UNIVERSITY OF SCIENCE
Informatics Center

Project 1

Customer Segmentation

HO NGUYEN THIEN BAO

NGUYEN ANH KHOA

(04 / 2025)

Table of Contents.

I. Dataset overview

II. EDA

III. Models

   i. Manual segmentation

   ii. K-Means with Scikit-learn

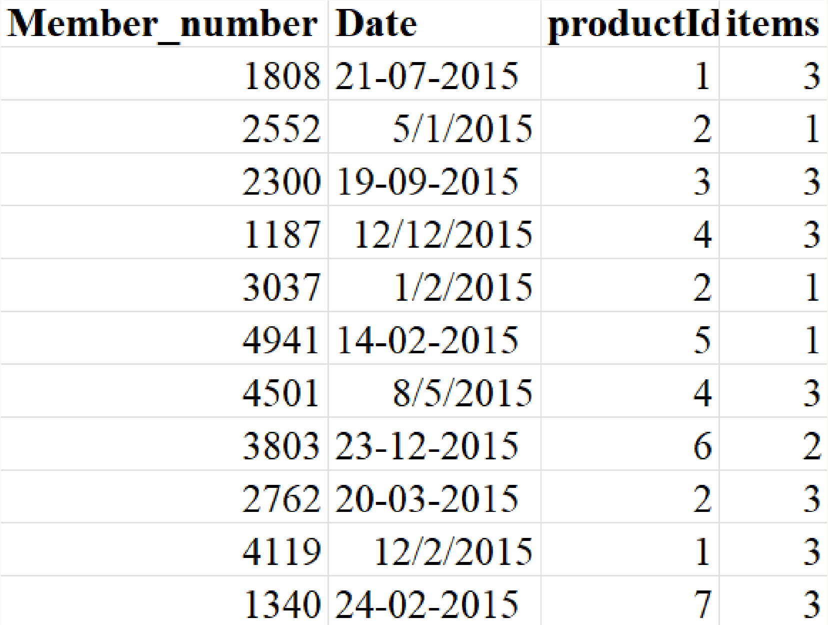
   iii. GMM

   iv. K-Means PySpark

IV. Conclusion

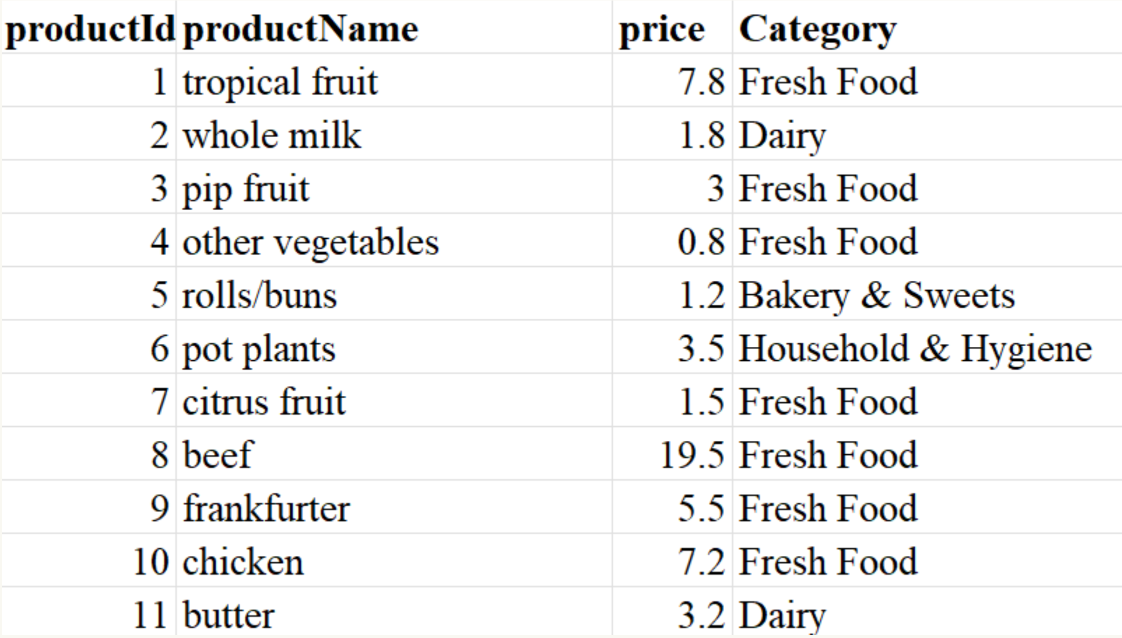
OVERVIEW

Dataset overview.



Member_number	Date	productId	items
1808	21-07-2015	1	3
2552	5/1/2015	2	1
2300	19-09-2015	3	3
1187	12/12/2015	4	3
3037	1/2/2015	2	1
4941	14-02-2015	5	1
4501	8/5/2015	4	3
3803	23-12-2015	6	2
2762	20-03-2015	2	3
4119	12/2/2015	1	3
1340	24-02-2015	7	3

Transaction.csv
4 cols
38.765 rows

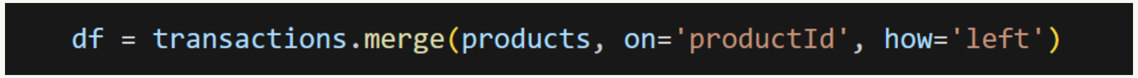


productId	productName	price	Category
1	tropical fruit	7.8	Fresh Food
2	whole milk	1.8	Dairy
3	pip fruit	3	Fresh Food
4	other vegetables	0.8	Fresh Food
5	rolls/buns	1.2	Bakery & Sweets
6	pot plants	3.5	Household & Hygiene
7	citrus fruit	1.5	Fresh Food
8	beef	19.5	Fresh Food
9	frankfurter	5.5	Fresh Food
10	chicken	7.2	Fresh Food
11	butter	3.2	Dairy

Products_with_Categories.csv
4 cols
167 rows

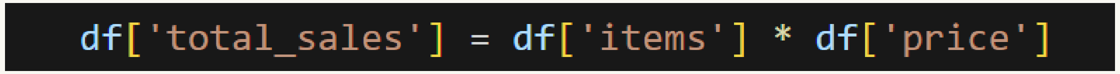
Data preparation.

- Merging 2 files to 1 Dataframe ‘**df**’



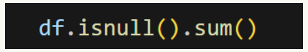
```
df = transactions.merge(products, on='productId', how='left')
```

- Compute ‘**total_sales**’ column



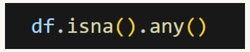
```
df['total_sales'] = df['items'] * df['price']
```

- Checking **Null** values



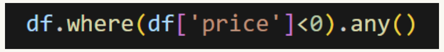
```
df.isnull().sum()
```

- Checking **NaN** values

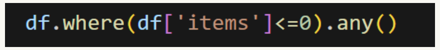


```
df.isna().any()
```

- Check **negative** values

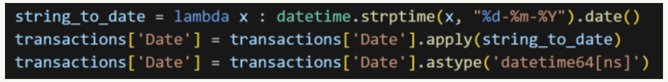


```
df.where(df['price']<0).any()
```



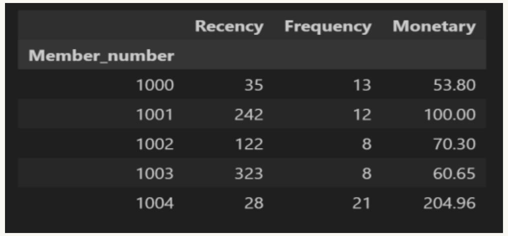
```
df.where(df['items']<=0).any()
```

- Change ‘Date’ to **datetime**



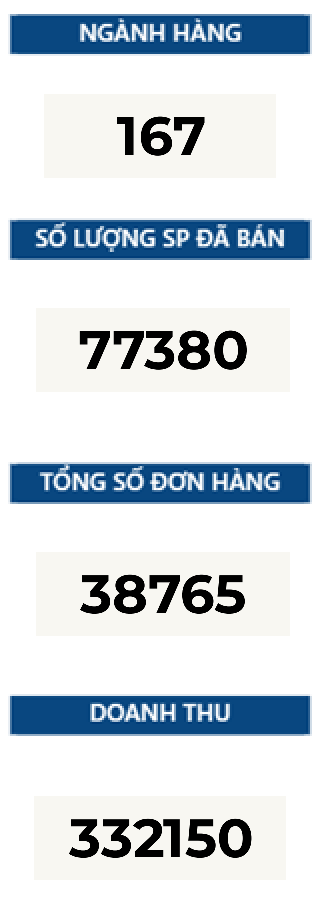
```
string_to_date = lambda x : datetime.strptime(x, "%d-%m-%Y").date()
transactions['Date'] = transactions['Date'].apply(string_to_date)
transactions['Date'] = transactions['Date'].astype('datetime64[ns]')
```

- Create ‘**df_RFM**’ with 3 cols ‘**Recency**’ , ‘**Frequency**’, ‘**Monetary**’ group by ‘Member_number’



	Recency	Frequency	Monetary
Member_number			
1000	35	13	53.80
1001	242	12	100.00
1002	122	8	70.30
1003	323	8	60.65
1004	28	21	204.96

