Sephora Products and Reviews

Final Project – Data Analysis Course

Project Objective

This project analyzes Sephora product data to identify top-performing products based on popularity, satisfaction, and availability.
and to understand how pricing, discounts, and online-only status affect customer behavior and product success.

Dataset Overview

The dataset used in this project was sourced from Kaggle and contains product-level and user-generated review data from Sephora. It includes over 8,000 beauty products and nearly 1 million reviews in the skincare category. The two main components of the dataset are:

- Product Information includes attributes such as product name, brand, pricing, availability, exclusivity, and highlights.
- Review Data includes attributes such as User feedback, ratings, recommendations, review text, skin characteristics, and feedback helpfulness.

Link to the dataset: [Sephora Products and Skincare Reviews]

Relevance

1. **Personal Interest in Beauty & Skincare**
   Our team has a strong passion for makeup and skincare, which makes this dataset especially engaging and encourages us to explore the insights more deeply and enthusiastically.
2. **Rich Consumer Feedback for Analysis**
   The dataset includes detailed product reviews and ratings, which allows us to explore customer sentiment, satisfaction trends, and product performance—perfect for extracting actionable business insights and key performance indicators (KPIs).
3. **Global Brand Recognition**
   Sephora is a leading international beauty retailer with a strong global presence. This makes the data relevant and relatable to a wide audience, including potential stakeholders in the beauty and retail industry.

Dataset Structure and Relationship

The data is organized in two related tables connected by the 'product_id' field.
This relationship allows us to merge product attributes with review behavior for deeper insight.
Schema: One-to-Many Relationship
- Product Info Table: One row per product.
- Reviews Table: Many reviews per product.

Data Dictionary

Product info

Feature	Description
product_id	A unique identifier assigned to each product on the website.
product_name	The name of the product.
brand_id	A unique identifier assigned to the brand of the product on the website.
brand_name	The name of the brand that produces the product.
loves_count	The number of users marked the product as a favorite.
rating	The average user rating of the product (on a scale from 1 to 5).
reviews	The total number of user reviews submitted for the product.
size	The size or quantity of the product, expressed in units like oz, mL, g, packs, etc., depending on the product type.
variation_type	The category of product variation such as size, color, or shade.
variation_value	The specific value associated with the variation type (e.g., “100 mL”, “Golden Sand”).

variation_desc	Additional descriptive information related to the variation (e.g., “for fair skin tones”).
ingredients	A list of ingredients included in the product, for example: [‘Product variation 1:’, ‘Water, Glycerin’, ‘Product variation 2:’, ‘Talc, Mica’] or if no variations [‘Water, Glycerin’]
price_usd	The regular (standard) price of the product in U.S. dollars.
value_price_usd	The estimated total value of the product, especially used for bundles or value sets. It shows how much it would cost to buy all items separately.
sale_price_usd	The current discounted price, if the product is on sale. If there's no discount, it might be the same as price_usd.
limited_edition	Indicates whether the product is a limited edition: 1 for true, 0 for false.
new	Indicates whether the product is newly launched: 1 for new, 0 for not new.
online_only	Indicates whether the product is only sold online or not: 1 for online only, 0 otherwise.
out_of_stock	Indicates whether the product is currently out of stock: 1 means it's out of stock, 0 means it’s available.
sephora_exclusive	Indicates whether the product is exclusive to Sephora: 1 for exclusive, 0 for non-exclusive.
highlights	A list of tags or features that highlight the product's attributes (e.g., “Vegan”, “Cruelty-Free”, “Matte Finish”).
primary_category	The first-level category assigned to the product in Sephora’s site hierarchy (breadcrumb navigation).
secondary_category	The second-level category in the breadcrumb navigation.
tertiary_category	The third-level category in the breadcrumb navigation.

child_count	The number of product variations (e.g., shades, sizes) available for this item.
child_max_price	The highest price among the variations of the product.
child_min_price	The lowest price among the variations of the product.

