

FINAL REPORT: PROJECT\_01

# CUSTOMER SEGMENTATION

DL07\_K306\_PHAM NGOC TRONG\_TRAN DINH HUNG

submitted: 23/08/2025

# AGENDA

**01. Business Objective**

**02. RFM analysis technique**

**03. Data Acquirement**

**04. Data Analysis**

**05. Project Structure**

**06. Data Workflow - Team task assignment**

**07. Visualization: Data overview**

**08. Visualization: RFM Quartile**

**09. Visualization: K-means (scikit-learn)**

**10. Visualization: Hierarchical clustering (scikit-learn)**

**11. Visualization: K-means (spark)**

**12. Results Comparison of 3 models**

**13. Recommendations**

**14. Our team**

# 01. Business Objective

Cửa hàng X chủ yếu bán các sản phẩm thiết yếu cho khách hàng như rau, củ, quả, thịt, cá, trứng, sữa, nước giải khát... Khách hàng của cửa hàng là khách hàng mua lẻ.

Chủ cửa hàng X mong muốn có thể bán được nhiều hàng hóa hơn cũng như giới thiệu sản phẩm đến **đúng đối tượng khách hàng**, chăm sóc và làm hài lòng khách hàng.



# 02. CUSTOMER SEGMENTATION

## RFM analysis technique

- **Recency:** Khoảng thời gian kể từ lần giao dịch gần nhất
- **Frequency:** Tần suất giao dịch của khách hàng
- **Monetary:** Tổng giá trị giao dịch

1

Manual RFM

2

Unsupervised Learning Algorithms

**K-Means clustering** với ML truyền thống (scikit-learn)

3

**Hierarchical clustering** với ML truyền thống (scikit-learn)

4

**K-Means clustering** bằng ML BigData (spark)

# 03. Data Acquisition

## Transactions.csv

- 38,765 records
- 4 attributes: Member\_number, Date, productId, items
- Date min: 01-01-2014
- Date max: 31-10-2015

	Member_number	Date	productId	items
0	1808	21-07-2015	1	3
1	2552	05-01-2015	2	1
2	2300	19-09-2015	3	3
3	1187	12-12-2015	4	3
4	3037	01-02-2015	2	1

## Products\_with\_Categories.csv

- 167 records = 167 products
- 4 attributes: productId, productName, price, Category

	productId	productName	price	Category
0	1	tropical fruit	7.8	Fresh Food
1	2	whole milk	1.8	Dairy
2	3	pip fruit	3.0	Fresh Food
3	4	other vegetables	0.8	Fresh Food
4	5	rolls/buns	1.2	Bakery & Sweets

# 04. Data Analysis

2 years of transactions:






































- 24 months
- 3,898 customers
- 167 products
- 14,963 orders (1 customer per date = 1 order)

\*\*DataFrame after combining, renaming columns, applying astype, calculate amount = product\_price \* quantity

	customer_id	order_date	product_id	quantity	product_name	product_price	product_category	amount	month	order_month
0	1808	2015-07-21	1	3	tropical fruit	7.8	Fresh Food	23.4	01-07-2015	2015-07-01
1	2552	2015-01-05	2	1	whole milk	1.8	Dairy	1.8	01-01-2015	2015-01-01
2	2300	2015-09-19	3	3	pip fruit	3.0	Fresh Food	9.0	01-09-2015	2015-09-01
3	1187	2015-12-12	4	3	other vegetables	0.8	Fresh Food	2.4	01-12-2015	2015-12-01
4	3037	2015-02-01	2	1	whole milk	1.8	Dairy	1.8	01-02-2015	2015-02-01



# 05. Project Structure



Name			
 data_input	Project_1		
 data_output	data_input		
 images	Products_with_Categories.csv		
 models	Transactions.csv		
 source_code	data_output		
 READ ME.txt	MLBD_01_df_trans_pre.csv	 MLBD_14_KMeans_segment_scatterplot.png	models
 topic1_Customer_Segmentation_19082025.pdf	MLBD_05_df_KMeans_cluster.csv	 MLBD_15_KMeans_segment_treemap.png	MLBD_kmeans_model_with_scale
	MLBD_06_df_KMeans_cluster_agg.csv	 MLBD_16_KMeans_segment_treemap_labeled.png	MLTT_hierachical_model_with_scale.pkl
	MLBD_07_df_KMeans_cluster_agg_labeled.csv	 MLBD_17_KMeans_segment_scatterplot_labeled.png	MLTT_kmeans_model_with_scale.pkl
	MLBD_08_df_KMeans_cluster_labeled.csv	 MLBD_18_KMeans_cluster_boxplot_labeled.png	source_code
	MLTT_01_df_trans_pre.csv	 MLTT_01_RFM_distributions_ori.png	project1_MLDuLieuLon.ipynb
	MLTT_02_df_RFM_scored.csv	 MLTT_02_RFM_pairplot_ori.png	project1_MLTruyenThong.ipynb
	MLTT_03_df_RFM_labeled.csv	 MLTT_03_RFM_correlation_matrix.png	Final_Report.pptx
	MLTT_04_df_RFM_labeled_agg.csv	 MLTT_04_RFM_check_outliers_before_scale_boxplot.png	READ ME.txt
	MLTT_05_df_KMeans_cluster.csv	 MLTT_05_RFM_distributions_after_scale.png	topic1_Customer_Segmentation_19082025.pdf
	MLTT_06_df_KMeans_cluster_agg.csv	 MLTT_06_RFM_map_sample.png	
	MLTT_07_df_KMeans_cluster_agg_labeled.csv	 MLTT_07_RFM_segment_barchart.png	
	MLTT_08_df_KMeans_cluster_labeled.csv	 MLTT_08_RFM_segment_boxplot.png	
	MLTT_09_df_Hierachical_cluster.csv	 MLTT_09_RFM_segment_scatterplot.png	
	MLTT_10_df_Hierachical_cluster_agg.csv	 MLTT_10_RFM_segment_treemap.png	
	MLTT_11_df_Hierachical_cluster_agg_labeled.csv	 MLTT_11_KMeans_optimal_k_with_elbow_silhouette.png	
	MLTT_12_df_Hierachical_cluster_labeled.csv	 MLTT_12_KMeans_segment_barchart.png	
	MLTT_13_df_RFM_after_scale.csv	 MLTT_13_KMeans_cluster_boxplot.png	
	MLTT_df_products_profile.html	 MLTT_14_KMeans_segment_scatterplot.png	
	MLTT_df_transactions_profile.html	 MLTT_15_KMeans_segment_treemap.png	
		 MLTT_16_KMeans_segment_treemap_labeled.png	
		 MLTT_17_KMeans_segment_scatterplot_labeled.png	
		 MLTT_18_KMeans_cluster_boxplot_labeled.png	
		 MLTT_19_Hierachical_optimal_k_with_silhouette.png	
		 MLTT_20_Hierachical_vs_Kmeans_silhouette.png	
		 MLTT_21_Hierachical_dendrogram.png	
		 MLTT_22_Hierachical_vs_KMeans_PCA_and_Clustering.png	
		 MLTT_23_Hierachical_segment_treemap_labeled.png	
		 MLTT_24_Hierachical_segment_scatterplot_labeled.png	
		 MLTT_25_Hierachical_cluster_boxplot_labeled.png	
		models	

# 06. Data Workflow

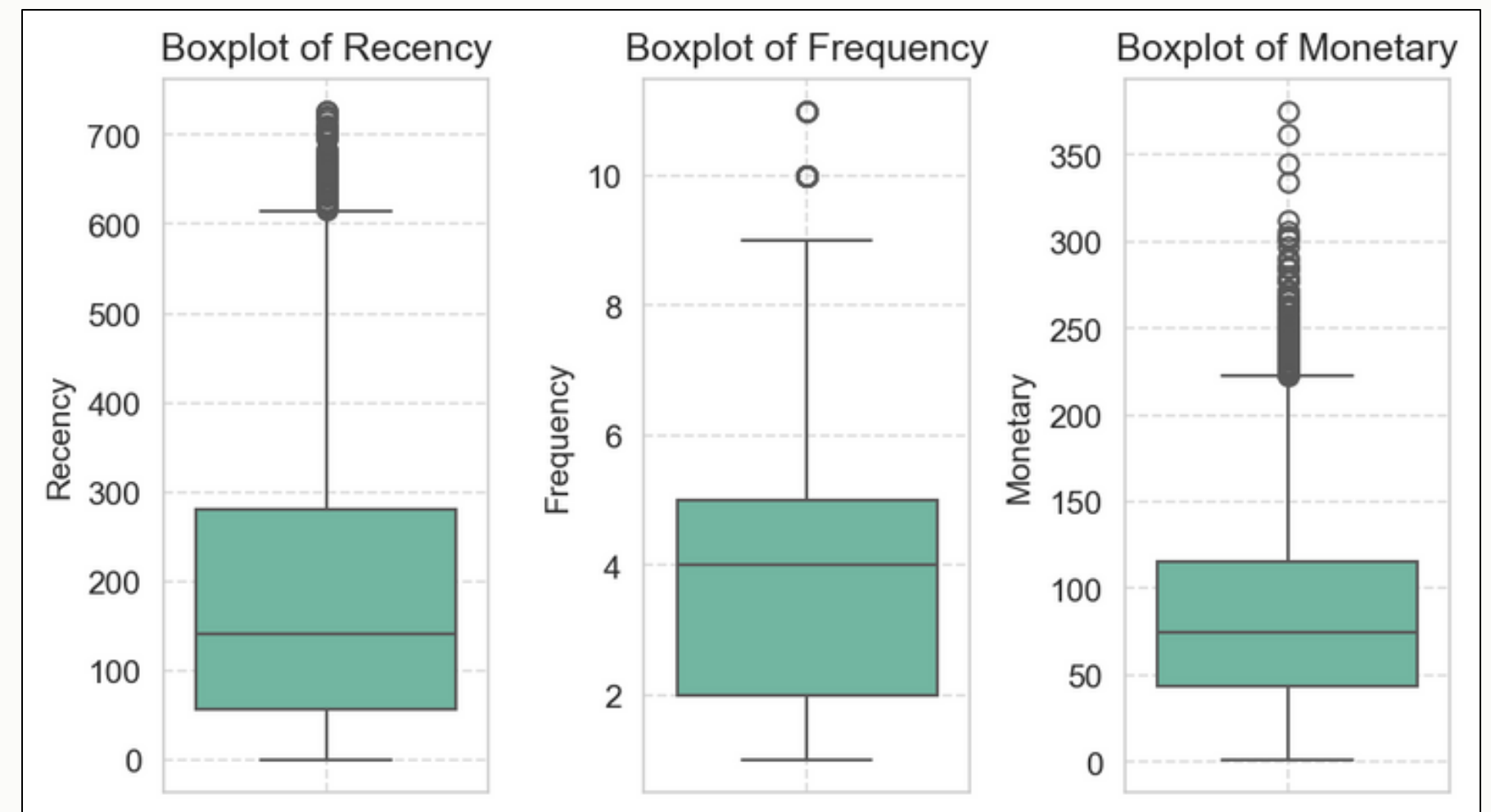
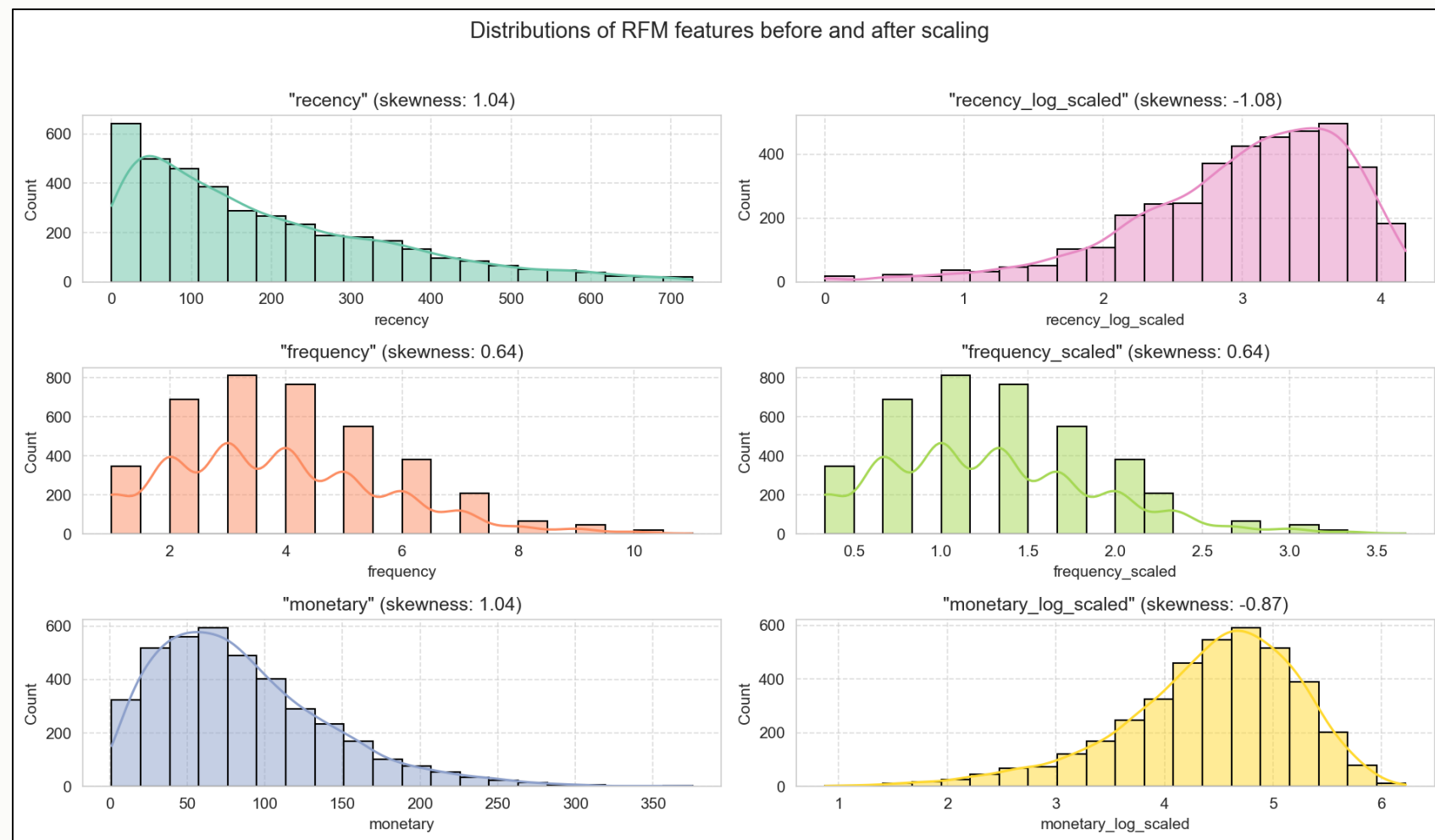
	RFM Manual	Unsupervised Learning Algorithms		
<b><u>ETL</u></b> <ul style="list-style-type: none"><li>• <b>Rename columns</b></li><li>• Check for Null / NaN values</li><li>• <b>Merge / join product data into transaction data</b></li><li>• Cast data types (order_date)</li><li>• <b>Validate records:</b><ul style="list-style-type: none"><li>• amount &lt;= 0</li><li>• order_date &gt;= today()</li><li>• product_price &lt;= 0</li></ul></li><li>• <b>Check for outliers</b></li></ul>	<b><u>RFM scoring format</u></b> <ul style="list-style-type: none"><li>• Check the data distribution of Recency, Frequency, and Monetary</li><li>• Examine the <b>correlation matrix</b> between Recency, Frequency, and Monetary</li><li>• <b>Calculate R-F-M scores</b> based on Recency, Frequency, and Monetary by grouping into quartiles</li><li>• <b>Create RFM_segment</b> using 4x4x4 group combinations</li></ul>	<b><u>Preprocessing for K-Means &amp; Hierachical clustering (scikit-learn)</u></b> <ul style="list-style-type: none"><li>• Apply <b>log transform</b> to Recency and Monetary</li><li>• Use <b>RobustScaler</b> to handle outliers</li><li>• <b>Re-evaluate</b> the distribution of Recency, Frequency, and Monetary after scaling</li></ul>		<b><u>Preprocessing for K-Means (spark)</u></b> <ul style="list-style-type: none"><li>• Log transform: Recency, Monetary</li><li>• <b>RobustScaler</b>: handle outliers using <b>VectorAssembler</b></li><li>• <b>Re-evaluate</b> the distribution of Recency, Frequency, and Monetary after scaling</li></ul>
	<b><u>RFM Segmentation Labeling</u></b> <ul style="list-style-type: none"><li>• Define <b>RFM based rules</b> (illustrated with an <b>R-FM sample map</b>)</li><li>• Assign labels to RFM segments based on the rules</li><li>• Summarize and visualize the results with charts</li></ul>	<b><u>K-Means Modeling (scikit-learn)</u></b> <ul style="list-style-type: none"><li>• Determine the <b>optimal k using Elbow</b> and <b>Silhouette</b> methods</li><li>• <b>Identify k = 4</b> as the best separation</li><li>• Build K-Means model with k = 4</li><li>• Summarize and visualize clusters with charts</li><li>• <b>Analyze cluster characteristics</b> based on Recency_mean, Frequency_mean, and Monetary_mean</li><li>• Assign appropriate <b>segment labels</b></li><li>• <b>Visualize clusters</b> after labeling</li></ul>	<b><u>Hierachical Clusterting Modeling (scikit-learn)</u></b> <ul style="list-style-type: none"><li>• Determine the <b>optimal k</b> using <b>Silhouette</b> methods</li><li>• <b>Identify k = 4</b> as the best separation</li><li>• Build Hierachical clustering model with k = 4</li><li>• Summarize and visualize clusters with charts</li><li>• Analyze cluster characteristics based on Recency_mean, Frequency_mean, and Monetary_mean</li><li>• Assign appropriate segment labels</li><li>• Visualize clusters after labeling</li></ul>	<b><u>K-means Modeling (spark)</u></b> <ul style="list-style-type: none"><li>• Determine the <b>optimal k using Elbow</b> and <b>Silhouette</b> methods</li><li>• <b>Identify k = 4</b> as the best separation</li><li>• Build K-Means model with k = 4</li><li>• Summarize and visualize clusters with charts</li><li>• <b>Analyze cluster characteristics</b> based on Recency_mean, Frequency_mean, and Monetary_mean</li><li>• Assign appropriate <b>segment labels</b></li><li>• <b>Visualize clusters</b> after labeling</li></ul>



