

SEPHORA Skincare Analysis

The Beauty Wishlist

Which skincare products deserve more attention?

Giovanna C. Villanueva



01. Business Problem

Scenario

Sephora's beauty marketing team wants to understand which skincare product characteristics are associated with higher consumer interest so they can create their next campaign.

Objective

Explore product popularity and compare three machine learning models to estimate the number of favorites for skincare products.

Target Variable

loves_count — the number of times a product has been marked as a favorite.

Business Value

The findings can help the team identify products to investigate for marketing campaigns and visibility/merchandising tests.



02. Dataset Overview

The original file had 8,494 products and 27 variables.

```
# Keep only skincare products
skincare = df[df["primary_category"] == "Skincare"].copy()
```

2,420
products

11
selected
variables

143
brands

Numerical Variables

price_usd, rating, reviews, loves_count

Categorical Variables

brand_name, secondary_category, limited_edition,
sephora_exclusive, online_only, product_id, product_name

Our evidence: 2,420 skincare products

```
# Select the columns used in this project
columns = [
    "product_id",
    "product_name",
    "brand_name",
    "secondary_category",
    "price_usd",
    "rating",
    "reviews",
    "loves_count",
    "limited_edition",
    "sephora_exclusive",
    "online_only"
]

skincare = skincare[columns]
```

03. Data Cleaning

Check	Finding	Decision
Missing rating/reviews	69 each	Training-set medians
Duplicate rows/IDs	0	Keep all rows
Invalid price/count/rating	0	No corrections needed
Extreme values (Outliers)	Price up to \$1,900	Inspect and retain



04. Exploratory Data Analysis | Histogram



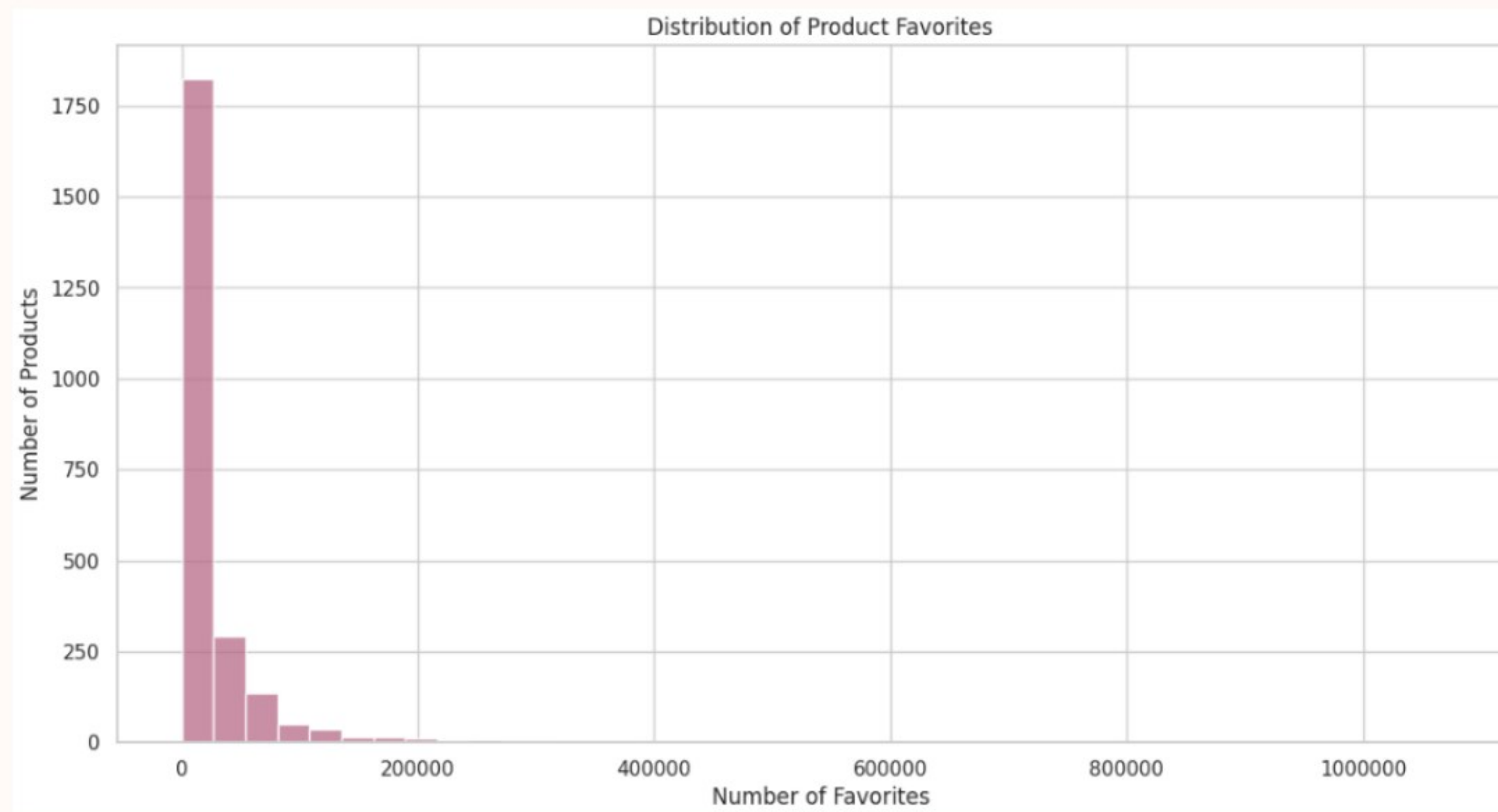
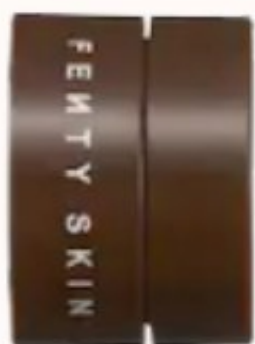
27,111

