Untitled

Andrew Estes

1/16/2022

Part 1

First, install (if necessary) and load the package cluster.datasets, along with tidyverse. Load the data set with data("nutrients.meat.fish.fowl.1959"). *Take a look at it, and read the documentation with ?nutrients.meat.fish.fowl.1959* Then make a copy of the data set. The clustering commands prefer labels to be row names, not specified in an actual column, so we'll do that too. Finally, nutrients are measured on different scales, so we'll rescale the variables. * meat <- nutrients.meat.fish.fowl.1959 %>% column_to_rownames("name") %>% scale()

```
library(tidyverse)
#install.packages("cluster.datasets")
library(cluster.datasets)
#?nutrients.meat.fish.fowl.1959
data("nutrients.meat.fish.fowl.1959")
df <- data.frame(nutrients.meat.fish.fowl.1959) %>%
  column_to_rownames("name") %>%
  scale()
```

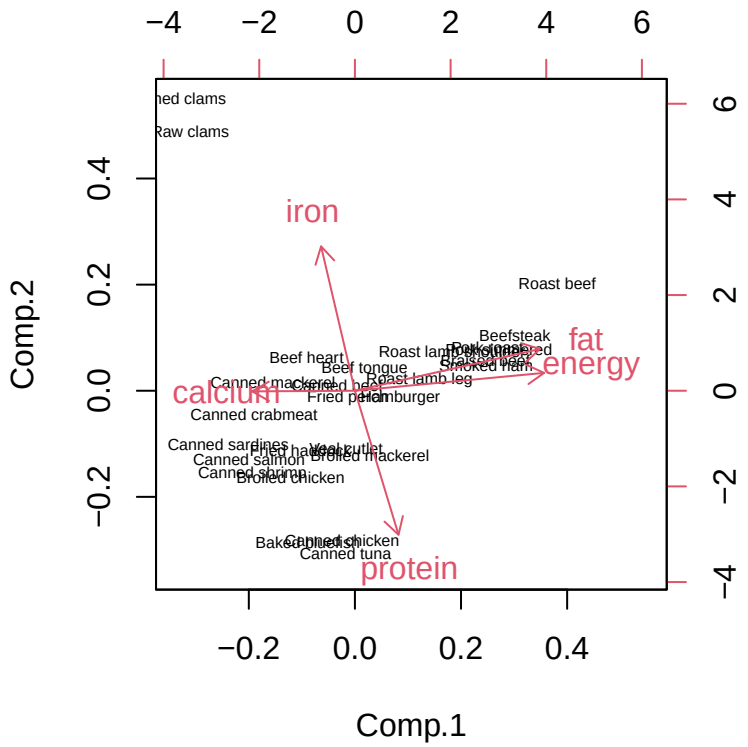
Part 2

Now, let's take a look at the data by using principal components. Using princomp and biplot, what can you say about the real-world meanings of the first two principal components? Which original variables seem positively or negatively correlated with each other? Do any natural groupings or patterns show up? Note: You can use the fig.width and fig.height chunk options to make the text on your graphs smaller or larger compared to the graph itself, in order to help readability. Another way is to change the cex option in the biplot command.

```
pca_df <- princomp(df)
summary(pca_df)
```

```
## Importance of components:
##                            Comp.1    Comp.2    Comp.3    Comp.4       Comp.5
## Standard deviation     1.4542964 1.0504520 0.9035024 0.8823425 0.0392724977
## Proportion of Variance 0.4392647 0.2291780 0.1695427 0.1616943 0.0003203299
## Cumulative Proportion  0.4392647 0.6684426 0.8379853 0.9996797 1.0000000000
```

```
biplot(pca_df, cex=c(.5, 1))
```



The first two components make up roughly 67% of the variance.

Calcium, Iron, Energy, Fat, Protein were highlighted. Fat and Energy were very closely related which makes sense knowing that 1 gram of fat is equivalent to 9 kCalories (or energy) whereas Protein is only 4 kCalories. Calcium was in direct oppoisition to Fat and Energy. Iron was in direct opposition to Protein.

Clams were grouped together, very far away from the other food types. The rest of the food items followed a slightly positive, slightly linear relationship going from seafood to chicken to red meat.

