Clustering for Better Breakfasts: A Comparison of K-Means and Hierarchical Algorithms.

1st Mustefa Abrahim
M.Tech in Data Science
Indian Institute of Technology Madras
Chennai, India
ge22m014@smail.iitm.ac.in

2nd Ahmed Shmels
M.Tech in Data Science
Indian Institute of Technology Madras
Chennai, India
ge22m009@smail.iitm.ac.in

3rd Kemal Mudie
M.Tech in Data Science
Indian Institute of Technology Madras
Chennai, India
ge22m010@smail.iitm.ac.in

***Abstract*—Breakfast is considered the most important meal of the day, and cereal is a popular and convenient choice. However, not all cereals are created equal, as they vary in their nutritional content. It is important to understand which cereals contain similar nutrients to help people make healthy choices and set them up for success for the day. The goal of this study is to use unsupervised learning methods of clustering to group cereals based on their nutritional attributes listed on cereal boxes. By clustering cereals into groups, we can explore what traits cause them to be similar and how this information can be used for marketing and product placement in stores. Additionally, individuals can use the clustering results to identify which cereals can be supplemented with others based on their similar nutritional attributes. The study will compare the effectiveness of two clustering algorithms, K-Means and Hierarchical, in predicting similar cereal varieties. Through this research, we aim to provide insights into how clustering can improve our understanding of breakfast cereal and its nutritional value, and ultimately promote healthy lifestyles.**

***Index Terms*—Unsupervised learning, K-Means Clustering, Hierarchical Clustering**

I. Introduction

The study involves the comparison of two clustering algorithms, hierarchical clustering and k-means clustering, for predicting similar breakfast cereal varieties. Four clustering methods were considered, including expectation-maximization and density clustering, but were not suitable for this study due to the specific dataset and desired objectives.

The hierarchical clustering method starts with every point as its own cluster and recursively joins clusters based on their similarity using a distance joining method such as variance, average, single, or complete (Alto, 2019). Divisive hierarchical clustering starts from one cluster and breaks it down. The advantage of hierarchical clustering is that subclusters are visible (Alto, 2019).

K-means clustering is a method that is centered on centroids. Each cluster has a centroid, and each point is attracted to a centroid based on its distance measure (Thompson, 2019). The final centroids and clusters are chosen by a recalculation of the means of all the points and distances of the points to the centers until the centroids do not move anymore. In this clustering method, all points will belong to a cluster, defined by a hard edge between the clusters. The number of clusters for k-means clustering is not predetermined, and two techniques, the elbow method (Fig.10) and the silhouette score (Fig.11) will be used to determine the optimal number of clusters for the cereal data (Thompson, 2019).

The choice of clustering method depends on the specific dataset and the desired objectives of the analysis. In this study, hierarchical clustering and k-means clustering were chosen as they provided better clustering results for predicting similar breakfast cereal varieties.

II. Methodology

A. Hierarchical Clustering

In this study, hierarchical agglomerative clustering was employed as the first clustering method, building clusters from the ground up. The dendrogram function of the SciPy library was utilized to perform hierarchical clustering, with different linkage methods (Single, Complete, Average, and Ward) being chosen. Among these methods, Ward's method (Fig. 1) was selected and is reported in this study. It works by minimizing the total within-cluster variance and is often preferred for its ability to produce compact and well-separated clusters (Murtagh & Legendre, 2014). Fig.1 displays the dendrogram of the cereal data, where each cereal begins as its own cluster and is then linked to the next based on the chosen linkage method, forming a hierarchy to the top. The dendrograms for the other linkage methods are available in Appendix C.

The distance on the y-axis representing dissimilarity is used to determine the optimal number of clusters. A line is drawn through the dendrogram, where the line can move up and down the longest distance before meeting a joining point. The number of lines that this cut goes through represents the optimum number of clusters. In this study, a red line was drawn through six lines in the dendrogram, resulting in six clusters, as depicted in the different colours in Fig.1. A breakdown of which cereals were clustered together can be seen in Appendix A.

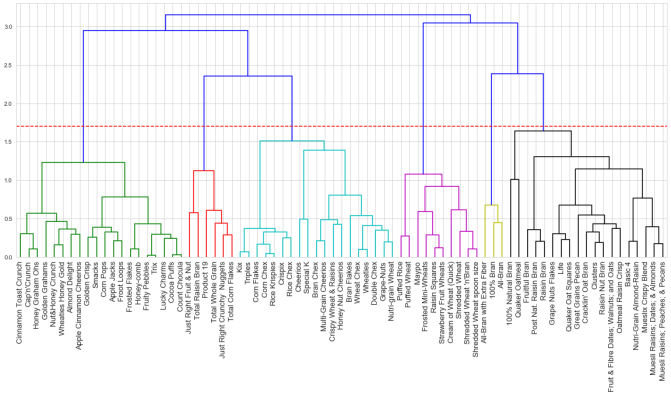

Fig. 1. Dendrogram with Ward's method

Examining the hierarchy of the clusters, it is interesting to see the sub-clusters take shape. There is a trend that the cereals tend to be combined with others that have similar names, such as in the 1st green cluster in Fig.1, where Nut & Honey Crunch was first combined with Wheaties Honey Gold. A few tiers later, it was combined with Honey Graham Ohs. Other similarities like this take place in the building of the hierarchy. This suggests that cereals with similar names may have similar traits or within-cluster variances that make them good candidates for clustering together.

