Ingredient Analysis

Victor Nguyen

November 7, 2022 - November 18, 2022

Contents

Purpose

Creating diets can be difficult as many people require diets specialized for certain tasks. These may include specific nutrients, which come from particular food ingredients. The purpose of this analysis is to find relationships between certain nutrients and properties of common types of food. 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

The `read_csv` function from tidyverse was used to load the data into R. 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. See the Debugging section for challenges faced during this process.

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

Three new variables were created to better analyze the data. Since there are 479 different categories, they can be difficult to visualize and learn from. The `Category.Broad` column aims to handle this by grouping these categories into 5 broader food categories. The process of grouping the categories was done manually, but the actual assigning of the labels was done using a loop with an if statement. The calculation of the `Data.Vitamins.TotalVitamin` data consisted of one statement, adding the vitamin columns together and changing units to micrograms if needed. Finally, the Vitamin B12 data was split into groups using a loop that calls a function which used a few if statements to check which group a value belonged to and added the correct group to the column.

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

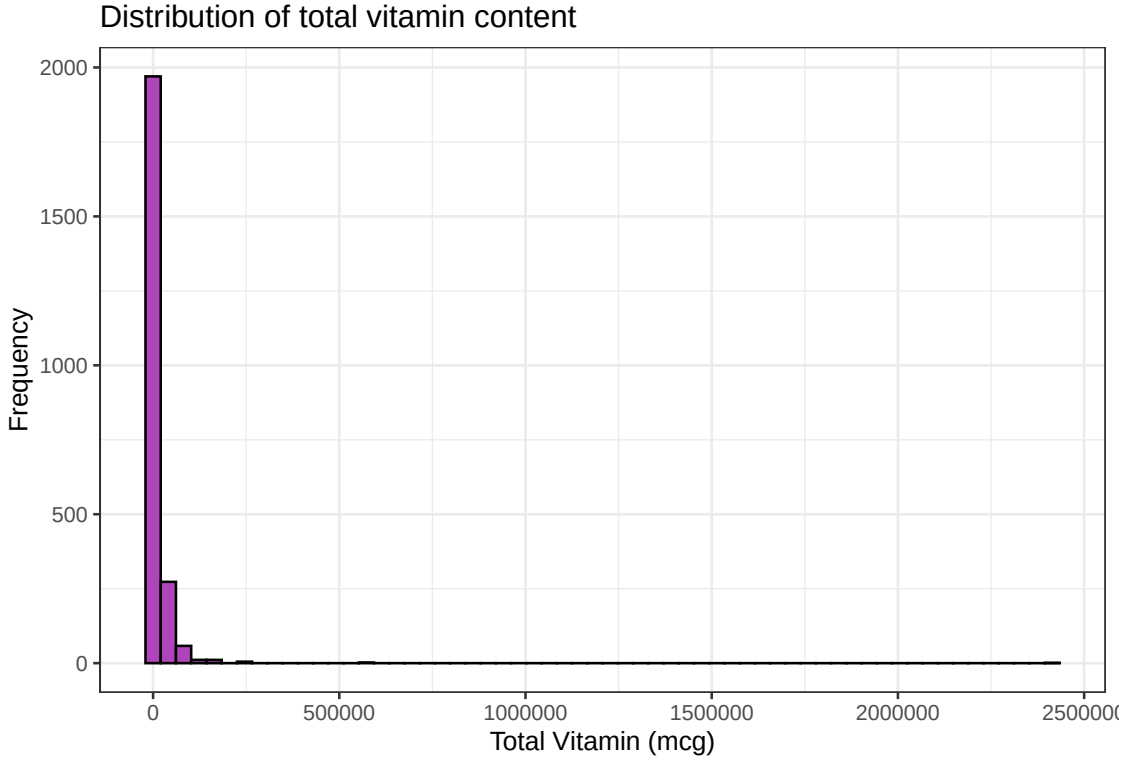
New Variable	Type	Description	Missing Values?
Data.Vitamins.VitaminB12.Group	factor	Group of Vitamin B12, based on the value of Data.Vitamins.VitaminB12. Grouped into “Less than 1”, “Between 1 and 2”, “Between 2 and 3”, and “Greater than 3”.	No

