2020 HAWAII UNIVERSITY INTERNATIONAL CONFERENCES
ARTS, HUMANITIES, SOCIAL SCIENCES, & EDUCATION JANUARY 6 - 8, 2020
HAWAII PRINCE HOTEL WAIKIKI, HONOLULU, HAWAII

STATISTICAL MODELING ANALYSIS OF CHOCOLATE NUTRITION SCIENCE AND CHEMISTRY

CHEN, CHARLES
MORRILL LEARNING CENTER
SANTA CLARA, CALIFORNIA

CHEN, MASON
STANFORD UNIVERSITY ONLINE HIGH SCHOOL
PALO ALTO, CALIFORNIA

Dr. Charles Chen
Morrill Learning Center
Santa Clara, California

Mr. Mason Chen
Stanford University Online High
School Palo Alto, California

Statistical Modeling Analysis of Chocolate Nutrition Science and Chemistry

Abstract

This paper adopts STEAM (Science, Technology, Engineering, Artificial Intelligence, Math) approach. The objectives of this paper are to use Multivariate Clustering Statistics to study the Chocolate Science and Products. Chocolate contains flavonoids and antioxidants which can prevent aging and beneficial to heart disease and diabetes patients. Antioxidants can prevent heart disease is because it reduces free radical formation. Data has been collected on 20+ chocolate ingredient nutrition contents from 60+ different types of chocolate. Both Clustering Variables and Principle Component Analysis methods are utilized to cluster (1) chocolate nutrition, (2) chocolate product types. Chocolate nutrition are clustered into four clusters which is consistent with Chocolate science research and can explain the common chocolate food science very well. Chocolate products can also be clustered into 4 clusters which can distinguish the major chocolate types (dark, milk, white). Five clustering distance algorithms are studied and compared based on the impact of clustering sequence and patterns. Number of clusters are also studied in order to determine which clustering distance algorithm can provide the best clustering pattern to explore the chocolate science research. This paper has demonstrated the effectiveness and power of adopting STEAM approach on the general Scientific Research.

Keywords: STEAM, Multivariate Statistics, Chocolate Science, Clustering, Principle Component Analysis, JMP

1. Introduction

Many people like eating chocolate but have concerns that chocolate is unhealthy. The objectives of this paper are (1) understand chocolate science and chemistry, (2) cluster chocolate nutrition, and (3) cluster chocolate products.

1.1 Adopt "STEAM" Approach

STEAM (Science, Technology, Engineering, Artificial Intelligence, Mathematics) methodology was applied on this project to help define the project scope. The chemistry "Science" studied was cocoa bean nutrition, flavonoids, flavanols, and antioxidants. "Technology" is the manufacturing process to produce the commercial chocolate products from coca beans. Systematic "Engineering" problem solving techniques such as 5 whys and SIPOCs were deployed to understand the root cause analysis. "Artificial Intelligence" algorithms such as clustering and principle component analysis are utilized to recognize the patterns hidden among Chocolate nutritions and products. "Math (Statistics)", Graphical Analytics are conducted to demonstrate the chocolate science and explore the product complexity. All 5 STEAM elements are critical to making this project successful.

1.2 Understand Chocolate Anti-Oxidant Science

Chocolate contains flavonoids and antioxidants which can prevent aging and beneficial to heart disease and diabetes patients [1-5]. Chocolate is a powerful source of antioxidant which prevents human body aging/heart disease since it increases blood flow. Apple and blueberry are well known fruits with rich amounts of antioxidant. In Figure 1, if chocolate's serving size is equal to apple, it contains the most antioxidant amount among all the foods listed below. Why can antioxidants prevent against heart disease?

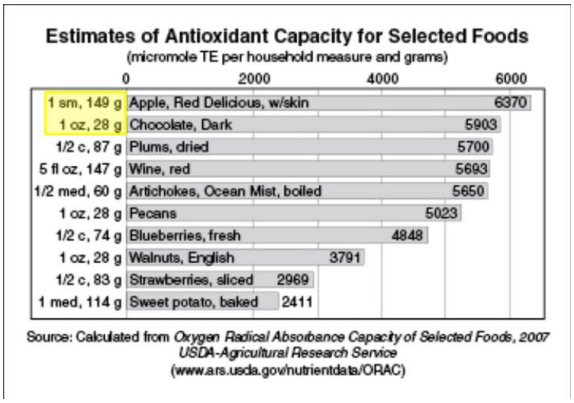

Figure 1: Estimates of Antioxidant Capacity

In Figure 2, based on research [6-11], cardiovascular heart disease (CHD) risk is lower if taking more than 3 Chocolate servings per week (1 serving = 30 g). A typical cardiovascular disease is Atrial Fibrillation (AF).

