Similar Breakfast Cereal

Torrey Capobianco

DSC550 Data Mining

Bellevue University

Summer 2020

There is a saying that breakfast is the most important meal of the day. Cereal is one of several common food choices to consume in the morning. Although this choice is popular and quick, this singular word, cereal, that represents this category of food in fact should not be created equal, as all cereals do not contain the same dietary contents. Understanding which nutrients the cereal contains can determine how one's day is set up for success and can contribute to a healthy life style. With many choices down the grocery store aisle, which cereals are similar in these dietary features? Through the unsupervised learning method of clustering, this study is looking at if cereals can be grouped based on different attributes that would appear from nutrition facts listed on cereal boxes. With the ability to cluster the cereals in groups, this paper will explore what traits cause them to be similar, and how this plays in to marketing with in store product placement. On personal interests, one can see what cereals can be supplemented with another based on their similar traits.

Related Work

When researching similar breakfast cereals, the most popular results are based on what has good taste, as represented in an article in the Los Angeles Times in 2019 (Kwan Peterson, 2019). To combat nutrition, however, there have been a few reports on the comparison of breakfast cereal, based on healthiness. ACalorieCounter (2019) shared their list of cereals and dietary facts side by side, for the observer to make their own observations. The most interesting one, however, has been implemented in Australia. Voluntary to the distributers, the government created a star rating program to display on their packaging that compares products. Kellogg has implemented this rating system that uses "an equation which takes into account the energy, saturated fat, sugar and sodium contents of the product and balances that against the positives of protein and fibre content" (Kellogg's, n.d.). The methods used in this paper will look at the

similar features that Kellogg did, but through different clustering means. The purpose is to find clusters of cereals and then within those clusters determine what attributes make them similar.

Methods

There are four main clustering methods, two of which will be applied in the study based on the type of data being used. The four types are expectation-maximization, density clustering, hierarchical clustering, and k-means clustering. Choosing the right methods often depends on what is trying to be achieved and what type of data is being used. Expectation-maximization is based on the probability that a point belongs to a cluster, and is assumed that all features have a Gaussian distribution, or is normally distributed (Thompson, 2019). Due to several features from this study not having normal distributions through the view of their histograms, it was determined expectation-maximization was not a clustering method to use. Density clustering does not look at every point but only looks at those that are densely close, leaving some points out (ODSC, 2018). Through the clustering technique, the density clustering algorithm determines how many clusters to output. The DBSCAN algorithm was tested in this study. Each time this method was tested with different parameters, it always resulted in one cluster. The density clustering method did not produce the types of clusters that the project was aiming for, thus additional methods were explored.

The two methods presented are hierarchical clustering and k-means clustering. Hierarchical clustering starts with every point as its own cluster. From there, a distance joining method is used, whether it be by variance, average, single (close points), or complete (distance between farthest points). Each point is joined with another, through each iteration, in a hierarchical manner, until they are joined at the top. This is called agglomerative hierarchical clustering, by starting at the bottom with individual clusters, vs. divisive starts from one cluster

and breaks them down (Alto, 2019). The advantage of hierarchical clustering is that sub clusters are visible. This type of clustering is uniquely displaced in a tree like form as opposed to clusters of groups in a scatter plot type graph. Through the use of a dendrogram, one can determine how many clusters are optimized for this method.

The second method used for this analysis is k-means clustering. K-means clustering is centered on centroids. Each cluster has a centroid and each point is attracted to a centroid based on its distance measure. 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 (Thompson, 2019). In this clustering method, all points will belong to a cluster, defined by a hard edge between the clusters. K-means clustering does not pick the number of clusters to be used. Two techniques, the elbow method and the silhouette score will be demonstrated to confirm the number of clusters the cereal should be split in to.

Data

The data that is used for this analysis consists of 77 different cereals and was retrieved from Kaggle (Crawford, 2018). The original data was cleaned by Petra Isenberg, Pierre Dragicevic, and Yvonne Jansen. It did not include missing or duplicated values. The data consists of 16 features, 9 of which will be used in the clustering techniques. These 9 variables that will be considered in the clusters are calories, protein, fat, sodium, fiber, carbohydrates, sugars, potassium, and vitamins. All are dietary continuous data used to evaluate the similarities between cereals. The vitamins feature consists of a percentage of the FDA recommended daily amount. Other features that are not included in the clustering but will be used to provide insights from results are the cereal name and the shelf it is located on at grocery stores. The selected features for clustering were normalized so there would be no skewedness in the clustering. Large

calorie amounts would portray a large distance from grams of fat or protein. All dimensions were put on the same scale for complete fairness in determining the Euclidean distance between points.

Exploratory Data Analysis

The data was first observed through exploratory data analysis (EDA) practices. A scatter plot of the dimensions of focus was created in 3D to get a feel of what the data looks like. By looking at figure 1, there might be a possible cluster in the left of the graph in a cube feature space. Figure 2 shows the scatter plot of the data in 2D.

