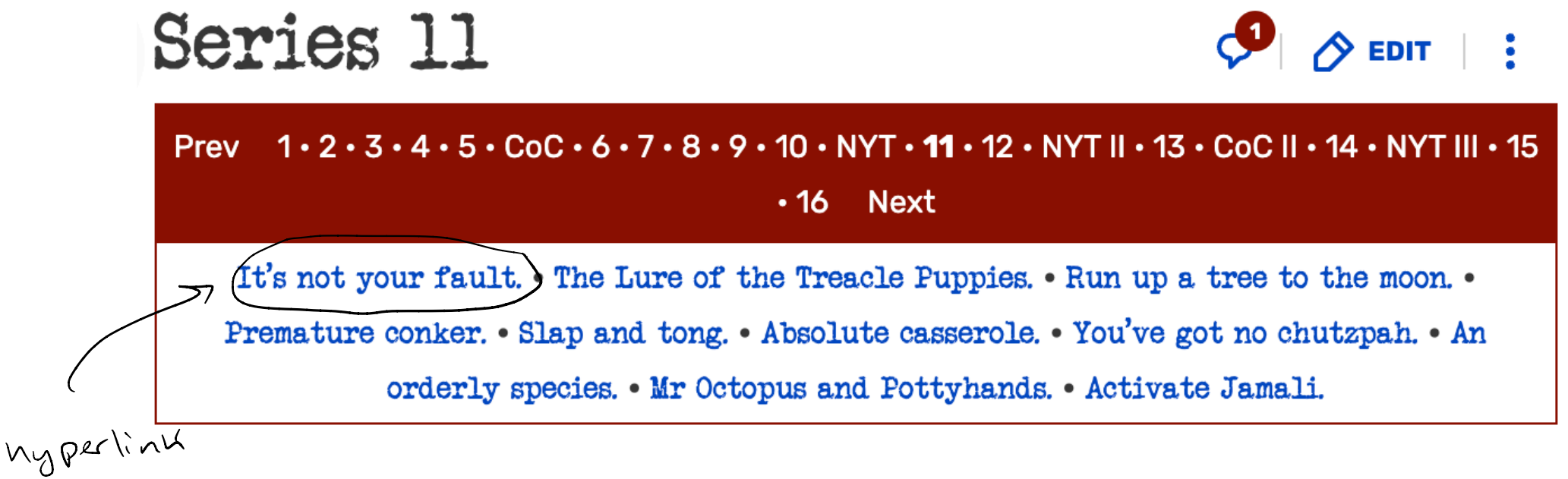
Lecture 19: Web scraping and data wrangling

Motivation: Taskmaster

Last time:

- `read_html` to import web page
- `html_elements` to extract elements of interest
- `html_table` to extract tabular data
- `html_text2` to extract text data

