ANALYZING PERFORMANCE OF CLUSTERING ALGORITHMS ON A REAL RETAIL DATASET

LEDION LIÇO, LEDION.LICO@FTI.EDU.AL

INDRIT ENESI, IENESI@FTI.EDU.AL

BETIM ÇIÇO, BCICO@UMT.EDU.AL

OBJECTIVE

- Finding groups of clients with similar buying patterns in a department store, to be used for targeted marketing

IMPLEMENTATION

- Performance analysis of K-means, K-Medoids, Agglomerative Clustering and DBSCAN will be applied in real department store dataset

- The main issue of the paper is to find the most appropriate clustering method for a real department store transactions dataset.

Dataset

- The dataset includes the transactions for year 2020 in a real department store and the evaluation of the algorithms will be performed on this real dataset.
- The transactions for 8844 clients during year 2020 were collected and aggregated in 3 main dimensions.
- Purchases are divided in 3 categories that are: clothes, Accessories Cosmetics and Others.
- Data are first normalized, then algorithms are performed to evaluate their performance.

AIMS OF THE PAPER

- To Obtain the right number of clusters - Elbow algorithm is implemented.
- The influence of noise and outliers of data over algorithms is analyzed as well as the correlation between dimensions.
- Time performance of clustering algorithms is considered .

CLUSTERING ALGORITHMS

- Clustering is an unsupervised learning needed to automatically decide on the grouping structure, no classes are predefined preliminarily.
- Clusters are used for finding data features and similarities among patterns of data.
- Based on the distances between them, data in the data set are grouped into classes.
- The data inside a class forms a cluster where there is a high intra class similarity and low inter class similarity

Classification of Clustering Algorithms

Partition	Hierarchical	Density	Grid
K-Means	Agglomerative (BIRCH, CHAMALEON)	DBSCAN	STING
K-Medoids (PAM, CLARA)	Divisive	DENCLUE	CLIQUE

Organization of clustering algorithms

- According to the resulting structure, they are classified in partitioning and hierarchical clustering,
- According to the modeling framework, they are classified in deterministic clustering, probabilistic models and Bayesian models
- Partitioning algorithms assign a data set points into a predefined number of clusters using iterative processes.
- The hierarchical algorithms are divided into agglomerative and divisive

The Elbow Method

- “Elbow” method is used to find the optimal number of clusters in the dataset
- It is based on the rule that increasing number of clusters, the sum of within-cluster variance will be decreased

- Steps of Elbow algorithms are:

1. Compute clustering algorithm for different values of k.
2. For each k, calculate the total within-cluster sum of square (wss).
3. Plot the curve of wss according to the number of clusters k.
4. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

- Elbow method usually is combined with clustering algorithms like K-means or K-medoids.

Implementation of The algorithms in a real RETAIL dataset

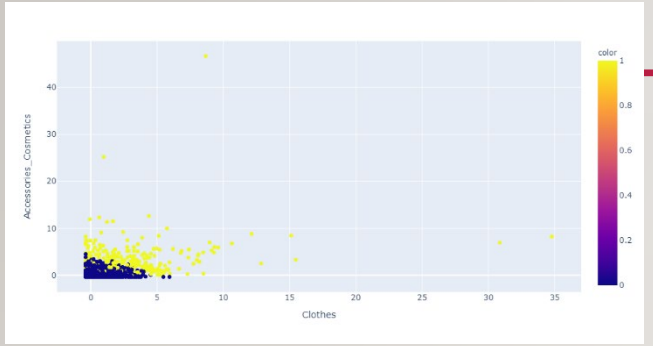
- Data Preprocessing
- *K-Means and K-Medoids*
- *Agglomerative Clustering*
- *DBSCAN*
- *Time Performance*

