TUNKU ABDUL RAHMAN UNIVERSITY OF MANAGEMENT AND TECHNOLOGY

FACULTY OF COMPUTING AND INFORMATION TECHNOLOGY

Nutrition Based Food Clustering

BMCS2114 MACHINE LEARNING

2023/2024

Student's name/ ID Number	:	Seow Yu Xuan/22WMR04098
Student's name/ ID Number	:	Tam Kok Yan/22WMR04101
Student's name/ ID Number	:	Tyuo Chen Moon/22WMR4106
Student's name/ ID Number	:	Wong Man Ee/22WMR04109
Programme	:	Bachelor of Computing Science in Data Science (RDS)
Tutorial Group	:	2
Tutor's name	:	Dr Lim Siew Mooi

Table of Contents

Abstract

In the pursuit of understanding dietary patterns and their impact on health, this study examines the realm of nutrition analysis with a focus on macronutrients, namely carbohydrates, proteins, and fats, concentrating on their complex interplay within the human diet. Leveraging a comprehensive dataset with varying nutritional profiles comprising 8790 rows and 51 features, encompassing various nutritional aspects, we use four different clustering algorithms: K-means clustering, hierarchical clustering, Gaussian Mixture Model, and Partitioning Around Medoids. Through thorough experimentation and analysis, we uncover distinct macronutrient patterns in the data, revealing light on underlying structures and relationships. Our comparative analysis of different clustering algorithms reveals their relative strengths and limits in capturing subtle dietary categories. This study not only increases understanding of macronutrient dynamics but also demonstrates the effectiveness of machine learning tools in interpreting complex nutritional data landscapes.

Keywords: nutrition, food clustering, silhouette score, k means, k-medoids

Introduction

Nutrition is a fundamental aspect of human health, and it plays a critical role in influencing overall well-being and susceptibility to various health conditions. The global prevalence of chronic diseases has been on the rise, and it is becoming increasingly evident that our diet plays a crucial role in both preventing and managing these conditions (Delvin, U.M. et. al., 2012) Malnutrition is a severe public health challenge that encompasses undernutrition, overweight, obesity, and noncommunicable diseases related to diet. Understanding the intricate relationship between nutrition and health is crucial to mitigating the risks associated with poor dietary choices and addressing a wide array of health conditions.

To gain a better understanding of this complex interplay, clustering analysis will be performed on a dataset containing different food product combinations and their respective descriptions of nutrition. The details of the dataset used are as follows:

Table 1: Description of Dataset

Item	**Description**
Scope/location	United States
Size of the dataset	2,014 KB/8790 rows
Size of the features	51

This analysis can aid in uncovering hidden patterns from the dataset and offer insights into the similarity of different food products based on their vitamin and mineral content and help us make informed decisions about our diet and make necessary changes to our food choices to improve our overall health and well-being.

Problem Statement

The issue of nutrition and health is increasingly critical worldwide, with a surge in chronic illnesses highlighting the significant role diet plays (Devlin et al., 2012). Many of these illnesses underscore the pressing need to change eating habits, regardless of whether they are caused by nutritional deficiencies or are made worse by bad dietary decisions. Malnutrition may take many different forms. These include undernutrition, which is typified by wasting or stunting, deficiencies in vital vitamins or minerals, and the widespread issues of overweight and obesity, which in turn lead to an increase in noncommunicable illnesses that are diet-related (*Nutrition,* n.d.). To tackle these intricate problems, we seek to implement clustering approaches in segmenting nutritional value to unveil hidden patterns from the nutrition dataset.

With that said, the **objectives** of performing cluster analysis on different food product combinations and the respective descriptions of their nutrition are as follows:

1. To perform clustering based on similar nutritional values
2. To analyse food clusters for balanced diet combinations
3. To uncover hidden nutritional information and underlying relationships
4. To assess clustering algorithm performance using specific metrics

Literature Review

K-Means Clustering

In an Irish context, a meal-based study suggested by O'Hara et al. (2022) utilised **K-means clustering.** Initially, the Irish food pyramid, consisting of six food groups, was used as the framework. Subsequently, a data-driven approach incorporating 12 nutrients validated by the Nutrient Rich Foods Index was adopted to redefine food groups, supplemented by additional groups to accommodate foods not covered by the pyramid. These **15 distinct food groups** later served as the fundamental direction to derive generic meals. This study found that classifying individuals based on generic meals resulted in similar classifications as classifying individuals based on their original meals in terms of meeting nutrient-based dietary guidelines.

This study (Atsa'am et al., 2021) diverges from the focus on the nutritional information of foods to examine cereal foods in West Africa. Employing K-means clustering analysis, the research identifies **six sub-groups** within the cereal category, wherein food items with similar nutritional profiles form clusters. The West Africa Food Composition Table (WAFCT) encompasses 472 food sources characterised by 28 nutrients, as outlined by the Food and Agriculture Organization of the United Nations (FAO). These food sources are further categorised into **13 groups**. The similarity is then evaluated using **Euclidean Distance**.

Hierarchical Clustering

In this study (Da Silva Torres et al., 2006), the approach is reversed, focusing on the composition of foods through hierarchical clustering. The analysis involved 53 food preparations sourced from four distinct restaurants. Distances between samples were computed using square **Euclidean Distances**. For instance, examining Restaurant 1, the initial division segregated foods into two groups, A and B, based on their calorie content. Subsequently, group A further divided into subgroup C, characterised by higher moisture content, and subgroup D, distinguished by lower caloric value. Conversely, group B comprised three subgroups, with "feijoada" notably standing out due to its high protein content. Groups E and F exhibited disparities in moisture content and calorific value. Foods from various dietary groups can therefore be combined to create a balanced diet.

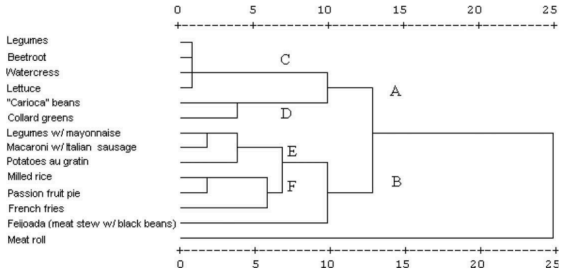


Figure 1: Clustering results from hierarchical clustering, depicted from:
Da Silva Torres, E. a. F., Garbelotti, M. L., & Neto, J. M. M. (2006)

Dalimunthe (2022) utilised **hierarchical clustering** to develop a food categorisation system focusing on nutritional attributes. They incorporated 10 features, encompassing calories, protein, fat, carbohydrates, calcium, phosphorus, iron, vitamin A, vitamin B, and vitamin C. The hierarchical approach has resulted in **six clusters**. Additionally, they employed the average linkage method alongside agglomerative clustering to identify similarities within the dataset effectively. The **Silhouette Index** was used as a reference to measure the best cluster, providing a quantitative measure of how well each data point fits within its cluster compared to other clusters. Given this, the team is implementing the Silhouette Index as a main reference to assess the quality of food clusters.

Gaussian Mixture Model (GMM)

Treitler et al. (2023) investigated the dietary patterns of adolescents in the United States by using citizen science projects to collect dietary data from high school students. The researchers performed a cluster analysis based on macro and micronutrients, revealing **9 food clusters** with distinct nutrient profiles. This research conducted two widely used methods - **K-means and GMM** to classify food items into clusters based on their similar nutrient profile. The research utilized the **Elbow Method** and the **Bayes Information Criterion** to estimate the number of clusters, which is consistent with the methodology used by the team working on this project. This paper highlights the potential of cluster analysis techniques in uncovering hidden patterns within dietary data and providing valuable insights into adolescent dietary habits.

Another research conducted by Balakrishna et al (2023) explores food clustering with **GMM** using the South African Food Composition Database (SAFCDB) focusing on nutrient content. The findings indicate that varying clustering criteria yield different numbers of clusters. For instance, when classifying food items by sodium content, **five classes** were identified, with class means ranging from 1.57 to 706.27 mg per 100 g. Conversely, **four classes** were distinguished based on the available carbohydrate content. In this study, the **Davies-Bouldin index** and **silhouette coefficient** were used as the performance metrics, providing insights into the clustering quality and the compactness of clusters.

Partitioning Around Medoids (PAM)

Following the initial application of K-means clustering to food groups, PAM was employed as a secondary algorithm (O'Hara et al., 2022). **PAM** was used to group individual meals, comprising various food groups, enabling clustering across both numerical and categorical variables. As PAM clustering necessitates a predetermined number of clusters, 24 different cluster numbers were applied to the data. The most frequently proposed cluster number among these indices was then chosen. The range of potential values considered for the **cluster number spanned from 4 to 15**.

