Ciência de Dados
(Big Data Processing and Analytics)

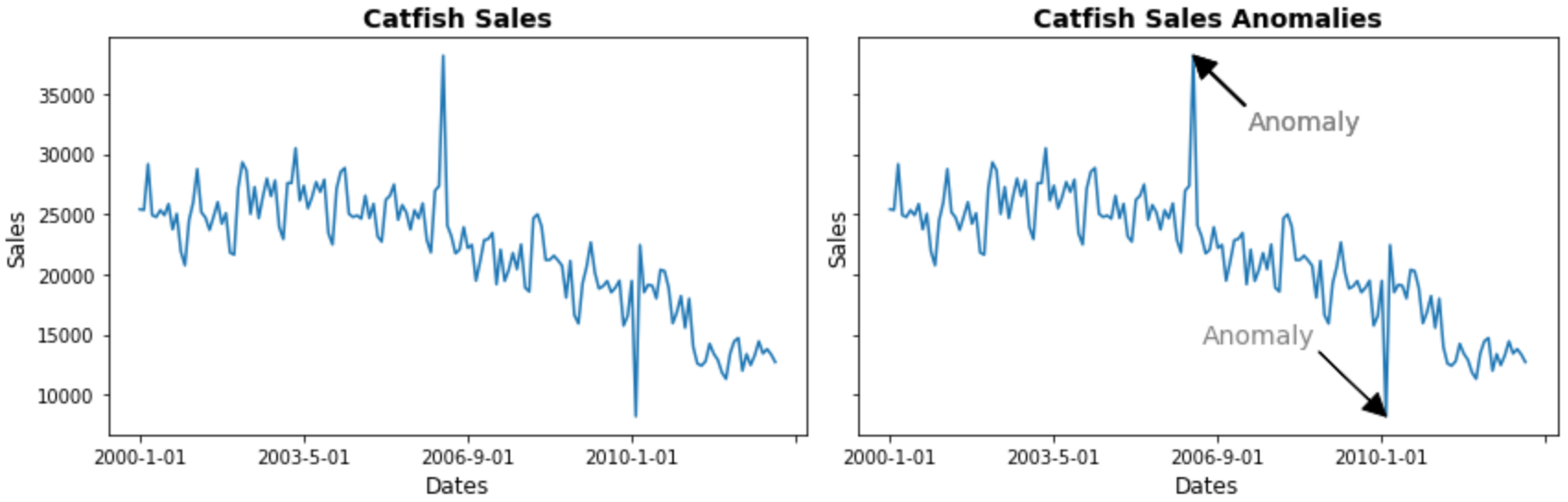
Big Data Analytics – Mineração e Análise de Dados

Professor curador
Prof. Dr. Rogério de Oliveira

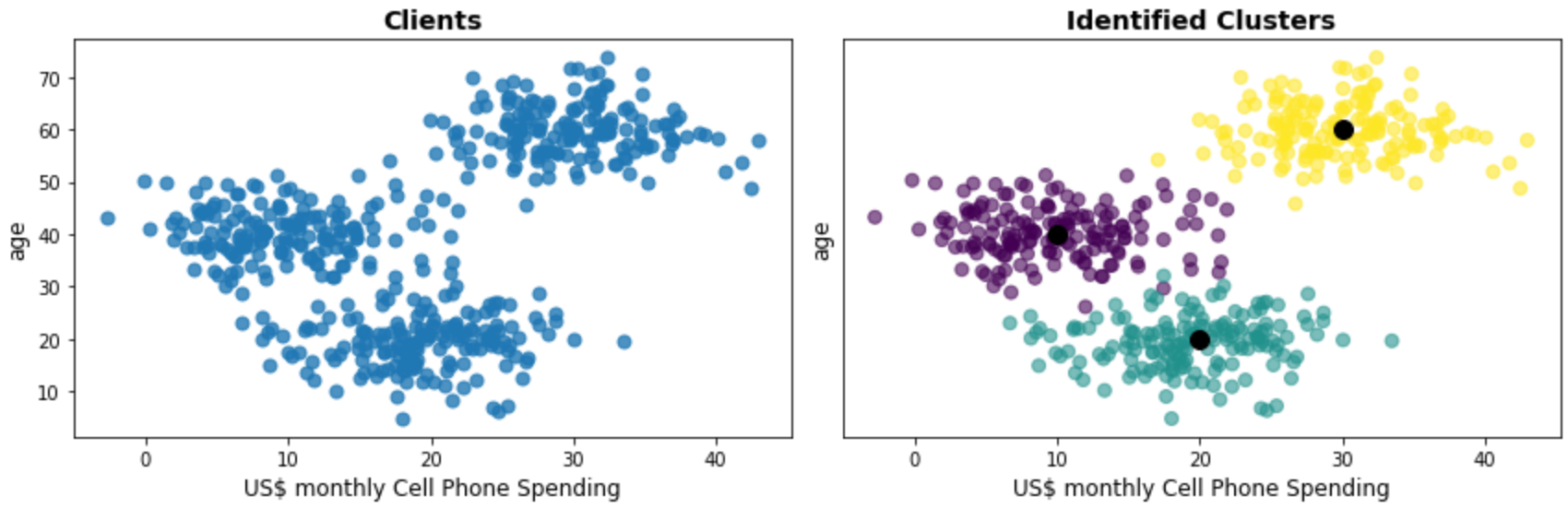
TRILHA 6
Aprendizado não Supervisionado: Clustering

