Math 189 Homework 7

Hien Bui, Cole Clisby, Cheolmin Hwang, Andrew Li, James Yu

Intro

In this report, we analyze a dataset containing information on women’s nutrient intake levels. Our goal is to determine the relationship between the intake of different nutrient variables and to apply dimensionality reduction and factor analysis in order to determine the latent factors behind the major nutritional features.

Body

Data

```
nutrient <- read.table('nutrient.txt', header=FALSE,
                       col.names=c('index', 'Calcium', 'Iron', 'Protein', 'VitaminA', 'V
itaminC'))
nutrient <- subset(nutrient, select=c('Calcium', 'Iron', 'Protein', 'VitaminA', 'Vitamin
C'))
head(nutrient)
```

```
##   Calcium   Iron Protein VitaminA VitaminC
## 1  522.29 10.188  42.561   349.13   54.141
## 2  343.32  4.113  67.793   266.99   24.839
## 3  858.26 13.741  59.933   667.90  155.455
## 4  575.98 13.245  42.215   792.23  224.688
## 5 1927.50 18.919 111.316   740.27   80.961
## 6  607.58  6.800  45.785   165.68   13.050
```

Source: Survey conducted by the United States Department of Agriculture (USDA) in 1985

URL: https://github.com/tuckermcelroy/ma189/blob/main/Data/nutrient.txt (https://github.com/tuckermcelroy/ma189/blob/main/Data/nutrient.txt)

Description: Contains 737 observations of nutrient levels from a survey about women in the United States aged 25-50 years old.

Variables:

1. Calcium (in milligrams)
2. Iron (in milligrams)
3. Protein (in grams)
4. Vitamin A (in micrograms)
5. Vitamin C (in milligrams)

Task 1

```
# level plot
library(lattice)
library(ellipse)
```

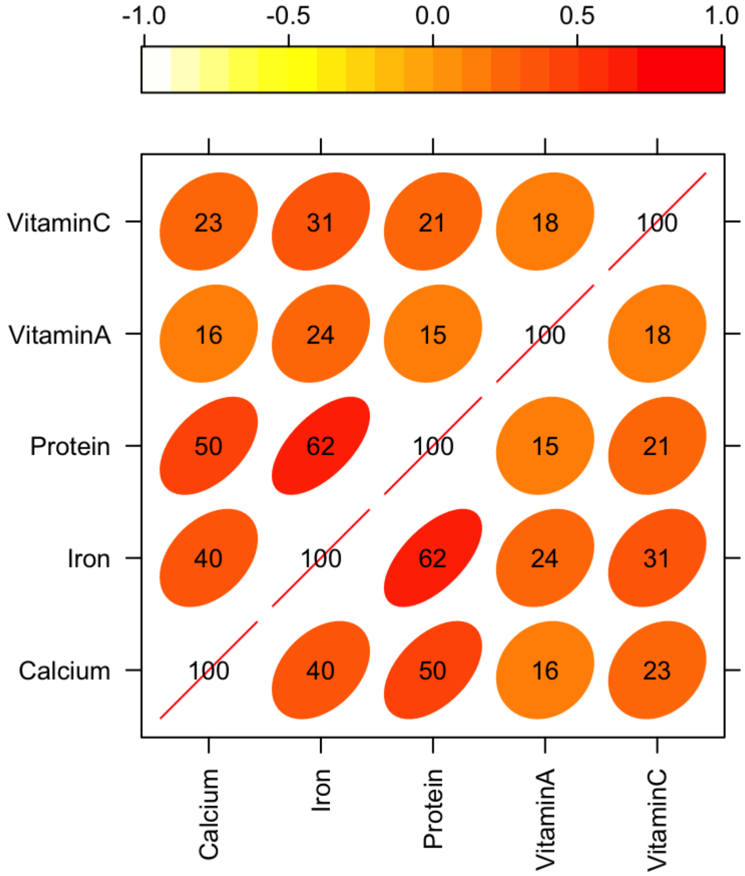
```
##
## Attaching package: 'ellipse'
```

```
## The following object is masked from 'package:graphics':
##
##     pairs
```

```
cor <- cor(nutrient)

panel.corrgram <- function(x, y, z, subscripts, at, level=0.9, label=FALSE, ...) {
  require("ellipse", quietly=TRUE)
  x <- as.numeric(x)[subscripts]
  y <- as.numeric(y)[subscripts]
  z <- as.numeric(z)[subscripts]
  zcol <- level.colors(z, at=at, ...)
  for (i in seq(along = z)) {
    ell = ellipse(z[i], level=level, npoints=50, scale=c(0.2,0.2), centre=c(x[i], y[i]))
    panel.polygon(ell, col=zcol[i], border=zcol[i], ...)
  }
  if (label)
    panel.text(x=x, y=y, lab=100*round(z,2), cex=0.8, col=ifelse(z<0, 'white', 'black'))
}

print(levelplot(cor[seq(1,5), seq(1,5)], at=do.breaks(c(-1.01,1.01), 20),
                xlab=NULL, ylab=NULL, colorkey=list(space='top'), col.regions=rev(heat.c
olors(100)),
                scales=list(x=list(rot=90)), panel=panel.corrgram, label=TRUE))
```



