Group 6

Dataset from Open Food Facts

# Clustering Food Products based on Nutritional Attributes

Presented by Tay Wei Rong & Lim Jo Sun

Open Food Facts

# Overview

Open Food Facts

# Problem Statement

## Problem

- Difficulties in accessing accurate nutritional content and relevant nutritional scores of food.
- Challenges in identifying potential health impacts associated with different food categories.

## Considerations

- **Data Quality**: Ensuring data quality through validation processes is paramount to maintain the integrity of analyses.
- **Nutritional Patterns**: Reveal patterns in dietary preferences, helping identify trends in consumption and nutritional habits.
- **Health Impacts**: Provide insights into the potential health impacts associated with different food categories.

Open Food Facts

# Objectives

- To Clean and pre-process the Open Food Facts dataset for reliable nutritional information.
- To Apply 3 clustering algorithms to group food products based on nutritional content, considering dietary preferences and health implications.
- To Analyse clusters to identify nutritional patterns, dietary trends, and health associations.
- To Visualize clusters using dimensionality reduction and data visualization techniques for easier interpretation.
- To Evaluate clustering algorithms' effectiveness using metrics like silhouette score and Davies–Bouldin index.

Open Food Facts

# Literature Reviews

**01**
K-Means

**02**
Fuzzy C-Means

**03**
DBSCAN

Open Food Facts

# 01. K-Means

## Overview

- K-Means clustering is widely utilized in food and beverage/nutrition research for its effectiveness in analysing data and identifying patterns.
- Its popularity stems from its centroid-based approach, which efficiently partitions data into clusters.
- Enables researchers to segment food products based on their nutritional content, dietary preferences, ingredients, and health implications, facilitating targeted analyses and insights.

## Effectiveness

- **Accurate Clustering:** K-Means algorithm effectively groups foods based on nutrition attributes.
- **Personalized Recommendations:** Enables personalized diet recommendations by clustering foods according to individual nutritional needs.

## Matrices

- Inertia
- Silhouette Score
- Davies–Bouldin Index
- Calinski-Harabasz Index

## Steps

1. Initialisation
2. Assignment step
3. Update step
4. Convergence Check
5. Finalisation

Sgroi, F., Sciortino, C., Baviera-Puig, A., & Modica, F. (2024). Analyzing consumer trends in functional foods: A cluster analysis approach. Journal of Agriculture and Food Research, 101041. https://doi.org/10.1016/j.jafr.2024.101041

Open Food Facts

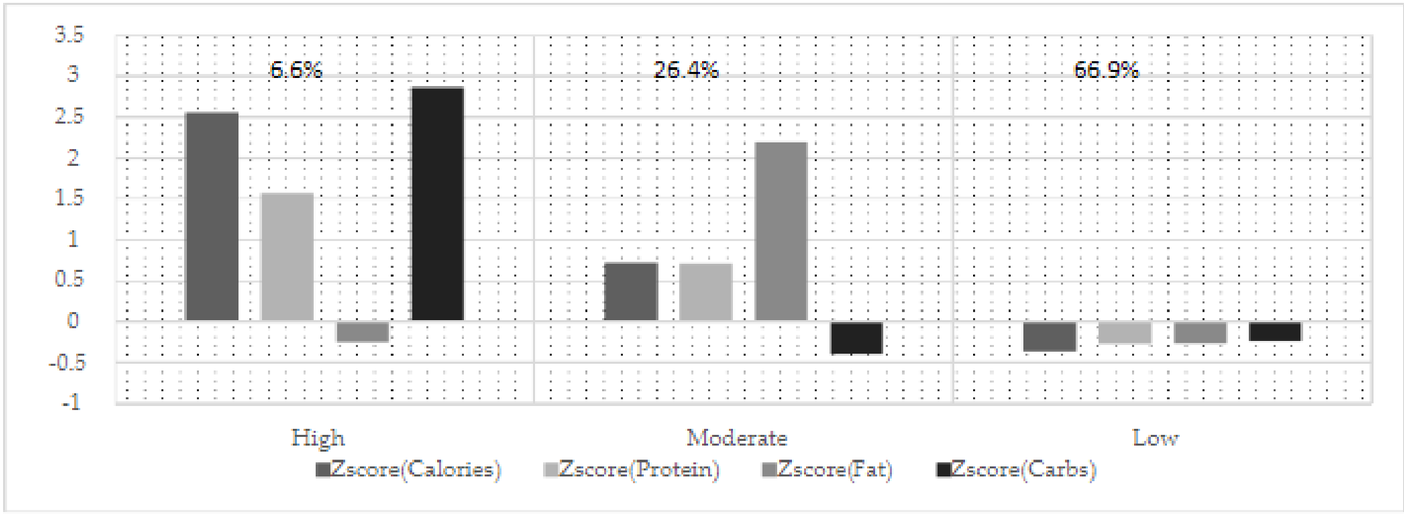
# 01. K-Means

## Application 1: Identifying Food Consumption Patterns

**Goal:** To categorise individuals into clusters based on their dietary intake patterns.

- Allows for the identification of groups of individuals with similar food consumption behaviours
- Enabling researchers to analyse various dietary preferences within the population.

**Figure 1**: Average Z-score of Macro-Nutrient category K-Means clusters.




Qasrawi, Radwan. (2021). Cluster analysis for food group consumption patterns in a national sample of Palestinian schoolchildren: Evidence from HBSC Survey 2013-2014. Al-Quds Journal for Academic Research. https://www.researchgate.net/publication/350560260_Cluster_analysis_for_food_group_consumption_patterns_in_a_national_sample_of_Palestinian_schoolchildren_Evidence_from_HBSC_Survey_2013-2014

Open Food Facts

# 01. K-Means

## Application 2: Deriving Dietary Patterns

**Goal:** To categorize individuals into specific dietary clusters based on their nutrient intake.

- Enables researchers to identify and characterise distinct dietary patterns within populations
- Facilitating the analysis of dietary habits and their potential associations with health outcomes.

Sauvageot, N., Schritz, A., Leite, S., Alkerwi, A., Stranges, S., Zannad, F., Streel, S., Hoge, A., Donneau, A. F., Albert, A., & Guillaume, M. (2017). Stability-based validation of dietary patterns obtained by cluster analysis. Nutrition Journal, 16(1). https://doi.org/10.1186/s12937-017-0226-9

Open Food Facts

# 02. Fuzzy C-Means

## Overview

- An algorithm allowing soft assignments of data points to clusters, making it ideal for scenarios where data may belong to multiple clusters.
- Its versatility lies in accommodating the degree of membership of data points to clusters, providing a nuanced understanding of data distribution.

## Effectiveness

FCM clustering excels in handling intricate datasets common in food and nutrition research, where data points may exhibit varying degrees of similarity to multiple clusters.

## Matrices

- Fuzzy Partition Coefficient (FPC)
- Silhouette Score
- Davies-Bouldin Index
- Calinski-Harabasz Index

## Steps

1. Initialisation
2. Membership Degree Calculation
3. Update Centroids
4. Convergence Check
5. Finalisation

Hashemi, S. E., Gholian-Jouybari, F., & Hajiaghaei–Keshteli, M. (2023). A fuzzy C-means algorithm for optimizing data clustering. Expert Systems With Applications, 227, 120377. https://doi.org/10.1016/j.eswa.2023.120377

Krasnov, D., Davis, D. A., Malott, K., Chen, Y., Shi, X., & Wong, A. (2023). Fuzzy C-Means Clustering: A review of applications in breast cancer detection. Entropy, 25(7), 1021. https://doi.org/10.3390/e25071021

Open Food Facts

# 02. Fuzzy C-Means

## Application 1: Clustering Food Productivity

**Goal:** To categorize food productivity data into clusters, enabling the identification of distinct patterns

- This facilitates the understanding of factors affecting food productivity and supports decision-making processes related to agricultural practices, resource allocation, and policy formulation.

**Table 17. Comparison of Rice Productivity for Fuzzy C-Means Clustering**

| $x_i$ | $C_1$ | 2014 (Ku/Ha) | 2015 (Ku/Ha) | $x_i$ | $C_2$ | 2014 (Ku/Ha) | 2015 (Ku/Ha) | $x_i$ | $C_3$ | 2014 (Ku/Ha) | 2015 (Ku/Ha) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_4$ | Riau | 36,35 | 36,63 | $x_1$ | Aceh | 48,39 | 50,56 | $x_{11}$ | DKI Jakarta | 53,86 | 55,95 |
| $x_5$ | BangkaBelitung | 23,62 | 22,85 | $x_2$ | SumateraUtara | 50,62 | 51,74 | $x_{12}$ | Jawa Barat | 58,82 | 61,22 |
| $x_{10}$ | Kepulauan Riau | 36,44 | 36,46 | $x_3$ | SumateraBarat | 50,06 | 50,25 | $x_{13}$ | Jawa Tengah | 53,57 | 60,25 |
| $x_{19}$ | NTT | 33,46 | 35,61 | $x_6$ | Jambi | 45,53 | 44,31 | $x_{14}$ | DIYogyakarta | 57,87 | 60,65 |
| $x_{20}$ | KalimantanBarat | 30,35 | 29,40 | $x_9$ | SumateraSelatan | 45,26 | 48,67 | $x_{15}$ | Jawa Timur | 59,81 | 61,13 |
| $x_{21}$ | KalimantanTengah | 34,57 | 35,07 | $x_7$ | Bengkulu | 40,20 | 44,92 | $x_{16}$ | Banten | 52,95 | 56,61 |
| $x_{24}$ | Kalimantan Utara | 36,05 | 27,27 | $x_8$ | Lampung | 51,18 | 51,49 | $x_{17}$ | Bali | 60,12 | 62,14 |
| $x_{32}$ | Maluku Utara | 34,01 | 35,11 | $x_{18}$ | NTB | 48,80 | 51,71 | $x_{28}$ | Gorontalo | 50,20 | 55,51 |
| Rice Productivity in $C_1$ < Indonesia = decrease | | | | $x_{22}$ | KalimantanSelatan | 42,05 | 41,87 | Rice Productivity in $C_3$ > Indonesia = increase | | | |
| | | | | $x_{23}$ | Kalimantan Timur | 42,55 | 41,20 | | | | |
| | | | | $x_{25}$ | Sulawesi Utara | 48,91 | 49,05 | | | | |
| | | | | $x_{26}$ | Sulawesi Tengah | 46,54 | 48,57 | | | | |
| | | | | $x_{27}$ | Sulawesi Selatan | 52,17 | 52,41 | | | | |
| | | | | $x_{29}$ | SulawesiTenggara | 46,84 | 47,07 | | | | |
| | | | | $x_{30}$ | Sulawesi Barat | 47,65 | 49,41 | | | | |
| | | | | $x_{31}$ | Maluku | 47,52 | 55,72 | | | | |
| | | | | $x_{33}$ | Papua Barat | 40,21 | 42,12 | | | | |
| | | | | $x_{34}$ | Papua | 43,09 | 43,95 | | | | |
| | | | | Rice Productivity in $C_2$ < Indonesia = decrease | | | | | | | |
| Indonesia | | 51,35 | 53,41 | Indonesia | | 51,35 | 53,41 | Indonesia | | 51,35 | 53,41 |

Adriyendi. (2016). Clustering using K-means and fuzzy C-means on food productivity. International Journal of u- and e- Service, Science and Technology. 9. 281-290. 10.14257/ijunesst.2016.9.12.26.

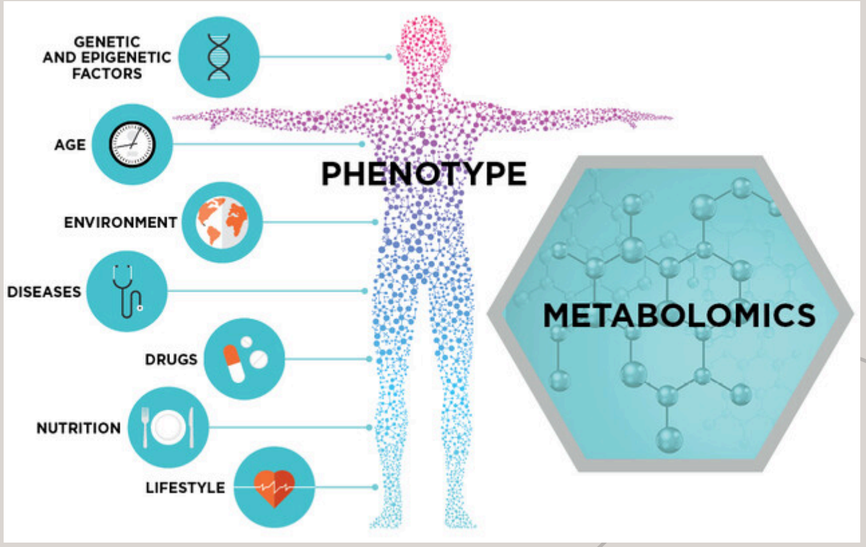
Open Food Facts

# 02. Fuzzy C-Means

## Application 2: Metabolomics Data Analysis

**Goal:** To analyse complex metabolomics data to identify patterns and group data points with soft assignments.

- This facilitates the discovery of underlying metabolic profiles and relationships within beverages, aiding in quality control, product development.
- Understanding the biochemical processes involved in beverage production.




