**PHÂN TÍCH RFM VÀ COHORT TRONG BỘ DỮ LIỆU RETAIL STORE CỦA OLIST**

**Objective:**

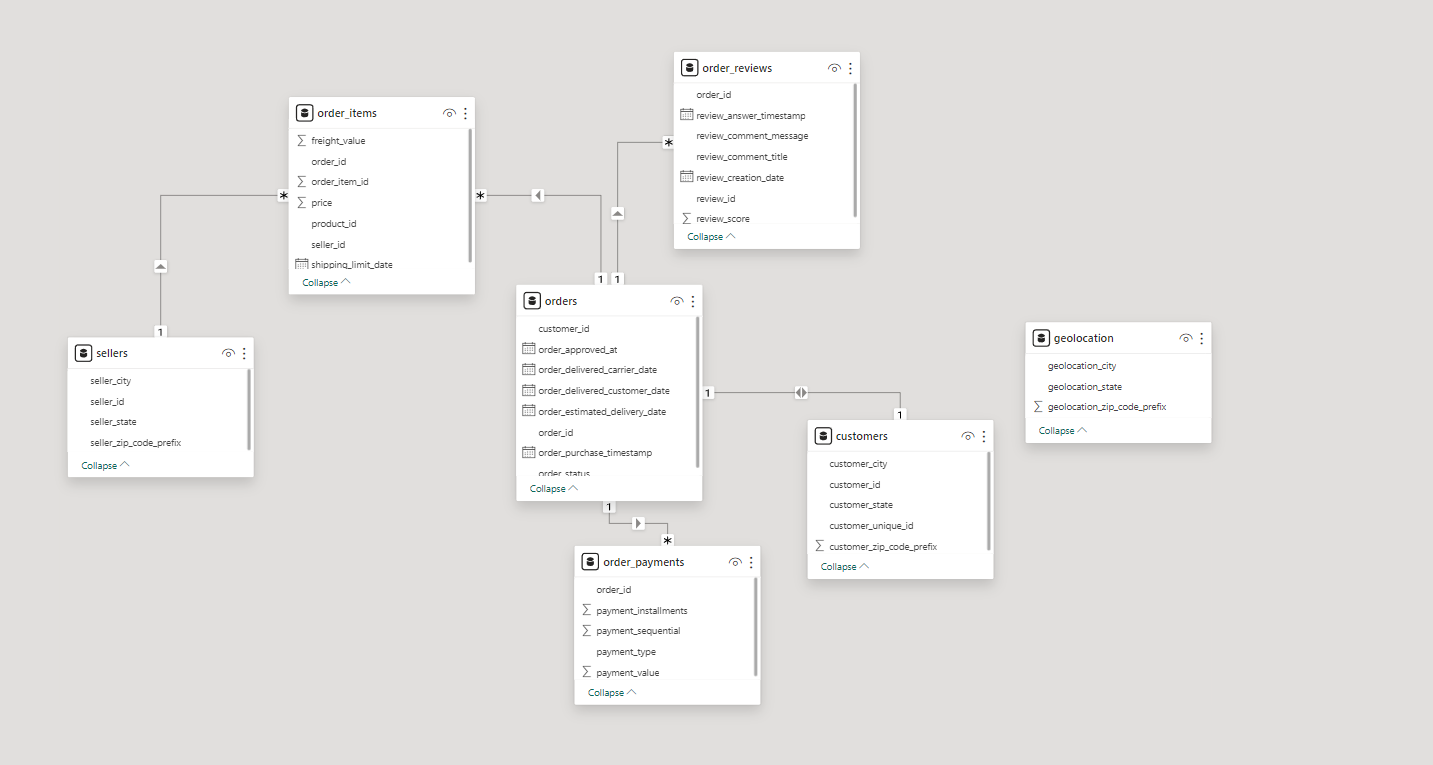
Mục tiêu của tôi trong dự án này là phân tích hành vi mua sắm của khách hàng dựa trên bộ dữ liệu của Olist - tập đoàn siêu thị lớn nhất Brazil thông qua việc ứng dụng phương pháp phân tích RFM và phân tích Cohort nhằm xác định phân khúc khách hàng giá trị, và đưa ra các khuyến nghị nhằm tăng tỉ lệ giữ chân khách hàng

**Demonstrated Skills:**

* **Data Cleaning and Preparation:** Tôi sẽ dùng python để kiểm tra và làm sạch dữ liệu bao gồm xử lý các dữ liệu bị thiếu, tìm các số liệu bất thường
* **RFM Analysis:** Tôi sẽ tính điểm RFM (Recency, Frequency, Monetary) để phân loại khách hàng thông qua DAX trong PowerBi
* **Cohort Analysis:** Tôi sẽ thực hiện phân tích Cohort để tính tỉ lệ giữ chân khách hàng thông qua SQL
* **Data Visualization:** Tôi sẽ sử dụng PowerBi để trực quan hóa các số liệu qua các biểu đồ cột, biểu đồ đường,....
* **Strategic Insights and Recommendations:** Dựa trên phân tích RFM và Cohort, tôi sẽ đưa ra các insight và khuyến nghị phù hợp

**Tools:**

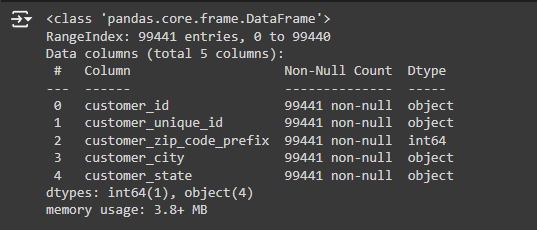
* **Google Colab:** Sử dụng Python thông qua google Colab để kiểm tra và làm sạch dữ liệu
* **Microsoft Sql Server:** Dùng để import dữ liệu và tính bảng phân tích Cohort
* **PowerBi:** Dùng để trực quan hóa dữ liệu, tính RFM

