

# Web scraping

Data Science in a Box  
[datasciencebox.org](http://datasciencebox.org)

Modified by Tyler George



# Scraping the web



# Scraping the web: what? why?

- Increasing amount of data is available on the web



# Scraping the web: what? why?

- Increasing amount of data is available on the web
- These data are provided in an unstructured format: you can always copy&paste, but it's time-consuming and prone to errors



# Scraping the web: what? why?

- Increasing amount of data is available on the web
- These data are provided in an unstructured format: you can always copy&paste, but it's time-consuming and prone to errors
- Web scraping is the process of extracting this information automatically and transform it into a structured dataset



# Scraping the web: what? why?

- Increasing amount of data is available on the web
- These data are provided in an unstructured format: you can always copy&paste, but it's time-consuming and prone to errors
- Web scraping is the process of extracting this information automatically and transform it into a structured dataset
- Two different scenarios:
  - Screen scraping: extract data from source code of website, with html parser (easy) or regular expression matching (less easy).
  - Web APIs (application programming interface): website offers a set of structured http requests that return JSON or XML files.



# Web Scraping with rvest



[datasciencebox.org](http://datasciencebox.org)

# Hypertext Markup Language

- Most of the data on the web is still largely available as HTML
- It is structured (hierarchical / tree based), but it's often not available in a form useful for analysis (flat / tidy).

```
<html>
  <head>
    <title>This is a title</title>
  </head>
  <body>
    <p align="center">Hello world!</p>
  </body>
</html>
```



# rvest

- The **rvest** package makes basic processing and manipulation of HTML data straight forward
- It's designed to work with pipelines built with %>%



# Core rvest functions

- `read_html` - Read HTML data from a url or character string
- `html_node` - Select a specified node from HTML document
- `html_nodes` - Select specified nodes from HTML document
- `html_table` - Parse an HTML table into a data frame
- `html_text` - Extract tag pairs' content
- `html_name` - Extract tags' names
- `htmlAttrs` - Extract all of each tag's attributes
- `html_attr` - Extract tags' attribute value by name



# SelectorGadget

- Open source tool that eases CSS selector generation and discovery
- Easiest to use with the Chrome Extension
- Find out more on the SelectorGadget vignette

## SelectorGadget: point and click CSS selectors



A screenshot of a computer screen showing the Hacker News homepage. The title bar says "SelectorGadget Screencast from Andrew Cantino". The main content area shows a list of news items. The 19th item in the list, which is highlighted with a yellow background, is "New theory may explain the notorious cold fusion experiment from two decades ago (discovermagazine.com)". A cursor arrow points to the number 19 next to this highlighted item.

| Rank | Title  | Source                 | Comments   |
|------|--|------------------------|--|
| 1.   | AnandTech: Microsoft Surface Review  | (anandtech.com)        | 77 points by bartolo 2 hours ago   37 comments         |
| 2.   | Wired's Review of the Microsoft Surface  | (wired.com)            | 42 points by colinplamondon 2 hours ago   16 comments  |
| 3.   | Zynga May Have Just Laid Off 100+ Employees From Its Austin Office               | (techcrunch.com)       | 386 points by hompawze 10 hours ago   14 comments      |
| 4.   | The Hardware Renaissance   | (techcrunch.com)       | 366 points by niquresh 11 hours ago   171 comments     |
| 5.   | Don't Call The New Microsoft Surface RT A Tablet, This Is A PC                   | (techcrunch.com)       | 23 points by vyrtek 2 hours ago   36 comments          |
| 6.   | Why we buy into ideas: how to convince others of our thoughts                    | (bufferapp.com)        | 6 points by sunas34 23 minutes ago   discuss           |
| 7.   | The rise of the "successful" unsustainable company                               | (asmartbear.com)       | 291 points by yannickmache 12 hours ago   105 comments |
| 8.   | Under the hood of Windows 8, or why desktop users should upgrade from Windows 7  | (extremetech.com)      | 261 points by eve_ 9:12 hours ago   170 comments       |
| 9.   | Marc Andreessen's Productivity Trick to Feeling Marvelously Efficient            | (idonethis.com)        | 106 points by mikemk 7 hours ago   34 comments         |
| 10.  | Show HN: Taurus.io - Create a product tour for your web app in 15 minutes        | (taurus.io)            | 31 points by etzio 3 hours ago   30 comments           |
| 11.  | The PC isn't dead  | (dendory.net)          | 9 points by dendory 1 hour ago   6 comments            |
| 12.  | Ceefax Final Broadcast: "Goodbye, cruel world."                                  | (h4ck.in)              | 76 points by learmers 7 hours ago   24 comments        |
| 13.  | Show HN: Fact check last night's Presidential debate with Quip                   | (quipvideo.com)        | 32 points by dimvaldran 4 hours ago   12 comments      |
| 14.  | Increasing wireless network speed by 1000%, by replacing packets with algebra    | (extremetech.com)      | 98 points by oliver 7 hours ago   30 comments          |
| 15.  | Amazon reopen wiped Kindle account   | (translate.google.com) | 2518 points by EwanTee 15 hours ago   137 comments     |
| 16.  | Zynga CEO Mark Pincus Confirms Layoffs: 5% of Workforce                          | (techcrunch.com)       | 47 points by nikunj 6 hours ago   11 comments          |
| 17.  | Stanford grad's site nets Southwest 'cease and desist'                           | (paloaltoonline.com)   | 21 points by cb33 4 hours ago   18 comments            |
| 18.  | OrderAhead is hiring a Marketing Associate                                       |                        | 2 hours ago  |
| 19.  | New theory may explain the notorious cold fusion experiment from two decades ago | (discovermagazine.com) |  |



# Using the SelectorGadget

The screenshot shows a web browser displaying the [IMDb Top 250](https://www.imdb.com/chart/top) chart. The page features a large banner at the top announcing "PREMIERES TONIGHT STARTING 8/7c CBS". Below the banner, the "IMDb Charts" section is visible, specifically the "Top Rated Movies" chart. The chart lists the top 250 movies based on user ratings. The first entry, "The Shawshank Redemption (1994)", is highlighted with a yellow selection box from the SelectorGadget extension. The extension's interface is overlaid on the bottom right of the page, showing options like "SHARE", "IMDb Charts", and "Top Rated Movies".

| Rank & Title                       | IMDb Rating | Your Rating |
|------------------------------------|-------------|-------------|
| 1. The Shawshank Redemption (1994) | 9.2         | ☆           |
| 2. The Godfather (1972)            | 9.2         | ☆           |
| 3. The Godfather: Part II (1974)   | 9.1         | ☆           |

IMDb Top 250 - IMDb

imdb.com/chart/top/

IMDb Menu All Search IMDb

SHARE

Sort by: Ranking

IMDb Rating Your Rating

Rank & Title

1. The Shawshank Redemption (1994) ★ 9.2 ★ +

2. The Godfather (1972) ★ 9.1 ★ +

3. The Godfather: Part II (1974) ★ 9.0 ★ +

4. The Dark Knight (2008) ★ 9.0 ★ +

5. 12 Angry Men (1957) ★ 8.9 ★ +

6. Schindler's List (1993) ★ 8.9 ★ +

7. The Lord of the Rings: The Return of the King (2003) ★ 8.9 ★ +

You Have Seen  
0/250 (0%)  
 Hide titles I've seen

IMDb Charts

Box Office  
Most Popular Movies  
Top Rated Movies  
Top Rated English Movies  
Most Popular TV  
Top Rated TV  
Top Rated Indian Movies  
Lowest Rated Movies

Top Rated Movies by Genre

Action  
Adventure  
Animation  
Biography  
Comedy  
Crime  
Drama  
Family  
Fantasy  
Film-Noir  
History  
Horror  
Music  
Musical  
Mystery  
Romance

Click on the app logo next to the search bar in your browser

IMDb Top 250 - IMDb

imdb.com/chart/top/

IMDb Charts

## Top Rated Movies

Top 250 as rated by IMDb Users

Showing 250 Titles

Sort by: Ranking

| Rank & Title  | IMDb Rating          | Your Rating |                                 |
|---|----------------------|-------------|---------------------------------|
| 1. The Shawshank Redemption (1994)                      | ★ 9.2                | ★           | [+]                             |
| 2. The Godfather (1972)                                 | ★ 9.1                | ★           | [+]                             |
| 3. The Godfather: Part II (1974)                        | ★ 9.0                | ★           | [+]                             |
| 4. The Dark Knight (2008)                               | ★ 9.0                | ★           | [+]                             |
| 5. 12 Angry Men (1957)                                  | ★ 8.9                | ★           | [+]                             |
| 6. Schindler's List (1993)                              | ★ 8.9                | ★           | [+]                             |
| 7. The Lord of the Rings: The Return of the King (2003) | No valid path found. |             | Clear Toggle Position XPath ? X |

SHARE

You Have Seen  
0/250 (0%)  
 Hide titles I've seen

IMDb Charts

- Box Office
- Most Popular Movies
- Top Rated Movies
- Top Rated English Movies
- Most Popular TV
- Top Rated TV
- Top Rated Indian Movies
- Lowest Rated Movies

