Clustering for E-commerce data using unsupervised ML techniques

Radwa Magdy Khairy Ammar Mohamed
Faculty of graduate studies of statistical research , Cairo University

***Abstract*—E-commerce system has become more popular and implemented in almost all business areas. E-commerce system is a platform for marketing and promoting the products to customer through online. Customer segmentation is known as a process of dividing the customers into groups which shares similar characteristics. The purpose of customer segmentation is to determine how to deal with customers in each category in order to increase the profit of each customer to the business. Segmenting the customers assist business to identify their profitable customer to satisfy their needs by optimizing the services and products. Therefore, customer segmentation helps E-commerce system to promote the right product to the right customer with the intention to increase profits. There are few types of customer segmentation factors which are demographic psychographic, behavioral. In this study, customer behavioral factor has been focused. Therefore, users will be analyzed using clustering algorithm in determining the purchase behavior of E-commerce system. The aim of clustering is to optimize the experimental similarity within the cluster and to maximize the dissimilarity in between clusters. In this research, the proposed approach analyzed the groups that share similar criteria to help vendors to identify and focus on the high profitable segment to the least profitable segment. This type of analysis can play important role in improving the business. Grouping their customer according to their similar behavioral factor to sustain their customer for long-term and increase their business profit. It also enables high exposure of the e-offer to gain attention of potential customers. In order to process the collected data and segment the customers, I used unsupervised machine learning techniques such as K-Means , hierarchical clustering , density-based spatial clustering of application with noise "DBSCAN" and gaussian mixture model "GMM". Depending on multiple clustering methods can assist in finding the best performed clustering technique.**

1. Introduction

E-commerce system is a platform for marketing and promoting the products to customer through online [1]. Customer segmentation is known as dividing the customers into groups which shares similar characteristics. The purpose of customer segmentation is to determine how to deal with customers in each category to increase the profit of each customer to the business. When customers receive too much information or unwanted details which is not related to their regular purchase or their interest on the products, it can cause confusion on deciding their needs. This might lead their customers to give up on purchasing the items they required and effect the business to lose their potential customers. The clustering analysis will help to categorize the E-commerce customer according to their spending habit, purchase habit or specific product or brand the customers interested in. In order to process the collected data and segment the customers, unsupervised machine learning techniques such as K-Means , hierarchical clustering , density-based spatial clustering of application with noise "DBSCAN" and gaussian mixture model "GMM". [2]. Online shopping is not anymore, a new, whereas most of the business are becoming online based. There are a number of online shopping platforms keep on increasing day by day. Since most traditional business started to implement E-commerce system in their business and E-commerce system has become trending, there are more competition in the field [3]. In order for a business to sustain for a longer term and be competitive, the business should know the ways to retain their customers. For an example, if an E-commerce system continuously display a customer the products that is expensive or above their shopping budget, then the customers may decide that this E-commerce system is not suitable for them. Therefore, they may look for another online shopping platforms which usually leads to high churn in E commerce platform. Customers differ in personality and have various preferences. There is evidence that inequalities in marketing exists. As a result, having the same approach and marketing for every consumer is not effective [4] and those consumers may be the most vital to the business [5]. Therefore, it can be stated that inequalities in marketing and inefficiencies may cause customer churn. It is critical for a company to segment its consumers and determine the distinctions between the customer segments. Market segmenting according to the customer purchase behaviour is important to decide the likelihood of the customer buying a specific product [6]. In this study it explains how a business can run for longer term by understanding their customer need and interest and satisfy them. The aim of this research is to conduct customer segmentation using the customer data and grouping their customers into groups that share similar criteria. Customer segmentation is carried out to find the potential and most profitable customer groups among the total customers [7]. Therefore, this helps to reduce the risk of losing the customer by selling the wrong product to the wrong customer group. Customer segmentation shows the way for E-commerce on how to make their business customer-focused and conquer a stable position in the business world. Rachmawati et al. [8] proclaim that for the large and sophisticated data information of today's E-commerce businesses, accurate and efficient customer segmentation management should be carried out. In this competitive and developing E-commerce business, it is important to analyze the customer need and apply market segmentation mine and analyse various target customers in the system to provide different customers with distinctive marketing methods and improve their customer loyalty and satisfaction. According to Shirole et al. [9], customer segmentation is based on discovering important differentiators that split customers into target groups. A customer segmentation model allows organizations to target specific groups of customers, allowing for more effective marketing resource allocation and the maximization of cross- and up-selling capability. Customer segmentation can also help to enhance customer service and increase customer loyalty and retention. There are several aspects of online shopping behavior can be found that can influence the strategic approach in E-commerce for longer term which are the security of seller and buyer E-commerce. The aim of this study is to conduct an effective segmentation of E-commerce customer. The customer segmentation proposed for this research is applying unsupervised machine learning techniques such as K-Means , hierarchical clustering , density-based spatial clustering of application with noise "DBSCAN" and gaussian mixture model "GMM". All these mentioned algorithms belong to clustering techniques. Clustering is one of the most frequently used forms of unsupervised learning. It automatically discovers natural grouping in data. Clustering is especially useful for exploring data you know nothing about. You might find connections you never would have thought of. Clustering can also be useful as a type of feature engineering, where existing and new examples can be mapped and labeled as belonging to one of the identified clusters in the data. The objectives and contributions of this research are: • Various data analysis techniques were shown to understand the dataset characteristics of customer purchase behavior.

