THE 80 CEREALS DATASET: UNDERSTANDING THE CONNECTION BETWEEN NUTRITION AND RATINGS THROUGH REGRESSION ANALYSES AND CLUSTERING

Manasa Bathina, Gabe Kass-Johnson, Harsh Patel, Sonam Sharma, John Sun

1 INTRODUCTION

The origin of the dataset or the year of its creation could not be conclusively determined. The dataset suggests that people rate healthier cereals higher, and according to recent time series studies for 2005-2011, on the trends of sugar and sodium, and other markers of the healthiness of the cereal, manufacturers have changed the ingredients of their cereals to make them healthier. Presumably, the cereal dataset predates these time series, and it foreshadowed these trends from 2005-2011. The time series report their units in quantities of 100g each. We do not have the information to standardize our dataset and compare it with the data in the time series plots from the literature. As a result of this lack of information, we cannot perform a longitudinal study.

Even if our analysis shows that people prefer to buy cereals with fewer calories per serving, lower sodium, or low sugar, that does not guarantee they'll be healthier since we cannot control how many servings they eat and the effect of advertising which increases consumption. For example, Neyens et al, demonstrated using statistics that "image magnification" of logos printed on cereal caused increased consumption among the study participants aged five to six.

According to a study conducted by IRI, a market research firm, Americans consumed around 8.4 billion units of cereal manufactured by Kelloggs(K), General Mills(GIS), Post Holdings(POST) and Quaker. These companies respectively hold assets totaling to around 9 figures or more. We chose this data to understanding why some cereals are rated more highly than others, and what effect the nutritional value has on these preferences.

Bialkova et al (2014) document the "health-pleasure trade-off effect" where consumers believe that healthier foods must taste worse. Our study would test that effect.

To evaluate whether product rating has a correlation on sales, one further research may investigate whether Nabisco, the manufacturer that produced higher-rated cereals gained more revenue. The behavior of people ultimately decide their health, and manufacturers providing cereals with lower calories per serving only can play a minor role in people's health.

2 FIGURES FROM THE LITERATURE

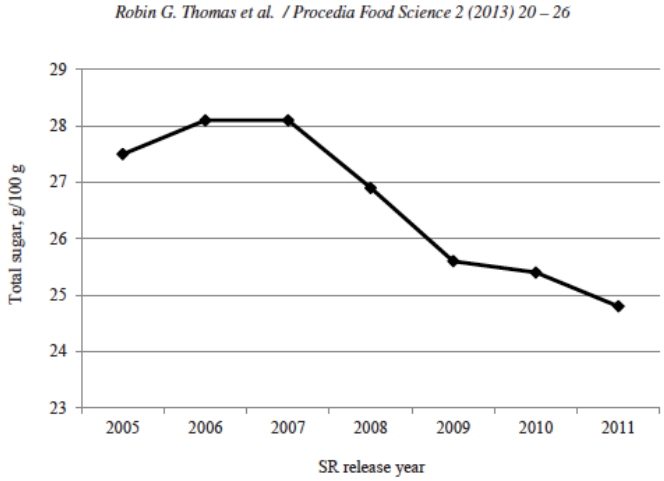


Fig.1. Trend for total sugar in RTE breakfast cereals. Mean values for General Mills and Kellogg cereals within USDA Standard Reference (SR) nutrient database.

[1]

Figure 1: Overall, the sugar decreased per year in non-monotonic fashion. This may suggest manufacturers experimented with the amount of sugar in cereal sugar to meet consumer preferences. The rate at which the manufacturer modified their ingredients (per month, per quarter) may give insight into other research questions including their quarterly stock prices and answer questions including whether increasing the amount of sodium in cereal by 1 gram decreased the sales of that cereal for that quarter of the year. The limitations of the figures shown include the fact that they do not have enough resolution to examine monthly changes in ingredients.

