Project 1
Course: Digital Strategies for Social Science
Computational Social Science

Submitted By: Mishu Dhar

Title: Scraping an Online Children's Toy Store to Collect Data and Mine Valuable Insights

Content:

Tools used: R (mainly), Python (a little bit).

1. Introduction:

In this project, I closely examined data from an online children's toy store. Utilizing both R and Python, I extracted a variety of information from the website, including product names and prices, customer reviews, and images associated with the products. I also employed scikit-learn to analyze the colors used in these images, as I believed this information was essential for analyzing and understanding the price ranges of the online toy shop. By processing and analyzing this data, I was able to identify key trends in product popularity and customer sentiment. The color analysis further provided insights into visual preferences dominating the toy market.

2. A Short Description of the Scraping Process

I utilized R (specifically the rvest and Selenium packages) to extract essential information such as product names, prices, customer reviews, and the dates of these reviews. Initially, I attempted to extract images associated with the products using R but was unsuccessful. Instead, I switched to Python to handle this task and successfully stored all images in a single folder. Subsequently, I analyzed these images to identify the dominant colors by comparing them to the nearest RGB values. The results were saved in a CSV file for further analysis. Additionally, I employed regular expressions at several stages to clean the extracted text, particularly the customer reviews.

3. Pricing of the products:

Research Question: Do the prices of toys in the online store demonstrate a normal distribution, or are there significant fluctuations?

In this part of the analysis, I attempted to visualize the products based on their prices. The Laser Tag Battle Pack, Hydraulic Cyborg Hand, Robot Factory, and Alphabet Farmyard Abacus emerged as the most expensive toys (fig 2). Conversely, the Exercise Dice, Push Popper, and World’s Smallest Microscope were identified as the cheapest

products (fig1).

