Customer Segmentation Using Clustering Algorithms

Pratik Bhangale, Computer Science, MIT ADT University, School of Engineering, pratiks.bhangale@gmail.com

Abstract - Clustering is an unsupervised problem of finding natural groups in the feature space of input data. In order to solve this problem, we can use a variety of algorithms. This report represents 10 of the most widely used clustering algorithms for customer segmentation data provided in an excel sheet.

Key Words – Clustering Algorithms, Skewness, Normalisation, Data Visualisation, SKLearn, Unsupervised Learning.

1. Introduction

Unsupervised learning potentially provides enormous potential to companies by allowing them the ability to cluster or group data based on similar characteristics or features. The algorithms that exist in unsupervised learning can be used based on the different scenario's, datasets, and computational limitations. The 10 algorithms used in this project are,

- Affinity Propagation
- Agglomerative Clustering
- BIRCH
- DBSCAN
- K-Means
- Mini-Batch K-Means
- Mean Shift
- OPTICS
- Spectral Clustering
- Mixture of Gaussians

Each algorithm offers a different approach to the challenge of discovering natural groups in data. There is no easy way to determine the best algorithm for your use case, but I have tried to determine the best algorithm for this use-case.

2. Engineering Methodology

This project is a study based on the results obtained by implementing the 10 algorithms given in the Introduction. The Python libraries used in this project aside from the ones for Clustering algorithms are as listed below,

- Pandas
- Numpy
- Matplotlib.pyplot
- Seaborn
- SKLearn
- StandardScaler
- Datetime

All these libraries are either pre-installed on the Pycharm platform and if not can be manually download through "pip install package_name" command on the python command prompt.

Implementing a Clustering Algorithm usually consists of the following few steps,

- ➢ Data Extraction – The data is generally stored in excel or csv files. In this step, we extract the data from these files to later manipulate it according to our needs.
- ➢ Data Manipulation – In this step, we sort and split the data into train and test sets. We also add a few parameters we may require later in the process.
- ➢ Skewness Reduction – We reduce the distortion or asymmetry in the data.

- Normalisation – For the algorithm to work at its best, the data needs to be as close to the range 0-1. This is what we aim to achieve in this step.
- Mean and Variance – We check if the mean and variance values are -1 and 0 respectively. If not, then we get them to the required values in this step.
- Model creation and fitting – The model is created using the required clustering algorithm, and the data we just processed is then fitted to the data.
- Visualising the obtained Clusters – We use libraries like matplotlib to visualise the results we have obtained after implementing the clustering algorithm.

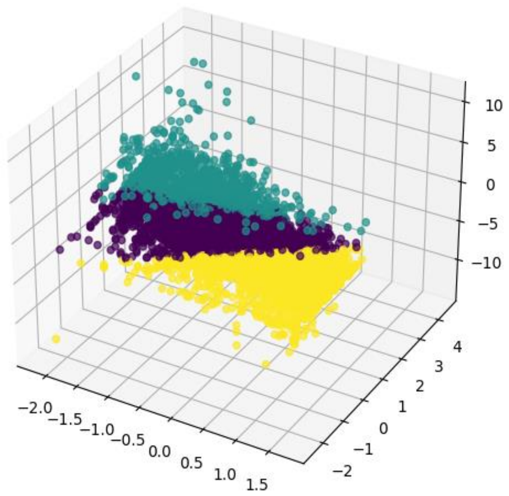
My project also follows the same methodology mentioned above. We have first extracted the data from the excel file, then sampled it, reduced it's skewness, normalised it, created a model, and we fit it with the processed data, and visualised it using a python library.