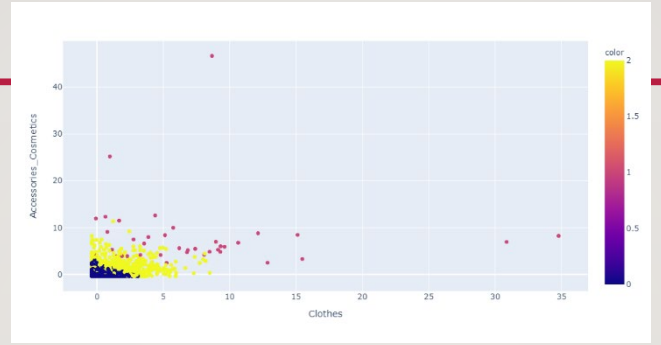
Data Preprocessing

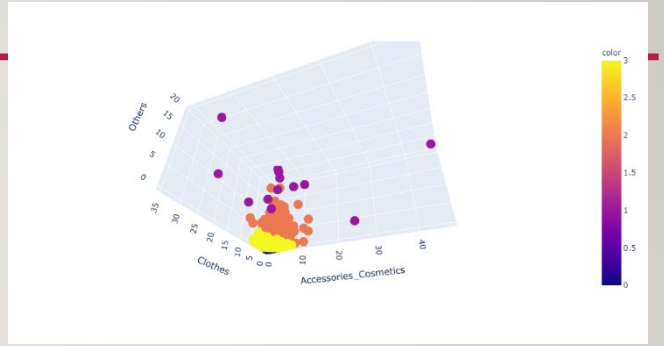
- The annual sales transactions of year 2020 for real department store were queried from the department store data warehouse.
- The transactions for 8844 clients during year 2020 were collected and aggregated in 3 main dimensions.
- Their purchases were divided in 3 categories that are: Clothes, Accessories Cosmetics and Others.
- The "Others" dimension represents purchases made in different shops that reside in the department store, use their POS-es but are not owned by the department store.
- Data will be normalized using the StandardScaler() function because usually different data components have different scales and their derivatives tend to align along directions with higher variance, reducing or slowing the convergence.
- In our case, the objective is that all the features are treated equally.

#	Column	Non-Null Count	Dtype
0	CLIENT	8844 non-null	int64
1	SEX	8844 non-null	object
2	CITY	8844 non-null	object
3	Age	8844 non-null	int64
4	Clothes	8844 non-null	float64
5	Accessories Cosmetics	8844 non-null	int64
6	Others	8844 non-null	int64

