## Part 1. Project Objective

The primary goal is to get actual hair care products information and customer reviews and analyze hair care best sellers from iHerb.

Project scope outline:

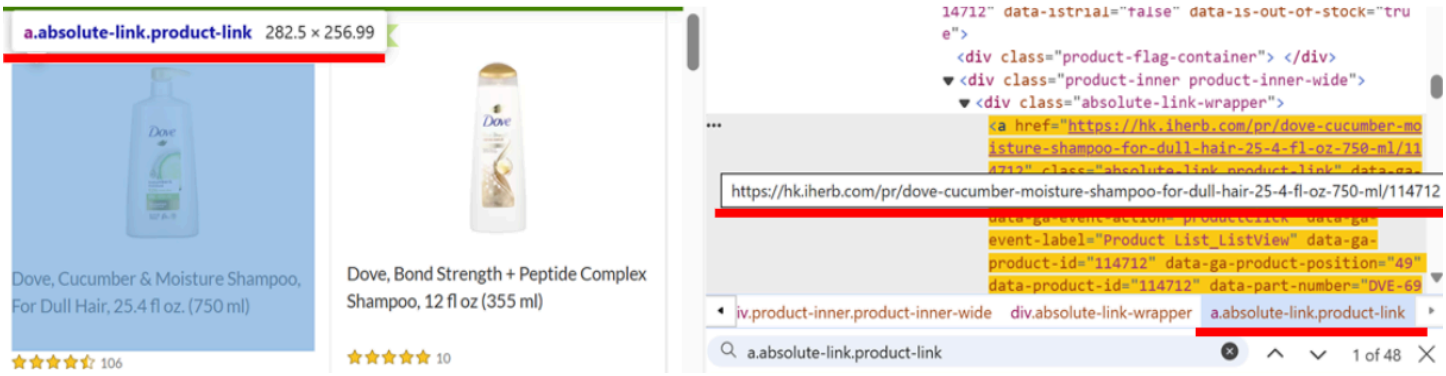
- Web Scraping (Product information and customer reviews)
- Data preprocessing
- Exploratory Data Analysis with Visualization (Dashboard)
- Sentiment Analysis with NLP

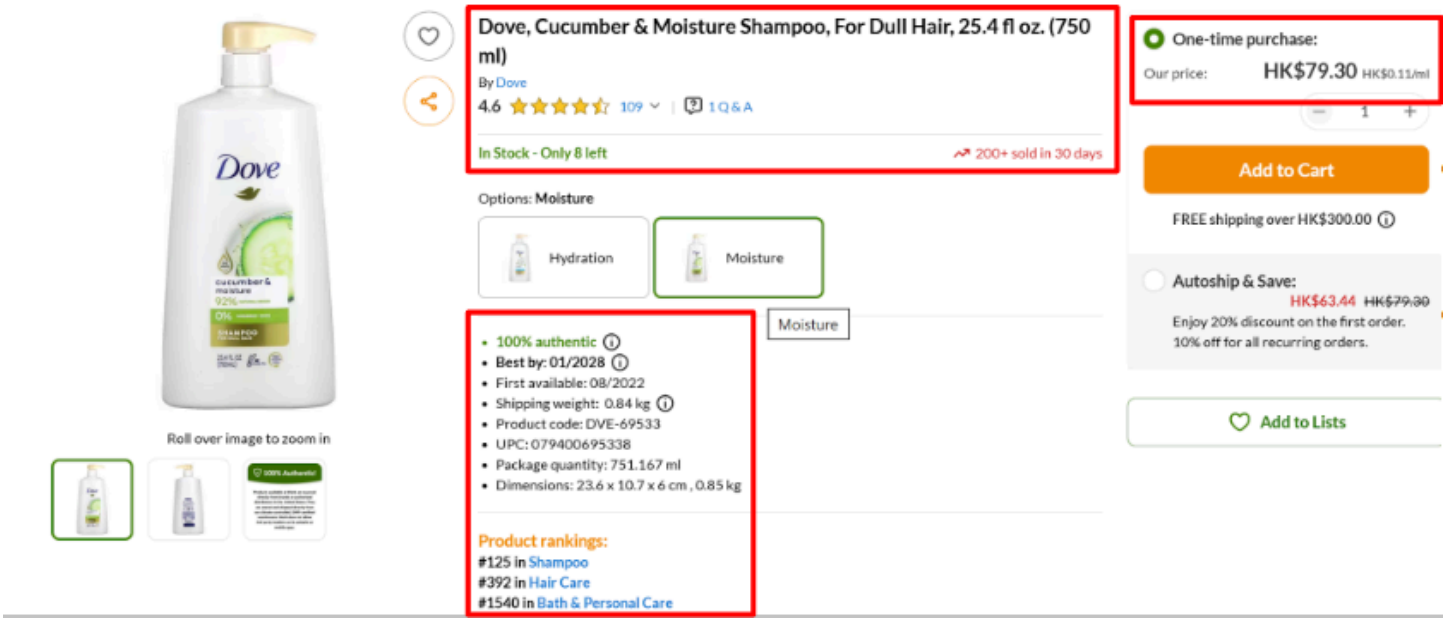
## Part 2. Web Scraping and Data Preprocessing of Product Information

### A. Scraping of Product Information

### (1) Methodology

- The dataset that we scraped from [Hair Care • Natural Hair Care Products | iHerb]
- The main web scraping tools we use are Selenium and BeautifulSoup.
- First, we scraped the product links for 630 targeted products and saved them in a CSV file (Product Link CSV). Since iHerb product pages detect bots using XParameter, we developed a semi-manual script and used VPN as a workaround.
- We defined the main function to scrape data from one product page at a time. Once the data was scraped, it was saved in another CSV file (Product Information CSV), and the corresponding product link was removed from the Product Link CSV. If data could not be scraped, the link remained in the Product Link CSV for future attempts. To scrape data from additional product pages, we changed the IP using the VPN and recalled the scraping function until we successfully collected data for all 630-product links.



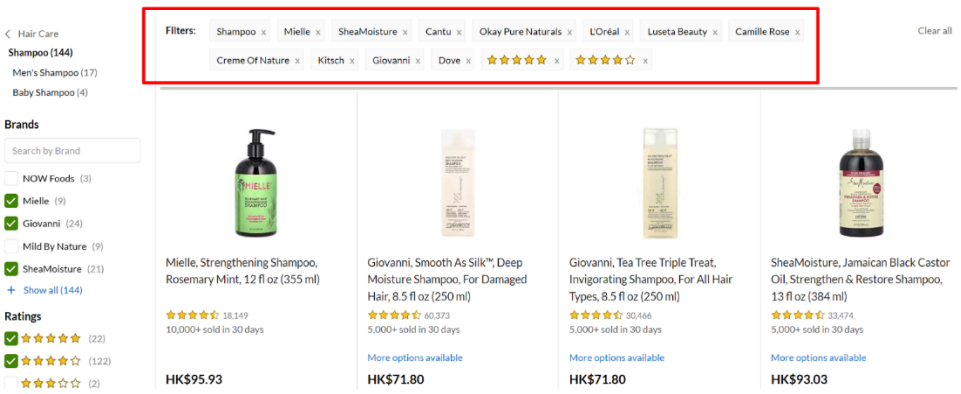



**(2) Other methods results**

| No. | Web scrape methods we have tried | | Results |
|---|---|---|---|
| 1 | Frameworks and Libraries | Scrapy | Fail |
| | | Playwright | Fail |
| | | BeautifulSoup | Success |
| 2 | Proxy and IP Management | Proxy Usage | Fail |
| | | Switch IP Addresses in Turn | Success |
| 3 | Browser Automation | Headless Browser (Bright Data Scraping Browser) | Fail |
| | | Selenium | Success |
| 4 | Data Extraction Techniques | Scrapfly | Fail |
| 5 | Request Management | Set User-Agent | Fail |
| | | Set Request Headers | Fail |
| | | Set Random Delay Time | Fail |
| 6 | No-Code Tools | Octoparse | Fail |
| | | Instant Data Scraper | Only got information from 4 columns |

**(3) Dataset Selection Criteria**

- For product information, we selected 630 products from 4 main categories and 11 well-known brands, each with over 40 hair care products available, all with product ratings of 4 or 5 out of 5.




| Category | Number of products | Percentage |
|---|---|---|
| Conditioner | 196 | 31% |
| Shampoo | 149 | 24% |
| Treatments | 144 | 23% |
| Styling | 141 | 22% |

**B. Data Preprocessing of Product Information**

**(1) Dataset Overview**

- The initial data is 630 product records with 12 columns.
  Out of 630 records, we identified 608 unique products. The remaining 22 instances are duplicates, which exist because a product can fall into multiple categories. Given our interest in understanding product distribution across categories, we have chosen to retain these duplicates in our dataset.

