Starbucks Coffee Statistical Analysis

Anna Wu
Mission San Jose High School
Fremont, CA 94539, USA
anna.dong.wu@gmail.com

Abstract

The purpose of this STEM project is to determine which Starbucks drinks among all coffee and tea options are best for cardiovascular disease (CVD) prevention. In order to do this, a health index was constructed considering different variables, including: saturated fat, cholesterol, sodium, carbohydrates, dietary fiber, sugars, protein, and caffeine. Each variable was assigned a weighting coefficient, with lower coefficients assigned to the factors that are more harmful and higher ones to those that are more beneficial. Therefore, drinks with the highest health index are determined to be the most beneficial to preventing CVD. Principal Components 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. PCA was successfully able to decompose the dominant sources of variability in relation to the Health Index, 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 on the basis of the first two Principal Components. Prin 1 did a good job grouping the data, separating Frappuccino Blended and Espresso beverages in one cluster, and mainly Cold Brew, Freshly Brewed, and Tea in another. Prin 2 largely grouped data based on cholesterol and fat content, and held less explanatory power than Principal Component 1. The health index originally derived on the basis of the scientific research, largely corroborated the results of PCA 1 vs. Drink/ Drink Category. Hierarchical Clustering was used to form 3 clusters across drink categories, and results were taken together with the Health Index/ PCA to investigate which combined set of factors contributed most to CVD prevention. This project sheds light on smarter ordering at Starbucks, making people more aware of how diet ultimately affects health and more specifically, how smart drink choices can promote CVD prevention.

Keywords
STEM, Starbucks Coffee, cardiovascular disease

1. Introduction and Literature Research

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 has also been associated with a diminished risk of CVD. 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. Antioxidant activity of flavonoids reduce free radical formation and scavenge free radicals, 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. Blood vessels absorb and deposit cholesterol, which may initiate the formation of an atherosclerotic lesion, causing blood vessel blockage. Coffee and tea provide an abundant amount of antioxidants that reduce oxidative stress that can damage cells. 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. 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.

2. 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 and sorted. After the defects are removed, coffee production is complete. In tea production, tea leaves are first plucked and laid into troughs. 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.

3. Data Collection

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, and Teas. 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.

3.1 Collect Data

Table 1. Espresso Data (15 drinks randomly selected)

Drink Name	Calories	Total Fat (g)	Saturated Fat (g)	Cholesterol (mg)	Sodium (mg)	Total Carb(g)	Dietary Fiber (g)	Sugars (g)	Protein (g)	Caffeine (mg)
Iced Cinnamon Dolce Latte	290	12	8	40	115	39	0	36	8	150
Pumpkin Spice Chai Tea Latte	290	4.5	2.5	20	180	54	0	53	10	95
Cappuccino	120	4	2	15	100	12	0	10	8	150
Toasted White Chocolate Mocha	420	15	9	45	380	58	0	53	13	150
Skinny Mocha	160	1.5	1	5	140	24	4	15	14	150
Caramel Macchiato	250	7	4.5	25	150	35	0	33	10	150
Iced Vanilla Latte	190	4	2	15	100	30	0	28	7	150
Iced White Chocolate Mocha	420	20	13	60	200	50	0	49	11	150

Drink Name	Calories	Total Fat (g)	Saturated Fat (g)	Cholesterol (mg)	Sodium (mg)	Total Carb(g)	Dietary Fiber (g)	Sugars (g)	Protein (g)	Caffeine (mg)
Iced Coffee Mocha	350	17	11	55	100	39	4	30	10	175
Cinnamon Dolce Latte	340	13	9	50	160	44	0	41	12	150
Iced Caramel Brulee Latte	420	15	9	55	210	65	0	49	9	150
Latte Macchiato	220	11	6	35	150	19	0	17	12	225
Caffe Mocha	360	15	9	50	150	44	4	35	13	175
Iced Caffe Americano	15	0	0	0	15	3	0	0	1	225
Iced Caffe Latte	130	4.5	2.5	20	115	13	0	11	8	150

4. Analysis and Results

After collecting all data, it was analyzed by constructing a health index and running Principal Component Analysis (and Clustering) to determine the best drinks for CVD prevention. Analysis was performed using JMP 13 Software (© 2017 SAS Institute, Inc.)

4.1 Health Index Analysis

The Health Index was developed 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 theindex, the less beneficial.

Table 2. Health Index Coefficients

Category (Factor)	Description	Health Index Coefficient
Calories	Calorie intake should match the amount of calories burned each day to help reduce the chance of gaining too much weight which is associated with CVD.[8]	-2
Total Fat	High intake of fats tends to increase susceptibility to CVD.[6]	-2
Saturated Fat	Higher intakes of the most common saturated fats are associated with a boost in the risk of coronary artery disease of up to 18%. Replacing just 1% of those fats with the same amount of calories from polyunsaturated fats or plant proteins is associated with a 6% to 8% lower risk.[6]	-2
Cholesterol	Cholesterol builds up in the walls of the arteries, causing them to become more narrow and slow blood flow. This can cause atherosclerosis (the building of calcium and plaques in the arteries).[1]	-2
Sodium	Sodium increases blood pressure. Hypertension is a major risk factor for heart attacks, stroke, and other cardiovascular problems.[10]	-2

