The American University in Cairo

DSCI 2411

Food Data Central

[Asmaa Elabasy, ID: 900205076]

1 Introduction

The FoodData Central dataset, provided by the USDA, is a comprehensive resource that contains food composition data with an emphasis on nutritional values and components in different food types. This data provides a clear understanding of the relationship between different constituents of nutrients in the food. In addition, detailed nutritional information and food components are useful to researchers, policymakers, nutritionists, and product developers.

The primary objective of the dataset is to provide transparent, scientifically rigorous, and easily accessible information about nutrients and other food components. This helps answer questions about the nutritional value of foods and allows people to make informed dietary choices. The dataset covers multiple data types to cater to various uses, with each type having a unique purpose and data collection method.

2 Problem Definition

One of the challenges in the food industry and public health is the variation in nutrient composition between branded and unbranded foods. Branded foods often have specific nutrient values derived from food labels, which may be influenced by marketing and packaging. However, unbranded foods or raw ingredients are often measured through analytical testing or survey data that may not include packaging or brand-specific details. Understanding these differences is crucial to providing accurate nutritional advice and guidelines. On the other hand, understanding the composition of nutrients and their relationships with Energy in food, could be the base for experimental food to obtain a good composition.

This analysis is aiming to nvestigate the relationships between food composition, calorie content, and serving sizes, particularly when distinguishing branded foods from other food types.

2.1 Description of the Dataset

The dataset includes data from multiple sources, such as food manufacturers (for branded foods), analytical testing (for foundation foods), research studies (for experimental foods), and dietary surveys (FNDDS).

Variables in the dataset cover nutrient content such as calories, protein, fat, carbohydrates, food categories as fruits, grains, dairy, serving sizes, and product-specific details (for branded foods).

Key Differences Between Branded Foods and Other Food Types in the Dataset
The dataset categorizes food items into different types, each having distinct characteristics and data sources.

1. Branded Foods: Branded foods are sourced from food manufacturers or companies that produce packaged, branded products. These manufacturers provide the nutrient content based on the information found on food labels.

2. Foundation Foods: The data comes from analytical testing of the food itself. For example, a generic, unbranded banana might have its nutrient data determined through lab tests, which provide the basic nutrient breakdown of the food.

3. Experimental Foods: Data for experimental foods is collected from scientific research studies. These foods might be tested under experimental conditions like testing genetically modified crops, new farming techniques, or altered food production methods).

4. FNDDS (Food and Nutrient Database for Dietary Studies)

   The FNDDS data is derived from survey data collected through the National Health and Nutrition Examination Survey (NHANES). The data represents the dietary intake of the general population, reflecting what people actually eat.

5. SR Legacy: SR Legacy contains historical data derived from the USDA's National Nutrient Database (NNDB), which has been in use for decades. The data comes from older food composition records that were used for research purposes.

3 Data Summarization

Understanding the distribution of food types, their nutritional composition, and average energy contributions helps create a comprehensive view of the dataset. The following visualizations provide insights into the dataset structure and highlight key differences in nutrient values and energy distribution.

3.1 Count of Types of Food in the Dataset

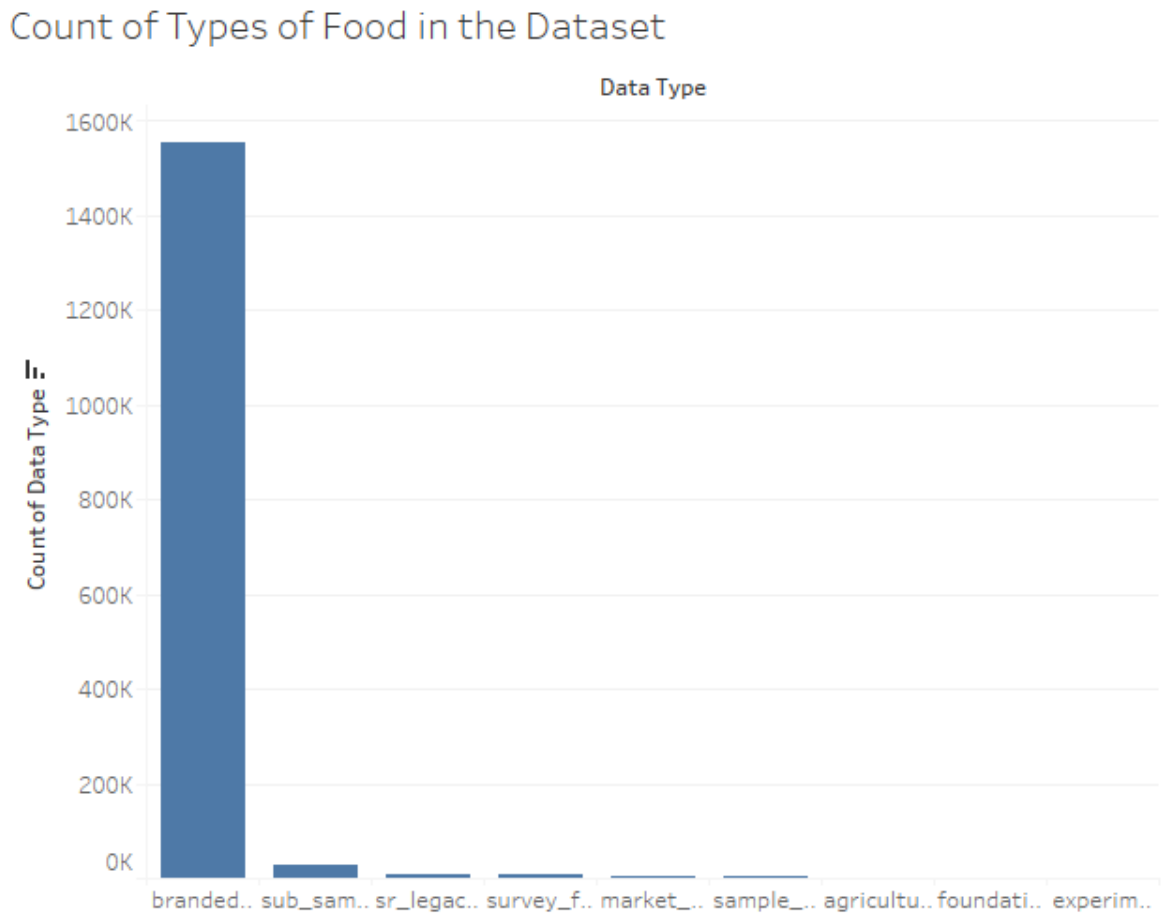

Figure 1: Count of Types of Food in the Dataset

This bar chart illustrates the number of entries for each food type (e.g., Branded Foods, Foundation Foods, SR Legacy, etc.).

- Branded Foods dominate the dataset, highlighting the extensive data provided by food manufacturers.

- Categories such as Foundation Foods and SR Legacy contain fewer records, which aligns with their more specialized and research-focused nature.

