CUSTOMER SEGMENTATION REPORT

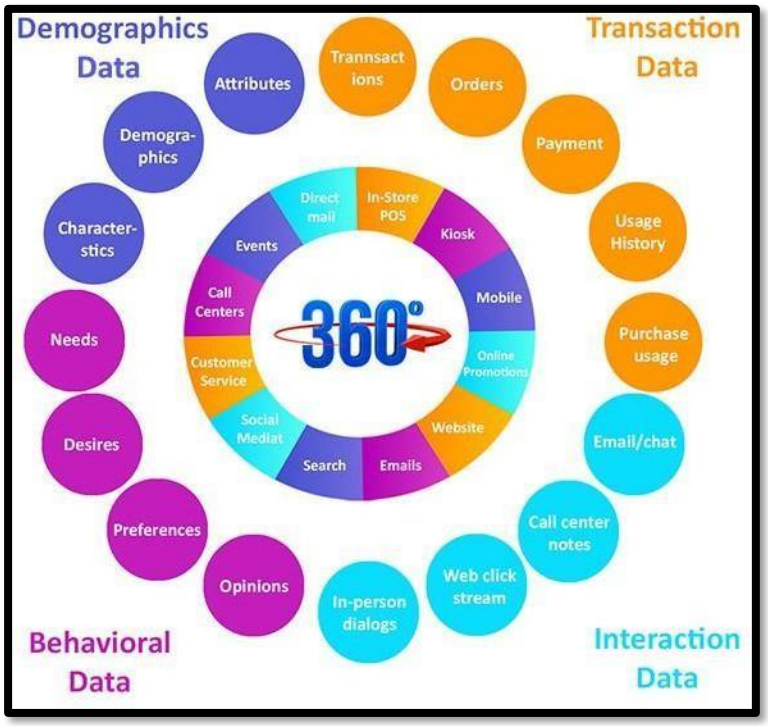
Group 1 – Data Customer

NGUYEN DUC MINH - 21020698
NGUYEN PHAN NAM SON - 21020704
PHAM MINH HIEU - 21021587

12/05/2024 - 28/05/2024

I. PHƯƠNG PHÁP

CUSTOMER360

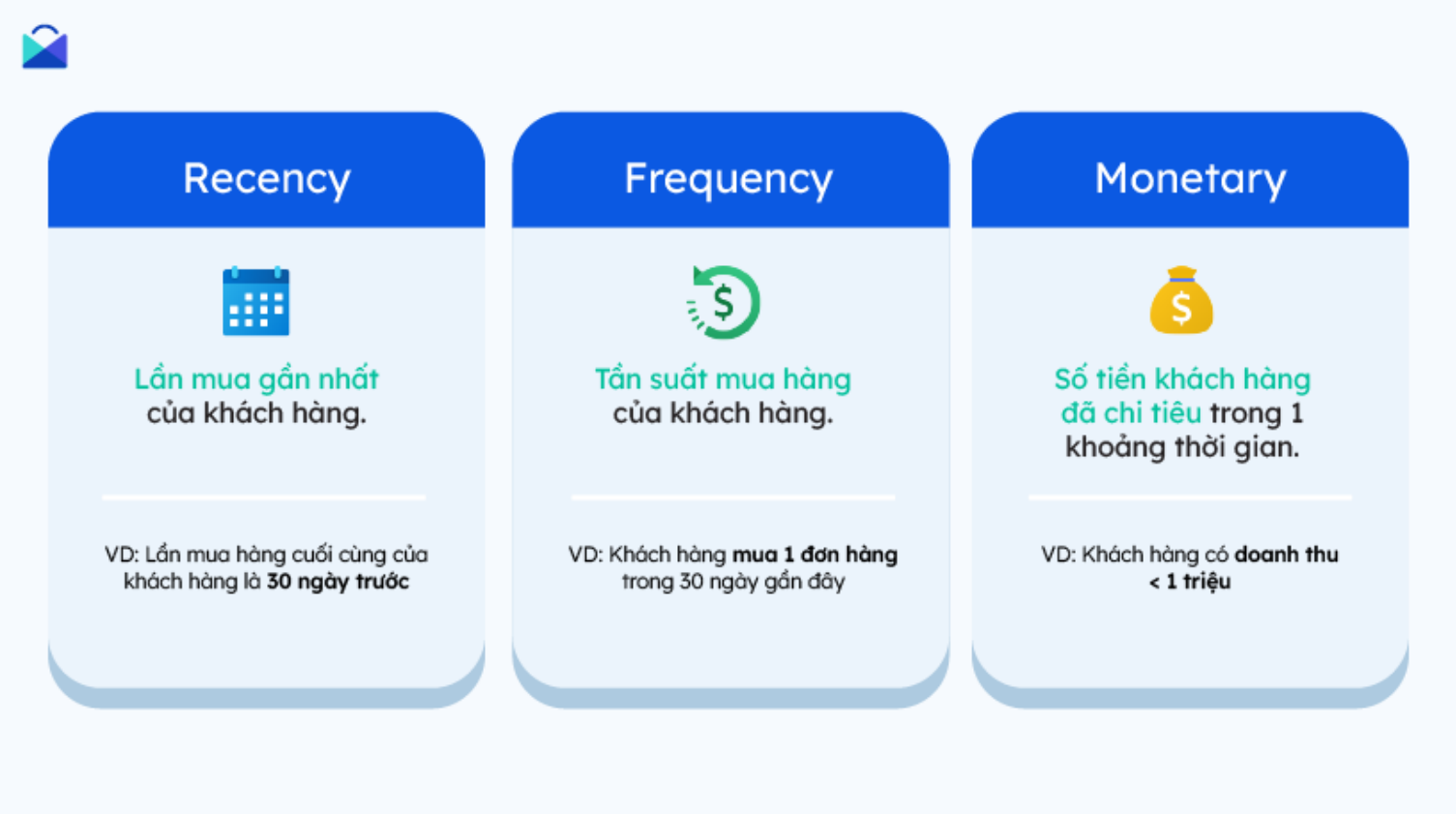


Customer360 là một chiến lược phân tích xoay quanh khách hàng dựa trên bộ dữ liệu về giao dịch (**Transaction Data**); dữ liệu tương tác với website hoặc ứng dụng (**Interaction Data**); hành vi tiêu dùng, nhu cầu của khách hàng (**Behavioral Data**) và dữ liệu về tuổi tác, nhân khẩu học của khách hàng (**Demographics Data**).

Phân khúc khách hàng dựa trên bộ dữ liệu về khía cạnh giao dịch (**Transaction Data**) của khách hàng trong 3 tháng **(01/06/2022 – 01/09/2022)**. Mô hình phân tích dữ liệu khách hàng được sử dụng là **RFM (Recency – Frequency – Monetary).**

MÔ HÌNH RFM

