STAT 118: Notes P

Webscraping Text with `rvest`

Emily Malcolm-White

```
#LOAD PACKAGES
library(tidyverse)
library(rvest)
```

Webscraping Text... and sometimes making it into tables

Let's look at the top 100 feature films released in 2020 and 2021 available here: https://www.imdb.com/search/title/?count=100&release_date=2020,2020&title_type=feature

```
URL <- read_html("https://www.imdb.com/search/title/?count=100&release_date=2020,2021&titl
```

Scrape IMBD for the titles of the 100 most popular feature films in 2020.

```
title_data <- URL %>%
  html_elements(".lister-item-header a") %>%
  html_text()

title_data
```

```
[1] "Ghostbusters: Afterlife"
[2] "Dune"
[3] "Tenet"
[4] "Spider-Man: No Way Home"
[5] "Bull"
```

```
 [6] "The Suicide Squad"
 [7] "Ammonite"
 [8] "Nobody"
 [9] "The Matrix Resurrections"
[10] "Promising Young Woman"
[11] "No Time to Die"
[12] "365 Days"
[13] "Don't Look Up"
[14] "Old"
[15] "Fatale"
[16] "Red Notice"
[17] "Sing 2"
[18] "F9: The Fast Saga"
[19] "The Hunt"
[20] "A Quiet Place Part II"
[21] "Extraction"
[22] "Zack Snyder's Justice League"
[23] "The French Dispatch"
[24] "Cinderella"
[25] "West Side Story"
[26] "Eternals"
[27] "Encanto"
[28] "After We Collided"
[29] "Girl in the Basement"
[30] "Free Guy"
[31] "CODA"
[32] "Pleasure"
[33] "The Voyeurs"
[34] "Wonder Woman 1984"
[35] "Last Night in Soho"
[36] "Vacation Friends"
[37] "Venom: Let There Be Carnage"
[38] "Wrath of Man"
[39] "Licorice Pizza"
[40] "The Worst Person in the World"
[41] "The Black Phone"
[42] "Black Widow"
[43] "The Last Duel"
[44] "Birds of Prey"
[45] "The King's Man"
[46] "The Little Things"
[47] "Run Hide Fight"
[48] "Shang-Chi and the Legend of the Ten Rings"
```

```
[49] "The Fallout"
[50] "Bill & Ted Face the Music"
[51] "The Devil All the Time"
[52] "Hamilton"
[53] "Cruella"
[54] "Army of the Dead"
[55] "House of Gucci"
[56] "Nightmare Alley"
[57] "The Power of the Dog"
[58] "The Many Saints of Newark"
[59] "The Father"
[60] "Army of Thieves"
[61] "Mortal Kombat"
[62] "The Old Guard"
[63] "Minari"
[64] "Occupation: Rainfall"
[65] "Malignant"
[66] "The Guilty"
[67] "Another Round"
[68] "The Tomorrow War"
[69] "Soul"
[70] "The Invisible Man"
[71] "Godzilla vs. Kong"
[72] "Greyhound"
[73] "Benedetta"
[74] "Palm Springs"
[75] "I'm Thinking of Ending Things"
[76] "Enola Holmes"
[77] "After We Fell"
[78] "The Conjuring: The Devil Made Me Do It"
[79] "The Empty Man"
[80] "Mulan"
[81] "Spiral"
[82] "Luca"
[83] "Greenland"
[84] "Sonic the Hedgehog"
[85] "Shiva Baby"
[86] "The Sadness"
[87] "King Richard"
[88] "Poppy"
[89] "Halloween Kills"
[90] "Pieces of a Woman"
[91] "Don't Breathe 2"
```

```
 [92] "Raya and the Last Dragon"
 [93] "The Harder They Fall"
 [94] "Werewolves Within"
 [95] "The Green Knight"
 [96] "Jungle Cruise"
 [97] "The Last Letter from Your Lover"
 [98] "Belfast"
 [99] "We Can Be Heroes"
[100] "Underwater"
```

Scrape IMBD for the runtime of the 100 most popular feature films in 2020.

```
library(readr)
# need this package for parse_number()
```