3.2 Average Amount of Nutrients in Different Food Categories

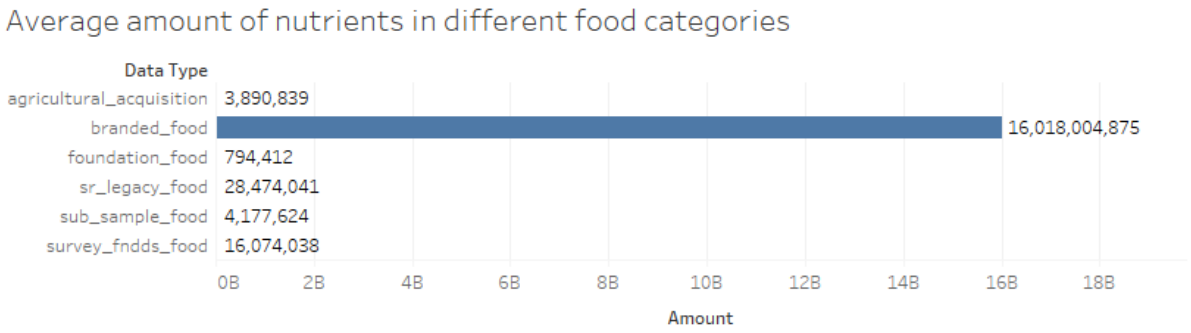

Figure 2: Average Amount of Nutrients in Different Food Categories

The bar chart highlights the average amount of nutrients for each food category. Branded Food account for the largest share of nutrients, followed by SR Legacy Foods. Categories such as Agricultural Acquisition and Sub-Sample Foods contribute significantly less to the overall nutrient count.

3.3 Average Amount of Energy in Different Food Categories

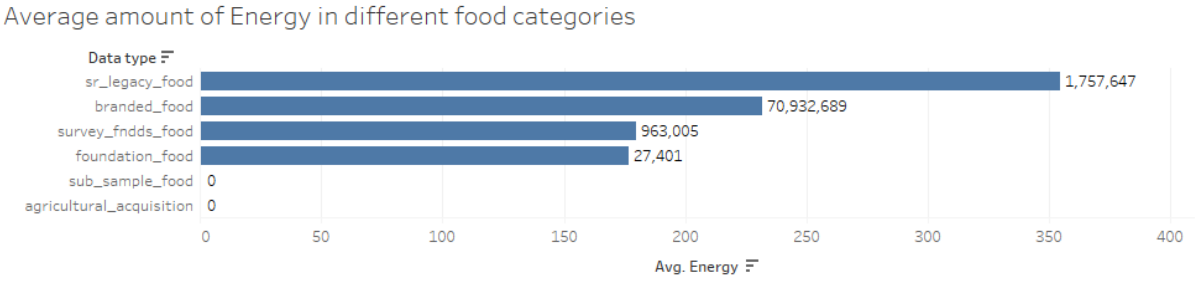

Figure 3: Average Amount of Energy in Different Food Categories

This chart compares the average energy (calories) across different food categories. The SR Legacy food type shows the highest average energy, followed by Branded Foods and FNDDS (Food and Nutrient Database for Dietary Studies). Foundation Foods have significantly lower energy values, and certain categories (e.g., Agricultural Acquisition and Sub-Sample Foods) show negligible or zero energy values.

3.4 Distribution of Energy (Calories) in all Foods

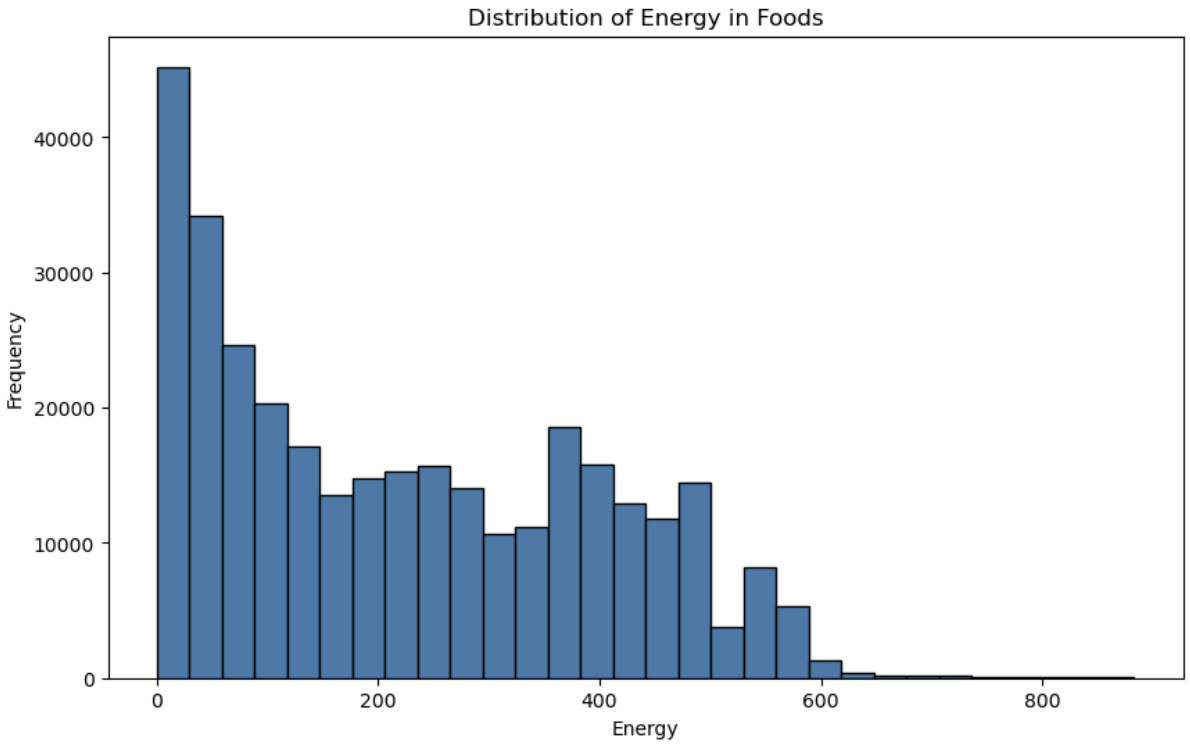

Figure 4: Distribution of Energy in Food dataset

The histogram of energy (calories) distribution shows a right-skewed distribution, with most food items having lower energy values. The majority of food items fall within the 0–200 calorie range per serving. Only a small proportion of foods have higher caloric values exceeding 400 calories. The skewed nature of the data highlights that the dataset contains a mix of low- and high-calorie items, with the former being more common.

3.5 Distribution of Nutrient Components in Food dataset

The distributions of different nutrient components (Carbohydrates, Protein, Fat, Sodium, and Sugars) across the dataset are similarly skewed, with most food items having lower values for these nutrients. The histograms reveal:

- Carbohydrates: The majority of food items contain less than 50 grams of carbohydrates per serving, but a small number of foods exceed this amount, likely representing carbohydrate-rich processed foods.

