Analysis of nutrient information from USDA National Nutrient Database using PCA

by Gap Kim

INTRODUCTION

A dataset from United States Department of Agriculture (USDA) National Nutrient Database for Standard Reference (SR) Release 28 has been analyzed. The dataset included 2223 food items with 46 nutrient information. In the exploratory data anlaysis section, distributions of proximates, minerals, vitamins and other selected nutrients have been analyzed where highly right-skewed distributions are observed. To better understand the dataset, a principal component analysis has been performed. The scree plot suggests that the cutoff number of principal component is 7 where the cumulative variance explained is at 56.5%. The importance of nutrients are discussed in terms of their contributions to the principal components. Moreover, the nutrients have been grouped as having a common traits based on their impact on the principal components. Primary and secondary relationships among nutrients have been identified.

EXPLORATORY DATA ANALYSIS

The USDA National Nutrient Database for Standard Reference (SR) is the major source of food composition data in the United States. It provides the foundation for most food composition databases in the public and private sectors. The documentation verison, Release 28 (SR28), used in this report contains data on 8,789 food items and up to 150 food components. After cleaning the data and focusing on only the nutrient variables, the analysis was conducted on a subdataset with 2223 food items with 46 nutrient information.

Proximates

First, the distribution of proximate components, which refer to water(moisture), protein, total lipid (fat), total carbohydrate, and ash, are examined and plotted in Figure 1.

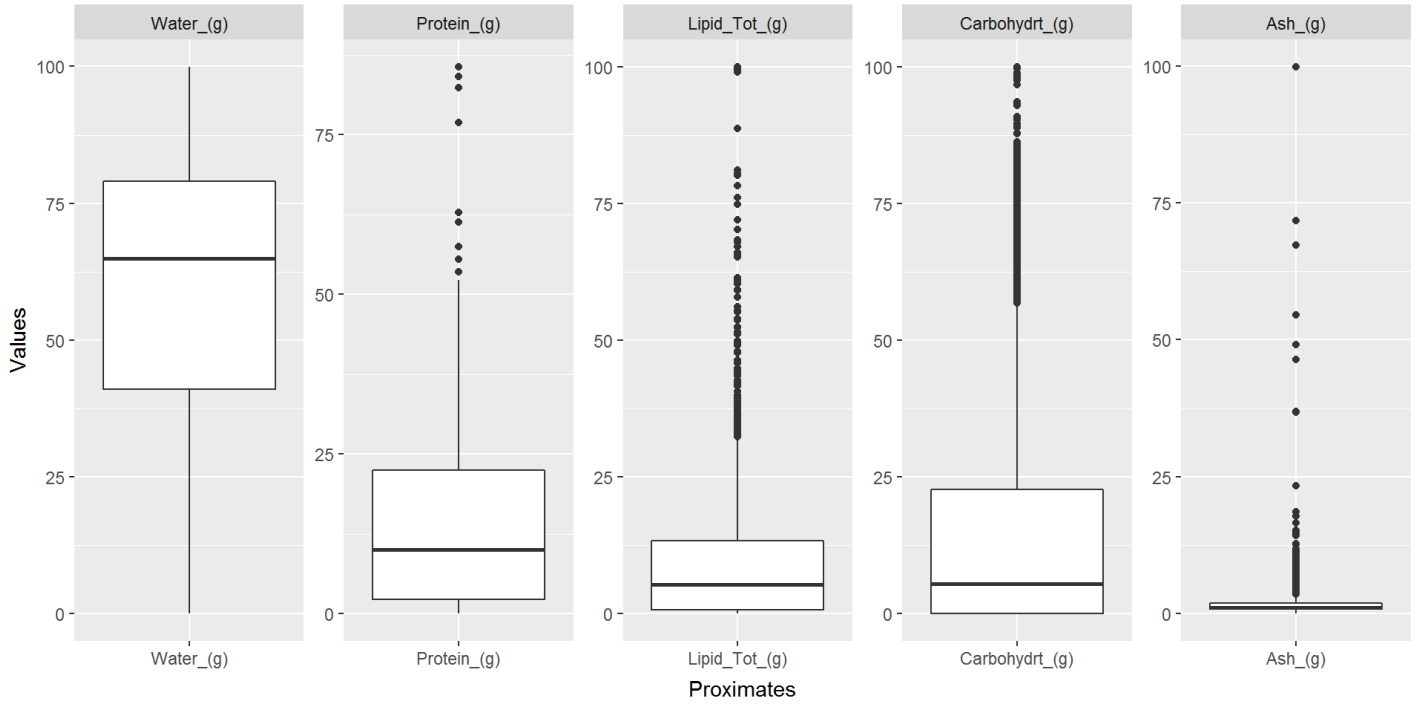

Figure 1: Boxplots of proximates in all foods

The five products that contain the highest protein amount are:

```
[1] "GELATINS,DRY PDR,UNSWTND"
[2] "EGG,WHITE,DRIED,STABILIZED,GLUCOSE RED"
[3] "EGG,WHITE,DRIED,PDR,STABILIZED,GLUCOSE RED"
[4] "EGG,WHITE,DRIED,FLAKES,STABILIZED,GLUCOSE RED"
[5] "COD,ATLANTIC,DRIED&SALTED"
```

As expected, the white region of eggs has high protein content. It is interesting that dried powdered gelatin and dried Atlantic cod have high protein content as well.

Minerals

The minerals included in the dataset are calcium, iron, magnesium, phosphorus, potassium, sodium, zinc, copper, manganese, and selenium, and their respective distributions are given in Figure 2.

