Final Report

Daniel Kojis

December 9, 2018

Abstract

This report sets out to explore the nutrition of common foods we eat. Not all foods were included, but 41 foods that are commonplace in an American diet were analyzed on a calorie for calorie basis. The data was analyzed collectively, as well as looking at macronutrient and micronutrient compositions individually at times. Among the many methods applied were starplots, multidimensional scaling, principal component analysis, and clustering methods. There are many takeaways from the analysis, some of which are included in the report and some not, but a few were more interesting than others. Contrary to my hypothesis, there appears to be no relationship between the macronutrient and micronutrient composition of a food. There was a slight positive relationships among micronutrients, however, as foods high in one micronutrient were more likely to be high in another. Fruits and vegetables were the main sources of vitamins and minerals; dairy products serve as a good source of calcium, but the other food groups offer little nutritional value in this department. Spinach was a truly unique food providing exceptionally high vitamin and mineral values across the board. The food groups that we biologically categorize foods into also do an impressive job at grouping foods that are nutritionally similar. The exceptions are sometimes intuitive, and sometimes not. Particular benefits and disadvantages of certain foods and food groups are highlighted throughout the report.

Introduction

Nutrition is a fundamental aspect of our lives. The foods we eat can have a substantial impact on our health and well-being. While people value their health and diet differently, at the end of the day, everybody eats and is impacted by the foods they choose. The effect of diet has both immediate and long-lasting effects.

Because of nutrition's universal importance, it's probably a good idea that we understand it. This is the motivation for this report: to evaluate the nutrition of common foods we eat.

About the Data

To understand the nutrients we put into our bodies, I set out to collect data from www.verywellfit.com. I selected 41 different food items in total (see next section), and they were chosen because of their commonplace in my personal diet. While you might find some of these foods obscure, I would assume that most of these foods are commonplace in the American diet. Each food item had a corresponding nutrition label on the website, which is where the data was web scraped from. Originally, the data contained 16 variables for one serving of the food item: Food Item, Food Group, Weight (g), Calories, Total Fat (g), Saturated Fat (g), Cholesterol (mg), Sodium (mg), Total Carbs (g), Sugar (g), Fiber (g), Protein (g), Vitamin A (daily value %), Vitamin C (DV%), Iron (DV%), and Calcium (DV%). While there are certainly other variables that could be considered for nutritional analysis, these variables are usually considered among the most important, and that is why they are readily available on most nutritional labels.

Because the data was web scraped from a consistently formatted website, there were little issues with missing or inaccurate data. The only exception to this was that four foods (sweet corn, green beans, onions, and ham) did not contain information on saturated fat. Since I wanted to preserve these data points and column variables, the missing data was imputed. Mean imputation was ruled out because saturated fat content varies quite a bit from food to food, so these values would likely be inaccurate. Rather, because saturated fat content in highly predictable from total fat content, I used deterministic regression imputation to replace the missing values. Total fat and cholesterol were the two predictors used to make linear regression predictions for the saturated fat value. While this has the potential to reduce the variance of saturated fat and possibly inflate the correlation between saturated fat and the predictor variables, I am not concerned with these issues. While other nutrition labels have slight differences of nutrient values of the same food, they are still very close. The imputed values were compared and nearly identical to similar nutrition labels of the same food item.

Because the serving sizes vary from food to food, it's not reasonable to compare the data in its current state. There are two ways I could choose to scale the dataset to make it more comparable: scaling to the same calorie value or to the same weight value. Each choice has its pros and cons and will lead to a slightly different interpretation of the data. Scaling to the weight value might be more intuitive if you are concerned with the nutrients obtained from eating a serving of the food, as weight values tend to predict how filling a food is. Scaling to the calorie value might be more beneficial to someone concerned with meeting their daily nutrient needs on a limited calorie intake (such as someone watching their weight). I chose the latter method out of personal interest, and all nutrient values were scaled to a 300 calorie serving of the food. It's important to remember the drawback of the calorie for calorie analysis, and note that lower calorie foods will have larger servings in terms of weight. I prefer this method; I think it provides a better nutritional evaluation if we aren't concerned with factoring in hunger, and I find it slightly easier to conceptualize than weights of food.

Food item reference

The following 41 foods items will be referenced in the report, most of which are abbreviated: corn, boneless chicken breast (chkn), ground beef 90% lean (beef), baked potato (ptto), 1% milkfat cottage cheese (cttg chs), broccoli (broc), plain whole milk yogurt (ygrt), green beans (gbean), whole wheat bread (brd), 2% milk (mlk), apple, banana (bna), grapes (grp), cherries (chrrs), orange (orng), onion, red bell pepper (pppr), salmon (slm), swiss cheese (cheese), carrots (crrt), whole wheat bagel (bgl), pear (prs), turkey breast (trky), steak (stk), ham, peanuts (pnut), almonds (almnd), brown rice (rice), oats, granola (grnla), celery (clry), snap peas (peas), peaches (pchs), watermelon (wmel), cauliflower (clflwr), spinach (spch), cucumber (ccmbr), black beans (bbean), bacon (bcn), egg, and tomato (tmto).

Goals

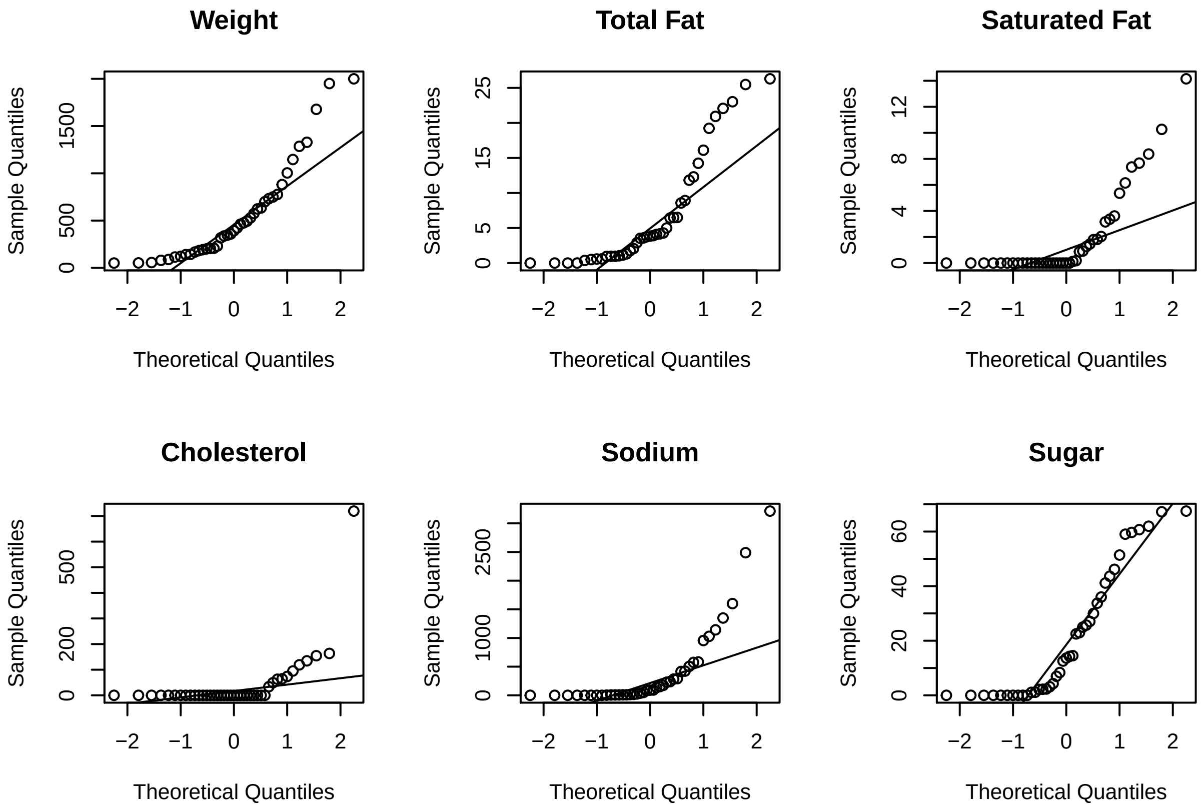
Put simply, the goal of this analysis is to explore the data and discover new things about the foods we eat and their nutrition. There are some particular questions we might want answers to. Which nutrients are related? Are macronutrient and micronutrient composition related? Which foods are rich in vitamins and minerals? Perhaps there are potential "superfoods" that stand out from the rest that are hidden in our dataset! What foods should I be eating to meet my daily value requirements? Which foods are most similar, and which foods are most different? The food in our dataset is categorized into a food group: vegetables, fruits, grains, dairy, meat, and legumes/nuts. These classification were formed on a biological basis, but how does their nutritional makeup compare? Do some foods seem like they belong in a different group?

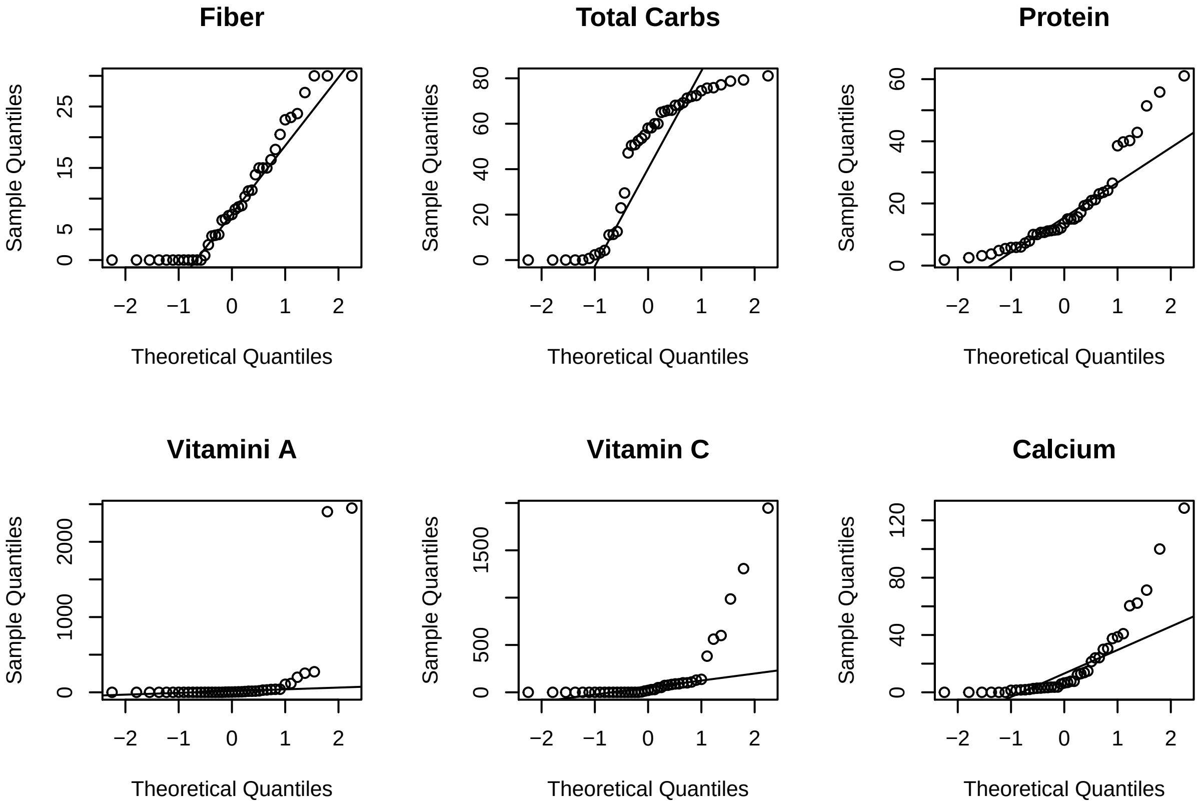
These are just some of the questions I hope to answer with this analysis. Much of the analysis will not deal with actual units of the nutrients, as many of them have little meaning to someone unfamiliar with nutrition. Rather, we'll be discussing the nutrients in mostly relative terms.