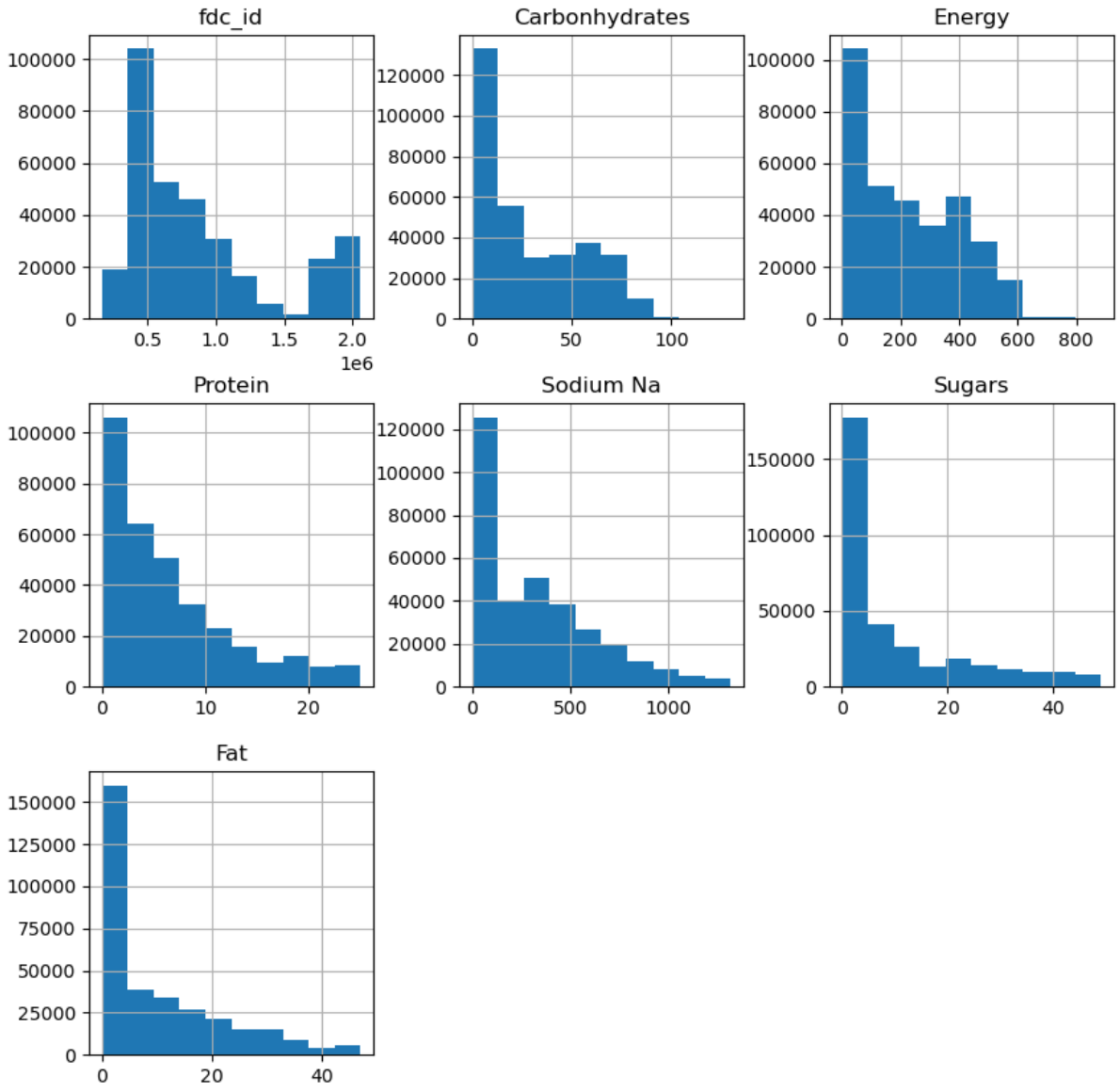

Figure 5: Distribution of Nutrient Components in Food Dataset

- Carbohydrates: The majority of food items contain less than 50 grams of carbohydrates per serving, but a small number of foods exceed this amount, likely representing carbohydrate-rich processed foods.
- Protein: Protein content is generally low, with most food items containing less than 10 grams per serving. Foods with higher protein content likely include meat, dairy, or protein-rich branded products.
- Fat: Most foods contain less than 20 grams of fat, with only a small number of items exceeding this threshold. High-fat foods could include snacks, processed meats, or desserts.
- Sodium: Sodium content varies widely, with a noticeable spike for low-sodium foods. Higher sodium values are likely associated with processed and branded foods.
- Sugars: Sugars are concentrated at lower levels (below 10 grams per serving), though a few items exceed 40 grams, likely reflecting desserts, beverages, or sweetened branded products.

3.6 Trends in Food Categories Over Time

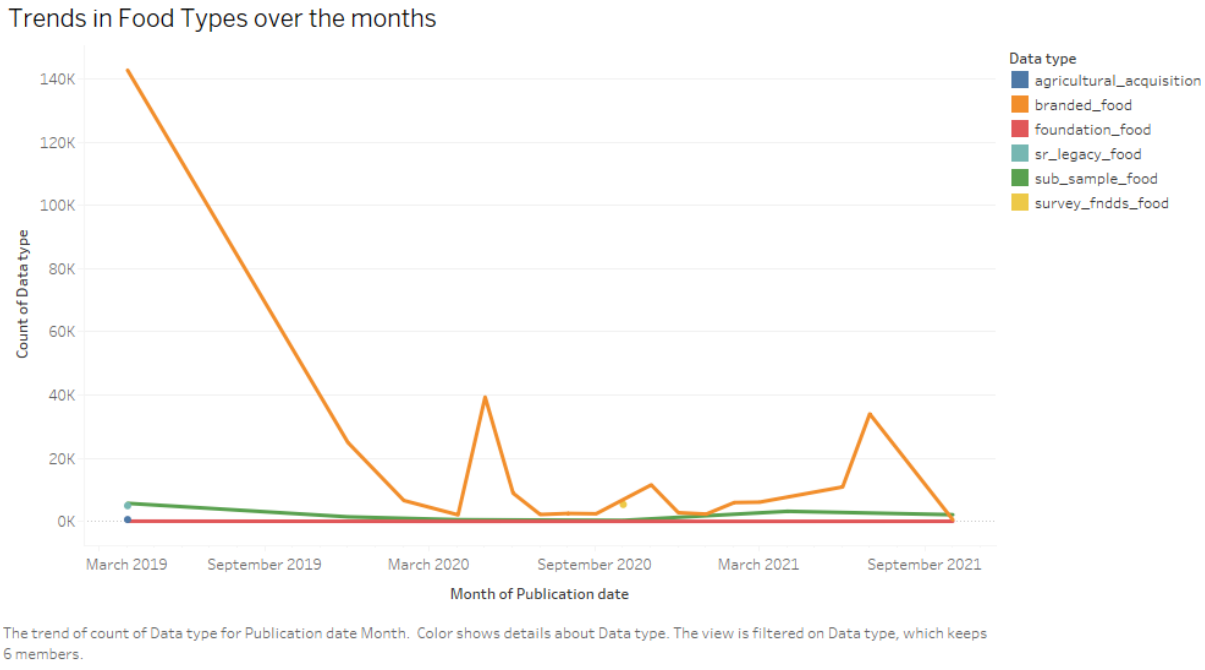

Figure 6: Trends in Food Categories Over Time

This line chart shows the count of data entries for different food categories over time (based on the month of publication). The majority of entries belong to Branded Foods, with a significant peak in early 2019, followed by a sharp decline and fluctuating values over subsequent years. Other food categories (e.g., Foundation Foods, SR Legacy) have relatively stable but minimal contributions over time.

One possible reason behind the declination in 2020 is the change in the market during pandemic of COVID-19.

However, this declination did not affect the composition of nutrients in different Branded Food Category as the distribution of the components is almost the same over time as show in figure 7.

