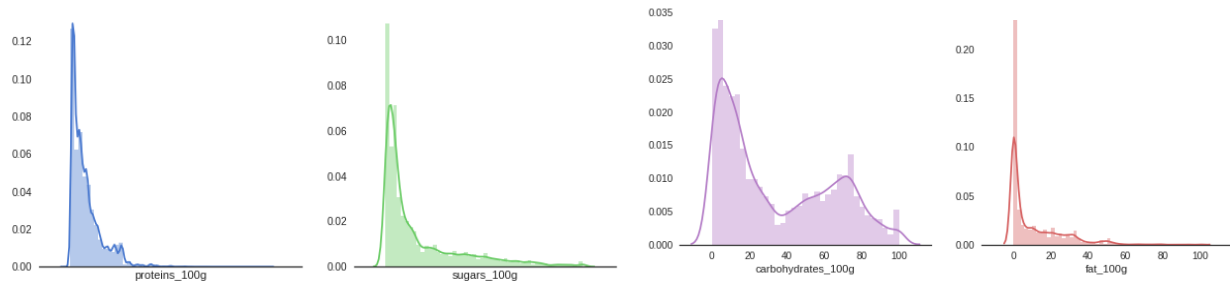


## Food Clustering

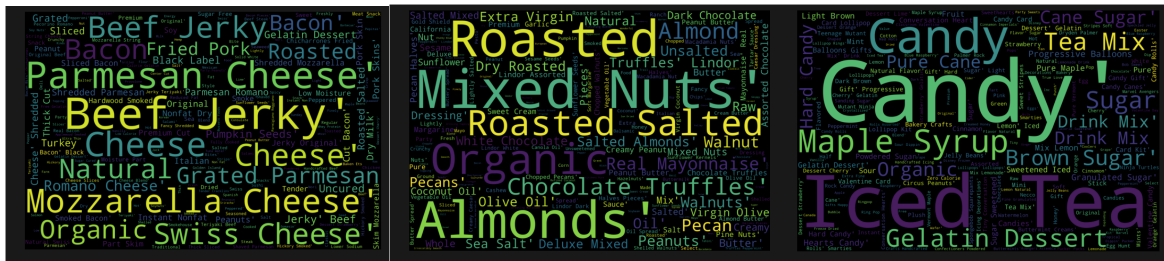
Food and nutrition are key to have good health. These two are really important for people to maintain a healthy diet especially for who has dietary restrictions. At this time, food clustering is a useful way to help people to choose food that they actually need or not. For instance, diabetic patients can utilize clustering result to avoid foods with high sugar food. And patients with simple fatty liver can use this result to choose foods with low fat and low sugar. In this report, I will introduce the method of food clustering.

Ingesting the dataset is the first step of clustering. I used food features dataset as the data frame to do the following exploration. There are about 40k rows in this data frame.

The second step is utilizing data visualization to come up with more insights.

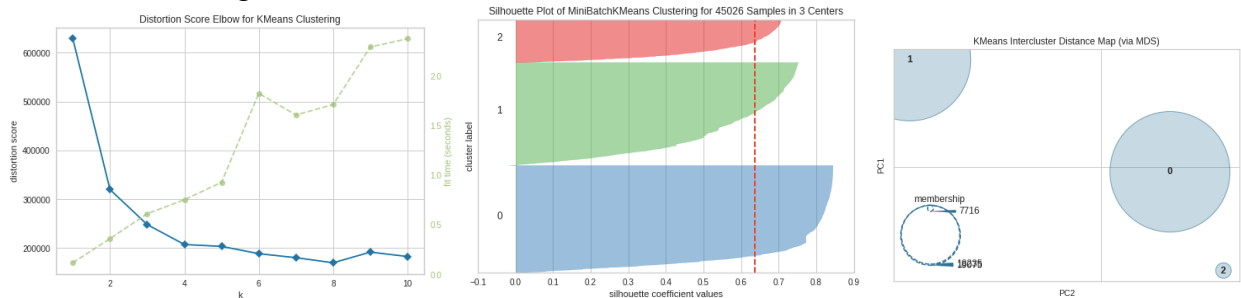


(These four histograms show out the generate distributions based on energy type: proteins, sugars, carbohydrates, and fat.)

