Question #1

Perform PCA on USDA National Nutrient Database and summarize findings

Preparing Data

1. **Check correlation between numeric variables.** "Nutrient_USRDA" is highly correlated to "Nutrient_g"/"Nutrient_mg"/"Nutrient_mcg" with correlation 1. "Nutrient_USRDA" is redundant and should be removed.

2. **Delete ID and non-numeric variables** since PCA only applied on numerical data.

3. View the new data created, there's **no missing value**.

4. **Explore distribution of data.** Histograms of most of the variables are skewed right. Consider transformations to "improve" the distributions and hopefully produce better correlations for PCA. Common transformations for right-skewed data are square root, cube root and log. Histograms look better with cube transformation.

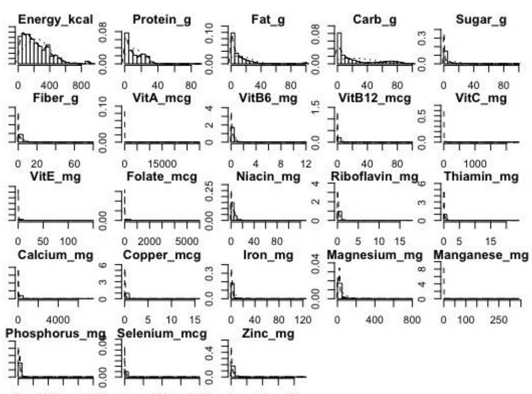


Figure 1 Histogram of origin

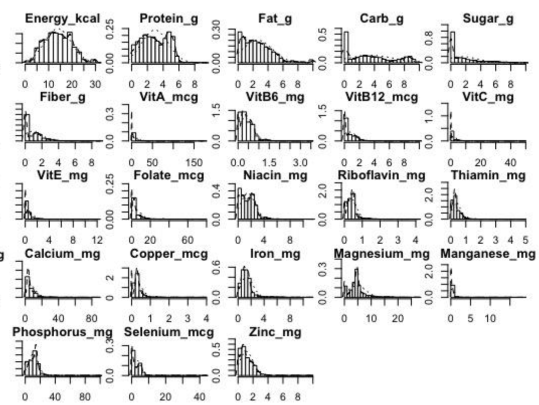


Figure 2 Histogram of square root

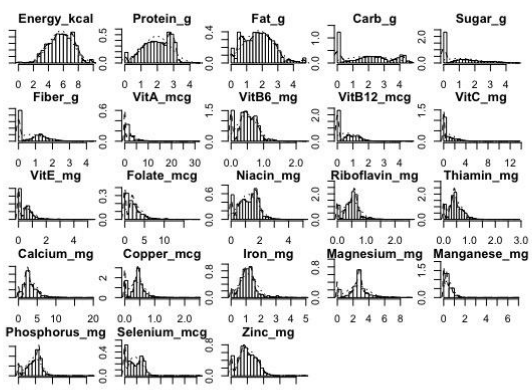


Figure 3 Histogram of cube

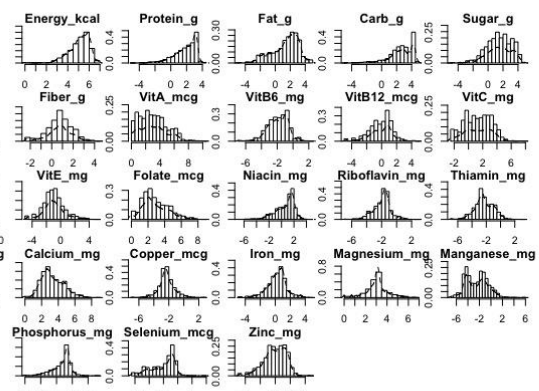


Figure 4 Histogram of log

5. It’s usually beneficial for each variable to be centered at zero for PCA, due to the fact that it makes comparing each principal component to the mean straightforward. This also eliminates potential problems with the scales of each variables, variables are measured in g, mg and mcg differently**. Standardizing each variable** will fix this issue.

Implement PCA in R

1. The **prcomp() function provides standard deviation and rotation**. The rotation measure provides the relationship between the initial variables and the principal components.

2. Summary table of principal components shows standard deviation, proportion of variance and cumulative proportion.

3. Square every standard deviation in Table 1 will get variance(eigenvalues) of each principal component. Add all variance up, then get the total variance of 22.99968.