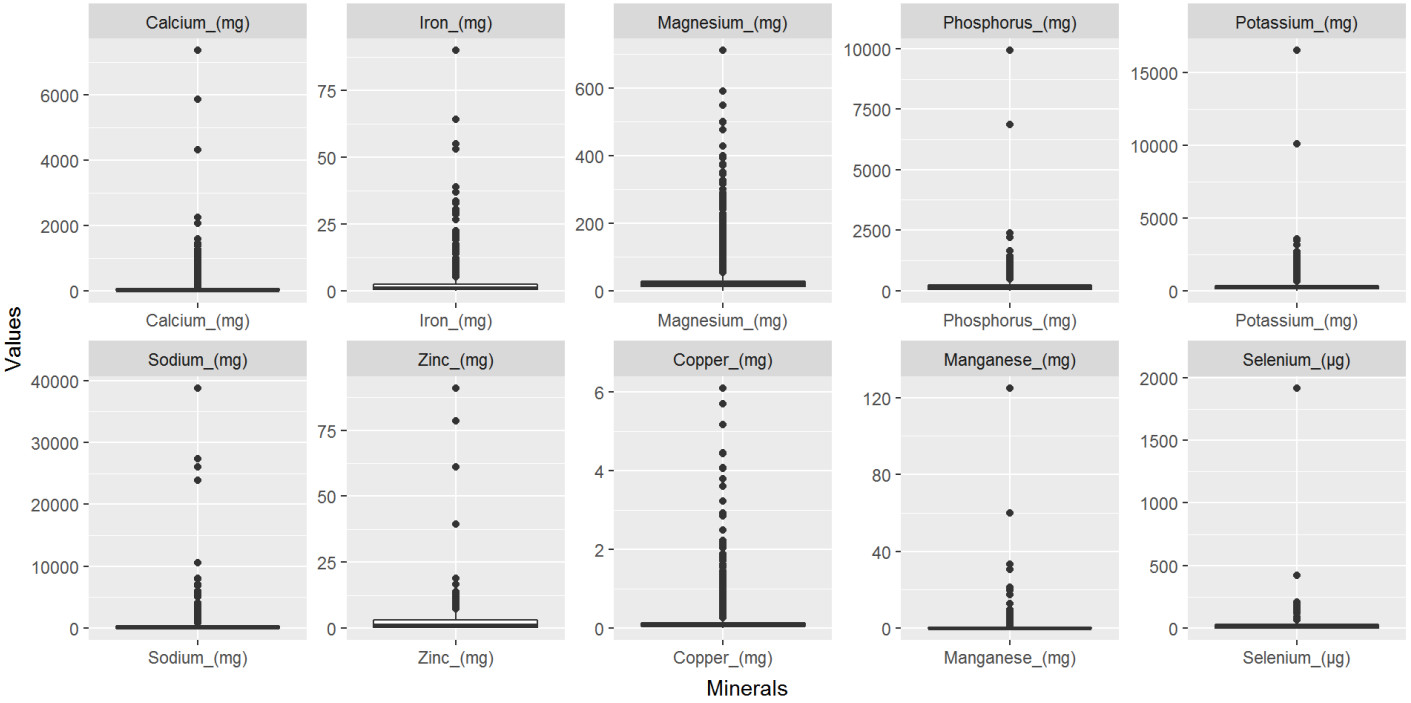

Figure 2: Bolxplots of minerals in all foods

The baking powder came out on top containing the highest amount of calcium:

```
[1] "LEAVENING AGENTS,BAKING PDR,DOUBLE-ACTING,STRAIGHT PO4"
[2] "LEAVENING AGENTS,BAKING PDR,DOUBLE-ACTING,NA AL SULFATE"
[3] "LEAVENING AGENTS,BAKING PDR,LOW-SODIUM"
[4] "SPICES,BASIL,DRIED"
[5] "WHEY,ACID,DRIED"
```

Vitamins and other nutrients

Among other nutrients included in the dataset, distributions of vitamin A, C, D and E and folate are plotted in Figure 3.

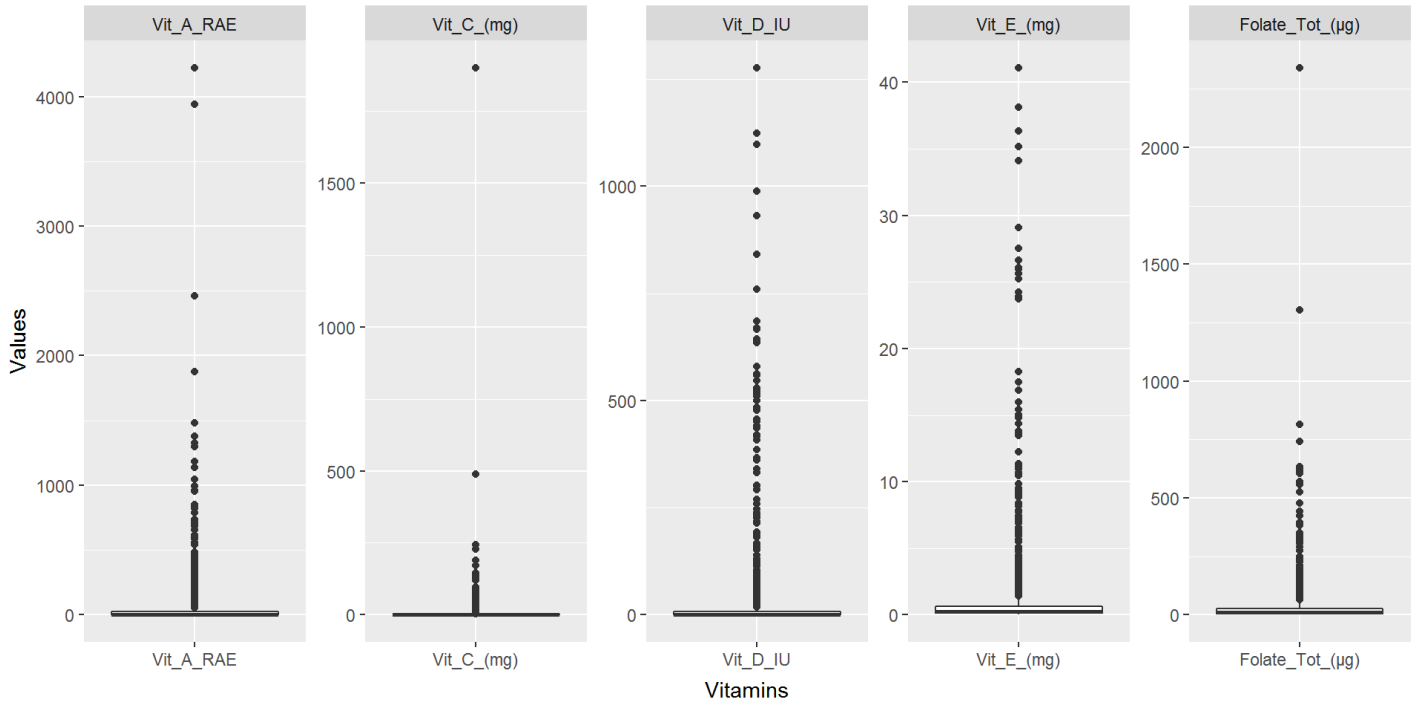

Figure 3: Bolxplots of vitamins and other nutrients in all foods

There seems to be quite a few outliers in most of nutrients as observed from the boxplotes. This implies that specific food products are available that contain specific targetted nutrients. For consumers with specific dietary requirement, the information would be helpful.

PRINCIPAL COMPONENT ANALYSIS

Scree plot and cumulative proportion of variance explained

A prinicpal component analysis has been performed on the dataset with 2223 food products with 46 nutrient information to get a better understanding. The resulting proportion of variance explained by each prinicpal component and their cumulative sum are plotted in Figure 4. Based on the scree plot, it seems not much explanation of variance is gained after 7th principal component, which explains 56.5% of the total variance in the dataset.

