



# Đồ án thực hành cuối kỳ



Môn: Lập trình cho Khoa học dữ liệu - Nhóm 15



# Nội dung



Giới thiệu thành viên



Sơ lược về dataset



Quá trình thực hiện

- Thu thập dữ liệu
- Khám phá dữ liệu
- Đặt câu hỏi và trả lời



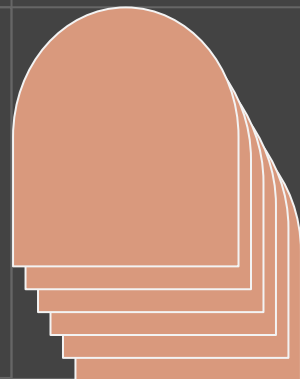
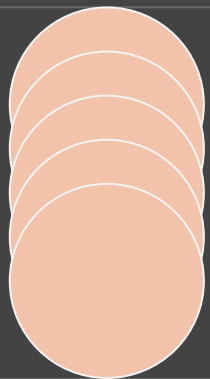
Đánh giá công việc

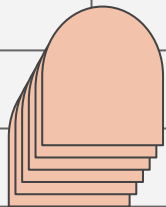
- Khó khăn gặp phải
- Kinh nghiệm rút ra



01

# Giới thiệu thành viên





Đinh Thị Hoàng Linh - 20120130

Phạm Trần Gia Phú - 20120348

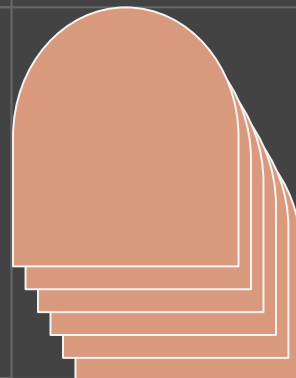
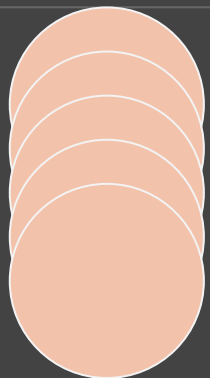
Nguyễn Anh Tuấn - 20120395

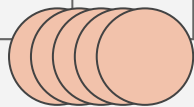
Phạm Khánh Hoàng Việt - 20120626



# 02

## Sơ lược về dataset



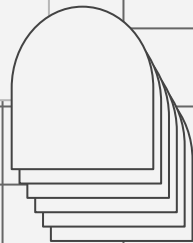


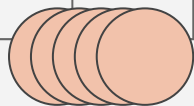
# Sơ lược về dataset



## Chemicals in cosmetics

- Mỹ phẩm là một trong những sản phẩm được sử dụng nhiều nhất hiện nay.
- Vậy nguyên liệu và thành phần của nó bao gồm những gì?
- Dataset này chứa thông tin bao gồm tên những chất hóa học được sử dụng để tạo ra mỹ phẩm và một vài các thông số của nó.





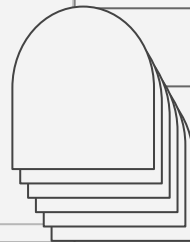
# Sơ lược về dataset



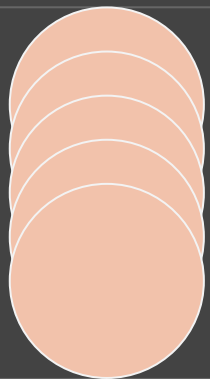
## Research Ideas

Dataset này có thể dùng để:

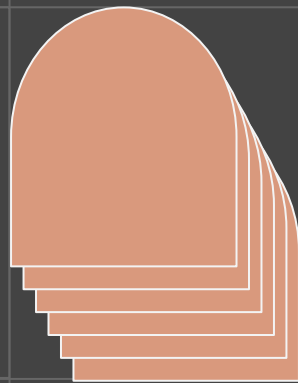
- Tìm ra các chất hóa học được sử dụng trong công nghiệp mỹ phẩm
- Các công ty sử dụng các chất nào cho các sản phẩm của họ
- Các yếu tố để lựa chọn sử dụng mỹ phẩm một cách an toàn.
- ..V..V..