Overall, hierarchical clustering provides a visual representation of the relationships between the cereals and can be useful for identifying subclusters and patterns in the data.

B. K-Means Clustering

The second clustering method presented in this study is K-means clustering. K-means clustering works by clustering the data around a centered point or centroids. Starting with a pre-determined number of k clusters, k centroids are randomly chosen, and data points are assigned to each cluster based on their closest distance. The centroids are then moved again, and after several iterations of minimizing the Euclidean distances from the points to the centroids, the final centroids are determined with the best clusters Fig.12 and Fig.13 depict the same. The goal of K-means clustering is to maximize intra-cluster cohesion and inter-cluster separation (Bonaccorso, 2020).

To determine the optimal number of clusters for the cereal dataset, the elbow test was conducted. The elbow test looks at the inertia or "average distance between samples and centroid" for different clusters by computing the sum of squared errors of each point to a cluster. As seen in Fig.10, the elbow test produces a graph that has a slight bend at six clusters, which is the number of clusters used in the K-means clustering analysis.

With six clusters chosen for k, the normalized data was put through the K-means clustering algorithm. Fig.2 shows a 3D space with the first three features, calories, protein, and fat. The stars represent the centroids, and the six colors represent the six different clusters. However, it was challenging to distinguish the different clusters due to overlapping colors. To address this issue, a test was done to see how the clusters would turn out with dimensionality reduction. Principal component analysis (PCA) is a dimensionality reduction technique for unsupervised learning methods that uses linear transformation to reduce dimensions (Ahmad, 2020). After conducting PCA and clustering with K-means, the six clusters were easier to distinguish in the 3-dimensional graph, as seen in Fig.3.

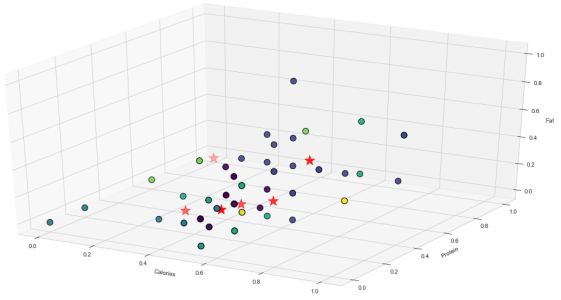

Fig. 2. K-means clustering with 6 clusters

Although the graphs look different between the two methods, the cereals were put into the same clusters for dimensionality reduction and with no dimensionality reduction. The cereals assigned to each cluster can be seen in Appendix B.

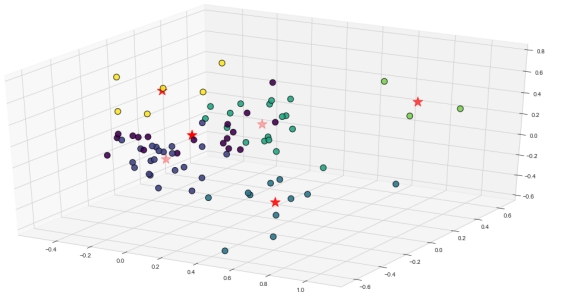

Fig. 3. K-means clustering using PCA dimensionality reduction

Overall, K-means clustering provides a way to partition the cereals into distinct groups based on their dietary features and can be useful for identifying patterns and similarities in the data.

III. Data

The dataset used in this analysis comprises 77 different cereals (Fig.4) sourced from Kaggle (Crawford, 2018). Prior to analysis, the dataset underwent cleaning to ensure that no missing or duplicated values were present.

The dataset's columns include calories per serving, grams of protein and fat, milligrams of sodium and potassium,

grams of dietary fiber and complex carbohydrates, grams of sugars, and vitamins and minerals expressed as 0, 25, or 100, indicating the typical percentage of FDA recommended intake. Additionally, the shelf feature provides information on the display location of the cereal, with values of 1, 2, or 3 indicating the shelf position counting from the floor. Out of

name	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
100%_Bran	70	4	1	130	10	5	6	280	25	3	1	0.33	68.402973
100%_Natural_Bra	120	3	5	15	2	8	8	135	0	3	1	1	33.983679
All-Bran	70	4	1	260	9	7	5	320	25	3	1	0.33	59.425505
All-Bran_with_Ext	50	4	0	140	14	8	0	330	25	3	1	0.5	93.704912
Almond_Delight	110	2	2	200	1	14	8		25	3	1	0.75	34.384843
Apple_Cinnamon_	110	2	2	180	1.5	10.5	10	70	25	1	1	0.75	29.509541
Apple_Jacks	110	2	0	125	1	11	14	30	25	2	1	1	33.174094
Basic_4	130	3	2	210	2	18	8	100	25	3	1.33	0.75	37.038562
Bran_Chex	90	2	1	200	4	15	6	125	25	1	1	0.67	49.120253

Fig. 4. Example of the first ten rows and 10 columns of the data-set

the dataset's 16 features, 9 variables were selected for use in the clustering techniques: calories, protein, fat, sodium, fiber, carbohydrates, sugars, potassium, and vitamins, as shown in Fig.4. These variables are continuous dietary data used to evaluate the similarities between cereals. The vitamin feature is expressed as a percentage of the FDA-recommended daily amount.

