Chemicals in Cosmetics.

Technical Case Study

2nd/3rd Week of December

Agenda

- Introduction & Objective
- Data Dictionary & Overview
- Data Pre-Processing & Cleaning
  - Missing values imputation
  - Feature Engineering
- Exploratory Data Analysis (EDA).
  - Univariate and Multivariate Analysis
- Key Insights, and Recommendations

Introduction & Objective

Technical Case Study

Chemical in Cosmetics Dataset:

These data reflect 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.

For all cosmetic products sold in California, the California Safe Cosmetics Act (“the Act”) requires the manufacturer, packer, and/or distributor named on the product label to provide to the CSCP a list of all cosmetic products that contain any ingredients known or suspected to cause cancer, birth defects, or other developmental or reproductive harm.

The data table consists of label names of cosmetic/personal care products, company/manufacturer names, product brand names, product categories, Chemical Abstracts Service registry numbers (CAS#) of the reported chemical ingredients, names of reported chemical ingredients, the number of reported chemicals for each product, and dates of reporting, product discontinuation or reformulation if applicable.

Assignment:

- Understanding the Data
- Preprocessing and Cleaning
- Exploratory Data Analysis (EDA)
- Key Insights and Analysis

Data Dictionary

Column Name	Description
CDPHId	Likely an identifier associated with a specific regulatory or health-related database. “CDPH” often stands for the "California Department of Public Health"
ProductName	The name of the product being referenced. This is generally the commercial or common name used in the marketplace.
CSFId	"Chemical Source File" ID. In many regulatory contexts, “CSF” is used to denote a record or ID related to the composition of a product or a formula associated with it.
CSF	The actual Confidential Statement of Formula (if provided) or a textual reference to a formula listing. It may also be a code indicating the type of chemical composition
CompanyId	A unique identifier for the company associated with the product. Useful for joining or linking this data to a separate table of company information.
CompanyName	The name of the company that manufactures or distributes the product.
BrandName	The brand under which the product is sold, which might differ from the parent company’s name if they own multiple brands.
PrimaryCategoryId	A unique identifier for the product’s primary category.
PrimaryCategory	The textual name or label of the product’s primary category (e.g., "Personal Care").
SubCategoryId	A unique identifier for a more specific classification that falls under the primary category. For instance, if the PrimaryCategory is "Personal Care," a SubCategory could be "Shampoo."
SubCategory	The textual name of the sub-category (e.g., "Shampoo").
CasId	An internal or database-specific ID tied to a chemical. This may not be the same as the CAS Registry Number, but rather a unique identifier in the dataset.
CasNumber	The CAS (Chemical Abstracts Service) Registry Number, a unique numerical identifier assigned to every chemical substance reported in the open scientific literature. It is commonly used to identify a chemical unambiguously.
ChemicalId	A database-specific unique identifier for a particular chemical. This might be used to link to a chemical details table.
ChemicalName	The common or standardized name of the chemical in the product. Often corresponds to the IUPAC name or another widely recognized chemical name.
InitialDateReported	The first date when this product-chemical combination was reported. It might correspond to when it was first introduced, registered, or discovered in the dataset.
MostRecentDateReported	The latest date on record when this product-chemical information was updated or confirmed. Useful for historical tracking of chemical usage in the product.
DiscontinuedDate	If applicable, the date when the product was discontinued or the chemical was no longer used in the product.
ChemicalCreatedAt	The timestamp or date when the chemical record was created in the database. This is a metadata field used for internal data management.
ChemicalUpdatedAt	The timestamp or date of the most recent update to the chemical record. Another metadata field important for audit and version control.
ChemicalDateRemoved	The date the chemical was removed from the dataset or from the product’s formulation. This might correspond to regulatory changes, reformulation, or product retirement.
ChemicalCount	Count of how many chemicals are associated with a product, or how frequently this particular chemical appears across different products.