On the other hand, Budiaji et al. (2021) investigated the Desirable Dietary Pattern (DDP) index, which is calculated based on an individual's total energy intake derived from protein, fat, and carbohydrate consumption. The dataset was gathered from responses to a questionnaire provided by 14 individuals. In contrast to the 2-dimensional plot utilized in this project, researchers employed a 3-dimensional plot to create **three to four groups**. Five different similarity distance techniques were applied to the **PAM algorithm and the Simple K-Medoids (SKM) algorithm**. Each cluster was interpreted based on the characteristics of its members. For instance, cluster 2 was identified as "carbohydrates" due to the highest frequency of members exhibiting this dietary trait.

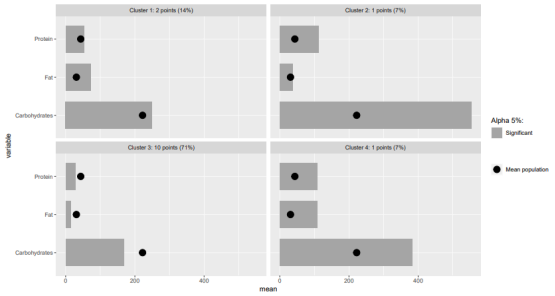


Figure 5. Barplot of the 4 clusters ($k = 4$).

Figure 2: Clustering results from PAM, depicted from:
Budiaji, W., Riyanto, R. A., & Suherna. (2021).

Research Methodology

Flowchart

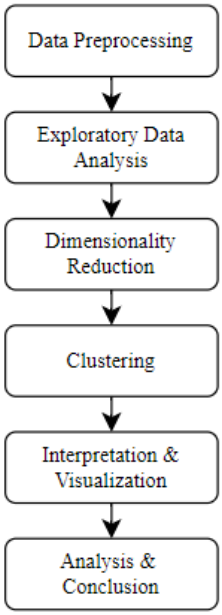


Figure 3: Flowchart of the project

Data Preprocessing

Before conducting data preprocessing, data acquisition is carried out to load the dataset in a variable 'NutritionList'. The information on each feature in the dataset is displayed for reference purposes.

Extract Relevant Data: Macronutrient

Since the dataset contains too many features, the analysis was narrowed down to focus solely on macronutrients to simplify the clustering process.

Data Cleaning

- Handling Missing Values
    - Filling missing values with specific value, 0
    - Filling missing values with mean of column
    - Filling missing values with median of column
    - Filling missing value with mode of column
    - Filling missing value with forward fill method
    - Filling missing value with backward fill method
    - Filling missing values with linear interpolation method
    - Remove the rows with missing values

- Handling Outliers
- Handling Contaminated Data
- Handling Inconsistent Data
- Handling Invalid Data
- Handling Duplicate Data
- Handling Data Type Issues
- Handling Structural Errors

MinMaxScaler

MinMaxScaler is a normalisation technique that converts the values to a range of 0 and 1. It is useful in preserving the shape of the original distribution using the following formula:

$$X_{\text{scaled}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

Where X = original feature value
X_{min} = minimum feature value
X_{max} = maximum feature value
X_{scaled} = scaled feature value

Figure 4: Mathematics formula involved in MinMax Scaler, depicted from [What is the MinMax Scaler? | Data Basecamp]

Exploratory Data Analysis

Method 1: Heatmap

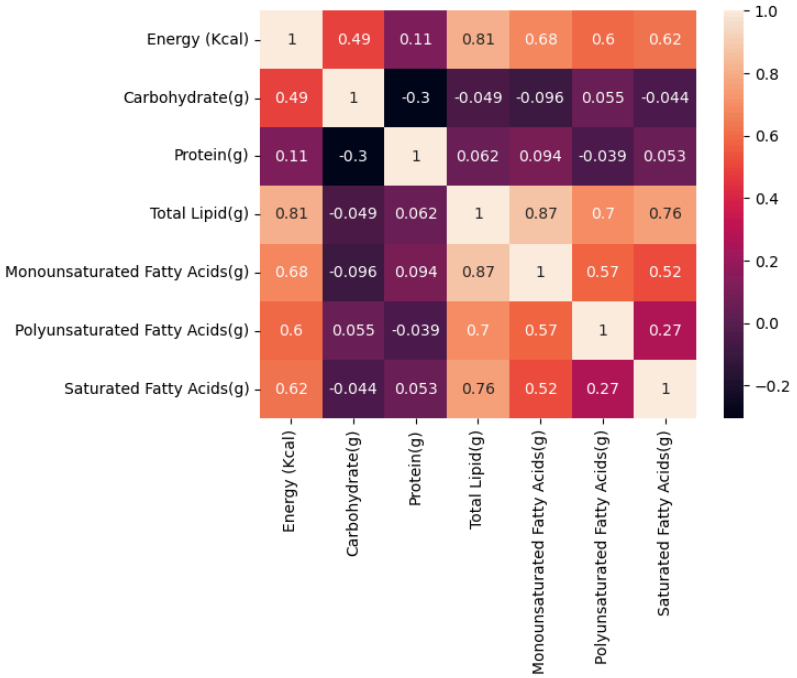


Figure 5: Correlation of the features using heatmap

Figure 5 shows findings of different correlations between different pairs of features in the macronutrients. These findings suggest that the pair of Energy and Total Lipid and the pair of Total Lipid and Monounsaturated Fatty Acid bring a high correlation (>0.8), which indicates that fats contain more energy per gram than carbohydrates.

Method 2: Violin Plot

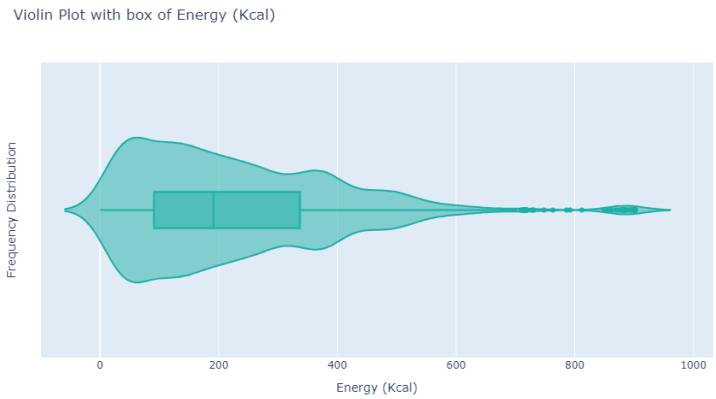


Figure 6: Violin plot to depict distributions of numeric data (Energy field)

Figure 6 shows one of the findings on the distribution of data. The findings show that almost all features are positively skewed, which means the most extreme values are on the right side.

Method 3: Histogram

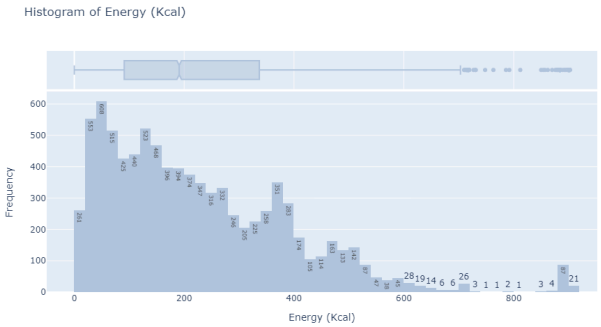


Figure 7: Histogram data distribution for Energy

These histograms offer a comprehensive insight into data distribution, effectively capturing outliers, data spread, and frequency distribution within each bin.

Method 4: Scatter Plot

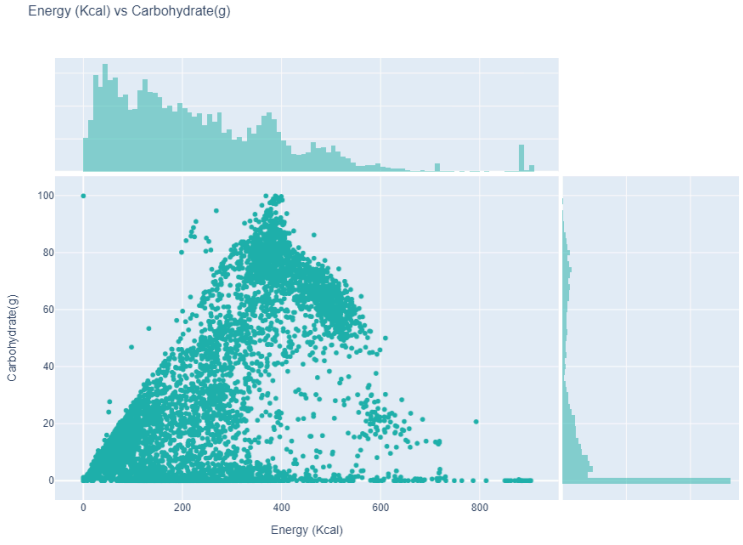


Figure 8: Scatter plot of correlations between Energy vs Carbohydrate

Through scatter plots, findings for each pair of features may bring out different meanings. For example, in Figure 8, findings show that the higher the energy, the higher the carbohydrate, but after energy reaches around 400 Kcal, the relationship from a positive correlation becomes a negative correlation.

