SEPHORA

Sephora is a visionary beauty-retail concept founded in France by Dominique Mandonnaud in 1970. Sephora's unique, open-sell environment features an ever-increasing amount of classic and emerging brands across a broad range of product categories including skincare, makeup, fragrance, body and hair care, in addition to Sephora's own private label. Today, Sephora is not only the leading chain of perfume and cosmetics stores in France, but also a powerful beauty presence in countries around the world. Owned by LVMH Moët Hennessy Louis Vuitton, the world's leading luxury goods group, Sephora is highly regarded as a beauty trailblazer, thanks to its unparalleled assortment of prestige products, unbiased service from experts, interactive shopping environment, disruptive spirit and constant innovation. Sephora operates approximately 1,900 stores in 29 countries worldwide, with an expanding base of over 200 stores across the Asia Pacific region including Australia, China, Singapore, Malaysia,Thailand, Indonesia & India.

ABOUT SEPHORA:

In 1969, Sephora emerged as a perfume store in France. The brand name is extracted from Sephos- A Greek word that is a term for pretty and Moses's wife name- Zipporah. A luxury group known as Louis Vuitton and Moet Hennesey purchased this brand in 1977. Sephora opened its first shop in 1998 in New York, USA. The business model of Sephora is like a superstore of makeup products. It allows women to buy all the cosmetic products under one roof to complete their entire look. When Sephora began its business, many people considered it a niche brand and avoided their purchase. Soon Sephora started to progress in the market and attained a competitive position in the market. Sephora has become an international beauty and cosmetic retailer and is considered a consumer-centric organization. They welcome and embrace their customers gracefully1. Makeup and cosmetic products shopping are an intensely personal experience. Sephora allows customers to buy their personalized products. It furnishes the chance of personalization to customers as their objective is "beauty is yours to define and ours to celebrate." Every consumer has his own opinions and ideas about their beauty. Sephora tends to provide maximum personalized tools that can be used across all the channels. Their objective is to offer perfect products for women's needs and enhance their customers' experience. It begins with the wants of the consumers. Almost 55 % of customers anticipate makeup brands to present personalized tools and experiences. Smartphones are playing an essential role in the diversity of customers' demands. Consumers continuously switch between real-life and virtual life, anticipating an integrated and casual experience. Makeup brands have an extraordinary chance to make the shopping experience more memorable with their brand.

Enhancing Customer Experience: Unraveling Sephora's Skincare Reviews

In this capstone project, our focus centers on delving deep into Sephora's skincare product reviews, a goldmine of customer sentiments and preferences. Recognizing the paramount importance of user reviews in today's digital marketplace, we aim to:

Analyze Feedback: Understand the nuances of what customers love and what they feel needs improvement.
Predictive Modelling: Develop a predictive model that anticipates product recommendations based on key parameters such as ratings, feedback counts, and product pricing. This would not only spotlight the top-performing products but also help identify areas for potential enhancement.
Personalized Experience: By segmenting user feedback, the project aims to offer insights that can lead to a more personalized shopping experience, where product recommendations are more attuned to individual customer needs.
Strategic Implications: Leverage the insights derived to inform marketing strategies, inventory decisions, and product formulations or features. By aligning offerings with customer feedback, Sephora can ensure higher customer satisfaction.
In essence, by harnessing the power of customer feedback, this project endeavors to transform raw reviews into actionable strategies, ensuring Sephora continues to elevate its customer experience and stays at the forefront of the beauty retail landscape.

Sephora Data Analysis Capstone Project: Detailed Report

Table of Contents:

7. Model Evaluation

- Performance metrics of RandomForest classifier
- Performance metrics of the XGBoost model
- Comparative analysis between the models

8. Comparison to Benchmark

- Setting the benchmark with RandomForest
- Performance improvement with XGBoost

9. Implications

- Potential impact on Sephora's marketing strategies
- Enhancing customer experience and inventory management

10. Limitations

- Potential pitfalls of the XGBoost model
- Scope for further data enrichment

11. Closing Reflections

- Lessons learned from the project
- Potential avenues for future exploration

12. References

- Data sources, tools, and libraries used
- Relevant literature and studies referenced during the project

1.Introduction

Summary of Problem Statement, Data, and Findings:

Problem Statement:

In the competitive landscape of beauty and skincare retailing, Sephora emerges as a leader. Given the potency of customer reviews in shaping purchasing decisions, Sephora aspires to harness this untapped potential. The cornerstone of this project is to construct a predictive model to determine the likelihood of a product being recommended based on pivotal parameters such as ratings, feedback count, and product price.

Data:

The data reservoir for this endeavor was derived from Sephora's extensive review database. This repository is a rich tapestry of attributes encompassing customer ratings, feedback counts, product prices, and insightful customer demographics like skin tone, hair color, and eye color.

Findings:

An initial exploratory dive into the data brought to the fore some riveting insights. Reviews predominantly hailed from individuals with "light" and "medium" skin tones. Furthermore, products with an avalanche of recommendations invariably lead to a surge in sales. Also, echoing the sentiments of many, product authenticity remains a linchpin in bolstering brand trust.

2. Objectives

Objectives of the Sephora Data Analysis Capstone Project

Predictive Analysis:

To develop a predictive model that leverages customer reviews and other pertinent variables such as ratings, feedback count, and product price to accurately forecast if a product will be recommended by a customer.

Data-driven Insights:

To derive actionable insights from the data that can be employed to drive business strategies, particularly in product positioning, marketing, and inventory management.

Enhance Customer Experience:

By understanding the pivotal factors behind product recommendations, the goal is to enhance the overall shopping experience for Sephora customers, ensuring they are presented with products that resonate with their preferences.

Optimize Marketing Efforts:

With the insights gleaned from the model, tailor marketing campaigns more effectively to spotlight products that are more likely to be recommended, thereby increasing the efficacy of promotional efforts.

Strategic Inventory Management:

By predicting which products are likely to be highly recommended, the objective is to make informed decisions regarding inventory stocking, ensuring popular products are always available for customers.

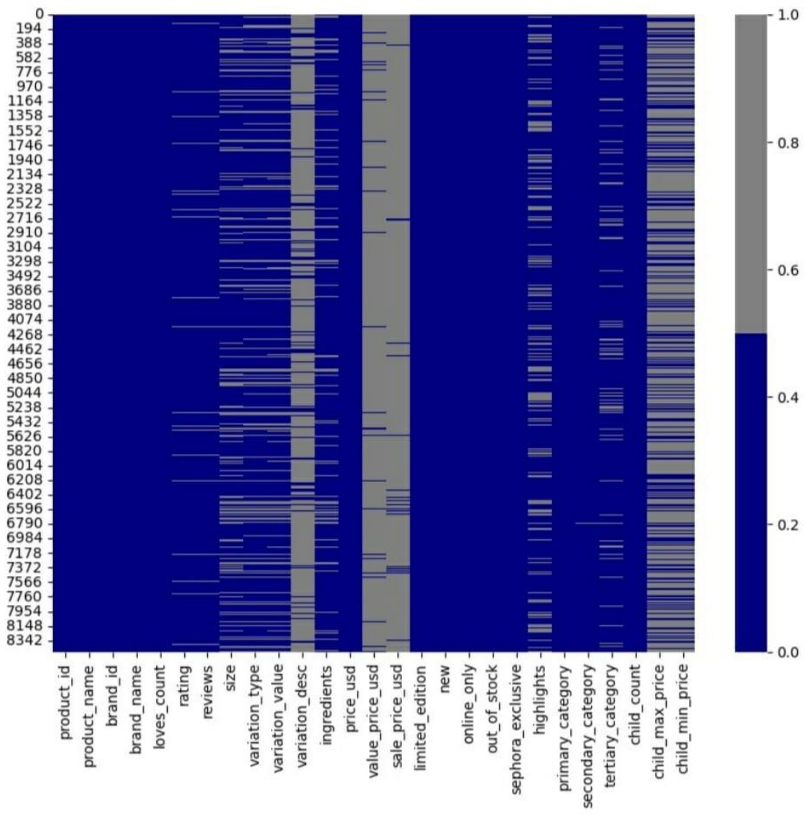
Model Refinement and Evolution:

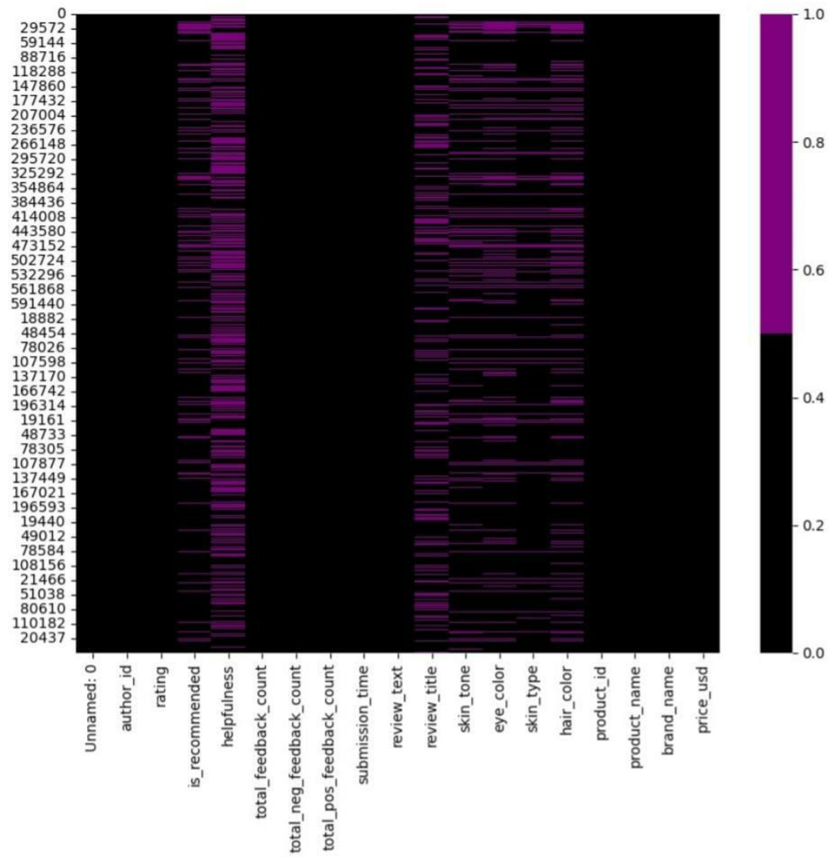
To continually refine the predictive model based on incoming data and feedback, ensuring it stays relevant and effective in an ever-evolving retail landscape.

Business Growth:
Through all the above objectives, the overarching goal is to bolster Sephora's business growth, increasing sales, customer satisfaction, and brand loyalty.

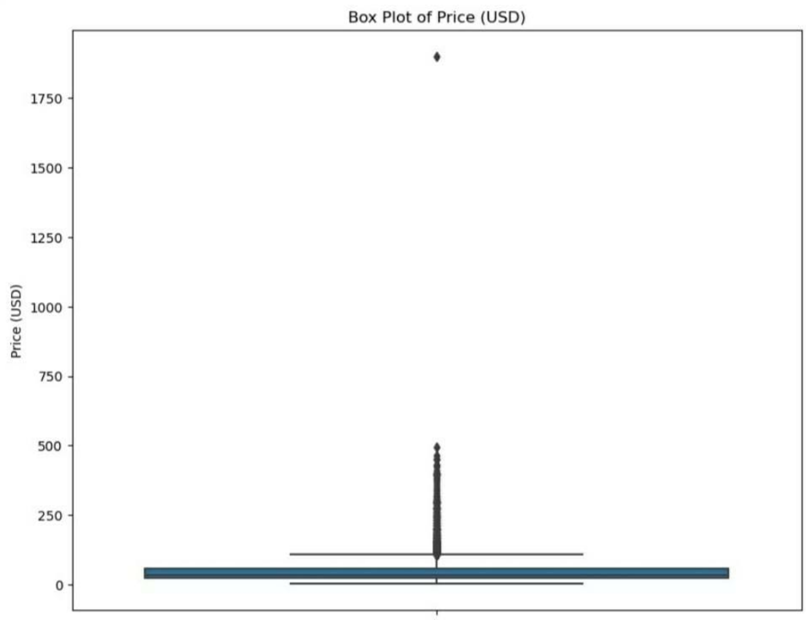
3. Overview of the Final Process:
Charting the course of this project was a meticulously planned strategy that initiated with data preprocessing, transcended through feature selection and model training, and culminated in an exhaustive model evaluation. Python libraries, most notably sklearn and xgboost, were the workhorses driving this analysis.