Mô hình RFM (Recency – Frequency – Monetary) được sử dụng trong marketing để phân phân loại tệp khách hàng. Mô hình phân tích dữ liệu khách hàng dựa trên 3 yếu tố:



Recency: Thời điểm sử dụng dịch vụ gần nhất cách ngày báo cáo.

⟹ Xác định, nhận diện khách hàng mới thông qua thời điểm sử dụng dịch vụ.

Frequency: Tần suất sử dụng dịch vụ của mỗi khách hàng trong khoảng thời gian báo cáo.

⟹ Xác định, nhận diện khách hàng thường xuyên sử dụng dịch vụ.

Monetary: Tổng doanh thu có được từ mỗi khách hàng trong khoảng thời gian báo cáo.

⟹ Xác định, nhận diện khách hàng tiềm năng với mức chi tiêu cao.

II. TÍNH TOÁN VÀ PHÂN TÍCH

DỮ LIỆU

- Bảng Customer_Registered

Column name	Data type	Meaning
ID	Bigint	Ex: 0,1, 2, 3, …
Contract	Varchar	Mã hợp đồng
LocationID	Int	Mã vị trí
BranchCode	Tinyint	Mã chi nhánh
Status	Tinyint	Trạng thái
Created_date	Datetime	Ngày đăng ký
Stop_date	Datetime	Ngày hủy