Adriyendi. (2016). Clustering using K-means and fuzzy C-means on food productivity. International Journal of u- and e- Service, Science and Technology. https://www.researchgate.net/publication/24395807_Application_of_Fuzzy_c-Means_Clustering_in_Data_Analysis_of_Metabolomics

Open Food Facts

# 03. DBSCAN

## Overview

- A leading algorithm in density-based clustering, renowned for its capability to identify clusters of arbitrary shape and effectively handle noise.
- Its key advantage lies in its ability to discover clusters based on the density of data points, rather than assuming a fixed number of clusters or specific shapes.

IEEE Journals & Magazine. (2021).Comparative analysis Review of pioneering DBSCAN and successive Density-Based clustering algorithms. | IEEE Xplore. https://ieeexplore.ieee.org/document/9453785

## Effectiveness

- It suitable for analysing large-scale datasets commonly encountered in food, beverages, and nutrition research.
- Its ability to process data efficiently without requiring a predetermined number of clusters enhances its applicability in diverse research contexts.

## Key Components

- Data Points
- Distance Matrix (Optional)
- Core Points, Border Points & Noise
- Cluster Assignments

## Steps

1. Parameter Selection
2. Neighbourhood Search
3. Core Point Identification
4. Cluster Expansion
5. Border Point Assignment
6. Noise Identification
7. Finalisation

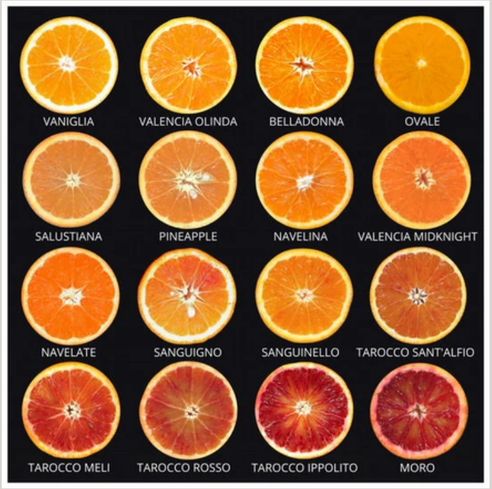
Open Food Facts

# 03. DBSCAN

## Application 1: Quality Assessment of Orange Hybrids

**Goal:** To classify orange hybrids into distinct quality classes based on their characteristics, enabling the assessment of oranges with similar attributes.

- This facilitates quality assessment processes within the food and beverage industry.
- Aiding in the selection and management of orange varieties for cultivation and consumption.




Alqorni, (2021). Application of Density Based Spatial Clustering Application With Noise (DBSCAN) in Determining the Quality of Keprok Orange and Siam Orange Hybrid in the Research Center of Orange and Subtropic Plants Batu City. Journal of Information Technology and Computer Science. https://www.researchgate.net/publication/351236137_Application_of_Density_Based_Spatial_Clustering_Application_With_Noise_DBSCAN_in_Determining_the_Quality_of_Keprok_Orange_and_Siam_Orange_Hybrid_in_the_Research_Center_of_Orange_and_Subtropic_Plants_B

Open Food Facts

# 03. DBSCAN

## Application 2: Analysing Nutritional Data

**Goal:** To identify and characterize dietary patterns within a population based on nutritional data.

- Detect clusters of arbitrary shapes and effectively handle outliers, researchers aim to uncover complex patterns in dietary habits and preferences.
- Facilitating a deeper understanding of nutrition-related behaviours and their potential impacts on health outcomes.

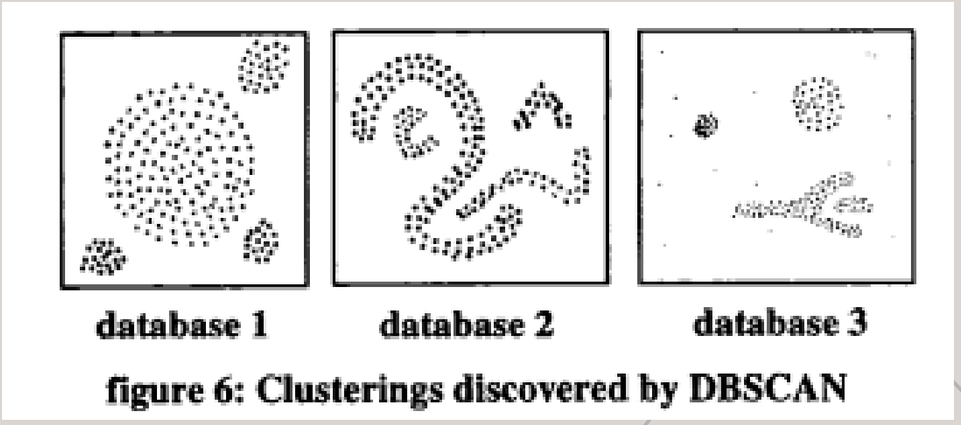


figure 6: Clusterings discovered by DBSCAN

Ester, M, Kriegel, P, Sander, J, Xu (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. https://file.biolab.si/papers/1996-DBSCAN-KDD.pdf

Open Food Facts

# Methodology

A. Data Collection and Preparation

B. Exploratory Data Analysis (EDA)

Open Food Facts

# Data Collection and Preparation

- Dataset Overview: 356,027 rows, 163 columns.
- Dropping Uninteresting Columns: Removal of text data, repeated information, and irrelevant columns.
- Handling Missing Data: Importance of complete data for accurate machine learning models.
- Data Validation and Outliers Detection: Impact of invalid data on model performance; validation and outlier detection techniques.

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. Journal of Big Data, 8(1). https://doi.org/10.1186/s40537-021-00516-9

Redman, T. C. (2018, April 3). If your data is bad, your machine learning tools are useless. Harvard Business Review. https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless

Open Food Facts

# Data Collection and Preparation

- Data Type Handling: Encoding categorical variables, standardizing numerical features for uniformity.
- Handling Inconsistent Data: Standardizing serving size, filling missing values.
- Handling Duplicate Data: Removing duplicated rows and columns for integrity and accuracy.

Data Headhunters (2024, January 7). Handling inconsistent data: Strategies for Standardization. https://dataheadhunters.com/academy/handling-inconsistent-data-strategies-for-standardization/

GeeksforGeeks. (2023, October 17). Handling inconsistent data. GeeksforGeeks. https://www.geeksforgeeks.org/handling-inconsistent-data/

Open Food Facts

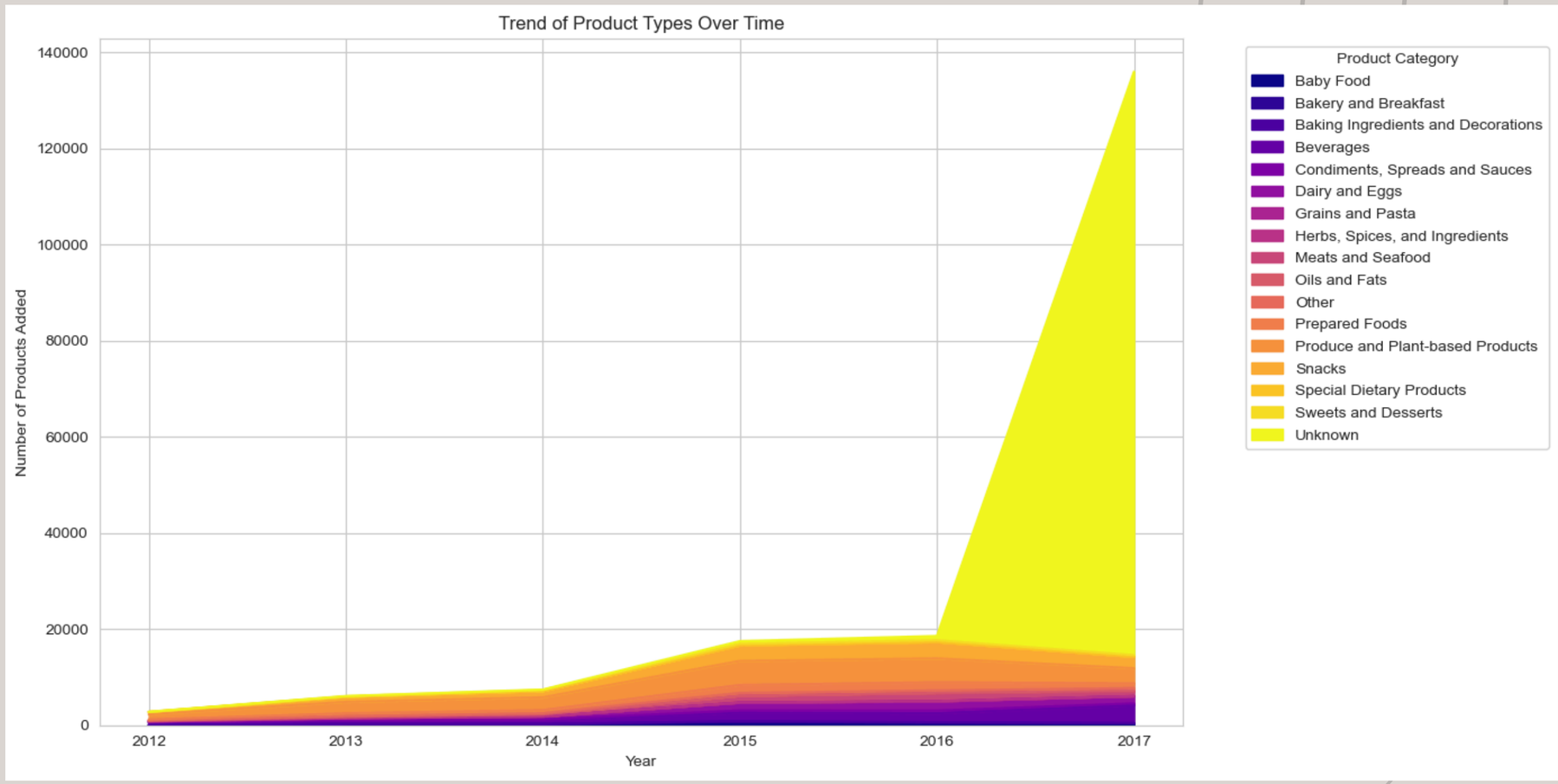
# Exploratory Data Analysis (EDA)

- A statistical method for analysing data sets to identify their main characteristics.
- The goal of EDA is to develop an understanding of the data, rather than to confirm statistical hypotheses.
- An iterative process that involves asking questions about the data, visualising and transforming the data, and using the findings to refine the questions.

JMP (n.d.). Exploratory data analysis. JMP. https://www.jmp.com/en_my/statistics-knowledge-portal/exploratory-data-analysis.html