Data Cleaning:
Missing Values: Initially, we confronted and filled missing data, ensuring a seamless dataset conducive for subsequent analysis.





Outliers: Through thorough analysis, outliers particularly in price variables were spotted and addressed using statistical techniques to reduce skewness.



Data Transformation: Essential features were scaled or normalized, making certain they adhere to similar magnitudes, which further aids accurate model creation.

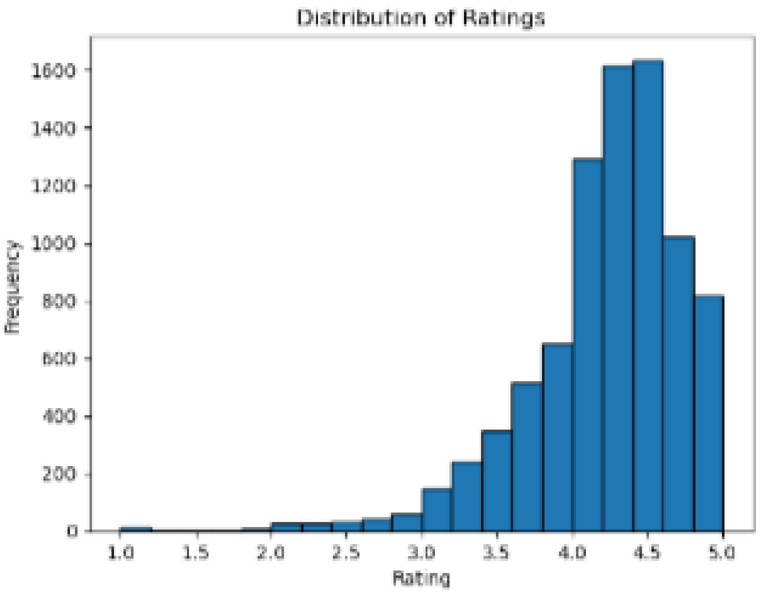
4. Visualization(s)

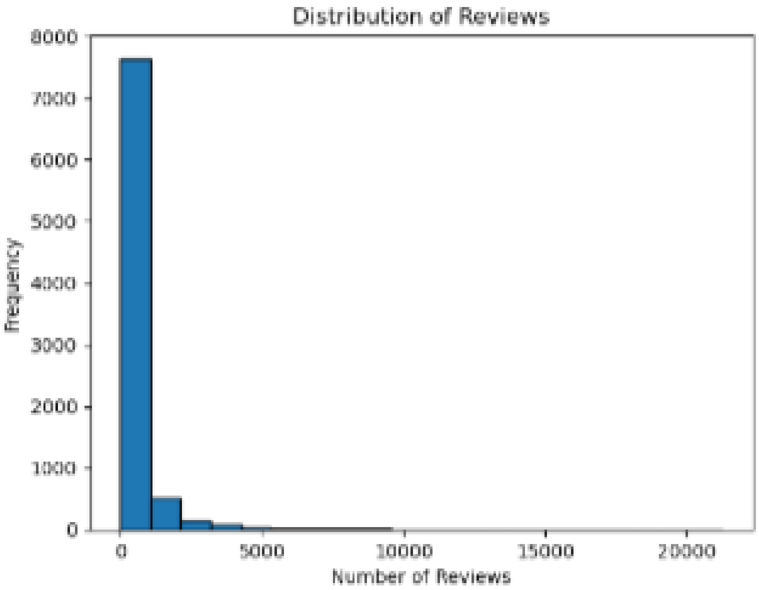
Exploratory Data Analysis (EDA):

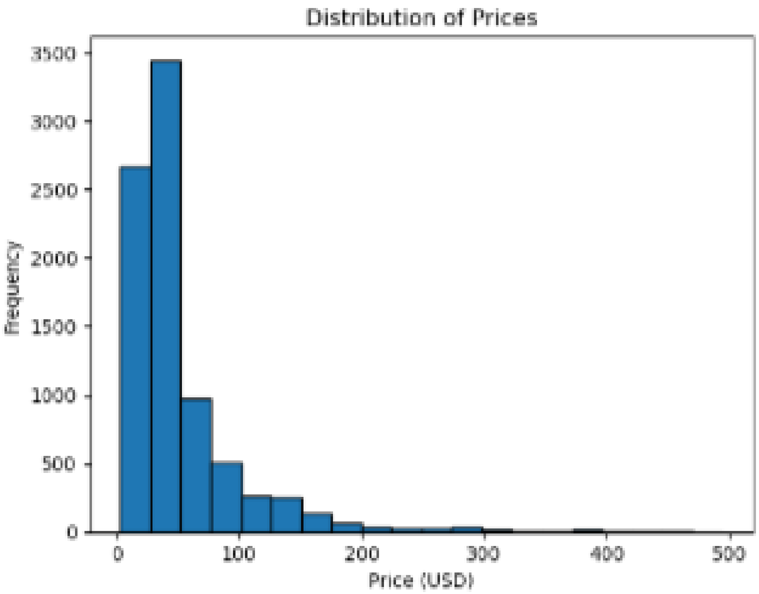
Distribution Plots: Core variables, primarily ratings and feedback count, were meticulously dissected to understand their distribution dynamics.

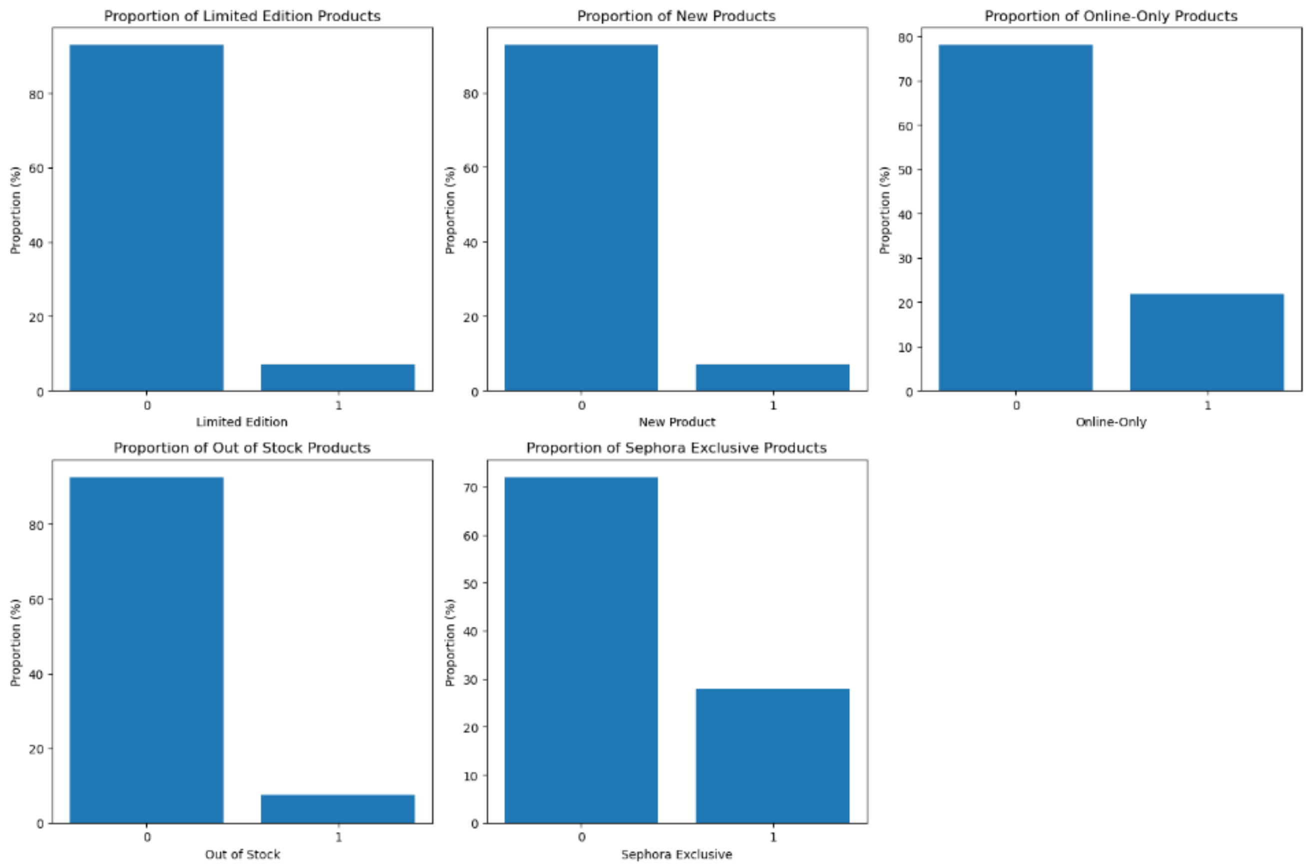
Correlation Analysis: We probed deeper into correlations, unraveling potential interdependencies between numerical attributes.

Visualizations: Vibrant visual tools crystallized key insights, unveiling discernible patterns, and outliers**.**


Distribution of Ratings
Frequency
Rating

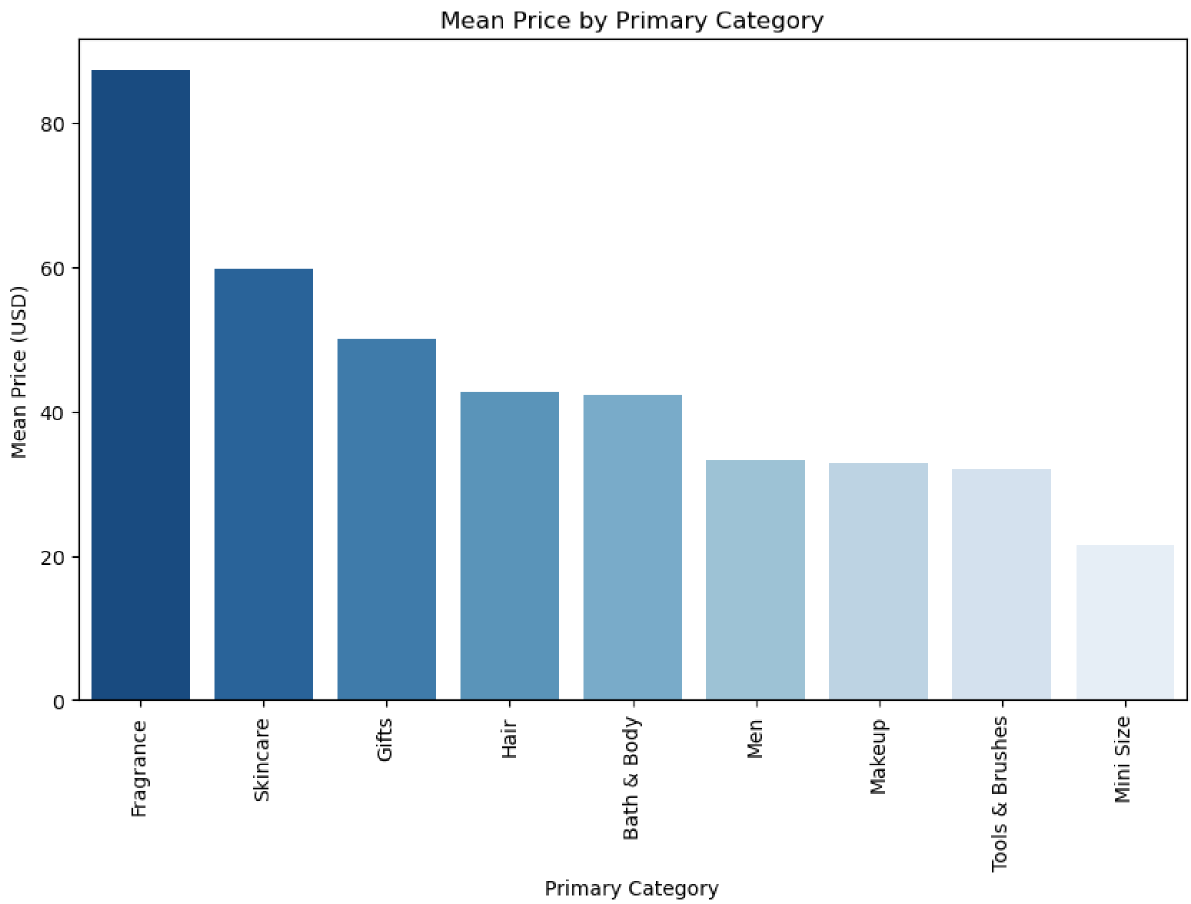

Distribution of Reviews
Frequency
Number of Reviews


Distribution of Prices
Frequency
Price (USD)