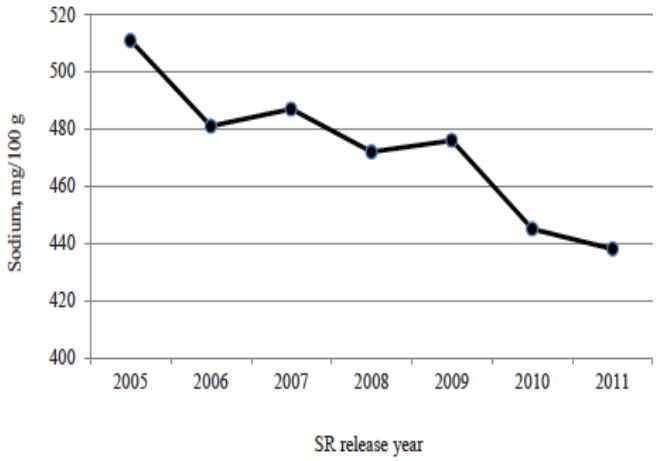


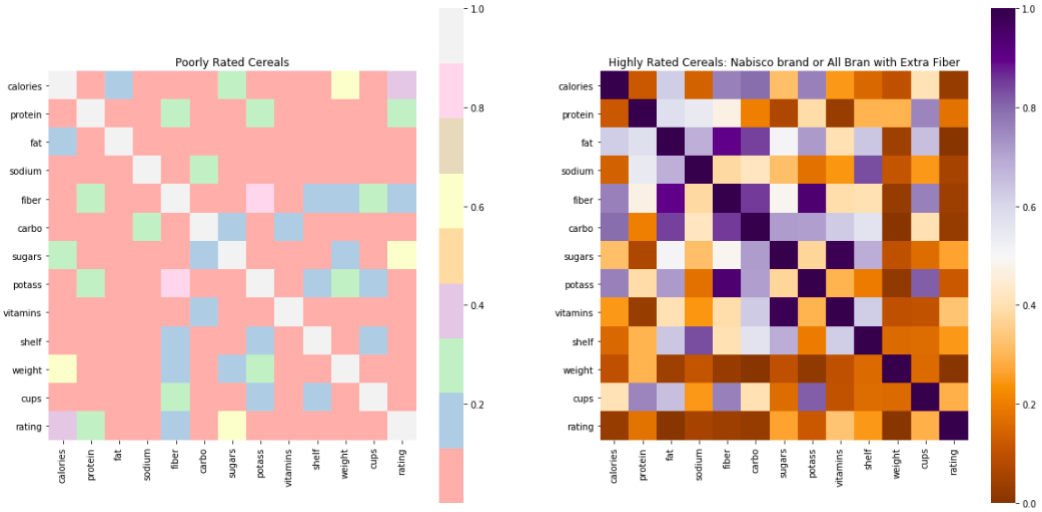
Fig.3. Trend for sodium in RTE breakfast cereals. Mean values for General Mills and Kellogg cereals within USDA Standard Reference (SR) nutrient database.

[2]

Figure 2: We infer that manufacturers experimented more with sodium content, perhaps as a result of the greater importance of sodium content in cereal and high sensitivity in consumer preferences for sodium. It would be riskier as a cereal business to put the wrong amount of sodium. So, it fluctuated but overall still decreased.

[1] Thomas (2013)
[2] Ibid

3 Heat Maps on Coefficient of Determination (R^2) Matrices



Overall the higher-ranked cereal from Nabisco had more close to one coefficient of determinations (R^2 than the lower-ranked cereals.

Figure 4: Nabisco was the manufacturer with the second-highest mean. The variables are more correlated than in the lowest tier rated cereals. The higher the shelf in the grocery store, the higher the sodium content, protein, fiber, sugar, vitamins.

Figure 5: Higher calorie, high fat, high sodium, high sugar, high vitamin, heavy, high cups cereal were rated lower. High protein, fiber, potassium cereals were rated higher. Fiber and potassium are positively correlated. Cereals' sodium level sorted by shelf. The negative correlation with calories on rating surprised us.

Protein, fiber, fat and sugar explain the ratings of the cereals.

4 Boxplots

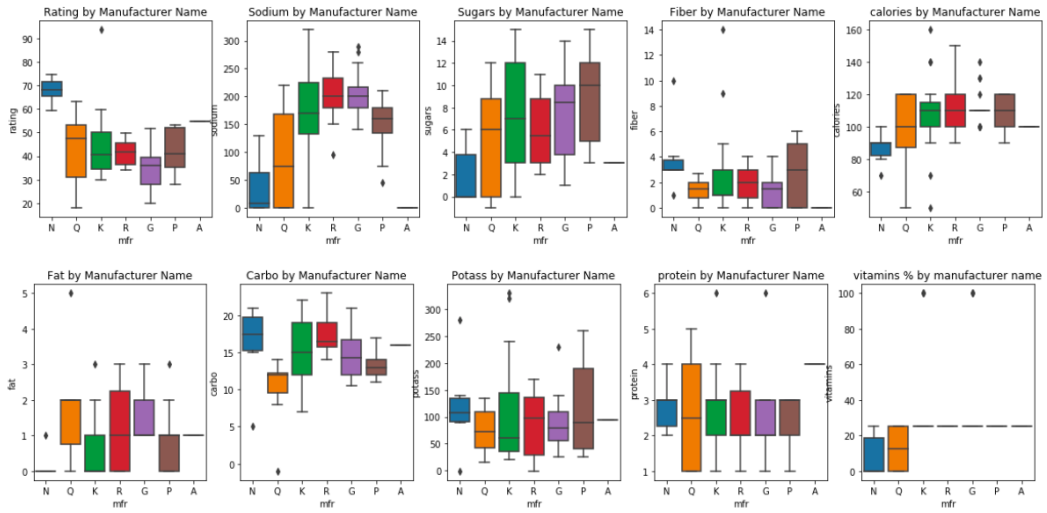


Figure 6: Bar charts comparing the features with the cereal manufacturers to see what the trends are.

According to a survey conducted by the 2010 dietary guidelines committee, there may be some amount of truth that shows that dietary fiber is somewhat essential for digestion and also reduces cardiovascular diseases, obesity and diabetes to some extent accounting for high fiber content cereals

gaining more ratings than other cereals. Strong medical evidence shows that lower sodium can lead to reduced blood pressure, hypertension, and kidney disease explaining the popularity of lower sodium cereals.

5 PCA-KMEANS

A medium amount of sodium doesn't ruin a cereal. Low fat, low carbs, medium protein, low cups/serving, low calories is associated with higher rating.

6 Gaussian Naive Bayes vs KNN vs SVM vs KNN vs Logistic Regression

Unsurprisingly, NB performed worse with PCA dimension reduced data.