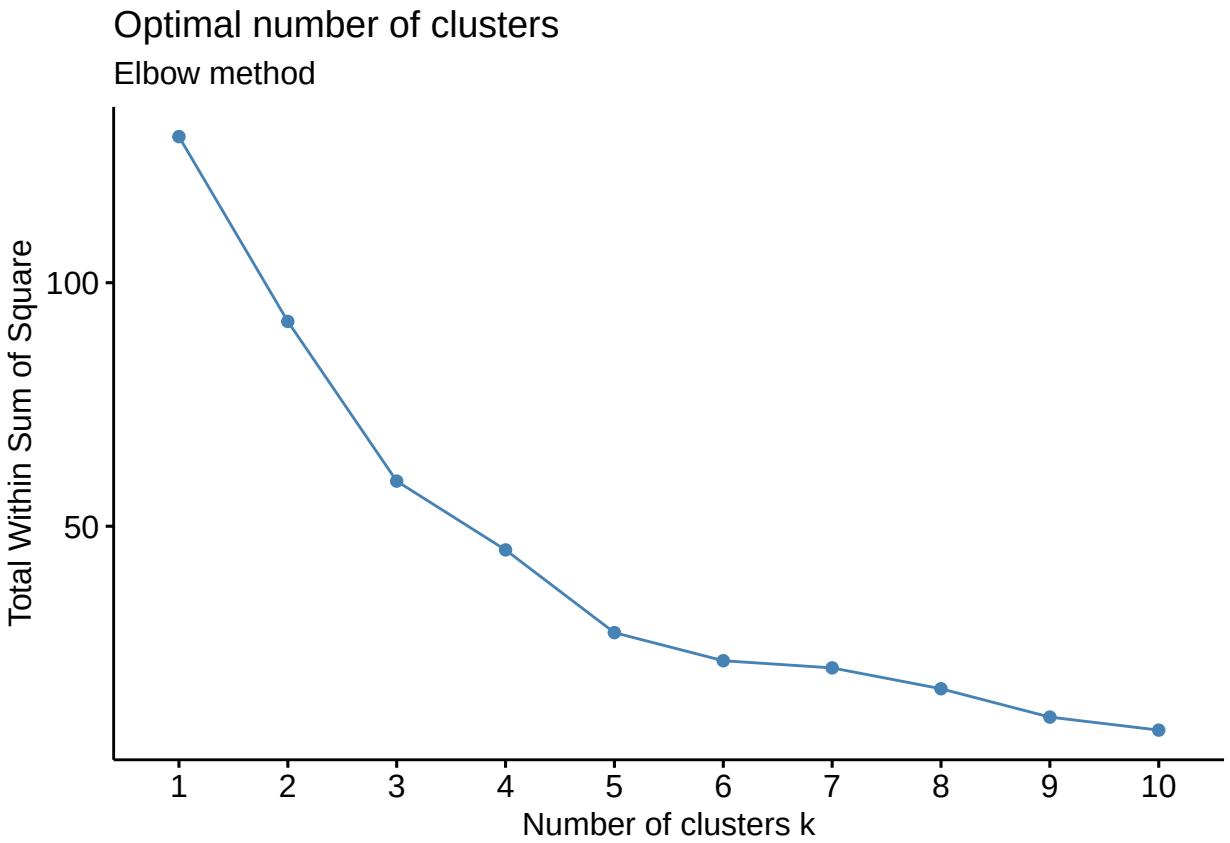
Part 3

Use either the wssplot from the notes or the command fviz_nbclust from the factoextra package to get a sense for an optimal number of clusters within the data set. Does this match any intuition you gained from the PCA?

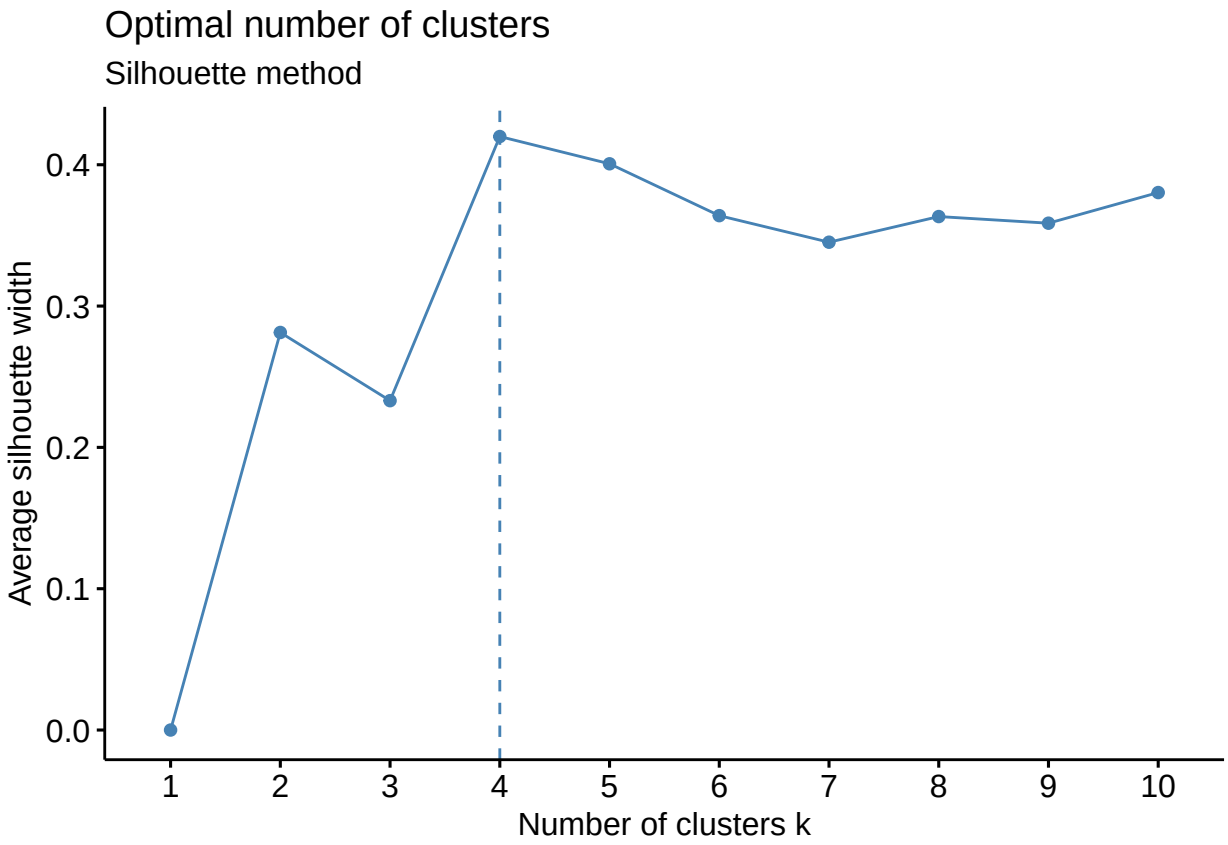
Use either the wssplot from the notes or the command fviz_nbclust from the factoextra package to get a sense for an optimal number of clusters within the data set. Does this match any intuition you gained from the PCA?

```
library(factoextra)

#http://www.sthda.com/english/articles/29-cluster-validation-essentials/96-determiningthe-optimal-numbe
# Elbow method
fviz_nbclust(df, kmeans, method = "wss") +
  labs(subtitle = "Elbow method")
```




```
# Silhouette method
fviz_nbclust(df, kmeans, method = "silhouette")+
  labs(subtitle = "Silhouette method")
```

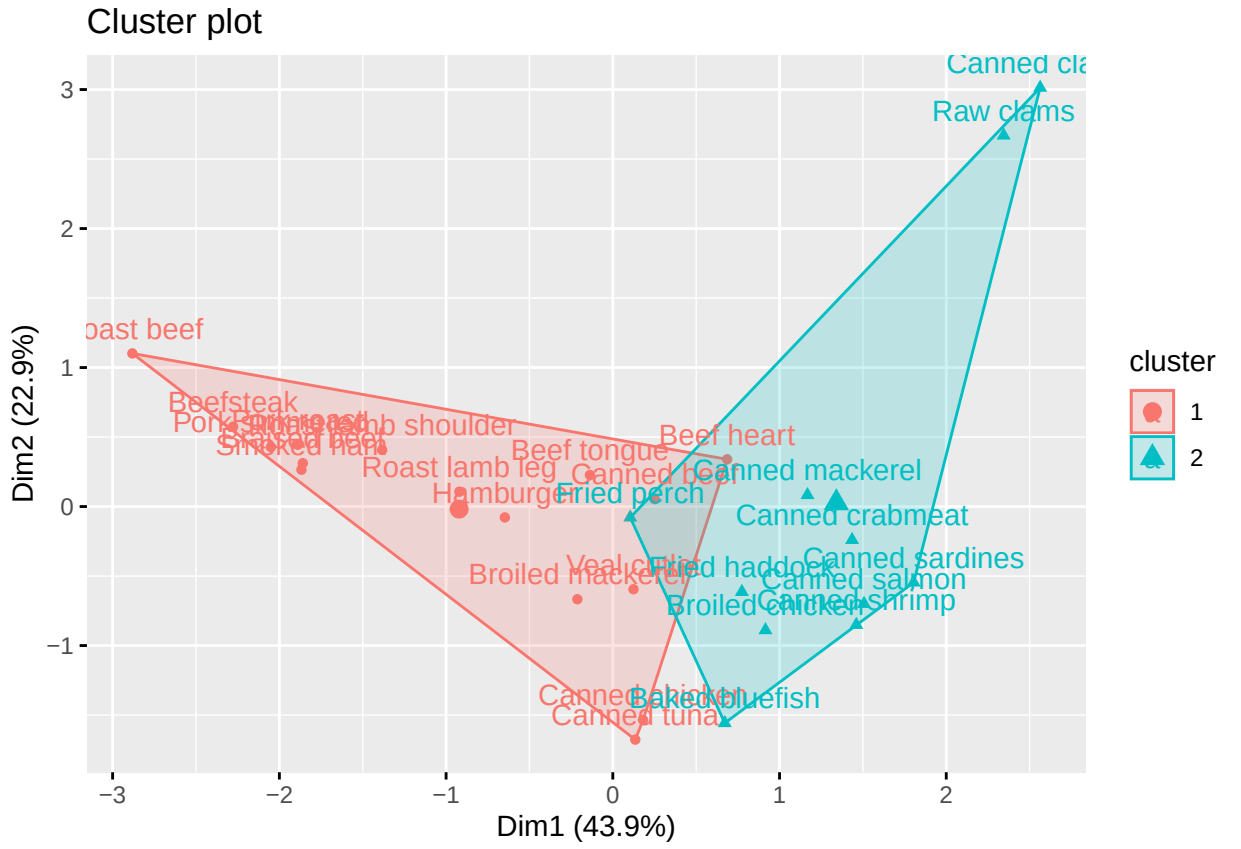


The PCA made it pretty clear that 4 was the ideal number of components to include, with a variance explanation of 99%. Based off the PCA, 3 variables explained 83% of the variance and 2 variables explained 67%. Depending on my goals, I would choose 2, 3, or 4 predictors for the model.

