**TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT**

**KHOA HỆ THỐNG THÔNG TIN**

**A blue circle with red and white text

Description automatically generated\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**BÁO CÁO GIỮA KỲ**

**PHƯƠNG PHÁP NGHIÊN CỨU LIÊN NGÀNH**

**CHỦ ĐỀ: ỨNG DỤNG PHÂN TÍCH PHÂN KHÚC KHÁCH HÀNG BẰNG RFM TRONG LẬP TRÌNH**

**Giảng viên hướng dẫn:**

***TS. Nguyễn Thôn Dã***

**Nhóm SIUUU**

**1. Nguyễn Đặng Hoài Nam**

**2. Lê Thanh Quý Hải**

**3. Lương Chí Trung**

**4. Cao Phan Trung Hiếu**

**5. Trần Đức Lương**

**6. Nguyễn Văn Tài**

Thành phố Hồ Chí Minh, ngày 22 tháng 10 năm 2023

|  |  |  |
| --- | --- | --- |
| **STT** | **HỌ VÀ TÊN** | **MSSV** |
| **1** | Nguyễn Đặng Hoài Nam | K224060798 |
| **2** | Lê Thanh Quý Hải | K224060780 |
| **3** | Lương Chí Trung | K224060818 |
| **4** | Cao Phan Trung Hiếu | K224060784 |
| **5** | Trần Đức Lương | K224060793 |
| **6** | Nguyễn Văn Tài | K224060809 |

**THÀNH VIÊN NHÓM**

**MỤC LỤC**

[**Mô hình RFM (Recency - Frequency - Monetary model)** 4](#_Toc148900359)

[**Khai báo thư viện** 5](#_Toc148900360)

[**Xử lý file .csv** 5](#_Toc148900361)

[**Trực quan hóa:** 10](#_Toc148900362)

[*1. Histogram* 10](#_Toc148900363)

[*2. Boxplot* 12](#_Toc148900364)

[*3. Pairplot* 13](#_Toc148900365)

[**RFM** 15](#_Toc148900366)

[1. **Phân khúc khách hàng với RFM Score** 15](#_Toc148900367)

[**2.** **Chuẩn hóa dữ liệu** 24](#_Toc148900368)

[**3.** **Chọn cụm tối ưu với Elbow** 26](#_Toc148900369)

[**4.** **Đánh giá** 29](#_Toc148900370)

[**5.** **Biểu đồ 3D** 30](#_Toc148900371)

