Predicting the Price of Beauty: A Multi-Model Analysis of Sephora Products

Jose Velasquez, Gabrielle Dominique

Skincare & Beauty Infographics

1

Introduction

2

Data Summary

3

Models/ Results

4

Conclusion

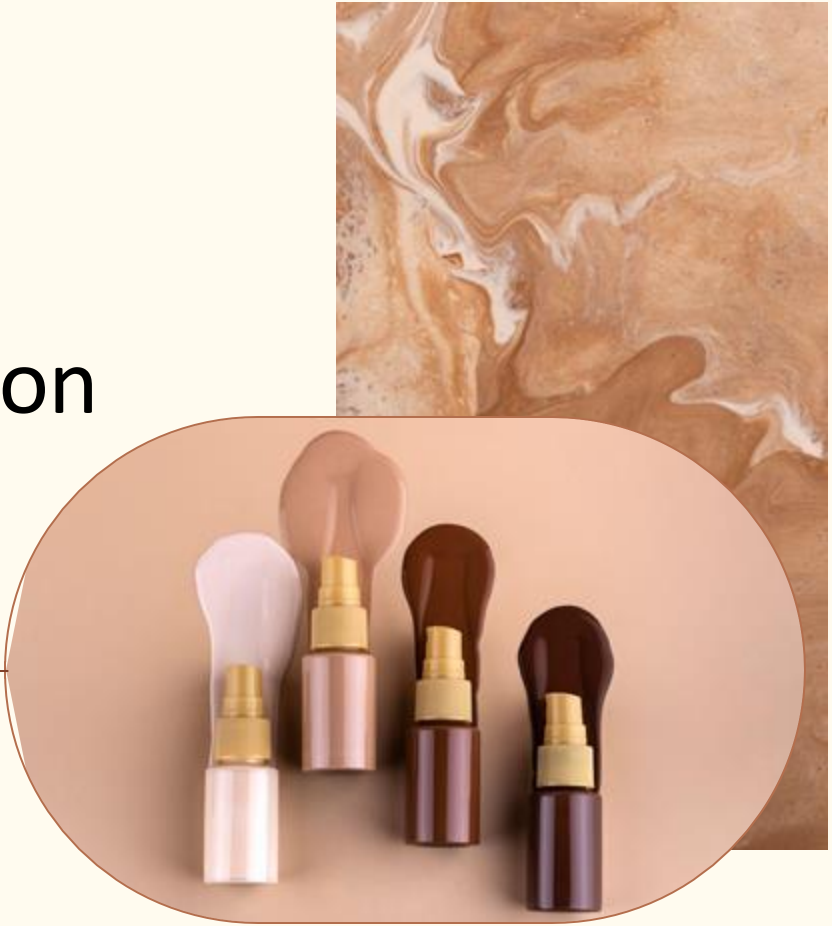
5

Questions

6

Appendix

Introduction



Our Team

Jose Velasquez Castellano
Lead Data Analyst

Gabrielle Dominique
Lead Data Analyst

Background

In 2025, Mckinsey & Co valued the beauty industry at around 450 billion dollars, clearly outlining the significance of the market.

As consumers become more vannlue-conscious and selective, the industry now faces pressure to justify product prices and prove real product performance.

This makes it important to understand what factors (like price, size, brand, or product category) actually influence consumer engagement and ratings.

