Ingredient Analysis

Victor Nguyen

2022/11/07 - 2022/11/

Contents

Purpose

The purpose of this investigation is to find relationships between certain nutrients. Discovering these correlations can make designing healthy diets easier and more efficient.

Data

The data set analyzed is an ingredient data set from CORGIS (The Collection of Really Great, Interesting, Situated Datasets [1]). It includes nutritional information on various food ingredients, collected from the United States Department of Agriculture's (USDA) Food Composition Database [2].

Obtaining the data set was simple: the CSV file was accessible through a download link on the website. It is a public data set and does not need any forms or contact with the authors to access.

Data Processing

To load the data into R, the `read_csv` function from tidyverse was used. Looking at the data, the last row contains details about Vitamin D as an ingredient, but since this analysis is only interested in food items, it was removed. Additionally, its data does not have any nonzero values, so it does not have much purpose for this evaluation. The data type of the category column was changed from character to factor, and the spaces from the column names were removed for easier handling. The types of the other columns automatically

assigned by R already properly represented their data. Finally, the Vitamin A column was renamed to be more concise. There were no unknown values in the data set.

Cleaned Data

The cleaned data consists of information in a tabular format. Each row is an observation of a food item or ingredient. The columns are named in a hierarchical manner where the category, description, and identifiers are separate from the numerical columns. This breaks down into more levels, such as vitamins and minerals.

Variable of Interest	Type	Description	Missing Values?
Data.Vitamins.VitaminB12	double	Amount of Vitamin B12, measured in micrograms (mcg)	No
Data.Vitamins.VitaminB6	double	Amount of Vitamin B6, measured in miiligrams (mg)	No
Data.MajorMinerals.Copper	double	Amount of copper, measured in milligrams (mg)	No
Data.MajorMinerals.Zinc	double	Amount of zinc, measured in milligrams (mg)	No
Data.Water	double	Amount of water, measured in grams (g)	No
Data.Fat.TotalLipid	double	Total lipid content, measured in grams (g)	No
Data.Fiber	double	Amount of fiber, measured in grams (g)	No

New Variables

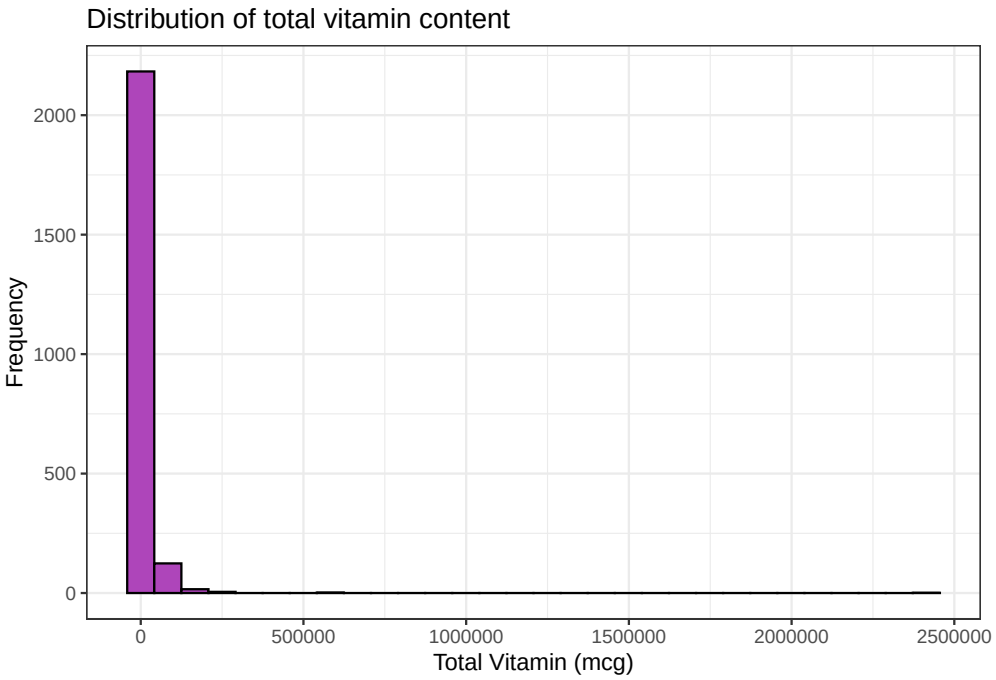
New Variable	Type	Description	Missing Values?
Category.Broad	factor	Broad category that groups the 479 categories (after removing Vitamin D) into 5 broader categories. One of Dairy/Fatty, Meat, Fruits/Vegetables/Plants, Cereals/Grains, Other. Manually grouped the 479 categories then read into R for automatic labeling.	No
Data.Vitamins.TotalVitamin	double	Total vitamin content, measured in micrograms (mcg). Calculated by adding all of the columns starting with Data.Vitamins. (Vitamin A, B12, B6, C, E, and K)	No
Data.Vitamins.VitaminB12.Group	factor	Vitamin B12, based on the value of Data.Vitamins.VitaminB12. Grouped into “Less than 1”, “Between 1 and 2”, “Between 2 and 3”, and “Greater than 3”.	No

