Food Clustering

Insun Lee, Kim Nguyen, Chao Zheng

Introduction

- Normalizing the dataset
- Principal Components Analysis (PCA)
- Apply K-means clustering
- Hierarchical clustering and dendrograms

Dataset

- USDA nutrient database- contains various types of foods with their corresponding nutritional contents
  - 4 food groups: Cereal-Grain-Pasta, Finfish-Shellfish, Vegetables, Fats-Oils.
  - Total 1164 different foods, and 151 different types of nutrients for each food.

- Many values are marked as ‘Nan’.
- Some columns contained 0 values.

Normalizing

- Numerical values vary widely across different types of nutrients.
  - Small numerical values in some micro-nutrients may characterize the food items
  - Larger numerical values do the same in macro-nutrients

- Therefore it is important to normalize the nutrient values to transform the features to be in the range[0,1]

$$\text{Normalized}(Xij) = (Xij - \min(X{\cdot}j)) \ / (\max(X{\cdot}j) - \min(X{\cdot}j))$$

Normalizing

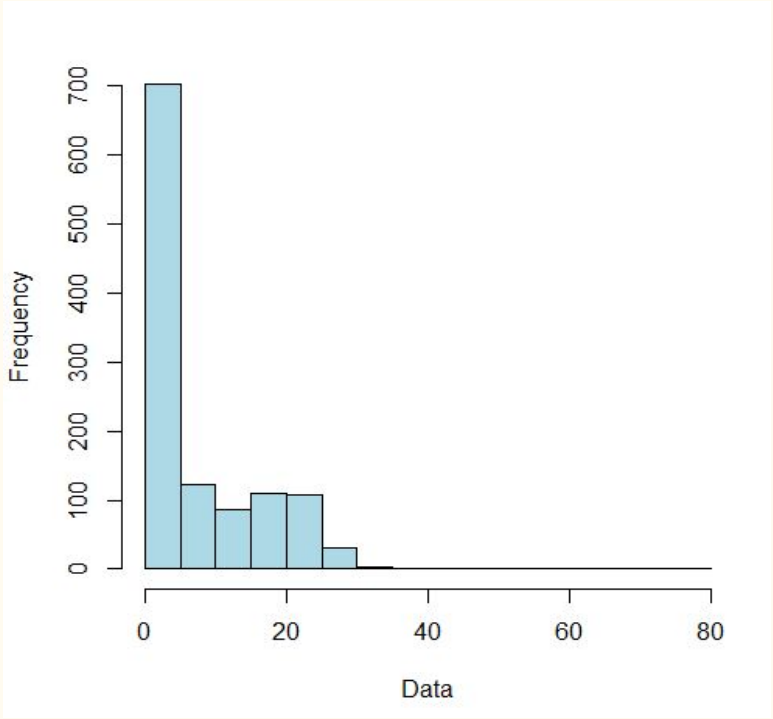
	name	Protein	Total.lipid..fat.	Carbohydrate..by.difference	Ash	Energ
1	WHEAT FLR,WHITE (INDUSTRIAL),10% PROT,BLEACHED...	9.71	1.48	76.22	0.58	
2	WHEAT FLR,WHITE,ALL-PURPOSE,UNENR	10.33	0.98	76.31	0.47	
3	MACARONI,DRY,UNENRICHED	13.04	1.51	74.67	0.88	
4	NOODLES,EGG,CKD,UNENR,W/ SALT	4.54	2.07	25.16	0.50	
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	