Top Rated Movies by Genre

- Action
- Adventure
- Animation
- Biography
- Comedy
- Crime
- Drama
- Family
- Fantasy
- Film-Noir
- History
- Horror

Box will open in the bottom right of the browser

Click on a page element, and it will turn green

IMDb Charts

## Top Rated Movies

Top 250 as rated by IMDb Users

Showing 250 Titles

Sort by: Ranking

|   | IMDb Rating | Your Rating |     |
|---|-------------|-------------|-----|
| 1. The Shawshank Redemption (1994)                      | 9.2         | ★           | [+] |
| 2. The Godfather (1972)                                 | 9.1         | ★           | [+] |
| 3. The Godfather: Part II (1974)                        | 9.0         | ★           | [+] |
| 4. The Dark Knight (2008)                               | 9.0         | ★           | [+] |
| 5. 12 Angry Men (1957)                                  | 8.9         | ★           | [+] |
| 6. Schindler's List (1993)                              | 8.9         | ★           | [+] |
| 7. The Lord of the Rings: The Return of the King (2003) | 8.9         | ★           | [+] |
| 8. Pulp Fiction (1994)                                  |             |             |     |

SHARE

You Have Seen  
0/250 (0%)  
 Hide titles I've seen

IMDb Charts

- Box Office
- Most Popular Movies
- Top Rated Movies
- Top Rated English Movies
- Most Popular TV
- Top Rated TV
- Top Rated Indian Movies
- Lowest Rated Movies

Top Rated Movies by Genre

- Action
- Adventure
- Animation
- Biography
- Comedy
- Crime
- Drama
- Family
- Fantasy
- Film-Noir
- History
- Horror
- Music
- Musical
- Mystery

.titleColumn

Clear (250) Toggle Position XPath ? X

selectorbad get will generate a minimal CSS selector for that element, and will highlight everything that is matched by the selector in yellow

IMDb Top 250 - IMDb

imdb.com/chart/top/

IMDb Menu All Search IMDb

SHARE

Sort by: Ranking

IMDb Rating Your Rating

Rank & Title

1. The Shawshank Redemption (1994) ★ 9.2

2. The Godfather (1972) ★ 9.1

3. The Godfather: Part II (1974) ★ 9.0

4. The Dark Knight (2008) ★ 9.0

5. 12 Angry Men (1957) ★ 8.9

6. Schindler's List (1993) ★ 8.9

7. The Lord of the Rings: The Return of the King

You Have Seen  
0/250 (0%)  
 Hide titles I've seen

IMDb Charts

- Box Office
- Most Popular Movies
- Top Rated Movies
- Top Rated English Movies
- Most Popular TV
- Top Rated TV
- Top Rated Indian Movies
- Lowest Rated Movies

Top Rated Movies by Genre

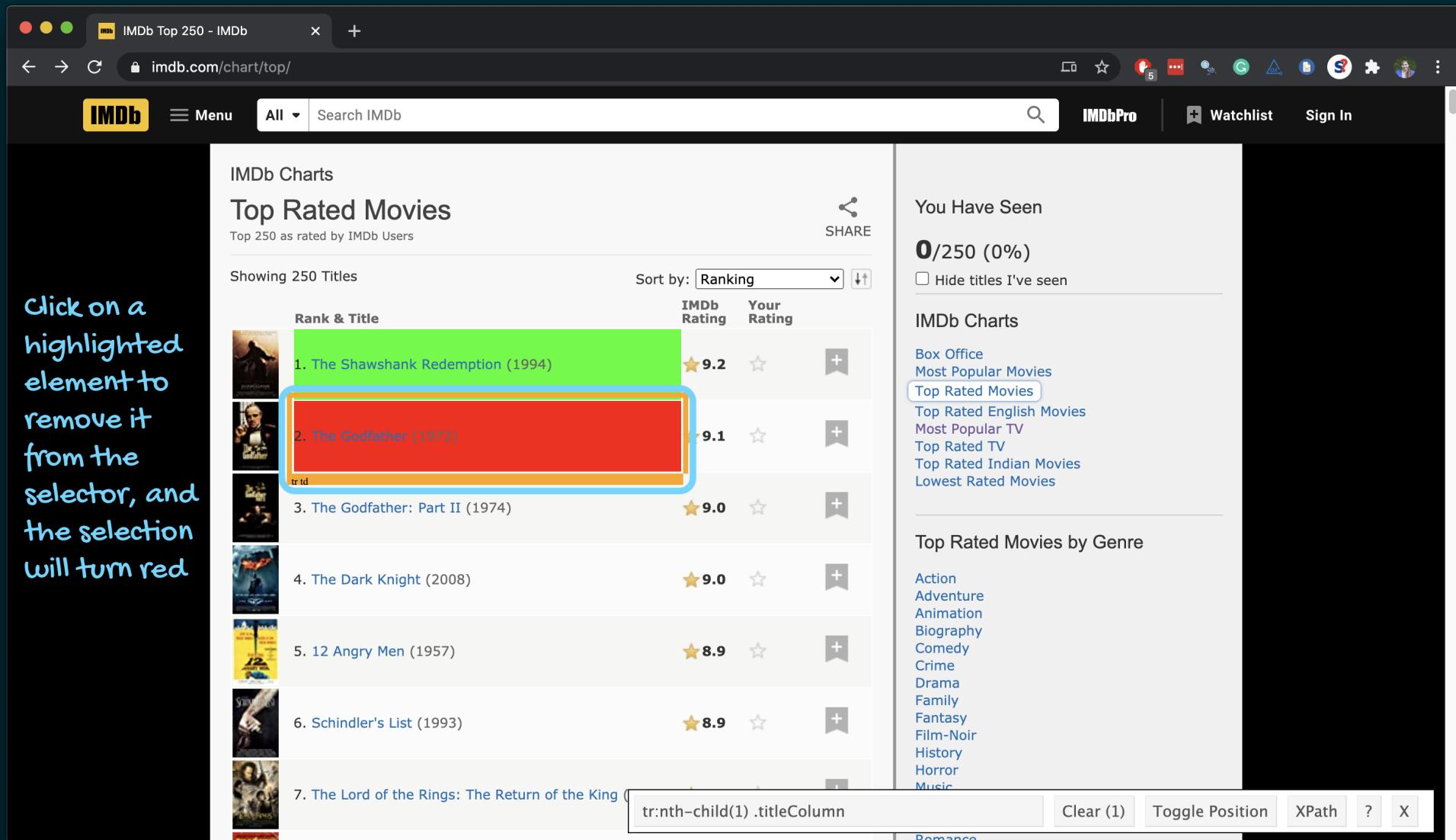
- Action
- Adventure
- Animation
- Biography
- Comedy
- Crime
- Drama
- Family
- Fantasy
- Film-Noir
- History
- Horror
- Music

tr:nth-child(1) .titleColumn

Romance

Clear (1) Toggle Position XPath ? X

Click on a highlighted element to remove it from the selector, and the selection will turn red



IMDb Top 250 - IMDb

imdb.com/chart/top/

IMDb Charts

## Top Rated Movies

Top 250 as rated by IMDb Users

Showing 250 Titles

Sort by: Ranking

| Rank & Title  | IMDb Rating | Your Rating |
|---|-------------|-------------|
| 1. The Shawshank Redemption (1994)                      | 9.2         | ☆           |
| 2. The Godfather (1972)                                 | 9.1         | ☆           |
| 3. The Godfather: Part II (1974)                        | 9.0         | ☆           |
| 4. The Dark Knight (2008)                               | 9.0         | ☆           |
| 5. 12 Angry Men (1957)                                  | 8.9         | ☆           |
| 6. Schindler's List (1993)                              | 8.9         | ☆           |
| 7. The Lord of the Rings: The Return of the King (2003) | 8.9         | ☆           |

SHARE

You Have Seen  
0/250 (0%)  
 Hide titles I've seen

IMDb Charts

- Box Office
- Most Popular Movies
- Top Rated Movies
- Top Rated English Movies
- Most Popular TV
- Top Rated TV
- Top Rated Indian Movies
- Lowest Rated Movies

