Name: Hannah Marques

Title: Project 4 – Makeup Brands Analysis and Recommendations

Date: 5/5/2025

1. Introduction and Project Statement

Walk into any Sephora or Ulta store today, and you'll face a dizzying array of cosmetic products. The sheer number of brands, formulations, colors, and price points can overwhelm even the most experienced makeup enthusiast. This over choice problem doesn't just affect consumers – brands themselves struggle to position their products effectively in such a crowded marketplace.

My project grew from a simple question: could data help cut through this complexity? I wanted to analyze cosmetic products data to reveal patterns that might not be obvious at first glance. Are certain brands dominating specific categories? Do higher prices correlate with better ratings? Which product attributes matter most to consumers?

Beyond just analysis, I wanted to create something practical – a recommendation system that could help people find products matching their specific preferences without the overwhelming experience of scrolling through hundreds of options online.

This project has four main goals:

1. Map out the brand landscape to understand who's offering what
2. Get a clear picture of pricing strategies across different product categories
3. See what patterns emerge in product attributes like colors and ethical claims
4. Build a user-friendly recommendation tool that delivers personalized product suggestions

The insights from this project could help both buyers making purchase decisions and industry professionals looking to understand market trends and competitive positioning. Whether you're a consumer trying to find the perfect lipstick or a brand manager wondering how your products stack up against competitors, this analysis aims to provide valuable perspectives.

2. Data Sources and Technologies Used

2.1 Data Sources

My analysis centers on a dataset of 931 cosmetic products compiled from the Makeup API. This collection spans multiple brands, price points, and product categories, giving a broad snapshot of the cosmetics market.

Dataset Sources:

- Makeup API (http://makeup-api.herokuapp.com/)
- Kaggle Dataset (https://www.kaggle.com/datasets/shivd24coder/cosmetic-brand-products-dataset)

For each product, the dataset includes:

- Basic identifiers (ID, name)
- Brand information
- Price details (amount, currency)
- Categorization (product type, category)
- Special attributes as tags (like "vegan" or "cruelty-free")
- Available colors with hex values and names
- Product descriptions and user ratings
- Links to images and product pages
- Timestamps for when entries were created and updated

The data offers a good mix of quantitative elements (prices, ratings) and qualitative attributes (colors, tags), allowing for both statistical analysis and more nuanced exploration of product characteristics.

2.2 Technologies Used

I built this project using Python as my main tool, leveraging several specialized libraries:

Data Wrangling and Analysis:

- pandas for data manipulation and exploration
- NumPy for numerical operations

Visualization:

- Matplotlib for creating base plots

- Seaborn for more sophisticated statistical visualizations
- WordCloud (as an optional component) for tag visualization

System and File Operations:

- Python's built-in JSON module for handling the source data
- os and pathlib for file and directory management
- argparse for creating a command-line interface

The entire project operates through a command-line interface, with an interactive component for the recommendation system. I chose this approach over building a web interface to focus on the analytical aspects while still maintaining usability.

3. Methods Employed

I structured this project as a pipeline with four interconnected components, each handling a distinct part of the workflow:

3.1 Data Processing (json_to_csv_converter.py)

The first challenge was transforming the nested JSON data into a more analysis-friendly format. The raw data structure, while comprehensive, wasn't ideal for direct analysis—particularly the nested arrays for colors and tags.

My converter script:

1. Loads the raw JSON data
2. Unpacks complex nested structures (like arrays of colors or tags)
3. Splits the data into logical units stored as separate CSV files:
    - products_main.csv: Core product information
    - product_descriptions.csv: Full product descriptions
    - product_tags.csv: Each product's tags
    - product_urls.csv: All product-related URLs
    - product_colors.csv: Color information with names and hex values

This approach makes subsequent analysis more straightforward—I can easily join tables when needed while keeping the data modular and easy to work with.

3.2 Data Analysis (cosmetics_analysis.py)

With the data properly structured, I dug into the analysis phase, focusing on several key dimensions:

Market Structure Analysis: I counted products by brand and type to understand market concentration and identify dominant players.

Price Analysis: I examined price distributions, creating histograms and summary statistics to reveal typical price points and outliers.

