C02_017EA","S0101_C02_017M","S0101_C02_017MA","S0101_C02_018E","S0101_C02_01
8EA","S0101_C02_018M","S0101_C02_018MA","S0101_C02_019E","S0101_C02_019EA","S0101_C02_019M","S0101_C02_019MA","S0101_C02_020E","S0101_C02_020EA","S0101_C02_020M","S0101_C02_020MA","S0101_C02_021E","S0101_C02_021EA","S0101_C02_021M","S0101_C02_021MA","S0101_C02_022E","S0101_C02_022EA","S0101_C02_022M","S0101_C02_022MA","S0101_C02_023E","S0101_C02_023EA",
"S0101_C02_023M","S0101_C02_023MA","S0101_C02_024E","S0101_C02_024EA","S0101_C02_024M","S0101_C02_024MA","S0101_C02_025E","S0101_C02_025EA","S0101_C02_025M","S0101_C02_025MA","S0101_C02_026E","S0101_C02_026EA","S0101_C02_026M","S0101_C02_026MA","S0101_C02_027E","S0101_C02_027EA","S0101_C02_027M","S0101_C02_027MA","S0101_C02_028E","S0101_C02_028EA","S010
1_C02_028M","S0101_C02_028MA","S0101_C02_029E","S0101_C02_029EA","S0101_C02_029M","S0101_C02_029MA","S0101_C02_030E","S0101_C02_030EA","S0101_C02_030M","S0101_C02_030MA","S0101_C02_031E","S0101_C02_031EA","S0101_C02_031M","S0101_C02_031MA","S0101_C02_032E","S0101_C02_032EA","S0101_C02_032M","S0101_C02_032MA","S0101_C02_033E","S0101_C02_033EA","S0101_C02
_033M","S0101_C02_033MA","S0101_C02_034E","S0101_C02_034EA","S0101_C02_034M","S0101_C02_034MA","S0101_C02_035E","S0101_C02_035EA","S0101_C02_035M","S0101_C02_035MA","S0101_C02_036E","S0101_C02_036EA","S0101_C02_036M","S0101_C02_036MA","S0101_C02_037E","S0101_C02_037EA","S0101_C02_037M","S0101_C02_037MA","S0101_C02_038E","S0101_C02_038EA","S0101_C02_038M
","S0101_C02_038MA","S0101_C03_001E","S0101_C03_001EA","S0101_C03_001M","S0101_C03_001MA","S0101_C03_002E","S0101_C03_002EA","S0101_C03_002M","S0101_C03_002MA","S0101_C03_003E","S0101_C03_003EA","S0101_C03_003M","S0101_C03_003MA","S0101_C03_004E","S0101_C03_004EA","S0101_C03_004M","S0101_C03_004MA","S0101_C03_005E","S0101_C03_005EA","S0101_C03_005M","S0
101_C03_005MA","S0101_C03_006E","S0101_C03_006EA","S0101_C03_006M","S0101_C03_006MA","S0101_C03_007E","S0101_C03_007EA","S0101_C03_007M","S0101_C03_007MA","S0101_C03_008E","S0101_C03_008EA","S0101_C03_008M","S0101_C03_008MA","S0101_C03_009E","S0101_C03_009EA","S0101_C03_009M","S0101_C03_009MA","S0101_C03_010E","S0101_C03_010EA","S0101_C03_010M","S0101_C
03_010MA","S0101_C03_011E","S0101_C03_011EA","S0101_C03_011M","S0101_C03_011MA","S0101_C03_012E","S0101_C03_012EA","S0101_C03_012M","S0101_C03_012MA","S0101_C03_013E","S0101_C03_013EA","S0101_C03_013M","S0101_C03_013MA","S0101_C03_014E","S0101_C03_014EA","S0101_C03_014M","S0101_C03_014MA","S0101_C03_015E","S0101_C03_015EA","S0101_C03_015M","S0101_C03_01
5MA","S0101_C03_016E","S0101_C03_016EA","S0101_C03_016M","S0101_C03_016MA","S0101_C03_017E","S0101_C03_017EA","S0101_C03_017M","S0101_C03_017MA","S0101_C03_018E","S0101_C03_018EA","S0101_C03_018M","S0101_C03_018MA","S0101_C03_019E","S0101_C03_019EA","S0101_C03_019M","S0101_C03_019MA","S0101_C03_020E","S0101_C03_020EA","S0101_C03_020M","S0101_C03_020MA",
"S0101_C03_021E","S0101_C03_021EA","S0101_C03_021M","S0101_C03_021MA","S0101_C03_022E","S0101_C03_022EA","S0101_C03_022M","S0101_C03_022MA","S0101_C03_023E","S0101_C03_023EA","S0101_C03_023M","S0101_C03_023MA","S0101_C03_024E","S0101_C03_024EA","S0101_C03_024M","S0101_C03_024MA","S0101_C03_025E","S0101_C03_025EA","S0101_C03_025M","S0101_C03_025MA","S010
1_C03_026E","S0101_C03_026EA","S0101_C03_026M","S0101_C03_026MA","S0101_C03_027E","S0101_C03_027EA","S0101_C03_027M","S0101_C03_027MA","S0101_C03_028E","S0101_C03_028EA","S0101_C03_028M","S0101_C03_028MA","S0101_C03_029E","S0101_C03_029EA","S0101_C03_029M","S0101_C03_029MA","S0101_C03_030E","S0101_C03_030EA","S0101_C03_030M","S0101_C03_030MA","S0101_C03
_031E","S0101_C03_031EA","S0101_C03_031M","S0101_C03_031MA","S0101_C03_032E","S0101_C03_032EA","S0101_C03_032M","S0101_C03_032MA","S0101_C03_033E","S0101_C03_033EA","S0101_C03_033M","S0101_C03_033MA","S0101_C03_034E","S0101_C03_034EA","S0101_C03_034M","S0101_C03_034MA","S0101_C03_035E","S0101_C03_035EA","S0101_C03_035M","S0101_C03_035MA","S0101_C03_036E
","S0101_C03_036EA","S0101_C03_036M","S0101_C03_036MA","S0101_C03_037E","S0101_C03_037EA","S0101_C03_037M","S0101_C03_037MA","S0101_C03_038E","S0101_C03_038EA","S0101_C03_038M","S0101_C03_038MA","S0101_C04_001E","S0101_C04_001EA","S0101_C04_001M","S0101_C04_001MA","S0101_C04_002E","S0101_C04_002EA","S0101_C04_002M","S0101_C04_002MA","S0101_C04_003E","S0
101_C04_003EA","S0101_C04_003M","S0101_C04_003MA","S0101_C04_004E","S0101_C04_004EA","S0101_C04_004M","S0101_C04_004MA","S0101_C04_005E","S0101_C04_005EA","S0101_C04_005M","S0101_C04_005MA","S0101_C04_006E","S0101_C04_006EA","S0101_C04_006M","S0101_C04_006MA","S0101_C04_007E","S0101_C04_007EA","S0101_C04_007M","S0101_C04_007MA","S0101_C04_008E","S0101_C
04_008EA","S0101_C04_008M","S0101_C04_008MA","S0101_C04_009E","S0101_C04_009EA","S0101_C04_009M","S0101_C04_009MA","S0101_C04_010E","S0101_C04_010EA","S0101_C04_010M","S0101_C04_010MA","S0101_C04_011E","S0101_C04_011EA","S0101_C04_011M","S0101_C04_011MA","S0101_C04_012E","S0101_C04_012EA","S0101_C04_012M","S0101_C04_012MA","S0101_C04_013E","S0101_C04_01
3EA","S0101_C04_013M","S0101_C04_013MA","S0101_C04_014E","S0101_C04_014EA","S0101_C04_014M","S0101_C04_014MA","S0101_C04_015E","S0101_C04_015EA","S0101_C04_015M","S0101_C04_015MA","S0101_C04_016E","S0101_C04_016EA","S0101_C04_016M","S0101_C04_016MA","S0101_C04_017E","S0101_C04_017EA","S0101_C04_017M","S0101_C04_017MA","S0101_C04_018E","S0101_C04_018EA",
"S0101_C04_018M","S0101_C04_018MA","S0101_C04_019E","S0101_C04_019EA","S0101_C04_019M","S0101_C04_019MA","S0101_C04_020E","S0101_C04_020EA","S0101_C04_020M","S0101_C04_020MA","S0101_C04_021E","S0101_C04_021EA","S0101_C04_021M","S0101_C04_021MA","S0101_C04_022E","S0101_C04_022EA","S0101_C04_022M","S0101_C04_022MA","S0101_C04_023E","S0101_C04_023EA","S010
1_C04_023M","S0101_C04_023MA","S0101_C04_024E","S0101_C04_024EA","S0101_C04_024M","S0101_C04_024MA","S0101_C04_025E","S0101_C04_025EA","S0101_C04_025M","S0101_C04_025MA","S0101_C04_026E","S0101_C04_026EA","S0101_C04_026M","S0101_C04_026MA","S0101_C04_027E","S0101_C04_027EA","S0101_C04_027M","S0101_C04_027MA","S0101_C04_028E","S0101_C04_028EA","S0101_C04
_028M","S0101_C04_028MA","S0101_C04_029E","S0101_C04_029EA","S0101_C04_029M","S0101_C04_029MA","S0101_C04_030E","S0101_C04_030EA","S0101_C04_030M","S0101_C04_030MA","S0101_C04_031E","S0101_C04_031EA","S0101_C04_031M","S0101_C04_031MA","S0101_C04_032E","S0101_C04_032EA","S0101_C04_032M","S0101_C04_032MA","S0101_C04_033E","S0101_C04_033EA","S0101_C04_033M