Dataset Overview

SIZE

114K x 22

Records Features

KEY FEATURES

ProductName ChemicalName ChemicalCount
PrimaryCategory SubCategory BrandName
CompanyName CASNumber CDPHId

DATA TYPES

int64	8
object	8
date	6

DATA QUALITY

DataFrame w/o duplicates: 114381 rows, 22 columns

Number of duplicate rows removed: 254

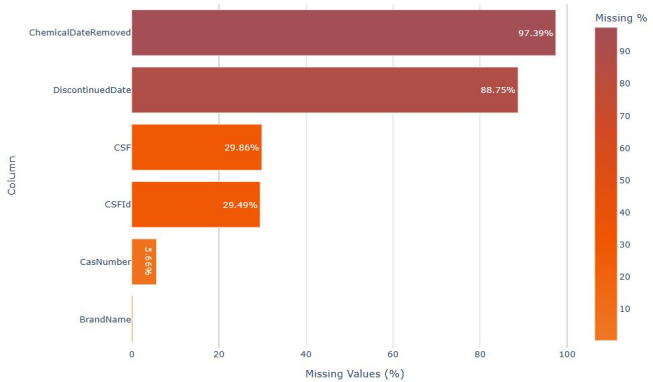
Section	Field	Description
Entities and Identifiers	CDPHId	Likely an identifier related to the California Department of Public Health or a similar regulatory body.
	ProductName	The name of the chemical product.
	CSFId & CSF	Could refer to a specific classification system or standard used for categorizing chemicals or products.
	CompanyId & CompanyName	Identifiers and names of companies that manufacture or distribute the products.
	BrandName	The brand under which the product is marketed.
Categorization	PrimaryCategoryId & PrimaryCategory	The main category under which the product falls.
	SubCategoryId & SubCategory	A more detailed classification within the primary category.
Chemical Information	CasId & CasNumber	References to the Chemical Abstracts Service (CAS) registry, which provides unique identifiers for chemicals.
	ChemicalId & ChemicalName	Identifiers and names of the specific chemicals contained in the product.
Date Fields	InitialDateReported	When the product or chemical was first reported or registered.
	MostRecentDateReported	The latest date of reporting or update.
	DiscontinuedDate	When the product was discontinued, if applicable.
	ChemicalCreatedAt & ChemicalUpdatedAt	Timestamps for when the chemical data was created and last updated in the system.
	ChemicalDateRemoved	When the chemical was removed from the product or system.
Quantitative Data	ChemicalCount	Possibly the number of chemicals in a product or the count of reports related to a chemical.

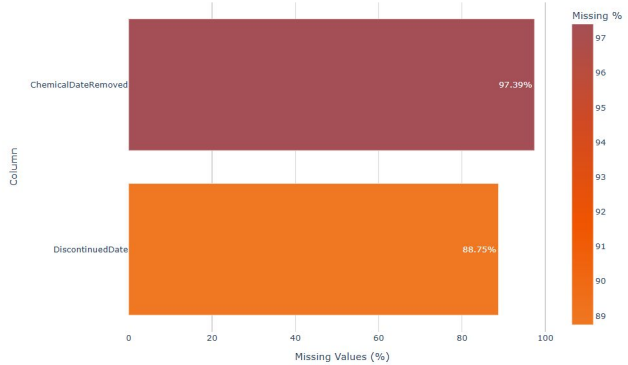
Sample Data

index	57827	5140	69208	110741
CDPHId	20995	1616	25222	40272
ProductName	Rimmel - London Salon Pro with Lycra Nail Polish	Paint My Moji-toes Red	SHISEIDO SHEER AND PERFECT FOUNDATION	Ethereal Elements - Beauty Boss Lip Gloss
CSFId	31227	0	39736	62503
CSF	Raggae Splash	Missing_CSF	I100	Pink Slip
CompanyId	171	11	1011	1332
CompanyName	Coty	OPI PRODUCTS INC.	Shiseido Americas Corporation	Stila Style LLC.
BrandName	Rimmel - London	OPI Nail Lacquer	SHISEIDO	Stila Styles LLC
PrimaryCategoryId	59	59	106	44
PrimaryCategory	Nail Products	Nail Products	Sun-Related Products	Makeup Products (non-permanent)
SubCategoryId	65	65	107	52
SubCategory	Nail Polish and Enamel	Nail Polish and Enamel	Sunscreen (making a cosmetic claim)	Lip Gloss/Shine
CasId	104	656	656	656
CasNumber	1333-86-4	13463-67-7	13463-67-7	13463-67-7
ChemicalId	31941	17492	41438	65920
ChemicalName	Carbon black (airborne, particles of respirable size)	Titanium dioxide	Titanium dioxide	Titanium dioxide
InitialDateReported	10-02-2013	09/24/2009	01/20/2015	11-01-2019
MostRecentDateReported	10-02-2013	07-01-2019	06-01-2015	11-01-2019
DiscontinuedDate	NaN	NaN	NaN	NaN
ChemicalCreatedAt	10-02-2013	06-04-2010	05/19/2015	11-01-2019
ChemicalUpdatedAt	10-02-2013	07-01-2019	05/19/2015	11-01-2019
ChemicalDateRemoved	NaN	NaN	NaN	NaN
ChemicalCount	2	1	1	1
DiscontinuedFlag	0	0	0	0
ChemicalRemovedFlag	0	0	0	0