03 ✨

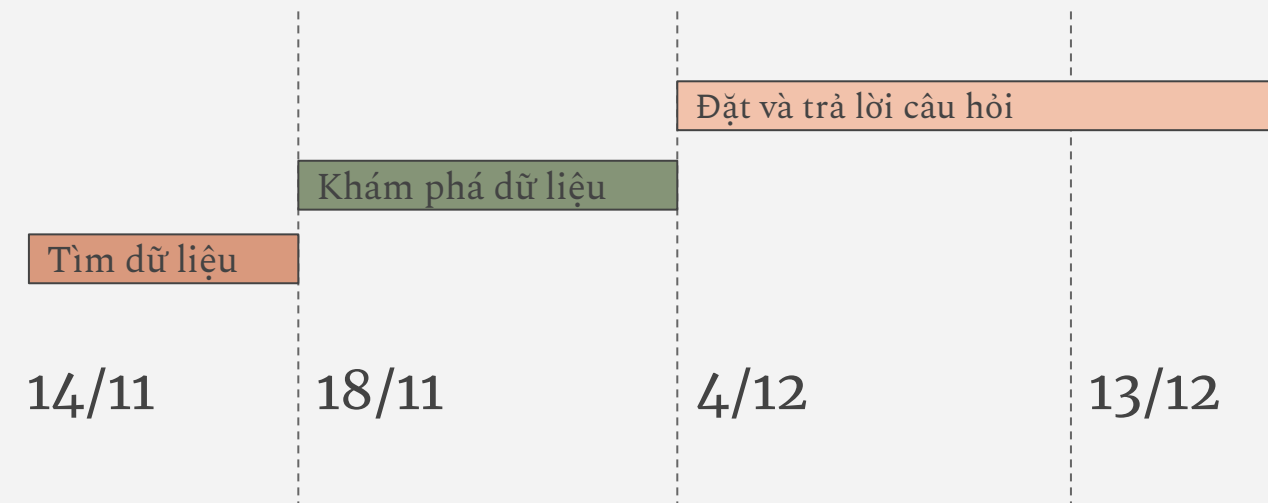
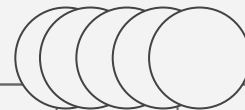


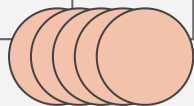
# Các bước thực hiện





# Quá trình thực hiện





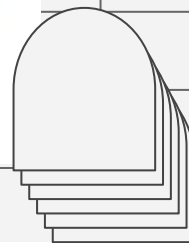
# Thu thập dữ liệu

Dữ liệu sử dụng có chủ đề về  
hóa chất trong mỹ phẩm

Dữ liệu được lấy từ Kaggle

DOWNLOADS

579



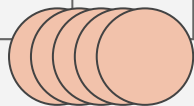
Compatibility · 100%

- ✓ License
- ✓ File Format
- ✓ File Description
- ✓ Column Description

# Khám phá dữ liệu

Ở bước này ta sẽ tập trung trả lời các câu hỏi để khám phá xem dataset có gì đặc biệt không như:

- Dữ liệu có bao nhiêu dòng và cột ?
- Mỗi dòng có ý nghĩa gì ?
- Dữ liệu có chứa dòng bị lặp không ?
- Mỗi cột có ý nghĩa gì ?
- Mỗi cột có kiểu dữ liệu gì ?
- Phân bố của các loại dữ liệu (numeric và non-numeric).



# Khám phá dữ liệu

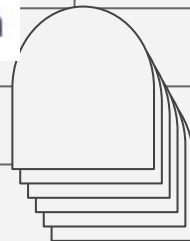


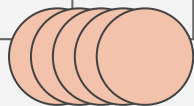
Kiểm tra kích thước dữ liệu

Dữ liệu có 114298 dòng và 23 cột

Mỗi dòng có ý nghĩa gì?

Mỗi dòng đại diện cho 1 thông tin của một sản phẩm

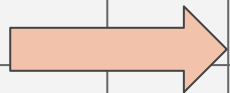




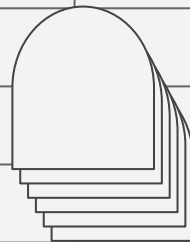
# Khám phá dữ liệu

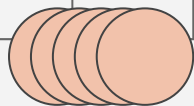


Kiểm tra dữ liệu có dòng bị lặp không?



Dữ liệu không có chứa dòng bị lặp !



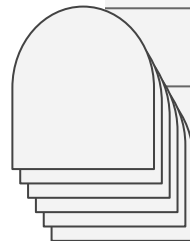


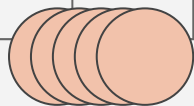
# Khám phá dữ liệu



Mỗi cột có ý nghĩa gì?

