RWorksheet5(Narra,Pelaez,Saria)

Agatha Hazel D. Narra, Riza Angelique F. Pelaez, Christine Pauline Saria

library(polite) library(httr) library(rvest) library(dplyr) library(stringr) library(magrittr) library(ggplot2)

1. Extracting TV Shows

```
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) %>%
  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, Popularity
})

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

```
show_url_df
```

```
##                                          Show_Link  User_Reviews Critic_Reviews Popularity_Ratin
## 1   https://imdb.com/title/tt0903747/?ref_=chttvtp_t_1 5,086 reviews            175                2
## 2   https://imdb.com/title/tt0903747/?ref_=chttvtp_t_1 5,086 reviews            175                2
## 3   https://imdb.com/title/tt5491994/?ref_=chttvtp_t_2   158 reviews              6            1,12
## 4   https://imdb.com/title/tt5491994/?ref_=chttvtp_t_2   158 reviews              6            1,12
## 5   https://imdb.com/title/tt0795176/?ref_=chttvtp_t_3   111 reviews             10            2,01
## 6   https://imdb.com/title/tt0795176/?ref_=chttvtp_t_3   111 reviews             10            2,01
## 7   https://imdb.com/title/tt0185906/?ref_=chttvtp_t_4 1,056 reviews             34              17
## 8   https://imdb.com/title/tt0185906/?ref_=chttvtp_t_4 1,056 reviews             34              17
## 9   https://imdb.com/title/tt7366338/?ref_=chttvtp_t_5 3,530 reviews             88              17
## 10  https://imdb.com/title/tt7366338/?ref_=chttvtp_t_5 3,530 reviews             88              17
## 11  https://imdb.com/title/tt0306414/?ref_=chttvtp_t_6   787 reviews             77              10
## 12  https://imdb.com/title/tt0306414/?ref_=chttvtp_t_6   787 reviews             77              10
## 13  https://imdb.com/title/tt0417299/?ref_=chttvtp_t_7   998 reviews             57              37
## 14  https://imdb.com/title/tt0417299/?ref_=chttvtp_t_7   998 reviews             57              37
## 15  https://imdb.com/title/tt6769208/?ref_=chttvtp_t_8    53 reviews              9            4,41
## 16  https://imdb.com/title/tt6769208/?ref_=chttvtp_t_8    53 reviews              9            4,41
## 17  https://imdb.com/title/tt0141842/?ref_=chttvtp_t_9   962 reviews             93               3
## 18  https://imdb.com/title/tt0141842/?ref_=chttvtp_t_9   962 reviews             93               3
## 19 https://imdb.com/title/tt2395695/?ref_=chttvtp_t_10   205 reviews             12            1,49
## 20 https://imdb.com/title/tt2395695/?ref_=chttvtp_t_10   205 reviews             12            1,49
## 21 https://imdb.com/title/tt0081846/?ref_=chttvtp_t_11    80 reviews              8            3,86
## 22 https://imdb.com/title/tt0081846/?ref_=chttvtp_t_11    80 reviews              8            3,86
## 23 https://imdb.com/title/tt9253866/?ref_=chttvtp_t_12   245 reviews             15            2,76
## 24 https://imdb.com/title/tt9253866/?ref_=chttvtp_t_12   245 reviews             15            2,76
## 25 https://imdb.com/title/tt0944947/?ref_=chttvtp_t_13 5,899 reviews            368               1
## 26 https://imdb.com/title/tt0944947/?ref_=chttvtp_t_13 5,899 reviews            368               1
## 27 https://imdb.com/title/tt7678620/?ref_=chttvtp_t_14   367 reviews              4              41
## 28 https://imdb.com/title/tt7678620/?ref_=chttvtp_t_14   367 reviews              4              41
## 29 https://imdb.com/title/tt0071075/?ref_=chttvtp_t_15   126 reviews              5            2,62
## 30 https://imdb.com/title/tt0071075/?ref_=chttvtp_t_15   126 reviews              5            2,62
## 31 https://imdb.com/title/tt1355642/?ref_=chttvtp_t_16   466 reviews             16              50
## 32 https://imdb.com/title/tt1355642/?ref_=chttvtp_t_16   466 reviews             16              50
## 33 https://imdb.com/title/tt2861424/?ref_=chttvtp_t_17   909 reviews             94              13
## 34 https://imdb.com/title/tt2861424/?ref_=chttvtp_t_17   909 reviews             94              13
## 35 https://imdb.com/title/tt1533395/?ref_=chttvtp_t_18    12 reviews              9            3,45
## 36 https://imdb.com/title/tt1533395/?ref_=chttvtp_t_18    12 reviews              9            3,45
## 37 https://imdb.com/title/tt8420184/?ref_=chttvtp_t_19   541 reviews             28            1,52
## 38 https://imdb.com/title/tt8420184/?ref_=chttvtp_t_19   541 reviews             28            1,52
## 39 https://imdb.com/title/tt0052520/?ref_=chttvtp_t_20   213 reviews             85              35
## 40 https://imdb.com/title/tt0052520/?ref_=chttvtp_t_20   213 reviews             85              35
## 41 https://imdb.com/title/tt1877514/?ref_=chttvtp_t_21   175 reviews             13            2,02
## 42 https://imdb.com/title/tt1877514/?ref_=chttvtp_t_21   175 reviews             13            2,02
## 43 https://imdb.com/title/tt1475582/?ref_=chttvtp_t_22 1,095 reviews            121              17
## 44 https://imdb.com/title/tt1475582/?ref_=chttvtp_t_22 1,095 reviews            121              17
## 45 https://imdb.com/title/tt2560140/?ref_=chttvtp_t_23 2,359 reviews             64               6
## 46 https://imdb.com/title/tt2560140/?ref_=chttvtp_t_23 2,359 reviews             64               6
## 47 https://imdb.com/title/tt0103359/?ref_=chttvtp_t_24   219 reviews             25              52
## 48 https://imdb.com/title/tt0103359/?ref_=chttvtp_t_24   219 reviews             25              52
## 49 https://imdb.com/title/tt0386676/?ref_=chttvtp_t_25 1,774 reviews             76               5
## 50 https://imdb.com/title/tt0386676/?ref_=chttvtp_t_25 1,774 reviews             76               5
```