# 06. Team task assignment

Vai trò:			
		Phạm Ngọc Trọng	Project Lead
		Trần Đình Hùng	Commercial Business Domain Advisor
Phân task:			
		 Trọng	 Hùng
1	- Problem analysis - Dataset analysis - Project timeline	x	x
2	- Data combine & ETL, transform & scale, detect outliers, check distribution, handle nulls, validate records	Hardcode, plot charts, write functions, save cleaned dataset as .csv files	Review distribution, suggest transformation, apply scaling
3	RFM-based rules, segmentation, scoring	Propose RFM-based rules, implement segmentation and scoring, visualize results with charts	Review practicality of Trọng's RFM-based rules, adjust and refine based on real business context
4	- RFM scoring & labeling	Calculate RFM scores, assign labels, visualize charts, save cleaned datasets to CSV	
5	- Choose the optimal k-clusters for the models	Plot Elbow and Silhouette charts, analyze results, and justify selecting k=4 for the models	Advise, critique, and discuss with Trọng to validate the choice
6	- Modeling, feature statistics, label selection, cluster labeling, chart visualization, and file export	Hardcode, plot charts, write functions, and save the cleaned dataset as .csv files	
7	- Compare results, optimize the model, and fine-tune model parameters	Perform model optimization	Participate in comparison and evaluation
8	- Prepare a presentation	Write version_2 of Final report.pptx	Write version_1 of Final report.pptx
9	- Organize files/folders and sync the latest versions	x	
10	- Present and defend the project	Present model results	Present the project outline

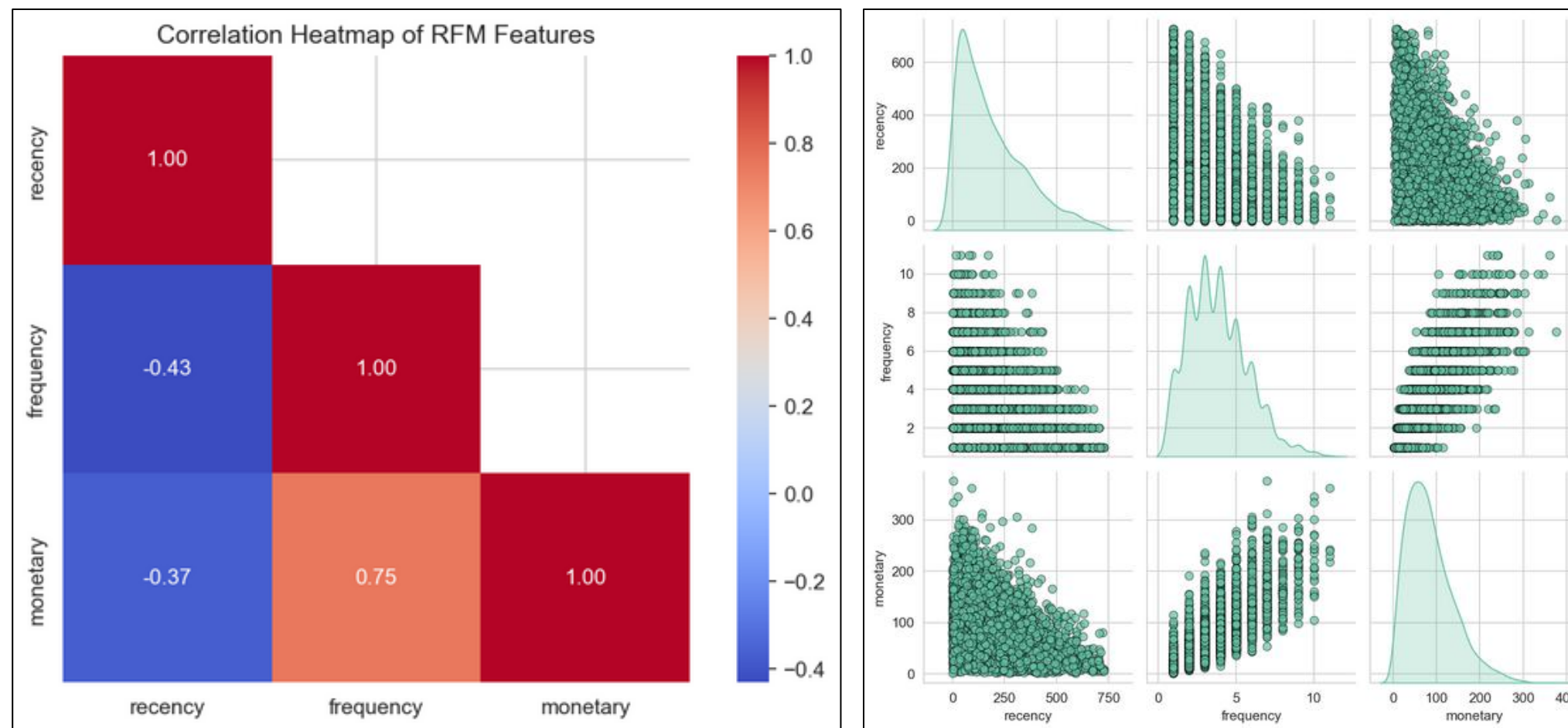
# 07. Visualization: Data overview



Dimension	Nhận xét biểu đồ
Recency	Lệch phải, đa số mới mua, đa phần là khách không quay lại
Frequency	Chủ yếu 2–4 lần, ít khách hàng mua nhiều
Monetary	Phần lớn chi tiêu thấp–trung bình

Feature	Nhận xét	Transform	Scaler
Recency	Phân bố trải rộng, có nhiều outlier, lệch phải	Log1p (giảm skew)	RobustScaler
Frequency	Giá trị nhỏ (1–10), có 2 outlier	–	RobustScaler
Monetary	Phân phối lệch phải mạnh, nhiều outlier chi tiêu cao, lệch phải	Log1p (giảm skew)	RobustScaler

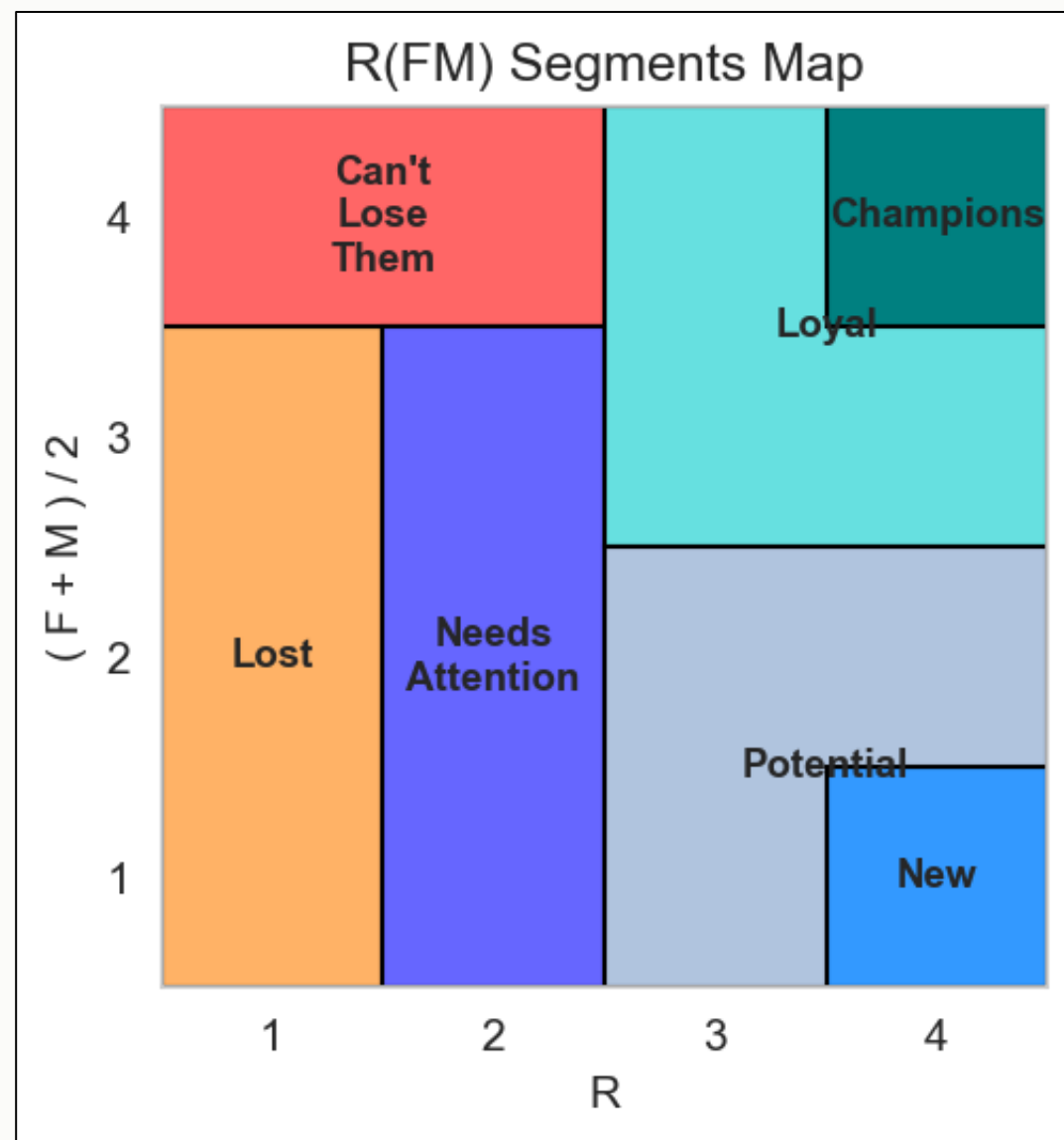
# 07. Visualization: Data overview



Nhận xét:

- Chỉ có F và M tương tác thuận mạnh

# 08. Visualization: RFM Quartile



## Phân nhóm khách hàng RFM rút gọn

Trong mô hình RFM (*Recency – Frequency – Monetary*), số nhóm khả dĩ có thể lên đến:

$$4 \times 4 \times 4 = 64$$

Việc phân chia chi tiết như vậy gây khó khăn trong phân tích và truyền đạt kết quả. Để đơn giản hóa, nhóm em đề xuất:

- Kết hợp **Frequency (F)** và **Monetary (M)** thành chỉ số FM đại diện (ở corr heatmap bên trên, F và M có tương quan thuận mạnh 0.75):

$$FM = \frac{F + M}{2}$$

- Giữ **Recency (R)** làm trục độc lập.
- Từ đó, giảm số nhóm còn **7 phân khúc chính**, giúp trực quan và dễ ứng dụng trong xử lý bài toán.

