What Do Nutrients Mean and How Can We Use Nutritional Value to Determine What to Eat for Certain Diets

By Tyler Mogensen

I. ○ Introduction/Problem definition

Food is one of the most important aspects in anybody's life. Food is what provides our bodies with energy to maintain essential body function, repair ourselves, and support us in any other efforts when it comes to living our lives. What we put in our bodies is typically due to a variety of causes such as personal taste, convenience, cost, or health concerns. With so many different options, what's the difference? There are many diets out there that are determined by individual goals and aspirations. For example, a bodybuilder who is bulking will want to retain muscle mass while eating in a surplus, thus they will be eating high calorie and high protein. Somebody who is looking to lose weight while staying heart conscious, will be looking for foods that are typically lower in sodium and calories. There are those that are diabetic that may want a low carb diet. There are also those that are lactose intolerant and looking for dairy alternatives with similar nutritional value. There are many different combinations of diets and foods that can be consumed to fit your needs, the problem for the average person is determining what foods they should be eating and staying consistent with their diet. Having access to this knowledge is essential.

I am looking to determine what is essential when it comes to what it means for a food to fit your diet. This can be in terms of overall calories, macronutrient preference, or micronutrient needs. I think that I will also need to classify foods into certain categories that will be helpful for those that are looking for a variety of similar foods that will fit their needs.

III. ○ Survey on what has been done and what is different in your project.

Having personally tried a variety of diets, it is not always easy to find a diet that is both enjoyable and sustainable. Diets are rarely personalized and are just a baseline for somebody who has never been on a true diet and just eats what is convenient. For example, the current USA recommended diet suggests nearly the same diet for everyone based on their findings and what a majority of Americans are deficient in. Many google searches give blanket diets as well and several dives and some basic understanding are needed to construct something that will be helpful. There are too many places to look and too much that isn't accounted for other than calories and macronutrients. Many people look to nutritionists who are not actually as qualified as dieticians and lack credentials such as Registered Dietitian approval, coursework, and certifications. While nutritionist information is typically better than nothing, the best way to find a diet is talking with a dietitian, but you will be paying for the price of quality of information.

IV. ○ Proposed method

Seeing as calories are usually the most important thing to consider in many of these diets, we will start by seeing what exactly tends to make for high and low calorie foods. We can then use this information to predict the calories for foods that are not already in our dataset. From there, I think that classifying the foods from our data set will be a helpful start for those that are looking for foods that will fit their needs. This will be mostly through clustering and categorizing foods based on the nutritional value. For example, I will label foods High, Medium, and Low based on their protein, heart healthiness, glycemic index, and calorie density. If someone is looking for a specific diet, then they can group these foods based on those categories. This will give them options that give them variety that will likely help them better stick to their diet than eating the same thing all the time and getting bored. Foods that would be later added to the data set could also be fit into these categories and clusters. The dataset that I am begging with was found on kaggle. I had to start with cleaning the data which included many regular expression functions to get rid of the labels for each variable. (https://www.kaggle.com/trolukovich/nutritional-values-for-common-foods-and-products?select=nutrition.csv)

I categorized a food as high in protein if it contained more than 20% of its weight in proteins, moderate from 10% - 20% and low if it contained less than 10% and None if there was no protein on its nutritional label. Various sources suggested this, unfortunately, there is not a standard for nutritional labels by the FDA for label claims.

For glycemic index, this is a measure of how quickly a food can cause blood sugar levels to increase. This is due to the ratio of fiber and other carbohydrates in food. Foods that are high in glycemic index will tend to have high carbohydrate content and low fiber. Fortunately, since all of the foods are measured at 100g, we can standardize this easily.

As for sodium, typically for a food weighing in at 100g, it can be considered heart healthy if it has less than 100 mg of sodium, but I wanted to include cholesterol in there as well since it plays a factor in heart health as well. There are two kinds of cholesterol and if I were able to find the amount for each, I would be able to better determine heart healthiness.

Finally, what many people may be looking for first when it comes to diets, is calories. Since the main factor for many diets is foods, I wanted to include a category for these. While you can only have so many calories for many diets, this gives a starting point to what foods to eat before others depending on goals. If they both are satiating, then the low calorie foods would be optimal for losing weight and higher calorie foods would be optimal for gaining weight.