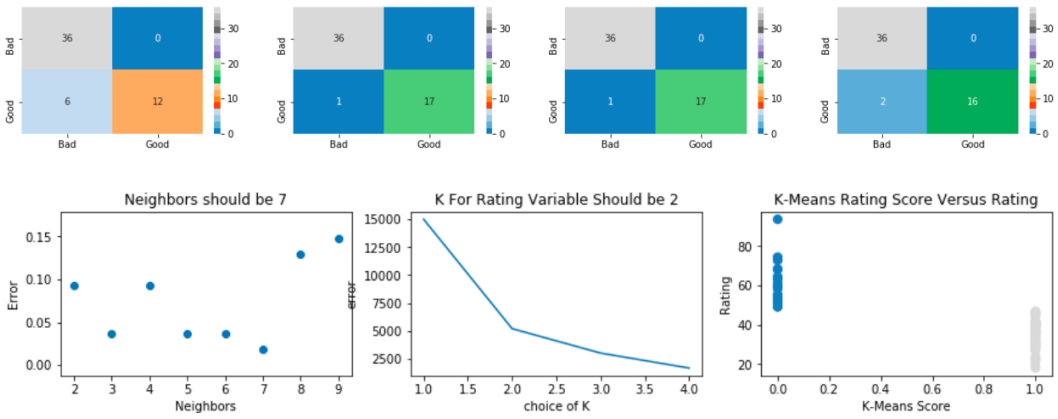


Table: Left to Right GSNB with not PCA processed features vs PCA processed KNN, Sigmoid Kernel 0.1 Soft Margin SVC, and Logistic Regression

7 K-MEANS BASED RECOMMENDATION SYSTEM

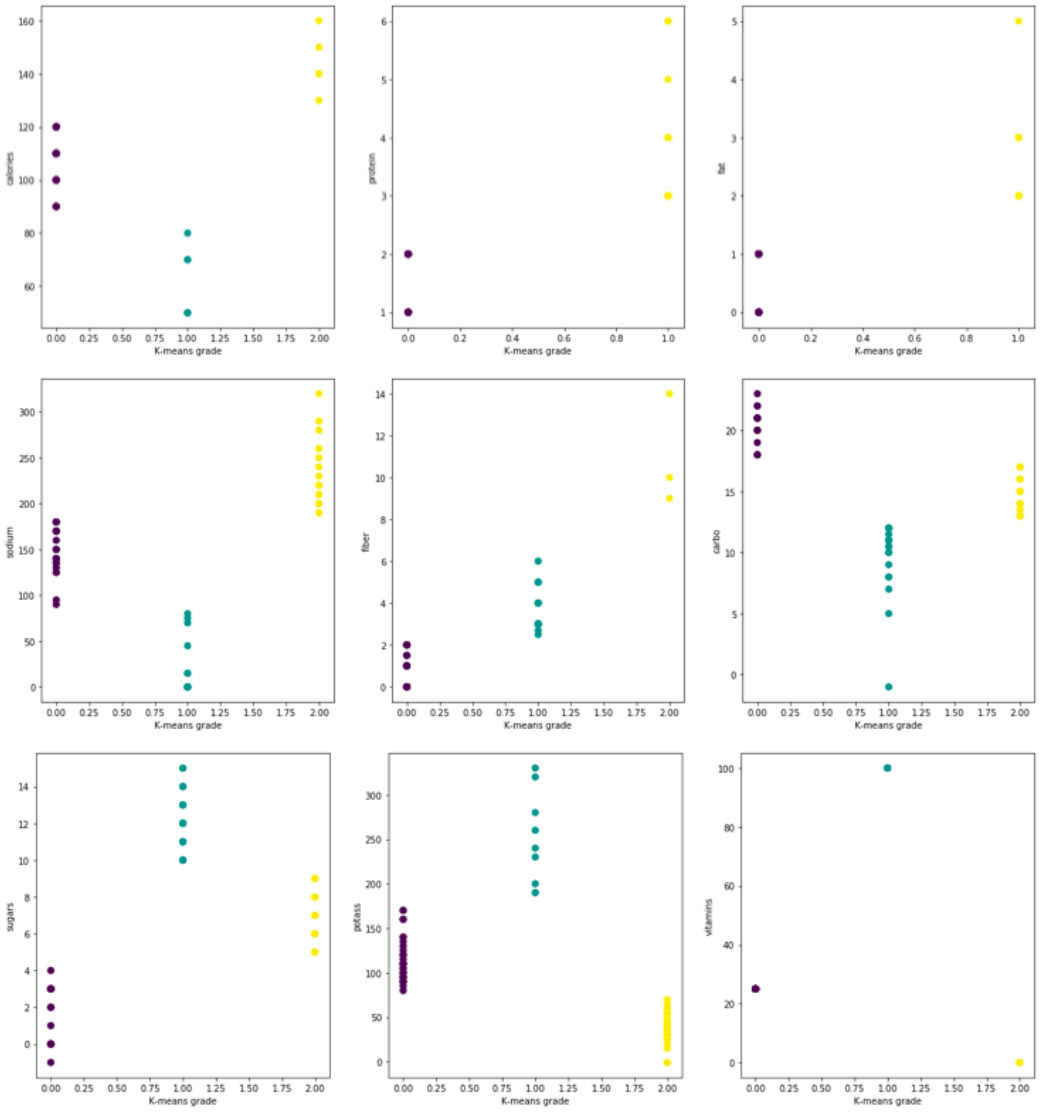


Figure 7:

Some variables have a higher variance than others indicating a greater diversity of preference among consumers.