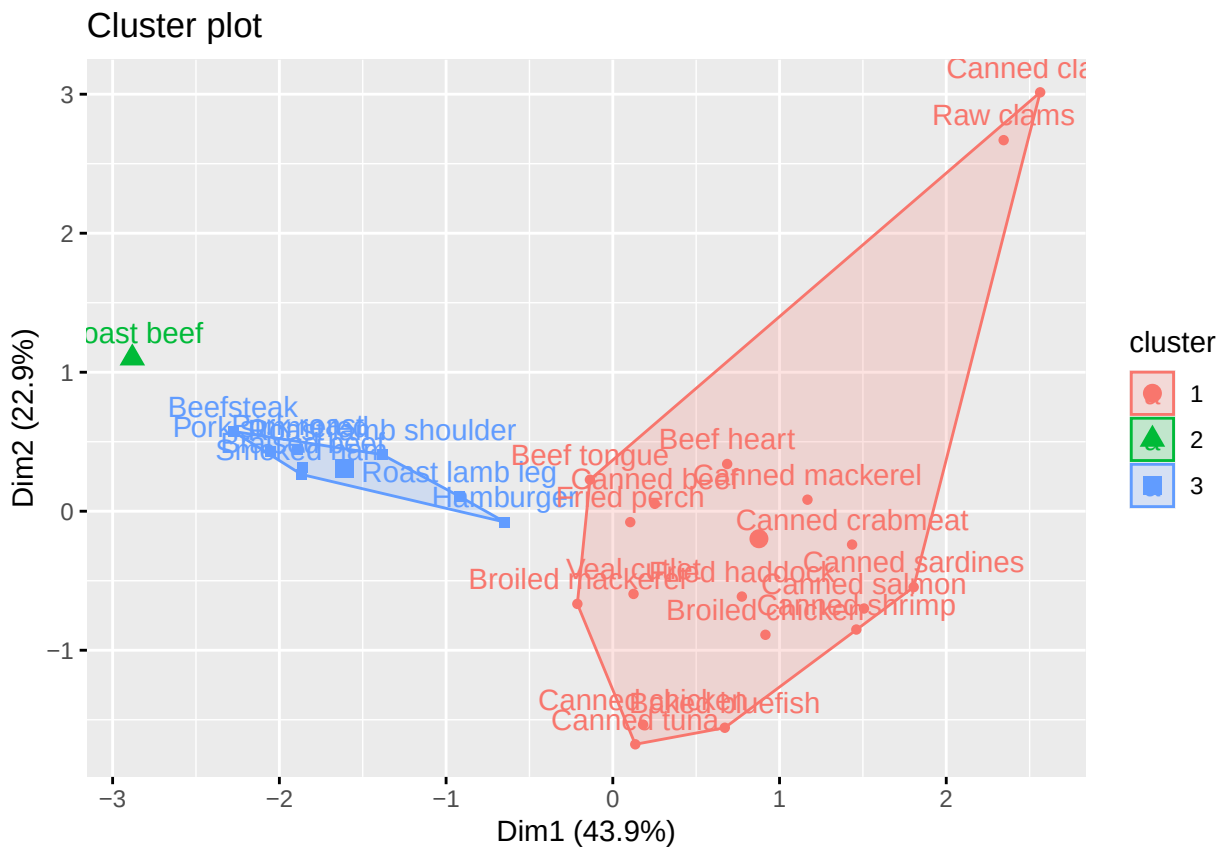
Part 4

Run k-means clustering with your data and the number of clusters you decided on. Plot the resulting clusters, with points color-coded, either "by hand" using the plot command or with the fviz_cluster command.

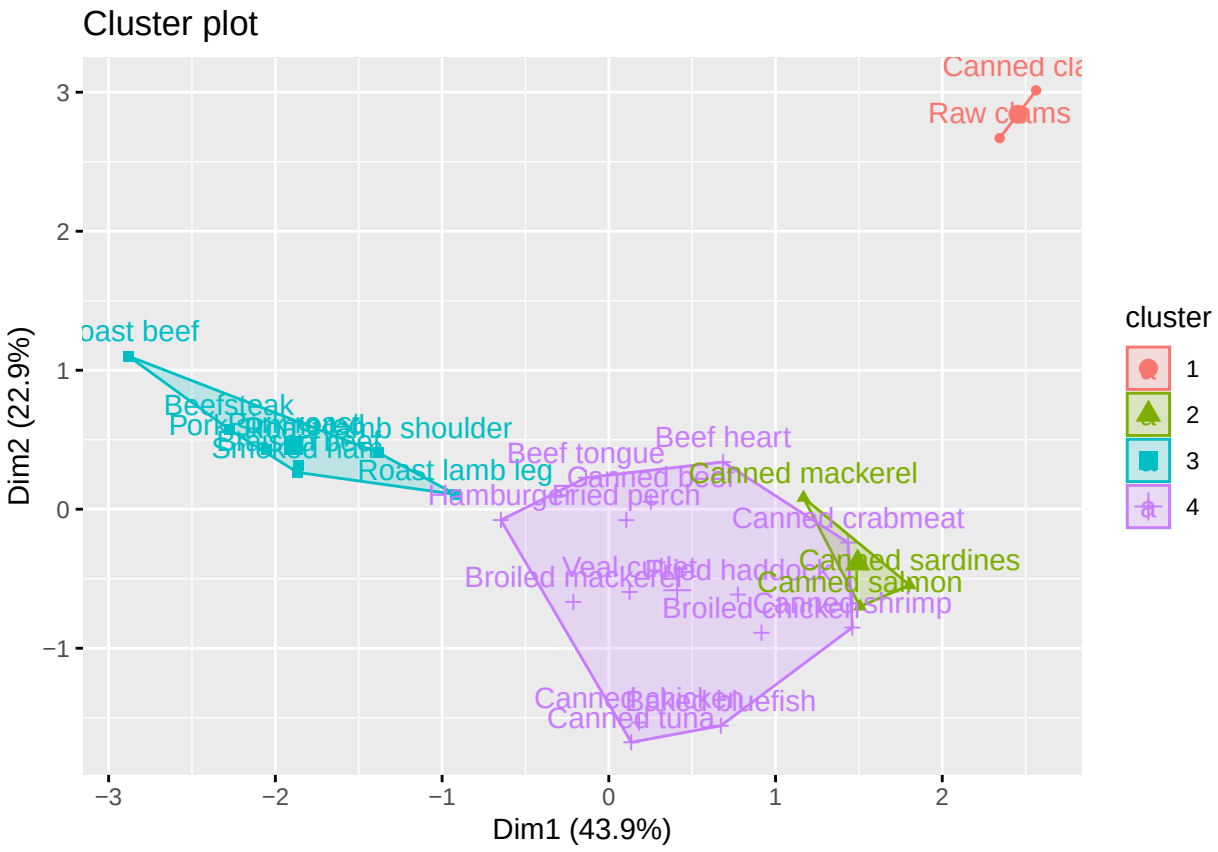
```
df.kmeans2 <- kmeans(df, centers = 2)
fviz_cluster(df.kmeans2, df)
```




```
df.kmeans3 <- kmeans(df, centers = 3)
fviz_cluster(df.kmeans3, df)
```



