**ĐẠI HỌC QUỐC GIA TP.HCM**

**TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT**

---🙦🟍🙤---

****

**BÁO CÁO NHÓM GIỮA KỲ**

**COHORT ANALYSIS & RFM**

**Môn học:** Toán ứng dụng trong kinh tế, quản trị và tài chính 2

**Giảng viên:** Thầy Nguyễn Phúc Sơn

**Thành viên nhóm 1:**

Nguyễn Trần Ngọc Bích K194131648

Phạm Tiến Đạt K194131651

Phạm Lam Hiếu K194131661

Võ Ngọc Thảo Nguyên K194131684

Trần Ngọc Phấn K194131690

Nguyễn Thị Quỳnh Như K204131842

*Tp.HCM, tháng 12 năm 2022*

Mục lục

**[1. Cohort Analysis và RFM 2](#_Toc20121)**

[1.1 Cohort Analysis 2](#_Toc27391)

[1.2 RFM 2](#_Toc25768)

**[2. Dữ liệu 4](#_Toc18767)**

[2.1 Cohort 4](#_Toc11236)

[2.2 RFM 4](#_Toc572)

**[3. Data Pre-Processing 5](#_Toc13790)**

[3.1 Cohort 5](#_Toc9258)

[3.2 RFM 5](#_Toc30237)

**[4. Phân tích Cohort 5](#_Toc16497)**

**[5. Mô hình hóa dữ liệu bằng mô hình RFM 8](#_Toc7386)**

**[6. Phân khúc khách hàng thông qua mô hình RFM theo hai chỉ số Recency và Monetary 10](#_Toc8608)**

**[7. Phân khúc khách hàng thông qua mô hình RFM theo từng chi nhánh cửa hàng 15](#_Toc8383)**

# **Cohort Analysis và RFM**

## **1.1 Cohort Analysis**

* *Định nghĩa:*

Cohort Analysis (phân tích tổ hợp) là một kỹ thuật phân tích trong Marketing tập trung vào việc phân tích hành vi của một hay một nhóm khách hàng có chung một đặc điểm trong khoảng thời gian nhất định, từ đó đo lường được mức độ tương tác của khách hàng theo thời gian, giúp chúng ta hiểu biết sâu sắc hơn về những trải nghiệm của khách hàng và cải thiện chúng một cách tốt nhất.

* *Phân loại:*

Có 2 loại Cohort Analysis là Acquisition cohorts (Nhóm tổ hợp chuyển đổi) và Behavioral cohorts (Nhóm tổ hợp về hành vi)

* Acquisition cohorts: Nhóm Cohort này phân chia khách hàng dựa trên thời điểm mà họ đăng ký hay sử dụng sản phẩm, có thể là theo dõi hàng ngày, hàng tuần hoặc hàng tháng, từ đó giúp chúng ta xác định được những khách hàng nào rời đi và thời điểm họ rời đi.
* Behavioral cohort: Nhóm này sẽ dựa trên những hoạt động của khách hàng trong một khoảng thời gian nhất định để đưa ra được lý do rời đi của khách hàng đối với sản phẩm họ đang sử dụng.

Như vậy ta thấy phân tích tổ hợp Cohort Analysis không chỉ giúp chúng ta biết có bao nhiêu khách hàng từ bỏ sản phẩm, từ bỏ khi nào mà còn giúp ta hiểu được lý do họ từ bỏ sản phẩm của mình để từ đó có thể khắc phục, cải thiện sản phẩm tốt hơn. Đây chính là cách để chúng ta có thể xác định được mức độ giữ chân khách hàng và các yếu tố quan trọng góp phần thúc đẩy mức độ tương tác của khách hàng và doanh thu đối với sản phẩm của mình.

## **1.2 RFM**

* *Định nghĩa:*

RFM (Recency - Frequency - Monetary Value) là một phần không thể thiếu của Marketing Analysis, được sử dụng để phân tích giá trị khách hàng (Customer Value), từ đó giúp cho doanh nghiệp biết rõ từng nhóm khách hàng của mình và đưa ra những chiến dịch marketing phù hợp.

* *Cấu trúc mô hình RFM:*

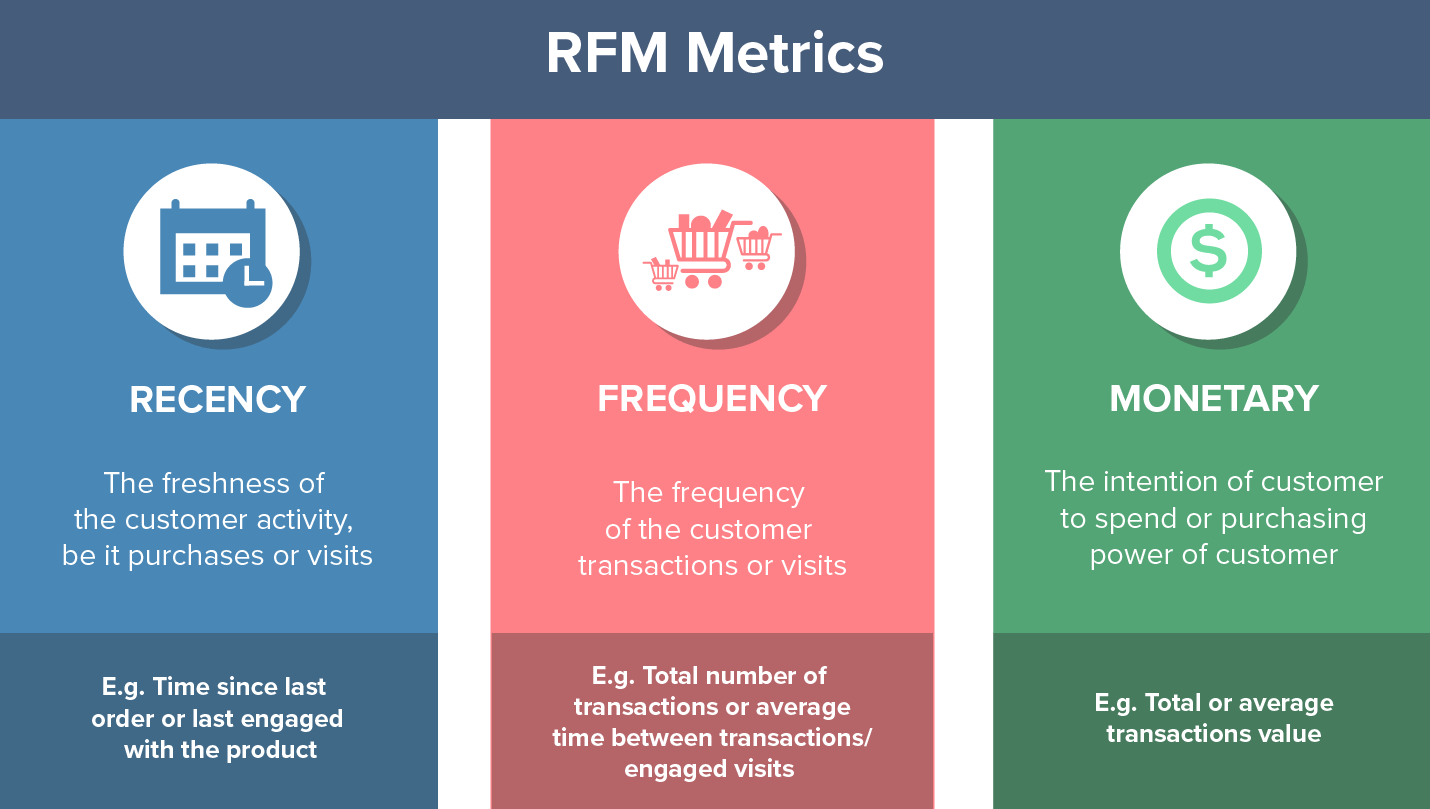
RFM được tạo nên từ ba yếu tố: R (Recency) - F (Frequency) - M (Monetary Value)