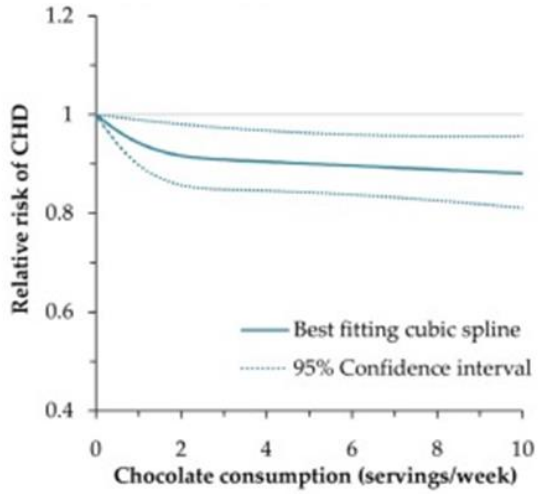

Figure 2: Relative Risk of CHD

Antioxidants can prevent heart disease is because it reduces free radical formation. Free radicals are atoms with an odd number of electrons as shown in Figure 3. When radicals form, they become highly reactive which causes cells to function poorly or die [12]. Excess free radicals initiate Cardiovascular disease **(**CVD**)** by damaging blood vessel. Bad cholesterol, Low-Density Lipoproteins (LDL), can also and only cause CVD after the oxidation of free radicals [13].

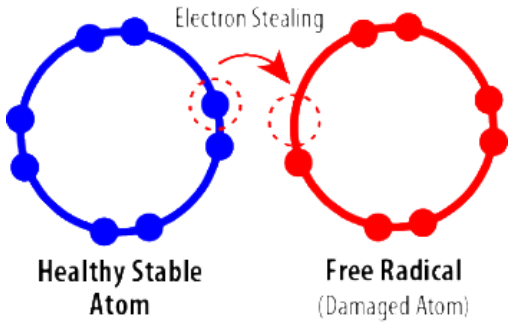

Figure 3: healthy stable atom and free radical

The oxidized components attract macrophages which absorb & deposit cholesterol [14] as show in Figure 4.

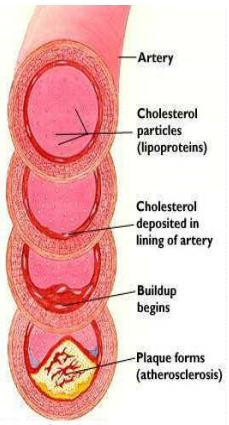

Figure 4: Diagram showing the effect of plaque buildup in arteries

1.2 Chocolate Product Research

There are three main commercial chocolate product types available in most stores: dark, milk, and white. The characteristics of dark chocolate are plenty amounts of soluble fiber, rich minerals (iron, magnesium, copper, manganese, potassium, phosphorus, zinc, selenium), powerful source of antioxidant, improve blood flow and lower blood pressure, increases High Density High-Density Lipoproteins (HDL, good cholesterol) and decreases Low-Density Lipoproteins (LDL, bad cholesterol), lower risk of cardiovascular disease, and improves brain function [15,16]. The side risks are that dark chocolate may cause migraines, kidney stones, and caffeine. The characteristics of milk chocolate are some of dark chocolate plus calcium, heart healthy, boosts brain functions, slows signs of aging, fights colds, stops tooth decay, lowers blood pressure, and reduces stress [17]. The main concerns are lots of sugar, and some side effects like caffeine [18,19]. The characteristics of white chocolate are rich calcium, prevents hypertension and heart failure, increases blood flow, maintains cholesterol level, and reduces breast cancer [20]. The problems are the enormous amounts of sugar, obesity, and diabetes [21]. Authors have previously done the multivariate statistics on understanding the Chocolate Science [22-24]. This paper will utilize more Advanced Clustering and Principle Component algorithms to further explore the chocolate science.

2. Clustering Methods

Section 2 will explain the different clustering methods and algorithms that were experimented to further study Chocolate Nutrition Science.

2.1 Clustering Chocolate Nutrition Variables

In order to analyze Chocolate Science and Nutrition pattern, the "Variable Clustering" method [25] is used for grouping similar nutrition variables into representative clusters which are a linear combination of all variables in the same cluster. The cluster can be represented most by the variables identified to be the most representative members (higher R-Square with own cluster in Figure 5). The most representative variable in the cluster can be used to explain most of the variation in the data analyzed. Typically, dimension reduction using Cluster Variables is often more interpretable than dimension reduction using principal components. Based on JMP Clustering Variable analysis, four clusters are identified as following:

Cluster 1: The higher the saturated fat, the higher the total fat, and the higher the calories. This cluster represents higher Fat/higher Calories in Chocolate products.

Cluster 2: Calcium and cocoa percent should have a negative correlation. This has indicated the main difference between Dark Chocolate (lower Calcium but higher cocoa percent) and Milk Chocolate (higher Calcium but lower cocoa percent).

Cluster 3: The higher the sugar, the higher the carbohydrates. Sugar and Carbohydrates are strongly correlated across most food products.

Cluster 4: Iron and Dietary have strong correlations due to higher Cocoa Percent (Citation).

