# Data science with R

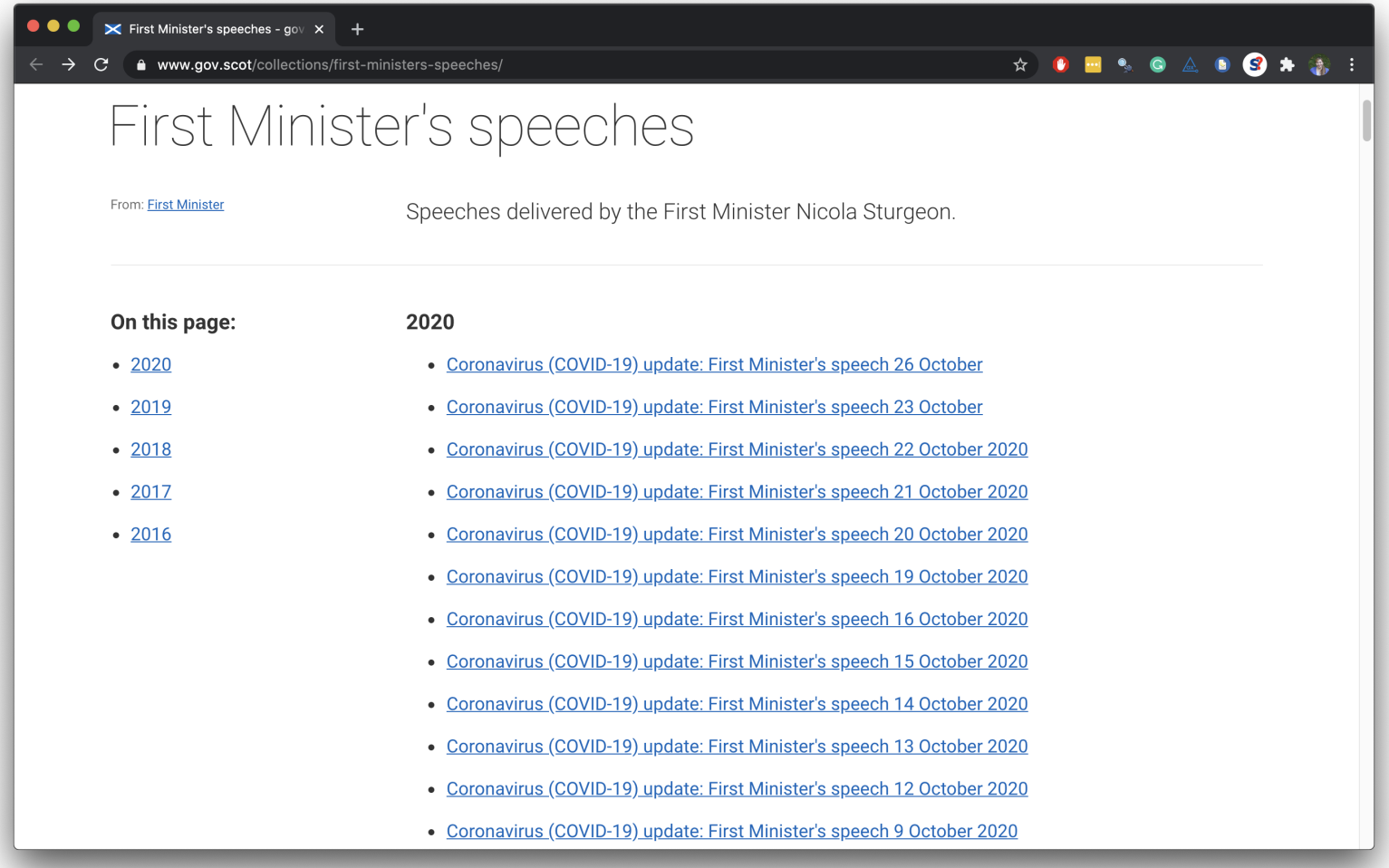
## Functions

Prfo. Dr. Jan Kirenz

*The following content is based on Mine Çetinkaya-Rundel's excellent book Data Science in a Box*