Method 5: Text

Figure 9: Word Cloud insight from the Description field

From WordCloud, we may notice that there are a lot of descriptions having fat. Besides, a higher portion of food may still be in the raw state.

Analysing via Categories

Calories

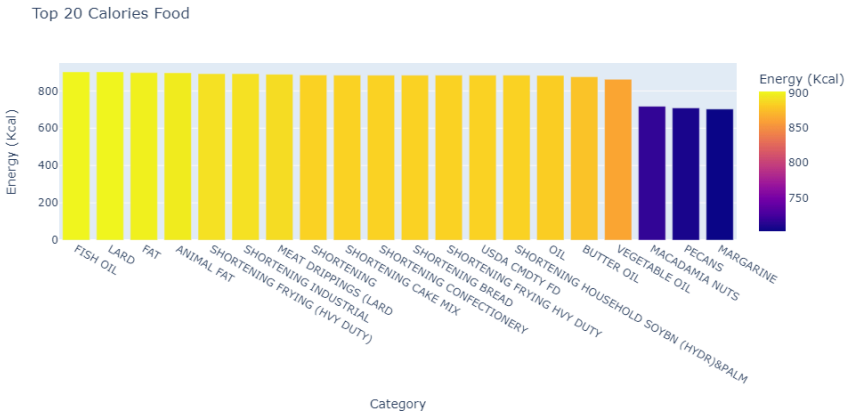


Figure 10: Calories Ranking Bar

The visualisations above indicate that fish oil and lard exhibit the highest calorie content, with 902K calories being recorded.

Carbohydrates

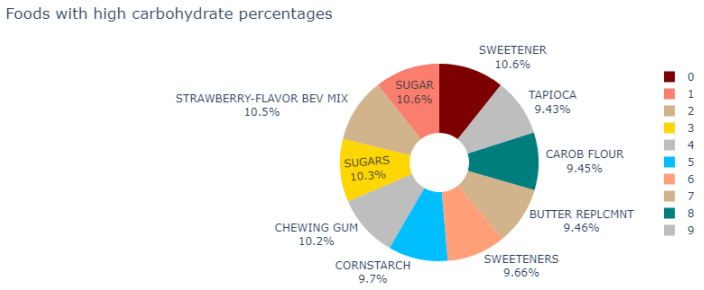


Figure 11: Pie Chart of Foods with High Carbohydrates Percentages

The pie chart tells us that sweeteners and sugar contributed to high carbohydrates.

Lipid

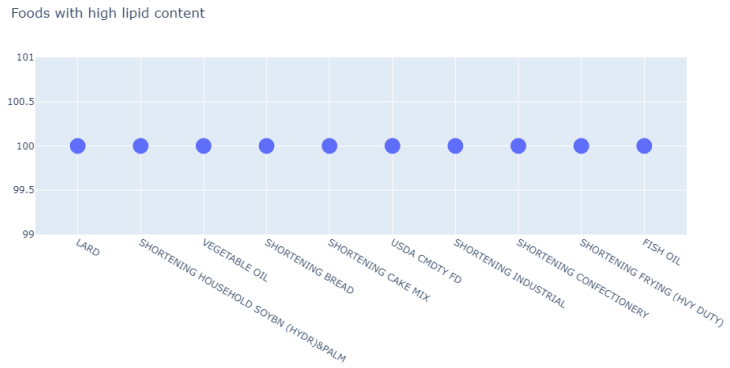


Figure 12: Foods with High Lipid Content

Findings suggest that it is advisable to limit the consumption of these foods due to their high lipid content (all with 100g).

Protein

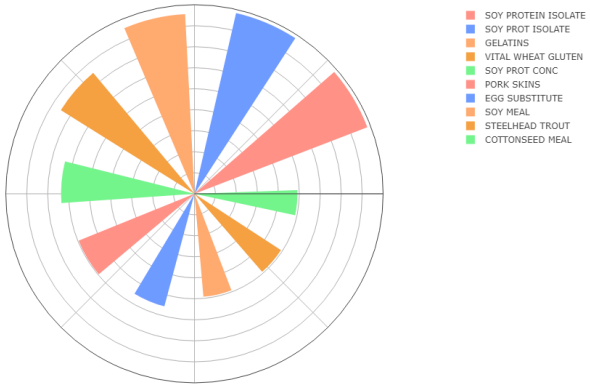


Figure 13: Circular Plot of Top 10 Foods Rich in Protein

The findings suggest that soy, gelatin, wheat, egg and pork skin are the highest protein foods, therefore those foods can be recommended to people who are in the gym and keeping fit.

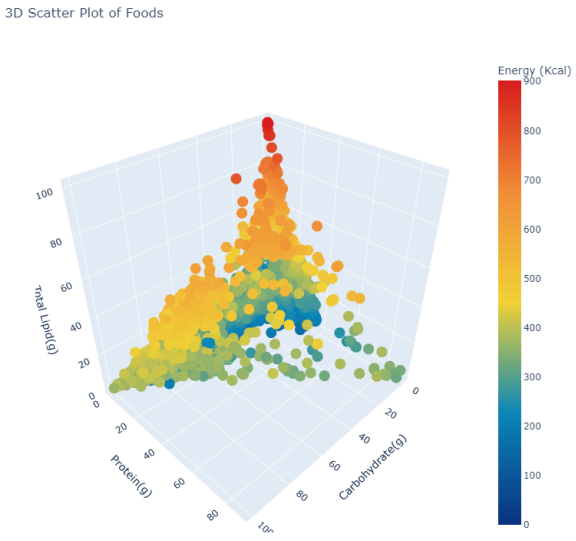


Figure 14: 3D Bubble plot of the relationship between food carbs, food protein and lipids

Findings show that each of the foods seems to have a balanced nutrition content in terms of weight where the total of 3 nutrients will not exceed 100g. There is no such thing as a food that has high carbs, lipids and protein at the same time. The bubble plot fully obeys the theory of energy equilibrium

Dimensionality Reduction

Data reduction techniques are invaluable when exploring high-dimensional datasets. With the current

dataset comprising 7 dimensions, the team aims to condense it into either 2 or 3 dimensions for visualization purposes. Among the plethora of dimensionality reduction techniques available, the team has narrowed their focus to two: Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE).

As discussed by Wakayama et al. (2023), t-SNE demonstrates robust performance when handling sufficiently large parameters. The article also highlights previous studies indicating that PCA may struggle to adequately separate certain clusters. This observation is reinforced by the 2D visualizations in the code, where PCA plots exhibit poor separation (as shown in Figure 15 and Figure 16). Furthermore, it is noted that t-SNE excels in capturing non-linear relationships that PCA might overlook. As a result, the team is determined to utilize t-SNE over PCA.

