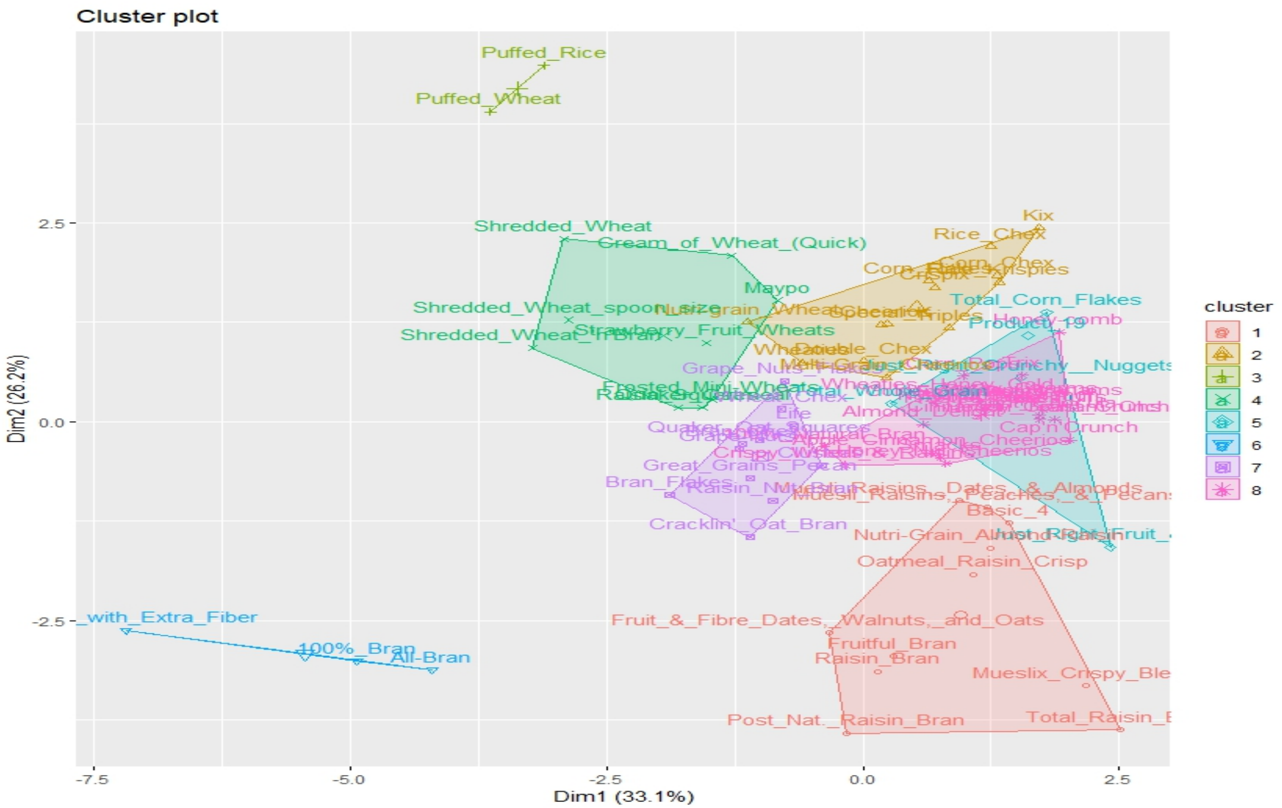
Clustering for Better Breakfasts: A Comparison of K-Means and Hierarchical Algorithms

AHMED GE22M009
KEMAL GE22M010
MUSTEFA GE22M014

Indian Institute of Technology Madras

MS4610: Introduction to Data Analytics Project
Date: 28-04-2023



INTRODUCTION

SIMILAR CEREALS

INTRODUCTION

Finding Similar Cereals

- Breakfast most important meal of the day
- Cereal is widely consumed, but not all are created equal
- What cereals have similar dietary features?
- Knowing similar cereals, we can supplement one for another
- Project presents clustering cereals to find similarities