• To analyze the connection between customer purchase behavior and the customer segmentation. The purpose of this study is to analyze the connection between the purchase behavior of E-commerce platform customer and the customer segmentation.

• To comprehend how emerging technologies enable marketers to better meet the requirements and desires of consumers. This project also comprehends how the emerging technologies enable marketers to better meet the requirements and desires of consumers where technology has the ability to influence and change customer behavior.

• To understand the consumer behavior and focus on high profitable segment other objective of this project is to obtain understanding on the consumer behavior and focus on high profitable segment. Customer segmentation objective is to enhance sales by providing customized experience tailored to each segment.

• Proposal of multiple ML clustering techniques for customer segmentation using UCI's E-commerce customer purchase behavior dataset.

• Data visualization and Dashboards for visualizing the results of customer segmentation to the end users.

2. Related Work

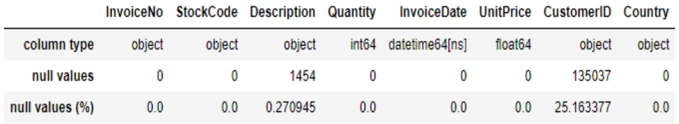
Previous approaches for comparing the performance of clustering algorithms can be divided according to the nature of used datasets. While some studies use either real-world or artificial data, others employ both types of datasets to compare the performance of several clustering methods. A comparative analysis using real world dataset is presented in several works [[10], [11], [12] , [13], [14], [15]]. More specifically, clustering algorithms are evaluated in terms of a combination of clustering measurements. Their results show that no algorithm can achieve the best performance on all measurements for any dataset and, for this reason, it is mandatory to use more than one performance measure to evaluate clustering algorithms. he following algorithms were compared: k-means, random swap, expectation-maximization, hierarchical clustering, self-organized maps (SOM) and fuzzy c-means. The authors found that the most important factor for the success of the algorithms is the model order, which represents the number of centroid or Gaussian components (for Gaussian models-based approaches) considered. Overall, the recognition accuracy was similar for clustering algorithms focused in minimizing a distance based objective function. The methods were evaluated in terms of both scalability and accuracy. In the former, the running time of both algorithms were compared for different number of instances and features. In addition, the authors assessed the ability of the methods in finding adequate sub-spaces for each cluster. Another common approach for comparing clustering algorithms considers using a mixture of real world and artificial data (e.g. [[16] , [17], [18], [19]]). In [18], the performance of k-means, single linkage and simulated annealing (SA) was evaluated, considering different partitions obtained by validation indexes. The authors proposed a new validation index called I index that measures the separation based on the maximum distance between clusters and compactness based on the sum of distances between objects and their respective centroids. They found that such an index was the most reliable among other considered indices, reaching its maximum value when the number of clusters is properly chosen.

3. Data

Typically, e-commerce datasets are proprietary and consequently to find among publicly available data. However, The UCI Machine Learning Repository has made this dataset containing actual transactions from 2010 and 2011. The dataset is maintained on their site, where it can be found by the title "Online Retail". "This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers."

