Departamento de Sistemas e Computação – FURB
Curso de Ciência da Computação
Trabalho de Conclusão de Curso – 2022/1

UTILIZAÇÃO DE CLUSTERIZAÇÃO PARA SEGMENTAÇÃO DE CLIENTES A PARTIR DE DADOS DE VAREJO

Aluno: Henrique José Wilbert

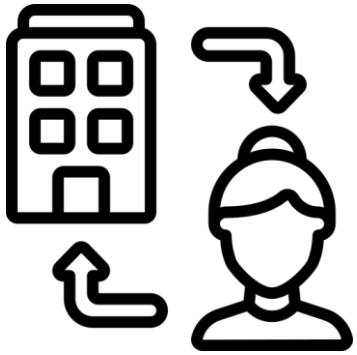
Orientador: Aurélio Faustino Hoppe

Roteiro

- Introdução
- Objetivos
- Fundamentação Teórica
- Trabalhos Correlatos
- Requisitos
- Especificação
- Implementação
- Análise dos Resultados
- Conclusões e Sugestões

Introdução(1/2)

- Crescimento da informatização de processos.

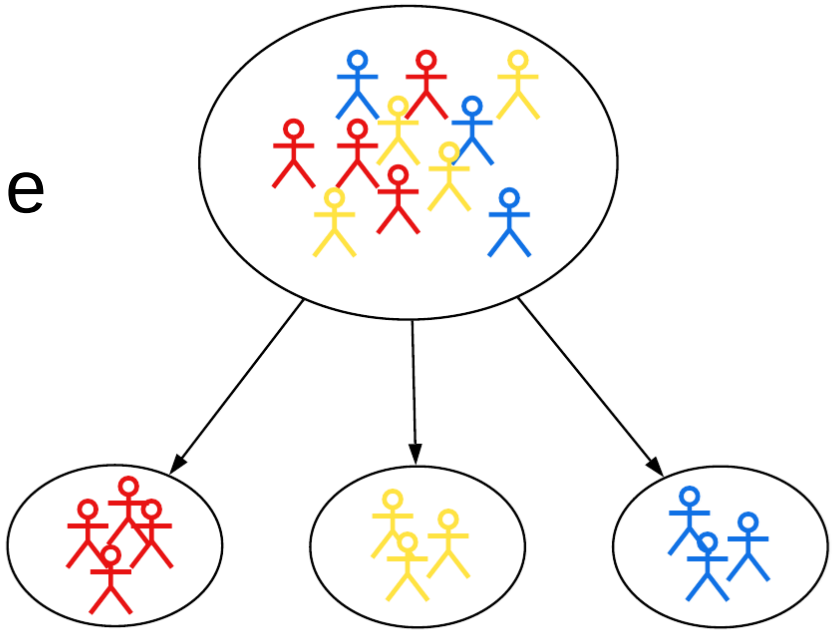


- Possibilidades de tratar os dados além do nível operacional.

- Foco nas relações empresa-cliente.

Introdução(2/2)

- Segmentação
- (R)ecência, (F)requência e (M)onetização
- *Clustering*
- Índices de Validação



Objetivos

Criação de um protótipo que utilize os atributos do modelo RFM em conjunto com o algoritmo de clusterização K-means.

i. Identificar diferentes segmentos de clientes com base em seu comportamento, com foco na definição da quantidade (k) de clusters.

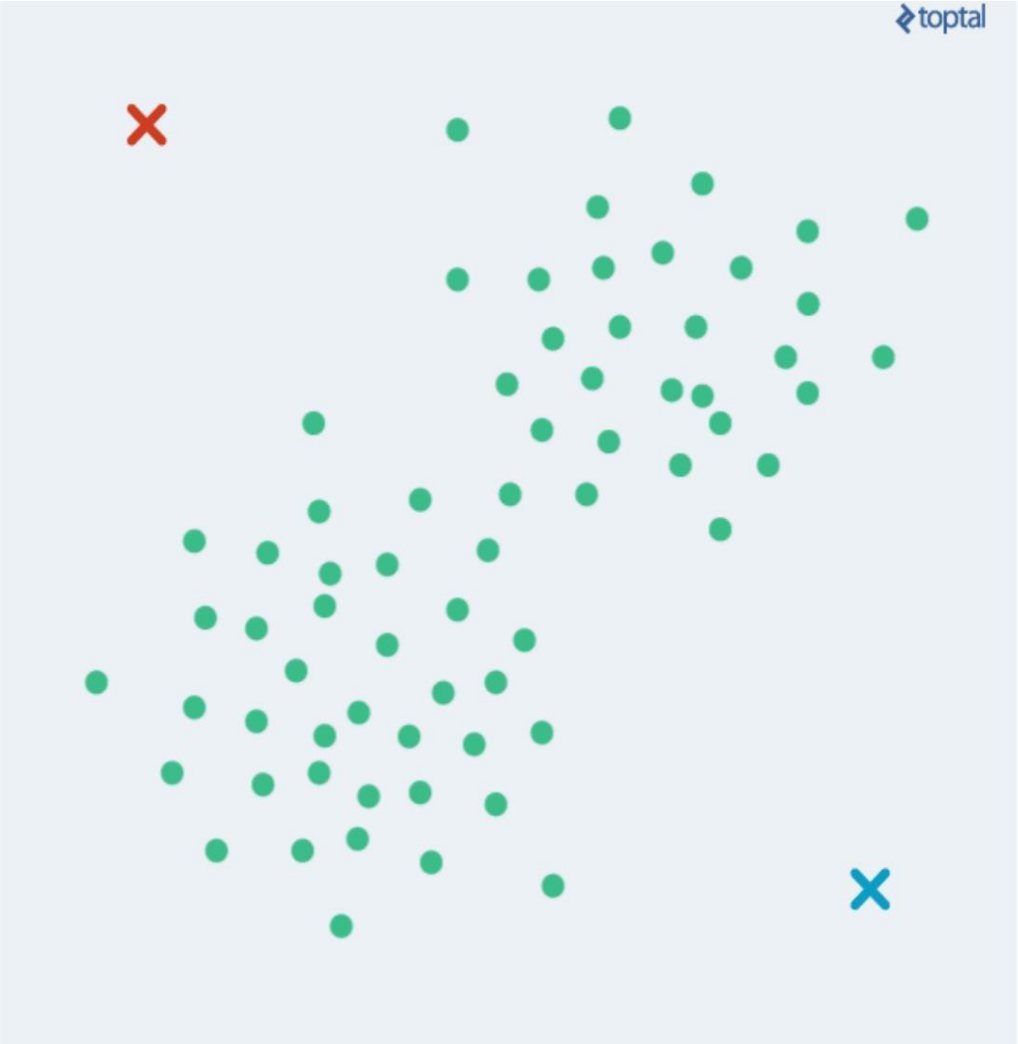
ii. Aplicar três índices de validação internos (Silhouette, Calinski-Harabasz e Davies-Bouldin)

iii. Aplicar três índices de validação externos (estabilidade global, estabilidade por cluster e estabilidade SLSa - Segment Level Stability across solutions)