**(2) Data Preprocessing**

- We can see that only 40% (252/630) of the products have data for sale_in_30_days, with values ranging from 100 to 40,000. Therefore, we choose to interpret that 60% (378/630) of the scrapped products did not have any sales in 30 days and we filled those none records as 0.

- Also, we filled null values in "stock_alert" with the string "in stock".
- By checking the product weight or volume, we found that there are 2 Scalp Exfoliator in the dataset. And we chose to drop them from the dataset (628 records left).
- Then, we standard the capacity and created a new column: volume_in_ml
- Finally, we created 3 more new columns as below.
  - Price per ml (discount_price / volume_in_ml)
  - Discount % ((list_price - discount_price) / list_price)*100)
  - Sales revenue (discount_price*sale_in_30days)

**(3) Data Schema for** iherb_hair_care_clean_dataset (Red represents the newly added columns)

| No. | Column | Description |
|---|---|---|
| 1 | product_id | The product's unique ID |
| 2 | product_name | Product name |
| 3 | brand_name | Product brand name |
| 4 | category | Product category |
| 5 | volume_in_ml | The total volume of a product measured in milliliters |
| 6 | list_price | Market value before any discounts or promotions are applied |
| 7 | discount_price | Market value after any discounts or promotions are applied |
| 8 | price_per_ml | the cost of a product measured per milliliter |
| 9 | discount% | The percentage reduction applied to the original list price of a product |
| 10 | sale_in_30days | The total number of units sold for a product within the last 30 days |
| 11 | sales_revenue | The total income from sales is calculated by multiplying the price of each product by the number of products sold over a 30-day period (Sales Revenue = product price*sale in 30 days) |
| 12 | rating | Total rating of product |
| 13 | no_of_reviews | The total number of customer reviews submitted for a product |
| 14 | first_available | When a product was first made available for sale |
| 15 | stock_alert | The availability status of a product |
| 16 | ranking_shampoo | The position of a shampoo product within its category |
| 17 | ranking_conditioner | The position of a conditioner product within its category |
| 18 | ranking_hair_treatments | The position of a treatment product within its category |
| 19 | ranking_hair_styling | The position of a hair styling product within its category |

## Part 3. Exploratory Data Analysis with Visualization (Dashboard)

**(1) Insights on Data Exploration**

- 5 brands that account for a relatively large number of datasets, indicating their prominence in the market. They are Cantu, Giovanni, SheaMoisture, Dove and Mielle.
- The stock alert situation indicates that nearly 90% of the items are currently in stock, while 9% are out of stock. Additionally, 1% of the items are unavailable in Hong Kong. This distribution highlights the overall availability of products and identifies

potential supply issues that may need to be addressed.

- The volume of the products is measured in milliliters, with a mean volume of 294 ml. The maximum volume recorded is 750 ml, which is approximately 2.5 times higher than the average. This information provides insights into the range of product sizes available and highlights the variability in volume among the items.
- The price per milliliter is analyzed, revealing a mean price of $0.42. The maximum price recorded is $3.14, which is approximately 7 times higher than the average. This data illustrates the variation in pricing and may indicate differences in product quality or brand positioning within the market.
- There is a notable gap in the first available date, ranging from 2007 to 2024. This shows that our dataset includes both long-established and newer products.
- As we filtered products with ratings between 4 and 5, the minimum rating in our dataset is 3.8 and the mean is 4.56.

**(2) Heatmap**

| | brand_name | category | discount% | discount_price | no_of_reviews | price_per_ml | rating | sale_in_30days | stock_alert | volume_in_ |
|---|---|---|---|---|---|---|---|---|---|---|
| brand_name | 1.000 | 0.235 | 0.185 | 0.283 | 0.109 | 0.301 | 0.132 | 0.000 | 0.193 | 0.356 |
| category | 0.235 | 1.000 | 0.154 | 0.145 | 0.123 | 0.288 | 0.142 | 0.018 | 0.092 | 0.297 |
| discount% | 0.185 | 0.154 | 1.000 | -0.213 | -0.303 | -0.109 | 0.018 | -0.238 | 0.194 | -0.004 |
| discount_price | 0.283 | 0.145 | -0.213 | 1.000 | -0.039 | 0.394 | 0.061 | -0.024 | 0.086 | 0.159 |
| no_of_reviews | 0.109 | 0.123 | -0.303 | -0.039 | 1.000 | -0.040 | -0.063 | 0.721 | 0.000 | 0.001 |
| price_per_ml | 0.301 | 0.288 | -0.109 | 0.394 | -0.040 | 1.000 | -0.094 | -0.006 | 0.040 | -0.796 |
| rating | 0.132 | 0.142 | 0.018 | 0.061 | -0.063 | -0.094 | 1.000 | -0.037 | 0.179 | 0.119 |
| sale_in_30days | 0.000 | 0.018 | -0.238 | -0.024 | 0.721 | -0.006 | -0.037 | 1.000 | 0.000 | -0.035 |
| stock_alert | 0.193 | 0.092 | 0.194 | 0.086 | 0.000 | 0.040 | 0.179 | 0.000 | 1.000 | 0.057 |

| | brand_name | category | discount% | discount_price | no_of_reviews | price_per_ml | rating | sale_in_30days | stock_alert | volume_in_ |
|---|---|---|---|---|---|---|---|---|---|---|
| brand_name | 1.000 | 0.235 | 0.185 | 0.283 | 0.109 | 0.301 | 0.132 | 0.000 | 0.193 | 0.356 |
| category | 0.235 | 1.000 | 0.154 | 0.145 | 0.123 | 0.288 | 0.142 | 0.018 | 0.092 | 0.297 |
| discount% | 0.185 | 0.154 | 1.000 | -0.213 | -0.303 | -0.109 | 0.018 | -0.238 | 0.194 | -0.004 |
| discount_price | 0.283 | 0.145 | -0.213 | 1.000 | -0.039 | 0.394 | 0.061 | -0.024 | 0.086 | 0.159 |
| no_of_reviews | 0.109 | 0.123 | -0.303 | -0.039 | 1.000 | -0.040 | -0.063 | 0.721 | 0.000 | 0.001 |
| price_per_ml | 0.301 | 0.288 | -0.109 | 0.394 | -0.040 | 1.000 | -0.094 | -0.006 | 0.040 | -0.796 |
| rating | 0.132 | 0.142 | 0.018 | 0.061 | -0.063 | -0.094 | 1.000 | -0.037 | 0.179 | 0.119 |
| sale_in_30days | 0.000 | 0.018 | -0.238 | -0.024 | 0.721 | -0.006 | -0.037 | 1.000 | 0.000 | -0.035 |
| stock_alert | 0.193 | 0.092 | 0.194 | 0.086 | 0.000 | 0.040 | 0.179 | 0.000 | 1.000 | 0.057 |

**Insights:**

1. **No. of reviews** is inversely proportional to **discount%** (-0.303)

   The number of reviews is inversely proportional to the discount percentage, with a correlation of -0.303. This means products with fewer reviews often have greater discounts, it suppose iHerb try to promote those unpopular products by offering discounts.