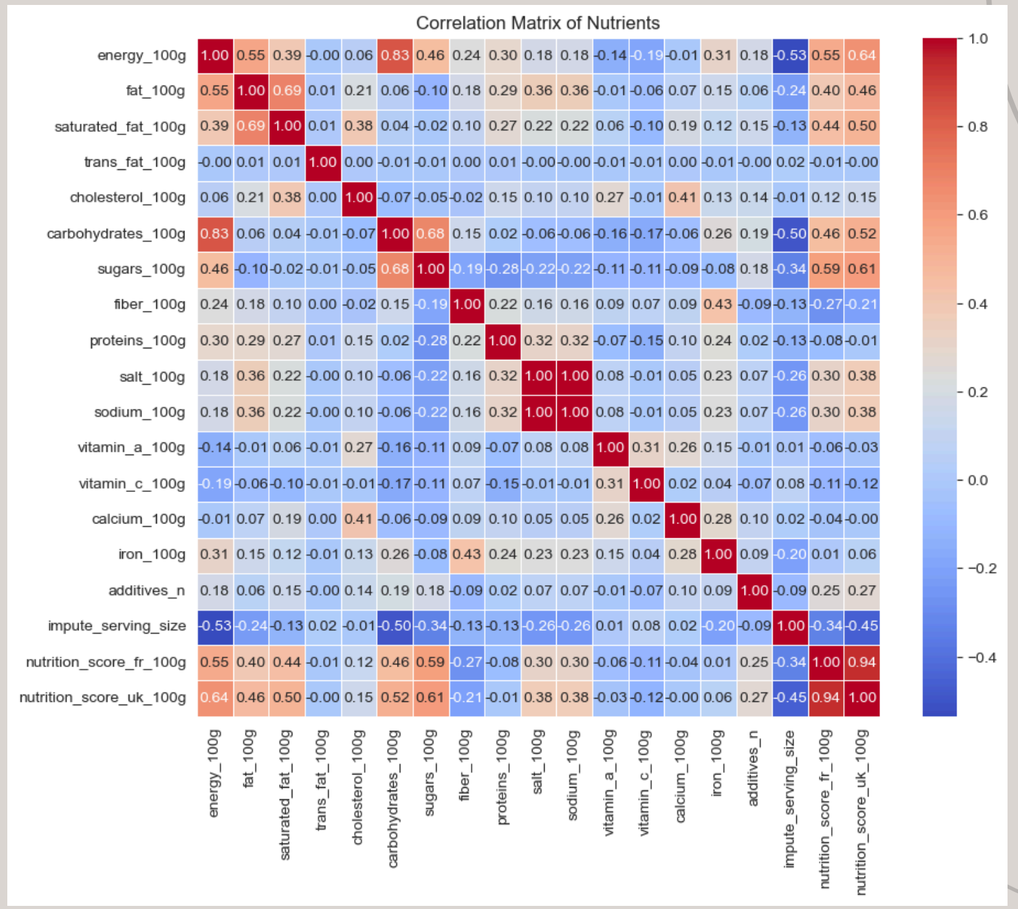
Open Food Facts

# Trend of Product Types Overtime



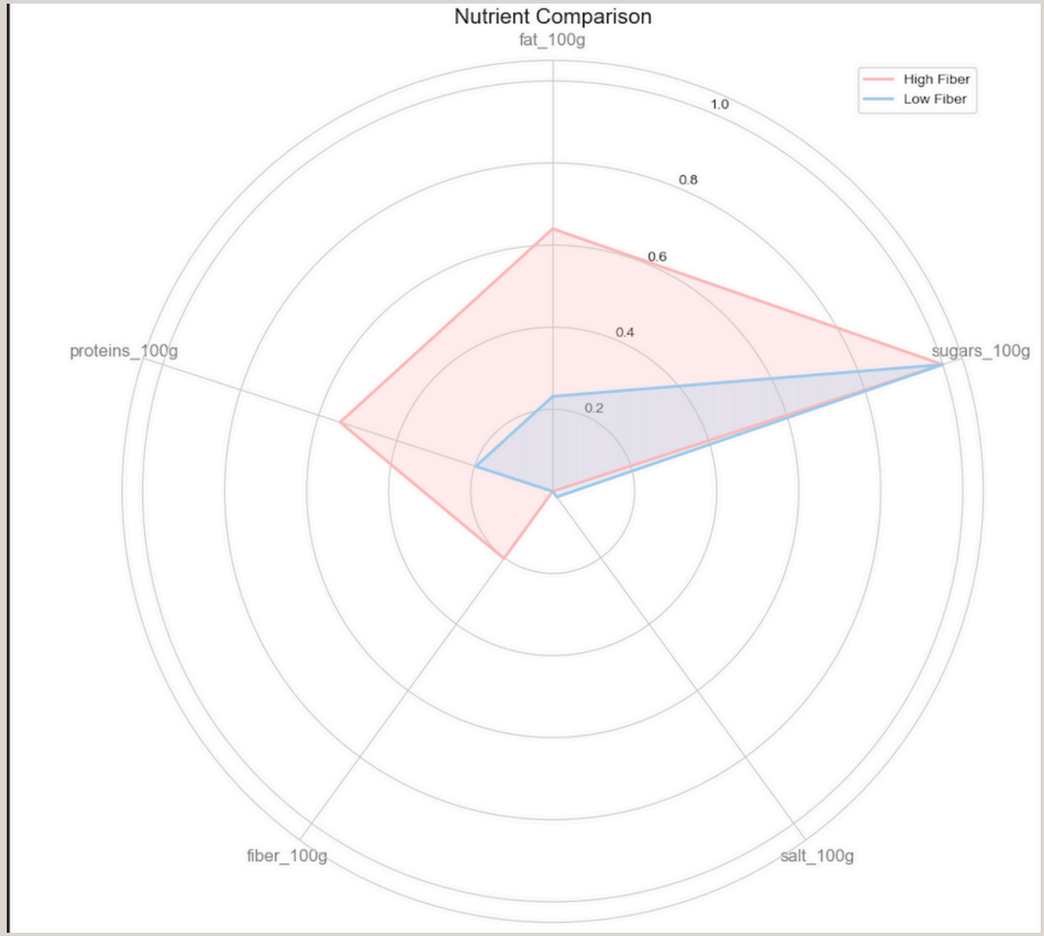
Open Food Facts

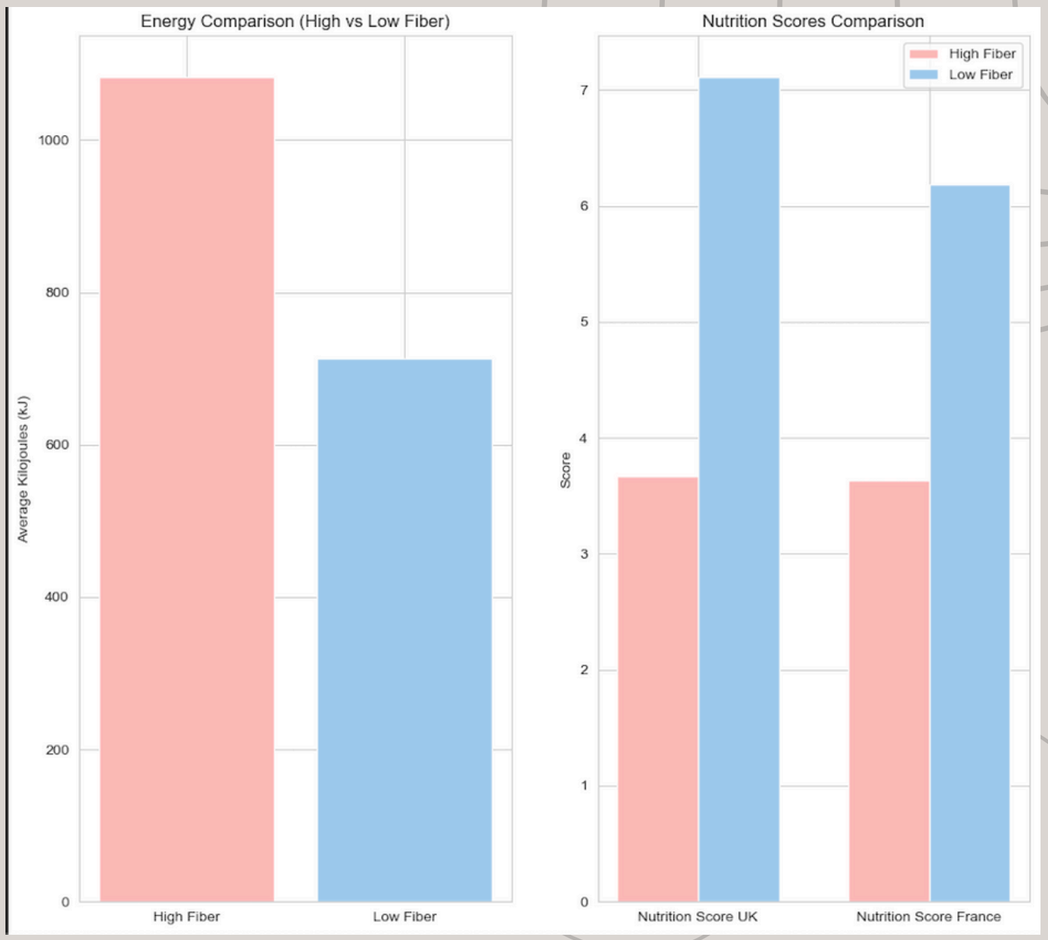
# Correlation Matrix of Nutrients



Open Food Facts

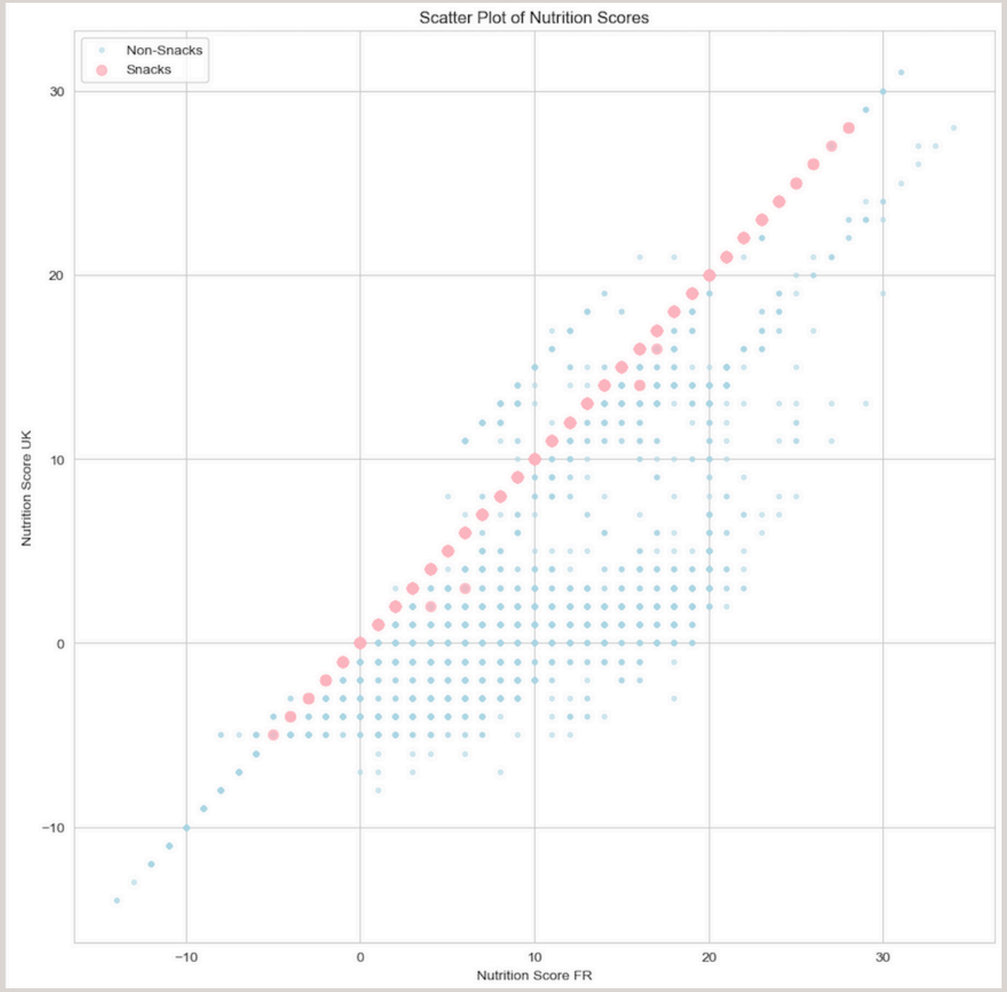
# Radar Chart of Nutrients Comparison

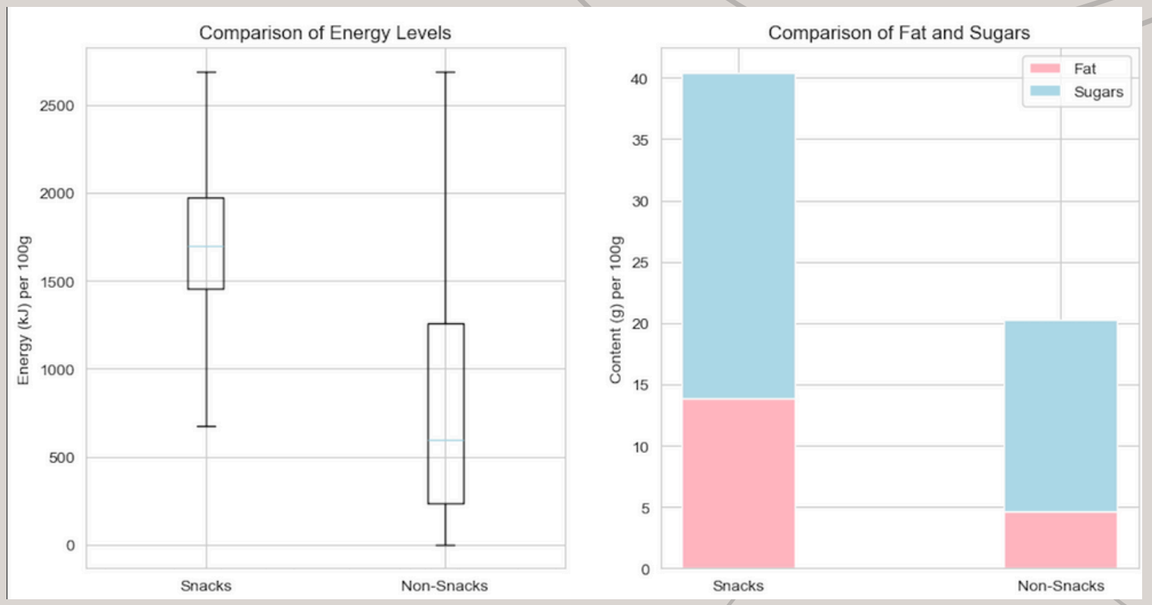


Open Food Facts

# Scatter Plot of Nutrition Score



Open Food Facts

# Results

**01**
K-Means

**02**
Fuzzy C-Means

**03**
DBSCAN

Open Food Facts

# K-Means

## Hyperparameters

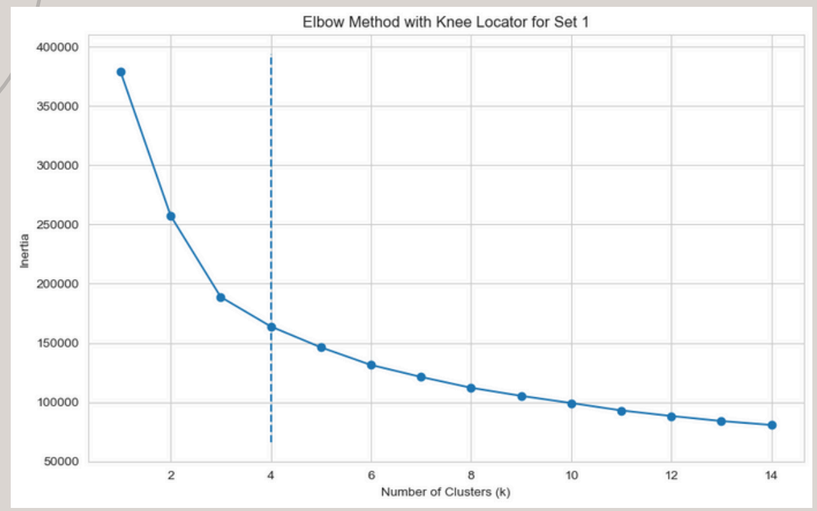
- n_clusters: Specifies the number of clusters to form.
- init: Specifies the method for initializing centroids.
- max_iter: Maximum number of iterations the algorithm will run.
- tol: Tolerance to declare convergence.
- algorithm: Specifies the algorithm to use. Options include ‘auto’, ‘full’, ‘elkan’. ‘auto’
- random_state: Controls the random seed to reproduce the results.

