Strategic Enhancement of Sephora's Recommendations through Data-Driven Insights and Machine Learning

Samantha Grace S. Aniversario, Darius Vincent C. Ardales, Ayiana Louise G. Romero
Dr. Andrew L. Tan Data Science Institute, De La Salle University
samantha_aniversario@dlsu.edu.ph, darius_ardales@dlsu.edu.ph, ayiana_romero@dlsu.edu.ph

ABSTRACT

Sephora, a leader in beauty e-commerce, strives to enhance its customer experience through advanced digital tools and personalized recommendations. This project leverages Sephora's extensive product and review datasets to develop and refine recommender systems using content-based filtering, collaborative filtering with Singular Value Decomposition (SVD), and logistic regression. The study focuses on extracting insights from review characteristics to improve recommendation accuracy and customer satisfaction. Content-based filtering analyzes product attributes to suggest similar items, while collaborative filtering with SVD uncovers hidden patterns in user preferences. The logistic regression model adds a layer of prediction to estimate the likelihood of product recommendations based on numerical features. The models achieved high accuracy, demonstrating their effectiveness in tailoring product suggestions. Future enhancements could include integrating additional data sources and continuous model updates to adapt to evolving consumer preferences.

Keywords

Recommender Systems, Machine Learning, Logistic Regression, Sephora, Singular Value Decomposition (SVD)

1. Introduction

Sephora, an industry leader in beauty e-commerce, has become synonymous with beauty and style [3]. Founded in 1970 by Dominique Mandonnaud, this French multinational retailer offers personal care and beauty products. Now owned by LVMH Moët Hennessy Louis Vuitton, Sephora operates globally with a robust online presence. Renowned for digital innovation and providing a top-tier omnichannel experience, Sephora has pioneered mobile use in the beauty industry. The company is dedicated to delivering a seamless and engaging shopping experience across all channels, including their website, mobile app, and social media platforms. Their extensive e-commerce platform features a wide range of products, customer reviews, and beauty advice [1].

Despite being a beauty and personal care retailer, Sephora has become a digital-first player in the brick-and-mortar industry by leveraging technology to serve its customers better [4]. Among Sephora's digital transformation strategies is the use of augmented reality and artificial intelligence to provide facial analysis and visualization tools that help customers digitally try on products and let them receive product recommendations based on their skin. Aside from that, Seophora utilizes shopper data from mobile engagements and other digital initiatives to enhance the customer journey, providing them with more personalized advertisements and product recommendations.

2. Problem Statement

In the highly competitive beauty and personal care market, consumers often feel overwhelmed by the vast array of product choices, further complicated by their diverse and complex needs. Sephora, a leading retailer in this industry, offers a wide range of brands and products and receives numerous customer reviews daily. To improve sales, customer experience, and satisfaction, it is crucial to swiftly and accurately provide customers with product recommendations that suit their needs and preferences. By utilizing Sephora's product and review datasets, this project aims to develop recommender systems using content-based and collaborative filtering techniques supported by machine learning algorithms such as logistic regression, lasso, and ridge.

3. Objectives

The project aims to achieve the following:

- **Analyze and utilize Sephora's products and reviews data:** Extract and interpret valuable insights from Sephora's dataset, focusing on consumer buying behavior to inform relevant business decision-making.
- **Develop content-based and collaborative filtering recommender systems:** Build recommender systems that use both product attributes and user behavior patterns to provide personalized product suggestions.
- **Enhance recommendation accuracy through logistic regression:** Implement Logistic Regression with Lasso and Ridge penalties to predict the likelihood of product recommendations.

Through these objectives, the project seeks to develop a robust recommendation system that aligns with individual user preferences and provides actionable insights for business strategy, thereby improving the overall customer experience.

4. Exploratory Data Analysis

The 'Sephora Products and Skincare Reviews dataset was sourced from Kaggle and includes Sephora data from 2008 to March 2023. The dataset is divided into two parts: product data and review data [5]. The product dataset contains information on the product, such as its id, name, brand, price, and ingredients. The reviews dataset contains information on customer reviews, such as the author, date posted, the review text and title, and product rating. The reviews dataset comprises five CSV files. Hence, the first step in the pre-processing is appending all the reviews data into one DataFrame. Upon reviewing the reviews DataFrame, it was found that there were duplicates of reviews (different products but the exact review details). During data collection, items purchased together by a user at the same time (or received as part of a promotion) were processed to share the same review instead of each item having its own individual review. These duplicates were then dropped for better accuracy. Removing outliers or extreme values from this dataset is optional, as they can provide meaningful insights into a user's strong feelings about a product, whether they love or hate it.

