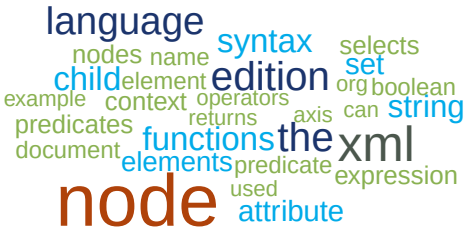
Introduction to Web Scraping with R

Scraping HTML Tables



Simon Munzert | IPSDS

HTML tables are everywhere

Purchased Equipments (June, 2006)			
Item Num#	Item Picture	Item Description	Price
		Shipping Handling, Installation, etc	Expense
1.		IBM Clone Computer.	$ 400.00
		Shipping Handling, Installation, etc	$ 20.00
2.		1GB RAM Module for Computer.	$ 50.00
		Shipping Handling, Installation, etc	$ 14.00
Purchased Equipments (June, 2006)			

DATES	POLLSTER	GRADE	SAMPLE	WEIGHT	APPROVE	DISAPPROVE	ADJUSTED	
• DEC. 28-30	Gallup	B-	1,500 A	1.03	40%	55%	41%	53%
• DEC. 26-28	Rasmussen Reports/Pulse Opinion Research	C+	1,500 LV	0.85	45%	53%	40%	53%
• DEC. 24-28	Ipsos	A-	1,519 A	2.01	37%	58%	37%	57%
• DEC. 23-27	Gallup	B-	1,500 A	0.58	38%	56%	39%	54%
• DEC. 24-26	YouGov	B	1,500 A	1.13	38%	52%	39%	55%

Built	Building	City	Country	Roof		Floors	Pinnacle		Current status
1870	Equitable Life Building	New York City	United States	043 m	142 ft	8			Destroyed by fire in 1912
1889	Auditorium Building	Chicago		082 m	269 ft	17	106 m	349 ft	Standing
1890	New York World Building	New York City		094 m	309 ft	20	106 m	349 ft	Demolished in 1955
1894	Philadelphia City Hall	Philadelphia		155.8 m	511 ft	9	167 m	548 ft	Standing
1908	Singer Building	New York City		187 m	612 ft	47			Demolished in 1968
1909	Met Life Tower			213 m	700 ft	50			Standing
1913	Woolworth Building			241 m	792 ft	57			Standing
1930	40 Wall Street					70	283 m	927 ft	Standing
1930	Chrysler Building			282.9 m	927 ft	77	319 m	1,046 ft	Standing
1931	Empire State Building			381 m	1,250 ft	102	443 m	1,454 ft	Standing
1972	World Trade Center (North Tower)			417 m	1,368 ft	110	527.3 m	1,730 ft	Destroyed in 2001 in the September 11 attacks
1974	Willis Tower (formerly Sears Tower)	Chicago		442 m	1,450 ft	108	527 m	1,729 ft	Standing
1996	Petronas Towers	Kuala Lumpur	Malaysia	379 m	1,242 ft	88	452 m	1,483 ft	Standing
2004	Taipei 101	Taipei	Taiwan	449 m	1,474 ft	101	509 m	1,671 ft	Standing
2010	Burj Khalifa	Dubai	United Arab Emirates	828 m	2,717 ft	163	829.8 m	2,722 ft	Standing

Recall: HTML tables share the same basic syntax

Syntax

- HTML tables are easy to spot in the wild
- just look for `<table>` tags

Example

```
<table>
  <tr> <th>Rank</th> <th>Nominal GDP</th> <th>Name</th> </tr>
  <tr> <th></th> <th>(per capita, USD)</th> <th></th> </tr>
  <tr> <td>1</td> <td>170,373</td> <td>Lichtenstein</td> </tr>
  <tr> <td>2</td> <td>167,021</td> <td>Monaco</td> </tr>
  <tr> <td>3</td> <td>115,377</td> <td>Luxembourg</td> </tr>
  <tr> <td>4</td> <td>98,565</td> <td>Norway</td> </tr>
  <tr> <td>5</td> <td>92,682</td> <td>Qatar</td> </tr>
</table>
```

Scraping HTML tables in R

Scraping HTML tables in R

A neat `rvest` solution

- exactly because scraping tables is an easy and repetitive task, there is a dedicated function for it (literally from the documentation):

