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
```

```
library(ggplot2)

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

titles_df <- as.data.frame(title_elements[3:52], stringsAsFactors = FALSE)
```

```r
colnames(titles_df) <- "Ranked_Titles"

split_titles <- strsplit(as.character(titles_df$Ranked_Titles), "\\.", fixed = FALSE)
titles_split_df <- data.frame(do.call(rbind, split_titles), stringsAsFactors = FALSE)

colnames(titles_split_df) <- c("Rank", "Title")
titles_split_df <- titles_split_df %>% select(Rank, Title)
titles_split_df$Title <- trimws(titles_split_df$Title)

rank_title <- titles_split_df

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

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

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

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_elements,
  Voters = voters_cleaned,
  Episodes = episodes_count,
  Year = years
)

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

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

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

  show_page <- read_html(complete_link)
  review_link <- show_page %>%
    html_nodes('a.isReview') %>%
    html_attr("href")
```

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

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

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

  return(data.frame(
    Show_Link = complete_link,
    User_Reviews = user_reviews_count,
    Critic_Reviews = critic_df,
    Popularity_Rating = popularity_score
  ))
})

show_details_df <- do.call(rbind, show_details_list)

final_shows_df <- cbind(top_tv_shows, show_details_df)

print(final_shows_df)
```

```
##               Rank                          Title Rating Voters Episodes
## 1                1                   Breaking Bad    9.5   2.2M       62
## 2                2                Planet Earth II    9.5   162K        6
## 3                3                   Planet Earth    9.4   223K       11
## 4                4                Band of Brothers    9.4   545K       10
## 5                5                      Chernobyl    9.3   905K        5
## 6                6                       The Wire    9.3   390K       60
## 7                7     Avatar: The Last Airbender    9.3   389K       62
## 8                8                  Blue Planet II    9.3    48K        7
## 9                9                   The Sopranos    9.2   497K       86
## 10              10    Cosmos: A Spacetime Odyssey    9.2   131K       13
## 11              11                         Cosmos    9.3    45K       13
## 12              12                     Our Planet    9.2    53K       12
## 13              13                Game of Thrones    9.2   2.4M       74
## 14              14                          Bluey    9.3    33K      194
## 15              15                The World at War    9.2    31K       26
## 16              16 Fullmetal Alchemist Brotherhood    9.1   208K       68
## 17              17                  Rick and Morty    9.1   626K       78
## 18              18                            Life    9.1    43K       11
## 19              19                The Last Dance    9.1   159K       10
## 20              20              The Twilight Zone    9.0    96K      156
## 21              21                The Vietnam War    9.1    29K       10
## 22              22                        Sherlock    9.1     1M       15
## 23              23                Attack on Titan    9.1   559K       98
## 24              24    Batman: The Animated Series    9.0   122K       85
```

```
## 25              25                      The Office    9.0   745K      188
## 26 Recently viewed                  Recently viewed    9.5   2.2M       62
## 27            <NA>                            <NA>    9.5   162K        6
## 28            <NA>                            <NA>    9.4   223K       11
## 29            <NA>                            <NA>    9.4   545K       10
## 30            <NA>                            <NA>    9.3   905K        5
## 31            <NA>                            <NA>    9.3   390K       60
## 32            <NA>                            <NA>    9.3   389K       62
## 33            <NA>                            <NA>    9.3    48K        7
## 34            <NA>                            <NA>    9.2   497K       86
## 35            <NA>                            <NA>    9.2   131K       13
## 36            <NA>                            <NA>    9.3    45K       13
## 37            <NA>                            <NA>    9.2    53K       12
## 38            <NA>                            <NA>    9.2   2.4M       74
## 39            <NA>                            <NA>    9.3    33K      194
## 40            <NA>                            <NA>    9.2    31K       26
## 41            <NA>                            <NA>    9.1   208K       68
## 