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 the tv show, tv rating, the number of people who voted, the number of episodes, the year it was released. #It will also include the number of user reviews and the number of critic reviews, as well as the popularity rating for each tv shows.

```
library(polite)
library(httr)
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
```

```
url <- "https://www.imdb.com/chart/toptv/?sort=rank%2Casc"
session <- bow(url, user_agent = "Educational")
session
```

```
## <polite session> https://www.imdb.com/chart/toptv/?sort=rank%2Casc
##     User-agent: Educational
##     robots.txt: 35 rules are defined for 3 bots
##    Crawl delay: 5 sec
##   The path is scrapable for this user-agent
```

```
title_list <- read_html(url) %>%
  html_nodes('.ipc-title__text') %>%
  html_text()

title_list_sub <- as.data.frame(title_list[3:27], stringsAsFactors = FALSE)
colnames(title_list_sub) <- "ranks"
```

```r
split_df <- strsplit(as.character(title_list_sub$ranks), "\\.", fixed = FALSE)
split_df <- data.frame(do.call(rbind, split_df), stringsAsFactors = FALSE)

colnames(split_df) <- c("rank", "title")
split_df <- split_df %>% select(rank, title)
split_df$title <- trimws(split_df$title)

rank_title <- split_df

rating_ls <- read_html(url) %>%
  html_nodes('.ipc-rating-star--rating') %>%
  html_text()

voter_ls <- read_html(url) %>%
  html_nodes('.ipc-rating-star--voteCount') %>%
  html_text()
clean_votes <- gsub('[()]', '', voter_ls)

eps_ls <- read_html(url) %>%
  html_nodes('span.sc-5bc66c50-6.OOdsw.cli-title-metadata-item:nth-of-type(2)') %>%
  html_text()
clean_eps <- gsub('[eps]', '', eps_ls)
num_eps <- as.numeric(clean_eps)

years <- read_html(url) %>%
  html_nodes('span.sc-5bc66c50-6.OOdsw.cli-title-metadata-item:nth-of-type(1)') %>%
  html_text()

top_tv_shows <- data.frame(
  Rank = rank_title[,1],
  Title = rank_title[,2],
  Rating = rating_ls,
  Voters = clean_votes,
  Episodes = num_eps,
  Year = years
)

home_link <- 'https://www.imdb.com/chart/toptv/'
main_page <- read_html(home_link)

links <- main_page %>%
  html_nodes("a.ipc-title-link-wrapper") %>%
  html_attr("href")

show_data <- lapply(links, function(link) {
  complete_link <- paste0("https://imdb.com", link)

  usrv_link <- read_html(complete_link)
  usrv_link_page <- usrv_link %>%
    html_nodes('a.isReview') %>%
    html_attr("href")

  critic <- usrv_link %>%
```

```
    html_nodes("span.score") %>%
    html_text()
  critic_df <- data.frame(Critic_Reviews = critic[2], stringsAsFactors = FALSE)

  pop_rating <- usrv_link %>%
    html_nodes('[data-testid="hero-rating-bar__popularity__score"]') %>%
    html_text()

  usrv <- read_html(paste0("https://imdb.com", usrv_link_page[1]))
  usrv_count <- usrv %>%
    html_nodes('[data-testid="tturv-total-reviews"]') %>%
    html_text()

  return(data.frame(Show_Link = complete_link, User_Reviews = usrv_count, Critic = critic_df, Popularity
})

show_url_df <- do.call(rbind, show_data)
print(show_url_df)
```

```
##                                          Show_Link  User_Reviews
## 1   https://imdb.com/title/tt0903747/?ref_=chttvtp_t_1 5,084 reviews
## 2   https://imdb.com/title/tt0903747/?ref_=chttvtp_t_1 5,084 reviews
## 3   https://imdb.com/title/tt5491994/?ref_=chttvtp_t_2   158 reviews
## 4   https://imdb.com/title/tt5491994/?ref_=chttvtp_t_2   158 reviews
## 5   https://imdb.com/title/tt0795176/?ref_=chttvtp_t_3   111 reviews
## 6   https://imdb.com/title/tt0795176/?ref_=chttvtp_t_3   111 reviews
## 7   https://imdb.com/title/tt0185906/?ref_=chttvtp_t_4 1,055 reviews
## 8   https://imdb.com/title/tt0185906/?ref_=chttvtp_t_4 1,055 reviews