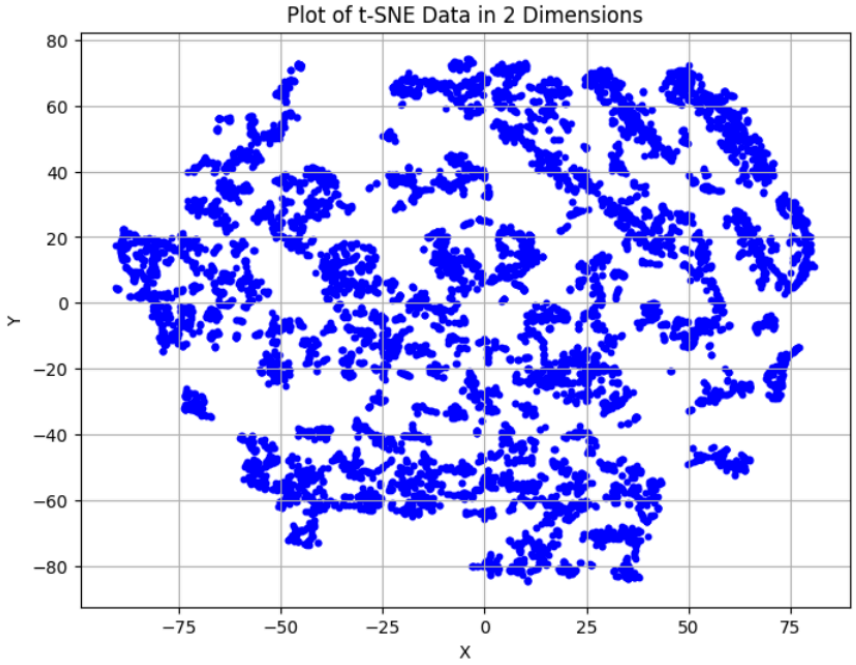


Figure 15: A 2 dimension t-SNE Plot captured from one of the datasets

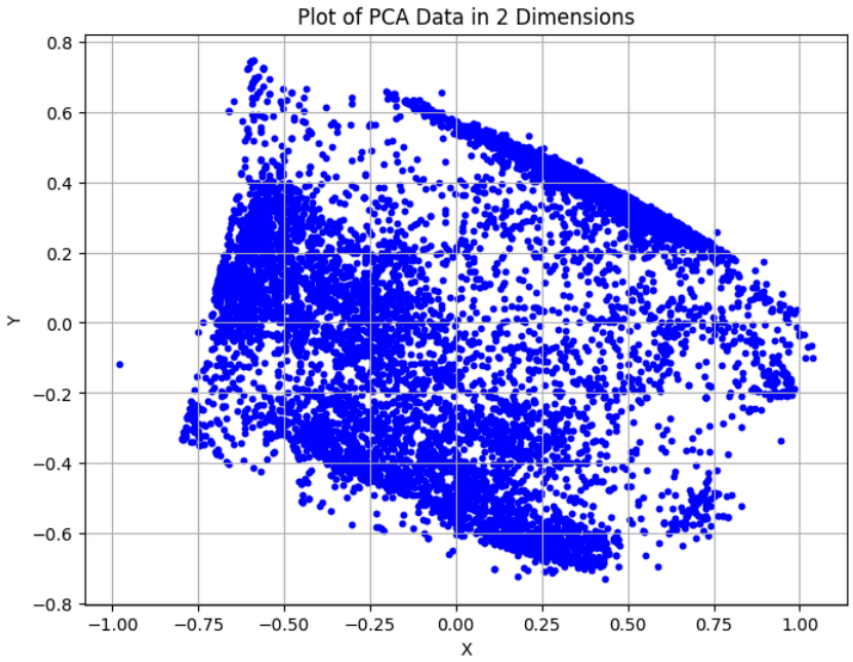


Figure 16: A 2 dimension PCA Plot captured from one of the datasets

Algorithm

K-Means Clustering

K-means stands out as a widely used clustering method that divides a dataset into K clusters. It aims to minimise the distances between data points within clusters while maximising the distances between clusters. However, K-means may encounter challenges when dealing with outlier-rich datasets (Botyarov & Miller, 2022).

The K-means algorithm operates as follows:

1. Initialize K with a random number, where K represents the desired number of clusters.
2. Assign data points to clusters whose centroids are closest.
3. Calculate new centroids for each cluster by computing the mean of the points assigned to that cluster. Then, update the assignment of data points.
4. Repeat steps 2 and 3 until convergence.
   ** Convergence criteria:*
   - *Centroids no longer exhibit significant changes.*
   - *Maximum number of iterations is reached.*
5. Obtain K clusters, each comprising a set of similar data points.

Agglomerative Clustering

Hierarchical clustering is a method aimed at hierarchically structuring data, reflecting relationships between data points. Through an iterative process, it calculates distances, typically using metrics like Euclidean Distance, to measure the similarity between objects. The result of hierarchical clustering is represented by a dendrogram, resembling a tree plot. There are two primary methodologies within hierarchical clustering: agglomerative and divisive. Agglomerative clustering progressively combines similar clusters, while divisive clustering divides data into increasingly smaller clusters (Botyarov & Miller, 2022).

In this project, the agglomerative approach is applied, with the steps stated as follows:

1. Consider each data point as a single-point cluster
2. Use the single linkage approach to combine the two closest distance clusters into a single cluster.
3. Continue from step 2 until there is just a single cluster.

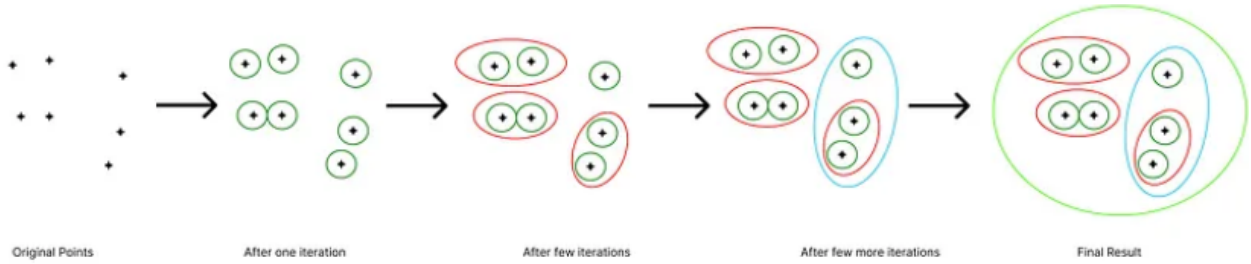


Figure 17: Concept of Agglomerative Clustering, depicted from Everything to know about Hierarchical Clustering; Agglomerative Clustering & Divisive Clustering. | by Chandra Prakash Bathula | Medium

Gaussian Mixture Model (GMM)

The GMM is a statistical technique that assumes the data points adhere to a Gaussian (normal) distribution. Similar to K-Means, GMM necessitates a predefined number of clusters, denoted as "n." GMM aims to optimise the model's fit to the data. The probability distribution function of a Gaussian Distribution with "d" features is expressed as follows:

$$N(\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

Where: μ = Mean
Σ = Covariance Matrix of the Gaussian
d = The number of features in our dataset
x = the number of data points

Figure 18: Mathematics formula involved in GMM, depicted from Treitler, J. T., Tekle, S., Ushe, J., Zanin, L., Capshaw, T. L., Tardieu, G., Libin, A., & Zeng, Q. (2023)

Partitioning Around Medoids (PAM)

The primary objective of the algorithm is to minimise the average dissimilarity between objects and their nearest neighbours. The PAM algorithm facilitates clustering concerning any designated distance matrix, thereby exhibiting reduced sensitivity to outliers. Compared to k-means, PAM is deemed more robust due to its utilisation of data points as medoids, avoiding random introduction (Botyarov & Miller, 2022).

1. Once the optimal number of clusters is identified, the first medoid is assigned as the data point that has the smallest distance to all other data points, making it the centre of the data set. The assignment data point is based on a chosen distance metric, typically Euclidean distance.
2. For each cluster, the algorithm tries to find a new medoid that minimises the total dissimilarity (or distance) between the data points in the cluster and the medoid. This is done by iteratively trying to swap each non-medoid point with the current medoid and select the one that minimises the total dissimilarity. If such a swap reduces the total dissimilarity, the medoid is updated.
3. Repeat this process until it converges. The convergence criteria are identical to the K-means clustering algorithm.

