BA820 Final Deliverable

Data Cleaning & Preprocessing:

In the process of cleaning our dataset consisting primarily of products and reviews tables, we streamlined the data by dropping columns with high null value proportions or deemed non-essential, such as `sale_price_usd` and `value_price_usd`. Categorical columns were handled by replacing nulls with `NA` or `Unknown` for `ingredients` and `tertiary_category`, respectively. For products lacking reviews, we updated the review count with data from the reviews table or set it to zero otherwise. Missing values in the `ratings` column were imputed with the median due to its left-skewed distribution. In the reviews table, concatenating spreadsheets and eliminating redundant columns like `helpfulness` facilitated data consolidation. Post-cleaning, outlier detection was performed by utilizing boxplots to examine distributions of continuous variables like `price` and `reviews`. An identified outlier with a maximum price of $1900 was dropped, while duplicates were discerned based on product and brand names, accounting for differences in product size categorization on Sephora's platform. Additionally, `submission_time` was converted to datetime format, and numerical columns in both datasets were normalized to achieve a mean of 0 and standard deviation of 1.

EDA:

EDA provided valuable insights into the products and reviews available on Sephora's platform. The numerical features showcased significant variability in the products, with `loves_count`, ratings, reviews, and prices spanning a wide range **(Table 1)**. While the majority of products are not limited edition or exclusive to Sephora, a substantial portion is available exclusively online, indicating the platform's diverse offerings and accessibility. The correlation between `loves_count` and `reviews` underscores the importance of user engagement and satisfaction in driving review activity **(Fig 2).** The most loved and highly rated products encompass a variety of makeup, skincare, and fragrance items, reflecting customer preferences for quality and efficacy across different product categories. Analysis of brand popularity, category distribution, and price variations sheds light on Sephora's competitive landscape and customer preferences.
All product categories (excluding gifts), exhibit a wide price range **(Table 2),** indicating significant price variation across different categories, suggesting diverse pricing strategies and consumer preferences. Most skin types provide more positive feedback than negative ones, implying that Sephora effectively caters to the needs of various skin types **(Fig 3)**. The widening gap between positive and negative reviews throughout the years highlights Sephora's improvement **(Fig 4).** Overall, these findings provide valuable insights for both customers and Sephora in understanding product trends, brand perception, and customer satisfaction levels.

Analysis Plan for Project:

- **Market Basket Analysis:** We'll find product associations from products reviewed within the same week to observe relationships between the products that are purchased together.
- **Natural Language Processing (NLP):** We aimed to analyze user reviews for sentiment, entities, topics and trends. We'll leverage sentiment analysis to categorize user `reviews` into positive and negative sentiments, allowing us to analyze the prominence of different entities and their associations with sentiment. We'll utilize TF-IDF to uncover key topics and conduct sentiment trend analysis to track sentiment changes over time.
- **Clustering:** We'll use the results of NLP to cluster reviews based on sentiment, and to cluster products together, based on the `price`, `number_of_reviews` and `average_rating`, and then identify similar themes that products within a cluster share. We'll do this using KMeans, using the elbow method to determine the optimal number of clusters for the product, and using the silhouette method to further refine our results.

Results:

- **Market Basket Analysis:** We grouped products that were reviewed within the same week as a single purchase, to increase rule generation. From there, we initially set a support threshold of **0.003** to filter the rules. Moving our notebook to the Google Cloud provided us with higher computational power and memory, which allowed us to experiment with different thresholds. Ultimately, we saw two strong associations: *1. Products and their mini sizes and 2. Products within the same category* **(Table 3).** The identified product rules will potentially help Sephora leverage sales by recommending the other when an audience shows interest in one.
- **Natural Language Processing (TF-IDF):** Moving to the Google Cloud Platform allowed us to explore more demanding models such as NLP like TF-IDF. We cleaned the data by converting our `review_text` and `review_title` columns to lowercase, removing punctuation, correcting typos and stemming the words. We tokenized the columns, since it assisted in breaking down large chunks of text into manageable pieces. We then used TF-IDF to convert it into frequency. We used cosine similarity to validate if the TF-IDF worked well, by picking a review and checking if the Top 3 reviews with the highest score were actually similar **(Fig 6).**
- **Sentiment Analysis (GloVe):** We conducted a sentiment analysis on the reviews to explore customers' feelings about Sephora's products. We generated word clouds for the most positive and negative reviews **(Fig 8 & Fig 9).** Skincare products seem to provoke very diverging customer opinions. Sephora seems to explore how to serve the needs of customers who have different skin tones and skin types as customers seem to have either a very positive or negative experience across all skin features **(Fig 10)**.
- **Clustering of Reviews:** In order to understand how variables in the review table might influence reviews, we performed a KMeans clustering analysis. We found that the feedback count variables influence the clustering the most (**Fig 11)**. Sentiment scores have a minimal influence on the clustering and the price point only mattered for highly priced products. In short, Sephora may want to look into highly priced products as they seem to have an impact on reviews.