Parte A

Aprendizado Não Supervisionado: Anomaly Detection



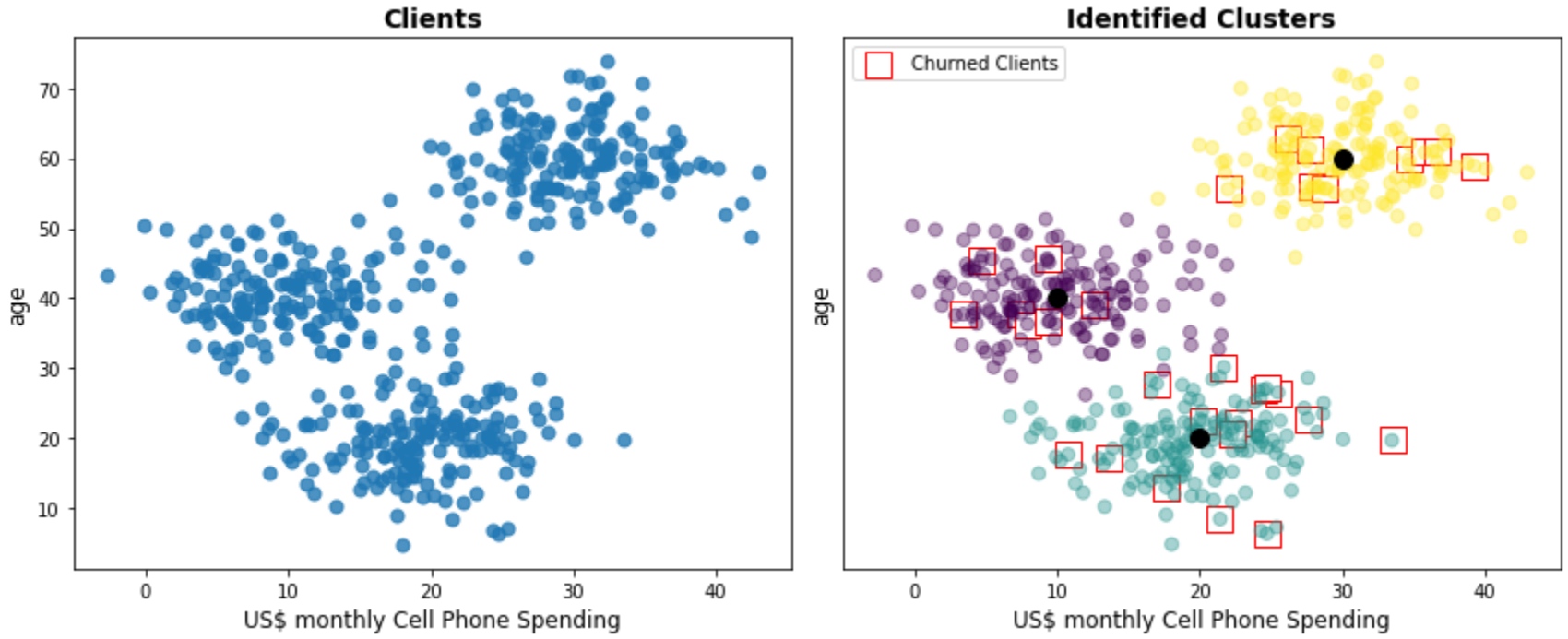
Aprendizado Não Supervisionado: Clustering