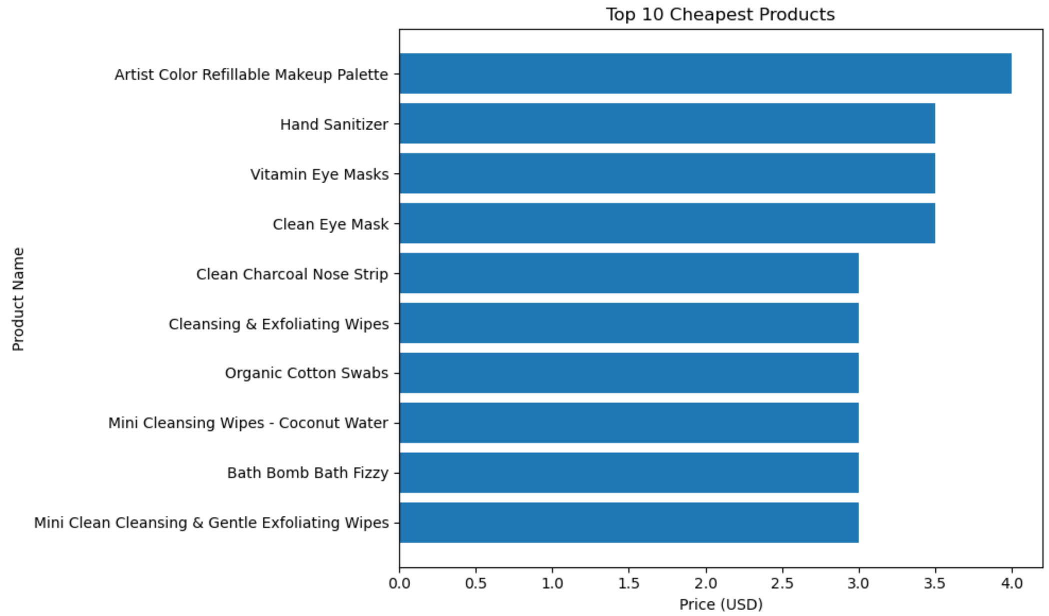
Most products aren't "Limited Edition" or "New".
About 1 in 5 products are "Out of Stock".
Roughly 1 in 3 products are "Sephora Exclusive".
A quarter of the products are sold "Online-Only".

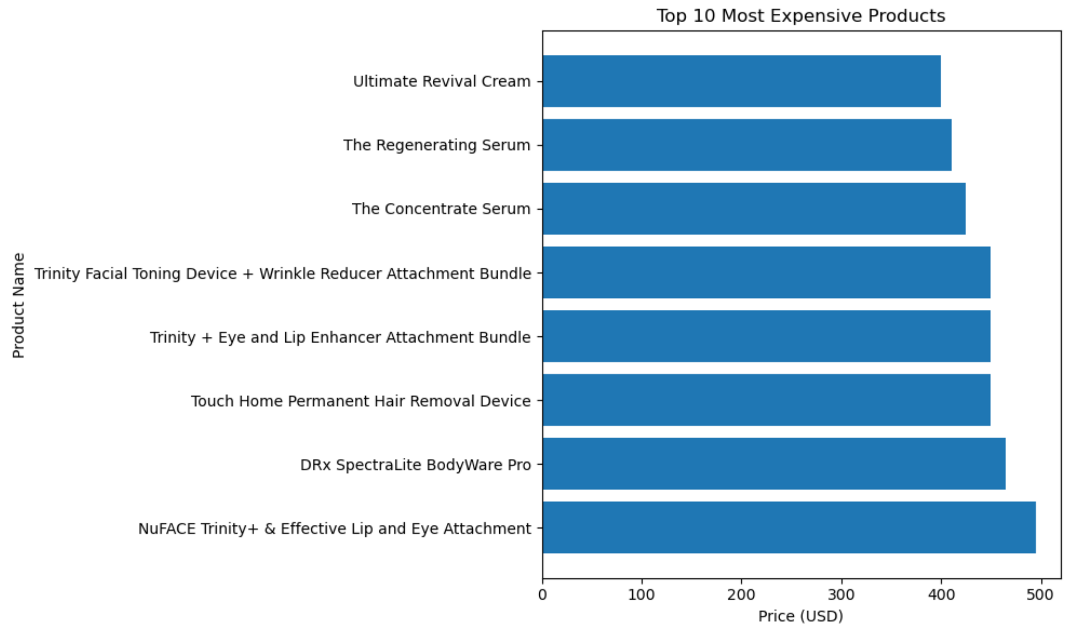
In short, most products are regular items, with a few being exclusive to Sephora or online-only.



Interpretation of the Mean Price by Primary Category Visualization:

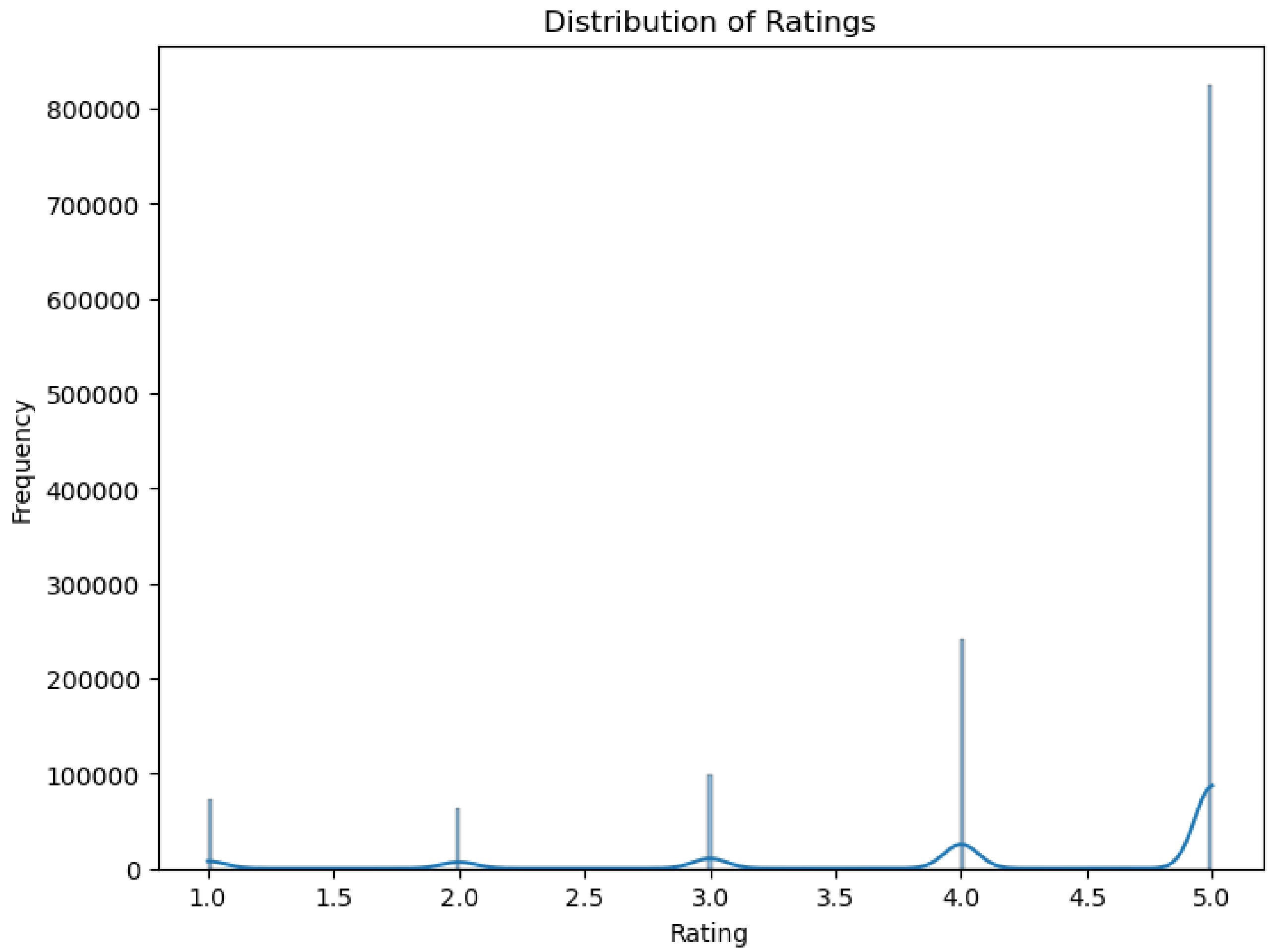
The bar chart represents the average prices of various product categories at Sephora. The standout observation is that "Fragrance" commands the highest average price, reflecting its premium positioning. In contrast, "Mini Size" items, likely smaller or travel-sized versions of products, have the lowest average cost. The middle ground is occupied by categories like "Skincare" and "Gifts", indicating a moderate pricing strategy for these items.



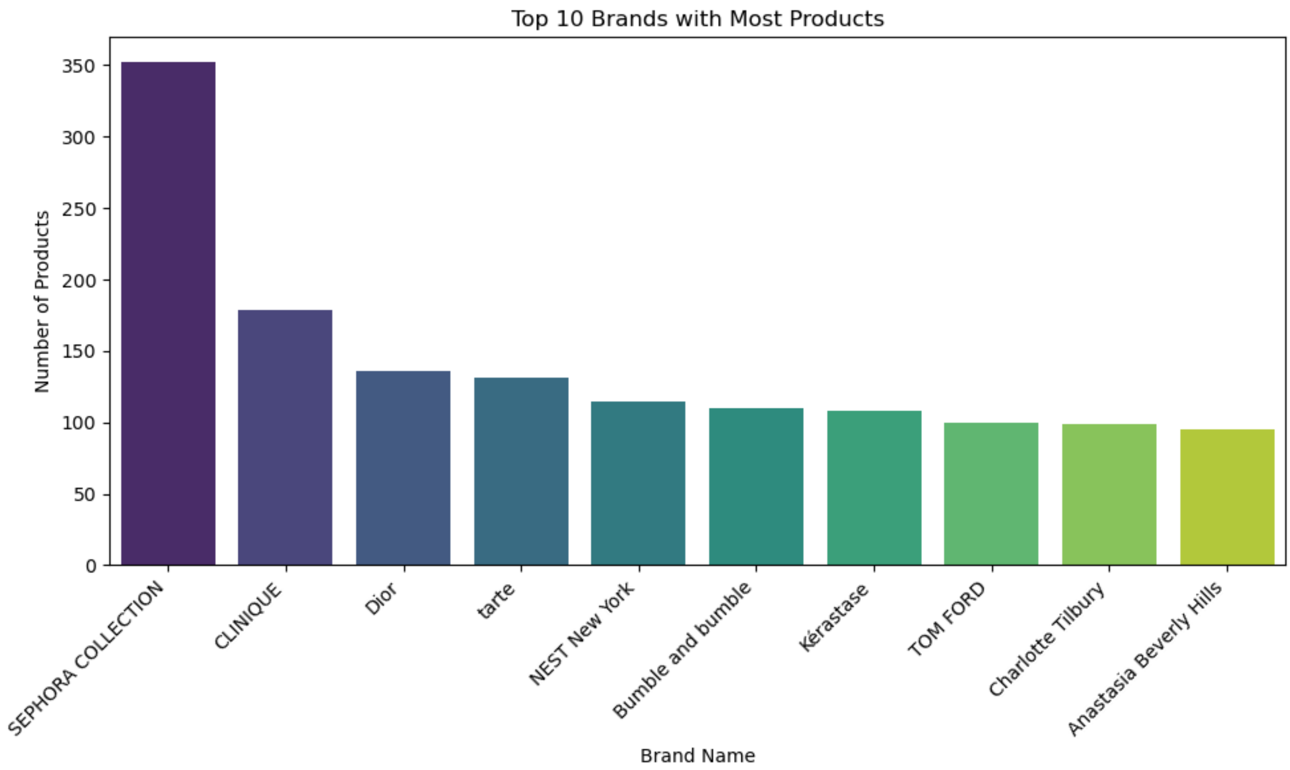


The graphs display the top 10 cheapest and most expensive products at Sephora.