Accessing HTML attributes



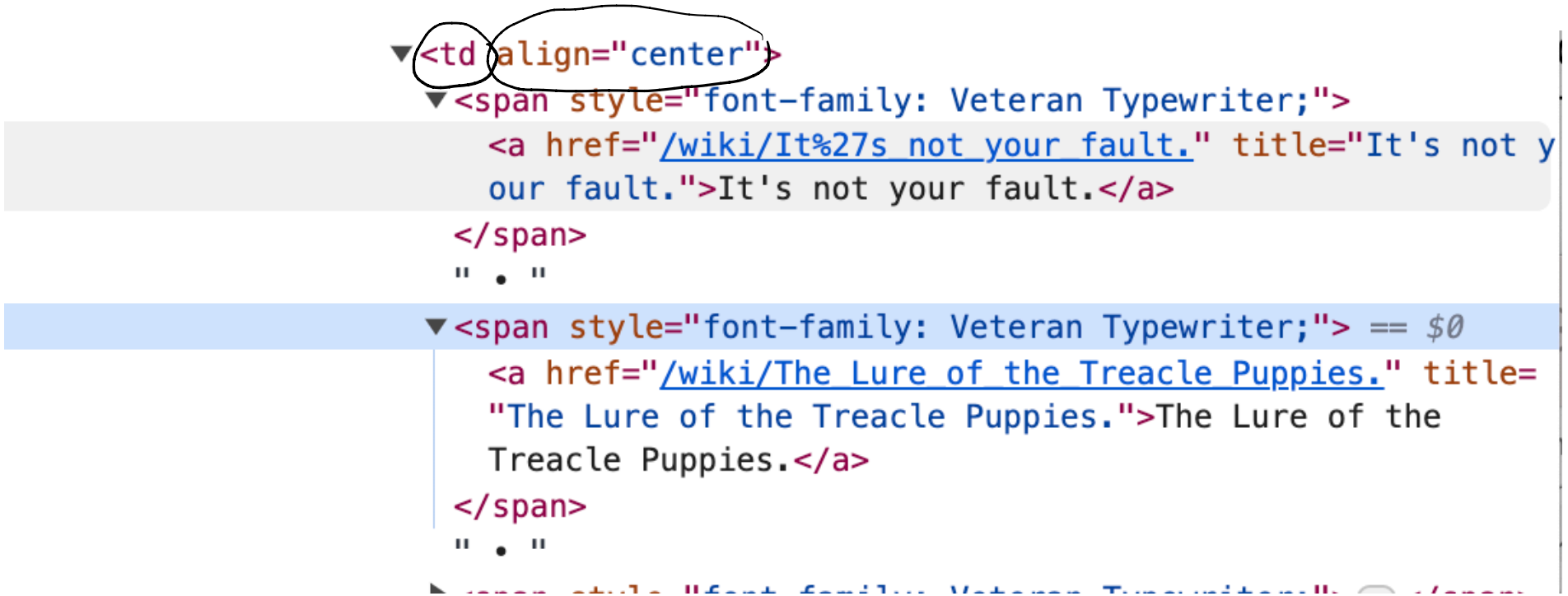
within this element, we have nested elements for each episode

Accessing HTML attributes

```
1 read_html("https://taskmaster.fandom.com/wiki/Series_11") |>
2   html_elements("td")
```

```
{xml_nodeset (436)}
 [1] <td>\n<table class="toccolours" align="center" style="background:
#89110 ...
 [2] <td>\n<a href="/wiki/Series_10" title="Series 10"><span
style="color: #F ...
 [3] <td align="center">\n<span style="font-family: Veteran
Typewriter;"><a h ...
 [4] <td class="pi-horizontal-group-item pi-data-value pi-font pi-
border-colo ...
 [5] <td class="pi-horizontal-group-item pi-data-value pi-font pi-
border-colo ...
 [6] <td colspan="7">Episode 1: <span style="font-family: Veteran
Typewriter; ...
 [7] <td>\n<a href="/wiki/Best_thing_you_can_carry,_but_only_just"
```

Accessing HTML attributes



Accessing HTML attributes

```
1 read_html("https://taskmaster.fandom.com/wiki/Series_11") |>
2   html_elements("td[align='center']")
```

```
{xml_nodeset (1)}
[1] <td align="center">\n<span style="font-family: Veteran Typewriter;">
<a hr ...
```

td element where align='center'

Accessing HTML attributes

td [align='center'] > span

```
▼<td align="center">
  ▼<span style="font-family: Veteran Typewriter;">
      <a href="/wiki/It%27s_not_your_fault." title="It's not y
      our fault.">It's not your fault.</a>
    </span>
    " • "
  ▼<span style="font-family: Veteran Typewriter;"> == $0
      <a href="/wiki/The_Lure_of_the_Treacle_Puppies." title=
      "The Lure of the Treacle Puppies.">The Lure of the
      Treacle Puppies.</a>
    </span>
    " • "
```

span

find span elements that are nested inside td align='center' element

(later, find title or hyperlink info)