2. **Price per ml** is influenced by **brand** (0.301)

   Indicating that certain brands are particularly more expensive.

3. **Discount price** (product prices) is proportional to **price per ml** (0.394)

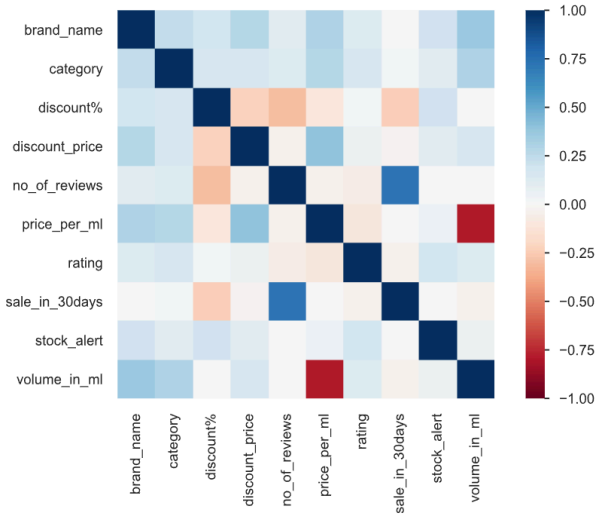
   Higher product prices typically lead to higher unit prices.

4. **Volume in ml** is **highly correlated** with **price per ml (-0.796)**

   Product volume increases, the price per milliliter tends to decrease.

5. **No. of reviews** is **highly correlated** with **sale in 30 days (0.721)**

   As the number of reviews increases, sales tend to increase as well.




**(3) New metrics created**

To conduct a more holistic analysis, we created additional metrics in the dataset.

- Product Volume in mL:

  We found that some products were missing volume information but had weight data, while some had both volume and weight. To address this, we calculated the average ratio of weight to volume by product category in order to convert weight into volume.

- Sales Revenue = product price*sale in 30 days

  Calculated as the product price after discount multiplied by the quantity sold

- Popularity = rating*no. of reviews / first available date of the products to 2024/9/30

  This score is calculated using the average user ratings, the number of reviews, and the time since availability. This helps us achieve a balanced view of product appeal, avoiding bias while considering trendiness.

  Components:

  (1) Rating: Reflects customer satisfaction and product quality. Higher ratings generally indicate better product performance.

  (2) Number of Reviews: A larger number of reviews can signify greater customer engagement and trust in the product.

  (3) Time Factor: Dividing by the day count from the first available date to September 2024 normalizes the score over time, allowing us to compare products regardless of how long they've been available.

**(4) Additional column splitting of product name**

We found that the product name column contained various pieces of information, such as brand name, volume or weight, flavor, and product type. To better organize this information, we decided to split the product names using a comma as the delimiter. The split information has been utilized in the dashboard.

**(5) Dashboard design, visual selection & methodology**

We designed a three-page dashboard using Power BI. Here are the key questions we aimed to address with the dashboard:

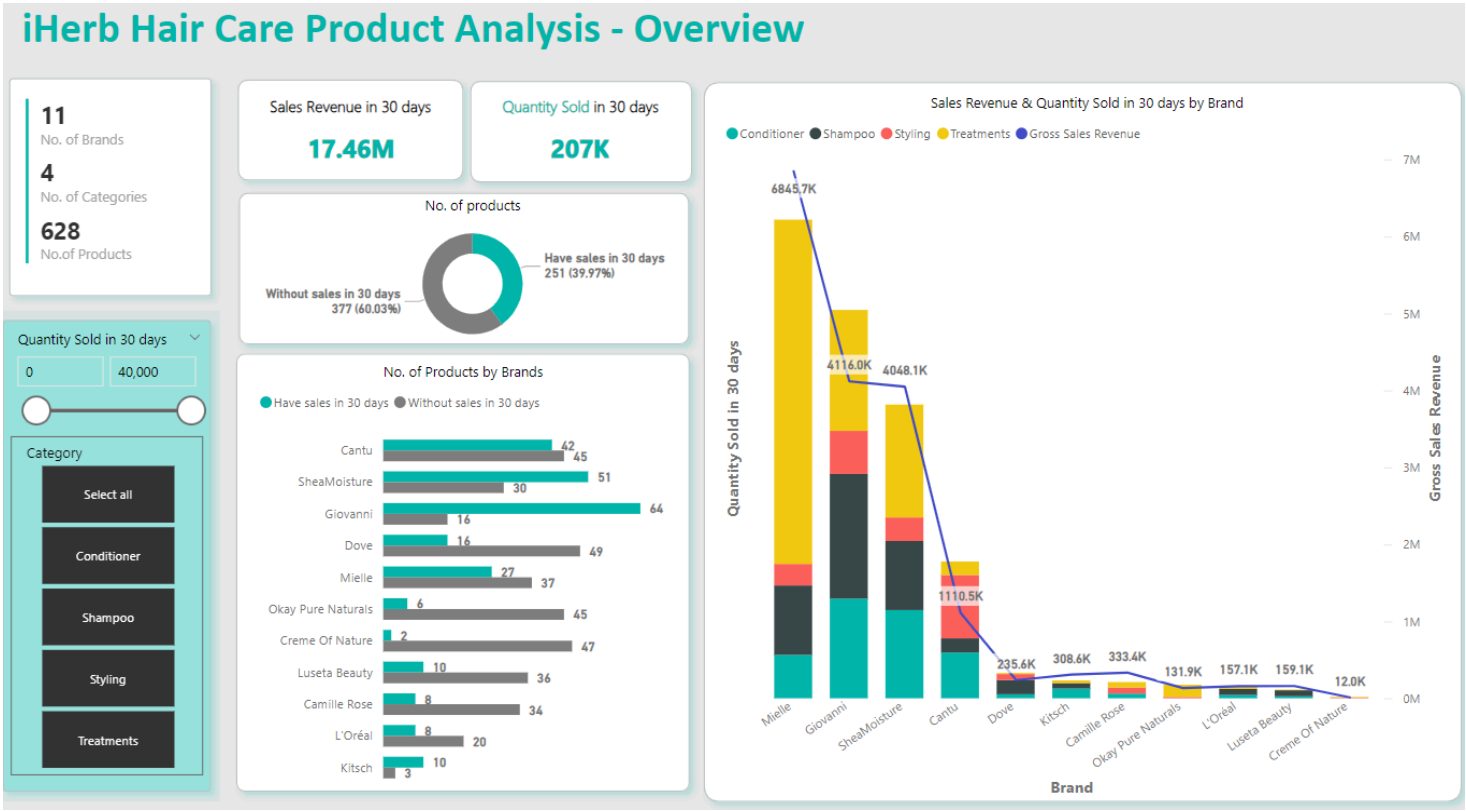
1. **Low Product Visibility**
   Many products may not be receiving adequate exposure, particularly those with low sales figures.

2. **Price Sensitivity / Elasticity**
   We sought to understand how price impacts sales, especially for products with varying ratings.

3. **Popularity vs. Sales Correlation**
   Hair care products are typical fast-moving consumer goods. The characteristics of low price elasticity, low differentiation, and low switching costs make “branding” and “reputation” crucial for companies.

## 5.1 Overview Page

The first page serves as an **Overview**, featuring filters for quantity sold and product categories to help users navigate the data effectively.




**Key Performance Indicators (KPIs)**:
At the top of the page, we display essential KPIs using KPI cards which can provide a concise overview of critical metrics:

- **Number of Brands**, **Number of Product Categories**, **Total Products**, **Sales Revenue**, **Quantity Sold in Last 30 Days**

**Product Availability Distribution**:
We utilize a **donut chart** to visually represent the distribution of products based on their sales performance in the last 30 days. The chart distinguishes between products with sales (green) and those without sales (grey).

- **Key Insight**: Only 40% of products had sales in the last 30 days, indicating potential issues with product visibility and marketing strategies.