* R (Recency): Khoảng thời gian mà khách hàng mua hàng gần đây nhất hoặc về đơn đặt hàng cuối cùng của khách hàng khi nào. Nó có nghĩa là số ngày kể từ khi khách hàng thực hiện giao dịch mua cuối cùng.
* F (Frequency): Tần suất mua hàng của khách hàng, là về số lần mua trong một khoảng thời gian nhất định. Có thể là 3 tháng, 6 tháng hoặc 1 năm. Vì vậy, chúng ta có thể hiểu giá trị này là tần suất hoặc số lượng khách hàng sử dụng sản phẩm của công ty. Giá trị càng lớn thì khách hàng càng gắn bó.

Chỉ số này thường được quan tâm nhiều bởi các doanh nghiệp có dịch vụ, sản phẩm đem lại lợi nhuận thấp vì họ tồn tại được là nhờ khách hàng sử dụng dịch vụ, sản phẩm thường xuyên. Ví dụ như dịch vụ Grab, máy bán nước tự động.

* M (Monetary Value): Tổng số tiền mà khách hàng đã chi tiêu trong khoảng thời gian nhất định đó.

Đây là chỉ số dùng để tính toán giá trị vật chất mà doanh nghiệp có được mỗi khi khách hàng sử dụng sản phẩm/ dịch vụ. Chỉ số này được quan tâm nhiều trong các loại hình kinh doanh dịch vụ mà khách hàng không sử dụng thường xuyên nhưng mỗi lần sử dụng sẽ chi một khoản tiền tương đối cao như du lịch, bất động sản,...



Mô hình RFM

# **Dữ liệu**

## **2.1 Cohort**

Bộ dữ liệu để chạy phân tích Cohort là bộ dữ liệu bán lẻ của một công ty bao gồm các thông tin về ['Số phiếu', 'Mã KH', 'Mã hàng', 'Tên hàng', 'Nhóm hàng', 'Ngày', 'Lý do', 'Số lượng', 'Đơn giá', 'Trả lại', 'Hạng thẻ', 'Chi nhánh']. Gồm có 811897 quan sát với 13 thuộc tính.

## **2.2 RFM**

Bộ dữ liệu được dùng là dữ liệu lịch sử bán lẻ của các chi nhánh bán lẻ chính của một công ty. Bộ dữ liệu được quan sát từ tháng 1-2021 đến ngày 13-07-2022, có 889071 quan sát với 39 thuộc tính như ['Mã hàng', 'Tên hàng', 'Ngày', 'Số phiếu', 'Lý do', 'Số lượng', 'Đơn giá', 'Đơn giá sau thuế', 'Thuế suất', 'ĐVT', 'Chiết khấu($)', 'Thành tiền', 'Thành tiền trước thuế cộng chiết khẩu', 'Mã NV', 'Kho', 'Khách hàng', 'Mã KH', 'Trả lại', 'Thuế', 'Chi nhánh', 'Hình thức TT', 'Mã NCC', 'Khu vực', 'Hạng thẻ', 'CK KHTT(%)', 'CK(%)', 'Nhóm C1', 'Nhóm C2', 'Nhóm C3', 'Xuất xứ', 'Doanh thu', 'Số phiếu mới', 'Brand', 'TT Bán >0 va HBTL <> 0', 'Mã KM', 'Nguồn đơn hàng', 'TT trước thuế', 'User', 'NCC'].

# **Data Pre-Processing**

## **3.1 Cohort**

* Kiểm tra và làm sạch dữ liệu khuyết bằng cách loại bỏ những quan sát bị khuyết.
* Kiểm tra và làm sạch dữ liệu bị lặp lại (duplicates) bằng cách loại bỏ những quan sát bị lặp lại.
* Thống kê mô tả bộ dữ liệu, và xem xét kết quả thống kê có phù hợp với điều kiện của thuộc tính hay không. Nếu không thì xử lý những giá trị đó. Như loại bỏ những giá trị nhỏ hơn 0 tại cột [‘Số lượng’]. Loại bỏ những giá trị là khách lẻ ‘KL’ tại cột [‘Mã KH’], bởi vì những khách lẻ được xem là outlier trong bộ dữ liệu và những khách lẻ này họ không có đăng ký thông tin trên hệ thống.
* Định dạng thuộc tính của dữ liệu. Định dạng cột [‘Ngày’], [‘Ngày được tạo’] thành dữ liệu datetime.

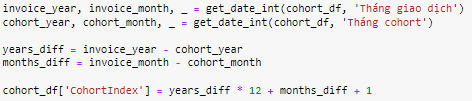
## **3.2 RFM**

* Loại bỏ những cột không cần thiết trong bộ dữ liệu.
* Lọc và lấy dữ liệu chính. Loại bỏ những thông tin quan sát về các tặng phẩm, khuyến mãi.
* Định dạng thuộc tính của dữ liệu. Định dạng cột [‘Ngày’] thành dữ liệu datetime và chỉ lấy dữ liệu ngày (bỏ dữ liệu thời gian).

# **Phân tích Cohort**

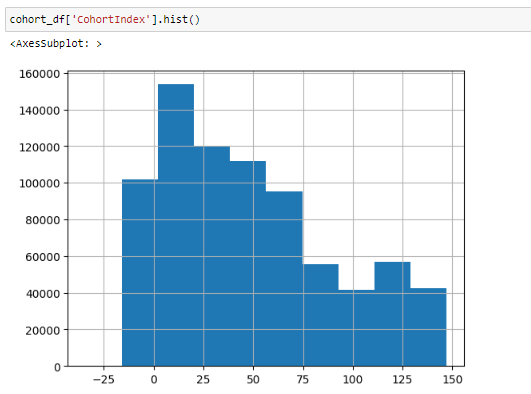
***Tạo ra 3 cột mới để chuẩn hóa dữ liệu:***