Analysis

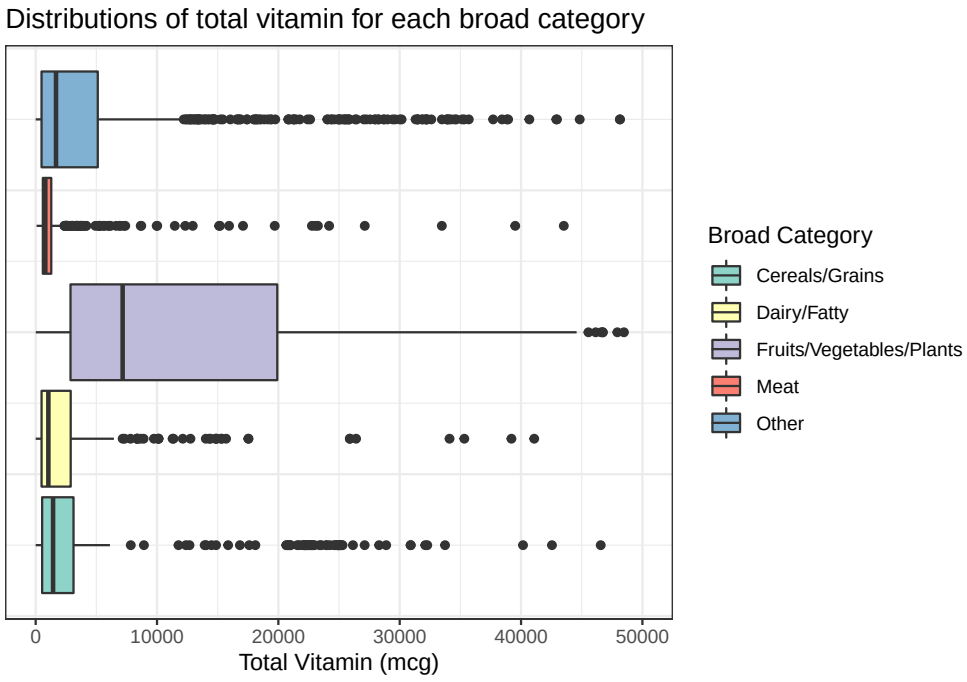
In this section, the total vitamin content, vitamins in relation to minerals, and water content are examined.

Total Vitamin Content

When the total vitamin content is looked at, there the graph is skewed to the right and unimodal. For this data set, the total vitamin content ranges from 0 mcg to 2.41402×10^6 mcg with a median of 1733.01 mcg.



The graph below shows five box plots, each for one of the 5 broad categories (some outliers were left out of the visualization). Fruits, vegetables, and other plant products tend to have the highest vitamin content (median: 9170.5 mcg) while meat products have the lowest (median: 794.21 mcg).

