PROJECT REPORT*

CUSTOMER SEGMENTATION

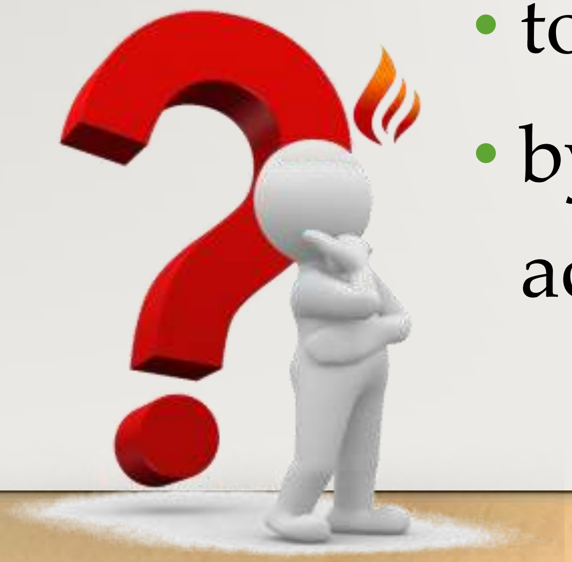
GOKAY BULUT

**Note: Since the company owns all intellectual property rights, this is not the original presentation provided to the company. Rather, it is a report explaining my work.*

AGENDA

- Introduction
- Data Wrangling & EDA
- Cohort Analysis
- RFMT Analysis
- K-Means Clustering
- Conclusion

Introduction



Problem:

- group customers into segments
- to understand high level trends better
- by providing insights on metrics
  across product / service and customer lifecycle.

Introduction

Data set:

- 98,572 rows of customer transactions,
- provided by the company.

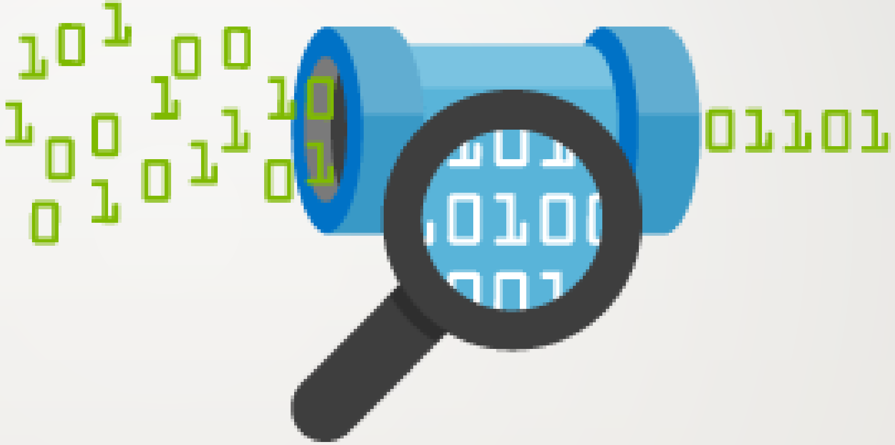
Introduction

Features:

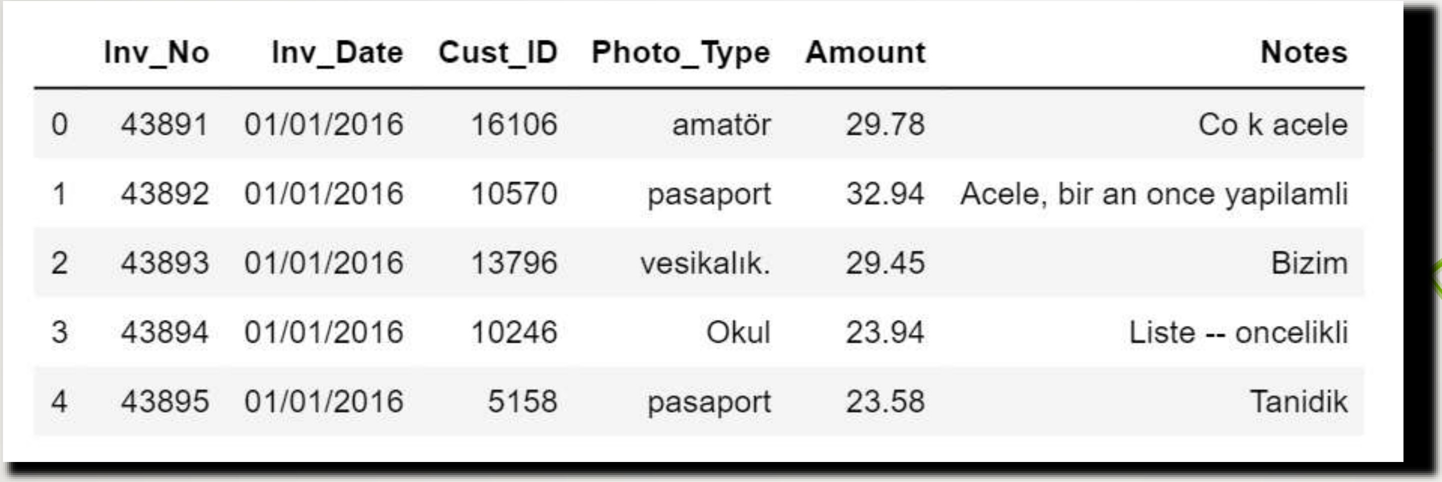
- 'Inv_No' - number for each transaction (integer).
- 'Inv_Date' - time of the transaction (string).
- 'name’ - name of the customer (string).
- 'lastname' - lastname of the customer (string).
- 'Cust_ID' - unique number for identifying the customers (integer).
- 'Photo_Type' - types of photos taken (string).
- 'Amount' - amount paid by the customer for the transaction (float).
- 'Notes’ - notes on the transaction (string).

AGENDA

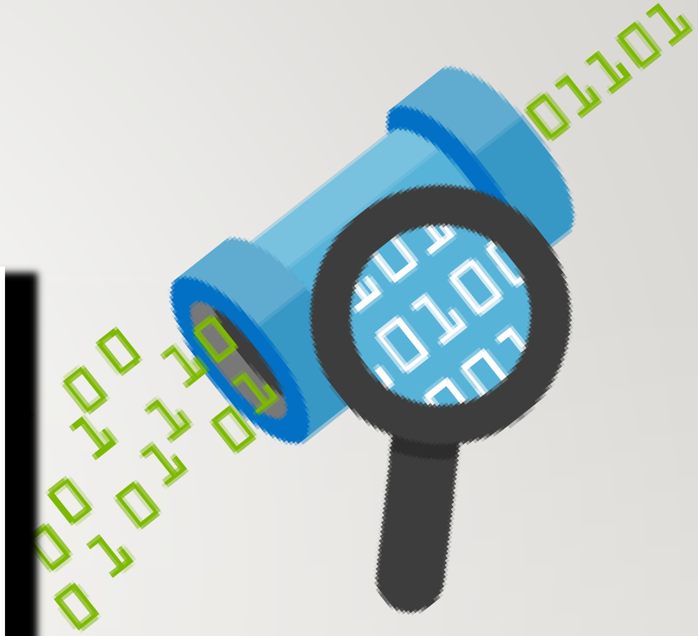
- Introduction
- Data Wrangling & EDA
- Cohort Analysis
- RFMT Analysis
- K-Means Clustering
- Conclusion



