Are you what you eat?

An Analysis of the National Health and Nutrition Examination Survey

Allie Collins & Erika Tyagi

12.11.2019
MACS 40800 | Unsupervised Machine Learning | Fall, 2019
Repository: https://github.com/erika-tyagi/clustering-nhanes

Contribution Statement:
Both authors jointly worked on the project and write-up of the report with a high degree of overlap. For the paper in particular, Allie focused on the literature review and predictive analysis and Erika on the methodology and dimension reduction / clustering analysis segments.

I. INTRODUCTION

Most individuals would agree that what they eat impacts their health. Yet, it's also true that there are many other contributing factors – a vast body of literature exists on the social determinants of health, defined as "conditions in the environments in which people are born, live, learn, work, play, worship, and age that affect a wide range of health, functioning, and quality-of-life outcomes and risks."[1] Our project seeks to investigate the nutritional profiles of Americans and their implications, leveraging the National Health and Nutrition Examination Survey (NHANES) conducted by the CDC.

The survey asks about many facets of individuals' health, including detailed dietary intake questions that facilitate this analysis. We used dimension reduction and clustering techniques to group nutritional profiles and subsequently link back to health outcomes and socio-demographic factors. We also use regression and supervised machine learning techniques in an effort to contribute to the literature seeking to understand the interdependent relationship between nutrition, socio-demographic factors, and health.

II. LITERATURE REVIEW

A review of existing literature on the topic suggested links between social determinants and individuals' health outcomes. For example, a study in the Journal of Nutrition Education and Behavior titled "Factors Affecting Low-Income Women's Food Choices and the Perceived Impact of Dietary Intake and Socioeconomic Status on Their Health and Weight" suggests challenges with healthy food options being accessible in low-income areas, as well as participants suffering from a lack of education on available and affordable healthy options. These women had a higher prevalence of obesity, diabetes, and hypertension.[2]

[1] https://www.healthypeople.gov/2020/topics-objectives/topic/social-determinants-of-health
[2] Dammann, K. W., & Smith, C. Factors Affecting Low-income Women's Food Choices and the Perceived Impact of Dietary Intake and Socioeconomic Status on Their Health and Weight.

In a similar vein, numerous studies have examined the relationship of food insecurity to weight through lenses such as food stamps. While the statistical evidence varies across studies, a meta-analysis titled "The Food Insecurity–Obesity Paradox: A Review of the Literature and the Role Food Stamps May Play" suggests that there is an association between food insecurity and being overweight among adults – and particularly among women.[3] More broadly, there is a growing recognition of the importance of social determinants of health, as recognized by the World Health Organization's creation of a new Department of Social Determinants of Health, for example, that seeks to disentangle the relationship between poverty and health outcomes.[4]

Another relevant body of literature to our investigation includes a handful of previous attempts to cluster nutritional profiles. One particular study published in the Journal of Epidemiology and Community Health used data similar to ours (i.e., intake over 2 days) to cluster on lifestyle and dietary variables from the Dutch National Food Consumption Survey. They found 8 clusters, four of which were characterized by poor quality food choices. Notably, the approach relied on using the particular foods that individuals ate, rather than the nutritional composition as we do here. Two adjustments made by the researchers included scaling by calories and removing children – both of which we incorporated into our data processing. A study on European adolescents called HELENA (Healthy Lifestyle in Europe by Nutrition in Adolescence) sought to cluster lifestyle risks along with diet and socio-demographic factors to find associations with obesity and other chronic diseases. Another study on an older demographic in Bordeaux leveraged k-means clustering to identify 5 dietary clusters for each gender and then associated these to various health conditions (for instance, depression).

In sum, our investigation into the literature suggested that clustering nutritional profiles can be a useful technique – and that considering the relationship between these clusters to socio-demographic

[3] Dinour, L. M., Bergen, D., & Ming-Chin, Y.
[4] https://www.who.int/social_determinants/strategic-meeting/en/

characteristics and health outcomes would be fruitful. This literature also guided our particular data processing and clustering model specifications.

III. EMPIRICAL STRATEGY

DATA

As previously noted, we relied on the CDC's National Health and Nutrition Examination Survey (NHANES). Since 1999, the survey has explored the health and nutritional status of Americans through a nationally representative sample of about 5,000 persons each year. The survey is particularly unique in that it involves a combination of interviews and a standard physical examination.

The survey includes five components – demographics, dietary, examination, laboratory, and questionnaire. We specifically leveraged the dietary component for data on dietary recall. Beginning in 2002, two dietary recalls were requested from all NHANES examinees, where respondents are asked to provide detailed descriptions (i.e., the type, form, brand, name, and amount consumed) of the foods that they consumed over two 24-hour periods. These responses are then translated into the specific amounts of nutrient intake over those two days. A full list of nutrient variables we used is included in Appendix A. These data provided the basis for our understanding of the nutritional profiles of Americans.