Analysis

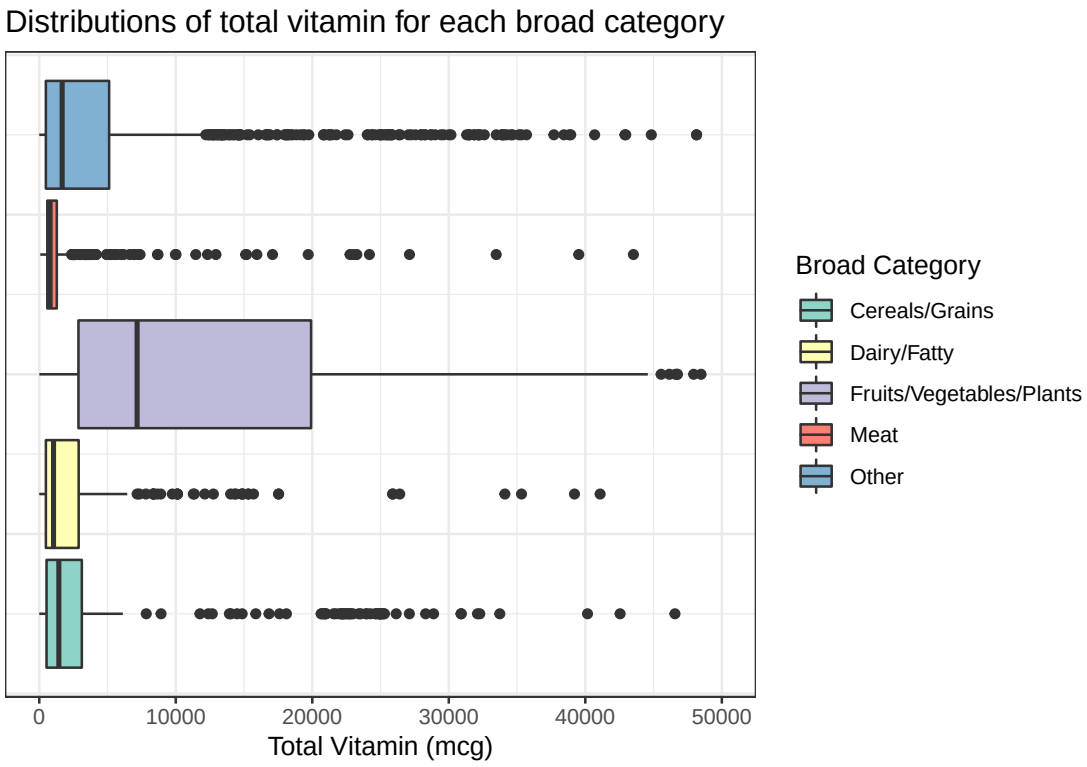
In this section, the total vitamin content, vitamins in relation to minerals, and water content are examined. Histograms and scatter plots are made using two functions (one for histograms and one for scatter plots) because the code is similar across different variables. The function for scatter plots uses an if statement to check for the vertical axis limits.

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. There are many more ingredients (in the data) with low vitamin content than with high vitamin content.



The graph below shows five box plots, each for one of the 5 broad categories (some outliers were excluded from the visualization). Fruits, vegetables, and other plant products tend to have the highest vitamin content (median: 9170.5 mcg) while meat products tend to have the lowest (median: 794.21 mcg). Therefore, fruits and vegetables are necessary for a high vitamin diet.