# First Minister's COVID speeches

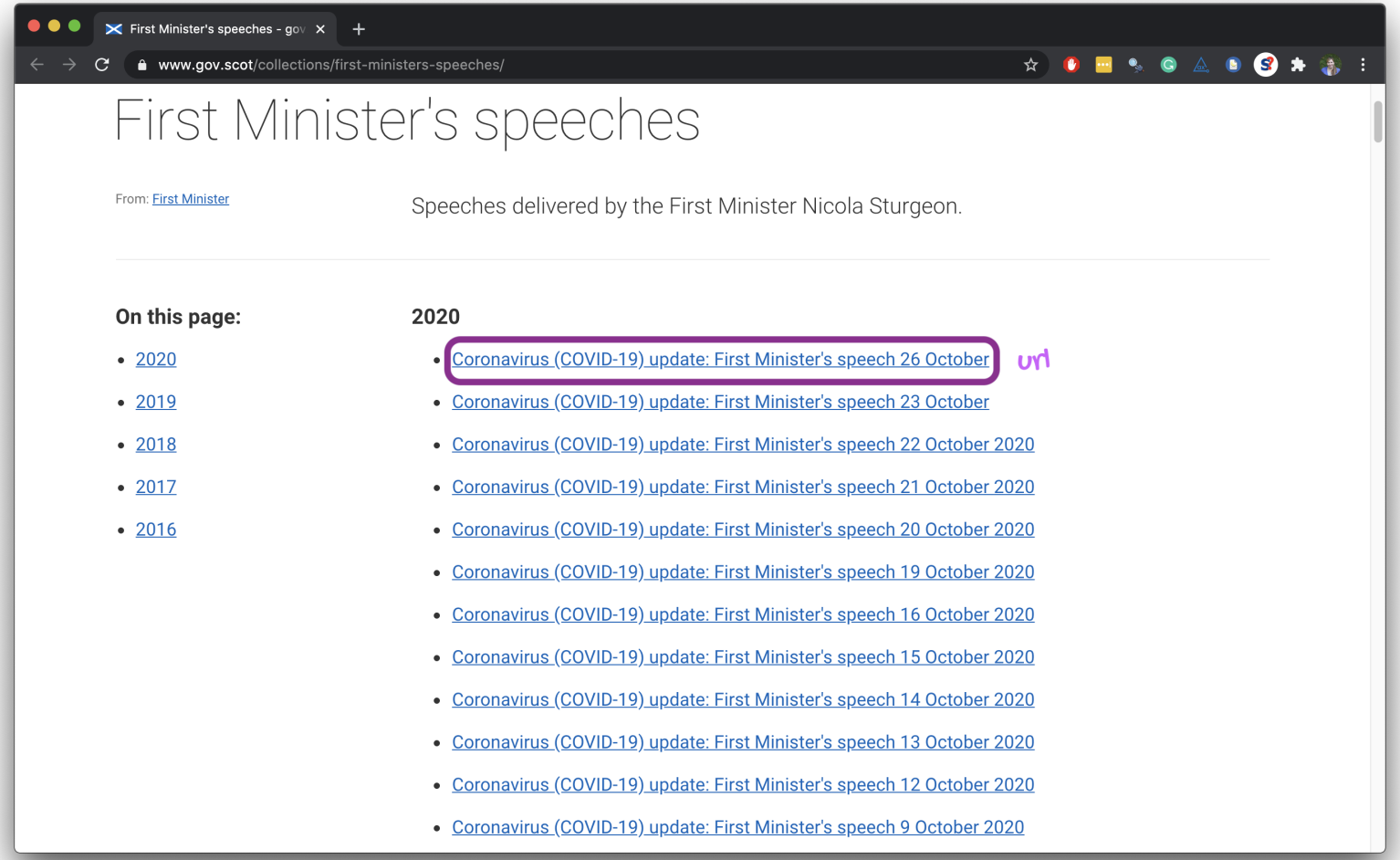
# 🏁 Start with


First Minister's speeches - gov
www.gov.scot/collections/first-ministers-speeches/
First Minister's speeches
From: First Minister
Speeches delivered by the First Minister Nicola Sturgeon.
On this page:
2020
2019
2018
2017
2016
2020
Coronavirus (COVID-19) update: First Minister's speech 26 October
Coronavirus (COVID-19) update: First Minister's speech 23 October
Coronavirus (COVID-19) update: First Minister's speech 22 October 2020
Coronavirus (COVID-19) update: First Minister's speech 21 October 2020
Coronavirus (COVID-19) update: First Minister's speech 20 October 2020
Coronavirus (COVID-19) update: First Minister's speech 19 October 2020
Coronavirus (COVID-19) update: First Minister's speech 16 October 2020
Coronavirus (COVID-19) update: First Minister's speech 15 October 2020
Coronavirus (COVID-19) update: First Minister's speech 14 October 2020
Coronavirus (COVID-19) update: First Minister's speech 13 October 2020
Coronavirus (COVID-19) update: First Minister's speech 12 October 2020
Coronavirus (COVID-19) update: First Minister's speech 9 October 2020