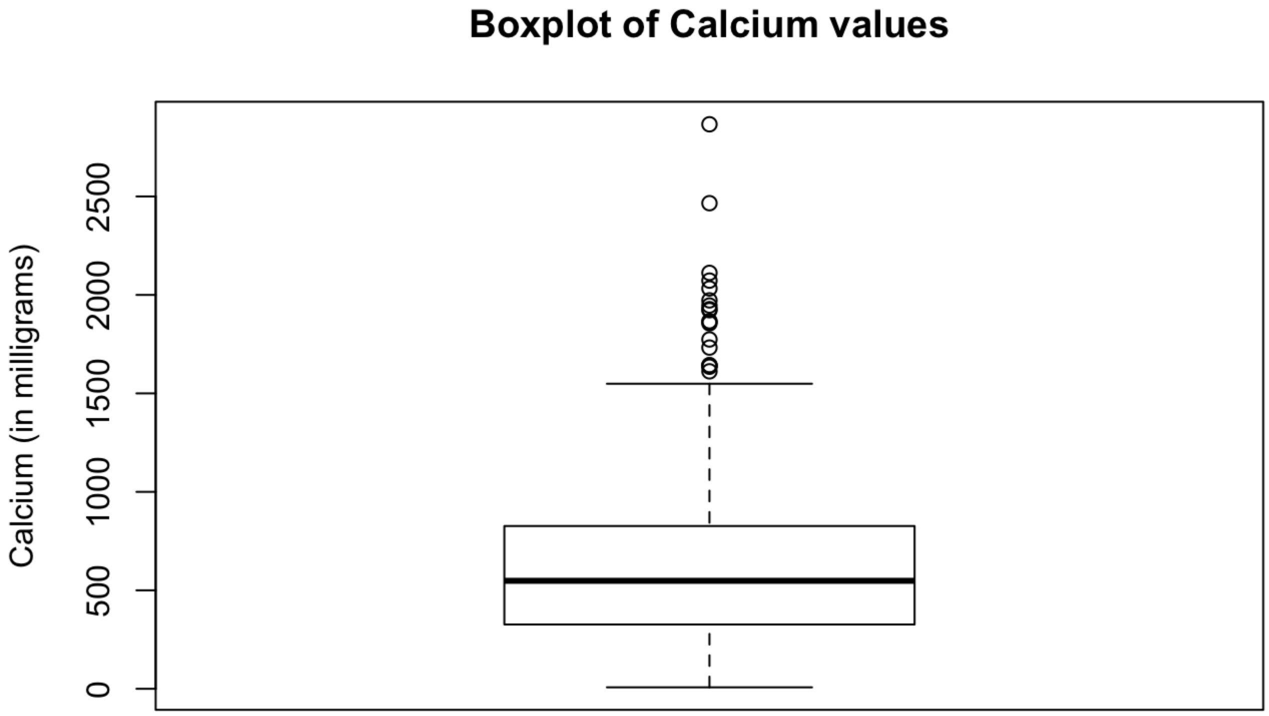
The level plot shows us the correlation between variables through the use of the hue and shape, darker colors and narrower ovals indicating stronger correlations between the corresponding two variables. By looking at the level plot it appears that Protein and Iron are strongly correlated, along with Protein and Calcium being somewhat correlated. The rest of the variables have much weaker correlations with each other.