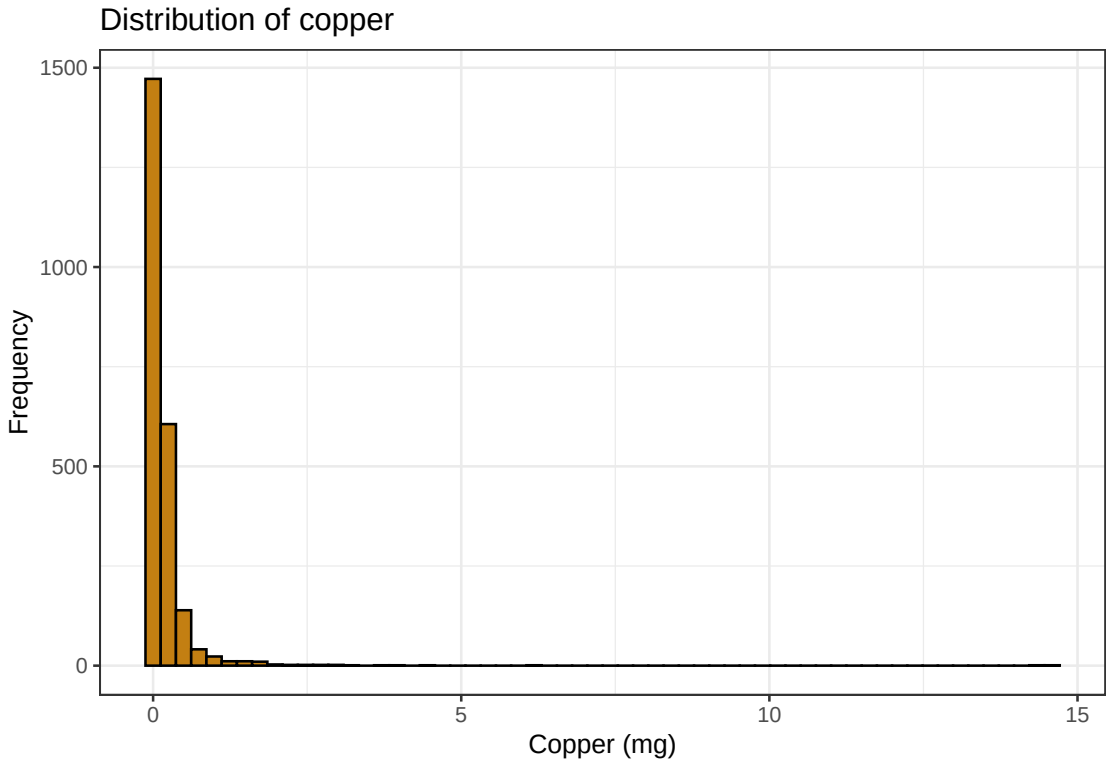
Vitamins in Relation to Minerals

While vitamins are organic, and minerals are inorganic [3], investigating the correlations in the amounts that appear in ingredients can provide explanations as 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.

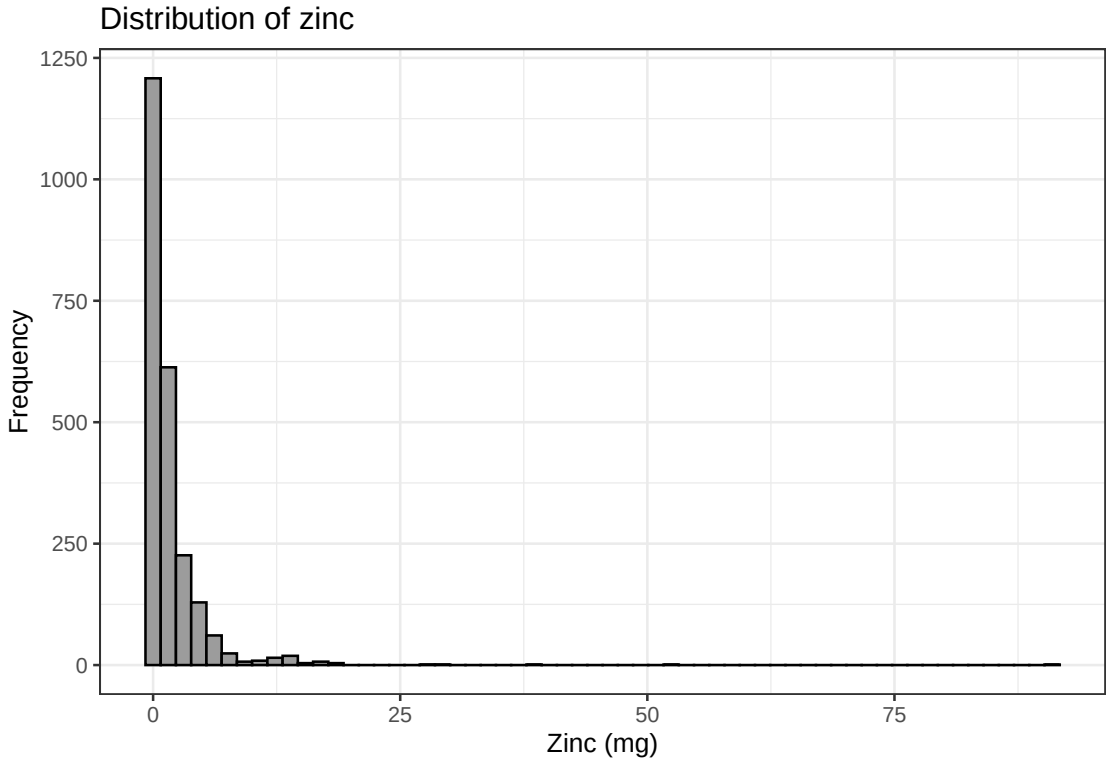
Minerals

The minerals examined in this analysis are copper and zinc.

The distribution of copper is 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. Most of the ingredients (in the data) have low copper compared to the large range of values.

