Sean Dube
Project 3 Report
ID: 200037417
Group 3

Is it possible to predict food diets from micro/macro nutrients?

Intro:

Nutrition is a topic of recent interest as many are investigating different diets to accommodate their health needs. For example, a diet of recent popularity is the plant-based diet due to its health benefits such as increased energy, lower cholesterol, and lower fat intake. Most individuals have grown up to learn that government issued food pyramids represent the most ideal diet for humans. Researchers have recently come up with alternative versions for healthier diets, as they view the authorities' version as a business model to encourage consumers. This research will look into food groups and try to determine if there are certain food groups associated with a specific diet. The goal in this analysis is to use multivariate analysis, more specifically principal component analysis and cluster analysis to identify these groups. This research can be beneficial for those wanting to gather an understanding of what diets are associated with certain food groups and also which food groups are associated with the healthier nutritious value.

Data:

The USDA National Nutrient DataBase (Release 27) is a comprehensive dataset that encompasses information about various foods and their nutritional values (Kelly, 2016). The data consists of 8618 entries of 45 variables, including both categorical and numerical information. The character variables include the type of food, food group out of 25 categorical options, description, common name, scientific name, and manufacturer. The original data set contains all nutrient values except from starch, fluoride, betaine, vitamin D2 and D3, added vitamin E, added vitamin B12, alcohol, caffeine, theobromine, phytosterols, individual amino acids, individual fatty acids, or individual sugars, please see the bibliography for the original data set (US Department of Agriculture, Agricultural Research Service, 2016. The numerical variables comprise of 38 nutritional value measurements for 100g of food, such as kilocalories, grams of carbs, fat, protein, sugar, and fiber, as well as various nutrients in milligrams, micrograms, or as a percentage– for example, 0.4 is 40%– of the U.S. Recommended Daily Allowance. It is worth mentioning that some RDA were eliminated in the wrangled dataset– for example, the US RDA for Iron varies based on age and sex– thus, it was omitted for simplicity.

Methods:

First step to this analysis is to perform basic exploratory analysis to get to know the data. Since we plan to implement PCA, the most helpful exploration is into the correlations between all the attributes.

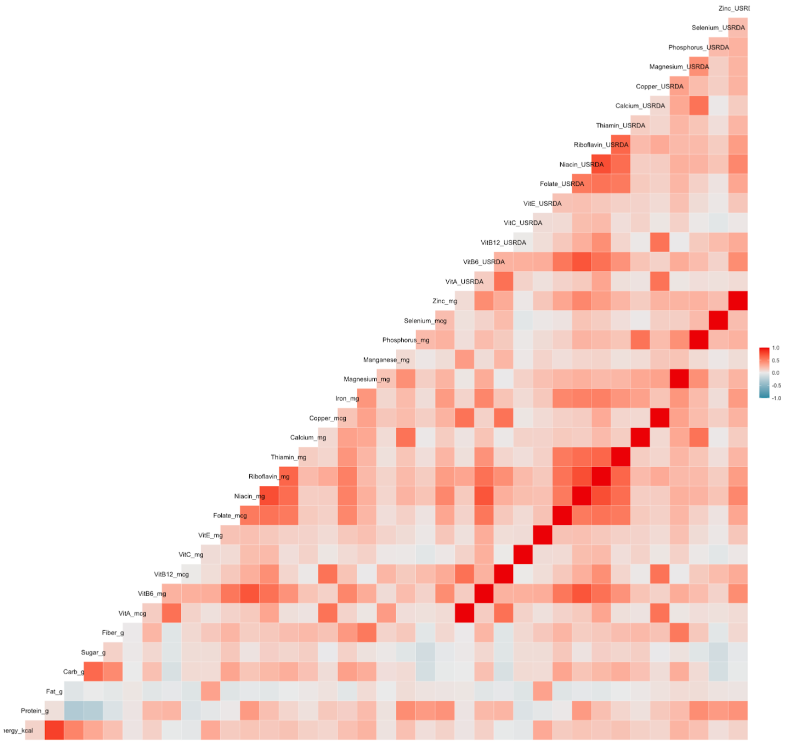


Figure 1: Correlation Matrix of all numerical attributes in the Nutrition Dataset

Figure 1 displays a correlation plot of all the attributes in the dataset. Right away, we notice heavy correlations between many of the attributes. Since, this is a high dimensional dataset with high correlations between attributes, there is a case to conduct PCA. Principal Component Analysis (PCA) is a statistical technique used to transform a set of variables, often correlated, into a smaller set of uncorrelated variables called principal components. The goal of PCA is to reduce the dimensionality of a dataset while retaining as much of the original variation as possible. In order to determine how many components to reduce to, we can look at scree plots and cumulative proportion of variance explained.