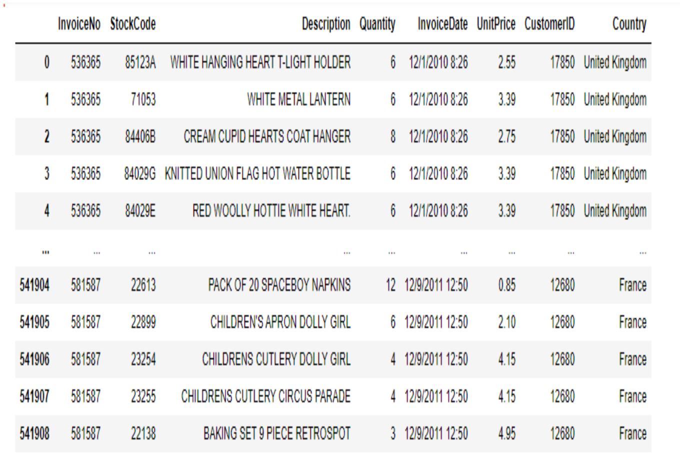
Dataset Columns:

Invoice No: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter ’c’, it indicates a cancellation.
Stock Code: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
Description: Product (item) name. Nominal.
Quantity: The quantities of each product (item) per transaction. Numeric.
Invoice Date: Invoice Date and time. Numeric, the day and time when each transaction was generated. Unit Price: Unit price. Numeric, Product price per unit in sterling.
Customer ID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
Country: Country name. Nominal, the name of the country where each customer resides.



	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
column type	object	object	object	int64	datetime64[ns]	float64	object	object
null values	0	0	1454	0	0	0	135037	0
null values (%)	0.0	0.0	0.270945	0.0	0.0	0.0	25.163377	0.0

Figure 1. Data Types with Null values included in dataset’s columns



	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850	United Kingdom
...	...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	12/9/2011 12:50	0.85	12680	France
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	12/9/2011 12:50	2.10	12680	France
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	12/9/2011 12:50	4.15	12680	France
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	12/9/2011 12:50	4.15	12680	France
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	12/9/2011 12:50	4.95	12680	France

Figure 2. Head of dataset

4. Problem Formulation

The aim of this study is to use multiple clustering methods in customer’s segmentation by dividing the customers into segments based on their behavior that can help E-commerce systems to understand the customer’s needs in each category to retain customers for longer term and achieve high profits. In this section, The main focus will be on demonstration all the stages have done till achieving the clustering model. Therefore, the first stage is to start to understand the business needs, working on business issues, finding the business criticality in E-commerce domain. Understanding the importance for clustering method that will add to business and the benefits of using it. Also, knowing what can clustering technique achieve in order to contribute in solving business problems, helping in covering any gaps in business that needs to work on it. The second stage is to work on exploring data and data analysis by checking all the columns of the dataset, checking data types, finding the duplicated values, finding the missing values, checking the distribution for numerical data, finding if there is any skewed data in any numeric data, detecting anomalies , checking imbalanced data and this is helping to extract insights from data while working on data anaylsis. After business understanding, data analysis, then the the third stage comes to clean your data, so if there are duplicated values , then it needs to be removed. if there are missing values, it needs to be handled either by removing the missing values if it’s un-needed data or not consuming high percentage of the data size or imputing missing values by choosing the suitable imputing technique for handling missing values and in this study, I decided to remove the missing values because they are un-needed data and not consuming high percentage from the dataset.Also, if there are some anomalies detected, then it needs to be removed but in this study, I kept the outliers because it refers to real online transactions , so it’s part of the main data that can’t be neglected.last thing in this stage is to filter the high variation data , so in this dataset, we have high variation such as customer ID, invoice number, stock code , description of the purchased products , so I decided to use them in analysis section as they are really much important in finding out insights , however I removed them before applying data modelling as this can cause poor performance if they are added to the ML algorithm.The forth stage after data cleaning is to visualise your data to find out insights and based on understanding the data types and the content of the data, the needed visualized graph can be selected using two libraries in python [Matplotlib and Seaborn]. The fifth stage is data preprocessing or data transformation or data preparing, so in this step I used feature engineering by creating new features such as from invoice date feature, I’ve created some new features such as day, month, year, weekday, day name, hours, minutes quarter, season, periods of day. Also, from both Quantity and Unit price , I’ve created total price feature and from total price, I’ve created two new features ”Amount” which refers to total amount for transactions for each customer and ”Frequency” which refers to the total number of transactions for each customer. Then, finding out the string data which needs to be encoded, Then working on normalization or standardizing the scale of data. There are multiple techniques can be applied for scaling the data such as standard scaling, minmax scaling , robust scaling. I used standard scaling to normalize the data. Also, I used principle component analysis for dimension reduction. All these steps needs to be done successfully before going to data modelling stage. The sixth stage is data modelling, All is needed is to select the right algorithms in machine learning or deep learning based on the previous stages , understanding the data and the target of the study.