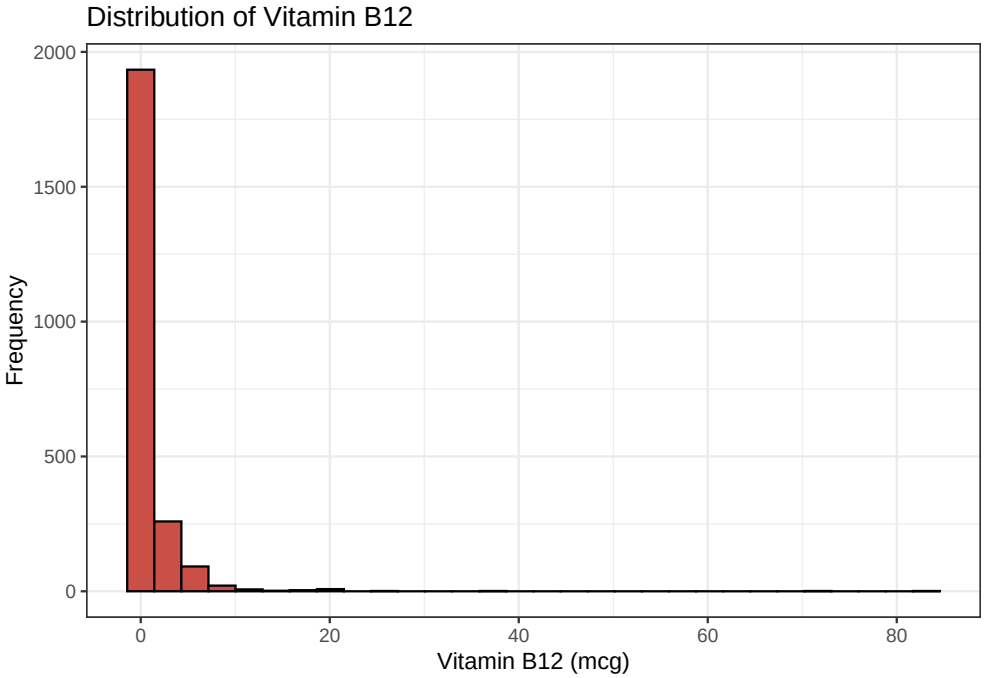


Vitamins in Relation to Minerals

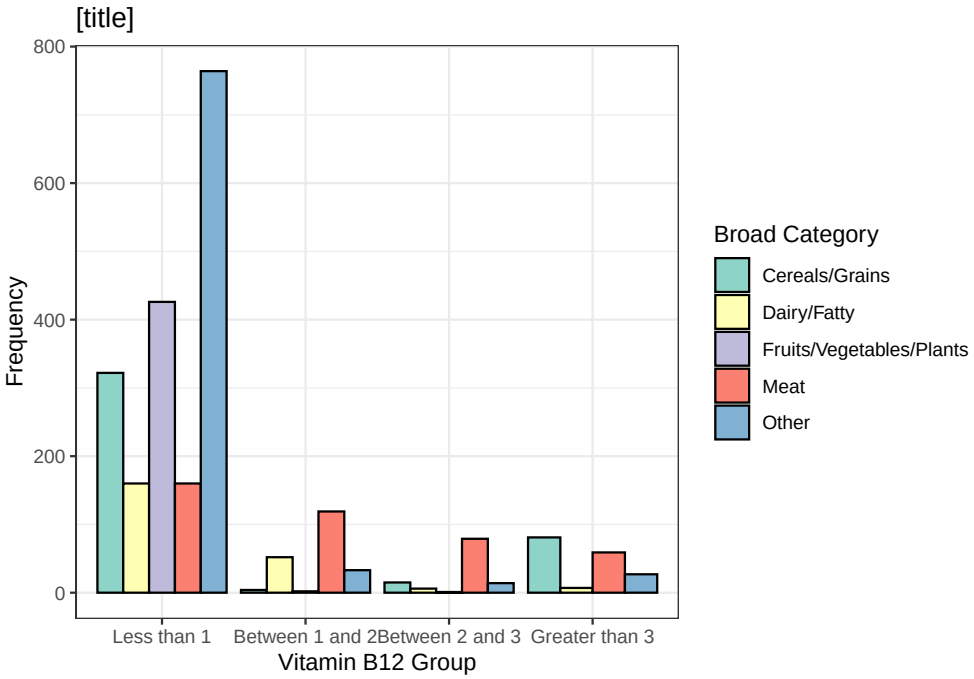
While vitamins are organic, and minerals are inorganic [3], investigating the correlations in the amounts that appear in ingredients can provide explanations to why some foods are better than others for specific tasks. Vitamin B12 and Vitamin B6 were the vitamins explored, and copper and zinc were the minerals examined.

Vitamin B12

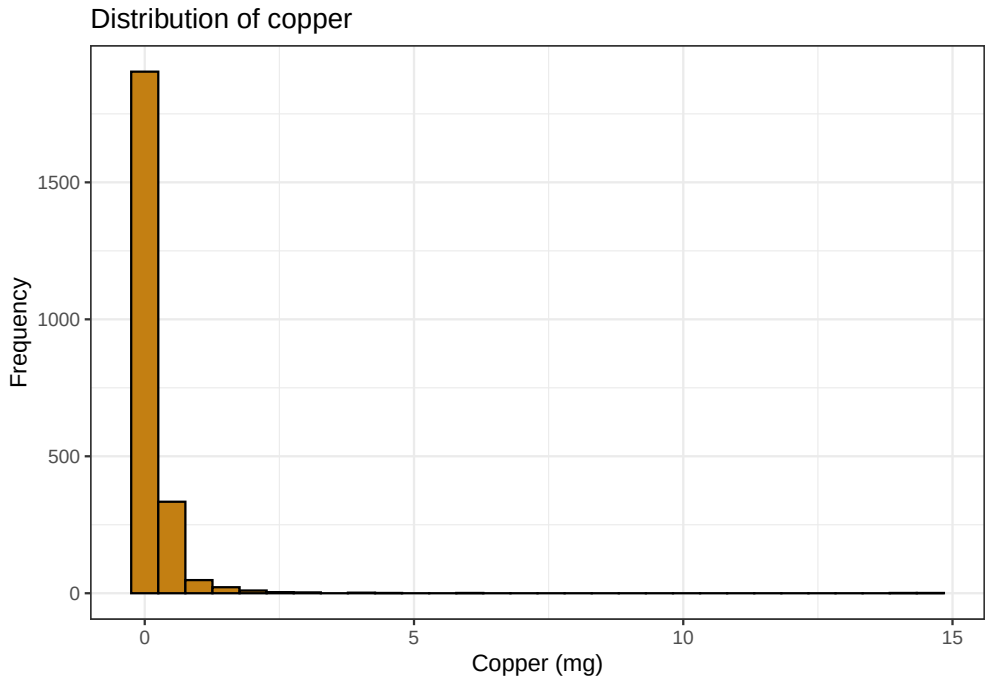
The distribution of Vitamin B12 in this data set is unimodal and skewed right, with no outliers on the lower side and many on the greater side. The values range from 0 mcg to 83.13 mcg with a median of 0.06 mcg.



