Instructions for CIS graduate class Term Paper Manuscript

Christopher Lee

Department of Information Systems, California State University

Los Angeles

e-mail : clee219@caltstatela.edu

**Abstract:** The beauty and skincare industry is rapidly changing, driven by new consumer demands, digital influence, and the desire for ethical, personalized products. This study uses two Kaggle datasets, Most Used Beauty & Cosmetics Products in the World and Sephora Products and Skincare Reviews, to analyze these trends. The datasets provide product details, brand attributes, user ratings, ingredients, pricing, and demographics. Hive QL was used to clean, merge, and analyze data, addressing product popularity, ingredient satisfaction correlation, and the impact of brand and price on purchases.

**1. Introduction**

The global beauty and cosmetics industry has experienced unprecedented growth over the past decade, with skincare emerging as a dominant segment fueled by social media influence, growing consumer awareness, and the increasing demand for personalized self-care. What was once a market focused on traditional notions of beauty has now evolved into a multifaceted ecosystem, where customer experiences, product transparency, and online engagement shape brand loyalty and market success. In this context, data analytics offers a powerful lens through which we can better understand consumer preferences, evaluate product performance, and explore the factors that drive purchasing behavior.

In 2024, the beauty industry is expected to reach new heights, driven by trends such as clean beauty, cruelty-free product development, and hybrid skincare-cosmetics. According to a recent industry report by GMI Research (2024), the global skincare market is projected to grow at a compound annual growth rate (CAGR) of over 4.5% between 2023 and 2028, largely driven by consumer demand for organic ingredients and sustainable formulations. This trend highlights a shift toward conscious consumerism, where buyers actively seek products that align with their personal values and skin-specific needs. The rise in e-commerce platforms such as Sephora has enabled customers to access a wide variety of products while reading in-depth user reviews, thereby influencing their decision-making with peer-generated content.

Similarly, Deloitte’s (2023) global beauty industry trends analysis highlights the increasing importance of digital engagement and personalization. In particular, the study reveals that over 60% of Gen Z and Millennial consumers rely on user-generated content, such as product reviews and social media endorsements, before purchasing beauty products. This insight underscores the value of harnessing real-time customer feedback, ratings, and demographic patterns to better understand what drives product popularity and brand trust. As customer data becomes more accessible through online platforms, analyzing this information with tools like Python and pandas becomes essential for extracting meaningful patterns and actionable business intelligence.

This project aims to investigate customer behavior and product trends by analyzing two comprehensive datasets: the Sephora Products and Skincare Reviews dataset and the Top Beauty & Cosmetics Products Worldwide 2024 dataset. These datasets offer rich information such as product ratings, price points, brand reputation, ingredients, demographic details, and user feedback metrics. By combining and cleaning these datasets, we aim to uncover key insights related to product popularity, pricing dynamics, ingredient influence, and demographic preferences.

Through detailed data cleaning, summary statistics, and insightful visualizations, this project addresses several core research questions: Which products are most popular globally and why? How do price and ingredients correlate with product popularity? What role do brand origin and demographic factors play in shaping consumer ratings? And finally, which ingredients are most favored by users around the world?

By answering these questions, this analysis will provide a deeper understanding of the evolving beauty landscape from a data-driven perspective. It will also offer valuable insights for marketers, product developers, and retailers aiming to meet customer expectations in a highly competitive market. This project not only highlights the practical value of data analytics in consumer industries but also underscores the growing need for ethical, transparent, and customer-centered approaches in product design and marketing.

2. Related Work

In my data-driven project it delivers a robust and comprehensive analysis of global skincare trends by examining two major datasets, Sephora product reviews and top-rated global cosmetics. This project stands apart through its integration of Hadoop Distributed File System (HDFS), Hive, and Python-based data processing tools to clean, normalize, and analyze over a million records. By synthesizing structured fields such as price, rating, ingredients, skin type, and consumer feedback, My work explores multifaceted business questions: What products are most favored across global markets? How do demographics and ingredients affect product ratings? What role does geographic origin play in product pricing and consumer trust? The project culminates in a detailed set of visualizations addressing multiple analytical questions, reinforcing its value not only in academic analysis but also in real-world marketing, product design, and inventory decisions.

In contrast, many publicly available skincare projects focus on narrower goals and use less complex pipelines. For instance, a project published on Medium by Dzakiah Putri focuses on building a skincare product recommendation system based on content similarity.3 The core technique is TF-IDF vectorization of product descriptions, followed by cosine similarity to suggest alternative products. While functional and useful for consumer-facing applications, the scope of this project is limited to metadata matching and does not explore factors like ingredient efficacy, user sentiment, or price-performance comparisons.