# **Mô hình RFM (Recency - Frequency - Monetary model)**

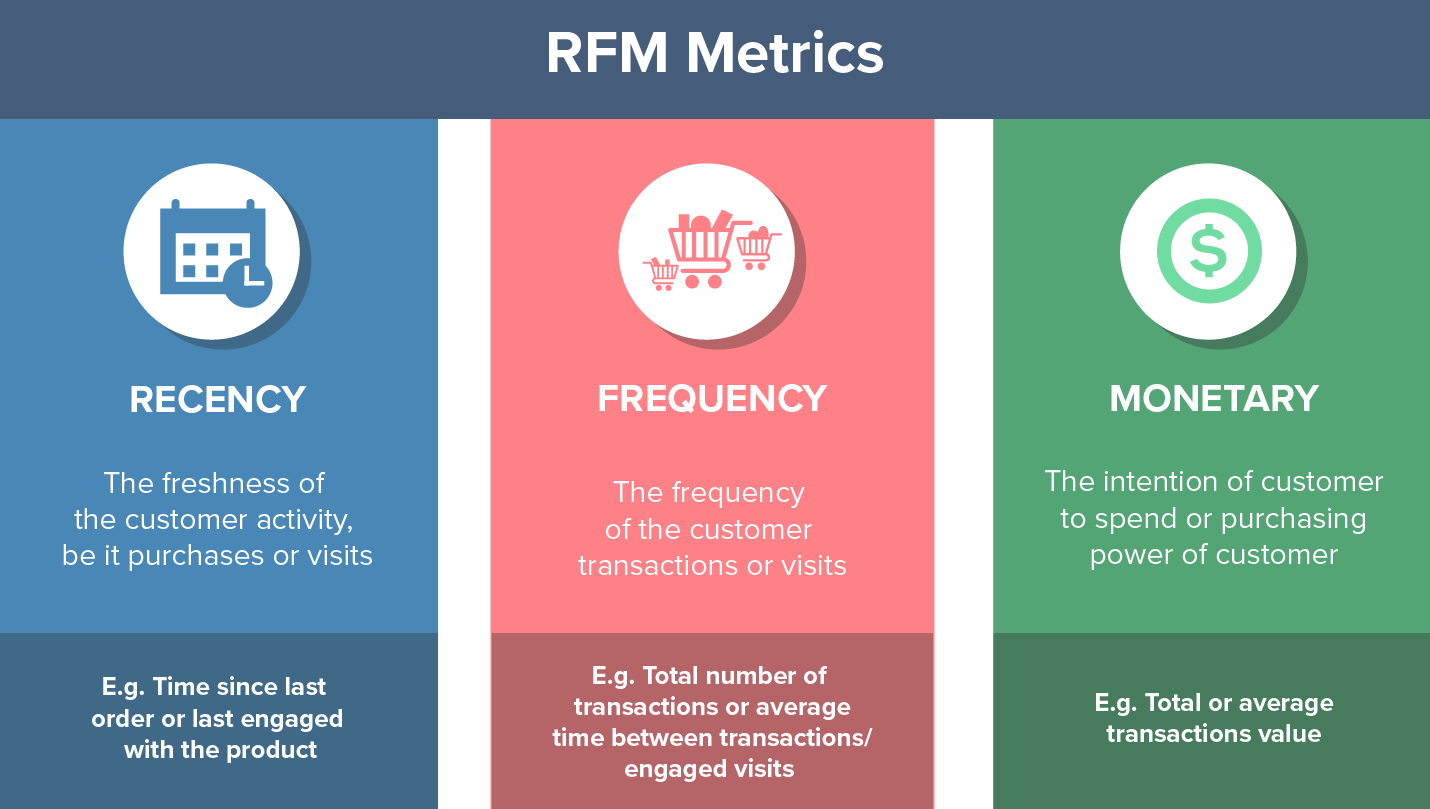
Recency - Frequency - Monetary là một mô hình được sử dụng nhiều trong các nghiên cứu ở trong và ngoài nước trong phân khúc khách hàng. Nó thường được sử dụng trong marketing cơ sở dữ liệu (kiểu như dựa vào dữ liệu về khách hàng để tiếp thị sản phẩm) và marketing trực tiếp và đã nhận được sự chú ý đặc biệt trong ngành bán lẻ và dịch vụ.

RFM định lượng giá trị của một khách hàng dựa trên 3 yếu tố chính:

-        **Recency:** Khoảng thời gian mua hàng gần đây nhất là bao lâu. Cho biết khách hàng có đang thực sự hoạt động gần thời điểm đánh giá. Chỉ số này càng lớn càng cho thấy xu hướng rời bỏ của khách hàng càng cao. Đó là một cảnh báo cho doanh nghiệp nên thay đổi sản phẩm để đáp ứng thị hiếu khách hàng hoặc thay đổi chính sách để nâng cao chất lượng phục vụ.

-        **Frequency:** Tần suất mua hàng của khách hàng. Nếu khách hàng mua càng nhiều đơn thì giá trị về doanh số mang lại cho công ty càng cao và tất nhiên giá trị của họ càng lớn. Tuy nhiên nếu chỉ xét dựa trên tần suất mua hàng thì cũng chưa đánh giá được đầy đủ mức độ tác động lên doanh thu bởi bên cạnh đó, giá trị đơn hàng cũng là yếu tố trực tiếp cho thấy khách hàng tiềm năng như thế nào.

-        **Monetary:** Là số tiền chi tiêu của khách hàng. Đây là yếu tố trực quan nhất ảnh hưởng tới doanh số. Hay nói cách khác, doanh nghiệp quan tâm nhất là khách hàng đã dành bao nhiêu tiền để mua sắm sản phẩm của công ty? Monetary sẽ tác động trực tiếp tới doanh thu và bị tác động gián tiếp thông qua 2 yếu tố còn lại là Recency và Frequency.

Với mỗi thuộc tính, khách hàng được chấm điểm theo thang điểm từ 1 đến 5 (số càng cao, kết quả càng tốt). Khách hàng "tốt nhất" sẽ nhận được điểm số cao nhất trong mọi danh mục. Do đó, khách hàng lý tưởng sẽ có số điểm là 5, 5, 5. Những khách hàng khác có điểm số thấp hơn có thể được xác định để cải thiện.