Bussiness Overview



NGÀNH HÀNG

167

SỐ LƯỢNG SP ĐÃ BÁN

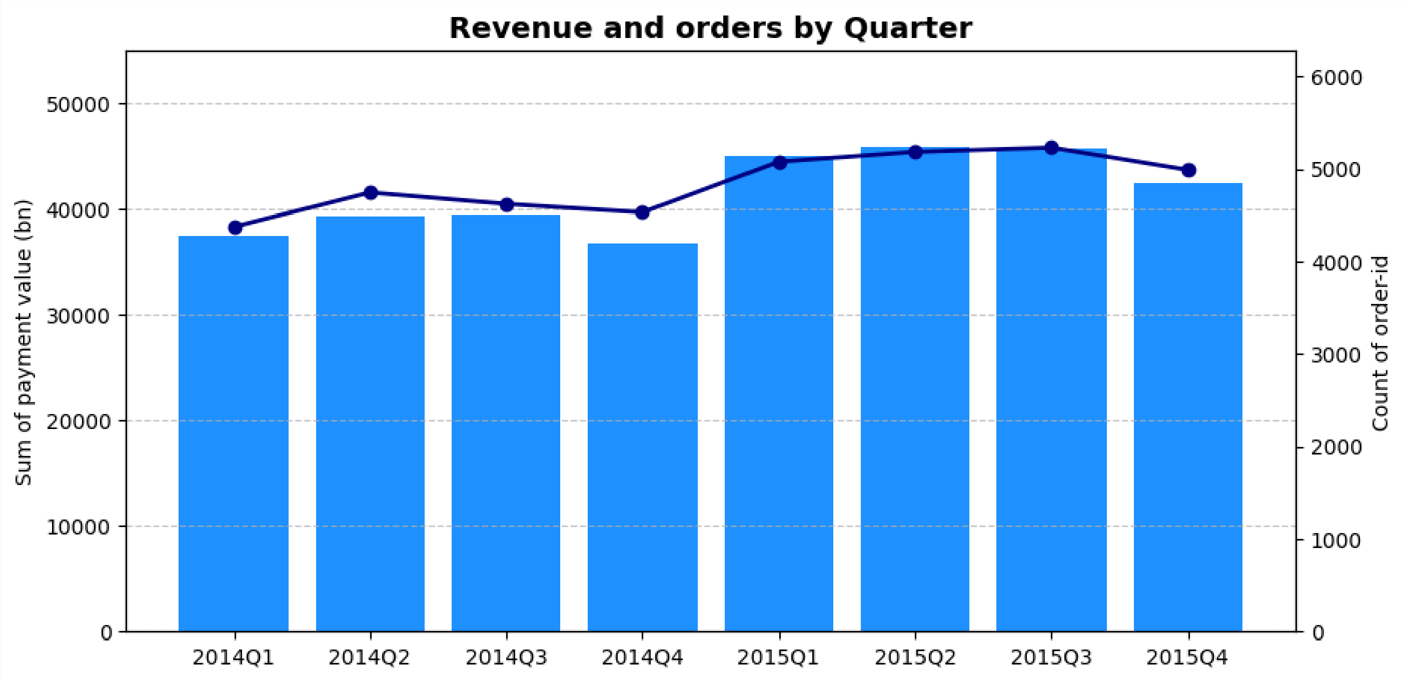
77380

TỔNG SỐ ĐƠN HÀNG

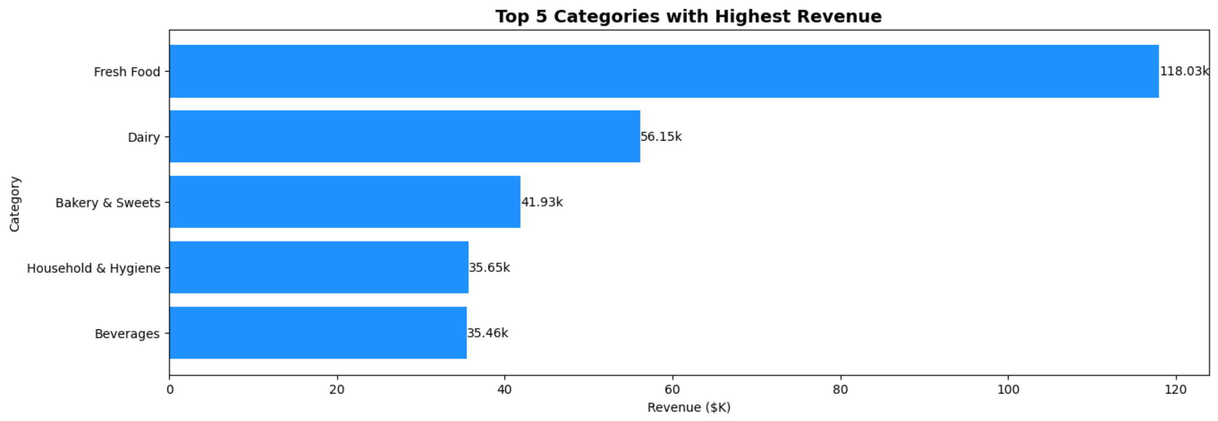
38765

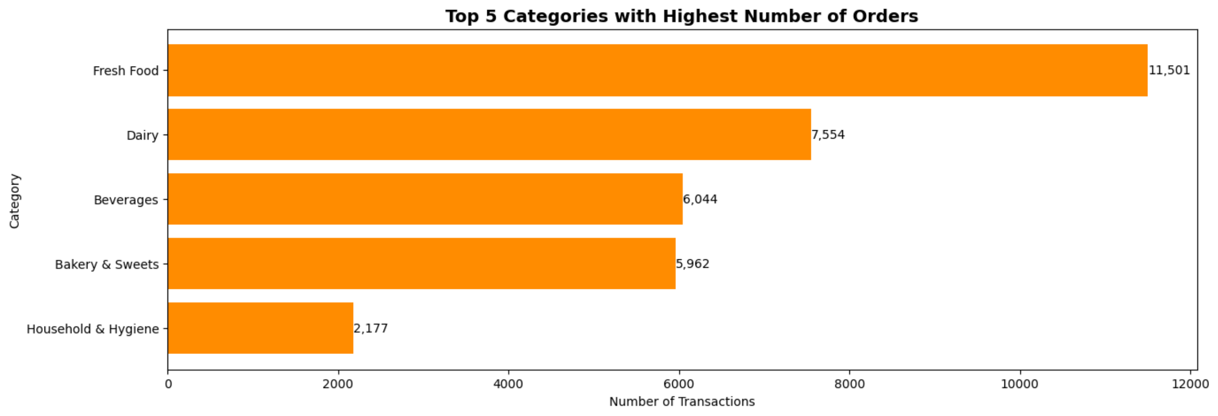
DOANH THU

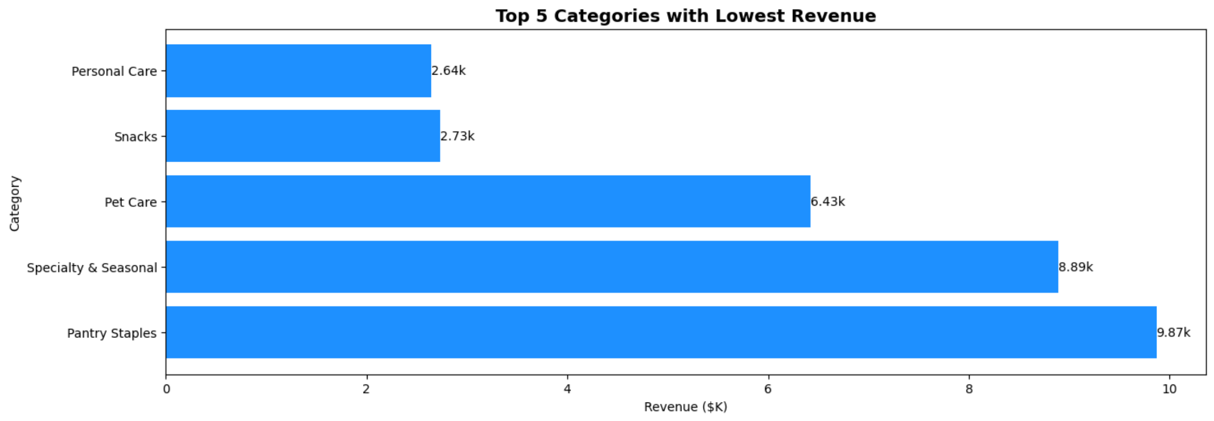
332150

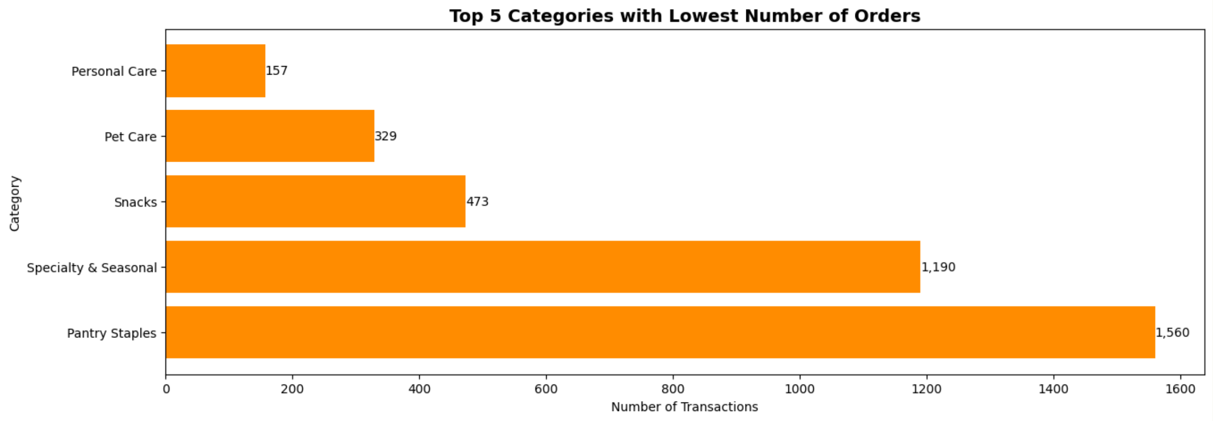


→ **Revenue and number of orders are proportional to each other**


Top 5 Categories with Highest Revenue

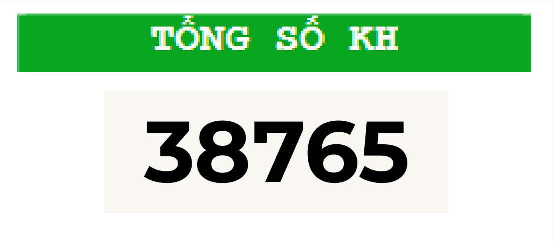

Top 5 Categories with Highest Number of Orders


Top 5 Categories with Lowest Revenue


Top 5 Categories with Lowest Number of Orders

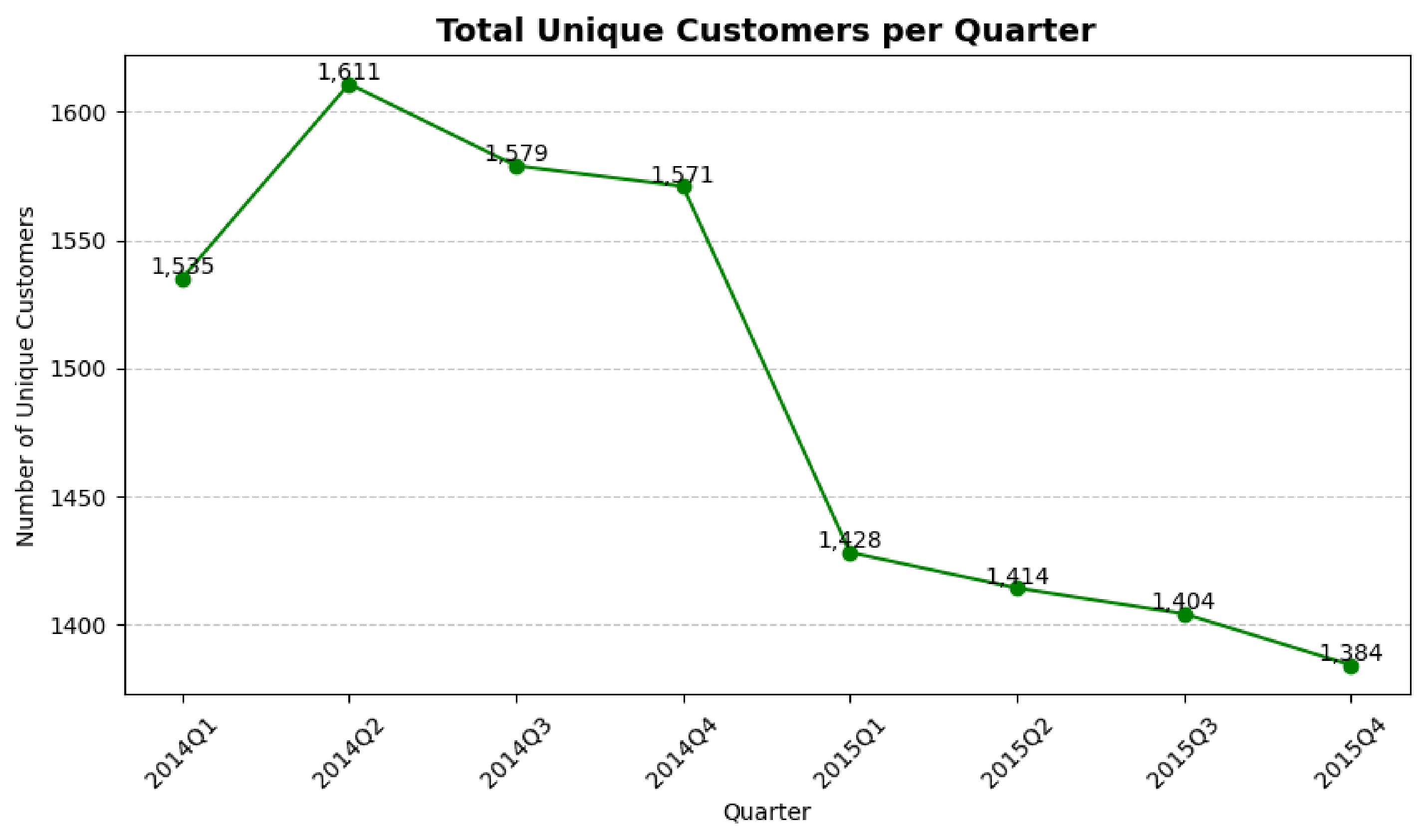
→ There are SIMILARITIES between categories when comparing the number of orders and revenue

Customer Analysis.

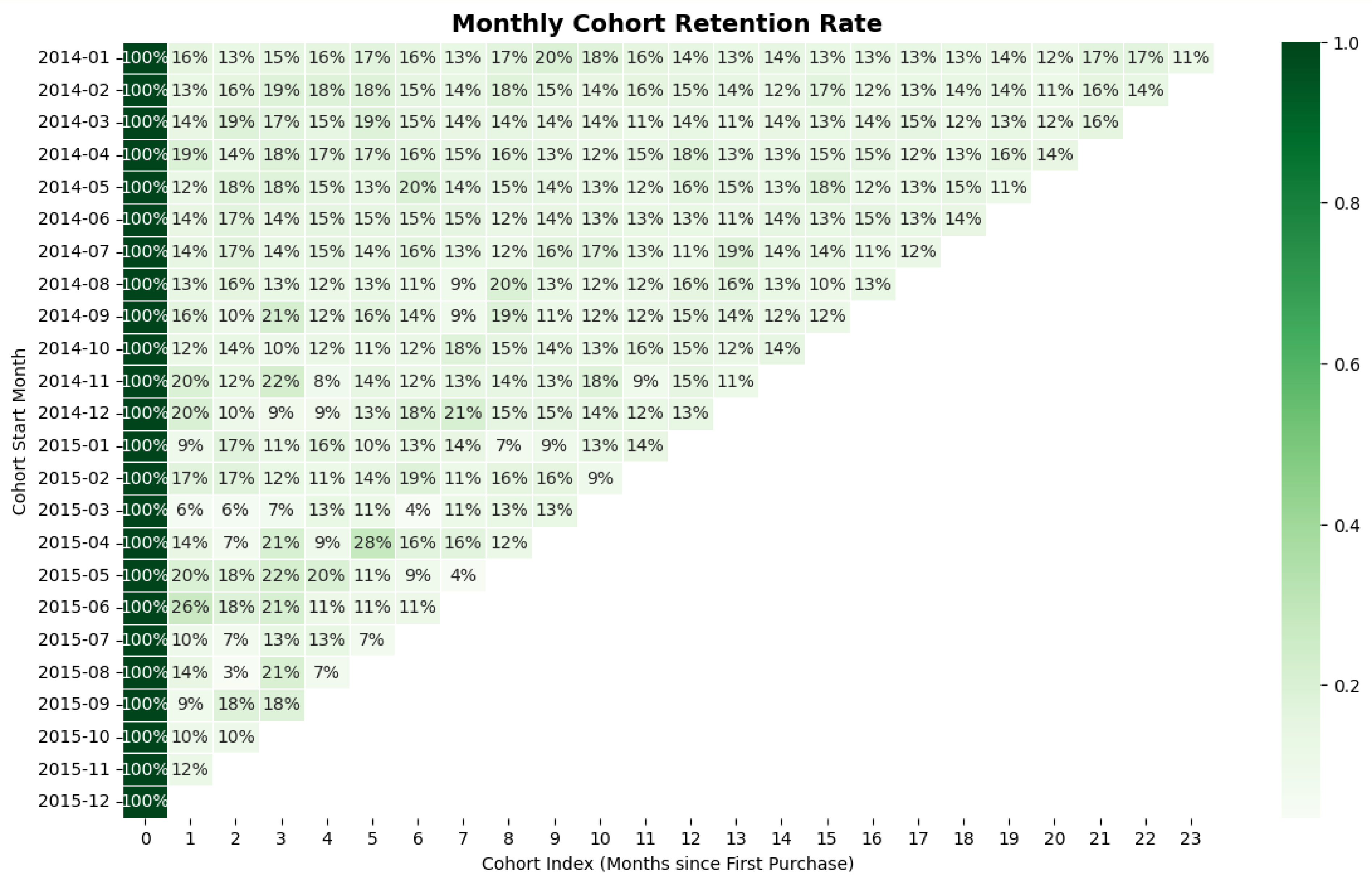


TỔNG SỐ KH

38765



Customer Analysis.



Monthly Cohort Retention Rate

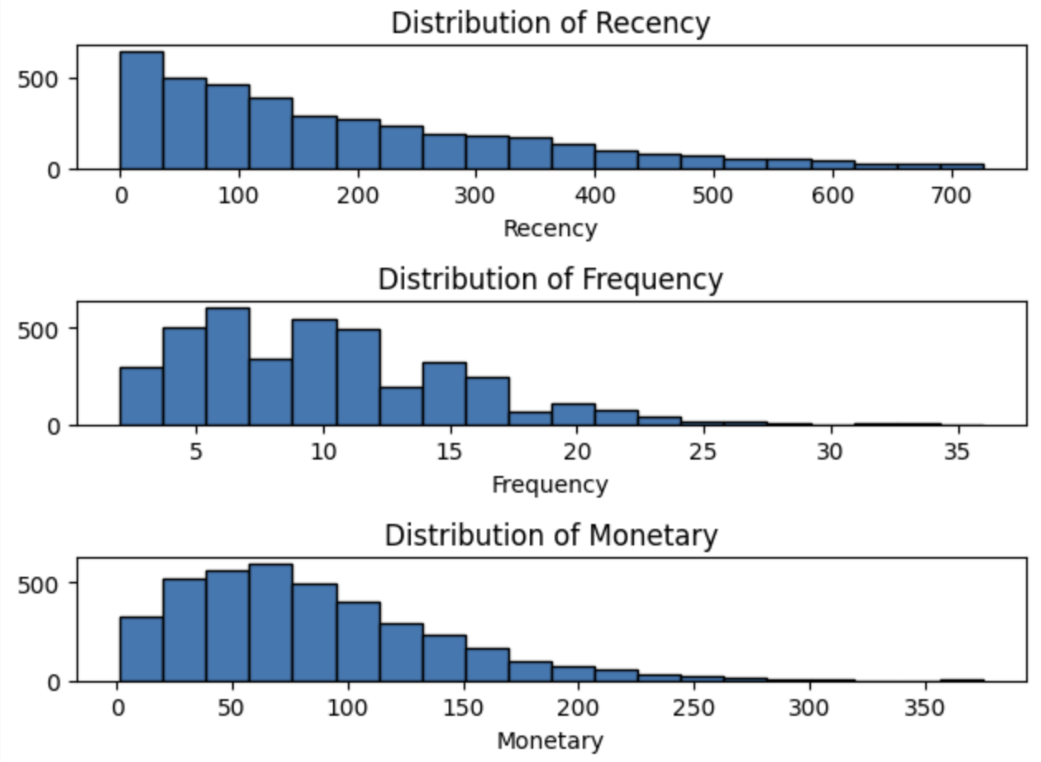
Cohort Start Month	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
2014-01	100%	16%	13%	15%	16%	17%	16%	13%	17%	20%	18%	16%	14%	13%	14%	13%	13%	13%	13%	14%	12%	17%	17%	11%
2014-02	100%	13%	16%	19%	18%	18%	15%	14%	18%	15%	14%	16%	15%	14%	12%	17%	12%	13%	14%	14%	11%	16%	14%	
2014-03	100%	14%	19%	17%	15%	19%	15%	14%	14%	14%	14%	11%	14%	11%	14%	13%	14%	15%	12%	13%	12%	16%		