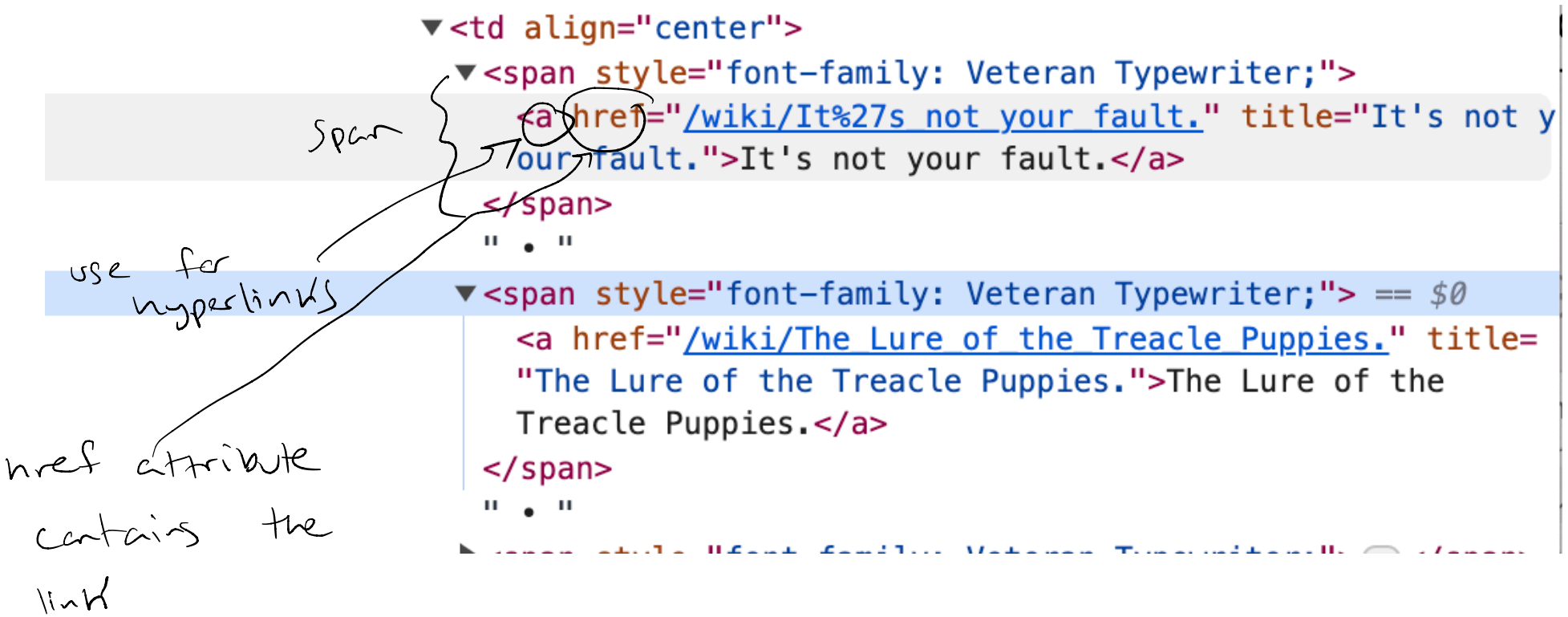
Accessing HTML attributes

```
1 read_html("https://taskmaster.fandom.com/wiki/Series_11") |>
2   html_elements("td[align='center'] > span")
```

```
{xml_nodeset (10)}
 [1] <span style="font-family: Veteran Typewriter;"><a
href="/wiki/It%27s_not ...
 [2] <span style="font-family: Veteran Typewriter;"><a
href="/wiki/The_Lure_o ...
 [3] <span style="font-family: Veteran Typewriter;"><a
href="/wiki/Run_up_a_t ...
 [4] <span style="font-family: Veteran Typewriter;"><a
href="/wiki/Premature_ ...
 [5] <span style="font-family: Veteran Typewriter;"><a
href="/wiki/Slap_and_t ...
 [6] <span style="font-family: Veteran Typewriter;"><a
href="/wiki/Absolute_c ...
 [7] <span style="font-family: Veteran Typewriter;"><a
```

