Introduction to web scraping

Motivation: Great British Bake Off

Show information

- 15 full series (series 16 currently in progress)
- Each series involves around 12 contestants, with one contestant eliminated each episode
- Each episode involves 3 baking challenges

Episode information

https://en.wikipedia.org/wiki/The_Great_British_Bake_Off_series_15

Episodes [edit]

Episode 1: Cake [edit]

The first episode is Cake Week.[17] For the first challenge, the bakers were given two hours to produce an elevated version of their signature loaf cake. For the technical

Episode 2: Biscuits [edit]

The second episode is Biscuit Week. For the first challenge, the bakers were given two hours to produce twelve Viennese whirls sandwiched in any shape, flavour, and filling of

Goal and required steps

Goal: Explore the different baking challenges across all the different series. Which challenges appear most frequently? Which challenges appear earliest in the series?

- Scrape episode information from each series
- Combine and clean
- Summarize the data, do statistics!

Scraping the data

```
1 library(rvest)
2 library(tidyverse)
3
4 gbbo <- read_html("https://en.wikipedia.org/wiki/The_Great_British_Ba
5 gbbo
```

read an HTML page

URL

```
{html_document}
<html class="client-nojs vector-feature-language-in-header-enabled
vector-feature-language-in-main-page-header-disabled vector-feature-
page-tools-pinned-disabled vector-feature-toc-pinned-clientpref-1
vector-feature-main-menu-pinned-disabled vector-feature-limited-width-
clientpref-1 vector-feature-limited-width-content-enabled vector-
feature-custom-font-size-clientpref-1 vector-feature-appearance-pinned-
clientpref-1 vector-feature-night-mode-enabled skin-theme-clientpref-
day vector-sticky-header-enabled vector-toc-available" lang="en"
dir="ltr">
[1] <head>\n<meta http-equiv="Content-Type" content="text/html;
charset=UTF-8 ...
[2] <body class="skin--responsive skin-vector skin-vector-search-vue
```

HTML basics

< > : tag

```
<html>
  <head>
    <title>Page title</title>
  </head>
  <body>
    <h1 id='first'>A heading</h1>
    <p>Some text &amp;amp; <b>some bold text.</b></p>
    <img src='bakeoff_0.jpg' width='300' height='200'>
  </body>
</html>
```

<html> ← start of the HTML document

<head> ← start of header

</head> ← end of header

header : metadata about the page

<body>

body : stuff that gets displayed on the page

<img> : image

attributes: src (source file), width, height, etc.

</body>

</html> ← end of the HTML document

HTML basics

```
<html>
  <head>
    <title>Page title</title>
  </head>
  <body>
    <h1 id='first'>A heading</h1>
    <p>Some text &amp;amp; <b>some bold
text.</b></p>
    <img src='bakeoff_0.jpg'
width='300' height='200'>
  </body>
</html>
```

A heading

Some text & **some bold text.**

Markdown equivalents:

** ** bold

h1

h2

Some HTML elements

- `<html>`: start of the HTML page
- `<head>`: header information (metadata about the page)
- `<body>`: everything that is on the page
- `<h1>` to `<h6>`: headings
- `<p>`: paragraphs
- `<b>`: bold
- `<table>`: table

Extracting HTML elements

Episodes [edit]

Episode 1: Cake [edit]

The first episode is Cake Week.[17] For the first challenge, the bakers were given two hours to produce an elevated version of their signature loaf cake. For the technical

It looks like the episode name is contained in a heading. How can we extract it?