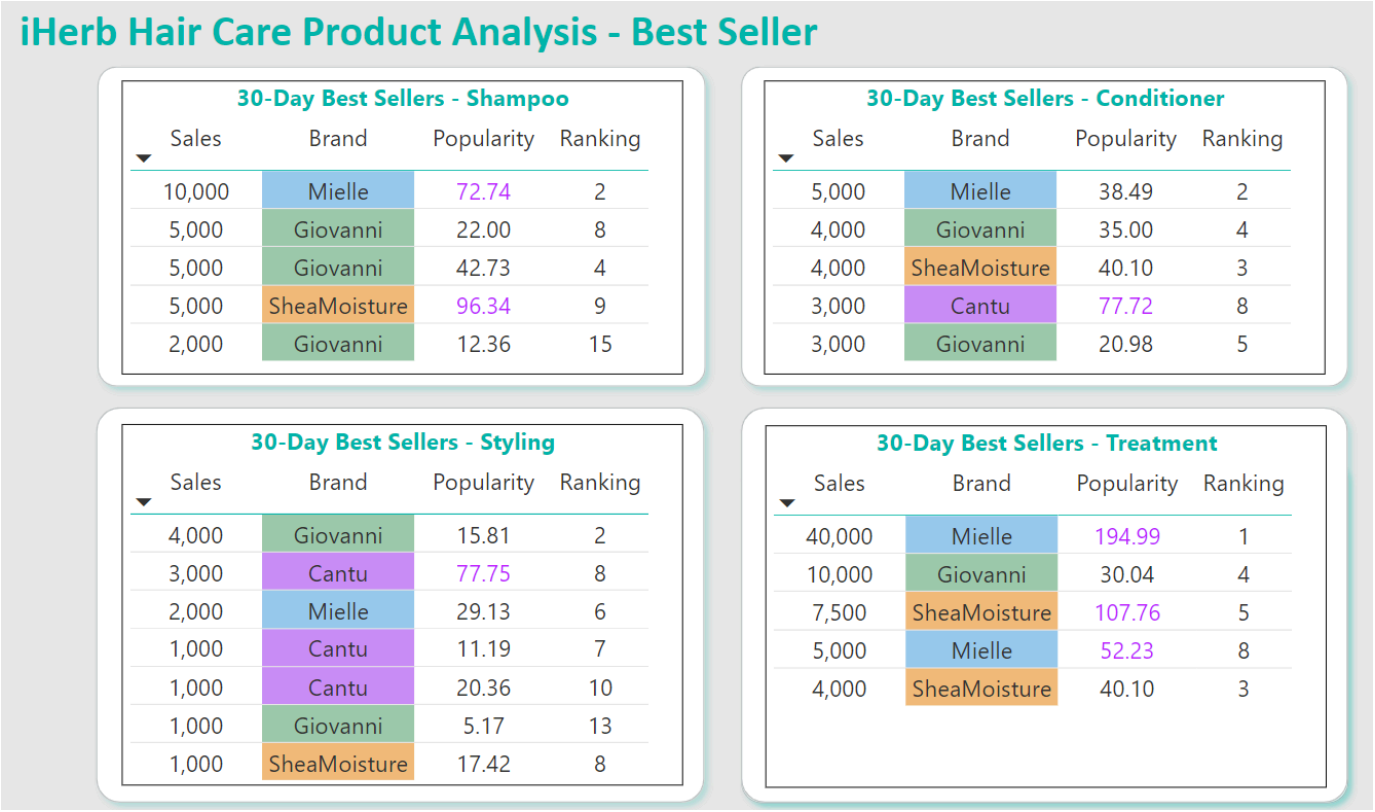
**Sales Revenue and Quantity Sold by Brand**:
On the right side of the Overview page, we present a combined chart of stacked columns and a line graph. The stacked columns represent the quantity sold by brand, while the line indicates total sales revenue.

The legend categorizes products by color (yellow, red, black, green), making it easy for viewers to differentiate between product categories.

- **Key Insight**:
  From a brand perspective, Mielle, Giovanni, SheaMoisture, Cantu, and Dove are contributed the largest sales. This information can guide marketing efforts and inventory decisions.
  From a category perspective, treatment products lead in sales, followed closely by shampoo and conditioner. Styling products, however, show the least sales performance. This insight can help prioritize product development and promotional strategies.

## 5.2 Best Sellers by Category

On the second page of the dashboard, we focus on the **Best Sellers by Category**. We define best sellers as the top five products with the highest sales over the past 30 days, and we present this information in four separate tables, each representing a product category.



iHerb Hair Care Product Analysis - Best Seller

30-Day Best Sellers - Shampoo

| Sales | Brand | Popularity | Ranking |
|---|---|---|---|
| 10,000 | Mielle | 72.74 | 2 |
| 5,000 | Giovanni | 22.00 | 8 |
| 5,000 | Giovanni | 42.73 | 4 |
| 5,000 | SheaMoisture | 96.34 | 9 |
| 2,000 | Giovanni | 12.36 | 15 |

30-Day Best Sellers - Conditioner

| Sales | Brand | Popularity | Ranking |
|---|---|---|---|
| 5,000 | Mielle | 38.49 | 2 |
| 4,000 | Giovanni | 35.00 | 4 |
| 4,000 | SheaMoisture | 40.10 | 3 |
| 3,000 | Cantu | 77.72 | 8 |
| 3,000 | Giovanni | 20.98 | 5 |

30-Day Best Sellers - Styling

| Sales | Brand | Popularity | Ranking |
|---|---|---|---|
| 4,000 | Giovanni | 15.81 | 2 |
| 3,000 | Cantu | 77.75 | 8 |
| 2,000 | Mielle | 29.13 | 6 |
| 1,000 | Cantu | 11.19 | 7 |
| 1,000 | Cantu | 20.36 | 10 |
| 1,000 | Giovanni | 5.17 | 13 |
| 1,000 | SheaMoisture | 17.42 | 8 |

30-Day Best Sellers - Treatment

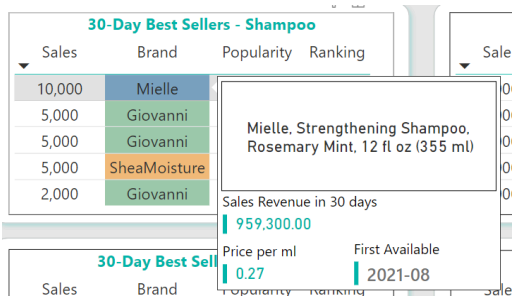
| Sales | Brand | Popularity | Ranking |
|---|---|---|---|
| 40,000 | Mielle | 194.99 | 1 |
| 10,000 | Giovanni | 30.04 | 4 |
| 7,500 | SheaMoisture | 107.76 | 5 |
| 5,000 | Mielle | 52.23 | 8 |
| 4,000 | SheaMoisture | 40.10 | 3 |

In each table, we display 4 key metrics: **Sales**, **Brand**, **Popularity** and **Ranking (in Category)**.

To enhance visual clarity, we use distinct colors for different brands throughout the tables. For example, Giovanni is represented in green and has three products listed among the best sellers in the shampoo category. Mielle, depicted in blue, stands out by occupying the top position in three of the four categories, indicating its strong market presence.

For the **Popularity** metric, we observed significant variation among the best-selling products. To highlight this, we color-coded popularity scores above 50 in purple, allowing users to easily identify products with high popularity at a glance.

Since we wanted to ensure the tables remained uncluttered, we opted not to display all product details directly. Instead, we incorporated tooltips that provide additional information when users hover over specific entries. These tooltips include crucial details such as product name, sales revenue, price per milliliter, and first availability, offering supplementary context to users.