We used the K-means data to develop a recommendation system by sorting cereals. We searched for cereals in the optimal clusters, i.e. lowest K-Means calories class, lowest K-Means sodium class, and the only cereal that met all criteria was All-Bran Extra Fiber cereal.

name	mfr	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	rating
All-Bran with Extra Fiber	K	50	4	0	140	14.0	8.0	0	330	25	93.704912

Figure 8: Healthy Across the Board

To simplify the analysis, we loosened the search criteria and only look for cereals in the clusters of the lowest sodium and sugar as we believed that people would consume equal the standard 2000/2500 calories, protein, carbohydrates daily even if their cereal had lower calories. In the conclusion of the K-Means section, in addition to recommending All-Bran Extra Fiber cereal, we recommend the following as alternatives based on their lower sodium and sugar content.

name	mfr	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	rating
Cream of Wheat (Quick)	N	100	3	0	80	1.0	21.0	0	-1	0	64.533816
Great Grains Pecan	P	120	3	3	75	3.0	13.0	4	100	25	45.811716
Maypo	A	100	4	1	0	0.0	16.0	3	95	25	54.850917
Puffed Rice	Q	50	1	0	0	0.0	13.0	0	15	0	60.756112
Puffed Wheat	Q	50	2	0	0	1.0	10.0	0	50	0	63.005645
Quaker Oatmeal	Q	100	5	2	0	2.7	-1.0	-1	110	0	50.828392
Shredded Wheat	N	80	2	0	0	3.0	16.0	0	95	0	68.235885
Shredded Wheat 'n'Bran	N	90	3	0	0	4.0	19.0	0	140	0	74.472949
Shredded Wheat spoon size	N	90	3	0	0	3.0	20.0	0	120	0	72.801787

Figure 9: Low Sugar and Sodium Cereals

name	mfr	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	rating
All-Bran with Extra Fiber	K	50	4	0	140	14.0	8.0	0	330	25	93.704912
Cheerios	G	110	6	2	290	2.0	17.0	1	105	25	50.764999
Corn Chex	R	110	2	0	280	0.0	22.0	3	25	25	41.445019
Corn Flakes	K	100	2	0	290	1.0	21.0	2	35	25	45.863324
Cream of Wheat (Quick)	N	100	3	0	80	1.0	21.0	0	-1	0	64.533816
Crispix	K	110	2	0	220	1.0	21.0	3	30	25	46.895644
Grape-Nuts	P	110	3	0	170	3.0	17.0	3	90	25	53.371007
Great Grains Pecan	P	120	3	3	75	3.0	13.0	4	100	25	45.811716
Kix	G	110	2	1	260	0.0	21.0	3	40	25	39.241114
Maypo	A	100	4	1	0	0.0	16.0	3	95	25	54.850917
Nutri-grain Wheat	K	90	3	0	170	3.0	18.0	2	90	25	59.642837
Product 19	K	100	3	0	320	1.0	20.0	3	45	100	41.503540
Puffed Rice	Q	50	1	0	0	0.0	13.0	0	15	0	60.756112
Puffed Wheat	Q	50	2	0	0	1.0	10.0	0	50	0	63.005645
Quaker Oatmeal	Q	100	5	2	0	2.7	-1.0	-1	110	0	50.828392
Rice Chex	R	110	1	0	240	0.0	23.0	2	30	25	41.998933
Rice Krispies	K	110	2	0	290	0.0	22.0	3	35	25	40.560159
Shredded Wheat	N	80	2	0	0	3.0	16.0	0	95	0	68.235885
Shredded Wheat 'n'Bran	N	90	3	0	0	4.0	19.0	0	140	0	74.472949
Shredded Wheat spoon size	N	90	3	0	0	3.0	20.0	0	120	0	72.801787
Special K	K	110	6	0	230	1.0	16.0	3	55	25	53.131324
Total Corn Flakes	G	110	2	1	200	0.0	21.0	3	35	100	38.839746
Total Whole Grain	G	100	3	1	200	3.0	16.0	3	110	100	46.658844
Triples	G	110	2	1	250	0.0	21.0	3	60	25	39.106174
Wheat Chex	R	100	3	1	230	3.0	17.0	3	115	25	49.787445
Wheaties	G	100	3	1	200	3.0	17.0	3	110	25	51.592193

Figure 10: Low Sugar Cereals

In reality, people on a low sodium diet would suffer from low iodine, and to make a fair recommendation, we only need to look for cereals in the lowest tier of sugar.

8 NAIVE-BAYES VS RANDOM FOREST

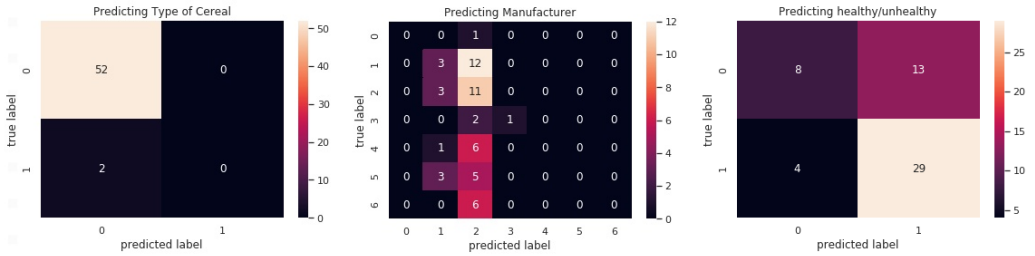


Figure 11: Confusion Matrix shown as a heatmap for three Categorical variables.

Naive-Bayes is one of the supervised learning algorithms which using Bayes theorem assuming no correlation among the features in the dataset exists. Random Forests uses the process of bagging to randomly sampling subsets of training data and then uses them for aggregating the predictions. We performed NB and Random forest along with Tfid Vectorizer using a pipeline to predict different labels using different features. The Tfid vectorizer,term frequency-inverse document frequency, will tokenize the document, learn the vocabulary, and compute the inverse frequency. Inversion means downscaling the most frequent words appearing across documents.

