RWorksheet#5_group(Corvera, Paclibar, Sabarillo)

Rotciv Corvera, Jhon Albert Paclibar, Kirk Axl Dend Sabarillo

2024-11-11

1. Extracting 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(stringr)
library(magrittr)
library(ggplot2)

url <- "https://www.imdb.com/chart/toptv/?sort=rank%2Casc"
#1
#get the ranks and titles
title_list <- read_html(url) %>%
  html_nodes('.ipc-title__text') %>%
  html_text()

#Clean extracted text
title_list_sub <- as.data.frame(title_list[3:27], stringsAsFactors = FALSE)
colnames(title_list_sub) <- "ranks"

split_df <- strsplit(as.character(title_list_sub$ranks), "\\.", fixed = FALSE)
split_df <- data.frame(do.call(rbind, split_df), stringsAsFactors = FALSE)

colnames(split_df) <- c("rank", "title")
split_df <- split_df %>% select(rank, title)

split_df$title <- trimws(split_df$title)

rank_title <- split_df

#get tv rating, the number of people who voted, the number of episodes, and the year it was released.
rating_ls <- read_html(url) %>%
```

```r
  html_nodes('.ipc-rating-star--rating') %>%
  html_text()

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

#get the number of episodes
eps_ls <- read_html(url) %>%
  html_nodes('span.sc-5bc66c50-6.OOdsw.cli-title-metadata-item:nth-of-type(2)') %>%
  html_text()
clean_eps <- gsub('[eps]', '', eps_ls)
num_eps <- as.numeric(clean_eps)

#get year released
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_ls,
  Voters = clean_votes,
  Episodes = num_eps,
  Year = years,
  stringsAsFactors = FALSE
)

#Number of user reviews
home_link <- 'https://www.imdb.com/chart/toptv/'
main_page <- read_html(home_link)

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

#get link of each show's page
show_data <- lapply(links, function(link) {
  complete_link <- paste0("https://imdb.com", link)

  #get the link for user review page
  usrv_link <- read_html(complete_link)
  usrv_link_page <- usrv_link %>%
    html_nodes('a.isReview') %>%
    html_attr("href")

  #get critic reviews
  critic <- usrv_link %>%
    html_nodes("span.score") %>%
    html_text()
  critic_df <- data.frame(Critic_Reviews = critic[2], stringsAsFactors = FALSE)
  #get pop rating
  pop_rating <- usrv_link %>%
    html_nodes('[data-testid="hero-rating-bar__popularity__score"]') %>%
    html_text()

  #get user reviews of each shows
  usrv <- read_html(paste0("https://imdb.com", usrv_link_page[1]))
  usrv_count <- usrv %>%
    html_nodes('[data-testid="tturv-total-reviews"]') %>%
    html_text()

  return(data.frame(Show_Link = complete_link, User_Reviews = usrv_count, Critic = critic_df, Popularit
})

show_url_df <- do.call(rbind, show_data)
show_url_df
```

```
##                                                Show_Link  User_Reviews
## 1    https://imdb.com/title/tt0903747/?ref_=chttvtp_t_1 5,091 reviews
## 2    https://imdb.com/title/tt0903747/?ref_=chttvtp_t_1 5,091 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 1,056 reviews
## 8    https://imdb.com/title/tt0185906/?ref_=chttvtp_t_4 1,056 reviews
## 9    https://imdb.com/title/tt7366338/?ref_=chttvtp_t_5 3,532 reviews
## 10   https://imdb.com/title/tt7366338/?ref_=chttvtp_t_5 3,532 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   998 reviews
## 14   https://imdb.com/title/tt0417299/?ref_=chttvtp_t_7   998 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   963 reviews
## 18   https://imdb.com/title/tt0141842/?ref_=chttvtp_t_9   963 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 5,899 reviews
## 26  https://imdb.com/title/tt0944947/?ref_=chttvtp_t_13 5,899 reviews
## 27  https://imdb.com/title/tt7678620/?ref_=chttvtp_t_14   367 reviews
## 28  https://imdb.com/title/tt7678620/?ref_=chttvtp_t_14   367 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   466 reviews
## 32  https://imdb.com/title/tt1355642/?ref_=chttvtp_t_16   466 reviews
## 33  https://imdb.com/title/tt2861424/?ref_=chttvtp_t_17   909 reviews
## 34  https://imdb.com/title/tt2861424/?ref_=chttvtp_t_17   909 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   541 reviews
## 38  https://imdb.com/title/tt8420184/?ref_=chttvtp_t_19   541 reviews
## 39  https://imdb.com/title/tt0052520/?ref_=chttvtp_t_20   213 reviews
## 40  https://imdb.com/title/tt0052520/?ref_=chttvtp_t_20   213 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 1,095 reviews
## 44  https://imdb.com/title/tt1475582/?ref_=chttvtp_t_22 1,095 reviews
## 45  https://imdb.com/title/tt2560140/?ref_=chttvtp_t_23 2,359 reviews
## 46  https://imdb.com/title/tt2560140/?ref_=chttvtp_t_23 2,359 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 1,944 reviews
## 50 https://imdb.com/title/tt11126994/?ref_=chttvtp_t_25 1,944 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             53                 15
## 50             53                 15
```

```
shows <- cbind(top_tv_shows, show_url_df)
shows
```

```
##    rank                            title Rating Voters Episodes      Year
## 1     1                     Breaking Bad    9.5   2.2M       62 2008-2013
## 2     2                  Planet Earth II    9.5   162K        6      2016
## 3     3                     Planet Earth    9.4   223K       11      2006
## 4     4                 Band of Brothers    9.4   545K       10      2001
## 5     5                        Chernobyl    9.3   906K        5      2019
## 6     6                         The Wire    9.3   390K       60 2002-2008
## 7     7       Avatar: The Last Airbender    9.3   389K       62 2005-2008
## 8     8                   Blue Planet II    9.3    49K        7      2017
## 9     9                     The Sopranos    9.2   498K       86 1999-2007
## 10   10      Cosmos: A Spacetime Odyssey    9.2   131K       13      2014
## 11   11                           Cosmos    9.3    46K       13      1980
## 12   12                       Our Planet    9.2    54K       12 2019-2023
## 13   13                  Game of Thrones    9.2   2.4M       74 2011-2019
## 14   14                            Bluey    9.3    33K      194    2018-
## 15   15                 The World at War    9.2    31K       26 1973-1974
## 16   16 Fullmetal Alchemist: Brotherhood    9.1   208K       68 2009-2010
## 17   17                   Rick and Morty    9.1   626K       78    2013-
## 18   18                             Life    9.1    44K       11      2009
## 19   19                   The Last Dance    9.1   159K       10      2020
## 20   20                The Twilight Zone    9.0    97K      156 1959-1964
## 21   21                  The Vietnam War    9.1    29K       10      2017
## 22   22                         Sherlock    9.1     1M       15 2010-2017
## 23   23                  Attack on Titan    9.1   560K       98 2013-2023
## 24   24      Batman: The Animated Series    9.0   122K       85 1992-1995
## 25   25                           Arcane    9.0   300K       18 2021-2024
## 26    1                     Breaking Bad    9.5   2.2M       62 2008-2013
## 27    2                  Planet Earth II    9.5   162K        6      2016
## 28    3                     Planet Earth    9.4   223K       11      2006
## 29    4                 Band of Brothers    9.4   545K       10      2001
## 30    5                        Chernobyl    9.3   906K        5      2019
## 31    6                         The Wire    9.3   390K       60 2002-2008
## 32    7       Avatar: The Last Airbender    9.3   389K       62 2005-2008
## 33    8                   Blue Planet II    9.3    49K        7      2017
## 34    9                     The Sopranos    9.2   498K       86 1999-2007
## 35   10      Cosmos: A Spacetime Odyssey    9.2   131K       13      2014
## 36   11                           Cosmos    9.3    46K       13      1980
## 37   12                       Our Planet    9.2    54K       12 2019-2023
## 38   13                  Game of Thrones    9.2   2.4M       74 2011-2019
## 39   14                            Bluey    9.3    33K      194    2018-
## 40   15                 The World at War    9.2    31K       26 1973-1974
## 41   16 Fullmetal Alchemist: Brotherhood    9.1   208K       68 2009-2010
## 42   17                   Rick and Morty    9.1   626K       78    2013-
## 43   18                             Life    9.1    44K       11      2009
## 44   19                   The Last Dance    9.1   159K       10      2020
## 45   20                The Twilight Zone    9.0    97K      156 1959-1964
## 46   21                  The Vietnam War    9.1    29K       10      2017
## 47   22                         Sherlock    9.1     1M       15 2010-2017
## 48   23                  Attack on Titan    9.1   560K       98 2013-2023
## 49   24      Batman: The Animated Series    9.0   122K       85 1992-1995
## 50   25                           Arcane    9.0   300K       18 2021-2024
##                                                Show_Link  User_Reviews
## 1    https://imdb.com/title/tt0903747/?ref_=chttvtp_t_1 5,091 reviews
## 2    https://imdb.com/title/tt0903747/?ref_=chttvtp_t_1 5,091 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 1,056 reviews
## 8    https://imdb.com/title/tt0185906/?ref_=chttvtp_t_4 1,056 reviews
## 9    https://imdb.com/title/tt7366338/?ref_=chttvtp_t_5 3,532 reviews
## 10   https://imdb.com/title/tt7366338/?ref_=chttvtp_t_5 3,532 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   998 reviews
## 14   https://imdb.com/title/tt0417299/?ref_=chttvtp_t_7   998 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   963 reviews
## 18   https://imdb.com/title/tt0141842/?ref_=chttvtp_t_9   963 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 5,899 reviews
## 26  https://imdb.com/title/tt0944947/?ref_=chttvtp_t_13 5,899 reviews
## 27  https://imdb.com/title/tt7678620/?ref_=chttvtp_t_14   367 reviews
## 28  https://imdb.com/title/tt7678620/?ref_=chttvtp_t_14   367 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   466 reviews
## 32  https://imdb.com/title/tt1355642/?ref_=chttvtp_t_16   466 reviews
## 33  https://imdb.com/title/tt2861424/?ref_=chttvtp_t_17   909 reviews
## 34  https://imdb.com/title/tt2861424/?ref_=chttvtp_t_17   909 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   541 reviews
## 38  https://imdb.com/title/tt8420184/?ref_=chttvtp_t_19   541 reviews
## 39  https://imdb.com/title/tt0052520/?ref_=chttvtp_t_20   213 reviews
## 40  https://imdb.com/title/tt0052520/?ref_=chttvtp_t_20   213 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 1,095 reviews
## 44  https://imdb.com/title/tt1475582/?ref_=chttvtp_t_22 1,095 reviews
## 45  https://imdb.com/title/tt2560140/?ref_=chttvtp_t_23 2,359 reviews
## 46  https://imdb.com/title/tt2560140/?ref_=chttvtp_t_23 2,359 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 1,944 reviews
## 50 https://imdb.com/title/tt11126994/?ref_=chttvtp_t_25 1,944 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              53                 15
## 50              53                 15
```