- **Products Clustering:** We clustered products based on their price, the number of reviews it received, and the average rating across all reviews. The findings revealed that Cluster 0, primarily composed of fragrances, has a higher minimum product price compared to other clusters, with an average rating of around 4.3. Cluster 1, with a high number of products, exhibits the widest range of ratings, the lowest average number of reviews, and the lowest average price, suggesting a correlation between affordability and mixed quality ingredients. Cluster 2, with the highest number of products across all nine categories, has a mid-price range and a decent average rating range from 3.8 to 5. Lastly, Cluster 3, with the fewest products and uneven category distribution, has the highest number of reviews **(Fig 12)**.

Challenges:

During our project, we encountered several challenges. Initially, we struggled with the computational power on Google Colab due to the large volume of data, over 1 million rows, after joining our tables. The process was time-consuming, prompting us to transition to the Google Cloud Platform for better efficiency.

In our Market Basket Analysis, we observed that a product and its mini size or travel-sized version are frequently purchased together. While this could be a genuine shopping pattern, we also considered the possibility that on the Sephora website, a review uploaded for a product might automatically apply to all sizes of that product. This could have influenced the rule generation, which was based on reviews.

Additionally, we faced difficulties with collaborating on GitHub, primarily due to our unfamiliarity with the platform and lack of prior training. Unexpected issues such as our SSH keys getting deleted added to the challenges. However, we managed to overcome these obstacles by continually learning through tutorials and reinstalling our SSH Keys each time they were deleted. This experience, although challenging, has been a valuable learning journey for us.

Next Steps:

For the next steps in this project, we plan to refine our approach to better capture the nuances of Sephora's specific domain wording. We aim to make the ideal positive and ideal negative vectors more inclusive of these unique expressions, such as including words like "smooth", "transformative", "effective", and "improved" for positive adjectives and words like "greasy", and "irritate" for negative adjectives. We could also use the themes of the TF-IDF results and create vectors based on this.

Additionally, we can implement Named Entity Recognition (NER) to identify specific entities within the data. This will allow us to accurately pinpoint product names and brands, providing a more detailed and precise analysis. By incorporating these steps, we hope to enhance the accuracy and relevance of our project outcomes.

Contribution Table

Team Member	Coding Contribution for this Deliverable
Oumou	Sentiment Analysis
Jessica	NLP: TF-IDF and cosine similarity
Alima	Clustering on Product Table
Asra	Clustering on Reviews Table
Sonya	NLP: data pre-processing and TF-IDF

Link to Dataset: https://rb.gy/25glxh

Appendix

Table 1: Basic Statistics for Numerical Features

Statistic	Loves Count	Rating	Reviews	Price (USD)
count	8494.0	8494.0	8494.0	8494.0
mean	29179.56592889098	4.197616811867202	433.8650812338121	51.65559453732046
std	66092.1225898942	0.5084476938088001	1086.7317719275052	53.669234407455846
min	0.0	1.0	0.0	3.0
25%	3758.0	4.0	22.0	25.0
50%	9880.0	4.289350000000001	112.0	35.0
75%	26841.25	4.5225	402.0	58.0
max	1401068.0	5.0	21281.0	1900.0