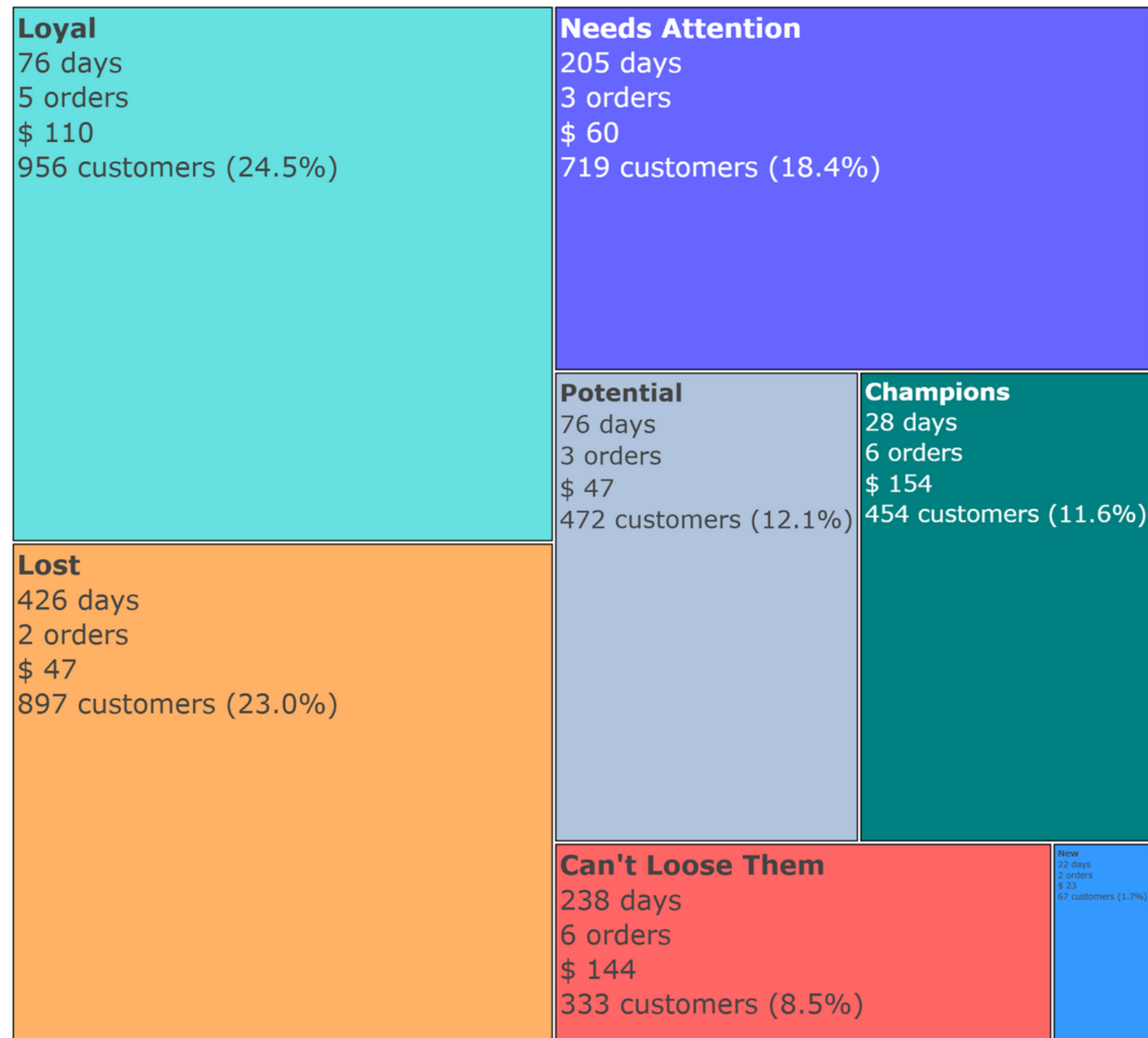
# 08. Visualization: RFM Quartile

Segment	Customer count	Customer %	Revenue %	Recency mean	Frequency mean	Monetary mean	Nhận xét chính
Champions	454	11.6%	21.0%	28	6	154	Khách hàng mua gần đây nhất, mua thường xuyên, và chi tiêu trung bình cao. Đóng góp tỷ lệ doanh thu vượt trội.
Loyal	956	24.5%	31.7%	76	5	110	Nhóm lớn nhất cả về số lượng và doanh thu. Tần suất mua ổn định, chi tiêu trung bình khá cao.
Potential	472	12.1%	6.7%	76	3	47	Khách regular + potential gộp lại
New	67	1.7%	0.5%	22	2	23	Nhóm nhỏ nhất, đóng góp không đáng kể, vì họ chỉ mới mua hàng gần đây.
Can't Lose Them	333	8.5%	14.4%	238	6	144	Khách hàng từng chi tiêu cao và mua nhiều lần nhưng đã lâu không quay lại. Doanh thu đóng góp tương tự như nhóm Champions.
Needs Attention	719	18.4%	13.0%	205	3	60	Nhóm chiếm tỷ trọng khách hàng lớn nhưng mức chi tiêu trung bình và tần suất mua thấp. 7 tháng họ chưa mua lại.
Lost	897	23.0%	12.8%	426	2	47	Tần suất không cao, doanh thu có tính chất của khách regular + potential đã từng mua hàng trước đó 1,5 năm.

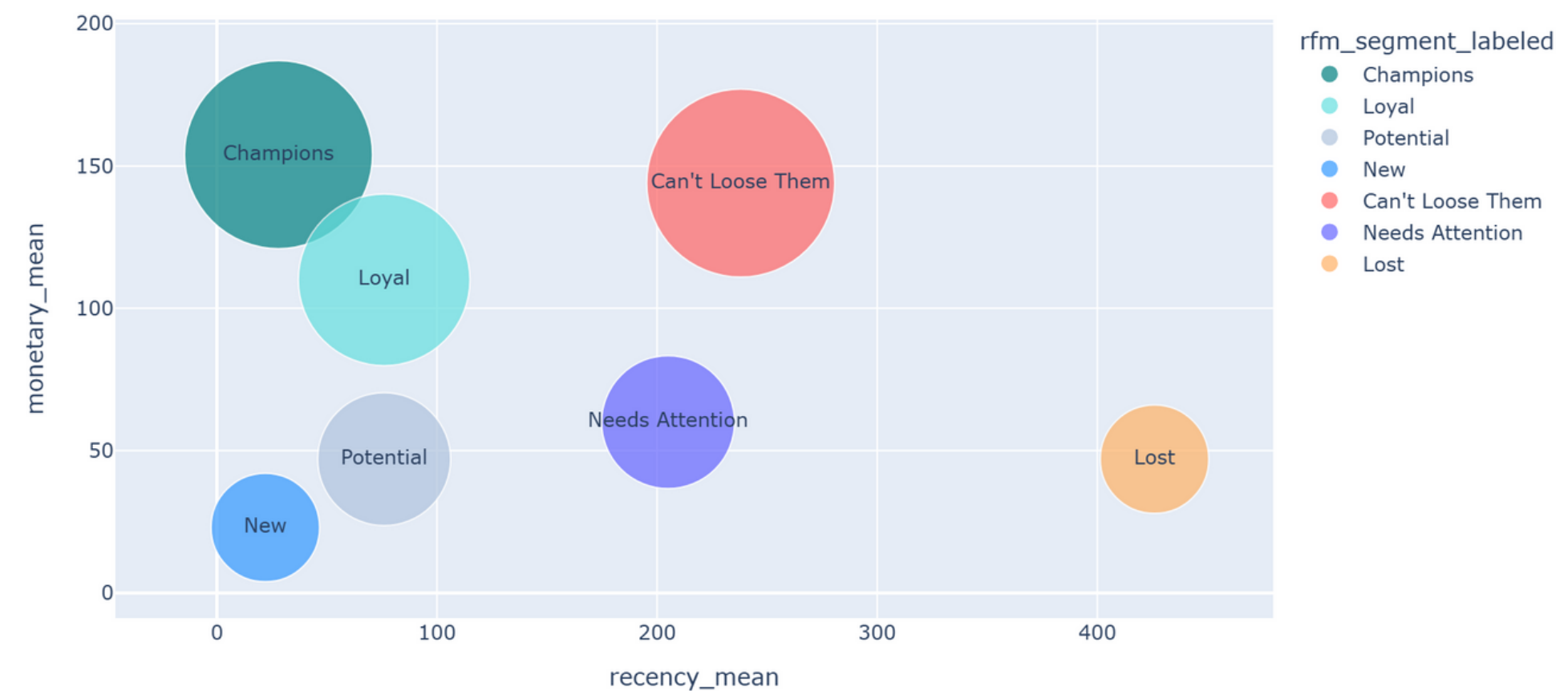


# 08. Visualization: RFM Quartile

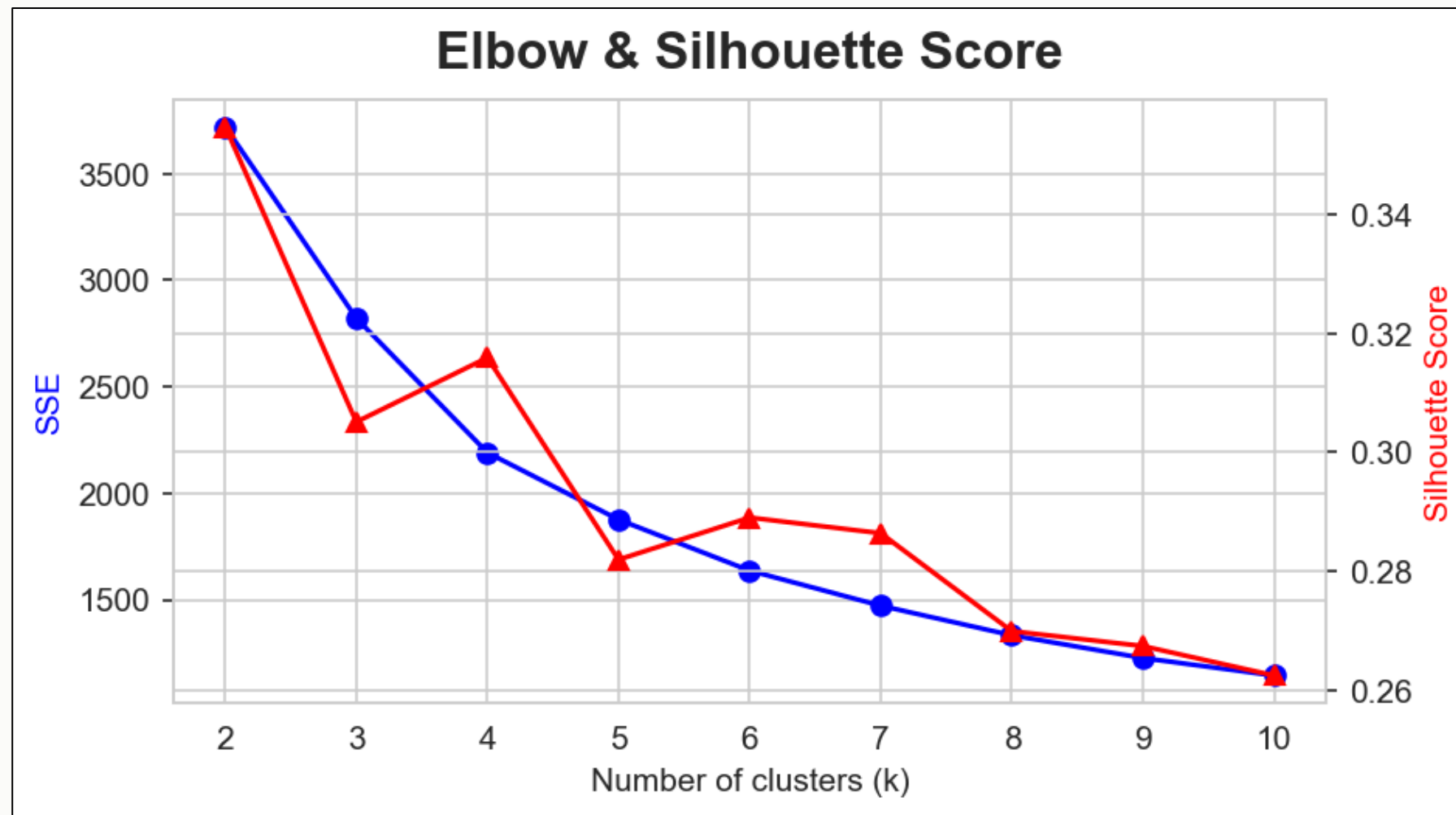
Customer Segmentation Distribution (Treemap)



Customer Segmentation by RFM (Recency, Frequency, Monetary)



# 09. Visualization: K-means (scikit-learn)



Phương pháp	Nhận xét
Elbow	Đường cong gãy rõ ở k=4, sau đó giảm chậm, chọn k=4.
Silhouette Score	Cao nhất ở k=2 nhưng quá ít cụm. Đỉnh tiếp theo ở k=4.

