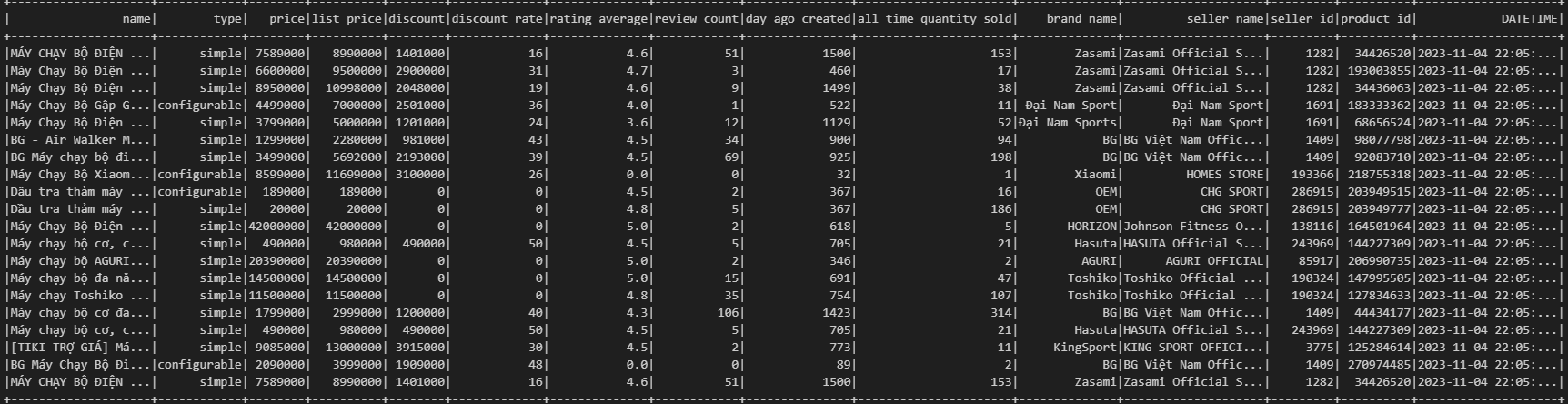
[Project Tiki Gym Gear Crawling](https://sites.google.com/view/tiki-data-crawling/trang-ch%E1%BB%A7)

1. Giới thiệu:

* Tiki là trang thương mại điện tử lọt top 2 tại Việt Nam và top 6 tại khu vực Đông Nam Á. Ở đây, các loại mặt hàng rất đa dạng, chứa nhiều cơ hội cho cả người bán lẫn người mua. Vì vậy, việc crawl data từ sàn thương mại điện tử này có thể cung cấp các dữ liệu có giá trị trong việc lựa quyết định mở cửa hàng tại đây như thông tin đối thủ, khẩu vị của khách hàng.