```r
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  https://imdb.com/title/tt
## 2     2                  Planet Earth II    9.5   162K        6      2016  https://imdb.com/title/tt
## 3     3                     Planet Earth    9.4   223K       11      2006  https://imdb.com/title/tt
## 4     4                 Band of Brothers    9.4   545K       10      2001  https://imdb.com/title/tt
## 5     5                        Chernobyl    9.3   906K        5      2019  https://imdb.com/title/tt
## 6     6                         The Wire    9.3   390K       60 2002-2008  https://imdb.com/title/tt
## 7     7       Avatar: The Last Airbender    9.3   389K       62 2005-2008  https://imdb.com/title/tt
## 8     8                   Blue Planet II    9.3    48K        7      2017  https://imdb.com/title/tt
## 9     9                     The Sopranos    9.2   497K       86 1999-2007  https://imdb.com/title/tt
## 10   10      Cosmos: A Spacetime Odyssey    9.2   131K       13      2014  https://imdb.com/title/tt
## 11   11                           Cosmos    9.3    45K       13      1980  https://imdb.com/title/tt
## 12   12                       Our Planet    9.2    53K       12 2019-2023  https://imdb.com/title/tt
## 13   13                 Game of Thrones    9.2   2.4M       74 2011-2019  https://imdb.com/title/tt
## 14   14                            Bluey    9.3    33K      194    2018-   https://imdb.com/title/tt
## 15   15                The World at War    9.2    31K       26 1973-1974  https://imdb.com/title/tt
## 16   16 Fullmetal Alchemist: Brotherhood    9.1   208K       68 2009-2010  https://imdb.com/title/tt
## 17   17                  Rick and Morty    9.1   626K       78    2013-   https://imdb.com/title/tt
## 18   18                            Life    9.1    43K       11      2009  https://imdb.com/title/tt
## 19   19                  The Last Dance    9.1   159K       10      2020 https://imdb.com/title/tt23
## 20   20               The Twilight Zone    9.0    96K      156 1959-1964 https://imdb.com/title/tt23
## 21   21                 The Vietnam War    9.1    29K       10      2017 https://imdb.com/title/tt0
## 22   22                        Sherlock    9.1     1M       15 2010-2017 https://imdb.com/title/tt0
## 23   23                  Attack on Titan    9.1   560K       98 2013-2023 https://imdb.com/title/tt9
## 24   24      Batman: The Animated Series    9.0   122K       85 1992-1995 https://imdb.com/title/tt9
## 25   25                      The Office    9.0   745K      188 2005-2013 https://imdb.com/title/tt0
## 26    1                     Breaking Bad    9.5   2.2M       62 2008-2013 https://imdb.com/title/tt0
## 27    2                  Planet Earth II    9.5   162K        6      2016 https://imdb.com/title/tt7
## 28    3                     Planet Earth    9.4   223K       11      2006 https://imdb.com/title/tt7
## 29    4                 Band of Brothers    9.4   545K       10      2001 https://imdb.com/title/tt0
## 30    5                        Chernobyl    9.3   906K        5      2019 https://imdb.com/title/tt0
## 31    6                         The Wire    9.3   390K       60 2002-2008 https://imdb.com/title/tt1
## 32    7       Avatar: The Last Airbender    9.3   389K       62 2005-2008 https://imdb.com/title/tt1
## 33    8                   Blue Planet II    9.3    48K        7      2017 https://imdb.com/title/tt2
## 34    9                     The Sopranos    9.2   497K       86 1999-2007 https://imdb.com/title/tt2
## 35   10      Cosmos: A Spacetime Odyssey    9.2   131K       13      2014 https://imdb.com/title/tt1
## 36   11                           Cosmos    9.3    45K       13      1980 https://imdb.com/title/tt1
## 37   12                       Our Planet    9.2    53K       12 2019-2023 https://imdb.com/title/tt8
## 38   13                 Game of Thrones    9.2   2.4M       74 2011-2019 https://imdb.com/title/tt8
## 39   14                            Bluey    9.3    33K      194    2018-  https://imdb.com/title/tt0
## 40   15                The World at War    9.2    31K       26 1973-1974 https://imdb.com/title/tt0
## 41   16 Fullmetal Alchemist: Brotherhood    9.1   208K       68 2009-2010 https://imdb.com/title/tt1
## 42   17                  Rick and Morty    9.1   626K       78    2013-  https://imdb.com/title/tt1
## 43   18                            Life    9.1    43K       11      2009 https://imdb.com/title/tt1
## 44   19                  The Last Dance    9.1   159K       10      2020 https://imdb.com/title/tt1
## 45   20               The Twilight Zone    9.0    96K      156 1959-1964 https://imdb.com/title/tt2
## 46   21                 The Vietnam War    9.1    29K       10      2017 https://imdb.com/title/tt2
## 47   22                        Sherlock    9.1     1M       15 2010-2017 https://imdb.com/title/tt0
## 48   23                  Attack on Titan    9.1   560K       98 2013-2023 https://imdb.com/title/tt0
## 49   24      Batman: The Animated Series    9.0   122K       85 1992-1995 https://imdb.com/title/tt0
## 50   25                      The Office    9.0   745K      188 2005-2013 https://imdb.com/title/tt0
```

```
##    Popularity_Rating
## 1                 20
## 2                 20
## 3              1,121
## 4              1,121
## 5              2,011
## 6              2,011
## 7                171
## 8                171
## 9                173
## 10               173
## 11               108
## 12               108
## 13               373
## 14               373
## 15             4,415
## 16             4,415
## 17                33
## 18                33
## 19             1,499
## 20             1,499
## 21             3,866
## 22             3,866
## 23             2,765
## 24             2,765
## 25                14
## 26                14
## 27               411
## 28               411
## 29             2,627
## 30             2,627
## 31               508
## 32               508
## 33               137
## 34               137
## 35             3,455
## 36             3,455
## 37             1,521
## 38             1,521
## 39               354
## 40               354
## 41             2,022
## 42             2,022
## 43               172
## 44               172
## 45                60
## 46                60
## 47               527
## 48               527
## 49                55
## 50                55
```

```
#2.
# Define URL for Breaking Bad
```

```r
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)

# 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]
```

```r
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           7
## 12          Permalink Apr 22, 2020           2
## 13           xpinerhd Sep 22, 2018          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                                                                                                    Slightly overrat
## 14                                                                                                  So boring I died t
## 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 C
## 20                                                                                                         Every bit a
##
## 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 ever
## 11
## 12
## 13
## 14
## 15
## 16
## 17
## 18
## 19
## 20
```