Data Summary

Data Source

Dataset

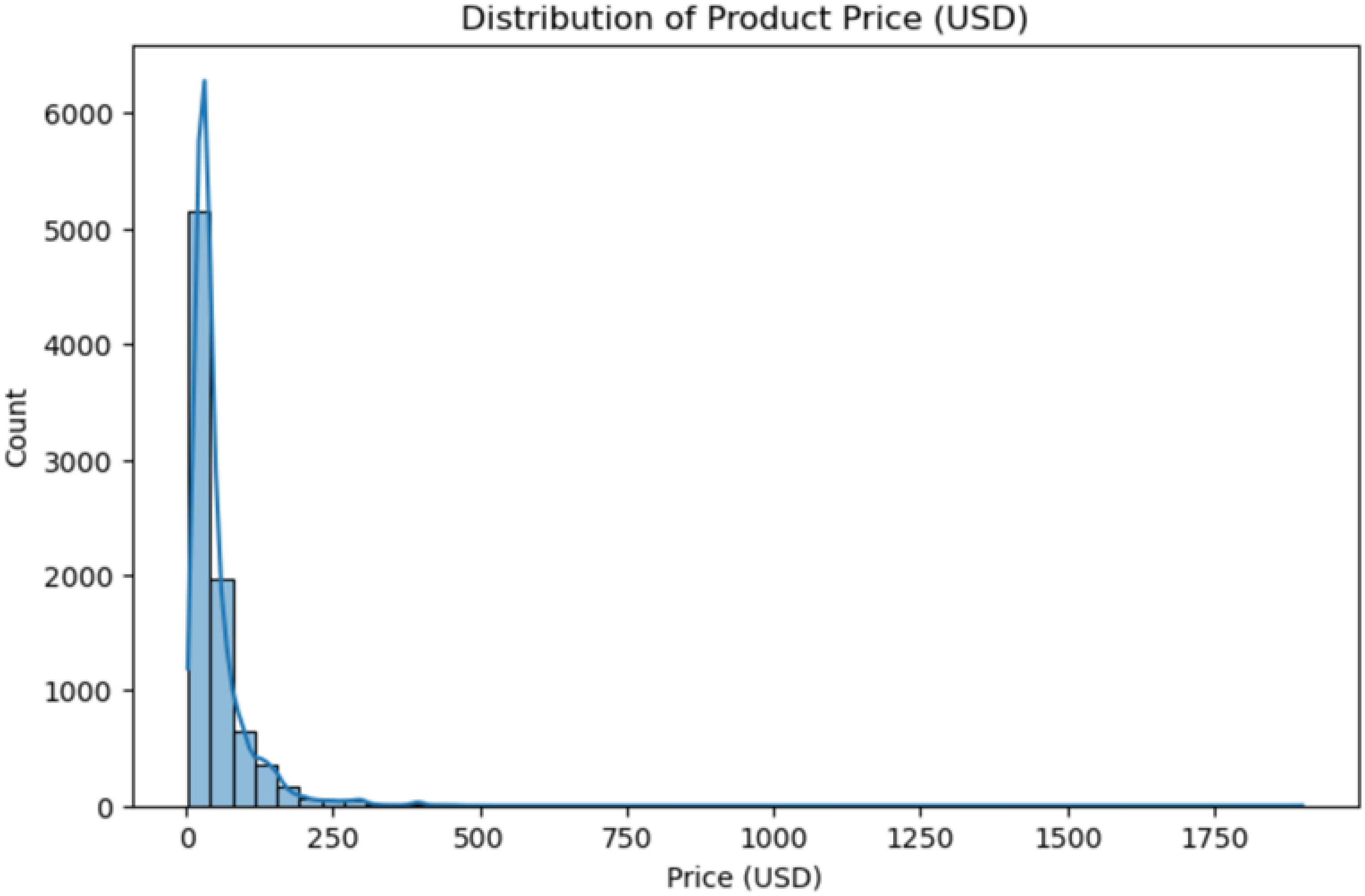
- Kaggle: Sephora Products & Skincare Reviews
- **8,000+ products** with price, brand, size, ratings, features
- **1M+ reviews** (skincare)

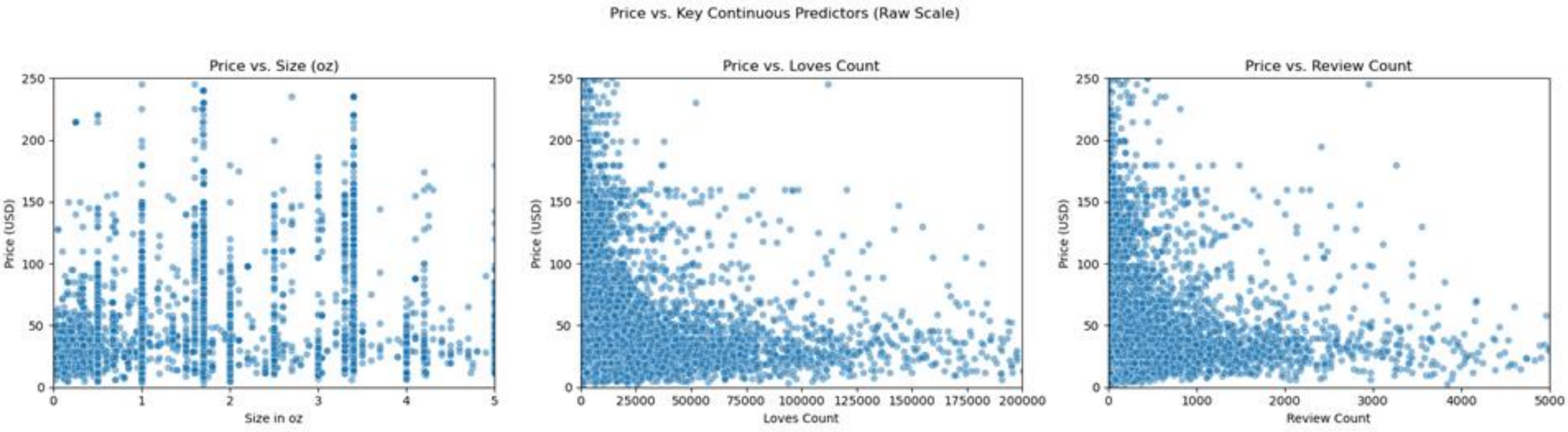
Data Cleaning

- Removed rows missing: **rating**, **reviews**, **size**
- Final modeling dataset: **~6,000 products**