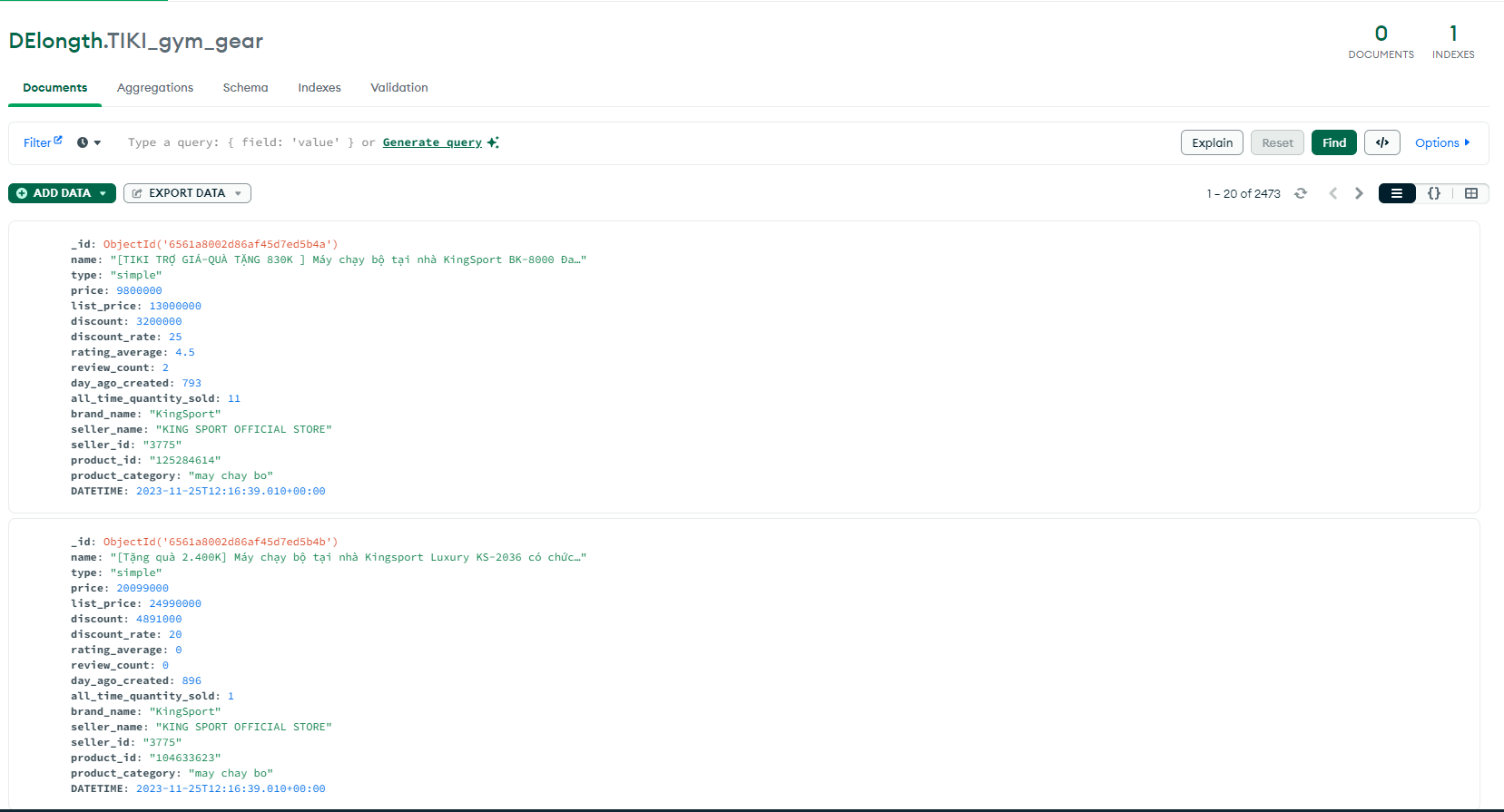
1. Phương pháp:

* Ban đầu, sử dụng thư viện Beautiful Soup để crawl danh sách các product\_id. Với product\_id, sử dụng url của Tiki để đi tới trang thông tin sản phẩm tương ứng với từng product\_id và lấy các trường dữ liệu quan trọng như: thông tin người bán, rating, giá thành sản phẩm, vv.
* Sau đó, transform raw data bằng Pyspark: chọn lọc một số trường thông tin quan trọng như tên người bán, tên sản phẩm, nhãn hiệu, giá thành:



Hình 1. Clean data sau khi crawl và transform.