The dataset also includes other features such as cereal name, shelf location at grocery stores, serving weight in ounces, number of cups in one serving, and a cereal rating, which may be from Consumer Reports. However, these features were not used in the clustering analysis but were utilized to gain insights from the results.

To eliminate skewness in the clustering, the selected variables were normalized. This was necessary as large values in one variable, such as calories, could have portrayed a large distance from grams of fat or protein. All dimensions were standardized to the same scale to ensure fairness in determining the Euclidean distance between points.

A. Exploratory Data Analysis

Exploratory data analysis (EDA) was conducted on the data to gain insights into its characteristics. A 3D scatter plot was created to visualize the dimensions of focus, revealing a possible cluster on the left of the graph in a cube feature space, as shown in Fig.5. Additionally, a 2D scatter plot was generated, as seen in Fig.6.

Histograms were used to assess the distribution of the data, with some variables displaying a normal distribution while others had no clear distribution. Consequently, the expectation-maximization method for clustering was not used, given that not all variables had a Gaussian distribution as seen in Fig.7.

A scatter plot matrix was employed to identify any visual relationships between certain variables, revealing a positive relationship between fiber and potassium, as shown in Fig.8. Finally, a covariance matrix with a heat map (Fig.9) was used, confirming the strong correlation between fiber and potassium.

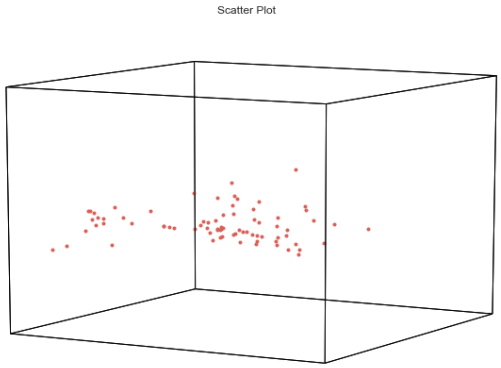

Fig. 5. depicts Cube scatter plot of data

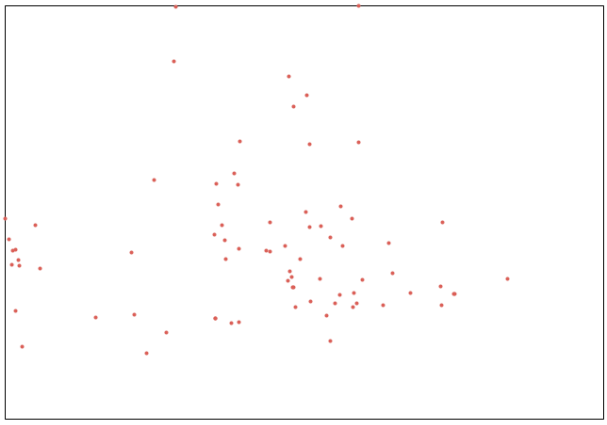

Fig. 6. depicts 2D scatter plot of data

Other variables that exhibited a decent correlation include protein and potassium, protein and fiber, calories and sugars, and calories and fat.

IV. Model Evaluation

The comparison between the two clustering methods, hierarchical with agglomerative Ward's method and K-means, revealed that both techniques produced nearly similar clustering results for the cereals, dividing them into six groups. Appendix D illustrates the clustering results of both methods, with slight variations represented by different colors. However, four cereals were clustered differently between the two methods. In hierarchical clustering, two cereals in Cluster B were clustered in Cluster A in K-means, and two cereals in Cluster C were present in Cluster B and D in K-means. Clusters E and F were consistent across both methods. For more detailed information on the clustering differences by cereal names, please refer to Appendix A and B.

