FOOD CLUSTERING

Insun Lee, Kim Nguyen, Chao Zheng

WASHINGTON STATE UNIVERSITY
CPTS 483 – FALLL 2016

Abstract

This project uses nutrient databases from US Department of Agriculture (USDA), which contains various type of foods with their corresponding nutritional contents. We will experiment four major type of food groups using the following data analysis tools: normalization, principal components analysis, K-means, and hierarchical clustering. Our findings show that major food groups can be further categorized and clustered in hierarchies.

Introduction

This project analyzes data from US Department of Agriculture Nutrient Database, which consists of various type of foods with their corresponding nutritional contents. The data contains detailed categorizations of each food item. For example, there are nine categories of Kale. The range includes raw, frozen and unprepared, cooked and boiled, cooked and drained without salt, etc. In addition, common knowledge suggests that major food groups can be further categorized. Vegetables can be leaves/stems, roots, or buds. Based on their nutritional contents, we expect to see these items clustered in hierarchies, from the major food groups, subgroups, and finally to variants of the same food items. The experiment is conducted on four major type of food groups: Cereal-Grain-Pasta, Finfish-Shellfish, Vegetables, Fats-Oils. We will apply the techniques of normalization, principal components analysis, K-means, and hierarchical clustering to analyze the dataset. Our findings show that major food groups can be further categorized.

Problem Definition

Our problem is clustering a large dataset. The work includes normalizing the data, applying principal components analysis (PCA), k-mean clustering, and hierarchical clustering. At end of the project, we expect our data to be clustered into four major groups. Also, after the first cluster we will continue to analyze the current result to determine if it can be future categorized into subgroup.

Models/Algorithms/Measures

First we have to normalize the data to transform the features to be in the range [0,1]. For jth dimension, the value of the ith data point will become:

$$Normalized(X_{ij}) = \frac{X_{ij} - \min{(X_j)}}{\max(X_j) - \min{(X_j)}} \qquad (1)$$

where min and max for X_j are calculated over the jth attribute/dimension on the complete dataset.

The second technique is the Principal Components Analysis (PCA). PCA produces a low-dimensional representations of the variables that have maximal variance, and mutually uncorrelated. It is also a tool for data visualization.

The first principal component of a set of features:

X_1, X_2, … X_p is the normalized linear combination of the features:

$$Z_1 = \varphi_{11}X_1 + \varphi_{21}X_2 + \cdots + \varphi_{p1}X_p \qquad (2)$$

with φ is the loading of the first principal component.

The second principal component states that linear combination of X_1, X_2, … X_p that has maximal variance among all linear combinations that are uncorrelated with Z_1.

The next tool the is K-mean method with four basic steps:

1. Randomly pick k centroids (centers of clusters)
2. Assign each data point to the closest centroid
3. Recompute cluster centroids (average location of data points) in light of current cluster assignments
4. Repeat step 2 and 3 until assignment do not change or change very little

The final concept is the proportion variance explained (PVE). The total variance in a dataset is defined as:

$$\sum_{j=1}^{p} Var(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2 \qquad (3)$$

and the variance explained by the m^th principal component is:

$$Var(Z_m) = \frac{1}{n} \sum_{i=1}^{n} x_{im}^2 \qquad (4)$$

Therefore, the PVE of the m^th principal component is given by the positive quantity between 0 and 1:

$$\frac{\sum_{i=1}^{n} z_{im}^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2} \qquad (5)$$

Implementation/Analysis

As stated in the introduction, we will analyze USDA nutritional database on Cereal-Grain-Pasta, Finfish-Shellfish, Vegetables, Fats-Oils. Before performing any analysis, we combine the content of these four food groups into one big text file. The original data we first encounter is very dirty so the first thing we need to do is cleaning up the dataset. We remove all of the columns that contains all 0's and ignore all the NaN (stands for not a number, usually defines as 0/0) and NA (stands for not applicable datatype).

Numerical values vary widely across different types of nutrients. Small numerical values in some micronutrients such as minerals and vitamins may characterize the food items, whereas larger numerical values do the same in macronutrients like protein and carbohydrates. Therefore, it is important to first normalize the nutrient values to transform the features to be in the range [0,1] (Figure 1, 2, 3, and 4).

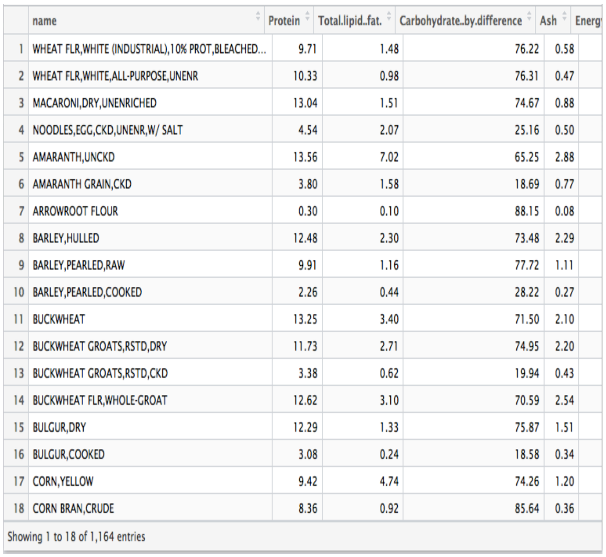
For the Principal Component Analysis, we visualize the data using only the first two principal components. Examine the first and second principal components to find their five highest absolute weights and their corresponding nutrients (Figure 6).

Next, we apply K-means clustering on the original data, with different sizes of cluster k = [4, 6, 8, 10, 12]. We display the result on the reduced dimensions (2-D) with a different color for each cluster (Figure 7, 8, 9 10, and 11).

Finally, we visualize the possible structures of food using hierarchical clustering and dendrograms. Thirty food items are randomly select from each groups. Then we create a dendrogram and review the labels of the food items from the dendrograms. We identify any distinct clusters and the corresponding food items, and obtain actual clusters from hierarchical clustering (Figure 12, 13, 14).

Results and Discussion

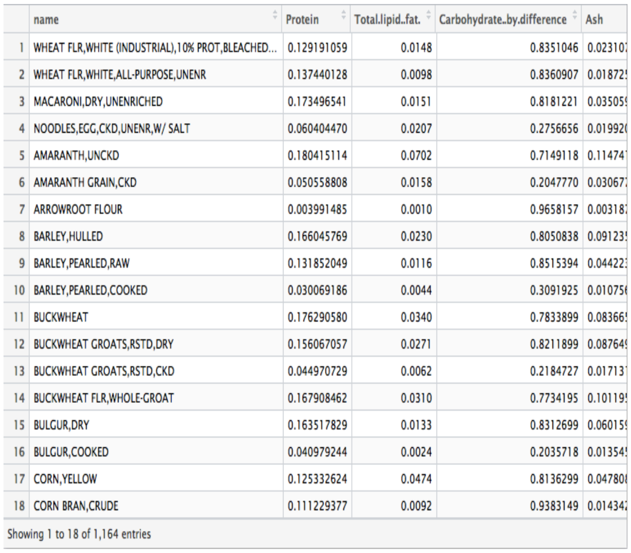
Below is the result before and after the normalization process of data:



	name	Protein	Total.lipid..fat.	Carbohydrate..by.difference	Ash	Energ
1	WHEAT FLR,WHITE (INDUSTRIAL),10% PROT,BLEACHED...	9.71	1.48	76.22	0.58	
2	WHEAT FLR,WHITE,ALL-PURPOSE,UNENR	10.33	0.98	76.31	0.47	
3	MACARONI,DRY,UNENRICHED	13.04	1.51	74.67	0.88	
4	NOODLES,EGG,CKD,UNENR,W/ SALT	4.54	2.07	25.16	0.50	
5	AMARANTH,UNCKD	13.56	7.02	65.25	2.88	
6	AMARANTH GRAIN,CKD	3.80	1.58	18.69	0.77	
7	ARROWROOT FLOUR	0.30	0.10	88.15	0.08	
8	BARLEY,HULLED	12.48	2.30	73.48	2.29	
9	BARLEY,PEARLED,RAW	9.91	1.16	77.72	1.11	
10	BARLEY,PEARLED,COOKED	2.26	0.44	28.22	0.27	
11	BUCKWHEAT	13.25	3.40	71.50	2.10	
12	BUCKWHEAT GROATS,RSTD,DRY	11.73	2.71	74.95	2.20	
13	BUCKWHEAT GROATS,RSTD,CKD	3.38	0.62	19.94	0.43	
14	BUCKWHEAT FLR,WHOLE-GROAT	12.62	3.10	70.59	2.54	
15	BULGUR,DRY	12.29	1.33	75.87	1.51	
16	BULGUR,COOKED	3.08	0.24	18.58	0.34	
17	CORN,YELLOW	9.42	4.74	74.26	1.20	
18	CORN BRAN,CRUDE	8.36	0.92	85.64	0.36	

Showing 1 to 18 of 1,164 entries

Figure 1. Data before normalization



	name	Protein	Total.lipid..fat.	Carbohydrate..by.difference	Ash
1	WHEAT FLR,WHITE (INDUSTRIAL),10% PROT,BLEACHED...	0.129191059	0.0148	0.8351046	0.02310
2	WHEAT FLR,WHITE,ALL-PURPOSE,UNENR	0.137440128	0.0098	0.8360907	0.01872
3	MACARONI,DRY,UNENRICHED	0.173496541	0.0151	0.8181221	0.03505
4	NOODLES,EGG,CKD,UNENR,W/ SALT	0.060404470	0.0207	0.2756656	0.01992
5	AMARANTH,UNCKD	0.180415114	0.0702	0.7149118	0.11474
6	AMARANTH GRAIN,CKD	0.050558808	0.0158	0.2047770	0.03067
7	ARROWROOT FLOUR	0.003991485	0.0010	0.9658157	0.00318
8	BARLEY,HULLED	0.166045769	0.0230	0.8050838	0.09123
9	BARLEY,PEARLED,RAW	0.131852049	0.0116	0.8515394	0.04422
10	BARLEY,PEARLED,COOKED	0.030069186	0.0044	0.3091925	0.01075
11	BUCKWHEAT	0.176290580	0.0340	0.7833899	0.08366
12	BUCKWHEAT GROATS,RSTD,DRY	0.156067057	0.0271	0.8211899	0.08764
13	BUCKWHEAT GROATS,RSTD,CKD	0.044970729	0.0062	0.2184727	0.01713
14	BUCKWHEAT FLR,WHOLE-GROAT	0.167908462	0.0310	0.7734195	0.10119
15	BULGUR,DRY	0.163517829	0.0133	0.8312699	0.06015
16	BULGUR,COOKED	0.040979244	0.0024	0.2035718	0.01354
17	CORN,YELLOW	0.125332624	0.0474	0.8136299	0.04780
18	CORN BRAN,CRUDE	0.111229377	0.0092	0.9383149	0.01434

Showing 1 to 18 of 1,164 entries

Figure 2. Data after normalization

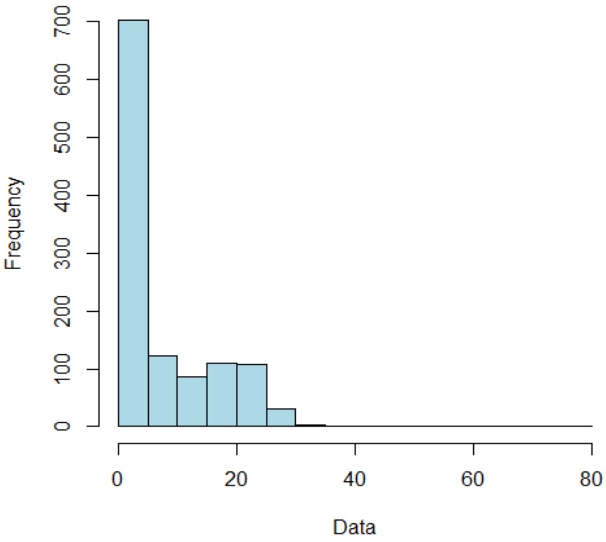


Figure 3. Data plot before normalization

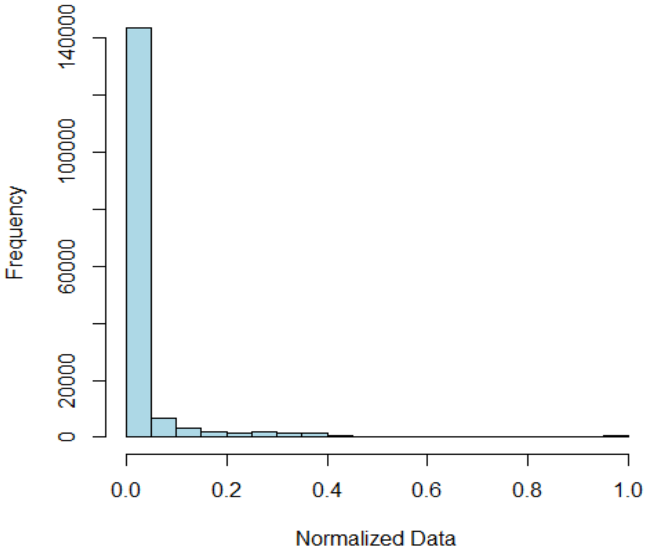


Figure 4. Data plot after normalization

Biplot is an enhanced scatter plot that uses both points and vectors to represent structure.