	ID	Contract	LocationID	BranchCode	Status	created_date	stopdate
1	0	SGDN00215	8	1	0	2011-11-25 10:48:13.860000	2012-01-05 10:02:10.000
2	1	SGDN00214	8	1	0	2012-06-14 18:55:25.517000	[NULL]
3	2	SGD374348	8	1	0	2012-11-01 18:59:04.603000	[NULL]
4	3	SGD022064	8	1	2	2011-06-22 14:54:30.997000	2013-05-29 13:57:51.000
5	4	SGD041015	8	5	2	2011-12-17 12:58:58.460000	2014-11-11 09:40:39.460

- Bảng Customer_Transaction

Column name	Data type	Meaning
ID	Bigint	Mã giao dịch
CustomerID	Varchar	Mã khách hàng
Purchase_Date	Datetime	Ngày giao dịch
GMV	Bigint	Giá trị giao dịch

	Transaction_ID	CustomerID	Purchase_Date	GMV
1	0	1,327,813	6/1/2022	95,000
2	1	1,157,830	6/1/2022	75,000
3	2	873,915	7/1/2022	95,000
4	3	3,505,071	7/1/2022	90,000
5	4	2,930,918	7/1/2022	109,091

CÁCH LÀM

- Lọc những khách hàng vẫn tiếp tục sử dụng dịch vụ (Join 2 bảng Customer_Registered và Customer_Transaction cùng với điều kiện Stop_date is null)
- Chỉ số recency = ngày báo cáo – ngày sử dụng dịch vụ gần nhất
- Chỉ số frequency = đếm số lần sử dụng dịch vụ
- Chỉ số monetary = tổng giá trị đã sử dụng dịch vụ

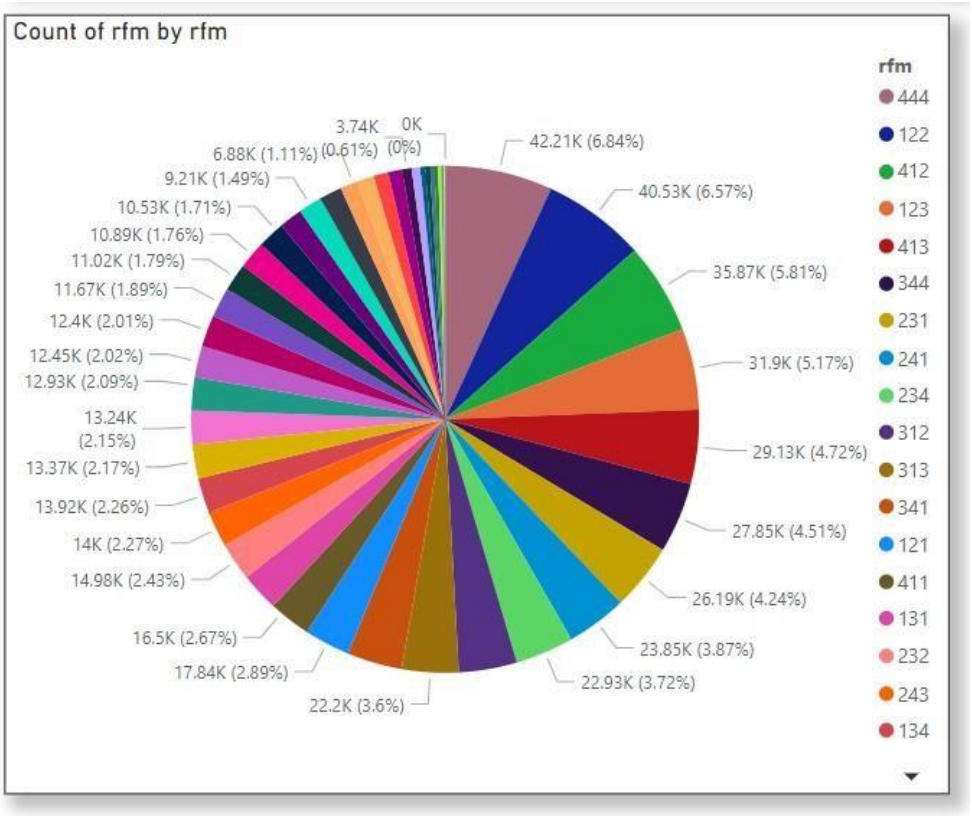
- Sử dụng windowfunction ntile phân chia bộ dữ liệu.
- Từ kết quả R – F – M, ta có bảng phân điểm như sau:

Điểm (IQR)	1	2	3	4
Recency	≥92 ngày	62 – 91 ngày	32 – 61 ngày	1 – 31 ngày
Frequency	1 lần	2 lần	3 lần	≥4 lần
Monetary	<70k	70k – 75k	75k – 95k	≥95k

Từ kết quả tính toán, với **617,164** khách hàng được chia thành **48** tổ hợp RFM khác nhau.