","S0101_C04_033MA","S0101_C04_034E","S0101_C04_034EA","S0101_C04_034M","S0101_C04_034MA","S0101_C04_035E","S0101_C04_035EA","S0101_C04_035M","S0101_C04_035MA","S0101_C04_036E","S0101_C04_036EA","S0101_C04_036M","S0101_C04_036MA","S0101_C04_037E","S0101_C04_037EA","S0101_C04_037M","S0101_C04_037MA","S0101_C04_038E","S0101_C04_038EA","S0101_C04_038M","S0
101_C04_038MA","S0101_C05_001E","S0101_C05_001EA","S0101_C05_001M","S0101_C05_001MA","S0101_C05_002E","S0101_C05_002EA","S0101_C05_002M","S0101_C05_002MA","S0101_C05_003E","S0101_C05_003EA","S0101_C05_003M","S0101_C05_003MA","S0101_C05_004E","S0101_C05_004EA","S0101_C05_004M","S0101_C05_004MA","S0101_C05_005E","S0101_C05_005EA","S0101_C05_005M","S0101_C
05_005MA","S0101_C05_006E","S0101_C05_006EA","S0101_C05_006M","S0101_C05_006MA","S0101_C05_007E","S0101_C05_007EA","S0101_C05_007M","S0101_C05_007MA","S0101_C05_008E","S0101_C05_008EA","S0101_C05_008M","S0101_C05_008MA","S0101_C05_009E","S0101_C05_009EA","S0101_C05_009M","S0101_C05_009MA","S0101_C05_010E","S0101_C05_010EA","S0101_C05_010M","S0101_C05_01
0MA","S0101_C05_011E","S0101_C05_011EA","S0101_C05_011M","S0101_C05_011MA","S0101_C05_012E","S0101_C05_012EA","S0101_C05_012M","S0101_C05_012MA","S0101_C05_013E","S0101_C05_013EA","S0101_C05_013M","S0101_C05_013MA","S0101_C05_014E","S0101_C05_014EA","S0101_C05_014M","S0101_C05_014MA","S0101_C05_015E","S0101_C05_015EA","S0101_C05_015M","S0101_C05_015MA",
"S0101_C05_016E","S0101_C05_016EA","S0101_C05_016M","S0101_C05_016MA","S0101_C05_017E","S0101_C05_017EA","S0101_C05_017M","S0101_C05_017MA","S0101_C05_018E","S0101_C05_018EA","S0101_C05_018M","S0101_C05_018MA","S0101_C05_019E","S0101_C05_019EA","S0101_C05_019M","S0101_C05_019MA","S0101_C05_020E","S0101_C05_020EA","S0101_C05_020M","S0101_C05_020MA","S010
1_C05_021E","S0101_C05_021EA","S0101_C05_021M","S0101_C05_021MA","S0101_C05_022E","S0101_C05_022EA","S0101_C05_022M","S0101_C05_022MA","S0101_C05_023E","S0101_C05_023EA","S0101_C05_023M","S0101_C05_023MA","S0101_C05_024E","S0101_C05_024EA","S0101_C05_024M","S0101_C05_024MA","S0101_C05_025E","S0101_C05_025EA","S0101_C05_025M","S0101_C05_025MA","S0101_C05
_026E","S0101_C05_026EA","S0101_C05_026M","S0101_C05_026MA","S0101_C05_027E","S0101_C05_027EA","S0101_C05_027M","S0101_C05_027MA","S0101_C05_028E","S0101_C05_028EA","S0101_C05_028M","S0101_C05_028MA","S0101_C05_029E","S0101_C05_029EA","S0101_C05_029M","S0101_C05_029MA","S0101_C05_030E","S0101_C05_030EA","S0101_C05_030M","S0101_C05_030MA","S0101_C05_031E
","S0101_C05_031EA","S0101_C05_031M","S0101_C05_031MA","S0101_C05_032E","S0101_C05_032EA","S0101_C05_032M","S0101_C05_032MA","S0101_C05_033E","S0101_C05_033EA","S0101_C05_033M","S0101_C05_033MA","S0101_C05_034E","S0101_C05_034EA","S0101_C05_034M","S0101_C05_034MA","S0101_C05_035E","S0101_C05_035EA","S0101_C05_035M","S0101_C05_035MA","S0101_C05_036E","S0
101_C05_036EA","S0101_C05_036M","S0101_C05_036MA","S0101_C05_037E","S0101_C05_037EA","S0101_C05_037M","S0101_C05_037MA","S0101_C05_038E","S0101_C05_038EA","S0101_C05_038M","S0101_C05_038MA","S0101_C06_001E","S0101_C06_001EA","S0101_C06_001M","S0101_C06_001MA","S0101_C06_002E","S0101_C06_002EA","S0101_C06_002M","S0101_C06_002MA","S0101_C06_003E","S0101_C
06_003EA","S0101_C06_003M","S0101_C06_003MA","S0101_C06_004E","S0101_C06_004EA","S0101_C06_004M","S0101_C06_004MA","S0101_C06_005E","S0101_C06_005EA","S0101_C06_005M","S0101_C06_005MA","S0101_C06_006E","S0101_C06_006EA","S0101_C06_006M","S0101_C06_006MA","S0101_C06_007E","S0101_C06_007EA","S0101_C06_007M","S0101_C06_007MA","S0101_C06_008E","S0101_C06_00
8EA","S0101_C06_008M","S0101_C06_008MA","S0101_C06_009E","S0101_C06_009EA","S0101_C06_009M","S0101_C06_009MA","S0101_C06_010E","S0101_C06_010EA","S0101_C06_010M","S0101_C06_010MA","S0101_C06_011E","S0101_C06_011EA","S0101_C06_011M","S0101_C06_011MA","S0101_C06_012E","S0101_C06_012EA","S0101_C06_012M","S0101_C06_012MA","S0101_C06_013E","S0101_C06_013EA",
"S0101_C06_013M","S0101_C06_013MA","S0101_C06_014E","S0101_C06_014EA","S0101_C06_014M","S0101_C06_014MA","S0101_C06_015E","S0101_C06_015EA","S0101_C06_015M","S0101_C06_015MA","S0101_C06_016E","S0101_C06_016EA","S0101_C06_016M","S0101_C06_016MA","S0101_C06_017E","S0101_C06_017EA","S0101_C06_017M","S0101_C06_017MA","S0101_C06_018E","S0101_C06_018EA","S010
1_C06_018M","S0101_C06_018MA","S0101_C06_019E","S0101_C06_019EA","S0101_C06_019M","S0101_C06_019MA","S0101_C06_020E","S0101_C06_020EA","S0101_C06_020M","S0101_C06_020MA","S0101_C06_021E","S0101_C06_021EA","S0101_C06_021M","S0101_C06_021MA","S0101_C06_022E","S0101_C06_022EA","S0101_C06_022M","S0101_C06_022MA","S0101_C06_023E","S0101_C06_023EA","S0101_C06
_023M","S0101_C06_023MA","S0101_C06_024E","S0101_C06_024EA","S0101_C06_024M","S0101_C06_024MA","S0101_C06_025E","S0101_C06_025EA","S0101_C06_025M","S0101_C06_025MA","S0101_C06_026E","S0101_C06_026EA","S0101_C06_026M","S0101_C06_026MA","S0101_C06_027E","S0101_C06_027EA","S0101_C06_027M","S0101_C06_027MA","S0101_C06_028E","S0101_C06_028EA","S0101_C06_028M
","S0101_C06_028MA","S0101_C06_029E","S0101_C06_029EA","S0101_C06_029M","S0101_C06_029MA","S0101_C06_030E","S0101_C06_030EA","S0101_C06_030M","S0101_C06_030MA","S0101_C06_031E","S0101_C06_031EA","S0101_C06_031M","S0101_C06_031MA","S0101_C06_032E","S0101_C06_032EA","S0101_C06_032M","S0101_C06_032MA","S0101_C06_033E","S0101_C06_033EA","S0101_C06_033M","S0
101_C06_033MA","S0101_C06_034E","S0101_C06_034EA","S0101_C06_034M","S0101_C06_034MA","S0101_C06_035E","S0101_C06_035EA","S0101_C06_035M","S0101_C06_035MA","S0101_C06_036E","S0101_C06_036EA","S0101_C06_036M","S0101_C06_036MA","S0101_C06_037E","S0101_C06_037EA","S0101_C06_037M","S0101_C06_037MA","S0101_C06_038E","S0101_C06_038EA","S0101_C06_038M","S0101_C
06_038MA","us"],
["United States","0100000US","United
States","333287562",null,"-555555555","*****","18358199",null,"18313",null,"19770314",null,"74288",null,"21220214",null,"73259",null,"21785658",null,"36858",null,"22446040",null,"36578",null,"22007628",null,"35988",null,"23225636",null,"27473",null,"22335859",null,"71580",null,"21711785",null,"64053",null,"19640925",null,"22570",null,"20835521",null,"24
037",null,"20553615",null,"65466",null,"21573853",null,"65896",null,"18559497",null,"48925",null,"15338575",null,"47915",null,"11005461",null,"38819",null,"6758839",null,"37355",null,"6159943",null,"36366",null,"40990528",null,"35051",null,"12976875",null,"17315",null,"72325602",null,"31891",null,"31254823",null,"43699",null,"133512606",null,"31556",nul
l,"269555318",null,"51570",null,"260961960",null,"31876",null,"247659649",null,"76518",null,"79396168",null,"72416",null,"70685247",null,"60943",null,"57822315",null,"24459",null,"23924243",null,"17561",null,"39.0",null,"0.1",null,"98.3",null,"0.1",null,"64.1",null,"0.1",null,"28.5",null,"0.1",null,"35.6",null,"0.1",null,"-888888888","