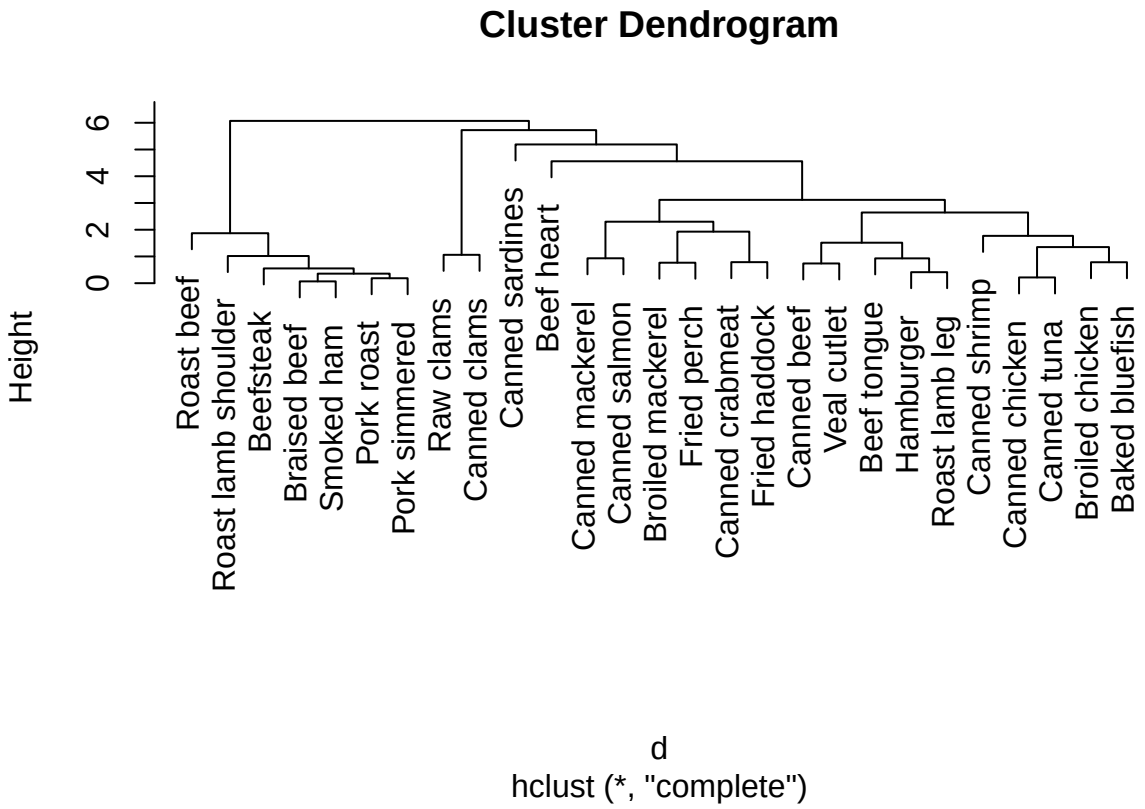

```
df.kmeans4 <- kmeans(df, centers = 4)
fviz_cluster(df.kmeans4, df)
```


Cluster plot
Canned cla
Raw clams
oast beef
Beefsteak
Roast lamb shoulder
Roast lamb leg
Beef tongue
Beef heart
Canned mackerel
Hamburger
Fried perch
Canned crabmeat
Canned sardines
Canned salmon
Broiled mackerel
Broiled haddock
Canned shrimp
Canned tuna
Baked bluefish
Dim2 (22.9%)
Dim1 (43.9%)
cluster
1
2
3
4

