TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
TRUNG TÂM TIN HỌC

ĐỒ ÁN TỐT NGHIỆP DATA SCIENCE

PROJECT
CUSTOMER SEGEMENTATION

Nguyễn Nhật Tố Trân vs Nguyễn Vũ Mai Phương

TABLE OF CONTENTS

01

BUSINESS UNDERSTANDING

BUSINESS UNDERSTANDING

Problem

- Một cửa hàng tạp hóa bán sản phẩm thiết yếu và khách hàng là người mua lẻ
- Cần tăng doanh thu, cải thiện mức độ hài lòng của khách hàng

Solution

- Xây dựng hệ thống phân cụm khách hàng
- Đề xuất phương án tiếp cận phù hợp dựa trên đặc điểm của từng nhóm khách hàng

PROCESS

1

EDA

Tiền xử lý dữ liệu và khám phá insights

2

RFM

Phân nhóm khách hàng dựa trên thông tin mua hàng trước đó

3

SOLUTIONS

Đề xuất giải pháp tiếp cận khách hàng phù hợp

MANUAL SEGMENTATION

UNSUPERVISED MODELING

- KMeans (PySpark)
- GMM
- DBSCAN
- Hierarchical Clustering

02

EDA

```
<class 'pandas.core.frame.DataFrame'>
Index: 38521 entries, 0 to 38764
Data columns (total 8 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   productId        38521 non-null  int64
 1   productName      38521 non-null  object
 2   price            38521 non-null  float64
 3   Category         38521 non-null  object
 4   Member_number    38521 non-null  int64
 5   Date             38521 non-null  datetime64[ns]
 6   items            38521 non-null  int64
 7   purchase_amount  38521 non-null  float64
dtypes: datetime64[ns](1), float64(2), int64(3), object(2)
memory usage: 2.6+ MB
```

Transactions timeframe from 2014-01-01 00:00:00 to 2015-12-30 00:00:00

0 transactions don't have a member number

3,898 unique Member_number

Data Pre-processing

- 38.521 entries (có duplicates, không có missing values, chuyển data type của “Date”)
- 8 features (“purchase_amount” được tạo mới)
- Dữ liệu được ghi nhận trong 2 năm
- 3898 khách hàng

```
df[num_cols].describe()
```

	price	items	purchase_amount
count	38521.000000	38521.000000	38521.000000
mean	4.305367	1.996729	8.590893
std	4.320088	0.817539	9.952770
min	0.300000	1.000000	0.300000
25%	1.250000	1.000000	2.400000
50%	2.500000	2.000000	5.200000
75%	6.100000	3.000000	10.500000
max	28.500000	3.000000	85.500000

```
df[cat_cols].describe()
```

	productName	Category
count	38521	38521
unique	167	11
top	whole milk	Fresh Food
freq	2455	11426

Data Pre-processing

- Phần lớn sản phẩm có giá thấp, và một số ít sản phẩm có giá rất cao
- Số lượng mỗi sản phẩm mỗi khách hàng mua là từ 1-3
- 167 sản phẩm, và “whole milk” được khi nhận nhiều nhất
- 11 ngành hàng, phổ biến nhất là Fresh Food

```
# Total purchases
total_purchases = df['items'].sum()
print(total_purchases)
```

```
76916
```

```
# Total revenue
total_revenue = (df['items'] * df['price']).sum()
print(total_revenue)
```

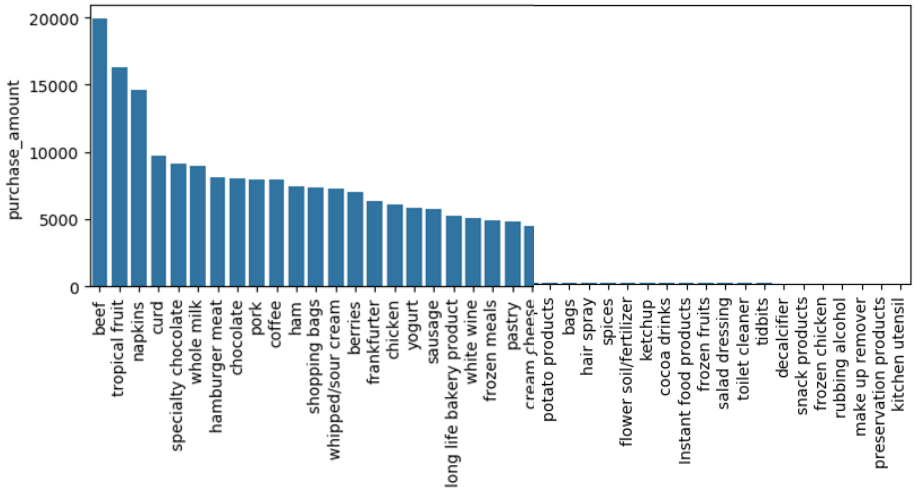
```
330929.80000000005
```

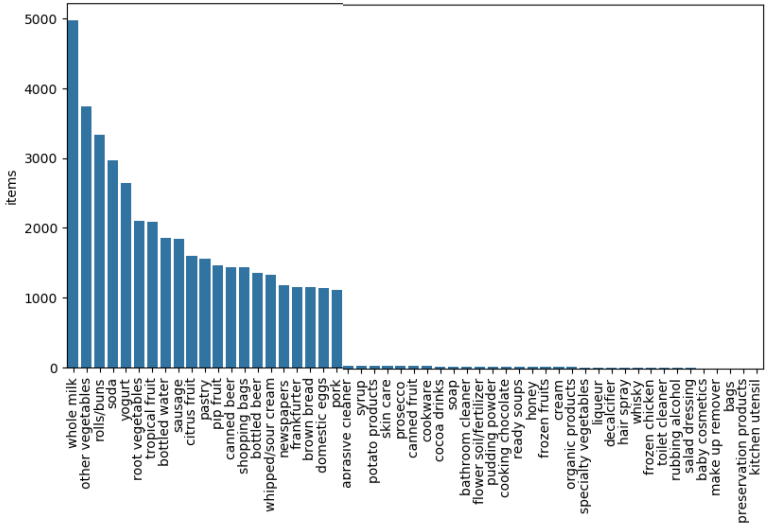
Insights

Trong vòng 2 năm

- Tổng sản phẩm được bán: 77.000
- Tổng doanh thu : 331.000$

EDA





Insights

- Sản phẩm mang lại doanh thu nhiều nhất : Thịt bò, trái cây nhiệt đới, khăn giấy, phô mai tươi, sô cô la đặc sản
- Sản phẩm bán chạy nhất : Sữa tươi nguyên chất, rau củ khác, bánh mì cuộn, soda, sữa chua
- Sản phẩm bán chậm nhất : Gà đông lạnh, cồn sát trùng, nước tẩy trang, sản phẩm bảo quản, dụng cụ nhà bếp

EDA

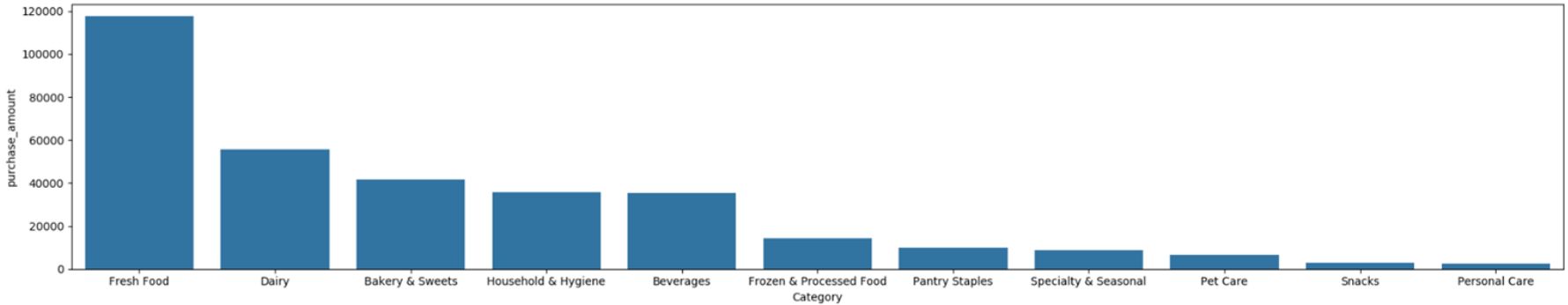
Category	items	purchase_amount	percent_amount
Fresh Food	22786	117566.20	35.526024
Dairy	14968	55748.24	16.845941
Bakery & Sweets	11796	41780.78	12.625270
Household & Hygiene	4339	35555.16	10.744019
Beverages	11963	35391.10	10.694443
Frozen & Processed Food	3689	14371.92	4.342891
Pantry Staples	3134	9857.45	2.978713
Specialty & Seasonal	2360	8864.30	2.678604
Pet Care	641	6425.80	1.941741
Snacks	928	2727.40	0.824163
Personal Care	312	2641.45	0.798190

Insights

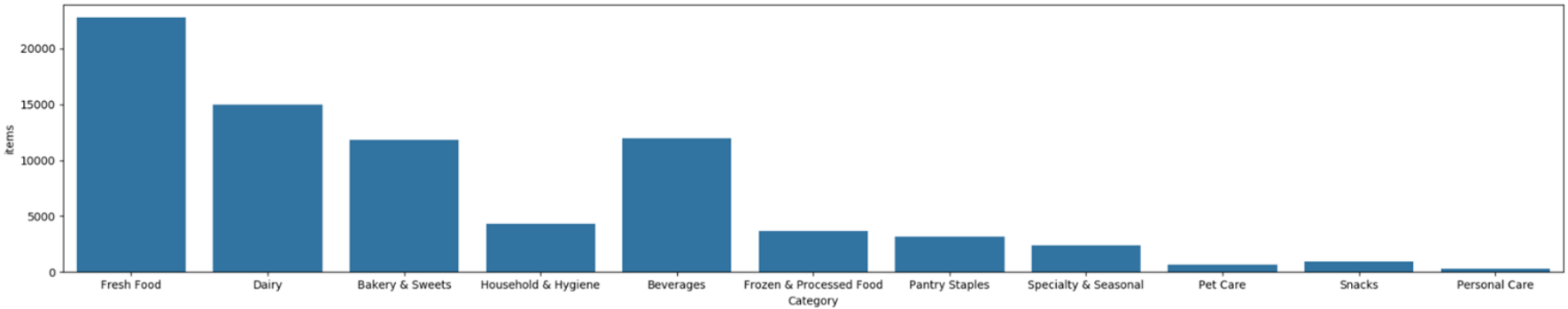
- Fresh Food đóng góp 1/3 doanh thu
- Ngàng hàng bán chạy nhất (60% doanh thu) : Thực phẩm tươi sống, Sản phẩm từ sữa, Bánh mì & đồ ngọt
- Sản phẩm chăm sóc thú cưng, đồ ăn vặt, chăm sóc cá nhân là các ngành hàng bán chậm nhất của cửa hàng

EDA

Biểu đồ doanh thu từng ngành hàng



Biểu đồ sản lượng mua từng ngành hàng



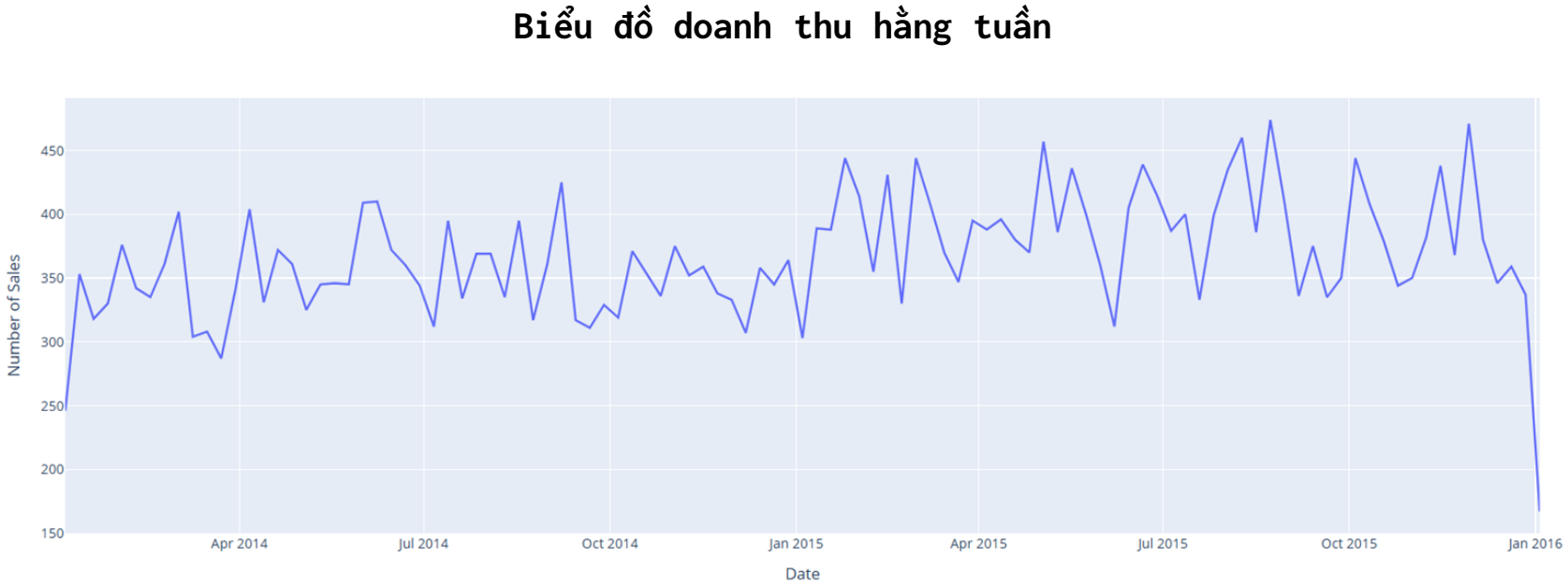
Member_number	items	purchase_amount
2193	63	361.45
1793	52	345.10
3289	63	334.15
2433	57	316.81
2743	41	312.46
...	...	...
1560	5	1.90
1221	2	1.70
4029	2	1.60
1250	2	1.30
4565	1	1.10

Danh sách sản lượng và doanh thu theo từng khách hàng

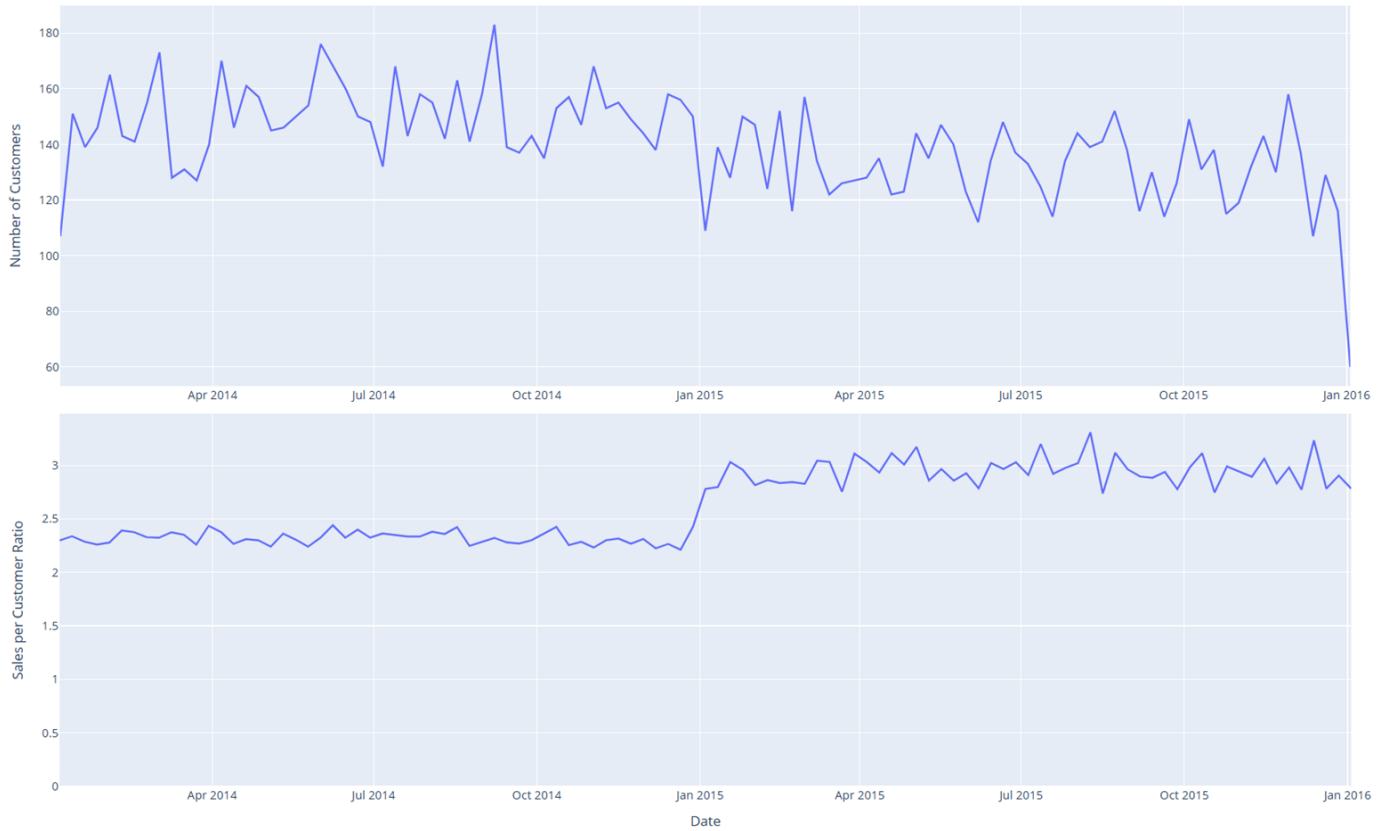
Insights

- Khách hàng chi tiêu nhiều nhất với **63** sản phẩm trong 2 năm tương ứng tổng số tiền là **361$**
- Xác định được khách hàng giá trị cao

EDA


Biểu đồ doanh thu hằng tuần
Number of Sales
450
400
350
300
250
200
150
Apr 2014
Jul 2014
Oct 2014
Jan 2015
Apr 2015
Jul 2015
Oct 2015
Jan 2016
Date