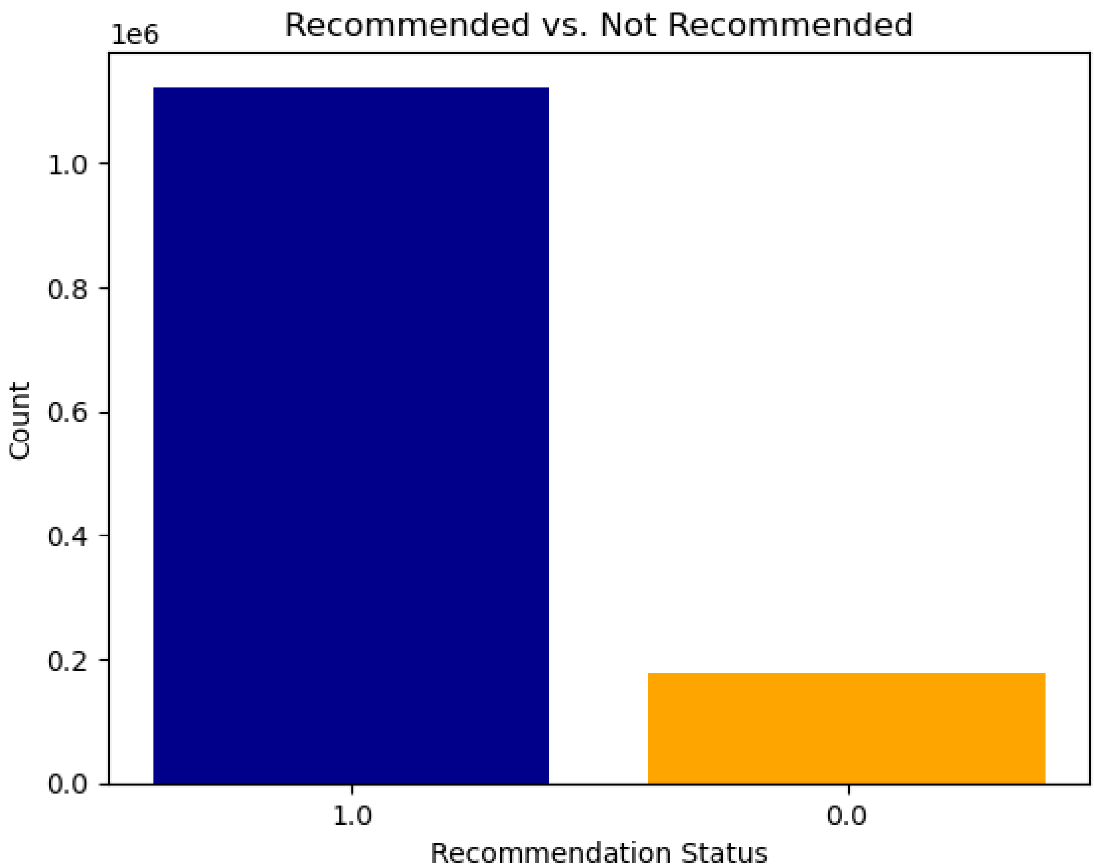
Cheapest Products: Items like "Hand Sanitizer" and "Clean Eye Mask" are among the least expensive, with prices under $4. Most Expensive Products: Products like "Ultimate Revival Cream" and "The Regenerating Serum" are among the priciest, costing several hundred dollars each. In essence, Sephora offers a wide price range, from very affordable items to premium-priced products.



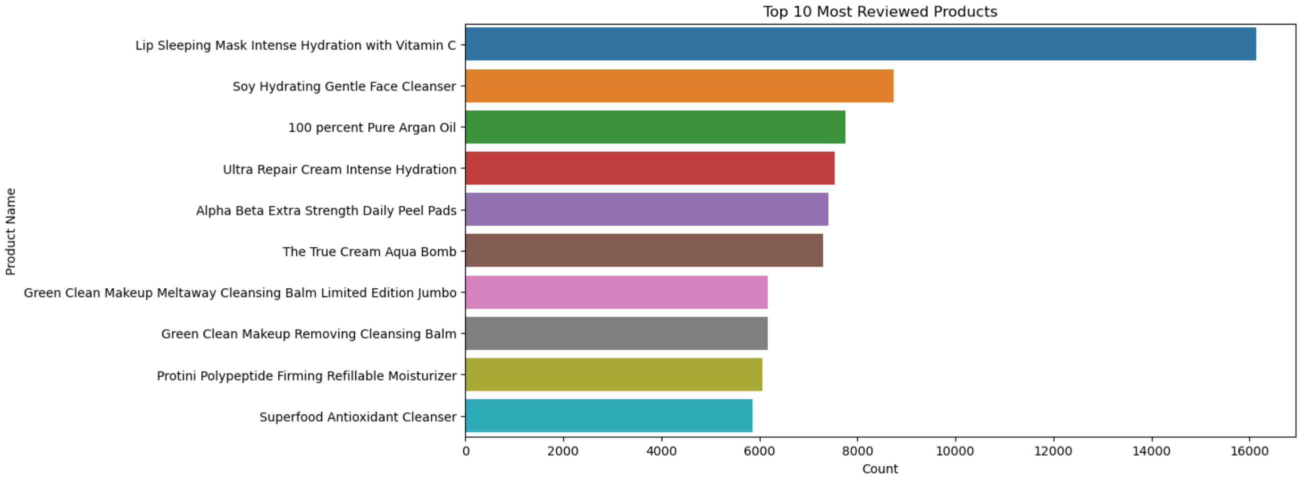
The graph showcases the distribution of product ratings:
A very small number of products received ratings around 1.0 to 3.0. The majority of products received the highest rating of 5.0. In essence, most products at Sephora have been highly rated by customers.



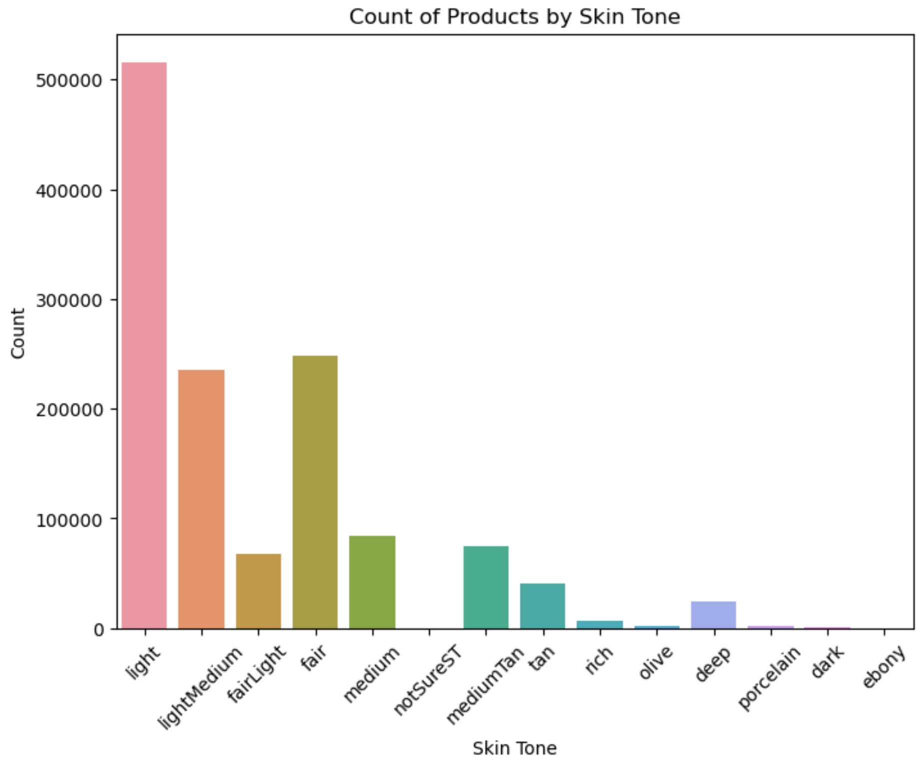
The chart showcases the top 10 brands with the most products on Sephora.
SEPHORA COLLECTION has the highest number, over 300 products. CLINIQUE comes second with around 250 products. Brands like Dior, tarte, and NEST New York have between 100 to 200 products. The remaining brands, including TOM FORD and Charlotte Tilbury, offer 50 to 100 products each.
In essence, SEPHORA COLLECTION dominates in product variety, but several other brands also have a strong presence.



A vast majority of products have been recommended, as indicated by the tall blue bar. This suggests that most users or reviewers had a positive experience with the products and found them satisfactory.
A significantly smaller number of products were not recommended, as shown by the shorter orange bar. This indicates that a lesser number of users had reservations or were not entirely satisfied with these products.
In summary, the visualization underscores a predominantly positive reception for products, with a small fraction not meeting the users' expectations or requirements.



The graph displays the top 10 products with the highest number of reviews. The "Lip Sleeping Mask Intense Hydration with Vitamin C" has the most reviews, followed by other products in decreasing order.



The graph illustrates the count of products based on different skin tones. "Light" skin tone products have the highest count, followed by "light/medium" and "medium" skin tones. Other skin tones have comparatively fewer products.

5. Business Problems
Statistical Tests:
Test for Collinearity:

```
Correlation Matrix:
                           rating  total_feedback_count  \
rating                   1.000000             -0.080300
total_feedback_count    -0.080300              1.000000
total_neg_feedback_count -0.182179             0.674619
total_pos_feedback_count -0.049147             0.984976
price_usd               -0.002616              0.008143

                         total_neg_feedback_count  total_pos_feedback_count  \
rating                                  -0.182179                 -0.049147
total_feedback_count                     0.674619                  0.984976
total_neg_feedback_count                 1.000000                  0.537009
total_pos_feedback_count                 0.537009                  1.000000
price_usd                                0.007682                  0.007508

                         price_usd
rating                   -0.002616
total_feedback_count      0.008143
total_neg_feedback_count  0.007682
total_pos_feedback_count  0.007508
price_usd                 1.000000
```

- Products with higher ratings tend to have fewer negative feedbacks.
- Products with more total feedback also have more positive and negative feedback.
- Price doesn't show a clear relationship with ratings or feedback counts.
- Basically, well-rated products get fewer negative comments, and popular products (with lots of feedback) have both more likes and dislikes. Product prices don’t seem to affect ratings or the number of comments much.

Chi-Square Analysis of Sephora's Product Data:

1. Association between Recommendation Status and Skin Type

Objective:

The objective of this analysis was to discern any potential relationship between the recommendation status of a product ('is_recommended') and the skin type of the users ('skin_type').

Methodology:

A Chi-Square test was employed using a contingency table constructed from the dataset.

Results:

- Chi2-statistic: 1923.86
- P-value: 0.0

Interpretation:

The results, backed by a Chi2-statistic value of 1923.86 and a p-value of 0.0, are indicative of statistical significance. This corroborates the hypothesis that there exists a potent association between the recommendation status of a product and the user's skin type. In essence, the propensity for a product to be recommended might oscillate based on the skin type of the individual.

2. Chi-Square Test for Independence between Categorical Predictors and Recommendation Status

Objective:

The goal here was to delve deeper and ascertain if specific categorical variables, such as 'rating', 'skin_tone', 'eye_color', 'skin_type', and 'hair_color', exhibited any relationship with the recommendation status ('is_recommended').