**Tổng quan dữ liệu**

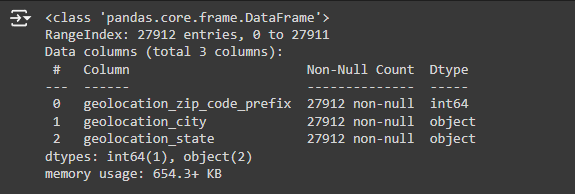
Đây là bảng relationship biểu diễn mối quan hệ giữa các bảng trong dataset của OLIST, tổng quan dữ liệu bao gồm các dữ liệu về đơn hàng, khách hàng, sản phẩm, người bán và thanh toán

**Task 1: Kiểm tra dữ liệu**

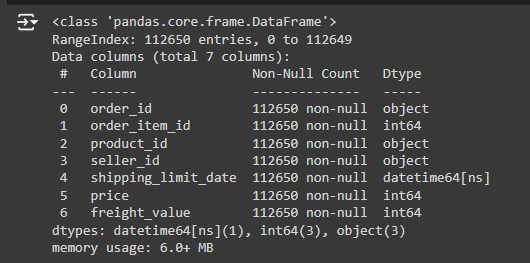
Trước hết, tôi sẽ kiểm tra loại dữ liệu của các dataframe xem liệu chúng đã hợp lệ chưa



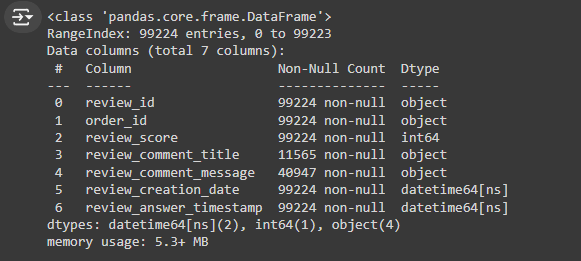
Bảng . Customers



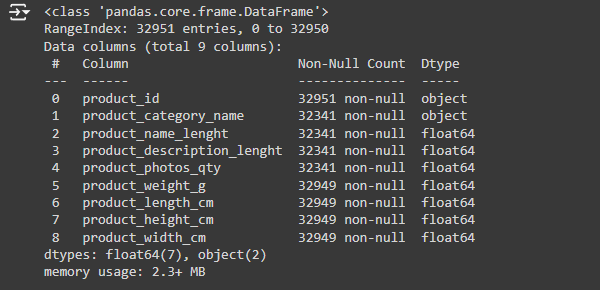
Bảng . Geolocation



Bảng . order\_items



Bảng . order\_reviews



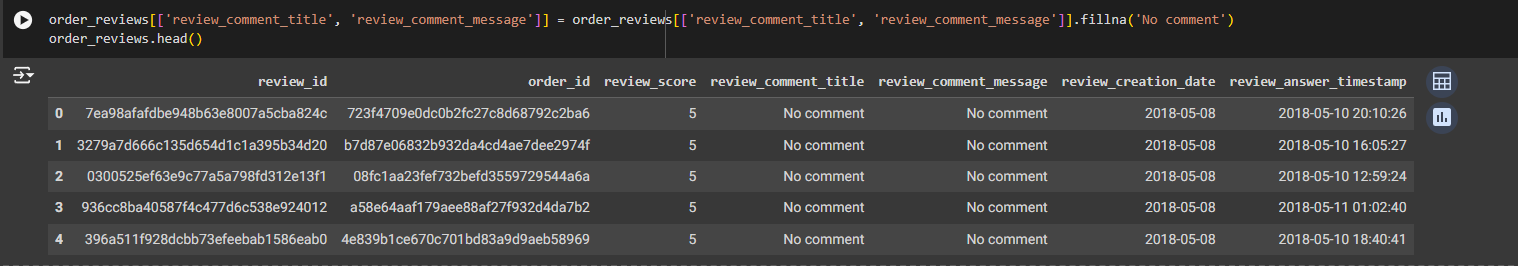
Bảng . products

‘

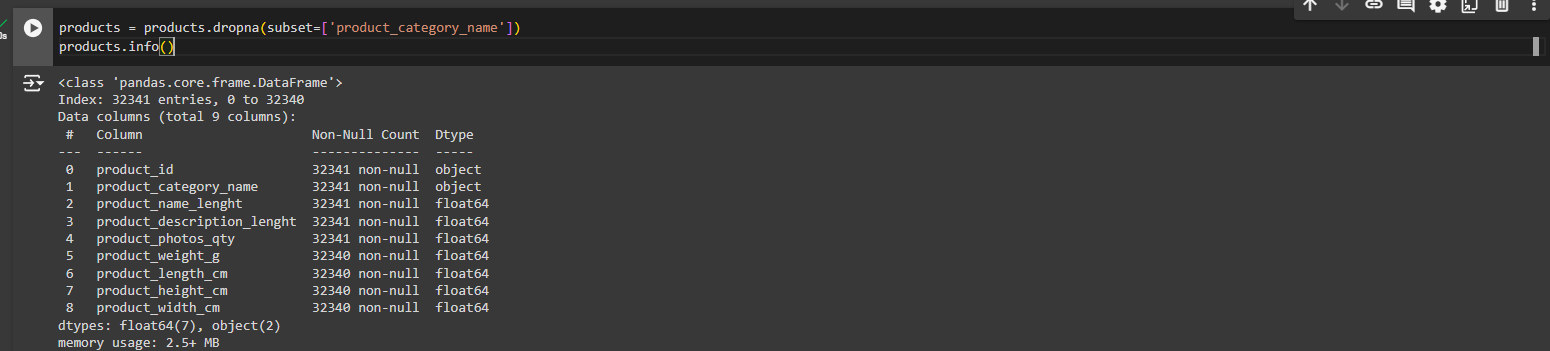
Vậy là sau khi kiểm tra các bảng thì dtype của chúng đều hợp lệ, hầu hết các bảng đều không có dữ liệu null, tuy nhiên thì ở bảng order\_reviews, mục review\_comment\_message và title bị trống, khả năng cao là do người dùng đánh giá mà không để lại comment, mục products cũng có sản phẩm bị thiếu

