---

title: “RWorksheet_#5a” author: “Espadon, Gerona, Bernasol” date: “2024-12-06” output: pdf_document

Load necessary libraries

```
library(rvest)
library(httr)
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(polite)
library(stringr)
library(ggplot2)
```

Define target URL

```
url <- 'https://www.imdb.com/chart/toptv/'

# Create a polite session
session <- bow(url, user_agent = "Educational")

session
```

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

#Extracting the ranks and titles

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

title_list
```

```
##  [1] "IMDb Charts"
##  [2] "Top 250 TV Shows"
##  [3] "1. Breaking Bad"
##  [4] "2. Planet Earth II"
##  [5] "3. Planet Earth"
##  [6] "4. Band of Brothers"
##  [7] "5. Chernobyl"
```

```
##  [8] "6. The Wire"
##  [9] "7. Avatar: The Last Airbender"
## [10] "8. Blue Planet II"
## [11] "9. The Sopranos"
## [12] "10. Cosmos: A Spacetime Odyssey"
## [13] "11. Cosmos"
## [14] "12. Our Planet"
## [15] "13. Game of Thrones"
## [16] "14. Bluey"
## [17] "15. The World at War"
## [18] "16. Fullmetal Alchemist: Brotherhood"
## [19] "17. Rick and Morty"
## [20] "18. Life"
## [21] "19. The Last Dance"
## [22] "20. The Twilight Zone"
## [23] "21. The Vietnam War"
## [24] "22. Sherlock"
## [25] "23. Attack on Titan"
## [26] "24. Batman: The Animated Series"
## [27] "25. Arcane"
## [28] "Recently viewed"
```

#Cleaning 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
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
```

Scrape Ratings

#Extracting tv rating, the number of people who voted, the number of episodes, and the year it was released.

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

rating_ls
```

```
##  [1] "9.5" "9.5" "9.4" "9.4" "9.3" "9.3" "9.3" "9.3" "9.2" "9.2" "9.3" "9.2"
## [13] "9.2" "9.3" "9.2" "9.1" "9.1" "9.1" "9.0" "9.0" "9.1" "9.1" "9.1" "9.0"
## [25] "9.0"
```

Scrape Vote Counts

```
voter_ls <- read_html(url) %>%
html_nodes('.ipc-rating-star--voteCount') %>%
html_text()

clean_votes <- gsub('[()]', '', voter_ls)
# Check if vote counts were extracted correctly
print(voter_ls)
```

```
##  [1] " (2.2M)" " (163K)" " (224K)" " (547K)" " (912K)" " (392K)" " (392K)"
##  [8] " (49K)"  " (501K)" " (132K)" " (46K)"  " (54K)"  " (2.4M)" " (34K)"
## [15] " (32K)"  " (210K)" " (629K)" " (44K)"  " (161K)" " (98K)"  " (30K)"
## [22] " (1M)"   " (567K)" " (123K)" " (341K)"
```

#extracted the number of episodes

```
eps_ls <- read_html(url) %>%
html_nodes('span.sc-300a8231-7.eaXxft.cli-title-metadata-item:nth-of-type(2)') %>%
html_text()
clean_eps <- gsub('[eps]', '', eps_ls)

num_eps <- as.numeric(clean_eps)

print(num_eps)
```

```
##  [1]  62   6  11  10   5  60  62   7  86  13  13  12  74 194  26  68  78  11  10
## [20] 156  10  15  98  85  18
```

```
#note to self, use gsub() to remove constant strings appearing in the dataset.
```

#extracted the year released

```
years <- read_html(url) %>%
html_nodes('span.sc-300a8231-7.eaXxft.cli-title-metadata-item:nth-of-type(1)') %>%
html_text()

years
```

```
##  [1] "2008-2013" "2016"      "2006"      "2001"      "2019"      "2002-2008"
##  [7] "2005-2008" "2017"      "1999-2007" "2014"      "1980"      "2019-2023"
## [13] "2011-2019" "2018- "    "1973-1974" "2009-2010" "2013- "    "2009"
## [19] "2020"      "1959-1964" "2017"      "2010-2017" "2013-2023" "1992-1995"
## [25] "2021-2024"
```

```
top_tv_shows <- data.frame(
Title = rank_title[,2],
Rating = rating_ls,
Voters = clean_votes,
Episodes = num_eps,
Year = years)


top_tv_shows
```

```
##                              Title Rating Voters Episodes      Year
## 1                      Breaking Bad    9.5   2.2M       62 2008-2013
## 2                   Planet Earth II    9.5   163K        6      2016
## 3                      Planet Earth    9.4   224K       11      2006
## 4                  Band of Brothers    9.4   547K       10      2001
## 5                         Chernobyl    9.3   912K        5      2019
## 6                         The Wire     9.3   392K       60 2002-2008
## 7        Avatar: The Last Airbender    9.3   392K       62 2005-2008
## 8                      Blue Planet II    9.3    49K        7      2017
## 9                       The Sopranos    9.2   501K       86 1999-2007
## 10      Cosmos: A Spacetime Odyssey    9.2   132K       13      2014
## 11                           Cosmos    9.3    46K       13      1980
## 12                       Our Planet    9.2    54K       12 2019-2023
## 13                  Game of Thrones    9.2   2.4M       74 2011-2019
## 14                            Bluey    9.3    34K      194    2018-
## 15                 The World at War    9.2    32K       26 1973-1974
## 16 Fullmetal Alchemist: Brotherhood    9.1   210K       68 2009-2010
## 17                   Rick and Morty    9.1   629K       78    2013-
## 18                             Life    9.1    44K       11      2009
## 19                   The Last Dance    9.0   161K       10      2020
## 20                The Twilight Zone    9.0    98K      156 1959-1964
## 21                  The Vietnam War    9.1    30K       10      2017
## 22                         Sherlock    9.1     1M       15 2010-2017
## 23                  Attack on Titan    9.1   567K       98 2013-2023
## 24      Batman: The Animated Series    9.0   123K       85 1992-1995
## 25                           Arcane    9.0   341K       18 2021-2024
```

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

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

links
```

```
##  [1] "/title/tt0903747/?ref_=chttvtp_t_1"
##  [2] "/title/tt5491994/?ref_=chttvtp_t_2"
##  [3] "/title/tt0795176/?ref_=chttvtp_t_3"
##  [4] "/title/tt0185906/?ref_=chttvtp_t_4"
##  [5] "/title/tt7366338/?ref_=chttvtp_t_5"
##  [6] "/title/tt0306414/?ref_=chttvtp_t_6"
##  [7] "/title/tt0417299/?ref_=chttvtp_t_7"
##  [8] "/title/tt6769208/?ref_=chttvtp_t_8"
##  [9] "/title/tt0141842/?ref_=chttvtp_t_9"
## [10] "/title/tt2395695/?ref_=chttvtp_t_10"
## [11] "/title/tt0081846/?ref_=chttvtp_t_11"
## [12] "/title/tt9253866/?ref_=chttvtp_t_12"
## [13] "/title/tt0944947/?ref_=chttvtp_t_13"
## [14] "/title/tt7678620/?ref_=chttvtp_t_14"
## [15] "/title/tt0071075/?ref_=chttvtp_t_15"
## [16] "/title/tt1355642/?ref_=chttvtp_t_16"
## [17] "/title/tt2861424/?ref_=chttvtp_t_17"
## [18] "/title/tt1533395/?ref_=chttvtp_t_18"
## [19] "/title/tt8420184/?ref_=chttvtp_t_19"
## [20] "/title/tt0052520/?ref_=chttvtp_t_20"
## [21] "/title/tt1877514/?ref_=chttvtp_t_21"
## [22] "/title/tt1475582/?ref_=chttvtp_t_22"
## [23] "/title/tt2560140/?ref_=chttvtp_t_23"
## [24] "/title/tt0103359/?ref_=chttvtp_t_24"
## [25] "/title/tt11126994/?ref_=chttvtp_t_25"
```

Loop to get link of each show's page

```r
show_data <- lapply(links, function(link) {
complete_link <- paste0("https://imdb.com", link)

#loop to 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")

#loop to extract critic reviews
critic <- usrv_link %>%
html_nodes("span.score") %>%
html_text()
critic_df <- data.frame(Critic_Reviews = critic[2], stringsAsFactors = FALSE)

#loop to extract pop rating
pop_rating <- usrv_link %>%
```