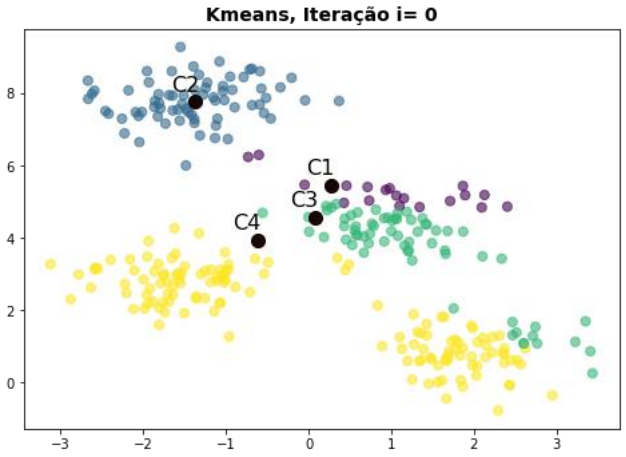
Aprendizado Supervisionado X Não Supervisionado

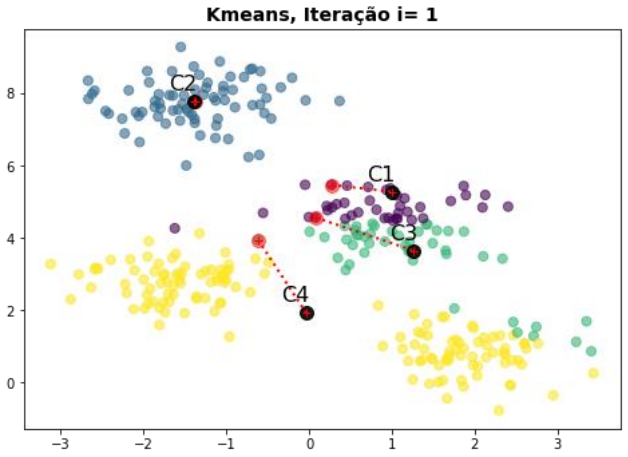
Característica	Aprendizado Supervisionado	Aprendizado não Supervisionado
Conjunto de Treinamento	Dados rotulados, entradas e saídas	Ausência de um conjunto de treinamento, há apenas dados não rotulados
Tipo de Tarefa	Preditivo	Analítico
Tarefas	Regressão e Classificação	Clusterização, detecção de anomalias, redução de dimensionalidade e mineração de regras de associação
Algoritmos	Regressão linear e logística, K-vizinhos mais próximos, Árvores de Decisão, SVM, Random Forest, Naive Bayes etc.	Clusterização Hierárquica, Kmédias, Floresta de Isolamento, PCA, SVD, apriori etc.
Complexidade computacional	Em geral mais simples	Em geral computacionalmente mais complexo