**Task 2:** **Xử lý dữ liệu bị thiếu**

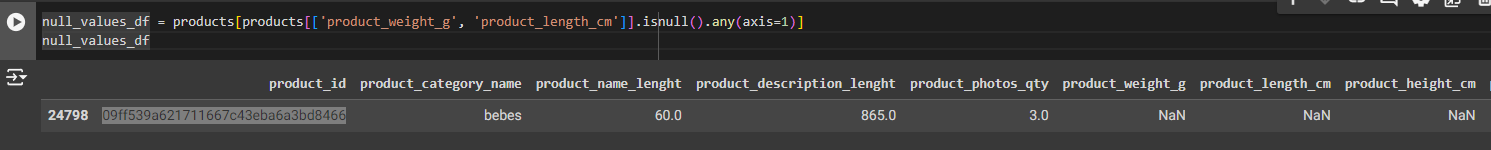
Đầu tiên là tôi sẽ thay thế các giá trị null trong bảng order\_reviews bằng “No comment”, thể hiện rằng đánh giá không để lại comment



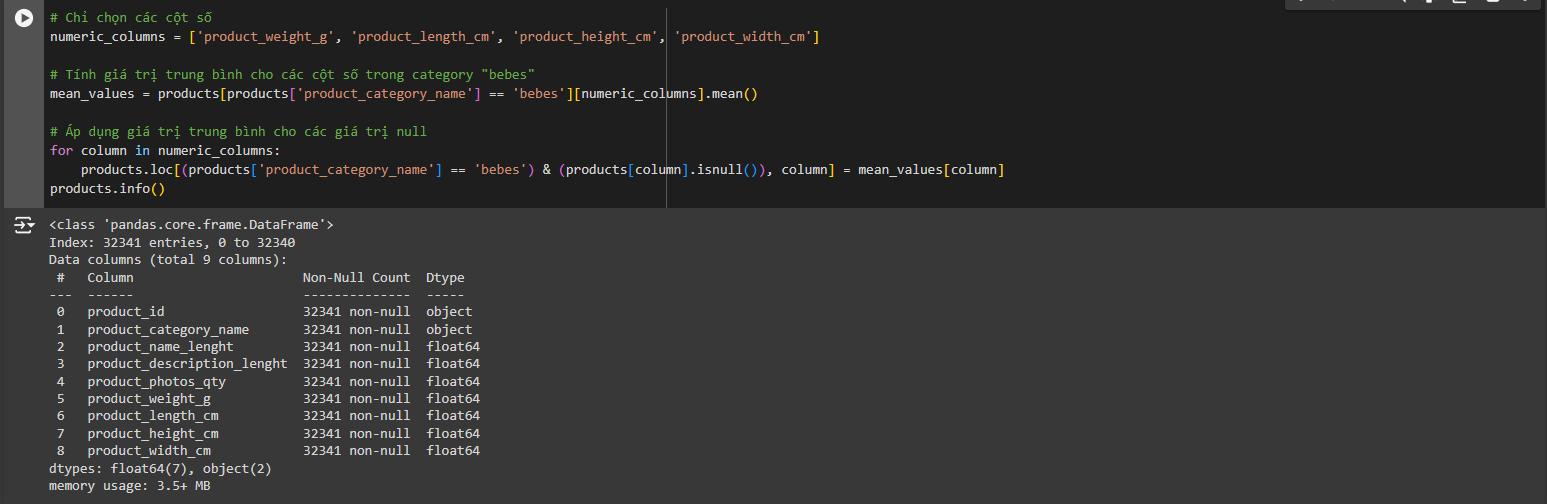
Tiếp theo tôi sẽ xóa các hàng mà cột product\_category\_name bị thiếu trong bảng products vì các tên sản phẩm đã được encode nên ta không thể phân tích bằng tên sản phẩm mà thiếu product\_category\_name được



Sau khi xóa các dòng chứa dữ liệu null ở cột product\_category\_names thì còn sót lại duy nhất 1 giá trị null ở cột product\_weight\_g và product\_length\_cm



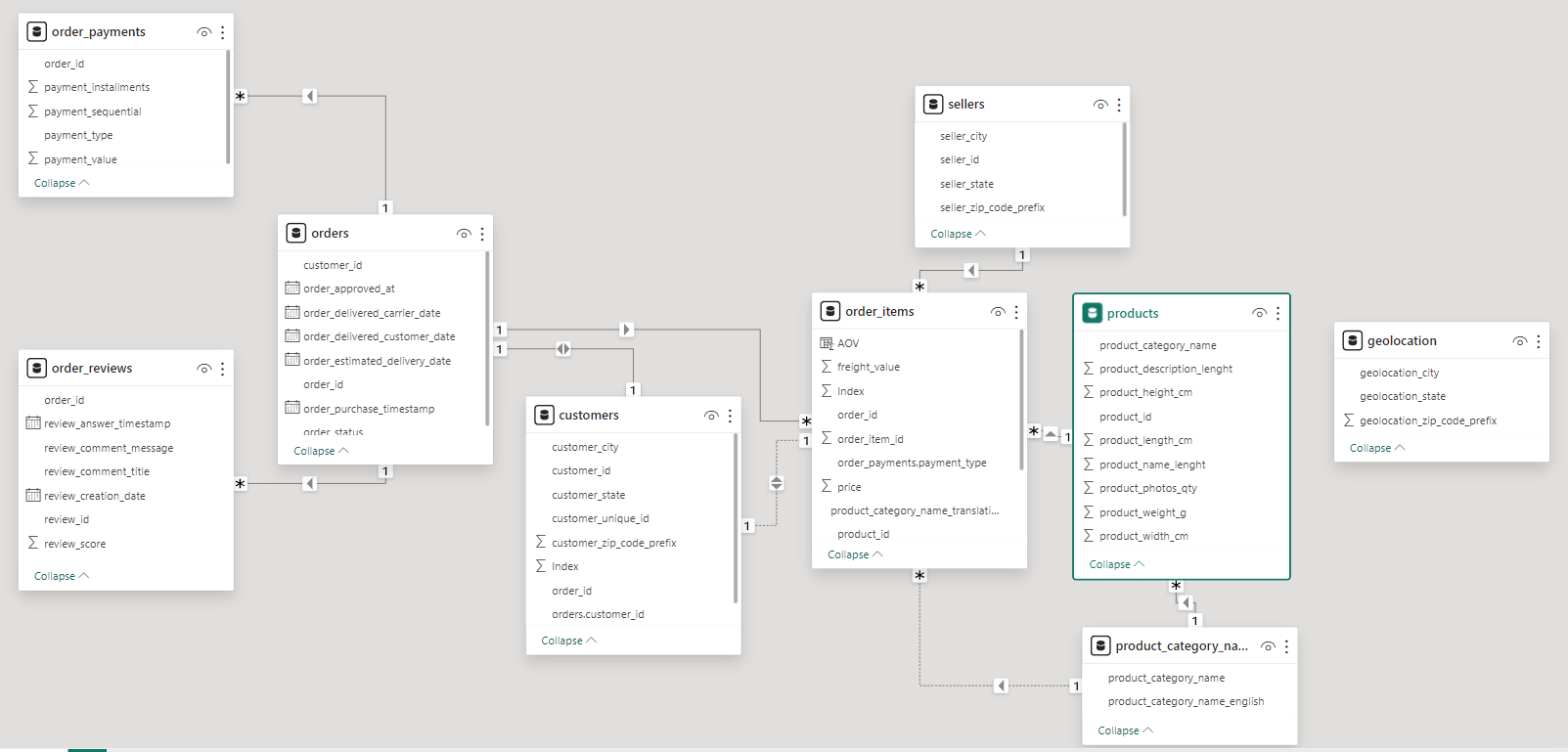
Chỉ còn lại một quan sát duy nhất chứa dữ liệu null và nó nằm ở hàng bebes, ta sẽ lấy trung bình siêu thị bebes và replace vào cột bị thiếu