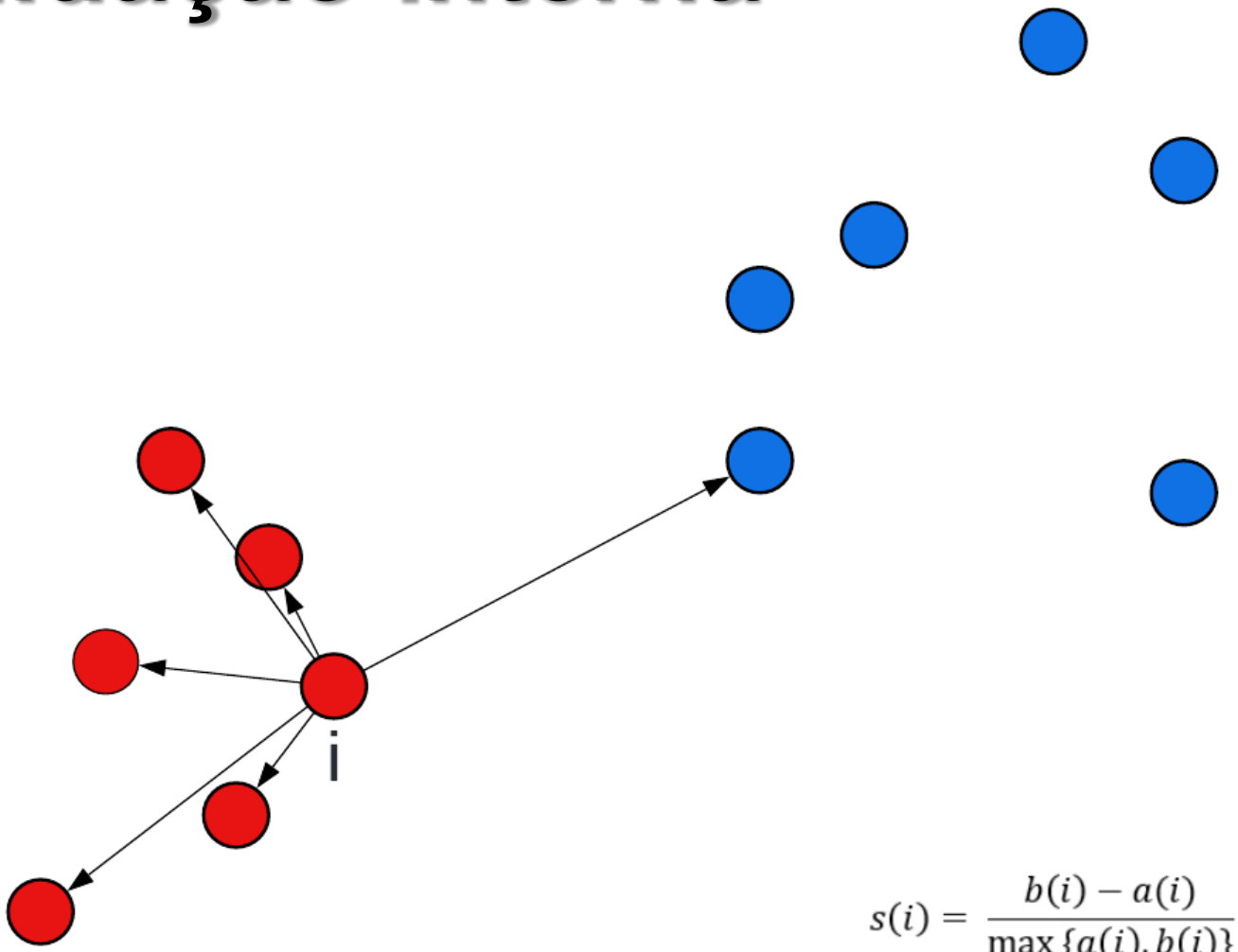
Fundamentação Teórica

- Clustering por K-means
- Índices de validação Interna
- Índices de validação Externa

Clustering por K-means

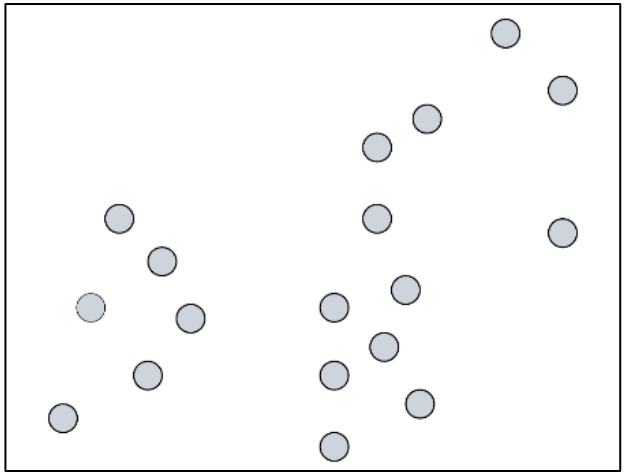


Validação Interna

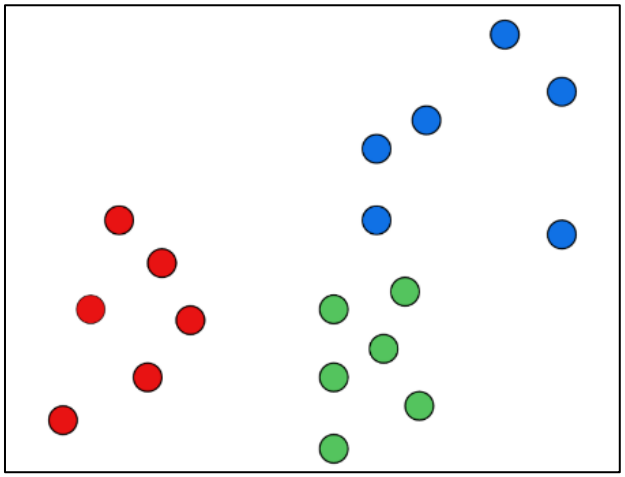


$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

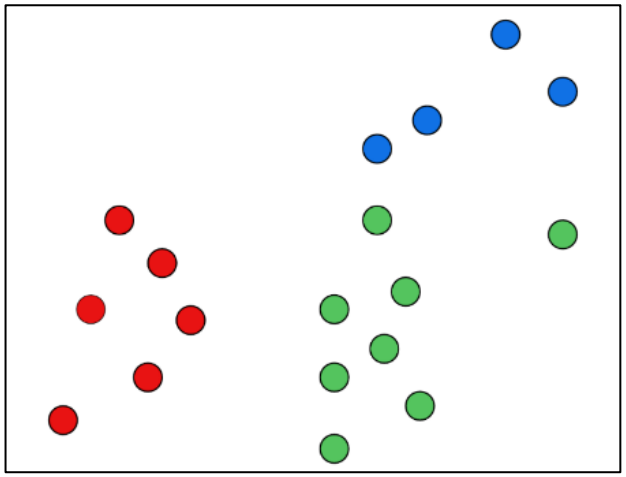
Validação Externa (Comparação)



Conjunto não clusterizado



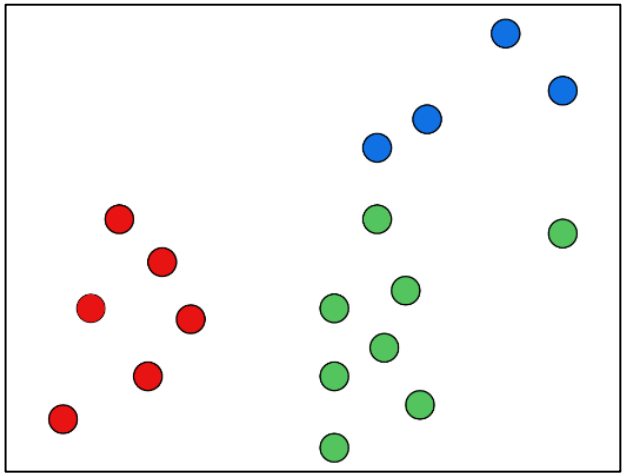
“True Labels” - clusters reais



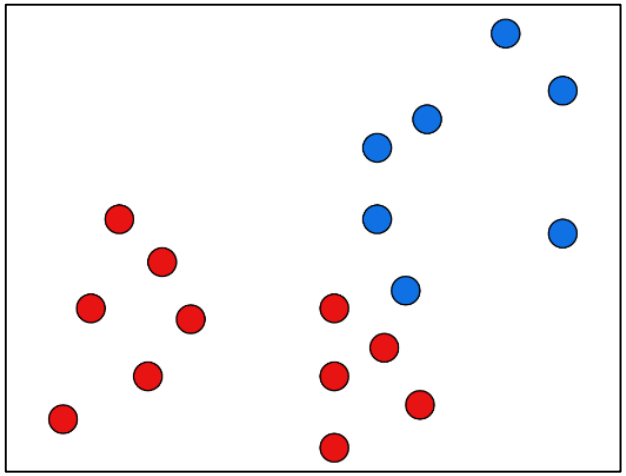
Resultado da Clusterização

Validação Externa (Estabilidade) (1/2)

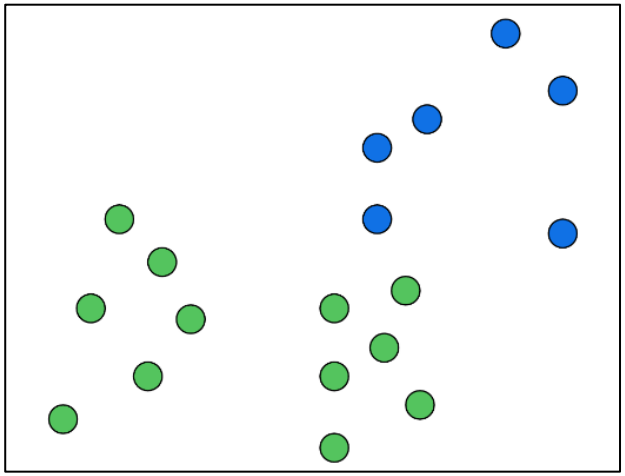
Comparação de várias execuções de um algoritmo:



1ª Execução



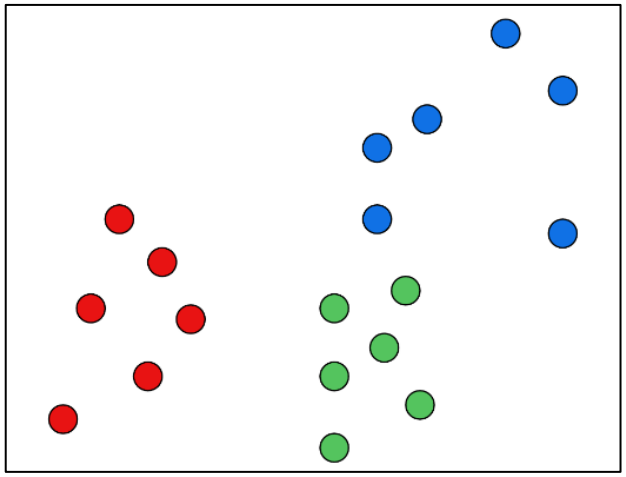
2ª Execução



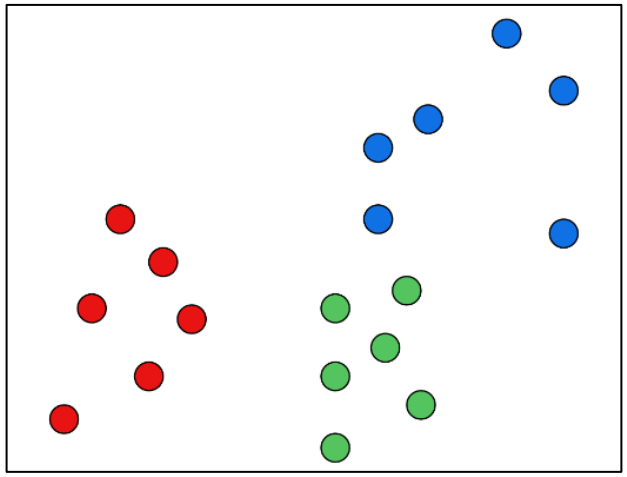
3ª Execução

Validação Externa (Estabilidade) (2/2)

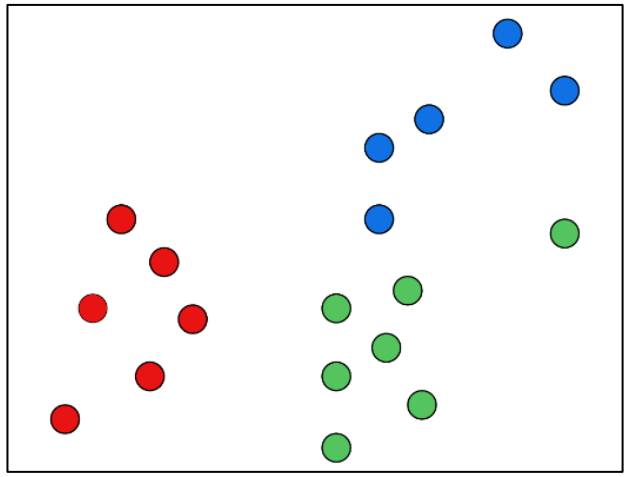
Estabilidade desejável:



1ª Execução



2ª Execução



3ª Execução

Trabalhos correlatos (1/3)

Título: LRFMP Model For Customer Segmentation In The Grocery Retail Industry: A Case Study Peker, Kocyigit e Eren (2017)

Características \ Trabalhos	Peker, Kocyigit e Eren (2017)
Alvo da clusterização	Clientes
Modelo utilizado	LRFMP
Objetivo da segmentação	Gerenciamento das relações com cliente
Algoritmo de clusterização utilizado	K-means
Foco metodológico	Formulação de um modelo novo e análise dos resultados
Número de dados (clientes)	16.024
Quantidade de índices de validação interna	3
Quantidade de clusters gerados	5
Inferências sobre os dados	Sim

Trabalhos correlatos (2/3)

Título: Segmentation of university customers loyalty based on RFM analysis using Fuzzy c-means clustering Hidayat et al. (2020)

Características \ Trabalhos	Hidayat et al. (2020)
Alvo da clusterização	Escolas
Modelo utilizado	RFM
Objetivo da segmentação	Lealdade na cooperação escola-universidade
Algoritmo de clusterização utilizado	Fuzzy c-means (FCM)
Foco metodológico	Análise dos resultados
Número de dados (escolas)	2.043
Quantidade de índices de validação interna	1
Quantidade de clusters gerados	3
Inferências sobre os dados	Sim

Trabalhos correlatos (3/3)

Título: Customer Segmentation And Strategy Development Based On User Behavior Analysis, Rfm Model And Data Mining Techniques: A Case Study

Tavakoli et al. (2018)

Características \ Trabalhos	Tavakoli *et al*. (2018)
Alvo da clusterização	Clientes
Modelo utilizado	R+FM
Objetivo da segmentação	Gerenciamento das relações com cliente
Algoritmo de clusterização utilizado	K-means
Foco metodológico	Formulação de um modelo novo e campanha de ofertas
Número de dados (clientes)	~3.000.000
Quantidade de índices de validação interna	-
Quantidade de clusters gerados	10
Inferências sobre os dados	Sim

Requisitos

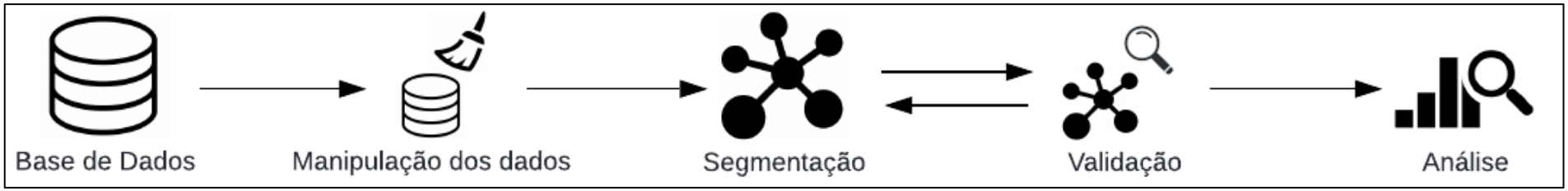
Funcionais

RF01 - Adquirir os dados transacionais de clientes a partir de um banco de dados.
RF02 - Filtrar os clientes com informações irregulares.
RF03 - Extrair dos clientes as características (recência, frequência e monetária) utilizadas no modelo RFM.
RF04 - Normalizar os dados para evitar disparidades nas escalas dos atributos.
RF05 - Exibir num gráfico 3D os clientes com sua localização definida pela pontuação do cliente nas características RFM.
RF06 - Segmentar em clusters os clientes com base nos atributos RFM.