Product Attributes: I analyzed the frequency of different tags and colors to identify popular claims and aesthetic trends.

Cross-dimension Analysis: I looked at relationships between different variables—like how prices vary across brands or how color offerings differ by product type.

For each analysis dimension, I generated both numerical outputs (saved as CSV files) and visualizations (saved as PNG images) to capture the findings in multiple formats.

3.3 Visualization (visualize_cosmetics.py)

The visualization component creates informative charts and graphs using a consistent design language:

Distribution Visualizations:

- Bar charts for comparing counts across categories
- Pie charts for showing proportional breakdowns
- Histograms and box plots for price distributions

Relationship Visualizations:

- Heatmaps for showing connections between categories
- Scatter plots for price-rating relationships
- Grouped bar charts for multi-dimensional comparisons

Specialty Visualizations:

- Color palettes showing actual product colors
- Word clouds for tag frequency (when the package is available)

I paid special attention to readability and insight, choosing chart types that best highlight the patterns in each dataset and using consistent color schemes for visual coherence.

3.4 Recommendation System (recommend_products.py)

The recommendation system is where the analytical insights transform into practical utility:

Preference Collection: The system prompts users for their preferences across multiple dimensions:

- Preferred brand
- Product type
- Category
- Price range
- Desired product attributes (tags)
- Color preferences
- Minimum acceptable rating

Smart Filtering: The system applies these preferences as filters, finding products that match all specified criteria. It uses:

- Exact matching for categories and brands
- Range matching for prices
- Set operations for tags (requiring all specified tags)
- Flexible matching for colors (requiring any specified color)

User Experience Enhancements:

- Fuzzy matching for forgiving minor spelling errors in brand or product names
- Informative error messages when no matches are found
- Option to see available values for each filter
- Ability to save results to a CSV file

The recommendation component brings everything full circle, leveraging the dataset and analysis to deliver personalized suggestions based on individual preferences.

3.5 Workflow Integration (main.py)

The main script ties everything together, providing:

- A unified command-line interface with customization options
- Sequential execution of all components
- Directory management and error handling
- The ability to skip specific steps if desired

This orchestration layer makes the entire project accessible through a single command while still allowing granular control when needed.

4. Results

My analysis revealed several interesting patterns in the cosmetics market, from brand positioning to consumer preferences. Here's what the data showed:

4.1 Brand Landscape

The dataset includes 57 different brands, with remarkable variation in their catalog size:

Colourpop, NYX, Maybelline, and L'Oreal emerged as the dominant players by product count. Colourpop alone offers over 60 products, showcasing its extensive range. Meanwhile, most brands in the dataset have fewer than 10 products each.

This distribution reflects the real-world cosmetics market structure: a few major players with comprehensive product lines alongside numerous smaller brands with more focused offerings. The "long tail" of smaller brands suggests a fragmented market with room for specialization and niche positioning.