Model Coefficients

Variable	Description
product_id	Product identification number
primary_category	Makeup, Skincare, Hair, Fragrance, Mini size
loves_count	Number of “loves” a product receives
rating	Customer rating score
reviews	Number of reviews a product receives
ingredients	Contents of a product
price_usd	Cost of a product
limited_edition	1 = item is limited edition
online_only	1 = Online-exclusive product
Size_oz	Product size in ounces


Distribution of Product Price (USD)
Count
6000
5000
4000
3000
2000
1000
0
0
250
500
750
1000
1250
1500
1750
Price (USD)



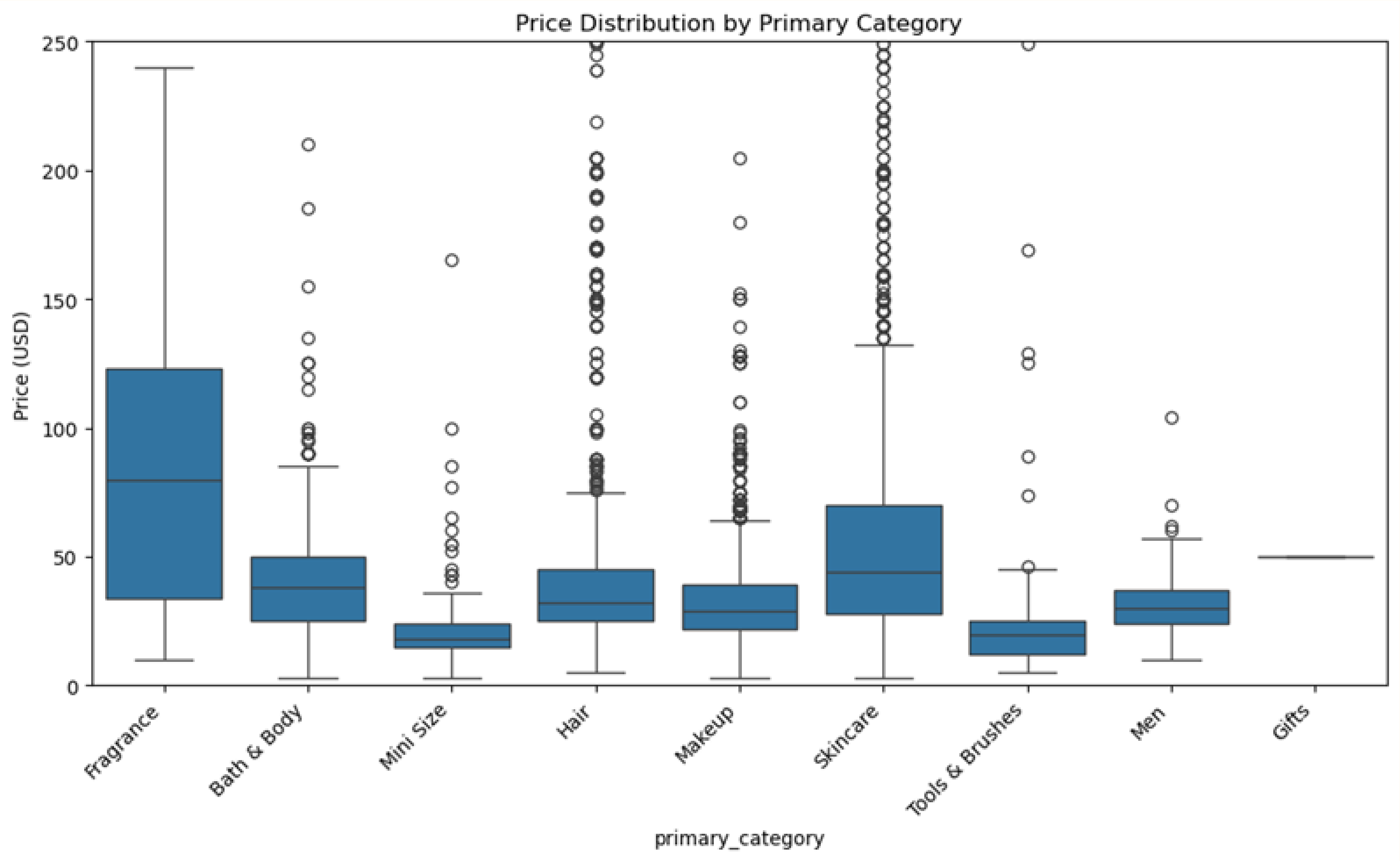
Price vs. Key Continuous Predictors (Raw Scale)