- Sử dụng phương pháp IQR (Interquartile Range) để tính điểm R – F – M.

Nhìn chung, các tổ hợp không có sự chênh lệch quá lớn về số lượng nhưng rất đa dạng về các tổ hợp với đủ các đặc điểm khác nhau. Điều bất ngờ khi tổ hợp (**444**) tương ứng với giá trị chi tiêu cao, thường xuyên sử dụng dịch vụ gần đây có số lượng khách hàng cao nhất với **42k**.



Khoảng một nửa số khách hàng tới từ 20% nhóm tổ hợp

Số khách hàng từ 10 tổ hợp đầu tiên rất đa dạng, mức chi tiêu trải dài từ thấp tới cao nhưng chỉ có 2 trong số 10 tổ hợp là có mức chi tiêu **dưới 70k** (231, 241). Bên cạnh đó, đa phần số khách hàng này đều sử dụng dịch vụ khá gần đây, không quá xa so với ngày báo cáo (01/09/2022) và mức chi tiêu tương đối khá.

rfm	total_users	CumulativeSum	Ratio (%)
444	42210	42210	6.84
122	40528	82738	13.41
412	35866	118604	19.22
123	31895	150499	24.39
413	29132	179631	29.11
344	27849	207480	33.62
231	26185	233665	37.86
241	23854	257519	41.73
234	22928	280447	45.44
312	22878	303325	49.15
313	22202	325527	52.75
341	21720	347247	56.26

Hơn 30% doanh thu tới từ số khách hàng trong 4 tổ hợp (444, 344, 122, 123) với mức chi tiêu chủ yếu là khá cùng với tần suất sử dụng tương đối.

Trong 4 tổ hợp thì tổng doanh thu từ **nhóm 444** khi chiếm gần **13%** tổng danh thu vì mức chi tiêu của nhóm này rất cao và thường xuyên sử dụng dịch vụ. Bên cạnh đó, doanh thu từ tệp khách hàng đã sử dụng dịch vụ từ lâu là tương đối cao. Điều đáng chú ý là **nhóm 344** tuy có số lượng khách hàng **top 6** nhưng mang lại doanh thu **top 2**, chứng tỏ mức chi tiêu trong nhóm này là cao hơn hẳn **nhóm 444**. *Vì vậy có thể tập trung vào nhóm này để phân tích và phát triển chiến lược riêng.*

rfm	revenue	CumulativeSum	ratio (%)
444	7357313751	7357313751	12.66
344	4508124329	11865438080	20.42
122	3068466459	14933904539	25.70
123	2999608275	17933512814	30.86
413	2726036886	20659549700	35.55
412	2723289591	23382839291	40.24
234	2344709444	25727548735	44.27
313	2096957366	27824506101	47.88
231	1859140473	29683646574	51.08
312	1717994206	31401640780	54.04
241	1676622681	33078263461	56.92
341	1558345601	34636609062	59.61

PHÂN LOẠI KHÁCH HÀNG

Dựa trên lý thuyết ma trận BCG về đánh giá tiềm năng phát triển danh mục sản phẩm để phân loại khách hàng thành 4 nhóm:

Phân khúc Stars – Nhóm khách hàng VIP: Thị phần cao, mức tăng trưởng lớn.