The results below are what appeared interesting and worthy of discussion. Other analysis was conducted, but these are the highlights which you might find useful.

Main Results

To begin, let's examine our variables individually. To assess whether or not any of the quantitative variables follow a normal distribution, a corresponding qq plot was plotted for each variable.


Weight
Total Fat
Saturated Fat
Cholesterol
Sodium
Sugar
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2
0
500
1500
5
15
25
4
8
12
200
1000
2500
20
40
60


Fiber
Total Carbs
Protein
Sample Quantiles
0
5
15
25
0
20
40
60
80
0
20
40
60
Theoretical Quantiles
−2
−1
0
1
2
Vitamini A
Vitamin C
Calcium
0
1000
2000
0
500
1500
0
40
80
120

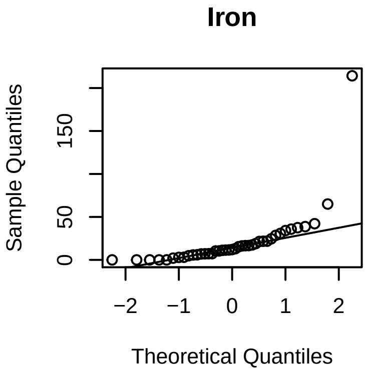
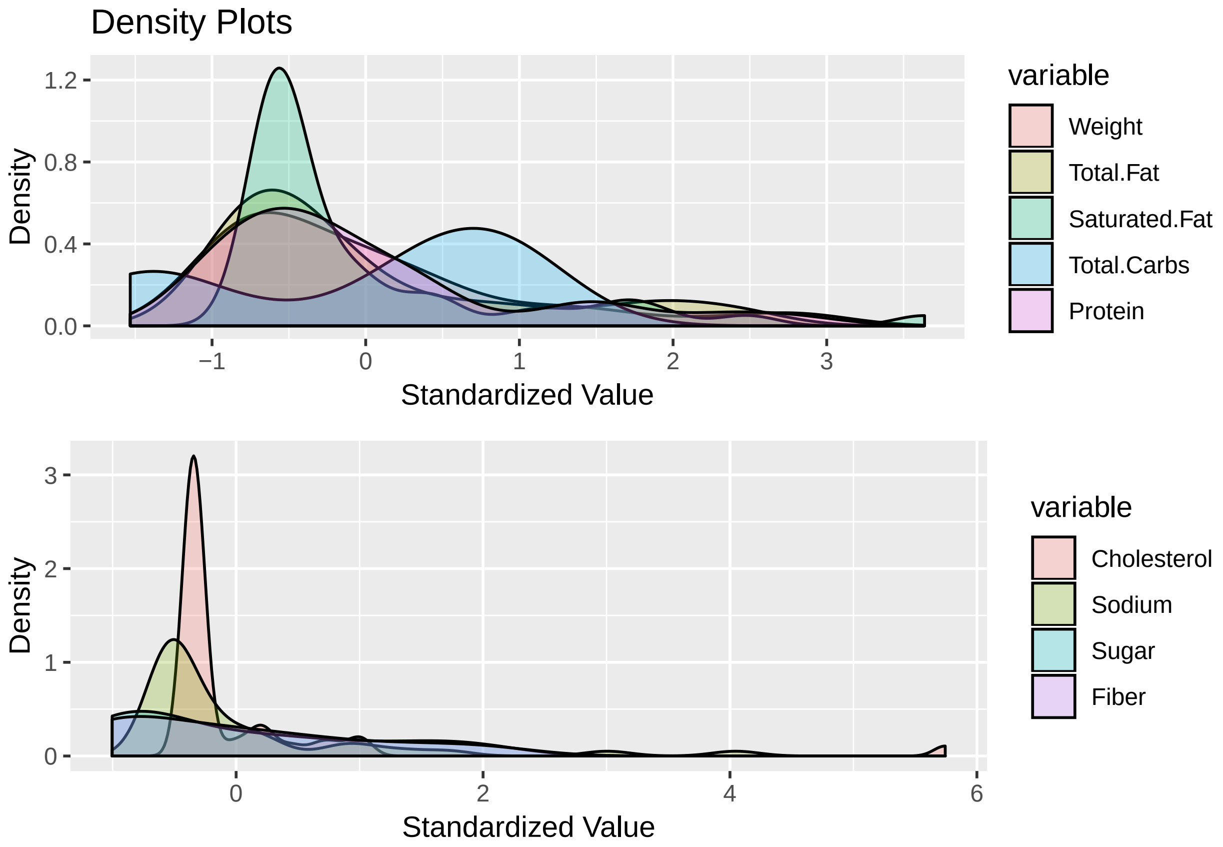


Figure 1: Normal qq plots for all quantitative variables.

The plots clearly indicated that none of the variables follow a normal distribution. This is further validated by the Shapiro-Wilk Test conducted on all 13 variables, resulting in all p-values below the 0.001 threshold and rejecting the null hypothesis that they come from a normal distribution. The qq plots also indicate a heavy tail in many of the variables.

To inspect this further, let's plot the density distributions of these variables. The variables were scaled by subtracting the column and mean and dividing by the column standard deviation for each value (i.e. standardization) so they could be more easily compared (with the exception of vitamin A). The density distribution of each variable is represented on one of the following four graphs.


Density Plots
Density
1.2
0.8
0.4
0.0
−1
0
1
2
3
Standardized Value
variable
Weight
Total.Fat
Saturated.Fat
Total.Carbs
Protein
Density
3
2
1
0
0
2
4
6
Standardized Value
variable
Cholesterol
Sodium
Sugar
Fiber

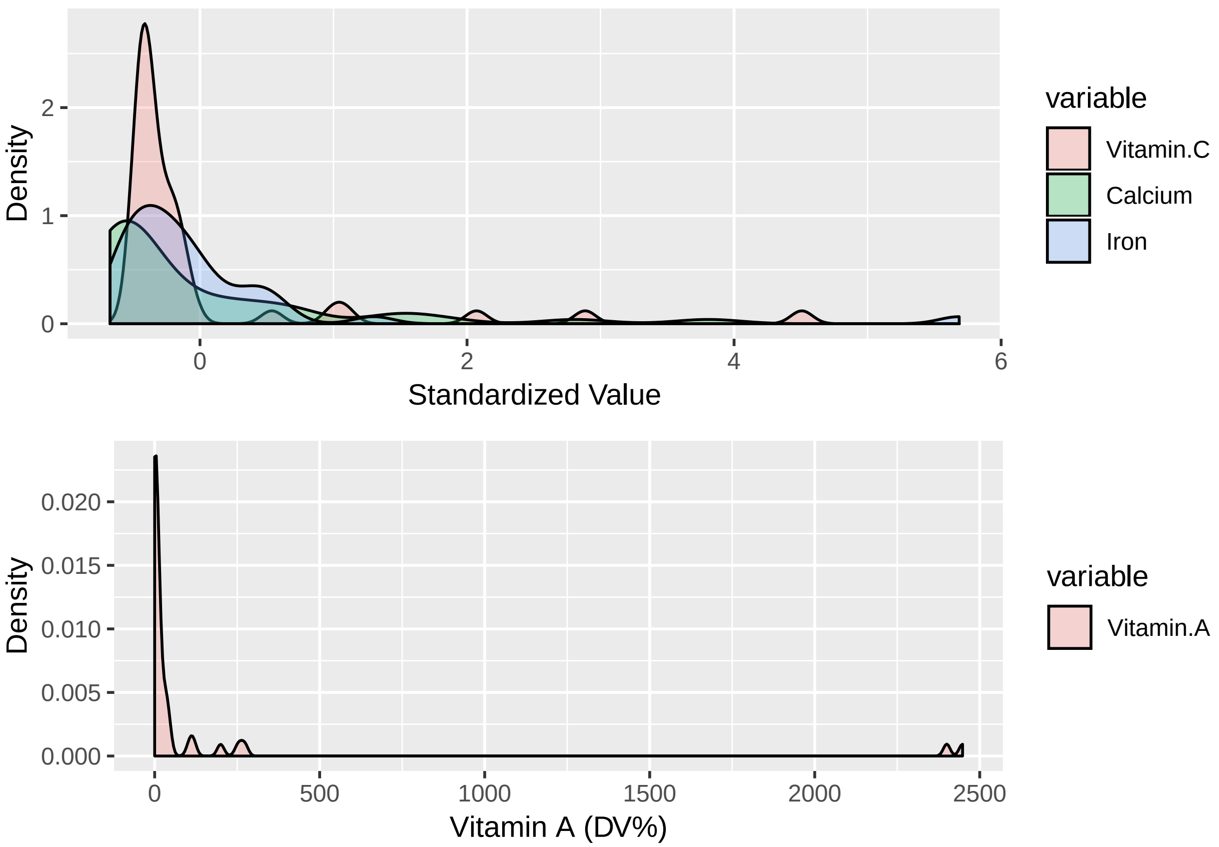


Figure 2: all variables are standardized with the exception of Vitamin A

The initial takeaway from these density plots is that most the variables have a right skew to them. Total Carbs might be the one exception, which seem to follow a form of bimodal distribution. These heavy tailed distributions are likely partly due to the fact that the variables have lots of zero values.

Taking a closer look at the third and fourth graph in Figure 2, this is especially true for the vitamins and minerals. Most values seem to fall close together with little to no nutritional value, with a few obvious exceptions at much higher values. This is especially clear in for Vitamin A, which has two very extreme outliers at nearly 2500% daily value, and a few other food items standing apart from the majority which have low values.

These distributions indicate that outliers are present in the data, so the analysis must be conscious of this in the process.

Relationship Between Variables

Next, let's examine the relationship among the variables through a correlation matrix. Because the sample size is relatively small, I calculated correlation confidence intervals and p-values to help determine whether or not a relationship was statistically significant. To help account for the non-normality of the data, I used bootstrap sample with 1000 replications for these calculations. All variables discussed as having a meaningful relationship had correlation values that were significant at the 1% level. The following plot depicts the information from the correlation matrix.

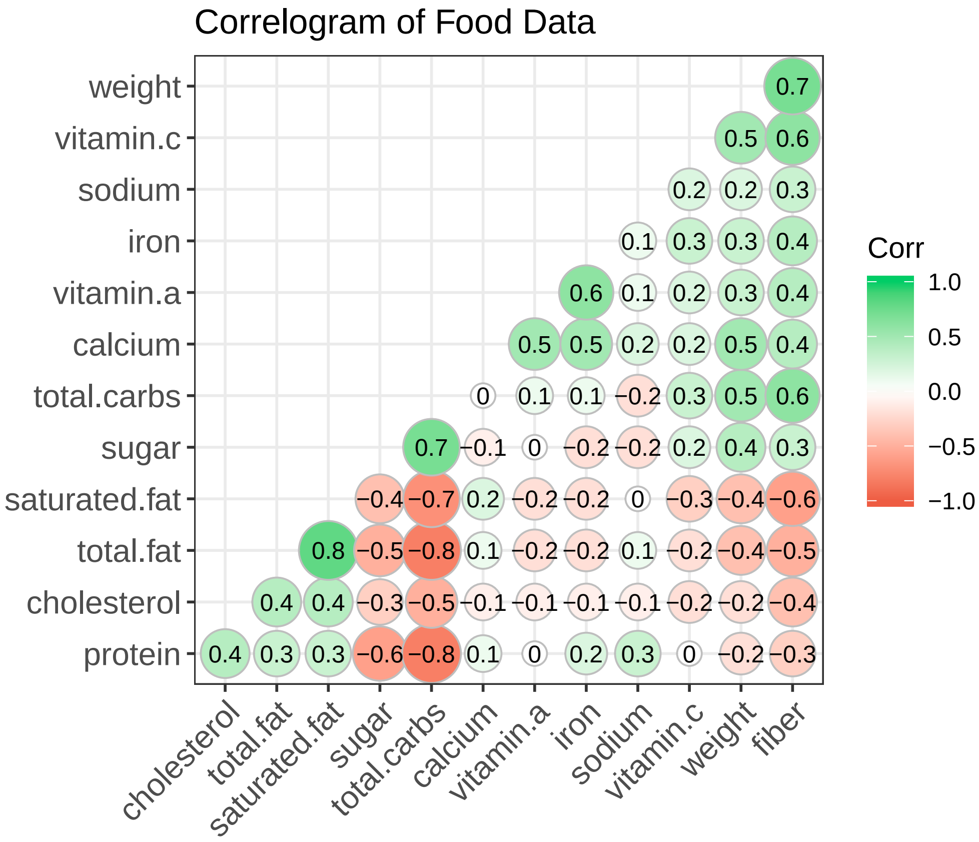


Figure 3: values rounded to first decimal place.