METHODS

FOR CLUSTERING

METHODS

Four methods looked at for unsupervised clustering

- Expectation-Maximization
- Density
- Hierarchical
- K-Means

METHODS

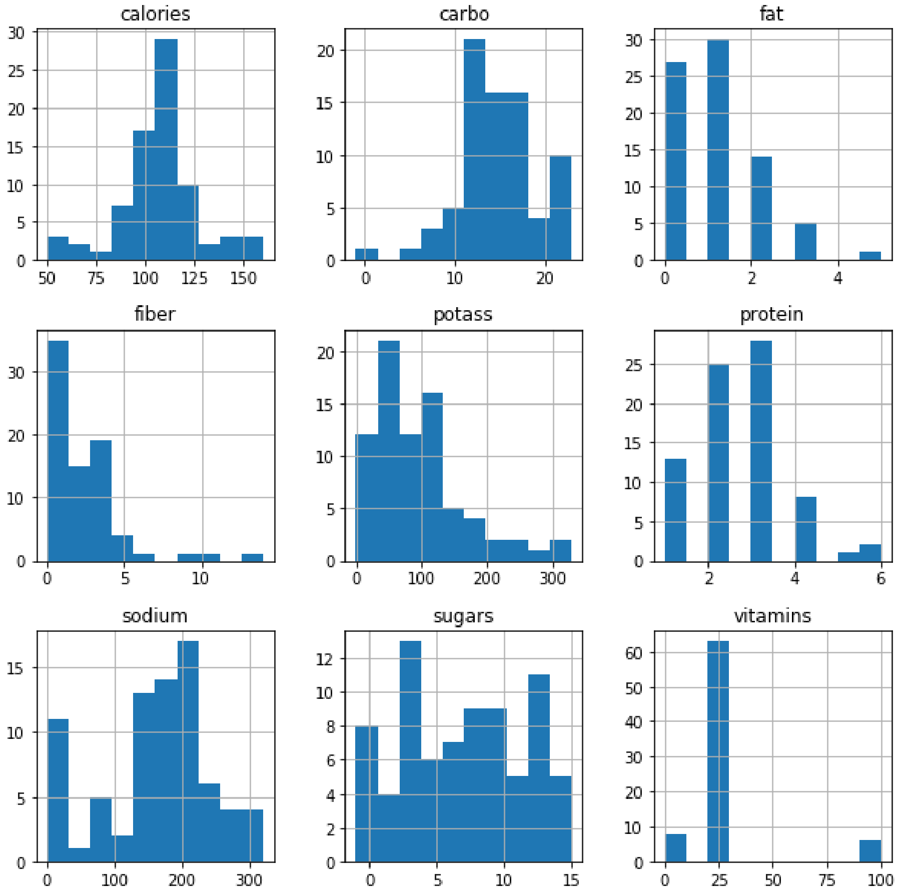
Expectation-Maximization

- Probability that all points belong to a cluster
- Assumes a normal distribution
- Not all variables met this criteria through their histograms

Density

- Looks at only points that are densely close
- DBSCAN was tested with different parameters and produced one cluster each time

Histograms of Each Variable



METHODS

Hierarchical

- Agglomerative hierarchical clustering builds from the ground up
- Points are joined based on distance measure to create a tree like structure

K-Means

- Points are centered around centroids
- Points are attracted to it's centroids based on a distance measure

Tools	
Scikit-Learn	Agglomerative Clustering KMeans Silhouette_score
SciPy	Dendrogram
Matplotlib	Scatterplot
Hypertools	Scatterplot

DATA

CEREALS

DATA OVERVIEW

Insights

- 77 Cereals
- 16 Features
- Hot and Cold

Dietary features used for clustering

- Calories
- Protein
- Fat
- Sodium
- Fiber
- Carbohydrates
- Sugars
- Potassium
- Vitamins

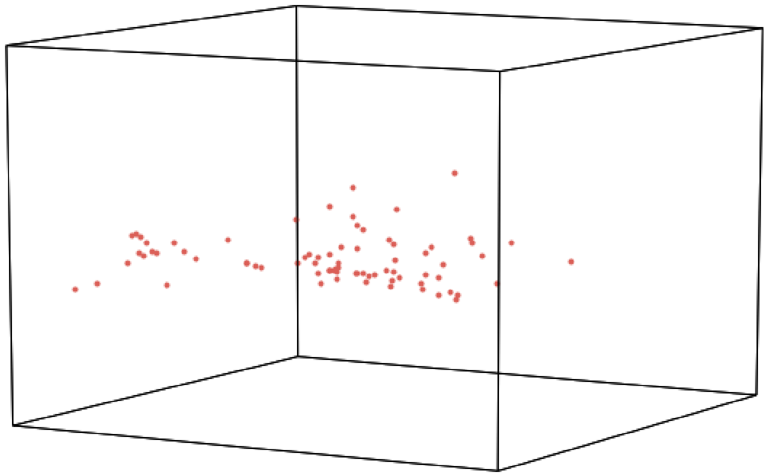
DATA PREPARATION

Normalization

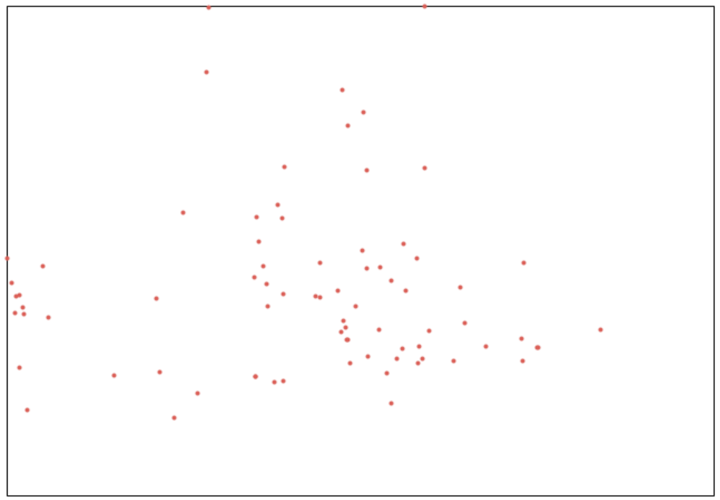
- Features used in clustering methods were normalized for the distances between each feature to be on the same scale
- Prevents skewness