Price vs. Size (oz)

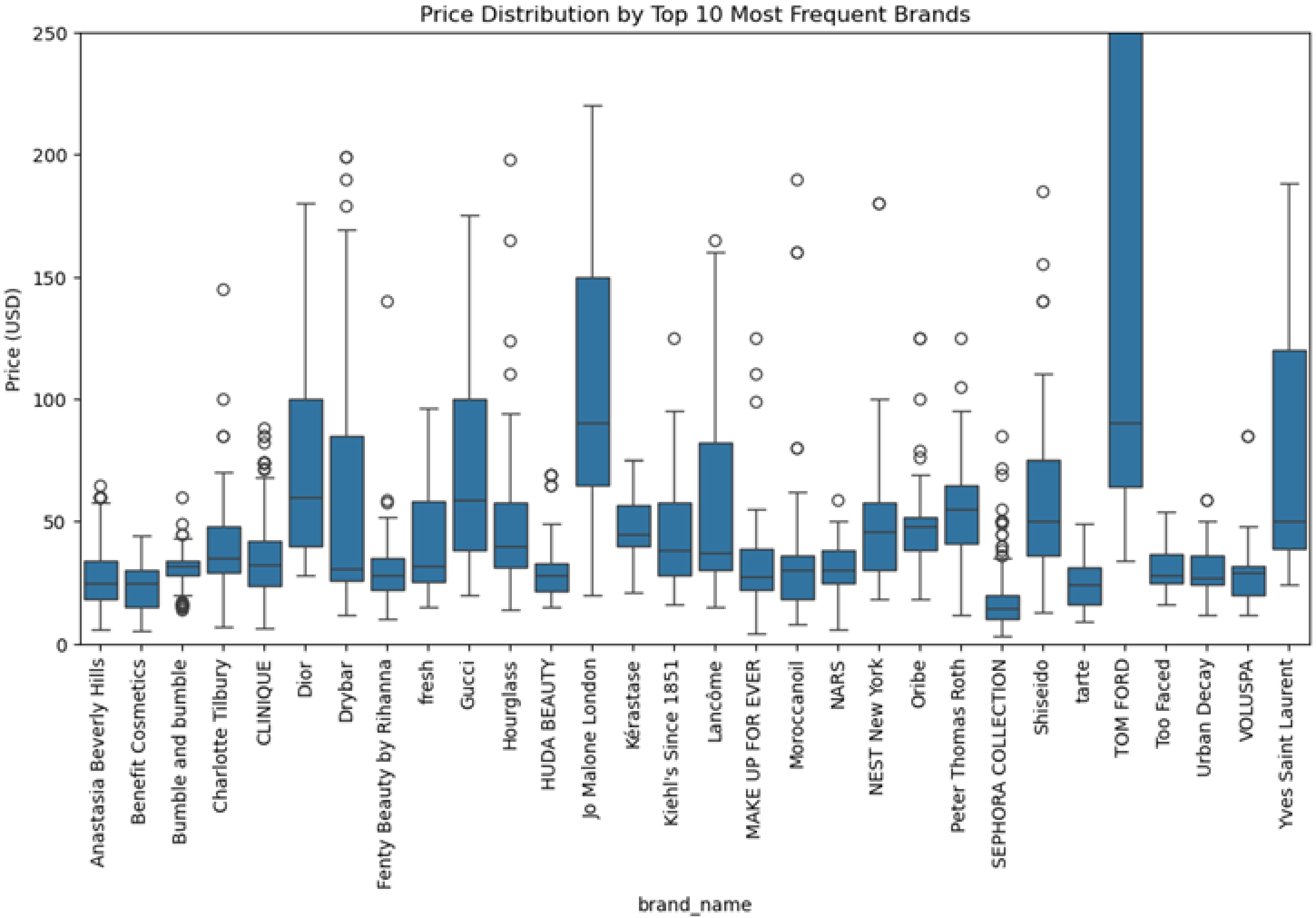
Price vs. Loves Count

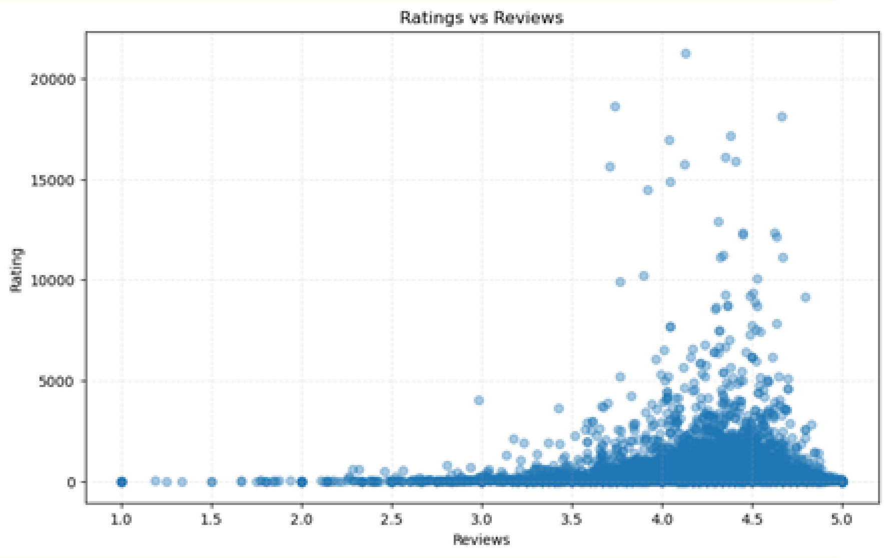
Price vs. Review Count

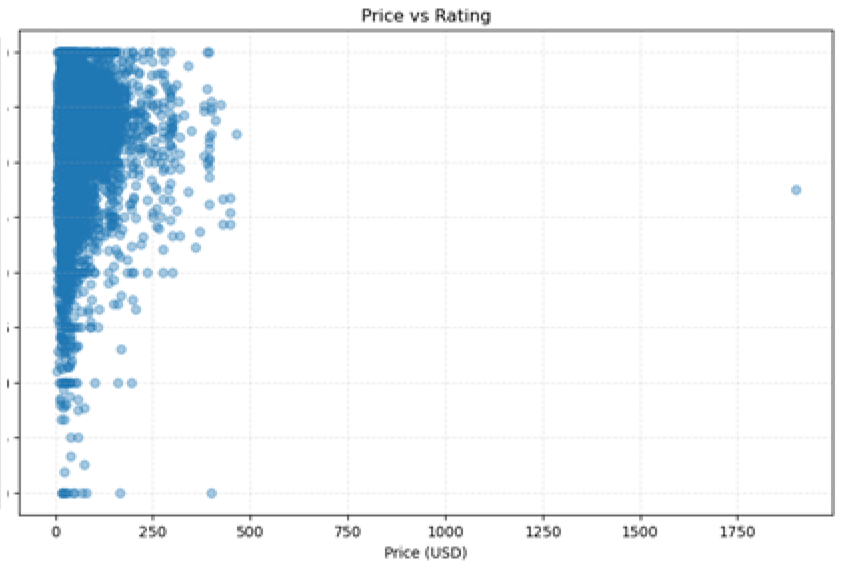
Back to Presentation



Price Distribution by Primary Category
Price (USD)
0
50
100
150
200
250
Fragrance
Bath & Body
Mini Size
Hair
Makeup
Skincare
Tools & Brushes
Men
Gifts
primary_category


Price Distribution by Top 10 Most Frequent Brands
Price (USD)
250
200
150
100
50
0
Anastasia Beverly Hills
Benefit Cosmetics
Bumble and bumble
Charlotte Tilbury
CLINIQUE
Dior
Drybar
Fenty Beauty by Rihanna
fresh
Gucci
Hourglass
HUDA BEAUTY
Jo Malone London
Kérastase
Kiehl's Since 1851
Lancôme
MAKE UP FOR EVER
Moroccanoil
NARS
NEST New York
Oribe
Peter Thomas Roth
SEPHORA COLLECTION
Shiseido
tarte
TOM FORD
Too Faced
Urban Decay
VOLUSPA
Yves Saint Laurent
brand_name


Ratings vs Reviews
Rating
20000
15000
10000
5000
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Reviews


Price vs Rating
0
250
500
750
1000
1250
1500
1750
Price (USD)

Models and Results

Q1: Can we predict the rating of a product?

Model:

- **Target (Y): Continuous product rating.**
- **Features (Xi): Base metrics (loves_count, reviews, price_usd, online_only, size_oz) and One-Hot Encoded primary categories (Hair, Makeup, Mini Size,Skincare)**

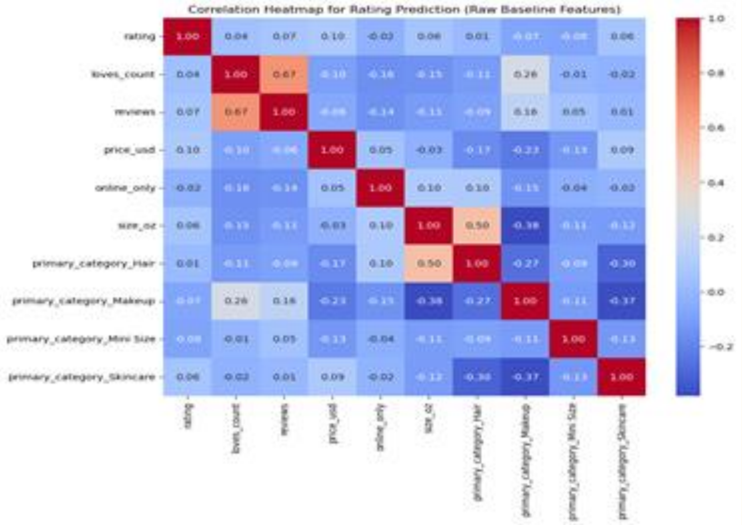
Q1: Can we predict the rating of a product?

