Kledja Caushi

COMP 3125

Dr. Weijie Pang

April 06, 2025

Draft Report - Sephora Data Analysis

Introduction

This project aims to explore consumer behavior and preferences regarding beauty products on Sephora.com, particularly focusing on how consumer attributes (such as skin type, hair color, and eye color) and review characteristics (ratings, helpfulness, and feedback counts) influence purchasing decisions. The primary research questions include:

1. How do product reviews, ratings, and the number of likes influence the likelihood of a purchase on Sephora.com?
2. How does the price of a product compare to similar products on other websites, and does this impact the likelihood of purchase?
3. How do specific ingredients in beauty products affect purchasing decisions, especially when considering potentially harmful or controversial ingredients?
4. Can we build a predictive model to estimate the probability of a product being purchased based on its reviews, likes, price, and ingredients?

The hypothesis driving this research is that consumer personal attributes significantly influence their product satisfaction and review ratings, potentially affecting other consumers' purchasing decisions.

Datasets

The dataset was scraped from Sephora.com and contains detailed product reviews alongside extensive user-provided context data. The initial dataset includes user reviews with structured

context data such as skin type, skin tone, eye color, hair color, and indicators if the review was incentivized or authored by a Sephora employee.

Initially, Ingredients were not scraped off the products. We instead opted on a public Kaggle dataset which had additional metrics such as:

- Product pricing (regular, value, and sale prices)
- Product details (ingredients, size, variation type)
- Popularity metrics (loves count, number of reviews, average rating)

Data preprocessing involved flattening nested JSON fields (ContextDataValues) using DuckDB for structured analysis, and subsequent analysis and visualization using Python's Pandas and Seaborn libraries. No extensive imputation was performed, as missing context attributes were clearly represented as null values.

Methodology

To answer the research questions, the following methods were employed:

1. **Data Flattening**: Using DuckDB to extract nested JSON fields (hairColor, eyeColor) into a flat tabular format suitable for analysis.
2. **Exploratory Data Analysis (EDA)**: Using Pandas to group, summarize, and calculate average ratings by categorical user attributes.
3. **Visualization**: Utilizing Seaborn bar plots to visually represent differences in average ratings by hair and eye color.
4. **Correlation Analysis**: Examining correlations between reviewer attributes and product ratings to identify potential trends or biases.

Future analyses will incorporate predictive modeling methods, such as logistic regression or random forest classifiers, to predict purchasing likelihood based on reviews, user engagement metrics, pricing, and ingredients.

Results

Preliminary results indicated observable differences in product ratings across different hair and eye color categories. For instance, users with specific attributes showed slight preferences or dissatisfaction with certain products. Detailed statistical analysis and visualization results will be presented in the final report.

Discussion

These preliminary findings suggest that consumer physical attributes potentially impact their product experiences and satisfaction. Understanding these nuanced preferences can allow Sephora and other retailers to optimize product recommendations, marketing strategies, and product development tailored to diverse customer segments. Further analysis incorporating pricing and ingredient data will provide more comprehensive insights, aligning with existing consumer behavior research emphasizing personalization in product marketing.

Future research could explore deeper sentiment analysis, cross-platform price comparisons, and predictive analytics to further refine consumer targeting and engagement strategies.