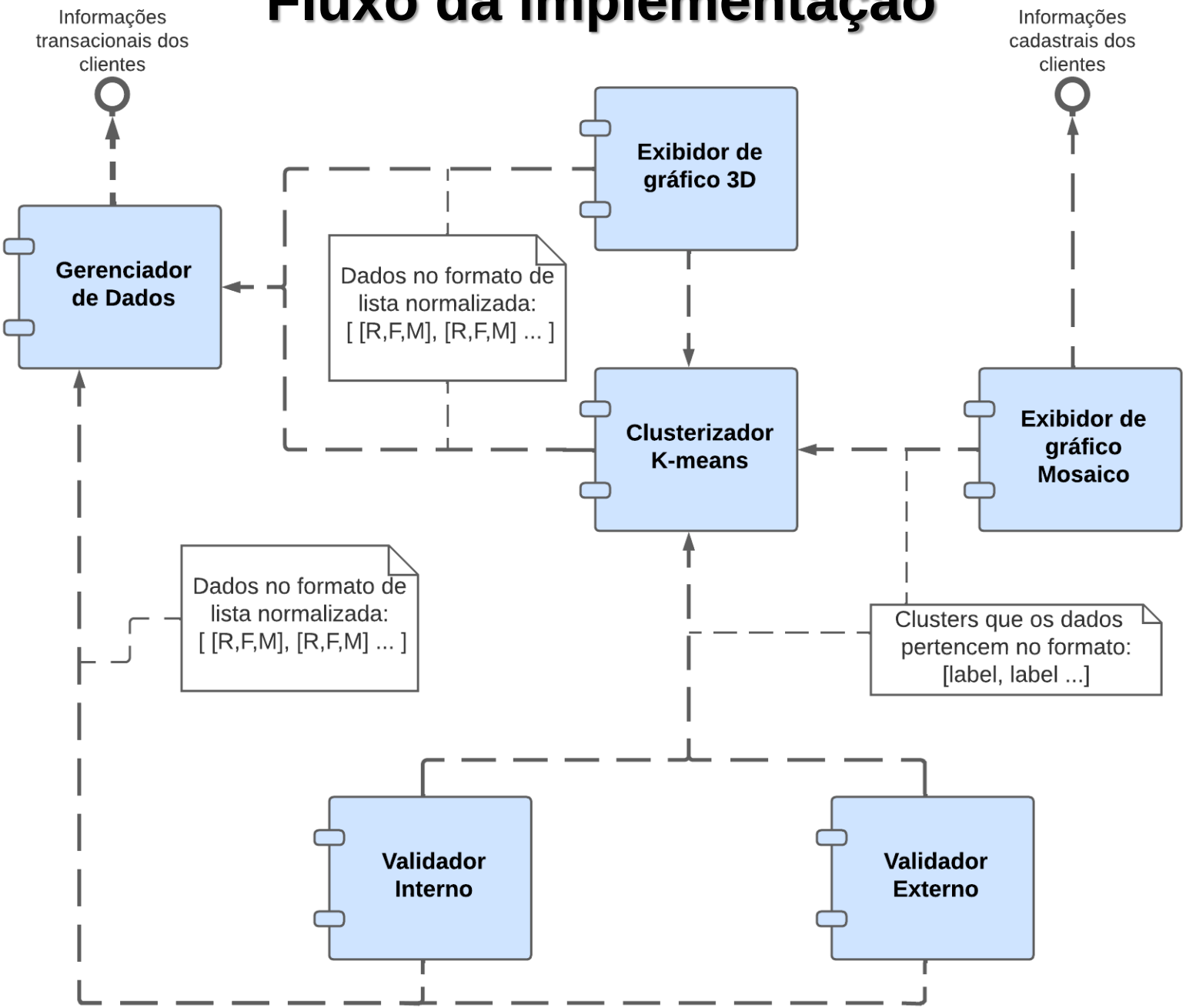
Não Funcionais

RNF01 - Utilizar o algoritmo de clusterização K-means para a segmentação dos clientes.
RNF02 - Aplicar os índices de validação interna Silhouette, Calinski-Harabasz e Davies-Bouldin para validação da qualidade dos clusters.
RNF03 - Aplicar os índices de validação externa de estabilidade global, estabilidade por cluster e estabilidade SLSa.
RNF04 - Utilizar a linguagem Python para o desenvolvimento do protótipo.

Etapas da implementação



Fluxo da implementação



Implementação (1/10)

Foram extraídos 1748 clientes de uma base de dados de loja de roupas masculinas e femininas:

ID	Recência	Frequência	Monetário
38	139	65	37176

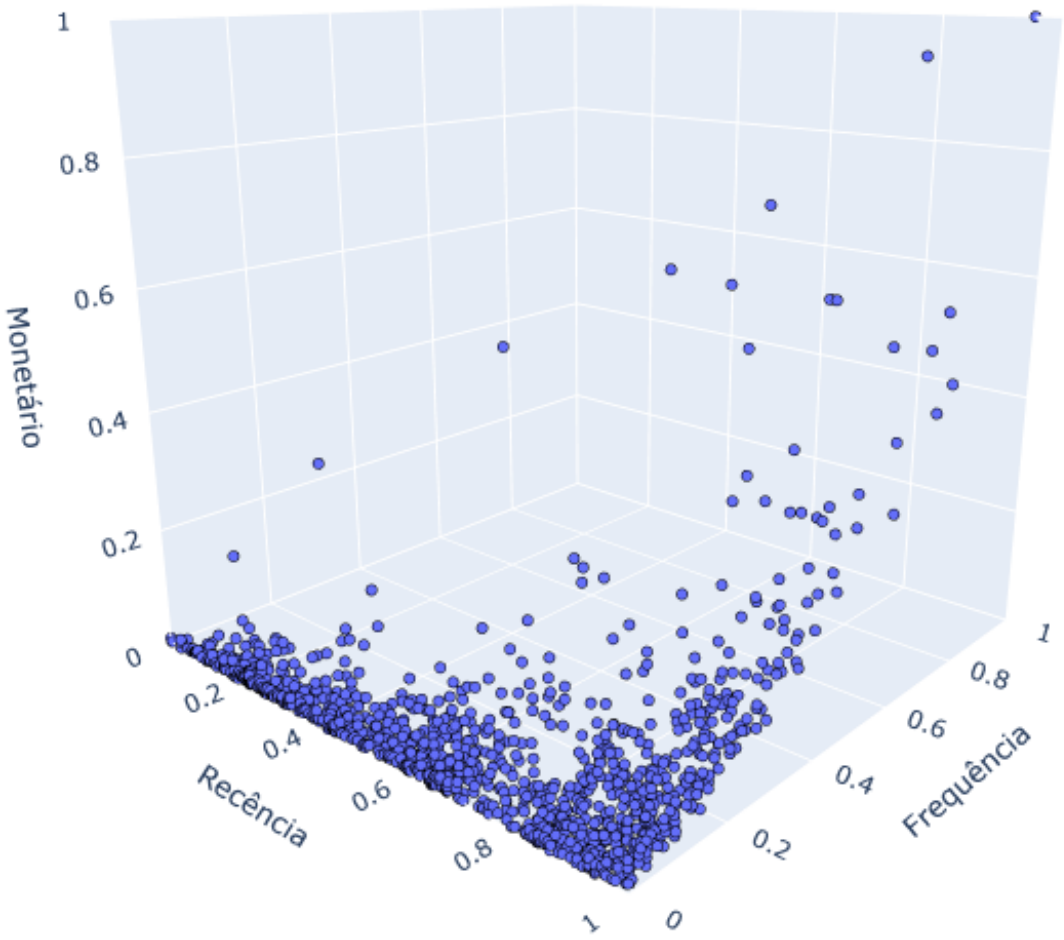
Após aplicada a normalização Min-Max:

ID	Recência	Frequência	Monetário
38	0,0074928	0,71910112	0.43890863

$$\frac{v - min_A}{max_A - min_A}\left(novo_{max_A} - novo\_min_A\right) + novo\_min_A$$

Implementação (2/10)

Distribuição dos clientes no gráfico 3D (RFM):



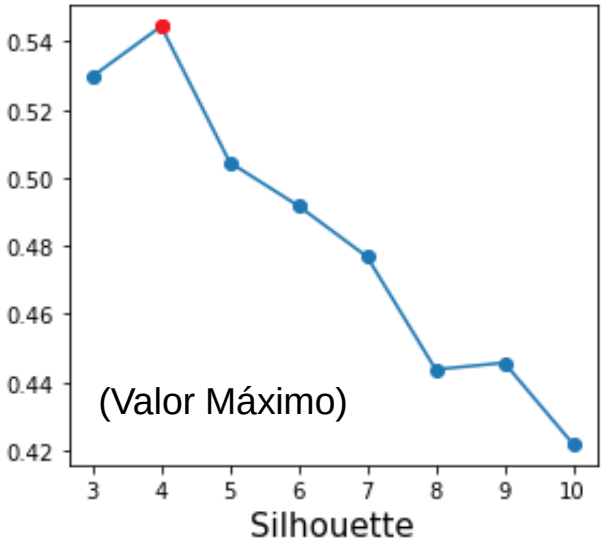
Implementação (3/10)

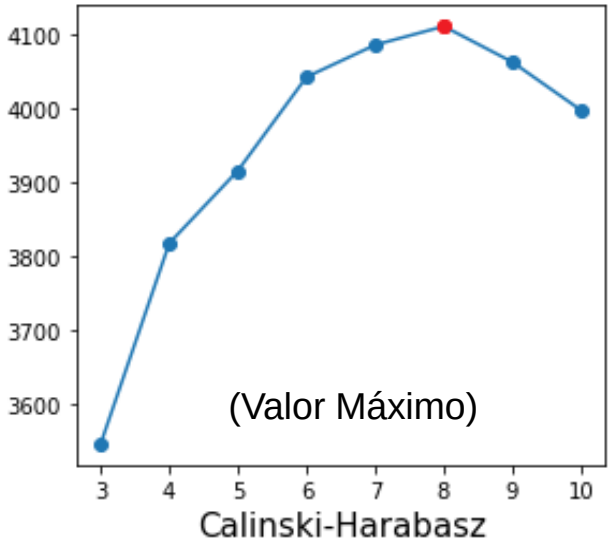
- Clusterização por K-means com k= 3 até k=10
- Silhouette (cada dado) $$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$
- Calinski-Harabasz (índice geral) $$VRC = \frac{BGSS}{\mathrm{k} - 1} / \frac{WGSS}{\mathrm{n} - \mathrm{k}}$$
- Davies-Bouldin (cluster) $$R_{ij} \equiv \frac{S_i + S_j}{M_{ij}} \quad \bar{R} \equiv \frac{1}{N} \sum_{i=1}^{N} R_i$$