Fig 1: Distribution of Loves Count, Rating, Reviews

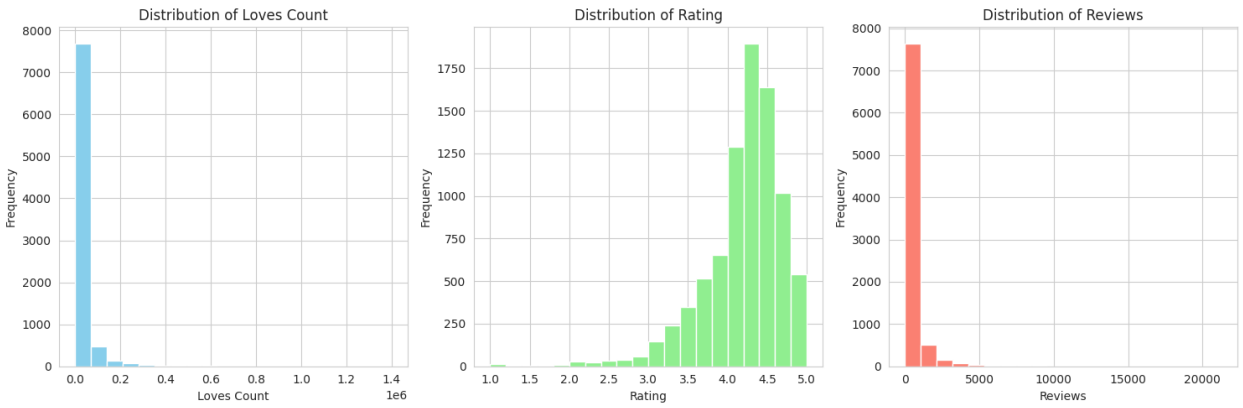

Fig 2: Correlation Matrix of Numeric Features

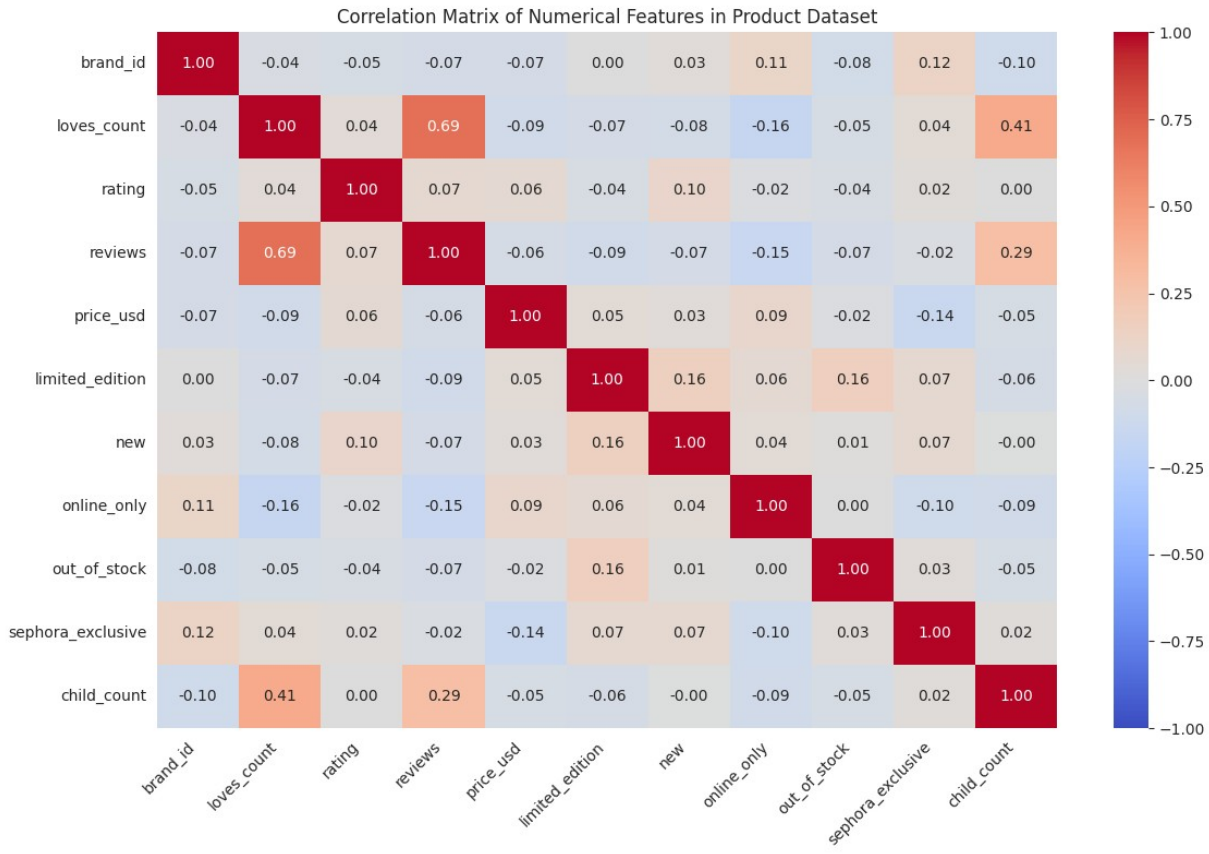

Table 2: Primary Categories Price Range

primary_category	min_price	max_price	avg_price
Bath & Body	3	300	42.2333
Fragrance	10	395	87.2626
Gifts	50	50	50
Hair	5	399	42.7867
Makeup	3	320	32.758
Men	10	104	33.2
Mini Size	3	165	21.3976
Skincare	3	495	59.7521
Tools & Brushes	4.95	249	31.9221