# **Khai báo thư viện**

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ

Mô tả được tạo tự động

- Khai báo thư viện Pandas để xử lý file dữ liệu .csv

- Khai báo thư viện matplotlib, seaborn, squarify để trực quan hóa dữ liệu lên các biểu đồ

# **Xử lý file .csv**

A close-up of a computer code

Description automatically generated

- Định dạng cột \_CustomerID về dạng chuỗi để xử lý

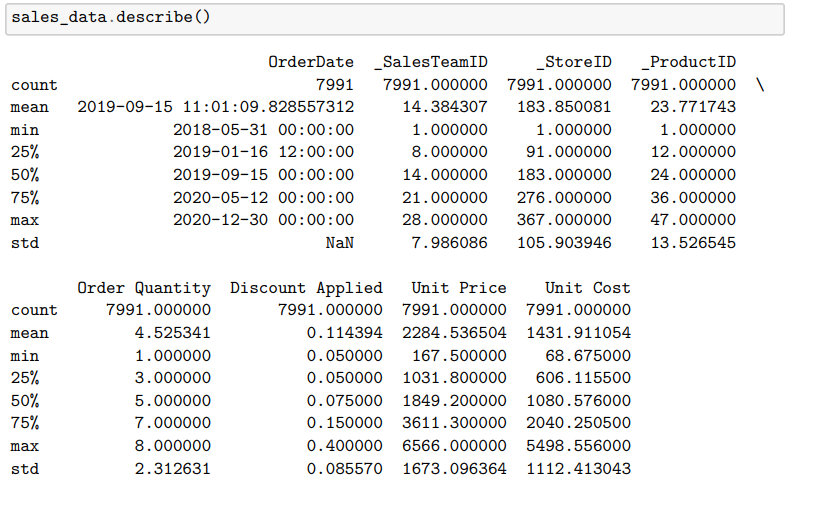
- Đọc file ‘sales\_data.csv’ về thành một DataFrame có tên là sales\_data.

- dtype=data\_type: Đây là một tham số tùy chọn trong phương thức read\_csv. Nó được sử dụng để chỉ định kiểu dữ liệu của các cột trong DataFrame. data\_type là một biến hoặc một dict mô tả kiểu dữ liệu cho từng cột.

- parse\_dates=['OrderDate']: Đây là một tham số tùy chọn khác trong phương thức read\_csv. Nó được sử dụng để chỉ định cột hoặc danh sách các cột chứa dữ liệu ngày tháng trong định dạng chuỗi. Trong trường hợp này, cột 'OrderDate' trong tệp CSV được chuyển đổi thành kiểu dữ liệu ngày tháng trong DataFrame.



- Mô tả file dữ liệu ‘sales\_data.csv’ như: trung bình, trung vị, số dòng, độ lệch chuẩn, q1, q2, q3, … của các cột trong file ‘sales\_data.csv’.



Tiếp theo:

A close up of a computer code

Description automatically generated

- Tạo một DataFrame mới có tên df từ DataFrame ‘sales\_data’

- Tạo một cột mới có tên ‘Revenue’ và các dữ liệu từ cột sẽ là kết quả từ công thức từ các cột của DataFrame df:

(‘Unit Price’ - (‘Unit Price’ \* ‘Discount Applied’) - ‘Unit Cost’) \* ‘Order Quantity’

A screenshot of a computer code

Description automatically generated

- Tạo các cột mới (Series) cho DataFrame ‘df’: ‘OrderNumber’, ‘\_CustomerID’, ‘OrderDate’, ‘Revenue’

- Sử dụng các cột đấy cho DataFrame ‘df\_dataset’

- Đặt ngày mua hàng ở thời điểm gần nhất là ‘2021 - 01 - 01’

- Tạo một DataFrame mới ‘rfm\_dataset’ bằng cách nhóm các cột của ‘df\_dataset’

- rfm\_dataset: Đây là tên của đối tượng DataFrame mới sẽ được tạo ra từ phương thức groupby và agg.

- df\_dataset: Đây là tên của đối tượng DataFrame ban đầu chứa dữ liệu bán hàng.

groupby('\_CustomerID'): Phương thức groupby được gọi trên đối tượng DataFrame df\_dataset để nhóm dữ liệu theo cột '\_CustomerID'. Điều này tạo ra các nhóm dữ liệu dựa trên giá trị của cột '\_CustomerID'.

- agg({...}): Phương thức agg được gọi trên đối tượng GroupBy để áp dụng các hàm tổng hợp lên từng nhóm dữ liệu. Trong trường hợp này, một dict được truyền vào phương thức agg để chỉ định các hàm tổng hợp và cột mục tiêu tương ứng.

- 'OrderDate': lambda v: (today\_date - v.max()).days: Đây là một khối mã lambda được sử dụng làm hàm tổng hợp cho cột 'OrderDate'. Nó tính toán số ngày từ ngày cuối cùng trong nhóm tới ngày hiện tại (gọi là Recency - thời gian gần nhất). today\_date là biến đại diện cho ngày hiện tại, và v.max() trả về ngày lớn nhất trong nhóm.

