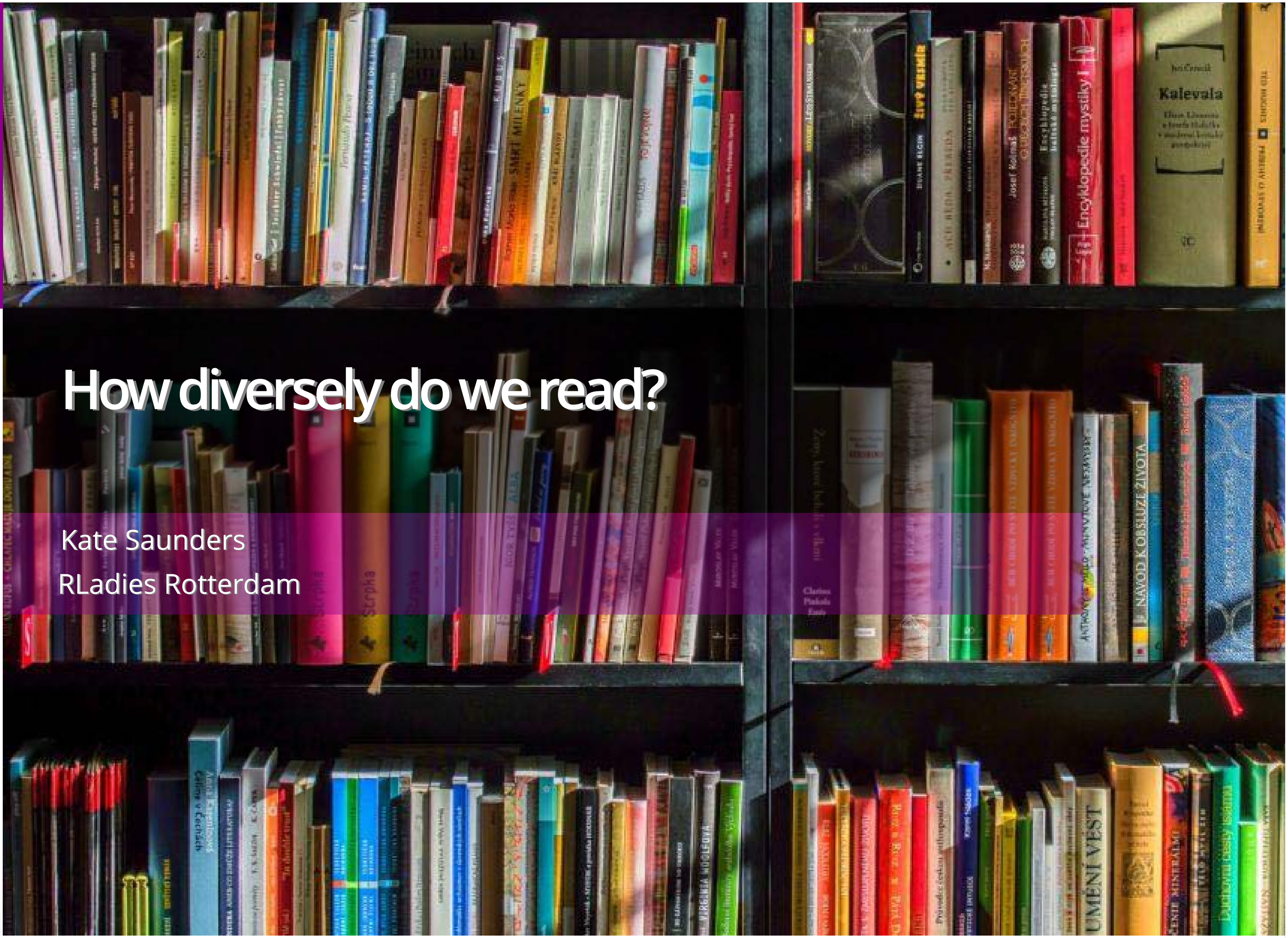
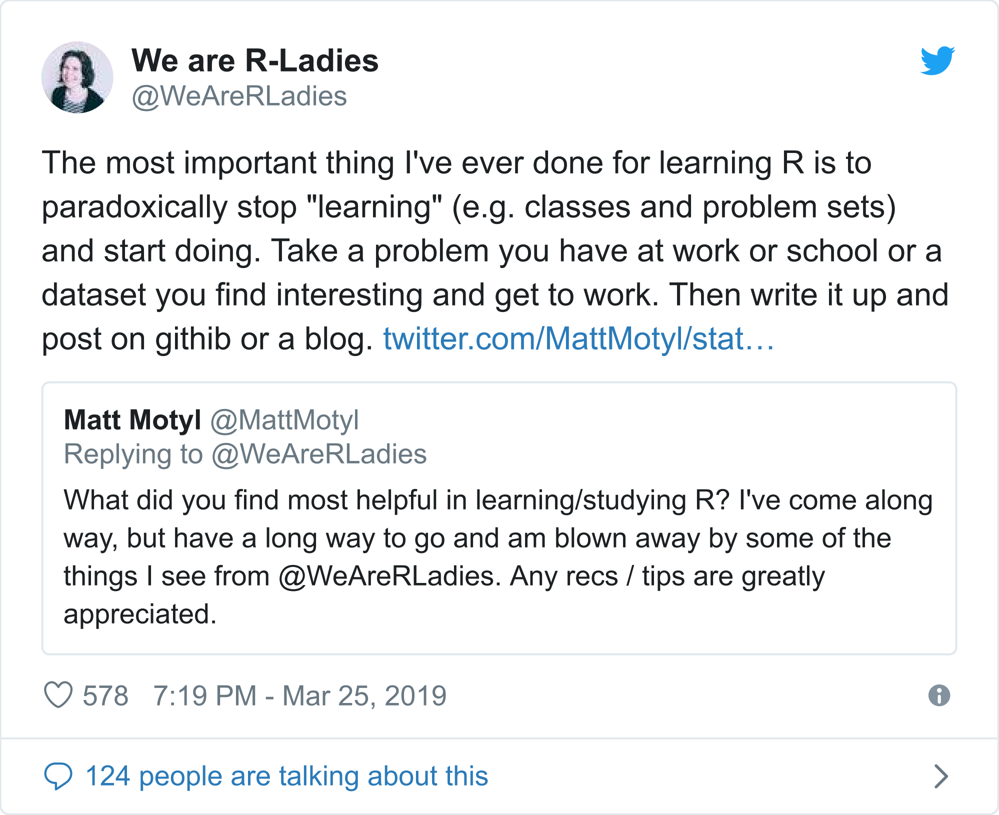


How diversely do we read?

Kate Saunders

RLadies Rotterdam

Motivation



We are R-Ladies
@WeAreRLadies

The most important thing I've ever done for learning R is to paradoxically stop "learning" (e.g. classes and problem sets) and start doing. Take a problem you have at work or school or a dataset you find interesting and get to work. Then write it up and post on githib or a blog. twitter.com/MattMotyl/stat…