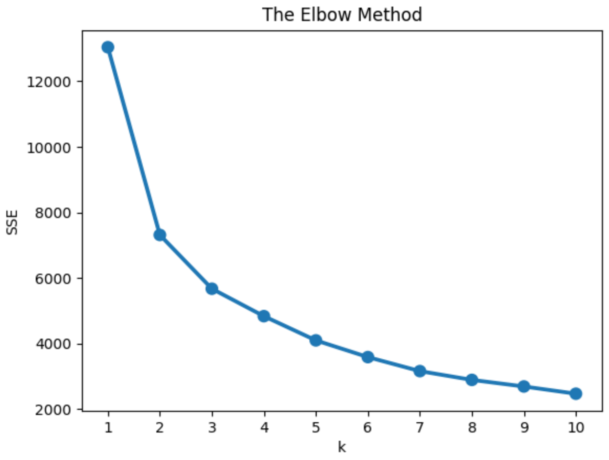
3. Analysis

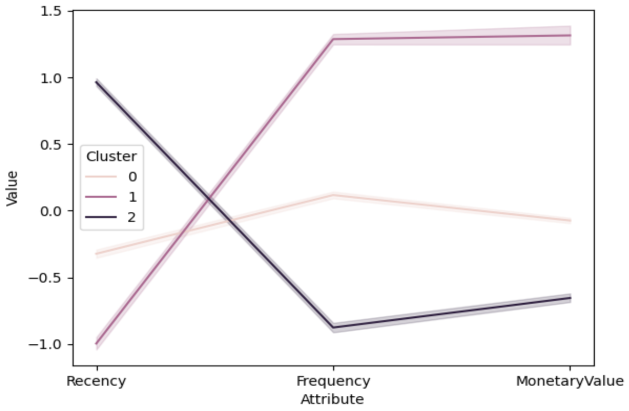
The dataset that we have used is consisting of the buying data of people. This dataset has about 3 lakh records of purchases of customers, and we have sorted customers into generally 3 different groups using clustering algorithms as follows.

- Kmeans Clustering Algorithm

The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm. We calculated k using the elbow method for this and all other algorithm that require the number of clusters to be defined.





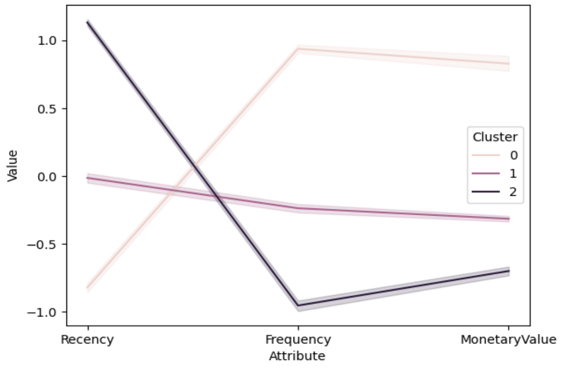


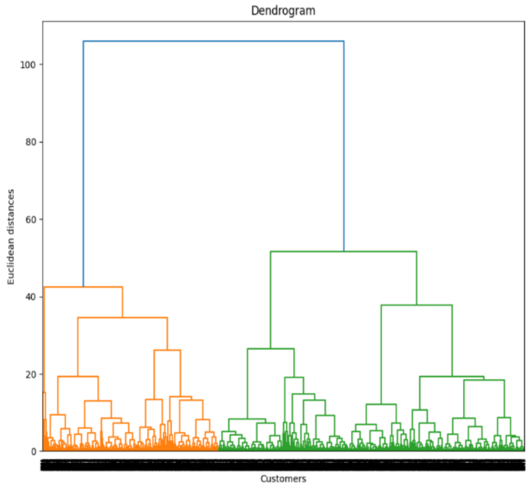
We observe that the slope is linear after the value of k is 3, hence we shall take the optimum number of clusters to be 3.

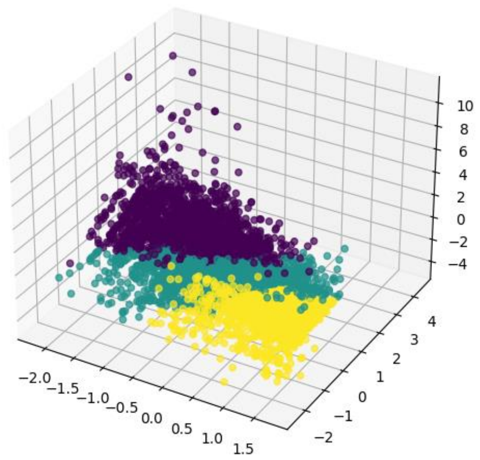
Kmeans algorithm is according to our observations the algorithm that had the best performance out of all the algorithms, based on the visualisation of the data.

