RWorksheet#5_group(freires,rocillo,sanico)

2024-11-06

Extracting TV Shows Reviews

1. Each group needs to extract the top 50 tv shows in Imdb.com. It will include the rank, the title of the tv show, tv rating, the number of people who voted, the number of episodes, the year it was released. It will also include the number of user reviews and the number of critic reviews, as well as the popularity rating for each tv shows.

```
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(kableExtra)
```

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

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

```
library(rmarkdown)

url <- 'https://www.imdb.com/chart/toptv/?ref_=nv_tvv_250'

session <- bow(url,
               user_agent = "Educational")
session
```

```
## <polite session> https://www.imdb.com/chart/toptv/?ref_=nv_tvv_250
##     User-agent: Educational
##     robots.txt: 35 rules are defined for 3 bots
##    Crawl delay: 5 sec
##   The path is scrapable for this user-agent
```

```
library(rvest)
library(dplyr)

title <- read_html(url) %>%
```

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


data_ <- data.frame(
    titleDf = title[1:25])

data_
```

```
##                                titleDf
## 1                           IMDb Charts
## 2                      1. Breaking Bad
## 3                    2. Planet Earth II
## 4                       3. Planet Earth
## 5                   4. Band of Brothers
## 6                          5. Chernobyl
## 7                           6. The Wire
## 8         7. Avatar: The Last Airbender
## 9                       8. Blue Planet II
## 10                        9. The Sopranos
## 11      10. Cosmos: A Spacetime Odyssey
## 12                           11. Cosmos
## 13                       12. Our Planet
## 14                  13. Game of Thrones
## 15                             14. Bluey
## 16                 15. The World at War
## 17 16. Fullmetal Alchemist: Brotherhood
## 18                   17. Rick and Morty
## 19                              18. Life
## 20                   19. The Last Dance
## 21                20. The Twilight Zone
## 22                  21. The Vietnam War
## 23                         22. Sherlock
## 24                  23. Attack on Titan
## 25      24. Batman: The Animated Series
```

```
title_rank <- as.data.frame(data_, stringsAsFactors = FALSE)
colnames(title_rank) <- "rank"

split_df <- strsplit(as.character(title_rank$rank), "\\.", 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)

title_rank <- split_df
title_rank
```

```
##          rank                         title
## 1  IMDb Charts                   IMDb Charts
## 2           1                  Breaking Bad
## 3           2                Planet Earth II
```

```
## 4            3                     Planet Earth
## 5            4                 Band of Brothers
## 6            5                        Chernobyl
## 7            6                         The Wire
## 8            7       Avatar: The Last Airbender
## 9            8                   Blue Planet II
## 10           9                     The Sopranos
## 11          10      Cosmos: A Spacetime Odyssey
## 12          11                           Cosmos
## 13          12                       Our Planet
## 14          13                 Game of Thrones
## 15          14                           Bluey
## 16          15                The World at War
## 17          16 Fullmetal Alchemist: Brotherhood
## 18          17                   Rick and Morty
## 19          18                             Life
## 20          19                 The Last Dance
## 21          20               The Twilight Zone
## 22          21                 The Vietnam War
## 23          22                         Sherlock
## 24          23                 Attack on Titan
## 25          24      Batman: The Animated Series
```

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

rating_ <- data.frame(
    ratingDf = rating[2:25])

rating_
```

```
##    ratingDf
## 1       9.5
## 2       9.4
## 3       9.4
## 4       9.3
## 5       9.3
## 6       9.3
## 7       9.3
## 8       9.2
## 9       9.2
## 10      9.3
## 11      9.2
## 12      9.2
## 13      9.3
## 14      9.2
## 15      9.1
## 16      9.1
## 17      9.1
## 18      9.0
## 19      9.0
## 20      9.1
## 21      9.1
## 22      9.1
```

```
## 23      9.0
## 24      9.0
```

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

  voted_ <- data.frame(
    votedDf = voted[2:25])

voted_
```

```
##    votedDf
## 1   (163K)
## 2   (224K)
## 3   (547K)
## 4   (910K)
## 5   (392K)
## 6   (391K)
## 7    (49K)
## 8   (501K)
## 9   (132K)
## 10   (46K)
## 11   (54K)
## 12  (2.4M)
## 13   (34K)
## 14   (32K)
## 15  (210K)
## 16  (629K)
## 17   (44K)
## 18  (160K)
## 19   (97K)
## 20   (30K)
## 21    (1M)
## 22  (566K)
## 23  (123K)
## 24  (334K)
```

```
episodes <- read_html(url) %>%
  html_nodes("span.sc-300a8231-7.eaXxft.cli-title-metadata-item:nth-of-type(2)") %>%
  html_text()


episodes_ <- data.frame(
    episodesDf = episodes[1:25])

episodes_
```

```
##    episodesDf
## 1      62 eps
## 2       6 eps
## 3      11 eps
## 4      10 eps
## 5       5 eps
```

```
## 6      60 eps
## 7      62 eps
## 8       7 eps
## 9      86 eps
## 10     13 eps
## 11     13 eps
## 12     12 eps
## 13     74 eps
## 14    194 eps
## 15     26 eps
## 16     68 eps
## 17     78 eps
## 18     11 eps
## 19     10 eps
## 20    156 eps
## 21     10 eps
## 22     15 eps
## 23     98 eps
## 24     85 eps
## 25     18 eps
```

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

years_ <- data.frame(
    years_releases = years[1:25])

years_
```

```
##    years_releases
## 1       2008-2013
## 2            2016
## 3            2006
## 4            2001
## 5            2019
## 6       2002-2008
## 7       2005-2008
## 8            2017
## 9       1999-2007
## 10           2014
## 11           1980
## 12      2019-2023
## 13      2011-2019
## 14         2018-
## 15      1973-1974
## 16      2009-2010
## 17         2013-
## 18           2009
## 19           2020
## 20      1959-1964
## 21           2017
## 22      2010-2017
## 23      2013-2023
## 24      1992-1995
```

```
## 25      2021-2024
```

```r
tv_shows <- data.frame(
  Rank = title_rank[,1],
  Title = title_rank[,2],
  Rating = rating,
  Voters = voted,
  Episodes = episodes,
  Year = years
)
tv_shows
```

```
##           Rank                          Title Rating  Voters Episodes
## 1  IMDb Charts                    IMDb Charts    9.5  (2.2M)   62 eps
## 2            1                   Breaking Bad    9.5  (163K)    6 eps
## 3            2                 Planet Earth II    9.4  (224K)   11 eps
## 4            3                    Planet Earth    9.4  (547K)   10 eps
## 5            4                Band of Brothers    9.3  (910K)    5 eps
## 6            5                       Chernobyl    9.3  (392K)   60 eps
## 7            6                       The Wire     9.3  (391K)   62 eps
## 8            7      Avatar: The Last Airbender    9.3   (49K)    7 eps
## 9            8                  Blue Planet II    9.2  (501K)   86 eps
## 10           9                    The Sopranos    9.2  (132K)   13 eps
## 11          10     Cosmos: A Spacetime Odyssey    9.3   (46K)   13 eps
## 12          11                          Cosmos    9.2   (54K)   12 eps
## 13          12                      Our Planet    9.2  (2.4M)   74 eps
## 14          13                Game of Thrones     9.3   (34K)  194 eps
## 15          14                           Bluey    9.2   (32K)   26 eps
## 16          15               The World at War     9.1  (210K)   68 eps
## 17          16 Fullmetal Alchemist: Brotherhood    9.1  (629K)   78 eps
## 18          17                  Rick and Morty     9.1   (44K)   11 eps
## 19          18                            Life     9.0  (160K)   10 eps
## 20          19                  The Last Dance     9.0   (97K)  156 eps
## 21          20               The Twilight Zone     9.1   (30K)   10 eps
## 22          21                 The Vietnam War     9.1    (1M)   15 eps
## 23          22                        Sherlock     9.1  (566K)   98 eps
## 24          23                 Attack on Titan     9.0  (123K)   85 eps
## 25          24     Batman: The Animated Series     9.0  (334K)   18 eps
##         Year
## 1  2008-2013
## 2       2016
## 3       2006
## 4       2001
## 5       2019
## 6  2002-2008
## 7  2005-2008
## 8       2017
## 9  1999-2007
## 10      2014
## 11      1980
## 12 2019-2023
## 13 2011-2019
## 14    2018-
## 15 1973-1974
## 16 2009-2010
```

```
## 17    2013-
## 18      2009
## 19      2020
## 20 1959-1964
## 21      2017
## 22 2010-2017
## 23 2013-2023
## 24 1992-1995
## 25 2021-2024
```

```
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")

show_url_df <- do.call(rbind, lapply(links, function(link) {
  complete_link <- paste0("https://imdb.com", link)

 usrv_link <- read_html(complete_link)
  usrv_link_page <- usrv_link %>%
    html_nodes('a.isReview') %>%
    html_attr("href")

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

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

   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_Reviews = critic[2], Po
}))

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

```
##           Rank                          Title Rating  Voters Episodes
## 1  IMDb Charts                    IMDb Charts    9.5  (2.2M)   62 eps
## 2            1                   Breaking Bad    9.5  (163K)    6 eps
## 3            2                 Planet Earth II    9.4  (224K)   11 eps
## 4            3                    Planet Earth    9.4  (547K)   10 eps
## 5            4                Band of Brothers    9.3  (910K)    5 eps
## 6            5                       Chernobyl    9.3  (392K)   60 eps
## 7            6                        The Wire    9.3  (391K)   62 eps
## 8            7      Avatar: The Last Airbender    9.3   (49K)    7 eps
## 9            8                  Blue Planet II    9.2  (501K)   86 eps
```