* [‘Tháng giao dịch’]: một chuỗi đại diện cho năm và tháng của một giao dịch/hóa đơn. Và đưa về cùng 1 ngày là ngày đầu tiên của tháng thực hiện giao dịch/ hóa đơn đó. Ví dụ: hóa đơn được thực hiện vào ngày 09-12-2011, thì chuẩn hóa về là ngày 01-12-2011.
* [‘Tháng cohort’]): một chuỗi đại diện cho năm và tháng mua hàng đầu tiên của khách hàng (ngày được tạo). Nhãn này phổ biến trên tất cả các hóa đơn cho một khách hàng cụ thể. Và đưa về cùng 1 ngày là ngày đầu tiên của tháng thực hiện giao dịch/ hóa đơn đầu tiên đó. Ví dụ: đơn hàng đầu tiên của khách hàng có ID là 123456 được thực hiện vào ngày 09-10-2011, thì chuẩn hóa về là ngày 01-10-2011.
* [‘CohortIndex’]: một số nguyên thể hiện giai đoạn của một khách hàng trong “vòng đời” của nó. Con số đại diện cho số tháng trôi qua kể từ lần mua đầu tiên. Được tính bằng cách lấy năm tại cột [‘Tháng giao dịch’] trừ cho năm tại cột [‘Tháng cohort’] tất cả nhân với 12 để ra được số tháng và cộng với tháng tại cột [‘Tháng giao dịch’] trừ cho tháng tại cột [‘Tháng cohort’] sau đó cộng với 1.



***Tạo bảng Pivot của Cohort bằng cách đếm số lượng khách hàng từ mỗi Cohort***

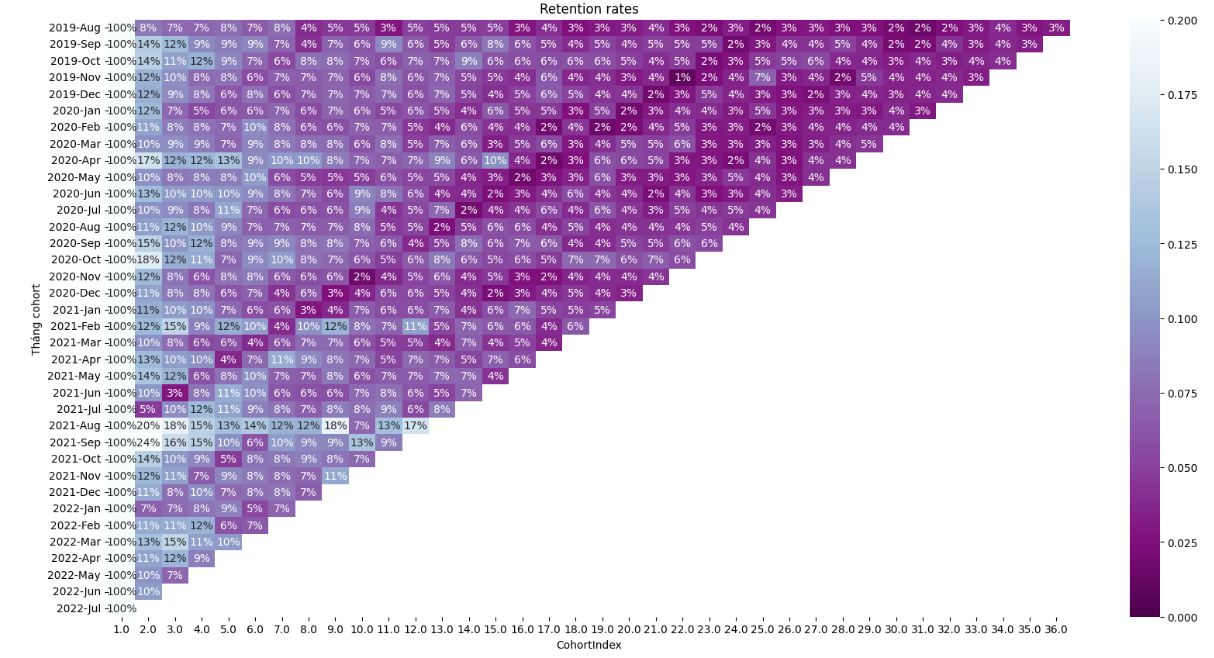
* Vì chỉ xem xét trong 36 thời đoạn nên lấy [‘CohortIndex’] từ (0;37)



* Nhóm cột [‘Tháng cohort’] và [‘CohortIndex’] lại. Tiếp theo nhóm dữ liệu vừa tạo đó theo cột [‘Mã KH’] để tạo bảng lượng khách hàng trong từng thời đoạn

***Tính tỷ lệ giữ chân khách hàng (Retention Rate)***

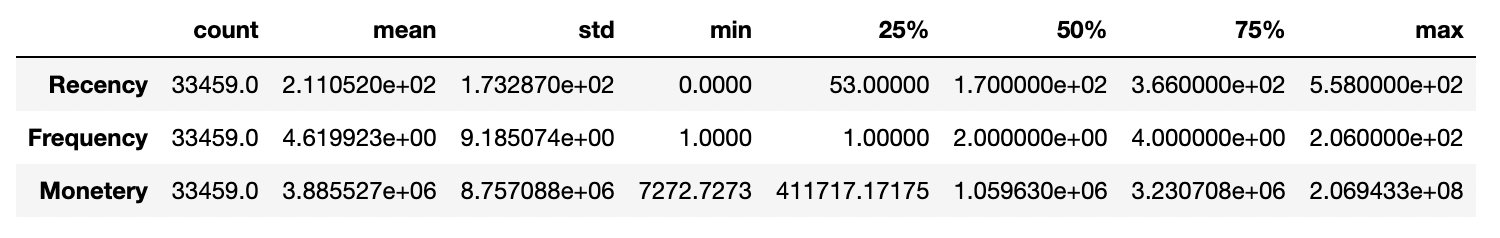
* Tỷ lệ giữ chân khách hàng là một số liệu rất hữu ích để hiểu có bao nhiêu trong số tất cả khách hàng vẫn đang hoạt động. Tỷ lệ giữ chân khách hàng cho biết tỷ lệ phần trăm khách hàng đang hoạt động so với tổng số khách hàng.
* Tỷ lệ giữ chân khách hàng được tính bằng cách lấy (Số lượng khách hàng trong thời đoạn / số lượng khách hàng đầu kỳ)\*100. Ví dụ: Nhóm khách hàng đăng ký mới tại Ngày 01-12-2010, trong tháng đầu tiên (thời đoạn thứ nhất) có 885 khách hàng. Trong tháng thứ hai (thời đoạn thứ hai) của nhóm này có lượng khách giao dịch là 324. Tỷ lệ khách hàng ở lại trong tháng thứ 2 kể từ khi mua hàng lần đầu vào ngày 01-12-2010 là (324/885)\*100 = 36.6%.