Fig 3: Skin Type Feedback

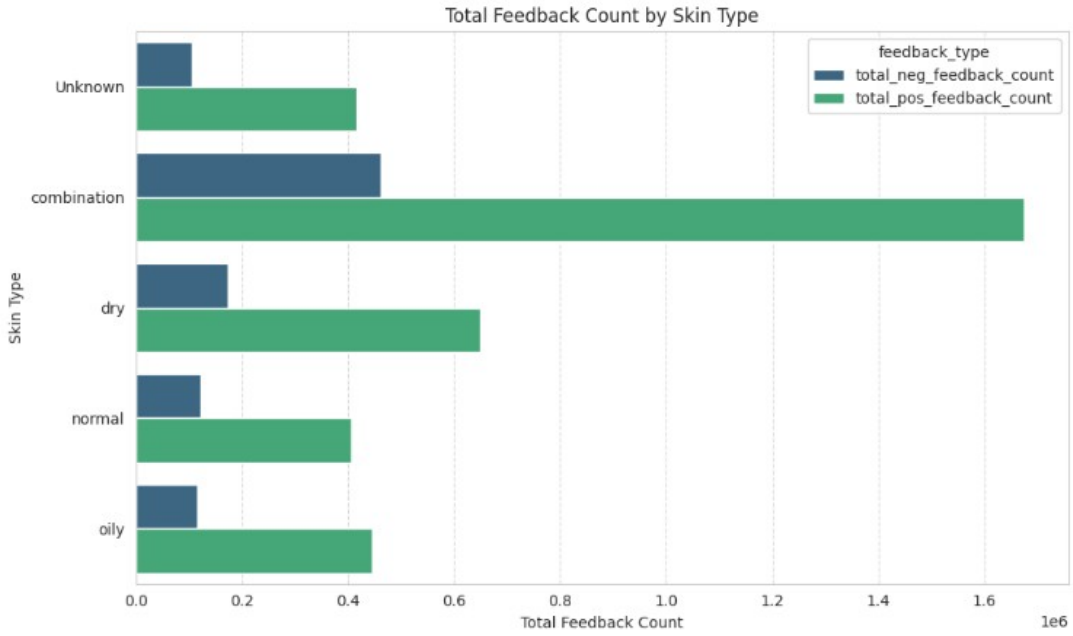

Fig 4: Feedback Trend

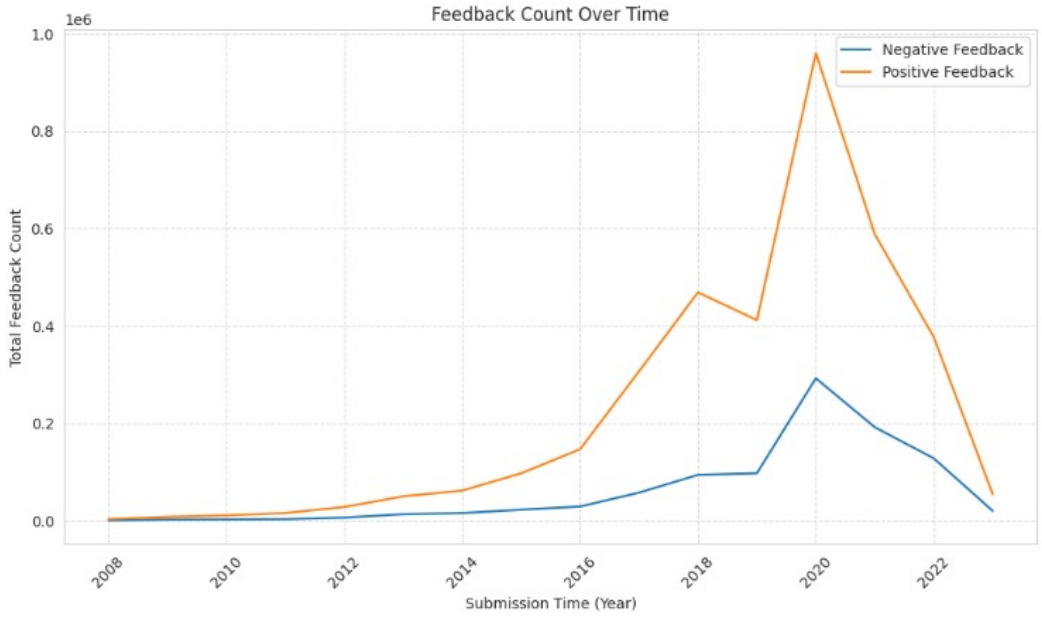

Table 3: Market Basket Example Results

Entry	Antecedents	Consequents	Support	Confidence	Lift
3	'Green Clean Makeup Meltaway Cleansing Balm'	['Green Clean Makeup Removing Cleansing Balm']	0.0054	0.9959	140.45
14	'**Mini** Superfood Antioxidant Cleanser'	['Superfood Antioxidant Cleanser']	0.0049	0.9879	149.58
1	'Daily Microfoliant Exfoliator'	['**Mini** Daily Microfoliant Exfoliator']	0.0046	0.9939	188.53
28	'(Set, Glow Face Mist)'	['Beauty Elixir Prep']	0.0034	0.9785	256.99

Fig 6: NLP Clustering

```
my skin GLOWS after using this serum!!! By far this is my favorite one ever. I have noticed a difference in my skin shade before and after using t
his product and wow. just wow. I recommend to anyone! Gifted by Sunday Riley
Perfect moisturizer for oily skin. Hydrates and balances oil leaving skin perfect for makeup application.
I have always used La Mer gel crème but wanted something that would be comparable in texture and consistency to wear under makeup but wouldn't bre
ak the bank. I think I have found it. I'll continue to use my La Mer (b/c nothing really compares) but this is great as another everyday option. L
ightweight, no smell, moisturizing.
```

