

Lecture 20: Web scraping and data wrangling

Last time: Taskmaster data

1 https://taskmaster.fandom.com/wiki/Series_11

TASKMASTER WIKI						
EXPLORE						
SERIES 16						
PREVIOUS SERIES						
INTERNATIONAL / OTHER						
ARTICLES						
Spoilers! This section contains spoilers.						
Task	Description	Charlotte Ritchie	Jamali Maddix	Lee Mack	Mike Wozniak	Sarah Kendall
Episode 1: It's not your fault. (18 March 2021)						
1	Prize: Best thing you can carry, but only just.	1	2	4	5	3
2	Do the most impressive thing under the table with one hand. You must be looking at the camera and waving with your other hand throughout your impressive thing.	2	3 ^[1]	3	5	4
3	Catch the rat. You must be at least three meters from the rat when you catch it. The rat will run across the red green in 30 minutes.	DQ	1	5	3 ^[2]	3 ^[2]
4	Deliver all the plates to Alex. When carrying the plates you must be traveling by scooter, or bicycle or hoverboard. For every plate fallen	2	1	5	3 ^[2]	4 ^[2]

Scraping the tabular data

```
1 read_html("https://taskmaster.fandom.com/wiki/Series_11") |>
2   html_element(".tmtable") |>
3   html_table()
```

```
# A tibble: 75 × 7
```

	Task	Description	`Charlotte Ritchie`	`Jamali Maddix`
	`Lee Mack`			
	<chr>	<chr>	<chr>	<chr>
	<chr>			
1	Episode 1: It's n...	Episode 1:...	Episode 1: It's no...	Episode 1: It'...
	Episode 1...			
2	1	Prize: Bes...	1	2
3	2	Do the mos...	2	3[1]
4	3	Catch the ...	DQ	1
5	4	Deliver al...	2	1
6	5	Live: Stac...	0	0
7	Total	Total	5	7
8	Episode 2: The Lu...	Episode 2:...	Episode 2: The Lur...	Episode 2: The...

Wrangling

Here's what we have so far:

	Task	Description	`Charlotte Ritchie`	`Jamali Maddix`	`I
2	Episode 1: It's...	Episode 1:...	Episode 1: It's no...	Episode 1: It'...	Ep:
3	1	Prize: Bes...	1	2	4
4	2	Do the mos...	2	3[1]	3
5	3	Catch the ...	DQ	1	5
6	4	Deliver al...	2	1	5
7	5	Live: Stac...	0	0	0
8	Total	Total	5	7	17
9	Episode 2: The ...	Episode 2:...	Episode 2: The Lur...	Episode 2: The...	Ep:
10	1	Prize: Bes...	5	1	2
11	2	Make the b...	0	5	0

What changes do you think we should make to this format?

Wrangling

What we ultimately want:

	Task	Description	episode	episode_name	air_date	contestant	score
1							
2	1 1	Prize: Best th...	1	"It's not y...	18 Marc...	Charlotte...	1
3	2 1	Prize: Best th...	1	"It's not y...	18 Marc...	Jamali Ma...	2
4	3 1	Prize: Best th...	1	"It's not y...	18 Marc...	Lee Mack	4
5	4 1	Prize: Best th...	1	"It's not y...	18 Marc...	Mike Wozn...	5
6	5 1	Prize: Best th...	1	"It's not y...	18 Marc...	Sarah Ken...	3
7	6 2	Do the most im...	1	"It's not y...	18 Marc...	Charlotte...	2
8	7 2	Do the most im...	1	"It's not y...	18 Marc...	Jamali Ma...	3[1
9	8 2	Do the most im...	1	"It's not y...	18 Marc...	Lee Mack	3
10	9 2	Do the most im...	1	"It's not y...	18 Marc...	Mike Wozn...	5
11	10 2	Do the most im...	1	"It's not y...	18 Marc...	Sarah Ken...	4

colnames: Task, Description, episode, episode_name,
air_date, contestant, score, series

Wrangling

Intermediate step:

1	Task	Description	episode	contestant	score	series
2	1	Prize: Best thing...	Episode 1...	Charlotte...	1	11
3	1	Prize: Best thing...	Episode 1...	Jamali Ma...	2	11
4	1	Prize: Best thing...	Episode 1...	Lee Mack	4	11
5	1	Prize: Best thing...	Episode 1...	Mike Wozn...	5	11
6	1	Prize: Best thing...	Episode 1...	Sarah Ken...	3	11
7	2	Do the most...	Episode 1...	Charlotte...	2	11
8	2	Do the most...	Episode 1...	Jamali Ma...	3	11
9	2	Do the most...	Episode 1...	Lee Mack	3	11
10	2	Do the most...	Episode 1...	Mike Wozn...	5	11
11	2	Do the most...	Episode 1...	Sarah Ken...	4	11

Wrangling

Here's what we have so far:

1	Task	Description	`Charlotte Ritchie`	`Jamali Maddix`	`I
2	Episode 1: It's...	Episode 1:...	Episode 1: It's no...	Episode 1: It'...	Ep:
3	1	Prize: Bes...	1	2	4
4	2	Do the mos...	2	3[1]	3
5	3	Catch the ...	DQ	1	5
6	4	Deliver al...	2	1	5
7	5	Live: Stac...	0	0	0
8	Total	Total	5	7	17
9	Episode 2: The ...	Episode 2:...	Episode 2: The Lur...	Episode 2: The...	Ep:
10	1	Prize: Bes...	5	1	2
11	2	Make the b...	0	5	0

What wrangling steps do we need to take?