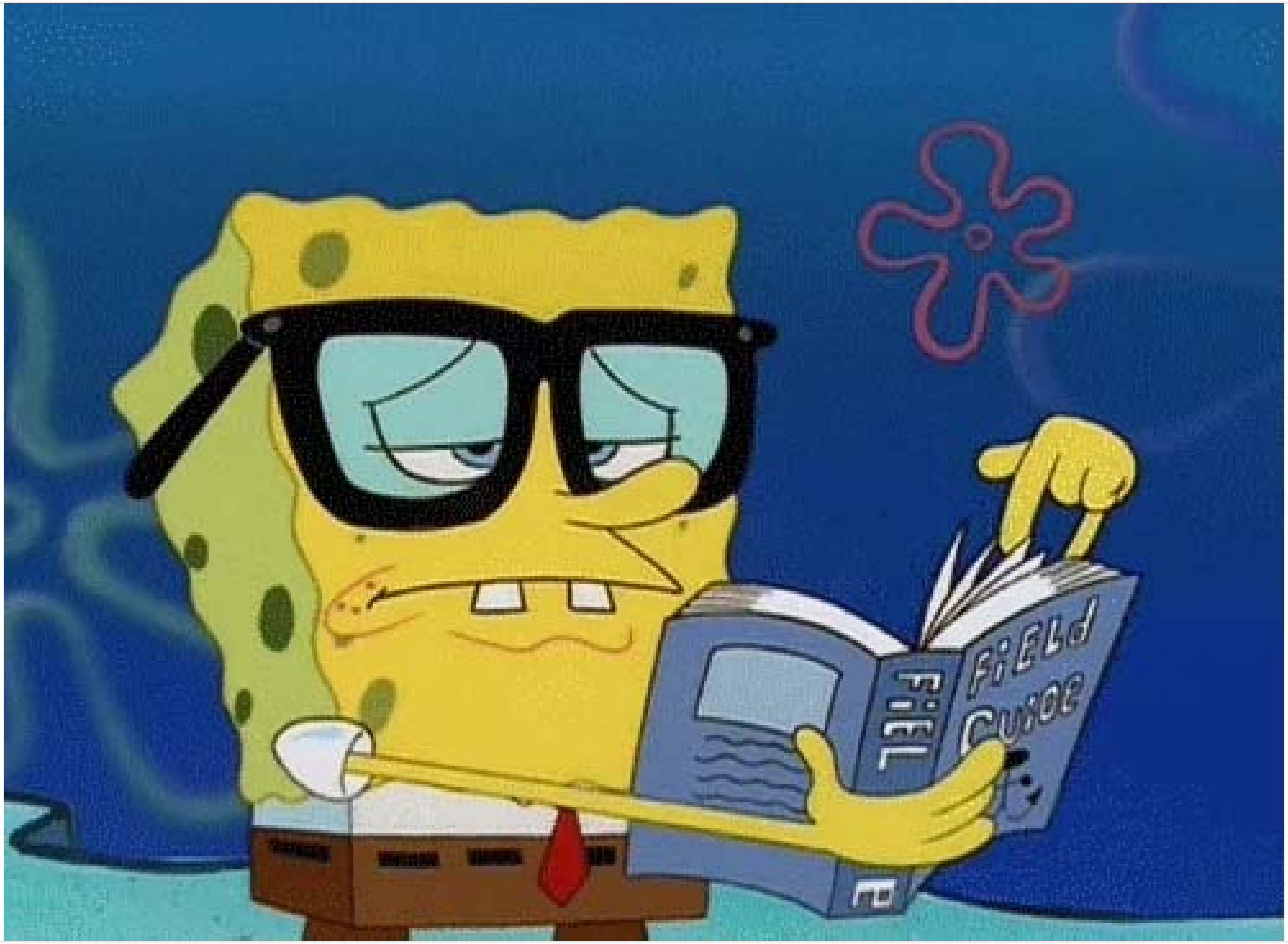
Matt Motyl @MattMotyl
Replying to @WeAreRLadies

What did you find most helpful in learning/studying R? I've come along way, but have a long way to go and am blown away by some of the things I see from @WeAreRLadies. Any recs / tips are greatly appreciated.

♡ 578 7:19 PM - Mar 25, 2019

124 people are talking about this

After PhD - I wanted to read more!



via GIPHY

Question

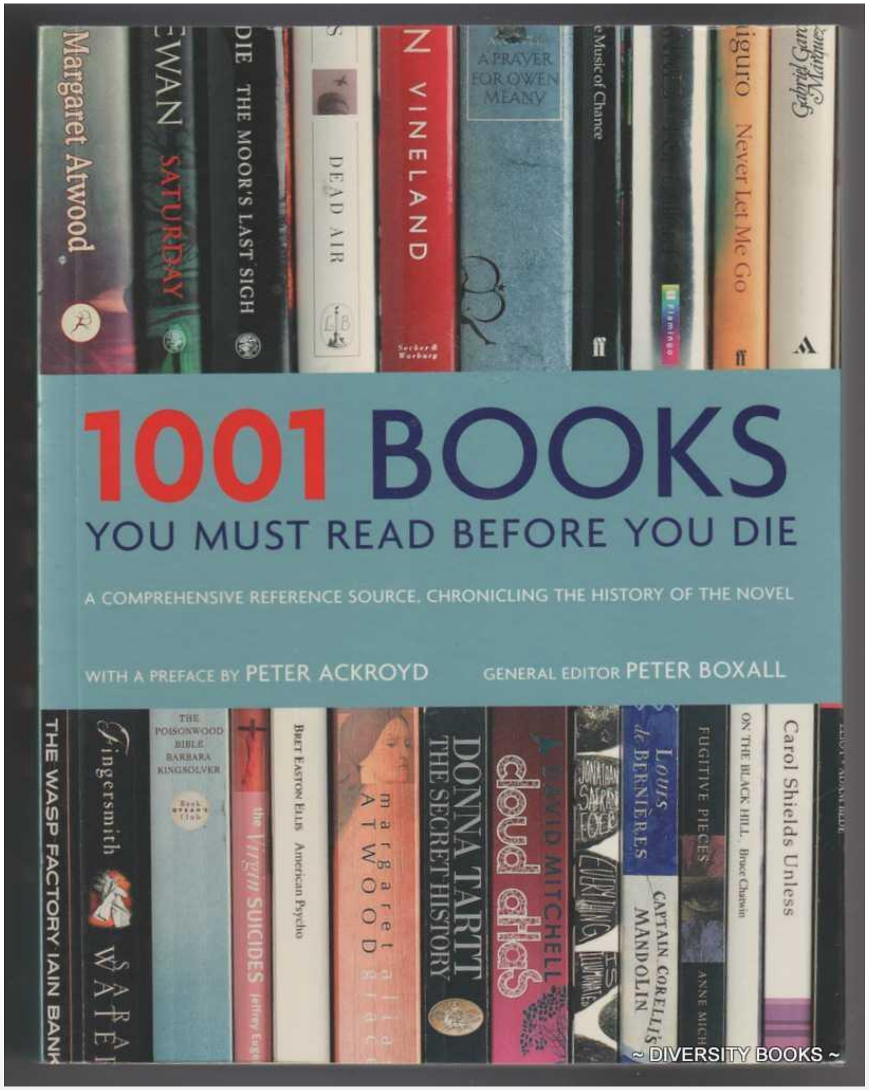
How should I choose what to read?

Question

How should I choose what to read?

Am I reading diversely?

Picking from lists


1001 BOOKS
YOU MUST READ BEFORE YOU DIE
A COMPREHENSIVE REFERENCE SOURCE, CHRONICLING THE HISTORY OF THE NOVEL
WITH A PREFACE BY PETER ACKROYD
GENERAL EDITOR PETER BOXALL
~ DIVERSITY BOOKS ~

Questions for our data

Are lists like this biased?

Outline

1. Get to know the basics of webpages

2. Look at some examples of webscraping

3. Get some data to answer my questions

Packages we need

```
library(rvest)
library(tidyverse)
library(plotly)
```

Disclaimer

I am not a web scraping expert

Disclaimer

I am not a web scraping expert

And sometimes I write naughty code ...