Many of these were found from FDA Codes and Regulations: Specific Requirements for Nutrient Content Claims

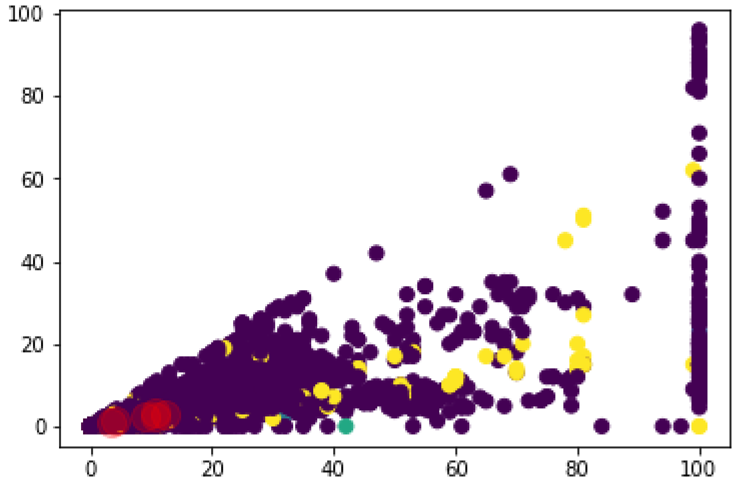
(https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?CFRPart=101&showFR=1&subpartNode=21:2.0.1.1.2.4)

V. ■ Intuition - why should it be better than state of the art?

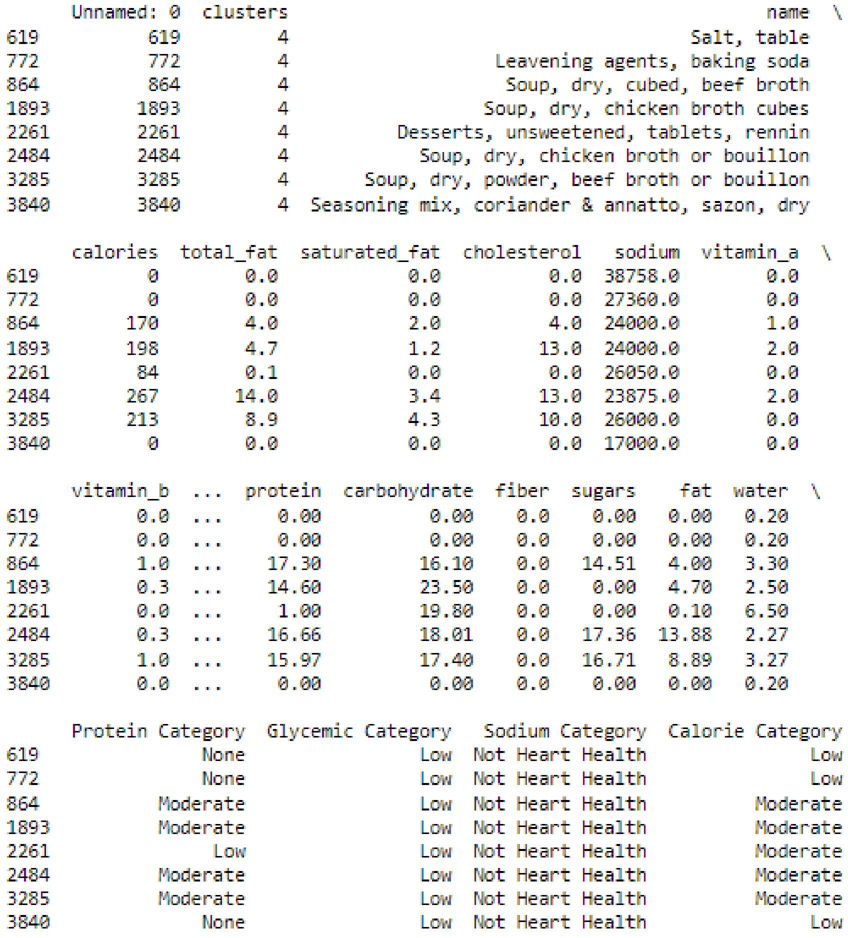
If somebody were looking to get their diet in order, the best way to have a consultation with a dietitian. However, these appointments tend to be expensive, which excludes many people from figuring out what they may need to achieve their personal goal and fuel their day to day lives. A simple google search will tell you what foods you should be eating if you are looking for say high protein or low glycemic foods, but they are typically giving you the top 5 or so foods that have the most nutritional value. These foods can be expensive as well, and another google search would be needed to figure out those alternatives. This needs to be done several times over to figure out some semblance of a diet. Using my model and data, we can quickly figure out what foods should be eaten for a certain diet and have a complete list of variety quickly and at a low cost. I think that if I were to take these findings and organize them into an application, users could have found a variety of foods that fit their needs and learn what each could do for them instead of just looking at calories or just eating what they think is healthy.

