Worksheet#5_group(Berja,Bibit,Buenvenida)

Berja,Bibit,Buenvenida

2024-11-06

Extracting Amazon Product Reviews

```
# 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
```

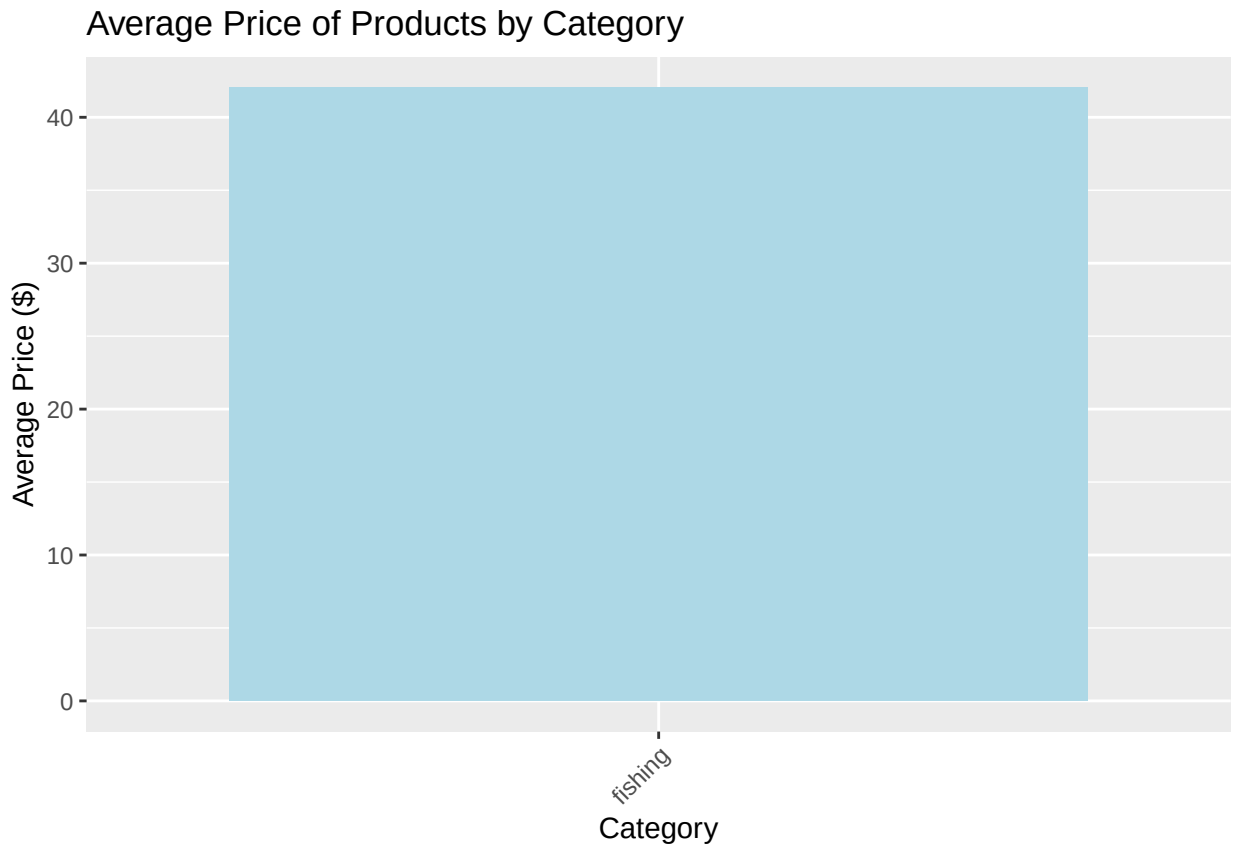
```
library(purrr)
library(ggplot2)  # Load ggplot2 for plotting

# 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)
```

```
}

# Example category URLs (you need to adjust these)
categories <- list(
  fishing = 'https://www.amazon.com/s?k=fishing',
  electronics = 'https://www.amazon.com/s?k=electronics',
  books = 'https://www.amazon.com/s?k=books',
  home_kitchen = 'https://www.amazon.com/s?k=home+kitchen',
  clothing = 'https://www.amazon.com/s?k=clothing'
)

# Initialize an empty data frame to store all products
all_products <- tibble()

# Loop through categories and scrape 30 products from each
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)

  # Limit to 30 products
  if (nrow(all_products) > 30) {
    all_products <- all_products %>% slice(1:30)
  }
}

# Convert Price and Rating to numeric for analysis
all_products$Price <- as.numeric(gsub("\\$", "", gsub(",", "", all_products$Price)))
all_products$Rating <- as.numeric(gsub(" out of 5 stars", "", all_products$Rating))

# Display the extracted data
print(all_products)
```