Sau khi thay thế giá trị null thì giờ ta đã có một dataset hoàn chỉnh, tiếp theo tôi sẽ thực hiện join các bảng lại với nhau trong PowerBi. Vì các bảng đều có cột chung nên việc join các bảng sẽ khá đơn giản và PowerBi có thể làm việc đó nhanh và gọn, đồng thời chúng cũng sẽ thuận tiện để giúp tôi trực quan hóa dữ liệu.

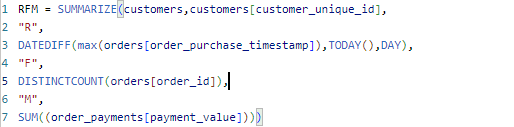
**Task 3: Join các bảng lại với nhau bằng PowerBi**

Sau khi join các bảng lại với nhau ta có relationship table giữa các bảng dưới đây, việc join các bảng lại với nhau sẽ giúp chúng ta dễ tham chiếu thông tin hơn, đồng thời sẽ thuận tiện trong quá trình visualize các dữ liệu tương quan với nhau

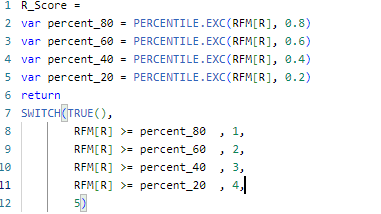
****

**Task 4: Tính RFM bằng PowerBi**

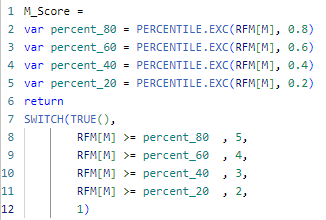
Đầu tiên ta sẽ tạo bảng RFM dựa trên các dự liệu đã có trong bảng. Cột R- Recency sẽ tương ứng với lần mua gần đây nhất, ta dùng hàm datediff để tính số ngày từ lần cuối mua đến hiện tại. Cột F-Frequency sẽ là tần suất giao dịch của khách hàng, ta sẽ đếm số lần mua hàng của khách bằng hàm DISTINCTCOUNT. Cột M sẽ là tổng số tiền khách hàng đã chi, ta sẽ dùng hàm sum để tính chúng

****

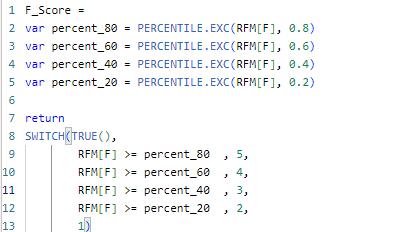
Tiếp theo, chúng ta sẽ tạo cột R\_Score để gán điểm cho khách hàng dựa trên giá trị Recency. Đầu tiên, chúng ta tính các phần trăm vị trí (80%, 60%, 40%, 20%) của giá trị Recency trong tập dữ liệu. Dựa vào các ngưỡng này, mỗi khách hàng sẽ được gán một điểm từ 1 đến 5: điểm 1 cho khách hàng ít mua gần đây nhất (cao hơn hoặc bằng 80%) và điểm 5 cho khách hàng mua gần đây nhất (dưới 20%

****

Tiếp tục với cột M\_Score chúng ta sẽ gán điểm cho khách hàng dựa trên giá trị Monetary. Đầu tiên, chúng ta tính các phần trăm vị trí 80%, 60%, 40%, và 20% của giá trị Monetary trong tập dữ liệu. Dựa vào các ngưỡng này, mỗi khách hàng sẽ được gán một điểm từ 1 đến 5: điểm 5 cho khách hàng có giá trị mua hàng cao nhất (trên 80%) và điểm 1 cho khách hàng có giá trị mua hàng thấp nhất (dưới 20%)



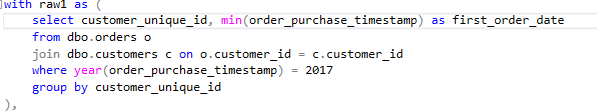
Tương tự với cột F\_Score, ta cũng sẽ đặt ra các ngưỡng, mỗi khách hàng sẽ được gán một điểm từ 1 đến 5: điểm 5 cho khách hàng có tần suất mua hàng cao nhất (trên 80%) và điểm 1 cho khách hàng có tần suất mua hàng thấp nhất (dưới 20%)



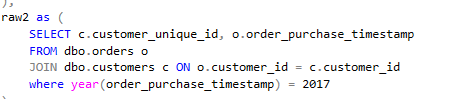
**Task 5: Tạo bảng Cohort**

Ta sẽ tiến hành tạo bảng Cohort bằng SQL để tính lượng khách hàng ở lại và rời đi. Ở đây tôi sẽ tính lượng khách hàng rời đi và ở lại hàng tháng theo từng năm và tôi sẽ bỏ qua năm 2016 vì dữ liệu của năm 2016 chỉ có 2 tháng, các bước cụ thể như sau:

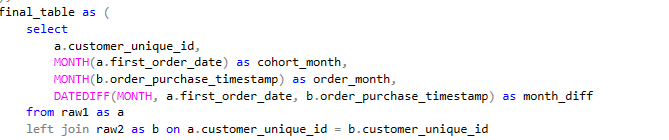
Đầu tiên tôi sẽ tạo 1 CTE chứa thông tin khách hàng và ngày đầu tiên mua hàng của họ trong năm bằng cách chọn ngày mua hàng nhỏ nhất sau khi đã join dữ liệu khách hàng với dữ liệu order.