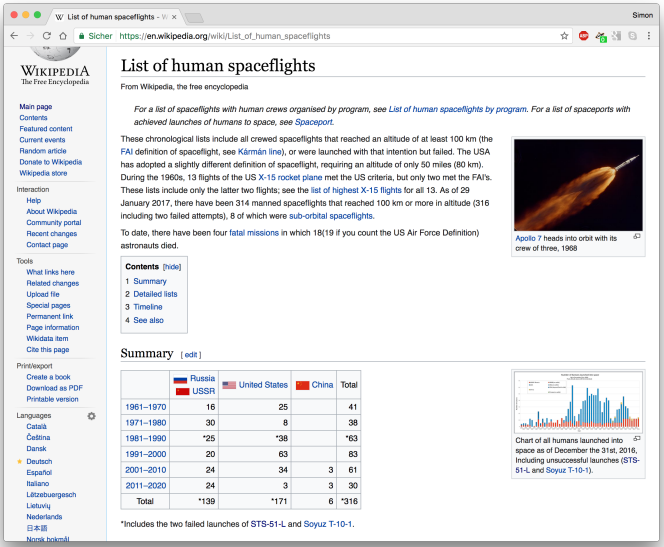
```
html_table(x, header = NA, trim = TRUE, fill = FALSE, dec = ".")
```

`x`	A node, node set or document
`header`	Use first row as header? If `NA`, will use first row if it consists of `<th>` tags
`trim`	Remove leading and trailing whitespace within each cell?
`fill`	If `TRUE`, automatically fill rows with fewer than the maximum number of columns with `NA`s
`dec`	The character used as decimal mark

Example

Example

- source: https://en.wikipedia.org/wiki/List_of_human_spaceflights



List of human spaceflights - W

Simon

Sicher https://en.wikipedia.org/wiki/List_of_human_spaceflights

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book
Download as PDF
Printable version

Languages
Català
Čeština
Dansk
Deutsch
Español
Italiano
Lëtzebuergesch
Lietuvių
Nederlands
日本語
Norsk bokmål

List of human spaceflights

From Wikipedia, the free encyclopedia

For a list of spaceflights with human crews organised by program, see List of human spaceflights by program. For a list of spaceports with achieved launches of humans to space, see Spaceport.

These chronological lists include all crewed spaceflights that reached an altitude of at least 100 km (the FAI definition of spaceflight, see Kármán line), or were launched with that intention but failed. The USA has adopted a slightly different definition of spaceflight, requiring an altitude of only 50 miles (80 km). During the 1960s, 13 flights of the US X-15 rocket plane met the US criteria, but only two met the FAI's. These lists include only the latter two flights; see the list of highest X-15 flights for all 13. As of 29 January 2017, there have been 314 manned spaceflights that reached 100 km or more in altitude (316 including two failed attempts), 8 of which were sub-orbital spaceflights.

To date, there have been four fatal missions in which 18(19 if you count the US Air Force Definition) astronauts died.

Apollo 7 heads into orbit with its crew of three, 1968

Contents [hide]
1 Summary
2 Detailed lists
3 Timeline
4 See also

Summary [edit]

	Russia / USSR	United States	China	Total
1961–1970	16	25		41
1971–1980	30	8		38
1981–1990	*25	*38		*63
1991–2000	20	63		83
2001–2010	24	34	3	61
2011–2020	24	3	3	30
Total	*139	*171	6	*316

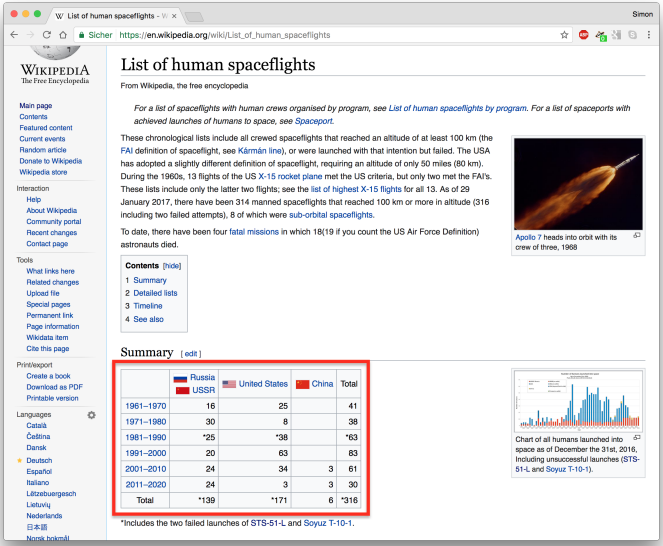
*Includes the two failed launches of STS-51-L and Soyuz T-10-1.

Chart of all humans launched into space as of December the 31st, 2016, Including unsuccessful launches (STS-51-L and Soyuz T-10-1).

