Leading Nutrition and Healthy choice

FINAL PROJECT REPORT

AMCS602

TEAM1
Bingyu Zhang
Jie Luo
Yichen Zhao
Yishan Shen

FINAL VERSION

12/12/2021

TABLE OF CONTENTS

1. PROJECT SUMMARY

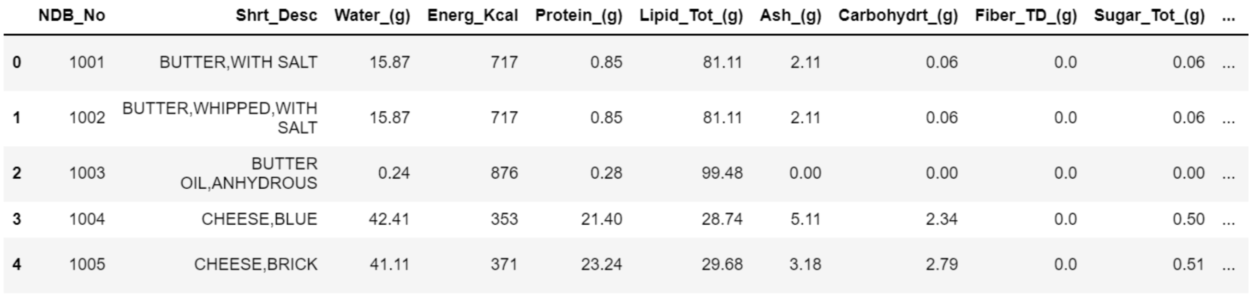
Adequate nutrition is one of the pillars of health, and the mission of the project aims to solve two problems. One is how to select food sensibly to supplement the nutrition we need, and the other one is how to match food to improve the quality of our diet.

As for the source of the data, we chose an abbreviated version of the Standard Reference (SR) from the USDA National Nutrient Database, which is the major source of food composition data in the United States and is owned by the United States Department of Agriculture (USDA). After preprocessing the data, we mainly used PCA and K-means methods to analyze the data to draw our conclusions.

2. PROJECT ELABORATION

2.1 DATA PREVIEW

DATA DISPLAY



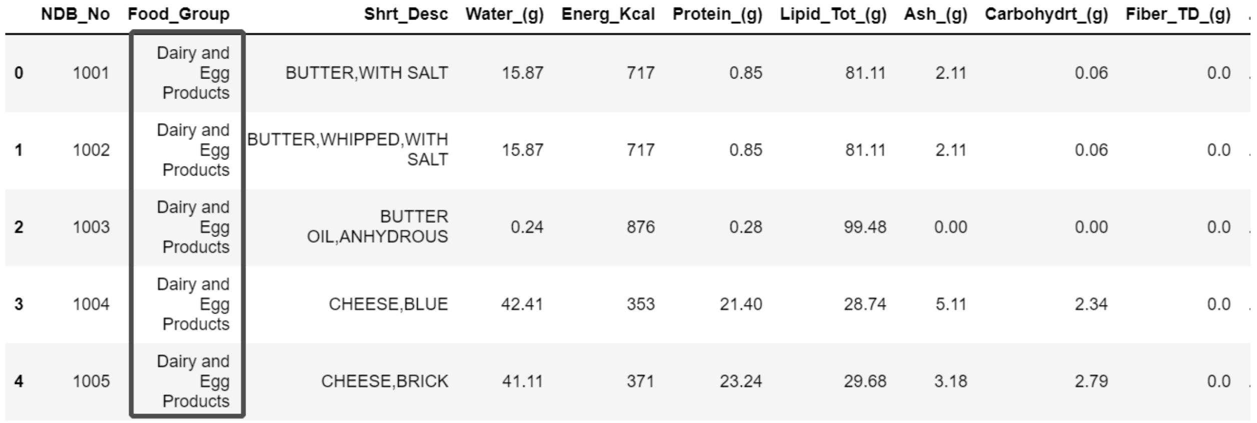
	NDB_No	Shrt_Desc	Water_(g)	Energ_Kcal	Protein_(g)	Lipid_Tot_(g)	Ash_(g)	Carbohydrt_(g)	Fiber_TD_(g)	Sugar_Tot_(g)	...
0	1001	BUTTER,WITH SALT	15.87	717	0.85	81.11	2.11	0.06	0.0	0.06	...
1	1002	BUTTER,WHIPPED,WITH SALT	15.87	717	0.85	81.11	2.11	0.06	0.0	0.06	...
2	1003	BUTTER OIL,ANHYDROUS	0.24	876	0.28	99.48	0.00	0.00	0.0	0.00	...
3	1004	CHEESE,BLUE	42.41	353	21.40	28.74	5.11	2.34	0.0	0.50	...
4	1005	CHEESE,BRICK	41.11	371	23.24	29.68	3.18	2.79	0.0	0.51	...

EXPLANATION

The data set contains 8,618 samples, which are the mixed food items. Each food item is numbered and measured up to 150 food components (it is 51 in the abbreviated version) such as water, protein, carbohydrate, minerals and vitamins content as illustrated above.

2.2 DATA PRE-PROCESSING

FOOD GROUPS



	NDB_No	Food_Group	Shrt_Desc	Water_(g)	Energ_Kcal	Protein_(g)	Lipid_Tot_(g)	Ash_(g)	Carbohydrt_(g)	Fiber_TD_(g)	.
0	1001	Dairy and Egg Products	BUTTER,WITH SALT	15.87	717	0.85	81.11	2.11	0.06	0.0	.
1	1002	Dairy and Egg Products	BUTTER,WHIPPED,WITH SALT	15.87	717	0.85	81.11	2.11	0.06	0.0	.
2	1003	Dairy and Egg Products	BUTTER OIL,ANHYDROUS	0.24	876	0.28	99.48	0.00	0.00	0.0	.
3	1004	Dairy and Egg Products	CHEESE,BLUE	42.41	353	21.40	28.74	5.11	2.34	0.0	.
4	1005	Dairy and Egg Products	CHEESE,BRICK	41.11	371	23.24	29.68	3.18	2.79	0.0	.

EXPLANATION

The USDA National Nutrition Database divided the 8,618 sample data into 25 food groups (e.g. dairy and egg products, nut and seed products, vegetables and vegetable products, beef products). Here we add the label of food groups to our original data as show above.

