1. Scraping Tabular Data

In [1]:

```
# load the library (install the library if it's not already installed)
library(rvest)
```

In [8]:

```
# specify the URL of the page you want to scrape
url <- "https://www.the-numbers.com/movie/budgets/all"
```

In [9]:

```
# read the HTML content of the page
html_content <- read_html(url)
```

In [10]:

```
# use the html_nodes() function to extract specific elements, such as divs or tables, using CSS selectors
data <- html_content %>%
  html_nodes("table") %>%
  html_table()
```

In [11]:

```
# the extracted data will be stored as a data frame
# you can inspect the data frame using the head() function
head(data)
```

A tibble: 100 × 6

1.

	ReleaseDate	Movie	ProductionBudget	DomesticGross	WorldwideGross
<int>	<chr>	<chr>	<chr>	<chr>	<chr>
1	Dec 9, 2022	Avatar: The Way of Water	$460,000,000	$641,726,731	$2,175,722,357
2	Apr 23, 2019	Avengers: Endgame	$400,000,000	$858,373,000	$2,794,731,755
3	May 20, 2011	Pirates of the Caribbean: On Stranger Tides	$379,000,000	$241,071,802	$1,045,713,802
4	Apr 22, 2015	Avengers: Age of Ultron	$365,000,000	$459,005,868	$1,395,316,979
5	May 17, 2023	Fast X	$340,000,000	$0	$0
6	Dec 16, 2015	Star Wars Ep. VII: The Force Awakens	$306,000,000	$936,662,225	$2,064,615,817
7	Apr 25, 2018	Avengers: Infinity War	$300,000,000	$678,815,482	$2,048,359,754
8	May 24, 2007	Pirates of the Caribbean: At World's End	$300,000,000	$309,420,425	$960,996,492
9	Nov 13, 2017	Justice League	$300,000,000	$229,024,295	$655,945,209
10	Oct 6, 2015	Spectre	$300,000,000	$200,074,175	$879,077,344
11	Jul 12, 2023	Mission: Impossible Dead Reckoning Part One	$290,000,000	$0	$0
12	Dec 18, 2019	Star Wars: The Rise of Skywalker	$275,000,000	$515,202,542	$1,072,767,997
13	May 23, 2018	Solo: A Star Wars Story	$275,000,000	$213,767,512	$393,151,347
14	Mar 7, 2012	John Carter	$263,700,000	$73,058,679	$282,778,100
15	Mar 23, 2016	Batman v Superman: Dawn of Justice	$263,000,000	$330,360,194	$872,395,091
16	Dec 13, 2017	Star Wars Ep. VIII: The Last Jedi	$262,000,000	$620,181,382	$1,331,635,141
17	Jul 11, 2019	The Lion King	$260,000,000	$543,638,043	$1,647,733,638
18	Nov 24, 2010	Tangled	$260,000,000	$200,821,936	$583,777,242
19	May 4, 2007	Spider-Man 3	$258,000,000	$336,530,303	$894,860,230
20	Apr 22, 2016	Captain America: Civil War	$250,000,000	$408,084,349	$1,151,899,586
21	Jul 1, 2022	Thor: Love and Thunder	$250,000,000	$343,256,830	$760,928,081
22	Jul 15, 2009	Harry Potter and the Half-Blood Prince	$250,000,000	$302,089,278	$929,411,069
23	Dec 12, 2013	The Hobbit: The Desolation of Smaug	$250,000,000	$258,241,522	$959,358,436
24	Dec 10, 2014	The Hobbit: The Battle of the Five Armies	$250,000,000	$255,119,788	$940,323,039
25	Apr 7, 2017	The Fate of the Furious	$250,000,000	$225,764,765	$1,236,703,796
26	Sep 29, 2021	No Time to Die	$250,000,000	$160,891,007	$759,959,662
27	Dec 17, 2009	Avatar	$237,000,000	$785,221,649	$2,899,384,102
28	Jun 28, 2006	Superman Returns	$232,000,000	$200,120,000	$391,081,192
29	Jul 19, 2012	The Dark Knight Rises	$230,000,000	$448,139,099	$1,082,228,107
30	May 23, 2017	Pirates of the Caribbean: Dead Men Tell No Tales	$230,000,000	$172,558,876	$794,861,794
⋮	⋮	⋮	⋮	⋮	⋮
71	Jul 31, 2019	Fast & Furious Presents: Hobbs & Shaw	$200,000,000	$173,956,935	$760,732,926
72	May 20, 2021	F9: The Fast Saga	$200,000,000	$173,005,945	$720,752,238
73	Dec 17, 2010	Tron: Legacy	$200,000,000	$172,062,763	$399,866,199
74	Oct 19, 2022	Black Adam	$200,000,000	$168,152,111	$391,261,706
75	Nov 12, 2009	2012	$200,000,000	$166,112,167	$757,677,748
76	Nov 3, 2021	Eternals	$200,000,000	$164,870,264	$401,731,759
77	Nov 14, 2018	Fantastic Beasts: The Crimes of Grindelwald	$200,000,000	$159,555,901	$648,455,339
78	May 21, 2009	Terminator Salvation	$200,000,000	$125,322,469	$365,491,792
79	Jun 15, 2022	Lightyear	$200,000,000	$118,307,188	$218,768,299
80	Jul 28, 2021	Jungle Cruise	$200,000,000	$116,987,516	$210,469,803
81	Jun 17, 2011	Green Lantern	$200,000,000	$116,601,172	$219,535,492
82	Apr 6, 2022	Fantastic Beasts: The Secrets of Dumbledore	$200,000,000	$95,850,844	$404,560,145
83	May 28, 2010	Prince of Persia: Sands of Time	$200,000,000	$90,759,676	$336,359,676
84	Jun 5, 2019	Dark Phoenix	$200,000,000	$65,845,974	$246,356,895
85	Feb 28, 2020	Onward	$200,000,000	$61,555,145	$133,357,601
86	Dec 16, 2020	Wonder Woman 1984	$200,000,000	$46,801,036	$166,360,232
87	Sep 4, 2020	Mulan	$200,000,000	$0	$69,973,540
88	Jul 2, 2021	The Tomorrow War	$200,000,000	$0	$19,220,000
89	Jul 13, 2022	The Gray Man	$200,000,000	$0	$451,178
90	Jun 29, 2011	Transformers: Dark of the Moon	$195,000,000	$352,390,543	$1,123,794,079
91	Jun 2, 2017	The Mummy	$195,000,000	$80,101,125	$409,953,905
92	Feb 27, 2013	Jack the Giant Slayer	$195,000,000	$65,187,603	$197,687,603
93	Apr 1, 2015	Furious 7	$190,000,000	$353,007,020	$1,514,553,486
94	May 16, 2013	Star Trek Into Darkness	$190,000,000	$228,778,661	$467,381,584
95	Jun 19, 2013	World War Z	$190,000,000	$202,706,711	$531,861,650

