American Journal of Engineering Research (AJER)
e-ISSN: 2320-0847 p-ISSN: 2320-0936
Volume-15, Issue-11, pp-01-07
www.ajer.org
Research Paper Open Access

Macronutrient patterns and data quality in 2395 food records an exploratory clustering study

Aishwarya Varadarajula[1]
[1]BASIS Independent, Fremont, California, United States

Abstract
Background: The Food Nutrition Dataset contains enough variables for a broad nutritional analysis, including vitamin and mineral fields. Whether those values can be compared reliably is less clear. This study examined the quality of its records and used relative macronutrient quantities to explore food groupings.
Methods: Five files were combined to give 2,395 food entries with 34 numerical attributes. Checks covered missing and duplicate records, nutrient-component consistency and agreement with the documented quantity basis. Fat, carbohydrate and protein values were converted to shares of their combined total for entries with a positive total. K-means solutions with two to eight groups were compared using silhouette scores and Davies–Bouldin indices. Ten refits on 80% subsets and a separate filtered analysis assessed the stability of the selected partition.
Results: Although every numerical field was complete, 80 entries failed at least one component-consistency check. Water exceeded 100 in 821 records, conflicting with the stated grams-per-100-g basis. Seventeen entries had a zero-macronutrient total and could not be clustered. Among the remaining 2,378 entries, the three-group solution separated carbohydrate-dominant (n = 1,512), fat-dominant (n = 359) and protein-dominant (n = 507) profiles. Its silhouette score of 0.608 was only slightly higher than the two-group score of 0.600; the Davies–Bouldin criterion favored two groups. Median adjusted Rand agreement across subset refits was 0.994. After filtering, 2,216 entries remained and agreement with the original partition was 0.957.
Conclusions: Relative macronutrient values produced stable groups, but the underlying records contain unresolved measurement problems. The next step is to verify the existing nutrient fields and their quantity basis before expanding the analysis to micronutrients or dietary adequacy.

Keywords: *food composition; macronutrients; micronutrients; K-means; compositional data; data quality; exploratory analysis*

Date of Submission: 27-10-2026 Date of acceptance: 06-11-2026

I. Introduction

Reliable food-composition data are needed to connect questions about food access with questions about nutritional value. This connection matters in settings where food insecurity and illness occur together. Fitzpatrick and Willis examined hunger among people experiencing homelessness [1], while Vozoris and Tarasuk found an association between household food insufficiency and poorer health [6]. Douglas and colleagues showed how food insecurity affected the day-to-day management of long-term health conditions [5]. These studies concern people and their circumstances; a food database can contribute information about composition, but cannot on its own account for those circumstances.

Macronutrients provide only part of that information. Work on unaccompanied homeless youth and adults experiencing homelessness has drawn attention to diet quality and nutritional status [7,8], and the review by Bhadra and Deb addresses nutritional anemia [3]. A food may supply energy without supplying an adequate range of micronutrients. Assessing that distinction requires measurements whose definitions and units can be traced to their source.

Composition also needs to be distinguished from processing. Moodie and colleagues examined the market and political practices surrounding ultra-processed foods [2]; Sinha and Haque considered the health consequences associated with excess processed-food consumption [4]. Similar proportions of fat, carbohydrate

and protein do not establish similar ingredients or preparation methods. This study therefore treats macronutrient composition as one descriptive feature of food, without assigning processing categories.

Clustering compares foods through the measurements chosen to describe them [10]. Before interpreting the groups, the quantity basis needs attention. A large serving and a small serving of the same food will contain different nutrient amounts even when their composition is identical. FoodData Central records this basis explicitly, while FAO/INFOODS guidance addresses the identification of comparable foods [11,12]. Here, the three macronutrients were expressed as shares of their subtotal. This removes a multiplier common to all three amounts, but also fixes their sum at one, a constraint that matters when analyzing compositional data [13].

The study began with an assessment of the five files: their completeness, their component totals and the information available to interpret each nutrient field. Clustering then examined the relative amounts of fat, carbohydrate and protein. Refits on subsets and selected exclusions tested how much the grouping changed. Together, these steps were used to assess the dataset's suitability for macronutrient grouping and to identify the work needed before including its vitamin and mineral columns.