2014-04	100%	19%	14%	18%	17%	17%	16%	15%	16%	13%	12%	15%	18%	13%	13%	15%	15%	12%	13%	16%	14%			
2014-05	100%	12%	18%	18%	15%	13%	20%	14%	15%	14%	13%	12%	16%	15%	13%	18%	12%	13%	15%	11%				
2014-06	100%	14%	17%	14%	15%	15%	15%	15%	12%	14%	13%	13%	13%	11%	14%	13%	15%	13%	14%					
2014-07	100%	14%	17%	14%	15%	14%	16%	13%	12%	16%	17%	13%	11%	19%	14%	14%	11%	12%						
2014-08	100%	13%	16%	13%	12%	13%	11%	9%	20%	13%	12%	12%	16%	16%	13%	10%	13%							
2014-09	100%	16%	10%	21%	12%	16%	14%	9%	19%	11%	12%	12%	15%	14%	12%	12%								
2014-10	100%	12%	14%	10%	12%	11%	12%	18%	15%	14%	13%	16%	15%	12%	14%									
2014-11	100%	20%	12%	22%	8%	14%	12%	13%	14%	13%	18%	9%	15%	11%										
2014-12	100%	20%	10%	9%	9%	13%	18%	21%	15%	15%	14%	12%	13%											
2015-01	100%	9%	17%	11%	16%	10%	13%	14%	7%	9%	13%	14%												
2015-02	100%	17%	17%	12%	11%	14%	19%	11%	16%	16%	9%													
2015-03	100%	6%	6%	7%	13%	11%	4%	11%	13%	13%														
2015-04	100%	14%	7%	21%	9%	28%	16%	16%	12%															
2015-05	100%	20%	18%	22%	20%	11%	9%	4%																
2015-06	100%	26%	18%	21%	11%	11%	11%																	
2015-07	100%	10%	7%	13%	13%	7%																		
2015-08	100%	14%	3%	21%	7%																			
2015-09	100%	9%	18%	18%																				
2015-10	100%	10%	10%																					
2015-11	100%	12%																						
2015-12	100%																							

