

RWorksheet#5_Group(Basa,Llanera,Tuares)

2024-11-06

```
# Load necessary libraries
library(polite)
library(rvest)
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(purrr)
library(ggplot2)

# Function to scrape products from a given category URL
scrape_amazon_category <- function(category_url, category_name) {
  # Create a polite session
  session <- bow(category_url, user_agent = "Educational")

  # Scrape the page
  page <- scrape(session)

  # Extract product details
  products <- page %>%
    html_nodes(".s-main-slot .s-result-item") %>%
    map_df(~ {
      tibble(
        Title = .x %>% html_node("h2") %>% html_text(trim = TRUE),
        Price = .x %>% html_node(".a-price .a-offscreen") %>% html_text(trim = TRUE),
        Description = .x %>% html_node(".a-text-normal") %>% html_text(trim = TRUE),
        Rating = .x %>% html_node(".a-icon-alt") %>% html_text(trim = TRUE),
        Reviews = .x %>% html_node(".a-size-small .a-link-normal") %>% html_text(trim = TRUE),
        Category = category_name # Add category name here
      )
    })

  return(products)
}

categories <- list(
  computers = 'https://www.amazon.com/s?k=computers',
```

```

luggage = 'https://www.amazon.com/s?k=luggage',
dress = 'https://www.amazon.com/s?k=dress2',
TV = 'https://www.amazon.com/s?k=TV',
ballpen = 'https://www.amazon.com/s?k=ballpen'
)

all_products <- tibble()

for (category_name in names(categories)) {
  category_url <- categories[[category_name]]
  category_products <- scrape_amazon_category(category_url, category_name)
  all_products <- bind_rows(all_products, category_products)

  if (nrow(all_products) > 30) {
    all_products <- all_products %>% slice(1:30)
  }
}

all_products$Price <- as.numeric(gsub("\\$", "", gsub(",", "", all_products$Price)))
all_products$Rating <- as.numeric(gsub(" out of 5 stars", "", all_products$Rating))

## Warning: NAs introduced by coercion
print(all_products)

```

```

## # A tibble: 30 x 6
##   Title                                Price Description Rating Reviews Category
##   <chr>                                <dbl> <chr>         <dbl> <chr>   <chr>
## 1 <NA>                                NA    <NA>         NA    <NA>   compute~
## 2 "Results"                          NA    "Check eac~   NA    <NA>   compute~
## 3 "Acer Aspire 3 A315-24P-R7VH Slim ~ 280. "Acer Aspi~   4.4 3 capa~ compute~
## 4 "HP Portable Laptop, Student and B~ 203. "HP Portab~   4.1 5 capa~ compute~
## 5 "HP 2022 Newest All-in-One Desktop~ 425. "HP 2022 N~   4.3 937    compute~
## 6 "More results"                      NA    <NA>         NA    <NA>   compute~
## 7 "HP Newest 14\" Ultral Light Lapto~ 250. "HP Newest~   4    1,924  compute~
## 8 "Dell Vostro 3910 Desktop Computer~ 500. "Dell Vost~   4.4 76     compute~
## 9 "HP 14 Laptop, Intel Celeron N4020~ 178. "HP 14 Lap~   4    1,844  compute~
## 10 "HP Elite Desktop PC Computer Inte~ 125  "HP Elite ~   3.8 7,490  compute~
## # i 20 more rows

```

```

ggplot(all_products, aes(x = Title, y = Price, fill = Category)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  theme_minimal() +
  labs(title = "Price of Products by Category", x = "Product", y = "Price ($)")

```

```

## Warning: Removed 9 rows containing missing values or values outside the scale range
## (`geom_bar()`).

```

reen Slim Thin Light Laptop Computer, 8-Core MediaTek Kompanio 520 Processor, 4GB RAM, 64GB eMMC, WiFi 6, Business, 14" HD Display, Intel Quad-Core N4120, 16GB DDR4 RAM, 64GB eMMC, 1 Year Office 365, Webcam, RJ-45, Computer, 21.5" FHD Display, Intel Celeron J4025 Processor, 16GB RAM, 256GB SSD, Webcam, HDMI, RJ-45, Wired Business, Intel Quad-Core N4120, 8GB RAM, 192GB Storage(64GB eMMC+128GB Micro SD), 1 Year Office 365, Wktop PC Computer Intel Core i5 3.1-GHz, 8 gb Ram, 1 TB Hard Drive, DVDRW, 19 Inch LCD Monitor, Keyboard, Mouse, Microsoft Office Trial . FHD Display . 32GB RAM . 1TB Storage (512GB SSD + 512GB External HD) . Intel Celeron FHD Display, Intel Celeron J4025 Processor, 16GB RAM, 512GB PCIe SSD, Webcam, HDMI, RJ-45, Wired Keyboard HP 17 Business Laptop, 17.3" HD+ Display, 11th Gen Intel Core i3-1125G4 Processor, 32GB RAM, 1TB SSD, Wi-Fi 6 64 GB Storage, 14-inch Micro-edge HD Display, Windows 11 Home, Thin & Portable, 4K Graphics, One Year of Micro Gen Intel 4-Core Processor | 16GB DDR4 RAM | 256GB SSD + 1TB HDD | Intel UHD Graphics 730 | Wi-Fi 6 | DVD-R Writer Desktop PC, Intel Core i5 3rd Gen 3.2 GHz, 16GB RAM, 2TB HDD, New 22 Inch LED Monitor, RGB Keyboard and with Intel Core i7-4770 Upto 3.9GHz, HD Graphics 4600 4K Support, 32GB RAM, 1TB SSD, DisplayPort, HDMI, Wi-Fi 0 SFF Desktop PC Intel i7-7700 4-Cores 3.60GHz 32GB DDR4 1TB SSD Wi-Fi BT HDMI Dual Monitor Support Wind

