RWorksheet#5_group.Rmd

Baylon_Calvario_Calzado

2024-11-01

```
library(rvest)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(stringr)
library(polite)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
library(knitr)

link = "https://www.imdb.com/chart/toptv/"
page = read_html(link)
session <- bow(link, user_agent = "Educational")
        session
```

```
## <polite session> https://www.imdb.com/chart/toptv/
##     User-agent: Educational
##     robots.txt: 35 rules are defined for 3 bots
##    Crawl delay: 5 sec
##   The path is scrapable for this user-agent
```

```
nam <- page %>% html_nodes(".ipc-title__text") %>% html_text()
name <- nam[!grepl("Top 250 TV Shows|IMDb Charts|Recently viewed|More to explore", nam, ignore.case = TR
name
```

```
##  [1] "1. Breaking Bad"
##  [2] "2. Planet Earth II"
##  [3] "3. Planet Earth"
##  [4] "4. Band of Brothers"
##  [5] "5. Chernobyl"
```

```
##  [6] "6. The Wire"
##  [7] "7. Avatar: The Last Airbender"
##  [8] "8. Blue Planet II"
##  [9] "9. The Sopranos"
## [10] "10. Cosmos: A Spacetime Odyssey"
## [11] "11. Cosmos"
## [12] "12. Our Planet"
## [13] "13. Game of Thrones"
## [14] "14. Bluey"
## [15] "15. The World at War"
## [16] "16. Fullmetal Alchemist: Brotherhood"
## [17] "17. Rick and Morty"
## [18] "18. Life"
## [19] "19. The Last Dance"
## [20] "20. The Twilight Zone"
## [21] "21. The Vietnam War"
## [22] "22. Sherlock"
## [23] "23. Attack on Titan"
## [24] "24. Batman: The Animated Series"
## [25] "25. The Office"
```

```
rank <- str_extract(name, "^\\d+\\.")
rank
```

```
##  [1] "1."  "2."  "3."  "4."  "5."  "6."  "7."  "8."  "9."  "10." "11." "12."
## [13] "13." "14." "15." "16." "17." "18." "19." "20." "21." "22." "23." "24."
## [25] "25."
```

```
title <- str_replace(name, "^\\d+\\.", "")
title
```

```
##  [1] " Breaking Bad"                    " Planet Earth II"
##  [3] " Planet Earth"                    " Band of Brothers"
##  [5] " Chernobyl"                       " The Wire"
##  [7] " Avatar: The Last Airbender"      " Blue Planet II"
##  [9] " The Sopranos"                    " Cosmos: A Spacetime Odyssey"
## [11] " Cosmos"                          " Our Planet"
## [13] " Game of Thrones"                 " Bluey"
## [15] " The World at War"                " Fullmetal Alchemist: Brotherhood"
## [17] " Rick and Morty"                  " Life"
## [19] " The Last Dance"                  " The Twilight Zone"
## [21] " The Vietnam War"                 " Sherlock"
## [23] " Attack on Titan"                 " Batman: The Animated Series"
## [25] " The Office"
```

```
yea = page %>% html_nodes(".sc-5bc66c50-6.OOdsw.cli-title-metadata-item") %>% html_text()
year <- str_extract_all(yea, "\\b\\d{4}(?:-\\d{4})?\\b") %>% unlist()
year
```

```
##  [1] "2008-2013" "2016"      "2006"      "2001"      "2019"      "2002-2008"
##  [7] "2005-2008" "2017"      "1999-2007" "2014"      "1980"      "2019-2023"
## [13] "2011-2019" "2018"      "1973-1974" "2009-2010" "2013"      "2009"
## [19] "2020"      "1959-1964" "2017"      "2010-2017" "2013-2023" "1992-1995"
## [25] "2005-2013"
```

```
rating = page %>% html_nodes(".ipc-rating-star--rating") %>% html_text()
rating
```

```
##  [1] "9.5" "9.5" "9.4" "9.4" "9.3" "9.3" "9.3" "9.3" "9.2" "9.2" "9.3" "9.2"
## [13] "9.2" "9.3" "9.2" "9.1" "9.1" "9.1" "9.1" "9.0" "9.1" "9.1" "9.1" "9.0"
## [25] "9.0"
```

```
episode <- page %>% html_nodes(".sc-5bc66c50-6.OOdsw.cli-title-metadata-item") %>%
html_text()
episodes <- str_extract_all(episode, "\\b\\d+ eps\\b") %>% unlist()
episodes
```

```
##  [1] "62 eps"  "6 eps"   "11 eps"  "10 eps"  "5 eps"   "60 eps"  "62 eps"
##  [8] "7 eps"   "86 eps"  "13 eps"  "13 eps"  "12 eps"  "74 eps"  "194 eps"
## [15] "26 eps"  "68 eps"  "78 eps"  "11 eps"  "10 eps"  "156 eps" "10 eps"
## [22] "15 eps"  "98 eps"  "85 eps"  "188 eps"
```

```
vote = page %>% html_nodes(".ipc-rating-star--voteCount") %>% html_text()
vote
```

```
##  [1] " (2.2M)" " (162K)" " (223K)" " (544K)" " (904K)" " (390K)" " (388K)"
##  [8] " (48K)"  " (496K)" " (131K)" " (45K)"  " (53K)"  " (2.4M)" " (33K)"
## [15] " (31K)"  " (208K)" " (625K)" " (43K)"  " (159K)" " (96K)"  " (29K)"
## [22] " (1M)"   " (558K)" " (122K)" " (744K)"
```

```
max_length <- max(length(rank), length(title), length(year), length(rating), length(episodes), length(v
rank <- c(rank, rep(NA, max_length - length(rank)))
title <- c(title, rep(NA, max_length - length(title)))
year <- c(year, rep(NA, max_length - length(year)))
rating <- c(rating, rep(NA, max_length - length(rating)))
episodes <- c(episodes, rep(NA, max_length - length(episodes)))
vote <- c(vote, rep(NA, max_length - length(vote)))
max_length
```

```
## [1] 25
```

```
movies = data.frame(rank, title, year, rating, episodes, vote, stringsAsFactors = FALSE)
write.csv(movies, "movies.csv")
print(head(movies))
```

```
##   rank             title      year rating episodes    vote
## 1   1.      Breaking Bad 2008-2013    9.5   62 eps  (2.2M)
## 2   2.   Planet Earth II      2016    9.5    6 eps  (162K)
## 3   3.      Planet Earth      2006    9.4   11 eps  (223K)
## 4   4.  Band of Brothers      2001    9.4   10 eps  (544K)
## 5   5.         Chernobyl      2019    9.3    5 eps  (904K)
## 6   6.          The Wire 2002-2008    9.3   60 eps  (390K)
```

```
movies %>%
  kable("latex", booktabs = TRUE) %>%
  kable_styling(latex_options = "scale_down")
```