- 'OrderNumber': 'count': Đây là một hàm tổng hợp được chỉ định bằng chuỗi 'count'. Nó đếm số lượng mục trong cột 'OrderNumber' trong mỗi nhóm, đại diện cho tần suất mua hàng (Frequency).

- 'Revenue': 'sum': Đây là một hàm tổng hợp được chỉ định bằng chuỗi 'sum'. Nó tính tổng giá trị của cột 'Revenue' trong mỗi nhóm, đại diện cho giá trị tiền đã chi tiêu (Monetary).

rename(columns={...}): Phương thức rename được gọi trên đối tượng DataFrame rfm\_dataset để đổi tên các cột.

- columns={...}: Đây là một tham số trong phương thức rename và một dict được truyền vào để chỉ định các cặp tên cột cũ và tên cột mới.

- 'OrderDate': 'Recency': Đây là một cặp key-value trong dict, trong đó 'OrderDate' là tên cột cũ và 'Recency' là tên cột mới.

- 'OrderNumber': 'Frequency': Đây là một cặp key-value khác trong dict, trong đó 'OrderNumber' là tên cột cũ và 'Frequency' là tên cột mới.

- 'Revenue': 'Monetary': Đây là một cặp key-value khác trong dict, trong đó 'Revenue' là tên cột cũ và 'Monetary' là tên cột mới.

- inplace=True: Đây là một tham số tùy chọn trong phương thức rename. Khi inplace=True, đối tượng DataFrame rfm\_dataset sẽ được thay đổi trực tiếp (thay vì trả về một bản sao đã được đổi tên).

# **Trực quan hóa:**

## *1. Histogram*

A graph of a function

Description automatically generated with medium confidence

*Biểu đồ phân phối của các yếu tố Recency, Frequency, Monetary*

Biểu đồ dạng "phân phối" (distribution plot) cho ba chỉ số RFM (Recency, Frequency, Monetary) từ tập dữ liệu rfm\_dataset. Biểu đồ này sẽ hiển thị phân phối của các chỉ số này trên cùng một hình (figure) có 3 biểu đồ, mỗi biểu đồ trong một hàng khác nhau. 

Biểu đồ cho chỉ số Recency:

* Dựa vào dữ liệu, giá trị trung bình (mean) của Recency là 14.38.
* Độ lệch chuẩn (standard deviation) là 7.99.
* Phân phối của Recency có thể được thể hiện bằng một đường cong. Nó cho biết mức độ tập trung của Recency trong dữ liệu và mức độ biến động xung quanh giá trị trung bình.

Biểu đồ cho chỉ số Frequency:

* Giá trị trung bình (mean) của Frequency là 23.77.
* Độ lệch chuẩn (standard deviation) là 13.53.
* Biểu đồ Frequency cũng có thể là một đường cong, thể hiện phân phối của số lần giao dịch của khách hàng. Nó cho biết mức độ tập trung và phân tán của số lần giao dịch trong dữ liệu.

Biểu đồ cho chỉ số Monetary:

* Giá trị trung bình (mean) của Monetary là 4,525.34.
* Độ lệch chuẩn (standard deviation) là 2,312.63.
* Biểu đồ Monetary sẽ hiển thị phân phối của doanh thu (tổng giá trị của các giao dịch) của khách hàng. Nó cung cấp cái nhìn về tập trung và độ biến động của doanh thu trong dữ liệu.

## *2A diagram of a diagram Description automatically generated with medium confidence. Boxplot*

Dựa vào biểu đồ boxplot trên ta nhận thấy được:

* Biểu đồ Recency: khách hàng mua hàng chủ yếu ở khoảng thời gian gần đây, chỉ có vài giá trị ở xa, cho thấy doanh nghiệp thu hút được đông đảo
* Biểu đồ Frequency: Tần suất mua hàng của khách hàng trung bình là kha khá cho thấy được công ty có chính sách đãi ngộ và thu hút khách hàng khá tốt,  nhưng cần tiếp tục phát huy thêm thế mạnh của công ty để tăng thêm tần suất mua hàng
* Biểu đồ Monetary:  Lượng tiền trung bình khách hàng chi cho sản phẩm của doanh nghiệp  là khá cao, tuy nhiên vẫn còn nằm ở mức kha khá, cần nâng cao thêm chất lượng sản phẩm, dịch vụ,...

Tóm lại qua biểu đồ boxplot ta có thể hiểu biết được mức phân phối dữ liệu, độ biến động, các giá trị ngoại biên, và nhận thấy được xu hướng của nhóm dữ liệu, từ đó đưa ra được các insight quan trọng.