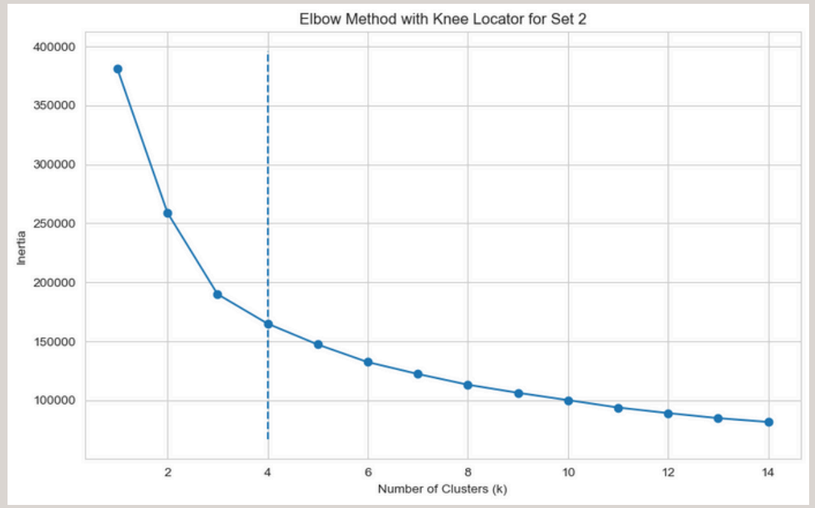
## Evaluation Metrics

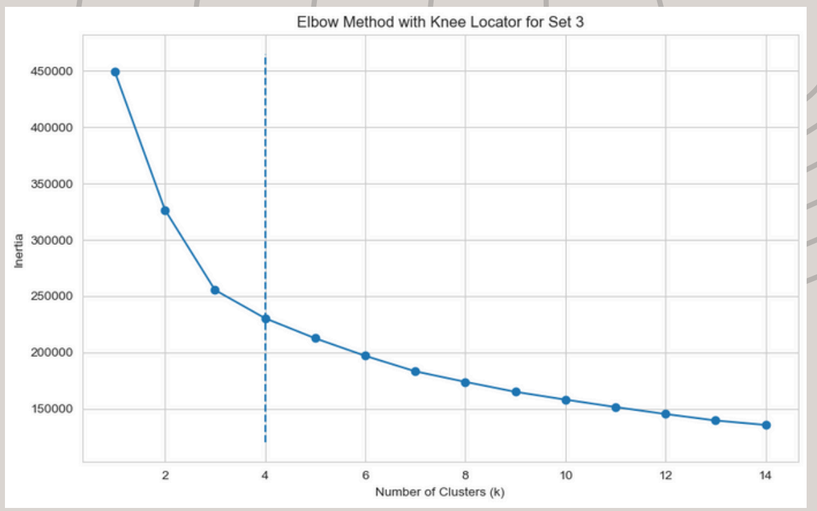
- Inertia: It measures the compactness of the clusters.
- Silhouette Score: It measures how similar an object is to its cluster compared to other clusters.
- Davies–Bouldin Index: It evaluates the clustering quality based on cluster compactness and separation.
- Calinski-Harabasz Index: Also known as the Variance Ratio Criterion, it measures the ratio of between-cluster dispersion to within-cluster dispersion.

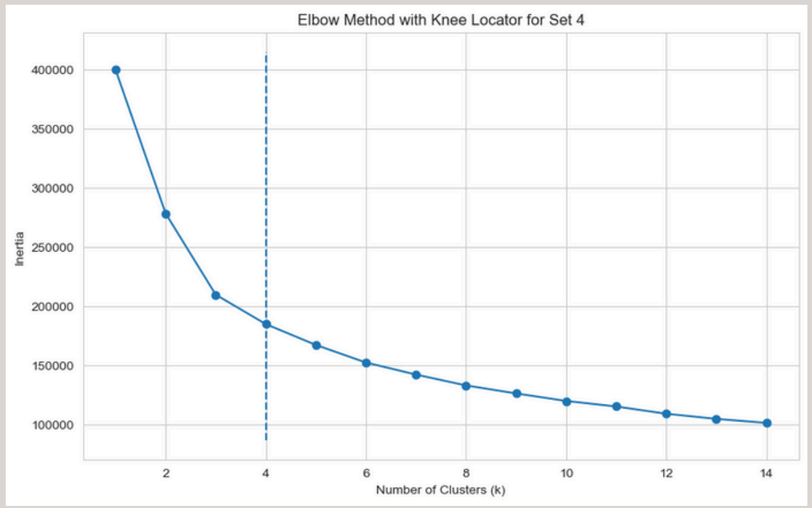
# Open Food Facts

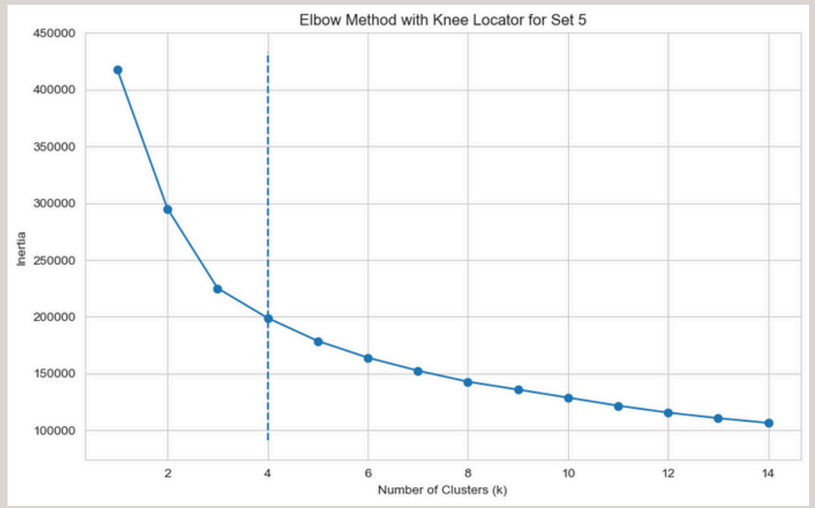
# **Feature Sets**

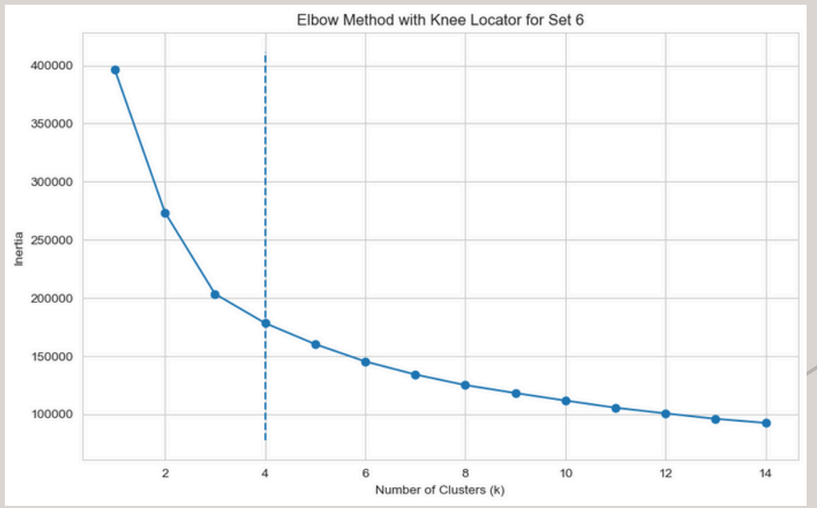


Open Food Facts

# Observations

- Upon observation, all models exhibit a similar optimal cluster count (4).
- They yield different inertia scores. Sets 2 through 6 consistently display higher inertia, suggesting a more dispersed dataset.
- In contrast, Feature Set 1 demonstrates superior clustering performance in terms of compactness.

Open Food Facts

# Similar Characteristics Across Feature Sets:

- **Dense Vertical Line:** Captures most variance in Principal Component 1.
- **Spread Out Clusters:** Represents diversity across features.
- **Colour Gradient (Cluster Labels):** Indicates visually overlapping clusters.

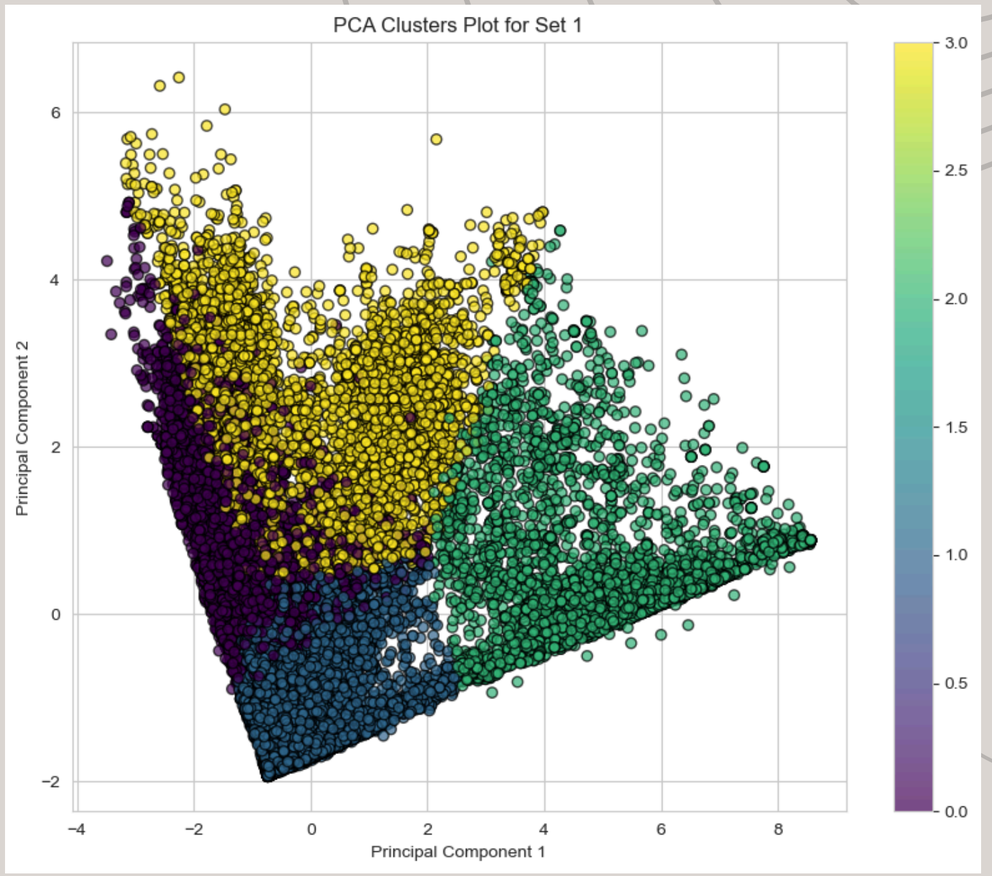
Open Food Facts

# PCA Cluster Plot

- **Dense Line:**
  a. Concentration of points around lower values of PC1 and PC2
  b. Indicates significant variance captured within dataset

- **Spread Out Clusters:**
  a. Cluster 1 (blue) at bottom left extending to center
  b. Represents products with distinct nutritional profiles

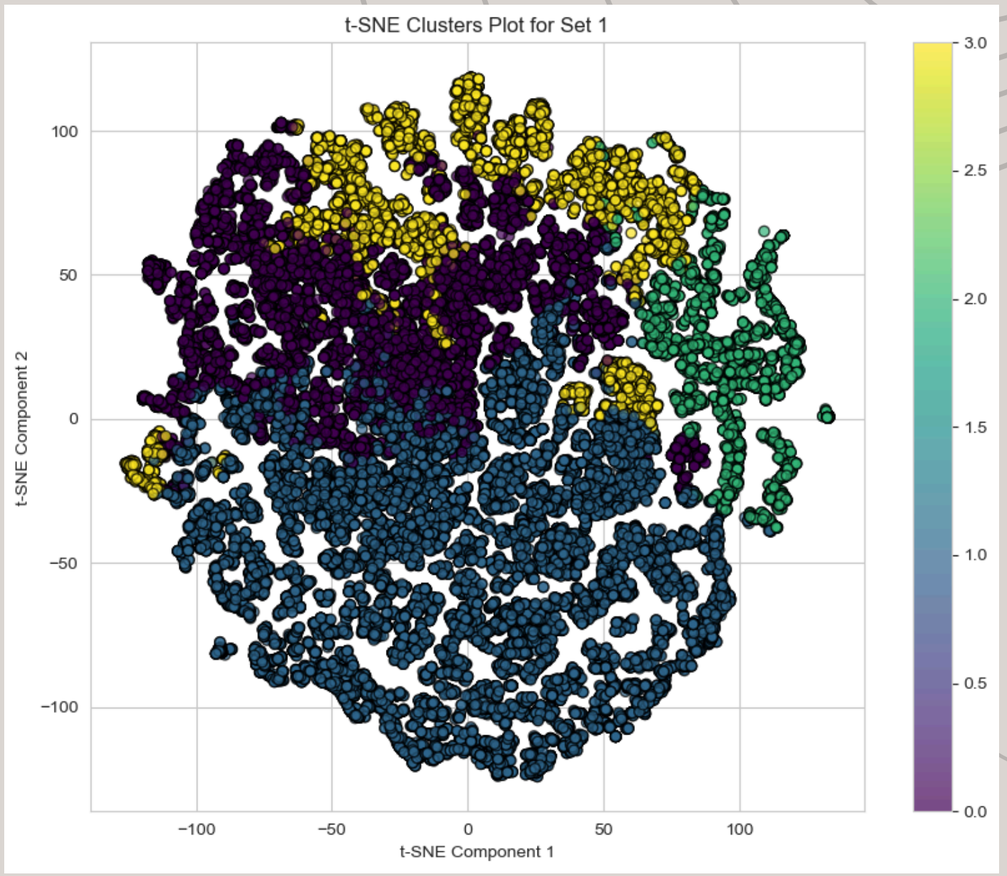
- **Overlapping:**
  a. Color gradient represents K-Means assigned clusters
  b. Visual overlap suggests need for clearer clustering



Open Food Facts

# t-SNE Cluster Plot

- **Density and Separation:** Some clusters tightly packed with overlap, others more dispersed.
- **Outliers:** Isolated points may represent unique properties.



