Principal Component Analysis on USDA National Nutrient Database data

By: Kriti Srivastava

Date: 10/27/2020

- **Specific dataset :** nndb_flat.csv
- **Scope:**
    - o Dimensions: 8,618 rows, 45 columns
    - o COLUMN NAME
        - ▪ **1 Unique/Integer**: id
        - ▪ **6 Catrgorical/String**: foodgroup, shortdescrip, descript, commonname, mfgname, scientificname
        - ▪ **38 Continuous/Decimal:** energy_kcal, protein_g, fat_g, carb_g, sugar_g, fiber_g, vita_mcg, vitb6_mg, vitb12_mcg, vitc_mg, vite_mg, folate_mcg, niacin_mg, riboflavin_mg, thiamin_mg, calcium_mg, copper_mcg, iron_mg, magnesium_mg, manganese_mg, phosphorus_mg, selenium_mcg, zinc_mg, vita_usrda, vitb6_usrda, vitb12_usrda, vitc_usrda, vite_usrda, folate_usrda, niacin_usrda, riboflavin_usrda, thiamin_usrda, calcium_usrda, copper_usrda, magnesium_usrda, phosphorus_usrda, selenium_usrda, zinc_usrda
    - o Data units: Each record is for 100 grams. The nutrient columns end with the units, so:
        - ▪ Nutrient_g is in grams
        - ▪ Nutrient_mg is in milligrams
        - ▪ Nutrient_mcg is in micrograms
        - ▪ Nutrient_USRDA is in the percentage of US Recommended Daily Allows (e.g. 0.50 is 50%)

For PCA Analysis, I am using columns : energy_kcal, protein_g, fat_g, carb_g, sugar_g, fiber_g, vita_mcg, vitb6_mg, vitb12_mcg, vitc_mg, vite_mg, folate_mcg, niacin_mg, riboflavin_mg, thiamin_mg, calcium_mg, copper_mcg, iron_mg, magnesium_mg, manganese_mg, phosphorus_mg, selenium_mcg, zinc_mg

The columns with 0 entries are considered as missing, therefore, omitting zero.

Exploratory Data Analysis

Initial dimensions of the sample set:

Number of variables: 45
Number of rows: 8618

Dimensions of the sample only with the columns being used for PCA:

Number of variables: 23
Number of rows: 8618
Number of missing values: 8214
Number of rows without missing values: 404

Cleaned sample after removing missing values:

Number of variables: 23
Number of rows: 404
Several rows per column: 17 samples per column.
Therefore, the sample looks sufficient enough at this stage to continue with PCA.

R Code:

```
#Set Working Directory
#--------------------------------------
setwd('C:/Users/Kriti/Downloads')


#Read in Datasets
 #-------------------------------------------------------------
nndb <- read.csv(file="nndb_flat.csv", header=TRUE, sep=",")
head(nndb)
# Initial Check on Sample Size and Number of Variables
#-------------------------------------------------------------
dim(nndb)
# [1] 8618 45 names(nndb)
 # Considering the columns energy_kcal, protein_g, fat_g, carb_g, sugar_g,
#fiber_g,vita_mcg, vitb6_mg, vitb12_mcg, vitc_mg, vite_mg, folate_mcg,
#niacin_mg, riboflavin_mg, thiamin_mg, calcium_mg, copper_mcg, iron_mg,
 #magnesium_mg, manganese_mg, phosphorus_mg, selenium_mcg, zinc_mg
#-----------------------------------------------------------------
nndb_features <- nndb[,c(8:30)]
head(nndb_features)
names(nndb_features)


#Show for first 6 rows of data
#-----------------------------------------
 head(nndb_features)
```

```
#Show the headers
#-----------------------------------------

names(nndb_features)

#Check for Missing Values (i.e. 0s) #checking for rows that has missing values (0 represents missing)
 #-----------------------------------------------------------------
NROW((nndb_features[rowSums(nndb_features == 0) !=0, ]))
#8214 rows has zero in it.

#checking for rows that do not has missing values (0 represents missing)
#-----------------------------------------------------------------
NROW((nndb_features[rowSums(nndb_features == 0)==0, ]))
# 404 rows has no 0 in it.

 #omitting rows with zero in it
 #-----------------------------------------

nndb_features1<-data.frame(nndb_features[rowSums(nndb_features == 0)==0, ])
dim(nndb_features1)
```

Normality test:

- If skewness is close to 0, the distribution is normal.
- If Kurtosis is -3 or 3, the distribution is normal.
- If skewness is less than -1 or greater than 1, the distribution is highly skewed.
- If skewness is between -1 and -0.5 or between 0.5 and 1, the distribution is - moderately skewed.
- If skewness is between -0.5 and 0.5, the distribution is approximately symmetric

Most of the variables where highly skewed However, the following variables are highly skewed as shown in the fig:1 below:

```
> #show description
> #----------------------
> library(psych)
> describe(nndb_features1)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Energy_kcal	1	404	257.92	142.58	267.00	255.52	180.88	23.00	542.00	519.00	-0.01	-1.23	7.09
Protein_g	2	404	7.13	4.90	5.89	6.55	4.27	0.60	30.00	29.40	1.29	2.21	0.24
Fat_g	3	404	9.23	8.47	6.42	7.95	6.99	0.20	57.85	57.65	1.65	4.15	0.42
Carb_g	4	404	37.65	28.01	28.00	35.75	29.37	0.54	91.30	90.76	0.47	-1.26	1.39
Sugar_g	5	404	15.16	16.44	7.11	12.45	8.41	0.08	71.00	70.92	1.21	0.61	0.82
Fiber_g	6	404	2.61	3.04	1.60	2.01	1.48	0.10	29.30	29.20	3.28	17.69	0.15
VitA_mcg	7	404	155.23	269.83	43.00	86.57	50.41	1.00	1324.00	1323.00	2.38	5.02	13.42
VitB6_mg	8	404	0.51	1.16	0.09	0.23	0.08	0.01	12.00	11.99	4.68	30.84	0.06
VitB12_mcg	9	404	1.50	3.06	0.32	0.76	0.36	0.01	21.00	20.99	3.52	15.37	0.15
VitC_mg	10	404	10.18	33.61	1.10	3.53	1.33	0.10	489.90	489.80	8.96	108.59	1.67
VitE_mg	11	404	1.86	5.76	0.50	0.66	0.52	0.01	51.92	51.91	6.02	40.10	0.29
Folate_mcg	12	404	197.39	411.76	39.00	86.84	50.41	1.00	2331.00	2330.00	3.04	9.64	20.49
Niacin_mg	13	404	5.08	9.14	1.78	2.96	2.16	0.03	69.00	68.97	3.59	16.96	0.45
Riboflavin_mg	14	404	0.45	0.77	0.18	0.27	0.16	0.01	5.86	5.85	3.60	17.19	0.04
Thiamin_mg	15	404	0.41	0.73	0.13	0.25	0.15	0.01	5.18	5.17	3.54	16.40	0.04
Calcium_mg	16	404	122.49	220.62	71.50	83.90	76.35	3.00	3333.00	3330.00	8.84	113.36	10.98
Copper_mcg	17	404	0.15	0.15	0.10	0.12	0.08	0.00	1.41	1.41	3.37	17.69	0.01
Iron_mg	18	404	4.60	8.77	1.53	2.33	1.51	0.03	62.10	62.07	3.36	13.54	0.44
Magnesium_mg	19	404	34.01	47.84	21.00	24.46	13.34	2.00	473.00	471.00	5.00	32.12	2.38
Manganese_mg	20	404	1.18	13.39	0.24	0.31	0.20	0.00	269.10	269.10	19.82	393.63	0.67
Phosphorus_mg	21	404	148.77	127.71	115.00	128.72	85.99	10.00	1150.00	1140.00	3.00	14.64	6.35
Selenium_mcg	22	404	11.25	10.35	7.85	9.71	7.93	0.20	67.50	67.30	1.74	4.42	0.52
Zinc_mg	23	404	2.46	5.69	0.84	1.11	0.67	0.08	51.70	51.62	5.56	38.66	0.28

Fig:1

R Code:

```
#show description
 #----------------------
 library(psych) describe(nndb_features1)
```

Transforming all the columns by performing log on all the columns and re-run the description.

R code:

```
 # transforming all the columns to log as dataset log_nndb_features1 and re-run the describe
#-------------------------------------------
 log_nndb_features1 <- nndb_features1
for(j in 1:ncol(log_nndb_features1)){
log_nndb_features1[,j] <- log(log_nndb_features1[,j])
}
 library(psych)
describe(log_nndb_features1)
```