mean favorites



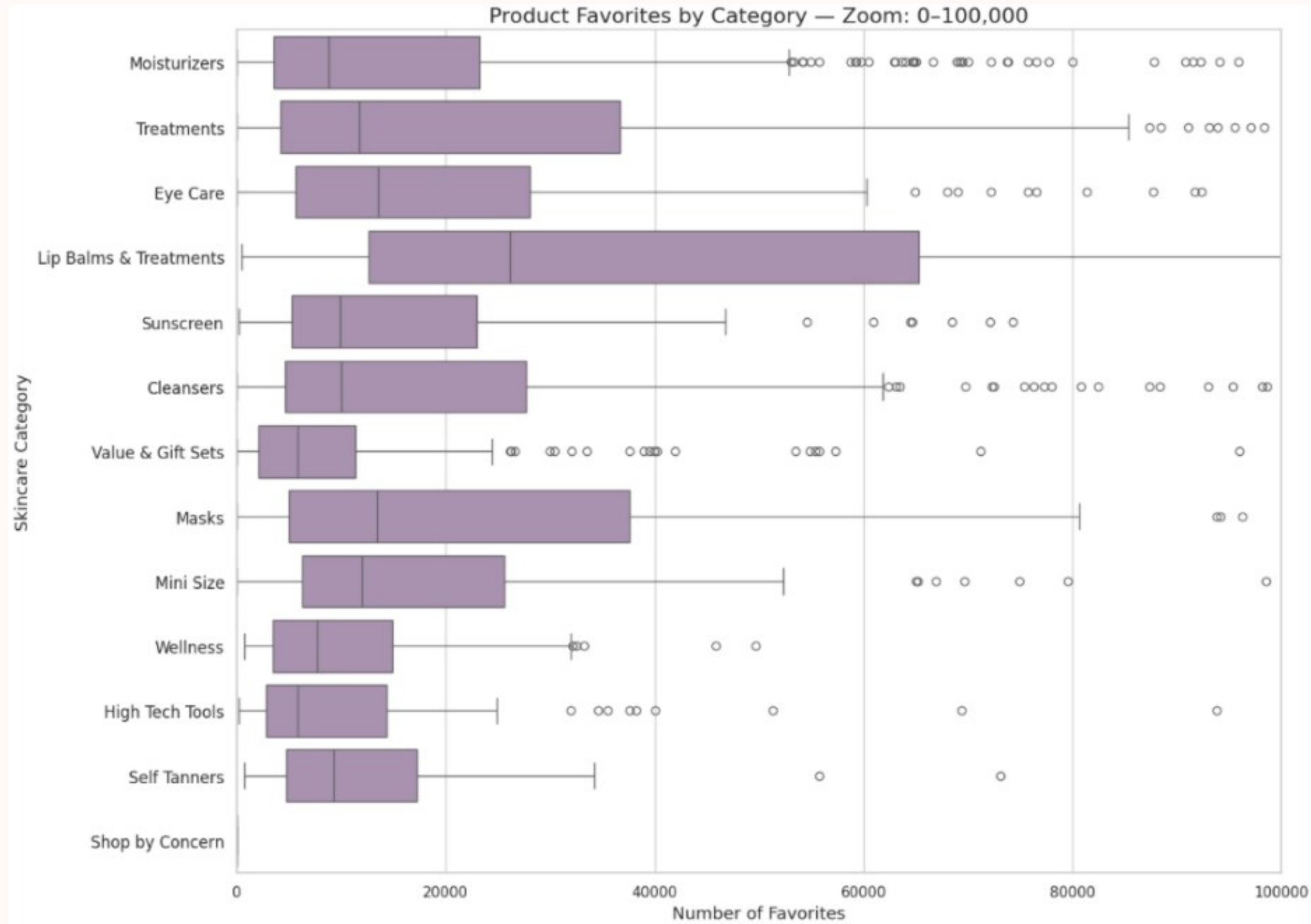
10,074

median favorites



- Most products have relatively few favorites
- The popular products pull the average above the median
- The median better represents a typical product.