```r
#2.
# Define URL for Breaking Bad
BreakingBad_urls <- "https://www.imdb.com/title/tt0903747/reviews/?ref_=tt_ov_urv"

# Initialize list to store data frames
df <- list()
df_names <- "Breaking_Bad"

# Read HTML session for the current URL
session <- read_html(BreakingBad_urls)

# Scrape reviewer names
reviewer_name <- session %>%
  html_nodes(".ipc-link.ipc-link--base") %>%
  html_text() %>%
  head(20)

# Scrape review dates
review_date <- session %>%
  html_nodes(".ipc-inline-list__item.review-date") %>%
  html_text() %>%
  head(20)

# Scrape user ratings (update CSS selector)
user_rating <- session %>%
  html_nodes(".ipc-rating-star--rating") %>%  # Example selector, verify it in the HTML
  html_text() %>%
  head(20)

# Scrape reviews' titles
review_title <- session %>%
  html_nodes(".ipc-title__text") %>%
  html_text() %>%
  head(20)

# Scrape helpful reviews
helpful_reviews <- session %>%
  html_nodes(".ipc-voting__label__count.ipc-voting__label__count--up") %>%
  html_text() %>%
  head(20)

# Scrape not helpful reviews
not_helpful_reviews <- session %>%
  html_nodes(".ipc-voting__label__count.ipc-voting__label__count--down") %>%
  html_text() %>%
  head(20)
```

```r
# Scrape text reviews
text_reviews <- session %>%
  html_nodes(".ipc-html-content-inner-div") %>%
  html_text() %>%
  head(20)

# Ensure each column has exactly 20 entries, filling with NA if fewer than 20 were scraped
reviewer_name <- c(reviewer_name, rep(NA, 20 - length(reviewer_name)))[1:20]
review_date <- c(review_date, rep(NA, 20 - length(review_date)))[1:20]
user_rating <- c(user_rating, rep(NA, 20 - length(user_rating)))[1:20]
review_title <- c(review_title, rep(NA, 20 - length(review_title)))[1:20]
helpful_reviews <- c(helpful_reviews, rep(NA, 20 - length(helpful_reviews)))[1:20]
not_helpful_reviews <- c(not_helpful_reviews, rep(NA, 20 - length(not_helpful_reviews)))[1:20]
text_reviews <- c(text_reviews, rep(NA, 20 - length(text_reviews)))[1:20]

# Create a temporary data frame for the current URL
dfTemp <- data.frame(
  reviewer_name = reviewer_name,
  review_date = review_date,
  user_rating = user_rating,
  review_title = review_title,
  helpful_reviews = helpful_reviews,
  not_helpful_reviews = not_helpful_reviews,
  text_reviews = text_reviews,
  stringsAsFactors = FALSE

)

# Append the temporary data frame to the list with a custom name
df[[df_names]] <- dfTemp

# View the data frame for "Breaking Bad"
print(df$Breaking_Bad)
```

```
##         reviewer_name  review_date user_rating
## 1             FiRE010  Jul 3, 2021          10
## 2           Permalink  Mar 6, 2019          10
## 3          bruhperson Jul 29, 2021          10
## 4           Permalink Feb 18, 2020          10
## 5        KinoKoopaKid  Nov 8, 2021          10
## 6           Permalink May 30, 2019          10
## 7         jehuschultz Nov 15, 2019          10
## 8           Permalink  Dec 8, 2022          10
## 9      Supermanfan-13 Jul 17, 2021          10
## 10          Permalink Nov 12, 2017          10
## 11  manishsingh-03299  Aug 5, 2022           6
## 12          Permalink Mar 13, 2021           5
## 13           xpinerhd  Mar 7, 2021          10
## 14          Permalink  Dec 8, 2022          10
## 15            Rob1331 Jan 11, 2014          10
## 16          Permalink  Nov 8, 2021          10
## 17 dhanushreddy-14919 Aug 11, 2021          10
## 18          Permalink May 19, 2019          10
## 19  TheLittleSongbird  May 4, 2021          10
## 20          Permalink Jun 23, 2021          10
##
## 1
## 2
## 3
## 4
## 5
## 6
## 7                                                                                         Those days a
## 8                                                                                                    I
## 9
## 10                                                                                                  A
## 11                                                                         Among the best and most ac
## 12                                                                                                  A
## 13
## 14                                                                                               Pret
## 15                                                                                               Once
## 16 If you mix Scarface, Robin Hood and maybe Tyler Durden with enough meth - you'll get a mean cockta
## 17                                                                               By far the greatest
## 18                                                                                            in a ca
## 19                                                        Since GOT is over, this is Officially the G
## 20                                                                                        Every bit a
##    helpful_reviews not_helpful_reviews
## 1             <NA>                <NA>
## 2             <NA>                <NA>
## 3             <NA>                <NA>
## 4             <NA>                <NA>
## 5             <NA>                <NA>
## 6             <NA>                <NA>
## 7             <NA>                <NA>
## 8             <NA>                <NA>
## 9             <NA>                <NA>
## 10            <NA>                <NA>
## 11            <NA>                <NA>
## 12            <NA>                <NA>
## 13            <NA>                <NA>
## 14            <NA>                <NA>
## 15            <NA>                <NA>
## 16            <NA>                <NA>
## 17            <NA>                <NA>
## 18            <NA>                <NA>
## 19            <NA>                <NA>
## 20            <NA>                <NA>
##
## 1
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10 'Breaking Bad' is one of the most popular rated shows on IMDb, is one of those rarities where eve
## 11
## 12
## 13
## 14
## 15
## 16
## 17
## 18
## 19
## 20
```

```r
# Define URL for Planet Earth II
PlanetEarthII_urls <- "https://www.imdb.com/title/tt5491994/reviews/?ref_=tt_ov_urv"

# Initialize list to store data frames
df <- list()
df_names <- "Planet_Earth_II"

# Read HTML session for the current URL
session <- read_html(PlanetEarthII_urls)

# Scrape reviewer names
reviewer_name <- session %>%
  html_nodes(".ipc-link.ipc-link--base") %>%
  html_text() %>%
  head(20)

# Scrape review dates
review_date <- session %>%
  html_nodes(".ipc-inline-list__item.review-date") %>%
  html_text() %>%
  head(20)

# Scrape user ratings (update CSS selector)
# First, inspect the correct selector for user rating from the page structure.
user_rating <- session %>%
  html_nodes(".ipc-rating-star--rating") %>%  # Adjust this selector if needed (check the page source)
  html_text() %>%
  head(20)

# Scrape reviews' titles
review_title <- session %>%
  html_nodes(".ipc-title__text") %>%
  html_text() %>%
  head(20)

# Scrape helpful reviews
helpful_reviews <- session %>%
  html_nodes(".ipc-voting__label__count.ipc-voting__label__count--up") %>%
  html_text() %>%
  head(20)

# Scrape not helpful reviews
not_helpful_reviews <- session %>%
  html_nodes(".ipc-voting__label__count.ipc-voting__label__count--down") %>%
  html_text() %>%
  head(20)

# Scrape text reviews
text_reviews <- session %>%
  html_nodes(".ipc-html-content-inner-div") %>%
  html_text() %>%
  head(20)

# Handle case where some elements might be missing, ensuring we have exactly 20 entries
reviewer_name <- c(reviewer_name, rep(NA, 20 - length(reviewer_name)))[1:20]
review_date <- c(review_date, rep(NA, 20 - length(review_date)))[1:20]
user_rating <- c(user_rating, rep(NA, 20 - length(user_rating)))[1:20]
review_title <- c(review_title, rep(NA, 20 - length(review_title)))[1:20]
helpful_reviews <- c(helpful_reviews, rep(NA, 20 - length(helpful_reviews)))[1:20]
not_helpful_reviews <- c(not_helpful_reviews, rep(NA, 20 - length(not_helpful_reviews)))[1:20]
text_reviews <- c(text_reviews, rep(NA, 20 - length(text_reviews)))[1:20]

# Create a temporary data frame for the current URL
dfTemp <- data.frame(
  reviewer_name = reviewer_name,
  review_date = review_date,
  user_rating = user_rating,
  review_title = review_title,
  helpful_reviews = helpful_reviews,
  not_helpful_reviews = not_helpful_reviews,
  text_reviews = text_reviews,
  stringsAsFactors = FALSE
)