There a few observations I want to point out.

The highest correlation was 0.79 between total fat and saturated fat. This shouldn't surprise us, as total fat is composed of saturated and unsaturated fat. However, this point might be used to counter the argument of "not all fat is bad" that you sometimes hear people say. While this point has merit (unsaturated fats are indeed good for you), it seems like total fat is a good indicator of the level of unhealthy fat. Generally speaking, fat is bad. "Not all fat is bad" doesn't have much application in our dataset, and likely only applies to unique foods that were excluded from this analysis.

Sugar and fiber, being types of carbs, have decently strong correlation with total carbs ($r = 0.70$ and $r = 0.58$, respectively). Protein, being the third macronutrient after fat and carbs, has strong negative correlation with total carbs ($r = -0.76$). These observations made so far shouldn't be surprising and won't be the focus of the analysis despite some of the largest correlation values. However, it's important to note this relationship with those unfamiliar with macronutrient composition of foods. One could also takeaway to try and eat higher protein foods if they are trying to limit their sugar intake (correlation of sugar and protein is -0.57).

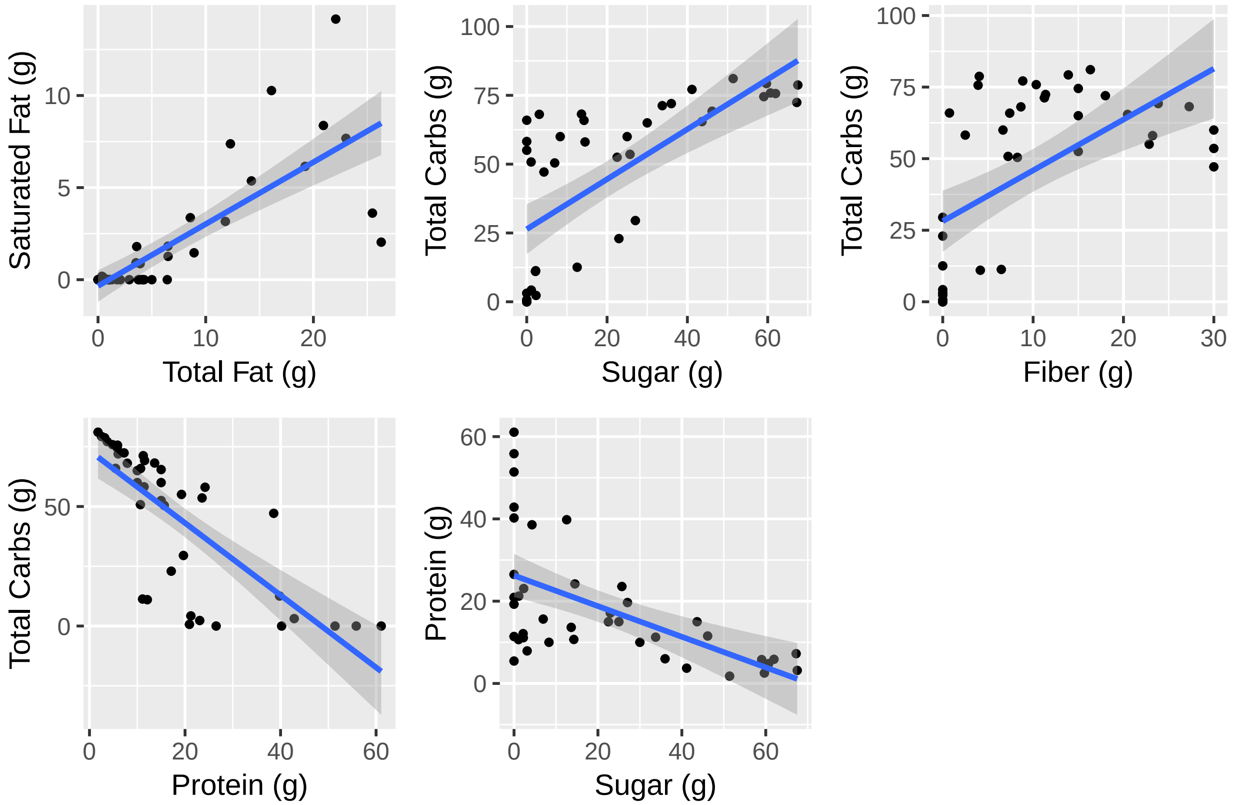


Figure 4: scatterplots of previously discussed variables.

Let's look at our vitamin & minerals (i.e. vitamin A, vitamin C , iron, and calcium). From the correlation matrix, there appears to be no relationship between vitamins & minerals and the macronutrient composition (i.e. total carbs, total fat, protein) of the food. The 12 corresponding variables correlation values fell between -0.3 and 0.3, and none of these values were significant at the 5% level. A canonical correlation analysis between the three macronutrient variables and four vitamin & mineral variables confirmed this observation of no relationship. This is an interesting observation, and rules out further analysis for the hypotheses that macronutrient composition is related to vitamin & mineral composition.

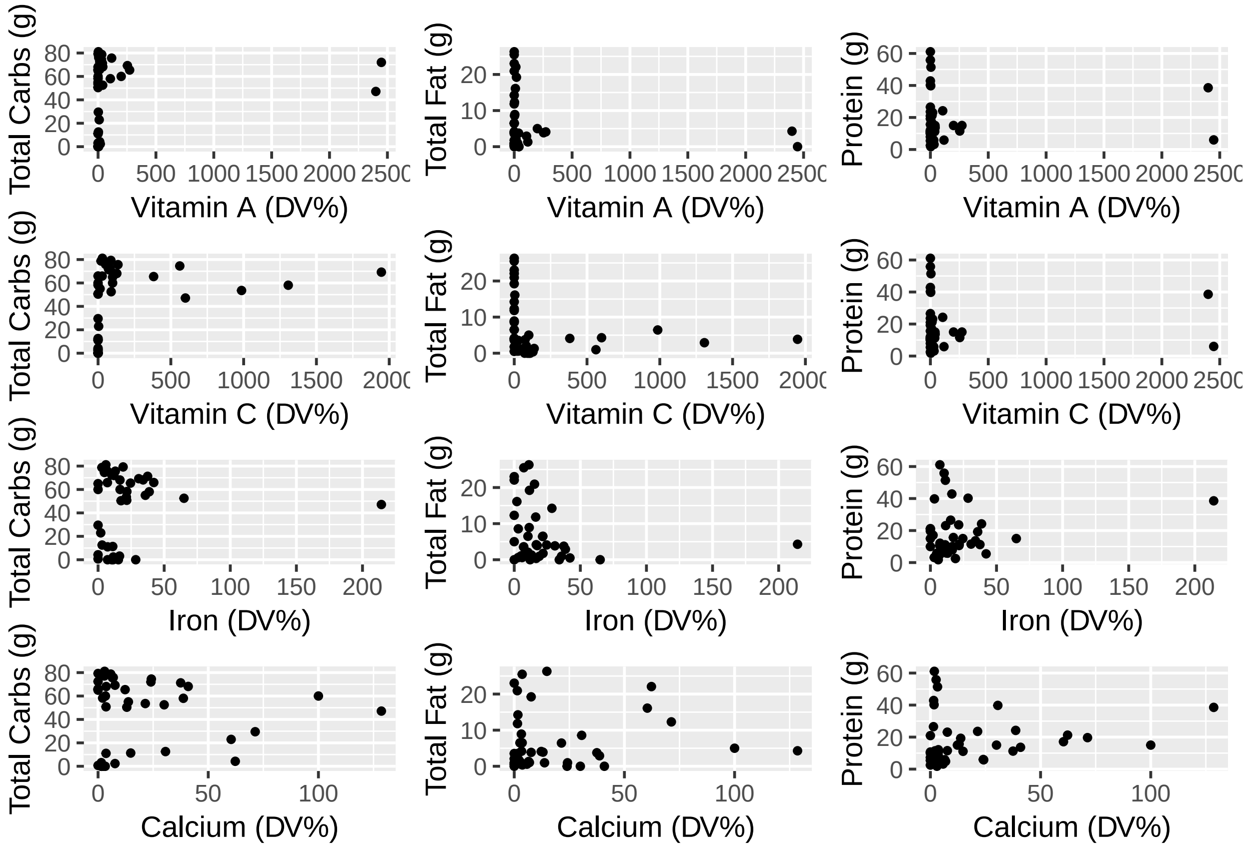


Figure 6: scatterplots of previously discussed variables.

Note that while it may appear a lot of these values take the value 0, there are a fair amount of observations with just relatively small values under 10%.

Another observation from the correlation matrix is that sodium stands alone as the sole variable with no meaningful relationships with any other variables; no correlation values were significant at the 5% level.

With the exception of Vitamin C, there seems to be a decently strong association between the other 3 vitamins & minerals. Upon examining the bivariate boxplots of these variables, however, there is an interesting observation. These plots typically show little relationship with the exception of a few key outliers. After removing these outliers, the correlation coefficients are considerably lower. A bivariate boxplot of vitamin A versus iron indicates spinach and carrots as two potential outliers. After removing them, the correlation coefficient decreased from 0.62 to 0.18.

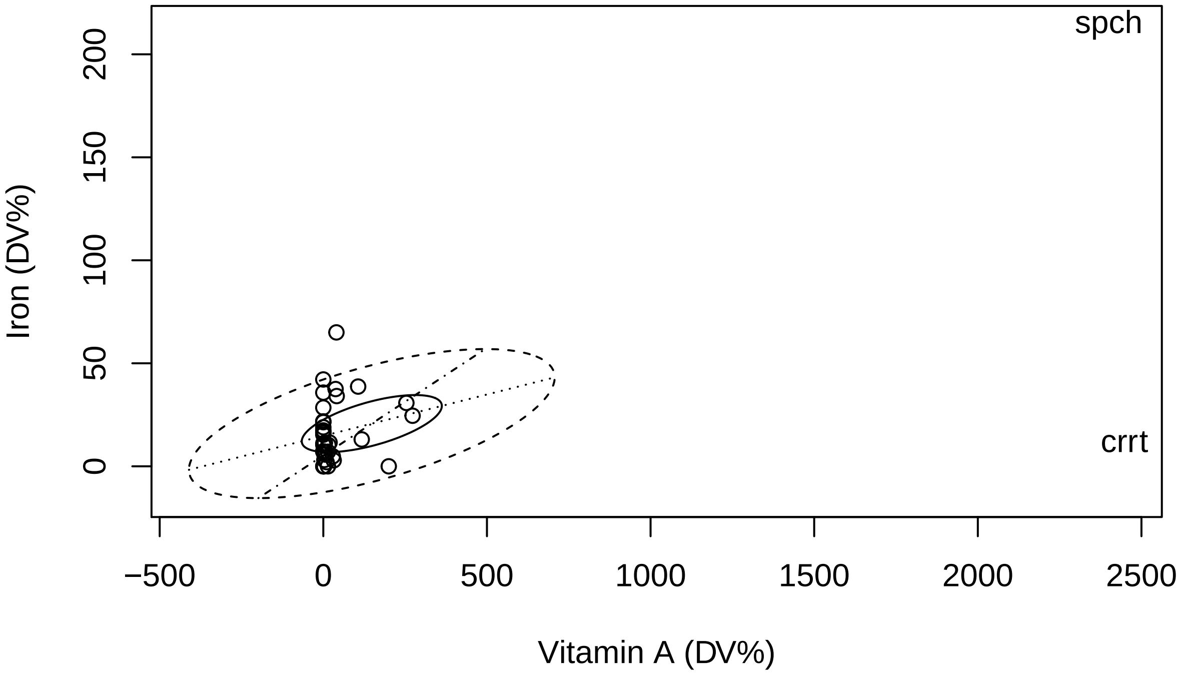


Figure 7: Nonrobust bivariate boxplot of vitamin A vs iron

The scatterplot of vitamin A versus calcium indicates that spinach, carrots, milk, yogurt, cheese, and celery are all potential outliers. Removing them decreased the correlation coefficient from 0.48 to 0.21.