```
## 10           9                       The Sopranos    9.2  (132K)   13 eps
## 11          10      Cosmos: A Spacetime Odyssey    9.3   (46K)   13 eps
## 12          11                            Cosmos    9.2   (54K)   12 eps
## 13          12                        Our Planet    9.2  (2.4M)   74 eps
## 14          13                   Game of Thrones    9.3   (34K)  194 eps
## 15          14                             Bluey    9.2   (32K)   26 eps
## 16          15                  The World at War    9.1  (210K)   68 eps
## 17          16 Fullmetal Alchemist: Brotherhood    9.1  (629K)   78 eps
## 18          17                    Rick and Morty    9.1   (44K)   11 eps
## 19          18                              Life    9.0  (160K)   10 eps
## 20          19                   The Last Dance    9.0   (97K)  156 eps
## 21          20                 The Twilight Zone    9.1   (30K)   10 eps
## 22          21                   The Vietnam War    9.1    (1M)   15 eps
## 23          22                          Sherlock    9.1  (566K)   98 eps
## 24          23                   Attack on Titan    9.0  (123K)   85 eps
## 25          24      Batman: The Animated Series    9.0  (334K)   18 eps
## 26 IMDb Charts                       IMDb Charts    9.5  (2.2M)   62 eps
## 27           1                      Breaking Bad    9.5  (163K)    6 eps
## 28           2                   Planet Earth II    9.4  (224K)   11 eps
## 29           3                      Planet Earth    9.4  (547K)   10 eps
## 30           4                  Band of Brothers    9.3  (910K)    5 eps
## 31           5                         Chernobyl    9.3  (392K)   60 eps
## 32           6                          The Wire    9.3  (391K)   62 eps
## 33           7       Avatar: The Last Airbender    9.3   (49K)    7 eps
## 34           8                    Blue Planet II    9.2  (501K)   86 eps
## 35           9                      The Sopranos    9.2  (132K)   13 eps
## 36          10      Cosmos: A Spacetime Odyssey    9.3   (46K)   13 eps
## 37          11                            Cosmos    9.2   (54K)   12 eps
## 38          12                        Our Planet    9.2  (2.4M)   74 eps
## 39          13                   Game of Thrones    9.3   (34K)  194 eps
## 40          14                             Bluey    9.2   (32K)   26 eps
## 41          15                  The World at War    9.1  (210K)   68 eps
## 42          16 Fullmetal Alchemist: Brotherhood    9.1  (629K)   78 eps
## 43          17                    Rick and Morty    9.1   (44K)   11 eps
## 44          18                              Life    9.0  (160K)   10 eps
## 45          19                   The Last Dance    9.0   (97K)  156 eps
## 46          20                 The Twilight Zone    9.1   (30K)   10 eps
## 47          21                   The Vietnam War    9.1    (1M)   15 eps
## 48          22                          Sherlock    9.1  (566K)   98 eps
## 49          23                   Attack on Titan    9.0  (123K)   85 eps
## 50          24      Batman: The Animated Series    9.0  (334K)   18 eps
##         Year                                          Show_Link  User_Reviews
## 1  2008-2013   https://imdb.com/title/tt0903747/?ref_=chttvtp_t_1 5,109 reviews
## 2       2016   https://imdb.com/title/tt0903747/?ref_=chttvtp_t_1 5,109 reviews
## 3       2006   https://imdb.com/title/tt5491994/?ref_=chttvtp_t_2   158 reviews
## 4       2001   https://imdb.com/title/tt5491994/?ref_=chttvtp_t_2   158 reviews
## 5       2019   https://imdb.com/title/tt0795176/?ref_=chttvtp_t_3   111 reviews
## 6  2002-2008   https://imdb.com/title/tt0795176/?ref_=chttvtp_t_3   111 reviews
## 7  2005-2008   https://imdb.com/title/tt0185906/?ref_=chttvtp_t_4 1,058 reviews
## 8       2017   https://imdb.com/title/tt0185906/?ref_=chttvtp_t_4 1,058 reviews
## 9  1999-2007   https://imdb.com/title/tt7366338/?ref_=chttvtp_t_5 3,534 reviews
## 10      2014   https://imdb.com/title/tt7366338/?ref_=chttvtp_t_5 3,534 reviews
## 11      1980   https://imdb.com/title/tt0306414/?ref_=chttvtp_t_6   787 reviews
## 12 2019-2023   https://imdb.com/title/tt0306414/?ref_=chttvtp_t_6   787 reviews
```

```
## 13 2011-2019   https://imdb.com/title/tt0417299/?ref_=chttvtp_t_7 1,003 reviews
## 14    2018-    https://imdb.com/title/tt0417299/?ref_=chttvtp_t_7 1,003 reviews
## 15 1973-1974   https://imdb.com/title/tt6769208/?ref_=chttvtp_t_8    53 reviews
## 16 2009-2010   https://imdb.com/title/tt6769208/?ref_=chttvtp_t_8    53 reviews
## 17    2013-    https://imdb.com/title/tt0141842/?ref_=chttvtp_t_9   966 reviews
## 18      2009   https://imdb.com/title/tt0141842/?ref_=chttvtp_t_9   966 reviews
## 19      2020  https://imdb.com/title/tt2395695/?ref_=chttvtp_t_10   205 reviews
## 20 1959-1964  https://imdb.com/title/tt2395695/?ref_=chttvtp_t_10   205 reviews
## 21      2017  https://imdb.com/title/tt0081846/?ref_=chttvtp_t_11    80 reviews
## 22 2010-2017  https://imdb.com/title/tt0081846/?ref_=chttvtp_t_11    80 reviews
## 23 2013-2023  https://imdb.com/title/tt9253866/?ref_=chttvtp_t_12   245 reviews
## 24 1992-1995  https://imdb.com/title/tt9253866/?ref_=chttvtp_t_12   245 reviews
## 25 2021-2024  https://imdb.com/title/tt0944947/?ref_=chttvtp_t_13 5,907 reviews
## 26 2008-2013  https://imdb.com/title/tt0944947/?ref_=chttvtp_t_13 5,907 reviews
## 27      2016  https://imdb.com/title/tt7678620/?ref_=chttvtp_t_14   369 reviews
## 28      2006  https://imdb.com/title/tt7678620/?ref_=chttvtp_t_14   369 reviews
## 29      2001  https://imdb.com/title/tt0071075/?ref_=chttvtp_t_15   126 reviews
## 30      2019  https://imdb.com/title/tt0071075/?ref_=chttvtp_t_15   126 reviews
## 31 2002-2008  https://imdb.com/title/tt1355642/?ref_=chttvtp_t_16   468 reviews
## 32 2005-2008  https://imdb.com/title/tt1355642/?ref_=chttvtp_t_16   468 reviews
## 33      2017  https://imdb.com/title/tt2861424/?ref_=chttvtp_t_17   910 reviews
## 34 1999-2007  https://imdb.com/title/tt2861424/?ref_=chttvtp_t_17   910 reviews
## 35      2014  https://imdb.com/title/tt1533395/?ref_=chttvtp_t_18    12 reviews
## 36      1980  https://imdb.com/title/tt1533395/?ref_=chttvtp_t_18    12 reviews
## 37 2019-2023  https://imdb.com/title/tt8420184/?ref_=chttvtp_t_19   542 reviews
## 38 2011-2019  https://imdb.com/title/tt8420184/?ref_=chttvtp_t_19   542 reviews
## 39    2018-   https://imdb.com/title/tt0052520/?ref_=chttvtp_t_20   214 reviews
## 40 1973-1974  https://imdb.com/title/tt0052520/?ref_=chttvtp_t_20   214 reviews
## 41 2009-2010  https://imdb.com/title/tt1877514/?ref_=chttvtp_t_21   175 reviews
## 42    2013-   https://imdb.com/title/tt1877514/?ref_=chttvtp_t_21   175 reviews
## 43      2009  https://imdb.com/title/tt1475582/?ref_=chttvtp_t_22 1,098 reviews
## 44      2020  https://imdb.com/title/tt1475582/?ref_=chttvtp_t_22 1,098 reviews
## 45 1959-1964  https://imdb.com/title/tt2560140/?ref_=chttvtp_t_23 2,370 reviews
## 46      2017  https://imdb.com/title/tt2560140/?ref_=chttvtp_t_23 2,370 reviews
## 47 2010-2017  https://imdb.com/title/tt0103359/?ref_=chttvtp_t_24   219 reviews
## 48 2013-2023  https://imdb.com/title/tt0103359/?ref_=chttvtp_t_24   219 reviews
## 49 1992-1995 https://imdb.com/title/tt11126994/?ref_=chttvtp_t_25 2,227 reviews
## 50 2021-2024 https://imdb.com/title/tt11126994/?ref_=chttvtp_t_25 2,227 reviews
##    Critic_Reviews Popularity_Rating
## 1             175                19
## 2             175                19
## 3               6               985
## 4               6               985
## 5              10             1,837
## 6              10             1,837
## 7              34               162
## 8              34               162
## 9              88               143
## 10             88               143
## 11             77               108
## 12             77               108
## 13             57               294
## 14             57               294
## 15              9             4,270
```

