Principal Component Analysis of Foods Nutrition Science Research

Mason Chen[1], Charles Chen[2], Patrick Giuliano[3]
[1]Stanford OHS, Palo Alto, CA 94303
[2,3]Morill Learning Center, San Jose, CA 95132

Abstract
The purpose of this project is to determine which Starbucks drinks among all coffee and tea options are best for cardiovascular disease (CVD) prevention and overall good health. A Science-based Health Index was constructed considering different coffee/tea constituents, including: saturated fat, cholesterol, sodium, carbohydrates, dietary fiber, sugars, protein, and caffeine. Antioxidant activity of flavonoids from Caffeine contained within Coffee/Tea can reduce free radical formation and scavenge free radicals. Principal Component Analysis (PCA) was used to explore all factors in the analysis and to inform on the utility of the health index in relation to its link to CVD prevention and good health. Principal Component 1 is more relevant to most unhealthy components such as sugars, carbohydrates, saturated fat, and total fat. Principal Component 2 is more related to healthy Caffeine nutrition. Additionally, Dietary Fiber and Caffeine are most opposite against the other unhealthy components along the direction of the both 1st and 2nd Principal Components. PCA Eigen Analysis is very powerful computational and visual diagnostic tool for discrimination and classification of coffee product types based on patterns in nutritional constituents. The PCA-based Health Index was derived based on the eigenvalues and eigenvectors of the first two Principal Components. The PCA-based Health Index was also compared and correlated to the Science-based Health Index (about 70%-80% R-Square Curve Fitting). The other Foods like Candy, Chocolate, Cereal are studied and their Nutrition Eigen plots of the first two principal components are quite different each other.

Key Words: PCA, Starbucks Coffee, Statistics, Anti-Oxidant, Cardiovascular Disease

1. INTRODUCTION

Studies show that a moderate intake of coffee, from 3-5 cups per day, shows an inverse relationship with cardiovascular disease. Regular consumption of tea, coffee and chocolate has also been associated with a diminished risk of CVD [Chen 2018, Wu,2018]. Conditions that lead to heart disease include high cholesterol, high blood pressure, and other chronic health problems including diabetes. A heart-healthy diet is typically low in cholesterol, trans fats, sodium, and saturated fat. Coffee and tea are rich in polyphenols with antioxidative properties in the form of flavonoids. Drinking coffee has also been proven to reduce the chance of type 2 diabetes, which often accompanies CVD. A healthy diet is one that is low in saturated fat and trans fats and restricts consumption to less than 300mg of dietary cholesterol and less than 1500mg of sodium per day.

1.1 Flavonoid Anti-Oxidant Science

Antioxidant activity of flavonoids reduce free radical formation and scavenge free radicals [Corliss,2014] as shown in Figure 1. Free radicals are atoms or groups of electrons which are highly reactive with important cellular components and cause cells to function poorly or die. Excess free radicals are thought to initiate atherosclerosis by damaging blood vessel walls, thus contributing to CVD progression. LDL cholesterol has also been implicated in heart disease, causing damage to blood vessels once oxidized by free radicals as they move across the endothelial membrane into the arterial wall. Blood vessels absorb and deposit cholesterol/ plaque, which may initiate the formation of an atherosclerotic lesion, causing blood vessel blockage. Coffee and tea provide an abundant quantity of antioxidants that reduce oxidative stress which can damage cells. Oxidation contributes to disease progression. Plaque rupture is an acute clinical event which causes heart attack or stroke in and very often requires immediate life-saving medical intervention. Regular consumption of coffee/tea is also associated with a diminished risk of heart disease by reducing LDL Cholesterol and triglycerides. The purpose of this project is to determine which Starbucks coffee and tea drinks—when considering all ingredients within them—are most beneficial to CVD prevention and therefore to overall health.