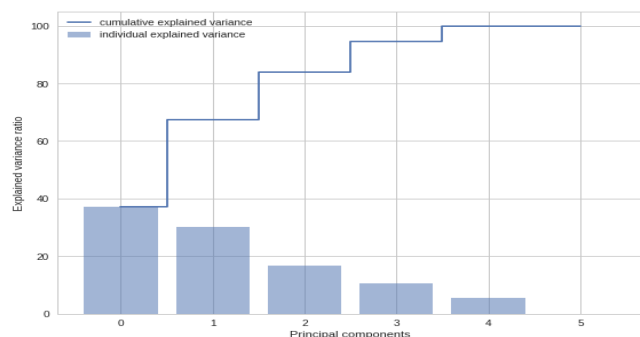


(These are word clouds about each feature. From left to right: High protein foods, High Fat foods, High Sugar foods. According to these word clouds, we can figure out that the Beef Jerky and cheese are the popular food with high protein, the Nuts are the popular food with high fat, and the Candy is the popular food with high sugar.)

Thirdly, building up the clustering model and define the food new clusters. The following are several cluster diagnostics:

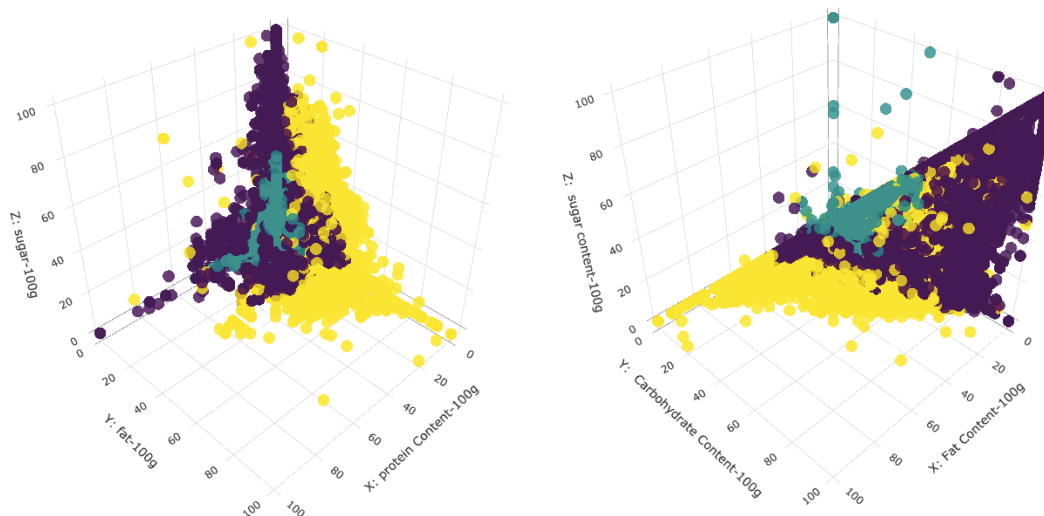


According to these charts, we can conclude that 3 is the appropriate number of clusters in our food features dataset. Hence, we utilized this size for the following clustering. In order to reduce the dimensions, I used the PCA to complete it.



Based on the left chart, the cumulative\_explained\_variance line shows out that when the PCA components increased into 4, we can get larger than 90% accuracy. Hence, I set the component as 4 for the following clustering exploration.

Finally, after did the PCA, I used the 3D scatter plots to show out the clustering result of food features.



(Left: clustering result of sugar-fat-protein food. Right: clustering result of sugar-carb-fat food)

In conclusion, in the sugar-fat-protein food clustering 3D-scatter plot: green spots are foods with low-sugar/low-fat/low-protein, purple spots are foods with high-sugar/low-fat/high-protein, and yellow spots are foods with high-sugar/high-fat/mid-protein. And in the sugar-carb-fat food clustering 3D-scatter plot: green spots indicate low-fat/mid-sugar/mid-carb, yellow spots indicate low-sugar, and yellow spots are foods with low-fat.

My Colab Link:

[https://colab.research.google.com/drive/1oU\\_IsiSbYZU7EwkfOFSWw1AHr7ep37Dk](https://colab.research.google.com/drive/1oU_IsiSbYZU7EwkfOFSWw1AHr7ep37Dk)