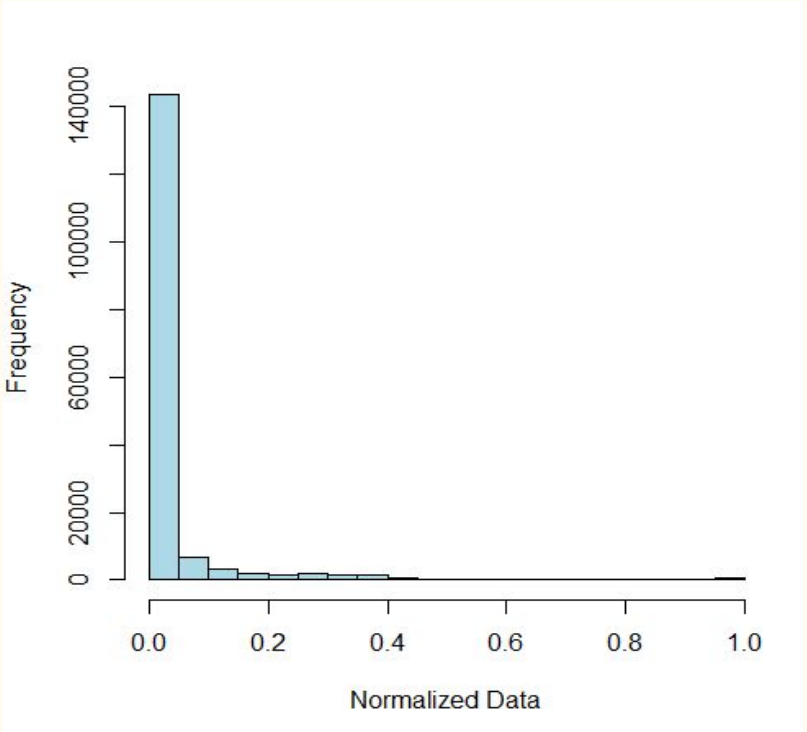
Showing 1 to 18 of 1,164 entries

	name	Protein	Total.lipid..fat.	Carbohydrate..by.difference	Ash
1	WHEAT FLR,WHITE (INDUSTRIAL),10% PROT,BLEACHED...	0.129191059	0.0148	0.8351046	0.0231
2	WHEAT FLR,WHITE,ALL-PURPOSE,UNENR	0.137440128	0.0098	0.8360907	0.0187
3	MACARONI,DRY,UNENRICHED	0.173496541	0.0151	0.8181221	0.0350
4	NOODLES,EGG,CKD,UNENR,W/ SALT	0.060404470	0.0207	0.2756656	0.0199
5	AMARANTH,UNCKD	0.180415114	0.0702	0.7149118	0.1147
6	AMARANTH GRAIN,CKD	0.050558808	0.0158	0.2047770	0.0306
7	ARROWROOT FLOUR	0.003991485	0.0010	0.9658157	0.0031
8	BARLEY,HULLED	0.166045769	0.0230	0.8050838	0.0912
9	BARLEY,PEARLED,RAW	0.131852049	0.0116	0.8515394	0.0442
10	BARLEY,PEARLED,COOKED	0.030069186	0.0044	0.3091925	0.0107
11	BUCKWHEAT	0.176290580	0.0340	0.7833899	0.0836
12	BUCKWHEAT GROATS,RSTD,DRY	0.156067057	0.0271	0.8211899	0.0876
13	BUCKWHEAT GROATS,RSTD,CKD	0.044970729	0.0062	0.2184727	0.0171
14	BUCKWHEAT FLR,WHOLE-GROAT	0.167908462	0.0310	0.7734195	0.1011
15	BULGUR,DRY	0.163517829	0.0133	0.8312699	0.0601
16	BULGUR,COOKED	0.040979244	0.0024	0.2035718	0.0135
17	CORN,YELLOW	0.125332624	0.0474	0.8136299	0.0478
18	CORN BRAN,CRUDE	0.111229377	0.0092	0.9383149	0.0143

Showing 1 to 18 of 1,164 entries

Data plots


Frequency
0
100
200
300
400
500
600
700
0
20
40
60
80
Data


Frequency
0
20000
60000
100000
140000
0.0
0.2
0.4
0.6
0.8
1.0
Normalized Data

Principal Components Analysis (PCA)

- Produces a low-dimensional representations of the variables that have maximal variance, and mutually uncorrelated
- It's also a tool for data visualization

The first principal component of a set of features:

X1, X2, . . . , Xp is the normalized linear combination of the features:

$$Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + \ldots + \phi_{p1} X_p$$