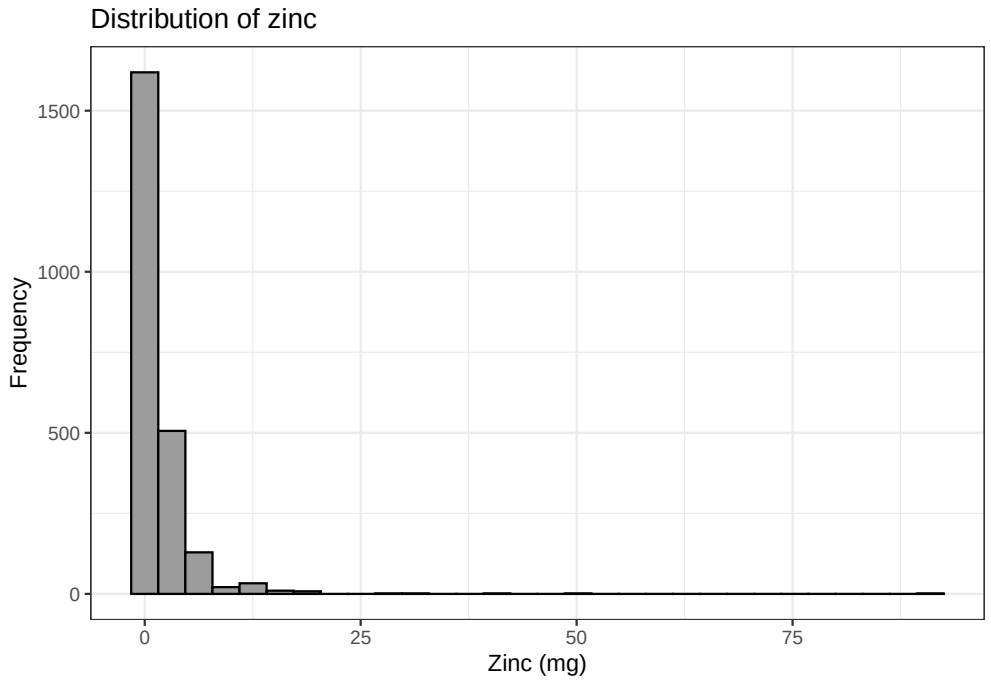
Interesting comparisions can be made when splitting this distribution. When the data is split into groups of Vitamin B12 and displayed by broad category, it can be seen that plants and miscellaneous ingredients are very significant in the "Less than 1" section but almost nonexistent in the others. This reveals the possibility that dairy and meat products have more Vitamin B12 than plant products.



The distribution of copper looks like that of Vitamin B12—unimodal and skewed right, with many outliers on the greater side. Its values range from 0 mg to 14.588 mg with a median of 0.088 mg.

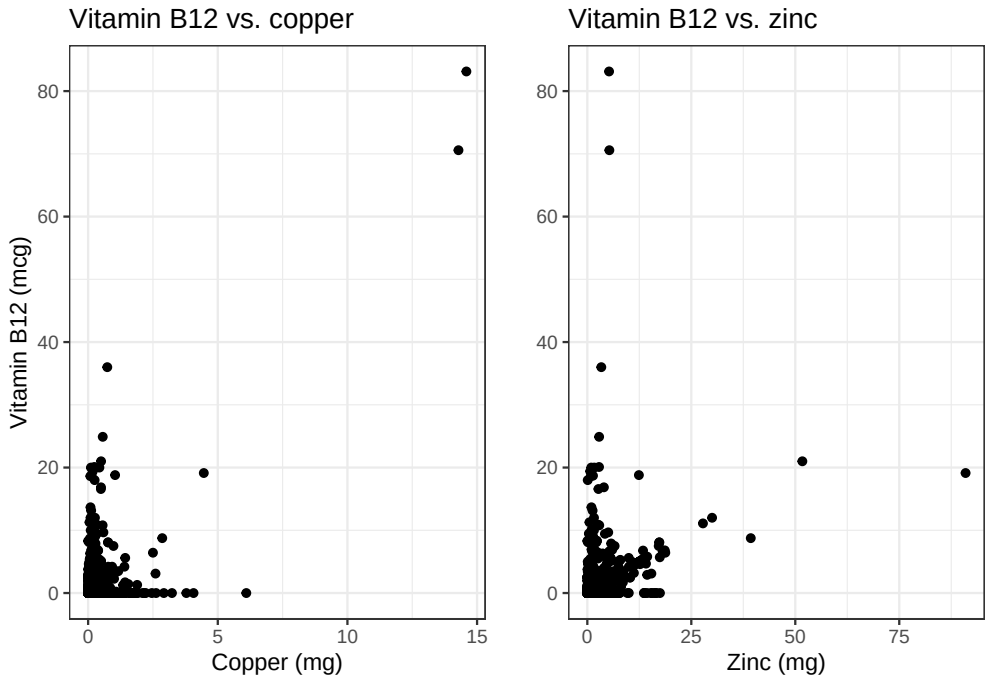


The data for zinc is similar to that of Vitamin B12 and copper in that it is unimodal and skewed right, again with many outliers on the right. The values range from 0 mg to 90.95 mg with a median of 0.7 mg.