I decided to use clustering methods in machine learning. I chose multiple clustering models such as K-Means , hierarchical clustering , density-based spatial clustering of application with noise “DBSCAN” and Gaussian mixture model “GMM” and compared the results in order to find out the best performed clustering technique.The seventh stage is to evaluate each algorithm, test the behavior of the model and getting the results and the last stage when everything is done and the model can achieve high results and reaching to the target. then deployment needs to be done over web site or application tool so that the model can be used.

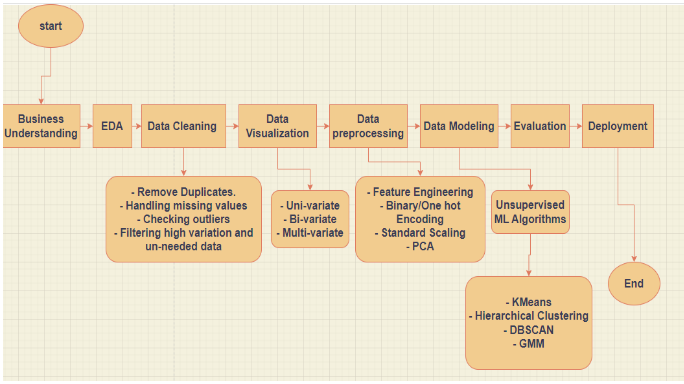


Figure 3. Problem Stages

5. Model

There are many clustering algorithms to choose. It is a good idea to explore a range of clustering algorithms and different configurations. It might take some time to figure out which type of clustering algorithm works the best for the given data, but when you do, you’ll get invaluable insight on your data.

5.1. Centroid-based:

These types of algorithms separate data points based on multiple centroids in the data. Each data point is assigned to a cluster based on its squared distance from the centroid.This is the most commonly used type of clustering. K-Means algorithm is one of the centroid based clustering algorithms. Here k is the number of clusters and is a hyperparameter to the algorithm.

K-Means Clustering may be the most widely known clustering algorithm and involves assigning examples to clusters in an effort to minimize the variance within each cluster. It’s a centroid-based algorithm and the simplest unsupervised learning algorithm. The algorithm tries to minimize the variance of data points within a cluster. It’s also how most people are introduced to unsupervised machine learning. K-means++ (default init parameter for K-Means in sklearn) is the algorithm which is used to overcome the drawback posed by the k-means algorithm. The goal is to spread out the initial centroid by assigning the first centroid randomly then selecting the rest of the centroids based on the maximum squared distance. The idea is to push the centroids as far as possible from one another.

Although the initialization in K-means++ is computationally more expensive than the standard K-means algorithm, the run-time for convergence to optimum is drastically reduced for K-means++. This is because the centroids that are initially chosen are likely to lie in different clusters already.

