

RWorksheet#5

Barrientos, Delfin, Infiesto

2024-11-06

#Extracting TV Shows Reviews

*#1. Each group needs to extract the top 50 tv shows in Imdb.com. It will include the rank, the title of
#It will also include the number of user reviews and the number of critic reviews, as well as the popul*

```
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(rvest)
```

```
library(polite)
```

```
library(tidyr)
```

```
library(httr)
```

```
url <- "https://www.imdb.com/chart/toptv/?ref_=nv_tvv_250"
```

```
session <- bow(url, user_agent = "Educational Purposes Only")
```

```
page <- scrape(session)
```

```
titles <- page %>%
```

```
  html_nodes('.ipc-title__text') %>%
```

```
  html_text()
```

```
ranks <- page %>%
```

```
  html_nodes('a.ipc-title-link-wrapper') %>%
```

```
  html_text() %>%
```

```
  gsub("[^0-9]", "", .)
```

```

links <- page %>%
  html_nodes('a.ipc-title-link-wrapper') %>%
  html_attr('href') %>%
  paste0("https://www.imdb.com", .)

titles <- titles[!titles %in% c("IMDb Charts", "Top 250 TV Shows")]
titles <- titles[1:50]
ranks <- ranks[1:50]
links <- links[1:50]

rank_title <- data.frame(
  rank = ranks,
  title = titles,
  link = links,
  stringsAsFactors = FALSE
)

scrape_show_details <- function(show_url) {
  Sys.sleep(1) # Be polite, avoid overwhelming IMDb's servers
  page <- tryCatch(read_html(show_url), error = function(e) return(NULL))

  if (is.null(page)) return(NULL)

  rating <- page %>%
    html_node('.ipc-rating-star--rating') %>%
    html_text(trim = TRUE)

  votes <- page %>%
    html_node('div.sc-7ab21ed2-3.dPVcnq') %>%
    html_text(trim = TRUE)
  clean_votes <- gsub('[()]', '', votes)

  episodes <- page %>%
    html_node('a[href*="episodes?"]') %>%
    html_text(trim = TRUE)

  year <- page %>%
    html_node('span.sc-8c396aa2-2') %>%
    html_text(trim = TRUE)

  user_reviews <- page %>%
    html_node('span[data-testid="reviews-header"]') %>%
    html_text(trim = TRUE)

  critic_reviews <- page %>%
    html_node('span[class*="score"]') %>%
    html_text(trim = TRUE)

  popularity <- page %>%
    html_node('.sc-39d285cf-1 dxqvqi') %>%

```

```

    html_text(trim = TRUE)

return(data.frame(
  rating = ifelse(is.na(rating), "N/A", rating),
  votes = ifelse(is.na(votes), "N/A", votes),
  episodes = ifelse(is.na(episodes), "N/A", episodes),
  year = ifelse(is.na(year), "N/A", year),
  user_reviews = ifelse(is.na(user_reviews), "N/A", user_reviews),
  critic_reviews = ifelse(is.na(critic_reviews), "N/A", critic_reviews),
  popularity = ifelse(is.na(popularity), "N/A", popularity),
  stringsAsFactors = FALSE
))
}

show_details <- lapply(rank_title$link, function(link_url) {
  scrape_show_details(link_url)
})

show_details <- Filter(Negate(is.null), show_details)
final_data <- do.call(rbind, show_details)

final_result <- cbind(rank_title, final_data)

write.csv(final_result, file = "top_50_tv_shows_imdb.csv", row.names = FALSE)

final_result

```

```

##      rank                title
## 1      1          1. Breaking Bad
## 2      2          2. Planet Earth II
## 3      3          3. Planet Earth
## 4      4          4. Band of Brothers
## 5      5          5. Chernobyl
## 6      6          6. The Wire
## 7      7          7. Avatar: The Last Airbender
## 8      8          8. Blue Planet II
## 9      9          9. The Sopranos
## 10     10         10. Cosmos: A Spacetime Odyssey
## 11     11          11. Cosmos
## 12     12          12. Our Planet
## 13     13          13. Game of Thrones
## 14     14          14. Bluey
## 15     15          15. The World at War
## 16     16 16. Fullmetal Alchemist Brotherhood
## 17     17          17. Rick and Morty
## 18     18          18. Life
## 19     19          19. The Last Dance
## 20     20          20. The Twilight Zone