NON-NUMERICAL and NON-IMPORTANT COLUMNS

For the non-numerical columns “GmWt_Desc1” and “GmWt_Desc2”, we directly remove them from data since the result will not be affected without them as they merely serve as a description of the amount of ingredients. For those features that are in high proportion but irrelevant to our focus, such as the feature “water_(g)”, we remove them as well.

MISSING DATA

For features with missing data, we first delete those features with the number of missing data greater than 3,500 directly since it accounts for roughly 40% of the overall data. Meanwhile, for those features with insignificant number of missing data, we filled them with the mean value of the feature from the same food group.

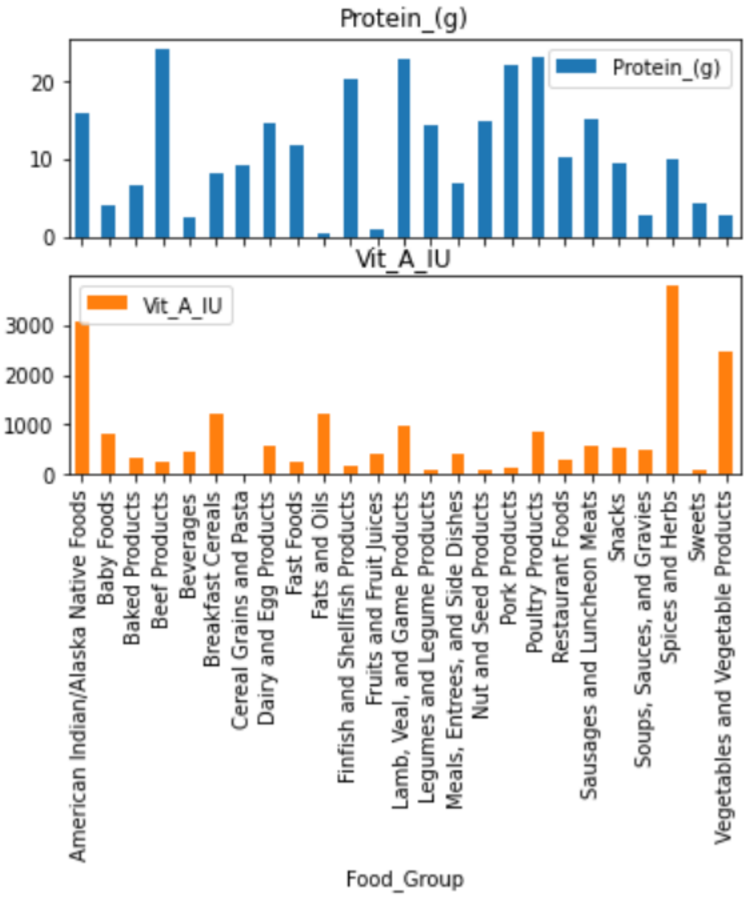
SMALL VARIANCE

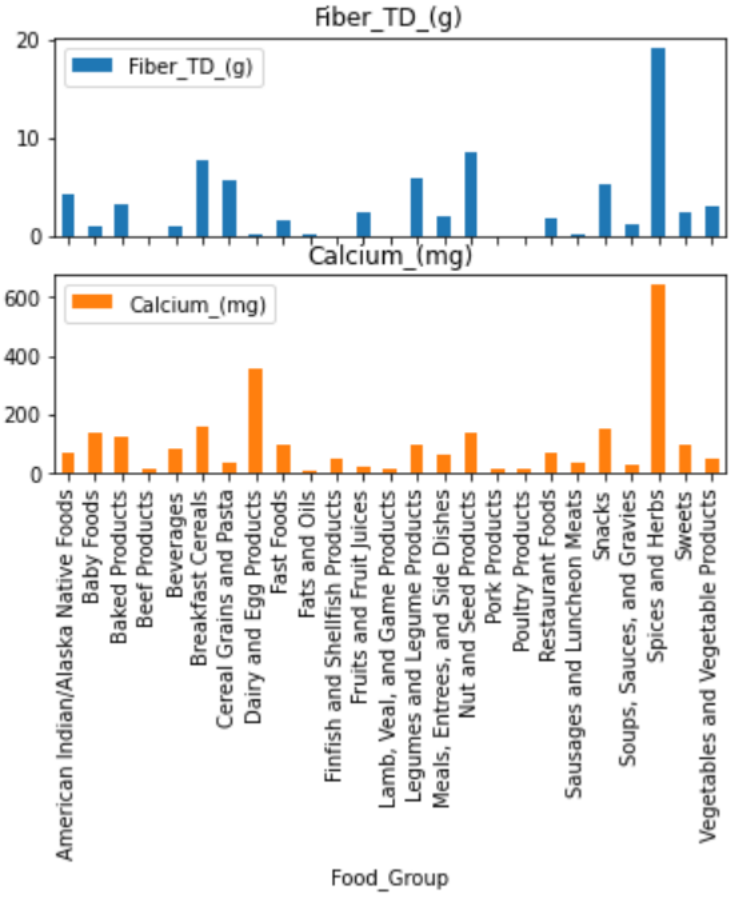
We would also like to delete those features with small variance, because the distribution of those features will not contribute to the result of our analysis. It shows that there is no such feature, so all remaining features are reserved through this step.

2.3 DATA EXPLORATION

In order to further understand the distribution of the features, we selected four representative features (e.g. fiber, calcium, protein and vitamin A) and checked whether there are differences among different food groups.