```
> library(psych)
> describe(log_nndb_features1)
                vars   n  mean   sd median trimmed  mad   min  max range  skew kurtosis   se
Energy_kcal        1 404  5.33 0.75   5.59    5.40 0.60  3.14 6.30  3.16 -0.79    -0.53 0.04
Protein_g          2 404  1.72 0.75   1.77    1.75 0.77 -0.51 3.40  3.91 -0.45    -0.10 0.04
Fat_g              3 404  1.75 1.08   1.86    1.80 1.15 -1.61 4.06  5.67 -0.45    -0.45 0.05
Carb_g             4 404  3.25 0.97   3.33    3.33 1.22 -0.62 4.51  5.13 -0.57    -0.29 0.05
Sugar_g            5 404  1.96 1.38   1.96    2.02 1.80 -2.53 4.26  6.79 -0.31    -0.62 0.07
Fiber_g            6 404  0.43 1.08   0.47    0.47 1.03 -2.30 3.38  5.68 -0.30     0.07 0.05
VitA_mcg           7 404  3.74 1.73   3.76    3.74 1.66  0.00 7.19  7.19 -0.01    -0.45 0.09
VitB6_mg           8 404 -2.03 1.46  -2.38   -2.17 1.04 -5.12 2.48  7.60  0.96     0.17 0.07
VitB12_mcg         9 404 -1.03 1.72  -1.14   -1.06 1.46 -4.61 3.04  7.65  0.25    -0.35 0.09
VitC_mg           10 404  0.40 1.88   0.10    0.29 1.93 -2.30 6.19  8.50  0.56    -0.44 0.09
VitE_mg           11 404 -0.67 1.42  -0.69   -0.71 1.09 -4.61 3.95  8.55  0.36     1.13 0.07
Folate_mcg        12 404  3.68 1.82   3.66    3.59 1.92  0.00 7.75  7.75  0.36    -0.66 0.09
Niacin_mg         13 404  0.50 1.57   0.58    0.50 1.34 -3.54 4.23  7.77 -0.01    -0.46 0.08
Riboflavin_mg     14 404 -1.66 1.27  -1.69   -1.70 0.99 -4.71 1.77  6.48  0.32    -0.09 0.06
Thiamin_mg        15 404 -1.94 1.45  -2.02   -2.00 1.64 -4.83 1.64  6.47  0.32    -0.71 0.07
Calcium_mg        16 404  4.14 1.16   4.27    4.15 1.14  1.10 8.11  7.01 -0.05    -0.23 0.06
Copper_mcg        17 404 -2.32 0.94  -2.30   -2.30 0.91 -6.91 0.34  7.25 -0.44     1.23 0.05
Iron_mg           18 404  0.46 1.41   0.43    0.39 1.15 -3.51 4.13  7.64  0.36     0.00 0.07
Magnesium_mg      19 404  3.09 0.88   3.04    3.07 0.71  0.69 6.16  5.47  0.28     1.03 0.04
Manganese_mg      20 404 -1.52 1.45  -1.44   -1.47 1.00 -5.81 5.60 11.40 -0.15     1.81 0.07
Phosphorus_mg     21 404  4.71 0.79   4.74    4.74 0.74  2.30 7.05  4.74 -0.30     0.18 0.04
Selenium_mcg      22 404  1.97 1.04   2.06    2.02 1.20 -1.61 4.21  5.82 -0.44    -0.19 0.05
Zinc_mg           23 404 -0.02 1.16  -0.18   -0.13 0.86 -2.53 3.95  6.47  0.91     0.99 0.06
```

Fig:2

The skewness of all the columns either between -0.5 and 0.5 oe between -1 and -0.5 or between 0.5 and 1, that means the distribution is either the distribution is approximately symmetric or -moderately skewed.

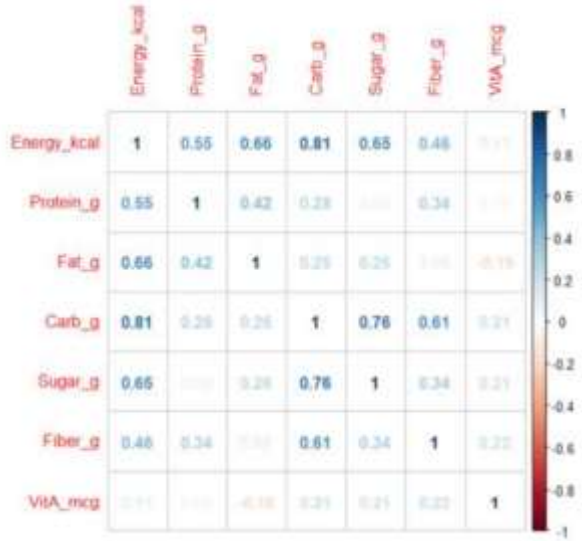

Fig: 3

Since Carb_g is highly correlated to and Energy_kcal and Sugar_g. So removing Energy_ kcal and Sugar_g as we know calories are dependent on many other factors and sugar is dependent on carbs.

R code:

```
set1 <- log_nndb_features1[,1:7]
 cor.set1 = cor(set1)
set1
corrplot(cor.set1, method="number")
```

After removing Energy_kcal and Sugar_g and checking the correlation again, we can see a high correlation between Niacin_mg and vitB6_mg, and Folate_mcg in fig:4, therefore removing Niacin_mg.

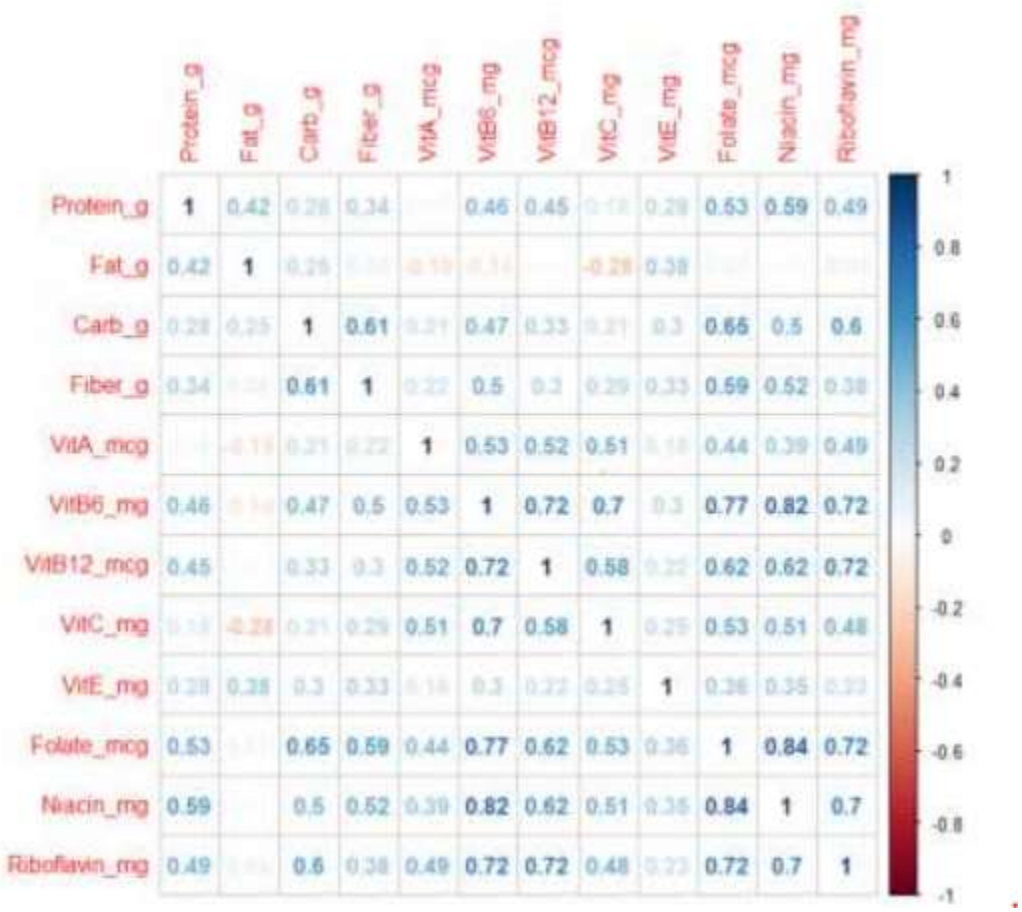

Fig:4

After removing Niacin_mg and checking the correlation again, we again can see a high correlation between Folated_mg and Thiamin_mg, and Iron_mg in fig:5, therefore removing Folated_mg,.

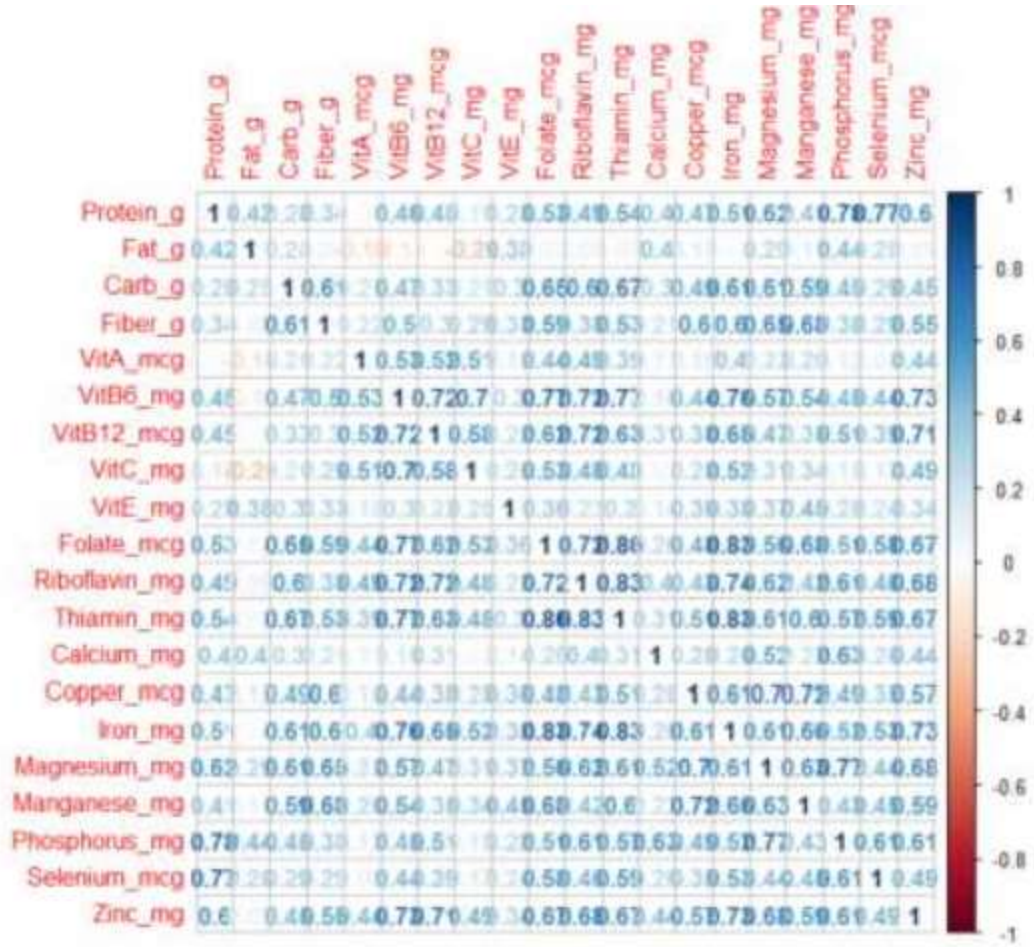