Evaluation Metrics

Silhouette Score

The Silhouette Score determines how well a data point fits into its cluster and how distinct it is from other clusters. A high silhouette score indicates good clustering as the data point is far from the closest neighboring cluster and near the average distance of its cluster. A low silhouette score indicates a poor grouping since the data point is near another cluster and distant from the average distance inside its cluster. The formula for the Silhouette Score is as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)}$$

Where a(i) = average distance of each data point to other data points within the same cluster
b(i) = average distance of each data point to all other clusters it doesn't belong to

Figure 19:Mathematics formula involved in Silhouette score, depicted from [The silhouette method - Training Systems Using Python Statistical Modeling [Book]]

Bayesian Information Criterion (BIC)

The BIC is a criterion for model selection among a finite set of models. It balances the goodness of fit of the model with the number of parameters in the model, penalising models that are more complex. The BIC is defined as:

$$\mathrm{BIC} = \ln(n)k - 2\ln(\hat{L}).$$

Bayesian Information Criterion formula

$\hat{L}$ is the maximized value of the likelihood function of the model
n is the number of data points
k is the number of free parameters to be estimated

Figure 20: Mathematics formula involved in BIC, depicted from [What is Bayesian Information Criterion (BIC)? | by Analyttica Datalab | Medium]

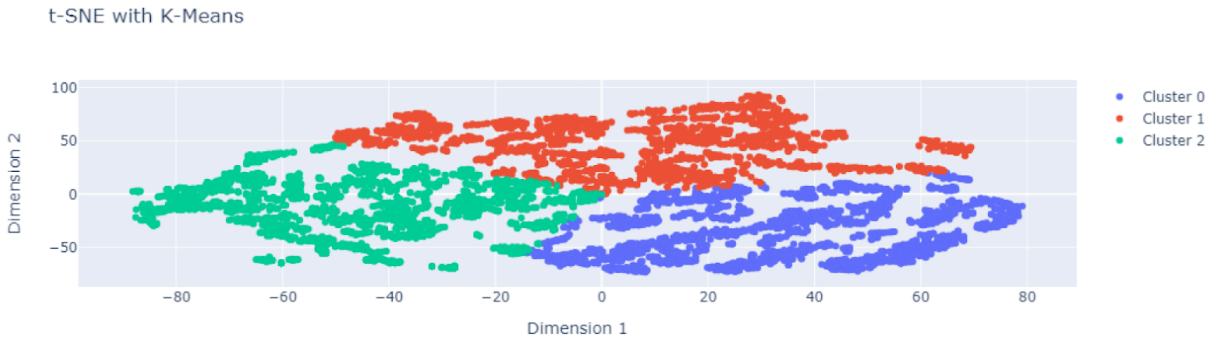
Table 2: Comparison of Silhouette Score of Clustering Techniques for Different Datasets

Dataset / Silhouette Score	K-Means	Agglomerative Hierarchical Clustering	GMM	PAM
Filling missing values with specific value, 0	0.43077356	0.4223424	0.4299104	0.43088004
Filling missing values with mean of column	0.41820893	0.37998518	0.37181818	0.41803378
Filling missing values with median of column	0.43591416	0.43591416	0.43544555	0.4359154
Filling missing value with mode of column	0.4307466	0.4223424	0.4299104	0.43088004
Filling missing value with forward fill method	0.4257467	0.40226936	0.42262316	0.42564258
Filling missing value with backward fill method	0.42386162	0.40360528	0.42308033	0.4204784
Filling missing values with linear interpolation method	0.42919	0.37387103	0.35802835	0.42898467
Remove the rows with missing values	0.4268393	0.39661264	0.42593384	0.42475685

Results for Different Dataset using same technique (K-Means)

Comparing clustering outcomes in a single dataset. Consider the following dataset of filling missing values with a specific value, 0 while implementing the K-Means technique.

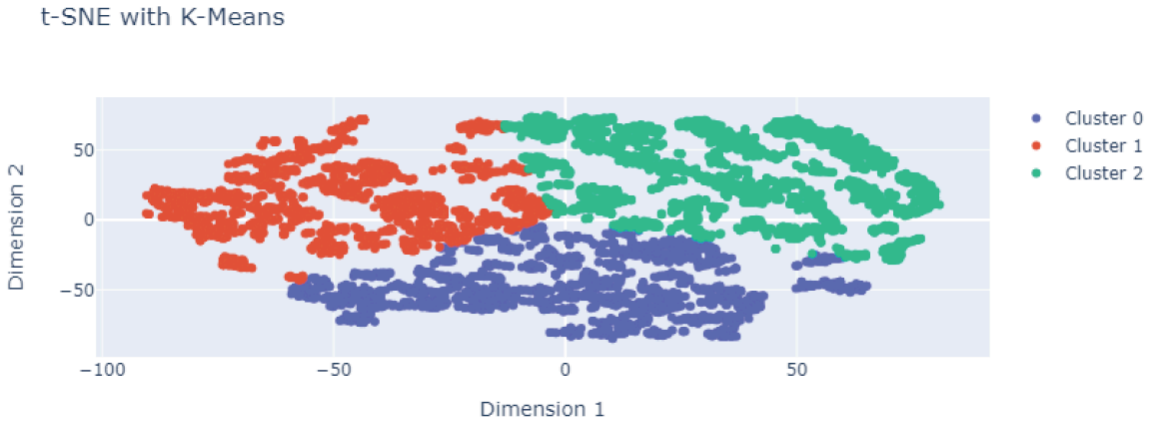
Dataset 1: Filling missing values with specific values, 0



Insights:

- In Cluster 0, on average is Low Carborhydrate but High Protein.
- In Cluster 1, on average is High in Energy and Carbohydrate, with a moderate level of protein and lipid.
- In Cluster 2, on average is Lowest Level of Lipid and Fatty Acids.

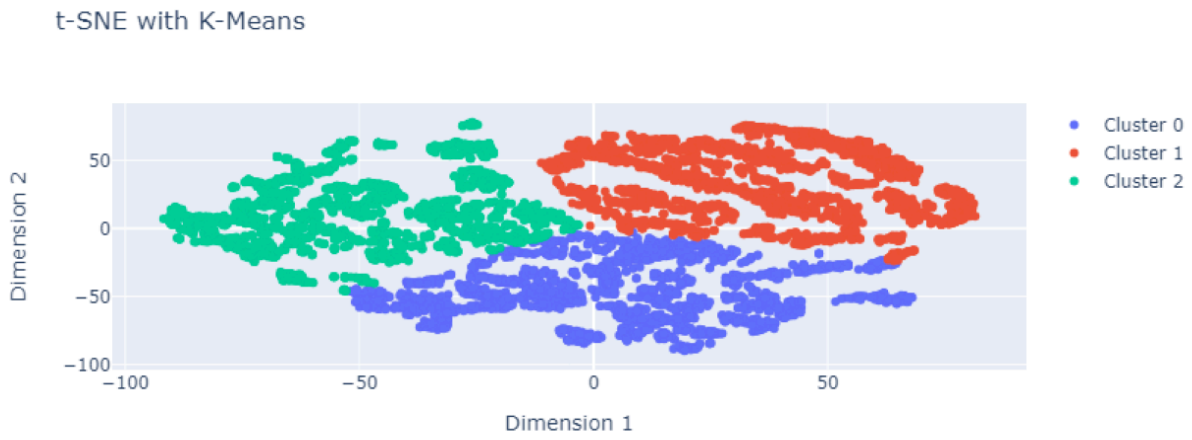
Dataset 2: Filling missing values with mean of column



Insights:

- Cluster 0: This cluster seems to segment foods that are with high lipid (fats).
- Cluster 1: The maximum values for the fatty acids in this cluster were the lowest among all the cluster. Hence can conclude that this cluster is food that low in fatty acids. This suggests that this cluster may consists of healthy food.
- Cluster 2: This cluster predominantly consists of low-carb foods, with approximately half of the data falling into this category. The statistical analysis reveals that this cluster exhibits the highest Q1 range for protein, reaching 0.63. Consequently, it can be inferred that Cluster 0 is primarily composed of protein-rich foods.

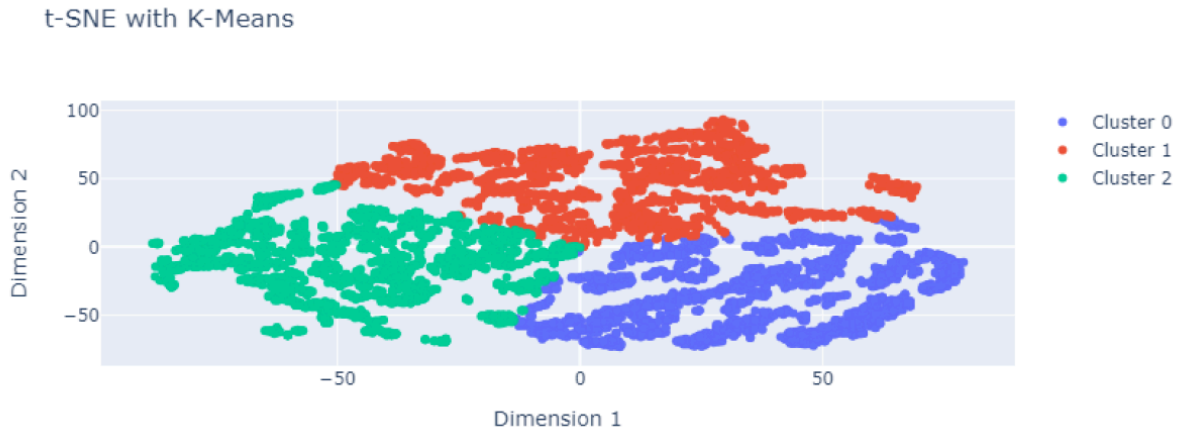
Dataset 3: Filling missing values with median of column



Insights:

- Cluster 0: It appears that this cluster contains foods heavy in lipids and energy. The energy value minimum of 0.656 is the greatest minimum value observed across all clusters.
- Cluster 1: This cluster predominantly consists of low-carb foods, with approximately half of the data falling into this category. The statistical analysis reveals that this cluster exhibits the highest Q1 range for protein, reaching 0.66. Consequently, it can be inferred that Cluster 1 is primarily composed of protein-rich foods.
- Cluster 2: The maximum values for the fatty acids in this cluster were the lowest among all the cluster. Hence can conclude that this cluster is food that low in fatty acids. This suggests that this cluster may consists of healthy food.