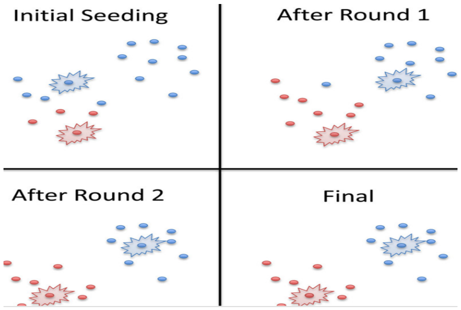


Figure 4. K-Means Technique

5.2. Hierarchical-based (Connectivity-based):

The idea is based on the core idea of objects being more related to nearby objects than to objects farther away.It builds a tree of clusters so everything is organized from the top-down. Initially each data point is considered as an individual cluster. At each iteration, the similar clusters merge with other clusters until one cluster or K clusters are formed.Hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis that seeks to build a hierarchy of clusters. It’s used to group objects in clusters based on how similar they are to each other. In general, the merges and splits are determined in a greedy manner. The results of hierarchical clustering are usually presented in a dendrogram.This technique is specific to the agglomerative hierarchical method of clustering. The method starts by considering each point as a separate cluster and starts joining points to clusters in a hierarchical fashion based on their distances. To get the optimal number of clusters for hierarchical clustering, we make use a dendrogram which is tree-like chart that shows the sequences of merges or splits of clusters. Hierarchical clustering is particularly useful in situations where you have a few observations you are particularly interested in and you want to be able to identify observations that are similar to those observations.

Hierarchical clustering can be:

- Agglomerative: it starts with an individual element and then groups them into single clusters.

- Divisive: it starts with a complete dataset and divides it into partitions.

Agglomerative clustering is best at finding small clusters. The end result looks like a dendrogram so that you can easily visualize the clusters when the algorithm finishes.

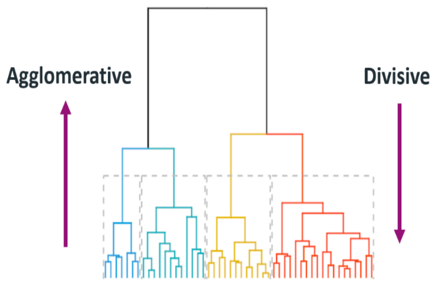


Figure 5. Hierarchical clustering

5.3. Density-based:

Data is grouped by areas of high concentrations of data points surrounded by areas of low concentrations of data points. Basically the algorithm finds the places that are dense with data points and calls those clusters.The clusters can be any shape. You aren't constrained to expected conditions. The clustering algorithms under this type don't try to assign outliers to clusters, so they get ignored.
DBSCAN stands for density-based spatial clustering of applications with noise. It's a density-based clustering algorithm. It is able to find irregular-shaped clusters. It separates regions by areas of low-density so it can also detect outliers really well. This algorithm is better than k-means when it comes to working with oddly shaped data.

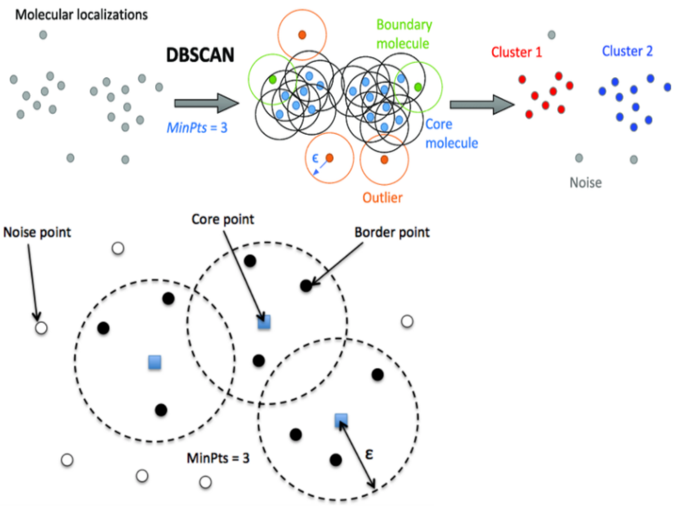


Figure 6. DBSCAN clustering

5.4. Distribution-based:

It is a clustering model in which we will fit the data on the probability that how it may belong to the same distribution.There is a center-point established. As the distance of a data point from the center increases, the probability of it being a part of that cluster decreases.This model works well on synthetic data and diversely sized clusters.The method suffers from overfitting, unless constraints are put on the model complexity. A more complex model will usually be able to explain the data better, which makes choosing the appropriate model complexity inherently difficult.
A Gaussian mixture model (GMM) attempts to find a mixture of multi-dimensional Gaussian probability distributions that best model any input dataset. In the simplest case, GMMs can be used for finding clusters in the same manner as k-means, but because GMM contains a probabilistic model under the hood, it is also possible to find probabilistic cluster assignments.The Gaussian mixture model uses multiple Gaussian distributions to fit arbitrarily shaped data.