Data Pre-Processing & Cleaning

Missing Values





Missing value imputation

`BrandName → Missing_BRAND_NAME`

`CasNumber → MISSING_CAS_NUMBER`

`CSFId → 0 (zero for missing ids)`

`CSF → Missing_CSF`

Feature Engineering

`DiscontinuedFlag` : Created based on **DiscontinuedDate,** if date exist 1 or else 0

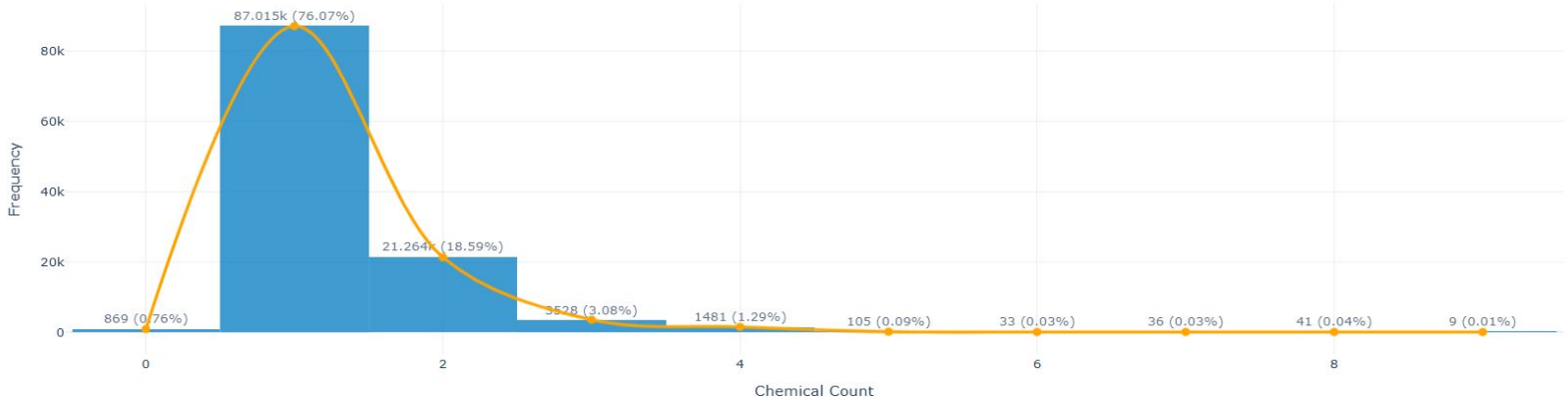
`ChemicalRemovedFlag` : Created based on **ChemicalDateRemoved,** if date exist 1 or else 0

Exploratory Data Analysis (EDA).

Uni-variate Analysis

Uni-variate Analysis

Feature : ChemicalCount



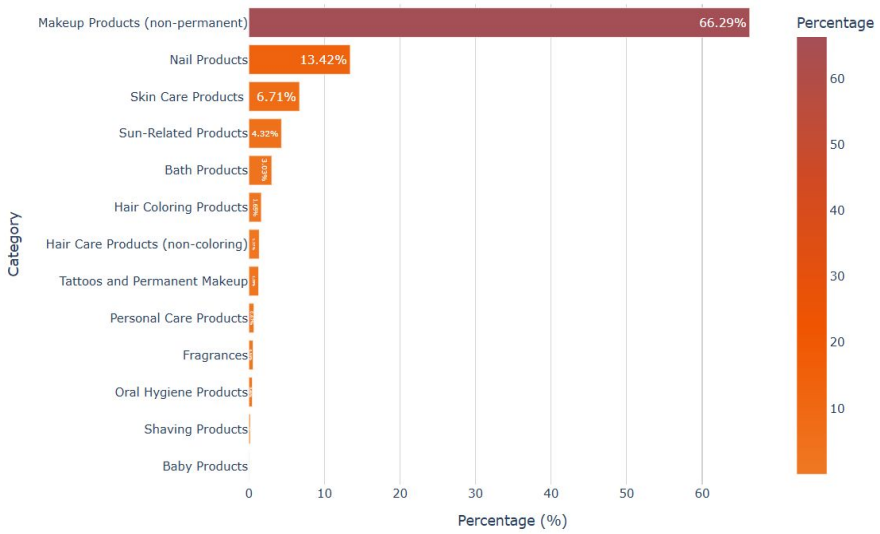
ChemicalCount	Stats
count	114381
mean	1.288982
std	0.636974
min	0
25%	1
50%	1
75%	1
max	9

