---

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

```r
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] "9.1"
```

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] " (2.3M)" " (164K)" " (225K)" " (550K)" " (916K)" " (394K)" " (394K)"
##  [8] " (50K)"  " (505K)" " (133K)" " (47K)"  " (55K)"  " (2.4M)" " (35K)"
## [15] " (32K)"  " (212K)" " (632K)" " (45K)"  " (162K)" " (99K)"  " (31K)"
## [22] " (1M)"   " (572K)" " (124K)" " (358K)"
```

#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   2.3M       62 2008-2013
## 2                   Planet Earth II    9.5   164K        6      2016
## 3                      Planet Earth    9.4   225K       11      2006
## 4                  Band of Brothers    9.4   550K       10      2001
## 5                         Chernobyl    9.3   916K        5      2019
## 6                          The Wire    9.3   394K       60 2002-2008
## 7        Avatar: The Last Airbender    9.3   394K       62 2005-2008
## 8                    Blue Planet II    9.3    50K        7      2017
## 9                      The Sopranos    9.2   505K       86 1999-2007
## 10      Cosmos: A Spacetime Odyssey    9.2   133K       13      2014
## 11                           Cosmos    9.3    47K       13      1980
## 12                       Our Planet    9.2    55K       12 2019-2023
## 13                  Game of Thrones    9.2   2.4M       74 2011-2019
## 14                            Bluey    9.3    35K      194    2018-
## 15                 The World at War    9.2    32K       26 1973-1974
## 16 Fullmetal Alchemist: Brotherhood    9.1   212K       68 2009-2010
## 17                   Rick and Morty    9.1   632K       78    2013-
## 18                             Life    9.1    45K       11      2009
## 19                   The Last Dance    9.0   162K       10      2020
## 20                The Twilight Zone    9.0    99K      156 1959-1964
## 21                  The Vietnam War    9.1    31K       10      2017
## 22                         Sherlock    9.1     1M       15 2010-2017
## 23                  Attack on Titan    9.1   572K       98 2013-2023
## 24      Batman: The Animated Series    9.0   124K       85 1992-1995
## 25                           Arcane    9.1   358K       18 2021-2024
```

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


links <- main_page %>%
```

```r
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 5,129 reviews
## 2    https://imdb.com/title/tt0903747/?ref_=chttvtp_t_1 5,129 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 1,066 reviews
## 8    https://imdb.com/title/tt0185906/?ref_=chttvtp_t_4 1,066 reviews
## 9    https://imdb.com/title/tt7366338/?ref_=chttvtp_t_5 3,545 reviews
## 10   https://imdb.com/title/tt7366338/?ref_=chttvtp_t_5 3,545 reviews
## 11   https://imdb.com/title/tt0306414/?ref_=chttvtp_t_6   789 reviews
## 12   https://imdb.com/title/tt0306414/?ref_=chttvtp_t_6   789 reviews
## 13   https://imdb.com/title/tt0417299/?ref_=chttvtp_t_7 1,006 reviews
## 14   https://imdb.com/title/tt0417299/?ref_=chttvtp_t_7 1,006 reviews
## 15   https://imdb.com/title/tt6769208/?ref_=chttvtp_t_8    55 reviews
## 16   https://imdb.com/title/tt6769208/?ref_=chttvtp_t_8    55 reviews
## 17   https://imdb.com/title/tt0141842/?ref_=chttvtp_t_9   972 reviews
## 18   https://imdb.com/title/tt0141842/?ref_=chttvtp_t_9   972 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    81 reviews
## 22  https://imdb.com/title/tt0081846/?ref_=chttvtp_t_11    81 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 5,921 reviews
## 26  https://imdb.com/title/tt0944947/?ref_=chttvtp_t_13 5,921 reviews
## 27  https://imdb.com/title/tt7678620/?ref_=chttvtp_t_14   372 reviews
## 28  https://imdb.com/title/tt7678620/?ref_=chttvtp_t_14   372 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   910 reviews
## 34  https://imdb.com/title/tt2861424/?ref_=chttvtp_t_17   910 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   214 reviews
## 40  https://imdb.com/title/tt0052520/?ref_=chttvtp_t_20   214 reviews
## 41  https://imdb.com/title/tt1877514/?ref_=chttvtp_t_21   176 reviews
## 42  https://imdb.com/title/tt1877514/?ref_=chttvtp_t_21   176 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 2,384 reviews
## 46  https://imdb.com/title/tt2560140/?ref_=chttvtp_t_23 2,384 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 2,362 reviews
## 50 https://imdb.com/title/tt11126994/?ref_=chttvtp_t_25 2,362 reviews
##    Critic_Reviews Popularity_Rating
## 1             176                23
## 2             176                23
## 3               6             1,066
## 4               6             1,066
## 5              10             1,872
## 6              10             1,872
## 7              34               150
## 8              34               150
## 9              88               145
## 10             88               145
## 11             77               106
## 12             77               106
## 13             57               335
## 14             57               335
## 15              9             4,334
## 16              9             4,334
## 17             93                32
## 18             93                32
## 19             12             1,450
## 20             12             1,450
## 21              8             3,826
## 22              8             3,826
## 23             15             2,815
## 24             15             2,815
## 25            368                16
## 26            368                16
## 27              4               274
## 28              4               274
## 29              5             2,474
## 30              5             2,474
## 31             16               470
## 32             16               470
## 33             94               110
## 34             94               110
## 35              9             3,254
## 36              9             3,254
## 37             28             1,486
## 38             28             1,486
## 39             85               316
## 40             85               316
## 41             13             1,632
## 42             13             1,632
## 43            121               174
## 44            121               174
## 45             65                56
## 46             65                56
## 47             25               520
## 48             25               520
## 49             59                11
## 50             59                11
```

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