When charted (Figure 1) from 2008 to March 2023, CLINIQUE is Sephora's most reviewed brand, with a total review count of 49,002. It is closely followed by Tatcha (with 46,768 reviews), Drunk Elephant (with 42,385 reviews), fresh (with 40,878 reviews), and The Ordinary (with 35,930 reviews). When the dataset is grouped by month and year (Figure 2), in April 2020, Sephora garnered the highest count of reviews at 23,924. This can be attributed to Sephora's annually anticipated spring sale [2]. Lastly, when grouped by year (Figure 3), the review count gradually increased from 2008 to 2020 but gradually decreased in the succeeding years. However, it's to be noted that the count for 2023 does not accurately represent the total count for that year, as the dataset ends in March 2023. The group created a dynamic chart to improve the visualization of how review counts by brand fluctuate throughout the year, as shown in Figure 4.

5. Modeling

In this project, the group implemented two recommender systems, content-based filtering, and item-based collaborative filtering, using Sephora's product and review datasets.

The group started the content-based filtering by preprocessing the dataset to fill in missing values and combine relevant features into a single column, combined_features. This column was then vectorized using Term Frequency-Inverse Document Frequency (TF-IDF) to create the item profile matrix. The similarity between items was calculated using cosine similarity, which allowed us to recommend products similar to those the user has shown interest in.

For item-based collaborative filtering, the group constructed an item-user matrix where rows represent items and columns represent users, with the matrix values representing the ratings given by users to items. We calculated item similarity using cosine similarity and identified the top k similar items to those rated by the user. Ratings were predicted based on the weighted average of similar items, and the top-rated items were recommended.

The group employed Singular Value Decomposition (SVD), a matrix factorization technique that decomposes the item-user matrix into three matrices to enhance collaborative filtering. This approach allowed us to reduce dimensionality and improve recommendation accuracy using latent features.

Lastly, the group also implemented a Logistic Regression model to predict the likelihood of a product being recommended based on a set of features. This model complements our existing recommender systems, which suggest similar items of interest to users and identify related top-rated products. The Logistic Regression model adds an extra layer of insight by assessing whether a specific item is likely to be recommended based on the characteristics of a given review. This approach is particularly valuable for addressing the significant number of missing values in the "is_recommended" feature. Reducing these null values is crucial, as it can substantially enhance the accuracy and effectiveness of the recommender system.

6. Discussion of Findings

6.1 Content-based Filtering

In the content-based filtering model, the system provided personalized recommendations by focusing on product attributes such as product name, brand, highlights, and categories. We first handled missing data and combined essential features into a single column. Then, the group used TF-IDF vectorization to transform the combined product information into numerical feature vectors for easier analysis.

For example, for a user who showed interest in products like "Fragrance Discovery Set" and "La Habana Eau de Parfum," the recommender system was able to identify key features such as "scent", "gift", "set", "discovery", and "fragrance". Based on these, the model recommended similar products like "Discovery Set," "Discovery Sampler Set," and "Peace Perfume Discovery Set." These recommendations mainly included different fragrance discovery sets and sampler kits, which indicates that the users prefer exploring various perfumes.

With this model, users can find new products that match their tastes and preferences, which improves their shopping experience. However, the content-based system faced challenges with less popular products that had limited feature data, making it less accurate for the recommendation on these items. For example, niche products with fewer descriptions and reviews were more challenging to match accurately with user preferences.

Despite this, this model showed significant potential in enhancing product discovery and providing personalized recommendations in the beauty e-commerce space, such as Sephora, as it can identify and suggest products that align with user preferences.

6.2 Collaborative Filtering with Singular Value Decomposition (SVD)

In applying the collaborative filtering model with Singular Value Decomposition (SVD), its goal is to provide personalized recommendations by analyzing user-product interactions. We created an item-user matrix from the rating data and then filled in missing values with the average rating for each product to handle gaps.

For instance, when the group applied SVD, the group decomposed the item-user matrix into several components and used these to reconstruct the matrix with fewer factors. This approach helped us discover hidden patterns in user preferences and product features.