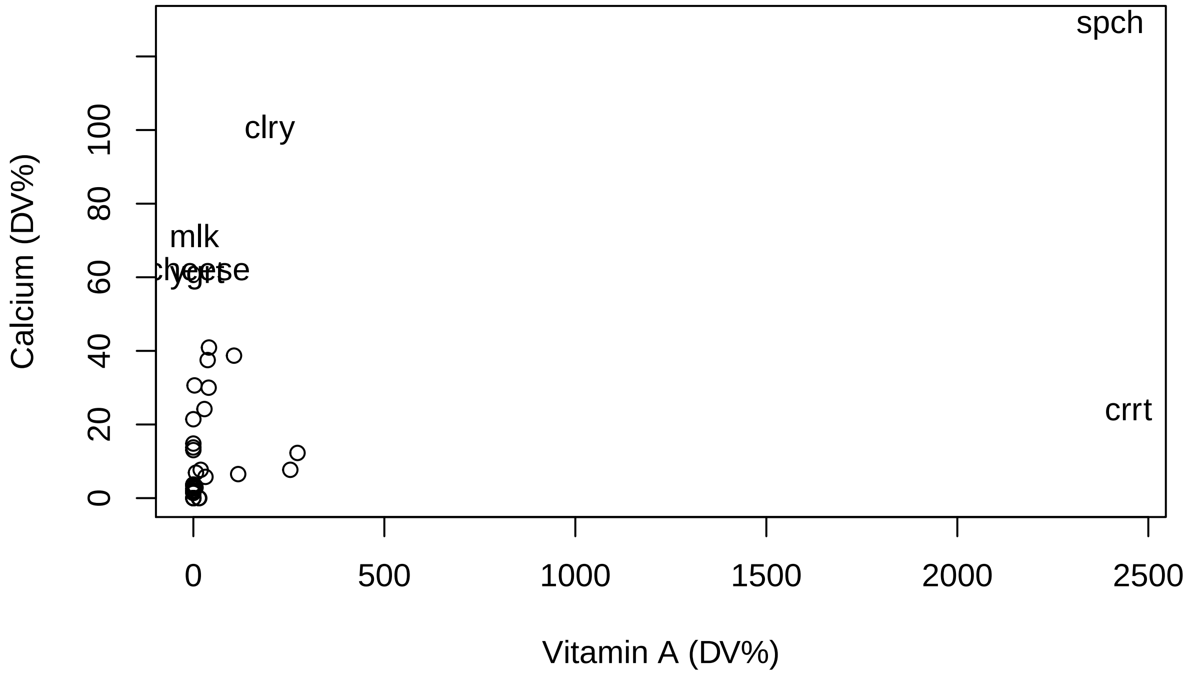


Figure 8: scatterplot of vitamin A vs calcium.

The bivariate boxplot of iron vs calcium indicate spinach, celery, milk, cheese, yogurt, and peas as potential outliers. Removing them decreased the correlation coefficient from 0.53 to 0.39.

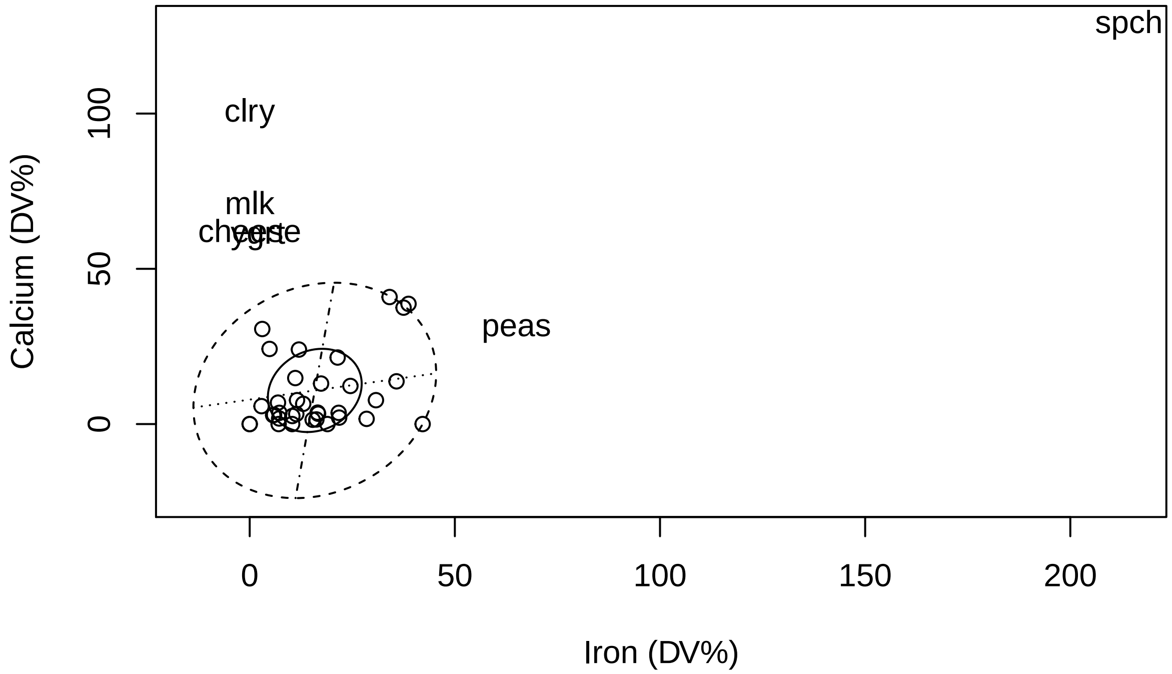


Figure 9: robust bivariate boxplot of iron vs calcium

These observation shouldn't lead us to the conclusion that these variables aren't related, however. We don't want to exclude these outliers from our analysis, but rather, we should be focusing on them. Instead of disregarding them, these should be the foods we are keying in on as foods to eat, as they often have a high values of multiple vitamins and minerals. They may distort the original correlation coefficient to make the relationship appear stronger, but I would argue there is still a relationship.

When the outliers are removed for vitamin C and weight, there is an opposite effect. The correlation becomes much stronger, increasing from 0.46 to 0.83. Ignoring the foods with quite high values, there is a very strong association at the bottom end for foods with small values of vitamin c and weight.

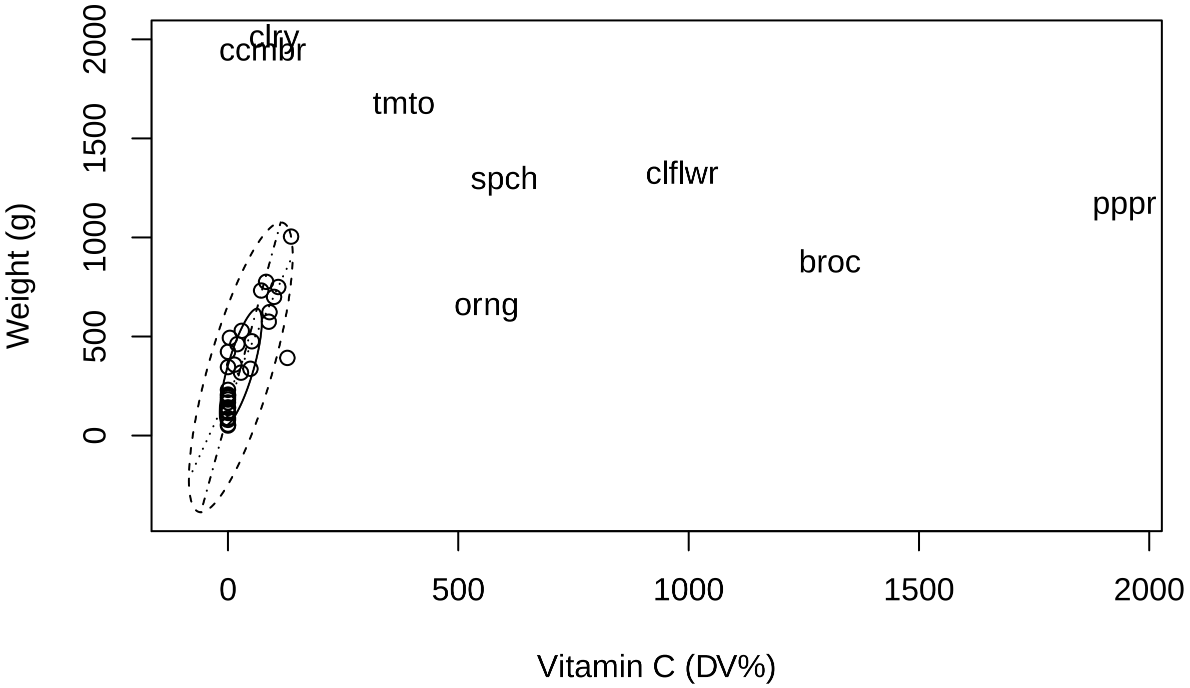


Figure 10: robust bivariate boxplot of vitamin C vs weight.

The bivariate boxplots for the other variables did not highlight any key outliers that would distort the interpretation of the correlation coefficients.

Star Plots

Now that we have a general sense of some of the relationships that these nutrients have, what about how they relate to the food items specifically? Using a star plot, let's take a closer look at how the foods stack up in terms of what they have to offer for vitamins and minerals. Spinach was removed from these plots because it has exceptionally high values for all four categories.

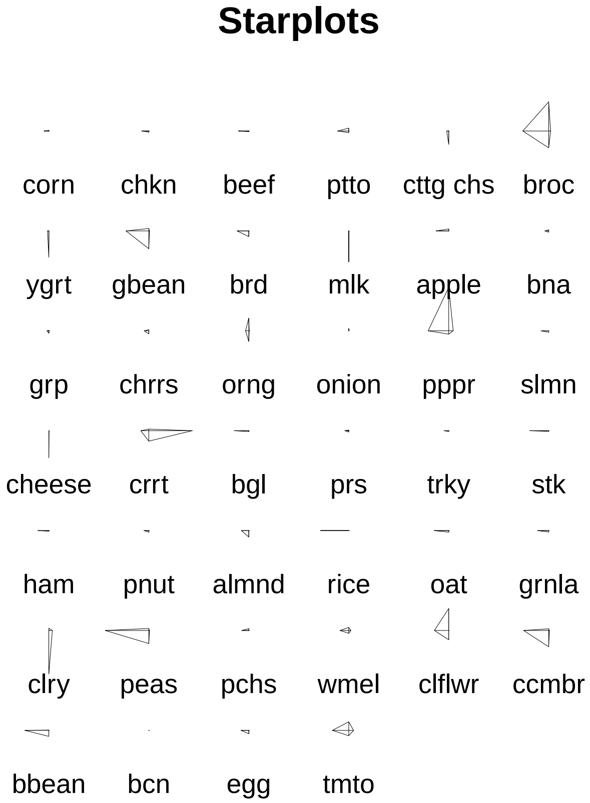


Figure 11: Starplot of all food items except spinach. The axis pointing up corresponds to vitamin C, to the right is vitamin A, toward the bottom is calcium, and to the left is iron.