Figure 1: Artwork by @allisonhorst

```
runtime_data <- URL %>%
  html_nodes(".text-muted .runtime") %>%
  html_text() %>%
  parse_number() %>% #this picks out only the numbers (and drops characters, in this case,
  as.numeric()

runtime_data
```

```
 [1] 124 155 150 148  88 132 117  92 148 113 163 114 138 108 102 118 110 143
[19]  90  97 116 242 107 113 156 156 102 105  88 115 111 109 116 151 116 103
[37]  97 119 133 128 103 134 152 109 131 128 109 132  96  91 138 160 134 148
[55] 158 150 126 120  97 127 110 125 115 128 111  90 117 138 100 124 113  91
[73] 131  90 134 123  98 112 137 115  93  95 119  99  77  99 144  98 105 126
[91]  98 107 139  97 130 127 110  98 100  95
```

Scrape IMBD for the ratings of the 100 most popular feature films in 2020

```
rating_data <- URL %>%
  html_elements(".ratings-imdb-rating strong") %>%
  html_text() %>%
  as.numeric()

rating_data
```

```
 [1] 7.1 8.0 7.3 8.2 6.5 7.2 6.5 7.4 5.7 7.5 7.3 3.3 7.2 5.8 5.4 6.3 7.4 5.2
[19] 6.5 7.2 6.8 7.9 7.1 4.4 7.2 6.3 7.2 5.0 6.3 7.1 8.0 6.4 6.0 5.4 7.0 6.3
[37] 5.9 7.1 7.1 7.8 6.9 6.7 7.4 6.1 6.3 6.3 6.3 7.4 7.0 5.9 7.1 8.3 7.3 5.8
[55] 6.6 7.0 6.8 6.3 8.2 6.4 6.1 6.7 7.4 4.7 6.2 6.3 7.7 6.5 8.0 7.1 6.3 7.0
[73] 6.7 7.4 6.6 6.6 4.7 6.3 6.2 5.8 5.2 7.4 6.4 6.5 7.1 6.4 7.5 7.3 5.5 7.0
[91] 6.0 7.3 6.6 6.0 6.6 6.6 6.7 7.3 4.7 5.9
```

Scrape IMBD for the number of votes of the 100 most popular feature films in 2020

```
votes_data <- URL %>%
  html_elements(".sort-num_votes-visible span:nth-child(2)") %>%
  html_text() %>%
  parse_number() %>%
  as.numeric()

votes_data
```

```
 [1] 198010 709968 555276 826008   6403 383945  21157 302111 265931 193635
[11] 427055  94889 568880 146362  13338 298774  78517 153670 122544 256405
[21] 249149 422272 140465  44477  90368 367477 243548  35529  11431 400132
[31] 151526  20419  26484 282635 157034  25314 245745 195224 131595  84389
[41] 173792 406059 169919 254735 162209 114048  25896 413094  29962  51152
[51] 145521 105148 250181 180715 149667 156177 186099  59456 175703  84208
[61] 183652 176623  89969  15096 101200 139314 179647 218123 354833 242273
[71] 224246 106404  23078 173023  94927 208431  17770 128880  34479 154925
[81]  60461 179540 126347 151693  27331  16275 126038     79  92837  53833
[91]  67002 163067  68396  22875 110048 203851  20394  82876  16349  89759
```

Scrape IMBD for the gross earnings of the 100 most popular feature films in 2020

```
gross_data <- URL %>%
  html_elements(".ghost~ .text-muted+ span") %>%
  html_text() %>%
```

```
  parse_number() %>%
  as.numeric()

gross_data
```

```
 [1] 129.36 108.33  58.46 804.75  55.82  27.27 160.87  48.28 162.79 173.20
[11]   5.81 160.07 164.87  96.09   2.39 121.63  46.37 213.55  27.47  90.12
[21] 183.65  10.85  84.16  37.18 224.54  86.10   1.00  53.81  42.20  13.39
[31]  70.41 100.92  65.63 148.97  92.00  32.64  54.72 116.99  17.29
```

```
#Notes this one has less entries than the rest and we can't figure out which one goes with
```