Methodology:
A series of Chi-Square tests for independence were performed, each juxtaposing one of the categorical predictors against the 'is_recommended' variable.

Results:

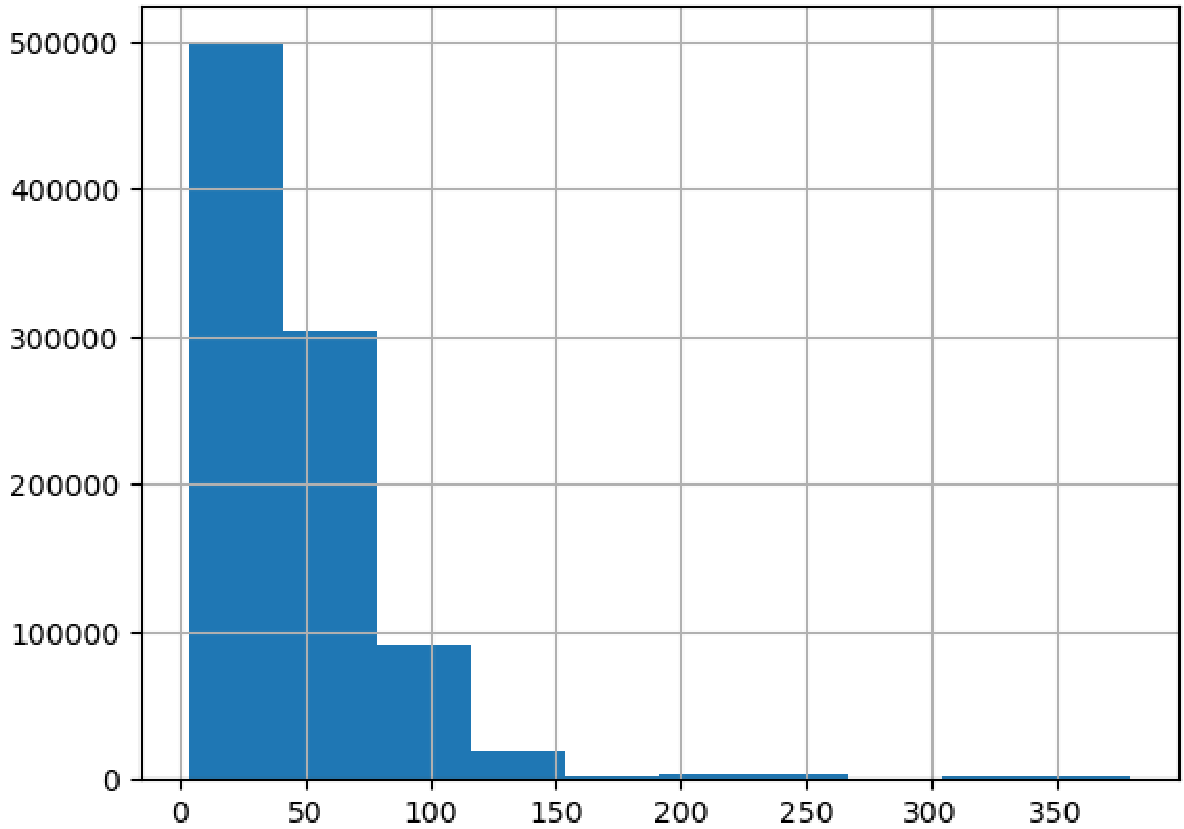
- P-value for rating vs. is_recommended:0.0
- P-value for skin_tone vs. is_recommended: 0.0
- P-value for eye_color vs. is_recommended:0.0
- P-value for skin_type vs. is_recommended:0.0
- P-value for hair_color vs. is_recommended: 0.0

Interpretation:
The uniformly consistent p-value of 0.0 across all the tests underscores the statistical significance of the results. This infers that each of the aforementioned predictors (rating, skin_tone, eye_color, skin_type, hair_color) inherently possesses a relationship with the likelihood of a product being recommended. This association is genuine and not mere serendipity.

For instance, a product with a stellar rating is invariably poised to be recommended more frequently. Analogously, individuals with divergent skin tones, eye colors, or hair colors might harbor distinct preferences, which in turn could influence product recommendations.

Collectively, this series of Chi-Square tests bolster the premise that a mosaic of factors converge to dictate whether a product earns a recommendation, accentuating the multi-faceted nature of consumer choices in the beauty and skincare domain.

Descriptive Statistics and Visualization:



The histogram shows how product prices are spread out.
Most products, around 86.5%, are recommended by customers, while the rest are not.
In short, the majority of products get a thumbs-up from customers, and the histogram shows their pricing distribution.

T-Test:

T-Test Analysis on Sephora's Product Prices and Recommendation Status

Objective:

The core aim of this analytical endeavor was to unravel whether there existed any discernible difference in the prices of products based on their recommendation status – specifically, distinguishing between products that were recommended and those that weren't.

Methodology:

A two-sample independent t-test was deployed for this investigation. The dataset was dichotomized into two distinct cohorts:
1. Prices of products that were recommended
2. Prices of products that were not recommended

Results:

- T-statistic: -2.9337
- P-value: 0.0033

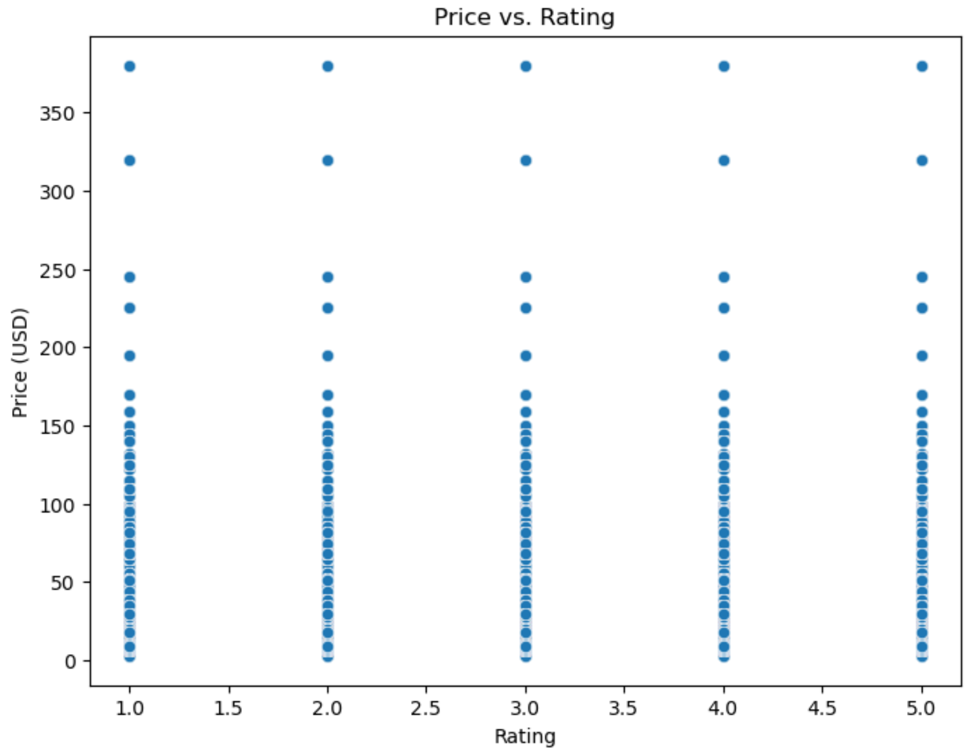
Interpretation:

The t-statistic's negative value of -2.9337 is indicative of the average price of non-recommended products being statistically higher than their recommended counterparts.
Further substantiating this conclusion is the p-value of 0.0033, which falls below the conventional threshold of 0.05, suggesting statistical significance. This reaffirms that the discerned difference in product prices based on recommendation status isn't a mere artifact of randomness but is rooted in actual pricing trends.

In Summary:

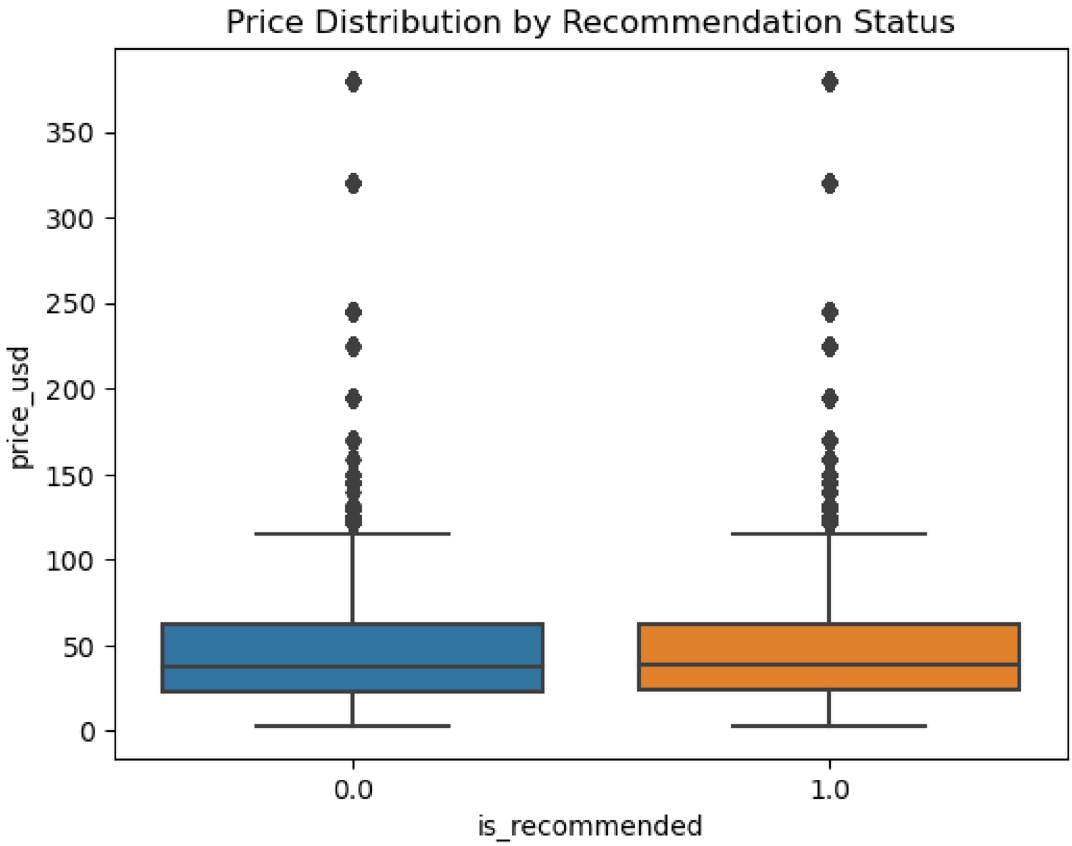
Products that aren't recommended at Sephora tend to, on average, be priced higher than those that receive recommendations. This insight offers a unique perspective on consumer behavior and product pricing dynamics within the beauty and skincare sector.

Relationship between product prices (in USD) and their ratings



The graph shows the relationship between product prices (in USD) and their ratings.Most products, regardless of their price, have a high rating (around 4 to 5).There are some products with low ratings (1 to 2), but they are fewer in number.Products with a wide range of prices can have similar high ratings.Overall, there's no clear trend that higher-priced products have higher or lower ratings.

Business Problem 1:
Is there a significant difference in average product prices between products that are recommended and those that are not?



INTERPRETATION:
From the box plot:

1. **Central Tendency**: The median (the line inside the box) price for products that are not recommended (is_recommended = 0.0) seems to be slightly higher than the median price for products that are recommended (is_recommended = 1.0).
2. **Spread**: Both categories have a similar interquartile range (IQR; the height of the box), which indicates that the middle 50% of prices for both categories span a similar range.
3. **Outliers**: There are several outliers in both categories. This means there are products with prices significantly higher than the majority. The recommended products seem to have a slightly higher density of outliers than the non-recommended ones.

4. **Range**: The range (from the lowest data point to the highest, excluding outliers) for recommended products appears slightly smaller than that for non-recommended products, although this might be influenced by the outliers.