Cohort Index (Months since First Purchase)

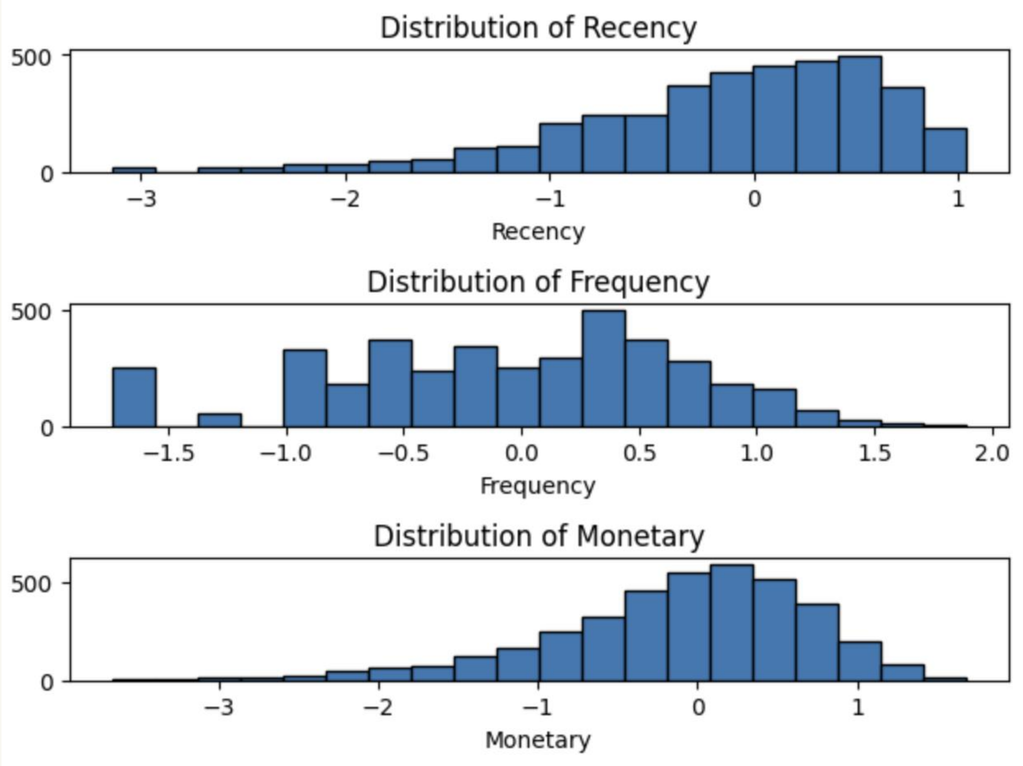
EDA

Scaling Data

All 3 columns Recency, Frequency and Monetary are **right skew** and not normal distribution --> Using **Log transformation** to reduce skewness

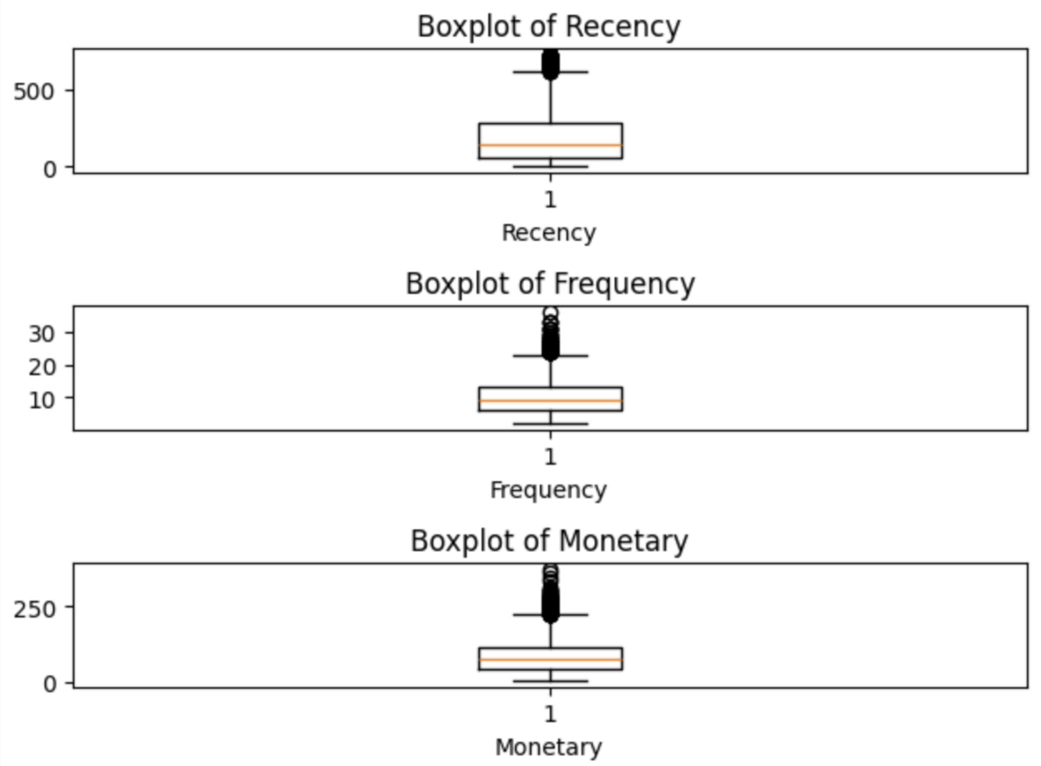


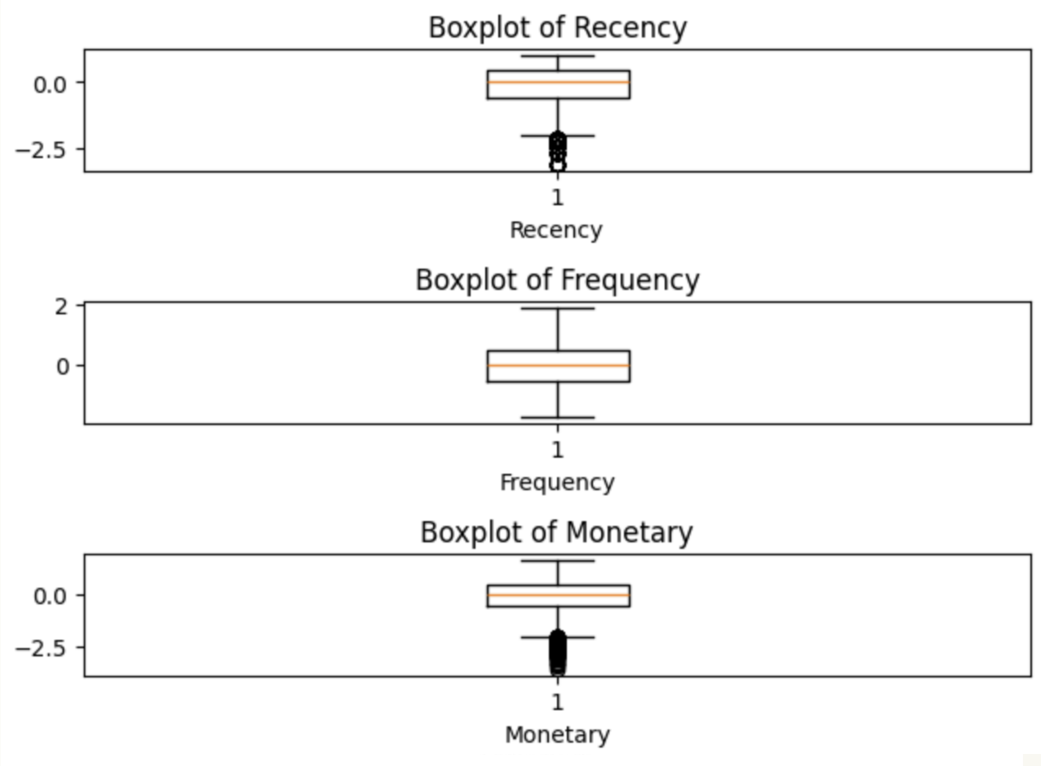
→



Scaling Data

All 3 columns Recency, Frequency and Monetary have many **upper outliers**
--> Using **Robust Scaler** to reduce the impact of outliers

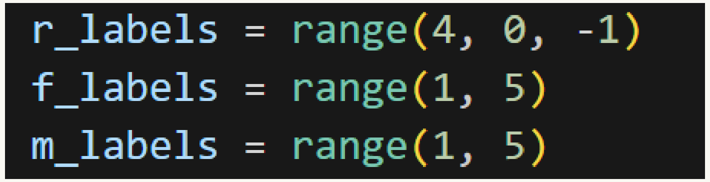




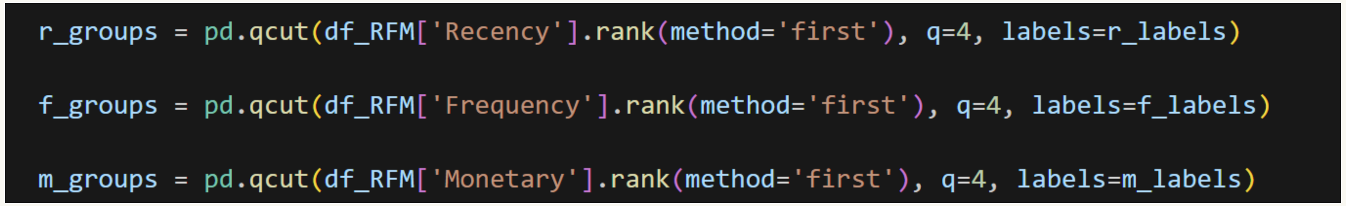
MODELS

Manual Segmentation.