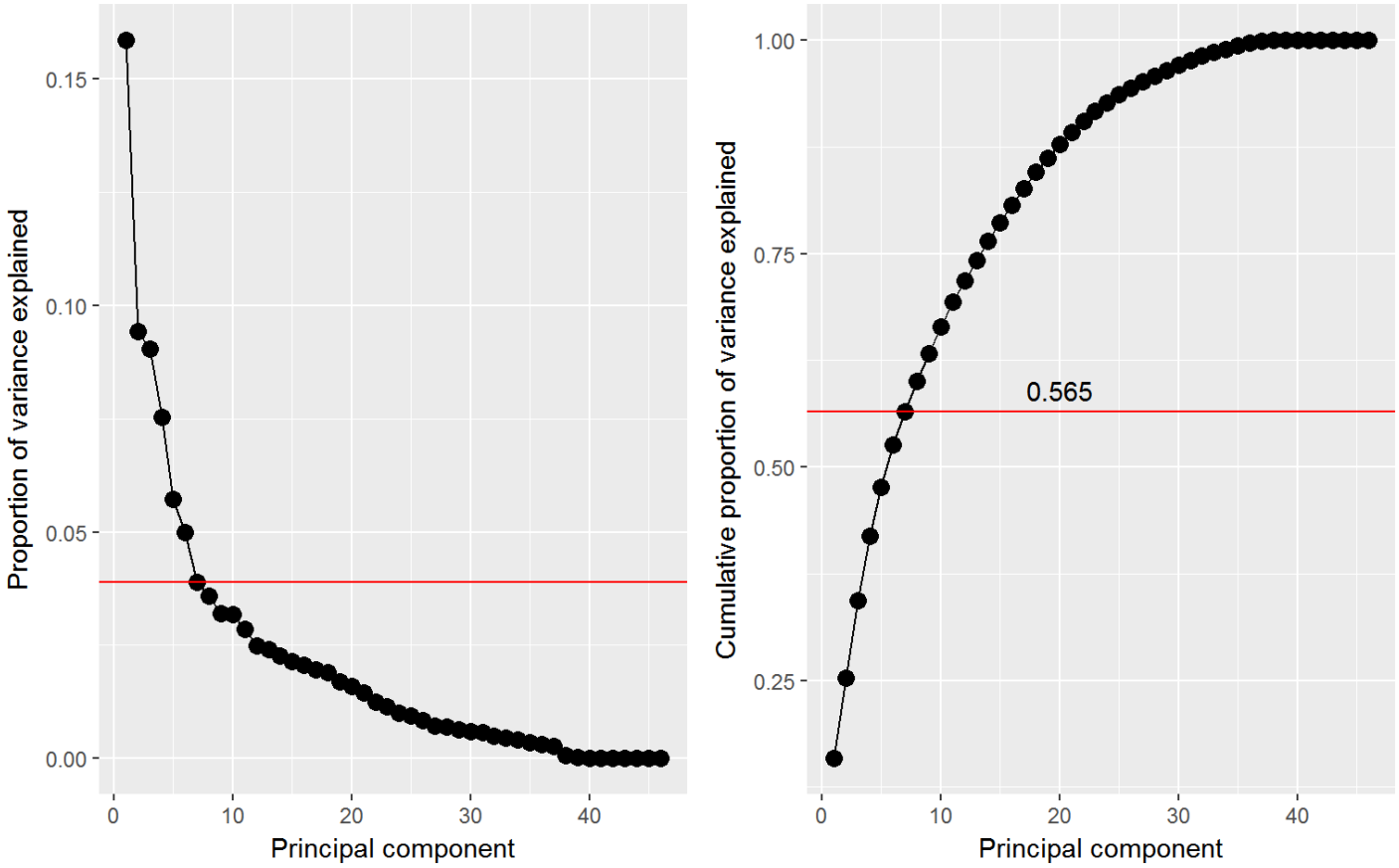

Figure 4: Proportion of variance explained (left) and their cumulative sum from PCA (right

Heatmap of nutrients and principal components

To get an overall pattern on how the 46 nutrient variables contribute to the first 20 principal components, which explains 87.8% of variation, a heatmap has been generated as shown in Figure 5. It is noted that the first principal component (PC1) consists of opposing loadings between water and the other nutrients.

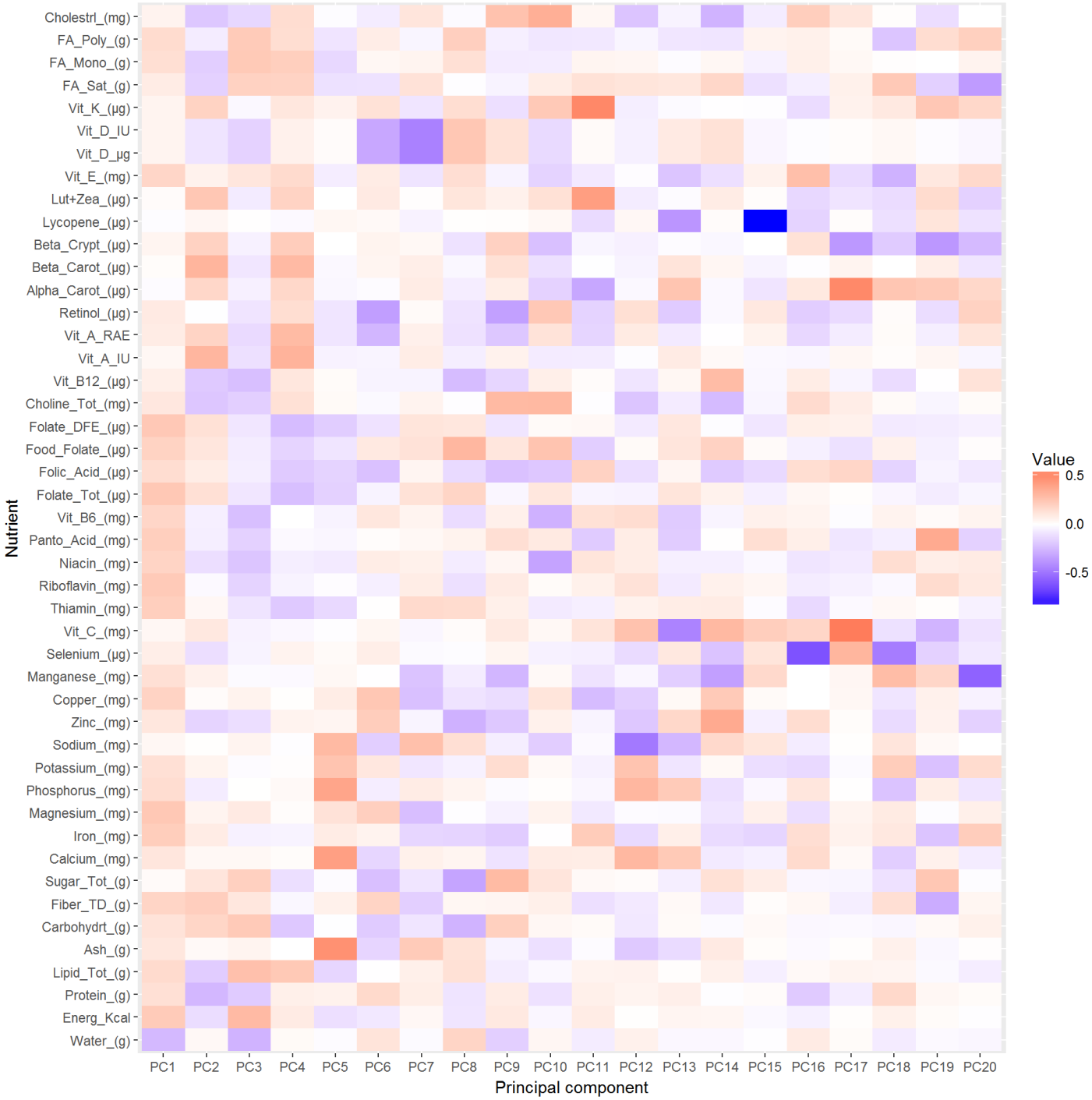

Figure 5: Heatmap of nutrients and first 20 principal components

A full spectrum of the nutrient loadings in the first two principal components are provided in Figure 6. As noted from the heatmap in Figure 5, the first principal component is marked by opposing signs of loadings between water and the rest of nutrients.