## 9   https://imdb.com/title/tt7366338/?ref_=chttvtp_t_5 3,531 reviews
## 10  https://imdb.com/title/tt7366338/?ref_=chttvtp_t_5 3,531 reviews
## 11  https://imdb.com/title/tt0306414/?ref_=chttvtp_t_6   786 reviews
## 12  https://imdb.com/title/tt0306414/?ref_=chttvtp_t_6   786 reviews
## 13  https://imdb.com/title/tt0417299/?ref_=chttvtp_t_7   997 reviews
## 14  https://imdb.com/title/tt0417299/?ref_=chttvtp_t_7   997 reviews
## 15  https://imdb.com/title/tt6769208/?ref_=chttvtp_t_8    53 reviews
## 16  https://imdb.com/title/tt6769208/?ref_=chttvtp_t_8    53 reviews
## 17  https://imdb.com/title/tt0141842/?ref_=chttvtp_t_9   961 reviews
## 18  https://imdb.com/title/tt0141842/?ref_=chttvtp_t_9   961 reviews
## 19 https://imdb.com/title/tt2395695/?ref_=chttvtp_t_10   205 reviews
## 20 https://imdb.com/title/tt2395695/?ref_=chttvtp_t_10   205 reviews
## 21 https://imdb.com/title/tt0081846/?ref_=chttvtp_t_11    80 reviews
## 22 https://imdb.com/title/tt0081846/?ref_=chttvtp_t_11    80 reviews
## 23 https://imdb.com/title/tt9253866/?ref_=chttvtp_t_12   245 reviews
## 24 https://imdb.com/title/tt9253866/?ref_=chttvtp_t_12   245 reviews
## 25 https://imdb.com/title/tt0944947/?ref_=chttvtp_t_13 5,894 reviews
## 26 https://imdb.com/title/tt0944947/?ref_=chttvtp_t_13 5,894 reviews
## 27 https://imdb.com/title/tt7678620/?ref_=chttvtp_t_14   366 reviews
## 28 https://imdb.com/title/tt7678620/?ref_=chttvtp_t_14   366 reviews
## 29 https://imdb.com/title/tt0071075/?ref_=chttvtp_t_15   126 reviews
## 30 https://imdb.com/title/tt0071075/?ref_=chttvtp_t_15   126 reviews
## 31 https://imdb.com/title/tt1355642/?ref_=chttvtp_t_16   465 reviews
## 32 https://imdb.com/title/tt1355642/?ref_=chttvtp_t_16   465 reviews
## 33 https://imdb.com/title/tt2861424/?ref_=chttvtp_t_17   908 reviews
```

```
## 34 https://imdb.com/title/tt2861424/?ref_=chttvtp_t_17   908 reviews
## 35 https://imdb.com/title/tt1533395/?ref_=chttvtp_t_18    12 reviews
## 36 https://imdb.com/title/tt1533395/?ref_=chttvtp_t_18    12 reviews
## 37 https://imdb.com/title/tt8420184/?ref_=chttvtp_t_19   541 reviews
## 38 https://imdb.com/title/tt8420184/?ref_=chttvtp_t_19   541 reviews
## 39 https://imdb.com/title/tt0052520/?ref_=chttvtp_t_20   213 reviews
## 40 https://imdb.com/title/tt0052520/?ref_=chttvtp_t_20   213 reviews
## 41 https://imdb.com/title/tt1877514/?ref_=chttvtp_t_21   175 reviews
## 42 https://imdb.com/title/tt1877514/?ref_=chttvtp_t_21   175 reviews
## 43 https://imdb.com/title/tt1475582/?ref_=chttvtp_t_22 1,094 reviews
## 44 https://imdb.com/title/tt1475582/?ref_=chttvtp_t_22 1,094 reviews
## 45 https://imdb.com/title/tt2560140/?ref_=chttvtp_t_23 2,353 reviews
## 46 https://imdb.com/title/tt2560140/?ref_=chttvtp_t_23 2,353 reviews
## 47 https://imdb.com/title/tt0103359/?ref_=chttvtp_t_24   219 reviews
## 48 https://imdb.com/title/tt0103359/?ref_=chttvtp_t_24   219 reviews
## 49 https://imdb.com/title/tt0386676/?ref_=chttvtp_t_25 1,773 reviews
## 50 https://imdb.com/title/tt0386676/?ref_=chttvtp_t_25 1,773 reviews
##    Critic_Reviews Popularity_Rating
## 1             175                20
## 2             175                20
## 3               6             1,121
## 4               6             1,121
## 5              10             2,011
## 6              10             2,011
## 7              34               171
## 8              34               171
## 9              88               173
## 10             88               173
## 11             77               108
## 12             77               108
## 13             57               373
## 14             57               373
## 15              9             4,415