We also leveraged the demographic component to understand how these profiles map to the socio-demographic characteristics of individuals. We specifically considered race, gender, age, and the ratio of family income relative to federal poverty guidelines. Finally, in considering how nutritional profiles relate to health outcomes, we included two variables included in the examination and questionnaire components, respectively – the individual's body mass index (BMI) and whether the individual had ever been told that they have high blood pressure or hypertension.

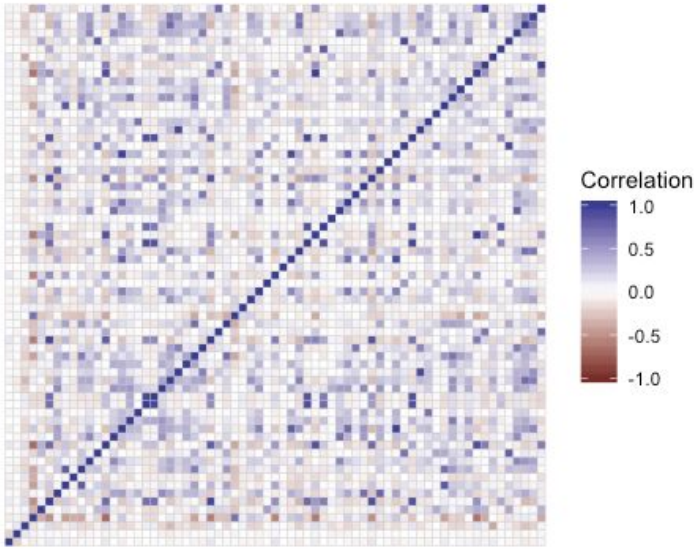
Given the inherent concerns associated with using two days of dietary recall to extrapolate information about the nutritional profiles of individuals, we first subset and aggregated these data to partly address these concerns. Specifically, we first limited our data to adults (i.e., respondents 18 years or older) and those who completed both dietary recall days. We also limited to individuals with reliable recall status – leveraging the CDC's criteria for evaluating whether an individual's responses should be considered reliable – and individuals who reported that the food that they consumed across both days was usual relative to their typical consumption patterns. This yielded a final dataset of 13,263 individuals.

METHODOLOGY

After collating and limiting our dataset as described above, we first sought to reduce the dimension of our feature space of nutritional intake data. First, we did so by simply summing the amount of each nutrient consumed across the two days (i.e., rather than include two features for the amount of each nutrient for each day, we included one feature for the total amount of that nutrient across the two recall days). Then, we standardized these data by dividing each nutrient by the individual's total caloric intake over the two days –as done by previous studies. This allowed us to better understand the nutritional breakdown of each person's intake – rather than simply capture the amount that each person ate. This yielded 67 features for each of the 13,263 individuals in the dataset, with each feature corresponding to a nutrient (e.g., total fat (gm) / energy (kcal) or magnesium (mg) / energy (kcal)).

Our next step was to use principal component analysis (PCA) to reduce the feature space to its latent dimensions. Given the context of our data, we expected several of our features to be highly correlated (e.g., cholesterol and saturated fat). This can be seen from the correlation matrix of these 67 nutrients, presented in Figure 1 below.

Figure 1: Correlation Matrix of Scaled Nutritional Variables



At this point, we could use these principal components to first diagnose the clusterability of our data – or to see if natural groupings of nutritional profiles exist in these data – and then perform clustering analysis to identify these groups. We specifically consider two clustering algorithms (k-means and PAM), and a variety of clusters (2 through 5) and determine the optimal specification using an internal validation approach considering connectivity, Dunn index, and silhouette width.

We then use these clusters in three ways. First, we seek to interpret them based on their inputs (e.g., by visualizing their distribution against the original nutrients) and based on their relationship to our demographic variables of interest to understand the characteristics of individuals who fall into the two groups.

Second, we seek to understand the degree to which these clusters – and the principal components that produced them – can be useful features in predicting health outcomes. Specifically, we build supervised

machine learning classifiers to predict whether an individual is obese and whether an individual has high blood pressure. We build these models with and without the cluster assignment and components as features to evaluate their importance as features and the additional predictive accuracy they yield.

Finally, as suggested by the literature described earlier, it's highly probably that socio-demographic factors likely affect both the health outcomes of interest (i.e., whether an individual is obese and whether an individual has high blood pressure) and also their nutritional intake. Thus, we use a regression analysis approach to identify the relationship between the nutritional profiles and the health outcomes after controlling for these demographic characteristics.

In sum, we use a combination of methodological tools – dimension reduction, clustering, supervised analysis, and regression analysis – to explore the nutritional profiles of Americans, and then consider how these profiles relate to socio-demographic characteristics and health outcomes.

IV. ANALYSIS & RESULTS

As described in detail in the previous section, our methodology followed four steps –

1. Using PCA to reduce the dimension of our nutritional intake feature space
2. Using clustering analysis to identify natural groupings from these components
3. Using supervised machine learning to predict health outcomes from these clusters
4. Using regression analysis to partial out the explanatory power of these clusters

This section presents the results of our analysis and is divided along these steps.