Second Principal Component

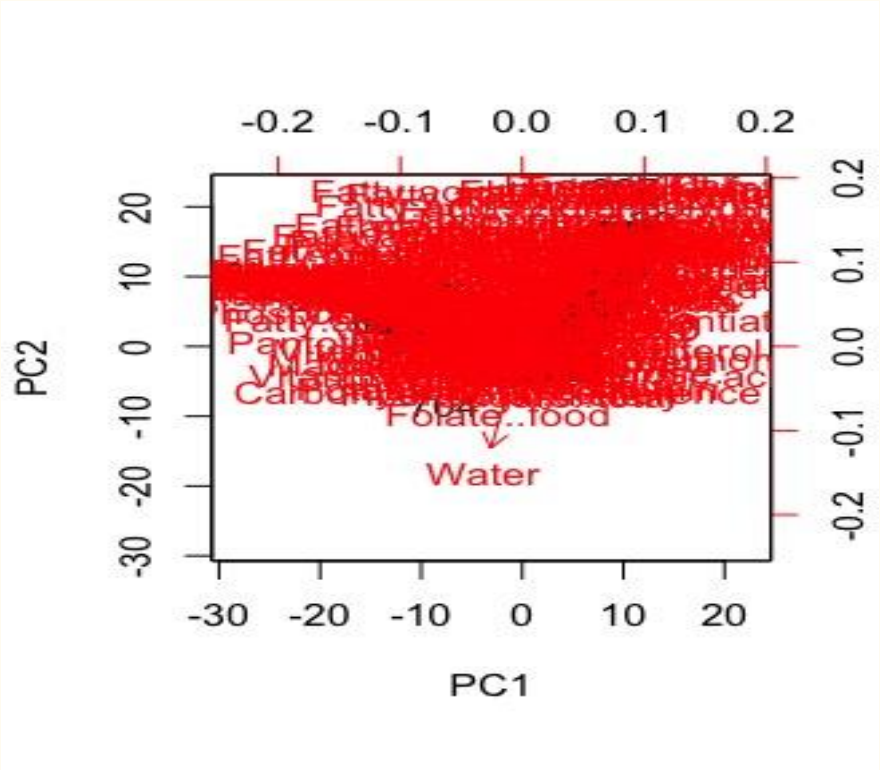
- Linear combination of X1, X2, . . . , Xp that has maximal variance among all linear combinations that are uncorrelated with Z1.

Example from our dataset:

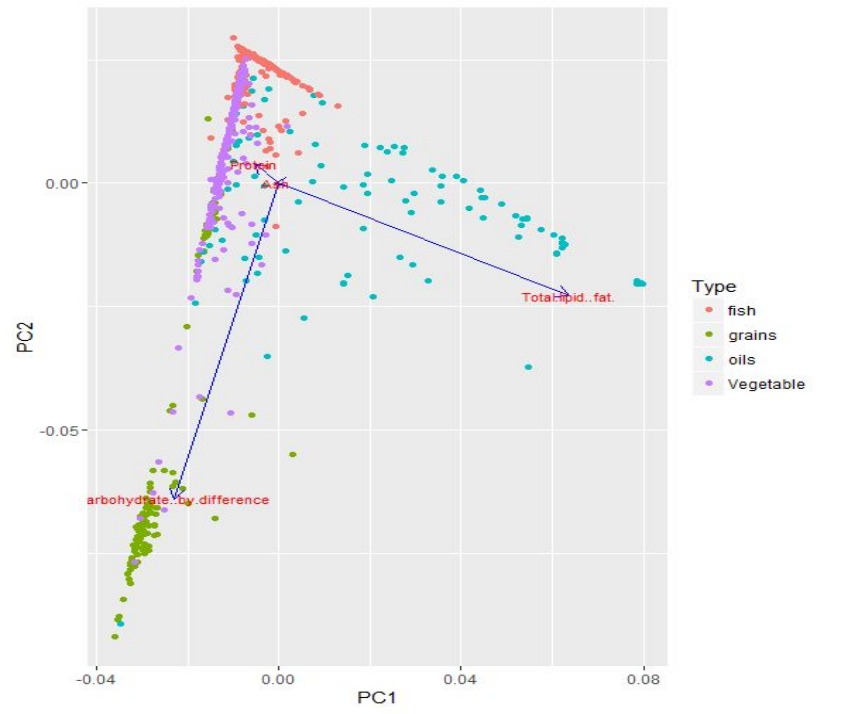
- The principal component score vectors have length n=1164(food names) - row
- The principal component loading vectors have length p = 151(nutrients) - column

Biplot

- The black food names represent the scores for the first two principal components.
- The red arrows indicate the first two principal component loading vectors.