II. Materials And Methods

Data source and unit of analysis

Food Nutrition Dataset was contributed to Kaggle by Utsav Dey [16]. The analysis used version 1, released on 29 June 2024 and listed under the CC0 Public Domain license. On 18 September 2026, the five supplied CSVs were confirmed to match the corresponding version-1 files byte for byte. FOOD-DATA-GROUP1.csv through FOOD-DATA-GROUP5.csv contained 551, 319, 571, 232 and 722 rows, respectively. All had the same 37 columns and were combined by appending rows. The GROUP labels identified source files, rather than verified food categories. Descriptive summaries and figures were available in the archive, but the accompanying metadata did not identify the primary measurement sources or sampling frame.

Each row described a food. After removing the two "Unnamed:" index columns, the dataset comprised food names and 34 numeric fields. These included energy, fat and its subtypes, carbohydrate-related measures, protein, water, cholesterol, vitamins, minerals and Nutrition Density. The analysis retained each row's original filename and position for traceability. The source description states that nutrient quantities generally refer to 100 g, with macronutrients and water reported in grams. That statement provided a basis for consistency checks, rather than proof that every value had been recorded correctly. Separate food names could also refer to closely related products or preparation variants.

Data quality assessment

Missing entries, negative amounts and repeated food names were checked first. Duplicate-record checks omitted the exported indices and the fields added for traceability. Recorded zeros were retained because the source did not explain their meaning. An unusually high value was likewise retained unless a stated exclusion applied; its magnitude alone gave no basis for choosing a replacement. Nutrition Density was not used as a feature because its description did not include a formula. Once the per-100-g description had been reviewed, a further check counted water and individual macronutrient values above 100. These counts are reported as documentation inconsistencies and were not used to select records for the sensitivity refit.

The component audit used four comparisons. Saturated fat was compared with total fat, and sugars and dietary fiber were each compared with carbohydrates. A fourth comparison summed saturated, monounsaturated and polyunsaturated fats and compared that sum with total fat. An excess greater than 0.1 recorded units triggered the first three flags; the tolerance for the summed fats was 0.3. These small allowances were chosen for possible rounding differences. They were not validated cutoffs, and each comparison assumed compatible units and denominators. Definitions also matter, particularly for fiber, so the flags identify entries requiring review rather than establish the cause of an error. A complete food-composition audit would require the broader checks described by FAO/INFOODS [15].

Descriptive analysis

Numeric fields were summarized using counts, ranges, quartiles, means, standard deviations and the frequency of zeros. A log10(value + 1) plot was used to examine the skew in Caloric Value. Spearman correlations described its relationships with fat, carbohydrates, protein, sugars, dietary fiber and water. These calculations used the recorded amounts without a serving-size adjustment. Because the sampling population and inclusion process were unknown, the analysis did not attach population-level confidence intervals or p-values to these summaries.

Macronutrient representation and clustering

For each food i, the macronutrient subtotal was $T_i = F_i + C_i + P_i$, where F, C and P are the recorded fat, carbohydrate and protein amounts. Entries with $T_i > 0$ were represented by $x_i = (F_i/T_i, C_i/T_i, P_i/T_i)$. Each

component thus describes its share of the three-macronutrient subtotal. This is a mass-based representation, conditional on compatible units and a common quantity basis within the row. It does not express the percentage of the whole food or the percentage of energy supplied by each nutrient.

Dividing by the subtotal removes a multiplier shared by all three amounts. It cannot resolve a wrong value or a unit mismatch affecting only one component. The 17 entries with a zero subtotal were excluded from clustering. The shares were already on a common scale, so no further standardization was applied. Food names, file labels, energy, micronutrients and Nutrition Density were not used as model features.

K-means was fitted with Euclidean distance for k = 2 to 8, using 30 initializations and random seed 42 for each model. The solution with the highest full-data mean silhouette score was selected [14]. Davies–Bouldin indices and inertia were examined alongside that score to show where the criteria agreed or differed. After fitting, cluster identifiers were ordered by decreasing mean carbohydrate share. Each group was named for its largest mean macronutrient share, and the five entries nearest its centroid were retained as examples.