DATA EXPLORATION

Scatter Plot Cube



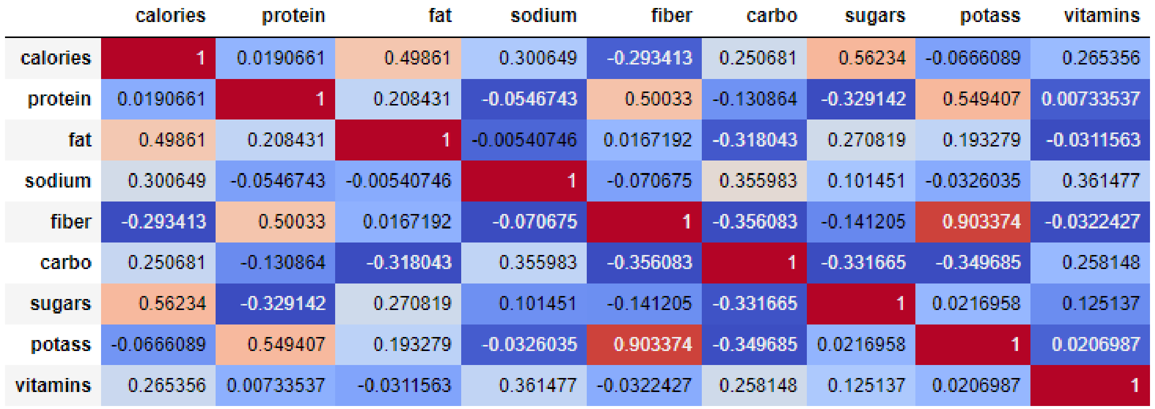
Scatter Plot 2D



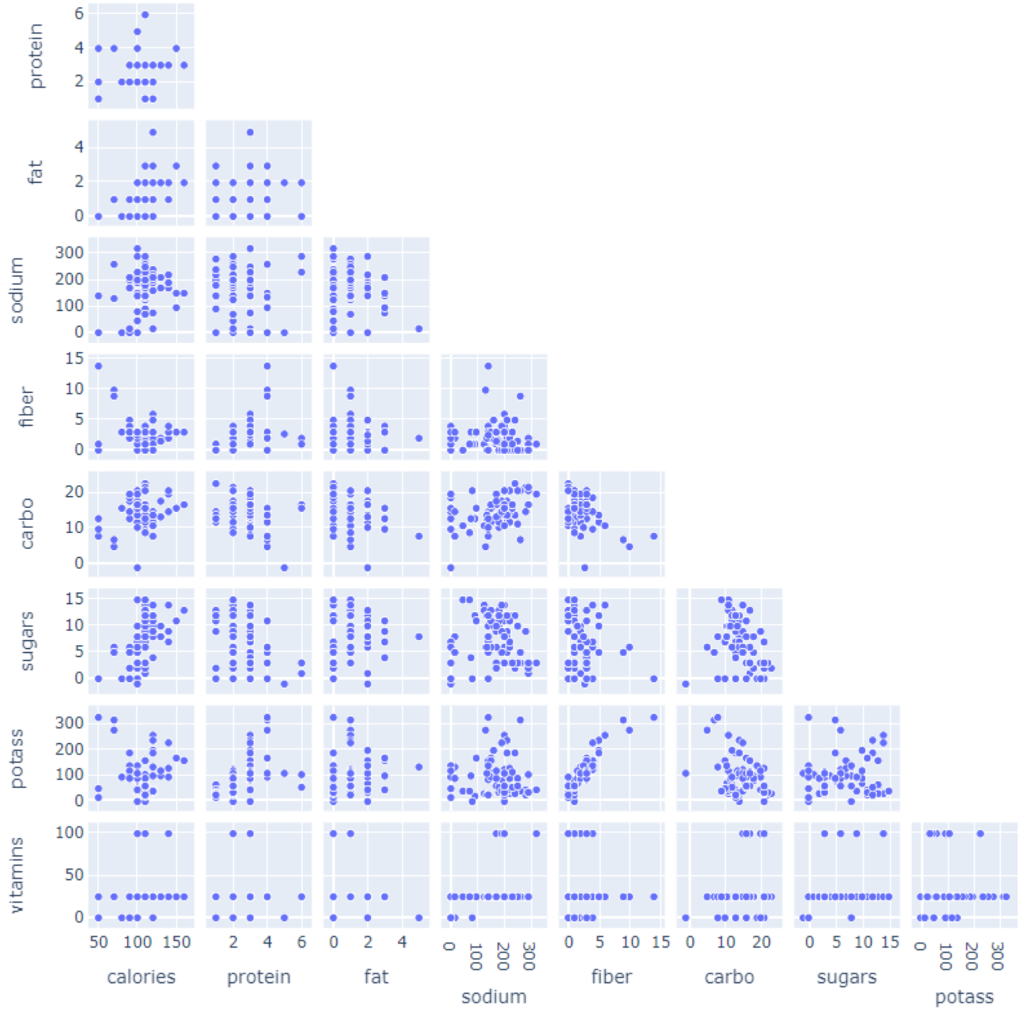
DATA EXPLORATION

Correlation

- Highly correlated:
  - Potassium/Fiber
- Slightly correlated:
  - Fiber/Protein
  - Sugar/Calories
  - Potassium/Protein