Using 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

Proportion Variance Explained (PVE)

The total variance present in a data set is defined as:

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

and the variance explained by the mth principal component is:

$$\mathrm{Var}(Z_m) = \frac{1}{n} \sum_{i=1}^{n} z_{im}^2.$$

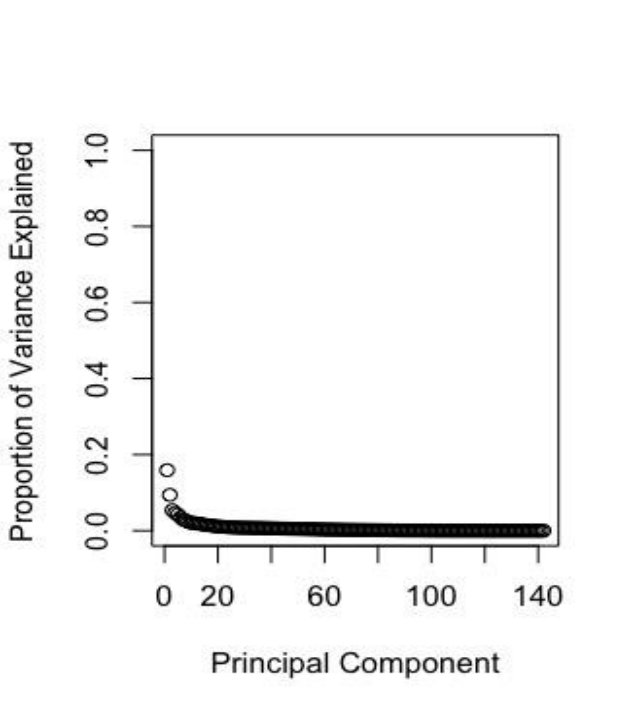
Proportion Variance Explained (PVE)

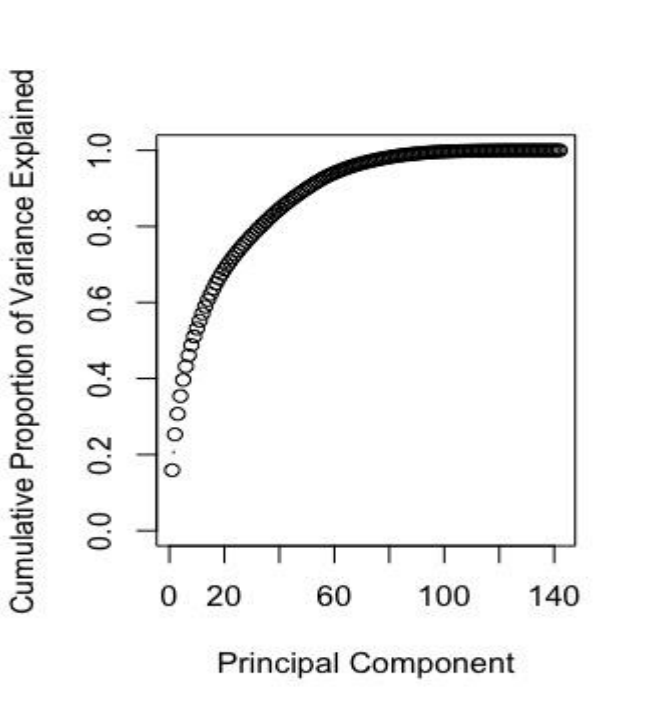
Therefore, the PVE of the mth principal component is given by the positive quantity between 0 and 1:

$$\frac{\sum_{i=1}^{n} z_{im}^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2}.$$

The PVEs sum to one. We sometimes display the cumulative PVEs.