```

```

## 21 21 21. The Vietnam War
## 22 22 22. Sherlock
## 23 23 23. Attack on Titan
## 24 24 24. Batman: The Animated Series
## 25 25 25. The Office
## 26 <NA> Recently viewed
## 27 <NA> <NA>
## 28 <NA> <NA>
## 29 <NA> <NA>
## 30 <NA> <NA>
## 31 <NA> <NA>
## 32 <NA> <NA>
## 33 <NA> <NA>
## 34 <NA> <NA>
## 35 <NA> <NA>
## 36 <NA> <NA>
## 37 <NA> <NA>
## 38 <NA> <NA>
## 39 <NA> <NA>
## 40 <NA> <NA>
## 41 <NA> <NA>
## 42 <NA> <NA>
## 43 <NA> <NA>
## 44 <NA> <NA>
## 45 <NA> <NA>
## 46 <NA> <NA>
## 47 <NA> <NA>
## 48 <NA> <NA>
## 49 <NA> <NA>
## 50 <NA> <NA>

```

```

## link rating votes
## 1 https://www.imdb.com/title/tt0903747/?ref_=chttvtp_t_1 9.0 N/A
## 2 https://www.imdb.com/title/tt5491994/?ref_=chttvtp_t_2 9.3 N/A
## 3 https://www.imdb.com/title/tt0795176/?ref_=chttvtp_t_3 9.3 N/A
## 4 https://www.imdb.com/title/tt0185906/?ref_=chttvtp_t_4 9.3 N/A
## 5 https://www.imdb.com/title/tt7366338/?ref_=chttvtp_t_5 9.0 N/A
## 6 https://www.imdb.com/title/tt0306414/?ref_=chttvtp_t_6 9.2 N/A
## 7 https://www.imdb.com/title/tt0417299/?ref_=chttvtp_t_7 8.3 N/A
## 8 https://www.imdb.com/title/tt6769208/?ref_=chttvtp_t_8 9.0 N/A
## 9 https://www.imdb.com/title/tt0141842/?ref_=chttvtp_t_9 9.5 N/A
## 10 https://www.imdb.com/title/tt2395695/?ref_=chttvtp_t_10 9.3 N/A
## 11 https://www.imdb.com/title/tt0081846/?ref_=chttvtp_t_11 9.2 N/A
## 12 https://www.imdb.com/title/tt9253866/?ref_=chttvtp_t_12 9.3 N/A
## 13 https://www.imdb.com/title/tt0944947/?ref_=chttvtp_t_13 9.5 N/A
## 14 https://www.imdb.com/title/tt7678620/?ref_=chttvtp_t_14 8.4 N/A
## 15 https://www.imdb.com/title/tt0071075/?ref_=chttvtp_t_15 9.1 N/A
## 16 https://www.imdb.com/title/tt1355642/?ref_=chttvtp_t_16 9.0 N/A
## 17 https://www.imdb.com/title/tt2861424/?ref_=chttvtp_t_17 9.5 N/A
## 18 https://www.imdb.com/title/tt1533395/?ref_=chttvtp_t_18 9.0 N/A
## 19 https://www.imdb.com/title/tt8420184/?ref_=chttvtp_t_19 9.3 N/A
## 20 https://www.imdb.com/title/tt0052520/?ref_=chttvtp_t_20 7.7 N/A
## 21 https://www.imdb.com/title/tt1877514/?ref_=chttvtp_t_21 9.0 N/A
## 22 https://www.imdb.com/title/tt1475582/?ref_=chttvtp_t_22 9.3 N/A
## 23 https://www.imdb.com/title/tt2560140/?ref_=chttvtp_t_23 8.9 N/A