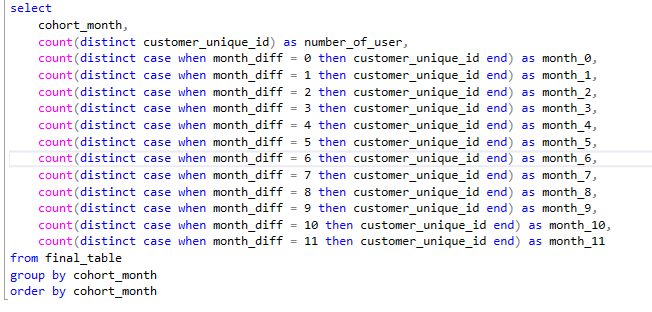
Tiếp theo tôi sẽ tạo thêm 1 CTE chứa ID khách hàng và tất cả các ngày mua hàng của họ trong năm.



Bước tiếp theo chúng ta sẽ join 2 bảng vừa tạo lại với nhau để tìm ra khoảng cách của ngày mua hàng đầu tiên với những ngày mua hàng tiếp theo.bằng hàm datediff



Cuối cùng tôi sẽ tổng hợp lại số khách hàng trong mỗi cohort và sẽ dùng hàm count distinct để trả về số lượng còn lại sau mỗi tháng sau đó sẽ nhóm và sắp xếp lại theo cohort\_month để ta có thể phân tích xu hướng của khách hàng thuận tiện hơn



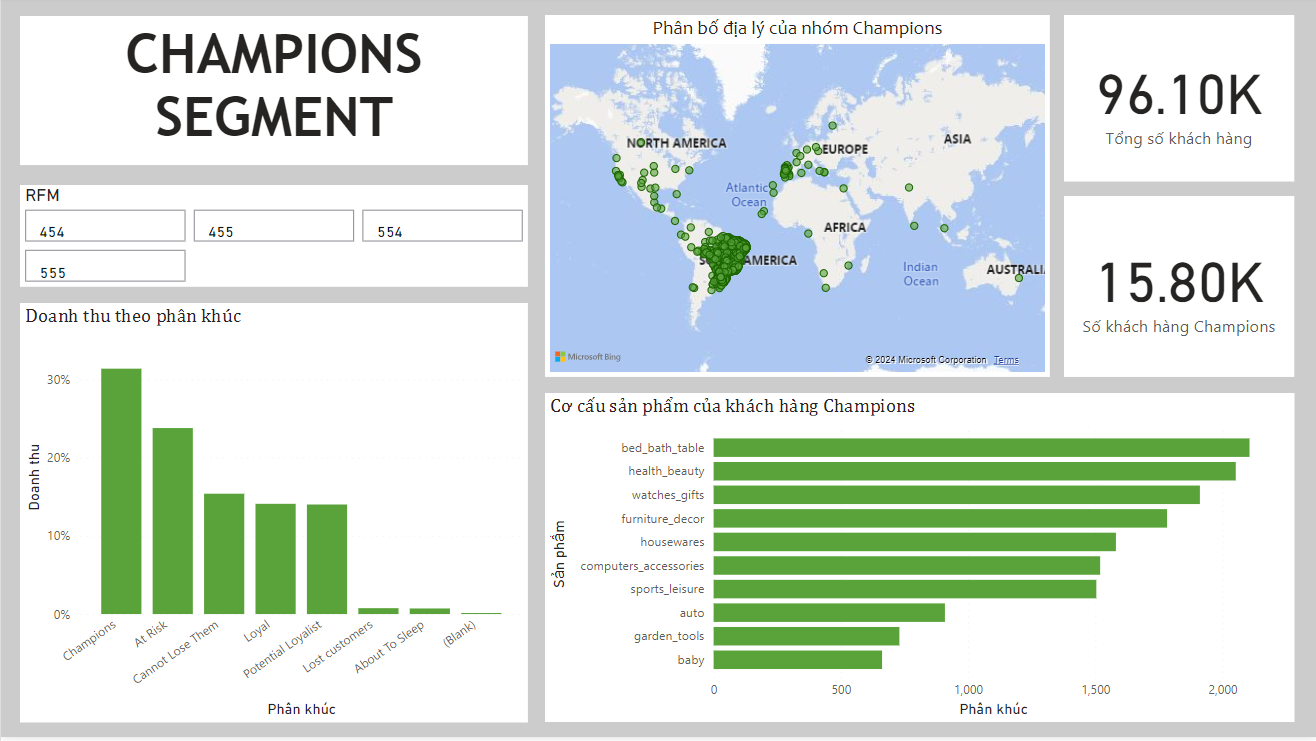
Ta sẽ thực hiện tương tự để tính số lượng khách hàng còn lại trong năm 2018. Sau khi có được bảng cohort, tôi sẽ sao chép chúng vào file excel để tổng hợp, chi tiết các bạn có thể vào github của tôi để xem.