DIMENSION REDUCTION

Applying PCA to our original set of 67 nutritional intake features confirms that the feature space can be

reduced to a reduced set of latent components. Figure 2 below shows the percentage of variance explained as the number of dimensions included increases. Using an 'elbow method' heuristic, roughly seven components are most helpful in capturing this feature space – as the marginal contribution in variation from including additional dimensions tapers off after this point. We thus use just the first seven components to run our clustering analysis.

Figure 2: Scree Plot of Scaled Nutritional Features

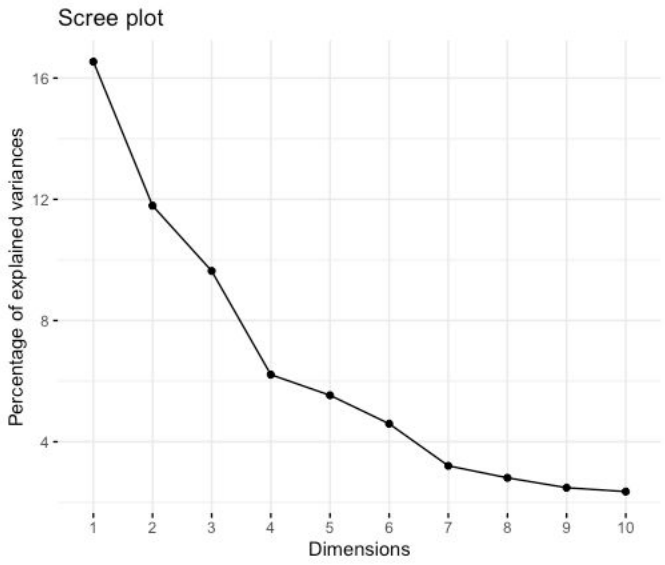
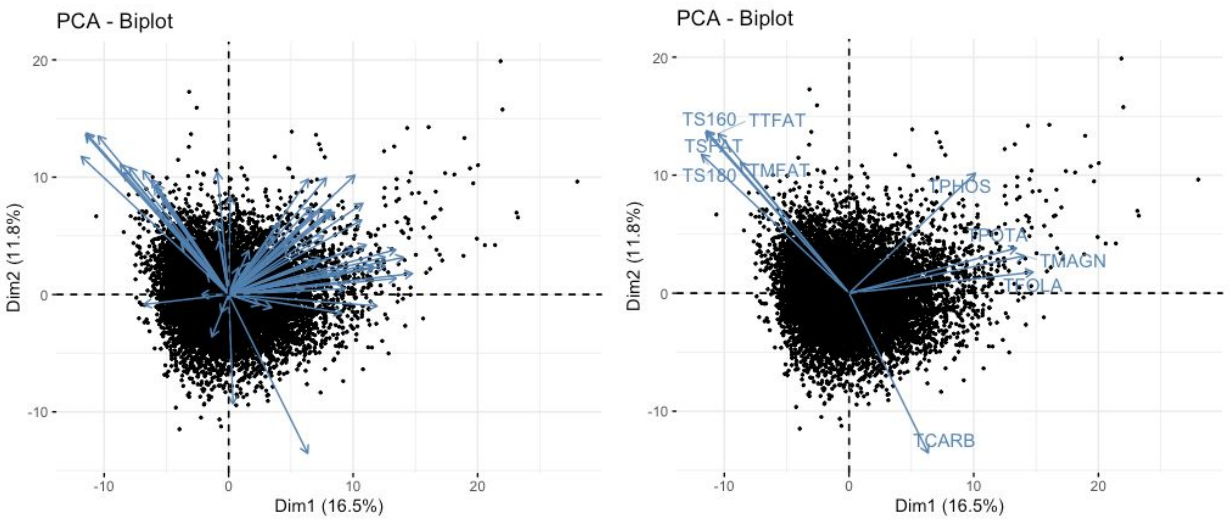


Figure 3 shows how the nutritional features map against the full set of observation across the first two principal component dimensions. The plot on the left shows the full set of 67 features, while the plot on the right simply labels the 10 variables with the highest contribution to the first dimension.

Figure 3: Principal Component Biplots



Looking at the biplot on the right yields insights into an interpretation of the component directions. A series of variables corresponding to 'healthy' nutrients are grouped together (i.e., phosphorous, potassium, magnesium, and folate) as are a series of 'unhealthy' nutrients (i.e., total fat, saturated fat, monounsaturated fat, and two fatty acids).

CLUSTERING ANALYSIS

With the first seven principal components, we then diagnose the clusterability of our data to get a sense of whether natural groupings exist, and if so, how many. We then use cross-validation to guide our model selection – specifically the choice of clustering algorithm and the number of centers. The results are presented in Figure 4 below, and suggest that using a k-means algorithm with 2 clusters would be a strong approach.