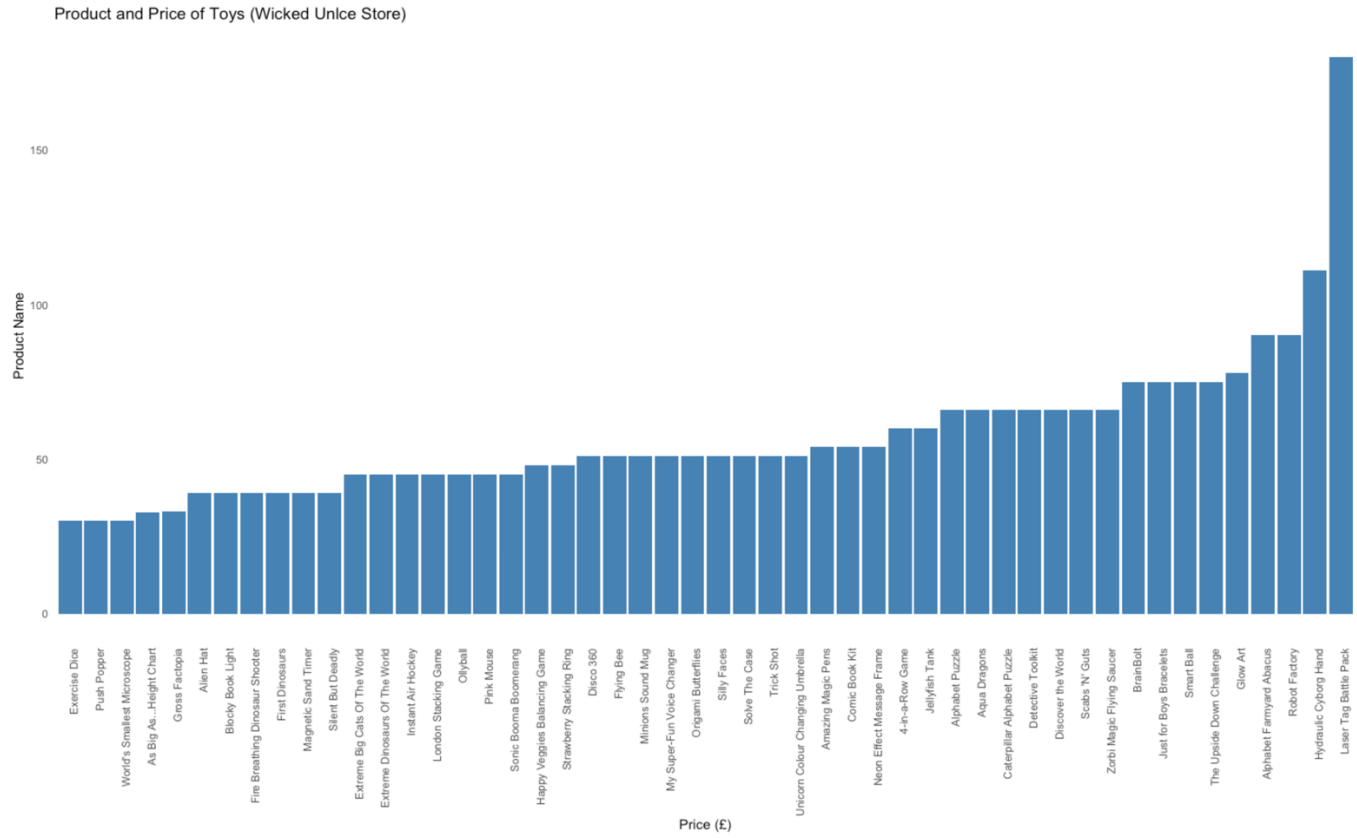

Figure 1 Price of the Various Products

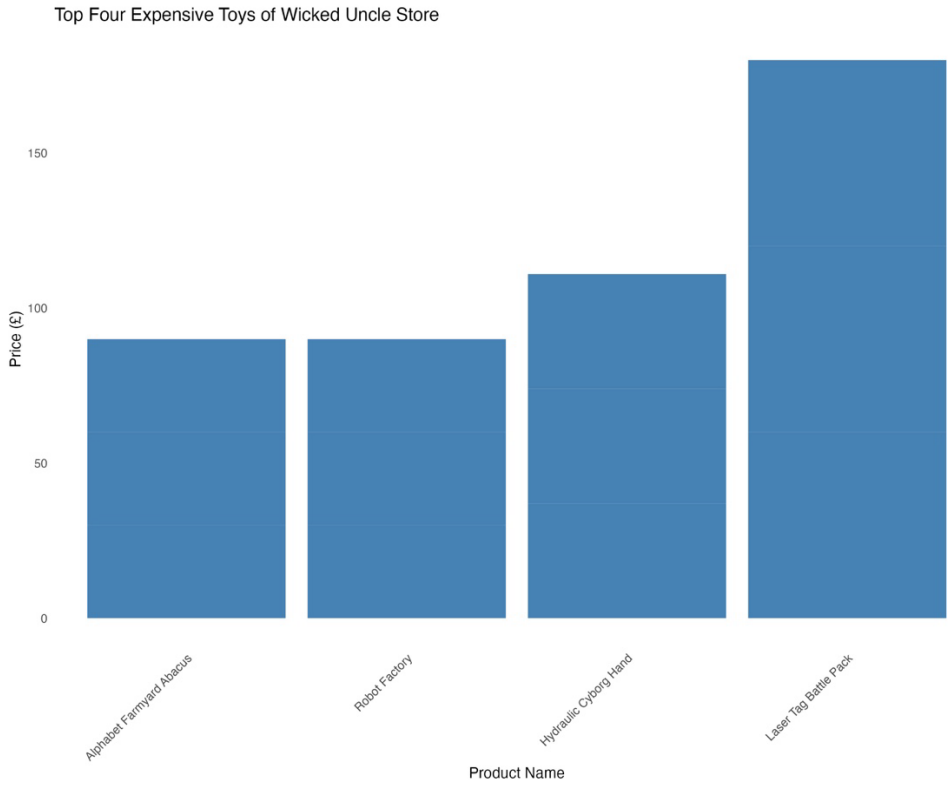

Figure 2 Top Four Expensive Products

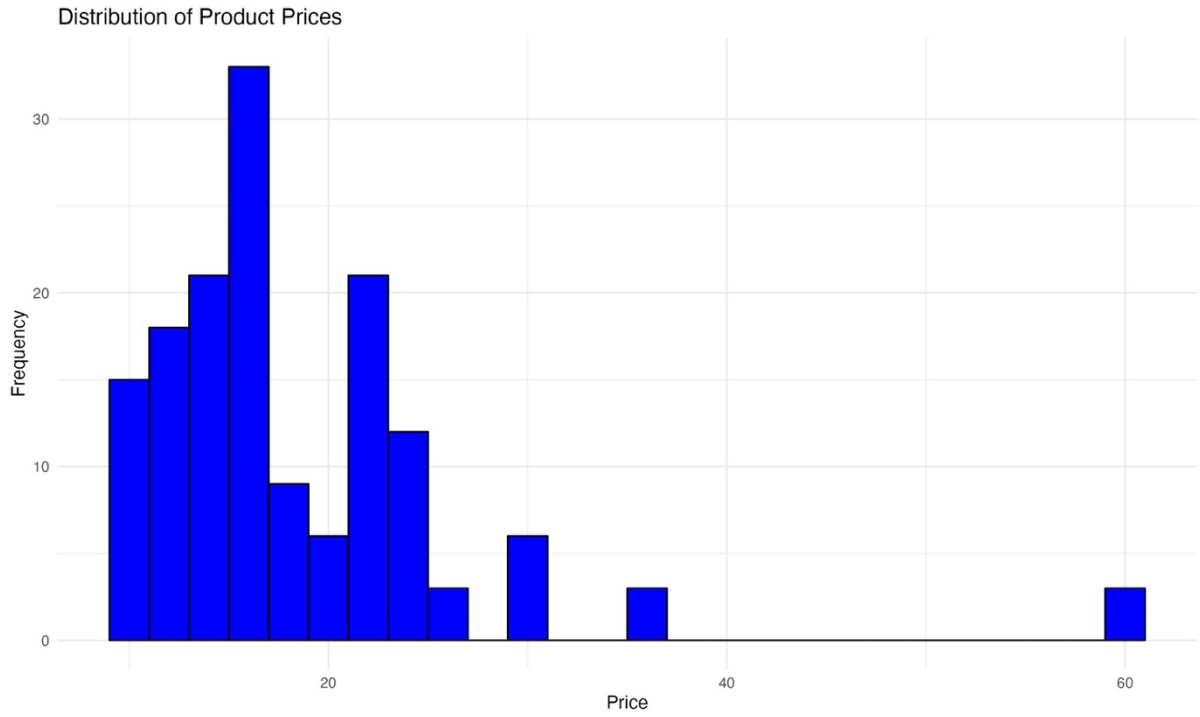

Figure 3 Distribution of Product Prices

The histogram (fig3) displays a multimodal distribution of toy prices, indicating that they do not follow a normal distribution. Prices are grouped distinctly, with most toys priced under £20 and peaks at approximately £10 and £20. Higher price points at £30 and £60 are less frequent, suggesting a strategic targeting of various customer segments by the store.

4. Basic Sentiment Analysis

Research Question: How do customer sentiments expressed in online reviews correlate with the perceived quality and satisfaction of toys sold by the online store?

Sentiment analysis of customer reviews revealed distinct customer satisfaction patterns and improvement areas. This analysis allowed me to understand the emotional tone behind customer interactions, highlighting which aspects of products and services resonated positively or negatively with consumers.