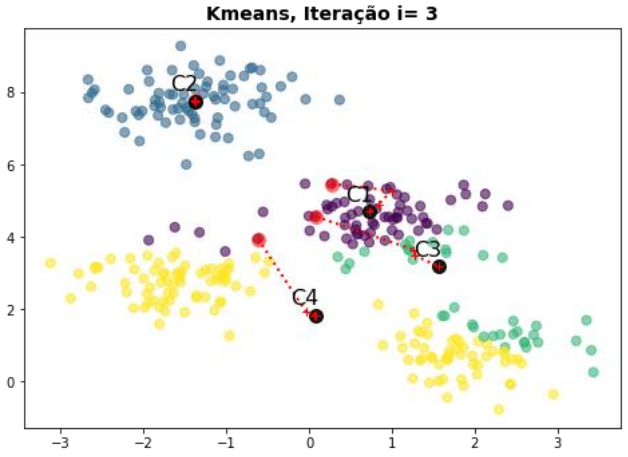
Classificação X Clusterização

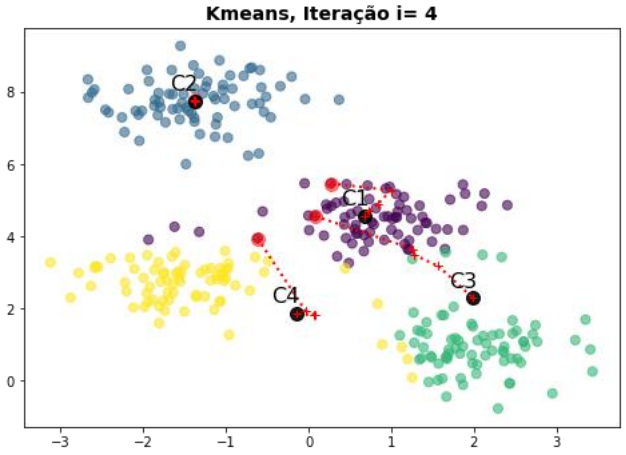


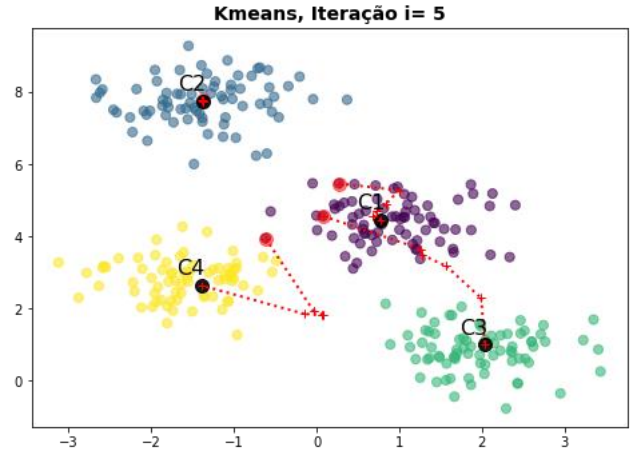
Clusterização Kmédias

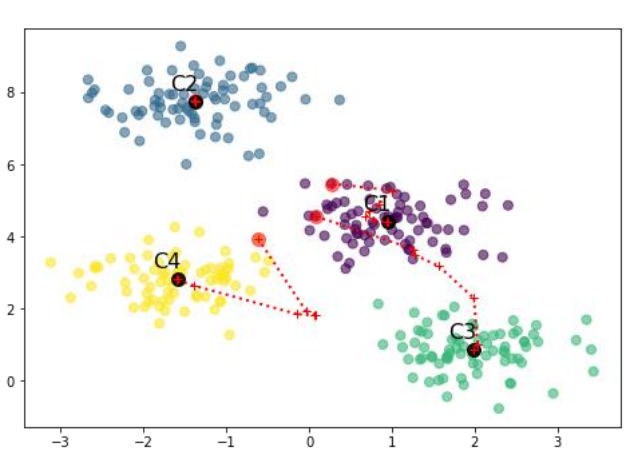






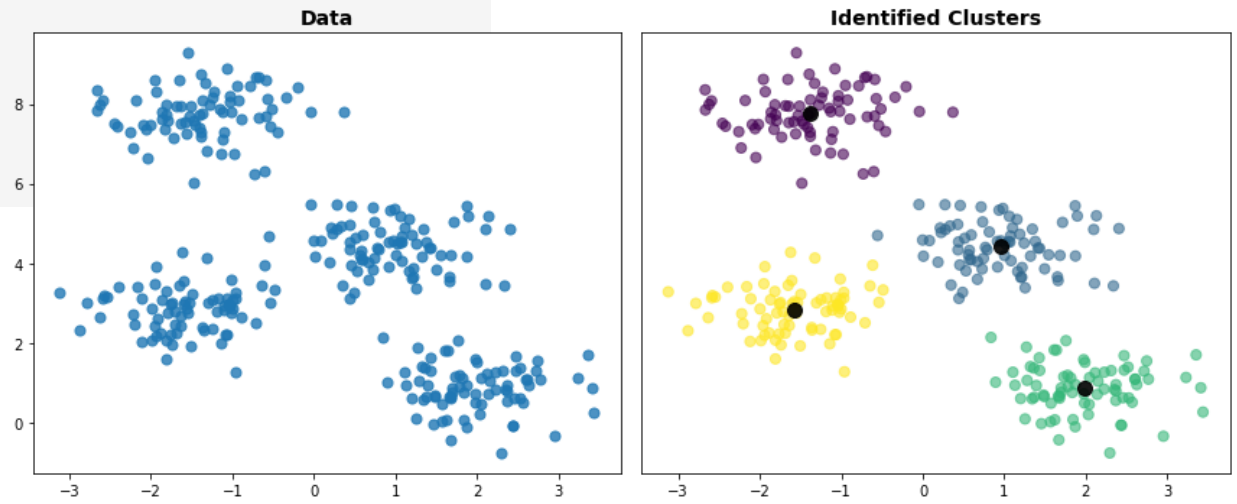






Clusterização Kmédias scikit-learn

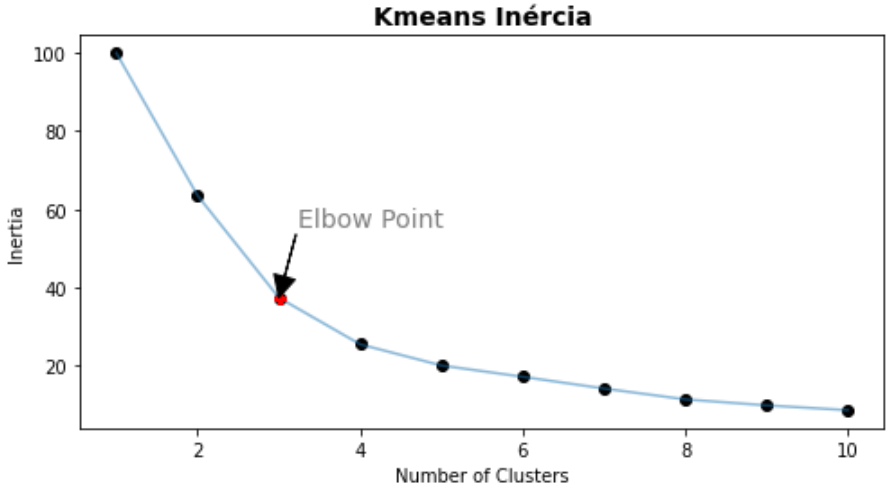
```
# Prepara os dados de entrada do estimador
X, _ = make_blobs(n_samples=300, centers=4, cluster_std=0.60, random_state=0)

# Configura e instancia o estimador
clf = KMeans(n_clusters = 4 , random_state= 1984)  # seed, para a reprodutibilidade

# Ajusta o estimador aos dados
clf.fit(X)

# Obtém os resultados do modelo
labels = clf.labels_
centroids = clf.cluster_centers_

print(labels)
print(centroids)
```