## 16              9             4,415
## 17             93                33
## 18             93                33
## 19             12             1,499
## 20             12             1,499
## 21              8             3,866
## 22              8             3,866
## 23             15             2,765
## 24             15             2,765
## 25            368                14
## 26            368                14
## 27              4               411
## 28              4               411
## 29              5             2,627
## 30              5             2,627
## 31             16               508
## 32             16               508
## 33             94               137
## 34             94               137
## 35              9             3,455
## 36              9             3,455
```

```
## 37             28            1,521
## 38             28            1,521
## 39             85              354
## 40             85              354
## 41             13            2,022
## 42             13            2,022
## 43            121              172
## 44            121              172
## 45             64               60
## 46             64               60
## 47             25              527
## 48             25              527
## 49             76               55
## 50             76               55
```

```
shows <- cbind(top_tv_shows, show_url_df)
shows
```

```
##    Rank                          Title Rating Voters Episodes      Year
## 1     1                   Breaking Bad    9.5   2.2M       62 2008-2013
## 2     2                 Planet Earth II    9.5   162K        6      2016
## 3     3                    Planet Earth    9.4   223K       11      2006
## 4     4                Band of Brothers    9.4   544K       10      2001
## 5     5                      Chernobyl    9.3   905K        5      2019
## 6     6                       The Wire    9.3   390K       60 2002-2008
## 7     7     Avatar: The Last Airbender    9.3   388K       62 2005-2008
## 8     8                  Blue Planet II    9.3    48K        7      2017
## 9     9                    The Sopranos    9.2   497K       86 1999-2007
## 10   10    Cosmos: A Spacetime Odyssey    9.2   131K       13      2014
## 11   11                         Cosmos    9.3    45K       13      1980
## 12   12                     Our Planet    9.2    53K       12 2019-2023
## 13   13                Game of Thrones    9.2   2.4M       74 2011-2019
## 14   14                          Bluey    9.3    33K      194    2018-
## 15   15               The World at War    9.2    31K       26 1973-1974
## 16   16 Fullmetal Alchemist Brotherhood    9.1   208K       68 2009-2010
## 17   17                  Rick and Morty    9.1   626K       78    2013-
## 18   18                            Life    9.1    43K       11      2009
## 19   19                The Last Dance    9.1   159K       10      2020
## 20   20              The Twilight Zone    9.0    96K      156 1959-1964
## 21   21                The Vietnam War    9.1    29K       10      2017
## 22   22                       Sherlock    9.1     1M       15 2010-2017
## 23   23                Attack on Titan    9.1   559K       98 2013-2023
## 24   24    Batman: The Animated Series    9.0   122K       85 1992-1995
## 25   25                     The Office    9.0   744K      188 2005-2013
## 26    1                   Breaking Bad    9.5   2.2M       62 2008-2013
## 27    2                 Planet Earth II    9.5   162K        6      2016
## 28    3                    Planet Earth    9.4   223K       11      2006
## 29    4                Band of Brothers    9.4   544K       10      2001
## 30    5                      Chernobyl    9.3   905K        5      2019
## 31    6                       The Wire    9.3   390K       60 2002-2008
## 32    7     Avatar: The Last Airbender    9.3   388K       62 2005-2008
## 33    8                  Blue Planet II    9.3    48K        7      2017
## 34    9                    The Sopranos    9.2   497K       86 1999-2007
## 35   10    Cosmos: A Spacetime Odyssey    9.2   131K       13      2014
```