Top Rated Movies by Genre

- Action
- Adventure
- Animation
- Biography
- Comedy
- Crime
- Drama
- Family
- Fantasy
- Film-Noir
- History
- Horror
- Music

tr~ tr+ tr .titleColumn , tr:nth-child(1) .titleColumn

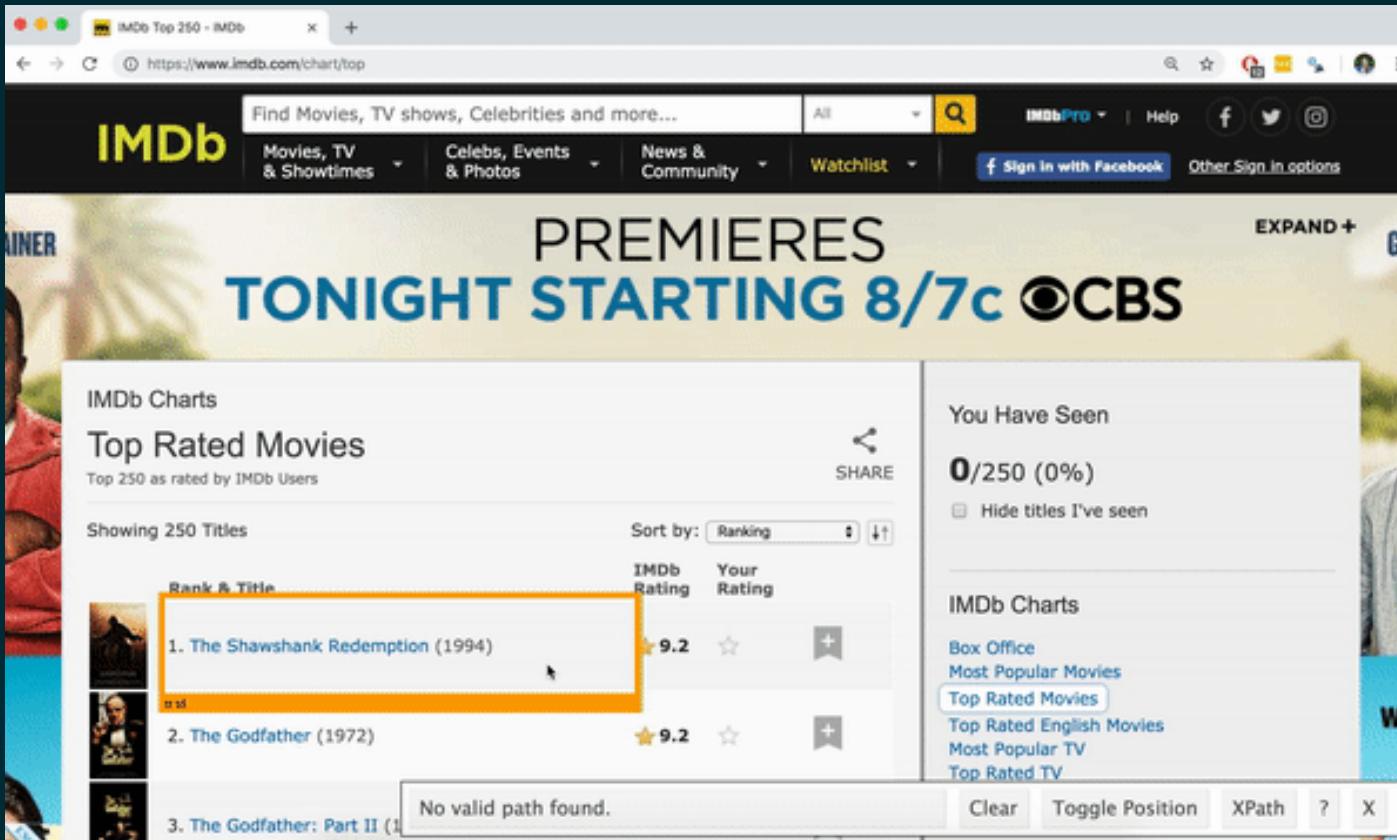
Clear (249) Toggle Position XPath ? X

Romance

Click on an unhighlighted element to add it to the selector and it will turn green

# Using the SelectorGadget

Through this process of selection and rejection, SelectorGadget helps you come up with the appropriate CSS selector for your needs



# Top 250 movies on IMDB

Take a look at the source code, look for the tag table tag (control +u in chrome):  
<http://www.imdb.com/chart/top>

IMDb Charts

## Top Rated Movies

Top 250 as rated by IMDb Users

Showing 250 Titles

Sort by: **Ranking**

| Rank & Title                                       | IMDb Rating | Your Rating |  |
|--|-------------|-------------|--|
| 1. <a href="#">The Shawshank Redemption (1994)</a> | 9.2         |             |  |
| 2. <a href="#">The Godfather (1972)</a>            | 9.1         |             |  |
| 3. <a href="#">The Godfather: Part II (1974)</a>   | 9.0         |             |  |

```
599   <div class="desc">Showing <span>250</span> Titles</div>
600   </div>
601   <br class="clear">
602   <table class="chart full-width" data-caller-name="chart-top250movie">
603     <colgroup>
604       <col class="chartTableColumnPoster"/>
605       <col class="chartTableColumnTitle"/>
606       <col class="chartTableColumnIMDbRating"/>
607       <col class="chartTableColumnYourRating"/>
608       <col class="chartTableColumnWatchlistRibbon"/>
609     </colgroup>
610     <thead>
611       <tr>
612         <th></th>
613         <th>Rank & Title</th>
614         <th>IMDb Rating</th>
615         <th>Your Rating</th>
616         <th></th>
617       </tr>
618     </thead>
619     <tbody class="lister-list">
620       <tr>
621         <td class="posterColumn">
622           <span name="rk" data-value="1"></span>
623           <span name="ir" data-value="9.222796866017044"></span>
624           <span name="us" data-value="7.791552811"></span>
625           <span name="nv" data-value="2297666"></span>
626           <span name="ur" data-value="-1.7772031339829564"></span>
627         <a href="/title/tt0111161/?pf_rd_m=A2FGELUUNQJNL&pf_rd_p=e31d89dd-322d-4646-8962-
628           327b42fe94b1&pf_rd_r=RP41R6C3PS7J108DRNN6pf_rd_s=center-
629           1&pf_rd_t=15506&pf_rd_i=top&ref_=chttp_tt_1"
630           > 
633           </a>      </td>
```



# First check if you're allowed!

```
library(robotstxt)
paths_allowed("http://www.imdb.com")
```

```
## [1] TRUE
```

vs. e.g.

```
paths_allowed("http://www.facebook.com")
```

```
## [1] FALSE
```



# Plan

IMDb Charts

## Top Rated Movies

Top 250 as rated by IMDb Users

Showing 250 Titles

Sort by: Ranking

| Rank & Title                                       | IMDb Rating | Your Rating         |
|--|-------------|---------------------|
| 1. <a href="#">The Shawshank Redemption</a> (1994) | ★ 9.2       | ★ <a href="#">+</a> |
| 2. <a href="#">The Godfather</a> (1972)            | ★ 9.1       | ★ <a href="#">+</a> |
| 3. <a href="#">The Godfather: Part II</a> (1974)   | ★ 9.0       | ★ <a href="#">+</a> |
| 4. <a href="#">The Dark Knight</a> (2008)          | ★ 9.0       | ★ <a href="#">+</a> |
| 5. <a href="#">12 Angry Men</a> (1957)             | ★ 8.9       | ★ <a href="#">+</a> |
| 6. <a href="#">Schindler's List</a> (1993)         | ★ 8.9       | ★ <a href="#">+</a> |

| title | year | rating |
|-------|------|--------|
|       |      |        |

# Plan

1. Read the whole page
2. Scrape movie titles and save as `titles`
3. Scrape years movies were made in and save as `years`
4. Scrape IMDB ratings and save as `ratings`
5. Create a data frame called `imdb_top_250` with variables `title`, `year`, and `rating`



# Step 1. Read the whole page



# Read the whole page

```
page <- read_html("https://www.imdb.com/chart/top/")  
page
```

```
## {html_document}  
## <html xmlns:og="http://ogp.me/ns#" xmlns:fb="http://www.facebook.com/2008/fbml">  
## [1] <head>\n<meta http-equiv="Content-Type" content="text/html" ...  
## [2] <body id="styleguide-v2" class="fixed">\n                <img ...
```



# A webpage in R

- Result is a list with 2 elements

```
typeof(page)
```

```
## [1] "list"
```



# A webpage in R

- Result is a list with 2 elements

```
typeof(page)
```

```
## [1] "list"
```

- that we need to convert to something more familiar, like a data frame....

```
class(page)
```

```
## [1] "xml_document" "xml_node"
```



# Step 2. Scrape movie titles and save as titles



# Scrape movie titles

The screenshot shows a web browser displaying the [IMDb Top 250 - IMDb](https://www.imdb.com/chart/top/) page. The main content is the "Top Rated Movies" chart, showing the top 250 movies as rated by IMDb users. The table includes columns for Rank & Title, IMDB Rating, and Your Rating. The first four rows are highlighted with green boxes, corresponding to the movie titles listed in the code below. The developer tools' element inspector is overlaid on the page, with the title ".titleColumn a" selected for the first row. The right sidebar shows "You Have Seen" statistics and links to other IMDb Charts like Box Office, Most Popular Movies, and Top Rated TV.

| Rank & Title                                       | IMDb Rating | Your Rating |
|--|-------------|-------------|
| 1. <a href="#">The Shawshank Redemption</a> (1994) | ★ 9.2       | ☆           |
| 2. <a href="#">The Godfather</a> (1972)            | ★ 9.1       | ☆           |
| 3. <a href="#">The Godfather: Part II</a> (1974)   | ★ 9.0       | ☆           |
| 4. <a href="#">The Dark Knight</a> (2008)          | ★ 9.0       | ☆           |
| 5. <a href="#">12 Angry Men</a> (1957)             |             |             |