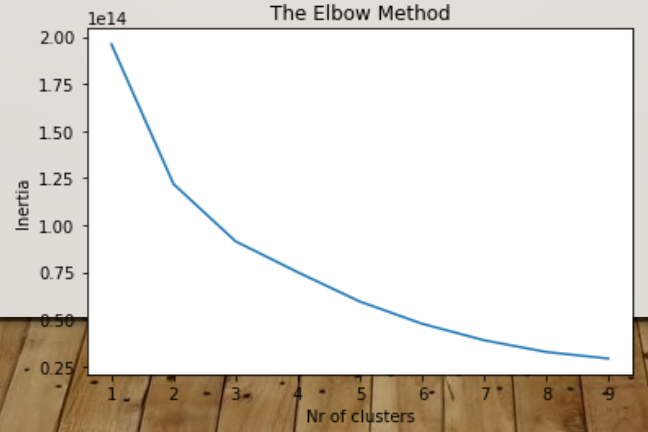
K-Means



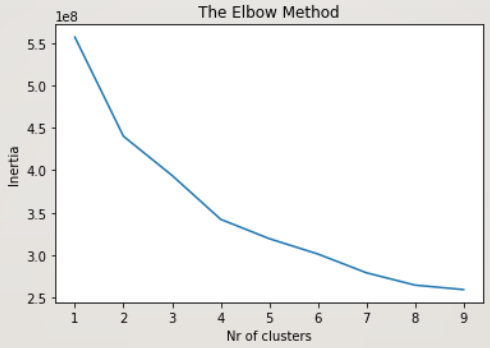


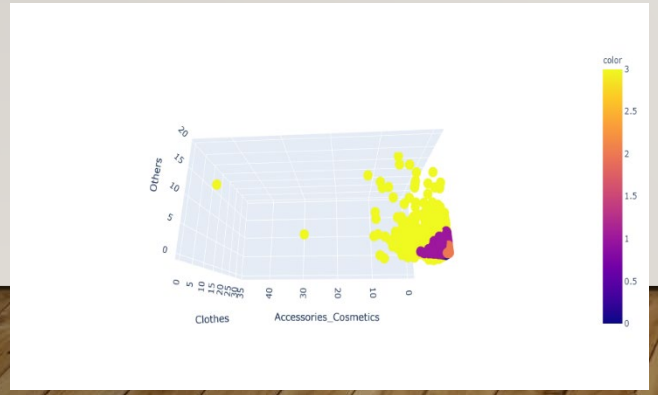


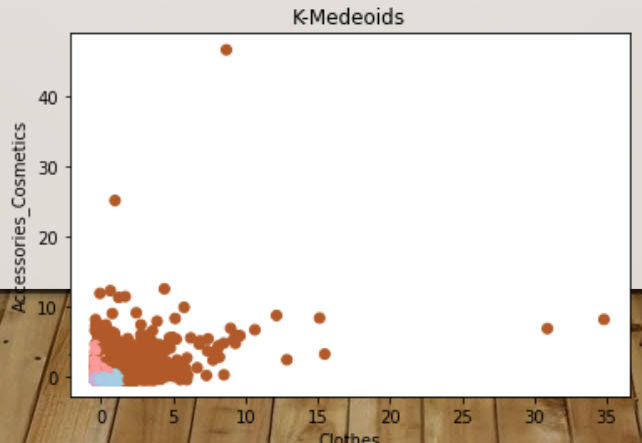
- best number of clusters, the elbow method is performed, the inertia is plotted for different number of clusters and the results are analyzed

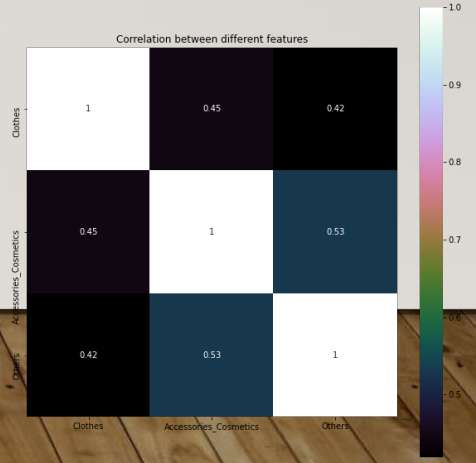


K-Medoids


The Elbow Method
1e8
Inertia
Nr of clusters

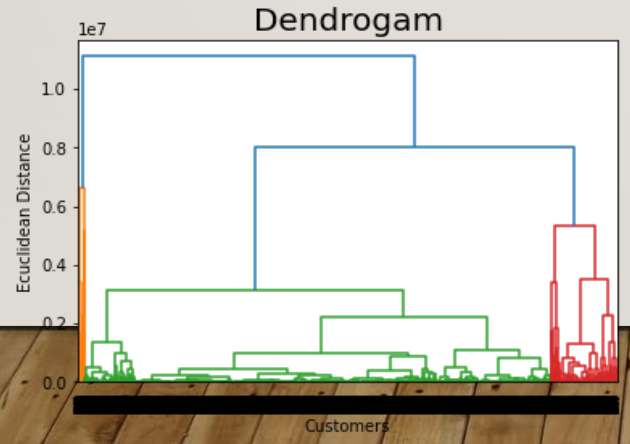

color
Others
Clothes
Accessories_Cosmetics

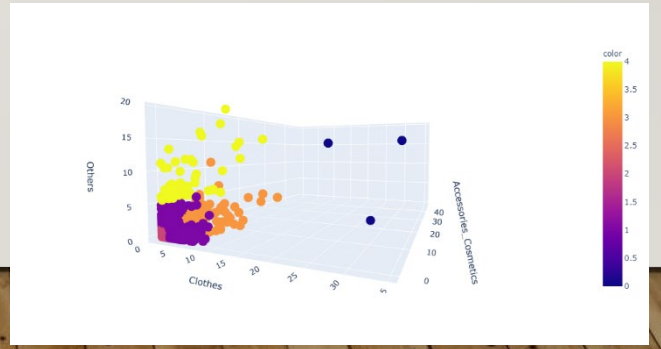

K-Medeoids
Accessories_Cosmetics
Clothes


Correlation between different features
Clothes
Accessories_Cosmetics
Others
1
0.45
0.42
0.45
1
0.53
0.42
0.53
1