Similarly, Melissa Monfared’s Kaggle notebook emphasizes exploratory data analysis and basic sentiment assessment of skincare reviews using Python libraries like TextBlob.2 Her work offers insight into rating distributions and product categories, using visualizations like bar charts and word clouds to represent review sentiment. However, the dataset used is significantly smaller, and there is no comparison across global markets or application of distributed systems. Monfared’s work is an accessible introduction to Exploratory Data Analysis but lacks the scale, rigor, and business implications demonstrated in my project.

A third project by Usama Khalid hosted on GitHub pushes deeper into natural language processing by applying a fine-tuned DistilBERT model to classify Sephora review sentiments.1 This approach leverages transformer-based deep learning models for binary sentiment classification and includes performance evaluations and predictions. While Khalid’s work is technically advanced in its application of machine learning, it focuses solely on text-based classification and does not incorporate structured fields like product category, pricing, or brand origin. As such, it provides valuable insights into customer tone but is not designed for business intelligence or comparative market analysis.

What distinguishes my project is its end-to-end scope: ingestion of large-scale datasets via Hive, data engineering using best practices (e.g., COALESCE for null handling, mode replacement for categorical variables), and strategic insights derived from a variety of structured data points. While the external projects each offer useful methods recommendation via content filtering,3 sentiment analysis using rule-based NLP,2 or transformer-based classification1 none match the holistic and business-oriented nature of my project. For future improvements, my project could incorporate sentiment models like those developed by Khalid or integrate similarity recommendations to further enhance user experience and decision-making capabilities.

3. Specifications

The table below provides a structured overview of the datasets utilized in this analysis, including key schema attributes extracted from each source. Specifically, two datasets were examined: Sephora Reviews and Top World Products, both sourced from publicly available Kaggle repositories. The Sephora Reviews dataset includes fields such as product information and pricing information. Meanwhile, the Top World Products dataset captures comparable attributes including ingredients and skin type information.

To enable large-scale processing and comparative analysis, all data was ingested and analyzed using a Hadoop Distributed File System (HDFS) architecture with Hive for querying. This approach allowed for efficient data handling across a distributed environment, supporting both normalization and harmonization of schema elements between the two datasets. Schema alignment was a critical step in ensuring analytical consistency across variables such as price, rating, ingredients, and product metadata. The total size of the processed data is approximately 4.27 GB.

By harmonizing schemas and leveraging distributed computing, this project ensures scalable, efficient processing and establishes a strong foundation for subsequent statistical analysis and visualization. The schema specifications also assist in determining field compatibility and enable more meaningful cross-dataset comparisons.

**3.1 Hardware Specification**

The table labeled Hardware Specifications outlines the computing infrastructure used to process and analyze the skincare and cosmetics datasets on a Hadoop-based distributed system. Two master nodes are responsible for managing and coordinating the distributed computing tasks across the cluster. In Hadoop, master nodes typically handle resource allocation (via YARN), file system metadata (via HDFS NameNode), and job scheduling. Having two master nodes can indicate high availability and fault tolerance. Three worker nodes that perform the actual data processing tasks. In a Hadoop environment, worker nodes run different services for storage or processing. The three worker nodes provide the parallel computing power necessary for distributed data querying and transformation. CPU Speed at 2.45 GHz indicates the processing speed of the CPUs across the nodes. A speed of 2.45 GHz suggests a mid-tier performance level sufficient for running Hive queries and parallel data operations without excessive latency. 155 GB is the total storage capacity available across the cluster for holding datasets and intermediate files. While 155 GB is relatively modest for big data applications, it is sufficient for the size of this project (4.27 GB dataset), allowing for multiple transformations and staging tables within Hive. Together, these specifications support efficient consumption, transformation, and querying of large datasets in a distributed architecture. The use of multiple nodes enhances scalability and fault tolerance, while the CPU and disk resources are aligned with the moderate data volume and analytical complexity of the project.

|  |  |
| --- | --- |
| **Hardware Specifications** | |
| Master Nodes | 2 |
| Worker Nodes | 3 |
| CPU Speed | 2.45 GHz |
| Disk Space | 155 GB |

**3.2 Data Specifications**

This section describes the overall size and sources of the datasets. The combined data volume used for analysis is 4.27 GB, which includes both datasets before transformation and normalization. This dataset, sourced from Kaggle, contains product and user review information specific to items sold on Sephora's platform. This dataset includes widely used cosmetic products from around the world, with various metadata such as category, ingredients, and ratings. Below is an exemplar data dictionary of important values utilized in the analysis of the data.