Data Wrangling & EDA



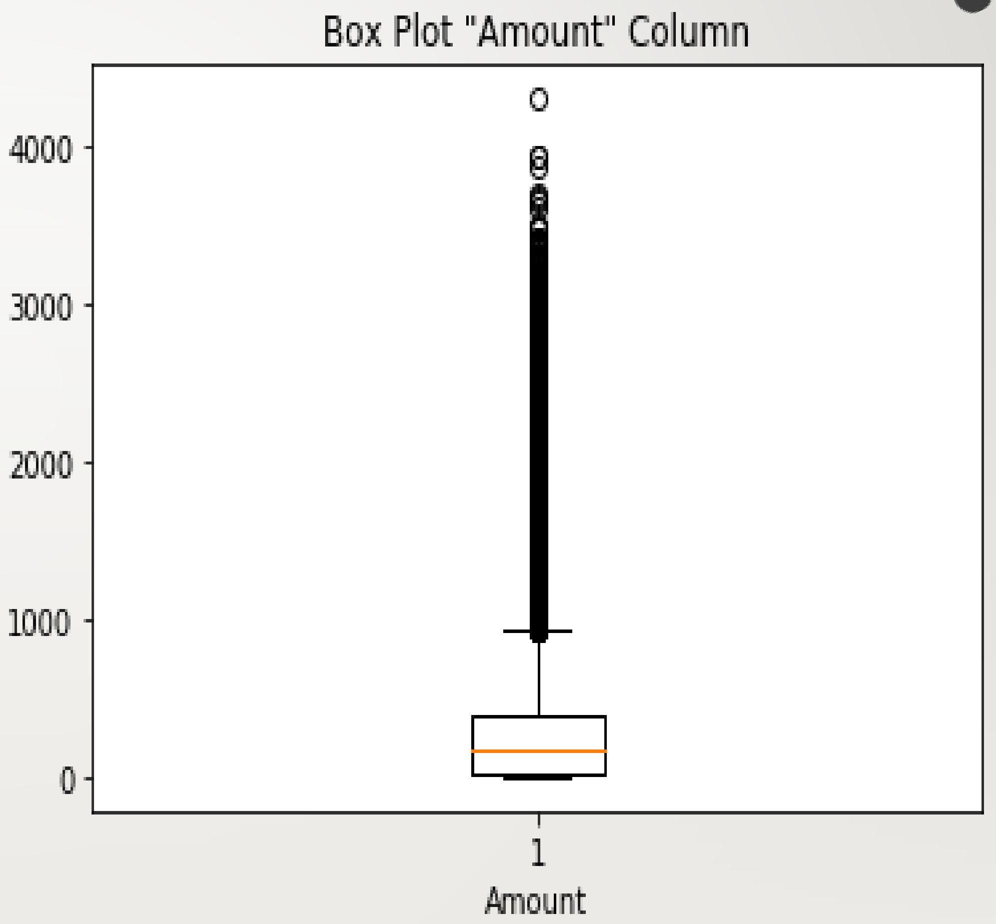
	Inv_No	Inv_Date	Cust_ID	Photo_Type	Amount	Notes
0	43891	01/01/2016	16106	amatör	29.78	Co k acele
1	43892	01/01/2016	10570	pasaport	32.94	Acele, bir an once yapilamli
2	43893	01/01/2016	13796	vesikalık.	29.45	Bizim
3	43894	01/01/2016	10246	Okul	23.94	Liste -- oncelikli
4	43895	01/01/2016	5158	pasaport	23.58	Tanidik



- Dropped redundant columns & rows,
- 'Photo_Type' column cleaned (lowercased & stripped dots),
- ‘Inv_Date' column → datetime type for better analysis,

Data Wrangling & EDA

- Sales ‘Amount’ mean = 331.2.
- Returning customers.
- Invoice numbers are unique.
- Invoice dates cover 1 year.
- 7 unique photo types.