# End with 🛑

```
## # A tibble: 160 x 6
##    title     date       location   abstract     text     url
##    <chr>     <date>     <chr>      <chr>        <chr>    <chr>
##  1 Coronavi… 2020-11-25 St Andrew… Statement g… "Good a… https:/…
##  2 Coronavi… 2020-11-24 Scottish … Statement g… "Thanks… https:/…
##  3 Coronavi… 2020-11-23 St Andrew… Statement g… "\nGood… https:/…
##  4 Coronavi… 2020-11-20 St Andrew… Statement g… "\nThan… https:/…
##  5 Coronavi… 2020-11-18 St Andrew… Statement g… "Thanks… https:/…
##  6 Coronavi… 2020-11-17 Scottish … Statement g… "Presid… https:/…
##  7 Coronavi… 2020-11-16 St Andrew… Statement g… "Thank … https:/…
##  8 Coronavi… 2020-11-16 St Andrew… Statement g… "\nThan… https:/…
##  9 Coronavi… 2020-11-10 Scottish … Statement g… "Presid… https:/…
## 10 Coronavi… 2020-11-09 St Andrew… Statement g… "\nThan… https:/…
## 11 Coronavi… 2020-11-06 St Andrew… Statement g… "Thanks… https:/…
## 12 Coronavi… 2020-11-04 St Andrew… Statement g… "\nThan… https:/…
## 13 Coronavi… 2020-11-03 St Andrew… Statement g… "Thanks… https:/…
## 14 Coronavi… 2020-11-02 St Andrew… Statement g… "\nThan… https:/…
## 15 Coronavi… 2020-10-31 <NA>       Statement g… "The Fi… https:/…
## # … with 145 more rows
```

# www.gov.scot/collections/first-ministers-speeches



Coronavirus (COVID-19) updat

www.gov.scot/publications/coronavirus-covid-19-update-first...

# Coronavirus (COVID-19) update: First Minister's speech 26 October

title

Published 26 Oct 2020

date

From: First Minister

Part of: Coronavirus in Scotland, Public safety and emergencies

Delivered by: First Minister Nicola Sturgeon

Location: St Andrew's House, Edinburgh

location

Statement given by First Minister Nicola Sturgeon at a media briefing in St Andrew's House on Monday 26 October 2020.

abstract

This document is part of a collection


Good afternoon, and thanks for joining us. I want to start with the usual daily report on the COVID statistics.

The total number of positive cases reported yesterday was 1,122.

This represents 7.1% of the total number of tests carried out. 428 of the new cases were in Greater Glasgow and Clyde, 274 in Lanarkshire, 105 in Lothian and

text

# Plan

1. Scrape `title`, `date`, `location`, `abstract`, and `text` from a few COVID-19 speech pages to develop the code

2. Write a function that scrapes `title`, `date`, `location`, `abstract`, and `text` from COVID-19 speech pages

3. Scrape the `url`s of COVID-19 speeches from the main page

4. Use this function to scrape from each individual COVID-19 speech from these `url`s and create a data frame with the columns `title`, `date`, `location`, `abstract`, `text`, and `url`

# Scrape data from a few COVID-19 speech pages

# Read page for 26 Oct speech

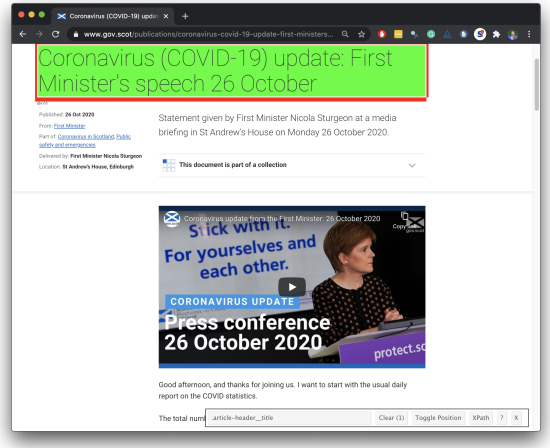
```
url <- "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-spe
speech_page <- read_html(url)
```

```
speech_page
```

```
## {html_document}
## <html dir="ltr" lang="en">
## [1] <head>\n<meta http-equiv="Content-Type" content
## [2] <body class="fontawesome site-header__container
```