Figure 4: Internal Validation Across Hard Partitioning Methods

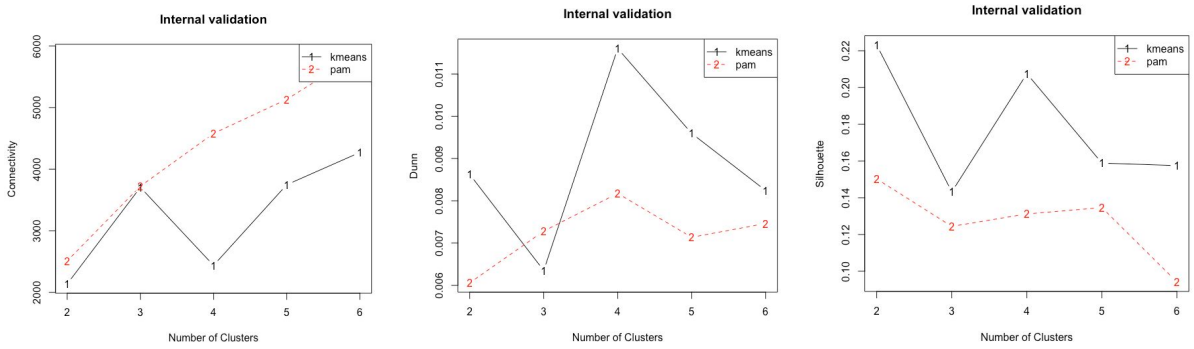
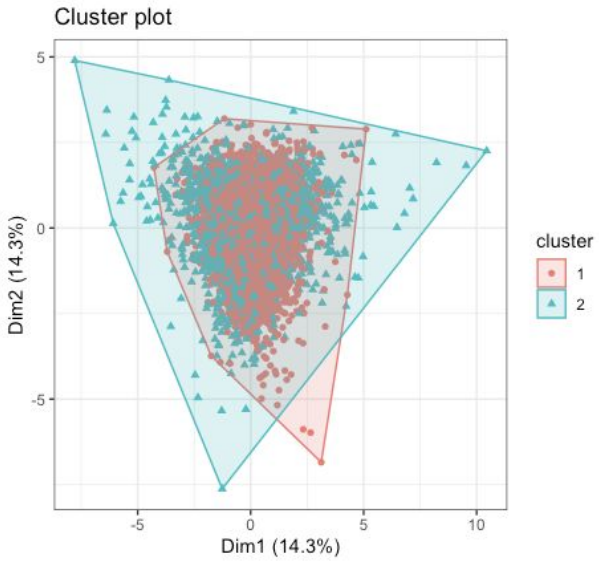


Figure 5 shows the distribution of observations from fitting this k-means algorithm across two dimensions.