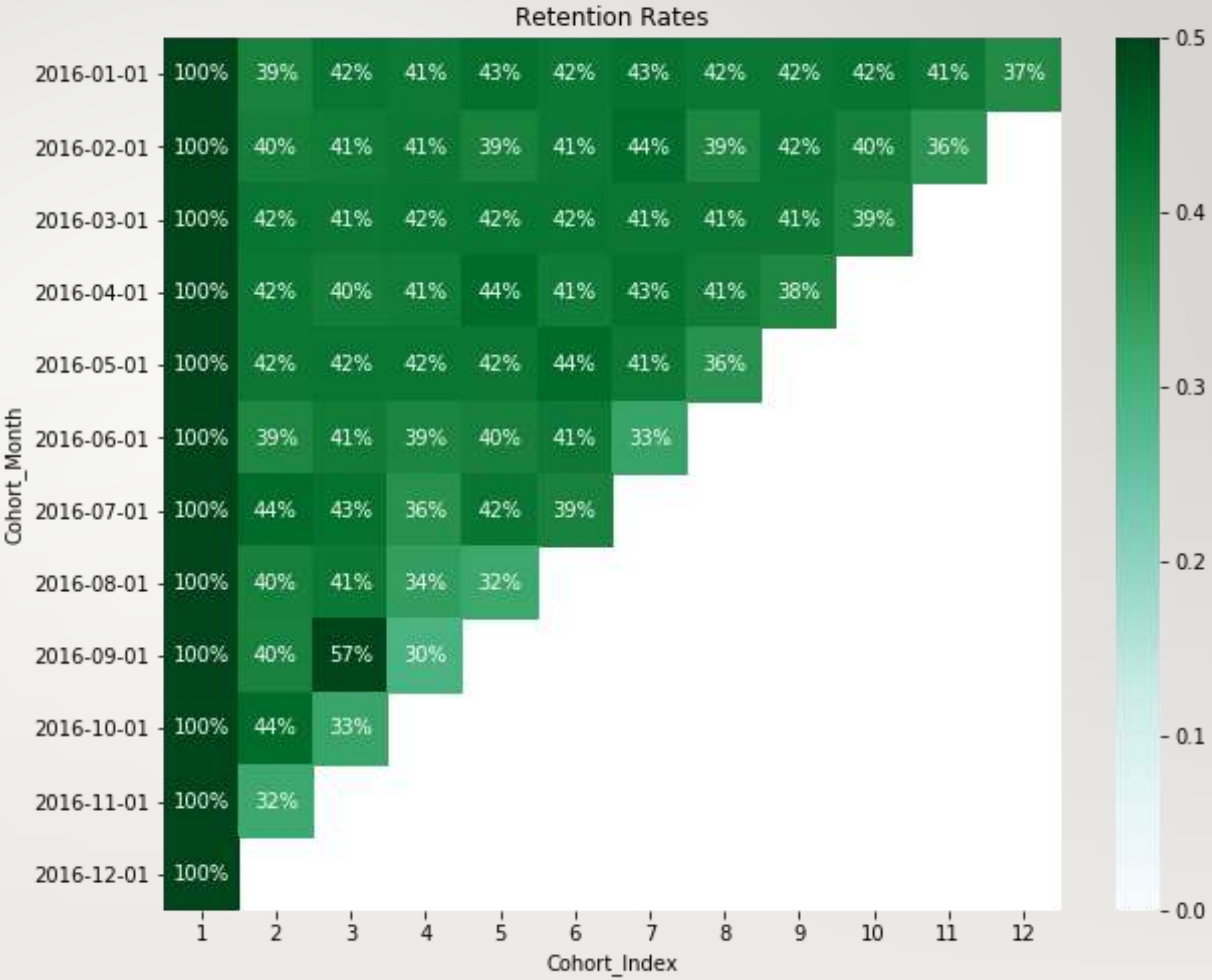


AGENDA

- Introduction
- Data Wrangling & EDA
- Cohort Analysis
- RFMT Analysis
- K-Means Clustering
- Conclusion

Retention

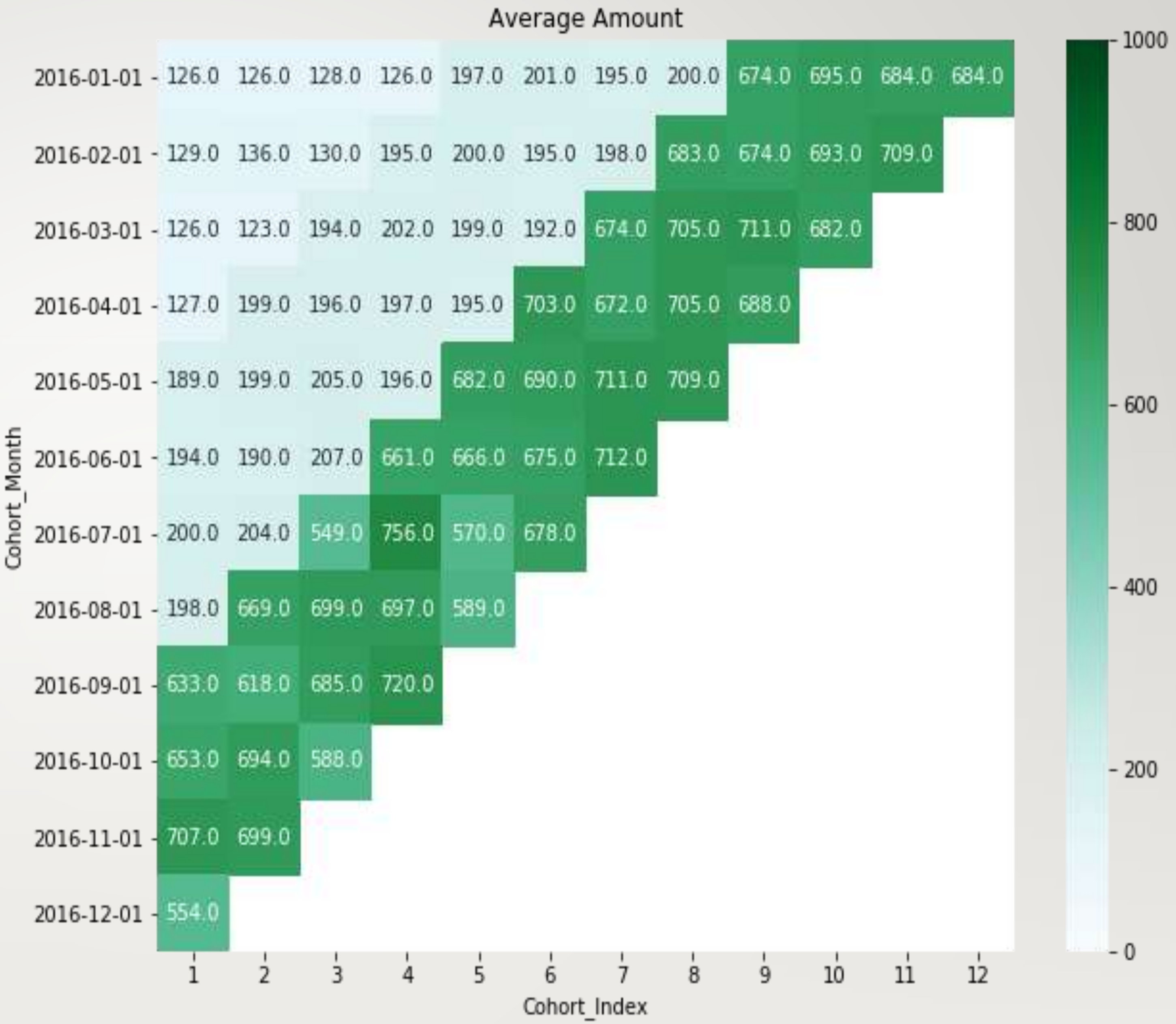
- While customer retention rate remains relatively constant around 40%,



Cohort_Month	1	2	3	4	5	6	7	8	9	10	11	12
2016-01-01	100%	39%	42%	41%	43%	42%	43%	42%	42%	42%	41%	37%
2016-02-01	100%	40%	41%	41%	39%	41%	44%	39%	42%	40%	36%	
2016-03-01	100%	42%	41%	42%	42%	42%	41%	41%	41%	39%		
2016-04-01	100%	42%	40%	41%	44%	41%	43%	41%	38%			
2016-05-01	100%	42%	42%	42%	42%	44%	41%	36%				
2016-06-01	100%	39%	41%	39%	40%	41%	33%					
2016-07-01	100%	44%	43%	36%	42%	39%						
2016-08-01	100%	40%	41%	34%	32%							
2016-09-01	100%	40%	57%	30%								
2016-10-01	100%	44%	33%									
2016-11-01	100%	32%										
2016-12-01	100%											