rank	title	year	rating	episodes	vote
1.	Breaking Bad	2008–2013	9.5	62 eps	(2.2M)
2.	Planet Earth II	2016	9.5	6 eps	(162K)
3.	Planet Earth	2006	9.4	11 eps	(223K)
4.	Band of Brothers	2001	9.4	10 eps	(544K)
5.	Chernobyl	2019	9.3	5 eps	(904K)
6.	The Wire	2002–2008	9.3	60 eps	(390K)
7.	Avatar: The Last Airbender	2005–2008	9.3	62 eps	(388K)
8.	Blue Planet II	2017	9.3	7 eps	(48K)
9.	The Sopranos	1999–2007	9.2	86 eps	(496K)
10.	Cosmos: A Spacetime Odyssey	2014	9.2	13 eps	(131K)
11.	Cosmos	1980	9.3	13 eps	(45K)
12.	Our Planet	2019–2023	9.2	12 eps	(53K)
13.	Game of Thrones	2011–2019	9.2	74 eps	(2.4M)
14.	Bluey	2018	9.3	194 eps	(33K)
15.	The World at War	1973–1974	9.2	26 eps	(31K)
16.	Fullmetal Alchemist: Brotherhood	2009–2010	9.1	68 eps	(208K)
17.	Rick and Morty	2013	9.1	78 eps	(625K)
18.	Life	2009	9.1	11 eps	(43K)
19.	The Last Dance	2020	9.1	10 eps	(159K)
20.	The Twilight Zone	1959–1964	9.0	156 eps	(96K)
21.	The Vietnam War	2017	9.1	10 eps	(29K)
22.	Sherlock	2010–2017	9.1	15 eps	(1M)
23.	Attack on Titan	2013–2023	9.1	98 eps	(558K)
24.	Batman: The Animated Series	1992–1995	9.0	85 eps	(122K)
25.	The Office	2005–2013	9.0	188 eps	(744K)