- **CDPHId**: Số nhận dạng nội bộ của Bộ Y tế Công cộng California (CDPH) cho sản phẩm. Lưu ý rằng CDPHId có thể xuất hiện nhiều lần nếu một sản phẩm có nhiều Màu sắc/Mùi hương/Hương vị, nhiều Danh mục hoặc nhiều Tên hóa chất/số CAS được báo cáo.
- **ProductName**: Tên của sản phẩm mà hóa chất được sử dụng.
- **CSFId**: Số nhận dạng nội bộ CDPH cho màu sắc/mùi hương/hương vị.
- **CSF**: Hệ số an toàn mỹ phẩm, thước đo mức độ an toàn của hóa chất được sử dụng trong mỹ phẩm.
- **CompanyId** : Mã số nhận dạng nội bộ CDPH cho công ty.
- **CompanyName**: Tên của công ty sản xuất sản phẩm mà hóa chất được sử dụng.
- **BrandName**: Tên thương hiệu của sản phẩm mà hóa chất được sử dụng.
- **PrimaryCategoryId**: Số nhận dạng nội bộ CDPH cho danh mục.
- **PrimaryCategory**: Loại mỹ phẩm chính mà hóa chất được sử dụng.
- **SubCategoryId**: Số nhận dạng nội bộ CDPH cho tiểu thể loại.
- **SubCategory**: Danh mục phụ của mỹ phẩm trong đó hóa chất được sử dụng.
- **CASId**: Số nhận dạng nội bộ CDPH cho hóa chất.



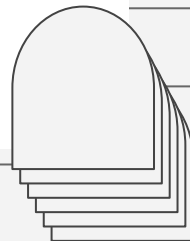


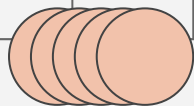
# Khám phá dữ liệu



Mỗi cột có ý nghĩa gì?

- **CasNumber:** Số CAS của hóa chất.
- **ChemicalId:** Số nhận dạng nội bộ CDPH cho hồ sơ hóa chất cụ thể cho sản phẩm này.
- **ChemicalName:** Tên hóa chất.
- **InitialDateReported:** Ngày hóa chất được báo cáo lần đầu tiên.
- **MostRecentDateReported:** Ngày hóa chất được báo cáo gần đây nhất.
- **DiscontinuedDate:** Ngày ngừng sản xuất sản phẩm sử dụng hóa chất.
- **ChemicalCreatedAt:** Ngày hóa chất được tạo ra.
- **ChemicalUpdatedAt:** Ngày hóa chất được cập nhật lần cuối.
- **ChemicalDateRemoved:** Ngày loại bỏ hóa chất khỏi sản phẩm.
- **ChemicalCount:** Số lượng hóa chất trong sản phẩm.





# Khám phá dữ liệu

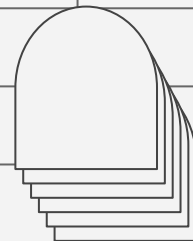


Mỗi cột hiện đang có kiểu dữ liệu gì?

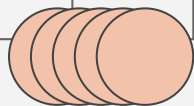
index	int64
CDPHId	int64
ProductName	object
CSFId	float64
CSF	object
CompanyId	int64
CompanyName	object
BrandName	object
PrimaryCategoryId	int64
PrimaryCategory	object
SubCategoryId	int64
SubCategory	object
CasId	int64



CasNumber	object
ChemicalId	int64
ChemicalName	object
InitialDateReported	object
MostRecentDateReported	object
DiscontinuedDate	object
ChemicalCreatedAt	object
ChemicalUpdatedAt	object
ChemicalDateRemoved	object
ChemicalCount	int64







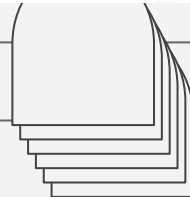
# Khám phá dữ liệu

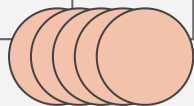


Cột nào đang có kiểu dữ liệu chưa phù hợp?

Các cột mang thông tin thời gian ngày, tháng, năm nhưng có kiểu object chưa phù hợp

InitialDateReported	object
MostRecentDateReported	object
DiscontinuedDate	object
ChemicalCreatedAt	object
ChemicalUpdatedAt	object
ChemicalDateRemoved	object





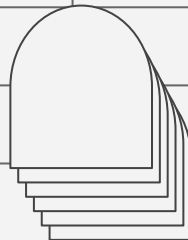
# Khám phá dữ liệu

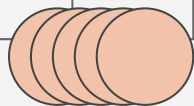


Chuyển kiểu dữ liệu các cột 'Date' sang datetime

Mục đích là để các cột này có thể xử lý như kiểu datetime chứ không phải kiểu str, tiện lợi cho quá trình trả lời câu hỏi ở phần sau

Phương pháp là sử dụng hàm `to_datetime()` kết hợp với `stack()` và `unstack()` để chuyển kiểu dữ liệu của nhiều cột cùng lúc.