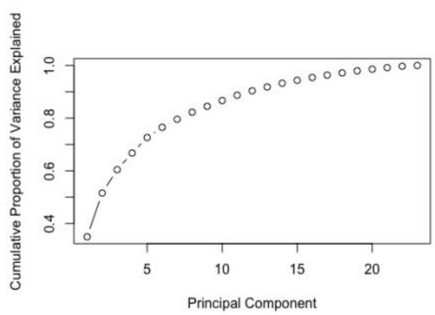
Component	**Standard deviation**	**Variance (eigenvalue)**	**Proportion of Variance**	**Cumulative Proportion**
PC1	2.8339	8.03098921	0.3492	0.3492
PC2	1.9554	3.82358916	0.1663	0.5154
PC3	1.42845	2.040469403	0.08872	0.60415
PC4	1.2067	1.45612489	0.06331	0.66746
PC5	1.16141	1.348873188	0.05865	0.7261
PC6	0.94899	0.90058202	0.03916	0.76526
PC7	0.83328	0.694355558	0.03019	0.79545
PC8	0.78742	0.620030256	0.02696	0.82241
PC9	0.71797	0.515480921	0.02241	0.84482
PC10	0.70991	0.503972208	0.02191	0.86673
PC11	0.6839	0.46771921	0.02034	0.88707
PC12	0.61641	0.379961288	0.01652	0.90359
PC13	0.58064	0.33714281	0.01466	0.91824
PC14	0.573	0.328329	0.01428	0.93252
PC15	0.50804	0.258104642	0.01122	0.94374
PC16	0.4904	0.24049216	0.01046	0.9542
PC17	0.45888	0.210570854	0.00916	0.96335
PC18	0.43902	0.19273856	0.00838	0.97173
PC19	0.42375	0.179564063	0.00781	0.97954
PC20	0.38409	0.147525128	0.00641	0.98595
PC21	0.36158	0.130740096	0.00568	0.99164
PC22	0.35291	0.124545468	0.00542	0.99705

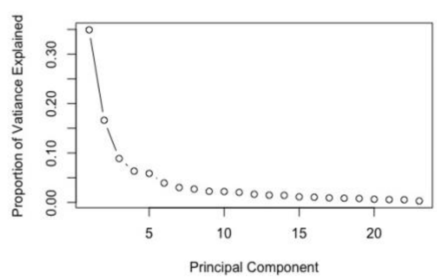
PC23	0.26034	0.067776916	0.00295	1
Total		22.99967701		

Table1

4. **The proportion of variance explained by each eigenvalue** is given in the fourth column in Table1. For example, 8.03098921 divided by 22.99968 equals 0.3492. About 34.92% of the variation is explained by the first eigenvalue. The cumulative percentage explained is obtained by adding the successive proportions of variation. For example, 0.3492 plus 0.1663 equals 0.5154, about 51.54% of the variation is explained by the first two eigenvalues together. The first two principal components explain 51.54% of the variance in the original variables.
5. There are several methods to **determine how many principal components to use**. Below presents scree plot and parallel analysis:

   a. In **scree plot**, we look for the point where the proportion of variance explained significantly drops off. The cumulated proportion of variance explained increases moderately after 5, so we stop at the fifth component. When using 5 out of 23 components, about 72.61% of the variance is accounted for and this is an acceptably large percentage.





   b. Run a **parallel analysis** to decide how many factors to retain. The third column of Table 2 shows how large eigenvalues can be as a result of just using randomly generated datasets. If the eigenvalue from actual data is greater than the generated eigenvalue, then have support to retain that factor. Since first five eigenvalues are greater than generated eigenvalues, we have support to **retain the first five component**. Same as the result from scree plot.

Component	Eigenvalue	0.95	Eigenvalue>0.95?
1	8.03098921	1.105	T
2	3.82358916	1.088	T
3	2.0404694	1.076	T
4	1.45612489	1.064	T
5	1.34887319	1.057	T
6	0.90058202	1.049	F

7	0.69435556	1.04	F
8	0.62003026	1.033	F
9	0.51548092	1.025	F
10	0.50397221	1.019	F
11	0.46771921	1.012	F
12	0.37996129	1.005	F
13	0.33714281	0.998	F
14	0.328329	0.991	F
15	0.25810464	0.984	F
16	0.24049216	0.977	F
17	0.21057085	0.971	F
18	0.19273856	0.964	F
19	0.17956406	0.957	F
20	0.14752513	0.949	F
21	0.1307401	0.942	F
22	0.12454547	0.932	F
23	0.06777692	0.922	F

Table2

Interpreting each component

1. First Principal Component Analysis-**PC1**
   The correlation between the first principal component and the original variables are copied into the Table3.
   First component can be viewed as the food that are **high** in *phosphorus, zinc, magnesium, iron, selenium, copper, calcium, manganese, niacin, riboflavin, vitB6, thiamin, vitB12, folate* and **low** in *sugar, vitC*.

$$PC1 = 0.2993 \times \left(\mathrm{Phosphorus_{mg}}\right) + 0.2971 \times \left(\mathrm{Zinc_{mg}}\right) + 0.2943 \times \left(\mathrm{Niacin_{mg}}\right) + \cdots + 0.0363 \times \left(\mathrm{Fiber_{g}}\right) + 0.0014 \times \left(\mathrm{Carb_{g}}\right) - 0.0121 \times \left(\mathrm{VitC_{mg}}\right) - 0.0389 \times (\mathrm{Sugar\_g})$$