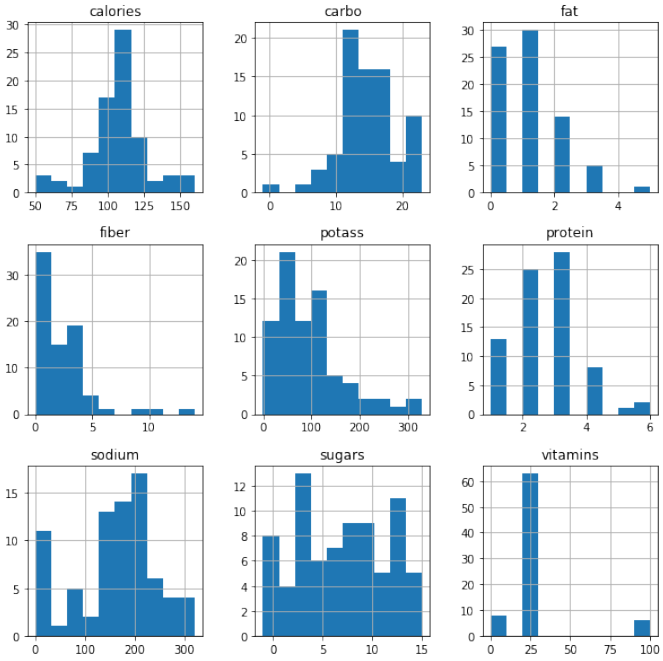

Fig. 7. depicts Histograms of each feature

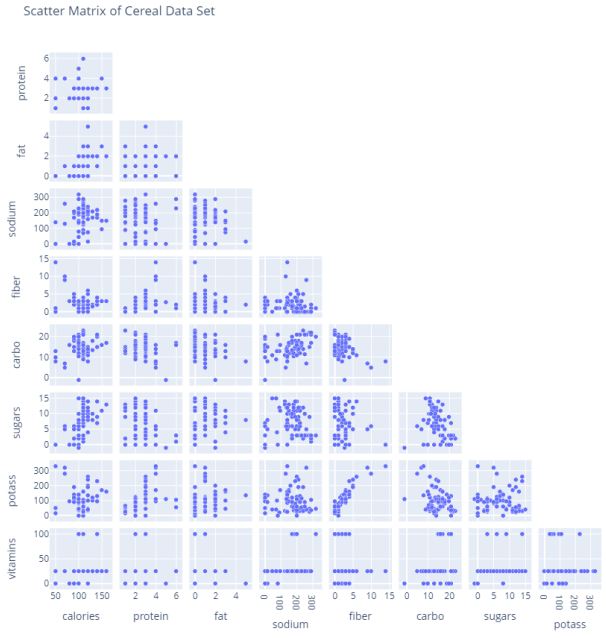

Fig. 8. Depicts Scatter plot matrix between each feature

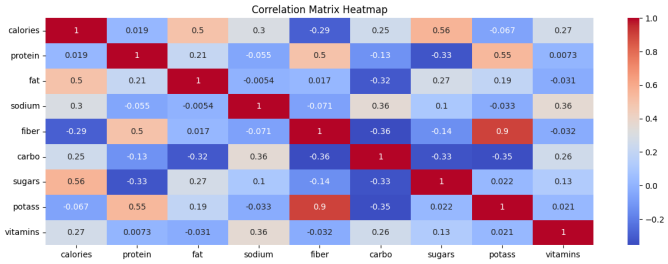

Fig. 9. Depicts Covariance heat map between each feature

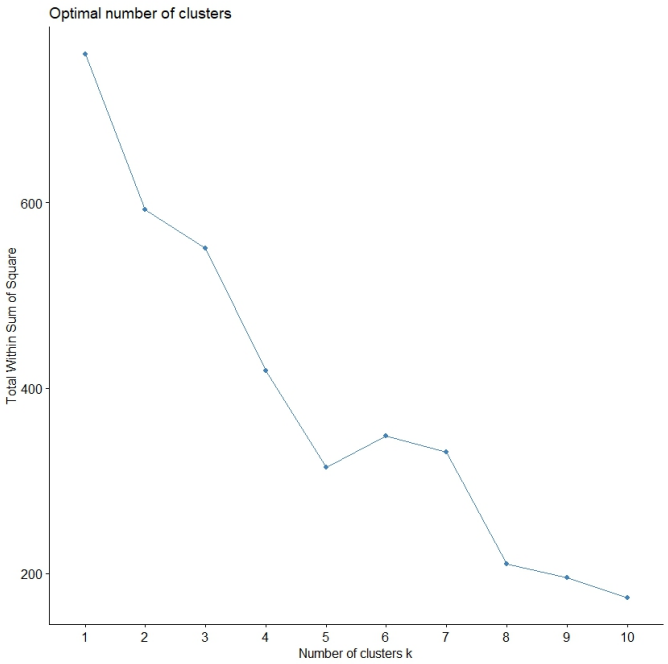

Fig. 10. Elbow Test

A. *Accuracy*

Unsupervised clustering aims to differentiate the data points as accurately as possible. The silhouette score was computed (Fig.11) to assess the clustering performance and the K-means method's accuracy. The silhouette coefficient measures the "closeness of each point in a particular cluster with respect to the other points in the neighboring clusters" (Ahmad, 2020). The scores range from 0 to 1 and are categorized into four ranges for accuracy: excellent, reasonable, weak, and no clustering has been found (Ahmad, 2020).

The closest score to 1 indicates the optimum number of clusters. As shown in Fig.11, the silhouette score for six clusters was the highest, which was consistent with the number of clusters determined by the elbow test. However, the score for six clusters was approximately 0.32, which falls within the weak category (range of 0.26-0.50), indicating that the quality of clusters is not very reliable (Ahmad, 2020).