# Scrape the nodes

```
page %>%  
  html_nodes(".titleColumn a")
```

```
## {xml_nodeset (250)}  
## [1] <a href="/title/tt0111161/?pf_rd_m=A2FGELUUNOQJNL&pf_...>  
## [2] <a href="/title/tt0068646/?pf_rd_m=A2FGELUUNOQJNL&pf_...>  
## [3] <a href="/title/tt0071562/?pf_rd_m=A2FGELUUNOQJNL&pf_...>  
## [4] <a href="/title/tt0468569/?pf_rd_m=A2FGELUUNOQJNL&pf_...>  
## [5] <a href="/title/tt0050083/?pf_rd_m=A2FGELUUNOQJNL&pf_...>  
## [6] <a href="/title/tt0108052/?pf_rd_m=A2FGELUUNOQJNL&pf_...>  
## [7] <a href="/title/tt0167260/?pf_rd_m=A2FGELUUNOQJNL&pf_...>  
## [8] <a href="/title/tt0110912/?pf_rd_m=A2FGELUUNOQJNL&pf_...>  
## [9] <a href="/title/tt0060196/?pf_rd_m=A2FGELUUNOQJNL&pf_...>  
## [10] <a href="/title/tt0120737/?pf_rd_m=A2FGELUUNOQJNL&pf_...>  
## [11] <a href="/title/tt0137523/?pf_rd_m=A2FGELUUNOQJNL&pf_...>  
## [12] <a href="/title/tt0109830/?pf_rd_m=A2FGELUUNOQJNL&pf_...>  
## [13] <a href="/title/tt1375666/?pf_rd_m=A2FGELUUNOQJNL&pf_...>  
## [14] <a href="/title/tt0167261/?pf_rd_m=A2FGELUUNOQJNL&pf_...>  
## [15] <a href="/title/tt0080684/?pf_rd_m=A2FGELUUNOQJNL&pf_...>  
## [16] <a href="/title/tt0133093/?pf_rd_m=A2FGELUUNOQJNL&pf_...>  
...>
```

The screenshot shows the IMDb Top Rated Movies chart. The page title is "IMDb Charts" and the section title is "Top Rated Movies". It displays the top 250 movies as rated by IMDb users. The results are sorted by ranking. The first four movies listed are:

| Rank | Title                    | Year   | IMDb Rating | Your Rating |
|------|--------------------------|--------|-------------|-------------|
| 1.   | The Shawshank Redemption | (1994) | 9.2         | ...         |
| 2.   | The Godfather            | (1972) | 9.1         | ...         |
| 3.   | The Godfather: Part II   | (1974) | 9.0         | ...         |
| 4.   | The Dark Knight          | (2008) | 9.0         | ...         |

The search bar at the bottom of the page contains the query ".titleColumn a". The right sidebar includes sections for "You Have Seen" (0/250), "IMDb Charts" (Box Office, Most Popular Movies, Top Rated Movies, etc.), and "Top Rated Movies by Gen" (Action, Adventure, Animation).

# Extract the text from the nodes

```
page %>%  
  html_nodes(".titleColumn a") %>%  
  html_text()
```

```
## [1] "The Shawshank Redemption"  
## [2] "The Godfather"  
## [3] "The Godfather: Part II"  
## [4] "The Dark Knight"  
## [5] "12 Angry Men"  
## [6] "Schindler's List"  
## [7] "The Lord of the Rings: The Return of the King"  
## [8] "Pulp Fiction"  
## [9] "The Good, the Bad and the Ugly"  
## [10] "The Lord of the Rings: The Fellowship of the Ring"  
## [11] "Fight Club"  
## [12] "Forrest Gump"  
## [13] "Inception"  
## [14] "The Lord of the Rings: The Two Towers"  
## [15] "Star Wars: Episode V - The Empire Strikes Back"  
## [16] "The Matrix"  
...  
...
```

The screenshot shows a web browser window displaying the 'IMDb Charts' page for 'Top Rated Movies'. The table lists the top 250 movies, with columns for Rank & Title, IMDb Rating, and Your Rating. The movie 'The Shawshank Redemption' is highlighted with a red box in the first row. The browser's developer tools are open, showing the selected element as '.titleColumn a'. The right sidebar shows 'You Have Seen' and 'IMDb Charts' sections.

| Rank & Title                                       | IMDb Rating | Your Rating |
|--|-------------|-------------|
| 1. <a href="#">The Shawshank Redemption</a> (1994) | 9.2         |             |
| 2. <a href="#">The Godfather</a> (1972)            | 9.1         |             |
| 3. <a href="#">The Godfather: Part II</a> (1974)   | 9.0         |             |
| 4. <a href="#">The Dark Knight</a> (2008)          | 9.0         |             |
| 5. <a href="#">12 Angry Men</a> (1957)             | 9.0         |             |

# Save as titles

```
titles <- page %>%
  html_nodes(".titleColumn a") %>%
  html_text()

titles

## [1] "The Shawshank Redemption"
## [2] "The Godfather"
## [3] "The Godfather: Part II"
## [4] "The Dark Knight"
## [5] "12 Angry Men"
## [6] "Schindler's List"
## [7] "The Lord of the Rings: The Return of the King"
## [8] "Pulp Fiction"
## [9] "The Good, the Bad and the Ugly"
## [10] "The Lord of the Rings: The Fellowship of the Ring"
## [11] "Fight Club"
## [12] "Forrest Gump"
## [13] "Inception"
## [14] "The Lord of the Rings: The Two Towers"
...
...
```

The screenshot shows a web browser displaying the IMDb Top 250 chart at [imdb.com/chart/top](https://imdb.com/chart/top). The page lists the top 250 movies as rated by IMDb users. The title 'The Shawshank Redemption' is highlighted with a red box. The chart includes columns for Rank & Title, IMDb Rating, and Your Rating. To the right, there are sections for 'You Have Seen' (0/250), 'IMDb Charts' (Box Office, Most Popular Movies, Top Rated Movies, etc.), and 'Top Rated Movies by Gen' (Action, Adventure, Animation). A search bar at the bottom contains the XPath expression '.titleColumn a'.

| Rank & Title                                       | IMDb Rating | Your Rating |
|--|-------------|-------------|
| 1. <a href="#">The Shawshank Redemption</a> (1994) | ★ 9.2       | ☆           |
| 2. <a href="#">The Godfather</a> (1972)            | ★ 9.1       | ☆           |
| 3. <a href="#">The Godfather: Part II</a> (1974)   | ★ 9.0       | ☆           |
| 4. <a href="#">The Dark Knight</a> (2008)          | ★ 9.0       | ☆           |
| ...  |             |             |
| <a href="#">12 Angry Men</a> (1957)                | ★ 9.0       | ☆           |

# Step 3. Scrape year movies were made and save as years



# Scrape years movies were made in

The screenshot shows a browser window displaying the IMDb Top Rated Movies chart. The page title is "IMDb Charts" and the section title is "Top Rated Movies". It says "Showing 250 Titles" and "Sort by: Ranking". The main content is a table with columns for Rank & Title, IMDB Rating, and Your Rating. The first four rows are:

| Rank & Title                       | IMDB Rating | Your Rating |
|------------------------------------|-------------|-------------|
| 1. The Shawshank Redemption (1994) | ★ 9.2       | ☆           |
| 2. The Godfather (1972)            | ★ 9.1       | ☆           |
| 3. The Godfather: Part II (1974)   | ★ 9.0       | ☆           |
| 4. The Dark Knight (2008)          | ★ 9.0       | ☆           |

A red box highlights the year "1994" in the first row. The developer tools' element inspector is open over the same cell, showing the full HTML structure: <td><span>(1994)</span></td>. The sidebar on the right shows "You Have Seen 0/250 (0%)" and a list of other IMDb Charts categories.

# Scrape the nodes

```
page %>%  
  html_nodes(".secondaryInfo")
```

```
## {xml_nodeset (250)}  
## [1] <span class="secondaryInfo">(1994)</span>  
## [2] <span class="secondaryInfo">(1972)</span>  
## [3] <span class="secondaryInfo">(1974)</span>  
## [4] <span class="secondaryInfo">(2008)</span>  
## [5] <span class="secondaryInfo">(1957)</span>  
## [6] <span class="secondaryInfo">(1993)</span>  
## [7] <span class="secondaryInfo">(2003)</span>  
## [8] <span class="secondaryInfo">(1994)</span>  
## [9] <span class="secondaryInfo">(1966)</span>  
## [10] <span class="secondaryInfo">(2001)</span>  
## [11] <span class="secondaryInfo">(1999)</span>  
## [12] <span class="secondaryInfo">(1994)</span>  
## [13] <span class="secondaryInfo">(2010)</span>  
## [14] <span class="secondaryInfo">(2002)</span>  
## [15] <span class="secondaryInfo">(1980)</span>  
## [16] <span class="secondaryInfo">(1999)</span>  
...  
.
```

The screenshot shows a web browser displaying the 'Top Rated Movies' section of the IMDb Top 250 chart. The page title is 'IMDb Charts' and the sub-section is 'Top Rated Movies'. The chart lists the top 250 movies based on IMDb users' ratings. The first movie, 'The Shawshank Redemption', has its year '(1994)' highlighted with a red box. The interface includes a sidebar on the right titled 'You Have Seen' showing '0/250 (0%)' and a 'SHARE' button. Below the chart, there are sections for 'IMDb Charts', 'Box Office', and 'Top Rated Movies by Gen'. A search bar at the bottom contains the query '.secondaryInfo'.