Average Sales

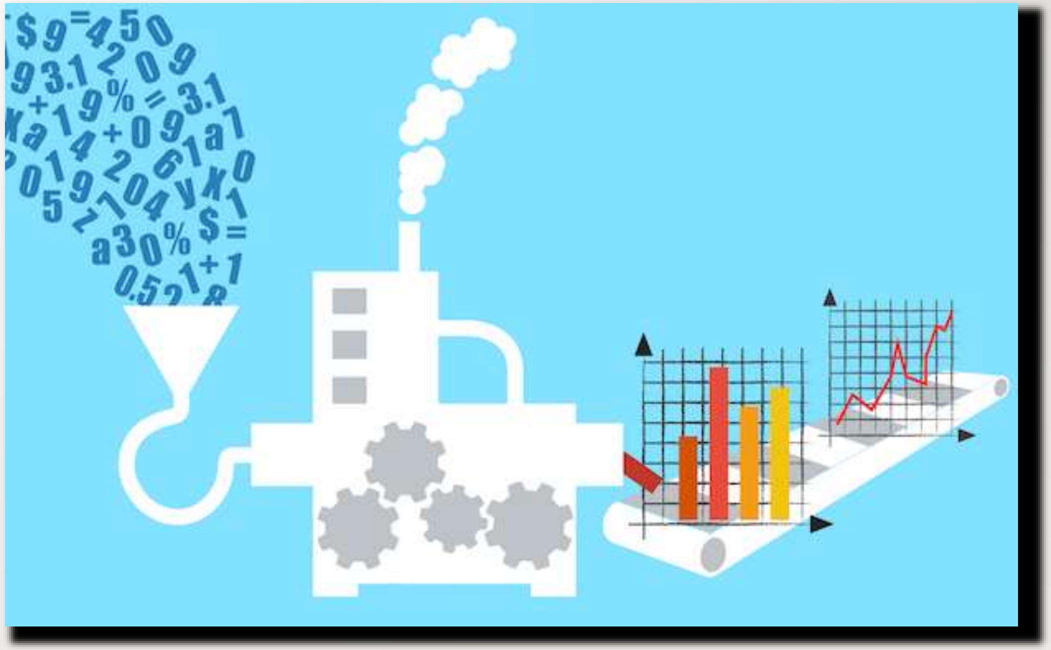
- There is an increase in sales towards Christmas.



	1	2	3	4	5	6	7	8	9	10	11	12
2016-01-01	126.0	126.0	128.0	126.0	197.0	201.0	195.0	200.0	674.0	695.0	684.0	684.0
2016-02-01	129.0	136.0	130.0	195.0	200.0	195.0	198.0	683.0	674.0	693.0	709.0	
2016-03-01	126.0	123.0	194.0	202.0	199.0	192.0	674.0	705.0	711.0	682.0		
2016-04-01	127.0	199.0	196.0	197.0	195.0	703.0	672.0	705.0	688.0			
2016-05-01	189.0	199.0	205.0	196.0	682.0	690.0	711.0	709.0				
2016-06-01	194.0	190.0	207.0	661.0	666.0	675.0	712.0					
2016-07-01	200.0	204.0	549.0	756.0	570.0	678.0						
2016-08-01	198.0	669.0	699.0	697.0	589.0							
2016-09-01	633.0	618.0	685.0	720.0								
2016-10-01	653.0	694.0	588.0									
2016-11-01	707.0	699.0										
2016-12-01	554.0											

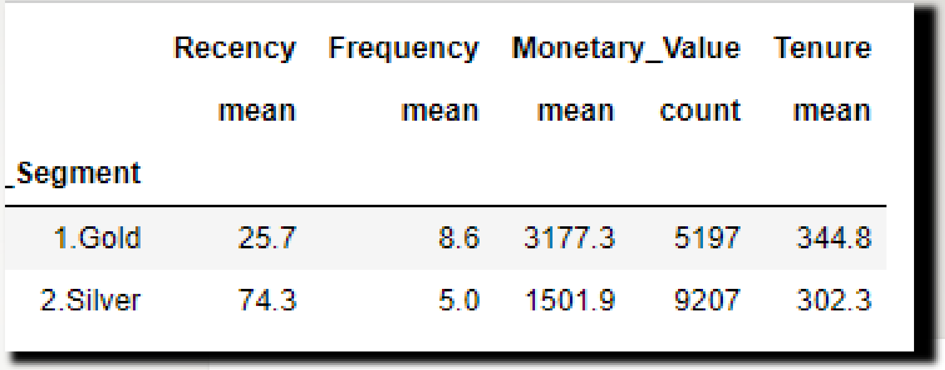
AGENDA

- Introduction
- Data Wrangling & EDA
- Cohort Analysis
- RFMT Analysis
- K-Means Clustering
- Conclusion

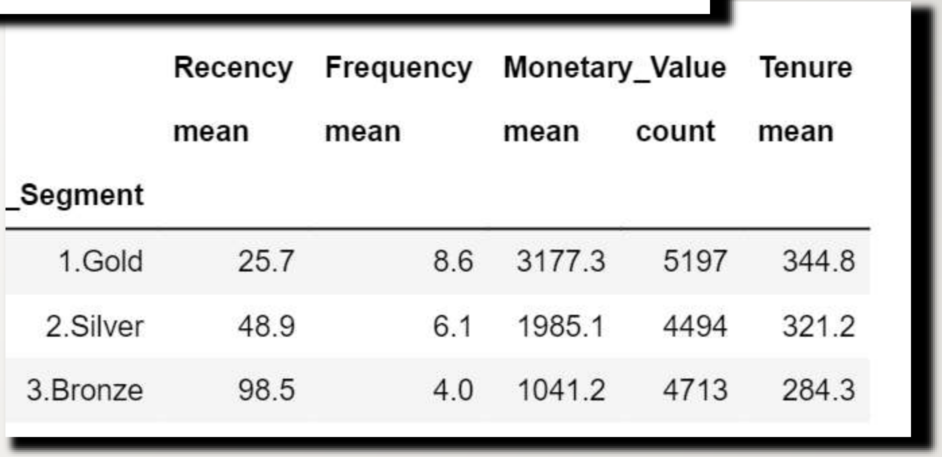


RFMT Analysis

- Recency → how recent
- Frequency → how many times
- Monetary Value → how much
- Tenure → for how long



	Recency	Frequency	Monetary_Value		Tenure
	mean	mean	mean	count	mean
_Segment					
1.Gold	25.7	8.6	3177.3	5197	344.8
2.Silver	74.3	5.0	1501.9	9207	302.3



	Recency	Frequency	Monetary_Value		Tenure
	mean	mean	mean	count	mean
_Segment					
1.Gold	25.7	8.6	3177.3	5197	344.8
2.Silver	48.9	6.1	1985.1	4494	321.2
3.Bronze	98.5	4.0	1041.2	4713	284.3

- segment 1 remained in both.
- segment 2 → divided in 2

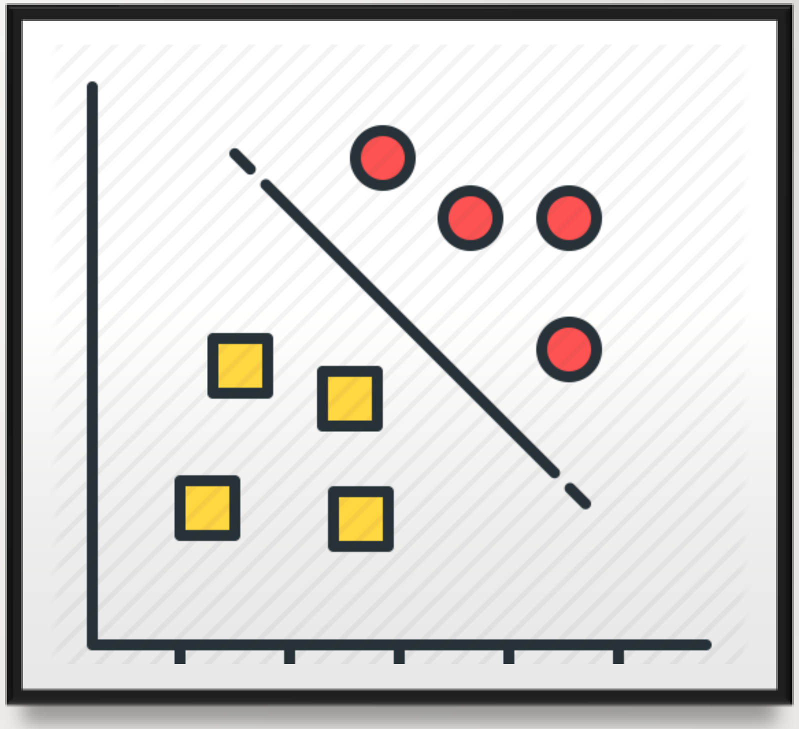
AGENDA

- Introduction
- Data Wrangling & EDA
- Cohort Analysis
- RFMT Analysis
- K-Means Clustering
- Conclusion