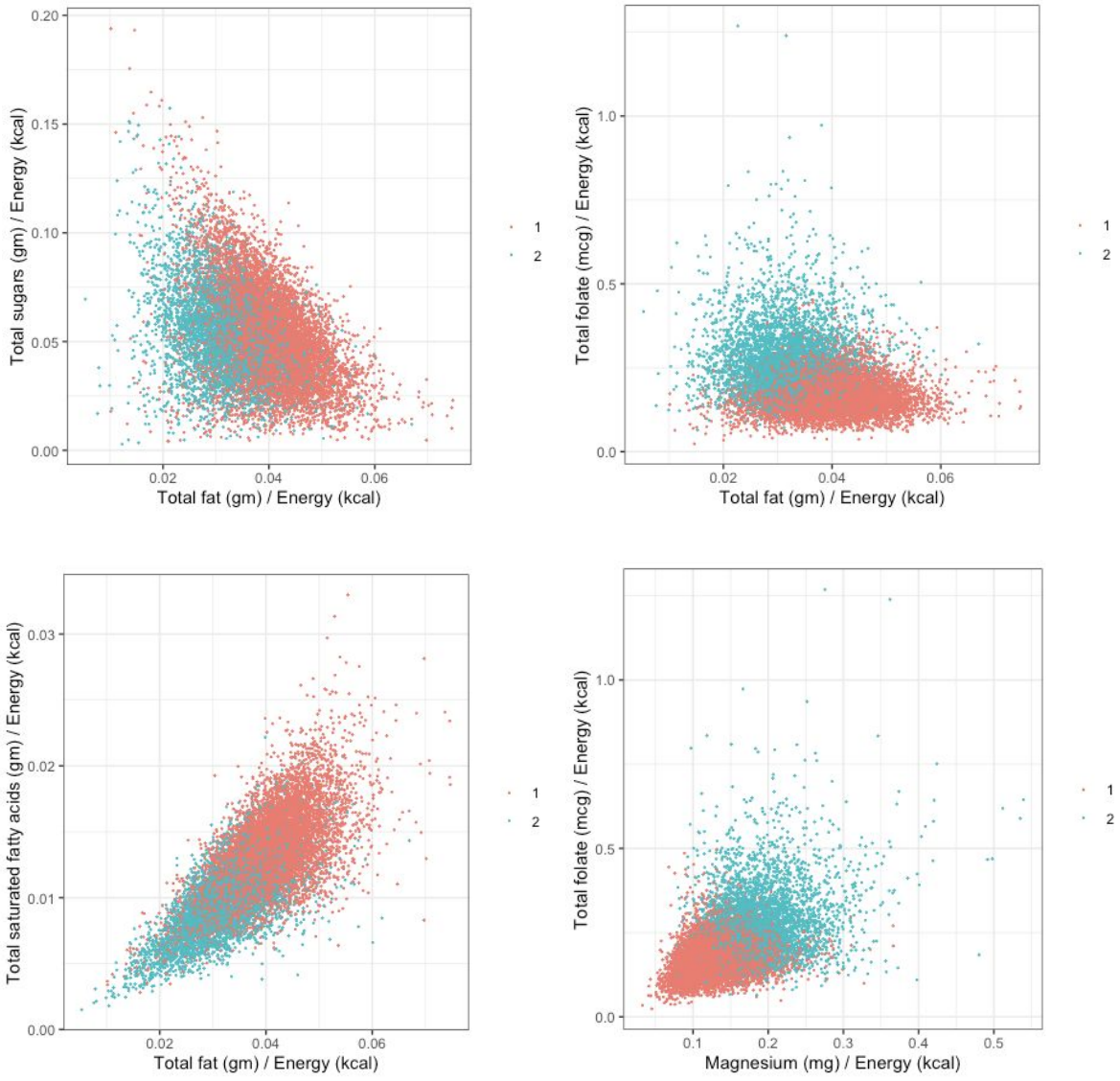
Figure 5: K-Means Cluster Plot



To better understand how the original nutritional intake features are reflected across the clusters, we visualize these distributions across a selected few scatterplots below. Perhaps unsurprisingly, the two clusters roughly correspond to ‘healthy’ and ‘unhealthy’ nutritional profiles – with Cluster 1 being the

'unhealthy' profile and Cluster 2 being the 'healthy' profile. Individuals in Cluster 1 tended to consume more fat and sugar, for example, while individuals in Cluster 2 tended to consume more folate and magnesium (both of which are found in dark green leafy vegetables, seeds, beans, peas, nuts, etc.).

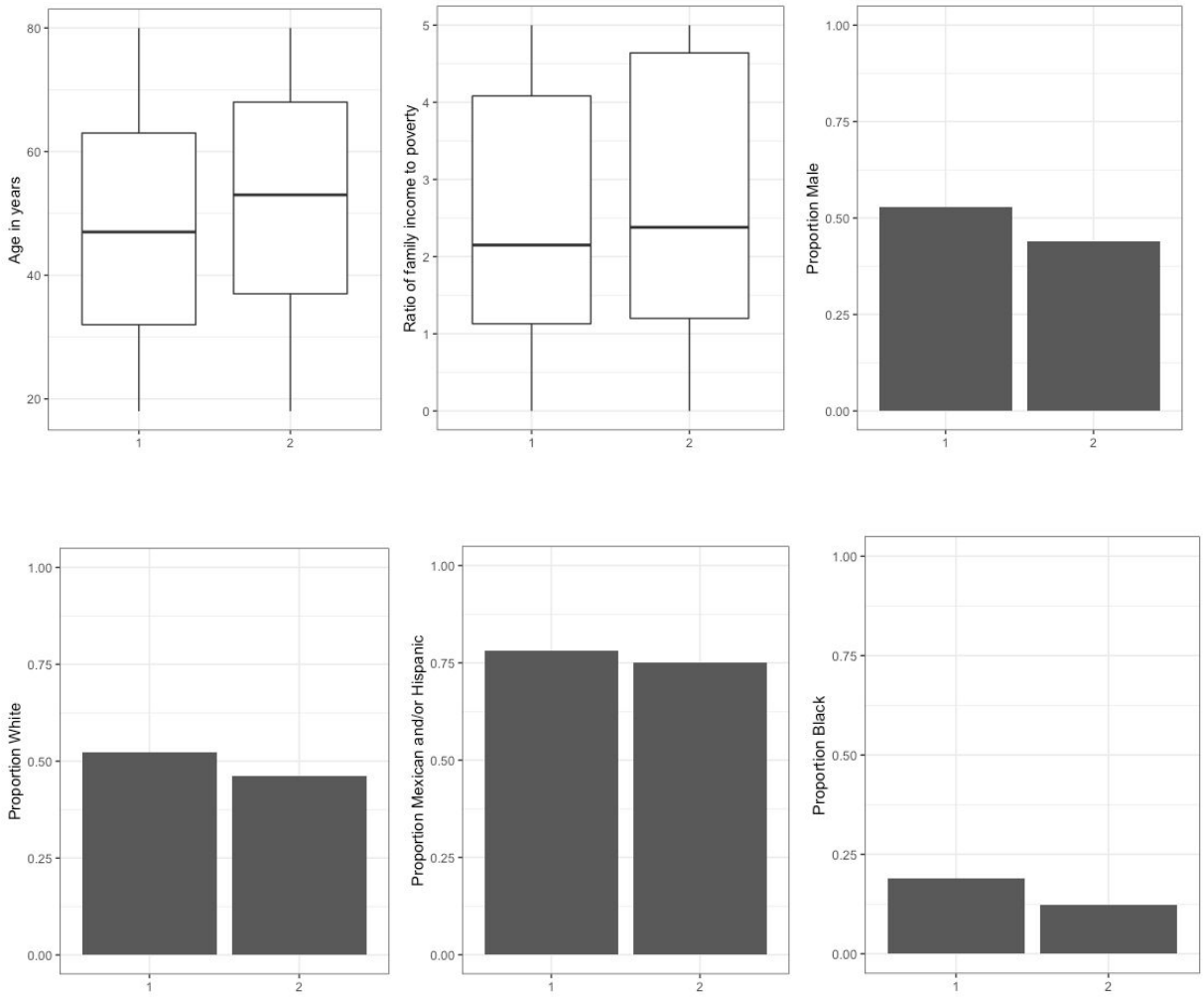
Figure 6: Distribution of Clusters Across Nutritional Intake Features