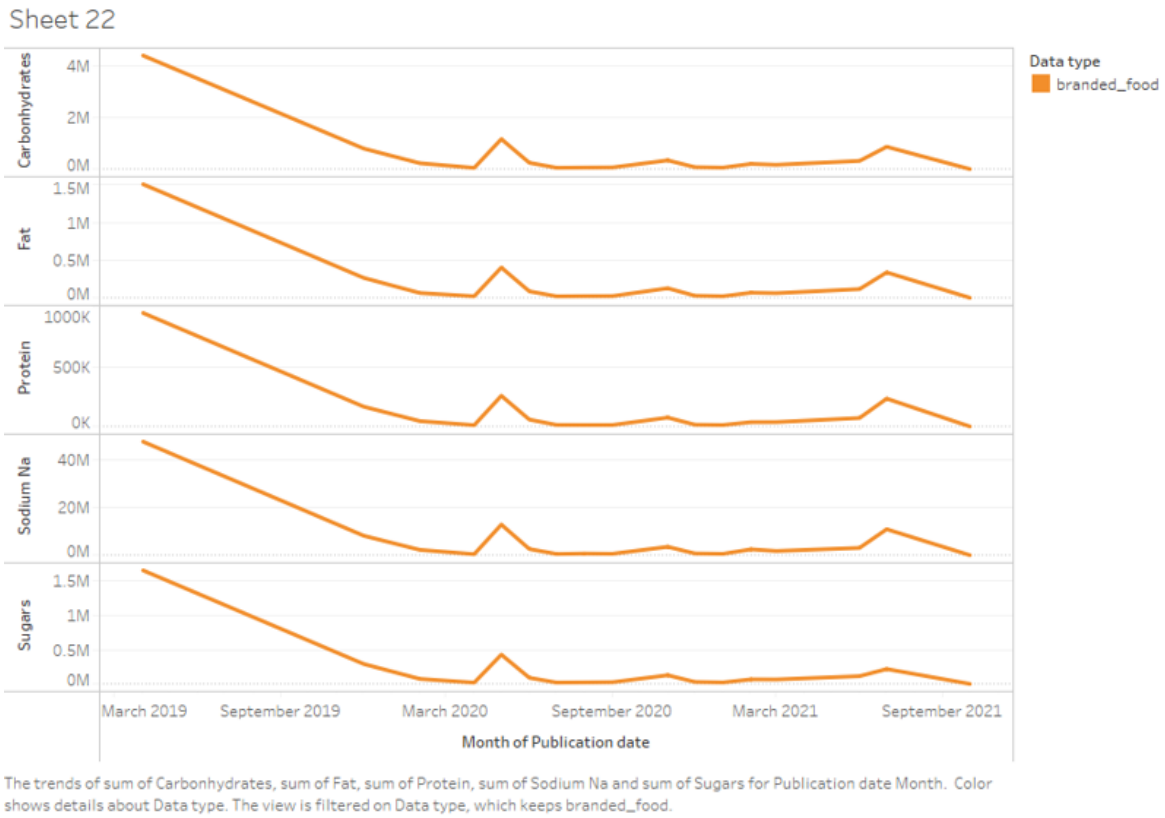

Figure 7: Trends of nutrients in Food Categories Over Time

4 Determining the Relationships between Attributes

This section is mainly focused on Branded Foods to understand how different nutrient components—such as carbohydrates, protein, fat, sugars, and sodium—relate to each other and to Energy (calories), since Branded Foods make up the largest portion of the dataset.

4.1 Count of Branded Food Categories

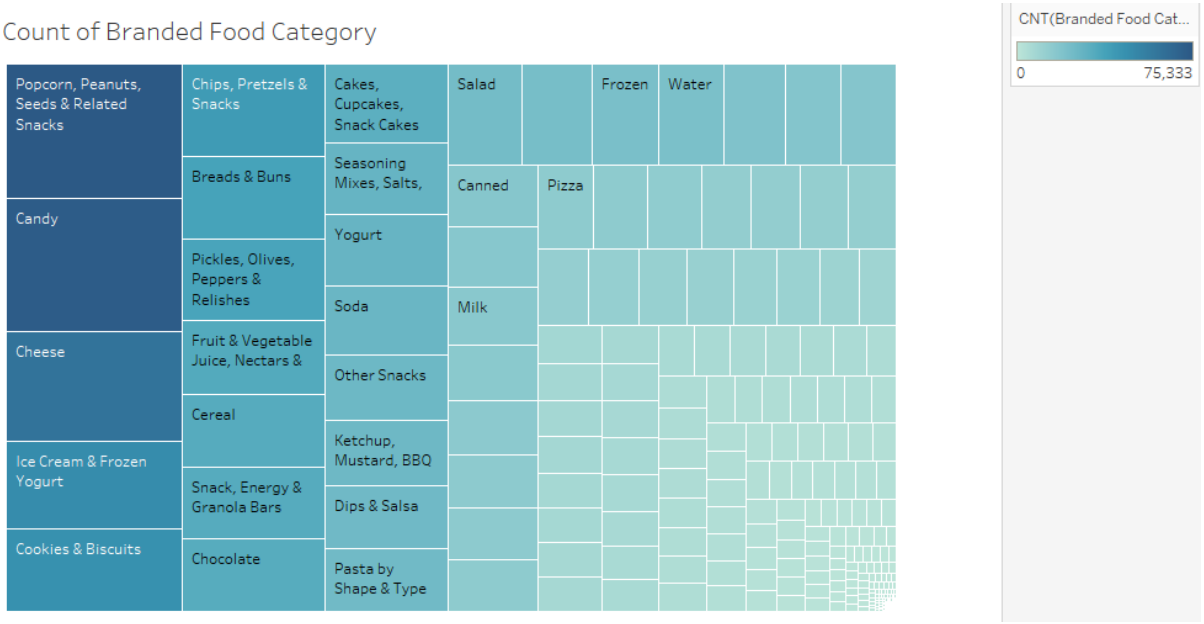

Figure 8: Count of produced Branded Food Categories

The tree map shows the distribution of Branded Food categories by count. Categories such as Popcorn, Pretzels, Snacks, Candy, and Cheese dominate in terms of production frequency, while categories like Frozen Food and milk have significantly fewer products.

This reflects consumer preferences for processed and convenient foods, which are commonly mass-produced.

Lower counts in categories like Frozen Food and Canned Food may indicate niche markets or less diversity in branded options.