Open Food Facts

# Comparison of Centroid (Before & After Tuning)

| Nutrient | Pre-tuned (Cluster 0) | Post-tuned (Cluster 0) | Difference |
|---|---|---|---|
| Energy (kJ) | Moderate (841) | Slightly lower (796) | -45 |
| Fat (g) | Moderately low (6.04) | Slightly higher (7.24) | +1.20 |
| Saturated Fat (g) | Lower (1.53) | Slightly higher (1.92) | +0.39 |
| Sugars (g) | Low (3) | Slightly higher (2.93) | -0.07 |
| Proteins (g) | Higher (11.47) | Slightly lower (12.11) | +0.64 |
| Fiber (g) | Moderate (1.24) | Slightly lower (0.92) | -0.32 |
| Salt (g) | High (1.25) | Lower (1.53) | +0.28 |

| Nutrient | Pre-tuned (Cluster 1) | Post-tuned (Cluster 1) | Difference |
|---|---|---|---|
| Energy (kJ) | Low (354) | Slightly higher (430) | +76 |
| Fat (g) | Very low (1.6) | Slightly higher (1.83) | +0.23 |
| Saturated Fat (g) | Very low (0.55) | Slightly higher (0.58) | +0.03 |
| Sugars (g) | Moderate (6.48) | Slightly higher (6.06) | -0.42 |
| Proteins (g) | Low (2.31) | Slightly higher (3.08) | +0.77 |
| Fiber (g) | Low (0.54) | Slightly higher (0.73) | +0.19 |
| Salt (g) | Very low (0.17) | Slightly higher (0.20) | +0.03 |

| Nutrient | Pre-tuned (Cluster 2) | Post-tuned (Cluster 2) | Difference |
|---|---|---|---|
| Energy (kJ) | High (1359) | Slightly lower (1348) | -11 |
| Fat (g) | Low (2.68) | Slightly lower (2.30) | -0.38 |
| Saturated Fat (g) | Low (0.94) | Lower (0.87) | -0.07 |
| Sugars (g) | Very high (62.44) | Slightly lower (62.85) | +0.41 |
| Proteins (g) | Low (2.03) | Slightly lower (1.96) | -0.07 |
| Fiber (g) | Low (0.61) | Similar (0.59) | -0.02 |
| Salt (g) | Very low (0.16) | Similar (0.16) | 0 |

| Nutrient | Pre-tuned (Cluster 3) | Post-tuned (Cluster 3) | Difference |
|---|---|---|---|
| Energy (kJ) | Very high (1469) | Slightly higher (1518) | +49 |
| Fat (g) | High (19.64) | Slightly lower (19.93) | -0.29 |
| Saturated Fat (g) | Moderate (4.90) | Moderate (4.86) | -0.04 |
| Sugars (g) | Moderate (11.72) | Slightly lower (13.22) | +1.50 |
| Proteins (g) | High (8.06) | Slightly lower (7.63) | -0.43 |
| Fiber (g) | Moderate (1) | Similar (1.07) | +0.07 |
| Salt (g) | Moderate (1.06) | Slightly lower (0.95) | -0.11 |

Open Food Facts

# Fuzzy C-Means

## Hyperparameters

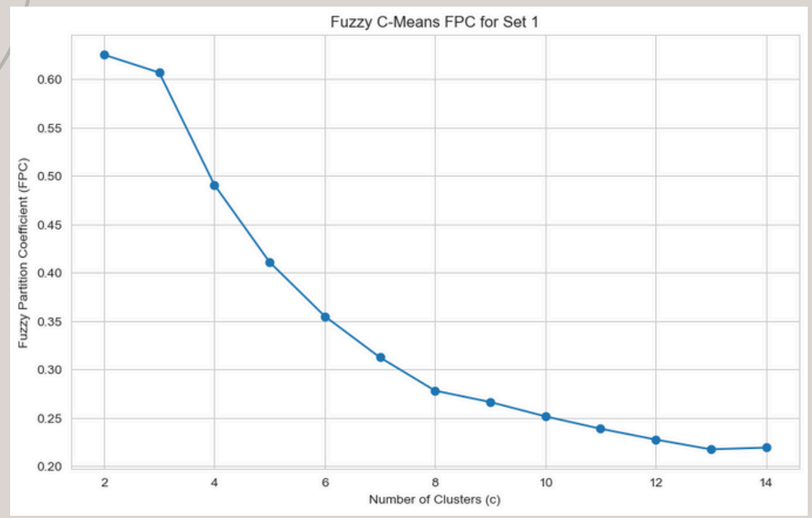
- Number of Clusters (K): Like K-Means, FCM requires specifying the number of clusters prior.
- Fuzziness Coefficient (m): The fuzziness coefficient controls the degree of fuzziness in the cluster assignments.
- Maximum Number of Iterations: FCM iteratively updates cluster centroids and membership values until convergence.
- Termination Criterion: FCM terminates when either the maximum number of iterations is reached.

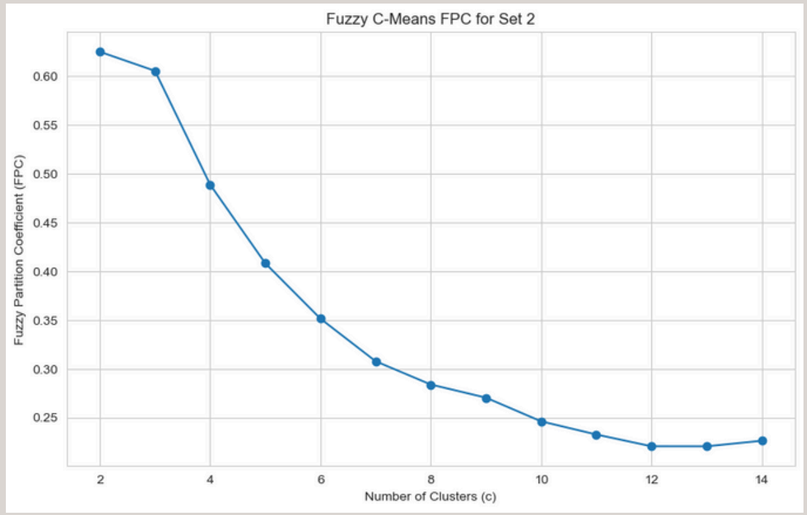
## Evaluation Metrics

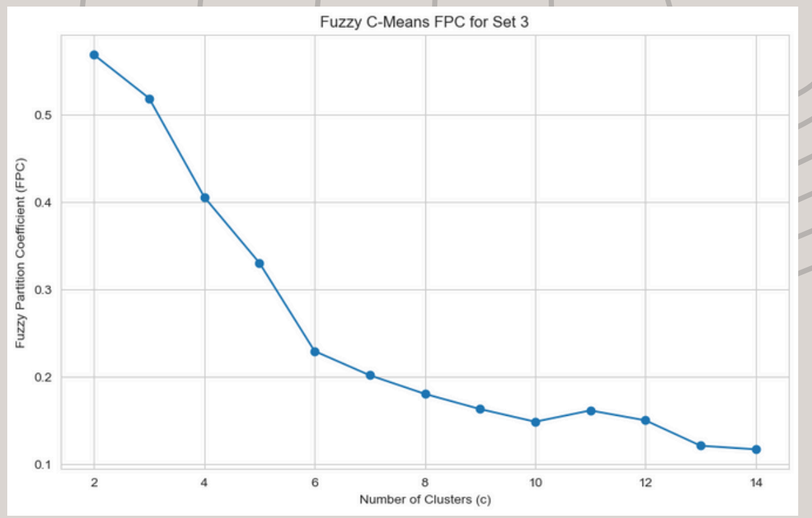
- Fuzzy Partition Coefficient (FPC): This metric measures the fuzziness of the clustering result.
- Silhouette Score: The Silhouette Score measures the compactness and separation of clusters.
- Davies-Bouldin Index: Assesses cluster quality by measuring the average similarity.
- Calinski-Harabasz Index: Measures the ratio of between-cluster dispersion to within-cluster dispersion for all clusters.