# **Mô hình hóa dữ liệu bằng mô hình RFM**

***Tính các chỉ số RFM:***

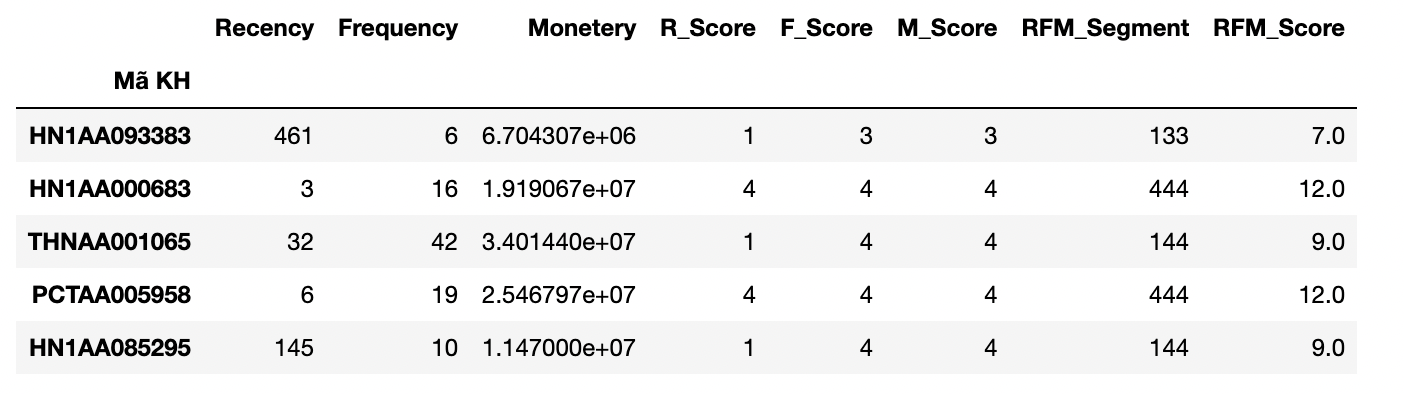
* Recency: được tính bằng cách nhóm cột [‘Mã KH’] và lấy ngày cuối cùng thực hiện giao dịch của khách hàng trừ cho ngày xuất dữ liệu làm mốc là ngày 13 tháng 7.
* Frequency: nhóm cột [‘Mã KH’] theo từng số phiếu mà họ đã giao dịch, sau đó đếm mã đơn hàng và tạo cột giá trị tần suất khách hàng [‘Frequency’] và gộp vào bộ dữ liệu chung
* Monetary Value: được tính bằng cách nhóm [‘Mã KH’] theo doanh thu và tính tổng các giá trị đơn hàng của từng khách hàng.



***Xếp hạng khách hàng chung bằng mô hình RFM***

Các chỉ số sẽ được tính điểm từ 1 đến 4 (dùng tứ phân vị).

* Xếp hạng “Recency”:
* Nhóm khách hàng có Recency lớn hơn 30 ngày là 1 điểm
* Nhóm khách hàng có Recency lớn hơn 14 và nhỏ hơn hoặc bằng 30 là 2 điểm
* Nhóm khách hàng có Recency lớn hơn 7 và nhỏ hơn hoặc bằng 14 là 3 điểm
* Nhóm khách hàng có Recency lớn hơn hoặc bằng 7 là 4 điểm
* Xếp hạng “Frequency”:
* Nhóm khách hàng có Frequency nhỏ hơn hoặc bằng 2 là 1 điểm
* Nhóm khách hàng có Frequency lớn hơn 2 và nhỏ hơn hoặc bằng 3 là 2 điểm
* Nhóm khách hàng có Frequency lớn hơn 3 và nhỏ hơn hoặc bằng 8 là 3 điểm
* Nhóm khách hàng có Frequency lớn hơn 8 là 4 điểm
* Xếp hạng “Monetary”: chia điểm theo tứ phân vị, 25% phân vị đầu cho 1 điểm, 25 - 50% cho 2 điểm, 50 - 75% cho 3 điểm và 75 - 100% cho 4 điểm



Trong đó:

* Cột [‘RFM\_Segment’]: được thực hiện bằng cách kết hợp các điểm xếp hạng của “Recency”, “Frequency” và “Monetary Value” lại với nhau tạo thành các nhóm phân khúc.
* Cột [‘RFM\_Score’]: được thực hiện bằng cách tính tổng các điểm xếp hạng của “Recency”, “Frequency” và “Monetary Value” lại với nhau.

# **Phân khúc khách hàng thông qua mô hình RFM theo hai chỉ số Recency và Monetary**

*Lưu ý hình trên phân khúc theo 5 bậc hạng trong dự án này chỉ phân khúc theo 4 bậc hạng nên sẽ có chút khác biệt*

* Recency từ 1-2 + Frequency từ 1-2 = hiberating

Ý nghĩa nhóm khách hàng có tần suất giao dịch được 2 lần cộng với việc từ 15 đến hơn 90 ngày không giao dịch đã không giao dịch thì được xếp vào nhóm kế thừa

* Recency từ 1-2 + Frequency từ 2-3 = at\_risk

Ý nghĩa nhóm khách hàng có tần suất giao dịch được 3 đến 8 lần cộng với việc từ 15 đến hơn 90 ngày đã không giao dịch thì được xếp vào nhóm rủi ro.

* Recency từ 1-2 + Frequency 4 = Can't lose

Ý nghĩa nhóm khách hàng có tần suất giao dịch từ 9-206 lần công với việc đã không giao dịch được khoảng từ 15 đến hơn 90 ngày thì xếp vào nhóm không thể đánh mất.

* Recency 2 + Frequency từ 1-2 = about to sleep

Ý nghĩa nhóm khách hàng có tần suất giao dịch từ 2-3 lần với việc đã từ 15-30 ngày không giao dịch thì được xếp vào nhóm sắp ngủ.

* Recency 2 + Frequency 2 = need attention

Ý nghĩa nhóm khách hàng có tần suất giao dịch được 3 lần với việc đã từ 15-30 ngày không giao dịch thì được xếp vào nhóm cần được quan tâm.

* Recency từ 2 đến 3 + Frequency từ 3 đến 4 = loyal customers

Ý nghĩa nhóm khách hàng có tần suất giao dịch từ 3 đến 8 lần với việc đã từ 8-30 ngày không giao dịch thì được xếp vào diện khách hàng trung thành.

* Recency 3 + Frequency 1 = Promising

Ý nghĩa nhóm khách hàng có tần suất giao dịch từ 4 đến 8 lần với việc đã từ 30 đến hơn 90 ngày không giao dịch thì được xếp vào diện khách hàng đầy hứa hẹn.

* Recency 4 + Frequency 1 = new customers

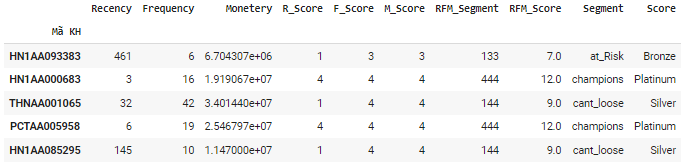
Ý nghĩa nhóm khách hàng có tần suất giao dịch được 2 lần với việc đã từ 0 đến 7 ngày không giao dịch thì được xếp vào diện khách hàng mới.

* Recency từ 3 đến 4 + Frequency từ 1 đến 2 = potential loyalists

Ý nghĩa nhóm khách hàng có tần suất giao dịch từ 2 đến 3 lần với việc từ 0 đến 14 ngày không giao dịch thì được xếp vào diện khách hàng tiềm năng trung thành.