```
## 36   11                          Cosmos    9.3    45K       13      1980
## 37   12                       Our Planet    9.2    53K       12 2019-2023
## 38   13                 Game of Thrones    9.2   2.4M       74 2011-2019
## 39   14                           Bluey    9.3    33K      194    2018-
## 40   15                The World at War    9.2    31K       26 1973-1974
## 41   16 Fullmetal Alchemist Brotherhood    9.1   208K       68 2009-2010
## 42   17                  Rick and Morty    9.1   626K       78    2013-
## 43   18                            Life    9.1    43K       11      2009
## 44   19                 The Last Dance    9.1   159K       10      2020
## 45   20               The Twilight Zone    9.0    96K      156 1959-1964
## 46   21                The Vietnam War    9.1    29K       10      2017
## 47   22                       Sherlock    9.1     1M       15 2010-2017
## 48   23                Attack on Titan    9.1   559K       98 2013-2023
## 49   24     Batman: The Animated Series    9.0   122K       85 1992-1995
## 50   25                     The Office    9.0   744K      188 2005-2013
##                                            Show_Link  User_Reviews
## 1   https://imdb.com/title/tt0903747/?ref_=chttvtp_t_1 5,084 reviews
## 2   https://imdb.com/title/tt0903747/?ref_=chttvtp_t_1 5,084 reviews
## 3   https://imdb.com/title/tt5491994/?ref_=chttvtp_t_2   158 reviews
## 4   https://imdb.com/title/tt5491994/?ref_=chttvtp_t_2   158 reviews
## 5   https://imdb.com/title/tt0795176/?ref_=chttvtp_t_3   111 reviews
## 6   https://imdb.com/title/tt0795176/?ref_=chttvtp_t_3   111 reviews
## 7   https://imdb.com/title/tt0185906/?ref_=chttvtp_t_4 1,055 reviews
## 8   https://imdb.com/title/tt0185906/?ref_=chttvtp_t_4 1,055 reviews
## 9   https://imdb.com/title/tt7366338/?ref_=chttvtp_t_5 3,531 reviews
## 10  https://imdb.com/title/tt7366338/?ref_=chttvtp_t_5 3,531 reviews
## 11  https://imdb.com/title/tt0306414/?ref_=chttvtp_t_6   786 reviews
## 12  https://imdb.com/title/tt0306414/?ref_=chttvtp_t_6   786 reviews
## 13  https://imdb.com/title/tt0417299/?ref_=chttvtp_t_7   997 reviews
## 14  https://imdb.com/title/tt0417299/?ref_=chttvtp_t_7   997 reviews
## 15  https://imdb.com/title/tt6769208/?ref_=chttvtp_t_8    53 reviews
## 16  https://imdb.com/title/tt6769208/?ref_=chttvtp_t_8    53 reviews
## 17  https://imdb.com/title/tt0141842/?ref_=chttvtp_t_9   961 reviews
## 18  https://imdb.com/title/tt0141842/?ref_=chttvtp_t_9   961 reviews
## 19 https://imdb.com/title/tt2395695/?ref_=chttvtp_t_10   205 reviews
## 20 https://imdb.com/title/tt2395695/?ref_=chttvtp_t_10   205 reviews
## 21 https://imdb.com/title/tt0081846/?ref_=chttvtp_t_11    80 reviews
## 22 https://imdb.com/title/tt0081846/?ref_=chttvtp_t_11    80 reviews
## 23 https://imdb.com/title/tt9253866/?ref_=chttvtp_t_12   245 reviews
## 24 https://imdb.com/title/tt9253866/?ref_=chttvtp_t_12   245 reviews
## 25 https://imdb.com/title/tt0944947/?ref_=chttvtp_t_13 5,894 reviews
## 26 https://imdb.com/title/tt0944947/?ref_=chttvtp_t_13 5,894 reviews
## 27 https://imdb.com/title/tt7678620/?ref_=chttvtp_t_14   366 reviews
## 28 