Fig 7: Sentiment Scores Over Time

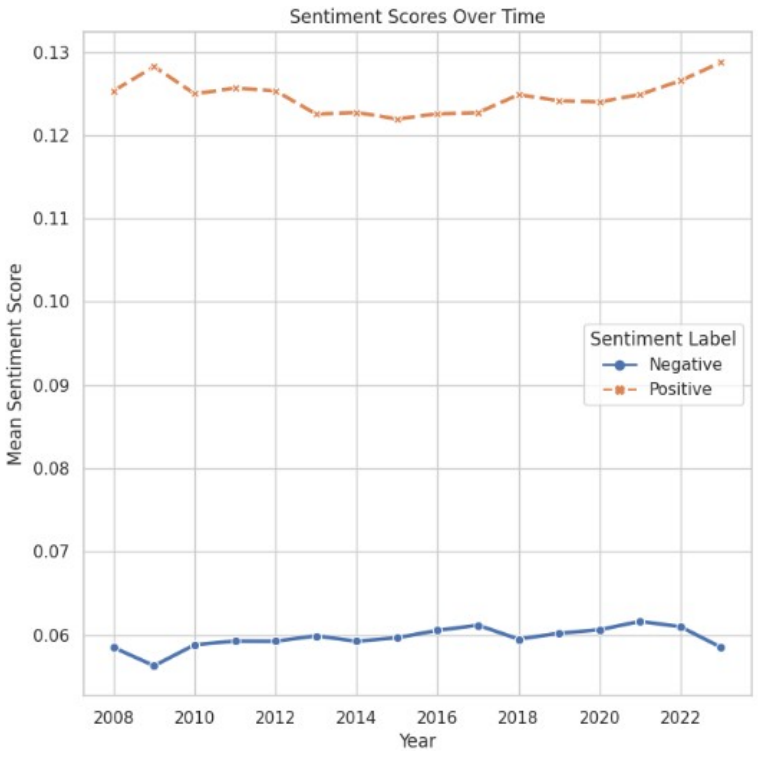

Fig 8: Word Cloud for Negative Reviews

Fig 9: Word Cloud for Positive Reviews

Fig 10: Sentiment Distribution by Skin Tone

