PROJECT REPORT

The Beauty of Safety - A Comprehensive Analysis of Chemicals in Cosmetic Products

AUTHORS:

Aditya Sairam Govindan

Anjana Deivasigamani

Jennisha Christina Martin

Naveen Kavitha Gunasekaran

Priyadharshan Sengutuvan

SUMMARY

These days, there's a growing concern over the safety of chemicals in beauty products, sparking interest from both the public and scientific communities. With a wide range of beauty products available, the need for a thorough evaluation of their chemical constituents has never been more pressing. The main objective of our project is to analyze the chemical composition of beauty products thereby identifying their toxicity rate to assess potential health implications to enhance public awareness about the hazardous ingredients used in the preparation of such beauty products and inform safer consumer choices. The “Chemicals in Cosmetics” dataset that we have selected provides insights into the diversity of cosmetic products available in the market and comprises detailed listings of the chemicals used in the cosmetic products and other product information such as the product name, company, brand, category, and reporting dates. In our project, we will be identifying and highlighting potentially harmful chemicals, delving into their types, assessing how frequently they have been used in the preparation of personal beauty care products individually, and keeping an eye out for any unusual trends or irregularities in the data, as well as exploring how the most common chemical ingredients used in the products, might impact our health.

The dataset we have chosen for analysis is sourced from data.gov. This dataset provides a comprehensive overview of the cosmetics and personal care products available in the market, containing chemicals that are known or suspected to cause cancer, birth defects, or other developmental or reproductive issues. It covers a wide range of products and a broad spectrum of chemical ingredients used in those products like diethanolamine, titanium dioxide, cocamide, etc, which are known to be hazardous. Moreover, this dataset encompasses detailed information such as product names, company and manufacturer details, brand names, categories, Chemical Abstracts Service (CAS) numbers, names of reported chemicals, the count of reported chemicals per product, and relevant reporting dates, including information on product discontinuation, thereby being

instrumental in our objective to dissect and analyze the chemical composition of cosmetics and to understand the prevalence and potential health risks associated with hazardous chemicals used in the beauty products being sold.

METHODS

About the Data:

The data reflects information that has been reported to the California Safe Cosmetics Program (CSCP) in the California Department of Public Health (CDPH). The primary purpose of the CSCP is to collect information on hazardous and potentially hazardous ingredients in cosmetic products sold in California and to make this information available to the public.

Data Import and Preliminary Analysis:

The data is imported into RStudio by using the read.csv command. An overview of the data is obtained using the head() function and the summary() function in R. A brief look at the dataset gives us some basic insights.

1. The dataset consists of 114635 rows and 22 columns.

2. The following columns contain only numeric data: CDPHId, CompanyId, PrimaryCategoryId, SubCategoryId, CasId, ChemicalId, ChemicalCount.

3. The data frame contains 254 duplicated rows which can be removed.

4. The following columns have a lot of NaN values that need to be handled.

Column	**Percentage**
ChemicalDateRemoved	97
DiscontinuedDate	88
CSF	29
CSFId	29

Table. 1 Number of Null Values

5. The following fields have data strings representing date formats: ***InitialDateReported, MostRecentDateReported, ChemicalCreatedAt, and ChemicalUpdatedAt***.

Data Cleaning:

As reported in the previous step, there are a lot of inconsistencies in the data that need to be cleaned and formatted so that the analysis and the modeling steps render accurate results.

1. The first step is to drop all the duplicate rows that we identified in the previous step. This can easily be done in R using the dplyr package.

2. The next step is to deal with the columns that mostly have NaN data. For example, ChemicalDateRemoved contains about 97% NaN data, suggesting that this column can be removed. Similarly, 'DiscontinuedDate' is also removed.

3. The rest of the NaN values can be removed by using the na.omit() function in R.

4. After these basic data cleaning steps, the data frame shape is now the following: 76584 rows and 20 columns.

5. To make our final model more accurate, we remove the following columns as they do not contribute much to the final model or further analysis— CDPHId, CSFId, CompanyId, PrimaryCategoryId, ChemicalId, SubCategoryId, CasNumber, CasId.