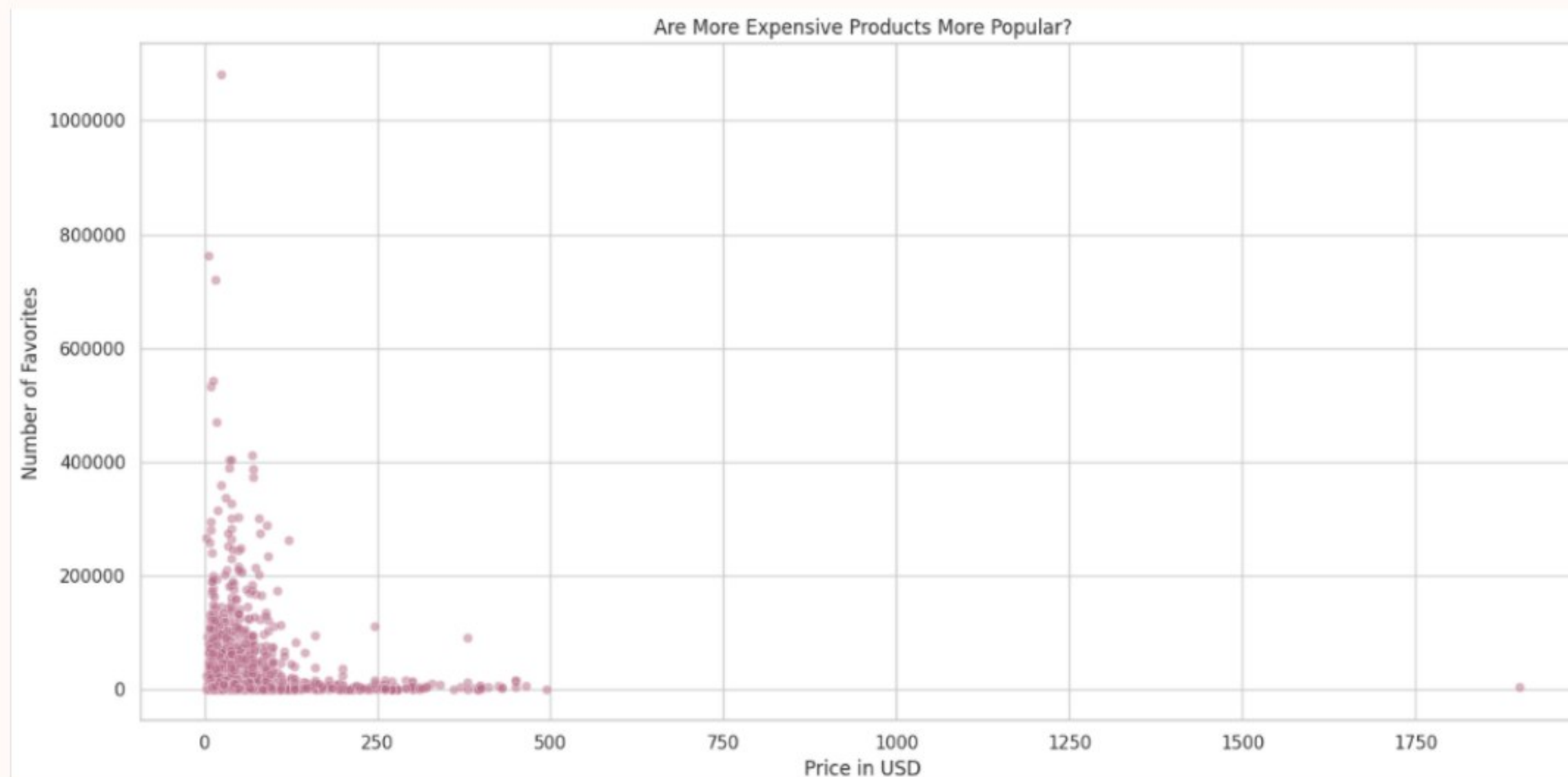
Lip care has the highest median favorites | Boxplot



Lip care: median 26,121 across 61 products.
Moisturizers: 8,788 across 551 products.



Higher prices do not guarantee more favorites | Scatter Plot



-0.111

Pearson correlation

More expensive products do not necessarily have more favorites