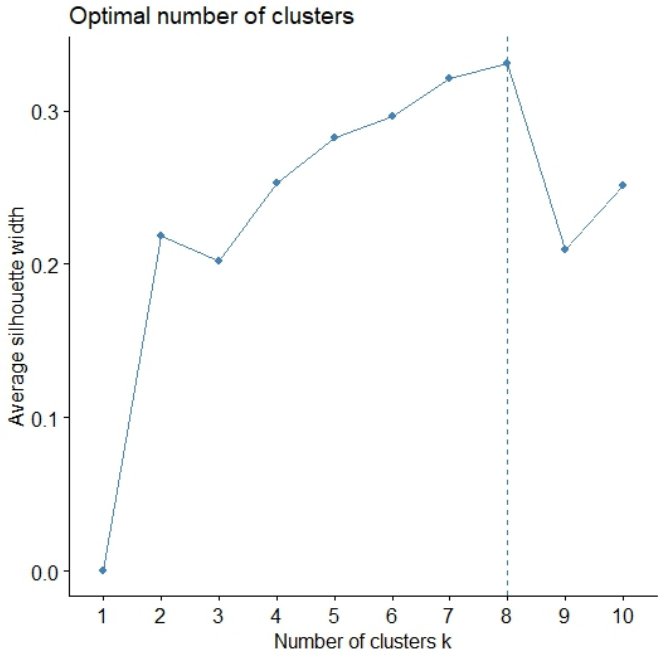

Fig. 11. Silhouette Score

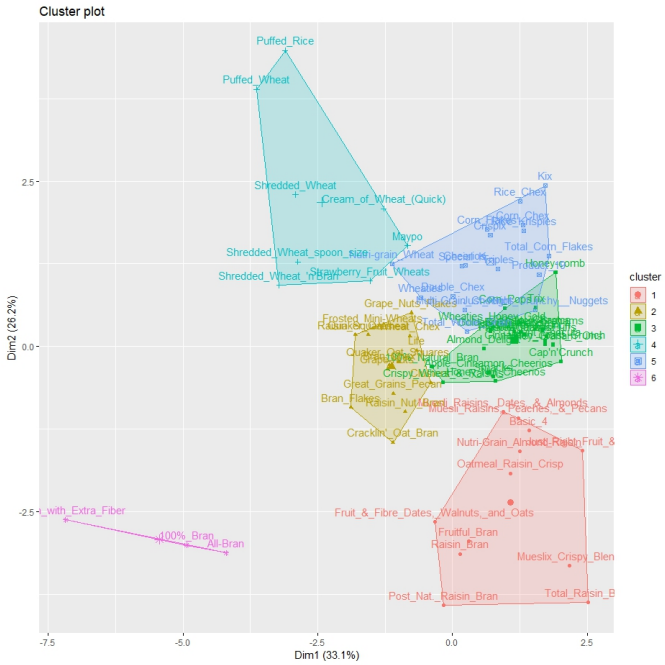

Fig. 12. K-Means with 6 cluster

V. Results and Insights

To gain insights into why certain cereals were assigned to specific clusters, histograms of each K-means cluster and attributes were generated to identify the frequent variables. This provided valuable insights into why these cereals were similar.

Cluster A was characterized by cereals with no or very low fiber and protein, high sugar content, and calorie counts ranging from 100-110. The cereals in this cluster included Froot Loops, Fruity Pebbles, Lucky Charms, and Trix, which are popular with children and are considered unhealthy.

Cluster B included cereals with low fiber content, around 100 calories, and higher levels of sodium. The cereals in this cluster contained words such as ”nuts,” ”grains,” ”chex,” and ”rice.” Cluster C was characterized by cereals with 3 grams of protein, higher fat and calorie content, and included Cream of Wheat, Shredded Wheat, Shredded Wheat ’n’Bran, Shredded Wheat spoon size, Puffed Wheat, and Frosted Mini-Wheats, along with a few others without wheat in the name.

Cluster D included cereals with no or low sodium, sugar, and fat content, and low fiber content. The cereals in this cluster included Cream of Rice, Rice Krispies, Rice Chex, and Crispix.

Cluster E was characterized by low-calorie cereals, high potassium and fiber content, and contained the word ”bran” in their names. This cluster included Bran Flakes, Raisin Bran, and Total Whole Grain. Lastly, Cluster F included cereals with low-fat content, moderate sugar content, and high levels of sodium. The cereals in this cluster included Cheerios, Kix, and Wheaties.