Fig:5

After removing Folated_mg and Thiamin_mg

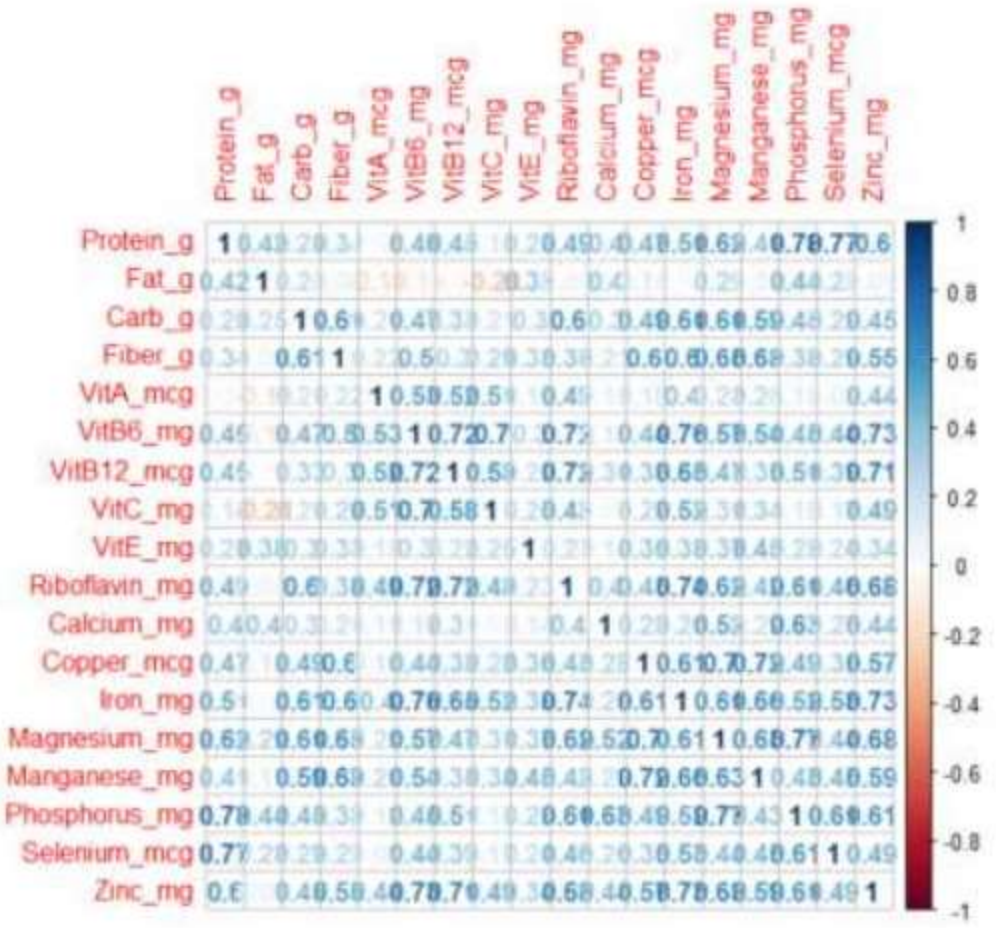

Fig: 6

Assumptions for factorability

The entire data set:

- **KMO Test:** Overall MSA = 0.89 (Since it is greater than .7 that means we are good)
- **Bartlett's Test of Sphericity**: p-value < 2.22e-16 (very small which means we have enough variance in the data so we can perform factor analysis)
- **Reliability Analysis using Cronbach's Alpha:** raw_alpha = 0.92 Hence, The sample data is good to perform PCA.

R Code:

#Test KMO Sampling Adequacy