VI. ■ Description of your approaches: algorithms, user interfaces, etc.

I think that my work on this data set could be a good starting point for an application similar to MyFitnessPal, but includes a database with categories that users could input their own foods to further improve the unsupervised clustering algorithm. This application could compare foods, so that users can make better food choices and include quick descriptions of micro and macro nutrients and what they do within the human body. This would be much more comprehensive than many other food tracking systems and provide information to the typical user who would be mostly interested in just calories. I would start with the user inputting their personal goals and physical statistics and then recommend foods from there.

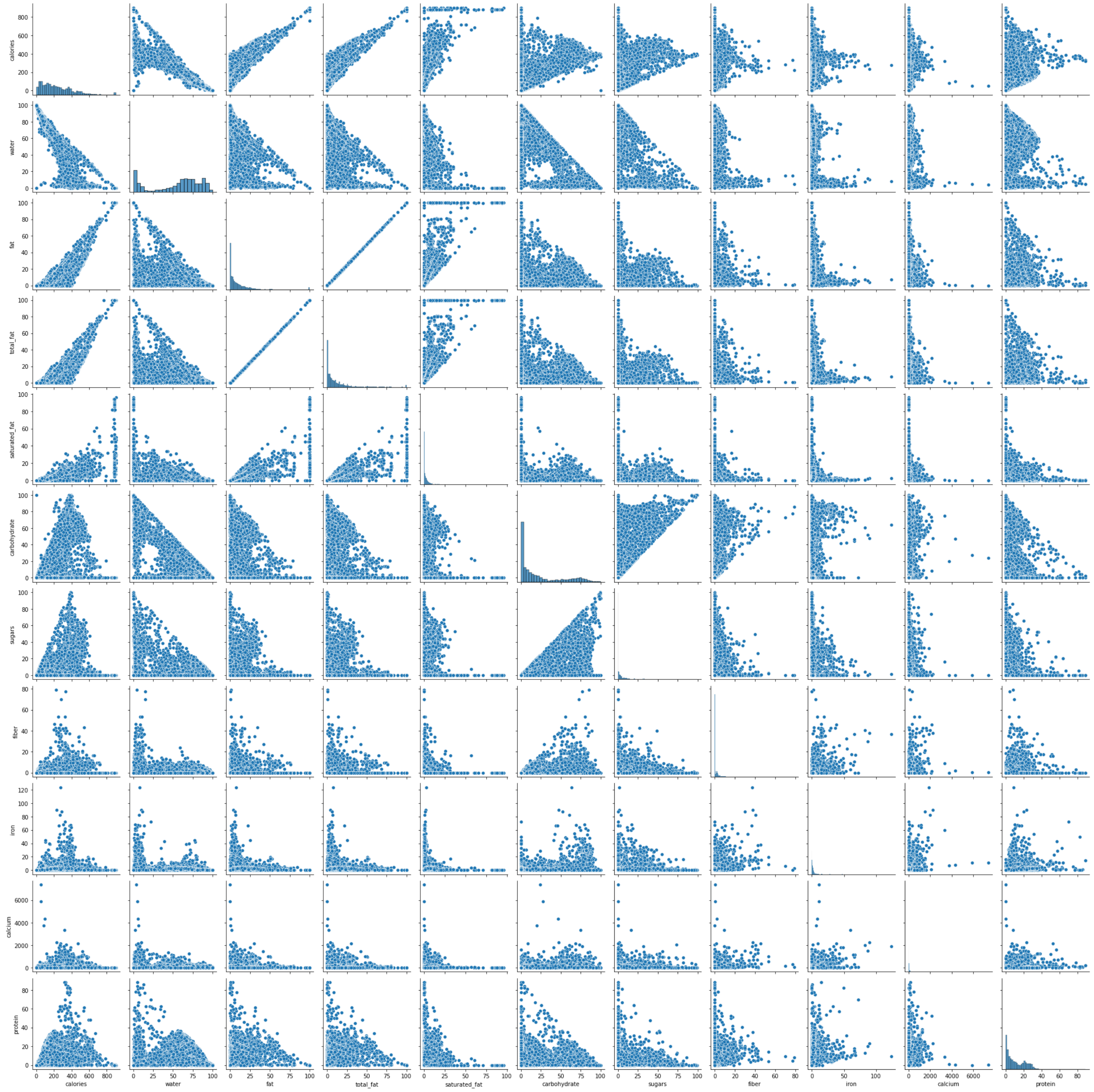


Using the KMeans Clustering Algorithm, I decided that I would put the nearly 8800 foods into 6 categories. This being for each section on the food pyramid that many people are familiar with when learning about food. This clustering method worked fairly well by putting similar foods into each category, but there was a disparity between the size of each cluster.

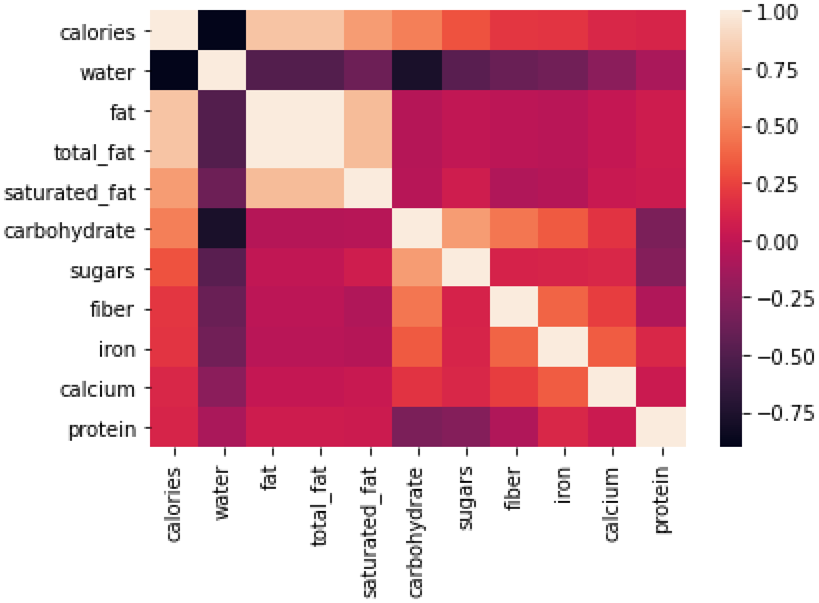


```
      Unnamed: 0  clusters                                       name  \
619          619         4                                Salt, table   
772          772         4                   Leavening agents, baking soda   
864          864         4                    Soup, dry, cubed, beef broth   
1893        1893         4                  Soup, dry, chicken broth cubes   
2261        2261         4          Desserts, unsweetened, tablets, rennin   
2484        2484         4              Soup, dry, chicken broth or bouillon   
3285        3285         4          Soup, dry, powder, beef broth or bouillon   
3840        3840         4  Seasoning mix, coriander & annatto, sazon, dry   

      calories  total_fat  saturated_fat  cholesterol   sodium  vitamin_a  \
