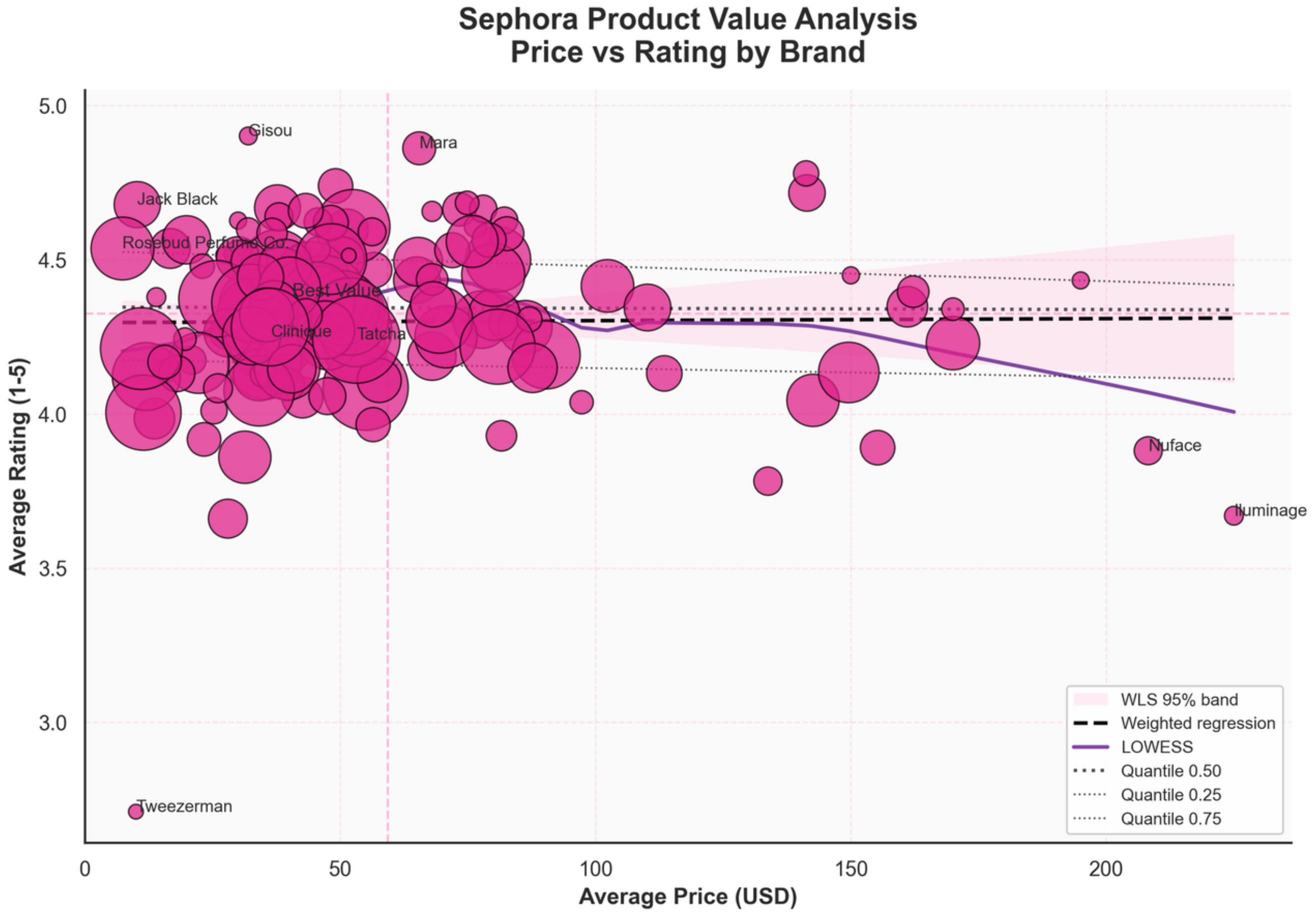
Sephora Product Value Analysis

Ⅰ. Introduction

This project is based on real beauty product and review data from the Sephora platform (data source: Kaggle – Sephora Products and Skincare Reviews).
The goal of this analysis is to visualize and explore the relationship between product prices and user ratings across different brands, in order to examine whether higher prices correspond to higher customer satisfaction, and to identify brands that demonstrate strong value-for-money performance or outstanding consumer reputation.

