

**Phân tích dữ liệu**

**sức khỏe làm đẹp và doanh số bán hàng**

**MÔN HỌC: KỸ THUẬT XỬ LÝ DỮ LIỆU LỚN**

**GIÁO VIÊN HƯỚNG DẪN:** PGS.TS. Nguyễn Thanh Bình

**THÀNH VIÊN NHÓM:** 22C01016 - Tống Lương Hương Quỳnh

22C01026 - Nguyễn Ngọc Thảo Uyên

22C01034 - Đào Thị Minh Lý



**MỤC LỤC**



[**1. Tổng Quan Về Dữ Liệu 3**](#_9neim29297n1)

[1.1. Nguồn Dữ Liệu 3](#_f7r88mpw52bv)

[1.2. Mục Đích Sử Dụng 3](#_arqbqws9duln)

[1.3. Thông tin bộ dữ liệu Hasaki 3](#_76oytu82cdp8)

[**2. Pipeline 4**](#_2w0nrslev7ho)

[2.1. Crawl dữ liệu từ trang web Hasaki bằng Python: 4](#_dt0x84xcw775)

[2.2. Data Preprocessing 5](#_5xxqgdyhwrnp)

[2.3. Kết nối và lưu trữ dữ liệu bằng SQL Server 6](#_lijga0hwu2cr)

[2.4. Trực quan hóa dữ liệu bằng Power BI 6](#_9zz814ouruc6)

[**3. Dashboard 7**](#_y9kys125ruhc)

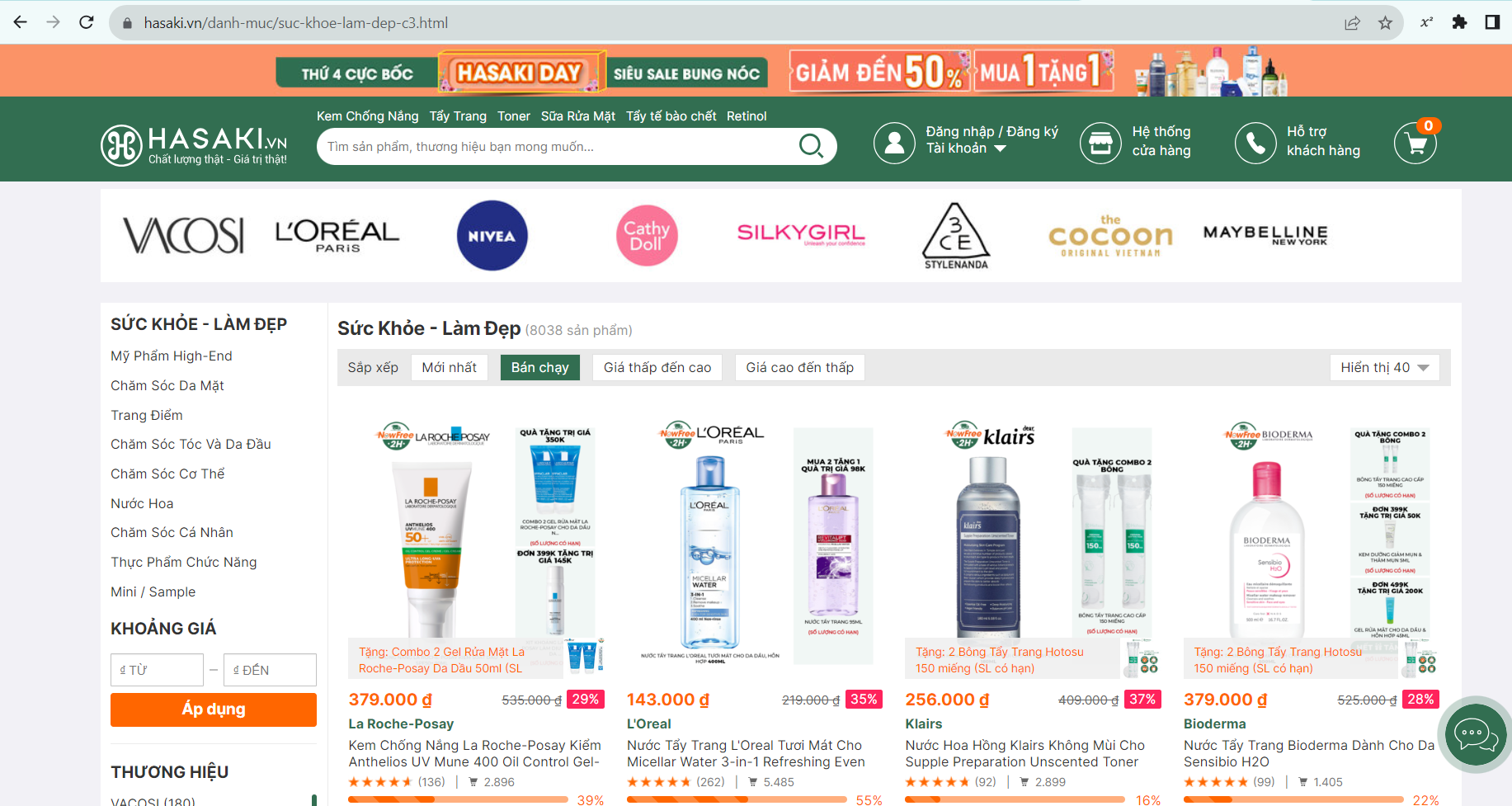
# 

# 

# **1. Tổng Quan Về Dữ Liệu**

## **1.1. Nguồn Dữ Liệu**

Trong project này, nhóm đã crawl bộ dữ liệu từ trang web chính thức của Hasaki, một thương hiệu hàng đầu trong lĩnh vực mỹ phẩm và chăm sóc sức khỏe tại Việt Nam. Trang web Hasaki.vn cung cấp một loạt sản phẩm chất lượng và được tin dùng bởi người tiêu dùng.