- Compute **R, F and M score**

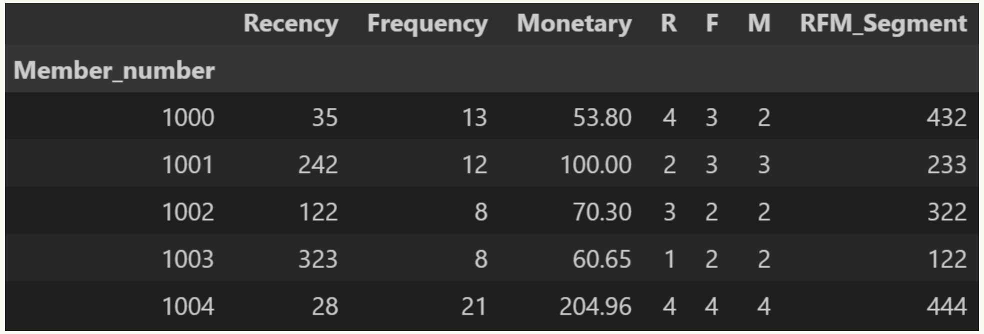


```
r_labels = range(4, 0, -1)
f_labels = range(1, 5)
m_labels = range(1, 5)
```



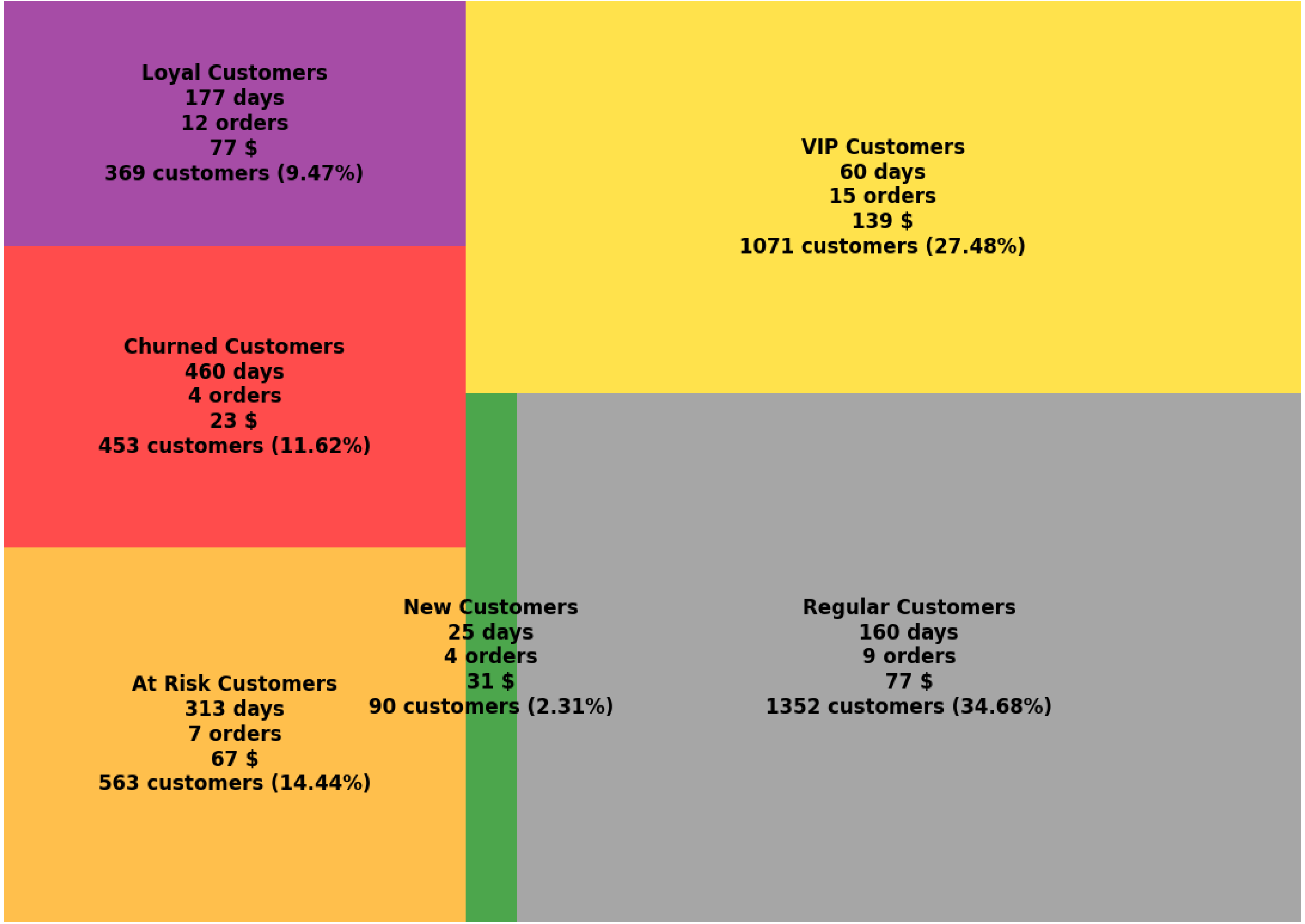
```
r_groups = pd.qcut(df_RFM['Recency'].rank(method='first'), q=4, labels=r_labels)

f_groups = pd.qcut(df_RFM['Frequency'].rank(method='first'), q=4, labels=f_labels)

m_groups = pd.qcut(df_RFM['Monetary'].rank(method='first'), q=4, labels=m_labels)
```

- Assign into df_RFM and concat to create '**RFM_segment**'

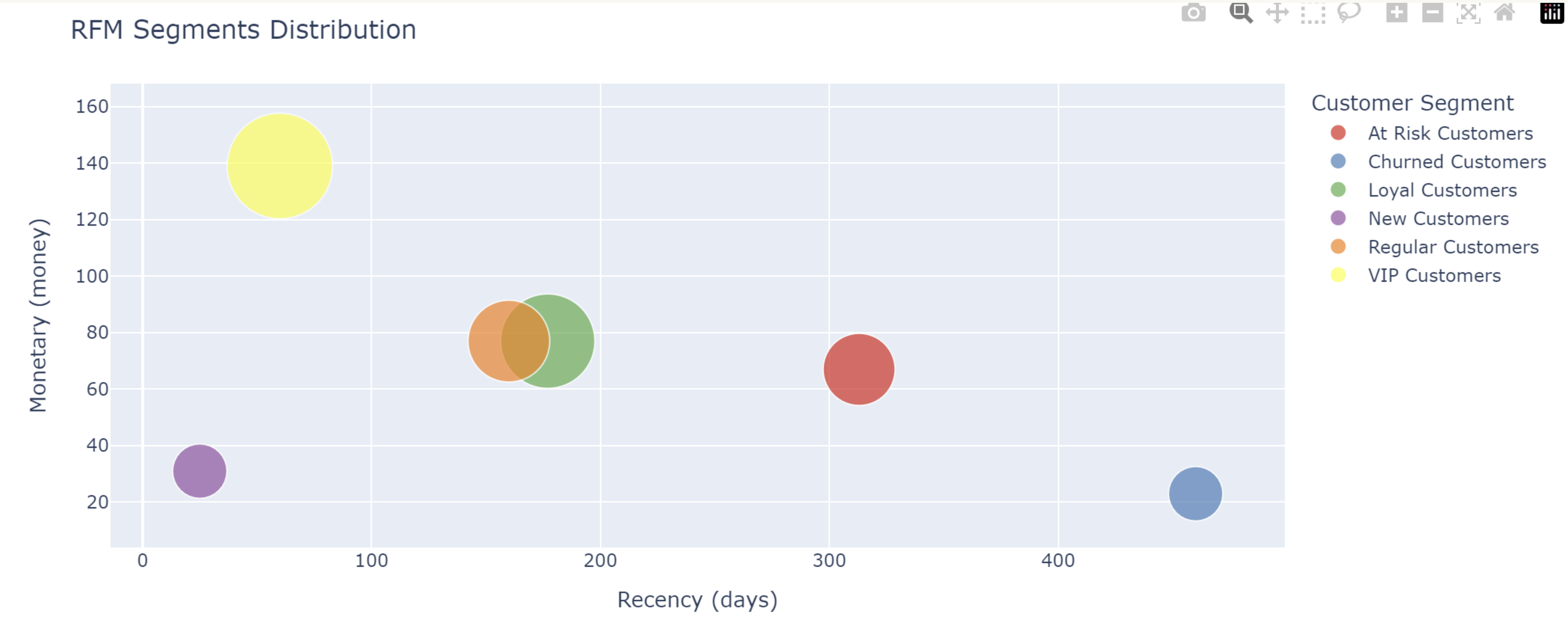


	Recency	Frequency	Monetary	R	F	M	RFM_Segment
Member_number							
1000	35	13	53.80	4	3	2	432
1001	242	12	100.00	2	3	3	233
1002	122	8	70.30	3	2	2	322
1003	323	8	60.65	1	2	2	122
1004	28	21	204.96	4	4	4	444