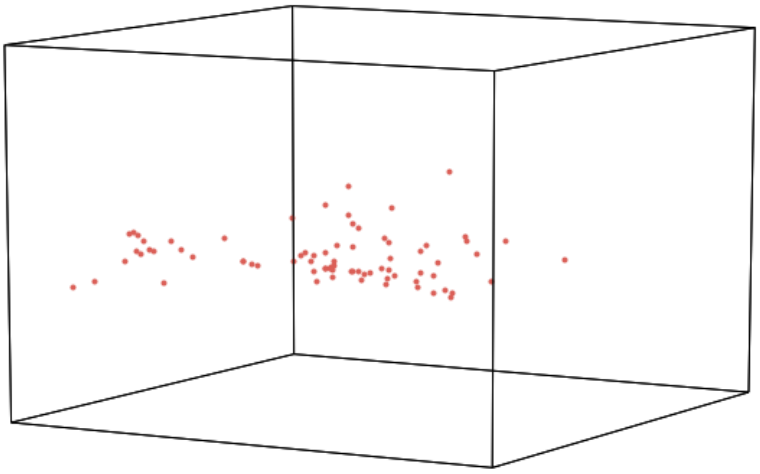

Figure 1: Cube scatter plot of data

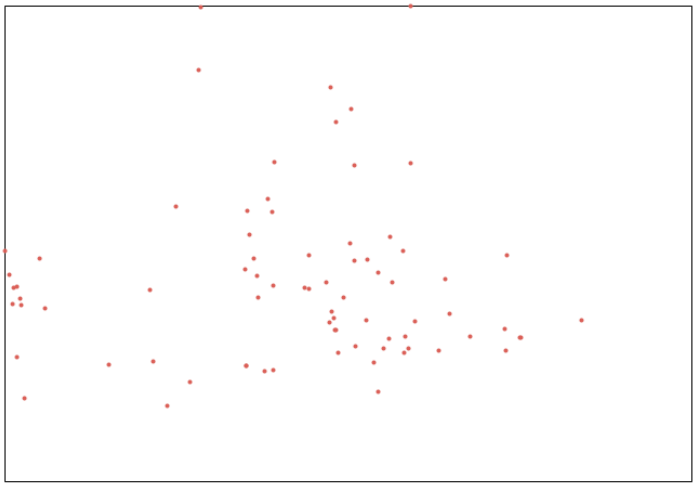

Figure 2: 2D scatter plot of data

Viewing the histograms in figure 3, one can see that there is a mixture of normally distributed data and nonparametric data, having no distribution. This was the determination of not to use the expectation-maximization method for clustering, of everything not having a Gaussian distribution.

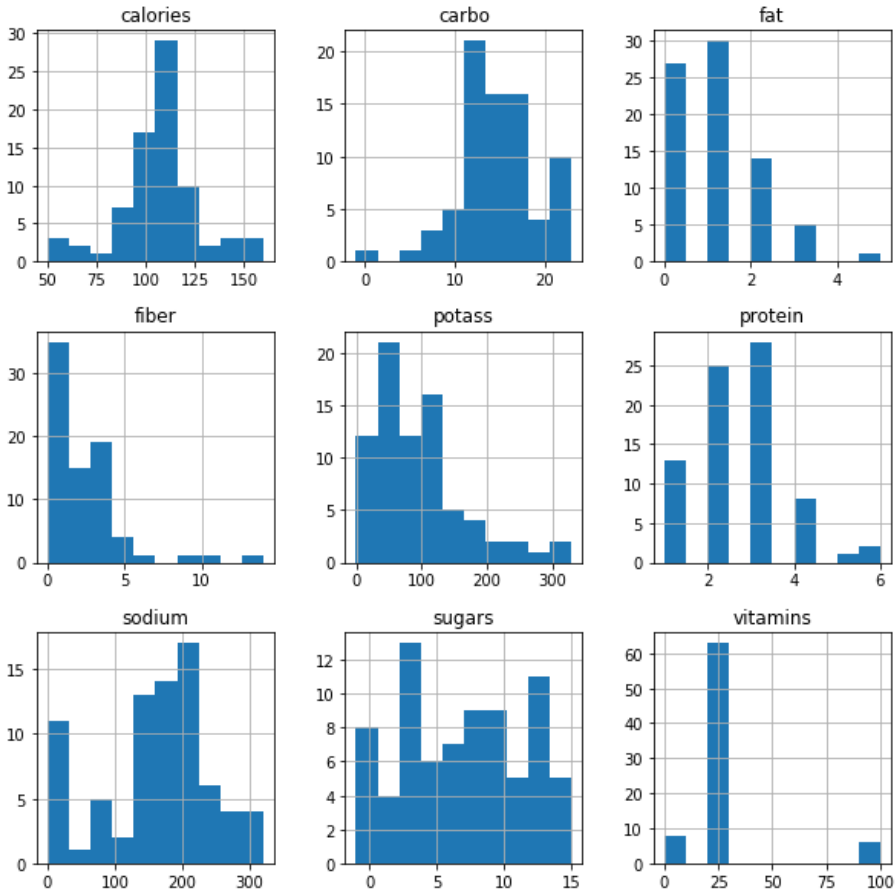

Figure 3: Histograms of each feature

A scatter plot matrix was used to look at any visual relationships between certain variables. There is an apparent positive relationship between fiber and potassium as seen in figure 4. The last exploratory process that was used was a covariance matrix with a heat map. This confirmed what was seen in the scatter plots between fiber and potassium as having a strong correlation. Other variables that have a decent correlation are protein and potassium, protein and fiber, calories and sugars, and calories and fat.