```
##                               Title Rating Voters Episodes      Year
## 1                      Breaking Bad    9.5   2.3M       62 2008-2013
## 2                   Planet Earth II    9.5   164K        6      2016
## 3                      Planet Earth    9.4   225K       11      2006
## 4                  Band of Brothers    9.4   550K       10      2001
## 5                         Chernobyl    9.3   916K        5      2019
## 6                          The Wire    9.3   394K       60 2002-2008
## 7        Avatar: The Last Airbender    9.3   394K       62 2005-2008
## 8                    Blue Planet II    9.3    50K        7      2017
## 9                      The Sopranos    9.2   505K       86 1999-2007
## 10      Cosmos: A Spacetime Odyssey    9.2   133K       13      2014
## 11                           Cosmos    9.3    47K       13      1980
## 12                       Our Planet    9.2    55K       12 2019-2023
## 13                  Game of Thrones    9.2   2.4M       74 2011-2019
## 14                            Bluey    9.3    35K      194    2018-
## 15                 The World at War    9.2    32K       26 1973-1974
## 16 Fullmetal Alchemist: Brotherhood    9.1   212K       68 2009-2010
## 17                   Rick and Morty    9.1   632K       78    2013-
## 18                             Life    9.1    45K       11      2009
## 19                   The Last Dance    9.0   162K       10      2020
## 20                The Twilight Zone    9.0    99K      156 1959-1964
## 21                  The Vietnam War    9.1    31K       10      2017
## 22                         Sherlock    9.1     1M       15 2010-2017
## 23                  Attack on Titan    9.1   572K       98 2013-2023
## 24      Batman: The Animated Series    9.0   124K       85 1992-1995
## 25                           Arcane    9.1   358K       18 2021-2024
## 26                     Breaking Bad    9.5   2.3M       62 2008-2013
## 27                  Planet Earth II    9.5   164K        6      2016
## 28                     Planet Earth    9.4   225K       11      2006
## 29                 Band of Brothers    9.4   550K       10      2001
## 30                        Chernobyl    9.3   916K        5      2019
## 31                         The Wire    9.3   394K       60 2002-2008
## 32       Avatar: The Last Airbender    9.3   394K       62 2005-2008
## 33                   Blue Planet II    9.3    50K        7      2017
## 34                     The Sopranos    9.2   505K       86 1999-2007
## 35      Cosmos: A Spacetime Odyssey    9.2   133K       13      2014
## 36                           Cosmos    9.3    47K       13      1980
## 37                       Our Planet    9.2    55K       12 2019-2023
## 38                  Game of Thrones    9.2   2.4M       74 2011-2019
## 39                            Bluey    9.3    35K      194    2018-
## 40                 The World at War    9.2    32K       26 1973-1974
## 41 Fullmetal Alchemist: Brotherhood    9.1   212K       68 2009-2010
## 42                   Rick and Morty    9.1   632K       78    2013-
## 43                             Life    9.1    45K       11      2009
## 44                   The Last Dance    9.0   162K       10      2020
## 45                The Twilight Zone    9.0    99K      156 1959-1964
## 46                  The Vietnam War    9.1    31K       10      2017
## 47                         Sherlock    9.1     1M       15 2010-2017
## 48                  Attack on Titan    9.1   572K       98 2013-2023
## 49      Batman: The Animated Series    9.0   124K       85 1992-1995
## 50                           Arcane    9.1   358K       18 2021-2024
##                                                 Show_Link  User_Reviews
## 1    https://imdb.com/title/tt0903747/?ref_=chttvtp_t_1 5,129 reviews
## 2    https://imdb.com/title/tt0903747/?ref_=chttvtp_t_1 5,129 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 1,066 reviews
## 8    https://imdb.com/title/tt0185906/?ref_=chttvtp_t_4 1,066 reviews
## 9    https://imdb.com/title/tt7366338/?ref_=chttvtp_t_5 3,545 reviews
## 10   https://imdb.com/title/tt7366338/?ref_=chttvtp_t_5 3,545 reviews
## 11   https://imdb.com/title/tt0306414/?ref_=chttvtp_t_6   789 reviews
## 12   https://imdb.com/title/tt0306414/?ref_=chttvtp_t_6   789 reviews
## 13   https://imdb.com/title/tt0417299/?ref_=chttvtp_t_7 1,006 reviews
## 14   https://imdb.com/title/tt0417299/?ref_=chttvtp_t_7 1,006 reviews
## 15   https://imdb.com/title/tt6769208/?ref_=chttvtp_t_8    55 reviews
## 16   https://imdb.com/title/tt6769208/?ref_=chttvtp_t_8    55 reviews
## 17   https://imdb.com/title/tt0141842/?ref_=chttvtp_t_9   972 reviews
## 18   https://imdb.com/title/tt0141842/?ref_=chttvtp_t_9   972 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    81 reviews
## 22  https://imdb.com/title/tt0081846/?ref_=chttvtp_t_11    81 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 5,921 reviews
## 26  https://imdb.com/title/tt0944947/?ref_=chttvtp_t_13 5,921 reviews
## 27  https://imdb.com/title/tt7678620/?ref_=chttvtp_t_14   372 reviews
## 28  https://imdb.com/title/tt7678620/?ref_=chttvtp_t_14   372 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   910 reviews
## 34  https://imdb.com/title/tt2861424/?ref_=chttvtp_t_17   910 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   214 reviews
## 40  https://imdb.com/title/tt0052520/?ref_=chttvtp_t_20   214 reviews
## 41  https://imdb.com/title/tt1877514/?ref_=chttvtp_t_21   176 reviews
## 42  https://imdb.com/title/tt1877514/?ref_=chttvtp_t_21   176 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 2,384 reviews
## 46  https://imdb.com/title/tt2560140/?ref_=chttvtp_t_23 2,384 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 2,362 reviews
## 50 https://imdb.com/title/tt11126994/?ref_=chttvtp_t_25 2,362 reviews
##    Critic_Reviews Popularity_Rating
## 1             176                23
## 2             176                23
## 3               6             1,066
## 4               6             1,066
## 5              10             1,872
## 6              10             1,872
## 7              34               150
## 8              34               150
## 9              88               145
## 10             88               145
## 11             77               106
## 12             77               106
## 13             57               335
## 14             57               335
## 15              9             4,334
## 16              9             4,334
## 17             93                32
## 18             93                32
## 19             12             1,450
## 20             12             1,450
## 21              8             3,826
## 22              8             3,826
## 23             15             2,815
## 24             15             2,815
## 25            368                16
## 26            368                16
## 27              4               274
## 28              4               274
## 29              5             2,474
## 30              5             2,474
## 31             16               470
## 32             16               470
## 33             94               110
## 34             94               110
## 35              9             3,254
## 36              9             3,254
## 37             28             1,486
## 38             28             1,486
## 39             85               316
## 40             85               316
## 41             13             1,632
## 42             13             1,632
## 43            121               174
## 44            121               174
## 45             65                56
## 46             65                56
## 47             25               520
## 48             25               520
## 49             59                11
## 50                 59                    11
```

```r
#knitr::kable()

library(kableExtra)
```

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

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

```r
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	2.3M	62	2008–2013	https://imdb.com/title/tt0903747/?ref_=chttvtp_t_1	5,129 reviews	136	23
Planet Earth II	9.5	164K	6	2016	https://imdb.com/title/tt0903747/?ref_=chttvtp_t_1	5,129 reviews	136	23
Planet Earth	9.4	225K	11	2006	https://imdb.com/title/tt5491994/?ref_=chttvtp_t_2	158 reviews	61	1,066
Band of Brothers	9.4	550K	10	2001	https://imdb.com/title/tt5491994/?ref_=chttvtp_t_2	158 reviews	61	1,066
Chernobyl	9.3	916K	5	2019	https://imdb.com/title/tt0795176/?ref_=chttvtp_t_3	111 reviews	10	1,872
The Wire	9.3	394K	60	2002–2008	https://imdb.com/title/tt0795176/?ref_=chttvtp_t_3	111 reviews	10	1,872
Avatar: The Last Airbender	9.3	394K	62	2005–2008	https://imdb.com/title/tt0185906/?ref_=chttvtp_t_4	1,066 reviews	34	150
Blue Planet II	9.3	50K	7	2017	https://imdb.com/title/tt0185906/?ref_=chttvtp_t_4	1,066 reviews	34	150
The Sopranos	9.2	505K	86	1999–2007	https://imdb.com/title/tt7366338/?ref_=chttvtp_t_5	3,545 reviews	88	145
Cosmos: A Spacetime Odyssey	9.2	133K	13	2014	https://imdb.com/title/tt7366338/?ref_=chttvtp_t_5	3,545 reviews	88	145
Cosmos	9.3	47K	13	1980	https://imdb.com/title/tt0306414/?ref_=chttvtp_t_6	789 reviews	77	106
Our Planet	9.2	55K	12	2019–2023	https://imdb.com/title/tt0306414/?ref_=chttvtp_t_6	789 reviews	77	106
Game of Thrones	9.2	2.4M	74	2011–2019	https://imdb.com/title/tt0417299/?ref_=chttvtp_t_7	1,006 reviews	57	335
Bluey	9.3	35K	194	2018–	https://imdb.com/title/tt0417299/?ref_=chttvtp_t_7	1,006 reviews	57	335
The World at War	9.2	32K	26	1973–1974	https://imdb.com/title/tt6769208/?ref_=chttvtp_t_8	55 reviews	9	4,334
Fullmetal Alchemist: Brotherhood	9.1	212K	68	2009–2010	https://imdb.com/title/tt6769208/?ref_=chttvtp_t_8	55 reviews	9	4,334
Rick and Morty	9.1	632K	78	2013–	https://imdb.com/title/tt0141842/?ref_=chttvtp_t_9	972 reviews	93	32
Life	9.1	45K	11	2009	https://imdb.com/title/tt0141842/?ref_=chttvtp_t_9	972 reviews	93	32
The Last Dance	9.0	162K	10	2020	https://imdb.com/title/tt2395695/?ref_=chttvtp_t_10	205 reviews	12	1,450
The Twilight Zone	9.0	99K	156	1959–1964	https://imdb.com/title/tt2395695/?ref_=chttvtp_t_10	205 reviews	12	1,450