Disclaimer

I am not a web scraping expert

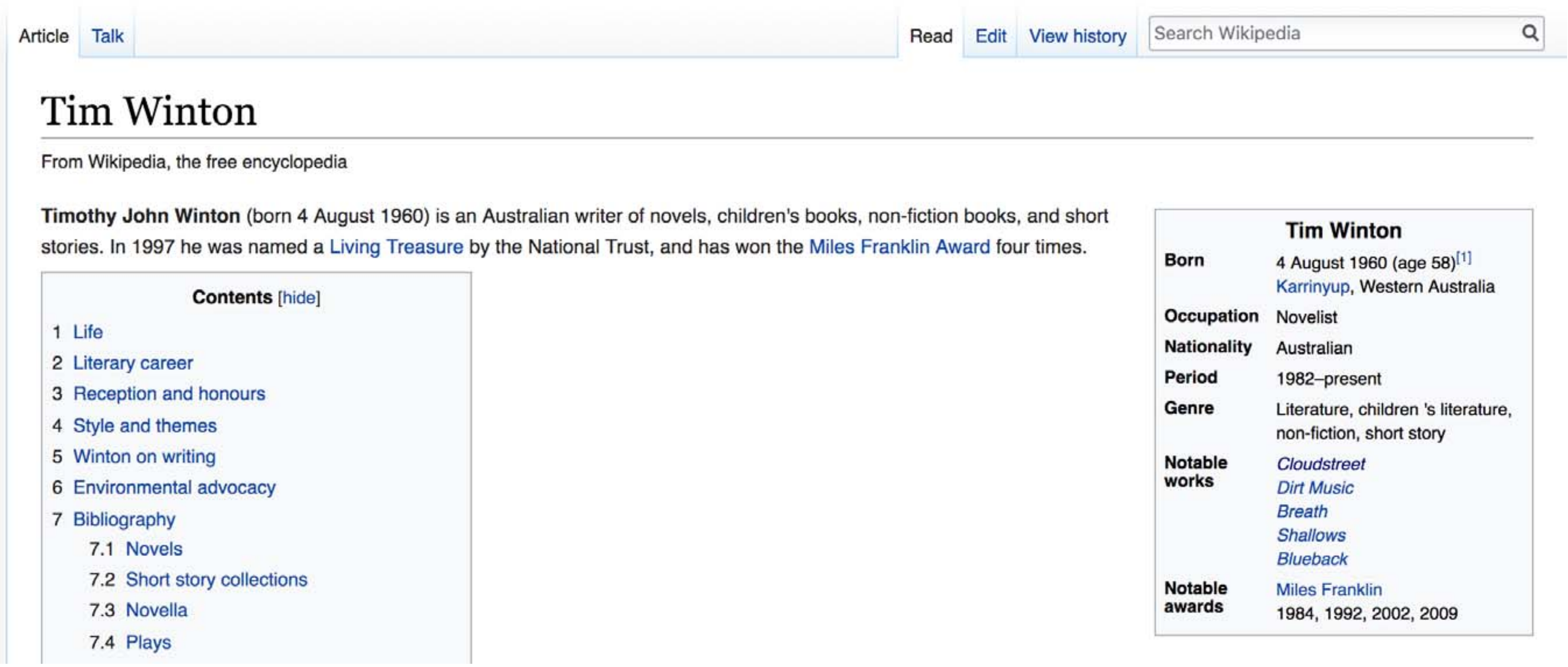
And sometimes I write naughty code …

But I can share what I know

Webpage basics

Go to a webpage

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



Tim Winton

From Wikipedia, the free encyclopedia

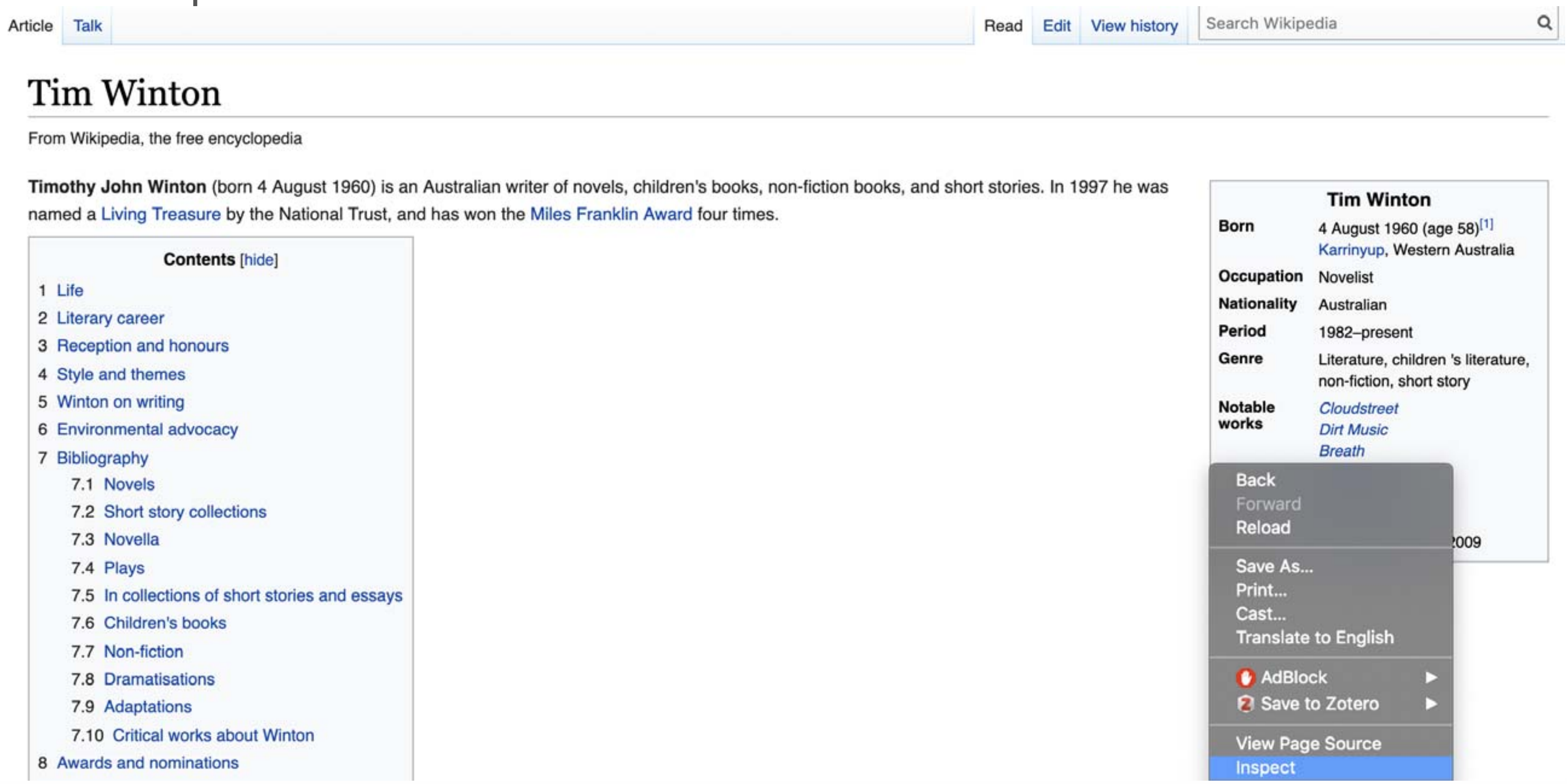
Timothy John Winton (born 4 August 1960) is an Australian writer of novels, children's books, non-fiction books, and short stories. In 1997 he was named a Living Treasure by the National Trust, and has won the Miles Franklin Award four times.

Contents [hide]

Tim Winton	
Born	4 August 1960 (age 58)[1] Karrinyup, Western Australia
Occupation	Novelist
Nationality	Australian
Period	1982–present
Genre	Literature, children 's literature, non-fiction, short story
Notable works	*Cloudstreet* *Dirt Music* *Breath* *Shallows* *Blueback*
Notable awards	Miles Franklin 1984, 1992, 2002, 2009

View html code in Chrome

- Right click the part of the page you want
- Select inpsect



Tim Winton

From Wikipedia, the free encyclopedia

Timothy John Winton (born 4 August 1960) is an Australian writer of novels, children's books, non-fiction books, and short stories. In 1997 he was named a Living Treasure by the National Trust, and has won the Miles Franklin Award four times.

Contents [hide]

1 Life
2 Literary career
3 Reception and honours
4 Style and themes
5 Winton on writing
6 Environmental advocacy
7 Bibliography
7.1 Novels
7.2 Short story collections
7.3 Novella
7.4 Plays
7.5 In collections of short stories and essays
7.6 Children's books
7.7 Non-fiction
7.8 Dramatisations
7.9 Adaptations
7.10 Critical works about Winton
8 Awards and nominations