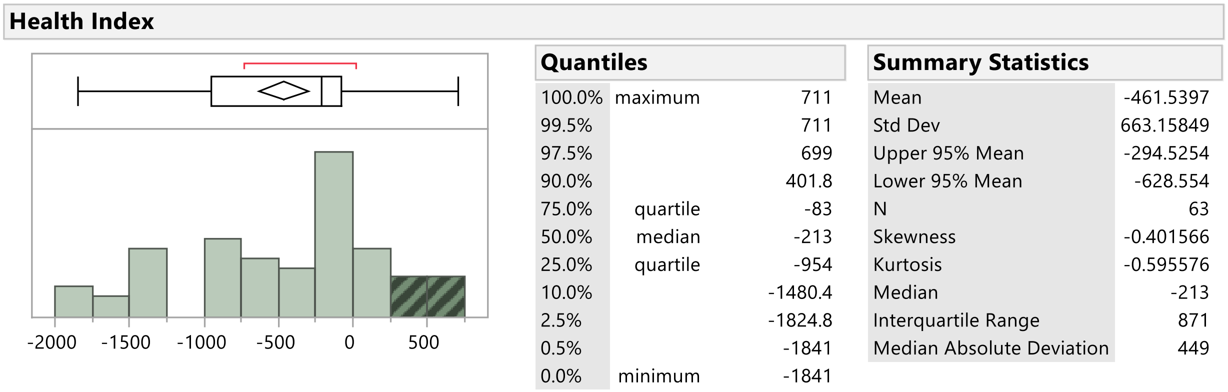
Category (Factor)	Description	Health Index Coefficient
Carbohydrates	Excessive carbohydrate intake is primary dietary factor that is bad for heart health.[2]	-1
Dietary Fiber	Dietary fiber from whole grains, as part of an overall healthy diet, may help improve blood cholesterol levels, and lower risk of heart disease, stroke, obesity and type 2 diabetes.[12]	+2
Sugars	Sugar-sweetened beverages can raise blood pressure and can stimulate the liver to dump more harmful fats into the bloodstream, which are both known to reduce heart health.[1]	-2
Proteins	Nutrients in low-fat protein can help lower cholesterol and blood pressure and help maintain a healthy weight. By choosing these proteins over high-fat meat options, risk of heart attack and stroke decreases.[3]	+1
Caffeine	Moderate coffee consumption was inversely significantly associated with CVD risk, with the lowest CVD risk at 3 to 5 cups/day, and heavy coffee consumption was not associated with elevated CVD risk.[4]	+2

After coefficients were assigned to each variable and an equation was developed, the health index value was calculated for each drink.

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

Drinks were first plotted on a histogram with summary statistics using JMP's Distribution Platform. Then, in order to interpret all drinks (and drink categories) on the basis of the same scale, a Z-transformation was applied to the distribution of Health Index, resulting in a Standardized Index, again plotted in the Distribution Platform with summary statistics. Note that the Z-transformation simply takes each individual value subtracts off the mean of all values, and divides by the standard deviation of all values.

The standardization revealed the 9 healthiest drinks according to Health Index Rating, as listed below, where they were, generally: freshly brewed, cold brew and iced coffees and had among the lowest calories, fat, saturated fat, cholesterol and sugar content, as well as the highest protein content with moderate to higher caffeine content.



Health Index

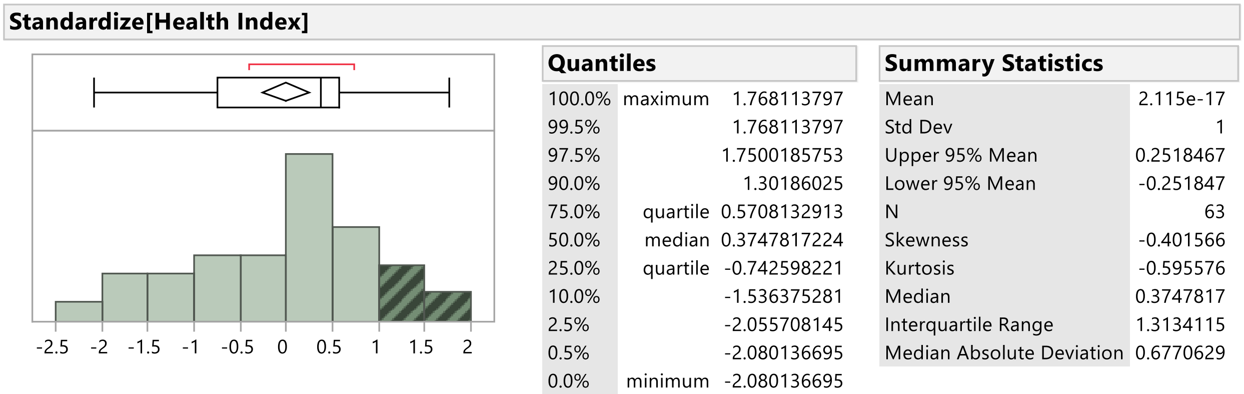
Quantiles

100.0%	maximum	711
99.5%		711
97.5%		699
90.0%		401.8
75.0%	quartile	-83
50.0%	median	-213
25.0%	quartile	-954
10.0%		-1480.4
2.5%		-1824.8
0.5%		-1841
0.0%	minimum	-1841