The Vietnam War	9.1	31K	10	2017	https://imdb.com/title/tt0081846/?ref_=chttvtp_t_11	81 reviews	3,826
Sherlock	9.1	1M	15	2010–2017	https://imdb.com/title/tt0081846/?ref_=chttvtp_t_11	81 reviews	3,826
Attack on Titan	9.1	572K	98	2013–2023	https://imdb.com/title/tt9253866/?ref_=chttvtp_t_12	245 reviews	2,815
Batman: The Animated Series	9.0	124K	85	1992–1995	https://imdb.com/title/tt9253866/?ref_=chttvtp_t_12	245 reviews	2,815
Arcane	9.1	358K	18	2021–2024	https://imdb.com/title/tt0944947/?ref_=chttvtp_t_13	5,921 reviews	16
Breaking Bad	9.5	2.3M	62	2008–2013	https://imdb.com/title/tt0944947/?ref_=chttvtp_t_13	5,921 reviews	16
Planet Earth II	9.5	164K	6	2016	https://imdb.com/title/tt7678620/?ref_=chttvtp_t_14	372 reviews	274
Planet Earth	9.4	225K	11	2006	https://imdb.com/title/tt7678620/?ref_=chttvtp_t_14	372 reviews	274
Band of Brothers	9.4	550K	10	2001	https://imdb.com/title/tt0071075/?ref_=chttvtp_t_15	126 reviews	2,474
Chernobyl	9.3	916K	5	2019	https://imdb.com/title/tt0071075/?ref_=chttvtp_t_15	126 reviews	2,474
The Wire	9.3	394K	60	2002–2008	https://imdb.com/title/tt1355642/?ref_=chttvtp_t_16	468 reviews	470
Avatar: The Last Airbender	9.3	394K	62	2005–2008	https://imdb.com/title/tt1355642/?ref_=chttvtp_t_16	468 reviews	470
Blue Planet II	9.3	50K	7	2017	https://imdb.com/title/tt2861424/?ref_=chttvtp_t_17	910 reviews	110
The Sopranos	9.2	505K	86	1999–2007	https://imdb.com/title/tt2861424/?ref_=chttvtp_t_17	910 reviews	110
Cosmos: A Spacetime Odyssey	9.2	133K	13	2014	https://imdb.com/title/tt1533395/?ref_=chttvtp_t_18	12 reviews	3,254
Cosmos	9.3	47K	13	1980	https://imdb.com/title/tt1533395/?ref_=chttvtp_t_18	12 reviews	3,254
Our Planet	9.2	55K	12	2019–2023	https://imdb.com/title/tt8420184/?ref_=chttvtp_t_19	541 reviews	1,486
Game of Thrones	9.2	2.4M	74	2011–2019	https://imdb.com/title/tt8420184/?ref_=chttvtp_t_19	541 reviews	1,486
Bluey	9.3	35K	194	2018–	https://imdb.com/title/tt0052520/?ref_=chttvtp_t_20	214 reviews	316
The World at War	9.2	32K	26	1973–1974	https://imdb.com/title/tt0052520/?ref_=chttvtp_t_20	214 reviews	316
Fullmetal Alchemist: Brotherhood	9.1	212K	68	2009–2010	https://imdb.com/title/tt1877514/?ref_=chttvtp_t_21	176 reviews	1,632
Rick and Morty	9.1	632K	78	2013–	https://imdb.com/title/tt1877514/?ref_=chttvtp_t_21	176 reviews	1,632
Life	9.1	45K	11	2009	https://imdb.com/title/tt1475582/?ref_=chttvtp_t_22	1,098 reviews	174
The Last Dance	9.0	162K	10	2020	https://imdb.com/title/tt1475582/?ref_=chttvtp_t_22	1,098 reviews	174
The Twilight Zone	9.0	99K	156	1959–1964	https://imdb.com/title/tt2560140/?ref_=chttvtp_t_23	2,384 reviews	56
The Vietnam War	9.1	31K	10	2017	https://imdb.com/title/tt2560140/?ref_=chttvtp_t_23	2,384 reviews	56
Sherlock	9.1	1M	15	2010–2017	https://imdb.com/title/tt0103359/?ref_=chttvtp_t_24	219 reviews	520
Attack on Titan	9.1	572K	98	2013–2023	https://imdb.com/title/tt0103359/?ref_=chttvtp_t_24	219 reviews	520
Batman: The Animated Series	9.0	124K	85	1992–1995	https://imdb.com/title/tt11126994/?ref_=chttvtp_t_25	2,362 reviews	11
Arcane	9.1	358K	18	2021–2024	https://imdb.com/title/tt11126994/?ref_=chttvtp_t_25	2,362 reviews	11

```r
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	2.3M	62	2008–2013	https://imdb.com/title/tt0903747/?ref_=chttvtp_t_1	5,129 reviews	176	23
Planet Earth II	9.5	164K	6	2016	https://imdb.com/title/tt0903747/?ref_=chttvtp_t_1	5,129 reviews	176	23
Planet Earth	9.4	225K	11	2006	https://imdb.com/title/tt5491994/?ref_=chttvtp_t_2	158 reviews	61	1,066
Band of Brothers	9.4	550K	10	2001	https://imdb.com/title/tt5491994/?ref_=chttvtp_t_2	158 reviews	61	1,066
Chernobyl	9.3	916K	5	2019	https://imdb.com/title/tt0795176/?ref_=chttvtp_t_3	111 reviews	150	1,872

#Extracting Amazon Product Reviews

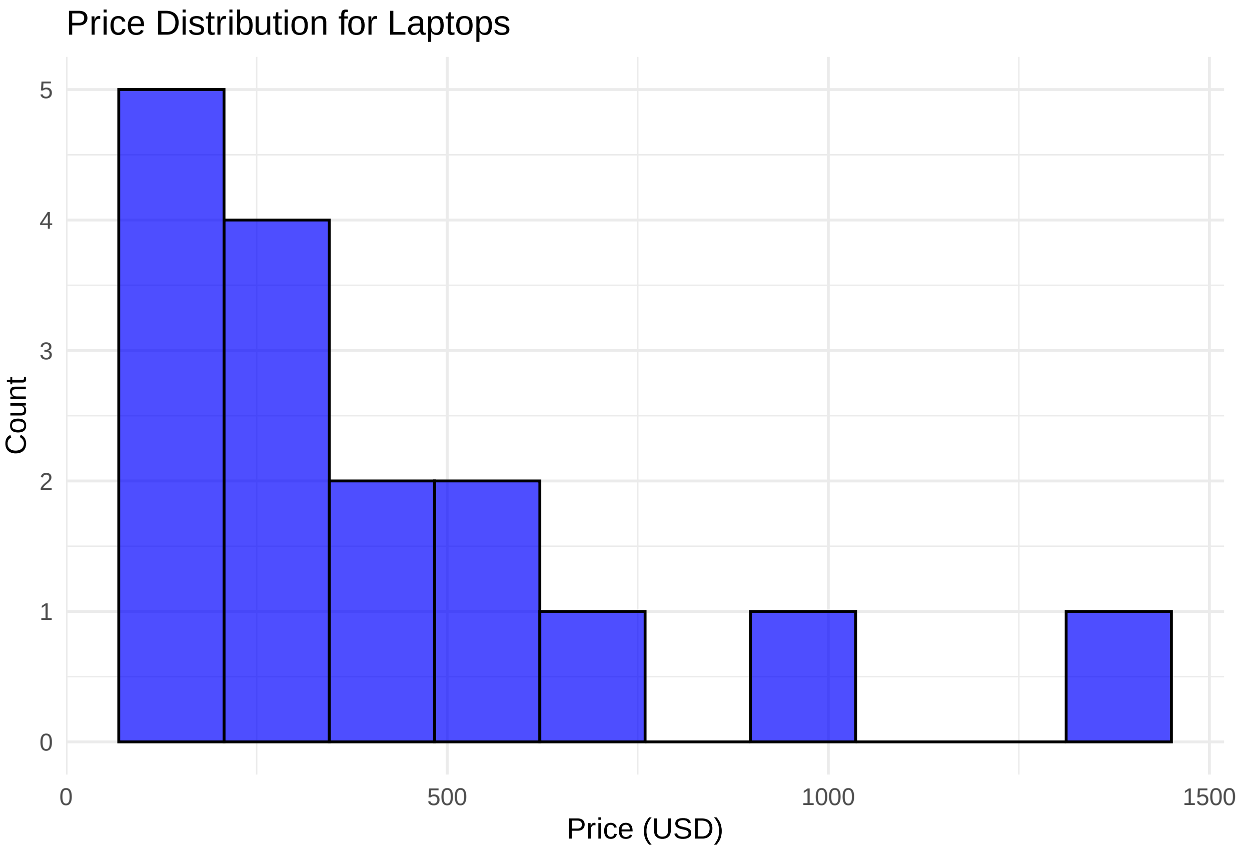
```r
url <- "https://www.amazon.com/"