○ Agglomerative Clustering Algorithm

An agglomerative clustering algorithm is a type of hierarchical clustering algorithm where each individual element to be clustered is in its own cluster. These clusters are merged iteratively until all the elements belong to one cluster. The number of clusters are determined by plotting a dendrogram and then selecting the number where the largest distance is covered by the lines together.



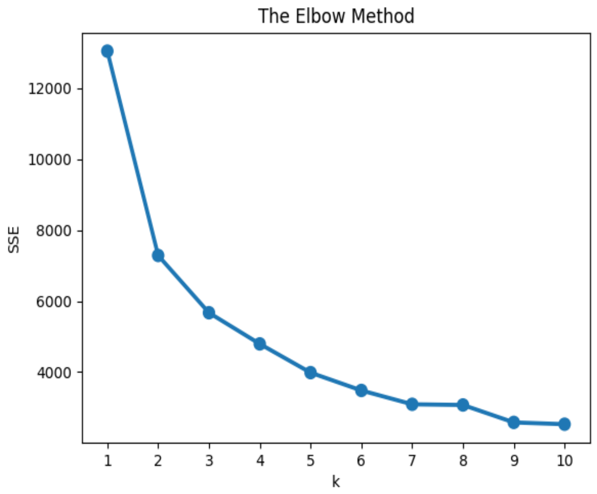


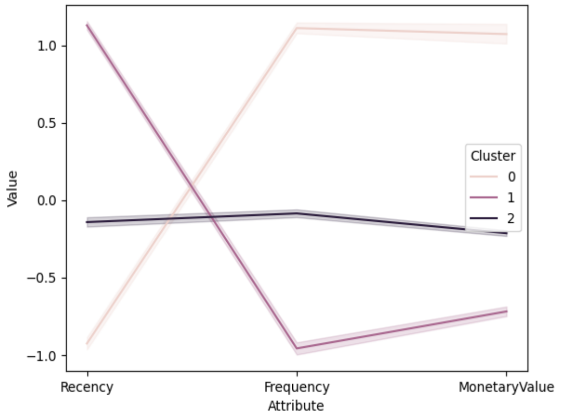


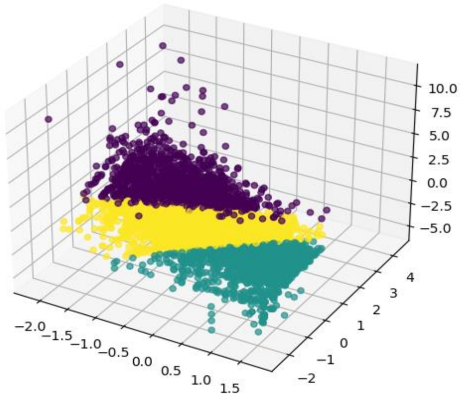
As we can see, the results are almost on par with the Kmeans clustering algorithm, minus a few exceptions.

○ Mini-Batch Kmeans Clustering Algorithm

Mini Batch K-means algorithm's main idea is to use small random batches of data of a fixed size, so they can be stored in memory. Each iteration a new random sample from the dataset is obtained and used to update the clusters and this is repeated until convergence. This algorithm gained popularity because it does not require the entire dataset to be stored in the memory at the same time.



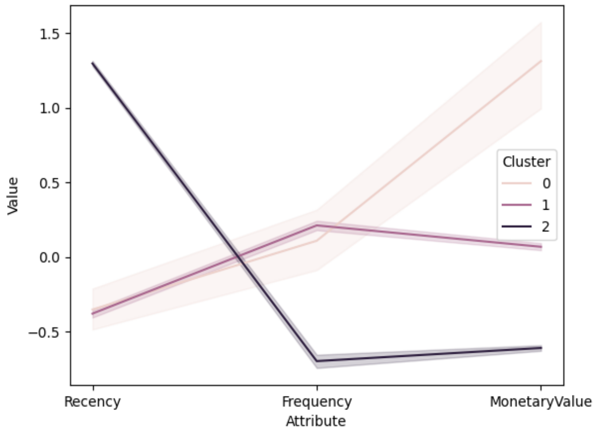


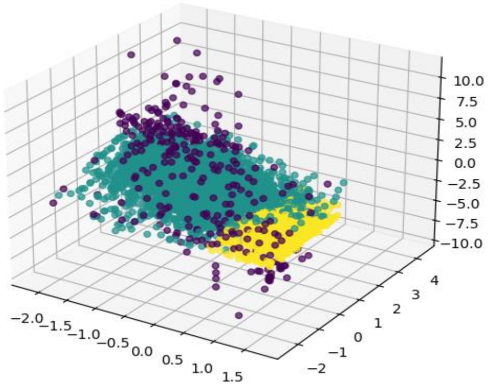


We observe that the results are more or less the same as regular kmeans algorithm, while using less resources as compared to it. The number of clusters are again decided using the elbow method.

- Gausian Mixture Clustering Algorithm

The Gausian Mixture Algorithm works much the same way as Kmeans, bar 2 major differences. Kmeans does not account for the varience, wheareas Gausian Mixture does, and Kmeans performs Soft Classification, in contrast to the Hard Classification performed by Gausian Mixture.





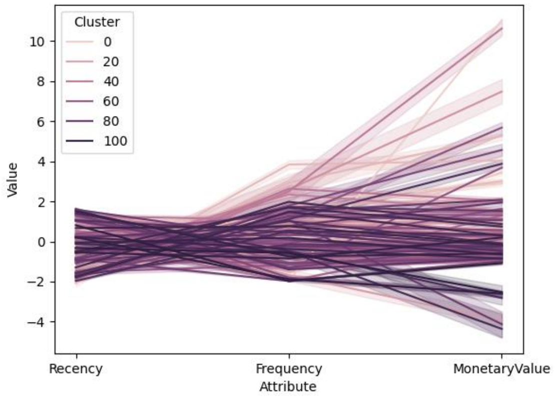
We observe that the clusters are not as well defined or as well spaced out as the ones that we observed before, but we can still make out 3 separate clusters have been formed.

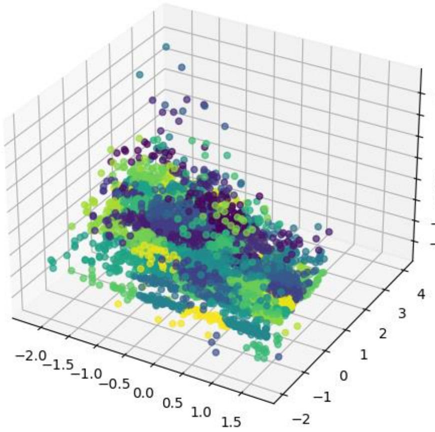
- Affinity Propogation

Affinity Propagation creates clusters by sending messages between data points until convergence. Unlike clustering algorithms such as k-means or k-medoids, affinity propagation does not require the number of clusters to be determined or estimated before running the algorithm.

A dataset is described using a small number of exemplars, ‘exemplars’ are members of the input set that are representative of clusters. The messages sent between pairs represent the suitability for one sample to be the exemplar of the other, which is updated in response to the values from other pairs.

This happens iteratively till the clusters converge.

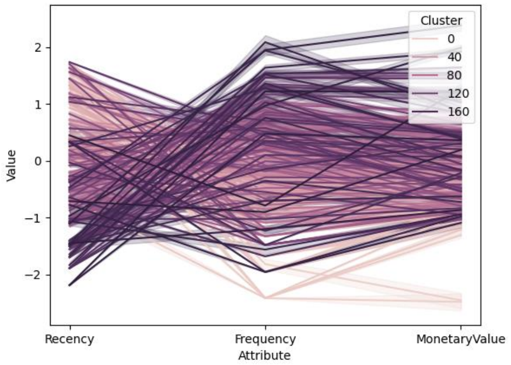


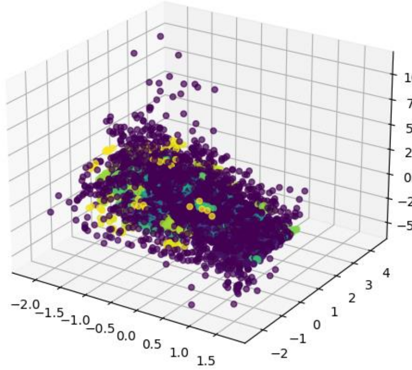


We can clearly observe that clustering has not occurred proprerly, and the data is all over the place. This algorithm is not reccomended for this usecase.

○ Optics Clustering Algorithm

This clustering technique is different from other clustering techniques in the sense that this technique does not explicitly segment the data into clusters. Instead, it produces a visualization of Reachability distances and uses this visualization to cluster the data.

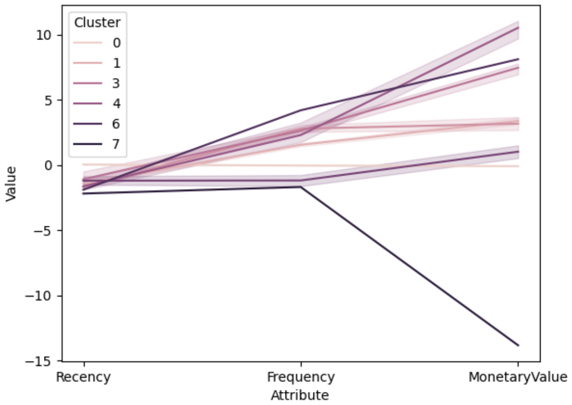


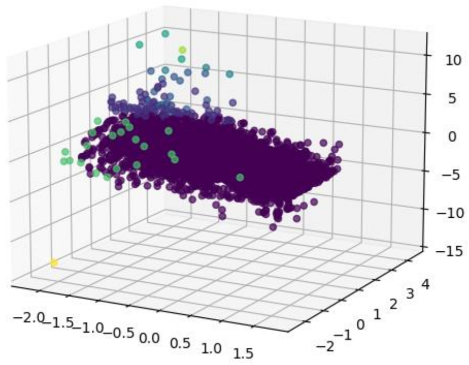


As we observe, the algorithm fails to properly cluster the data, and an unexpected output is achived.

○ MeanShift Clustering Algorithm

Given a set of data points, the algorithm iteratively assigns each data point towards the closest cluster centroid and direction to the closest cluster centroid is determined by where most of the points nearby are at. So each iteration each data point will move closer to where the most points are at, which is or will lead to the cluster center. When the algorithm stops, each point is assigned to a cluster.

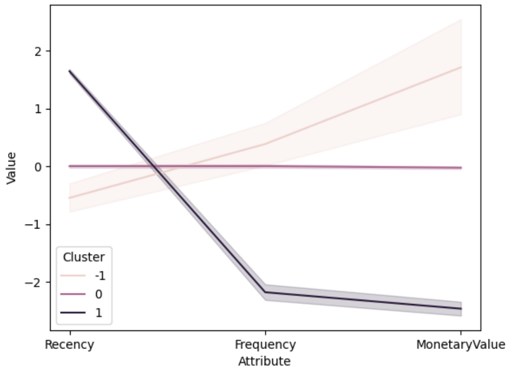


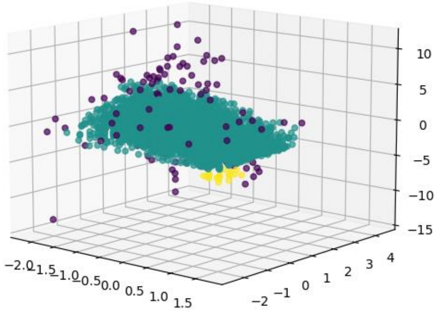


As we can observe, one big cluster has formed, and the rest, which is a small part is scattered away from this cluster. This algorithm will also not suffice for this use case.

○ DB Scan Clustering Algorithm

The DBSCAN algorithm is based on this intuitive notion of “clusters” and “noise”. The key idea is that for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points. This algorithm has a benefit that it is able to perform will even in the presence of noise, over kmeans and agglomerative clustering that perform poorly when noise is introduced.