EDA

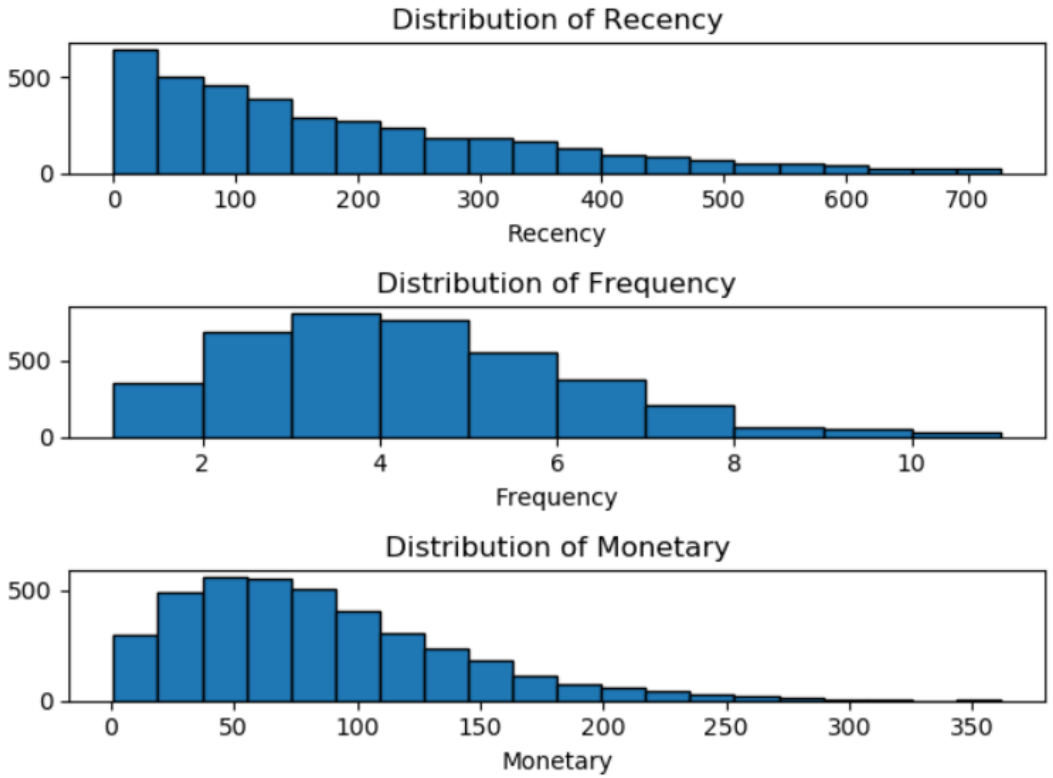


Biểu đồ số lượng và doanh thu mỗi khách hàng hằng tuần

03

RFM

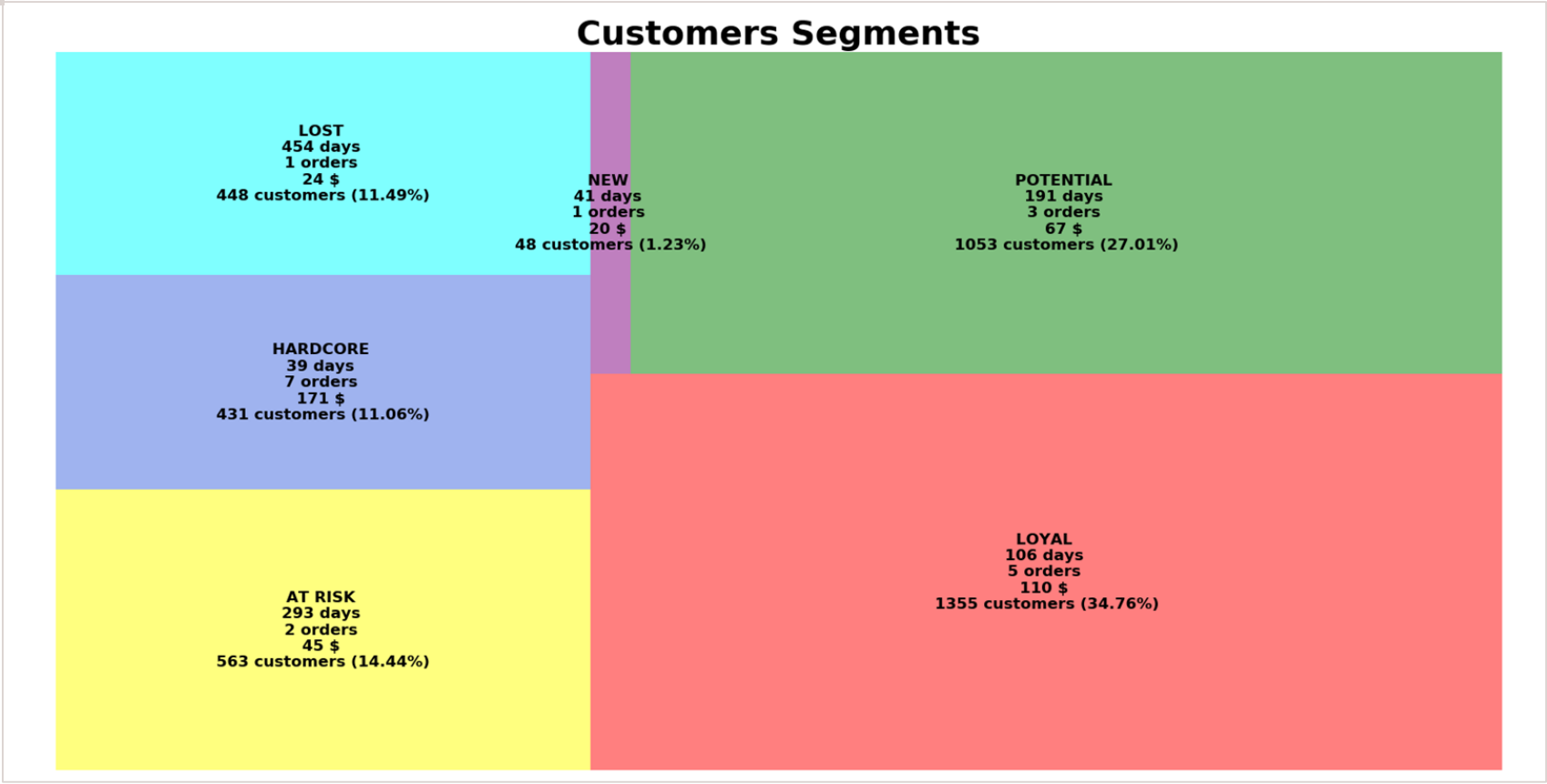
RFM


Distribution of Recency
500
0
0
100
200
300
400
500
600
700
Recency
Distribution of Frequency
500
0
2
4
6
8
10
Frequency
Distribution of Monetary
500
0
0
50
100
150
200
250
300
350
Monetary

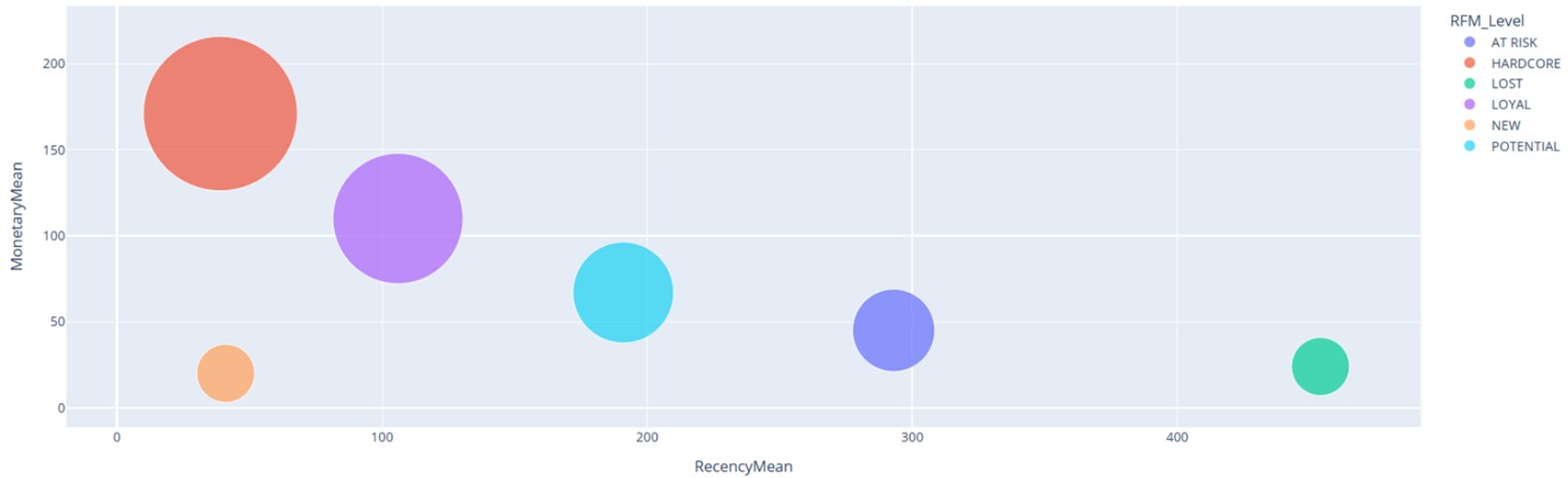
04

MANUAL SEGMENTATION

MANUAL SEGMENTATION



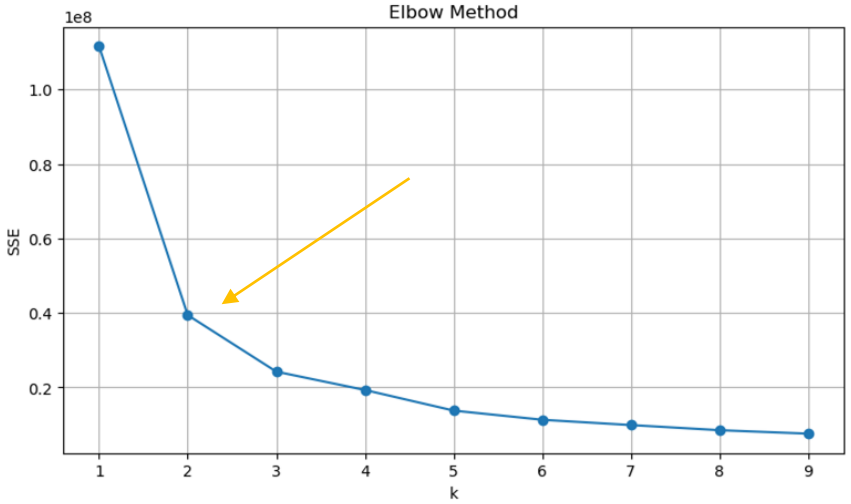
MANUAL SEGMENTATION



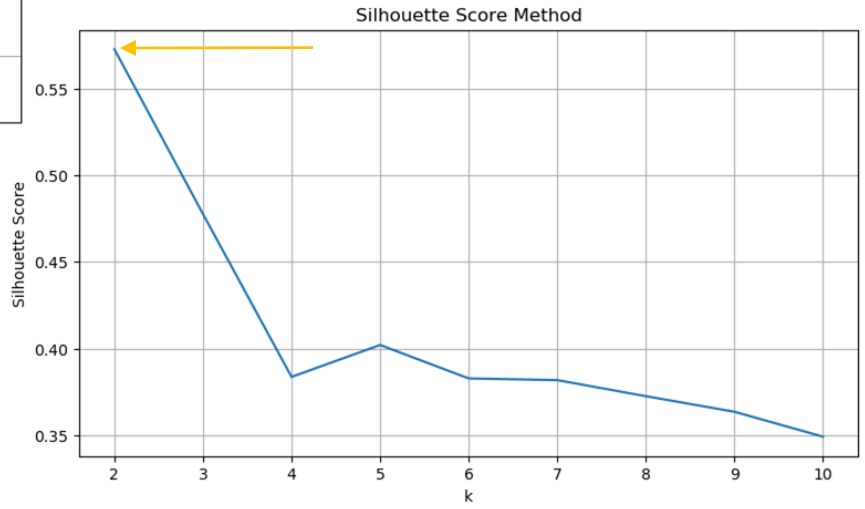
05

UNSUPERVISED MODELING

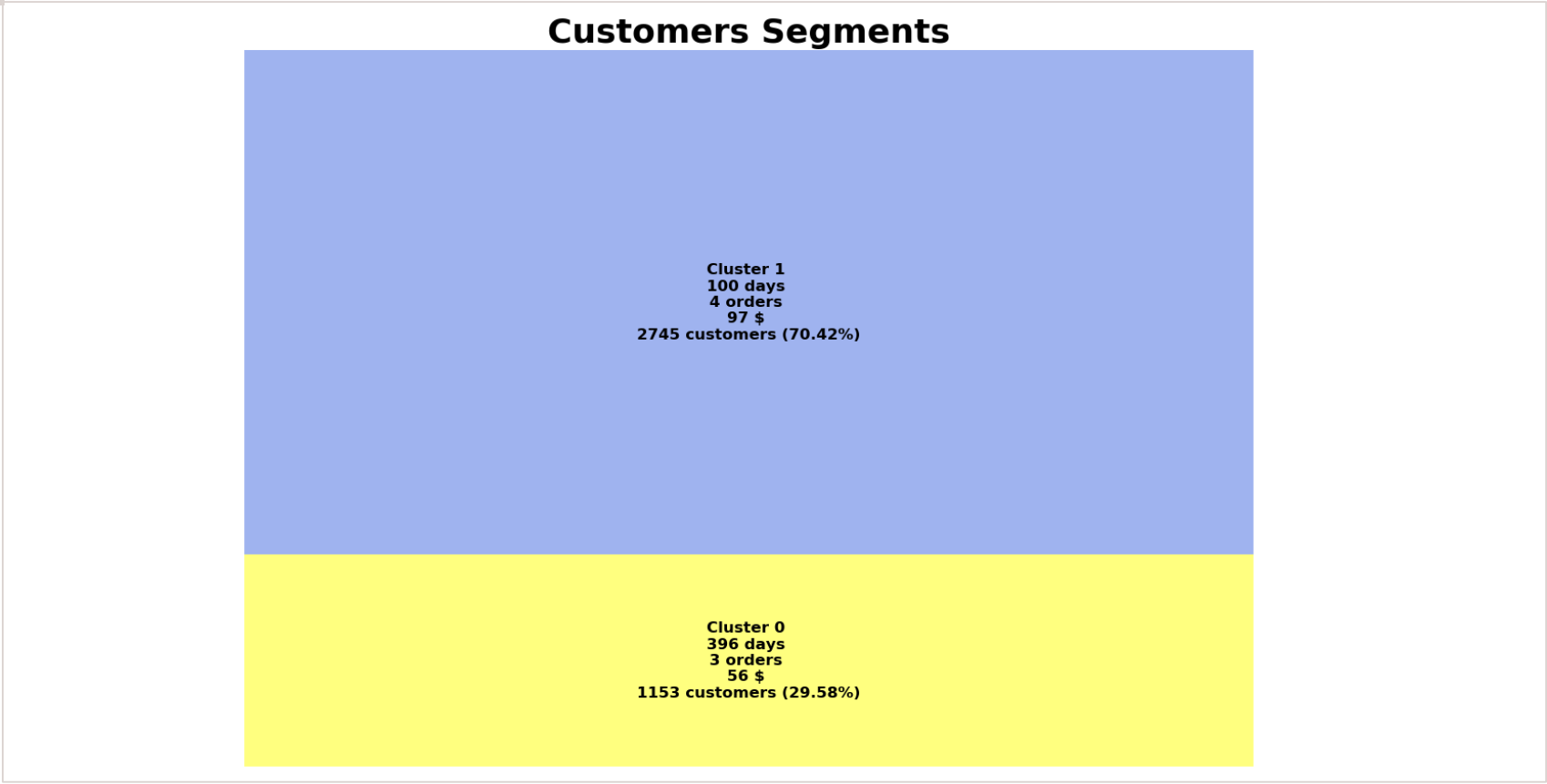
KMEANS


Elbow Method
1e8
SSE
k
1.0
0.8
0.6
0.4
0.2
1
2
3
4
5
6
7
8
9

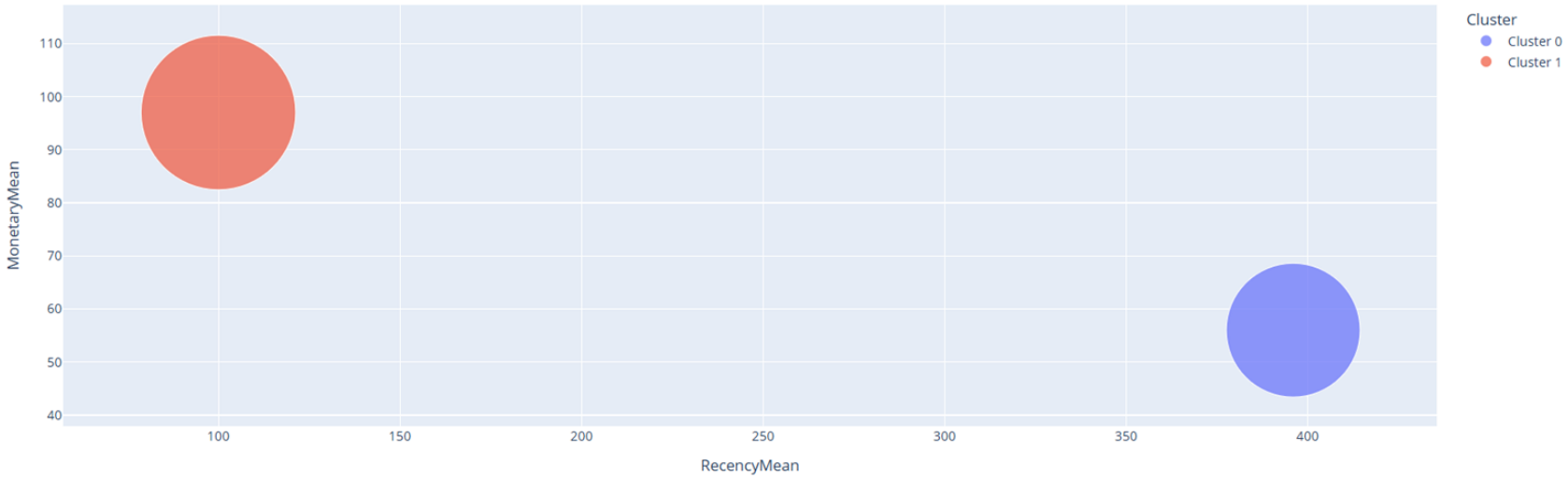
k=2


Silhouette Score Method
Silhouette Score
k
0.55
0.50
0.45
0.40
0.35
2
3
4
5
6
7
8
9
10

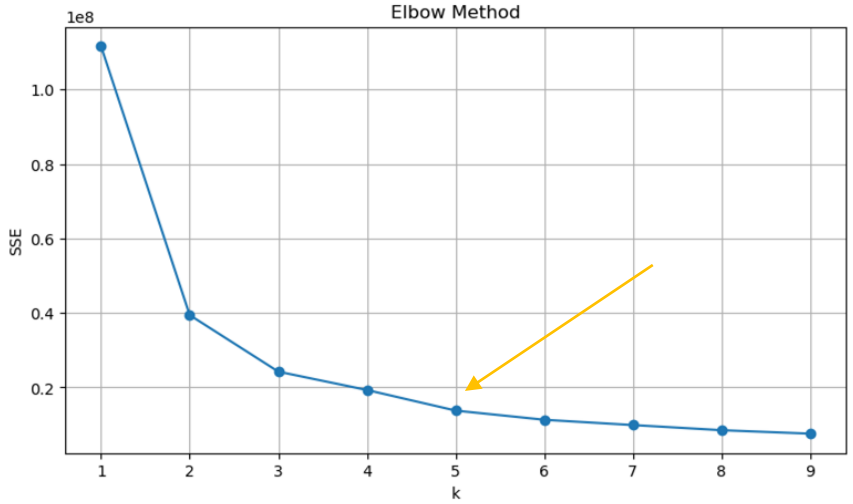
KMEANS


Customers Segments
Cluster 1
100 days
4 orders
97 $
2745 customers (70.42%)
Cluster 0
396 days
3 orders
56 $
1153 customers (29.58%)

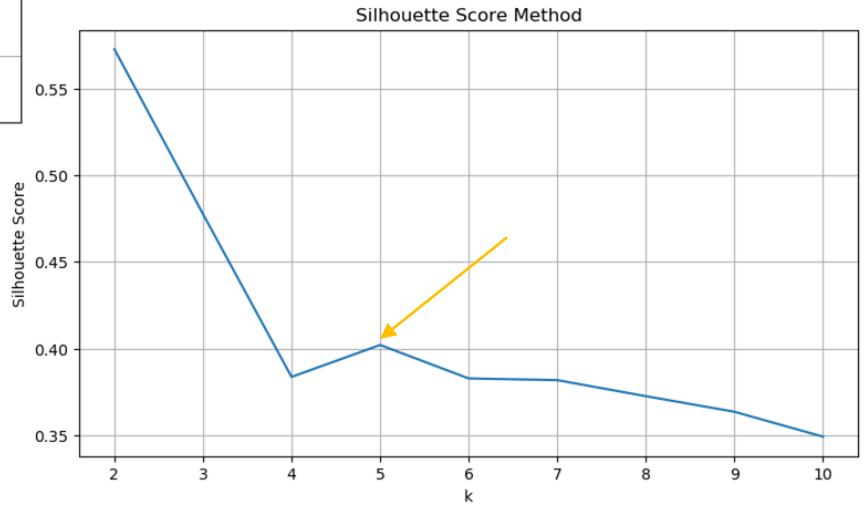
KMEANS


Cluster
Cluster 0
Cluster 1
MonetaryMean
110
100
90
80
70
60
50
40
100
150
200
250
300
350
400
RecencyMean