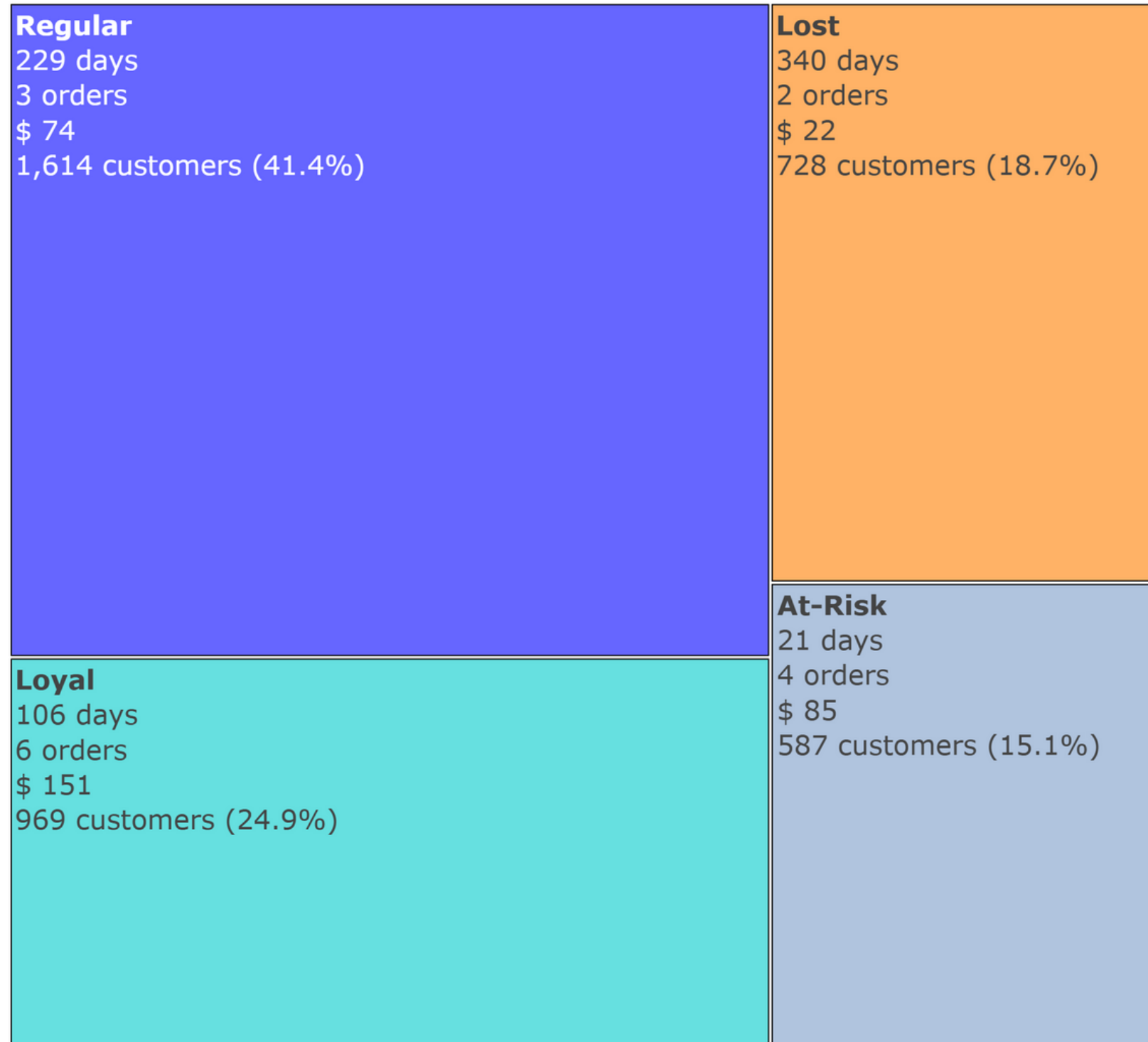
kết hợp 2 phương pháp, chọn **k = 4**

# 09. Visualization: K-means (scikit-learn)

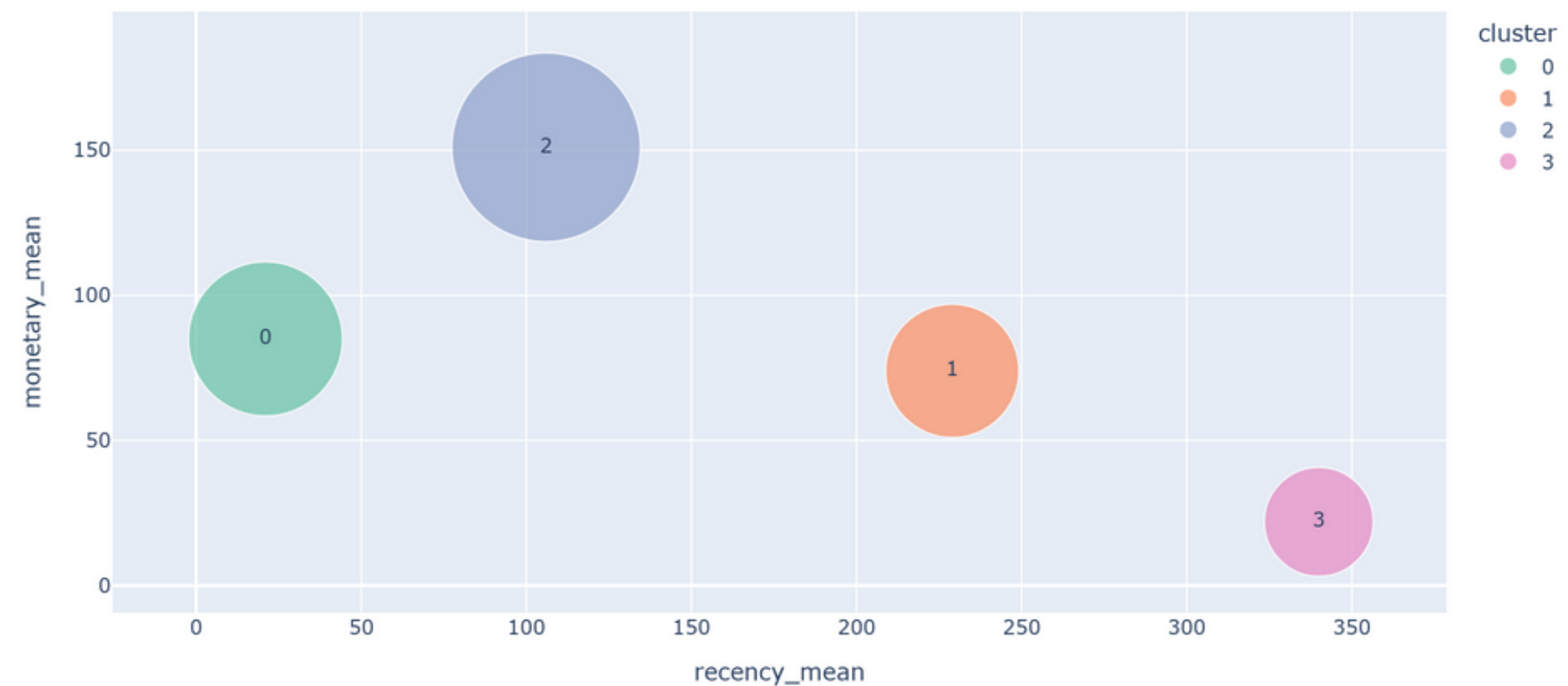
Cluster	Segment	Customer count	Customer %	Revenue %	Recency mean	Frequency mean	Monetary mean	Nhận xét chính
2	Loyal	969	24.9%	44.2%	106	6	151	Nhóm khách hàng trung thành: mua khá thường xuyên, chi tiêu cao, đóng góp doanh thu lớn nhất.
1	Regular	1614	41.4%	36.0%	229	3	74	Nhóm đông nhất: mua ở mức ổn định, chi tiêu trung bình khá, đóng góp doanh thu cao thứ hai.
0	At-Risk	587	15.1%	14.9%	21	4	85	Khách từng mua gần đây, tần suất vừa phải, chi tiêu trung bình. Tuy nhiên tỷ trọng nhỏ, cần theo dõi.
3	Lost	728	18.7%	4.9%	340	2	22	Nhóm khách hàng gần như mất: lâu không quay lại, mua rất ít, chi tiêu thấp, đóng góp doanh thu không đáng kể.

# 09. Visualization: K-means (scikit-learn)

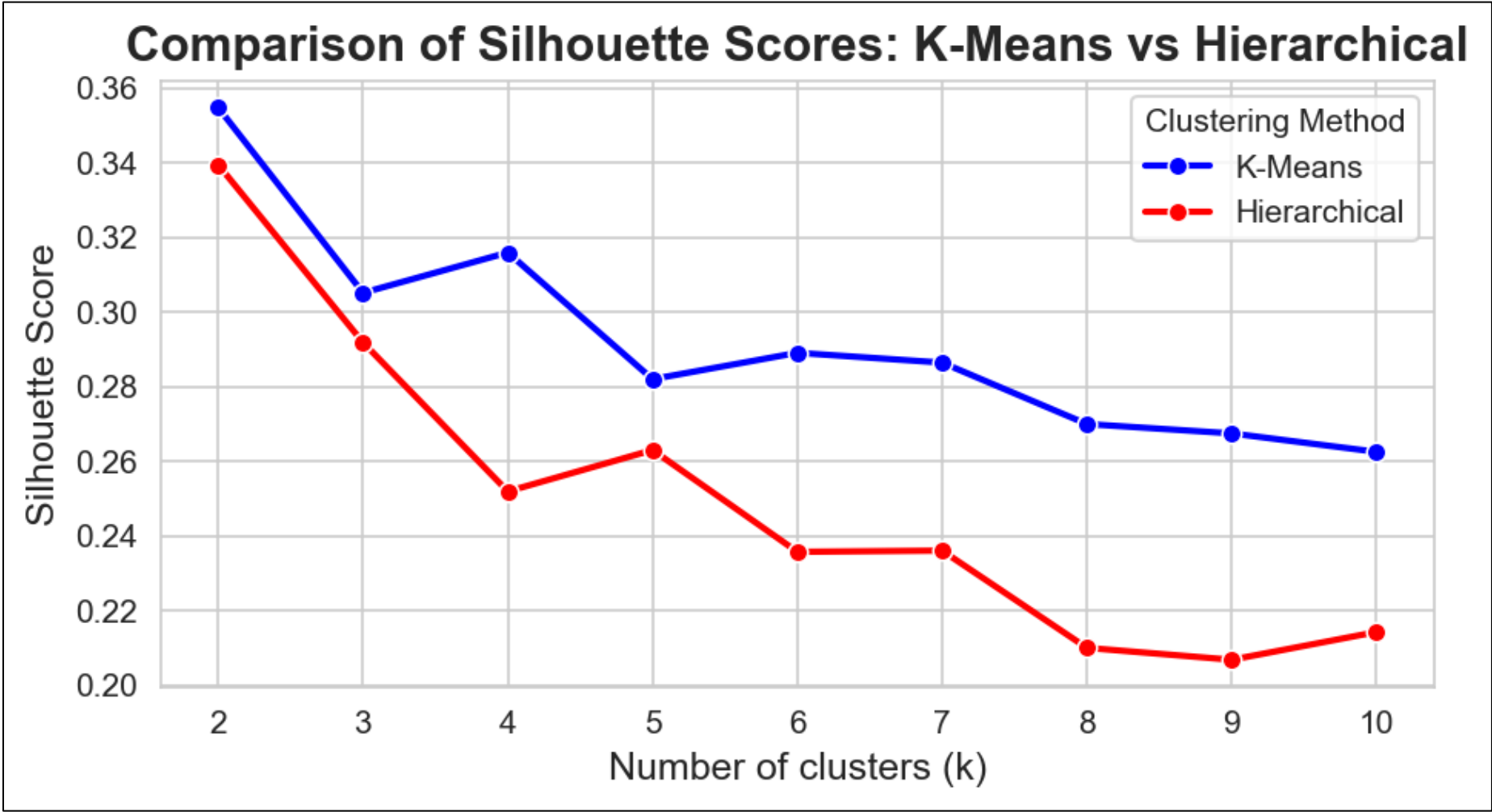
Customer Segmentation Distribution (Treemap)



Customer Segmentation by KMeans (Recency, Frequency, Monetary)



# 10. Visualization: Hierarchical clustering (scikit-learn)



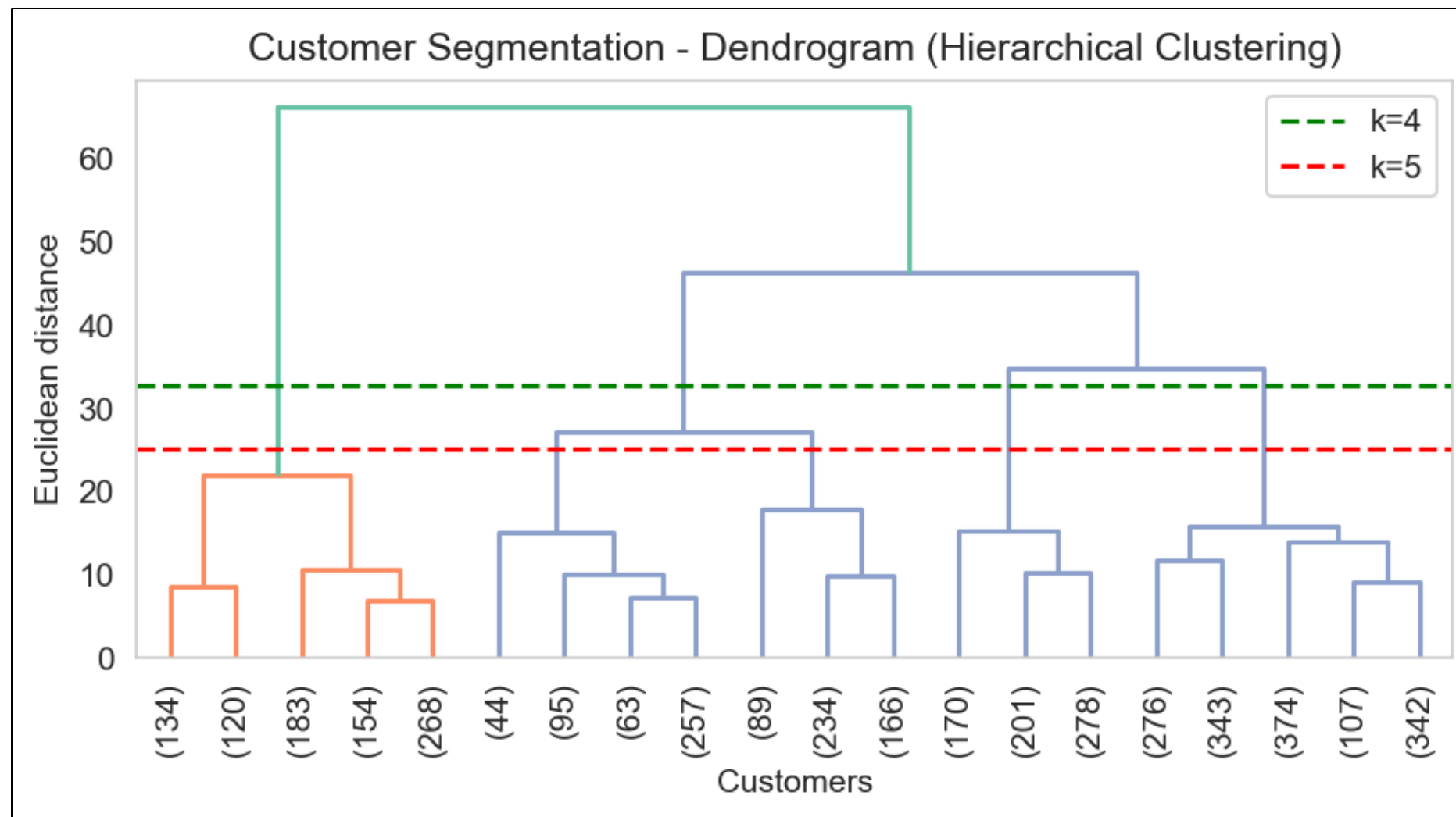
**Nhận xét** Silhouette score của Hierarchical