Extracting HTML elements

Episodes [edit]

Episode 1: Cake [edit]

The first episode is Cake Week.[17] For the first challenge, the bakers were given two hours to produce an elevated version of their signature loaf cake. For the technical

specify html elements to look at

```
1 gbbo |>
2   html_elements("h1")
```

```
{xml_nodeset (1)}
[1] <h1 id="firstHeading" class="firstHeading mw-first-
heading">\n<i>The Grea ...
```

Extracting HTML elements

Episodes [edit]

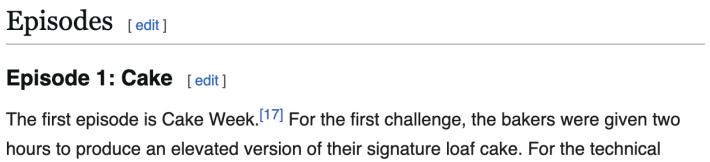
Episode 1: Cake [edit]

The first episode is Cake Week.[17] For the first challenge, the bakers were given two hours to produce an elevated version of their signature loaf cake. For the technical

```
gbbo |>
  html_elements("h2")
```

```
{xml_nodeset (7)}
[1] <h2 class="vector-pinnable-header-label">Contents</h2>
[2] <h2 id="Bakers">Bakers</h2>\n
[3] <h2 id="Results_summary">Results summary</h2>\n
[4] <h2 id="Episodes">Episodes</h2>\n
[5] <h2 id="Specials">Specials</h2>\n
[6] <h2 id="Ratings">Ratings</h2>\n
[7] <h2 id="References">References</h2>\n
```

Extracting HTML elements



Episodes [edit]

Episode 1: Cake [edit]

The first episode is Cake Week.[17] For the first challenge, the bakers were given two hours to produce an elevated version of their signature loaf cake. For the technical

```
gbbo |>
  html_elements("h3")
```

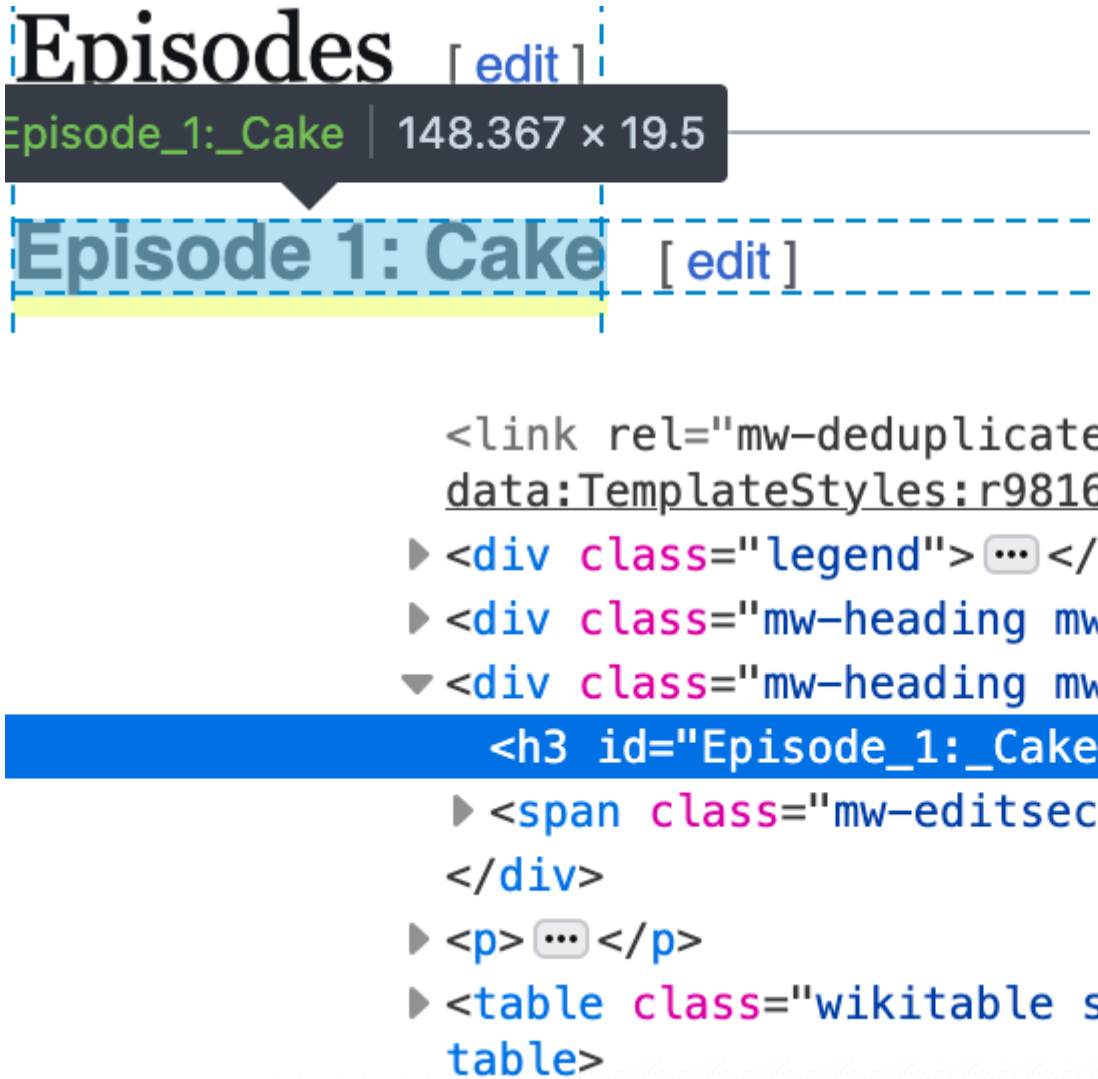
```
{xml_nodeset (12)}
 [1] <h3 id="Episode_1:_Cake">Episode 1: Cake</h3>\n
 [2] <h3 id="Episode_2:_Biscuits">Episode 2: Biscuits</h3>\n
 [3] <h3 id="Episode_3:_Bread">Episode 3: Bread</h3>\n
 [4] <h3 id="Episode_4:_Caramel">Episode 4: Caramel</h3>\n
 [5] <h3 id="Episode_5:_Pastry">Episode 5: Pastry</h3>\n
 [6] <h3 id="Episode_6:_Autumn">Episode 6: Autumn</h3>\n
 [7] <h3 id="Episode_7:_Desserts">Episode 7: Desserts</h3>\n
 [8] <h3 id="Episode_8:_The_'70s_(Quarterfinals)">\n<span
id="Episode_8:_The_ ...
 [9] <h3 id="Episode_9:_Patisserie_(Semifinal)">\n<span
id="Episode_9:_Patiss ...
[10] <h3 id="Episode_10:_Final">Episode 10: Final</h3>\n
```