# Append the temporary data frame to the list with a custom name
df[[df_names]] <- dfTemp

# View the data frame for "Planet Earth II"
print(df$Planet_Earth_II)
```

```
##        reviewer_name  review_date user_rating
## 1        arjanhylkema  Nov 7, 2016          10
## 2           Permalink  Nov 5, 2016          10
## 3            Wentloog  Nov 5, 2016          10
## 4           Permalink  Nov 9, 2016          10
## 5        john-m-madsen  Nov 5, 2016          10
## 6           Permalink  Nov 8, 2016          10
## 7        thespookybuz Nov 17, 2016          10
## 8           Permalink Nov 13, 2016          10
## 9         pjdickinson  Nov 6, 2016          10
## 10          Permalink Dec 31, 2016          10
## 11           dbijis33 Nov 19, 2016          10
## 12          Permalink Dec 28, 2016           7
## 13     dhanrajjughead May 19, 2019          10
## 14          Permalink Sep 29, 2017          10
## 15         NeilBarnett Nov 22, 2016          10
## 16          Permalink Oct 12, 2017          10
## 17      salmanu-27386  Dec 4, 2016          10
## 18          Permalink Oct 20, 2018          10
## 19 panagiotiskatsanos Apr 23, 2020          10
## 20          Permalink  Jan 5, 2017          10
##
## 1
## 2
## 3                                                                              At once awe-inspiring a
## 4                                                       Yet another masterpiece from BBC Nature & Davi
## 5
## 6
## 7                                                                                              Danger
## 8                                                                                 Greatest documentary
## 9                                                                           Best thing on TV since las
## 10
## 11                                                                     One of the best documentaries
## 12                                                                                 In times of climate
## 13
## 14                                                         More irritated with IMDb for the bias than
## 15                                                                            Should be required view
## 16                                                                            What a Beautiful Planet
## 17 Like the first 'Planet Earth', does for nature and our planet as 'Walking with Dinosaurs' did with
## 18                                                                          This masterpiece deserves
## 19
## 20                                                                                              Absol
##    helpful_reviews not_helpful_reviews
## 1             <NA>                <NA>
## 2             <NA>                <NA>
## 3             <NA>                <NA>
## 4             <NA>                <NA>
## 5             <NA>                <NA>
## 6             <NA>                <NA>
## 7             <NA>                <NA>
## 8             <NA>                <NA>
## 9             <NA>                <NA>
## 10            <NA>                <NA>
## 11            <NA>                <NA>
## 12            <NA>                <NA>
## 13            <NA>                <NA>
## 14            <NA>                <NA>
## 15            <NA>                <NA>
## 16            <NA>                <NA>
## 17            <NA>                <NA>
## 18            <NA>                <NA>
## 19            <NA>                <NA>
## 20            <NA>                <NA>
##
## 1
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16 Absolutely adore the first 'Planet Earth' from 2007, one of the best documentaries ever made and a
## 17
## 18
## 19
## 20
```

```
# Define URL for Planet Earth
PlanetEarth_urls <- "https://www.imdb.com/title/tt0795176/reviews/?ref_=tt_ov_urv"

# Initialize list to store data frames
df <- list()
df_names <- "Planet_Earth"

# Read HTML session for the current URL
session <- read_html(PlanetEarth_urls)

# Scrape reviewer names
reviewer_name <- session %>%
  html_nodes(".ipc-link.ipc-link--base") %>%
  html_text() %>%
  head(20)

# Scrape review dates
review_date <- session %>%
  html_nodes(".ipc-inline-list__item.review-date") %>%
  html_text() %>%
  head(20)

# Scrape user ratings (corrected CSS selector)
user_rating <- session %>%
  html_nodes(".ipc-rating-star--rating") %>%  # Adjust this selector if needed (inspect page for correc
  html_text() %>%
  head(20)

# Scrape reviews' titles
review_title <- session %>%
  html_nodes(".ipc-title__text") %>%
  html_text() %>%
  head(20)

# Scrape helpful reviews
helpful_reviews <- session %>%
  html_nodes(".ipc-voting__label__count.ipc-voting__label__count--up") %>%
  html_text() %>%
  head(20)
```

```r
# Scrape not helpful reviews
not_helpful_reviews <- session %>%
  html_nodes(".ipc-voting__label__count.ipc-voting__label__count--down") %>%
  html_text() %>%
  head(20)

# Scrape text reviews
text_reviews <- session %>%
  html_nodes(".ipc-html-content-inner-div") %>%
  html_text() %>%
  head(20)


# Handle case where some elements might be missing, ensuring we have exactly 20 entries
reviewer_name <- c(reviewer_name, rep(NA, 20 - length(reviewer_name)))[1:20]
review_date <- c(review_date, rep(NA, 20 - length(review_date)))[1:20]
user_rating <- c(user_rating, rep(NA, 20 - length(user_rating)))[1:20]
review_title <- c(review_title, rep(NA, 20 - length(review_title)))[1:20]
helpful_reviews <- c(helpful_reviews, rep(NA, 20 - length(helpful_reviews)))[1:20]
not_helpful_reviews <- c(not_helpful_reviews, rep(NA, 20 - length(not_helpful_reviews)))[1:20]
text_reviews <- c(text_reviews, rep(NA, 20 - length(text_reviews)))[1:20]

# Create a temporary data frame for the current URL
dfTemp <- data.frame(
  reviewer_name = reviewer_name,
  review_date = review_date,
  user_rating = user_rating,
  review_title = review_title,
  helpful_reviews = helpful_reviews,
  not_helpful_reviews = not_helpful_reviews,
  text_reviews = text_reviews,
  stringsAsFactors = FALSE
)

# Append the temporary data frame to the list with a custom name
df[[df_names]] <- dfTemp

# View the data frame for "Planet Earth"
print(df$Planet_Earth)
```

```
##      reviewer_name  review_date user_rating
## 1     robert-kamer  Feb 8, 2007          10
## 2        Permalink Nov 19, 2008          10
## 3         jim-1409  Jan 4, 2009          10
## 4        Permalink Dec 15, 2006          10
## 5  ccthemovieman-1  Sep 1, 2007          10
## 6        Permalink Aug 27, 2006          10
## 7         cmcoveos Apr 30, 2006          10
## 8        Permalink Jun 29, 2015           9
## 9         Loordssm Jul 20, 2006          10
## 10       Permalink Jan 28, 2009          10
## 11        ultimorn  Jun 1, 2015           7
## 12       Permalink  Oct 8, 2020           3
## 13     bob the moo  Dec 4, 2007          10
## 14       Permalink Jan 15, 2007          10
## 15           alfeu Jul 30, 2008          10
## 16       Permalink Dec 25, 2017           9
## 17         Cabrone Sep 14, 2009          10
## 18       Permalink May 31, 2020           9
## 19        berndt65 Jul 27, 2014          10
## 20       Permalink  Jan 4, 2023          10
##                                                              review_title
## 1                                                            User reviews
## 2                                                            11 out of 10
## 3                                          A masterpiece of a documentary
## 4                                                      In A Word: Amazing
## 5      The most amazing achievement in natural history TV has ever given
## 6                                                    Simply put, stunning
## 7                               An amazing trip around our beautiful planet.
## 8  A visually impressive and memorable look at the world that we live in
## 9                                     Is it real? I mean, actual footagge?
## 10                                                              Beautiful
## 11                                             Are you kidding me people?
## 12                                   It doesn't get any better than this.
## 13                                          Only 4 Eps can touch my soul!
## 14                            Should be called "BBC - Yeah, animals suck"
## 15                                           Brilliant Documentary Series
## 16                             Explanation to those low-rating reviews...
## 17                                                      Truly Astonishing
## 18                                               The Greatest Series Ever
## 19                                                                 beauty
## 20                                                Absolutely Mindblowing!
##    helpful_reviews not_helpful_reviews
## 1             <NA>                <NA>
## 2             <NA>                <NA>
## 3             <NA>                <NA>
## 4             <NA>                <NA>
## 5             <NA>                <NA>
## 6             <NA>                <NA>
## 7             <NA>                <NA>
## 8             <NA>                <NA>
## 9             <NA>                <NA>
## 10            <NA>                <NA>
## 11            <NA>                <NA>
## 12            <NA>                <NA>
## 13            <NA>                <NA>
## 14            <NA>                <NA>
## 15            <NA>                <NA>
## 16            <NA>                <NA>
## 17            <NA>                <NA>
## 18            <NA>                <NA>
## 19            <NA>                <NA>
## 20            <NA>                <NA>
##
## 1
## 2
## 3
## 4
## 5
## 6
## 7  As the influence of man expands across the globe, fewer and fewer truly untouched wilderness exis
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16
## 17
## 18
## 19
## 20
```

```r
# Define URL for Band Of Brothers
BandOfBrothers_urls <- "https://www.imdb.com/title/tt0185906/reviews/?ref_=tt_ov_urv"

# Initialize list to store data frames
df <- list()
df_names <- "Band_Of_Brothers"

# Read HTML session for the current URL
session <- read_html(BandOfBrothers_urls)

# Scrape reviewer names
reviewer_name <- session %>%
  html_nodes(".ipc-link.ipc-link--base") %>%
  html_text() %>%
  head(20)

# Scrape review dates
review_date <- session %>%
  html_nodes(".ipc-inline-list__item.review-date") %>%
  html_text() %>%
  head(20)

# Scrape user ratings (corrected CSS selector)
user_rating <- session %>%
  html_nodes(".ipc-rating-star--rating") %>%
  html_text() %>%
  head(20)

# Scrape reviews' titles
review_title <- session %>%
  html_nodes(".ipc-title__text") %>%
  html_text() %>%
  head(20)

# Scrape helpful reviews
helpful_reviews <- session %>%
  html_nodes(".ipc-voting__label__count.ipc-voting__label__count--up") %>%
  html_text() %>%
  head(20)

# Scrape not helpful reviews
not_helpful_reviews <- session %>%
  html_nodes(".ipc-voting__label__count.ipc-voting__label__count--down") %>%
  html_text() %>%
  head(20)

# Scrape text reviews
text_reviews <- session %>%
  html_nodes(".ipc-html-content-inner-div") %>%
  html_text() %>%
  head(20)


# Handle case where some elements might be missing, ensuring we have exactly 20 entries
reviewer_name <- c(reviewer_name, rep(NA, 20 - length(reviewer_name)))[1:20]
review_date <- c(review_date, rep(NA, 20 - length(review_date)))[1:20]
user_rating <- c(user_rating, rep(NA, 20 - length(user_rating)))[1:20]
review_title <- c(review_title, rep(NA, 20 - length(review_title)))[1:20]
helpful_reviews <- c(helpful_reviews, rep(NA, 20 - length(helpful_reviews)))[1:20]
not_helpful_reviews <- c(not_helpful_reviews, rep(NA, 20 - length(not_helpful_reviews)))[1:20]
text_reviews <- c(text_reviews, rep(NA, 20 - length(text_reviews)))[1:20]