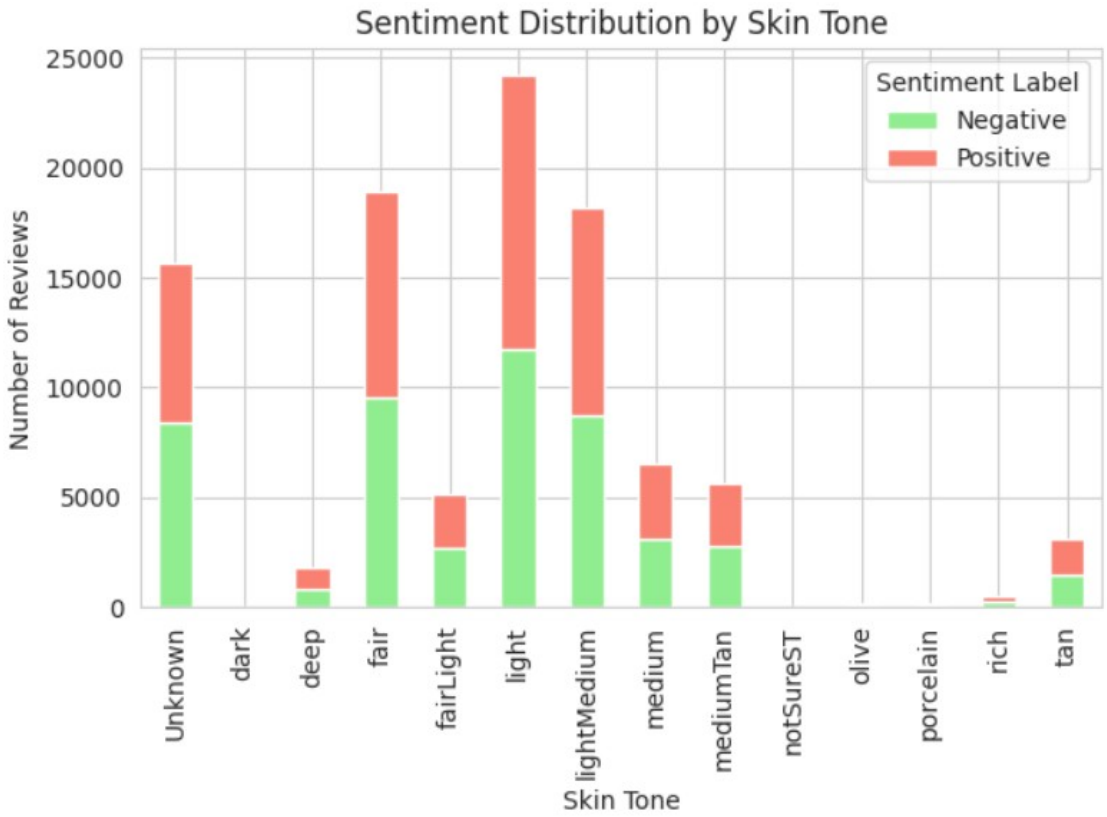

Fig 11: Feature Influencing Reviews Clustering

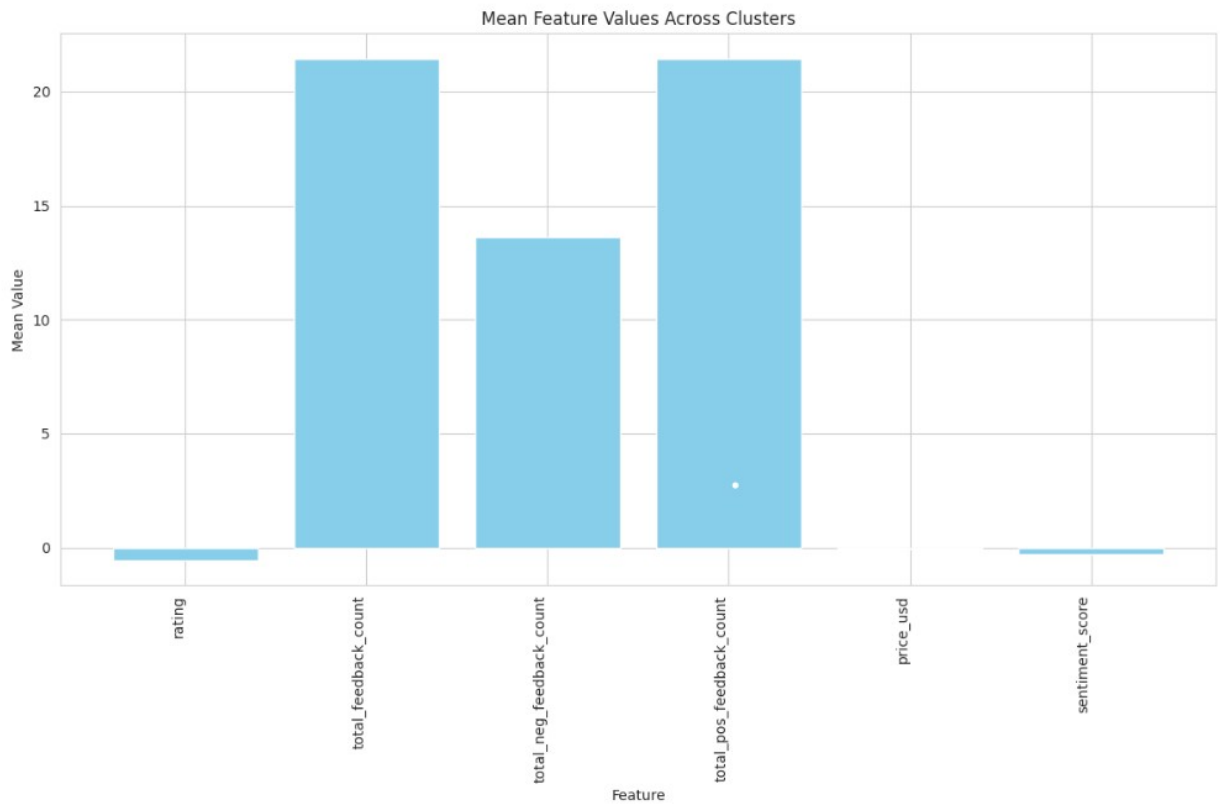

Fig 12: Category Distribution Across Clusters