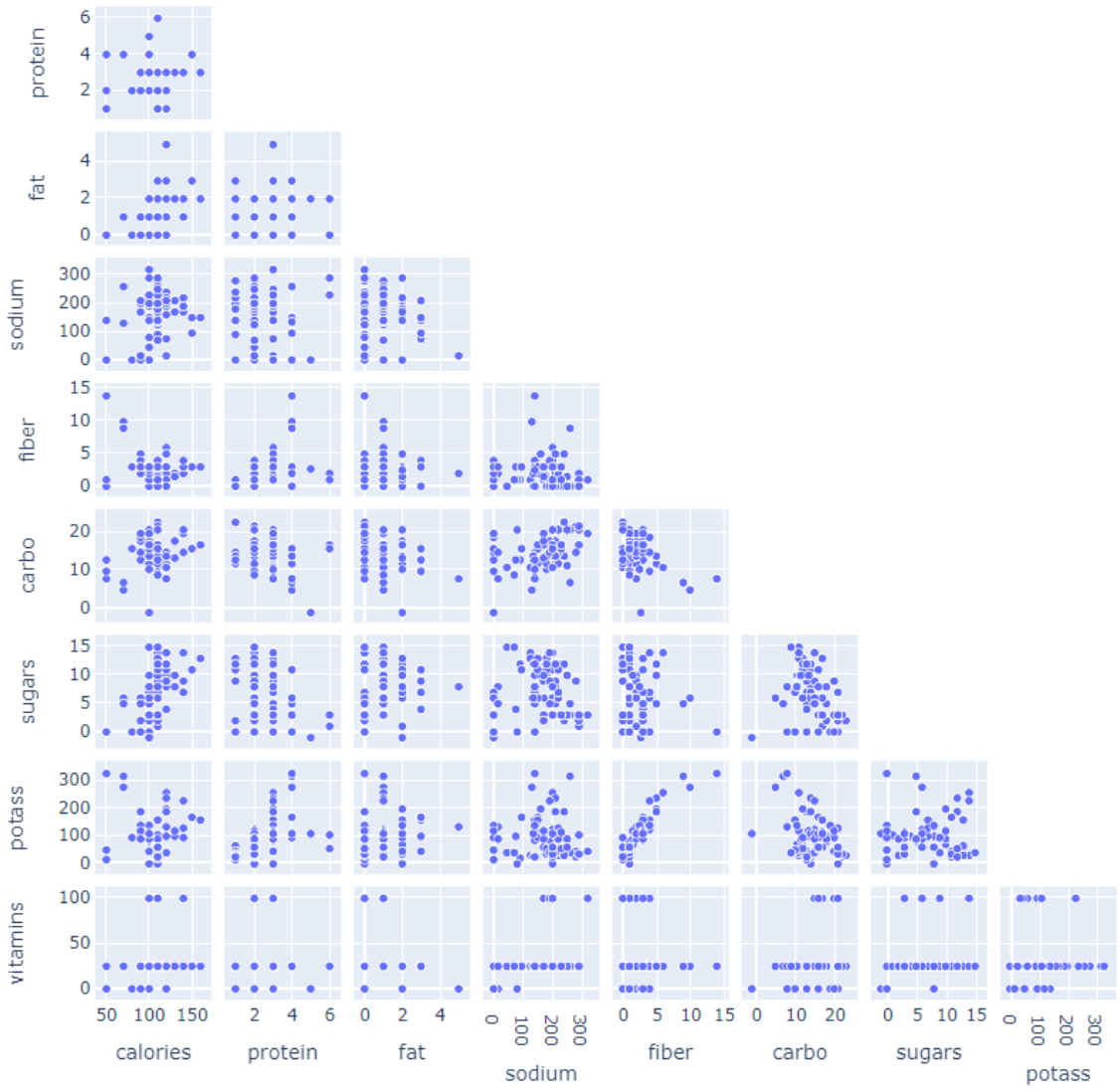

Figure 4: Scatter plot matrix between each feature

	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins
calories	1	0.0190661	0.49861	0.300649	-0.293413	0.250681	0.56234	-0.0666089	0.265356
protein	0.0190661	1	0.208431	-0.0546743	0.50033	-0.130864	-0.329142	0.549407	0.00733537
fat	0.49861	0.208431	1	-0.00540746	0.0167192	-0.318043	0.270819	0.193279	-0.0311563
sodium	0.300649	-0.0546743	-0.00540746	1	-0.070675	0.355983	0.101451	-0.0326035	0.361477
fiber	-0.293413	0.50033	0.0167192	-0.070675	1	-0.356083	-0.141205	0.903374	-0.0322427
carbo	0.250681	-0.130864	-0.318043	0.355983	-0.356083	1	-0.331665	-0.349685	0.258148
sugars	0.56234	-0.329142	0.270819	0.101451	-0.141205	-0.331665	1	0.0216958	0.125137
potass	-0.0666089	0.549407	0.193279	-0.0326035	0.903374	-0.349685	0.0216958	1	0.0206987
vitamins	0.265356	0.00733537	-0.0311563	0.361477	-0.0322427	0.258148	0.125137	0.0206987	1

