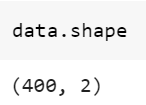
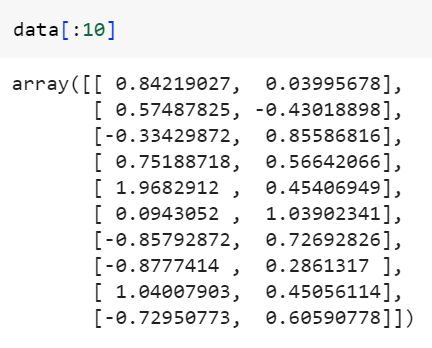
Pragya Bhatia

[W1968251@westminster.ac.uk](mailto:W1968251@westminster.ac.uk)

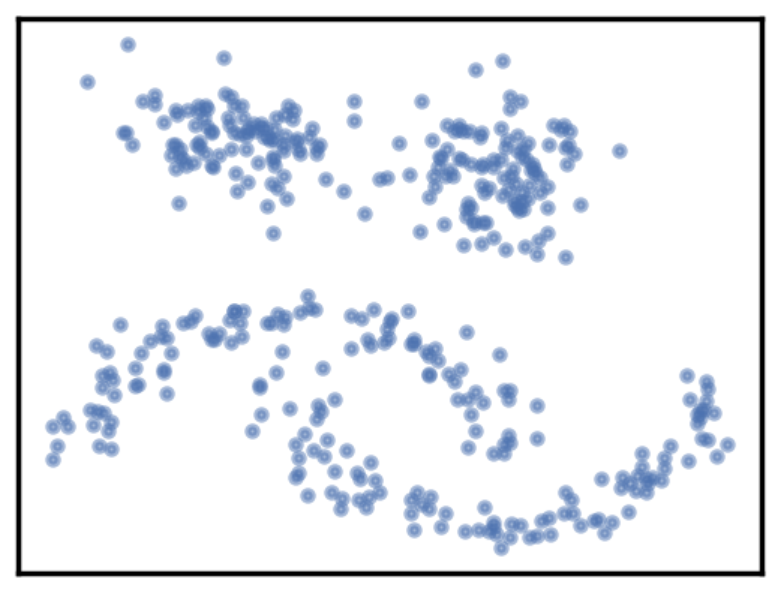
TASK 1: Preparing your test Environment:

1. Show evidence of Loading the 'data.npy' data file.

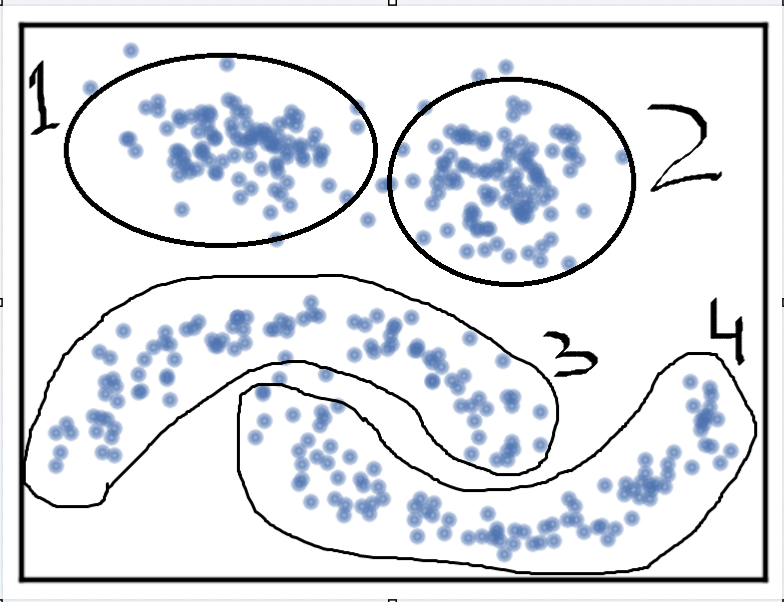




1. Scatter plot the 'data.npy' data.



1. On your Scatter plot, annotate the various groups of data as you observe them.



1. Establish the plot\_clusters (data, algorithm, args, kwds) little utility function call (syntax only), which does the clustering and the plotting of the results for each of the algorithms: K-Means, DBSCAN, Birch, OPTICS, Affinity Propagation, HDBSCAN, Agglomerative Clustering, Spectral Clustering

|  |  |
| --- | --- |
| Algorithm Name | Syntax |
| K-Means | plot\_clusters (data, KMeans, (), {‘n\_clusters’:value, ‘init’:value, ‘max\_iter’:value, ‘algorithm’:value}) |
| Affinity Propagation | plot\_clusters (data, AffinityPropagation, (), {‘preference’:value, ‘damping’:value, ‘convergence\_iter’:value, ‘max\_iter’:value, ‘affinity’=value'}) |
| DBSCAN | plot\_clusters (data, DBSCAN, (), {‘eps’:value, ‘min\_samples’:value, ‘metric’:value}) |
| Birch | plot\_clusters (data, Birch, (), {‘n\_clusters’:value, ‘branching\_factor’:value, ‘threshold’:value }) |
| OPTICS | plot\_clusters (data, OPTICS, (), {‘min\_samples’:value, ‘metric’:value, ‘max\_eps’:value }) |
| HDBSCAN | plot\_clusters (data, HDBSCAN, (), {‘min\_cluster\_size’:value, ‘metric’:value }) |
| Agglomerative Clustering | plot\_clusters (data, AgglomerativeClustering, (), {‘n\_clusters’:value, ‘metric’:value, ‘linkage’:value}) |
| Spectral Clustering | plot\_clusters (data, SpectralClustering, (), {‘n\_clusters’:value, ‘affinity’:value }) |

