
Cluster
Analysis

Cluster Analysis

- In data science, we often think about how to use data to make predictions on new data points. This is called "supervised learning."

- Sometimes, however, rather than 'making predictions', we instead want to categorize data into buckets. This is termed "unsupervised learning."

https://towardsdatascience.com/an-introduction-to-clustering-algorithms-in-python-123438574097

Cluster Analysis

- To illustrate the difference, let's say we're at a major pizza chain and we've been tasked with creating a feature in the order management software that will predict delivery times for customers.
- In order to achieve this, we are given a dataset that has delivery times, distances traveled, day of week, time of day, staff on hand, and volume of sales for several deliveries in the past.
- From this data, we can make predictions on future delivery times. This is a good example of supervised learning.

https://towardsdatascience.com/an-introduction-to-clustering-algorithms-in-python-123438574097

Cluster Analysis

- Now, let's say the pizza chain wants to send out targeted coupons to customers.
- It wants to segment its customers into 4 groups: large families, small families, singles, and college students. We are given prior ordering data (e.g. size of order, price, frequency, etc) and we're tasked with putting each customer into one of the four buckets.
- This would be an example of "unsupervised learning" since we're not making predictions; we're merely categorizing the customers into groups.

https://towardsdatascience.com/an-introduction-to-clustering-algorithms-in-python-123438574097

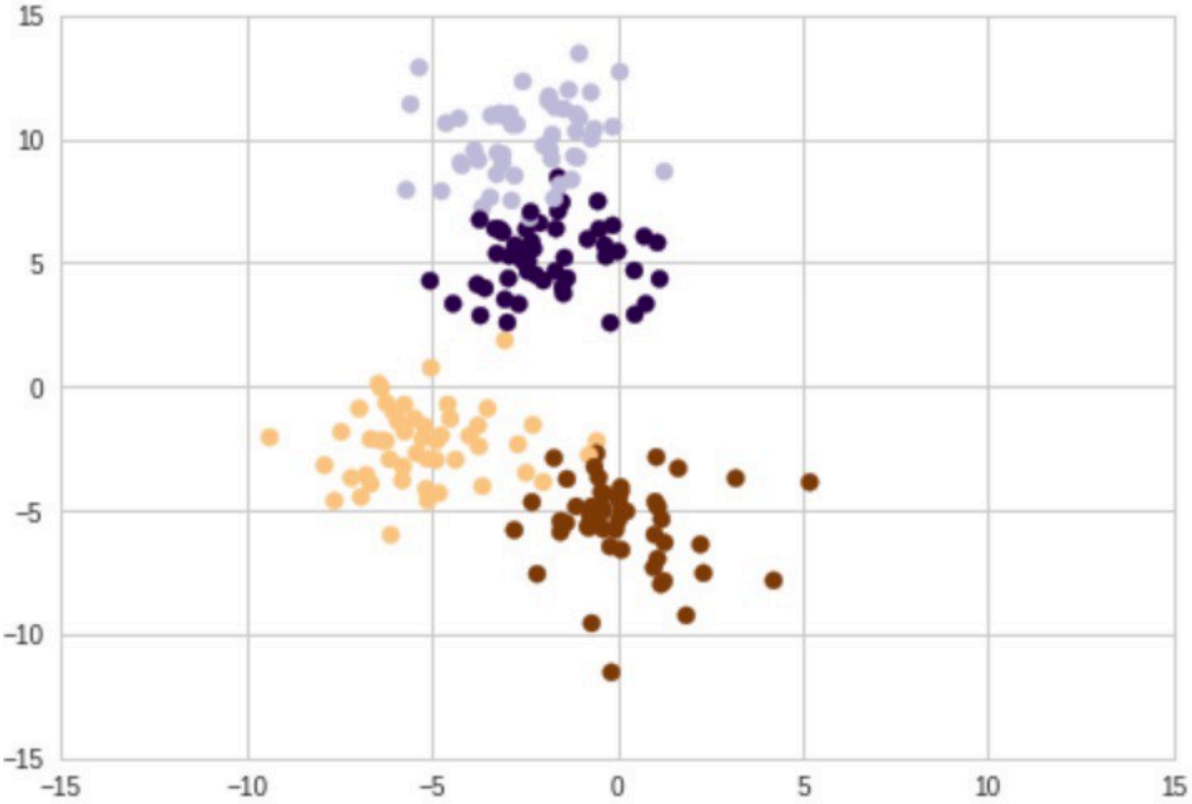
Cluster Analysis

- Clustering is one of the most frequently utilized forms of unsupervised learning.
- We explore two of the most common forms of clustering:
  - k-means and
  - Hierarchical clustering