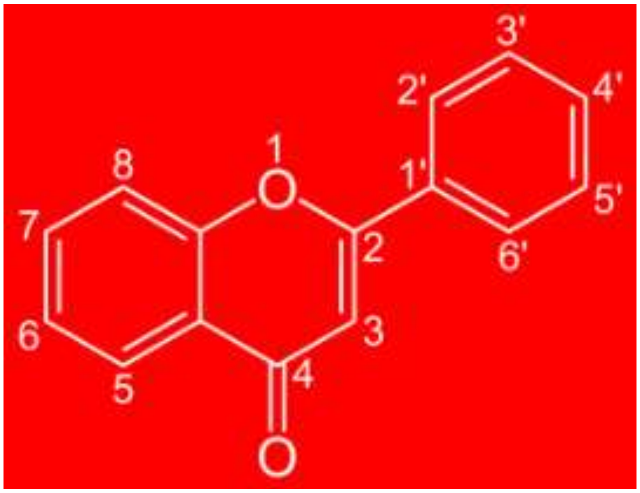


Figure 1: Flavonoid Chemical Structure

1.2 Coffee and Tea Technology

To begin coffee production, cocoa cherries are harvested, spread out, and washed to remove the pulp and parenchyma. They are then hulled, and polished, graded, sorted and finally roasted. After select beans which are considered defective are removed, coffee production is complete. In tea production, tea leaves are first plucked and laid into troughs [Ding,2018]. From there, they are blown with hot air to dry, and are “fixed” or heated to make them more aromatic. The leaves are then placed into a temperature-controlled room and are left to brown to get a more intense flavor. Tea leaves are then rolled tightly to preserve their flavor and subjected to aging and fermentation. Tea production is then considered complete.

2 DATA COLLECTION PLAN AND PCA METHODOLOGY

2.1 Data Collection

To accomplish this, data was collected from the Starbucks online menu, and each ingredient listed in the nutrition facts was made a variable (also known as a “factor”). In this project, we will be basing the health benefits (also known as the “response”) of each kind of drink on these variables, also known as constituents. Starbucks provides an online menu with nutrition facts for most of their drinks. We decided to focus on their most popular ones: Espressos, Frappuccinos, Freshly Brewed Coffee, Cold Brew and Iced Coffees, Refreshers, Teas, and Seasonal Drinks (featured during Thanksgiving/ Christmas) as shown in Figure 2. From these categories, we selected 15 drinks to analyze by designating each drink with a value and using a random number generator to obtain numbers corresponding to each drink. From the random selection of drinks within each category, the amount of calories, total fat, saturated fat, cholesterol, sodium, total carbohydrates, dietary fiber, sugar, protein, and caffeine was studied.

1
Black and White Mocha Crème Frappuccino®
2
Black and White Mocha Frappuccino®
3
Black and White Mocha Light Frappuccino®
4
Brown Sugar Shortbread Creme Frappuccino®
5
Brown Sugar Shortbread Frappuccino®
6
Brown Sugar Shortbread Light Frappuccino®
7
Caffè Vanilla Frappuccino® Blended Coffee
8
Caffè Vanilla Light Frappuccino® Blended Coffee
9
Caramel Brulée Crème Frappuccino® Blended Crème
10
Caramel Brulée Frappuccino® Blended Beverage
11
Caramel Brulée Light Frappuccino® Blended Beverage
12
Caramel Cocoa Cluster Frappuccino® Blended Coffee

Figure 2: Coffee Tea Product

2.2 Principle Component Analysis (PCA)

Principle Component Analysis will transform and reduce the original factor space dimensions through Linear Algebra Eigen methodology [Golub, 1965 & 2000]. Principle components are derived from Eigenvalues and Eigenvectors. PCA reduces the dimensionality of the correlated variables in the dataset into principal components (where N components are created for N variables), where each principal component is an independent linear combination of all of the input variables. Through PCA, 10 original nutritional factors can be condensed to 2-3 Principal components which can explain most of the variability in Coffee/Tea nutritional factors and reveal patterns in the relationships between the factors themselves from the principal components and their relative directions. Principal components are Orthogonal to each other due to the intrinsic behavior of Eigen Vectors. The main limitation of PCA is the linear combinations of original components’ dimensions, causing them to not be directly interpretable on the original input variables. There has been a long debate between the linear PCA methodology and the aggressive Neural networks methodology. Both methods apply data transformations on the original dataset; however, the PCA approach is more transparent and can be explained better in the context of domain knowledge, while the Neural algorithm is like Black Box but very powerful in fitting the dataset.