Figure 5: Covariance heat map between each feature

Hierarchical Clustering

The first method of clustering presented is hierarchical agglomerative clustering, building from the ground up. To conduct hierarchical clustering, SciPy's dendrogram function was used. The type of linkage is that was used is Ward's method. Ward's method clusters by minimizing the total within-cluster variance (Wikipedia, n.d.). Figure 6 depicts this dendrogram of the data. As one can see, each cereal starts off as its own cluster and is then linked to the next one based

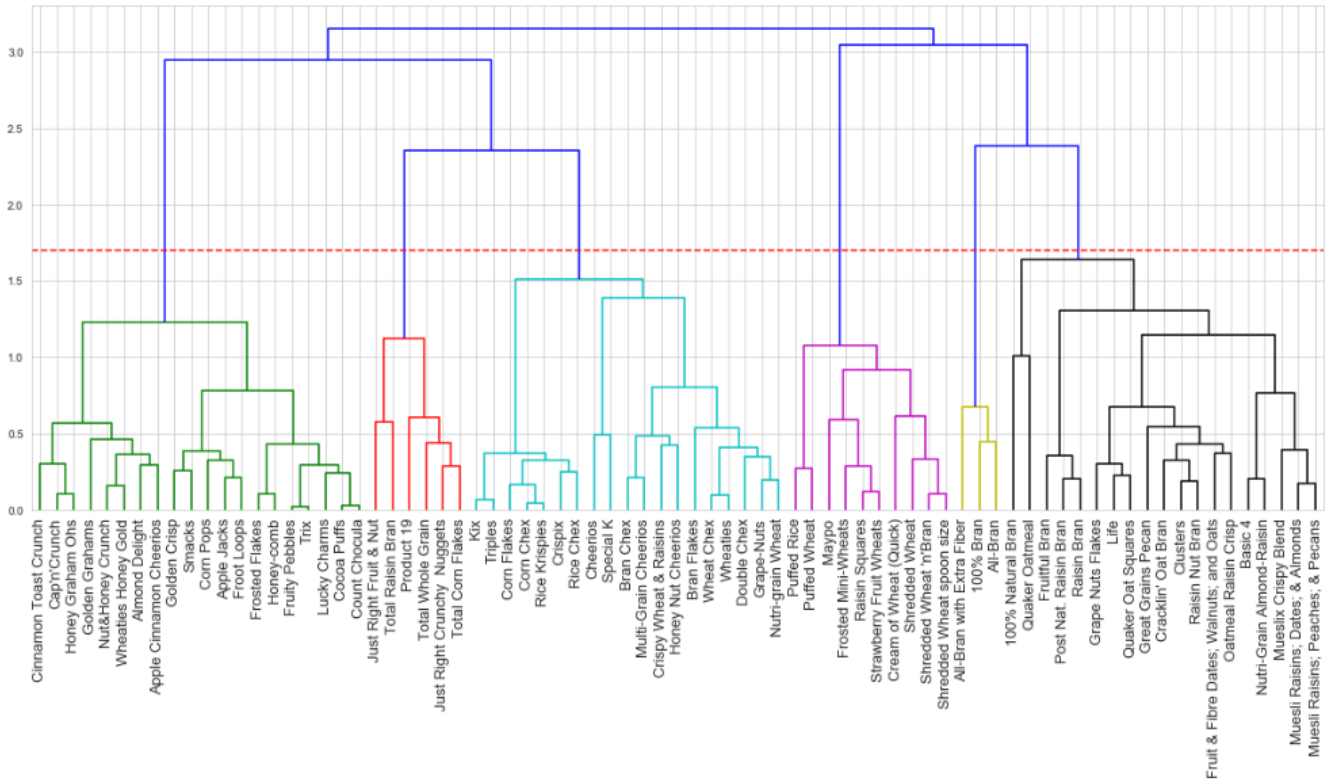

Figure 6: Dendrogram with Ward's method

on the method or linkage chosen, building up the hierarchy to the top. To find the optimal number of clusters, the distance on the y axis representing dissimilarity is used. One should draw a line through the dendrogram where the line can move up and down the longest distance before meeting a joining point. The number of lines this cut goes through represents the optimum number of clusters. For the cereal data, the red line is drawn through six lines, resulting in six

clusters as depicted in the different colors in figure 6. A breakdown of which cereals were clustered together can be seen in appendix A.

Looking at the hierarchy, 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 figure 6, 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 this building of the hierarchy. Their features have similar traits or within-cluster variances.

K-Means Clustering

The second method presented is K-means clustering. K-means clustering clusters the data around a centered point, or centroids. Starting off with the pre-determined k-clusters, k centroids are randomly chosen. The data points are assigned to each cluster by its closest distance. The randomly chosen centroids are then moved again, and after several iterations of minimizing the Euclidian distances from the points to the centroids, the final centroids are determined with the best clusters. Ultimately, k-means clustering is maximizing the intra-cluster cohesion and the inter-cluster separation (Bonaccorso, 2020). To determine the optimal number of clusters for the cereal data set, 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 (Bonaccorso, 2020). As seen in figure 7, the elbow test produces a graph that has a slight bend at six clusters. This is the number of clusters that is used in the K-means clustering analysis.