Summary Statistics

Mean	-461.5397
Std Dev	663.15849
Upper 95% Mean	-294.5254
Lower 95% Mean	-628.554
N	63
Skewness	-0.401566
Kurtosis	-0.595576
Median	-213
Interquartile Range	871
Median Absolute Deviation	449

Figure 1a. Non-Standardized Health Index (before Z-transformation)

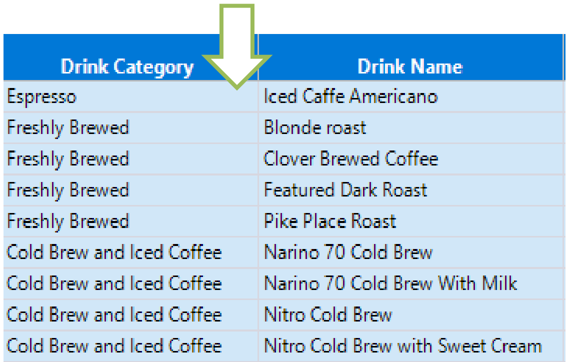


Standardize[Health Index]

Quantiles		
100.0%	maximum	1.768113797
99.5%		1.768113797
97.5%		1.7500185753
90.0%		1.30186025
75.0%	quartile	0.5708132913
50.0%	median	0.3747817224
25.0%	quartile	-0.742598221
10.0%		-1.536375281
2.5%		-2.055708145
0.5%		-2.080136695
0.0%	minimum	-2.080136695

Summary Statistics	
Mean	2.115e-17
Std Dev	1
Upper 95% Mean	0.2518467
Lower 95% Mean	-0.251847
N	63
Skewness	-0.401566
Kurtosis	-0.595576
Median	0.3747817
Interquartile Range	1.3134115
Median Absolute Deviation	0.6770629

Figure 1b. Standardized Health Index (after Z-transformation)

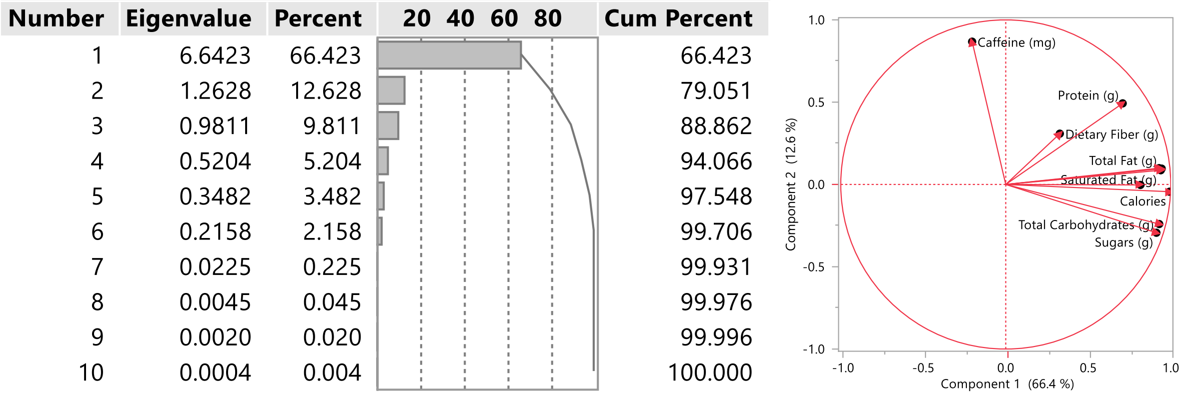


Drink Category	Drink Name
Espresso	Iced Caffe Americano
Freshly Brewed	Blonde roast
Freshly Brewed	Clover Brewed Coffee
Freshly Brewed	Featured Dark Roast
Freshly Brewed	Pike Place Roast
Cold Brew and Iced Coffee	Narino 70 Cold Brew
Cold Brew and Iced Coffee	Narino 70 Cold Brew With Milk
Cold Brew and Iced Coffee	Nitro Cold Brew
Cold Brew and Iced Coffee	Nitro Cold Brew with Sweet Cream

Figure 1c. 9 healthiest drinks from distribution selection in JMP of Standardized Health Index

4.2 Principal Components Analysis

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 on the basis of Principal Components 1 (Prin 1) and 2 (Prin 2), respectively. Therefore, 79% of the total variation was explained the first two Principal Components. Prin 1 and Prin 2 were then saved as columns and charted in the Graph Builder Platform in JMP.