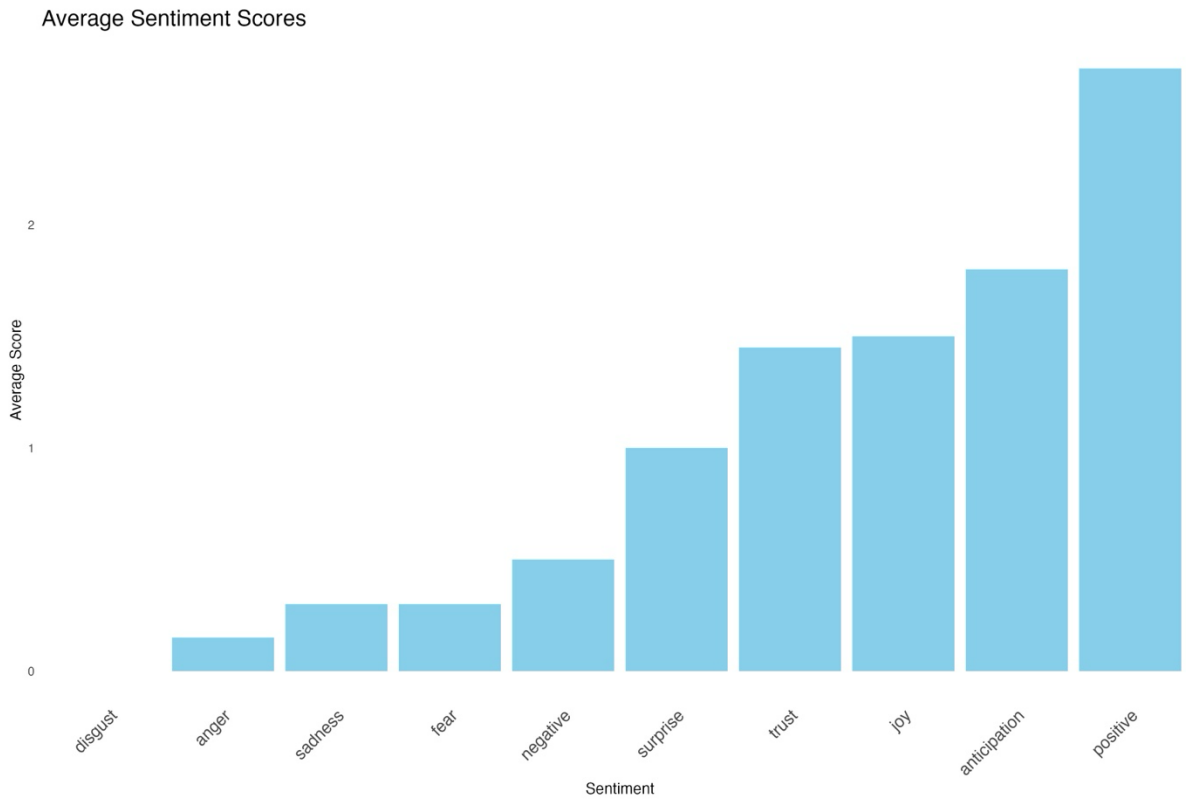

Figure 4 Average Sentiment Score of the Customer Reviews

Figure 5 Wordcloud of the most frequently used words in customer reviews.

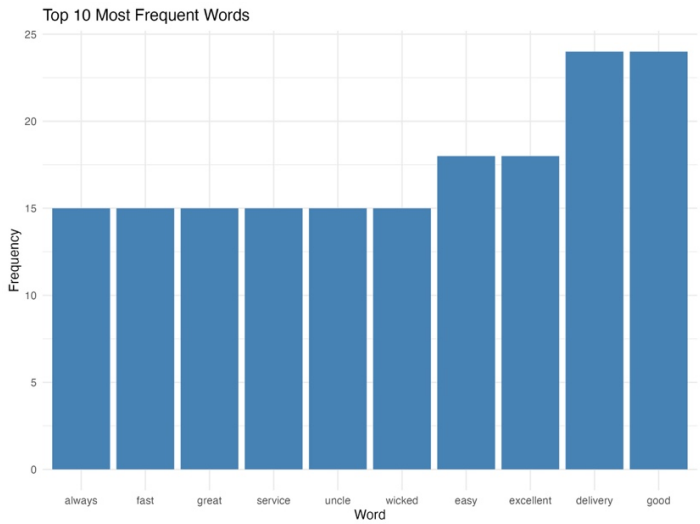

Figure 6 Top Most Frequently Used Words in Customer Review

The "Average Sentiment Scores" (fig 4) chart shows a predominance of positive sentiments like joy, trust, anticipation, and positivity, significantly outnumbering negatives such as sadness and anger. This indicates that most customers view the products from the online toy store favorably, suggesting high satisfaction and perceived quality.

Additionally, the "Top 10 Most Frequent Words" (fig 6) bar chart and word cloud (fig 5) feature positive terms like "great," "excellent," "fast," and "good," commonly used in customer reviews, which underscores approval and satisfaction.