Clustering Variables method can effectively explore the Chocolate nutrition clustering patterns which can explain the common foods science well. Adopting this dimension-reduction clustering algorithm can help simplify the predictive modeling by enhancing the signal-noise ratio, particularly in a very complicated/coupled design or system behavior.

Cluster Members

Cluster	Members	RSquare with Own Cluster	RSquare with Next Closest	1-RSquare Ratio
1	Calories (g)	0.789	0.314	0.308
1	Calories_from_Fat (g)	0.976	0.456	0.044
1	Total_Fat (g)	0.977	0.426	0.04
1	Saturated_Fat (g)	0.935	0.361	0.101
2	Cocoa_Percent	0.742	0.366	0.406
2	Cholesterol (mg)	0.811	0.387	0.309
2	Vitamin_A	0.505	0.126	0.566
2	Vitamin_C	0.412	0.016	0.598
2	Calcium	0.726	0.079	0.297
3	Sodium (mg)	0.345	0.013	0.664
3	Carbs (g)	0.876	0.185	0.152
3	Sugar (g)	0.874	0.416	0.216
4	Dietary_Fiber (g)	0.888	0.403	0.187
4	Protein (g)	0.73	0.358	0.421
4	Iron	0.803	0.269	0.269

Figure 5: Cluster Chocolate Nutrition Variables

2.2 Principle Component Analysis (PCA)

The Principle Component Analysis (PCA) method [26-28] was further conducted to be compared with the Section 2.1 Clustering Variable method. PCA method derives a small number of independent linear combinations (principal components) that capture as much of the variability in the original variables as possible. The Bi-plot in Figure 6 graphs the matrix between the Chocolate Nutrition variables and the first two principle components.

The relative distance of nutritions can indicate the affinity among them. For example, the Cocoa Percent, Iron and Dietary Fiber are located within the marked blue zone which is consistent with Section 2.1 Cluster 4. The Sugar and Calcium components are located almost in the opposite direction of Cocoa Percent which is consistent with Section 2.1 Cluster 2. The nutrition components located closer to the center (0,0) such as Vitamin C and Protein have less effect on representing Chocolate Science. PCA and Clustering Variables methods are complimentary to uncover any hidden scientific pattern.

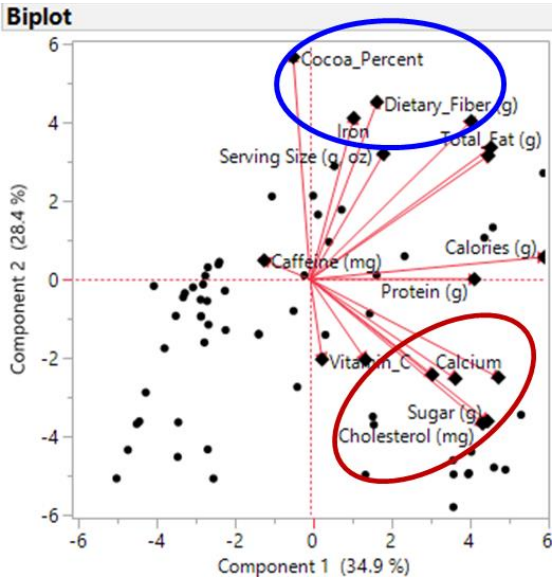

Figure 6: PCA Biplot of Chocolate Nutrition

2.3 Hierarchical Clustering and Dendrogram

Based on the results of Section 2.1 Clustering Variables and Section 2.2 Principle Component Analysis, 8 presentative Nutrition Components are selected for further pattern recognition analysis. Unlike previous two methods clustering on Chocolate Nutritions, Section 2.3 Hierarchical Clustering is a multivariate technique that groups Chocolate Products together that share similar values across a number of nutrition variables. Hierarchical clustering combines clusters successively. The method begins by treating each Chocolate Product as its own cluster. Then, at each step, the two Chocolate Products that are closest in terms of distance are combined into a single cluster. The result is depicted as a tree, called a Dendrogram shown in Figure 7.

The Dendrogram gives information about the degree of dissimilarity of clusters. Based on Cree Plot, three clusters (Blue, Red, Green) were identified among 23 commercial chocolate products.

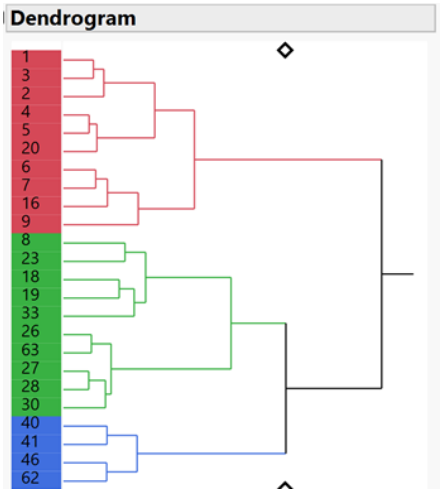

Figure 7: Hierarchical Clustering Dendrogram

Understanding the three-clustering patterns may further help study the Chocolate Product and Foods Science. Figure 8 plots the 8 critical Chocolate Nutritions parallel across three identified product clusters. Cluster 1 (red curve) seems to be the healthy one (higher Cocoa Percent/Fiber/Iron, lower Cholesterol/Sodium/Sugar/Calcium). Cluster 3 (blue curve) seems to be the unhealthy one (opposite trend against the 1st cluster).