OUTCOME NAME	The distributions of protein and vitamin A



OUTCOME NAME	The distributions of fiber and calcium
	

2.4 NORMALIZATION

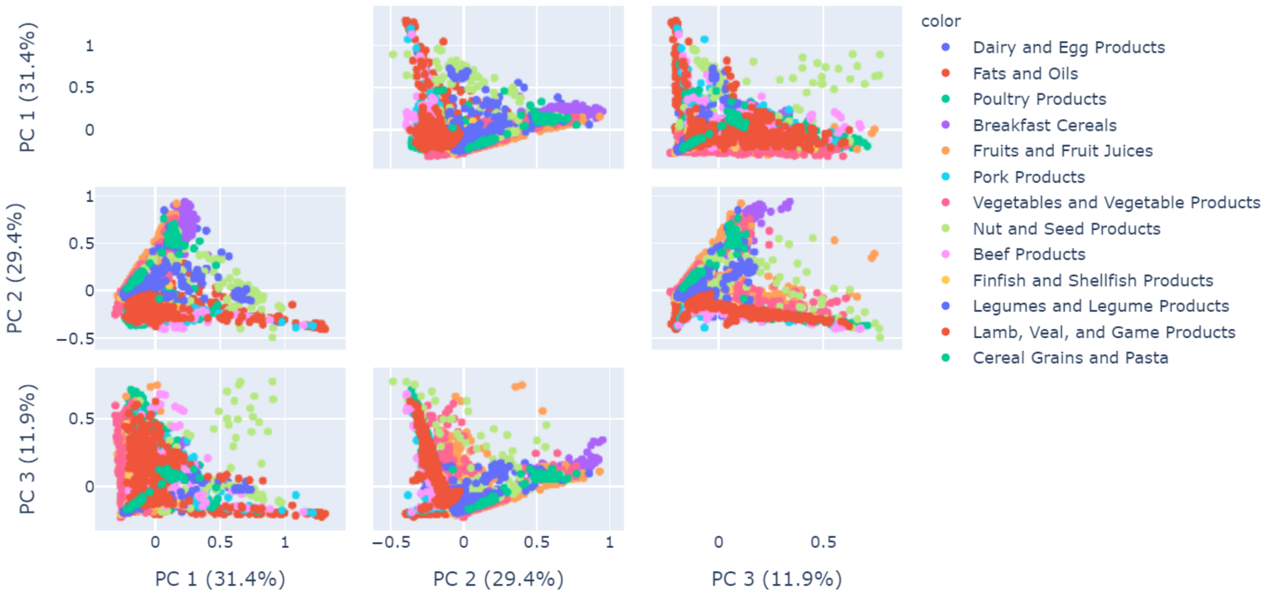
This section details how we normalize our dataset before applying PCA.

Before applying PCA, we noticed a variety in range of values among features. The difference between the max and min values in some feature is much larger than in others. By adopting Min-Max feature scaling method, we bring values of all features into the range [0, 1]. The corresponding formula is as follows:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

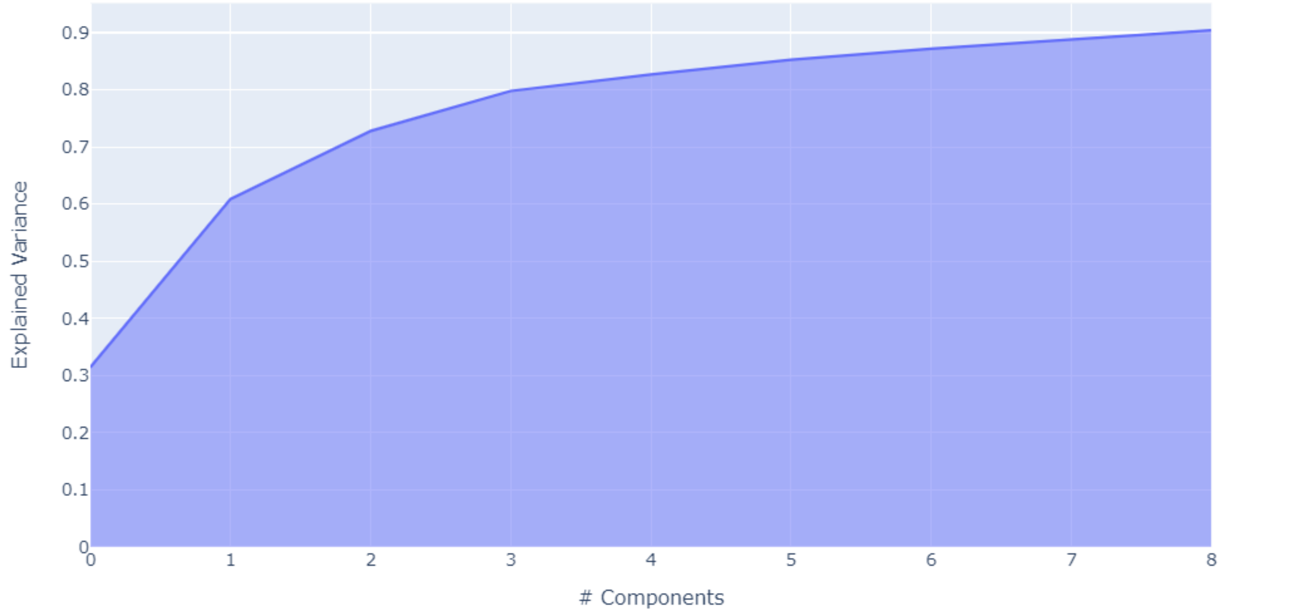
2.5 PCA

In order to solve the problems we raised at the beginning, it is more reasonable to filter grouped food. The reason is that we found some food groups are not suitable as a generally accepted food to balance people's daily diet in the original food groups, such as baby food. Therefore, we performed PCA processing on the data containing only 13 food groups that we considered reasonable. Bellowing are the results of applying PCA.



EXPLANATION

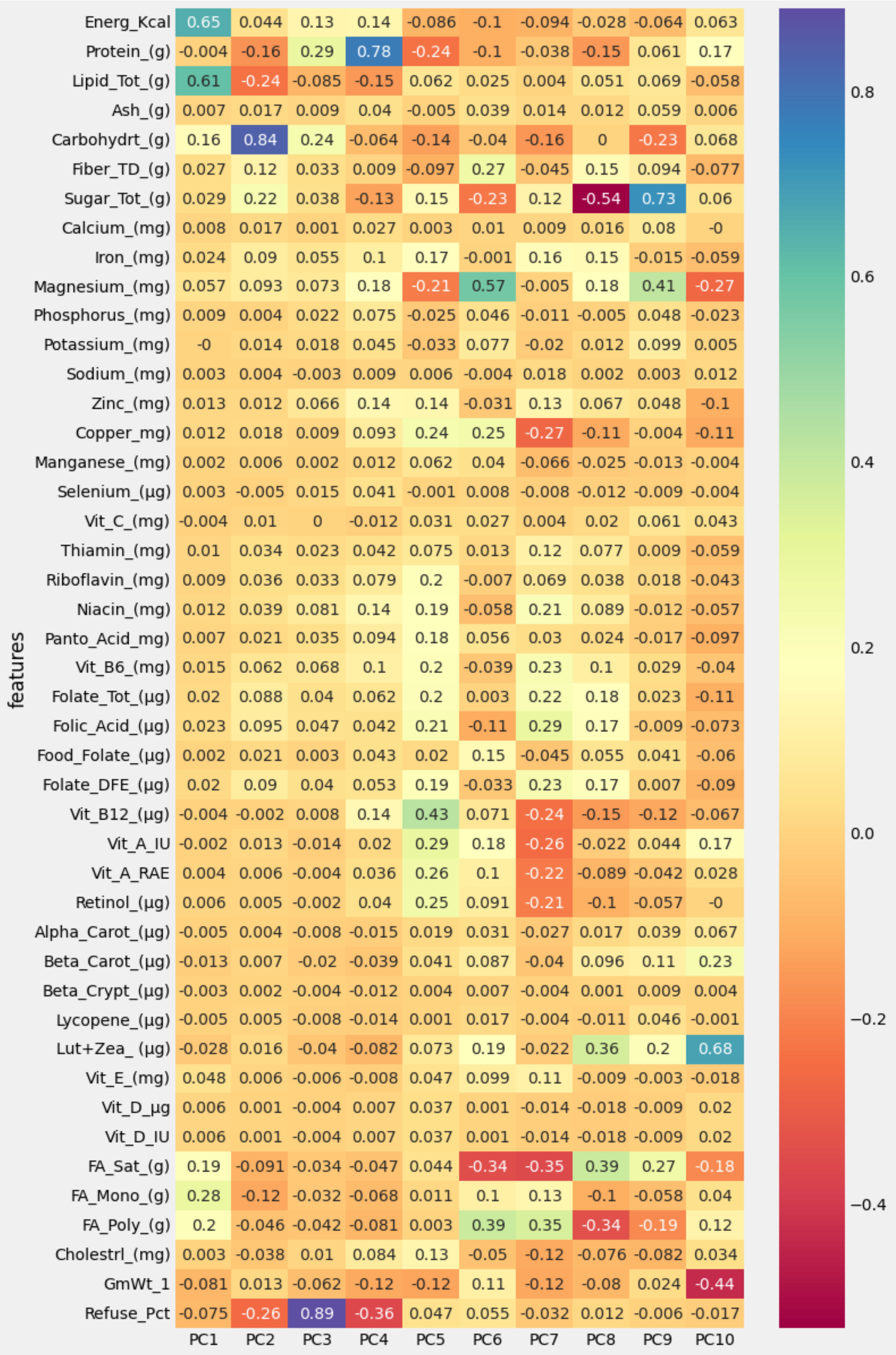
These graphs are the projection graphs of the first three principal components. Each shows the projection of original data points onto the plane composed of two of the first three principal components. The pattern shown in each graph indicates the direction of maximized variance for its corresponding principal components.



EXPLANATION

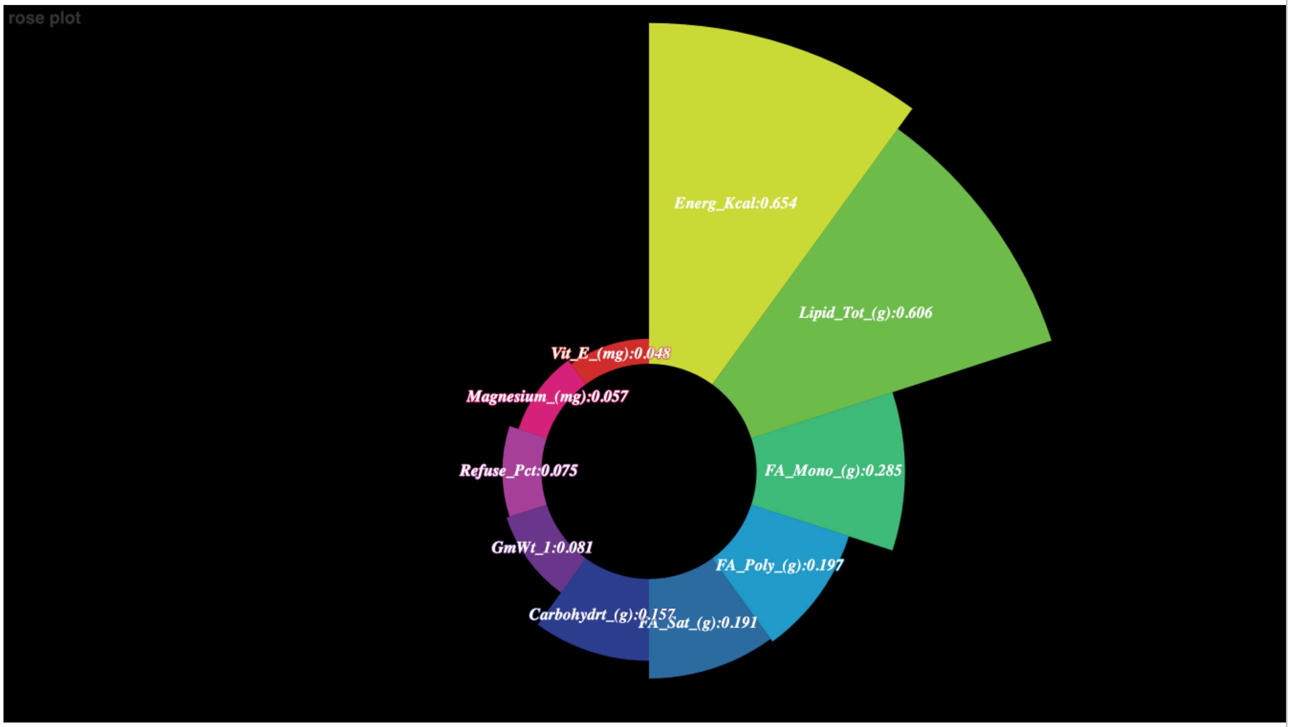
The figure shows the accumulated varience explained by principal components as the number of considered components increases. The explanation reaches about 90% when the first 7 components come into consideration.