Proportion Variance Explained (PVE)





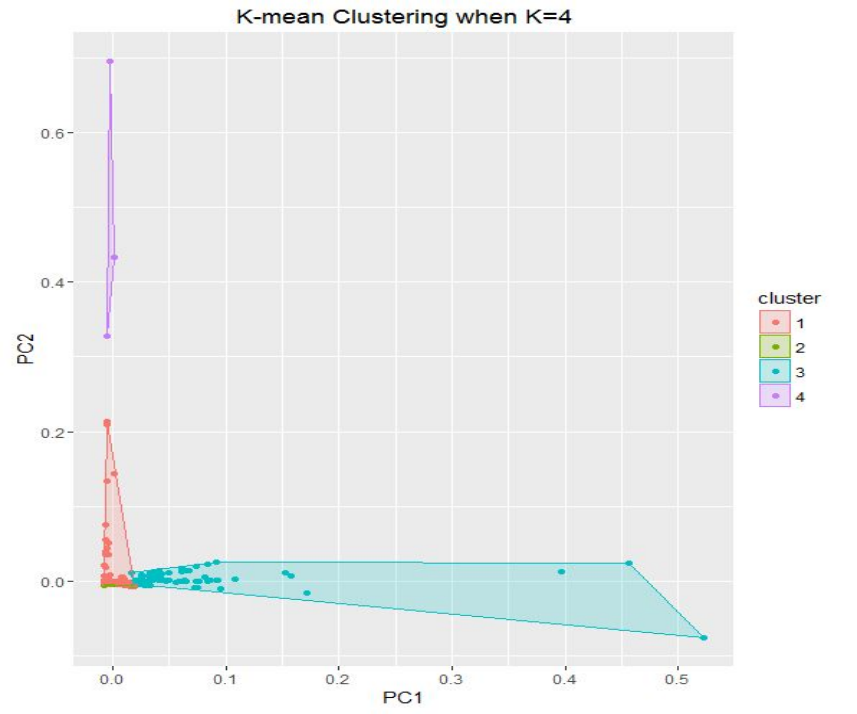
Five highest absolute weights

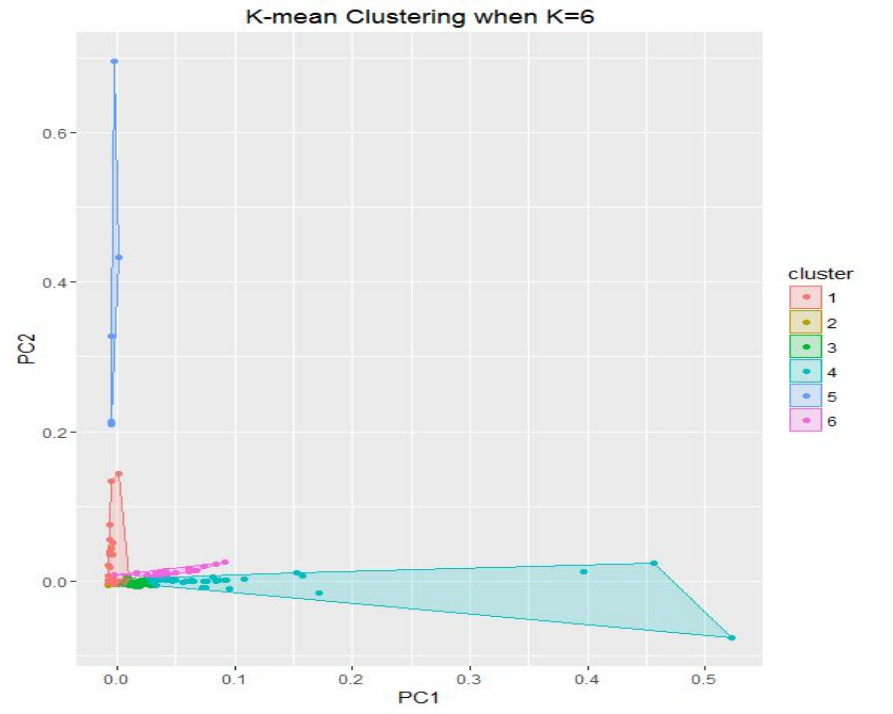
```
> topN <- 5

>pca.object <- prcomp(mydata, center = TRUE, scale.=TRUE)

> load.rot <- pca.object$rotation

> load.rot <- pr.out$rotation

> names(load.rot[,1][order(abs(load.rot[,1]),decreasing=TRUE)][1:topN])

[1] "Phenylalanine"   [2] "Serine"

[3] "Valine"

[4] "Leucine"

[5] "Isoleucine"
```

K-mean

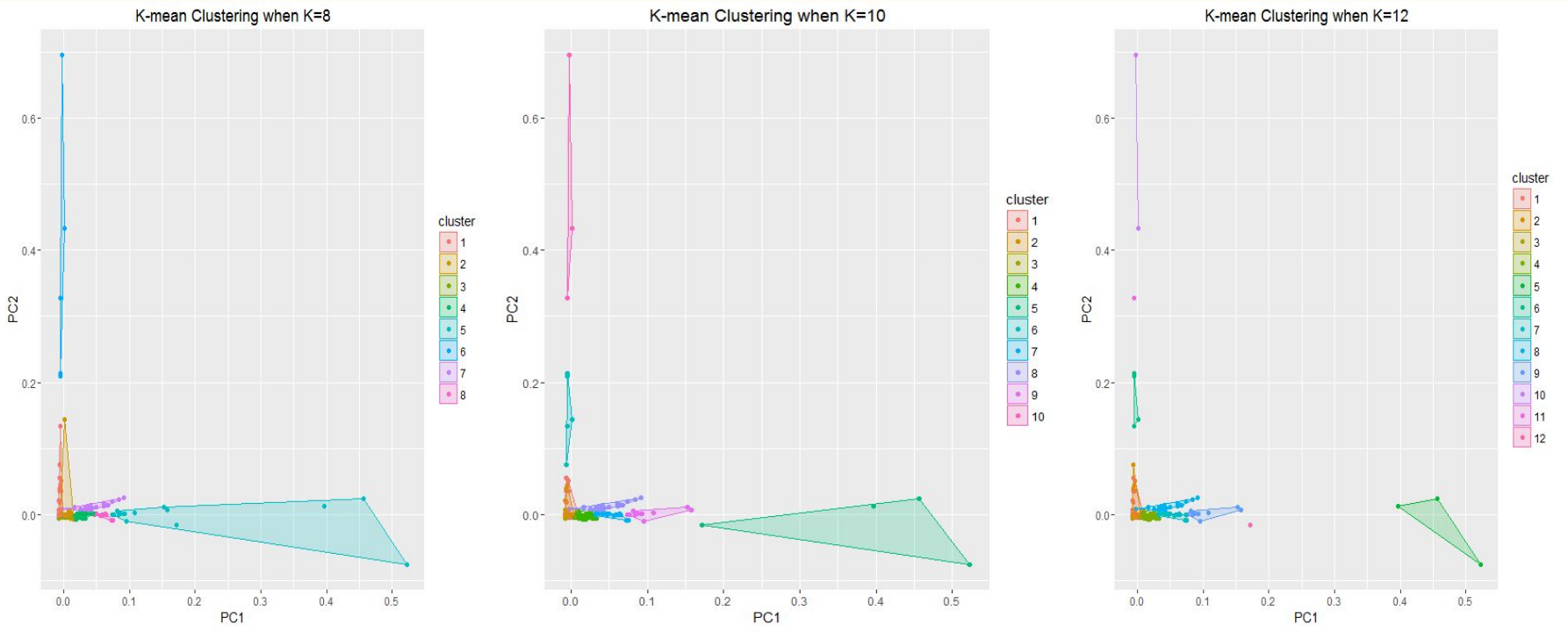
1. Randomly pick k centroids (centers of clusters)
2. Assign each data point to the closest centroid.
3. Recompute cluster centroids (average location of data points) in light of current cluster assignments.
4. Repeat Steps 2 and 3 until assignments don’t change or change very little.