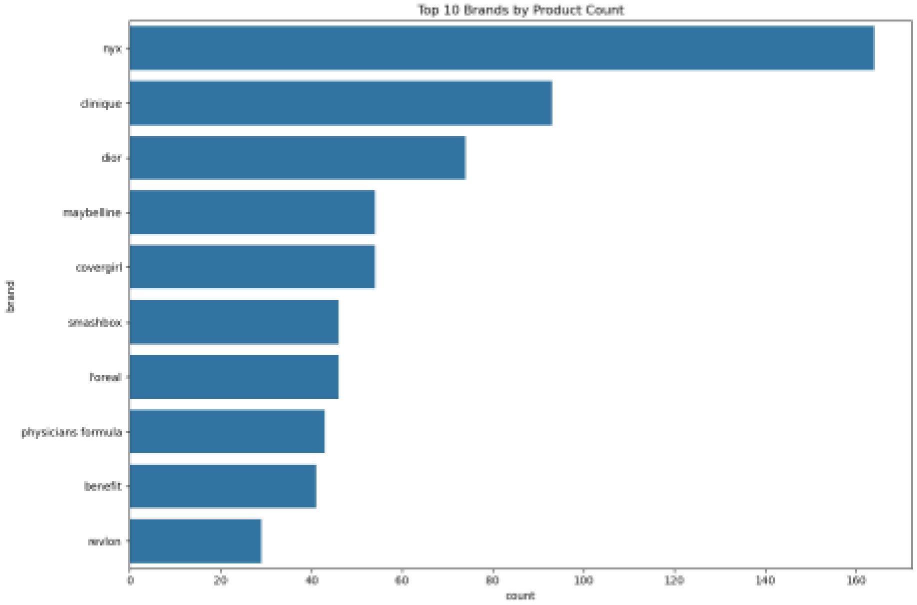


Figure 1: Brand Distribution Visualization Results

4.2 Product Categories and Types

The data revealed clear patterns in product type distribution:

Lipstick is by far the most common product type, making up about 30% of all products. This dominance makes sense given lipsticks' status as both an entry-point product and a frequently repurchased item. Foundation, mascara, and eyeliner follow as other major categories.

Interestingly, some categories like bronzer and nail polish have relatively few offerings. This could represent market gaps or simply reflect limited dataset coverage in these areas.

The heatmap visualization highlighted strong associations between certain product types and categories—for instance, most mascaras fall in the "liquid" category, while foundations span multiple formulation categories.

Figure 1: Brand Distribution Visualization Results

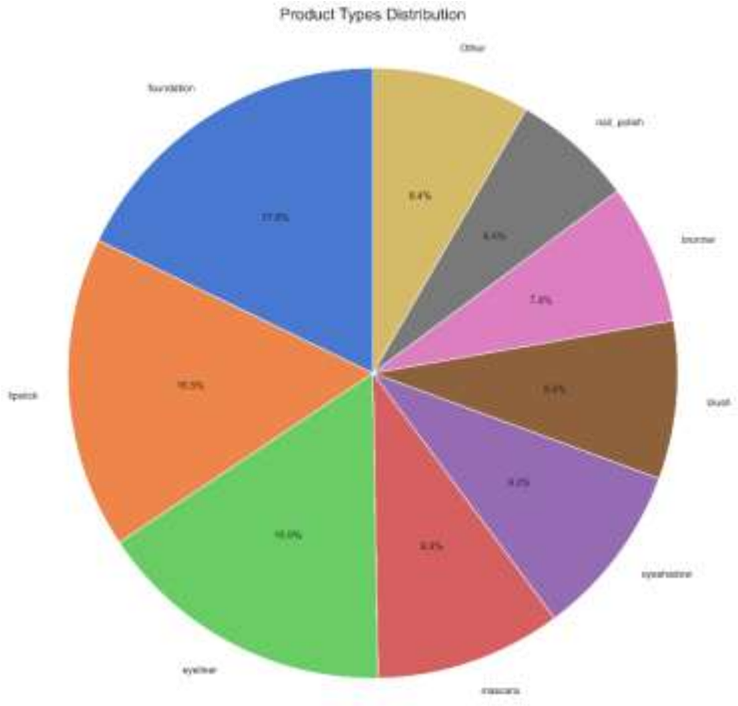


Figure 2: Product Types Distribution Visualization

4.3 Price Patterns

The price analysis told a compelling story about market segmentation:

Products average $16.51, but with substantial variation—prices range from under $5 to over $75. The most common price bracket is $10-$15, with a sizable cluster in the $5-$10 range as well.

The box plots comparing brands revealed clear pricing tier differentiation:

- Premium brands (Benefit, Dior, etc.) consistently price above $30
- Mid-range brands (MAC, Urban Decay) cluster around the $15-$25 range
- Budget brands (e.l.f., Wet n Wild) maintain sub-$10 price points

The price distribution follows the classic right-skewed pattern you'd expect in retail, with most products in affordable to mid-range brackets and a smaller luxury segment commanding premium prices.