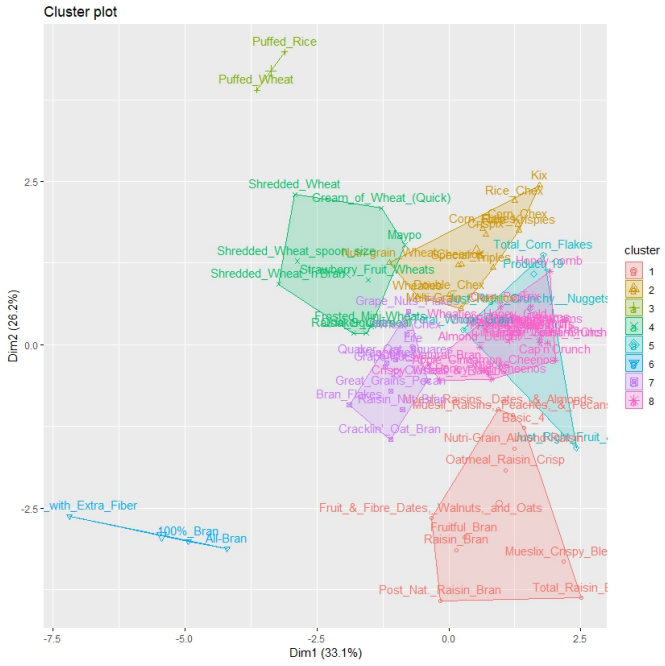

Fig. 13. K-Means with optimum cluster(8 cluster

Additionally, it was interesting to compare the final clusters to the in-store shelf placement of the cereals. The shelf placement was excluded from the algorithms to cluster solely on dietary means. The frequency of shelf placement per cluster indicated that Clusters E and F, with the majority of Cluster C, were located on the top shelf. These cereals had healthier attributes and were somewhat out of sight. Cluster B spanned the 1st and 3rd shelf, and Clusters A and D spanned all three shelves, with A being prominent on the middle shelf. These were the unhealthier cereals targeted towards children and were placed at eye level. Figure 12 shows histograms for calories, sugars, sodium, and fiber, while Figure 13 presents the frequency of shelf placement per cluster.

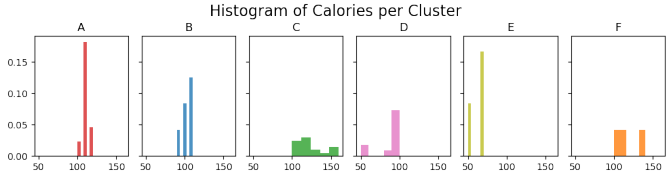

Fig. 14. Histograms by Cluster for Calories

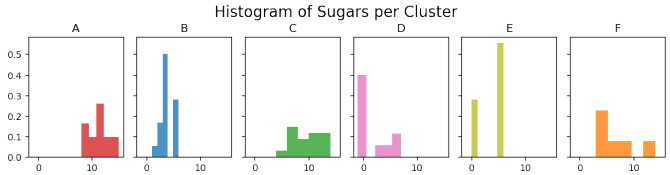

Fig. 15. Histograms by Cluster for Sugars

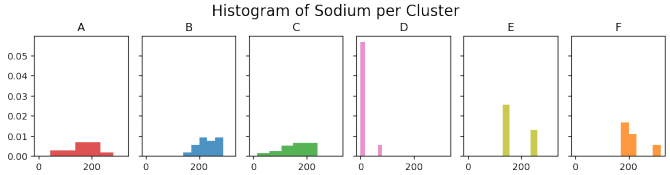

Fig. 16. Histograms by Cluster for Sodium

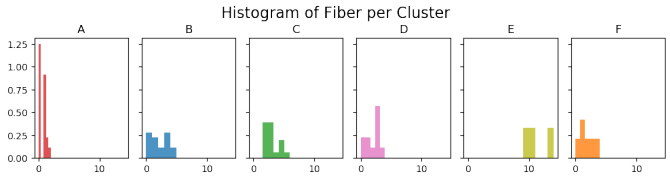

Fig. 17. Histograms by Cluster for Fiber

VI. Conclusion

This paper presents two methods for clustering the cereals, hierarchical clustering and K-means clustering. Hierarchical clustering minimized within cluster variance, building a tree

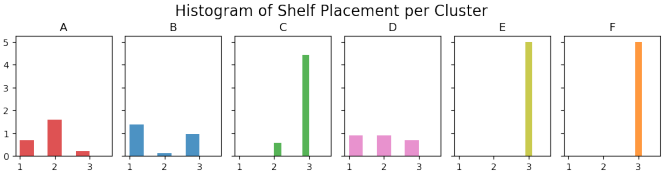

Fig. 18. Histogram of shelf placement per cluster

like graph, adding to each grouping as the tree grew as seen in Fig.1.

K-means clustering grouped cereals by minimizing the distance to the nearest centroids. Although the means to develop their clusters were different, both clustering methods produced very similar clusters Fig.12 and Fig.13 depict how K-means works when the K value is 6 and 8. The goal of clustering the cereals was to see if similar cereals could be grouped together based on their dietary attributes. These groupings would then determine what cereals are grouped as healthy and unhealthier. Although the accuracy test of the clusters was on the weak scale, the histograms of each attribute of these clusters told a different story. There is underlying information on which cereals are similar based their attributes, of similar calories, fat, protein, sugars, sodium, fiber, and potassium. These findings also were represented in the product placement in store aisles. Cereals that have a healthier diet are placed on the shelf that is further away than those that have less nutrients. It was also found that cereals with common words in names are also similar.

Based on these findings, the individual consumer can see where one cereal can be substituted for another within each cluster. Individual taste can act on the deciding factor of which cereal to consume based on a similar cereal.

Appendix

Appendix A

Cereals Clustered by Hierarchical Clustering Each cluster contains the cereal that the hierarchical method grouped it with. Highlighted colors signify the cereals grouped differently than in K-means clustering.

Appendix B

Cereals Clustered by K-Means Clustering Each cluster contains the cereal that the K-means method grouped it with. Highlighted colors signify the cereals grouped differently than in hierarchical clustering.