chip: Built for Apple Intelligence, 13.6-inch Liquid Retina Display, 8GB RAM, 256GB SSD Storage, Backlit Keyboard, M2 chip: Built for Apple Intelligence, 13.6-inch Liquid Retina Display, 8GB RAM, 256GB SSD Storage, Backlit Keyboard

PS Display | AMD Ryzen 3 7320U Quad-Core Processor | AMD Radeon Graphics | 8GB LPDDR5 | 128GB NVMe SSD

```
ggplot(all_products, aes(x = Title, y = Rating, fill = Category)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  theme_minimal() +
  labs(title = "Ratings of Products by Category", x = "Product", y = "Rating (out of 5)")
```

```
## Warning: Removed 9 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```

reen Slim Thin Light Laptop Computer, 8-Core MediaTek Kompanio 520 Processor, 4GB RAM, 64GB eMMC, WiFi 6, Business, 14" HD Display, Intel Quad-Core N4120, 16GB DDR4 RAM, 64GB eMMC, 1 Year Office 365, Webcam, RJ-45, Computer, 21.5" FHD Display, Intel Celeron J4025 Processor, 16GB RAM, 256GB SSD, Webcam, HDMI, RJ-45, Wired Business, Intel Quad-Core N4120, 8GB RAM, 192GB Storage(64GB eMMC+128GB Micro SD), 1 Year Office 365, Wktop PC Computer Intel Core i5 3.1-GHz, 8 gb Ram, 1 TB Hard Drive, DVDRW, 19 Inch LCD Monitor, Keyboard, Mouse, Microsoft Office Trial . FHD Display . 32GB RAM . 1TB Storage (512GB SSD + 512GB External HD) . Intel Celeron FHD Display, Intel Celeron J4025 Processor, 16GB RAM, 512GB PCIe SSD, Webcam, HDMI, RJ-45, Wired Keyboard HP 17 Business Laptop, 17.3" HD+ Display, 11th Gen Intel Core i3-1125G4 Processor, 32GB RAM, 1TB SSD, Wi-Fi 6 64 GB Storage, 14-inch Micro-edge HD Display, Windows 11 Home, Thin & Portable, 4K Graphics, One Year of Micro Gen Intel 4-Core Processor | 16GB DDR4 RAM | 256GB SSD + 1TB HDD | Intel UHD Graphics 730 | Wi-Fi 6 | DVD-R Writer Desktop PC, Intel Core i5 3rd Gen 3.2 GHz, 16GB RAM, 2TB HDD, New 22 Inch LED Monitor, RGB Keyboard and with Intel Core i7-4770 Upto 3.9GHz, HD Graphics 4600 4K Support, 32GB RAM, 1TB SSD, DisplayPort, HDMI, Wi-Fi 0 SFF Desktop PC Intel i7-7700 4-Cores 3.60GHz 32GB DDR4 1TB SSD Wi-Fi BT HDMI Dual Monitor Support Wind

chip: Built for Apple Intelligence, 13.6-inch Liquid Retina Display, 8GB RAM, 256GB SSD Storage, Backlit Keyboard, M2 chip: Built for Apple Intelligence, 13.6-inch Liquid Retina Display, 8GB RAM, 256GB SSD Storage, Backlit Keyboard

PS Display | AMD Ryzen 3 7320U Quad-Core Processor | AMD Radeon Graphics | 8GB LPDDR5 | 128GB NVMe SSD

```
ranked_by_price <- all_products %>%
  group_by(Category) %>%
  arrange(Price) %>%
  mutate(Price_Rank = row_number())

ranked_by_rating <- all_products %>%
  group_by(Category) %>%
  arrange(desc(Rating)) %>%
  mutate(Rating_Rank = row_number())
```