	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



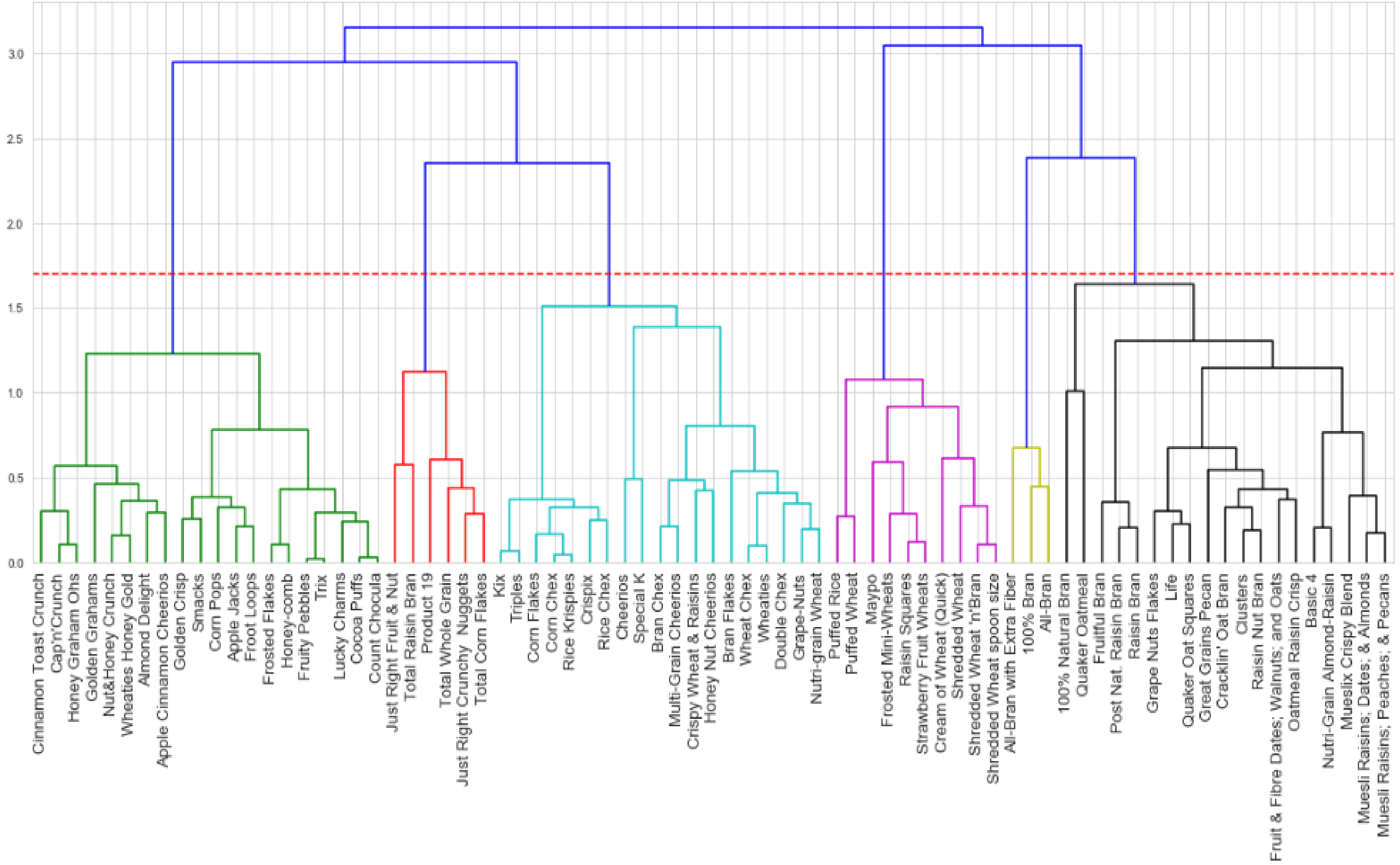
CLUSTERING

EXPERIMENTS

CLUSTERING: HIERARCHICAL

Hierarchical Clustering

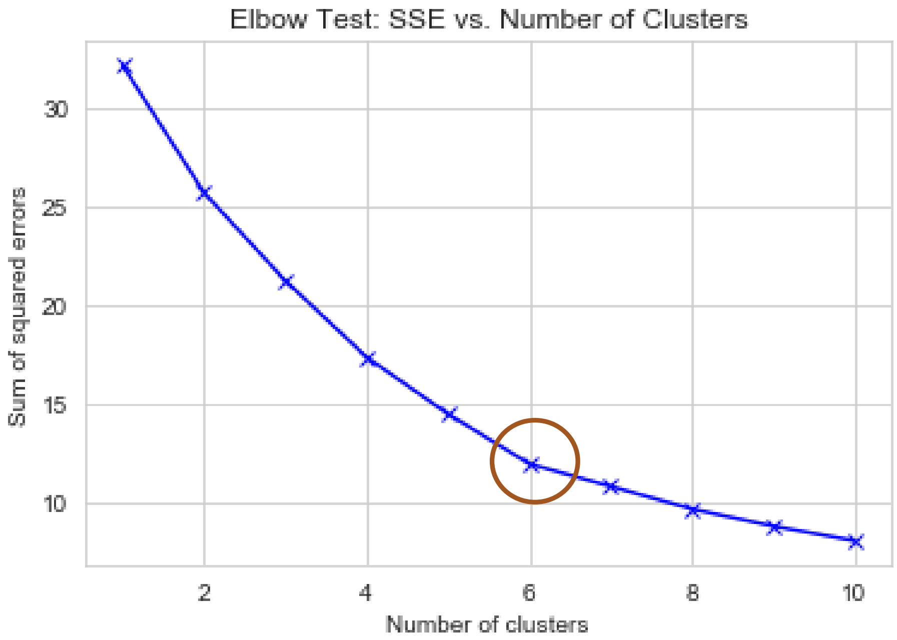
- Agglomerative
- Linkage: Ward’s Method