K-Means Clustering

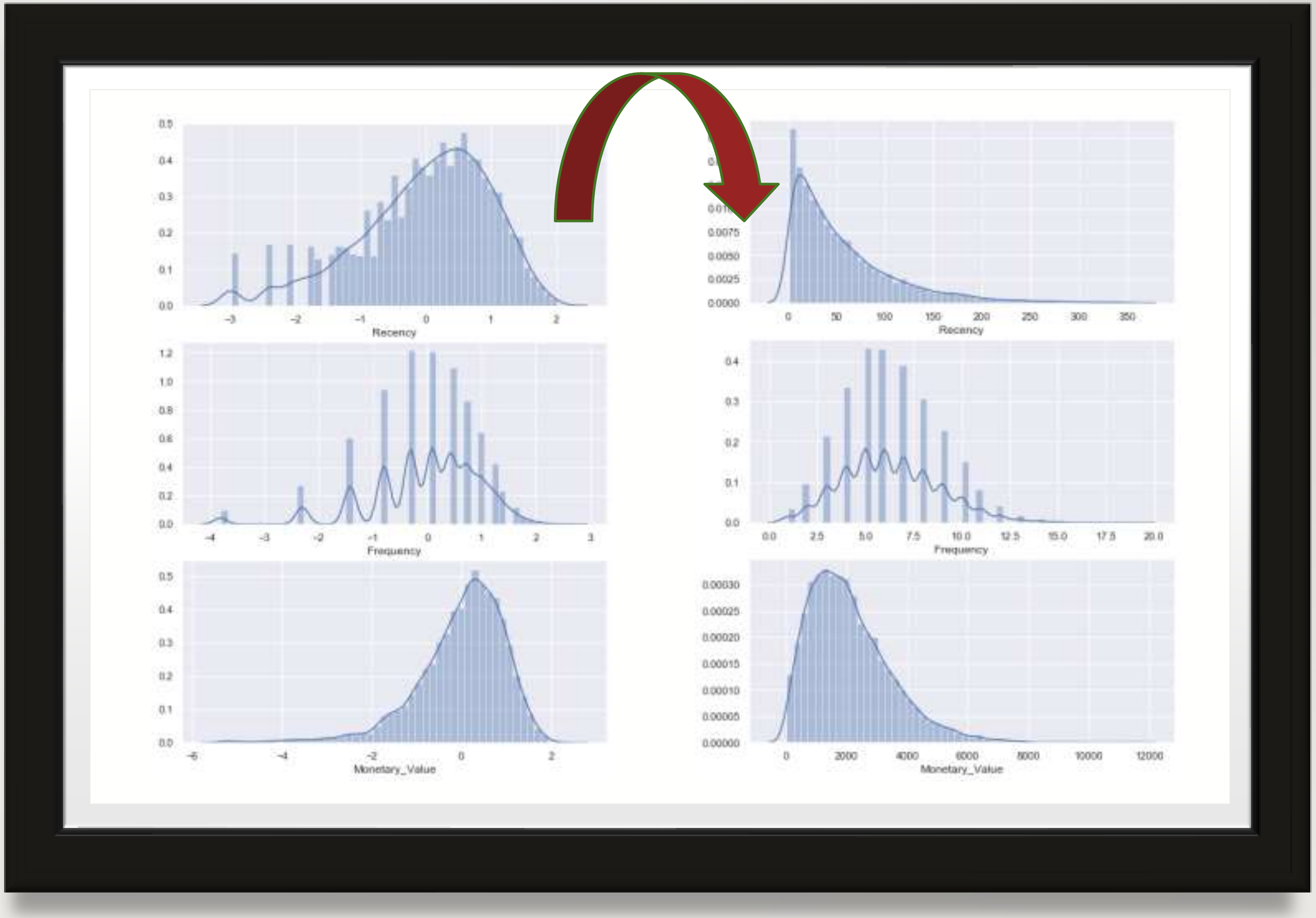
- Partitions n observations into k clusters,
- Each observation belongs to cluster with nearest mean.

- Assumptions:
  - variables symmetrically distributed,
  - have the same mean and variance.