10 episodes in Season 11

Accessing HTML attributes



Accessing HTML attributes

```
1 read_html("https://taskmaster.fandom.com/wiki/Series_11") |>
2   html_elements("td[align='center'] > span > a")
```

```
{xml_nodeset (10)}
 [1] <a href="/wiki/It%27s_not_your_fault." title="It's not your
fault.">It's ...
 [2] <a href="/wiki/The_Lure_of_the_Treacle_Puppies." title="The Lure of
the  ...
 [3] <a href="/wiki/Run_up_a_tree_to_the_moon." title="Run up a tree to
the m ...
 [4] <a href="/wiki/Premature_conker." title="Premature
conker.">Premature co ...
 [5] <a href="/wiki/Slap_and_tong." title="Slap and tong.">Slap and
tong.</a>
 [6] <a href="/wiki/Absolute_casserole." title="Absolute
casserole.">Absolute ...
 [7] <a href="/wiki/You%27ve_got_no_chutzpah." title="You've got no
```

Still need to get hyperlinks

Accessing HTML attributes

Accessing HTML attributes

```
1 read_html("https://taskmaster.fandom.com/wiki/Series_11") |>
2   html_elements("td[align='center'] > span > a") |>
3   html_attr("href")
```

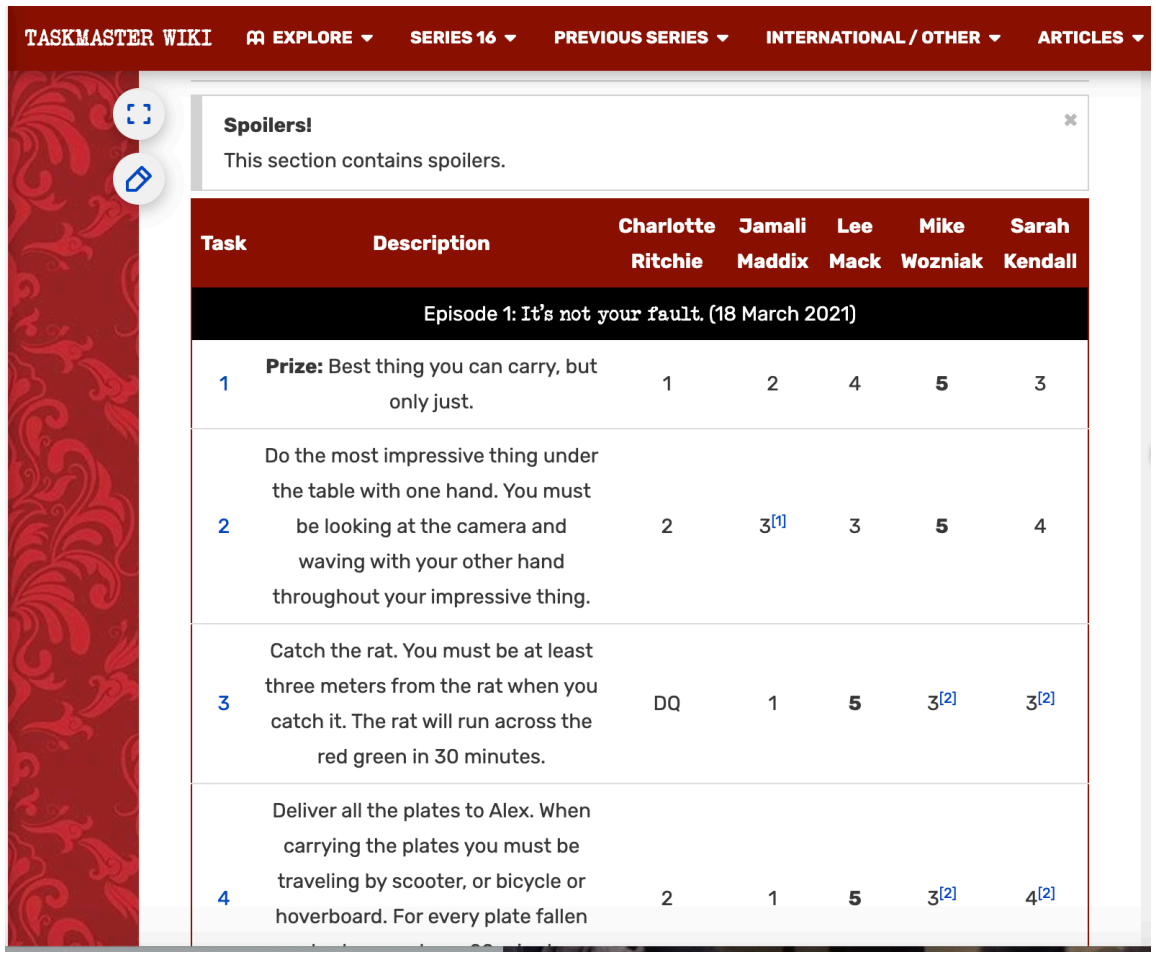
```
 [1] "/wiki/It%27s_not_your_fault."
 [2] "/wiki/The_Lure_of_the_Treacle_Puppies."
 [3] "/wiki/Run_up_a_tree_to_the_moon."
 [4] "/wiki/Premature_conker."
 [5] "/wiki/Slap_and_tong."
 [6] "/wiki/Absolute_casserole."
 [7] "/wiki/You%27ve_got_no_chutzpah."
 [8] "/wiki/An_orderly_species."
 [9] "/wiki/Mr_Octopus_and_Pottyhands."
[10] "/wiki/Activate_Jamali."
```