Number	Eigenvalue	Percent	Cum Percent
1	6.6423	66.423	66.423
2	1.2628	12.628	79.051
3	0.9811	9.811	88.862
4	0.5204	5.204	94.066
5	0.3482	3.482	97.548
6	0.2158	2.158	99.706
7	0.0225	0.225	99.931
8	0.0045	0.045	99.976
9	0.0020	0.020	99.996
10	0.0004	0.004	100.000

Figure 2. Principal Components Eigenvalues (Variance Decomposition) and Bi-plot of Component 2 vs 1

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. The formulates for Prin 1 and Prin 2, and the graphs of their values by drink type and category (generated in Graph Builder), are shown in the analysis below:

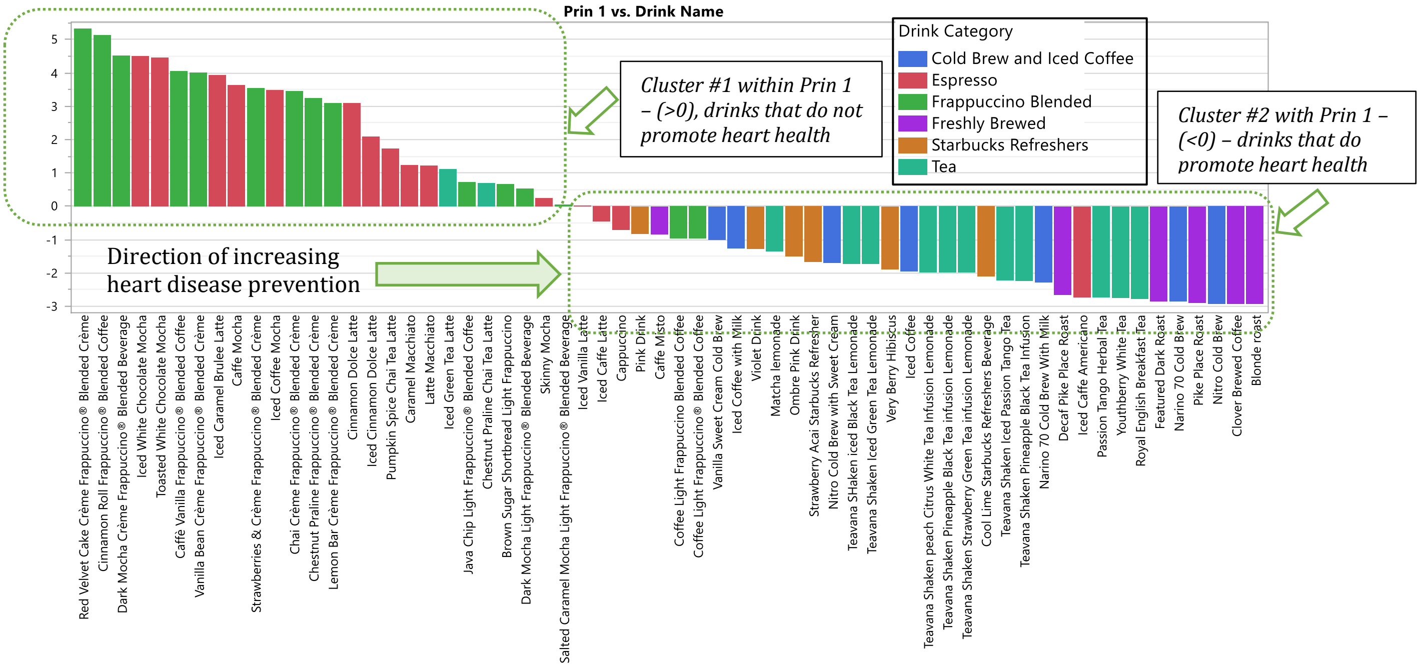


Figure 3a. Bar Chart of Principal Component 1 (sorted in descending order)

*Prin 1= 0.00263 *"Calories"+ 0.05689*"Total Fat (g)" + 0.09036* "Saturated Fat (g)"+ 0.016585* "Cholesterol (mg)"+ 0.00284*"Sodium (mg)" + 0.016667*"Total Carbohydrates (g)"+ 0.118207* "Dietary Fiber (g)"+ 0.01730*"Sugars (g)" + 0.06479*"Protein (g)"+ -0.0008423* "Caffeine (mg)"+ (-2.736)*