In summary, sentiment analysis and textual review content reveal that customers are generally satisfied with their purchases, positively impacting the store's reputation and customer retention. This trend highlights not only the quality of the products but also the effectiveness of the store's marketing strategies.

5. Dominant Colors

Research Question: What are the dominant colors present in the product images on the online toy store's website?

Colors significantly influence consumer perception and purchasing decisions. By understanding the dominant colors in toy images, we can gain insights into what visually appeals to customers, potentially driving their buying choices.

For this part, I used a dataset extracted using a Python script. The Python script and the extracted dataset will be submitted with the assignment. As mentioned in the

assignment instructions to ‘seek help beyond this course,’ I utilized my experience with Python for this task.

The script provided utilizes Python's Imaging Library (PIL) and scikit-learn’s KMeans clustering algorithm to analyze and determine the dominant colors in a collection of images. Each image is resized to reduce complexity and reshaped into an array of RGB values for processing. The KMeans algorithm, set to identify five clusters, finds the centroid with the highest frequency in each image, representing the dominant color. This RGB value is then converted to its closest named CSS3 color for interpretability. Results from all images are aggregated, and frequencies of each named color are compiled into a DataFrame, which is then saved as a CSV file. This process automates the identification of prevalent colors across images, streamlining data analysis and presentation.

Note: Both of these Python scripts and datasets are attached to the assignment.

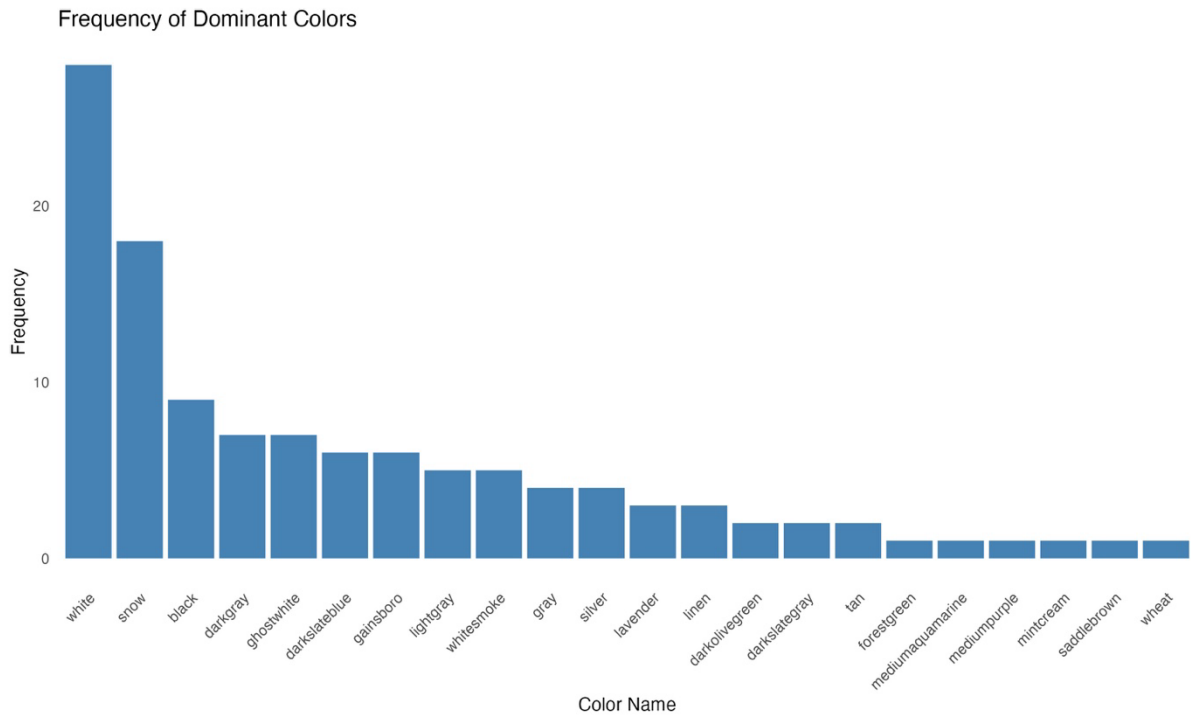

Figure 7 Dominant Colors

The "Frequency of Dominant Colors" chart shows that white, snow, and black are the most prevalent colors in product images from the online toy store, indicating a preference for neutral backgrounds that emphasize the products. Other neutral shades like dark gray and ghost-white are also common, supporting a trend towards using subtle color palettes that ensure toys stand out. This strategic use of color likely aims to enhance visual appeal and influence customer purchasing decisions.