```
# 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,
```

```r
  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                                                    Yo
## 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 Like the first 'Planet Earth', does for nature and our
## 18          Permalink Oct 20, 2018          10
## 19 panagiotiskatsanos Apr 23, 2020          10
## 20          Permalink  Jan 5, 2017          10
##    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
```

```r
# 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
```

```r
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"
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                                        A masterpiece of a
## 4        Permalink Dec 15, 2006          10                                                    In A V
## 5  ccthemovieman-1  Sep 1, 2007          10     The most amazing achievement in natural history TV ha
## 6        Permalink Aug 27, 2006          10                                                  Simply p
## 7         cmcoveos Apr 30, 2006          10                              An amazing trip around our beau
## 8        Permalink Jun 29, 2015           9 A visually impressive and memorable look at the world tha
```

```
## 9         Loordssm Jul 20, 2006          10                                Is it real? I mean, actu
## 10       Permalink Jan 28, 2009          10
## 11        ultimorn  Jun 1, 2015           7                                          Are you kiddi
## 12       Permalink  Oct 8, 2020           3                                It doesn't get any bett
## 13     bob the moo  Dec 4, 2007          10                                       Only 4 Eps can t
## 14       Permalink Jan 15, 2007          10                      Should be called "BBC - Yeah, a
## 15           alfeu Jul 30, 2008          10                                  Brilliant Docume
## 16       Permalink Dec 25, 2017           9                       Explanation to those low-ratin
## 17         Cabrone Sep 14, 2009          10                                              Truly
## 18       Permalink May 31, 2020           9                                    The Greatest
## 19        berndt65 Jul 27, 2014          10
## 20       Permalink  Jan 4, 2023          10                                   Absolutely
##
## 1
## 2
## 3
## 4
## 5
## 6
## 7  As the influence of man expands across the globe, fewer and fewer truly untouched wilderness exist
## 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() %>%
```

```r
  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
)
```

```r
# 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                          Possibly the finest 10 hours
## 4           Permalink Apr 18, 2004          10                             One of the best war movies
## 5         wildcatt268 Feb 13, 2003          10
## 6           Permalink Jan 23, 2005          10
## 7             arjay24 Sep 16, 2004          10                    One of, if not the best, mini serie
## 8           Permalink  May 6, 2022          10           This series is so unbelievably realistic, s
## 9           rbverhoef  Nov 4, 2019          10                          One of the best mini-series
## 10          Permalink  Nov 5, 2001          10                                              Probably
## 11         yodaschoda Aug 25, 2004          10                           Realistic WWII Drama With Wa
## 12          Permalink May 30, 2015           7                                                   wa
## 13 philip_vanderveken Apr 10, 2021           5                                         You can't l
## 14          Permalink  May 2, 2006          10
## 15     Supermanfan-13  Jun 3, 2019          10                                       Not very real
## 16          Permalink Jan 26, 2005          10                 Without Doubt, the Best Mini-Series
## 17          thiagoutp  May 3, 2022          10                                                 Grea
## 18          Permalink Oct 24, 2018           9 A series like this won't be made again (see below), so
## 19         bsmith5552  Dec 7, 2002          10                                        Share With
## 20          Permalink Nov 25, 2002          10                                           Best Mini
##
## 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 sti
## 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()
```

```r
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)

# 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]
```

```r
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                                                    They
## 3           stelmakh  May 9, 2019          10                                                  Goosebu
## 4          Permalink May 14, 2019          10                                       I highly recomme
## 5       natashapekar  May 7, 2019          10                                     No hero wakes up w
## 6          Permalink May 20, 2019          10                                              So far le
## 7         m-porpaczi  May 6, 2019          10
## 8          Permalink May 13, 2019          10                          Bleak, Unsettling, Haunting A
## 9           Lladerat  May 6, 2019          10
## 10         Permalink Nov 27, 2019          10                                                  HBO
## 11          jfirebug May 23, 2019           7
## 12         Permalink Jan 27, 2024           5
## 13           thegldt Jun 15, 2019           8                       Unveiling Human Errors and Poli
## 14         Permalink May 20, 2019          10 The movie is far from thuth. A lot of fake info to crea
## 15 alexander-phoenix May 30, 2019          10                                              Emotional
## 16         Permalink  Jun 7, 2019          10                                                   Just
## 17       wmeduardowm  May 6, 2019           9                          Now you look like the mini
## 18         Permalink Sep 27, 2022           9
## 19    Leofwine_draca May 26, 2019           9
## 20         Permalink Jul 10, 2022          10              It is hard to overestimate the importance
##
## 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
```

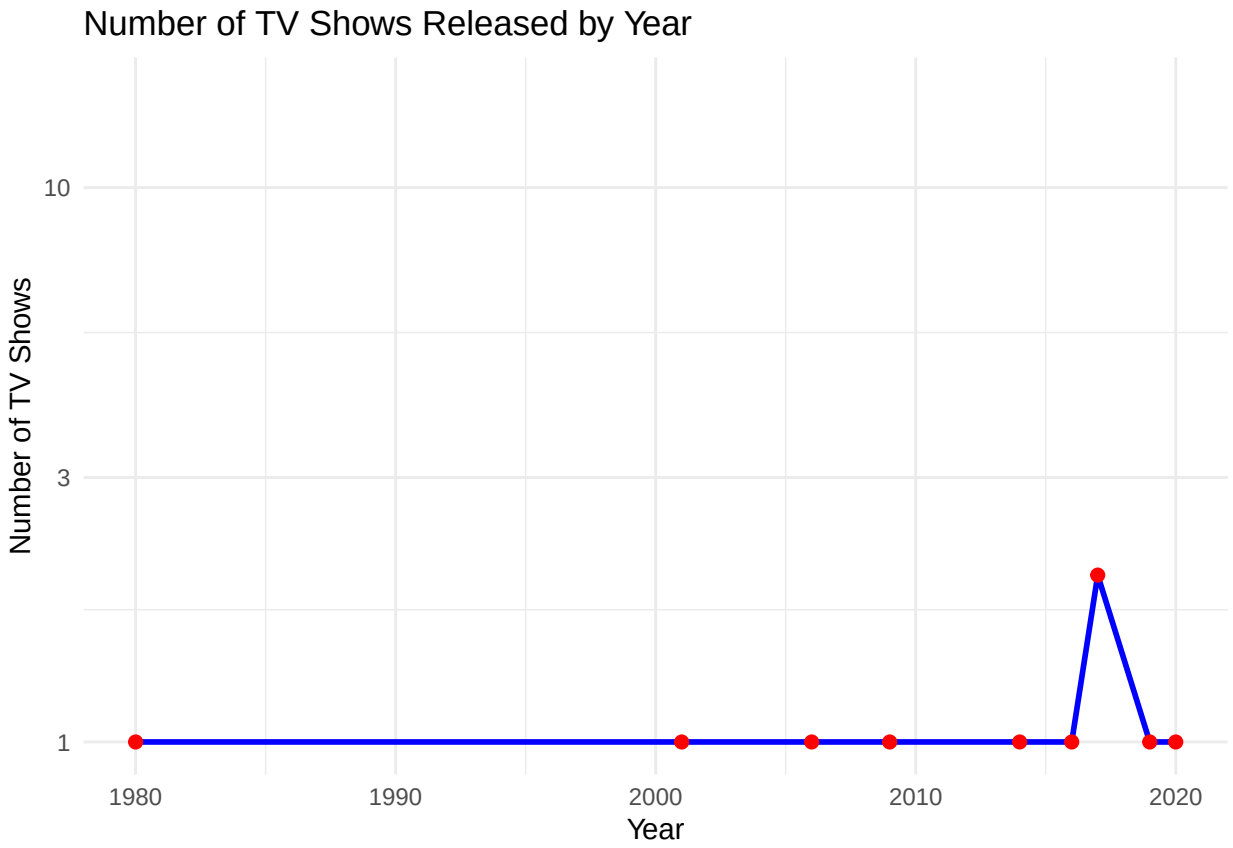
```
## 12
## 13
## 14
## 15
## 16
## 17
## 18
## 19
## 20
```