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

#loop to 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, Popularity_
})

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

```
##                                          Show_Link  User_Reviews
## 1    https://imdb.com/title/tt0903747/?ref_=chttvtp_t_1 5,116 reviews
## 2    https://imdb.com/title/tt0903747/?ref_=chttvtp_t_1 5,116 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 3,536 reviews
## 10   https://imdb.com/title/tt7366338/?ref_=chttvtp_t_5 3,536 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 5,911 reviews
## 26  https://imdb.com/title/tt0944947/?ref_=chttvtp_t_13 5,911 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 1,098 reviews
## 44  https://imdb.com/title/tt1475582/?ref_=chttvtp_t_22 1,098 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 2,268 reviews
## 50 https://imdb.com/title/tt11126994/?ref_=chttvtp_t_25 2,268 reviews
##    Critic_Reviews Popularity_Rating
## 1             175                19
## 2             175                19
## 3               6               985
## 4               6               985
## 5              10             1,837
## 6              10             1,837
## 7              34               162
## 8              34               162
## 9              88               143
## 10             88               143
## 11             77               108
## 12             77               108
## 13             57               294
## 14             57               294
## 15              9             4,270
## 16              9             4,270
## 17             93                27
## 18             93                27
## 19             12             1,439
## 20             12             1,439
## 21              8             3,817
## 22              8             3,817
## 23             15             2,632
## 24             15             2,632
## 25            368                12
## 26            368                12
## 27              4               331
## 28              4               331
## 29              5             2,696
## 30              5             2,696
## 31             16               479
## 32             16               479
## 33             94               124
## 34             94               124
## 35              9             3,030
## 36              9             3,030
## 37             28             1,530
## 38             28             1,530
## 39             85               319
## 40             85               319
## 41             13             1,776
```

```
## 42             13             1,776
## 43            121               167
## 44            121               167
## 45             65                43
## 46             65                43
## 47             25               510
## 48             25               510
## 49             59                 1
## 50             59                 1
```

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

```
##                              Title Rating Voters Episodes      Year
## 1                      Breaking Bad    9.5   2.2M       62 2008-2013
## 2                    Planet Earth II    9.5   163K        6      2016
## 3                       Planet Earth    9.4   224K       11      2006
## 4                   Band of Brothers    9.4   547K       10      2001
## 5                         Chernobyl     9.3   912K        5      2019
## 6                          The Wire     9.3   392K       60 2002-2008
## 7         Avatar: The Last Airbender    9.3   392K       62 2005-2008
## 8                      Blue Planet II    9.3    49K        7      2017
## 9                       The Sopranos    9.2   501K       86 1999-2007
## 10      Cosmos: A Spacetime Odyssey    9.2   132K       13      2014
## 11                           Cosmos    9.3    46K       13      1980
## 12                        Our Planet    9.2    54K       12 2019-2023
## 13                  Game of Thrones    9.2   2.4M       74 2011-2019
## 14                            Bluey    9.3    34K      194    2018-
## 15                 The World at War    9.2    32K       26 1973-1974
## 16 Fullmetal Alchemist: Brotherhood    9.1   210K       68 2009-2010
## 17                    Rick and Morty    9.1   629K       78    2013-
## 18                             Life    9.1    44K       11      2009
## 19                    The Last Dance    9.0   161K       10      2020
## 20                The Twilight Zone     9.0    98K      156 1959-1964
## 21                  The Vietnam War     9.1    30K       10      2017
## 22                          Sherlock    9.1     1M       15 2010-2017
## 23                  Attack on Titan     9.1   567K       98 2013-2023
## 24      Batman: The Animated Series    9.0   123K       85 1992-1995
## 25                           Arcane     9.0   341K       18 2021-2024
## 26                      Breaking Bad    9.5   2.2M       62 2008-2013
## 27                   Planet Earth II    9.5   163K        6      2016
## 28                      Planet Earth    9.4   224K       11      2006
## 29                  Band of Brothers    9.4   547K       10      2001
## 30                        Chernobyl     9.3   912K        5      2019
## 31                         The Wire     9.3   392K       60 2002-2008
## 32        Avatar: The Last Airbender    9.3   392K       62 2005-2008
## 33                     Blue Planet II    9.3    49K        7      2017
## 34                      The Sopranos    9.2   501K       86 1999-2007
## 35      Cosmos: A Spacetime Odyssey    9.2   132K       13      2014
## 36                           Cosmos    9.3    46K       13      1980
## 37                        Our Planet    9.2    54K       12 2019-2023
## 38                  Game of Thrones    9.2   2.4M       74 2011-2019
## 39                            Bluey    9.3    34K      194    2018-
## 40                 The World at War    9.2    32K       26 1973-1974
## 41 Fullmetal Alchemist: Brotherhood    9.1   210K       68 2009-2010
```

```
## 42                   Rick and Morty    9.1   629K       78    2013-
## 43                             Life    9.1    44K       11      2009
## 44                  The Last Dance    9.0   161K       10      2020
## 45                The Twilight Zone    9.0    98K      156 1959-1964
## 46                  The Vietnam War    9.1    30K       10      2017
## 47                         Sherlock    9.1     1M       15 2010-2017
## 48                  Attack on Titan    9.1   567K       98 2013-2023
## 49      Batman: The Animated Series    9.0   123K       85 1992-1995
## 50                           Arcane    9.0   341K       18 2021-2024
##                                                  Show_Link  User_Reviews
## 1    https://imdb.com/title/tt0903747/?ref_=chttvtp_t_1 5,116 reviews
## 2    https://imdb.com/title/tt0903747/?ref_=chttvtp_t_1 5,116 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 3,536 reviews
## 10   https://imdb.com/title/tt7366338/?ref_=chttvtp_t_5 3,536 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 5,911 reviews
## 26  https://imdb.com/title/tt0944947/?ref_=chttvtp_t_13 5,911 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 1,098 reviews
## 44  https://imdb.com/title/tt1475582/?ref_=chttvtp_t_22 1,098 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 2,268 reviews
## 50 https://imdb.com/title/tt11126994/?ref_=chttvtp_t_25 2,268 reviews
##    Critic_Reviews Popularity_Rating
## 1             175                19
## 2             175                19
## 3               6               985
## 4               6               985
## 5              10             1,837
## 6              10             1,837
## 7              34               162
## 8              34               162
## 9              88               143
## 10             88               143
## 11             77               108
## 12             77               108
## 13             57               294
## 14             57               294
## 15              9             4,270
## 16              9             4,270
## 17             93                27
## 18             93                27
## 19             12             1,439
## 20             12             1,439
## 21              8             3,817
## 22              8             3,817
## 23             15             2,632
## 24             15             2,632
## 25            368                12
## 26            368                12
## 27              4               331
## 28              4               331
## 29              5             2,696
## 30              5             2,696
## 31             16               479
## 32             16               479
## 33             94               124
## 34             94               124
## 35              9             3,030
## 36              9             3,030
## 37             28             1,530
## 38             28             1,530
## 39             85               319
## 40             85               319
## 41             13             1,776
## 42             13             1,776
## 43            121               167
## 44            121               167
## 45             65                43
## 46             65                43
## 47             25               510
```

```
## 48              25                510
## 49              59                  1
## 50              59                  1
```

```
#knitr::kable()

library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

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