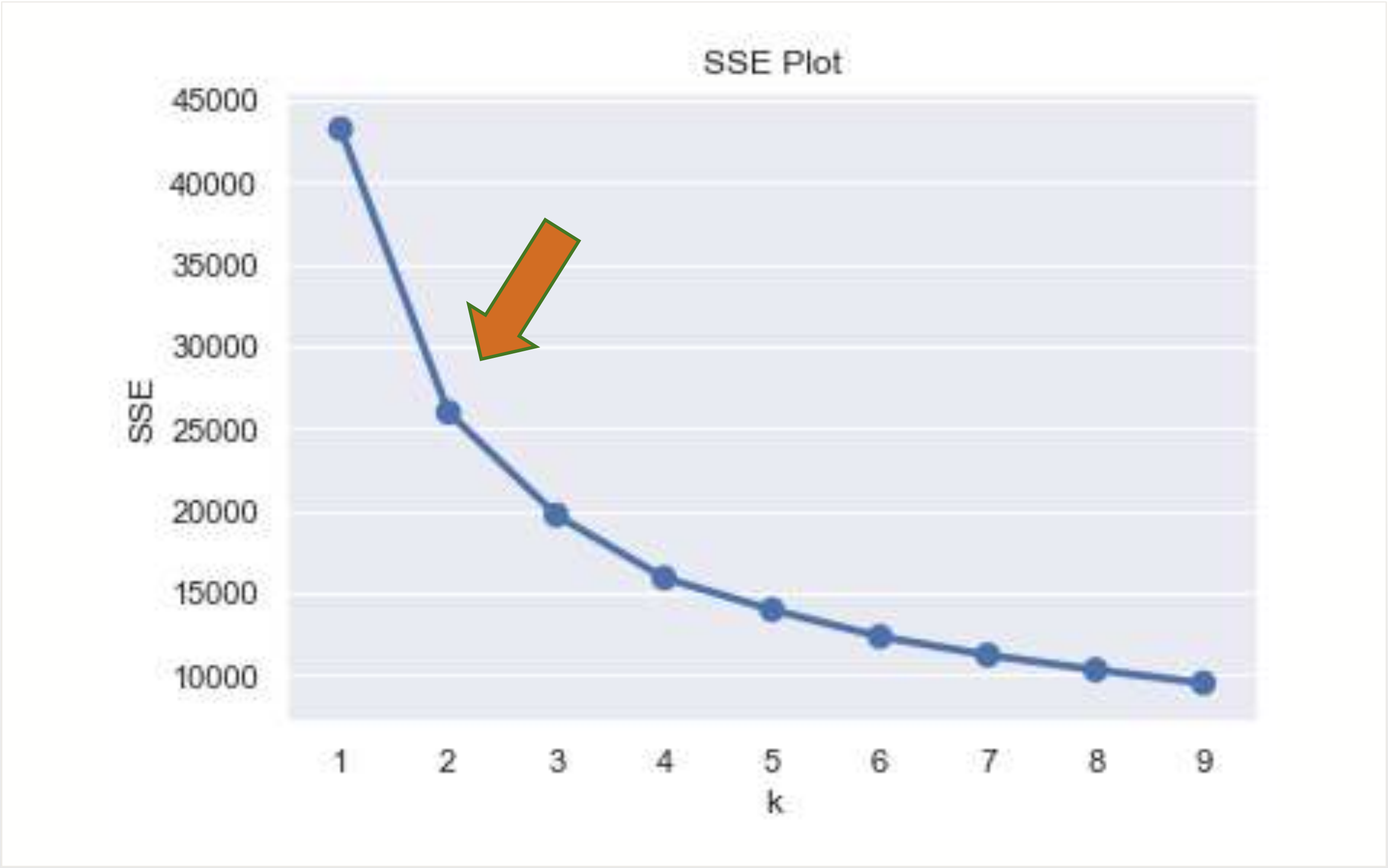
K-Means Clustering

- Log transformation to unskew the data
- Standardized to same mean
- Scaled to same std



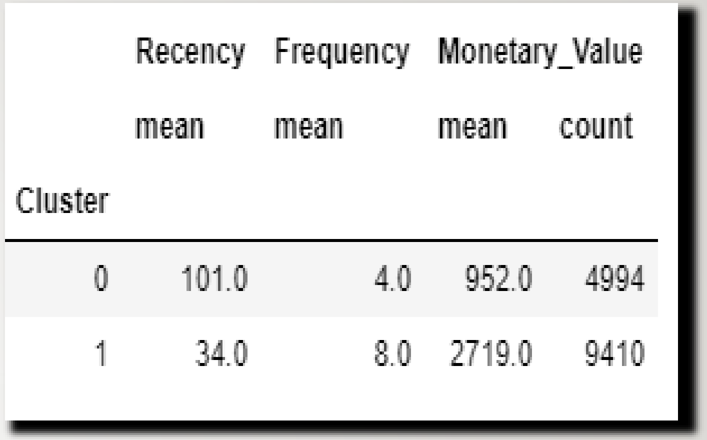
K-Means Clustering

- k = Elbow point (where decrease in SSE slows down) & next point



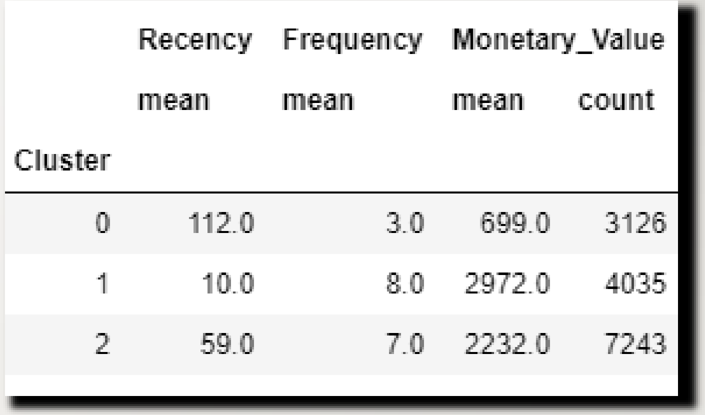
K-Means Clustering

- k=2, clusters more distinct, cluster 1 = 2x cluster 0 in size.



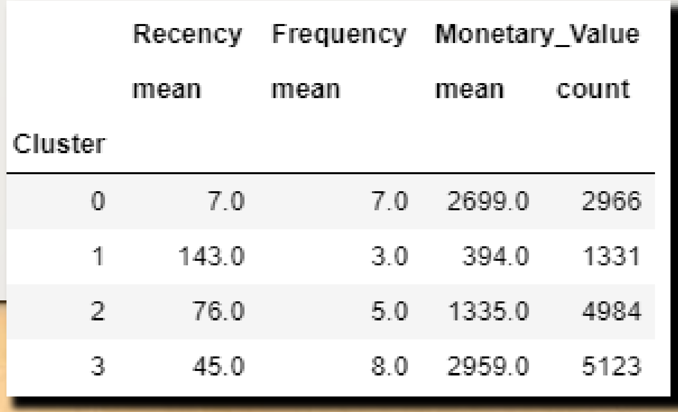
	Recency	Frequency	Monetary_Value	
	mean	mean	mean	count
Cluster				
0	101.0	4.0	952.0	4994
1	34.0	8.0	2719.0	9410

- k=3, clusters 0 and 2 F & M close.



	Recency	Frequency	Monetary_Value	
	mean	mean	mean	count
Cluster				
0	112.0	3.0	699.0	3126
1	10.0	8.0	2972.0	4035
2	59.0	7.0	2232.0	7243

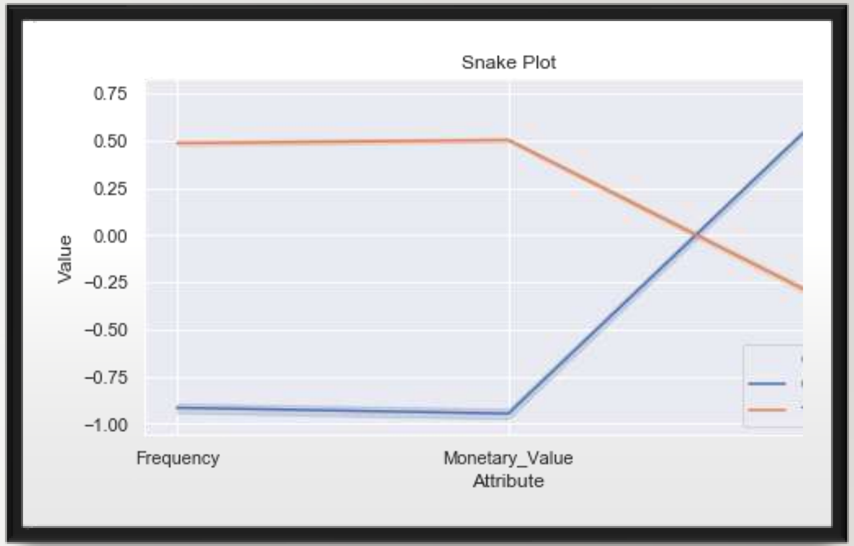
- k=4, clusters 0 and 3 F & M very close.