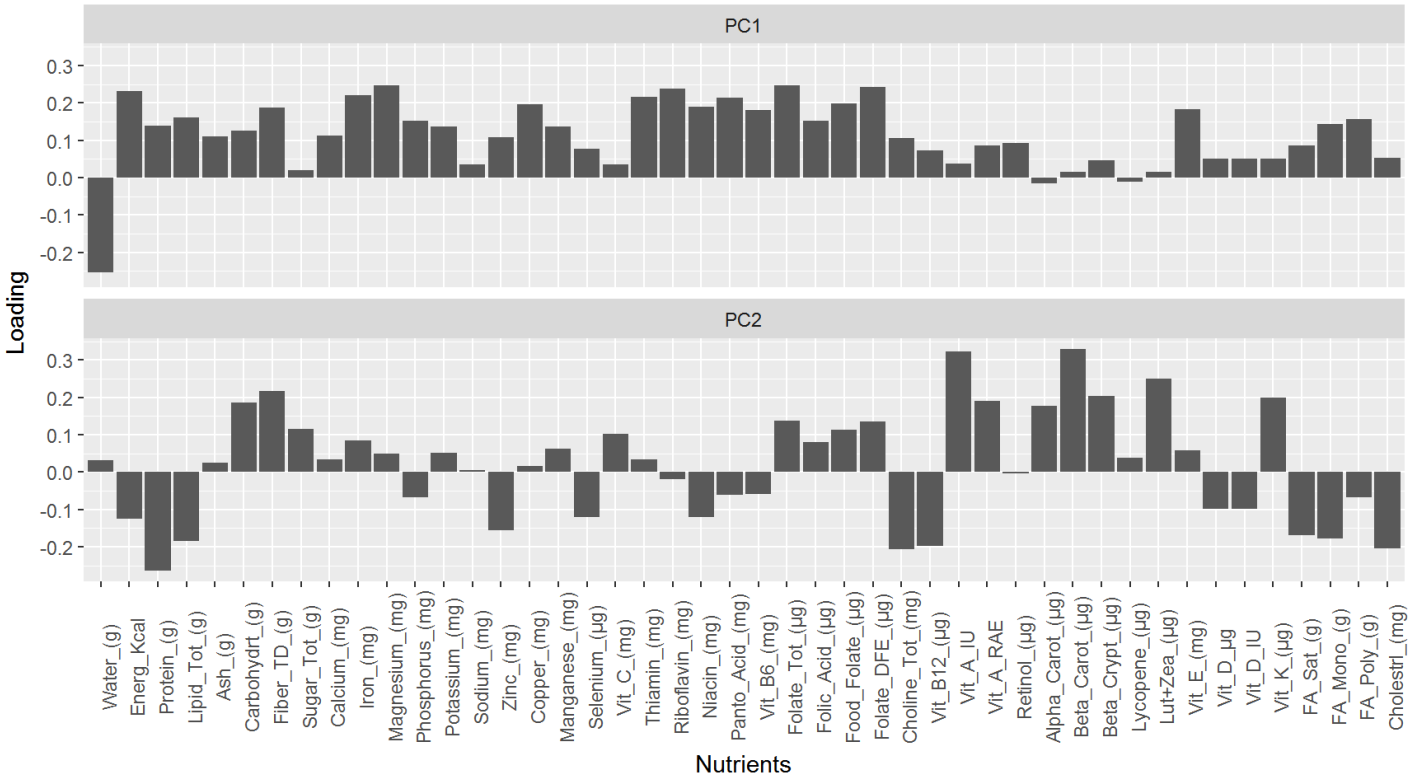

Figure 6: Full spectrum of the nutrient loadings in PC1 and PC2

Selection of nutrients based on their contribtuion to the principal components

The principal component itself does not select variables since each principal component contains all variable information. By comparing the magnitude of the loadings, however, the contribution from each variable can be observed. For example, for the second principal component (PC2) in Figure 6, the water and ash are not as important as energy, protein and lipid since their loadings are smaller. Therefore, the following quantity may be used to compare the relative importance of each variable for the given principal component:

$$\text{Nutrient importance} = \text{Nutrient loading}^2 \times \text{Proportion of variance explained by the principal component}$$

Adding the 'Nutrient importance' values across all the principal components for each nutrient variable, and then summing those values for all variables will be 1. Based on the scree plot analysis, the 'Nutrient importance' values were summed up to 7th principal component. The results have been reordered in descending ordr and summarized in Figure 7.

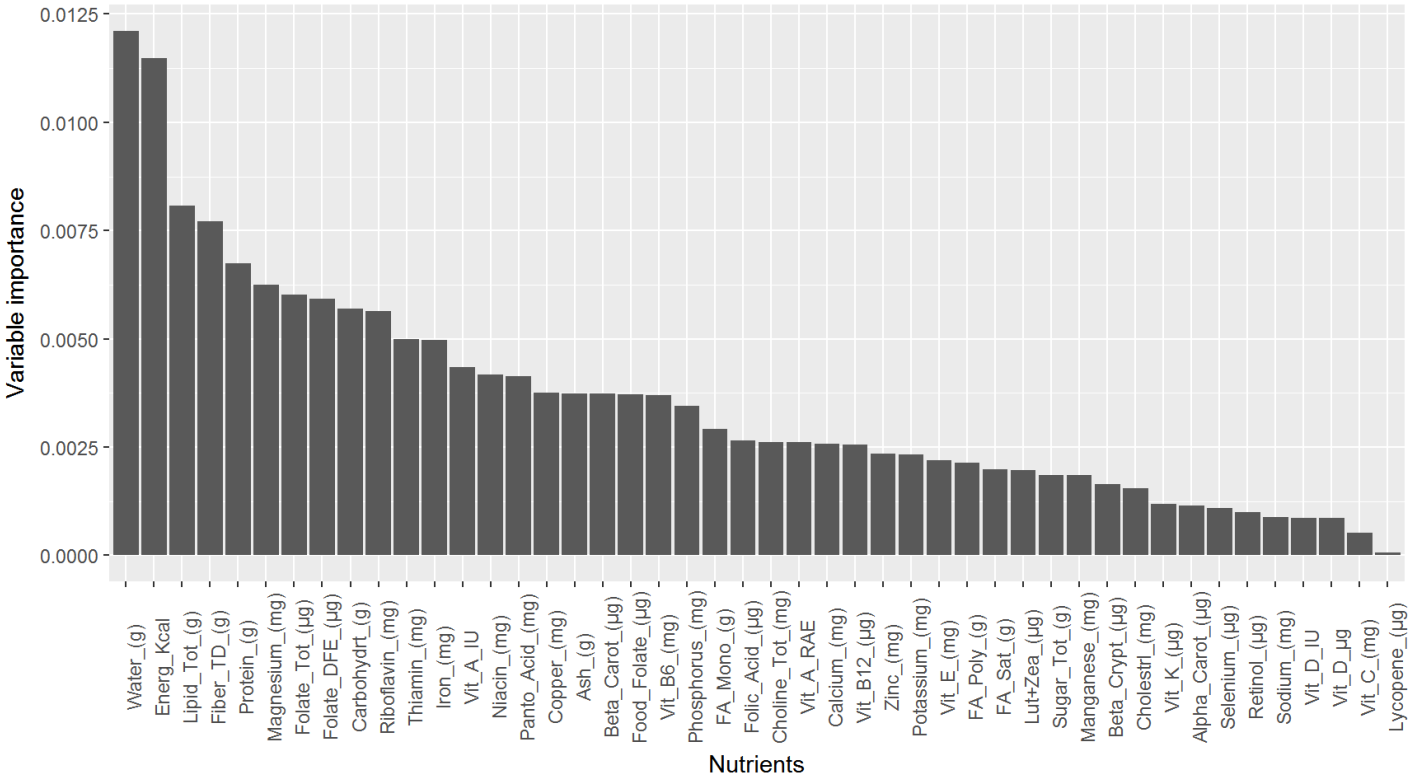

Figure 7: Contribution of nutrients on the first 7 prinicpal components

We can observe that the 6 most influential nutrients on the first 7 principal components are:

```
[1] "Water_(g)"      "Energ_Kcal"     "Lipid_Tot_(g)"  "Fiber_TD_(g)"
[5] "Protein_(g)"    "Magnesium_(mg)"
```

Biplot of selected nutrients

The biplot is created showing the directions of the 6 most influential nutrients as shown in Figure 8. Plotting all 46 nutrient directions will clutter the biplot with too much information. The selected 6 nutrients capture some aspects of scattered patterns shown in PC1 and PC2.