```
## 16              9             4,270
## 17             93                27
## 18             93                27
## 19             12             1,439
## 20             12             1,439
## 21              8             3,817
## 22              8             3,817
## 23             15             2,632
## 24             15             2,632
## 25            368                12
## 26            368                12
## 27              4               331
## 28              4               331
## 29              5             2,696
## 30              5             2,696
## 31             16               479
## 32             16               479
## 33             94               124
## 34             94               124
## 35              9             3,030
## 36              9             3,030
## 37             28             1,530
## 38             28             1,530
## 39             85               319
## 40             85               319
## 41             13             1,776
## 42             13             1,776
## 43            121               167
## 44            121               167
## 45             64                43
## 46             64                43
## 47             25               510
## 48             25               510
## 49             59                 1
## 50             59                 1
```

2. From the 50 tv shows, select at least 5 tv shows to scrape 20 user reviews that will include the reviewer's name, date of reviewed, user rating, title of the review, the numbers for "is helpful" and "is not helpful", and text reviews.

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

show_urls <- c(
  'https://www.imdb.com/title/tt0903747/reviews/?ref_=tt_ov_urv',   # Breaking Bad
  'https://www.imdb.com/title/tt5491994/reviews/?ref_=tt_ov_ql_2',  # Planet Earth II
  'https://www.imdb.com/title/tt0795176/reviews/?ref_=tt_ov_ql_2',  # Planet Earth
  'https://www.imdb.com/title/tt0185906/reviews/?ref_=tt_ov_ql_2',  # Band of Brothers
  'https://www.imdb.com/title/tt7366338/reviews/?ref_=tt_ov_ql_2'   # Chernobyl
)

scrape_reviews <- function(show_url) {

  review_page <- read_html(show_url)
```

```
  show_name <- review_page %>%
    html_nodes('h2') %>%
    html_text() %>%
    trimws()

  reviewers <- review_page %>%
    html_nodes('a.ipc-link--base[data-testid="author-link"]') %>%
    html_text()

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

  user_ratings <- review_page %>%
    html_nodes('.ipc-rating-star--rating') %>%
    html_text() %>%
    as.numeric()

  review_titles <- review_page %>%
    html_nodes('h3.ipc-title__text') %>%
    html_text()

  helpful_count <- review_page %>%
    html_nodes('.ipc-voting__label__count--up') %>%
    html_text() %>%
    as.numeric()

  not_helpful_count <- review_page %>%
    html_nodes('.ipc-voting__label__count--down') %>%
    html_text() %>%
    as.numeric()

  review_text <- review_page %>%
    html_nodes('.ipc-html-content-inner-div') %>%
    html_text()

  review_text <- trimws(review_text)

  reviews <- data.frame(
    Show = show_name,
    Reviewer = reviewers[1:20],
    Date = review_dates[1:20],
    UserRating = user_ratings[1:20],
    ReviewTitle = review_titles[1:20],
    HelpfulCount = helpful_count[1:20],
    NotHelpfulCount = not_helpful_count[1:20],
    ReviewText = review_text[1:20]
  )

  return(reviews)
}

all_reviews <- lapply(show_urls, scrape_reviews)
```

```r
reviews_df <- bind_rows(all_reviews)
print(reviews_df)
```

```
##                 Show             Reviewer         Date UserRating
## 1       Breaking Bad              FiRE010  Jul 3, 2021         10
## 2       Breaking Bad            bruhperson  Mar 6, 2019         10
## 3       Breaking Bad         KinoKoopaKid Jul 29, 2021         10
## 4       Breaking Bad           jehuschultz Feb 18, 2020         10
## 5       Breaking Bad        Supermanfan-13  Nov 8, 2021         10
## 6       Breaking Bad     manishsingh-03299 May 30, 2019         10
## 7       Breaking Bad              Rob1331  Dec 8, 2022         10
## 8       Breaking Bad             xpinerhd Nov 15, 2019         10
## 9       Breaking Bad    dhanushreddy-14919 Jul 17, 2021         10
## 10      Breaking Bad           tushv-31482  Dec 8, 2022         10
## 11      Breaking Bad               dyarutd Sep 28, 2024          7
## 12      Breaking Bad            reebokroot Feb 14, 2021          5
## 13      Breaking Bad          alpierce1991 May 18, 2014         10
## 14      Breaking Bad     TheLittleSongbird Nov 12, 2017         10
## 15      Breaking Bad           gogoschka-1 Jan 11, 2014         10
## 16      Breaking Bad           FishDrowned  Nov 8, 2021         10
## 17      Breaking Bad            joegalgano Aug 11, 2021         10
## 18      Breaking Bad           agatt-87232 May 19, 2019         10
## 19      Breaking Bad        Leofwine_draca  May 4, 2021         10
## 20      Breaking Bad         dristysultana Jun 23, 2021         10
## 21   Planet Earth II          arjanhylkema  Nov 7, 2016         10
## 22   Planet Earth II              Wentloog  Nov 5, 2016         10
## 23   Planet Earth II          john-m-madsen  Nov 5, 2016         10
## 24   Planet Earth II           thespookybuz  Nov 9, 2016         10
## 25   Planet Earth II            pjdickinson  Nov 5, 2016         10
## 26   Planet Earth II               dbijis33  Nov 8, 2016         10
## 27   Planet Earth II        dhanrajjughead Nov 17, 2016         10
## 28   Planet Earth II           NeilBarnett Nov 13, 2016         10
## 29   Planet Earth II          salmanu-27386  Nov 6, 2016         10
## 30   Planet Earth II    panagiotiskatsanos Dec 31, 2016         10
## 31   Planet Earth II                ianrobo Nov 19, 2016         10
## 32   Planet Earth II             adamonIMDb Dec 28, 2016          7
## 33   Planet Earth II             tinyfordst May 19, 2019         10
## 34   Planet Earth II           larask-21775 Oct 20, 2018         10
## 35   Planet Earth II            BobFillmore Sep 29, 2017         10
## 36   Planet Earth II         farshidkarimi Nov 22, 2016         10
## 37   Planet Earth II     TheLittleSongbird Oct 12, 2017         10
## 38   Planet Earth II      myersei-165-4350  Dec 4, 2016         10
## 39   Planet Earth II     fierceeagle-40009 Apr 23, 2020         10
## 40   Planet Earth II          adam-whitmore  Jan 5, 2017         10
## 41      Planet Earth          robert-kamer  Feb 8, 2007         10
## 42      Planet Earth              jim-1409 Nov 19, 2008         10
## 43      Planet Earth       ccthemovieman-1  Jan 4, 2009         10
## 44      Planet Earth              cmcoveos Dec 15, 2006         10
## 45      Planet Earth             Loordssm  Sep 1, 2007         10
## 46      Planet Earth             ultimorn Aug 27, 2006         10
## 47      Planet Earth           bob the moo Apr 30, 2006         10
## 48      Planet Earth                 alfeu Jun 29, 2015          9
## 49      Planet Earth               Cabrone Jul 20, 2006         10
## 50      Planet Earth             berndt65 Jan 28, 2009         10
```

```
## 51      Planet Earth        planktonrules  Jun 1, 2015          7
## 52      Planet Earth      dakuchonekobing  Oct 8, 2020          3
## 53      Planet Earth             rooprect  Dec 4, 2007         10
## 54      Planet Earth                bs3dc Jan 15, 2007         10
## 55      Planet Earth                Nerte Jul 30, 2008         10
## 56      Planet Earth                ortz3 Dec 25, 2017         10
## 57      Planet Earth        solon-stewart Sep 14, 2009          9
## 58      Planet Earth   bellino-angelo2014 Sep 20, 2020         10
## 59      Planet Earth      krational-66550 May 31, 2020          9
## 60      Planet Earth     edwardashton4545 Jul 27, 2014         10
## 61  Band of Brothers              Rob1331 Sep 27, 2022         10
## 62  Band of Brothers         sanderson777 Oct 14, 2001         10
## 63  Band of Brothers          wildcatt268 Jan 18, 2002         10
## 64  Band of Brothers              arjay24 Apr 18, 2004         10
## 65  Band of Brothers            rbverhoef Feb 13, 2003         10
## 66  Band of Brothers           yodaschoda Jan 23, 2005         10
## 67  Band of Brothers   philip_vanderveken Sep 16, 2004         10
## 68  Band of Brothers       Supermanfan-13  May 6, 2022         10
## 69  Band of Brothers            thiagoutp  Nov 4, 2019         10
## 70  Band of Brothers           bsmith5552  Nov 5, 2001         10
## 71  Band of Brothers faded_english_monkey Aug 25, 2004         10
## 72  Band of Brothers        planktonrules May 30, 2015          7
## 73  Band of Brothers       stilonkostrzyn Apr 10, 2021          5
## 74  Band of Brothers          faded_Glory  May 2, 2006         10
## 75  Band of Brothers              jazmodo  Jun 3, 2019         10
## 76  Band of Brothers             kait2007 Jan 26, 2005         10
## 77  Band of Brothers          mickman91-1  May 3, 2022         10
## 78  Band of Brothers          kipmcmillan Oct 24, 2018          9
## 79  Band of Brothers       erwan_ticheler  Dec 7, 2002         10
## 80  Band of Brothers            grahamsj3 Nov 25, 2002         10
## 81         Chernobyl     curiosityonmars May 23, 2019         10
## 82         Chernobyl             stelmakh May 10, 2019         10
## 83         Chernobyl        natashapekar  May 9, 2019         10
## 84         Chernobyl          m-porpaczi May 14, 2019         10
## 85         Chernobyl             Lladerat  May 7, 2019         10
## 86         Chernobyl             jfirebug May 20, 2019         10
## 87         Chernobyl              thegldt  May 6, 2019         10
## 88         Chernobyl    alexander-phoenix May 13, 2019         10
## 89         Chernobyl          wmeduardowm  May 6, 2019         10
## 90         Chernobyl       Leofwine_draca Nov 27, 2019         10
## 91         Chernobyl         Jamie_Seaton May 23, 2019          1
## 92         Chernobyl             piter554 Jun 29, 2019          8
## 93         Chernobyl              frimark May 20, 2019         10
## 94         Chernobyl      krzysztof-18241 May 30, 2019         10
## 95         Chernobyl           ahmetkozan  Jun 7, 2019          9
## 96         Chernobyl              Rob1331 Sep 27, 2022         10
## 97         Chernobyl        stephenpdodds  May 6, 2019          9
## 98         Chernobyl       Supermanfan-13 Jul 10, 2022          9
## 99         Chernobyl            emholberg May 26, 2019         10
## 100        Chernobyl        josephwhitton May 15, 2019          7
##
## 1
## 2
## 3
```

```
## 4
## 5
## 6                                                                                         Those days
## 7
## 8
## 9
## 10                                                                                               On
## 11                                                                                      My Review F
## 12
## 13                                                                     Easily the most overrated telev
## 14                                                                            Among the best and most
## 15  If you mix Scarface, Robin Hood and maybe Tyler Durden with enough meth - you'll get a mean cock
## 16                                                                                  By far the greates
## 17                                                                                              in a
## 18                                                             Since GOT is over, this is Officially the
## 19                                                                                        Every bit
## 20
## 21
## 22                                                                                     At once awe-i
## 23                                                                 Yet another masterpiece from BBC Nat
## 24
## 25
## 26
## 27                                                                                         Greatest d
## 28                                                                                    Best thing on TV
## 29
## 30                                                                               One of the best docum
## 31                                                                                        In times
## 32
## 33                                                                     More irritated with IMDb for the
## 34
## 35                                                                                     Should be req
## 36                                                                                     What a Beauti
## 37         Like the first 'Planet Earth', does for nature and our planet as 'Walking with Dinosaurs
## 38                                                                                   This masterpiec
## 39
## 40                                                                               Peerless evocation of
## 41
## 42                                                                                           A mast
## 43
## 44                                                             The most amazing achievement in natural h
## 45
## 46                                                                                An amazing trip arou
## 47                                                         A visually impressive and memorable look at t
## 48                                                                                      Is it real?
## 49
## 50                                                                                               Ar
## 51                                                                                      It doesn't g
## 52                                                                                            Only
## 53                                                                                 Should be called "B
## 54                                                                                            Bril
## 55                                                                                 Explanation to tho
## 56
## 57
```

```
## 58                                                                      Words fail me to
## 59
## 60
## 61
## 62                                                                     Possibly the fine
## 63                                                                        One of the bes
## 64
## 65
## 66                                                                One of, if not the best
## 67                                                        This series is so unbelievably
## 68                                                                     One of the best m
## 69
## 70                                                                      Realistic WWII D
## 71
## 72
## 73
## 74                                                                                     N
## 75                                                              Without Doubt, the Best M
## 76
## 77                                                   A series like this won't be made again (se
## 78
## 79
## 80                                                                      A-1, TOPS, the B
## 81
## 82
## 83                                                                               I hig
## 84                                                                             No hero
## 85
## 86
## 87                                                                    Bleak, Unsettling,
## 88
## 89
## 90
## 91
## 92
## 93
## 94
## 95                                                                       Now you look li
## 96
## 97
## 98
## 99                                                        It is hard to overestimate the
## 100
##     HelpfulCount NotHelpfulCount
## 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
## 21            NA             NA
## 22            NA             NA
## 23            NA             NA
## 24            NA             NA
## 25            NA             NA
## 26            NA             NA
## 27            NA             NA
## 28            NA             NA
## 29            NA             NA
## 30            NA             NA
## 31            NA             NA
## 32            NA             NA
## 33            NA             NA
## 34            NA             NA
## 35            NA             NA
## 36            NA             NA
## 37            NA             NA
## 38            NA             NA
## 39            NA             NA
## 40            NA             NA
## 41            NA             NA
## 42            NA             NA
## 43            NA             NA
## 44            NA             NA
## 45            NA             NA
## 46            NA             NA
## 47            NA             NA
## 48            NA             NA
## 49            NA             NA
## 50            NA             NA
## 51            NA             NA
## 52            NA             NA
## 53            NA             NA
## 54            NA             NA
## 55            NA             NA
## 56            NA             NA
## 57            NA             NA
## 58            NA             NA
## 59            NA             NA
## 60            NA             NA
## 61            NA             NA
## 62            NA             NA
## 63            NA             NA
## 64            NA             NA
```

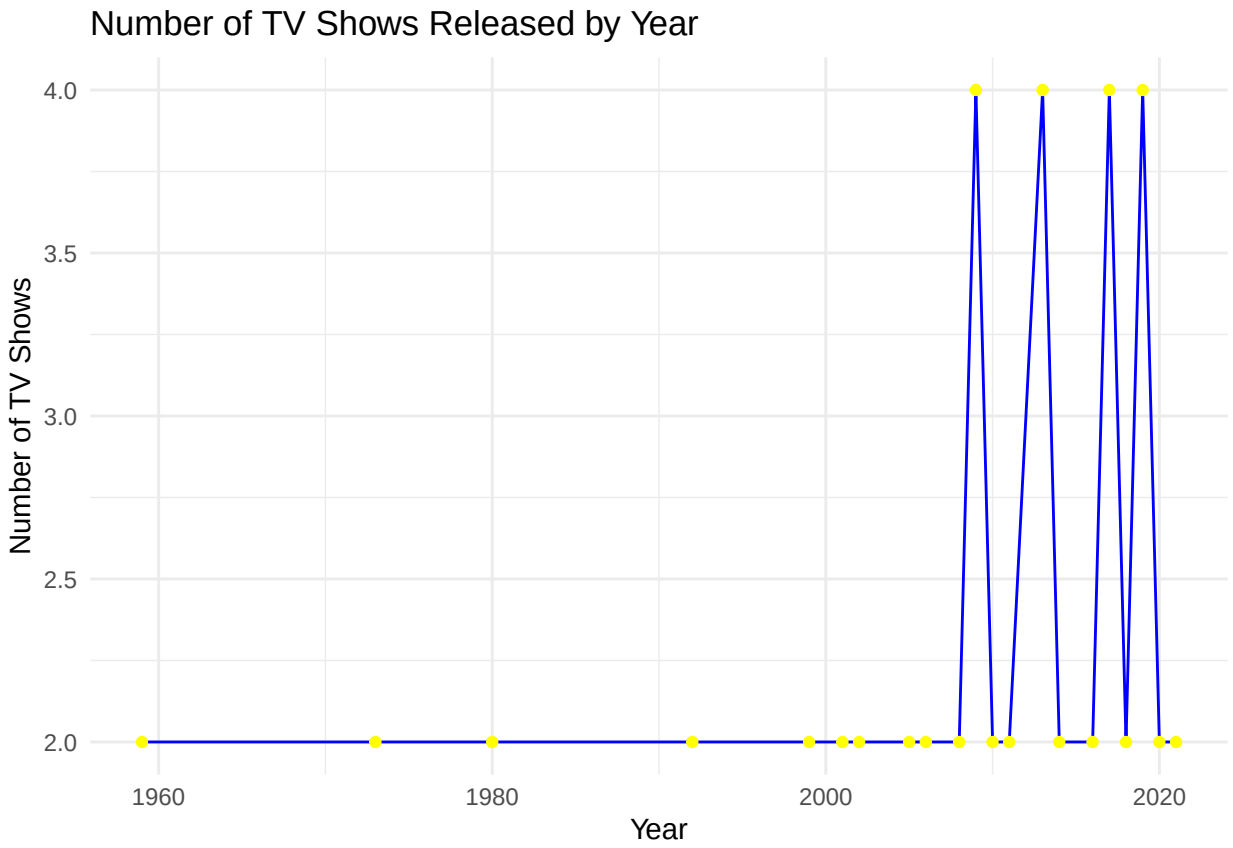
```
## 65            NA              NA
## 66            NA              NA
## 67            NA              NA
## 68            NA              NA
## 69            NA              NA
## 70            NA              NA
## 71            NA              NA
## 72            NA              NA
## 73            NA              NA
## 74            NA              NA
## 75            NA              NA
## 76            NA              NA
## 77            NA              NA
## 78            NA              NA
## 79            NA              NA
## 80            NA              NA
## 81            NA              NA
## 82            NA              NA
## 83            NA              NA
## 84            NA              NA
## 85            NA              NA
## 86            NA              NA
## 87            NA              NA
## 88            NA              NA
## 89            NA              NA
## 90            NA              NA
## 91            NA              NA
## 92            NA              NA
## 93            NA              NA
## 94            NA              NA
## 95            NA              NA
## 96            NA              NA
## 97            NA              NA
## 98            NA              NA
## 99            NA              NA
## 100           NA              NA
##
## 1
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16
## 17
```

```
## 18
## 19
## 20
## 21
## 22
## 23
## 24
## 25
## 26
## 27
## 28
## 29
## 30
## 31
## 32
## 33
## 34
## 35
## 36
## 37  Absolutely adore the first 'Planet Earth' from 2007, one of the best documentaries ever made and
## 38
## 39
## 40
## 41
## 42
## 43
## 44
## 45
## 46
## 47
## 48
## 49
## 50
## 51
## 52
## 53
## 54
## 55
## 56
## 57
## 58
## 59
## 60
## 61
## 62
## 63
## 64
## 65
## 66
## 67
## 68
## 69
## 70
## 71
```

```
## 72
## 73
## 74
## 75
## 76
## 77
## 78
## 79
## 80
## 81
## 82
## 83
## 84
## 85
## 86
## 87
## 88
## 89
## 90
## 91
## 92
## 93
## 94
## 95
## 96
## 97
## 98
## 99
## 100
```

3. Create a time series graph for the tv shows released by year. Which year has the most number of tv shows released?

```
library(ggplot2)

shows$Year <- as.numeric(substr(shows$Year, 1, 4))
shows <- shows %>% filter(!is.na(Year))

shows_by_year <- shows %>%
  group_by(Year) %>%
  summarize(Number_of_Shows = n(), .groups = "drop")

ggplot(shows_by_year, aes(x = Year, y = Number_of_Shows)) +
  geom_line(color = "blue") +
  geom_point(color = "yellow") +
  labs(title = "Number of TV Shows Released by Year",
       x = "Year",
       y = "Number of TV Shows") +
  theme_minimal()
```



4. Select 5 categories from Amazon and select 30 products from each category.
5. Extract the price, description, ratings and reviews of each product.