# Extract title

```
title <- speech_page %>%
    html_node(".article-header__title") %>%
    html_text()

title
```




```
## [1] "Coronavirus (COVID-19) update: First Minister's speech 26 October"
```

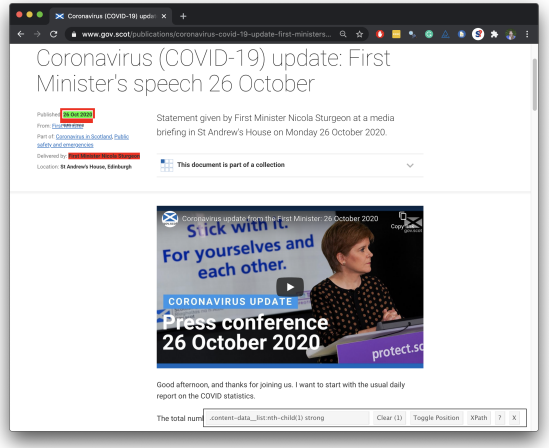
# Extract date

```
library(lubridate)

speech_page %>%
    html_node(".content-data__list:nth-child(1) strong") %>%
    html_text()
```

```
## [1] "26 Oct 2020"
```

```
date <- speech_page %>%
    html_node(".content-data__list:nth-child(1) strong") %>%
    html_text() %>%
    dmy()
date
```

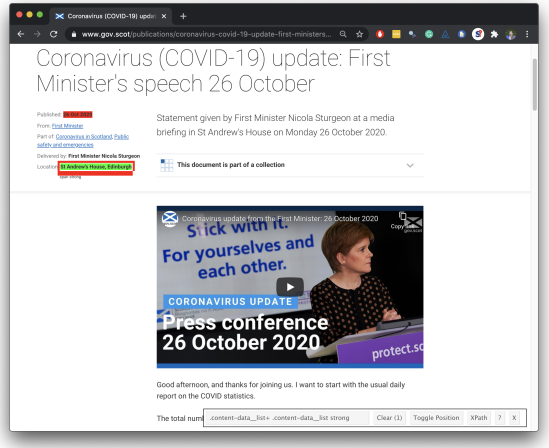
```
## [1] "2020-10-26"
```



# Extract location

```
location <- speech_page %>%
    html_node(".content-data__list+ .content-data__list strong") %>%
    html_text()

location
```

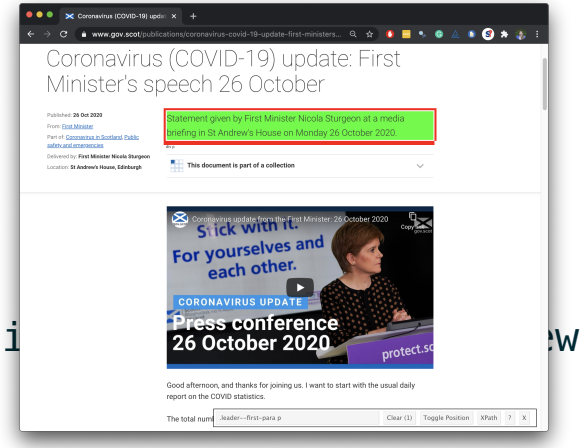
```
## [1] "St Andrew's House, Edinburgh"
```



# Extract abstract

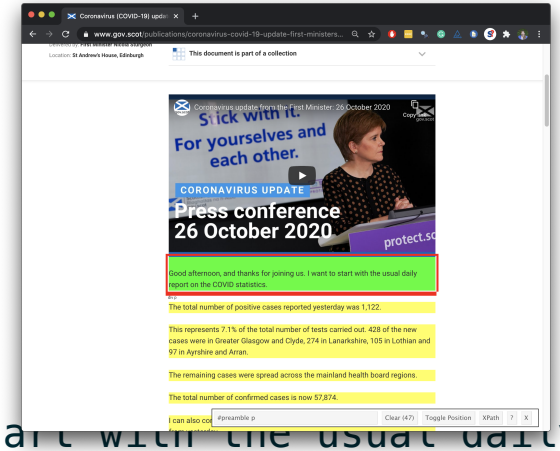
```
abstract <- speech_page %>%
    html_node(".leader--first-para p") %>%
    html_text()

abstract
```

```
## [1] "Statement given by First Minister Nicola Sturgeon at a medi                ew's Ho
```