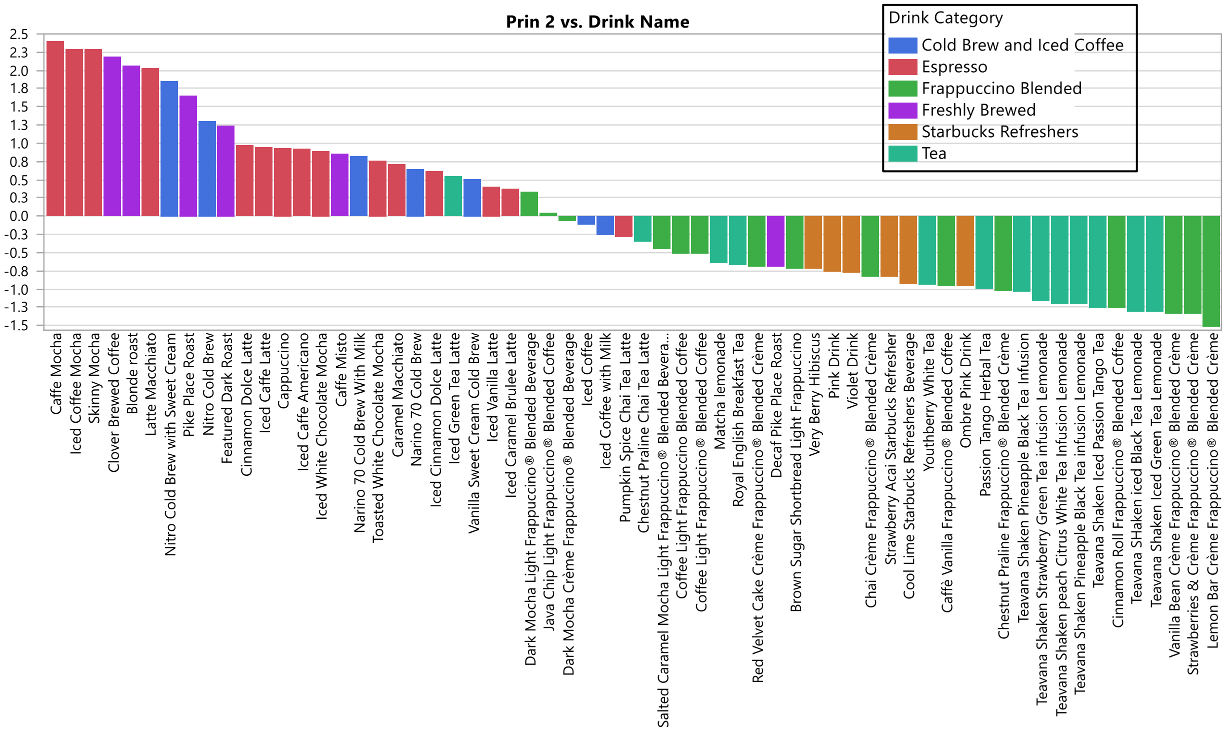


Figure 3b. Bar Chart of Principal Component 2 (sorted in descending order)

Prin 2 = -0.000268*Calories + 0.01328 * Total Fat (g) + 0.01904* Saturated Fat (g) + 0.004056 * Cholesterol (mg) + -0.0000237* Sodium (mg) + -0.00989* Total Carbohydrates (g) + 0.25534*Dietary Fiber (g) + -0.012747* Sugars (g) +0.10329* Protein (g) + 0.0082256* (Caffeine (mg)) + (-0.997149)

Principal Component 1 does a very good job at separating the drinks in a similar way as the scientifically-based Health Index, grouping drinks as indicated by Category for the most part per the Health Index (note in the case of the Health Index – higher scores indicate that drinks promote heart disease prevention). Principal Component 2 provides a different grouping result, largely because it groups drinks by category more so on the basis of cholesterol and fat content. Since Principal Component 2 has less explanatory power than Prin 1, (it only accounts for 12.6% of the overall variance in grouping of all variables, whereas Prin 1 accounts for over 66.4%), it’s not surprising that it doesn’t provide a clear differentiation among drinks and drink categories since it accounts for far less variation. In fact, Prin 2 accounts for 81% less total explained variation compared to Prin 1.

$$\%\ decrease, explained\ variation = \frac{Variace\ Explained_{Prin\ 1} - Variace\ Explained_{Prin\ 2}}{Variace\ Explained_{Prin\ 1}} = \frac{66.4 - 12.6}{66.4} = 81\%$$

Principal Component Analysis shows that Freshly Brewed, Cold and Ice Brewed Coffees, Teas, but only a few Espressos tend to be more beneficial in terms of heart disease prevention.

4.3 Clustering Analysis

We used JMPs Hierarchical Clustering platform (Ward method) to produce a Dendrogram, which is a factor tree plot which allows for the dynamic selection of the clusters; in this case, we selected 3-clusters to make the analysis more interpretable, and where the same input variables were considered as previously.

Clustering is a technique of grouping rows together that share similar values across a number of variables. The clusters were then colored in the dendrogram and across all rows of the data table (again for ease of interpretation), and the results output for the dendrogram is produced below, where Clusters 1, 2, and 3 are red, green, and blue, respectively.

Finally, we constructed a Parallel Plot separating the clusters – using JMPs Graph Builder platform, where all input variables were plotted on X, and the parallel plot selection was chosen in Graph Builder. Cluster #1 groups drinks which tended to span the range of calories, total fat, saturated fat, cholesterol, sugars and tended to be at the extreme ends of dietary fiber. Also, drinks grouped in this cluster had moderate to high levels of protein and moderate levels of caffeine.