With six chosen for k-clusters, the normalized data was put through the K-means clustering algorithm. Figure 8 reflects this in a 3D feature space, with the first three features, calories, protein and fat. The stars represent the centroids and the six colors represent the six

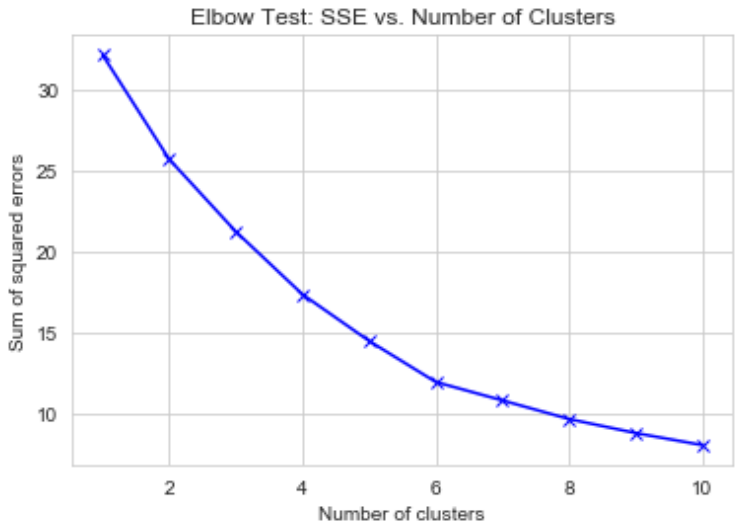

Figure 7: Elbow Test

different clusters. Looking at the graph, it was challenging to see the different clusters do to overlapping colors. From here, a test was done to see how the clusters would turn out with dimensionality reduction. Principal component analysis, or PCA, is a dimensionality reduction technique for unsupervised learning methods. It uses linear transformation to reduce dimensions

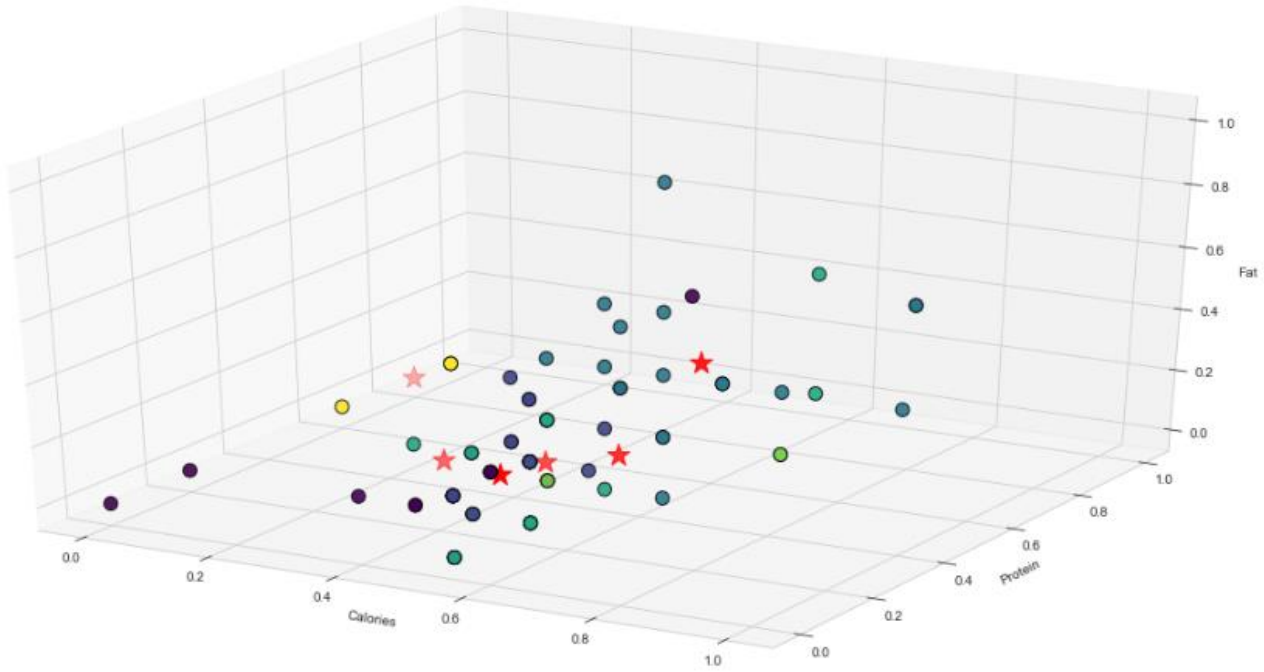

Figure 8: K-means clustering with 6 clusters

(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 figure 9. Although the graph looks completely different between the two, after analyzing the cereals in tabular format, 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.