TASK 2: Choosing each algorithm's parameter values

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm Name** | **Args or kwd used** | **Args or kwd values** | **Justification** |
| K-Means | n\_clusters | 4 | Observable cluster are 4 and with Elbow Method we can find best cluster value |
| init | k-means++ | k-means++ will be default - and centroids are chosen with a probability proportional to the square of the distance from the point to the nearest existing centroid, where as in Random it will random centroids - slower convergence speed. [1] [2] |
| max\_iter | Default: 300 | Although it is showing the same result for the 100 value as well, but it is safe to use the default value.[3] |
| algorithm | Default: Lloyd | Elkan is faster in processing than Lloyd as it uses a lower & upper-bound approach. Lloyd spends a lot of processing time computing the distances between each of the k cluster centres and the n data points. Since points usually stay in the same clusters after a few iterations, much of this work is unnecessary, making the naive implementation very inefficient. [4] |
| Elkan |

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm Name** | **Args or kwd used** | **Args or kwd values** | **Justification** |
| Affinity Propagation | preference | -50 | The preference is suggested to be the median of the similarity matrix. However, a median preference value may not produce the optimal clustering result. Decreasing too much is giving us or increasing Hence, we can estimate the preference that generates the optimal solution. [5] |
| damping | Range [0.5-1) | [6] |
| Max\_iter | Default: 200 | 30 |
|  | affinity | Euclidean | No effect |

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm Name** | **Args or kwd used** | **Args or kwd values** | **Justification** |
| Agglomerative Clustering | N\_clusters | 4 | We can see 4 distinct clusters, so using the value 4. |
| metric | Euclidean | If linkage is a ward, then the metric to be used is Euclidean only [7] |
| linkage | Complete, ward | Ward is better as it causes minimum increase in information loss. Moreover, complete merges the close groups of clusters (as it determines the most significant distance between any two objects in the different clusters) [8] |

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm Name** | **Args or kwd used** | **Args or kwd values** | **Justification** |
| Spectral Clustering | n\_clusters | 4 | We can see 4 distinct clusters, so using the value 4. |
| affinity | random |  |
| k-means++ |  |
| gamma | Default: 300 |  |
|  | 100 |  |
| algorithm | Default: Lloyd |  |
| elkan |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm Name** | **Args or kwd used** | **Args or kwd values** | **Justification** |
| DBSCAN | eps | 4 | We can see 4 distinct clusters, so using the value 4. |
| Min\_samples | random |  |
| k-means++ |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm Name** | **Args or kwd used** | **Args or kwd values** | **Justification** |
| HDBSCAN | min\_cluster\_size | 8 | Set it to the smallest size grouping that you wish to consider a cluster i.e. 4. Increasing the min\_cluster\_size to 8 reduces the number of clusters, merging some together. This is a result of HDBSCAN\* reoptimizing which flat clustering provides greater stability under a slightly different notion of what constitutes a cluster.[9] |
| Min\_samples | 6 | min\_samples does is provide a measure of how conservative you want you clustering to be.[9] |
|  | cluster\_selection\_epsilon | 0.2 | 0.2 units ensures that clusters below the given threshold are not split up any further[9] |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm Name** | **Args or kwd used** | **Args or kwd values** | **Justification** |
| OPTICS | min\_cluster\_size | 0.019 | Minimum number of samples in an OPTICS cluster – Obtained by |
| Min\_samples | 0.1 |  |
| k-means++ |  |
|  | Xi |  |  |

**References:**

[1] [Matthew Mayo](https://www.kdnuggets.com/author/matt-mayo), 2022.Centroid Initialization Methods for k-means Clustering. KDnuggets

[2] [Nitish Kumar Thakur](https://medium.com/@nitishkthakur?source=post_page-----d5ddd8b0350e--------------------------------), 2020. k-Means Clustering: Comparison of Initialization strategies

[3] https://holypython.com/k-means/k-means-optimization-parameters/

[4] Alejandra Ornelas Barajas. 2015. K-Means clustering accelerated algorithms using the triangle inequality. COMP 5703 Advanced Algorithms

[5] Yang, K.C., Yu, C.H. and Wang, J.S., 2012. Robust affinity propagation using preference estimation.  
[6] Frey B J, Dueck D. Clustering by Passing Messages Between Data Points. Science, 2007, 315(5814), 972-976. <http://www.psi.toronto.edu/affinitypropagation>