Task 2

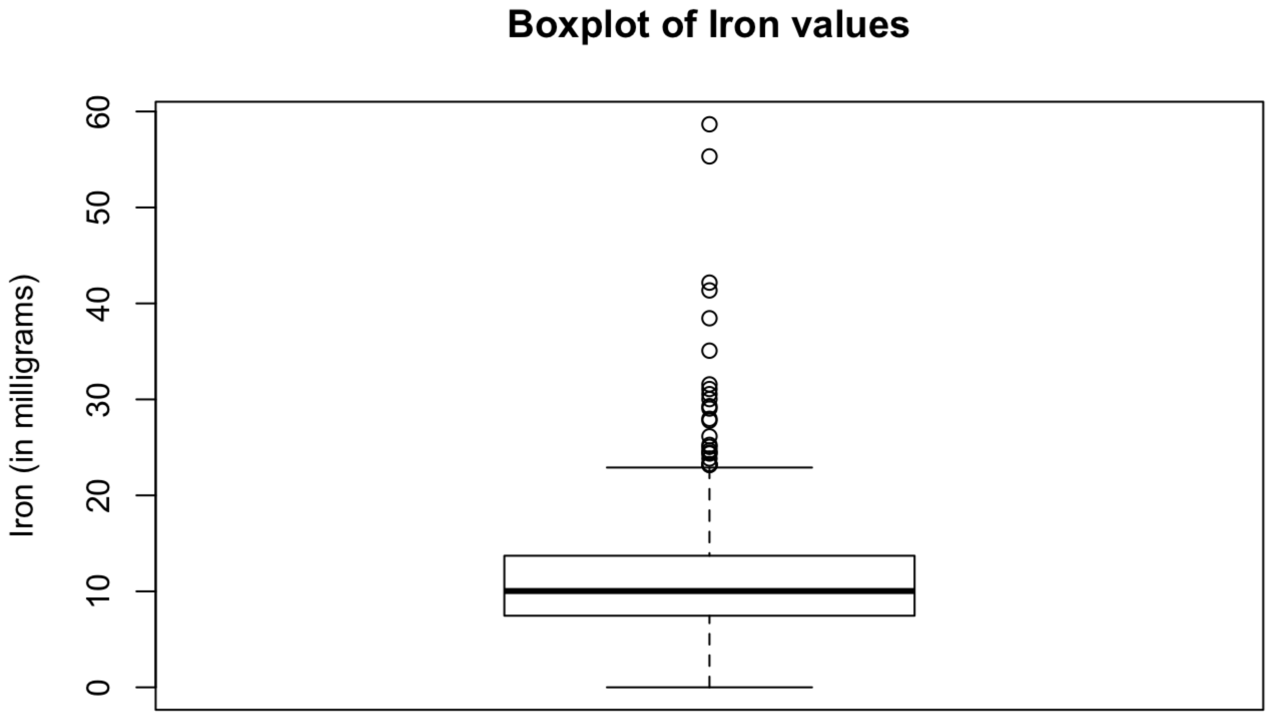
PCA

Underlying Assumptions for PCA: The underlying assumptions for PCA are that the variables are continuous and there are no significant outliers. It also assumes that there is a linear relationship between all variables, our dataset is large enough, and our variables are adequately correlated.

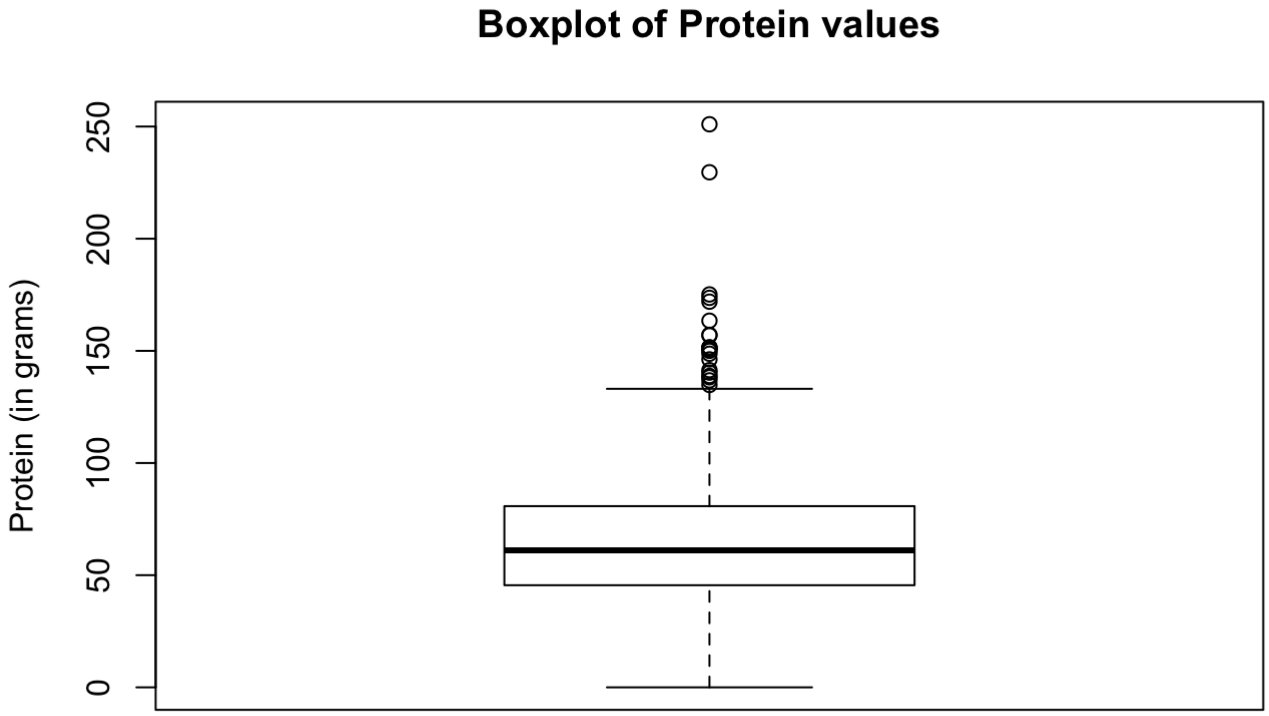
```
boxplot(nutrient$Calcium, data=nutrient, ylab='Calcium (in milligrams)')
title('Boxplot of Calcium values')
```