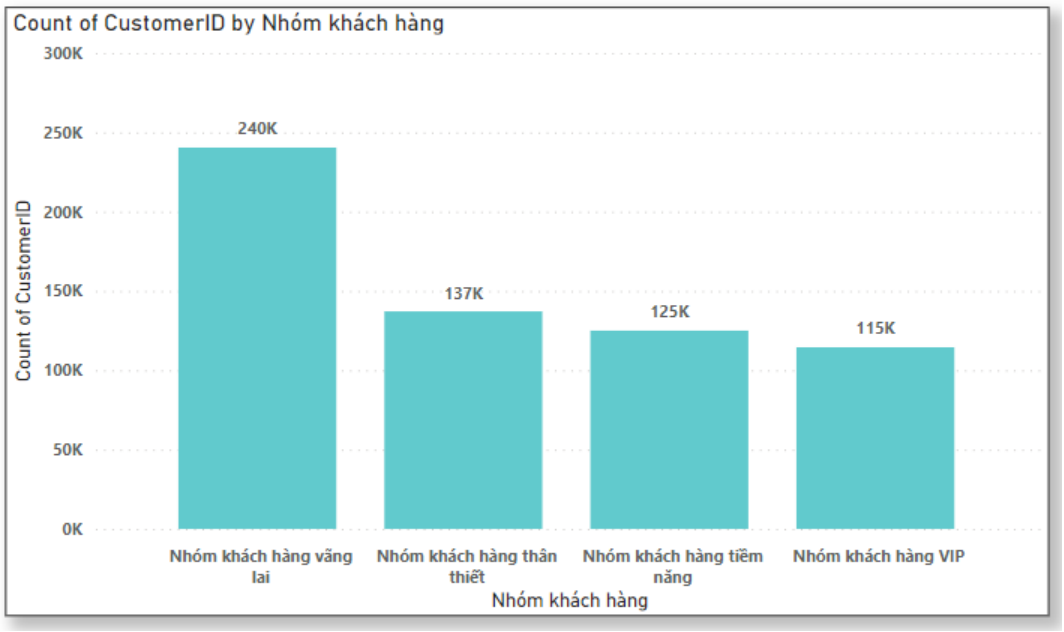
Phân khúc Question Marks – Nhóm khách hàng tiềm năng: Thị phần thấp, mức tăng trưởng lớn.

Phân khúc Cash Cows – Nhóm khách hàng thân thiết: Thị phần lớn, mức tăng trưởng thấp.

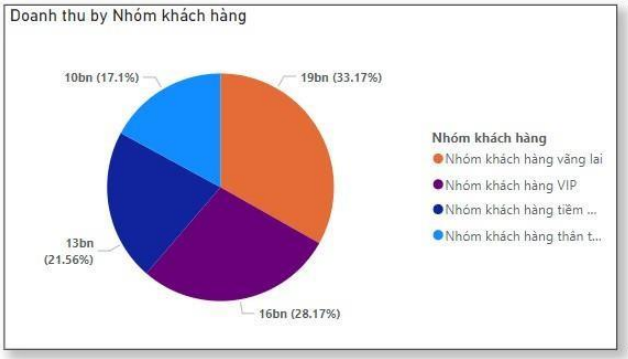
Phân khúc Dogs – Nhóm khách hàng vãng lai (mới): Thị phần thấp, mức tăng trưởng thấp.

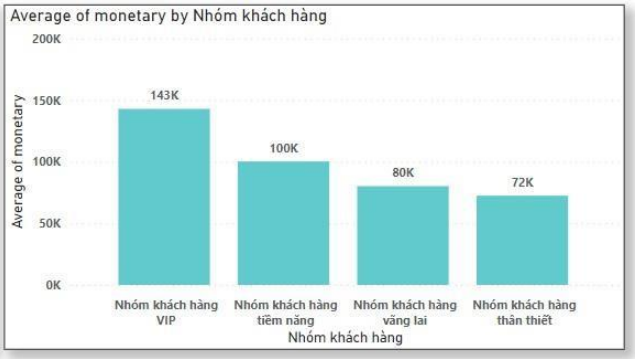
Nhóm khách hàng	Mô tả	Các tổ hợp RFM
Nhóm khách hàng VIP	Là nhóm khách hàng sử dụng dịch vụ thường xuyên, chi tiêu cao và gần đây hoặc khá gần đây	444, 443, 343, 344, 333, 334, 244, 243
Nhóm khách hàng tiềm năng	Là nhóm khách hàng sử dụng dịch vụ không thường xuyên nhưng chi tiêu cao và gần đây hoặc khá gần đây	223, 224, 233, 234, 413, 414, 423, 424
Nhóm khách hàng thân thiết	Là nhóm khách hàng sử dụng dịch vụ thường xuyên, có mức chi tiêu thấp và khá gần đây	221, 222, 231, 232, 241, 242, 331, 332, 341, 342, 441, 442, 421, 422
Nhóm khách hàng vãng lai	Là nhóm khách hàng sử dụng dịch vụ với mức chi tiêu trải đều và đã từ lâu hoặc mới tiếp cận với mức chi tiêu thấp	121, 122, 123, 124, 131, 132, 133, 134, 141, 142, 143, 144, 311, 312, 411, 412