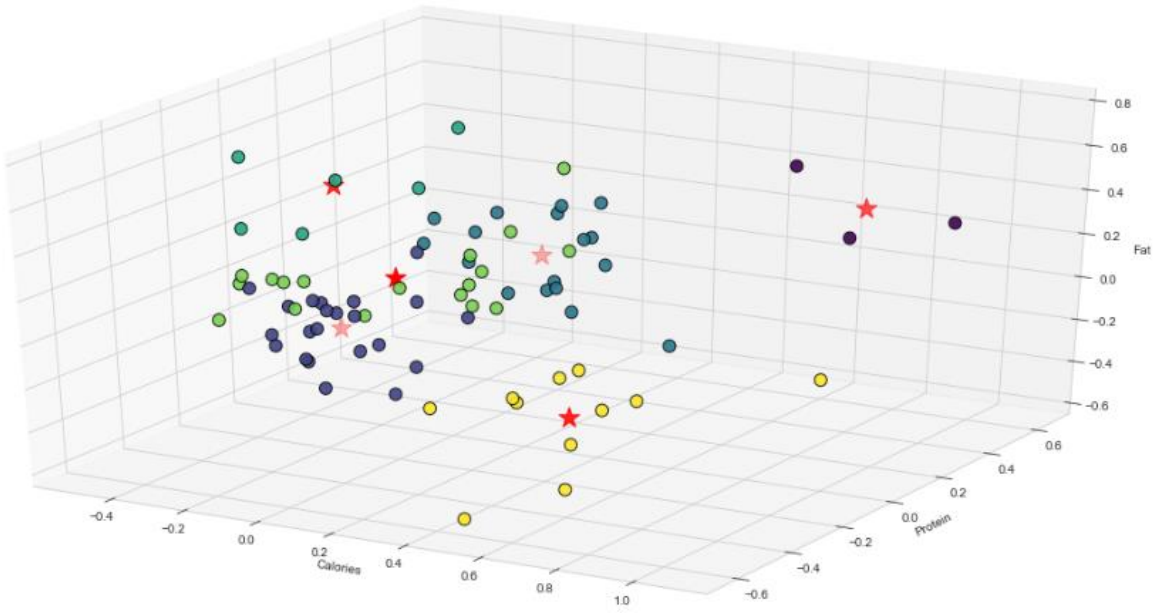

Figure 9: K-means clustering using PCA dimensionality reduction

Evaluation

Comparing the two clustering methods, hierarchical with agglomerative Ward's method and K-means, both clustering techniques came close to clustering the cereals in the same six groups. Figure 10 depicts the two methods, with the slight variation in clustering represented by the colors. There were four cereals that were clustered differently. Cluster B had two cereals in hierarchical that were in cluster A in K-means. Cluster C had two cereals in hierarchical that were present in cluster B and D in K-means. Clusters E and F were the same across both methods. Further details of the clustering differences by cereal names are in Appendix A and B.

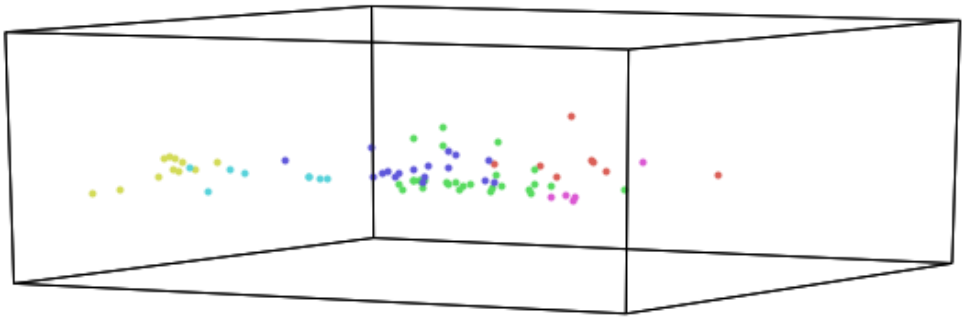

Figure 10 (a): Hierarchical clustering

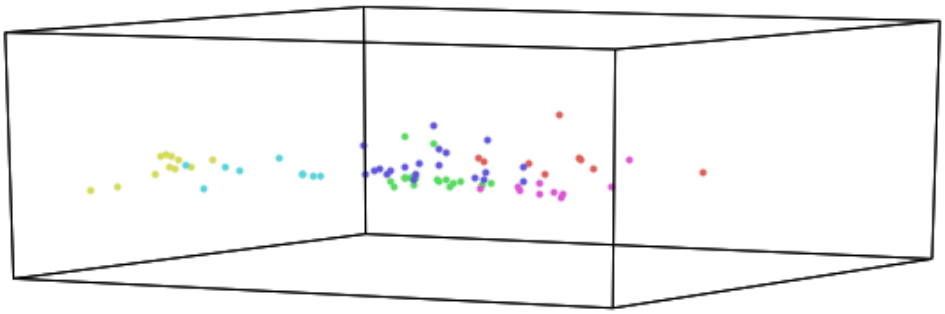

Figure 10 (b): K-means clustering

Accuracy

The goal of un-supervised clustering is to differentiate the points as best as possible. To evaluate the clusters and test the accuracy of the K-means method, the silhouette score was conducted. The silhouette coefficient's plot shows the "closeness of each point in a particular cluster has with respect to the other points in the neighboring clusters" (Ahmad, 2020). Scores can be from 0 to 1, and are segmented into four ranges for accuracy: excellent, reasonable, weak, and no clustering has been found (Ahmad, 2020). Looking at the silhouette score for the cereal data set, the closest number to 1 is the optimum number of clusters. As seen in figure 11, six clusters show the highest score closest to 1, which was the number of clusters used from the elbow test. However, the score for six clusters is close to 0.32. According to Ahmad (2020), this

falls within the weak category (range of 0.26-0.50), meaning the quality of clusters are not very reliable.