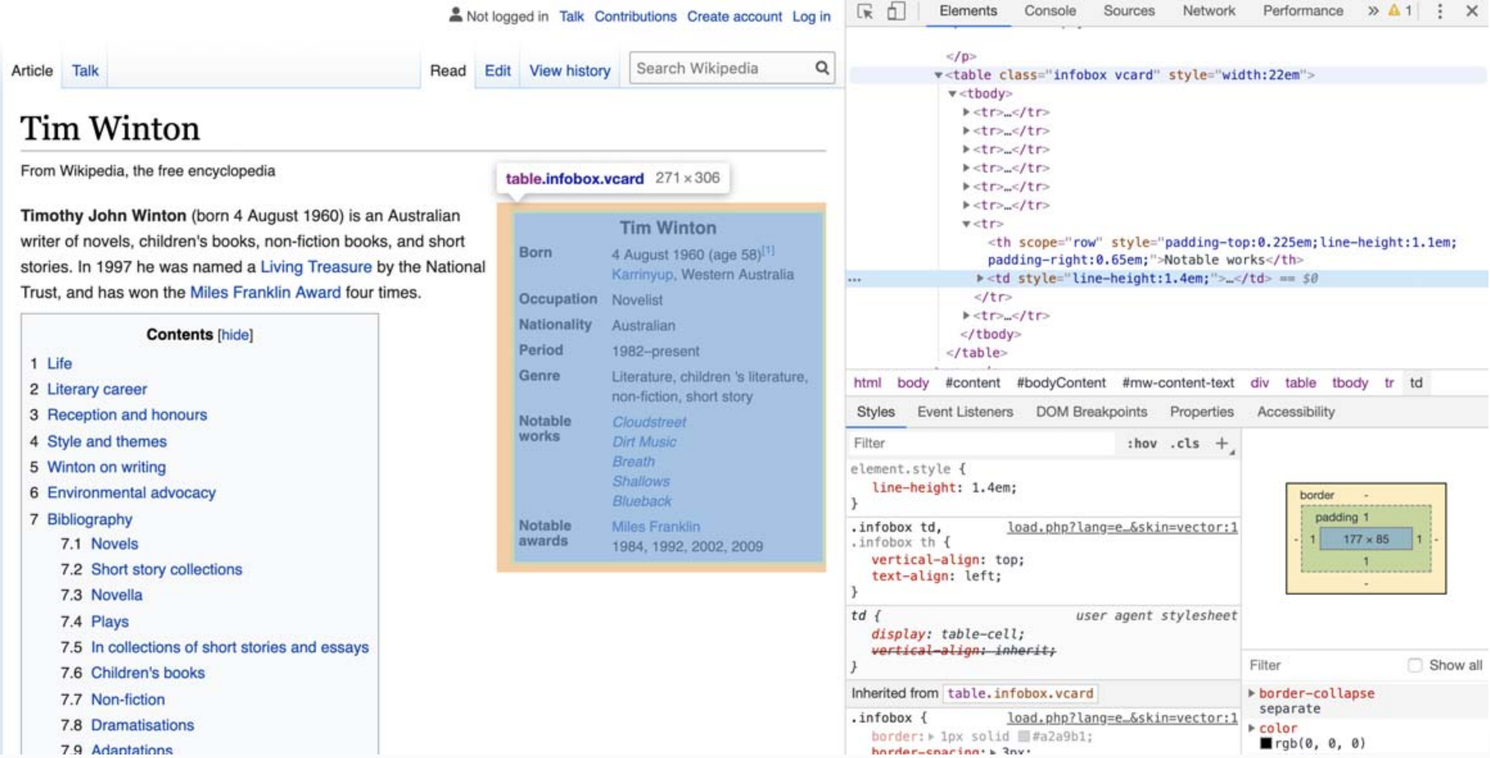
Tim Winton

Born	4 August 1960 (age 58)[1] Karrinyup, Western Australia
Occupation	Novelist
Nationality	Australian
Period	1982–present
Genre	Literature, children 's literature, non-fiction, short story
Notable works	*Cloudstreet* *Dirt Music* *Breath*

2009

Html code

- Brings up the html code
- Highlights the piece of html code related to your click
- Hover over html code to see other features of the web page



Tim Winton

From Wikipedia, the free encyclopedia

Timothy John Winton (born 4 August 1960) is an Australian writer of novels, children's books, non-fiction books, and short stories. In 1997 he was named a Living Treasure by the National Trust, and has won the Miles Franklin Award four times.

Contents [hide]

1 Life
2 Literary career
3 Reception and honours
4 Style and themes
5 Winton on writing
6 Environmental advocacy
7 Bibliography
7.1 Novels
7.2 Short story collections
7.3 Novella
7.4 Plays
7.5 In collections of short stories and essays
7.6 Children's books
7.7 Non-fiction
7.8 Dramatisations
7.9 Adaptations

Tim Winton	
Born	4 August 1960 (age 58)[1] Karrinyup, Western Australia
Occupation	Novelist
Nationality	Australian
Period	1982–present
Genre	Literature, children 's literature, non-fiction, short story
Notable works	*Cloudstreet* *Dirt Music* *Breath* *Shallows* *Blueback*
Notable awards	Miles Franklin 1984, 1992, 2002, 2009

```
</p>
▾<table class="infobox vcard" style="width:22em">
  ▾<tbody>
    ▸<tr>…</tr>
    ▸<tr>…</tr>
    ▸<tr>…</tr>
    ▸<tr>…</tr>
    ▸<tr>…</tr>
    ▸<tr>…</tr>
    ▾<tr>
       <th scope="row" style="padding-top:0.225em;line-height:1.1em;
       padding-right:0.65em;">Notable works</th>
...   ▸<td style="line-height:1.4em;">…</td> == $0
     </tr>
    ▸<tr>…</tr>
   </tbody>
</table>
```

html body #content #bodyContent #mw-content-text div table tbody tr td

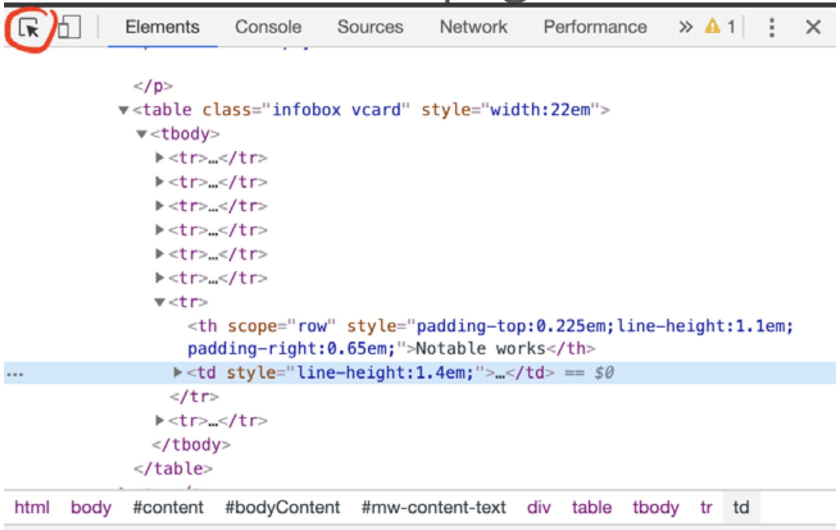
Styles | Event Listeners | DOM Breakpoints | Properties | Accessibility

```
Filter                                  :hov .cls +
element.style {
   line-height: 1.4em;
}
.infobox td,             load.php?lang=e…&skin=vector:1
.infobox th {
   vertical-align: top;
   text-align: left;
}
td {                         user agent stylesheet
   display: table-cell;
   ~~vertical-align: inherit;~~
}
Inherited from table.infobox.vcard
.infobox {               load.php?lang=e…&skin=vector:1
   border: ▸ 1px solid ■#a2a9b1;
   border-spacing: ▸ 3px;
```

```
▸ border-collapse
  separate
▸ color
  ■rgb(0, 0, 0)
```

Inpsect button

- Similarly, click the top left button in the side panel
- Explore related features of the webpage and html code



```
        </p>
      ▼<table class="infobox vcard" style="width:22em">
        ▼<tbody>
          ▶<tr>…</tr>
          ▶<tr>…</tr>
          ▶<tr>…</tr>
          ▶<tr>…</tr>
          ▶<tr>…</tr>
          ▶<tr>…</tr>
          ▼<tr>
              <th scope="row" style="padding-top:0.225em;line-height:1.1em;
              padding-right:0.65em;">Notable works</th>
...         ▶<td style="line-height:1.4em;">…</td> == $0
            </tr>
          ▶<tr>…</tr>
          </tbody>
        </table>
```

Basic html types

Structure: <tag> Some stuff </tag>

Basic tag types are:

div - Division or section

table - Table

p - Paragraph elements

h - Heading

Webscraping basics

Read a webpage

```r
library(rvest)
author_url <- "https://en.wikipedia.org/wiki/Tim_Winton"
wiki_data <- read_html(author_url) # Scrape the data from the webpage
wiki_data
```

```
## {xml_document}
## <html class="client-nojs" lang="en" dir="ltr">
## [1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset= ...
## [2] <body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-sub ...
```