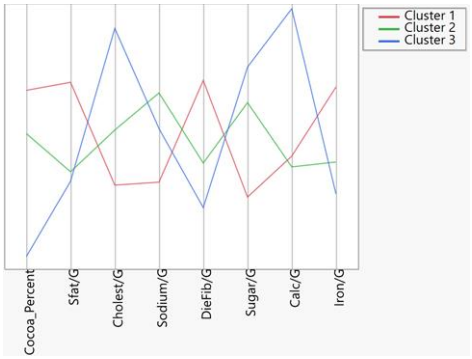

Figure 8: Parallel Plot of Nutrition Distributions

If these three clusters are indicating the product health, how these three clusters are separated? For each variable, gives the R-Square value that represents the proportion of variation explained by the clusters. This number is the R-Square value for a regression of the variable on the clusters. Figure 9, based on Hierarchical Clustering Algorithm, calculates the R-Square of each 8 critical nutrition component on the contribution amount of clustering pattern. It's not surprised that Cocoa percent and Calcium are top two deciding nutritions to separate Chocolate Products. This result is similar to the Section 2.1 Cluster 3 on strong negative correlations between these two nutrition factors. In addition to the grouping patterns, Section 2.3 Hierarchical Clustering method can further provide the relative ranking of how these critical nutrition variables could impact the chocolate product types (Dark, Milk, White).

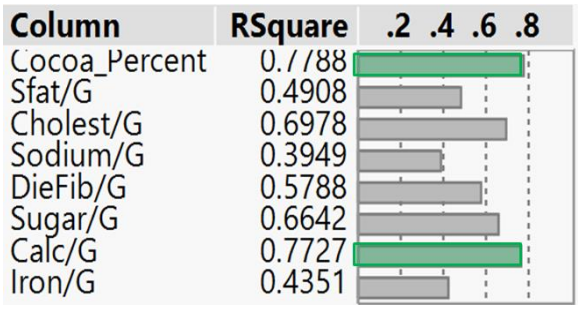

Figure 8: Hierarchical Clustering Ranking

3. Clustering Algorithms

Section 2.3 clustering method has identified three chocolate product groups separated by nutritions. The clustering patterns were identified based on clustering distance algorithm of calculating the dissimilarity of nutritions among chocolate products. Section 3 will further compare various clustering distance algorithms that were experimented to further study Chocolate Nutrition Science.

3.1 Clustering Distance Algorithms

There are several cluster algorithms: (1) Average, (2) Centroid, (3) Ward, (4) Single, and (5) Complete (Citation). Will these 5 different clustering algorithms have the same results? If different, how to select which algorithm to explore the clustering patterns best? We don't want take any risk as if these algorithms were applied to human life, especially for heart disease patient.

In Figure 9, three existing clusters (Green, Yellow, Red) are going to join next. Which two clusters should bond first? The joining sequence is determined by the clustering distance algorithms. Centroid, Single, and Complete algorithms are compared show in Figure 9. The Centroid algorithm connects Green cluster and Yellow cluster through the purple line connecting the two cluster means (purple triangles). The Single algorithm groups Green cluster and Red cluster by the closest points between these two clusters. The Complete algorithm groups Yellow cluster and Green cluster by the farthest points between these two clusters. Depending on which distance algorithm chosen, the clustering sequence and pattern may be different. We must dive into the mathematical calculations for each clustering distance algorithm and understand the benefits and limitations of each algorithm in order to choose the best algorithm to draw reliable clustering patterns and results.