3. DATA ANALYSIS AND DERIVE HEALTH INDEX

3.1 PCA of Coffee/Tea Nutritions

We used JMPs Principal Components Analysis (PCA) platform across all factors in the dataset (e.g. 'Sugars', 'Protein', 'Caffeine'), where 66.4% and 12.6% of variation were attributable to Principal Components 1 (Prin 1) and 2 (Prin 2), respectively. Therefore, 79% of the total variation was explained the first two Principal Components out of 10 Original Components. This PCA has aligned well with 20%-80% Pareto Concept. In Figure 3, Bi-plot of Component 2 vs Component 1 is shown. The length in radial directions are contributing to the Eigenvalues. The orientation in angular directions are related to Eigenvectors. The Principal Component 1 is more relevant to unhealthy nutrition components such as Sugars, Total Carbohydrates, Calories, Saturated Fat, total Fat… especially, the Calories and Saturated Fat are almost oriented in the Component 1 Axis. It may indicate that one major marketing segment is on the higher-energy Coffee/Tea product. The Principal Component 2 is mainly consisted of healthy Caffeine, Protein and Dietary Fiber nutritional constituents. The Caffeine constituent is the most opposite against the Sugars and total Carbohydrates. It may indicate the Coffee/Tea makers may differentiate their product flavors in order to attract customers from both segments: Healthy and Tasty. Like eating Dark Chocolate, some Consumers may drink Black Coffee and Black Tea.

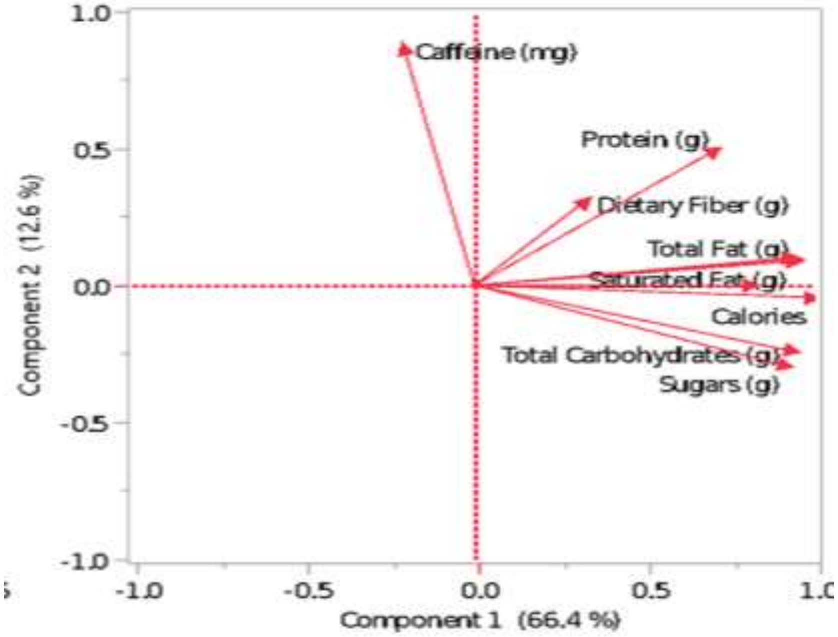


Figure 3: Biplot of Coffee/Tea Nutritions [8]

3.2 Derive Science-Heath Index and PCA-Health Index

The Science-Health Index was developed based on the basis of each of the factors, taking into account the scientific research and applying weighting coefficients with a positive or negative sign depending on whether each factor contributed to (positive) or was detrimental to (negative) CVD prevention. Therefore, the higher health index the more beneficial in terms of heart disease prevention, and the lower the index, the less beneficial. After coefficients were assigned to each variable and an equation was developed, the health index value was calculated for each drink. The Coffee/Tea Science-Health Index is shown below:

Science-Health Index =-2 * Calories + -2 * "Total Fat (g)" + -2 * "Saturated Fat (g)" + -2 *"Cholesterol (mg)" + -2 * "Sodium (mg)" + -1 *"Total Carbohydrates (g)") + 2 * "Dietary Fiber (g)" + -2 *"Sugars (g)") + 1 * "Protein (g)" + 2 * "Caffeine (mg)"

The PCA-Healthy Index can also be calculated based on the PCA Eigenvectors. To avoid the larger variance bias effects, *Z*-standardization transformation was conducted on the original data. Prin 1 and Prin 2 Eigenvectors shown in Figures 4 and 5. are derived from JMP PCA analysis. PCA-Health Index= -Eigenvalue 1* Prin 1 Eigenvector + Eigenvalue 2* Prin 2Eigenvector.

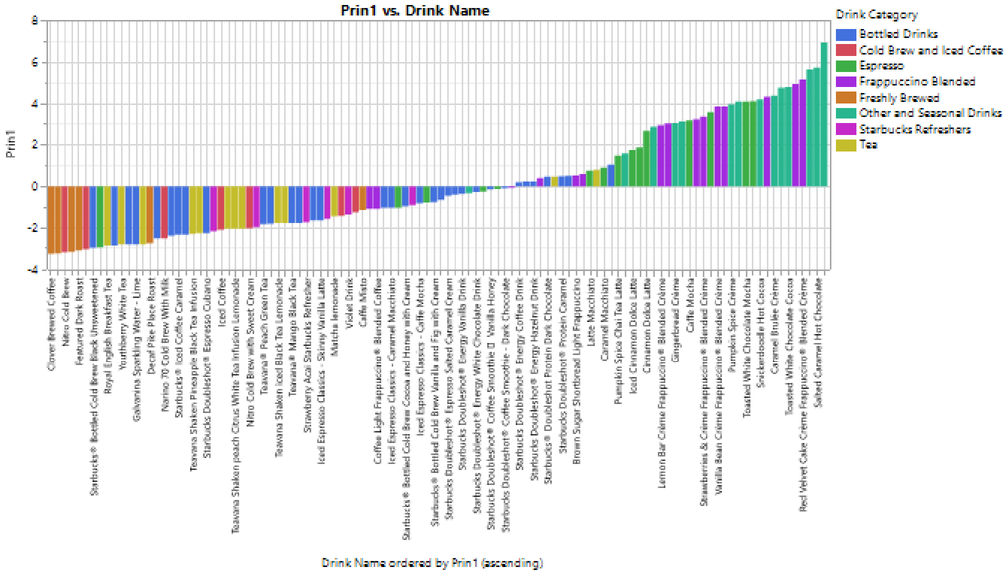


Figure 4: 1st Principal Component of Coffee/Tea Nutritions

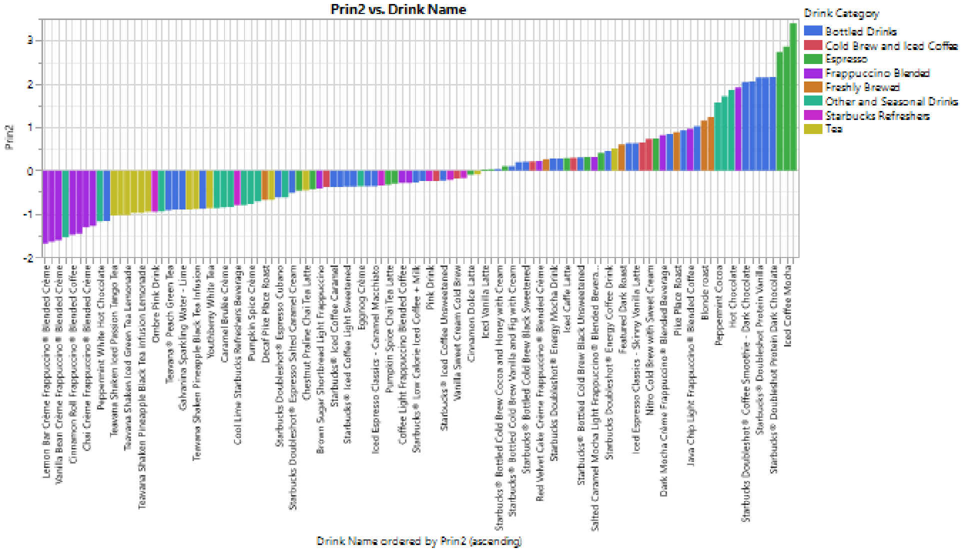
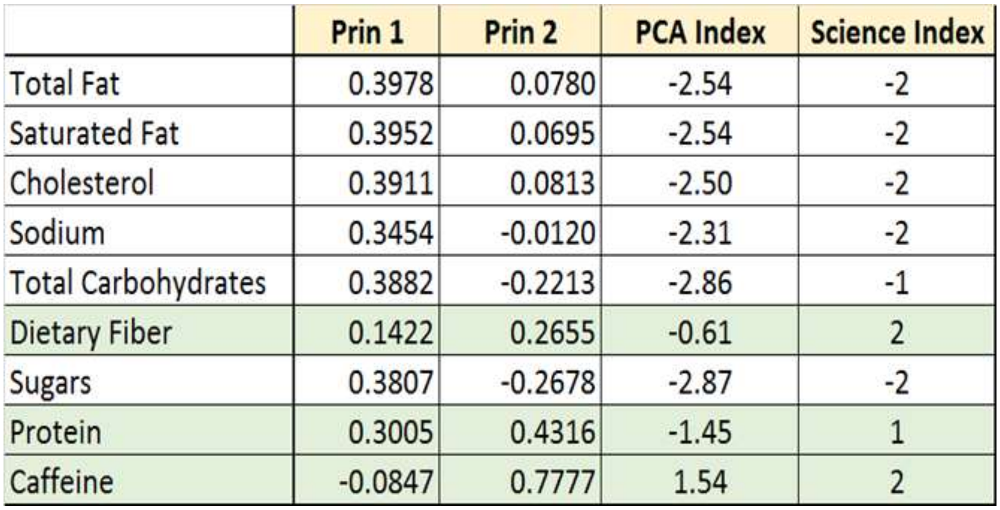


Figure 5: 2nd Principal Component of Coffee/Tea Nutritions

The relative ranking between two Health Index Methods are significantly different each other as shown in Table 1. The first approach is based on Scientific Literature Research and the second approach is based on empirical PCA limited to Starbucks Product. Even two Health Index Methods have shown different formulas, the follow-up model validation can be further conducted in larger sampling and wider product types. Though, this is out of scope of this paper. The Health Index derived in this paper is still very subjective without direct research or clinical test validation.

Table 1: Compare PCA-Health Index vs. Science-Health Index [8]



	Prin 1	Prin 2	PCA Index	Science Index
Total Fat	0.3978	0.0780	-2.54	-2
Saturated Fat	0.3952	0.0695	-2.54	-2
Cholesterol	0.3911	0.0813	-2.50	-2
Sodium	0.3454	-0.0120	-2.31	-2
Total Carbohydrates	0.3882	-0.2213	-2.86	-1
Dietary Fiber	0.1422	0.2655	-0.61	2
Sugars	0.3807	-0.2678	-2.87	-2
Protein	0.3005	0.4316	-1.45	1
Caffeine	-0.0847	0.7777	1.54	2

3.3 Compare Different Foods Products

Four different Foods products were compared based on the Eigen Plot as shown in Figure 6. The health nutrition components are highlighted in the green box and the unhealthy nutrition components are highlighted in the red box. The relative Green-Red distribution patterns may indicate which Foods product family is relatively healthier in general since more nutrition variances are plotted based on the first two principal components. Among four products, Dark Chocolate seems has most the healthy nutrition components along the 1st Principal Component (positive direction). While the other unhealthy nutrition components are either oriented in the opposite of the 1st component or in the 2nd principal component. Coffee/Tea products have most unhealthy nutrition components near the 1st Principal Component which may indicate most Starbucks Coffee/Tea products are not healthy. The Eigen Plot nutrition pattern may help judge which foods product is considering to segment their product types based on the healthy nutrition factors.