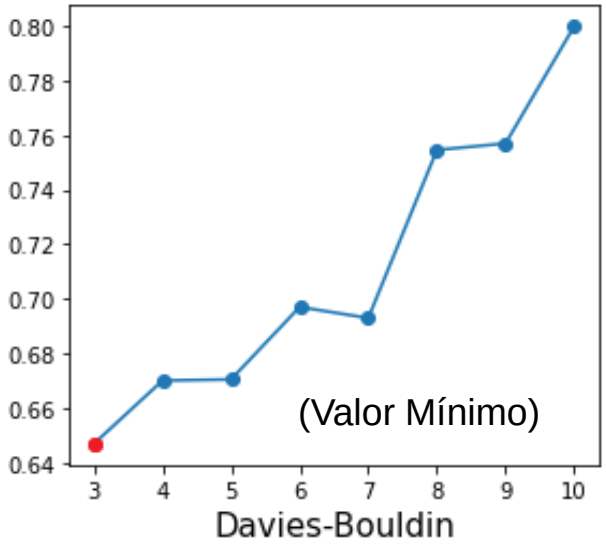
Implementação (4/10)

Foram aplicados os índices ao longo das 8 soluções.

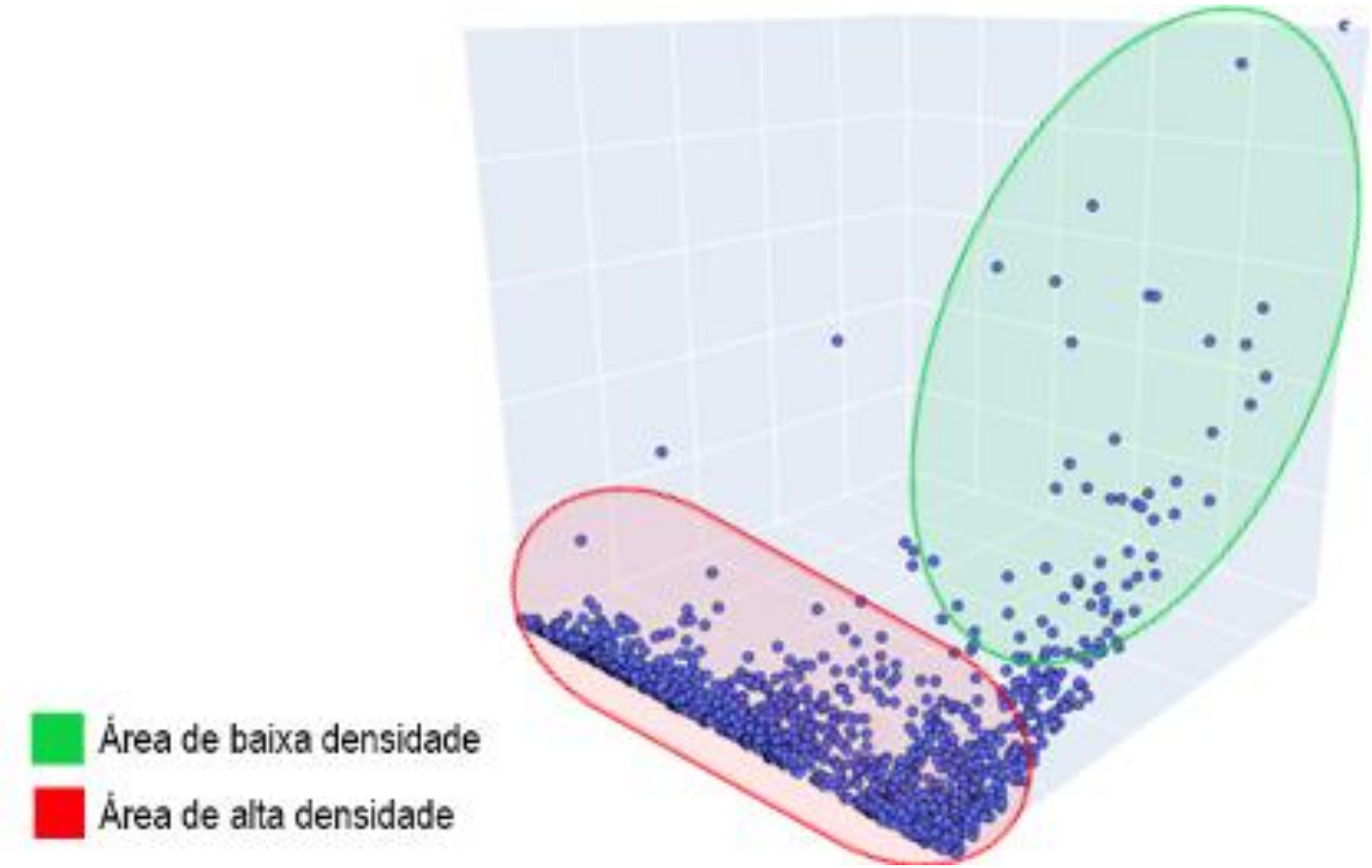
Recomendação dos índices:







Implementação (5/10)



Como não há estrutura natural de clusters, ela será imposta nos dados.

Implementação (6/10)

- Estabilidade global (Adjusted Rand Index – ARI)

$$RI = \frac{a+d}{a+b+c+d}$$

- Estabilidade por cluster (Jaccard Index)

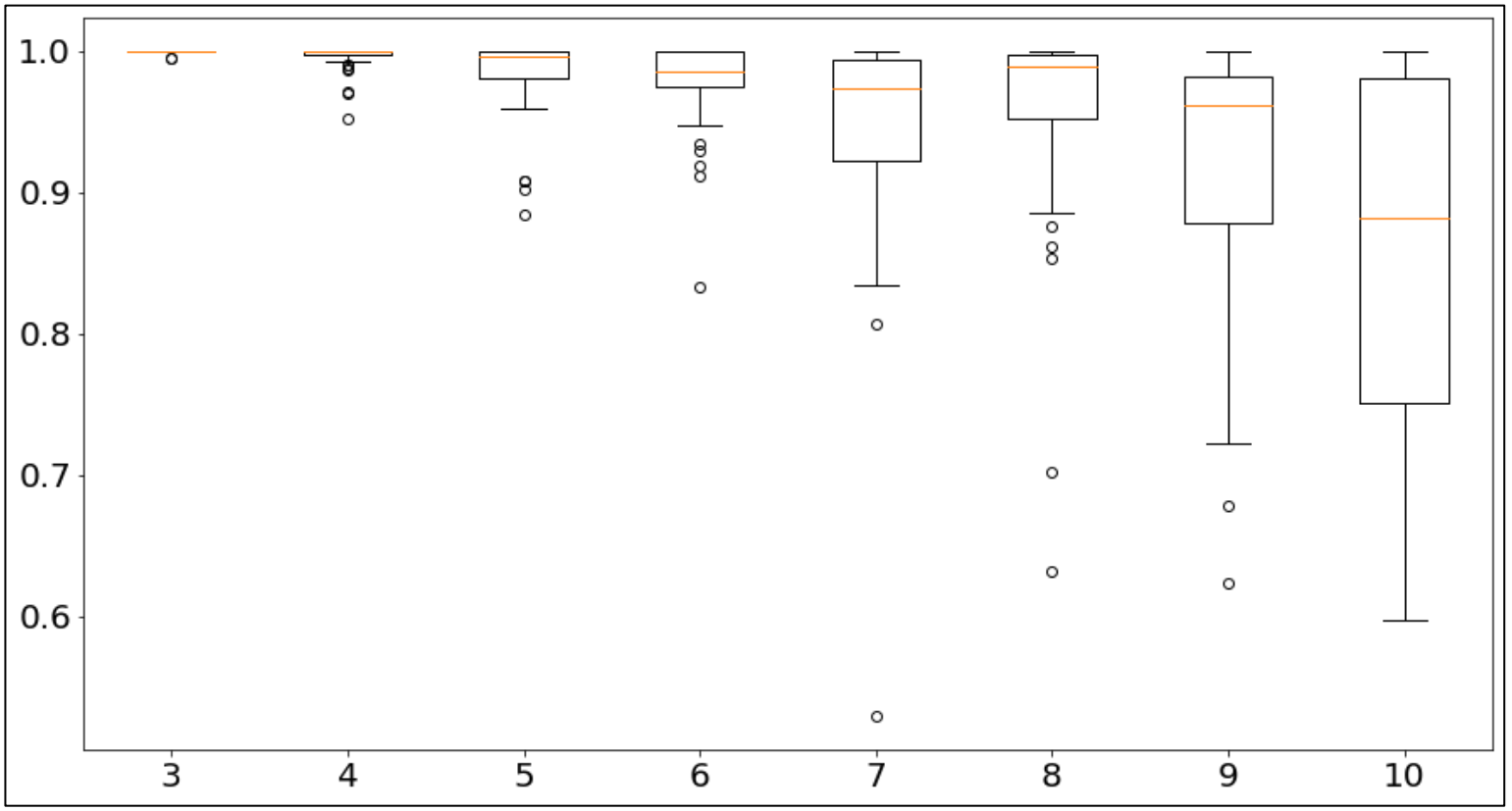
$$Jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