|  |  |
| --- | --- |
| **Data Specification** | |
| Size | 4.27 GB |
| Sephora Products | <https://www.kaggle.com/datasets/nadyinky/sephora-products-and-skincare-reviews> |
| Global Cosmetic Products | <https://www.kaggle.com/datasets/waqi786/most-used-beauty-cosmetics-products-in-the-world> |

|  |  |
| --- | --- |
| **Sephora Reviews Data Dictionary** | |
| Product\_ID | Unique identifier for each product |
| Product\_Name | Name of the cosmetic product |
| Brand\_ID | Numeric ID representing the brand |
| Brand\_Name | Textual name of the brand |
| Loves | Count of users who have "loved" or favorited the product |
| Rating | Average customer rating (e.g., out of 5) |
| Ingredients | List of ingredients for each product |
| Price\_USD | Product price in U.S. dollars |
|  |  |
| **Top World Products Data Dictionary** | |
| Product\_Name | Name of the global beauty product |
| Brand\_Name | Name of the brand offering the product |
| Category | Type of product (e.g., serum, cleanser, foundation) |
| Price\_USD | Listed price in U.S. dollars |
| Rating | User rating score |
| Reviews | Total number of reviews |
| Skin\_Type | Intended skin type (e.g., oily, dry, sensitive) |
| Ingredients | Key ingredients used in the product |

**4. Process**

The visual workflow provided outlines the complete data engineering and analytics pipeline employed in the project, which focuses on analyzing skincare and cosmetic product data from both Sephora and a global collection of beauty brands. This multi-step process begins with the acquisition of raw datasets from external repositories such as Kaggle. The Sephora dataset in particular was fragmented into several CSV files, requiring a preliminary step of file consolidation using Python. Through the Pandas library, the data was appended into a unified structure to facilitate further processing and ensure schema consistency across datasets.

Following the consolidation, the unified files were uploaded into the Hadoop Distributed File System (HDFS), a scalable and fault-tolerant storage framework ideal for processing large volumes of data. The use of HDFS enabled distributed computing, which significantly improved the performance of query execution, data transformation, and large-scale file handling. Hive was used as the primary querying tool within the Hadoop ecosystem, offering SQL-like operations to clean, transform, and analyze the data in a structured format.

The wrangling phase involved handling missing values, identifying and correcting inconsistencies, and preparing the datasets for integration. At this point, both datasets were evaluated for structural alignment, which necessitated renaming various columns to create a standardized schema. For example, columns such as “product\_name\_x” were renamed to a simplified and more interpretable format like “Product\_Name.” This renaming process was essential for merging and comparing the Sephora dataset with the global product dataset in a meaningful way.

Columns that did not contribute analytical value or were redundant were then dropped. These included fields like author IDs, timestamps, and other metadata not relevant to pricing, reviews, ratings, or ingredient analysis. Subsequently, value normalization was applied across all string-based fields to ensure uniform casing and remove extraneous spacing, allowing for more reliable filtering, grouping, and aggregation. This type of normalization is critical in preparing the data for cross-dataset comparison and reducing the risk of analytical errors due to inconsistent formatting.

After normalization, the cleaned datasets were exported from Hive using the INSERT OVERWRITE DIRECTORY function and stored back in HDFS as structured CSVs. These exported datasets served as the foundation for the final stage of the project: data visualization. Using tools such as Power BI and Python libraries like Matplotlib and Seaborn, the team created visual narratives that explored patterns in pricing, product popularity, rating distributions, ingredient frequency, and regional brand performance. These visualizations enabled business-oriented insights and comparisons between Sephora-exclusive products and globally recognized cosmetics.

This pipeline reflects a modern, scalable approach to data analysis that leverages big data tools for ingestion and processing, scripting languages for data wrangling and cleaning, and business intelligence platforms for storytelling. It demonstrates how technical architecture, such as distributed computing with Hadoop and Hive, can support data harmonization and empower large-scale analytical projects.4

**Project Worklfow**

**5. Data Cleansing**

The data cleansing process carried out in this project was central to ensuring the accuracy, consistency, and interpretability of the analysis across two large datasets: the Sephora product reviews dataset and the Top World Products dataset. Given the size of the combined data, exceeding 4 GB, it was critical to adopt a methodical and technically robust approach to data preparation. The process began with file consolidation. The Sephora dataset, originally split across multiple CSV files, was appended using Python's pandas library into a unified table. This consolidation step was immediately followed by a key-based join operation, combining product-level metadata with review-level details via the Product\_ID field. This created a comprehensive dataset ready for ingestion into Hive via Hadoop Distributed File System (HDFS), allowing for scalable processing.

Once within Hive, schema harmonization took place. Column names across both datasets were normalized for consistency and clarity. Ambiguous suffixes like \_x or \_y which were remnants of previous merges, were removed, and attributes were renamed using standardized, readable formats (e.g., product\_name\_x became Product\_Name). This schema alignment was essential not only for the integrity of downstream queries but also for enabling meaningful comparisons between Sephora-exclusive products and globally recognized cosmetics.