ChemicalCount:

- The **relatively small standard deviation** (0.64) suggests that the values of ChemicalCount don't vary too widely from the mean of 1.29. Most of the values are close to 1, with only a few data points stretching up to a maximum of 9.
- The **maximum value** is 9, which is the highest recorded chemical count. This indicates that the ChemicalCount values range from 0 to 9, with the majority of values clustered around 1.
- **Possible Outliers**: While most values are clustered around 1, the presence of values like 9 might indicate outliers or rare events where a higher chemical count is recorded.

Uni-variate Analysis

Feature : PrimaryCategory



- Each product is categorized among 13 primary categories
- Makeup Products (non-permanent) has highest count 66.29% followed by Nail and Skin care products

Feature : ChemicalName

ChemicalName (Top 10)	Count	Prop
Titanium dioxide	93479	81.73%
Silica, crystalline	2816	2.46%
Retinol/retinyl esters,	2154	1.88%
Mica	1919	1.68%
Butylated hydroxyanisole	1888	1.65%
Carbon black	1758	1.54%
Talc	1549	1.35%
Cocamide diethanolamine	1397	1.22%
Retinyl palmitate	1181	1.03%
Vitamin A palmitate	971	0.85%

- Titanium dioxide is the most prevalent chemical, accounting for 81.73% of the total chemicals in the dataset. This indicates its significant representation relative to the others.
- Unique 123 chemicals
- The remaining 9 chemicals collectively make up 13.66% of the total chemicals.

Uni-variate Analysis

Feature : SubCategory

SubCategory (Top 10)	Count	Prop
Lip Color - Lipsticks, Liners, and Pencils	16553	14.47%
Eye Shadow	15744	13.76%
Foundations and Bases	14095	12.32%
Nail Polish and Enamel	9945	8.69%
Lip Gloss/Shine	8972	7.84%
Eyeliner/Eyebrow Pencils	5948	5.20%
Face Powders	4927	4.31%
Sunscreen (making a cosmetic claim)	4557	3.98%
Blushes	4082	3.57%
Skin Moisturizers (making a cosmetic claim)	3364	2.94%

- Unique 89 sub-categories, lip Color - Lipsticks, Liners, and Pencils accounts for 14.47% of the total count, making it the largest subcategory.
- The subcategories Eye Shadow (13.76%) and Foundations and Bases (12.32%) also represent significant portions of the total, together contributing 26.08%.