  - Minimizes total within-cluster variance
- Dendrogram
  - 6 clusters
- Trends:
  - Cereals clustered with similar names



CLUSTERING: K-MEANS

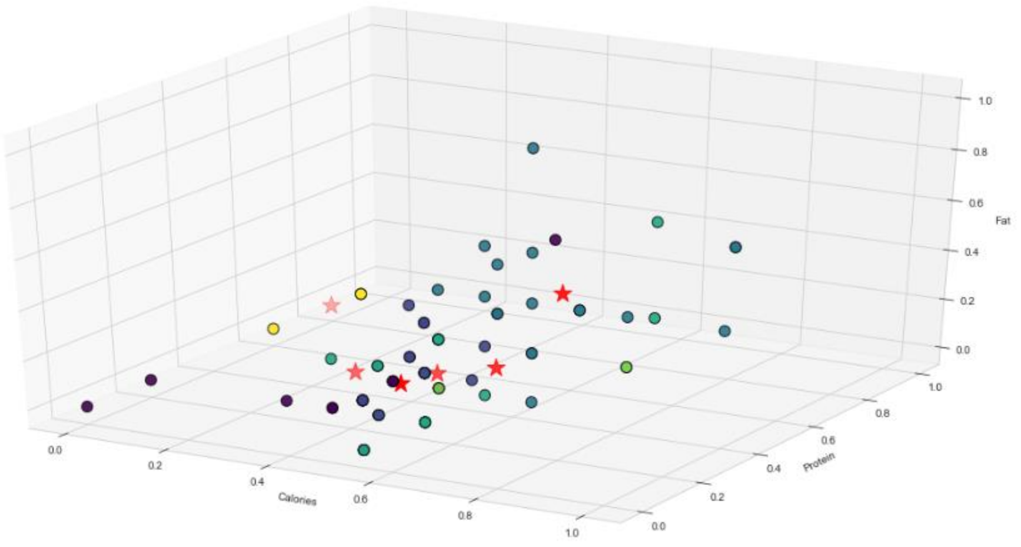
Elbow Test

- Inertia
  - Average distance between samples and centroid
- Bend in the curve
  - 6 clusters