K-mean


K-mean Clustering when K=4
PC2
PC1
0.0
0.2
0.4
0.6
0.1
0.3
0.5
cluster
1
2
3
4


K-mean Clustering when K=6
PC2
PC1
0.0
0.2
0.4
0.6
0.1
0.3
0.5
cluster
1
2
3
4
5
6

K-mean


K-mean Clustering when K=8
PC2
PC1
cluster
1
2
3
4
5
6
7
8
K-mean Clustering when K=10
PC2
PC1
cluster
1
2
3
4
5
6
7
8
9
10
K-mean Clustering when K=12
PC2
PC1
cluster
1
2
3
4
5
6
7
8
9
10
11
12

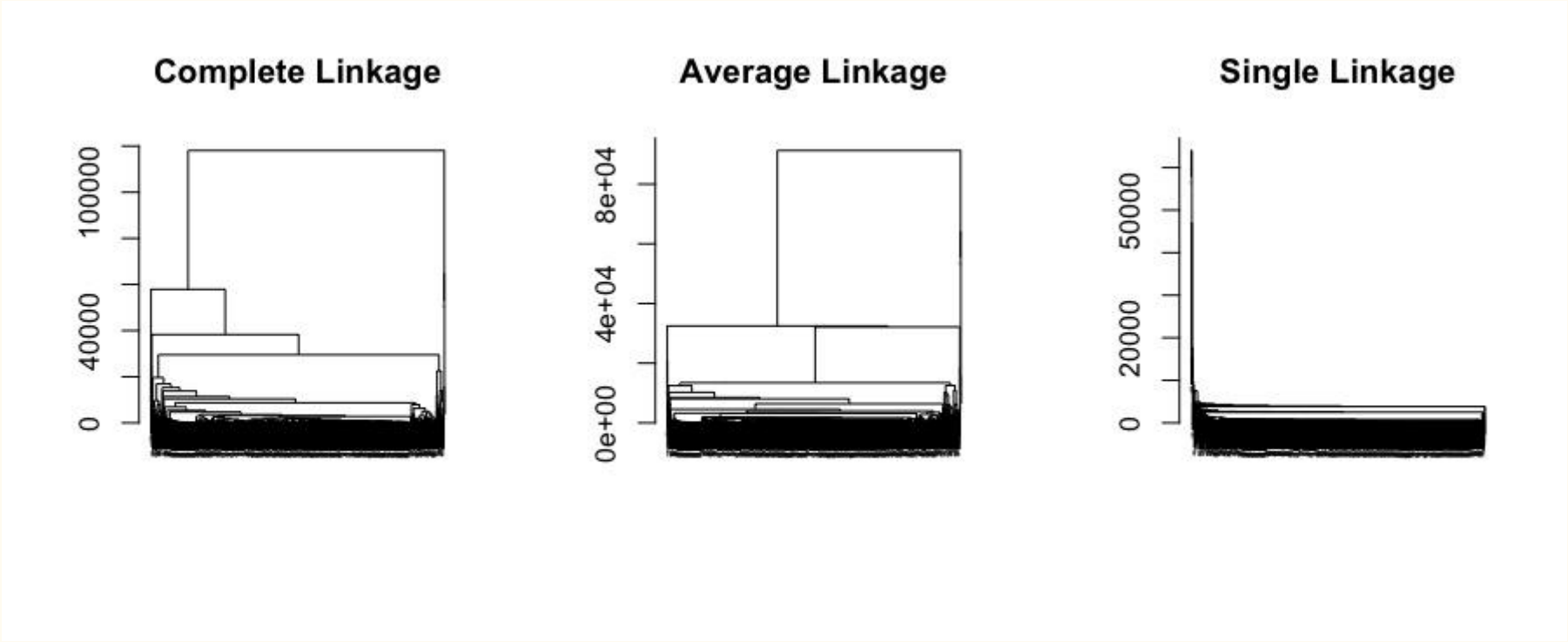
Hierarchical clustering

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.