How to scrape a table - html_table()

```r
table_data <- wiki_data %>%
  rvest::html_table() #Get all tables on the webpage

length(table_data)
```

```
## [1] 3
```

```r
str(table_data[[1]])
```

```
## 'data.frame': 7 obs. of 2 variables:
## $ Tim Winton: chr "Born" "Occupation" "Nationality" "Period" ...
## $ Tim Winton: chr "4 August 1960 (1960-08-04) (age 58)[1]Karrinyup, Western
Australia" "Novelist" "Australian" "1982–present" ...
```

Other approaches - html_nodes()

```
table_data_eg1 <- wiki_data %>%
 rvest::html_nodes("table") %>% # get all the nodes of type table
 purrr::pluck(1) %>% #pull out the first one
 rvest::html_table(header = FALSE) #convert it to table type
str(table_data_eg1)
```

```
## 'data.frame': 8 obs. of 2 variables:
## $ X1: chr "Tim Winton" "Born" "Occupation" "Nationality" ...
## $ X2: chr "Tim Winton" "4 August 1960 (1960-08-04) (age 58)[1]Karrinyup,
Western Australia" "Novelist" "Australian" ...
```

Other approaches - html_node()

```r
table_data_eg2 <- wiki_data %>%
  rvest::html_node("table") %>% # just get the first table match
  rvest::html_table(header = FALSE) #convert it to table type
str(table_data_eg2)
```

```
## 'data.frame': 8 obs. of 2 variables:
## $ X1: chr "Tim Winton" "Born" "Occupation" "Nationality" ...
## $ X2: chr "Tim Winton" "4 August 1960 (1960-08-04) (age 58)[1]Karrinyup,
Western Australia" "Novelist" "Australian" ...
```

Get the Nationality

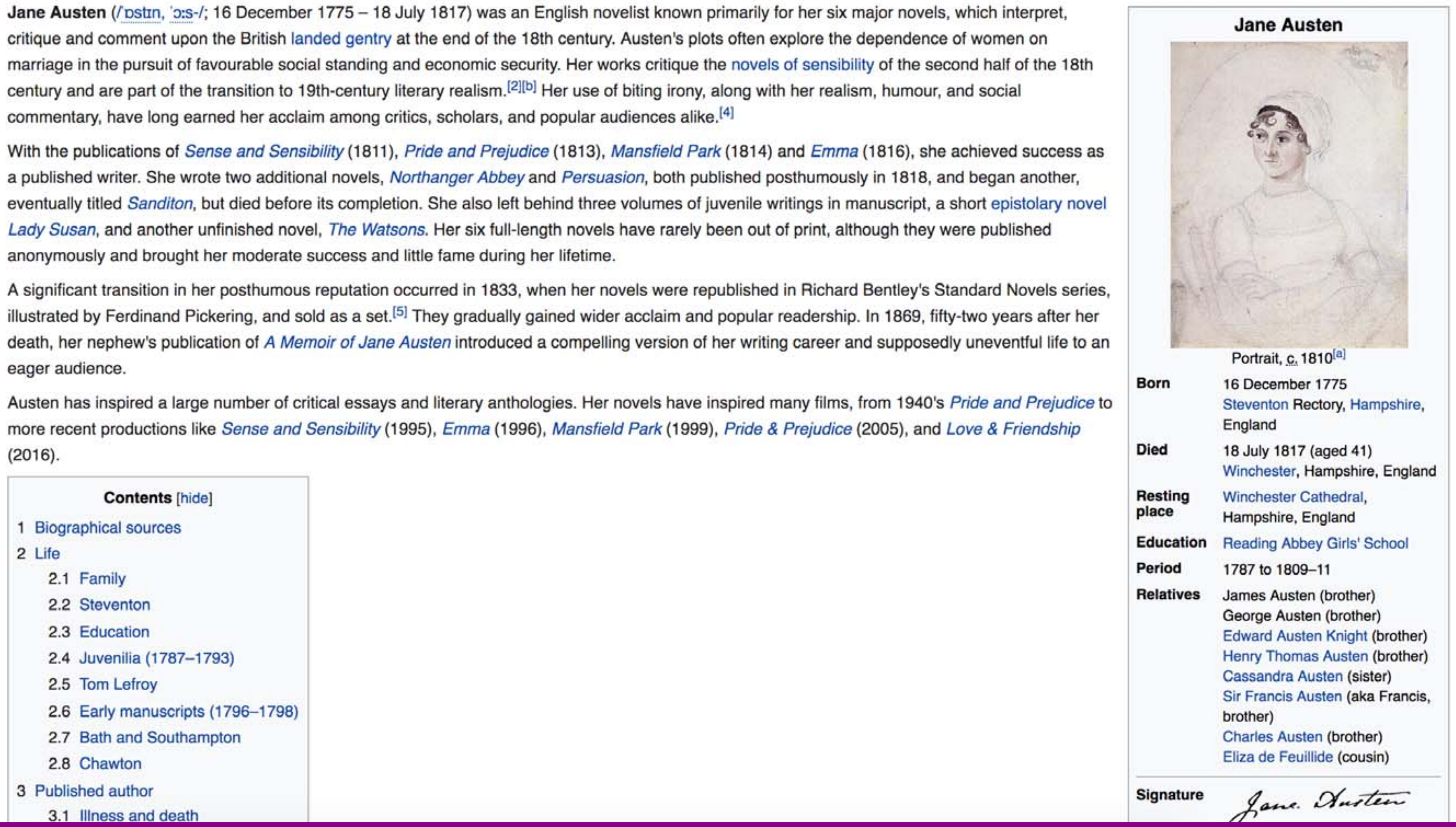
```
author_nationality = table_data_eg2 %>%
  dplyr::rename(Category = X1, Response = X2) %>%
  dplyr::filter(Category == "Nationality") %>%
  dplyr::select(Response) %>%
  as.character()
author_nationality
```

```
## [1] "Australian"
```

Can we generalise?

Same example - Different author

"https://en.wikipedia.org/wiki/Jane_Austen"



Jane Austen (/ˈɒstɪn, ˈɔːs-/; 16 December 1775 – 18 July 1817) was an English novelist known primarily for her six major novels, which interpret, critique and comment upon the British landed gentry at the end of the 18th century. Austen's plots often explore the dependence of women on marriage in the pursuit of favourable social standing and economic security. Her works critique the novels of sensibility of the second half of the 18th century and are part of the transition to 19th-century literary realism.[2][b] Her use of biting irony, along with her realism, humour, and social commentary, have long earned her acclaim among critics, scholars, and popular audiences alike.[4]

With the publications of *Sense and Sensibility* (1811), *Pride and Prejudice* (1813), *Mansfield Park* (1814) and *Emma* (1816), she achieved success as a published writer. She wrote two additional novels, *Northanger Abbey* and *Persuasion*, both published posthumously in 1818, and began another, eventually titled *Sanditon*, but died before its completion. She also left behind three volumes of juvenile writings in manuscript, a short epistolary novel *Lady Susan*, and another unfinished novel, *The Watsons*. Her six full-length novels have rarely been out of print, although they were published anonymously and brought her moderate success and little fame during her lifetime.

A significant transition in her posthumous reputation occurred in 1833, when her novels were republished in Richard Bentley's Standard Novels series, illustrated by Ferdinand Pickering, and sold as a set.[5] They gradually gained wider acclaim and popular readership. In 1869, fifty-two years after her death, her nephew's publication of *A Memoir of Jane Austen* introduced a compelling version of her writing career and supposedly uneventful life to an eager audience.

Austen has inspired a large number of critical essays and literary anthologies. Her novels have inspired many films, from 1940's *Pride and Prejudice* to more recent productions like *Sense and Sensibility* (1995), *Emma* (1996), *Mansfield Park* (1999), *Pride & Prejudice* (2005), and *Love & Friendship* (2016).