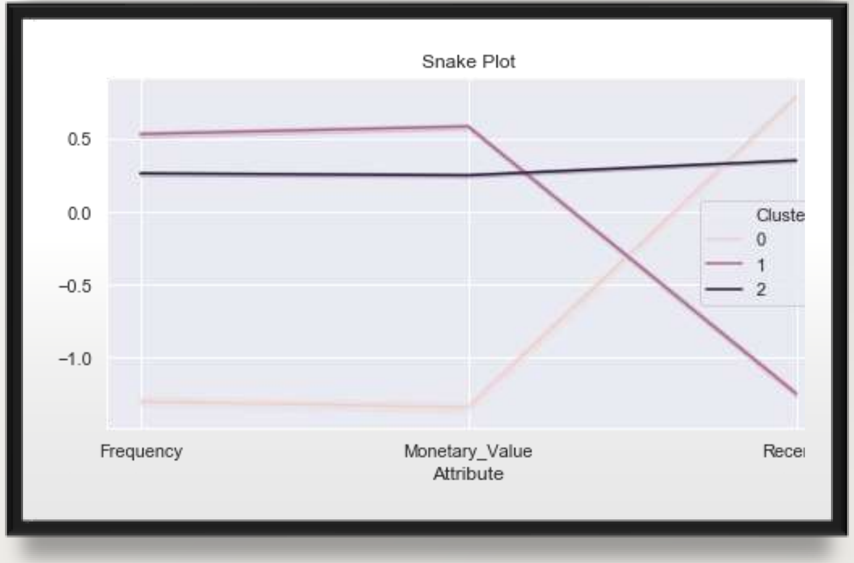


	Recency	Frequency	Monetary_Value	
	mean	mean	mean	count
Cluster				
0	7.0	7.0	2699.0	2966
1	143.0	3.0	394.0	1331
2	76.0	5.0	1335.0	4984
3	45.0	8.0	2959.0	5123

Snake Plots

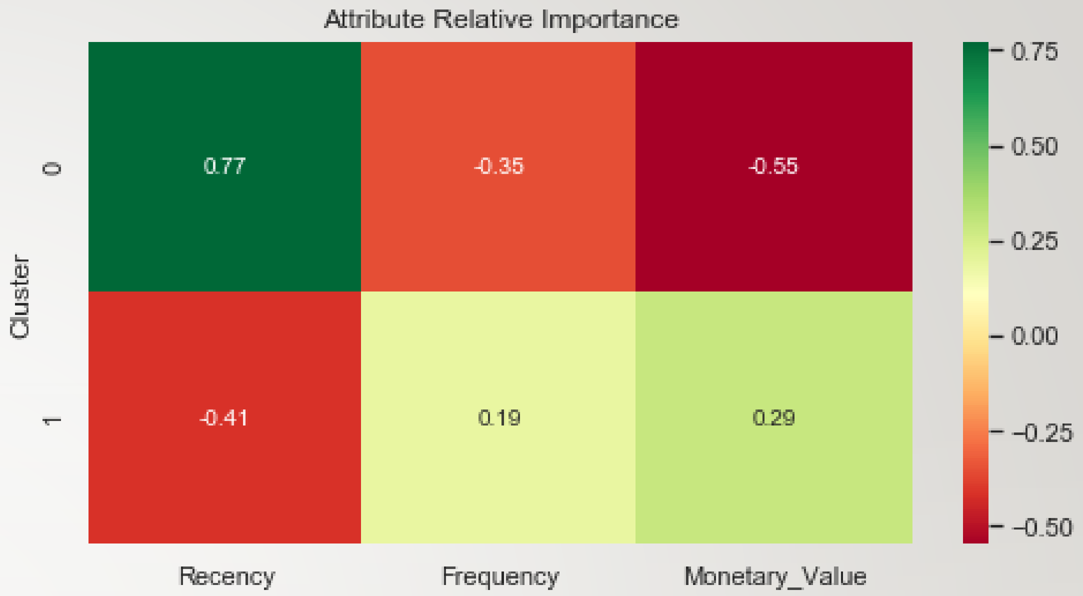
- Tool for visualizing clusters
- Some overlap with 3 clusters.

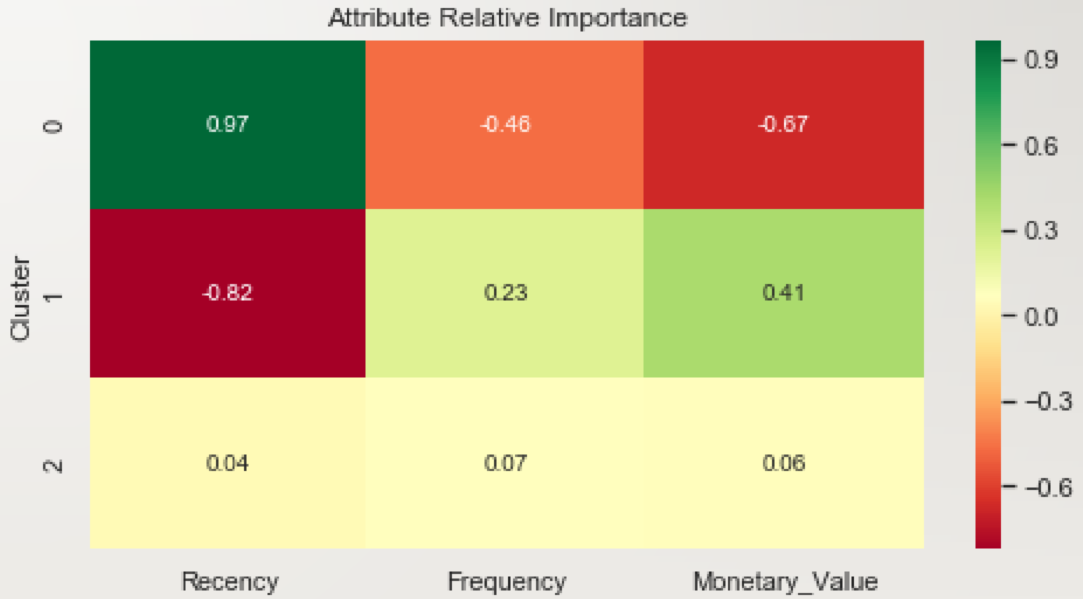




Relative Importance of Segment Attributes

- proportion of cluster average to population average
- indicates none of the attributes are important for defining Cluster 2, compared to the population average.