# Create a temporary data frame for the current URL
dfTemp <- data.frame(
  reviewer_name = reviewer_name,
  review_date = review_date,
  user_rating = user_rating,
  review_title = review_title,
  helpful_reviews = helpful_reviews,
  not_helpful_reviews = not_helpful_reviews,
  text_reviews = text_reviews,
  stringsAsFactors = FALSE
)

# Append the temporary data frame to the list with a custom name
df[[df_names]] <- dfTemp

# View the data frame for "band of brothers"
print(df$Band_Of_Brothers)
```

```
##         reviewer_name  review_date user_rating
## 1             Rob1331 Sep 27, 2022          10
## 2           Permalink Oct 14, 2001          10
## 3        sanderson777 Jan 18, 2002          10
## 4           Permalink Apr 18, 2004          10
## 5         wildcatt268 Feb 13, 2003          10
## 6           Permalink Jan 23, 2005          10
## 7             arjay24 Sep 16, 2004          10
## 8           Permalink  May 6, 2022          10
## 9           rbverhoef  Nov 4, 2019          10
## 10          Permalink  Nov 5, 2001          10
## 11         yodaschoda Aug 25, 2004         10
## 12          Permalink May 30, 2015          7
## 13 philip_vanderveken Apr 10, 2021          5
## 14          Permalink  May 2, 2006         10
## 15     Supermanfan-13  Jun 3, 2019         10
## 16          Permalink Jan 26, 2005         10
## 17          thiagoutp  May 3, 2022         10
## 18          Permalink Oct 24, 2018          9
## 19         bsmith5552  Dec 7, 2002         10
## 20          Permalink Nov 25, 2002         10
##                                                           review_title
## 1                                                         User reviews
## 2                                                         Incredible!!
## 3                            Possibly the finest 10 hours ever created
## 4                               One of the best war movies/series ever
## 5                                                            Realistic
## 6                                                            Excellent
## 7                      One of, if not the best, mini series' ever made
## 8              This series is so unbelievably realistic, so authentic.
## 9                            One of the best mini-series ever created!
## 10                                              Probably the best ever
## 11                            Realistic WWII Drama With Warts Included
## 12                                                      war, no frills
## 13                                             You can't beat this....
## 14                                                         Overrated??
## 15                                           Not very realistic at all
## 16                   Without Doubt, the Best Mini-Series Ever Recorded
## 17                                                    Great Miniseries
## 18 A series like this won't be made again (see below), so treasure it
## 19                                            Share With Your Children
## 20                                               Best Mini series ever
##    helpful_reviews not_helpful_reviews
## 1             <NA>                <NA>
## 2             <NA>                <NA>
## 3             <NA>                <NA>
## 4             <NA>                <NA>
## 5             <NA>                <NA>
## 6             <NA>                <NA>
## 7             <NA>                <NA>
## 8             <NA>                <NA>
## 9             <NA>                <NA>
## 10            <NA>                <NA>
## 11            <NA>                <NA>
## 12            <NA>                <NA>
## 13            <NA>                <NA>
## 14            <NA>                <NA>
## 15            <NA>                <NA>
## 16            <NA>                <NA>
## 17            <NA>                <NA>
## 18            <NA>                <NA>
## 19            <NA>                <NA>
## 20            <NA>                <NA>
##
## 1
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14 Lots of people applaud this series for its realism, but I can't really agree. I think there is st
## 15
## 16
## 17
## 18
## 19
## 20
```

```r
# Define URL for Chernobyl
Chernobyl_urls <- "https://www.imdb.com/title/tt7366338/reviews/?ref_=tt_ov_urv"

# Initialize list to store data frames
df <- list()
df_names <- "Chernobyl"

# Read HTML session for the current URL
session <- read_html(Chernobyl_urls)

# Scrape reviewer names
reviewer_name <- session %>%
  html_nodes(".ipc-link.ipc-link--base") %>%
  html_text() %>%
  head(20)

# Scrape review dates
review_date <- session %>%
  html_nodes(".ipc-inline-list__item.review-date") %>%
  html_text() %>%
  head(20)

# Scrape user ratings (corrected CSS selector)
user_rating <- session %>%
  html_nodes(".ipc-rating-star--rating") %>%
  html_text() %>%
  head(20)

# Scrape reviews' titles
review_title <- session %>%
  html_nodes(".ipc-title__text") %>%
  html_text() %>%
  head(20)
```

```r
# Scrape helpful reviews
helpful_reviews <- session %>%
  html_nodes(".ipc-voting__label__count.ipc-voting__label__count--up") %>%
  html_text() %>%
  head(20)

# Scrape not helpful reviews
not_helpful_reviews <- session %>%
  html_nodes(".ipc-voting__label__count.ipc-voting__label__count--down") %>%
  html_text() %>%
  head(20)

# Scrape text reviews
text_reviews <- session %>%
  html_nodes(".ipc-html-content-inner-div") %>%
  html_text() %>%
  head(20)

# Handle case where some elements might be missing, ensuring we have exactly 20 entries
reviewer_name <- c(reviewer_name, rep(NA, 20 - length(reviewer_name)))[1:20]
review_date <- c(review_date, rep(NA, 20 - length(review_date)))[1:20]
user_rating <- c(user_rating, rep(NA, 20 - length(user_rating)))[1:20]
review_title <- c(review_title, rep(NA, 20 - length(review_title)))[1:20]
helpful_reviews <- c(helpful_reviews, rep(NA, 20 - length(helpful_reviews)))[1:20]
not_helpful_reviews <- c(not_helpful_reviews, rep(NA, 20 - length(not_helpful_reviews)))[1:20]
text_reviews <- c(text_reviews, rep(NA, 20 - length(text_reviews)))[1:20]

# Create a temporary data frame for the current URL
dfTemp <- data.frame(
  reviewer_name = reviewer_name,
  review_date = review_date,
  user_rating = user_rating,
  review_title = review_title,
  helpful_reviews = helpful_reviews,
  not_helpful_reviews = not_helpful_reviews,
  text_reviews = text_reviews,
  stringsAsFactors = FALSE
)

# Append the temporary data frame to the list with a custom name
df[[df_names]] <- dfTemp

# View the data frame for "Chernobyl"
print(df$Chernobyl)
```

```
##        reviewer_name  review_date user_rating
## 1    curiosityonmars May 23, 2019          10
## 2          Permalink May 10, 2019          10
## 3           stelmakh  May 9, 2019          10
## 4          Permalink May 14, 2019          10
## 5       natashapekar  May 7, 2019          10
## 6          Permalink May 20, 2019          10
## 7         m-porpaczi  May 6, 2019          10
## 8          Permalink May 13, 2019          10
## 9           Lladerat  May 6, 2019          10
## 10         Permalink Nov 27, 2019          10
## 11          jfirebug May 23, 2019           7
## 12         Permalink Jan 27, 2024           1
## 13           thegldt Jun 29, 2019           8
## 14         Permalink May 20, 2019          10
## 15 alexander-phoenix May 30, 2019          10
## 16         Permalink  Jun 7, 2019          10
## 17       wmeduardowm  May 6, 2019           9
## 18         Permalink Sep 27, 2022           9
## 19    Leofwine_draca May 26, 2019           9
## 20         Permalink Jul 10, 2022          10
##                                                   review_title helpful_reviews
## 1                                                 User reviews            <NA>
## 2                                            They got it right            <NA>
## 3                                         Goosebumps and tears            <NA>
## 4                                I highly recommend this film!            <NA>
## 5                              No hero wakes up wanting to die            <NA>
## 6                                       So far looks excellent            <NA>
## 7                                                   Incredible            <NA>
## 8                   Bleak, Unsettling, Haunting All Throughout            <NA>
## 9                                                 Unbelievable            <NA>
## 10                                           HBO did it again!            <NA>
## 11                                                   Exemplary            <NA>
## 12                                                    Amazing!            <NA>
## 13            Unveiling Human Errors and Political Shadows            <NA>
## 14                                           How cost the lie?            <NA>
## 15                                      Emotionally drained...            <NA>
## 16                                           Just watch it (!)            <NA>
## 17                     Now you look like the minister of coal!            <NA>
## 18                                                   Cracking.            <NA>
## 19                                                 Must Watch!            <NA>
## 20 It is hard to overestimate the importance of this show.            <NA>
##    not_helpful_reviews
## 1                 <NA>
## 2                 <NA>
## 3                 <NA>
## 4                 <NA>
## 5                 <NA>
## 6                 <NA>
## 7                 <NA>
## 8                 <NA>
## 9                 <NA>
## 10                <NA>
## 11                <NA>
## 12                <NA>
## 13                <NA>
## 14                <NA>
## 15                <NA>
## 16                <NA>
## 17                <NA>
## 18                <NA>
## 19                <NA>
## 20                <NA>
```

```
##
## 1
## 2
## 3
## 4  As my mother tells it, the weather was quite nice, the sky was clear without any sign of clouds in
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16
## 17
## 18
## 19
## 20
```

```
#3.