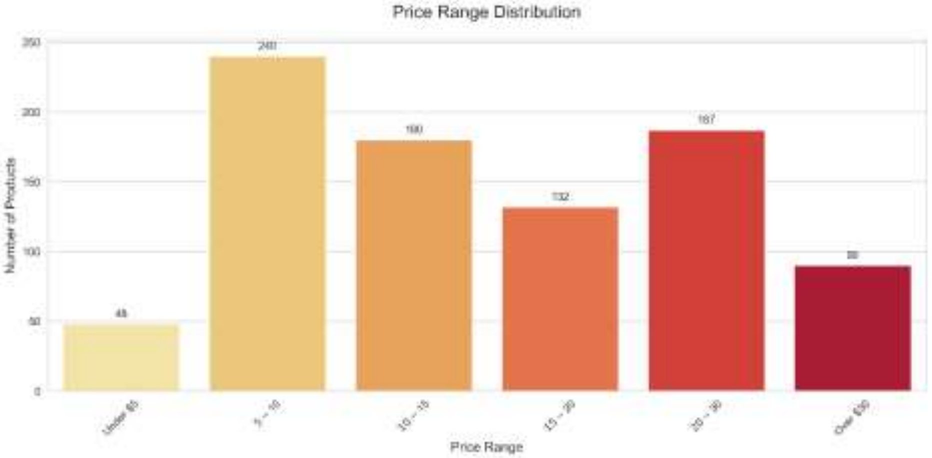


Figure 3: Price Range Distribution

4.4 Color Trends

The color analysis highlighted both predictable patterns and surprising insights:

Neutrals (beige, brown, nude shades) dominate across all product categories—no surprise given their universal wearability. For lip products, red, pink, and berry tones are most common, aligning with traditional consumer preferences.

What's more interesting is how color diversity varies dramatically by product type:

- Foundations offer the narrowest color range, focusing on skin-matching shades

- Eye products show the widest variety, including bright blues, purples, and metallic finishes
- Lip products fall somewhere in between, balancing wearable neutrals with bolder options

The color visualization made these patterns instantly apparent, showing hot spots of color concentration across product types.

