FUNDAMENTALS OF DATA MINING

Clustering & Frequent Pattern Mining

Semiha Hazel Çavuş 150115045
Zeynep Kumralbaş 150115051

Problem Definition

- Analyze 3 clustering algorithms on ‘seeds data set’.
- Analyze 3 frequent pattern mining algorithms.

Seeds Data Set

Kernels belonging to 3 different varieties of wheat:

- Kama
- Rosa
- Canadian

70 samples for each class

Data Set: Attributes, Data Characteristics

- Area, A
- Perimeter, P
- Compactness, C=4*pi*A/P^2
- Length of kernel
- Width of kernel
- Asymmetry coefficient
- Length of kernel Groove

Number of instances: 210 Number of attributes: 7

Attribute characteristics: real valued, continuous

Data Preprocessing

- No missing values
- All values are normalized into 0-1

IDE/Environment

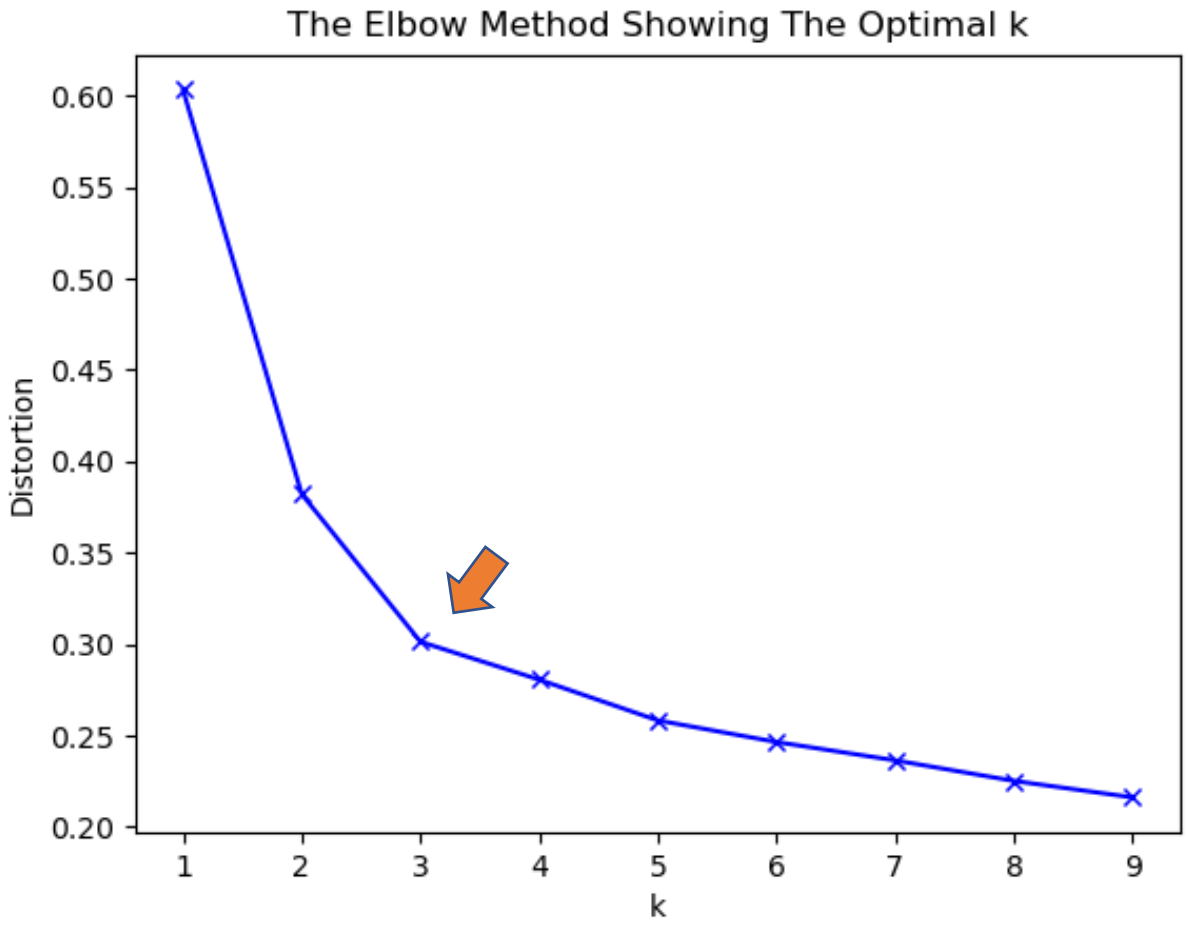
Environment:

IDE:

Libraries

ECLAT: https://github.com/adoku14/Eclat-Python-Implementation/blob/master/eclat.py

K-Means: Determining Number of Clusters



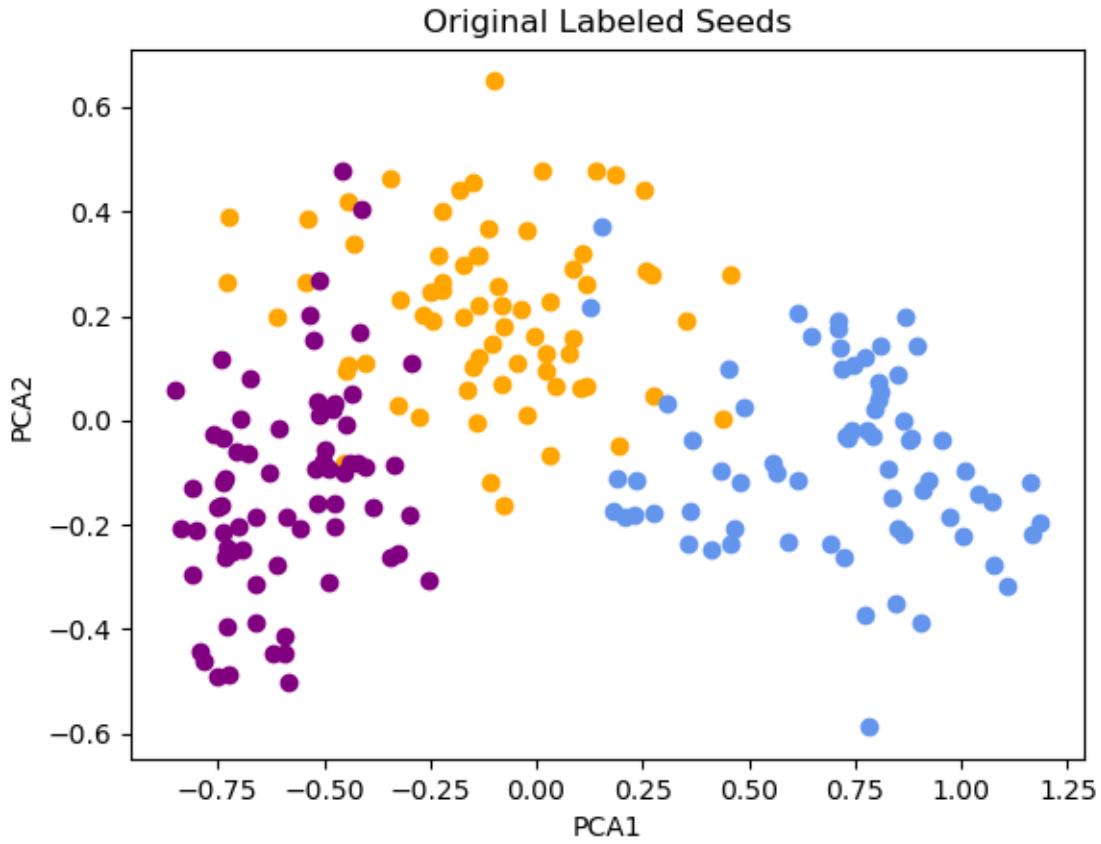
K-Means: Parameters

Parameters of Kmeans method:

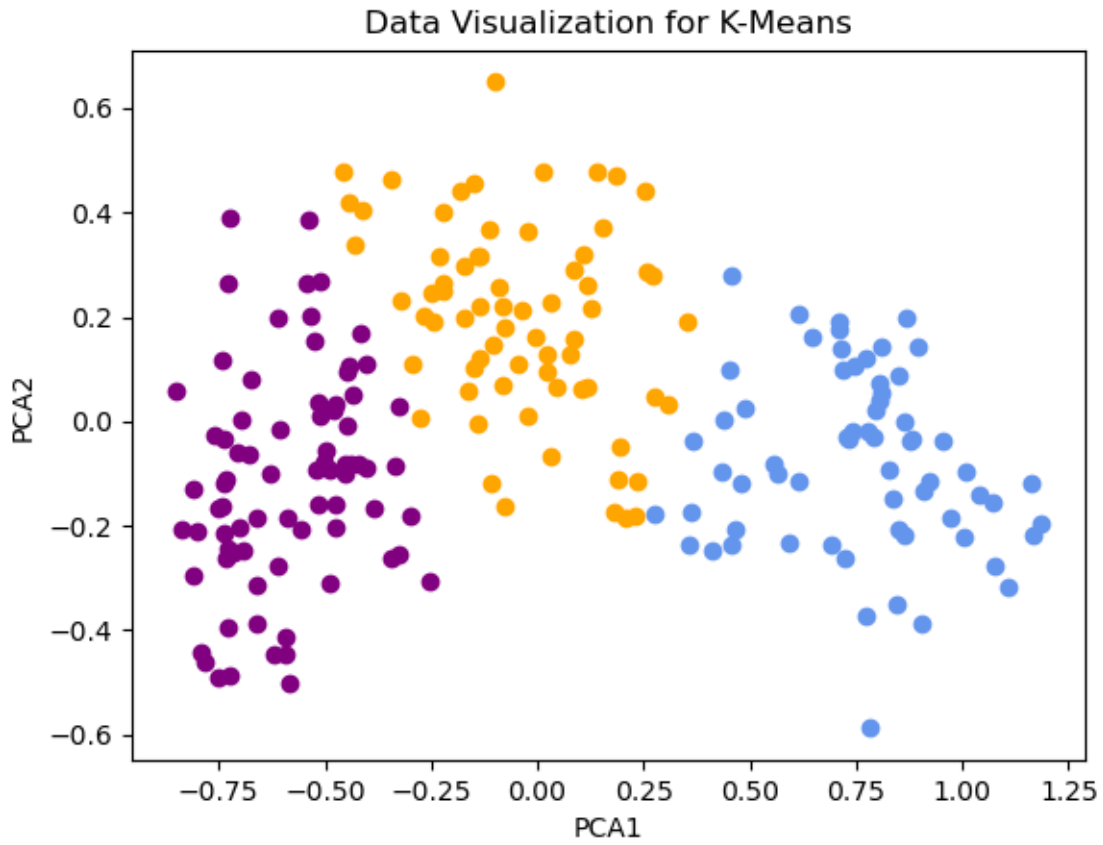
- n_clusters=3
- init='k-means++’
- n_initint, default=10
- max_iterint, default=300
- tolfloat, default=1e-4

sklearn.cluster.Kmeans(n_clusters=3)