```
knitr::kable(shows,caption = "Extracting Rating, VoteCount, Episodes, Year and Reviews") %>%
kable_classic(full_width = T, html_font = "Cambria") %>%
kable_styling(font_size = 8)
```

Table 1: Extracting Rating, VoteCount, Episodes, Year and Reviews

Title	Rating	Voters	Episodes	Year	Show_Link	User_Reviews	Critic_Reviews	Popularity	Rating
Breaking Bad	9.5	2.2M	62	2008–2013	https://imdb.com/title/tt0903747/?ref_=ch_tvtvtp_t_1	5,116 reviews	137	19	1
Planet Earth II	9.5	163K	6	2016	https://imdb.com/title/tt0903747/?ref_=ch_tvtvtp_t_1	5,116 reviews	137	19	1
Planet Earth	9.4	224K	11	2006	https://imdb.com/title/tt5491994/?ref_=ch_tvtvtp_t_2	958 reviews	85	98	2
Band of Brothers	9.4	547K	10	2001	https://imdb.com/title/tt5491994/?ref_=ch_tvtvtp_t_2	958 reviews	85	98	2
Chernobyl	9.3	912K	5	2019	https://imdb.com/title/tt0795176/?ref_=ch_tvtvtp_t_3	1,111 reviews	83	137	3
The Wire	9.3	392K	60	2002–2008	https://imdb.com/title/tt0795176/?ref_=ch_tvtvtp_t_3	1,111 reviews	83	137	3
Avatar: The Last Airbender	9.3	392K	62	2005–2008	https://imdb.com/title/tt0898266/?ref_=ch_tvtvtp_t_4	1,059 reviews	62	162	4
Blue Planet II	9.3	49K	7	2017	https://imdb.com/title/tt0898266/?ref_=ch_tvtvtp_t_4	1,059 reviews	62	162	4
The Sopranos	9.2	501K	86	1999–2007	https://imdb.com/title/tt0736338/?ref_=ch_tvtvtp_t_5	3,536 reviews	143	143	5
Cosmos: A Spacetime Odyssey	9.2	132K	13	2014	https://imdb.com/title/tt0736338/?ref_=ch_tvtvtp_t_5	3,536 reviews	143	143	5
Cosmos	9.3	46K	13	1980	https://imdb.com/title/tt0077414/?ref_=ch_tvtvtp_t_6	787 reviews	108	108	6
Our Planet	9.2	54K	12	2019–2023	https://imdb.com/title/tt0077414/?ref_=ch_tvtvtp_t_6	787 reviews	108	108	6
Game of Thrones	9.2	2.4M	74	2011–2019	https://imdb.com/title/tt0457299/?ref_=ch_tvtvtp_t_7	1,001 reviews	204	204	7
Bluey	9.3	34K	194	2018–	https://imdb.com/title/tt0457299/?ref_=ch_tvtvtp_t_7	1,001 reviews	204	204	7
The World at War	9.2	32K	26	1973–1974	https://imdb.com/title/tt0769208/?ref_=ch_tvtvtp_t_8	53 reviews	4	270	8
Fullmetal Alchemist: Brotherhood	9.1	210K	68	2009–2010	https://imdb.com/title/tt0769208/?ref_=ch_tvtvtp_t_8	53 reviews	4	270	8
Rick and Morty	9.1	629K	78	2013–	https://imdb.com/title/tt0149842/?ref_=ch_tvtvtp_t_9	968 reviews	27	27	9
Life	9.1	44K	11	2009	https://imdb.com/title/tt0149842/?ref_=ch_tvtvtp_t_9	968 reviews	27	27	9
The Last Dance	9.0	161K	10	2020	https://imdb.com/title/tt2395695/?ref_=ch_tvtvtp_t_10	205 reviews	1	439	10
The Twilight Zone	9.0	98K	156	1959–1964	https://imdb.com/title/tt2395695/?ref_=ch_tvtvtp_t_10	205 reviews	1	439	10

The Vietnam War	9.1	30K	10	2017	https://imdb.com/title/tt0081846/reviews?ref_=ttexrv_t_11	3,817	3,817
Sherlock	9.1	1M	15	2010–2017	https://imdb.com/title/tt0081846/?ref_=ttexrv_t_11	3,817	3,817
Attack on Titan	9.1	567K	98	2013–2023	https://imdb.com/title/tt9253866/?ref_=ttexrv_t_12	2,632	2,632
Batman: The Animated Series	9.0	123K	85	1992–1995	https://imdb.com/title/tt9253866/?ref_=ttexrv_t_12	2,632	2,632
Arcane	9.0	341K	18	2021–2024	https://imdb.com/title/tt0944947/?ref_=ttexrv_t_13	5,911	5,911
Breaking Bad	9.5	2.2M	62	2008–2013	https://imdb.com/title/tt0944947/?ref_=ttexrv_t_13	5,911	5,911
Planet Earth II	9.5	163K	6	2016	https://imdb.com/title/tt7678620/?ref_=ttexrv_t_14	331	331
Planet Earth	9.4	224K	11	2006	https://imdb.com/title/tt7678620/?ref_=ttexrv_t_14	331	331
Band of Brothers	9.4	547K	10	2001	https://imdb.com/title/tt0071075/?ref_=ttexrv_t_15	2,696	2,696
Chernobyl	9.3	912K	5	2019	https://imdb.com/title/tt0071075/?ref_=ttexrv_t_15	2,696	2,696
The Wire	9.3	392K	60	2002–2008	https://imdb.com/title/tt1355642/?ref_=ttexrv_t_16	479	479
Avatar: The Last Airbender	9.3	392K	62	2005–2008	https://imdb.com/title/tt1355642/?ref_=ttexrv_t_16	479	479
Blue Planet II	9.3	49K	7	2017	https://imdb.com/title/tt2861424/?ref_=ttexrv_t_17	124	124
The Sopranos	9.2	501K	86	1999–2007	https://imdb.com/title/tt2861424/?ref_=ttexrv_t_17	124	124
Cosmos: A Spacetime Odyssey	9.2	132K	13	2014	https://imdb.com/title/tt1533395/?ref_=ttexrv_t_18	3,030	3,030
Cosmos	9.3	46K	13	1980	https://imdb.com/title/tt1533395/?ref_=ttexrv_t_18	3,030	3,030
Our Planet	9.2	54K	12	2019–2023	https://imdb.com/title/tt8420184/?ref_=ttexrv_t_19	1,530	1,530
Game of Thrones	9.2	2.4M	74	2011–2019	https://imdb.com/title/tt8420184/?ref_=ttexrv_t_19	1,530	1,530
Bluey	9.3	34K	194	2018–	https://imdb.com/title/tt0052520/?ref_=ttexrv_t_20	319	319
The World at War	9.2	32K	26	1973–1974	https://imdb.com/title/tt0052520/?ref_=ttexrv_t_20	319	319
Fullmetal Alchemist: Brother-hood	9.1	210K	68	2009–2010	https://imdb.com/title/tt1877514/?ref_=ttexrv_t_21	1,776	1,776
Rick and Morty	9.1	629K	78	2013–	https://imdb.com/title/tt1877514/?ref_=ttexrv_t_21	1,776	1,776
Life	9.1	44K	11	2009	https://imdb.com/title/tt1475582/?ref_=ttexrv_t_22	167	167
The Last Dance	9.0	161K	10	2020	https://imdb.com/title/tt1475582/?ref_=ttexrv_t_22	167	167
The Twilight Zone	9.0	98K	156	1959–1964	https://imdb.com/title/tt2560140/?ref_=ttexrv_t_23	43	43
The Vietnam War	9.1	30K	10	2017	https://imdb.com/title/tt2560140/?ref_=ttexrv_t_23	43	43
Sherlock	9.1	1M	15	2010–2017	https://imdb.com/title/tt0102359/?ref_=ttexrv_t_24	510	510
Attack on Titan	9.1	567K	98	2013–2023	https://imdb.com/title/tt0102359/?ref_=ttexrv_t_24	510	510
Batman: The Animated Series	9.0	123K	85	1992–1995	https://imdb.com/title/tt1126994/?ref_=ttexrv_t_25	2,268	2,268
Arcane	9.0	341K	18	2021–2024	https://imdb.com/title/tt1126994/?ref_=ttexrv_t_25	2,268	2,268