	ReleaseDate	Movie	ProductionBudget	DomesticGross	WorldwideGross
<int>	<chr>	<chr>	<chr>	<chr>	<chr>
96	May 10, 2013	The Great Gatsby	$190,000,000	$144,840,419	$353,640,419
97	Nov 6, 2009	Disney's A Christmas Carol	$190,000,000	$137,855,863	$315,709,697
98	Jul 11, 2013	Pacific Rim	$190,000,000	$101,802,906	$411,002,906
99	Dec 16, 2021	The Matrix Resurrections	$190,000,000	$40,463,197	$159,197,755
100	Nov 25, 2015	The Good Dinosaur	$187,500,000	$123,087,120	$333,771,037

In this example, the html_nodes() function is used with the argument "table" to extract all tables from the HTML content of the page. The html_table() function is then used to convert the extracted HTML tables into R data frames.

Note that you may need to modify the CSS selector depending on the structure of the table on the webpage you're trying to scrape. You can inspect the HTML source code of the page to determine the correct CSS selector to use.

---

2. Scraping Non-Tabular Data

In [13]:

```
# Scrape the webpage
webpage <- read_html("https://archive.org/details/internetarchivebooks")
```

In [14]:

```
# Extract the text data
text_data <- html_text(html_nodes(webpage, "p"))
```

In [15]:

```
# Store the text data in a data frame
data_frame <- data.frame(text = text_data)
```

In [16]:

```
# the extracted data will be stored as a data frame
# you can inspect the data frame using the head() function
head(data_frame)
```

A data.frame: 6 × 1

	text
	<chr>
1	Due to a planned power outage on Friday, 1/14, between 8am-1pm PST, some services may be impacted.
2	Search the history of over 780 billion web pages on the Internet.
3	Capture a web page as it appears now for use as a trusted citation in the future.
4	Please enter a valid web address
5	Books contributed by the Internet Archive.
6	Total Views 203,325,781 (Older Stats)

---

3. Data Sources We Can Consider

https://archive.org/ (https://archive.org/) **<- CHECK THIS OUT**

https://data.world/arcadeanalytics/best-500-albums-amazon-neptune (https://data.world/arcadeanalytics/best-500-albums-amazon-neptune)

https://en.wikipedia.org/wiki/List_of_online_music_databases (https://en.wikipedia.org/wiki/List_of_online_music_databases)