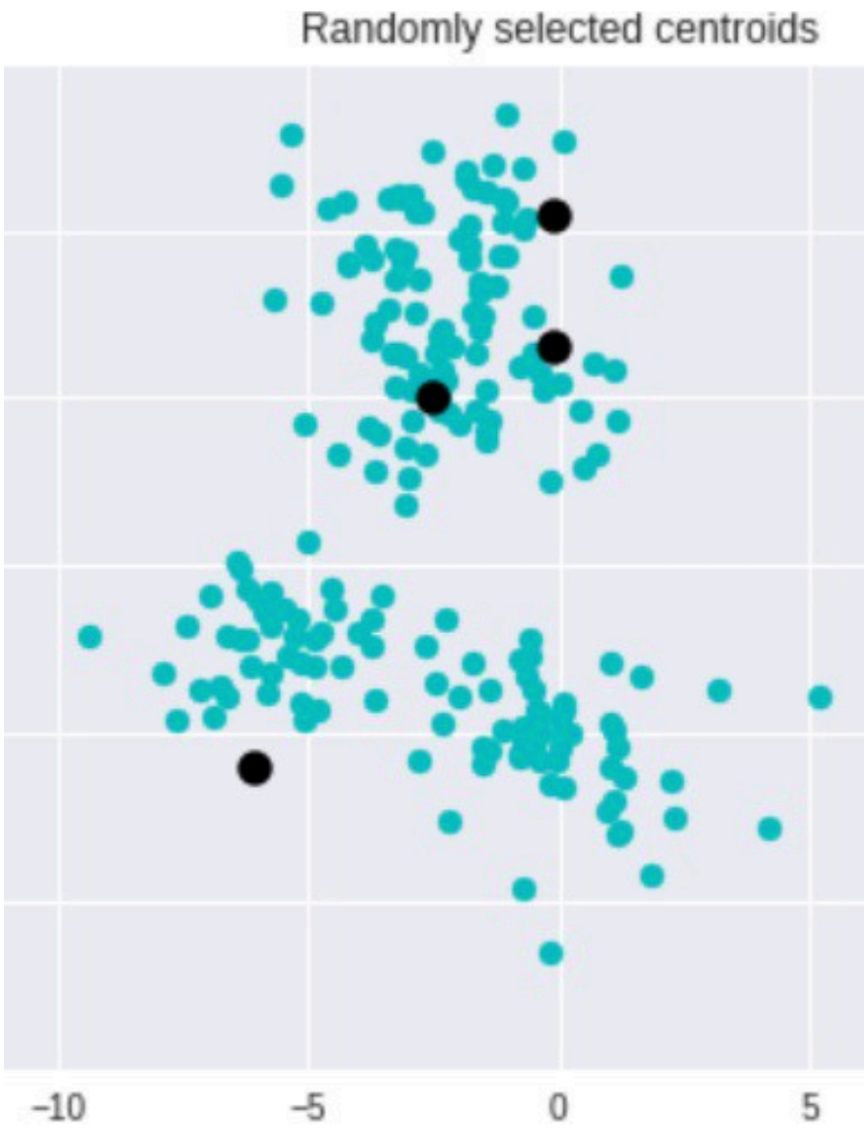
https://towardsdatascience.com/an-introduction-to-clustering-algorithms-in-python-123438574097

