# Statistics Application on the Study of Chocolate Science with Heart Disease

Mason Chen[1], Charles Chen[2]

[1]Stanford University OHS, Palo Alto, CA, USA, mason05@ohs.stanford.edu
[2]Applied Materials, Sunnyvale CA, USA, charles.chen.training@gmail.com

**Abstract**

The objectives of this paper are to use Multivariate Statistics to define a health biometric on choosing a healthy chocolate to patients with heart disease. Chocolate, made from cocoa beans, contains flavonoids which contain antioxidants. Flavonoids are the most abundant polyphenols in the human diet. Polyphenols have antioxidant properties which can prevent aging and is also beneficial to heart disease and diabetes patients. People with heart diseases should eat less of saturated fat, trans fat, sodium, and cholesterol. They should eat more dietary fiber. Cocoa flavanols promote healthy blood flow circulation from head to toe. The heart, brain, and muscle depend on a healthy circulatory system. Data has been collected on 20+ chocolate ingredient contents from 60+ different types of chocolate. Multivariate correlation study has found that (1) strong negative correlation between Cocoa and Sugar, and (2) a strong positive correlation between Dietary Fiber and Iron. Most dark chocolate contains more cocoa and less sugar. Dietary fiber and iron are high in correlation because of the high cocoa percent. The above two correlations can be further explained by conducting the Hierarchical Clustering Analysis on separating the Dark Chocolate, Milk Chocolate, and White Chocolate. The Cocoa and Calcium are the deciding factors to separate these three Chocolates.

**Keywords:** STEM, Flavonoids, Chocolate, Statistics, Antioxidant, Clustering

## 1. Introduction

Many people like eating chocolate but have concerns that chocolate is unhealthy. Are they sure whether eating chocolate is unhealthy? The objectives of this paper are to find out if eating chocolate is unhealthy, what diseases can be prevented by chocolate, why can chocolate prevent those diseases, what chocolate nutrition help prevent those diseases, and how to select the best chocolate for preventing those diseases.

### 1.1 Chocolate and Atrial Fibrillation Research

The objective of this research is to evaluate the relative risk of CVD (cardiovascular disease). Research shows that individuals consuming chocolate > once per week have a lower risk of AF (Atrial Fibrillation: a type of common cardiovascular heart disease) than individuals consuming chocolate regularly. In Figure 1, the x-axis is the chocolate consumption (30 g / week) while the y-axis is the risk of CVD. There was little further reduction in the risk of heart disease greater than 3 servings per week. This proves that chocolate, especially dark [1], may be inversely associated with AF and a healthy snacking option. A lot of time was taken to find this chart that showed chocolate does help prevent heart disease [2]. However, this chart has many loopholes such as it does not tell us which is the best chocolate type. This

project can fill in those loopholes and demonstrate which is the best chocolate to eat. Now that we know chocolate does help prevent heart disease, what in chocolate helps prevent it?

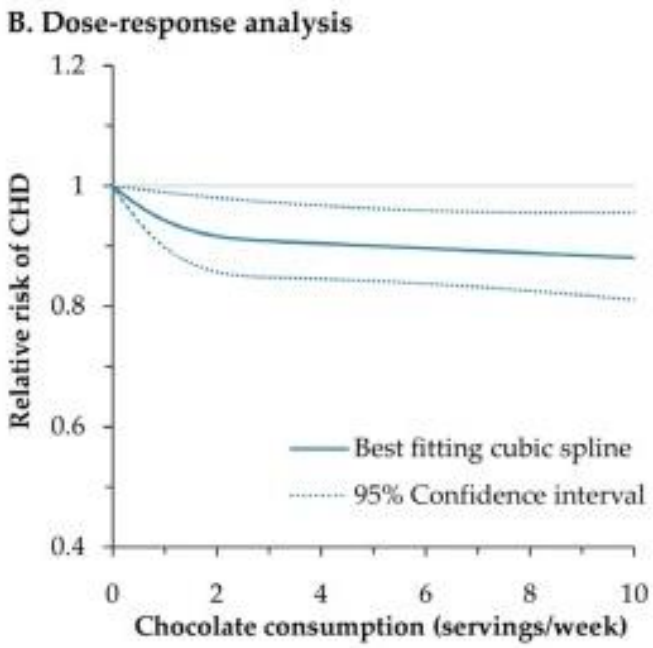


**Figure 1:** Chocolate Consumption vs Risk of Cardiovascular Heart Disease [3] (1 serving = 30 g)

### 1.2 Flavonoids Science and Structure

Now that chocolate is proven to help prevent against heart disease, what is the critical ingredient in chocolate that prevents against heart disease? Flavonoids are the most abundant polyphenols in the human diet [4], representing 2/3 of those digested [5]. Polyphenols are compounds found abundantly in natural food sources that have antioxidant properties. Flavonoids have the general structure of a 15- carbon skeleton as shown in Figure 2. The structure is abbreviated as C6-C3-C6 and consists of two phenyl rings (A and B) and a heterocyclic ring (C). There are seven different types of flavonoids (classified based on its chemical structure): flavones, flavanol, flavanones, isoflavones, anthocyanidins, chalcones, and catechins. Chocolate flavonoids are flavanols. Antioxidants are the critical ingredient that prevents against heart disease. The higher the antioxidant, the higher the cocoa percent. Dark chocolate has the highest cocoa percent, so it should be the healthiest chocolate [6].

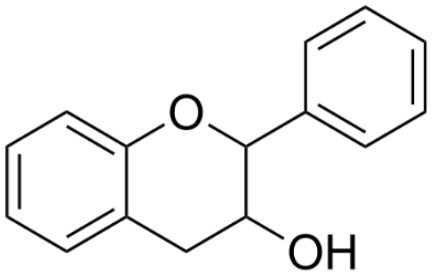


**Figure 2:** Flavonoids Structure

Why do antioxidants prevent heart disease? Cocoa flavanols which contain antioxidants promote healthy blood flow from head to toe [7]. The heart, brain, and muscle depend on a healthy circulatory system [8] as shown in Figure 3. Supporting healthy blood flow is essential to helping maintain exceptional health throughout life [9]. Flavanol benefits include a longer life, weight control, and prevention of cardiovascular disease, cancer, diabetes, and neurodegenerative disease [10]. People with heart diseases should eat less saturated fat, trans fat, sodium, and cholesterol [11]. They should eat more dietary fiber [12]. We know that chocolate contains antioxidants which are the key ingredient that prevents heart disease [13]; however, is chocolate the only food that prevents heart disease?

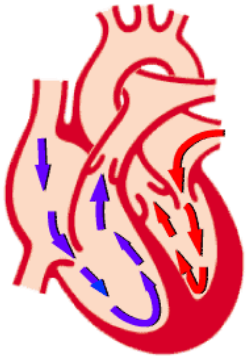

**Figure 3:** Healthy Circulatory System and Blood Flow

### 1.3 Chocolate vs Coffee Research

Chocolate was studied on its effect on heart disease, what about coffee? Do they have the same effect on heart disease? Coffee is commonly suggested as heart-healthy as well [14]. Research was conducted comparing and contrasting chocolate and coffee to find out if there were any similarities between coffee and chocolate and which one was more heart-healthy. Research has not shown whether coffee or chocolate is better at preventing heart disease. The greater the intake of both coffee and chocolate, the smaller their effects on heart disease. Their effects on the human body, however, differ. Coffee, as we all know, contains caffeine which keeps us awake while chocolate contains an enzyme called phenylethylamine which boosts our mood. We finished our research about chocolate, coffee, and heart disease. How do we find out which dark chocolate is the healthiest chocolate and prevents heart disease the most?

## 2. Graphical and Multivariate Statistical Analysis

Section 1 explained the scientific research used to help define the Health Index. Section 2 will explain the statistical modeling used to help define the Health Index in section 4. The data collection plan will be explained. Variable clustering method will be implemented to back up the 8 variables chosen by scientific research. Statistical analyses such as distribution, correlation, and dendrogram analysis were used to find any correlations between the ingredients and help define the health index.

### 2.1 Data Collection and Transformation

It's critical to collect the right chocolate data. Target was chosen since it had plenty of chocolate products (enough sample size) and was extremely convenient for collecting data. 60+ different types of chocolates were collected, and each had 20 variables. Ensuring good data quality is critical to screen out noise data. Not all 20 variables were used; instead, only 8 variables that were crucial to heart disease based on Define Phase were used as shown in Figure 4.