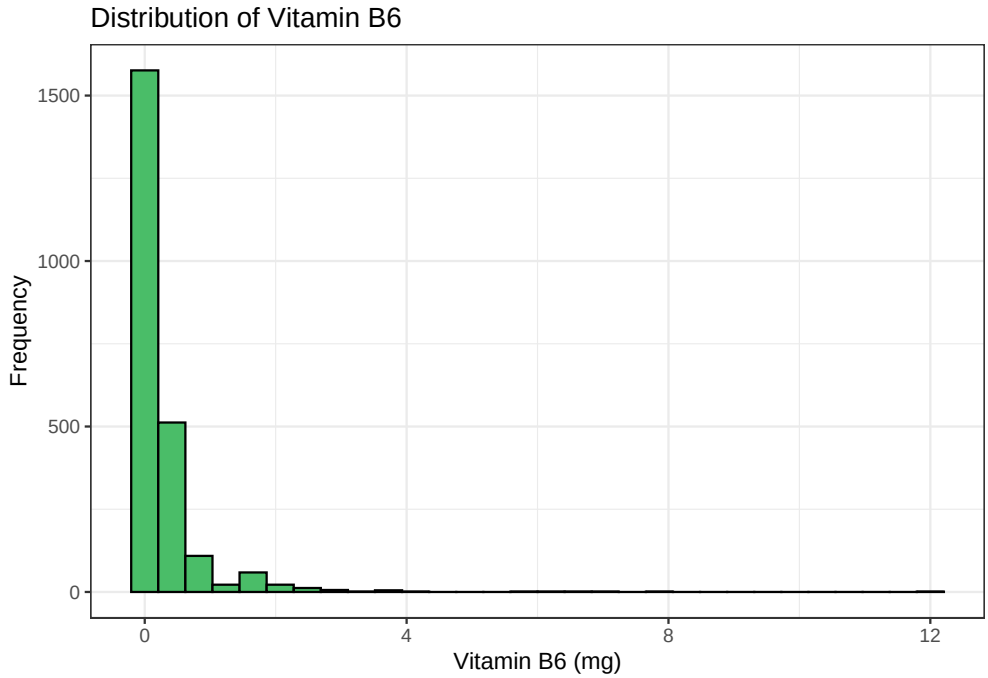
Vitamin B12 was found to have a strong, positive, correlation with copper (0.60273) and a weak, positive correlation with zinc (0.34739).

Note: the relationships may be difficult to see in the scatter plots below.

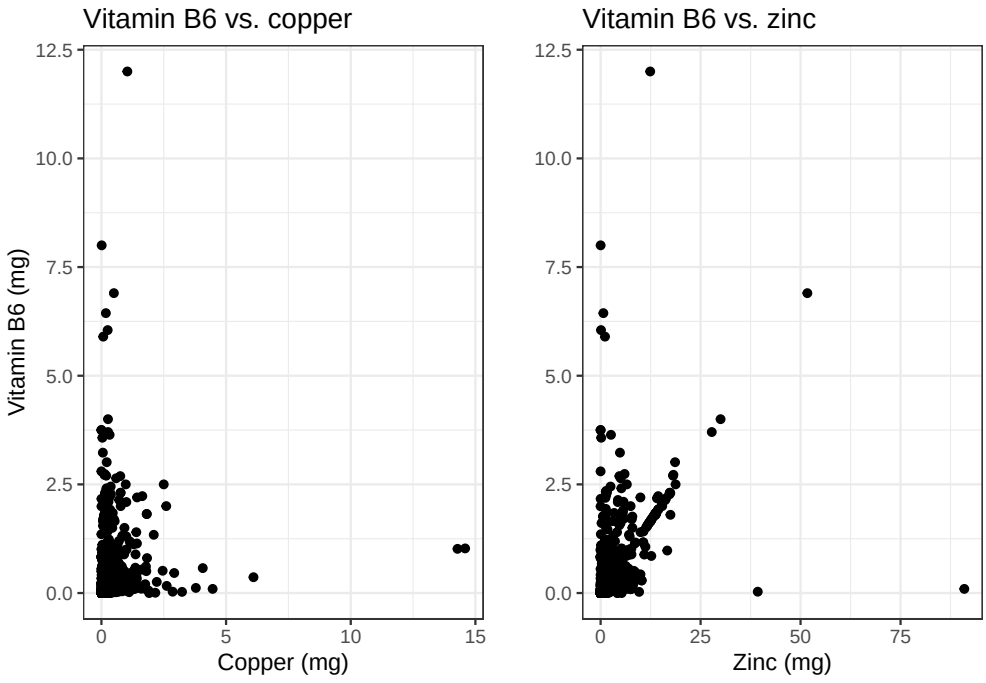


Vitamin B6

Like Vitamin B12, Vitamin B6 has a distribution that is unimodal and skewed right, with many outliers on the right side. This data ranges from 0 mg to 12 mg, and its median is 0.1 mg.



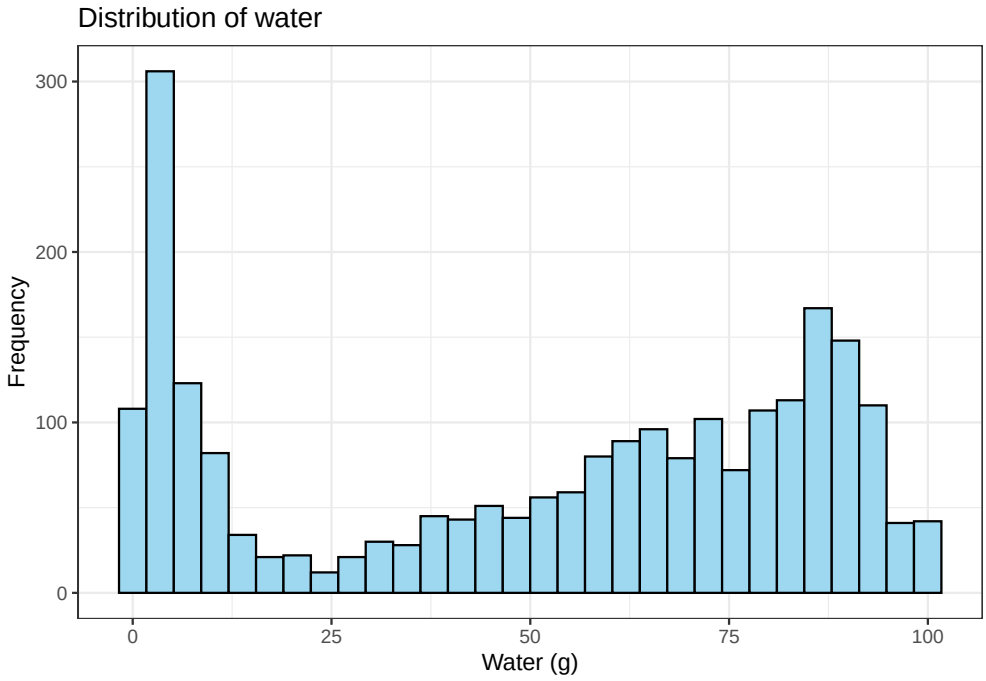
Unlike Vitamin B12, Vitamin B6 has a very weak, positive correlation with copper (0.14172) and a moderate, positive correlation with zinc (0.47162).



Water Content

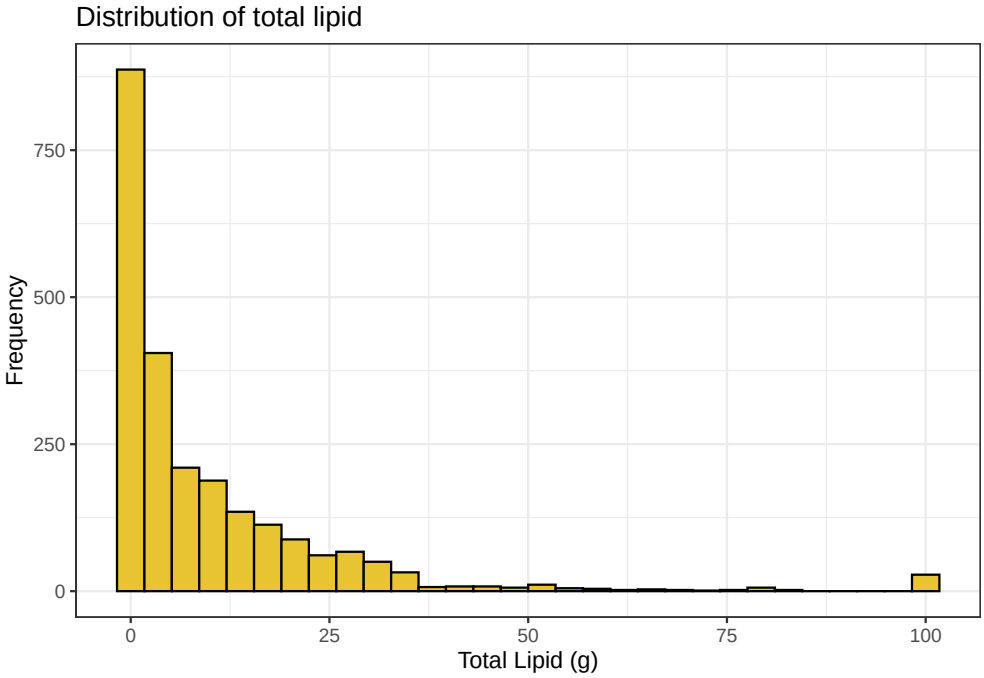
Water content is another significant factor to consider as food types will have varying levels of water. The human body is also made up of about 60% water [4], so finding what ingredients have high water can be useful in replenishing thirst in the form of food.

The distribution of water looks different than many of the other variables of interest because it is bimodal. It ranges from 0 g to 99.98 g and has a median of 60.5 g.

