Identification of Hazardous Ingredients in Cosmetic Products for Reformulation

Olivia Folly-Gah

10/18/2024

Introduction

My project has focused on identifying hazardous ingredients in cosmetic products that are known or suspected to cause cancer, birth defects, or reproductive harm. Through the analysis of a dataset provided by the California Department of Public Health (CDPH), I aim to identify and highlight the most frequently used harmful chemicals present in these products. The broader goal is to provide insights that will assist in the reformulation of cosmetic products to be safer. This report outlines the progress made in the analysis so far, the challenges I encountered, and the steps I plan to take moving forward.

Summary of Work Done

During this reporting period, I concentrated on cleaning and preprocessing the dataset, which consists of 114,635 rows and 22 columns. The primary focus was on products categorized as skin or hair care items, filtering out other irrelevant categories. I cross-referenced hazardous chemicals in these products with known lists of chemicals that are linked to carcinogenicity, reproductive harm, and birth defects. To visualize the findings, I created several graphs illustrating the distribution of hazardous chemicals across different product categories, providing a clearer understanding of how widespread these harmful substances are in skin and hair care products.

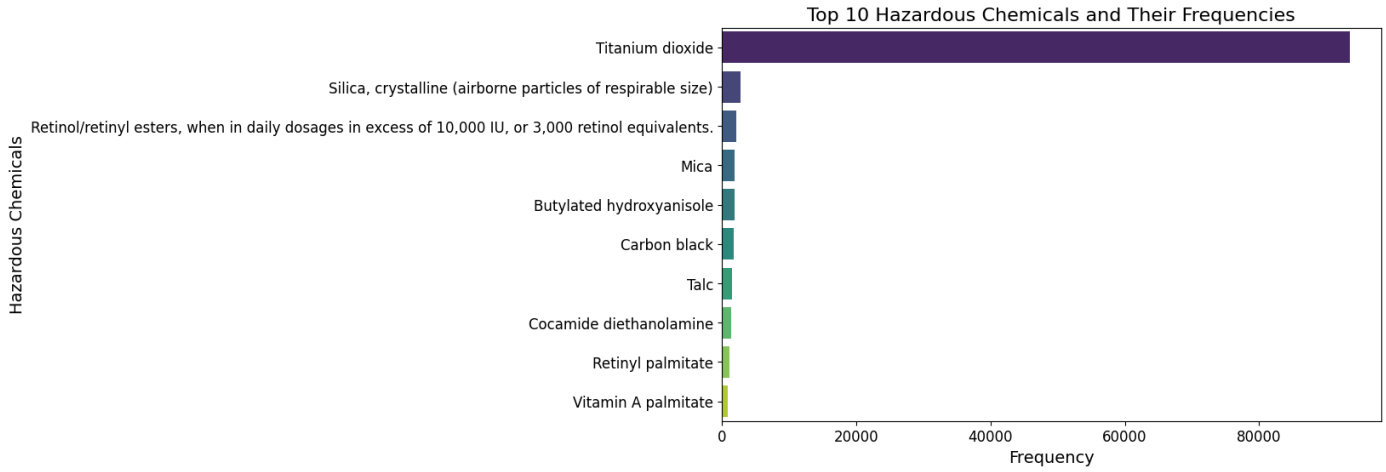

Figure 1: Distribution of hazardous ingredients across products.