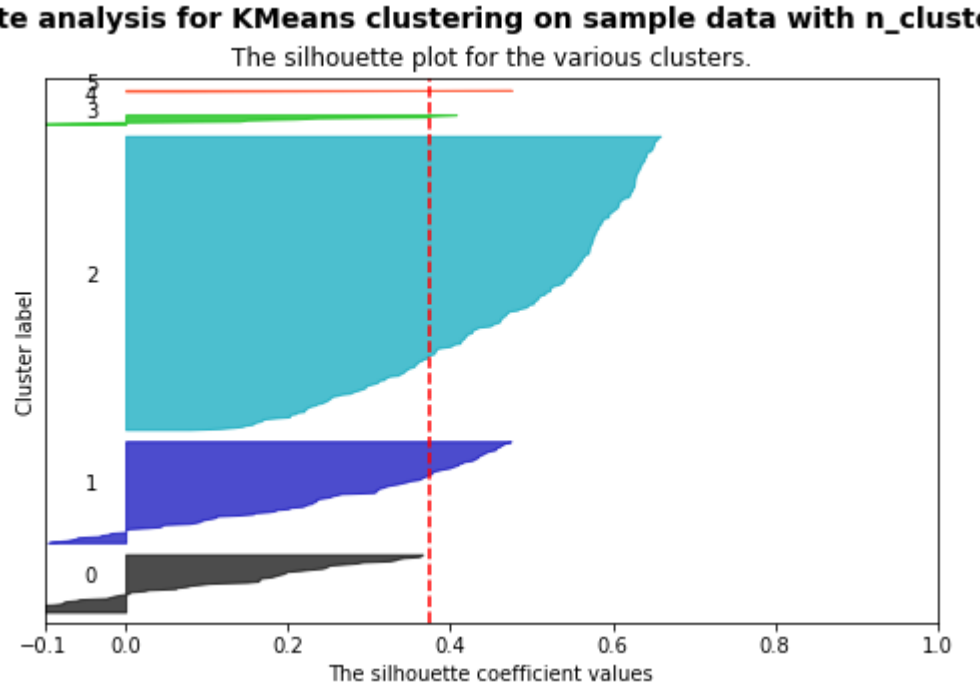


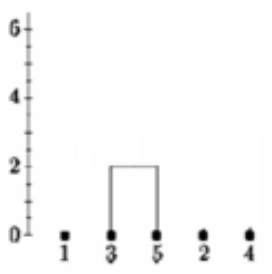
TRILHA 6
Aprendizado não Supervisionado: Clustering

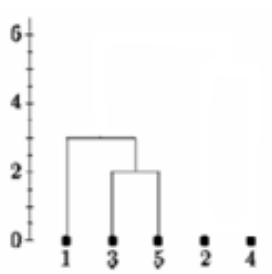
Parte B

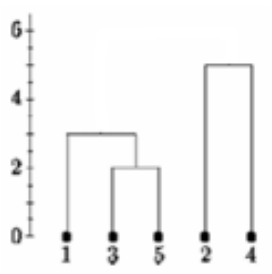
Melhor Número de Clusters

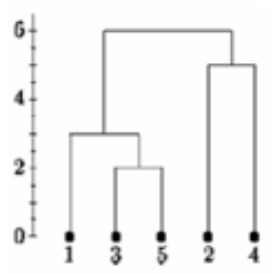












Clusterização Hierárquica: Dendograma

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

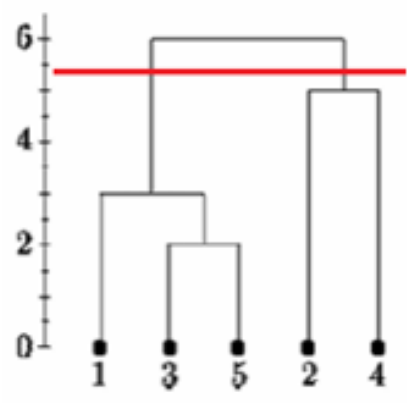
	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

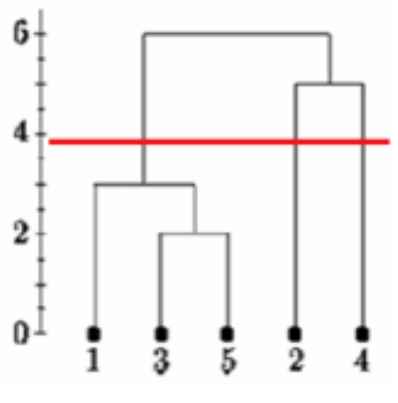
	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