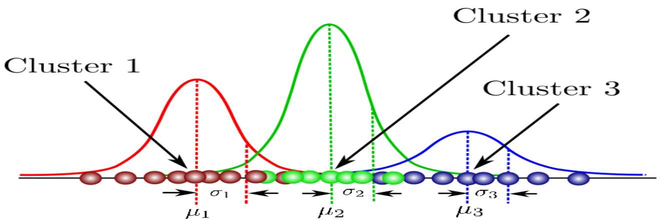


Figure 7. Gaussian Mixture Model

6. Results

The work described in this study is based on a database providing details on purchases made on an E-commerce platform over a period of one year. Each entry in the dataset describes the purchase of a product, by a particular customer and at a given date. In total, approximately 4000 clients appear in the database. I chose to work on this case by using clustering techniques in machine learning such as K-Means , hierarchical clustering , density-based spatial clustering of application with noise DBSCAN and Gaussian mixture model on sample of dataset and results are different in each model.

6.1. K-Means Algorithm:

By using elbow method to determine the best number of clustering the data after reprocessing, I able to find 7 clusters is the best number based on this metric and below scatter plot shows the data points belong to each cluster based on two features Amount and Frequency with the size of each cluster.

6.2. Hierarchical Clustering:

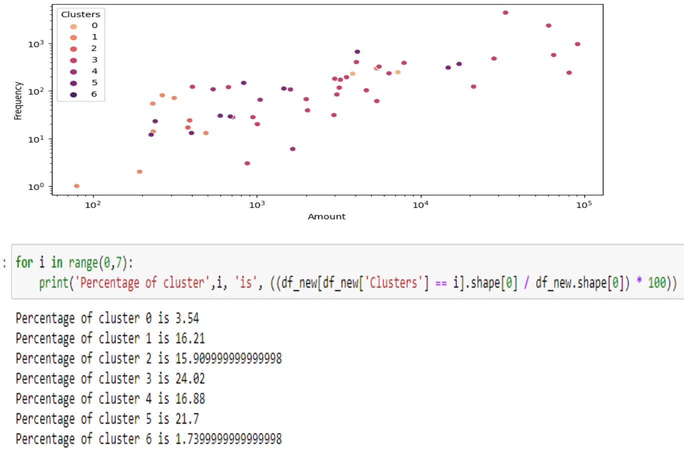
By using Silhouette metric, we are able to cluster the 10000 records to 2 clusters.

6.3. DBSCAN:

By applying the algorithm on 1000 records and given the hyper parameters, epsilon = 5 , min. samples = 4 , we found all the data belongs to one cluster.

6.4. Gaussian Mixture Model:

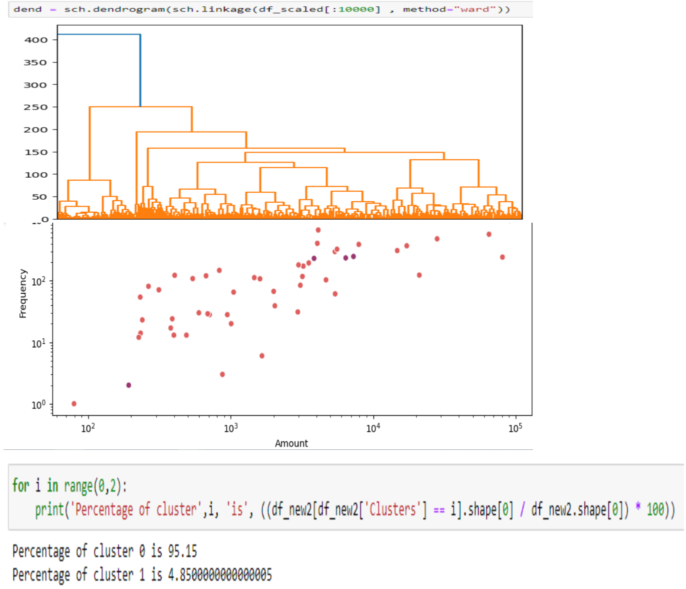
By applying the algorithm on the whole dataset, given number of clustering = 8 , we can find some overlapping between clusters which is not expected as this algorithm is famous of applying clustering using soft assignment.



```
for i in range(0,7):
    print('Percentage of cluster',i, 'is', ((df_new[df_new['Clusters'] == i].shape[0] / df_new.shape[0]) * 100))
```

```
Percentage of cluster 0 is 3.54
Percentage of cluster 1 is 16.21
Percentage of cluster 2 is 15.909999999999998
Percentage of cluster 3 is 24.02
Percentage of cluster 4 is 16.88
Percentage of cluster 5 is 21.7
Percentage of cluster 6 is 1.7399999999999998
```

