RWorksheet#5_group(Leysa,Calambro)

Camilo Leysa

2024-11-11

Loading needed libraries:

```
library(rvest)
```

```
## Warning: package 'rvest' was built under R version 4.4.2
```

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

```
## Warning: package 'httr' was built under R version 4.4.2
```

```
library(stringr)
```

```
polite::use_manners(save_as = "polite_scrape_tvshows.R")
```

```
## v Setting active project to "C:/Worksheet_#5".
```

```
url <- "https://www.imdb.com/chart/toptv/?ref_=nv_tvv_250"
session <- bow(url, user_agent = "Educational")
session
```

```
## <polite session> https://www.imdb.com/chart/toptv/?ref_=nv_tvv_250
##     User-agent: Educational
##     robots.txt: 35 rules are defined for 3 bots
##    Crawl delay: 5 sec
##   The path is scrapable for this user-agent
```

Getting the TV Show title.

```
#Title
title_lists <- scrape(session) %>% html_nodes("h3.ipc-title__text") %>% html_text(trim = TRUE)
#filter unwanted
title_lists <- title_lists[!grepl("Recently viewed", title_lists)]
title_lists
```

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

List of top 50 TV Shows

```
class(title_lists)
```

```
## [1] "character"
```

```
title_List <- as.data.frame(title_lists[2:51])
title_List
```

```
##                      title_lists[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. 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
## 26                                <NA>
## 27                                <NA>
## 28                                <NA>
## 29                                <NA>
## 30                                <NA>
## 31                                <NA>
## 32                                <NA>
## 33                                <NA>
## 34                                <NA>
## 35                                <NA>
## 36                                <NA>
## 37                                <NA>
## 38                                <NA>
## 39                                <NA>
## 40                                <NA>
## 41                                <NA>
## 42                                <NA>
## 43                                <NA>
## 44                                <NA>
## 45                                <NA>
## 46                                <NA>
## 47                                <NA>
## 48                                <NA>
## 49                                <NA>
## 50                                <NA>
```

Seperating the rank number and the TV Show title.

```
colnames(title_List) <- "ranks"
split_df <- strsplit(as.character(title_List$ranks),".",fixed = TRUE)
split_df <- data.frame(do.call(rbind,split_df))
split_df <- split_df[-c(3:4)]
colnames(split_df) <- c("Ranks","Title")
str(split_df)
```

```
## 'data.frame':    50 obs. of  2 variables:
##  $ Ranks: chr  "1" "2" "3" "4" ...
##  $ Title: chr  " Breaking Bad" " Planet Earth II" " Planet Earth" " Band of Brothers" ...
```

The Rank and the Title of the TV Shows

```
class(split_df)
```

```
## [1] "data.frame"
```

```
split_df
```

```
##    Ranks                           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                       The Office
## 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>
```

Top 50 TV Show Rating

```
rating <- scrape(session) %>% html_nodes("span.ipc-rating-star--rating") %>% html_text
tv_rating <- as.data.frame(rating [1:50])
tv_rating
```

```
##    rating[1:50]
## 1           9.5
## 2           9.5
## 3           9.4
## 4           9.4
## 5           9.3
## 6           9.3
## 7           9.3
## 8           9.3
## 9           9.2
## 10          9.2
## 11          9.3
## 12          9.2
## 13          9.2
## 14          9.3
## 15          9.2
## 16          9.1
## 17          9.1
## 18          9.1
## 19          9.1
## 20          9.0
## 21          9.1
## 22          9.1
## 23          9.1
## 24          9.0
## 25          9.0
## 26         <NA>
## 27         <NA>
## 28         <NA>
## 29         <NA>
## 30         <NA>
## 31         <NA>
## 32         <NA>
## 33         <NA>
## 34         <NA>
## 35         <NA>
## 36         <NA>
## 37         <NA>
## 38         <NA>
## 39         <NA>
## 40         <NA>
## 41         <NA>
## 42         <NA>
## 43         <NA>
## 44         <NA>
## 45         <NA>
## 46         <NA>
## 47         <NA>
```

```
## 48          <NA>
## 49          <NA>
## 50          <NA>
```

Number of People who Voted

```
tv_votes <- scrape(session) %>% html_nodes("span.ipc-rating-star--voteCount") %>% html_text
total_tv_votes <- as.data.frame(tv_votes[1:50])
total_tv_votes
```

```
##    tv_votes[1:50]
## 1          (2.2M)
## 2          (162K)
## 3          (223K)
## 4          (544K)
## 5          (905K)
## 6          (390K)
## 7          (388K)
## 8           (48K)
## 9          (497K)
## 10         (131K)
## 11          (45K)
## 12          (53K)
## 13         (2.4M)
## 14          (33K)
## 15          (31K)
## 16         (208K)
## 17         (626K)
## 18          (43K)
## 19         (159K)
## 20          (96K)
## 21          (29K)
## 22           (1M)
## 23         (559K)
## 24         (122K)
## 25         (744K)
## 26            <NA>
## 27            <NA>
## 28            <NA>
## 29            <NA>
## 30            <NA>
## 31            <NA>
## 32            <NA>
## 33            <NA>
## 34            <NA>
## 35            <NA>
## 36            <NA>
## 37            <NA>
## 38            <NA>
## 39            <NA>
## 40            <NA>
## 41            <NA>
## 42            <NA>
```

```
## 43           <NA>
## 44           <NA>
## 45           <NA>
## 46           <NA>
## 47           <NA>
## 48           <NA>
## 49           <NA>
## 50           <NA>
```