Next, we concerned about the specific features contained in each component. It helps us find the most complementary or the least relevant nutrients for achieving the best contribution when matching the diet once we identify them.



features	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Energ_Kcal	0.65	0.044	0.13	0.14	-0.086	-0.1	-0.094	-0.028	-0.064	0.063
Protein_(g)	-0.004	-0.16	0.29	0.78	-0.24	-0.1	-0.038	-0.15	0.061	0.17
Lipid_Tot_(g)	0.61	-0.24	-0.085	-0.15	0.062	0.025	0.004	0.051	0.069	-0.058
Ash_(g)	0.007	0.017	0.009	0.04	-0.005	0.039	0.014	0.012	0.059	0.006
Carbohydrt_(g)	0.16	0.84	0.24	-0.064	-0.14	-0.04	-0.16	0	-0.23	0.068
Fiber_TD_(g)	0.027	0.12	0.033	0.009	-0.097	0.27	-0.045	0.15	0.094	-0.077
Sugar_Tot_(g)	0.029	0.22	0.038	-0.13	0.15	-0.23	0.12	-0.54	0.73	0.06
Calcium_(mg)	0.008	0.017	0.001	0.027	0.003	0.01	0.009	0.016	0.08	-0
Iron_(mg)	0.024	0.09	0.055	0.1	0.17	-0.001	0.16	0.15	-0.015	-0.059
Magnesium_(mg)	0.057	0.093	0.073	0.18	-0.21	0.57	-0.005	0.18	0.41	-0.27
Phosphorus_(mg)	0.009	0.004	0.022	0.075	-0.025	0.046	-0.011	-0.005	0.048	-0.023
Potassium_(mg)	-0	0.014	0.018	0.045	-0.033	0.077	-0.02	0.012	0.099	0.005
Sodium_(mg)	0.003	0.004	-0.003	0.009	0.006	-0.004	0.018	0.002	0.003	0.012
Zinc_(mg)	0.013	0.012	0.066	0.14	0.14	-0.031	0.13	0.067	0.048	-0.1
Copper_mg)	0.012	0.018	0.009	0.093	0.24	0.25	-0.27	-0.11	-0.004	-0.11
Manganese_(mg)	0.002	0.006	0.002	0.012	0.062	0.04	-0.066	-0.025	-0.013	-0.004
Selenium_(µg)	0.003	-0.005	0.015	0.041	-0.001	0.008	-0.008	-0.012	-0.009	-0.004
Vit_C_(mg)	-0.004	0.01	0	-0.012	0.031	0.027	0.004	0.02	0.061	0.043
Thiamin_(mg)	0.01	0.034	0.023	0.042	0.075	0.013	0.12	0.077	0.009	-0.059
Riboflavin_(mg)	0.009	0.036	0.033	0.079	0.2	-0.007	0.069	0.038	0.018	-0.043
Niacin_(mg)	0.012	0.039	0.081	0.14	0.19	-0.058	0.21	0.089	-0.012	-0.057
Panto_Acid_mg)	0.007	0.021	0.035	0.094	0.18	0.056	0.03	0.024	-0.017	-0.097
Vit_B6_(mg)	0.015	0.062	0.068	0.1	0.2	-0.039	0.23	0.1	0.029	-0.04
Folate_Tot_(µg)	0.02	0.088	0.04	0.062	0.2	0.003	0.22	0.18	0.023	-0.11
Folic_Acid_(µg)	0.023	0.095	0.047	0.042	0.21	-0.11	0.29	0.17	-0.009	-0.073
Food_Folate_(µg)	0.002	0.021	0.003	0.043	0.02	0.15	-0.045	0.055	0.041	-0.06
Folate_DFE_(µg)	0.02	0.09	0.04	0.053	0.19	-0.033	0.23	0.17	0.007	-0.09
Vit_B12_(µg)	-0.004	-0.002	0.008	0.14	0.43	0.071	-0.24	-0.15	-0.12	-0.067
Vit_A_IU	-0.002	0.013	-0.014	0.02	0.29	0.18	-0.26	-0.022	0.044	0.17
Vit_A_RAE	0.004	0.006	-0.004	0.036	0.26	0.1	-0.22	-0.089	-0.042	0.028
Retinol_(µg)	0.006	0.005	-0.002	0.04	0.25	0.091	-0.21	-0.1	-0.057	-0
Alpha_Carot_(µg)	-0.005	0.004	-0.008	-0.015	0.019	0.031	-0.027	0.017	0.039	0.067
Beta_Carot_(µg)	-0.013	0.007	-0.02	-0.039	0.041	0.087	-0.04	0.096	0.11	0.23
Beta_Crypt_(µg)	-0.003	0.002	-0.004	-0.012	0.004	0.007	-0.004	0.001	0.009	0.004
Lycopene_(µg)	-0.005	0.005	-0.008	-0.014	0.001	0.017	-0.004	-0.011	0.046	-0.001
Lut+Zea_ (µg)	-0.028	0.016	-0.04	-0.082	0.073	0.19	-0.022	0.36	0.2	0.68
Vit_E_(mg)	0.048	0.006	-0.006	-0.008	0.047	0.099	0.11	-0.009	-0.003	-0.018
Vit_D_µg	0.006	0.001	-0.004	0.007	0.037	0.001	-0.014	-0.018	-0.009	0.02
Vit_D_IU	0.006	0.001	-0.004	0.007	0.037	0.001	-0.014	-0.018	-0.009	0.02
FA_Sat_(g)	0.19	-0.091	-0.034	-0.047	0.044	-0.34	-0.35	0.39	0.27	-0.18
FA_Mono_(g)	0.28	-0.12	-0.032	-0.068	0.011	0.1	0.13	-0.1	-0.058	0.04
FA_Poly_(g)	0.2	-0.046	-0.042	-0.081	0.003	0.39	0.35	-0.34	-0.19	0.12
Cholestrl_(mg)	0.003	-0.038	0.01	0.084	0.13	-0.05	-0.12	-0.076	-0.082	0.034
GmWt_1	-0.081	0.013	-0.062	-0.12	-0.12	0.11	-0.12	-0.08	0.024	-0.44
Refuse_Pct	-0.075	-0.26	0.89	-0.36	0.047	0.055	-0.032	0.012	-0.006	-0.017

Although we have the diagram above, which specifically gives out which features are composed of each component and the specific content ratio of each feature in each conponent, the rose plot below can make it much more intuitive and understandable for each conponent. The top 10 important (use absolute value of coeffience) features of PC1 are placed in order as follows:



2.6 K-MEANS CLUSTERING

Clustering is one of the most common exploratory data analysis technique used to get an intuition about the structure of the data. Clustering analysis can be done on the basis of features where we try to find subgroups of samples based on features or on the basis of samples where we try to find subgroups of features based on samples. Based on the results from PCA, we could further solve our problems via K-means Clustering, in order to cluster existing food items into more reasonable categories.

In other words, we try to find homogeneous subgroups within the food items such that food items in each cluster are as similar as possible according to a similarity measure such as composition of protein, fat, etc. Before we get into the implementation part, we first take a quick look at the K-means algorithm.