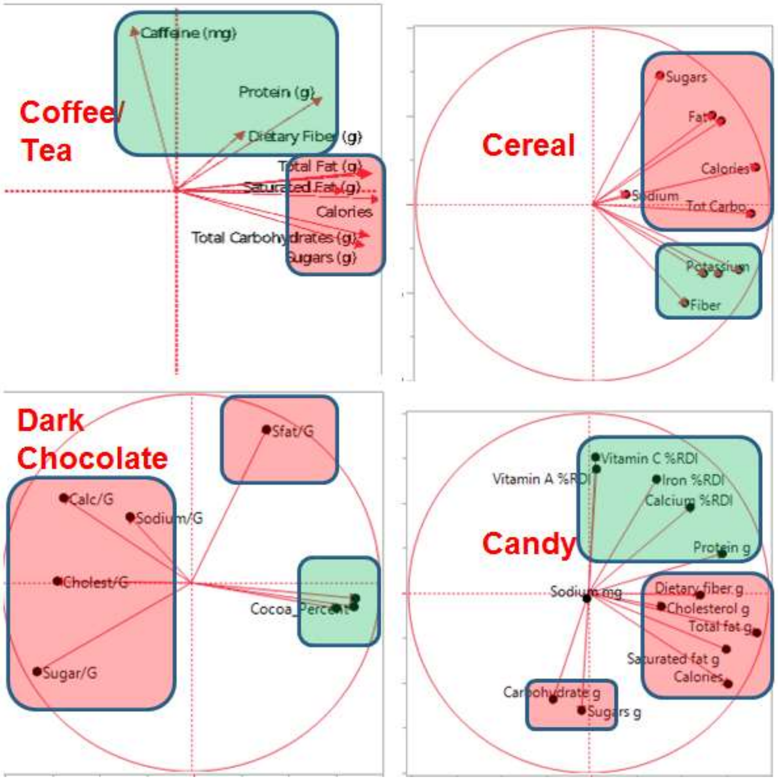


Figure 6: Eigen Plot of four different Foods Products

4. DATA ANALYSIS AND DERIVE HEALTH INDEX

This paper has utilized the Principal Component Eigen Analysis to study the Coffee and Tea nutritions. The first two principal components have contributed to total 79% variance based on the top two Eigenvalues. The first principal component is attributed to the unhealthy nutritions such as Sugars, Total Fat… The second principal component is related to the healthy nutritions such as Caffeine and Dietary Fiber. Two health indexes are derived: (1) by scientific research, and (2) by PCA method. Two methods have about 70%-80% correlation. Eigen Plot can judge the Foods Product on the health situation. The PCA method has shown a great potential to help conduct scientific research.

Acknowledgements

Authors would like to thank Anna Wu for collecting the Coffee/Tea Nutrition Data.

References

Chen, Mason, (2018 July) "Multivariate Statistics of Antioxidant Chocolate", IWSM Bristol Proceedings, Vol 2 37-40

Chen, Mason, (2018 July), "Choose Healthy Chocolate", IEOM Europe Proceedings, 434-441

Wu, Anna Dong. "Starbucks and Cardiovascular Disease Prevention." IEOM, IEOM Society, 26 July 2018, Corliss, J., Eating Too Much Added Sugar Increases the Risk of Dying with Heart Disease, Harvard Health Publishing, 201402067021 February 6, 2014.

Ding, M., Bhupathiraju, S., Satija, A., van Dam, R., & Hu, F., Long-Term Coffee Consumption and Risk of Cardiovascular Disease: A Systematic Review and a Dose-Response Meta-Analysis of Prospective Cohort Studies, Circulation, vol. 137, no. 13, pp. 31, 2018.

Golub, G.H., Kahan, W. (1965), “Calculating the singular values and pseudo-inverse of a matrix,” Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis 2:2, 205–224.

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.