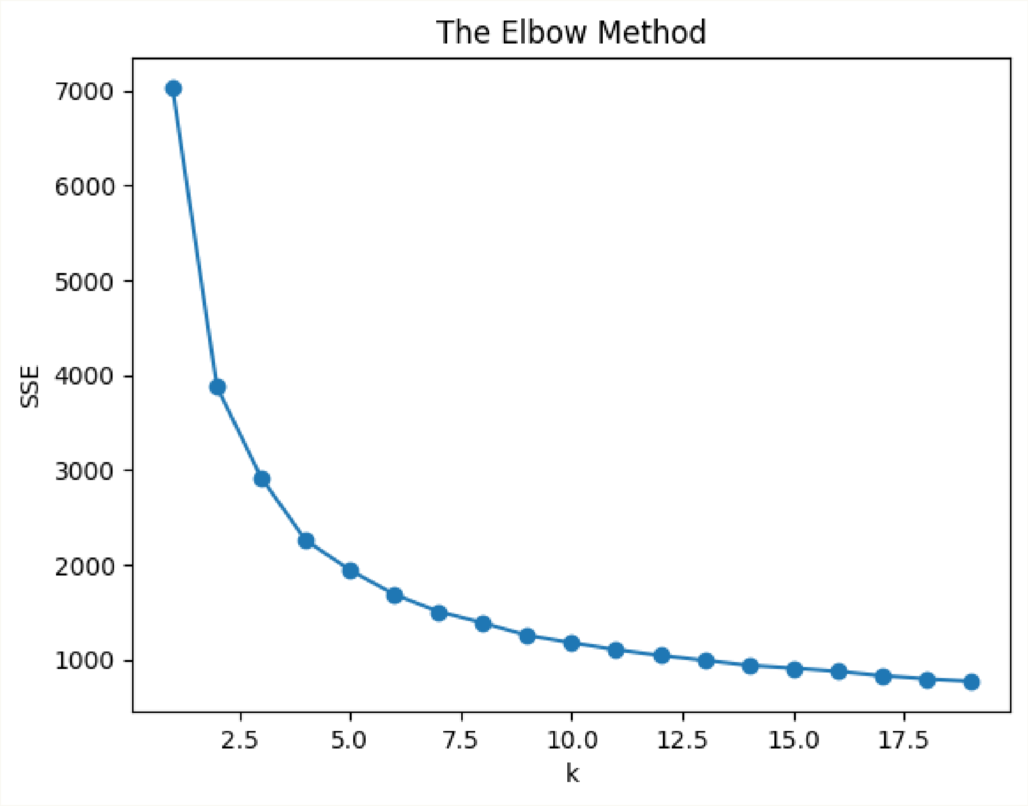
Customer Segments - RFM Analysis

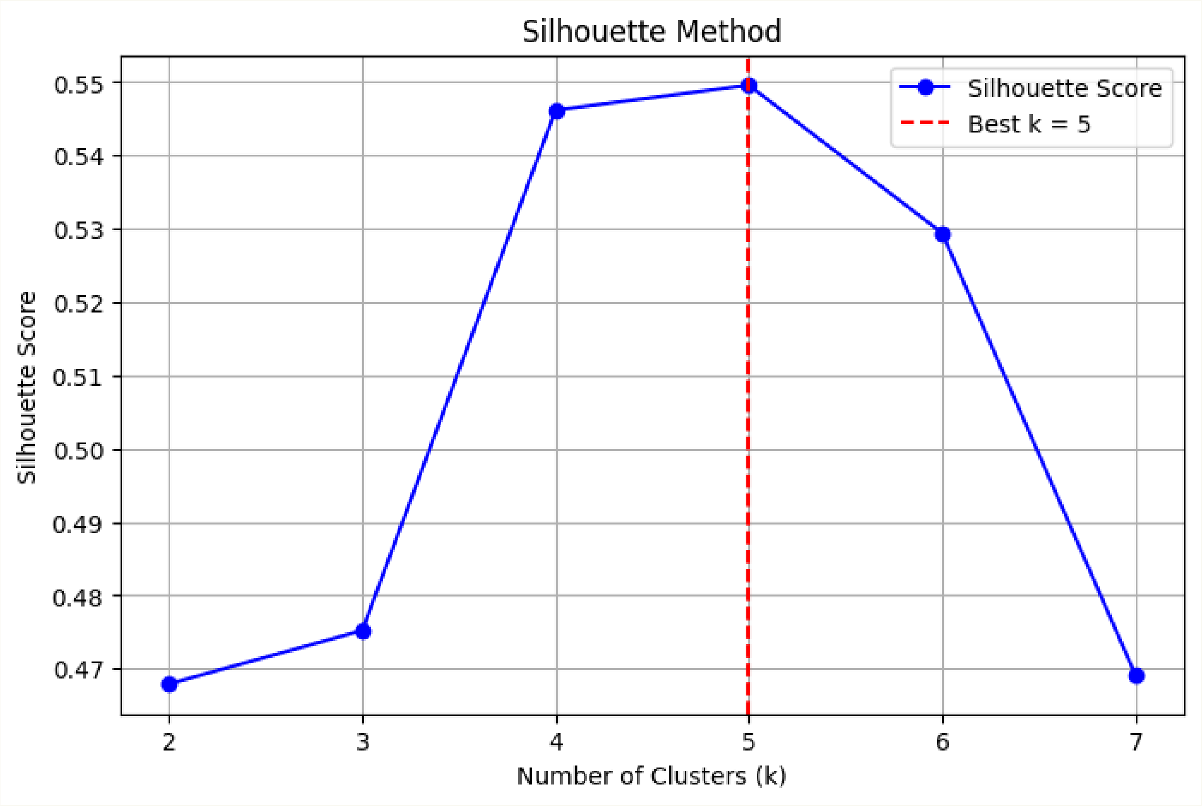


Customer Segments - RFM Analysis



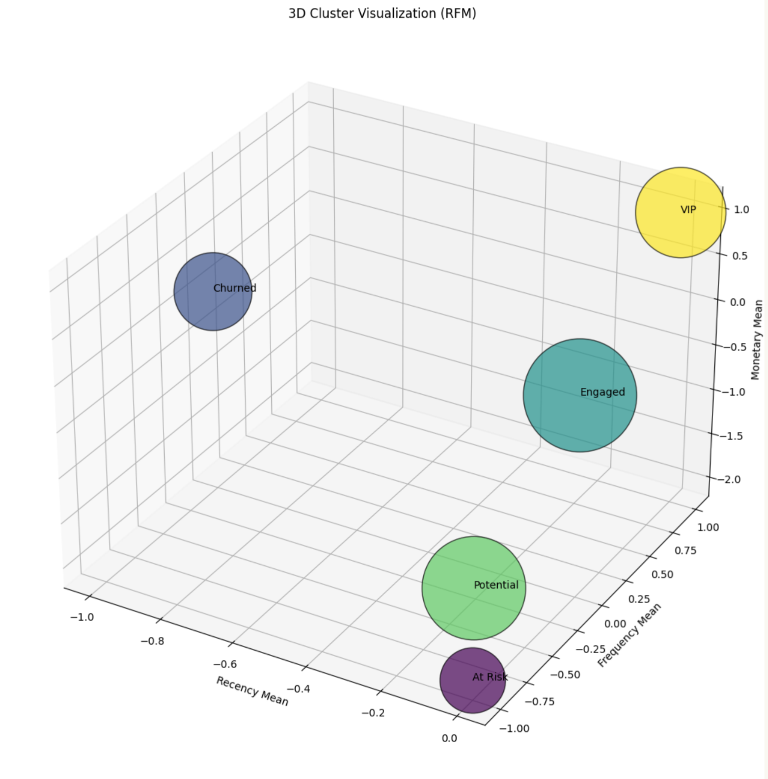
Kmeans with Scikit-learn

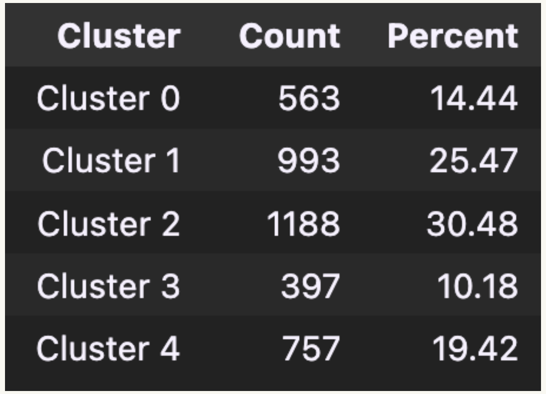




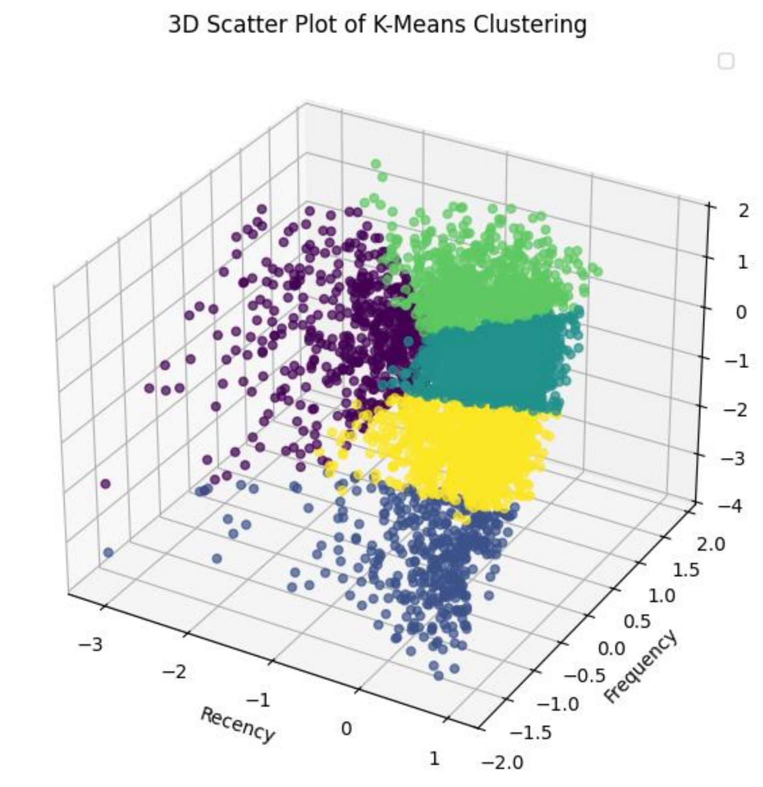
→ k = 5 is best for both methods

Kmeans with Scikit-learn

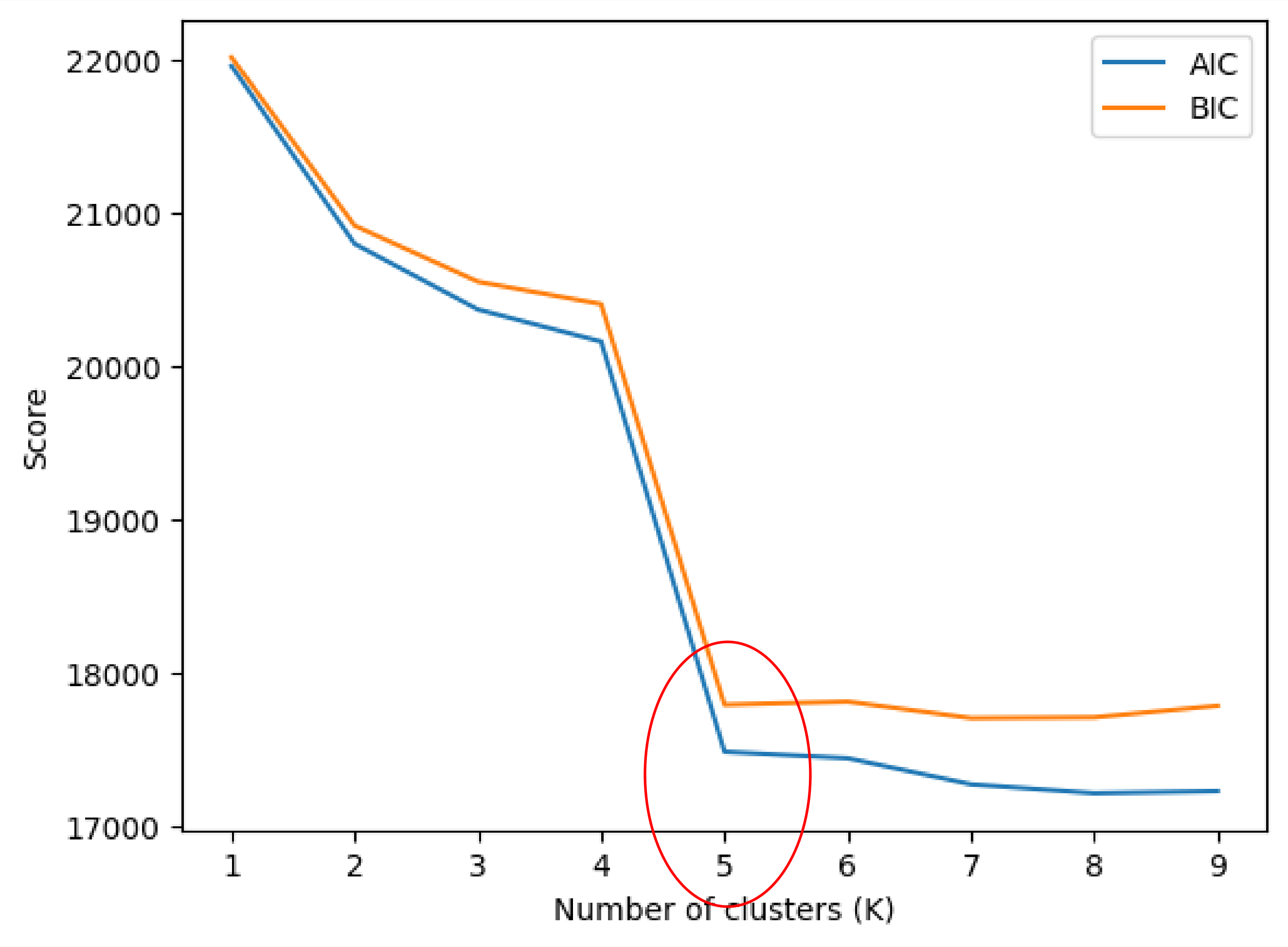




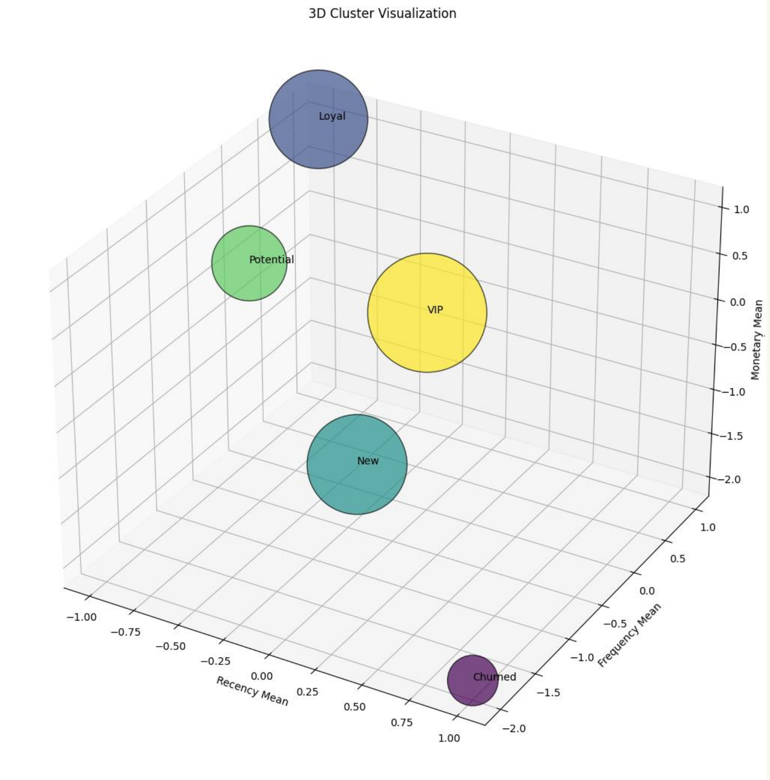
Cluster	Count	Percent