https://imdb.com/title/tt7678620/?ref_=chttvtp_t_14   366 reviews
## 29 https://imdb.com/title/tt0071075/?ref_=chttvtp_t_15   126 reviews
## 30 https://imdb.com/title/tt0071075/?ref_=chttvtp_t_15   126 reviews
## 31 https://imdb.com/title/tt1355642/?ref_=chttvtp_t_16   465 reviews
## 32 https://imdb.com/title/tt1355642/?ref_=chttvtp_t_16   465 reviews
## 33 https://imdb.com/title/tt2861424/?ref_=chttvtp_t_17   908 reviews
## 34 https://imdb.com/title/tt2861424/?ref_=chttvtp_t_17   908 reviews
## 35 https://imdb.com/title/tt1533395/?ref_=chttvtp_t_18    12 reviews
## 36 https://imdb.com/title/tt1533395/?ref_=chttvtp_t_18    12 reviews
## 37 https://imdb.com/title/tt8420184/?ref_=chttvtp_t_19   541 reviews
## 38 https://imdb.com/title/tt8420184/?ref_=chttvtp_t_19   541 reviews
```

```
## 39 https://imdb.com/title/tt0052520/?ref_=chttvtp_t_20   213 reviews
## 40 https://imdb.com/title/tt0052520/?ref_=chttvtp_t_20   213 reviews
## 41 https://imdb.com/title/tt1877514/?ref_=chttvtp_t_21   175 reviews
## 42 https://imdb.com/title/tt1877514/?ref_=chttvtp_t_21   175 reviews
## 43 https://imdb.com/title/tt1475582/?ref_=chttvtp_t_22 1,094 reviews
## 44 https://imdb.com/title/tt1475582/?ref_=chttvtp_t_22 1,094 reviews
## 45 https://imdb.com/title/tt2560140/?ref_=chttvtp_t_23 2,353 reviews
## 46 https://imdb.com/title/tt2560140/?ref_=chttvtp_t_23 2,353 reviews
## 47 https://imdb.com/title/tt0103359/?ref_=chttvtp_t_24   219 reviews
## 48 https://imdb.com/title/tt0103359/?ref_=chttvtp_t_24   219 reviews
## 49 https://imdb.com/title/tt0386676/?ref_=chttvtp_t_25 1,773 reviews
## 50 https://imdb.com/title/tt0386676/?ref_=chttvtp_t_25 1,773 reviews
##    Critic_Reviews Popularity_Rating
## 1             175                20
## 2             175                20
## 3               6             1,121
## 4               6             1,121
## 5              10             2,011
## 6              10             2,011
## 7              34               171
## 8              34               171
## 9              88               173
## 10             88               173
## 11             77               108
## 12             77               108
## 13             57               373
## 14             57               373
## 15              9             4,415
## 16              9             4,415
## 17             93                33
## 18             93                33
## 19             12             1,499
## 20             12             1,499
## 21              8             3,866
## 22              8             3,866
## 23             15             2,765
## 24             15             2,765
## 25            368                14
## 26            368                14
## 27              4               411
## 28              4               411
## 29              5             2,627
## 30              5             2,627
## 31             16               508
## 32             16               508
## 33             94               137
## 34             94               137
## 35              9             3,455
## 36              9             3,455
## 37             28             1,521
## 38             28             1,521
## 39             85               354
## 40             85               354
## 41             13             2,022
```

```
## 42              13             2,022
## 43             121               172
## 44             121               172
## 45              64                60
## 46              64                60
## 47              25               527
## 48              25               527
## 49              76                55
## 50              76                55
```