# Extract the text from the nodes

```
page %>%  
  html_nodes(".secondaryInfo") %>%  
  html_text()
```

```
## [1] "(1994)" "(1972)" "(1974)" "(2008)" "(1957)"  
## [7] "(2003)" "(1994)" "(1966)" "(2001)" "(1999)"  
## [13] "(2010)" "(2002)" "(1980)" "(1999)" "(1990)"  
## [19] "(1954)" "(1995)" "(1991)" "(2002)" "(1997)"  
## [25] "(1977)" "(1998)" "(2014)" "(2001)" "(1999)"  
## [31] "(1994)" "(1962)" "(2002)" "(1991)" "(1995)"  
## [37] "(1960)" "(1994)" "(1936)" "(1998)" "(1988)"  
## [43] "(2014)" "(2000)" "(2006)" "(2011)" "(2006)"  
## [49] "(1968)" "(1954)" "(1988)" "(1979)" "(1979)"  
## [55] "(1981)" "(1940)" "(2006)" "(2012)" "(1957)"  
## [61] "(2008)" "(2018)" "(1957)" "(1980)" "(2018)" "(1964)"  
## [67] "(1997)" "(2003)" "(2019)" "(2016)" "(2017)" "(2012)"  
## [73] "(1986)" "(1984)" "(2018)" "(2019)" "(1981)" "(1963)"  
## [79] "(2009)" "(1999)" "(1995)" "(1984)" "(1995)" "(2009)"  
## [85] "(2020)" "(1997)" "(1983)" "(1968)" "(1992)" "(1931)"  
## [91] "(1958)" "(2007)" "(1985)" "(1941)" "(2012)" "(2000)"  
...
```

The screenshot shows the IMDb Top Rated Movies page. The page title is "IMDb Charts" and the section title is "Top Rated Movies". It displays the top 250 movies as rated by IMDb users. The first four movies listed are: 1. The Shawshank Redemption (1994) with a rating of 9.2; 2. The Godfather (1972) with a rating of 9.1; 3. The Godfather: Part II (1974) with a rating of 9.0; and 4. The Dark Knight (2008) with a rating of 9.0. A red box highlights the ".secondaryInfo" node in the bottom right corner of the page. The URL in the browser address bar is "imdb.com/chart/top".

# Clean up the text

We need to go from "(1994)" to 1994:

- Remove ( and ): string manipulation
- Convert to numeric: `as.numeric()`



# stringr

- **stringr** provides a cohesive set of functions designed to make working with strings as easy as possible
- Functions in stringr start with `str_*`( ), e.g.
  - `str_remove()` to remove a pattern from a string

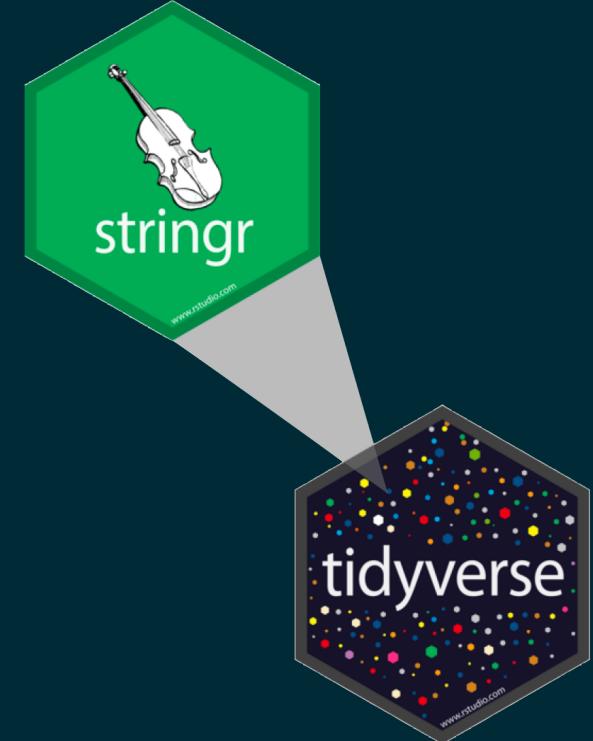
```
str_remove(string = "jello", pattern = "el")
```

```
## [1] "jlo"
```

- `str_replace()` to replace a pattern with another

```
str_replace(string = "jello", pattern = "j", replacement = "h")
```

```
## [1] "hello"
```



# Clean up the text

```
page %>%
  html_nodes(".secondaryInfo") %>%
  html_text() %>%
  str_remove("\\"(") # remove (
```

```
## [1] "1994)" "1972)" "1974)" "2008)" "1957)" "1993)" "2003)"
## [8] "1994)" "1966)" "2001)" "1999)" "1994)" "2010)" "2002)"
## [15] "1980)" "1999)" "1990)" "1975)" "1954)" "1995)" "1991)"
## [22] "2002)" "1997)" "1946)" "1977)" "1998)" "2014)" "2001)"
## [29] "1999)" "2019)" "1994)" "1962)" "2002)" "1991)" "1995)"
## [36] "1985)" "1960)" "1994)" "1936)" "1998)" "1988)" "1931)"
## [43] "2014)" "2000)" "2006)" "2011)" "2006)" "1942)" "1968)"
## [50] "1954)" "1988)" "1979)" "1979)" "2000)" "1981)" "1940)"
## [57] "2006)" "2012)" "1957)" "1950)" "2008)" "2018)" "1957)"
## [64] "1980)" "2018)" "1964)" "1997)" "2003)" "2019)" "2016)"
## [71] "2017)" "2012)" "1986)" "1984)" "2018)" "2019)" "1981)"
## [78] "1963)" "2009)" "1999)" "1995)" "1984)" "1995)" "2009)"
## [85] "2020)" "1997)" "1983)" "1968)" "1992)" "1931)" "1958)"
## [92] "2007)" "1985)" "1941)" "2012)" "2000)" "1952)" "1959)"
## [99] "2004)" "1948)" "1952)" "1962)" "1921)" "1987)" "2016)"
...
```



# Clean up the text

```
page %>%
  html_nodes(".secondaryInfo") %>%
  html_text() %>%
  str_remove("\\\\(") %>% # remove (
  str_remove("\\\\)") # remove )
```

```
## [1] "1994" "1972" "1974" "2008" "1957" "1993" "2003" "1994"
## [9] "1966" "2001" "1999" "1994" "2010" "2002" "1980" "1999"
## [17] "1990" "1975" "1954" "1995" "1991" "2002" "1997" "1946"
## [25] "1977" "1998" "2014" "2001" "1999" "2019" "1994" "1962"
## [33] "2002" "1991" "1995" "1985" "1960" "1994" "1936" "1998"
## [41] "1988" "1931" "2014" "2000" "2006" "2011" "2006" "1942"
## [49] "1968" "1954" "1988" "1979" "1979" "2000" "1981" "1940"
## [57] "2006" "2012" "1957" "1950" "2008" "2018" "1957" "1980"
## [65] "2018" "1964" "1997" "2003" "2019" "2016" "2017" "2012"
## [73] "1986" "1984" "2018" "2019" "1981" "1963" "2009" "1999"
## [81] "1995" "1984" "1995" "2009" "2020" "1997" "1983" "1968"
## [89] "1992" "1931" "1958" "2007" "1985" "1941" "2012" "2000"
## [97] "1952" "1959" "2004" "1948" "1952" "1962" "1921" "1987"
## [105] "2016" "1960" "2010" "2020" "1944" "1971" "1927" "1976"
```

...