We can combine all this data into one data frame:

```
movies<-data.frame(Title = title_data,
Runtime = runtime_data,
Rating = rating_data
)


movies
```

```
                                    Title Runtime Rating
1                   Ghostbusters: Afterlife     124    7.1
2                                    Dune     155    8.0
3                                   Tenet     150    7.3
4                  Spider-Man: No Way Home     148    8.2
5                                    Bull      88    6.5
6                         The Suicide Squad     132    7.2
7                                 Ammonite     117    6.5
8                                   Nobody      92    7.4
9                  The Matrix Resurrections     148    5.7
10                   Promising Young Woman     113    7.5
11                          No Time to Die     163    7.3
12                                365 Days     114    3.3
13                         Don't Look Up     138    7.2
14                                   Old     108    5.8
15                                Fatale     102    5.4
16                           Red Notice     118    6.3
17                               Sing 2     110    7.4
```

18	F9: The Fast Saga	143	5.2
19	The Hunt	90	6.5
20	A Quiet Place Part II	97	7.2
21	Extraction	116	6.8
22	Zack Snyder's Justice League	242	7.9
23	The French Dispatch	107	7.1
24	Cinderella	113	4.4
25	West Side Story	156	7.2
26	Eternals	156	6.3
27	Encanto	102	7.2
28	After We Collided	105	5.0
29	Girl in the Basement	88	6.3
30	Free Guy	115	7.1
31	CODA	111	8.0
32	Pleasure	109	6.4
33	The Voyeurs	116	6.0
34	Wonder Woman 1984	151	5.4
35	Last Night in Soho	116	7.0
36	Vacation Friends	103	6.3
37	Venom: Let There Be Carnage	97	5.9
38	Wrath of Man	119	7.1
39	Licorice Pizza	133	7.1
40	The Worst Person in the World	128	7.8
41	The Black Phone	103	6.9
42	Black Widow	134	6.7
43	The Last Duel	152	7.4
44	Birds of Prey	109	6.1
45	The King's Man	131	6.3
46	The Little Things	128	6.3
47	Run Hide Fight	109	6.3
48	Shang-Chi and the Legend of the Ten Rings	132	7.4
49	The Fallout	96	7.0
50	Bill & Ted Face the Music	91	5.9
51	The Devil All the Time	138	7.1
52	Hamilton	160	8.3
53	Cruella	134	7.3
54	Army of the Dead	148	5.8
55	House of Gucci	158	6.6
56	Nightmare Alley	150	7.0
57	The Power of the Dog	126	6.8
58	The Many Saints of Newark	120	6.3
59	The Father	97	8.2
60	Army of Thieves	127	6.4

61	Mortal Kombat	110	6.1
62	The Old Guard	125	6.7
63	Minari	115	7.4
64	Occupation: Rainfall	128	4.7
65	Malignant	111	6.2
66	The Guilty	90	6.3
67	Another Round	117	7.7
68	The Tomorrow War	138	6.5
69	Soul	100	8.0
70	The Invisible Man	124	7.1
71	Godzilla vs. Kong	113	6.3
72	Greyhound	91	7.0
73	Benedetta	131	6.7
74	Palm Springs	90	7.4
75	I'm Thinking of Ending Things	134	6.6
76	Enola Holmes	123	6.6
77	After We Fell	98	4.7
78	The Conjuring: The Devil Made Me Do It	112	6.3
79	The Empty Man	137	6.2
80	Mulan	115	5.8
81	Spiral	93	5.2
82	Luca	95	7.4
83	Greenland	119	6.4
84	Sonic the Hedgehog	99	6.5
85	Shiva Baby	77	7.1
86	The Sadness	99	6.4
87	King Richard	144	7.5
88	Poppy	98	7.3
89	Halloween Kills	105	5.5
90	Pieces of a Woman	126	7.0
91	Don't Breathe 2	98	6.0
92	Raya and the Last Dragon	107	7.3
93	The Harder They Fall	139	6.6
94	Werewolves Within	97	6.0
95	The Green Knight	130	6.6
96	Jungle Cruise	127	6.6
97	The Last Letter from Your Lover	110	6.7
98	Belfast	98	7.3
99	We Can Be Heroes	100	4.7
100	Underwater	95	5.9

Make a list OR Make a plot!

```
ggplot(movies, aes(x=runtime_data, y=rating_data)) +
  geom_point()
```