Metadata

Title:

Analysis of Chemicals in Cosmetic Products with a focus on Baby Products

Author:

Name: Sumuk Shashidhar

Class: Class XII A

School: Sri Kumaran Children's Home

Dataset Used

Chemicals in Cosmetics by Sumuk Shashidhar

Imports and Requirements

For this project, we need the data sets and some python libraries

In [1]:
```
import pandas as pd
import numpy as np
import plotly.express as px
import plotly.io as pio
import psutil
from IPython.display import Image
```

Reading and cleaning the data

In [2]:
```
df_original = pd.read_csv('./data/chemicals-in-cosmetics.csv')
df = df_original.drop_duplicates()
print('The original data had ', df_original.shape[0], "rows")
print('After removing duplicates, the data has', df.shape[0], "rows")
```

```
The original data had  112870 rows
After removing duplicates, the data has 112616 rows
```

Sampling the data

In [3]:
```
df.head()
```

Out[3]:

	CDPHId	ProductName	CSFId	CSF	CompanyId	CompanyName	BrandName	PrimaryCategory

	CDPHId	ProductName	CSFId	CSF	CompanyId	CompanyName	BrandName	PrimaryCategory
0	2	ULTRA COLOR RICH EXTRA PLUMP LIPSTICK-ALL SHADES	NaN	NaN	4	New Avon LLC	AVON	
1	3	Glover's Medicated Shampoo	NaN	NaN	338	J. Strickland & Co.	Glover's	
2	3	Glover's Medicated Shampoo	NaN	NaN	338	J. Strickland & Co.	Glover's	
3	4	PRECISION GLIMMER EYE LINER-ALL SHADES �	NaN	NaN	4	New Avon LLC	AVON	
4	5	AVON BRILLIANT SHINE LIP GLOSS-ALL SHADES �	NaN	NaN	4	New Avon LLC	AVON	

5 rows × 22 columns

Analysis

Let us look at the total number of chemicals that we have in our dataset

```
In [4]: df['ChemicalName'].value_counts().size
```

```
Out[4]: 123
```

It seems that we have a total of 123 chemicals in our given data

Trends and Averages

Let us see what is the average number of reported chemicals, as well as the maximum and minimum for each product

```
In [5]: df['ChemicalCount'].describe()
```

```
Out[5]: count    112616.000000
        mean          1.282402
        std           0.629696
        min           0.000000
        25%           1.000000
        50%           1.000000
        75%           1.000000
        max           9.000000
        Name: ChemicalCount, dtype: float64
```

This tells us that there are some products with no reported chemicals, and there arae some with as many as 9 reported chemicals.

However, the average seems to be around 1 chemical

Removing some bias from our observations

It doesn't make sense that some products have no reported chemicals at all, so let us closely examine what we have

In [6]:

```
df.loc[df.ChemicalCount==0].head()
```

Out[6]:

	CDPHId	ProductName	CSFId	CSF	CompanyId	CompanyName	BrandName	PrimaryCatego
31	24	White Premium Lotion Soap	NaN	NaN	181	GOJO Industries, Inc.	GOJO�	
497	333	Gentle Cleanser	NaN	NaN	71	Sunrider Manufacturing, L.P.	Kandesn	
498	334	Cleansing Foam	NaN	NaN	71	Sunrider Manufacturing, L.P.	Kandesn	
499	334	Cleansing Foam	NaN	NaN	71	Sunrider Manufacturing, L.P.	Kandesn	
500	334	Cleansing Foam	NaN	NaN	71	Sunrider Manufacturing, L.P.	Kandesn	

5 rows × 22 columns

The number of chemicals being equal to zero suggests that the chemicals were removed from the product (reported in 'ChemicalDateRemoved'). This can be verified by checking if there are NaN values in this column.

In [7]:

```
df.loc[df.ChemicalCount==0]['ChemicalDateRemoved'].isnull().max()
```

Out[7]:

```
False
```

In [8]:

```
df_n0 = df.loc[(df.ChemicalCount>0) & (df['DiscontinuedDate'].isna())]
```

The maximum number of chemicals that is reported in a product is 9. We can find these products:

In [9]:

```
df_n0.loc[df.ChemicalCount==9]
```

Out[9]:

	CDPHId	ProductName	CSFId	CSF	CompanyId	CompanyName	BrandName	PrimaryCat
60819	22212	Moisturizing Shampoo	NaN	NaN	165	Regis Corporation	Regis Design Line	

	CDPHId	ProductName	CSFId	CSF	CompanyId	CompanyName	BrandName	PrimaryCat
60820	22212	Moisturizing Shampoo	NaN	NaN	165	Regis Corporation	Regis Design Line	
60821	22212	Moisturizing Shampoo	NaN	NaN	165	Regis Corporation	Regis Design Line	
60822	22212	Moisturizing Shampoo	NaN	NaN	165	Regis Corporation	Regis Design Line	
60823	22212	Moisturizing Shampoo	NaN	NaN	165	Regis Corporation	Regis Design Line	
60824	22212	Moisturizing Shampoo	NaN	NaN	165	Regis Corporation	Regis Design Line	
60825	22212	Moisturizing Shampoo	NaN	NaN	165	Regis Corporation	Regis Design Line	
60826	22212	Moisturizing Shampoo	NaN	NaN	165	Regis Corporation	Regis Design Line	
60827	22212	Moisturizing Shampoo	NaN	NaN	165	Regis Corporation	Regis Design Line	

9 rows × 22 columns

Uh oh!

