Exploring Nutritional Patterns in Common Foods Using the CORGIS Dataset

Analysis of Macronutrients and Clustering of Food Groups

Eldana Hailemichael

Department of Computing and Data Science

Wentworth Institute of Technology

Boston, MA, USA

hailemichaele@wit.edu

ABSTRACT

This report analyzes nutritional information for a wide range of foods using the CORGIS Food dataset. The aim is to reveal patterns in protein, fat, and carbohydrate content, identify nutritional outliers, and apply K-means clustering to group similar foods. The study demonstrates the power of exploratory data analysis, visualization, and clustering for understanding food nutrient profiles at a population level.

KEYWORDS

Nutrition, Food Data, Macronutrients, Data Visualization, Clustering

1. INTRODUCTION

Eating patterns and food choices are central to human health and disease prevention. This project investigates the nutritional composition, specifically, protein, fat, and carbohydrate content, of common foods using the CORGIS open nutrition dataset. My motivation was to explore which foods stand out for their nutrient content, how different food categories compare, and whether foods can be grouped based on similarities in their nutrient profiles.

Public health research underscores the influence of dietary macronutrient balance on long-term health. Leveraging public datasets and data science tools helps identify both expected and surprising trends in food nutrition, with practical implications for dietary guidelines and consumer choices.

2. Data

2.1 Source of Data Set

The dataset used in this project is the CORGIS Food dataset, downloaded from the official CORGIS repository (https://corgis-edu.github.io/corgis/csv/food/). CORGIS compiles its food data from trusted sources such as the USDA and other food composition tables. It is widely cited and frequently updated, ensuring credible analysis.

2.2 Characteristics of the Data Set

The dataset is in CSV format and contains over 5,000 foods with 38 variables each. Key columns used in this project are:

Column Name	Description	Unit
Description	Name of food item	-
Category	Food group/category	-
Data.Protein	Protein content	grams/100g
Data.Fat.Total Lipid	Total fat content	grams/100g

| Data.Carbohydrate | Carbohydrate content | grams/100g

Data munging included focusing on the ten most frequent food categories, filling missing values with zeros for calculations, and standardizing units (grams per 100g serving). No merging of datasets was performed. New categories for analysis were not created, but summary statistics and groupings were calculated and visualized.

3. **Methodology**

Exploratory Data Analysis (EDA) was carried out by ranking and visualizing foods by their individual nutrient content, and by computing average protein, fat, and carbohydrate values for the top categories. Relationships between nutrients were explored with scatterplots.

For advanced analysis, I applied K-means clustering using the `KMeans` module from Scikit-learn. I standardized the protein, fat, and carbohydrate features, set k=3 for interpretability, and compared cluster assignments visually. K-means assumes clusters are roughly spherical and may be biased by outliers, but it is easy to interpret and effective for revealing broad patterns in nutrition data.

4. **Results**

The analysis found significant variation in nutrient content across foods. Dried fish and protein powders had the highest protein values, while oils and drinks contained almost none. Animal-based foods dominated the upper end of both protein and fat, while carbohydrate-rich foods were mostly breads, cereals, and sweets.

Bar charts showed distinct differences in average nutrient content by category. Scatterplots revealed a tendency for protein and fat to rise together in certain foods, and clustering grouped foods into high-protein/fat, high-carb, and intermediate clusters that align with common dietary patterns.

5. **Discussion**

The findings demonstrate how exploratory data analysis and visualization can quickly surface key trends for nutritional comparison. For example, clustering made it easy to see which foods cluster nutritionally, and the plots highlighted the differences between food groups.

These results matter because clear visualization of nutrient content helps inform healthier choices, both at the consumer level and for policy or menu planning. Future work could expand to other nutrients or dietary factors (such as sodium or micronutrients), or apply more sophisticated modeling or grouping.

6. **Coding & Reference**

- All code, data, and plots are included in the GitHub repository under the appropriate folders: /data (dataset), /graphs (visualizations), /codes (Jupyter notebook), report (pdf & docs.

- Dataset: CORGIS Food Dataset – https://corgis-edu.github.io/corgis/csv/food/