The plots party reinforce the idea that fruits and veggies are rich in these nutrients, as the food items that stand out are all from these groups. An issue with these plots, however, is that some foods with high values dominate a particular axis. For example, because 300 calories of carrots provides 2448% of your daily value of vitamin A, most other foods appear to offer next to nothing from the graphs (even though this is not true). To address this issue, and address another nutritional question, values that exceeded 100 percent daily value were converted down to 100 percent. This analysis might provide more insight to someone who is simply worried about meeting their daily value requirements, and doesn't care about an excess amount of a particular vitamin or mineral. The star plots are as follows:

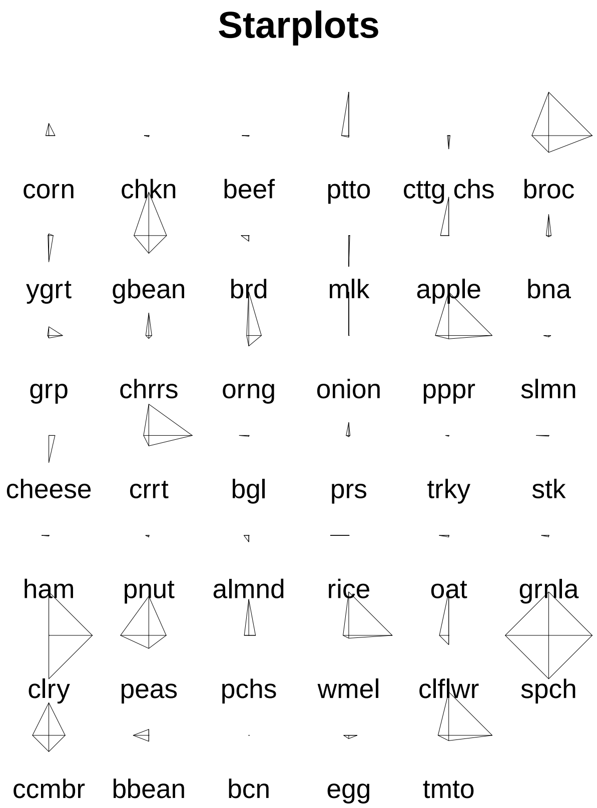


Figure 12: Starplot of all food items with daily values exceeding 100% converted down to 100%. The axis pointing up corresponds to vitamin c, to the right is vitamin A, toward the bottom is calcium, and to the left is iron.

There are a lot of interpretations that can be taken from this graph. One can imagine meeting their vitamin A, vitamin C, iron, and calcium needs by choosing a combination of foods that axes all add up to the max length. Spinach stands out from the rest, as a 300 calories serving would satisfy your daily values for all four of these vitamins and minerals. It's also clear that meats provide little nutritional value in this department. While yogurt, milk, and cheese offer a decent source of calcium, the other nutrients are predominantly found in fruits and vegetables.

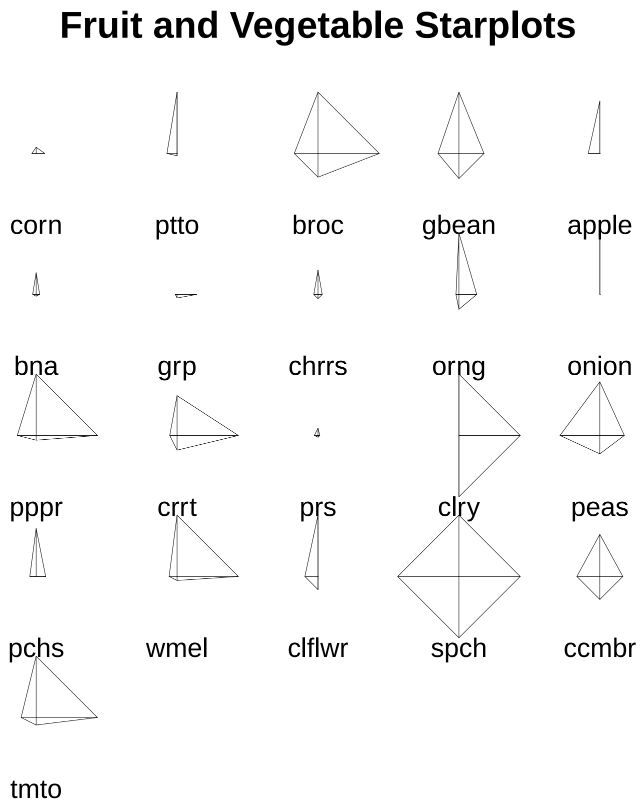


Figure 13: starplot of fruits and vegetables with values exceeding 100% converted to 100%. The axis pointing up corresponds to vitamin c, to the right is vitamin A, toward the bottom is calcium, and to the left is iron.

Taking a closer look, spinach, broccoli, green beans, peas, carrots, peppers and tomatoes have a lot to offer in terms of being well rounded or especially high in one or two of the vitamins and minerals. It's also interesting to see that cherries, corn, pears, bananas and grapes have fairly low values of these vitamins and minerals. It's important not to dismiss their nutritional value completely, though, as they might offer useful nutrients, such as potassium or antioxidants, that are not present in the analysis. Eating these foods, however, won't help you meet some of your most salient nutritional needs. Vitamin C also appears to be much more common, as it's present in a handful of these fruits and vegetables. Iron and calcium, on the other hand, are not nearly as plentiful in fruits and veggies.

Multidimensional Scaling

More generally speaking, we might wonder which of these foods are more similar or distinct. One way to address this is through multidimensional scaling. Since I am concerned about the magnitude of their similarities and not just their ranked order, I will use classical multidimensional scaling. To calculate the proximity matrix, I used the Euclidean distances between the rows and the standardized data. The data was projected onto two dimensions.

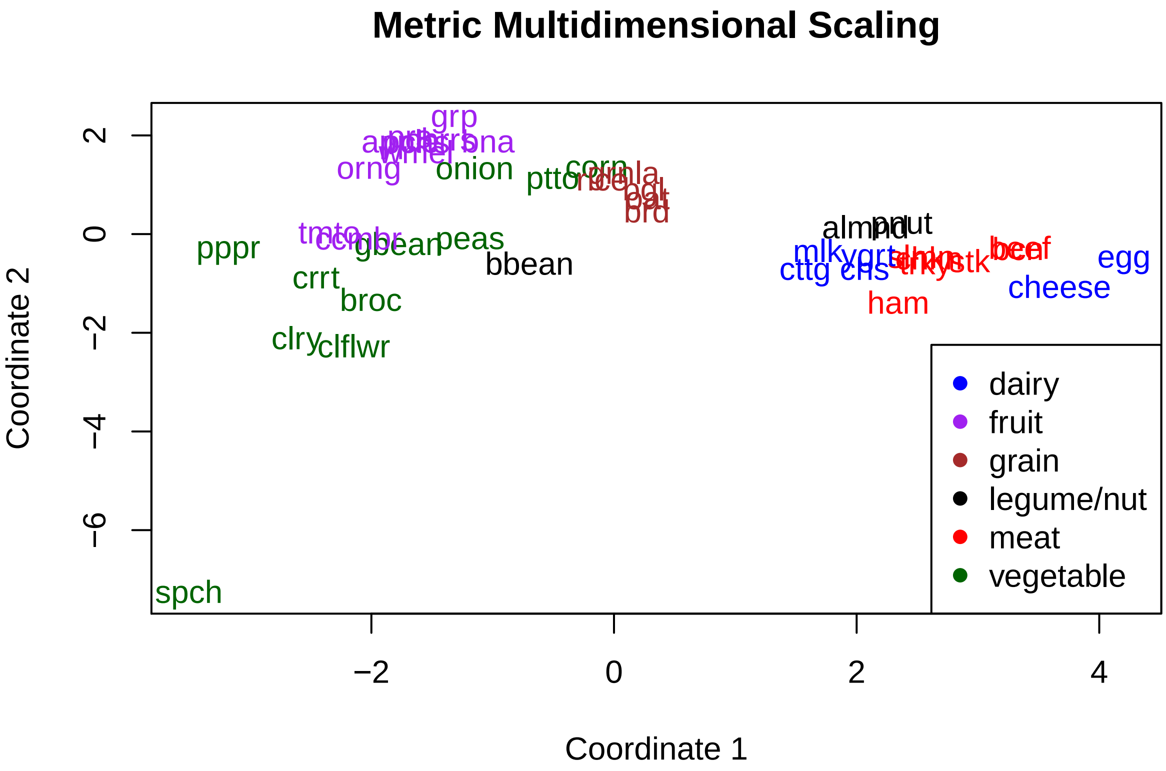


Figure 14

The first observation is one that we have been seeing: spinach is truly a unique food as it's off in nowhere land by itself.

It also becomes clear that most of the data is clustered around its own food group with a few exceptions. Tomatoes and cucumbers, which are technically classified as fruits, are grouped farther from the fruit group and closer to the vegetable group. Corn and potatoes also find themselves closer to grain foods than their own vegetable group. This makes some intuitive sense, as these foods are often linked to food groups they don't technically belong in. There seems to be an actual nutrient based explanation for this, and it's not solely just the taste and physical evaluation our brain makes. These foods truly do behave like foods in different food groups at the nutrient level. This is also observed in black beans, which has a closer resemblance to other vegetables than peanuts and almonds, and it is not too far from the grain group. Perhaps slightly less intuitive is that onions find themselves closer to the fruit group than vegetables. Humans usually don't think of an onion as a fruit from its taste and appearance, but it appears its nutrient composition is more similar to this group.

On the grander scale, there is clear divide down the middle of the graph. Meat, dairy, and nuts are found exclusively to the right, while fruits and vegetables are founds exclusively to the left. Grains find themselves closer to fruits and vegetables, but are more toward the center than any other group. All in all, multidimensional scaling divided the food into clusters that matched much of our preconceived notions of which of these foods are similar and dissimilar. While foods are grouped by biological reasons, such as containing seeds or being the flesch of an animal, their nutrient composition is reflected fairly accurately by these groupings as well.

To further explore these natural groupings and the possible reasoning, I will resort to a similar approach in principal component analysis. To examine the macronutrient relationship among these foods, I removed the four vitamins and minerals as well as sodium for analysis. Because the macronutrient data are not on relatively similar scales, I will use the correlation matrix instead of the covariance matrix to calculate the principal components. The data was plotted on the first two principal components, with the first component accounting for 52% of the variance and the second component accounting for an additional 18%.