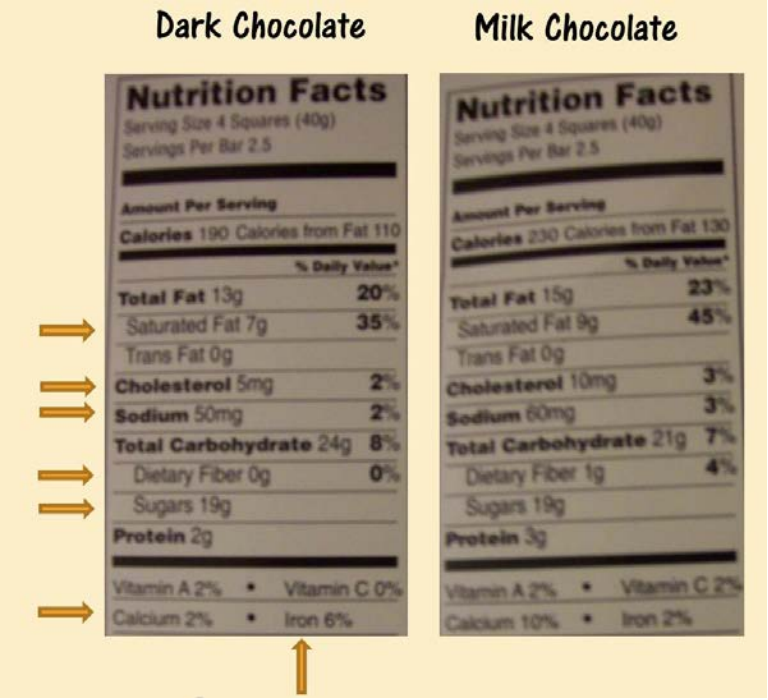


**Figure 4.** 8 variables chosen to be used in statistical modeling based on scientific research

The raw data collected have different distributions (sample mean and sample standard deviation). In order to eliminate the bias of the central tendency and spread, the raw data was transformed to become Z- Standardized Data as shown in Table 1. After Z-transformation, all variables have new sampled distributions of mean, 0, and standard deviation. The objective of this transformation is to eliminate any larger variation bias in building the statistical modeling of deriving the chocolate health index.

**Table 1:** A Subset of Z- Standardized Data

| Chocolate%_1 | Chocolate%_1 | Calorie/Size_1 | CalFat/Size_1 | Tfat/Size_1 | Sfat/Size_1 | Cholest/Size_1 | Sodium/Size_1 | Carbs/Size_1 |
|---|---|---|---|---|---|---|---|---|
| 1.46313048 | 1.46313048 | 0.88587249 | 1.741211934 | 1.965213277 | 1.427055953 | -1.112379727 | -0.72659527 | -2.484040893 |
| 1.208673005 | 1.208673005 | 0.518015041 | 1.176780474 | 0.909641981 | 0.750357618 | -1.112379727 | -0.573353644 | -1.703445541 |
| 1.2595645 | 1.2595645 | 0.231903692 | 1.897998451 | 2.023856131 | 1.539839005 | -1.112379727 | -1.033078522 | -2.137109629 |
| 0.547083571 | 0.547083571 | 0.88587249 | 1.14542317 | 1.085570534 | 0.33681974 | -1.112379727 | -1.033078522 | -1.44324709 |
| 0.547083571 | 0.547083571 | -0.422065106 | 0.39284789 | 0.381856333 | 0.33681974 | -1.112379727 | -1.033078522 | -0.286809533 |
| 0.445300581 | 0.445300581 | 1.253729938 | 1.176780474 | 1.173534805 | 1.088706786 | -1.112379727 | -0.72659527 | -1.18304864 |
| 0.547083571 | 0.547083571 | -0.198338938 | 0.835150906 | 0.881863788 | 0.590086963 | -1.112379727 | -1.033078522 | -0.9502395 |
| 0.445300581 | 0.445300581 | 0.518015041 | 1.176780474 | 1.173534805 | 1.088706786 | -1.112379727 | 1.418787496 | -0.922850189 |
| -0.063614369 | -0.063614369 | 1.458095188 | 1.709854631 | 1.73064188 | 1.915782527 | -0.109165248 | -0.692541575 | -1.269781453 |
| | | -0.972775673 | -0.947264232 | -0.784827734 | -0.834541114 | 0.075637418 | 1.547833076 | 0.967012239 |
| | | -1.076033903 | -0.610585817 | -0.673714964 | -0.715822105 | -1.112379727 | 0.192854487 | 0.638340516 |
| | | -1.076033903 | -0.610585817 | -0.673714964 | -0.715822105 | -1.112379727 | 0.192854487 | 0.638340516 |
| | | -1.076033903 | -0.986873458 | -1.025572059 | -0.715822105 | -1.112379727 | -1.033078522 | 1.332203055 |
| | | 1.866825687 | 1.709854631 | 1.73064188 | 1.915782527 | 0.141638375 | 0.499337739 | -1.269781453 |
| | | -1.076033903 | -0.986873458 | -1.025572059 | -1.166954338 | -1.112379727 | -1.033078522 | 1.332203055 |
| -0.063614369 | -0.063614369 | 4.09626477 | -0.029050373 | -0.001987771 | 0.104418294 | -1.112379727 | -1.033078522 | 0.0706348 |
| | | 1.458095188 | 1.709854631 | 1.73064188 | 1.915782527 | -0.109165248 | -0.692541575 | -1.269781453 |
| 0.547083571 | 0.547083571 | 0.612118109 | 0.579533231 | 0.578241688 | 0.175950897 | -1.112379727 | 2.530680224 | -0.765520894 |
| -0.063614369 | -0.063614369 | -0.072267843 | -0.208045551 | -0.403685101 | -0.138792515 | -1.112379727 | 1.105176727 | 0.20265939 |
| 0.954215531 | 0.954215531 | 0.88587249 | 1.176780474 | 1.173534805 | 1.088706786 | -1.112379727 | -1.033078522 | -1.703445541 |
| | | 1.458095188 | 1.396281597 | 1.73064188 | 1.915782527 | -1.112379727 | -0.692541575 | -1.269781453 |
| | | 1.458095188 | 1.709854631 | 1.73064188 | 1.915782527 | 0.141638375 | -0.692541575 | -1.269781453 |
| 0.445300581 | 0.445300581 | 0.518015041 | 0.894564744 | 1.173534805 | 0.750357618 | -0.209486695 | 1.725270748 | -0.922850189 |
| | | -1.076033903 | -1.363161098 | -1.025572059 | -0.715822105 | -1.112379727 | -1.033078522 | 1.332203055 |
| | | 1.049364689 | 1.082708564 | 1.144213378 | 1.539839005 | 0.141638375 | 1.520948579 | -0.980672071 |

