RWorksheet#5_group(1&9)

GROUP 1&9 BSIT-2B

2023-11-30

##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.

```
library(rvest)
```

```
## Warning: package 'rvest' was built under R version 4.3.2
```

```
library(httr)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.3.2
```

```
##
## 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(polite)
```

```
## Warning: package 'polite' was built under R version 4.3.2
```

```
polite::use_manners(save_as = 'polite_scrape.R')
```

```
## v Setting active project to 'C:/Users/missy/OneDrive/Documents/Worksheet#5'
```

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

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

```
## <polite session> https://m.imdb.com/chart/toptv/?ref_=nv_tvv_250
##     User-agent: Educational
##     robots.txt: 23 rules are defined for 2 bots
##    Crawl delay: 5 sec
##   The path is scrapable for this user-agent
```

```
title_show <- character(0)
list_year_ep <-character(0)

title_show <- scrape(session) %>%
  html_nodes('h3.ipc-title__text') %>%
  html_text

title_show_only <- as.data.frame(title_show[2:51])
title_show_only
```

```
##                           title_show[2:51]
## 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. The World at War
## 15                        15. Rick and Morty
## 16                                16. Bluey
## 17     17. Fullmetal Alchemist Brotherhood
## 18                          18. The Last Dance
## 19                                19. Life
## 20                     20. The Twilight Zone
## 21                               21. Sherlock
## 22                         22. The Vietnam War
## 23          23. Batman: The Animated Series
## 24                         24. Attack on Titan
## 25  25. Scam 1992: The Harshad Mehta Story
## 26                              26. The Office
## 27                                  27. Arcane
## 28                         28. The Blue Planet
## 29                        29. Better Call Saul
## 30                            30. Human Planet
## 31                                 31. Firefly
## 32                           32. Frozen Planet
## 33                         33. Clarkson's Farm
## 34                              34. Death Note
## 35            35. Only Fools and Horses....
## 36                         36. Hunter x Hunter
## 37                           37. The Civil War
```

```
## 38                      38. True Detective
## 39                           39. Seinfeld
## 40               40. The Beatles: Get Back
## 41                            41. Dekalog
## 42                           42. Sahsiyet
## 43                              43. Fargo
## 44                        44. Cowboy Bebop
## 45                       45. Gravity Falls
## 46                      46. Nathan for You
## 47  47. Last Week Tonight with John Oliver
## 48                     48. When They See Us
## 49                          49. Succession
## 50 50. Apocalypse: La 2ème guerre mondiale
```

```
colnames(title_show_only) <- "Rank"

show_df <- strsplit(as.character(title_show_only$Rank),".",fixed = TRUE)
show_df <- data.frame(do.call(rbind,show_df))
```

```
## Warning in (function (..., deparse.level = 1) : number of columns of result is
## not a multiple of vector length (arg 1)
```

```
show_df <- show_df[-c(3:5)]

#renaming column 1 and 2
colnames(show_df) <- c("Rank","Title")


list_year_ep <- scrape(session) %>%
  html_nodes('span.sc-479faa3c-8.bNrEFi.cli-title-metadata-item
') %>%
  html_text

years_only <- c()
for (i in seq(1, length(list_year_ep), by = 3)) {
  years_only <- c(years_only, list_year_ep[i])
}
Year <- years_only[1:50]


ep_only <- c()
for (i in seq(2, length(list_year_ep), by = 3)) {
  ep_only <- c(ep_only, list_year_ep[i])
}
Episode <- ep_only[1:50]


df_title_ep <- data.frame(Episode,Year)
colnames(df_title_ep) <- c("Number Of Episodes","Year Released")


list_rating <- scrape(session) %>%
  html_nodes('span.ipc-rating-star.ipc-rating-star--base.ipc-rating-star--imdb.ratingGroup--imdb-rating
```

```
  html_text()

wholeRATING <- as.data.frame(list_rating[1:50])
colnames(wholeRATING) <- "Rating"

wholeRATING$Rating <- gsub("\\s*\\([^)]+\\)\\s*", "", wholeRATING$Rating)
wholeRATING$Vote_Count <- gsub(".*\\(([^)]+)\\)", "\\1", list_rating[1:50])


df_rating_vote <- wholeRATING
colnames(df_rating_vote) <- c("Rating","Vote Count")

final_df <- cbind(show_df, df_rating_vote,df_title_ep )
final_df
```

```