(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","
(X)","5.5",null,"0.1",null,"5.9",null,"0.1",null,"6.4",null,"0.1",null,"6.5",null,"0.1",null,"6.7",null,"0.1",null,"6.6",null,"0.1",null,"7.0",null,"0.1",null,"6.7",null,"0.1",null,"6.5",null,"0.1",null,"5.9",null,"0.1",null,"6.3",null,"0.1",null,"6.2",null,"0.1",null,"6.5",null,"0.1",null,"5.6",null,"0.1",null,"4.6",null,"0.1",null,"3.3",null,"0.1",nul
l,"2.0",null,"0.1",null,"1.8",null,"0.1",null,"12.3",null,"0.1",null,"3.9",null,"0.1",null,"21.7",null,"0.1",null,"9.4",null,"0.1",null,"40.1",null,"0.1",null,"80.9",null,"0.1",null,"78.3",null,"0.1",null,"74.3",null,"0.1",null,"23.8",null,"0.1",null,"21.2",null,"0.1",null,"17.3",null,"0.1",null,"7.2",null,"0.1",null,"-888888888","(X)","-888888888","
(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","0.0",null,"-888888888.0","(X)","1.4",null,"-888888888.0","
(X)","165228214",null,"33974",null,"9394890",null,"17175",null,"10110917",null,"44770",null,"10892415",null,"44625",null,"11167522",null,"29157",null,"11488577",null,"25511",null,"11245260",null,"22926",null,"11785090",null,"18217",null,"11322522",null,"46238",null,"10939843",null,"43458",null,"9853198",null,"19288",null,"10447394",null,"17510",null,"10
163454",null,"41061",null,"10492488",null,"39776",null,"8798968",null,"28988",null,"7149850",null,"29313",null,"4901587",null,"27680",null,"2861152",null,"22549",null,"2213087",null,"21772",null,"21003332",null,"21855",null,"6655455",null,"18325",null,"37053677",null,"31397",null,"16000644",null,"24967",null,"67948814",null,"26513",null,"132607016",null
,"41773",null,"128174537",null,"26318",null,"121344241",null,"51806",null,"36417132",null,"41674",null,"32135422",null,"37727",null,"25924644",null,"13171",null,"9975826",null,"10712",null,"37.9",null,"0.2",null,"-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","
(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","
(X)","5.7",null,"0.1",null,"6.1",null,"0.1",null,"6.6",null,"0.1",null,"6.8",null,"0.1",null,"7.0",null,"0.1",null,"6.8",null,"0.1",null,"7.1",null,"0.1",null,"6.9",null,"0.1",null,"6.6",null,"0.1",null,"6.0",null,"0.1",null,"6.3",null,"0.1",null,"6.2",null,"0.1",null,"6.4",null,"0.1",null,"5.3",null,"0.1",null,"4.3",null,"0.1",null,"3.0",null,"0.1",nul
l,"1.7",null,"0.1",null,"1.3",null,"0.1",null,"12.7",null,"0.1",null,"4.0",null,"0.1",null,"22.4",null,"0.1",null,"9.7",null,"0.1",null,"41.1",null,"0.1",null,"80.3",null,"0.1",null,"77.6",null,"0.1",null,"73.4",null,"0.1",null,"22.0",null,"0.1",null,"19.4",null,"0.1",null,"15.7",null,"0.1",null,"6.0",null,"0.1",null,"-888888888","(X)","-888888888","
(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","
(X)","168059348",null,"33976",null,"8963309",null,"16776",null,"9659397",null,"50753",null,"10327799",null,"50648",null,"10618136",null,"27752",null,"10957463",null,"23972",null,"10762368",null,"22118",null,"11440546",null,"18915",null,"11013337",null,"44732",null,"10771942",null,"42627",null,"9787727",null,"14299",null,"10388127",null,"15537",null,"103
90161",null,"41393",null,"11081365",null,"42138",null,"9760529",null,"33797",null,"8188725",null,"31650",null,"6103874",null,"28196",null,"3897687",null,"29430",null,"3946856",null,"26394",null,"19987196",null,"23380",null,"6321420",null,"17493",null,"35271925",null,"31368",null,"15254179",null,"25427",null,"65563792",null,"27137",null,"136948302",null,
"35991",null,"132787423",null,"21770",null,"126315408",null,"44412",null,"42979036",null,"46925",null,"38549825",null,"39097",null,"31897671",null,"15875",null,"13948417",null,"11126",null,"40.1",null,"0.1",null,"-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","
(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","
(X)","5.3",null,"0.1",null,"5.7",null,"0.1",null,"6.1",null,"0.1",null,"6.3",null,"0.1",null,"6.5",null,"0.1",null,"6.4",null,"0.1",null,"6.8",null,"0.1",null,"6.6",null,"0.1",null,"6.4",null,"0.1",null,"5.8",null,"0.1",null,"6.2",null,"0.1",null,"6.2",null,"0.1",null,"6.6",null,"0.1",null,"5.8",null,"0.1",null,"4.9",null,"0.1",null,"3.6",null,"0.1",nul
l,"2.3",null,"0.1",null,"2.3",null,"0.1",null,"11.9",null,"0.1",null,"3.8",null,"0.1",null,"21.0",null,"0.1",null,"9.1",null,"0.1",null,"39.0",null,"0.1",null,"81.5",null,"0.1",null,"79.0",null,"0.1",null,"75.2",null,"0.1",null,"25.6",null,"0.1",null,"22.9",null,"0.1",null,"19.0",null,"0.1",null,"8.3",null,"0.1",null,"-888888888","(X)","-888888888","
(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","-888888888","(X)","1"]]
```

Web Scraping

Over the next two lectures we'll cover the main differences between the two approaches and general workflows.

However, I want to forewarn that web scraping typically involves a fair bit of **detective work**, iterating and adjusting steps

- According to the type of data you want
- To match the specifics of a given website

In short, web scraping involves **as much art as it does science.**

The good news, though: **if you can see it, you can scrape it.**

Ethical + Legal Considerations

The last line brings up an important consideration: just because you **can** scrape it, doesn't mean you **should**.

Legality

- In short, it is **currently legal** to scrape data from the web using automated tools as long as the data are **publicly available** (hiQ Labs vs. Linkedin)
  - We'll chat more later on about what this means for "hidden" APIs
- May get blocked due to violating a site's Terms of Service preventing scraping

Ethicality

- Need to consider the impact your scraper will have on the host server
  - Easy to write a function that can overwhelm a website's host with rapid requests
  - We'll return to the "be nice" mantra later on

Scraping Static Websites

Static Scraping: Preliminaries

Today we'll be using **SelectorGadget**, which is a Chrome extension that makes it easy to discover CSS selectors.

- Install the extension directly **here**.

Please note that SelectorGadget is only available for **Chrome**. If you prefer using **Firefox**, then you can try **ScrapeMate**.

Static Scraping: Preliminaries

The primary R package that we'll be using today is **rvest**, a simple webscraping library inspired by Python's **Beautiful Soup**, but with extra tidyverse functionality.

rvest is designed to work with webpages that are built server-side and thus requires knowledge of the relevant CSS selectors...

Which means that now is probably a good time for us to cover what these are in more detail.

CSS Selectors and SelectorGadget