**Visualisation**

Ở phần trực quan hóa dữ liệu, tôi sẽ sử dụng PowerBi để tạo Dashboard. Ở đây tôi sẽ tập trung phân tích vào phân khúc khách hàng dựa trên RFM

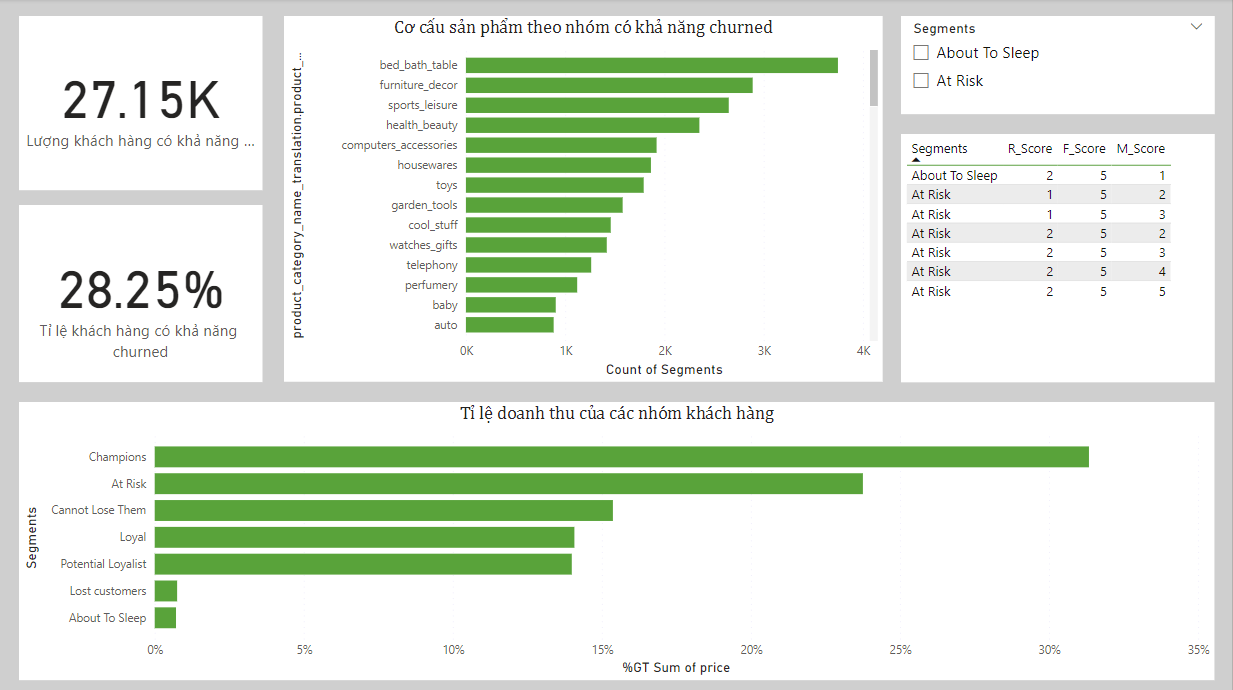
Dưới đây là bảng tổng quan phân tích phân khúc Champions - Phân khúc được coi là giá trị nhất của doanh nghiệp:

* Doanh thu của tệp khách hàng này chiếm tới 30% tổng doanh thu của siêu thị, khẳng định được giá trị của tệp khách này.
* So với quy mô khách hàng, tệp khách Champions có số lượng là 15 nghìn so với 96 nghìn tổng số khách hàng, chiếm hơn 15% tổng số khách. Đây là một con số khá là tích cực, cho thấy doanh nghiệp có một lượng khách trung thành ổn định
* Khách hàng sẽ tập trung chủ yếu ở Nam Mỹ và rải rác ở các vùng lân cận như Bắc Mỹ và một số quốc giá Châu Âu
* Danh mục sản phẩm được mua nhiều nhất của nhóm khách hàng này là đồ nội thất gia dụng (bed\_bath\_table, furniture\_decor, housewares), tiếp đến là các danh mục đồ liên quan đến sức khỏe và làm đẹp, cuối cùng là đồng hồ. Sự kết hợp giữa nội thất, làm đẹp và trang sức gợi ý rằng phân khúc này tập trung vào lối sống chất lượng cao

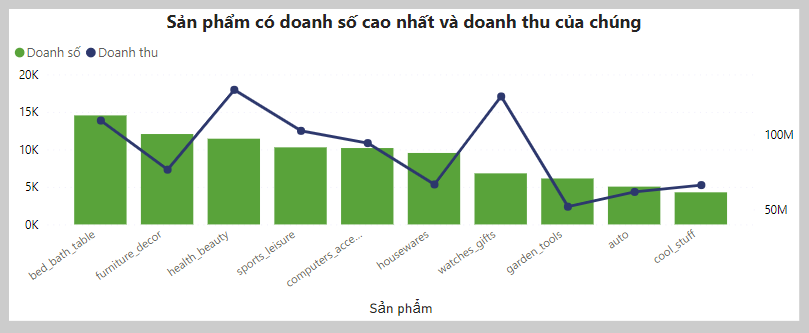
****

Tiếp theo ta sẽ phân tích nhóm khách hàng có khả năng rời bỏ cao nhất, đó là nhóm At Risk và About to sleep, đây là 2 nhóm có lần mua gần nhất xa nhất. Cụ thể ở dashboard dưới đây:

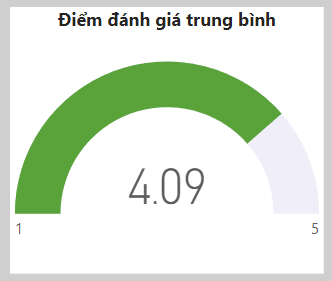
* Tổng số lượng của các tệp này là hơn 27 nghìn khách hàng, chiếm 28.25% tổng số khách hàng. Đây là một tỉ lệ khá đáng kể, cho thấy doanh nghiệp cần có chiến lược giữ chân khách hàng hiệu quả
* Cơ cấu sản phẩm của các nhóm khách hàng này chủ yếu là đồ gia dụng và đồ thể thao, trong đó cơ cấu của bed\_bath\_table chiếm đáng kể trong số này. Một trong những sản phẩm được mua ít nhất của nhóm này là baby, auto, perfumery; Những mặt hàng này cũng ko nằm trong top sản phẩm được mua nhiều nhất của tệp Champions, siêu thị nên phân phối lại danh mục hàng hóa nhằm tăng doanh thu và tỉ lệ giữ chân khách hàng
* Tuy là nhóm khách hàng có tỉ lệ rời bỏ cao, tệp khách At Risk lại có chiếm 20% tỉ lệ doanh thu của siêu thị, chỉ xếp sau tệp Champions, điều này chứng tỏ là lượng khách vãng lai của siêu thị là rất nhiều và họ cũng là một nguồn doanh thu ổn định tiềm năng. siêu thị nên đưa ra các chiến lược để giữ chân tệp khách này lại