4.2 Relationship between Serving Size and Energy in Branded Food Category

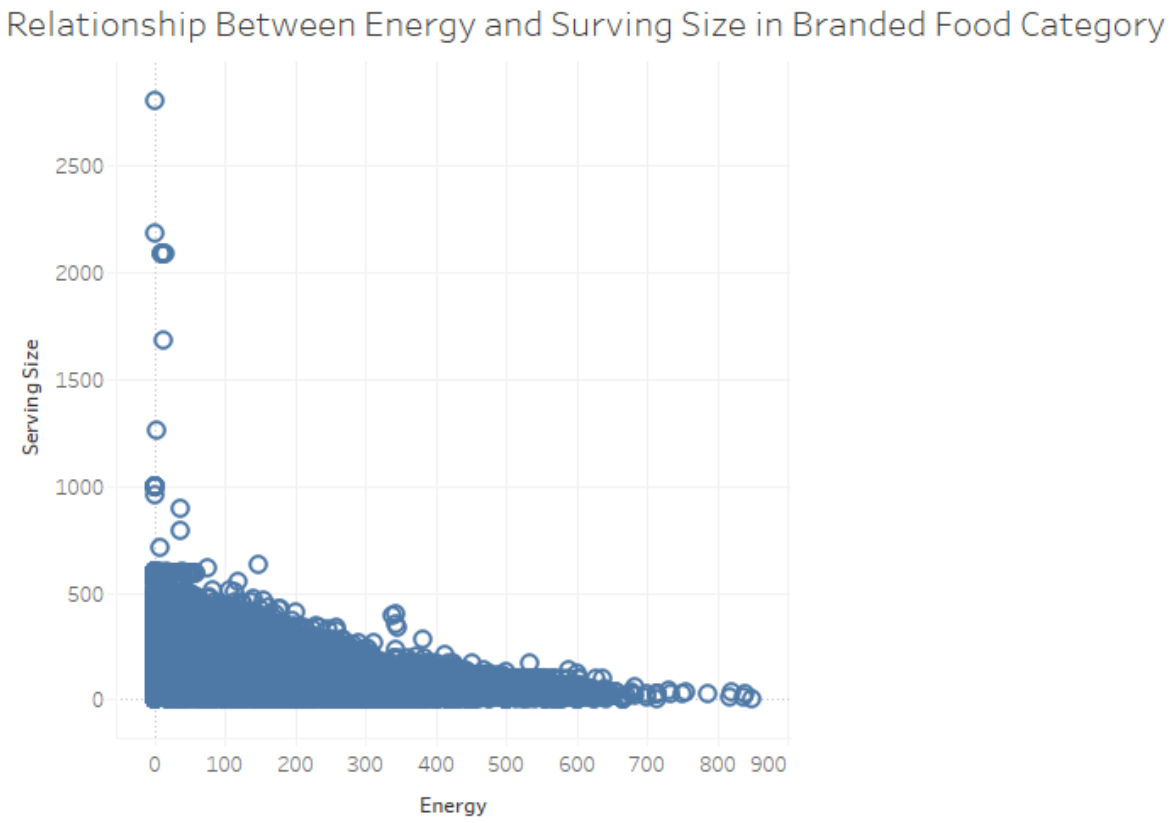

Figure 9: Relationship between Serving size in Branded Food category and Energy

Figure 9 highlights the relationship between serving size and energy in Branded Foods. Most foods cluster around lower energy values (0-200 calories) and smaller serving sizes (0-200 grams)

The lack of a strong linear relationship suggests serving size is not a consistent predictor of energy density across all branded foods.

Average Serving Sizes Across Branded Food Categories

The box plot below highlights serving size variations across different branded food categories, with Powdered Drinks standing out as having significantly higher average serving sizes.

Powdered Drinks: Their larger serving sizes may reflect the inclusion of mixing instructions.