From the results, the group generated recommendations for a specific user. For *author_id = 11741593524*, the top recommendations included:

- **Product ID *P504322:*** Rated 5.0, suggesting a high interest from the user.
- **Product ID *P420652:*** Rated 2.0, indicating the user favors it.

This approach helped us recommend products based on the user's past ratings and the patterns found through SVD. However, due to memory limitations, the group had to work with a smaller dataset, which limited the number of recommendations the group could generate. Ideally, using a larger database provides more comprehensive results.

The SVD-based model effectively helps identify products the user might like, though there's room for improvement if the group can manage larger datasets.

6.3 Logistic Regression

Before creating the Logistic Regression model, selecting the most relevant features for training was essential. Since the goal was to add a layer of insight to the recommender systems and minimize the null values in the "is_recommended" feature, the group opted to use only numerical features. This decision was based on the observation that string and object features did not require transformation into numerical representations suitable for the current Logistic Regression model.

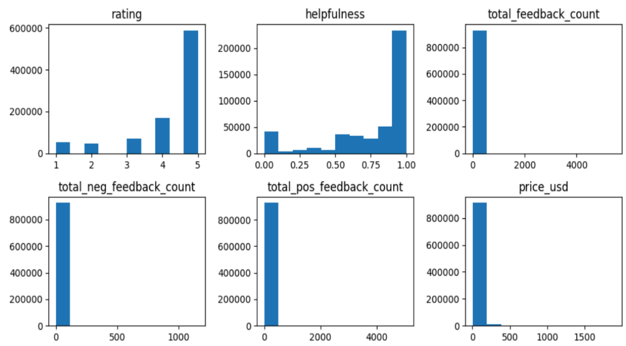


Figure 5. Feature Distribution

The numerical features were rating, helpfulness, total feedback count, total negative feedback count, total positive feedback count, and price in USD. However, these fields contained null values, which could negatively impact the model's performance. To address this, the group performed data imputation. To determine whether mean or median imputation was appropriate, the group examined the skewness of each feature's distribution using histograms (Figure 5). Since all features exhibited skewed distributions, the group imputed the missing values using the median. Additionally, the target feature, "is_recommended," had many null values. We decided to drop these null entries, as retaining them would be uninformative for training the model, and the large dataset size mitigated the impact of this decision on model accuracy.

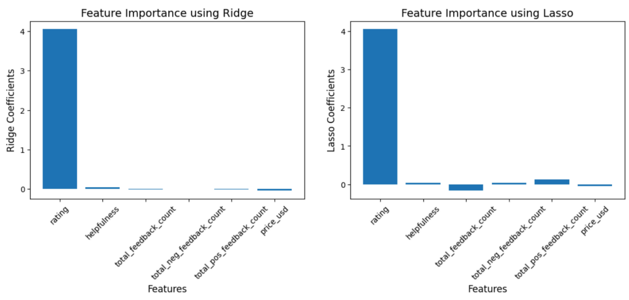


Figure 6. Feature Importance

We employed Ridge and Lasso regression (L2 and L1 regularization for Logistic Regression, respectively) to determine the important features. As shown in Figure 6, 'rating' emerged as the most important feature, followed by 'helpfulness,' 'price_usd,' and 'total_pos_feedback_count.' We selected these four features to train the model.

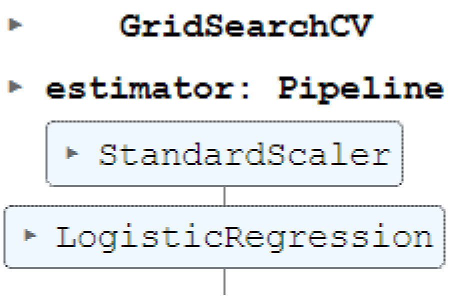
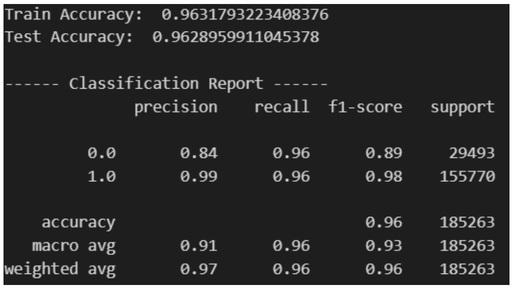


Figure 7. Logistic regression Pipeline

For the Logistic Regression model, the group implemented a pipeline with a standard scaler followed by the Logistic Regression algorithm (Figure 7). This approach ensured that the data was scaled consistently, reducing the impact of outliers and improving model performance.