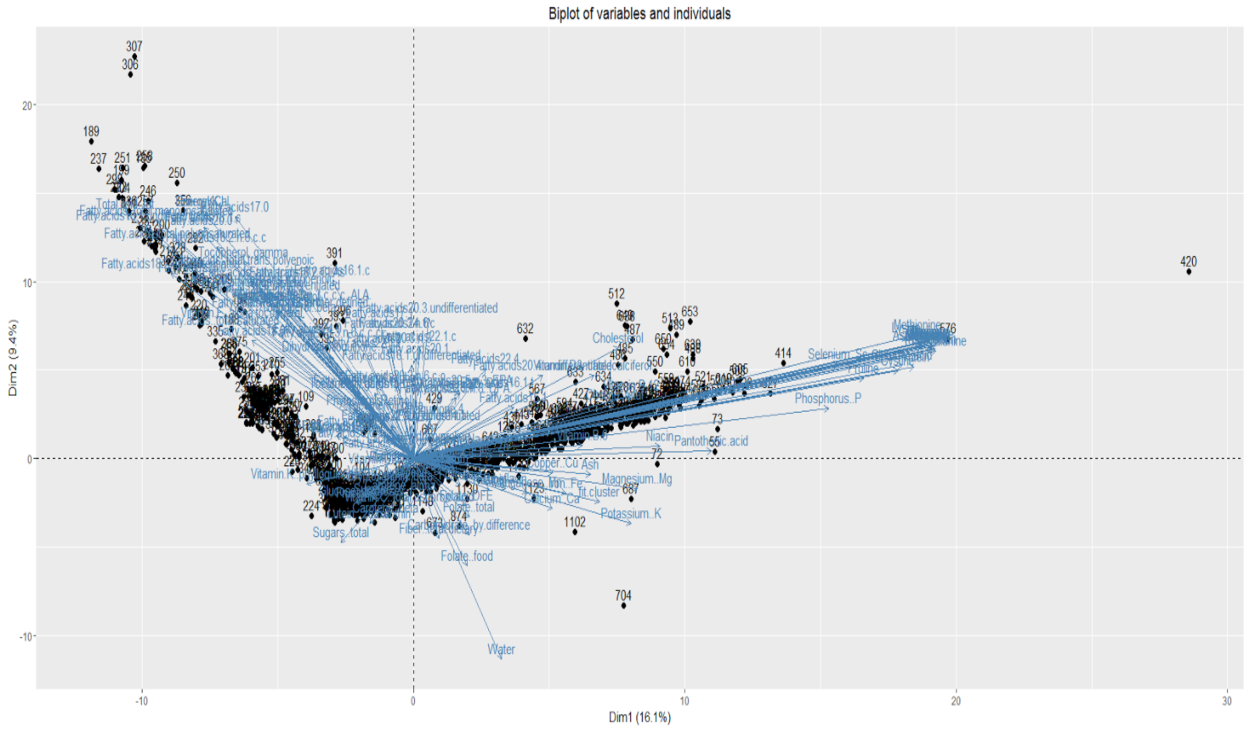


Figure 5. Biplot of the nutrient dataset

- The black dots (food names) represent the scores for the first two principal components.
- The blue arrows indicate the first two principal component loading vectors.

Next, we use Load command to interpret PCA

- Property of Loadings
  - Sum of squares within each component are eigenvalues (components' variances)
  - Coefficients in linear combination predicting a variable by standardized components

The graph below shows that oil contains high total.lipid fat as nutrition. Whereas, most contain carbohydrate by difference is grains. Both vegetable and fish contains protein.

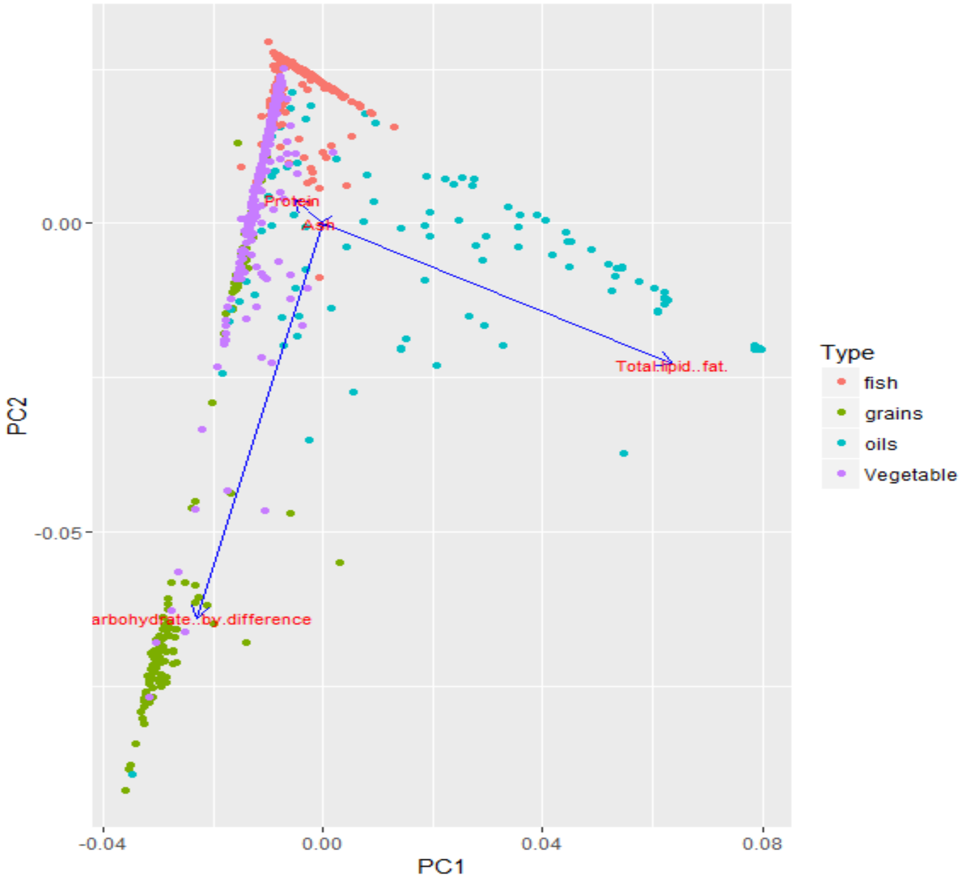


Figure 6. Principal Component Analysis of the dataset

In our project, we did 5 clusters with 5 different k values: 4, 6, 8, 10, and 12.

K = 4:

R Command: `autoplot(fanny(mydata[-5], 4), frame = TRUE)`

K = 6:

R Command: `autoplot(fanny(mydata[-5], 6), frame = TRUE)`

K = 8:

R Command: `autoplot(fanny(mydata[-5], 8), frame = TRUE)`

K = 10:

R Command: `autoplot(fanny(mydata[-5], 10), frame = TRUE)`

K = 12:

R Command: `autoplot(fanny(mydata[-5], 12), frame = TRUE)`

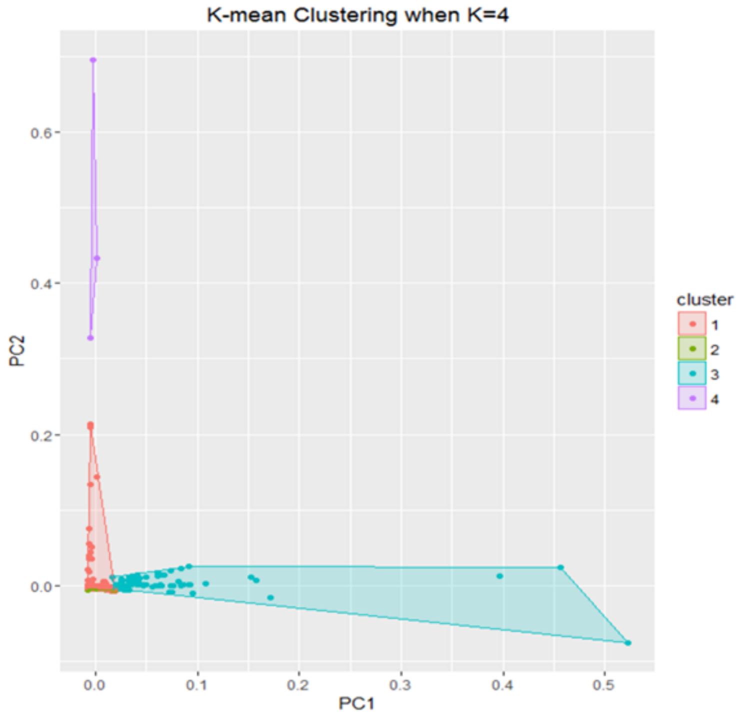


Figure 7. K-mean Clustering when K = 4

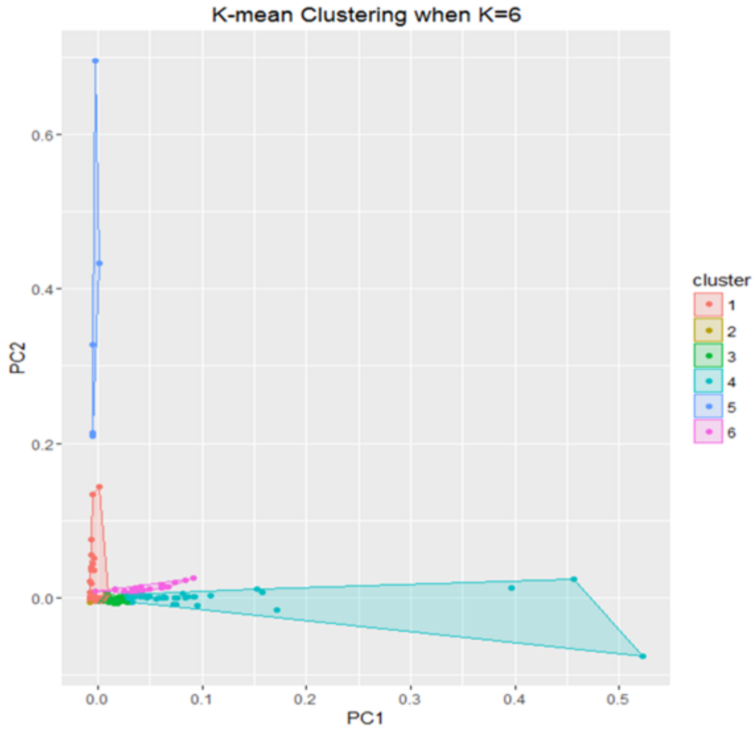


Figure 8. K-mean Clustering when K = 6

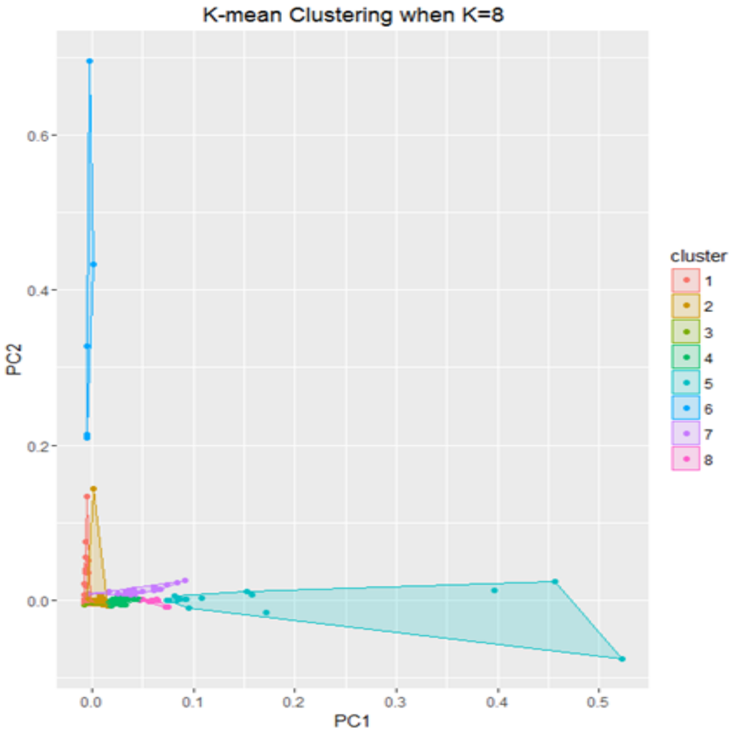


Figure 9. K-mean Clustering when K = 8

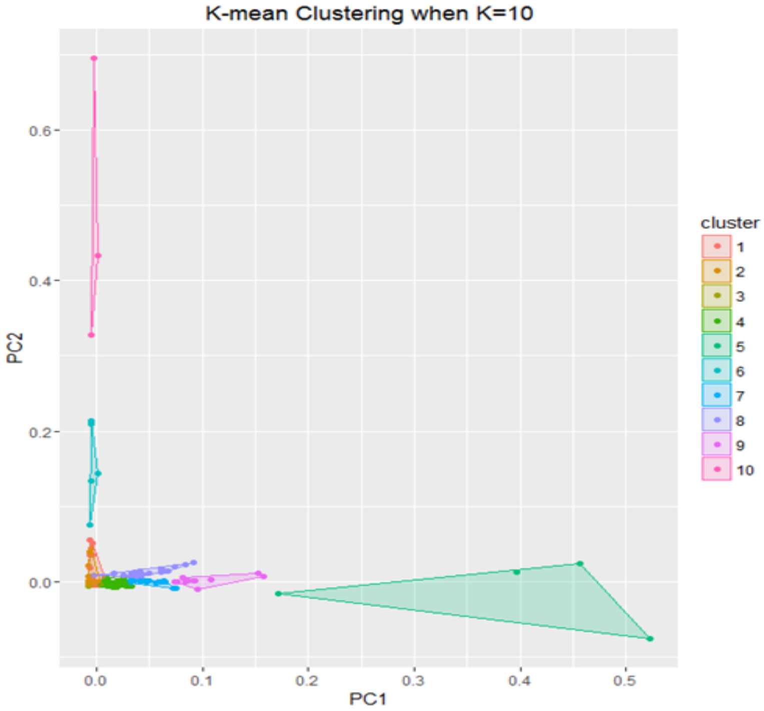


Figure 10. K-mean Clustering when K = 10

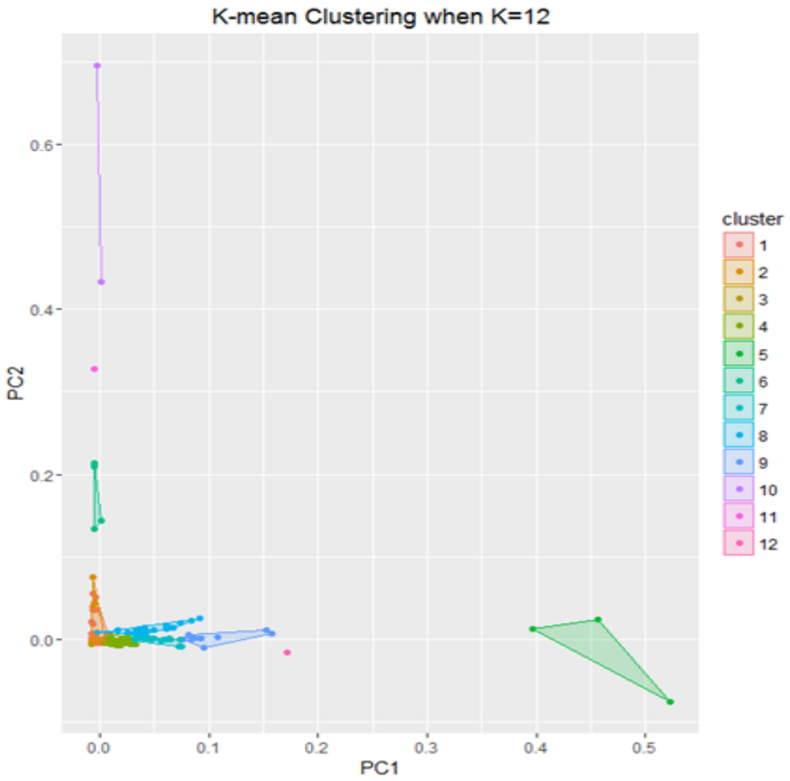


Figure 11. K-mean Clustering when K = 12

One of the process of doing K-mean clustering is to test different k-values, and pick the one with the best result. In our case, all the result from k = 4, 6, 8, 10, 12 are seems good. For each plot, all data points are well assigned to one of the k groups and labeled in different colors. We did not find any k values that would give a plot with intersect or cross over among the different groups. The reason for why our dataset can be well clustered into many different groups is: beyond the four major groups, these foods can be further clustered into subgroups. As an example, under the vegetable group, foods can be clustered into: leaves, roots, and buds base on their nutritional contents. So we could expect that the results for even higher k values would still return a good plot.

K-means clustering requires us to pre-specify the number of clusters K. This can be a disadvantage. Hierarchical clustering is an alternative approach which does not require that we commit to a particular choice of K. One approaching of doing hierarchical clustering is “bottom-up clustering”, this is the one was demoed during the lecture. There are three “Linkage” types when we were using “hclust()” command in R: “complete”, “average”,”single”.

Complete: Maximal inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the largest of these dissimilarities.