[7] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html#sklearn.cluster.AgglomerativeClustering

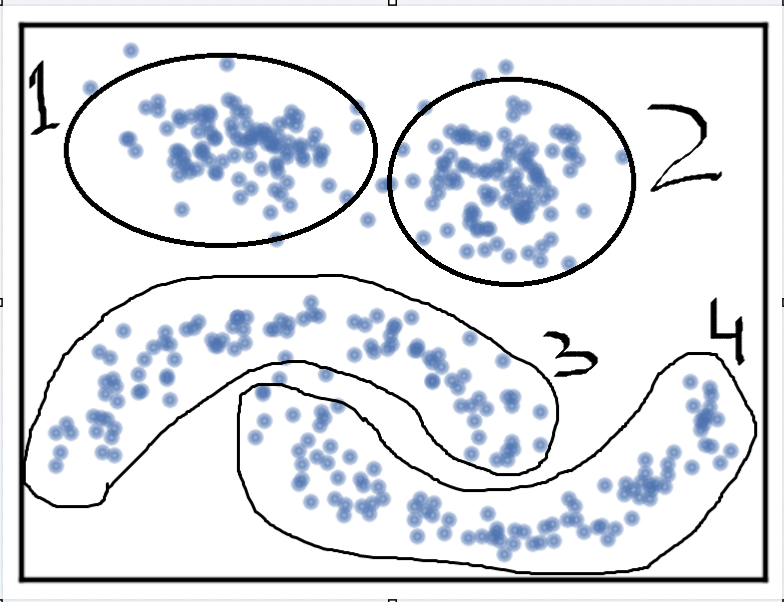
[8] [Adji Achmad Rinaldo Fernandes](https://www.frontiersin.org/people/u/1363712), Solimun Solimun. 2022. Comparison of the Use of Linkage in Cluster Integration With Path Analysis Approach

[9] Campello, Moulavi, and Sander. Density-Based Clustering Based on Hierarchical Density Estimates https://hdbscan.readthedocs.io/en/latest/parameter\_selection.html

1. The given domain

Business Domain

* We have a non-globular cluster with some noise/outliers.
* non-globular clusters might represent diverse purchasing behaviors or patterns that cannot be easily characterized by a simple, spherical model.
* For example - We can connect to the non-globular clusters with a supply-chain customer domain specific

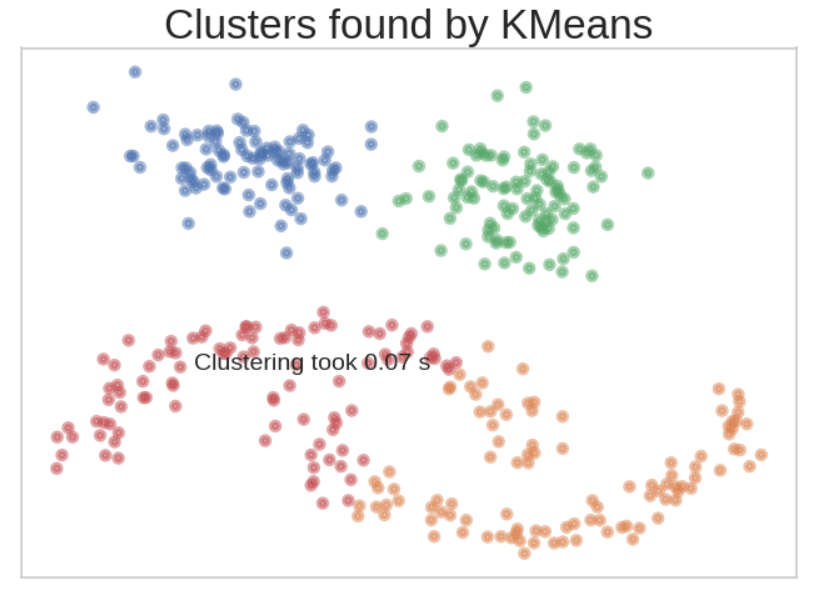


1. No-Buying Customers unsatisfied customers.
2. Regular Buying customers satisfied customers.
3. Cluster where non-Buying habit changes from un-satisfied to satisfied.
4. Long/Short term buyers – Online, Store at 1 time more shopping, Season, Holidays
5. Noises/Outliers – Behavioral Pattern could not be identified.

Technical Domain –

Looking at the Problem statement – DBSCAN, Hierarchal & spectral clustering could be useful for the identification of non-globular patterns & noise.

1. Specified algorithms.
2. K-Means - The number of cluster are chosen are 4 with the validating same value using elbow method. Instead of using random we are using k-means++ in which centroids are chosen with a probability proportional to the square of the distance from the point to the nearest existing centroid to increase convergence speed.



1. Intra-Cluster & Inter Cluster Distance

|  |  |
| --- | --- |
| Algorithm | Intra-Cluster & Inter Cluster Distance |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |