# Analysis of calories in various food products

Name student: Jane Doe

Student number: 123456

Date: 2025-04-10

## Introduction

In today's society, there's a widespread awareness regarding food consumption. Quantifying caloric intake and reducing lipid consumption are common recommendations from dietitians and nutritionists. Therefore, knowledge about the nutrient composition of various foods is essential. The dataset analyzed in this report comprises a comma-separated values (CSV) file with data from over 300 food products. For each item, the quantities of calories, lipids, proteins, saturated fatty acids, carbohydrates, and dietary fibers are labeled. Furthermore, the food products are classified into different categories, such as desserts, vegetables, fruits, and so on.

## Dataset

The dataset originates from Kaggle ([Nutritional Facts for most common foods](https://www.kaggle.com/datasets/niharika41298/nutrition-details-for-most-common-foods)). The dataset uses Wikipedia as a reference ([Table of food nutrients - Wikipedia](https://en.wikipedia.org/wiki/Table_of_food_nutrients)).

## Aim

The primary goal of the data analysis of this dataset is to quantify and compare the calorie content between different food groups and within individual food groups.

The analysis specifically aims to provide the following insights:

* Identification of the food group(s) with the highest average calorie content. This helps us determine which categories of food generally contribute the most to total energy intake.
* Identification of the specific food item within the food group that has the highest average calorie count. This pinpoints individual food items that could significantly contribute to high calorie consumption.
* Identification of the item within the entire dataset with the highest number of calories.
* Highlighting items with a caloric amount higher than the median.

## Data import

First, the data was viewed in Visual Studio Code. A screenshot can be seen below.

A screen shot of a computer program

AI-generated content may be incorrect.

Figure 1: analysis of the csv file in Visual Studio.

Looking at the file, it's a CSV file with the comma as the column separator and decimals separated by a period.

Next, I loaded the file into Excel via Power Query:

A screenshot of a computer

AI-generated content may be incorrect.

Figure 2. Data import via Power Query.

Next, I changed the table selection name (Formulas > Name Manager) to "data" to make future selections easier.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 3. Name manager. Selection name changed.

## Cleaning

First, I checked for empty values in the Calories column using the COUNTBLANK function.

It turned out there was 1 missing value.

I then used XLOOKUP to identify it:

A white grid with black text

AI-generated content may be incorrect.

Figure 4. Missing values.

An inspection of the data reveals that some rows contain data ranges:

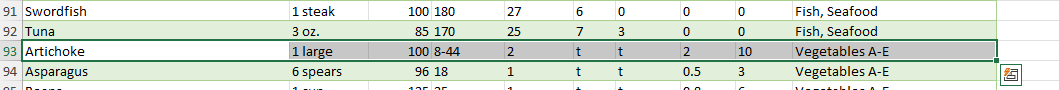


Figure 5. Data range instead of value.

I therefore first filtered and removed these values.

Figure 6. Empy cell for Frozen peas.

Frozen peas had an empty value for this column. This row was also removed.

I removed this data from the source file (CSV) and reloaded it so that the calories are no longer interpreted as a range.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 7. Cleaned values loaded again.

With this cleaned data, a pivot chart can now be created.

Identification of the Food Group(s) with the Highest and Lowest Average Calorie Content

To identify food groups with the highest fat content, I created a pivot chart:

A graph of food categories

AI-generated content may be incorrect.

Figure 8. Pivot chart of calories plotted against categories.

As seen in this chart, seeds and nuts have the highest average calorie content, while vegetables have the lowest.

To gain a bit more insight into the data's distribution, I also created a Boxplot:

A graph with blue squares and black text

AI-generated content may be incorrect.

Figure 9. Boxplot of calories in different food categories.

Here too, it's clear that seeds and nuts come out highest (highest median). Bread and breakfast shows a large spread.

Identification of the Specific Food Item within the Food Group with the Highest Average Calorie Content

To identify food items with the highest calorie count within the highest average category, I first created a bar plot of all items:

A graph with blue lines and white text

AI-generated content may be incorrect.

Figure 10. Barplot of calories versus all items.

And then filtered by seeds and nuts:

A graph of calories per item

AI-generated content may be incorrect.

Figure 11. Seeds and nuts.

This shows that Brazil nuts and Roasted and salted can be classified as having the highest calories.

I then also created a pie chart for this analysis:

A pie chart with numbers and a number of calories per item

AI-generated content may be incorrect.

Figure 12. PIE chart of the data.

Identification of the item with the highest calorie count within the entire dataset.

Since we initially looked at each category separately, it's possible that an individual item has a very high categorical value but doesn't fall into the same category as the items with the highest average categorical value.

Because the initial analysis was done per category, it's possible that a single item exists with an exceptionally high caloric value, even if it doesn't belong to the category that has the highest average caloric value.

Therefore, an analysis was also conducted on the highest and lowest values. For this, I used the MIN, MAX, and XLOOKUP functions. The following results were obtained:

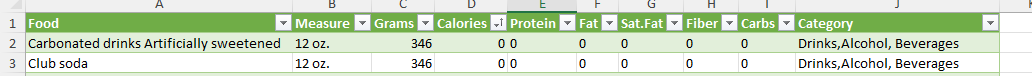
A screenshot of a graph

AI-generated content may be incorrect.

*Figure 13: Identification of the food item with the maximum calories using XLOOKUP.*

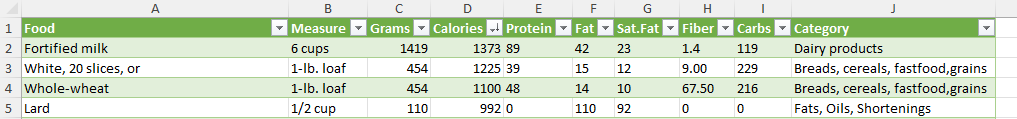
These were checked by sorting:

Two items were found with 0 calories:



*Figure 14: Sorted by calorie amount.*

And indeed, the same item was found for the maximum number of calories:



*Figure 15: Sorted by calorie amount.*

Marking items with a higher caloric quantity than the median.

To flag which food items have a high caloric value, I used the median. The median is more robust than the average and less sensitive to individual outliers. First, I calculated the median of the calorie count. Additionally, I counted how many rows were greater than the median: A white grid with black text

AI-generated content may be incorrect.

*Figure 16: Median of calorie count and the number of rows higher than the median.*

Next, I used this value with Conditional Formatting to indicate which items have a high caloric value. The dataset is sorted alphabetically by food items.

A screenshot of a computer

AI-generated content may be incorrect.

*Figure 17: Conditional formatting for highlighting high-calorie items.*

Conclusion and Discussion:

The analysis of calorie content across over 300 food items, categorized into various groups, reveals significant variations. Seeds and nuts were identified as the food group with the highest average calorie count, while vegetables generally showed the lowest calorie content. Within the seeds and nuts category, Brazil nuts and roasted, salted nuts contained the highest individual calorie counts. Notably, pure vegetable oil was identified as the food item with the absolute maximum calories in the entire dataset. Conversely, two food items were found to have a calorie content of zero.

By calculating the median calorie content and applying it for conditional formatting, food items with a relatively high calorie content were visually highlighted in the dataset. These findings underscore the wide range of calories within different food groups and individual products.

The results can be valuable for consumers looking to monitor their calorie intake, as well as for professionals in the food industry and healthcare. The wide spread within some categories, such as bread and cereals, suggests that further, more detailed analyses within these groups could be beneficial.

## Checklist:

A screenshot of a computer

AI-generated content may be incorrect.

It appears that two elements are still missing from the portfolio:

* Clustered bar plot
* XY scatterplot