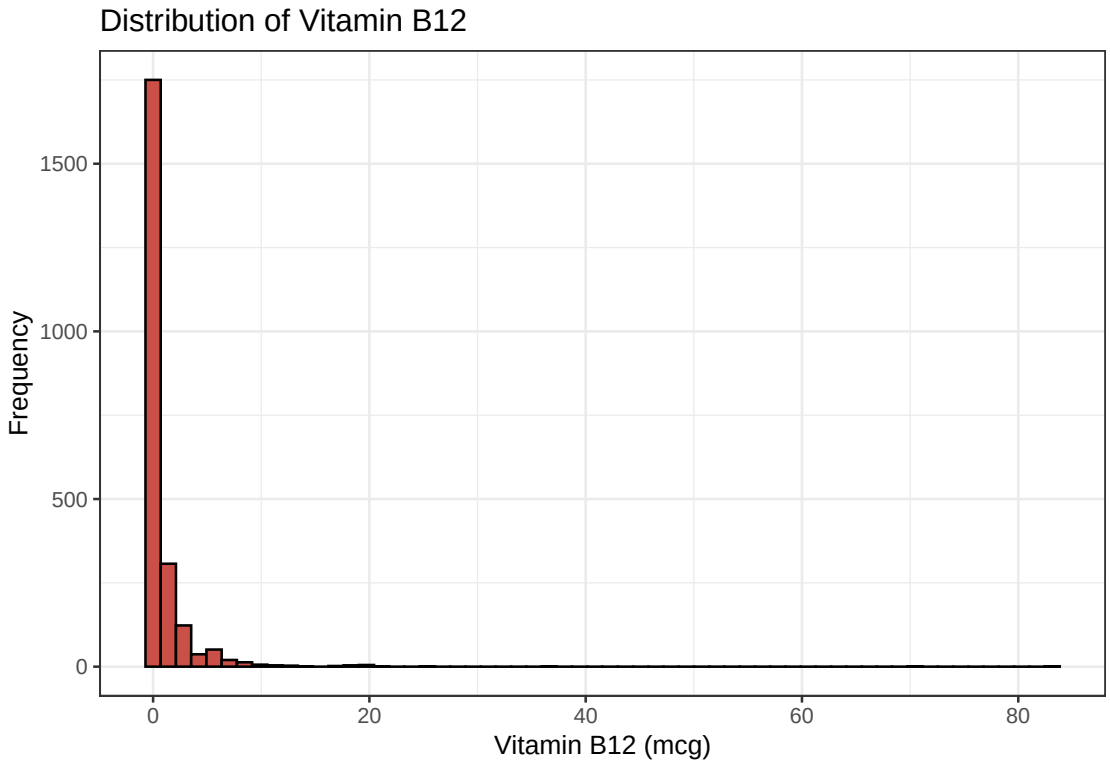


The data for zinc is similar to that of copper in that it is unimodal and skewed right, again with many outliers on the right. The values range from 0 mg to 90.95 mg with a median of 0.7 mg. Like many of the other variables seen, many ingredients (in the data) have low zinc content while few have high zinc content.