It turns out it is only one product, where each chemical is separately reported.

The following code is used to generate the bar chart showing the number of products per number of chemicals. In counting the number of products, different color, scent and/or flavor of the product are neglected (e.g. 'Professional Eyeshadow Base' can be beige or bright, but it is counted only once with the identification number 'CDPHId'=26).

```
In [10]: df_n0.loc[df['CDPHId']==26]
```

Out[10]:

	CDPHId	ProductName	CSFId	CSF	CompanyId	CompanyName	BrandName	PrimaryCateg
32	26	Professional Eyeshadow Base	337.0	Beige	27	CHANEL, INC	CHANEL	
33	26	Professional Eyeshadow Base	338.0	Bright	27	CHANEL, INC	CHANEL	

2 rows × 22 columns

In [11]:

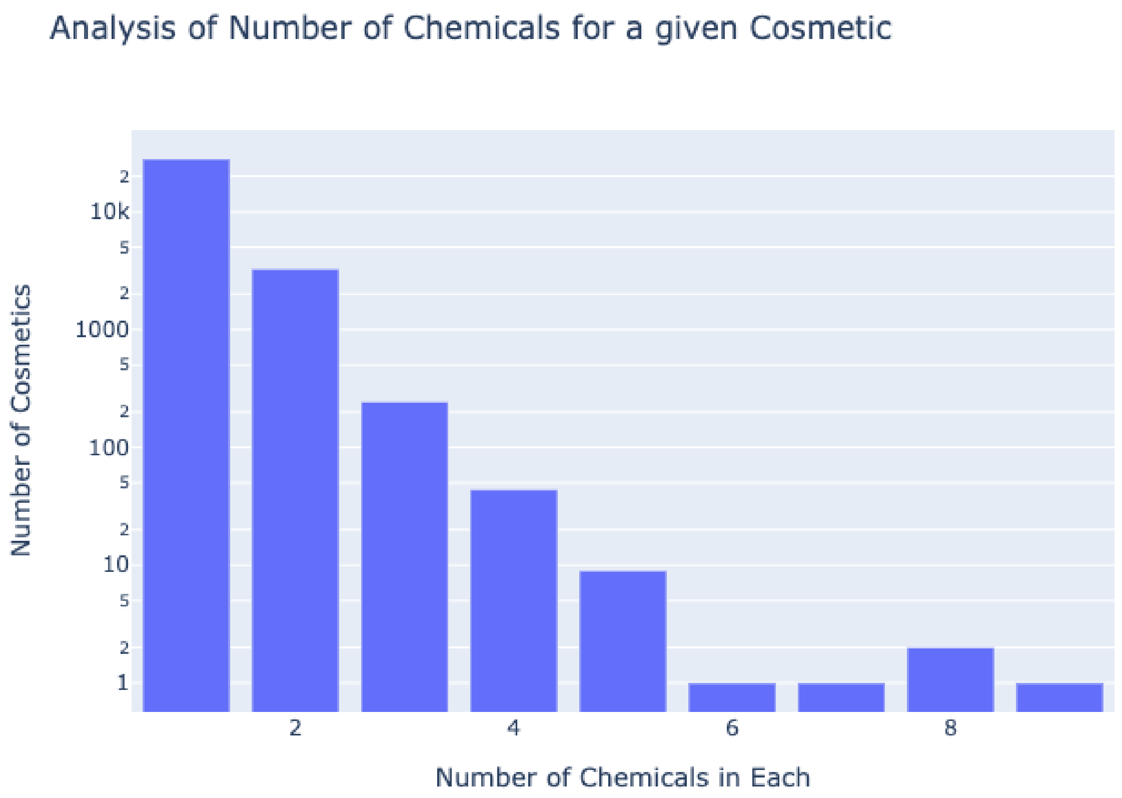
```
data = df_n0.groupby(['ChemicalCount']).nunique()['CDPHId']
```

We have grouped everything by unique CDPHId, so that we have no outlier values

In [12]:

```
fig = px.bar(data, x=data.index, y=data.values, log_y=True, labels={
    "y":"Number of Cosmetics",
    "ChemicalCount":"Number of Chemicals in Each"
}, title="Analysis of Number of Chemicals for a given Cosmetic")
img_bytes = fig.to_image(format="png")
Image(img_bytes)
```

Out[12]:



Baby Products

We are starting the next part of our analysis which deals with baby products

In [13]:

```
baby_prod = df_n0.loc[df_n0['PrimaryCategory']=='Baby Products']
baby_prod.head()
```

Out[13]:

	CDPHId	ProductName	CSFId	CSF	CompanyId	CompanyName	BrandNam
14178	3195	Baby Don't Cry Shampoo	22468.0	Fragrance/parfum	174	John Paul Mitchell Systems	John Pau Mitche System
19139	4654	Harmon Zinc Oxide Ointment 2oz	NaN	NaN	266	Harmon Stores Inc.	Harmo Face Value

	CDPHId	ProductName	CSFId	CSF	CompanyId	CompanyName	BrandNam
19140	4654	Harmon Zinc Oxide Ointment 2oz	NaN	NaN	266	Harmon Stores Inc.	Harmo Face Value
20078	5092	Balmex Multi-Purpose Healing Ointment	NaN	NaN	60	Chattem, Inc.	Balme
20083	5096	Balmex Prevention Baby Powder	NaN	NaN	60	Chattem, Inc.	Balme

5 rows × 22 columns

Commonality of Chemicals

Let us see which chemicals are present in these products

In [14]:

```
baby_prod_chem = baby_prod['ChemicalName'].value_counts()
print(baby_prod_chem)
```

```
Titanium dioxide
16
Cocamide DEA
5
Retinyl palmitate
3
Retinol/retinyl esters, when in daily dosages in excess of 10,000 IU, or 3,000 r
etinol equivalents.     3
Cocamide diethanolamine
2
Lead
1
Talc
1
Cadmium and cadmium compounds
1
Styrene
1
Acetaldehyde
1
Butylated hydroxyanisole
1
Formaldehyde (gas)
1
Name: ChemicalName, dtype: int64
```

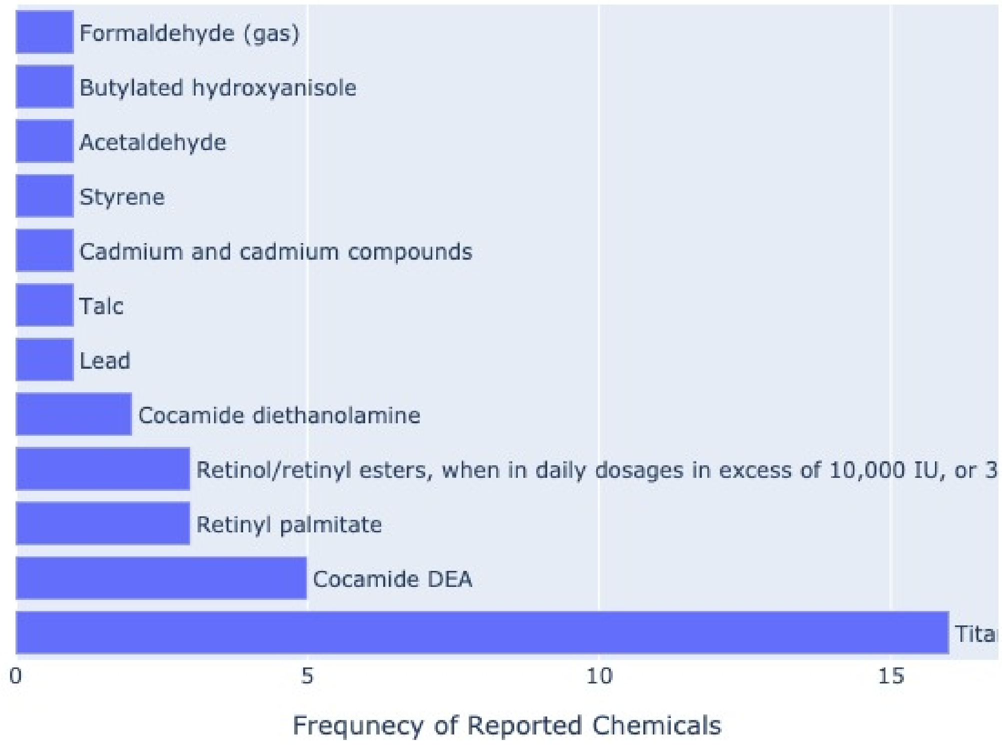
We see that Titanium Dioxide is present in a lot of baby cosmetic products

Fortunately, according to this%20is,are%20increasingly%20manufactured%20and%20used.) resource, we see that TiO_2 is inert and safe

In [15]:

```
fig = px.bar(y=baby_prod_chem.index, x=baby_prod_chem.values, text=baby_prod_che
    "x":"Frequnecy of Reported Chemicals"
})
fig.update_traces(texttemplate='', textposition='outside')
fig.update_layout(uniformtext_minsize=8, uniformtext_mode='hide')
fig.update_yaxes(visible=False)
img_bytes = fig.to_image(format="jpeg")
Image(img_bytes)
```

Out[15]:



Conclusion

In essence, most chemicals that are found in Baby Chemical Products are safe. However, concerned parents and other stakeholders can consult this data to be sure and be wary of the various chemicals present in their children's cosmetic products