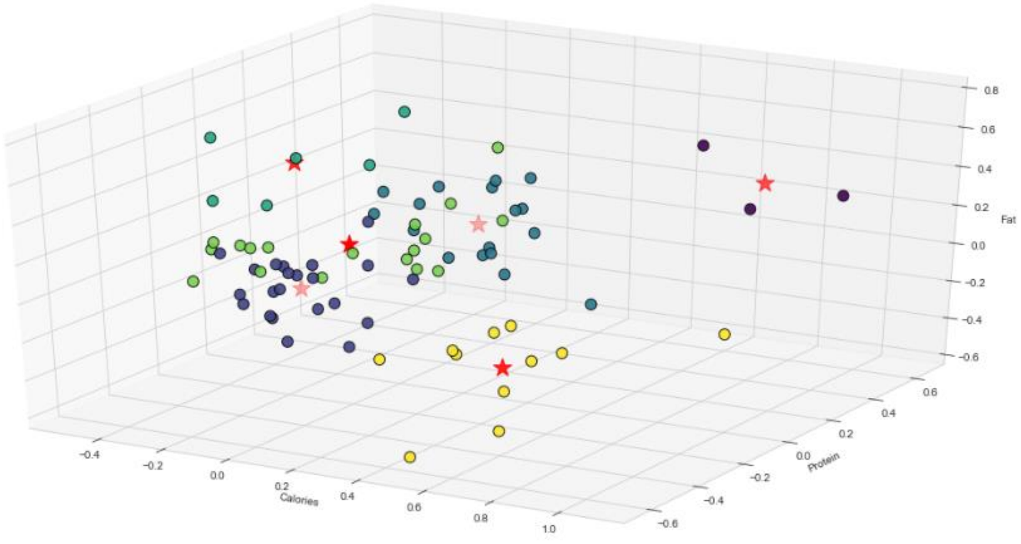
CLUSTERING: K-MEANS

K-Means Clusters



- Euclidean Distance
- Challenging to see distinct clusters

K-Means Clusters with PCA



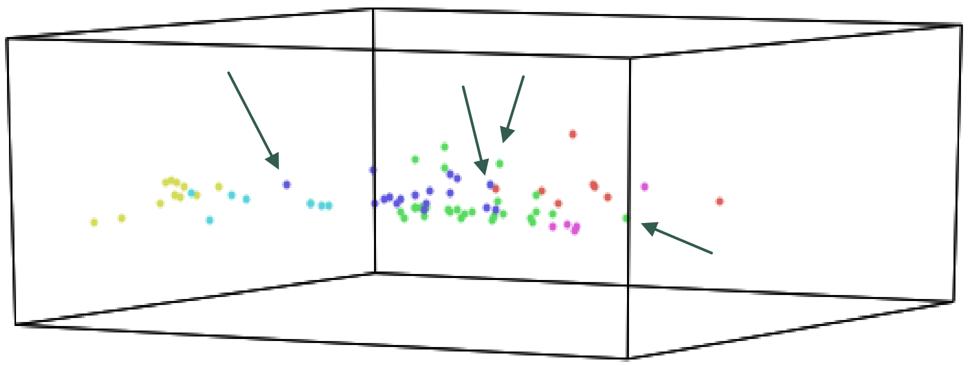
- Distinct clusters graphed with dimensionality reduction
- Resulted in cereals put in the same clusters