## 2.2 Variable Clustering

After raw data has been Z-transformed, JMP variable clustering (Figure 5) was conducted to find out and analyze the correlations between the ingredients within the cluster. The red squares near each form different clusters. There is one large red group on the top left, two medium groups in the middle, and another large one located in the bottom right. The top 2-3 highest correlations, based on highest r-squares (Figure 6), were analyzed. The 8 variables chosen by scientific research were saturated fat, cholesterol, iron, dietary fiber, sugar, calcium, cocoa percent, and sodium. Will the top 8 variables chosen by JMP clustering match the 8 variables chosen by science? The first cluster was grouped because the higher the saturated fat, the higher the total, and therefore, the higher the calories. Saturated fat was chosen since it had a high r-square which matches one of the 8 variables chosen by research. The second cluster justified that calcium and cocoa percent should have a negative correlation since dark chocolate contains the most cocoa and should have the least amount of milk. The lower the cocoa percentage, the higher the milk since white chocolate contains the most milk but the least amount of cocoa. The three highest r-squares from the second cluster were cocoa-percent, cholesterol, and calcium which all match the 8 variables chosen by science. For the third cluster, sugar and carbohydrates have the highest r-square and sugar matches with the top 8 variables. For the last cluster, dietary fiber and iron had the highest r-square and also matches with the 8 variables. In conclusion, the 8 variables chosen by statistics do match the 8 variables chosen by science. Statistics can back up science.

