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.

```
library(polite)
library(httr)
library(rvest)
library(dplyr)
```

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

```
##
## 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(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[3:27], 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$Title <- trimws(titles_split_df$Title)

rank_title <- titles_split_df
rank_title
```

```
##    Rank                           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
```

```r
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()
```

```r
episodes_cleaned <- gsub('[eps]', '', episode_elements)
episodes_count <- as.numeric(episodes_cleaned)

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()

min_length <- min(nrow(rank_title), length(rating_elements), length(voters_cleaned), length(episodes_cou


rank_title <- rank_title[1:min_length, ]
rating_elements <- rating_elements[1:min_length]
voters_cleaned <- voters_cleaned[1:min_length]
episodes_count <- episodes_count[1:min_length]
years <- years[1:min_length]

top_tv_shows <- data.frame(
  Rank = rank_title$Rank,
  Title = rank_title$Title,
  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 Year
## 1     1 Breaking Bad    9.5   2.2M       NA <NA>
## 2     1 Breaking Bad    9.5   2.2M       NA <NA>
## 3     1 Breaking Bad    9.5   2.2M       NA <NA>
## 4     1 Breaking Bad    9.5   2.2M       NA <NA>
## 5     1 Breaking Bad    9.5   2.2M       NA <NA>
## 6     1 Breaking Bad    9.5   2.2M       NA <NA>
## 7     1 Breaking Bad    9.5   2.2M       NA <NA>
## 8     1 Breaking Bad    9.5   2.2M       NA <NA>
## 9     1 Breaking Bad    9.5   2.2M       NA <NA>
## 10    1 Breaking Bad    9.5   2.2M       NA <NA>
## 11    1 Breaking Bad    9.5   2.2M       NA <NA>
## 12    1 Breaking Bad    9.5   2.2M       NA <NA>
## 13    1 Breaking Bad    9.5   2.2M       NA <NA>
## 14    1 Breaking Bad    9.5   2.2M       NA <NA>
## 15    1 Breaking Bad    9.5   2.2M       NA <NA>
## 16    1 Breaking Bad    9.5   2.2M       NA <NA>
## 17    1 Breaking Bad    9.5   2.2M       NA <NA>
## 18    1 Breaking Bad    9.5   2.2M       NA <NA>
## 19    1 Breaking Bad    9.5   2.2M       NA <NA>
## 20    1 Breaking Bad    9.5   2.2M       NA <NA>
## 21    1 Breaking Bad    9.5   2.2M       NA <NA>
## 22    1 Breaking Bad    9.5   2.2M       NA <NA>
## 23    1 Breaking Bad    9.5   2.2M       NA <NA>
## 24    1 Breaking Bad    9.5   2.2M       NA <NA>
## 25    1 Breaking Bad    9.5   2.2M       NA <NA>
## 26    1 Breaking Bad    9.5   2.2M       NA <NA>
## 27    1 Breaking Bad    9.5   2.2M       NA <NA>
## 28    1 Breaking Bad    9.5   2.2M       NA <NA>
## 29    1 Breaking Bad    9.5   2.2M       NA <NA>
## 30    1 Breaking Bad    9.5   2.2M       NA <NA>
```

```
## 31    1 Breaking Bad    9.5   2.2M       NA <NA>
## 32    1 Breaking Bad    9.5   2.2M       NA <NA>
## 33    1 Breaking Bad    9.5   2.2M       NA <NA>
## 34    1 Breaking Bad    9.5   2.2M       NA <NA>
## 35    1 Breaking Bad    9.5   2.2M       NA <NA>
## 36    1 Breaking Bad    9.5   2.2M       NA <NA>
## 37    1 Breaking Bad    9.5   2.2M       NA <NA>
## 38    1 Breaking Bad    9.5   2.2M       NA <NA>
## 39    1 Breaking Bad    9.5   2.2M       NA <NA>
## 40    1 Breaking Bad    9.5   2.2M       NA <NA>
## 41    1 Breaking Bad    