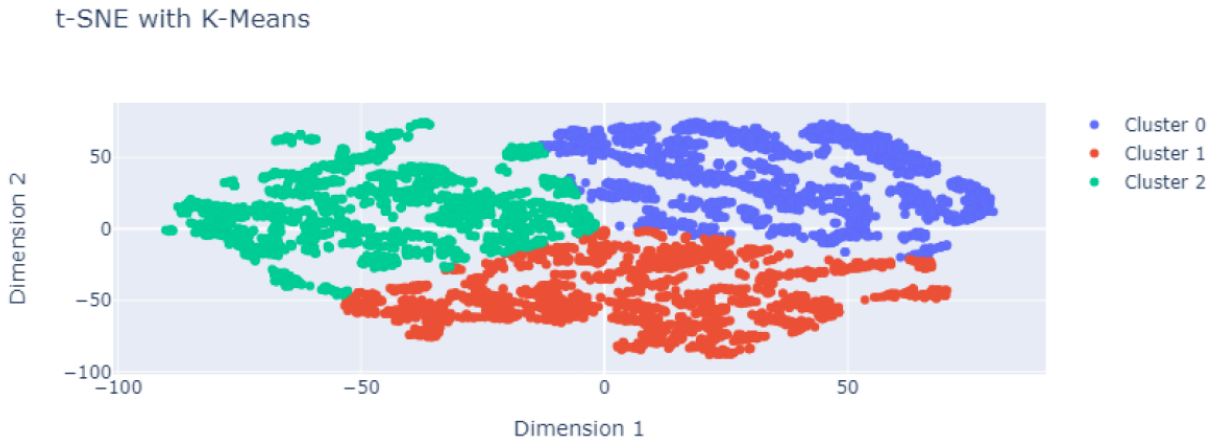
Dataset 4: Filling missing value with mode of column



Insights:

- In Cluster 0, on average is Slightly High in Energy and Protein, with a moderate level of Total Lipid.
- In Cluster 1, on average is High in Energy and Carbohydrate, with a moderate level of protein and lipid.
- In Cluster 2, on average is Slightly High in Energy with a moderate level of Carbohydrate.

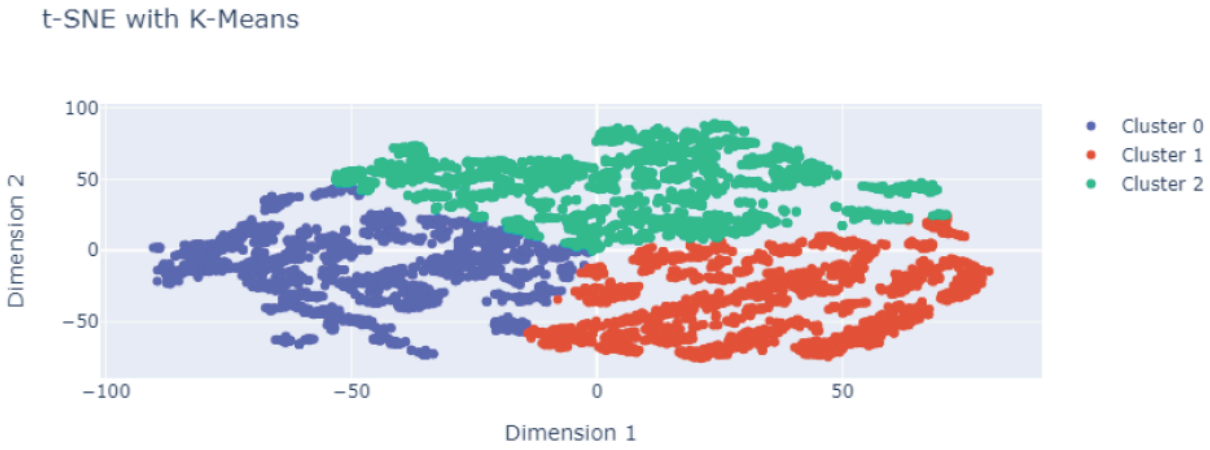
Dataset 5: Filling missing value with forward fill method



Insights:

- Cluster 0 indicates low value in carbohydrates and high value in protein.
- Cluster 1 shows very high values in energy, with moderate to high values in carbohydrates and total lipids on average.
- Cluster 2 exhibits moderate values in energy and carbohydrates, slightly low values in protein, and very low values in total lipids.

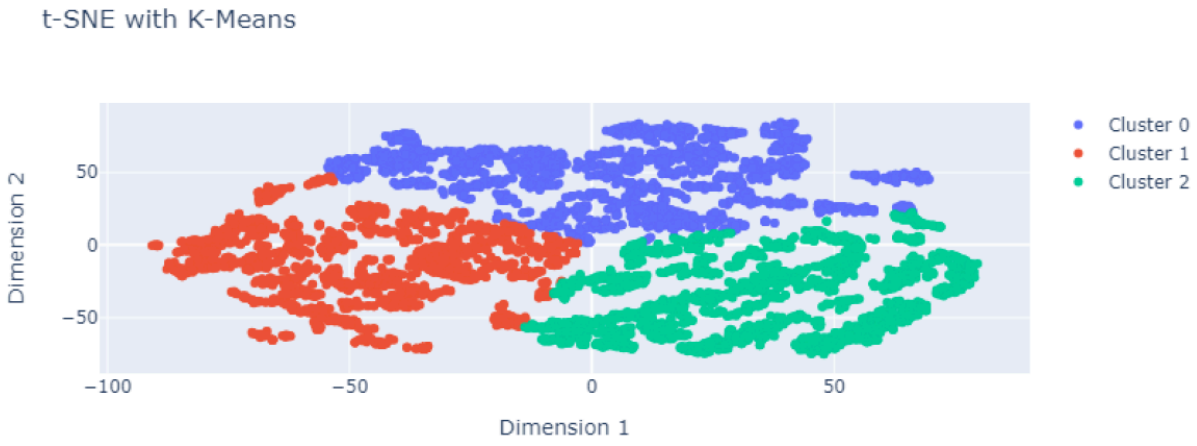
Dataset 6: Filling missing value with backward fill method



Insights:

- Cluster 0 shows very low values in total lipids among all clusters, with low values in protein. This cluster is probably related to foods that are generally healthier than those in the other two clusters.
- Cluster 1 indicates very low values for carbohydrates and moderate to high values in protein 19 and lipid, suggesting a high protein and low carbohydrate food group for this cluster.
- Cluster 2 exhibits the highest values in energy and carbohydrates and slightly moderate to low values in protein.

Dataset 7: Filling missing values with a linear interpolation method

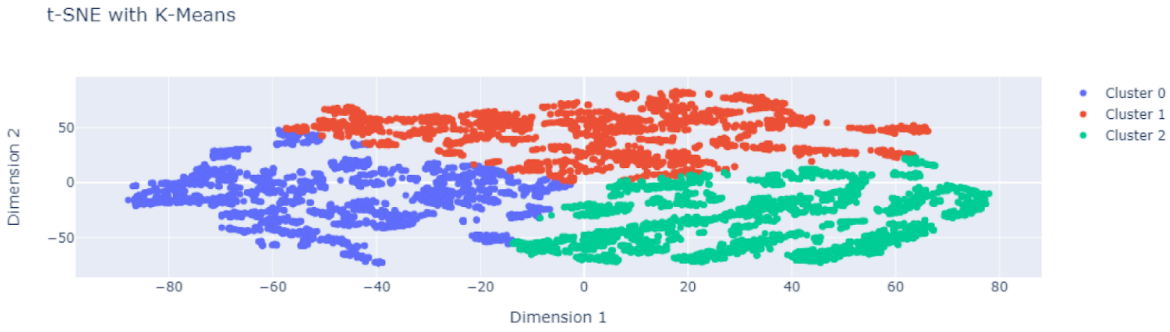


Insights:

- Cluster 0 contains the highest mean of energy values following with the highest values of carbohydrates.