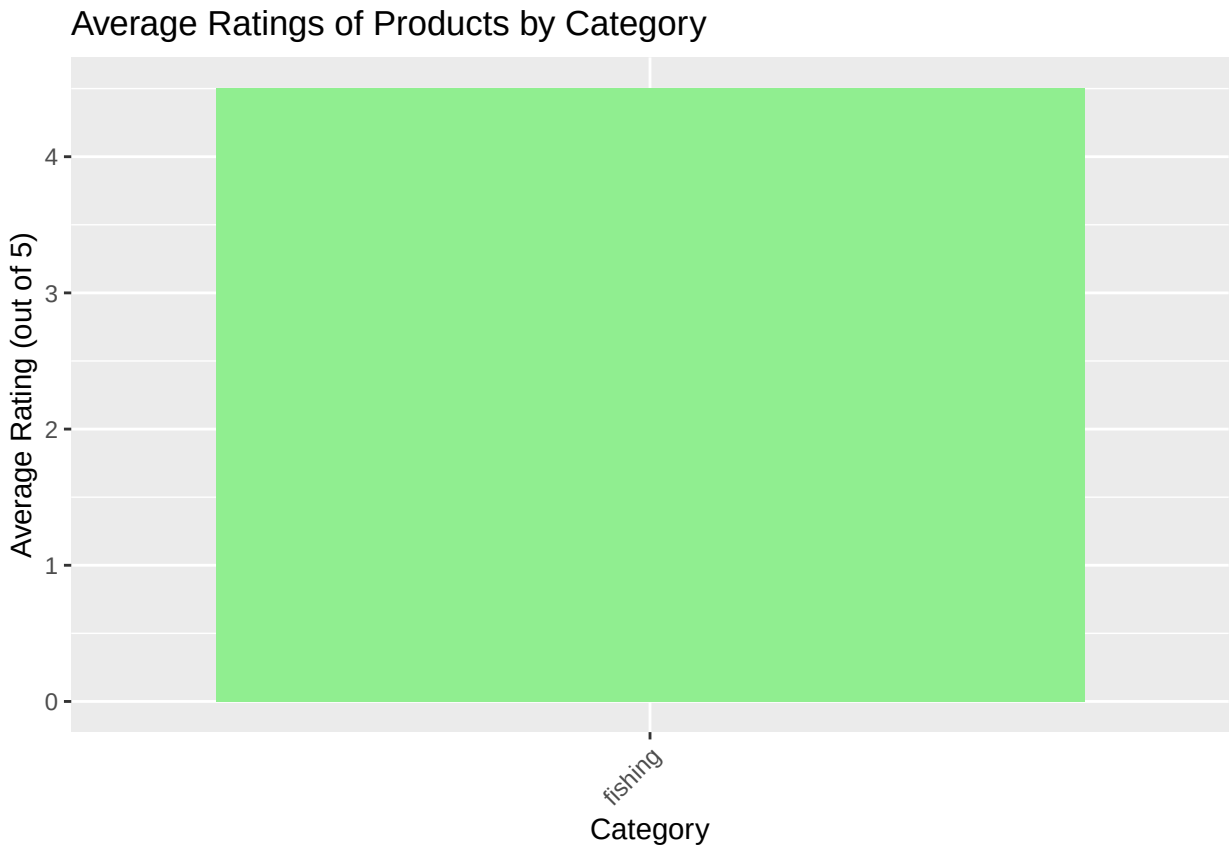
```
## # A tibble: 30 x 6
##    Title                                Price Description Rating Reviews Category
##    <chr>                                <dbl> <chr>        <dbl> <chr>   <chr>
##  1 <NA>                                  45.0 <NA>          NA   <NA>    fishing
##  2 Results                               NA   Check each~   NA   <NA>    fishing
##  3 KastKing SteelStream 6pc Fishing T~   32.0 KastKing S~    4.6 434     fishing
##  4 PLUSINNO 264/397pcs Fishing Access~   28.5 PLUSINNO 2~    4.6 1,174   fishing
##  5 Sougayilang Fishing Rod Combos wit~   49.9 Sougayilan~    4.3 5,701   fishing
##  6 More results                          NA   <NA>          NA   <NA>    fishing
##  7 275-Piece Fishing Lure Kit - Frogs~   17.8 275-Piece ~    4.4 4,144   fishing
##  8 PLUSINNO Fishing Lures, 137Pcs Tac~   14.4 PLUSINNO F~    4.4 351     fishing
##  9 ZACX Fish Lip Gripper Pliers - Upg~   16.0 ZACX Fish ~    4.7 14,981  fishing
## 10 Trump Topwater Fishing Lure           14.0 Trump Topw~    4.8 519     fishing
## # i 20 more rows
```

```
# Prepare data for bar plots
avg_price <- all_products %>%
  group_by(Category) %>%
  summarize(Average_Price = mean(Price, na.rm = TRUE))

avg_rating <- all_products %>%
```

```
  group_by(Category) %>%
  summarize(Average_Rating = mean(Rating, na.rm = TRUE))

# Create a bar plot for Average Price by Category using ggplot2
ggplot(avg_price, aes(x = reorder(Category, Average_Price), y = Average_Price)) +
  geom_bar(stat = "identity", fill = "lightblue") +
  labs(title = "Average Price of Products by Category",
       x = "Category",
       y = "Average Price ($)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
# Create a bar plot for Average Rating by Category using ggplot2
ggplot(avg_rating, aes(x = reorder(Category, Average_Rating), y = Average_Rating)) +
  geom_bar(stat = "identity", fill = "lightgreen") +
  labs(title = "Average Ratings of Products by Category",
       x = "Category",
       y = "Average Rating (out of 5)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




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

# Rank products by Ratings
ranked_by_rating <- all_products %>%
  group_by(Category) %>%
  arrange(desc(Rating)) %>%
  mutate(Rating_Rank = row_number())