finds the hyperlink element for each episode

extract the hyperlink attribute

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                4
 3 2                   Do the mos… 2                   3[1]             3
 4 3                   Catch the … DQ                  1                5
 5 4                   Deliver al… 2                   1                5
 6 5                   Live: Stac… 0                   0                0
 7 Total               Total       5                   7                17
 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
1	Episode 1: It's…	Episode 1:…	Episode 1: It's no…	Episode 1: It'…	Epi
2	1	Prize: Bes…	1	2	4
3	2	Do the mos…	2	3[1]	3
4	3	Catch the …	DQ	1	5
5	4	Deliver al…	2	1	5
6	5	Live: Stac…	0	0	0
7	Total	Total	5	7	17
8	Episode 2: The …	Episode 2:…	Episode 2: The Lur…	Episode 2: The…	Epi
9	1	Prize: Bes…	5	1	2
10	2	Make the b…	0	5	0

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

episode episodename Task Description Contestant Score

1

2

.

Wrangling

What we ultimately want:

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

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

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 wrangling steps do we need to take?

- put contestant names into a column (pivot_longer) (in process, gives us a column of scores)
- create a new column for episode (mutate)
- get rid of "Episode" & "Total" rows (filter)
- create a new column for series (mutate)

Wrangling

Step 1: create a separate column for episode:

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

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

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

What R code would we use to remove these rows?

```
filter( Task %in% c(1, 2, 3, 4, 5) )
```

or

```
filter( !startswith(Task, "Episode"), !Task %in%
        c("Total", "Grand Total"))
```

Wrangling

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

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

```
# 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
```

Wrangling

Step 4: Pivot

create a contestant column

```
# 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?

```
pivot_longer( cols = -c(Task, Description, episode),
              names_to = "contestant",
              values_to = "score")
```

Wrangling

Step 4: Pivot

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

```
# A tibble: 270 × 6
   Task  Description                                    episode contestant
score series
   <chr> <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