Finding the right selectors

(Demo)

Finding the right selectors

1. Open the web page in your browser (I find that Firefox or Chrome tend to work best)
2. Right-click on the element you want, and click “Inspect”


Episodes [edit]
Episode_1:_Cake | 148.367 × 19.5
Episode 1: Cake [edit]
<link rel="mw-deduplicated-inline-style" href="mw-data:TemplateStyles:r981673959">
…
… flow-r
 flow-root
<h3 id="Episode_1:_Cake">Episode 1: Cake</h3>
…

…
<table class="wikitable sortable jquery-tablesorter
table>

Finding the right selector

```
gbbo |>
  html_elements("h3")
```

```
{xml_nodeset (12)}
 [1] <h3 id="Episode_1:_Cake">Episode 1: Cake</h3>\n
 [2] <h3 id="Episode_2:_Biscuits">Episode 2: Biscuits</h3>\n
 [3] <h3 id="Episode_3:_Bread">Episode 3: Bread</h3>\n
 [4] <h3 id="Episode_4:_Caramel">Episode 4: Caramel</h3>\n
 [5] <h3 id="Episode_5:_Pastry">Episode 5: Pastry</h3>\n
 [6] <h3 id="Episode_6:_Autumn">Episode 6: Autumn</h3>\n
 [7] <h3 id="Episode_7:_Desserts">Episode 7: Desserts</h3>\n
 [8] <h3 id="Episode_8:_The_'70s_(Quarterfinals)">\n<span id=" ...
 [9] <h3 id="Episode_9:_Patisserie_(Semifinal)">\n<span id="Ep ...
[10] <h3 id="Episode_10:_Final">Episode 10: Final</h3>\n
[11] <h3 id="The_Great_Christmas_Bake_Off"><i>The Great Christ ...
[12] <h3 id="The_Great_New_Year_Bake_Off"><i>The Great New Yea ...
```