Command:

```
hc.complete = hclust(dist(x), method=”complete”)

plot(hc.complete,main=”Complete Linkage”,xlab=””,sub=””,cex=0.9)
```

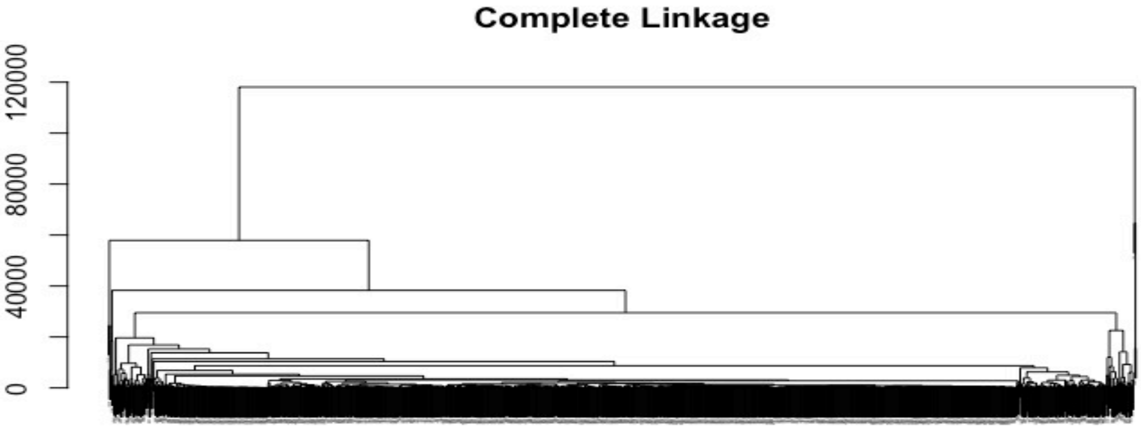


Figure 12. Complete Linkage dendrogram

Average: Mean inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the average of these dissimilarities.

Command:

```
hc.average=hclust(dist(x), method="average")

plot(hc.average, main="Average Linkage", xlab="", sub="", cex=.9)
```

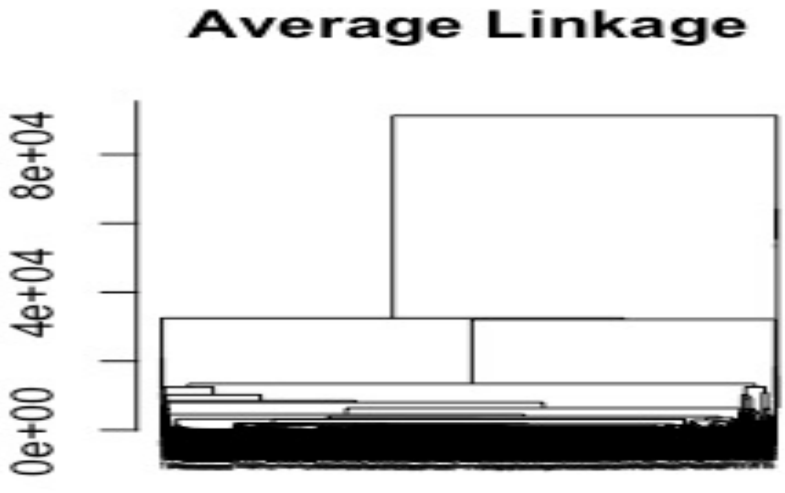


Figure 13. Average Linkage dendrogram

Single: Minimal inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the smallest of these dissimilarities.

Command:

```
hc.single=hclust(dist(x), method="single")

plot(hc.single, main="Single Linkage", xlab="", sub="", cex=.9)
```

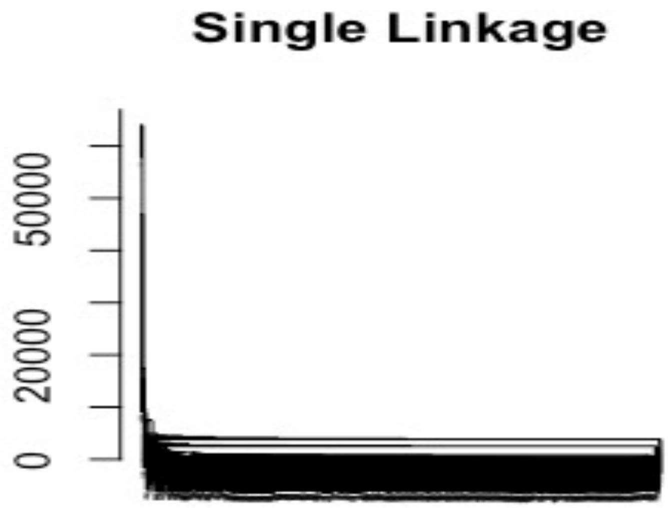


Figure 14. Single Linkage dendrogram

Our dataset was combined from four type of foods: Cereal-Grain-Pasta, Finfish-Shellfish, Vegetables, Fats-Oils. So after we were done with the Hierarchical clustering, we should expect that within a certain level in the dendrogram, our dataset would cluster into 4 groups.

Look at the dendrograms with the “Complete Linkage”, at the level 30,000 to 40,000, our dataset cluster into 4 groups as we expected. Here is an additional plot using the same dataset and “complete linkage”, after using the Scaled Features in R:

Command:

```
xsc=scale(x)

plot(hclust(dist(xsc), method="complete"), main="Hierarchical Clustering with Scaled Features")
```