```
#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: 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() %>%
    head(30)


  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(30)


  dfTemp <- data.frame(Product_Name = product_name[1:30],
                       Description = product_description[1:30],
                       Rating = ratings[1:30],
                       Price = price[1:30],
                       stringsAsFactors = FALSE)

  df[[i]] <- dfTemp
}

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

```
##
## 1                                      JanSport SuperBreak One Backpacks - Durable, Lightweight Boo
## 2  MATEIN Travel Laptop Backpack, Business Anti Theft Slim Sturdy Laptops Backpack with USB Charging
## 3                                                    JanSport Laptop Backpack - Computer Bag wi
## 4    Taygeer Travel Backpack for Women, Carry On Backpack with USB Charging Port & Shoe Pouch, TSA 15
## 5
## 6       YOREPEK Travel Backpack, Extra Large 50L Laptop Backpacks for Men Women, Water Resistant Coll
## 7                                                  JanSport Right Pack Backpack - Durable Daypa
## 8
## 9         Laptop Backpack,Business Travel Anti Theft Slim Durable Laptops Backpack with USB Charging
## 10                                                                                                ZO
## 11
```

```
## 12
## 13
## 14                                                                                                THE N
## 15                                                                                  Amazon Basics Transparent Scho
## 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 44.93
## 6         <NA>    4.7 55.00
## 7         <NA>    3.7 23.99
## 8         <NA>    4.7  8.99
## 9         <NA>    4.7 12.00
## 10        <NA>    4.6 27.99
## 11        <NA>    4.6 39.99
## 12        <NA>    4.6 44.00
## 13        <NA>    4.8 64.99
## 14        <NA>    4.7 23.99
## 15        <NA>    4.7 29.99
## 16        <NA>    4.7 23.99
## 17        <NA>    4.6 99.00
## 18        <NA>    4.8 65.00
## 19        <NA>     NA 66.20
## 20        <NA>     NA 70.00
## 21        <NA>     NA 80.00
## 22        <NA>     NA 16.49
## 23        <NA>     NA 38.91
## 24        <NA>     NA 60.00
## 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-Co
```

```
## 2                         HP Portable Laptop, Student and Business, 14" HD Display, Intel Quad-Core I
## 3              HP Newest 255 G10 Laptop for Home or Work, 16GB RAM, 1TB SSD, 15.6" Full HD, Ryzen 3 73
## 4
## 5        HP 14 Laptop, Intel Celeron N4020, 4 GB RAM, 64 GB Storage, 14-inch Micro-edge HD Display, Wi
## 6            HP Newest 14" Ultral Light Laptop for Students and Business, Intel Quad-Core N4120, 8GB I
## 7       Lenovo IdeaPad 1 Student Laptop, Intel Dual Core Processor, 20GB RAM, 1TB SSD + 128GB eMMC, 15
## 8                                                                HP 17 Laptop, 17.3" HD+ Display, 11t
## 9                                  15.6 FHD Student Laptop Computer, 16GB RAM 1TB SSD, Backlit Keyboard
## 10              HP 15.6" Portable Laptop (Include 1 Year Microsoft 365), HD Display, Intel Quad-Core I
## 11                               HP 15.6 FHD Display G9 Laptop • 32GB RAM • 1TB Storage (512GB SSD & 50G
## 12 HP 2024 Newest Laptop for Business and Student, 15.6" HD Touchscreen, Intel 6-Core i3-1215U Proces
## 13        HP 15.6" Business & Student Laptop Computer, Intel Core i5, Windows 11 Pro Laptop with Micro
## 14                                                                          HP Probook X360 11 G2 1
## 15                                                                       Lenovo 14 FHD Laptop - AMD
## 16              HP Newest Pavilion 15.6" Touchscreen Laptop • Intel 12th-Gen 6 Core Processor • 40GB RA
## 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>    4.2 321.99
## 3         <NA>    4.4 197.36
## 4         <NA>    4.1 269.00
## 5         <NA>    4.4 399.99
## 6         <NA>    4.1 599.00
## 7         <NA>    4.0 147.68
## 8         <NA>    4.0 176.00
## 9         <NA>    4.1 209.99
## 10        <NA>    4.2 243.99
## 11        <NA>    5.0 399.00
## 12        <NA>    4.1 475.00
## 13        <NA>    4.4 389.00
## 14        <NA>    4.3 439.00
## 15        <NA>    4.1 279.90
## 16        <NA>    3.4 419.58
## 17        <NA>    2.7 499.00
## 18        <NA>    4.8 399.00
## 19        <NA>     NA 449.00
## 20        <NA>     NA 689.99
## 21        <NA>     NA 134.99
## 22        <NA>     NA 154.99
## 23        <NA>     NA 114.99
## 24        <NA>     NA 145.00
```

```