6. Unsuccessful Attempt:

I began scraping data from the most popular online bookstore in Bangladesh. The scraping was successful. However, I encountered an interesting problem: all the information on the website was in the Bengali language and stored in Bengali fonts. Due to my limited technical ability and lack of experience working with non-English data, I couldn't proceed further. This situation presented a new challenge and highlighted the need for me to develop skills to handle multilingual data for future research.

Description: df [10 × 5]

	Title <chr>	Author <chr>	OriginalPrice <chr>	DiscountedPrice <chr>
1	What Do Citizens Think	নজরুল ইসলাম	৩০৬.০০ টাকা	৩৪০.০০ টাকা
2	স্মৃতিপটে বাংলাদেশ	শাহাবুদ্দিন চৌধুরী	১,৫৬০.০০ টাকা	২,০০০.০০ টাকা
3	দেশভাগের গল্প: তামিল, ইংরেজি, পাঞ্জাবী, মারাঠি, কানাড়ি, মালয়ালাম	জাভেদ ইকবাল (অনুবাদক), মোস্তফা আজিজ জয় (অনুবাদক)	৩১২.০০ টাকা	৪০০.০০ টাকা
4	দেশভাগের গল্প: উর্দু	জাভেদ ইকবাল (অনুবাদক), মোস্তফা আজিজ জয় (অনুবাদক)	৩৫৮.৮০ টাকা	৪৬০.০০ টাকা
5	বখতিয়ারের ঘোড়া	রাশেদ রহমান	২১৮.৪০ টাকা	২৮০.০০ টাকা
6	শাহ আবদুল করিম : লোকগানের ভিন্নধর্মী উত্তরাধিকার	মোহাম্মদ শেখ সাদী	৩১২.০০ টাকা	৪০০.০০ টাকা
7	পয়ার ও লাচাড়ি	প্রশান্ত মৃধা	২১৮.৪০ টাকা	২৮০.০০ টাকা
8	একজন পুরুষের স্থান	মোরশেদুর রহমান (অনুবাদক)	২১৮.৪০ টাকা	২৮০.০০ টাকা
9	আমার স্ত্রী আমাকে আর ভালোবাসে না ও অন্যান্য কবিতা	আব্দালিব রাশদী	২৭৩.০০ টাকা	৩৫০.০০ টাকা
10	হাফিজ: শিরাজের বুলবুল	রফিকুল রনি	৪৮৩.৬০ টাকা	৬২০.০০ টাকা