# Convert to numeric

```
page %>%  
  html_nodes(".secondaryInfo") %>%  
  html_text() %>%  
  str_remove("\\(") %>% # remove (  
  str_remove("\\)") %>% # remove )  
  as.numeric()
```

```
## [1] 1994 1972 1974 2008 1957 1993 2003 1994 1966 2001 1999 1994  
## [13] 2010 2002 1980 1999 1990 1975 1954 1995 1991 2002 1997 1946  
## [25] 1977 1998 2014 2001 1999 2019 1994 1962 2002 1991 1995 1985  
## [37] 1960 1994 1936 1998 1988 1931 2014 2000 2006 2011 2006 1942  
## [49] 1968 1954 1988 1979 1979 2000 1981 1940 2006 2012 1957 1950  
## [61] 2008 2018 1957 1980 2018 1964 1997 2003 2019 2016 2017 2012  
## [73] 1986 1984 2018 2019 1981 1963 2009 1999 1995 1984 1995 2009  
## [85] 2020 1997 1983 1968 1992 1931 1958 2007 1985 1941 2012 2000  
## [97] 1952 1959 2004 1948 1952 1962 1921 1987 2016 1960 2010 2020  
## [109] 1944 1971 1927 1976 2011 1955 1973 1983 2000 2019 2001 1962  
## [121] 2010 1965 2009 1989 1995 1997 1961 1985 1988 1950 2018 2004  
## [133] 1975 2021 1950 1959 2005 1992 1997 2004 2013 1961 1963 2007  
## [145] 1995 1948 2006 2001 2009 1980 1975 1974 1988 1998 2010 1925  
...
```



# Save as years

```
years <- page %>%
  html_nodes(".secondaryInfo") %>%
  html_text() %>%
  str_remove("\\(") %>% # remove (
  str_remove("\\)") %>% # remove )
  as.numeric()
```

```
years
```

```
## [1] 1994 1972 1974 2008 1957 1993 2003 1994 1966
## [13] 2010 2002 1980 1999 1990 1975 1954 1995 1991
## [25] 1977 1998 2014 2001 1999 2019 1994 1962 2002
## [37] 1960 1994 1936 1998 1988 1931 2014 2000 2006
## [49] 1968 1954 1988 1979 1979 2000 1981 1940 2006
## [61] 2008 2018 1957 1980 2018 1964 1997 2003 2019
## [73] 1986 1984 2018 2019 1981 1963 2009 1999 1995
## [85] 1984 2020 1997 1983 1968 1992 1931 1958 2007
## [97] 1995 1952 1959 2004 1948 1952 1962 1921 1987
## [109] 2004 1944 1971 1927 1976 2011 1955 1973 1983
## [121] 2009 2010 1965 2009 1989 1995 1997 1961 1985
...  
[121] 2004 1944 1971 1927 1976 2011 1955 1973 1983
[121] 2009 2010 1965 2009 1989 1995 1997 1961 1985
```

The screenshot shows a browser window displaying the IMDb Top Rated Movies chart. The chart lists the top 250 movies based on user ratings. The first four entries are:

| Rank & Title                       | IMDb Rating | Your Rating |
|------------------------------------|-------------|-------------|
| 1. The Shawshank Redemption (1994) | 9.2         |             |
| 2. The Godfather (1972)            | 9.1         |             |
| 3. The Godfather: Part II (1974)   | 9.0         |             |
| 4. The Dark Knight (2008)          | 9.0         |             |

The interface includes a sidebar with links to other charts like 'Top Rated English Movies' and 'Top Rated Indian Movies'. A search bar at the top right says 'Search IMDb'.

# Step 4. Scrape IMDB ratings and save as ratings



# Scrape IMDB ratings

The screenshot shows a web browser window displaying the [IMDb Top 250 - IMDb](https://www.imdb.com/chart/top/) page. The main content is titled "Top Rated Movies" and lists the top 250 movies as rated by IMDb users. The table includes columns for Rank & Title, IMDB Rating, and Your Rating. The first four rows of the table are highlighted, showing the following data:

| Rank & Title                       | IMDB Rating | Your Rating |
|------------------------------------|-------------|-------------|
| 1. The Shawshank Redemption (1994) | 9.2         |             |
| 2. The Godfather (1972)            | 9.1         |             |
| 3. The Godfather: Part II (1974)   | 9.0         |             |
| 4. The Dark Knight (2008)          | 9.0         |             |

A red box highlights the "9.2" rating for "The Shawshank Redemption". On the right side of the page, there is a sidebar titled "You Have Seen" showing "0/250 (0%)". Below that is a list of "IMDb Charts" including Box Office, Most Popular Movies, Top Rated Movies (which is selected), Top Rated English Movies, Most Popular TV, Top Rated TV, Top Rated Indian Movies, and Lowest Rated Movies. At the bottom of the sidebar, there is a section for "Top Rated Movies by Gen" with links for Action, Adventure, Animation, and Crime. A developer tool's inspection panel is visible at the bottom, showing the element for the "9.2" rating.

# Scrape the nodes

```
page %>%  
  html_nodes("strong")
```

```
## {xml_nodeset (250)}  
## [1] <strong title="9.2 based on 2,498,614 user ratings">9.2</strong>  
## [2] <strong title="9.1 based on 1,722,841 user ratings">9.1</strong>  
## [3] <strong title="9.0 based on 1,195,759 user ratings">9.0</strong>  
## [4] <strong title="9.0 based on 2,448,983 user ratings">9.0</strong>  
## [5] <strong title="8.9 based on 738,054 user ratings">8.9</strong>  
## [6] <strong title="8.9 based on 1,279,061 user ratings">8.9</strong>  
## [7] <strong title="8.9 based on 1,726,847 user ratings">8.9</strong>  
## [8] <strong title="8.8 based on 1,928,854 user ratings">8.8</strong>  
## [9] <strong title="8.8 based on 724,318 user ratings">8.8</strong>  
## [10] <strong title="8.8 based on 1,748,127 user ratings">8.8</strong>  
## [11] <strong title="8.8 based on 1,965,915 user ratings">8.8</strong>  
## [12] <strong title="8.7 based on 1,928,857 user ratings">8.7</strong>  
## [13] <strong title="8.7 based on 2,197,508 user ratings">8.7</strong>  
## [14] <strong title="8.7 based on 1,561,025 user ratings">8.7</strong>  
## [15] <strong title="8.7 based on 1,216,407 user ratings">8.7</strong>  
## [16] <strong title="8.6 based on 1,782,159 user ratings">8.6</strong>  
...  
...
```

The screenshot shows the IMDb Top Rated Movies chart. The top four entries are displayed:

| Rank & Title                       | IMDb Rating | Your Rating |
|------------------------------------|-------------|-------------|
| 1. The Shawshank Redemption (1994) | 9.2         | ☆           |
| 2. The Godfather (1972)            | 9.1         | ☆           |
| 3. The Godfather: Part II (1974)   | 9.0         | ☆           |
| 4. The Dark Knight (2008)          | 9.0         | ☆           |

A search bar at the bottom contains the text "strong".



# Extract the text from the nodes

```
page %>%  
  html_nodes("strong") %>%  
  html_text()
```

```
## [1] "9.2" "9.1" "9.0" "9.0" "8.9" "8.9" "8.9" "8.9" "8.9"  
## [11] "8.8" "8.7" "8.7" "8.7" "8.7" "8.6" "8.6" "8.6" "8.6"  
## [21] "8.6" "8.6" "8.6" "8.6" "8.6" "8.6" "8.6" "8.5" "8.5"  
## [31] "8.5" "8.5" "8.5" "8.5" "8.5" "8.5" "8.5" "8.5" "8.5"  
## [41] "8.5" "8.5" "8.5" "8.5" "8.5" "8.5" "8.5" "8.5" "8.5"  
## [51] "8.4" "8.4" "8.4" "8.4" "8.4" "8.4" "8.4" "8.4" "8.4"  
## [61] "8.4" "8.4" "8.4" "8.4" "8.4" "8.4" "8.3" "8.3" "8.3"  
## [71] "8.3" "8.3" "8.3" "8.3" "8.3" "8.3" "8.3" "8.3" "8.3"  
## [81] "8.3" "8.3" "8.3" "8.3" "8.3" "8.3" "8.3" "8.3" "8.3"  
## [91] "8.3" "8.3" "8.3" "8.3" "8.3" "8.3" "8.3" "8.3" "8.3"  
## [101] "8.3" "8.3" "8.2" "8.2" "8.2" "8.2" "8.2" "8.2" "8.2"  
## [111] "8.2" "8.2" "8.2" "8.2" "8.2" "8.2" "8.2" "8.2" "8.2"  
## [121] "8.2" "8.2" "8.2" "8.2" "8.2" "8.2" "8.2" "8.2" "8.2"  
## [131] "8.2" "8.2" "8.2" "8.2" "8.2" "8.2" "8.2" "8.2" "8.2"  
## [141] "8.2" "8.2" "8.2" "8.2" "8.2" "8.2" "8.1" "8.1" "8.1"  
## [151] "8.1" "8.1" "8.1" "8.1" "8.1" "8.1" "8.1" "8.1" "8.1"  
...
```

The screenshot shows the IMDb Top Rated Movies chart. The top four results are displayed:

| Rank & Title                       | IMDb Rating | Your Rating |
|------------------------------------|-------------|-------------|
| 1. The Shawshank Redemption (1994) | 9.2         | ☆           |
| 2. The Godfather (1972)            | 9.1         | ☆           |
| 3. The Godfather: Part II (1974)   | 9.0         | ☆           |
| 4. The Dark Knight (2008)          | 9.0         | ☆           |

A red box highlights the rating '9.2' for The Shawshank Redemption. The search bar at the bottom contains the word 'strong'.