AGGLOMERATIVE CLUSTERING

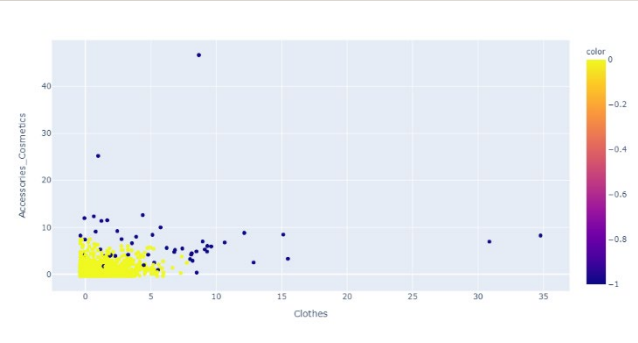
- The dendrogram of agglomerative clustering shows similar results with K-means. The longest range without juncture of the clusters is between Euclidian Distance of 0.3 and 0.5 that coincides with the number of clusters equal to 5.
- The agglomerative algorithm is performed for 5 clusters and the data is plotted. The 3D plot is made for a better visualization as shown in figure 12. It is observed that again the clusters are impacted by the outliers.

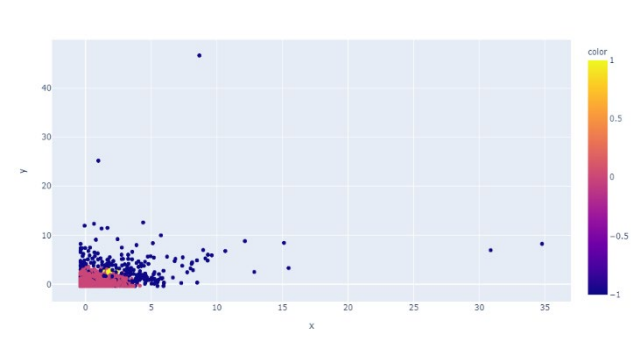


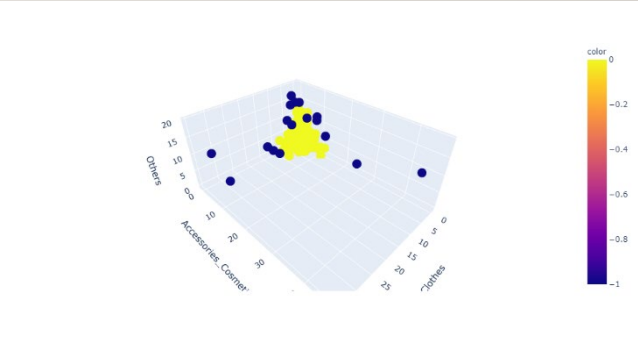


DBSCAN

- DBSCAN algorithm is peformed with different radiuses and min_samples.
- For the radius=2 and the min_samples=10, we have one major group, and the other data is considered as noise







Time Performance

- Times are calculated when running the algorithms in Jupiter Notebooks in a Intel(R) Core (TM) i5-6300U CPU @ 2.40GHz 2.50 GHz system.

Algorithm	Nr. of Iterations/Runs	Time is seconds
K-Means	10	3.2
K-Medoids	10	73
Dendrogram/Agglomerative	1	629
DBSCAN	10	8

CONCLUSIONS

- The usage of Clustering in CRM systems is a very effective technique for customer grouping producing very important information for business analysts.
- The implementations and tests are done on a real department store dataset in difference with most previous tests and evaluations that are done on predefined datasets that have known patterns.
- Results show that in case of noise and outliers in the data, K-Medoids is more robust than other algorithms and it is recommended to be used.
- We suggest using it for customer segmentation in department stores as usually the data are affected by outliers.
- Agglomerative Clustering is also affected by outliers.
- Simulations show that density-based clustering algorithm does not have a good performance when the density is similar in most of the dataset.

Conclusions - Continue

- The correlation between dimensions is argued by the physical location of “Accessories “and “Others”.
- The other data obtained by the clustering algorithm are delivered for further analysis to the marketing department.
- In terms of time performance in our dataset, K-Means and DBSCAN are the most performant as they do simple calculations.
- Computing K-Medoids requires pairwise distance to be calculated for each point in the dataset so it is computationally expensive.
- The same happens in computing a dendrogram as all possible cluster combinations are tested.
- We conclude that for large retail datasets, K-Means and DBSCAN are more time effective.

- Thank you for your attention