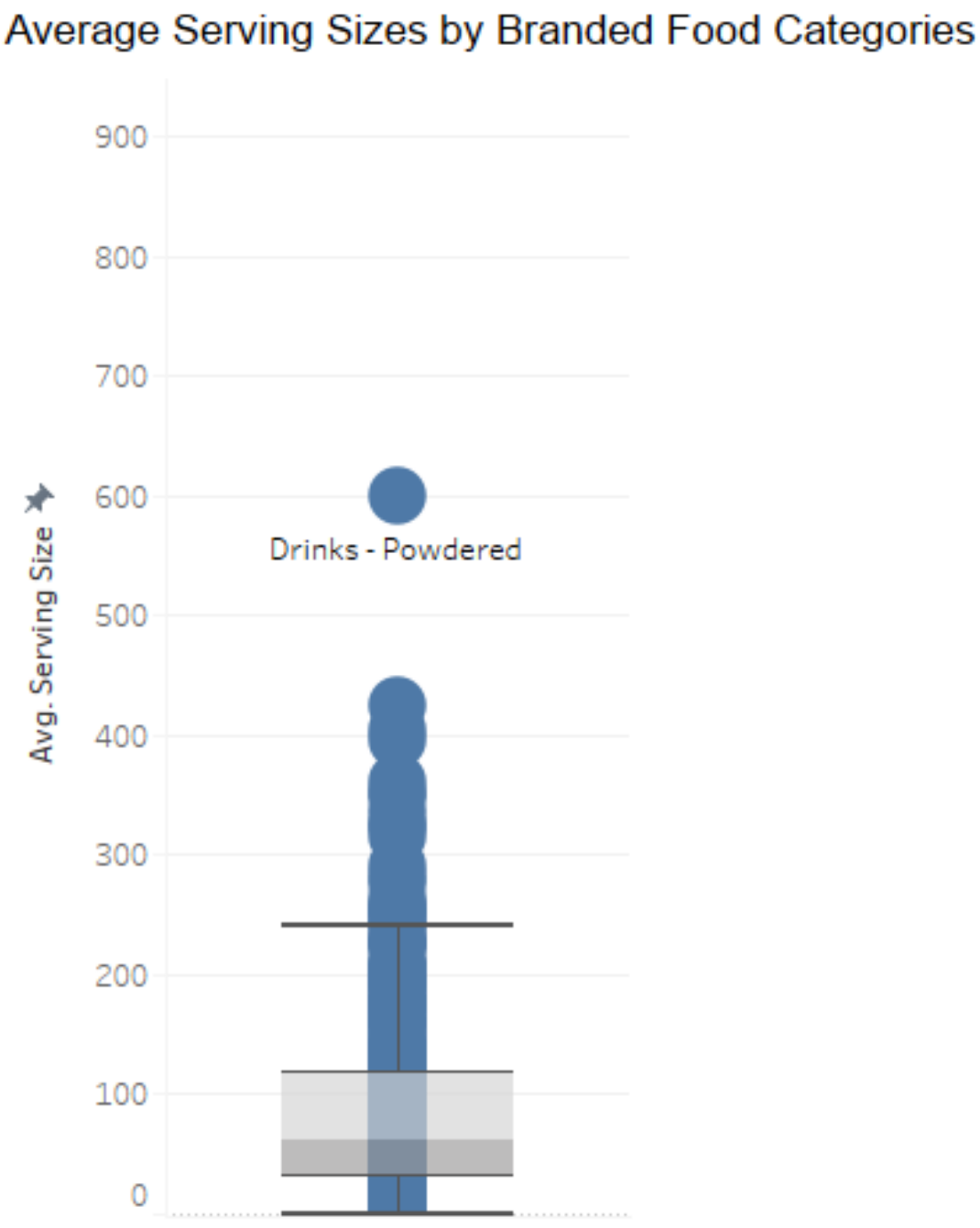

Figure 10: Average Serving Sizes Across Branded Food Categories

4.3 Nutrients in Different Branded Food Categories

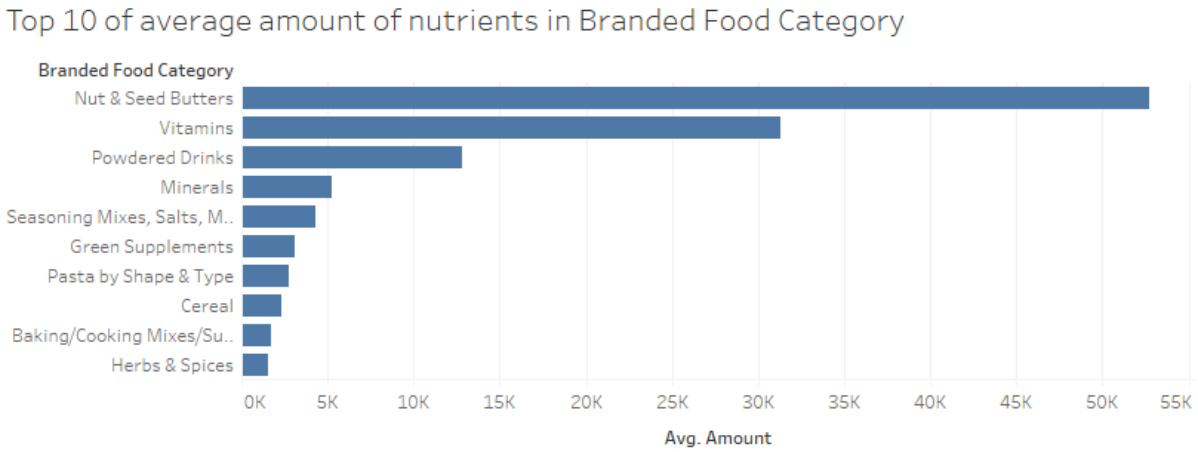

Figure 11: Branded Food Categories with highest Average amount of Nutrients

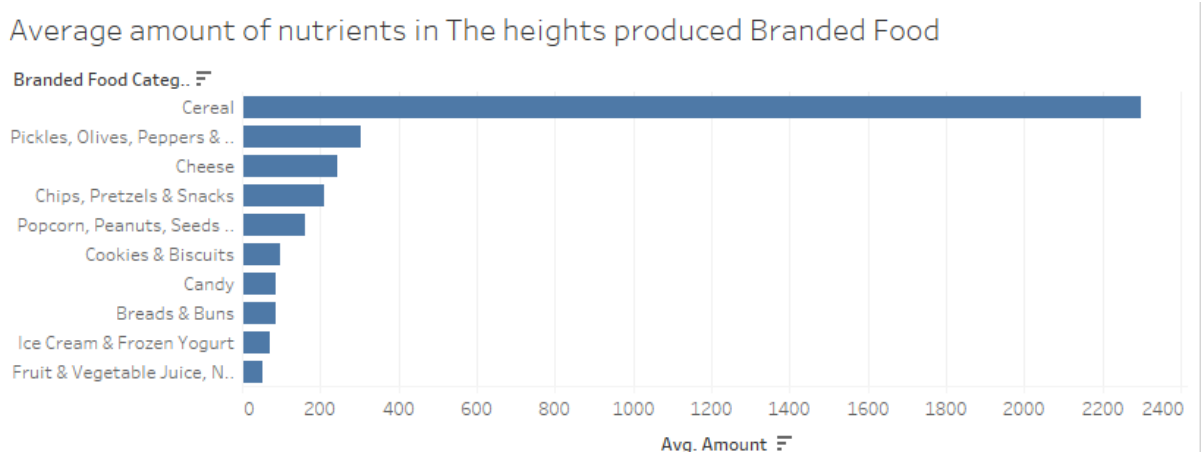

Figure 12: Average amount of nutrients in the highest produced Branded Food

Figure 11 shows the top 10 Branded Food categories with the highest average nutrient amounts. Categories like Nut and Seed Butters, Vitamins, and Powdered Drinks lead in nutrient density, while Cereal and Herbs and Spices also show high nutrient values.

Nutrient-dense products such as Nut and Seed Butters and Powdered Drinks cater to specific dietary needs, such as high-protein diets or fortified supplements. Categories with a focus on health or specialty diets (e.g., vitamins, minerals) rank higher in nutrient content despite having lower production counts compared to snacks and candy.

On the other hand, figure 12 shows that Cereal, a category with high production, ranks significantly lower in terms of average nutrient value.

Thus, high production does not necessarily equate to high nutritional value. Branded

Foods that are mass-produced, such as Cereal and Candy, are often energy-dense but lack significant amounts of beneficial nutrients.

4.4 The correlation between different food nutrients and Energy

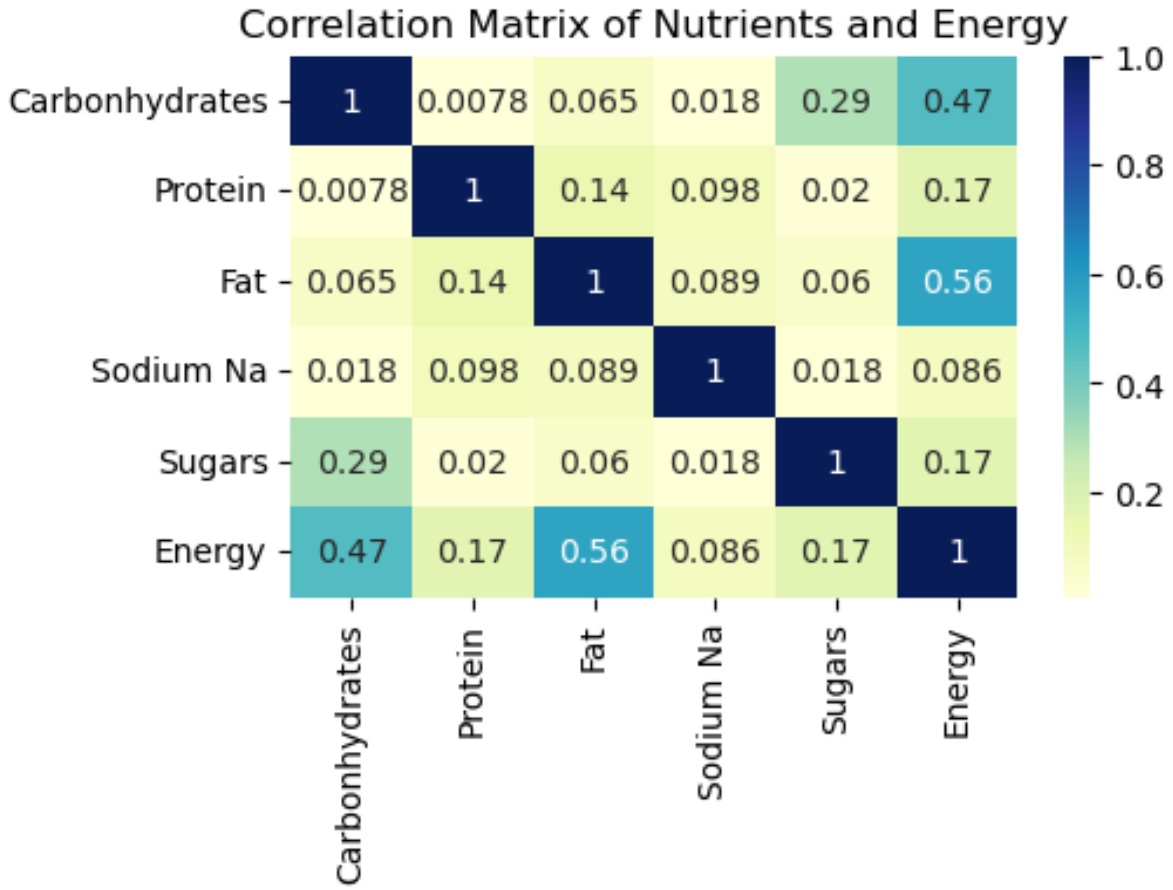

Figure 13: Correlation coefficient between different nutrients and Energy

- Fat and Energy: Correlation Value: 0.56 (strongest positive correlation)
  Fat content has the strongest correlation with energy.

- Carbohydrates and Energy: Correlation Value: 0.47 (moderate positive correlation)
  Insight: Branded Foods with higher carbohydrate content tend to have higher energy (calorie) levels. This aligns with the fact that carbohydrates contribute around 4 kcal per gram as in the nutrient calories conversion file in the dataset.

- Fat and Energy: Correlation Value: 0.56 (strongest positive correlation) Insight: Fat content has the strongest correlation with energy. T
- Protein and Energy: Correlation Value: 0.17 (weak positive correlation)
- Sugars and Energy: Correlation Value: 0.17 (weak positive correlation)
- Sodium and Energy: Correlation Value: 0.086 (very weak positive correlation)

However; the relationship between Sugar, Sodium and Energy is not linear.