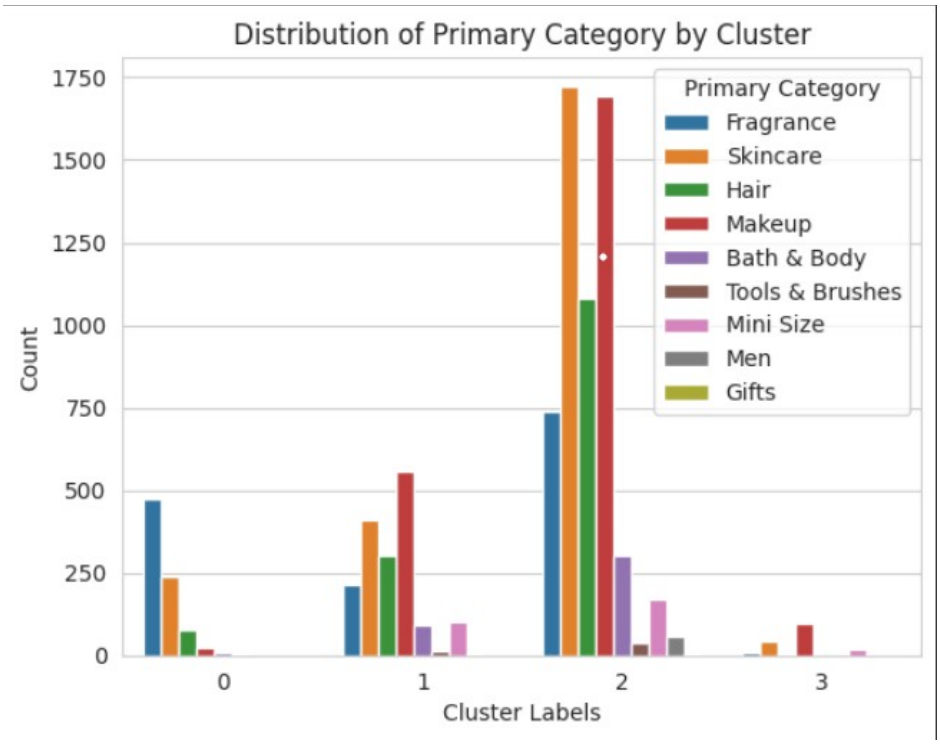

Fig 13: Clustering Using Elbow Optimal

Fig 14: Clustering Using Silhouette Optimal

Fig 15: Category Distribution Across Clusters