K-Means: Original Clusters

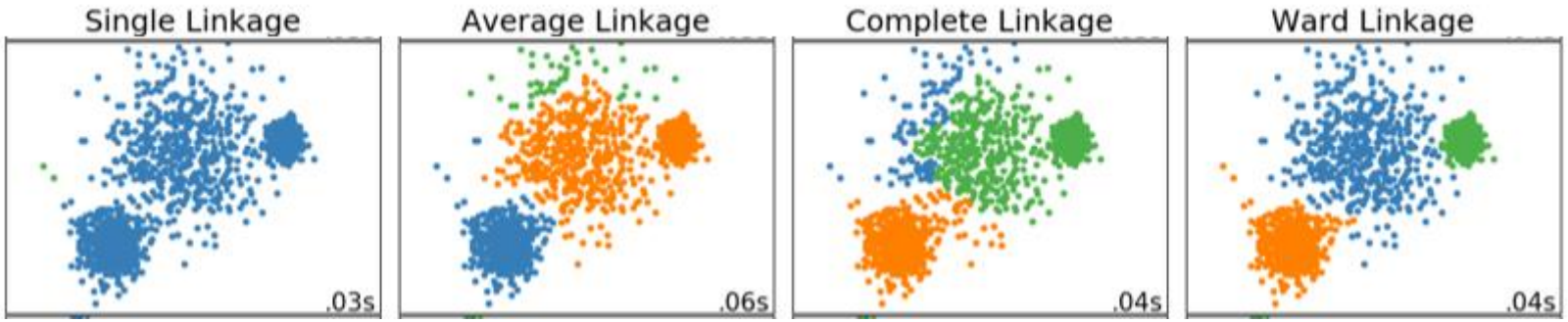


K-Means: Clusters



AGNES: Dendrogram Parameters

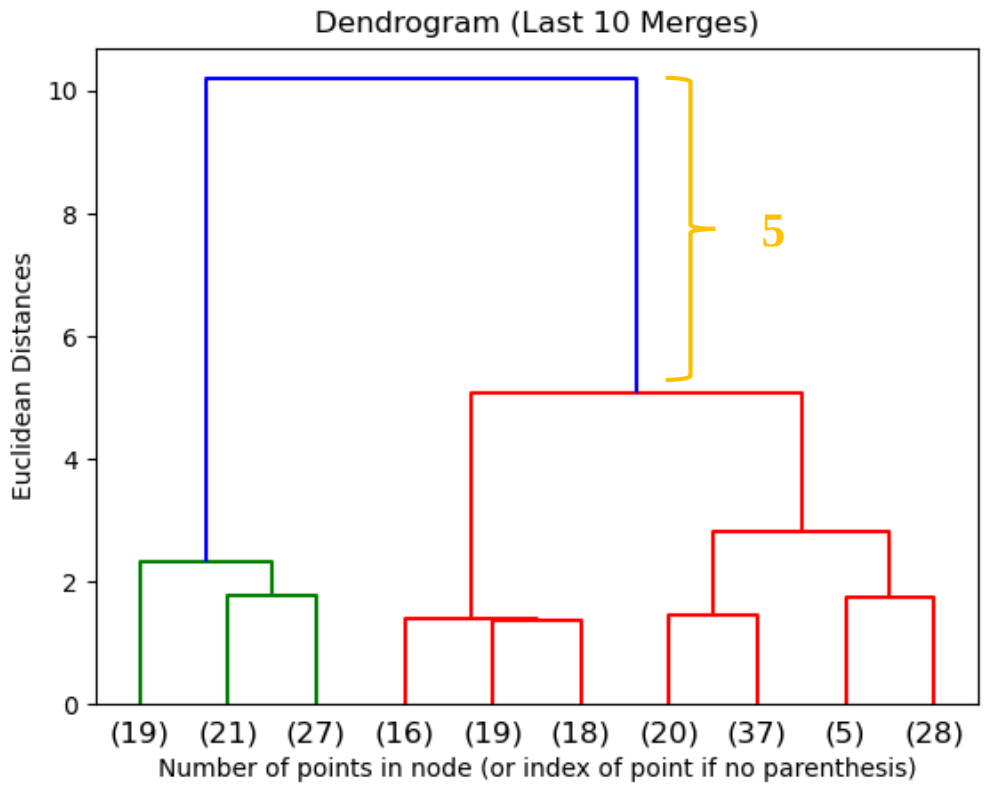
dendrogram(linkage(data, method='ward', metric='euclidean'))

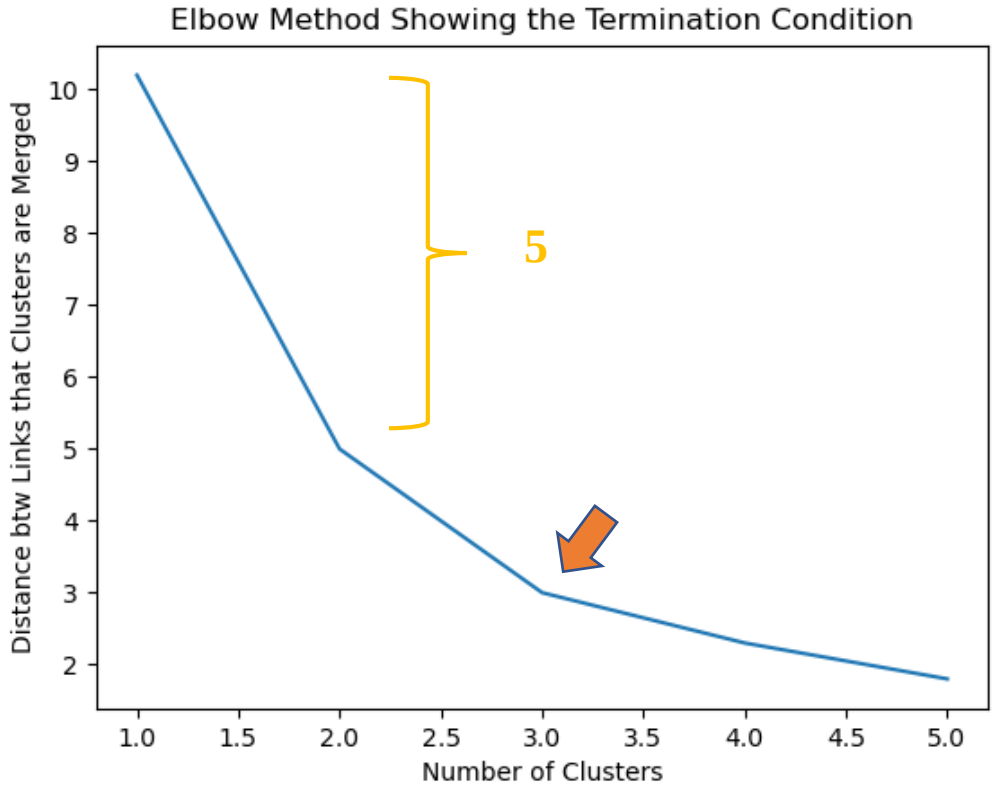


Different linkage types [1]