```
print(ranked_by_price)
```

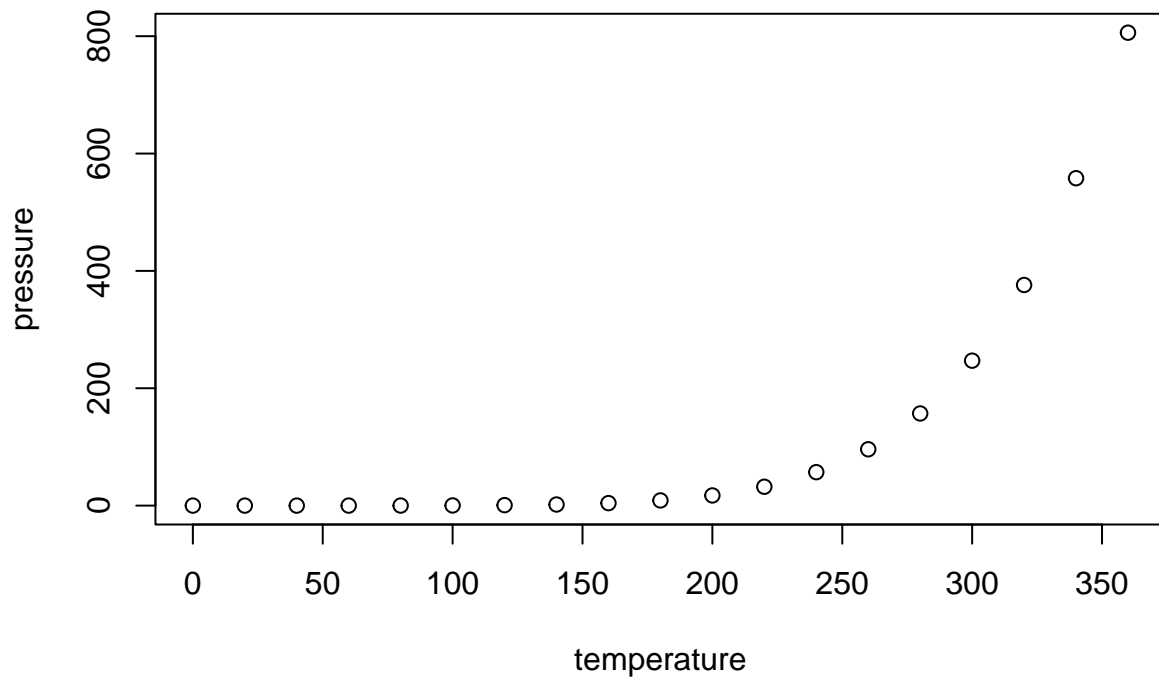
```
## # A tibble: 30 x 7
## # Groups:   Category [2]
##   Title Price Description Rating Reviews Category Price_Rank
##   <chr> <dbl> <chr> <dbl> <chr> <chr> <int>
## 1 Amazon Basics 105. Expandable~ 4.5 5 sizes luggage 1
## 2 HP Elite Desktop PC Com~ 125 HP Elite D~ 3.8 7,490 compute~ 1
## 3 Cosbarn 130. 5-Piece Lu~ 4.4 2 sizes luggage 2
## 4 Cosbarn 130. Luggage Se~ 4.2 67 luggage 3
## 5 Samsonite 140. Freeform H~ 4.4 5 sizes luggage 4
## 6 Dell OptiPlex Computer ~ 172. Dell OptiP~ 3.6 2,041 compute~ 2
## 7 HP 14 Laptop, Intel Cel~ 178. HP 14 Lapt~ 4 1,844 compute~ 3
## 8 Coolife 190. Expandable~ 4.6 4 sizes luggage 5
## 9 Dell Optiplex 9020 Smal~ 200 Dell Optip~ 4 3,556 compute~ 4
## 10 Lenovo Newest Flagship ~ 202. Lenovo New~ 4.3 486 compute~ 5
## # i 20 more rows
```

```
print(ranked_by_rating)
```

```
## # A tibble: 30 x 7
## # Groups:   Category [2]
##   Title Price Description Rating Reviews Category Rating_Rank
##   <chr> <dbl> <chr> <dbl> <chr> <chr> <int>
## 1 "HP Newest All-in-One ~ 400. "HP Newest~ 4.8 8 sizes compute~ 1
## 2 "Apple 2022 MacBook Ai~ 749 "Apple 202~ 4.7 1,067 compute~ 2
## 3 "Apple 2022 MacBook Ai~ 749 "Apple 202~ 4.7 550 compute~ 3
## 4 "Coolife" 190. "Expandabl~ 4.6 4 sizes luggage 1
## 5 "Amazon Basics" 105. "Expandabl~ 4.5 5 sizes luggage 2
## 6 "Acer Aspire 3 A315-24~ 280. "Acer Aspi~ 4.4 3 capa~ compute~ 4
## 7 "Dell Vostro 3910 Desk~ 500. "Dell Vost~ 4.4 76 compute~ 5
## 8 "Cosbarn" 130. "5-Piece L~ 4.4 2 sizes luggage 3
## 9 "Samsonite" 140. "Freeform ~ 4.4 5 sizes luggage 4
## 10 "HP 2022 Newest All-in~ 425. "HP 2022 N~ 4.3 937 compute~ 6
## # i 20 more rows
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.