Class Activity 17

Arthur Viegas Eguoa

May 03 2024

Group Activity 1

1. Go to the the numbers webpage and extract the table on the front page.

```
session1 <- bow(url = "https://www.the-numbers.com/movie/budgets/all") %>%
  scrape() %>%
  html_table() %>%
  purrr::pluck(1) %>%
  janitor::clean_names() %>%
  mutate(across(everything(), as.character))
```

2. Find out the number of pages that contain the movie table, while looking for the changes in the url in the address bar. How does the url changes when you go to the next page?

Answer: The first is 1, the second 101, the third is 201. So the increment is by 100, starting from 1. The last page is 6401.

3. Write a for loop to store all the data in multiple pages to a single data frame. Also, do the same using `purrr::map_df`.

```
library(tidyverse)
library(rvest)

new_urls <- "https://www.the-numbers.com/movie/budgets/all/"

# Create an empty list
df1 <- list()

# Generate a vector of indices
index <- seq(1, 6401, 100)
```

```
#This is bad, it takes a long time
for(i in 1:length(index)){
  url <- str_glue({new_urls}, {index[i]})

  webpage <- read_html(url) #This is the same as bow() %>% scrape()
  table_new <- html_table(webpage)[[1]] %>% #gets the ith table from the list
    janitor::clean_names() %>%
```

```
    mutate(across(everything(), as.character))
  df1[[i]] <- table_new
}

final_data <- bind_rows(df1)

dplyr::glimpse(final_data)
Rows: 6,497
Columns: 6
$ x                <chr> "1", "2", "3", "4", "5", "6", "7", "8", "9", "10", "~
$ release_date     <chr> "Dec 9, 2022", "Apr 23, 2019", "May 20, 2011", "Apr ~
$ movie            <chr> "Avatar: The Way of Water", "Avengers: Endgame", "Pi~
$ production_budget <chr> "$460,000,000", "$400,000,000", "$379,000,000", "$36~
$ domestic_gross   <chr> "$684,075,767", "$858,373,000", "$241,071,802", "$45~
$ worldwide_gross  <chr> "$2,317,514,386", "$2,788,912,285", "$1,045,713,802"~

#This is the base r way of doing that, we should use dplyr
# do.call("rbind", df1)
# reduce(df1, rbind)
```

```
#.x is the object, .f is the function
# urls <- map(.x = index, .f = function(i) str_glue({new_urls}, {index[i]}))
# urls

urls <- map(index, ~str_glue({new_urls}, {.x}))
urls
[[1]]
https://www.the-numbers.com/movie/budgets/all/1

[[2]]
https://www.the-numbers.com/movie/budgets/all/101

[[3]]
https://www.the-numbers.com/movie/budgets/all/201

[[4]]
https://www.the-numbers.com/movie/budgets/all/301

[[5]]
https://www.the-numbers.com/movie/budgets/all/401

[[6]]
https://www.the-numbers.com/movie/budgets/all/501

[[7]]
https://www.the-numbers.com/movie/budgets/all/601

[[8]]
https://www.the-numbers.com/movie/budgets/all/701

[[9]]
https://www.the-numbers.com/movie/budgets/all/801

[[10]]
```

```
https://www.the-numbers.com/movie/budgets/all/901

[[11]]
https://www.the-numbers.com/movie/budgets/all/1001

[[12]]
https://www.the-numbers.com/movie/budgets/all/1101

[[13]]
https://www.the-numbers.com/movie/budgets/all/1201

[[14]]
https://www.the-numbers.com/movie/budgets/all/1301

[[15]]
https://www.the-numbers.com/movie/budgets/all/1401

[[16]]
https://www.the-numbers.com/movie/budgets/all/1501

[[17]]
https://www.the-numbers.com/movie/budgets/all/1601

[[18]]
https://www.the-numbers.com/movie/budgets/all/1701

[[19]]
https://www.the-numbers.com/movie/budgets/all/1801

[[20]]
https://www.the-numbers.com/movie/budgets/all/1901

[[21]]
https://www.the-numbers.com/movie/budgets/all/2001

[[22]]
https://www.the-numbers.com/movie/budgets/all/2101

[[23]]
https://www.the-numbers.com/movie/budgets/all/2201

[[24]]
https://www.the-numbers.com/movie/budgets/all/2301