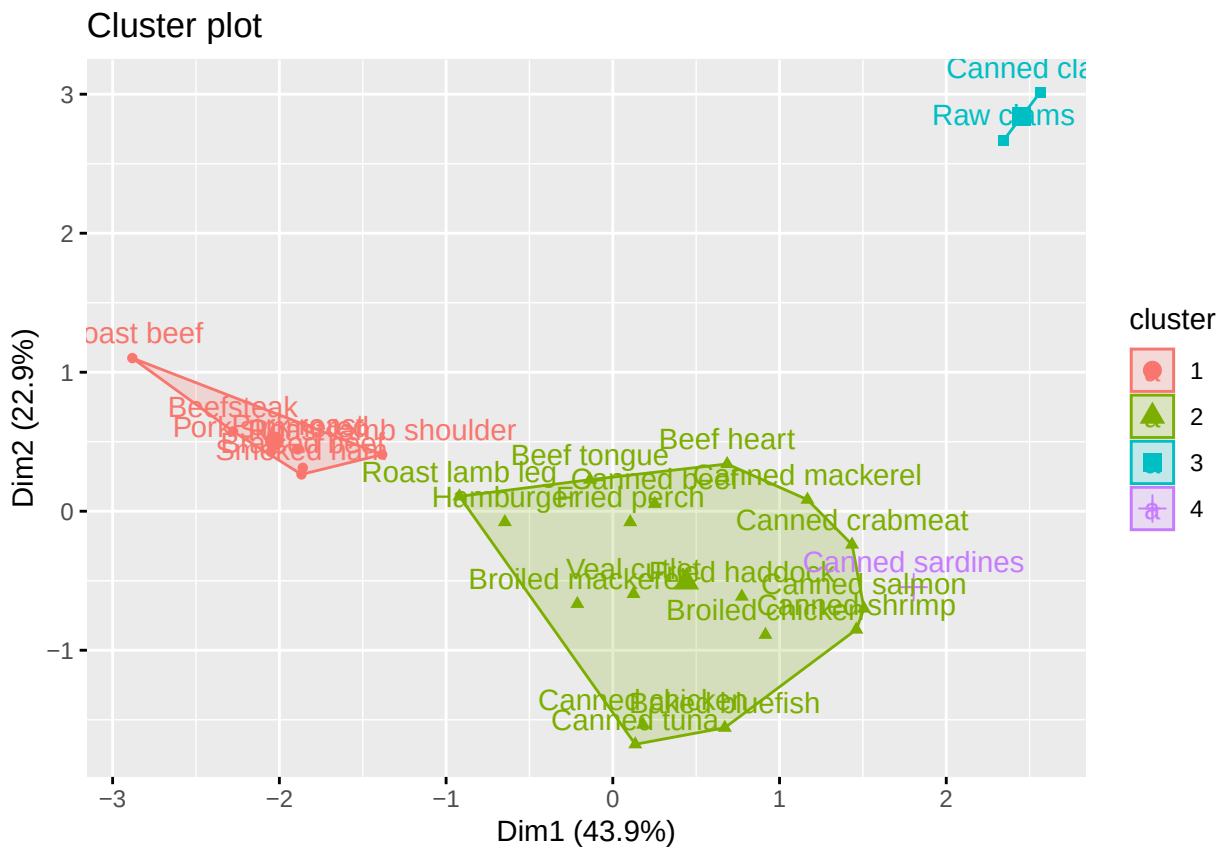
Part 5

Run a hierarchical clustering on your data. Plot the dendrogram, and make another scatter plot with points color-coded into the same number of clusters as the kmeans clustering used. Do the clusters make sense? How would you describe the nature of the various clusters in plain language?

```
d <- dist(df, method="euclidean")
df.cluster <- hclust(d, method="complete")
plot(df.cluster)
```




```
sub_grp4 <- cutree(df.cluster, k = 4)
fviz_cluster(list(data=df, cluster=sub_grp4))
```



Plotting components 2, 3, and 4 shows that "canned sardines" and "beef heart" play an important role. They are the 3rd cluster in the 3-component kmeans and they are the 4th cluster in the 4-component kmeans. In the dendrogam, they are in a tier separated greatly from the other meat, fish, and fowl foods.