Tiếp theo tôi sẽ đánh giá doanh thu của các sản phẩm có doanh số cao nhất. Đứng đầu trong doanh số cao nhất là bed\_bath\_table và funiture\_decor, doanh thu của bed\_bath\_table khá cao nhưng của funiture\_decor lại có phần hơi thấp, điều này chứng tỏ AOV (Average order value) của dòng sản phẩm này thấp. Ngược lại, sản phẩm health\_beauty có doanh số xếp thứ 3 nhưng doanh thu lại nắm vị trí có doanh thu cao gần nhất sau watches\_gift, một sản phẩm cũng có doanh số thấp nhưng doanh thu cao ngất ngưởng, điều này chứng tỏ AOV của health\_beauty và watches\_gift là rất cao. Dòng sản phẩm Garden\_tools,auto,cool\_stuff lại vừa có doanh số thấp vừa doanh thu thấp, siêu thị nên cân nhắc phân phối lại sản phẩm



Để đánh giá khách quan thì tôi sẽ thêm một biểu đồ nho nhỏ là điểm review của khách hàng



Ta có thể thấy điểm đánh giá nằm ở mức khá cao là 4.09/5, ở ngưỡng trên trung bình. Điều này phản ánh chất lượng sản phẩm và dịch vụ là rất tốt, đồng thời giúp nhà hàng xây dựng hình ảnh uy tín và có độ tin cậy cao trong mắt khách hàng. Vì vậy vấn đề cần giải quyết của siêu thị sẽ không phải là cải thiện chất lượng sản phẩm hay dịch vụ mà là ở những khía cạnh khác.