# Convert the 'Year' column to numeric if it isn't already
top_tv_shows$Year <- as.numeric(top_tv_shows$Year)
```

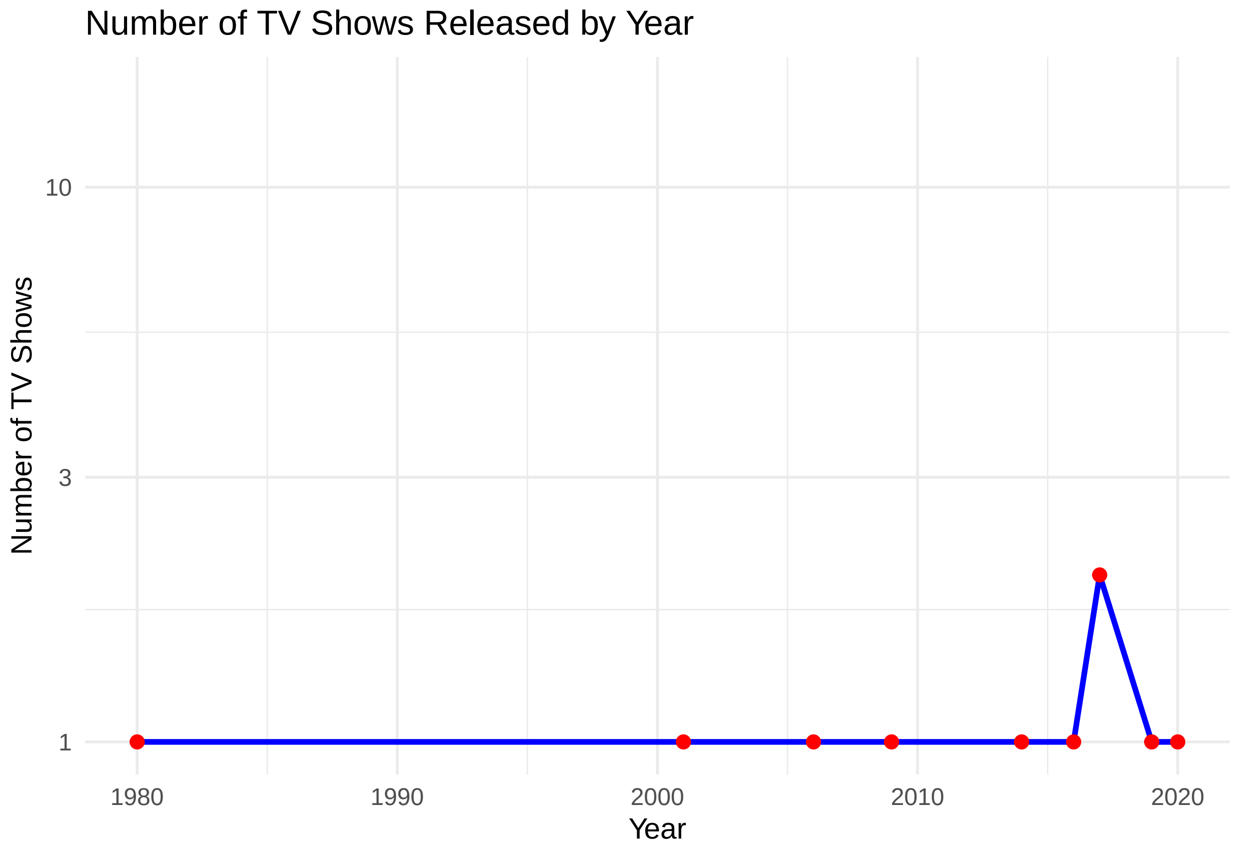
```
## Warning: NAs introduced by coercion
```

```
# Group the data by Year and count the number of shows per year
shows_by_year <- top_tv_shows %>%
  group_by(Year) %>%
  summarise(Count = n())

# Plot the number of shows released by year
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") +
  scale_y_log10() +  # Use log scale for y-axis
  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.
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_line()`).
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```



```
# Find the year with the most TV shows released
most_shows_year <- shows_by_year %>%
  filter(Count == max(Count))

# Print the year with the most releases
print(most_shows_year)
```

```
## # A tibble: 1 x 2
##    Year Count
##   <dbl> <int>
## 1    NA    15
```

2. Extracting Amazon Product Reviews

```
#4. URLs
urls <- c('https://www.amazon.com/s?k=backpacks&crid=35ZQ1H72MC3G9&sprefix=backpacks%2Caps%2C590&ref=nb_
          'https://www.amazon.com/s?k=laptops&crid=L7MQBW7MD4SX&sprefix=laptopb%2Caps%2C1304&ref=nb_sb_
          'https://www.amazon.com/s?k=phone+case&dc&crid=1VPDCJ87S93TL&sprefix=phone+cas%2Caps%2C451&re
          'https://www.amazon.com/s?k=mountain+bike&crid=1ZQR71S8XHZN6&sprefix=mountain+bik%2Caps%2C499
          'https://www.amazon.com/s?k=tshirt&crid=2RQIP7MP6IYAW&sprefix=tshirt%2Caps%2C443&ref=nb_sb_no
```

```
#5
df <- list()

for (i in seq_along(urls)) {

  session <- bow(urls[i], user_agent = "Educational")

  product_name <- scrape(session) %>% html_nodes('h2.a-size-mini') %>% html_text() %>% head(30)
  product_description <- scrape(session) %>% html_nodes('div.productDescription') %>% html_text() %>% he


  product_rating <- scrape(session) %>% html_nodes('span.a-icon-alt') %>% html_text() %>% head(30)
  ratings <- as.numeric(str_extract(product_rating, "\\d+\\.\\d"))


  product_price <- scrape(session) %>% html_nodes('span.a-price') %>%  html_text() %>% head(30)
  price <- as.numeric(str_extract(product_price, "\\d+\\.\\d+"))


  product_review <- scrape(session) %>% html_nodes('div.review-text-content') %>% html_text() %>% head(3


  dfTemp <- data.frame(Product_Name = product_name[1:30], Description = product_description[1:30], Ratin

  df[[i]] <- dfTemp
}

print(df[[1]])
```

```
##
## 1                                       JanSport SuperBreak One Backpacks - Durable, Lightweight Bo
## 2  MATEIN Travel Laptop Backpack, Business Anti Theft Slim Sturdy Laptops Backpack with USB Charging
## 3    Taygeer Travel Backpack for Women, Carry On Backpack with USB Charging Port & Shoe Pouch, TSA 1
## 4
## 5       YOREPEK Travel Backpack, Extra Large 50L Laptop Backpacks for Men Women, Water Resistant Col
## 6   Lapsouno Travel Backpack, Large Carry on Backpack, 17 Inch Laptop Backpack, Big Backpack, Extra
## 7
## 8         Laptop Backpack,Business Travel Anti Theft Slim Durable Laptops Backpack with USB Charging
## 9                                                                                               Z
## 10 LOVEVOOK Laptop Backpack for Women, 15.6 Inch Work Business Backpacks Purse with USB Port, Large
## 11
## 12                                                        MIYCOO Backpack - Ultra Lightweight Pack
## 13                                                                  Amazon Basics Transparent Scho
## 14
## 15                                                     JanSport Right Pack Backpack - Durable Dayp
## 16
## 17
## 18
## 19
## 20
## 21
## 22
## 23
## 24
## 25
## 26
## 27
## 28
## 29
## 30
##    Description Rating Price
## 1         <NA>    4.5 31.99
## 2         <NA>    4.3 38.00
## 3         <NA>    4.6 21.99
## 4         <NA>    4.7 39.96
## 5         <NA>    4.7 23.99
## 6         <NA>    4.6 23.99
## 7         <NA>    4.7 27.99
## 8         <NA>    4.5 39.99
## 9         <NA>    3.7 23.98
## 10        <NA>    4.6 49.99
## 11        <NA>    4.6  8.99
## 12        <NA>    4.8 12.00
## 13        <NA>    4.8 29.99
## 14        <NA>    4.8 23.99
## 15        <NA>    4.6 29.98
## 16        <NA>    4.7 99.00
## 17        <NA>    4.7 16.99
## 18        <NA>    4.7 16.49
## 19        <NA>     NA 65.00
## 20        <NA>     NA 44.00
## 21        <NA>     NA 64.99
## 22        <NA>     NA 80.00
## 23        <NA>     NA    NA
## 24        <NA>     NA    NA
## 25        <NA>     NA    NA
## 26        <NA>     NA    NA
## 27        <NA>     NA    NA
## 28        <NA>     NA    NA
## 29        <NA>     NA    NA
## 30        <NA>     NA    NA
```

```r
print(df[[2]])
```

```
##
## 1    Acer Aspire 3 A315-24P-R7VH Slim Laptop | 15.6" Full HD IPS Display | AMD Ryzen 3 7320U Quad-Cor
## 2             HP Newest 255 G10 Laptop for Home or Work, 16GB RAM, 1TB SSD, 15.6" Full HD, Ryzen 3 73
## 3       HP 14 Laptop, Intel Celeron N4020, 4 GB RAM, 64 GB Storage, 14-inch Micro-edge HD Display, Wi
## 4
## 5                                                                   HP 17 Laptop, 17.3" HD+ Display, 11
## 6  HP 2024 Newest Laptop for Business and Student, 15.6" HD Touchscreen, Intel 6-Core i3-1215U Proces
## 7          Apple 2024 MacBook Air 13-inch Laptop with M3 chip: Built for Apple Intelligence, 13.6-inc
## 8                              HP 15.6 FHD Display G9 Laptop • 32GB RAM • 1TB Storage (512GB SSD & 50
## 9  HP Newest Pavilion 15.6" Touchscreen Laptop with 12 Months Microsoft • 40GB RAM • 1TB Storage (51
## 10                        HP Portable Laptop, Student and Business, 14" HD Display, Intel Quad-Core N
## 11             HP 15.6" Portable Laptop (Include 1 Year Microsoft 365), HD Display, Intel Quad-Core
## 12          HP Newest 14" Ultral Light Laptop for Students and Business, Intel Quad-Core N4120, 8GB
## 13                                              ASUS Lightweight 15.5" Full HD Laptop, Windows 11 Home O
## 14    Lenovo IdeaPad 1 Student Laptop, Intel Dual Core Processor, 20GB RAM, 1TB SSD + 128GB eMMC, 15
## 15                              acer Gateway Chromebook 311 CBO311-1H-C1MX Laptop | Intel Celeron N4500
## 16    HP Newest 14" LED Business Laptop Computer, 16GB RAM 320GB Storage (64GB eMMC+256GB SD Card),
## 17
## 18
## 19
## 20
## 21
## 22
## 23
## 24
## 25
## 26
## 27
## 28
## 29
## 30
##    Description Rating  Price
## 1         <NA>    4.2 279.99
## 2         <NA>     NA 321.99
## 3         <NA>    4.3 448.88
## 4         <NA>    4.4 599.00
## 5         <NA>    4.4 176.00
## 6         <NA>    4.0 209.99
## 7         <NA>    4.1 146.97
## 8         <NA>    4.2 475.00
## 9         <NA>    4.3 399.00
## 10        <NA>    4.7 422.00
## 11        <NA>    4.4 899.00
## 12        <NA>    4.8  99.00
## 13        <NA>    4.1 419.58
## 14        <NA>    4.1 499.00
## 15        <NA>    4.0 599.00
## 16        <NA>    4.1 204.99
## 17        <NA>    4.1 299.00
## 18        <NA>    4.5 279.90
## 19        <NA>    4.0 249.99
## 20        <NA>     NA 169.99
## 21        <NA>     NA 399.00
## 22        <NA>     NA 167.99
## 23        <NA>     NA 199.99
## 24        <NA>     NA 299.99
## 25        <NA>     NA 399.99
## 26        <NA>     NA     NA
## 27        <NA>     NA     NA
## 28        <NA>     NA     NA
## 29        <NA>     NA     NA
## 30        <NA>     NA     NA
```

```r
print(df[[3]])
```

```
##
## 1                         ESR for iPhone 14 Case/iPhone 13 Case, Compatible with MagSafe, Shockproof
## 2  BENTOBEN Magnetic for iPhone 13 Case & iPhone 14 Case [Compatible with Magsafe] Translucent Matte
## 3                                                                            OtterBox iPhone 15, iPho
## 4                                                                     ORNARTO Compatible with iPhone
## 5
## 6                                                                                          OtterBox i
## 7                              Temdan for iPhone 16 Pro Case Clear, [Compatible with Magsafe][Anti-Ye
## 8                                                                                               Otter
## 9                                                                 FABSPARK Phone Case for iPhone 13 C
## 10                                                                                     OtterBox iPhon
## 11                                                                         OtterBox iPhone 13 Pro Ma
## 12                           elago Compatible with iPhone 14 Pro Case, Liquid Silicone Case, Full Bo
## 13                               elago Compatible with iPhone 14 Case, Liquid Silicone Case, Full Bo
## 14                       elago Compatible with iPhone 14 Pro Max Case, Liquid Silicone Case, Full Bo
## 15               elago Liquid Silicone Case Compatible with iPhone 13 Pro Case (6.1"), Premium Silic
## 16
## 17
## 18
## 19
## 20
## 21
## 22
## 23
## 24
## 25
## 26
## 27
## 28
## 29
## 30
##    Description Rating Price
## 1         <NA>    3.5 14.99
## 2         <NA>    3.5 30.99
## 3         <NA>    3.5 12.98
## 4         <NA>    4.6 34.90
## 5         <NA>    4.5 39.95
## 6         <NA>    4.6  9.99
## 7         <NA>    4.5 20.99
## 8         <NA>    3.9 44.95
## 9         <NA>    4.7 34.63
## 10        <NA>    4.6 39.95
## 11        <NA>    4.7  8.99
## 12        <NA>    4.3 29.95
## 13        <NA>    4.7  9.99
## 14        <NA>    4.6 19.95
## 15        <NA>    4.4 39.95
## 16        <NA>    4.5 29.95
## 17        <NA>    4.5 39.95
## 18        <NA>    4.4 12.99
## 19        <NA>    4.7 12.99
## 20        <NA>     NA 12.99
## 21        <NA>     NA 12.99
## 22        <NA>     NA 31.95
## 23        <NA>     NA 39.95
## 24        <NA>     NA    NA
## 25        <NA>     NA    NA
## 26        <NA>     NA    NA
## 27        <NA>     NA    NA
## 28        <NA>     NA    NA
## 29        <NA>     NA    NA
## 30        <NA>     NA    NA
```

```
print(df[[4]])
```

```
##
```

```
## 1  WEIZE Mountain Bike, 24/26/27.5 inch Outdoor Cycling Bike,18-Speed/High-Carbon Steel/Dual Full Sus
## 2                          Racer Electric Bike for Adults - 21-Speed Mountain Lightweight Ebike with
## 3                                          Mongoose Flatrock 21-Speed Hardtail Mountain Bike, 24 to 2
## 4   Schwinn Traxion Mountain Bike for Adult Men Women, 29-Inch Wheels, Full Suspension, 24-Speed Shi
## 5                                   Dynacraft Magna Echo Ridge Mountain Bike - Rugged and Durable Des
## 6                             Mongoose Malus Mens and Women Fat Tire Mountain Bike, 26-Inch Bicycle W
## 7                                          Schwinn High Timber Mountain Bike for Adult Youth Men Women
## 8                       Huffy Stone Mountain Hardtail Mountain Bike for Boys/Girls/Men/Women, 20"/24"/2
## 9                      Outroad 26 Inch Mountain Bike, 21-Speed/High-Carbon Steel/Aviation Grade Frame,
## 10                                                   Mountain Bike 24/26/27.5 Inch, Dual Full Suspens
## 11                                                                       Schwinn Bonafide Men and Wo
## 12    isinwheel M10 Electric Bike Adult 500W, 26" Commuting Electric Mountain Bike 20MPH Max Range 5
## 13                                                Dynacraft Vertical Alpine Eagle Mountain Bike - Ru
## 14                            Grafton Mountain Bike for Adult and Youth Men and Women, 24/26 / 27.5-
## 15                                      26/27.5" Mountain Bike 21 Speed Bikes for Adults, Men & Women
## 16                                                        Ktaxon Mountain Bike 26/27.5/29 Inch Men &
## 17
## 18
## 19
## 20
## 21
## 22
## 23
## 24
## 25
## 26
## 27
## 28
## 29
## 30
##    Description Rating  Price
## 1         <NA>    4.0 179.99
## 2         <NA>    4.3 199.99
## 3         <NA>    4.1 199.99
## 4         <NA>    4.3 309.99
## 5         <NA>    4.0 629.99
## 6         <NA>    4.4 674.99
## 7         <NA>    4.1 158.94
## 8         <NA>    3.7 169.99
## 9         <NA>    3.9 492.99
## 10        <NA>    4.3 519.99
## 11        <NA>    4.2 499.99
## 12        <NA>    4.3 229.99
## 13        <NA>    4.3  99.98
## 14        <NA>    3.5 179.99
## 15        <NA>    2.5 612.53
## 16        <NA>    4.2 389.99
## 17        <NA>     NA 349.99
## 18        <NA>     NA  99.99
## 19        <NA>     NA 199.99
## 20        <NA>     NA 199.99
## 21        <NA>     NA     NA
## 22        <NA>     NA     NA
## 23        <NA>     NA     NA
## 24        <NA>     NA     NA
## 25        <NA>     NA     NA
## 26        <NA>     NA     NA
## 27        <NA>     NA     NA
## 28        <NA>     NA     NA
## 29        <NA>     NA     NA
## 30        <NA>     NA     NA
```