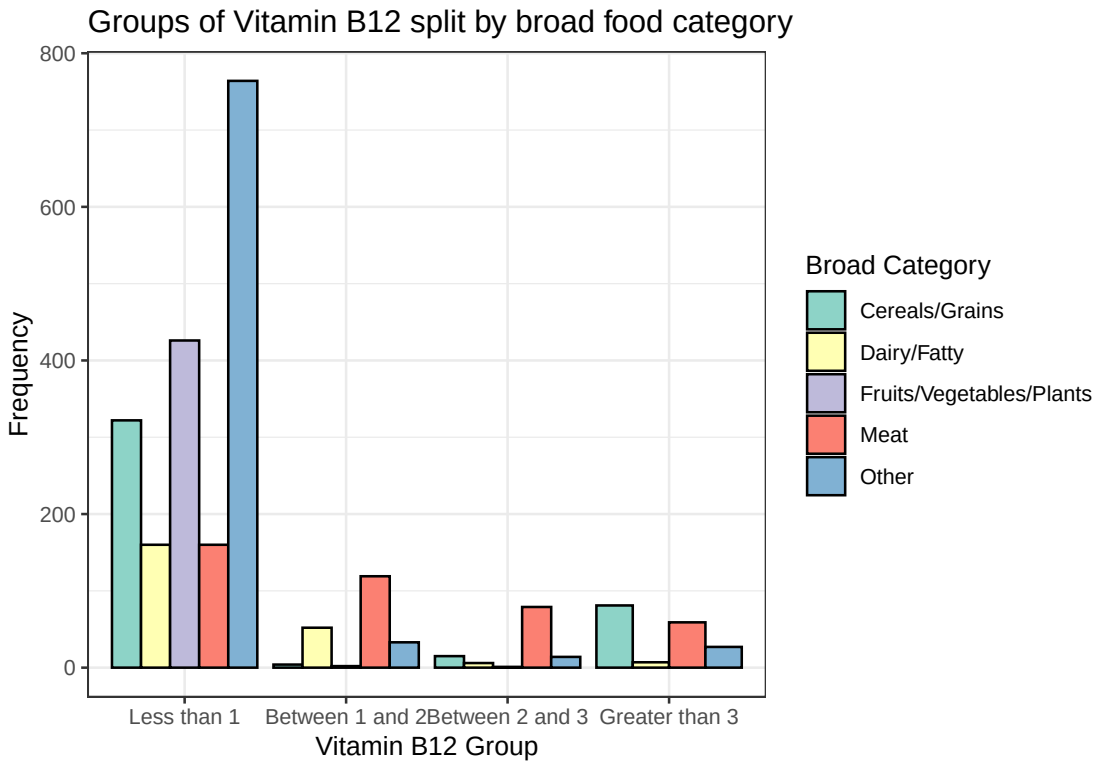


Vitamin B12

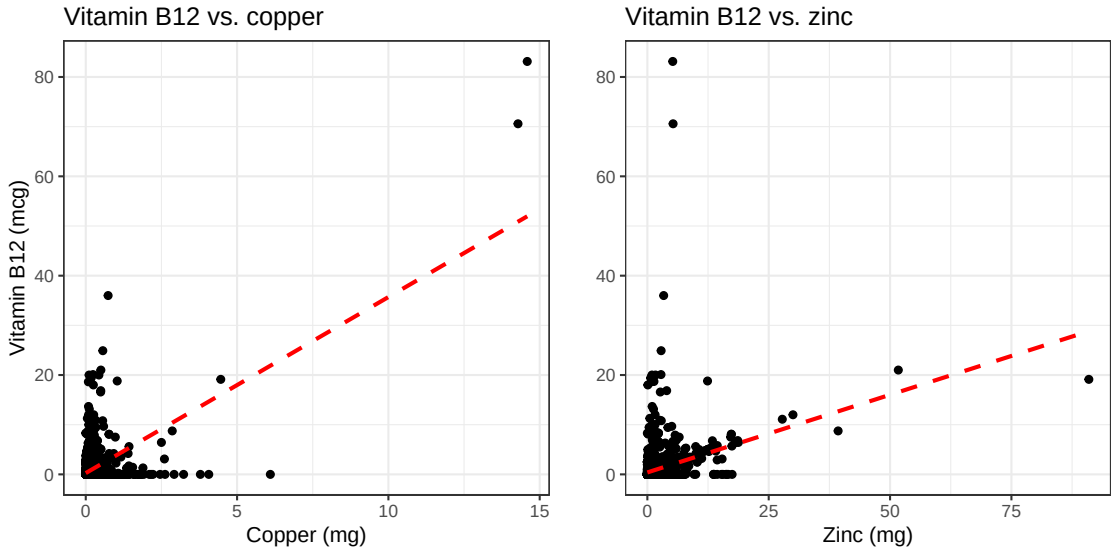
The distribution of Vitamin B12 in this data set has the same general shape as those of copper and zinc—unimodal and skewed right, with no outliers on the lower side and many on the greater side. The values range from 0 mcg to 83.13 mcg with a median of 0.06 mcg. Many food items (in the data) have low amounts of Vitamin B12 compared to the few that have high concentrations.



Interesting comparisons 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 while meat ingredients have a prevalence in all groups. This reveals the possibility that dairy and meat products have more Vitamin B12 than plant products. Animal products are efficient options for having plentiful sources of Vitamin B12 [4].