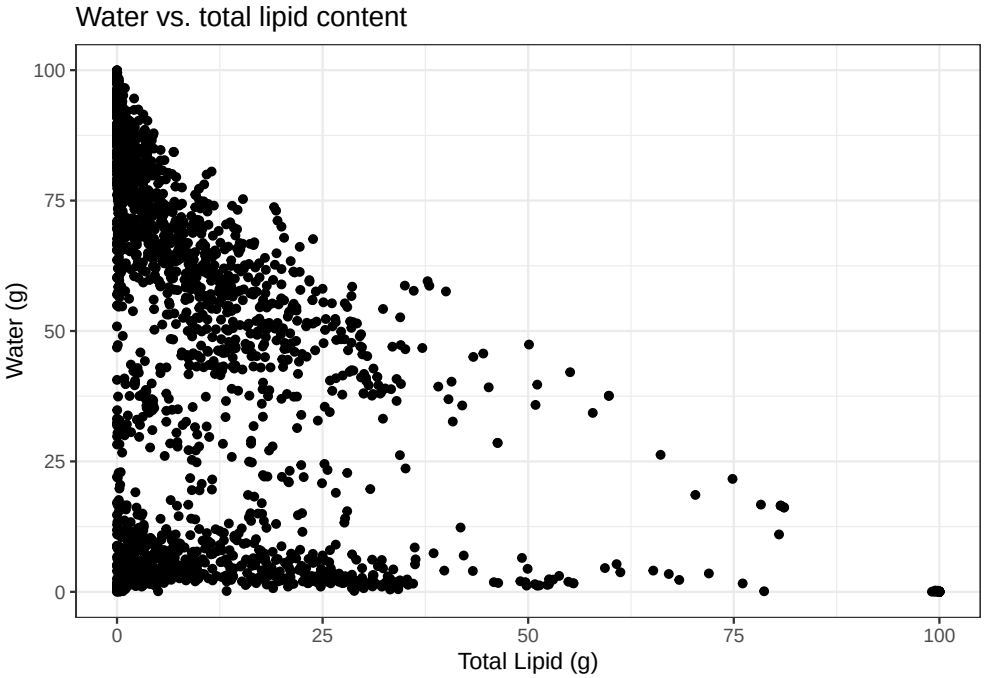


Water vs. Lipids

Looking at the fat content of an ingredient could give insight to its water content. The distribution of total lipid content, like most of the other variables, is unimodal and skewed right, but it is less skewed than those graphs. It has a minimum of 0 g and a maximum of 100 g with a median of 3.8 g.

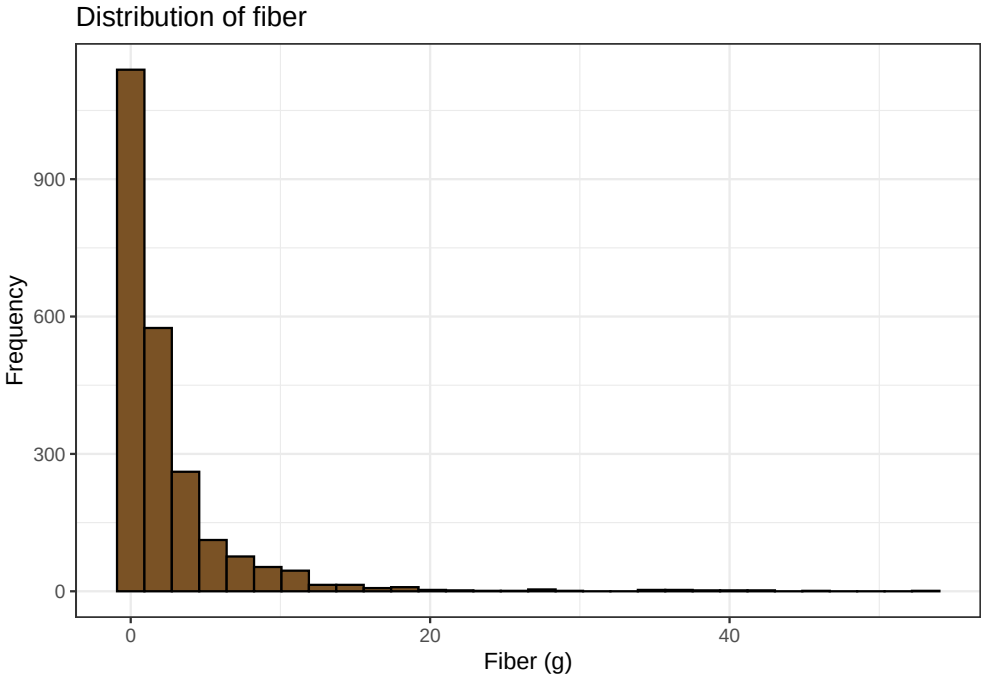


A moderate, negative correlation between total lipid content and water was found (-0.45999). As fat content increases, water content decreases.

