ANALYSIS OF FOOD NUTRIENT COMPOSITION USING MULTIVARIATE STATISTICAL METHODS

NEELESH VASNANI

FINAL PROJECT

161.762 MULTIVARIATE ANALYSIS FOR BIG DATA

DR. MATTHEW PAWLEY

MASSEY UNIVERSITY

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Obesity and malnutrition are two of the most concerning nutrition-related causes of chronic illness and death. This study employs multivariate statistical techniques to analyze food nutrient composition and better understand food nutrition in light of concerning nutritional issues such as obesity, malnutrition, and the overall impact of nutrition on health in general. The data used in the study was sourced from the official USDA food database which contains thousands of food items and their nutritional values.

A variety of statistical methods were used to understand the food nutrition data. Principal components analysis (PCA), factor analysis, and correspondence analysis were performed to understand how the nutritional variables are related to each other and to the different types of foods. Cluster analysis was used to group foods that are similar in terms of their nutrient composition. Canonical correlation analysis was carried out to determine how different sets of nutritional variables relate to each other. Finally, discriminant analysis and partial least squares analysis were used to create models that predict foods that are high in disease-related nutritional variables.

The results from the different multivariate analyses produced meaningful insights. PCA and Factor analyses were both able to reduce dimensionality by 71% and explained which foods are associated with which variables. For instance, the variables TotalFat, SaturatedFat, and Carbohydrates were most closely related to Calories. Foods such as condiments/sauces/oils, snacks, and bread/desserts/sweets were associated with having a high number of calories. Moreover, strong canonical correlations were found between vitamins and minerals and between micronutrients and macronutrients. Cluster analysis identified three groups of foods with similar vitamin and micronutrient composition. Lastly, three predictive models were finalized using discriminant analysis to classify whether foods are high in calories, sugar, or cholesterol.

1. INTRODUCTION AND RATIONALE

Countries and organizations typically use health indicators such as life expectancy and mortality rates to track and monitor the overall health levels of certain groups or populations. The literature is abundant with studies that investigate patterns and movements in these health indicators at a surface level. However, it is perhaps more important to be concerned about the underlying levers and root causes that drive these health indicators. Nutrition is one driver that plays a key role in an individual's overall health and greatly impacts these health indicators. The importance of proper nutrition cannot be undermined as it is one of the most important factors underpinning a healthy lifestyle.

According to a report by CDC (2018), 75% of all deaths occur from the top 10 leading causes of death. Some of the top leading causes of death globally are major diseases such as heart disease (23.1%), diabetes (2.9%), cancer (21.7%), and kidney disease (1.8%). Combined, these four causes account for about half (49.5%) of all deaths. Daousi et al. (2006) noted that a major treatable risk factor that is common across these major diseases - especially diabetes and heart disease - is obesity. This implies that obesity is a major risk factor for more than half of all deaths. It is important to emphasize that obesity, in turn, is primarily caused by unhealthy diets and poor nutrition habits. Despite being preventable, obesity has tripled since 1975 and kills more people than underweight (WHO, 2020). Other top leading causes of death are diarrhoeal diseases, respiratory illnesses, and malaria. One of the major preventable risk factors which is invariably common across these three causes of death is malnutrition (WHO, 2020). In fact, a century-concluding study by Rice et al. (2000) noted that malnutrition accounted for 50% of all deaths in children. These statistics illuminate that obesity and malnutrition are two of the most concerning nutrition-related causes of chronic illness and ultimately death. In this way, obesity, malnutrition, and nutrition in general greatly impact health indicators such as life expectancy and mortality rate.

Understanding nutrition from a consumption standpoint requires a thorough investigation of available foods and their nutrient composition. To address growing nutrition concerns around the world, several health and weight loss applications have been developed. In fact, one in six individuals actively uses some sort of health application on their smartphone (CDC, 2018). These applications are generally powered by official food databases maintained by organizations such as the United States Department of Agriculture (USDA). The USDA food database contains nutritional information of thousands of food items including multiple nutritional variables such as protein, fat, carbohydrates, and calories, among other nutrients. As such, the multivariate, matrix-like nature of this data merits the use of multivariate statistical methods for efficient analysis.

Several multivariate analysis studies such as De Moraes et al. (2012) and North and Emmet (2000) have looked into specific diet combinations and their associations with various socio-economic variables. However, not much work has been done in using multivariate analysis to study nutritional variables intrinsically based on the totality of foods available for consumption. This work explores food nutrient composition using multivariate statistical methods to better understand food nutrition at its core in light of the aforementioned themes such as obesity, malnutrition, and the overarching impact of nutrition on health in general. The analysis is hinged on the following aims:

- How are the several nutritional variables related to each other?
- Which food items are closely associated with which nutrients?
- Which types of food are associated with higher amounts of each nutrient?
- Which types of food are discernably similar to each other in terms of nutrient composition?
- How are groups of nutritional variables related to each other? Are there canonical correlations between vitamins and minerals? and between macronutrients and micronutrients?
- How can nutritional variables be modeled to best predict foods that are high in disease-related variables such as calories, sugar, and cholesterol?

2. METHODS

This section outlines the data collection and analysis methods used for the study.

2.1. DATA PREPARATION

The official USDA food database is the primary dataset used for the analysis. The data is sourced from the United States Department of Agriculture National Nutrient Database for Standard Reference. After treatment of missing values and outliers, the cleaned database contains 4,124 unique food items with their nutrient composition represented by 14 nutritional variables. However, this data in its raw form is not sufficient to address the research aims. As such, the food database is augmented with additional variables to enable deeper analyses.

First, a novel 10-category classification of foods is developed and curated based on the official food category groups recommended by ODPHP (2021) and NIA (2017). These categories are blended into the dataset and each food item is assigned one of these food categories. This allows simplified analysis by enabling meaningful inspection of relevant food types as opposed to

the granular obstacle of investigating and comparing thousands of food items. Second, for each of the 14 nutritional variables, a corresponding binary variable is added to classify whether or not the food item is high in the respective nutrient. The thresholds used to determine whether the amount of nutrients in a food item is high are based on equivalents of the dietary intake standards recommended by the National Institute of Health (NIH, 2021). Table 1 below summarizes the structure of the final dataset used for analysis in terms of the variables, variable types, variable definitions, and data sources.

Table 1: Structure of Final Dataset

Variables (31)	Definition	Type	Source
ID	Unique identifier for 4,124 observations	index	USDA Food database (USDA, 2021)
Description	Detailed name of food	qualitative	USDA Food database (USDA, 2021)
Category	Category of food	qualitative	Blended from ODPHP (2021), NIA (2017)
Calories	Number of Calories (kcal) in 100g of the food	continuous	USDA Food database (USDA, 2021)
Protein	Amount of Protein (g) in 100g of the food	continuous	USDA Food database (USDA, 2021)
TotalFat	Amount of TotalFat (g) in 100g of the food	continuous	USDA Food database (USDA, 2021)
Carbohydrate	Amount of Carbohydrates (g) in 100g of the food	continuous	USDA Food database (USDA, 2021)
Sodium	Amount of Sodium (mg) in 100g of the food	continuous	USDA Food database (USDA, 2021)
SaturatedFat	Amount of Saturated Fat (g) in 100g of the food	continuous	USDA Food database (USDA, 2021)
Cholesterol	Amount of Cholesterol (mg) in 100g of the food	continuous	USDA Food database (USDA, 2021)
Sugar	Amount of Sugar (g) in 100g of the food	continuous	USDA Food database (USDA, 2021)
Calcium	Amount of Calcium (g) in 100g of the food	continuous	USDA Food database (USDA, 2021)
Iron	Amount of Iron (mg) in 100g of the food	continuous	USDA Food database (USDA, 2021)
Potassium	Amount of Potassium (mg/d) in 100g of the food	continuous	USDA Food database (USDA, 2021)
VitaminC	Amount of Vitamin C (mg) in 100g of the food	continuous	USDA Food database (USDA, 2021)
VitaminE	Amount of Vitamin E (mg) in 100g of the food	continuous	USDA Food database (USDA, 2021)
VitaminD	Amount of Vitamin D (ug) in 100g of the food	continuous	USDA Food database (USDA, 2021)
HighCalories	1 if high in Calories, 0 if not; Threshold: 667kcal	discrete-binary	Derived; threshold based on NIH (2021)
HighProtein	1 if high in Protein, 0 if not; Threshold: 26.7g	discrete-binary	Derived; threshold based on NIH (2021)
HighTotalFat	1 if high in Total Fat, 0 if not; Threshold: 25.7g	discrete-binary	Derived; threshold based on NIH (2021)
HighCarbohydrate	1 if high in Carbohydrates, 0 if not; Threshold: 66.7g	discrete-binary	Derived; threshold based on NIH (2021)
HighSodium	1 if high in Sodium, 0 if not; Threshold: 767mg	discrete-binary	Derived; threshold based on NIH (2021)
HighSaturatedFat	1 if high in Saturated Fat, 0 if not; Threshold: 7.33g	discrete-binary	Derived; threshold based on NIH (2021)
HighCholesterol	1 if high in Cholesterol, 0 if not; Threshold: 100mg	discrete-binary	Derived; threshold based on NIH (2021)
HighSugar	1 if high in Sugar, 0 if not; Threshold: 10g	discrete-binary	Derived; threshold based on NIH (2021)
HighCalcium	1 if high in Calcium, 0 if not; Threshold: 333mg	discrete-binary	Derived; threshold based on NIH (2021)
HighIron	1 if high in Iron, 0 if not; Threshold: 2.67mg	discrete-binary	Derived; threshold based on NIH (2021)
HighPotassium	1 if high in Potassium, 0 if not; Threshold: 1167mg/d	discrete-binary	Derived; threshold based on NIH (2021)
HighVitaminC	1 if high in Vitamin C, 0 if not; Threshold: 30mg	discrete-binary	Derived; threshold based on NIH (2021)
HighVitaminE	1 if high in Vitamin E, 0 if not; Threshold: 5mg	discrete-binary	Derived; threshold based on NIH (2021)
HighVitaminD	1 if high in Vitamin D, 0 if not; Threshold: 6.67ug	discrete-binary	Derived; threshold based on NIH (2021)

2.2. STATISTICAL METHODOLOGY

A variety of multivariate statistical techniques are used to understand the data and facilitate meaningful analysis. Primarily, the multivariate analysis procedures available in the software SAS are used to carry out the analysis. In procedures where variables are not automatically standardized, the variables are standardized before being analyzed. The following methods are selected for analysis:

- **Principal Components Analysis (PCA):** PCA is used to reduce the number of variables and visualize the data to understand the variables through principal components. This analysis would help visualize in an efficient manner which foods are associated with which nutritional variables and how the variables interrelate with each other.
- **Factor Analysis:** Factor analysis is also a dimensionality reduction method similar to PCA and is used to identify underlying factors that determine the correlations between variables. Factor analysis is carried out primarily to validate and supplement the PCA and to see if any latent factors exist among the nutritional variables.
- **Correspondence Analysis (CA):** Correspondence analysis is used to analyze associations between the levels of multiple categorical variables. This analysis is chosen to address the categorical nature of some variables; for instance, to see which food categories are associated with high amounts of each nutrient based on frequency counts of foods that are classified as high in a particular nutrient.
- **Canonical Correlation Analysis:** Canonical correlation analysis is used to measure and visualize correlations between sets of variables. This method is used to see how nutritional variable groups are correlated to each other; for example, to see how vitamins (Vitamin C, Vitamin D, and Vitamin E) are related to minerals (Calcium, Iron, Potassium, and Sodium).
- **Cluster Analysis:** Cluster analysis is an unsupervised classification technique that reduces data complexity by grouping data into clusters. This method is chosen to determine the similarity of different foods based on their nutrient characteristics. While a PCA or correspondence analysis can already visualize the similarity of foods based on individual variables, this technique can efficiently delineate which foods are most similar to each other based on overall vitamin or macronutrient content.
- **Discriminant Analysis:** Discriminant analysis is a supervised classification technique that uses discriminant functions to differentiate groups and ultimately predict group membership when membership is unknown. This technique is chosen to create a number of predictive models that may be useful in the context of nutrition issues such as obesity. For

example, discriminant analysis can be used to predict whether a food is high in calories or high in sugar based on its macronutrient composition - in the event that only this information is available. For this method, the dataset is divided into train and test subsets in proportions of 75% for the training dataset (3092 observations) and 25% for the test dataset (1032 observations). For each model, stepwise discriminant analysis is used to determine important variables for classification. Depending on the test of equality of covariance matrices, either a fisher linear discriminant function or quadratic discriminant function is created based on the training dataset using the variables deemed important by the stepwise discriminant analysis. The discriminant functions are tested on the test dataset to inspect error rates as summarized in the confusion matrix.

- **Partial least Squares (PLS):** PLS is a technique similar to discriminant analysis in that it is used to create models based on variable importance. This technique comes in handy especially when there are multiple response variables such as when predicting a set of dependent variables from a set of independent variables. The PLS method is chosen primarily to tackle multiple-response models where discriminant analysis would not be suitable. For example, PLS can efficiently create a model where the set of dependent variables comprises of multiple macronutrient variables (e.g. protein, fat, carbohydrates) and the set of independent variables comprises of several micronutrient variables.

3. RESULTS AND ANALYSIS

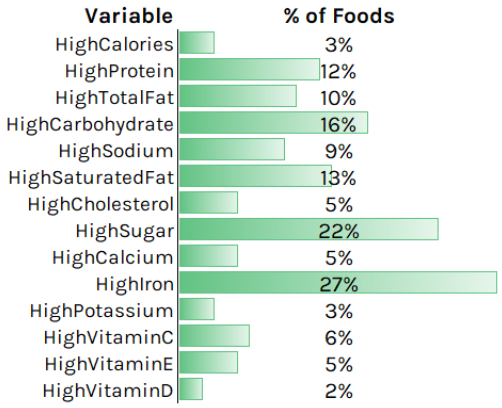
Before proceeding into the multivariate analyses, it is perhaps important to get a feel of the data. The summary statistics of the data are shown below.

Table 2. Summary Statistics of Nutritional Variables

	Calories	Protein	TotalFat	Carbohydrate	Sodium	SaturatedFat	Cholesterol	Sugar	Calcium	Iron	Potassium	VitaminC	VitaminE	VitaminD
unit	kcal	g	g	g	mg	g	mg	g	mg	mg	mg/d	mg	mg	ug
Mean	234.1	7.7	11.3	26.3	386.8	3.8	16.9	10.0	108.6	2.4	286.2	14.1	1.8	0.4
Std	142.4	6.8	13.0	21.9	297.8	4.1	29.9	7.8	120.5	2.5	68.8	16.3	2.4	0.4
High Threshold	666.7	26.7	25.7	66.7	766.7	7.3	100.0	10.0	333.3	2.7	1166.7	30.0	5.0	6.7
Beverages	108.8	1.3	1.0	18.4	53.0	0.6	0.9	12.4	60.4	0.7	260.7	42.5	0.2	0.2
Bread/desserts/sweets	328.7	5.9	10.7	54.3	434.6	4.3	11.9	27.2	119.5	2.2	265.2	0.9	0.6	0.1
Condiments/sauces/oils	464.2	3.4	44.7	15.5	706.8	12.7	5.6	7.0	72.2	2.5	205.0	3.2	7.6	0.3
Dairy	228.5	13.9	14.1	11.7	466.7	8.5	42.8	8.6	431.4	0.7	271.5	1.7	0.4	0.9
Fruits	85.8	1.4	1.4	19.2	56.1	0.7	0.0	14.0	23.2	0.8	289.3	41.9	0.7	0.0
Grains/nuts/cereals	306.6	10.2	7.4	53.5	293.6	1.2	1.2	11.9	134.3	9.3	348.4	10.0	2.7	0.8
Meat/Poultry/Seafood	212.5	23.2	12.5	0.8	261.6	4.5	93.9	0.2	20.9	2.0	307.1	0.6	0.4	0.9
Plants/Vegetables	72.5	3.2	1.5	13.2	151.0	0.3	2.5	3.0	72.6	1.8	354.0	18.1	0.7	0.2
Snacks	422.9	8.8	15.6	63.1	438.5	4.0	3.3	13.7	117.8	3.2	394.0	18.8	4.3	0.2
Soups/mixed dishes	110.4	5.6	3.9	13.5	1006.1	1.1	6.6	2.2	33.7	1.1	167.0	3.2	0.4	0.1

The table above shows the overall mean and standard deviation of each variable, further broken down by food category. Upon initial inspection, some high-level observations can be made. The average number of calories in 100 grams of a food item is 234 kcal. Moreover, the values in red indicate undesirable nutrient amounts while the values in green indicate healthy nutrient amounts. For example, condiments/snacks/oils and snacks are very high in calories whereas fruits and plants/vegetables have fewer calories. It is also apparent that soups/mixed dishes are very high in sodium. These types of associations are further explored in the multivariate analyses to follow.

Figure 1. Percentage of Foods High in Each Nutrient



The figure on the left shows a summary of the derived response variables which indicate if foods are high in any of the nutritional variables based on recommended daily intake equivalents. It is observable that 22% of all food items are high in Sugar, which is concerning given the growing rates of diabetes globally. Additionally, around 16% of foods are high in Carbohydrates, 27% of foods are high in Iron, and only 2% are high in Vitamin D.

Figure 2. Correlation Matrix of Nutritional Variables

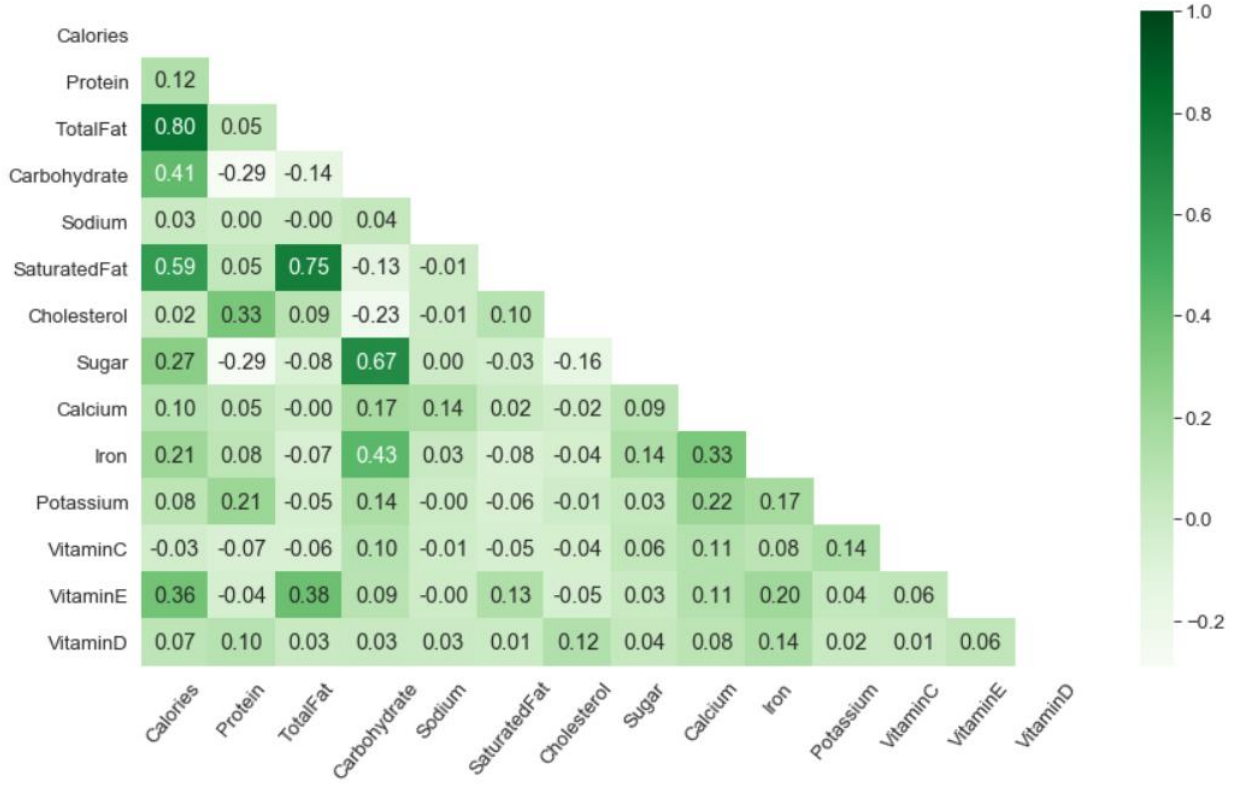
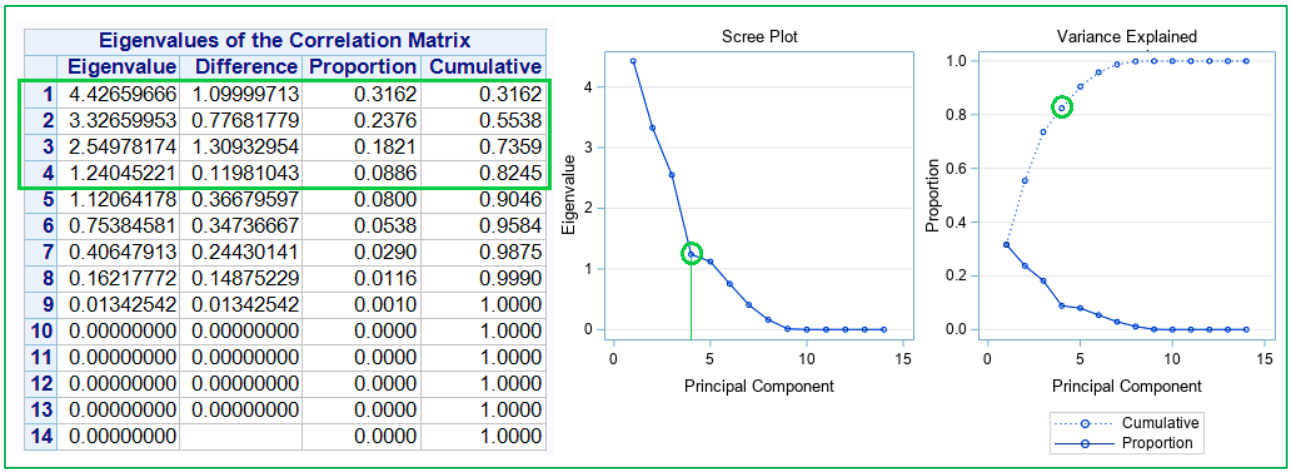


Figure 2 above shows pairwise correlations between nutritional variables. Values in darker shades of green indicate higher positive correlations while values approaching white indicate negative correlations. It is immediately apparent that the variable TotalFat is strongly correlated with Calories. The correlation of 0.80 between these two variables is the highest among all pairs. Carbohydrates and SaturatedFat are also moderately correlated with the number of Calories. It would be interesting to validate if these variables are indeed deemed strong predictors of Calories in the discriminant analysis. Other notable positive correlations occur between Sugar and Carbohydrates, Iron and Carbohydrates, SaturatedFat and TotalFat, Cholesterol and Protein, Vitamin E and Protein/Fat, and Iron and Calcium. Moderate negative correlations are apparent between Sugar and Protein, Carbohydrates and Protein, and Cholesterol and protein.

3.1. PRINCIPAL COMPONENTS ANALYSIS

The figure below shows the output generated from the PCA performed on the correlation matrix of the food nutrition data.

Figure 3. Principal Components Analysis SAS Output



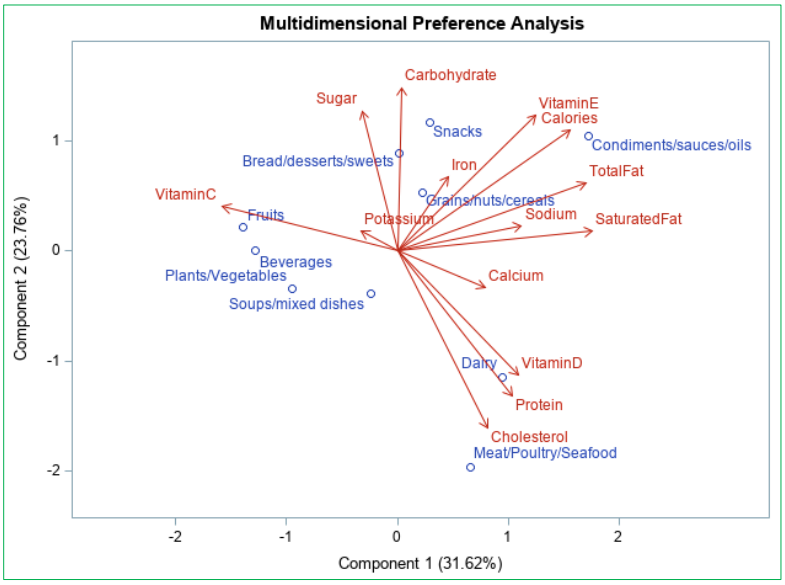
Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	4.42659666	1.09999713	0.3162	0.3162
2	3.32659953	0.77681779	0.2376	0.5538
3	2.54978174	1.30932954	0.1821	0.7359
4	1.24045221	0.11981043	0.0886	0.8245
5	1.12064178	0.36679597	0.0800	0.9046
6	0.75384581	0.34736667	0.0538	0.9584
7	0.40647913	0.24430141	0.0290	0.9875
8	0.16217772	0.14875229	0.0116	0.9990
9	0.01342542	0.01342542	0.0010	1.0000
10	0.00000000	0.00000000	0.0000	1.0000
11	0.00000000	0.00000000	0.0000	1.0000
12	0.00000000	0.00000000	0.0000	1.0000
13	0.00000000	0.00000000	0.0000	1.0000
14	0.00000000		0.0000	1.0000

The first two eigenvalues account for more than half of the variation and the first four out of 14 eigenvalues explain 82% of the total variation. The scree plot shows a sharp elbow at the fourth principal component. The PCA reduced the number of dimensions significantly by 71% - from 14 variables initially to four components explaining more than 80% of the variation.

Figure 4 below shows the biplot of the PCA and reveals which types of foods are associated with which types of nutritional variables. Food categories are used as data points in plotting to clearly show associations between foods and nutrients.

Figure 4. PCA Biplot of Foods and Nutritional Variables



Several insights can be derived from the biplot above. Component 2 is heavily loaded in the variables: Carbohydrates and Sugar, and in the food categories: snacks and bread/dessert/sweets. Hence, snacks and breads/desserts/sweets are positively associated with Sugar and Carbohydrates but negatively related to Protein and Vitamin D. Condiments/sauces/oils can be said to be fairly unhealthy as they are strongly associated TotalFat, SaturatedFat, Sodium, and Calories. Also, fruits are positively associated with Vitamin C and grains/nuts/cereals are associated with Iron. Meat/poultry/seafood and dairy are related positively to Protein, Cholesterol, Calcium, and Vitamin D and negatively to Sugar and Carbohydrates.

Table 3. Summary of Associations based on PCA Component Patterns and Component Scores

Strong negative associations	Principal Component	Strong positive associations
VitaminC *Fruits, Beverages, Plants/vegetables*	PC 1	**Calories, TotalFat, SaturatedFat** *Condiments/sauces/oils, Dairy, Meat/poultry/seafood*
Protein, Cholesterol, VitaminD *Dairy, Meat/poultry/seafood*	PC 2	**Calories, Carbohydrates, Sugar, VitaminE** *Snacks, Breads/desserts/sweets, Condiments/sauces/oils*
Sodium, SaturatedFat *Soups/mixed dishes, Condiments/sauces/oils*	PC 3	**Protein, Carbohydrates, Iron, Potassium, VitaminD** *Snacks, Grains/nuts/cereals, Meat/Poultry/Seafood*
VitaminE, Iron, Potassium *Meat/poultry/seafood, Condiments/sauces/oils*	PC 4	**Sugar, Calcium** *Bread/sweets/desserts, Dairy*
nutritional variables (component patterns) *Food categories (component scores)*	Legend	**nutritional variables (component patterns)** *Food categories (component scores)*

The table above summarizes key observations from the component pattern and component score plots generated by the PCA (See Appendix A for these plots). Principal component 1 primarily describes calorie-dense and fatty foods such as condiments/sauces/oils, dairy, and

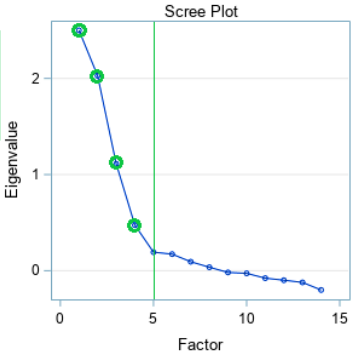
meat/poultry/seafood. Principal component 2 appears to be related to calorie-dense foods that are high in carbohydrates and sugar but low in protein. Component 3 appears to include more nutritious foods such as grains/nuts/cereals. Component 3 is also positively related to healthy nutrients like Iron, Potassium, and Vitamin D and negatively related to undesirable variables like Sodium and SaturatedFat which are found in foods such as condiments/sauces/oils and soups/mixed dishes. This insight was not visible in the biplot which only included two components. Finally, principal component 4 appears to be negatively correlated with micronutrients including Iron, Potassium, and Vitamin E and strongly associated with Sugar and Calcium, which are found in bread/sweets/desserts and dairy products, respectively.

3.2. FACTOR ANALYSIS

Figure 5. Factor Selection for Factor Analysis on Nutritional Variables

	Eigenvalue	Difference	Proportion	Cumulative
1	2.50036812	0.47225448	0.4124	0.4124
2	2.02811364	0.91685216	0.3345	0.7469
3	1.11126148	0.63725267	0.1833	0.9302
4	0.47400881	0.28167092	0.0782	1.0084
5	0.19233788	0.02006494	0.0317	1.0401
6	0.17227295	0.07904884	0.0284	1.0685
7	0.09322411	0.05720813	0.0154	1.0839
8	0.03601598	0.05402840	0.0059	1.0899
9	-.01801243	0.00971355	-0.0030	1.0869
10	-.02772597	0.04981225	-0.0046	1.0823
11	-.07753823	0.02145384	-0.0128	1.0695
12	-.09899206	0.02287135	-0.0163	1.0532
13	-.12186341	0.07883660	-0.0201	1.0331
14	-.20070001		-0.0331	1.0000

4 factors will be retained by the PROPORTION criterion.



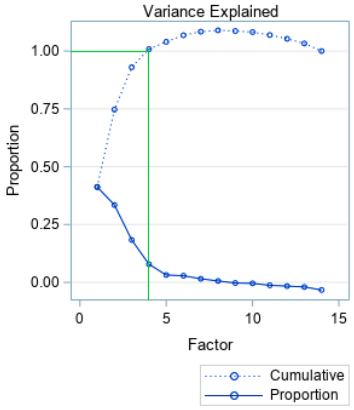


Figure 6. Factor Pattern Loadings on Variables

Rotated Factor Pattern							
	Factor1		Factor2		Factor3		Factor4	
Calories	87	*	.		.		.	
Protein	.		-44	*	.		59	*
TotalFat	95	*	.		.		.	
Carbohydrate	.		89	*	.		.	
Sodium	.		.		.		.	
SaturatedFat	82	*	.		.		.	
Cholesterol	.		.		.		64	*
Sugar	.		84	*	.		.	
Calcium	.		.		64	*	.	
Iron	.		.		59	*	.	
Potassium	.		.		66	*	.	
VitaminC	.		.		46	*	.	
VitaminE	50	*	.		.		.	
VitaminD	.		.		.		57	*

Printed values are multiplied by 100 and rounded to the nearest integer. Absolute values greater than 0.4 are flagged with an '*'. Absolute values less than 0.4 are not printed.

As shown in Figure 5 above, four factors were retained based on the proportion criterion and scree plot. The factor analysis resulted in largely similar dimensionality reduction results with the PCA. Comparing Table 2 to Figure 6, both Factor 1 and PC 1 are positively associated with Calories, TotalFat, and SaturatedFat and hence describe calorie-dense fatty foods. Both Factor 2 and PC 2 are positively associated with Carbohydrate and Sugar and negatively with Protein, describing energy-rich foods. Factor 3 seems to be describing micronutrient-rich foods as it is related to Calcium, Iron, Potassium, and Vitamin C. PC 3 is also strongly associated with Iron and Potassium. Factor 4 is associated with Protein, Cholesterol, and Vitamin D and appears to be describing meat-based and non-vegetarian foods. Factor 4 is the only factor with different variable loadings compared to the PCA since PC 4 is associated with Sugar and Calcium. That said, factor 4 is more interpretable than PC 4. Overall, both PCA and Factor analysis yielded similar results – reducing dimensionality by 71% and having three out of four dimensions with cohesive constructs. However, the factor analysis model is slightly more conclusive on variable loadings.

3.3. CORRESPONDENCE ANALYSIS

Figure 7. Correspondence Analysis

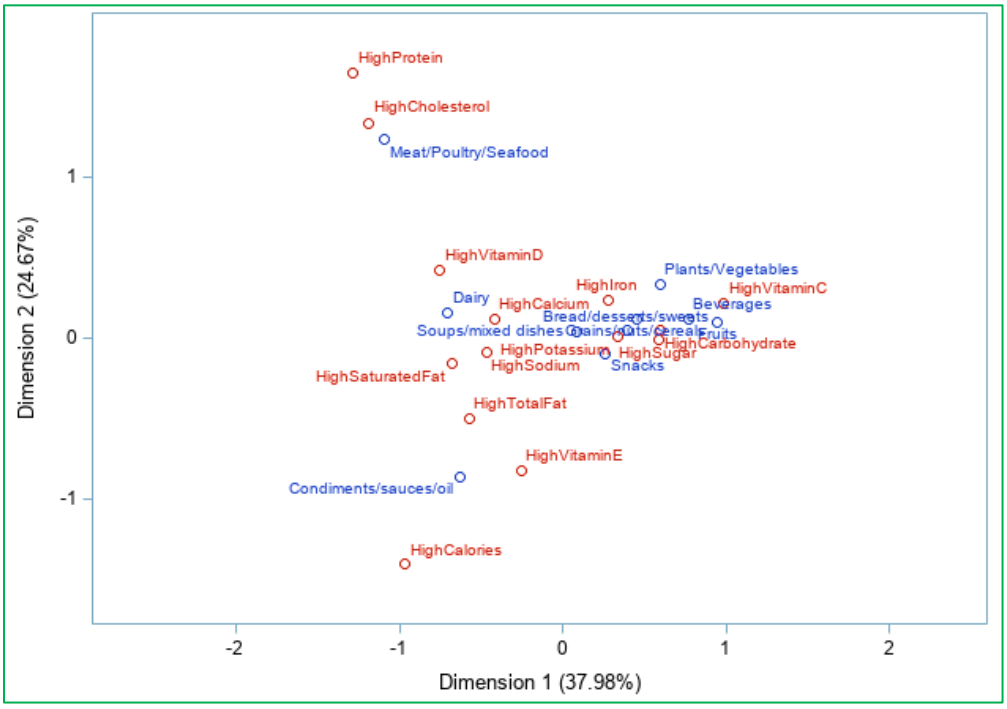


Figure 7 above shows the results for the correspondence analysis between foods and the "high" classifier variables corresponding to each nutrient. For example, food is classified as high in protein if the amount of protein in 100 grams of the food exceeds the 100-gram equivalent of the recommended daily intake for protein. So, if a food item is associated with one of these classifier

variables, it means there are relatively more foods under that category classified as such. The first two dimensions account for 63% of the total inertia implying a decent approximation. The following insights can be observed from the plot:

- Meat/poultry/seafoods and Condiments/sauces/oils have the most distinct associations relative to other foods. The former is strongly associated with HighProtein and HighCholesterol while the latter is strongly associated with HighCalories.
- HighSugar and HighCarbohydrates variables have similar food associations based on their location. Both are related to breads/sweets/desserts, snacks, and fruits.
- Fruits, plants/vegetables, and beverages have similar excess-nutrient associations. They are primarily related to HighVitaminC.
- HighSaturatedFat is associated with dairy products and condiments/sauces/oils.
- Dairy products are associated with HighCalcium.
- Grains/nuts/cereals are associated with HighIron.

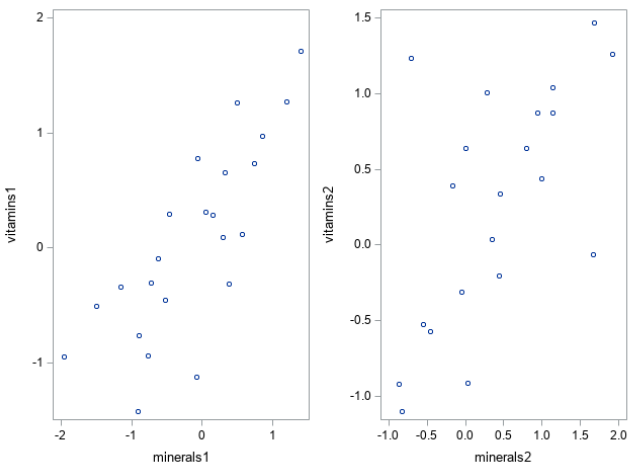
3.4. CANONICAL CORRELATION ANALYSIS

Canonical correlation analysis is undertaken to determine whether correlations exist between the following sets of nutritional variables:

- **Vitamins**: *VitaminC, VitaminD, Vitamin* and
  Minerals: *Calcium, Iron, Potassium, Sodium*
- **Macronutrients**: *Carbohydrates, Protein, TotalFat, Sugar* and
  Micronutrients: *Calcium, Iron, Potassium, Sodium, VitaminC, VitaminD, VitaminE*

Vitamins vs Minerals:

Figure 8: Canonical Correlation SAS Output– Vitamins vs Minerals



	Canonical Correlation	Adjusted Canonical Correlation
1	0.835901	0.722198
2	0.689003	0.624289

Correlations Between the Vitamin and Their Canonical Variables

	vitamins1	vitamins2
VitaminC	0.8019	-0.4876
VitaminD	0.1164	0.9791
VitaminE	-0.4987	0.1759

Correlations Between the Mineral and Their Canonical Variables

	minerals1	minerals2
Calcium	0.0051	0.7748
Iron	-0.1246	0.6120
Potassium	0.3947	0.2758
Sodium	-0.9588	0.0072

Correlations Between the Vitamin and the Canonical Variables of the Mineral

	minerals1	minerals2
VitaminC	0.6703	-0.3359
VitaminD	0.0973	0.6746
VitaminE	-0.4169	0.1212

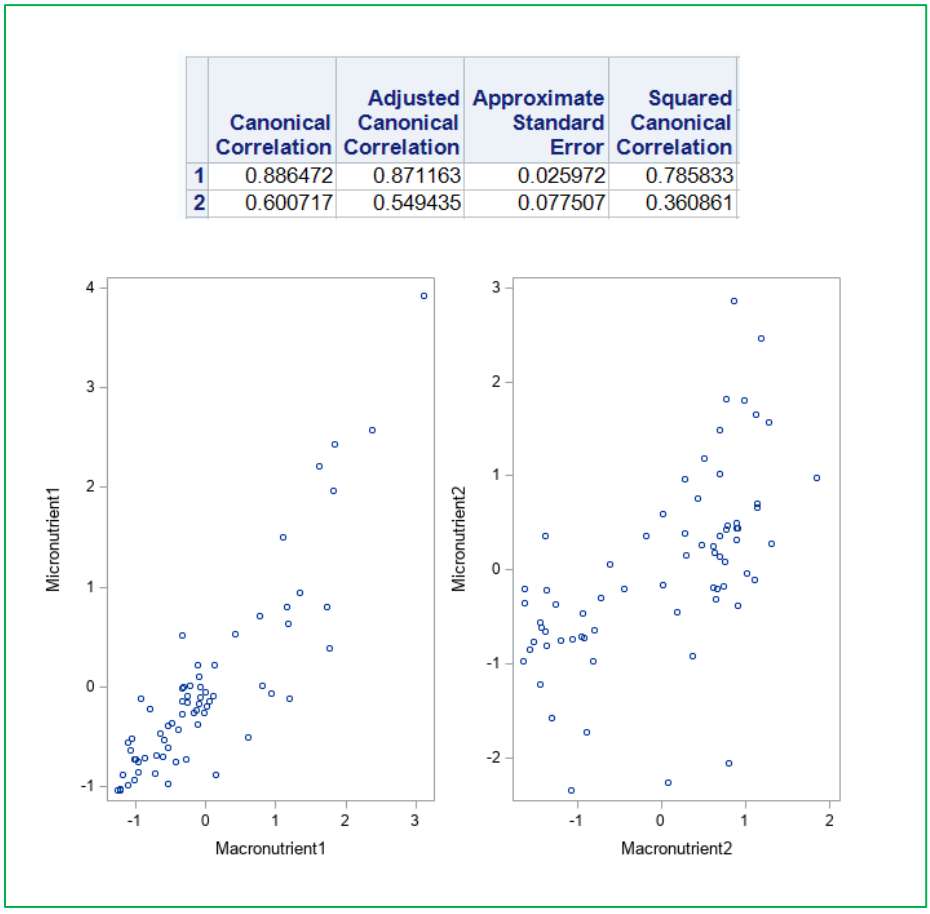
Correlations Between the Mineral and the Canonical Variables of the Vitamin

	vitamins1	vitamins2
Calcium	0.0043	0.5339
Iron	-0.1041	0.4217
Potassium	0.3299	0.1900
Sodium	-0.8015	0.0050

Two significant canonical correlations exist between vitamins and minerals. The first vitamin canonical variable is positively related to Vitamin C and negatively related to Vitamin E. This variable is likely related to Fruits and plants/vegetables. The second Vitamin canonical variable is related positively to Vitamin D and negatively to Vitamin C, so it is likely related to meat products. The first mineral canonical variable is related positively to Potassium and negatively to Sodium. The second mineral canonical variable is mainly related to Calcium and Iron. Based on the cross-correlations between canonical variate pairs, Vitamin C is related to Potassium through the first canonical variate pair and Vitamin D is related to Calcium and Iron, based on the second canonical variate pair.

Micronutrients vs Macronutrients:

Figure 9: Canonical Correlation SAS Output – Micronutrients vs Macronutrients



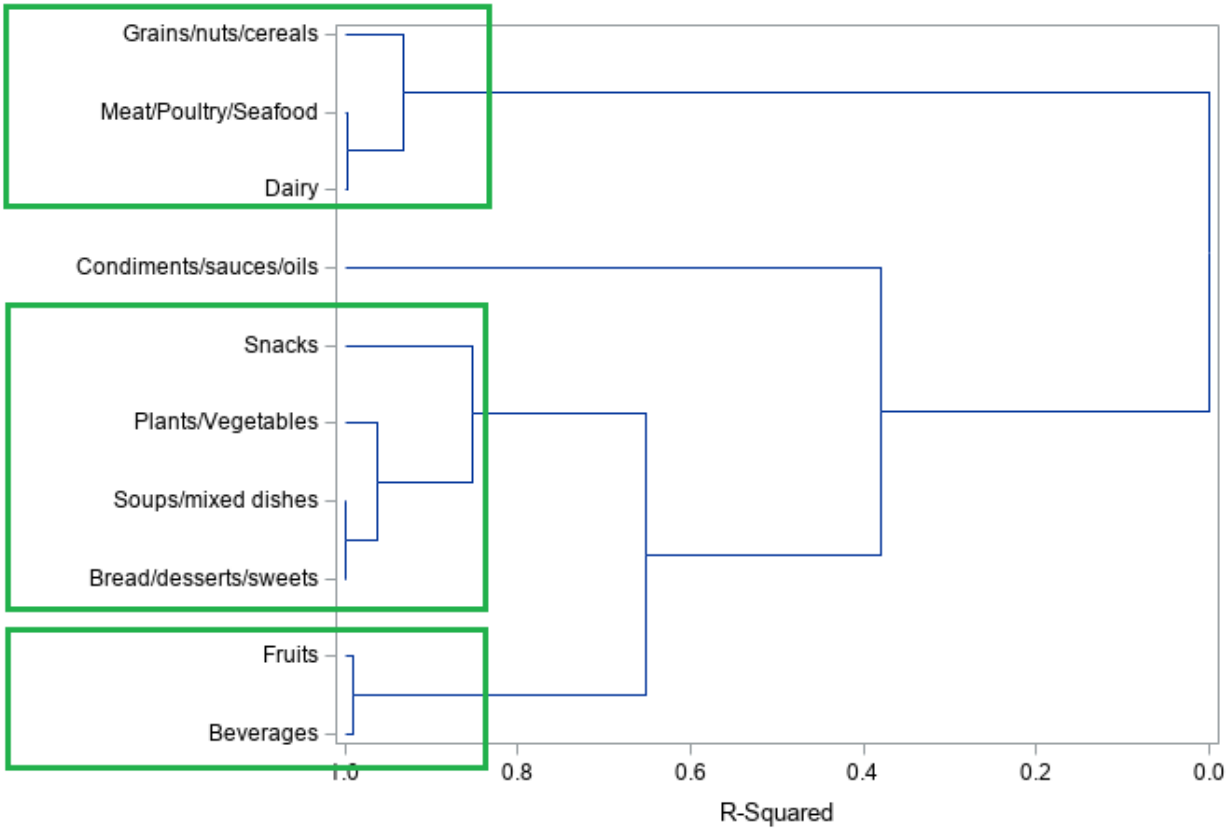
	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation
1	0.886472	0.871163	0.025972	0.785833
2	0.600717	0.549435	0.077507	0.360861

Similarly, two significant canonical correlations are also found between the micronutrient and macronutrient set of variables. The plot above shows strong positive associations between the two pairs of canonical variables.

3.5. CLUSTER ANALYSIS

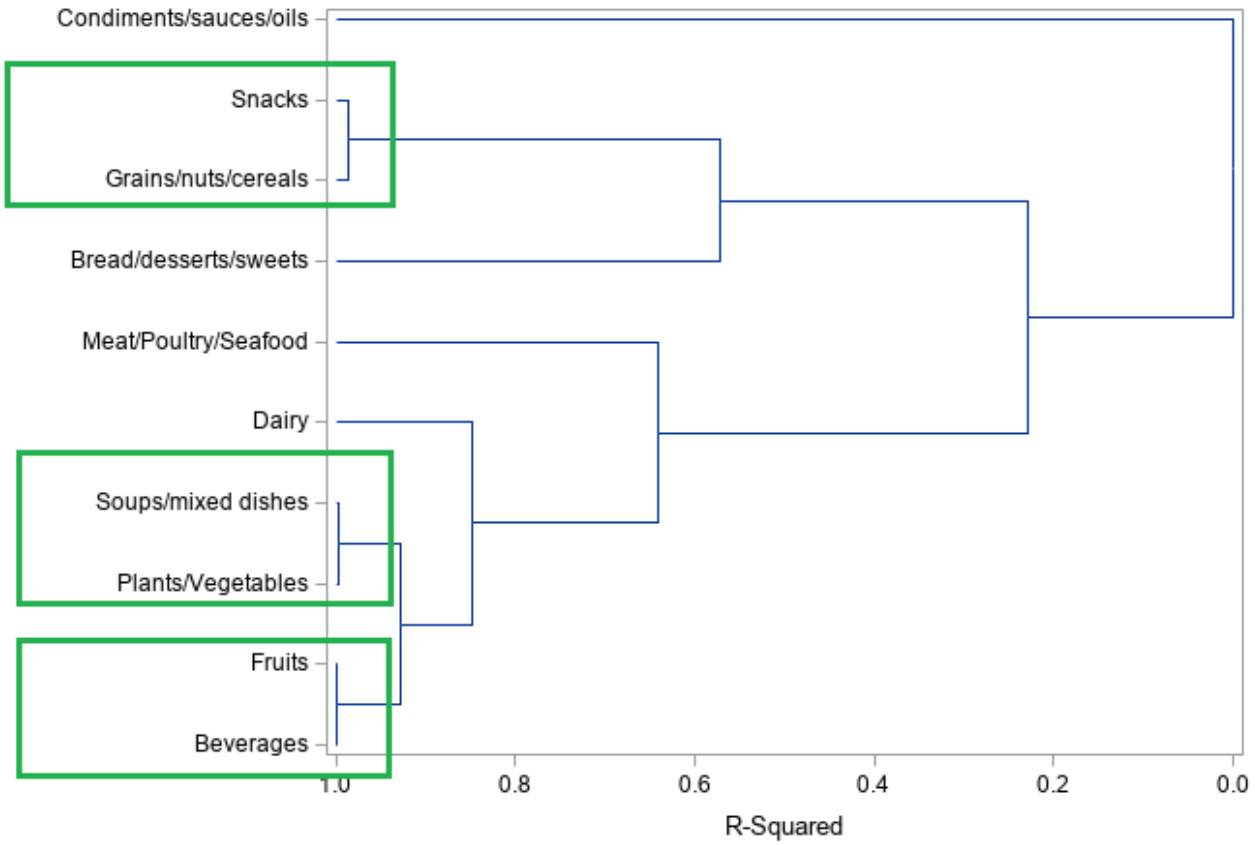
Cluster analysis was used to determine whether there are groups of foods similar to one another based on nutrient composition. Two cluster analyses resulted in discernable and significant cluster groupings.

Figure 10: Clusters of Foods Based on Similarity in Vitamin Composition



In the first model shown in Figure 10 above, three main clusters of more than one type of food were identified based on similarity in their overall vitamin content. For example, as seen in the first cluster, dairy products, meat/poultry/seafood, and grains/nuts/cereals generally have similar vitamin composition. Similarly, the other two clusters contain food groups with similar vitamin nutrient contents. Among all food groups, soups/mixed dishes and breads/desserts/sweets are the most similar in terms of vitamin content as they represent the first cluster formed. Technically, the category condiments/sauces/oils also belongs to a cluster of its own, but it does not have similar vitamin content to any other food category.

Figure 11: Clusters of Foods Based on Similarity in Macronutrient Composition



In the second model which used macronutrient variables, three cluster groups of two food groups each were identified based on similarities in macronutrient content. For example, snacks and grains/nuts/cereals have similar macronutrient composition and belong to the same cluster.

3.6. DISCRIMINANT ANALYSIS

Discriminant analysis was used to create predictive models which classify foods into disease-related nutritional variables such as HighCalories, HighSugar, and HighCholesterol. These models are developed for scenarios when the underlying nutritional variable may not be readily available. To ensure usability, the models were trained using 75% of the observations and then tested on the remaining 25% of the data. Moreover, a necessary assumption is made wherein the underlying nutritional variable is discarded from the model; for example, the number of calories must not be used when classifying food as HighCalorie. The use case of these models goes beyond proving that nutrient relationships exist in that the models enable the classification of foods when nutritional data on certain variables is incomplete, not readily available, or dubious. The discriminant analysis resulted in three final models which are described below.

Model 1:

Predicting whether a food should be classified as HighCalorie based on its macronutrients

Predictor variables deemed significant by Stepwise Discriminant Analysis:
TotalFat, Protein, Sugar, Carbohydrate

Discriminant function:
Quadratic

Figure 12: Model 1 Performance on Test Data (Confusion Matrix)

The DISCRIM Procedure
Classification Summary for Test Data: WORK.NUTRITIONTEST
Classification Summary using Quadratic Discriminant Function

Observation Profile for Test Data	
Number of Observations Read	1032
Number of Observations Used	1032

Number of Observations and Percent Classified into HighCalories			
From HighCalories	0	1	Total
0	993	4	997
	99.60	0.40	100.00
1	0	35	35
	0.00	100.00	100.00
Total	993	39	1032
	96.22	3.78	100.00
Priors	0.97542	0.02458	

Error Count Estimates for HighCalories			
	0	1	Total
Rate	0.0040	0.0000	0.0039
Priors	0.9754	0.0246	

As evident in the confusion matrix, this model performed very well with an overall error rate estimate of only 0.39%. The discriminant function only misclassified 4 out of 1032 observations in the test data. The model certifies that the macronutrient profile of a food is a strong predictor of whether or not it is high in calories.

Model 2:

Predicting whether a food should be classified as HighSugar based on nutritional variables (excluding Sugar content)

Predictor variables deemed significant by Stepwise Discriminant Analysis:
TotalFat, Protein, Carbohydrate, VitaminD, Iron, Potassium, VitaminE, SaturatedFat

Discriminant function:
Linear

Figure 13: Model 2 Performance on Test Data (Confusion Matrix)

The DISCRIM Procedure
Classification Summary for Test Data: WORK.NUTRITIONTEST
Classification Summary using Linear Discriminant Function

Observation Profile for Test Data	
Number of Observations Read	1032
Number of Observations Used	1032

Number of Observations and Percent Classified into HighSugar			
From HighSugar	0	1	Total
0	721	72	793
	90.92	9.08	100.00
1	91	148	239
	38.08	61.92	100.00
Total	812	220	1032
	78.68	21.32	100.00
Priors	0.78525	0.21475	

Error Count Estimates for HighSugar			
	0	1	Total
Rate	0.0908	0.3808	0.1531
Priors	0.7853	0.2147	

This model yielded an overall misclassification rate estimate of 15.31% on the test dataset.

Figure 14: Model 3 Performance on Test Data (Confusion Matrix)

Model 3:

Predicting whether a food should be classified as HighCholesterol based on nutritional variables (excluding Sugar content)

Predictor variables deemed significant by Stepwise Discriminant Analysis:
Protein, Carbohydrate, Potassium, VitaminE, SaturatedFat

Discriminant function:
Linear

The DISCRIM Procedure
Classification Summary for Test Data: WORK.NUTRITIONTEST
Classification Summary using Quadratic Discriminant Function

Observation Profile for Test Data	
Number of Observations Read	1032
Number of Observations Used	1032

Number of Observations and Percent Classified into HighCholesterol

From HighCholesterol	0	1	Total
0	711	282	993
	71.60	28.40	100.00
1	11	28	39
	28.21	71.79	100.00
Total	722	310	1032
	69.96	30.04	100.00
Priors	0.94696	0.05304	

Error Count Estimates for HighCholesterol

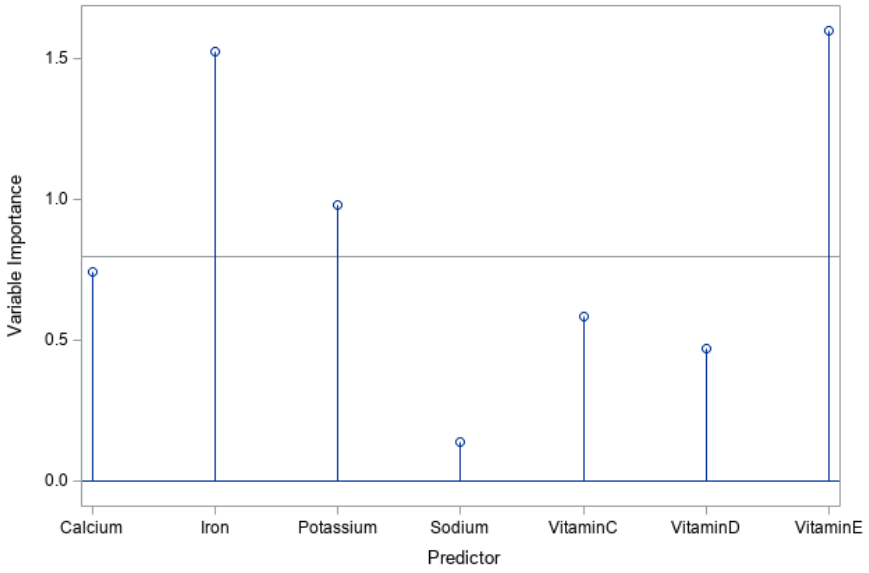
	0	1	Total
Rate	0.2840	0.2821	0.2839
Priors	0.9470	0.0530	

This model returned an overall misclassification rate estimate of 28.39% on the test dataset.

3.7. PARTIAL LEAST SQUARES REGRESSION

A multiple-response PLS model was used to extend the results from the canonical correlation analysis where it was found that macronutrients are correlated with micronutrients. PLS was used primarily to inspect the variable importance of micronutrients (set of independent variables) in predicting macronutrients (set of dependent variables)

Figure 15. Variable Importance Plot of Micronutrients in Predicting Macronutrients



Based on the variable importance plot above, the micronutrients Iron, Potassium, and Vitamin E are the strongest predictors of macronutrients.

Figure 16. Profiles of Centered and Scaled Estimates – Micronutrients and Macronutrients

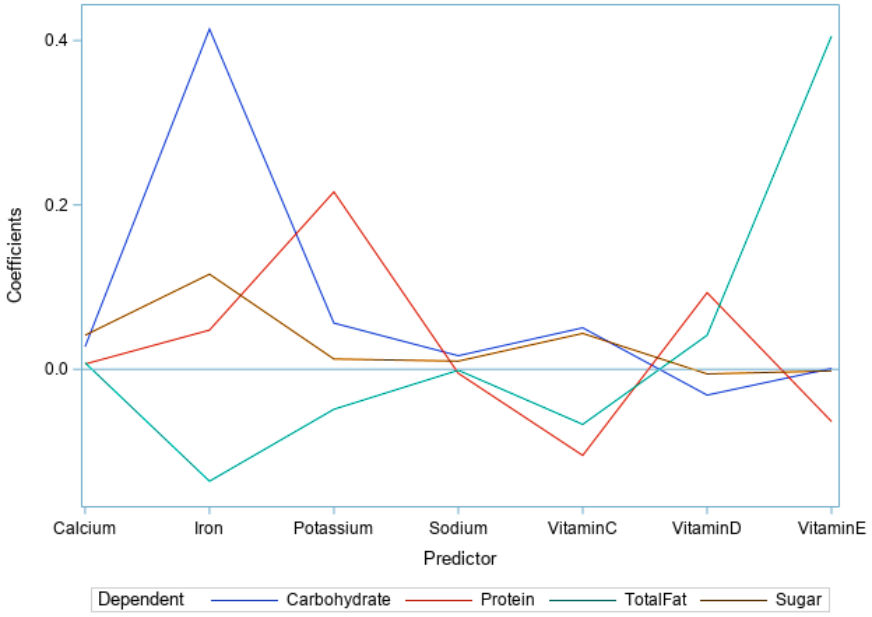


Figure 16 above shows how well each micronutrient predicts each macronutrient. All micronutrients except Vitamin D and Vitamin E are strong predictors of Carbohydrates and Sugar. Vitamin E is a strong predictor of TotalFat. For protein, the best micronutrient predictors are Iron, Potassium, and Vitamin D.

4. DISCUSSION AND CONCLUSIONS

The results from the different multivariate analyses present several insights and implications related to nutrition issues. The Principal Components Analysis and Factor Analysis explained how nutritional variables are related to each other as well as to different kinds of foods. The variables TotalFat, SaturatedFat, and Carbohydrates were most closely related to Calories. Foods such as condiments/sauces/oils, snacks, and bread/desserts/sweets were associated with having a high number of calories. The correspondence analysis also showed that condiments/sauces/oils are most associated with exceeding recommended caloric intakes. Excess caloric-intake over time is an underestimated factor that leads to obesity (NIH, 2021). It follows that refraining from foods associated with high calories and avoiding foods with high amounts of total fat, saturated fat, and carbohydrates would considerably reduce the likelihood of obesity in individuals. On the other hand, these high-caloric options should be preferred when dealing with

malnutrition. The third principal component illuminated that grains/nuts/cereals and meat/poultry/seafood are strongly associated with healthy micronutrients despite being high in calories, which makes them suitable food groups to combat malnutrition. As such, it is vital to choose the right foods with the right nutrient composition depending on health requirements.

Canonical correlation analyses showed positive correlations among variable groups. Vitamins were correlated with minerals and micronutrients were correlated with macronutrients. The PLS model built on top of this finding revealed that among micronutrients, Iron, Potassium, and Vitamin E are the strongest predictors of macronutrients. Understanding these variable groups is important because the body needs them in different amounts. According to WSU (2021), most adults get sufficient amounts of macronutrients and micronutrients without having to track intake. However, adults with medical conditions such as diabetes need to pay special attention to their intake. Understanding how micronutrients relate to macronutrients can help in planning diet combinations and regulating intake for people with special medical conditions or risk factors. Decisions pertaining to diet combinations can also draw from the cluster analysis in which it was found that foods can be categorized into three major groups in terms of similar vitamins and macronutrient composition. It was found that grains/nuts/cereals, dairy products, and meat/poultry/seafood belong to a cluster with similar vitamins and macronutrient composition as they are associated with Protein, Vitamin D, and Cholesterol. For example, people lacking in these nutrients would benefit from consuming foods that belong to this food cluster.

Excess intake of Calories, Sugar, and Cholesterol is a leading risk factor for obesity and several health problems (WSU, 2021). In light of this, discriminant analysis was used to create reasonably accurate models that predict whether a food is high in these three nutritional variables. TotalFat, Protein, Sugar, and Carbohydrate were found to be significantly strong predictors of high-caloric foods. Both micronutrients and macronutrients except Vitamin C, Calcium, and Sodium were strong predictors of high-sugar foods. For high cholesterol, the strongest predictors were protein, carbohydrate, potassium, Vitamin E, and saturated fat. Understanding these nutritional variables from a multivariate angle can help mitigate the uninformed consumption of foods that are high in calories, sugar, or cholesterol.

The multivariate statistical techniques used to analyze the data were selected with the purpose of understanding food nutrient composition in the context of rising nutritional concerns such as obesity and malnutrition. The results have revealed interesting patterns and relationships among the nutritional variables that not only make for interesting insights but also pose new questions worth addressing. This analysis can be used as a starting point for more extensive

analysis of food nutrient composition using multivariate analysis. First, subcategories can be added to the USDA food nutrition data to enable the same forms of multivariate analyses and deeper insights within each food category. For example, it may be meaningful to study specific kinds of meats within the meat/poultry/seafoods category. Second, although arduous on the data collection front, additional nutritional variables such as Fiber and other types of essential vitamins can be blended into the dataset for deeper analysis. Finally, more intricate derived variables such as macronutrient ratios can be added to the dataset. With nutrition-related health issues rising exponentially, the need to employ efficient multivariate statistical methods to draw out insights will only emulsify. The degree to which analysis projects in this space are rooted in solving pressing nutritional issues would ultimately determine project success.

REFERENCES

CDC. (2018). Leading Causes of Death. *Centers for Disease Control and Prevention*. Retrieved from: https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm

Daousi, C., Casson, F., Gill, C., MacFarlane, A., Wilding, J., Pinkney, J. (2006). Prevalence of obesity in type 2 diabetes in secondary care: association with cardiovascular risk factors. Postgraduate Medical Journal, 82(966), 280–284. doi:10.1136/pmj.2005.039032

De Moraes, A. C. F., Adami, F., & Falcão, M. C. (2012). Understanding the correlates of adolescents' dietary intake patterns. A multivariate analysis. Appetite, 58(3), 1057–1062. doi:10.1016/j.appet.2012.01.024

NIA. (2017). Know Your Food Groups. *National Institute on Aging*. Retrieved from: https://www.nia.nih.gov/health/know-your-food-groups

NIH. (2021). Nutrient Recommendations: Dietary Reference Intakes (DRI). *National Institutes of Health*. Retrieved from: https://ods.od.nih.gov/HealthInformation/Dietary_Reference_Intakes.aspx#:~:text=Recommended%20Dietary%20Allowance%20(RDA)%3A,assumed%20to%20ensure%20nutritional%20adequacy.

North, K., Emmet, P. (2000). Multivariate analysis of diet among three-year-old children and associations with socio-demographic characteristics. European Journal of Clinical Nutrition, 54, 73-80

ODPHP. (2021). Major categories and subcategories used in DGAC analyses of WWEIA Food Categories. *Office of Disease Prevention and Health Promotion*. Retrieved from: https://health.gov/our-work/food-nutrition/previous-dietary-guidelines/2015/advisory-report/appendix-e-2/appendix-e-27

Rice, A., Sacco, L., Hyder, A., Black, R. (2000). Malnutrition as an underlying cause of childhood deaths associated with infectious diseases in developing countries. *Bulletin of the World Health Organization*.

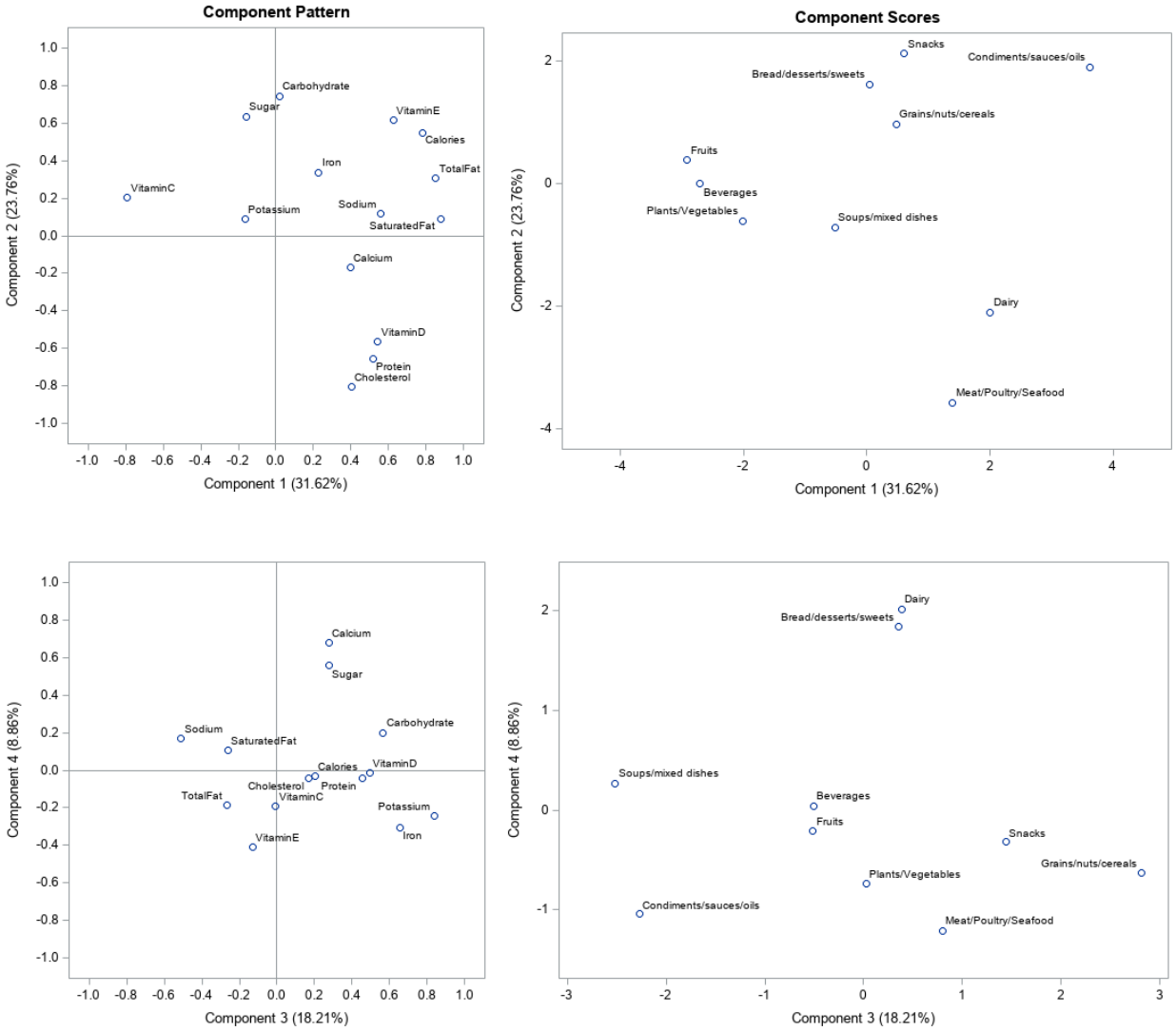
USDA (2021). FoodData Central Data. *U.S. Department of Agriculture*. Retrieved from: https://fdc.nal.usda.gov/download-datasets.html

WHO. (2020). Children: improving survival and well-being. *World Health Organization*.

WSU. (2021). Nutrition Basics. Washington State University. Retrieved from: https://mynutrition.wsu.edu/nutrition-basics

APPENDICES

Appendix A – Component Pattern and Component Score Plots for PCA



Appendix B – SAS Code

```
/* FOOD NUTRIENT COMPOSITION MULTIVARIATE ANALYSIS*/

/*Importing datasets*/
proc import datafile="C:\Users\user\Google Drive\School\Master of
Analytics\161.762 Multivariate Analysis for Big Data\Final
Project\USDAfoodnutrition.csv"
out=nutrition dbms=csv replace; getnames=yes; run;

proc import datafile="C:\Users\user\Google Drive\School\Master of
Analytics\161.762 Multivariate Analysis for Big Data\Final
Project\USDAfoodnutritiontrain.csv"
out=nutritiontrain dbms=csv replace;  getnames=yes; run;

proc import datafile="C:\Users\user\Google Drive\School\Master of
Analytics\161.762 Multivariate Analysis for Big Data\Final
Project\USDAfoodnutritiontest.csv"
out=nutritiontest dbms=csv replace;  getnames=yes; run;

proc import datafile="C:\Users\user\Google Drive\School\Master of
Analytics\161.762 Multivariate Analysis for Big Data\Final
Project\USDAfoodnutritioncategorycrosstab.csv"
out=nutritioncategory dbms=csv replace;  getnames=yes; run;

proc import datafile="C:\Users\user\Google Drive\School\Master of
Analytics\161.762 Multivariate Analysis for Big Data\Final
Project\USDAfoodnutritioncategory.csv"
out=nutritioncategorytable dbms=csv replace;  getnames=yes; run;

proc import datafile="C:\Users\user\Google Drive\School\Master of
Analytics\161.762 Multivariate Analysis for Big Data\Final
Project\usdafoodnutritionhighscorecrosstab.csv"
out=nutritioncategoryhigh dbms=csv replace;  getnames=yes; run;

proc import datafile="C:\Users\user\Google Drive\School\Master of
Analytics\161.762 Multivariate Analysis for Big Data\Final
Project\usdafoodnutritionhighscore.csv"
out=nutritioncategoryhightable dbms=csv replace;  getnames=yes; run;

proc import datafile="C:\Users\user\Google Drive\School\Master of
Analytics\161.762 Multivariate Analysis for Big Data\Final
Project\usdafoodnutritiondummy.csv"
out=nutritiondummy dbms=csv replace;  getnames=yes; run;

/*Variable Groupings*/
%let Vitamins = VitaminC VitaminD VitaminE;
%let Minerals = Calcium Iron Potassium Sodium;
%let Macronutrients = Carbohydrate Protein TotalFat Sugar;
%let Micronutrients = Calcium Iron Potassium Sodium VitaminC VitaminD VitaminE;

/*Data description*/
proc contents data=nutrition;
run;


/*Draftsman plot of scatterplot matrix*/
```

```
ods graphics / reset;
title 'Draftsman Plot of Nutrition Data ';
proc sgscatter data=nutrition;
matrix Calories -- VitaminD / diagonal=(histogram normal);
run;

ods select Cov PearsonCorr;
proc corr data=nutrition noprob outp=OutCorr nomiss cov;
var Calories -- VitaminD;
run;

/*Princomp procedure for initial PCA*/
ods graphics / reset;
title 'PCA on nutrition dataset';
proc princomp data=nutritioncategory out=prin;
var Calories -- VitaminD;
run;
title;

/*PCA graphical exploration*/
ods graphics/ imagemap;
proc princomp data=nutritioncategory n=4 out=prin4 prefix=comp
plots=all;
var Calories -- VitaminD;
id Category;
run;

/*PCA Biplot*/
proc prinqual data=nutritioncategory mdpref=2;
transform identity(Calories -- VitaminD);
id Category;
run;

/*Preliminary factor analysis with scree and variance plots*/
proc factor data=nutrition  plots=all priors=smc;
title 'factor analysis';
var Calories -- VitaminD;
run;

/*Using an Orthogonal varimax rotation with 2 factors specified*/
ods select orthrotfactpat patternplot;
proc factor data=nutrition r=v n=4 plots=all
flag=0.4 fuzz=0.4; /* assumption made to eliminate nonsignificant values*/
title2 'varimax rotation';
var Calories -- VitaminD;
run;


/*Correspondence analysis*/
ods graphics on;
proc corresp data=nutritioncategoryhightable cross=rows observed rp short;
      tables Category, Variable;
      weight Average;
run;


/*Canonical correlation*/
```

```
ods output cancorr=a;
proc cancorr data=nutritioncategory ncan=2 vprefix=vitamins wprefix=minerals
vname='Vitamin' wname= 'Mineral' out=vitminout;
var &Vitamins;
with &Minerals;
run;

proc sgscatter data=vitminout;
plot vitamins1*minerals1 vitamins2*minerals2;
run;

proc sgplot data=a;
series y=squcancorr x=number /markers;
xaxis integer; run;

ods output cancorr=a;
proc cancorr data=nutritioncategory ncan=2 vprefix=Micronutrient
wprefix=Macronutrient
vname='Micronutrient' wname= 'Macronutrient' out=micromacroout;
var &Micronutrients;
with &Macronutrients;
run;

proc sgscatter data=micromacroout;
plot Micronutrient1*Macronutrient1
       Micronutrient2*Macronutrient2;
run;

proc sgplot data=a;
series y=squcancorr x=number /markers;
xaxis integer; run;

/*Canonical Discriminant Analysis*/

/*Predicting High Calories Model*/

/*Stepwise Discriminant Analysis to determine vars*/
proc stepdisc data=nutritiontrain method=stepwise;
class HighCalories;
var &Macronutrients;
run;

/*Quadratic Discriminant Analysis*/
proc discrim data=nutritiontrain pool=test
testdata=nutritiontest testlisterr slpool=.05;
title1 'Test for equality of covariance matrices';
title2 'and quadratic discriminant analysis';
class HighCalories;
priors prop;
var &_STDVAR;
run; title;


/*Predicting High Sugar Model*/

/*Stepwise Discriminant Analysis to determine vars*/
proc stepdisc data=nutritiontrain method=stepwise;
```

```
class HighSugar;
var Calories Protein TotalFat SaturatedFat Sodium Carbohydrate
      VitaminC VitaminD VitaminE Calcium Iron Potassium Cholesterol;
run;

/*Fisher LDA*/

proc discrim data=nutritiontrain pool=yes
testdata=nutritiontest testlisterr;
class HighSugar;
priors prop;
var &_STDVAR;
run;

/*Predicting High Cholesterol Model*/

/*Stepwise Discriminant Analysis to determine vars*/
proc stepdisc data=nutritiontrain method=stepwise;
class HighCholesterol;
var Calories Protein TotalFat SaturatedFat Sugar Carbohydrate
      VitaminC VitaminD VitaminE Calcium Iron Potassium Sodium;
run;

/*Fisher LDA*/
proc discrim data=nutritiontrain pool=no
testdata=nutritiontest testlisterr;
class HighCholesterol;
priors prop;
var &_STDVAR;
run;


/*Cluster Analysis*/

/*Vitamin cluster groups*/

/*Standardizing selected variables*/
proc standard data=nutritioncategory mean=0 std=1 out=stan;
var Calories -- VitaminD;
run;

data stanclus;
      set stan;
      drop &Macronutrients &Minerals Calories SaturatedFat Cholesterol;
run;


/*Clustering*/
ods graphics on;
proc cluster data=stanclus method=single ccc pseudo outtree=tree print=15
plots=den(height=rsq);
id Category;
run;

/*Macronutrient cluster groups*/

/*Standardizing selected variables*/
```

```
proc standard data=nutritioncategory mean=0 std=1 out=stan;
var Calories -- VitaminD;
run;

data stanclus;
      set stan;
      drop &Vitamins &Minerals Calories SaturatedFat Cholesterol;
run;


/*Clustering*/
ods graphics on;
proc cluster data=stanclus method=single ccc pseudo outtree=tree print=15
plots=den(height=rsq);
id Category;
run;


/*PLS Analysis*/


/*Multiple response PLS Regression*/
ods graphics on;
proc pls data = nutrition method=pls(algorithm=nipals)
cv=one cvtest(seed=608789001)
plot=(vip xyscores yscores xscores yweightplot xweightplot parmprofiles dmodxy);
model &Macronutrients = &Micronutrients;
run;


/*Single response PLS Regression*/
ods graphics on;
proc pls data = nutrition method=pls(algorithm=nipals)
cv=one cvtest(seed=608789001)
plot=(vip xyscores xscores parmprofiles dmod);
model Calories = Protein -- VitaminD;
run;
```

Appendix C – Python Code for Supplementary Visualizations

```
import matplotlib
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import seaborn as sns
from pylab import rcParams
from bs4 import BeautifulSoup
from matplotlib import cm

#setting up plotting options

%matplotlib inline
%reload_ext autoreload
%autoreload 2

rcParams['figure.figsize'] = 14,8
rcParams['font.size'] = 20
rcParams['axes.facecolor'] = 'white'
plt.style.use('seaborn-white')
sns.set_style('white')
sns.set_context('paper', font_scale=1.5)
palette = sns.color_palette("Greens")

#display options for Pandas

pd.set_option('max_columns', 20)
pd.set_option('max_rows', 20)

#import data

usdanutrition = pd.read_csv('USDAfoodnutrition.csv')
usdanutritioncategory = pd.read_csv('USDAfoodnutritioncategory.csv')
usdanutritioncategorycrosstab = pd.read_csv('USDAfoodnutritioncategorycrosstab.csv')
usdanutritionhighscore = pd.read_csv('USDAfoodnutritionhighscore.csv')
usdanutritionhighscorecrosstab = pd.read_csv('USDAfoodnutritionhighscorecrosstab.csv')

#visualizations

usdanutritionforcorr = usdanutrition.drop(columns=[ 'HighCalories',
                                                    'HighProtein',
                                                    'HighTotalFat',
                                                    'HighCarbohydrate',
                                                    'HighSodium',
                                                    'HighSaturatedFat',
                                                    'HighCholesterol',
                                                    'HighSugar',
                                                    'HighVitaminC',
                                                    'HighVitaminD',
                                                    'HighVitaminE',
```

```
                                    'HighCalcium',
                                    'HighIron',
                                    'HighPotassium',
                                    'ID'])
usdanutritioncorr = usdanutritionforcorr.corr()


#cutting out redundant values from upper right triangle in heatmap

mask = np.zeros_like(usdanutritioncorr)
triangle_indices = np.triu_indices_from(mask)
mask[triangle_indices]=True

#plotting into heatmap

sns.heatmap(usdanutritioncorr, cmap="Greens", mask=mask, annot=True, fmt='.2f')
plt.title("Correlation Matrix of Nutritional Variables\n", size=17)
plt.xticks(rotation=50)

usdanutritioncategorycrosstab.plot(kind="bar", subplots=True,
                        x="Category", sharey=False,
                        layout=(5,3), color="#33bb77", figsize=(16,13))
```