Example

Example

- source: https://en.wikipedia.org/wiki/List_of_human_spaceflights
- not entirely clean table: some cells empty, images, links



List of human spaceflights - W

Simon

Sicher https://en.wikipedia.org/wiki/List_of_human_spaceflights

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book
Download as PDF
Printable version

Languages
Català
Čeština
Dansk
Deutsch
Español
Italiano
Lëtzebuergesch
Lietuvių
Nederlands
日本語
Norsk bokmål

List of human spaceflights

From Wikipedia, the free encyclopedia

For a list of spaceflights with human crews organised by program, see List of human spaceflights by program. For a list of spaceports with achieved launches of humans to space, see Spaceport.

These chronological lists include all crewed spaceflights that reached an altitude of at least 100 km (the FAI definition of spaceflight, see Kármán line), or were launched with that intention but failed. The USA has adopted a slightly different definition of spaceflight, requiring an altitude of only 50 miles (80 km). During the 1960s, 13 flights of the X-15 rocket plane met the US criteria, but only two met the FAI's. These lists include only the latter two flights; see the list of highest X-15 flights for all 13. As of 29 January 2017, there have been 314 manned spaceflights that reached 100 km or more in altitude (316 including two failed attempts), 8 of which were sub-orbital spaceflights.

To date, there have been four fatal missions in which 18(19 if you count the US Air Force Definition) astronauts died.

Apollo 7 heads into orbit with its crew of three, 1968

Contents [hide]
1 Summary
2 Detailed lists
3 Timeline
4 See also

Summary [edit]

	Russia / USSR	United States	China	Total
1961–1970	16	25		41
1971–1980	30	8		38
1981–1990	*25	*38		*63
1991–2000	20	63		83
2001–2010	24	34	3	61
2011–2020	24	3	3	30
Total	*139	*171	6	*316

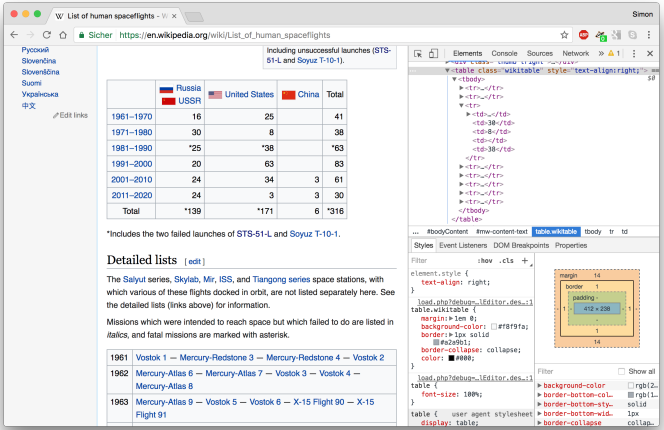
*Includes the two failed launches of STS-51-L and Soyuz T-10-1.

Chart of all humans launched into space as of December the 31st, 2016, Including unsuccessful launches (STS-51-L and Soyuz T-10-1).

Example

Example

- source: https://en.wikipedia.org/wiki/List_of_human_spaceflights
- not entirely clean table: some cells empty, images, links
- HTML code is straightforward



Example

Example

- source: https://en.wikipedia.org/wiki/List_of_human_spaceflights
- not entirely clean table: some cells empty, images, links
- HTML code is straightforward

```
▼<table class="wikitable" style="text-align:right;">
  ▼<tbody>
    ▶<tr>…</tr>
    ▶<tr>…</tr>
    ▼<tr>
      ▶<td>…</td>
        <td>30</td>
        <td>8</td>
        <td></td>
        <td>38</td>
      </tr>
    ▶<tr>…</tr>
    ▶<tr>…</tr>
    ▶<tr>…</tr>
    ▶<tr>…</tr>
    ▶<tr>…</tr>
  </tbody>
</table>
```

Example

Scraping the table with R

R code

```r
library(rvest)
url_p <- read_html("https://en.wikipedia.org/wiki/List_of_human_spaceflights")
tables <- html_table(url_p, header = TRUE, fill = TRUE)
spaceflights <- tables[[1]]
spaceflights
```

```
            Russia Soviet Union United States China Total
1 1961-1970                 16            25    NA    41
2 1971-1980                 30             8    NA    38
3 1981-1990                *25           *38    NA   *63
4 1991-2000                 20            63    NA    83
5 2001-2010                 24            34     3    61
6 2011-2020                 28             3     3    34
7     Total                *143          *171     6  *320
```

end

Summary

Summary

- scraping HTML tables with R is straightforward
- the high-level `html_table()` function is easy to use and generally robust
- sometimes, HTML tables are a bit more complex. You might encounter:
  - tables where cells span multiple rows
  - tables where headers are not in the first row
- in such cases, you might want to check out the `htmltab` package, which is more complex to use but also provides more flexbility