	PC1
Phosphorus_mg	0.299277691
Zinc_mg	0.297107939
Niacin_mg	0.294338566
Riboflavin_mg	0.292815483
VitB6_mg	0.283511886
Magnesium_mg	0.270882614
Thiamin_mg	0.263032729
Protein_g	0.256891877
Iron_mg	0.255886779
Selenium_mcg	0.240638848
Copper_mcg	0.232833074
VitB12_mcg	0.20389979
Folate_mcg	0.20184984
Energy_kcal	0.150119339
Calcium_mg	0.148573035
Manganese_mg	0.145469225
Fat_g	0.114979685
VitE_mg	0.097376183
VitA_mcg	0.097345439
Fiber_g	0.036329846
Carb_g	0.001380964
VitC_mg	-0.012087581
Sugar_g	-0.038926437

Table3

	PC2
VitB12_mcg	0.27526006
Protein_g	0.23956295
Selenium_mcg	0.20580229
Fat_g	0.13779409
Zinc_mg	0.1272124
Niacin_mg	0.07174708
Phosphorus_mg	0.06047325
VitB6_mg	0.0265521
Riboflavin_mg	-0.0265569
Energy_kcal	-0.0304766
VitE_mg	-0.0835671
Copper_mcg	-0.0850684
VitA_mcg	-0.106607
Iron_mg	-0.1239173
Thiamin_mg	-0.1252729
Magnesium_mg	-0.1512906
Calcium_mg	-0.2094101
Folate_mcg	-0.2261434
Manganese_mg	-0.2415448
VitC_mg	-0.2593686
Sugar_g	-0.3350501
Fiber_g	-0.4108812
Carb_g	-0.4438004

Table4

2. Second Principal Component Analysis-**PC2**

   The correlation between the second principal component and the original variables are copied into the Table4.
   Second component can be viewed as the food that are **high** in *vitB12, protein, selenium* and **low** in *carb, fiber*.

3. Third Principal Component Analysis-**PC3**

	PC3
Energy_kcal	0.57234716
Fat_g	0.51446348
Sugar_g	0.20967826
Carb_g	0.20601933
Fiber_g	0.10659897
VitE_mg	0.09959312
Thiamin_mg	0.0919126
Protein_g	0.07434351
Iron_mg	0.06512707
Phosphorus_mg	0.03457768
Niacin_mg	0.01303043
Riboflavin_mg	-0.0056597
Calcium_mg	-0.0291242
Zinc_mg	-0.0395245
Magnesium_mg	-0.0586505
VitB12_mcg	-0.0841187
Selenium_mcg	-0.0887486
Manganese_mg	-0.1273156
VitB6_mg	-0.1419006
Folate_mcg	-0.159748
Copper_mcg	-0.1981232
VitA_mcg	-0.2031897
VitC_mg	-0.3423963

Table5

The correlation between the third principal component and the original variables are copied into the Table5.
Third component can be viewed as the food that are **high** in *energy, fat* and **low** in *vitC.*

4. Fourth Principal Component Analysis-**PC4**

The correlation between the fourth principal component and the original variables are copied into the Table6.
Fourth component can be viewed as the food that are **high** in *manganese, magnesium, copper* and **low** in *vitA.*

	PC4
Manganese_mg	0.36836763
Magnesium_mg	0.2736802
Copper_mcg	0.26595095
Fiber_g	0.15359432
Phosphorus_mg	0.1402365
Protein_g	0.13216424
Zinc_mg	0.05825884
Selenium_mcg	0.05711588
Carb_g	0.02732854
Iron_mg	-0.0104385
Energy_kcal	-0.0372095
Folate_mcg	-0.0468822
Calcium_mg	-0.0482742
Thiamin_mg	-0.0704216
Niacin_mg	-0.0839546
Fat_g	-0.0990789
VitB6_mg	-0.1424603
Riboflavin_mg	-0.2027302
Sugar_g	-0.2690974
VitC_mg	-0.2808463
VitE_mg	-0.2810481
VitB12_mcg	-0.3035216
VitA_mcg	-0.4887787

Table6

5. Fifth Principal Component Analysis-**PC5**

	PC5
VitE_mg	0.53767278
Copper_mcg	0.29529953
Fat_g	0.28766398
Manganese_mg	0.27775607
VitA_mcg	0.26184507
Selenium_mcg	0.16607056
Calcium_mg	0.13207705
Energy_kcal	0.13000716
Magnesium_mg	0.0908521
Folate_mcg	0.0639547
Phosphorus_mg	0.02213777
VitC_mg	0.02058759
Zinc_mg	-0.0001702
VitB12_mcg	-0.0374386
Sugar_g	-0.0826554
Protein_g	-0.0829129
Fiber_g	-0.1091013
VitB6_mg	-0.1192366
Carb_g	-0.138676
Riboflavin_mg	-0.2239977
Iron_mg	-0.2366658
Niacin_mg	-0.2542374
Thiamin_mg	-0.2979572

Table7

The correlation between the fifth principal component and the original variables are copied into the Table7. Fifth component can be viewed as the food that are **high** in *vitE* and **low** in *thiamin, niacin, iron, riboflavin.*