- khi  $k=2, k=3$ : điểm cao nhất nhưng ít ý nghĩa khi phân cụm.
- khi  $k=5$  silhouette của Hierarchical cao hơn  $k=4$ , gây mâu thuẫn với K-Means model.

k	K-Means	Hierarchical
2	cao nhất	cao nhất
3	cao nhì	cao nhì
4	cao hơn k=5	thấp hơn k=5
5	thấp hơn k=4	cao hơn k=4

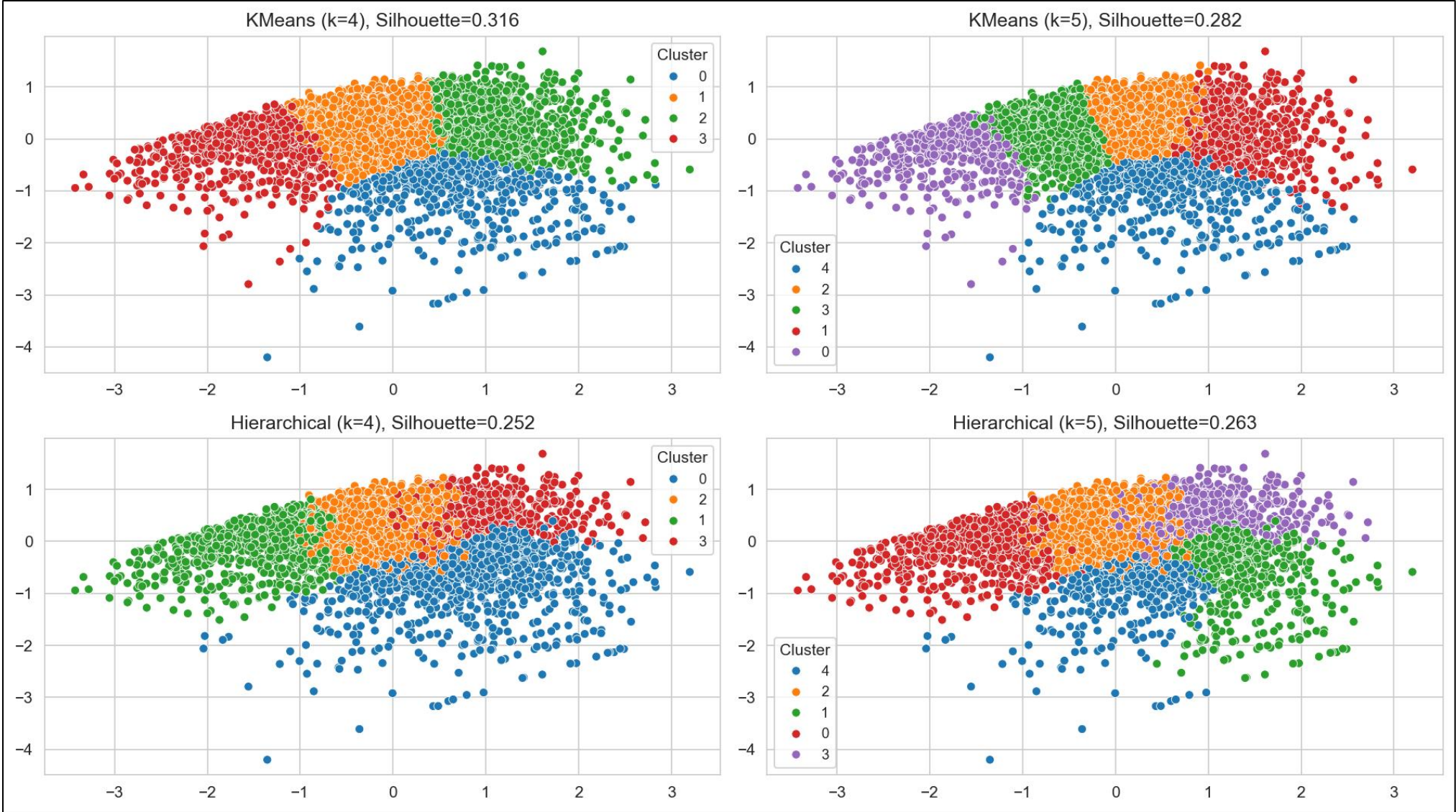


# 10. Visualization: Hierarchical clustering (scikit-learn)



# 10. Visualization: Hierarchical clustering

## (scikit-learn)



Thêm 1 cách tham khảo khác, chúng em sử dụng thêm kỹ thuật **giảm chiều PCA** (Principal Component Analysis):

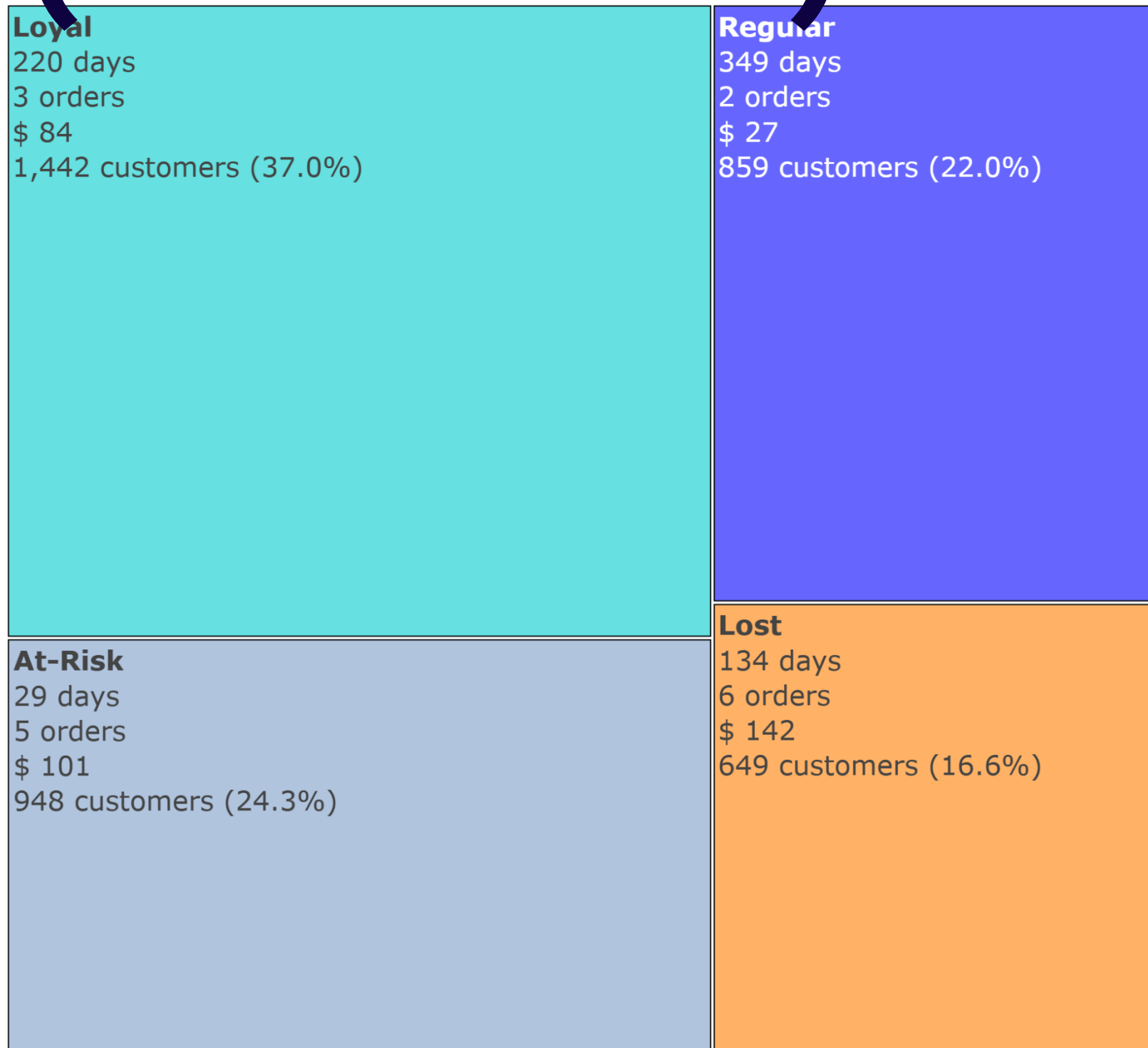
- để so sánh sự phân cụm giữa K-Means và Hierarchical
- khi:
  - k=4
  - và k=5 sẽ phân cụm ra sao

Model	k	Silhouette Score	Nhận xét
K-Means	4	0.316	Silhouette cao hơn, ranh giới cụm rõ nhất
K-Means	5	0.282	Điểm silhouette giảm, cụm kém rõ ràng
Hierarchical	4	0.252	Silhouette thấp, nhiều điểm bị lẫn cụm
Hierarchical	5	0.263	Silhouette cao hơn k=4 nhưng vẫn kém K-Means
Rút ra			<b>Chọn k=4</b> để giảm số lượng group: <ul style="list-style-type: none"><li>- dễ gán nhãn cho từng segment</li><li>- dễ so sánh kết quả giữa các model</li></ul>

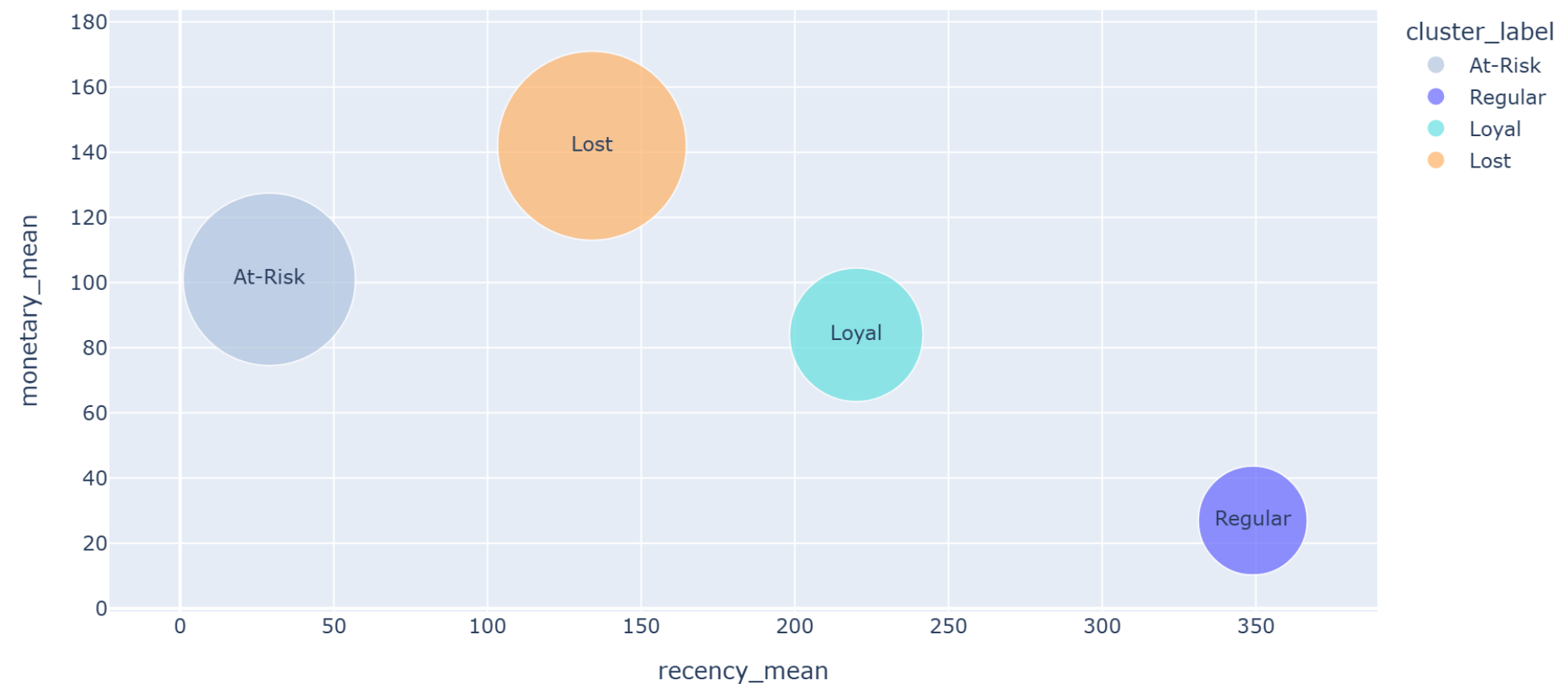
# 10. Visualization: Hierarchical clustering (scikit-learn)

Cluster	Segment	Customer count	Customer %	Revenue %	Recency mean	Frequency mean	Monetary mean	Nhận xét chính
3	Loyal	649	16.6%	27.7%	134	6	142	Nhóm khách hàng trung thành: mua khá thường xuyên, chi tiêu cao.
2	Regular	1442	37.0%	36.6%	220	3	84	Nhóm đông nhất: mua ở mức ổn định, chi tiêu trung bình khá.
0	At-Risk	948	24.3%	28.8%	29	5	101	Khách từng mua gần đây, tần suất vừa phải, chi tiêu trung bình. Tuy nhiên tỷ trọng nhỏ, cần theo dõi.
1	Lost	859	22.0%	6.9%	349	2	27	Nhóm khách hàng gần như mất: lâu không quay lại, mua rất ít, chi tiêu thấp, đóng góp doanh thu không đáng kể.

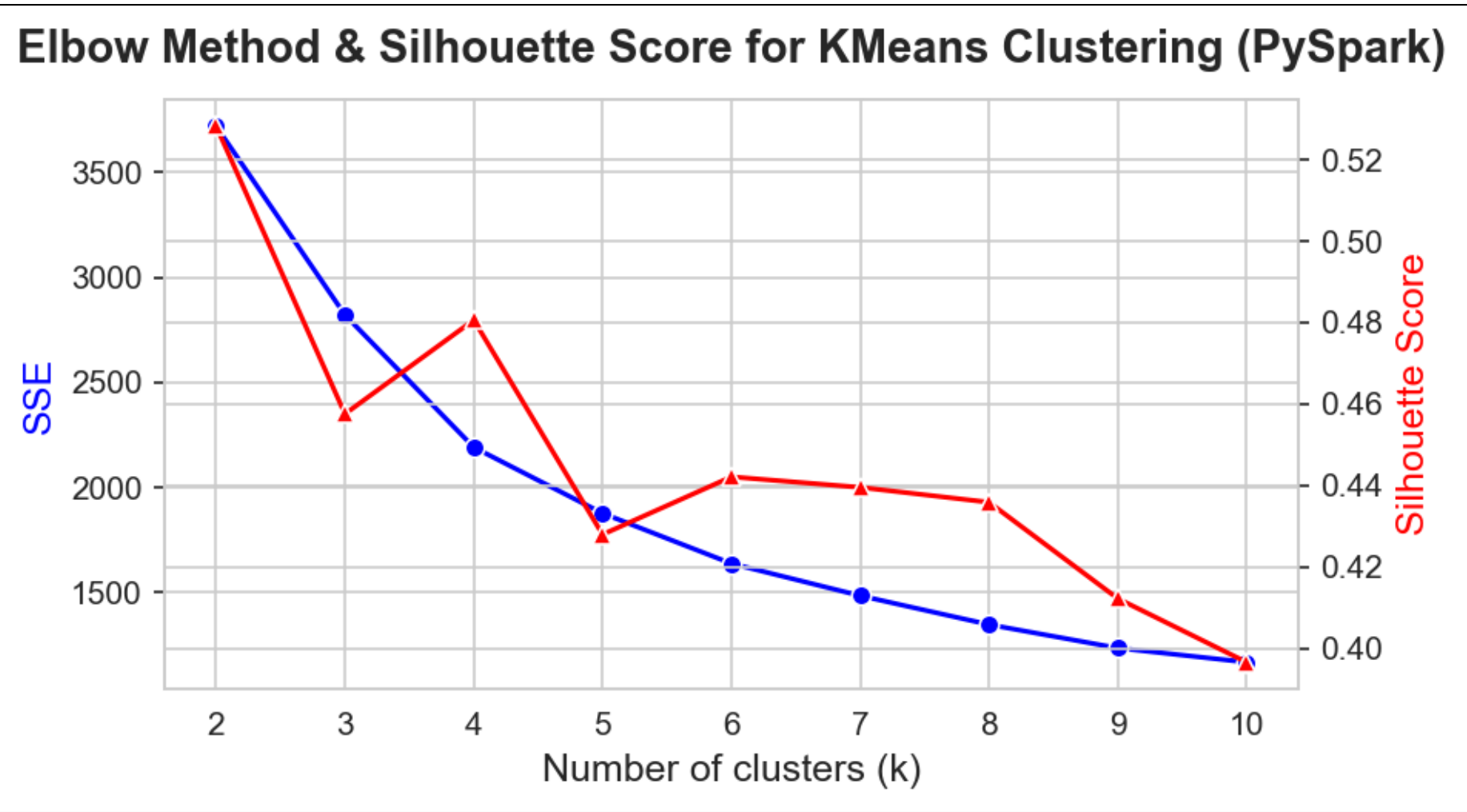
# 10. Visualization: Hierarchical clustering (scikit-learn)



Customer Segmentation by Hierarchical Clustering (Recency, Frequency, Monetary)



# 11. Visualization: K-means (spark)



Phương pháp	Nhận xét
Elbow	Đường cong gãy rõ ở k=4, sau đó giảm chậm, chọn k=4.
Silhouette Score	Cao nhất ở k=2 nhưng quá ít cụm. Đỉnh tiếp theo ở k=4.

kết hợp 2 phương pháp, chọn **k = 4**

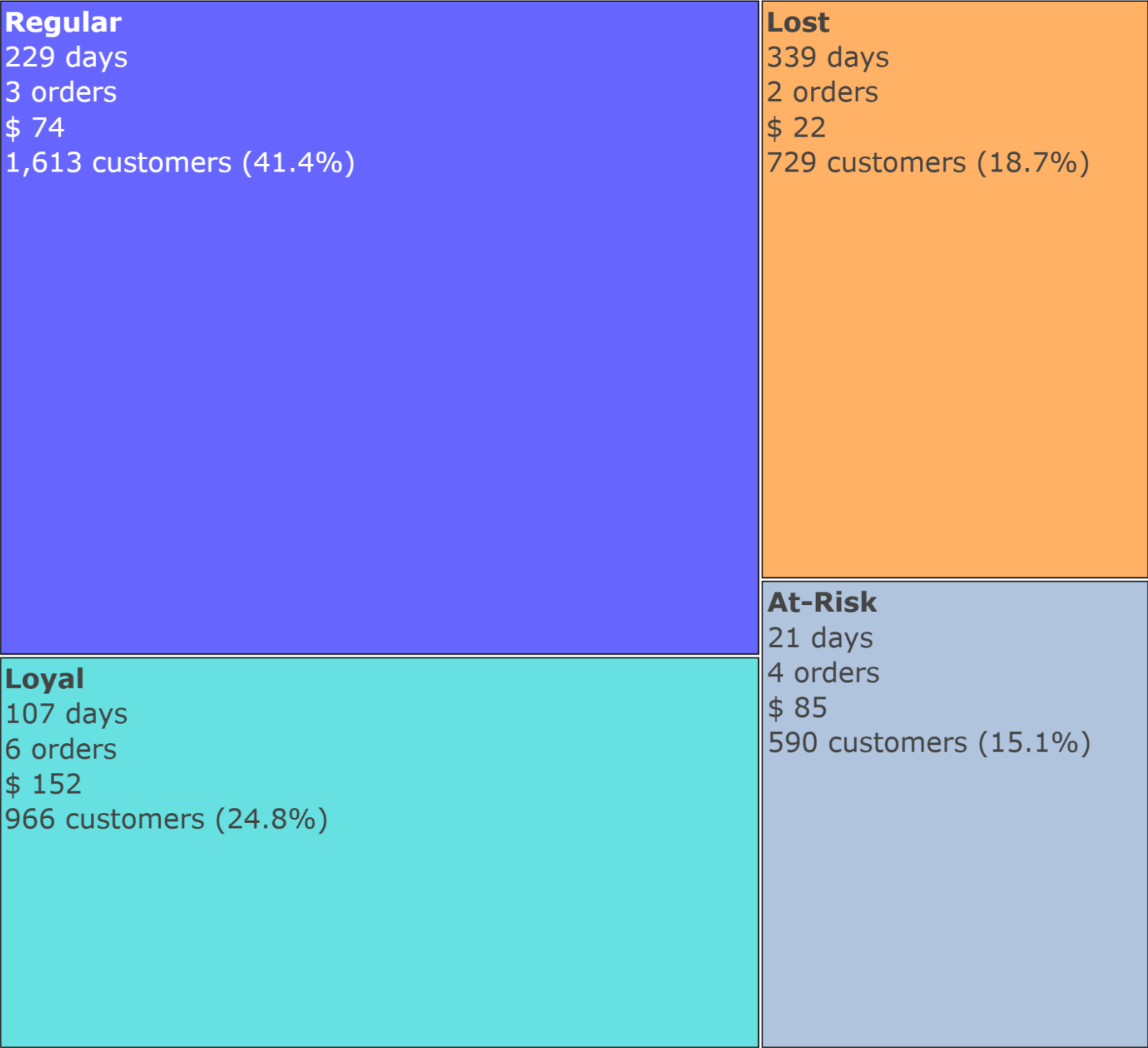


# 11. Visualization: K-means (spark)

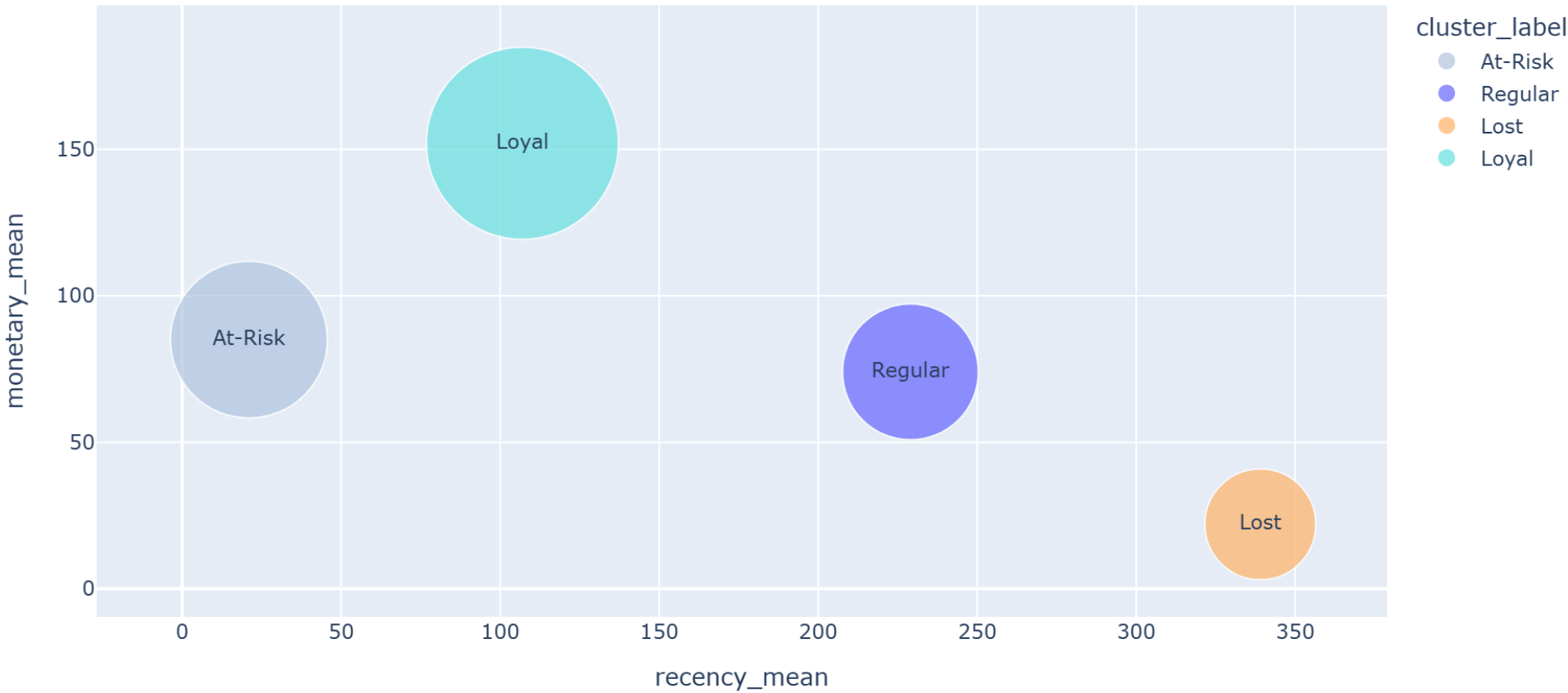
Cluster	Segment	Customer count	Customer %	Revenue %	Recency mean	Frequency mean	Monetary mean	Nhận xét chính
3	Loyal	966	24.8%	44.1%	107	6	152	Nhóm khách hàng trung thành: mua thường xuyên, chi tiêu cao, đóng góp doanh thu lớn nhất.
1	Regular	1613	41.4%	35.9%	229	3	74	Nhóm đông nhất: mua ở mức ổn định, chi tiêu trung bình khá, đóng góp doanh thu cao thứ hai.
0	At-Risk	590	15.1%	15.1%	21	4	85	Khách từng mua gần đây, tần suất vừa phải, chi tiêu trung bình. Tuy nhiên tỷ trọng nhỏ, cần theo dõi.
2	Lost	729	18.7%	4.9%	339	2	22	Nhóm khách hàng gần như mất: lâu không quay lại, mua rất ít, chi tiêu thấp, đóng góp doanh thu không đáng kể.

# 11. Visualization: K-means (spark)

Customer Segmentation Distribution (Treemap)



Customer Segmentation by KMeans (Recency, Frequency, Monetary)

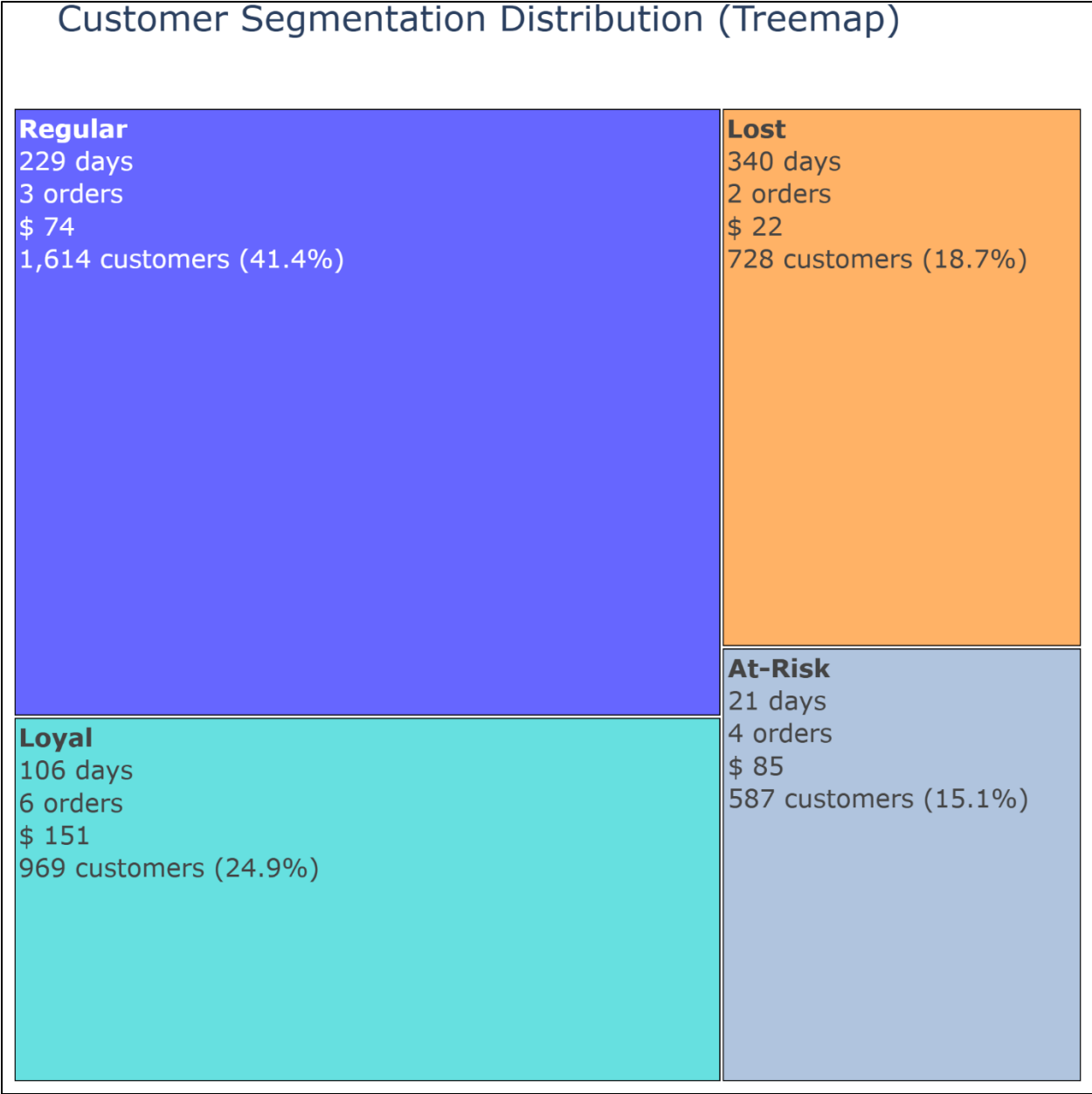


# 12. Differences in results between 3 models

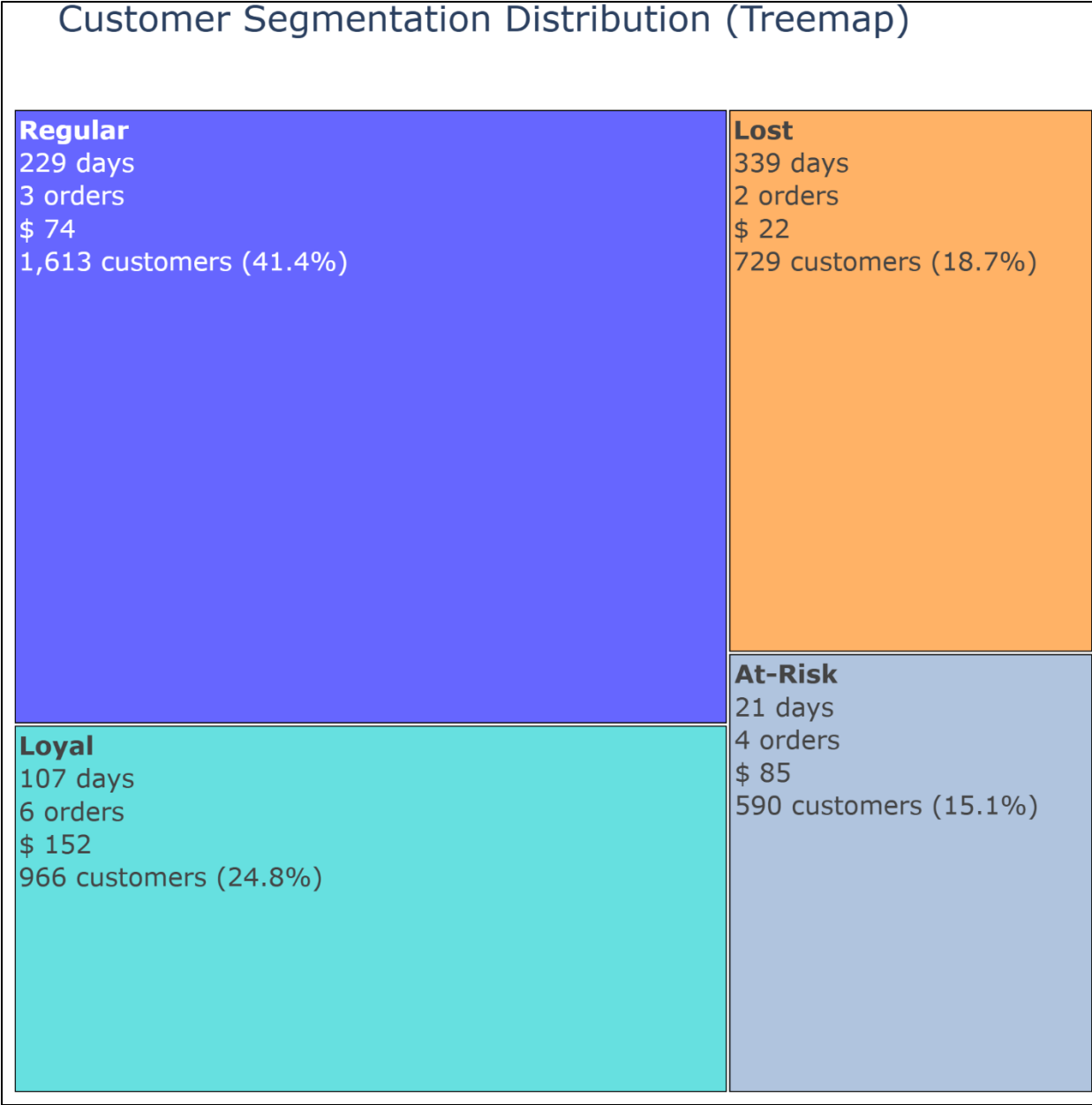
K-Means (scikit-leans)  K-Means (spark)	<table><tr><th>Cluster</th><th>Segment</th><th>Customer count</th><th>Customer %</th><th>Revenue %</th><th>Recency mean</th><th>Frequency mean</th><th>Monetary mean</th><th>Nhận xét chính</th></tr><tr><td>2</td><td>Loyal</td><td>969</td><td>24.9%</td><td>44.2%</td><td>106</td><td>6</td><td>151</td><td>Nhóm khách hàng trung thành: mua khá thường xuyên, chi tiêu cao, đóng góp doanh thu lớn nhất.</td></tr><tr><td>1</td><td>Regular</td><td>1614</td><td>41.4%</td><td>36.0%</td><td>229</td><td>3</td><td>74</td><td>Nhóm đông nhất: mua ở mức ổn định, chi tiêu trung bình khá, đóng góp doanh thu cao thứ hai.</td></tr><tr><td>0</td><td>At-Risk</td><td>587</td><td>15.1%</td><td>14.9%</td><td>21</td><td>4</td><td>85</td><td>Khách từng mua gần đây, tần suất vừa phải, chi tiêu trung bình. Tuy nhiên tỷ trọng nhỏ, cần theo dõi.</td></tr><tr><td>3</td><td>Lost</td><td>728</td><td>18.7%</td><td>4.9%</td><td>340</td><td>2</td><td>22</td><td>Nhóm khách hàng gần như mất: lâu không quay lại, mua rất ít, chi tiêu thấp, đóng góp doanh thu không đáng kể.</td></tr></table>	Cluster	Segment	Customer count	Customer %	Revenue %	Recency mean	Frequency mean	Monetary mean	Nhận xét chính	2	Loyal	969	24.9%	44.2%	106	6	151	Nhóm khách hàng trung thành: mua khá thường xuyên, chi tiêu cao, đóng góp doanh thu lớn nhất.	1	Regular	1614	41.4%	36.0%	229	3	74	Nhóm đông nhất: mua ở mức ổn định, chi tiêu trung bình khá, đóng góp doanh thu cao thứ hai.	0	At-Risk	587	15.1%	14.9%	21	4	85	Khách từng mua gần đây, tần suất vừa phải, chi tiêu trung bình. Tuy nhiên tỷ trọng nhỏ, cần theo dõi.	3	Lost	728	18.7%	4.9%	340	2	22	Nhóm khách hàng gần như mất: lâu không quay lại, mua rất ít, chi tiêu thấp, đóng góp doanh thu không đáng kể.	<b>Kết quả K-Means:</b> - K-Means với scikit-learn - K-Means với spark Kết quả phân cụm gần như giống nhau, sai khác rất ít, thể hiện rõ đặc trưng của các cụm.
	Cluster	Segment	Customer count	Customer %	Revenue %	Recency mean	Frequency mean	Monetary mean	Nhận xét chính																																						
2	Loyal	969	24.9%	44.2%	106	6	151	Nhóm khách hàng trung thành: mua khá thường xuyên, chi tiêu cao, đóng góp doanh thu lớn nhất.																																							
1	Regular	1614	41.4%	36.0%	229	3	74	Nhóm đông nhất: mua ở mức ổn định, chi tiêu trung bình khá, đóng góp doanh thu cao thứ hai.																																							
0	At-Risk	587	15.1%	14.9%	21	4	85	Khách từng mua gần đây, tần suất vừa phải, chi tiêu trung bình. Tuy nhiên tỷ trọng nhỏ, cần theo dõi.																																							
3	Lost	728	18.7%	4.9%	340	2	22	Nhóm khách hàng gần như mất: lâu không quay lại, mua rất ít, chi tiêu thấp, đóng góp doanh thu không đáng kể.																																							
<table><tr><th>Cluster</th><th>Segment</th><th>Customer count</th><th>Customer %</th><th>Revenue %</th><th>Recency mean</th><th>Frequency mean</th><th>Monetary mean</th><th>Nhận xét chính</th></tr><tr><td>3</td><td>Loyal</td><td>966</td><td>24.8%</td><td>44.1%</td><td>107</td><td>6</td><td>152</td><td>Nhóm khách hàng trung thành: mua thường xuyên, chi tiêu cao, đóng góp doanh thu lớn nhất.</td></tr><tr><td>1</td><td>Regular</td><td>1613</td><td>41.4%</td><td>35.9%</td><td>229</td><td>3</td><td>74</td><td>Nhóm đông nhất: mua ở mức ổn định, chi tiêu trung bình khá, đóng góp doanh thu cao thứ hai.</td></tr><tr><td>0</td><td>At-Risk</td><td>590</td><td>15.1%</td><td>15.1%</td><td>21</td><td>4</td><td>85</td><td>Khách từng mua gần đây, tần suất vừa phải, chi tiêu trung bình. Tuy nhiên tỷ trọng nhỏ, cần theo dõi.</td></tr><tr><td>2</td><td>Lost</td><td>729</td><td>18.7%</td><td>4.9%</td><td>339</td><td>2</td><td>22</td><td>Nhóm khách hàng gần như mất: lâu không quay lại, mua rất ít, chi tiêu thấp, đóng góp doanh thu không đáng kể.</td></tr></table>	Cluster	Segment	Customer count	Customer %	Revenue %	Recency mean	Frequency mean	Monetary mean	Nhận xét chính	3	Loyal	966	24.8%	44.1%	107	6	152	Nhóm khách hàng trung thành: mua thường xuyên, chi tiêu cao, đóng góp doanh thu lớn nhất.	1	Regular	1613	41.4%	35.9%	229	3	74	Nhóm đông nhất: mua ở mức ổn định, chi tiêu trung bình khá, đóng góp doanh thu cao thứ hai.	0	At-Risk	590	15.1%	15.1%	21	4	85	Khách từng mua gần đây, tần suất vừa phải, chi tiêu trung bình. Tuy nhiên tỷ trọng nhỏ, cần theo dõi.	2	Lost	729	18.7%	4.9%	339	2	22	Nhóm khách hàng gần như mất: lâu không quay lại, mua rất ít, chi tiêu thấp, đóng góp doanh thu không đáng kể.		
Cluster	Segment	Customer count	Customer %	Revenue %	Recency mean	Frequency mean	Monetary mean	Nhận xét chính																																							
3	Loyal	966	24.8%	44.1%	107	6	152	Nhóm khách hàng trung thành: mua thường xuyên, chi tiêu cao, đóng góp doanh thu lớn nhất.																																							
1	Regular	1613	41.4%	35.9%	229	3	74	Nhóm đông nhất: mua ở mức ổn định, chi tiêu trung bình khá, đóng góp doanh thu cao thứ hai.																																							
0	At-Risk	590	15.1%	15.1%	21	4	85	Khách từng mua gần đây, tần suất vừa phải, chi tiêu trung bình. Tuy nhiên tỷ trọng nhỏ, cần theo dõi.																																							
2	Lost	729	18.7%	4.9%	339	2	22	Nhóm khách hàng gần như mất: lâu không quay lại, mua rất ít, chi tiêu thấp, đóng góp doanh thu không đáng kể.																																							
Hierarchical clustering	<table><tr><th>Cluster</th><th>Segment</th><th>Customer count</th><th>Customer %</th><th>Revenue %</th><th>Recency mean</th><th>Frequency mean</th><th>Monetary mean</th><th>Nhận xét chính</th></tr><tr><td>3</td><td>Loyal</td><td>649</td><td>16.6%</td><td>27.7%</td><td>134</td><td>6</td><td>142</td><td>Nhóm khách hàng trung thành: mua khá thường xuyên, chi tiêu cao.</td></tr><tr><td>2</td><td>Regular</td><td>1442</td><td>37.0%</td><td>36.6%</td><td>220</td><td>3</td><td>84</td><td>Nhóm đông nhất: mua ở mức ổn định, chi tiêu trung bình khá.</td></tr><tr><td>0</td><td>At-Risk</td><td>948</td><td>24.3%</td><td>28.8%</td><td>29</td><td>5</td><td>101</td><td>Khách từng mua gần đây, tần suất vừa phải, chi tiêu trung bình. Tuy nhiên tỷ trọng nhỏ, cần theo dõi.</td></tr><tr><td>1</td><td>Lost</td><td>859</td><td>22.0%</td><td>6.9%</td><td>349</td><td>2</td><td>27</td><td>Nhóm khách hàng gần như mất: lâu không quay lại, mua rất ít, chi tiêu thấp, đóng góp doanh thu không đáng kể.</td></tr></table>	Cluster	Segment	Customer count	Customer %	Revenue %	Recency mean	Frequency mean	Monetary mean	Nhận xét chính	3	Loyal	649	16.6%	27.7%	134	6	142	Nhóm khách hàng trung thành: mua khá thường xuyên, chi tiêu cao.	2	Regular	1442	37.0%	36.6%	220	3	84	Nhóm đông nhất: mua ở mức ổn định, chi tiêu trung bình khá.	0	At-Risk	948	24.3%	28.8%	29	5	101	Khách từng mua gần đây, tần suất vừa phải, chi tiêu trung bình. Tuy nhiên tỷ trọng nhỏ, cần theo dõi.	1	Lost	859	22.0%	6.9%	349	2	27	Nhóm khách hàng gần như mất: lâu không quay lại, mua rất ít, chi tiêu thấp, đóng góp doanh thu không đáng kể.	<b>Kết quả Hierarchical clustering:</b> Kết quả phân cụm khác với K-Means, <b>đường biên giới không rõ và có lẫn nhóm vào nhau giữa các cụm.</b> - Nhóm At-Risk tăng - Nhóm Regular + Royal giảm
Cluster	Segment	Customer count	Customer %	Revenue %	Recency mean	Frequency mean	Monetary mean	Nhận xét chính																																							
3	Loyal	649	16.6%	27.7%	134	6	142	Nhóm khách hàng trung thành: mua khá thường xuyên, chi tiêu cao.																																							
2	Regular	1442	37.0%	36.6%	220	3	84	Nhóm đông nhất: mua ở mức ổn định, chi tiêu trung bình khá.																																							
0	At-Risk	948	24.3%	28.8%	29	5	101	Khách từng mua gần đây, tần suất vừa phải, chi tiêu trung bình. Tuy nhiên tỷ trọng nhỏ, cần theo dõi.																																							
1	Lost	859	22.0%	6.9%	349	2	27	Nhóm khách hàng gần như mất: lâu không quay lại, mua rất ít, chi tiêu thấp, đóng góp doanh thu không đáng kể.																																							