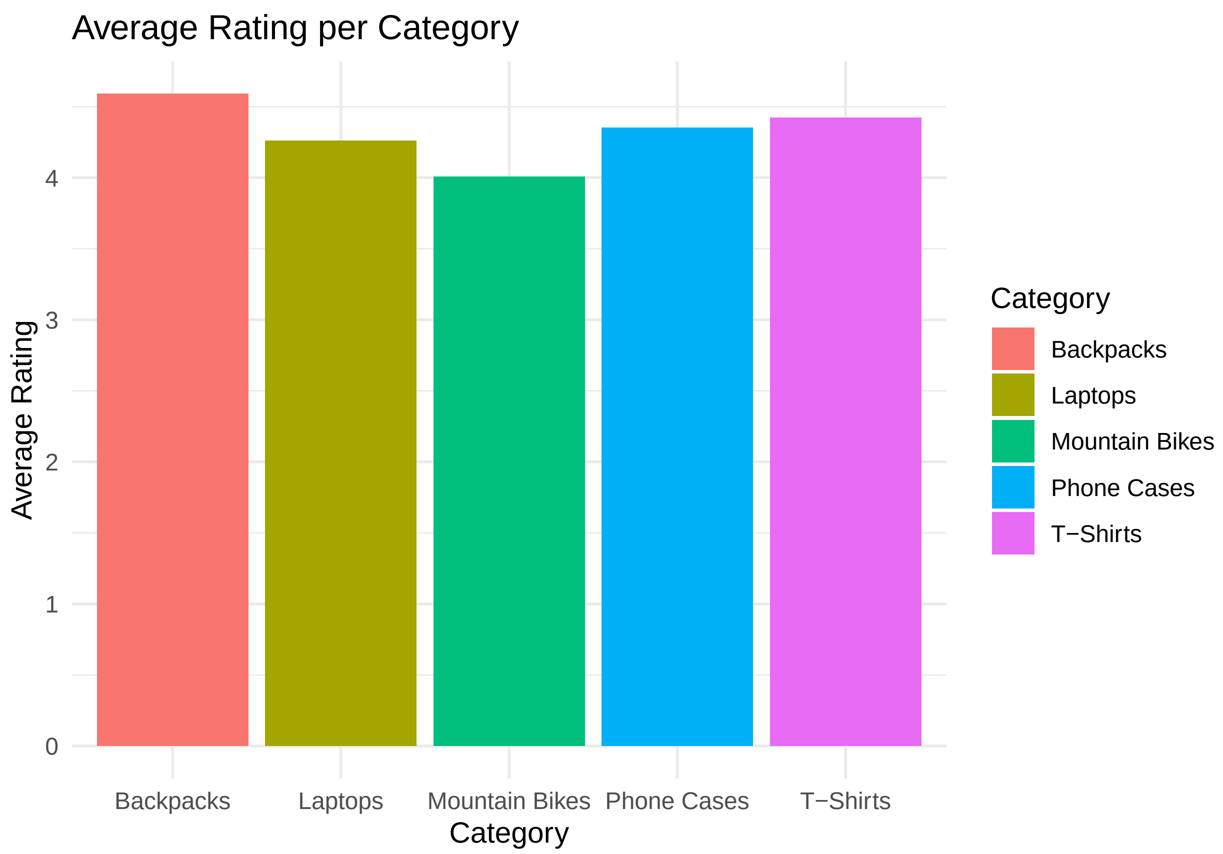
```r
print(df[[5]])
```

```
## 
## 1
## 2                       Men's Pocket Undershirt Pack, Cotton Crew Neck T-Shirt, Moisture Wicking Tee
## 3
## 4                        Men's T-Shirt, Beefy-T Heavyweight Cotton Crewneck Tee, 1 or 2 Pack, Availab
## 5
## 6                                                                      Men's Eversoft Cotton Stay Tuc
## 7
## 8                                                                    Men's Eversoft Cotton Stay Tucke
## 9
## 10                                                          Unisex Adult Ultra Cotton T-Shirt, Style
## 11
## 12                                                              Adult Heavyweight Short Sleeve Tee, S
## 13
## 14                                   Men's Short Sleeve T-Shirt Pack, Essentials Crewneck Cotton T-Sh
## 15
## 16                                                                      Men's V-Neck T-Shirts, Multi
## 17
## 18                                                            Men's Loose Fit Heavyweight Short-Slee
## 19
## 20                                                              3 Pack, Men's Short Sleeve Crew Neck
## 21
## 22                                                          Unisex Adult Heavy Cotton T-Shirt, Style
## 23
## 24                        Men's Eversoft Cotton T Shirts, Breathable & Moisture Wicking with Odor Con
## 25
## 26 3 Pcs Men's Oversized Heavy Cotton Summer T-Shirts Vintage Tee Loose Fit Short Sleeve Casual Tshi
## 27
## 28                                                              1 Pack, Men's Short Sleeve Crew Neck
## 29
## 30                        5 Pack Men's Dry Fit T Shirts, Athletic Running Gym Workout Short Sleeve T
##    Description Rating Price
## 1         <NA>    4.4 21.98
## 2         <NA>    4.5  6.70
## 3         <NA>    4.6 12.00
## 4         <NA>    4.6 21.48
## 5         <NA>    4.5 17.42
## 6         <NA>    4.5 18.49
## 7         <NA>    4.4 14.46
## 8         <NA>    4.5  9.65
## 9         <NA>    4.6 14.99
## 10        <NA>    4.1 17.58
## 11        <NA>    4.5 26.00
## 12        <NA>    4.5 19.79
## 13        <NA>    4.4 21.99
## 14        <NA>    4.0 19.99
## 15        <NA>    4.3 56.99
## 16        <NA>    4.3 89.97
## 17        <NA>    4.2  9.89
## 18        <NA>    4.5 10.99
## 19        <NA>    4.5 11.56
## 20        <NA>    4.6 16.99
## 21        <NA>    4.3 35.99
## 22        <NA>    4.5 38.99
## 23        <NA>    4.5 29.99
## 24        <NA>    4.5 26.99
## 25        <NA>    4.4 14.99
## 26        <NA>    4.4 14.99
## 27        <NA>    4.4 10.64
## 28        <NA>    4.3 14.49
## 29        <NA>    4.4 30.34
## 30        <NA>    4.4 34.50
```