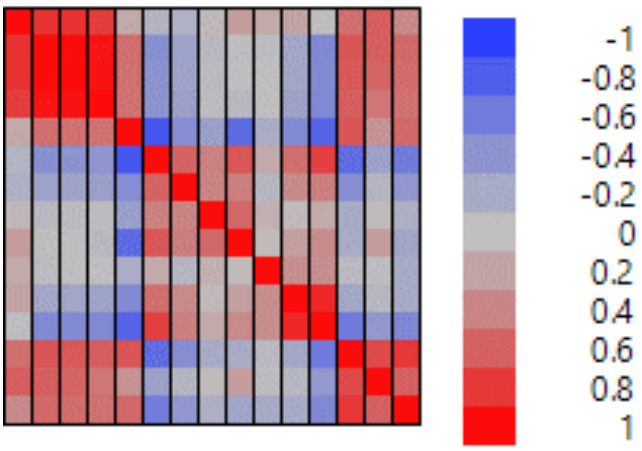


**Figure 5:** Correlations Color Map

| Cluster | Members | RSquare with Own Cluster |
|---|---|---|
| 1 | Calories (g) | 0.789 |
| 1 | Calories_from_Fat (g) | 0.976 |
| 1 | Total_Fat (g) | 0.977 |
| 1 | Saturated_Fat (g) | 0.935 |
| 2 | Cocoa_Percent | 0.742 |
| 2 | Cholesterol (mg) | 0.811 |
| 2 | Vitamin_A | 0.505 |
| 2 | Vitamin_C | 0.412 |
| 2 | Calcium | 0.726 |
| 3 | Sodium (mg) | 0.345 |
| 3 | Carbs (g) | 0.876 |
| 3 | Sugar (g) | 0.874 |
| 4 | Dietary_Fiber (g) | 0.888 |
| 4 | Protein (g) | 0.73 |
| 4 | Iron | 0.803 |

**Figure 6:** Cluster Members and R-Square

### 2.3 Distribution Analysis

JMP interactive graphical analysis was conducted in order to uncover the comprehensive chocolate nutrition distributions. The objective is to find any patterns in how the chocolate producers made today's chocolate product types in order to provide us better insight regarding the chocolate health index. The first dark chocolate sample distribution (Figure 7) was conducted on the eight variables and chocolate type to compare the different chocolate types' nutrition facts.

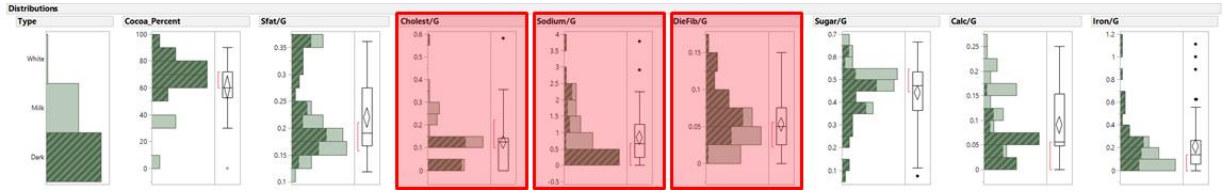


**Figure 7:** Dark Chocolate Distribution

After looking at the interactive graphical mode of dark chocolate nutrition distribution (dark chocolate is selected in dark green in Figure 7), some interesting correlations were found. Dark chocolates have relatively low cholesterol, low sodium, and high dietary fiber. This helps prove that the hypothesis (dark chocolate is healthier than milk chocolate) may be correct. Milk chocolate (in Figure 8), on the other hand, does not show any significant correlation patterns among the variables analyzed. Most sampled distributions are near random (white noise). This observation may indicate there is no health requirement on formulating the chocolate nutrition ingredients for the milk chocolate. The distribution contrast between dark chocolate and milk chocolate has provided first-hand information on how to derive the Chocolate Health Index.

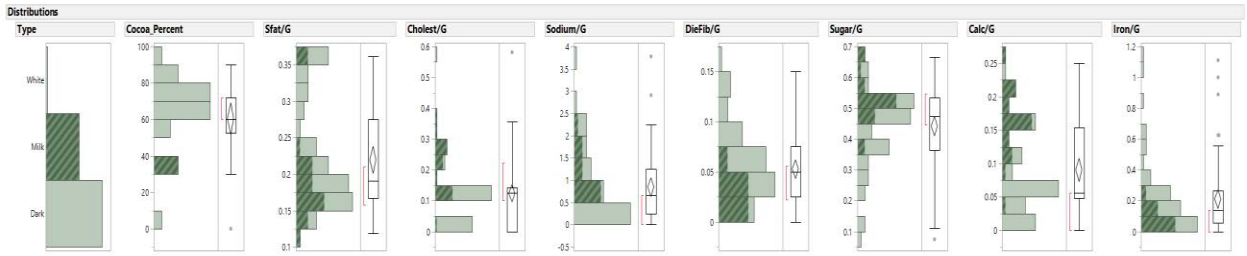


**Figure 8:** Milk Chocolate Distribution

## 2.4 Dark Chocolate Correlation Analysis

This JMP multivariate correlation study [15] shown in Figure 9 and Table 2 was further done to see if any chocolate type has any strong correlation(s) between healthy and unhealthy. Correlations between <-0.75 and >0.75 thresholds were set to identify any strong nutrition correlation pattern.