Model:

- Target (Y): Continuous product rating.
- Features (Xi): Base metrics (loves_count, reviews, price_usd, online_only, size_oz) and One-Hot Encoded primary categories (Hair, Makeup, Mini Size,Skincare)

Results:

- Mean Absolute Error (MAE): 0.35065513221861994
- Mean Squared Error (MSE): 0.21904120618387735
- Root Mean Squared Error (RMSE): 0.4680183823140682
- R-squared (R2): 0.013241098076260727



Q2: Can we predict the price of a product?

Model:

- **Target (Y): Logarithm of Price (Log(Price_Usd))**
- **Features (Xi): Log(Product Size in oz), Log(Loves Count), Log(Review Count), Rating (Raw Value), Limited Edition (0/1), Online Only (0/1), 35 One-Hot Encoded Dummies (Primary Category, Top 30 Brands)**

Why Log?

- **Log transformation applied to continuous features to handle high skewness and improve linearity. Top 30 Brand Grouping was used to manage complexity. Visually, our data was heavily skewed towards the center. If interested, see appendix.**
- **For discrete values: 100* (e^(coefficient) - 1) = impact**
- **For continuous values: the coefficient is interpreted as an elasticity**

Q2: Can we predict the price of a product?

Model:

- **Target (Y): Logarithm of Price (Log(Price_Usd))**
- **Features (Xi): Log(Product Size in oz), Log(Loves Count), Log(Review Count), Rating (Raw Value), Limited Edition (0/1), Online Only (0/1), 35 One-Hot Encoded Dummies (Primary Category, Top 30 Brands)**

Results:

- **R-squared (R2): 0.4657**
- **RMSE (Log-Scale): 0.5014**
- **MSE: 0.2514**
- **MAE: 0.3785**

Feature	Coefficient	Price Impact	Insight
brand_TOM FORD	+0.968	+163%	Being this brand more than doubles the expected price compared to the baseline brand.
primary_category_ Fragrance	+0.811	+125%	The strongest categorical driver; Fragrance products are 2.25 times the price of the baseline category.
brand_Jo Malone London	+0.615	+85%	A significant jump associated with a luxury brand.
brand_Kérastase	+0.581	+79%	Indicates a strong premium in the Hair category.

Feature	Coefficient	Price Impact	Insight
brand_The Ordinary	-1.211	-70%	Strongest negative factor
brand_The INKEY List	-1.040	-65%	Confirms the enormous, measurable effect of budget-brand positioning.
brand_SEPHORA COLLECTION	-0.865	-58%	The in-house brand is strongly correlated with value pricing.
primary_category_Mini Size	-0.374	-31%	Unsurprisingly, this category consistently reduces the predicted price.

Feature	Coefficient	Interpretation
Rating	+0.1095	A one-point increase in rating slightly increases the price by 11%
Size_oz	+0.0957	Size is essential: A 1% increase in size per ounce leads to an approximate 9.6% increase in price .
Love_count	-0.0169	Popularity is irrelevant: The number of "loves" has almost no measurable effect on price.

Q2: Can we predict the price of a product?

Model:

- **Random Forest attempt**
  - **We used an ensemble model composed of 100 individual decision trees.**
  - **Bagging Method to reduce overfitting**
  - **Each of the 100 trees votes on the final price, and we took the average of those 100 predictions as our final Y value.**
  - **Trains 80% and tested 20%**

Q2: Can we predict the price of a product?