9.5   2.2M       NA <NA>
## 42    1 Breaking Bad    9.5   2.2M       NA <NA>
## 43    1 Breaking Bad    9.5   2.2M       NA <NA>
## 44    1 Breaking Bad    9.5   2.2M       NA <NA>
## 45    1 Breaking Bad    9.5   2.2M       NA <NA>
## 46    1 Breaking Bad    9.5   2.2M       NA <NA>
## 47    1 Breaking Bad    9.5   2.2M       NA <NA>
## 48    1 Breaking Bad    9.5   2.2M       NA <NA>
## 49    1 Breaking Bad    9.5   2.2M       NA <NA>
## 50    1 Breaking Bad    9.5   2.2M       NA <NA>
##                                                Show_Link  User_Reviews
## 1    https://imdb.com/title/tt0903747/?ref_=chttvtp_t_1 5,118 reviews
## 2    https://imdb.com/title/tt0903747/?ref_=chttvtp_t_1 5,118 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 1,059 reviews
## 8    https://imdb.com/title/tt0185906/?ref_=chttvtp_t_4 1,059 reviews
## 9    https://imdb.com/title/tt7366338/?ref_=chttvtp_t_5 3,538 reviews
## 10   https://imdb.com/title/tt7366338/?ref_=chttvtp_t_5 3,538 reviews
## 11   https://imdb.com/title/tt0306414/?ref_=chttvtp_t_6   787 reviews
## 12   https://imdb.com/title/tt0306414/?ref_=chttvtp_t_6   787 reviews
## 13   https://imdb.com/title/tt0417299/?ref_=chttvtp_t_7 1,001 reviews
## 14   https://imdb.com/title/tt0417299/?ref_=chttvtp_t_7 1,001 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   968 reviews
## 18   https://imdb.com/title/tt0141842/?ref_=chttvtp_t_9   968 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 5,914 reviews
## 26  https://imdb.com/title/tt0944947/?ref_=chttvtp_t_13 5,914 reviews
## 27  https://imdb.com/title/tt7678620/?ref_=chttvtp_t_14   369 reviews
## 28  https://imdb.com/title/tt7678620/?ref_=chttvtp_t_14   369 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   468 reviews
## 32  https://imdb.com/title/tt1355642/?ref_=chttvtp_t_16   468 reviews
## 33  https://imdb.com/title/tt2861424/?ref_=chttvtp_t_17   911 reviews
```

```
## 34  https://imdb.com/title/tt2861424/?ref_=chttvtp_t_17   911 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   542 reviews
## 38  https://imdb.com/title/tt8420184/?ref_=chttvtp_t_19   542 reviews
## 39  https://imdb.com/title/tt0052520/?ref_=chttvtp_t_20   214 reviews
## 40  https://imdb.com/title/tt0052520/?ref_=chttvtp_t_20   214 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 1,099 reviews
## 44  https://imdb.com/title/tt1475582/?ref_=chttvtp_t_22 1,099 reviews
## 45  https://imdb.com/title/tt2560140/?ref_=chttvtp_t_23 2,377 reviews
## 46  https://imdb.com/title/tt2560140/?ref_=chttvtp_t_23 2,377 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/tt11126994/?ref_=chttvtp_t_25 2,288 reviews
## 50 https://imdb.com/title/tt11126994/?ref_=chttvtp_t_25 2,288 reviews
##    Critic_Reviews Popularity_Rating
## 1             176                24
## 2             176                24
## 3               6             1,082
## 4               6             1,082
## 5              10             2,066
## 6              10             2,066
## 7              34               173
## 8              34               173
## 9              88               146
## 10             88               146
## 11             77               112
## 12             77               112
## 13             57               327
## 14             57               327
## 15              9             4,399
## 16              