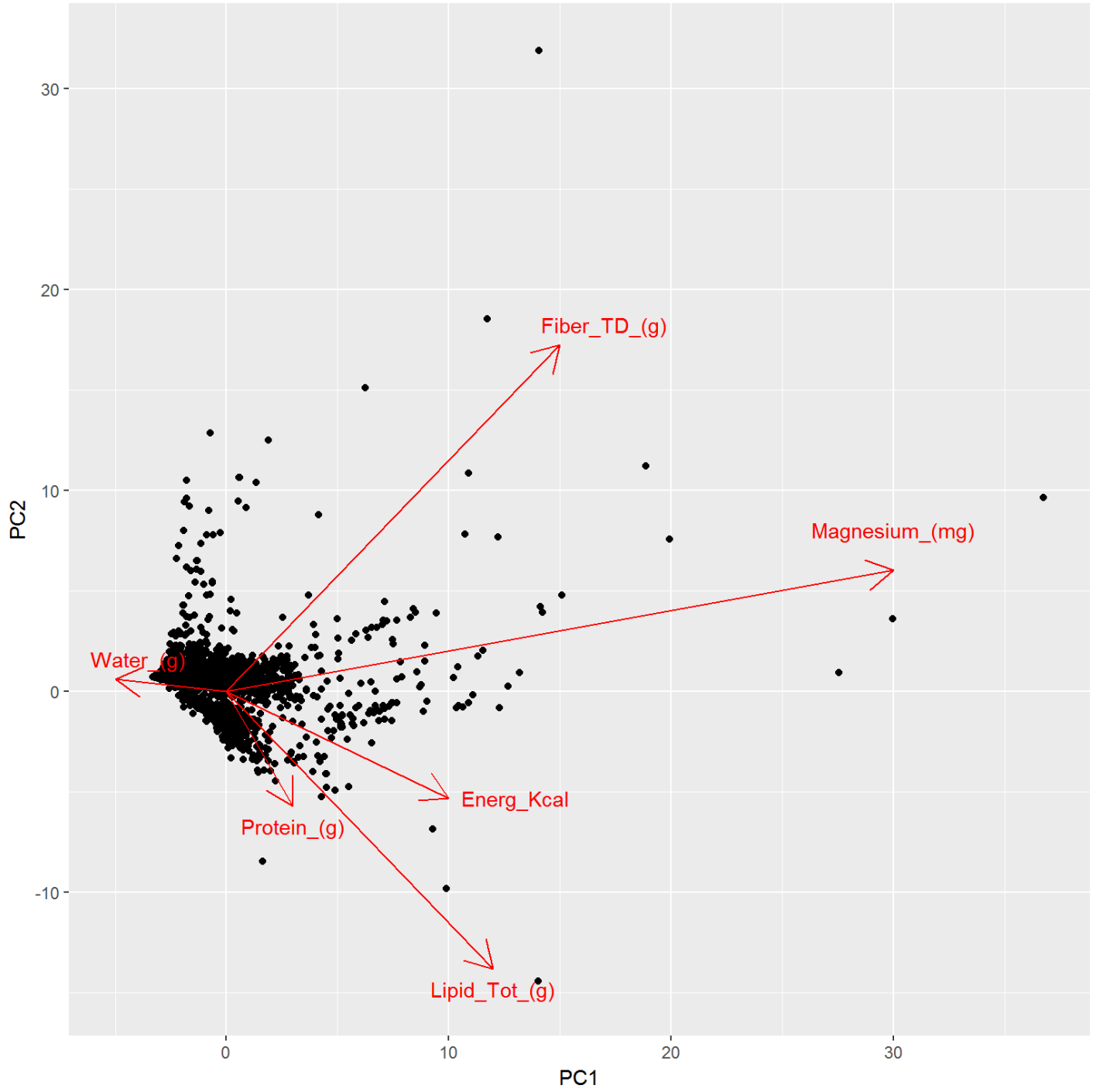

Figure 8: Biplot of 6 most significant nutrients impacting the first 7 principal components

Interactive plots

From the biplot in Figure 8, the water shows opposite direction in relation to other varaibles. Using the interactive plot provided in Figure 9, which shows the water content on the first three principal component axis, the direction of water content can be analyzed. It can be observed that foods with high water content are clustered near PC1 value of 0 and negative side whereas foods with lower water content are mostly found on positive values of PC1.

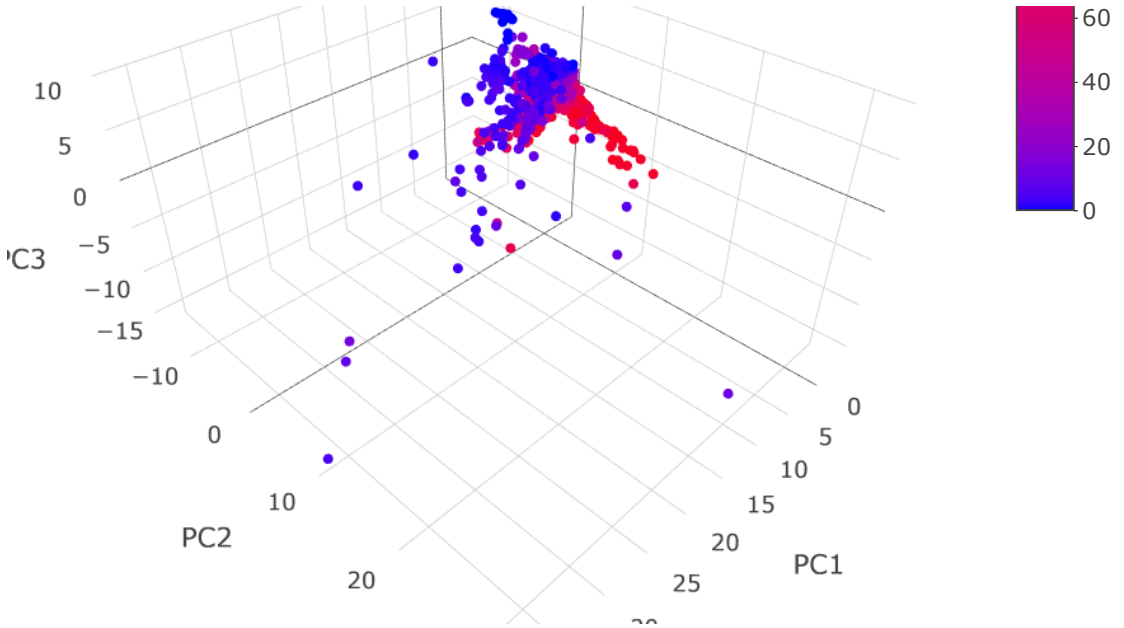

Figure 9: Interactive plot of water amount on the first three principal components

Since the 6 nutrients were selected based on their contributions to the first 7 principal components, their contribution may be better observed by choosing different principal components. For example, the direction of total lipid in Figure 8 is not so apparent in terms of explaing the pattern in PC1 and PC2. Looking at the heatmap in Figure 5, the total lipid may be more prominent in PC6 and PC7. With the interactive plot given in Figure 10, the direction of total lipid loading vector can be better observed when plotted on PC2, PC6 and PC7.