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


makeup_url <- "https://www.amazon.com/s?k=lipstick&crid=1T047TIYTCQV2&sprefix=lip%2Caps%2C436&ref=nb_sb

session1 <- bow(makeup_url,
                user_agent = "Educational")
session1
```

```
## <polite session> https://www.amazon.com/s?k=lipstick&crid=1T047TIYTCQV2&sprefix=lip%2Caps%2C436&ref=n
##     User-agent: Educational
##     robots.txt: 138 rules are defined for 5 bots
##    Crawl delay: 5 sec
##   The path is scrapable for this user-agent
```

```
library(rvest)

amazondf1 <- data.frame()

page1 <- scrape(session1)
```

```
price1 <- page1 %>%
  html_nodes('.a-price .a-offscreen') %>%
  html_text(trim = TRUE) %>%
  gsub("[^0-9\\.]", "", .) %>%
  head(30)

description1 <- page1 %>%
  html_nodes('.a-color-base.a-text-normal') %>%
  html_text() %>%
  head(30)

ratings1 <- page1 %>%
  html_nodes('.a-icon-alt') %>%
  html_text() %>%
  gsub(" out of 5 stars", "", .) %>%
  head(30)

amazondf1 <- data.frame(
  Prices = price1,
  Descriptions = description1,
  Ratings = ratings1,
   stringsAsFactors = FALSE)

amazondf1
```

```
##    Prices
## 1    4.99
## 2   38.38
## 3    9.99
## 4    6.33
## 5  633.00
## 6    9.49
## 7    5.24
## 8   40.31
## 9   17.50
## 10  17.50
## 11  25.00
## 12   7.49
## 13  53.50
## 14  12.99
## 15   5.24
## 16  40.31
## 17   9.99
## 18  16.50
## 19 137.50
## 20  22.00
## 21   5.70
## 22  51.82
## 23  12.09
## 24   4.99
## 25   0.81
## 26   9.79
## 27   3.40
```

```
## 28  24.29
## 29   7.49
## 30   7.99
##
## 1
## 2                                Maybelline Super Stay Vinyl Ink Longwear No-Budge Liquid Lipcolor Mak
## 3                                                                                L'Oréal Paris Colour Ri
## 4  Flower Knows Strawberry Rococo Cloud Lip Cream Matte Liquid Lipstick-Long Lasting&Highly Pigmentec
## 5
## 6                    Miraculous Ladybug Cosmetic Lipstick Set - 7 Fruit Scented Color-Changing Lipstic
## 7                                                        Rimmel London Provocalips 16hr Kiss-Proof
## 8                                                                      6Pcs Matte Liquid Lipsti
## 9                           wet n wild Silk Finish Lipstick, Hydrating Rich Buildable Lip Color, Formu
## 10                                                    COVERGIRL Outlast, 10 Sugey Girl, Lipstain, Smoo
## 11                                                 firstfly Pack of 6 Crystal Flower Jelly Lipstick, I
## 12                                                       Maybelline Color Sensational Lipstick, Lip
## 13                         stila Stay All Day® Liquid Lipstick, Matte | Long-Lasting Color Wear, No Ti
## 14
## 15                                                  REVLON Lipstick, ColorStay Suede Ink, Built-in
## 16                                                          Maybelline Super Stay Matte Ink Liquic
## 17                                Rimmel Lasting Finish Lipstick - Up to 8 Hours of Intense Lip Colo
## 18                                                                                L'Oreal Pa
## 19
## 20                   Urban Decay Vice Hydrating Lipstick, Long-lasting Cream Matte or Shine Finish, Ci
## 21
## 22                                                  Maybelline Super Stay 24, 2-Step Liquid Lipsti
## 23                                                     Revlon Super Lustrous Lipstick, Lip Makeup
## 24                                                       Peripera Ink the Velvet Lip Tint - High I
## 25                                                                          LAURA GELLER NEW YORI
## 26                     wet n wild Mega Last High-Shine Lipstick Lip Color, Infused with Seed Oils For a
## 27    QiBest 7Pcs Matte Liquid Lipstick + 1Pcs Lip Plumper Makeup Set Kit, Pigmented Long Lasting Lip
## 28                                                              Rimmel Lasting Finish Matte Lipst
## 29                                  Maybelline Super Stay Ink Crayon Lipstick Makeup, Precision Tip I
## 30                                           Revlon ColorStay Overtime Liquid Lipstick with Clear Lip
##    Ratings
## 1      4.6
## 2      4.3
## 3      4.4
## 4      4.8
## 5      4.5
## 6      4.6
## 7      4.5
## 8      4.0
## 9      4.2
## 10     3.9
## 11     4.2
## 12     4.4
## 13     4.3
## 14     4.4
## 15     4.3
## 16     4.4
## 17     4.3
## 18     4.4
## 19     4.6
```

```
## 20     4.4
## 21     4.5
## 22     4.5
## 23     4.5
## 24     4.1
## 25     4.3
## 26     4.0
## 27     4.2
## 28     4.4
## 29     4.4
## 30     4.5
```

```
skincare_url <- 'https://www.amazon.com/s?k=face+mask&rh=n%3A11060451&ref=nb_sb_noss'

session2 <- bow(skincare_url,
                user_agent = "Educational")
session2
```

```
## <polite session> https://www.amazon.com/s?k=face+mask&rh=n%3A11060451&ref=nb_sb_noss
##     User-agent: Educational
##     robots.txt: 138 rules are defined for 5 bots
##    Crawl delay: 5 sec
##   The path is scrapable for this user-agent
```

```
library(rvest)

amazondf2 <- data.frame()

page2 <- scrape(session2)

price2 <- page2 %>%
  html_nodes('.a-price .a-offscreen') %>%
  html_text(trim = TRUE) %>%
  gsub("[^0-9\\.]", "", .) %>%
  head(30)

description2 <- page2 %>%
  html_nodes('.a-color-base.a-text-normal') %>%
  html_text() %>%
  head(30)

ratings2 <- page2 %>%
  html_nodes('.a-icon-alt') %>%
  html_text() %>%
  gsub(" out of 5 stars", "", .) %>%
  head(30)

amazondf2 <- data.frame(
  Prices = price2,
  Descriptions = description2,
  Ratings = ratings2,
   stringsAsFactors = FALSE)

amazondf2
```

```
##    Prices
```

```
## 1    8.39
## 2    0.52
## 3   16.00
## 4   17.49
## 5    0.45
## 6   24.99
## 7   11.89
## 8    0.50
## 9   16.99
## 10  12.90
## 11   3.23
## 12  19.00
## 13   8.39
## 14   0.52
## 15  16.00
## 16  11.89
## 17   0.50
## 18  16.99
## 19   7.97
## 20   1.14
## 21   9.99
## 22  14.99
## 23   1.50
## 24  25.00
## 25  11.96
## 26   1.71
## 27  14.95
## 28   9.99
## 29   0.83
## 30  14.99
##
## 1                                                                                                BIODANCE Bio-Collag
## 2    DERMAL 16 Combo Pack A Collagen Essence Korean Face Mask - Hydrating & Soothing Facial Mask witl
## 3     DERMAL 24 Combo Pack A Collagen Essence Korean Face Mask - Hydrating & Soothing Facial Mask wi
## 4    ZealSea Face Masks Skincare, Facial Masks for Women Skin Care, Sheet Masks Beauty with Natural
## 5                                 COSRX Snail Mucin Sheet Mask 10 EA with Snail Mucin Serum, Self Care
## 6                                       FACETORY K Beauty Face Mask Skin Care - BEST OF 7 COLLECTIO
## 7                                                                                                    Cela
## 8    DERMAL 39 Combo Pack A Collagen Essence Korean Face Mask - Hydrating & Soothing Facial Mask witl
## 9       Freeman Naturals Facial Mask 12 Piece Variety Bundle, Peel-Off, Gel & Cream Face Masks, Hydra
## 10  Kiehl's Rare Earth Deep Pore Cleansing Mask, Pore-Minimizing Face Mask for Clogged Pores, Detoxi
## 11                                                        Vitamin C Face Mask with Kaolin Clay and
## 12   MAREE Collagen Face Mask with Hyaluronic Acid - Sheet Face Masks Skincare with Green & Red Alg
## 13
## 14
## 15      LOOPS DOUBLE TAKE - Glow Hydrogel Face Mask - Calms and Soothes Skin’s Surface - Helps Refi
## 16
## 17    Ebanel 15 Pack Collagen Peptide Hydrating Face Masks, Instant Brightening Firming Anti Aging
## 18                           SUNGBOON EDITOR Deep Collagen Overnight mask 37gx4ea | Real Collagen 2,16
## 19   New York Biology Dead Sea Mud Mask for Face and Body - Spa Quality Pore Reducer for Acne, Blac
## 20
## 21                        14 Pack Sheet Face Masks Skincare for All Skin Types, Hydrating Face Masks
## 22 DERMAL 32 Combo A+B Set Collagen Essence Korean Face Mask - Hydrating and Soothing Facial Mask wi
## 23  Face Mask Skin Care Hydrating Face Masks Sheets, Hyaluronic Acid Sheets Pack Deep Moisturizing Fa
```