The three shares occupy a two-dimensional simplex because their sum is fixed [13]. Euclidean K-means provided a baseline for examining that space. A log-ratio analysis was not undertaken: it would first require a defensible treatment of zero values. PCA was used to draw the two-dimensional projection, while clustering operated on the shares themselves. The workflow was implemented in Python [9], using established clustering concepts discussed by Malik and Tuckfield [10].

Stability sensitivity and reproducibility

For the repeatability check, a NumPy random generator with seed 42 selected ten subsets. Each draw included 80% of the modeled records without replacement. K-means used the selected k, 20 initializations and seeds 0–9 for the ten refits. Every refitted model assigned labels to the full modeled dataset, and the adjusted Rand index (ARI) compared those labels with the original partition. This comparison included records used to fit the subset models. It therefore measured stability within the available data, rather than agreement on an independent validation sample.

For the sensitivity analysis, records were retained only if $T_i \geq 1$ and none of the four component checks had flagged them. The model was refitted at the selected k with 30 initializations and seed 42, and its labels were compared with the original labels for those same records. The threshold of 1 recorded unit tested the influence of very small subtotals, for which rounding may affect ratios; it was not a minimum portion size. The software environment comprised Python 3.12.13, pandas 2.2.3, NumPy 2.3.5, scikit-learn 1.8.0 and matplotlib 3.10.8. The analysis package retained the executable script and source-row identifiers.

III. Results

Dataset completeness and quality flags

The initial checks found no blank cells, negative amounts, duplicate names or exact duplicate food records among the 2,395 entries. The component checks gave a different picture: 80 records, or 3.34%, had at least one flag (Table 1). One readily identifiable example was “Coconut whole coconut,” with total fat recorded as 71.5 and saturated fat as 672.0. The source assigns both fields the same grams-per-100-g basis, under which these values are inconsistent. No record-level explanation was available. Values above 100 also occurred in the water field for 821 foods and in at least one of the fat, carbohydrate or protein fields for 123 foods. These counts can overlap. They show that the stated mass basis does not fit all rows.

Audit item	Entries
Any component-consistency flag	80
Saturated fat > total fat + 0.1	25
Sugars > carbohydrates + 0.1	8
Fiber > carbohydrates + 0.1	10
Summed fat subtypes > total fat + 0.3	64
Zero three-macronutrient total	17
Positive three-macronutrient total below 1	98

Table 1. Audit counts for the 2,395 records. One record may trigger several flags. Checks use the recorded numbers; the source units remain subject to verification.

The file schema contained twelve vitamin fields and ten mineral fields, including sodium. The problem was how to interpret them consistently. For sodium, the heading gave grams but the descriptive text gave milligrams. Vitamin A had a similar conflict between milligrams and micrograms. The field called Vitamin B11 was described as folic acid, without a mapping to other folate measures. Consequently, the presence of numeric micronutrient values did not provide a sufficient basis for ranking foods by their concentrations.

Recorded nutrient distributions

Attribute	Median	25th–75th percentile	Range
Caloric Value	117	44.5–258	0–6,077
Fat	2.1	0.3–9.4	0–550.7
Carbohydrates	6.8	0.5–25.05	0–390.2
Protein	3.5	0.8–13.3	0–560.3
Water	39.6	6.5–129.55	0–1,875.9

Table 2. Distribution of selected fields as recorded. Values have not been corrected or converted; inconsistencies with the source's stated per-100-g basis remain unresolved.

Spearman correlations with Caloric Value were 0.744 for fat, 0.712 for protein and 0.349 for carbohydrates. Sugars showed almost no association ($\rho = -0.000045$). These summaries refer to the recorded amounts across the selected foods. Without a reliable common quantity basis, portion differences and recording problems may affect those relationships. In particular, the near-zero sugar correlation cannot be interpreted as evidence about sugar's contribution to dietary energy.