Finally, we consider how a select handful of demographic features vary across the two clusters – specifically, age, gender, race and ethnicity, and poverty (measured as the ratio of family income to the

federal poverty level). Figure 7 shows that Cluster 1 (the 'unhealthy' profile) is younger, slightly less wealthy, and slightly more male than Cluster 2. Cluster 1 is also slightly more White and more Black than Cluster 2.

Figure 7: Distribution of Clusters Across Demographic Characteristics



Importantly, the figures below are useful for an exploratory – rather than confirmatory – approach to understanding meaningful differences between the clusters (i.e., these differences are not necessarily statistically significant). In sum, we used two unsupervised approaches – PCA and clustering – to try to capture the nutritional profiles of individuals based on their dietary recall information. Using these

profiles, we explore the relationship between nutrition and health outcomes in the following sections using both supervised machine learning techniques and regression analysis.

SUPERVISED ANALYSIS

Taking both the clusters and principal components, we leverage supervised machine learning classifiers to try to (separately) predict two health outcomes – (1) whether a person is obese and (2) whether a person has had high blood pressure. The intent in this analysis is to understand whether the nutritional profiles are helpful features in predicting these outcomes when also considering the socio-demographic factors – which as described in the literature review have been found to be important. As described earlier, we consider ethnicity, age, gender, and income relative to poverty level – as well as either cluster assignment or the seven principal components of the nutritional features.

We looked at 4 cases:

- Random forest models leveraging demographic variables + cluster assignments
- Random forest models leveraging demographic variables + PCA factors
- SVM models leveraging demographic variables + cluster assignments
- SVM models leveraging demographic variables + PCA factors

Full results (including model evaluation and feature importance) across all models is included in our repository. In general, the models were able to predict with reasonable accuracy, largely driven by the demographic features. Specifically, the demographic variables – particularly the individual's age, ethnicity, and poverty level – saw the highest feature importance. The nutritional profiles when considered as a binary cluster assignment provided little additional predictive power, as did the the components themselves.

This broadly suggests that demographic variables yield higher predictive power than the nutritional

profiles. However, a supervised machine learning approach doesn't allow us to identify the relative contribution of each variable in explaining the dependent variable after controlling for the other variables included in the model. Lifestyle factors associated with being low-income may include eating 'unhealthy' food, for example, but this approach doesn't allow us to identify the relative contribution of one's nutritional profile in determining health outcomes after controlling for the other included demographic variables. The regression analysis approach below seeks to further explore this question. Nonetheless, the low feature importance (and minimal increase in predictive power) associated with including the nutritional profiles as features suggests that these are not meaningfully contributing to predicting the health outcomes of interest.

REGRESSION ANALYSIS

Following our initial supervised machine learning efforts to investigate the association between demographic information, nutritional intake, and health outcomes, we moved to a regression approach to better understand the relative contributions of each variable after controlling for the others.

We ran models looking at the same outcomes as the machine learning classifiers (e.g. whether an individual is obese and then whether an individual has had high blood pressure), with three separate sets of variables:

- Demographic variables only
- Demographic variables + cluster assignment
- Demographic variables + PCA factors

At a high level – and as can be seen in the regression output in Appendix B – the relationship between the two dependent variables (i.e., the health outcomes) and the independent variables is much more nuanced than the supervised approach above would suggest. In the model for obesity, the demographic variables

are highly statistically significant (as suggested by the supervised approach), however, the cluster assignment and components are also highly statistically significant after controlling for these demographic variables. Similarly, in the model where the dependent variable is whether an individual has high blood pressure, the coefficients on the demographic variables are statistically significant – and the cluster assignment and several of the components are also (albeit weakly) significant. Notably, however, the low R-squared value in the model for obesity suggests that this model specification may not be particularly robust.

Overall, the regression approach suggests that while their predictive power in a supervised machine learning context may be low, the relationship between nutritional profiles and health outcomes – beyond what is already captured by differences in socio-demographic variables – is worthy of future exploration.

V. DISCUSSION

Several limitations were imposed by the dataset available to us. Firstly, the survey only includes two days of individuals' food consumption. While we leveraged checks on the individual's dietary recall and whether their consumption was typical for them, two days still does not present a complete view. A longitudinal study may be better equipped to paint a better picture of individuals' dietary intake and thus be better served to interpret links to socio-demographic indicators and health outcomes.Moreover, responding to questions regarding food intake is likely to be impacted by the Hawthorne effect - that is, respondents adjusting their behavior in response to being surveyed – manifesting here as perhaps reporting a healthier version of what they ate or under-reporting things they may have been embarrassed to have consumed (such as a lot of sweets).

Additionally, in linking to health outcomes, we were again constrained by the data on hand with respect to the disease outcomes we could incorporate . For instance, the ideal question to ask may be whether an

individual's food intake can predict whether they will become obese, yet what we were answering is whether they are currently obese (and similarly with high blood pressure). With a dataset that included intake over a longer period of time and sequential health outcomes (e.g. diabetes, with several previous periods of food consumption), one could try to discern a more causal relationship between individuals' nutritional profiles and the associated outcome.

Moreover, our intent in incorporating regression analysis was to hone in on the relative contributions of various demographic and nutritional variables in explaining variance in health outcomes. Future work could seek to build on this idea, and incorporate additional factors to more clearly determine the additive information provided by an understanding of what people are consuming with respect to the myriad other factors that affect their social and physical well-being.

VI. CONCLUSION

In sum, our analysis of the NHANES data indicated that nutritional profiles can be grouped into 'healthy' and 'unhealthy' clusters per our review of the literature and prior beliefs – albeit with less clear differentiation than we initially expected. We then found that in initial efforts to relate to health outcomes, when controlling for demographic variables, the additional lift provided by our current nutritional clusters and components was not substantial – demographic factors did the heavy lifting in predicting health outcomes. However, there is still a relevant role played by nutrition in associations to health outcomes to explore. We look forward to further work that investigates longer time-horizon data regarding nutritional intake to further refine the understanding of the relationship between food, demographic indicators, and health.

REFERENCES

Dammann, K. W., & Smith, C. (2009). Factors Affecting Low-income Women's Food Choices and the Perceived Impact of Dietary Intake and Socioeconomic Status on Their Health and Weight. *Journal of Nutrition Education and Behavior*, *41*(4), 242–253.

Dinour, L. M., Bergen, D., & Ming-Chin, Y. (2007). The Food Insecurity–Obesity Paradox: A Review of the Literature and the Role Food Stamps May Play. *Journal of the American Dietetic Association*, 1952–1961.

Hulshof, K. F., Wedel, M., Lowik, M. R., Kok, F. J., Kistemaker, C., Hermus, R. J., … Ockhuisen, T. (1992). Clustering of dietary variables and other lifestyle factors (Dutch Nutritional Surveillance System). *Journal of Epidemiology & Community Health*, *46*(4), 417–424.

Ottevaere, C., Huybrechts, I., Benser, J., De Bourdeaudhuji, I., Cuenca-Garcia, M., Dallongeville, J., … DeHenauw, S. (2011). Clustering patterns of physical activity, sedentary and dietary behavior among European adolescents: The HELENA study. *BMC Public Health*, *11*(328). https://doi.org/10.1186/1471-2458-11-328

Samieri, C., Jutand, M.-A., Feart, C., Capuron, L., & Letenneur, L. (2008). Dietary Patterns Derived by Hybrid Clustering Method in Older People: Association with Cognition, Mood, and Self-Rated Health. *Journal of the American Dietetic Association*, *108*(9), 1461–1471.

APPENDIX

APPENDIX A: Full List of Included Nutrients

Alpha-carotene (mcg) Alcohol (gm) Added alpha-tocopherol (Vitamin E) (mg) Vitamin E as alpha-tocopherol (mg) Added vitamin B12 (mcg) Beta-carotene (mcg) Caffeine (mg) Calcium (mg) Carbohydrate (gm) Total choline (mg) Cholesterol (mg) Copper (mg) Beta-cryptoxanthin (mcg) Folic acid (mcg) Folate as dietary folate equivalents (mcg) Food folate (mcg) Dietary fiber (gm) Total Folate (mcg) Iron (mg) Energy (kcal) Lycopene (mcg Lutein + zeaxanthin (mcg) MFA 16:1 (Hexadecenoic) (gm)) MFA 18:1 (Octadecenoic) (gm) MFA 20:1 (Eicosenoic) (gm) MFA 22:1 (Docosenoic) (gm) Magnesium (mg) Total monounsaturated fatty acids (gm) Moisture (gm) Niacin (mg) PFA 18:2 (Octadecadienoic) (gm) PFA 18:3 (Octadecatrienoic) (gm) PFA 18:4 (Octadecatetraenoic) (gm Niacin (mg)) PFA 18:2 (Octadecadienoic) (gm)	PFA 18:3 (Octadecatrienoic) (gm) PFA 18:4 (Octadecatetraenoic) (gm) PFA 20:4 (Eicosatetraenoic) (gm) PFA 20:5 (Eicosapentaenoic) (gm) PFA 22:5 (Docosapentaenoic) (gm) PFA 22:6 (Docosahexaenoic) (gm) Total polyunsaturated fatty acids (gm Phosphorus (mg) Potassium (mg) Protein (gm) Retinol (mcg)) SFA 4:0 (Butanoic) (gm) SFA 6:0 (Hexanoic) (gm) SFA 8:0 (Octanoic) (gm) SFA 10:0 (Decanoic) (gm) SFA 12:0 (Dodecanoic) (gm) SFA 14:0 (Tetradecanoic) (gm) SFA 16:0 (Hexadecanoic) (gm) SFA 18:0 (Octadecanoic) (gm) Selenium (mcg) Total saturated fatty acids (gm) Sodium (mg) Total sugars (gm) Total fat (gm) Theobromine (mg) Vitamin A as retinol activity equivalents (mcg) Thiamin (Vitamin B1) (mg) Vitamin B12 (mcg) Riboflavin (Vitamin B2) (mg) Vitamin B6 (mg) Vitamin C (mg) Vitamin D (D2 + D3) (mcg) Vitamin K (mcg) Zinc (mg)

APPENDIX B: Regression Output

	Dependent variable:		
	has_high_bp		
	(1)	(2)	(3)
RIDAGEYR	0.012***	0.012***	0.012***
	(0.0002)	(0.0002)	(0.0002)
as.factor(RIAGENDR)2	0.005	0.007	0.014*
	(0.007)	(0.007)	(0.008)
INDFMPIR	-0.016***	-0.016***	-0.016***
	(0.002)	(0.002)	(0.002)
as.factor(RIDRETH1)2	0.013	0.013	0.018
	(0.016)	(0.016)	(0.016)
as.factor(RIDRETH1)3	0.030***	0.029***	0.031***
	(0.011)	(0.011)	(0.012)
as.factor(RIDRETH1)4	0.147***	0.145***	0.145***
	(0.013)	(0.013)	(0.013)
as.factor(RIDRETH1)5	0.005	0.008	0.018
	(0.015)	(0.015)	(0.015)
as.factor(assignment_kmeans)2		-0.014*	
		(0.008)	
PC1			-0.002*
			(0.001)
PC2			0.005***
			(0.001)
PC3			-0.003**
			(0.001)
PC4			-0.001
			(0.002)
PC5			-0.009***
			(0.002)
PC6			0.013***
			(0.002)
PC7			-0.0004
			(0.003)
Constant	-0.233***	-0.231***	-0.243***
	(0.014)	(0.014)	(0.015)
Observations	13,263	13,263	13,263
R2	0.220	0.220	0.225
Adjusted R2	0.219	0.220	0.224

```
=================================================================
                                  Dependent variable:
                              ---------------------------
                                         is_obese
                                (1)       (2)       (3)
-----------------------------------------------------------------
RIDAGEYR                     0.001***  0.001***  0.001***
                             (0.0002)  (0.0002)  (0.0002)

as.factor(RIAGENDR)2         0.056***  0.062***  0.077***
                              (0.008)   (0.008)   (0.008)

INDFMPIR                     -0.010*** -0.009*** -0.008***
                              (0.003)   (0.003)   (0.003)

as.factor(RIDRETH1)2         -0.086*** -0.082*** -0.074***
                              (0.018)   (0.018)   (0.018)

as.factor(RIDRETH1)3         -0.086*** -0.090*** -0.083***
                              (0.013)   (0.013)   (0.013)

as.factor(RIDRETH1)4           0.011     0.005     0.008
                              (0.015)   (0.015)   (0.015)

as.factor(RIDRETH1)5         -0.249*** -0.237*** -0.221***
                              (0.017)   (0.017)   (0.017)

as.factor(assignment_kmeans)2           -0.062***
                                        (0.009)

PC1                                                -0.010***
                                                    (0.001)

PC2                                                0.013***
                                                    (0.001)

PC3                                                -0.007***
                                                    (0.002)

PC4                                                 0.005**
                                                    (0.002)

PC5                                                -0.011***
                                                    (0.002)

PC6                                                0.026***
                                                    (0.002)

PC7                                                 0.007**
                                                    (0.003)

Constant                     0.386***  0.393***  0.356***
                              (0.016)   (0.016)   (0.017)

-----------------------------------------------------------------
Note:                                *p<0.1; **p<0.05; ***p<0.01
```