Many lower-priced products also have few favorites

Price alone offers limited guidance for choosing products to promote.

Review activity relates most strongly to favorites



Strongest

0.72

Reviews/Favorites

Weakest

-0.15

Online/Favorites

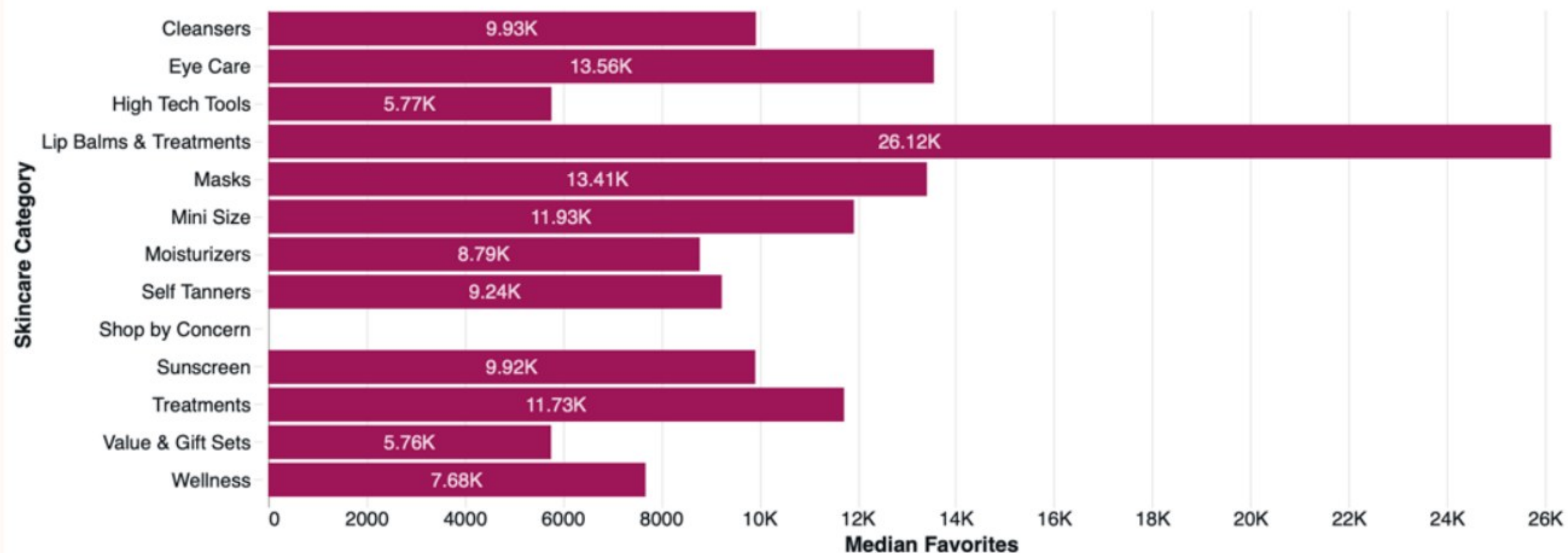
Little to no relationship: Rating/Loves_Count **0.01**