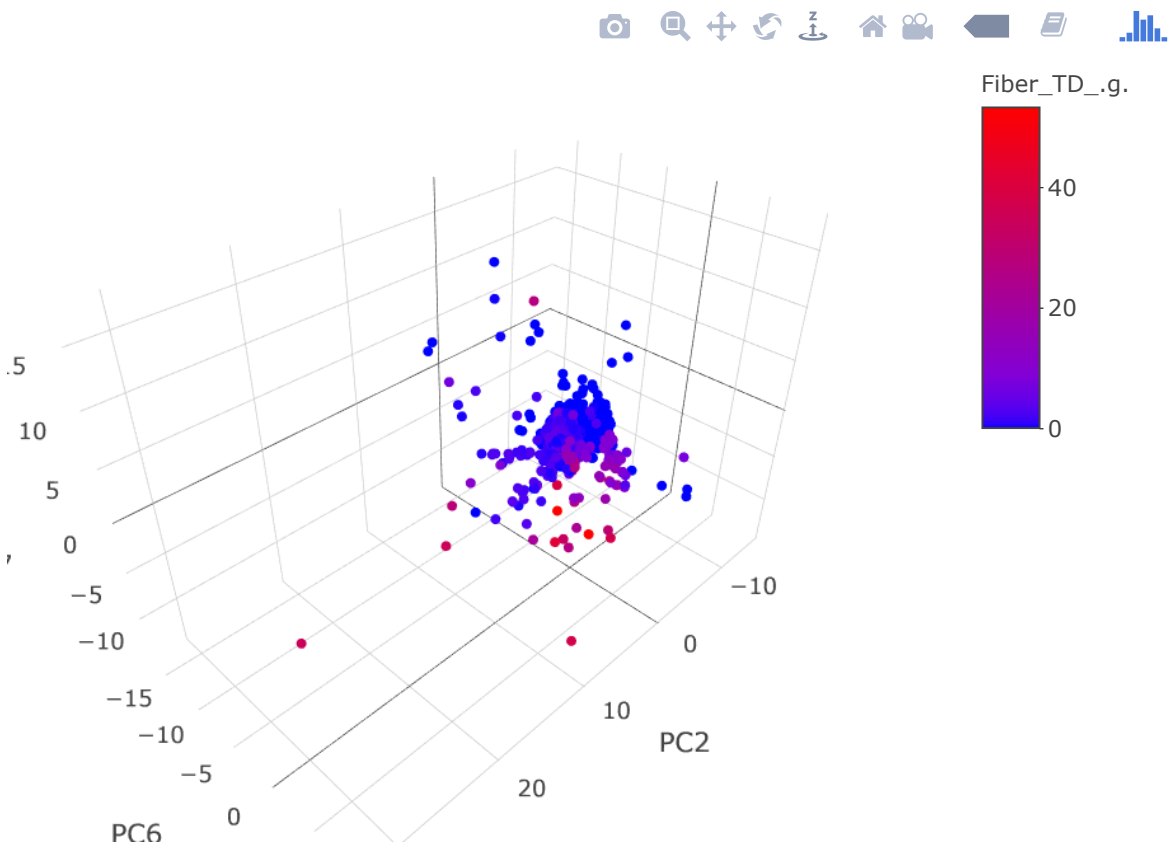

Figure 10: Interactive plot of total lipid amount on PC2, PC6 and PC7

Correlation among nutrients

Finally, a correlation matrix plot is given in Figure 11. Correlation between water and other nutrients are negative as noted by the red color in the figure. This agrees with observation seen in the heatmap in Figure 5 for PC1. Water is also strongly correlated with energy, total lipid, and carbohydrate.

There are highly positive correlations as well noted by the dark blue colored squares. One can observe cluster of postive correlation between Thiamin through Folate_DFE. This can also be observed in Figure 5 where the changes in loadings for first several principal components are similar, which is an indication of a strong correlation. Picking just one, Choline_Tot and Cholestrol are highly correlated in Figure 11. We can also verify this from Figure 5 where loadings for these two are very similar for most of the principal components.