Contents [hide]

Jane Austen

Portrait, c. 1810[a]

Born	16 December 1775 Steventon Rectory, Hampshire, England
Died	18 July 1817 (aged 41) Winchester, Hampshire, England
Resting place	Winchester Cathedral, Hampshire, England
Education	Reading Abbey Girls' School
Period	1787 to 1809–11
Relatives	James Austen (brother) George Austen (brother) Edward Austen Knight (brother) Henry Thomas Austen (brother) Cassandra Austen (sister) Sir Francis Austen (aka Francis, brother) Charles Austen (brother) Eliza de Feuillide (cousin)
Signature	Jane Austen

Generalise the web page

```
author_first_name = "Jane"
author_last_name = "Austen"
author_url <- paste("https://en.wikipedia.org/wiki/",
  author_first_name, "_", author_last_name, sep = "")
wiki_data <- read_html(author_url)
```

Let's get that table

```r
table_again <- wiki_data %>%
  rvest::html_nodes(".infobox.vcard") %>% #search for a class
  rvest::html_table(header = FALSE) %>%
  purrr::pluck(1)
head(table_again)
```

```
##                       X1
## 1          Jane Austen
## 2 Portrait, c. 1810[a]
## 3                 Born
## 4                 Died
## 5        Resting place
## 6            Education
##                                                                   X2
## 1                                                       Jane Austen
## 2                                              Portrait, c. 1810[a]
## 3 (1775-12-16)16 December 1775Steventon Rectory, Hampshire, England
## 4  18 July 1817(1817-07-18) (aged 41)Winchester, Hampshire, England
## 5                          Winchester Cathedral, Hampshire, England
```

Web scraping is tricky

No nationality category in Jane Austen's table

```
##  [1] "Jane Austen"          "Portrait, c. 1810[a]" "Born"                
##  [4] "Died"                 "Resting place"        "Education"           
##  [7] "Period"               "Relatives"            ""                    
## [10] "Signature"
```

Differnt way - Matching paragraphs

```r
para_data <- wiki_data %>%
  rvest::html_nodes("p") # get all the paragraphs
head(para_data)
```

```
## {xml_nodeset (6)}
## [1] <p class="mw-empty-elt">\n\n\n\n</p>
## [2] <p><b>Jane Austen</b> (<span class="nowrap"><span class="IPA nopopup ...
## [3] <p>With the publications of <i><a href="/wiki/Sense_and_Sensibility" ...
## [4] <p>A significant transition in her posthumous reputation occurred in ...
## [5] <p>Austen has inspired a large number of critical essays and literar ...
## [6] <p>There is little biographical information about Jane Austen's life ...
```

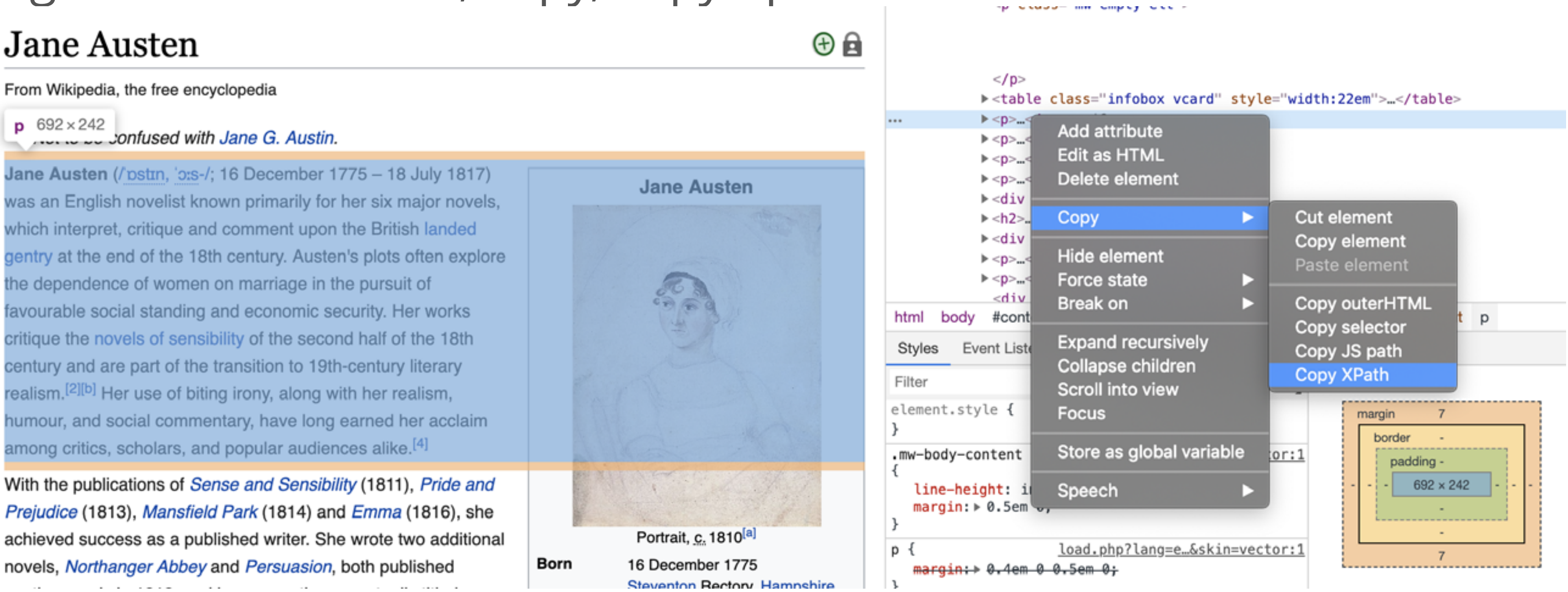
Get the text - html_text()

```r
text_data <- para_data %>%
  purrr::pluck(2) %>% # get the second paragraph
  rvest::html_text() # convert the paragraph to text
head(text_data)
```

```
## [1] "Jane Austen (/ˈɒstɪn, ˈɔːs-/; 16 December 1775 – 18 July 1817) was an
English novelist known primarily for her six major novels, which interpret,
critique and comment upon the British landed gentry at the end of the 18th
century. Austen's plots often explore the dependence of women on marriage in
the pursuit of favourable social standing and economic security. Her works
critique the novels of sensibility of the second half of the 18th century and
are part of the transition to 19th-century literary realism.[2][b] Her use of
biting irony, along with her realism, humour, and social commentary, have long
earned her acclaim among critics, scholars, and popular audiences alike.[4]"
```

Xpath Example

- Right click html code, copy, copy Xpath



Jane Austen

From Wikipedia, the free encyclopedia

p 692 × 242

confused with *Jane G. Austin*.

Jane Austen (/ˈɒstɪn, ˈɔːs-/; 16 December 1775 – 18 July 1817) was an English novelist known primarily for her six major novels, which interpret, critique and comment upon the British landed gentry at the end of the 18th century. Austen's plots often explore the dependence of women on marriage in the pursuit of favourable social standing and economic security. Her works critique the novels of sensibility of the second half of the 18th century and are part of the transition to 19th-century literary realism.[2][b] Her use of biting irony, along with her realism, humour, and social commentary, have long earned her acclaim among critics, scholars, and popular audiences alike.[4]

With the publications of *Sense and Sensibility* (1811), *Pride and Prejudice* (1813), *Mansfield Park* (1814) and *Emma* (1816), she achieved success as a published writer. She wrote two additional novels, *Northanger Abbey* and *Persuasion*, both published

Jane Austen

Portrait, c. 1810[a]

Born	16 December 1775 Steventon Rectory, Hampshire

```
  </p>
▸<table class="infobox vcard" style="width:22em">…</table>
▸<p>…
▸<p>…
▸<p>…
▸<p>…
▸<div
▸<h2>…
▸<div
▸<p>…
▸<p>…
 <div
```