Ⅱ. Main Figure

Figure: Sephora Price vs Rating by Brand



2.1 Relationship Between Average Price and Average Rating

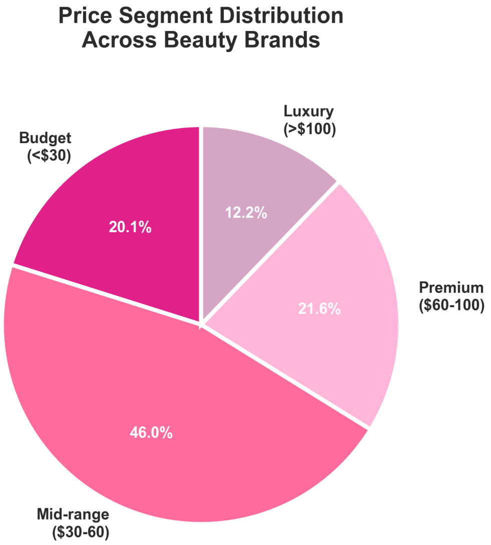
- Each pink bubble represents a beauty brand; bubble size reflects review volume.
- The black dashed line shows the weighted regression trend between price and rating.
- The purple LOWESS curve highlights non-linear patterns in the data.
- Text labels identify brands with notable ratings or visibility.

2.2 Legend Explanation

- **Pink bubbles:** Brand-level price and rating; size = review count.
- **Black dashed line:** WLS regression trend.
- **Purple curve:** LOWESS smoothing.
- **Shaded band:** 95% confidence interval for WLS.
- **X-axis:** Average price (USD).
- **Y-axis:** Average rating (1–5).
- **Text labels:** Selected standout brands.

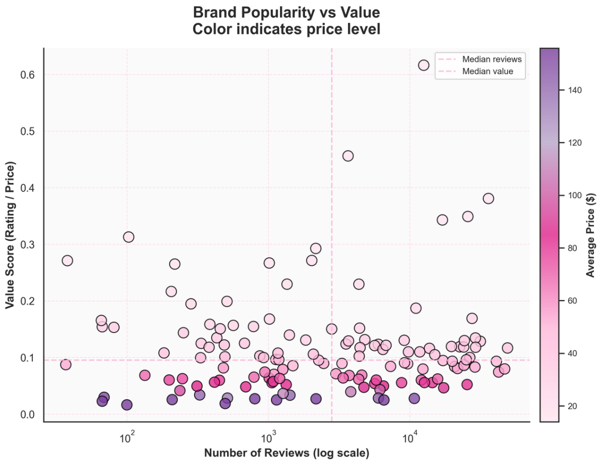
III. Supplementary Figures

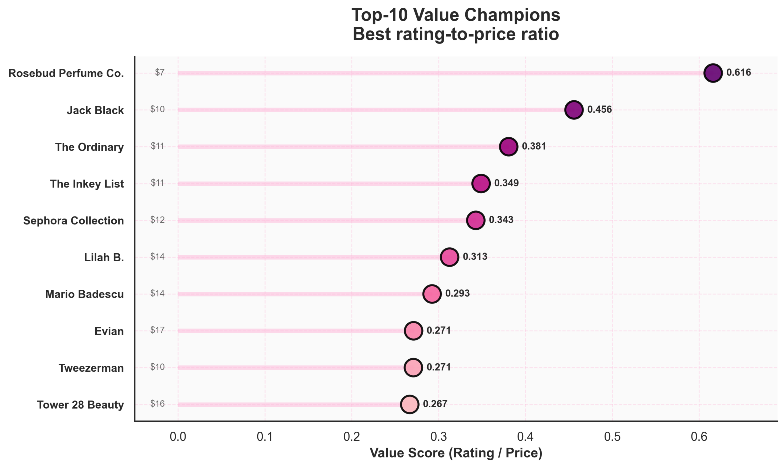
3.1 How Beauty Brands Are Distributed Across Price Tiers