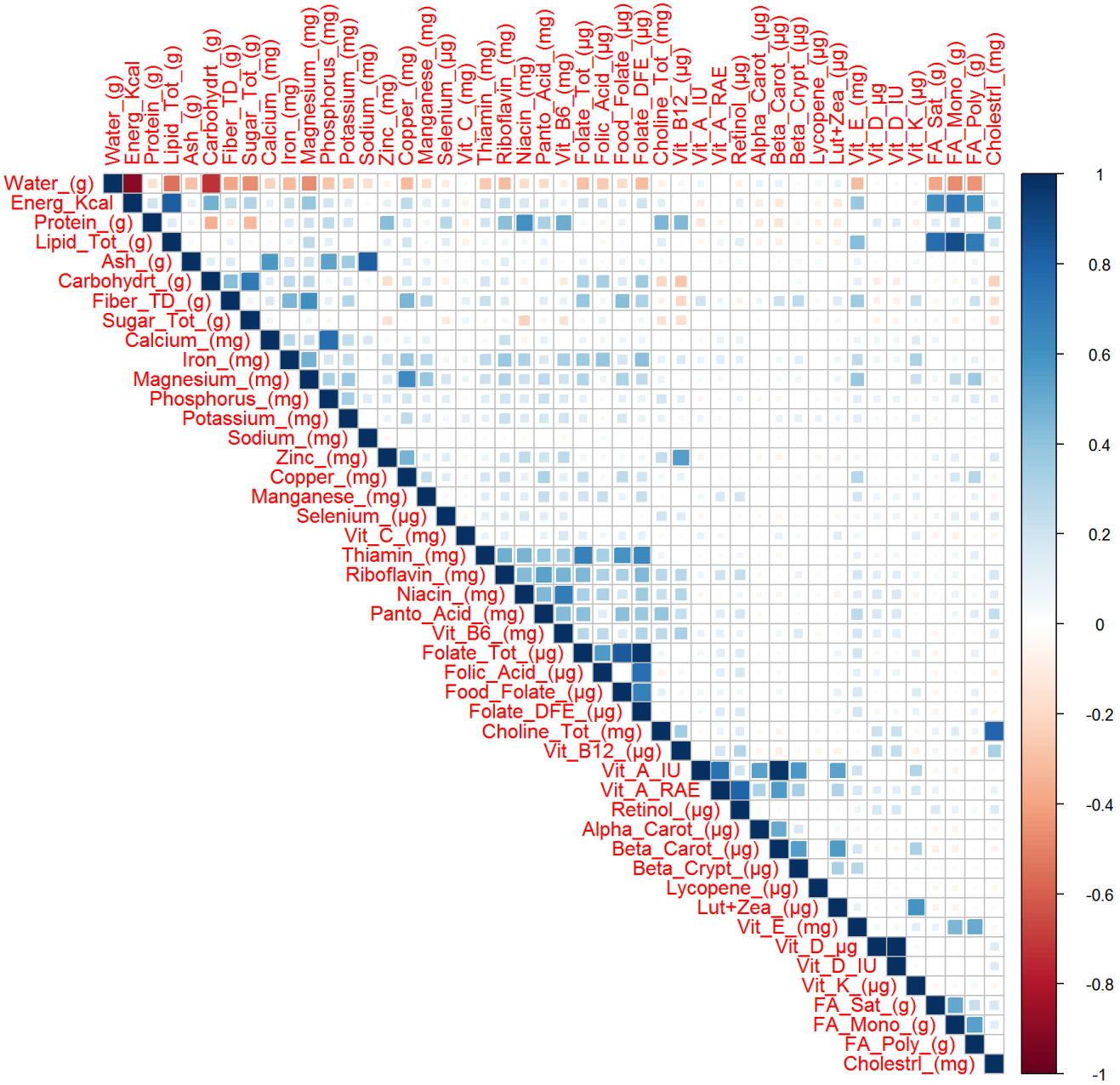

Figure 11: Correlation plot of the dataset

Grouping of similar nutrients

In Figure 5, the nutrients may be grouped into similar categories based on their influence on the prinicipal components. Nutrients with similar traits have similar influences on the principal components and can be categorized into a common group. With nutrient loading contribution to first 7 principal components and using the Minkowski distance with p=1, all nutrients with Minkowski distance equal or less than 0.4 have been identified:

```
Lipid_Tot_(g) : FA_Mono_(g)
Ash_(g) : Calcium_(mg) Sodium_(mg)
Carbohydrt_(g) : Sugar_Tot_(g)
Sugar_Tot_(g) : Carbohydrt_(g)
Calcium_(mg) : Ash_(g) Phosphorus_(mg)
Iron_(mg) : Manganese_(mg)
Magnesium_(mg) : Copper_(mg)
Phosphorus_(mg) : Calcium_(mg)
Sodium_(mg) : Ash_(g)
Zinc_(mg) : Selenium_(µg)
Copper_(mg) : Magnesium_(mg)
Manganese_(mg) : Iron_(mg)
Selenium_(µg) : Zinc_(mg)
Vit_C_(mg) : Lycopene_(µg)
Thiamin_(mg) : Folate_Tot_(µg) Food_Folate_(µg) Folate_DFE_(µg)
Riboflavin_(mg) : Panto_Acid_(mg)
Niacin_(mg) : Panto_Acid_(mg) Vit_B6_(mg)
Panto_Acid_(mg) : Riboflavin_(mg) Niacin_(mg) Vit_B6_(mg)
Vit_B6_(mg) : Niacin_(mg) Panto_Acid_(mg)
Folate_Tot_(µg) : Thiamin_(mg) Food_Folate_(µg) Folate_DFE_(µg)
Food_Folate_(µg) : Thiamin_(mg) Folate_Tot_(µg)
Folate_DFE_(µg) : Thiamin_(mg) Folate_Tot_(µg)
Choline_Tot_(mg) : Vit_B12_(µg) Cholestrl_(mg)
Vit_B12_(µg) : Choline_Tot_(mg) Cholestrl_(mg)
Vit_A_IU : Beta_Carot_(µg)
Alpha_Carot_(µg) : Beta_Crypt_(µg) Lut+Zea_(µg)
Beta_Carot_(µg) : Vit_A_IU Beta_Crypt_(µg) Lut+Zea_(µg)
Beta_Crypt_(µg) : Alpha_Carot_(µg) Beta_Carot_(µg) Lut+Zea_(µg)
Lycopene_(µg) : Vit_C_(mg)
Lut+Zea_(µg) : Alpha_Carot_(µg) Beta_Carot_(µg) Beta_Crypt_(µg)
Vit_E_(mg) : FA_Poly_(g)
Vit_D_µg : Vit_D_IU
Vit_D_IU : Vit_D_µg
FA_Sat_(g) : FA_Mono_(g)
FA_Mono_(g) : Lipid_Tot_(g) FA_Sat_(g) FA_Poly_(g)
FA_Poly_(g) : Vit_E_(mg) FA_Mono_(g)
```

One can compare the results above with Figure 5. For example, Lipid_Tot and FA_Mono has similar traits on the first 7 principal components. Similarly, iron and manganese have similar loadings on the principal components. These can be understood as the primary common nutrients since their common traits are within Minkowski distance of 0.4.

We can also evaluate secondary relationships. For example, Ash and Phosophorous have seconday common traits that is not as strong as the primary since the relationshp is found through Ash-Calcium-Phosophorous. The relationship is farther than Minkowski distance of 0.4. In Figure 5, Ash and Phosophorous share similar traits but are not as similar as those found in primary relationships.

CONCLUSIONS

Based on the USDA National Nutrient Database for Standard Reference (SR) Release 28, dataset with 2223 food items with 46 nutrient information has been analyzed. Following conclusions may be drawn from this study:

- The distributions of nutrients are highly right-skewed with outliers. This indicates that there are specific food items that provide very high specific nutrient content.
- The scree plot from PCA suggests that the cutoff number of principal component is 7 where cumulative variance explained is at 56.5%.
- It was also noted that the first principal component had interesting pattern where loadings between water and the rest of nutrients were opposite.
- The nutrients, “Water_(g)”, “Energ_Kcal”, “Lipid_Tot_(g)”, “Fiber_TD_(g)”, “Protein_(g)”, and “Magnesium_(mg)” were the 6 most significant contributors of the first 7 principal components.
- Based on the nutrient impact on the first 7 principal components, the nutrients have been grouped as having common traits where primary and secondary relationships among nutrients can be identified (see section “Grouping of similar nutrients for details”).