```
library(psych)
KMO(log_nndb_features4)
```

#Test Bartlett's Test of Sphericity

```
library(REdaS)
bart_spher(log_nndb_features4)
#p-value < 2.22e-16 (Very Small Number)
```

#Test for Reliability Analysis using Cronbach's Alpha

```
library(psych)
alpha(log_nndb_features4,check.keys=TRUE)
```

Selecting the components:

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10
Standard deviation     2.9060 1.5500 1.22185 0.99912 0.96728 0.82340 0.69657 0.64301 0.61984 0.58567
Proportion of Variance 0.4692 0.1335 0.08294 0.05546 0.05198 0.03767 0.02696 0.02297 0.02134 0.01906
Cumulative Proportion  0.4692 0.6026 0.68558 0.74103 0.79301 0.83068 0.85764 0.88061 0.90195 0.92101
                          PC11    PC12    PC13    PC14    PC15    PC16    PC17    PC18
Standard deviation     0.54954 0.48290 0.46305 0.40805 0.39144 0.36914 0.34138 0.31580
Proportion of Variance 0.01678 0.01295 0.01191 0.00925 0.00851 0.00757 0.00647 0.00554
Cumulative Proportion  0.93778 0.95074 0.96265 0.97190 0.98041 0.98798 0.99446 1.00000
> print(p)
Standard deviations (1, .., p=18):
 [1] 2.9060315 1.5499794 1.2218516 0.9991232 0.9672845 0.8234036 0.6965669 0.6430072 0.6198412 0.5856703
[11] 0.5495371 0.4828968 0.4630507 0.4080499 0.3914426 0.3691422 0.3413824 0.3158020
```