Number of Episodes of each TV Shows

```
episodes <- scrape(session) %>% html_nodes("span.sc-5bc66c50-6.OOdsw") %>% html_text
cl_episodes <- gsub("\\D", "", episodes)
cleaned_ep <- str_extract(episodes, "\\d+(?=\\s*eps)")
cleaned_ep <- as.numeric(cleaned_ep)
cleaned_ep <- cleaned_ep[!is.na(cleaned_ep)]
cleaned_episodes <- as.data.frame(cleaned_ep[1:25])
cleaned_episodes
```

```
##    cleaned_ep[1:25]
## 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              188
```

Year of TV Shows released

```
tv_years <- scrape(session) %>% html_nodes("span.sc-5bc66c50-6.OOdsw") %>% html_text
clyear <- gsub(".*?(\\d{4}(-\\d{4})?).*", "\\1", tv_years)
yeartv <- str_extract(tv_years, "\\b\\d{4}(-\\d{4})?\\b")
yeartv <- as.numeric(yeartv)
yeartv <- yeartv[!is.na(yeartv)]
```

```
tv_year_of_air <- as.data.frame(yeartv[1:25])
tv_year_of_air
```

```
##    yeartv[1:25]
## 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         2005
```

Data frame of TV Shows

```
final_data <- cbind(split_df,tv_rating,cleaned_episodes,tv_year_of_air)
colnames(final_data) <- c("Ranks", "TV Rating", "Number of Votes", "Number of Episodes", "Year Released
final_data
```

```
##    Ranks                        TV Rating Number of Votes Number of Episodes
## 1      1                     Breaking Bad             9.5                 62
## 2      2                  Planet Earth II             9.5                  6
## 3      3                     Planet Earth             9.4                 11
## 4      4                 Band of Brothers             9.4                 10
## 5      5                        Chernobyl             9.3                  5
## 6      6                         The Wire             9.3                 60
## 7      7       Avatar: The Last Airbender             9.3                 62
## 8      8                    Blue Planet II             9.3                  7
## 9      9                     The Sopranos             9.2                 86
## 10    10      Cosmos: A Spacetime Odyssey             9.2                 13
## 11    11                           Cosmos             9.3                 13
## 12    12                       Our Planet             9.2                 12
## 13    13                  Game of Thrones             9.2                 74
## 14    14                            Bluey             9.3                194
## 15    15                The World at War             9.2                 26
```

```
## 16    16  Fullmetal Alchemist Brotherhood            9.1                68
## 17    17                   Rick and Morty            9.1                78
## 18    18                             Life            9.1                11
## 19    19                  The Last Dance             9.1                10
## 20    20                The Twilight Zone            9.0               156
## 21    21                  The Vietnam War            9.1                10
## 22    22                         Sherlock            9.1                15
## 23    23                  Attack on Titan            9.1                98
## 24    24      Batman: The Animated Series            9.0                85
## 25    25                        The Office            9.0               188
## 26  <NA>                             <NA>           <NA>                62
## 27  <NA>                             <NA>           <NA>                 6
## 28  <NA>                             <NA>           <NA>                11
## 29  <NA>                             <NA>           <NA>                10
## 30  <NA>                             <NA>           <NA>                 5
## 31  <NA>                             <NA>           <NA>                60
## 32  <NA>                             <NA>           <NA>                62
## 33  <NA>                             <NA>           <NA>                 7
## 34  <NA>                             <NA>           <NA>                86
## 35  <NA>                             <NA>           <NA>                13
## 36  <NA>                             <NA>           <NA>                13
## 37  <NA>                             <NA>           <NA>                12
## 38  <NA>                             <NA>           <NA>                74
## 39  <NA>                             <NA>           <NA>               194
## 40  <NA>                             <NA>           <NA>                26
## 41  <NA>                             <NA>           <NA>                68
## 42  <NA>                             <NA>           <NA>                78
## 43  <NA>                             <NA>           <NA>                11
## 44  <NA>                             <NA>           <NA>                10
## 45  <NA>                             <NA>           <NA>               156
## 46  <NA>                             <NA>           <NA>                10
## 47  <NA>                             <NA>           <NA>                15
## 48  <NA>                             <NA>           <NA>                98
## 49  <NA>                             <NA>           <NA>                85
## 50  <NA>                             <NA>           <NA>               188
##    Year Released
## 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          2005
## 26          2008
## 27          2016
## 28          2006
## 29          2001
## 30          2019
## 31          2002
## 32          2005
## 33          2017
## 34          1999
## 35          2014
## 36          1980
## 37          2019
## 38          2011
## 39          2018
## 40          1973
## 41          2009
## 42          2013
## 43          2009
## 44          2020
## 45          1959
## 46          2017
## 47          2010
## 48          2013
## 49          1992
## 50          2005
```