CSS **selectors** are the **"what"** of the display rules. They identify which rules should be applied to which elements.

- E.g. Text elements that are selected as **".h1"** (i.e. top line headers) are usually larger and displayed more prominently than text elements selected as **".h2"** (i.e. sub-headers).

The key point is that if you can **identify the CSS selector(s)** of the content you want, then you can **isolate it from the rest** of the webpage content that you don't want.

CSS Selectors and SelectorGadget

This where SelectorGadget comes in. We'll work through an extended example (with a twist!) below, but I highly recommend looking over this **quick vignette** soon.

Here are two helpful links if you're interested in reading more about **CSS** (i.e Cascading Style Sheets) and **SelectorGadget**

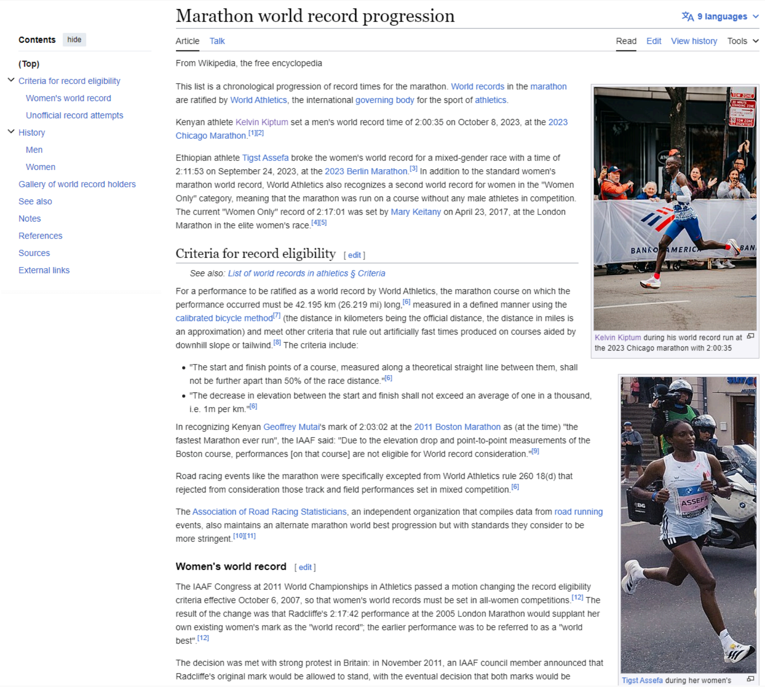
Application 1: Wikipedia

Let's say you watched the U.S. Olympic Marathon Trials last year and now want to scrape the wikipedia page on **marathon world record progression**

- Women's record: 3:40:22 in 1926 to 2:11:53 last year!

First, open up this page in your browser. Take a look at its structure: What type of objects does it contain? How many tables does it have? Do these tables all share the same columns? What row- and columns-spans? Etc.

Application 1: Wikipedia



Marathon world record progression

From Wikipedia, the free encyclopedia

This list is a chronological progression of record times for the marathon. World records in the marathon are ratified by World Athletics, the international governing body for the sport of athletics.

Kenyan athlete Kelvin Kiptum set a men's world record time of 2:00:35 on October 8, 2023, at the 2023 Chicago Marathon.[1][2]

Ethiopian athlete Tigst Assefa broke the women's world record for a mixed-gender race with a time of 2:11:53 on September 24, 2023, at the 2023 Berlin Marathon.[3] In addition to the standard women's marathon world record, World Athletics also recognizes a second world record for women in the "Women Only" category, meaning that the marathon was run on a course without any male athletes in competition. The current "Women Only" record of 2:17:01 was set by Mary Keitany on April 23, 2017, at the London Marathon in the elite women's race.[4][5]

Kelvin Kiptum during his world record run at the 2023 Chicago marathon with 2:00:35

Criteria for record eligibility [edit]

See also: List of world records in athletics § Criteria

For a performance to be ratified as a world record by World Athletics, the marathon course on which the performance occurred must be 42.195 km (26.219 mi) long,[6] measured in a defined manner using the calibrated bicycle method[7] (the distance in kilometers being the official distance, the distance in miles is an approximation) and meet other criteria that rule out artificially fast times produced on courses aided by downhill slope or tailwind.[8] The criteria include:

- "The start and finish points of a course, measured along a theoretical straight line between them, shall not be further apart than 50% of the race distance."[6]
- "The decrease in elevation between the start and finish shall not exceed an average of one in a thousand, i.e. 1m per km."[6]

In recognizing Kenyan Geoffrey Mutai's mark of 2:03:02 at the 2011 Boston Marathon as (at the time) "the fastest Marathon ever run", the IAAF said: "Due to the elevation drop and point-to-point measurements of the Boston course, performances [on that course] are not eligible for World record consideration."[9]

Road racing events like the marathon were specifically excepted from World Athletics rule 260 18(d) that rejected from consideration those track and field performances set in mixed competition.[6]

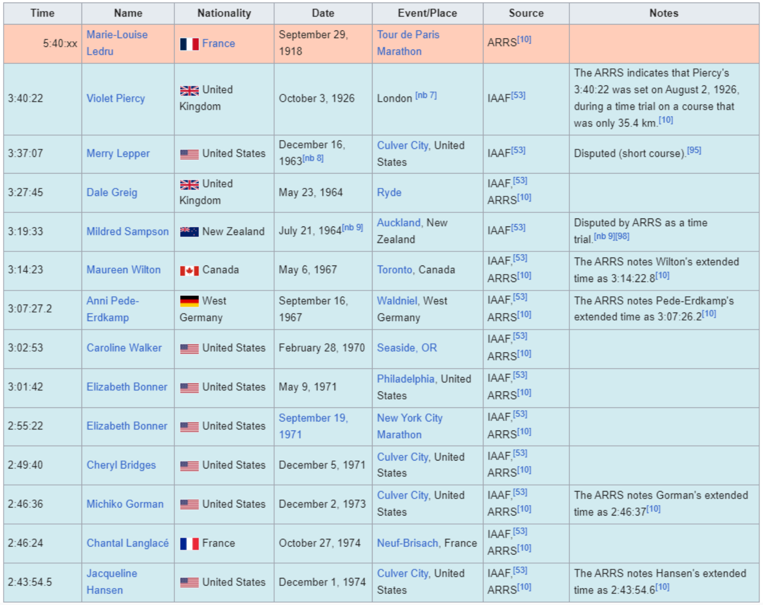
The Association of Road Racing Statisticians, an independent organization that compiles data from road running events, also maintains an alternate marathon world best progression but with standards they consider to be more stringent.[10][11]

Women's world record [edit]

The IAAF Congress at 2011 World Championships in Athletics passed a motion changing the record eligibility criteria effective October 6, 2007, so that women's world records must be set in all-women competitions.[12] The result of the change was that Radcliffe's 2:17:42 performance at the 2005 London Marathon would supplant her own existing women's mark as the "world record"; the earlier performance was to be referred to as a "world best".[12]

The decision was met with strong protest in Britain: in November 2011, an IAAF council member announced that Radcliffe's original mark would be allowed to stand, with the eventual decision that both marks would be

Tigst Assefa during her women's



Time	Name	Nationality	Date	Event/Place	Source	Notes
5:40:xx	Marie-Louise Ledru	France	September 29, 1918	Tour de Paris Marathon	ARRS[10]	
3:40:22	Violet Piercy	United Kingdom	October 3, 1926	London [nb 7]	IAAF[53]	The ARRS indicates that Piercy's 3:40:22 was set on August 2, 1926, during a time trial on a course that was only 35.4 km.[10]
3:37:07	Merry Lepper	United States	December 16, 1963[nb 8]	Culver City, United States	IAAF[53]	Disputed (short course).[95]
3:27:45	Dale Greig	United Kingdom	May 23, 1964	Ryde	IAAF,[53] ARRS[10]	
3:19:33	Mildred Sampson	New Zealand	July 21, 1964[nb 9]	Auckland, New Zealand	IAAF[53]	Disputed by ARRS as a time trial.[nb 9][98]
3:14:23	Maureen Wilton	Canada	May 6, 1967	Toronto, Canada	IAAF,[53] ARRS[10]	The ARRS notes Wilton's extended time as 3:14:22.8[10]
3:07:27.2	Anni Pede-Erdkamp	West Germany	September 16, 1967	Waldniel, West Germany	IAAF,[53] ARRS[10]	The ARRS notes Pede-Erdkamp's extended time as 3:07:26.2[10]
3:02:53	Caroline Walker	United States	February 28, 1970	Seaside, OR	IAAF,[53] ARRS[10]	
3:01:42	Elizabeth Bonner	United States	May 9, 1971	Philadelphia, United States	IAAF,[53] ARRS[10]	
2:55:22	Elizabeth Bonner	United States	September 19, 1971	New York City Marathon	IAAF,[53] ARRS[10]	
2:49:40	Cheryl Bridges	United States	December 5, 1971	Culver City, United States	IAAF,[53] ARRS[10]	
2:46:36	Michiko Gorman	United States	December 2, 1973	Culver City, United States	IAAF,[53] ARRS[10]	The ARRS notes Gorman's extended time as 2:46:37[10]
2:46:24	Chantal Langlacé	France	October 27, 1974	Neuf-Brisach, France	IAAF,[53] ARRS[10]	
2:43:54.5	Jacqueline Hansen	United States	December 1, 1974	Culver City, United States	IAAF,[53] ARRS[10]	The ARRS notes Hansen's extended time as 2:43:54.6[10]

Application 1: Wikipedia

Once you've familiarised yourself with the structure, read the whole page into R using the `rvest::read_html()` function.