619          0        0.0            0.0          0.0  38758.0        0.0   
772          0        0.0            0.0          0.0  27360.0        0.0   
864        170        4.0            2.0          4.0  24000.0        1.0   
1893       198        4.7            1.2         13.0  24000.0        2.0   
2261        84        0.1            0.0          0.0  26050.0        0.0   
2484       267       14.0            3.4         13.0  23875.0        2.0   
3285       213        8.9            4.3         10.0  26000.0        0.0   
3840         0        0.0            0.0          0.0  17000.0        0.0   

      vitamin_b  ...  protein  carbohydrate  fiber  sugars    fat  water  \
619         0.0  ...     0.00          0.00    0.0    0.00   0.00   0.20   
772         0.0  ...     0.00          0.00    0.0    0.00   0.00   0.20   
864         1.0  ...    17.30         16.10    0.0   14.51   4.00   3.30   
1893        0.3  ...    14.60         23.50    0.0    0.00   4.70   2.50   
2261        0.0  ...     1.00         19.80    0.0    0.00   0.10   6.50   
2484        0.3  ...    16.66         18.01    0.0   17.36  13.88   2.27   
3285        1.0  ...    15.97         17.40    0.0   16.71   8.89   3.27   
3840        0.0  ...     0.00          0.00    0.0    0.00   0.00   0.20   

      Protein Category  Glycemic Category   Sodium Category  Calorie Category
619               None                Low  Not Heart Health               Low
772               None                Low  Not Heart Health               Low
864           Moderate                Low  Not Heart Health          Moderate
1893          Moderate                Low  Not Heart Health          Moderate
2261               Low                Low  Not Heart Health          Moderate
2484          Moderate                Low  Not Heart Health          Moderate
3285          Moderate                Low  Not Heart Health          Moderate
3840              None                Low  Not Heart Health               Low
```

