Homework 6: Clustering on World Food Facts Dataset

https://colab.research.google.com/drive/1mCRON_gORrJqH4oEYzFv2DtQ0QfVK4uB

Overview

Utilizing the dataset from Kaggle, which gives the content of common nutrients(fat, carbohydrates, sugars, proteins, salt and energy) of each food product, this project establishes an unsupervised model to fit the products into different clusters. In order to better visualize the result of clustering, PCA is also performed to reduce the dimensionality of the data.

Technique

Step 1: EDA

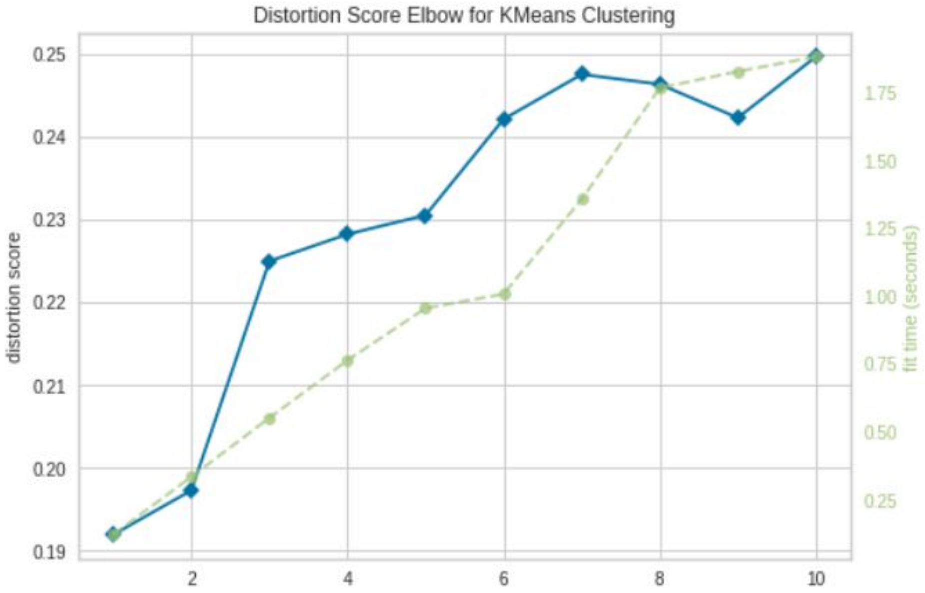
In this step, the goal is to better understand the relationship between variables and the features of each variable. Histograms are used to describe the distribution of fat, carbohydrates, sugars and proteins; word clouds are used to show the high-frequency food which is high of those four nutrients correspondingly.

Step 2: Scale the Data

In the original dataset, the units of different features are all calculated in gram per 100 gram. However, the magnitude and range can be very different. For example, the content of salt per 100 gram can be much less than that of carbohydrates. Therefore, I scale the data so that every feature has the same range.

Step 3: Clustering

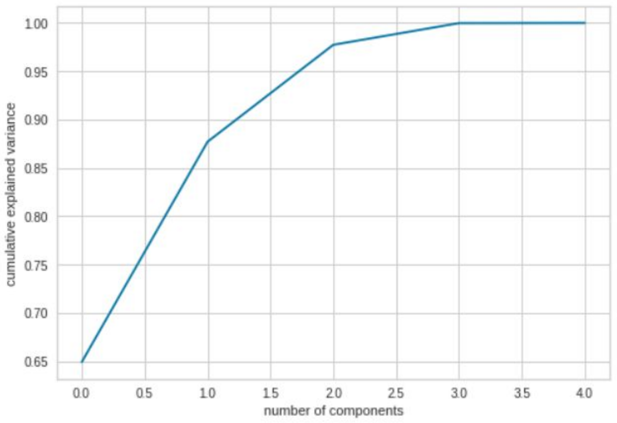
Based on K-Means algorithm, the graph below shows the best number of clusters is two, which is reasonable in both distortion score and fit time. Then I put all of the six features into a K-means model and get **two clusters**.



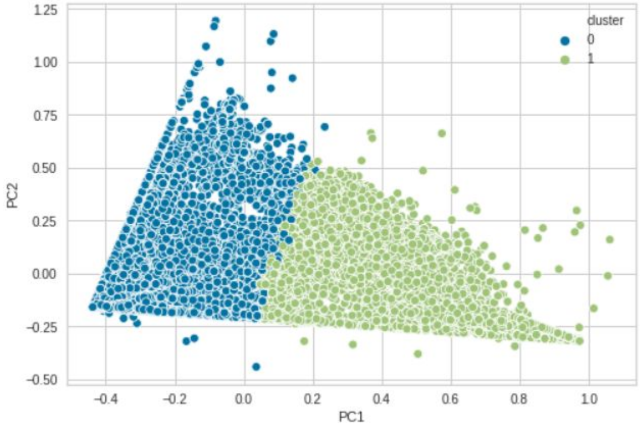
Step 4: PCA

Since there are six features in total, making it difficult to visualize the result within three dimensions, PCA is conducted to reduce the dimension. First I plot the cumulative explained variance to decide how many components to keep. And we can see from the below graph 98% variance can be retained if we keep **two components**. Through the coefficients got from the PCA process, I get the following information:

- PC1 represents food with high carbohydrates, sugars and energy
- PC2 represents food with high fat and energy



Using the first two components, I visualize the two clusters in a 2D graph below. We can see that cluster 1 is relatively high in PC1 and cluster 0 is relatively high in PC2.



Conclusion

Based on the information above, at least we can draw a conclusion that cluster 1 is “high-fat food”. But for cluster 0, there are not enough attributes to name it. Therefore, further analysis needs to be conducted on cluster 0.