Further Analysis and Feature Engineering:

Now the data is in a presentable format, but it requires further analysis and feature engineering. The following steps are taken to make it more accurate.

1. The columns MostRecentDateReported_Year and InitialDateReported_Year can be composed into a single column by taking the time difference between the two. This can help us in further analysis, making our data frame easier to navigate.

2. Similarly, the ChemicalUpdatedAt_Year and ChemicalCreatedAt_Year give information about the timeline for a chemical and hence can be composed into a single column by taking the time difference between the two.

3. Label Encoding helps to convert categorical data into numeric data. This can help in applying the machine learning model. This can be achieved using the tidyverse package.

4. The information for Toxicity Rating and Side Effects is obtained by referencing websites and packages online. This data is retrieved from https://www.nlm.nih.gov/toxnet/index.html, using the chemical names provided by the CASId in our dataset

5. The Toxicity Rating column represents how toxic a given product is based on the following factors: The toxicity of a chemical, Usage Frequency, and the number of chemicals present. This column is divided into buckets of 10 from 0-100.

Data Visualization and EDA:

After performing the above steps, Visualization is now performed to get deeper insights into the clean data at hand.

1. Top 20 chemicals used in industry: The visualization shows that the chemical Titanium dioxide is used extensively in the chemical industry surpassing all the chemicals significantly and standing at a frequency of more than 75000. It can also be seen that Silica, Crystalline, and Mica are also considerably used.

2. Annual reporting frequency for top 20 chemicals: This visualization shows which chemicals have been used over the years and provides insights into customer preferences and how the industry is handling the same. One interesting observation is that cocamide diethanolamine has been discontinued since 2016 and is not used as of now. A deeper investigation shows that the said chemical is included in Proposition List 65, as it has the possibility of causing cancer. There is a general dip in all the chemicals during the year 2012-2016. This is because, The California Office of Environmental Health Hazard Assessment added cocamide DEA to the list in 2012, requiring manufacturers to either warn consumers about it on labels or remove it from their products.

3. Distribution of toxicity rating: This bar plot shows the distribution of toxicity rating over the years. It can be seen that most of the products are rated at around 25, suggesting that they are not hazardous unless or otherwise used in huge quantities. But, there are a few products which are rated at 75% and even 90% suggesting that a good amount of caution is required before using the said product.

Modeling and Predictive Analysis :

When it comes to predictive analysis, we are trying to achieve two goals. To predict the toxicity rating of a product given its parameters and to predict the number of chemicals that might be used in a product.

1. To predict the toxicity rating, we used the Random Forest classifier. This classification algorithm by using a test-train split of 20-80. We were able to achieve an accuracy of about 97%.

2. To predict the chemical count, we used a combination of regression algorithms ranging from Linear Regression to the XGBoost algorithm. The LightGBM Regression model is the best-performing model among the ones that have been tested, standing at an accuracy of about 80.23%.

In summary, the LightGBM Regression model outperforms other models in terms of both cross-validation and test set performance, making it the best choice for predicting the Chemical Count in this dataset.

RESULTS

The following are the results that we found to be fascinating from our analysis:

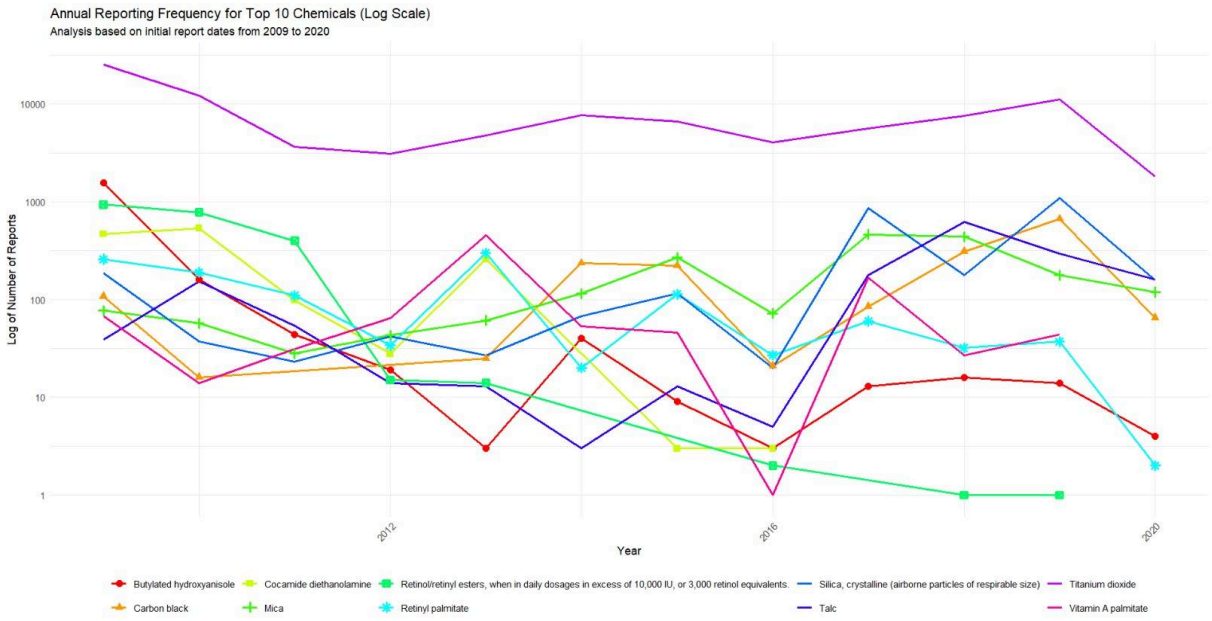

Fig 1. Annual Reporting Frequency for Top 10 Chemicals

This visualization (Fig.1) illustrates the yearly submissions of reports for the top 10 chemical substances that were frequently reported in cosmetic products between 2009 and 2020, with the data presented on a logarithmic scale. Overall, it shows the trends and fluctuations in the number of reports for each chemical reported annually in the top 10 list, across the years. From the graph below, it can be seen that the reporting frequency of the chemical titanium dioxide has been the highest when compared to the other chemicals, over the period, indicating that the use of this chemical in the beauty care products has been a concerning factor. Moreover, another noticeable trend that we observed was

that there was a sudden spike in the reports for Retinyl Palmitate in the middle of the year 2012, which makes it stand out compared to its otherwise fluctuating reporting frequency. Whereas, for the remaining chemicals such as Silica (crystalline), Retinol/retinyl esters, and Talc represented by the blue, green, and purple lines respectively—the frequency of reporting remained consistently low despite certain fluctuations over the years, indicating that their use in beauty products has been more or less stable throughout the given period. This provides valuable insights into how reporting trends evolve, potentially reflecting changes in usage patterns, regulatory requirements, and the detection of hazardous substances. This information could indeed be beneficial for stakeholders to identify any unusual trends that may require further investigation or action.