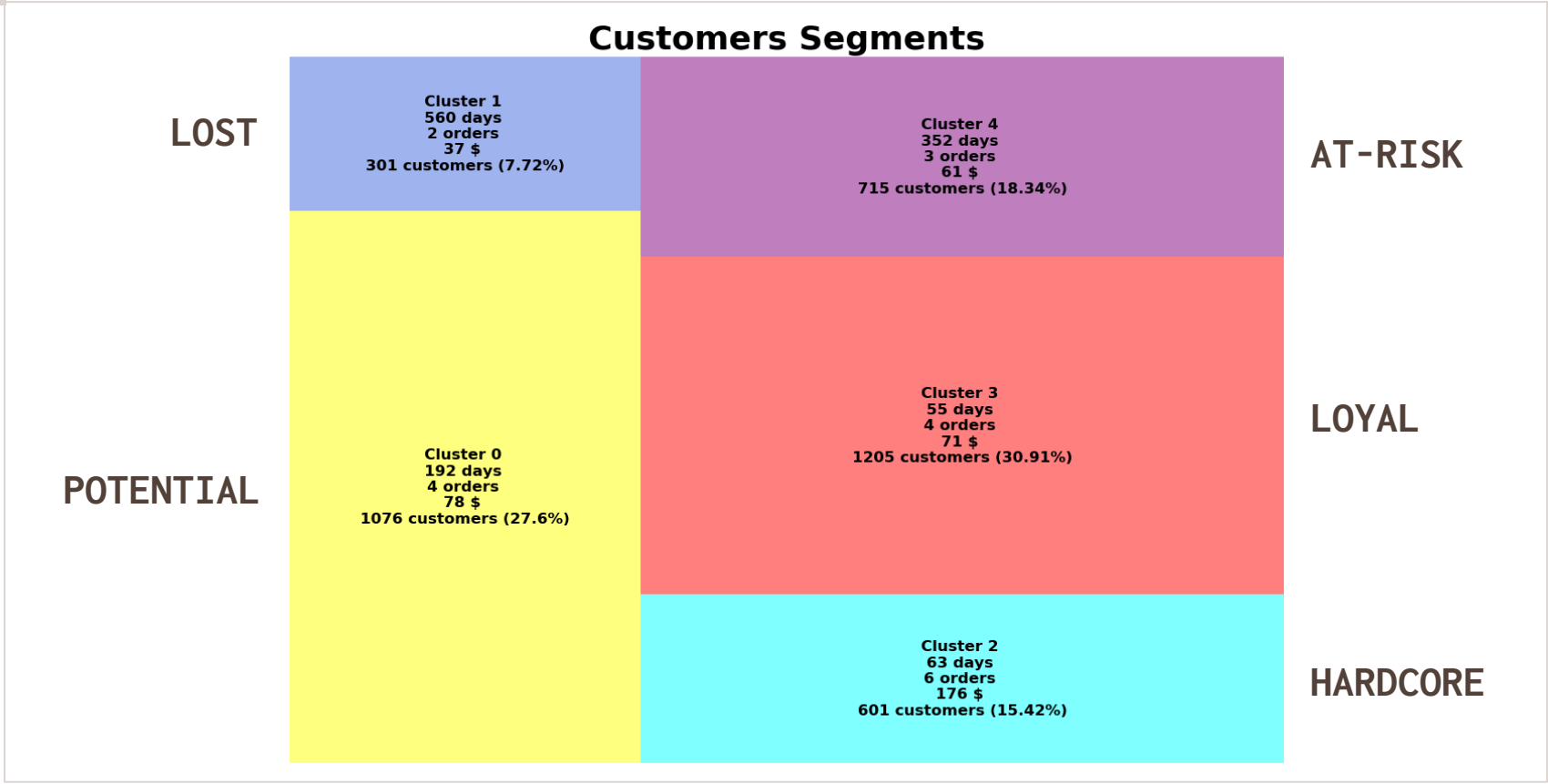
KMEANS


Elbow Method
1e8
SSE
k
1.0
0.8
0.6
0.4
0.2
1
2
3
4
5
6
7
8
9

k=5


Silhouette Score Method
Silhouette Score
k
0.55
0.50
0.45
0.40
0.35
2
3
4
5
6
7
8
9
10

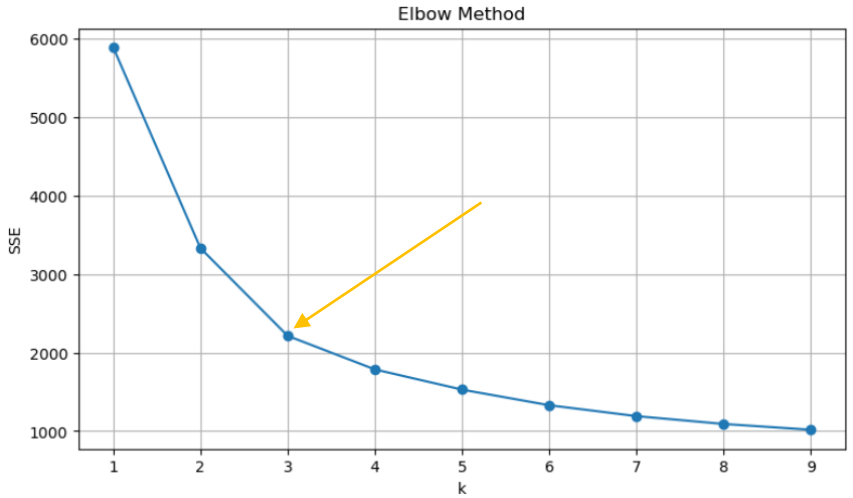
KMEANS


Customers Segments
LOST
POTENTIAL
AT-RISK
LOYAL
HARDCORE
Cluster 1
560 days
2 orders
37 $
301 customers (7.72%)
Cluster 4
352 days
3 orders
61 $
715 customers (18.34%)
Cluster 0
192 days
4 orders
78 $
1076 customers (27.6%)
Cluster 3
55 days
4 orders
71 $
1205 customers (30.91%)
Cluster 2
63 days
6 orders
176 $
601 customers (15.42%)

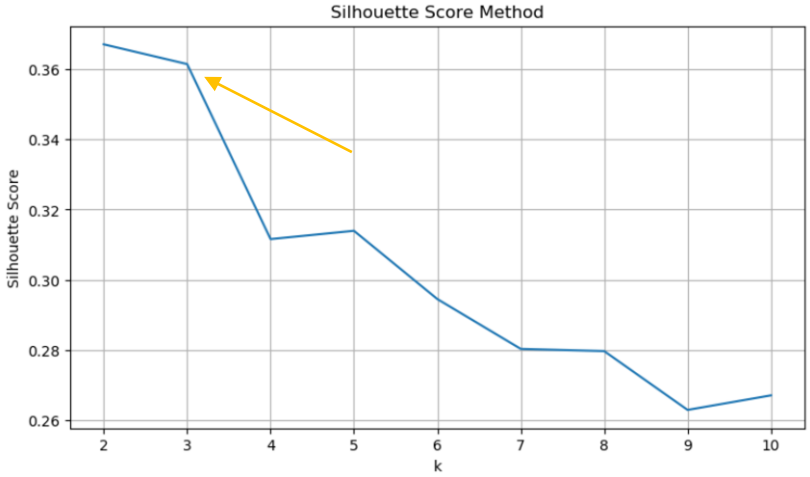
KMEANS


HARDCORE
POTENTIAL
AT-RISK
LOST
LOYAL
MonetaryMean
RecencyMean
0
50
100
150
200
0
100
200
300
400
500
600
Cluster
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4

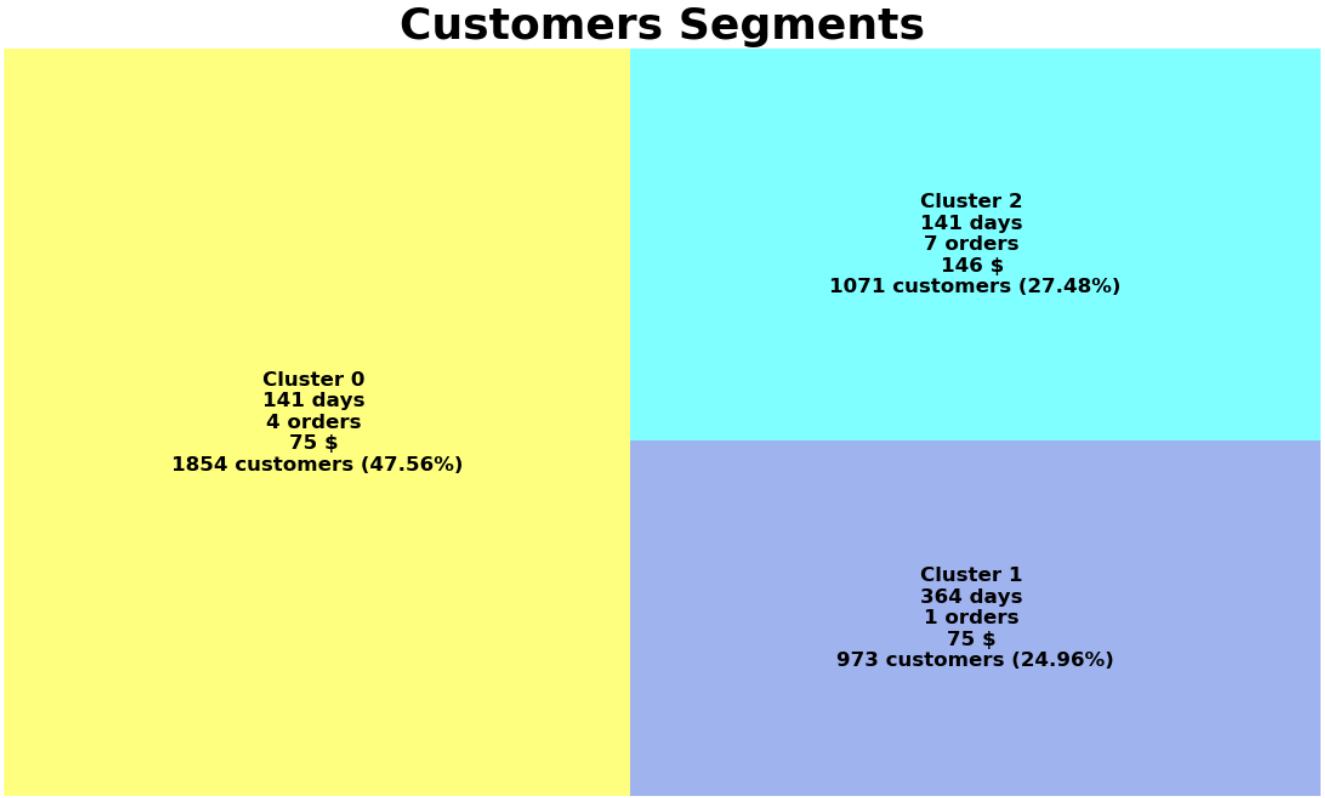
KMEANS (SCALE DATA)



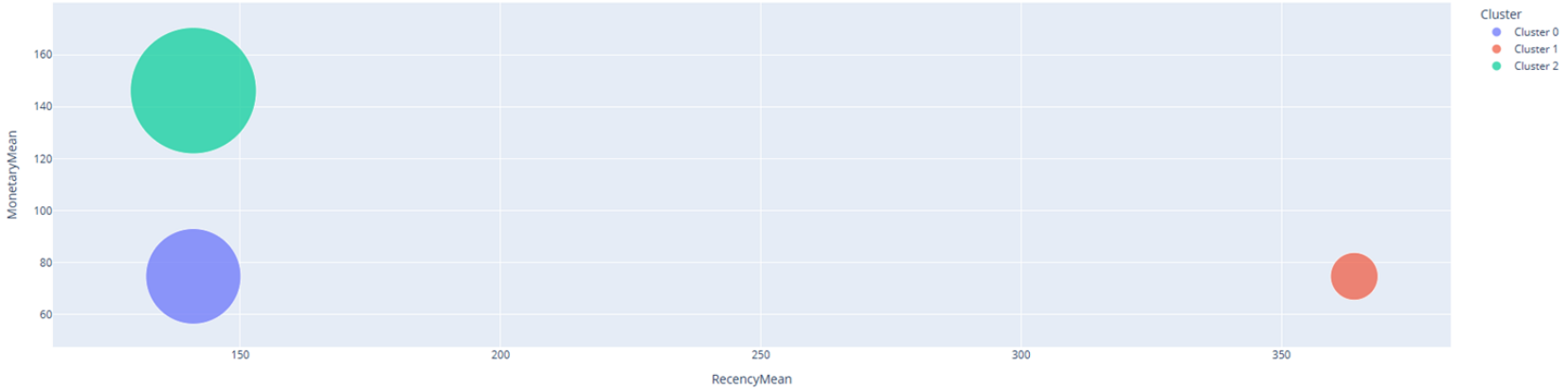
k=3



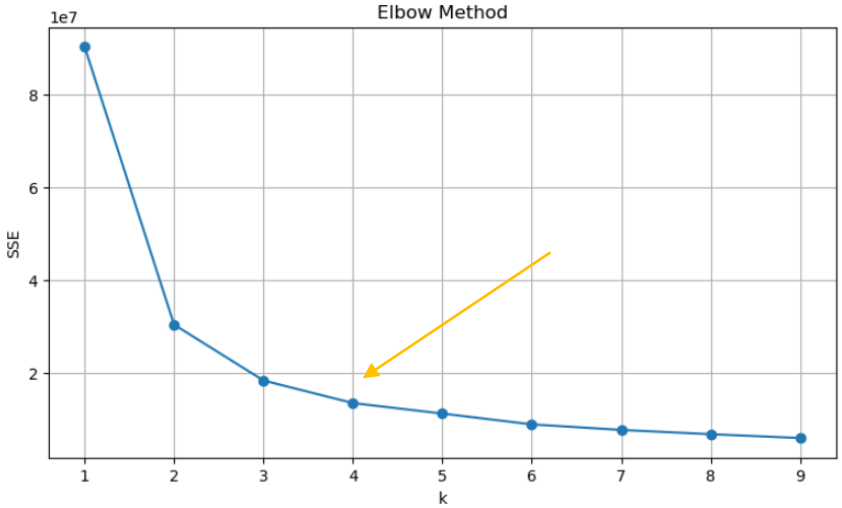
KMEANS (SCALE DATA)



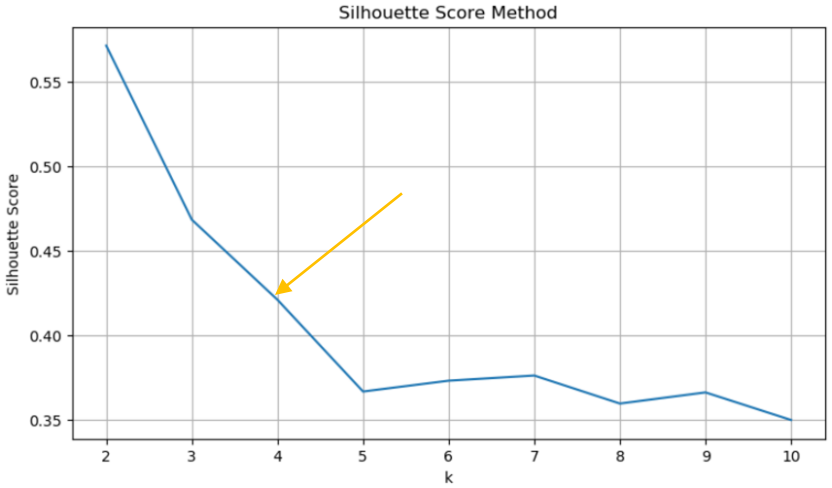
KMEANS (SCALE DATA)



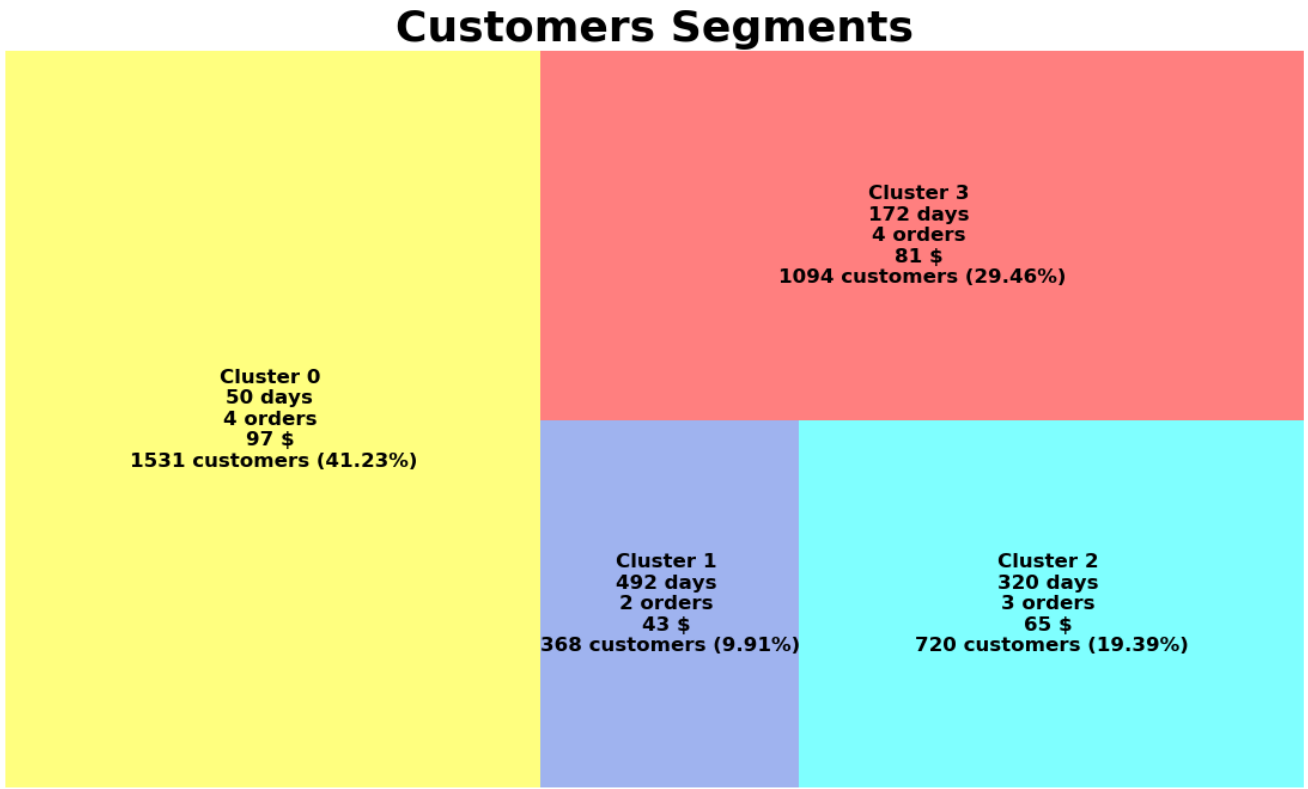
KMEANS (REMOVE OUTLIER)