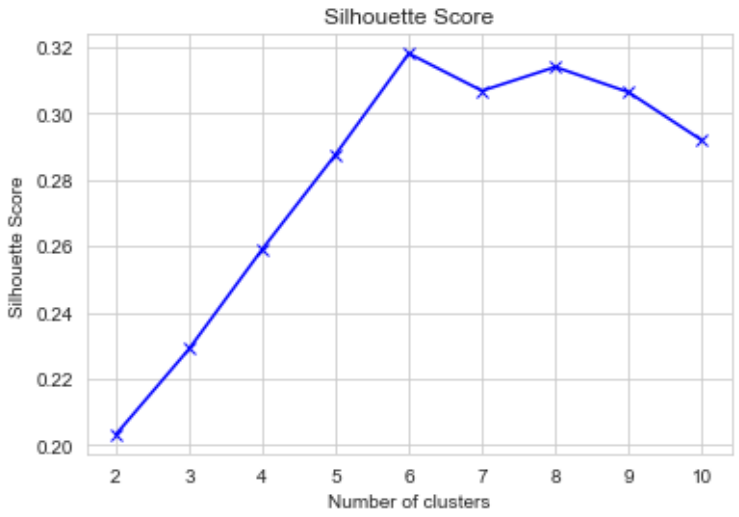

Figure 11: Silhouette Score

Results and Insights

To have insight to reasons why certain cereals were placed in their assigned cluster, histograms of each K-means cluster and attributes were conducted to see what variables were frequent. This gave great insight for why these cereals are similar. Cluster A had no or very low fiber and protein, was high in sugar, and calories frequented around 100-110. Cereal names for cluster A had Froot Loops, Fruity Pebbles, Lucky Charms, Trix, which are cereals known to be favorited by kids and are considered to be unhealthy. Cluster B had many cereals around 100 calories, low in fiber, but had more sodium. Cluster C had 3 grams of protein, and was higher on fat and calories. Cluster D had no or low sodium, sugar, and fat, and was low on fiber. These cereals had common names such as 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 E had low calories, 4 grams of protein, and high potassium and fiber, which also correlated together. Cluster E had all cereals with the word "bran" in them.

Lastly, Cluster F had low fat, moderate sugars, and high sodium. Figure 12 shows histograms for calories, sugars, sodium, and fiber.