```
library(kableExtra)

movies <- shows[c(1:5),]

knitr::kable(movies, caption = "IMDB Movies") %>%
kable_classic(full_width = T, html_font = "Arial Narrow") %>%
kable_styling(font_size = 8)
```

Table 2: IMDB Movies

Title	Rating	Voters	Episodes	Year	Show_Link	User_Reviews	Critic_Reviews	Popularity_Rating
Breaking Bad	9.5	2.2M	62	2008–2013	https://imdb.com/title/tt0903747/?ref_=ch_tvtp_t_1	5,116 reviews	137	19
Planet Earth II	9.5	163K	6	2016	https://imdb.com/title/tt0903747/?ref_=ch_tvtp_t_1	5,116 reviews	137	19
Planet Earth	9.4	224K	11	2006	https://imdb.com/title/tt5491994/?ref_=ch_tvtp_t_2	658 reviews	98	985
Band of Brothers	9.4	547K	10	2001	https://imdb.com/title/tt5491994/?ref_=ch_tvtp_t_2	658 reviews	98	985
Chernobyl	9.3	912K	5	2019	https://imdb.com/title/tt0795176/?ref_=ch_tvtp_t_3	1,111 reviews	183	837

#Extracting Amazon Product Reviews

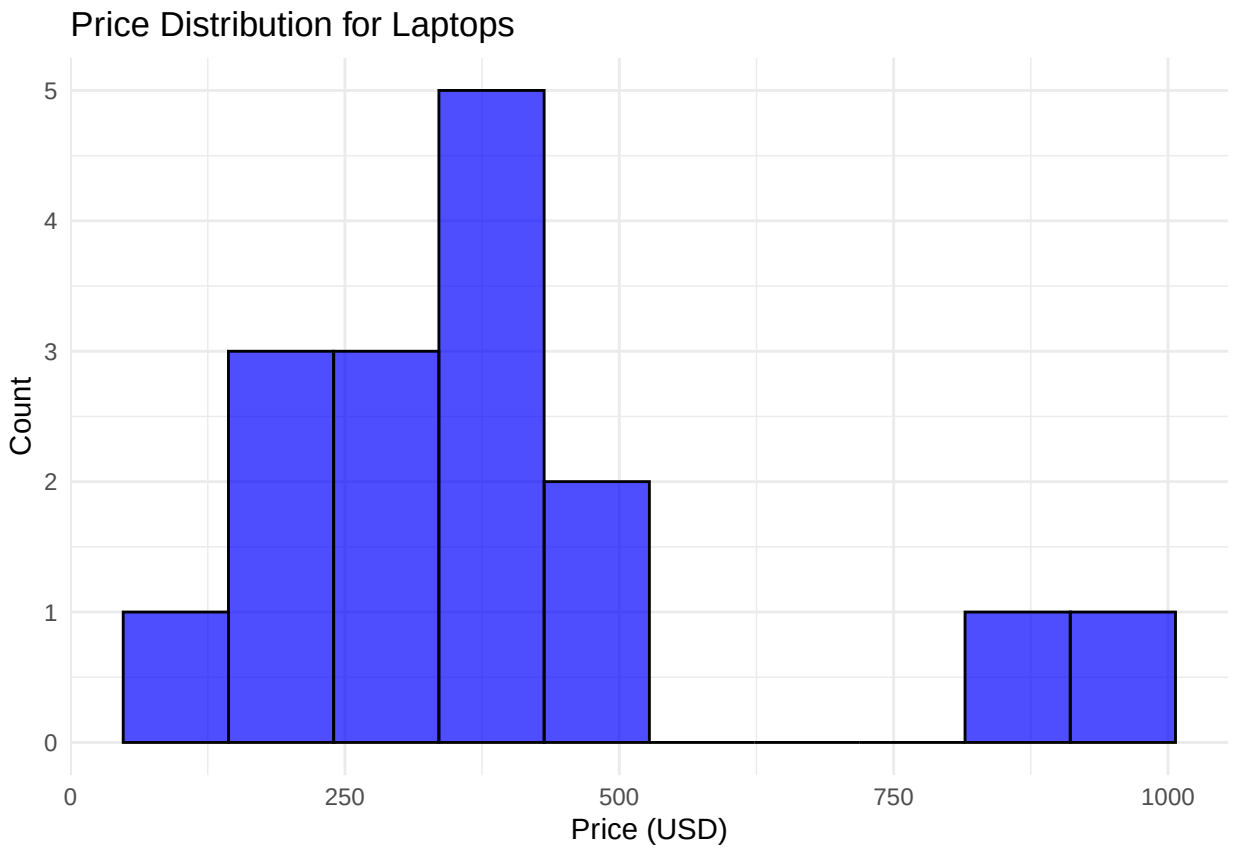
```
url <- "https://www.amazon.com/"

# Define the scraping function
scrape_amazon <- function(url) {
page <- read_html(url)

# Extract product details
products <- page %>% html_nodes(".s-title-instructions-style") %>% html_text(trim = TRUE)
prices <- page %>% html_nodes(".a-price-whole") %>% html_text(trim = TRUE)
ratings <- page %>% html_nodes(".a-icon-alt") %>% html_text(trim = TRUE)
reviews <- page %>% html_nodes(".s-underline-text") %>% html_text(trim = TRUE)

# Handle missing data by aligning lengths
max_length <- max(length(products), length(prices), length(ratings), length(reviews))
products <- c(products, rep(NA, max_length - length(products)))
prices <- c(prices, rep(NA, max_length - length(prices)))
ratings <- c(ratings, rep(NA, max_length - length(ratings)))
reviews <- c(reviews, rep(NA, max_length - length(reviews)))

# Create a data frame
return(data.frame(
Product = products,
Price = as.numeric(gsub("[^0-9.]", "", prices)),
Ratings = as.numeric(gsub("[^0-9.]", "", str_extract(ratings, "^[0-9.]+"))),
Reviews = as.numeric(gsub("[^0-9]", "", reviews)),
stringsAsFactors = FALSE
))
}
```

```
# Define URLs for categories
categories <- c("Laptops", "Books", "Shoes", "Televisions", "Fashion Bags")
urls <- c(
```

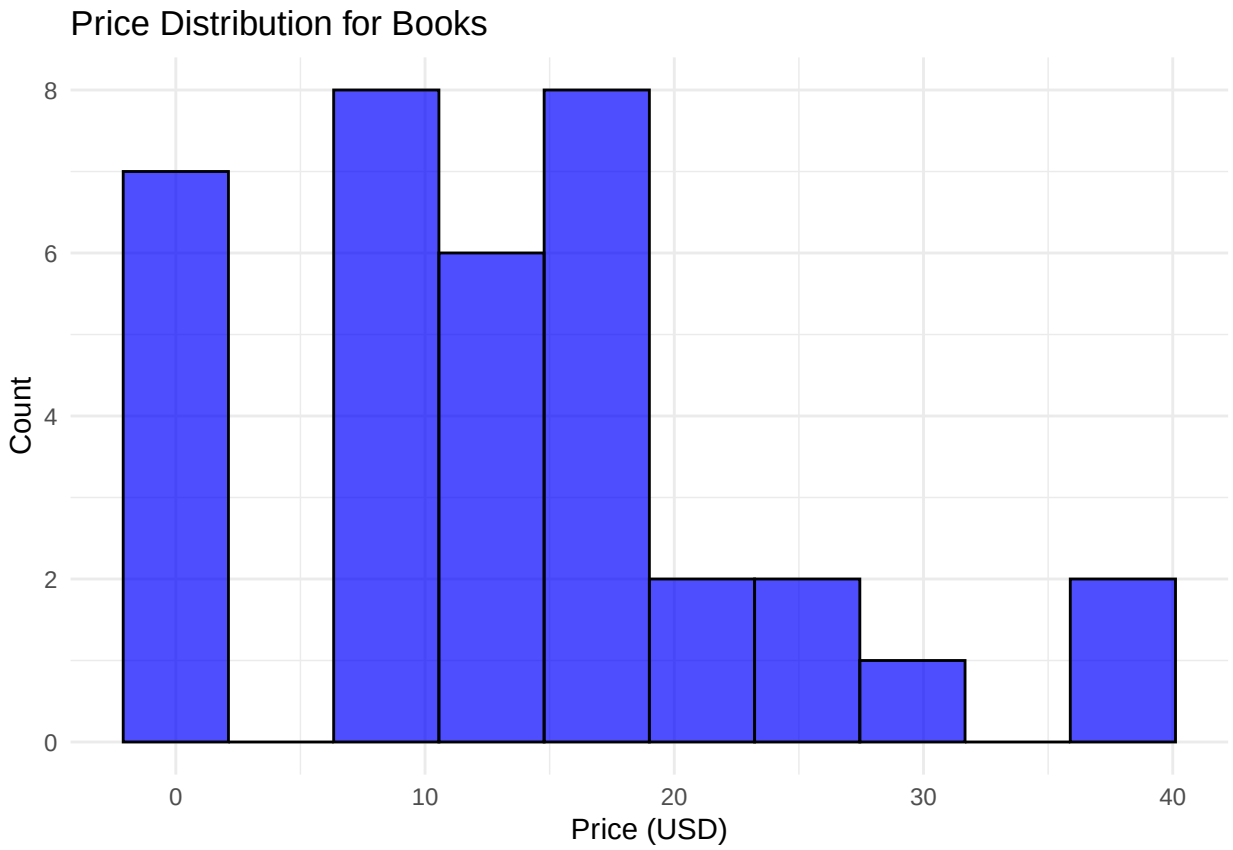
```
'https://www.amazon.com/s?k=laptop&crid=108GXR4VZZEMS&sprefix=lap%2Caps%2C680&ref=nb_sb_ss_ts-doa-p_1_3
'https://www.amazon.com/s?k=books&i=stripbooks-intl-ship&crid=3C5FBQTXKB575&sprefix=books%2Cstripbooks-
'https://www.amazon.com/s?k=shoes&i=stripbooks-intl-ship&crid=PWE5DZDD7EU7&sprefix=shoes%2Cstripbooks-in
'https://www.amazon.com/s?k=television&i=stripbooks-intl-ship&crid=O8JO99JDMGHY&sprefix=television%2Cstr
'https://www.amazon.com/s?k=fashion+bags&i=stripbooks-intl-ship&crid=3N3PC8YMHSW66&sprefix=fashion+bags%
)