9             4,399
## 17             93                26
## 18             93                26
## 19             12             1,457
## 20             12             1,457
## 21              8             3,902
## 22              8             3,902
## 23             15             2,798
## 24             15             2,798
## 25            368                16
## 26            368                16
## 27              4               368
## 28              4               368
## 29              5             2,290
## 30              5             2,290
## 31             16               505
## 32             16               505
## 33             94               126
## 34             94               126
## 35              9             3,107
## 36              9             3,107
```

```
## 37             28            1,593
## 38             28            1,593
## 39             85              366
## 40             85              366
## 41             13            1,841
## 42             13            1,841
## 43            121              172
## 44            121              172
## 45             65               52
## 46             65               52
## 47             25              518
## 48             25              518
## 49             59                2
## 50             59                2
```

```
View(final_shows_df)
```

```
#2.
library(rvest)
library(dplyr)

url_5shows <- c(
  "https://www.imdb.com/title/tt0903747/reviews/?ref_=ttexr_ql_2",
  "https://www.imdb.com/title/tt2098220/?ref_=chttvtp_t_33",
  "https://www.imdb.com/title/tt2861424/?ref_=chttvtp_t_17",
  "https://www.imdb.com/title/tt2560140/?ref_=chttvtp_t_23",
  "https://www.imdb.com/title/tt11126994/?ref_=chttvtp_t_25"
)

five_df <- data.frame(
  Title = c(
    "Breaking Bad",
    "Hunter x Hunter",
    "Rick and Morty",
    "Attack on Titan",
    "Arcane"
  ),
  URLs = url_5shows
)

scrape_reviews <- function(show_url) {
  page <- read_html(show_url)

  usernames <- page %>%
    html_nodes('[data-testid="author-link"]') %>%
    html_text()

  review_dates <- page %>%
    html_nodes('li.review-date') %>%
    html_text()

  user_rating <- page %>%
    html_nodes('span.ipc-rating-star--rating') %>%
    html_text()
```

```
  rev_title <- page %>%
    html_nodes('h3.ipc-title__text') %>%
    html_text()

  text_rev <- page %>%
    html_nodes('div.ipc-html-content-inner-div') %>%
    html_text()


  min_reviews <- min(length(usernames), length(review_dates), length(user_rating),
                     length(rev_title), length(text_rev))

  data.frame(
    Usernames = head(usernames, min_reviews),
    Dates = head(review_dates, min_reviews),
    User_Rating = head(user_rating, min_reviews),
    Review_Title = head(rev_title, min_reviews),
    Text_Reviews = head(text_rev, min_reviews)
  )
}

reviews_data <- lapply(five_df$URLs, scrape_reviews)
names(reviews_data) <- five_df$TitleS


reviews_data[["Breaking Bad"]]
```

```
## NULL
```

```
reviews_data[["Hunter x Hunter"]]
```

```
## NULL
```

```
reviews_data[["Rick and Morty"]]
```

```
## NULL
```

```
reviews_data[["Attack on Titan"]]
```

```
## NULL
```

```
reviews_data[["Arcane"]]
```

```
## NULL
```

```
#3.
library(ggplot2)
library(polite)
library(httr)
```

```r
library(rvest)
library(dplyr)

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
```