Reviews Data

Feature	Description
author_id	A unique identifier assigned to the user who wrote the review.
rating	The rating given by the author for the product (on a scale of 1 to 5).
is_recommended	Indicates whether the reviewer recommends the product: 1 for yes, 0 for no.
helpfulness	A ratio measuring how helpful users found the review: calculated as total_pos_feedback_count / total_feedback_count.
total_feedback_count	Total number of helpfulness votes (both positive and negative) left by users for the review.
total_neg_feedback_count	The number of users who rated the review as unhelpful.
total_pos_feedback_count	The number of users who rated the review as helpful.
submission_time	The date the review was posted on the website as 'yyyy-mm-dd' format
review_text	The full text content of the review written by the user.

review_title	The title of the review written by the author
skin_tone	The reviewer’s reported skin tone (e.g., fair, tan, deep).
eye_color	The reviewer’s eye color (e.g., brown, green, blue).
skin_type	The reviewer’s skin type (e.g., oily, dry, combination).
hair_color	The reviewer’s hair color (e.g., black, blonde, auburn).
product_id	A unique identifier linking the review to the specific product.

Target Audience

Stakeholder: E-commerce Optimization Specialist

We chose the **E-commerce Optimization Specialist** as our stakeholder due to the strong alignment between their responsibilities and the analytical value of our dataset. This role is directly responsible for enhancing the performance of the online store by making data-driven decisions related to Product performance, pricing strategies, inventory availability, and overall digital customer experience. Our dataset includes detailed product-level attributes (e.g., pricing, sale status, exclusivity, stock availability, category) alongside customer-generated data such as ratings, recommendations, and written reviews. This dual perspective enables us to uncover insights that are directly actionable for an e-commerce specialist, such as identifying high-performing online-only products, evaluating the impact of discounts on satisfaction, or determining whether top-rated products are consistently available in stock.

Persona Paragraph

Name: Sarah Hanks.
Demographics: 34 years old, based in New York City, married without children, holds a Master's degree in Digital Marketing and Analytics, with 7 years of experience in the beauty and retail industry.
Job Title: E-commerce Optimization Specialist at Sephora.
Goals & Motivations: Sarah has always been passionate about merging creativity with data. After starting her career in merchandising, she moved into e-commerce optimization to focus on the fast-growing online beauty market. Now at Sephora's headquarters, she works to improve the online store's performance by ensuring top products are visible, fairly priced, and consistently available. Her daily focus includes analyzing customer behavior, evaluating how discounts affect satisfaction, and identifying which product features drive conversions.
Pain Points: She struggles with out-of-stock issues for top-rated items, unclear performance of online-only products, and the difficulty of identifying which product features drive customer satisfaction.
Data Needs: Sarah relies on data to understand how products are performing online. She looks at customer ratings, loves count, and review sentiment, and compares them across different price levels, stock availability, and product variations like size or shade. She needs clear, easy-to-read KPIs that highlight what's working well and where improvements are needed on the website.

Audience Questions Mapped to Analysis

Question	How We'll Answer It
Are high-rated products often out of stock, and does this impact the user experience?	We'll compare out_of_stock status with ratings and recommendation levels to detect missed opportunities.
How does pricing (regular, value, and sale) affect customer satisfaction?	We'll calculate KPIs like price sensitivity and discount impact by comparing price_usd, sale_price_usd, and value_price_usd with rating and is_recommended.
Which product variations (e.g., size, color, shade) receive the highest satisfaction and engagement?	We'll analyze customer ratings and recommendation rates by variation_type and variation_value to identify which variations are most positively received.

KPIs

1. Price Sensitivity (Rating vs. Price vs. Sale Price)
   → Understands how price influences perception.

2. Out-of-Stock Frequency of Highly Rated Products
   → Helps identify inventory issues affecting revenue.

3. Performance of Online-Only Products (Loves Count + Rating)
   → Measures the success of items not sold in stores.

4. Helpfulness Ratio of Reviews (Positive Votes / Total Votes)
   → Identifies reviews that most influence purchase decisions.

5. Discounted Price Trends vs. Customer Feedback
   → Evaluates how markdowns affect satisfaction.

6. High vs. Low Performing Products by Availability Status
   → Explores the connection between inventory and customer experience.

EDA Summary (Exploratory Data Analysis)

While we focused most of our insights in the final dashboard, key EDA steps helped shape the direction of our analysis:

Product Dataset EDA:

- **Price Distribution:** Helped define product tiers and identify outliers.
- **Ratings Distribution:** Confirmed that most products were rated positively.
- **Loves Count:** Revealed customer engagement and preference patterns.
- **Child Count Distribution:** Offered insight into product variation complexity.