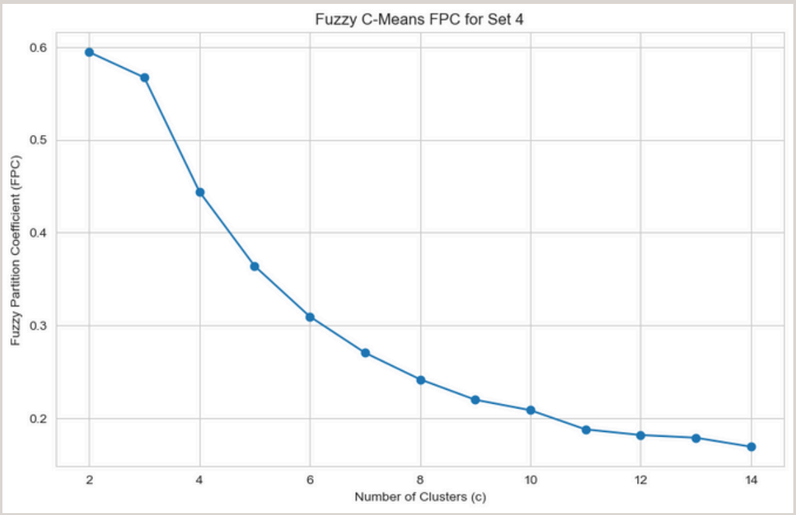
# Open Food Facts

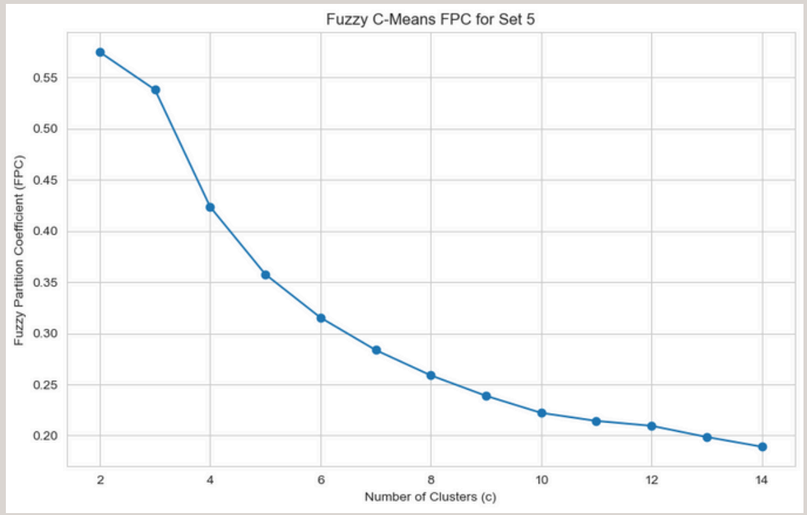
## **Feature Sets**

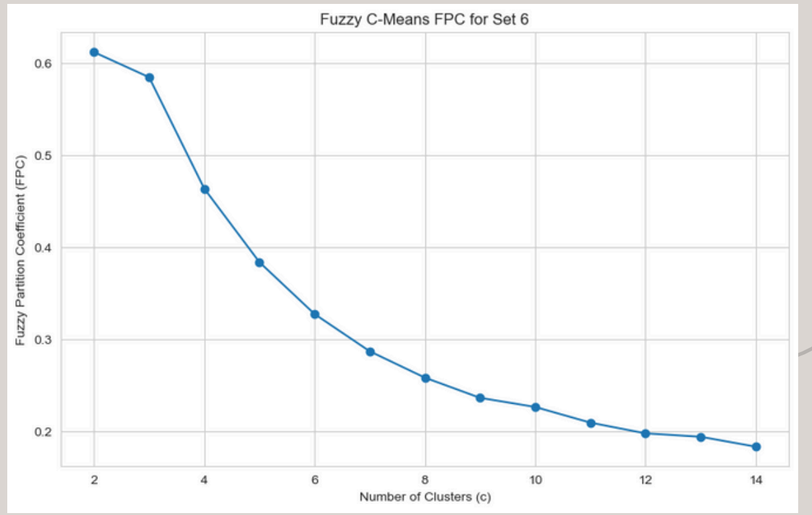


Open Food Facts

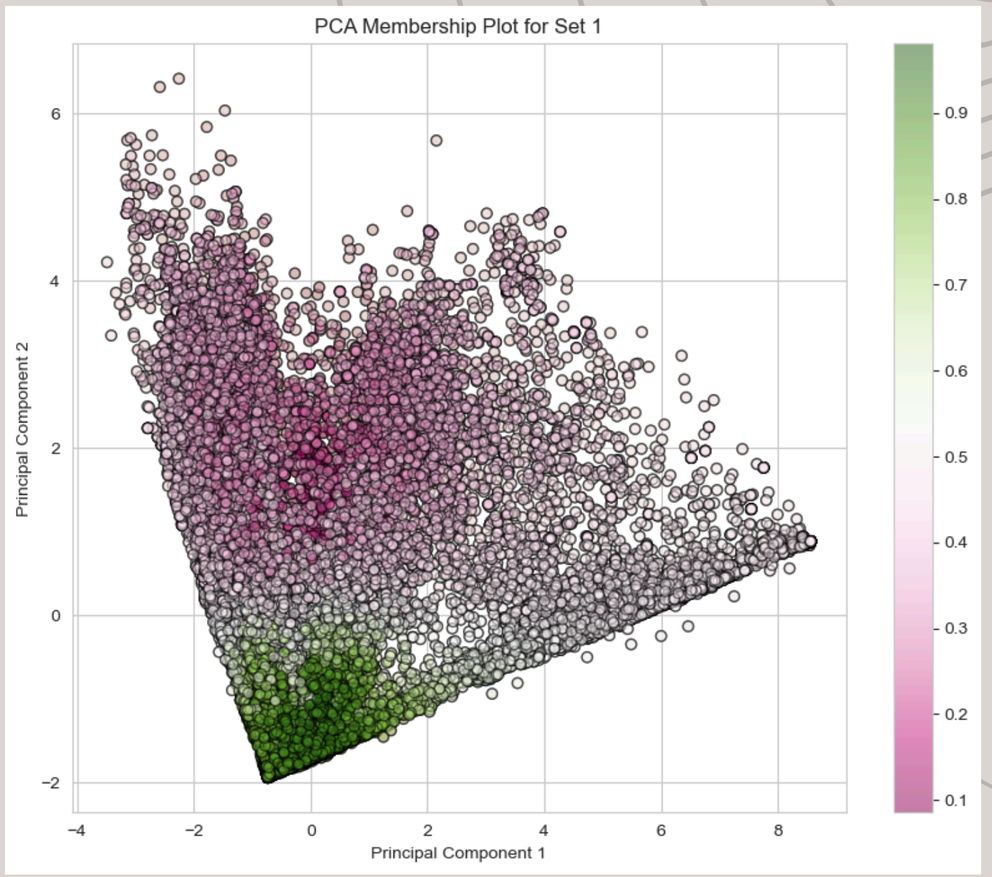
# Observations

- Optimal Number of Clusters: 2 determined for all datasets
- Fuzzy Partition Coefficients (FPC): Indicates some distinct separation but not highly defined. Values range from 0 to 1. Higher values suggest clearer, better-separated clusters
- FPC Comparison: 'Set 1' has highest FPC (0.6253), indicating clearest separation. 'Set 3' has lowest FPC (0.5685), indicating less clear separation
- Best Model Selection: Set 1 with cluster number 2 chosen based on FPC analysis
- Visualization: Model visualization to understand and validate clustering results

Open Food Facts

# PCA Membership Plot

- Utilizes Gradient: Green to pink gradient indicates low to high membership
- Clearly Defined Clusters: Green and pink clusters at top and bottom. Deep colors signify strong association with clusters
- Fuzzy Clustering Feature: Provides degree of membership rather than strict partitioning. Captures subtleties in data, avoiding overlooking nuances



Open Food Facts

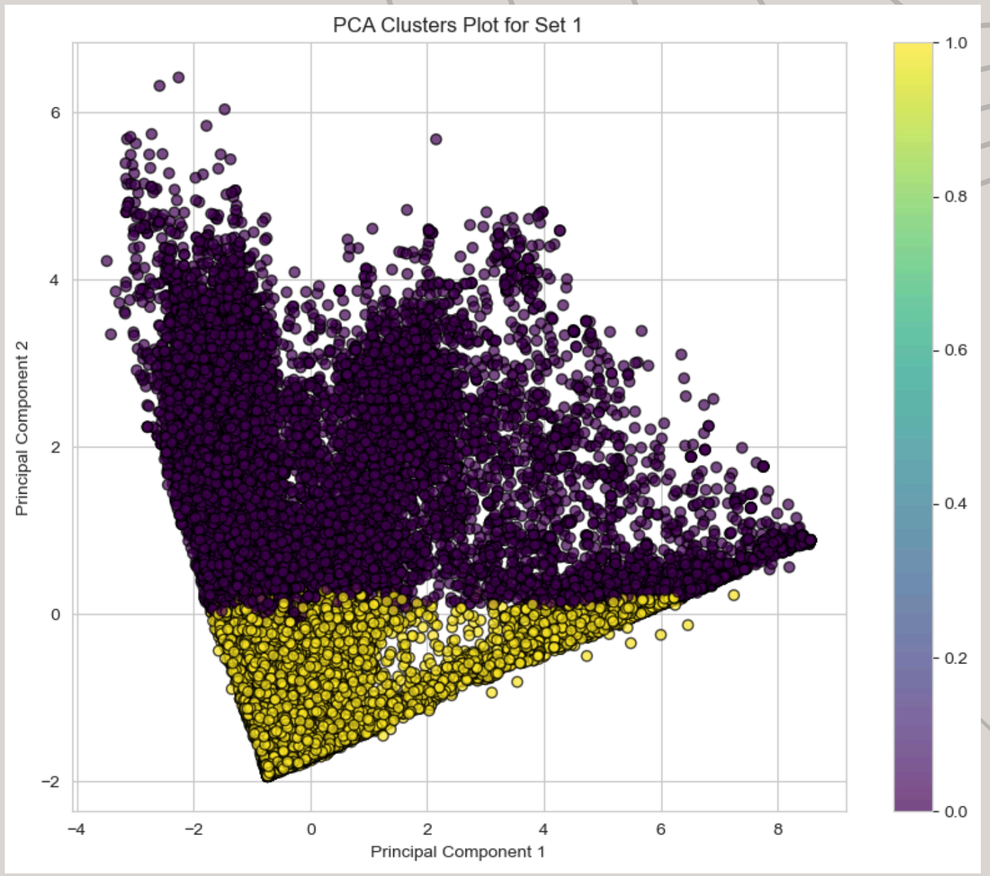
# PCA Cluster Plot

**Separation:**

- Purple cluster: Forms distinct group at top left
- Yellow cluster: Fans out from bottom left to top right
- Relatively sharp separation line, with some overlap at center where purple and yellow mix

**Spread:**

- Yellow cluster: Compact, suggesting tighter cohesion
- Purple cluster: More spread out, indicating wider variety of data points



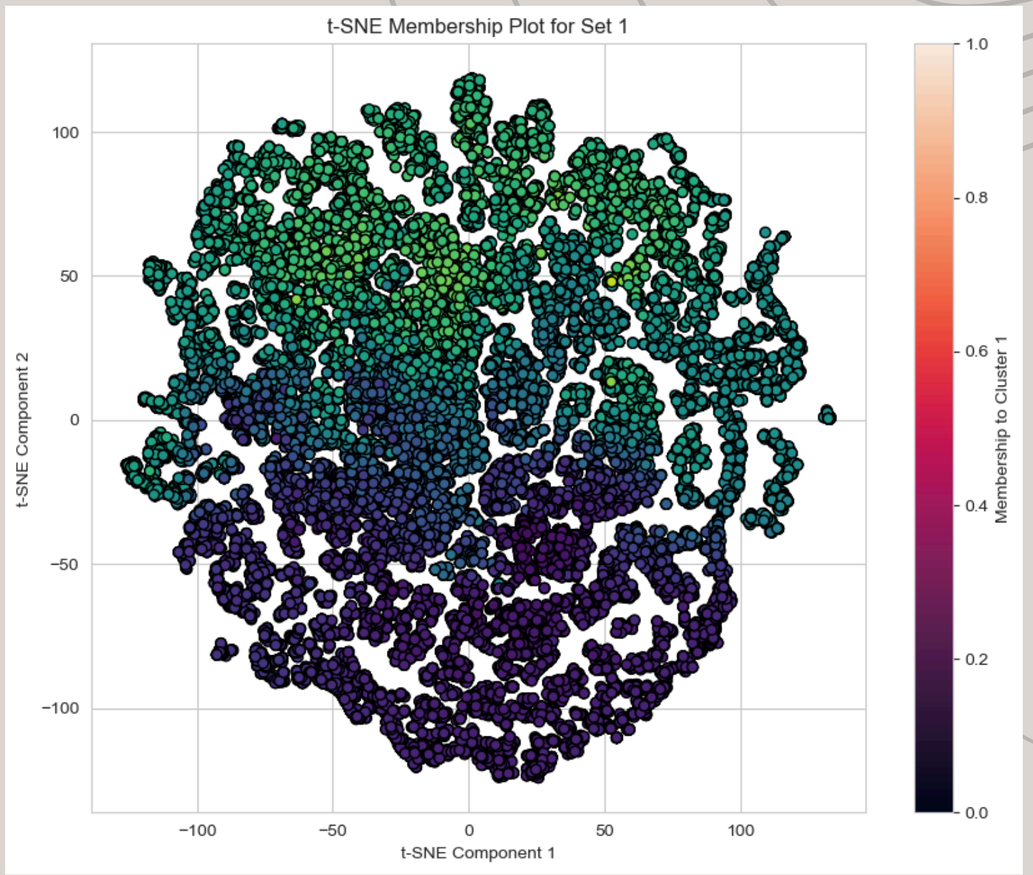
Open Food Facts

# t-SNE Membership Plot

**Strong Affiliation (Bright Green Points):**
- Highly characteristic of Cluster 1
- Share common features or behaviors representative of cluster's defining qualities

**Weak Affiliation (Purple Points):**
- Indicates weaker affiliation to Cluster 1
- Might belong more strongly to another cluster

**Moderate Membership (Shades of Teal):**
- Transitional areas between clusters
- Data points share attributes with multiple clusters



Open Food Facts

# t-SNE Cluster Plot

**Cluster Distribution:**

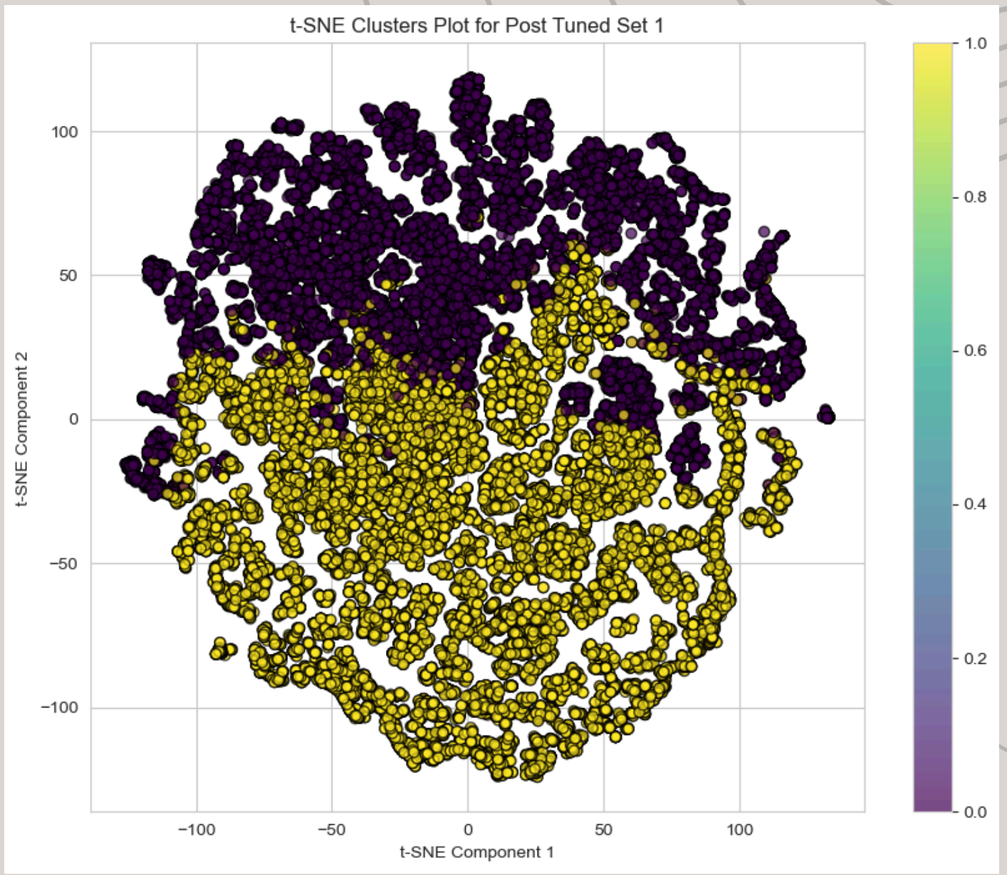
- Purple cluster: Highest membership in one cluster
- Yellow cluster: Points primarily belonging to the other cluster