## *3. Pairplot*

A graph of a number of blue and black bars

Description automatically generated with medium confidence

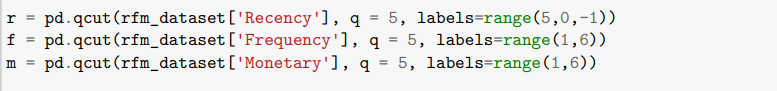
*Biểu đồ Pairplot với 3 yếu tố R, F, M*

* Biểu đồ Pairplot trên là một lưới của các biểu đồ phân tán (scatterplots), trong đó mỗi cặp cột được vẽ một biểu đồ phân tán riêng, giúp thấy được mối quan hệ giữa nhiều biến trong tập dữ liệu mà cụ thể ở đây là 'Recency', 'Frequency', và 'Monetary’.
* Tổng cộng 9 biểu đồ trong biểu đồ Pairplot sẽ được hiển thị. Có một biểu đồ cho mỗi cặp cột dữ liệu trong rfm. Mỗi cặp cột dữ liệu sẽ tạo ra một biểu đồ phân tán (scatter plot) và histogram cho cả hai biến. Do đó, chúng ta sẽ có các biểu đồ sau đây:
  + Recency vs. Recency (Histogram của Recency trên đường chéo chính).
  + Recency vs. Frequency (Biểu đồ phân tán của Recency và Frequency).
  + Recency vs. Monetary (Biểu đồ phân tán của Recency và Monetary).
  + Frequency vs.Recency (Biểu đồ phân tán của Frequency và Recency).
  + Frequency vs. Frequency (Histogram của Frequency trên đường chéo chính).
  + Frequency vs. Monetary (Biểu đồ phân tán của Frequency và Monetary).
  + Monetary vs. Recency (Biểu đồ phân tán của Monetary và Recency).
  + Monetary vs. Frequency (Biểu đồ phân tán của Monetary và Frequency).
  + Monetary vs. Monetary (Histogram của Monetary trên đường chéo chính).
* Biểu đồ Pairplot của 'Recency', 'Frequency', và 'Monetary’ giúp thực hiện một số công việc quan trọng:
  + *Phân tích Mối quan hệ:* Quan sát các scatterplots trong Pairplot để xem liệu có mối quan hệ nào giữa các biến 'Recency,' 'Frequency,' và 'Monetary.' Chúng ta có thể tìm hiểu xem liệu có mối tương quan giữa chúng. Như biểu đồ trên cho thấy rằng 'Frequency' và 'Monetary' có xu hướng tương quan dương (tăng Frequency thường đi kèm với tăng Monetary (mang tính tương đối)), điều này có thể mang ý nghĩa kinh doanh quan trọng.
  + *Xác định Biến quan trọng:* Pairplot có thể giúp quyết định xem biến nào quan trọng hơn trong việc đưa ra quyết định hoặc phân tích dữ liệu. Nếu một biến có mối tương quan mạnh với các biến khác, chúng ta có thể tập trung vào nó.
  + *Phát hiện Biến ngoại lai:* Xem Pairplot để tìm hiểu xem có sự xuất hiện của các giá trị ngoại lai trong dữ liệu của chúng ta. Các điểm dữ liệu cách xa khỏi phân phối chung có thể là biến ngoại lai.
  + *Trình bày Dữ liệu:* Pairplot cũng có thể được sử dụng để trình bày dữ liệu cho người khác một cách trực quan. Điều này có thể hữu ích khi chúng ta muốn chia sẻ thông tin về dữ liệu với đồng nghiệp hoặc cố vấn.
* Tóm lại, biểu đồ Pairplot là một công cụ mạnh mẽ trong việc phân tích và hiểu dữ liệu. Nó giúp thấy được mối quan hệ giữa các biến và xác định các đặc điểm quan trọng trong tập dữ liệu của chúng ta.

# **RFM**

* 1. Phân khúc khách hàng với RFM Score

Ở đây, chúng em chọn phương pháp chia khoảng ‘Quintiles’ để chia điểm cho từng giá trị R-F-M của mỗi khách hàng:



- pd.qcut: Đây là phương thức trong pandas được sử dụng để chia dữ liệu thành các nhóm dựa trên các khoảng cắt (quantiles). Nó chia dữ liệu thành các nhóm có số lượng mẫu gần như như nhau.

- rfm\_dataset['Recency']: Đây là cột 'Recency' trong đối tượng DataFrame rfm\_dataset. Chúng ta sẽ chia cột này thành các nhóm.