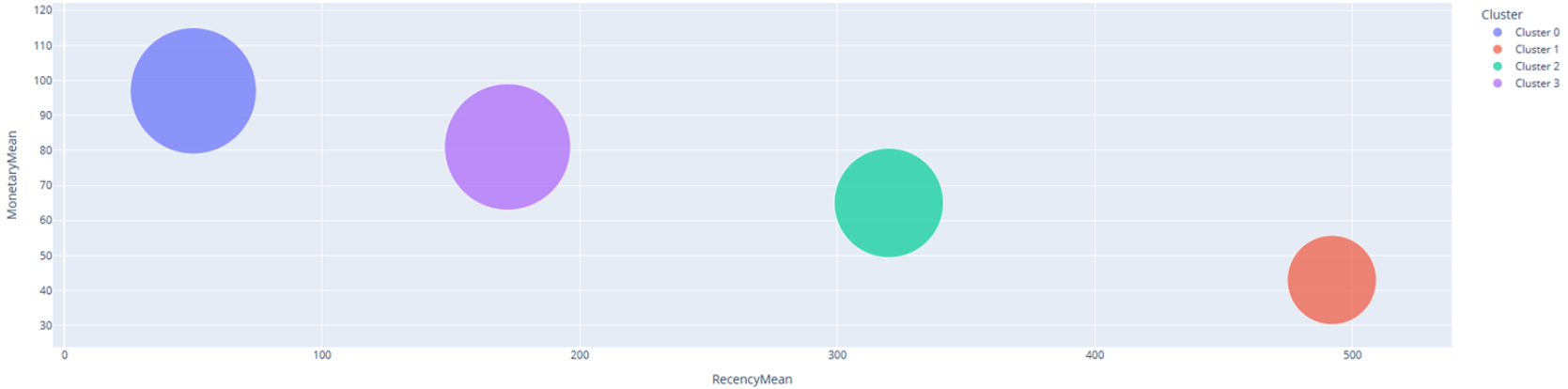
k=4



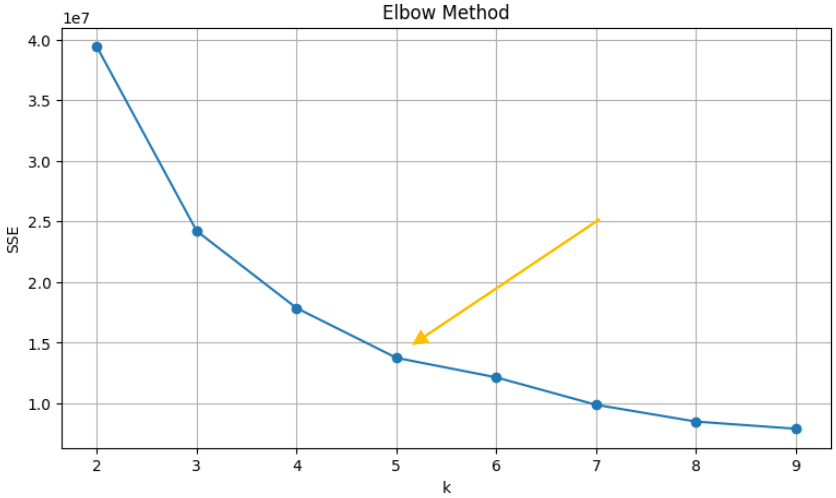
KMEANS (REMOVE OUTLIER)



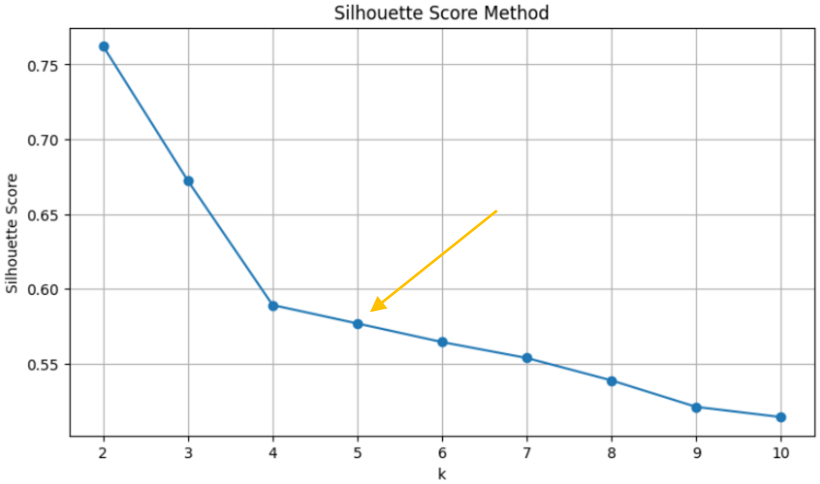
KMEANS (REMOVE OUTLIER)



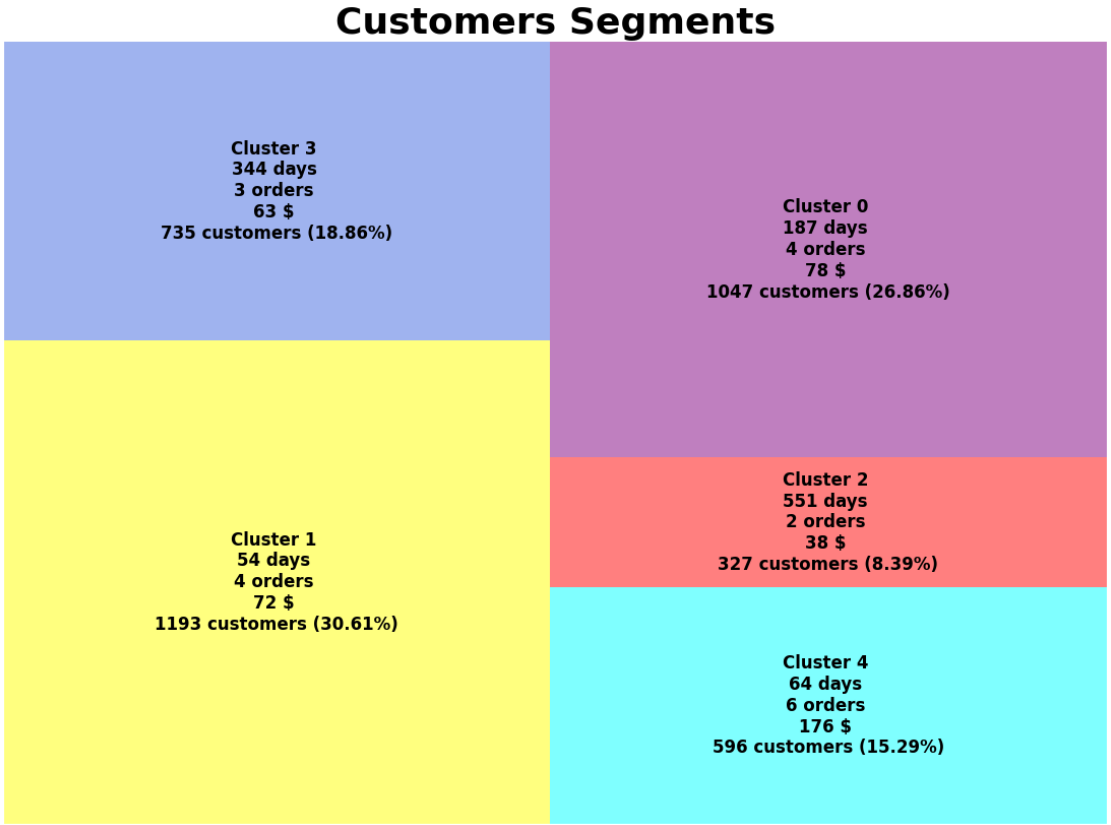
KMEANS (PYSPARK)



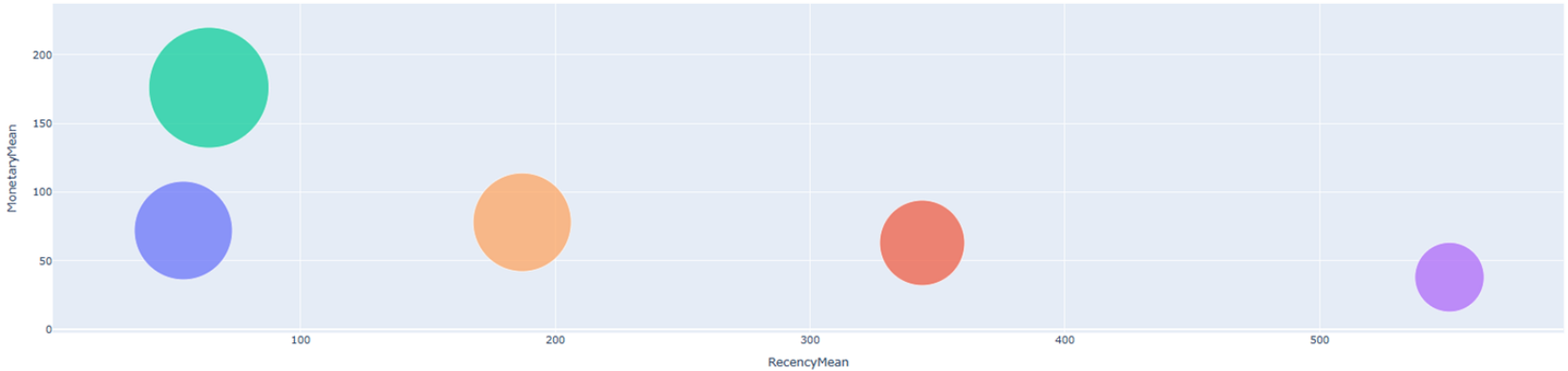
k=5



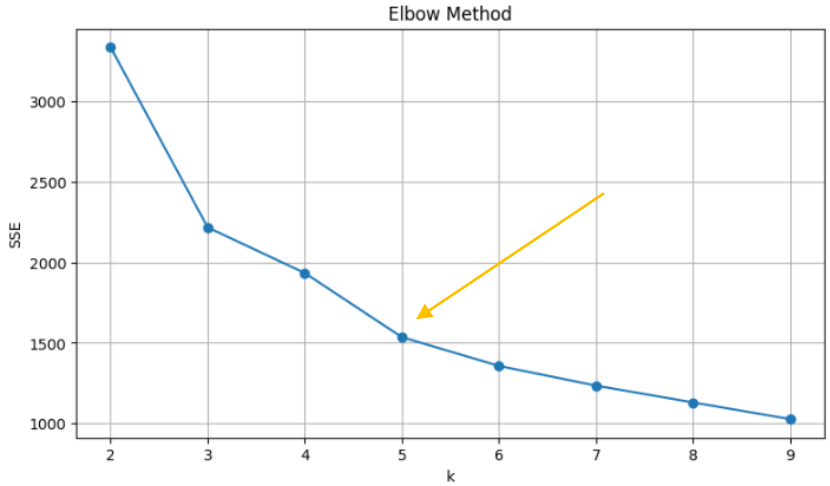
KMEANS (PYSPARK)



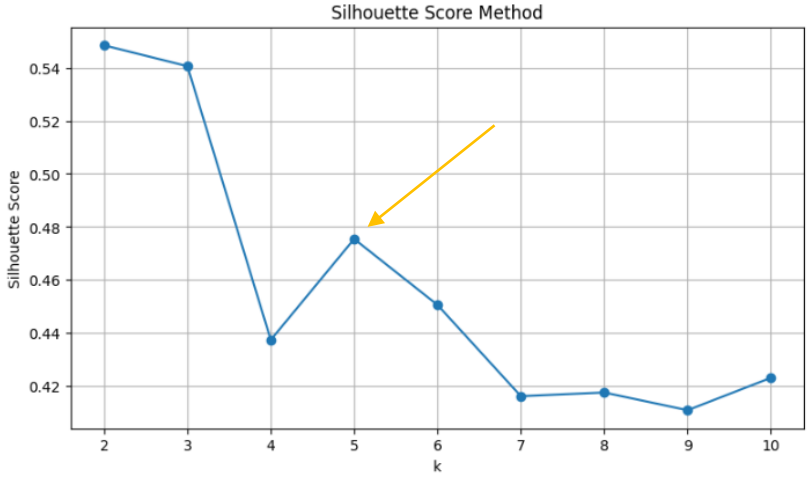
KMEANS (PYSPARK)



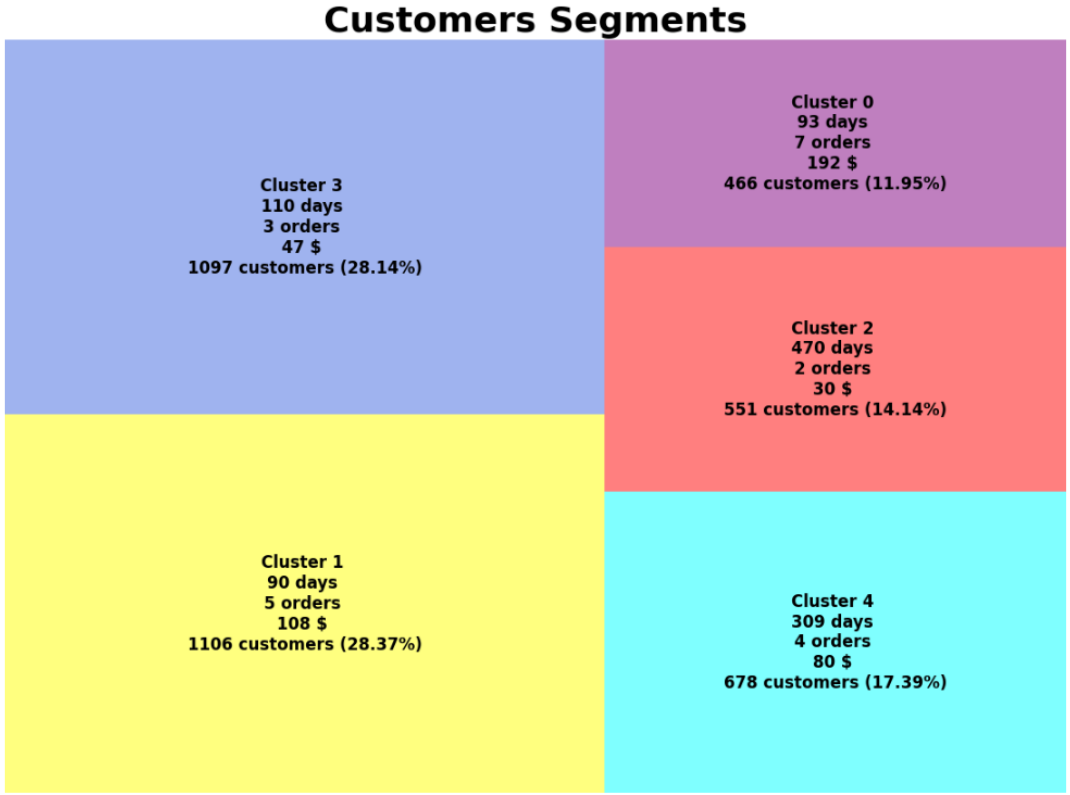
KMEANS (PYSPARK + SCALE DATA)



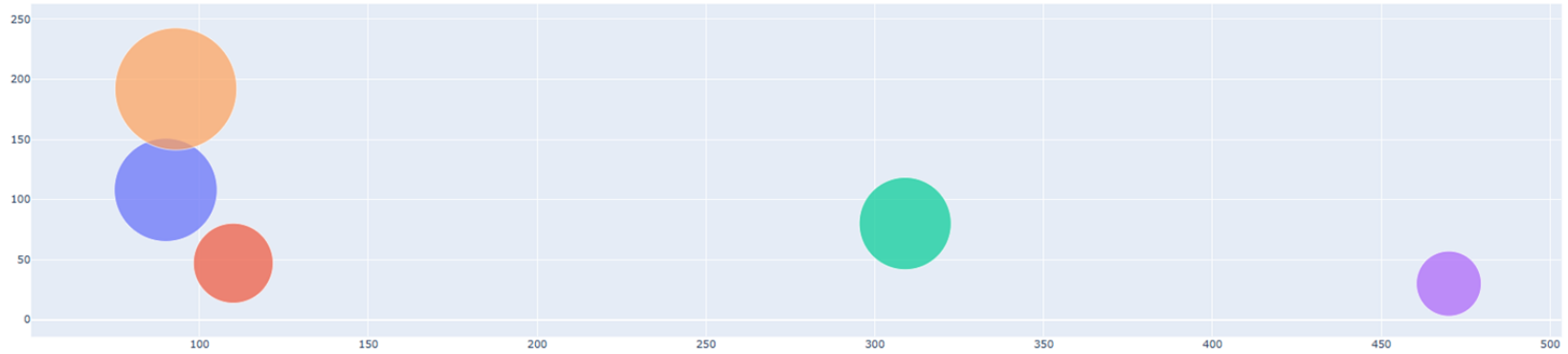
k=5



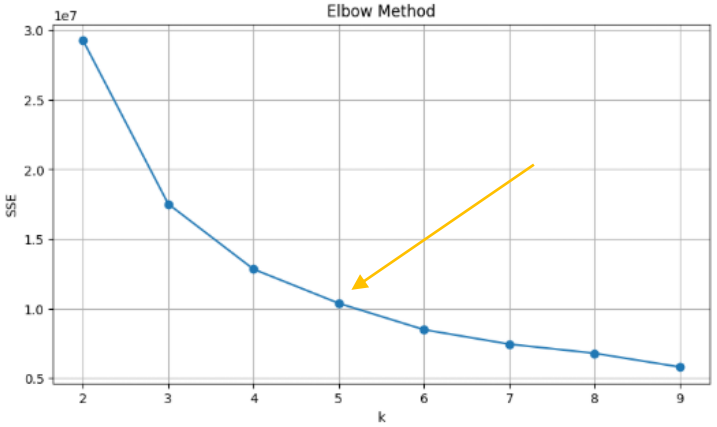
KMEANS (PYSPARK + SCALE DATA)



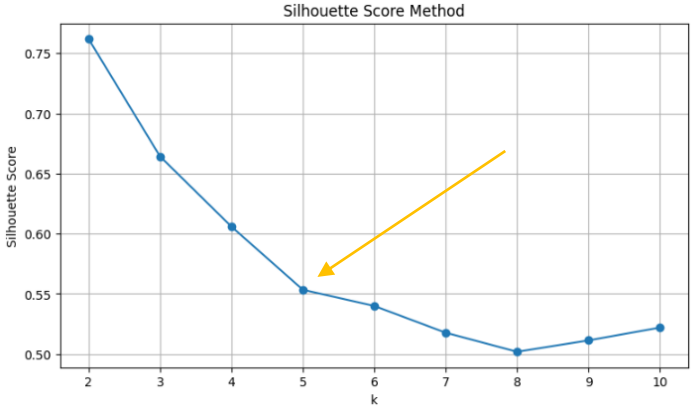
KMEANS (PYSPARK + SCALE DATA)



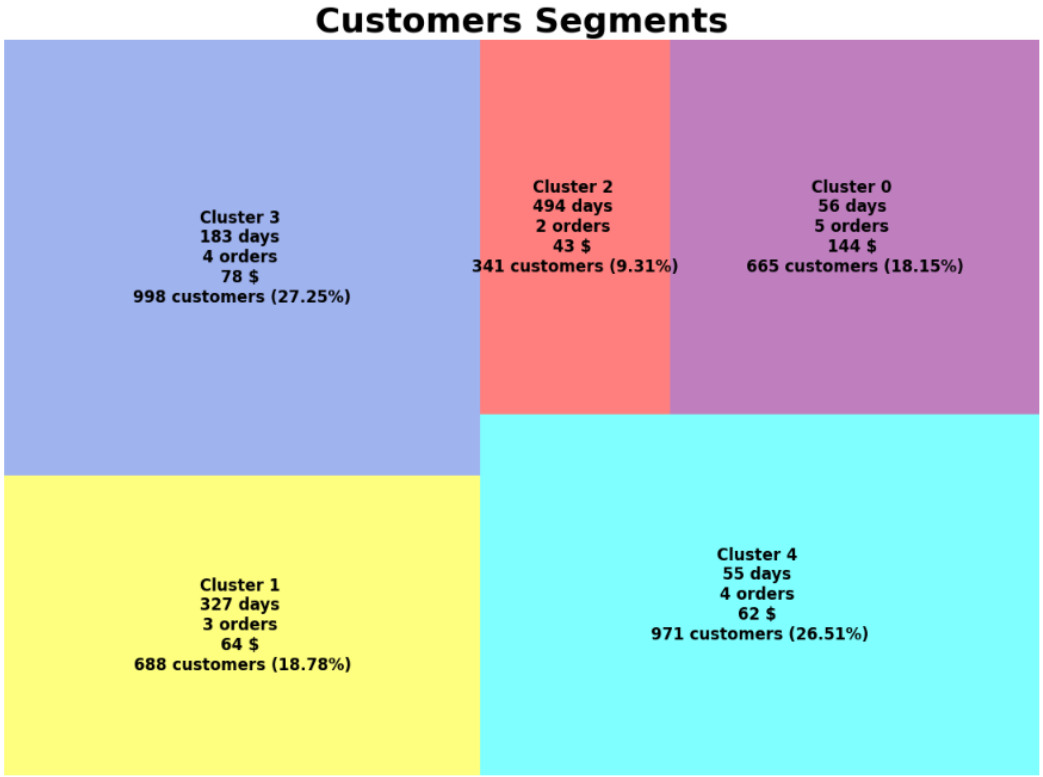
KMEANS (PYSPARK + REMOVE OUTLIERS)