## 25        <NA>     NA 484.00
## 26        <NA>     NA     NA
## 27        <NA>     NA     NA
## 28        <NA>     NA     NA
## 29        <NA>     NA     NA
## 30        <NA>     NA     NA
```

```
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, iPh
## 4                                                                              ORNARTO Compatible with iPhone
## 5                                                                                          OtterBox
## 6
## 7                                                                                                  Otte
## 8                                                                                        OtterBox iPho
## 9       OtterBox iPhone SE 3rd & 2nd Gen, iPhone 8 & iPhone 7 (not Compatible with Plus Sized Models)
## 10                                                                          FABSPARK Phone Case for iPhone 13
## 11               elago Liquid Silicone Case Compatible with iPhone 13 Pro Case (6.1"), Premium Silic
## 12
## 13                           elago Compatible with iPhone 14 Pro Case, Liquid Silicone Case, Full Bo
## 14               ESR for iPhone 14 Case/iPhone 13 Case, Military-Grade Protection, Shockproof Air-Gua
## 15                     Temdan for iPhone 16 Pro Max Case Clear, [Compatible with Magsafe][Anti-Yellow
## 16                       elago Compatible with iPhone 14 Pro Max Case, Liquid Silicone Case, Full Bo
## 17
## 18
## 19
## 20
## 21
## 22
## 23
## 24
## 25
## 26
## 27
## 28
## 29
## 30
##    Description Rating Price
## 1         <NA>    3.7 14.99
## 2         <NA>    3.6 30.99
## 3         <NA>    3.7 12.98
## 4         <NA>    4.6 35.19
## 5         <NA>    4.5 39.95
## 6         <NA>    4.6  9.99
## 7         <NA>    4.5 20.99
## 8         <NA>    4.7 34.63
## 9         <NA>    3.9 39.95
## 10        <NA>    4.7 44.95
## 11        <NA>    4.7 29.95
## 12        <NA>    4.7 27.09
## 13        <NA>    4.3 39.95
## 14        <NA>    4.4 27.25
```

```
## 15        <NA>    4.7 49.95
## 16        <NA>    4.5  9.99
## 17        <NA>    4.6 12.99
## 18        <NA>    4.6 31.99
## 19        <NA>    4.5 59.95
## 20        <NA>     NA 12.99
## 21        <NA>     NA  9.99
## 22        <NA>     NA  8.99
## 23        <NA>     NA 12.99
## 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                                        Mongoose Flatrock 21-Speed Hardtail Mountain Bike, 24 to 1
## 3                              Mongoose Malus Mens and Women Fat Tire Mountain Bike, 26-Inch Bicycle Wh
## 4   Schwinn Traxion Mountain Bike for Adult Men Women, 29-Inch Wheels, Full Suspension, 24-Speed Shim
## 5                                       Schwinn High Timber Mountain Bike for Adult Youth Men Women
## 6                                    Dynacraft Magna Echo Ridge Mountain Bike - Rugged and Durable Desi
## 7                                                                    Schwinn Bonafide Men and Wom
## 8                               Grafton Mountain Bike for Adult and Youth Men and Women, 24/26 / 27.5-I
## 9                                          26/27.5" Mountain Bike 21 Speed Bikes for Adults, Men & Women
## 10                     Huffy Stone Mountain Hardtail Mountain Bike for Boys/Girls/Men/Women, 20"/24"/20
## 11    isinwheel M10 Electric Bike Adult 500W, 26" Commuting Electric Mountain Bike 20MPH Max Range 55
## 12                                                 Dynacraft Vertical Alpine Eagle Mountain Bike - Rug
## 13                        Mongoose Impasse Full Suspension Mountain Bike, Men and Women, 18-Inch Alumin
## 14                                 Mongoose Salvo Comp, Trail, or Sport Mountain Bike for Adult Men and
## 15                                                                                 Hiland Full Susp
## 16                       Dynacraft Air Zone Aftershock 20" Mountain Bike - Rugged and Durable Design, I
## 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>    4.1 199.99
## 3         <NA>    4.4 309.99
## 4         <NA>    4.3 492.99
```

```
## 5         <NA>    4.1 519.99
## 6         <NA>    4.0 637.01
## 7         <NA>    4.2 674.99
## 8         <NA>    3.5 345.99
## 9         <NA>    2.5 499.99
## 10        <NA>    3.7 158.94
## 11        <NA>    4.3 169.99
## 12        <NA>    4.3 612.53
## 13        <NA>    4.0 349.99
## 14        <NA>    4.1  99.99
## 15        <NA>    4.2 199.99
## 16        <NA>    4.2 249.99
## 17        <NA>     NA 389.99
## 18        <NA>     NA 485.71
## 19        <NA>     NA 549.99
## 20        <NA>     NA 899.99
## 21        <NA>     NA 289.99
## 22        <NA>     NA 149.99
## 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
```