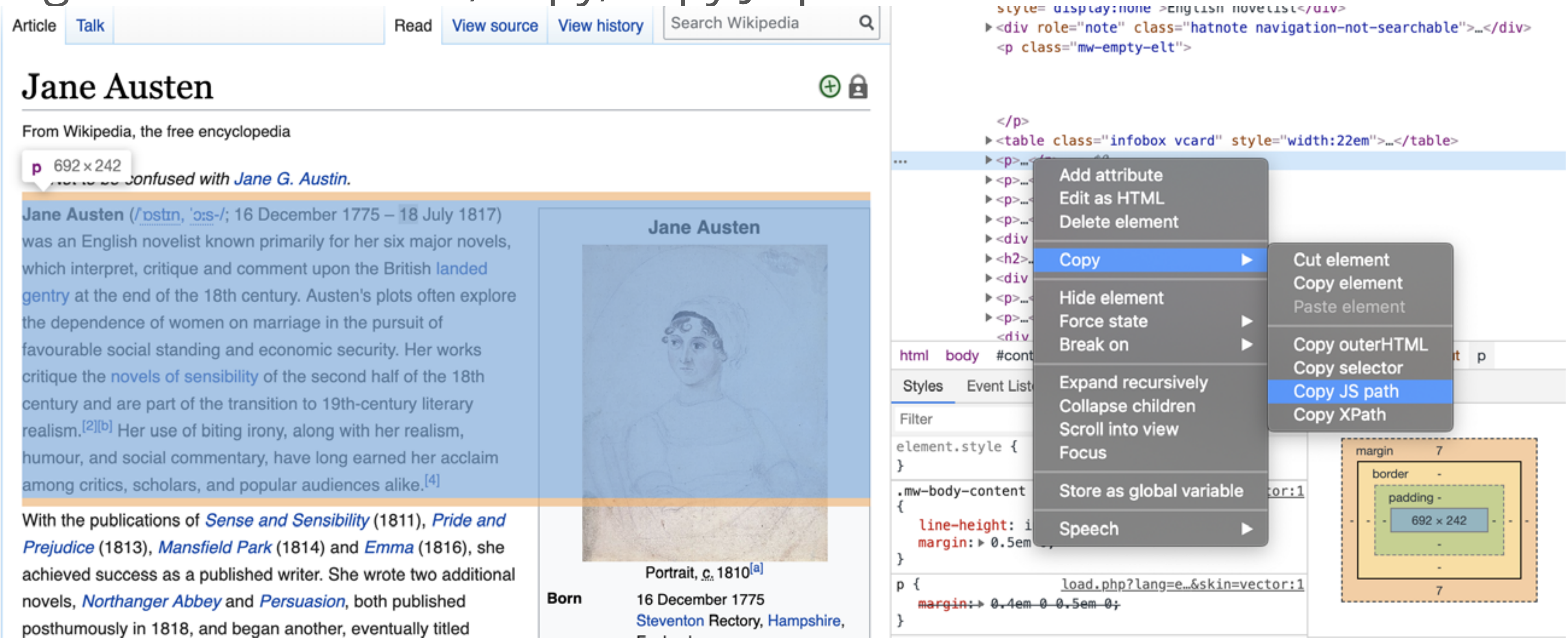
```
element.style {
}
.mw-body-content                        or:1
{
  line-height: i
  margin:▸ 0.5em 0;
}
p {                load.php?lang=e…&skin=vector:1
  ~~margin:▸ 0.4em 0 0.5em 0;~~
}
```

Using an Xpath

```
para_xpath = '//*[@id="mw-content-text"]/div/p[2]'
text_data <- wiki_data %>%
  rvest::html_nodes(xpath = para_xpath) %>%
  rvest::html_text()
text_data
```

JSpath Example

- Right click html code, copy, copy JS path



Jane Austen

From Wikipedia, the free encyclopedia

Not to be confused with *Jane G. Austin*.

Jane Austen (/ˈɒstɪn, ˈɔːs-/; 16 December 1775 – 18 July 1817) was an English novelist known primarily for her six major novels, which interpret, critique and comment upon the British landed gentry at the end of the 18th century. Austen's plots often explore the dependence of women on marriage in the pursuit of favourable social standing and economic security. Her works critique the novels of sensibility of the second half of the 18th century and are part of the transition to 19th-century literary realism.[2][b] Her use of biting irony, along with her realism, humour, and social commentary, have long earned her acclaim among critics, scholars, and popular audiences alike.[4]

With the publications of *Sense and Sensibility* (1811), *Pride and Prejudice* (1813), *Mansfield Park* (1814) and *Emma* (1816), she achieved success as a published writer. She wrote two additional novels, *Northanger Abbey* and *Persuasion*, both published posthumously in 1818, and began another, eventually titled

Jane Austen

Portrait, c. 1810[a]

Born	16 December 1775 Steventon Rectory, Hampshire,

```
style="display:none">English novelist</div>
▸<div role="note" class="hatnote navigation-not-searchable">…</div>
 <p class="mw-empty-elt">

 </p>
▸<table class="infobox vcard" style="width:22em">…</table>
▸<p>…</p>
▸<p>…
▸<p>…
▸<p>…
▸<div
▸<h2>…
▸<div
▸<p>…
▸<p>…
 <div
```

```
element.style {
}
.mw-body-content                                     tor:1
{
   line-height: i
   margin:▸ 0.5em 0;
}
p {                          load.php?lang=e…&skin=vector:1
   ~~margin:▸ 0.4em 0 0.5em 0;~~
}
```

Using CSS ID

```r
para_css = "#mw-content-text > div > p:nth-child(5)"
text_data <- wiki_data %>%
  rvest::html_nodes(css = para_css) %>%
  rvest::html_text()
text_data
```

```
## [1] "Jane Austen (/ˈɒstɪn, ˈɔːs-/; 16 December 1775 – 18 July 1817) was an
English novelist known primarily for her six major novels, which interpret,
critique and comment upon the British landed gentry at the end of the 18th
century. Austen's plots often explore the dependence of women on marriage in
the pursuit of favourable social standing and economic security. Her works
critique the novels of sensibility of the second half of the 18th century and
are part of the transition to 19th-century literary realism.[2][b] Her use of
biting irony, along with her realism, humour, and social commentary, have long
earned her acclaim among critics, scholars, and popular audiences alike.[4]"
```

Text Analysis

```r
possible_nationalities <- c("Australian", "Chinese", "Mexican", "English", "Ethiopian")

# Do any of these nationalities appear in the text?
count_values = str_count(text_data, possible_nationalities)
possible_nationalities[count_values == TRUE]
```

```
## [1] "English"
```

Learnt so far

- Know how to explore a web page with inspect
- Know some basics about how to get data

Also know:

- Can be hard to generalise
- Formats aren't always standard

Learnt so far

- Know how to explore a web page with inspect
- Know some basics about how to get data

Also know:

- Can be hard to generalise
- Formats aren't always standard

Back to the original question ...

Get our list

1001 books to read

Read the book list from a website

```
book_list_url <- "https://mizparker.wordpress.com/the-lists/1001-books-to-read-before-you-di
paragraph_data <- read_html(book_list_url) %>% # read the web page
  rvest::html_nodes("p") # get the paragraphs
head(paragraph_data)
```

```
## {xml_nodeset (6)}
## [1] <p>This list has appeared in several places around the internet, and ...
## [2] <p>If you would like to download a spreadsheet of the list and keep  ...
## [3] <p><strong>21st Century:</strong></p>
## [4] <p>1.  Never Let Me Go – Kazuo Ishiguro<br>\n2.  Saturday – Ian McEw ...
## [5] <p><strong>20th Century:</strong></p>
## [6] <p>70. Timbuktu – Paul Auster<br>\n71. The Romantics – Pankaj Mishra ...
```

Get the book list from the paragraphs

```
book_string <- paragraph_data %>%  #the list is in pieces
  purrr::pluck(4) %>% # get the first part of the list
  html_text(trim = TRUE) %>% # convert it to text, remove white space
  gsub("\n", "", .) #remove the newline character
head(book_string)
```

```
## [1] "1.  Never Let Me Go – Kazuo Ishiguro2.  Saturday – Ian McEwan3.  On
Beauty – Zadie Smith4.  Slow Man – J.M. Coetzee5.  Adjunct: An Undigest – Peter
Manson6.  The Sea – John Banville7.  The Red Queen – Margaret Drabble8.  The
Plot Against America – Philip Roth9.  The Master – Colm Toibin10.  Vanishing
Point – David Markson11.  The Lambs of London – Peter Ackroyd12.  Dining on
Stones – Iain Sinclair13.  Cloud Atlas – David Mitchell14.  Drop City – T.
Coraghessan Boyle15.  The Colour – Rose Tremain16.  Thursbitch – Alan Garner17.
The Light of Day – Graham Swift18.  What I Loved – Siri Hustvedt19.  The
Curious Incident of the Dog in the Night-Time – Mark Haddon20.  Islands – Dan
Sleigh21.  Elizabeth Costello – J.M. Coetzee22.  London Orbital – Iain
Sinclair23.  Family Matters – Rohinton Mistry24.  Fingersmith – Sarah Waters25.
The Double – Jose Saramago26.  Everything is Illuminated – Jonathan Safran
Foer27.  Unless – Carol Shields28.  Kafka on the Shore – Haruki Murakami29.
```

Let's put our list together

But …. web scraping often means string handling

We want to split the string by any numbers followed by a full stop

Careful:

- don't want to split book titles with numbers, like Catch 22,
- don't want to split authors with full stops, like J.R.R Tolkien

Actually bit tricky!