Review Dataset EDA:

- **Ratings Distribution:** Validated the overall tone of customer sentiment.
- **Helpfulness Distribution:** Assessed the trustworthiness and utility of reviews.
- **Reviews Volume:** Showed popularity trends

These exploratory analyses guided KPI selection and ensured that our dashboard focused on dimensions most relevant to the E-commerce Optimization Specialist.

Key findings

Best-Performing Online-Only Product

The Essence Skincare Boosting Treatment (Lunar New Year Edition)

This product leads as the top performer among online-only items, based on a formula that combines two metrics, Rating and Loves_count:
ZN([Rating]) * 20 + LOG(ZN([Loves Count]) + 1)

ZN() is short for **"Zero if Null"**
→ If the rating is missing (null), it treats it as **0**
→ If it has a value (e.g., 4.8), it keeps that value

ZN([Rating]) * 20
It turns the small 1–5 rating scale into a **bigger, more visible score**.
We want ratings to have a **strong influence** on the total performance.

LOG(ZN([Loves Count]) + 1)
The **logarithm** takes big numbers and **shrinks them down**
We also add +1 so that if the love count is 0, we get:
log(0 + 1) = log(1) = 0
Loves count can be **very big and unbalanced**. Log helps keep it fair.
It ensures that **popular products** still benefit, but **don't dominate** just because of raw numbers.

What does the whole formula mean?
Take the product's rating, make it powerful (by multiplying by 20), and add a scaled version of how much customers love it.

Helpfulness Ratio — 77%

A ratio measuring how helpful users found the review: calculated as total_pos_feedback_count / total_feedback_count.

- A 77% helpfulness ratio indicates that customer reviews are highly useful and trustworthy. This level of helpfulness enhances the shopping experience significantly
- Maintaining or improving this ratio should remain a focus to support confident purchase decisions and reduce returns.

% of High-Rated Products Out of Stock — 2%

- Only 2% of high-rated products are currently out of stock, which is a strong signal of effective inventory and demand forecasting.

- Continuous monitoring is advised to ensure that demand surges (e.g., due to discounts or viral popularity) don't disrupt this balance.

Customer Ratings by Price — Colored by Discount Status

Most highly rated products are priced under $100, indicating strong customer satisfaction within this range. There is no significant concentration of discounts among the highest-rated products, suggesting that discounts are not the primary driver of positive reviews. Instead, product value and quality likely play a more important role in shaping customer perception. Products above $200 receive scattered and often lower ratings, hinting at potential price sensitivity at higher tiers.

Most Reviewed Products

- The **"Lip Sleeping Mask Intense Hydration with Vitamin C"** stands out as the **most reviewed product**, with **7,822 reviews**, significantly higher than others.

- Other products with high review volume include:

    - **Niacinamide 10% + Zinc 1% Oil Control Serum** (~4,276 reviews)

    - **Daily Microfoliant Exfoliator** (~4,282 reviews)

This suggests strong customer engagement and popularity, making these products ideal candidates for promotion, bundling, or highlighting on the website.

Most Loved Products

- The same **"Lip Sleeping Mask Intense Hydration with Vitamin C"** also leads in **"Loves Count"**, crossing **1 million likes**, indicating not only high engagement but also very strong customer satisfaction.

- **Niacinamide 10% + Zinc 1% Oil Control Serum** again appears high on the list, reinforcing its popularity and perceived value.

This implies that high love counts correlate well with strong emotional connection and loyalty, making these products prime for repeat purchase strategies and influencer promotion.

→→The overlap between most-reviewed and most-loved products (especially **Lip Sleeping Mask** and **Niacinamide Serum**) indicates a **strong**

alignment between popularity and satisfaction.

Top Performing Brands

- **Erno Laszlo**, **Gisou**, and **MARA** rank as the **top three performing brands**, indicating strong overall customer satisfaction, engagement, and product quality across their offerings.

- Brands like **MACRENE Actives** and **DAMDAM** also show strong performance, suggesting growing consumer trust and appeal in their product lines.

- These rankings can help Sephora's e-commerce team prioritize which brands to feature prominently on the homepage, bundle in promotional campaigns, or stock more aggressively.

Online-only vs not online only

Customer satisfaction and performance levels are consistent across online-only and in-store products, indicating a balanced experience regardless of availability

Recommendations

1. Promote Top-Performing Products More Aggressively

- Highlight the Lip Sleeping Mask Intense Hydration with Vitamin C (and The Essence Skincare Boosting Treatment) in marketing campaigns, especially on social and email channels, as it drives both popularity and satisfaction.