* Recency 4 + Frequency từ 3 đến 4 = champions

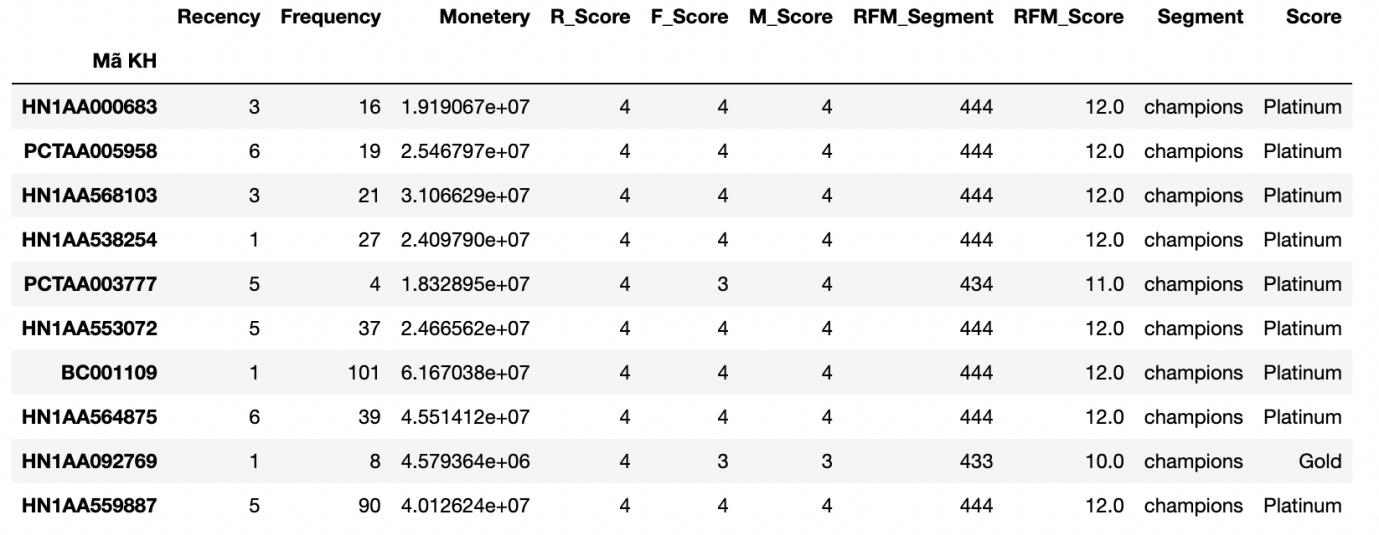
Ý nghĩa nhóm khách hàng có tần suất giao dịch từ 4 đến 206 lần cộng với việc từ 0 đến 7 ngày không giao dịch thì được xếp vào nhóm khách hàng vô địch.



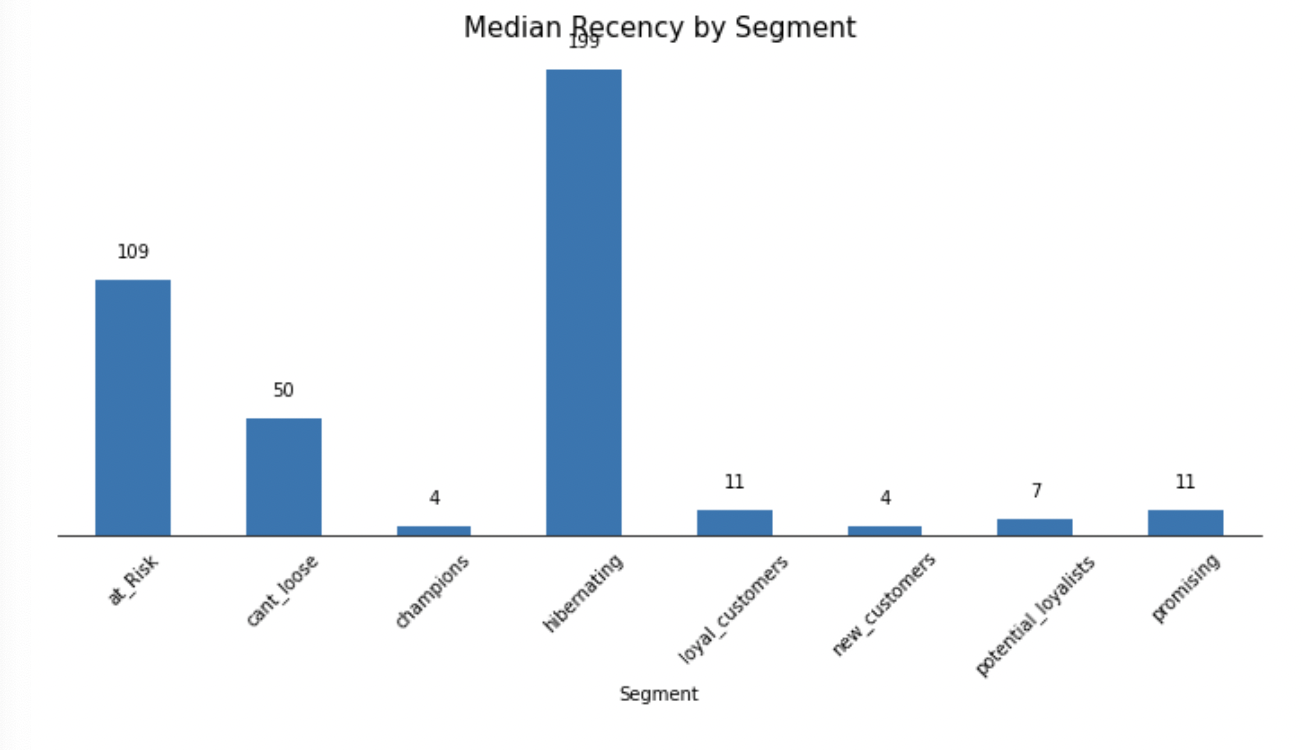
Phần điểm RFM được chia thành các mức xếp hạng:

* Platinum: Tổng điểm 3 tiêu chí > 10
* Gold : Tổng điểm 3 tiêu chí trên 9 và <= 10
* Silver : Tổng điểm 3 tiêu chí trên 7 và <= 9
* Bronze: Tổng điểm 3 tiêu trên 5 và <= 7
* Green : Tổng điểm 3 tiêu chí <= 5

Nhóm 10 khách hàng thuộc top đầu là [“champions”] (most valuable) như sau:

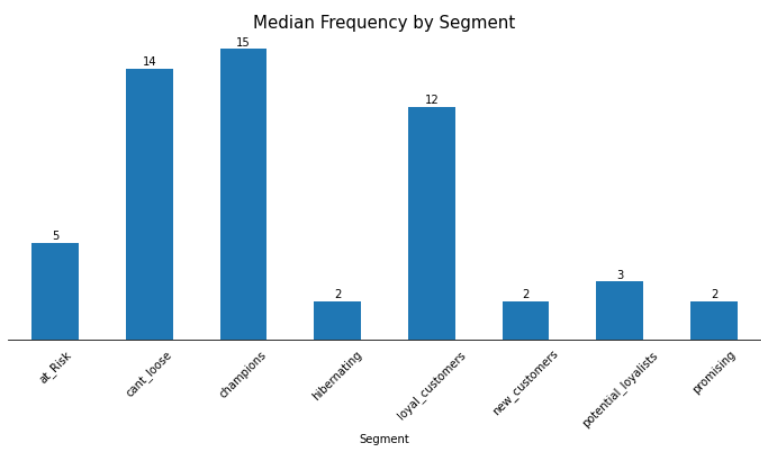


* Biểu đồ cột thể hiện giá trị trung vị Recency của từng nhóm khách hàng như sau:



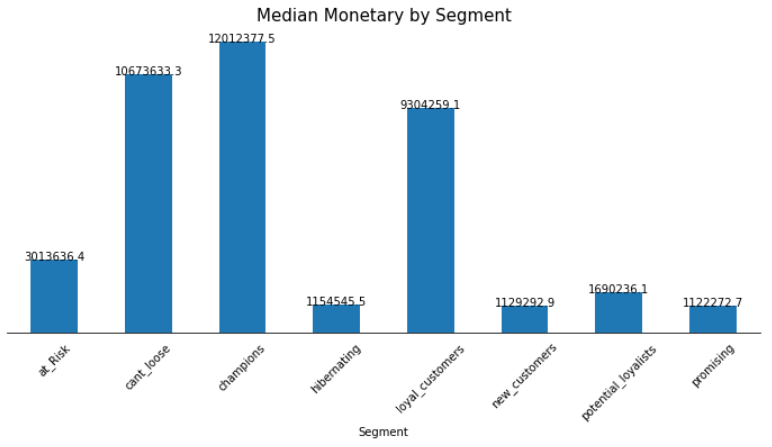
Từ đồ thị ta thấy 50% khách hàng thuộc nhóm *champions* có thời điểm gần nhất giao dịch với cửa hàng là từ 4 ngày đổ về (từ thời điểm 13/7/2022) và ở 50% nhóm khách hàng thuộc *hibernating(kế thừa)* là 199 ngày.

* Biểu đồ cột thể hiện giá trị trung vị Frequency của từng nhóm khách hàng như sau:



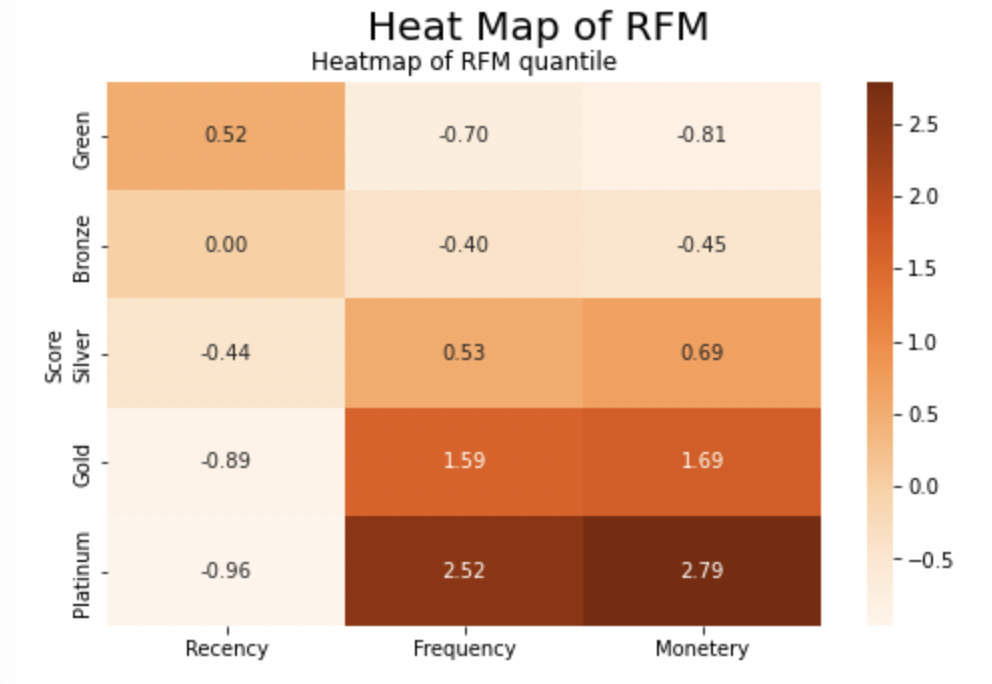
Từ đồ thị ta thấy khoảng 50% khách hàng thuộc nhóm *champions* có tần suất giao dịch với cửa hàng lớn nhất 15 lần, còn ở 50% nhóm khách hàng thuộc *hibernating (kế thừa)* là 2 lần

* Biểu đồ cột thể hiện giá trị trung vị Monetary của từng nhóm khách hàng như sau:



Từ đồ thị ta thấy 50% khách hàng thuộc nhóm *champions* có đơn hàng trị giá lớn nhất (khoảng 12 triệu đồng) và 50% nhóm khách hàng thuộc *hibernating(kế thừa)* có giá trị đơn hàng thấp nhất (khoảng 1 triệu 154 nghìn đồng).

* Mức độ tương quan mạnh yếu của từng nhóm khách hàng tương ứng thông qua ba chỉ số của RFM:



Ta có thể thấy các nhóm khách hàng thuộc nhóm *Platinum* có ô giá trị Recency rất thấp (-0.96, màu rất trắng) còn hai ô giá trị Frequency và Monetary (2.51 và 2.79, màu rất tối), điều này thể hiện rằng nhóm khách hàng có tần suất đang giao dịch rất thường xuyên và giá trị đơn hàng của họ rất lớn.

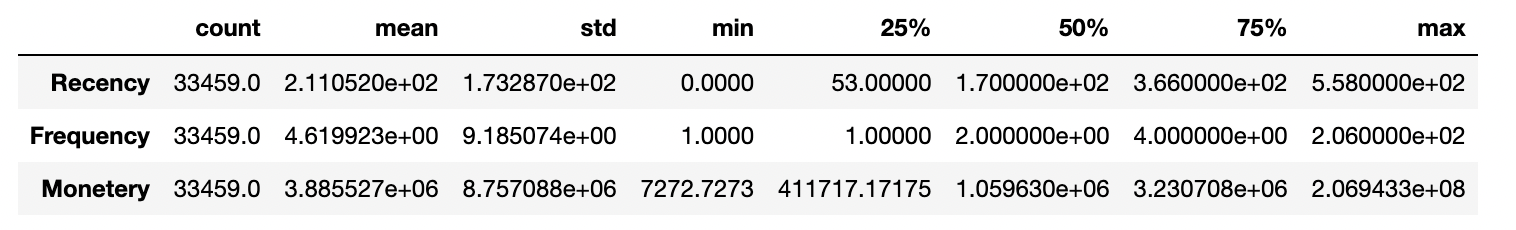
Ngược lại khi quan sát nhóm khách *Green*, tỷ lệ Recency mạnh tương đối ở cực dương (0.52) điều này cho thấy nhóm này đã tương đối lâu không quay lại các cửa hàng bán lẻ của SNB. Giá trị đơn hàng của hàng và tần suất giao dịch cũng thấp hơn so với nhóm giá trị đơn hàng và tần suất trung bình 0.2 lần.