# 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
))
}
```

```r
# 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-i
'https://www.amazon.com/s?k=television&i=stripbooks-intl-ship&crid=08J099JDMGHY&sprefix=television%2Cst
'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")
```

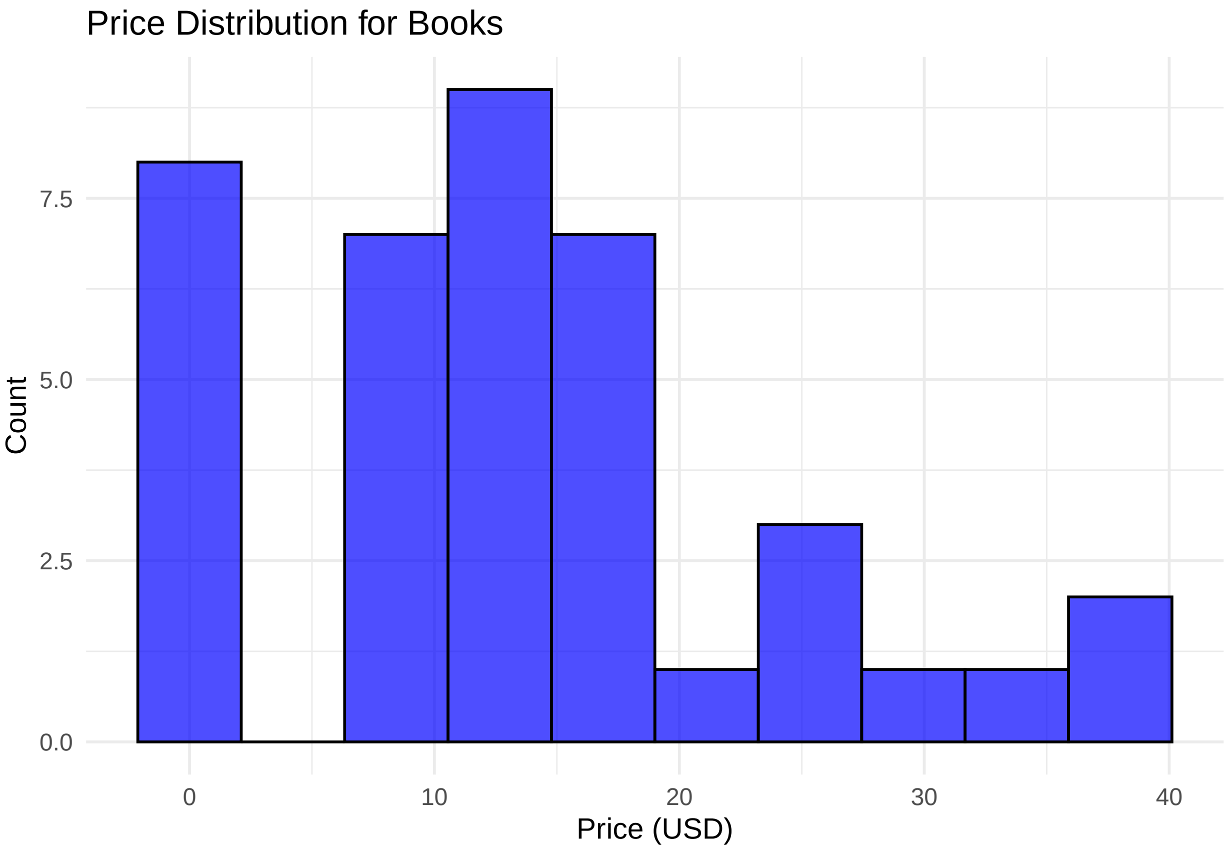
Plot price distributions