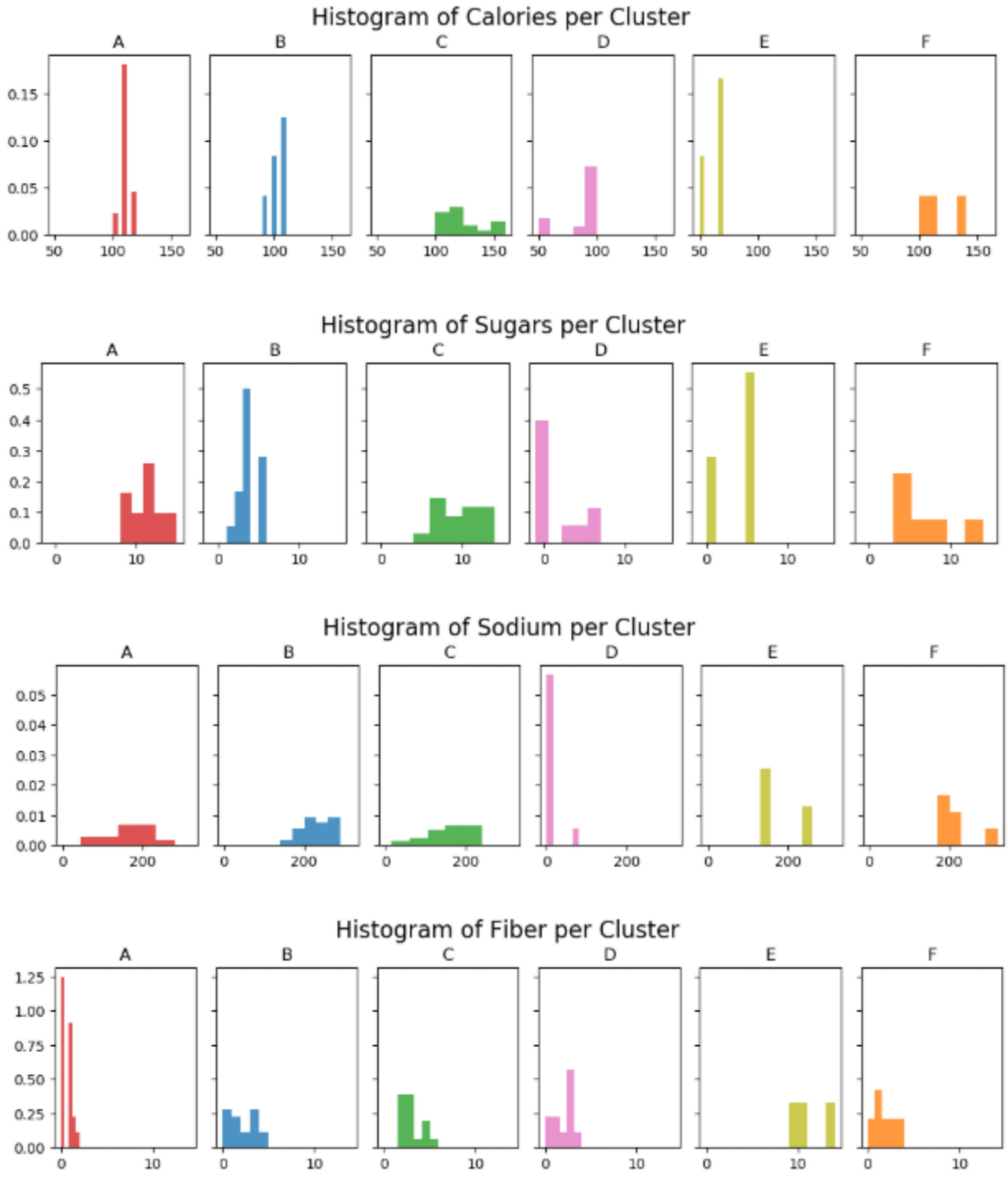

Figure 12: Histograms by Cluster for Calories, Sugars, Sodium, and Fiber

Secondly, it was also interesting to compare the final clusters to the in-store shelf placement of the cereals. The shelf placement was excluded from the algorithms, in order to cluster solely on dietary means. In the data, there are three shelves, with shelf 1 being on the bottom and shelf 3 on the top row. The frequency of shelf placement per cluster is presented in figure 13. Clusters E and F with the majority of cluster C are located on row 3, or the top shelf. These cereals have common words in their names, such as “bran,” “raisin,” “grain,” “nuts,” and specific nuts like walnuts, pecans, and almonds. These cereals have healthier attributes and are on the top shelf, somewhat out of sight. Cluster B has cereals that also have the words “nuts” and “grains” in them, as well as “chex,” and “rice.” Cluster B spans the 1st and 3rd shelf. Clusters A and D span all three shelves, with A being prominent on the middle shelf. As mentioned earlier, these are the unhealthier cereals targeted for kids. By their product placement, they are in a child’s eyesight level.

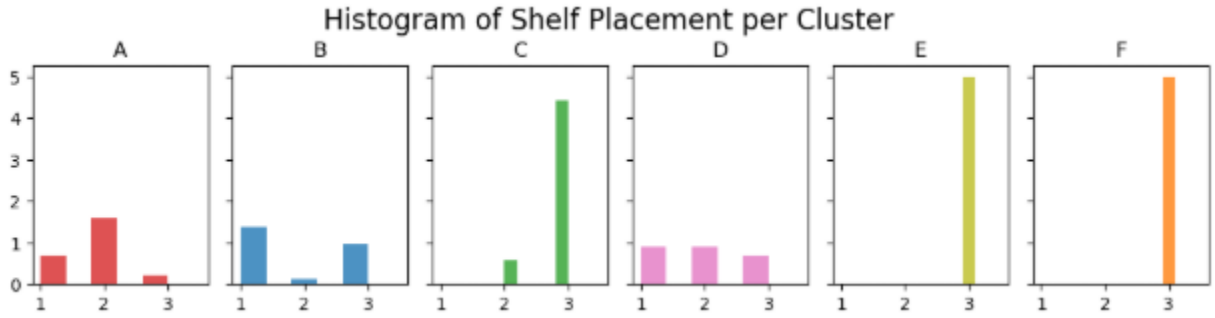

Figure 13: Histogram of shelf placement per cluster

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 like graph, adding to each grouping as the tree grew. 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. 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 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.

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		

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.

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		

Strawberry Fruit Wheats		

Resources

ACalorieCounter. (2019). Breakfast Cereal Compared: Cereals from Post, Kellogg's & General Mills. Retrieved from http://www.acaloriecounter.com/breakfast-cereal.php

Ahmad, Imran. (2020). *40 Algorithms Every Programmer Should Know*. Packt Publishing.

Alto, Valentina. (2019). Unsupervised Learning: K-means vs Hierarchical Clustering. Retrieved from https://towardsdatascience.com/unsupervised-learning-k-means-vs-hierarchical-clustering-5fe2da7c9554

Bonaccorso, Giuseppe. (2020). *Mastering Machine Learning Algorithms* (2nd ed.). Packt Publishing.

Crawford, Chris. (2018). 80 Cereals. Retrieved from https://www.kaggle.com/crawford/80-cereals?select=cereal.csv

Kellogg's. (n.d.). How does the Health Star Rating work? Retrieved from https://www.kelloggs.com.au/en_AU/health/howhealthstarswork.html

Kwan Peterson, Lucas. (2019). The official breakfast cereal power rankings: Part I. *Los Angeles Times*. Retrieved from https://www.latimes.com/food/la-official-breakfast-cereal-power-rankings-part-1-story.html

ODSC – Open Data Source. (2018). Three Popular Clustering Methods and When to Use Each. Retrieved from https://medium.com/predict/three-popular-clustering-methods-and-when-to-use-each-4227c80ba2b6

Thompson, Josh. (2019). Choosing the Right Clustering Algorithm for your Dataset. Retrieved from https://www.kdnuggets.com/2019/10/right-clustering-algorithm.html

Wikipedia. (n.d.). Ward's method. Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Ward%27s_method