Fig: 7

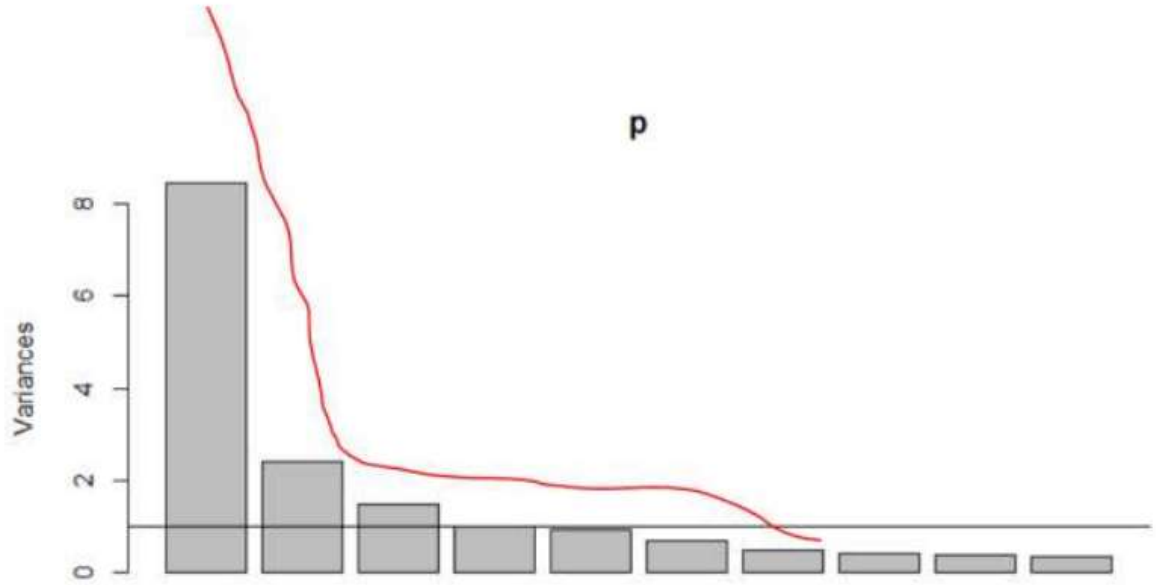

Fig: 8

RCode:

```
 ###Create PCA
p = prcomp(log_nndb_features4, center=T, scale=T)
p

#Check Scree Plot
plot(p)
abline(1, 0)
```

#Check PCA Summary Information
summary(p)
print(p)

According to PCS summary information (Fig:7), approx. 80% of the cumulative variance is explained by 5 components. Approx 47% of the cumulative variance is explained by first component itself. 3 components are determined by the Scree plot (Fig: 4) which is greater than 1 Eigenvalue.However, On performing the Knee test, there are 2 components. Since, 68% of the cumulative variance is determined by the 3 components therefore, I would use 3 components in the model.

The visualize the top 10 variables of the first components.

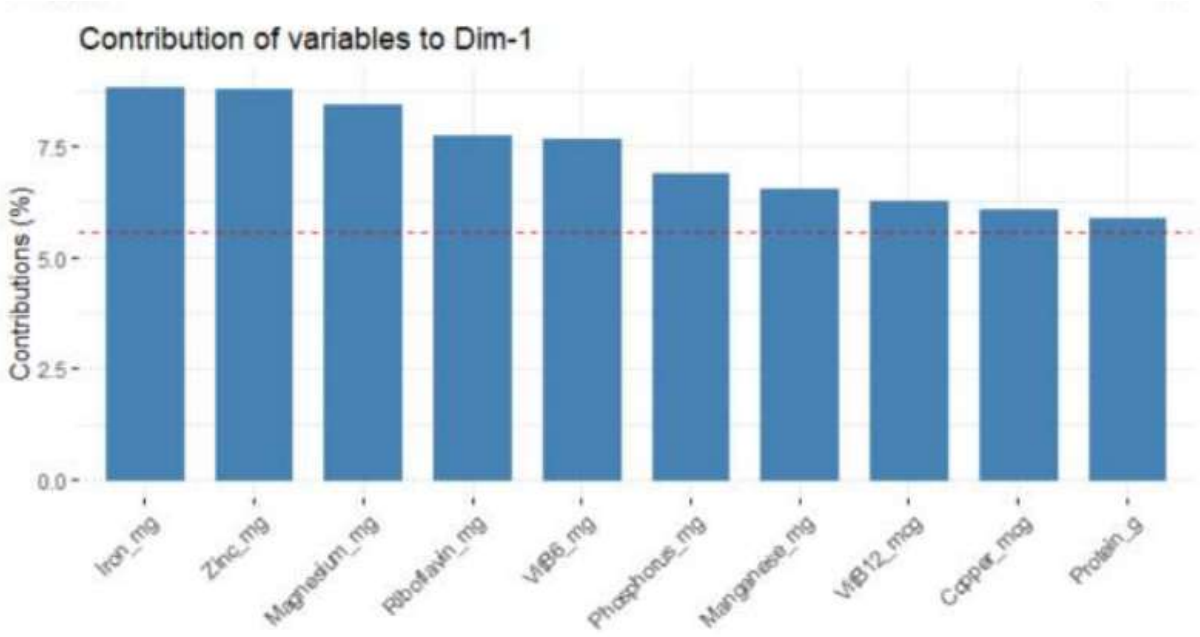

Fig:9

Assuming the variables are independent, the factor rotation used is “varimax” with nfactos = 3