Appendix C

Other commonly used linkage method in hierarchical clustering is Single linkage, which is the distance between two clusters as the minimum distance between any two points in the clusters. Average linkage is the distance between two clusters as the average distance between all pairs of points in the clusters, and Complete linkage, which is the distance between two clusters as the maximum distance between any two points in the clusters, produces compact, well-separated clusters but can be sensitive to outliers and noise.

Cluster A	Cluster B	Cluster C
Almond Delight	Bran Chex	100% Natural Bran
Apple Cinnamon Cheerios	Bran Flakes	Basic 4
Apple Jacks	Cheerios	Clusters
Cap'n'Crunch	Corn Chex	Cracklin' Oat Bran
Cinnamon Toast Crunch	Corn Flakes	Fruit & Fibre Dates; Walnuts; and Oats
Cocoa Puffs	Crispix	Fruitful Bran
Corn Pops	Crispy Wheat & Raisins	Grape Nuts Flakes
Count Chocula	Double Chex	Great Grains Pecan
Froot Loops	Grape-Nuts	Life
Frosted Flakes	Honey Nut Cheerios	Muesli Raisins; Dates; & Almonds
Fruity Pebbles	Kix	Muesli Raisins; Peaches; & Pecans
Golden Crisp	Multi-Grain Cheerios	Mueslix Crispy Blend
Golden Grahams	Nutri-grain Wheat	Nutri-Grain Almond-Raisin
Honey Graham Ohs	Rice Chex	Oatmeal Raisin Crisp
Honey-comb	Rice Krispies	Post Nat. Raisin Bran
Lucky Charms	Special K	Quaker Oat Squares
Nut&Honey Crunch	Triples	Quaker Oatmeal
Smacks	Wheat Chex	Raisin Bran
Trix	Wheaties	Raisin Nut Bran
Wheaties Honey Gold		
Cluster D	**Cluster E**	**Cluster F**
Cream of Wheat (Quick)	100% Bran	Just Right Crunchy Nuggets
Frosted Mini-Wheats	All-Bran	Just Right Fruit & Nut
Maypo	All-Bran with Extra Fiber	Product 19
Puffed Rice		Total Corn Flakes
Puffed Wheat		Total Raisin Bran
Raisin Squares		Total Whole Grain
Shredded Wheat		
Shredded Wheat 'n'Bran		
Shredded Wheat spoon size		
Strawberry Fruit Wheats		

Fig. 19. Appendex A

Cluster A	Cluster B	Cluster C
Almond Delight	Bran Chex	100% Natural Bran
Apple Cinnamon Cheerios	Bran Flakes	Basic 4
Apple Jacks	Cheerios	Clusters
Cap'n'Crunch	Corn Chex	Cracklin' Oat Bran
Cinnamon Toast Crunch	Corn Flakes	Fruit & Fibre Dates; Walnuts; and Oats
Cocoa Puffs	Crispix	Fruitful Bran
Corn Pops	Double Chex	Great Grains Pecan
Count Chocula	Grape Nuts Flakes	Life
Crispy Wheat & Raisins	Grape-Nuts	Muesli Raisins; Dates; & Almonds
Froot Loops	Kix	Muesli Raisins; Peaches; & Pecans
Frosted Flakes	Multi-Grain Cheerios	Mueslix Crispy Blend
Fruity Pebbles	Nutri-grain Wheat	Nutri-Grain Almond-Raisin
Golden Crisp	Rice Chex	Oatmeal Raisin Crisp
Golden Grahams	Rice Krispies	Post Nat. Raisin Bran
Honey Graham Ohs	Special K	Quaker Oat Squares
Honey Nut Cheerios	Triples	Raisin Bran
Honey-Comb	Wheat Chex	Raisin Nut Bran
Lucky Charms	Wheaties	
Nut&Honey Crunch		
Smacks		
Trix		
Wheaties Honey Gold		
Cluster D	**Cluster E**	**Cluster F**
Cream of Wheat (Quick)	100% Bran	Just Right Crunchy Nuggets
Frosted Mini-Wheats	All-Bran	Just Right Fruit & Nut
Maypo	All-Bran with Extra Fiber	Product 19
Puffed Rice		Total Corn Flakes
Puffed Wheat		Total Raisin Bran
Quaker Oatmeal		Total Whole Grain
Raisin Squares		
Shredded Wheat		
Shredded Wheat 'n'Bran		
Shredded Wheat spoon size		