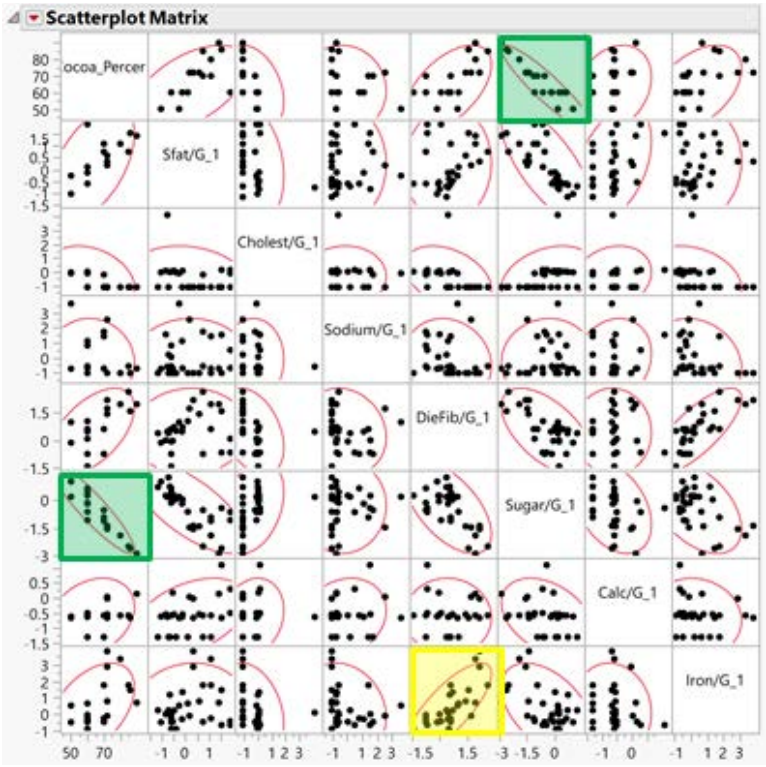


**Figure 9:** Scatterplot Matrix

**Table 2:** Multivariate Correlations Table

**Correlations**

| | Cocoa_Percent | Sfat/G_1 | Cholest/G_1 | Sodium/G_1 | DieFib/G_1 | Sugar/G_1 | Calc/G_1 | Iron/G_1 |
|---|---|---|---|---|---|---|---|---|
| Cocoa_Percent | 1.0000 | 0.5291 | -0.3114 | -0.0583 | 0.5482 | -0.9162 | 0.2625 | 0.4597 |
| Sfat/G_1 | 0.5291 | 1.0000 | -0.1980 | 0.0184 | 0.0341 | -0.7068 | 0.4161 | 0.0687 |
| Cholest/G_1 | -0.3114 | -0.1980 | 1.0000 | 0.0302 | -0.3666 | 0.3333 | 0.1732 | -0.3304 |
| Sodium/G_1 | -0.0583 | 0.0184 | 0.0302 | 1.0000 | -0.1344 | 0.0462 | 0.1667 | -0.1862 |
| DieFib/G_1 | 0.5482 | 0.0341 | -0.3666 | -0.1344 | 1.0000 | -0.5804 | -0.0207 | 0.7722 |
| Sugar/G_1 | -0.9162 | -0.7068 | 0.3333 | 0.0462 | -0.5804 | 1.0000 | -0.3696 | -0.4669 |
| Calc/G_1 | 0.2625 | 0.4161 | 0.1732 | 0.1667 | -0.0207 | -0.3696 | 1.0000 | -0.1037 |
| Iron/G_1 | 0.4597 | 0.0687 | -0.3304 | -0.1862 | 0.7722 | -0.4669 | -0.1037 | 1.0000 |

The scatterplot matrix shows a visual diagram of the nutrition correlations among commercial chocolate products. The straighter and more diagonal the line, the stronger the correlation is between any two nutrition variables. Sugar and cocoa have a strong negative correlation of -0.9162. This shows that the higher the chocolate percent is, the lower the sugar percent. Since dark chocolate has high chocolate percent and low sugar, this correlations study indirectly indicates that dark chocolate is the healthiest chocolate type. The other identified strong positive correlation is between dietary fiber and iron. One science research has shown in Figure 10 that most dark chocolate products with 70%-85% cocoa percent are rich in fiber and iron. Dietary fiber and iron are high because of the dark chocolate's high cocoa percent. Both graphical analyses have further provided why dark chocolate is healthier due to certain skewed nutrition preference.