Taking "brand names" to predict the two categorical variables present in the dataset that are Manufactures and type of cereal, we get some amazing results. While predicting categorical variable the manufacturer, the accuracy was too low which came next to 0.28, and while predicting the type we get almost 0.97 commendable accuracies. When we added another categorical variable namely Health rating that describes whether the cereal is healthy or unhealthy, we get 0.66 accuracy. Compared the the results from NB with the Random Forest, it performed identically for predicting cereal manufacturer and Type. It exceeded NB for predicting Healthy/unhealthy rating. The accuracy was almost 0.70. All other predictions for continuous variables faired worse. Having the accuracy of around 0.3. In conclusion, we can say that this method was not that reliable compared to the KNN on PCA Dimension reduced features which had 0.98 accuracy score and exceeded the NB accuracy score. For future scope what we can do is to add some more data in the dataset or we can use some Gradient boosting algorithm such as Adaboost to increase the accuracies.

9 MULTIPLE LINEAR REGRESSION

Variable	Estimate for Coefficient	P-value
Intercept	59.77353	0
calories	-.19984	0
protein	2.92629	0
fat	-1.97562	0
sodium	-0.05861	0
fiber	2.70013	0
carbo	1.046	0
sugars	-0.82197	0
vitamins to potassium ratio	-0.14522	.0036
shelf	-0.72164	.0003
weight	-4.6685	.0233
cups	-0.95566	.2060

We did not use a train-test split as we wanted to implement MLR as a inference statistical tool rather than a machine-learning tool to understand how the rating variable was scored.

The regression suggests our predicted variable serves as a proxy for the health rating of the cereal ceteris paribus. According to the regression analysis, consumers prefer healthier cereals, cereals with fewer calories, and sugar.

The data contradict the "health-pleasure trade-off effect" phenomenon cited by Bialkova as the survey participants rated the healthy foods higher.

Since ratings of cereals correlate with the health level of the cereal, this suggests, assuming a representative sample, that no commercials do not pressure consumers to buy cereals that have too high amounts of sugar, and cereal companies do not have the incentive to sell to customers highly sugary products, but rather have management who orient themselves towards health promotion. We had adjusted R^2 99.2 percent

10 Linear Regression

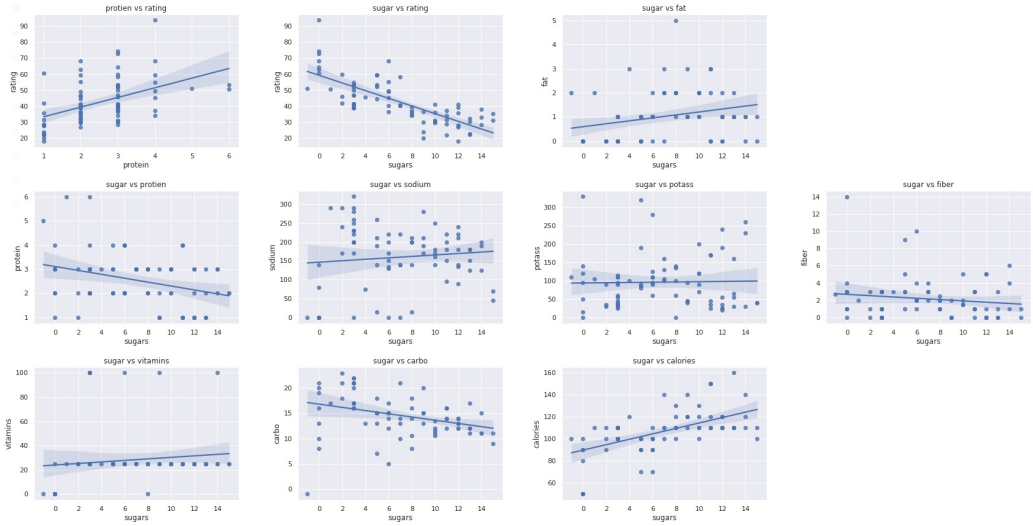


Fig 12: Different regression plots which shows how sugar is correlated with other variables.

While running the linear regression to check the most important relation that is sugar vs rating we got an R^2 57 percent and conclude that sugar is totally negatively correlated with the rating.

11 Clustering

Hierarchical Clustering

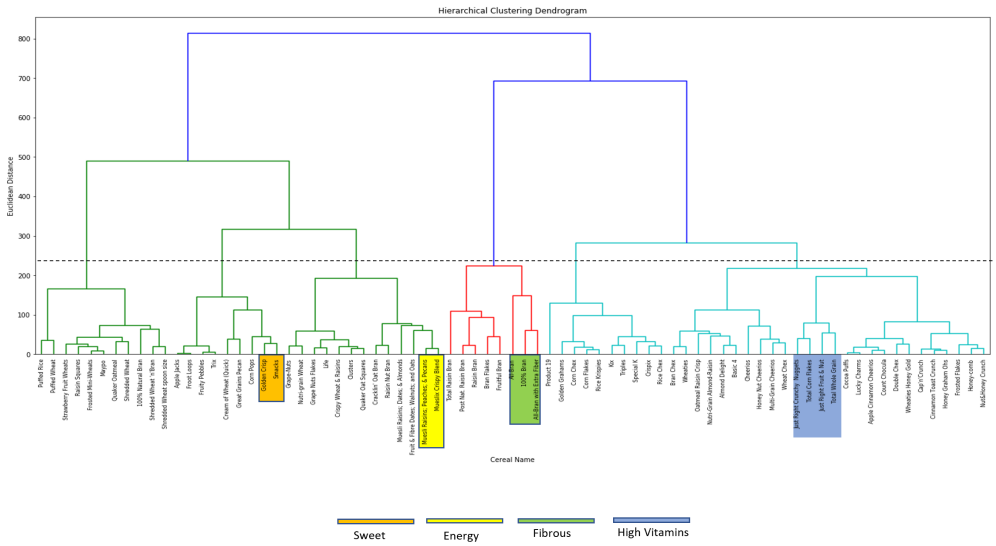


Fig 13: Dendrogram

While plotting the Dendrogram, cereals with common features are clustered together and we can observe clusters of Sugary cereals, high calorie (Energy) cereals, cereals with most fiber content and

high vitamins cereals. Using Hierarchical clustering method, we can see optimal number of clusters should be 6 when we consider all the features in the data set.