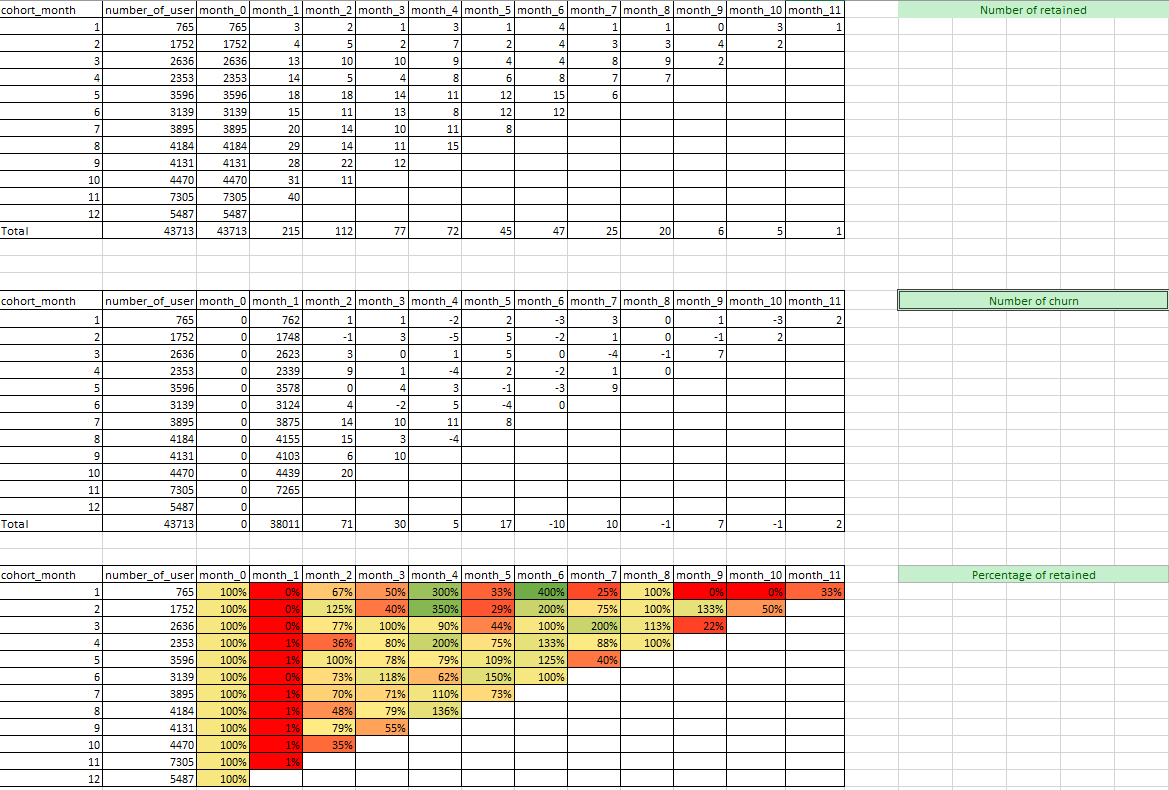
**Phân tích cohort**

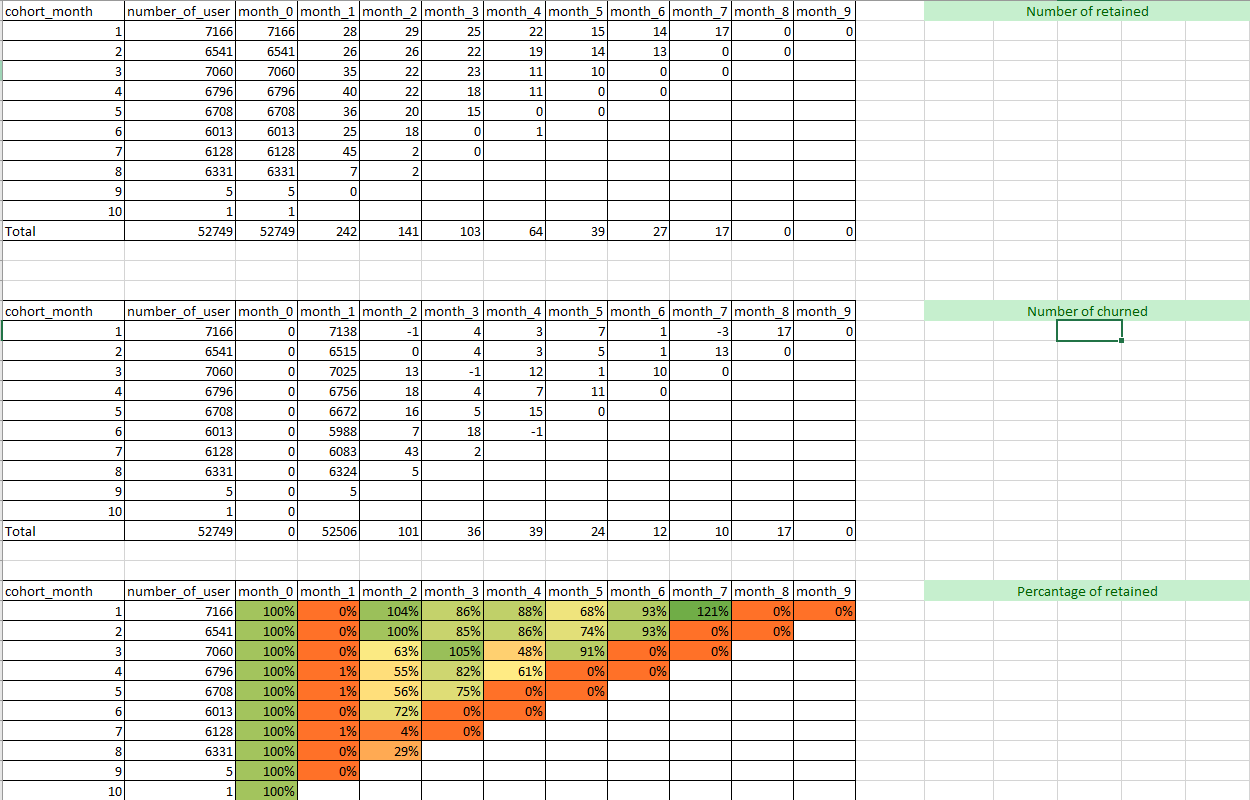
Ở phần này, tôi sẽ phân tích cohort như tôi đã đề cập ở phần trước. Sau khi xử lý bảng cohort bằng SQLthì tôi sẽ paste chúng vào excel để dễ nhìn hơn

Dưới đây là bảng cohort năm 2017 của siêu thị, bảng thứ nhất đại diện cho số khách hàng ở lại qua các tháng, bảng thứ hai đại diện cho số khách hàng rời đi qua các tháng và bảng thứ ba đại diện cho phần trăm người ở lại qua các tháng. Nhìn vào bảng retain ta có thể thấy số người dùng của siêu thị giảm mạnh sau tháng đầu tiên, tỉ lệ gần như lên tới 100%, từ tháng thứ 2 trở đi là cũng có xu hướng tương tự và ta cũng có thể thấy là có những tháng sẽ có khách hàng quay trở lại, thường là sau 2 tháng 1 nhưng mà số lượng không đáng kể

Tuy tỉ lệ giữ chân cực thấp theo bằng cohort, siêu thị vẫn có một lượng doanh thu ổn định, và như ta đã phân tích RFM ở trên, cửa hàng vẫn có một lượng khách champion khá lớn. Nguyên nhân của điều này có thể là do mặt hàng chuyên dụng của siêu thị là nội thất, đây là mặt hàng không được mua thường xuyên hoặc hàng tháng hàng ngày, trái ngược với các mặt hàng như thực phẩm.

Vì đây là những mặt hàng có tuổi thọ cao, nên siêu thị có thể cân nhắc mở rộng sản phẩm và dịch vụ như là trang trí nhà cửa, dịch vụ bảo trì hoặc nâng cấp nội thất.

****

Dưới đây là dữ liệu cohort năm 2018, ta có thể thấy là tỉ lệ retain có vẻ đã tăng đôi chút nhưng mà không đáng kể, có vẻ siêu thị vẫn mắc phải vấn đề ở năm 2017

**Đề xuất giải pháp cho siêu thị**

Phân phối lại dòng sản phẩm:

* Cân nhắc bỏ bớt các sản phẩm mang lại ít doanh thu cũng như doanh số của siêu thị như dòng sản phẩm baby, auto, cool\_stuff
* Tập trung vào những dòng sản phẩm tiềm năng như health\_beauty, computer\_accessory, watches\_gif

Mở rộng dịch vụ và sản phẩm bổ sung:

* Có thể nhập thêm sản phẩm liên quan đến sửa chữa nội thất, hoặc nhập thêm một số loại nội thất theo mùa nhằm giữ chân khách
* Kết hợp dịch vụ bảo trì và nâng cấp nội thất
* Kết hợp dịch vụ tư vấn thiết kế nội thất

Tạo chương trình khuyến mãi với chu kỳ dài:

* Tạo chương trình khách hàng thân thiết với phần thưởng dài hạn. Chẳng hạn, tích điểm không chỉ dựa trên số lần hay giá trị mua hàng mà còn dựa trên tương tác khác như đánh giá sản phẩm, chia sẻ trên mạng xã hội, tham gia sự kiện
* Tạo ưu đãi dài hạn cho khách hàng trung thành như là giảm giá hoặc quà tặng sau mỗi chu kỳ vài năm hoặc khi họ có nhu cầu nâng cấp, bảo trì nội thất

Một số yếu tố khác có thể cân nhắc:

* Cung và cầu của thị trường: Cân nhắc thị hiếu và nhu cầu của người dùng, xem xét liệu phong cách nội thất còn phù hợp không
* Đối thủ cạnh tranh: Nghiên cứu các sản phẩm và chiến lược của đối thủ cạnh tranh
* Yếu tố vận chuyển: Vì đồ nội thất khá cồng kềnh, siêu thị có thể xem xét các yếu tố liên quan đến vận chuyển như là chi phí vận chuyển, thời gian giao hàng, độ thỏa dụng của khách hàng,.....