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     1.9129163 1.1271286 0.9135128 0.77484728 0.64015333
## Proportion of Variance 0.5227498 0.1814884 0.1192151 0.08576976 0.05854233
## Cumulative Proportion  0.5227498 0.7042382 0.8234533 0.90922310 0.96776542
##                            Comp.6      Comp.7
## Standard deviation     0.45521194 0.135735529
## Proportion of Variance 0.02960256 0.002632019
## Cumulative Proportion  0.99736798 1.000000000
```

Table 1

```
## 
## Loadings:
##               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## total.fat      0.428  0.281  0.413                0.614  0.421
## saturated.fat  0.398  0.376  0.359        -0.234 -0.709       
## cholesterol    0.301  0.262 -0.642  0.570  0.316              
## sodium         0.131 -0.695  0.368  0.577  0.152              
## sugar         -0.390  0.240         0.544 -0.660  0.222       
## total.carbs   -0.510                       0.207 -0.252  0.790
## protein        0.372 -0.412 -0.387 -0.151 -0.578         0.431
## 
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.143  0.143  0.143  0.143  0.143  0.143  0.143
## Cumulative Var  0.143  0.286  0.429  0.571  0.714  0.857  1.000
```

Table 2

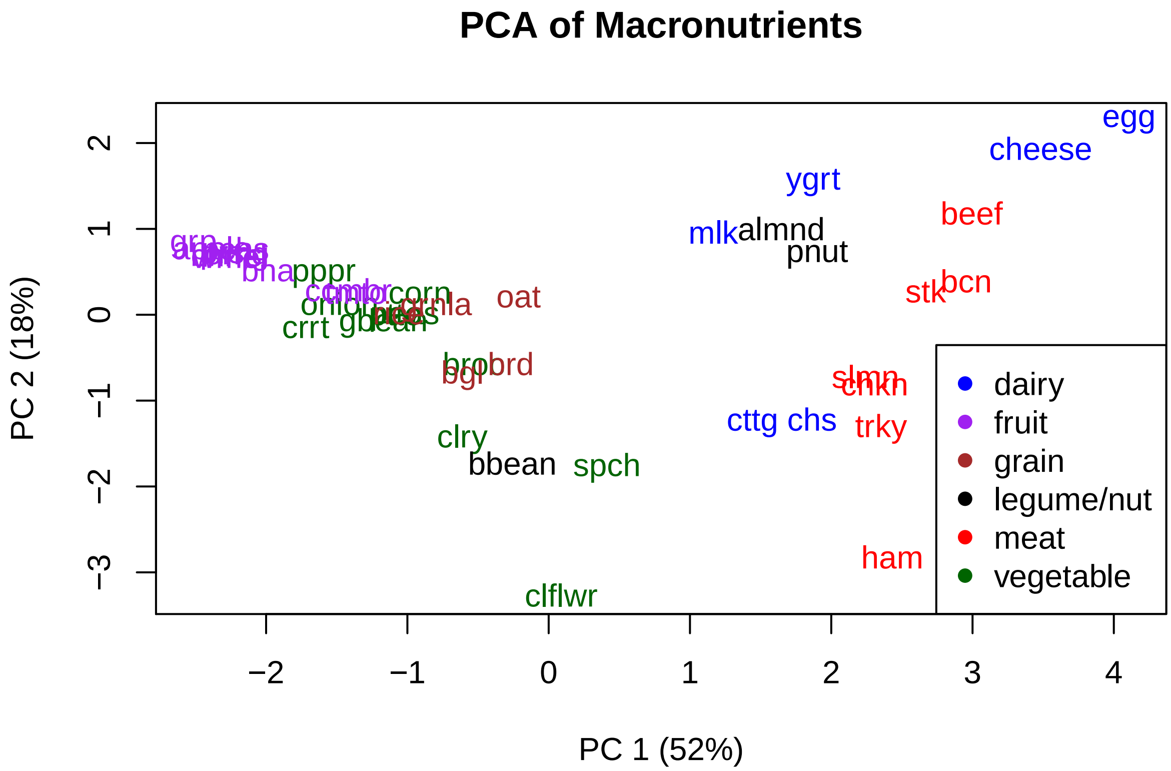


Figure 15

This graph shows a lot of the same patterns that the multidimensional scaling plot showed us, so I will not repeat them. After removing vitamins and minerals, however, we notice that spinach is no longer seen as an extreme outlier. From examining the PCA loadings (Table 2) of the first component, there's some insight as to why there is a clear division between meat, dairy, and nuts versus fruits, vegetables, and grain. The first component seems to represent the contrast that carbohydrates (total carbs, fiber, sugar) have with fats (total fat, saturated fat, cholesterol) and protein. The second component is also able to distinguish fruit from vegetables fairly effectively. From looking at the loadings and examining the data closer, this is mostly due to the distinct differences in sugar, fiber, and protein. Fruits tend to much higher in sugar content, whereas vegetables are far and away the best source of fiber. Fruits also have the lowest protein values among all the food groups, while vegetables provide some protein, sometimes in high amounts. For example, you may have heard the dubious saying that "broccoli has more protein than steak", which relies on a calorie for calorie analysis like ours. While my analysis disproves this notion, broccoli does offer a high value of 24g of protein compared to steak's 40g.

Turning the analysis back to vitamins and minerals, let's use the same principal component analysis approach on vitamin A, vitamin C, calcium, and iron. The data was plotted on the first two principal components, with the first component accounting for 55% of the variance and the second component accounting for an additional 23%.

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4
## Standard deviation     1.4792777 0.9533260 0.7350762 0.60213771
## Proportion of Variance 0.5470657 0.2272076 0.1350843 0.09064246
## Cumulative Proportion  0.5470657 0.7742733 0.9093575 1.00000000
```

Table 3

```
##
## Loadings:
##           Comp.1 Comp.2 Comp.3 Comp.4
## vitamin.a  0.554  0.221  0.536  0.598
## vitamin.c  0.284 -0.946         0.141
## calcium    0.519  0.232 -0.807  0.158
## iron       0.586         0.238 -0.773
##
##                Comp.1 Comp.2 Comp.3 Comp.4
## SS loadings      1.00   1.00   1.00   1.00
## Proportion Var   0.25   0.25   0.25   0.25
## Cumulative Var   0.25   0.50   0.75   1.00
```

Table 4

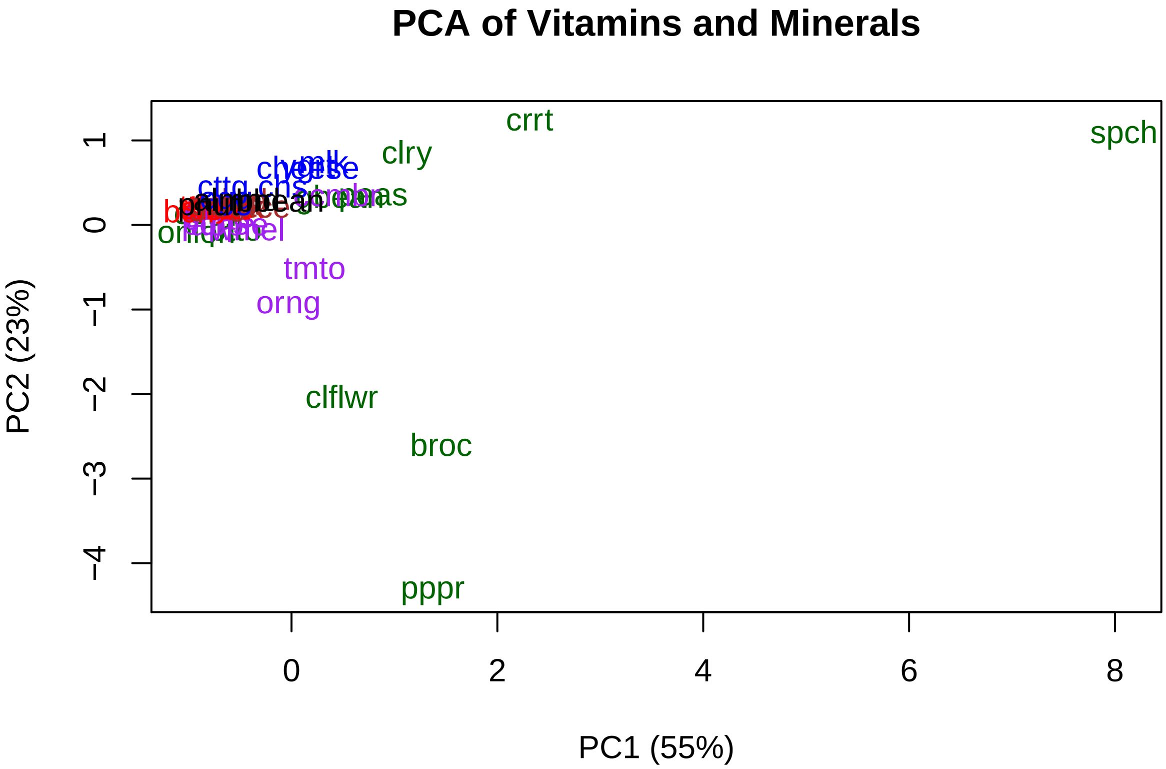


Figure 16

From the first component's loading in Table 4, we can tell that values higher on the first axis tend be higher vitamins and minerals. The second axis largely reflects values of vitamin C. Most of the data is clustered around (0,0), with a handful of exceptions. Celery, carrots, spinach, tomatoes, oranges, cauliflower, broccoli, and peppers are separated from the pack because of their high nutritional values. While this plot is able to highlight these exceptionally rich foods, it does a poor job offering nutritional insight on the rest of the data. This is not ideal, because many of these foods do offer key nutrients that can help a healthy diet. To get a better picture of these lower valued foods, the analysis was repeated after removing the eight previously mentioned foods that stood out from the cluster and plotted with a biplot.

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4
## Standard deviation     1.3162182 1.0185979 0.9208440 0.61812156
## Proportion of Variance 0.4331076 0.2593854 0.2119884 0.09551856
## Cumulative Proportion  0.4331076 0.6924930 0.9044814 1.00000000
```

Table 5

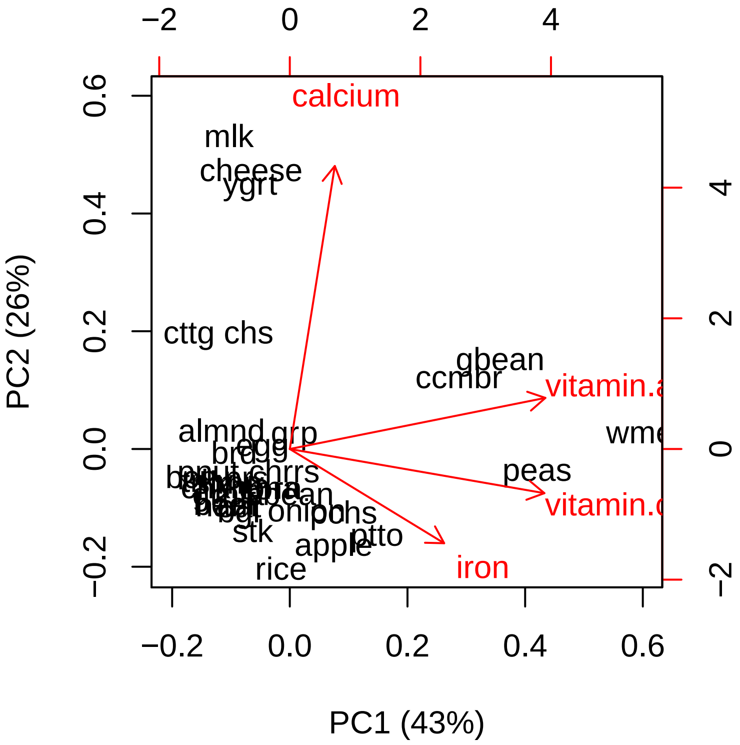


Figure 17: Biplot of vitamins and minerals with outliers removed.

The graph explains 69% of the total variance, with the first component explaining 43% and the second component explaining 23% of the total variance. From examining the PCA loadings depicted on the biplot, this has similar interpretation to the previous plot, as values clustered around the the bottom left have lower nutritional values. However, the second component largely reflects calcium values. Once again, we can see that dairy items serve as a good source of calcium as milk, cheese, cottage cheese, and yogurt all stand apart from the lower cluster. One of the several observations that could me made is that meat and grain products offer little vitamins and minerals. With the exception of rice and bread providing a slight amount of nutrition, the other foods are exceptionally low in these values. Some iron is present, but these foods should be eaten with macronutrient needs in consideration, not for vitamins and minerals. Cucumbers, watermelons, green beans, and peas also separate themselves as the “second tier” of nutrient rich foods.

Let’s bring the data back together and see if there any patterns we haven’t found with the entire dataset. Using a scree plot to graph the principal components versus the eigenvalues, it appears that three principal components could be reasonable from the elbow in the graph. It’s important to note that these three components only account for 66% of the total variation, so they do not tell the entire story. However, because I want to maintain interpretability and the principal components are harder to understand, I don’t think it’d be advantageous to examine 7 or 8 components to account for most of the variation.

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3     Comp.4
## Standard deviation     2.1713404 1.6402248 1.07633674 1.04507915
## Proportion of Variance 0.3626707 0.2069490 0.08911544 0.08401465
## Cumulative Proportion  0.3626707 0.5696197 0.65873516 0.74274981
##                            Comp.5     Comp.6     Comp.7     Comp.8
## Standard deviation     0.93851993 0.82181074 0.71534779 0.67994150
## Proportion of Variance 0.06775536 0.05195176 0.03936327 0.03556311
## Cumulative Proportion  0.81050517 0.86245693 0.90182020 0.93738331
##                           Comp.9    Comp.10    Comp.11     Comp.12
## Standard deviation     0.5351310 0.50309695 0.39789315 0.330016233
## Proportion of Variance 0.0220281 0.01946973 0.01217838 0.008377747
## Cumulative Proportion  0.9594114 0.97888114 0.99105952 0.999437267
##                             Comp.13
## Standard deviation     0.0855308733
## Proportion of Variance 0.0005627331
## Cumulative Proportion  1.0000000000
```