Wrangling

Step 1: create a separate column for episode:

```
1 read_html("https://taskmaster.fandom.com/wiki/Series_11") |>
2   html_element(".tmtable") |>
3   html_table() |>
4   mutate(episode = ifelse(startsWith(Task, "Episode"), Task, NA))
```

```
# A tibble: 75 × 2
```

Task	episode
<chr>	<chr>
1 Episode 1: It's not your fault. (18 March 2021)	Episode
1: It's ...	
2 1	<NA>
3 2	<NA>
4 3	<NA>
5 4	<NA>
6 5	<NA>
7 Total	<NA>
8 Episode 2: The Lure of the Treacle Puppies. (25 March 2021)	Episode
2: The L...	
9 1	<NA>

Wrangling

Step 2: fill in the episodes

```
1 read_html("https://taskmaster.fandom.com/wiki/Series_11") |>
2   html_element(".tmtable") |>
3   html_table() |>
4   mutate(episode = ifelse(startsWith(Task, "Episode"), Task, NA)) |>
5   fill(episode, .direction = "down")
```

Task	episode
Episode 1: It's...	Episode 1: It'...
1	Episode 1: It'...
2	Episode 1: It'...
3	Episode 1: It'...
4	Episode 1: It'...
5	Episode 1: It'...
6	Episode 1: It'...
7	Episode 1: It'...
Total	Episode 1: It'...
Episode 2: The Lure of...	Episode 2: The...
1	Episode 2: The...
2	Episode 2: The...

Wrangling

Step 3: remove the “Total” and “Episode” rows in the Task column

1	Task	episode
2	Episode 1: It's...	Episode 1: It'...
3	1	Episode 1: It'...
4	2	Episode 1: It'...
5	3	Episode 1: It'...
6	4	Episode 1: It'...
7	5	Episode 1: It'...
8	Total	Episode 1: It'...
9	Episode 2: The Lure of...	Episode 2: The...
10	1	Episode 2: The...
11	2	Episode 2: The...

What R code would we use to remove these rows?

Wrangling

Step 3: remove the “Total” and “Episode” rows in the Task column

```
1 read_html("https://taskmaster.fandom.com/wiki/Series_11") |>
2   html_element(".tmtable") |>
3   html_table() |>
4   mutate(episode = ifelse(startsWith(Task, "Episode"), Task, NA)) |>
5   fill(episode, .direction = "down") |>
6   filter(!startsWith(Task, "Episode"),
7          !(Task %in% c("Total", "Grand Total")))
```

A tibble: 54 × 8

	Task	Description	Charlotte Ritchie	Jamali Maddix
	Lee Mack			

1	1	Prize: Best thing you c...	1	2
2	2	Do the most impressive ...	2	3[1]
3	3	Catch the rat. You must...	DQ	1
4	4	Deliver all the plates ...	2	1
5	5	Live: Stack your bucket...	0	0
6	1	Prize: Best drinking ve...	5	1
7	2	Make the balloon hover ...	0	5
8	3	Team: Have an argument...	2	2

Wrangling

Step 4: Pivot

```
# A tibble: 54 × 8
  Task Description `Charlotte Ritchie` `Jamali Maddix`
`Lee Mack`
  <chr> <chr> <chr> <chr>
<chr>
1 1 Prize: Best thing you c... 1 2 4
2 2 Do the most impressive ... 2 3[1] 3
3 3 Catch the rat. You must... DQ 1 5
4 4 Deliver all the plates ... 2 1 5
5 5 Live: Stack your bucket... 0 0 0
6 1 Prize: Best drinking ve... 5 1 2
7 2 Make the balloon hover ... 0 5 0
8 3 Team: Have an argument... 2 2 5
9 4 Make the house haunted. 3 3 4
```

How should we pivot this data?

Wrangling

Step 4: Pivot

```
1 read_html("https://taskmaster.fandom.com/wiki/Series_11") |>
2   html_element(".tmtable") |>
3   html_table() |>
4   mutate(episode = ifelse(startsWith(Task, "Episode"), Task, NA)) |>
5   fill(episode, .direction = "down") |>
6   filter(!startsWith(Task, "Episode"),
7          !(Task %in% c("Total", "Grand Total"))) |>
8   pivot_longer(cols = -c(Task, Description, episode),
9                names_to = "contestant",
10               values_to = "score") |>
11   mutate(series = 11)
```

A tibble: 270 × 6

	Task	Description	episode	contestant	score
	<chr>	<chr>	<chr>	<chr>	<dbl>
1	1	Prize: Best thing you can carry, but o...	Episod...	Charlotte...	1
11					
2	1	Prize: Best thing you can carry, but o...	Episod...	Jamali Ma...	2
11					
3	1	Prize: Best thing you can carry, but o...	Episod...	Lee Mack	4
11					

4 1 Prize: Best thing you can carry, but o... Episod... Mike Wozn... 5
11
5 1 Prize: Best thing you can carry but o Episod Sarah Ken 3

Next steps

- Separate episode info into episode number, episode name, and air date columns
- Combine data from multiple series