Elbow method does not give us a clear elbow point when we consider all the features. In order to reduce number of features, we have eliminated highly correlated features and features with less impact on anyones daily diet. We believed that people would consume equal the standard 2000/2500 calories, protein, carbohydrates daily and as these features in our data set contributes to only about 10 percent of daily calories intake, these can be left out for purpose of our analysis. To form a cluster of healthy/unhealthy cereals we have only considered: 'fiber','sugars','fat','protein' Elbow Method Using k=2 , it was difficult to gain insights from the cluster. Using k=5 was the sweet spot to gain insight.

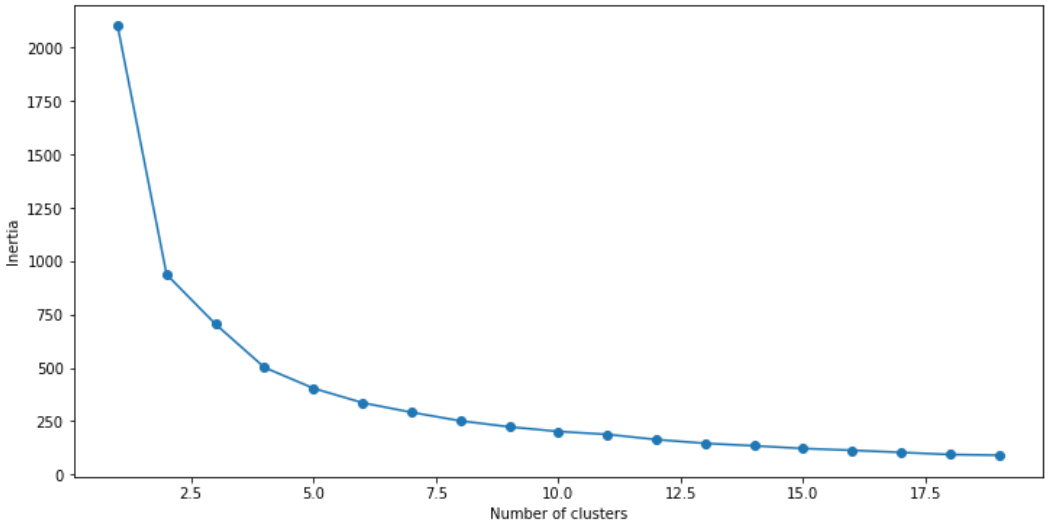


Fig 14: Elbow Curve

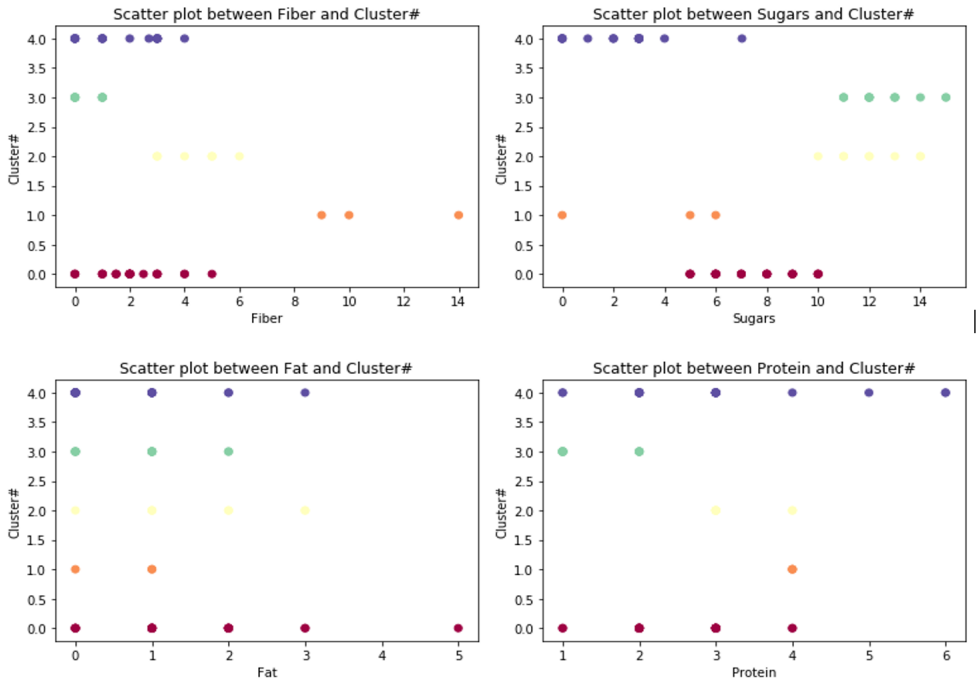


Fig 15: Scatterplots showing different clusters

By looking at the mean values of all features in all available clusters we can identify clusters of healthy and unhealthy cereals. In our analysis cluster 1 has high fiber and low sugar and is categorized as an healthy cluster, cluster 3 has the least fiber and the most sugar content and can be categorized as an unhealthy cluster.