Table 6

```
##                      Comp.1     Comp.2      Comp.3
## weight         0.338930397  0.1844344  0.10225322
## total.fat     -0.353488576  0.1432457 -0.08263025
## saturated.fat -0.348868975  0.1099747 -0.23403655
## cholesterol   -0.251769510  0.0723652 -0.05503838
## sodium         0.003956557  0.3270206  0.58240843
## sugar          0.291900640 -0.2722536 -0.09909457
## fiber          0.378853077  0.2137961  0.15240732
## total.carbs    0.411662847 -0.2329770 -0.09908976
## protein       -0.238434066  0.3416503  0.29176525
## vitamin.a      0.171102055  0.3544666 -0.41728110
## vitamin.c      0.226704332  0.1793665  0.31989077
## calcium        0.109283338  0.4423302 -0.32243021
## iron           0.165506959  0.4205735 -0.27804010
```

Table 7: loadings for first three principal components.

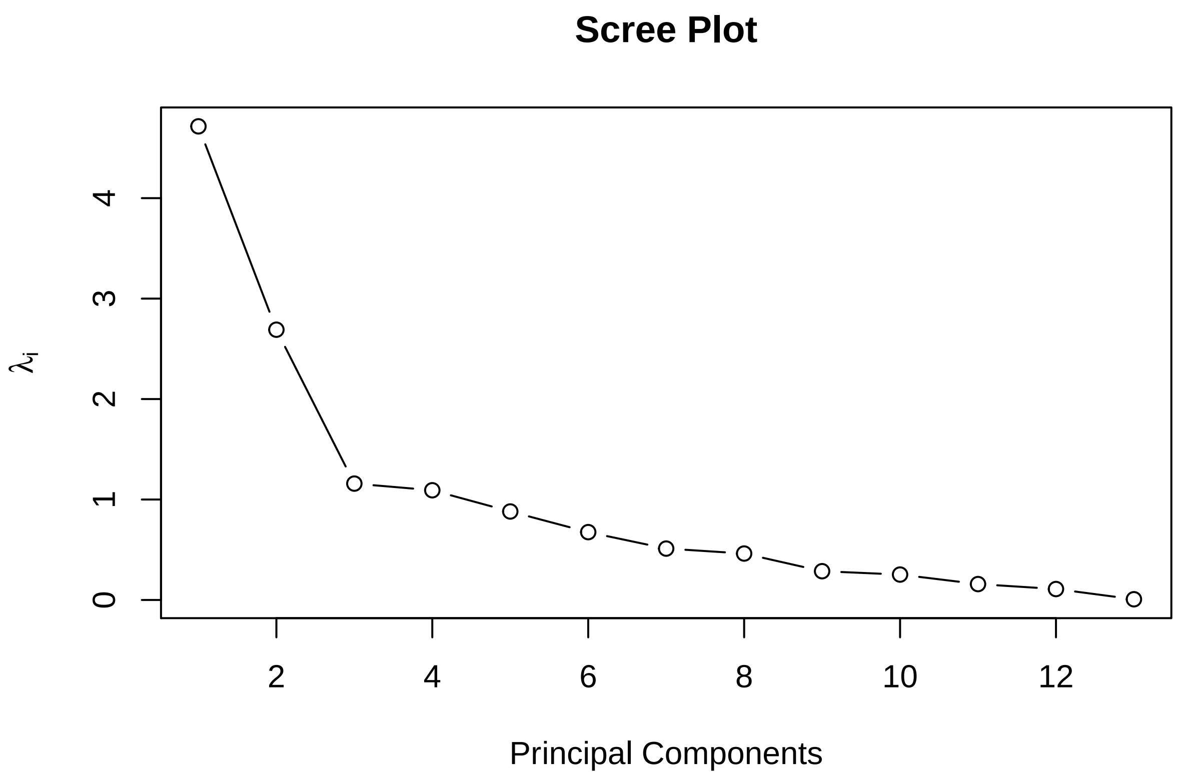


Figure 18

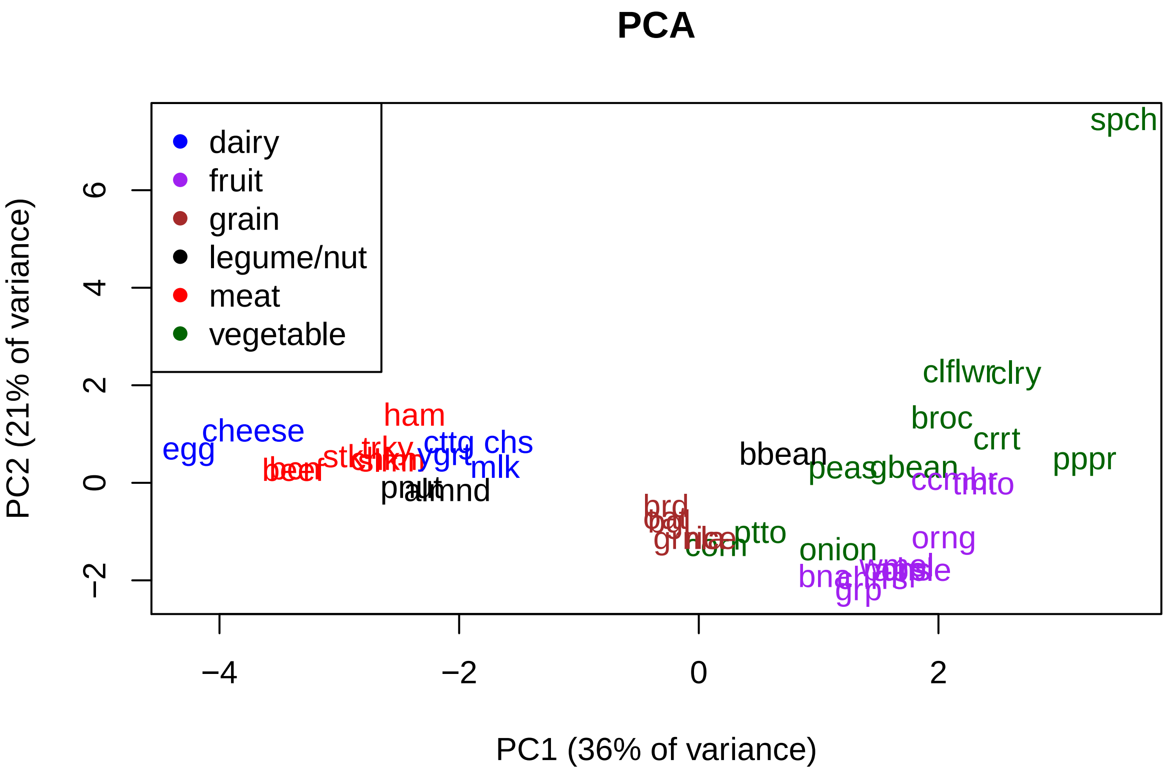

PCA
dairy
fruit
grain
legume/nut
meat
vegetable
spch
clflwr
clry
broc
crrt
pppr
bbean
peas
gbean
ccmbr
tmto
ham
cheese
egg
cttg chs
trky
ygrt
mlk
beef
stk
almnd
brd
ptto
corn
onion
orng
bna
grp
PC2 (21% of variance)
PC1 (36% of variance)
6
4
2
0
−2
−4
−2
0
2

Figure 19

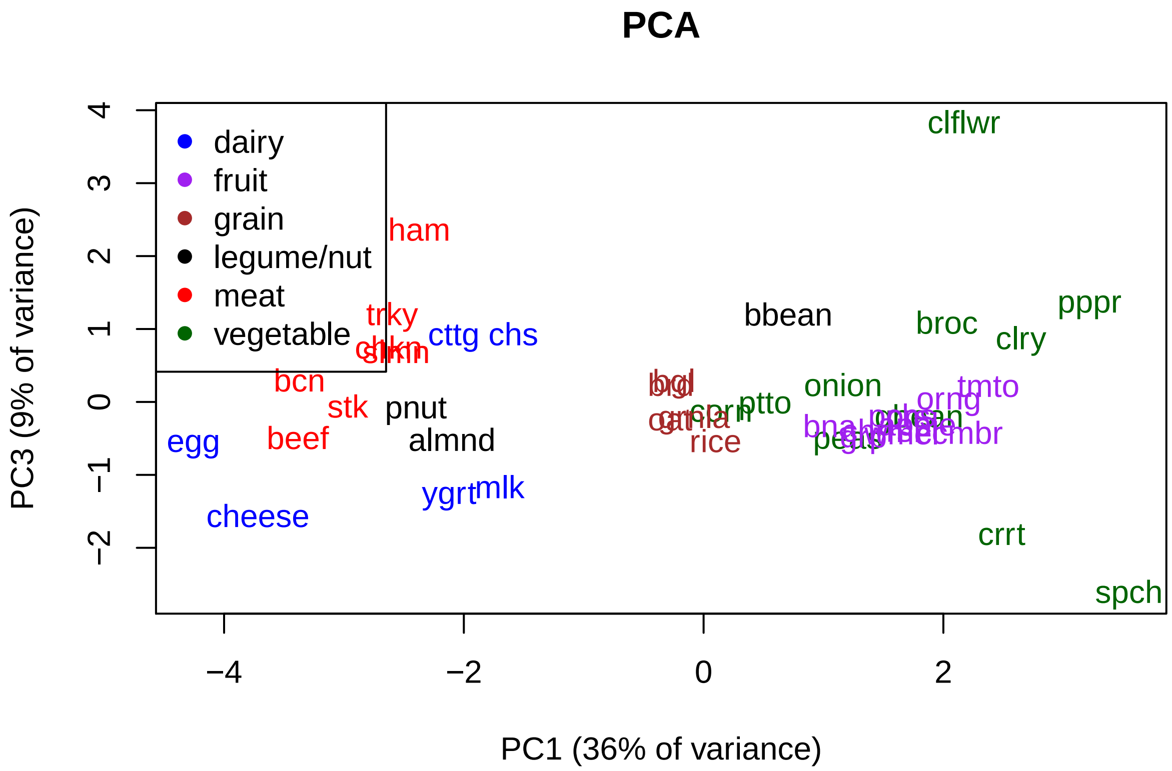

PCA
dairy
fruit
grain
legume/nut
meat
vegetable
PC3 (9% of variance)
PC1 (36% of variance)
−4
−2
0
2
4
3
2
1
0
−1
−2
clflwr
ham
bbean
pppr
broc
clry
trky
cttg chs
chkn
slmn
bcn
bgl
onion
tmto
orng
ptto
crn
stk
pnut
egg
beef
almnd
rice
ygrt
mlk
cheese
crrt
spch