- q=5: Đây là tham số tùy chọn trong phương thức qcut, chỉ định số lượng nhóm mong muốn. Trong trường hợp này, chúng ta muốn chia thành 5 nhóm.

- labels=range(5,0,-1): Đây là tham số tùy chọn khác trong phương thức qcut, cho phép chỉ định nhãn cho các nhóm. Trong trường hợp này, chúng ta sử dụng các nhãn từ 5 đến 1 (theo thứ tự giảm dần) để đại diện cho các nhóm.

- Tương tự, các biến f và m cũng được tạo bằng cách chia cột Frequency và Monetary thành các nhóm tương ứng.

A screenshot of a computer code

Description automatically generated- Sau khi thực hiện các dòng mã trên, biến r, f và m sẽ chứa các nhãn tương ứng với các nhóm đã được chia dựa trên cột Recency, Frequency và Monetary trong đối tượng DataFrame rfm\_dataset.

A screenshot of a computer code

Description automatically generated

Hàm tự định nghĩa ‘segment’: được dùng để chia khoảng dựa trên các điểm số của từng thành phần ‘Recency’ - ‘Frequency’ - ‘Monetary’ của mỗi khách hàng mua hàng. Có 11 loại khách hàng và đặc điểm tương ứng:

A screenshot of a document

Description automatically generated

Cuối cùng:

A screenshot of a computer code

Description automatically generated

- rfm = rfm\_dataset.assign(R=r.values, F=f.values, M=m.values): Dòng mã này tạo một đối tượng DataFrame mới có tên là rfm bằng cách sử dụng phương thức assign trên đối tượng DataFrame rfm\_dataset. Ba cột mới được thêm vào DataFrame rfm là 'R', 'F' và 'M', với giá trị lần lượt lấy từ biến r, f và m.

- rfm['rfm\_group'] = rfm['R'].astype(str) + rfm['F'].astype(str) + rfm['M'].astype(str): Dòng mã này tạo một cột mới trong DataFrame rfm có tên là 'rfm\_group'. Nó được tạo bằng cách ghép các giá trị trong cột 'R', 'F' và 'M' thành một chuỗi. Các giá trị trong cột 'R', 'F' và 'M' được chuyển đổi sang kiểu dữ liệu chuỗi (str) trước khi ghép.

- rfm['Score Total'] = rfm[['R', 'F', 'M']].sum(axis=1): Dòng mã này tạo một cột mới trong DataFrame rfm có tên là 'Score Total'. Nó tính tổng các giá trị trong cột 'R', 'F' và 'M' cho mỗi hàng và gán kết quả vào cột 'Score Total'.

- rfm['Class'] = rfm['rfm\_group'].apply(segment): Dòng mã này tạo một cột mới trong DataFrame rfm có tên là 'Class'. Nó áp dụng một hàm segment lên cột 'rfm\_group' cho mỗi hàng để xác định nhóm RFM tương ứng cho từng hàng.

- Cuối cùng, dòng mã print(rfm) được sử dụng để in ra DataFrame rfm với các cột và giá trị đã được cập nhật.

🡪Và thành quả:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A list of customer satisfaction survey

Description automatically generated with medium confidence

Để có thể trực quan hóa bao nhiêu phân khúc, chúng em đếm từng phân khúc cho khách hàng dựa theo điểm số có bao nhiêu mỗi phân khúc:

A screenshot of a computer code

Description automatically generated

- customer\_segment = rfm.groupby('Class').agg({'Recency': 'mean', 'Frequency': 'mean', 'Monetary': ['mean', 'count']}): Dòng mã này sử dụng phương thức groupby để nhóm dữ liệu trong DataFrame rfm theo cột 'Class'. Sau đó, phương thức agg được sử dụng để tính toán giá trị trung bình của cột 'Recency', 'Frequency' và 'Monetary' trong mỗi nhóm. Đồng thời, số lượng khách hàng trong mỗi nhóm cũng được tính bằng cách sử dụng hàm đếm count.

- .round(1): Phương thức .round(1) được áp dụng sau khi tính toán giá trị trung bình để làm tròn đến một chữ số thập phân.

- customer\_segment.columns = ['Mean\_R', 'Mean\_F', 'Mean\_M', 'Count']: Dòng mã này đặt lại tên các cột trong DataFrame customer\_segment thành 'Mean\_R', 'Mean\_F', 'Mean\_M' và 'Count' để mô tả các giá trị đã tính toán.

- customer\_segment = customer\_segment.sort\_values(by='Count'): Đoạn mã này sắp xếp DataFrame customer\_segment theo cột 'Count' (số lượng khách hàng) theo thứ tự tăng dần.