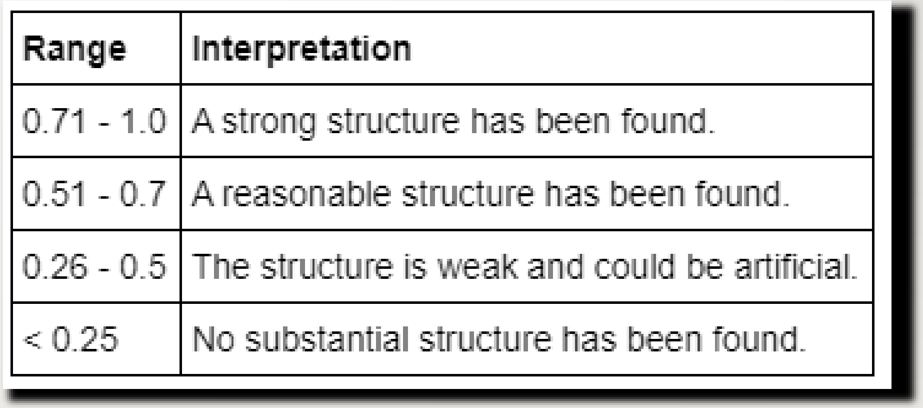




Silhouette Analysis

- measures how well each datapoint xi "fits" its assigned cluster,
- and also how poorly it fits into other clusters.
- axi = avg distance from xi to all other points within its own cluster k. The lower the value, the better.
- bxi = min avg distance from xi to points in a different cluster, minimized over clusters.

$$s(x_i) = \frac{b_{x_i} - a_{x_i}}{\max\left(a_{x_i}, b_{x_i}\right)}$$



Range	Interpretation
0.71 - 1.0	A strong structure has been found.
0.51 - 0.7	A reasonable structure has been found.
0.26 - 0.5	The structure is weak and could be artificial.
< 0.25	No substantial structure has been found.

Silhouette Analysis

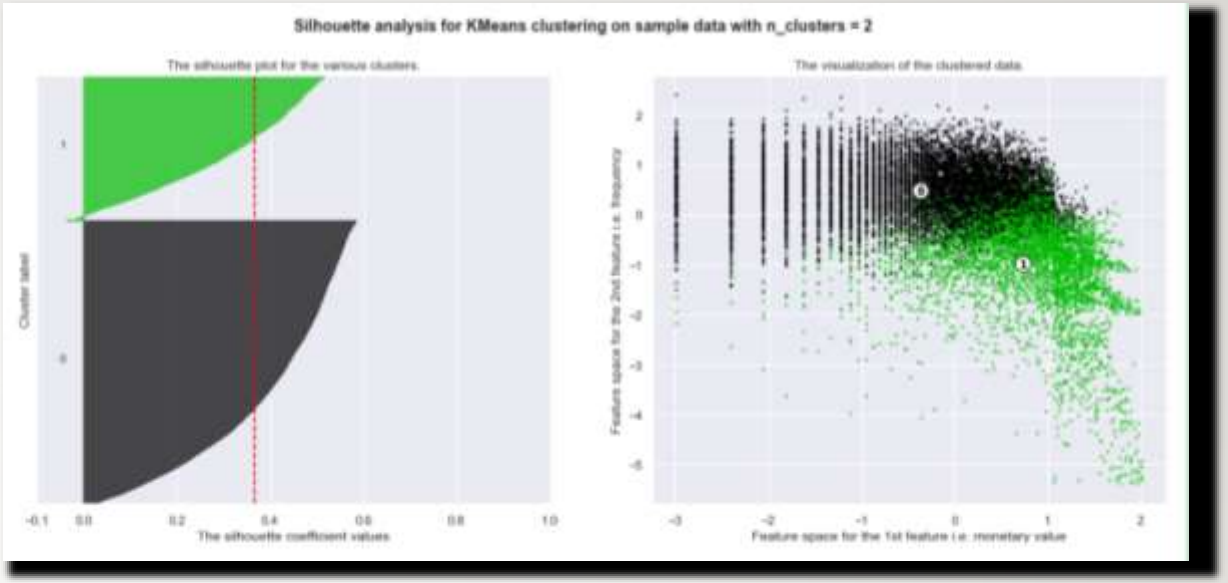
- n_clusters =2,

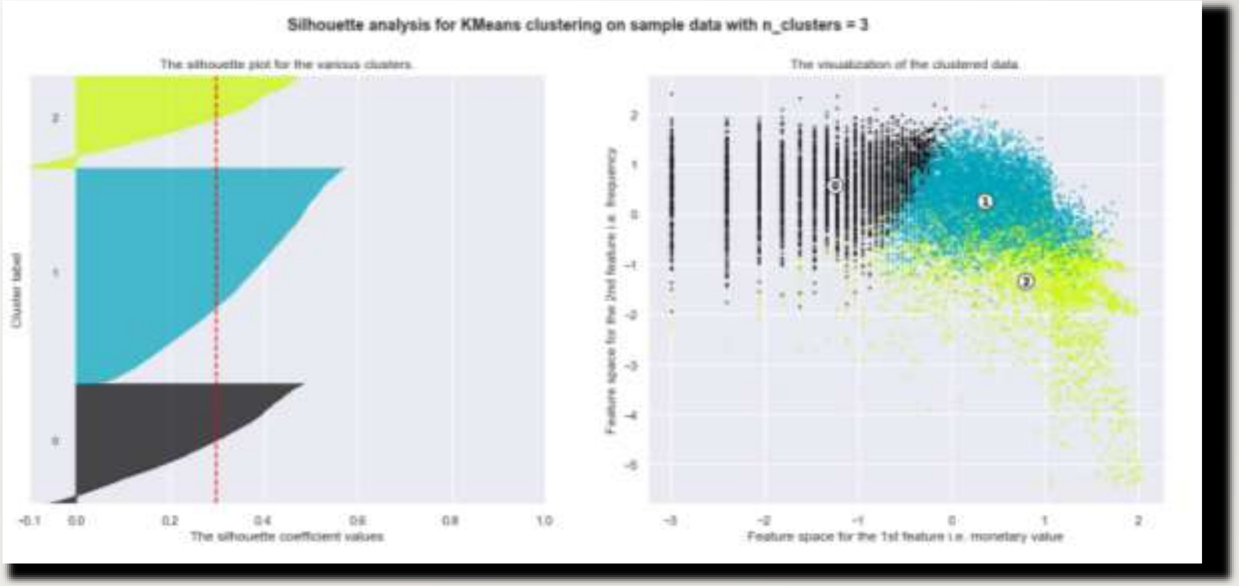
Silhouette Score = 0.37.

- n_clusters =3,

Silhouette Score = 0.3.

- Best score < 0.5 (with n_clusters=2),
- →Structure is weak & could be artificial





AGENDA

- Introduction
- Data Wrangling & EDA
- Cohort Analysis
- RFMT Analysis
- K-Means Clustering
- Conclusion

Conclusion

- Methods used:
  - RFMT Analysis,
  - K-Means clustering,
    - Snake Plots,
    - Relative Importance of Segment Attributes,
    - Silhouette scores.

Conclusion

- All methods identified two distinct clusters,
- While favoring 2 clusters, a 3 clusters option is also possible,
- RFMT Analysis → 3 almost equally distanced clusters feasable,

Conclusion

- For better capturing customer behavior,
- and more focused marketing to target diverse customers,
- suggest 3-clustered customer segmentation, pending managerial decision.