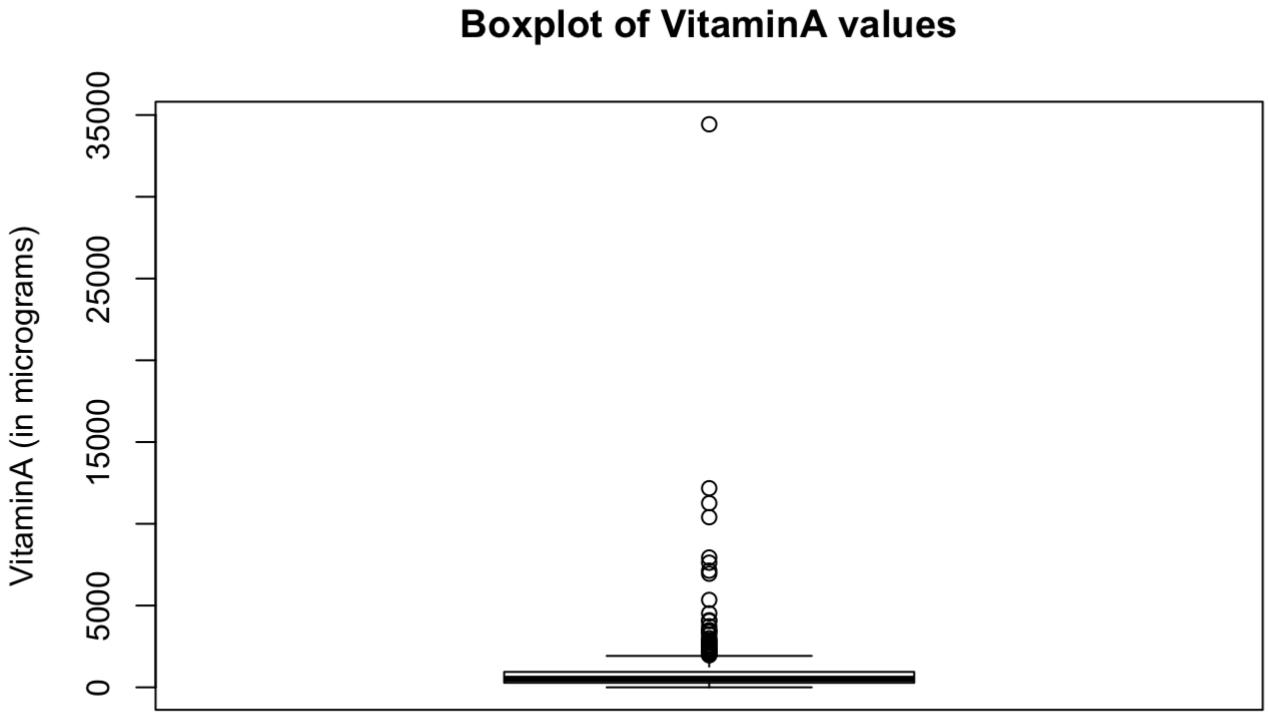
```
boxplot(nutrient$Iron, data=nutrient, ylab='Iron (in milligrams)')
title('Boxplot of Iron values')
```