cluster	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
0	108.9	2.5	1.4	158.7	2.0	14.2	7.4	99.6	29.6	2.4	1.0	0.7	40.1
1	63.3	4.0	0.7	176.7	11.0	6.7	3.7	310.0	25.0	3.0	1.0	0.4	73.8
2	135.0	3.2	1.6	174.4	4.2	14.4	12.1	202.5	34.4	2.9	1.3	0.8	36.2
3	110.7	1.4	0.8	149.3	0.4	12.1	12.6	37.5	25.0	1.8	1.0	0.9	28.2
4	98.8	2.8	0.6	159.8	1.5	18.1	2.3	73.3	27.0	1.9	1.0	0.9	51.8

Fig 16: Mean Values

name	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating	cluster
All-Bran with Extra Fiber	50.0	4.0	0.0	140.0	14.0	8.0	0.0	330.0	25.0	3.0	1.0	0.50	93.704912	1
100% Bran	70.0	4.0	1.0	130.0	10.0	5.0	6.0	280.0	25.0	3.0	1.0	0.33	68.402973	1
All-Bran	70.0	4.0	1.0	260.0	9.0	7.0	5.0	320.0	25.0	3.0	1.0	0.33	59.425505	1

Fig 17: Healthy Cluster

name	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating	cluster
Golden Crisp	100.0	2.0	0.0	45.0	0.0	11.0	15.0	40.0	25.0	1.0	1.0	0.88	35.252444	3
Smacks	110.0	2.0	1.0	70.0	1.0	9.0	15.0	40.0	25.0	2.0	1.0	0.75	31.230054	3
Apple Jacks	110.0	2.0	0.0	125.0	1.0	11.0	14.0	30.0	25.0	2.0	1.0	1.00	33.174094	3
Cocoa Puffs	110.0	1.0	1.0	180.0	0.0	12.0	13.0	55.0	25.0	2.0	1.0	1.00	22.736446	3
Count Chocula	110.0	1.0	1.0	180.0	0.0	12.0	13.0	65.0	25.0	2.0	1.0	1.00	22.396513	3
Froot Loops	110.0	2.0	1.0	125.0	1.0	11.0	13.0	30.0	25.0	2.0	1.0	1.00	32.207582	3
Cap'n'Crunch	120.0	1.0	2.0	220.0	0.0	12.0	12.0	35.0	25.0	2.0	1.0	0.75	18.042851	3
Corn Pops	110.0	1.0	0.0	90.0	1.0	13.0	12.0	20.0	25.0	2.0	1.0	1.00	35.782791	3
Fruity Pebbles	110.0	1.0	1.0	135.0	0.0	13.0	12.0	25.0	25.0	2.0	1.0	0.75	28.025765	3
Lucky Charms	110.0	2.0	1.0	180.0	0.0	12.0	12.0	55.0	25.0	2.0	1.0	1.00	26.734515	3
Trix	110.0	1.0	1.0	140.0	0.0	13.0	12.0	25.0	25.0	2.0	1.0	1.00	27.753301	3
Frosted Flakes	110.0	1.0	0.0	200.0	1.0	14.0	11.0	25.0	25.0	1.0	1.0	0.75	31.435973	3
Honey Graham Ohs	120.0	1.0	2.0	220.0	1.0	12.0	11.0	45.0	25.0	2.0	1.0	1.00	21.871292	3
Honey-comb	110.0	1.0	0.0	180.0	0.0	14.0	11.0	35.0	25.0	1.0	1.0	1.33	28.742414	3

Fig 18: Unhealthy Cluster

12 Neural Network Regressor

Fig 19: Learning and Validation curves for the neural network.

The graph above shows the learning capability of the neural network given varying test set sizes. When the test set is less than fifty samples, the network has difficulty making a good prediction on the validation set.

This model is an implementation of a Scikit-Learn MLPRegressor. It contains a minimal number of hidden layers and neurons. Since the dataset is relatively small, a gradient descent optimization is less efficient than the lbfgs solver, which is a quasi-newtonian method. The neural network was able to predict rating and extract meaning from the data that successfully corresponds to other methods. The analysis uses several built-in functions from sci-kit learn to preprocess data, and graph results. A min-max scaler was used to normalize the feature array before feeding it to the neural net. Another function was used to plot partial dependence. A pipeline was used to hold the model, and validation curves were made with built-in metrics.

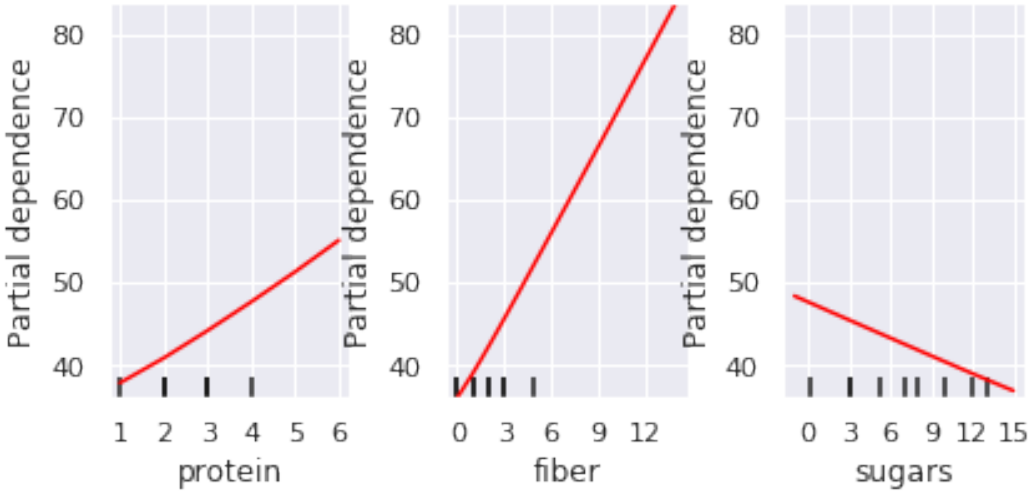


Fig 20: The partial dependence of three variables on the target feature, ratings.