2. Encourage More Helpful Reviews Through Incentives

- With a **77% helpfulness ratio**, reviews are a valuable decision-making tool. Offer incentives (e.g., loyalty points or discounts) for users who write detailed, helpful reviews.

- Highlight the “most helpful” reviews directly on product pages to build trust and reduce hesitation.

3. Maintain Stock for High-Rated Products — Especially During Promotions

- With only 2% of high-rated products out of stock, the current supply chain is strong. Maintain this efficiency by monitoring trends before major sales events.

- Implement real-time alerts for high-rated products nearing low stock levels to prevent missed revenue opportunities.

4. Explore Strategic Discounting for Mid-Priced Products

- The scatter plot shows limited discount impact at very low or high price points. Focus discount strategies on mid-range products where customer rating sensitivity to price is more elastic.

5. Expand and Test Online-Only Product Lines in Top Categories

- Online-only products show nearly equal ratings to the in-store options, indicating strong potential. Invest in exclusive online launches within high-performing brands (e.g., Erno Laszlo, Marcene), and monitor engagement closely.

Limitations and suggested data needs

1. **Incomplete Behavioral Data**
   There is a lack of detailed behavioral data that captures how users interact with products on the Sephora website.
   Data needs: To gain deeper insights, it is essential to collect and analyze customer behavior signals such as click-through rates, add-to-cart rates, cart abandonment, and overall conversion rates. Additionally, understanding the user journey and identifying where users drop off in the purchase funnel can help uncover friction points and optimize the shopping experience.

2. **No Access to Sales Volume or Revenue Data**
   While we assessed product performance using reviews, ratings, and loves count, we could not directly analyze actual sales or revenue — limiting our understanding of financial impact.
   Data needs: Integrate sales quantity, revenue per product, and margin data to prioritize products that are both loved and profitable.

3. **The product variations are not handled correctly in the dataset**
   Variants of the same base product—such as different sizes, colors, or formulations—are listed under entirely separate product names and IDs, without a shared identifier to link them together. This fragmentation makes it difficult to analyze a product's overall performance across its variations or to compare how specific variants (e.g., 50 mL vs. 100 mL) influence customer satisfaction, pricing sensitivity, or sales.
   Data needs: For more accurate and unified product analysis, future datasets should include a common product identifier that groups all related variations under a single parent product.

4. **No Demographic or Segmented Insights**
   User-level data (age, location, gender, loyalty status) was not available, which restricts targeted recommendations for specific customer groups.
   Data needs: Enable segmentation by demographics, loyalty tier, or purchase history to tailor promotions and optimize online-only inventory strategy.

Challenges

1. Large-Scale Review Data Across Multiple Files

One of the main challenges we faced was managing the review data, which was spread across five separate files, resulting in over one million rows when combined. This large volume of data significantly complicated the cleaning process and increased the risk of duplication and inconsistency.

To address this, we considered two possible solutions:

- Option 1: Filter the data to include only the most recent reviews, starting from 2020 onward.

- Option 2: Filter by specific products to focus on those with a high volume of reviews for deeper product-level analysis.

After evaluating both options, we chose to proceed with the first, using reviews from 2020 and later. This decision was based on two key reasons:

1. The number of reviews has increased in recent years, making the dataset both richer and more representative.

2. More recent reviews are likely to reflect the current state of the products, including any improvements or reformulations, and therefore offer more relevant and accurate sentiment.

2. Handling Missing and Misleading Data

Several fields, such as sale_price_usd, variation_type, and secondary_category, had a high number of missing values. In some cases, missing data was misleading — for example, the field helpfulness had missing values even though its component fields were complete. This required additional validation logic and careful treatment to avoid distorting insights.

3. SQL Environment and Data Integration Issues

We faced several technical difficulties when attempting to upload the reviews data into SQL environments:

- **XAMPP (phpMyAdmin):** The reviews file exceeded the 40MB upload limit. Attempts to increase the limit by editing PHP configuration failed due to system-level permission restrictions on Windows, and the changes were not applied.

- **PostgreSQL:** The system could not read the reviews file as UTF-8 encoded, despite being saved as such. Illegal characters persisted, even after cleaning the text in Python.

- **DB Browser for SQLite:** Although the file was uploaded, multiline review texts caused structural issues, pushing data into unintended columns. We resolved this by cleaning and reformatting the data in Python before reimporting.