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

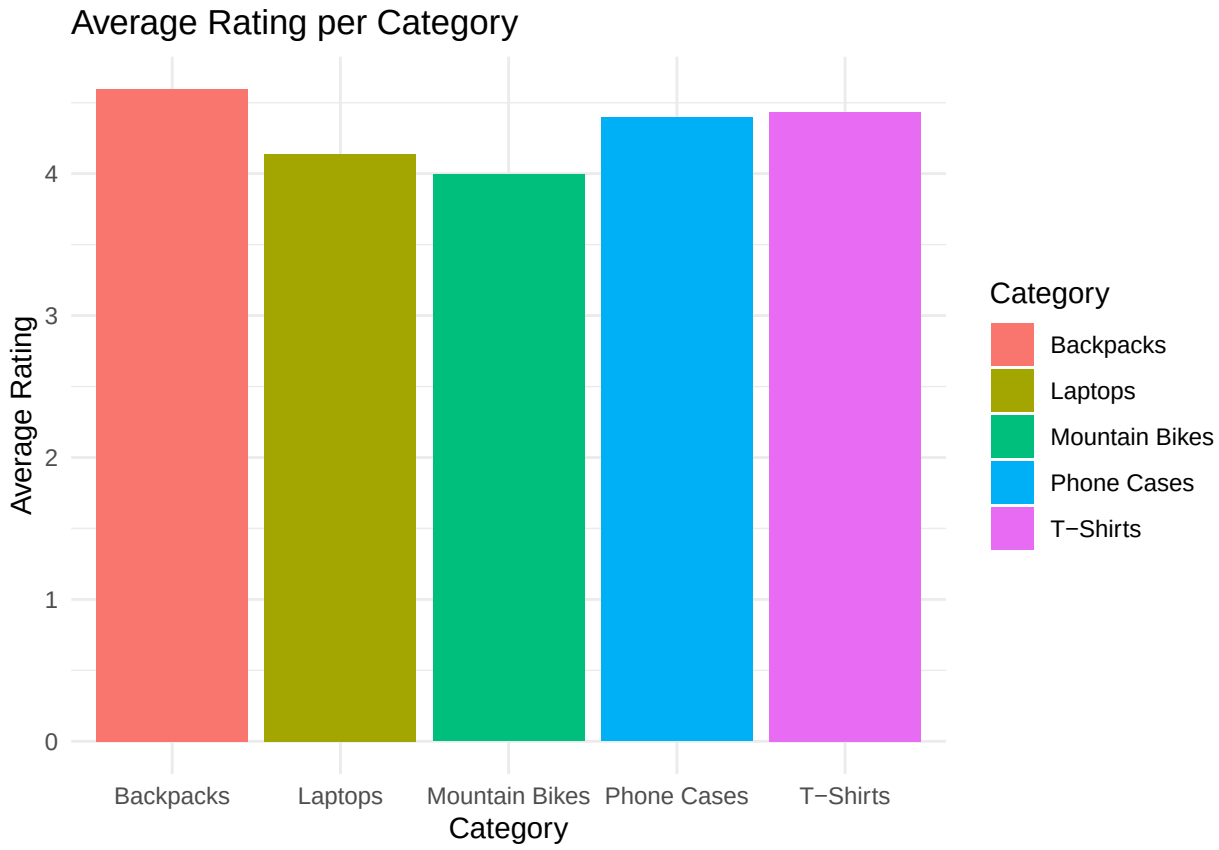
```
##                                                                                                 Produ
## 1
## 2                                                              Men's Crew T-Shirts, Multipack, Style 
## 3                                                                                            Fruit of t
## 4                                                            Men's Eversoft Cotton Stay Tucked Crew T-
## 5
## 6          Men's Pocket Undershirt Pack, Cotton Crew Neck T-Shirt, Moisture Wicking Tee, Assorted 6
## 7
## 8           Men's T-Shirt, Beefy-T Heavyweight Cotton Crewneck Tee, 1 or 2 Pack, Available in Tall 
## 9                                                                                            Fruit of t
## 10                                                         Men's Eversoft Cotton Stay Tucked V-Neck T-
## 11
## 12                                                  Unisex Adult Ultra Cotton T-Shirt, Style G2000, Mult
## 13
## 14                             Men's Cotton, Moisture-Wicking Crew Tee Undershirts, Multi-Packs Avai
## 15                                                                                               True 
## 16                                                     6 Pack, Men's Short Sleeve Crew Neck T-Shirt, S 
## 17                                                                                             Russell A
## 18 Men's Dri-Power Cotton Blend Short Sleeve Tees, Moisture Wicking, Odor Protection, UPF 30+, Sizes 
## 19                                                                                              Comfort
## 20                                                     Adult Heavyweight Short Sleeve Tee, Style G1717/
## 21                                                                                                   C
## 22                                                   Men's Loose Fit Heavyweight Short-Sleeve Pocket T-
## 23
## 24                        Men's Short Sleeve T-Shirt Pack, Essentials Crewneck Cotton T-Shirt, 4 or 6
## 25
```

```
## 26                                                    Men's V-Neck T-Shirts, Multipack, Style
## 27                                                                                        True
## 28                                                3 Pack, Men's Short Sleeve Crew Neck T-Shirt, S
## 29
## 30          Workout Shirts for Men Short Sleeve Quick Dry Athletic Gym Active T Shirt Moisture Wi
```

```r
#6.

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

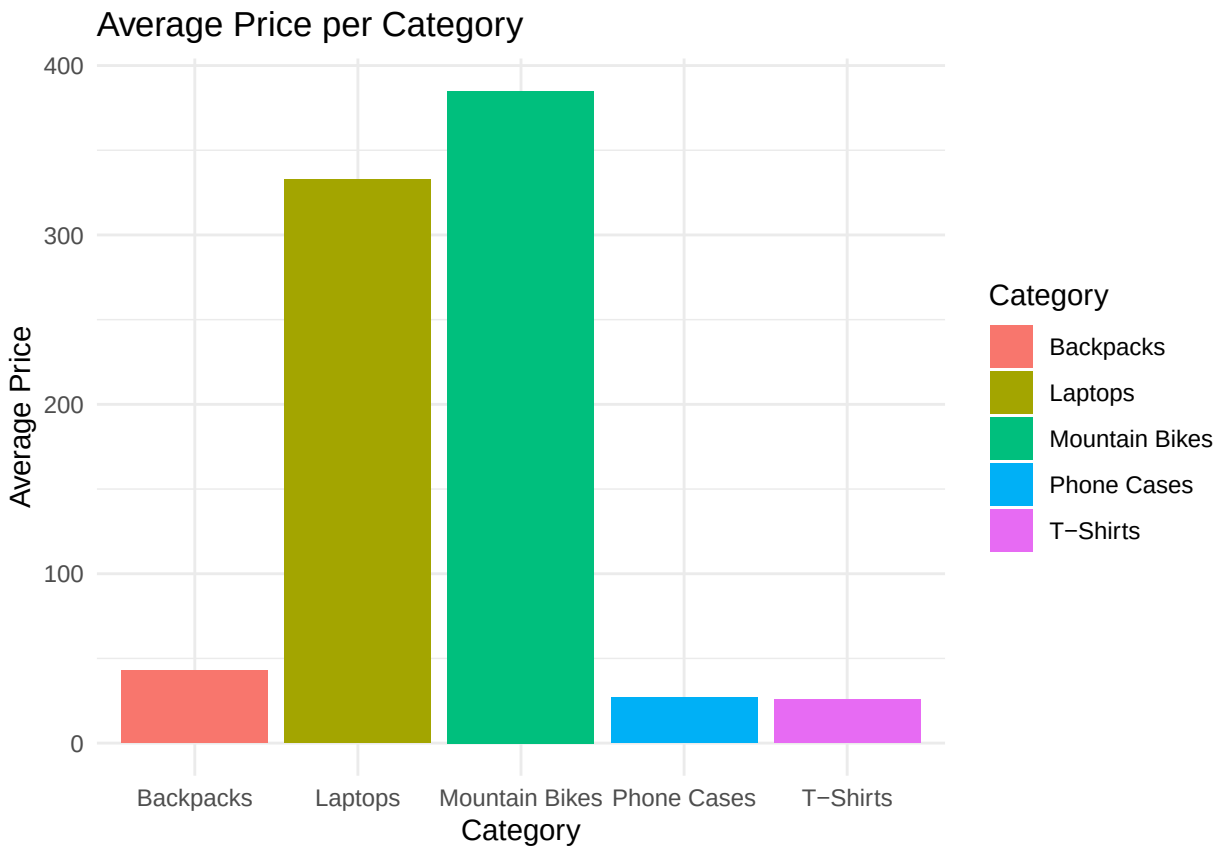
```r
#7