10 rows | 1–5 of 5 columns

Figure 8 A Glimpse of the Extracted Data from My First Scrap

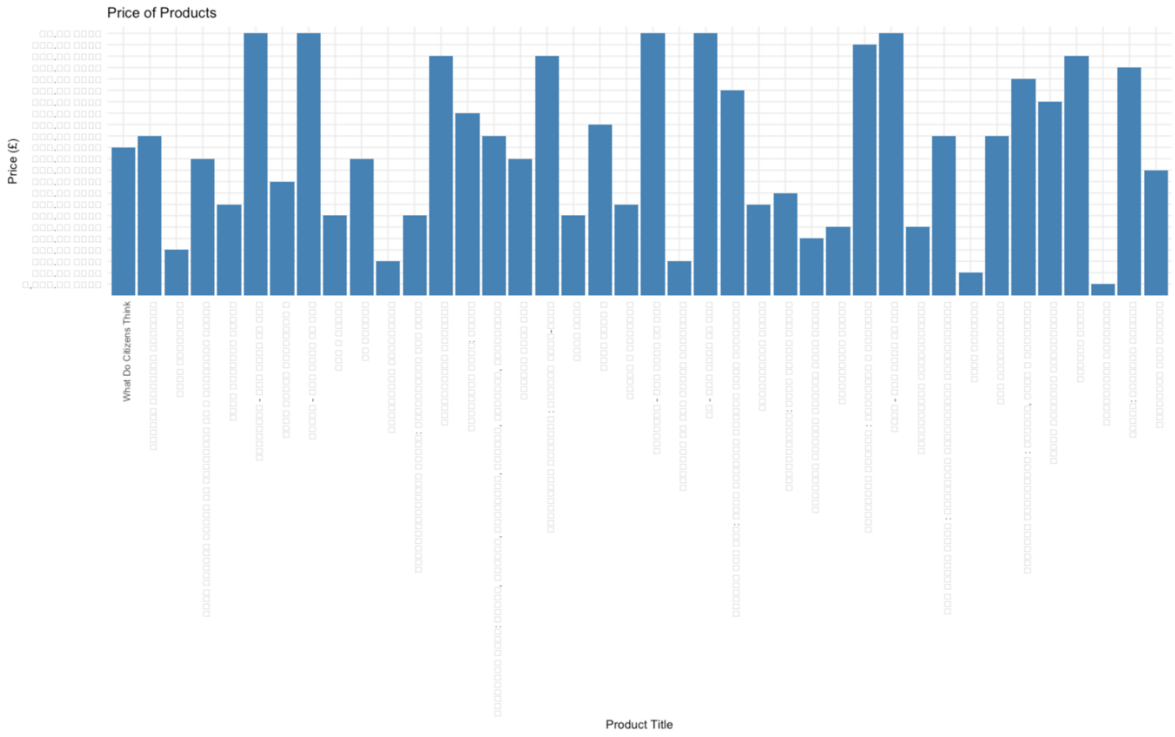

Figure 9 Wierd Plot of my First Attempt

Notably, the product titles appear in English, indicating that Bengali fonts are not displayed. This absence could be due to the plotting environment's limitations in supporting non-Latin scripts such as Bengali.

After this encounter, I stopped working with that website - Prothoma ,and started working with the Wicked Uncle.

7. Legal and Ethical Issues

The GDPR (General Data Protection Regulation) is a data privacy and security law passed by the EU (European Union) and put into effect on May 25, 2018. The main purpose of this regulation is to give EU citizens control over their personally identifiable information by putting limitations on organizations targeting and collecting this data.

The GDPR doesn't state that web scraping is illegal; however, it restricts what businesses can do with the contact data they wish to extract. For example, in some cases, to gather personal data and use it for various purposes, they have to receive explicit consent from the data subjects (Fatenite, 2024).

Based on this, I may have breached some ethical aspects. Of course, this is business data, which could be very sensitive for them. However, I was considerate towards the website and did not make any changes; I simply used HTML elements to extract the necessary information.

8. Limitations

The way the colors are extracted and counted are not the actual colors bur somehow close to those colors as it takes the nearest RGB value and and then define it as that color. So a color which is not genuinely for example white might be considered as white. This is sort of a limitation. When it comes to NLP algorithms, especially it is widely arguable issue how acurate these algorithms are. For example the algorithm behind the sentiment detection does not understand sarcasm. So, a sarcastic negative comment can be shown as a positive comment in this analysis which is an error. More careful analysis and interpretation are needed in these scenarios.

9. Conclusion:

Throughout this project, I successfully employed web scraping techniques to gather valuable data from the Wicked Uncle toy store website. Using both R and Python, I extracted product details, customer reviews, and image data, which allowed me to conduct further analyses. I analyzed dominant colors in product images to understand visual marketing strategies, and sentiment analysis of customer reviews revealed a predominantly positive customer perception. Additionally, I also tried to focus on the pricing ranges of various toys on this website. This project not only improved my technical skills but also taught me how to conduct various analyses on the extracted data.

10. Reference

[1] Web Scraping Web Scraping R for Data Science

[2] Web Scraping in R Stats and R

[3] Web Scraping in R: The Complete Guide 2024 Link

[4] Web Scraping With R ScrapingBee

[5] For the Python Part of this projec, my own code from a course named “Python and R programming” I just have cpmpleted with University of Dalarna.

[6] Gabija Fatenaite. 2024. Is Web Scraping Legal? retrived from