APPENDIX

All R codes used in producing the results are included below:

```
##############################
### Initial Setup

knitr::opts_chunk$set(comment=NA, echo=FALSE, warning=FALSE, message=FALSE,
                      fig.align="center")
options(digits=4)
rm(list=ls())

###
### Data cleaning
###

SR = read.table("ABBREV.txt", header=F, row.names=1, sep="^", quote="~")
SR = na.omit(SR) # remove rows with missing values
SR = SR[row.names(SR) != "13352",] # remove "duplicate" entry
row.names(SR) = SR[,1] # set more meaningful row names
SR = SR[,-1]

names(SR) = c("Water_(g)", "Energ_Kcal", "Protein_(g)", "Lipid_Tot_(g)", "Ash_(g)", "Carbohydrt_(g)",
 "Fiber_TD_(g)", "Sugar_Tot_(g)", "Calcium_(mg)", "Iron_(mg)", "Magnesium_(mg)", "Phosphorus_(mg)",
"Potassium_(mg)", "Sodium_(mg)", "Zinc_(mg)", "Copper_(mg)", "Manganese_(mg)", "Selenium_(µg)", "Vit_
C_(mg)", "Thiamin_(mg)", "Riboflavin_(mg)", "Niacin_(mg)", "Panto_Acid_(mg)", "Vit_B6_(mg)", "Folate_
Tot_(µg)", "Folic_Acid_(µg)", "Food_Folate_(µg)", "Folate_DFE_(µg)", "Choline_Tot_(mg)", "Vit_B12_(µ
g)", "Vit_A_IU", "Vit_A_RAE", "Retinol_(µg)", "Alpha_Carot_(µg)", "Beta_Carot_(µg)", "Beta_Crypt_(µ
g)", "Lycopene_(µg)", "Lut+Zea_(µg)", "Vit_E_(mg)", "Vit_D_µg", "Vit_D_IU", "Vit_K_(µg)", "FA_Sat_
(g)", "FA_Mono_(g)", "FA_Poly_(g)", "Cholestrl_(mg)", "GmWt_1", "GmWt_Desc1", "GmWt_2", "GmWt_Desc2",
 "Refuse_Pct")

SRp = SR[,c(1:46)] # restrict to just the nutrient variables

str(SRp)

##########################
### EDA
##########################

###
### Proximates
###

library(reshape2)
library(ggplot2)

proximates = subset(SRp, select=c("Water_(g)","Protein_(g)","Lipid_Tot_(g)",
                                  "Carbohydrt_(g)","Ash_(g)"))

melt_proximates = melt(proximates)

# Figure 1: Boxplots of proximates in all foods

ggplot(data=melt_proximates) +
   geom_boxplot(aes(x=variable, y=value)) +
   facet_wrap(~variable, scale="free", ncol=5) +
   labs(x="Proximates", y="Values")

rownames(SRp)[order(-SRp$`Protein_(g)`)[1:5]]
```

```
###
### Minerals
###

minerals = subset(SRp, select=c("Calcium_(mg)", "Iron_(mg)", "Magnesium_(mg)", "Phosphorus_(mg)", "Po
tassium_(mg)", "Sodium_(mg)", "Zinc_(mg)", "Copper_(mg)", "Manganese_(mg)", "Selenium_(µg)"))

melt_minerals = melt(minerals)

# Figure 2: Bolxplots of minerals in all foods

ggplot(data=melt_minerals) +
   geom_boxplot(aes(x=variable, y=value)) +
   facet_wrap(~variable, scale="free", ncol=5) +
   labs(x="Minerals", y="Values")

rownames(SRp)[order(-SRp$`Calcium_(mg)`)[1:5]]

###
### Vitamins and other nutrients
###

vitamin = subset(SRp, select=c("Vit_A_RAE", "Vit_C_(mg)", "Vit_D_IU", "Vit_E_(mg)", "Folate_Tot_(µg)"
))

melt_vitamin = melt(vitamin)

# Figure 3: Bolxplots of vitamins and other nutrients in all foods

ggplot(data=melt_vitamin) +
   geom_boxplot(aes(x=variable, y=value)) +
   facet_wrap(~variable, scale="free", ncol=5) +
   labs(x="Vitamins", y="Values")

##########################
### PCA
##########################

pr.out = prcomp(SRp, center=TRUE, scale=TRUE)  # With center=TRUE, scale=TRUE
pr.out$x = -pr.out$x
pr.out$rotation = -pr.out$rotation

pve = pr.out$sdev^2 / length(pr.out$sdev)  # Proportion of variance explained

library(ggplot2)
library(gridExtra)

# Figure 4: Proportion of variance explained and their cumulative sum from PCA

p1 = ggplot() +
   geom_line(aes(x=c(1:length(pr.out$sdev)), y=pve)) +
   geom_point(aes(x=c(1:length(pr.out$sdev)), y=pve), size=3) +
   geom_hline(yintercept=pve[7], color='red') +
   labs(x="Principal component", y="Proportion of variance explained")

p2 = ggplot() +
```

```
   geom_line(aes(x=c(1:length(pr.out$sdev)), y=cumsum(pve))) +
   geom_point(aes(x=c(1:length(pr.out$sdev)), y=cumsum(pve)), size=3) +
   geom_hline(yintercept=cumsum(pve)[7], color='red') +
   annotate("text", x=20, y=0.59, label=paste(round(cumsum(pve)[7],3))) +
   labs(x="Principal component", y="Cumulative proportion of variance explained")

grid.arrange(p1, p2, ncol=2)


###
### Heatmap
###

melt_pr.out = melt(pr.out$rotation[,c(1:20)])
colnames(melt_pr.out) = c("Nutrient", "PC", "Value")

# Figure 5: Heatmap of nutrients and first 20 principal components

ggplot(data=melt_pr.out) +
   geom_tile(aes(x=PC, y=Nutrient, fill=Value)) +
   scale_fill_gradient2(low='blue', mid='white', high='red', midpoint=0) +
   labs(x="Principal component")

library(reshape2)

melt_pr.out1 = melt(pr.out$rotation[,c(1:2)])
colnames(melt_pr.out1) = c("Nutrient", "PC", "Value")

# Figure 6: Full spectrum of the nutrient loadings in PC1 and PC2

ggplot(data=melt_pr.out1) +
   geom_col(aes(x=Nutrient, y=Value)) +
   facet_wrap(~PC, ncol=1) +
   theme(axis.text.x = element_text(angle = 90)) +
   labs(x="Nutrients", y="Loading")

#
# Sanity check
#
pve.mat = matrix(rep(pve, each = 46), nrow=length(pve))

nutrient.impact = apply(pr.out$rotation[,c(1:46)]^2 * pve.mat, 1, sum)
sum(nutrient.impact)  # Should equal 1
###
### Selection of nutrients
###

pve.mat = matrix(rep(pve, each = 7), nrow=length(pve))
nutrient.impact = apply(pr.out$rotation[,c(1:7)]^2 * pve.mat, 1, sum)
melt_NI = melt(nutrient.impact)

# Figure 7: Contribution of nutrients on the first 7 prinicpal components

ggplot(data=melt_NI) +
   geom_col(aes(x=reorder(rownames(melt_NI), -value), y=value)) +
   theme(axis.text.x = element_text(angle = 90)) +
   labs(x="Nutrients", y="Variable importance")
```

```
top6 = rownames(melt_NI)[order(-melt_NI$value)[1:6]]
top6

###
### Biplot
###

slopes = rep(NA, 6)

for(s in c(1:6)) {
   slopes[s] = pr.out$rotation[top6[s],2] / pr.out$rotation[top6[s],1]
}

# Figure 8: Biplot of 6 most significant nutrients impacting the first 6 principal components

ggplot() +
   geom_point(aes(x=pr.out$x[,1], y=pr.out$x[,2])) +
   geom_segment(aes(x=0, y=0, xend=-5, yend=slopes[1]*(-5)),
                arrow = arrow(length = unit(0.03, "npc")), color="red") +
   annotate("text", x=-4, y=slopes[1]*(-5)+1, label=top6[1], color="red") +
   geom_segment(aes(x=0, y=0, xend=10, yend=slopes[2]*10),
                arrow = arrow(length = unit(0.03, "npc")), color="red") +
   annotate("text", x=13, y=slopes[2]*10, label=top6[2], color="red") +
   geom_segment(aes(x=0, y=0, xend=12, yend=slopes[3]*12),
                arrow = arrow(length = unit(0.03, "npc")), color="red") +
   annotate("text", x=12, y=slopes[3]*12-1, label=top6[3], color="red") +
   geom_segment(aes(x=0, y=0, xend=15, yend=slopes[4]*15),
                arrow = arrow(length = unit(0.03, "npc")), color="red") +
   annotate("text", x=17, y=slopes[4]*15+1, label=top6[4], color="red") +
   geom_segment(aes(x=0, y=0, xend=3, yend=slopes[5]*3),
                arrow = arrow(length = unit(0.03, "npc")), color="red") +
   annotate("text", x=3, y=slopes[5]*3-1, label=top6[5], color="red") +
   geom_segment(aes(x=0, y=0, xend=30, yend=slopes[6]*30),
                arrow = arrow(length = unit(0.03, "npc")), color="red") +
   annotate("text", x=30, y=slopes[6]*30+2, label=top6[6], color="red") +
   labs(x="PC1", y="PC2")

###
### Interactive plots
###

# Figure 9: Interactive plot of water amount on the first three principal components

library(plotly)

pr.out_df = data.frame(cbind(pr.out$x[,c(1:7)], SRp))

plot_ly(pr.out_df, x =~PC1, y =~PC2, z =~PC3, color = ~Water_.g.,
        type = "scatter3d", mode = "markers", colors=c("blue", "red"),
        marker=list(size=3))

# Figure 10: Interactive plot of total lipid amount on PC2, PC6 and PC7

plot_ly(pr.out_df, x =~PC2, y =~PC6, z =~PC7, color = ~Fiber_TD_.g.,
        type = "scatter3d", mode = "markers", colors=c("blue", "red"),
        marker=list(size=3))
```

```
###
### Correlation
###

# Figure 11: Correlation plot of the dataset

library(corrplot)
SRp.cor = cor(SRp)
corrplot(SRp.cor, method="square", type="upper")

###
### Grouping of similar nutrients
###

sim.crit = 0.4
similar = list()

for(j in c(1:45)) {
   nutj = rownames(pr.out$rotation)[j]

   for(i in c(1:46)) {
      nuti = rownames(pr.out$rotation)[i]
      dist.scores = sum(abs(pr.out$rotation[j,c(1:7)] - pr.out$rotation[i,c(1:7)]))

      if ((dist.scores <= sim.crit) && (dist.scores != 0)) {
         similar[[nutj]] = append(similar[[nutj]], nuti)
      }
   }
}

for(n in c(1:length(similar))) {
   cat(names(similar[n]),":", similar[[n]],"\n")
}
```