Next, the group conducted a grid search to determine the model's optimal regularization parameter (C value). The results indicated that a C value of 0.01 was the best choice. The model coefficients revealed that 'rating' had a strong positive influence (coefficient: 3.979395) on the likelihood of a product being recommended, while 'price_usd' had a negative impact (coefficient: -0.035325), suggesting that higher prices may decrease the likelihood of a recommendation.



```
Train Accuracy:  0.9631793223408376
Test Accuracy:  0.9628959911045378

------ Classification Report ------
              precision    recall  f1-score   support

         0.0       0.84      0.96      0.89     29493
         1.0       0.99      0.96      0.98    155770

    accuracy                           0.96    185263
   macro avg       0.91      0.96      0.93    185263
weighted avg       0.97      0.96      0.96    185263
```

Figure 8. Logistic regression model evaluation

The model achieved a training accuracy of 96.32% and a test accuracy of 96.29%, indicating robust performance. As seen in Figure 8, the classification report further validates the model's effectiveness. For class 0 (not recommended), the precision is 0.84, with a high recall of 0.96, resulting in an F1-score of 0.89. The model performs exceptionally well for class 1 (recommended), with a precision of 0.99, a recall of 0.96, and an F1-score of 0.98. The model's overall accuracy is 96%, with a macro average F1-score of 0.93 and a weighted average F1-score of 0.96, underscoring the model's reliability across both classes.

7. Conclusion and Recommendations

This study has explored the different recommender systems to enhance Sephora's product suggestions. As mentioned, content-based filtering, item-based collaborative filtering with SVD, and logistic regression were utilized in this study. The content-based filtering approach could recommend products based on specific attributes like scent and branch matching the user's preferences. The item-based collaborative filtering with the SVD model helped us discover hidden patterns in user-product interactions, which aided in personalized recommendations, even if the group had a smaller dataset that posed some challenges. This logistics regression model added an extra layer of predictions, highlighting which products will likely be recommended based on factors such as rating and price. These models can improve Sephora users' shopping experience by offering them more personalized product recommendations.

To further improve the recommendation accuracy, integrating additional datasets, such as social media sentiment analysis or customer demographics, could offer a more comprehensive view of consumer behavior and preferences. Additionally, it is crucial to continuously and regularly retrain the models with new data to adapt to the evolving market and consumer preferences. This can be achieved through automated pipelines that regularly update the recommendation algorithms with the latest datasets. Incorporating a feedback loop could also be valuable as users can rate the recommendations they receive, thus helping further refine the algorithms.

8. References

[1] About Us | Sephora. (n.d.). Sephora. https://www.sephora.com/beauty/about-us

[2] Hou, K., Larkworthy, J., & Smith, E. (2020, April 17). What to Buy During Sephora Beauty Insider Savings Event 2020. The Cut. https://www.thecut.com/2020/04/sephora-beauty-insider-spring-savings-event-2020.html

[3] Schwanke, C. (2023, May 3). What Is the History of Sephora? LoveToKnow. https://www.lovetoknow.com/life/style/what-is-history-sephora

[4] Sephora: A Technology Foundation That's More Than Skin Deep, and a Digital Success Story That's More Than Meets the Eye - Digital Innovation and Transformation. (2020, February 11). Digital Innovation and Transformation. https://d3.harvard.edu/platform-digit/submission/sephora-a-technology-foundation-thats-more-than-skin-deep-and-a-digital-success-story-thats-more-than-meets-the-eye/

[5] Sephora Products and Skincare Reviews. (2023b, August 2). Kaggle. https://www.kaggle.com/datasets/nadyinky/sephora-products-and-skincare-reviews/data