## **1.2. Mục Đích Sử Dụng**

Bộ dữ liệu có tiềm năng sử dụng rộng rãi trong việc cho mục đích phân tích xu hướng mua sắm, đánh giá hiệu suất của các sản phẩm sức khỏe và làm đẹp, và phân tích doanh số bán hàng của Hasaki.vn. Các chuyên gia và nhà nghiên cứu có thể sử dụng bộ dữ liệu này để hiểu rõ hơn về thị trường mỹ phẩm và sức khỏe tại Việt Nam và tạo ra các chiến lược kinh doanh hiệu quả.

## **1.3. Thông tin bộ dữ liệu Hasaki**

Bộ dữ liệu này cung cấp một loạt thông tin chi tiết về sản phẩm và người mua, bao gồm:

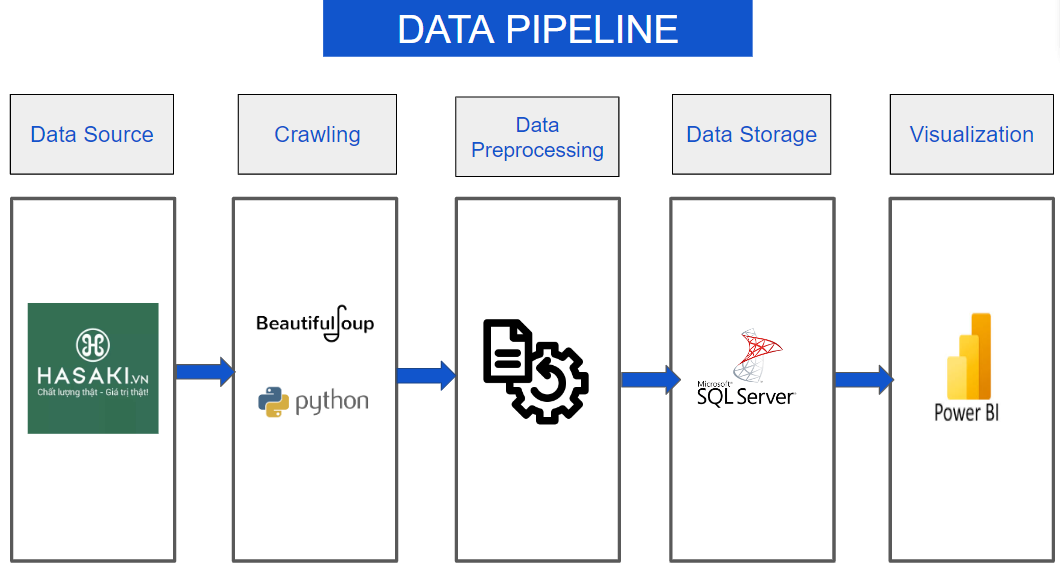
* ‘item\_name': Tên của sản phẩm.
* 'item\_brand': Thương hiệu của sản phẩm.
* 'item\_id': Mã số sản phẩm.
* 'item\_product': Loại sản phẩm.
* 'item\_price': Giá sản phẩm.
* 'item\_bought': Số lượng sản phẩm đã mua, được tính từ khi sản phẩm được đăng bán trên web đến thời điểm crawl dữ liệu.
* 'item\_discount': Giảm giá của sản phẩm.
* 'item\_price\_previous': Giá gốc của sản phẩm trước khi giảm giá.
* 'item\_star': Điểm đánh giá của sản phẩm.
* 'item\_comment\_number': Số lượng bình luận về sản phẩm.
* ‘xuat\_xu\_thuong\_hieu’: Thông tin về nơi sản xuất
* ‘gioi\_tinh’: Giới tính người sử dụng.

Bộ dữ liệu cho phép các nhà phân tích thực hiện các phân tích sâu về hành vi mua sắm và sự phân phối sản phẩm trong thị trường mỹ phẩm và chăm sóc sức khỏe tại Việt Nam.

Bộ dữ liệu được crawl trong 30 ngày (Từ ngày *30/08/2023* đến ngày *28/09/2023*), từ đó nhóm tiến hành xử lý dữ liệu để phân tích tình hình bán hàng trong 1 tháng dưới hình thức bán hàng trên trang Web chính thức của Hasaki.