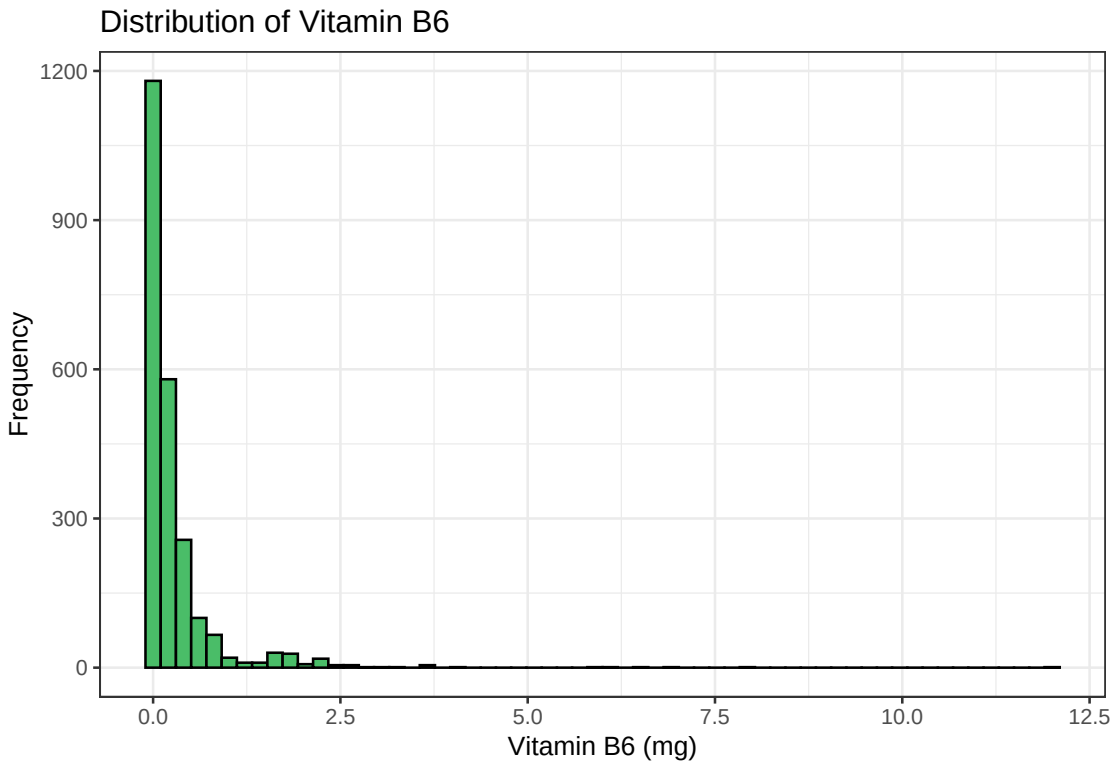


Vitamin B12 was found to have a strong, positive, correlation with copper (0.60273) and a weak, positive correlation with zinc (0.34739). This means that foods with high amounts of copper or zinc tend to have high amounts of Vitamin B12, and foods with low quantities of those minerals often have low quantities of Vitamin B12.

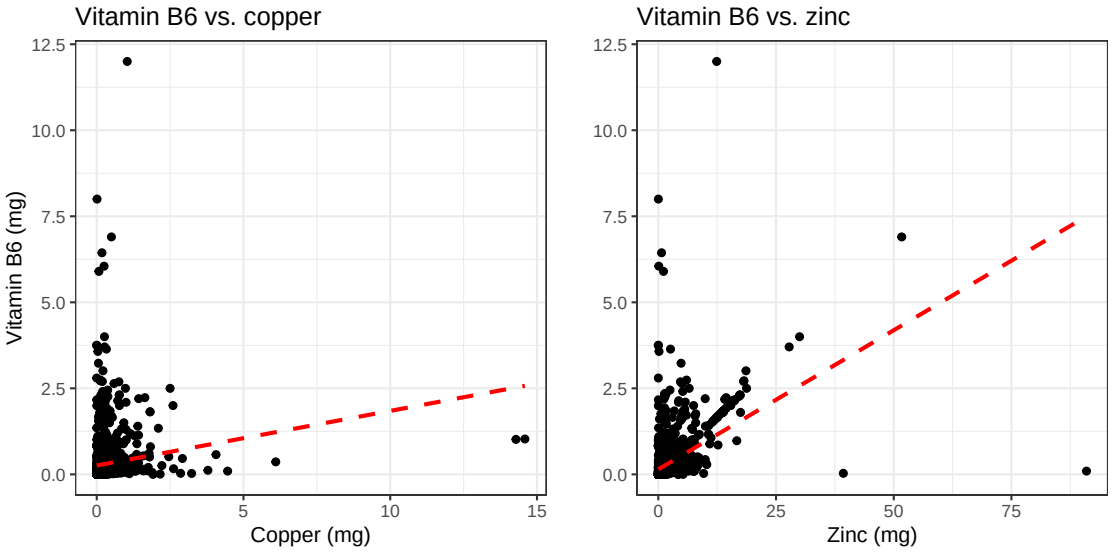


Vitamin B6

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. Most of the data has "low" Vitamin B6, but there are some outliers that greatly increase the range.