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

```r
#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()
```



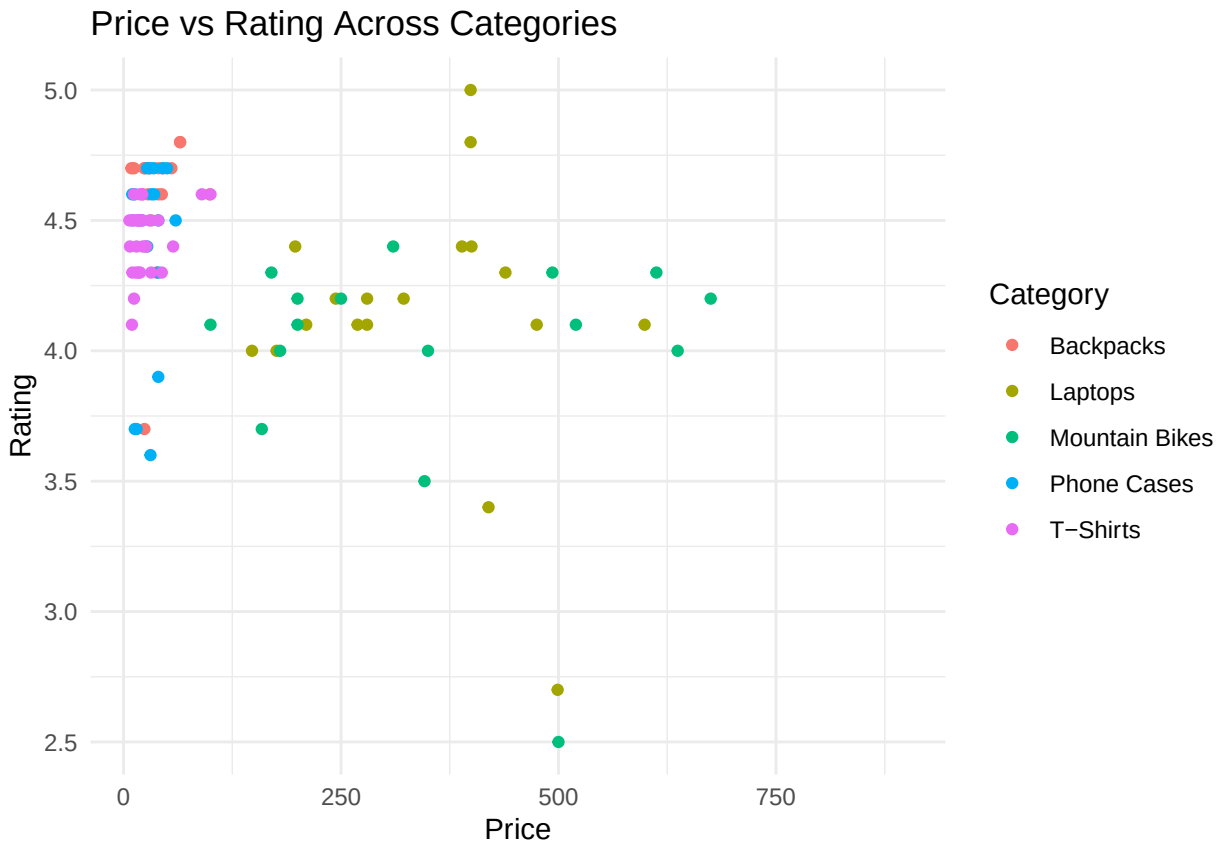

```r
avg_price <- combined_df %>%
```

```
  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()
```



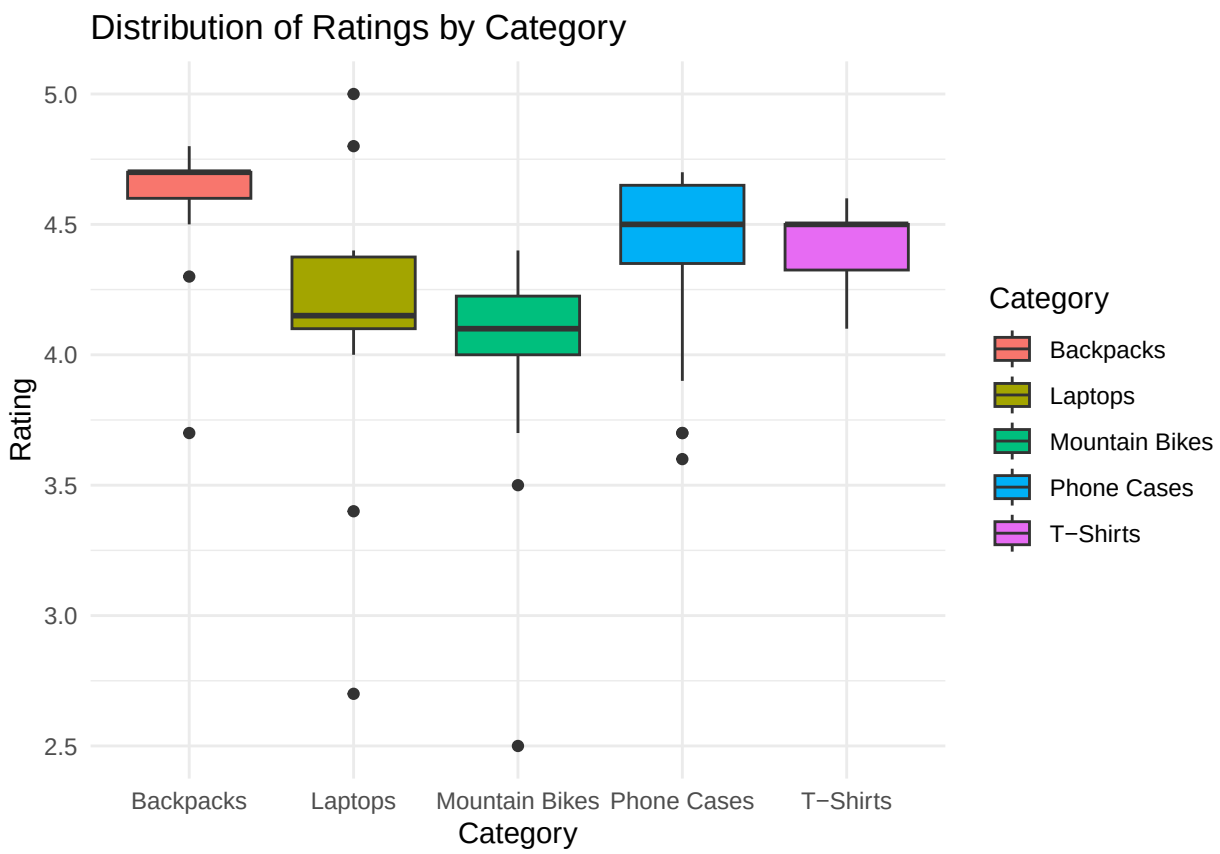

```
ggplot(combined_df, aes(x = Price, y = Rating, color = Category)) +
  geom_point() +
  labs(title = "Price vs Rating Across Categories", x = "Price", y = "Rating") +
  theme_minimal()
```