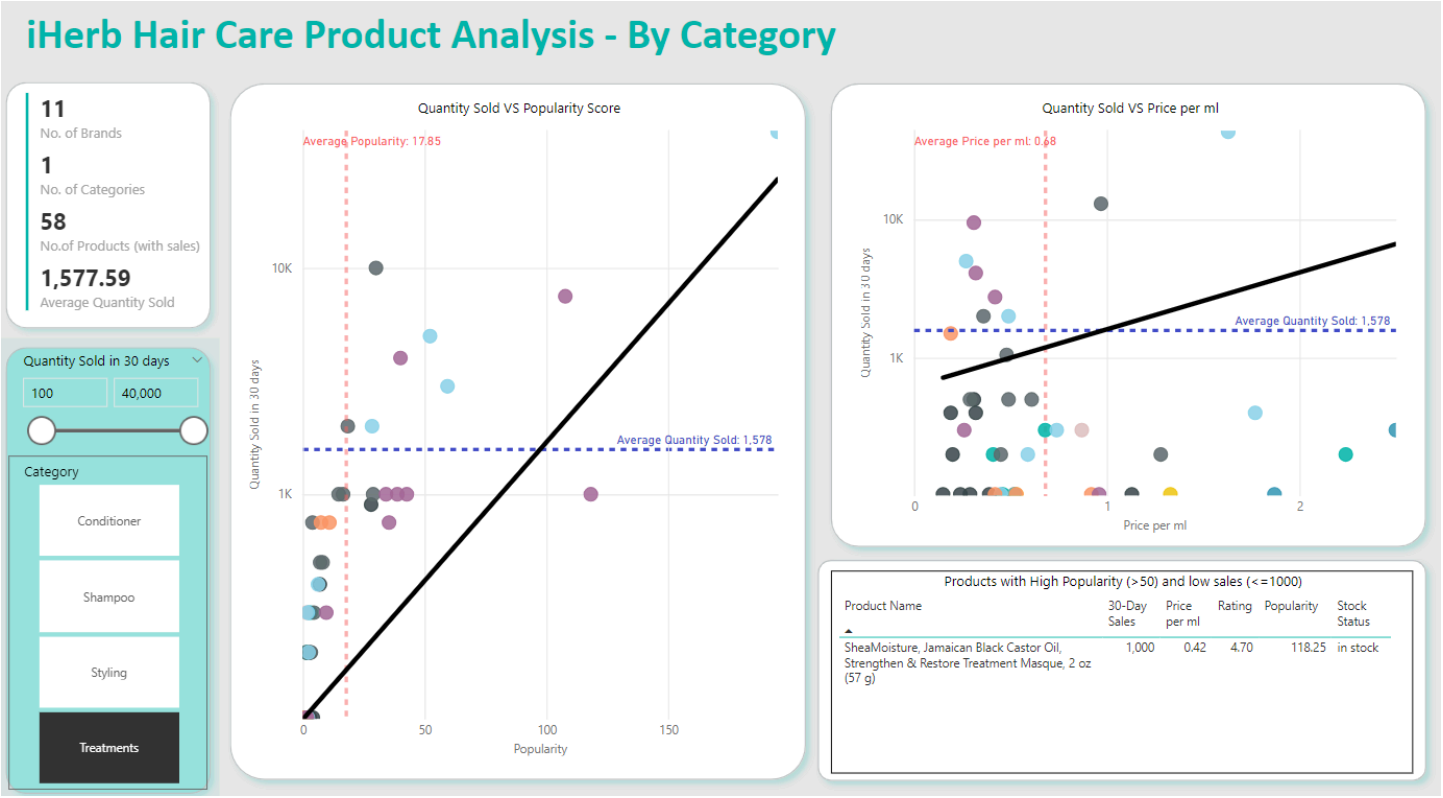



**Insights:** From our analysis, we can observe that most of the best-selling products rank highly within their respective categories. Notably, Mielle has achieved the best sales in the past 30 days, with three of its products ranking first across four product categories. Additionally, brands like SheaMoisture and Giovanni also feature prominently, with products listed in all four categories on the best-selling list. This presents an excellent opportunity for bundling these products together, potentially enhancing sales performance through effective cross-selling strategies. Lastly, Cantu demonstrates relatively strong sales specifically in the hair styling category, indicating a niche market presence.

Overall, these insights not only highlight the top-performing brands but also suggest actionable strategies for product bundling to optimize sales further.

**5.3 Category Analysis and Potential Star Products**

The final page features a product category filter, along with two scatter plots that visualize the relationships between quantity sold and popularity scores, as well as price per milliliter. Additionally, there is a table showcasing potential star products.




**Scatter Plots**:

The scatter plots visualize the relationship between quantity sold and popularity scores on one plot, and quantity sold versus price per ML on the other.

- The **y-axis** represents the quantity sold, and we've included a **blue horizontal reference line** indicating the average quantity sold across all products. This line facilitates quick comparisons to see which products exceed average performance.
- The **x-axis** for the first scatter plot shows popularity, while the second plot displays price per ML. Each plot has a **red vertical reference line** representing the average value for popularity or price per ML, clearly labeled with data values. These reference lines help identify products that stand out in terms of performance.
- **Trend Lines**:
  A **black trend line** is overlaid on each scatter plot, providing a visual representation of the correlation between the two variables. This trend line helps users quickly assess whether an increase in popularity or price is associated with higher sales.

**Insights from Scatter Plots**:

With the category filter applied, we observe that the trend line for quantity sold against popularity scores trends upward from left to right. This correlation is intuitive, as higher reviews and better ratings typically drive increased sales. In the second scatter plot, the trend line for quantity sold against price per ML reveals greater variation among categories. For shampoo and conditioner, the slope is relatively flat, suggesting that consumers are less price-sensitive in these categories. This insight could inform pricing strategies, indicating that brands may have more flexibility in their pricing. Conversely, the trend line for styling products shows a negative correlation, indicating that as the price per ML increases, the quantity sold tends to decrease. This suggests that consumers are more sensitive to price changes in this category, and higher prices may deter purchases. Interestingly, treatment products display a positive trend line, indicating that quantity sold is positively associated with price per ML. This unexpected finding suggests that consumers may view higher-priced treatment products as being of higher quality or effectiveness, and the average price per ML in this category is notably double that of others.

**Table of Potential Star Products**:

The accompanying table highlights products that have high popularity scores but relatively low sales figures. This design element serves as a call to action for marketing teams to promote these products, as they have untapped potential to drive additional sales. By focusing on these products, brands can capitalize on existing consumer interest while addressing sales gaps.

## Part 4. Sentiment Analysis on Customer Reviews

### A. Dataset Description

The review dataset comprises 1,000 customer reviews, focusing exclusively on the top-selling product from each of the four categories—Shampoo, Conditioner, Treatments, and Styling—over a 30-day period across 11 well-known brands. The dataset is curated to reflect a balanced selection of 250 reviews per best-selling product, aligning with the actual distribution of customer ratings. This approach ensures a representative and accurate portrayal of consumer sentiment. Each product is detailed in three columns: 'product_id' for product identification, 'rating' for the star rating assigned by the reviewer, and 'review_text' for the feedback provided by the reviewer.

The process of scraping customer reviews involves constructing URLs to access the review pages. A CSS selector is then used to locate all review containers on each page. When a review is truncated, the "Show more" button is identified, clicked, and the system waits for the full text to become visible. The complete review text is extracted from the corresponding HTML element and added to a collection list. This method efficiently collects a comprehensive dataset of customer feedback across multiple pages.


### B. Text Preprocessing

A series of steps were undertaken to prepare the customer reviews for analysis, including:

- Emoji Conversion: Emojis, which are prevalent in social media communications, were converted to their official textual descriptions, ensuring that the emotional content conveyed by emojis was preserved in a readable format.
- Lowercasing: All text was normalized to lowercase to maintain uniformity and facilitate easier text processing, reducing the complexity of the dataset by eliminating case sensitivity issues.
- Special Character Removal: Special characters like ,.:;@#^&_+-*/<>(){}®=…, which may interfere with text analysis, were removed, with the exception of emotive punctuations such as exclamation marks and question marks.
- Stopword Removal: Stopwords, which are common words that do not add much value to the meaning of the data like a, an, the, etc, were removed using the Natural Language Toolkit (NLTK) library, eliminating frequently occurring words that do not provide substantial information for sentiment analysis.
- Spelling Correction: Spelling errors were corrected using the TextBlob library, improving the overall quality of the text data and ensuring that misspelled words did not skew the analysis results.
- Lemmatization: Lemmatization was applied using the TextBlob library to transform each word into its fundamental form, thereby standardizing lexical variations and diminishing the complexity of the textual dataset.