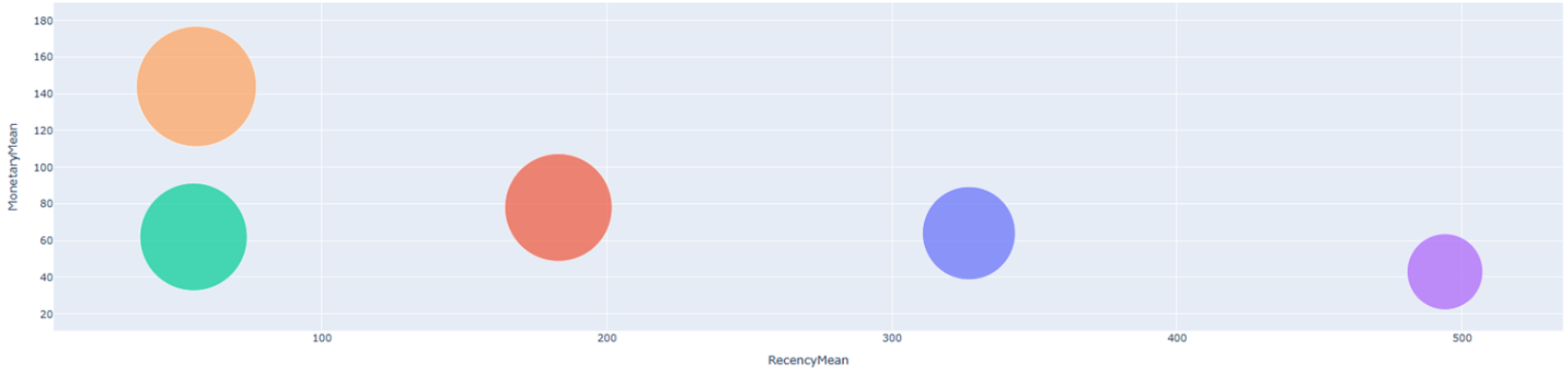
k=5



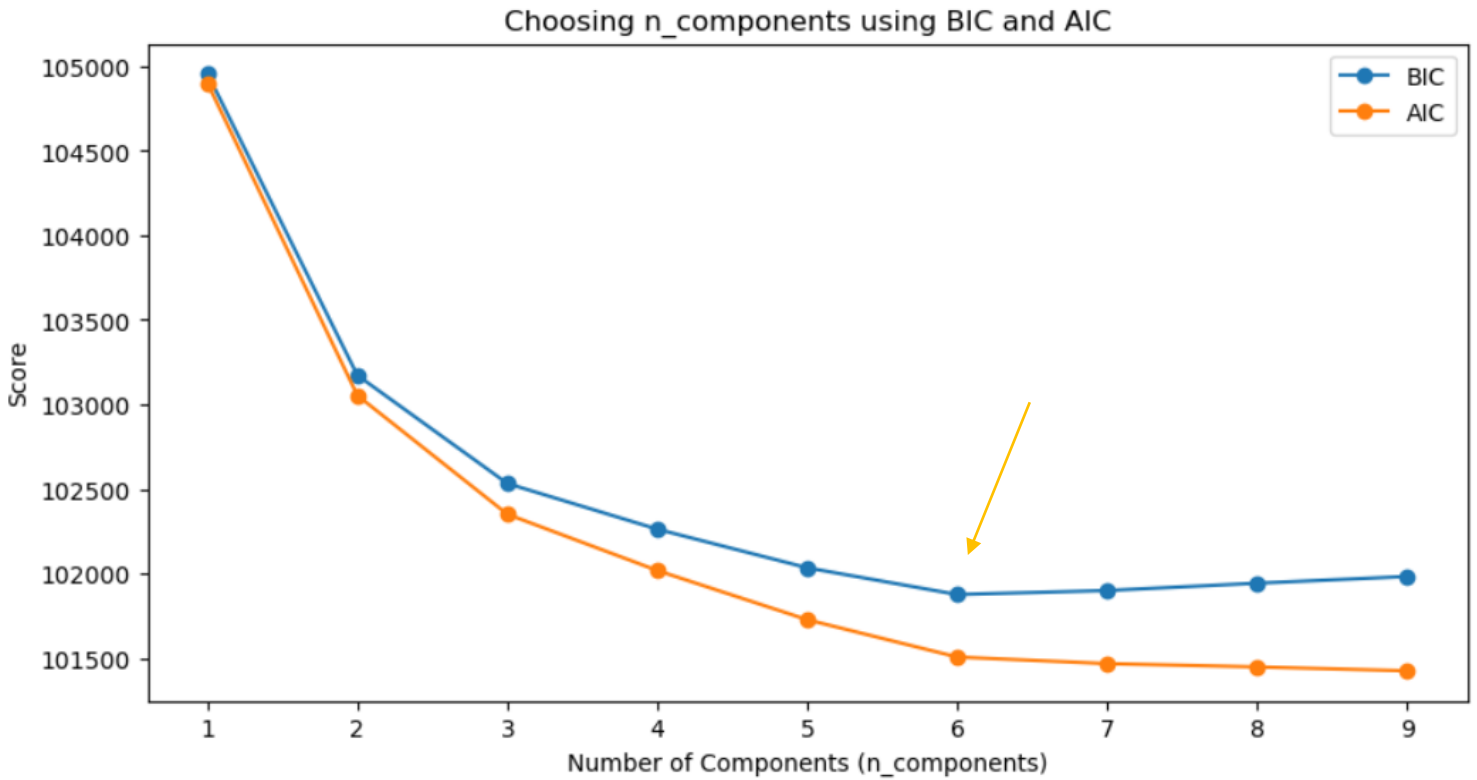
KMEANS (PYSPARK + REMOVE OUTLIERS)



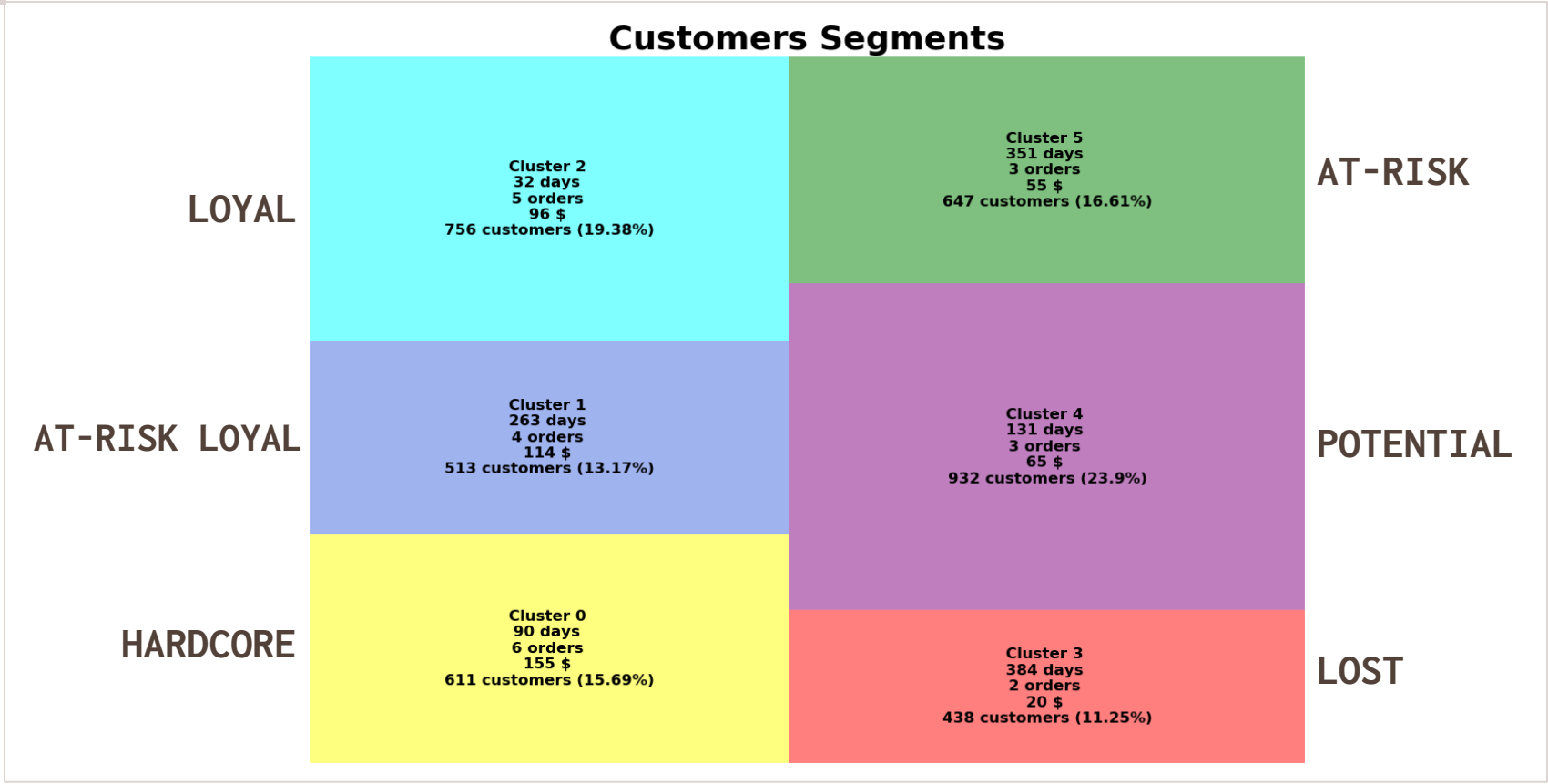
KMEANS (PYSPARK + REMOVE OUTLIERS)



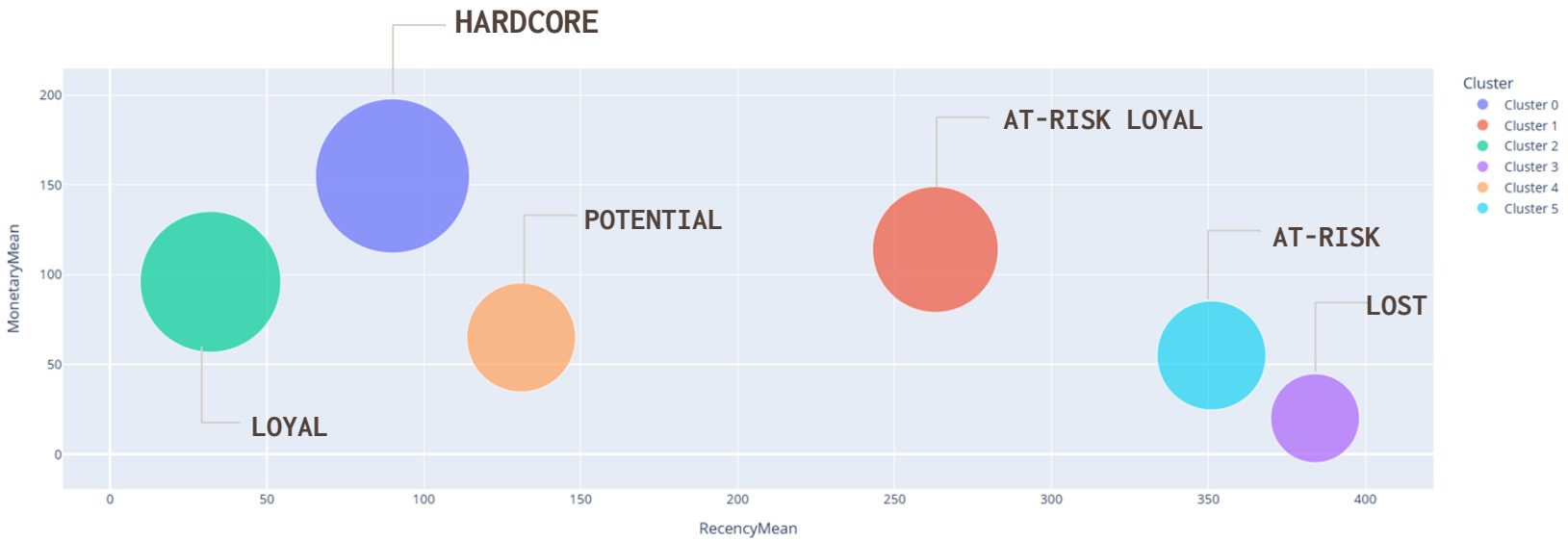
GMM


Choosing n_components using BIC and AIC
BIC
AIC
Score
105000
104500
104000
103500
103000
102500
102000
101500
1
2
3
4
5
6
7
8
9
Number of Components (n_components)

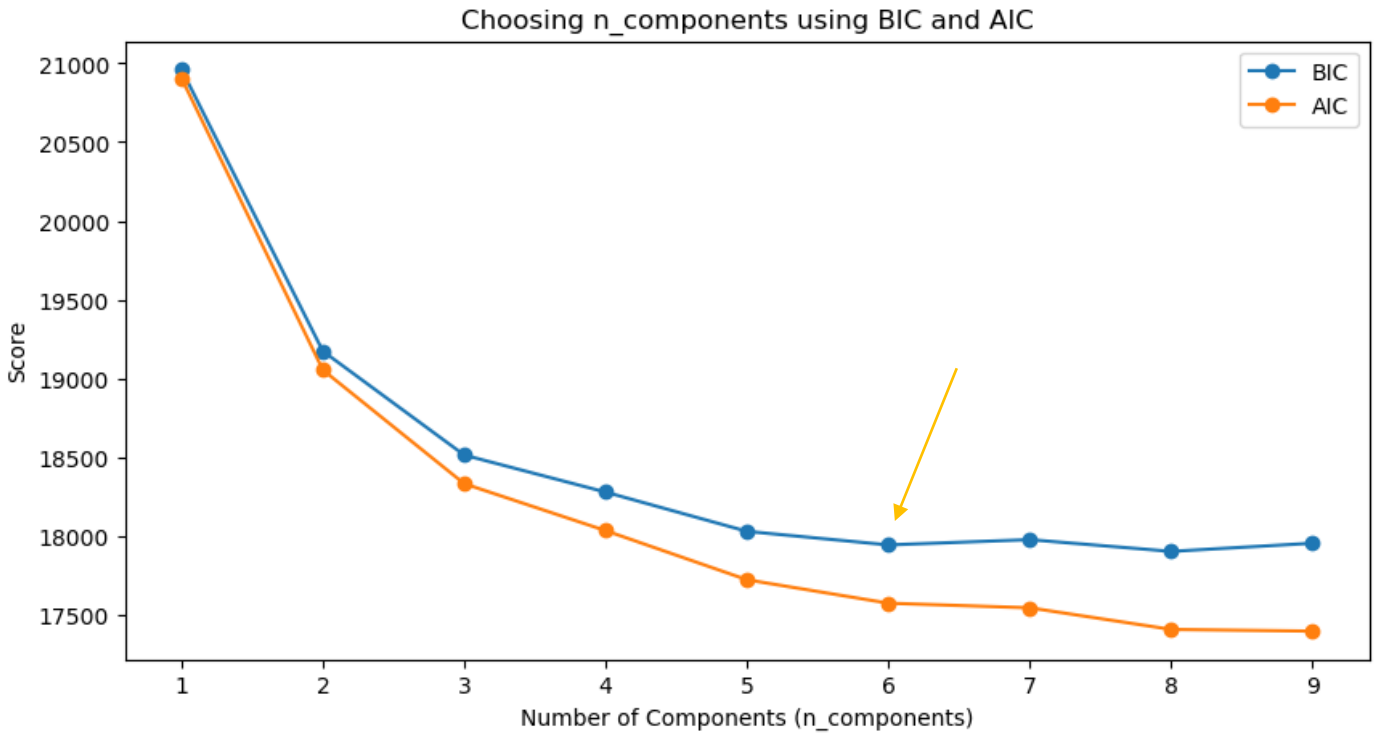
GMM


Customers Segments
LOYAL
AT-RISK LOYAL
HARDCORE
AT-RISK
POTENTIAL
LOST
Cluster 2
32 days
5 orders
96 $
756 customers (19.38%)
Cluster 5
351 days
3 orders
55 $
647 customers (16.61%)
Cluster 1
263 days
4 orders
114 $
513 customers (13.17%)
Cluster 4
131 days
3 orders
65 $
932 customers (23.9%)
Cluster 0
90 days
6 orders
155 $
611 customers (15.69%)
Cluster 3
384 days
2 orders
20 $
438 customers (11.25%)

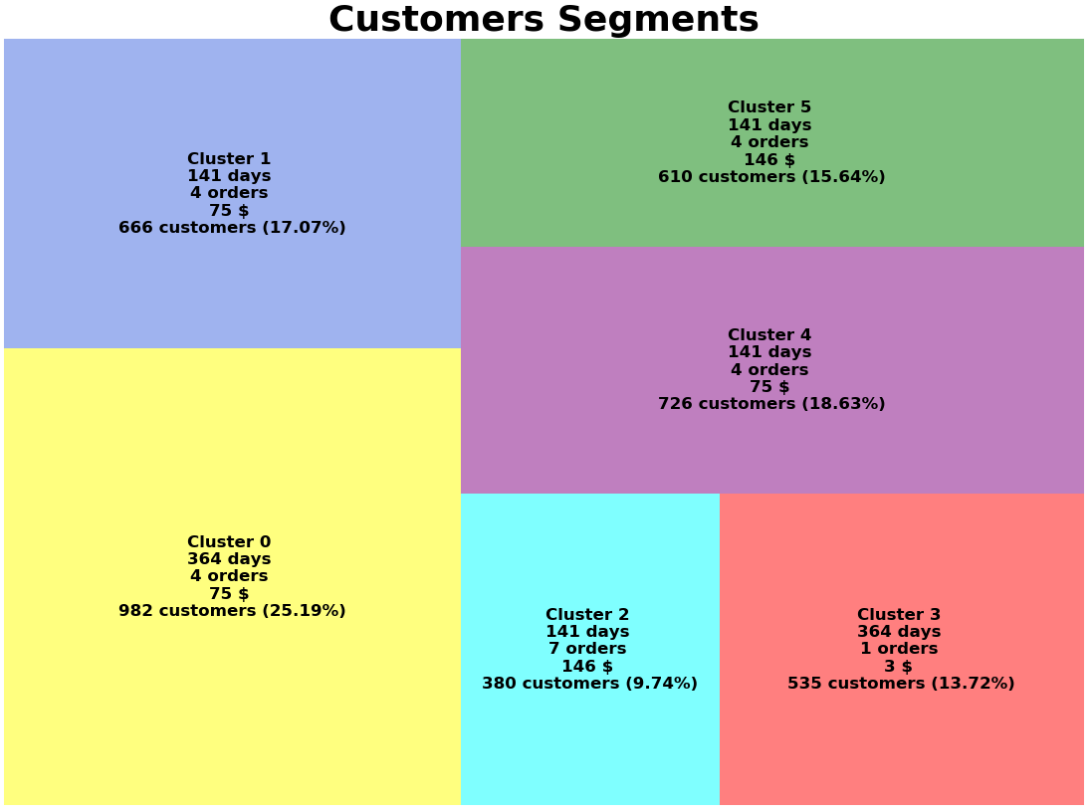
GMM


HARDCORE
AT-RISK LOYAL
POTENTIAL
AT-RISK
LOST
LOYAL
MonetaryMean
RecencyMean
0
50
100
150
200
250
300
350
400
Cluster
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5