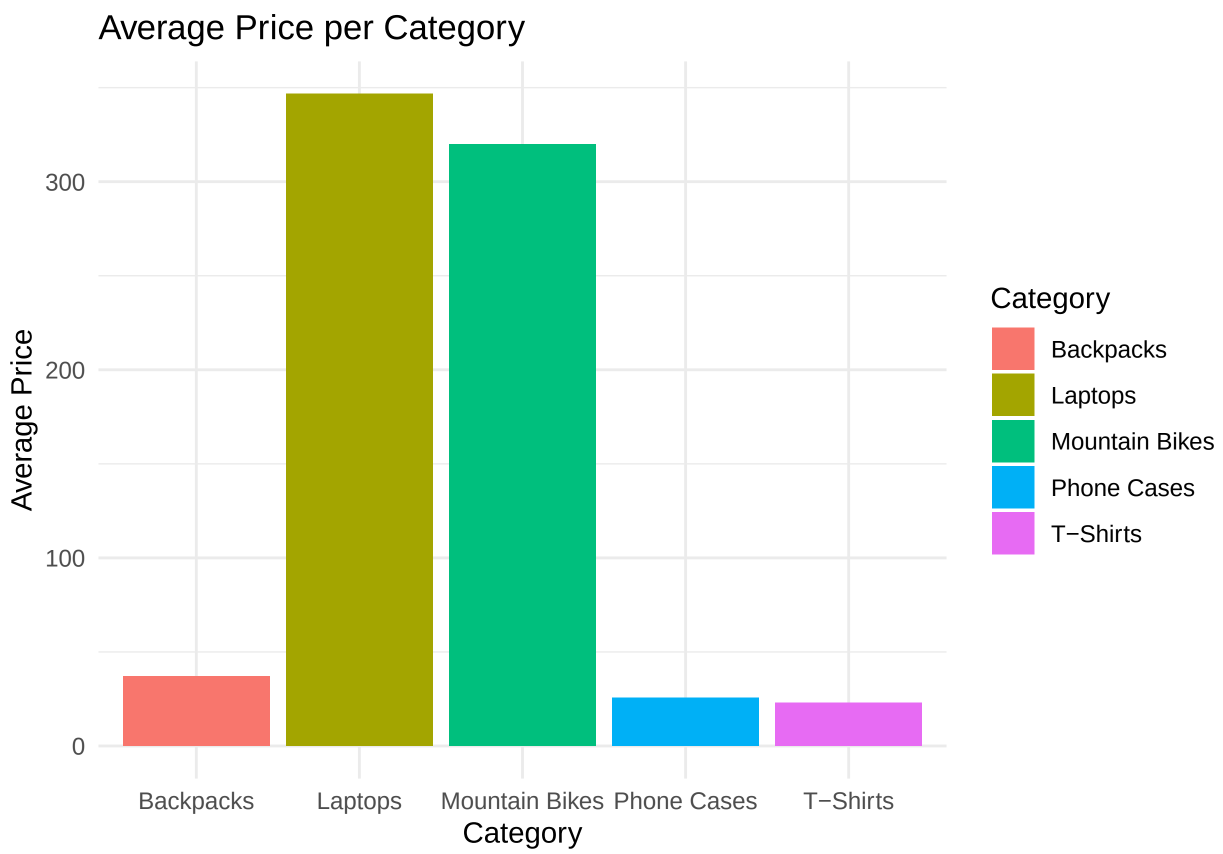
```
#6.

#The code extracts data from Amazon product listing pages based on different search queries, such as "b
```

```
#7

#This data can be used to compare product popularity, analyze price trends, examine the relationship be
```

```
#8
combined_df <- do.call(rbind, df)
combined_df$Category <- rep(c("Backpacks", "Laptops", "Phone Cases", "Mountain Bikes", "T-Shirts"), each

avg_rating <- combined_df %>%
  group_by(Category) %>%
  summarize(Average_Rating = mean(Rating, na.rm = TRUE))

ggplot(avg_rating, aes(x = Category, y = Average_Rating, fill = Category)) +
  geom_bar(stat = "identity") +
  labs(title = "Average Rating per Category", x = "Category", y = "Average Rating") +
  theme_minimal()
```

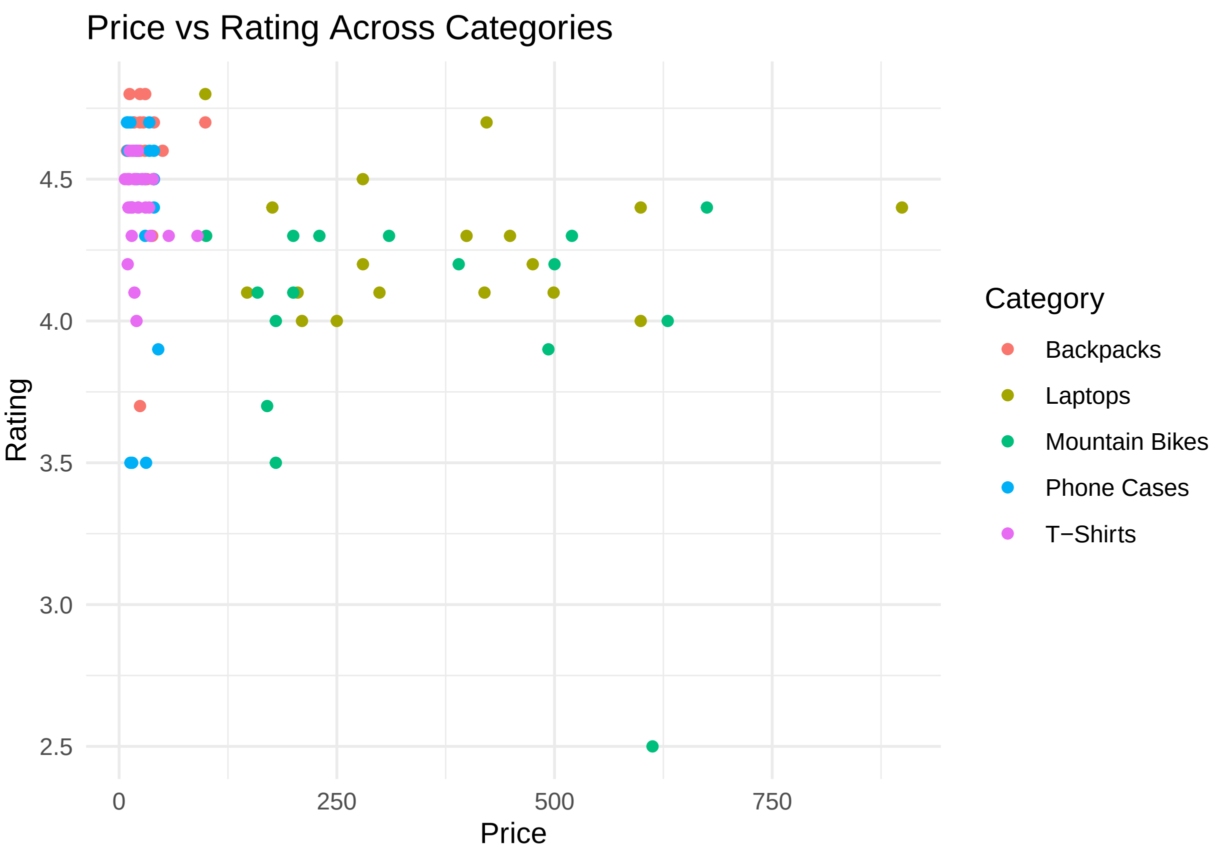


```r
avg_price <- combined_df %>%
  group_by(Category) %>%
  summarize(Average_Price = mean(Price, na.rm = TRUE))

ggplot(avg_price, aes(x = Category, y = Average_Price, fill = Category)) +
  geom_bar(stat = "identity") +
  labs(title = "Average Price per Category", x = "Category", y = "Average Price") +
  theme_minimal()
```