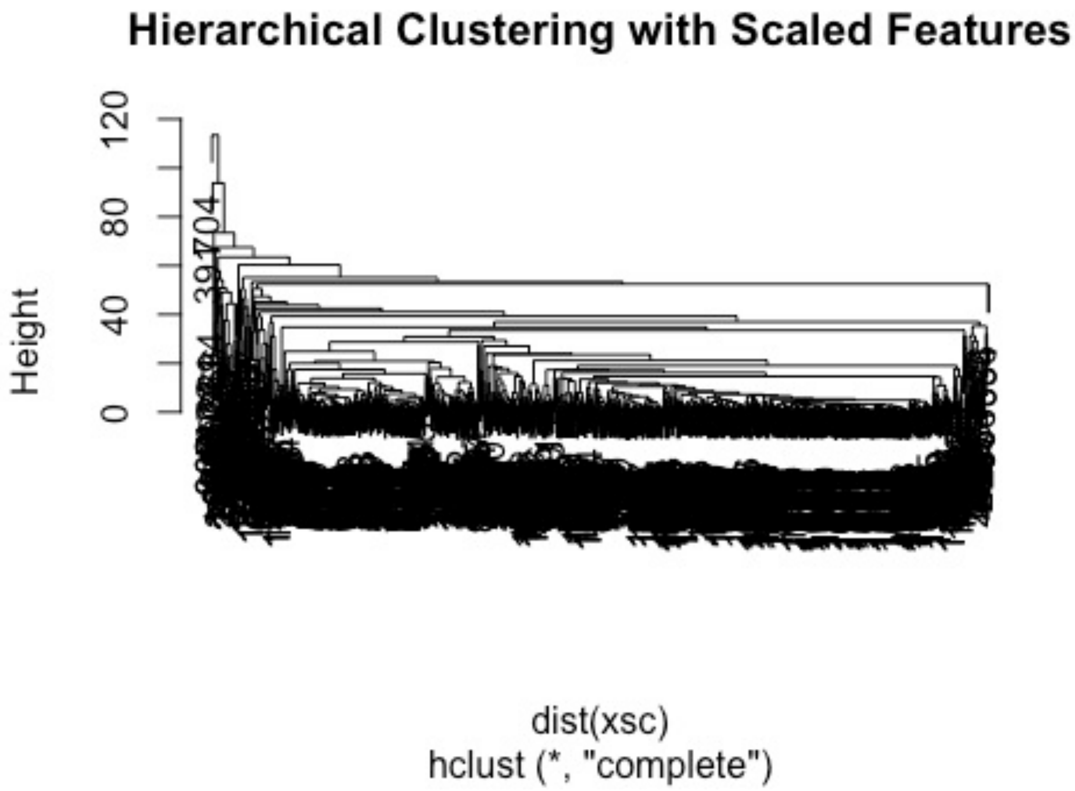


Figure 15. Hierarchical Clustering with Scaled Features

Related Work

Here is a wine clustering project that we found: https://rstudio-pubs-static.s3.amazonaws.com/33876_1d7794d9a86647ca90c4f182df93f0e8.html.

‘Wine Clustering’ project done by Gabriel Martos. His procedure was different from this project. He started by calling his data from a package called ‘rattle’. It contains 14 columns with 178 entries of data. Then, without any data wrangling nor normalization, he went straight to find K value for K-mean clustering. Then, without any explanation related to graph that he came out, he made 2D representation of K-mean clustering when k = 3. Hierarchical clustering was done after K-mean clustering. Based on his first procedure, he found out that when K=3, clustering will have best result. Thereby, he cut the tree into 3 parts. The difference is that, instead of finding best K value, this project experimented. Also, data was messy and large. Thereby, data normalization and wrangling are necessary. This project performs Principal Components Analysis before K-mean clustering. Overall, we explore data more than the author did in his project.

Conclusion

In conclusion, the four major food groups Cereal-Grain-Pasta, Finfish-Shellfish, Vegetables, and Fats-Oils can be further categorized and clustered in hierarchies. After performing the normalization and Principal Component Analysis, we test the K-mean clustering for k = 4, 6, 8, 10, 12 and all the plots return good results. We did not find any k values that would give a plot with intersect or cross over among the different groups. Since the K-means method requires us to pre-specify the number of clusters K, we avoid this disadvantage by using hierarchical clustering as an alternative approach. It has three types of linkage: complete, average, and single. The reason for why our dataset can be well clustered into many different groups is that beyond the four major groups, these foods can be further clustered into subgroups. As an example, under the vegetable group, foods can be clustered into: leaves, roots, and buds base on their nutritional contents.

Bibliography

Analytics Vidhya Content Team. (2016, March 21). Practical Guide to Principal Component Analysis (PCA) in R & Python. Retrieved December 11, 2016, from https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/

Martins, T. G. (2013, November 28). Computing and visualizing PCA in R. Retrieved December 11, 2016, from https://www.r-bloggers.com/computing-and-visualizing-pca-in-r/

Martos, G. (n.d.). Cluster Analysis with R. Retrieved December 11, 2016, from https://rstudio-pubs-static.s3.amazonaws.com/33876_1d7794d9a86647ca90c4f182df93f0e8.html

Plotting PCA (Principal Component Analysis). (n.d.). Retrieved December 10, 2016, from https://cran.r-project.org/web/packages/ggfortify/vignettes/plot_pca.html

V. (n.d.). Principal Component Analysis explained visually. Retrieved December 10, 2016, from http://setosa.io/ev/principal-component-analysis/

Zhang , Wenlu. (2016, October). data_2. Retrieved December 10, 2016, from https://www.dropbox.com/sh/820lxzd3ayuy2zu/AADtFkjMyltxx5_8R_2Kzibxa?dl=0