4.5 Energy vs. Nutrients

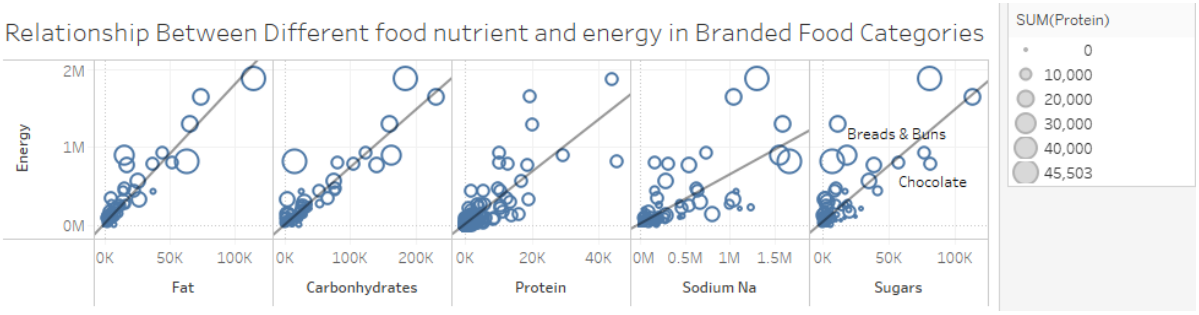

Figure 14: Relationship between Nutrients and Energy

- A strong linear trend is visible between fat, carbohydrates and energy. The larger bubbles correspond to foods with moderate to high protein content.
- Protein's relationship with energy is less pronounced, with fewer points in the high-protein and high-energy range.
- Sodium and Sugar has a weaker linear relationship with energy. Therefore, they do not follow the linear model.

Relation to Correlation Matrix: The near-zero correlation (0.086) confirms that sodium content does not significantly influence calorie levels. Sugars:

Observation: Foods high in sugar, such as chocolate, show a positive trend with energy. The presence of sugar-rich foods aligns with carbohydrate-heavy foods contributing to higher calories. Relation to Correlation Matrix: The weak correlation (0.17) suggests that while sugar contributes to energy, its impact is less significant than that of total carbohydrates or fat.

4.6 Carbohydrates, Proteins and Fats in Different Food Categories

Balanced Diet Vs. Branded Foods:

A Balanced Diet typically has a higher protein ratio compared to fat (Harvard T.H.). The generally recommended dietary composition suggests that protein should contribute more to the total calorie intake than fat to maintain muscle mass and other bodily functions. However, the data from Branded Foods does not follow this trend. Instead, it shows a higher carbohydrate contribution, with moderate levels of fat and protein, which is more reflective of high-carb, low-protein diets found in processed foods as in figure 10.

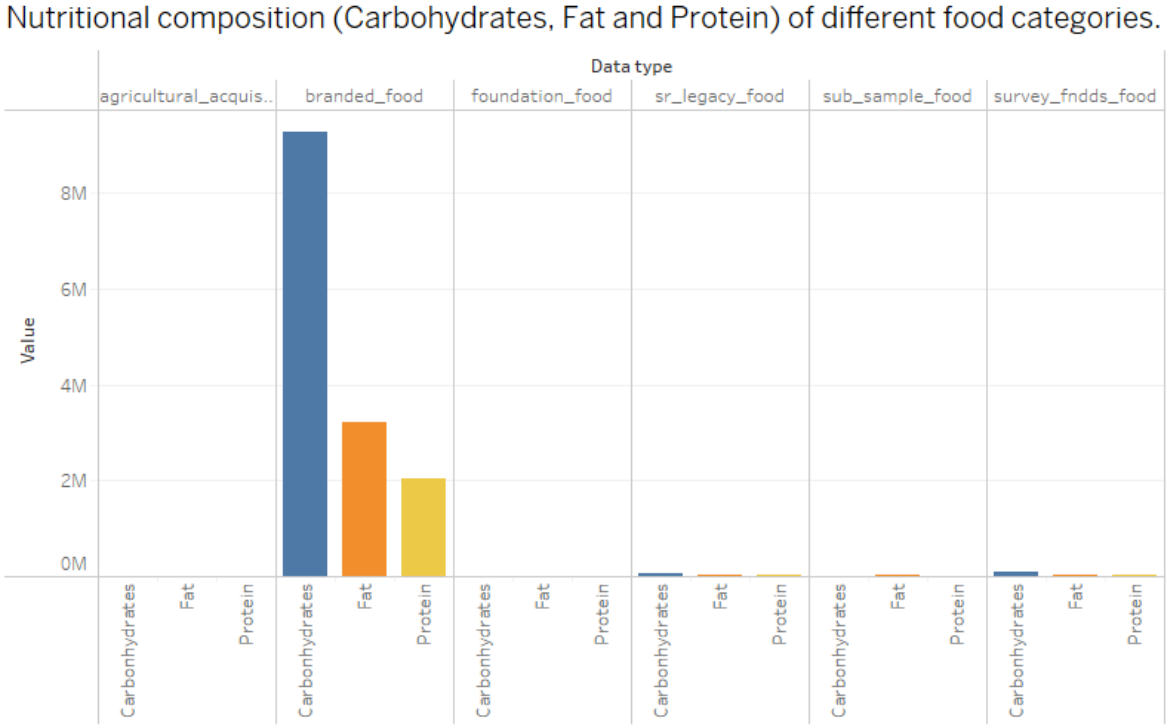

Figure 15: Nutrient Composition of Branded Foods

The stacked bar chart illustrates the composition of carbohydrates, fat, and protein across various food categories. Branded Foods contribute overwhelmingly to carbohydrate content, with moderate levels of protein and fat.

To address the composition of these nutrients in different food categories, especially SR Legacy Foods, eliminating Branded Food Category provides a clear picture of the distribution. SR Legacy Foods show a stronger trend towards balancing nutrients, particularly in terms of protein and fat. The experimental nature of SR Legacy Foods suggests that