# Scrape data for all categories
amazon_data <- lapply(urls, scrape_amazon)
names(amazon_data) <- categories

# Combine all data into a single data frame
combined_data <- bind_rows(amazon_data, .id = "Category")
```

Plot price distributions

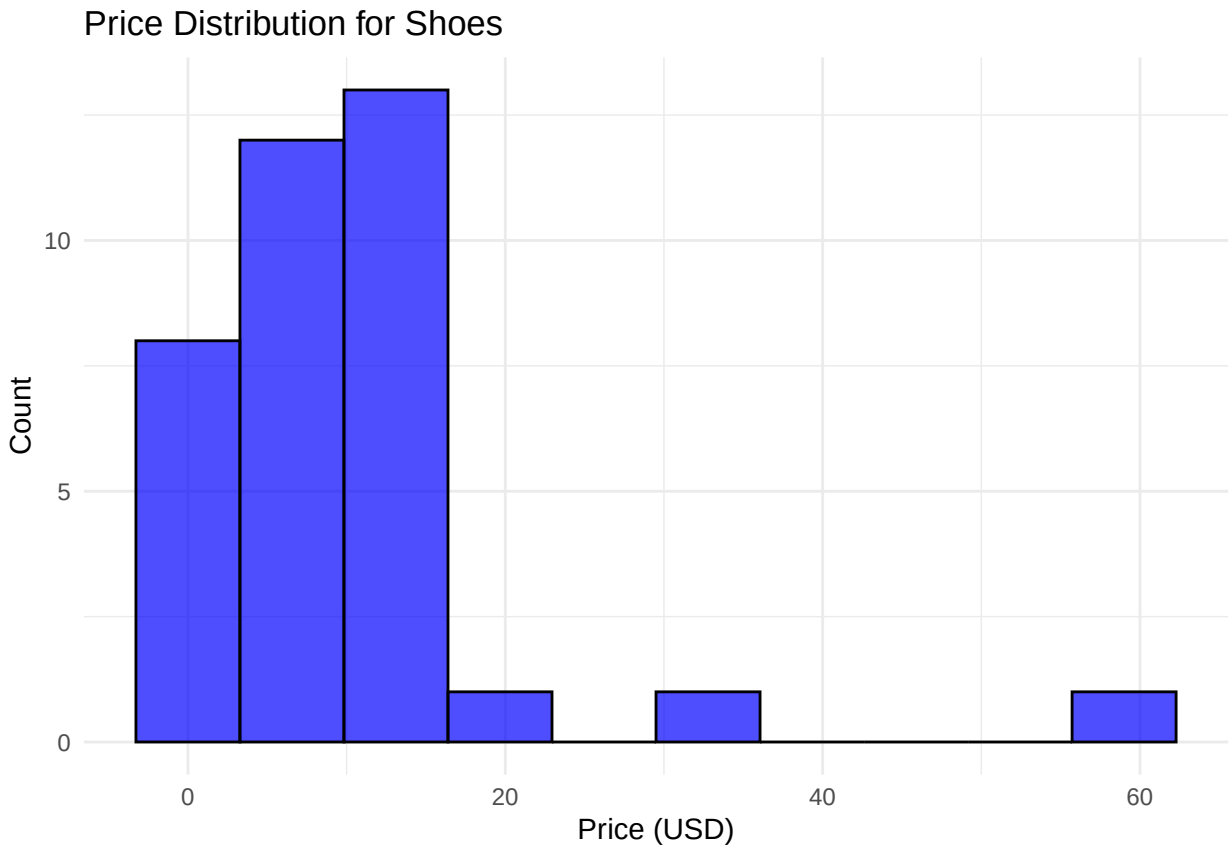
```
## Warning: Removed 73 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



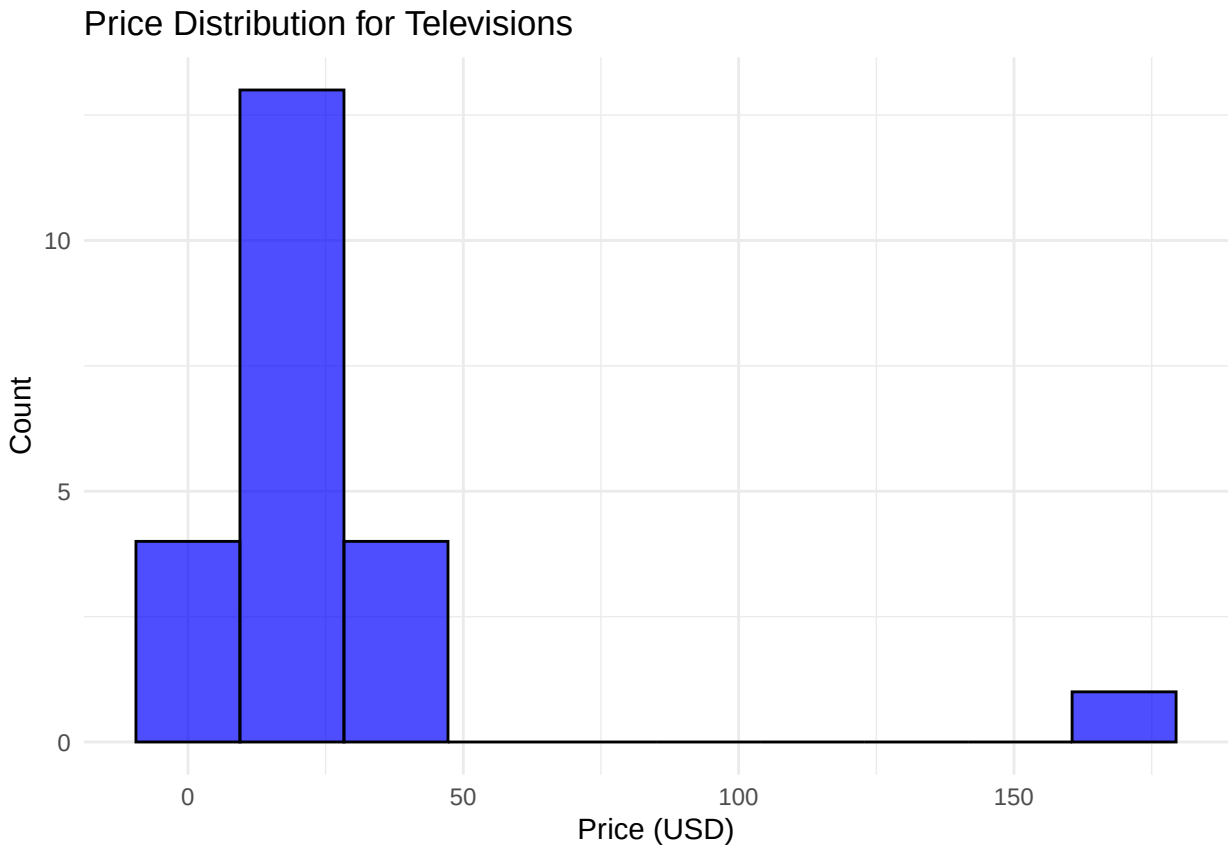

```
## Warning: Removed 133 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



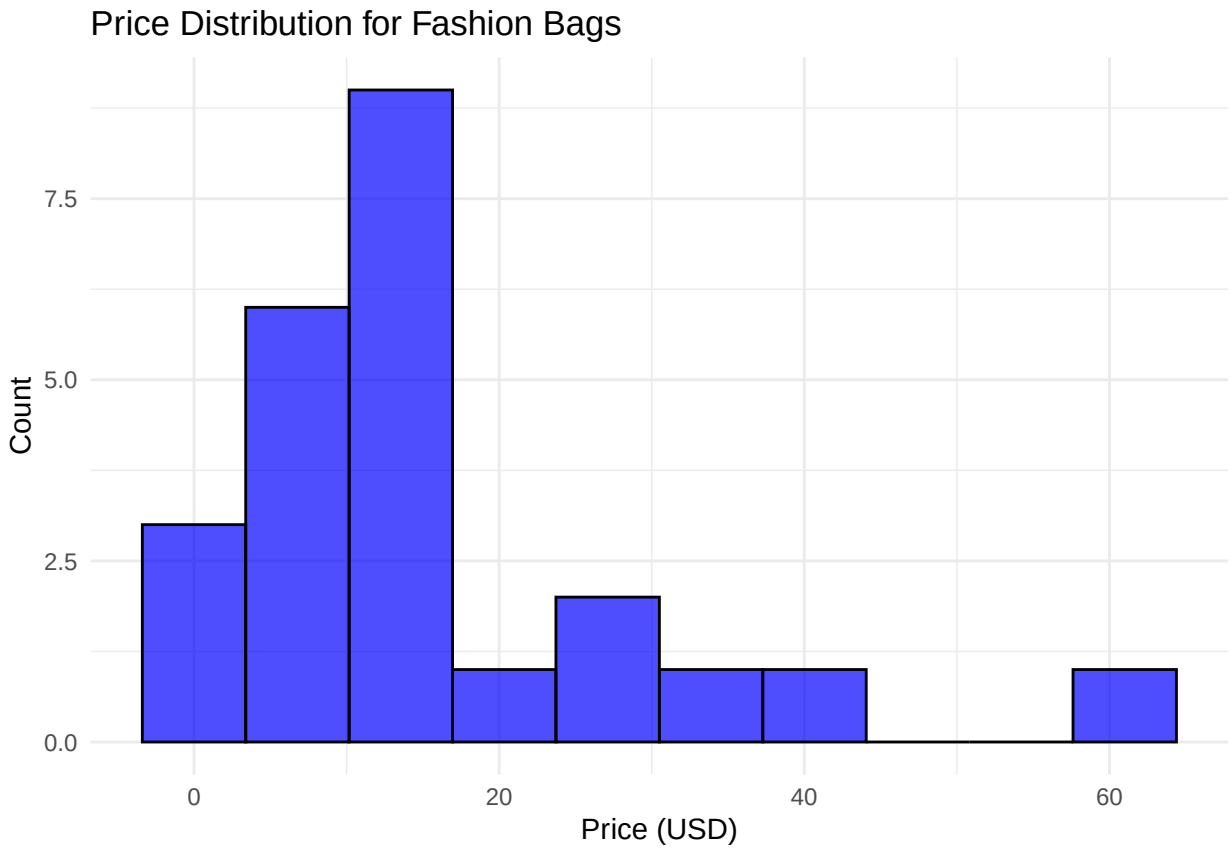