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

titles_df <- as.data.frame(title_elements[3:52], stringsAsFactors = FALSE)
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$Title <- trimws(titles_split_df$Title)

rank_title <- titles_split_df
rank_title
```

```
##               Rank                          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
## 26 Recently viewed                Recently viewed
## 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>
```

```
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-300a8231-7.eaXxft.cli-title-metadata-item:nth-of-type(1)') %>%
  html_text()

cat("Rank and Title:", nrow(rank_title), "\n")
```

```
## Rank and Title: 50
```

```
cat("Ratings:", length(rating_elements), "\n")
```

```
## Ratings: 1
```

```r
cat("Voters:", length(voters_cleaned), "\n")
```

```
## Voters: 25
```

```r
cat("Episodes:", length(episodes_count), "\n")
```

```
## Episodes: 0
```

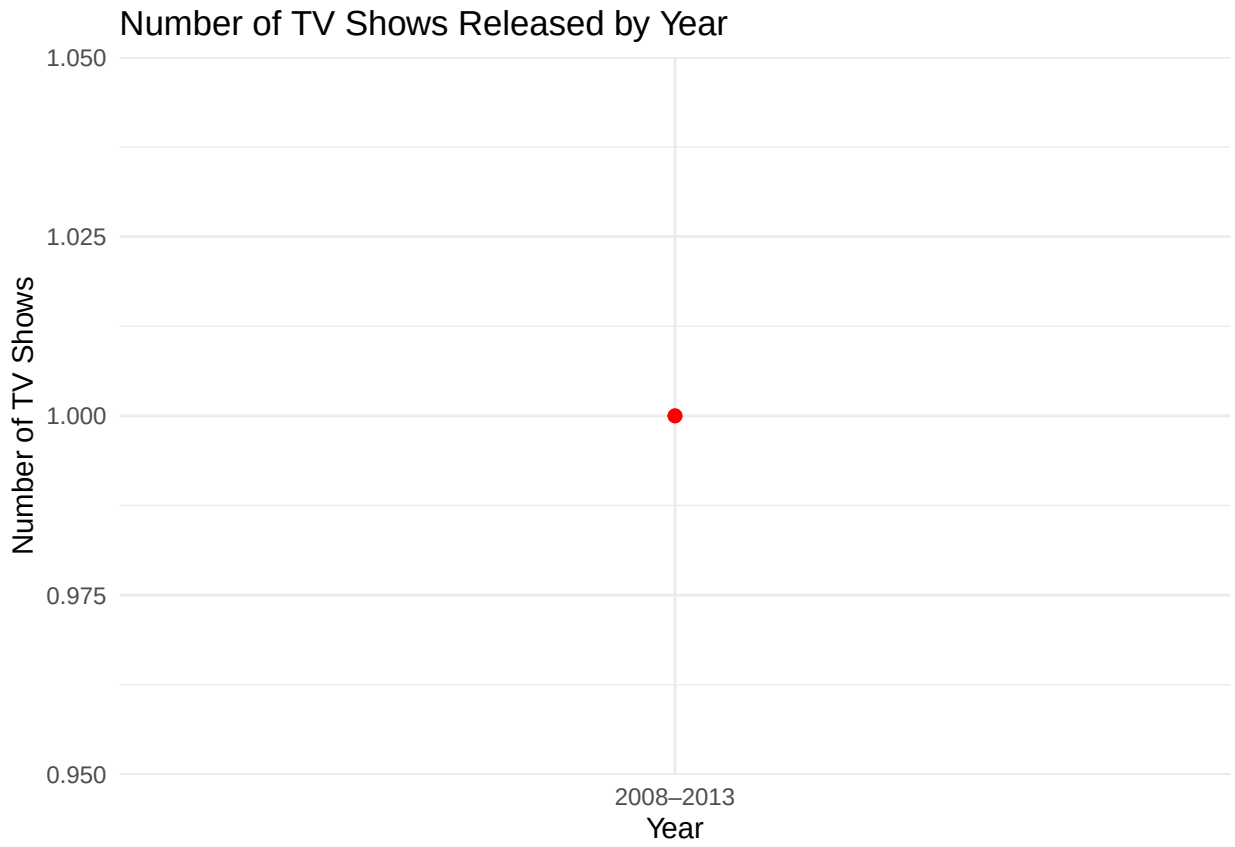
```r
cat("Years:", length(years), "\n")
```

```
## Years: 25
```

```r
min_length <- min(nrow(rank_title), length(rating_elements), length(voters_cleaned), length(episodes_co

rank_title <- rank_title[1:min_length, ]
rating_elements <- rating_elements[1:min_length]
voters_cleaned <- voters_cleaned[1:min_length]
episodes_count <- episodes_count[1:min_length]
years <- years[1:min_length]

top_tv_shows <- data.frame(
  Rank = rank_title[,1],
  Title = rank_title[,2],
  Rating = rating_elements,
  Voters = voters_cleaned,
  Episodes = episodes_count,
  Year = years
)


top_tv_shows <- top_tv_shows %>% filter(!is.na(Year) & Year != "")


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


ggplot(shows_by_year, aes(x = Year, y = Count)) +
  geom_line(color = "blue", size = 1) +
  geom_point(color = "red", size = 2) +
  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.
```

```
## `geom_line()`: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
```




```
most_shows_year <- shows_by_year[which.max(shows_by_year$Count), ]
cat("The year with the most number of TV shows released is:", most_shows_year$Year, "with", most_shows_
```

```
## The year with the most number of TV shows released is: 2008-2013 with 1 shows.
```

```
#4.
```