# 12. Differences in results between 3 models

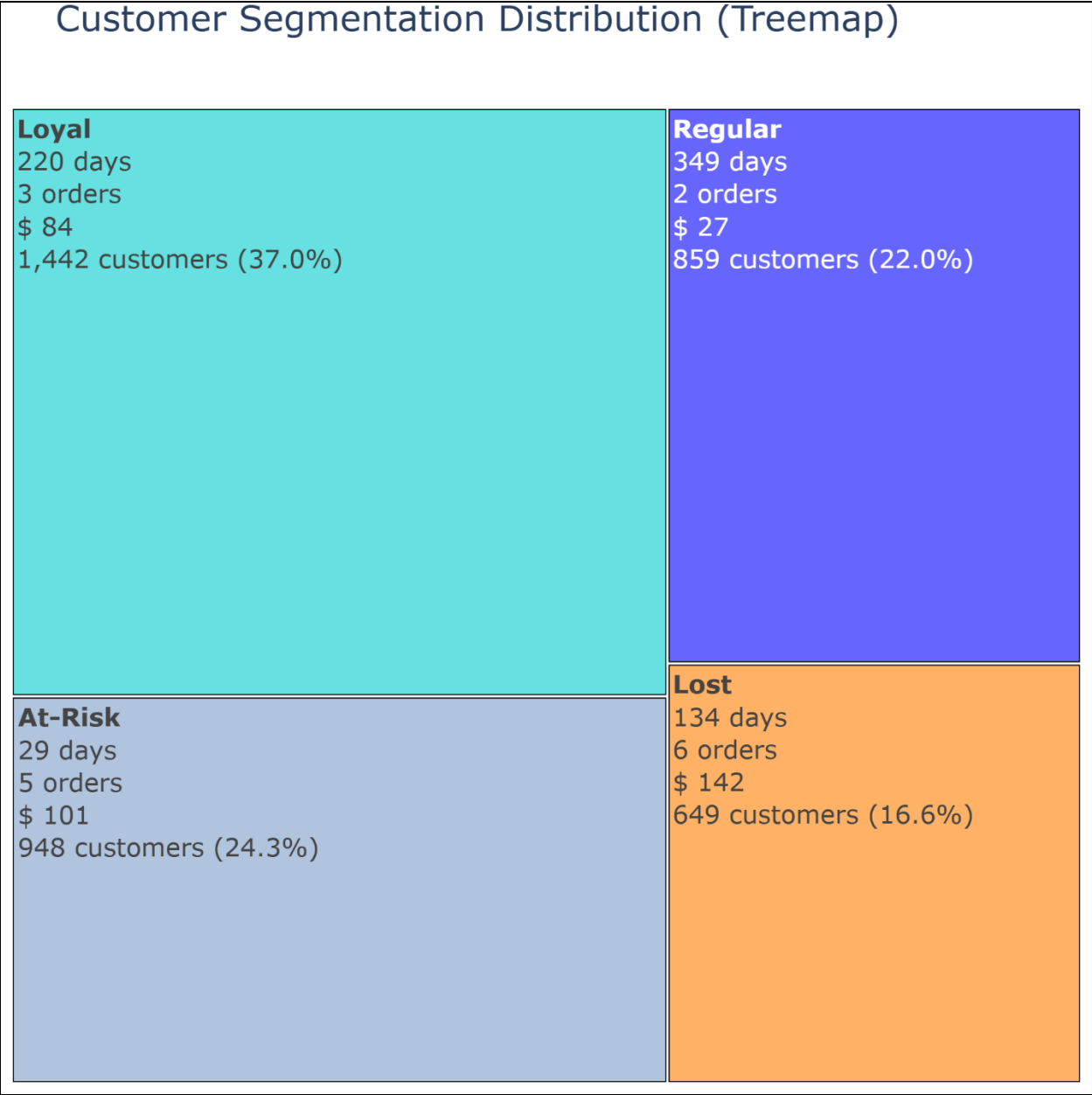
K-Means (scikit-leans)



K-Means (spark)



Hierachical clustering



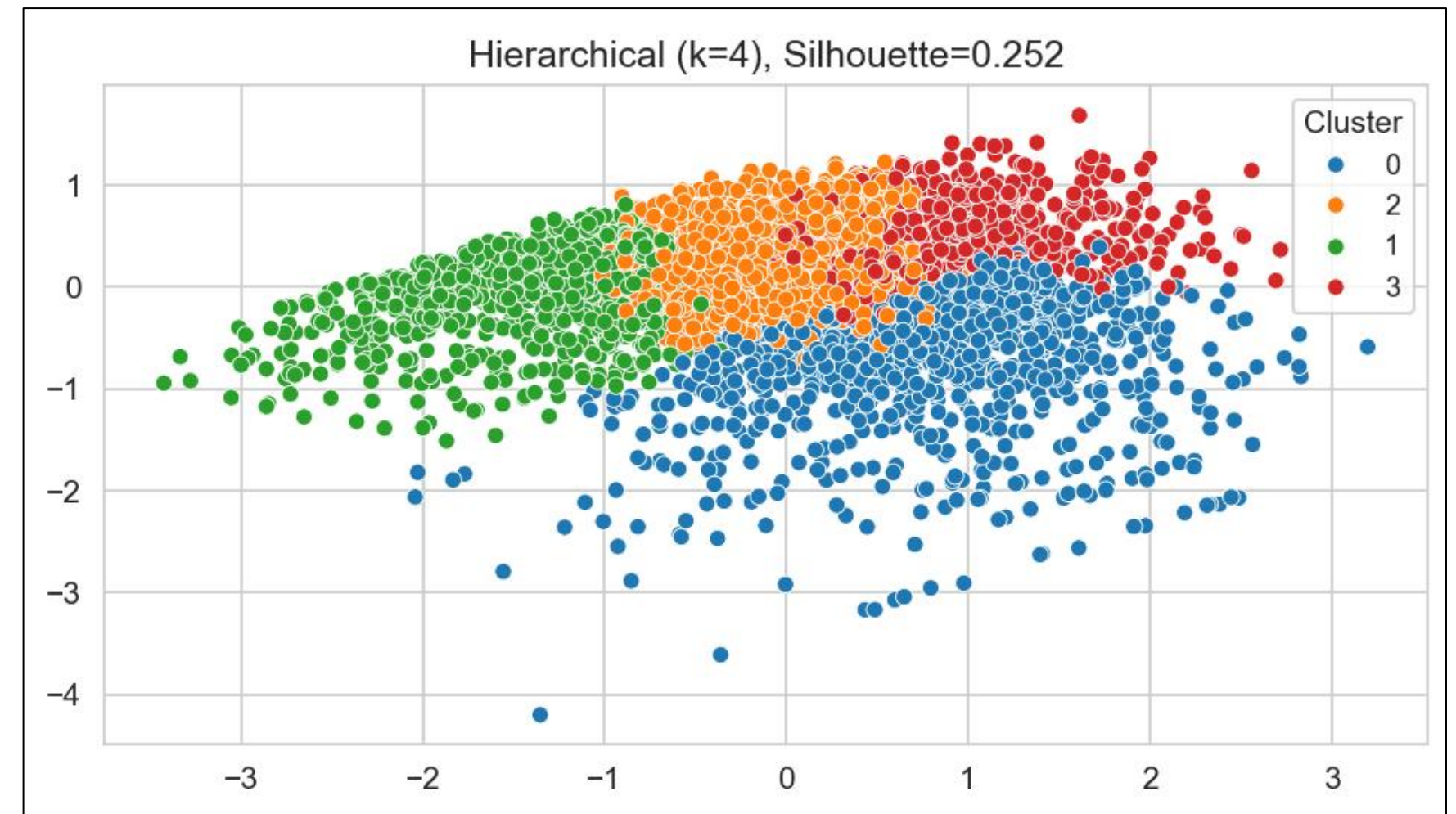
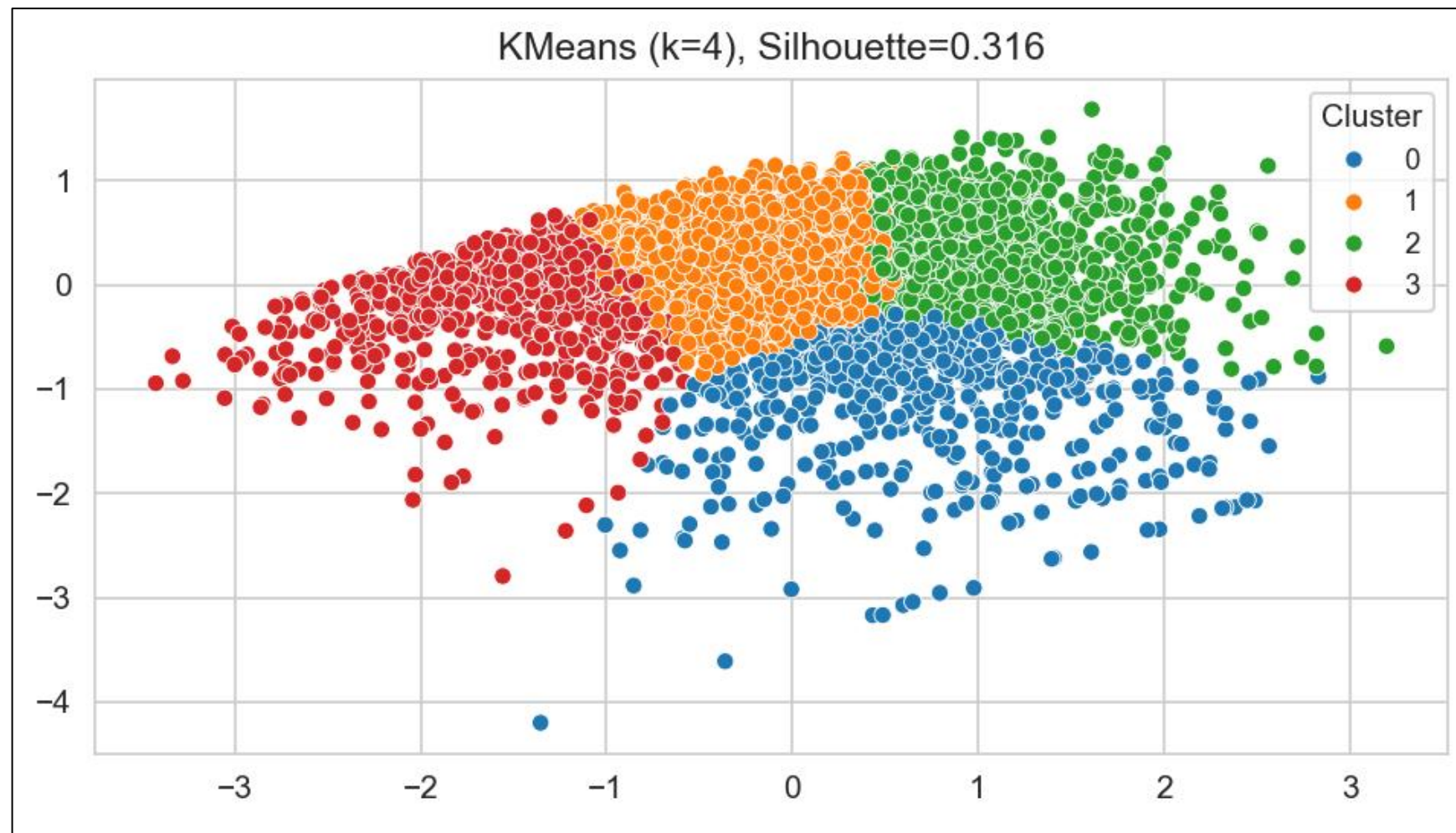


# 12. Differences in results between 3 models

K-Means (scikit-learn)

K-Means (spark)

Hierarchical  
clustering





# 13. Recommendations

Segment	Action	Mục tiêu	Measure
Loyal	Chương trình VIP, ưu đãi độc quyền, khuyến khích giới thiệu bạn bè; gói sản phẩm cá nhân hóa	Tăng tần suất mua sắm, gia tăng giá trị đơn hàng	Doanh thu tăng 10-20%
Regular	Email/SMS cá nhân hóa về xu hướng, flash sale, khuyến khích mua lặp lại	Chuyển đổi sang nhóm trung thành	Theo dõi mức tăng tần suất mua
At-Risk	Ưu đãi chào mừng, khảo sát phản hồi, nhắc nhở bỏ giỏ hàng tự động	Ngăn ngừa rời bỏ, xây dựng thói quen sử dụng	Tỷ lệ phản hồi & giảm tỷ lệ churn
Lost	Chiến dịch "We Miss You", ưu đãi giới hạn thời gian, giảm giá tái kích hoạt	Tái kích hoạt 10-15% khách hàng qua kênh chi phí thấp	Tỷ lệ quay lại; giảm ưu tiên nếu không phản hồi sau 2 lần

# OUR TEAM

Trần Đình Hùng  
Phạm Ngọc Trọng