```
## Warning: Removed 140 rows containing non-finite outside the scale range
## (`stat_bin()`).
```




```
## Warning: Removed 83 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

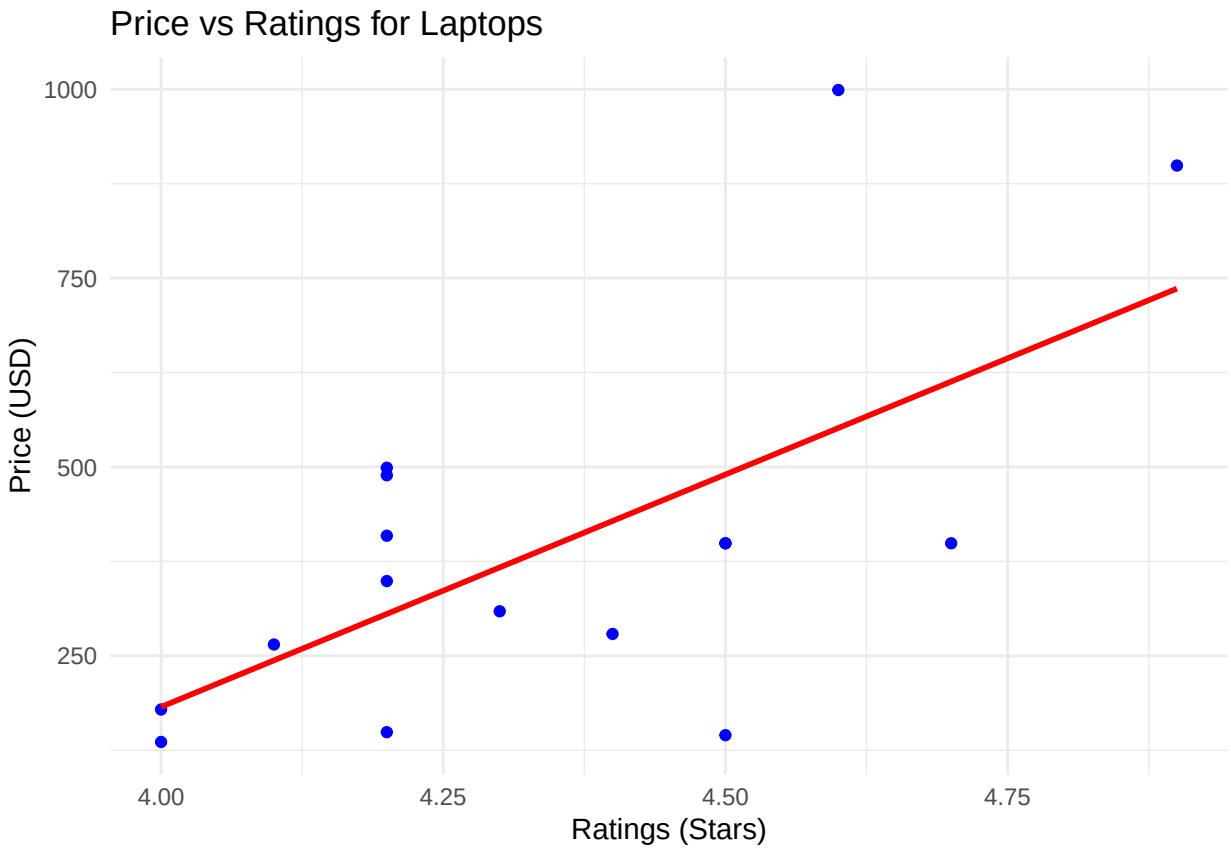



```