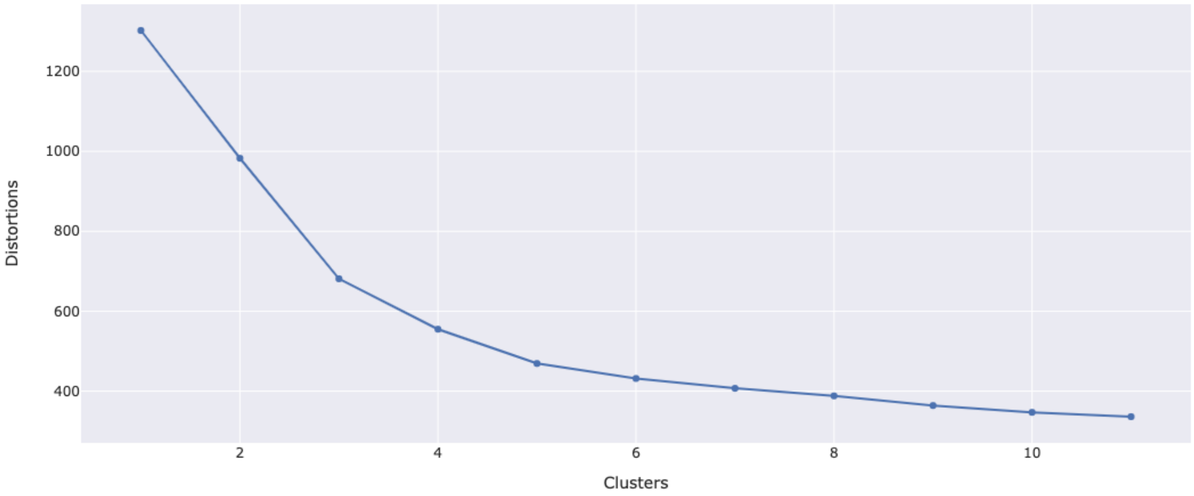
Algorithm: K-means

1. Specify the number of clusters K.
2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids, i.e assignment of data points to clusters isn't changing.
   - Assign each data point to the closest cluster (centroid) so that the sum of the squared distances between each data point and center of its assigned cluster is minimized.
   - Update the centroids of each cluster by computing the average of all the data points that belong to the cluster.

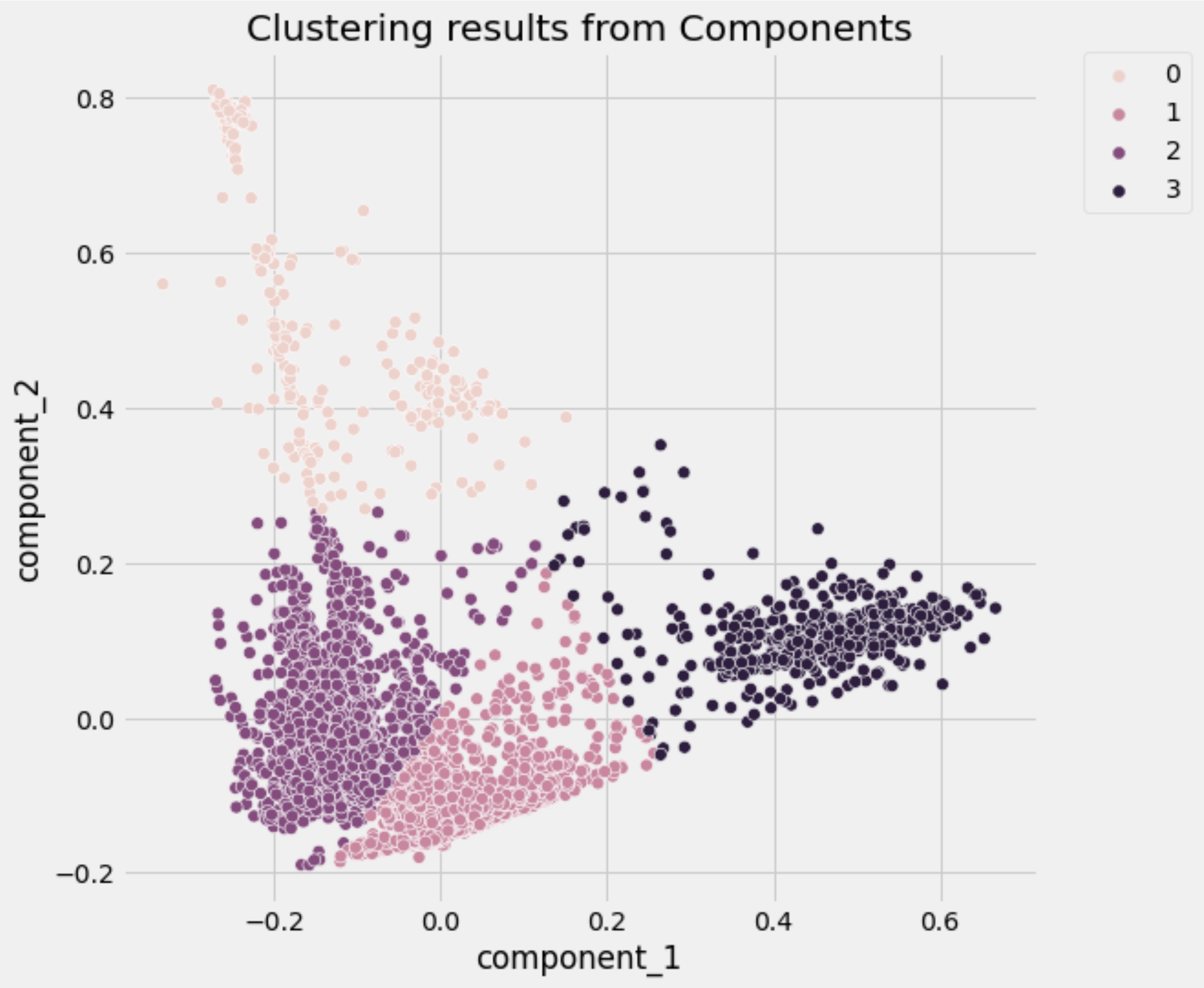
Next we considered to choose the appropriate number of clusters. We performed the elbow method, that is to try different values for K, run the algorithm and record the distortion loss.

Then we plotted the value of loss function versus number of clusters. We picked the elbow point on the curve, which in this case is K = 4.