Selection and composition of clusters

Table 3 shows why the choice of k needs qualification. Three groups gave the maximum silhouette value, 0.608447, but two groups gave 0.600219, a difference of approximately 0.0082. The Davies–Bouldin index favored two groups instead, with values of 0.678254 at k = 2 and 0.695806 at k = 3. The reported partition follows the silhouette selection rule. The competing result is retained because the data do not clearly establish three as the only useful number of groups.

k	Silhouette ↑	Davies–Bouldin ↓	Inertia ↓
2	0.600	0.678	227.51
3	0.608	0.696	132.54
4	0.505	0.737	89.06
5	0.506	0.699	63.76
6	0.453	0.744	51.36
7	0.476	0.719	41.63
8	0.480	0.710	34.76

Table 3. Candidate models fitted to the same 2,378 entries. Arrows indicate the preferred direction, although inertia ordinarily decreases as k increases.

Cluster	n (%)	Fat %	Carb. %	Protein %
Carbohydrate-dominant	1,512 (63.58)	9.53	76.97	13.50
Fat-dominant	359 (15.10)	65.76	12.80	21.43
Protein-dominant	507 (21.32)	21.26	6.58	72.16

Table 4. Cluster counts use the 2,378 modeled records as denominator. Shares are means of row-wise normalized values, not ratios of pooled nutrient totals or energy shares.

The three centroids describe familiar contrasts in macronutrient balance, while the nearby food names show the breadth of each group. Egg dinner roll, rice pudding and breadsticks lay near the carbohydrate-dominant centroid. Bacon meatless, chicken pork bologna and black walnut dried were examples near the fat-dominant centroid. The protein-dominant examples included lamb liver cooked, lamb heart cooked and chicken fajita strips. These names are given as recorded in the dataset. Their proximity reflects the three numerical shares; it does not establish similar ingredients, processing or use in a diet.

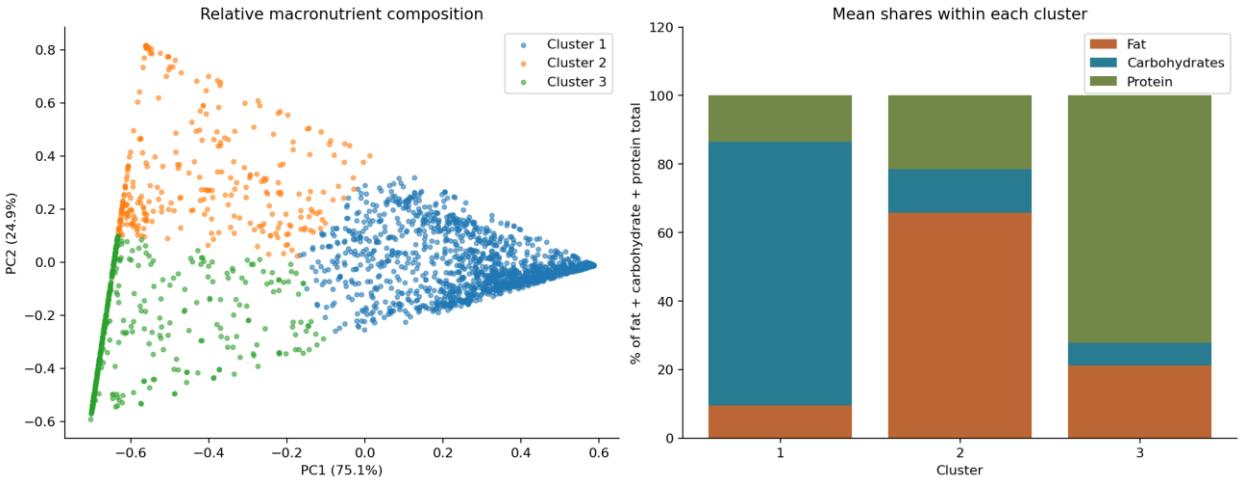


Figure 1. PCA projection of the macronutrient shares (left) and mean shares for the three clusters (right). PC1 explains 75.1% of the variance and PC2 explains 24.9%. Since the three shares sum to one, their variation lies in a two-dimensional plane.

Stability and quality filtered sensitivity

Subset refits closely reproduced the original partition. Median ARI was 0.993727, with a range of 0.984262–0.997241 across the ten refits. The filtered analysis retained 2,216 entries and gave an ARI of 0.957200 on those records. Thus, much of the grouping persisted after the specified exclusions. Errors common to the source files could still persist in both analyses, and these results do not provide an external validation estimate.