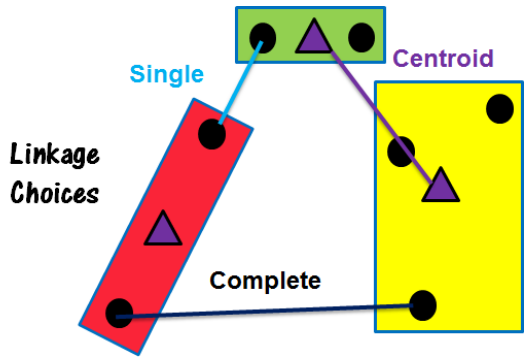

Figure 9: Diagram of the Centroid, Single, and Complete Clustering Methods

3.2 Clustering Algorithms

Section 3.2 will compare five major clustering distance algorithms [29-35]. The calculations of the five different clustering algorithms are shown in Figure 10. The first algorithm is Average which is the distance pair divided by the number of distances. Since the Average algorithm compares the average distances, it typically joins smaller and similar variances. The 2nd algorithm is centroid which calculates the distance between the cluster means. Among five algorithms mentioned, Centroid is the most robust algorithm to outliers. The 3rd algorithm Ward uses the ANOVA sum/mean of squares (between divided by within). The Ward algorithm is Centroid divided by the degree of freedom. Ward joins smaller numbers of observations and which is the most sensitive to outliers. The 4th Single uses the minimum distance, and therefore typically, joining larger variances/larger number. Clusters (favor in Single algorithm) are large in size, elongated or irregular. Those clusters may have shorter distances with other similar clusters than with small-sized clusters. The last algorithm Complete joins clusters based on the farthest distance. It is more sensitive to moderate outliers and, very different from single algorithm. Complete algorithm normally joins smaller variances/smaller numbers of clusters. How will these algorithms impact the clustering patterns?

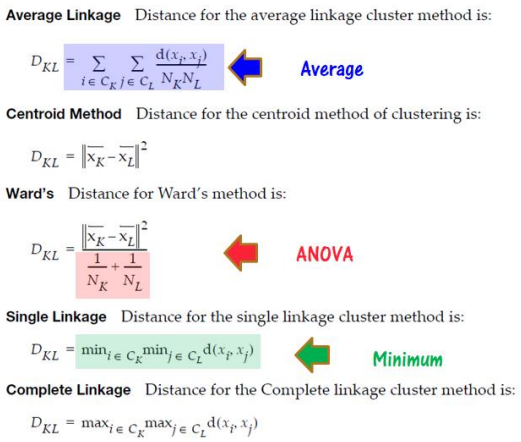

Figure 10: JMP Clustering Distance Algorithms [36]

In Figure 11, "Average" distance method would join smaller clusters while "Centroid" method is more robust to outliers. "Ward" method would also join smaller clusters though is very sensitive to outliers. "Single-Minimum" method will join larger and irregular/elongated clusters. "Complete-Maximum" method will join smaller clusters while moderately sensitive to outliers. It's critical to select the appropriate clustering distance methods in order to form the clusters which can effectively represent the cluster patterns.

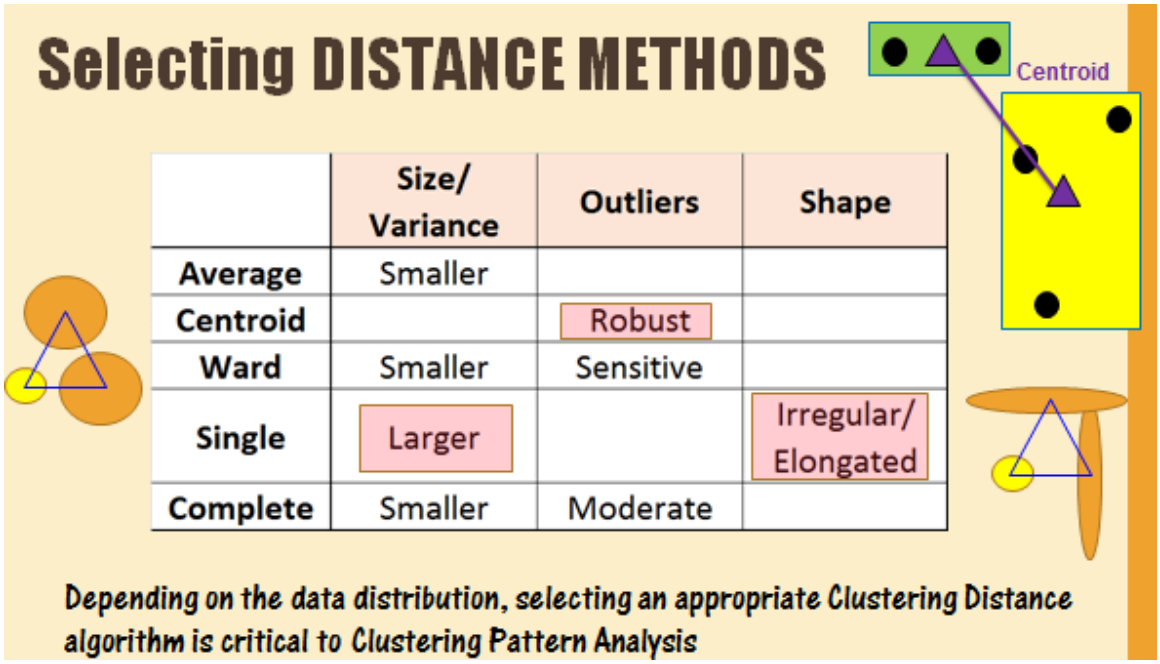

Figure 11: Selecting Clustering Methods

3.3 Constellation Plot

This constellation plots (shown in Figures 11) arrange the nutrition as endpoints and each cluster join as a new point. The lines represent membership in a cluster. The length of a line between cluster joins approximates the distance between the clusters that were joined. Using the constellation plot, it is possible to see which clusters are combined first for different clustering algorithms. Single algorithm had a different constellation plot than Ward, Average, Centroid, and Complete which had the same constellation plots. In Figure, Ward algorithm joins smaller clusters first. Though, at the same 10 clusters, Single algorithm joins larger clusters more. Two different clustering distance algorithms had shown very different Constellation patterns.