K-Means Clustering

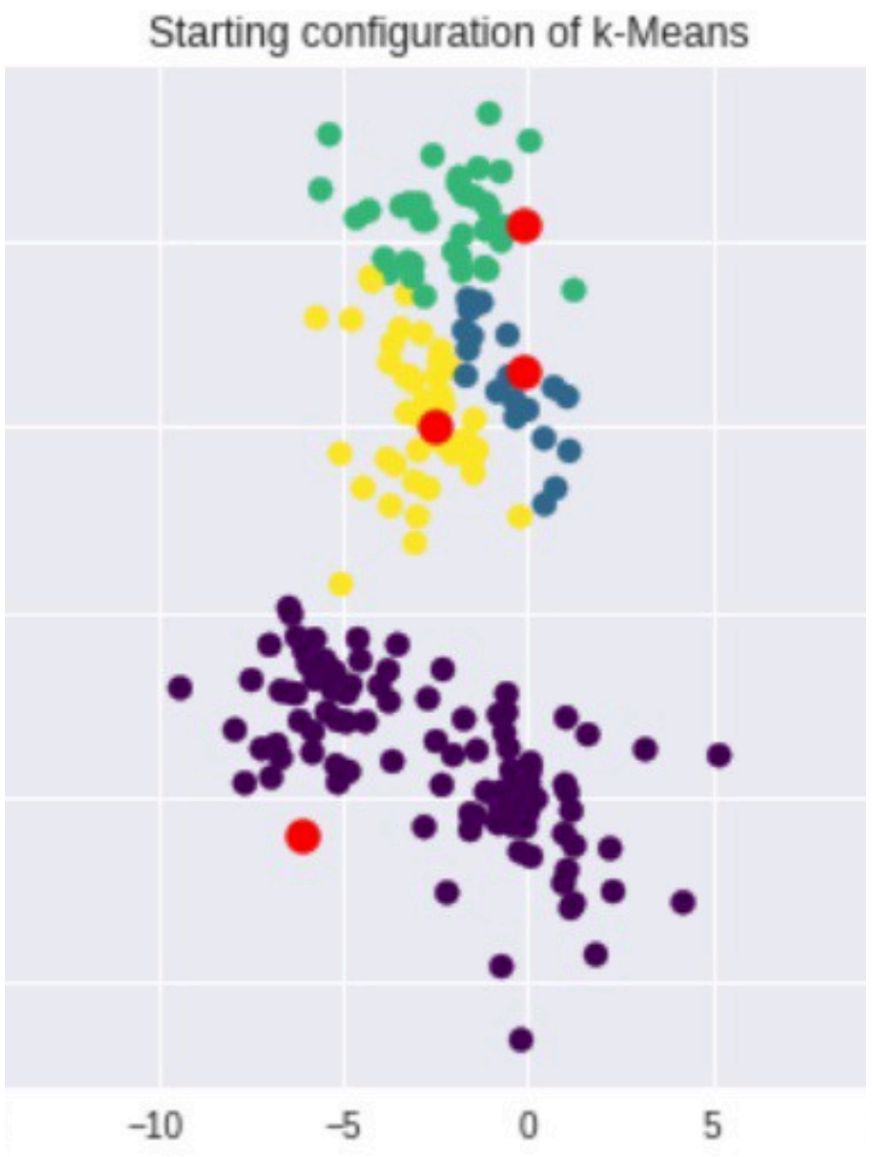
Initial Dataset



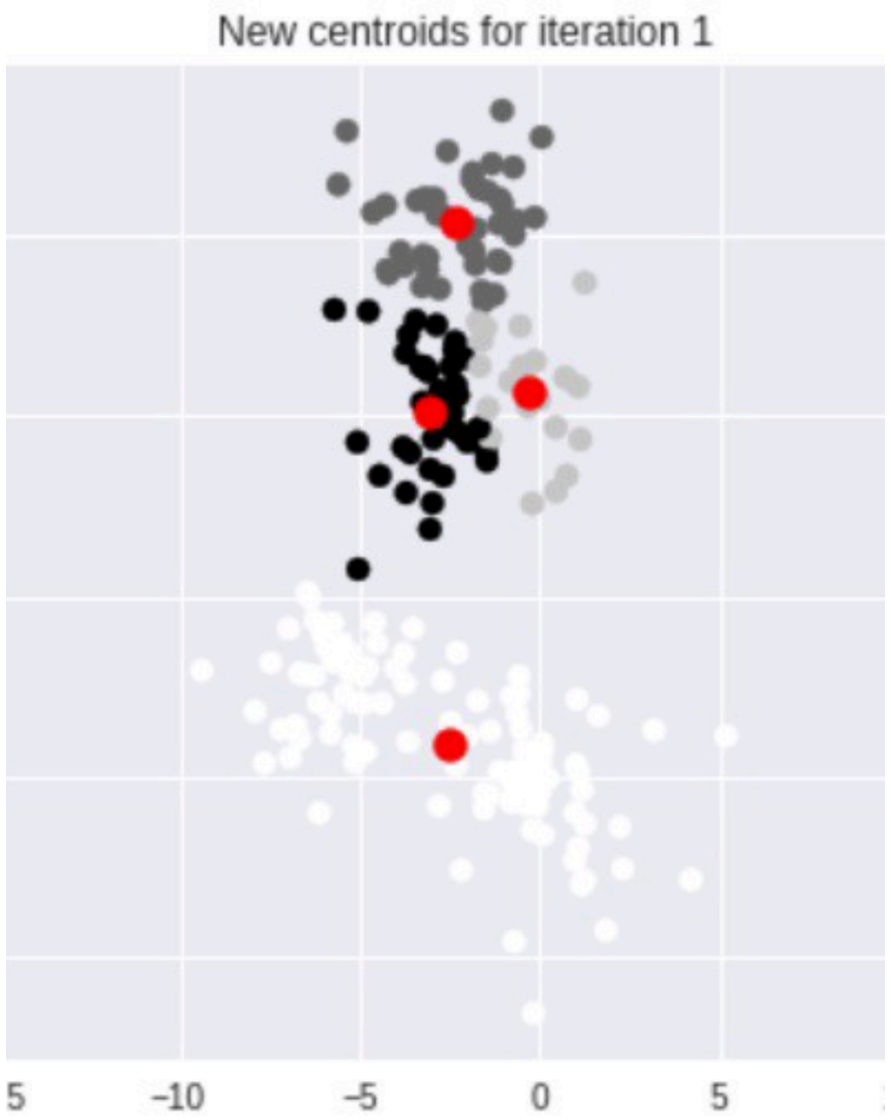
K-Means Clustering



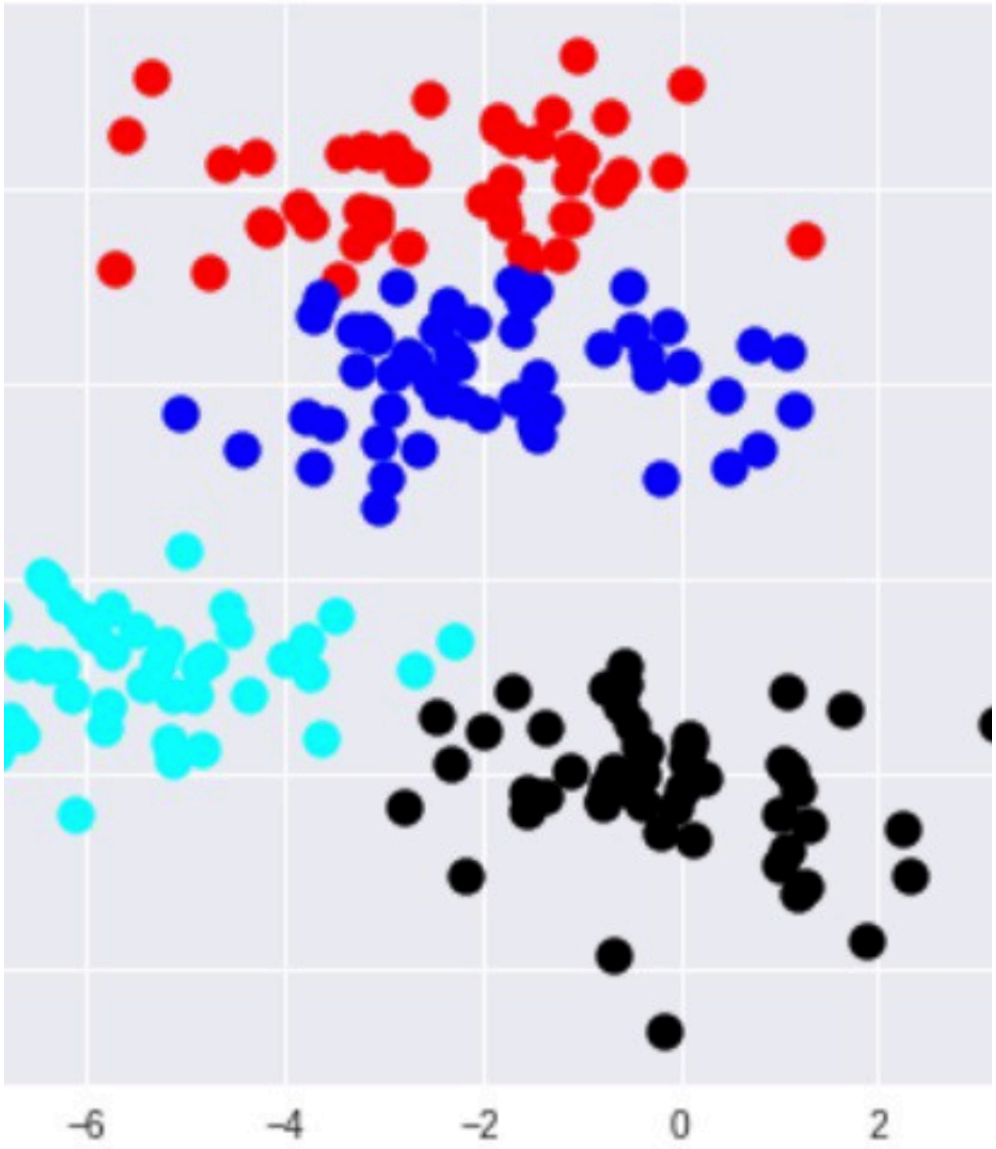
K-Means Clustering



K-Means Clustering



K-Means Clustering



Elbow Algorithm