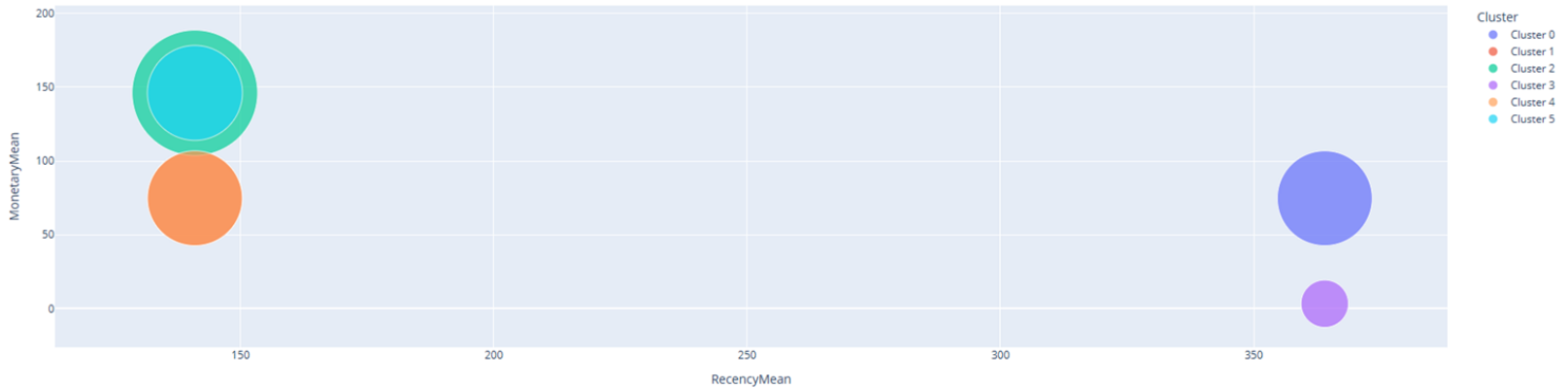
GMM (SCALE DATA)



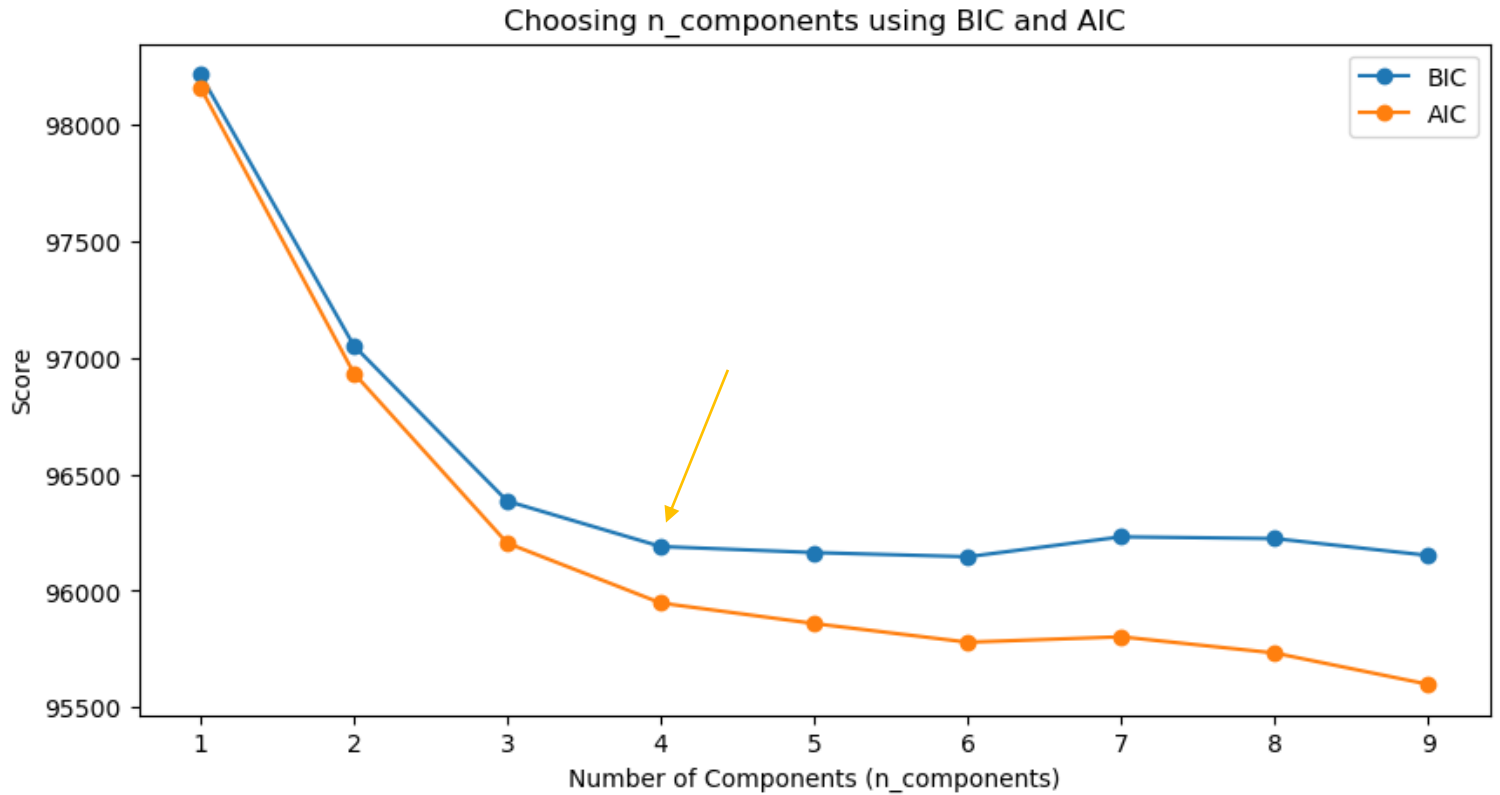
GMM (SCALE DATA)



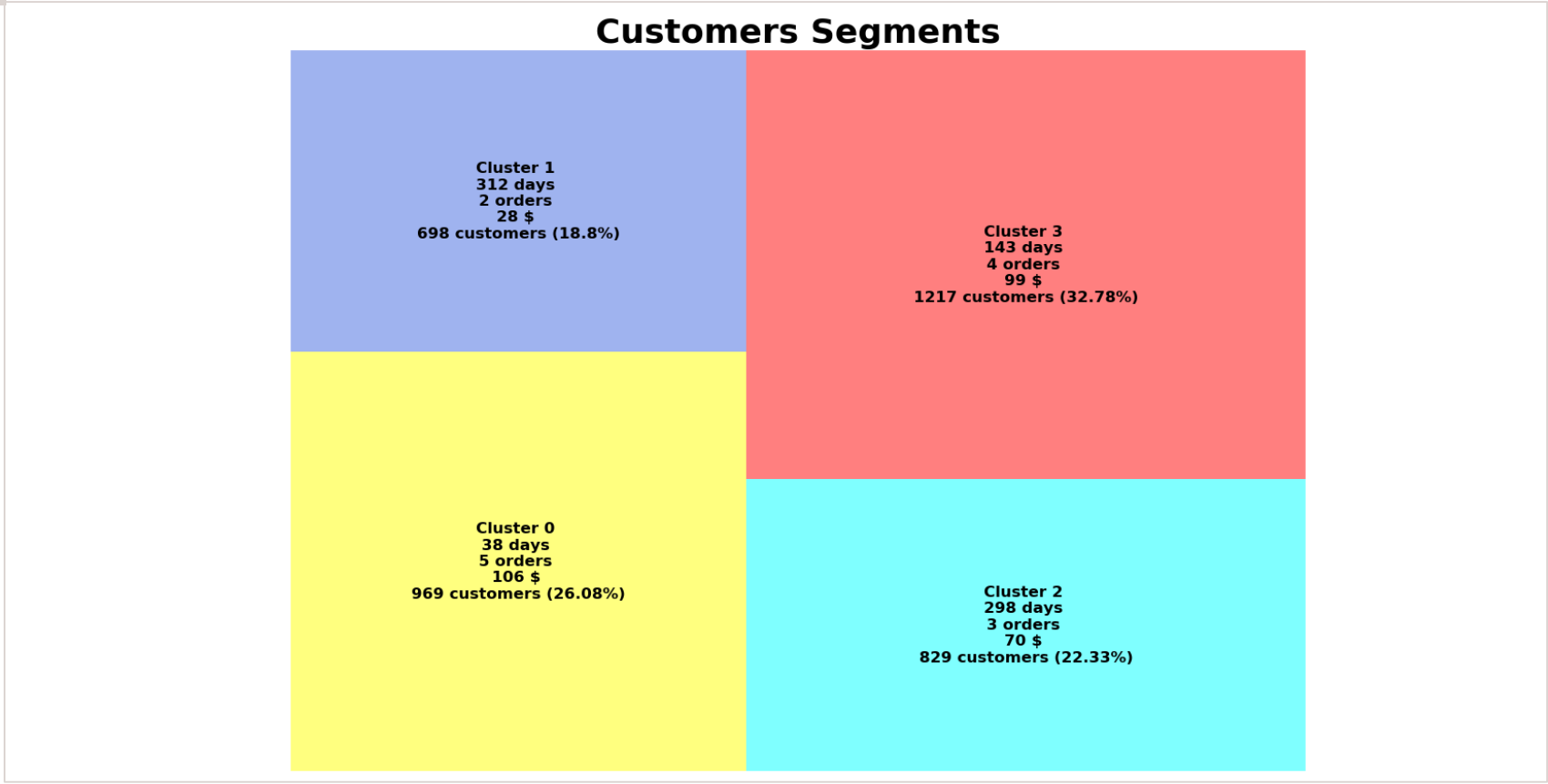
GMM (SCALE DATA)



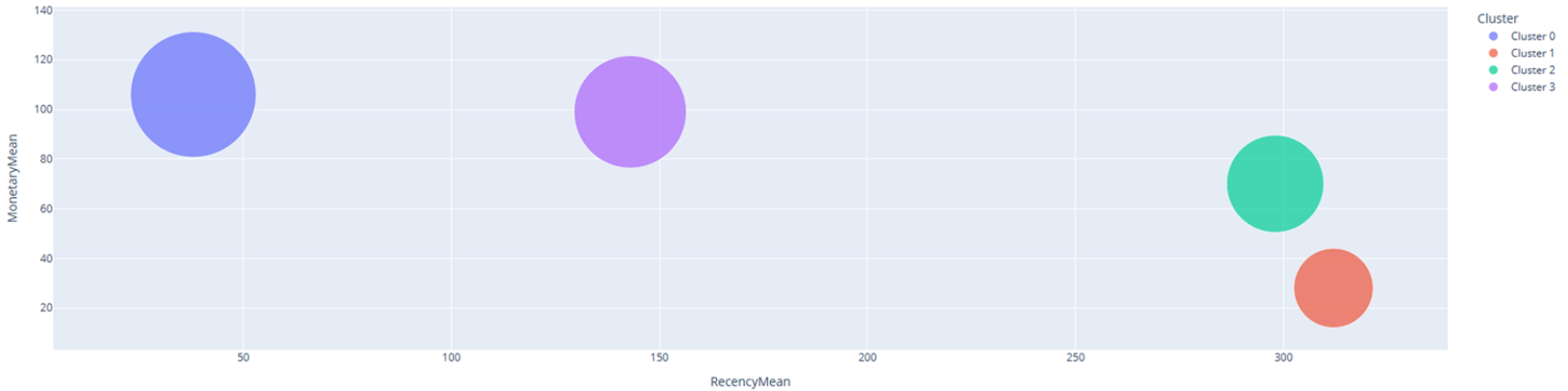
GMM (REMOVE OUTLIER)



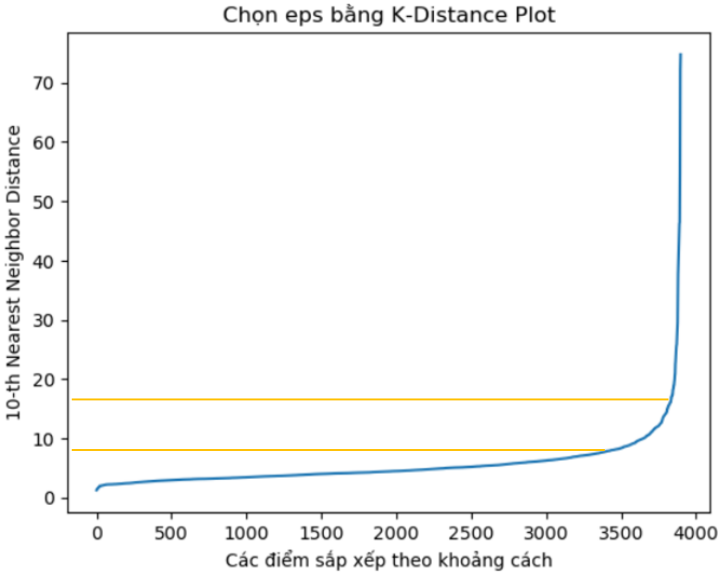
GMM (REMOVE OUTLIER)

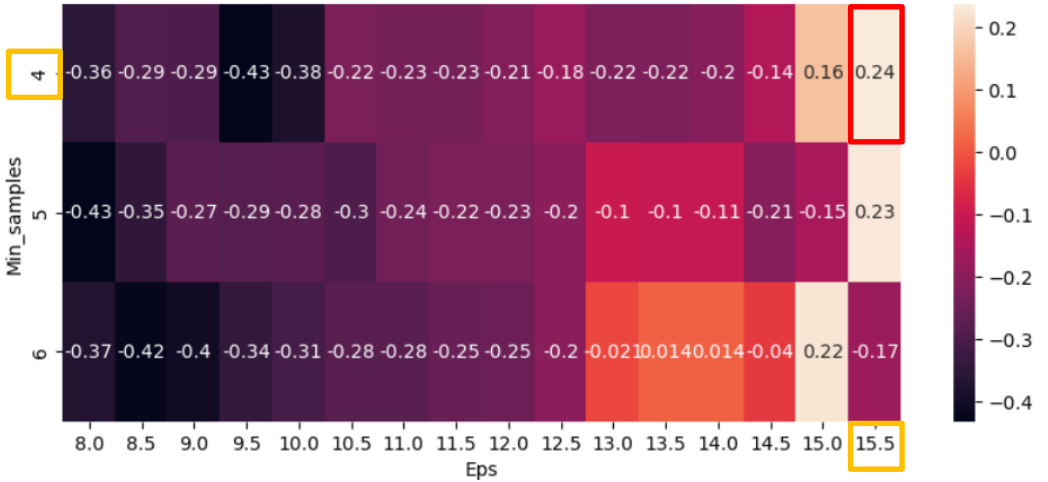


GMM (REMOVE OUTLIER)



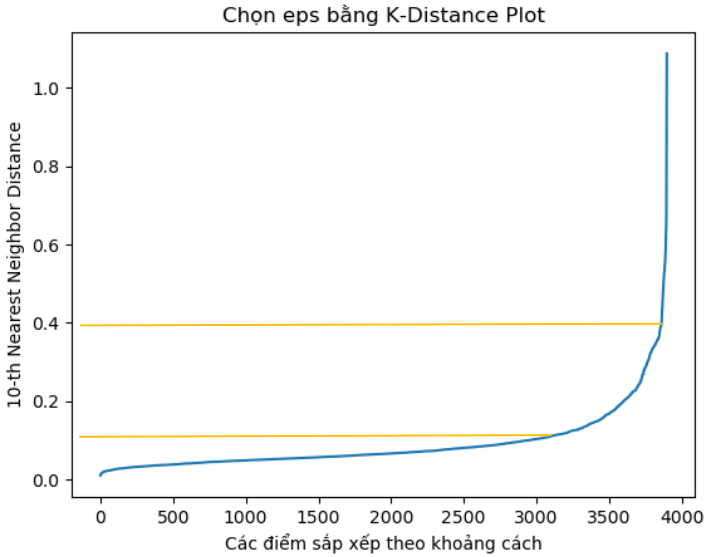
DBSCAN

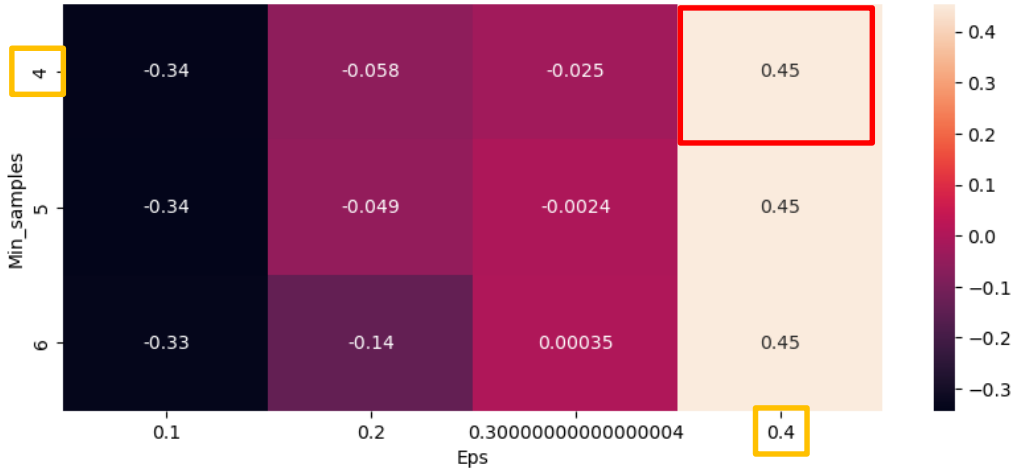




Nhiễu (-1): 53 điểm
Cụm 0: 3839 điểm
Cụm 1: 6 điểm

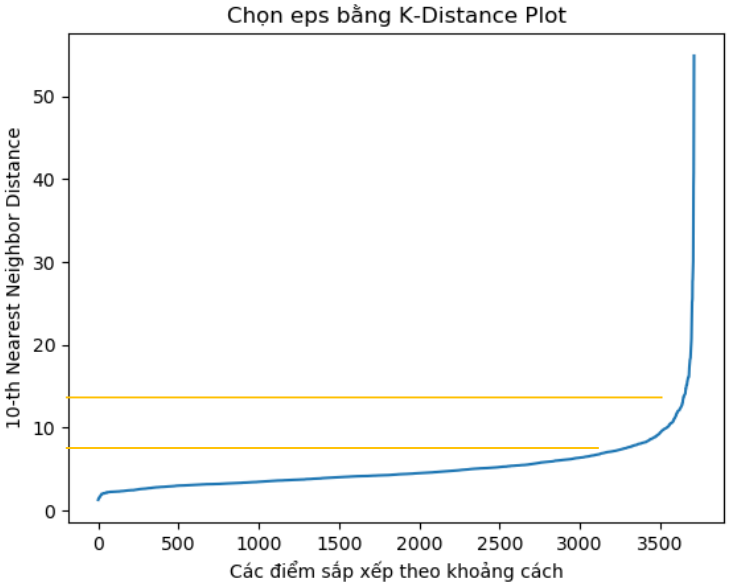
DBSCAN (SCALE DATA)