## Warning: Removed 102 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

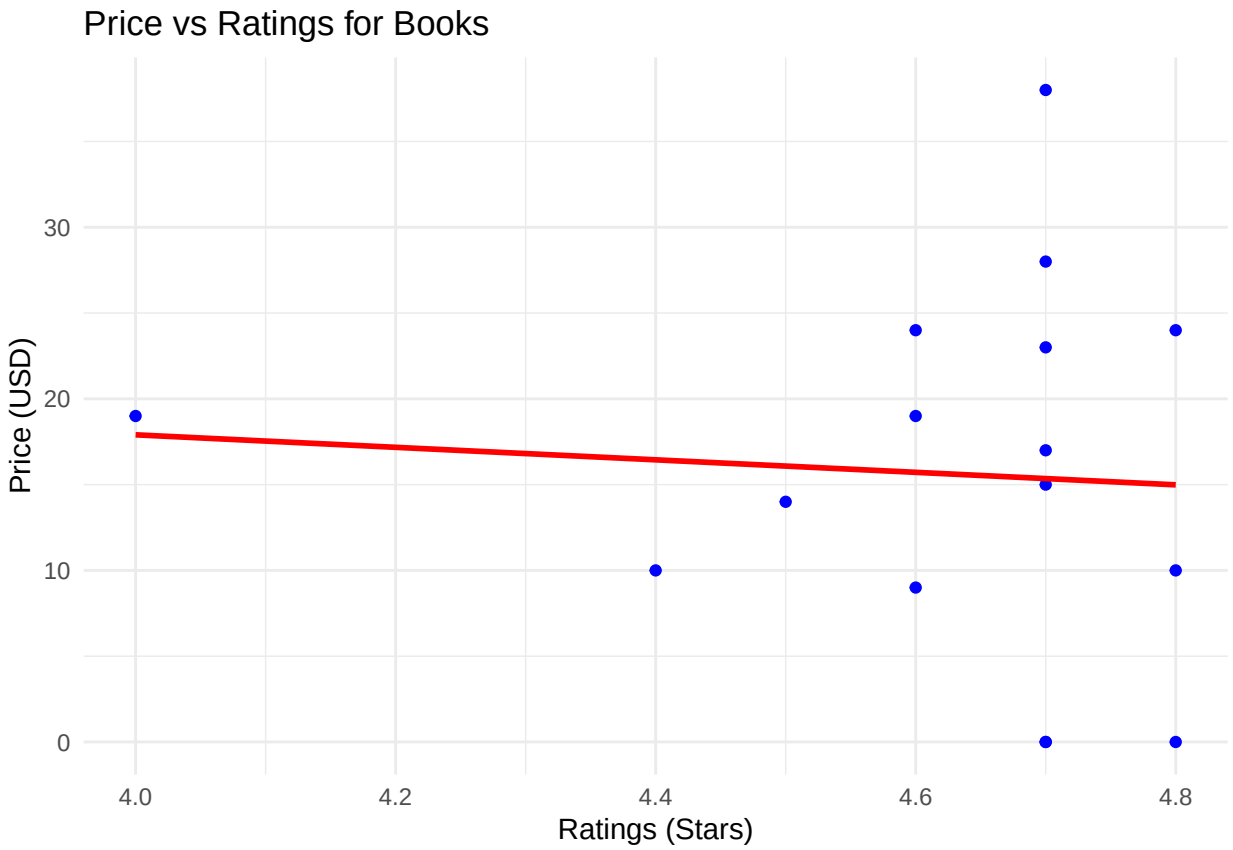


Plot price vs ratings

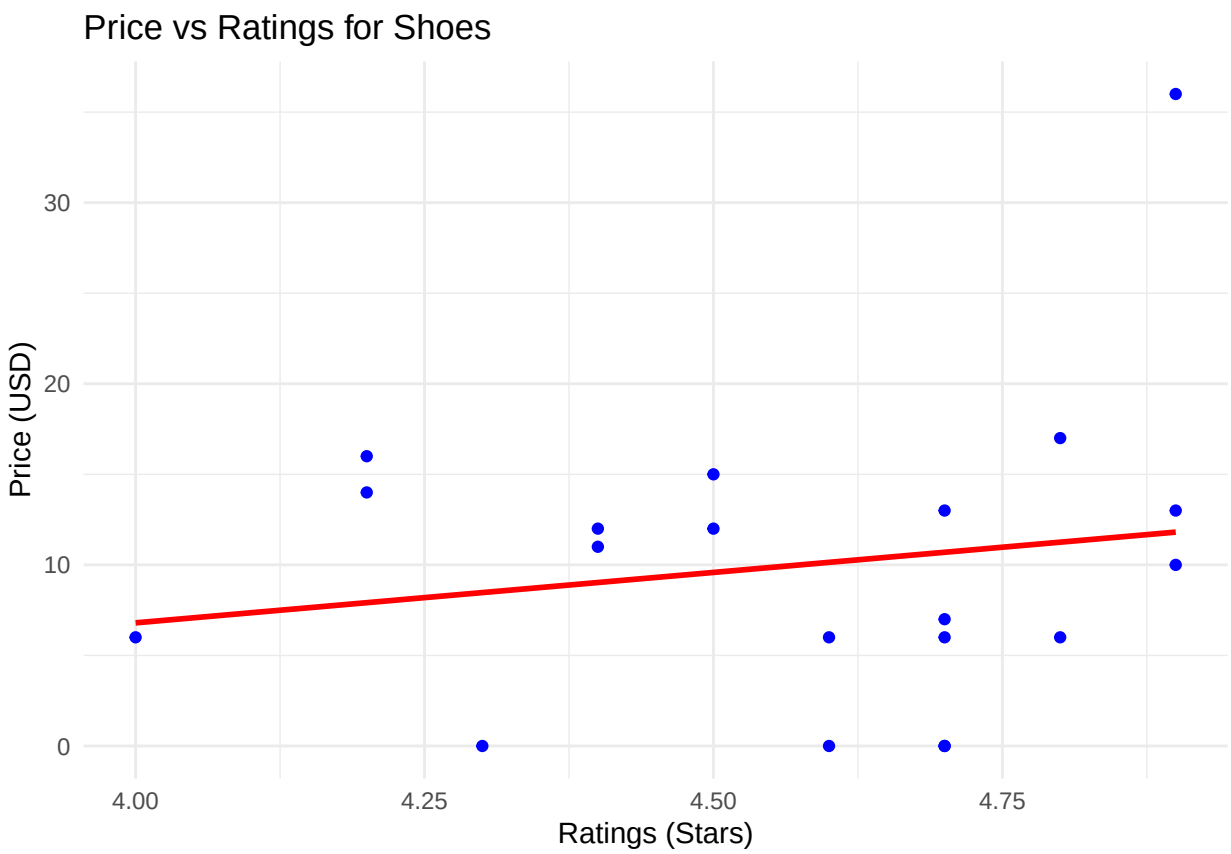
```
## `geom_smooth()` using formula = 'y ~ x'
```



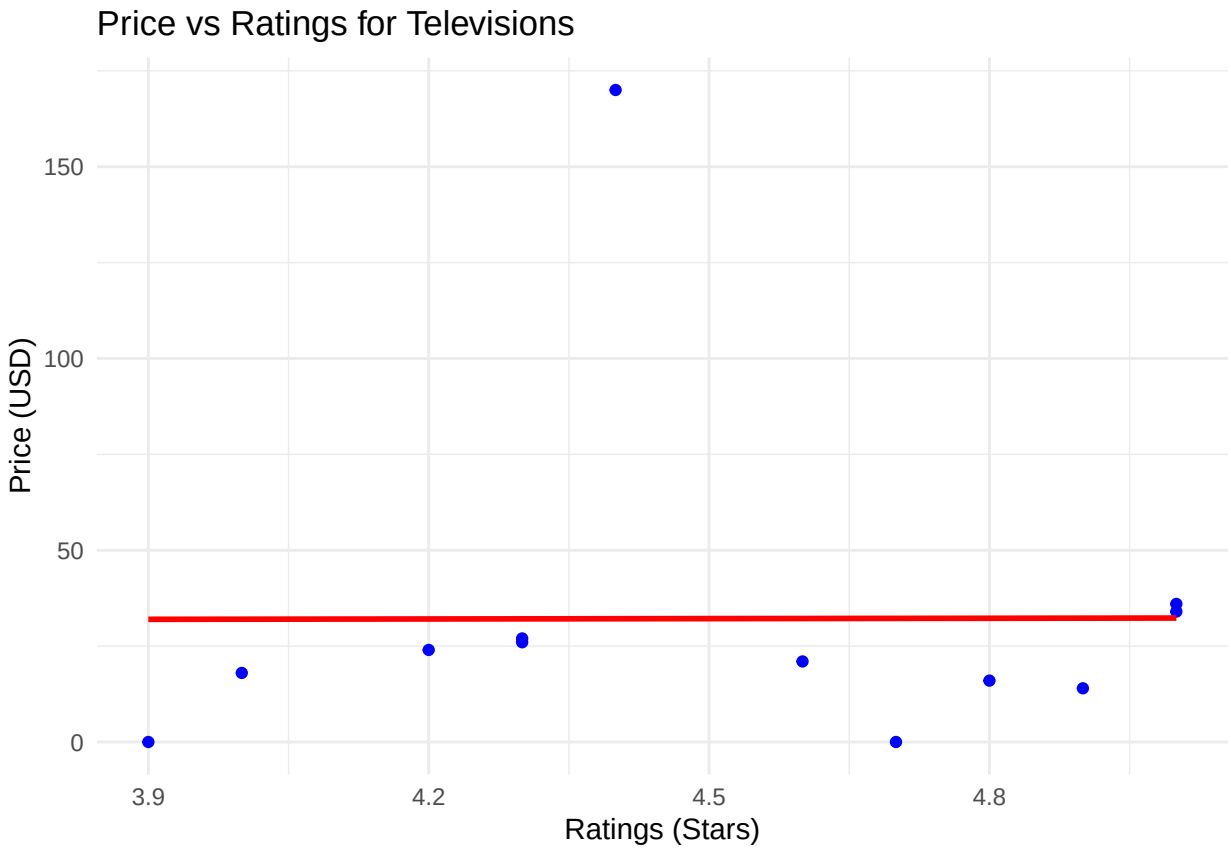