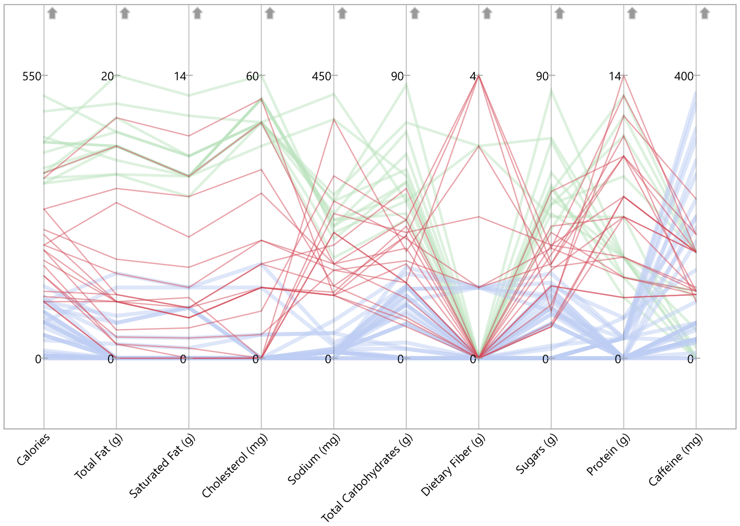


Figure 4a. Parallel Coordinate Plot, Cluster #1 selected

Drinks grouped in Cluster #2 tended to have have higher levels of calories, total fat, saturated fat, cholesterol, sodium, carbohydrates, sugars and protein, no dietary fiber, moderate/high protein, and low to moderate caffeine levels.

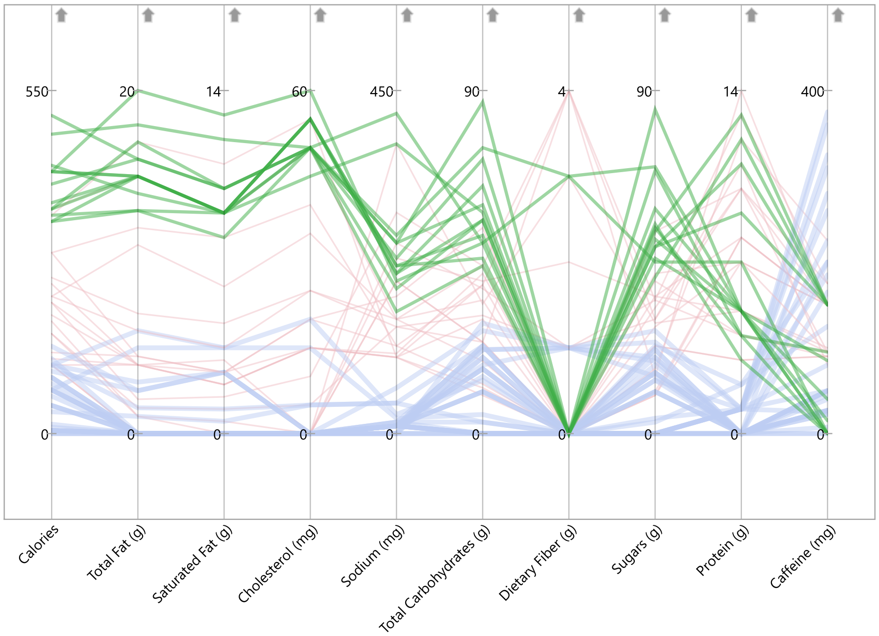


Figure 4b. Parallel Coordinate Plot, Cluster #2 selected

Drinks grouped in Cluster #3 tended to have low to moderate amounts of calories, total fat, saturated fat, cholesterol, sodium, and carbohydrates, dietary fiber, moderate sugar levels, moderate to high protein levels, and generally moderate to high caffeine levels.

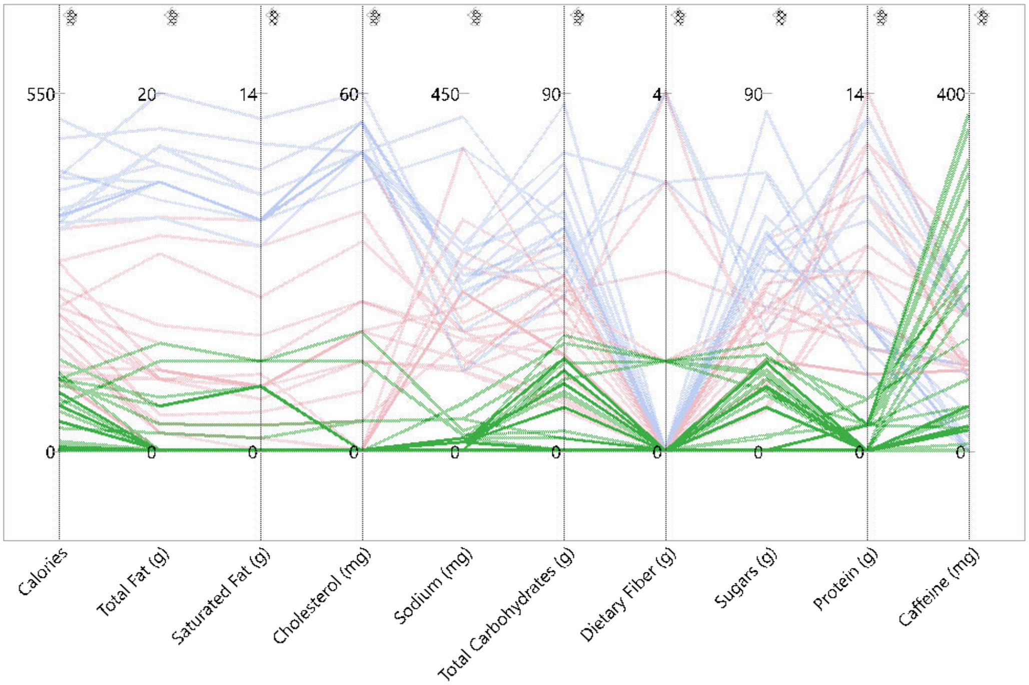


Figure 4. Cluster #2

Drinks grouped in Cluster #3 tended to have varying amounts of calories, total fat, saturated fat, cholesterol, sodium, and carbohydrates, low and high dietary fiber, moderate sugar levels, moderate to high protein levels, and generally moderate caffeine levels.