# Convert to numeric

```
page %>%  
  html_nodes("strong") %>%  
  html_text() %>%  
  as.numeric()
```

```
## [1] 9.2 9.1 9.0 9.0 8.9 8.9 8.9 8.8 8.8 8.8 8.8 8  
## [16] 8.6 8.6 8.6 8.6 8.6 8.6 8.6 8.6 8.6 8.6 8.6 8  
## [31] 8.5 8.5 8.5 8.5 8.5 8.5 8.5 8.5 8.5 8.5 8.5 8  
## [46] 8.5 8.5 8.4 8.4 8.4 8.4 8.4 8.4 8.4 8.4 8.4 8  
## [61] 8.4 8.4 8.4 8.4 8.4 8.4 8.3 8.3 8.3 8.3 8.3 8  
## [76] 8.3 8.3 8.3 8.3 8.3 8.3 8.3 8.3 8.3 8.3 8.3 8  
## [91] 8.3 8.3 8.3 8.3 8.3 8.3 8.3 8.3 8.3 8.3 8.3 8  
## [106] 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8  
## [121] 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8  
## [136] 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.1 8.1  
## [151] 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1  
## [166] 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1  
## [181] 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1  
## [196] 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1  
## [211] 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.0 8.0 8.0  
...
```

The screenshot shows a browser window displaying the IMDb Top 250 chart. The page title is "IMDb Charts" and the sub-section is "Top Rated Movies". The chart lists the top 250 movies based on IMDb users' ratings. The first four entries are:

| Rank & Title                       | IMDb Rating | Your Rating |
|------------------------------------|-------------|-------------|
| 1. The Shawshank Redemption (1994) | 9.2         |             |
| 2. The Godfather (1972)            | 9.1         |             |
| 3. The Godfather: Part II (1974)   | 9.0         |             |
| 4. The Dark Knight (2008)          | 9.0         |             |

A red box highlights the rating "9.2" for "The Shawshank Redemption". The search bar at the bottom contains the word "strong". On the right side of the page, there are sidebar links for "You Have Seen" (0/250), "IMDb Charts", "Box Office", "Most Popular Movies", "Top Rated Movies", "Top Rated English Movies", "Most Popular TV", "Top Rated TV", "Top Rated Indian Movies", "Lowest Rated Movies", and "Top Rated Movies by Gen" (Action, Adventure, Animation).

# Save as ratings

```
ratings <- page %>%
  html_nodes("strong") %>%
  html_text() %>%
  as.numeric()

ratings
```

```
## [1] 9.2 9.1 9.0 9.0 8.9 8.9 8.9 8.8 8.8 8.8 8.8 8
## [16] 8.6 8.6 8.6 8.6 8.6 8.6 8.6 8.6 8.6 8.6 8.6 8
## [31] 8.5 8.5 8.5 8.5 8.5 8.5 8.5 8.5 8.5 8.5 8.5 8
## [46] 8.5 8.5 8.4 8.4 8.4 8.4 8.4 8.4 8.4 8.4 8.4 8
## [61] 8.4 8.4 8.4 8.4 8.4 8.4 8.3 8.3 8.3 8.3 8.3 8
## [76] 8.3 8.3 8.3 8.3 8.3 8.3 8.3 8.3 8.3 8.3 8.3 8
## [91] 8.3 8.3 8.3 8.3 8.3 8.3 8.3 8.3 8.3 8.3 8.3 8
## [106] 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2
## [121] 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2
## [136] 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.1 8.1
## [151] 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1
## [166] 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1
...
...
```

The screenshot shows the IMDb Top 250 chart on a Mac OS X desktop. The chart lists the top 250 movies rated by IMDb users. The columns are Rank & Title, IMDb Rating, and Your Rating. The first four rows are highlighted with yellow boxes around the 'strong' rating values. The 'Your Rating' column for these rows shows a red 'strong' button instead of a numerical rating.

| Rank & Title                       | IMDb Rating | Your Rating |
|------------------------------------|-------------|-------------|
| 1. The Shawshank Redemption (1994) | 9.2         | strong      |
| 2. The Godfather (1972)            | 9.1         | strong      |
| 3. The Godfather: Part II (1974)   | 9.0         | strong      |
| 4. The Dark Knight (2008)          | 9.0         | strong      |
| ...                                | 8.2         | 8.2         |
| 5. 12 Angry Men (1957)             | 8.2         | 8.2         |

# Step 5. Create a data frame called imdb\_top\_250



# Create a data frame: `imdb_top_250`

```
imdb_top_250 <- tibble(  
  title = titles,  
  year = years,  
  rating = ratings  
)  
  
imdb_top_250
```

```
## # A tibble: 250 x 3  
##   title                 year  rating  
##   <chr>                <dbl>  <dbl>  
## 1 The Shawshank Redemption 1994    9.2  
## 2 The Godfather           1972    9.1  
## 3 The Godfather: Part II 1974     9  
## 4 The Dark Knight        2008     9  
## 5 12 Angry Men          1957    8.9  
## 6 Schindler's List        1993    8.9  
## # ... with 244 more rows
```



[Previous](#)

1

2

3

4

5

...

32

[Next](#)

|   | <b>title</b>                                  | ♦ | <b>year</b> ♦ | <b>rating</b> ♦ |
|---|---|---|---------------|-----------------|
| 1 | The Shawshank Redemption                      |   | 1994          | 9.2             |
| 2 | The Godfather                                 |   | 1972          | 9.1             |
| 3 | The Godfather: Part II                        |   | 1974          | 9               |
| 4 | The Dark Knight                               |   | 2008          | 9               |
| 5 | 12 Angry Men                                  |   | 1957          | 8.9             |
| 6 | Schindler's List                              |   | 1993          | 8.9             |
| 7 | The Lord of the Rings: The Return of the King |   | 2003          | 8.9             |
| 8 | Pulp Fiction                                  |   | 1994          | 8.8             |



# Clean up / enhance

May or may not be a lot of work depending on how messy the data are

- See if you like what you got:

```
glimpse(imdb_top_250)
```

```
## Rows: 250
## Columns: 3
## $ title  <chr> "The Shawshank Redemption", "The Godfather", "Th~
## $ year   <dbl> 1994, 1972, 1974, 2008, 1957, 1993, 2003, 1994, ~
## $ rating <dbl> 9.2, 9.1, 9.0, 9.0, 8.9, 8.9, 8.9, 8.8, 8.8, 8.8~
```

- Add a variable for rank

```
imdb_top_250 <- imdb_top_250 %>%
  mutate(rank = 1:nrow(imdb_top_250)) %>%
  relocate(rank)
```



```
## # A tibble: 250 x 4
##   rank title                               year rating
##   <int> <chr>
## 1     1 The Shawshank Redemption           1994    9.2
## 2     2 The Godfather                      1972    9.1
## 3     3 The Godfather: Part II             1974    9.0
## 4     4 The Dark Knight                   2008    9.0
## 5     5 12 Angry Men                     1957    8.9
## 6     6 Schindler's List                  1993    8.9
## 7     7 The Lord of the Rings: The Return of the King 2003    8.9
## 8     8 Pulp Fiction                    1994    8.8
## 9     9 The Good, the Bad and the Ugly    1966    8.8
## 10    10 The Lord of the Rings: The Fellowship of the Ring 2001    8.8
## 11    11 Fight Club                       1999    8.8
## 12    12 Forrest Gump                   1994    8.7
## 13    13 Inception                      2010    8.7
## 14    14 The Lord of the Rings: The Two Towers    2002    8.7
## 15    15 Star Wars: Episode V - The Empire Strikes Back 1980    8.7
## 16    16 The Matrix                      1999    8.6
## 17    17 Goodfellas                     1990    8.6
## 18    18 One Flew Over the Cuckoo's Nest    1975    8.6
## 19    19 Seven Samurai                  1954    8.6
## 20    20 Se7en                         1995    8.6
## # ... with 230 more rows
```



# What next?



[datasciencebox.org](http://datasciencebox.org)

Which years have the most movies on the list?



## Which years have the most movies on the list?

```
imdb_top_250 %>%  
  count(year, sort = TRUE)
```

```
## # A tibble: 86 x 2  
##   year     n  
##   <dbl> <int>  
## 1 1995     8  
## 2 1957     7  
## 3 1994     6  
## 4 1997     6  
## 5 2000     6  
## 6 2004     6  
## # ... with 80 more rows
```



Which 1995 movies made the list?