### C. Methodology

#### (1) Word clouds by rating

Remarkably, the Shampoo, Conditioner, and Treatments categories are dominated by a single, cohesive product line from Mielle:

- Strengthening Shampoo, Rosemary Mint (product_id: 108764)
- Strengthening Conditioner, Rosemary Mint Blend (product_id: 124721)
- Scalp & Hair Strengthening Oil, Rosemary Mint (product_id: 110487)

On the other hand, the Styling category's best-seller, L.A. Hold Styling Gel—Strong Hold (product_id: 6418), is from a different brand, Giovanni.
We performed a collective analysis on the best-selling products from the Shampoo, Conditioner, and Treatments categories, which all belong to the same brand and product line. The Styling product was analyzed separately. To visualize consumer sentiments, word clouds were created for these two groups, with each group containing three distinct word clouds corresponding to different rating scales: (i) 4 to 5 stars, (ii) 3 stars, and (iii) 1 to 2 stars.

**(2) <u>Sentiment analysis using different approaches</u>**

Sentiment analysis can be influenced by various factors, such as the nature of the text, the context in which words are used, and the subtleties of human language. By experimenting with two distinct approaches as listed below, we aimed to determine which approach would provide the most accurate and reliable results for our project:

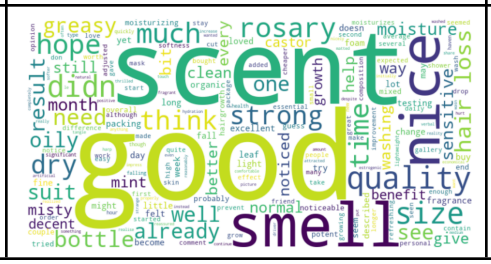
- NLTK **VADER** (Valence Aware Dictionary and sEntiment Reasoner): It is a lexicon and rule-based tool for sentiment analysis, particularly effective for social media text. It uses a lexicon to assign sentiment scores to words and applies rules to evaluate the sentiment of the text in context. VADER outputs 'pos', 'neu', and 'neg' scores, indicating the positive, neutral, and negative proportions of the text, which sum to 1. Additionally, it provides a 'compound' score ranging from -1 to +1, offering a quick measure of overall sentiment polarity.
- Huggingface **RoBERTa** Transformers: It is a deep learning model that uses the Transformer architecture to analyze complex textual data. It captures context and subtleties beyond surface-level sentiments. RoBERTa produces logits, which are then normalized into probabilities using Softmax, resulting in positive, neutral, and negative sentiment scores that sum to 1. These scores reflect the model's confidence in classifying the sentiment of the text, offering a nuanced understanding of its emotional tone.

To our surprise, both sentiment analysis approaches, VADER and RoBERTa, did not perform as expected on the preprocessed text, particularly when scoring sentiment on reviews with low ratings. This prompted us to further investigate the performance of these tools on our review dataset. To do this, we utilized boxplots to visualize the distribution of sentiment scores from both VADER and RoBERTa. This visualization technique offers insights into how sentiment scores are distributed across different review ratings and can help assess the impact of text preprocessing on the models' outputs.

In addition to visualizing the distribution of scores, we conducted error detection for both VADER and RoBERTa by examining reviews with the lowest ratings (1-star) that had the top 3 highest positive sentiment scores, and reviews with the highest ratings (5-star) that had the top 3 highest negative sentiment scores. This analysis was conducted for both preprocessed and non-preprocessed text to identify any patterns or discrepancies that might be attributable to the preprocessing steps.

**D. Analysis and Discussion**

**(1) <u>Word clouds by rating</u>**

<table>
<tr><td colspan="3">Mielle, Strengthening Hair Products, Rosemary Mint (Shampoo, Conditioner, Treatments Best Sellers)</td></tr>
<tr><td>4 to 5 stars</td><td>3 stars</td><td>1 to 2 stars</td></tr>
<tr><td>  </td><td>  </td><td>  </td></tr>
</table>

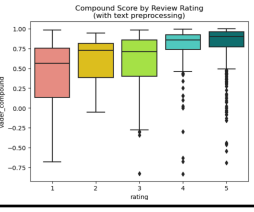
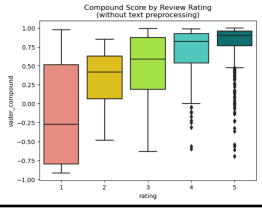
- For the 4 to 5 stars rating, the word cloud prominently featured words such as "smell", "good", "great", "scent", "size", "quality", and "packing" or "package". This indicates that customers who rated these products highly often praised their pleasant aroma, overall quality, and packaging. The emphasis on "smell" and "scent" suggests that the fragrance of these products is a significant factor in customer satisfaction.
- In the 3-star rating, while words like "scent", "good", and "smell" still appeared, indicating some level of satisfaction, words in medium size like "hair loss", "oily", and "greasy" also emerged. This suggests that customers with a neutral rating had mixed feelings about the products, appreciating the scent but also experiencing issues such as hair loss and the products making their hair oily or greasy.
- For the 1 to 2 stars rating, the word cloud highlighted "even" and "time" in the largest size, alongside "hair loss", "oily", and "greasy". This significant placement suggests that customers who rated these products poorly were particularly frustrated with the products' performance. The words "even" and "time" likely refer to customers' experiences of having to shampoo multiple

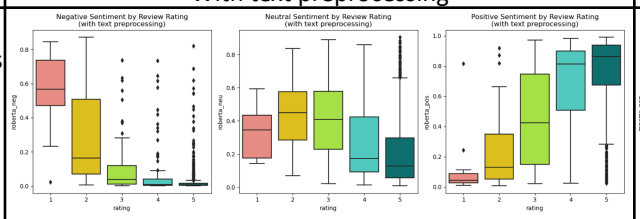
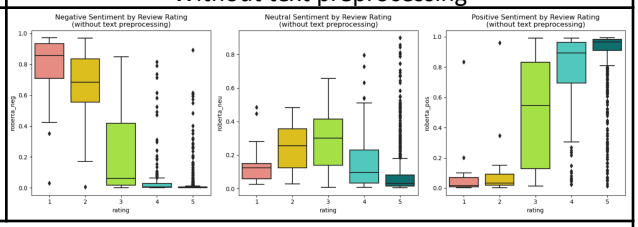
times due to the product leaving their hair feeling oily or greasy, which is a strong indicator of dissatisfaction.

| Giovanni, L.A. Hold Styling Gel, Strong Hold (**Styling Best Seller**) | | |
|---|---|---|
| 4 to 5 stars | 3 stars | 1 to 2 stars |
|  |  |  |

- For the 4 to 5 stars rating, the word cloud prominently displayed words such as "great", "good", "scent", "size", "package", and "hold". This indicates that customers who gave high ratings were particularly impressed with the product's performance, enjoying its scent, the size of the product, its packaging, and most notably, the strength of the hold it provided. The emphasis on "hold" suggests that the product's ability to maintain hairstyles was a significant factor in customer satisfaction.
- In the 3-star rating, the words "hold", "scent", and "size" appeared in larger sizes, indicating that customers with a neutral rating still recognized the product's key attributes, such as its scent and the size of the product, as well as its hold capabilities.
- For the 1 to 2 stars rating, the word cloud highlighted "dry" and "hold" in larger sizes. This indicates that customers who gave low ratings were particularly critical of the product's effect on their hair's texture, with "dry" being a significant concern. Despite the product's hold being mentioned, the negative connotation of "dry" suggests that the strength of the hold came at the cost of hair dryness, which was a major point of dissatisfaction for these customers.