Bộ dữ liệu ban đầu được chia thành 8 bảng theo 8 ngành hàng của Hasaki và được crawl về từ trang Web theo từng ngày, bao gồm: *ngành hàng* ***cá nhân****, ngành hàng* ***cơ thể****, ngành hàng* ***da mặt****, ngành hàng* ***high-end****, ngành hàng* ***nước hoa****, ngành hàng* ***thực phẩm chức năng (tpcn)****, ngành hàng* ***tóc****, ngành hàng* ***trang điểm****.*

# **2. Pipeline**



## **2.1. Crawl dữ liệu từ trang web Hasaki bằng Python:**

Bộ dữ liệu được crawl từ trang web Hasaki như sau:

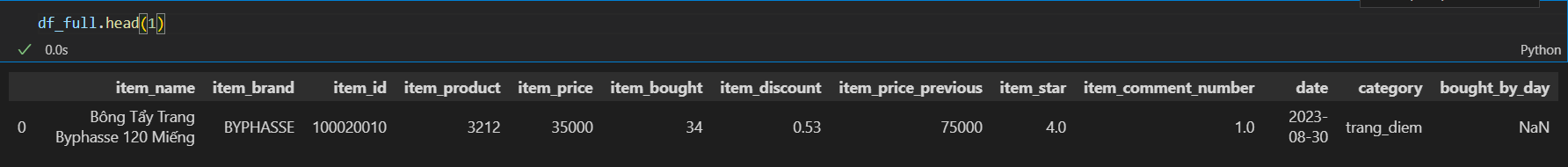
* Sử dụng Python và thư viện BeautifulSoup để crawl dữ liệu từ trang web Hasaki.
* Xác định các trang web con hoặc API endpoint cần crawl để lấy thông tin sản phẩm, giá cả, mô tả, và chi tiết khác.
* Sử dụng module Requests để thực hiện các HTTP requests đến các URL cần crawl.
* Lấy dữ liệu từ trang web bằng cách sử dụng các tag HTML. Sau khi lấy được dữ liệu từ trang web, tiến hành xử lý dữ liệu bằng Python để làm sạch và chuẩn hóa thông tin.

## **2.2. Data Preprocessing**

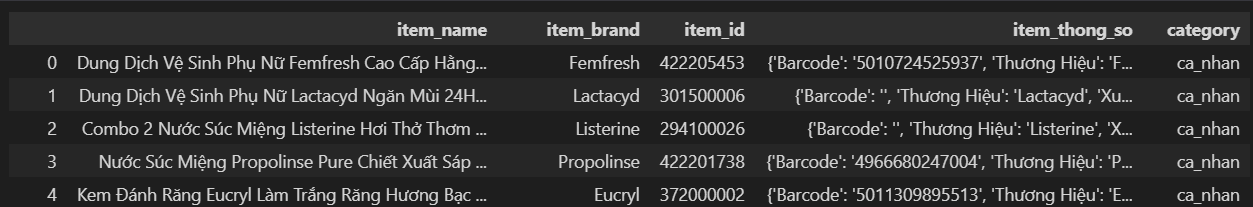
Sau khi đã thu thập đủ dữ liệu trong 30 ngày, nhóm sử dụng **Python Pandas** để làm sạch và chuẩn hóa thông tin của dữ liệu.

Đầu tiên tiến hành thêm cột **“category”** để phân biệt mỗi ngành hàng và gộp 8 bảng lại với nhau đồng thời crawl thêm dữ liệu thông số về xuất xứ của sản phẩm và giới tính của người có thể sử dụng sản phẩm. Từ đó có được 2 bảng dữ liệu là bảng ***Category*** và bảng ***Information***.

Bảng ***Category*** đã xử lý các giá trị NULL và thêm cột ***bought\_by\_day*** (lượng mua trong ngày). Bên cạnh đó, nhóm có chuyển đổi vài cột dữ liệu về đúng format để có thể xử lý dữ liệu. Cụ thể, cột ***item\_discount*** và ***item\_star*** vừa chứa thông tin *%* (ví dụ 40%), vừa chứa số thập phân (ví dụ 0.4). Hoặc cột ***item\_price\_previous*** ban đầu là chuỗi chứa cả *.* và *đ* (ví dụ 100.000đ). Ngoài ra, nhóm còn loại bỏ duplicate các dòng dữ liệu, mỗi ngày (***date***) chỉ giữ lại thông tin bán hàng của một sản phẩm.



Bảng ***Information*** trước khi xử lý dữ liệu:



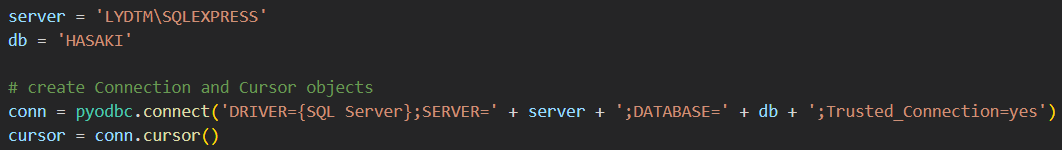
Nhóm tiến hành xử lý các giá trị NULL và tách cột ***item\_thong\_so*** thành các cột là ***xuat\_xu\_thuong\_hieu*** và ***gioi\_tinh****.* Làm rõ hơn, cột ***item\_thong\_so*** có nhiều thông tin như *barcode, thương hiệu, xuất xứ thương hiệu, đặc tính, giới tính*, … Nhưng có vài thông tin như *barcode, đặc tính* thì không phải sản phẩm nào cửa hàng cũng cập nhật lên web, nên nhóm nhận thấy rằng không thể sử dụng được các cột thông tin bị thiếu dữ liệu quá nhiều, do đó nhóm đã xóa các cột này đi, chỉ giữ lại những cột chứa các thông tin giá trị là ***xuat\_xu\_thuong\_hieu*** và ***gioi\_tinh***.