```

## 24	https://www.imdb.com/title/tt0103359/?ref_=chttvtp_t_24	8.4	N/A
## 25	https://www.imdb.com/title/tt0386676/?ref_=chttvtp_t_25	8.9	N/A
## 26	<NA>	9.0	N/A
## 27	<NA>	9.3	N/A
## 28	<NA>	9.3	N/A
## 29	<NA>	9.3	N/A
## 30	<NA>	9.0	N/A
## 31	<NA>	9.2	N/A
## 32	<NA>	8.3	N/A
## 33	<NA>	9.0	N/A
## 34	<NA>	9.5	N/A
## 35	<NA>	9.3	N/A
## 36	<NA>	9.2	N/A
## 37	<NA>	9.3	N/A
## 38	<NA>	9.5	N/A
## 39	<NA>	8.4	N/A
## 40	<NA>	9.1	N/A
## 41	<NA>	9.0	N/A
## 42	<NA>	9.5	N/A
## 43	<NA>	9.0	N/A
## 44	<NA>	9.3	N/A
## 45	<NA>	7.7	N/A
## 46	<NA>	9.0	N/A
## 47	<NA>	9.3	N/A
## 48	<NA>	8.9	N/A
## 49	<NA>	8.4	N/A
## 50	<NA>	8.9	N/A
##	episodes year user_reviews critic_reviews popularity		
## 1	Episodes62 N/A	N/A	5K N/A
## 2	Episodes6 N/A	N/A	158 N/A
## 3	Episodes11 N/A	N/A	111 N/A
## 4	Episodes10 N/A	N/A	1K N/A
## 5	Episodes5 N/A	N/A	3.5K N/A
## 6	Episodes60 N/A	N/A	786 N/A
## 7	Episodes62 N/A	N/A	997 N/A
## 8	Episodes7 N/A	N/A	53 N/A
## 9	Episodes86 N/A	N/A	960 N/A
## 10	Episodes13 N/A	N/A	205 N/A
## 11	Episodes13 N/A	N/A	80 N/A
## 12	Episodes12 N/A	N/A	245 N/A
## 13	Episodes74 N/A	N/A	5.8K N/A
## 14	Episodes194 N/A	N/A	366 N/A
## 15	Episodes26 N/A	N/A	126 N/A
## 16	Episodes68 N/A	N/A	464 N/A
## 17	Episodes78 N/A	N/A	908 N/A
## 18	Episodes11 N/A	N/A	12 N/A
## 19	Episodes10 N/A	N/A	541 N/A
## 20	Episodes156 N/A	N/A	213 N/A
## 21	Episodes10 N/A	N/A	175 N/A
## 22	Episodes15 N/A	N/A	1K N/A
## 23	Episodes98 N/A	N/A	2.3K N/A
## 24	Episodes85 N/A	N/A	219 N/A
## 25	Episodes188 N/A	N/A	1.7K N/A
## 26	Episodes62 N/A	N/A	5K N/A

## 27	Episodes6	N/A	N/A	158	N/A
## 28	Episodes11	N/A	N/A	111	N/A
## 29	Episodes10	N/A	N/A	1K	N/A
## 30	Episodes5	N/A	N/A	3.5K	N/A
## 31	Episodes60	N/A	N/A	786	N/A
## 32	Episodes62	N/A	N/A	997	N/A
## 33	Episodes7	N/A	N/A	53	N/A
## 34	Episodes86	N/A	N/A	960	N/A
## 35	Episodes13	N/A	N/A	205	N/A
## 36	Episodes13	N/A	N/A	80	N/A
## 37	Episodes12	N/A	N/A	245	N/A
## 38	Episodes74	N/A	N/A	5.8K	N/A
## 39	Episodes194	N/A	N/A	366	N/A
## 40	Episodes26	N/A	N/A	126	N/A
## 41	Episodes68	N/A	N/A	464	N/A
## 42	Episodes78	N/A	N/A	908	N/A
## 43	Episodes11	N/A	N/A	12	N/A
## 44	Episodes10	N/A	N/A	541	N/A
## 45	Episodes156	N/A	N/A	213	N/A
## 46	Episodes10	N/A	N/A	175	N/A
## 47	Episodes15	N/A	N/A	1K	N/A
## 48	Episodes98	N/A	N/A	2.3K	N/A
## 49	Episodes85	N/A	N/A	219	N/A
## 50	Episodes188	N/A	N/A	1.7K	N/A

[View\(final_result\)](#)