Choose some values of k and run the clustering algorithm

For each cluster, compute the within-cluster sum-of-squares between the centroid and each data point.

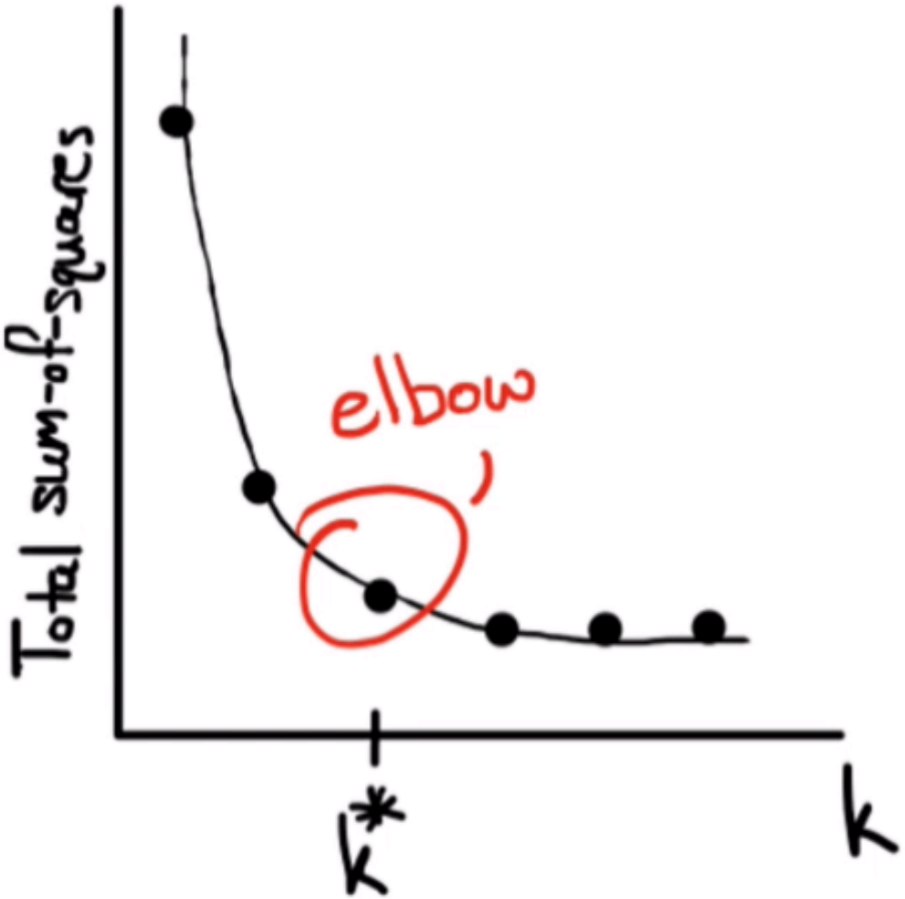
Sum up for all clusters, plot on a graph

Repeat for different values of k, keep plotting on the graph.

Then pick the elbow of the graph.

Home many Cluster to Start with?

- One way is to plot the data points and try different values to see what works the best. Another technique is called the **elbow** method.