[[25]]
https://www.the-numbers.com/movie/budgets/all/2401

[[26]]
https://www.the-numbers.com/movie/budgets/all/2501

[[27]]
https://www.the-numbers.com/movie/budgets/all/2601
```

```
[[28]]
https://www.the-numbers.com/movie/budgets/all/2701

[[29]]
https://www.the-numbers.com/movie/budgets/all/2801

[[30]]
https://www.the-numbers.com/movie/budgets/all/2901

[[31]]
https://www.the-numbers.com/movie/budgets/all/3001

[[32]]
https://www.the-numbers.com/movie/budgets/all/3101

[[33]]
https://www.the-numbers.com/movie/budgets/all/3201

[[34]]
https://www.the-numbers.com/movie/budgets/all/3301

[[35]]
https://www.the-numbers.com/movie/budgets/all/3401

[[36]]
https://www.the-numbers.com/movie/budgets/all/3501

[[37]]
https://www.the-numbers.com/movie/budgets/all/3601

[[38]]
https://www.the-numbers.com/movie/budgets/all/3701

[[39]]
https://www.the-numbers.com/movie/budgets/all/3801

[[40]]
https://www.the-numbers.com/movie/budgets/all/3901

[[41]]
https://www.the-numbers.com/movie/budgets/all/4001

[[42]]
https://www.the-numbers.com/movie/budgets/all/4101

[[43]]
https://www.the-numbers.com/movie/budgets/all/4201

[[44]]
https://www.the-numbers.com/movie/budgets/all/4301

[[45]]
https://www.the-numbers.com/movie/budgets/all/4401
```

```
[[46]]
https://www.the-numbers.com/movie/budgets/all/4501

[[47]]
https://www.the-numbers.com/movie/budgets/all/4601

[[48]]
https://www.the-numbers.com/movie/budgets/all/4701

[[49]]
https://www.the-numbers.com/movie/budgets/all/4801

[[50]]
https://www.the-numbers.com/movie/budgets/all/4901

[[51]]
https://www.the-numbers.com/movie/budgets/all/5001

[[52]]
https://www.the-numbers.com/movie/budgets/all/5101

[[53]]
https://www.the-numbers.com/movie/budgets/all/5201

[[54]]
https://www.the-numbers.com/movie/budgets/all/5301

[[55]]
https://www.the-numbers.com/movie/budgets/all/5401

[[56]]
https://www.the-numbers.com/movie/budgets/all/5501

[[57]]
https://www.the-numbers.com/movie/budgets/all/5601

[[58]]
https://www.the-numbers.com/movie/budgets/all/5701

[[59]]
https://www.the-numbers.com/movie/budgets/all/5801

[[60]]
https://www.the-numbers.com/movie/budgets/all/5901

[[61]]
https://www.the-numbers.com/movie/budgets/all/6001

[[62]]
https://www.the-numbers.com/movie/budgets/all/6101

[[63]]
https://www.the-numbers.com/movie/budgets/all/6201
```

```
[[64]]
https://www.the-numbers.com/movie/budgets/all/6301

[[65]]
https://www.the-numbers.com/movie/budgets/all/6401
map_df(urls, ~read_html(.x) %>%
         html_table() %>%
         .[[1]] %>%
         janitor::clean_names() %>%
         mutate(across(everything(), as.character)))
# A tibble: 6,497 x 6
   x     release_date movie     production_budget domestic_gross worldwide_gross
   <chr> <chr>        <chr>     <chr>             <chr>          <chr>
 1 1     Dec 9, 2022  Avatar: ~ $460,000,000      $684,075,767   $2,317,514,386
 2 2     Apr 23, 2019 Avengers~ $400,000,000      $858,373,000   $2,788,912,285
 3 3     May 20, 2011 Pirates ~ $379,000,000      $241,071,802   $1,045,713,802
 4 4     Apr 22, 2015 Avengers~ $365,000,000      $459,005,868   $1,395,316,979
 5 5     May 17, 2023 Fast X    $340,000,000      $146,126,015   $714,567,285
 6 6     Dec 16, 2015 Star War~ $306,000,000      $936,662,225   $2,064,615,817
 7 7     Apr 25, 2018 Avengers~ $300,000,000      $678,815,482   $2,048,359,754
 8 8     May 24, 2007 Pirates ~ $300,000,000      $309,420,425   $960,996,492
 9 9     Nov 13, 2017 Justice ~ $300,000,000      $229,024,295   $655,945,209
10 10    Jun 28, 2023 Indiana ~ $300,000,000      $174,480,468   $381,408,656
# i 6,487 more rows
```

Group Activity 2

1. Go to scrapethissite and extract the table on the front page.

```
session1 <- read_html("https://www.scrapethissite.com/pages/forms/") %>%
  html_table() %>%
  .[[1]]
```

2. Find out the number of pages that contain the movie table, while looking for the changes in the url in the address bar. How does the url changes when you go to the next page? It goes from 1 to 23, with increments of 1

3. Write a for loop to store all the data in multiple pages to a single data frame. Also, do the same using `purrr::map_df`.

```
new_urls <- "https://www.scrapethissite.com/pages/forms/?page_num="

# Create an empty list
df1 <- list()

# Generate a vector of indices
index <- seq(1, 23, 1)

urls <- map(index, ~str_glue({new_urls}, {.x}))
urls
[[1]]
https://www.scrapethissite.com/pages/forms/?page_num=1

[[2]]
https://www.scrapethissite.com/pages/forms/?page_num=2
```

```
[[3]]
https://www.scrapethissite.com/pages/forms/?page_num=3

[[4]]
https://www.scrapethissite.com/pages/forms/?page_num=4

[[5]]
https://www.scrapethissite.com/pages/forms/?page_num=5

[[6]]
https://www.scrapethissite.com/pages/forms/?page_num=6

[[7]]
https://www.scrapethissite.com/pages/forms/?page_num=7

[[8]]
https://www.scrapethissite.com/pages/forms/?page_num=8

[[9]]
https://www.scrapethissite.com/pages/forms/?page_num=9

[[10]]
https://www.scrapethissite.com/pages/forms/?page_num=10

[[11]]
https://www.scrapethissite.com/pages/forms/?page_num=11

[[12]]
https://www.scrapethissite.com/pages/forms/?page_num=12

[[13]]
https://www.scrapethissite.com/pages/forms/?page_num=13

[[14]]
https://www.scrapethissite.com/pages/forms/?page_num=14

[[15]]
https://www.scrapethissite.com/pages/forms/?page_num=15

[[16]]
https://www.scrapethissite.com/pages/forms/?page_num=16

[[17]]
https://www.scrapethissite.com/pages/forms/?page_num=17

[[18]]
https://www.scrapethissite.com/pages/forms/?page_num=18

[[19]]
https://www.scrapethissite.com/pages/forms/?page_num=19

[[20]]
https://www.scrapethissite.com/pages/forms/?page_num=20
```

```
[[21]]
https://www.scrapethissite.com/pages/forms/?page_num=21

[[22]]
https://www.scrapethissite.com/pages/forms/?page_num=22

[[23]]
https://www.scrapethissite.com/pages/forms/?page_num=23
map_df(urls, ~read_html(.x) %>%
         html_table() %>%
         .[[1]] %>%
         janitor::clean_names() %>%
         mutate(across(everything(), as.character))) -> final_data
```