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



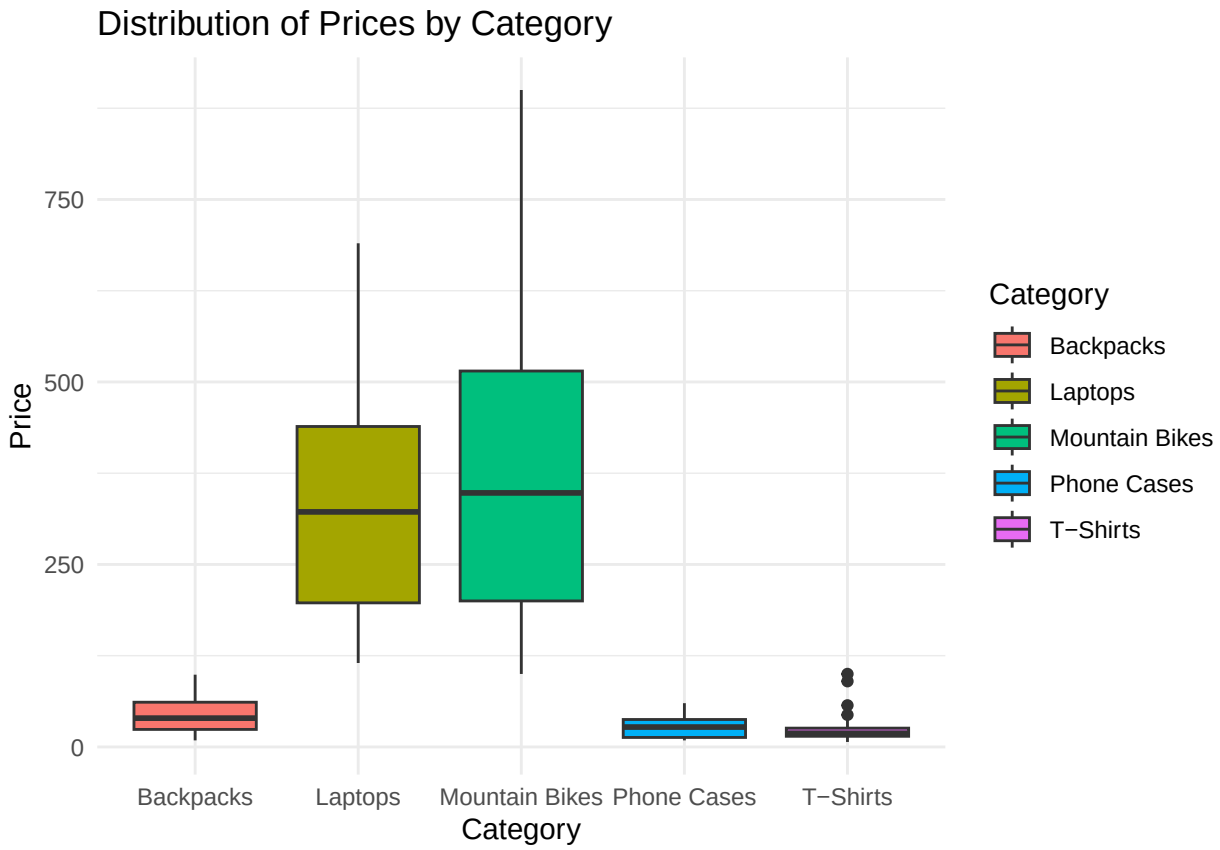

```
#9
ggplot(combined_df, aes(x = Category, y = Rating, fill = Category)) +
  geom_boxplot() +
  labs(title = "Distribution of Ratings by Category", x = "Category", y = "Rating") +
  theme_minimal()
```

```
## 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 26 rows containing non-finite outside the scale range (`stat_boxplot()`).
```




```r
#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
##    <int> <chr>
##  1     1 "adidas Unisex Prime 6 Backpack, Black, One Size  "
##  2     2  <NA>
```

```
##  3     3 "Abshoo Classical Basic Travel Backpack For School Water Resistant Bookbag  "
##  4     4 "Laptop Backpack,Business Travel Anti Theft Slim Durable Laptops Backpack with USB Charging
##  5     5 "THE NORTH FACE Women's Jester Luxe Everyday Laptop Backpack, Gardenia White/Burnt Coral Me
##  6     1 "HP 15.6 FHD Display G9 Laptop • 32GB RAM • 1TB Storage (512GB SSD & 500GB External) • AMD
##  7     2  <NA>
##  8     3 "HP Newest 255 G10 Laptop for Home or Work, 16GB RAM, 1TB SSD, 15.6\" Full HD, Ryzen 3 7330
##  9     4 "HP 15.6\" Business & Student Laptop Computer, Intel Core i5, Windows 11 Pro Laptop with Mi
## 10     5 "HP 14 Laptop, Intel Celeron N4020, 4 GB RAM, 64 GB Storage, 14-inch Micro-edge HD Display,
## # i 15 more rows
```