- Kết quả cuối cùng là DataFrame customer\_segment đã được tính toán và sắp xếp theo số lượng khách hàng trong mỗi nhóm.

Kết quả:

A table of numbers and a list of people

Description automatically generated

Sau khi trực quan hóa:

A graph with colorful bars

Description automatically generated with medium confidence

Biểu đồ Countplot thường được sử dụng để đếm số lượng mẫu trong từng phân đoạn của một biến rời rạc (discrete variable) và hiển thị chúng dưới dạng các cột dọc hoặc ngang. Trong trường hợpnày, biểu đồ là một Countplot dọc (cột dọc) với dữ liệu trên trục y.

Mục đích của biểu đồ này là hiển thị số lượng khách hàng trong từng phân đoạn (segment) của biến "Class." Thông qua biểu đồ Countplot có thể dễ dàng so sánh sự phân bố của các phân đoạn và xác định sự khác biệt trong số lượng mẫu giữa chúng.

A diagram of a company

Description automatically generated with medium confidence

Tree Map là một loại biểu đồ thống kê dùng để hiển thị dữ liệu phân phối dưới dạng các hình chữ nhật có diện tích thể hiện giá trị của dữ liệu. Diện tích của mỗi hình chữ nhật trong tree map thể hiện số lượng khách hàng ở từng phân khúc RFM. Điều này cho phép biết được phân khúc nào có số lượng khách hàng nhiều hơn, và phân khúc nào có số lượng khách hàng ít hơn. Hình chữ nhật lớn đại diện cho một phân khúc có nhiều khách hàng, trong khi hình chữ nhật nhỏ đại diện cho một phân khúc có ít khách hàng.

A colorful pie chart with text

Description automatically generatedĐây là biểu đồ hình tròn (pie chart) hiển thị phân phối của các phân khúc khách hàng dưới dạng biểu đồ hình tròn. Biểu đồ này được sử dụng để thể hiện cấu trúc phần trăm các phần trong tổng thể.

## **Chuẩn hóa dữ liệu**

Nhóm em chọn biến đổi Box-cox để biến đổi các giá trị.

Ảnh có chứa văn bản, Phông chữ, hàng, màu trắng

Mô tả được tạo tự động

Có một sự khác biệt rất lớn giữa ba giá trị này, có thể ảnh hưởng đến thời gian chạy mô hình và độ chính xác của thuật toán. Nghiên cứu đã tiến hành tiêu chuẩn hóa dữ liệu theo phương pháp phân phối tiêu chuẩn (Standard Score), còn được gọi là Z-score để đưa dữ liệu vào phạm vi phân phối trong đó giá trị trung bình của các quan sát là 0 và độ lệch chuẩn là 1. Sau khi chuẩn hóa dữ liệu, có thể kết luận rằng ba giá trị hiện tại, tần số và giá trị tiền tệ có trọng số bằng nhau khi được đưa vào phân tích trong mô hình phân cụm K-Means.

Dữ liệu sau chuẩn hóa như sau:

Ảnh có chứa văn bản, thực đơn, sách, ảnh chụp màn hình

Mô tả được tạo tự độngẢnh có chứa văn bản, thực đơn, sách, ảnh chụp màn hình

Mô tả được tạo tự động

## **Chọn cụm tối ưu với Elbow**

Ảnh có chứa văn bản, Phông chữ, hàng, màu trắng

Mô tả được tạo tự độngẢnh có chứa văn bản, hàng, ảnh chụp màn hình, Sơ đồ

Mô tả được tạo tự động  
Số lượng cụm được chia thành 3 cụm theo thuật toán K-means, mô tả các giá trị trung bình của Recency, Frequency, Monetary để đánh giá các đặc điểm của từng cụm.

Ảnh có chứa văn bản, biên lai, Phông chữ, ảnh chụp màn hình

Mô tả được tạo tự động

Ảnh có chứa ảnh chụp màn hình, Hình chữ nhật, biểu đồ, hàng

Mô tả được tạo tự độngẢnh có chứa ảnh chụp màn hình, Hình chữ nhật, biểu đồ, hàng

Mô tả được tạo tự độngẢnh có chứa ảnh chụp màn hình, văn bản, biểu đồ, Hình chữ nhật

Mô tả được tạo tự động

Cụm 0 có giá trị Monetary, Frequency và Recency cao nhất. Cụm 2 có giá trị Monetary, Frequency cao hơn cụm 1. Vì vậy, có thể kết luận rằng cụm 0 là khách hàng vàng, cụm 2 là khách hàng bạc và cụm 1 là khách hàng đồng.