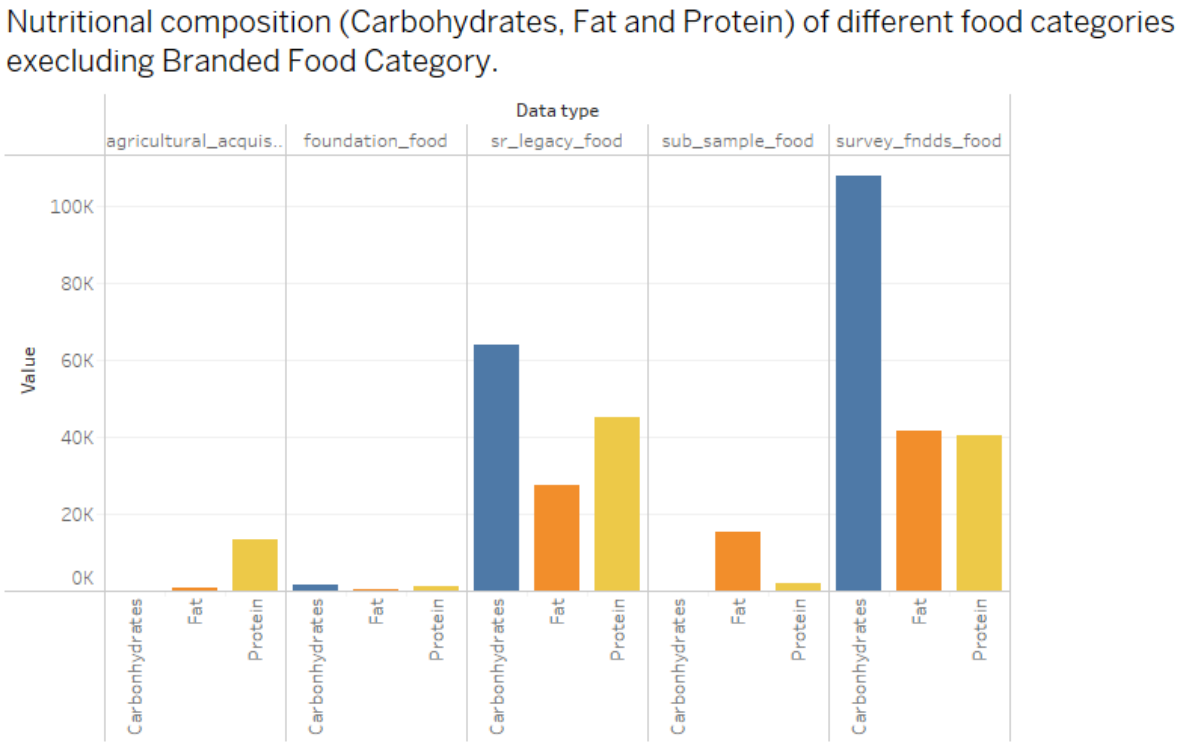

Figure 16: Nutrient Composition of Food Categories without Branded Food

there is an intentional effort to balance nutrient profiles, with protein and fat levels more closely aligned, which may indicate a shift in food science towards healthier, more balanced formulations.

4.7 Nutrient Composition in SR Legacy Foods

Figure 17 illustrates the amount of nutrients in the SR Legacy Foods category, with Sodium dominating the nutrient content compared to others like Protein, Fat, and Sugars.

Sodium: The significantly higher sodium levels reflect the inclusion of processed or preserved foods in SR Legacy data, which likely includes items with added salts for preservation.

Balanced Macronutrients: Lower but relatively consistent levels of carbohydrates, protein, and fats suggest SR Legacy Foods focus on well-rounded nutrition, aligning with its historical data collection goals for diverse foods.

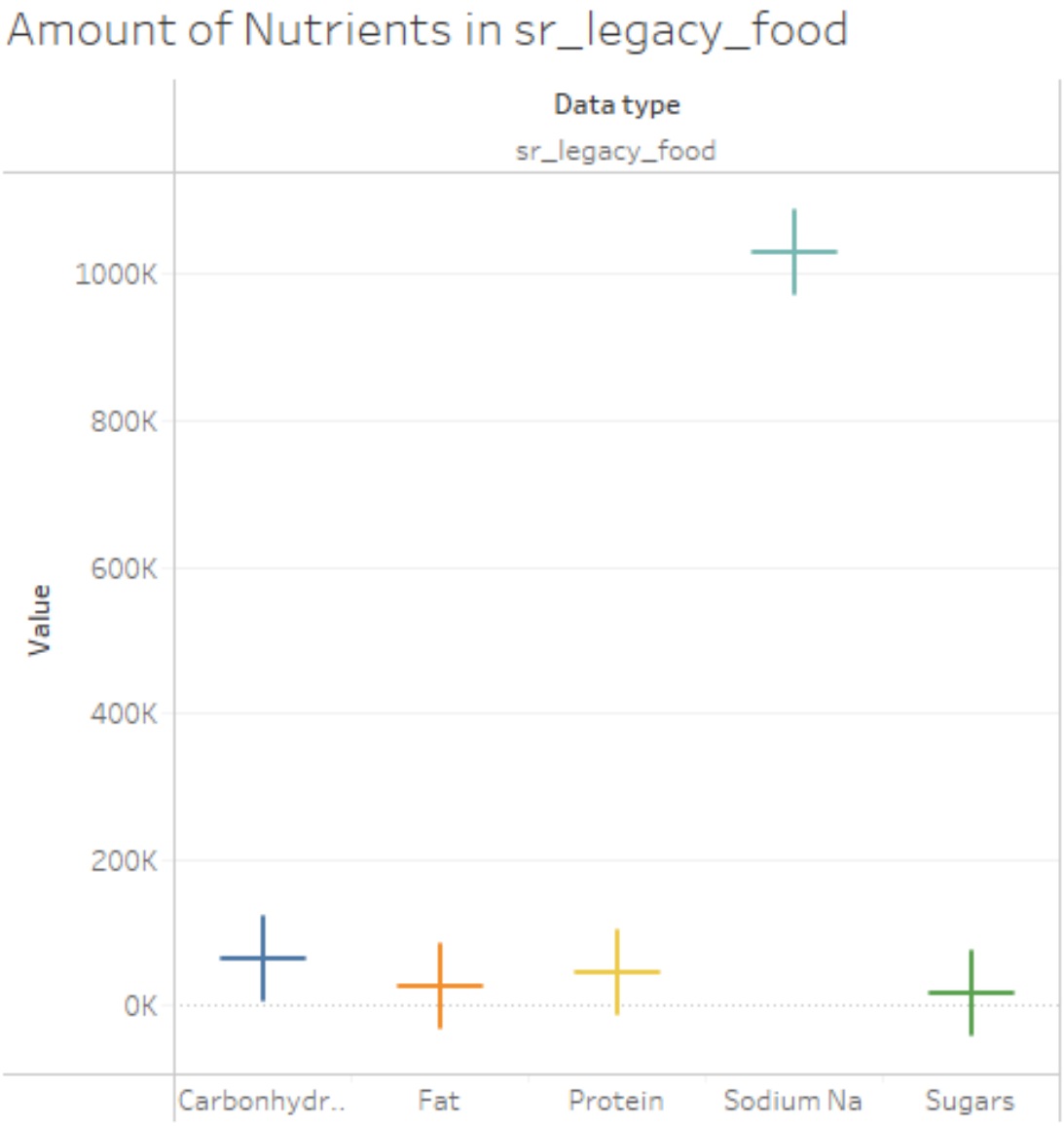

Figure 17: Nutrient Composition in SR Legacy Foods

5 Conclusion

Branded Foods dominate the marketplace. These are the snacks, drinks, and meals we see on store shelves. While these foods make up the majority of what's available, they often lack the nutritional balance we need.

They provide quick energy but can lead to overconsumption of calories. This imbalance doesn't align with the idea of a healthy diet, which emphasizes more protein and balanced macronutrients.

In contrast, foods in categories like Foundation Foods—raw or minimally processed items—offer a much better nutrient profile. They naturally provide more protein and fewer empty calories, helping us stay healthier. But these foods are less represented in the dataset and, often, in our diets.

There's a growing opportunity to reformulate popular products to include more protein, reduce excessive sugars, and create foods that are both convenient and nutrient-dense. For Public Health Advocates:

6 Refrences

- Harvard T.H. Chan School of Public Health. (n.d.). The healthy eating plate. The Nutrition Source. Retrieved December 9, 2024, from https://nutritionsource.hsph.harvard.edu/healthy-eating-plate/

- U.S. Department of Agriculture, Agricultural Research Service. (n.d.). Food-Data Central. U.S. Department of Agriculture. Retrieved December 9, 2024, from https://fdc.nal.usda.gov/