42            <NA>                            <NA>    9.1   626K       78
## 43            <NA>                            <NA>    9.1    43K       11
## 44            <NA>                            <NA>    9.1   159K       10
## 45            <NA>                            <NA>    9.0    96K      156
## 46            <NA>                            <NA>    9.1    29K       10
## 47            <NA>                            <NA>    9.1     1M       15
## 48            <NA>                            <NA>    9.1   559K       98
## 49            <NA>                            <NA>    9.0   122K       85
## 50            <NA>                            <NA>    9.0   745K      188
##         Year                                              Show_Link  User_Reviews
## 1  2008-2013  https://imdb.com/title/tt0903747/?ref_=chttvtp_t_1 5,086 reviews
## 2       2016  https://imdb.com/title/tt0903747/?ref_=chttvtp_t_1 5,086 reviews
## 3       2006  https://imdb.com/title/tt5491994/?ref_=chttvtp_t_2   158 reviews
## 4       2001  https://imdb.com/title/tt5491994/?ref_=chttvtp_t_2   158 reviews
## 5       2019  https://imdb.com/title/tt0795176/?ref_=chttvtp_t_3   111 reviews
## 6  2002-2008  https://imdb.com/title/tt0795176/?ref_=chttvtp_t_3   111 reviews
## 7  2005-2008  https://imdb.com/title/tt0185906/?ref_=chttvtp_t_4 1,056 reviews
## 8       2017  https://imdb.com/title/tt0185906/?ref_=chttvtp_t_4 1,056 reviews
## 9  1999-2007  https://imdb.com/title/tt7366338/?ref_=chttvtp_t_5 3,530 reviews
## 10      2014  https://imdb.com/title/tt7366338/?ref_=chttvtp_t_5 3,530 reviews
## 11      1980  https://imdb.com/title/tt0306414/?ref_=chttvtp_t_6   787 reviews
## 12 2019-2023  https://imdb.com/title/tt0306414/?ref_=chttvtp_t_6   787 reviews
## 13 2011-2019  https://imdb.com/title/tt0417299/?ref_=chttvtp_t_7   998 reviews
## 14    2018-   https://imdb.com/title/tt0417299/?ref_=chttvtp_t_7   998 reviews
## 15 1973-1974  https://imdb.com/title/tt6769208/?ref_=chttvtp_t_8    53 reviews
## 16 2009-2010  https://imdb.com/title/tt6769208/?ref_=chttvtp_t_8    53 reviews
## 17    2013-   https://imdb.com/title/tt0141842/?ref_=chttvtp_t_9   962 reviews
## 18      2009  https://imdb.com/title/tt0141842/?ref_=chttvtp_t_9   962 reviews
## 19      2020 https://imdb.com/title/tt2395695/?ref_=chttvtp_t_10   205 reviews
## 20 1959-1964 https://imdb.com/title/tt2395695/?ref_=chttvtp_t_10   205 reviews
## 21      2017 https://imdb.com/title/tt0081846/?ref_=chttvtp_t_11    80 reviews
## 22 2010-2017 https://imdb.com/title/tt0081846/?ref_=chttvtp_t_11    80 reviews
## 23 2013-2023 https://imdb.com/title/tt9253866/?ref_=chttvtp_t_12   245 reviews
## 24 1992-1995 https://imdb.com/title/tt9253866/?ref_=chttvtp_t_12   245 reviews
## 25 2005-2013 https://imdb.com/title/tt0944947/?ref_=chttvtp_t_13 5,898 reviews
## 26 2008-2013 https://imdb.com/title/tt0944947/?ref_=chttvtp_t_13 5,898 reviews
## 27      2016 https://imdb.com/title/tt7678620/?ref_=chttvtp_t_14   367 reviews
```

```
## 28      2006 https://imdb.com/title/tt7678620/?ref_=chttvtp_t_14   367 reviews
## 29      2001 https://imdb.com/title/tt0071075/?ref_=chttvtp_t_15   126 reviews
## 30      2019 https://imdb.com/title/tt0071075/?ref_=chttvtp_t_15   126 reviews
## 31 2002-2008 https://imdb.com/title/tt1355642/?ref_=chttvtp_t_16   465 reviews
## 32 2005-2008 https://imdb.com/title/tt1355642/?ref_=chttvtp_t_16   465 reviews
## 33      2017 https://imdb.com/title/tt2861424/?ref_=chttvtp_t_17   908 reviews
## 34 1999-2007 https://imdb.com/title/tt2861424/?ref_=chttvtp_t_17   908 reviews
## 35      2014 https://imdb.com/title/tt1533395/?ref_=chttvtp_t_18    12 reviews
## 36      1980 https://imdb.com/title/tt1533395/?ref_=chttvtp_t_18    12 reviews
## 37 2019-2023 https://imdb.com/title/tt8420184/?ref_=chttvtp_t_19   541 reviews
## 38 2011-2019 https://imdb.com/title/tt8420184/?ref_=chttvtp_t_19   541 reviews
## 39    2018-  https://imdb.com/title/tt0052520/?ref_=chttvtp_t_20   213 reviews
## 40 1973-1974 https://imdb.com/title/tt0052520/?ref_=chttvtp_t_20   213 reviews
## 41 2009-2010 https://imdb.com/title/tt1877514/?ref_=chttvtp_t_21   175 reviews
## 42    2013-  https://imdb.com/title/tt1877514/?ref_=chttvtp_t_21   175 reviews
## 43      2009 https://imdb.com/title/tt1475582/?ref_=chttvtp_t_22 1,095 reviews
## 44      2020 https://imdb.com/title/tt1475582/?ref_=chttvtp_t_22 1,095 reviews
## 45 1959-1964 https://imdb.com/title/tt2560140/?ref_=chttvtp_t_23 2,358 reviews
## 46      2017 https://imdb.com/title/tt2560140/?ref_=chttvtp_t_23 2,358 reviews
## 47 2010-2017 https://imdb.com/title/tt0103359/?ref_=chttvtp_t_24   219 reviews
## 48 2013-2023 https://imdb.com/title/tt0103359/?ref_=chttvtp_t_24   219 reviews
## 49 1992-1995 https://imdb.com/title/tt0386676/?ref_=chttvtp_t_25 1,774 reviews
## 50 2005-2013 https://imdb.com/title/tt0386676/?ref_=chttvtp_t_25 1,774 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
```