Figure 20

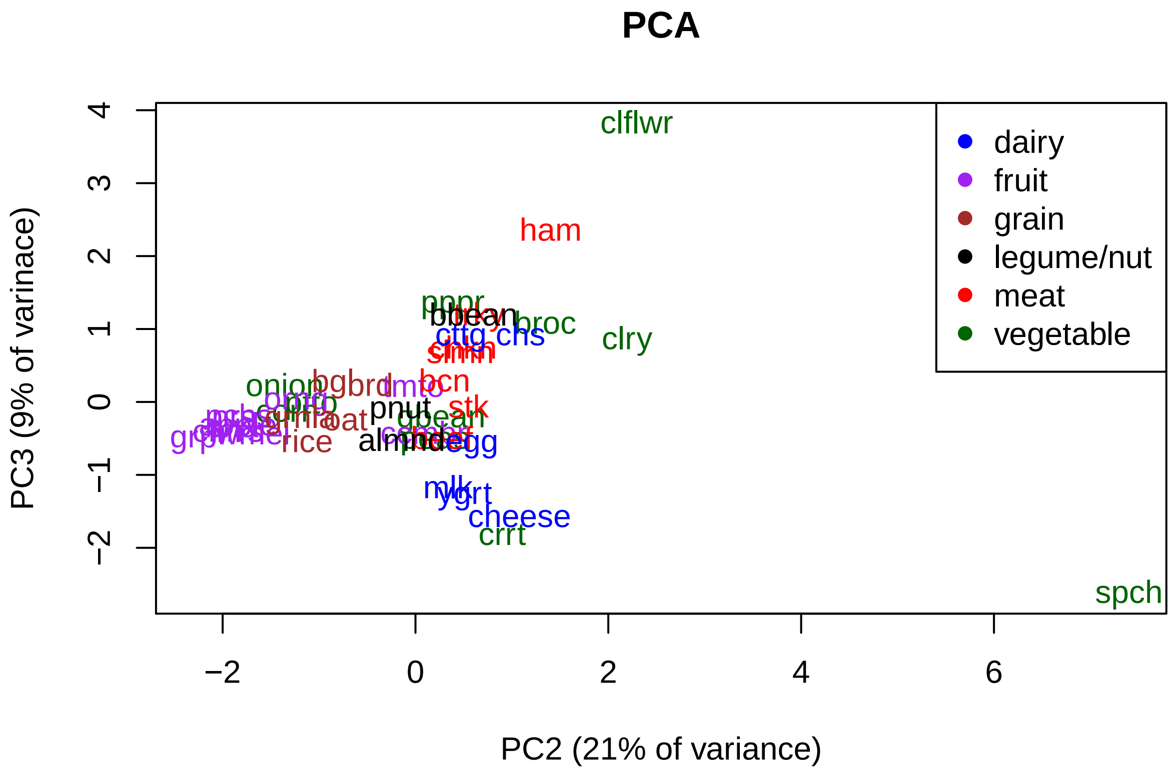


Figure 21

The third component in this analysis highlights cauliflower as distinct from the other food items. Upon further analysis, this is likely due to the insanely high sodium values it has. While cauliflower has been noted as an excellent source of vitamin C and also fiber, this may not be the ideal food for someone who needs to watch their sodium intake. Ham is also highlighted as by the third principal component, likely because of its high sodium values as well.

From comparing all three graphs, fruits appear to be consistently lumped together and have less variation than the other food groups. Grains follow this same pattern. Foods in the fruit and grain group could be considered more like the foods in their own group. On the other hand, if you look at vegetables, they tend to be more spread out, indicating a more diverse range of nutrients. A fruit or grain product has a more predictable nutrient composition than a vegetable. Dairy products also show some variation, but not nearly as much as vegetables.

Clustering

We've seen from multidimensional scaling and principal component analysis that the data generally tends to cluster around its assigned food group. If we apply our own clustering techniques, will these findings be confirmed or possibly lead us to new clusters and discoveries?

Initially, I attempted k means clustering analysis by minimizing the the within-group sum of squares over the standardized variables. The scree plot did not indicate a clear choice for k, so values from 3 to 10 were tested to see if they found interesting/insightful clusters. This approach didn't produce useful results, and it was often hard to understand why certain items were being clustered together. I assume this method didn't work well because of its tendency to cluster into shapes of about equal sizes, which is not ideal for this dataset considering there is likely different variances across clusters.

To address k-means shortcomings, I turned to a model-based clustering approach. BIC was calculated on cluster values from 1 to 8 for the various models. The lowest BIC value was from the “VII” model with 6 clusters. The “VII” model is characterized as having spherical clusters of unequal volume, partly confirming why k-means wasn’t satisfactory. The clusters are as displayed in print and also using multidimensional scaling.

```
## [1] "Cluster 1"
## [1] "corn"  "ptto"  "brd"   "bgl"   "rice"  "oat"   "grnla"
## [1] "Cluster 2"
## [1] "chkn" "slmn" "trky"
## [1] "Cluster 3"
## [1] "beef"     "cttg chs" "ygrt"     "mlk"      "cheese"   "stk"
## [7] "ham"      "bcn"
## [1] "Cluster 4"
##  [1] "broc"   "gbean"  "pppr"   "crrt"   "clry"   "peas"   "clflwr"
##  [8] "spch"   "ccmbr"  "bbean"  "egg"    "tmto"
## [1] "Cluster 5"
## [1] "apple" "bna"   "grp"   "chrrs" "orng"  "onion" "prs"   "pchs"  "wmel"
## [1] "Cluster 6"
## [1] "pnut"  "almnd"
```

Table 8: Cluster assignments

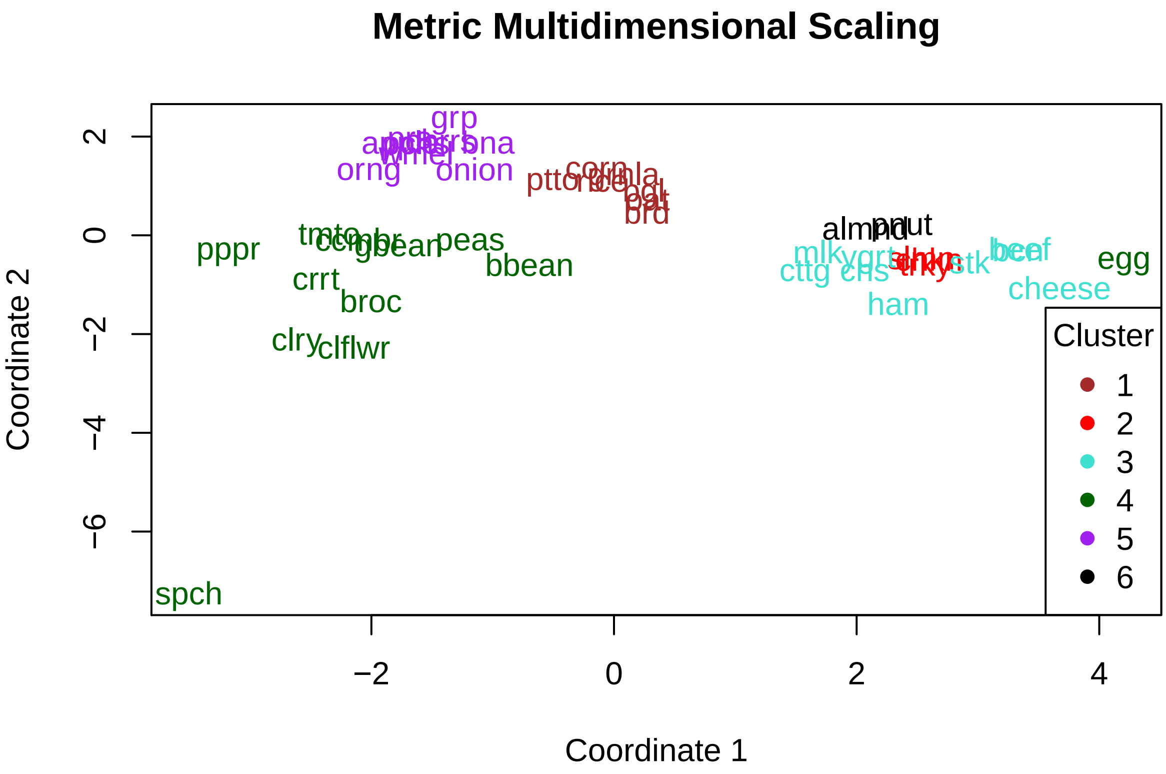


Figure 22

The cluster confirms some of the previous observations we’ve made. Cluster 1 contains all of the grain products and includes two vegetables, corn and potatoes, which humans commonly associate with grain.

Cluster 5 contains all fruits with the exception of onions, which has been discussed as being nutritionally similar to fruit. Tomatoes, cucumber, and black beans abandoned their technical food group as well, and fall in the Cluster 4 with the other vegetables. Surprisingly, eggs fall into this vegetable cluster, too. Cluster 2 is made up of turkey, salmon, and chicken. Under further inspection, these meats have higher protein values and can be categorized as the “lean meats” from their fat content. These foods are ideal for a health enthusiast trying to meet their protein needs, like a bodybuilder. Cluster 3 contains the rest of the meats along with the dairy products, likely due to their similar protein, fat, and cholesterol content. Cluster 6 is able to distinguish the two nuts from the food items, showing that these truly are a unique class of their own in terms of their nutritional composition.

The model based clustering approach did a good job of finding meaningful and relevant clusters. The one observation that didn’t provide more insight into the data was the classification of the egg in the cluster made up of mostly vegetables. This inspired further analysis, but it was to no avail. I could not come to any logical conclusions of why this food was clustered that way. Hopefully this doesn’t keep me up at night.

Conclusion

Many of the questions I originally asked were answered. We found that there wasn’t a relationship between the macronutrient and micronutrient composition of a food. Foods that were high in one micronutrient, however, were more likely to be higher in another. There were some foods that stood out as particularly nutritious for micronutrients. Spinach could be deemed a “superfood” with exceptionally high vitamin and mineral values. The notion that fruits and vegetables are healthy was reinforced, and the more nutrient rich fruits and vegetables were highlighted. We also saw that vegetables can be a good source of fiber and even protein, whereas fruits offer next to nothing in terms of protein. Meats and grains don’t contain many vitamins or minerals, but chicken, salmon, and turkey serve as great low fat options for protein.

There was also validation in the food groups these items get assigned to. It turns out these classifications extend far beyond biological reasons and do a pretty good job categorizing these food by their nutritional composition. We saw that black beans, tomatoes, cucumber, corn, and potatoes belonging to groups outside their technical classification, but matching our intuition of how we usually think about them. We found that onions were more similar to fruits than vegetables as well.

Another interesting finding was that vegetables were highly diverse, whereas fruit were just the opposite. When picking out a fruit, one could assume they have fairly similar nutrition. Vegetables, however, vary greatly from food to food on which nutrients they provide you with.

These weren’t all of the findings, but just some of the key points. Having a better understanding of we put into our body is important, and there’s much more room for analysis that could be done. For future research, it’d be interesting to include more food items to see how they compare and if they change any of the relationships or observations. One thing is for sure though: eat more spinach!