EVALUATION

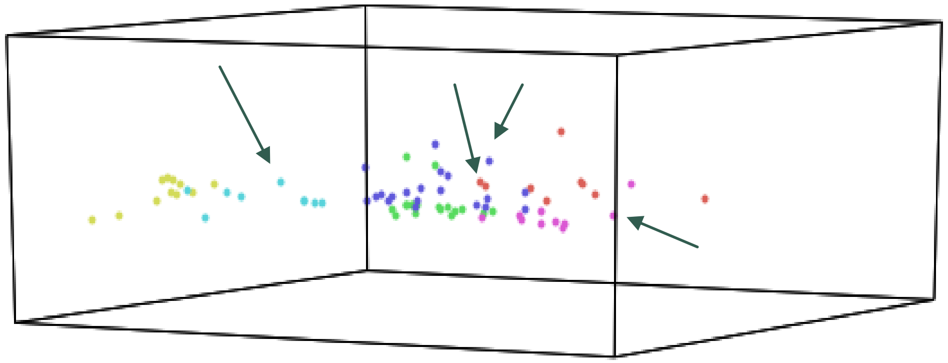
CLUSTERING

EVALUATION: COMPARISON

Hierarchical



K-Means



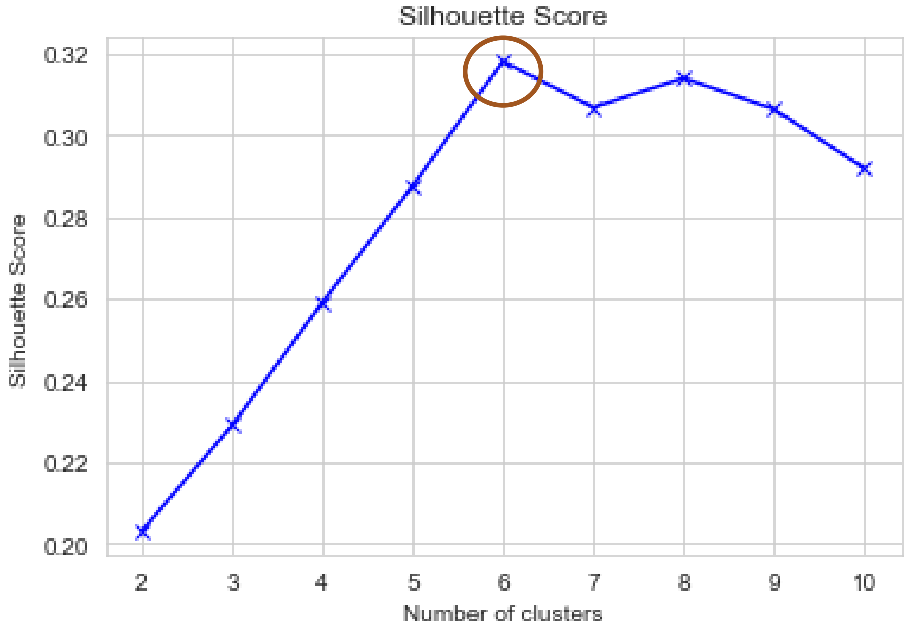
- Both methods clustered very similarly

- There were 4 cereals that were clustered differently

EVALUATION: RELIABILITY

Silhouette Score

- Tests accuracy of K-Means
- Plot shows closeness of points in each cluster in relation to other points in neighbor clusters
- Highest point is how many clusters
- Scale of 0-1 with four categories
- Score 0.32 → weak category
  - Clusters not very reliable



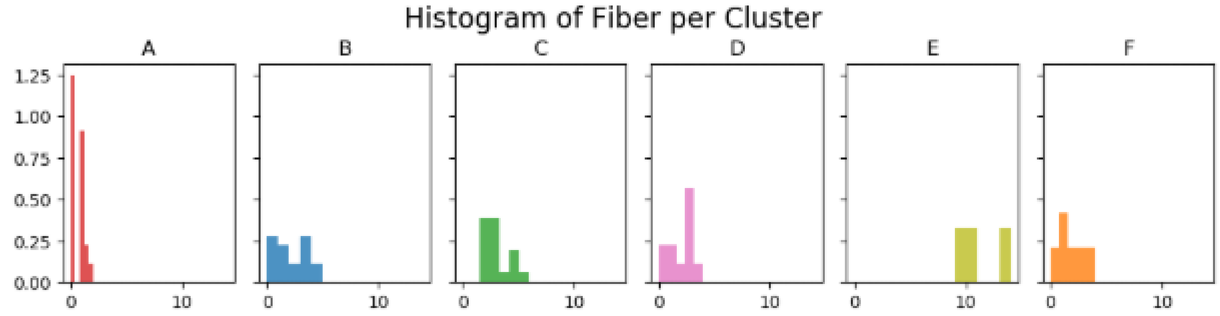
RESULTS & INSIGHTS

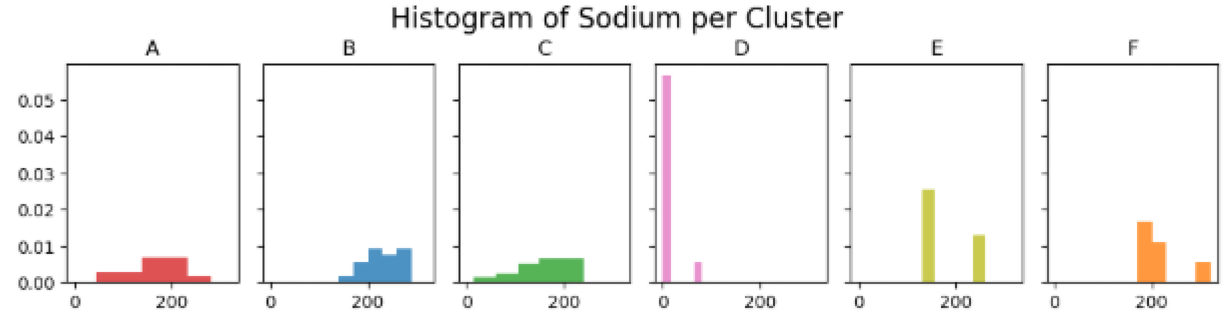
K-MEANS CLUSTERING

RESULTS & INSIGHTS

Histograms

- Histograms of each cluster’s features
- Fiber
  - Cluster A – least
  - Cluster E – most
- Sodium
  - Cluster D – least