Finding the right selector

CSS selector (specify HTML elements)

[ ] : specify information about elements attributes

^= : starts with

```
1 gbbo |>
2   html_elements("h3[id^='Episode']")
```

```
{xml_nodeset (10)}
 [1] <h3 id="Episode_1:_Cake">Episode 1: Cake</h3>\n
 [2] <h3 id="Episode_2:_Biscuits">Episode 2: Biscuits</h3>\n
 [3] <h3 id="Episode_3:_Bread">Episode 3: Bread</h3>\n
 [4] <h3 id="Episode_4:_Caramel">Episode 4: Caramel</h3>\n
 [5] <h3 id="Episode_5:_Pastry">Episode 5: Pastry</h3>\n
 [6] <h3 id="Episode_6:_Autumn">Episode 6: Autumn</h3>\n
 [7] <h3 id="Episode_7:_Desserts">Episode 7: Desserts</h3>\n
 [8] <h3 id="Episode_8:_The_'70s_(Quarterfinals)">\n<span
id="Episode_8:_The_ ...
 [9] <h3 id="Episode_9:_Patisserie_(Semifinal)">\n<span
id="Episode_9:_Patiss ...
[10] <h3 id="Episode_10:_Final">Episode 10: Final</h3>\n
```

Extracting the data

```
1 gbbo |>
2   html_elements("h3[id^='Episode']") |>
3   html_text2() <- extract text from the elements
```

```
 [1] "Episode 1: Cake"                          "Episode 2: Biscuits"
 [3] "Episode 3: Bread"                         "Episode 4: Caramel"
 [5] "Episode 5: Pastry"                        "Episode 6: Autumn"
 [7] "Episode 7: Desserts"                      "Episode 8: The '70s
(Quarterfinals)"
 [9] "Episode 9: Patisserie (Semifinal)"        "Episode 10: Final"
```

Extracting data from other elements

Release	
Original network	Channel 4
Original release	24 September – 26 November 2024
Series chronology	
← **Previous**	**Next** →
Series 14	Series 16

How would we scrape the air dates for the series?

Extracting data from other elements


```
1 gbbo |>
2   html_elements("table")
```

```
{xml_nodeset (18)}
 [1] <table class="infobox vevent"><tbody>\n<tr><th colspan="2 ...
 [2] <table class="wikitable sortable" style="text-align:cente ...
 [3] <table class="wikitable" style="text-align:center; font-s ...
```

Extracting data from other elements


```
gbbo |>
  html_elements("table[class='infobox vevent']")
```

```
{xml_nodeset (1)}
[1] <table class="infobox vevent"><tbody>\n<tr><th colspan="2"
class="infobox ...
```

Extracting data from other elements


```
1 gbbo |>
2   html_elements("table.infobox.vevent")
```

```
{xml_nodeset (1)}
[1] <table class="infobox vevent"><tbody>\n<tr><th colspan="2"
class="infobox ...
```

.infobox <=> [class~="infobox"]

Extracting data from other elements

```
gbbo |>
  html_element("table.infobox.vevent") |>
  html_table() ← read the table into R as a data frame
```

The Great British Bake Off	The Great British Bake Off
Series 15	Series 15
Presented by	Noel FieldingAlison Hammond
Judges	Paul HollywoodPrue Leith
No. of contestants	12
Winner	Georgie Grasso

The Great British Bake Off	The Great British Bake Off
Runners-up	Christiaan de VriesDylan Bachelet
Location	Welford Park, near Newbury, Berkshire
No. of max. bakes	30
No. of episodes	10
Release	Release
Original network	Channel 4
Original release	24 September (2024-09-24) –26 November 2024 (2024-11-26)

Class activity

- Work independently or with a neighbor on the class activity
- At the end of class, submit your work as an HTML file on Canvas (one per group, list all your names)