- Cluster 1 shows the lowest mean of the fat as well as the fatty acid, but moderate values formother nutrition.
- Cluster 2 has the high value of energy values and highest values of the protein, show that it is a good suggested food group for fitness people.

Dataset 8: Remove the rows with missing values



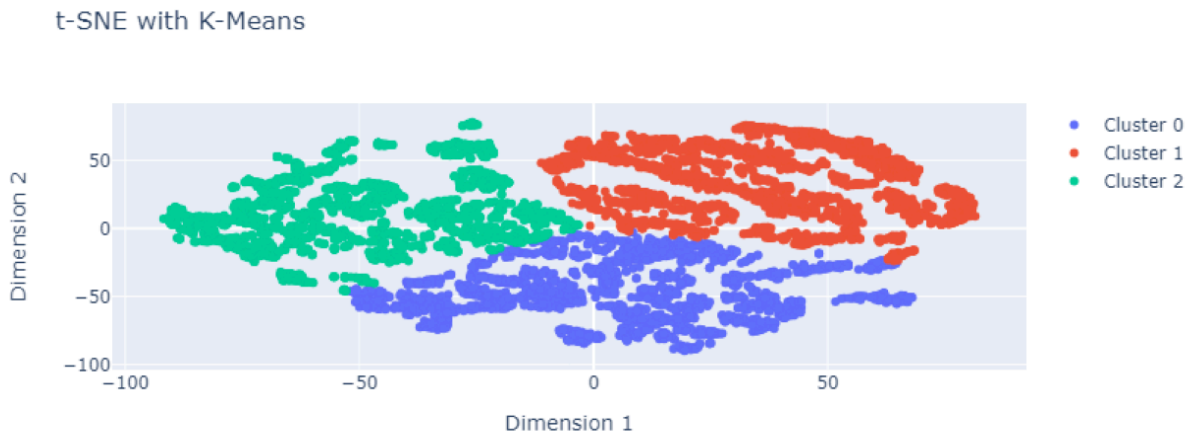
Insights:

- Cluster 0 shows the lowest mean of the fat as well as the fatty acid, but moderate values for other nutrition.
- Cluster 1 contains the highest mean of energy values followed by the highest values of carbohydrates.
- Cluster 2 has high value of energy values and highest value of protein, showing that it is a good suggested food group for fitness people

Results for Same Dataset Using Different Techniques

Dataset Used: Filling missing values with median of column

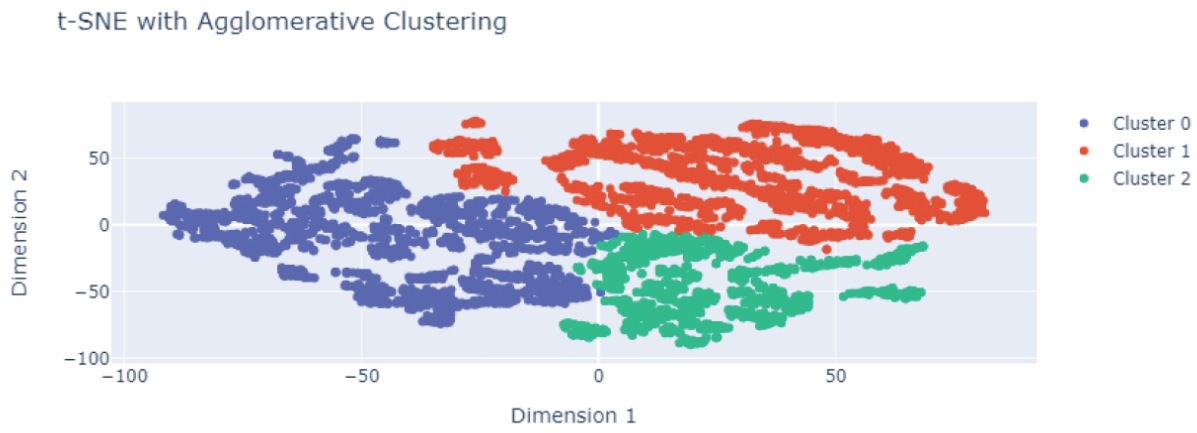
Technique 1: K-Means Clustering



Insights:

- Cluster 0 shows the lowest mean of the fat as well as the fatty acid, but moderate values for other nutrition.
- Cluster 1 contains the highest mean of energy values followed by the highest values of carbohydrates.
- Cluster 2 has high value of energy values and highest value of protein, showing that it is a good suggested food group for fitness people

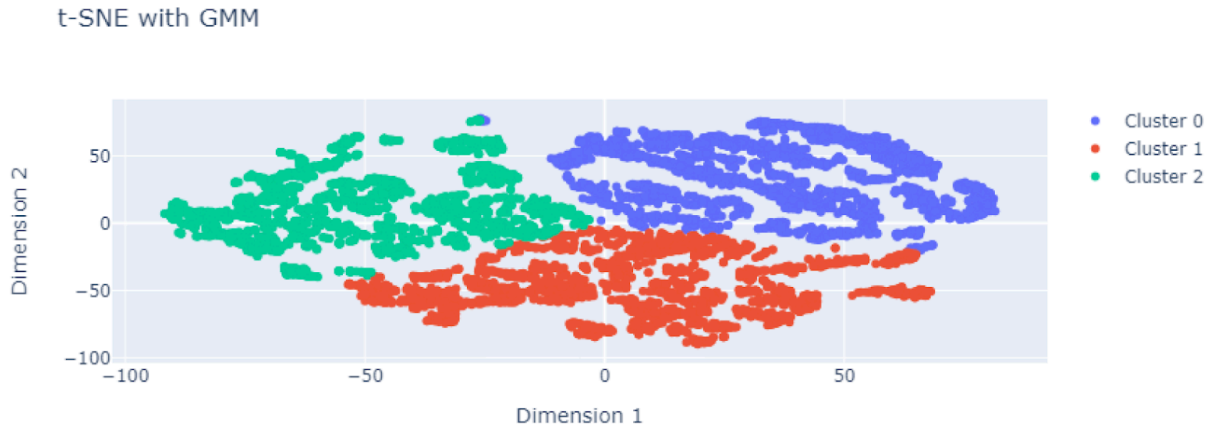
Technique 2: Agglomerative Clustering



Insights:

- Cluster 0: Among all the food clusters, this one has the lowest fat content, with lipids and three fatty acids each around value of 0.3 (excluding maximum value). This implies that there may be healthy foods in this food cluster.
- Cluster 1: This cluster seems to be a low-carb meal cluster, as Q1 and Q2 have carbohydrates values of 0. The greatest value of carbohydrates was about 0.75g (after log tansform and min max scaled), which is the lowest among all clusters.
- Cluster 2: This cluster may consist of foods high in energy because the first quartile of data was around 0.84, the highest of all the clusters. Based on the three fatty acid compositions, this cluster indicates that the meal in this cluster is maximal in fat. The fact that this food cluster has the highest value of total lipids among all clusters serves as additional support.

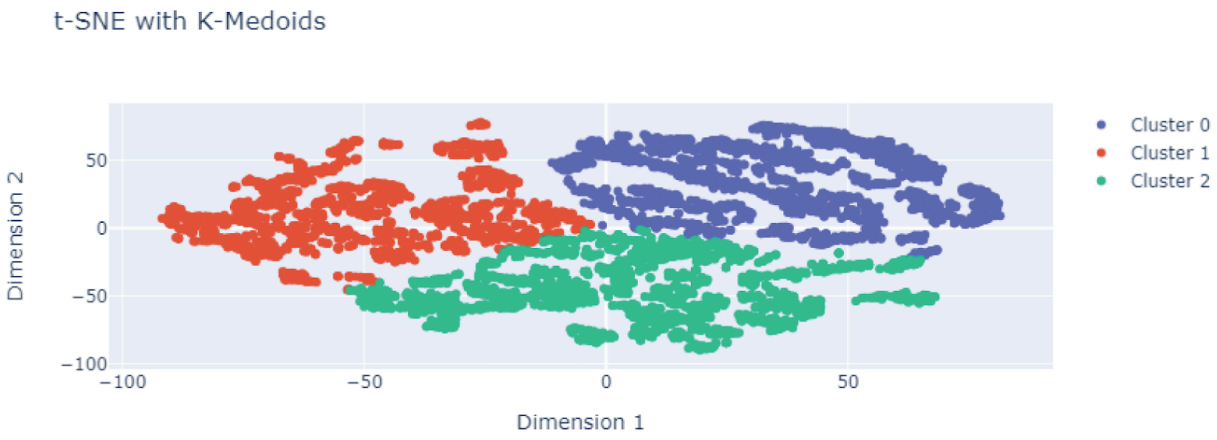
Technique 3: Gaussian Mixure Model (GMM)



Insights:

- Cluster 0: With a maximum value of 1 and the highest reported protein Q1 of 0.659, this cluster appears to have categorized foods based on protein.
- Cluster 1: This is a high-energy cluster with the first quartile of the data lying at 0.83 (the highest among all clusters).
- Cluster 2: Foods that fall under this category are thought to be lower in fat, or more healthful overall. This is due to the fact that 75% of the nutritional data for fats are less than 0.20 (0.17, 0.13, 0.11, 0.10 for lipid and 3 fatty acids, respectively).