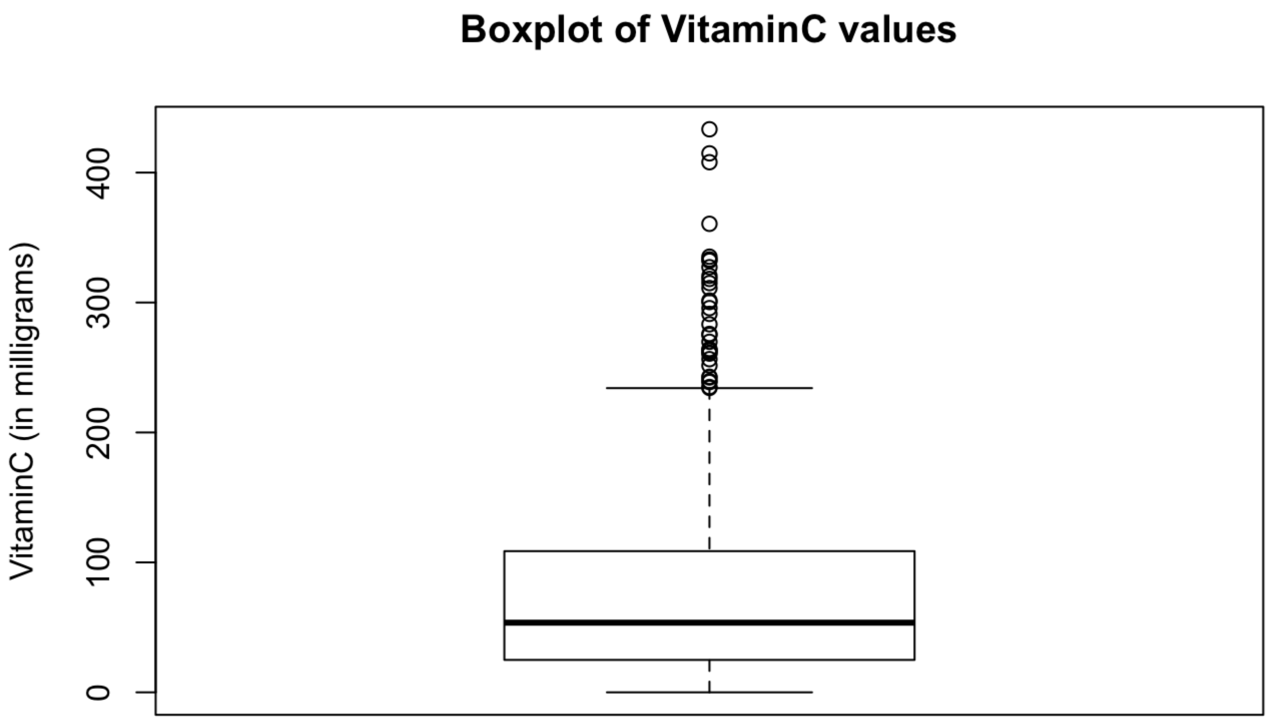
```
boxplot(nutrient$Protein, data=nutrient, ylab='Protein (in grams)')
title('Boxplot of Protein values')
```



```
boxplot(nutrient$VitaminA, data=nutrient, ylab='VitaminA (in micrograms)')
title('Boxplot of VitaminA values')
```



```
boxplot(nutrient$VitaminC, data=nutrient, ylab='VitaminC (in milligrams)')
title('Boxplot of VitaminC values')
```



However, when we create boxplots for each of the variables, we can see that the variables have a large number of outliers. The underlying assumptions for PCA are relatively weak, and we can still fit the factor model regardless.

```
# fit factor model using pca
pca_result <- prcomp(nutrient, scale=TRUE, rank = 2)
pca_result
```

```
## Standard deviations (1, .., p=5):
## [1] 1.5103824 0.9766802 0.8964675 0.7863928 0.5854685
##
## Rotation (n x k) = (5 x 2):
##                PC1         PC2
## Calcium  0.4725630 -0.26586441
## Iron     0.5431481 -0.09248598
## Protein  0.5370491 -0.34767498
## VitaminA 0.2724785  0.78259867
## VitaminC 0.3449756  0.43292481
```

```
pca_var <- pca_result$sdev^2

pve <- pca_var / sum(pca_var)
out2 <- cbind(pca_var, pve, cumsum(pve))
colnames(out2) <- c('Eigenvalue', 'Proportion', 'Cumulative')
#rownames(out2) <- c('PC1', 'PC2')

t(pca_result$rotation)
```

```
##        Calcium        Iron    Protein  VitaminA  VitaminC
## PC1  0.4725630  0.54314812  0.5370491 0.2724785 0.3449756
## PC2 -0.2658644 -0.09248598 -0.3476750 0.7825987 0.4329248
```

The first loading vector puts relatively equal weight on Calcium, Iron, and Protein, whereas it puts less weight on Vitamin A and Vitamin C.

The third loading vector also puts relatively equal weight on Calcium, Iron, and Protein, but does not follow the traces of the first loading vector as it puts huge significance on Vitamin A, and the least weight on Vitamin C.

MLE

Assumptions for MLE: We assume that the data are independently sampled from a multivariate normal distribution

```
# factor model using MLE
n_factors <- 2
fa_fit <- factanal(nutrient, n_factors, rotation='varimax')
loading <- fa_fit$loadings[,1:2]
t(loading)
```