```
for (category in categories) {
data <- amazon_data[[category]]
p <- ggplot(data, aes(x = Price)) +
geom_histogram(bins = 10, fill = "blue", color = "black", alpha = 0.7) +
labs(title = paste("Price Distribution for", category),
x = "Price (USD)", y = "Count") +
theme_minimal()
print(p) # Explicitly print plot
}
```

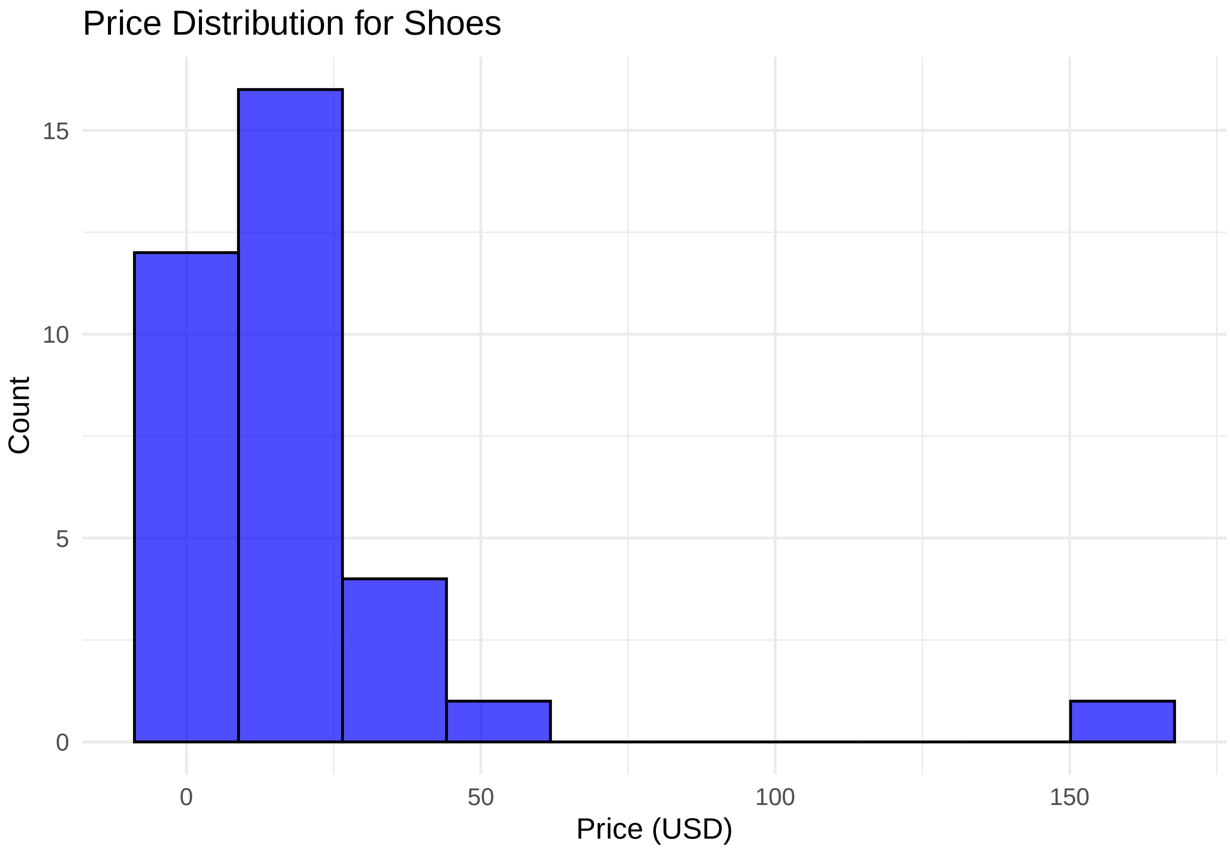
```
## Warning: Removed 62 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



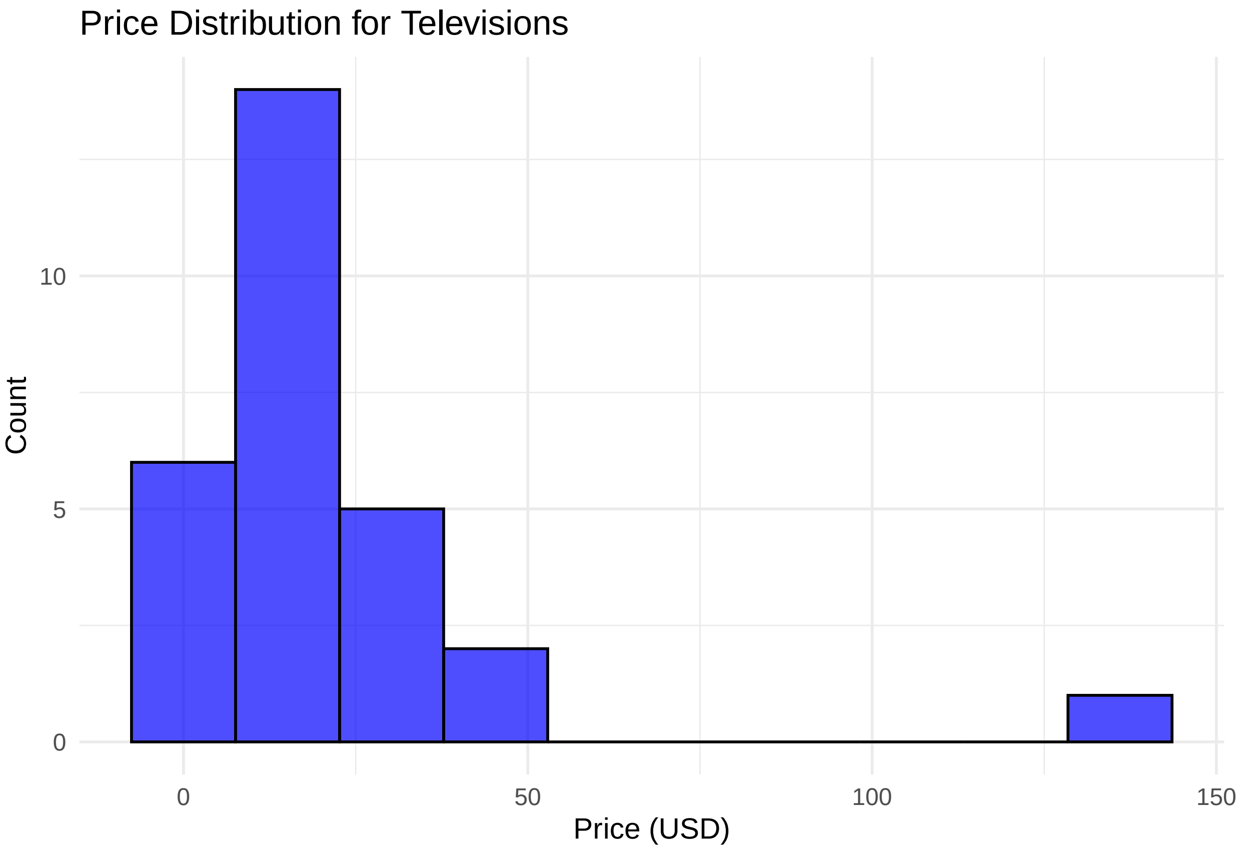
```
## Warning: Removed 121 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



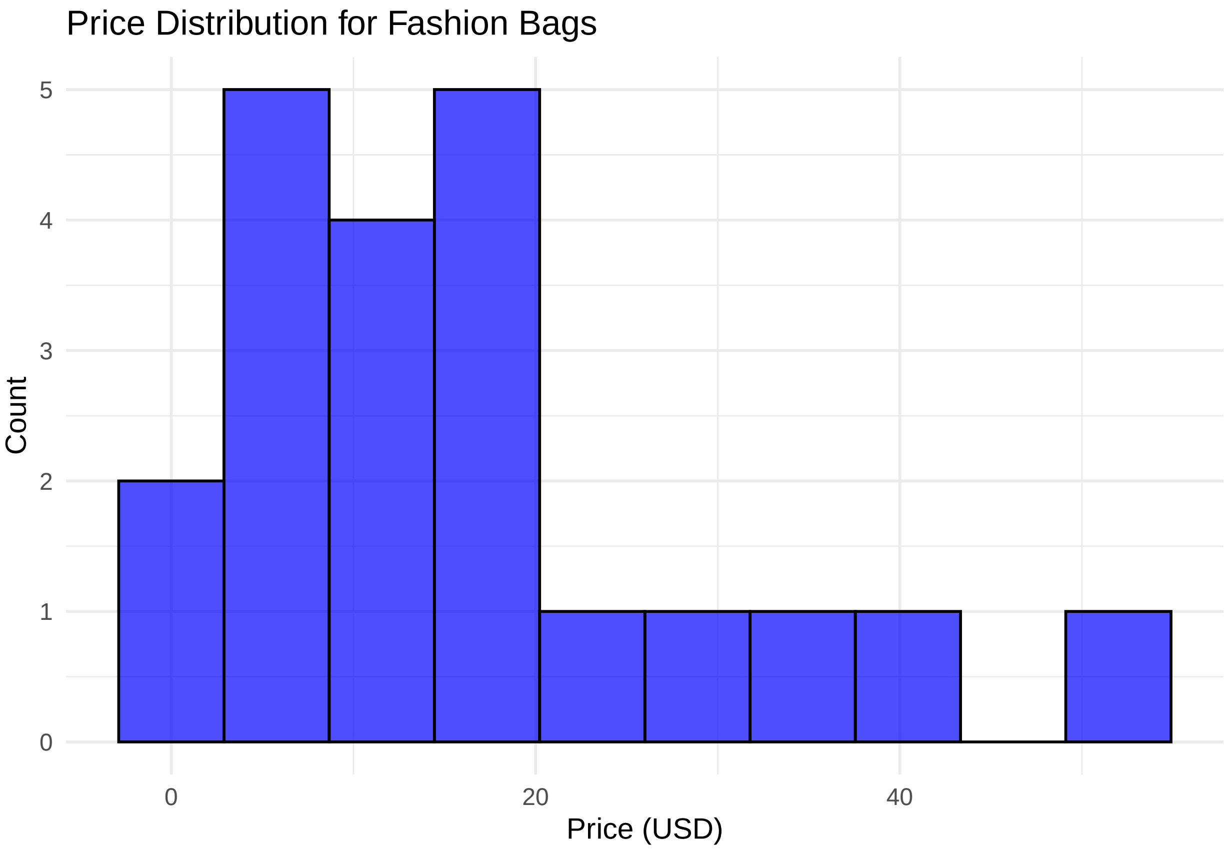
```
## Warning: Removed 121 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



```
## Warning: Removed 80 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



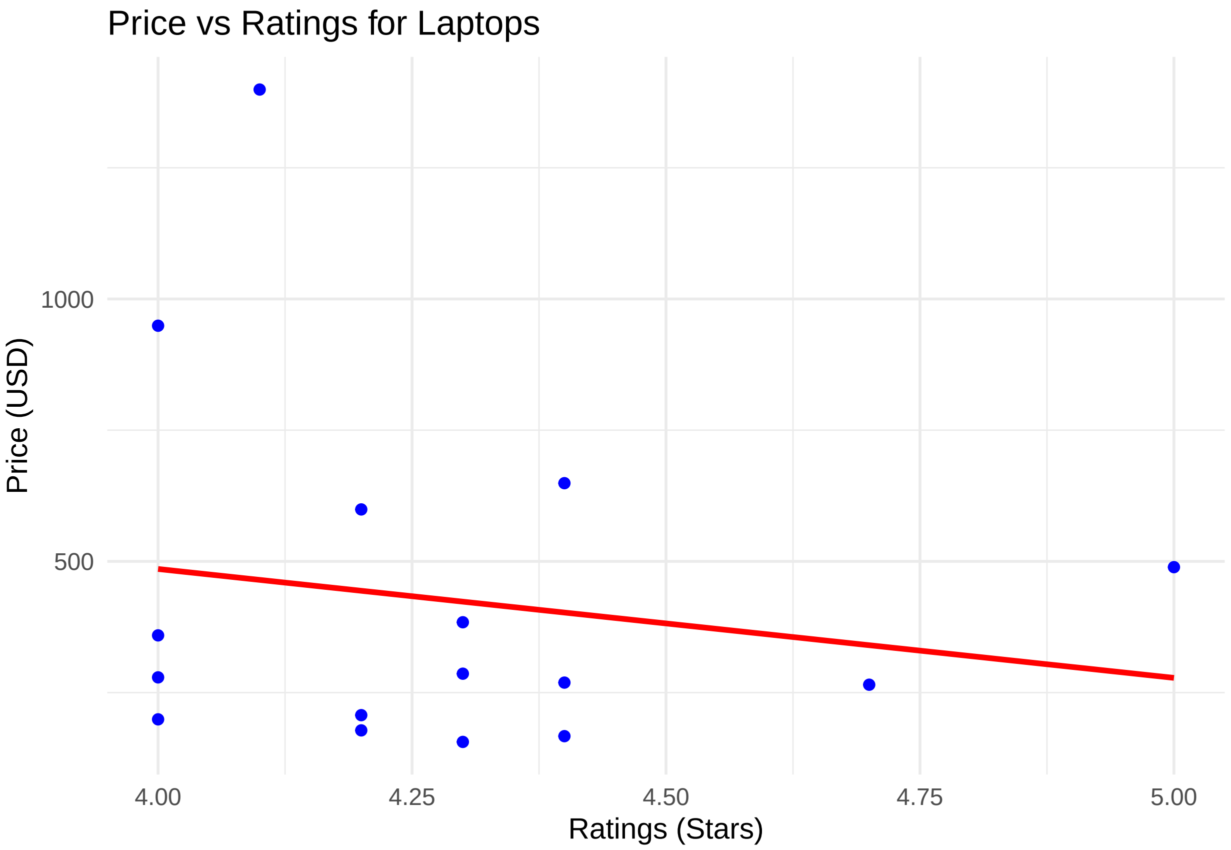
```
## Warning: Removed 84 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