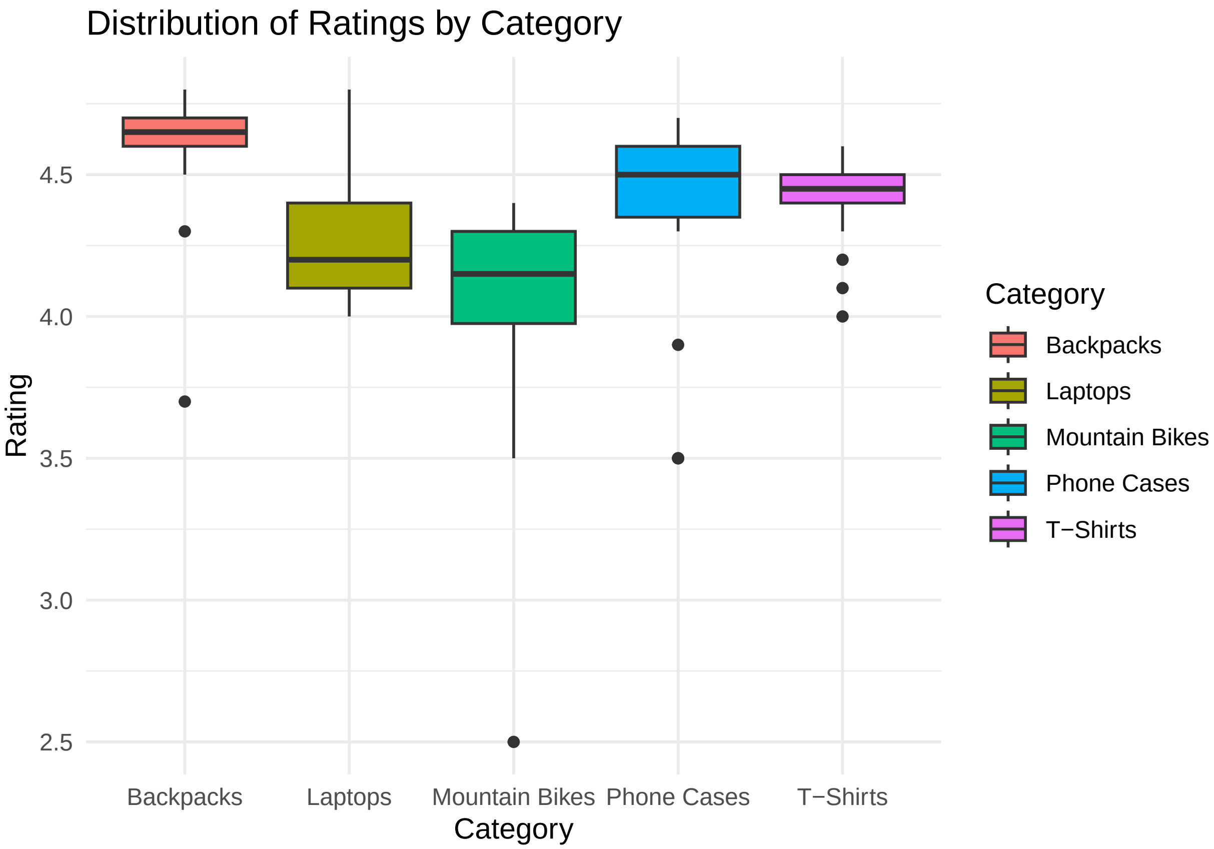
```
ggplot(combined_df, aes(x = Price, y = Rating, color = Category)) +
  geom_point() +
  labs(title = "Price vs Rating Across Categories", x = "Price", y = "Rating") +
  theme_minimal()
```

```
## Warning: Removed 49 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



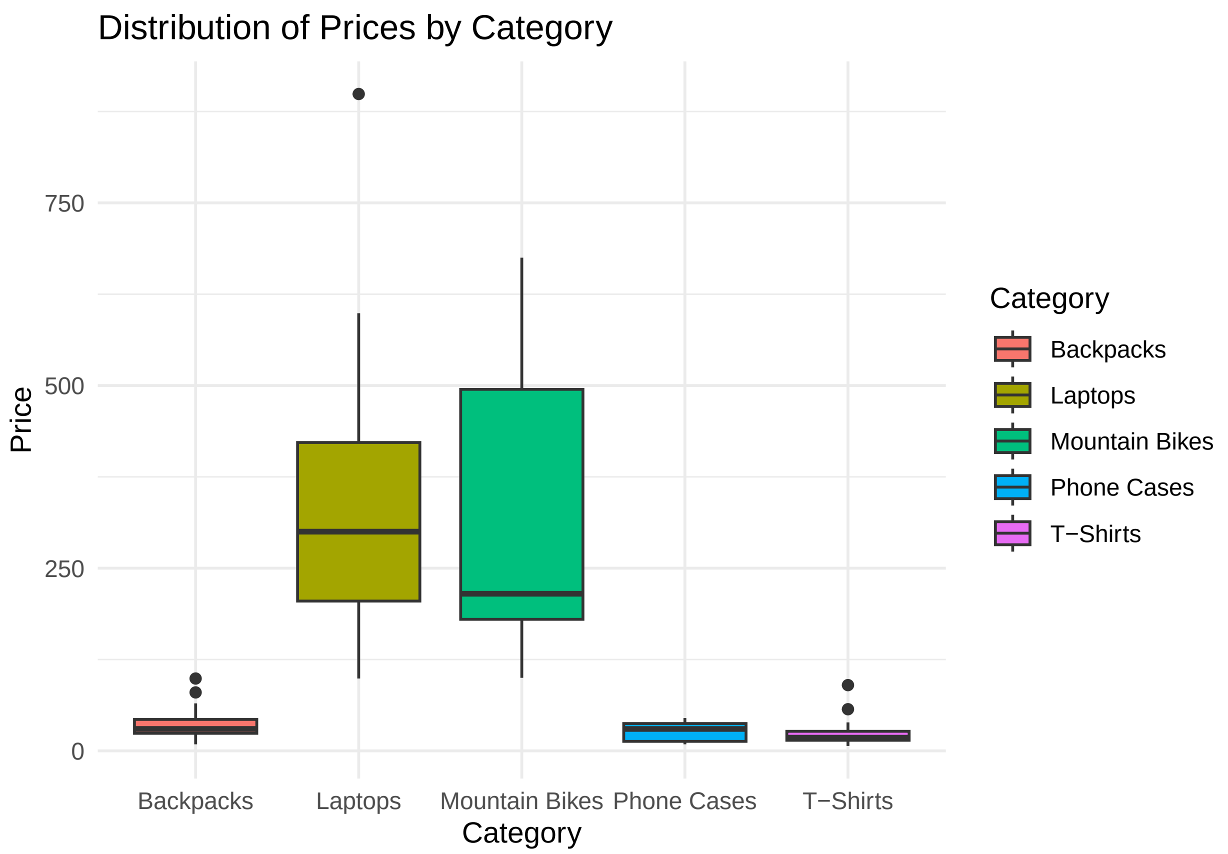
```r
#9
ggplot(combined_df, aes(x = Category, y = Rating, fill = Category)) +
  geom_boxplot() +
  labs(title = "Distribution of Ratings by Category", x = "Category", y = "Rating") +
  theme_minimal()
```

```
## Warning: Removed 49 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



```
ggplot(combined_df, aes(x = Category, y = Price, fill = Category)) +
  geom_boxplot() +
  labs(title = "Distribution of Prices by Category", x = "Category", y = "Price") +
  theme_minimal()
```

```
## Warning: Removed 30 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



```
#10
ranked_data <- lapply(df, function(df_category) {
  df_category %>%
    arrange(desc(Rating), Price) %>%
    mutate(Rank = row_number()) %>%
    select(Rank, everything())
})

categories <- c("Backpacks", "Laptops", "Phone Cases", "Mountain Bikes", "T-Shirts")
for (i in seq_along(ranked_data)) {
  ranked_data[[i]]$Category <- categories[i]
}

ranked_combined_df <- do.call(rbind, ranked_data)
ranked_combined_df <- ranked_combined_df %>%
  arrange(Category, Rank) %>%
  group_by(Category) %>%
  slice(1:5)

print(ranked_combined_df)
```

```
## # A tibble: 25 x 6
## # Groups:   Category [5]
##     Rank Product_Name                           Description Rating Price Category
##    <int> <chr>                                  <chr>        <dbl> <dbl> <chr>
##  1     1 "MIYCOO Backpack - Ultra Lightweight~ <NA>           4.8  12   Backpac~
##  2     2 "THE NORTH FACE Vault Everyday Lapto~ <NA>           4.8  24.0 Backpac~
##  3     3 "Amazon Basics Transparent School Ba~ <NA>           4.8  30.0 Backpac~
##  4     4  <NA>                                 <NA>           4.7  16.5 Backpac~
##  5     5  <NA>                                 <NA>           4.7  17.0 Backpac~
##  6     1 "HP Newest 14\" Ultral Light Laptop ~ <NA>           4.8  99   Laptops
##  7     2 "HP Portable Laptop, Student and Bus~ <NA>           4.7 422   Laptops
##  8     3  <NA>                                 <NA>           4.5 280.  Laptops
##  9     4 "HP 17 Laptop, 17.3" HD+ Display, 11~ <NA>           4.4 176   Laptops
## 10     5 "ASUS E410 Intel Celeron N4020 4GB 6~ <NA>           4.4 599   Laptops
## # i 15 more rows
```