- This pie chart shows how brands are distributed across four price tiers. Most brands fall into the mid-range category (46%), while luxury brands represent the smallest share (12%).
- Slice color = price tier; percentages = proportion of brands in each tier.

3.2 How Brand Popularity Relates to Product Value





- The scatter plot shows that brand popularity (review count) does not strongly predict product value (rating ÷ price), as many highly reviewed brands still cluster at low value scores. The Top-10 ranking on the right highlights affordable brands—such as Rosebud Perfume Co. and The Ordinary—that achieve the highest value scores, confirming that strong value is driven more by pricing efficiency than by popularity or brand visibility.

Legend Explanation

Scatter plot:	• **Top-10 chart:**
• Point color = price level • X-axis = review count (log scale) • Y-axis = value score • Dashed lines = median benchmarks	○ Bar length = value score ○ Color = average price ○ Labels = numeric score

IV. Key Findings

- price and average rating show only a weak relationship, indicating that higher-priced beauty brands do not necessarily earn higher customer satisfaction.
- Mid-range brands ($30–60) make up the largest share of the market, suggesting that most products target accessible price levels.
- Value scores (rating ÷ price) reveal large variation across brands, with many high-priced brands offering low value despite high ratings.
- Brand popularity (review count) does not strongly predict value; several highly reviewed brands still fall below the median value score.
- Affordable brands such as Rosebud Perfume Co., Jack Black, and The Ordinary consistently outperform premium and luxury brands in value efficiency.

V. Data & Methods

Data

- The dataset combines multiple Sephora review files (`reviews_*.csv`) with a product information file (`product_info.csv`).
- For each product, we use price, user rating, brand, and category information.
- Reviews are aggregated to the brand level, keeping only brands with at least **30 reviews** to ensure stable estimates.

Methods

- **Data cleaning and transformation**
  - Ratings and prices are converted to numeric values, and entries with missing or invalid values are removed.
  - Prices are winsorized at the **1st and 99th percentiles** to reduce the influence of extreme outliers.
  - Brand names are normalized (stripped and title-cased), and brand-level averages are computed for price, rating, and review count.
  - A **value score** is defined as *average rating ÷ average price*, and brands are assigned to four price tiers (Budget, Mid-range, Premium, Luxury) based on their average price.
- **Statistical modeling**
  - The linear relationship between price and rating is summarized using **weighted least squares (WLS)**regression, with review counts as weights.
  - A **bootstrap procedure** is used to estimate a 95% confidence band around the WLS regression line.
  - A **LOWESS smoother** captures potential non-linear patterns in the price–rating relationship, and **quantile regression** (25%, 50%, 75%) provides rating benchmarks across the price range.
  - Pearson correlation is reported as a global summary of the price–rating association.
- **Visualization**
  - All figures are produced in Python using **pandas, NumPy, Matplotlib, Seaborn, and statsmodels**, with a consistent pink-themed style.
  - Bubble sizes encode review volume, colors encode price level, and reference lines (medians, quantiles) are added to support interpretation.

VI. Significance Statement

Understanding how price, rating, popularity, and value interact in the beauty market helps identify whether consumers truly get better quality by paying more. By visualizing brand-level patterns, this analysis reveals important gaps between perception and actual product value, highlighting which brands offer exceptional performance relative to cost. These insights are useful not only for consumers making informed purchasing decisions, but also for brands aiming to position their products more effectively.