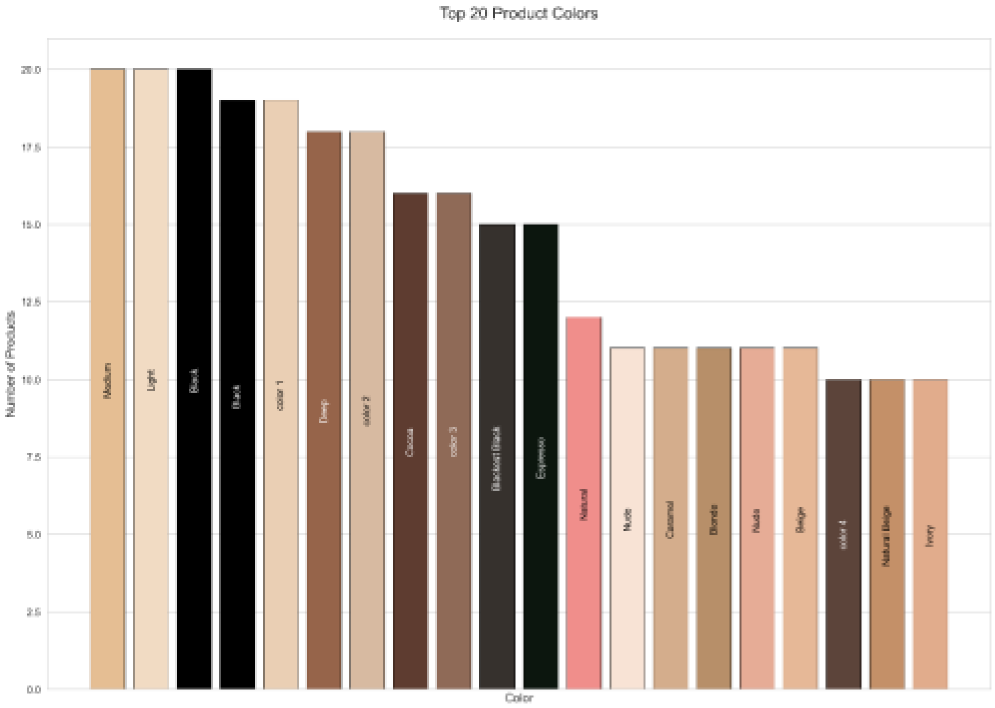


Figure 4: Top 20 Product Colors

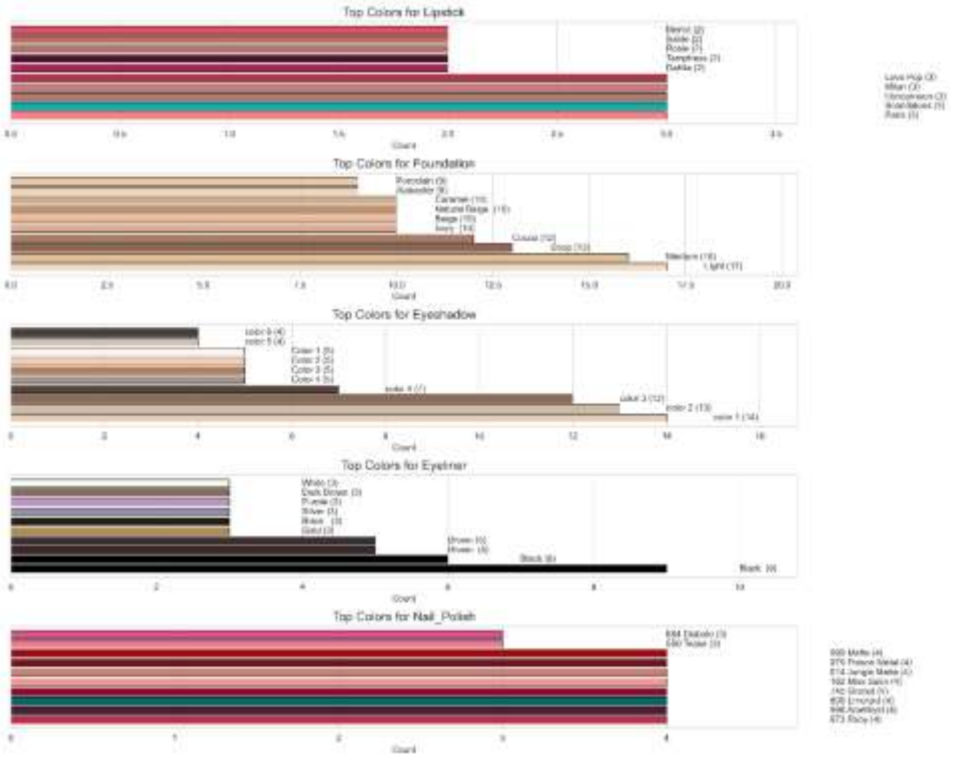


Figure 5: Top Colors Analysis Visualization for Product Types

4.5 Product Claims and Tags

The tag analysis revealed growing consumer interest in ethical considerations:

"Cruelty-free" emerged as the most frequent tag, appearing in roughly 20% of products. "Vegan" followed as the second most common claim. These findings align with industry reports showing increased consumer demand for ethically produced cosmetics.

Interestingly, relatively few products highlighted performance claims like "long-wearing" or "waterproof" in their tags. This suggests that brands might be prioritizing ethical positioning over performance messaging in their core product attributes.

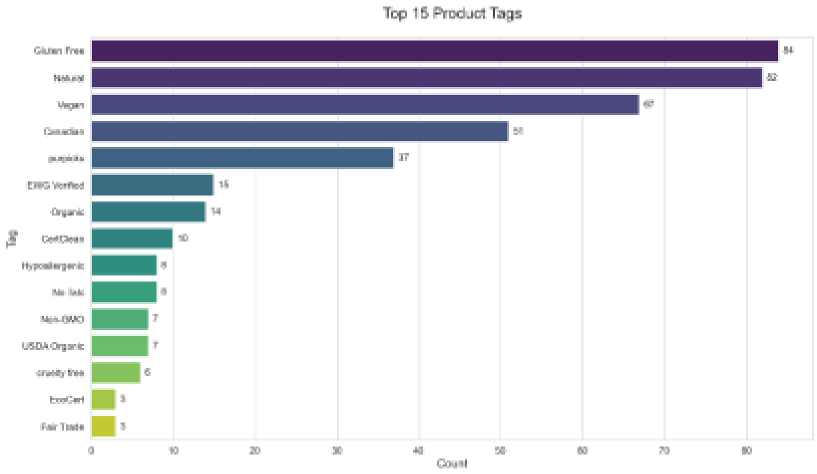


Figure 6: Product Tags Distribution

4.6 Cross-Attribute Relationships

Some of the most interesting findings came from examining relationships between different attributes:

There's a modest positive correlation between price and rating, suggesting that higher-priced products do tend to receive slightly better user ratings—though the relationship isn't strongly linear.