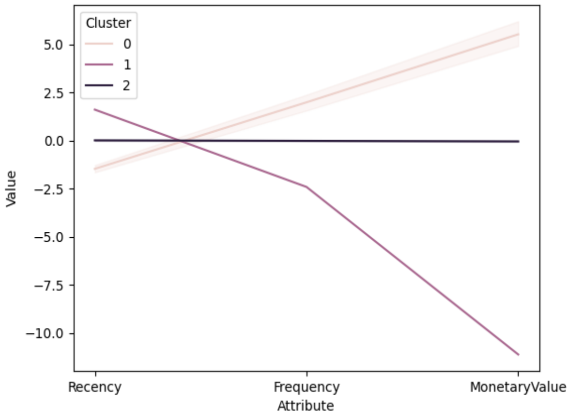


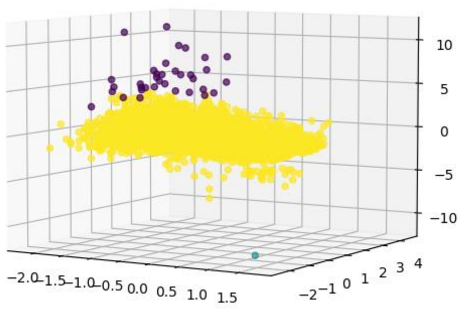


We observe again that a central large cluster appears, and a few points are scattered across. This algorithm is not suitable fo rthis usecase.

○ Birch Clustering Algorithm

Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) is a clustering algorithm that can cluster large datasets by first generating a small and compact summary of the the large dataset that retains as much information as possible. This smaller summary is then clustered instead of clustering the larger dataset.

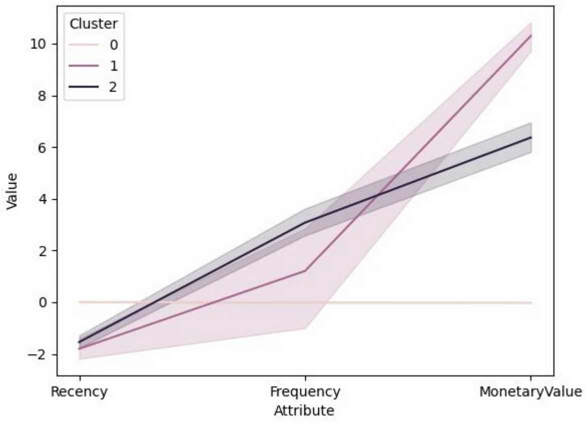


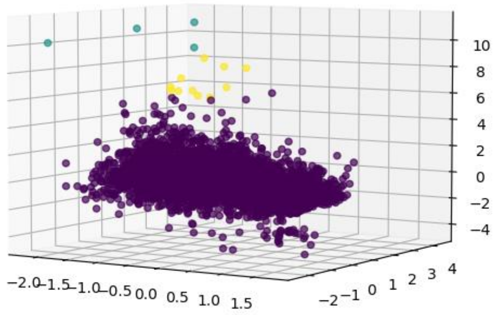


In a similar fashion to the recent few algorithms, a large central loab forms, and a few bits and bobs are scattered across. This algorithm is not suitable for this application.

○ Spectral Clustering Algorithm

It treats each data point as a graph-node and thus transforms the clustering problem into a graph-partitioning problem. It invloves 3 steps, namely, Building the Similarity Graph, Projecting the data onto a lower Dimensional Space, and Clustering the Data.





We observe again that a central loab appears, and a few points are scattered across. This algorithm is not suitable for this usecase.

4. Conclusion

In this paper, we studied the use and implementation of 10 of the most popular clustering algorithms. We can conclude that, for this particular use-case, Kmeans, Mini-Batch Kmeans, and Agglomerative clustering algorithms are the most well suited, based on their clustering visualisations.

5. Future Work

- Implementing more new Algorithms

- Optimising existing Algorithms
- Creating new Algorithms

6. Acknowledgement

I would like to thank Mr.Kaustubh Lambe for providing insights and their expertise in this field, which greatly assisted me in the completion of this project.

7. References

- Data Clustering: Algorithms and Applications – Charu C Agarwal
- https://www.geeksforgeeks.org
- https://www.tutorialspoint.com/
- https://stackoverflow.com/