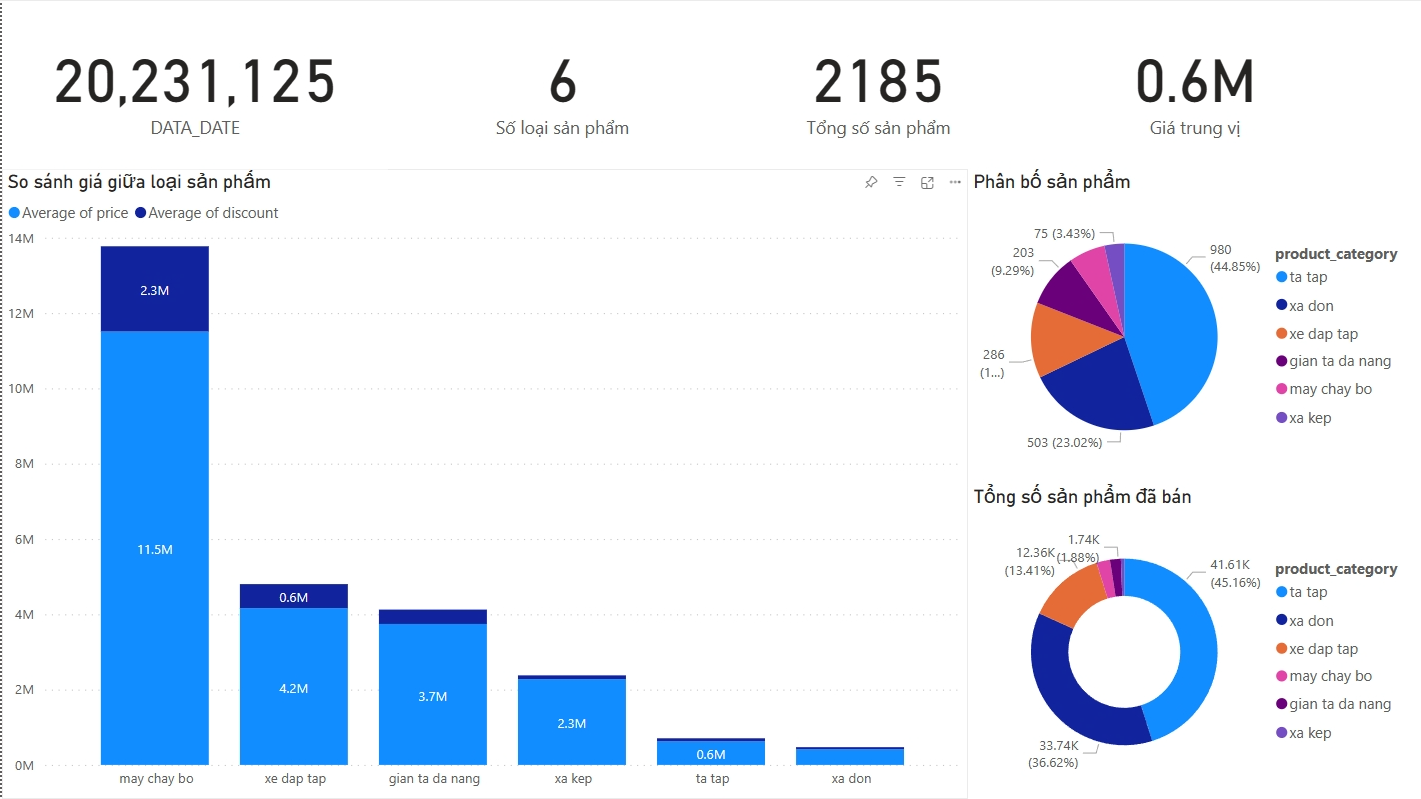
* Sau khi dữ liệu được làm sạch, connect python với MongoDB và lưu trữ data vào đây.



* + Sau đó, sử dụng upload data lên PowerBI service để phục vụ visualizing.

1. Key insights:
2. Tổng quan:

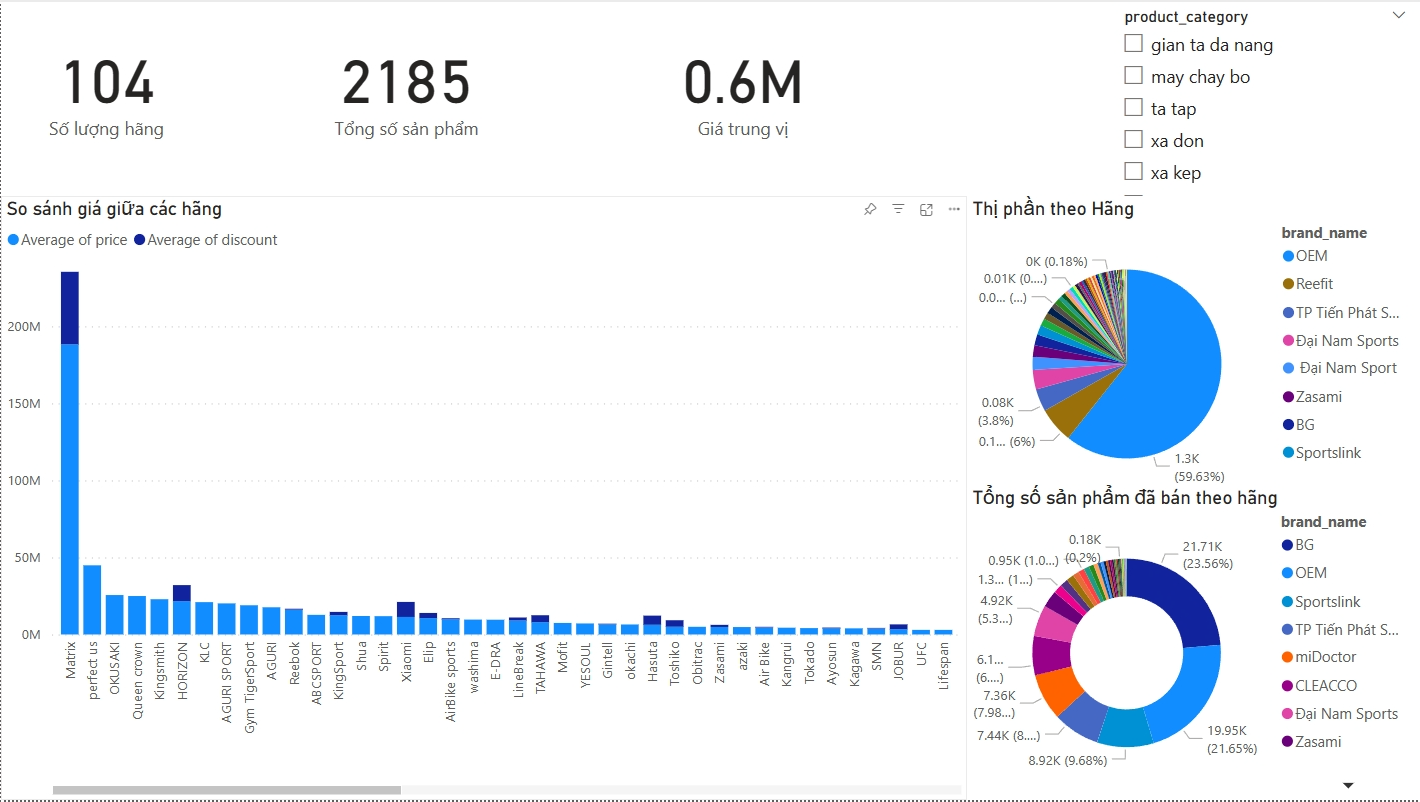
* Trong ngày 20231125, thu thập được thông tin từ 2185 sản phẩm, chia làm 6 nhóm sản phẩm chính: máy chạy, xe đạp, giàn tạ, xà đơn, xà kép, và tạ tập.