# Extract text

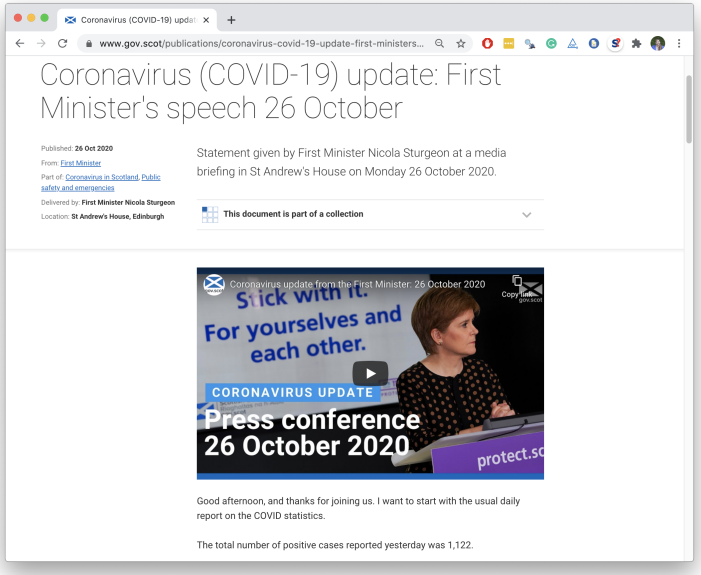
```
text <- speech_page %>%
    html_nodes("#preamble p") %>%
    html_text() %>%
    list()

text
```




```
## [[1]]
##  [1] "\nGood afternoon, and thanks for joining us. I want to start with the usual daily repo
##  [2] "The total number of positive cases reported yesterday was 1,122."
##  [3] "This represents 7.1% of the total number of tests carried out. 428 of the new cases we
##  [4] "The remaining cases were spread across the mainland health board regions. "
##  [5] "The total number of confirmed cases is now 57,874."
##  [6] "I can also confirm that 1,152 people are in hospital – that is an increase of 36 from
##  [7] "90 people are in intensive care, which is four more than yesterday."
##  [8] "And I regret to say that in the last 24 hours, one further death has been registered o
##  [9] "We also reported 11 deaths on Saturday, and one yesterday.  So since the last bri
## [10] "That reminds us again of how dangerous this virus can be and I want to send my condole
...
```

# Put it all in a data frame

```
oct_26_speech <- tibble(
  title    = title,
  date     = date,
  location = location,
  abstract = abstract,
  text     = text,
  url      = url
)

oct_26_speech
```




```
## # A tibble: 1 x 6
##   title       date       location   abstract     text  url
##   <chr>       <date>     <chr>      <chr>        <lis> <chr>
## 1 Coronaviru… 2020-10-26 St Andrew… Statement g… <chr… https://w…
```

# Read page for 23 Oct speech

```
url <- "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-spe
speech_page <- read_html(url)
```

```
speech_page
```

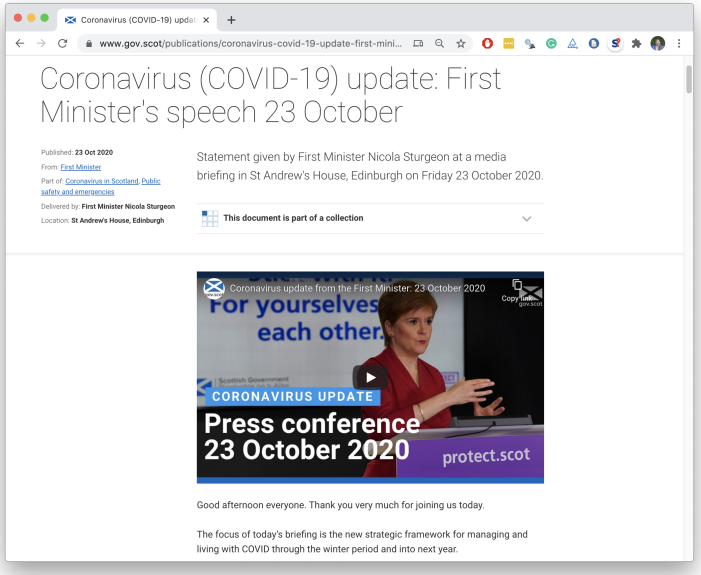
```
## {html_document}
## <html dir="ltr" lang="en">
## [1] <head>\n<meta http-equiv="Content-Type" content="text/html ...
## [2] <body class="fontawesome site-header__container">\n\n\n\n\ ...
```