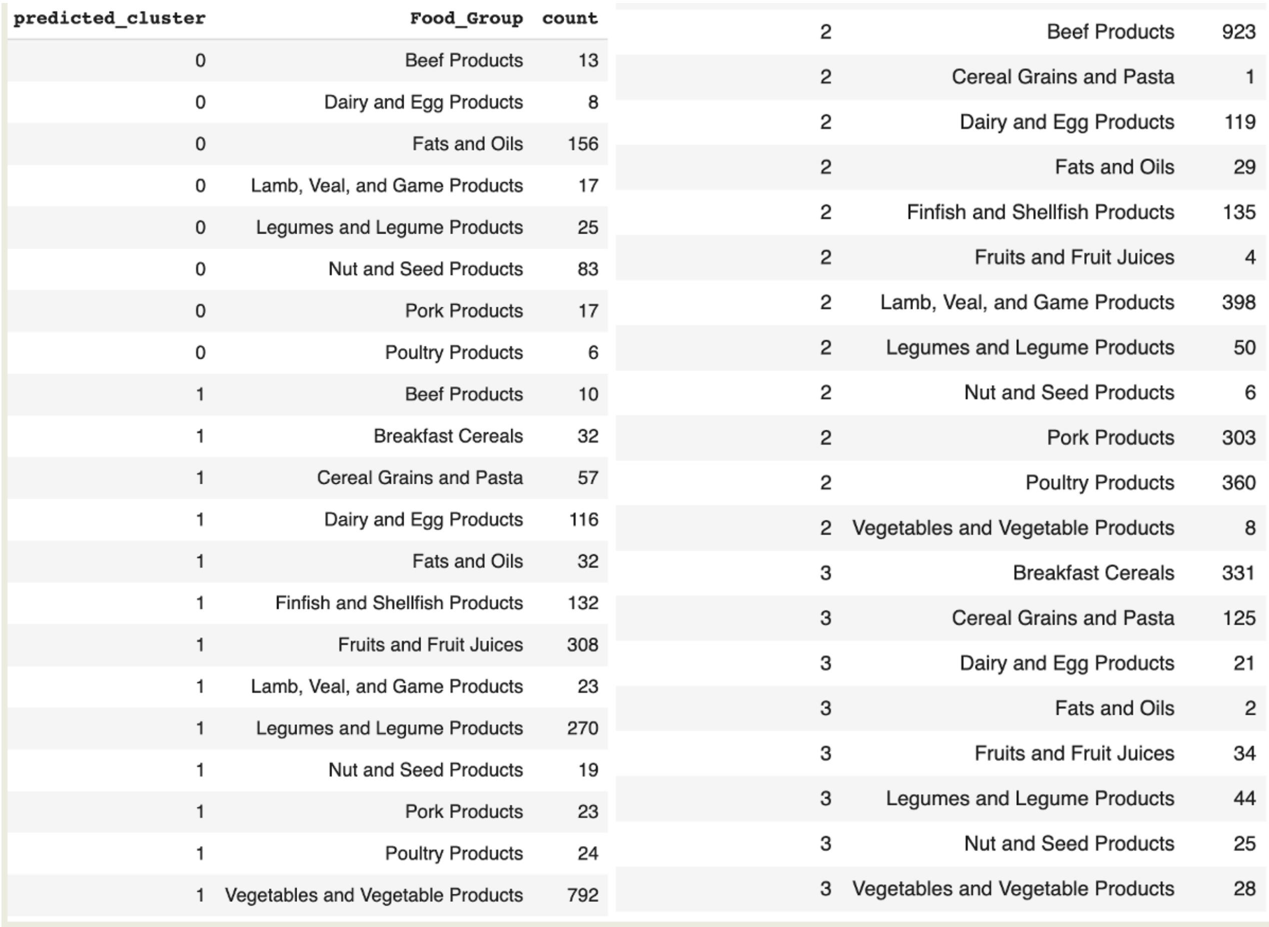
A lower K value will lead to greater bias while a higher K value will result to a higher variance when running the algorithm. Thus, we located the best hyperparameter due to bias-variance tradeoff.



The visual representation of the clusters confirms the results of the clustering evaluation metrics. The performance of K-means clustering was pretty good. The clusters only slightly overlapped, and cluster assignments were much better than random.

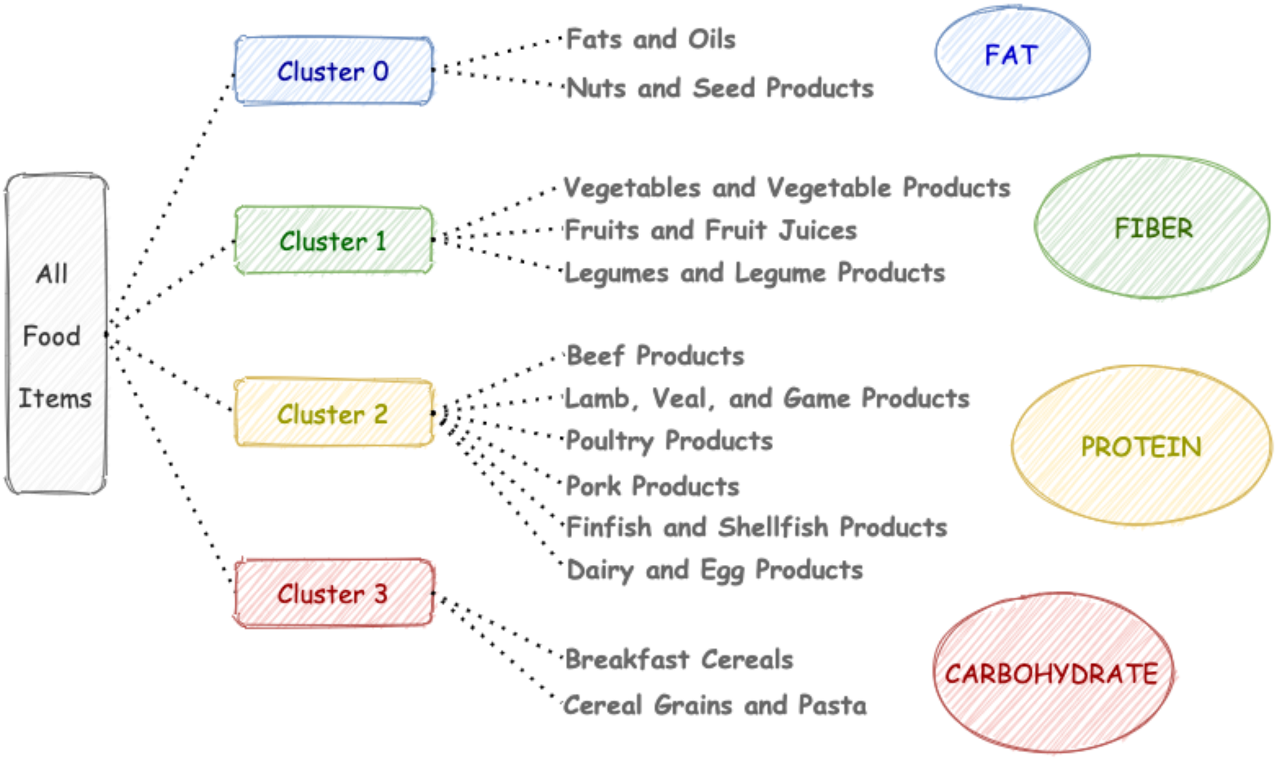


The following frame shows the clustering result: different food group content in each cluster.