* Từ hình mình họa, có thể rút ra được insight đầu tiên: sản phẩm tạ tập và xà đơn, do giá thành thấp (dưới 600k) và cấu tạo đơn giản, trở thành sản phẩm có luợt bán nhiều nhất, và hiện cũng được bày bán rộng rãi nhất.
* Ngược lại, máy chạy bộ, sản phẩm vốn đắt đỏ và cần nhiều không gian hơn, được bày bán ít hơn và có all-time sold thấp.
* Đáng chú ý, xà kép tuy có giá thành thấp thứ 3, nhưng lại không nhận được nhiều sự chú ý do khách hàng hiện tại hài lòng với xà đơn, một sản phẩm đem lại hiệu suất tập luyện ít hơn, nhưng lại tiện dụng, dễ lắp đặt và rẻ hơn đáng kể.

1. Tình hình sản phẩm:

* Nhìn vào tổng thể, BG và OEM có vẻ là 2 hãng lớn nhất khi cả số lượng sản phẩm đã bán ra của cả 2 chiếm gần một nửa thị trường.



* Tuy số lượng mẫu mã hãng OEM đang được bày bán khá nhiều (60%), nhưng số lượng các sản phẩm đã bán chỉ khoảng 20%. Điều này cho thấy hãng này chỉ đa dạng về sản phẩm, chứ không thực sự chiếm được sự tin dùng của thị trường.
* Ngược lại, hãng BG hiện chỉ có 40 model sản phẩm, chiếm chưa đến 1% tổng số lượng model đang bày bán, nhưng đã bán đến 21 nghìn sản phẩm, chiếm gần ¼ thị trường. Điều này cho thấy, đây là một hãng lớn, có sức hút riêng cho từng model của mình.

1. Phân tích người bán

* Có thể thấy, hiện nay, BG store đang dẫn đầu thị trường với 21 nghìn sản phẩm đã được bán ra. Nhìn vào rating của store, con số khá ổn với 5 nghìn lượt đánh giá và rating 3.54.
* Ngoài ra, có thể thấy là sản phẩm tạ tập đang được các cửa hàng rất chú trọng, khi số lượng model của sản phẩm này hoàn toàn vượt trội so với loại sản phẩm khác.
* Dựa vào chỉ số rating, có thể thấy một điểm sáng nữa là Boro sport, khi hãng này nhận được đánh giá khá cao ở 4.92 sau 4,500 lượt đánh giá. Hiệu suất vânj hành cũng thể hiện ở số lượng sản phẩm, khi hãng này đã bán tới 6,000 sản phẩm.
* Điều này tạo nên nghi vấn về việc: Liệu đa dạng hóa sản phẩm có phải là lựa chọn tốt, khi 2 shop dẫn đầu về số lượng model sản phẩm là Đại Nam, và Phúc Thành, chỉ nhận được rating rất thấp, kèm theo đó là số lượng bán ra cực kì thấp.