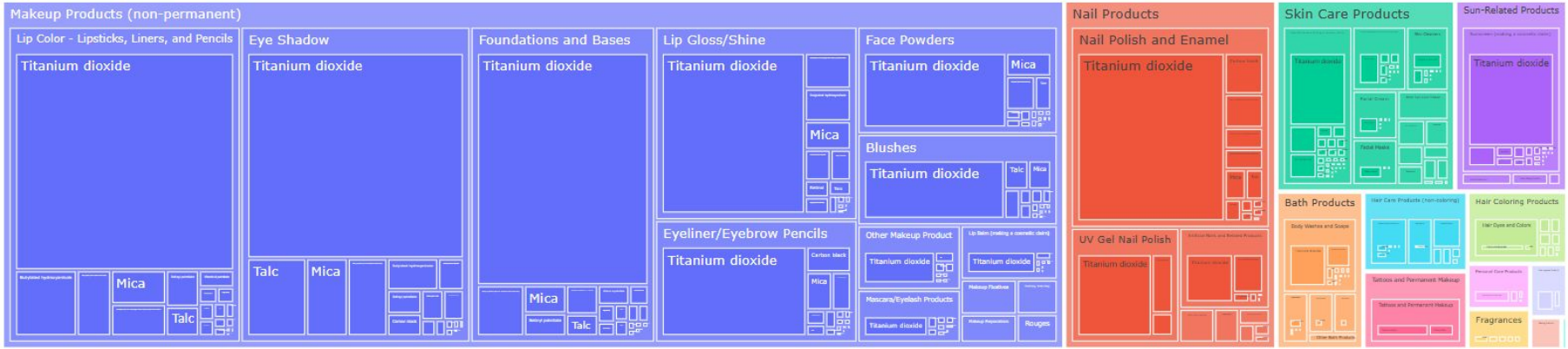
Feature : BrandName

BrandName (Top 20)	Count	Prop
SEPHORA	3394	2.97%
NYX	3227	2.82%
bareMinerals	3158	2.76%
Charlotte Tilbury	2451	2.14%
Revlon	2335	2.04%
NARS	2185	1.91%
Victoria's Secret Beauty	2106	1.84%
tarte	2101	1.84%
Sally Hansen	1834	1.60%
Sephora	1767	1.54%
No7	1749	1.53%
MAKE UP FOR EVER	1723	1.51%
CoverGirl	1711	1.50%
Parfums Christian Dior	1611	1.41%
Palladio	1519	1.33%
The Body Shop	1457	1.27%
CLARINS PARIS	1401	1.22%
Bath & Body Works	1387	1.21%
Rimmel - London	1362	1.19%
L'Oreal	1202	1.05%

2714 Unique brands

Chemical Analysis

Chemical Distribution Analysis: Categories to Specific Chemicals

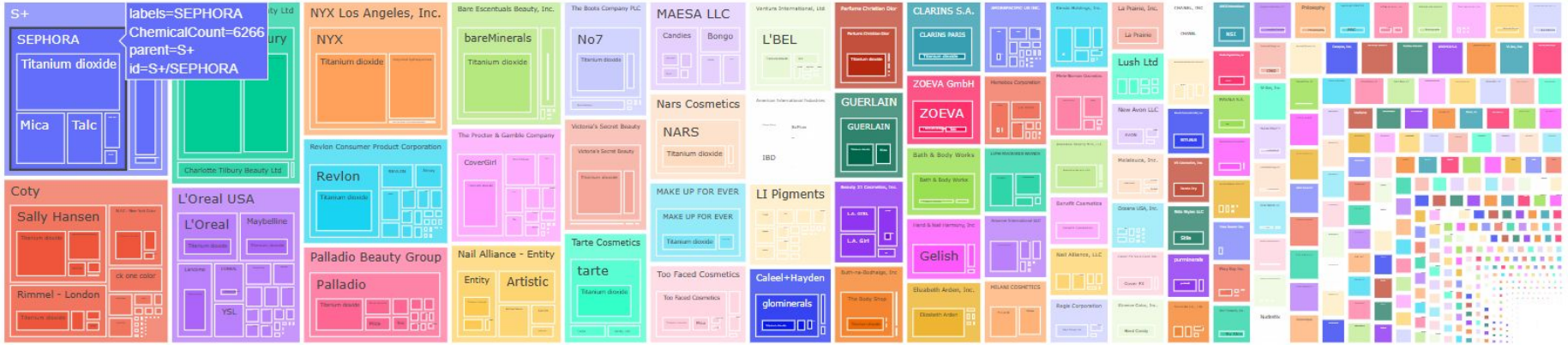


Key Observations & Insights on Chemical Usage Across Product Categories

- **Titanium dioxide** dominance across all major categories underscores its critical role in cosmetics and skincare, making it a **cornerstone** chemical in product formulations.
- The lower chemical diversity in categories like **fragrances** and **bath products** suggests simpler formulations, likely tailored to niche consumer needs or regulatory requirements
- The prominence of makeup products reflects the diversity and complexity of their formulations, driven by consumer demand for **high-performing** and **aesthetically pleasing cosmetics**.
- Makeup products represent the largest category with complex formulations, heavily relying on chemicals like Titanium dioxide, Mica, and Talc.