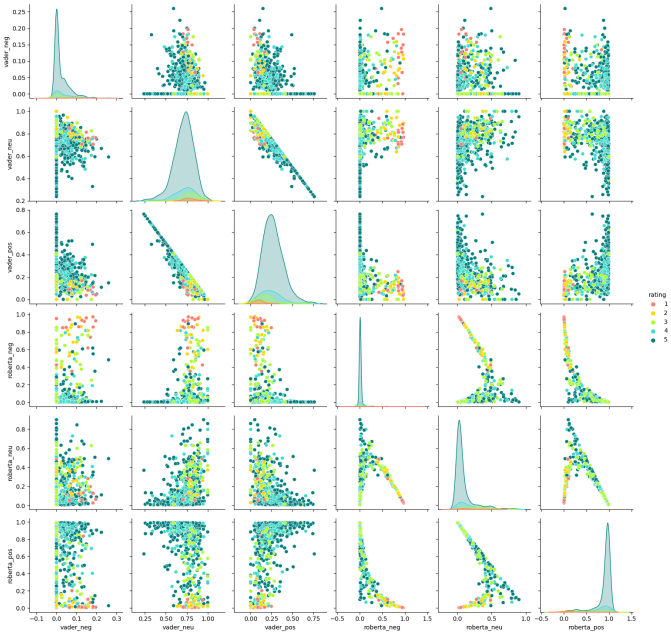
### (2) <u>Sentiment analysis using different approaches</u>

| **VADER** | With text preprocessing | Without text preprocessing |
|---|---|---|
| Boxplots - sentiment scores against review ratings |  |  |
| | The boxplots without preprocessing demonstrate a wider range of compound scores, particularly for the 1-star ratings where no negative compound scores were observed with preprocessing. This suggests that VADER is capable of detecting more subtle variations in the raw text that could be obscured by preprocessing steps. | |
| 1- star rating with the top 3 highest positive sentiment scores | Text preprocessing led to altered reviews, as seen in the top 1-star ratings where positive sentiment scores were 0.415, 0.332, and 0.284. This approach introduced issues such as incorrect spelling corrections, like "hype" to "hope", and the removal of negation words, such as "wouldn't", which likely skewed the sentiment analysis by not accurately reflecting the reviewer's intent. | VADER assigned lower positive sentiment scores of 0.248, 0.166, and 0.133 to some critical reviews. The original text preserved the reviewers' actual language, including negative expressions and proper spelling. This approach captures the reviewers' dissatisfaction more accurately, with complaints like greasy hair and excessive hair loss being clearly articulated. |
| 5- star rating with the top 3 highest negative sentiment scores | The top 5-star reviews with preprocessing exhibited higher negative sentiment scores (0.435, 0.4, 0.384), potentially due to the removal of context, such as "don't miss it", which turned positive statements into negatives. | Reviews without preprocessing had lower negative sentiment scores (0.26, 0.224, 0.2), accurately reflecting the reviewers' intent and preserving the context. |

| **RoBERTa** | With text preprocessing | Without text preprocessing |
|---|---|---|
| Boxplots - sentiment scores against review ratings |  |  |
| | The boxplots for sentiment scores without preprocessing show a more distinct separation across different review ratings, particularly in the positive sentiment category. The median scores progressively increase with higher review | |

| | | |
|---|---|---|
| | ratings, indicating a clearer differentiation of sentiments. In contrast, the boxplots with preprocessing exhibit overlapping interquartile ranges and closer medians, suggesting that preprocessing may obscure the inherent patterns in the data. | |
| 1- star rating with the top 3 highest positive sentiment scores | RoBERTa's sentiment analysis with preprocessing on 1-star reviews showed high positive sentiment scores of 0.8160, 0.2453, and 0.1145, which contradicts the low ratings due to text alterations. The preprocessing obscured nuances, missing the negative shift in a special case with a high positive score of 0.8160, where the review contained significant dissatisfaction. | RoBERTa's analysis without preprocessing on 1-star reviews resulted in scores of 0.8347, 0.2015, and 0.1009, capturing reviewers' authentic sentiment more accurately. However, the model still did not fully capture the sentiment shift in a review with a high positive score of 0.8347, indicating limited sensitivity to minor negative sentiment changes within mostly positive reviews. |
| 5- star rating with the top 3 highest negative sentiment scores | RoBERTa's sentiment analysis with preprocessing on 5-star reviews indicated minimal negative sentiment, with scores of 0.0236, 0.0282, and 0.0469. This suggests that it may have reduced the detection of negative language, despite the presence of negative phrases, leading to an underestimation of negative sentiment. | Without preprocessing, RoBERTa captured significantly higher negative sentiment scores of 0.8924, 0.6168, and 0.6070 in 5-star reviews. These elevated scores are justified, as the reviews contain predominantly negative wordings. The preservation of the original text's context more accurately reflects the negative feedback. |

In our analysis, both VADER and RoBERTa delivered better performance without text preprocessing, likely due to the preservation of essential sentiment cues in the raw text. VADER's strength lies in its ability to interpret direct sentiment indicators such as emojis and slang, which might be lost or altered during preprocessing. It also uses punctuation for sentiment intensity and capitalizes on words as emphasis signals. RoBERTa, with its contextual awareness, relies on the original text structure and natural syntax for accurate sentiment analysis, which may be compromised by preprocessing steps like lowercasing that obscure contextual clues. Furthermore, we observed that the NLTK library's stopwords list includes negation words critical for sentiment analysis; their removal during preprocessing could negatively impact the accuracy of both VADER and RoBERTa. Consequently, we determined to compare the performance of both tools in the situation of no text preprocessing and ultimately choose the best one to be used in our analysis.




The pairplot comparison between VADER and RoBERTa highlights the superior performance of RoBERTa across various metrics. RoBERTa's data points are more tightly clustered, indicating less variability and more consistent performance. This concentrated distribution suggests that RoBERTa provides more reliable and stable results, which is crucial for applications requiring dependable sentiment analysis.

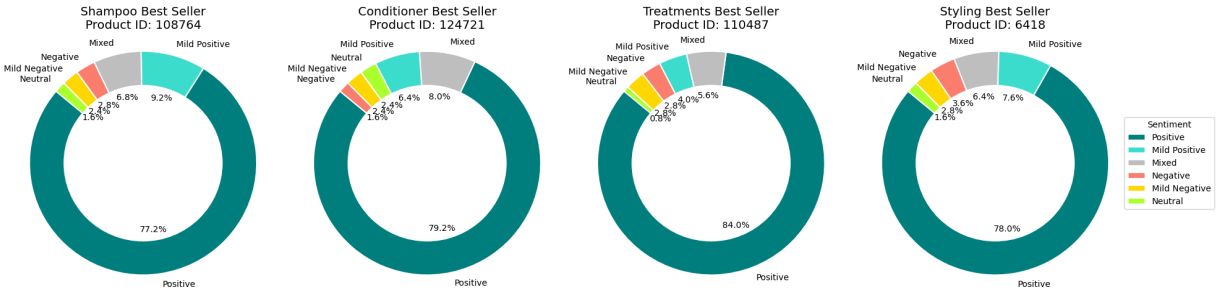
In contrast, VADER's points show a wider spread, reflecting higher variability in its performance. This inconsistency can lead to less predictable outcomes, making VADER less reliable for tasks that demand precision. The density plots further support this observation, with RoBERTa exhibiting sharp peaks that indicate a high frequency of scores within a narrow range, underscoring its reliability and precision.