##    Rank                                Title Rating Vote Count
## 1     1                        Breaking Bad   9.5        2.1M
## 2     2                     Planet Earth II   9.5        155K
## 3     3                        Planet Earth   9.4        217K
## 4     4                    Band of Brothers   9.4        506K
## 5     5                           Chernobyl   9.3        833K
## 6     6                           The Wire    9.3        366K
## 7     7          Avatar: The Last Airbender   9.3        348K
## 8     8                      Blue Planet II   9.3         45K
## 9     9                        The Sopranos   9.2        448K
## 10   10         Cosmos: A Spacetime Odyssey   9.3        127K
## 11   11                              Cosmos   9.3         43K
## 12   12                          Our Planet   9.3         49K
## 13   13                     Game of Thrones   9.2        2.2M
## 14   14                    The World at War   9.2         28K
## 15   15                      Rick and Morty   9.1        579K
## 16   16                               Bluey   9.4         22K
## 17   17     Fullmetal Alchemist Brotherhood   9.1        190K
## 18   18                      The Last Dance   9.1        145K
## 19   19                                Life   9.1         42K
## 20   20                  The Twilight Zone    9.1         90K
## 21   21                           Sherlock   9.1        975K
## 22   22                     The Vietnam War   9.1         27K
## 23   23         Batman: The Animated Series   9.0        113K
## 24   24                     Attack on Titan   9.1        473K
## 25   25  Scam 1992: The Harshad Mehta Story   9.3        153K
## 26   26                          The Office   9.0        679K
## 27   27                              Arcane   9.0        251K
## 28   28                     The Blue Planet   9.0         42K
## 29   29                    Better Call Saul   9.0        618K
## 30   30                        Human Planet   9.0         27K
## 31   31                             Firefly   9.0        277K
## 32   32                       Frozen Planet   9.0         32K
## 33   33                     Clarkson's Farm   9.0         52K
## 34   34                          Death Note   8.9        363K
## 35   35               Only Fools and Horses   9.0         55K
## 36   36                     Hunter x Hunter   9.0        125K
## 37   37                       The Civil War   9.0         18K
```

```
## 38   38                       True Detective   8.9        616K
## 39   39                             Seinfeld   8.9        342K
## 40   40                The Beatles: Get Back   9.0         27K
## 41   41                              Dekalog   9.0         27K
## 42   42                              Sahsiyet   9.0         45K
## 43   43                                 Fargo   8.9        395K
## 44   44                          Cowboy Bebop   8.9        133K
## 45   45                          Gravity Falls   8.9        128K
## 46   46                         Nathan for You   8.9         36K
## 47   47   Last Week Tonight with John Oliver   8.9         94K
## 48   48                       When They See Us   8.9        134K
## 49   49                              Succession   8.9        246K
## 50   50  Apocalypse: La 2ème guerre mondiale   9.0         14K
##    Number Of Episodes Year Released
## 1              62 eps     2008-2013
## 2               6 eps          2016
## 3              11 eps          2006
## 4              10 eps          2001
## 5               5 eps          2019
## 6              60 eps     2002-2008
## 7              62 eps     2005-2008
## 8               7 eps          2017
## 9              86 eps     1999-2007
## 10             13 eps          2014
## 11             13 eps          1980
## 12             12 eps     2019-2023
## 13             73 eps     2011-2019
## 14             26 eps     1973-1974
## 15             74 eps        2013-
## 16            171 eps        2018-
## 17             68 eps     2009-2010
## 18             10 eps          2020
## 19             11 eps          2009
## 20            156 eps     1959-1964
## 21             15 eps     2010-2017
## 22             10 eps          2017
## 23             85 eps     1992-1995
## 24             98 eps     2013-2023
## 25             10 eps          2020
## 26            188 eps     2005-2013
## 27             10 eps        2021-
## 28              8 eps          2001
## 29             63 eps     2015-2022
## 30              8 eps          2011
## 31             14 eps     2002-2003
## 32             10 eps     2011-2012
## 33             17 eps        2021-
## 34             37 eps     2006-2007
## 35             64 eps     1981-2003
## 36            148 eps     2011-2014
## 37              9 eps          1990
## 38             30 eps        2014-
## 39            173 eps     1989-1998
## 40              3 eps          2021
```

```
## 41             10 eps     1989-1990
## 42             16 eps        2018-
## 43             51 eps     2014-2024
## 44             26 eps     1998-1999
## 45             41 eps     2012-2016
## 46             32 eps     2013-2017
## 47            338 eps        2014-
## 48              4 eps          2019
## 49             39 eps     2018-2023
## 50              6 eps          2009
```