Clusterização Hierárquica: Dendograma e Clusters



$C_1 = 1,3,5$

$C_2 = 2,4$



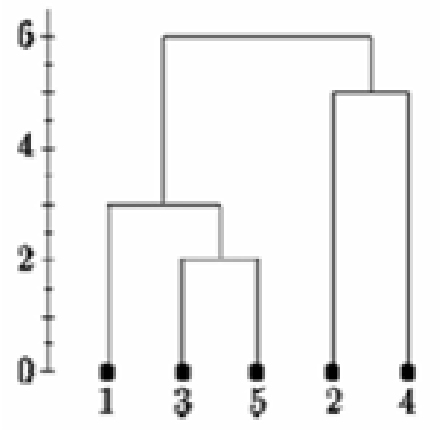
$C_1 = 1,3,5$

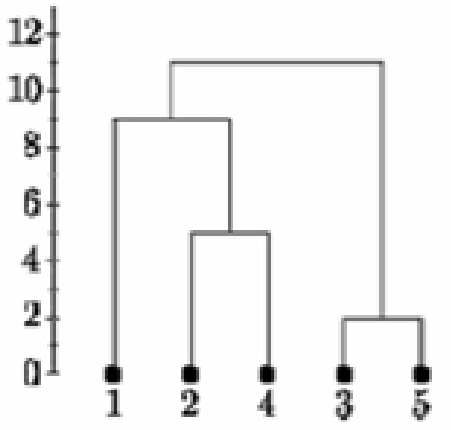
$S_1 = 2$

$S2 = 4$

Clusterização Hierárquica: Linkage

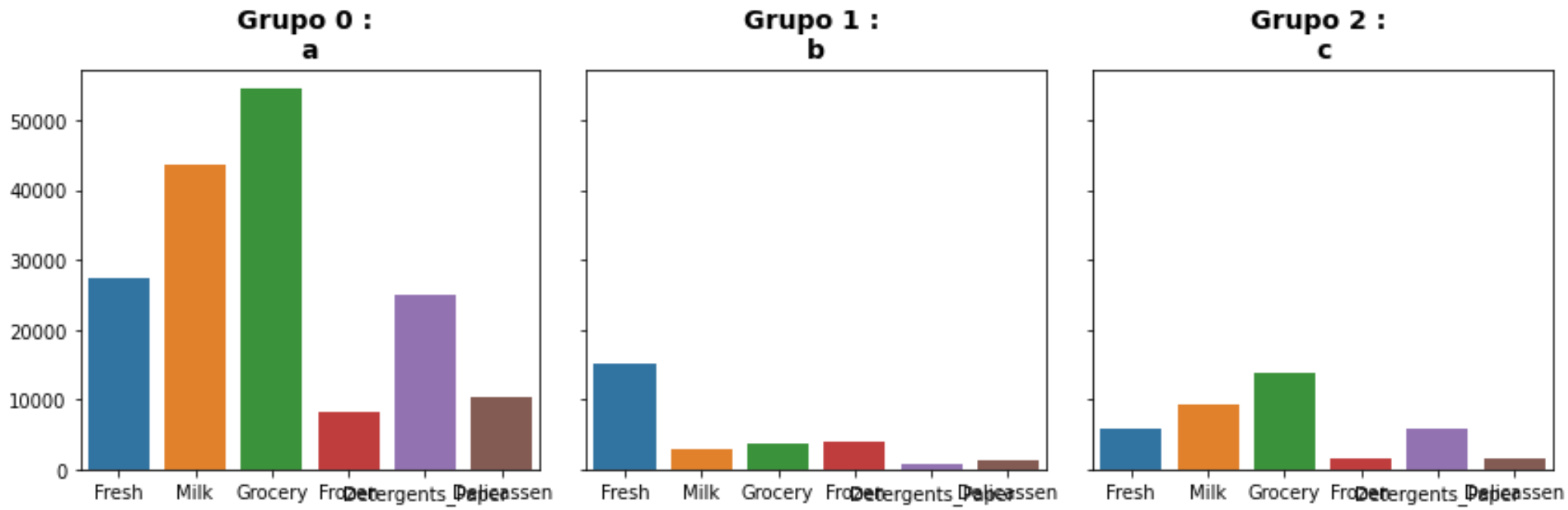
- $complete\_linkage(A, B) = \{\max\, d(a, b) : a \in A,\, b \in B\}$
- $single\_linkage(A, B) = \{\min\, d(a, b) : a \in A,\, b \in B\}$
- $average\_linkage(A, B) = \dfrac{1}{|A| \cdot |B|} \sum_{a \in A} \sum_{b \in B} d(a, b)$





	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

Caracterizando os Grupos: Quem está em cada grupo?



Universidade Presbiteriana
Mackenzie