05. Databricks confirms the category pattern

Median Favorites by Skincare Category

Shop by Concern contains only one product, so it should not be used to draw conclusions about the category.



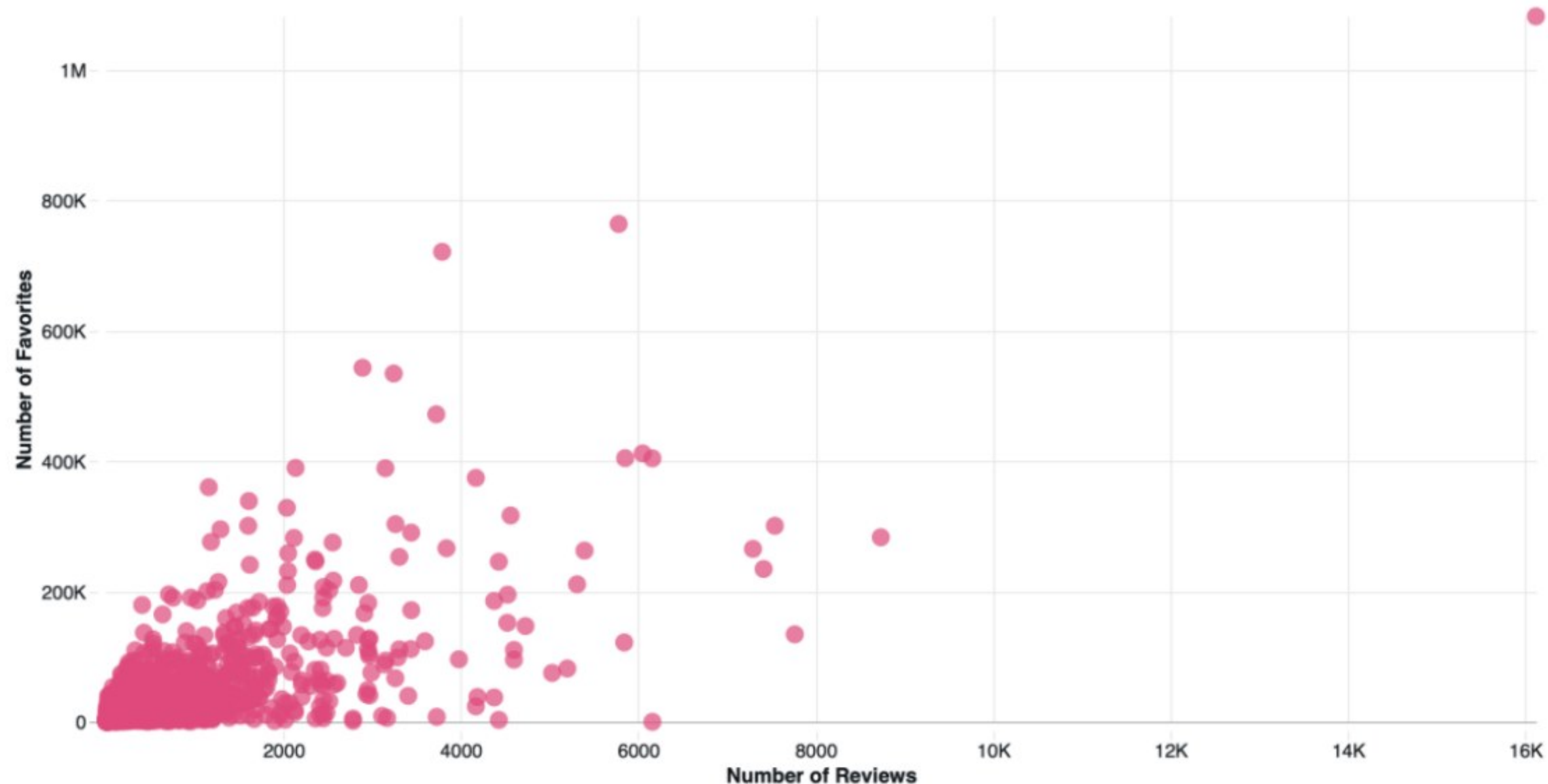
Lip care leads, while shop by concern has only one product, making category conclusions limited.