# Extract components of 23 Oct speech

```
title <- speech_page %>%
  html_node(".article-header__title") %>%
  html_text()

date <- speech_page %>%
  html_node(".content-data__list:nth-child(1) strong") %>%
  html_text() %>%
  dmy()

location <- speech_page %>%
  html_node(".content-data__list+ .content-data__list strong") %>%
  html_text()

abstract <- speech_page %>%
  html_node(".leader--first-para p") %>%
  html_text()

text <- speech_page %>%
  html_nodes("#preamble p") %>%
  html_text() %>%
  list()
```

# Put it all in a data frame

```
oct_23_speech <- tibble(
  title    = title,
  date     = date,
  location = location,
  abstract = abstract,
  text     = text,
  url      = url
)

oct_23_speech
```

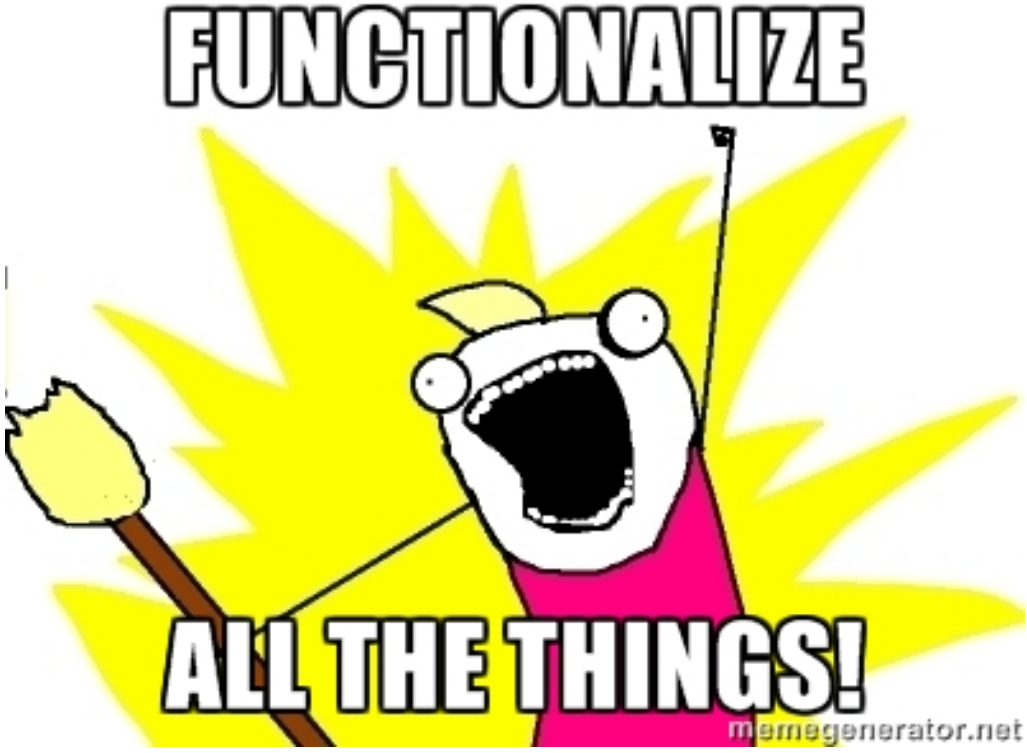



```
## # A tibble: 1 x 6
##   title       date       location   abstract     text  url
##   <chr>       <date>     <chr>      <chr>        <lis> <chr>
## 1 Coronaviru… 2020-10-23 St Andrew… Statement g… <chr… https://w…
```