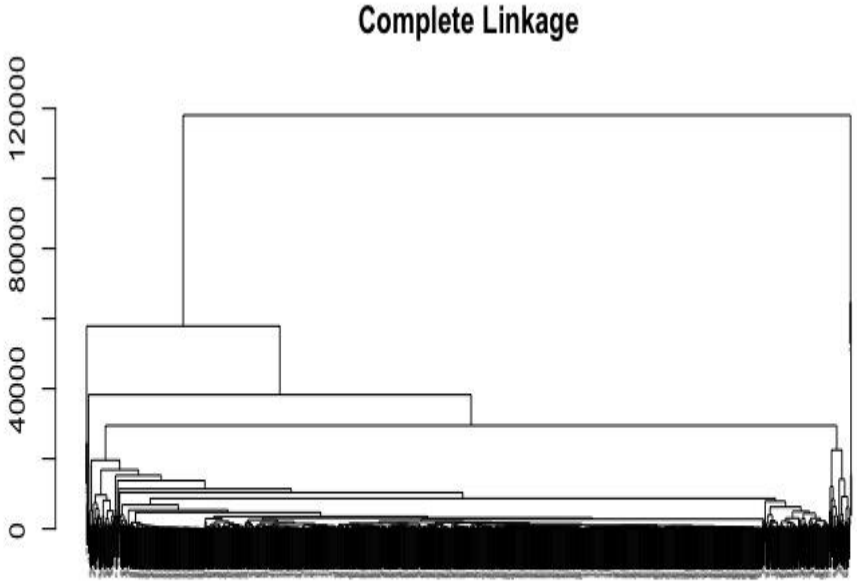
- bottom-up clustering.

Dendrograms



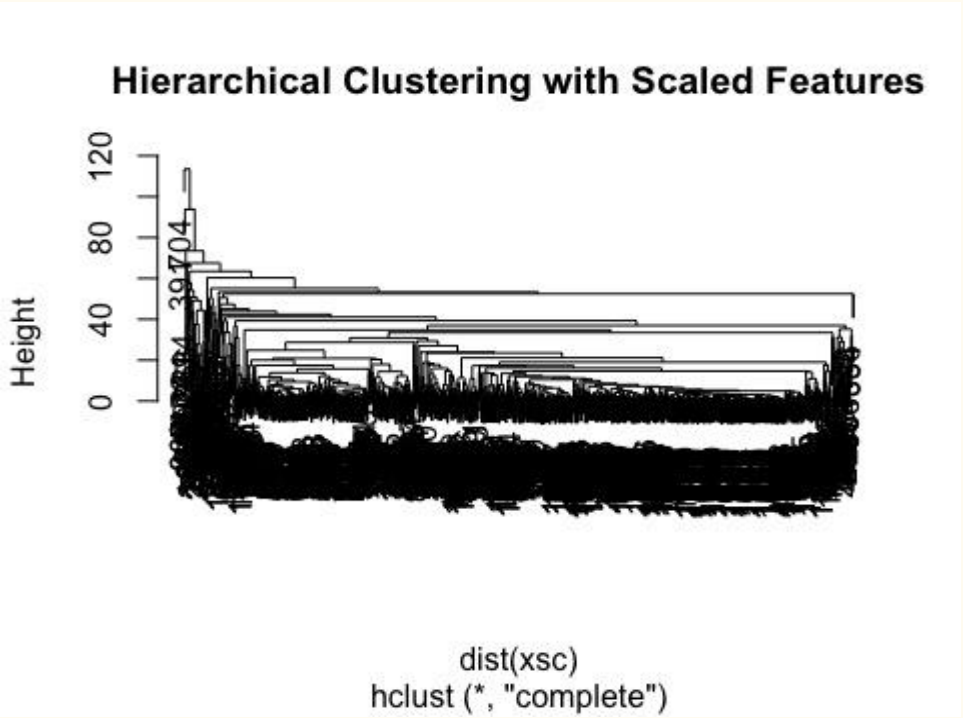
Dendrograms

Remind: there are four types food in our dataset.



Dendrograms

With Complete linkage and Scaled Features.



Any Question?