IV. Discussion

What the clusters establish

Carbohydrate-dominant foods accounted for 63.58% of modeled entries. This is a description of the database, whose food coverage and selection process are not known. It gives no estimate of how often those foods are eaten or their contribution to nutrient intake. The centroid examples also help define the meaning of a group: breadsticks and rice pudding, for instance, can be close in the three-share space while differing as food products. The clusters summarize the chosen composition variables, not every characteristic of the foods.

The repeated refits support a stable division of this particular share space. Whether that division is useful outside the dataset is a separate question. The three features emphasize relative dominance by construction, and K-means can divide a continuous spread of observations even without distinct natural classes. The close silhouette scores for two and three groups make this especially relevant here. Independent food characteristics or a verified external dataset would be needed to assess the nutritional usefulness of the labels.

Why micronutrient expansion requires validation

Further micronutrient work should start with the columns already available. The sodium and vitamin A unit conflicts must be resolved before their values enter a distance calculation. The folic acid field also needs a clear definition: naturally occurring folate, folic acid and dietary folate equivalents are not interchangeable labels. Standardizing the numeric columns would give them comparable statistical scales, but would leave these measurement questions unanswered. Once definitions and units have been checked, the need for additional nutrient data can be assessed.

Portion information creates another unresolved step. FoodData Central provides an example of how nutrient denominators and portion weights can be documented [11]. For an amount q measured in an edible portion of m grams, the conversion is 100q/m per 100 g. The conversion cannot be applied here simply by assuming a portion weight. Although the source states a per-100-g basis, the water values above 100 contradict that interpretation for some records. Recovering the original weights would be necessary if those records were

actually portion-based. Summing the nutrient columns would not recover food mass reliably, because units differ and some nutrient categories overlap. External matching would also need to record preparation state, edible portion, reference version and the confidence of each match [12].

Relevance to food insecurity and nutritional research

The food-insecurity studies cited in the introduction concern access, diet and health [1–8]. Applying the present analysis to those questions would require a link between food records and actual use. An adequacy assessment would need amounts consumed and the relevant nutrient requirements. A cost comparison would need prices collected in a specified location and period. Storage and cooking facilities could further restrict which foods are practical. None of these conditions is measured by the macronutrient groups, so they would have to be added explicitly in an applied study.

A manageable next question is whether verified micronutrients distinguish foods within each macronutrient group. That question could be tested against independent nutrient-quality criteria and checked in a separate dataset. Meal quantities and local prices could then support an analysis of feasible food choices. Studies extending to individual intake or health outcomes would need a participant protocol and the relevant institutional review.

V. Limitations and research contribution

The main uncertainty lies in the measurements and their source. Sampling coverage, primary measurement records and the meaning of zeros could not be established. The sensitivity analysis addresses only the four component flags and small macronutrient totals; records inconsistent with the per-100-g description may remain. Its agreement with the original model should be interpreted within that scope. In addition, k was fixed during refitting, so the analysis did not assess repeatability of cluster-number selection. Alternative feature representations, zero treatments, algorithms and external datasets were not compared.

The contribution is a reproducible baseline paired with specific, traceable data-quality findings. It shows which comparisons can be attempted using the recorded macronutrient shares and where further nutritional interpretation becomes uncertain. Extending the work requires verified quantities and nutrient definitions before more features are added. With those foundations, the next analysis could test the additional information provided by micronutrients rather than assume that a larger feature set produces a better classification.

VI. Conclusion

The dataset's complete numerical fields masked substantial inconsistencies in the nutrient records. Relative fat, carbohydrate and protein values produced repeatable groups among 2,378 foods, although the internal criteria did not strongly distinguish a three-group solution from a two-group one. These groups offer a limited descriptive summary. Reliable micronutrient comparisons will depend on correcting the quantity and unit problems identified in the audit and checking the existing values against their original sources.

Acknowledgments

We thank Utsav Dey for making the Food Nutrition Dataset publicly available through Kaggle [16].