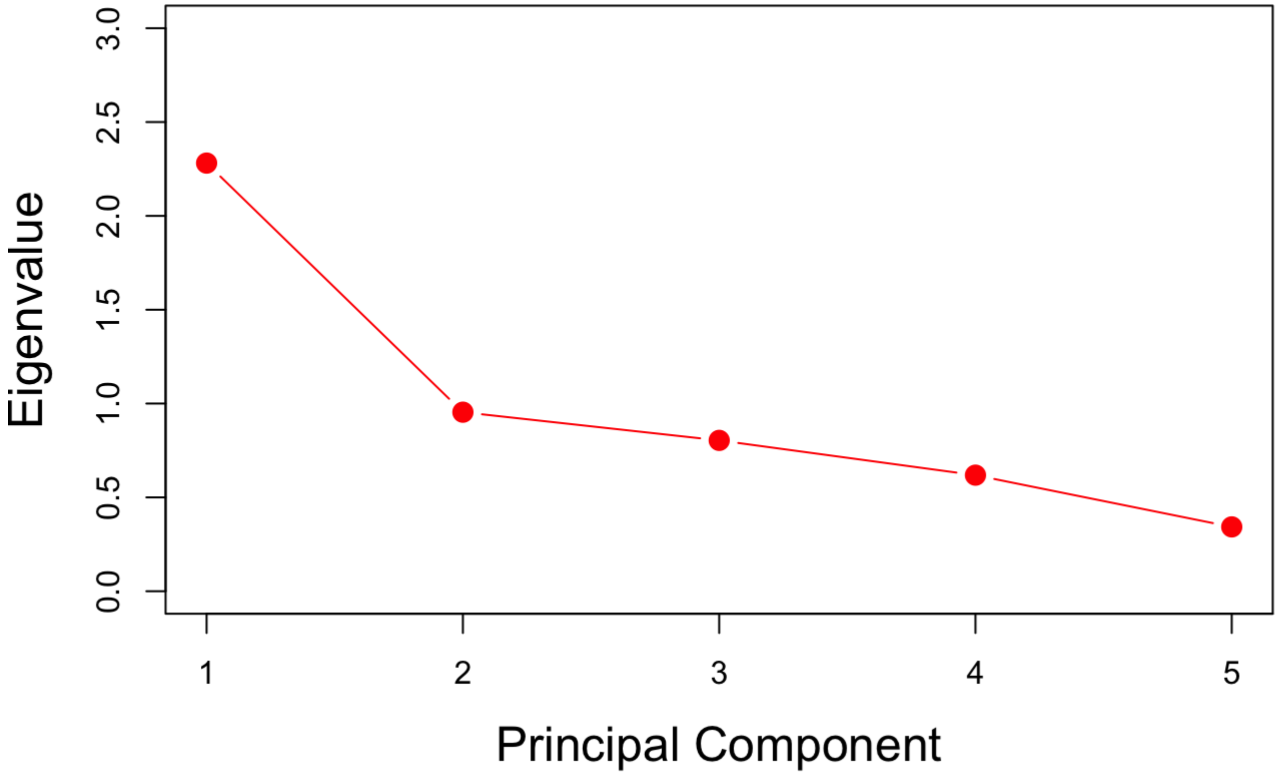
```
##           Calcium      Iron   Protein   VitaminA  VitaminC
## Factor1 0.4662298 0.5675046 0.9888518 0.09839555 0.1510572
## Factor2 0.2984038 0.4743206 0.1310440 0.37773096 0.4787664
```

We can observe that Vitamin C, Iron, and Vitamin A are weighted heavily by Factor 2, and that Protein is weighted lightly in Factor 2 . In contrast in Factor 1, Protein shows a weight close to 1, but Vitamin C and A show a very low weight.