Tệp khách hàng tiếp cận sử dụng dịch vụ trong 3 tháng tập trung chủ yếu là những **khách hàng mới** và **khách hàng đã sử dụng dịch vụ từ rất lâu**. Bên cạnh đó, 3 nhóm khách hàng còn lại không có sự khác biệt quá lớn, *điều này khá tích cực khi khả năng khách hàng tiềm năng có thể trở thành khách hàng VIP khá cao với chiến lược tiếp cận phù hợp.*



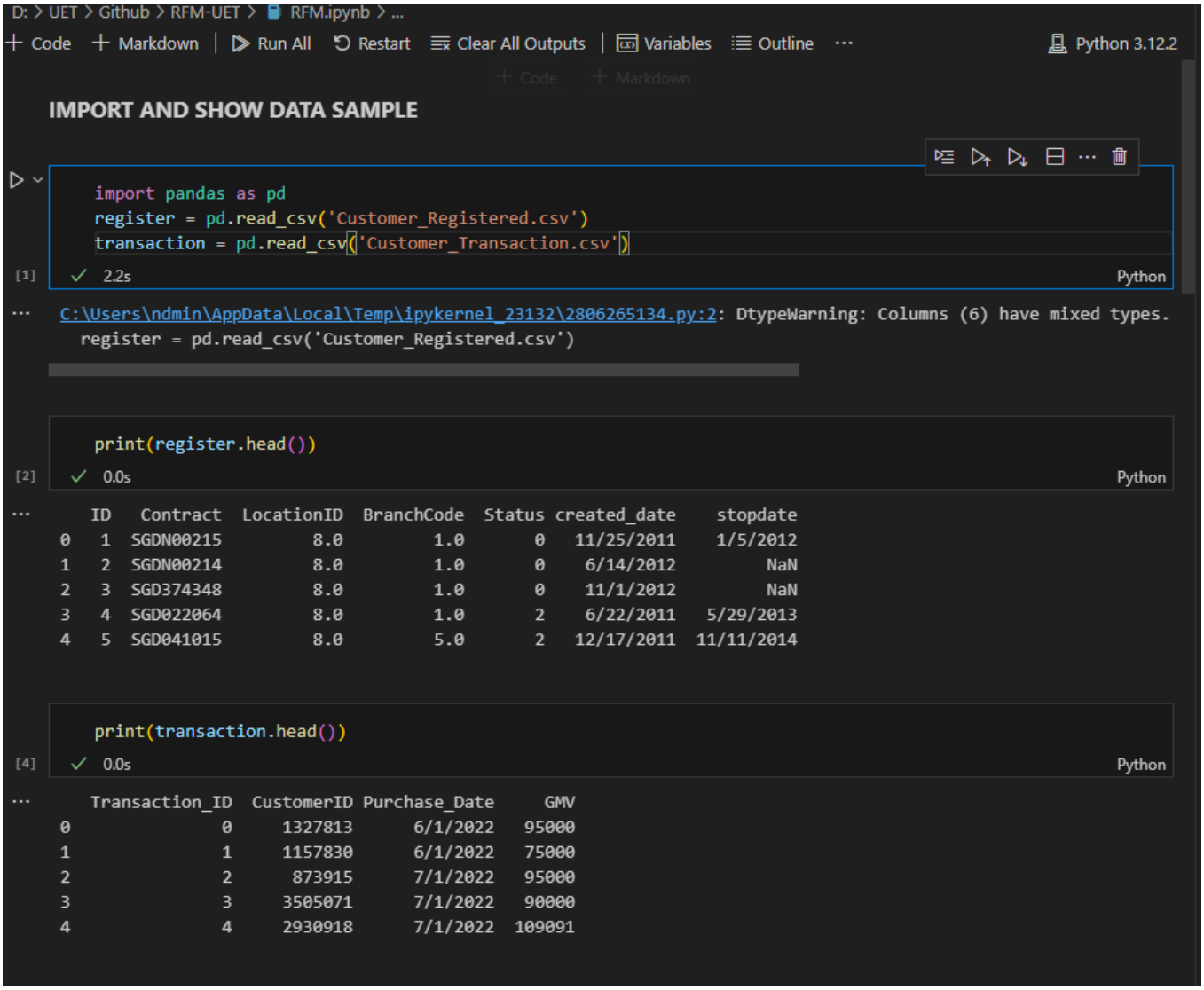
Hơn **30%** doanh thu xuất phát từ nhóm khách hàng khách hàng vãng lai với mức chi tiêu trung bình khoảng **80k.** Nguồn thu chính chủ yếu xuất phát từ những khách hàng mới tiếp cận sử dụng dịch vụ hoặc đã sử dụng từ rất lâu. *Với mức chi tiêu của nhóm khách hàng này ở mức điểm số 3 khá cao, vì vậy có thể lập kế hoạch tiếp cận nhóm đối tượng trên.*