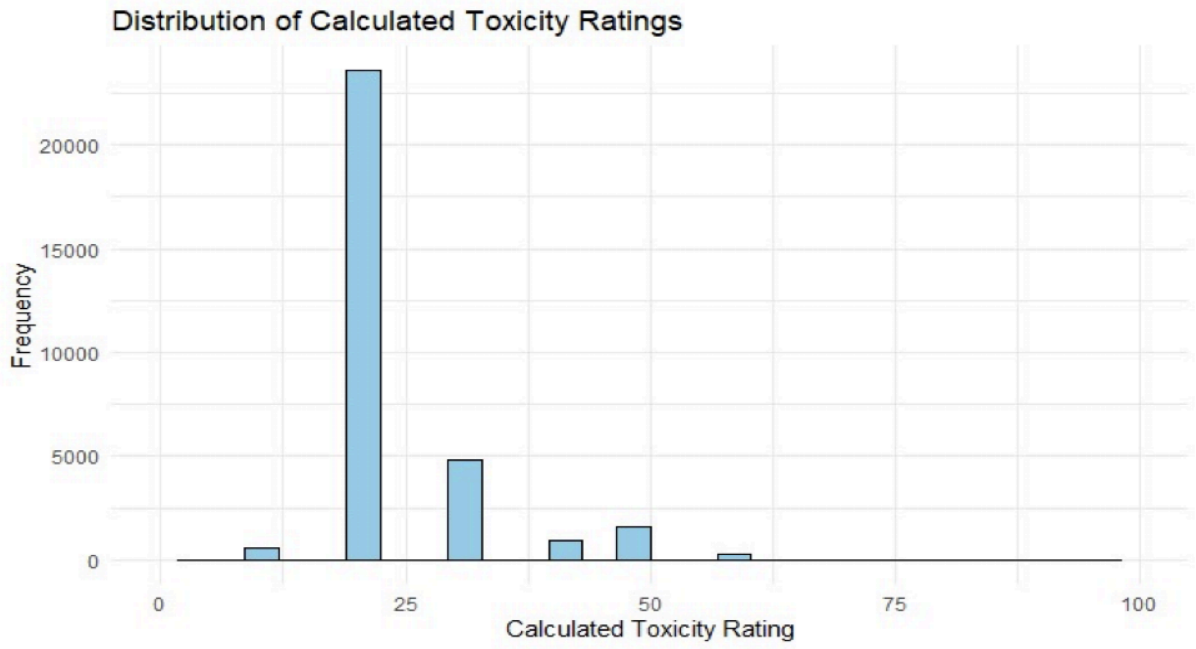

Fig 2. Distribution of Calculated Toxicity Rating

In Fig. 2, the visualization illustrates the distribution of toxicity ratings calculated for the cosmetic products sold in the market between the years 2009 and 2020. From the graph, it is clearly evident that the distribution is heavily skewed to the left, indicating that there are a considerable number of chemical substances having moderate to high toxicity ratings, while a significant proportion of the substances have lower toxicity ratings. This underscores the potential necessity for tighter regulations and enhanced surveillance of substances that possess elevated toxicity levels. This visualization could offer valuable perspectives for stakeholders, such as regulatory agencies or manufacturers, as it has the potential to shape risk evaluation and the establishment of safety guidelines for the use of chemicals. Overall, this visualization is useful for understanding the overall toxicity profile of a collection of chemical substances used in the preparation of beauty products.