**Separation:**

- Visible separation between clusters
- Some areas where clusters come close or slightly overlap
- Transitional data points share characteristics of both clusters

**Density and Outliers:**

- Both clusters have outliers
- Purple cluster: Sparse outliers far from main group



Open Food Facts

# Comparison of Centroid (Before & After Tuning)

| Nutrient | Before Tuning (Cluster 0) | After Tuning (Cluster 0) | Difference |
|---|---|---|---|
| Energy (kJ) | High (1104.65) | Higher (1171.07) | +66.42 |
| Fat (g) | Moderate (10.05) | Slightly higher (11.45) | +1.40 |
| Saturated Fat (g) | Moderate (2.63) | Slightly higher (2.99) | +0.36 |
| Trans Fat (g) | Low (0.000224) | Similar (0.000221) | -0.000003 |
| Cholesterol (g) | Low (0.000099) | Similar (0.000107) | +0.000008 |
| Carbohydrates (g) | Moderate (33.95) | Similar (34.19) | +0.24 |
| Sugars (g) | High (13.13) | Similar (13.46) | +0.33 |
| Fiber (g) | Moderate (1.04) | Slightly lower (1.02) | -0.02 |
| Proteins (g) | High (8.65) | Slightly higher (9.22) | +0.57 |
| Salt (g) | Moderate (1.01) | Slightly higher (1.11) | +0.10 |
| Sodium (g) | Low (0.396) | Similar (0.435) | +0.039 |
| Vitamin A (g) | Very low (0.000001) | Similar (0.000001) | 0 |
| Vitamin C (g) | Very low (0.000098) | Lower (0.000076) | -0.000022 |
| Calcium (g) | Very low (0.001914) | Lower (0.001621) | -0.000293 |
| Iron (g) | Very low (0.000046) | Similar (0.000047) | +0.000001 |

| Nutrient | Before Tuning (Cluster 1) | After Tuning (Cluster 1) | Difference |
|---|---|---|---|
| Energy (kJ) | Moderate (460.32) | Similar (464.86) | +4.54 |
| Fat (g) | Low (2.59) | Slightly lower (2.28) | -0.31 |
| Saturated Fat (g) | Low (0.76) | Lower (0.68) | -0.08 |
| Trans Fat (g) | Low (0.000110) | Similar (0.000120) | +0.000010 |
| Cholesterol (g) | Low (0.000052) | Similar (0.000049) | -0.000003 |
| Carbohydrates (g) | Moderate (17.91) | Similar (18.87) | +0.96 |
| Sugars (g) | Moderate (9.29) | Similar (9.21) | -0.08 |
| Fiber (g) | Low (0.59) | Similar (0.65) | +0.06 |
| Proteins (g) | Low (3.34) | Similar (3.35) | +0.01 |
| Salt (g) | Low (0.310) | Similar (0.299) | -0.011 |
| Sodium (g) | Low (0.122) | Similar (0.118) | -0.004 |
| Vitamin A (g) | Very low (0.000002) | Similar (0.000002) | 0 |
| Vitamin C (g) | Very low (0.000713) | Lower (0.000659) | -0.000054 |
| Calcium (g) | Very low (0.008867) | Lower (0.008241) | -0.000626 |
| Iron (g) | Very low (0.000020) | Similar (0.000020) | 0 |

Open Food Facts

# DBSCAN

## Hyperparameters:

- Epsilon (ε):
  - Defines radius for neighboring points in the same cluster
  - Core points within radius, outliers outside
- Minimum Samples (MinPts):
  - Specifies minimum points to form dense region or core point
  - Points with fewer than MinPts neighbors considered outliers
- Distance Metric:
  - Various options like Euclidean distance, Manhattan distance, or cosine similarity
  - Defines similarity measure between data points

## Evaluation Metrics

Silhouette Score:

- Measures compactness and separation of clusters
- Higher values signify better clustering structures

Davies-Bouldin Index:

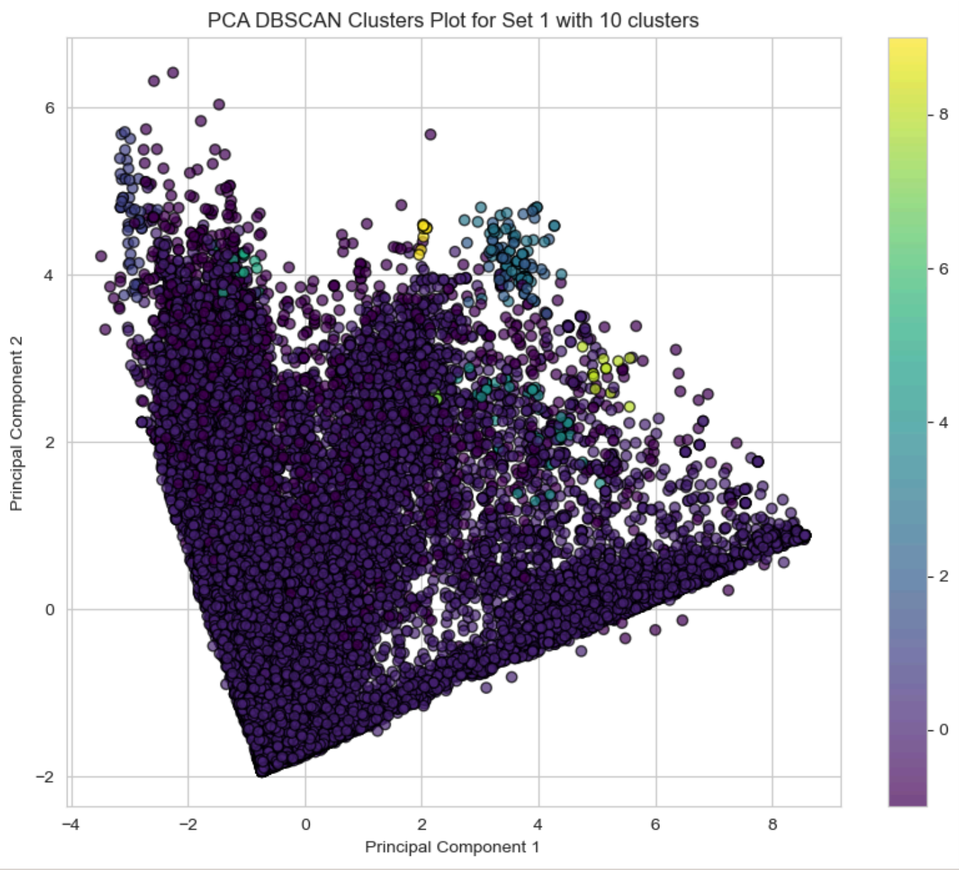
- Evaluates clustering quality based on within-cluster and between-cluster distances
- Lower values indicate better clustering

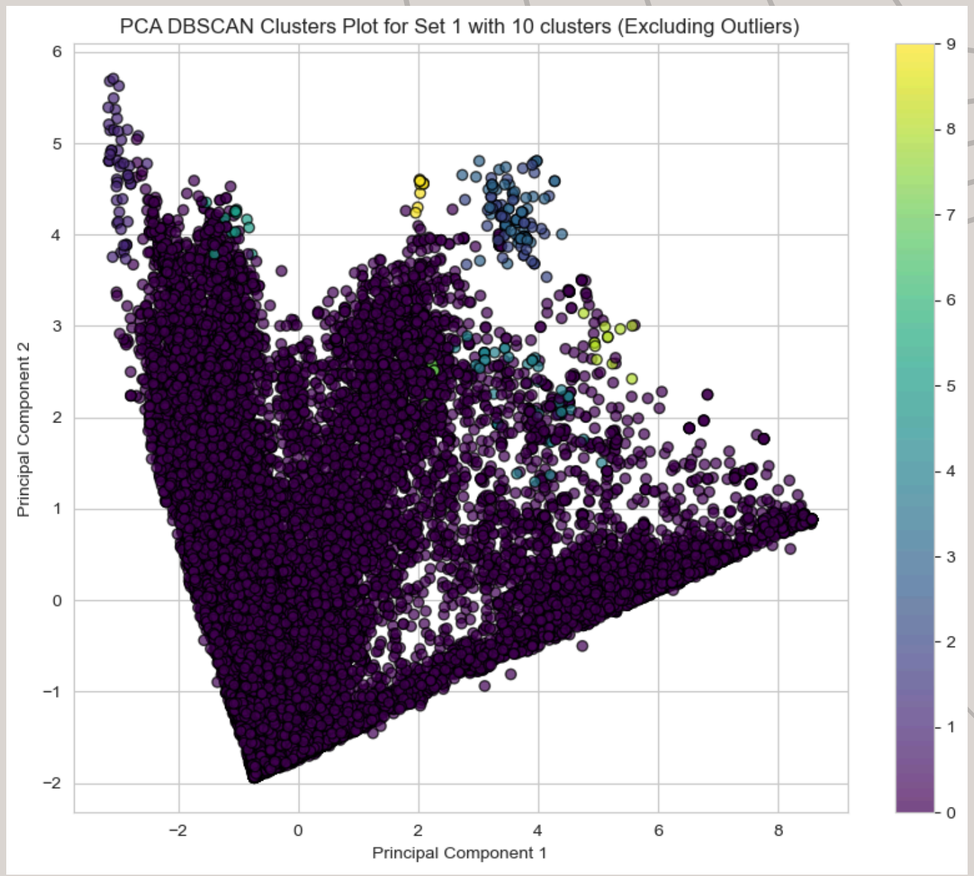
Calinski-Harabasz Index:

- Computes ratio of between-cluster dispersion to within-cluster dispersion

Open Food Facts

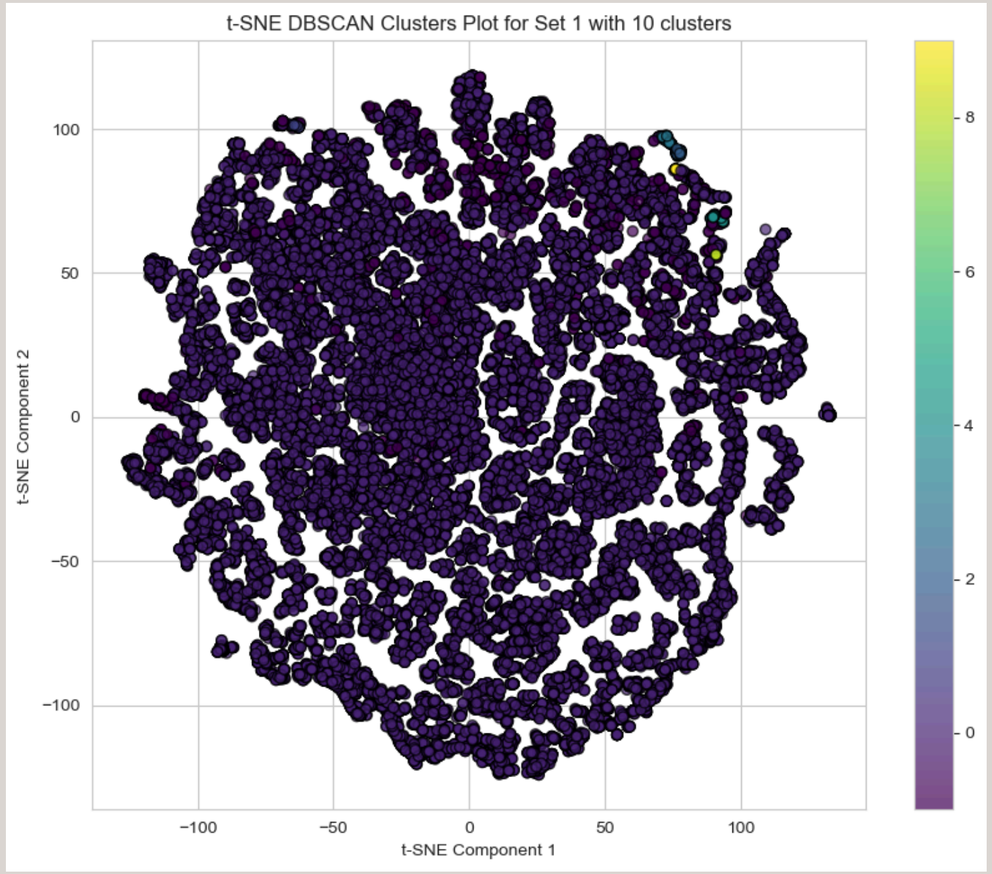
# PCA Cluster Plots

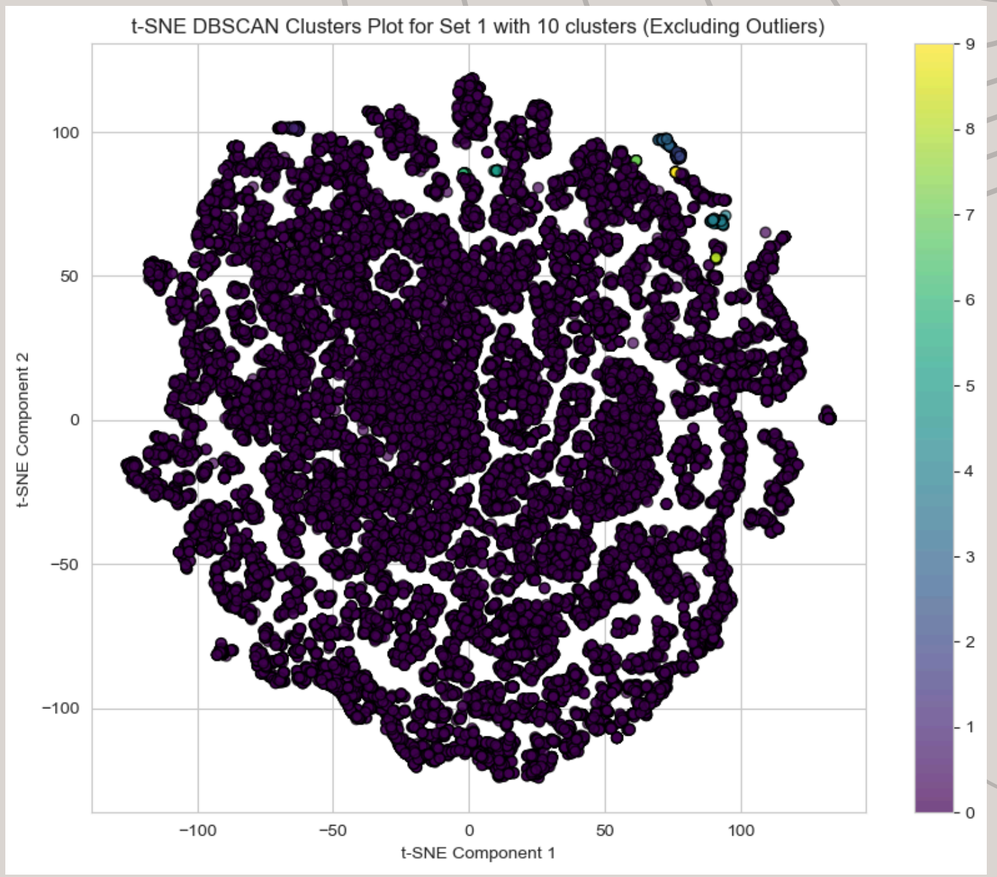


Open Food Facts

# t-SNE Cluster Plots



Open Food Facts

# Observation

**Cluster Distribution and Separation:**

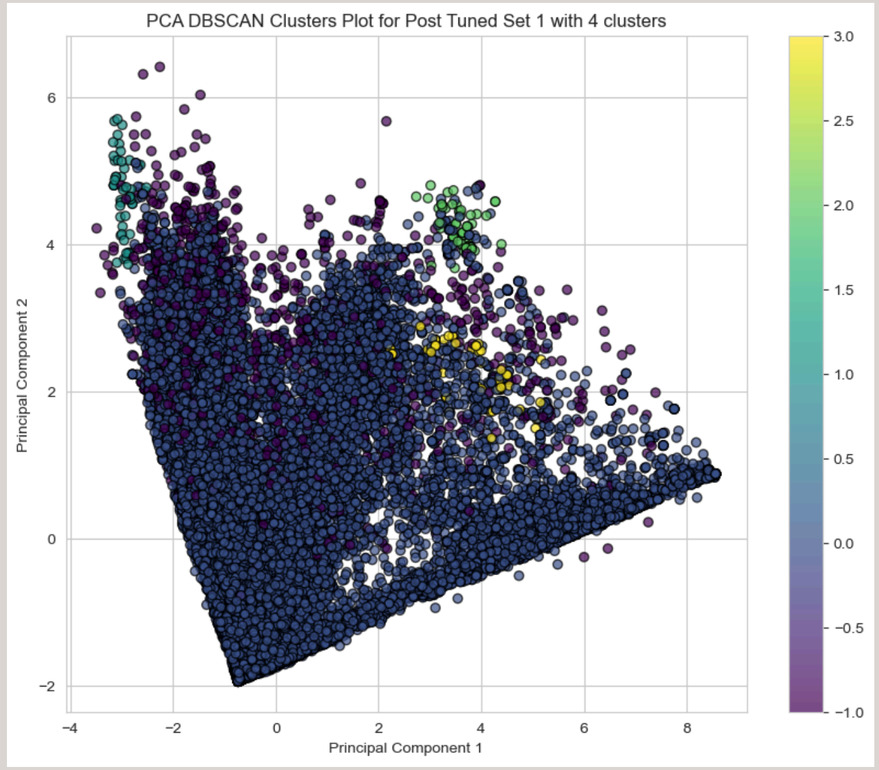
- Majority of data points tightly clustered in large central mass
  - Indicates high-density area with similar data points in first two principal components
- Several small clusters visible around periphery
  - Relatively well-separated from each other and main group
  - Some overlap observed

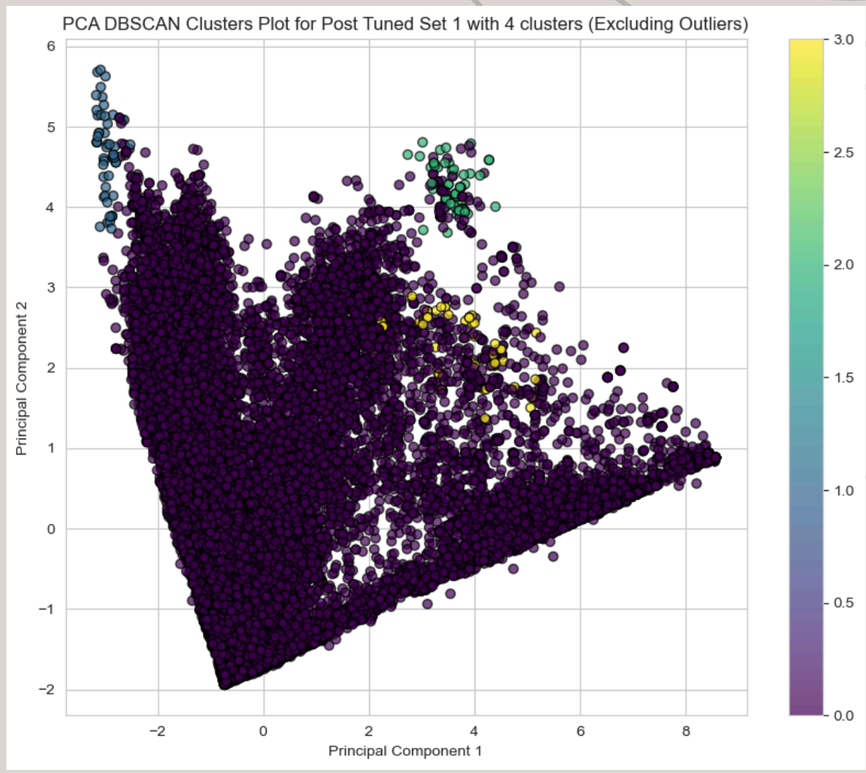
**Outliers and Noise:**

- Disappearance of various points in PCA plot (excluding outliers)
- These points may represent outliers or noise in dataset
- DBSCAN classifies some points as noise, not belonging to any cluster
- Likely points algorithm couldn't group into main or peripheral clusters

Open Food Facts

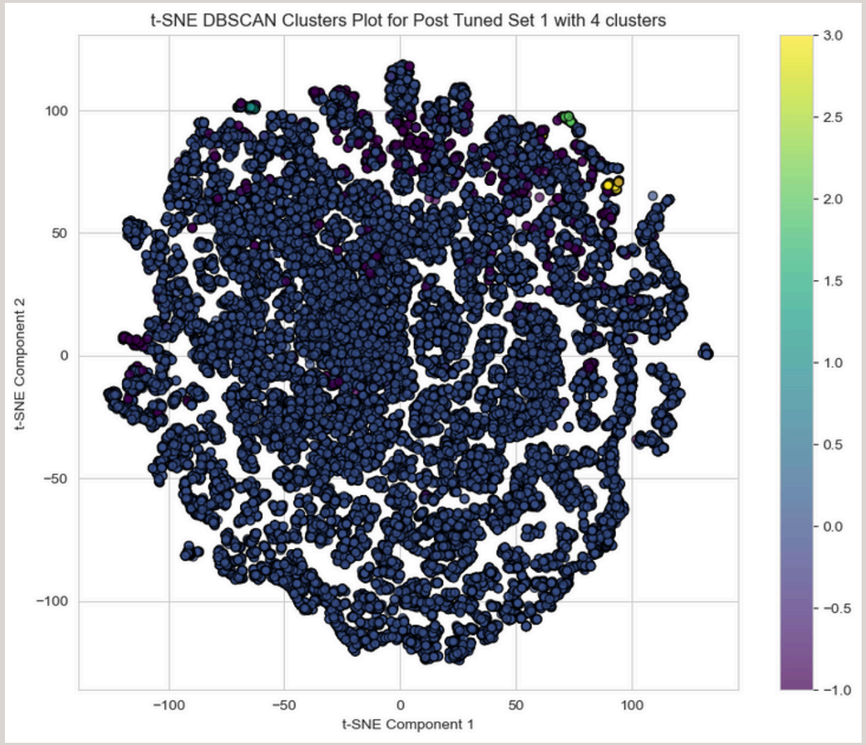
# PCA Cluster Plots (Post Tuned)

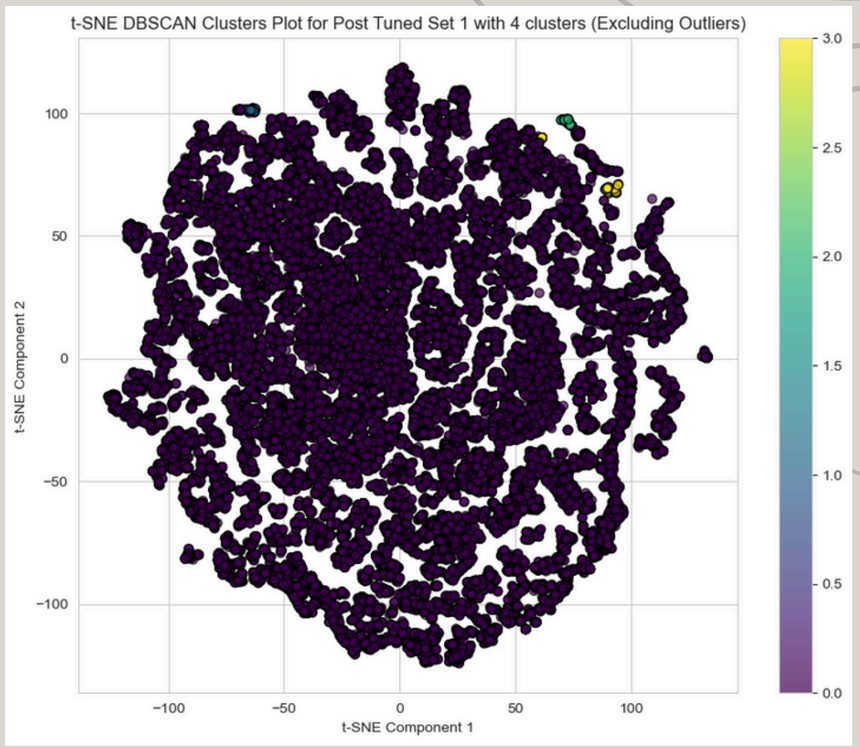


Open Food Facts

# t-SNE Cluster Plots (Post Tuned)



Open Food Facts

# Observation

**Cluster Distribution:**
- PCA and t-SNE plots show data assigned to four main clusters
- Clusters relatively distinct with some overlap
- Main cluster where most data points were assigned
- Clusters 1 and 2 (blue and green) in t-SNE plot very compact
- Cluster 3 (yellow) shows some spread

**Outlier Identification:**
- First chart includes outliers, second chart excludes them
- Significant presence of outliers in first plot
- Indicates many data points not fitting well into dense regions defined by eps and min_samples settings
  - - Suggests dataset has variance or noise, or distinct subgroups not conforming to larger cluster patterns

Open Food Facts

# Comparison of Centroid (Before & After Tuning)

| Cluster | Before Tuning | Cluster | After Tuning |
|---|---|---|---|
| 0 | Energy: Low<br>Fat: Low | 0 | Energy: Low<br>Fat: Low |
| 1 | Energy: Medium<br>Fat: Moderate<br>Protein: High | 1 | Energy: Slightly increased<br>Fat: Slightly increased<br>Protein: High |
| 2 and 3 | Energy: High<br>Fat: Very high<br>Carbohydrates: Very high | 2 | Similar to pre-tuning Cluster 2 and 3 but with refined separation |
| 4 to 9 | Varying from moderate to very high in calorie content with different levels of sugars and fats | 3 | Reflects similar attributes to pre-tuning Cluster 4 but with better differentiation in nutritional content, particularly protein and carbohydrates |

Open Food Facts

# Conclusion

**Best Clustering Model**: K-Means Algorithm on Feature Set 1

**Reasoning:**
- Optimal number of clusters (k=4)
- Lower inertia scores compared to other feature sets
- Superior performance consistently observed

**Evaluation Metrics:**
- Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index affirm clustering robustness

**Conclusion:**
- K-Means on Feature Set 1 balances distinct cluster separation and compactness
- Comprehensive evaluation and analysis support its effectiveness as the best clustering model

Open Food Facts

# Future Recommendation

**Ensemble Clustering:**
- Combine strengths of multiple algorithms to enhance performance

**Feature Engineering:**
- Conduct further feature engineering to extract more relevant features

**Advanced Dimensionality Reduction:**
- Explore techniques beyond PCA and t-SNE (e.g., UMAP, autoencoders)

**Semi-Supervised and Active Learning:**
- Incorporate domain knowledge or user feedback iteratively

**Dynamic Data Handling:**
- Implement online or incremental clustering algorithms

Group 6

# Thank You

Presented by Tay Wei Rong & Lim Jo Sun