From the 50 tv shows, select at least 5 tv shows to scrape the user reviews that will include the reviewer's name, date of reviewed, user rating, title of the review, and text reviews.

```
tv_show_urls <- c(
  "https://www.imdb.com/title/tt0081846/reviews",  #COSMOS
  "https://www.imdb.com/title/tt0903747/reviews", #BREAKING BAD
  "https://www.imdb.com/title/tt0185906/reviews", # BAND OF BROTHERS
  "https://www.imdb.com/title/tt7366338/reviews", #CHERNOBYL
  "https://www.imdb.com/title/tt0417299/reviews" #Avatar: The Last Airbender
)

all_reviews <- list()

for (url in tv_show_urls) {
  # Read HTML content
  page <- read_html(url)

  reviewers_name <- page %>% html_nodes(".display-name-link") %>% html_text()
  dates <- page %>% html_nodes("span.review-date") %>% html_text()
  user_ratings <- page %>% html_nodes("span.rating-other-user-rating") %>% html_text()
  text_reviews <- page %>% html_nodes("div.text") %>% html_text()

  reviews_df <- data.frame(
    Reviewer_Name = reviewers_name[1:5],
    Date = dates[1:5],
    User_Rating = user_ratings[1:5],
    Text_Review = text_reviews[1:5],
    stringsAsFactors = FALSE
  )

   all_reviews[[url]] <- reviews_df
}

final_reviews_df <- do.call(rbind, all_reviews)
colnames(final_reviews_df) <- c("Name", "Date of Review", "User Rating", "Text Reviews")

rownames(final_reviews_df) <- NULL

final_reviews_df
```

```
##                   Name    Date of Review                      User Rating
## 1           khatcher-2  22 February 2004  \n           9/10\n
```

```
## 2              Cari-8      22 July 1999 \n            10/10\n
## 3         Steve_Nyland    26 August 2007 \n            10/10\n
## 4             phynigan 19 September 2005 \n            10/10\n
## 5            Cheese-18     27 March 2001 \n            10/10\n
## 6              FiRE010       4 July 2021 \n            10/10\n
## 7       Supermanfan-13   9 November 2021 \n            10/10\n
## 8    TheLittleSongbird  13 November 2017 \n            10/10\n
## 9         KinoKoopaKid      30 July 2021 \n            10/10\n
## 10         jehuschultz  19 February 2020 \n            10/10\n
## 11           rbverhoef  14 February 2003 \n            10/10\n
## 12  philip_vanderveken 17 September 2004 \n            10/10\n
## 13          bsmith5552   6 November 2001 \n            10/10\n
## 14       planktonrules       31 May 2015 \n            10/10\n
## 15       DiCaprioFan13 28 September 2022 \n            10/10\n
## 16      Leofwine_draca  28 November 2019 \n            10/10\n
## 17            jfirebug       21 May 2019 \n            10/10\n
## 18          ahmetkozan       8 June 2019 \n            10/10\n
## 19       deepfrieddodo  6 September 2022 \n            10/10\n
## 20           emholberg       27 May 2019  \n             9/10\n
## 21          mjplysaght   2 February 2017 \n            10/10\n
## 22           cuzzinman 21 September 2018 \n            10/10\n
## 23 A_Different_Drummer 18 September 2016 \n            10/10\n
## 24          Quinoa1984    16 August 2009 \n            10/10\n
## 25        KineticSeoul   11 October 2010 \n            10/10\n
##
## 1
## 2
## 3  I was 13 years old when COSMOS premiered on PBS in September of 1980 and amazingly as this may sou
## 4
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16
## 17
## 18
## 19
## 20
## 21
## 22
## 23
## 24
## 25
```