III.CODE XỬ LÝ

NHẬP VÀ XEM DỮ LIỆU



Code và kết quả

TÍNH TOÁN CHỈ SỐ RFM

- Tính toán chỉ số R – F – M và chấm điểm
- Lưu kết quả RFM vào bảng Customer_RFM

```python
transaction['Purchase_Date'] = pd.to_datetime(transaction['Purchase_Date'])

valid_customers = register[(register['stopdate'].isna()) & (register['ID'] != 0)]

merged_df = pd.merge(valid_customers, transaction, left_on='ID', right_on='CustomerID')

reference_date = pd.to_datetime('2022-09-01')

rfm = merged_df.groupby('CustomerID').agg({
    'Purchase_Date': lambda x: (reference_date - x.max()).days,  # Recency
    'Transaction_ID': 'nunique',  # Frequency
    'GMV': 'sum'  # Monetary
}).reset_index()

rfm.columns = ['CustomerID', 'recency', 'frequency', 'monetary']

# Calculate R, F, M scores using quartiles
rfm['r_quartile'] = pd.qcut(rfm['recency'], 4, labels=False, duplicates='drop') + 1
rfm['f_quartile'] = pd.qcut(rfm['frequency'], 4, labels=False, duplicates='drop') + 1
rfm['m_quartile'] = pd.qcut(rfm['monetary'], 4, labels=False, duplicates='drop') + 1
rfm['f_quartile'] = 5 - rfm['f_quartile']

rfm['RFM_Score'] = rfm['r_quartile'].astype(str) + rfm['f_quartile'].astype(str) + rfm['m_quartile'].astype(str)

rfm['Date'] = reference_date.strftime('%Y-%m-%d')

print(rfm.head())
```

```
   CustomerID  recency  frequency  monetary  r_quartile  f_quartile  \
0       71739       62          1    105000           2           4
1       72014       62          2    254091           2           4
2       72052       31          1    145000           1           4
3       72657       62          1    200000           2           4
4       74549       62          1    125000           2           4

   m_quartile RFM_Score        Date
0           3       243  2022-09-01
1           4       244  2022-09-01
2           4       144  2022-09-01
3           4       244  2022-09-01
4           4       244  2022-09-01
```

```
✓ scores = rfm['RFM_Score']
```

```
print (scores.head())
```

```
[9] ✓ 0.0s
```

```
0    243
1    244
2    144
3    244
4    244
Name: RFM_Score, dtype: object
```