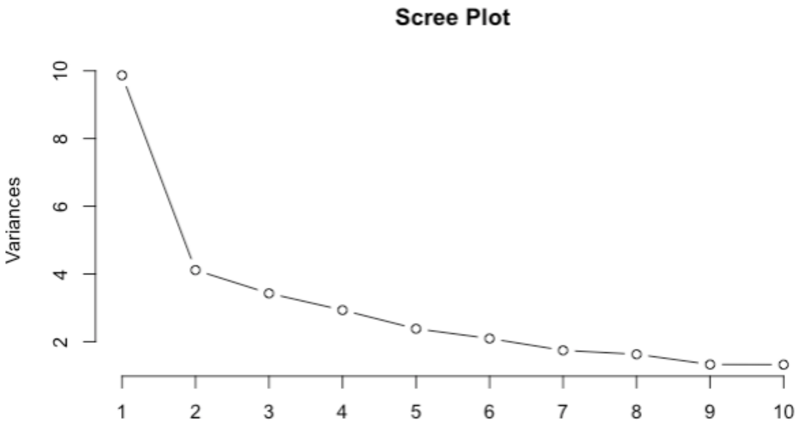


Figure 2: Scree Plot showing eigenvalues of the principal components on the y-axis and the component number on the x-axis

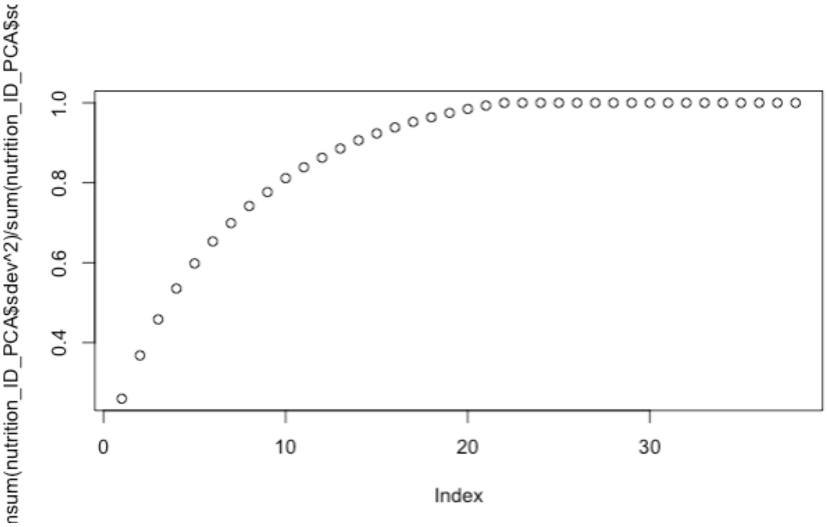


Figure 3: Plot showing the relationship of the cumulative proportion of variance explained vs principal component index

Figure 2 shows us that the ideal number of PCA's to retain is between 4-5 (point at which the line levels) and *Figure 3* shows that in order to keep more than 60% of the variability from the original dataset, we must retain at least 5 components. After reducing the scaling and reducing the dimensions to the dataset, the next step is to conduct clustering analysis. Next, we perform cluster analysis, the first method being K-means. K-means is a centroid-based clustering algorithm that partitions data into a predetermined number of clusters by iteratively minimizing the sum of squared distances between data points and their assigned cluster's centroid. In other words, k-means seeks to minimize the within-cluster sum of squares (WCSS) by adjusting the position of centroids until the WCSS cannot be reduced any further. After, we plot the results to investigate the groupings. Before we perform cluster analysis, we must find the optimal number of clusters.