[datasciencebox.org](http://datasciencebox.org)

## Which 1995 movies made the list?

```
imdb_top_250 %>%  
  filter(year == 1995) %>%  
  print(n = 8)
```

```
## # A tibble: 8 x 4  
##   rank title           year rating  
##   <int> <chr>          <dbl>  <dbl>  
## 1    20 Se7en            1995    8.6  
## 2    35 The Usual Suspects  1995    8.5  
## 3    81 Toy Story         1995    8.3  
## 4    83 Braveheart        1995    8.3  
## 5   125 Heat              1995    8.2  
## 6   145 Casino             1995    8.2  
## 7   194 Before Sunrise    1995    8.1  
## 8   224 La Haine           1995    8
```



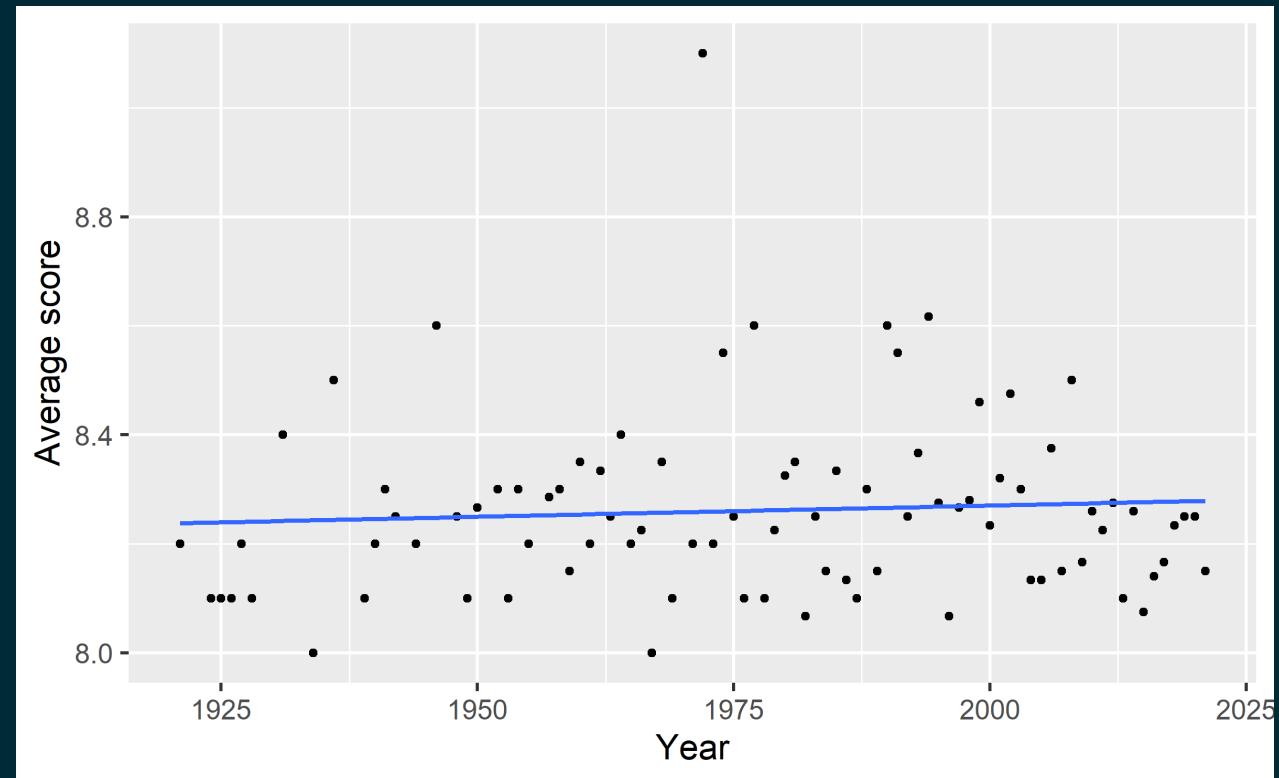
Visualize the average yearly rating for movies that made it on the top 250 list over time.



Visualize the average yearly rating for movies that made it on the top 250 list over time.

Plot

Code



Visualize the average yearly rating for movies that made it on the top 250 list over time.

Plot

Code

```
imdb_top_250 %>%
  group_by(year) %>%
  summarise(avg_score = mean(rating)) %>%
  ggplot(aes(y = avg_score, x = year)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Year", y = "Average score")
```



# "Can you?" vs "Should you?"

## Researchers just released profile data on 70,000 OkCupid users without permission

By Brian Resnick | @B\_resnick | brian@vox.com | May 12, 2016, 6:00pm EDT

A group of researchers has released a data set on nearly 70,000 users of the online dating site OkCupid. The data dump breaks the cardinal rule of social science research ethics: **It took identifiable personal data without permission.**

The information — while publicly available to OkCupid users — was collected by Danish researchers who never contacted OkCupid or its clientele about using it.

The data, collected from November 2014 to March 2015, includes user names, ages, gender, religion, and personality traits, as well as answers to the personal questions the site asks to help match potential mates. The users hail from a few dozen countries around the world.

The data dump did not reveal anyone's real name. But it's entirely possible to use clues from a user's location, demographics, and OkCupid user name to determine their identity.

If your OkC username is one you've used anywhere else, I now know your sexual preferences & kinks, your answers to thousands of questions.

— Scott B. Weingart (@scott\_bot) May 11, 2016

Source: Brian Resnick, Researchers just released profile data on 70,000 OkCupid users without permission, Vox.



# "Can you?" vs "Should you?"

**Emil OW Kirkegaard @KirkegaardEmil · May 8**  
The OKCupid paper has now been submitted. This means that the dataset is now public! Enjoy! :) [openpsych.net/forum/showthre...](http://openpsych.net/forum/showthread.php?1134-OKCupid-dataset-submitted)

**Ethan Jewett @esjewett · May 11**  
@KirkegaardEmil This data set is highly re-identifiable. Even includes usernames? Was any work at all done to anonymize it?

**Emil OW Kirkegaard**  
@KirkegaardEmil

@esjewett No. Data is already public.



# Challenges



[datasciencebox.org](http://datasciencebox.org)

# Unreliable formatting at the source

Screenshot of a Gumtree search results page for "Used Cars, Vans & Motorbikes" in Edinburgh.

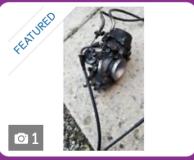
**Category:**

- < All Categories
- Motors
  - Cars 1,414
  - Parts 563
  - Accessories 505
  - Motorbikes & Scooters 164
  - Vans 142
  - Other Vehicles 47
  - Wanted 24
  - Campervans & Motorhomes 13
  - Caravans 8
  - Plant & Tractors 1

**Filters:**

Other options

- Urgent ads 3
- Feature ads 37
- Ads with pictures 2,793
- Search title & description

|  |  |
|--|--|
|    | <b>Suzuki DR650 SE OEM BST40 Carb</b> may fit KTM 640 LC4<br>Corstorphine, Edinburgh<br>Have upgraded my 1998 DR650SE with the TM40 so surplus is my OEM BST40 comes with throttle cables choke also mic...<br><b>£155</b><br>29 days ago  |
|    | <b>2006 Nissan Micra 1.2 Sport 5dr HATCHBACK Petrol M...</b><br>Joppa, Edinburgh<br>2006, NISSAN, MICRA, 1.2 Sport 5dr, 5 Doors, HATCHBACK, Grey, 1895GBP, Petrol, 1240, 60000 V5 Registration Docume...<br><b>£1,895</b><br>5 mins ago  |
|    | <b>Volvo, 960, Saloon, 1996, Automatic, 2922 (cc), 4 doors</b><br>Liberton, Edinburgh<br>This Volvo 960 is quite simply a lovely car; built when Volvos were real Volvos with a build quality, which in my opinion is second ...<br><b>1996   83,860 miles   Petrol   2,922 cc</b><br><b>£3,000</b><br>1 day ago |
|  | <b>Skoda octavia 2.0TDI 150 DSG</b><br>Easter Road, Edinburgh<br>Skoda octavia 2016 (66) 2.0 TDI 150 DSG 76 000 miles • Euro 6 • automatic gearbox 6-gear • Sat nav • Half-Leather • Bluetooth...<br><b>2016   76,000 miles   Diesel   1,968 cc</b><br><b>£7,500</b><br>63 days ago                              |

# Data broken into many pages

The screenshot shows a web browser displaying a search results page from [yelp.co.uk](https://yelp.co.uk/search?find_desc=Vegetarian&find_loc=Edinburgh&ns=1). The search parameters are set to "Vegetarian" and "Edinburgh". The results are filtered by "Restaurants". The page displays two cards for businesses:

- 9. The Edinburgh Larder**: 4.5 stars, 248 reviews. Category: Delis, Coffee & Tea Shops. Image: A plate of food including what looks like sausages and vegetables.
- 10. Hanam's**: 4.5 stars, 62 reviews. Category: Middle Eastern. Image: A plate of Middle Eastern cuisine, possibly kebabs or shawarma.

Below the cards is a navigation bar with numbers 1 through 9, followed by a right arrow, indicating there are 24 pages in total. The current page is highlighted with a purple border. To the right of the cards is a map of Edinburgh showing the locations of the top 10 restaurants. Each location is marked with a red circle containing a number from 1 to 10, corresponding to the ranking in the list.

# Workflow



[datasciencebox.org](http://datasciencebox.org)

# Screen scraping vs. APIs

Two different scenarios for web scraping:

- Screen scraping: extract data from source code of website, with html parser (easy) or regular expression matching (less easy)
- Web APIs (application programming interface): website offers a set of structured http requests that return JSON or XML files



# A new R workflow

- When working in an R Markdown document, your analysis is re-run each time you knit
- If web scraping in an R Markdown document, you'd be re-scraping the data each time you knit, which is undesirable (and not *nice*)!
- An alternative workflow:
  - Use an R script to save your code
  - Saving interim data scraped using the code in the script as CSV or RDS files
  - Use the saved data in your analysis in your R Markdown document