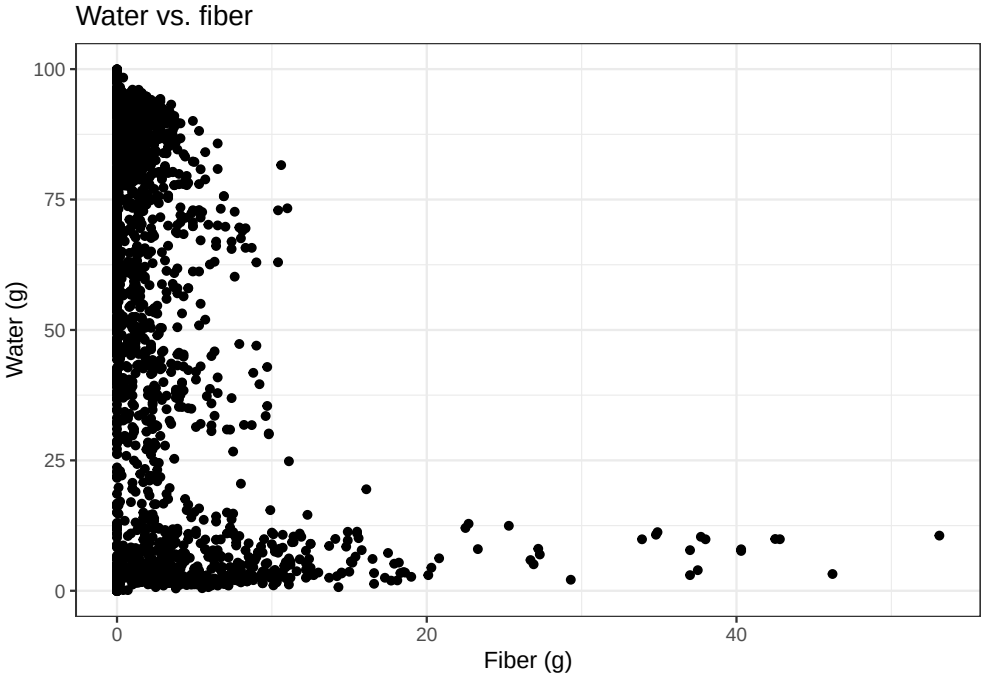


Water vs. Fiber

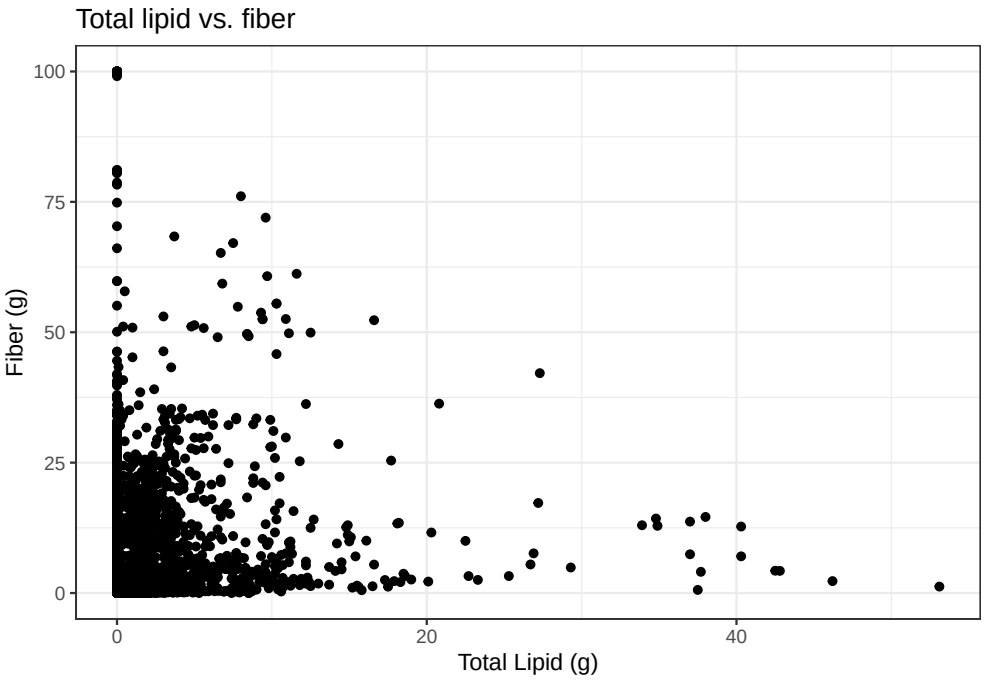
Fiber is the other predictor that was investigated in relation to the water content. Its distribution is similar to that of total lipid content, but it ranges from 0 g to 53.1 g and has a median of 1 g.



Fiber was found to have a moderate, negative correlation with water content (-0.39566). As fiber increases, water decreases.



Since the correlation of water and total fat are in the same direction as and similar magnitude to that of water and fiber, total fat as a function of fiber was also analyzed. However, no correlation was found (0.00372).



Debugging

Conclusion

References

[1] A. C. Bart, D. Kafura, C. A. Shaffer, J. Tibau, L. Gusukama, and E. Tilevich, "CORGIS," *CORGIS Datasets Project*. [Online]. Available: https://corgis-edu.github.io/corgis/

[2] R. Whitcomb, J. Min Choi, and B. Guan, "Ingredients CSV file," *CORGIS Datasets Project*. [Online]. Available: https://think.cs.vt.edu/corgis/csv/ingredients/

[3] "Vitamins and minerals," *The Nutrition Source*, Sep. 2012. [Online]. Available: https://www.hsph.harvard.edu/nutritionsource/vitamins/

[4] "The water in you: Water and the human body," *U.S. Geological Survey*, May 2019. [Online]. Available: https://www.usgs.gov/special-topics/water-science-school/science/water-you-water-and-human-body