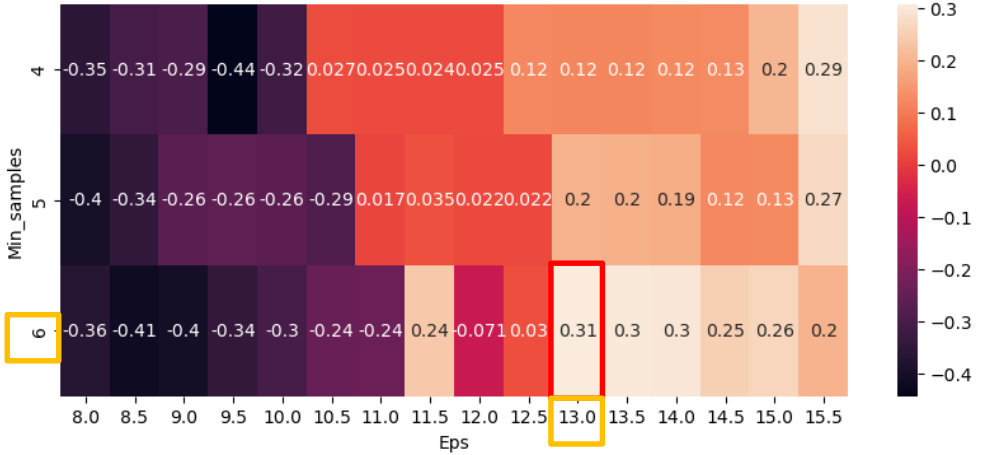





```
Nhiễu (-1): 25 điểm
Cụm 0: 3873 điểm
```

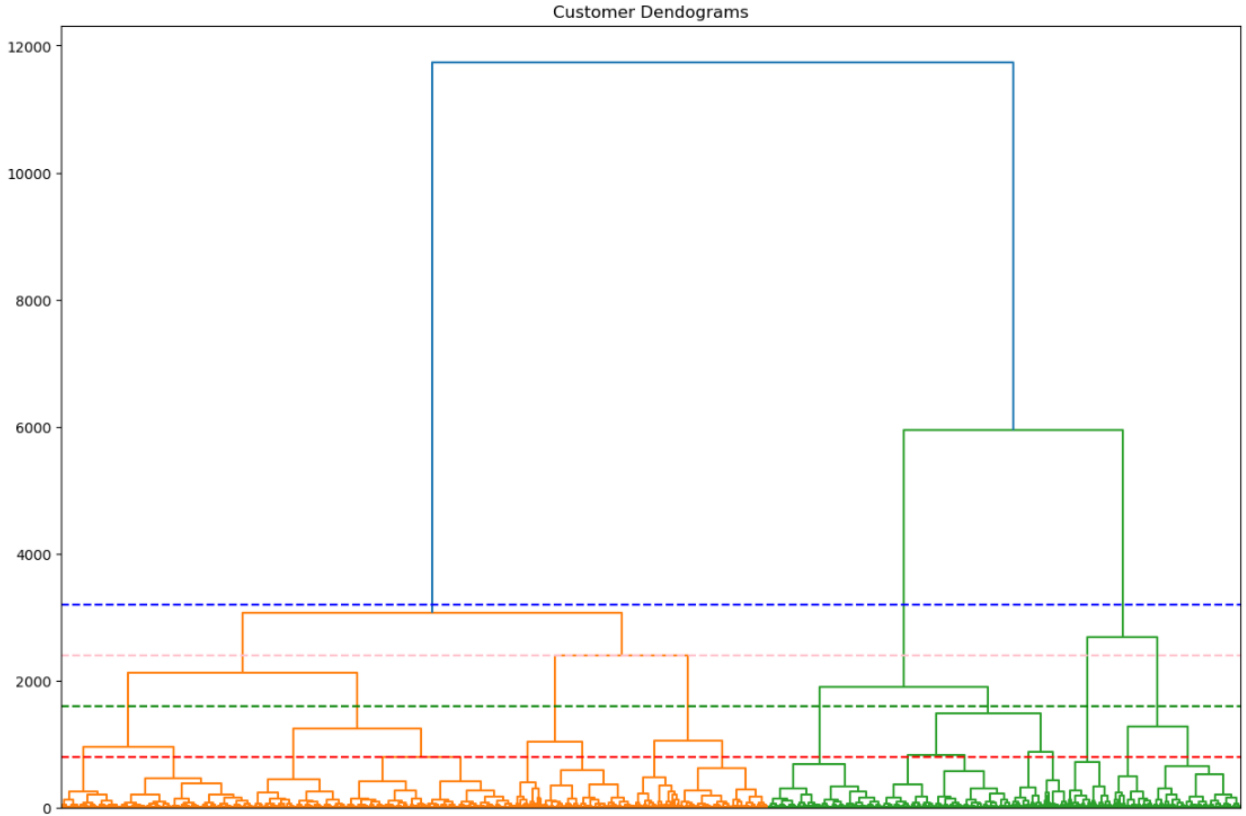
DBSCAN (REMOVE OUTLIERS)



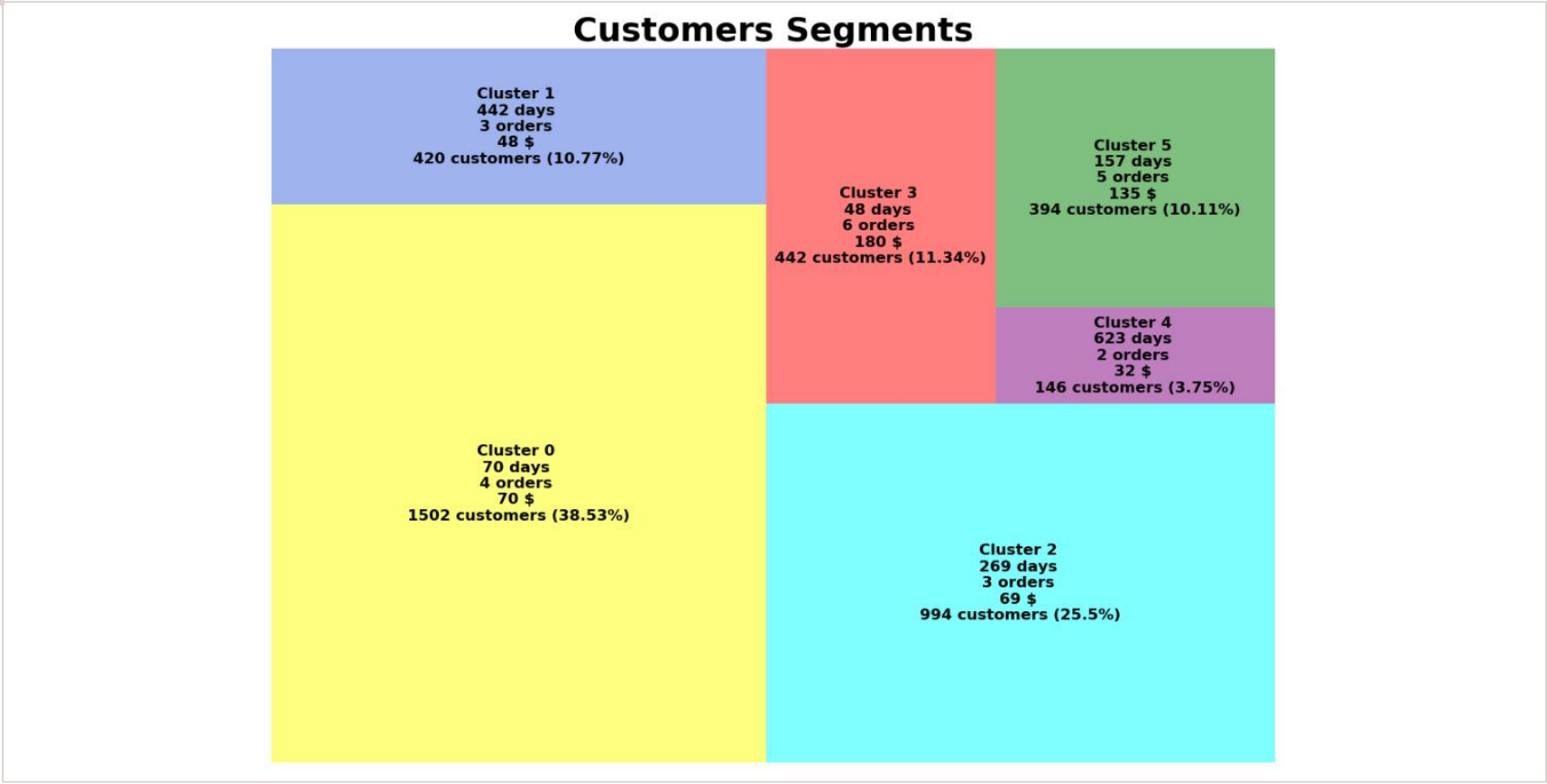



```
Nhiễu (-1): 79 điểm
Cụm 0: 3634 điểm
```

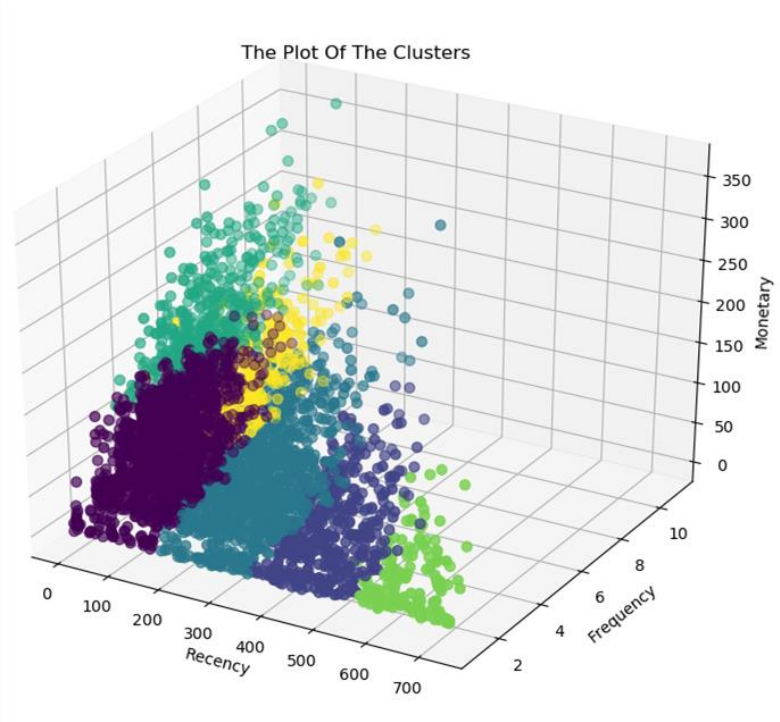
HIERARCHICAL CLUSTERING



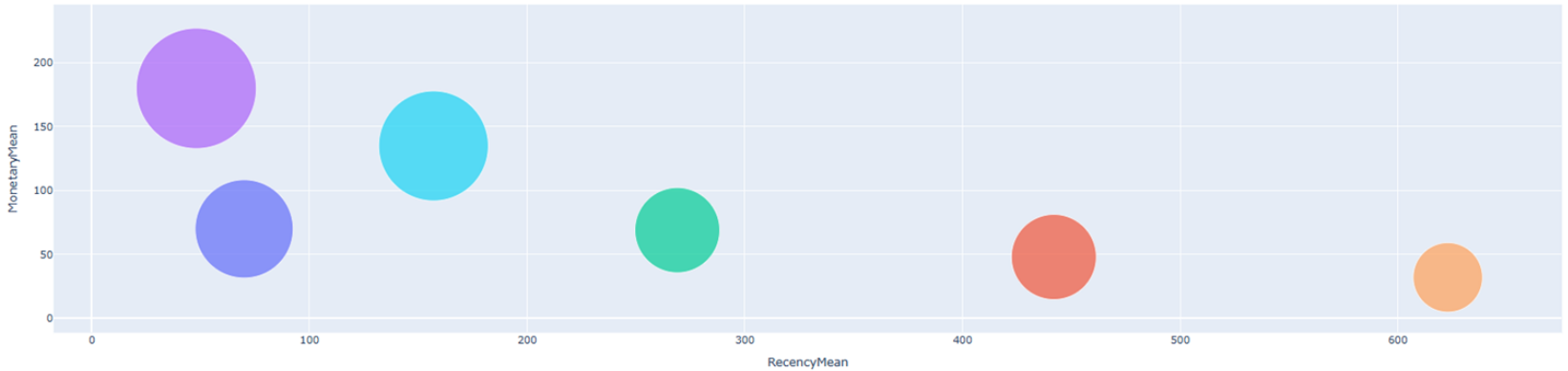
HIERARCHICAL CLUSTERING



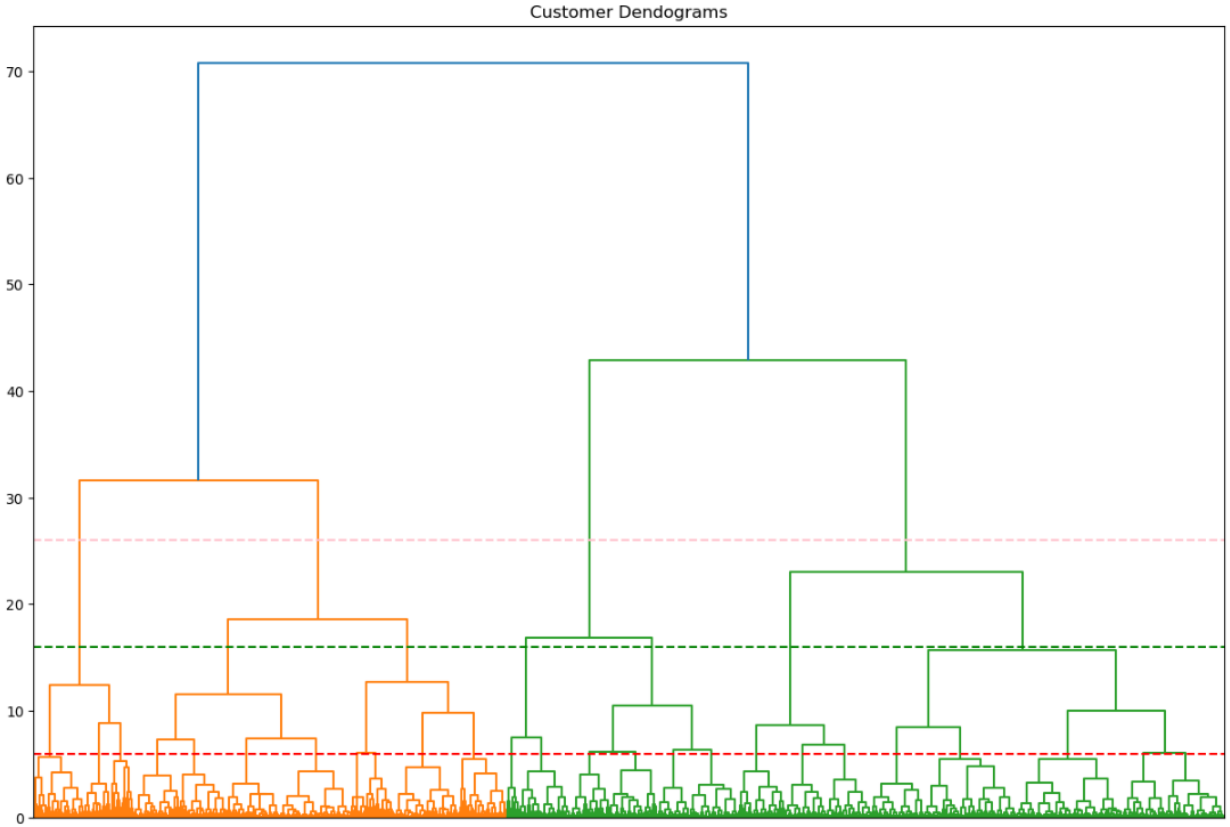
HIERARCHICAL CLUSTERING



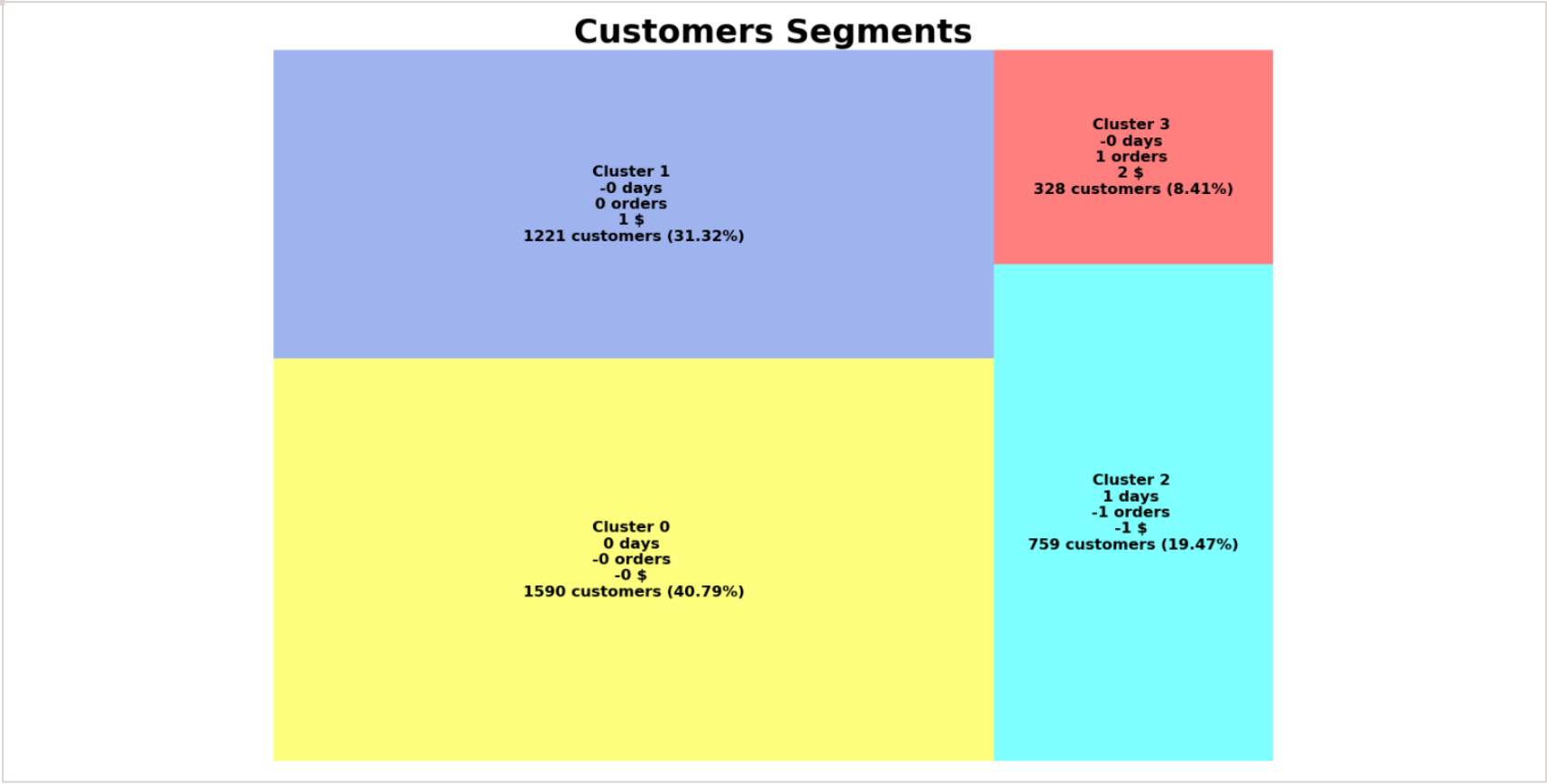
HIERARCHICAL CLUSTERING



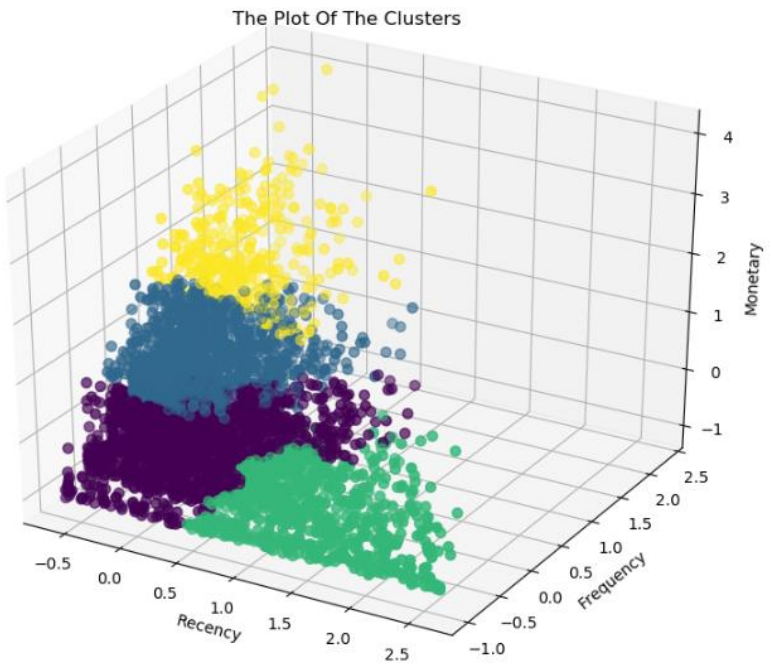
HIERARCHICAL CLUSTERING (SCALE DATA)