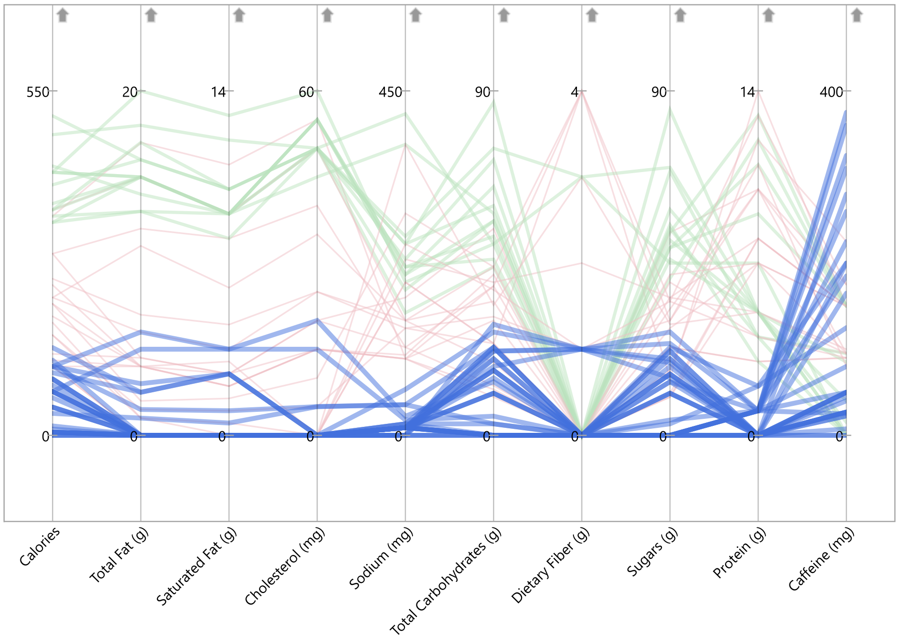


Figure 4c. Parallel Coordinate Plot, Cluster #3 selected

To complete the clustering analysis, we used Graph Builder, this time to group the average health index by Drink Category across each of the cluster levels, and to further explore the degree of association between the health index (standardized; developed previously) and hierarchical cluster level. Results of clustering analysis show that Freshly Brewed, Cold & Ice Brewed Coffees, Teas, Espressos, and Starbucks Refreshers tend to be more beneficial in terms of heart disease prevention.