- Estabilidade SLSa (Entropia)

$$H = -\sum_{i=1}^{n} p_j \log p_j$$

Implementação (7/10)

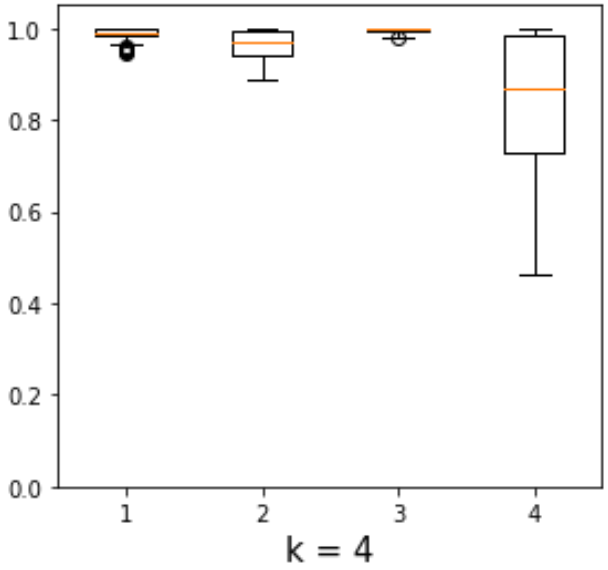
Estabilidade global (Adjusted Rand Index - ARI):

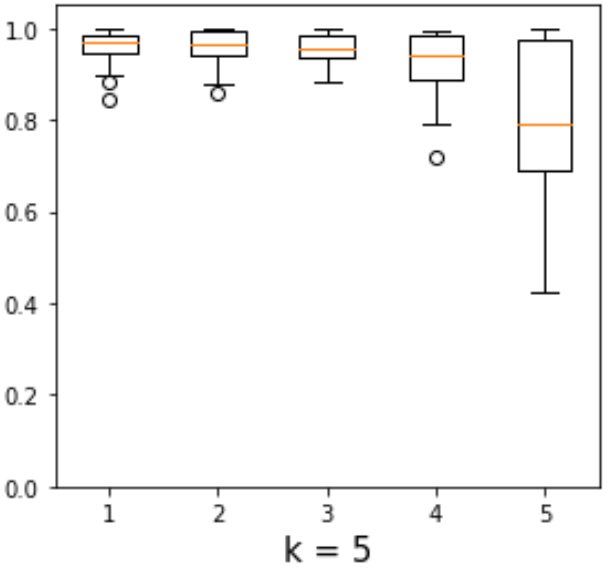


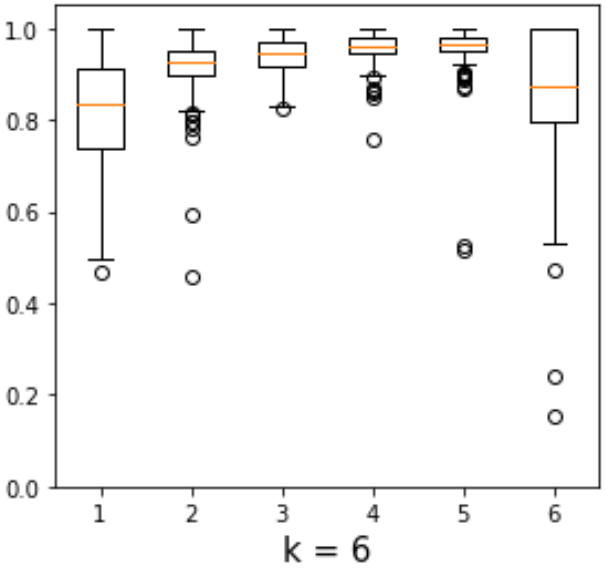
- Consiste em criar várias amostras, aplicar a clusterização delas, e comparar as soluções através do índice ARI.

Implementação (8/10)

Estabilidade por cluster (Jaccard Index):



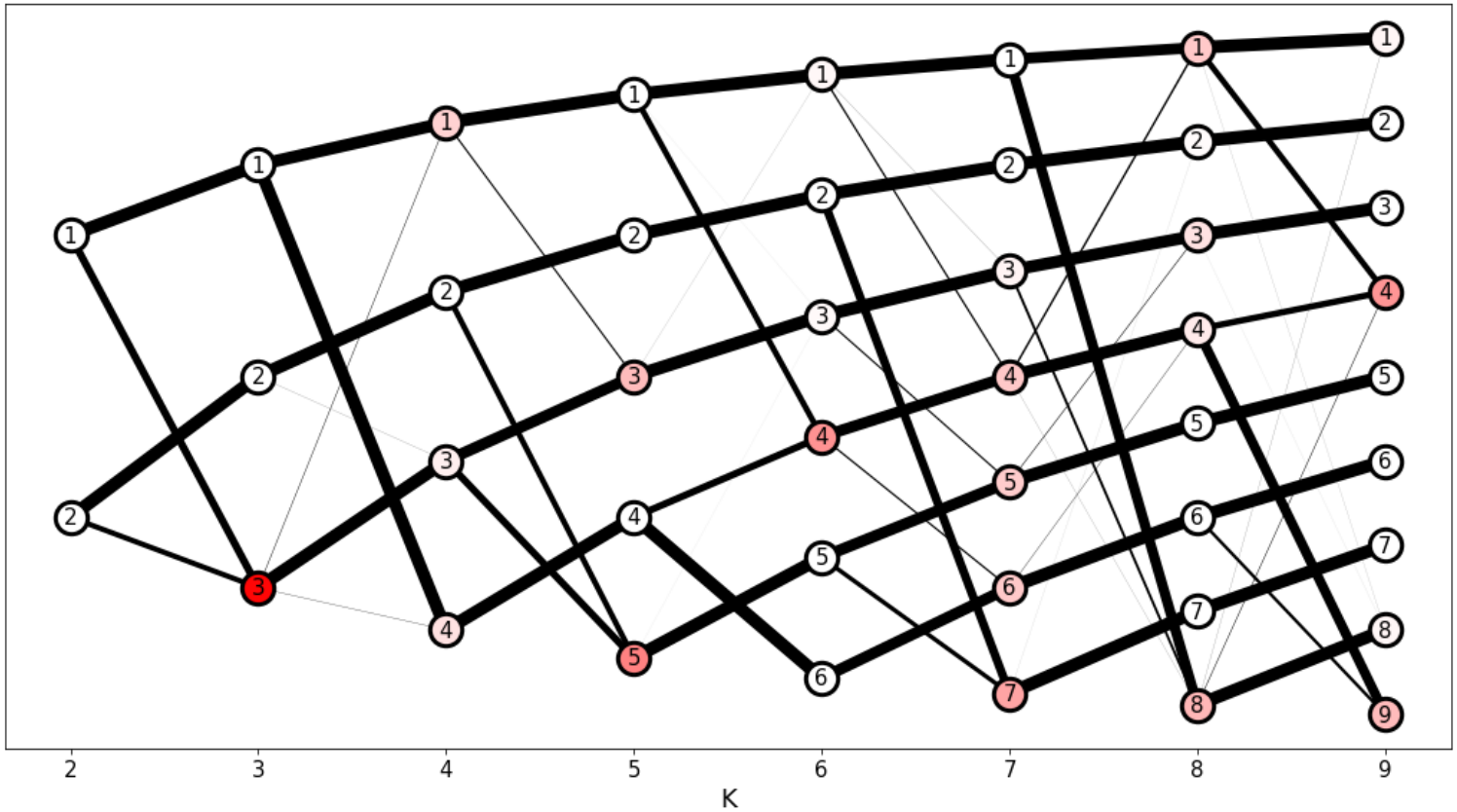




- Consiste em criar várias amostras com mesmo número de clusters, aplicar a clusterização delas, e comparar as soluções através do índice de Jaccard.