# Khám phá dữ liệu



Kiểm tra phân bố dữ liệu của các cột

Kiểu dữ liệu số

Số giá trị thiếu

Tỉ lệ giá trị thiếu

Giá trị nhỏ nhất

Giá trị lớn nhất

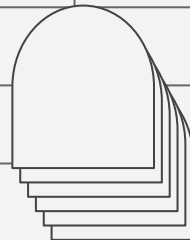
Kiểu dữ liệu phân loại

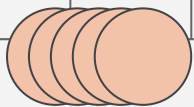
Số giá trị thiếu

Tỉ lệ giá trị thiếu

Số giá trị khác nhau

Các giá trị khác nhau



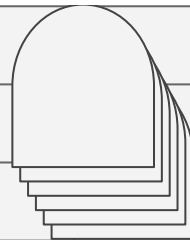


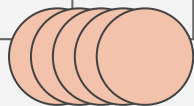
# Khám phá dữ liệu



Kiểm tra phân bố dữ liệu của các cột có kiểu số

	index	CDPHId	CSFId	CompanyId	PrimaryCategoryId	SubCategoryId	CasId	ChemicalId	ChemicalCount
<b>missing_values</b>	0.0	0.0	33916.000	0.0	0.0	0.0	0.0	0.0	0.0
<b>missing_ratio</b>	0.0	0.0	29.673	0.0	0.0	0.0	0.0	0.0	0.0
<b>min</b>	0.0	2.0	1.000	4.0	1.0	3.0	2.0	0.0	0.0
<b>max</b>	114297.0	41451.0	64883.000	1391.0	111.0	172.0	1242.0	67907.0	9.0





# Khám phá dữ liệu



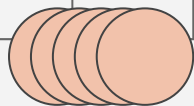
Kiểm tra phân bố dữ liệu của các cột có kiểu phân loại

	missing_num	missing_ratio	n_diff_vals	diff_vals
<b>ProductName</b>	0	0.0	33638	[ULTRA COLOR RICH EXTRA PLUMP LIPSTICK-ALL SHA...
<b>CSF</b>	34340	30.04427	34258	[5858-81-1, D&C RED 7 CALCIUM LAKE, D&C RED 28...
<b>CompanyName</b>	0	0.0	606	[New Avon LLC, J. Strickland & Co., OPI PRODUC...
<b>BrandName</b>	216	0.18898	2711	[AVON, Glover's, OPI, ABSOLUTE, ABSOLUTE FX, G...
<b>PrimaryCategory</b>	0	0.0	13	[Makeup Products (non-permanent), Hair Care Pr...
<b>SubCategory</b>	0	0.0	89	[Lip Color - Lipsticks, Liners, and Pencils, H...
<b>CasNumber</b>	6396	5.595898	125	[13463-67-7, 65996-92-1, 140-67-0, 68603-42-9,...
<b>ChemicalName</b>	0	0.0	123	[Titanium dioxide, Distillates (coal tar), Est...



v



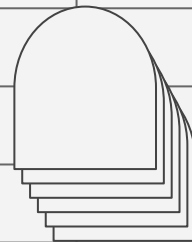


# Tiền xử lý



## Tác vụ 1

Đổi tên giá trị 'Cocamide diethanolamine' thành 'Cocamide DEA' tại cột 'ChemicalName' vì hai tên gọi đều của cùng một chất.




# Đặt câu hỏi và trả lời

Ở phần này nhóm mình sẽ dựa vào dataset đã khám phá phía trên để đưa ra những câu hỏi và trả lời, mục đích là để rút ra những insight từ dataset.





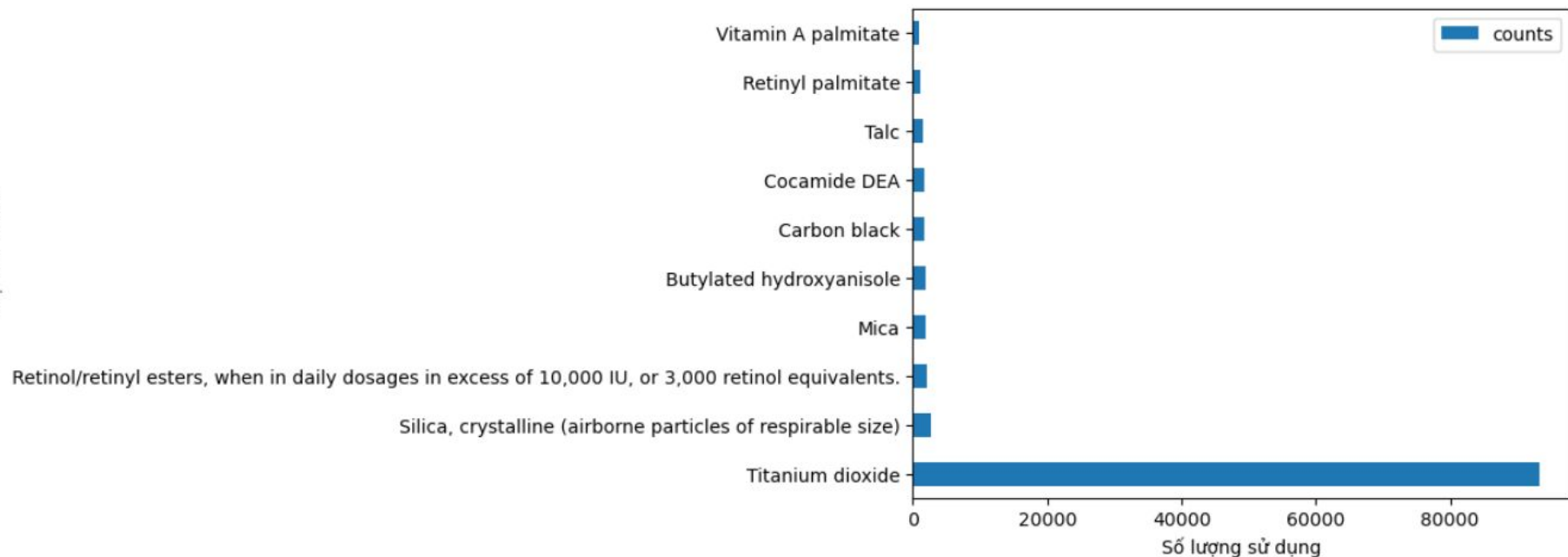
## Câu hỏi 1: Top 10 các loại hoá chất được sử dụng nhiều nhất