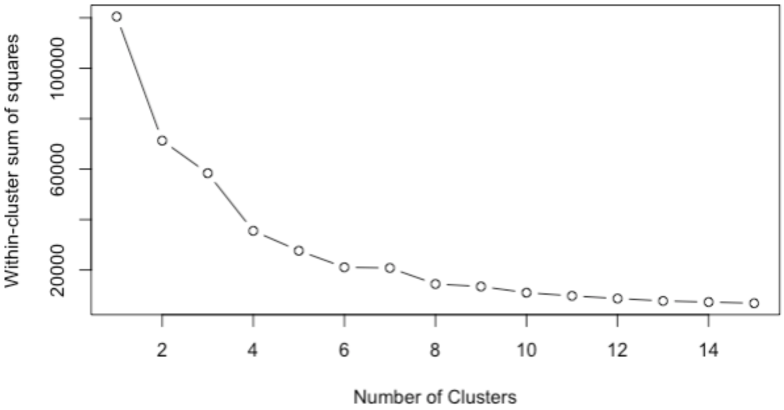


Figure 4: The relationship between within-cluster sum of squares the number of clusters

To find this optimum, we use the within-cluster sum of squares as our criterion. We then produce a plot of the relationship between the within-cluster sum of squares and the number of clusters shown by *Figure 2*. The optimal number of clusters is at the point where the value of within-cluster sum of squares levels off. In our case, the optimal number of clusters is between 7-11-. Another type of cluster analysis we can employ is hierarchical clustering. Hierarchical clustering is a bottom-up approach that builds a tree-like hierarchy of clusters. In hierarchical clustering, each data point is initially assigned to its own cluster, and then these clusters are iteratively merged into larger clusters based on the similarity between them. We use the same method, relationship between within-cluster sum of squares and number of clusters, to find the optimal number of clusters. After we plot these hierarchical clusterings on a cluster dendrogram.

Results:

From the PCA, we retained 5 components which maintained at least 60% of the original data's variability. Refer to *Figure 2* and *Figure 3* for more information. From our within-cluster sum of squares cluster number criterion, we first conducted K-means cluster analysis with K = 7,9,11. *Figure 5* displays results for each K:

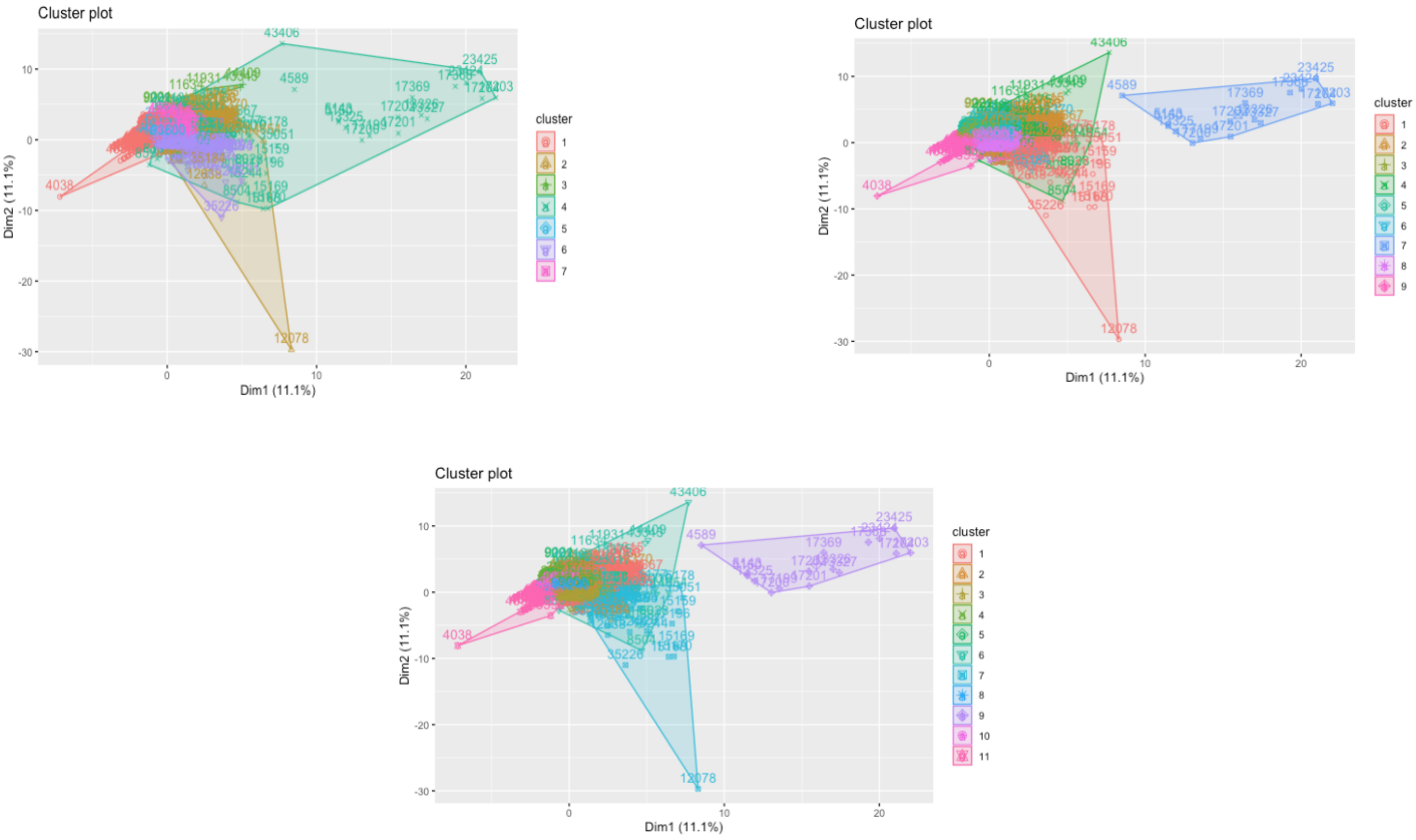


Figure 5: K means clustering plot with K = 7 (Top Left) K = 9 (Top Right) K = 11 (Bottom Middle)

To begin, although we were able to create different sets of clusters and plot them, the results are inconclusive as the plots are not interpretable. From *Figure 5,* we can see that with K-means for K = 7, 9, 11, all three cluster plots have massive congestion, and it is difficult to evaluate any clear cluster separation. For K = 7, all the clusters are connected with clusters 1, 2, and 4 stretched by outliers. For K = 9, cluster 7 is fully separated from the others. For K = 11, we see an almost identical plot, but are unable to draw any sound conclusions. Next we conducted hierarchical clustering with the same number of clusters based on the previous within-cluster sum of squares criterion. *Figure 6* displays the results:

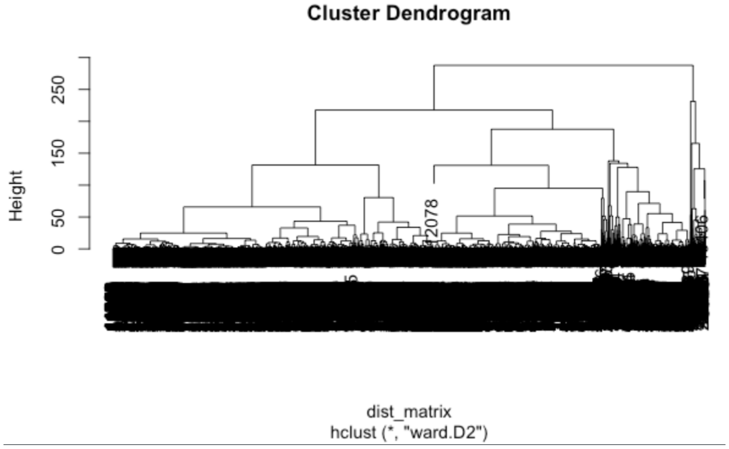


Figure 6: Cluster Dendrogram for Hierarchical Clustering with Ward's Linkage

Similar to K-means clustering, we are unable to interpret the results from hierarchical clustering. *Figure 6* displays the hierarchical clusters from the hierarchical clustering method. From height 0 to 50, it is difficult to understand each level.

Discussion:

The aims of this analysis were to employ clustering methods to be able to predict food groups based on nutrients. A massive issue with the analysis is that we were unable to gather any interpretable results for the cluster analyses. This failed attempt could be due to many things. The first is that we are working with a high dimensional data set, and we used PCA to reduce the dimensions. PCA is a great technique to use when conducting analysis on high dimensional data as one can reduce dimensions as covariates are highly correlated. This may be effective for many applications, but in our case from the PCA's retained for clustering, we were unable to identify certain cluster groups. Another issue with this analysis is that PCA requires scaling and centering numerical data only. An important feature in this data set was the pre categorization of food groups. Since this variable had to be dropped in order to reduce dimensionality, we lost vital information to be able to classify food groups after clustering. Some limitations of the research are due to high dimensionality. These include the "curse of dimensionality", computational complexity, and overfitting. The "curse of dimensionality" refers to how as the number of variables increases, the volume of the data space increases exponentially. This curse makes it difficult to find meaningful patterns in that data. Also as the number of variables increases, the computational complexity increases. This can lead to methods taking longer and becoming more difficult to compute. The last limitation of this research is overfitting. Since we have a lot of attributes, the model becomes very complex and ends up fitting the noise. This can be seen in *Figure 6*. The cluster dendrogram becomes very complicated and impossible to interpret for height 0-50.

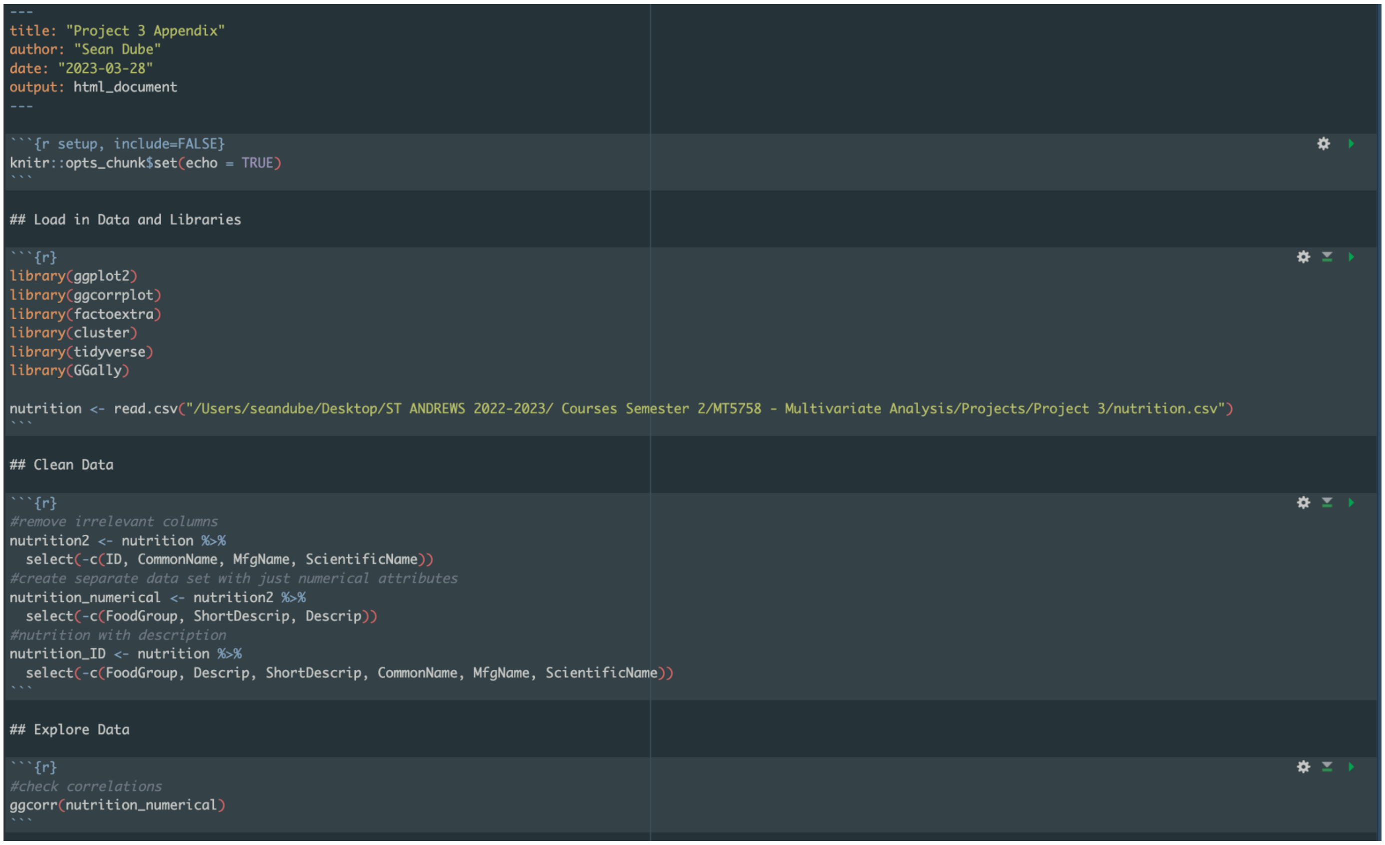
Conclusion:

The original aims of this research were to look into food groups and try to determine if there are certain food groups associated with a specific diet. Upon applying PCA and cluster analysis, we were not able to draw any conclusions about this research question. *Figure 4* shows we are able to generate clusters, but most clusters are difficult to visualize and we are unable to identify what the observations are associated with each cluster. The cluster plots do show that there are definite groupings of similar data which means there are certain foods that can be categorized together based on nutrients. This analysis can serve as an initial investigative look into whether clustering can be used to predict food groups. Moving forward, this analysis could benefit from using a dataset with fewer variables to avoid issues like computational complexity and the “curse of dimensionality”.

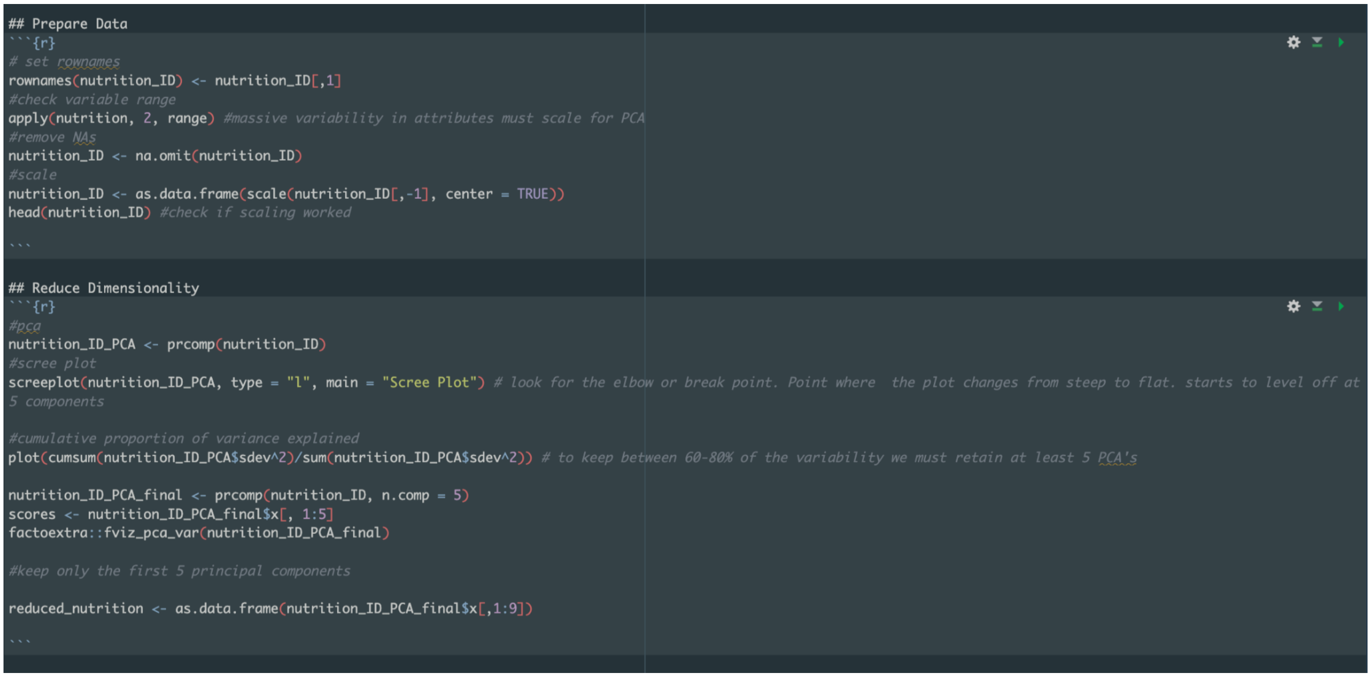
Bibliography

1. Kelly, C. (2016) USDA National Nutrient DB - dataset by Craig Kelly. Available at https://data.world/craigkelly/usda-national-nutrient-db (Accessed: March 1, 2023);

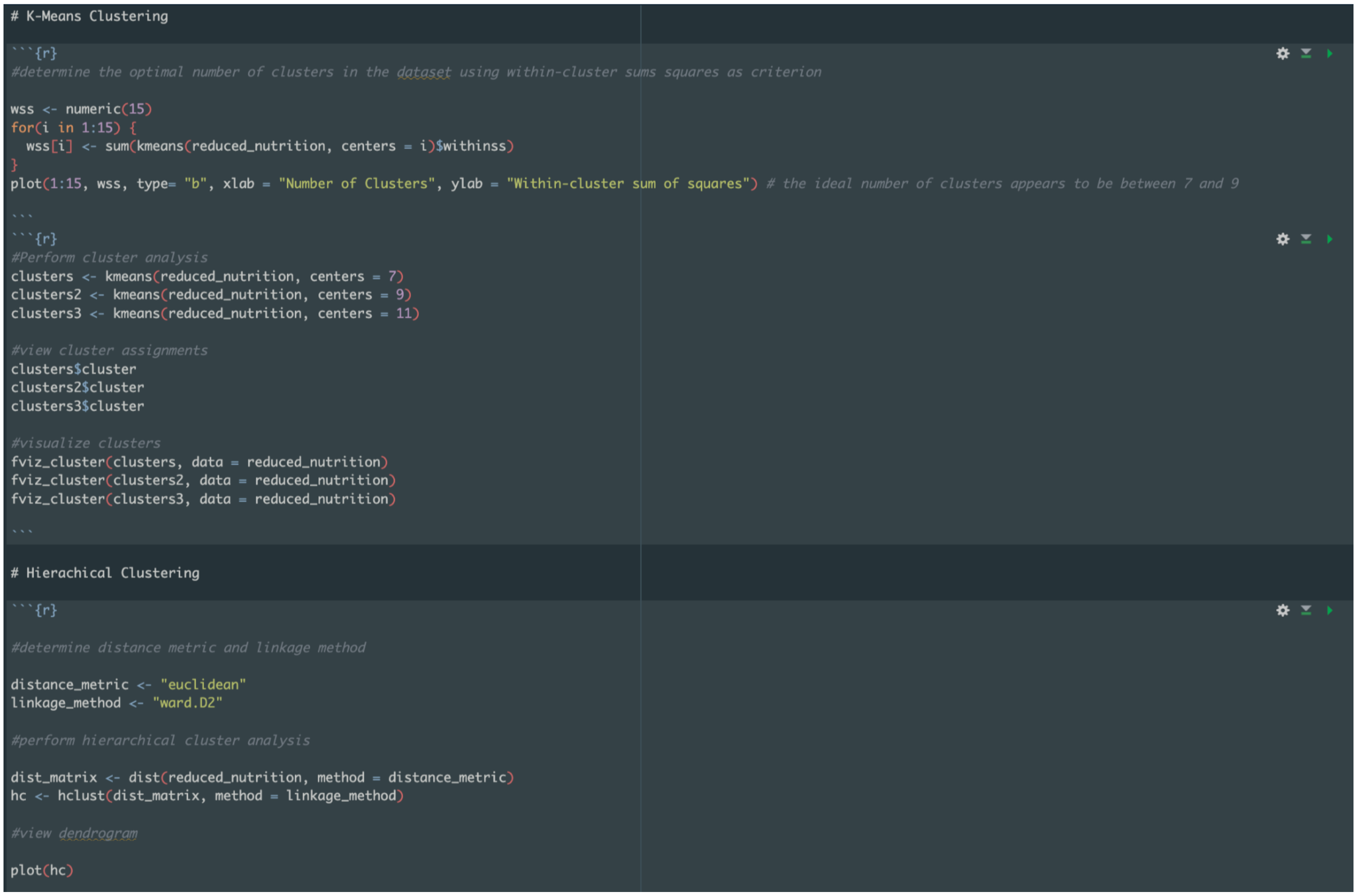
Appendix



~~~
---
title: "Project 3 Appendix"
author: "Sean Dube"
date: "2023-03-28"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

## Load in Data and Libraries