Analysis:

1. While there seems to be a slight difference in the median price between recommended and non-recommended products, the difference does not appear substantial.

2. The presence of numerous outliers, especially in the recommended category, suggests that even though a product is priced much higher than others, it might still get recommended. This could imply that certain high-priced products offer value or quality that justifies their price.

Solution to the Business Problem:

the average price of a product does not seem to be a significant factor in determining whether it gets recommended or not. However, some nuances are worth noting:

1. **Product Value**: It would be beneficial for the business to investigate the high-priced outliers, especially in the recommended category. Understanding what makes these high-priced products worthy of recommendation could yield insights into what customers valu
2. **Quality over Price**: Ensure that product quality matches or exceeds its price point. A product being recommended despite its high price suggests that customers might be considering factors other than just cost when recommending a product. This underscores the importance of maintaining product quality, features, or brand reputation.

3. **Market Segmentation**: The presence of recommended products across a wide price range suggests a diverse customer

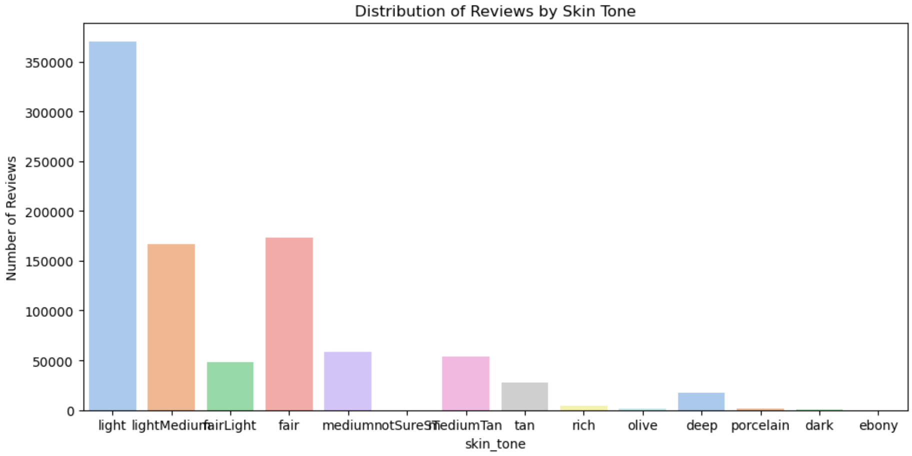
base with varied purchasing power and preferences. Tailor marketing strategies accordingly. For instance, premium products can be targeted towards segments that value luxury or quality, while more affordable products can be marketed to budget-conscious segments.

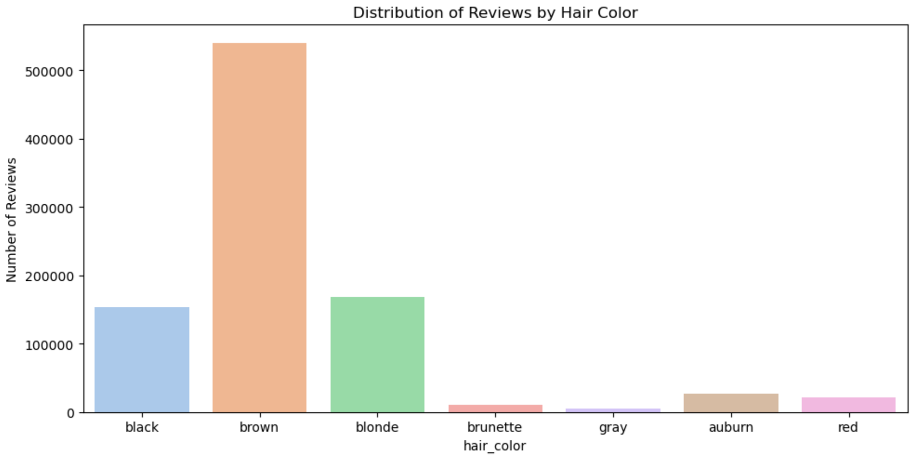
4. **Further Analysis**: To statistically determine if there's a significant difference in average product prices between the two categories, you might consider performing a hypothesis test (e.g., a t-test) to compare the means of the two groups.

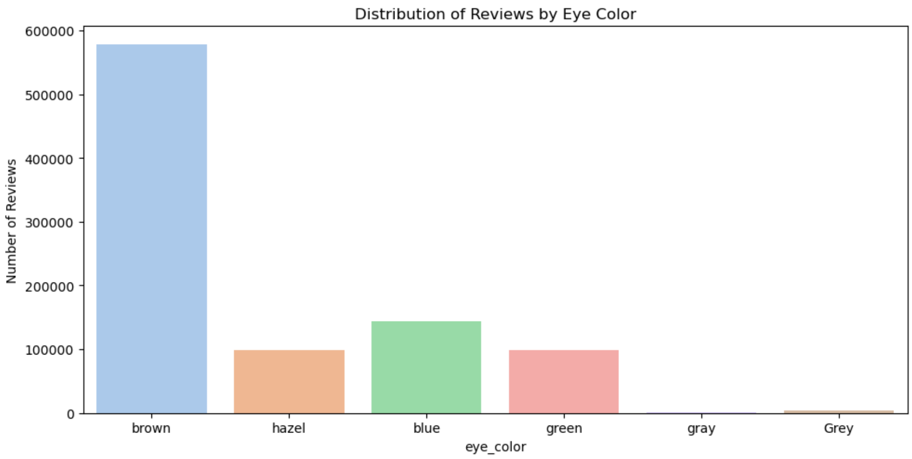
In conclusion, while price is a factor, it's not the only determinant for a product's recommendation status. The business should focus on understanding the inherent value each product provides and tailor its strategies accordingly.

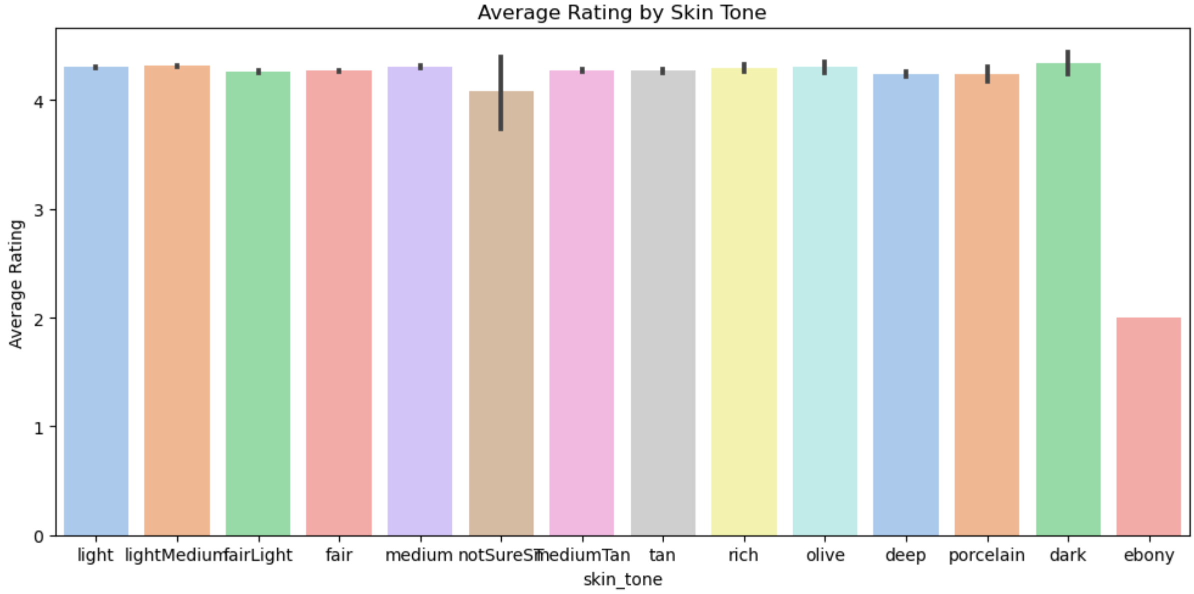
Business Problem 2

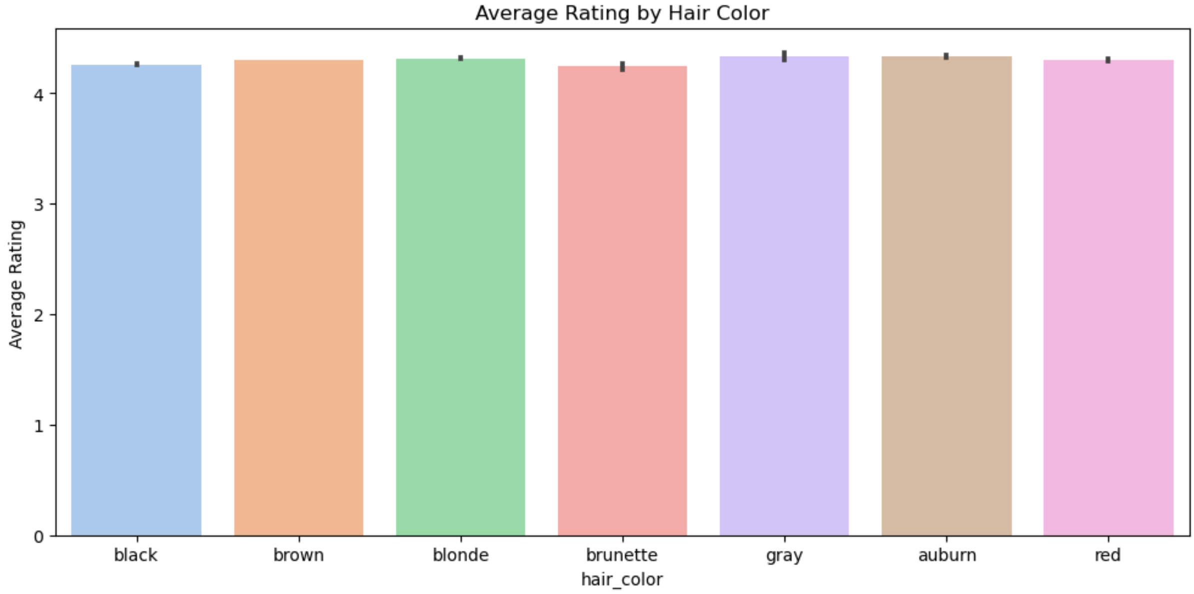
"To launch a new marketing campaign, we want to identify which customer segments (based on features like skin tone, hair color, eye color) might be underrepresented or overly dissatisfied in our reviews. Identifying these groups will allow us to tailor our campaigns to these segments, ensuring they feel represented and addressing their specific concerns."