The next phase of the cleansing process involved the targeted removal of irrelevant columns. Fields such as timestamp data, author IDs, and secondary product metadata were excluded to streamline the dataset and focus the analysis on price, rating, review count, brand, and ingredient-related features. This step reduced data noise and computational overhead while preserving the variables most relevant to the project’s research questions.

Missing values were addressed with a nuanced strategy. For categorical fields such as Skin\_Type and Ingredients, default values were assigned using domain logic, “normal” was used as a default for Skin\_Type based on its frequency in the dataset. Boolean fields like Is\_Recommended were filled with a neutral value of 0, indicating no recommendation in the absence of data. These decisions balanced data completeness with integrity, a critical tradeoff in large-scale analytics.

Further, all string-based fields were normalized using .lower().strip() functions in Python. This ensured uniformity in casing and formatting, preventing duplicative groupings in later analysis (e.g., treating “Dry”, “dry ”, and “DRY” as distinct values). The use of data type casting was equally important. Fields representing numerical values such as Price\_USD, Rating, and Reviews were explicitly converted to float or integer types to support accurate aggregation and visualization.

Finally, the cleansed datasets were exported from Hive using the INSERT OVERWRITE DIRECTORY function, creating well-structured CSVs ready for consumption in Power BI and Python visualization libraries. These final datasets enabled a variety of comparative analyses, from ingredient frequency distribution to pricing trends across regions. The rigor of the data cleansing process laid the foundation for all subsequent analytical insights and was instrumental in ensuring that the findings were not only statistically valid but also actionable.

This cleansing process exemplifies the importance of data quality in the broader data science and information systems community. Without such preprocessing, even the most sophisticated models and visualizations can yield misleading conclusions. As data increasingly drives decision-making across industries, the methodologies employed in this project reinforce the necessity of thoughtful, scalable, and reproducible data cleaning practices.

**6. Conclusion**

This project presented a comprehensive analysis of consumer trends in the global skincare and cosmetics market by leveraging large-scale datasets sourced from Sephora’s platform and international beauty product listings. Through a methodical pipeline built on distributed computing infrastructure, including Hadoop Distributed File System (HDFS), Hive, and Python, the project demonstrated how raw, unstructured data can be transformed into meaningful business intelligence. From ingesting and cleaning over 4 GB of product and review data to harmonizing schemas and visualizing key insights, each stage of the process highlighted the power of data engineering and analytics in uncovering patterns of consumer behavior, product performance, and market dynamics.

The work is both technically rigorous and highly relevant to the business, data science, and information systems communities. From a business perspective, the findings offer actionable insights into pricing strategies, consumer sentiment, and ingredient trends that could influence product development and marketing decisions. For instance, the ability to compare product ratings across regions or identify the most favored ingredients allows companies to better align their offerings with consumer preferences. In terms of data science, the project illustrates how advanced data wrangling, normalization, and data analysis can be applied in real-world scenarios to generate empirical evidence from messy, heterogeneous datasets. For the information systems field, the integration of Hive SQL querying with cloud-based data pipelines models an enterprise-level workflow, showing how analytical ecosystems can be scaled to support big data initiatives in retail and e-commerce domains.

Ultimately, what was built was not just a set of charts or tables, but a reusable and scalable pipeline that integrates structured data modeling, business-oriented metrics, and visual storytelling. The project provided deep insights into pricing distributions, ingredient popularity, and product performance, while also enabling cross-market comparisons. From this work, key lessons emerged: the importance of schema alignment in multi-source analysis, the value of string normalization for accurate aggregations, and the critical role that distributed systems play in processing data at scale. This project not only demonstrates the value of data analytics in the beauty industry, but also reinforces the broader importance of clean, interpretable, and well-modeled data in any domain seeking to derive strategic insights from consumer behavior.

### References

Works Cited

[1] Khalid, Usama. SemesterProject\_NLP. GitHub, 2022, <https://github.com/UsamaK98/SemesterProject_NLP>.

[2] Monfared, Melissa. “Skincare Products EDA & Sentiment Analysis.” Kaggle, 2022, <https://www.kaggle.com/code/melissamonfared/skincare-products-eda-sentiment-analysis>.

[3] Putri, Dzakiah. “Skincare Products Recommendation System.” Medium, 2022, <https://medium.com/@dzakiahptr/skincare-products-recommendation-system-7d3b2c94884c>.

[4] Dean, Jeffrey, and Sanjay Ghemawat. “MapReduce: Simplified Data Processing on Large Clusters.” Communications of the ACM, vol. 51, no. 1, 2008, pp. 107–113.

[5} Sephora Products Raw Dataset: <https://www.kaggle.com/datasets/nadyinky/sephora-products-and-skincare-reviews>

[6} Global Cosmetic Products Raw Dataset: <https://www.kaggle.com/datasets/waqi786/most-used-beauty-cosmetics-products-in-the-world>

[7] GitHub Repository: <https://github.com/yochris723/CIS5200_Project>