```{r}
library(ggplot2)
library(ggcorrplot)
library(factoextra)
library(cluster)
library(tidyverse)
library(GGally)

nutrition <- read.csv("/Users/seandube/Desktop/ST ANDREWS 2022-2023/ Courses Semester 2/MT5758 - Multivariate Analysis/Projects/Project 3/nutrition.csv")
```

## Clean Data

```{r}
#remove irrelevant columns
nutrition2 <- nutrition %>%
  select(-c(ID, CommonName, MfgName, ScientificName))
#create separate data set with just numerical attributes
nutrition_numerical <- nutrition2 %>%
  select(-c(FoodGroup, ShortDescrip, Descrip))
#nutrition with description
nutrition_ID <- nutrition %>%
  select(-c(FoodGroup, Descrip, ShortDescrip, CommonName, MfgName, ScientificName))
```

## Explore Data

```{r}
#check correlations
ggcorr(nutrition_numerical)
```
~~~



~~~
## Prepare Data
```{r}
# set rownames
rownames(nutrition_ID) <- nutrition_ID[,1]
#check variable range
apply(nutrition, 2, range) #massive variability in attributes must scale for PCA
#remove NAs
nutrition_ID <- na.omit(nutrition_ID)
#scale
nutrition_ID <- as.data.frame(scale(nutrition_ID[,-1], center = TRUE))
head(nutrition_ID) #check if scaling worked

~~~

Reduce Dimensionality
```{r}
#pca
nutrition_ID_PCA <- prcomp(nutrition_ID)
#scree plot
screeplot(nutrition_ID_PCA, type = "l", main = "Scree Plot") # look for the elbow or break point. Point where  the plot changes from steep to flat. starts to level off at
5 components

#cumulative proportion of variance explained
plot(cumsum(nutrition_ID_PCA$sdev^2)/sum(nutrition_ID_PCA$sdev^2)) # to keep between 60-80% of the variability we must retain at least 5 PCA's

nutrition_ID_PCA_final <- prcomp(nutrition_ID, n.comp = 5)
scores <- nutrition_ID_PCA_final$x[, 1:5]
factoextra::fviz_pca_var(nutrition_ID_PCA_final)

#keep only the first 5 principal components

reduced_nutrition <- as.data.frame(nutrition_ID_PCA_final$x[,1:9])

```
```



```
# K-Means Clustering

```{r}
#determine the optimal number of clusters in the dataset using within-cluster sums squares as criterion

wss <- numeric(15)
for(i in 1:15) {
  wss[i] <- sum(kmeans(reduced_nutrition, centers = i)$withinss)
}
plot(1:15, wss, type= "b", xlab = "Number of Clusters", ylab = "Within-cluster sum of squares") # the ideal number of clusters appears to be between 7 and 9

```
```{r}
#Perform cluster analysis
clusters <- kmeans(reduced_nutrition, centers = 7)
clusters2 <- kmeans(reduced_nutrition, centers = 9)
clusters3 <- kmeans(reduced_nutrition, centers = 11)

#view cluster assignments
clusters$cluster
clusters2$cluster
clusters3$cluster

#visualize clusters
fviz_cluster(clusters, data = reduced_nutrition)
fviz_cluster(clusters2, data = reduced_nutrition)
fviz_cluster(clusters3, data = reduced_nutrition)

```

Hierachical Clustering

```{r}

#determine distance metric and linkage method

distance_metric <- "euclidean"
linkage_method <- "ward.D2"

#perform hierarchical cluster analysis

dist_matrix <- dist(reduced_nutrition, method = distance_metric)
hc <- hclust(dist_matrix, method = linkage_method)

#view dendrogram

plot(hc)
```