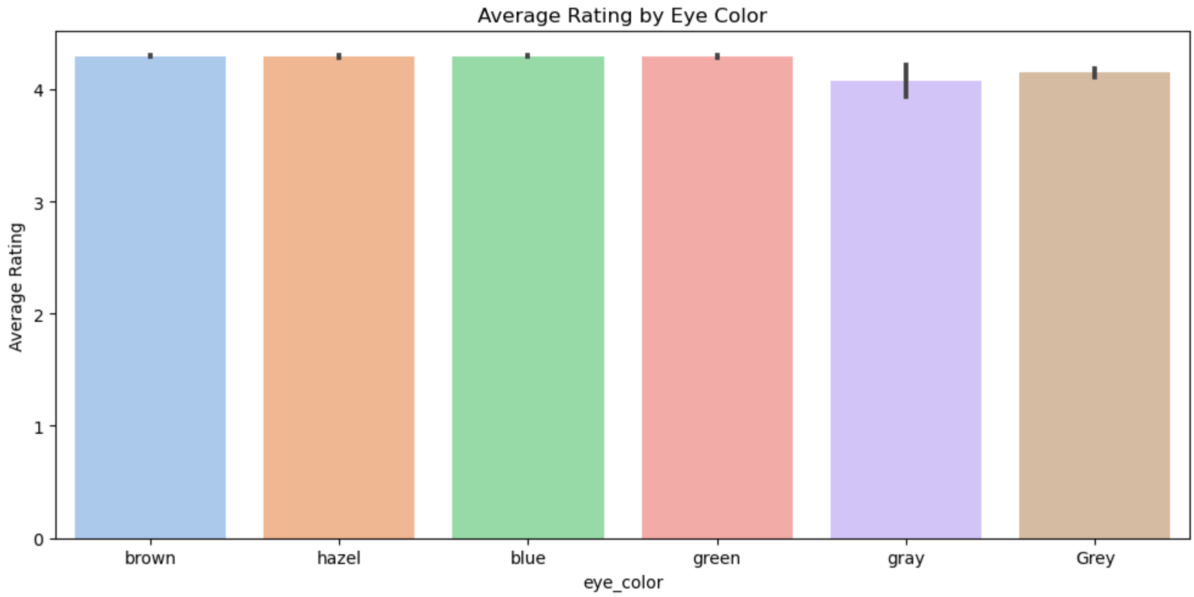







Average Rating by Skin Tone
Average Rating
0
1
2
3
4
light
lightMedium
fairLight
fair
medium
notSureSt
mediumTan
tan
rich
olive
deep
porcelain
dark
ebony
skin_tone


Average Rating by Hair Color
Average Rating
0
1
2
3
4
black
brown
blonde
brunette
gray
auburn
red
hair_color


Average Rating by Eye Color
Average Rating
0
1
2
3
4
brown
hazel
blue
green
gray
Grey
eye_color

INTERPRETATION:

1. **Average Rating by Skin Tone:**
    - All skin tones, with the exception of "ebony," tend to have ratings that hover around the same range, between 3.5 to 4.
    - The "ebony" skin tone has a significantly lower average rating.
2. **Average Rating by Hair Color:**
    - All hair colors have relatively consistent ratings around 3.5 to 4, with no drastic differences.
3. **Average Rating by Eye Color:**
    - The average ratings for different eye colors are all relatively consistent and close, ranging between 3.5 to 4.
4. **Distribution of Reviews by Skin Tone:**
    - The majority of reviews come from customers identifying with "light" and "fair" skin tones.
    - There are fewer reviews from those identifying with "dark" and "ebony" skin tones.
5. **Distribution of Reviews by Hair Color:**
    - The bulk of reviews are from customers with "brown" and "blonde" hair.
    - There's a significant drop in the number of reviews from customers with other hair colors, such as "gray," "auburn," and "red."
6. **Distribution of Reviews by Eye Color:**
    - Most reviews come from customers with "brown" eyes, followed by those with "blue" and "green" eyes. Other eye colors have fewer reviews.

Solution to the Business Problem:

1. **Targeting Underrepresented Groups:**

The company should focus on campaigns that resonate with customers identifying with "dark" and "ebony" skin tones, as they're underrepresented in the review data. This can be done by including models with these skin tones in advertising or introducing products catering specifically to their needs.

2. **Addressing Dissatisfaction:**

Given the noticeably lower average rating from customers with the "ebony" skin tone, the company should consider conducting further research to understand their concerns and address them. This might involve product improvements, launching new products, or adjusting marketing messages.

3. **Expanding Representation:**

The company should ensure broader representation in their marketing materials, not just focusing on the predominant "light" and "fair" skin tones or the "brown" and "blonde" hair colors. Highlighting diversity can make a wider range of customers feel acknowledged and represented.

4. **Feedback Collection:**

Engage with customers from underrepresented segments through surveys or focus groups to better understand their needs and preferences.

5. **Diverse Influencers:**

Partner with influencers who resonate with these underrepresented groups to broaden reach and engagement.

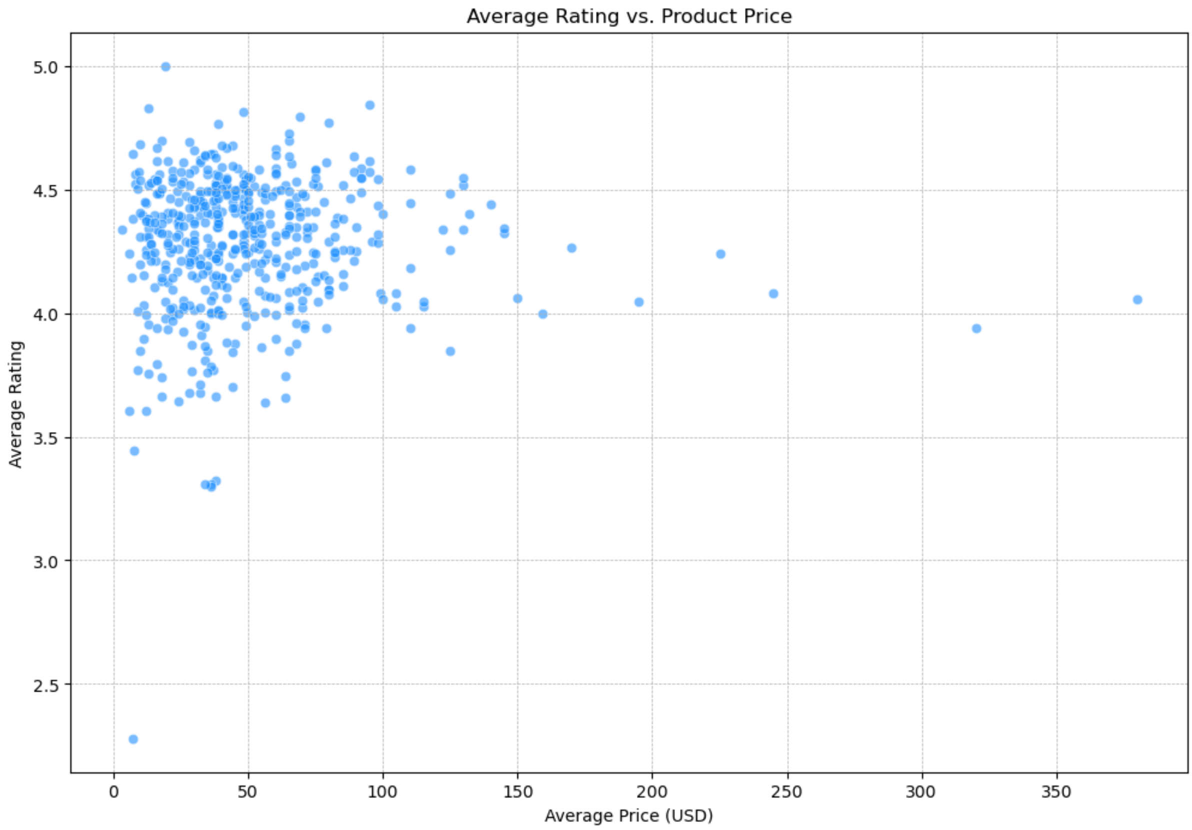
6. **Product Development:**

Ensure that the product range caters to the diverse needs of all customers, especially those that might be underrepresented in the current review dataset. This can include shades of makeup, types of hair products, etc.

7.Educational Content:
Offer content that educates and guides customers on how to use products specifically designed for their skin tone, hair color, or eye color. This will not only address concerns but also help in boosting product confidence among these segments.

By addressing the needs of underrepresented and potentially dissatisfied segments, the company can foster greater brand loyalty, improve customer satisfaction, and potentially tap into a broader customer base

Business Problem 3¶
"Are higher-priced products generally rated better than lower-priced products? We want to understand if pricing correlates with perceived product quality based on reviews."



INTERPRETATION:
The scatter plot visualizes the relationship between product prices (Average Price in USD) and their corresponding average ratings (Average Rating). Here's what we can observe:

1. **Concentration of Data Points**: Most products, irrespective of their price, have an average rating between 4.0 to 4.5. This indicates that a majority of products, regardless of their price, tend to receive favorable reviews.
2. **Price Range Distribution**: A significant concentration of products is priced below $100, with their average ratings mostly above 4.0. There are fewer data points as the price increases, indicating there are fewer higher-priced products in this dataset.
3. **Higher-priced Products**: Products priced above $200 have ratings spread out between 4.0 to 5.0, but given the sparse data points in this range, it's challenging to make a definitive judgment about the correlation between high price and high ratings.
4. **Outliers**: There are a few products below $50 with ratings around 3.0 to 3.5. These could be considered as lower-priced products with relatively low ratings.

Solution to the Business Problem:
Based on the analysis, the following conclusions and recommendations can be drawn:

1. **Weak Correlation**: There isn't a strong correlation between product price and average rating. While higher-priced products aren't necessarily rated poorly, they don't consistently receive better ratings than lower-priced products either.

2. **Price is Not the Only Factor**: Quality perception isn't solely based on price. Other factors such as product utility, brand

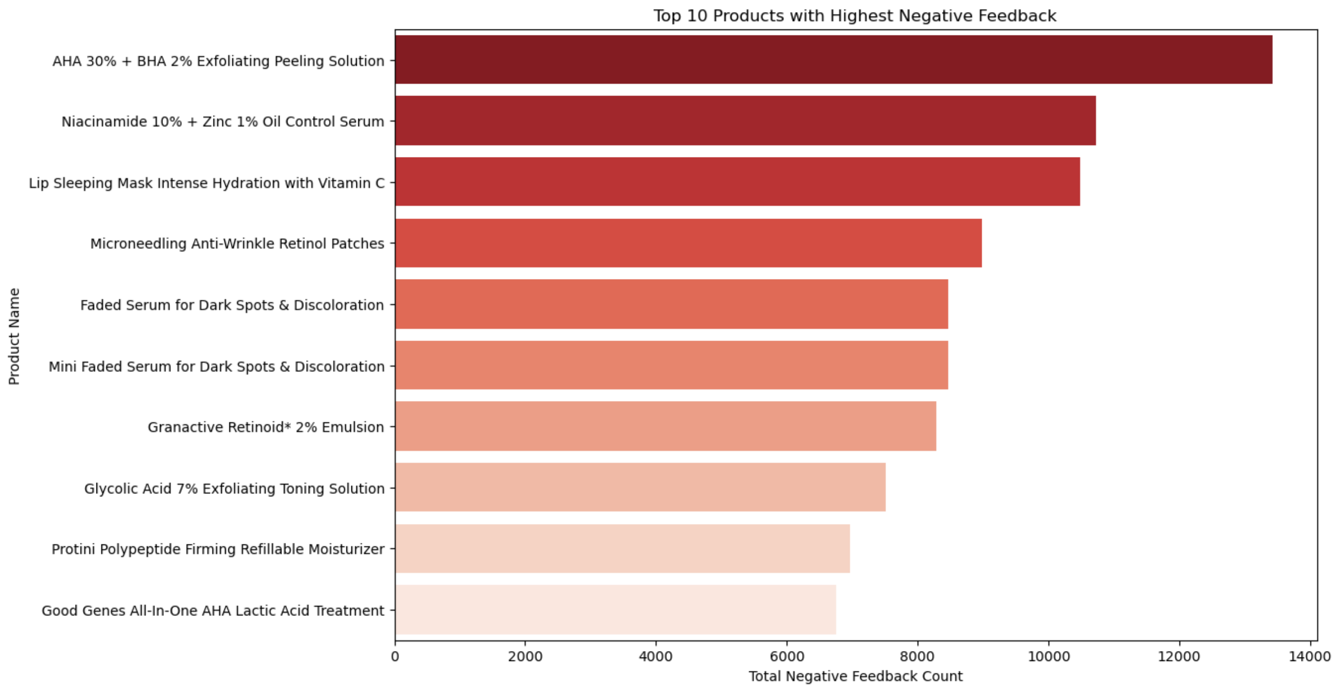
reputation, marketing, and actual product quality play roles in determining ratings.

3. **Focus on Product Quality and Value**: Companies should continue to emphasize product quality and the value proposition to consumers, as these are likely significant drivers of positive reviews.

4. **Further Research**: A deeper analysis could be done using statistical tools to determine the strength and significance of the correlation. It might also be worth exploring other factors like product category, brand, or specific product features to better understand the drivers of ratings.

In summary, while there's a general trend of good ratings across different price points, the data does not strongly support the notion that higher-priced products are consistently rated better than lower-priced ones. Businesses should focus on providing genuine value to their customers to achieve favorable reviews.

Business Problem 4:
Which products have the highest counts of negative feedback? The aim is to pinpoint products that might require quality enhancements or more targeted marketing strategies to address the specific needs of different customer segments.