A. Appendix

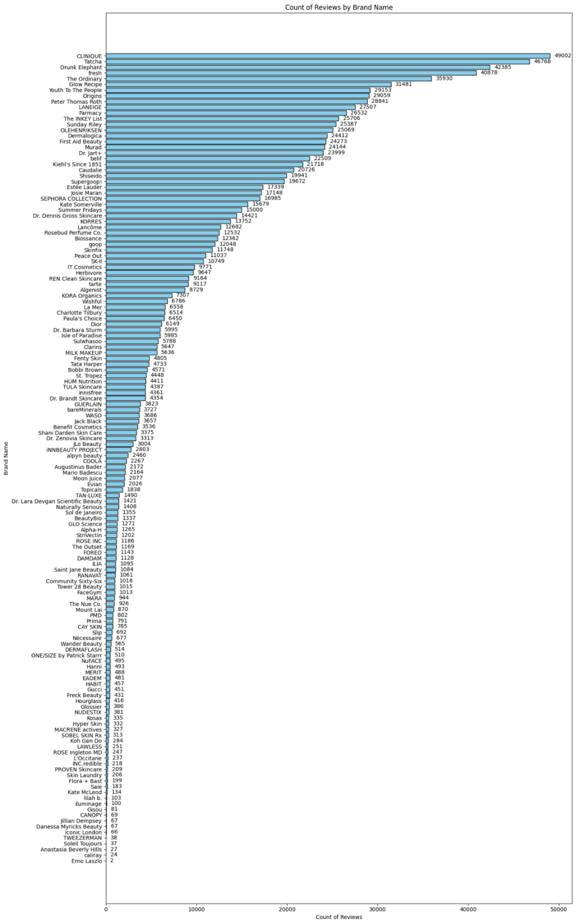


Figure 1. Count of Reviews by Brand Chart

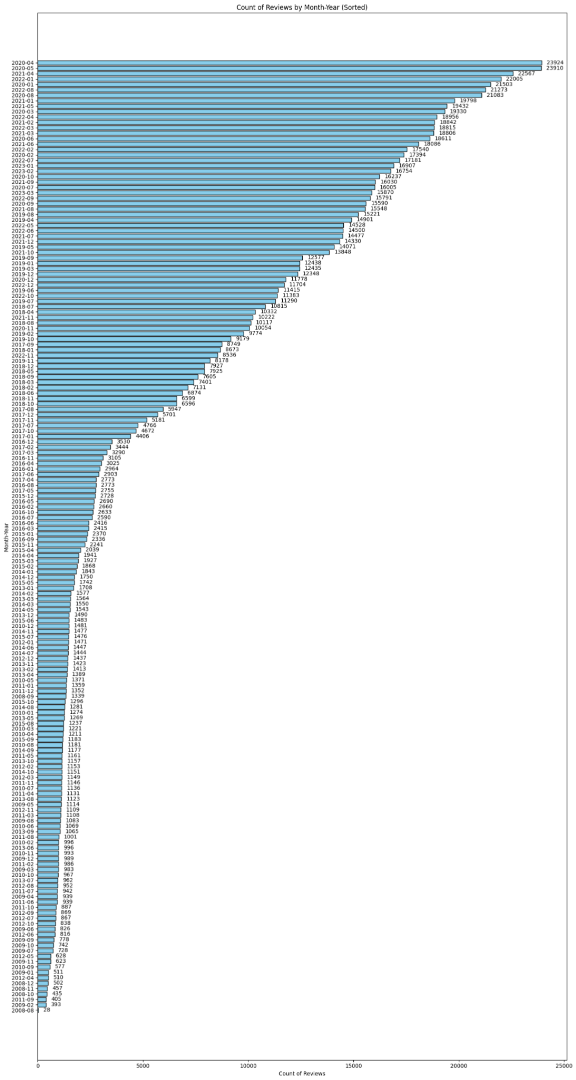


Figure 2. Count of Reviews by Brand Chart

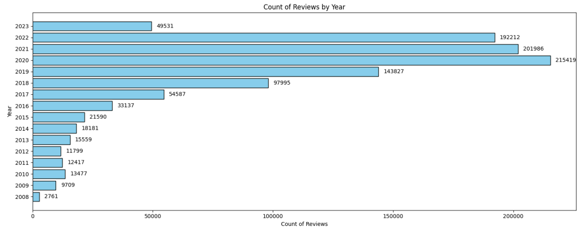
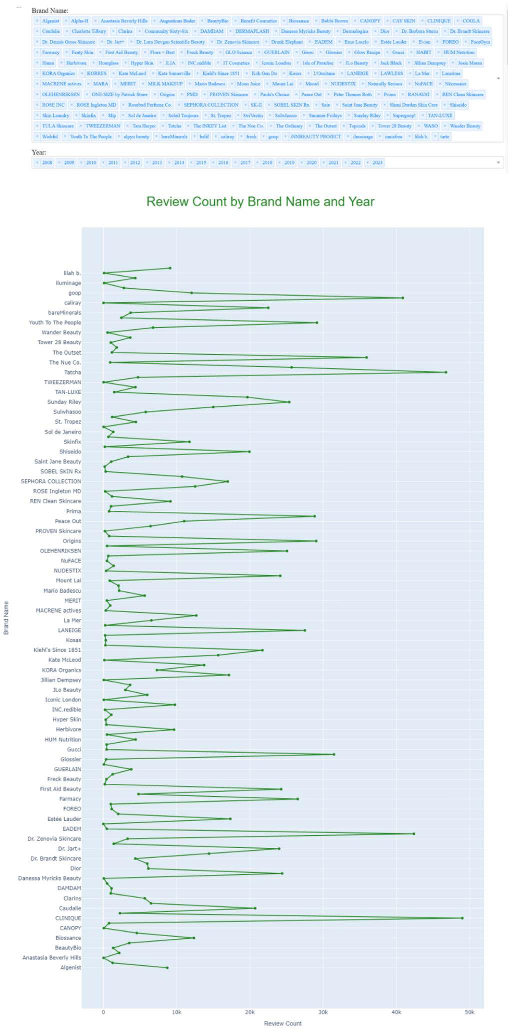


Figure 3. Count of Reviews by Year



Brand Name:

× Algenist × Alpha-H × Anastasia Beverly Hills × Augustinus Bader × BeautyBio × Benefit Cosmetics × Biossance × Bobbi Brown × CANOPY × CAY SKIN × CLINIQUE × COOLA
× Caudalie × Charlotte Tilbury × Clarins × Community Sixty-Six × DAMDAM × DERMAFLASH × Danessa Myricks Beauty × Dermalogica × Dior × Dr. Barbara Sturm × Dr. Brandt Skincare
× Dr. Dennis Gross Skincare × Dr. Jart+ × Dr. Lara Devgan Scientific Beauty × Dr. Zenovia Skincare × Drunk Elephant × EADEM × Erno Laszlo × Estée Lauder × Evian × FOREO × FaceGym
× Farmacy × Fenty Skin × First Aid Beauty × Flora + Bast × Freck Beauty × GLO Science × GUERLAIN × Gisou × Glossier × Glow Recipe × Gucci × HABIT × HUM Nutrition
× Hanni × Herbivore × Hourglass × Hyper Skin × ILIA × INC.redible × IT Cosmetics × Iconic London × Isle of Paradise × JLo Beauty × Jack Black × Jillian Dempsey × Josie Maran