PCA output:

```
Mean item complexity =  1.6
Test of the hypothesis that 3 components are sufficient.

The root mean square of the residuals (RMSR) is  0.06
 with the empirical chi square  481.5  with prob <  1.4e-50

Fit based upon off diagonal values = 0.98> print(p2$loadings, cutoff=.48, sort=T)

Loadings:
              RC3    RC2    RC1
Carb_g         0.709
Fiber_g        0.826
VitE_mg        0.547
Copper_mcg     0.756
Magnesium_mg   0.640         0.547
Manganese_mg   0.832
VitA_mcg              0.737
VitB6_mg              0.800
VitB12_mcg            0.782
VitC_mg               0.799
Riboflavin_mg         0.665
Iron_mg        0.581  0.596
Zinc_mg               0.603
Protein_g                    0.822
Fat_g                        0.641
Calcium_mg                   0.686
Phosphorus_mg                0.853
Selenium_mcg                 0.672

                 RC3   RC2   RC1
SS loadings    4.224 4.155 3.962
Proportion Var 0.235 0.231 0.220
Cumulative Var 0.235 0.465 0.686
```

Component **RC3** is formulated by Carg_g, Fiber_g , VitE_mg, Copper_mcg, Maganese_mg. Since, Magnesium_mg is explained by 64% of the variance in RC3 however only 54% by RC1 therefore, I choose to keep Magnesium_mg with RC3. Since this group of food is rich in fiber, Vitamin E as well as carbohydrates, therefore, I can name component RC3 as **FiberRichedCarbsFood.**

Formula for **FiberRichedCarbsFood (RC3) component**:

FiberRichedCarbsFood= 0.709*Carb_g + 0.826*Fiber_g + 0.547*VitE_mg + 0.756*Copper_mcg + 0.640*Magnesium_mg + 0.832*Manganese_mg

Component **RC2** is formulated by VitA_mcg, VitB6_mg, VitB12_mcg, VitC_mg, Riboflavin_mg, Zinc_mg. Since, Iron_mg is explained by 60% of variance in RC2 however only 58% by RC3 therefore, I choose to keep Iron_mg with RC2. Since, this group of food is rich in Vitamins, iron, and zinc. Foods rich in vitamin C helps in the absorption of iron therefore, they are good for anemic people therefore, I named component RC2 as **Anti-anemicFood**.

Formula for **Anti-anemicFood(RC2)** component:

Anti-anemicFood = 0.737*VitA_mcg + 0.800*VitB6_mg + 0.782*VitB12_mcg + 0.799*VitC_mg + Riboflavin_mg*0.665 + 0.596*Iron_mg + 0.603*Zinc_mg

Component **RC1** is formulated by Protein_g, Fat_g, Calcium_mg, Phosphorus_mg, Selenium_mcg. Since this group of food is rich in protein and fats, I named component RC1 as HighProtienFood. The formula for **HighProtienFood(RC1)** component:

HighProtienFood = 0.822*Protein_g + 0.641*Fat_g + 0.868*Calcium_mg + 0.853*Phosphorus_mg + 0.672*Selenium_mcg

#R Code for PCA Analysis:

```
library("FactoMineR")
p4 <- PCA(log_nndb_features4, graph = FALSE)
#IF graph is set to true, it will provide the individual and variable maps
variables <- get_pca_var(p4)
#Which variables contribute the most to the PCs?
#there are ll variables
head(variables$contrib, 11)

library("corrplot")
corrplot(variables$contrib, is.corr=FALSE)

# Contributions of variables to PC1
fviz_contrib(p4, choice = "var", axes = 1, top = 10)
```

 # PCA

```
p2 = psych::principal(log_nndb_features4, rotate="varimax", nfactors=3, scores=TRUE)
p2
print(p2$loadings, cutoff=.48, sort=T)
```

Highest and lowest values for each principal component

Components	Lowest Score	Highest Score
FiberRichedCarbsFood (RC3)	-3.23365	2.90078
Anti-anemicFood (RC2)	-1.8891	3.1326
HighProtienFood (RC1)	-3.4094	3.0945

Explanation:

FiberRichedCarbsFood: The least score for FiberRichedCarbsFood is -3.23365 which means the food item is very low in fiber and carbohydrates whereas the food item who scores the highest as 2.90078 is very good in the nutrients like fiber, carbohydrates, and other minerals.

Anti-anemicFood: The food item that scores lowest as -1.8891 looks like it has nutrients that are low in iron and vitamins however food item that has the highest score as 3.1326 is very good for people dealing with amnesia or has low vitamins as this food item is good in iron, zinc, and vitamins.

HighProtienFood: The food item that scores lowest as -3.4094 are very low in proteins and fats however food item that has the highest score as 3.0945 is a high source of proteins and fats like meat which are good for muscle building.

Code: Calculating scores

```
scores <- p2$scores
```

```
head(scores)

scores_1 <- scores[,1]
 summary(scores_1)

scores_2 <- scores[,2]
summary(scores_2)

scores_3 <- scores[,3]
summary(scores_3)
```