Nhóm xác định 3 cụm khách hàng dựa trên từng đặc điểm:

Khách hàng Vàng:

* + Monetary (Số tiền): Cao. Khách hàng vàng có giá trị Monetary cao, đây là điểm mạnh đặc trưng của họ.
  + Frequency (Tần suất): Cao. Họ mua sắm thường xuyên, thể hiện sự trung thành và cam kết.
  + Recency (Thời gian mua gần nhất): Cao. Họ duy trì mối quan hệ mua sắm gần đây với công ty, đóng góp đáng kể cho doanh nghiệp.

Khách hàng Bạc:

* + Monetary (Số tiền): Cao, nhưng không bằng khách hàng vàng. Họ có khả năng chi tiêu đáng kể cho các sản phẩm hoặc dịch vụ.
  + Frequency (Tần suất): Khá cao. Họ mua sắm đều đặn, mặc dù không thường xuyên như khách hàng vàng.
  + Recency (Thời gian mua gần nhất): Rất thấp. Có thể dạo này họ bận nên chưa quay lại mua hàng.

Khách hàng Đồng:

* + Monetary (Số tiền): Thấp. Họ chi tiêu ít tiền cho sản phẩm hoặc dịch vụ.
  + Frequency (Tần suất): Thấp. Họ mua sắm ít khi, chỉ khi cần.
  + Recency (Thời gian mua gần nhất): Ở mức vừa, có thể họ đã không mua sắm trong một thời gian dài hoặc đã mua sắm gần đây nhưng không thường xuyên

Số lượng của mỗi cụm khách hàng sau khi thực hiện thuật toán KMeans với 50 khách hàng:

-        Khách hàng đồng: 21

-        Khách hàng bạc: 15

-        Khách hàng vàng: 14

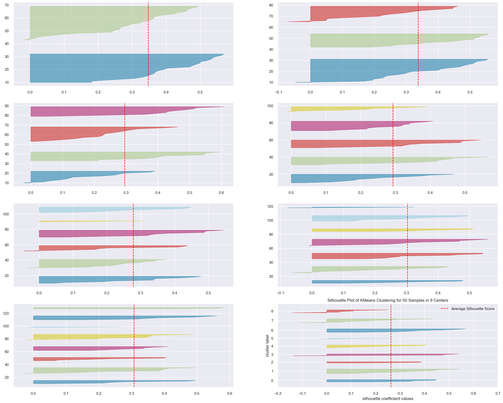
* 1. **Đánh giá**

Nhóm em sử dụng phân tích thuật toán Silhouette để đánh giá sự phù hợp của phân đoạn cụm.

A screen shot of a computer program

Description automatically generated

Code này tạo ra một hình ảnh chứa nhiều đồ thị silhouette, mỗi đồ thị cho một giá trị k (số lượng cụm) khác nhau. Điều này giúp đánh giá hiệu quả của thuật toán KMeans với các giá trị k khác nhau trong việc phân cụm dữ liệu.



## **Biểu đồ 3D**

Ở đây nhóm tụi em sử dụng Biểu đồ scatter 3D để hiển thị mô hình phân cụm của dữ liệu. Dữ liệu được biểu diễn trên ba trục: 'Recency' (thời gian gần nhất), 'Frequency' (tần suất) và 'Monetary' (giá trị tiền tệ).

A graph of a graph with red and black dots

Description automatically generated with medium confidence

Nhìn vào biểu đồ 3D trên có thể thấy được:

* Phân cụm dữ liệu: Mỗi điểm dữ liệu trên biểu đồ được đặt trong một màu sắc khác nhau dựa trên nhãn cụm của nó từ cột 'Cluster'
* Mối quan hệ giữa các biến: Biểu đồ scatter 3D cho phép chúng ta nhìn thấy mối quan hệ giữa ba biến 'Recency', 'Frequency' và 'Monetary'. Chúng ta có thể nhận thấy xu hướng và mức độ phân tán của các điểm dữ liệu trên ba trục này.
* Trung tâm cụm: Biểu đồ cũng hiển thị các điểm trung tâm cụm bằng các biểu tượng 'X' màu đỏ. Điểm trung tâm cụm đại diện cho một điểm trung bình của các mẫu trong cụm tương ứng.

Tổng quan, biểu đồ scatter 3D giúp chúng ta trực quan hóa mô hình phân cụm của dữ liệu và cung cấp cái nhìn toàn diện về mối quan hệ giữa các biến và sự phân tán của các mẫu trong dữ liệu.

--------------------HẾT--------------------