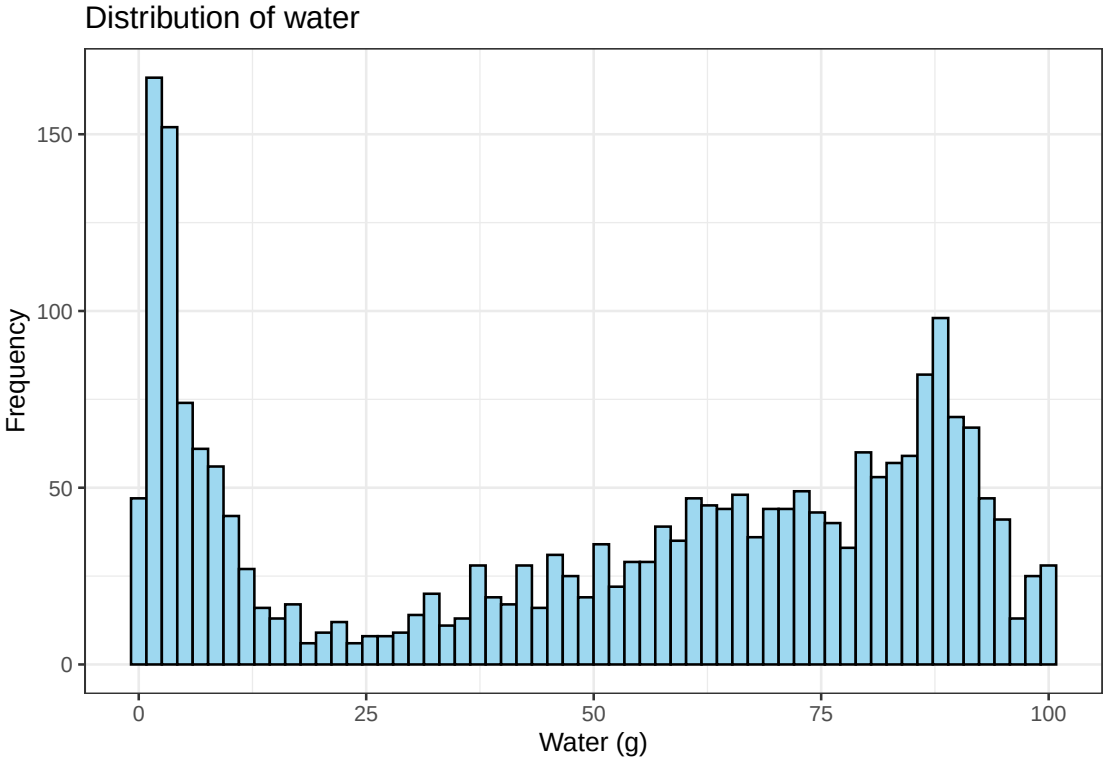
Unlike Vitamin B12, Vitamin B6 has a very weak, positive correlation with copper (0.14172) and a moderate, positive correlation with zinc (0.47162). This implies that copper content in an ingredient does not generally affect the amount of Vitamin B6. However, high concentrations of zinc often are associated with high levels of Vitamin B6.



Water Content

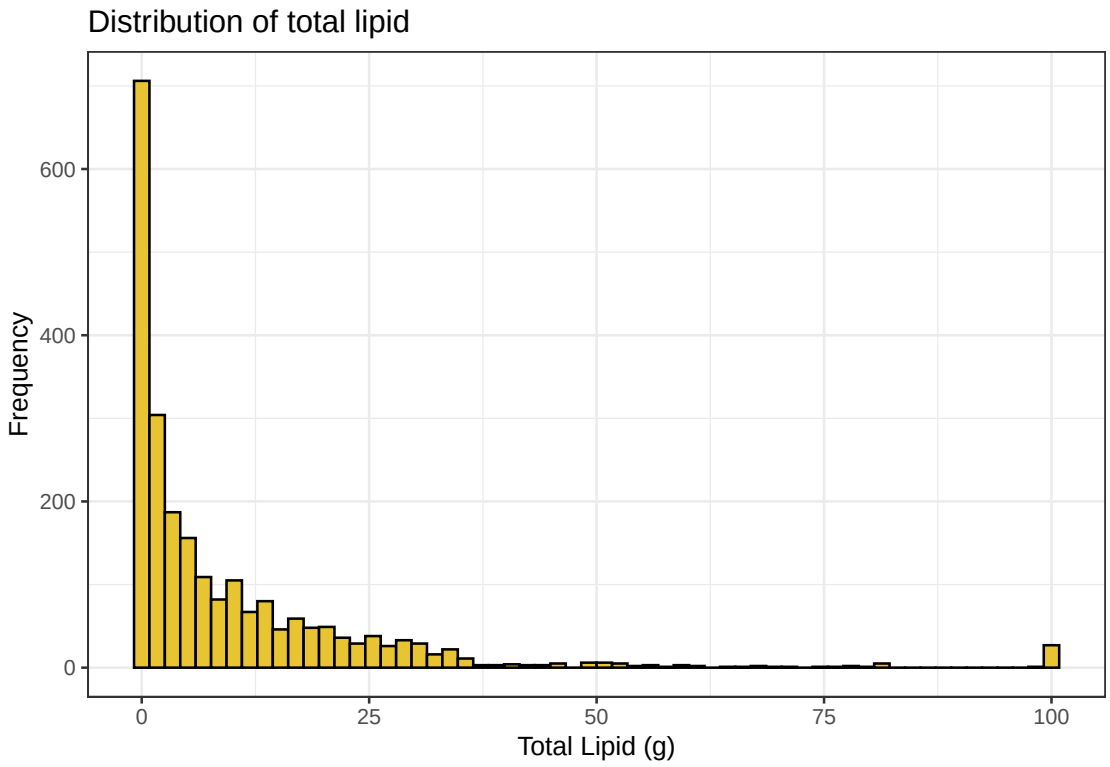
Water content is another significant factor to consider as food types will have varying levels of water. The human body is also made up of about 60% water [5], so finding what ingredients have high water can be useful in replenishing thirst in the form of food.