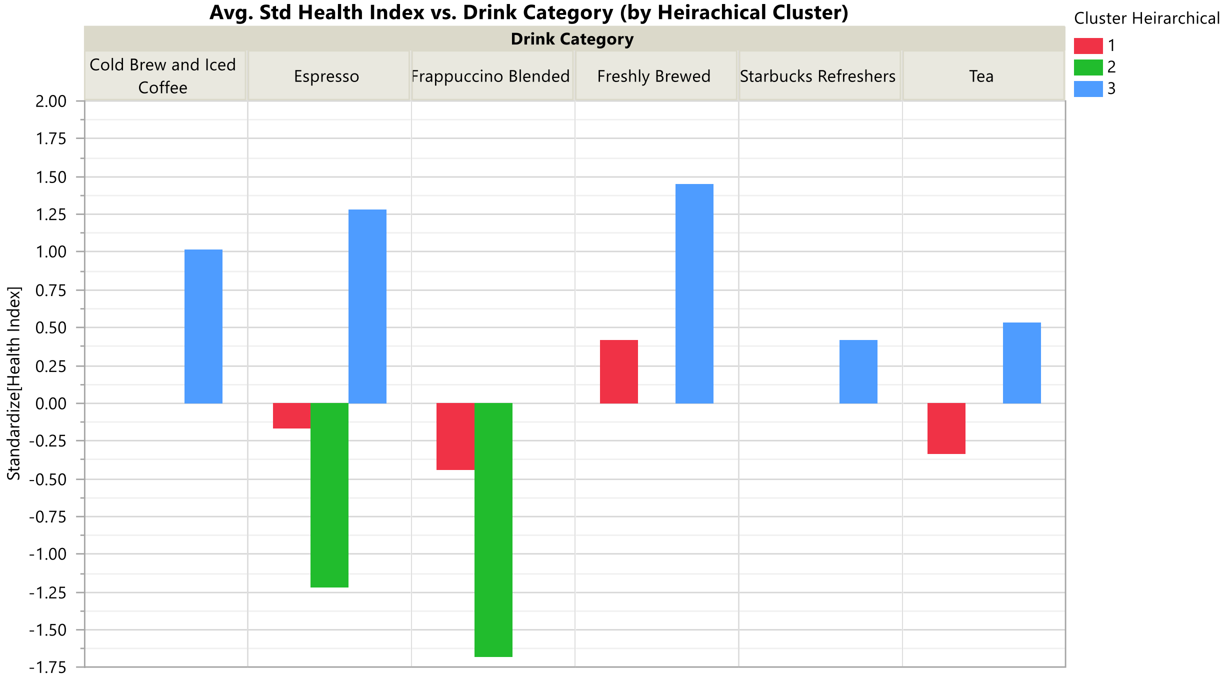


Figure 6. Avg. Std Health Index vs. Drink Category (by Hierarchical Cluster)

The results demonstrate that, on average, drinks in Cluster #3 are grouped into the highest health index ratings, those in Cluster # 1 are grouped into moderate to lower health index ratings, and those grouped into Cluster #2 have the lowest ratings. What we observe makes sense practically, since Drinks in Cluster #1 have generally lower sugar, fat, and cholesterol levels, and moderate/high caffeine levels. We know based on scientific research that caffeine present in prepared coffee beverages is associated heart disease prevention[4], and we also know that diets high in cholesterol, fat, and sugar are associated with higher risk of heart disease[7], therefore the results of scientific research corroborate the computational results yielded by the clustering approach – where drinks in drink categories within Cluster #3 are the most likely to facilitate heart disease prevention, where those in Cluster #1 are less likely, and those in Cluster #2 are less likely still or may even be detrimental to heart health.

5. Conclusion

Overall, the results of all three analyses corroborate one other and point to the following drink categories at Starbucks as the best choices for heart disease prevention: Cold & Ice Brewed Coffees, Freshly Brewed Coffees, and Starbucks Refreshers and Teas. Espressos are also a good choice only if they are lower in sugar, cholesterol, etc.

Next steps for research include a more granular assessment by specific drinks within each category, and a more precisely weighted (in terms of relative magnitude and mathematical structure of the coefficients) health index model. Also, we can plan to include additional factors in the analysis which science has suggested are correlated with heart disease prevention, such as antioxidants, flavonoids, and other substances present in coffee beverages.

Acknowledgements

Author would like to thank Patrick Guiliano for his mentorship and guidance throughout this entire project.

References

1. Cholesterol and Heart Disease, Physicians Committee for Responsible Medicine, Available: https://pcrm.org/health/health-topics/cholesterol-and-heart-disease
2. Corliss, J., Eating Too Much Added Sugar Increases the Risk of Dying with Heart Disease, Harvard Health Publishing, Available: https://www.health.harvard.edu/blog/eating-too-much-added-sugar-increases-the-risk-of-dying-with-heart-disease-201402067021 February 6, 2014.
3. Curr, A., The Effects of Protein Intake on Blood Pressure and Cardiovascular Disease, John Hopkins Medical Institutions, Available: https://www.ncbi.nlm.nih.gov/m/pubmed/12544662/
4. 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.
5. Drinking Coffee Could Lead to a Longer Life, Scientist Says: Whether it's caffeinated or decaffeinated, coffee is associated with lower mortality, which suggests the association is not tied to caffeine, University of Southern California, Available: www.sciencedaily.com/releases/2017/07/170710172118.htm, July 10, 2017.
6. Harvard Researchers Renew Warnings About Saturated Fat and Heart Disease, Harvard Health Publishing, Available: https://www.health.harvard.edu/heart-health/harvard-researchers-renew-warnings-about-saturated-fat-and-heart-disease January, 2017.
7. Higher coffee consumption associated with lower risk of early death, European Society of Cardiology, Available: www.sciencedaily.com/releases/2017/08/170827101750.htm, August 27, 2017.
8. Han, X., Ren, J., Caloric Restriction and Heart Function: Is There a Sensible Link?, NCBI, Available: https://ww.ncbi.nlmnih.gov/pmc/articles/PMC4002317/
9. Jephcote, B., The More Carbs You Eat, the Higher the Risk of Heart Disease, Diabetes.co, Available: https://www.diabetes.co.uk/in-depth/carbs-higher-risk-heart-disease-states-leading-cardiologist-dr-salim-yusuf/ February 20, 2017.
10. Shmerling, R., Sodium Studies Blur the Picture on What is Heart Healthy, Harvard Health Publishing, Available: https://www.health.harvard.edu/blog/sodium-studies-blur-picture-heart-healthy-201408157366 August 15, 2014.
11. Three to Four Cups of Coffee a Day Linked to Longer Life, Science Daily, Available: https://www.sciencedaily.com/releases/2017/11/171122190659.htm November 22, 2017.
12. Whole Grains and Fiber, American Heart Association, Available: https://www.heart.org/HEARTORG/HealthyLiving/HealthyEating

Biographies

Anna Wu is a junior at Mission San Jose High School and is currently taking AP Statistics. Anna is familiar with IBM SPSS/Modeler software and is an active user of SAS JMP software for applications in Biology/ Chemistry.