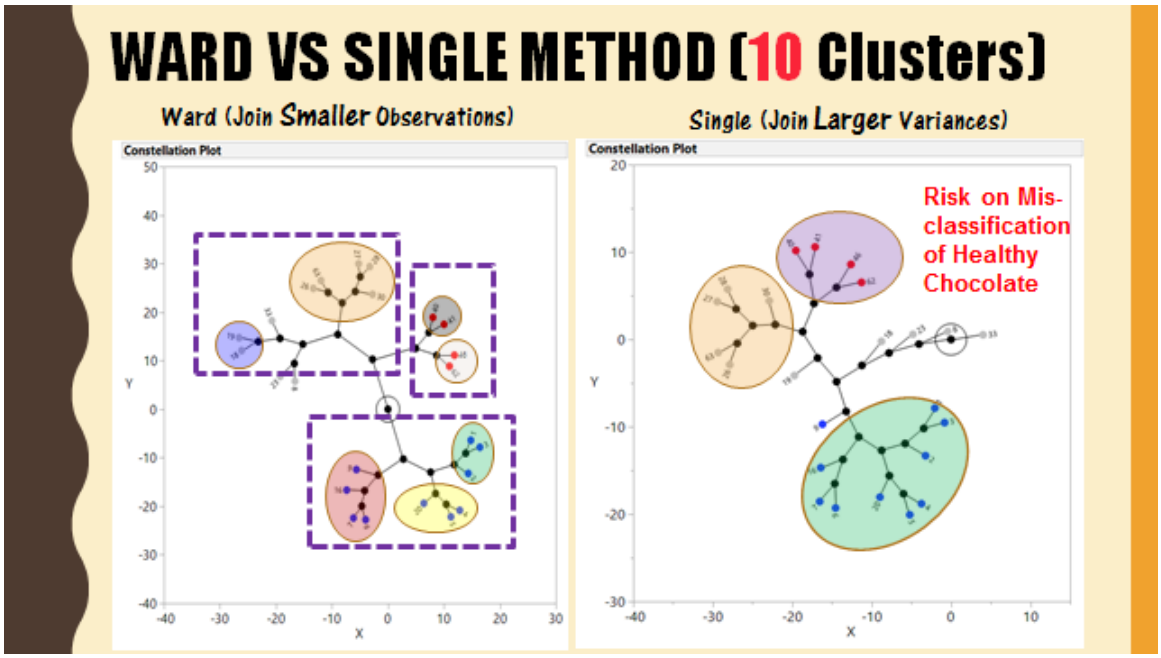

Figure 11: Constellation Plot for Ward vs. Single Clustering Method

3.4 Principle Component Analysis

Since some algorithms may join smaller clusters first while the other algorithms may join larger clusters first. The number of eventual clusters may impact the end result of clustering patterns. In order to optimize the number of clusters in clustering analysis, Principle Component Analysis (PCA) is utilized. The first principal component is the linear combination of the standardized original variables that has the greatest possible variance. Each subsequent principal component is the linear combination of the variables that has the greatest possible variance and is uncorrelated with all previously defined components.

Lists the eigenvalue that corresponds to each principal component in order from largest to smallest. The eigenvalues represent a partition of the total variation in the multivariate sample. In Figure 12, the amplitude of seven Eigenvalues are ranked from top to bottom. Based on Pareto concept, top four Eigenvalues would reach Cum Percent > 80%.

Eigenvalues

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	5.9254	34.855		34.855
2	4.8336	28.433		63.288
3	2.0252	11.913		75.201
4	1.3256	7.797		82.998
5	1.0255	6.032		89.030
6	0.9010	5.300		94.330
7	0.6484	3.814		98.144

Figure 12: Eigenvalues of PCA

In Figure 13, Scree plot also shows a graph of the eigenvalue for each component. This scree plot helps in visualizing the dimensionality of the data space. Scree plot also indicates 4 cluster components are best (near knee of transition).

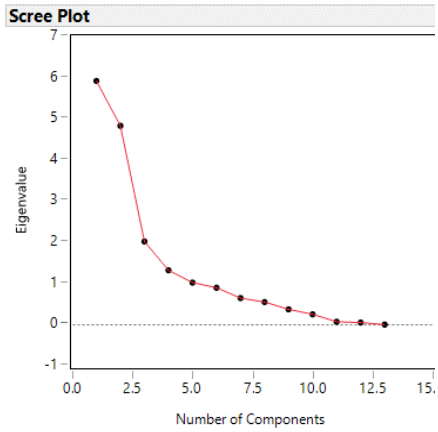

Figure 13: Scree Plot of Eigenvalues

3.5 Compare Clustering Distance Algorithms

The distribution analysis was conducted for comparing Ward algorithm and Single algorithm. In Figure 14a, only the first Ward cluster was selected which happens to be the healthier half of the dark chocolate (selected in dark green). The Single clusters are also listed aside for direct comparison. The second cluster for Ward consists of the unhealthy dark chocolate while the third consists of both milk chocolate and white chocolate. This direct comparison may indicate that Ward algorithm is better at dividing healthy and unhealthy clusters than Single algorithm.