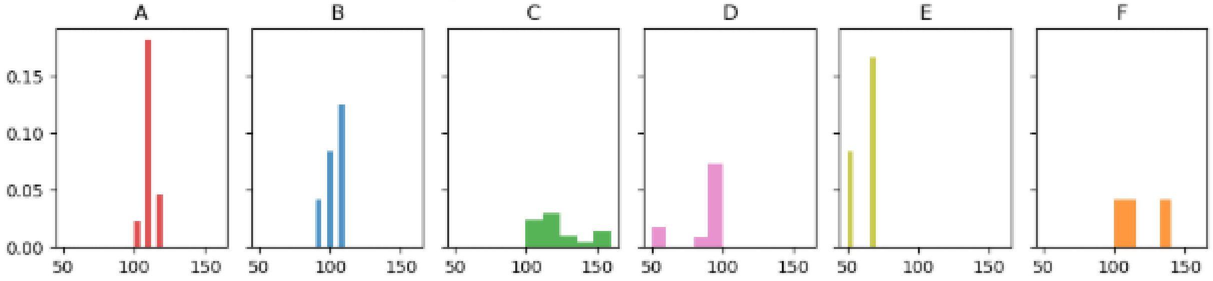


RESULTS & INSIGHTS

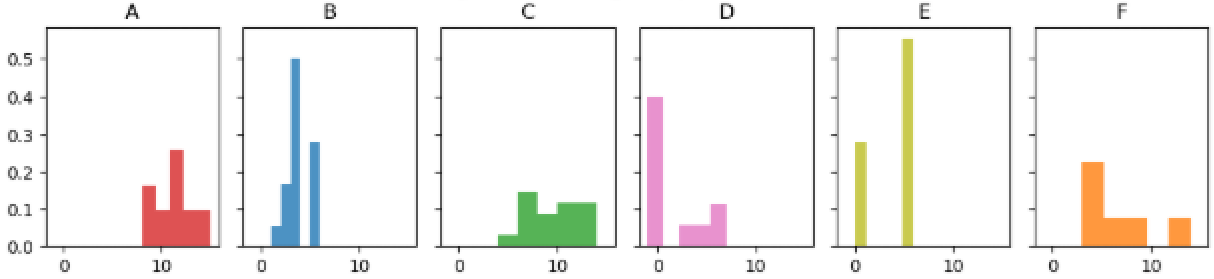
Histograms

- Calories
  - Cluster D & E – least
  - Cluster C & F – more on higher end
- Sugars
  - Cluster B, D, E – least
  - Cluster A, C, F - most

Histogram of Calories per Cluster



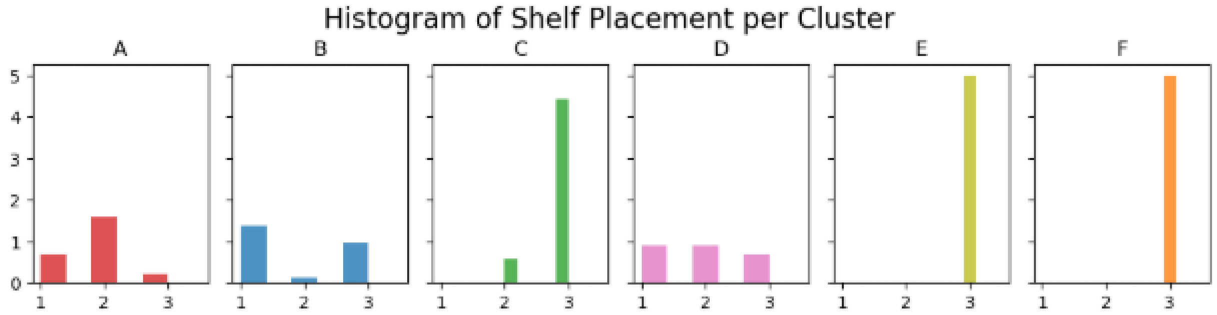
Histogram of Sugars per Cluster



RESULTS & INSIGHTS

Marketing

- Clusters C, E, F are located on the top shelf
  - ‘All-Bran’
  - ‘Life’
  - ‘Raisin Bran’
  - ‘Total Whole Grain’
- Cluster A is most dominant of middle shelf
  - ‘Trix’
  - ‘Froot Loops’
  - ‘Cocoa Puffs’
  - ‘Fruity Pebbles’



RESULTS & INSIGHTS

Cluster A	Cluster B	Cluster C	Cluster D	Cluster E	Cluster F
Almond Delight	Bran Chex	100% Natural Bran	Cream of Wheat (Quick)	100% Bran	Just Right Crunchy Nuggets
Apple Cinnamon Cheerios	Bran Flakes	Basic 4	Frosted Mini-Wheats	All-Bran	Just Right Fruit & Nut
Apple Jacks	Cheerios	Clusters	Maypo	All-Bran with Extra Fiber	Product 19
Cap'n'Crunch	Corn Chex	Cracklin' Oat Bran	Puffed Rice		Total Corn Flakes
Cinnamon Toast Crunch	Corn Flakes	Fruit & Fibre Dates; Walnuts; and Oats	Puffed Wheat		Total Raisin Bran
Cocoa Puffs	Crispix	Fruitful Bran	Quaker Oatmeal		Total Whole Grain
Corn Pops	Double Chex	Great Grains Pecan	Raisin Squares		
Count Chocula	Grape Nuts Flakes	Life	Shredded Wheat		
Crispy Wheat & Raisins	Grape-Nuts	Muesli Raisins; Dates; & Almonds	Shredded Wheat 'n'Bran		
Froot Loops	Kix	Muesli Raisins; Peaches; & Pecans	Shredded Wheat spoon size		
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					

CONCLUSION

CEREAL CLUSTERS

CONCLUSION

Similar Cereals

- Although the silhouette test scored low on reliable clusters, we were able to find similar cereals and understand where their dietary features lie
- With the clustered cereals, the consumer can see what cereals have healthier attributes
- Within these clusters, it is then up to the consumer to make their decision based on taste preferences while keeping in the same cluster

REFERENCE:

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.