*this is getting tiring...*

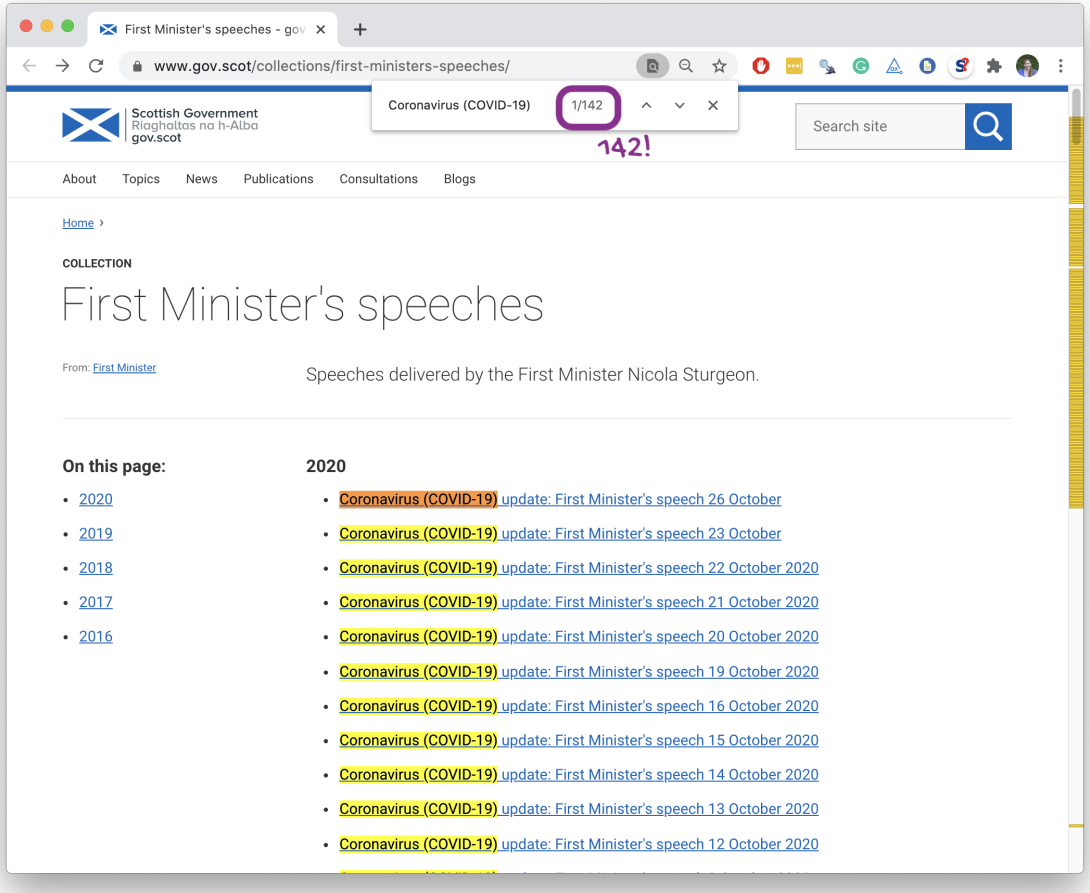
# Functions

# When should you write a function?




When you've copied and pasted a block of code more than twice.

# How many times will we need to copy and paste the code we developed to scrape data on all of First Minister's COVID-19 speeches?



# Why functions?

- Automate common tasks in a more powerful and general way than copy-and-pasting:
    - Give your function an evocative name that makes your code easier to understand
    - As requirements change, only need to update code in one place, instead of many
    - Eliminate chance of making incidental mistakes when you copy and paste (i.e. updating a variable name in one place, but not in another)
- Down the line: Improve your reach as a data scientist by writing functions (and packages!) that others use

Assuming that the page structure is the same for each speech page, how many "things" do you need to know for each speech page to scrape the data we want from it?

```
url_23_oct <- "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-speech-23-oct
speech_page <- read_html(url_23_oct)

title <- speech_page %>%
  html_node(".article-header__title") %>%
  html_text()

date <- speech_page %>%
  html_node(".content-data__list:nth-child(1) strong") %>%
  html_text() %>%
  dmy()

location <- speech_page %>%
  html_node(".content-data__list+ .content-data__list strong") %>%
  html_text()

abstract <- speech_page %>%
  html_node(".leader--first-para p") %>%
  html_text()

text <- speech_page %>%
  html_nodes("#preamble p") %>%
  html_text() %>%
  list()

tibble(
  title = title, date = date, location = location,
  abstract = abstract, text = text, url= url
)
```

# Turn your code into a function

- Pick a short but informative **name**, preferably a verb.

```
scrape_speech <-
```