Figure 8. Clustering results of using K-Means on 10000 records with the size of each cluster



```
dend = sch.dendrogram(sch.linkage(df_scaled[:10000] , method="ward"))
```

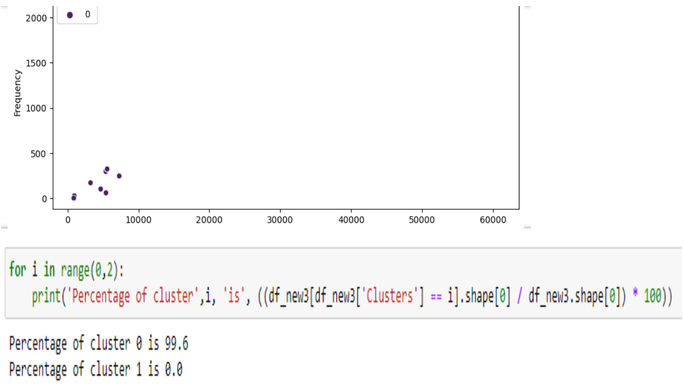
```
for i in range(0,2):
    print('Percentage of cluster',i, 'is', ((df_new2[df_new2['Clusters'] == i].shape[0] / df_new2.shape[0]) * 100))
```

```
Percentage of cluster 0 is 95.15
Percentage of cluster 1 is 4.8500000000000005
```

Figure 9. Hierarchical Clustering results on 10000 records with the size of each cluster

7. Conclusion:

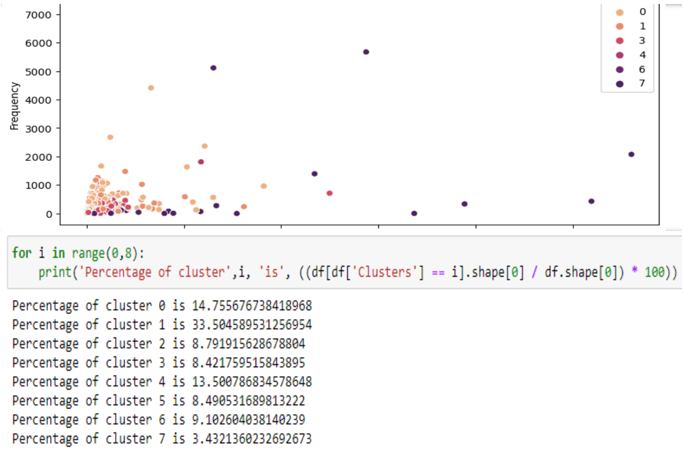
I am able to apply 4 different clustering methods in machine learning and compared the results based on each algorithm while using different metrics to determine the best number to cluster the data. It can be noticed that there is some overlapping between some of clusters and this can be improved if we used other cluster techniques in deep learning.



```
for i in range(0,2):
    print('Percentage of cluster',i, 'is', ((df_new3[df_new3['Clusters'] == i].shape[0] / df_new3.shape[0]) * 100))
```

```
Percentage of cluster 0 is 99.6
Percentage of cluster 1 is 0.0
```

Figure 10. DBSCAN Clustering results on 1000 records with the size of each cluster



```
for i in range(0,8):
    print('Percentage of cluster',i, 'is', ((df[df['Clusters'] == i].shape[0] / df.shape[0]) * 100))
```

```
Percentage of cluster 0 is 14.755676738418968
Percentage of cluster 1 is 33.504589531256954
Percentage of cluster 2 is 8.791915628678804
Percentage of cluster 3 is 8.421759515843895
Percentage of cluster 4 is 13.500786834578648
Percentage of cluster 5 is 8.490531689813222
Percentage of cluster 6 is 9.102604038140239
Percentage of cluster 7 is 3.4321360232692673
```

Figure 11. GMM Clustering results on the whole dataset with the size of each cluster

References

[1] G. I. Bhaskara and V. Filimonau, “The covid-19 pandemic and organisational learning for disaster planning and management: A perspective of tourism businesses from a destination prone to consecutive disasters,” *Journal of Hospitality and Tourism Management*, vol. 46, pp. 364–375, 2021.