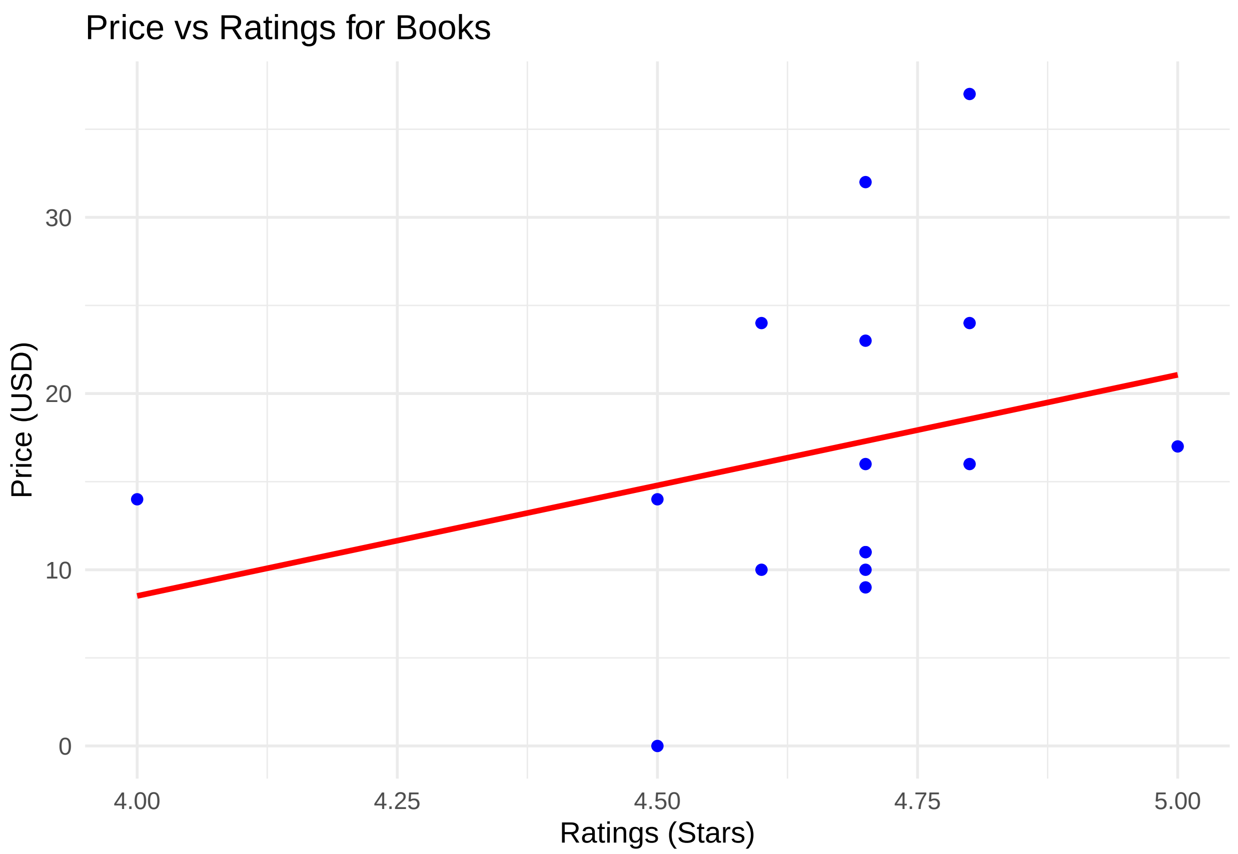
Plot price vs ratings

```r
# Plot price vs ratings with missing value handling
for (category in categories) {
data <- amazon_data[[category]]

# Remove rows with non-finite prices or ratings
data <- data %>% filter(!is.na(Price) & is.finite(Price) & !is.na(Ratings) & is.finite(Ratings))

p <- ggplot(data, aes(x = Ratings, y = Price)) +
geom_point(color = "blue") +
geom_smooth(method = "lm", se = FALSE, color = "red") +
labs(title = paste("Price vs Ratings for", category),
x = "Ratings (Stars)", y = "Price (USD)") +
theme_minimal()
print(p) # Explicitly print plot
}
```

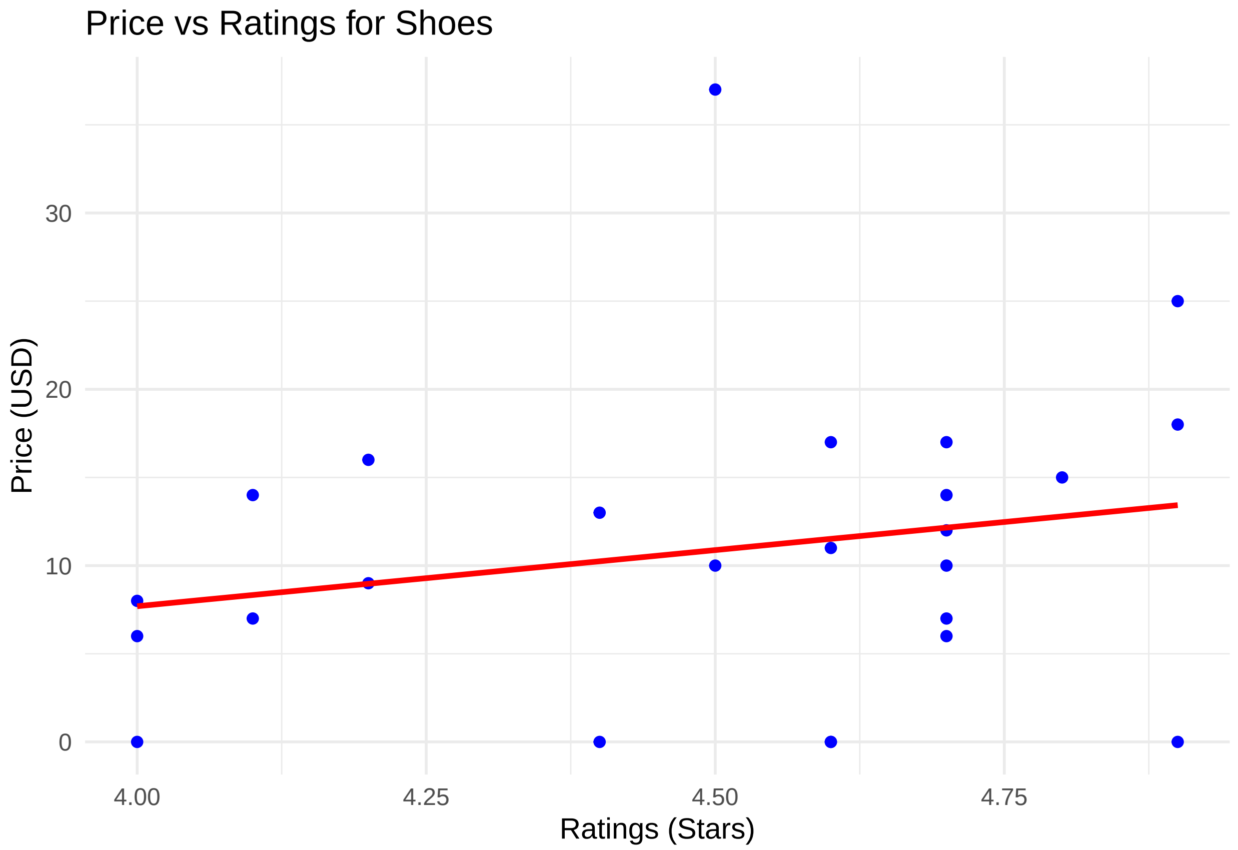
```
## `geom_smooth()` using formula = 'y ~ x'
```



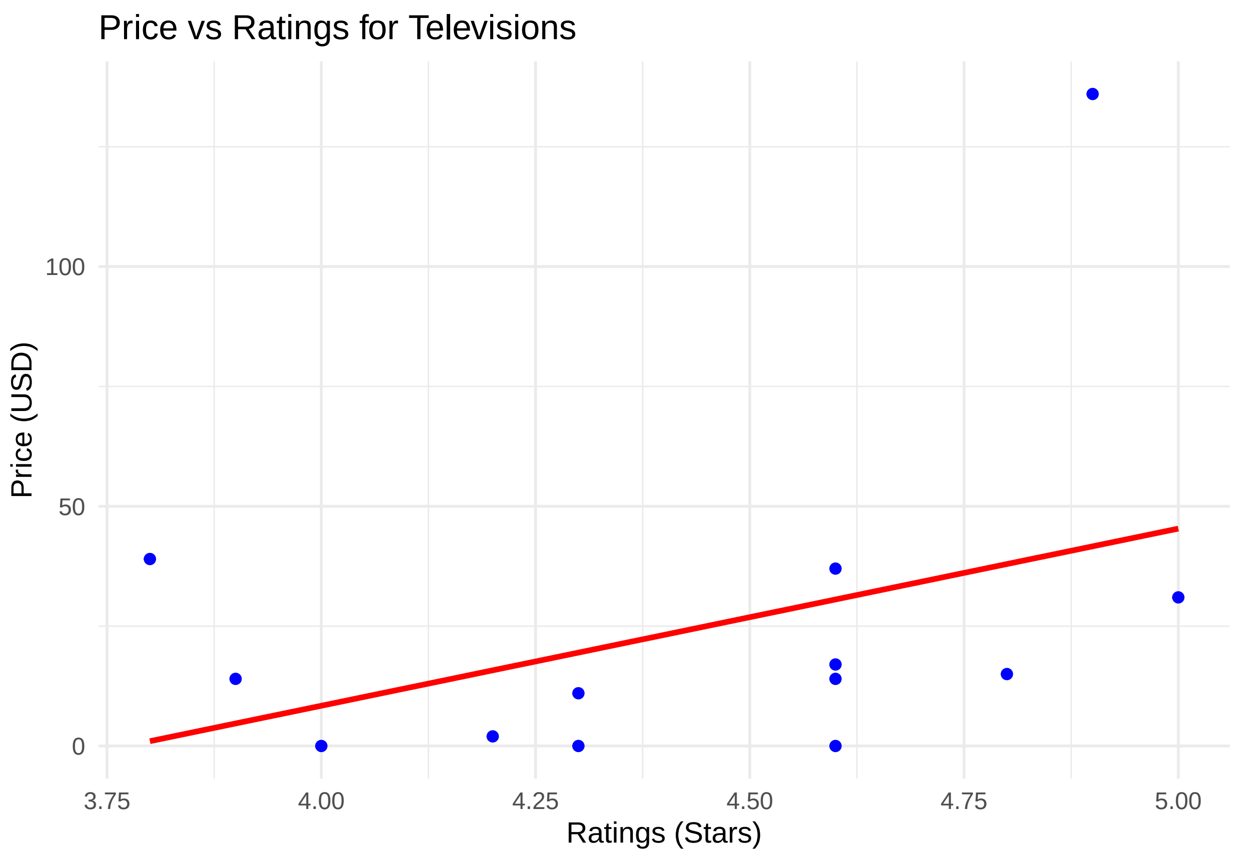
```
## `geom_smooth()` using formula = 'y ~ x'
```



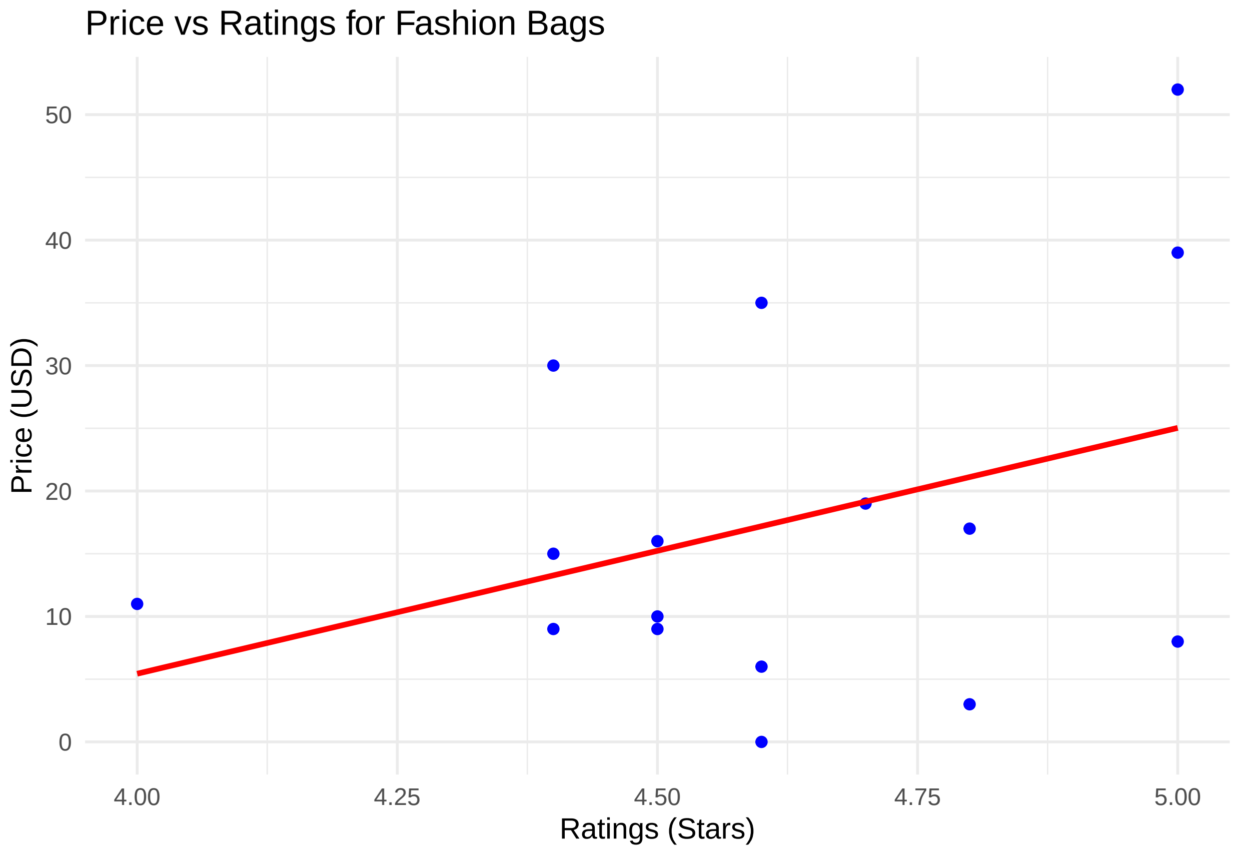
```
## `geom_smooth()` using formula = 'y ~ x'
```



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



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



```r
# 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 Dell Inspiron Touchscreen Laptop, 15.6" Business & Student Laptop Computer, Windows 11 Pro Laptop
## 2
## 3                                 HP 14" Ultral Light Laptop for Students and Business, Intel Quad-C
## 4                                                        HP Newest 14" HD Laptop, Windows 11, Intel
## 5                                                  ApoloSign 15.6" Full HD Laptop, 12GB RAM, 512GB S
##   Price Ratings Reviews Rank
## 1   489     5.0     116    1
## 2    NA     5.0    1252    2
## 3   265     4.7      81    3
## 4   167     4.4      NA    4
## 5   269     4.4    2015    5
##
## Top 5 Products in Books
##
## 1 Dog Man: Big Jim Begins: A Graphic Novel (Dog Man #13): From the Creator of Captain Underpants Book
## 2                                              Forgotten Home Apothecary : 250 Powerful Remedies at
## 3
## 4                                               Atomic Habits: An Easy & Proven Way to Build Good Ha
## 5                 Wind and Truth: Book Five of the Stormlight Archive Book 5 of 5: The Stormlight Ar
##   Price Ratings Reviews Rank
## 1    17     5.0      NA    1
## 2    16     4.8   13639    2
## 3    24     4.8      15    3
## 4    37     4.8      NA    4
## 5     9     4.7       2    5
##
## Top 5 Products in Shoes
##
## 1        1000 Sneakers: A Guide to the World's Greatest Kicks, from Sport to Street by Mathieu Le Mau
## 2                                                     Those Shoes by Maribeth Boelts  and Noah Z. Jo
## 3                                       XRHWomen's Slip on Shoes Non Slip Fashion Canvas Sneakers Lo
## 4                                             Shoe Dog: A Memoir by the Creator of Nike by Phil Knig
## 5 FeethitMens Slip On Walking Shoes Lightweight Breathable Non Slip Running Shoes Comfortable Fashio
##   Price Ratings     Reviews Rank
## 1     0     4.9 3.10000e+01    1
## 2    18     4.9 3.89200e+03    2
## 3    25     4.9 5.81900e+03    3
## 4    15     4.8 3.79938e+15    4
## 5     6     4.7 5.81900e+03    5
##
## Top 5 Products in Televisions
##
## 1                     3... 2...1... We're on the Air: An Inside Look at Sports Television, Journalis
## 2                           50 Years of Happy Days: A Visual History of an American Television Class
## 3 Behind the Screens: Illustrated Floor Plans and Scenes from the Best TV Shows of All Time by Iñaki
## 4                             Black TV: Five Decades of Groundbreaking Television from Soul Train to
## 5                                                                                                   T
##   Price Ratings Reviews Rank
## 1    31     5.0      NA    1
## 2   136     4.9      NA    2
## 3    15     4.8      NA    3
## 4     0     4.6      32    4
## 5    14     4.6     913    5
##
## Top 5 Products in Fashion Bags
##
## 1
## 2
## 3
## 4
## 5 A Fashion Coloring Book - Handbags: A coloring book for Adults and Teenagers, for stress Relief &
##   Price Ratings Reviews Rank
## 1     8     5.0      NA    1
## 2    39     5.0       8    2
## 3    52     5.0       3    3
## 4     3     4.8       3    4
## 5    17     4.8      NA    5
```