Technique 4: Partitioning Around Medoids



Insights:

- Cluster 0: This food group is low in carbs and high in protein, as seen by the values of the two nutrients, which are 0.66 and 0.74 for protein and 0 to 0.08 for carbohydrates, respectively.
- Cluster 1: Foods in this cluster have the lowest fat content, making them considered healthful.
- Cluster 3: This food cluster appeared to be food that provides more energy due to high carbohydrate composition (around 0.73 to 0.98) as compared to other food clusters.

Result and Discussions

As discussed in the section on Evaluation Metrics, the Silhouette Score is a crucial evaluation metrics that assess the quality of clusters. The higher the Silhouette Score, the higher the quality of each cluster. Hence, the team sought to evaluate the clustering performance through a comparison of the Silhouette Score obtained from different datasets. The results for each model were compiled in the table below.

Table 6: Comparison of Clustering Techniques for Different Datasets

Dataset	Best Clustering Technique	Silhouette Score
Filling missing values with specific value, 0	K-Medoids	0.43088004
Filling missing values with mean of column	K-Means	0.41820893
Filling missing values with median of column	K-Medoids	0.4359154
Filling missing value with mode of column	K-Medoids	0.43088004
Filling missing value with forward fill method	K-Means	0.4257467
Filling missing value with backward fill method	K-Means	0.42386162
Filling missing values with linear interpolation method	K-Means	0.42919
Remove the rows with missing values	K-Means	0.4268393

Based on the statistical analysis provided, both K-Medoids and K-Means emerged as the top-performing clustering techniques for this task. This conclusion is drawn from the observation that K-Medoids was deemed the best in three datasets, while K-Means prevailed in five, with no instances favoring the other two clustering methods. Furthermore, the dataset filled with the median of the columns yielded the **highest Silhouette Score of 0.4359154**, indicating that K-Medoids was the most effective clustering technique when compared to Hierarchical Clustering, Gaussian Mixture Model, and K-Means.

Furthermore, the optimal number of clusters according to the specific dataset, is shown to be 3, where each cluster is examined as follows:

- Food group with a high protein content and minimal carbs
- Food groupings with reduced fat content
- Food group of high-energy and carbohydrate

Conclusion and Technical Future Recommendations

This project aims to identify the most effective clustering algorithm for grouping foods based on their nutritional content, particularly focusing on macronutrients. On average, the data reveals approximately three distinct food groups characterized as follows:

- A food group with low carbohydrates and high protein content.
- A food group with minimal fat content, comprising very small amounts of lipids, monounsaturated fatty acids, polyunsaturated fatty acids, and saturated fatty acids.
- A food group that contributes significantly to energy and carbohydrate intake.

Identifying these distinct food groups is crucial for facilitating balanced dietary choices, aligning with the research objectives. By grouping similar nutritional profiles, the analysis yields valuable insights into food nutrition.

However, evaluating clustering algorithms poses challenges as each iteration may produce different results. This variability is inherent to unsupervised machine learning, which aims to uncover hidden patterns in data. Moreover, assessing generated clusters is subjective and may introduce bias, necessitating domain expertise. The interpretation task of each cluster can be time-intensive, further underscoring the complexities of this task.

In terms of prospective recommendations, delving into micronutrients alongside macronutrients, as opposed to focusing solely on macronutrients, could greatly enhance the depth of this study in the future. The methodology for determining the optimal number of clusters was somewhat restricted by the self-defined parameter search, particularly concerning the range for the number of clusters, due to limitations in resources. Hence, it is recommended to bolster hardware resources to facilitate attaining the optimal parameters for each modelling technique.

Reference

U.S. Healthcare Data. n.d.. Kaggle. https://www.kaggle.com/datasets/maheshdadhich/us-healthcare-data/?select=Nutritions_US.csv

Atsa'am, D. D., Oyelere, S. S., Balogun, O. S., Wario, R., & Blamah, N. (2021). K-means cluster analysis of the West African species of cereals based on nutritional value composition. *African Journal of Food, Agriculture, Nutrition and Development, 21*`(01), 17195–17212. https://doi.org/10.18697/ajfand.96.19775

Balakrishna, Y., Manda, S., Mwambi, H., & Van Graan, A. (2023). Determining classes of food items for health requirements and nutrition guidelines using Gaussian mixture models. *Frontiers in Nutrition, 10*. https://doi.org/10.3389/fnut.2023.1186221

Bathula, C. P.. (2023, June 26). *Everything to know about Hierarchical Clustering; Agglomerative Clustering & Divisive Clustering.* https://medium.com/@chandu.bathula16/everything-to-know-about-hierarchical-clustering-agglomerative-clustering-divisive-clustering-badf31ae047

Botyarov, M., & Miller, E. E. (2022). Partitioning around medoids as a systematic approach to generative design solution space reduction. Results in Engineering, 15, 100544. https://doi.org/10.1016/j.rineng.2022.100544

Budiaji, W., Riyanto, R. A., & Suherna. (2021). The application of medoid-based cluster validation in desirable dietary pattern data. *Journal of Physics: Conference Series, 1863*(1), 012069. https://doi.org/10.1088/1742-6596/1863/1/012069

Da Silva Torres, E. a. F., Garbelotti, M. L., & Neto, J. M. M. (2006). The application of hierarchical clusters analysis to the study of the composition of foods. *Food Chemistry, 99*(3), 622–629. https://doi.org/10.1016/j.foodchem.2005.08.032

Dalimunthe, S. (2022). Implementation of agglomerative hierarchical clustering based on the classification of food ingredients content of nutritional substances. *www.academia.edu.* https://www.academia.edu/71761646/Implementation_of_Agglomerative_Hierarchical_Clustering_Based_on_The_Classification_of_Food_Ingredients_Content_of_Nutritional_Substances

Datalab, A.. (2019, Jan 16). *What is Bayesian Information Criterion (BIC)?*. https://medium.com/@analyttica/what-is-bayesian-information-criterion-bic-b3396a894be6

Devlin, U., McNulty, B., Nugent, A. P., & Gibney, M. J. (2012). The use of cluster analysis to derive dietary patterns: methodological considerations, reproducibility, validity and the effect of energy mis-reporting. *Proceedings of the Nutrition Society, 71*(4), 599–609. https://doi.org/10.1017/s0029665112000729

Nutrition. n.d.. World Health Organization. https://www.who.int/health-topics/nutrition#tab=tab_1

O'Hara, C., O'Sullivan, A., & Gibney, E. R. (2022). A clustering approach to Meal-Based analysis of dietary intakes applied to population and individual data. *The Journal of Nutrition, 152*(10), 2297–2308. https://doi.org/10.1093/jn/nxac151

The silhouette method. (n.d.). O REILLY. https://www.oreilly.com/library/view/training-systems-using/9781838823733/f6058e32-7d77-4256-abb5-ebae0e679d56.xhtml

Treitler, J. T., Tekle, S., Ushe, J., Zanin, L., Capshaw, T. L., Tardieu, G., Libin, A., & Zeng, Q. (2023). Characterizing nutrient patterns of food items in adolescent diet using data from a novel citizen science project and the US National Health and Nutrition Examination Survey (NHANES). *Frontiers in Nutrition, 10*. https://doi.org/10.3389/fnut.2023.1233141

Wakayama, R., Takasugi, S., Honda, K., & Kanaya, S. (2023). Application of a Two-Dimensional Mapping-Based visualization technique: *Nutrient-Value-Based food grouping. Nutrients, 15*(23), 5006. https://doi.org/10.3390/nu15235006

What is the MinMax Scaler?. (2023, May 17). Data Base Camp. https://databasecamp.de/en/ml/minmax-scaler-en