CLUSTER ANALYSIS OF SALES TRANSACTIONS DATASET

Student Name: Bubly Babu **Student ID: 22031115**

Introduction :

Prototype-based clustering, which employs the Fuzzy C-means algorithm, and K-means clustering are the two methods of data grouping that are compared in this assignment. These techniques are crucial for identifying patterns in data without the need for pre-established categories. Fuzzy C-Means is versatile for complicated structures since it permits data points to partially belong to various groups. K-Means, on the other hand, organizes data points differently. The examination examines how these techniques perform in various contexts, taking into account their adaptability, readability, sensitivity to outliers, and usefulness. The intention is to assist users in identifying their areas of strength and choosing the approach that best suits their data analysis requirements.

Data and preprocessing:

Utilized a weekly sales dataset that included quantities for 800 distinct products and covered 52 weeks. Each row signifies a distinct product, while each column corresponds to a specific week, including normalized data in the latter half of the columns. The findings reveal fluctuations in the dataset from 0 to a peak of 51 over the weeks. Notably, there are no missing values in this dataset sourced from the UCI Machine Learning Repository. As part of the preparatory steps, standard scaling of sales data ensures uniform magnitude across values. Employing Z-scores facilitates the identification of outliers, and then transforms them. Subsequently, the dataset is primed for further scrutiny, enabling robust clustering methods to unveil inherent patterns in weekly sales transactions.

Cluster analysis:

This study used two alternative clustering techniques to find hidden patterns in the dataset of weekly sales transactions: K-means and the Fuzzy C-means (FCM) algorithm, also referred to as a prototype-based clustering strategy. As shown by FCM, prototype-based clustering addresses the inherent ambiguity in the dataset by allowing data points to simultaneously belong to numerous clusters. This is accomplished by enabling every data point to be a part of many clusters with varying levels of membership. By minimizing the sum of squares inside each cluster, K-means, in contrast, allocates each data point to a single cluster. The rationale behind selecting these techniques arose from their complementing advantages: FCM tackled ambiguity in the dataset and intricate features of customer purchasing patterns, whereas K-means offered a streamlined and transparent approach to classify products into discrete groups.