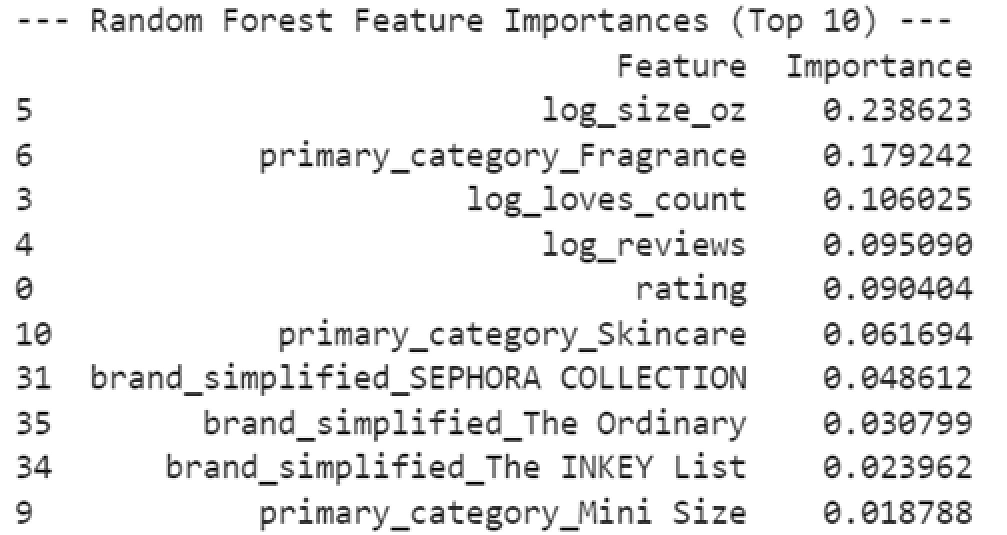
Model:

- **Random Forest attempt**
  - **We used an ensemble model composed of 100 individual decision trees.**
  - **Bagging Method to reduce overfitting**
  - **Each of the 100 trees votes on the final price, and we took the average of those 100 predictions as our final Y value.**
  - **Trains 80% and tested 20%**

Results:

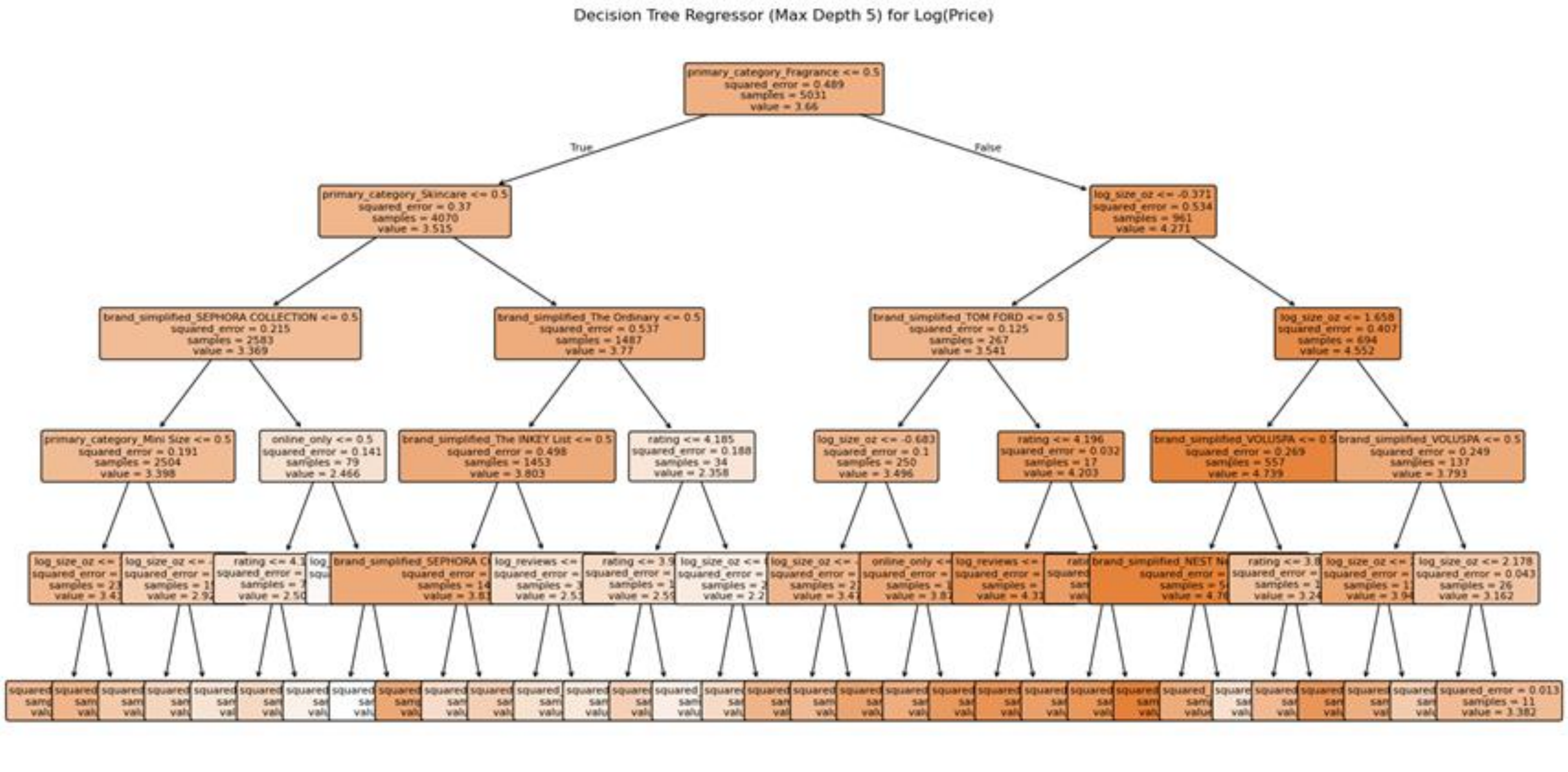
- **R-squared (R2): 0.5775**
- **RMSE (Log-Scale): 0.4459**