Code và kết quả

PHÂN LOẠI NHÓM KHÁCH HÀNG

```python
def classify_customer(rfm_score):
    vip = {'444', '443', '343', '344', '333', '334', '244', '243'}
    loyal = {'221', '222', '231', '232', '241', '242', '331', '332', '341', '342', '441', '442', '421', '422'}
    potential = {'223', '224', '233', '234', '413', '414', '423', '424', '313', '314'}

    if rfm_score in vip:
        return "Nhóm khách hàng VIP"
    elif rfm_score in loyal:
        return "Nhóm khách hàng thân thiết"
    elif rfm_score in potential:
        return "Nhóm khách hàng tiềm năng"
    else:
        return "Nhóm khách hàng vãng lai"

BGC_name = rfm['RFM_Score'].apply(classify_customer)

print(BGC_name.head())
```

Code

```
0          Nhóm khách hàng VIP
1          Nhóm khách hàng VIP
2    Nhóm khách hàng vãng lai
3          Nhóm khách hàng VIP
4          Nhóm khách hàng VIP
Name: RFM_Score, dtype: object
```

Kết quả

TÍNH TOÁN TỊNH TIẾN

- Tính toán số lượng khách hàng mỗi tổ hợp

⇒ Phục vụ mục đích tính toán tịnh tiến sau đó.

- Doanh thu tịnh tiến được áp dụng tương tự.

Số lượng khách hàng tịnh tiến

```python
rfm_total_users = rfm.groupby('RFM_Score').size().reset_index(name='total_users')

rfm_total_users['ID'] = rfm_total_users['total_users'].rank(method='first', ascending=False)

rfm_total_users['CumulativeSum'] = rfm_total_users['total_users'].cumsum()

# Calculate ratio (%) column
total_user_sum = rfm_total_users['total_users'].sum()
rfm_total_users['ratio (%)'] = (rfm_total_users['CumulativeSum'] / total_user_sum) * 100
rfm_total_users['ratio (%)'] = rfm_total_users['ratio (%)'].round(2)

result = rfm_total_users[['RFM_Score', 'total_users', 'CumulativeSum', 'ratio (%)']]

print(result)
```

Code

...	RFM_Score	total_users	CumulativeSum	ratio (%)
0	141	13744	13744	12.05
1	142	4648	18392	16.12
2	143	9271	27663	24.25
3	144	14386	42049	36.86
4	241	13743	55792	48.91
5	242	4627	60419	52.96
6	243	8952	69371	60.81
7	244	3673	73044	64.03
8	341	18322	91366	80.09
9	342	6354	97720	85.66
10	343	12085	109805	96.25
11	344	4276	114081	100.00

Kết quả

Doanh thu tịnh tiến

```python
rfm_revenue_users = rfm.groupby('RFM_Score')['monetary'].sum().reset_index(name='total_revenue')

rfm_revenue_users['ID'] = rfm_revenue_users['total_revenue'].rank(method='first', ascending=False)

rfm_revenue_users['CumulativeSum'] = rfm_revenue_users['total_revenue'].cumsum()

# Calculate ratio (%) column
total_revenue_sum = rfm_revenue_users['total_revenue'].sum()
rfm_revenue_users['ratio (%)'] = (rfm_revenue_users['CumulativeSum'] / total_revenue_sum) * 100
rfm_revenue_users['ratio (%)'] = rfm_revenue_users['ratio (%)'].round(2)

result = rfm_revenue_users[['RFM_Score', 'total_revenue', 'CumulativeSum', 'ratio (%)']]

print(result)
```

Code

...	RFM_Score	total_revenue	CumulativeSum	ratio (%)
0	141	973504768	973504768	8.75
1	142	392620627	1366125395	12.28
2	143	908425181	2274550576	20.45
3	144	2480266236	4754816812	42.74
4	241	980680368	5735497180	51.56
5	242	390906252	6126403432	55.07
6	243	877168721	7003572153	62.96
7	244	524794709	7528366862	67.68
8	341	1313774650	8842141512	79.49
9	342	536938514	9379080026	84.31
10	343	1184223766	10563303792	94.96
11	344	560957406	11124261198	100.00

Kết quả

PHÂN CÔNG NHIỆM VỤ

Tên	*Nhiệm vụ*
Nguyễn Đức Minh	Code, Report, Diagram
Nguyễn Phan Nam Sơn	Code (Main), Presentation
Phạm Minh Hiếu	Code, Research, Plan