Ý nghĩa: Hẳn là ai cũng muốn biết trong sản phẩm mình sử dụng chứa thành phần chính là gì đúng không? Hãy điểm qua top 10 loại hoá chất được sử dụng nhiều nhất trong mỹ phẩm để xem thử loại hoá chất nào được các nhãn hàng ưa chuộng nhất nhé.








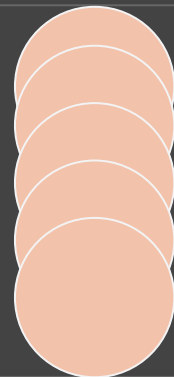
- Có thể thấy Titanium dioxide được sử dụng rất nhiều, có thể là đa phần trong các loại mỹ phẩm hiện nay.
- Titanium dioxide hay còn gọi là Titania, là một hợp chất tự nhiên. Khi sử dụng trong các sản phẩm mỹ phẩm tại Liên minh Châu u (EU), Titanium dioxide thô sau khi khai thác sẽ được các chuyên gia xử lý và tinh chế cho phù hợp với mục đích sản xuất.
- Titanium dioxide có dạng bột mịn màu trắng, có độ tương phản cao và có khả năng tạo ra sắc tố trắng sáng nên thường dùng làm nền cho các loại mỹ phẩm.

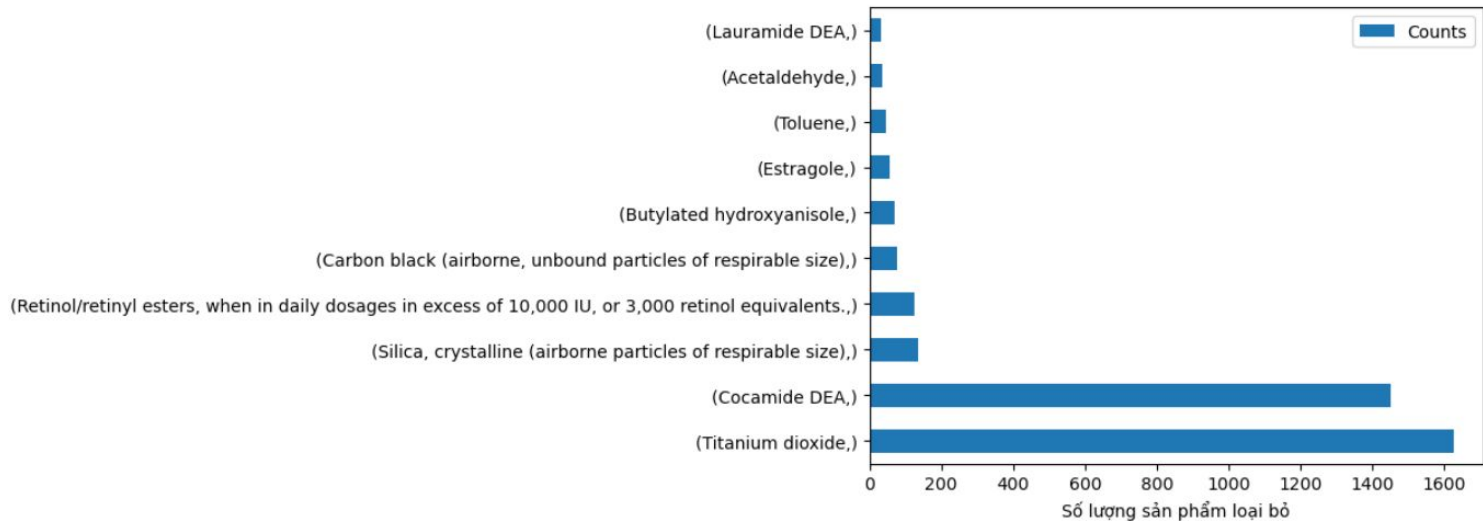


## Câu hỏi 2: Top 10 những hóa chất đã được loại bỏ nhiều nhất



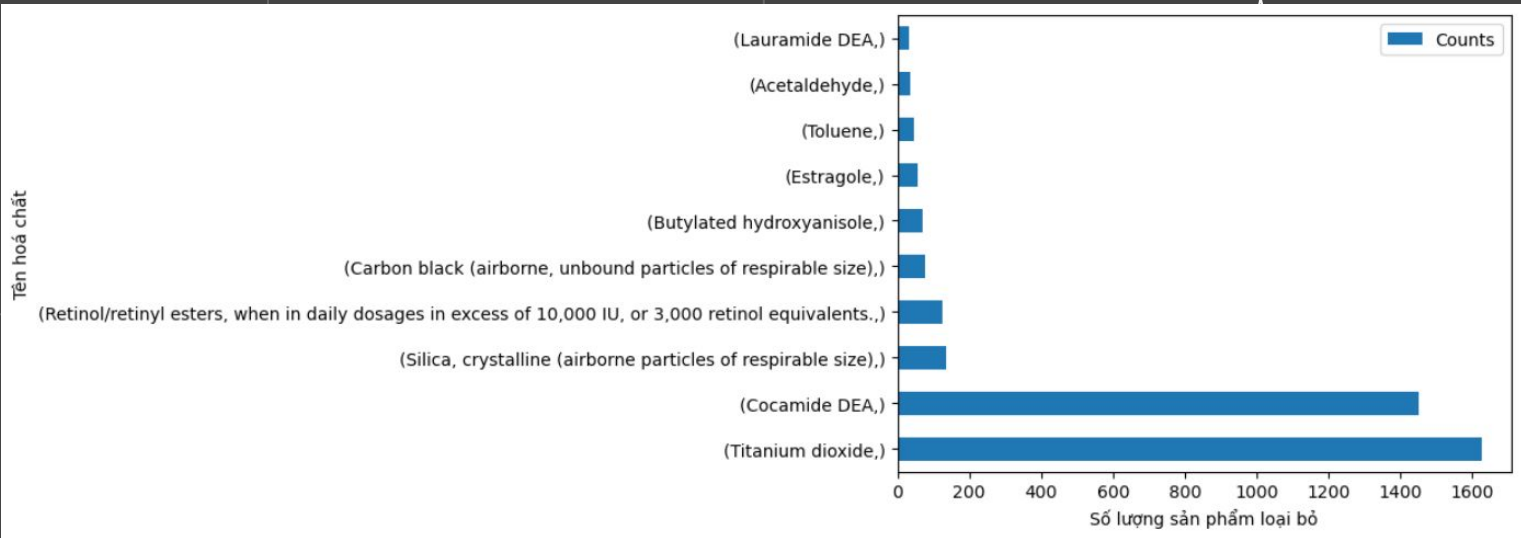