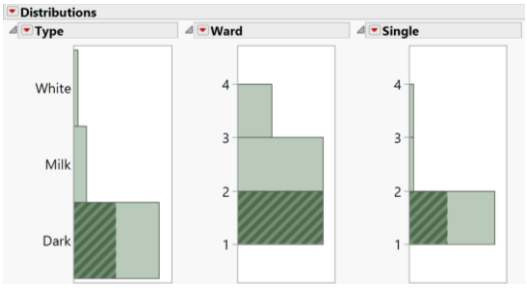

Figure 14a: Distribution Analysis for the 1st Ward Cluster (selected) among Chocolate Types

In Figure 14b, instead, all of the first Single cluster is selected. The first Single cluster consists of almost all of the data set while the second and third Single cluster contain very little data. As studied before, Single combines larger variance and large-sized clusters. Therefore, the first cluster is relatively large compared to the second and third cluster. In conclusion, Ward algorithm is more accurate at dividing healthy and unhealthy chocolate products (dark vs. milk/white).

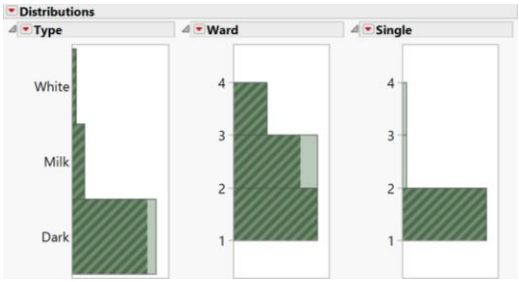

Figure 14b: Distribution Analysis for the 1st Single (selected) among Chocolate Types

Section 3.4 Principle Component Analysis suggests 4 clusters are a good choice to cluster Chocolate Products. Figure 14c further studies the effect of number of clusters between two cluster algorithms (Ward vs. Single). The objective is to investigate when both Ward and Single algorithms can converge to the similar clustering patterns when cluster number is higher. For Ward algorithm, when increasing the cluster numbers, the cluster is becoming more even (joining smaller clusters first). However, for Single algorithm, the first cluster still dominates the majority of the data populations (joining larger clusters first). The clustering pattern between two cluster algorithms are still different in the range of 3-5 clusters.

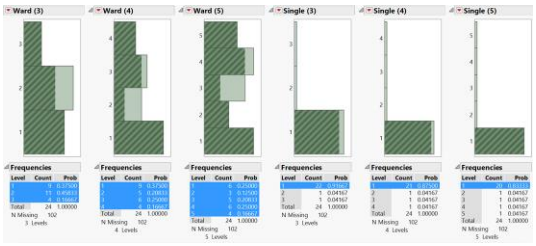

Figure 14c: Distribution Analysis vs. Cluster Algorithm and Cluster Number

4. Conclusions and Future Work

STEAM approach is very successful on understanding Chocolate Science Research and Nutrition Food Science. Modern Multivariate Statistics and Artificial Intelligence Algorithms can explore the Chocolate Science Patterns which can further help consumers pick their healthy chocolate products based on their preferred nutritions needed. Clustering distance algorithm is critical on deciding the clustering sequence and clustering patterns. Cluster number is also determined by Principle Component Analysis and Scree Chart. In order to pick the better Clustering algorithm to interpret the Chocolate Science and Chocolate Product, the direct comparison of two extreme clustering algorithms have been thoroughly studied. This STEAM approach can be applied to similar fields such as Coffee Science and Product, as well as to other Healthy Nutrition Study.

Acknowledgements

Authors would like to thank the Biology Advisor Mr. Patrick Giuliano for helping and supporting me throughout this project.