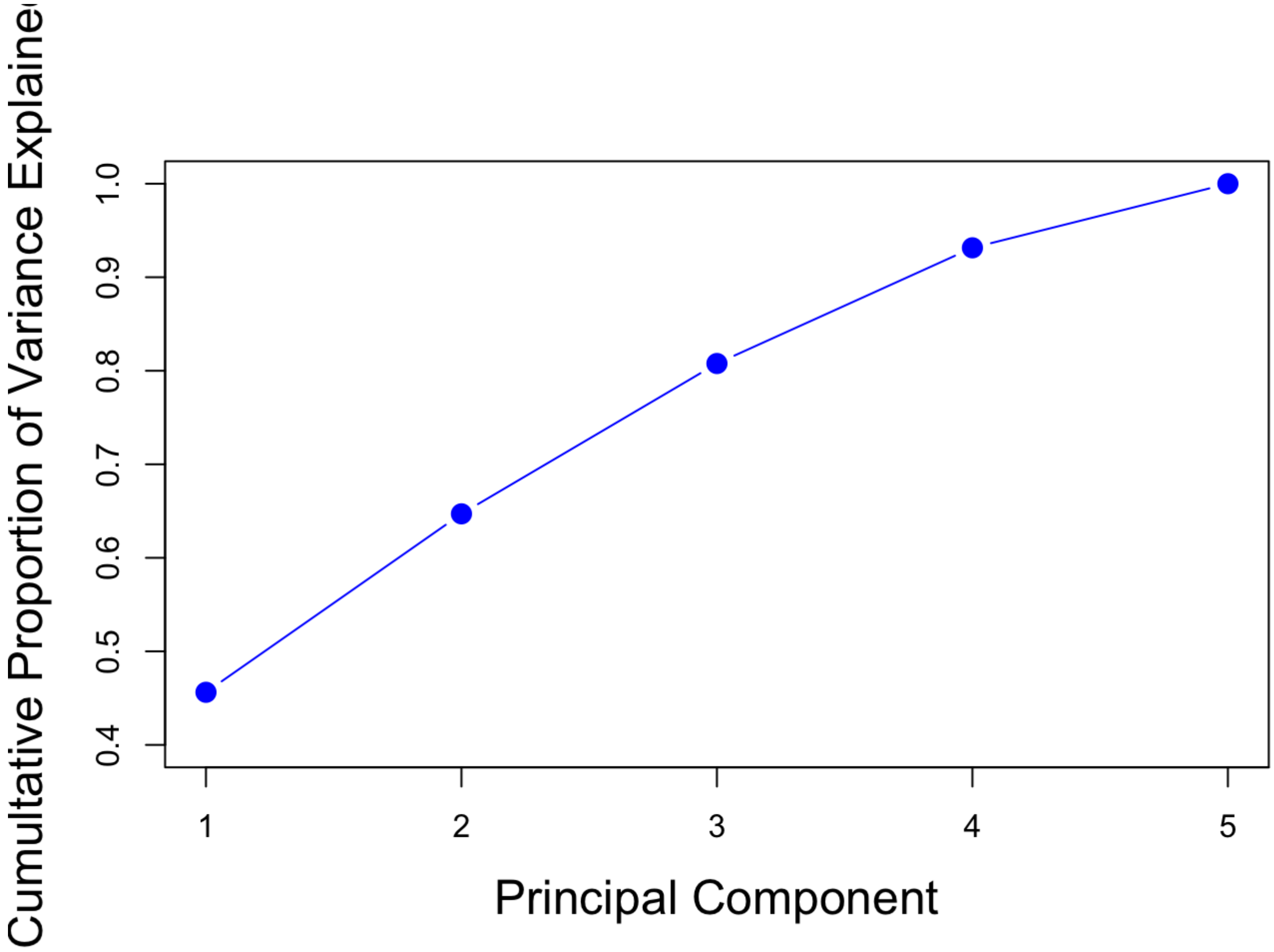
We prefer the PCA method because it presents to us a clear picture of which variables are affecting the outcome. It also outputs a clearer indication of grouping of variables, showing which variables are evenly weighted on each principle component.

Task 3

```
# scree plot
plot(pca_var, xlab='Principal Component', ylab='Eigenvalue', ylim=c(0,3), xaxt='n', type
='b', col='red', cex=2,
     pch=20, cex.lab=1.5)
axis(1, at=c(1,2,3,4,5,6), labels=c(1,2,3,4,5,6))
```



```
# cumultative proportion plot
plot(cumsum(pve), xlab='Principal Component', ylab='Cumultative Proportion of Variance E
xplained', ylim=c(0.4,1), xaxt='n',
     type='b', col='blue', cex=2, pch=20, cex.lab=1.5)
axis(1, at=c(1,2,3,4,5,6), labels=c(1,2,3,4,5,6))
```



We will reduce the dimensions from 5 to 2. We chose to use the first two principle components since together they can explain around 65% of the variance, and generating any more factors would provide us little information than what we would have at 2 factors.

Task 4

```
t(loading)
```

```
##           Calcium      Iron   Protein   VitaminA  VitaminC
## Factor1 0.4662298 0.5675046 0.9888518 0.09839555 0.1510572
## Factor2 0.2984038 0.4743206 0.1310440 0.37773096 0.4787664
```

```
pca_result <- prcomp(nutrient, scale=TRUE, rank = 2)
t(pca_result$rotation)
```