Databricks | Review Activity and Favorites

Review Count vs. Product Favorites

Products with more reviews tend to have more favorites. However, this does not prove that reviews cause favorites. Products that have been available longer may have had more time to collect both.



Products that have been available longer may have had more time to collect reviews and favorites.



08. Feature Engineering

1,936

training products

484

test products

Training-only preparation

Learn missing-value medians and category encoding from the 80% training split.

One-hot encoding

Convert brand and category to indicators without assigning them an artificial order.

preparing product information for modeling

8 original predictors become 155 numerical columns, and unseen categories receive zeros.

09. Multicollinearity (VIF)

Check	Finding
Category: Self Tanners	4.41
Brand: Sephora Collection	3.70
Brand: CLINIQUE	3.50
Category: Wellness	3.02
Sephora Exclusive	2.92

All values below 5

Overlapping information is not a major concern.

Kept all predictors at this stage.

10. Model Evaluation

Model	R^2	RMSE	MAE
Random Forest	0.62	24,300	12,464
Linear Regression	0.56	25,928	16,111
Decision Tree	0.53	26,722	14,474

R^2 : higher is better

RMSE and MAE: lower is better

Random Forest is our recommended model. Its average absolute error is 12,464 favorites



10. Feature Selection

Random Forest version	Columns	R^2	RMSE	MAE
All Predictors	155	0.615	24,300.145	12,464.281
Without Limited Edition	154	0.615	24,305.822	12,468.488

Tested limited_edition because its correlation with loves_count was weak (-0.06)

The same test set informed this choice.

Independent validation should confirm it.