[1] https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering

AGNES: Finding Termination Condition



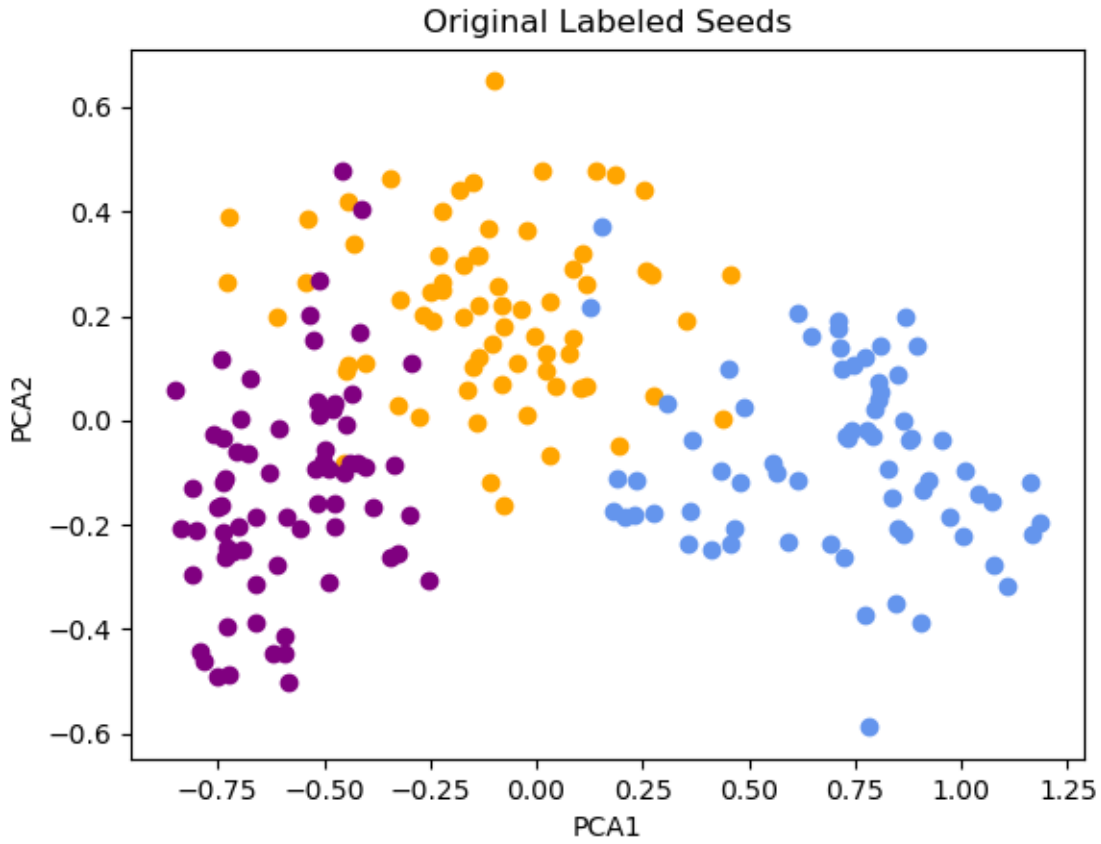


AGNES: Parameters

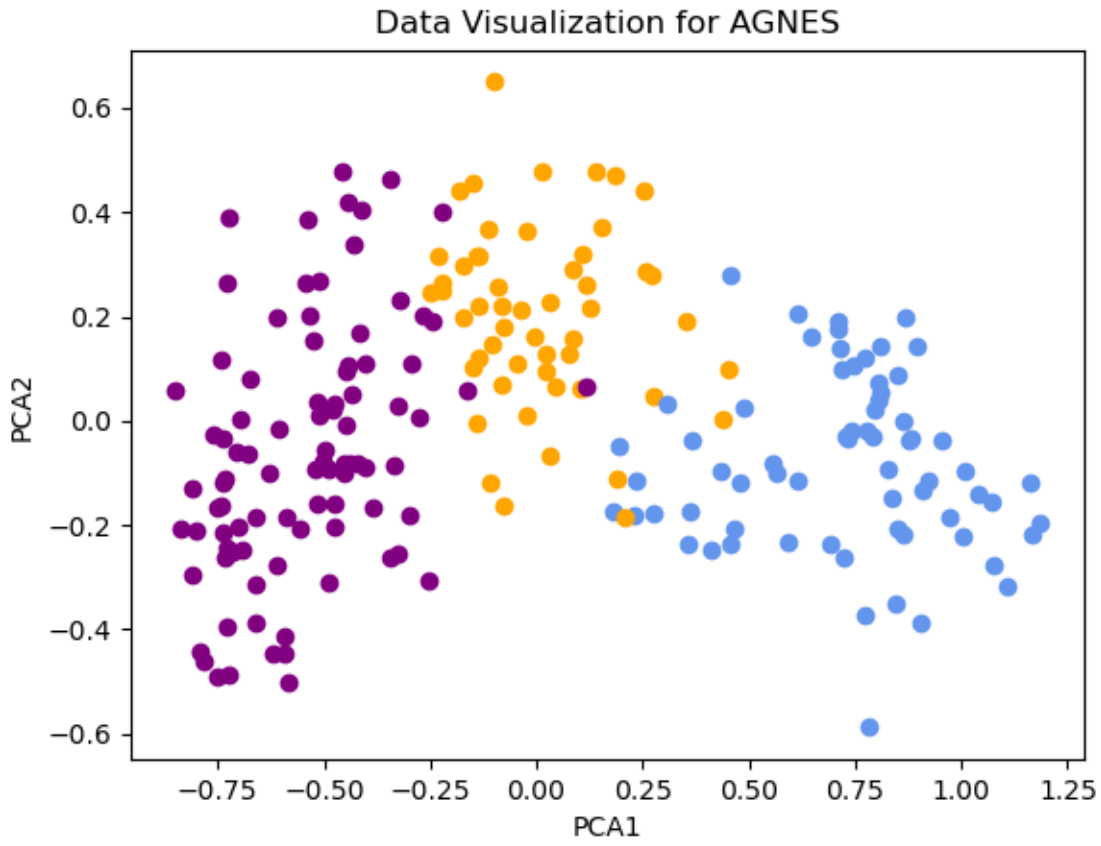
AgglomerativeClustering parameters:

- n_clusters=None
- distance_threshold=3
- affinity='euclidean’
- linkage='ward’)

AGNES: Original Clusters



AGNES: Clusters



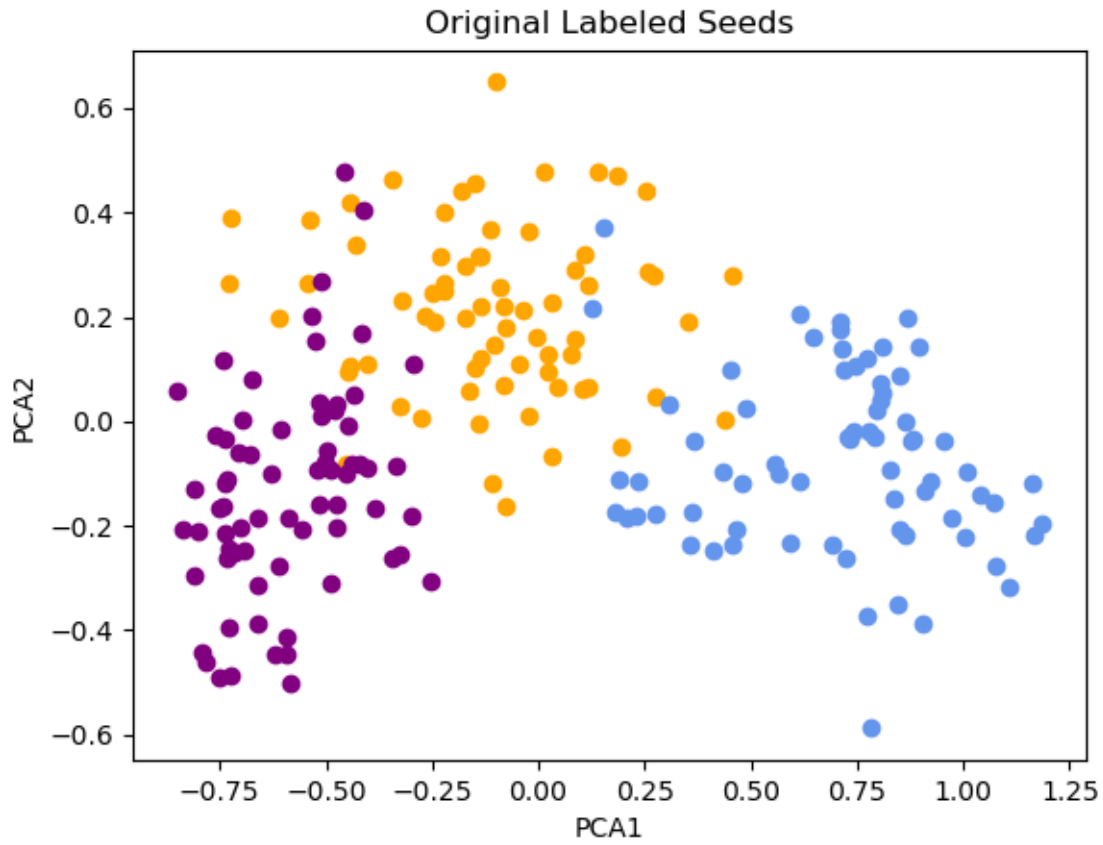
DBSCAN: Heuristic[2] for Choosing Epsilon

- ✓Calculate distances between each point and its k^{th} nearest neighbour.
- ✓Sort distances in descending order.
- ✓Plot the k-dist graph, where x axis is the point indexes and y-axis is the distances.
- ✓Find the elbow on the graph and set epsilon to corresponding distance.