Fig. 20. Appendex B

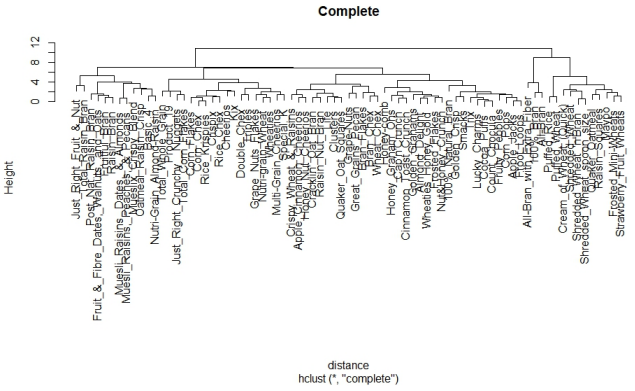

Fig. 21. Hierarchical with complete dendrogram method

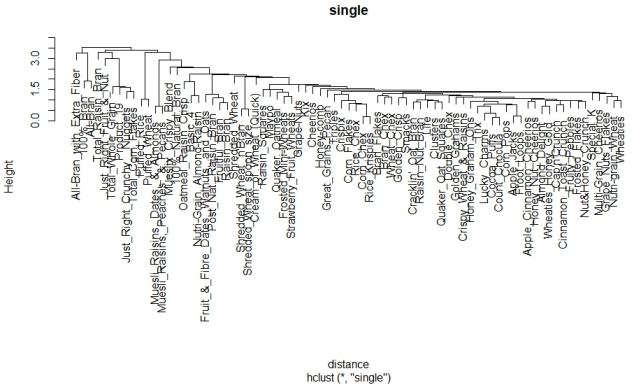

Fig. 22. Hierarchical with Single dendrogram method

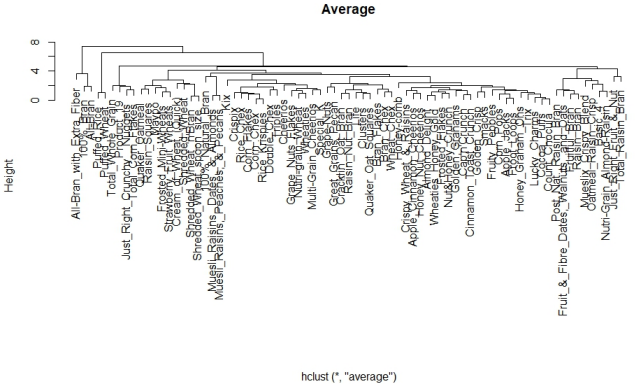

Fig. 23. Hierarchical with Average dendrogram method

APPENDIX D

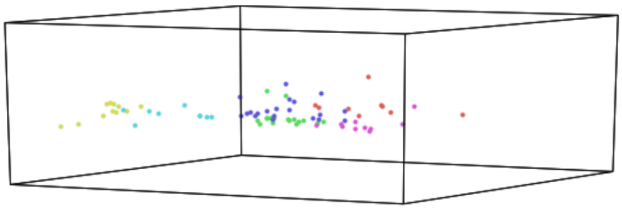

Fig. 24. K-means clustering

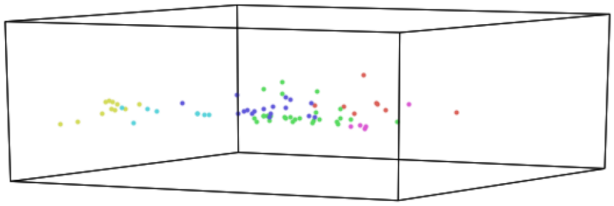

Fig. 25. Hierarchical clustering

REFERENCES

[1] Crawford, Chris. "80 cereals." Retrieved from https://www.Kaggle.Com/crawford/80-cereals? Select=cereal.Csv.

[2] Wati, M., Rahmah, W. H., Novirasari, N., Budiman, E. (2021, March). Analysis K-Means Clustering to Predicting Student Graduation. In *Journal of Physics: Conference Series* (Vol. 1844, No. 1, p. 012028). IOP Publishing.

[3] Ahmad, I. (2020). *40 Algorithms Every Programmer Should Know: Hone your problem-solving skills by learning different algorithms and their implementation in Python*. Packt Publishing Ltd.

[4] Arora, R. K., Badal, D. (2013). Evaluating student's performance using k-means clustering. *International Journal of Computer Science And Technology*, 4(2), 553-557.

[5] Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." *Journal of the royal statistical society. series c (applied statistics)* 28.1 (1979): 100-108.

[6] Pham, D. T., Dimov, S. S., Nguyen, C. D. (2005). Selection of K in K-means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 219(1), 103-119.

[7] Nielsen, F., Nielsen, F. (2016). Hierarchical clustering. *Introduction to HPC with MPI for Data Science*, 195-211.

[8] Wills, G. J. (1998, October). An interactive view for hierarchical clustering. In Proceedings IEEE Symposium on Information Visualization (Cat. No. 98TB100258) (pp. 26-31). IEEE.

[9] P. Mohan, B. Lee, T. Chaspari, and C. R. Ahn, "Capturing occupant routine behaviors in smart home environment using hierarchical clustering models," in Proceedings Construction Research Congress 2020: Computer Applications, pp. 1310-1318, American Society of Civil Engineers, Reston, VA, Nov. 2020.

[10] G. Bonaccorso, Mastering Machine Learning Algorithms, 2nd ed. Packt Publishing, 2020.

[11] C. Crawford, "80 Cereals," Retrieved from https://www.kaggle.com/crawford/80-cereals?select=cereal.cs, 2018.

[12] E. H. Ruspini, "A new approach to clustering," Information and Control, vol. 15, no. 1, pp. 22-32, 1969.