These graphs show the degree to which each of the three features contribute to the prediction of a high or low rating. Fiber has the strongest effect, while sugar has a clear negative correlation. These graphs confirm what we hypothesized from the correlation matrices and the linear regression graphs.

13 CONCLUSIONS

Making Hypotheses about the Data

The final outcomes of interpretations of the dataset are affected by several factors, including the choice of method, the tuning of parameters specific to that method, and the robustness of statistical analysis. But they should also be colored by human factors, like sociology, economics, and psychology. The nexus of these widely varying contexts will give a more fulsome appreciation of the real impact of the given data.

In the case of our cereal data set, it's origin in possible proximity to industry and enterprise research suggests a pattern of promoting healthy cereals. To gain more effective understanding of how cereals more broadly interact with people and society, it becomes necessary to collate many more data sets, and expand the analysis to include factors that can really explain human behavior, which is endlessly complex, even when simply choosing a cereal to eat.

14 CREDITS

Before the formal analysis, Harsh did exploratory data visualization with heatmaps and bar charts. Harsh reported on running Naive Bayes Multinomial Distribution with TFIDF vectorizer to predict various categorical varibles and compared the results with Random Forest Classifier and included heat-maps of the predicted versus true labels. He also extensively organized the project full text so that the order of discussion followed a logical progression.

John reported on multiple linear regression to predict cereal. John also made heatmaps on the correlation matrix of the variables and partitioned based on the manufacturer using boxplots. John ran several initial methods and created the latex document. John created a recommendation table suggesting what cereal to eat based on criteria from K-Means and compared PCA'd SVM, KNN, and gaussian NB using heat maps for predicting the rating variable using all other numerical variables, and PCA with K-means.

Sonam ran K-Means on several variables, ran Hierarchical clustering, recommendation system for and designed the entire power point presentation.

Manasa organized weekly meetings and sent the minutes, and did the Linear Regression part along with plotting the different regplots and also reviewed the report to see grammatical errors.

Gabe contributed text to the project proposal. Gabe implemented a simple sci-kit learn MLPRegressor, using an lbfgs solver, and will make further visualizations of the results. Gabe uploaded a correlation matrix with a different aesthetic. Gabe suggested a better way of using the heatmaps to show correlation, by squaring every element of the corr matrix and using better color schemes. Gabe contributed text to the final report and edited the final report for grammar and clarity, and formatting.

Harsh, John, Gabe, and Sonam wrote the figure captions on the heatmap and boxplots.

REFERENCES

Svetlana Bialkova, Lena Sasse, and Anna Fenko. The role of nutrition labels and advertising claims in altering consumers' evaluation and choice. *Appetite*, 96:38 – 46, 2016. ISSN 0195-6663. doi: https://doi.org/10.1016/j.appet.2015.08.030. URL http://www.sciencedirect.com/science/article/pii/S0195666315300040.

Chris Crawford. 80 cereals, 2017. URL www.kaggle.com/crawford/80-cereals/metadata.

Meghan R. Longacre, Keith M. Drake, Linda J. Titus, Jennifer Harris, Lauren P. Cleveland, Gail Langeloh, Kristy Hendricks, and Madeline A. Dalton. Child-targeted tv advertising and preschoolers' consumption of high-sugar breakfast cereals. *Appetite*, 108:295 – 302, 2017. ISSN 0195-6663. doi: https://doi.org/10.1016/j.appet.2016.10.014. URL http://www.sciencedirect.com/science/article/pii/S019566631630544X.

Venkat Murali. Cereal data factor analysis, 2018. URL https://www.kaggle.com/venky12347/cereal-data-factor-analysis.

E. Neyens, G. Aerts, and T. Smits. The impact of image-size manipulation and sugar content on children's cereal consumption. *Appetite*, 95:152 – 157, 2015. ISSN 0195-6663. doi: https://doi.org/10.1016/j.appet.2015.07.003. URL http://www.sciencedirect.com/science/article/pii/S0195666315003207.

Hrefna Palsdottir. Breakfast cereal:healthy or unhealthy, 2019. URL https://www.healthline.com/nutrition/are-breakfast-cereals-healthy.

365 Team. How to combine pca and k-means in python?, Mar 2020. URL https://365datascience.com/pca-k-means/.

Robin G. Thomas, Pamela R. Pehrsson, Jaspreet K.C. Ahuja, Erin Smieja, and Kevin B. Miller. Recent trends in ready-to-eat breakfast cereals in the u.s. *Procedia Food Science*, 2:20 – 26, 2013. ISSN 2211-601X. doi: https://doi.org/10.1016/j.profoo.2013.04.005. URL http://www.sciencedirect.com/science/article/pii/S2211601X13000060. 36th National Nutrient Databank Conference.

Peter G. Williams. The benefits of breakfast cereal consumption: A systematic review of the evidence base. *Adv. Nutr.*, 5:636S–673S, 2014. URL https://academic.oup.com/advances/article/5/5/636S/4565784.