Cluster 0	563	14.44
Cluster 1	993	25.47
Cluster 2	1188	30.48
Cluster 3	397	10.18
Cluster 4	757	19.42

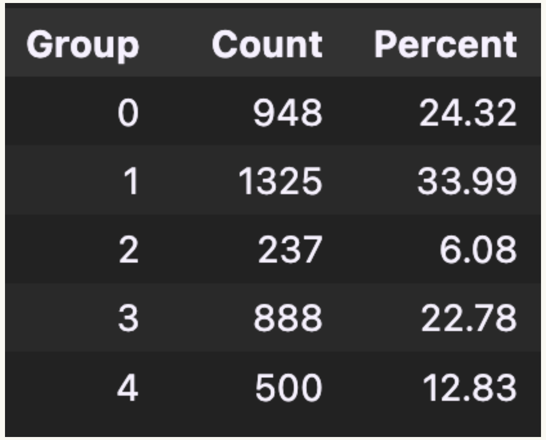


GMM

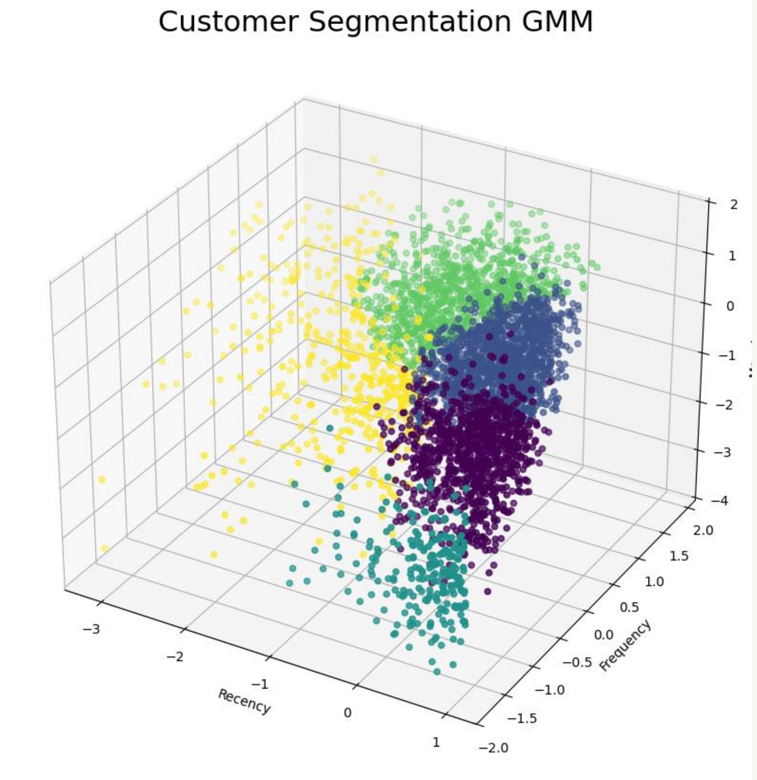

22000
21000
20000
19000
18000
17000
Score
1
2
3
4
5
6
7
8
9
Number of clusters (K)
AIC
BIC

GMM

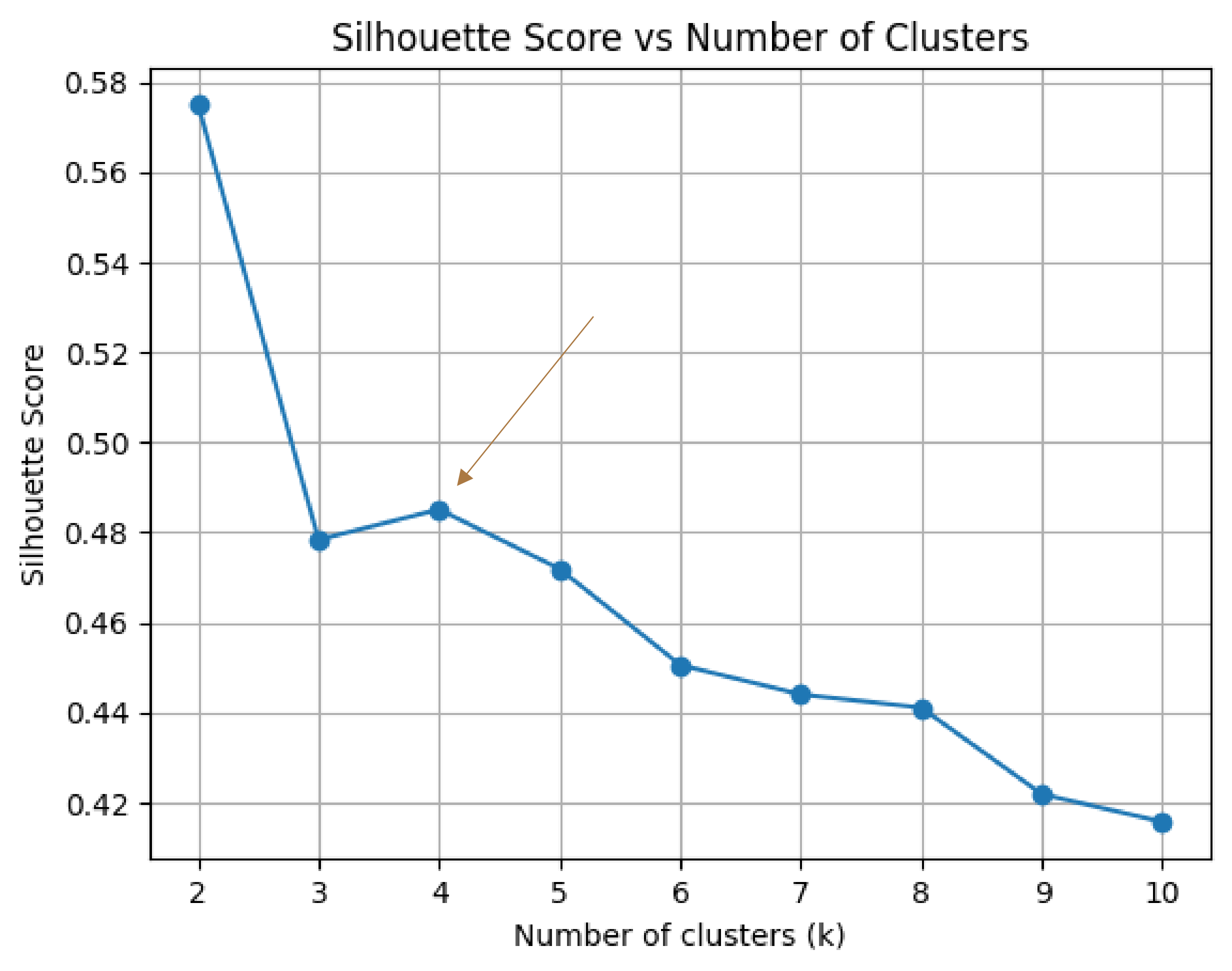

3D Cluster Visualization



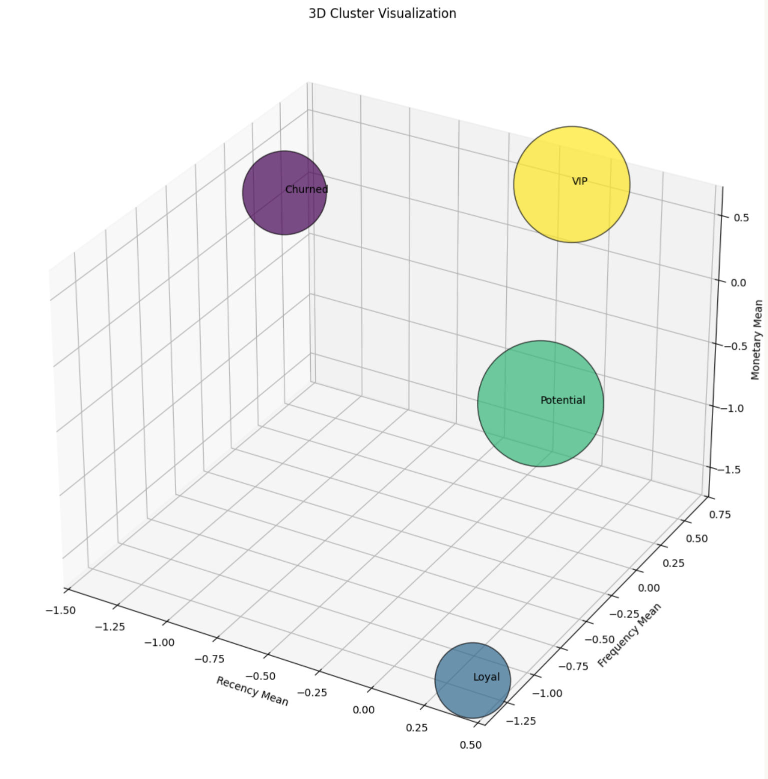
Group	Count	Percent
0	948	24.32
1	1325	33.99
2	237	6.08
3	888	22.78
4	500	12.83

