RWorksheet#5_group(Lomibao,rabago and andigan)

2024-11-18

bow and load the necesssary packages

```
library(kableExtra)
library("rvest")
library("polite")
library("dplyr")
```

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

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

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

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

```
library("stringr")

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

```
## v Setting active project to "/cloud/project".
```

```
url <- "https://www.imdb.com/chart/toptv/?ref_=nv_tvv_250v"
webpage <- read_html(url)
 session <- bow(url,
                user_agent = "Student education purpose")
 session
```

```
## <polite session> https://www.imdb.com/chart/toptv/?ref_=nv_tvv_250v
##     User-agent: Student education purpose
##     robots.txt: 35 rules are defined for 3 bots
##    Crawl delay: 5 sec
##   The path is scrapable for this user-agent
```

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

scraping the title

```
 title <- webpage%>%html_nodes('h3.ipc-title__text')%>%html_text()
 title <- title[2:26]
 title
```

```
##  [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. Arcane"
```

```
title_list <- as.data.frame(title[1:50])
colnames(title_list)<-"ranks"
```

spliting the data frame

```
split_df <- strsplit(as.character(title_list$ranks),".",fixed = TRUE)
split_df<- data.frame(do.call(rbind,split_df))
split_df
```

```
##      X1                              X2
## 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                             Arcane
## 26 <NA>                              <NA>
## 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>
```

renaming the columns

```
split_df<-split_df[-c(3,4)]
colnames(split_df)<- c("Ranks","Titles")
split_df
```

```
##    Ranks                            Titles
## 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                            Arcane
## 26  <NA>                              <NA>
## 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>
```

scraping the star- rating and saving in the data frame

```
ratings<- webpage %>%
  html_nodes('span.ipc-rating-star--rating') %>%
  html_text()
ratings <- as.data.frame(ratings)
```

scraping the numbers of vote

```
number_votes <- webpage %>%
  html_nodes("span.ipc-rating-star--voteCount") %>%
  html_text()
number_votes <- as.data.frame(number_votes)
```

scraping the number of episode

```
num_ep <- webpage %>%
  html_nodes('span.sc-300a8231-7.eaXxft.cli-title-metadata-item:nth-of-type(2)')%>%
  html_text()
num_ep
```

```
##  [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"  "18 eps"
```

cleanig the episode data.

```
episode_counts <- str_extract(num_ep, "\\d+ eps")
number_episode <- str_remove(episode_counts, " eps")
number_episode <- as.data.frame(number_episode)
colnames(number_episode) <- "Episode"
number_episode
```

```
##    Episode
## 1       62
## 2        6
## 3       11
## 4       10
## 5        5
## 6       60
## 7       62
## 8        7
## 9       86
## 10      13
## 11      13
## 12      12
## 13      74
## 14     194
## 15      26
## 16      68
## 17      78
## 18      11
## 19      10
## 20     156
## 21      10
## 22      15
## 23      98
## 24      85
## 25      18
```

scraping the year release

```
year <- webpage %>%
  html_nodes('span.sc-300a8231-7.eaXxft.cli-title-metadata-item') %>%html_text()
year
```

```
##  [1] "2008-2013" "62 eps"    "TV-MA"     "2016"      "6 eps"     "TV-G"
##  [7] "2006"      "11 eps"    "TV-PG"     "2001"      "10 eps"    "TV-MA"
## [13] "2019"      "5 eps"     "TV-MA"     "2002-2008" "60 eps"    "TV-MA"
## [19] "2005-2008" "62 eps"    "TV-Y7-FV"  "2017"      "7 eps"     "TV-G"
## [25] "1999-2007" "86 eps"    "TV-MA"     "2014"      "13 eps"    "TV-PG"
## [31] "1980"      "13 eps"    "TV-PG"     "2019-2023" "12 eps"    "TV-PG"
## [37] "2011-2019" "74 eps"    "TV-MA"     "2018- "    "194 eps"   "TV-Y"
## [43] "1973-1974" "26 eps"    "TV-PG"     "2009-2010" "68 eps"    "TV-14"
## [49] "2013- "    "78 eps"    "TV-MA"     "2009"      "11 eps"    "TV-G"
## [55] "2020"      "10 eps"    "TV-MA"     "1959-1964" "156 eps"   "TV-PG"
## [61] "2017"      "10 eps"    "TV-MA"     "2010-2017" "15 eps"    "TV-14"
## [67] "2013-2023" "98 eps"    "TV-MA"     "1992-1995" "85 eps"    "TV-PG"
## [73] "2021-2024" "18 eps"    "TV-14"
```

Exracting using the regex.

```
release_years <- str_extract(year, "\\d{4}")
release_years <- release_years[!is.na(release_years)]  # Remove NA values
release_years <- as.numeric(release_years)
relyear <- as.data.frame(release_years)
colnames(relyear) <-"Year"
relyear
```

```
##    Year
## 1  2008
## 2  2016
## 3  2006
## 4  2001
## 5  2019
## 6  2002
## 7  2005
## 8  2017
## 9  1999
## 10 2014
## 11 1980
## 12 2019
## 13 2011
## 14 2018
## 15 1973
## 16 2009
## 17 2013
## 18 2009
## 19 2020
## 20 1959
## 21 2017
## 22 2010
## 23 2013
## 24 1992
## 25 2021
```

creating csv file for every one of the data.

```
#title and ranks
#rank_title <- data.frame(
 # rank_title = split_df)
#write.csv(rank_title,file = "title.csv")
#ranting
#write.csv(rating,file = "star_rating.csv")
#vote count
 #write.csv(number_votes,file = "vote_count.csv")
#year
 # write.csv(relyear = "year.csv")
#number of episode
 #write.csv(number_episode = "number_episode.csv")
```

checking the length.

```
cat("Show Titles length: ", length(title), "\n")
```

```
## Show Titles length:  25
```

```r
cat("Show Ratings length: ", length(ratings), "\n")
```

```
## Show Ratings length:  1
```

```r
cat("Number of Votes length: ", length(number_votes), "\n")
```

```
## Number of Votes length:  1
```

```r
cat("Episode Counts length: ", length(number_episode), "\n")
```

```
## Episode Counts length:  1
```

```r
cat("Release Years length: ", length(release_years), "\n")
```

```
## Release Years length:  25
```

Combining them all to a data frame.

```r
 # imdb_top_tv_shows <- data.frame(
 # Title = rank_title,
#  Rating = ratings,
#  Votes = number_votes,
 # Episode = episodes,
 # Release_Year = release_years,
  #stringsAsFactors = FALSE
 # )
#imdb_top_tv_shows
```