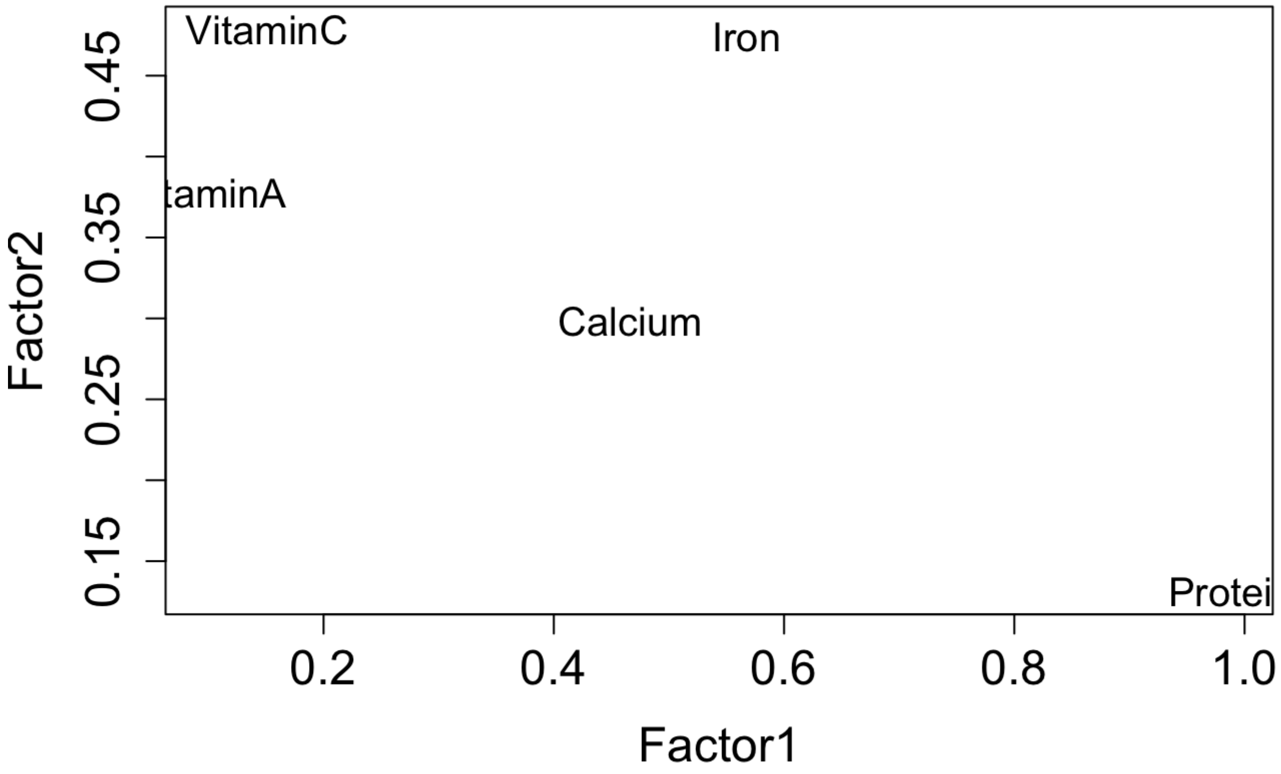
```
##        Calcium        Iron    Protein  VitaminA  VitaminC
## PC1  0.4725630  0.54314812  0.5370491 0.2724785 0.3449756
## PC2 -0.2658644 -0.09248598 -0.3476750 0.7825987 0.4329248
```

MLE: It looks like Calcium, Iron, and Protein have high loadings for Factor 1 and VitaminA, VitaminC have higher loadings for Factor 2. This demonstrates how Factor 1 is more associated with the macronutrients (Calcium, Iron, Protein) and Factor 2 is more associated with the micronutrients (VitaminA, VitaminC)

PCA: From the results when we reduce the dimensionality to 2, we can see similar results with MLE in that PC1 is more associated with the macronutrients of calcium, iron, and protein. However, compared to MLE the distrubtion of the nutrients is more even for the first factor. For PC2, the micronnutrients of Vitamin A and Vitamin C have more association than the macronutrients, similar to our results from MLE. Another important note is that Iron is extremely low for PC2.

Task 5

```
plot(loading, type='n', cex.axis=1.5, cex.lab=1.5)
text(loading, labels=names(nutrient), cex=1.25)
```



Shows that Factor 1 loads heavily on Protein, with Iron and Calcium being loaded as well. Vitamin A and C are not loaded much at all for Factor 1.

Factor 2 loads heavily on Vitamin A and C, with Iron and Calcium also being loaded as well. Protein is not loaded very much for Factor 2.

Both factors have a somewhat balanced loading on Iron and Calcium, with Factor 1 having slightly more relative to Factor 2 for Calcium. Regardless, this plot demonstrates what we observed in Task 4, which was that Factor 1 indicates the common variation of macronutrients and minerals (Protein, Calcium, Iron), and Factor 2 indicates the common variation of micronutrients/vitamins (Vitamin A, Vitamin C, Iron). Iron is a mineral, but it is still loaded relatively high for both factors

Task 6

We were able to reduce the dimensions of the nutrient dataset using PCA, which allowed us to reduce the dimensions of the dataset from 5 to 2, making it easier to visualize/understand, while still explaining the majority of the variance of the dataset.

We fit a factor model using MLE to create two factors, one associated with macronutrients (Calcium, Iron, Protein) and one associated with micronutrients (VitaminA, VitaminC)

Based on these results with two nutrient groups, we may be able to conduct further investigation on nutrient intake by looking for foods that commonly provide a number of nutrients from the same group.

Conclusion

Upon first analysis of the data, we used a level plot to look for correlations between any of the variables in the dataset, and found the strongest correlation (darkest value with the most oval-like shape) to be between Protein and Iron, with a value of 62. Afterwards, we used PCA and MLE to fit our factor model, but favored the PCA.