Worksheet#5a_group

Cadiz, Guion, Jacildo

2024-11-06

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

sesh <- bow(url_Imdb,
              user_agent = "Educational")
sesh
```

```
## <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 <- character(0)
rank <- character(0)
rating <- character(0)

title <- scrape(sesh) %>%
  html_nodes('h3.ipc-title__text') %>%
  html_text


dataB <- data.frame(
    titleDf = title[1:50])

dataB
```

```
##                                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
## 26                           25. The Office
## 27                         Recently viewed
## 28                                    <NA>
## 29                                    <NA>
## 30                                    <NA>
## 31                                    <NA>
## 32                                    <NA>
## 33                                    <NA>
```

```
## 34                                    <NA>
## 35                                    <NA>
## 36                                    <NA>
## 37                                    <NA>
## 38                                    <NA>
## 39                                    <NA>
## 40                                    <NA>
## 41                                    <NA>
## 42                                    <NA>
## 43                                    <NA>
## 44                                    <NA>
## 45                                    <NA>
## 46                                    <NA>
## 47                                    <NA>
## 48                                    <NA>
## 49                                    <NA>
## 50                                    <NA>
```

```r
colnames(dataB) <- "ranks"

split_df <- strsplit(as.character(dataB$ranks),".",fixed = TRUE)
split_df <- data.frame(do.call(rbind,split_df))

colnames(split_df) <- c("ranks","title")

head(split_df)
```

```
##         ranks             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
```

Extracting Amazon Product Reviews

4. Select 5 categories from Amazon and select 30 products from each category.

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

url <- 'https://www.amazon.com/s?rh=n%3A3760911%2Cn%3A11058281&dc&qid=1730855869&rnid=3760911&ref=sr_nr

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

```
## <polite session> https://www.amazon.com/s?rh=n%3A3760911%2Cn%3A11058281&dc&qid=1730855869&rnid=37609
##     User-agent: Educational
##     robots.txt: 137 rules are defined for 4 bots
##    Crawl delay: 5 sec
##   The path is scrapable for this user-agent
```

5. Extract the price, description, ratings and reviews of each product.

```
library(rvest)

page1 <- read_html(url)
price1 <- scrape(session) %>%
  html_nodes('.a-price .a-offscreen') %>%
  html_text

data1 <- data.frame(
    priceDf1 = price1[1:30])

data1
```

```
##    priceDf1
## 1      <NA>
## 2      <NA>
## 3      <NA>
## 4      <NA>
## 5      <NA>
## 6      <NA>
## 7      <NA>
## 8      <NA>
## 9      <NA>
## 10     <NA>
## 11     <NA>
## 12     <NA>
## 13     <NA>
## 14     <NA>
## 15     <NA>
## 16     <NA>
## 17     <NA>
## 18     <NA>
## 19     <NA>
## 20     <NA>
## 21     <NA>
## 22     <NA>
## 23     <NA>
## 24     <NA>
## 25     <NA>
## 26     <NA>
## 27     <NA>
## 28     <NA>
## 29     <NA>
## 30     <NA>
```

```
description1 <- page1 %>% html_nodes('span.a-size-large product-title-word-break') %>% html_text()

data2 <- data.frame(
    desDf1 = description1[1:30])

data2
```

```
##    desDf1
## 1    <NA>
## 2    <NA>
## 3    <NA>
## 4    <NA>
## 5    <NA>
## 6    <NA>
## 7    <NA>
## 8    <NA>
## 9    <NA>
## 10   <NA>
## 11   <NA>
## 12   <NA>
## 13   <NA>
## 14   <NA>
## 15   <NA>
## 16   <NA>
## 17   <NA>
## 18   <NA>
## 19   <NA>
## 20   <NA>
## 21   <NA>
## 22   <NA>
## 23   <NA>
## 24   <NA>
## 25   <NA>
## 26   <NA>
## 27   <NA>
## 28   <NA>
## 29   <NA>
## 30   <NA>
```

```
ratings1 <- scrape(session) %>%
  html_nodes('span.a-icon-alt') %>%
  html_text

data3 <- data.frame(
    ratingDf1 = ratings1[1:30])

data3
```

```
##    ratingDf1
## 1       <NA>
## 2       <NA>
## 3       <NA>
## 4       <NA>
## 5       <NA>
## 6       <NA>
```

```
## 7       <NA>
## 8       <NA>
## 9       <NA>
## 10      <NA>
## 11      <NA>
## 12      <NA>
## 13      <NA>
## 14      <NA>
## 15      <NA>
## 16      <NA>
## 17      <NA>
## 18      <NA>
## 19      <NA>
## 20      <NA>
## 21      <NA>
## 22      <NA>
## 23      <NA>
## 24      <NA>
## 25      <NA>
## 26      <NA>
## 27      <NA>
## 28      <NA>
## 29      <NA>
## 30      <NA>
```

```
reviews1 <- scrape(session) %>%
  html_nodes('#acrCustomerReviewText') %>%
  html_text

data4 <- data.frame(
    reviewsDf1 = reviews1[1:30])

data4
```

```
##    reviewsDf1
## 1        <NA>
## 2        <NA>
## 3        <NA>
## 4        <NA>
## 5        <NA>
## 6        <NA>
## 7        <NA>
## 8        <NA>
## 9        <NA>
## 10       <NA>
## 11       <NA>
## 12       <NA>
## 13       <NA>
## 14       <NA>
## 15       <NA>
## 16       <NA>
## 17       <NA>
## 18       <NA>
## 19       <NA>
## 20       <NA>
```

```
## 21       <NA>
## 22       <NA>
## 23       <NA>
## 24       <NA>
## 25       <NA>
## 26       <NA>
## 27       <NA>
## 28       <NA>
## 29       <NA>
## 30       <NA>
```

6. Describe the data you have extracted.

```
library(psych)
describe(data1)
```

```
## Warning in describe.1(x = x, na.rm = na.rm, interp = interp, skew = skew, : You
## were trying to describe a non-numeric data.frame or vector which describe
## converted to numeric.
```

```
## Warning in min(x, na.rm = na.rm): no non-missing arguments to min; returning
## Inf
```

```
## Warning in max(x, na.rm = na.rm): no non-missing arguments to max; returning
## -Inf
```

```
##          vars n mean sd median trimmed mad min  max range skew kurtosis se
## priceDf1    1 0  NaN NA     NA     NaN  NA Inf -Inf  -Inf   NA       NA NA
```

7. What will be your use case for the data you have extracted?

8. Create graphs regarding the use case. And briefly explain it.

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

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