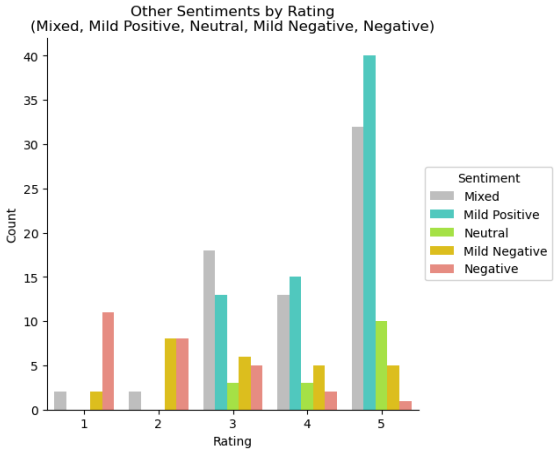
Overall, RoBERTa demonstrates more consistent and reliable performance compared to VADER, making it the preferable choice for sentiment analysis in our project.

### (3) Insights from reviews

To achieve a more nuanced understanding of customer sentiment, we have refined our sentiment classification by setting specific thresholds based on normalized scores derived from RoBERTa's sentiment analysis model:

- Positive: pos > 0.75 and neg < 0.25
- Mild Positive: pos > 0.5 and neg < 0.5
- Neutral: pos < 0.25 and neg < 0.25 and neu > 0.5
- Negative: neg > 0.75 and pos < 0.25
- Mild Negative: pos < 0.5 and neg > 0.5
- Mixed: Any instance that doesn't meet the criteria for the above categories (pos and neg are moderate; both around 0.3 to 0.5)




Our analysis has conclusively determined that ratings of 3 to 5 stars predominantly reflect positive sentiment, aligning with our expectations. Furthermore, it is a significant finding that the majority of mixed sentiments—our newly defined category encompassing both positive and negative feedback—also predominantly originate from these higher ratings.

This bar graph shows that mixed sentiments have high counts within the 3 to 5-star rating range. This pattern underscores our finding that even when sentiments are mixed—containing elements of both positivity and negativity—they tend to be associated with higher star ratings. This new category of mixed sentiment provides additional insight into customer feedback beyond the traditional positive or negative classification.

In this regard, we further examined the most positive, the most negative, as well as mixed reviews. We found that mixed reviews are more reliable because they provide a balanced view by capturing both positive and negative aspects of a product. This comprehensive perspective helps us understand the strengths and weaknesses of a product more clearly. Mixed reviews also offer detailed feedback, providing specific details on various product features.

Moreover, mixed reviews help set realistic expectations by presenting a balanced account, enabling potential buyers to have a more accurate understanding of what to expect, thus reducing the likelihood of disappointment. They are often seen as more trustworthy because they are neither overly optimistic or pessimistic, making them less biased and more credible. Mixed reviews also reflect the influence of ambiguity, acknowledging that different users may have varying experiences based on individual preferences or use cases. This adds to their overall reliability and usefulness, highlighting the value of mixed sentiments in customer feedback.

## Part 5. Conclusions

### A. Sales performance

- Mielle has the best sales in the past 30 days, with 3 of the 4 product categories ranking first.
- SheaMoisture, Giovanni and Mielle have four product categories on the best-selling list, which can be bundled into a series to sell to improve performance.
- Cantu's sales performance in hair styling is relatively good.
- Dove demonstrates poor sales performance on the iHerb platform.

*Dove did not rank as a best-seller. The dataset includes 72 Dove products, but 56 (or 78%) had zero sales in the past 30 days. Additionally, Dove's highest popularity score is only 2.

- There are 7 products rated 4 stars or above that are not available in Hong Kong. It is recommended to introduce these products to develop the Hong Kong market.

### B. Consumer preferences

- Consumers are most willing to spend money on treatments and are least willing to spend money on styling products.
- Consumers like good scent hair care products with great sizes.
- In shampoo, conditioner and treatments, hair loss, greasy and oily products are not popular with consumers.
- Consumers use styling products to hold the hairstyle or keep their hair curly, but don't want their hair to become dry after use.

## Part 6. Limitations and Challenges

### A. Limitations

**● Less key information available**

The iHerb website does not get much key information available, such as there are no repeated purchases, whether the customer is a loyal user, etc. If this information is available, an in-depth analysis of customer characteristics can be made.

**● Skewed rating distribution towards high scores**

The imbalanced distribution of ratings, which is heavily skewed towards high ratings, significantly hinders the development of robust models for machine learning applications. This lack of balance can lead to biased predictions and limit the model's ability to generalize effectively across different rating categories. Therefore, addressing this issue is crucial for improving the accuracy and reliability of sentiment analysis models.

### B. Challenges

**● Dataset selection**

The team initially had differing opinions on the capstone project theme, leading to delays due to an overload of ideas. After organizing our thoughts, we finally agreed on a theme. This experience highlighted the importance of time management in project work.

**● Web scraping challenge**

The anti-crawling measures on the iHerb website are very great. The team spent a lot of time and tried many methods to crawl to the required information. It was a big challenge for us to break through their anti-crawler measures, but we are happy to finally crawl the required data.

## Part 7. Future Work

### A. A more comprehensive analysis

Since it takes a lot of time to web scrape the iHerb website, and there are only a total of 7 one-star and two-star products on the iHerb website, which accounts for a small proportion of the total number of products (2563), the team did not crawl the one-star and two-star products. If we have enough time, the team hopes to analyze lower-rated products.

### B. Year-over-year comparisons

Because there are only sales of the product in the past 30 days, time-based comparisons cannot be made. For example: comparison of the sales of each brand in recent years. Most popular product changes etc.

### C. Create a customized list of stopwords to enhance sentiment analysis

To enhance sentiment analysis, it is essential to create a customized list of stopwords that specifically preserves negation words. This approach ensures that important contextual information is retained, allowing for more accurate interpretation of sentiments expressed in the text.

### D. Train sentiment models using a balanced dataset

To address overfitting issues, it is important to train sentiment models using a balanced dataset. This approach ensures that the models can generalize better to unseen data and improves their overall performance in sentiment classification tasks.

## Part 8. Key takeaways

### A. Key Insights on Sentiment Analysis Models

The following summary highlights the varying effectiveness of sentiment analysis models, particularly RoBERTa and VADER, in processing customer reviews and the importance of nuanced sentiment classifications.

- General text preprocessing steps may not always be required and largely depend on the chosen sentiment analysis approach.
- In our case, RoBERTa demonstrates more consistent and reliable performance across various metrics compared to VADER.
- In addition to Positive, Neutral, and Negative sentiments, the classification is further divided into Mild Positive, Mixed, and Mild Negative to provide deeper insights into customer reviews.
- The majority of mixed sentiments come from ratings between 3 and 5, making it worthwhile to investigate these reviews in more detail.

**B. "Result-oriented" Analysis**

● The team initially developed the dashboard with exploratory analysis but found the results unsatisfactory and lacking business context.

● By shifting the focus to defining key objectives and related metrics, the team redesigned the dashboard to deliver more meaningful and actionable insights.

**C. Teamwork and collaborative discussion**

● The project benefited from close collaboration and productive discussions within the team, where new ideas and perspectives were shared to improve the analysis.

● The team members effectively shared their domain knowledge with each other, facilitating the overall project progress.

● The supportive teamwork environment was crucial in driving the successful completion of the project.

## Part 9. Reference


Dataset
https://hk.iherb.com/c/hair-care

Use WebDriver to automate Microsoft Edge
Use WebDriver to automate Microsoft Edge - Microsoft Edge Developer documentation | Microsoft Learn

How to create Tooltip Pages in Power BI
https://www.youtube.com/watch?v=npaQ42K1sTs

Instant Data Scraper & Octoparse
https://www.youtube.com/watch?v=0xzTzw6GQiw&t=87s

NLP & Sentiment Analysis Tutorial
https://www.kaggle.com/code/furkannakdagg/nlp-sentiment-analysis-tutorial


## Part 10. Distribution of Work

Chung Yim Hung (Nicole):
Analysis methods research, no-code tools trying, presentation deck preparation and report writing

Chan Wing Ki (Teresa):
Analysis methods research, code writing, presentation deck preparation and report writing

Wong Wing Ying (Elaine):
Analysis methods research, code writing, presentation deck preparation and report writing