Declarations

Data and code availability: The source dataset is available from Utsav Dey on Kaggle under the listed CC0 Public Domain license [16]. The analyzed CSV files match version 1. Analysis code and derived outputs are retained in the study working files; a public code repository has not yet been deposited.

Ethics: This secondary analysis used food descriptions and nutrient records without participant-level observations.

Funding, competing interests and author contributions: No Competing Interest

AI assistance: ChatGPT (OpenAI) assisted with computational analysis and manuscript drafting and editing.

References

[1] Fitzpatrick, K. M., & Willis, D. E. (2021). Homeless and hungry: food insecurity in the land of plenty. Food Security, 13(1), 3–12. https://doi.org/10.1007/s12571-020-01115-x

[2] Moodie, R., Bennett, E., Kwong, E. J. L., Santos, T. M., Pratiwi, L., Williams, J., & Baker, P. (2021). Ultra-processed profits: the political economy of countering the global spread of ultra-processed foods—a synthesis review on the market and political practices of transnational food corporations and strategic public health responses. International Journal of Health Policy and Management, 10(12), 968–982. https://doi.org/10.34172/ijhpm.2021.45

[3] Bhadra, P., & Deb, A. (2020). A review on nutritional anemia. Indian Journal of Natural Sciences, 10(59), 18466–18474. [Provisional: supplied pagination retained; conflicting pagination found; verify against the journal PDF before submission.]

[4] Sinha, S., & Haque, M. (2022). Obesity, diabetes mellitus, and vascular impediment as consequences of excess processed food consumption. Cureus, 14(9), e28762. https://doi.org/10.7759/cureus.28762

[5] Douglas, F., MacIver, E., & Yuill, C. (2020). A qualitative investigation of lived experiences of long-term health condition management with people who are food insecure. BMC Public Health, 20, 1309. https://doi.org/10.1186/s12889-020-09299-9
[6] Vozoris, N. T., & Tarasuk, V. S. (2003). Household food insufficiency is associated with poorer health. The Journal of Nutrition, 133(1), 120–126. https://doi.org/10.1093/jn/133.1.120
[7] Hatsu, I., Gunther, C., Hade, E., Vandergriff, S., Slesnick, N., Williams, R., Bruno, R. S., & Kennel, J. (2019). Unaccompanied homeless youth have extremely poor diet quality and nutritional status. International Journal of Adolescence and Youth, 24(3), 319–332. https://doi.org/10.1080/02673843.2018.1538885
[8] Huang, C., Foster, H., Paudyal, V., Ward, M., & Lowrie, R. (2022). A systematic review of the nutritional status of adults experiencing homelessness. Public Health, 208, 59–67. https://doi.org/10.1016/j.puhe.2022.04.013
[9] McKinney, W. (2012). Python for data analysis: Data wrangling with pandas, NumPy, and IPython. O'Reilly Media. https://www.oreilly.com/library/view/python-for-data/9781449323592/
[10] Malik, A., & Tuckfield, B. (2019). Applied unsupervised learning with R: Uncover hidden relationships and patterns with k-means clustering, hierarchical clustering, and PCA. Packt Publishing. https://www.oreilly.com/library/view/applied-unsupervised-learning/9781789956399/C12628_FM_EPUB_Final_NT.xhtml
[11] USDA Agricultural Research Service. (n.d.). FoodData Central: Foundation Foods documentation. Accessed 8 September 2026. https://fdc.nal.usda.gov/Foundation_Foods_Documentation/
[12] FAO/INFOODS. (2012). Guidelines for food matching, version 1.2. Food and Agriculture Organization of the United Nations. https://www.fao.org/4/ap805e/ap805e.pdf
[13] Aitchison, J. (1982). The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B (Methodological), 44(2), 139–160. https://doi.org/10.1111/j.2517-6161.1982.tb01195.x
[14] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7
[15] FAO/INFOODS. (2012). Guidelines for checking food composition data prior to the publication of a user table/database, version 1.0. https://www.fao.org/infoods/infoods/standards-guidelines/en/
[16] Dey, U. (2024). Food Nutrition Dataset (Version 1) [Data set]. Kaggle. https://www.kaggle.com/datasets/utsavdey1410/food-nutrition-dataset (accessed 18 September 2026).