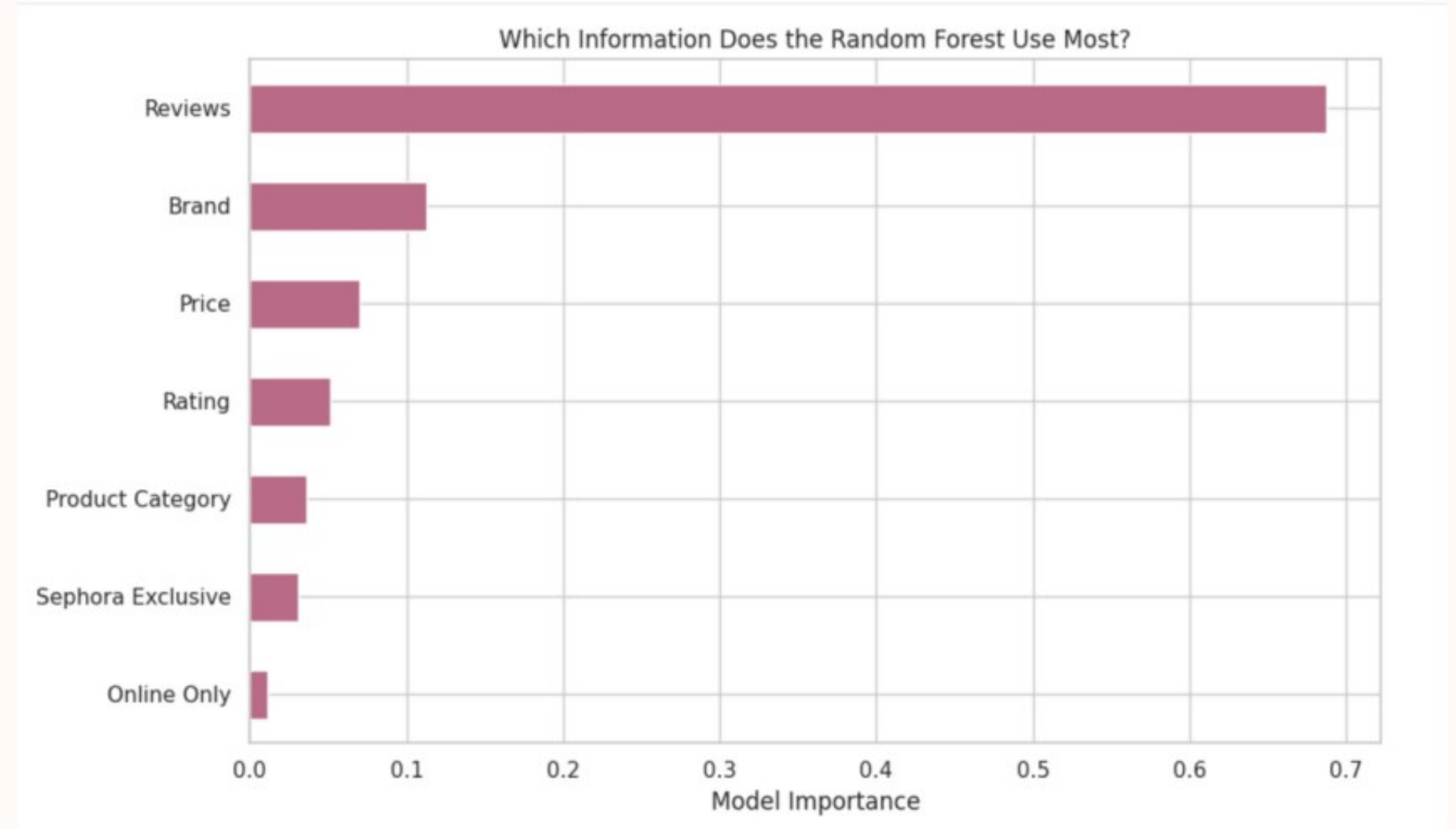
Model Interpretation

Which features the Random Forest used most?

Review Count

Brand

Price



****This does not mean that increasing reviews will cause more favorites.

12. Business Recommendations

Candidate Products

Investigate lip care, as it has the highest median number of favorites. Check stock and suitability.

Supporting Evidence

Use Random Forest estimates alongside brand and customer activity. Avoid price-only decisions.

Campaign Comparison

Compare added product visibility with the current shipping experience through a controlled test

Business Value

Measure purchase conversion and revenue. The dataset cannot estimate the sales lift in advance.



13. Conclusion

What We Learned

Category context and customer activity help explain differences in favorites.

Most Useful Insight

Lip care is a candidate for testing, while review count helps estimate existing-product interest.

Biggest Challenge

Highly skewed counts and many brands make comparisons and modeling harder

Future Improvements

Add product age, exposure and sales.



THANK YOU!

SEPHORA Skincare Analysis