```
## `geom_smooth()` using formula = 'y ~ x'
```



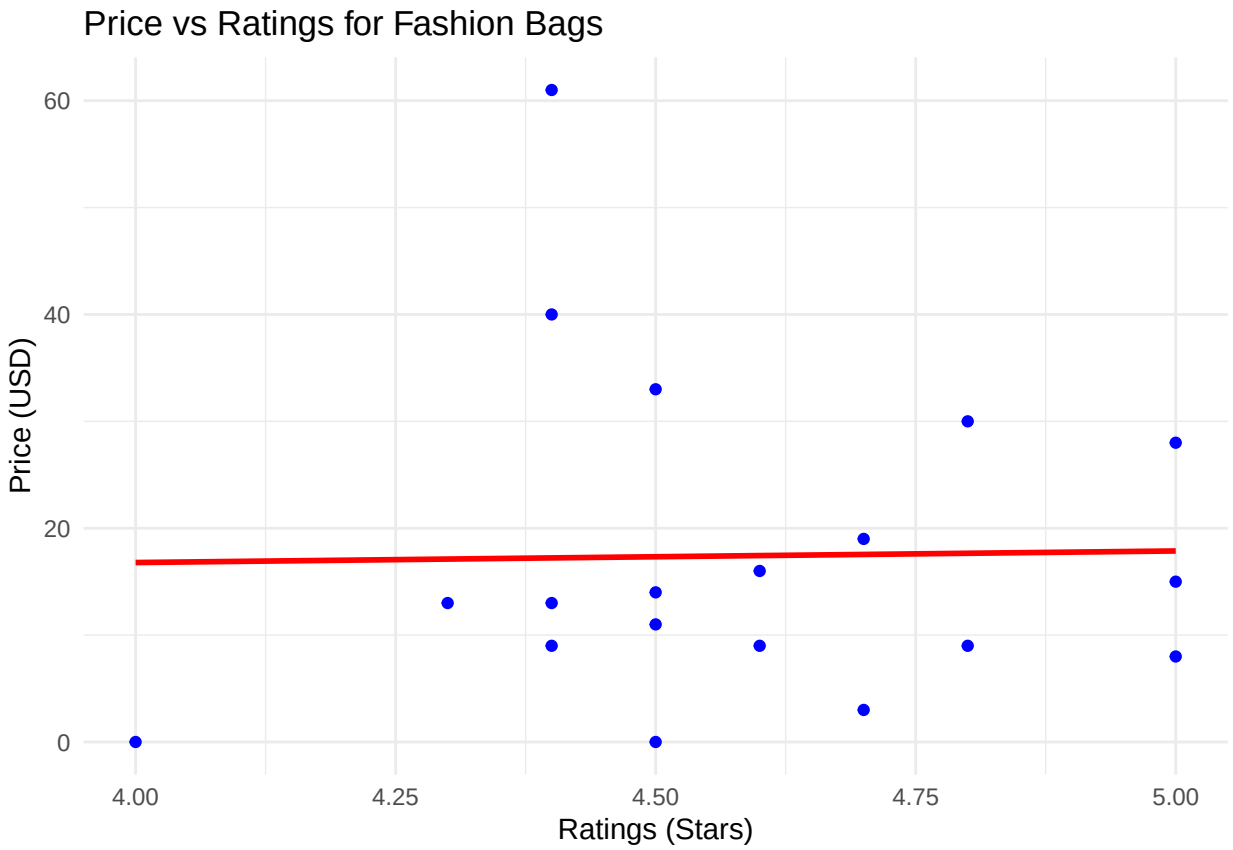

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
## `geom_smooth()` using formula = 'y ~ x'
```




```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# Rank products within each category
rank_products <- function(data) {
data <- data %>%
arrange(desc(Ratings), Price) %>%
mutate(Rank = row_number())
return(data)
}

ranked_data <- lapply(amazon_data, rank_products)

# Print top 5 products per category
for (category in categories) {
cat("\nTop 5 Products in", category, "\n")
print(head(ranked_data[[category]], 5))
}
```

```
##
## Top 5 Products in Laptops
##
## 1                                                    HP Lastest 255 G10 15.6" FHD Business Laptop
## 2 15.6 FHD Laptop, Win 11 Laptop Intel 12th Gen N100(Beat i3 1115G4), 16GB RAM 1TB SSD, WiFi 6, BT5.1
## 3       Apple 2024 MacBook Air 13-inch Laptop with M3 chip: Built for Apple Intelligence, 13.6-inch l
## 4                          acer Gateway Chromebook 311 CBO311-1H-C1MX Laptop | Intel Celeron N4500 | 1
## 5         NIMO 15.6'' FHD IPS Student Laptop, 16GB RAM 1TB SSD, Intel Pentium Quad-Core N100(Beat to
##   Price Ratings      Reviews Rank
## 1   899     4.9 3.090031e+19    1
## 2   399     4.7 1.441202e+25    2
```

```
## 3   999     4.6 2.799928e+19    3
## 4   145     4.5 1.900000e+01    4
## 5   399     4.5           NA    5
##
## Top 5 Products in Books
##
## 1 Little Corner: Coloring Book for Adults and Teens, Super Cute Designs of Cozy, Hygge Spaces for Rel
## 2                                                                                 Wind and Truth: Book Five of the
## 3       Girl Moments: Coloring Book for Adults and Teens Featuring Cute Cozy Daily Activities for Rel
## 4                                                  Dog Man: Big Jim Begins: A Graphic Novel (Dog Ma
## 5                                                                                                   Fo
##   Price Ratings Reviews Rank
## 1     0     4.8     279    1
## 2    10     4.8      NA    2
## 3    24     4.8      NA    3
## 4     0     4.7      33    4
## 5     0     4.7      NA    5
##
## Top 5 Products in Shoes
##
## 1 Red Lace, Yellow Lace: A Board/Picture Book For Kids About Learning to Tie Shoes and the Importance
## 2
## 3
## 4
## 5
##   Price Ratings      Reviews Rank
## 1    10     4.9 1.658166e+15    1
## 2    13     4.9           NA    2
## 3    36     4.9           NA    3
## 4     6     4.8           NA    4
## 5    17     4.8 1.239200e+04    5
##
## Top 5 Products in Televisions
##
## 1                                                                The Big Story: The Oral History of I
## 2                    3... 2...1... We're on the Air: An Inside Look at Sports Television, Journalism
## 3                           Black TV: Five Decades of Groundbreaking Television from Soul Train to
## 4 Behind the Screens: Illustrated Floor Plans and Scenes from the Best TV Shows of All Time  by Iñaki
## 5                                                  FILM/TV DIRECTOR'S FIELD MANUAL: Seventy Maxims
##   Price Ratings  Reviews Rank
## 1    34     5.0 14991499    1
## 2    36     5.0       NA    2
## 3    14     4.9      759    3
## 4    16     4.8 20252025    4
## 5     0     4.7       29    5
##
## Top 5 Products in Fashion Bags
##
## 1                                          Fashionary Bag Design: A Handbook for Accessories Des
## 2                                          Patchwork Quilted Bags: Totes, Purses and Accessories
## 3       Build a Bag Book: Tote Bags (paperback edition): Sew 15 stunning projects and endless variati
## 4 Half Yard Bags & Purses: Sew 12 Beautiful Bags and 12 Matching Purses  Part of: Half Yard (7 books)
## 5                                                Megan Hess: The Bag (Ultimate Fashion Ward
##   Price Ratings     Reviews Rank
```

```
## 1     8     5.0          NA    1
## 2    15     5.0 4.00040e+07    2
## 3    28     5.0 8.00000e+00    3
## 4     9     4.8          NA    4
## 5    30     4.8 6.19762e+15    5
```