Implementação (9/10)

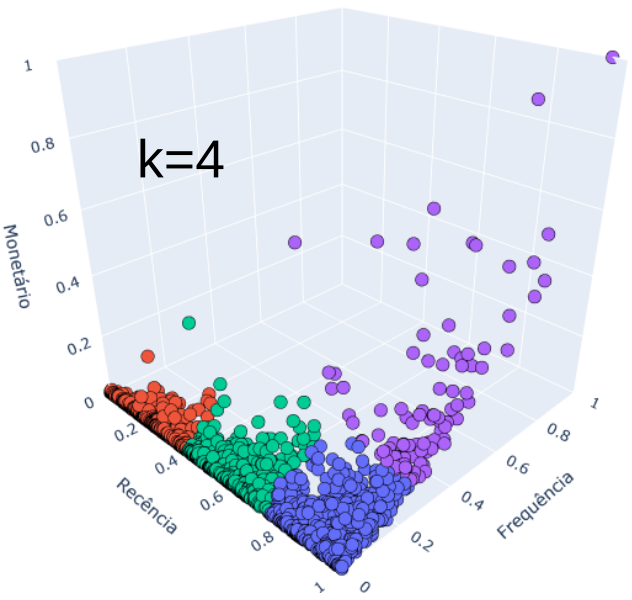
Estabilidade SLSa (Entropia):

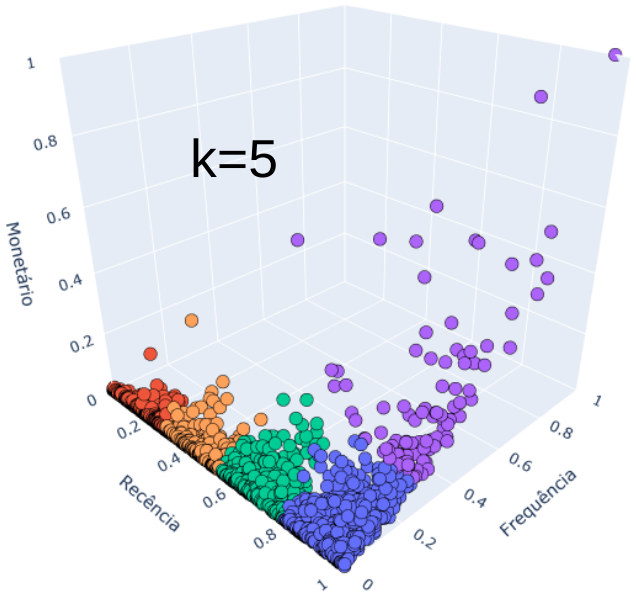


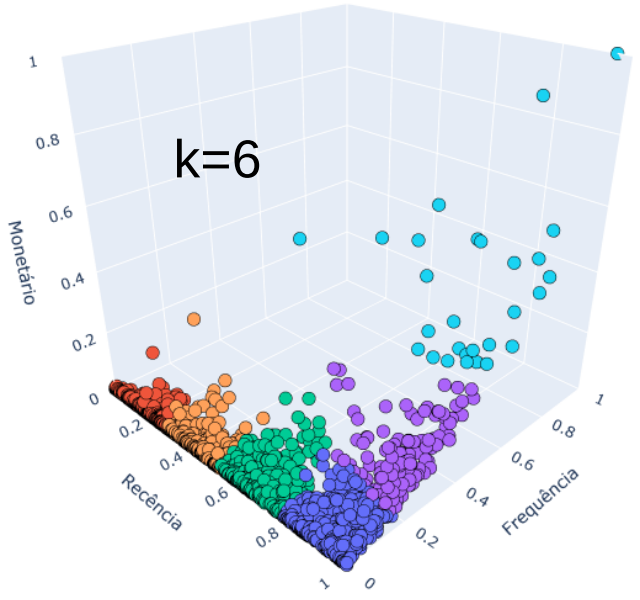
- Consiste no cálculo da medida de entropia entre cada cluster e os clusters da solução anterior.

Implementação (10/10)

Soluções k=4, k=5 e k=6:

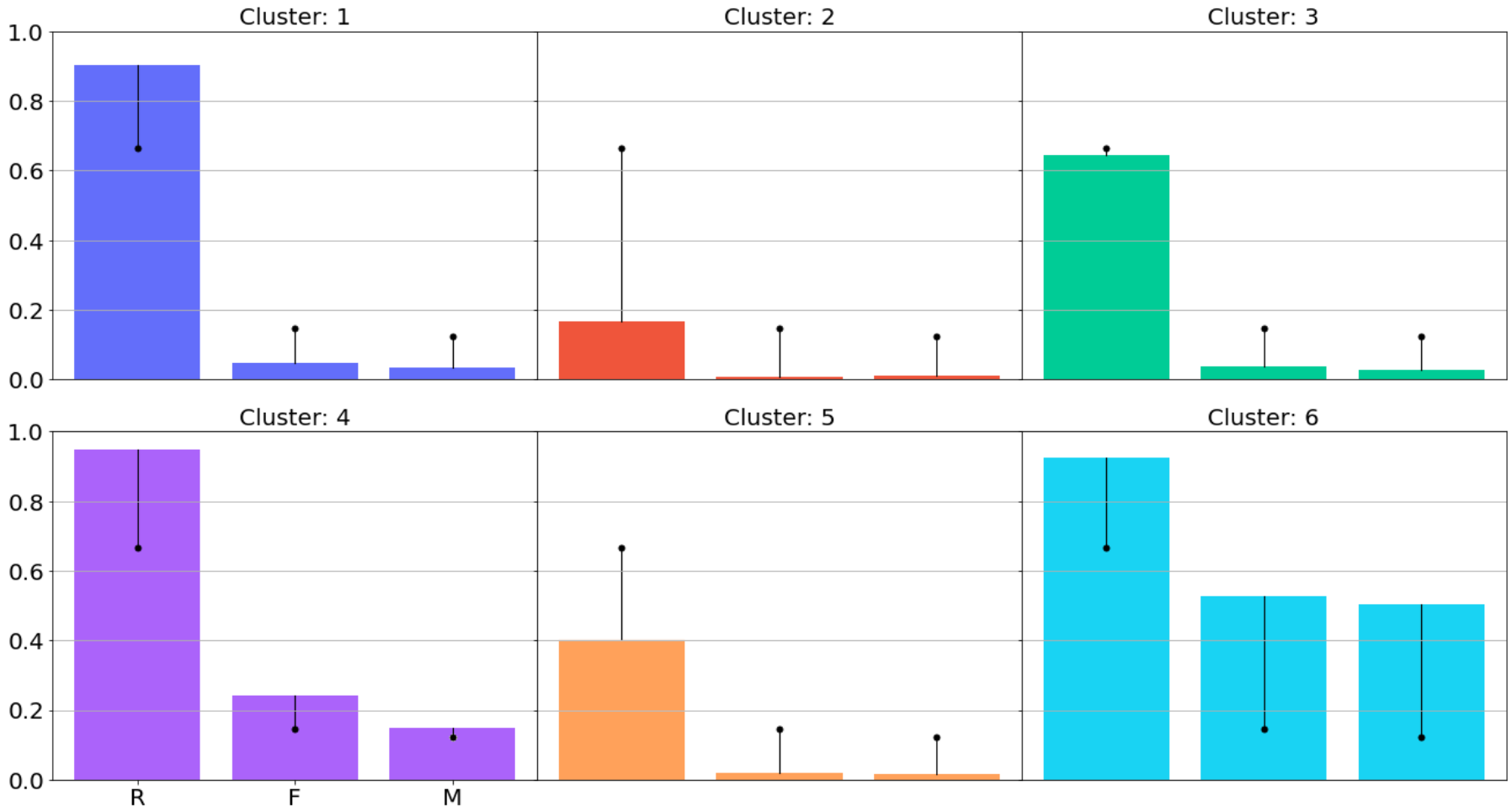






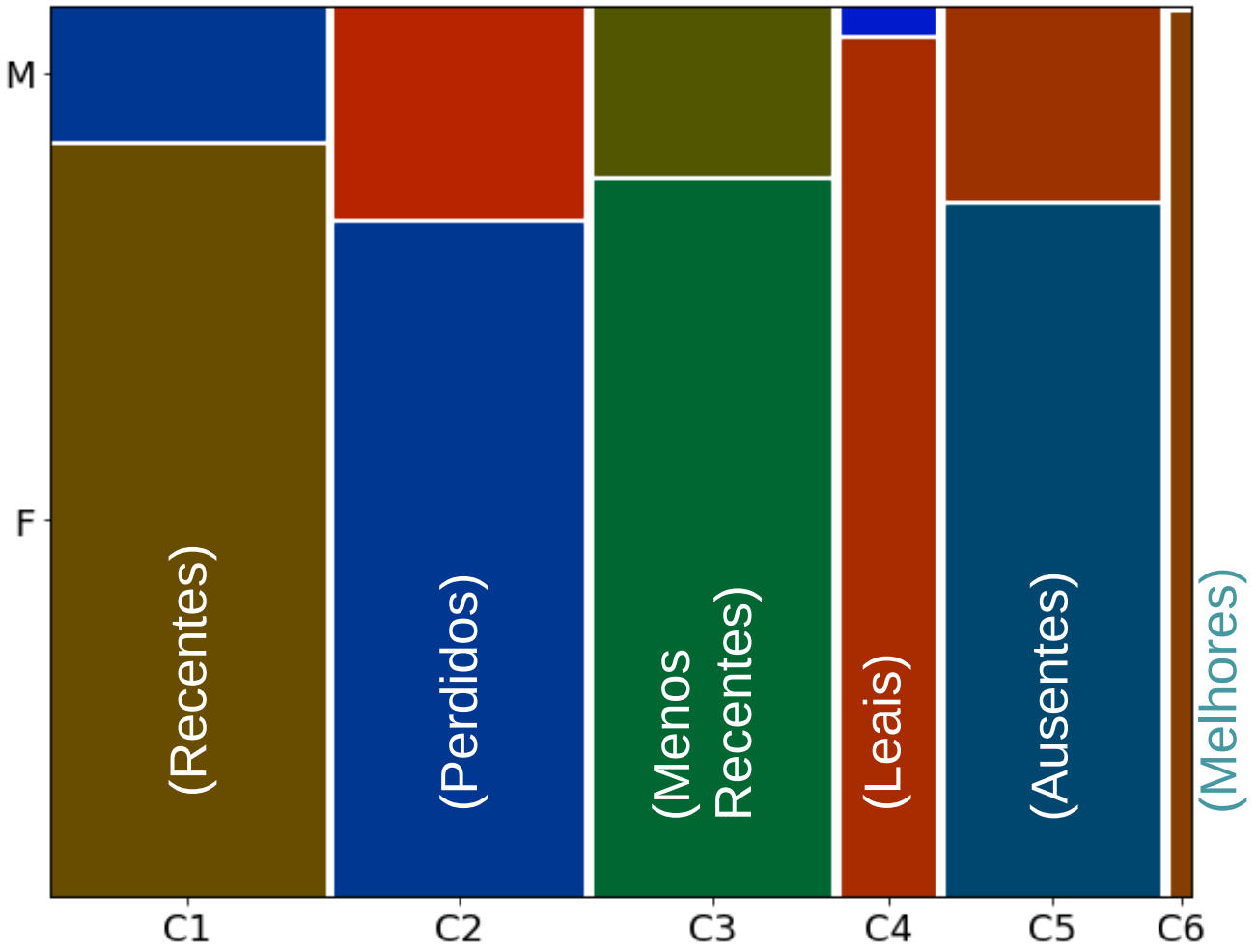
Análise dos Resultados

Gráfico de perfil:



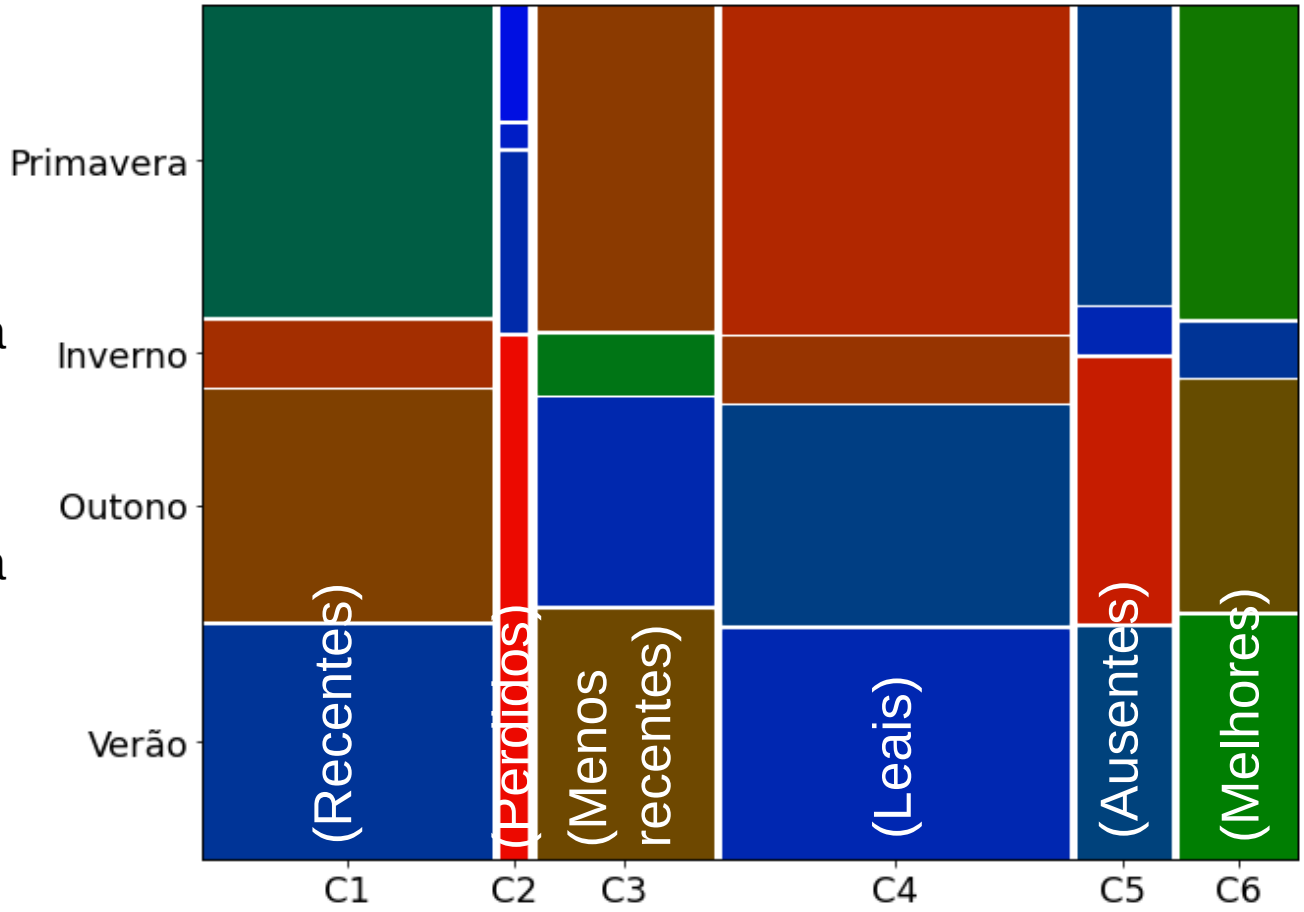
Mosaico (por gênero)

- Cluster 6 é composto apenas por mulheres
- Cluster 2 e 5 possuem quantidade anormalmente maior de homens



Mosaico (linha de roupa)

- Cluster 4 apresenta mais compras da linha de primavera
- Cluster 2 apresenta mais compras da linha de verão



Devoluções

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
9,09%	6,93%	8,11%	11,81%	8,68%	17,50%
(Recentes)	(Perdidos)	(Menos Recentes)	(Leais)	(Ausentes)	(Melhores)

Conclusões

- K-means, RFM, Validações Internas, Validações Externas
- Melhores clientes, clientes leais, clientes recentes, clientes menos recentes, clientes ausentes e clientes perdidos
- Dados comportamentais, preferências e tendências
- Nem sempre existem clusters naturais em bases reais

Sugestões

- Novo modelo de atributos (RFM)
- Bases diferentes (supermercados, imobiliárias concessionárias, entre outras..)
- Algoritmos diferentes (DBSCAN, hierárquico, Fuzzy C-means)

Obrigado!