Pros and Cons of K-Means

Hands-on Exercise with K-Means in Python

- Download Python script k-means.py

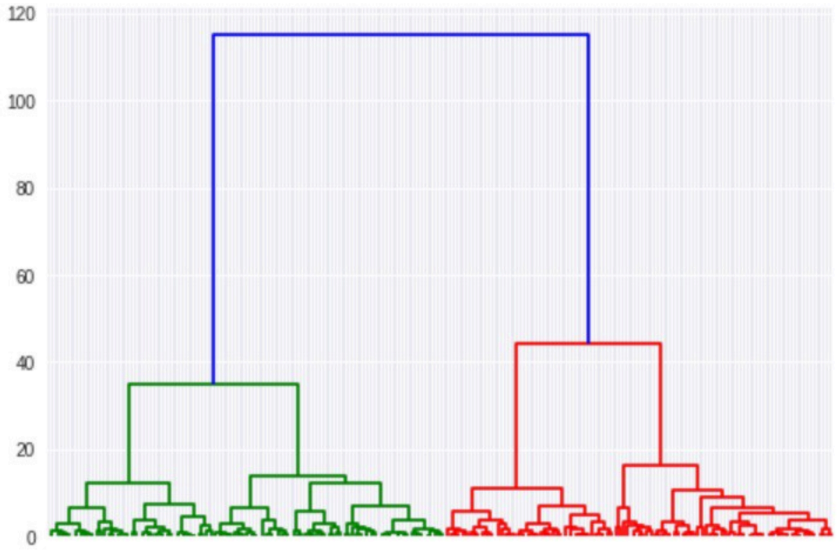
https://towardsdatascience.com/an-introduction-to-clustering-algorithms-in-python-123438574097

Agglomerative Hierarchical Clustering

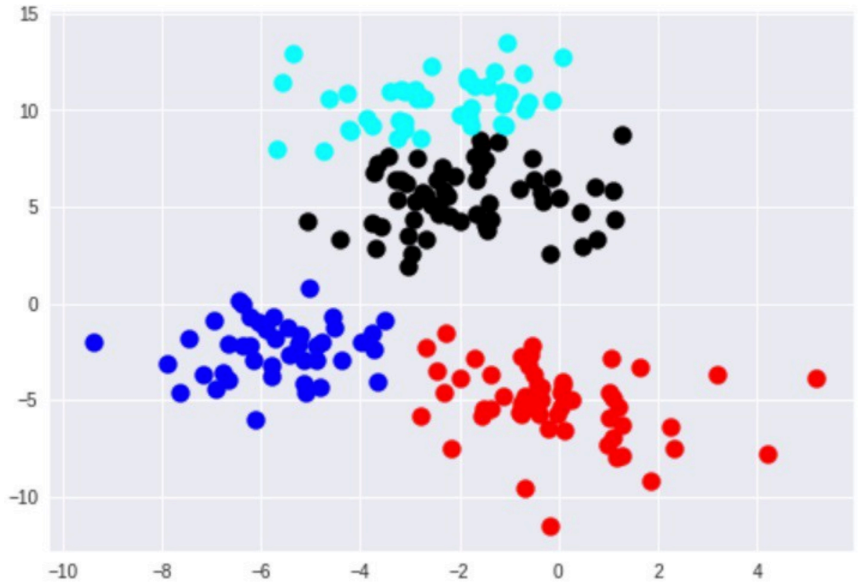
- Agglomerative hierarchical clustering differs from k-means in a key-way.
- Rather than choosing a number of clusters and starting out with random centroids, we instead begin with every point in our dataset as a "cluster."
- Then we find the two closest points and combine them into a cluster.
- Then, we find the next closest points, and those become a cluster.
- We repeat the process until we only have one big giant cluster.
- Along the way, we create what's called a dendrogram. This is our "history." You can see the dendrogram for our data points below to get a sense of what's happening.

https://towardsdatascience.com/an-introduction-to-clustering-algorithms-in-python-123438574097

Dendogram



Clusters



Hands-on Exercises with Python Cluster Analysis

A list of 10 of the more popular algorithms is as follows:

1. Affinity Propagation
2. Agglomerative Clustering
3. BIRCH
4. DBSCAN
5. K-Means
6. Mini-Batch K-Means
7. Mean Shift
8. OPTICS
9. Spectral Clustering
10. Mixture of Gaussians

https://machinelearningmastery.com/clustering-algorithms-with-python/

Hands-on Exercise in Python

- Download Python script clustering-algorithms.py

https://towardsdatascience.com/an-introduction-to-clustering-algorithms-in-python-123438574097