```r
mthn = read_html("https://en.wikipedia.org/wiki/Marathon_world_record_prog
mthn
```

```
## {html_document}
## <html class="client-nojs vector-feature-language-in-header-enabled vector-fe
## [1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UTF-
## [2] <body class="skin--responsive skin-vector skin-vector-search-vue mediawi
```

Application 1: Wikipedia

As you can see, this is an **XML** document that contains **everything** needed to render the Wikipedia page.[1]

It's kind of like viewing someone's entire LaTeX document (preamble, syntax, etc.) when all we want are the data from some tables in their paper.

[1] XML stands for Extensible Markup Language and is one of the primary languages used for encoding and formatting web pages.

First Table: Women's Records

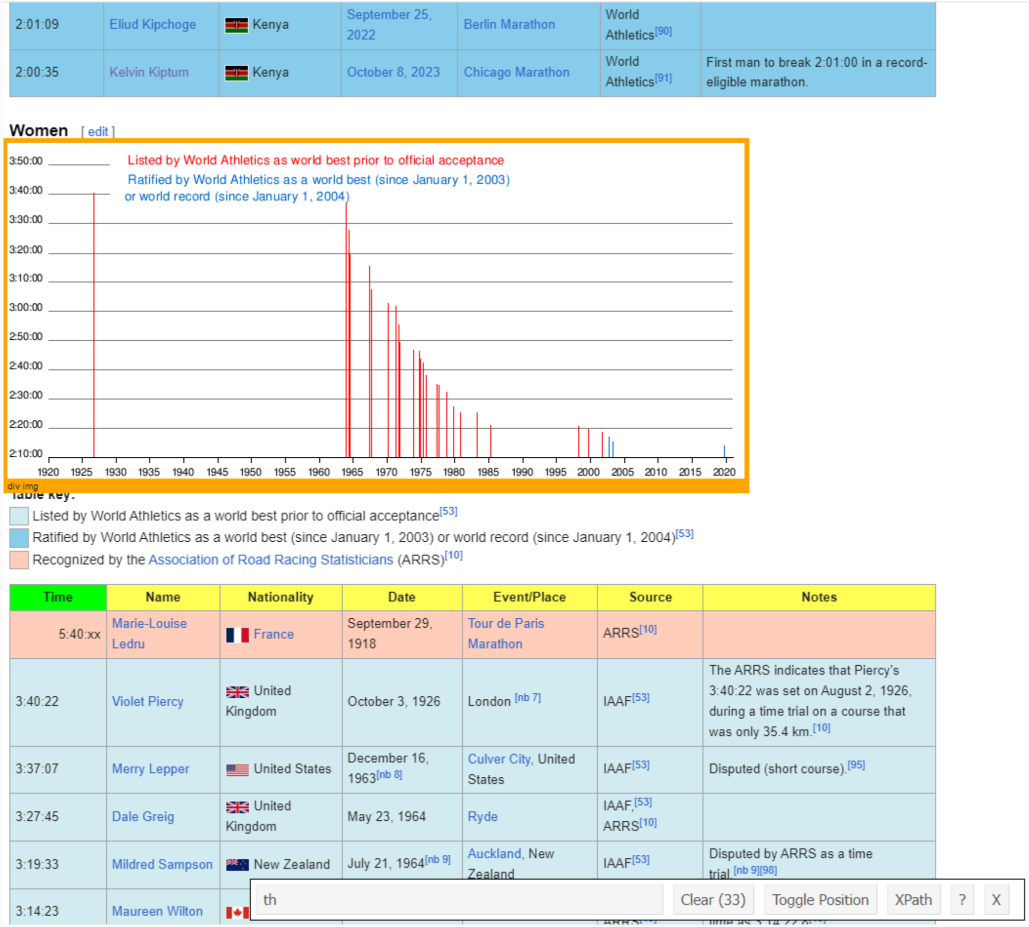
Let's start by scraping our first table from the page, which documents the **women's record progression**.

The first thing we need to do is identify the table's unique CSS selector using SelectorGadget.

Note: this *will* require trial and error - and a lot of clicking.

First Table: Women's Records

Start by activating SelectorGadget and clicking on a chart element.



2:01:09	Eliud Kipchoge	Kenya	September 25, 2022	Berlin Marathon	World Athletics[90]	
2:00:35	Kelvin Kiptum	Kenya	October 8, 2023	Chicago Marathon	World Athletics[91]	First man to break 2:01:00 in a record-eligible marathon.

Women [edit]

3:50:00
3:40:00
3:30:00
3:20:00
3:10:00
3:00:00
2:50:00
2:40:00
2:30:00
2:20:00
2:10:00
1920
1925
1930
1935
1940
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
Listed by World Athletics as world best prior to official acceptance
Ratified by World Athletics as a world best (since January 1, 2003)
or world record (since January 1, 2004)
div img

Table key:

□ Listed by World Athletics as a world best prior to official acceptance[53]

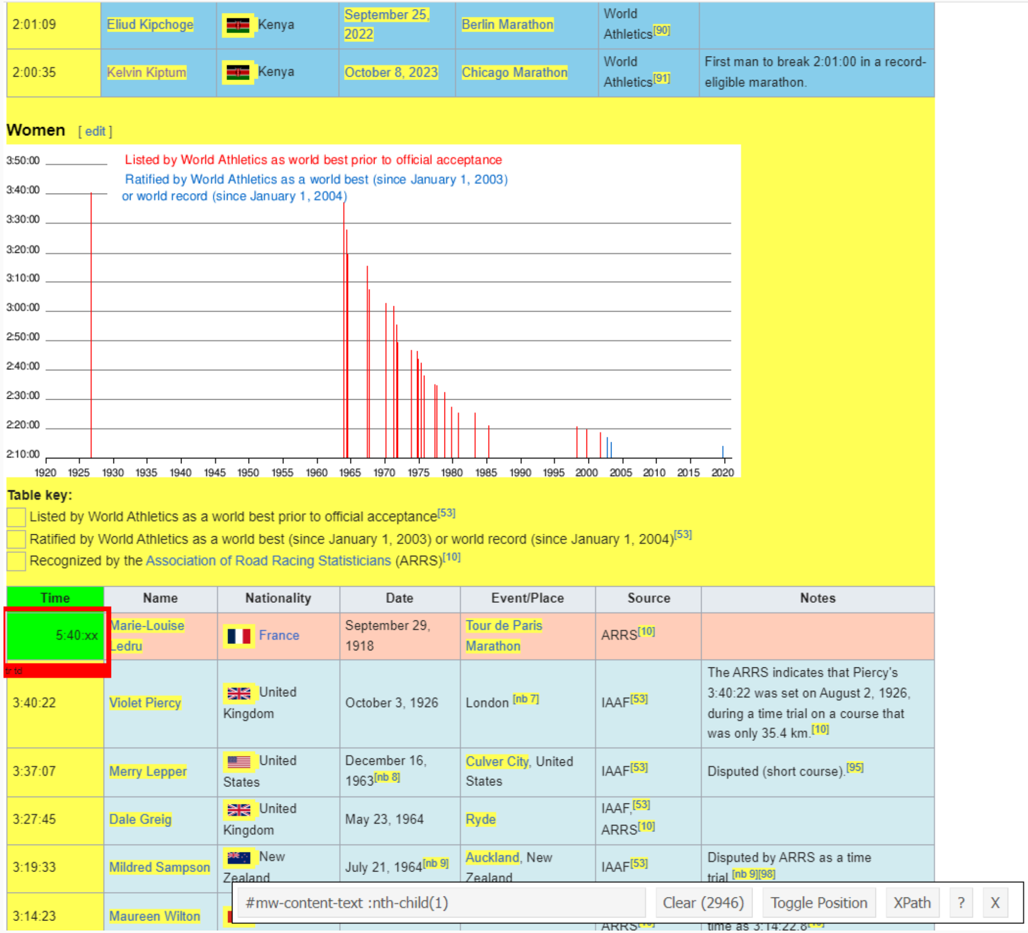
🟦 Ratified by World Athletics as a world best (since January 1, 2003) or world record (since January 1, 2004)[53]

🟧 Recognized by the Association of Road Racing Statisticians (ARRS)[10]

Time	Name	Nationality	Date	Event/Place	Source	Notes
5:40:xx	Marie-Louise Ledru	France	September 29, 1918	Tour de Paris Marathon	ARRS[10]	
3:40:22	Violet Piercy	United Kingdom	October 3, 1926	London [nb 7]	IAAF[53]	The ARRS indicates that Piercy's 3:40:22 was set on August 2, 1926, during a time trial on a course that was only 35.4 km.[10]
3:37:07	Merry Lepper	United States	December 16, 1963[nb 8]	Culver City, United States	IAAF[53]	Disputed (short course).[95]
3:27:45	Dale Greig	United Kingdom	May 23, 1964	Ryde	IAAF,[53] ARRS[10]	
3:19:33	Mildred Sampson	New Zealand	July 21, 1964[nb 9]	Auckland, New Zealand	IAAF[53]	Disputed by ARRS as a time trial.[nb 9][98]
3:14:23	Maureen Wilton					

First Table: Women's Records

Clicking on another chart element expands the selection (but too much!)



2:01:09	Eliud Kipchoge	Kenya	September 25, 2022	Berlin Marathon	World Athletics[90]	
2:00:35	Kelvin Kiptum	Kenya	October 8, 2023	Chicago Marathon	World Athletics[91]	First man to break 2:01:00 in a record-eligible marathon.

Women [edit]

Listed by World Athletics as world best prior to official acceptance
Ratified by World Athletics as a world best (since January 1, 2003)
or world record (since January 1, 2004)
3:50:00
3:40:00
3:30:00
3:20:00
3:10:00
3:00:00
2:50:00
2:40:00
2:30:00
2:20:00
2:10:00
1920
1925
1930
1935
1940
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020

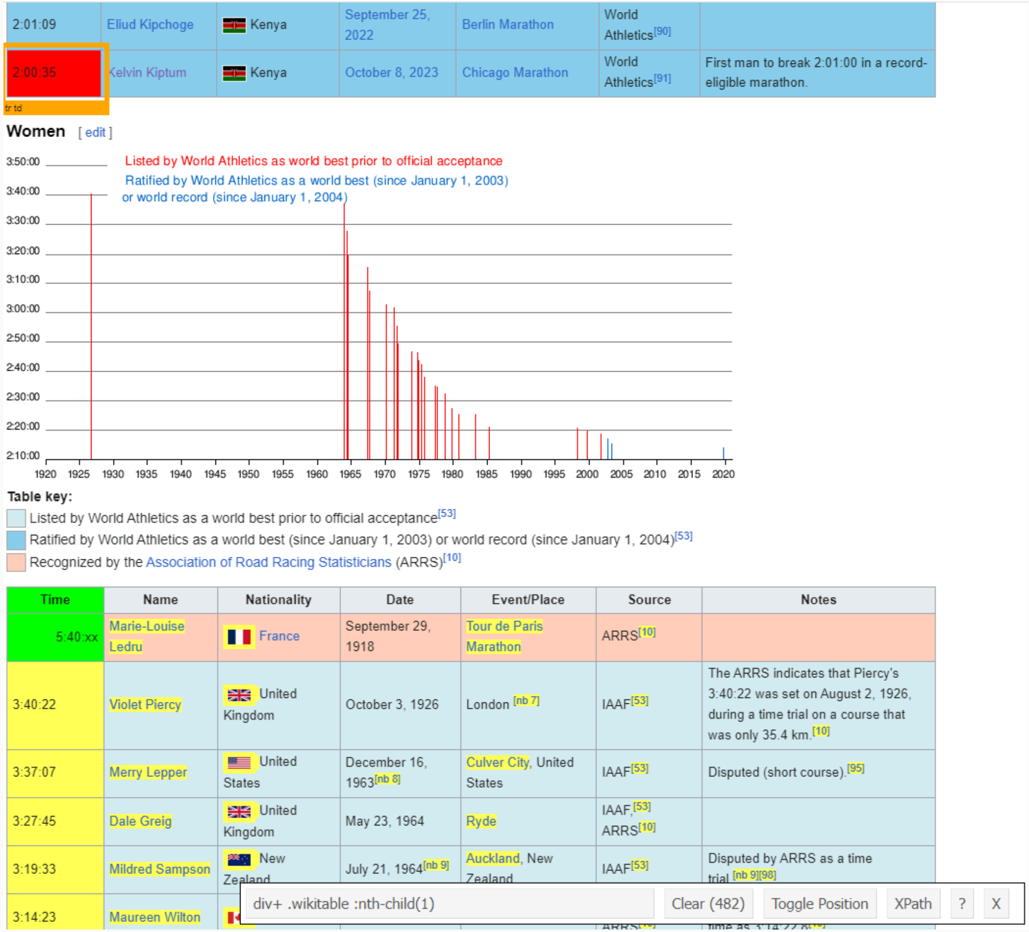
Table key:

- 🟨 Listed by World Athletics as a world best prior to official acceptance[53]
- 🟨 Ratified by World Athletics as a world best (since January 1, 2003) or world record (since January 1, 2004)[53]
- 🟨 Recognized by the Association of Road Racing Statisticians (ARRS)[10]

Time	Name	Nationality	Date	Event/Place	Source	Notes
5:40:xx	Marie-Louise Ledru	France	September 29, 1918	Tour de Paris Marathon	ARRS[10]	
3:40:22	Violet Piercy	United Kingdom	October 3, 1926	London [nb 7]	IAAF[53]	The ARRS indicates that Piercy's 3:40:22 was set on August 2, 1926, during a time trial on a course that was only 35.4 km.[10]
3:37:07	Merry Lepper	United States	December 16, 1963[nb 8]	Culver City, United States	IAAF[53]	Disputed (short course).[95]
3:27:45	Dale Greig	United Kingdom	May 23, 1964	Ryde	IAAF,[53] ARRS[10]	
3:19:33	Mildred Sampson	New Zealand	July 21, 1964[nb 9]	Auckland, New Zealand	IAAF[53]	Disputed by ARRS as a time trial [nb 9][98]
3:14:23	Maureen Wilton				ARRS[10]	time as 3:14:22.8[10]

First Table: Women's Records

Click the elements you *don't* want until they turn red:



2:01:09	Eliud Kipchoge	Kenya	September 25, 2022	Berlin Marathon	World Athletics[90]	
2:00:35	Kelvin Kiptum	Kenya	October 8, 2023	Chicago Marathon	World Athletics[91]	First man to break 2:01:00 in a record-eligible marathon.

Women [edit]

Table key:

- Listed by World Athletics as a world best prior to official acceptance[53]
- Ratified by World Athletics as a world best (since January 1, 2003) or world record (since January 1, 2004)[53]
- Recognized by the Association of Road Racing Statisticians (ARRS)[10]

Time	Name	Nationality	Date	Event/Place	Source	Notes
5:40:xx	Marie-Louise Ledru	France	September 29, 1918	Tour de Paris Marathon	ARRS[10]	
3:40:22	Violet Piercy	United Kingdom	October 3, 1926	London [nb 7]	IAAF[53]	The ARRS indicates that Piercy's 3:40:22 was set on August 2, 1926, during a time trial on a course that was only 35.4 km.[10]
3:37:07	Merry Lepper	United States	December 16, 1963[nb 8]	Culver City, United States	IAAF[53]	Disputed (short course).[95]
3:27:45	Dale Greig	United Kingdom	May 23, 1964	Ryde	IAAF,[53] ARRS[10]	
3:19:33	Mildred Sampson	New Zealand	July 21, 1964[nb 9]	Auckland, New Zealand	IAAF[53]	Disputed by ARRS as a time trial [nb 9][98]
3:14:23	Maureen Wilton					

div+ .wikitable :nth-child(1)

First Table: Women's Records

Working through this iterative process yields

`"div+ .wikitable :nth-child(1)"`.

We can use this unique CSS selector to isolate the women's record table content from the rest of the HTML document.

- **Extract the table content** with `html_element()`
- Parse the **HTML table into an R data frame** with `html_table()`

First Table: Women's Records

- **Extract the table content** with `html_element()`
- Parse the **HTML table into an R data frame** with `html_table()`

```
women ← mthn %>%
  html_element("div+ .wikitable :nth-child(1)") %>% ## select table eleme
  html_table()                                      ## convert to data fr

women
```

```
## # A tibble: 45 × 7
##    Time        Name                 Nationality    Date     Event/Place     Source No
##    <chr>       <chr>                <chr>          <chr>    <chr>            <chr>   <
##  1 5:40:xx     Marie-Louise Ledru France         Septe… Tour de Pari… ARRS[… "
##  2 3:40:22     Violet Piercy        United Kingdom Octob… London [nb 7] IAAF[… "
##  3 3:37:07     Merry Lepper         United States  Decem… Culver City,… IAAF[… "
##  4 3:27:45     Dale Greig           United Kingdom May 2… Ryde            IAAF,… "
##  5 3:19:33     Mildred Sampson      New Zealand    July … Auckland, Ne… IAAF[… "
##  6 3:14:23     Maureen Wilton       Canada         May 6… Toronto, Can… IAAF,… "
##  7 3:07:27.2 Anni Pede-Erdkamp West Germany   Septe… Waldniel, We… IAAF,… "
##  8 3:02:53     Caroline Walker      United States  Febru… Seaside, OR     IAAF,… "
```

First Table: Women's Records

Great, it worked! Now convert the date string to a format that R actually understands.[2]

```
women %>%
  mutate(Date = mdy(Date)) %>% ## convert string to date format
  head(4)
```

```
## # A tibble: 4 × 7
##   Time    Name               Nationality   Date       Event/Place   Source No
##   <chr>   <chr>              <chr>         <date>     <chr>         <chr>  <
## 1 5:40:xx Marie-Louise Ledru France        1918-09-29 Tour de Pari… ARRS[… "
## 2 3:40:22 Violet Piercy      United Kingd… 1926-10-03 London [nb 7] IAAF[… "
## 3 3:37:07 Merry Lepper       United States NA         Culver City,… IAAF[… "
## 4 3:27:45 Dale Greig         United Kingd… 1964-05-23 Ryde          IAAF,… "
```

[2] *Note:* If column name had spaces or capital letters, we could use the `janitor::clean_names()` convenience function to clean them. (Q: How else could we have done this?)

First Table: Women's Records

Alright that mostly worked, but there are a few hyperlink references in the dates leading to `NA`s.

- This is a case where using a regular expression is convenient: match the pattern `"[nb X]"` at the **end of the strings**

```
women <- women %>%
  mutate(across(where(is.character), # mutate across all character string
                ~str_replace(.x, "\\[nb [0-9]+\\]$", "")), # remove [nb 0
         Date = mdy(Date))  # convert string to date format

women
```

```
## # A tibble: 45 × 7
##    Time      Name              Nationality Date       Event/Place Source No
##    <chr>     <chr>             <chr>       <date>     <chr>         <chr>  <
##  1 5:40:xx   Marie-Louise Led… France      1918-09-29 "Tour de Par… ARRS[… "
##  2 3:40:22   Violet Piercy     United Kin… 1926-10-03 "London "     IAAF[… "
##  3 3:37:07   Merry Lepper      United Sta… 1963-12-16 "Culver City… IAAF[… "
##  4 3:27:45   Dale Greig        United Kin… 1964-05-23 "Ryde"        IAAF,… "
```

Table 2: Men's Records

We could stop here and plot the women's records, but while we're here let's grab the men's records as well.

Challenge: take a couple minutes to use SelectorGadget and find the CSS selector for the men's record table.

- Don't peek at next slide until you give it a shot!

```
men <- mthn %>%
  html_element("") %>% ## add the selector!
  html_table()    %>%
  mutate(across(where(is.character), # mutate across all character string
                ~str_replace(.x, "\\[nb [0-9]+\\]$", "")), # remove [nb 0
         Date = mdy(Date))
```

Table 2: Men's Records

```
men ← mthn %>%
  html_element("p+ .wikitable :nth-child(1)") %>% ## add the selector!
  html_table()    %>%
  mutate(across(where(is.character), # mutate across all character string
                ~str_replace(.x, "\\[nb [0-9]+\\]$", "")), # remove [nb 0
         Date = mdy(Date))
tail(men)
```

```
## # A tibble: 6 × 7
##   Time    Name           Nationality Date       Event/Place    Source    No
##   <chr>   <chr>          <chr>       <date>     <chr>            <chr>     <
## 1 2:03:38 Patrick Makau  Kenya       2011-09-25 Berlin Marathon  IAAF,[82… "
## 2 2:03:23 Wilson Kipsang Kenya       2013-09-29 Berlin Marathon  IAAF[85]… "
## 3 2:02:57 Dennis Kimetto Kenya       2014-09-28 Berlin Marathon  IAAF[87]… "
## 4 2:01:39 Eliud Kipchoge Kenya       2018-09-16 Berlin Marathon  IAAF[89]  "
## 5 2:01:09 Eliud Kipchoge Kenya       2022-09-25 Berlin Marathon  World At… "
## 6 2:00:35 Kelvin Kiptum  Kenya       2023-10-08 Chicago Marathon World At… "
```

Browser Inspection Tools

SelectorGadget is a great tool, but sometimes it **takes more work than necessary** and isn't available in all browsers.

Alternate approach: use the **inspect web element**

- Chrome: Right-click > "Inspect" (`Ctrl + Shift + I`)
- Scroll over source elements until the table of interest is highlighted
- Use the selector that pops up over the element
  - i.e. `"table.wikitable"`

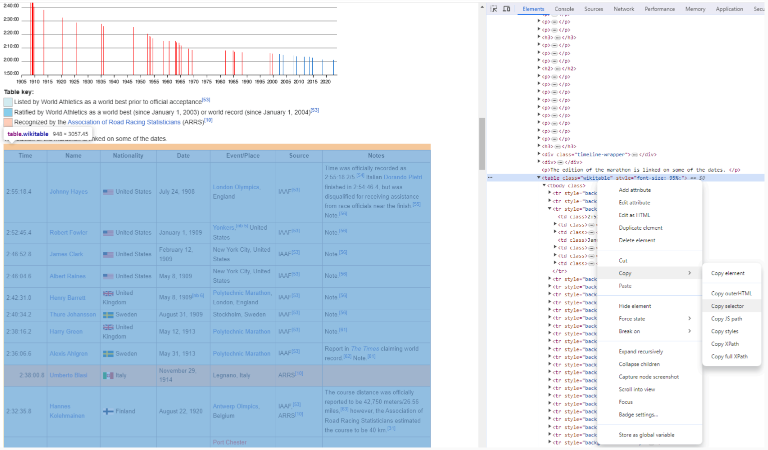


Table key:

- Listed by World Athletics as a world best prior to official acceptance[53]
- Ratified by World Athletics as a world best (since January 1, 2003) or world record (since January 1, 2004)[53]
- Recognized by the Association of Road Racing Statisticians (ARRS)[10]

...nked on some of the dates.

Time	Name	Nationality	Date	Event/Place	Source	Notes
2:55:18.4	Johnny Hayes	United States	July 24, 1908	London Olympics, England	IAAF[53]	Time was officially recorded as 2:55:18 2/5.[54] Italian Dorando Pietri finished in 2:54:46.4, but was disqualified for receiving assistance from race officials near the finish.[55] Note.[56]
2:52:45.4	Robert Fowler	United States	January 1, 1909	Yonkers,[nb 5] United States	IAAF[53]	Note.[56]
2:46:52.8	James Clark	United States	February 12, 1909	New York City, United States	IAAF[53]	Note.[56]
2:46:04.6	Albert Raines	United States	May 8, 1909	New York City, United States	IAAF[53]	Note.[56]
2:42:31.0	Henry Barrett	United Kingdom	May 8, 1909[nb 6]	Polytechnic Marathon, London, England	IAAF[53]	Note.[56]
2:40:34.2	Thure Johansson	Sweden	August 31, 1909	Stockholm, Sweden	IAAF[53]	Note.[56]
2:38:16.2	Harry Green	United Kingdom	May 12, 1913	Polytechnic Marathon	IAAF[53]	Note.[61]
2:36:06.6	Alexis Ahlgren	Sweden	May 31, 1913	Polytechnic Marathon	IAAF[53]	Report in *The Times* claiming world record.[62] Note.[61]
2:38:00.8	Umberto Blasi	Italy	November 29, 1914	Legnano, Italy	ARRS[10]	
2:32:35.8	Hannes Kolehmainen	Finland	August 22, 1920	Antwerp Olmpics, Belgium	IAAF,[53] ARRS[10]	The course distance was officially reported to be 42,750 meters/26.56 miles,[63] however, the Association of Road Racing Statisticians estimated the course to be 40 km.[31]
				Port Chester		

```
▸<p> ⋯ </p>
▸<p> ⋯ </p>
▸<h3> ⋯ </h3>
▸<p> ⋯ </p>
▸<p> ⋯ </p>
▸<p> ⋯ </p>
▸<h2> ⋯ </h2>
▸<p> ⋯ </p>
▸<p> ⋯ </p>
▸<p> ⋯ </p>
▸<p> ⋯ </p>
▸<p> ⋯ </p>
▸<p> ⋯ </p>
▸<p> ⋯ </p>
▸<p> ⋯ </p>
▸<p> ⋯ </p>
▸<h3> ⋯ </h3>
▸<div class="timeline-wrapper"> ⋯ </div>
▸<div> ⋯ </div>
 <p>The edition of the marathon is linked on some of the dates. </p>
▾<table class="wikitable" style="font-size: 95%;"> == $0
  ▾<tbody class>
    ▸<tr style="back
    ▸<tr style="back
    ▾<tr style="back
       <td class>2:5
      ▸<td class> ⋯ <
      ▸<td class> ⋯ <
       <td class>Jan
      ▸<td class> ⋯ <
      ▸<td class> ⋯ <
      ▸<td class> ⋯ <
     </tr>
    ▸<tr style="back
    ▸<tr style="back
    ▸<tr style="back
```

Duplicate Selectors

If we look at the source code a bit more we can see that the selector `table.wikitable` **isn't unique**!

In cases like this, use `html_elements()` to retrieve all matching elements as a list

```
mthn_tabs <- mthn %>%
  html_elements("table.wikitable") ## select all elements matching the se

# first match is men's results
men2 <- mthn_tabs[[1]] %>% html_table()

# second match is women's results
women2 <- mthn_tabs[[2]] %>% html_table()
```

Marathon Progression

Challenge: combine both tables into a single dataframe and plot the record progression

- Clean the `Time` variable (get rid of alphanumeric characters at the end)
- Separate the time variable into hours/minutes/seconds using `separate()`
- Use `hours(), minutes(), seconds()` from `lubridate` to create a time variable for the marathon time
- Create a plot with
  - Marathon time on the y-axis
    - Make sure to include a `scale_y_time()` layer!
  - Date on the x-axis
  - Linetype aesthetic mapping based on the group (men vs. women)
  - Use a nice theme

Answer in a few slides (no peeking until you've tried first)

Marathon Progression

```
men ← mutate(men, Group = "Men")
women ← mutate(women, Group = "Women")

mthn_prog ← rbind(men, women) %>%
  mutate(Time = str_replace_all(Time, "[:alpha:]", "")) %>%
   separate(col = "Time", into = c("Hours", "Minutes", "Seconds"), sep = '
  # add time variable using lubridate functions
  mutate(mthn_time = hours(Hours) + minutes(Minutes) + seconds(Seconds))


ggplot(mthn_prog, aes(x = Date, y = mthn_time)) +
  geom_line(aes(linetype = Group)) +
  scale_y_time() +
  labs(y = "Marathon World Record",
       x = "Year") +
  theme_minimal()
```

Marathon Progression

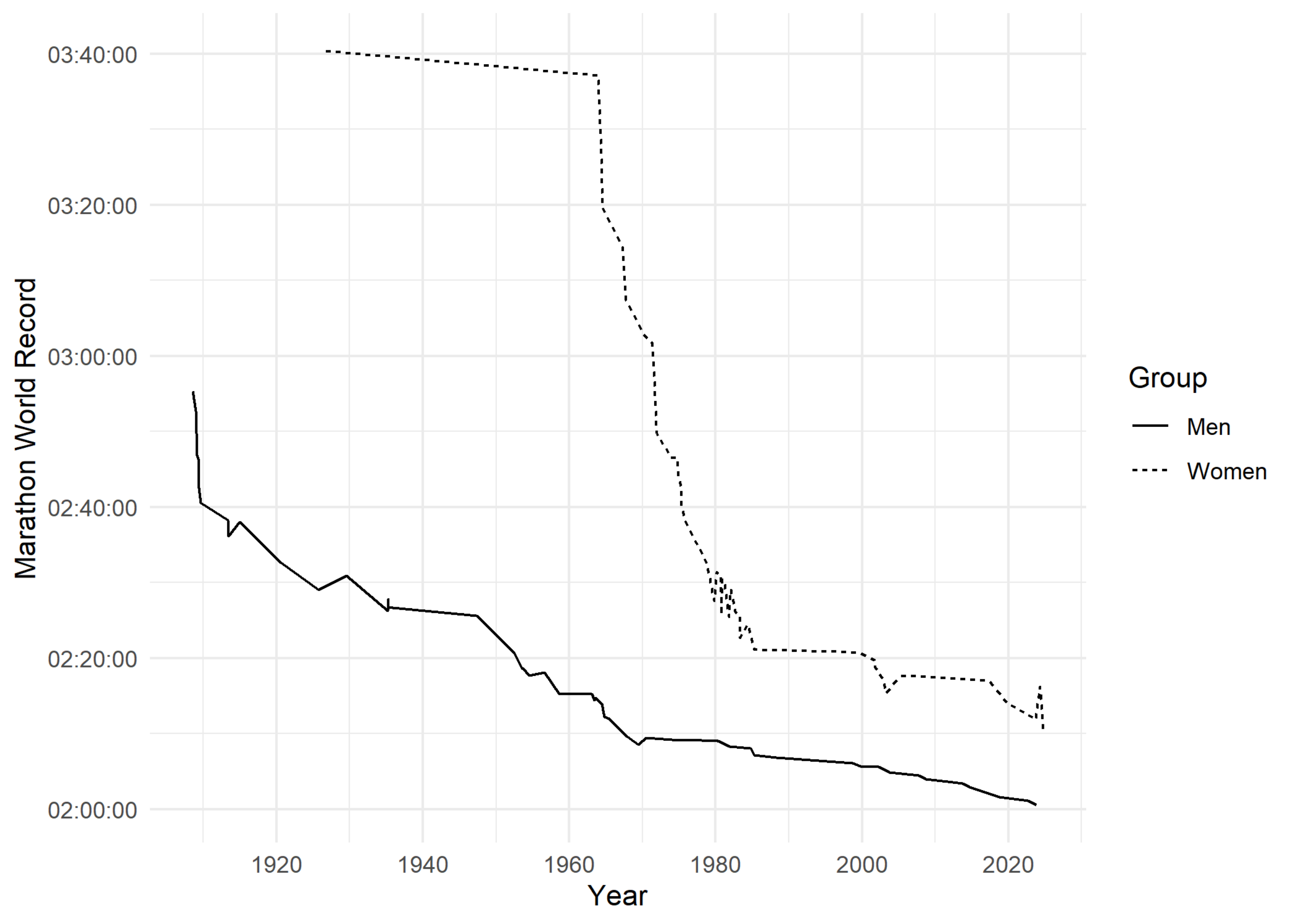


Table of Contents