```
## 24                                                    Vitamin C and Collagen Sheet Face Mask - Reduc
## 25                           Mario Badescu Clay Face Mask Skin Care for Men and Women, Pore Minimizer
## 26   Ebanel 10 Pack Collagen Face Mask, Instant Brightening & Hydrating Face Sheet Mask with Aloe Ver
## 27              Bliss Pumpkin Enzyme Face Mask | Pumpkin Powerhouse Resurfacing & Exfoliating Mask | V
## 28           Korean Face Mask - 10ct Snail Mucin Hydrating Face Masks Anti Wrinkle Anti Aging Deep Mo
## 29  The Face Shop At Home Aesthetics Vegan Collagen Face Mask, Korean Glass Skin Care, Original & Vit
## 30                                                                  Caudalie Instant Detox Mask P
##    Ratings
## 1      4.3
## 2      4.5
## 3      4.6
## 4      4.6
## 5      4.6
## 6      4.7
## 7      4.6
## 8      4.6
## 9      4.6
## 10     4.5
## 11     4.4
## 12     4.5
## 13     4.7
## 14     4.6
## 15     4.5
## 16     4.8
## 17     4.4
## 18     4.4
## 19     4.5
## 20     4.6
## 21     4.4
## 22     4.7
## 23     4.5
## 24     4.5
## 25     4.4
## 26     4.4
## 27     4.4
## 28     4.5
## 29     4.7
## 30     4.6
```

```
fragrance_url <- 'https://www.amazon.com/s?k=women&rh=n%3A11056591&ref=nb_sb_noss'

session3 <- bow(fragrance_url,
                user_agent = "Educational")
session3
```

```
## <polite session> https://www.amazon.com/s?k=women&rh=n%3A11056591&ref=nb_sb_noss
##     User-agent: Educational
##     robots.txt: 138 rules are defined for 5 bots
##    Crawl delay: 5 sec
##   The path is scrapable for this user-agent
```

```
library(rvest)

amazondf3 <- data.frame()
```

```r
page3 <- scrape(session3)

price3 <- page3 %>%
  html_nodes('.a-price .a-offscreen') %>%
  html_text(trim = TRUE) %>%
  gsub("[^0-9\\.]", "", .) %>%
  head(30)

description3 <- page3 %>%
  html_nodes('.a-color-base.a-text-normal') %>%
  html_text() %>%
  head(30)

ratings3 <- page3 %>%
  html_nodes('.a-icon-alt') %>%
  html_text() %>%
  gsub(" out of 5 stars", "", .) %>%
  head(30)

amazondf3 <- data.frame(
  Prices = price3,
  Descriptions = description3,
  Ratings = ratings3,
   stringsAsFactors = FALSE)

amazondf3
```

```
##    Prices
## 1   23.43
## 2    6.89
## 3   24.99
## 4   55.94
## 5   18.65
## 6   88.00
## 7   13.48
## 8    1.60
## 9   19.95
## 10  11.21
## 11   4.48
## 12  14.95
## 13  24.99
## 14   7.57
## 15  70.00
## 16  57.00
## 17  16.76
## 18  95.00
## 19 124.99
## 20  36.76
## 21 168.00
## 22  85.50
## 23 114.00
## 24  46.93
## 25  27.61
## 26  74.29
```

```
## 27  50.00
## 28  14.71
## 29 100.00
## 30  26.25
##
## 1
## 2
## 3
## 4                          Victoria's Secret Bombshell Mini Fragrance Mist, Notes of Purple Passio
## 5                                                                                                E
## 6                                                    Ralph Lauren - Ralph - Eau de Toilette -
## 7
## 8  Calvin Klein Eternity Eau de Parfum - Floral Women's Perfume - With Notes of Bergamot, White Lily
## 9
## 10
## 11                                                                        Victoria's Secret Fragr
## 12
## 13                                                  Ariana Grande Cloud Eau de Parfum - Warm Gourmand
## 14                                                    Victoria's Secret Tease Eau de Parfum, Women's P
## 15                                                          Mugler Alien - Eau de Parfum - Wome
## 16
## 17
## 18                                                                                               I
## 19
## 20
## 21                                                       Victoria's Secret Love Spell Mini
## 22
## 23
## 24
## 25
## 26
## 27                                                                                          Paul
## 28
## 29
## 30
##    Ratings
## 1      4.4
## 2      4.6
## 3      4.6
## 4      4.5
## 5      4.6
## 6      4.7
## 7      4.3
## 8      4.6
## 9      4.7
## 10     4.6
## 11     4.6
## 12     4.5
## 13     4.6
## 14     4.6
## 15     4.6
## 16     4.6
## 17     4.7
## 18     4.1
```

```
## 19     4.2
## 20     4.7
## 21     4.6
## 22     4.7
## 23     4.4
## 24     4.6
## 25     4.6
## 26     4.6
## 27     4.5
## 28     4.3
## 29     4.6
## 30     4.3
```

```
nailcare_url <- "https://www.amazon.com/s?k=nail+polish&rh=n%3A17242866011&ref=nb_sb_noss"

session4 <- bow(nailcare_url,
                user_agent = "Educational")
session4
```

```
## <polite session> https://www.amazon.com/s?k=nail+polish&rh=n%3A17242866011&ref=nb_sb_noss
##     User-agent: Educational
##     robots.txt: 138 rules are defined for 5 bots
##    Crawl delay: 5 sec
##   The path is scrapable for this user-agent
```

```
library(rvest)

amazondf4 <- data.frame()

page4 <- scrape(session4)

price4 <- page4 %>%
  html_nodes('.a-price .a-offscreen') %>%
  html_text(trim = TRUE) %>%
  gsub("[^0-9\\.]", "", .) %>%
  head(30)

description4 <- page4 %>%
  html_nodes('.a-color-base.a-text-normal') %>%
  html_text() %>%
  head(30)

ratings4 <- page4 %>%
  html_nodes('.a-icon-alt') %>%
  html_text() %>%
  gsub(" out of 5 stars", "", .) %>%
  head(30)

amazondf4 <- data.frame(
  Prices = price4,
  Descriptions = description4,
  Ratings = ratings4,
   stringsAsFactors = FALSE)

amazondf4
```

```
##    Prices
## 1    9.59
## 2   19.18
## 3   11.99
## 4    9.15
## 5   18.30
## 6   10.79
## 7    9.15
## 8   18.30
## 9   10.79
## 10  13.59
## 11   3.62
## 12  19.99
## 13  24.69
## 14  49.38
## 15  37.12
## 16  13.58
## 17  25.99
## 18   9.99
## 19   9.99
## 20  13.99
## 21   7.99
## 22   9.99
## 23  17.71
## 24  17.71
## 25  25.30
## 26   9.99
## 27   7.57
## 28  12.99
## 29   5.52
## 30  11.04
##
## 1
## 2                  beetles Gel Polish 44 PCS Gel Nail Polish Set-36 Colors Gel Polish Set Base Coat
## 3                                  JODSONE 36 PCS Gel Nail Polish Set-32 Colors Gel Polish Kit Base
## 4           Modelones Christmas Nail Polish Set 6 Colors Red Green Glitter Nail polish Set Gold Silv
## 5          Beetles Nail Blooming Gel 15ml Clear Uv Led Blossom Gel Polish for Spreading Effect Marbl
## 6                                                                    OPI Infinite Shine Long Wea
## 7                                              GAOY Rose Garden Jelly Gel Nail Polish of 6 Transp
## 8                                                                          OPI Nail Lacq
## 9                                  Morovan Fingernail Nail Polish Set: 15 Color Burgundy Red Fall Fi
## 10                 Morovan Fingernail Nail Polish Set: Holographic Metallic Lacquer Air Dry Nail Pol
## 11 Bright Nail Polish Set for Girls & Teens - 7 Vibrant, Non-Toxic, Kid-Safe Colors for Every Day of
## 12                                                                         OPI Nail Lacq
## 13
## 14       Beetles Red Gel Nail Polish, 6Pcs Candy Cane Christmas Gel Polish Glitter Burgundy Red Spar
## 15         Beetles Pearl Gel Nail Polish, 6 Colors Shimmer Pearl White Pink Purple Mermaid Nail Drau
## 16
## 17   3C4G Celestial 12-Pack Nail Polish Tower for Girls & Teens - 12 Vibrant, Non-Toxic Colors - Sar
## 18                   modelones Nail Polish 6 Colors Neutral Nude Nail Polish Set Nude Pink Quick Dry
## 19
## 20
## 21        Beetles Gel Polish 44 PCS Gel Nail Polish Set 36 Colors Gel Nail Polish with Base Coat Gl
## 22                                                                         OPI Nail Lac
```

```
## 23                                                                                  Sally H
## 24
## 25                                                  Morovan Brown Nail Polish Set: Fall Winter Nude Nail
## 26
## 27                                                                                     OPI Nail Lacquer N
## 28                                                                                  OPI Nail Lacquer Nail
## 29
## 30
##    Ratings
## 1      4.2
## 2      4.5
## 3      4.3
## 4      4.2
## 5      4.6
## 6      4.5
## 7      4.2
## 8      4.6
## 9      4.2
## 10     4.1
## 11     4.3
## 12     4.8
## 13     3.4
## 14     4.5
## 15     4.6
## 16     4.3
## 17     4.5
## 18     4.1
## 19     4.6
## 20     4.3
## 21     4.3
## 22     4.7
## 23     4.1
## 24     4.5
## 25     4.5
## 26     4.6
## 27     4.3
## 28     4.7
## 29     4.4
## 30     4.7
```

```
 Haircare_url <- "https://www.amazon.com/b/?node=11057971&ref_=Oct_d_odnav_d_11057241_0&pd_rd_w=5q3s3&c

session5 <- bow(Haircare_url, user_agent = "Educational")
session5
```

```
## <polite session> https://www.amazon.com/b/?node=11057971&ref_=Oct_d_odnav_d_11057241_0&pd_rd_w=5q3s3&
##     User-agent: Educational
##     robots.txt: 138 rules are defined for 5 bots
##    Crawl delay: 5 sec
##   The path is scrapable for this user-agent
```