Chemical Analysis

Chemical Distribution Hierarchy: Companies, Brands, and Specific Chemicals



Observations & Insights from the Treemap Visualization

- **Top-Level Categories:** Each color block represents a distinct company, brand, chemical name with its size corresponding to chemicalcount.
- **Diversity Across Brands**: Some brands show a broader variety of chemicals (more subdivisions within the block), while others focus on fewer chemicals, potentially indicating specialized or niche product lines.
- **Chemical Prevalence**: Specific chemicals like "Titanium dioxide" are repeatedly mentioned across many brands, indicating their widespread use in the cosmetics or beauty industry.
- **Chemical Composition**: Sub-blocks within each brand represent individual chemicals, such as "Titanium dioxide," "Mica," and "Talc," showing their presence or significance in the products offered by that brand.

Product Discontinuation Analysis

Discontinuation by Category, Subcategory, and Company (Top 10)

Products Category	Products SubCategory	CompanyName	Products	Brands	Chemicals	ChemicalsCount	Discontinued
Makeup Products (non-permanent)	Lip Color - Lipsticks, Liners, and Pencils	The Procter & Gamble Company	639	639	639	639	577
Makeup Products (non-permanent)	Foundations and Bases	The Procter & Gamble Company	598	598	598	598	553
Hair Coloring Products	Hair Dyes and Colors	The Procter & Gamble Company	309	309	309	317	309
Makeup Products (non-permanent)	Eye Shadow	The Procter & Gamble Company	293	293	293	301	282
Makeup Products (non-permanent)	Lip Gloss/Shine	The Procter & Gamble Company	363	363	363	363	280
Makeup Products (non-permanent)	Lip Gloss/Shine	MAESA LLC	316	316	316	828	276
Nail Products	Nail Polish and Enamel	The Procter & Gamble Company	258	258	258	329	258
Makeup Products (non-permanent)	Lip Gloss/Shine	Victoria's Secret Beauty	626	626	626	802	239
Makeup Products (non-permanent)	Eye Shadow	The Boots Company PLC	379	379	379	417	181
Makeup Products (non-permanent)	Lip Color - Lipsticks, Liners, and Pencils	Too Faced Cosmetics	335	335	335	499	177

Observations & Insights from the Product Discontinuation Analysis

- **Procter & Gamble Company Leads in Discontinuations** : Dominates the list with the highest counts in products, brands, chemicals, and discontinuations, such as Lip Color (639 products, 577 discontinued) and Foundations and Bases (598 products, 553 discontinued), indicating large-scale reformulations or portfolio cleanup.
- **MAESA LLC** (828 chemicals for 316 products) and **Victoria's Secret Beauty** (802 chemicals for 626 products) highlight high formulation complexity in their product lines.
- Top companies like Too Faced Cosmetics (177 discontinued out of 335 products, **~52%**), Victoria's Secret Beauty, and The Boots Company PLC show high discontinuation rates, suggesting a focus on reformulations or phasing out product lines.

Chemicals Removed Analysis

Chemicals Removed Distribution: Category, Subcategory, and Company (Top 10)

Products Category	Products SubCategory	CompanyName	Products	Brands	Chemicals	ChemicalsCount	ChemicalRemoved
Makeup Products (non-permanent)	Lip Color - Lipsticks, Liners, and Pencils	NYX Los Angeles, Inc.	1217	1217	1217	2299	98
Makeup Products (non-permanent)	Lip Color - Lipsticks, Liners, and Pencils	Sunrider Manufacturing, L.P.	134	134	134	149	94
Nail Products	Nail Polish and Enamel	Coty	1779	1779	1779	2339	74
Makeup Products (non-permanent)	Lip Color - Lipsticks, Liners, and Pencils	MAVALA S.A.	156	156	156	308	52
Makeup Products (non-permanent)	Eye Shadow	Arcadia Beauty Labs LLC	126	126	126	120	46
Nail Products	Nail Polish and Enamel	Beauty Selectives	86	86	86	86	43
Makeup Products (non-permanent)	Foundations and Bases	Kingdom Animalia, LLC.	140	140	140	173	38
Makeup Products (non-permanent)	Lip Color - Lipsticks, Liners, and Pencils	C.F.E.B. SISLEY	87	87	87	81	36
Skin Care Products	Skin Moisturizers (making a cosmetic claim)	Vi-Jon, Inc.	158	158	158	145	35
Makeup Products (non-permanent)	Foundations and Bases	LVMH FRAGRANCE BRANDS	337	337	337	455	34

