URL to Colab:

https://colab.research.google.com/drive/1J7ZzmavgRJDobPwG-EP2Y_XPP_dJMZWR

Project Explanation:

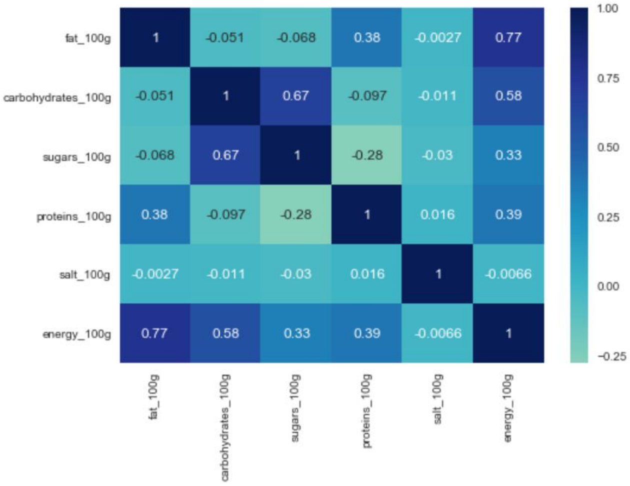
The project is to modify the current World Food Fact notebook, identify the correct attributes to cluster on, conduct Principal Components Analysis to identify the most important features and improve the food clusters using unsupervised machine learning models.

Ingestion and EDA

In our dataset, each food product has six features: fat, carbohydrates, sugar, protein, salt and energy. We first dropped the outliers in case they would affect the model building process. Then as we all know that energy is calculated by fat, carbohydrates and proteins, and sugar is considered as one type of carbohydrates, we computed and visualized the correlations between each features of food.

As we can see from the following correlation heatmap, the correlation between sugar and carbohydrate is 0.67, and the correlations between energy and fat, carbohydrate and protein are 0.77, 0.58 and 0.39 respectively. This indicates that before model building, it would be better to conduct Principal Component Analysis to remove the correlations between features.

However, before we started Principal Component Analysis, features standardization is needed to identify the components that maximize the variance. So, we removed the mean and scaled the features to unit variance.



Principal Components Analysis

Four components are needed for model building, and these four components will explain 94.53% of data variance. After composing the four components by using the elements of eigenvectors as the weights of each feature, we decided that the four principal components include comfort level, keto level, salty level, and energy level.

Data Modeling – Gaussian Mixture Model

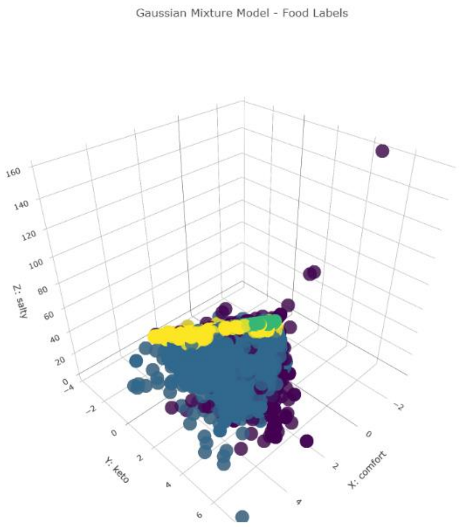
Since we cannot simply separate food into two groups – good or bad food, we decided to use Gaussian Mixture Model to cluster food and calculate their certainties to be in each cluster. By comparing the AIC of Gaussian Mixture Models with different components, we determine to cluster the food in our dataset into four clusters to achieve the smallest AIC.

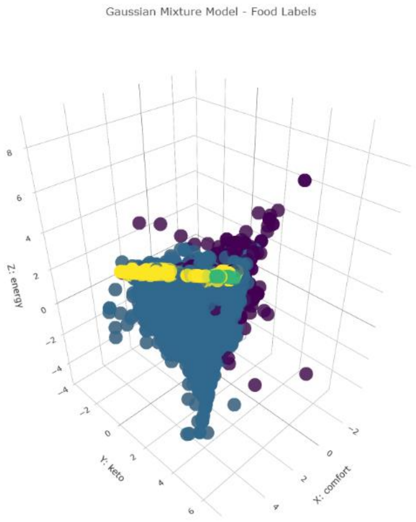
Recommendations to Management:

After we conducted the Gaussian Mixture Modeling, the following four food clusters and the means of their principal nutrition components are listed below:

Food Cluster	Comfort Level (Mean)	Keto Level (Mean)	Salty Level (Mean)	Energy Level (Mean)
Sweet treats and other high carb food	-0.40	-1.07	-0.06	0.08
Sauce and seasoning	-0.03	0.49	1.29	0.81
Fruit and fruit related product	-1.82	-0.26	-0.12	-0.33
Keto friendly food (Relatively healthy food)	0.65	0.42	-0.02	0.01

We also plotted the four food clusters based on the four food feature dimensions:





Project Conclusion:

To sum up our project, we have to first understand that there are no “good” or “bad” food. Gaussian Mixture Model is to build cluster food into four categories, which are sweet treats, sauce and seasoning, fruit and fruit related products, and keto friendly food. Before building the Gaussian Mixture Model, it is important to use Principal Component Analysis to remove the correlations among the features of food.