References

1. Petyaev, Ivan M., and Yuriy K. Bashmakov. “Dark Chocolate: Opportunity for an Alliance between Medical Science and the Food Industry?” Current Neurology and Neuroscience Reports., U.S. National Library of Medicine, 2017,
2. Magrone, Thea, et al. “Cocoa and Dark Chocolate Polyphenols: From Biology to Clinical Applications.” Current Neurology and Neuroscience Reports., U.S. National Library of Medicine, 2017,
3. Allen, R. R., Carson, L., Kwik-Uribe, C., Evans, E. M., & Erdman, J. W. (2008 April), “Daily consumption of a dark chocolate containing flavanols and added sterol esters affects cardiovascular risk factors in a normotensive population with elevated cholesterol”, Journal of Nutrition. 138(4):725-31.
4. “Is Dark Chocolate or Cocoa a Good Source of Iron?” ConsumerLab.com,
5. Rao, Linda. “Dark Chocolate Can Pack a Big Antioxidant Wallop.” Prevention, Prevention, 25 May 2018,
6. Wensem, van. “Overview of Scientific Evidence for Chocolate Health Benefits.” Environmental Toxicology and Chemistry, Wiley-Blackwell, 26 Dec. 2014.
7. Panche, A. N., et al. “Flavonoids: An Overview.” Current Neurology and Neuroscience Reports., U.S. National Library of Medicine, 2016,
8. Latif, R. (2013, March). Chocolate/cocoa and human health: a review. The Netherlands Journal of Medicine. 71(2):63-8.
9. Patel Wang, J., Varghese, M., Ono, K., Yamada, M., Levine, S., Tzavaras, N., Pasinetti, G. M. (2014). Cocoa extracts reduce Oligomerization of amyloid-β: implications for cognitive improvement in Alzheimer's disease. Journal of Alzheimer's disease, 41(2):643-50.
10. R. K., Brouner, J., & Spendiff, O. (2015, December). Dark chocolate supplementation reduces the oxygen cost of moderate intensity cycling. Journal of the International Society of Sports Nutrition 2015, 12:47.
11. Crichton, G. E., Elias, M. F., Alkerwi, A. (2016, May). Chocolate intake is associated with better cognitive function: The Maine-Syracuse Longitudinal Study. 100:126-32
12. “Antioxidants: In Depth.” National Center for Complementary and Integrative Health, U.S. Department of Health and Human Services, 4 May 2016
13. Maxwell, Simon R. J., and Gregory Y. H. Lip. Advances in Pediatrics., U.S. National Library of Medicine, Oct. 1997
14. Bentson, Jacob. “Mechanics of Plaque Formation.” Arteriosclerosis, Thrombosis, and Vascular Biology, 2014
15. LeWine, Howard, and M.D. “Sweet Dreams: Eating Chocolate Prevents Heart Disease.” Harvard Health Blog, 17 June 2015,
16. “7 Proven Health Benefits of Dark Chocolate.” Healthline, Healthline Media,
17. Cee, Jenna. “Health Risks of Dark Chocolate.” LIVESTRONG.COM, Leaf Group, 3 Oct. 2017,
18. “The Five Biggest Problems With Chocolate Milk Campaigns.” One Green Planet, 22 July 2014
19. “The Health Benefits of White Chocolate (Yes, They Exist) - Perfect Health Diet.” Perfect Health Diet, perfecthealthdiet.com/2014/03/white-chocolate/.
20. “22 Proven Health Benefits of White Chocolate (No.9 Shocking).” Dr. Heben, Dr. Heben, 9 Feb. 2017

21. “White Chocolate Disadvantages, Benefits of White Chocolate.” Tarragon Disadvantages, Benefits of Tarragon,
22. Mason C., (2018 July) “Multivariate Statistics of Antioxidant Chocolate”, IWSM Bristol Proceedings, Vol 2 37-40
23. Mason C., (2018 July), “Choose Healthy Chocolate”, IEOM Europe Proceedings, 434-441
24. Wu, Anna Dong. “Starbucks and Cardiovascular Disease Prevention.” IEOM, IEOM Society, 26 July 2018,
25. Harris, C.W. and Kaiser, H.F. (1964), “Oblique Factor Analytic Solutions by Orthogonal
26. Golub, G.H. and van der Vorst, H.A., (2000), “Eigenvalue Computation in the 20th Century,” Journal of Computational and Applied Mathematics 123, 35-65.
27. Jackson, J. Edward (2003), A User’s Guide to Principal Components, New Jersey: John Wiley and Sons.
28. Mardia, K., Kent, J., and Bibby, J. (1980), Multivariate Analysis, First Edition, New York: Academic Press.
29. Milligan, G.W. (1980), “An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms,” Psychometrika, 45, 325–342.
30. Hartigan, J.A. (1981), “Consistence of Single Linkage for High–Density Clusters,” Journal of the American Statistical Association, 76, 388–394.
31. Florek, K., Lukaszewicz, J., Perkal, J., and Zubrzycki, S. (1951a), “Sur La Liaison et la Division des Points d’un Ensemble Fini,” Colloquium Mathematica, 2, 282–285.
32. Jardine, N. and Sibson, R. (1971), Mathematical Taxonomy, New York: John Wiley and Sons.
33. McQuitty, L.L. (1957), “Elementary Linkage Analysis for Isolating Orthogonal and Oblique Types and Typal Relevancies,” Educational and Psychological Measurement, 17, 207–229.
34. Sokal, R.R. and Michener, C.D. (1958), “A Statistical Method for Evaluating Systematic Relationships,” University of Kansas Science Bulletin, 38, 1409–1438.
35. Sneath, P.H.A. (1957) “The Application of Computers to Taxonomy,” Journal of General Microbiology,17, 201–226.
36. “Correlations and Multivariate Techniques.” Multiple Linear Regression