Observations & Insights from Chemical Removed Distribution:

- **Companies with Intense Chemical Removal Efforts :** Sunrider Manufacturing, L.P. (134 products, 94 chemicals removed, **~70%**) and Coty (1,779 products, 74 chemicals removed, **~4%**) highlight contrasting strategies—Sunrider focusing on intensive reformulations, while Coty targets fewer changes across its vast portfolio.
- **Nail Products Reflect High Chemical Involvement :** Nail Polish and Enamel by Coty (1,779 products, 2,339 chemical counts) and Beauty Selectives highlights significant chemical involvement, reflecting complex formulations. Nail products likely face stricter scrutiny due to safety concerns over specific chemicals.

Company and Brand Analysis

CompanyName	BrandName	Products	ChemicalsCount	ChemicalsRemoved	Discontinued
S+	SEPHORA	3394	6266	42	0
NYX Los Angeles, Inc.	NYX	3227	5546	105	0
Bare Escentuals Beauty, Inc.	bareMinerals	3158	3267	5	13
Charlotte Tilbury Beauty Ltd	Charlotte Tilbury	2451	5909	6	0
Revlon Consumer Product Corporation	Revlon	2335	2343	9	86
Nars Cosmetics	NARS	2184	2422	0	191
Victoria's Secret Beauty	Victoria's Secret Beauty	2106	2572	5	1176
Tarte Cosmetics	tarte	2101	2193	9	0
Coty	Sally Hansen	1834	2921	18	36
S+	Sephora	1767	1889	0	1
The Boots Company PLC	No7	1749	2356	30	413
MAKE UP FOR EVER	MAKE UP FOR EVER	1723	2253	7	19
The Procter & Gamble Company	CoverGirl	1681	1750	0	1538
Parfums Christian Dior	Parfums Christian Dior	1611	1753	2	11

Observations & Insights from Company & Brand Analysis:

- **Product Diversity & Chemical Usage :** S+ (SEPHORA) leads with **3,394** products and a ChemicalsCount of **6,266**, highlighting its diverse and complex formulations. NYX Los Angeles, Inc. (NYX) follows with **3,227** products and **5,546** chemicals, emphasizing its expansive product line.
- **Chemical Removal Efforts :** NYX Los Angeles, Inc. leads in ChemicalsRemoved (105), reflecting a strong focus on reformulations or clean product initiatives. S+ (SEPHORA) follows with 42, emphasizing product safety or regulatory compliance, while The Boots Company PLC (No7) shows notable efforts with 30 chemicals removed.

Key Insights and Analysis

Insights and Observations

- **Titanium Dioxide Dominance** : Titanium dioxide is the most prevalent chemical, appearing in **81.73%** of the total products across various categories, underscoring its widespread importance in cosmetics and skincare formulations.
- **Makeup Products Dominate Categories** : **66.29%** of products belong to the "Makeup Products (non-permanent)" category, making it the largest primary category. Subcategories like "Lipsticks, Liners, and Pencils" and "Eye Shadow" together account for 26.08% of total products.
- **High Discontinuation Rates in Top Companies** : Procter & Gamble leads in product discontinuations, with **639** products (**90%**) in Lip Color subcategory reformulated or discontinued. This indicates active portfolio optimization and regulatory compliance efforts.
- **Chemical Diversity Across Brands** : Brands like Sephora and NYX show high chemical diversity, with NYX leading chemical removal efforts (105 chemicals removed). This reflects a commitment to cleaner formulations and regulatory adherence.

Additional Suggestions Based on Dataset

- **Highlight Emerging Trends in Chemical Use** : Analyze year-over-year data for InitialDateReported and MostRecentDateReported to uncover: Trends in the adoption of specific chemicals & Changes in chemical composition.
- **Product and Chemical Lifecycle Analysis** :
  - Examine DiscontinuedDate, ChemicalDateRemoved, and ChemicalRemovedFlag
  - Identify chemicals and products with a short lifecycle to understand the drivers (e.g., safety concerns or market trends).
- **Emerging Chemicals** : Identify newly reported chemicals using InitialDateReported, focusing on ChemicalCounts that could represent innovation.

Thank You!

Nagesh Ch.
ncheripally@outlook.com
+91-7680919292