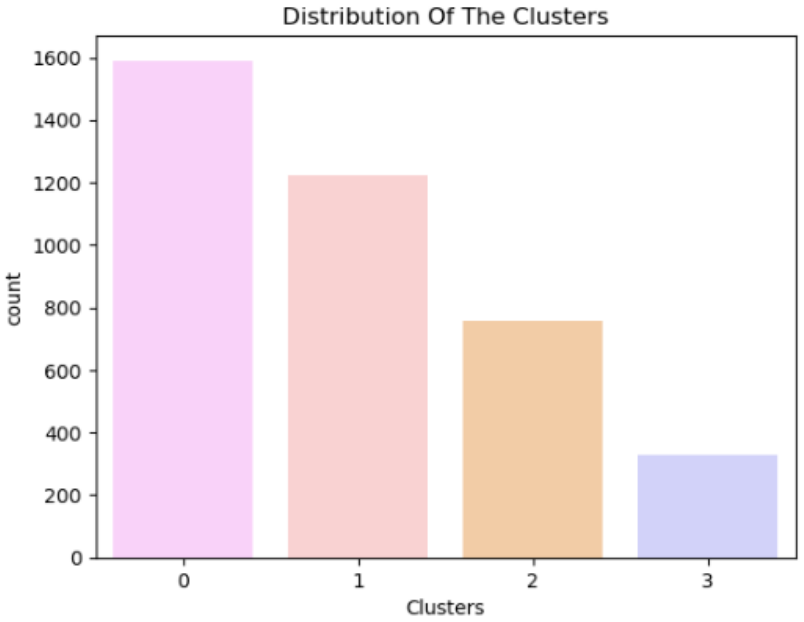


HIERARCHICAL CLUSTERING (SCALE DATA)

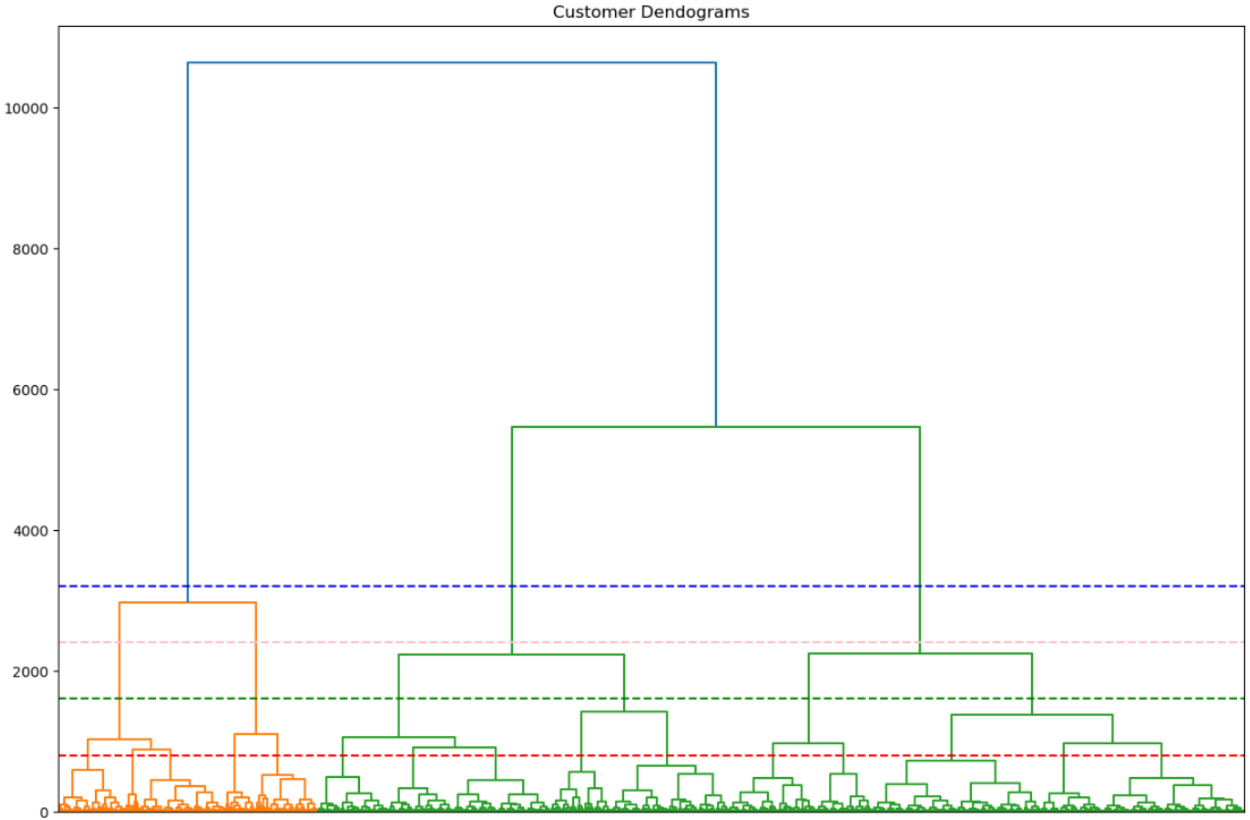


HIERARCHICAL CLUSTERING (SCALE DATA)

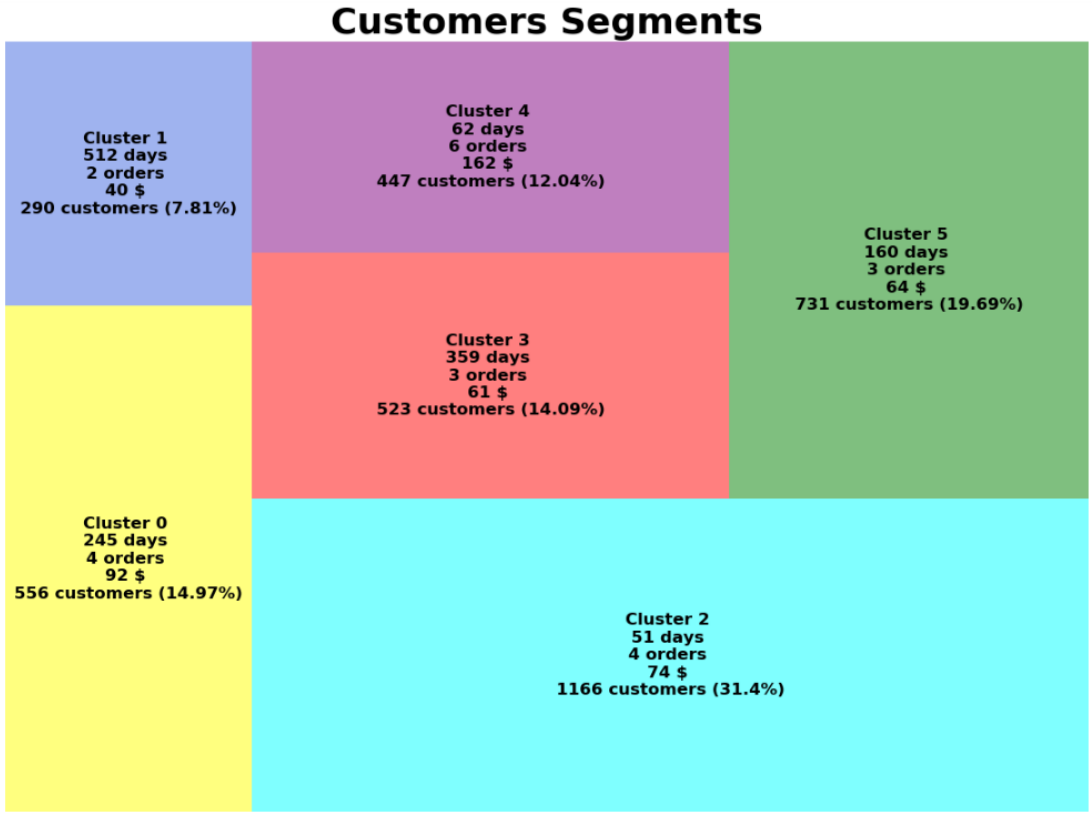




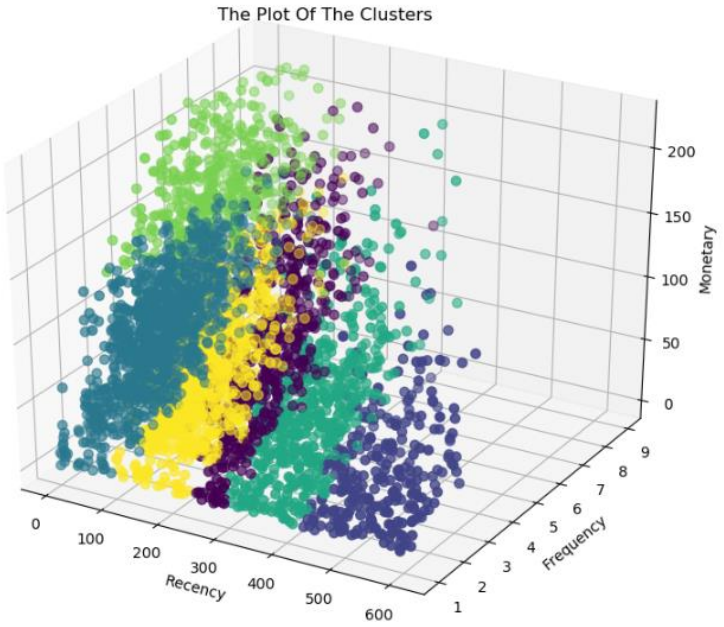
HIERARCHICAL CLUSTERING (REMOVE OUTLIERS)

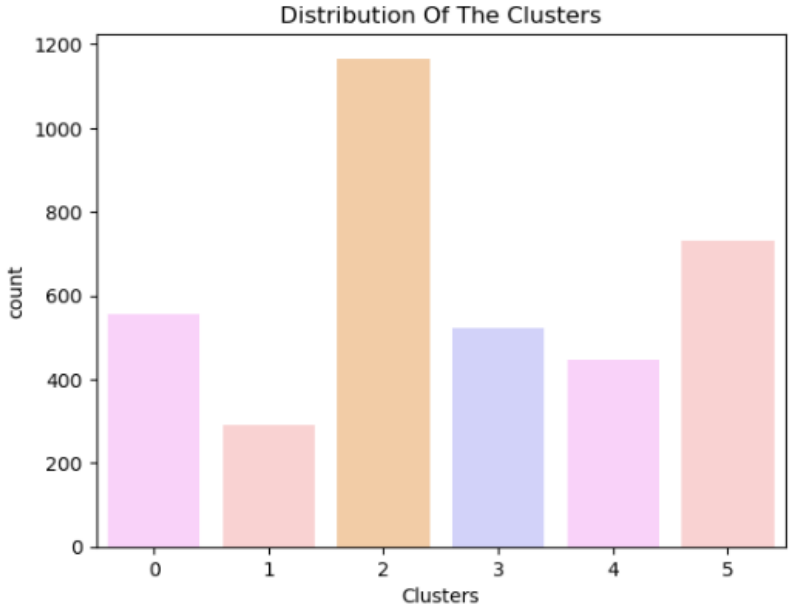


HIERARCHICAL CLUSTERING (REMOVE OUTLIERS)

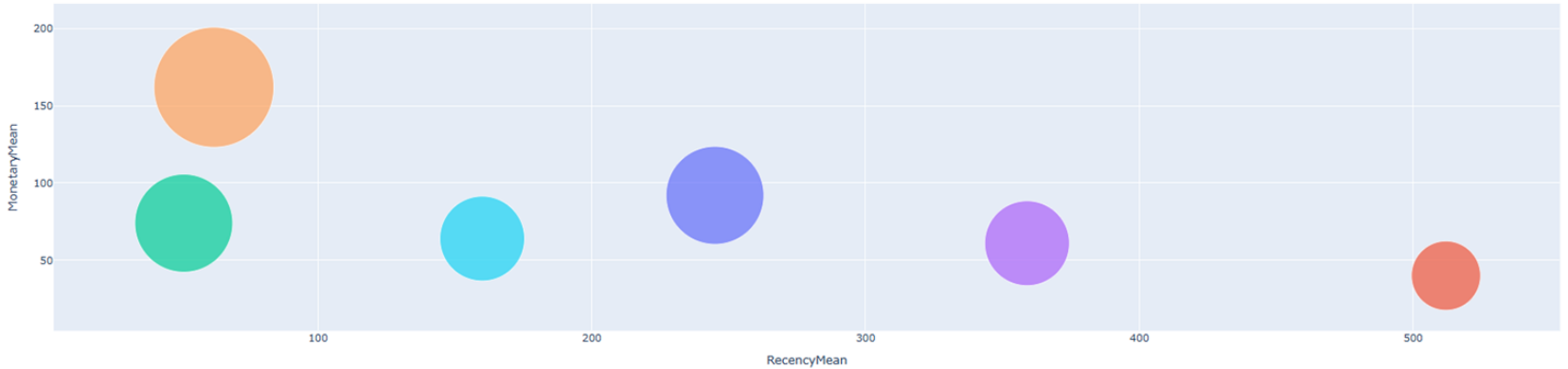


HIERARCHICAL CLUSTERING (REMOVE OUTLIERS)





HIERARCHICAL CLUSTERING (REMOVE OUTLIERS)



EVALUATION

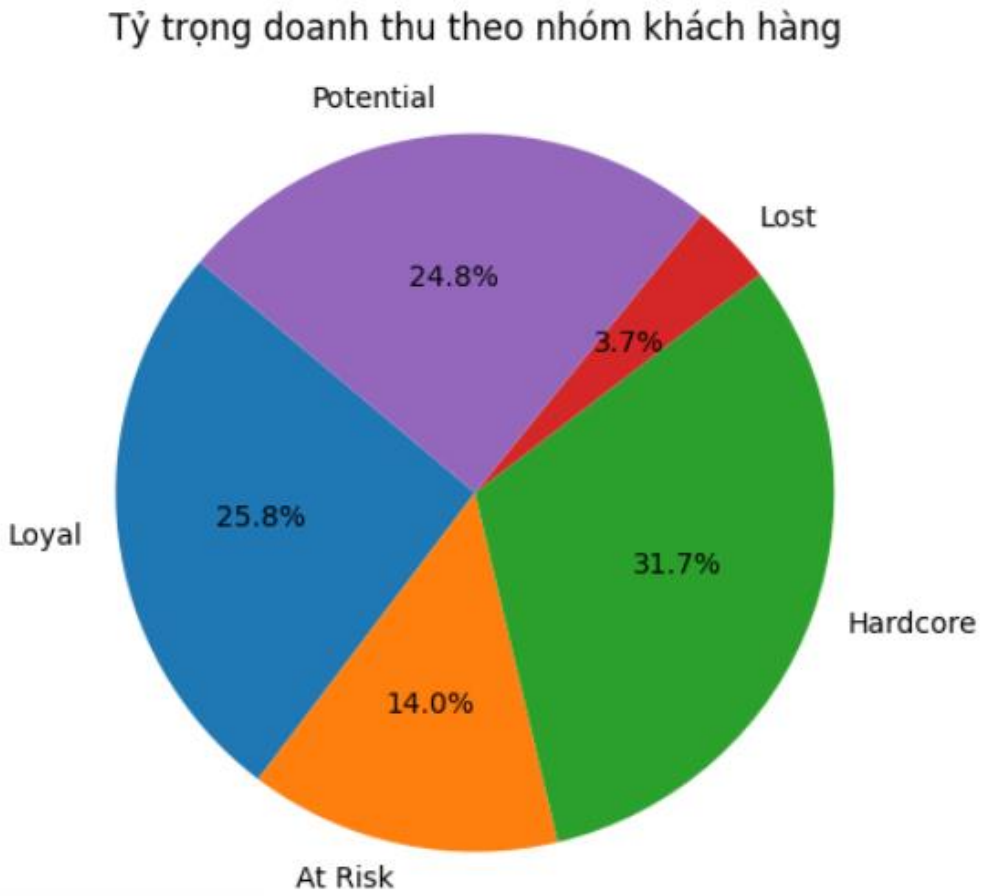
Phương pháp	Số cụm	Silhouette Score	Nhận xét
Manual Segmentation	6	-0.059	Thực hiện phân khúc thủ công dựa trên cảm tính, không phản ánh đúng cấu trúc dữ liệu
KMeans	2	0.57	Các cụm rõ ràng và tách biệt tốt, tuy nhiên 2 cụm không đủ để phản ánh sự đa dạng của khách hàng
KMeans	**5**	**0.4**	Cung cấp một sự phân chia chi tiết hơn k=2 mà vẫn giữ được tính chất phân tách tốt giữa các cụm
KMeans (PySpark)	5	0.58	Cùng chia 5 cụm nhưng kết quả Sihoutte tốt hơn machine learning truyền thống nhiều
GMM	6	0.2	Cách cụm khá gần nhau, phân tách chưa tốt
DBSCAN	2 + outliers	0.23	Phần lớn dữ liệu tập trung vào 1 cụm, không hiệu quả
Hierarchical Clustering	6	0.33	Cách cụm phân tách tốt hơn GMM, nhưng vẫn không tối ưu bằng KMeans

EVALUATION (REMOVE OUTLIERS & SCALE DATA)

Phương pháp	Remove outliers		Scale data		Nhận xét
	Số cụm	Silhouette Score	Số cụm	Silhouette Score	
KMeans	4	**0.42**	3	0.36	Khi loại bỏ outliers thì các cụm phân tách tốt
KMeans (PySpark)	6	0.45	5	0.48	Các cụm chồng chéo lên nhau
GMM	4	0.2	6	0.13	Các cụm không được phân tách rõ ràng
DBSCAN	1 + outliers	0.3	1+ outliers	0.45	Không thực hiện phân cụm được
Hierarchical Clustering	6	0.23	4	0.25	Kết quả phân cụm không tốt

06

SOLUTIONS


Tỷ trọng doanh thu theo nhóm khách hàng
Potential
24.8%
Lost
3.7%
Loyal
25.8%
31.7%
Hardcore
14.0%
At Risk

Segment	Tỷ lệ	Đặc điểm			Doanh thu	Mục tiêu	Đề xuất phương pháp tiếp cận
		R	F	M			
Hardcore	15	Cao	Cao	Cao	31.7%	Giữ chân, tăng giá trị đơn	Chương trình khách hàng thân thiết cao cấp, dịch vụ cá nhân hóa, cross-sell
Loyal	19	Cao	Cao	TB	25.8%	Tăng giá trị đơn	Tích điểm thành viên, khuyến mãi combo, Ưu đãi sinh nhật & dịp đặc biệt
Potential	23	TB	Cao	TB	24.8%	Tăng tần suất mua	Ưu đãi cá nhân hóa diễn ra trong thời gian ngắn, gửi tin SMS/mail
At Risk	17	Thấp	TB	Thấp	14	Lôi kéo trở lại	Khảo sát và cải thiện dịch vụ, tương tác trên nhiều channel, giảm giá đặc biệt hoặc quà tặng
Lost	12	Thấp	Thấp	Thấp	3.7%	Cân nhắc nguồn lực, có thể bỏ qua nhóm này	Khảo sát lý do rời bỏ, vì nhóm chiếm tỷ lệ khá ít nên có thể bỏ qua

THANK YOU!

FOR WATCHING