This is the fifth cluster from the clustering algorithm. As we can see many of these foods are related in many ways. First, I can see that they are all similar as they are mostly soups and broths. Next, the sodium content for these is very high, but the fats are low. They also happen to all be in the Not Heart Health category and Low Glycemic Index Category. However, each food has its own little differences such as calorie category which could be used to substitute foods from one to another in a diet.



Looking at the pairplot, we can see many relations between calories and micro and macro nutrients. Something that may be most important to see from this pairplot may be the relations between certain micronutrients and macronutrients. For instance, carbohydrates and iron are typically correlated since many grains are fortified with iron.



As for this heatmap, I think that it is more helpful to better visualize the relationships between variables, especially in terms of positive or negative relationships since it is harder to determine from some of the pairplots.

VII. ○ Evaluation

After analyzing the correlation between amounts of nutrients in food and the amount of calories in said food, I can narrow down the amount of variables needed to estimate the number of calories. These variables are water, fat, carbs, and calcium. These were found by finding the highest correlated variables, working through the backwards elimination model, and using the selectKbest algorithm in sklearn. The r-squared and adjusted r-squared values were astoundingly high at .989. There could be some overfitting, but after further research, it seems that the only variables that should be correlated with calories are protein, carbohydrates, and fats. These variables would be 4,4,and 9 calories per gram respectively, so this would make sense as to why it should be so accurate, but not with the variables that were used. This could be because there is typically some calcium and other phosphates in protein taking its place in the model. The same can be said for the testing and training data as well.

As for the clustering, it seems that similarly prepared foods are in similar clusters. Many foods that I would consider good fats are in the first category, micronutrient dense foods in the second, whole foods in the third, raw vegetables in the fourth, typically unhealthy foods in the fifth, and high protein foods in the sixth. I think that these would be a good starting point for grouping foods. There are some nuances to certain foods and I would have them moved, but in terms of diets, you could pick some foods from each category and eat certain amounts from each to fit your goals. I think that the first group could probably have the clustering algorithm run on it again since it is such a large cluster. This would further group more similar foods together and have more specific substitutes for certain foods.

Overall, I think that we can provide an accurate estimate for the amount of calories in a food when we are given the nutrients in said food using our model. We are also able to put these foods into clusters using our clustering algorithm to put similar foods together and conduct further analysis on those clusters. Using what we found, we can go further and create a recommendation system that gives users options for their diets that fit their goals and needs. Further, we could make more categories for high, moderate, and low for each nutrient. This way, if somebody were looking for the benefits of certain nutrients, they could filter the data and find foods in that category with other filters.