```
library(rvest)

amazondf5 <- data.frame()
```

```r
page5 <- scrape(session5)

price5 <- page5 %>%
  html_nodes('.a-price .a-offscreen') %>%
  html_text(trim = TRUE) %>%
  gsub("[^0-9\\.]", "", .) %>%
  head(30)

description5 <- page5 %>%
  html_nodes('.a-color-base.a-text-normal') %>%
  html_text() %>%
  head(30)

ratings5 <- page5 %>%
  html_nodes('.a-icon-alt') %>%
  html_text() %>%
  gsub(" out of 5 stars", "", .) %>%
  head(30)

amazondf5 <- data.frame(
  Prices = price5,
  Descriptions = description5,
  Ratings = ratings5,
   stringsAsFactors = FALSE)

amazondf5
```

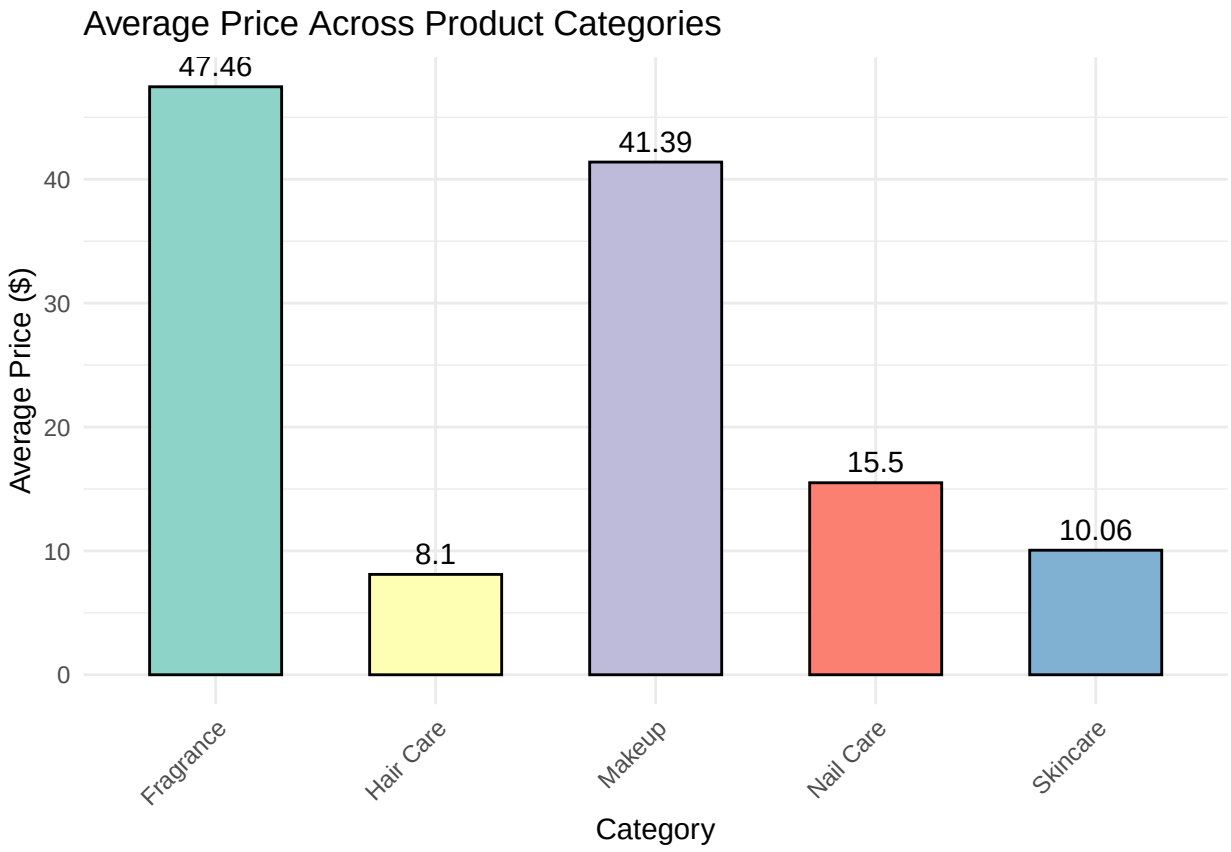
```
##    Prices
## 1    4.79
## 2    0.02
## 3    8.99
## 4   14.98
## 5    0.29
## 6   29.99
## 7    9.95
## 8    1.24
## 9    7.99
## 10   1.00
## 11   8.99
## 12   7.99
## 13   7.99
## 14  10.00
## 15  12.00
## 16   4.00
## 17  31.98
## 18   7.63
## 19   1.27
## 20   8.98
## 21   8.50
## 22   0.01
## 23   9.99
## 24   6.28
## 25   0.52
## 26   7.85
```

```
## 27   6.59
## 28   3.30
## 29   9.99
## 30   9.99
##
## 1  Goody Hair Accessories Kit, Everyday Essentials for Women - Ouchless Damage-free Hair Ties, Slidel
## 2                                                 LuSeren Hair Clips for Women 4.3 Inch Large Hair C
## 3                                               Matte Hair Clips for Women and Girls - Rectangle and
## 4       Kitsch Metal French Hair Pins for Women, Gold French Pins for Thick Hair, U Shaped Hair Pins
## 5                                                          Hicober Microfiber Hair Towel, 3 Pac
## 6      Kitsch Satin Hair Scrunchies for Women, Softer Than Silk Scrunchies for Hair, Satin Scrunchi
## 7             LUKACY 6 Pack Large Metal Hair Claw Clips - 4 Inch Big gold hair clips,Perfect Jaw ha
## 8                                                                                        Teens
## 9                                12 Pack Square Claw Clips, Big and Small Neutral Rectangle Hair Cl
## 10                                                        Silky Satin Oversized Hair Bows and H
## 11                                                TOCESS 8 Pack Big Hair Claw Clips for Women Large
## 12        Goody Ouchless Forever Polyband Hair Ties - 150 Ct, Clear, Hair Bands for Women's Hair, Ea
## 13                Kitsch Nylon Hair Elastics, No Tangle & No Breakage Hair Ties for Ponytails, Updo
## 14                                                                      Winkeyes Hair Styl
## 15                          Glitter Mini Butterfly Hair Claw Clips, Sparkly Colorful Hair Accessories
## 16 AIMIKE 6pcs Professional Hair Clips for Styling Sectioning, Non Slip No-Trace Duck Billed Hair Cli
## 17                          Aigee 6pcs Topsy Tail and Hair Loop Styling Tool Set - Ponytail Makers, H
## 18                                                                    Medium Claw Hair Cl
## 19                                              27Pcs Hair Styling Set, Hair Design Styling Tools, DIY H
## 20                                                    Teenitor Clear Elastic Hair Band
## 21                                                                      Flower Claw Clip
## 22                                Kitsch Gold Metal Claw Clips, Large Hair Clips for All Hair Types
## 23                               YANIBEST 5 Pcs Flat Hair Clips - Stylish Lay Flat Claw Clips for All
## 24                               8PCS Hair Clips for Women, Flower Claw Clips for Thick Hair, Non-Sli
## 25       8Pcs No Bend No Crease Hair Clips- Styling Duck Bill Clips Alligator Hair Barrettes for Sty
## 26                                          Funtopia Hair Clips for Girls,100 Pcs No Slip Metal Snap, H
## 27                                                                          Teenitor H
## 28                     Goody SlideProof Bobby Pins - 48 Ct, Brunette Brown, High Gloss Bobby Hair Pi
## 29                                                        Sparkling Crystal Stone Braided Hair
## 30  Camila Paris AD66/2 French Side Comb Large Curved Tortoiseshell Hair Combs for Women Fine Hair, 4
##    Ratings
## 1     4.5.
## 2     4.1.
## 3      4.2
## 4      4.7
## 5      4.6
## 6      4.4
## 7      4.6
## 8      4.6
## 9      4.6
## 10     4.6
## 11     4.7
## 12     4.5
## 13     4.7
## 14     4.3
## 15     4.5
## 16     3.9
## 17     4.6
## 18     4.6
```

```
## 19     4.4
## 20     4.5
## 21     3.9
## 22     4.4
## 23     4.6
## 24     4.4
## 25     4.3
## 26     4.4
## 27     4.3
## 28     4.7
## 29     4.5
## 30     4.6
```

6. Describe the data you have extracted.
7. What will be your use case for the data you have extracted?
8. Create graphs regarding the use case. And briefly explain it.

```r
library(ggplot2)
library(dplyr)

amazondf_all <- rbind(
cbind(Category = "Makeup", amazondf1),
cbind(Category = "Skincare", amazondf2),
cbind(Category = "Fragrance", amazondf3),
cbind(Category = "Nail Care", amazondf4),
cbind(Category = "Hair Care", amazondf5))

amazondf_all$Prices <- as.numeric(amazondf_all$Prices)

avg_prices <- amazondf_all %>%
group_by(Category) %>%
summarise(Average_Price = mean(Prices, na.rm = TRUE))

ggplot(avg_prices, aes(x = Category, y = Average_Price, fill = Category)) +
geom_bar(stat = "identity", show.legend = FALSE, color = "black", width = 0.6) +
geom_text(aes(label = round(Average_Price, 2)),
vjust = -0.5, size = 4, color = "black") +
labs(title = "Average Price Across Product Categories",
x = "Category",
y = "Average Price ($)") +
theme_minimal() +
scale_fill_brewer(palette = "Set3") +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