# Turn your code into a function

- Pick a short but evocative **name**, preferably a verb.
- List inputs, or **arguments**, to the function inside `function`. If we had more the call would look like `function(x, y, z)`.

```
scrape_speech <- function(x){



}
```

# Turn your code into a function

- Pick a short but informative **name**, preferably a verb.
- List inputs, or **arguments**, to the function inside `function`. If we had more the call would look like `function(x, y, z)`.
- Place the **code** you have developed in body of the function, a `{` block that immediately follows `function(...)`.

```
scrape_speech <- function(url){

  # code we developed earlier to scrape info
  # on single art piece goes here

}
```

# scrape_speech()

```
scrape_speech <- function(url) {

  speech_page <- read_html(url)

  title <- speech_page %>%
    html_node(".article-header__title") %>%
    html_text()

  date <- speech_page %>%
    html_node(".content-data__list:nth-child(1) strong") %>%
    html_text() %>%
    dmy()

  location <- speech_page %>%
    html_node(".content-data__list+ .content-data__list strong") %>%
    html_text()

  abstract <- speech_page %>%
    html_node(".leader--first-para p") %>%
    html_text()

  text <- speech_page %>%
    html_nodes("#preamble p") %>%
    html_text() %>%
    list()

  tibble(
    title = title, date = date, location = location,
    abstract = abstract, text = text, url = url
  )
}
```

# Function in action

```
scrape_speech(url = "https://www.gov.scot/publications/coronavirus-covid-19-update-first-
  glimpse()
```

```
## Rows: 1
## Columns: 6
## $ title    <chr> "Coronavirus (COVID-19) update: First Ministe…
## $ date     <date> 2020-10-26
## $ location <chr> "St Andrew's House, Edinburgh"
## $ abstract <chr> "Statement given by First Minister Nicola Stu…
## $ text     <list> [<"\nGood afternoon, and thanks for joining …
## $ url      <chr> "https://www.gov.scot/publications/coronaviru…
```

# Function in action

```
scrape_speech(url = "https://www.gov.scot/publications/coronavirus-covid-19-update-first-
  glimpse()
```

```
## Rows: 1
## Columns: 6
## $ title    <chr> "Coronavirus (COVID-19) update: First Ministe…
## $ date     <date> 2020-10-23
## $ location <chr> "St Andrew's House, Edinburgh"
## $ abstract <chr> "Statement given by First Minister Nicola Stu…
## $ text     <list> [<"\nGood afternoon everyone. Thank you very…
## $ url      <chr> "https://www.gov.scot/publications/coronaviru…
```

# Function in action

```
scrape_speech(url = "https://www.gov.scot/publications/coronavirus-covid-19-update-first-
  glimpse()
```

```
## Rows: 1
## Columns: 6
## $ title    <chr> "Coronavirus (COVID-19) update: First Ministe…
## $ date     <date> 2020-10-22
## $ location <chr> "St Andrew's House, Edinburgh"
## $ abstract <chr> "Statement given by First Minister Nicola Stu…
## $ text     <list> [<"\nGood afternoon, let me start as usual w…
## $ url      <chr> "https://www.gov.scot/publications/coronaviru…
```

# Writing functions

# What goes in / what comes out?

- They take input(s) defined in the function definition

```
function([inputs separated by commas]){
  # what to do with those inputs
}
```

- By default they return the last value computed in the function

```
scrape_page <- function(x){
  # do bunch of stuff with the input...

  # return a tibble
  tibble(...)
}
```

- You can define more outputs to be returned in a list as well as nice print methods (but we won't go there for now...)

## What is going on here?

```
add_2 <- function(x){
  x + 2
  1000
}
```

```
add_2(3)
```

```
## [1] 1000
```

```
add_2(10)
```

```
## [1] 1000
```

# Naming functions

> "There are only two hard things in Computer Science: cache invalidation and naming things." - Phil Karlton

# Naming functions

- Names should be short but clearly evoke what the function does
- Names should be verbs, not nouns
- Multi-word names should be separated by underscores (`snake_case` as opposed to `camelCase`)
- A family of functions should be named similarly (`scrape_page()`, `scrape_speech()` OR `str_remove()`, `str_replace()` etc.)
- Avoid overwriting existing (especially widely used) functions

```
# JUST DON'T
mean <- function(x){
  x * 3
  }
```