Bảng ***Information*** sau khi xử lý dữ liệu:

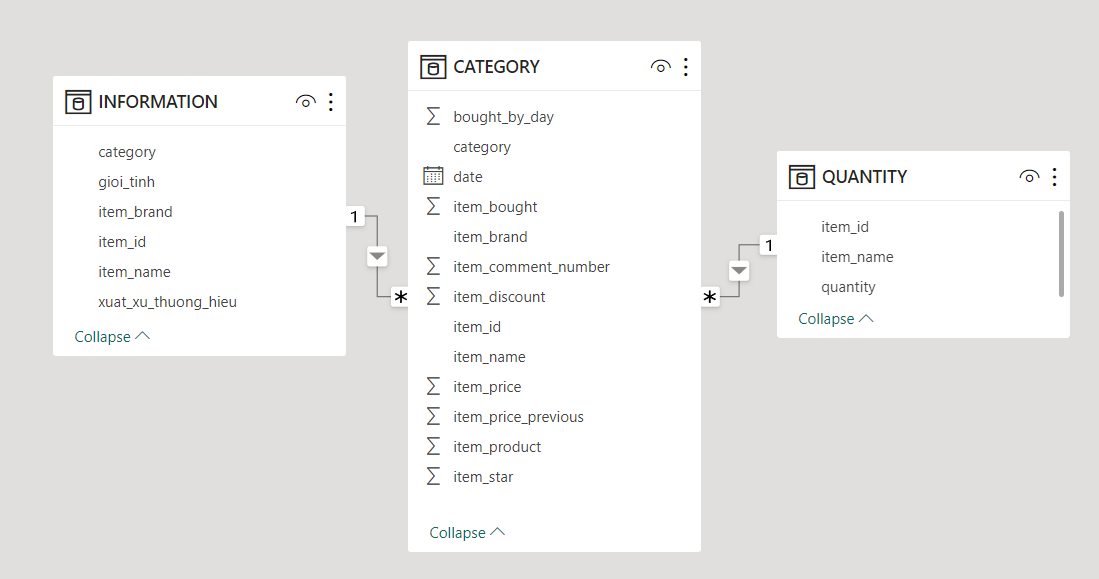


## **2.3. Kết nối và lưu trữ dữ liệu bằng SQL Server**

Sử dụng thư viện pyodbc để kết nối Python với cơ sở dữ liệu SQL Server. Cung cấp thông tin về máy chủ, cơ sở dữ liệu, tên người dùng, và mật khẩu để thiết lập kết nối.



Sau khi đã kết nối được với cơ sở dữ liệu SQL Server, nhóm sử dụng Python tiến hành đưa dữ liệu thu thập từ trang web Hasaki vào cơ sở dữ liệu SQL Server. *(Lưu ý xem lại cấu trúc bảng với các trường tương ứng với dữ liệu cần lưu trữ.)* Với dữ liệu Tiếng Việt, để lưu trữ và hiển thị đúng định dạng thì cần sử dụng kiểu dữ liệu là nvarchar, tức là tạo bảng dữ liệu với kiểu dữ liệu cho chính xác, sau đó mới thêm dữ liệu vào SQL Server.



Sau khi có kết nối và bảng cơ sở dữ liệu, sử dụng SQL để thực hiện các truy vấn dữ liệu.

## **2.4. Trực quan hóa dữ liệu bằng Power BI**

Sau khi đã lưu trữ dữ liệu ở SQL Server, nhóm thực hiện trực quan hóa dữ liệu bằng công cụ Power BI sau khi đã lấy dữ liệu từ SQL Server thông qua kết nối trực tiếp. Quy trình này giúp biến đổi dữ liệu từ cơ sở dữ liệu SQL Server thành các biểu đồ và báo cáo trực quan, hỗ trợ quyết định và phân tích dữ liệu hiệu quả.

**Kết nối với SQL Server:**

Đầu tiên, sử dụng Power BI để tạo một kết nối đến cơ sở dữ liệu SQL Server. Điều này cho phép truy cập dữ liệu trực tiếp từ cơ sở dữ liệu SQL Server mà không cần sao chép dữ liệu.

**Lấy dữ liệu, Xử lý và biến đổi dữ liệu:**

Sau khi thiết lập kết nối, lựa chọn cơ sở dữ liệu và bảng cần lấy dữ liệu từ SQL Server. Tùy chỉnh truy vấn để lấy dữ liệu phù hợp với mục đích phân tích. Sử dụng Power BI để thực hiện các biến đổi dữ liệu cần thiết, như lọc dữ liệu không cần thiết, kết hợp bảng, tính toán cột mới và định dạng dữ liệu.

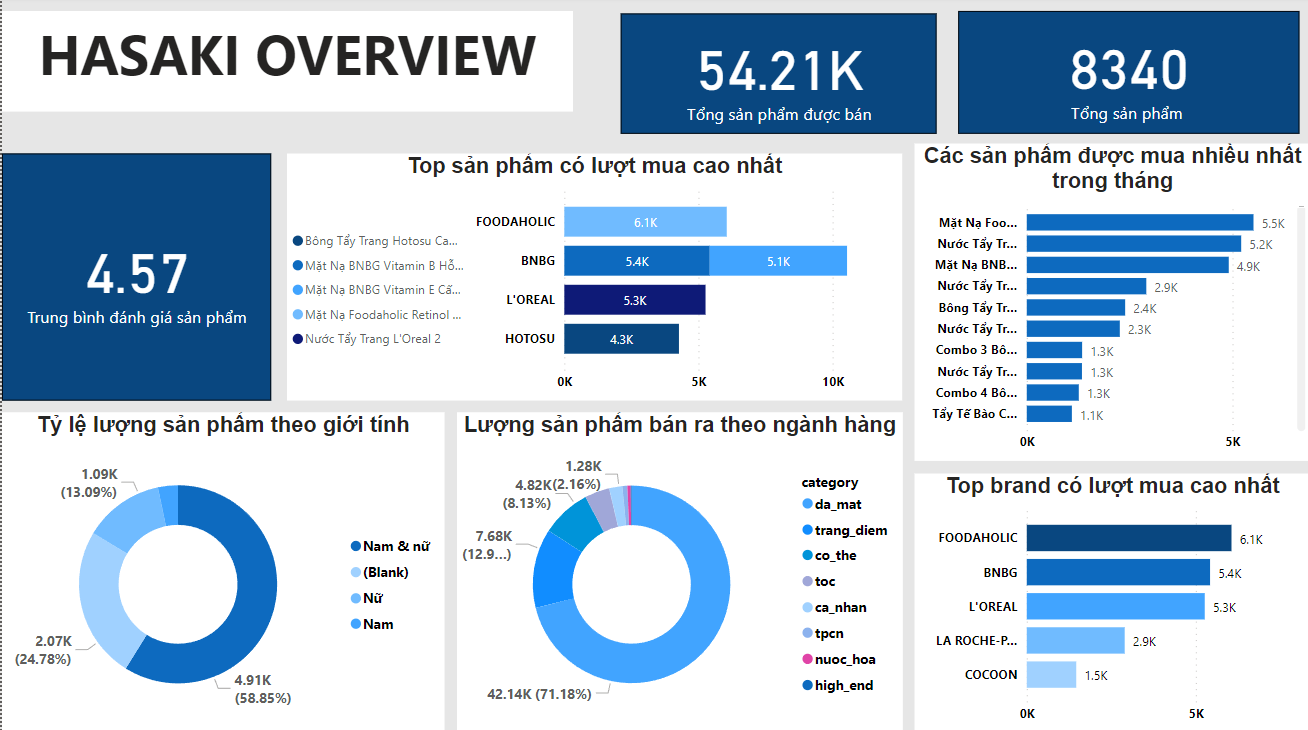
**Tạo báo cáo**

Tạo các biểu đồ và đồ thị trực quan dựa trên dữ liệu đã lấy. Sử dụng các tính năng của Power BI để tạo biểu đồ cột, biểu đồ bánh donut, biểu đồ dòng, và các loại biểu đồ khác để hiển thị thông tin một cách rõ ràng. Sắp xếp biểu đồ và thông tin để tạo các trang báo cáo. Thêm tiêu đề, mô tả, và các phần giải thích để giải thích dữ liệu.

**Chia sẻ và xuất báo cáo**

Cuối cùng, chia sẻ báo cáo hoặc xuất dữ liệu trực quan từ Power BI để chia sẻ với đồng nghiệp hoặc khách hàng. Có thể xuất báo cáo dưới dạng tệp PDF hoặc đăng tải lên dịch vụ Power BI để chia sẻ trực tuyến.

# **3. Dashboard**

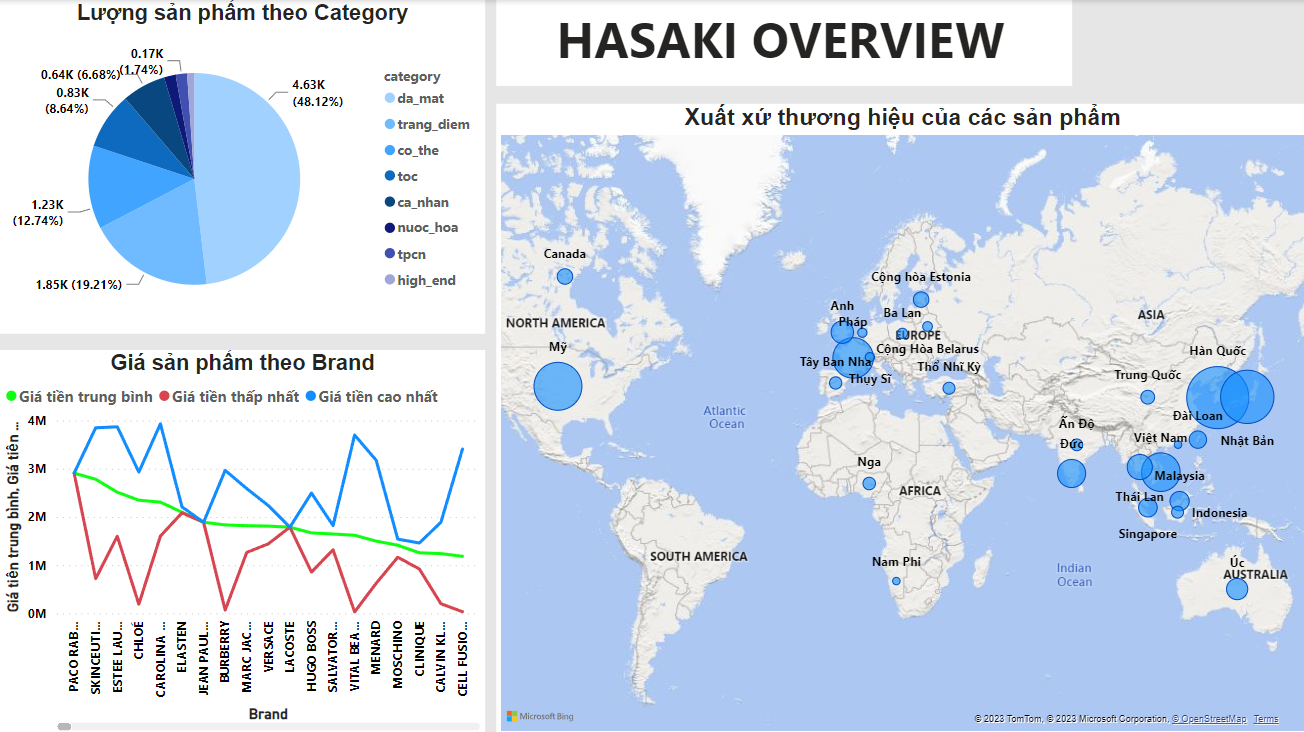


Dashboard Overview thể hiện cái nhìn tổng quan về các sản phẩm của trang Web Hasaki trong vòng 30 ngày. Hasaki quản lý tổng cộng 8,340 sản phẩm và trong vòng 30 ngày, họ đã bán tổng cộng 54,210 sản phẩm. Điểm trung bình đánh giá của các sản phẩm trên trang web của Hasaki là 4.57, được nhóm xem xét là một kết quả tương đối tích cực. Điều này cho thấy rằng Hasaki đã thu hút sự hài lòng của đa số người dùng.

Đến ngày 28/09, 5 sản phẩm có lượt mua cao nhất thuộc các brand Foodaholic, BNBG, L’oreal, Hotosu đã thu hút nhiều sự quan tâm từ người mua. Ngoài ra, 10 sản phẩm có lượt mua cao nhất không chỉ là 10 sản phẩm bán chạy trong thời gian gần đây mà cũng là 10 sản phẩm có lượng bán ra cao nhất từ trước đến giờ.

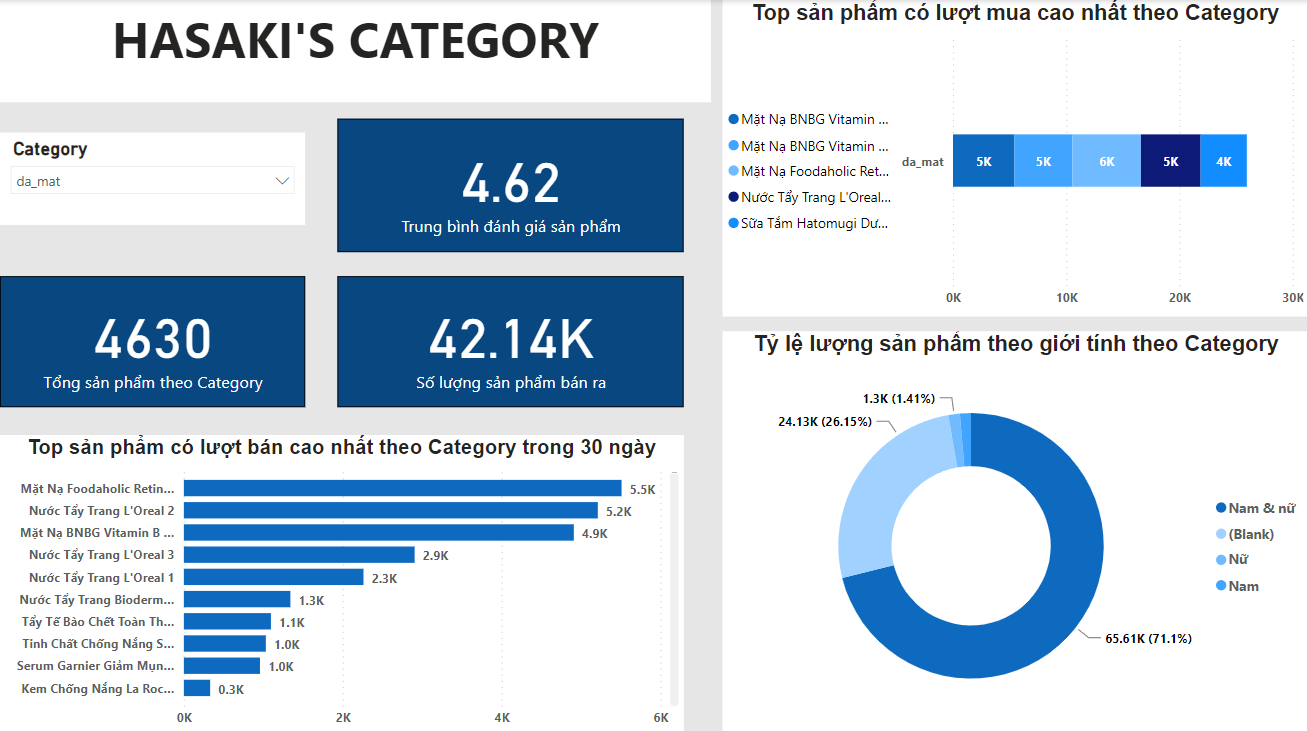
Hiện nay, khoảng 60% sản phẩm làm đẹp phục vụ cả nam và nữ. Điều này thể hiện sự tiến bộ trong quan điểm về chăm sóc sắc đẹp, không chỉ dành riêng cho phái nữ. Nam giới ngày càng quan tâm đến việc làm đẹp của họ. Sự thay đổi này được thúc đẩy bởi sự phát triển của thị trường làm đẹp nam giới, và dẫn đến sự phát triển mạnh mẽ của sản phẩm làm đẹp dành cho cả nam và nữ trong hiện tại và trong tương lai.

Ta có thể thấy có sự ưu tiên rõ ràng đối với việc chăm sóc da mặt, với 71,18% sản phẩm hướng nằm trong danh mục này. Điều này cho thấy mọi người quan tâm đến việc duy trì sức khỏe và làm đẹp cho làn da. Trong các thương hiệu được ưa chuộng, 2 brand đứng đầu là Foodaholic và BNBG có danh tiếng mạnh trong việc chăm sóc da mặt, đặc biệt là trong việc bán các loại mặt nạ, thể hiện sự quan tâm của mọi người đối với việc làm đẹp da mặt. Ngoài ra các sản phẩm của brand Cocoon - một thương hiệu mỹ phẩm thuần chay đến từ Việt Nam. Mặc dù mới ra mắt sau một thời gian, nhưng Cocoon đã có lượt bán tương đối cao, cho thấy sự quan tâm ngày càng gia tăng của người tiêu dùng đối với sản phẩm bảo vệ môi trường. Điều này thể hiện xu hướng quan tâm đến việc chăm sóc da mặt và đồng thời góp phần vào việc bảo vệ môi trường.



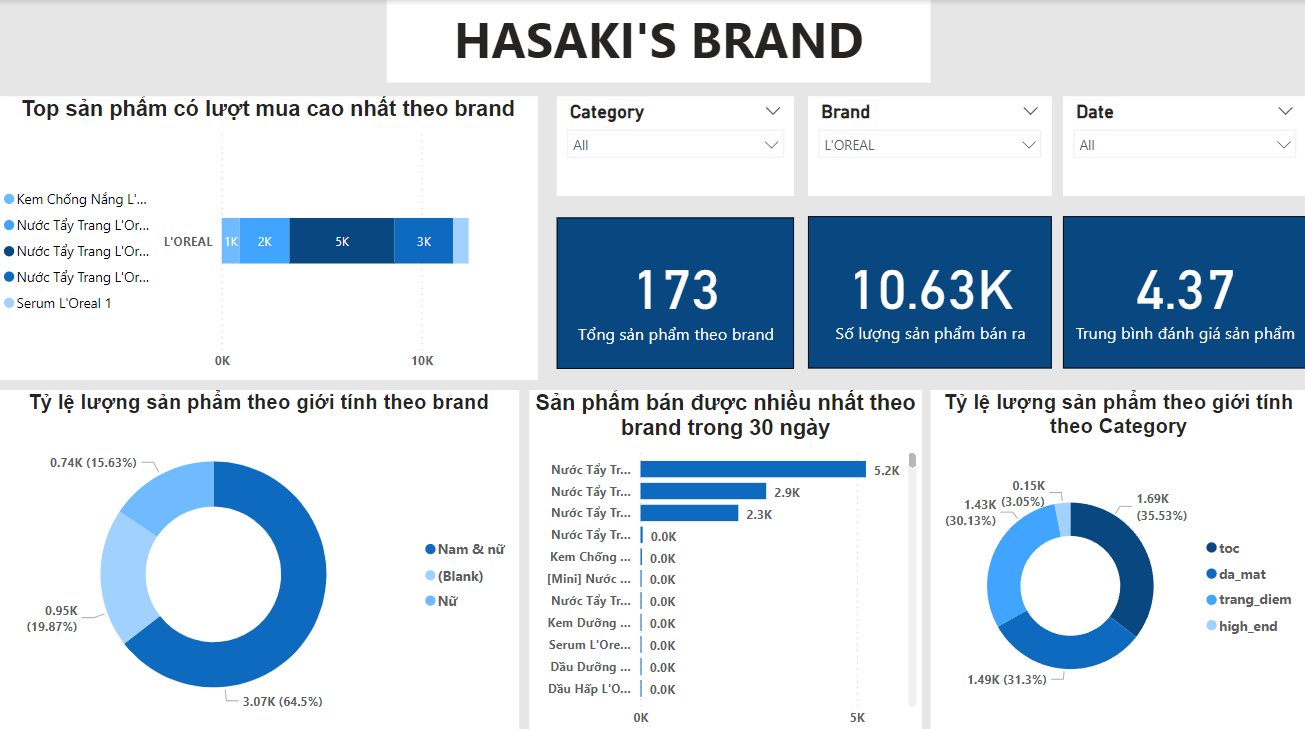
Nhóm đã xây dựng một biểu đồ để xác định xuất xứ của các thương hiệu mỹ phẩm, và ta thấy rằng phần lớn sản phẩm của các thương hiệu nổi tiếng đến từ các quốc gia như Hàn Quốc, Nhật Bản, Pháp và Mỹ. Tuy nhiên, không nên bỏ qua sự phát triển của thương hiệu mỹ phẩm Việt Nam, với nhiều sản phẩm được người tiêu dùng đón nhận điều đó được thể hiện qua việc Hasaki bán cũng có nhiều sản phẩm trong nước. Sự thịnh hành của các sản phẩm Hasaki trên thị trường nội địa cho thấy sự quan tâm ngày càng tăng đối với mỹ phẩm trong nước, và không chỉ dừng lại ở sản phẩm nước ngoài.