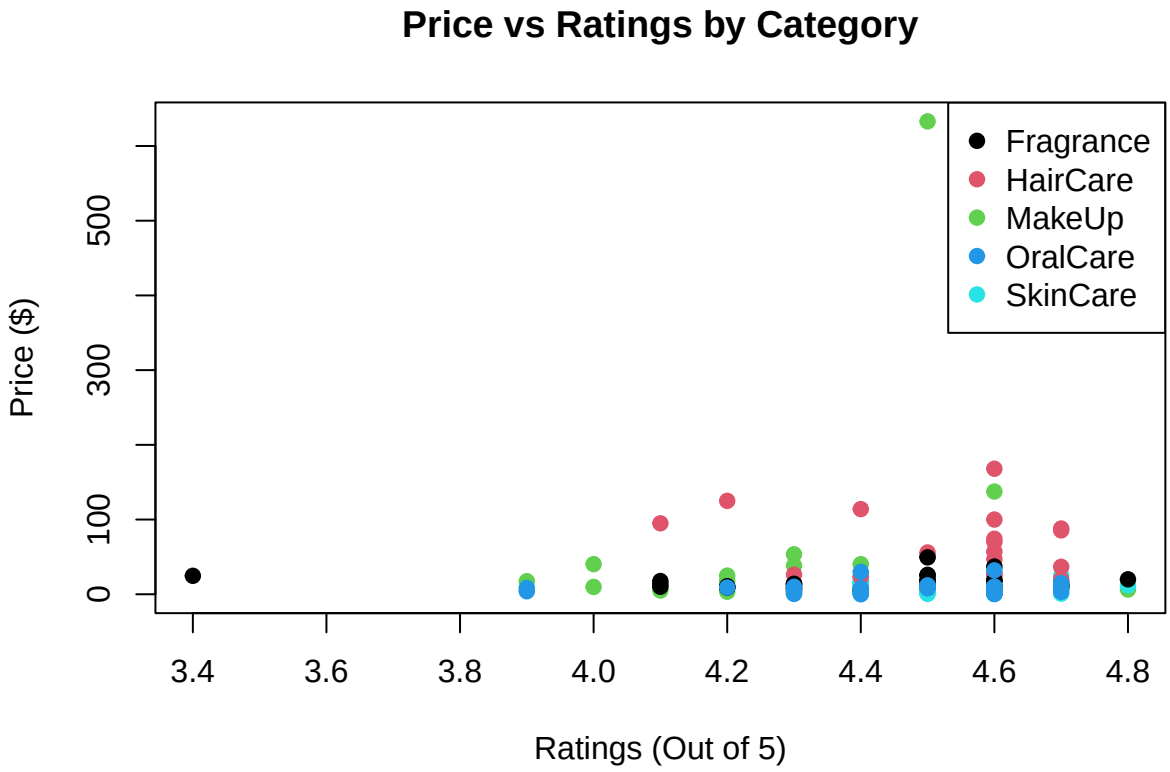


9. Graph the price and the ratings for each category. Use basic plotting functions and ggplot2 package.

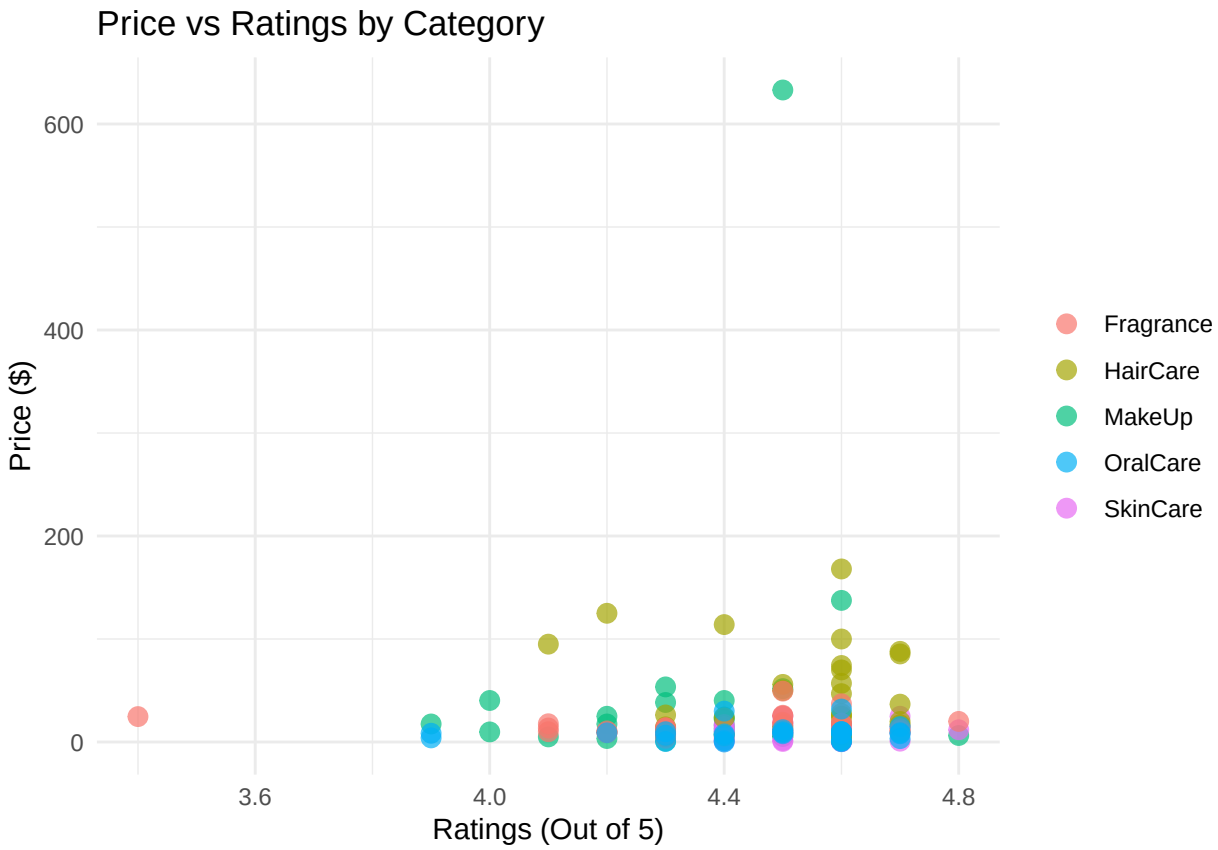
```
amazondf1$Category <- "MakeUp"
amazondf2$Category <- "SkinCare"
amazondf3$Category <- "HairCare"
amazondf4$Category <- "Fragrance"
amazondf5$Category <- "OralCare"


combined_df <- rbind(amazondf1, amazondf2, amazondf3, amazondf4, amazondf5)

combined_df$Prices <- as.numeric(combined_df$Prices)
combined_df$Ratings <- as.numeric(combined_df$Ratings)
```

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

```
plot_data <- combined_df[!is.na(combined_df$Prices) & !is.na(combined_df$Ratings), ]
category_colors <- as.factor(plot_data$Category)

#Basic Plotting
plot(plot_data$Ratings, plot_data$Prices,
col = category_colors,
pch = 19,
xlab = "Ratings (Out of 5)",
ylab = "Price ($)",
main = "Price vs Ratings by Category")

legend("topright", legend = levels(category_colors), col = 1:5, pch = 19)
```




```
library(ggplot2)

ggplot(plot_data, aes(x = Ratings, y = Prices, color = Category)) +
geom_point(size = 3, alpha = 0.7) +
labs(title = "Price vs Ratings by Category",
x = "Ratings (Out of 5)",
y = "Price ($)") +
theme_minimal() +
theme(legend.title = element_blank())
```



10. Rank the products of each category by price and ratings. Explain briefly

```
library(dplyr)

amazondf1$Category <- "Makeup"
amazondf2$Category <- "Skincare"
amazondf3$Category <- "Haircare"
amazondf4$Category <- "Fragrance"
amazondf5$Category <- "Oral Care"

combined_df <- rbind(amazondf1, amazondf2, amazondf3, amazondf4, amazondf5)

combined_df$Prices <- as.numeric(combined_df$Prices)
combined_df$Ratings <- as.numeric(combined_df$Ratings)
```

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

```
ranked_df <- combined_df %>%
arrange(Category, desc(Ratings), Prices) %>%
group_by(Category) %>%
mutate(Rank = row_number())

top_ranked <- ranked_df %>%
filter(Rank <= 5)

top_ranked
```

```
## # A tibble: 25 x 5
## # Groups:   Category [5]
```

```
##    Prices Descriptions                                      Ratings Category  Rank
##     <dbl> <chr>                                               <dbl> <chr>    <int>
##  1  20.0  OPI Nail Lacquer Nail Polish | Opaque Light Re~     4.8 Fragran~     1
##  2   9.99 OPI Nail Lacquer Nail Polish | Opaque Dark Red~     4.7 Fragran~     2
##  3  11.0  OPI Top Coat | Nail Polish Top Coat | Prevents~     4.7 Fragran~     3
##  4  13.0  OPI Nail Lacquer Nail Polish | Opaque Dark Bla~     4.7 Fragran~     4
##  5   9.99 Essie Gel Couture Longwear Nail Polish Kit, Sh~     4.6 Fragran~     5
##  6  16.8  Versace Yellow Diamond for Women 3.0 oz Eau de~     4.7 Haircare     1
##  7  20.0  Daisy By Marc Jacobs for Women Eau De Toilette~     4.7 Haircare     2
##  8  36.8  Gucci Bamboo by Gucci for Women 2.5 oz Eau de ~     4.7 Haircare     3
##  9  85.5  Ed Hardy Women's Perfume Fragrance by Christia~     4.7 Haircare     4
## 10  88    Ralph Lauren - Ralph - Eau de Toilette - Women~     4.7 Haircare     5
## # i 15 more rows
```

```r
library(ggplot2)

ggplot(top_ranked, aes(x = reorder(Descriptions, -Rank), y = Ratings, fill = Category)) +
geom_col(show.legend = FALSE) +
coord_flip() +
labs(
title = "Top 5 Ranked Products by Category",
x = "Product Description",
y = "Ratings (Out of 5)") +
facet_wrap(~Category, scales = "free_y") +
theme_minimal()
```