Random Forest Results



```
--- Random Forest Feature Importances (Top 10) ---
                                Feature  Importance
5                           log_size_oz    0.238623
6            primary_category_Fragrance    0.179242
3                       log_loves_count    0.106025
4                           log_reviews    0.095090
0                                rating    0.090404
10            primary_category_Skincare    0.061694
31  brand_simplified_SEPHORA COLLECTION    0.048612
35        brand_simplified_The Ordinary    0.030799
34      brand_simplified_The INKEY List    0.023962
9            primary_category_Mini Size    0.018788
```

Random Forest Model



Decision Tree Regressor (Max Depth 5) for Log(Price)

Q3: Can we predict the probability of the price being higher than the median?

Model:

- **Logistic Model**
  - **Y= 1 if the price > $35 and 0 Otherwise**

Q3: Can we predict the probability of the price being higher than the median?

Model:

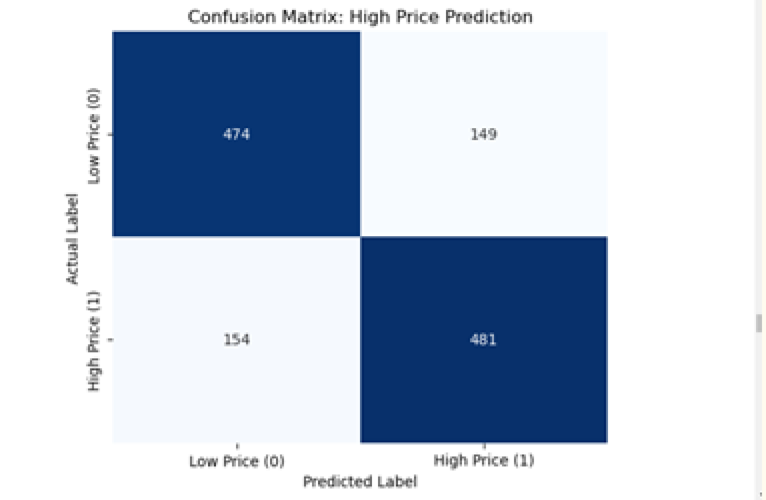
- **Logistic Model**
  - **Y= 1 if the price > $35 and 0 Otherwise**

Results:

- **Accuracy: 0.7591**
- **ROC-AUC Score: 0.8454**

Note:

- **Mean was in the 50s, which is why it was ignored**



Drawbacks and Limitations

- **Linear assumptions** don’t hold because beauty prices are nonlinear and brand-driven.
- **Missing key predictors** (ingredients, product type, packaging, marketing tier).
  - Ingredients are extremely diverse and inconsistent across products, making them difficult to categorize or one-hot encode
- **One-Hot Encoding Assumes Each Category Has a Fixed Average Effect**
  - Category and brand dummies oversimplify reality (e.g., “all fragrance is expensive”)
  - Category-level assumptions wash out nuance
- **Rating Prediction Was Extremely Poor**
  - Ratings have almost **no relationship** to price, size, category, or popularity.
  - Explained by subjective experience and marketing factors

Conclusion

Summary

1. **Why this data set?**
   - Beauty is a $450B industry, and this dataset lets us study what drives product pricing and consumer engagement.
2. **EDA Summary**
   - Fragrance and luxury brands are the most expensive. Minis and hair products are cheaper. Love count and reviews show weak relationships with price and ratings
3. **Model Results**
   - Rating Model had R^2 of 1.3%, Price regression was moderate ($R^2 \approx$ 46%), Random Forest improved prediction (($R^2 \approx$ 0.58), Logistic model classified high vs low price well (AUC ≈ 0.85).
4. **Interpretation**
   - Price is Objective, Rating is Subjective
     1. After it is out the store, is it the sellers problem?
   - Brand Equity is the Ultimate Price Driver
   - The Market Penalizes Size for Certain Products
     1. A high quality perfume vs a larger sized bottle of shampoo are valued differently, and price shows that
5. **Future Work: Focus on one specific brand and collect direct customer feedback (surveys, reviews) to understand how consumers actually perceive product quality**