[2] F. Nie, Z. Li, R. Wang, and X. Li, “An effective and efficient algorithm for k-means clustering with new formulation,” *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2022.

[3] P. Brandtner, F. Darbanian, T. Falatouri, and C. Udokwu, “Impact of covid-19 on the customer end of retail supply chains: A big data analysis of consumer satisfaction,” *Sustainability*, vol. 13, no. 3, p. 1464, 2021.

[4] D. W. Khong, “Rents: How marketing causes inequality by gerrit de geest,” *Asian Journal of Law and Policy*, vol. 1, no. 1, pp. 83–86, 2021.

[5] K. M. Manero, R. Rimiru, and C. Otieno, “Customer behaviour segmentation among mobile service providers in kenya using k-means algorithm,” *International Journal of Computer Science Issues (IJCSI)*, vol. 15, no. 5, pp. 67–76, 2018.

[6] S. Janardhanan and R. Muthalagu, “Market segmentation for profit maximization using machine learning algorithms,” in *Journal of Physics: Conference Series*, vol. 1706, p. 012160, IOP Publishing, 2020.

[7] V. Dawane, P. Waghodekar, and J. Pagare, “Rfm analysis using k-means clustering to improve revenue and customer retention,” in *Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021)*, 2021.

[8] I. K. Rachmawati, M. Bukhori, F. Nuryanti, S. P. FE, and S. Hidayatullah, “Collaboration technology acceptance model, subjective norms and personal innovations on buying interest online,” *International Journal of Innovative Science and Research Technology*, vol. 5, no. 11, 2020.

[9] R. Shirole, L. Salokhe, and S. Jadhav, “Customer segmentation using rfm model and k-means clustering,” 2021.

[10] Y. G. Jung, M. S. Kang, and J. Heo, “Clustering performance comparison using k-means and expectation maximization algorithms,” *Biotechnology & Biotechnological Equipment*, vol. 28, no. sup1, pp. S44–S48, 2014.

[11] T. Kinnunen, I. Sidoroff, M. Tuononen, and P. Fränti, “Comparison of clustering methods: A case study of text-independent speaker modeling,” *Pattern Recognition Letters*, vol. 32, no. 13, pp. 1604–1617, 2011.

[12] I. G. Costa, F. d. A. de Carvalho, and M. C. de Souto, “Comparative analysis of clustering methods for gene expression time course data,” *Genetics and Molecular Biology*, vol. 27, pp. 623–631, 2004.

[13] M. C. De Souto, I. G. Costa, D. S. De Araujo, T. B. Ludermir, and A. Schliep, “Clustering cancer gene expression data: a comparative study,” *BMC bioinformatics*, vol. 9, no. 1, pp. 1–14, 2008.

[14] G. Kou, Y. Peng, and G. Wang, “Evaluation of clustering algorithms for financial risk analysis using mcdm methods,” *Information sciences*, vol. 275, pp. 1–12, 2014.

[15] J. Erman, M. Arlitt, and A. Mahanti, “Traffic classification using clustering algorithms,” in *Proceedings of the 2006 SIGCOMM workshop on Mining network data*, pp. 281–286, 2006.

[16] H. Pirim, B. Ekşioğlu, A. D. Perkins, and Ç. Yüceer, “Clustering of high throughput gene expression data,” *Computers & operations research*, vol. 39, no. 12, pp. 3046–3061, 2012.

[17] E. R. Dougherty, J. Barrera, M. Brun, S. Kim, R. M. Cesar, Y. Chen, M. Bittner, and J. M. Trent, “Inference from clustering with application to gene-expression microarrays,” *Journal of Computational Biology*, vol. 9, no. 1, pp. 105–126, 2002.

[18] U. Maulik and S. Bandyopadhyay, “Performance evaluation of some clustering algorithms and validity indices,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 12, pp. 1650–1654, 2002.

[19] D. Verma and M. Meila, “A comparison of spectral clustering algorithms,” *University of Washington Tech Rep UWCSE030501*, vol. 1, pp. 1–18, 2003.