Ý nghĩa: Những hóa chất bị loại bỏ có thể do một vài nguyên nhân như có nhiều báo cáo không tốt cho sức khỏe con người hoặc các thành phần đó không tốt cho môi trường, vv.. Những hóa chất như vậy cũng đáng để ta quan tâm.





## Nhật xét:


- **Titanium dioxit** là hóa chất có nhiều sản phẩm loại bỏ nhất.
- Tuy **Cocamide DEA** có số loại bỏ đứng thứ hai nhưng lại có số lượng sản phẩm sử dụng nhỏ hơn rất nhiều so với Titanium dioxit (biểu đồ câu hỏi số 1). Điều đó dẫn đến câu hỏi tại sao Cocamide DEA lại bị loại bỏ nhiều như vậy?




Cocamide diethanolamine hay Cocamide DEA được sử dụng trong các sản phẩm sữa rửa mặt, dầu gội đầu hay nước rửa tay như một chất tạo bọt. Cocamide DEA được tuyên bố là an toàn ở nồng độ dưới 10%. Tuy nhiên, việc sử dụng thành phần này đã giảm trong những năm qua. Điều này do việc sử dụng nhiều và kéo dài Cocamide DEA có liên quan đến ung thư. Ngay ở số lượng nhỏ, nó có thể gây một số tác dụng phụ như ngứa. Hơn nữa Cocamide DEA có thể tác dụng với các chất khác tạo ra các chất có hại được gọi là nitrosamine. Cocamide DEA được xếp ở mức 7 trên thang 10 của EWG (trong đó 1 là thấp nhất, 10 là cao nhất về mức độ nguy hiểm).

(EWG - Environmental Working Group, là một tổ chức phi lợi nhuận của Mỹ chuyên nghiên cứu và vận động chính sách trong các lĩnh vực trợ cấp nông nghiệp, hóa chất độc hại, chất gây ô nhiễm nước uống và trách nhiệm giải trình của công ty.)

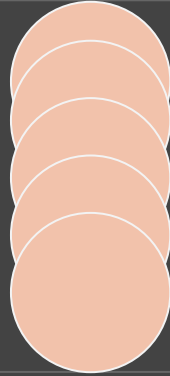
Vì vậy, mỗi lần mua sản phẩm có nhiều bọt như dầu gội đầu, nước rửa mặt, nước rửa, ... ta nên tránh những sản phẩm có thành phần Cocamide DEA.



Câu hỏi 3: Các hãng sản xuất mỹ phẩm có sản phẩm chứa nhiều hóa chất?



Ý nghĩa: Khám phá các hãng mỹ phẩm sử dụng nhiều loại hoá chất cho sản phẩm của mình nhé :v Để biết xem hãng nào có nhiều hoá chất và đưa ra lựa chọn phù hợp





Regis Corporation	96
Palladio Beauty Group	64
Bliss World LLC	24
Puritan's Pride	10
Vitamin World, Inc.	10
Cosmopharm Ltd.	9
MILANI COSMETICS	6
Good 'N Natural	5



Một số điều rút ra từ việc xem xét thông tin các sản phẩm chứa nhiều chất hóa học trong thành phần:


- Hãng 'Regis Corporation' có vẻ rất thích sử dụng hóa chất, các sản phẩm chứa nhiều hóa chất nhất trong dataset này đều do hãng này sản xuất...

- Kể đến chúng ta có hãng 'Palladio Beauty Group', nếu như bảng vàng các sản phẩm chứa từ 7 chất hóa học trở lên bị Regis Corporation độc chiếm thì Palladio Beauty Group cũng chiếm đa số trong các sản phẩm chứa từ 5 tới 6 chất hóa học.

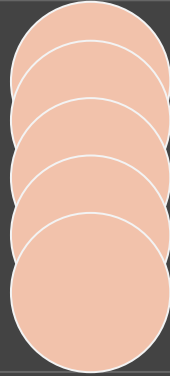
Nói chung là nếu thích sử dụng các loại mỹ phẩm thiên nhiên, thân thiện với môi trường... thì nên né 2 hãng này ra.

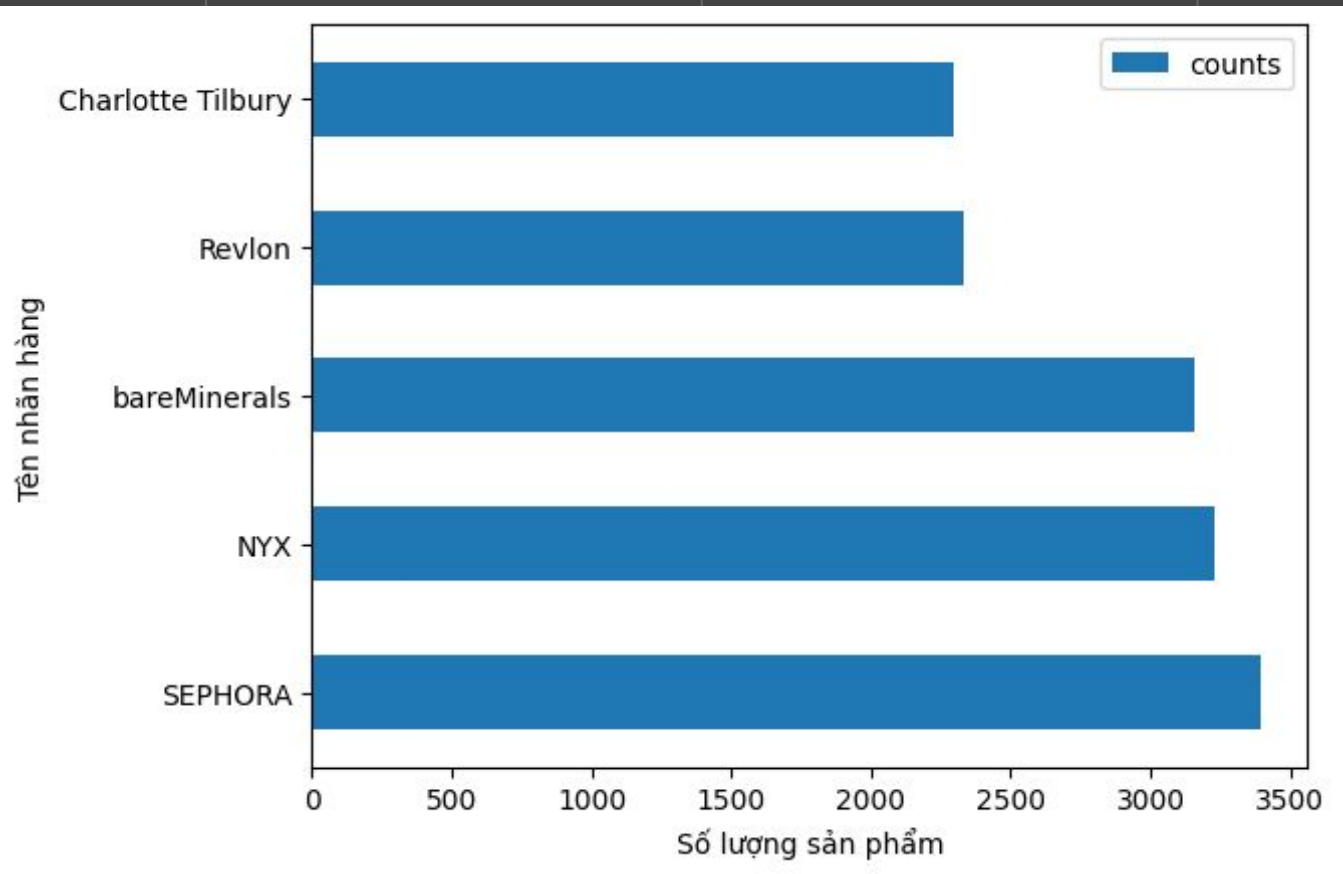


## Câu hỏi 4: Top 5 Các brand có nhiều sản phẩm nhất



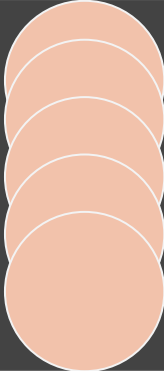


Ý nghĩa: Tìm hiểu top 5 hãng mỹ phẩm có nhiều sản phẩm trên thị trường nhất. Có thể ta sẽ dựa vào đó để tìm kiếm sản phẩm phù hợp



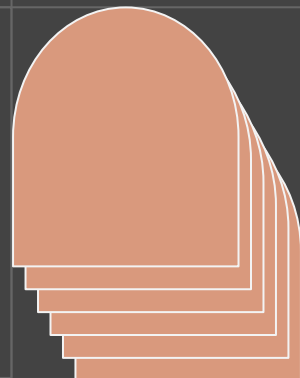
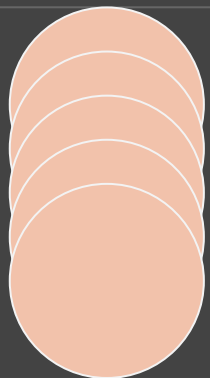




- 
- 
- Đây là 5 nhãn hàng mỹ phẩm có nhiều sản phẩm nhất
  - Đứng đầu là **SEPHORA** và thứ 5 là **Charlotte Tilbury**
  - Có thể thấy **SEPHORA**, **NYX**, **bareMinerals** là các nhãn hàng có nhiều sản phẩm nhất và chênh lệch số lượng sản phẩm giữa 3 nhãn hàng này là không nhiều.
  - Sau top 3 thì có sự chênh lệch khá đáng kể khoảng 800-1000 sản phẩm ở 2 vị trí tiếp theo là **Revlon** và **Charlotte Tilbury**
- 

04

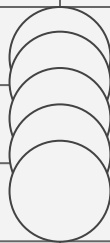
# Đánh giá công việc



# Đánh giá công việc

## Linh

- Khó khăn gặp phải: Tìm ra câu hỏi có giá trị ứng dụng cao.
- Bài học rút ra: Cách phân tích dữ liệu và sử dụng các thư viện của python để trả lời câu hỏi



## Tuấn

- Khó khăn gặp phải là chưa có nhiều kinh nghiệm sử dụng Notebook, Python.
- Điều học được: Sử dụng Jupyter, Python; Biết thêm một số hóa chất phổ biến.



# Đánh giá công việc

## Phú

- Khó khăn gặp phải: Gặp khó khăn trong vấn đề suy nghĩ đặt câu hỏi.
- Bài học rút ra: Được làm việc với 1 quy trình khoa học dữ liệu hoàn chỉnh

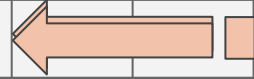


## Việt

- Khó khăn gặp phải: Chưa có nhiều kinh nghiệm với python
- Bài học rút ra: Nắm được cách sử dụng các thư viện của python vào xử lý dữ liệu



# Các nguồn tham khảo



- Tài liệu thư viện Numpy: [NumPy Documentation](#)
- Tài liệu thư viện Pandas: [pandas documentation — pandas 1.5.2 documentation\(pydata.org\)](#)
- Tài liệu thư viện Matplotlib: [Matplotlib 3.6.2 documentation\(matplotlib.org\)](#)



# Thanks for listening!