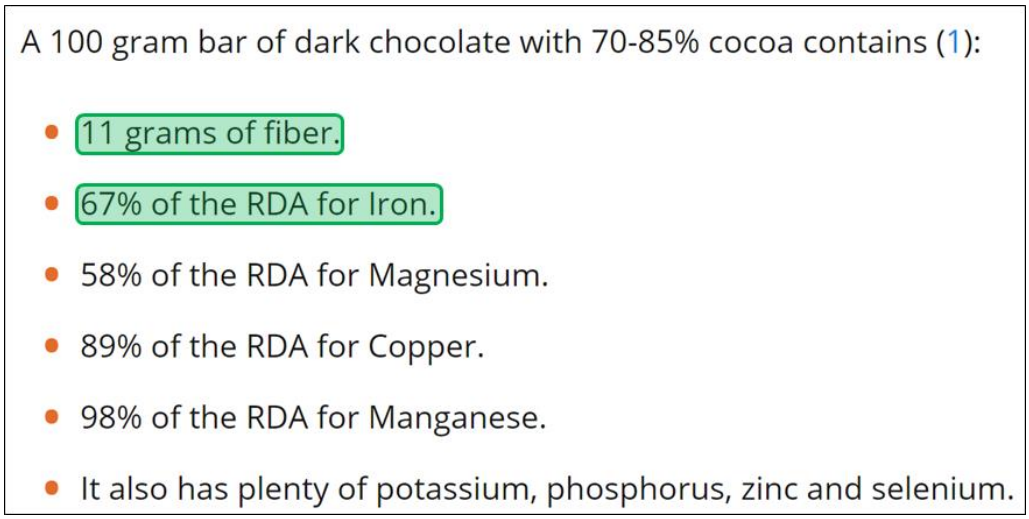


**Figure 10:** Iron and Dietary Fiber Correlation Research

### 2.5 Dendrogram and Cluster Distribution Analysis

After finding correlations between the ingredients, it was time for JMP to group the chocolates into three clusters as shown in Figure 11. Two of the clusters contained dark chocolate while the other cluster contained both milk and white chocolate as shown in Figure 12. The two dark chocolate clusters divided the healthier and unhealthier dark chocolate. This cluster analysis used the Ward method. In a separate principle deciding factor analysis, calcium was the main ingredient that separated the two dark chocolate clusters from milk and white. Cocoa percent was what divided all three clusters.

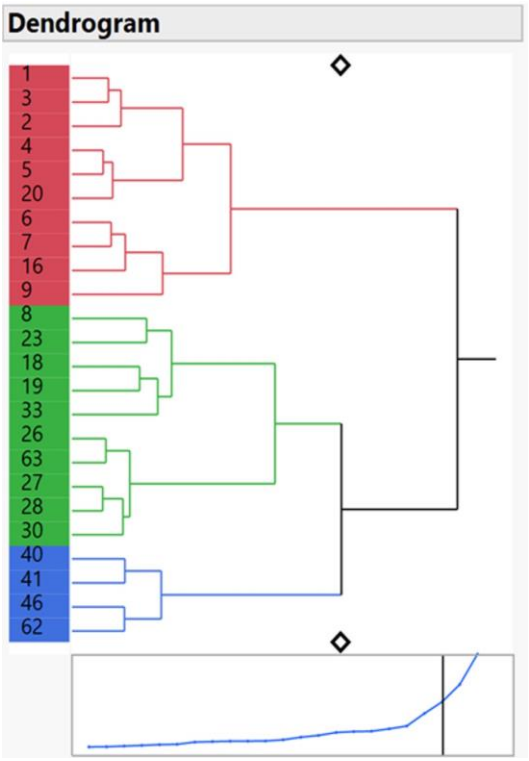


**Figure 11:** Dendrogram Analysis

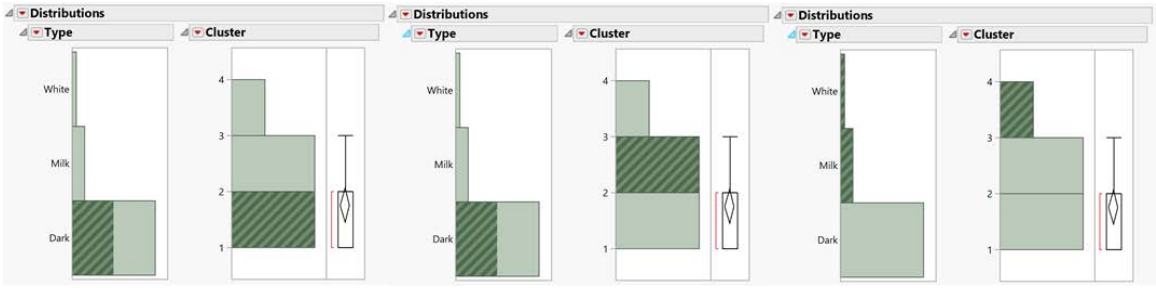


**Figure 12:** Cluster Distribution Analysis

### 3. Health Index and Results

The health index was determined based on the past research and statistical analysis. The eight critical variables were given multiplied by coefficients. Nutritions that have -2x or 2x mean that according to the Define Phase, these are positive (2) or negative (-2) to heart diseases. Nutritions that have -1 or 1 means that in general cases, these are positive (1) or negative (-1) to your overall health. The formula is further clarified in Figure 13.