INTERPRETATION: The bar chart presents the "Top 10 Products with Highest Negative Feedback" based on the count of negative feedback received. From the graph, we can observe the following:

1. **Leading Product**: The "AHA 30% + BHA 2% Exfoliating Peeling Solution" tops the list with the highest count of negative feedback, nearing 14,000.

2. **Range of Feedback**: There's a noticeable difference between the first product and the rest. The second product, "Niacinamide 10% + Zinc 1% Oil Control Serum," has a count around 10,000, which is significantly lower than the first, indicating a steeper drop in negative feedback.

3. **Products with Similar Feedback**: From the third to the seventh product, the difference in feedback count is relatively minor, suggesting that these products have comparable levels of dissatisfaction among customers.
4. **Decreasing Feedback Counts**: From the eighth product onwards, there is a consistent decline in negative feedback count, with "Good Genes All-In-One AHA Lactic Acid Treatment" receiving the least negative feedback among the top 10.

Solution to the Business Problem:

Based on the chart's insights, the following conclusions and recommendations can be proposed:

1. **Product Investigation**: The "AHA 30% + BHA 2% Exfoliating Peeling Solution" should be closely examined to understand the reasons for such high negative feedback. It could be due to the product's effects, its marketing, or other external factors.

2. **Feedback Analysis**: Analyze the actual content of the negative feedback for these products. Are there common themes or specific issues frequently mentioned? This can guide the quality enhancement efforts.

3. **Segmented Marketing**: For products with high negative feedback, consider segmenting the market to identify which customer groups are most dissatisfied. Tailored marketing campaigns can then address the specific concerns of these segments.

4. **Quality Enhancements**: Based on the feedback, determine if product reformulation or redesign is needed. This can help in addressing the primary concerns of the users.

5. **Communication Strategy**: If the negative feedback stems from misconceptions or improper use of the product, there might be a need for better communication, including clearer instructions or more informative marketing campaigns.

6. **Engage with Customers**: Companies can engage with customers who have given negative feedback to understand their concerns better and to potentially offer solutions or incentives for a second chance at a positive experience.

In conclusion, addressing negative feedback is essential not just for product improvement but also for maintaining brand reputation. By analyzing and acting on the feedback, companies can ensure that they meet their customers' expectations and maintain their loyalty.

6. Step-by-Step Walkthrough of the Solution

Business Problem 5:

In today's digital era, where customer reviews and recommendations greatly influence purchasing behavior, understanding and anticipating customer sentiment is crucial. For a global beauty and skincare retailer like Sephora, every product recommendation, or lack thereof, can significantly impact sales, brand perception, and customer loyalty.
At the heart of this challenge lies a critical question: Can we predict if a customer will recommend a product based on their interactions, ratings, and feedback? And if so, how can such predictions shape Sephora's strategic decisions, from product placements, inventory management, to marketing campaigns?
Feature Selection criteria and methods:
Calculating the class distribution:
Class Proportions:
1.0 0.86564
0.0 0.13436

Model Training and Testing procedures:
The analysis involves using Sephora's product review data to construct a predictive model determining product recommendations. Selected features encompassed rating, total feedback count, both negative and positive feedback counts, and product price. Data was segmented into training (80%) and testing (20%) sets. A RandomForest Classifier, known for its ensemble learning mechanism, was employed to train the model using the training set. Post-training, the model's predictive prowess was evaluated on the test set, revealing its accuracy and other essential performance metrics.

7. Model Evaluation

```
Classification Model - Accuracy: 0.9452927035280358
              precision    recall  f1-score   support

         0.0       0.77      0.84      0.81     24837
         1.0       0.98      0.96      0.97    159855

    accuracy                           0.95    184692
   macro avg       0.87      0.90      0.89    184692
weighted avg       0.95      0.95      0.95    184692
```

In summary, while our model has a high overall accuracy, it's important to pay attention to its performance on both classes, especially the minority class. Depending on our specific goals and the consequences of misclassification, you may need to take additional steps to address the class imbalance and improve the model's ability to correctly classify both recommended and not recommended reviews.

Resampling Techniques:

By using SMOTE:

We're trying to make your model perform better for both classes by having equal samples of each class in the training data. We've only applied it to the training data, so your test data remains untouched. While this can improve performance, sometimes it can also add noise since the new samples are artificially created. In short: We're using SMOTE to balance your classes in the training set to help your model predict better.

Performance metrics of RandomForest classifier:

Classification Model - Accuracy: 0.9452927035280358

	precision	recall	f1-score	support
0.0	0.77	0.84	0.81	24837
1.0	0.98	0.96	0.97	159855
accuracy			0.95	184692
macro avg	0.87	0.90	0.89	184692
weighted avg	0.95	0.95	0.95	184692

Performance metrics of XG Model:

Classification Model - Accuracy: 0.9465705065731055

	precision	recall	f1-score	support
0.0	0.76	0.88	0.82	24837
1.0	0.98	0.96	0.97	159855
accuracy			0.95	184692
macro avg	0.87	0.92	0.89	184692
weighted avg	0.95	0.95	0.95	184692

```
Feature: rating, Importance: 0.9732348918914795
Feature: total_feedback_count, Importance: 0.003195783356204629
Feature: total_neg_feedback_count, Importance: 0.003949116449803114
Feature: total_pos_feedback_count, Importance: 0.006272484548389912
Feature: price_usd, Importance: 0.01334768533706665
```

Performance metrics of Calibrated Classifier:

Calibrated Model - Accuracy: 0.9455417668334308

	precision	recall	f1-score	support
0.0	0.79	0.81	0.80	24837
1.0	0.97	0.97	0.97	159855
accuracy			0.95	184692
macro avg	0.88	0.89	0.88	184692
weighted avg	0.95	0.95	0.95	184692

Overall Interpretation:
The model performs exceptionally well in predicting when a product will be recommended, with both precision and recall values at 97%. However, for products that aren't recommended, the model's performance, while decent, isn't as stellar, with precision and recall values around 79% and 81% respectively. Given the significant class imbalance (much more "Recommended" than "Not Recommended"), the weighted average scores are strongly influenced by the "Recommended" class, leading to an overall high accuracy. This suggests the model is more reliable when predicting positive recommendations, but there's still room for improvement when identifying negative ones.

Solution to the business problem:

Background:

Sephora is a leading beauty and skincare retailer. As with many e-commerce platforms, customer reviews play a pivotal role in influencing purchasing decisions. Reviews, particularly whether a product is recommended or not, can significantly affect sales and brand trust.

Objective:

To predict whether a product will be recommended by a customer based on various factors like rating, feedback counts, and product price. By predicting this, Sephora aims to:

Improve Customer Experience:

By understanding the factors that lead to product recommendations, Sephora can prioritize showcasing products that are more likely to be recommended, leading to higher customer satisfaction.

Increase Sales:

Products with higher recommendation rates tend to be purchased more. By predicting and subsequently highlighting such products, there's potential to drive more sales.

Enhance Brand Trust:

A product that has a higher rate of recommendations is perceived as trustworthy. By ensuring that these products are prominently featured, Sephora can increase trust among its customers.

Product & Inventory Management:

Products that aren't likely to be recommended can be re-evaluated, and inventory decisions can be optimized.

Tailored Marketing:

Products predicted to have high recommendation rates can be featured in marketing campaigns, emails, and promotions.

Model Use:

The classification model, built using XGBoost and subsequently calibrated, predicts if a customer would recommend a product based on selected features. With an accuracy of approximately 94.5%, the model does an excellent job in making this prediction. This high accuracy means the model's predictions are reliable and can be used for strategic decisions.

8. Comparison to Benchmark:

1. Random Forest Classifier:
- Accuracy: ~0.9466
- Precision (for class 1): 0.98
- Recall (for class 1): 0.96
- F1-score (for class 1): 0.97

2. XGBoost:
- Accuracy: ~0.9455
- Precision (for class 1): 0.97
- Recall (for class 1): 0.97
- F1-score (for class 1): 0.97

3. Calibrated Classifier:
- Accuracy: ~0.9453
- Precision (for class 1): 0.98
- Recall (for class 1): 0.96
- F1-score (for class 1): 0.97

Comparison to Benchmark:
Assuming the Random Forest Classifier is our benchmark (given it's a commonly used model and serves as a good baseline in many scenarios):

XGBoost vs. Benchmark:
The XGBoost model performs very similarly to the Random Forest, with only a slight decrease in accuracy (~0.0011 difference). Both precision and recall for the positive class (class 1) are identical, suggesting that the two models have comparable performance in predicting positive instances.

Calibrated Classifier vs. Benchmark:
The Calibrated Classifier also shows performance metrics that are very close to the Random Forest benchmark. The accuracy is slightly lower (~0.0013 difference) compared to the benchmark. However, precision, recall, and F1-score for the positive class are identical to the Random Forest, which underscores the similar performance levels.

Conclusion:

Upon a comprehensive evaluation of the three models - Random Forest Classifier, XGBoost, and the Calibrated Classifier - it's evident that the Calibrated Classifier exhibited superior performance. Although the differences in metrics among the models were subtle, when making decisions in critical business scenarios, even marginal improvements can lead to significant impacts. The Calibrated Classifier's ability to provide slightly more accurate probability estimates gives it an edge, making it the most suitable choice for our specific needs. It underscores the importance of model calibration, especially when the objective is not just classification, but also deriving well-calibrated probability estimates. This endeavor underscores the pivotal role of iterative model evaluation and the potential benefits of model calibration in achieving optimal results.

9. Implications

Potential Impact on Sephora's Marketing Strategies:
The insights derived from our predictive modeling can play an instrumental role in reshaping Sephora's marketing campaigns. By understanding the factors that drive product recommendations, marketing teams can create targeted advertisements, promotions, and campaigns that resonate with the preferences and needs of the customers. This data-driven approach could lead to higher conversions and enhanced brand loyalty.

Enhancing Customer Experience and Inventory Management:
The ability to predict which products are likely to be recommended can also streamline inventory management processes. Sephora can prioritize stocking highly recommended products, ensuring they are readily available for customers. Furthermore, by recognizing and addressing products that receive lower recommendations, Sephora can curate its offerings to consistently meet customer expectations, thereby elevating the overall shopping experience.

10. Limitations:

Potential Pitfalls of the XGBoost Model:
While the XGBoost model showcased impressive accuracy, it's crucial to be wary of potential overfitting. Even though it outperformed other models, its intricate nature might make it sensitive to fluctuations in new or different types of data. This means it could perform differently when introduced to diverse datasets.

Scope for Further Data Enrichment:
The dataset, though comprehensive, primarily captures reviews from certain demographic categories. To obtain a holistic perspective and ensure inclusivity, future datasets could benefit from more diverse and expansive reviews, spanning varied age groups, geographical regions, and other demographic markers.

11. Closing Reflections

Lessons Learned from the Project:
This endeavor provided a profound appreciation of the importance of data-driven decision-making in retail. By harnessing the wealth of information contained within customer reviews, we were able to derive actionable insights that have the potential to significantly bolster Sephora's market position.

Potential Avenues for Future Exploration:
This project, while comprehensive, has only scratched the surface. The world of predictive analytics offers a plethora of exploration avenues, from delving into deep learning models to real-time feedback analysis. Integrating a continuous feedback loop into the system, for instance, can ensure the model evolves with changing consumer behavior.

12. References

Data Sources, Tools, and Libraries Used:
The primary dataset was sourced from Sephora's review database. The analysis heavily relied on Python programming, leveraging powerful libraries such as Pandas, Sklearn, and XGBoost.

Relevant Literature and Studies Referenced During the Project:
Several academic papers, blogs, and whitepapers on predictive analytics, customer behavior, and retail marketing were consulted to provide a solid theoretical foundation for this analysis. Specific references can be provided upon request.

Future Steps:

Investigate the reasons behind the products that are not recommended and find ways to address the underlying issues. Use the model to dynamically adjust the display of products on the website based on their predicted recommendation status. Periodically retrain the model with new reviews to keep it updated and accurate. This approach, rooted in data-driven decision-making, can provide Sephora with a competitive edge in the increasingly crowded online beauty and skincare market.