Resources:

- https://regexr.com/
- stringr cheatsheet from RStudio

Do some string handling

```r
strsplit("a123b", split = "\\d")
  #Split by digits \\d
strsplit("a123b", split = "\\d+")
  #Split by one or more digits \\d+
strsplit("a.b", split = "\\.")
  #Split by fullstop \\.
strsplit("a1.b", split = "\\d+\\.")
  #Split by digits and fullstop \\d+\\.
strsplit("a1.b", split = "\\d+?\\.")
  #Matches as few digits as possible \\d+? and fullstop \\.
```

Check out: https://regexr.com/

Split up our list

```r
split_book_string <- book_string %>%
  strsplit(split = "\\d+?\\.") %>%
    # split the string by any numbers followed by a full stop
  as.data.frame(stringsAsFactors = FALSE) %>%
    # make this a data frame
  dplyr::filter(. != "")
    # remove any empty rows
head(split_book_string)
```

```
##   c........Never.Let.Me.Go...Kazuo.Ishiguro......Saturday...Ian.McEwan...
## 1                                        Never Let Me Go - Kazuo Ishiguro
## 2                                                   Saturday - Ian McEwan
## 3                                                 On Beauty - Zadie Smith
## 4                                                 Slow Man - J.M. Coetzee
## 5                                     Adjunct: An Undigest - Peter Manson
## 6                                                 The Sea - John Banville
```

Split up our columns

```r
names(split_book_string) <- "book_string"
book_df <-split_book_string %>%
  tidyr::separate(book_string, sep = "\\-", into = c("book", "author"))
  # split our author and book into columns
  # very lucky that whoever coded this webpage used a long hash!
head(book_df)
```

```
##                       book          author
## 1        Never Let Me Go   Kazuo Ishiguro
## 2               Saturday       Ian McEwan
## 3              On Beauty      Zadie Smith
## 4               Slow Man     J.M. Coetzee
## 5   Adjunct: An Undigest     Peter Manson
## 6                The Sea    John Banville
```

Wrap the code chunks

Could vectorise it properly, but we leave that for later.
For now we'll just wrap that our code snippets in a function and use lapply

```r
Get_book_data <- function(para_ind){

  book_str <- paragraph_data %>%
    purrr::pluck(para_ind) %>%
    html_text(trim = TRUE) %>%
    gsub("\n", "", .)  #remove newline character

  book_df <- book_str %>%
    strsplit(split = "\\d+?\\.") %>% #match the number index
    as.data.frame(stringsAsFactors = FALSE) %>%
    dplyr::filter(. != "")  # remove empty first row

  names(book_df) <- "book_string"
  book_df <- book_df %>%
    tidyr::separate(book_string, sep = "\\-", into = c("book", "author"))
```

Put it together

```
book_data <- lapply(seq(4,12,2) %>% as.list(), Get_book_data) %>%
  do.call(rbind, .) %>%
  dplyr::mutate(author = str_trim(author))
nrow(book_data) # Has 1001 rows so let's assume we are all good!
```

```
## [1] 1001
```

```
head(book_data) # Looks pretty good at first glance
```

```
##                        book          author
## 1        Never Let Me Go  Kazuo Ishiguro
## 2               Saturday      Ian McEwan
## 3              On Beauty     Zadie Smith
## 4               Slow Man     J.M. Coetzee
## 5   Adjunct: An Undigest    Peter Manson
## 6                The Sea   John Banville
```

Now let's get the nationalities of all the authors!

Get nationalities

More wrapping

Also wrap our code up pieces to get the nationality

```
search_string = "Tim Winton"
wiki_data <- Read_wiki_page(search_string)
infocard <- Get_wiki_infocard(wiki_data)
if(is.null(infocard)){
  nationality <- "Missing infocard"
}else if(any(infocard[,1] == "Nationality")){
  nationality <- Get_nationality_from_infocard(infocard)
}else{
  first_paragraph <- Get_first_text(wiki_data)
  nationality <- Guess_nationality_from_text(first_paragraph,
    possible_nationalities)
}
nationality
```

What nationalities to search for?

We need a list of nationalities for searching.
Let's get one!

```
# Get table of nationalities
url <- "http://www.vocabulary.cl/Basic/Nationalities.htm"
xpath <- "/html/body/div[1]/article/table[2]"
nationalities_df <- url %>%
  read_html() %>%
  html_nodes(xpath = xpath) %>%
  html_table() %>%
  as.data.frame()

possible_nationalities = nationalities_df[,2]
head(possible_nationalities)
```

```
## [1] "Afghan"               "Albanian"             "Algerian"
## [4] "ArgentineArgentinian" "Australian"           "Austrian"
```

Manual fixing

```
fix_entry = "ArgentineArgentinian"
i0 = which(nationalities_df == fix_entry, arr.ind = TRUE)
new_row = nationalities_df[i0[1], ]
nationalities_df[i0] = "Argentine"
new_row[,2] = "Argentinian"
nationalities_df = rbind(nationalities_df, new_row)

fix_footnote1 = "Colombia *"
i1 = which(nationalities_df == fix_footnote1, arr.ind = TRUE)
nationalities_df[i1] = strsplit(fix_footnote1, split = ' ')[[1]][1]

fix_footnote2 = "American **"
i2 = which(nationalities_df == fix_footnote2, arr.ind = TRUE)
nationalities_df[i2] = strsplit(fix_footnote2, split = ' ')[[1]][1]

possible_nationalities = nationalities_df[,2]
```

Get Nationalities

```
nationality_from_author_search = sapply(book_data$author[1:20],
                                        function(search_string){
  nataionality = tryCatch( # Just in case!
     Query_nationality_from_wiki(search_string,
                                 possible_nationalities),
    error = function(e) NA)
  }) %>% unlist()
nationality_from_author_search
```

How diversely do we read

Run it!

```r
nationality_from_author_search = sapply(book_data$author %>% unique(),
                                        function(search_string){
  print(search_string)
  nataionality = tryCatch( # Just in case!
     Query_nationality_from_wiki(search_string,
                                 possible_nationalities),
    error = function(e) NA)
  })
author_nationality_df <- as.data.frame(nationality_from_author_search) %>%
  dplyr::mutate(author = rownames(.))
names(author_nationality_df) <- c("nationality", "author")
book_data <- book_data %>%
  dplyr::left_join(author_nationality_df)
head(book_data)

save(book_data, file = "book_data.RData")
```

Result

```r
load("book_data.RData")
table_nationalities <- book_data %>%
  dplyr::select(author, nationality) %>%
  dplyr::distinct() %>%
  dplyr::select(nationality) %>%
  unlist() %>%
  table() %>%
  as.data.frame(stringsAsFactors = FALSE)
names(table_nationalities ) = c("Nationality", "Frequency")
table_nationalities %>%
  arrange(desc(Frequency))
```

```
##                                            Nationality Frequency
## 1                                             American       114
## 2                                              English        68
## 3                                              British        57
## 4                                     Missing infocard        47
## 5                                               French        42
## 6                                                Irish        19
```

Let's take a look

```
pie_plot <- table_nationalities %>%
  plot_ly(labels = ~Nationality, values = ~Frequency) %>%
  add_pie(hole = 0.6) %>%
  layout(title = "Nationalities",  showlegend = F,
         xaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE),
         yaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE))
```

Plotting result

A Few thoughts

- So many ways to approach this problem
- Approach here was to use the standard rvest toolbox
- Not perfect - much needed cleaning of nationality strings
- A bit of quessing of nationalities

What else could we have done

- Can scrape more data from goodreads website
- Goodreads has an API
- Check out the repository by famguy/rgoodreads to get started
- Using this API makes querying things like year or gender straightforward
- But goodreads has no nationality, so this solution still is useful!

What else for webscraping

- There are easier ways to answer this same question
- Namely, RSelenium for pages with javascript
- Learning the hard way can be good sometimes though!