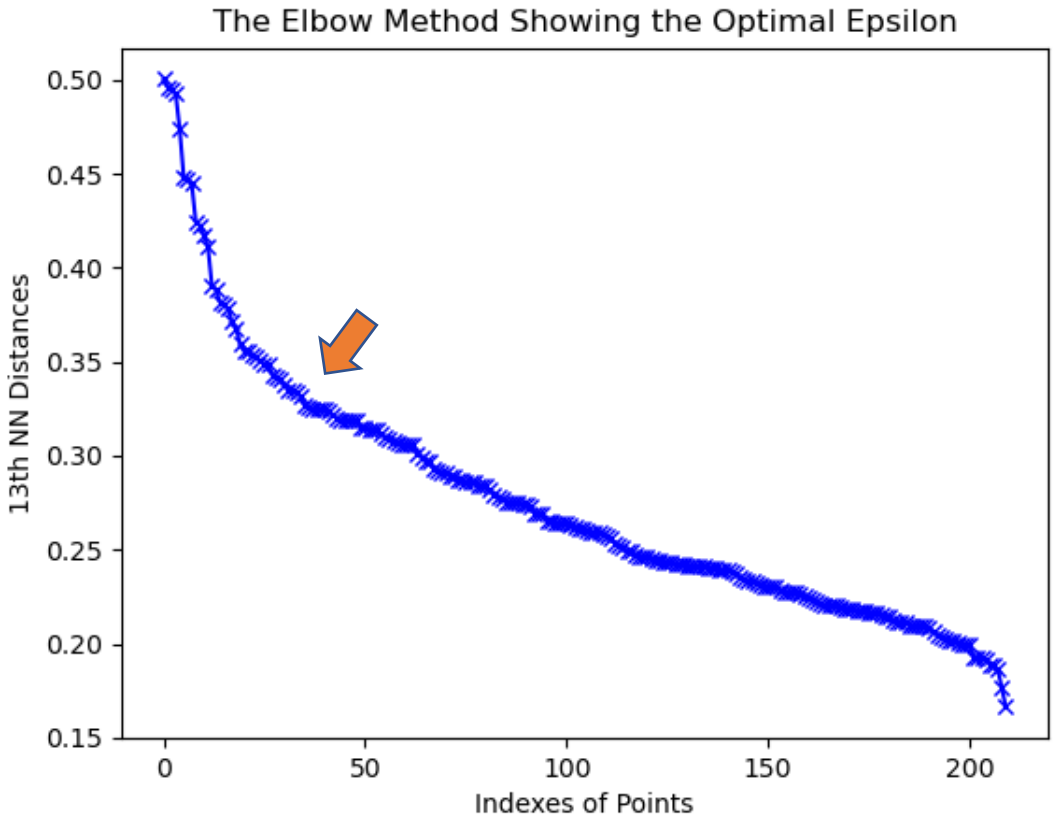
Rule of thumb [3]:	MinPts ≥ dimension + 1	k = MinPts - 1
	MinPts = 2*dimension	

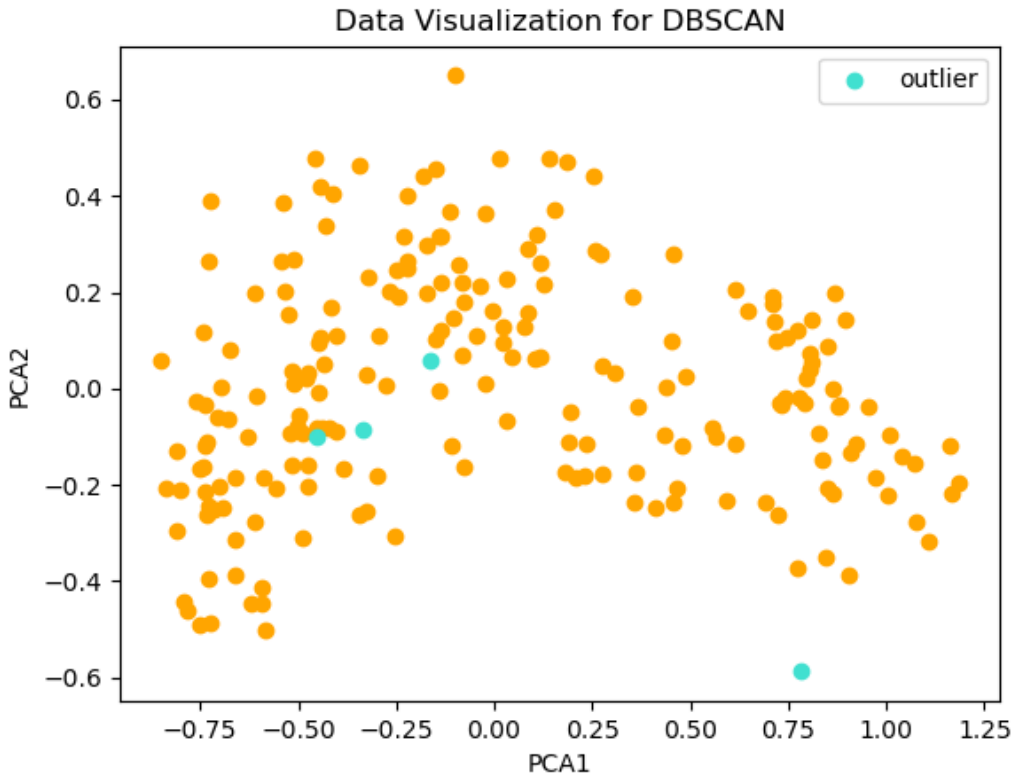
[2] ESTER, Martin, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Kdd*. 1996. p. 226-231.
[3] SCHUBERT, Erich, et al. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)*, 2017, 42.3: 1-21.