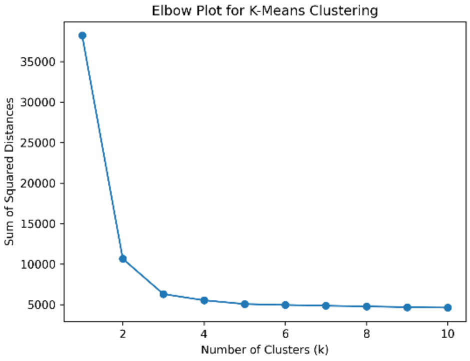


Fig 1 : Elbow Plot

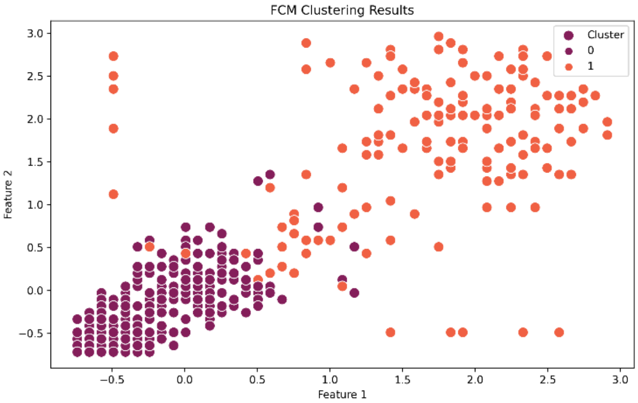


Fig 2 : FCM Clustering

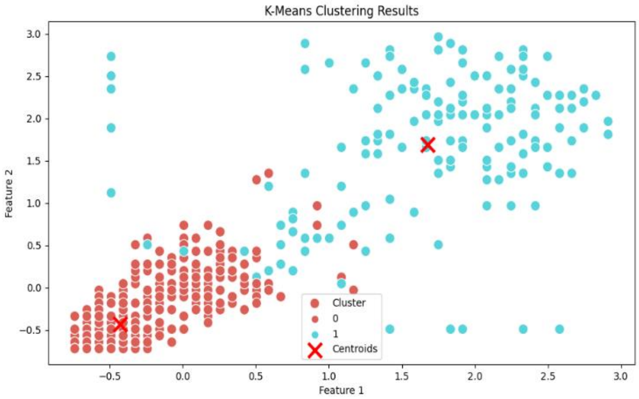


Fig 3 : K-means Clustering

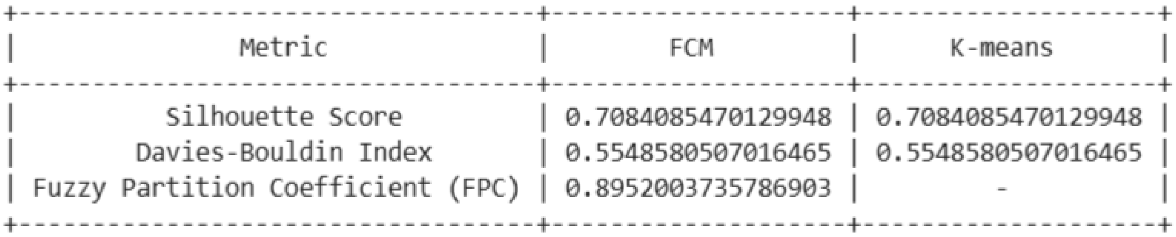
Notable trends were observed in the clustering results when the weekly sales transactions dataset was subjected to both the Fuzzy C-means (FCM) and K-means clustering techniques. With its design to manage fuzzy relationships, the FCM algorithm produced a significant Fuzzy Partition Coefficient (FPC) of 0.8952. This coefficient indicates that the dataset contains a significant amount of complexity, allowing products to show partial membership across multiple clusters. By contrast, K-means yielded an inertia value of 10,656.04, indicating a reduced degree of uncertainty because every product was clearly assigned to a single cluster.

Upon examination of the scatter plots generated by the two methods (Figs. 2 and 3), it was noted that the clusters displayed comparable patterns, underscoring the durability of the discovered structures. The FCM clusters had more fuzzy and subtle boundaries than the K-means clusters, which were more clearly defined. The dependability of identifying patterns in the sales transactions dataset is suggested by the consistency observed across clustering approaches. The agreement between the crisp clusters of K-means and the fuzzy clusters of FCM provides a comprehensive picture of the underlying trends in consumer purchase behaviour and boosts trust in the identified groupings.

Conclusion:

Strong clustering results for both the Fuzzy C-means (FCM) and K-means approaches are revealed by the evaluation metrics. Well-defined and cohesive clusters are indicated by a high Silhouette Score (0.7084), and their distinctiveness is confirmed by a low Davies-Bouldin Index (0.5549). Given the inherent uncertainty in the dataset, FCM's Fuzzy Partition Coefficient (FPC) of 0.8952 demonstrates how well it can manage uncertainty.

In conclusion, both methods do a good job of capturing the underlying structures of the weekly sales transactions dataset. The consistent FCM and K-means results, which provide meaningful information about the 52-week client buying patterns, reinforce the dependable clustering results.



```
+-----------------------------------+--------------------+--------------------+
|               Metric              |        FCM         |      K-means       |
+-----------------------------------+--------------------+--------------------+
|          Silhouette Score         | 0.7084085470129948 | 0.7084085470129948 |
|        Davies-Bouldin Index       | 0.5548580507016465 | 0.5548580507016465 |
| Fuzzy Partition Coefficient (FPC) | 0.8952003735786903 |         -          |
+-----------------------------------+--------------------+--------------------+
```

Table 1 : Evaluation Metric

Reference

archive.ics.uci.edu. (n.d.). *UCI Machine Learning Repository: Sales_Transactions_Dataset_Weekly Data Set*. [online] Available at: https://archive.ics.uci.edu/ml/datasets/Sales_Transactions_Dataset_Weekly.

Pang-Ning Tan (2020). *Introduction to data mining : international edition*. Harlow: Pearson Education.