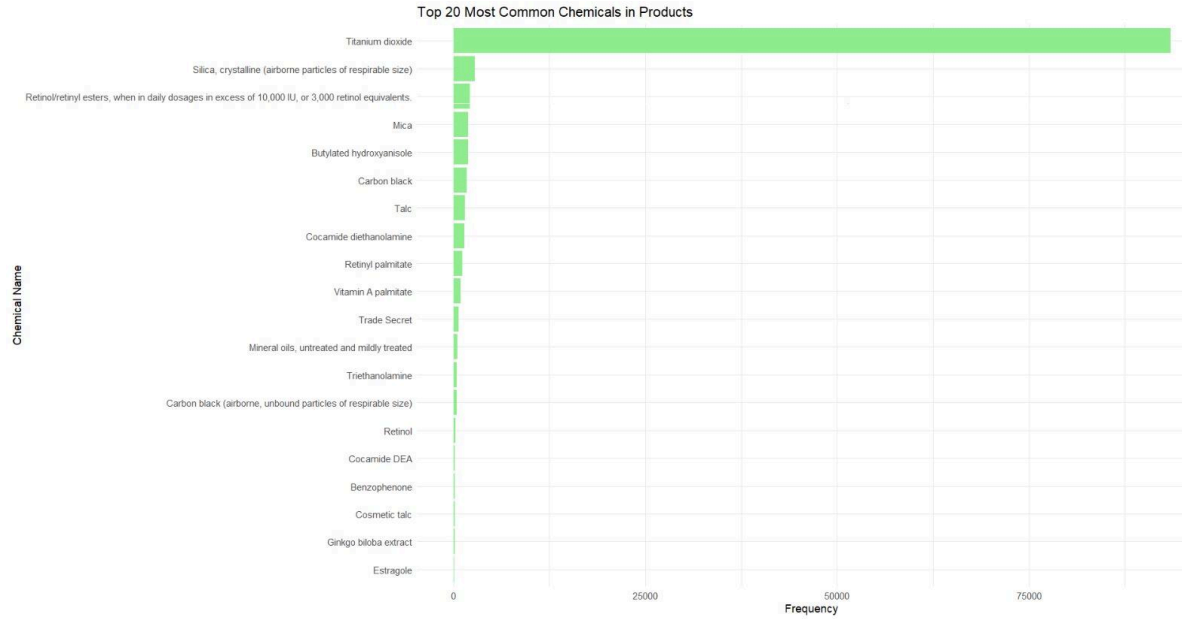

Fig 3. Top 20 Most Common Chemicals

In Fig. 3, This bar chart illustrates the frequency of the top 20 most common chemical compounds utilized in the making of the personal beauty care items, across the dataset. Overall, it can be seen that a chemical named titanium dioxide has been used in most of the makeup and skin care products when compared to the other chemical substances, titanium dioxide, although commonly used, can be hazardous if overexposed. Excessive exposure can potentially lead to DNA damage and chromosomal abnormalities, which may have serious health consequences, including death. The main purpose of this visualization is to enable consumers to quickly spot the chemicals that are most and least commonly used in the beauty care products that they purchase for personal/professional use.

DISCUSSIONS

The analysis of cosmetic product data provides valuable insights into the prevalence of potentially harmful chemicals, such as Titanium dioxide, Silica, and crystalline, and their impact on consumer safety. Our rigorous data analysis and visualization uncover trends in chemical usage over time, enabling us to prioritize consumer health.

These insights have implications for various stakeholders, including consumers, regulatory agencies, cosmetic manufacturers, and public health organizations. Consumers benefit from heightened awareness, enabling them to make informed purchasing decisions and advocate for safer products. Regulatory agencies have the opportunity to leverage these insights to enact policies that prioritize public health and enhance product safety standards. Cosmetic manufacturers can use the findings to reformulate products, prioritize safer ingredients, and enhance transparency in labeling, aligning with

consumer demand and regulatory requirements. Additionally, public health organizations can leverage the findings to advocate for comprehensive regulatory reforms, fostering a culture of safety and accountability within the cosmetics industry.

Moreover, our results facilitate better-informed decision-making processes at both individual and systemic levels. Consumers can navigate the cosmetics market more effectively, armed with knowledge about product safety. Regulatory agencies can integrate our findings into policy-making processes, strengthening oversight mechanisms and ensuring regulatory compliance. Cosmetic manufacturers can optimize product formulations based on identified trends, enhancing both safety and consumer trust. Moving forward, improvements in data cleaning techniques, deeper analysis, and collaboration with stakeholders can further enhance the relevance and impact of our efforts in promoting consumer safety and regulatory compliance within the cosmetics industry. Through ongoing refinement and collaboration, we aim to contribute significantly to the advancement of consumer safety and regulatory practices in the cosmetics sector.

STATEMENT OF CONTRIBUTIONS

Data Import and Preliminary Analysis: Aditya Sairam Govindan, Naveen Kavitha Gunasekaran

Data Cleaning: Priyadharshan Sengutuvan

Further Analysis And Feature Engineering: Anjana Deivasigamani

Data Visualization and EDA: Jennisha Christina Martin

Modeling and Predictive Analysis: Aditya Sairam Govindan, Naveen Kavitha Gunasekaran

REFERENCES

[1] Nohynek, G.J., Antignac, E., Re, T. and Toutain, H., 2010. Safety assessment of personal care products/cosmetics and their ingredients. Toxicology and applied pharmacology, 243(2), pp.239-259.

[2] California Department of Public Health (CDPH), Safe Cosmetics Program Chemical Database.

[3] Environmental Working Group's (EWG) Skin Deep Cosmetics Database.

[4] U.S. Food and Drug Administration. "Cosmetics - Safety and Regulation".

[5] Cosmetic Ingredient Review (CIR) Compendium.