DBSCAN: Original Clusters



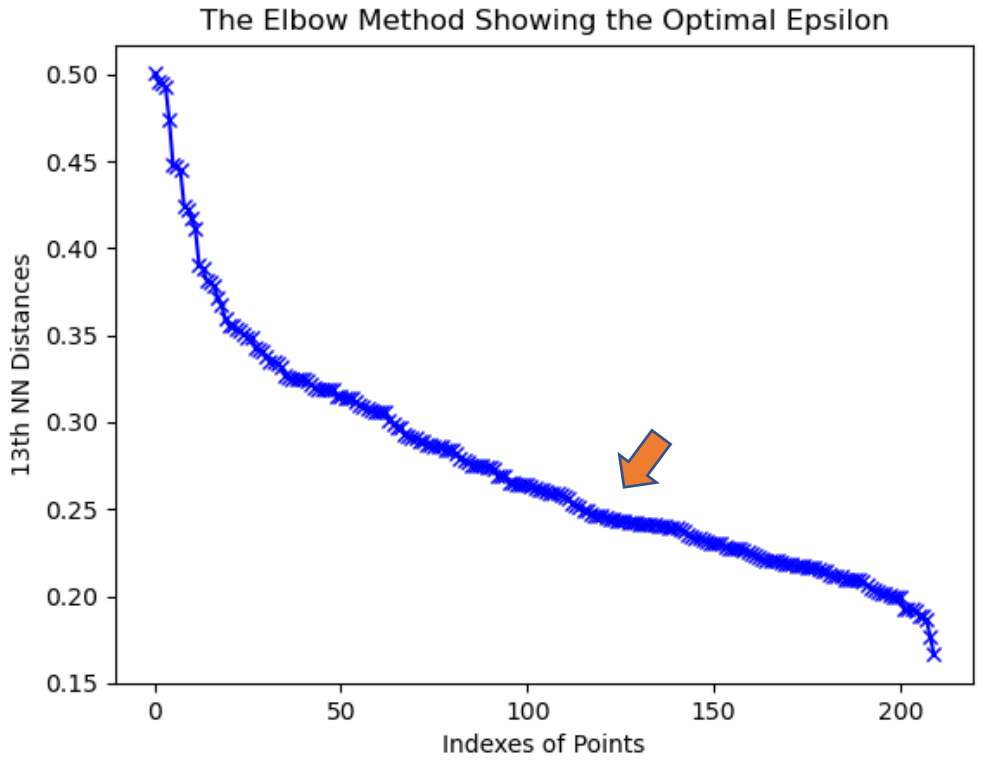
DBSCAN: k-dist Graph

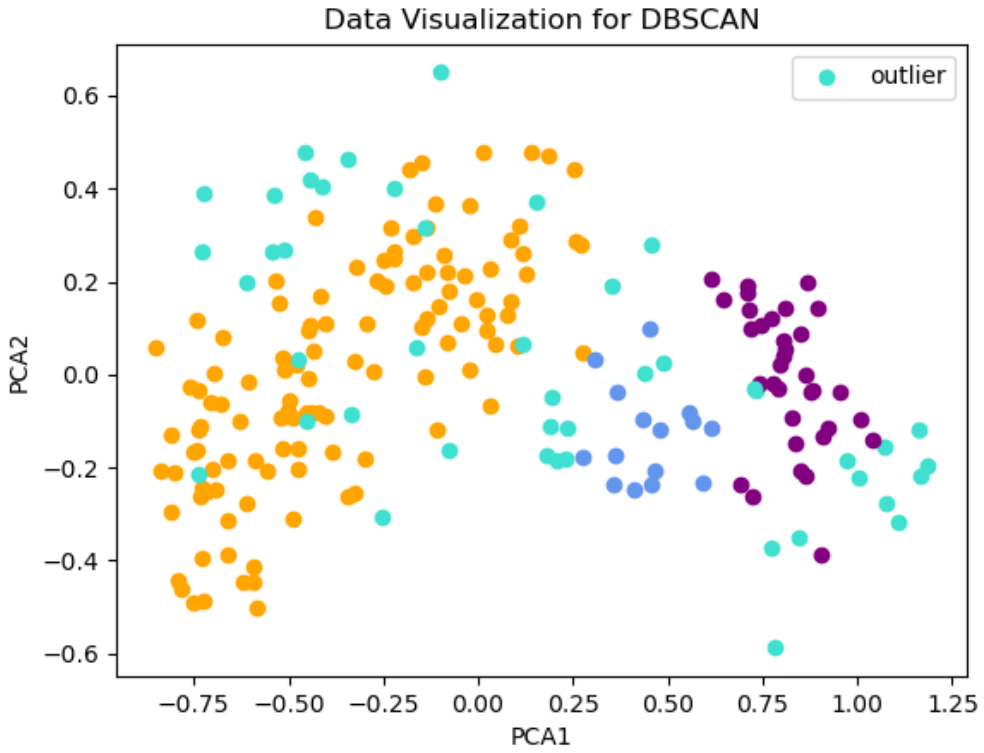




Eps = 0.32
min_samples = 14

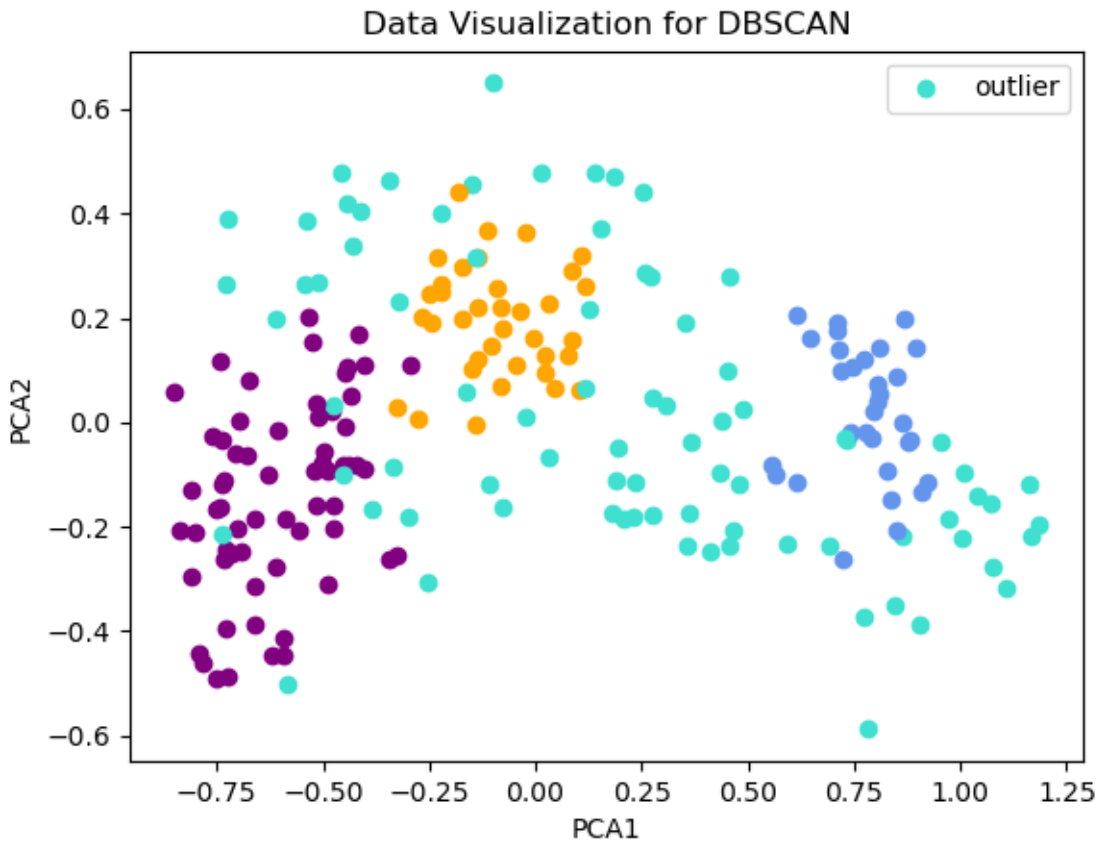
DBSCAN: k-dist Graph





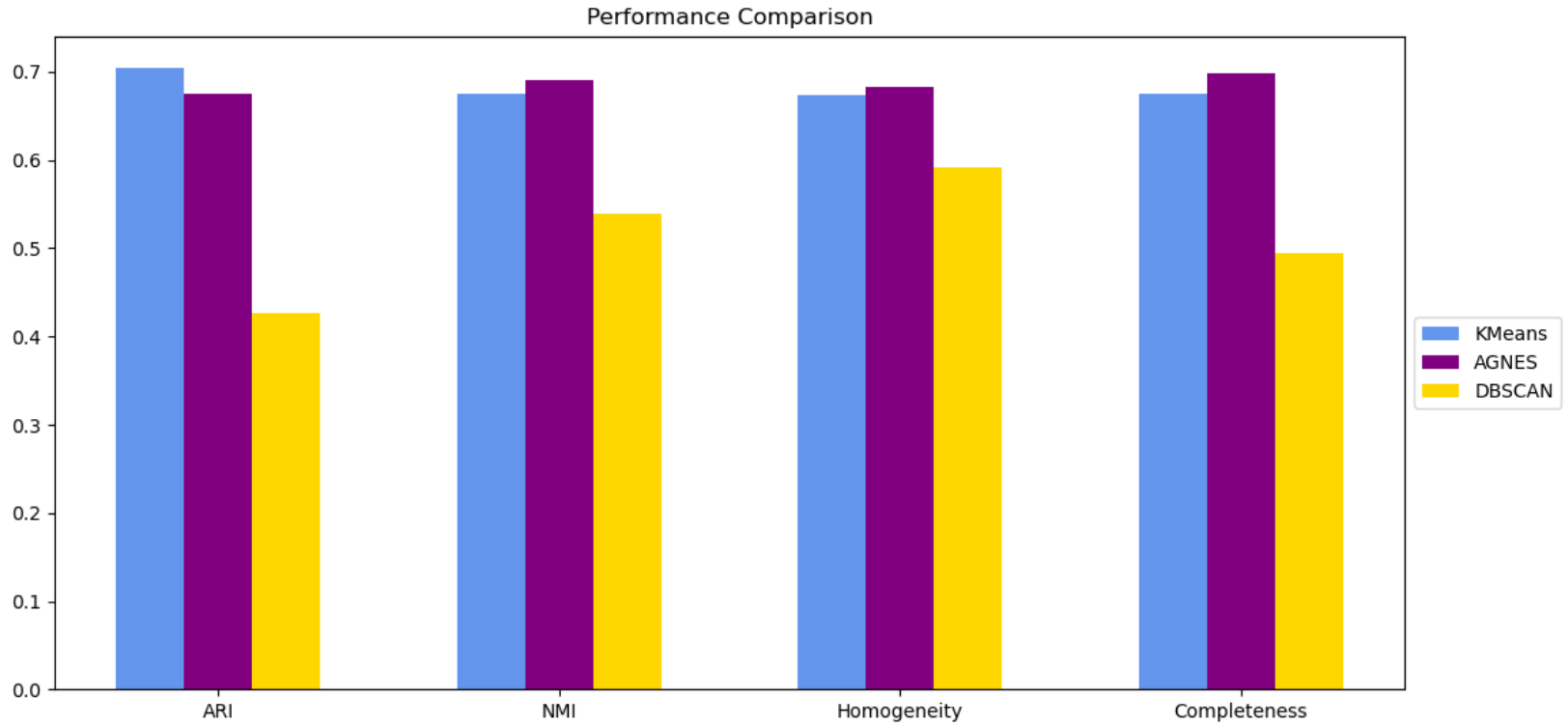
Eps = 0.24
min_samples = 14

DBSCAN: Clusters



Eps = 0.26
min_samples = 25

Comparison of Clustering Algorithms



Frequent Pattern Mining: Transaction Encoding

	Bread	Butter	Cheese	Coffee Powder	Ghee	Lassi	Milk	Panner	Sugar	Sweet	Tea Powder	Yougurt
0	False	True	True	True	True	True	False	False	False	False	False	True
1	False	False	False	True	True	False	False	False	False	False	False	False
2	False	True	True	False	False	True	False	False	False	False	True	False
3	True	True	True	True	False	False	False	True	False	False	True	False
4	False	True	True	True	False	False	False	False	True	True	False	True
...	...	...	...	...	...	...	...	...	...	...	...	...
12521	True	False	True	False	False	False	True	True	True	False	False	False
12522	True	False	True	True	False	False	False	False	True	False	False	False
12523	True	False	True	False	False	False	True	False	False	False	False	True