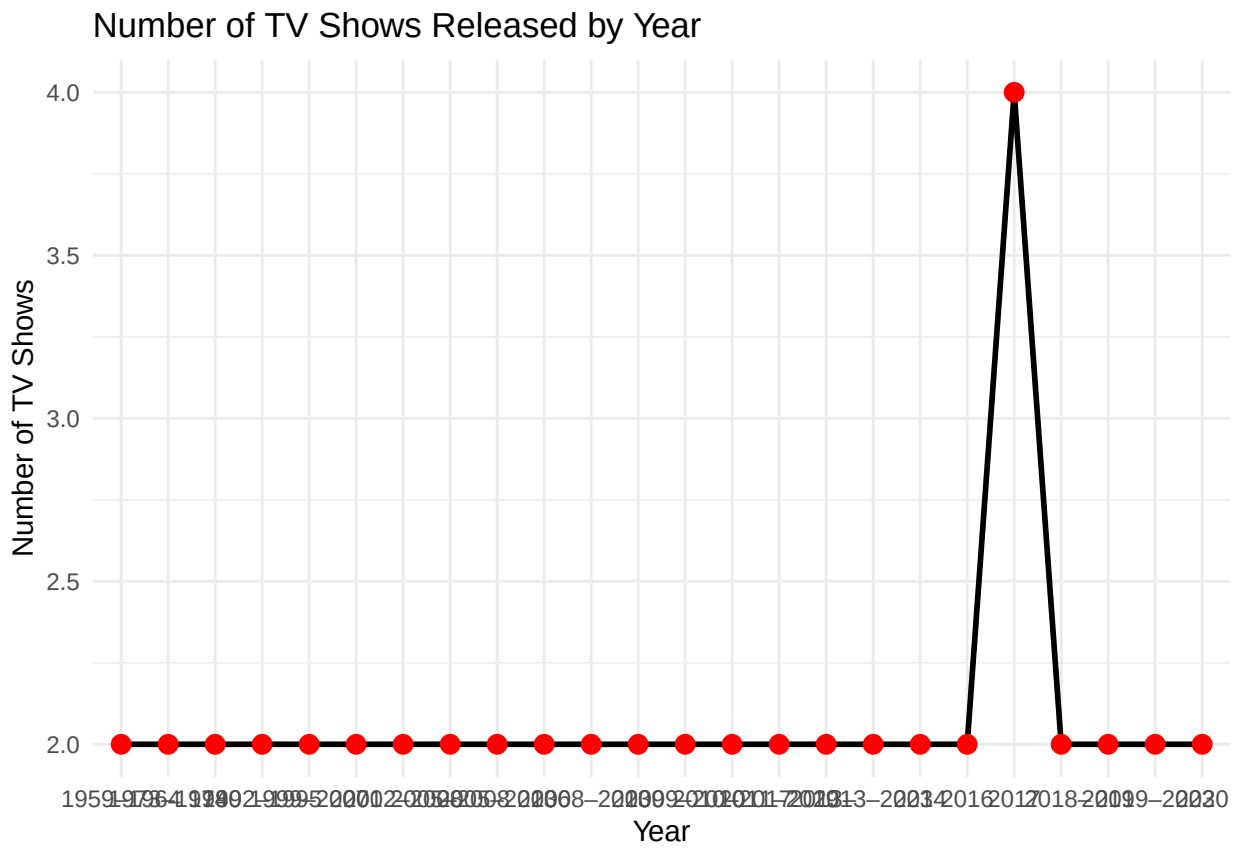
```r
View(final_shows_df)


#3. Create a time series graph for the tv shows released by year. Which year has the most number of tv

shows_by_year <- top_tv_shows %>%
  group_by(Year) %>%
  summarise(Count = n()) %>%
  arrange(Year)

ggplot(shows_by_year, aes(x = Year, y = Count, group = 1)) +
  geom_line(color = "black", size = 1) +
  geom_point(color = "red", size = 3) +
  labs(
    title = "Number of TV Shows Released by Year",
    x = "Year",
    y = "Number of TV Shows"
  ) +
  theme_minimal()
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```




```
max_shows_year <- shows_by_year %>%
  filter(Count == max(Count))

print(max_shows_year)
```

```
## # A tibble: 1 x 2
##   Year  Count
##   <chr> <int>
## 1 2017      4
```