# Display the ranked data
print(ranked_by_price)
```

```
## # A tibble: 30 x 7
## # Groups:   Category [1]
##    Title                    Price Description Rating Reviews Category Price_Rank
##    <chr>                    <dbl> <chr>        <dbl> <chr>   <chr>          <int>
##  1 Anezus Fishing Wire for~  6.99 Anezus Fis~    4.6 9,889   fishing            1
##  2 Easiest Fishing Knots --~  7.69 Easiest Fi~    4.7 2,319   fishing            2
##  3 KastKing Fish Scale, To~  8.96 KastKing F~    4.3 162     fishing            3
##  4 Trump Topwater Fishing ~ 14.0  Trump Topw~    4.8 519     fishing            4
##  5 Fishing Lures Kit for F~ 14.0  Fishing Lu~    4.4 5,332   fishing            5
##  6 PLUSINNO Fishing Lures,~ 14.4  PLUSINNO F~    4.4 351     fishing            6
##  7 ZACX Fish Lip Gripper P~ 16.0  ZACX Fish ~    4.7 14,981  fishing            7
##  8 KastKing Patented V15 V~ 16.0  KastKing P~    4.5 10,802  fishing            8
```

```
##  9 Ugly Stik Dock Runner S~ 17.4  Ugly Stik ~    4.5 3,580   fishing           9
## 10 275-Piece Fishing Lure ~ 17.8  275-Piece ~    4.4 4,144   fishing          10
## # i 20 more rows
```

```
print(ranked_by_rating)
```

```
## # A tibble: 30 x 7
## # Groups:   Category [1]
##    Title                   Price Description Rating Reviews Category Rating_Rank
##    <chr>                   <dbl> <chr>        <dbl> <chr>   <chr>           <int>
##  1 Trump Topwater Fishing~ 14.0  Trump Topw~    4.8 519     fishing             1
##  2 Columbia Unisex-Adult ~ 22.5  Columbia U~    4.8 7,603   fishing             2
##  3 ZACX Fish Lip Gripper ~ 16.0  ZACX Fish ~    4.7 14,981  fishing             3
##  4 Easiest Fishing Knots ~  7.69 Easiest Fi~    4.7 2,319   fishing             4
##  5 HUK Men's Rogue Wave S~ 68.0  HUK Men's ~    4.7 7,848   fishing             5
##  6 KastKing Karryall Fish~ 85.0  KastKing K~    4.7 1,031   fishing             6
##  7 KastKing SteelStream 6~ 32.0  KastKing S~    4.6 434     fishing             7
##  8 PLUSINNO 264/397pcs Fi~ 28.5  PLUSINNO 2~    4.6 1,174   fishing             8
##  9 Anezus Fishing Wire fo~  6.99 Anezus Fis~    4.6 9,889   fishing             9
## 10 KastKing BlowBak Tacti~ 30.0  KastKing B~    4.6 2,709   fishing            10
## # i 20 more rows
```