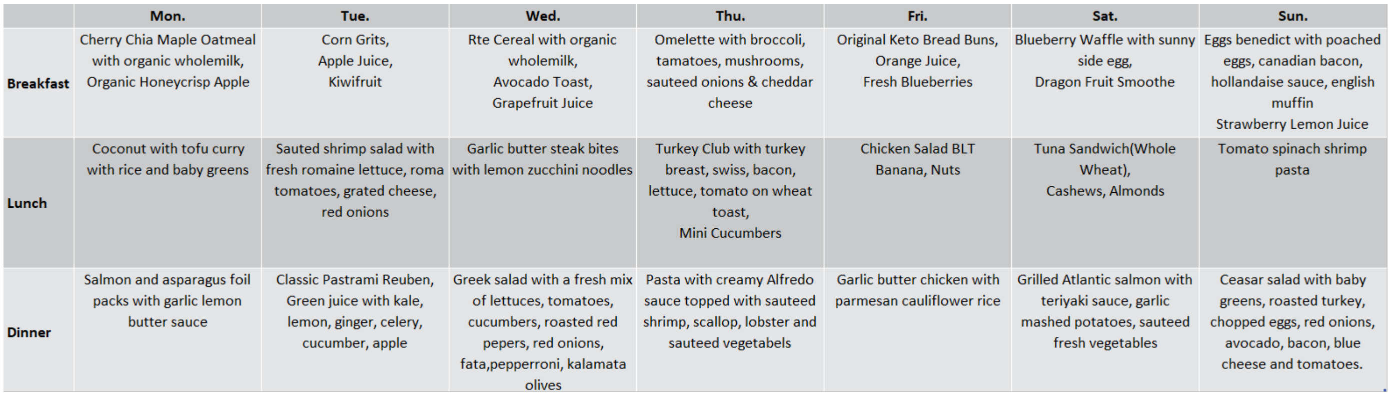
predicted_cluster	Food_Group	count
0	Beef Products	13
0	Dairy and Egg Products	8
0	Fats and Oils	156
0	Lamb, Veal, and Game Products	17
0	Legumes and Legume Products	25
0	Nut and Seed Products	83
0	Pork Products	17
0	Poultry Products	6
1	Beef Products	10
1	Breakfast Cereals	32
1	Cereal Grains and Pasta	57
1	Dairy and Egg Products	116
1	Fats and Oils	32
1	Finfish and Shellfish Products	132
1	Fruits and Fruit Juices	308
1	Lamb, Veal, and Game Products	23
1	Legumes and Legume Products	270
1	Nut and Seed Products	19
1	Pork Products	23
1	Poultry Products	24
1	Vegetables and Vegetable Products	792
2	Beef Products	923
2	Cereal Grains and Pasta	1
2	Dairy and Egg Products	119
2	Fats and Oils	29
2	Finfish and Shellfish Products	135
2	Fruits and Fruit Juices	4
2	Lamb, Veal, and Game Products	398
2	Legumes and Legume Products	50
2	Nut and Seed Products	6
2	Pork Products	303
2	Poultry Products	360
2	Vegetables and Vegetable Products	8
3	Breakfast Cereals	331
3	Cereal Grains and Pasta	125
3	Dairy and Egg Products	21
3	Fats and Oils	2
3	Fruits and Fruit Juices	34
3	Legumes and Legume Products	44
3	Nut and Seed Products	25
3	Vegetables and Vegetable Products	28

The following diagram briefly summarizes the clustering results from above. We pick the top numerous food groups in each cluster to be the representative ones. As we can see, these 4 clusters categorize these food items based on four main ingredients which are fat, fiber, protein and carbohydrate. The result is quite aligned with our common sense. Next, we will use it to build a much healthier diet recommendation.



3. WEEKLY DIET RECOMMENDATION

Based on the clustering results, we choose food items randomly from each cluster and make up this recipe:



	Mon.	Tue.	Wed.	Thu.	Fri.	Sat.	Sun.
Breakfast	Cherry Chia Maple Oatmeal with organic wholemilk, Organic Honeycrisp Apple	Corn Grits, Apple Juice, Kiwifruit	Rte Cereal with organic wholemilk, Avocado Toast, Grapefruit Juice	Omelette with broccoli, tamatoes, mushrooms, sauteed onions & cheddar cheese	Original Keto Bread Buns, Orange Juice, Fresh Blueberries	Blueberry Waffle with sunny side egg, Dragon Fruit Smoothe	Eggs benedict with poached eggs, canadian bacon, hollandaise sauce, english muffin Strawberry Lemon Juice
Lunch	Coconut with tofu curry with rice and baby greens	Sauted shrimp salad with fresh romaine lettuce, roma tomatoes, grated cheese, red onions	Garlic butter steak bites with lemon zucchini noodles	Turkey Club with turkey breast, swiss, bacon, lettuce, tomato on wheat toast, Mini Cucumbers	Chicken Salad BLT Banana, Nuts	Tuna Sandwich(Whole Wheat), Cashews, Almonds	Tomato spinach shrimp pasta
Dinner	Salmon and asparagus foil packs with garlic lemon butter sauce	Classic Pastrami Reuben, Green juice with kale, lemon, ginger, celery, cucumber, apple	Greek salad with a fresh mix of lettuces, tomatoes, cucumbers, roasted red pepers, red onions, fata,pepperroni, kalamata olives	Pasta with creamy Alfredo sauce topped with sauteed shrimp, scallop, lobster and sauteed vegetabels	Garlic butter chicken with parmesan cauliflower rice	Grilled Atlantic salmon with teriyaki sauce, garlic mashed potatoes, sauteed fresh vegetables	Ceasar salad with baby greens, roasted turkey, chopped eggs, red onions, avocado, bacon, blue cheese and tomatoes.

4. SUPPLEMENT

Attached are the complete codes for your reference.

AMCS_602_Final_Project.html