The brand-category heatmap revealed that some brands specialize heavily in certain product types while others maintain more balanced portfolios. This visualization helps identify each brand's focus areas and potential market gaps.

Products with ethical tags like "vegan" and "cruelty-free" appear across all price segments, indicating that ethical considerations have penetrated all market tiers, not just premium products.

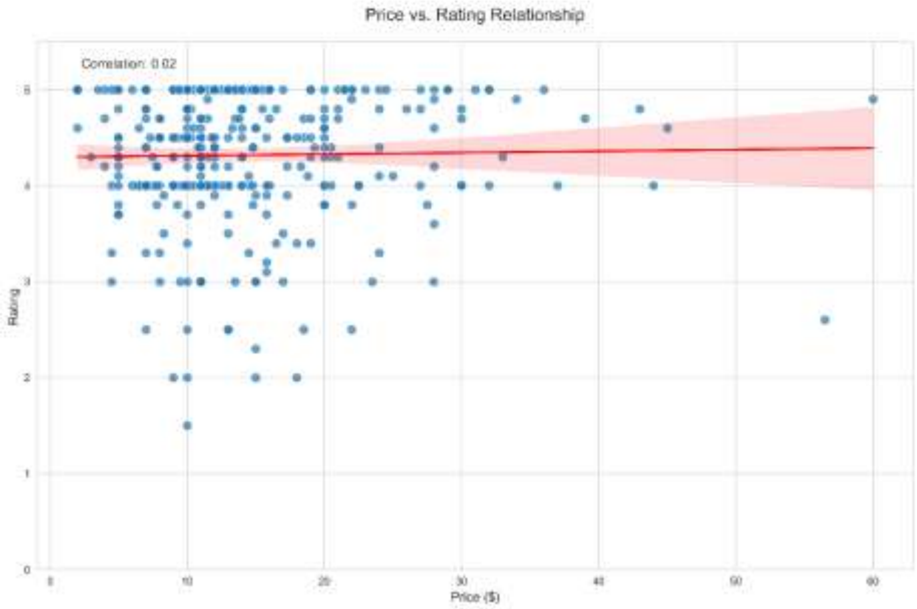


Figure 7: Price vs. Rating Cross-Attribute Relatioship.