× KORA Organics × KORRES × Kate McLeod × Kate Somerville × Kiehl's Since 1851 × Koh Gen Do × Kosas × L'Occitane × LANEIGE × LAWLESS × La Mer × Lancôme
× MACRENE actives × MARA × MERIT × MILK MAKEUP × Mario Badescu × Moon Juice × Mount Lai × Murad × NUDESTIX × Naturally Serious × NuFACE × Nécessaire
× OLEHENRIKSEN × ONE/SIZE by Patrick Starrr × Origins × PMD × PROVEN Skincare × Paula's Choice × Peace Out × Peter Thomas Roth × Prima × RANAVAT × REN Clean Skincare
× ROSE INC × ROSE Ingleton MD × Rosebud Perfume Co. × SEPHORA COLLECTION × SK-II × SOBEL SKIN Rx × Saie × Saint Jane Beauty × Shani Darden Skin Care × Shiseido
× Skin Laundry × Skinfix × Slip × Sol de Janeiro × Soleil Toujours × St. Tropez × StriVectin × Sulwhasoo × Summer Fridays × Sunday Riley × Supergoop! × TAN-LUXE
× TULA Skincare × TWEEZERMAN × Tata Harper × Tatcha × The INKEY List × The Nue Co. × The Ordinary × The Outset × Topicals × Tower 28 Beauty × WASO × Wander Beauty
× Wishful × Youth To The People × alpyn beauty × bareMinerals × belif × caliray × fresh × goop × iNNBEAUTY PROJECT × iluminage × innisfree × lilah b. × tarte

Year:

× 2008 × 2009 × 2010 × 2011 × 2012 × 2013 × 2014 × 2015 × 2016 × 2017 × 2018 × 2019 × 2020 × 2021 × 2022 × 2023

Figure 4. Dynamic Count of Reviews by Brand and Year Chart