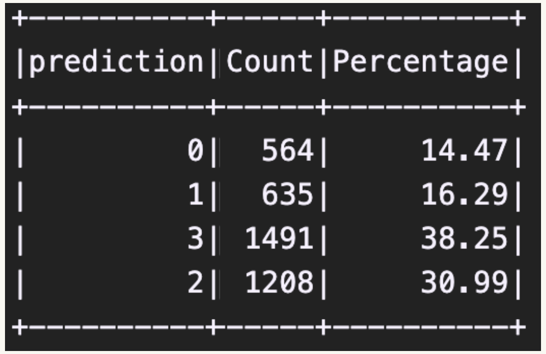

Customer Segmentation GMM

Kmeans pySpark

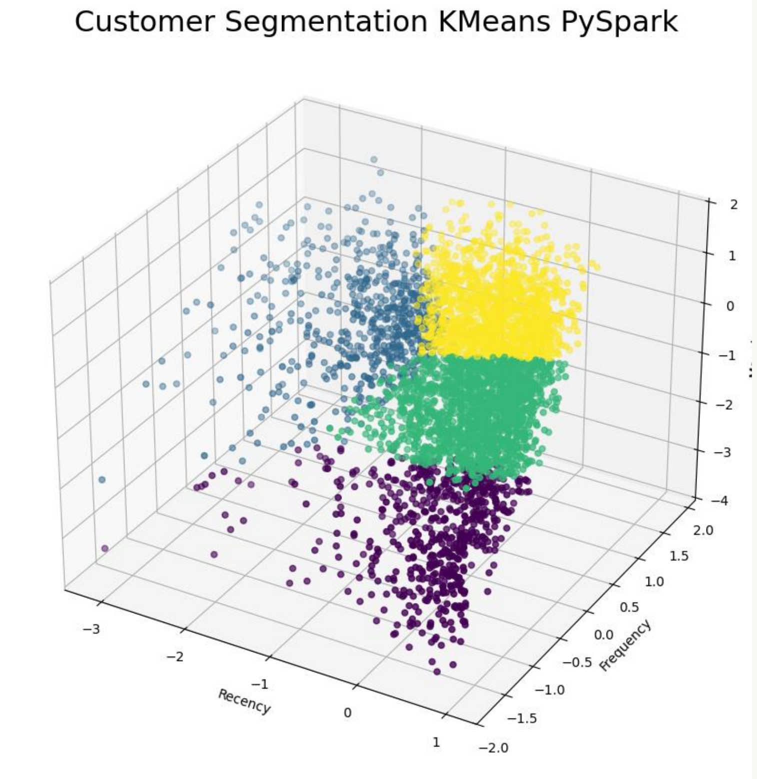


Kmeans pySpark





```
+----------+-----+----------+
|prediction|Count|Percentage|
+----------+-----+----------+
|         0|  564|     14.47|
|         1|  635|     16.29|
|         3| 1491|     38.25|
|         2| 1208|     30.99|
+----------+-----+----------+
```



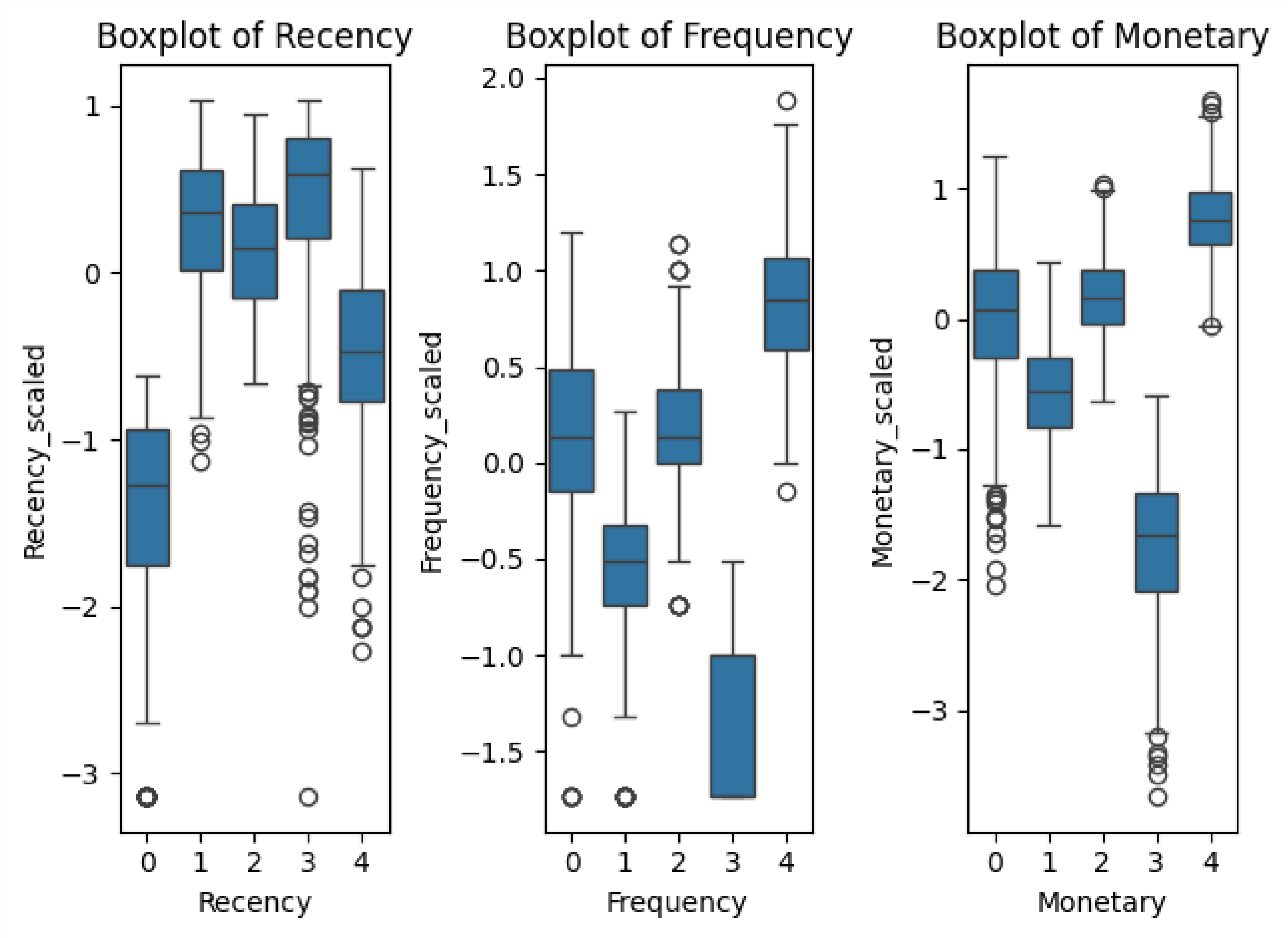
CONCLUSION

Conclusion

	Manual	K-Means Scikit	K-Means Pyspark	GMM
k	6	5	4	5

→ Choosing k = 5 for the final customer segmentation

Conclusion


Boxplot of Recency
Recency_scaled
1
0
−1
−2
−3
0 1 2 3 4
Recency
Boxplot of Frequency
Frequency_scaled
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0 1 2 3 4
Frequency
Boxplot of Monetary
Monetary_scaled
1
0
−1
−2
−3
0 1 2 3 4
Monetary

Group	Cluster	Number	% Revenue	Note
VIP	4	757	19.42%	Highest value customer group, frequent shoppers and big spenders. Need special offers and personalized care.
Engaged	2	1188	30.48%	Customers with stable interactions, frequent shopping. Can be retained with membership programs and promotions.
Potential	1	993	25.47%	Potential customer group, can be developed into loyal customers with appropriate incentive strategies.
At risk	3	397	10.18%	Customers are decreasing in interaction, at risk of leaving. Need reactivation strategies such as reminder emails, special offers.
Churned	0	563	14.44%	Customers who have almost left, difficult to attract back. Need to consider a strong discount campaign or remove from the marketing list.

Project Learnings.

1. Clear task delegation helps the team save time and optimize work.
2. Understanding customer insights → Building an appropriate model.
3. EDA determines the quality of clustering

   → Selecting the right variables is crucial.
4. The RFM method is very useful for customer segmentation

   → Developing strategies tailored to each group.

Division of work

	Thiên Bảo	Anh Khoa
Data preparation		x
Several information		x
Manual Segmentation		x
K-Means with Scikit-learn		x
GMM	x	
K-Means PySpark	x	
Conclusion	x	x
PowerPoint	x	
Steamlit	x	x
Presentation	x	x

About our team

Hồ Nguyễn Thiên Bảo

MEET

Technician
Biotechnology - Microbiology

THE

Nguyễn Anh Khoa

TEAM

Senior student
Data Science

Learnings from DL07

- Handling unexpected situations during project execution, such as working with big data and unmet hardware requirements.
- Understanding necessary concepts like RFM for Customer Segmentation and Content-based and Collaborative Filtering for Recommendation Systems.
- Working with Streamlit - enabling the rapid and simple creation of user interfaces.

Learnings from teamwork

- Effective communication is crucial for efficient teamwork.
- Knowing how to manage personal time and take responsibility for assigned tasks.
- Clearly defining the collaborative work of each individual within the team.

THANKS FOR YOUR KIND ATTENTION!