$$[(Chocolate\% * 2) + (Dietary\_Fiber * 2) - (Sugar * 2) + Calcium - (Saturated\_Fat * 2) - (Cholesterol * 2) - Sodium + Iron] = Health\ Index$$

**Figure 13:** Health Index Formula

Based on the health index, the top four healthiest chocolates were determined as shown in Figure 14. Those chocolates were Lindt Excellence 90% Cocoa, Lindt Excellence 85% Cocoa, Ghirardelli Midnight Reverie 86% Cocoa, and Equal Exchange Chocolate Organic Panama Extra Dark. They all have >80% cocoa, no cholesterol, high dietary fiber, and low sugar.

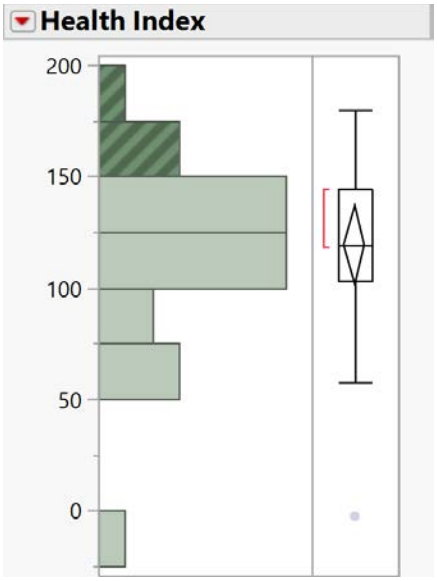


**Figure 14:** Health Index Results

## 4. Conclusions

The top four healthiest chocolates for heart disease were found in this paper. JMP software tools such as cluster analysis, correlation analysis, and distribution analysis were all applied to this project. Cocoa science, such as cocoa production, flavonoids, antioxidants, flavanol benefits, and the different types of chocolate, was learned throughout this paper. STEM approach was applied effectively to define the project scope by taking systematic scientific literature and engineering problem-solving techniques. Further research may consider different health indexes for other diseases (cancer, diabetes, etc). The STEM approach could be applied more in daily life. The other opportunity is that antioxidants and flavonoid science will be further researched and proven through the STEM framework.

## Acknowledgments

I would like to thank my statistics and math advisor Dr. Charles Chen and my 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, www.ncbi.nlm.nih.gov/pmc/articles/PMC5626948/.
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, www.ncbi.nlm.nih.gov/pmc/articles/PMC5465250/.

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, www.consumerlab.com/answers/is-dark-chocolate-or-cocoa-a-good-source-of-iron/dark_chocolate_cocoa_iron/.
5. Rao, Linda. "Dark Chocolate Can Pack a Big Antioxidant Wallop." Prevention, Prevention, 25 May 2018, www.prevention.com/food-nutrition/healthy-eating/a20454762/dark-chocolate-and-antioxidants-0/.
6. Wensem, van. "Overview of Scientific Evidence for Chocolate Health Benefits." *Environmental Toxicology and Chemistry*, Wiley-Blackwell, 26 Dec. 2014, setac.onlinelibrary.wiley.com/doi/full/10.1002/ieam.1594.
7. Panche, A. N., et al. "Flavonoids: An Overview." *Current Neurology and Neuroscience Reports.*, U.S. National Library of Medicine, 2016, www.ncbi.nlm.nih.gov/pmc/articles/PMC5465813/.
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. Mason C., (2018 July) "Multivariate Statistics of Antioxidant Chocolate", IWSM Bristol Proceedings, Vol 2 37-40
13. Mason C., (2018 July), "Choose Healthy Chocolate", IEOM Europe Proceedings, 434-441
14. Wu, Anna Dong. "Starbucks and Cardiovascular Disease Prevention." IEOM, IEOM Society, 26 July 2018, www.ieomsociety.org/paris2018/papers/424.pdf.
15. "Correlations and Multivariate Techniques." Multiple Linear Regression, www.jmp.com/support/help/14/correlations-and-multivariate-techniques.shtml.