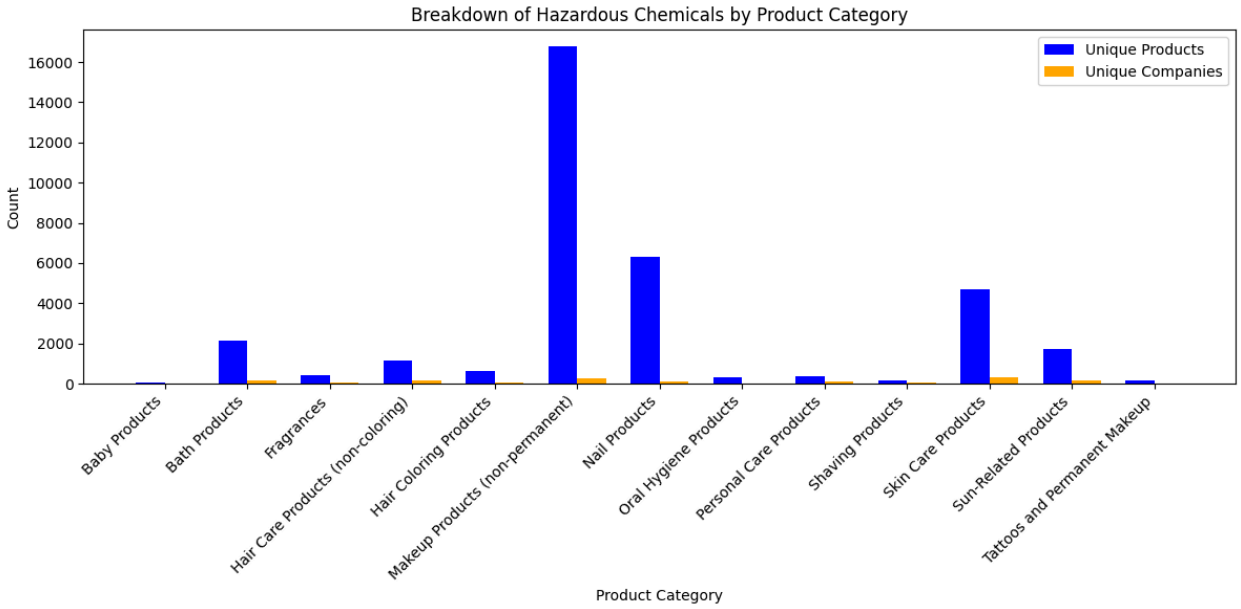

Figure 2: Breakdown of hazardous chemicals by product category.

Progress and Milestones

I have successfully preprocessed and cleaned the dataset by standardizing key columns such as product names and company names. For example, redundant names were merged to avoid duplication issues (e.g., "Alberto Culver USA, Inc." was standardized to "Alberto Culver, Inc."). Additionally, I identified the hazardous chemicals present based on their association with carcinogenicity or reproductive harm. Products related to skin and hair care were filtered for more focused analysis.

Going forward, I need to conduct a more detailed analysis of the most common hazardous chemicals found in the dataset. Another key task is to research safer alternatives to these hazardous ingredients and begin formulating recommendations for product reformulation. Additionally, I need to validate my findings by cross-referencing the chemicals with existing safety regulations and cosmetic safety studies.

Problem-Solving and Challenges

One of the main challenges I faced was the lack of data specific to products targeting melanated skin and curly/coily hair. The dataset does not include clear labeling or identification for these product categories, which has made it difficult to conduct a targeted analysis for this demographic. This is important because individuals with these hair and skin types may be more vulnerable to harmful chemicals due to the higher frequency of certain ingredients in products targeted toward them. Additionally, there were missing data points related to product discontinuation dates or reformulation details, which affected my ability to track when products were updated to remove hazardous chemicals.

Approaches and Solutions
To address these challenges, I focused my analysis on the broader skin and hair care categories, analyzing trends within these products. For missing data, I applied imputation methods where possible and retained records even if some information, such as the brand name, was missing. This approach improved the overall completeness of the dataset and allowed for a more thorough examination of hazardous chemicals across product types.

By broadening the analysis to general skin and hair care categories, I was able to develop a more comprehensive understanding of the prevalence of hazardous chemicals. This allowed for a more accurate identification of common harmful ingredients and improved the quality of the insights generated from the dataset.

Technical Depth and Accuracy
The technical work so far has involved extensive data cleaning and preprocessing. I addressed missing values, standardized company and product names, and filtered out irrelevant categories. This ensured that the data was ready for deeper analysis. Exploratory data analysis (EDA) techniques were applied to visualize trends in hazardous ingredient use across the product categories.

While this project is currently focused on data analysis, I plan to implement predictive modeling in the next phase to predict the likelihood that a product contains hazardous ingredients based on its ingredient list. Possible models to explore include the Random Forest Classifier, which is well-suited for handling mixed data types and provides insight into feature importance. Gradient Boosting Machines, such as XGBoost, may also be used for their high accuracy and ability to handle imbalanced data. I have also come across Naive Bayes classifiers in research papers related to chemical toxicity identification, which I may consider for additional experimentation.

Future Plans and Goals
Over the next two weeks, my goal is to conduct a deeper analysis of the most commonly used hazardous chemicals in the dataset. I will also begin researching safer alternatives to these chemicals, aiming to provide concrete reformulation recommendations that can reduce the risk of harm to consumers. I plan to explore the feasibility of implementing predictive models to identify potentially harmful products based on their ingredient lists. Furthermore, I intend to address the gap in data related to products designed for melanated skin and curly/coily hair by seeking additional external data sources or studies. For example, I recently attended a seminar hosted by Seppic titled “Navigate the Textured Hair Needs and Formulations with Proven-Efficacy Solutions,” which may provide relevant insights into this aspect of the project.