Nhìn chung xu hướng chung của biểu đồ nhiệt cho thấy rằng các khách hàng có độ cam kết với các cửa hàng bán lẻ thì họ sẽ càng có tần suất giao dịch nhiều và chi tiền mạnh cho các đơn hàng kèm theo việc thời gian giao dịch gần đây sẽ thấp.

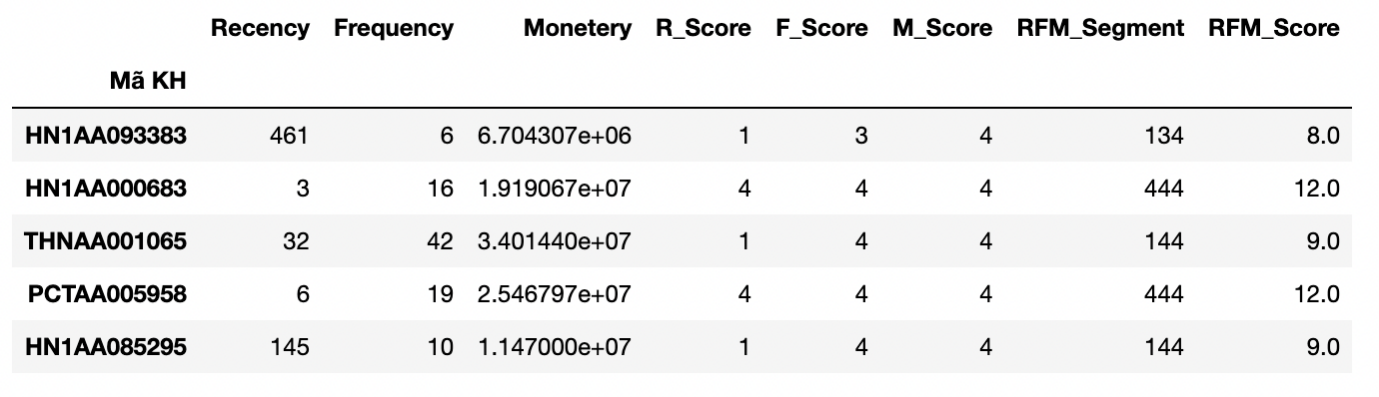
# **Phân khúc khách hàng thông qua mô hình RFM theo từng chi nhánh cửa hàng**

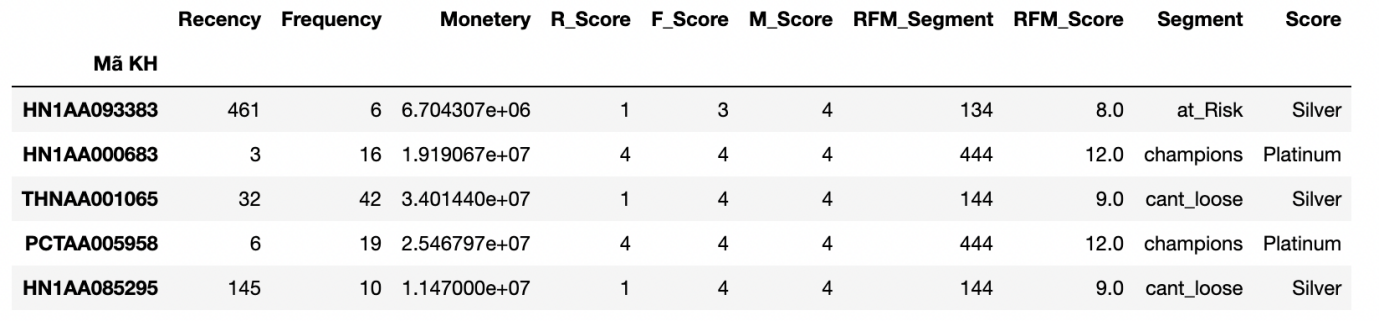
* Xây dựng mô hình RFM phân tích theo từng chi nhánh:

Ở bước này, nhóm nghiên cứu không lọc đi nhóm khách hàng lẻ có định danh (Frequency = 1)

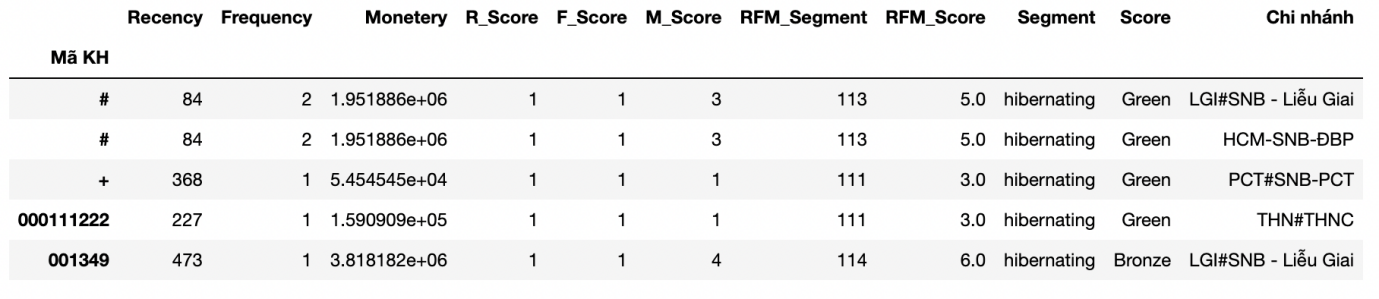


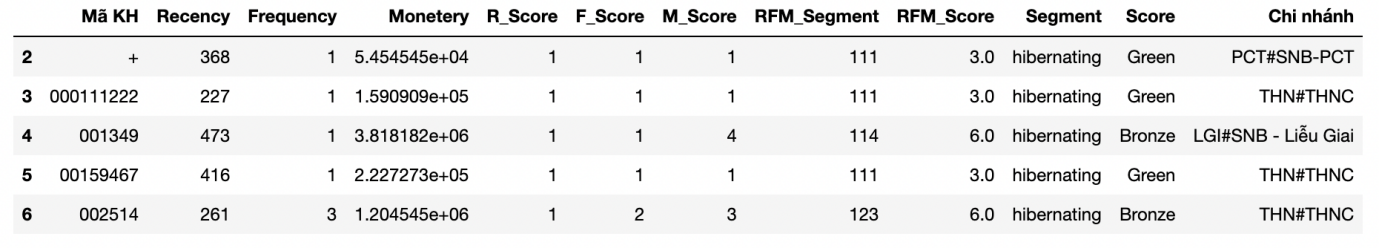
* Sau đó, tiến hành các bước tương tự như phân tích theo hai chỉ số Recency và Monetary mục 6, ta được các kết quả tương ứng:



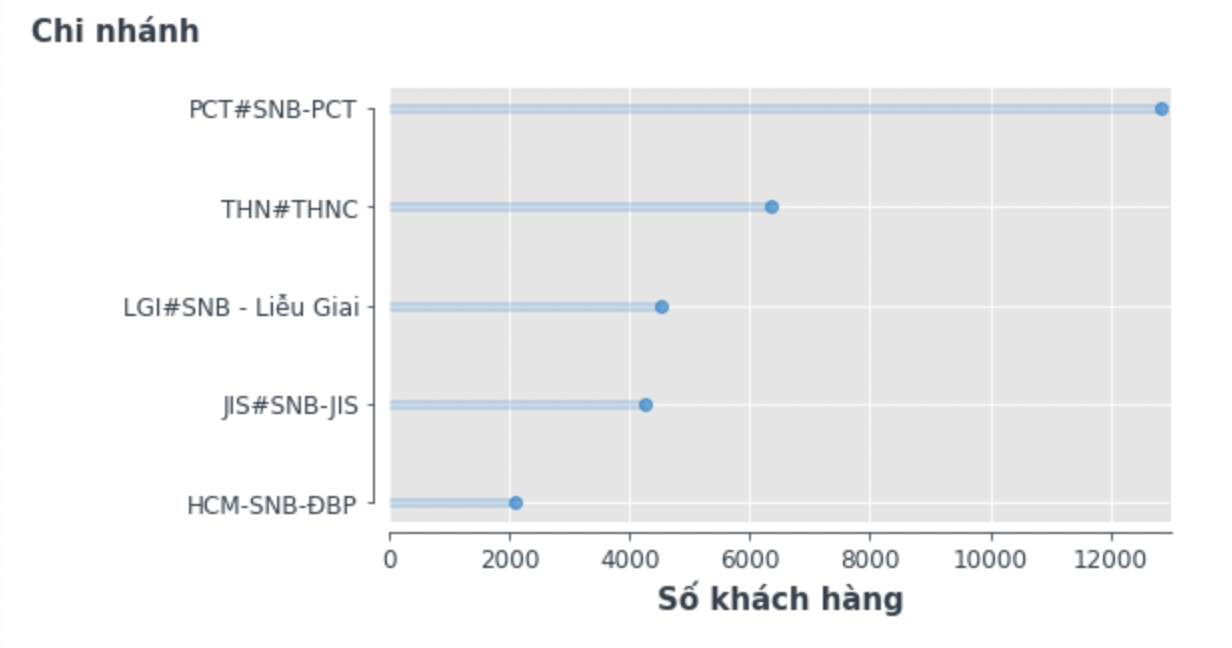


* Tiếp theo, đếm số quan sát chứa khách hàng từng giao dịch với chi nhánh nào và lọc ra nhóm khách hàng có giao dịch tại hai chi nhánh trở lên.





* Biểu đồ thể hiện số lượng khách hàng đến giao dịch từng chi nhánh trong suốt năm vừa rồi như sau:

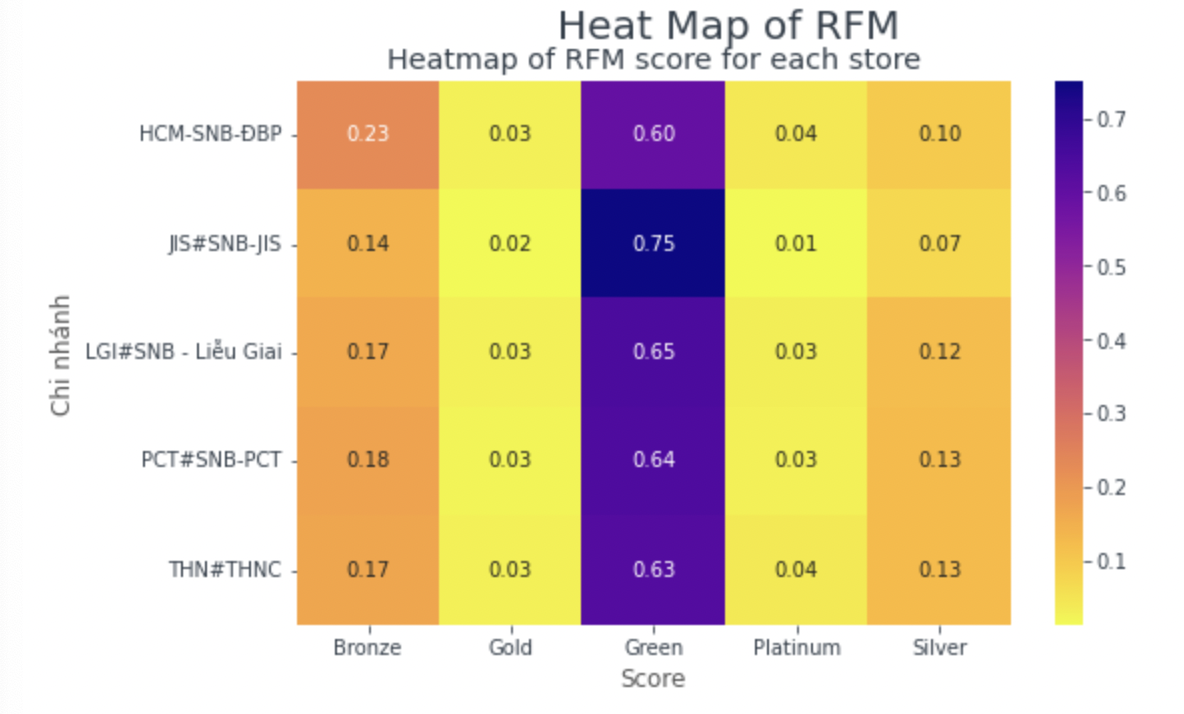


Ta nhận thấy chi nhánh có nhiều khách hàng nhất đó là chi nhánh Phan Chu Trinh Hà Nội với hơn 12850 khách hàng. Số lượng khách hàng của chi nhánh này gần như chiếm đến 43% tổng toàn bộ khách hàng của các hàng bán lẻ.

Hầu hết khách hàng của các cửa hàng đều là khách vãng lai (chiếm 75,4%), đây là nhóm khách hàng chỉ có đến cửa hàng trong thời điểm rất lâu trước đây và tần suất giao dịch của họ không vượt quá ba lần.

Ngoài ra 11,2% nhóm khách hàng được xem vào dạng có rủi ro (at\_risk), nhóm khách hàng đã có đến cửa hàng trong thời điểm rất lâu trước đây và tần suất giao dịch của họ từ 2 đến 8 lần với cửa hàng.

Như vậy, điều này cho thấy cách vận hành của của các cửa hàng bán lẻ phần lớn có sự thu hút đặc biệt với khách lẻ, chưa thể thu hút khách hàng để tạo sự commitment giữa hai bên. Phân loại của mô hình RFM cho thấy toàn bộ cửa hàng có tỷ lệ khách hàng trung thành không vượt quá 3%.



Đây là bảng pivot của khách hàng theo từng chi nhánh tùy theo mức xếp hạng điểm RFM.