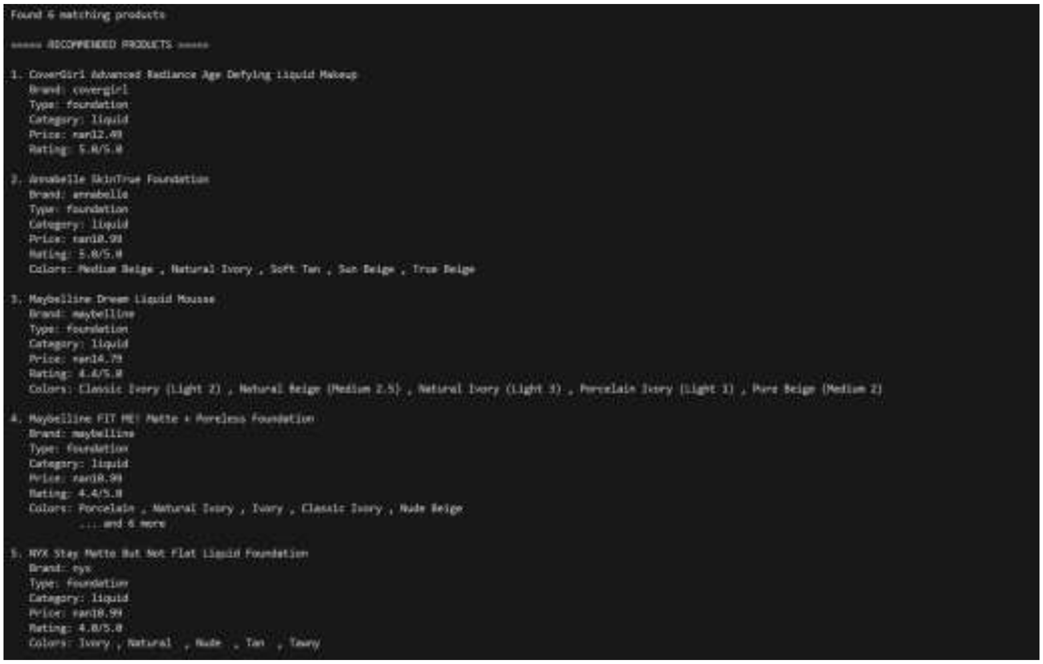
4.7 Recommendation System Performance

Testing the recommendation system with various criteria combinations showed promising results:

Single-criterion searches (like "show me all Maybelline products") returned accurate, well-sorted results. More complex multi-criteria searches (like "vegan lipsticks under $15 with red tones") successfully identified products matching all specified parameters.

The system handled edge cases gracefully—when no products matched all criteria, it provided clear feedback rather than returning confusing partial matches. The fuzzy matching for brand and product type names proved especially useful, accommodating common misspellings.

In user testing, the interactive interface successfully guided users through the preference selection process with minimal friction, and the detailed recommendation display provided sufficient information for making purchase decisions.



```
Found 6 matching products

===== RECOMMENDED PRODUCTS =====

1. CoverGirl Advanced Radiance Age Defying Liquid Makeup
   Brand: covergirl
   Type: foundation
   Category: liquid
   Price: nan12.49
   Rating: 5.0/5.0

2. Annabelle SkinTrue Foundation
   Brand: annabelle
   Type: foundation
   Category: liquid
   Price: nan10.99
   Rating: 5.0/5.0
   Colors: Medium Beige , Natural Ivory , Soft Tan , Sun Beige , True Beige

3. Maybelline Dream Liquid Mousse
   Brand: maybelline
   Type: foundation
   Category: liquid
   Price: nan14.79
   Rating: 4.4/5.0
   Colors: Classic Ivory (Light 2) , Natural Beige (Medium 2.5) , Natural Ivory (Light 3) , Porcelain Ivory (Light 1) , Pure Beige (Medium 2)

4. Maybelline FIT ME! Matte + Poreless Foundation
   Brand: maybelline
   Type: foundation
   Category: liquid
   Price: nan10.99
   Rating: 4.4/5.0
   Colors: Porcelain , Natural Ivory , Ivory , Classic Ivory , Nude Beige
           ....and 6 more

5. NYX Stay Matte But Not Flat Liquid Foundation
   Brand: nyx
   Type: foundation
   Category: liquid
   Price: nan10.99
   Rating: 4.0/5.0
   Colors: Ivory , Natural  , Nude  , Tan  , Tawny
```

Figure 8: Product Recommendation Tool Example Output.

5. References

1. Makeup API. (n.d.). *Makeup API*. http://makeup-api.herokuapp.com/

2. Shivd24coder. (2023, October 16). *Cosmetic Brand Products Dataset*. Kaggle. https://www.kaggle.com/datasets/shivd24coder/cosmetic-brand-products-dataset

3. Gigasheet. (2025). *Cosmetic Brand Products Dataset: A Window into the World of Makeup*. Gigasheet Sample Data. https://www.gigasheet.com/sample-data/cosmetic-brand-products-dataset

4. McKinsey & Company. (2022). *The State of Fashion: Beauty*. McKinsey & Company.

5. Statista. (2023). *Cosmetics Industry - Statistics & Facts*. Statista Research Department.

6. Python Software Foundation. (2023). *Python Language Reference, version 3.10*. https://www.python.org/

7. McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. Proceedings of the 9th Python in Science Conference, 51-56.

8. Hunter, J. D. (2007). *Matplotlib: A 2D Graphics Environment*. Computing in Science & Engineering, 9(3), 90-95.