Dữ liệu về giá cả của các sản phẩm được biểu thị dưới dạng biểu đồ đường, cho phép chúng ta thấy rõ sự chênh lệch về giá cả theo từng thương hiệu. Với khoảng giá trung bình từ vài triệu đến vài ngàn đồng. Dựa trên dữ liệu này, ta thấy Hasaki có thể đáp ứng nhu cầu mua sắm của các phân khúc khách hàng cụ thể, ví dụ như học sinh - sinh viên, và cung cấp các sản phẩm phù hợp với họ.



Tiếp theo, nhóm tiến sâu hơn vào việc xem xét từng danh mục sản phẩm một. Ví dụ, khi xem xét danh mục chăm sóc da mặt, chúng tôi tìm thấy tổng cộng 4,630 sản phẩm đã được kinh doanh, và trong số này, cửa hàng đã bán được 42,140 sản phẩm. Điều đáng chú ý là sản phẩm trong danh mục này đạt được điểm đánh giá trung bình là 4.62, thể hiện mức độ hài lòng của khách hàng và sự tin tưởng trong sản phẩm.

Dựa vào biểu đồ tròn, chúng ta có thể nhận thấy rằng, ngoại trừ nước hoa có đa số người tiêu dùng trong các danh mục này là nữ, các danh mục sản phẩm khác đều có sự đa dạng về sản phẩm dành cho cả nam và nữ. Lý do cho điều này là do nước hoa ít có loại dành cho cả nam và nữ (nước hoa unisex), nước hoa thường được phân chia rõ ràng giữa hai loại, dẫn đến việc phụ nữ chiếm đa số trong mua sắm sản phẩm này.



Nhóm tiếp tục tiến hành phân tích dựa trên từng thương hiệu cụ thể. Lấy ví dụ về thương hiệu L'Oréal, chúng tôi tìm thấy tổng cộng 173 sản phẩm thuộc thương hiệu này đã được kinh doanh, và trong số này, cửa hàng đã bán được 10,630 sản phẩm. Đáng chú ý, sản phẩm của thương hiệu L'Oréal có điểm đánh giá trung bình là 4.37, cho thấy mức độ hài lòng và sự tin tưởng của khách hàng đối với thương hiệu này.

Theo giới tính của khách hàng, chúng ta thấy rằng sản phẩm dành cho cả nam và nữ đều được ưa chuộng và có sự quan tâm. Loreal không có sản phẩm dành riêng cho nam. Về tỉ lệ sản phẩm, sản phẩm của L’oreal tập trung chủ yếu vào các danh mục chăm sóc tóc, da mặt và trang điểm chiếm tỷ trọng lớn trong cửa hàng. Ngoài ra, có một số sản phẩm dành cho các danh mục khác nhưng chiếm tỷ trọng nhỏ hơn. L’oreal còn được biết đến với các sản phẩm chất lượng và đa dạng, với một phần sản phẩm cao cấp phục vụ cho những khách hàng có nhu cầu sử dụng sản phẩm cao cấp.

Tổng số lượng sản phẩm L’oreal bán ra chiếm một phần quan trọng trong tổng số sản phẩm bán ra của cửa hàng, với tỷ lệ khoảng 20%. Sự thành công của L’oreal có thể được giải thích bằng sự phù hợp về giá cả, chất lượng và sự đa dạng của danh mục sản phẩm, đáp ứng nhu cầu của nhiều loại khách hàng khác nhau. L’oreal đã đạt được một lượng bán hàng đáng kể, cho thấy tầm quan trọng của thương hiệu này trong ngành công nghiệp làm đẹp.

Sau khi phân tích dữ liệu sức khỏe làm đẹp, ta thấy rằng ngành làm đẹp đang trở nên ngày càng quan trọng và phát triển nhanh chóng. Phân tích dữ liệu sức khỏe và làm đẹp có thể giúp cải thiện dịch vụ và sản phẩm trong ngành, giúp xác định xu hướng mới, thay đổi trong thị trường sức khỏe và làm đẹp, và cung cấp thông tin hữu ích cho quyết định kinh doanh.

Tất cả những điểm trên cho thấy tại sao phân tích dữ liệu sức khỏe và làm đẹp có thể là một đề tài quan trọng và hấp dẫn. Nó cung cấp lợi ích cho ngành công nghiệp, cá nhân và nghiên cứu trong lĩnh vực sức khỏe và làm đẹp.