12524	True	False	True	False	True	False	False	False	True	False	False	True
12525	True	False	False	False	False	False	False	True	False	False	False	True

Apriori

Apriori parameters:

- dataFrame
- min_support = 0.2
- use_colnames = True

support	itemsets
0.441162	(Milk)
0.439885	(Ghee)
0.439805	(Coffee Powder)
0.439326	(Yougurt)
0.437809	(Bread)
0.437730	(Sweet)
0.437650	(Sugar)
0.437570	(Butter)
0.437171	(Cheese)
0.434616	(Panner)
0.433658	(Lassi)
0.429746	(Tea Powder)
0.205812	(Coffee Powder, Ghee)
0.205652	(Sweet, Lassi)
0.205253	(Butter, Sugar)
0.204614	(Milk, Sugar)
0.203976	(Coffee Powder, Yougurt)
0.203577	(Panner, Bread)
0.203018	(Butter, Sweet)
0.202698	(Milk, Lassi)
0.202698	(Sweet, Bread)
0.202140	(Yougurt, Cheese)
0.201980	(Butter, Ghee)
0.201980	(Bread, Cheese)
0.201900	(Butter, Yougurt)
0.201900	(Sugar, Yougurt)
0.201820	(Bread, Coffee Powder)
0.201421	(Panner, Ghee)
0.201022	(Milk, Coffee Powder)
0.200942	(Coffee Powder, Cheese)
0.200942	(Milk, Bread)
0.200862	(Sugar, Ghee)
0.200623	(Milk, Yougurt)
0.200543	(Lassi, Coffee Powder)
0.200543	(Milk, Sweet)
0.200463	(Lassi, Ghee)
0.200463	(Milk, Ghee)
0.200144	(Bread, Yougurt)
0.200064	(Bread, Lassi)

FP-Growth

FP-Growth parameters:

- dataFrame
- min_support = 0.2
- use_colnames = True

support	itemsets
0.441162	(Milk)
0.439885	(Ghee)
0.439805	(Coffee Powder)
0.439326	(Yougurt)
0.437809	(Bread)
0.437730	(Sweet)
0.437650	(Sugar)
0.437570	(Butter)
0.437171	(Cheese)
0.434616	(Panner)
0.433658	(Lassi)
0.429746	(Tea Powder)
0.205812	(Coffee Powder, Ghee)
0.205652	(Sweet, Lassi)
0.205253	(Butter, Sugar)
0.204614	(Milk, Sugar)
0.203976	(Coffee Powder, Yougurt)
0.203577	(Panner, Bread)
0.203018	(Butter, Sweet)
0.202698	(Milk, Lassi)
0.202698	(Sweet, Bread)
0.202140	(Yougurt, Cheese)
0.201980	(Butter, Ghee)
0.201980	(Bread, Cheese)
0.201900	(Butter, Yougurt)
0.201900	(Sugar, Yougurt)
0.201820	(Bread, Coffee Powder)
0.201421	(Panner, Ghee)
0.201022	(Milk, Coffee Powder)
0.200942	(Coffee Powder, Cheese)
0.200942	(Milk, Bread)
0.200862	(Sugar, Ghee)
0.200623	(Milk, Yougurt)
0.200543	(Lassi, Coffee Powder)
0.200543	(Milk, Sweet)
0.200463	(Lassi, Ghee)
0.200463	(Milk, Ghee)
0.200144	(Bread, Yougurt)
0.200064	(Bread, Lassi)

Eclat

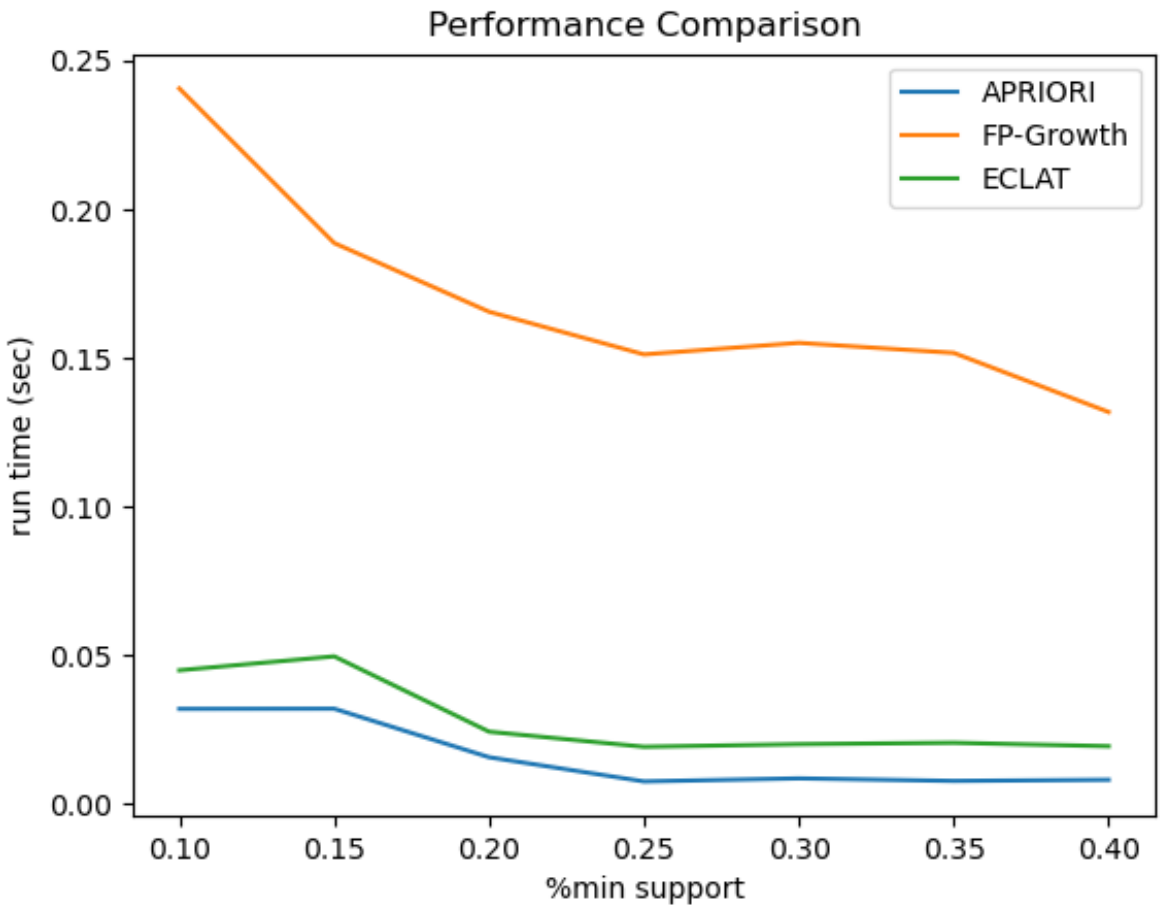
Eclat parameters:

- min_support = 0.2

```
['Milk'] : 0.4412
['Ghee'] : 0.4399
['Coffee Powder'] : 0.4398
['Yougurt'] : 0.4393
['Bread'] : 0.4378
['Sweet'] : 0.4377
['Sugar'] : 0.4376
['Butter'] : 0.4376
['Cheese'] : 0.4372
['Panner'] : 0.4346
['Lassi'] : 0.4337
['Tea Powder'] : 0.4297
['Coffee Powder', 'Ghee'] : 0.2058
['Sweet', 'Lassi'] : 0.2057
['Butter', 'Sugar'] : 0.2053
['Milk', 'Sugar'] : 0.2046
['Coffee Powder', 'Yougurt'] : 0.204
['Panner', 'Bread'] : 0.2036
['Butter', 'Sweet'] : 0.203
['Milk', 'Lassi'] : 0.2027
['Sweet', 'Bread'] : 0.2027
```

```
['Yougurt', 'Cheese'] : 0.2021
['Bread', 'Cheese'] : 0.202
['Butter', 'Ghee'] : 0.202
['Butter', 'Yougurt'] : 0.2019
['Sugar', 'Yougurt'] : 0.2019
['Bread', 'Coffee Powder'] : 0.2018
['Panner', 'Ghee'] : 0.2014
['Milk', 'Coffee Powder'] : 0.201
['Coffee Powder', 'Cheese'] : 0.2009
['Milk', 'Bread'] : 0.2009
['Sugar', 'Ghee'] : 0.2009
['Milk', 'Yougurt'] : 0.2006
['Lassi', 'Coffee Powder'] : 0.2005
['Milk', 'Sweet'] : 0.2005
['Lassi', 'Ghee'] : 0.2005
['Milk', 'Ghee'] : 0.2005
['Bread', 'Yougurt'] : 0.2001
['Bread', 'Lassi'] : 0.2001
```

Performance Comparison



THANKS FOR LISTENING