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. These two peaks indicate that many ingredients (in the data) have very low and very high water composition, but fewer have amounts in the middle of the distribution.

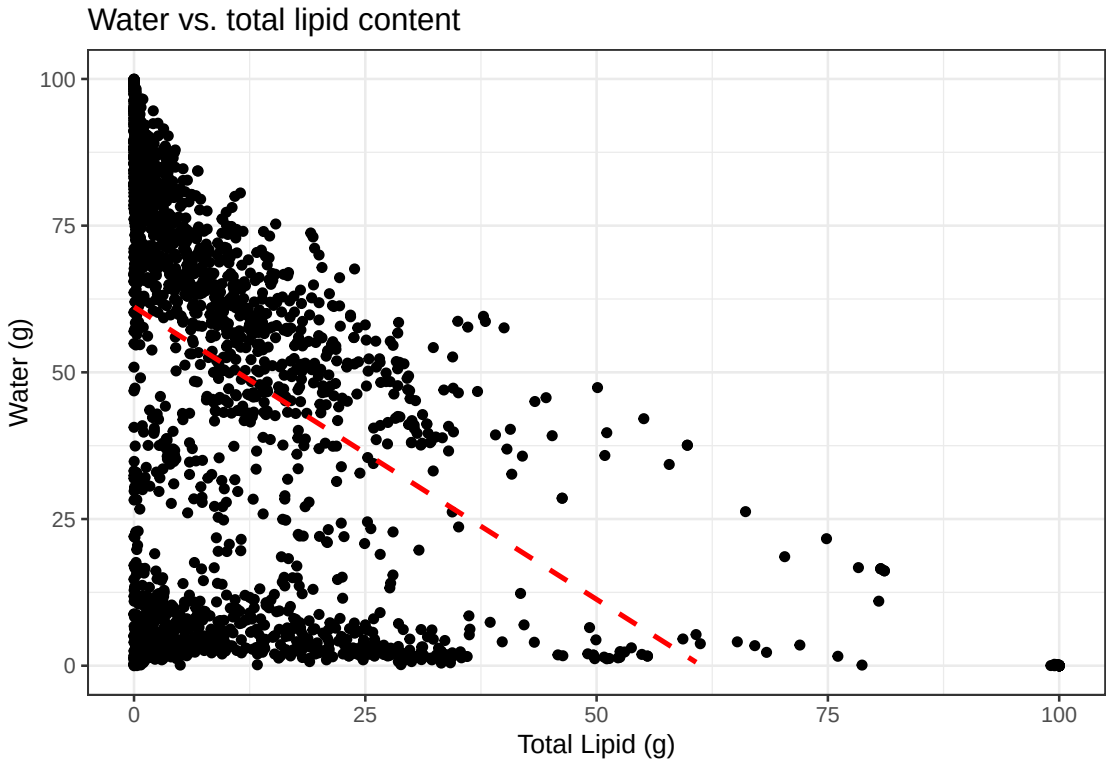


Water vs. Lipids

Looking at the fat content of an ingredient could give insight into 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. More ingredients (in the data) have low total lipid amounts than high amounts.

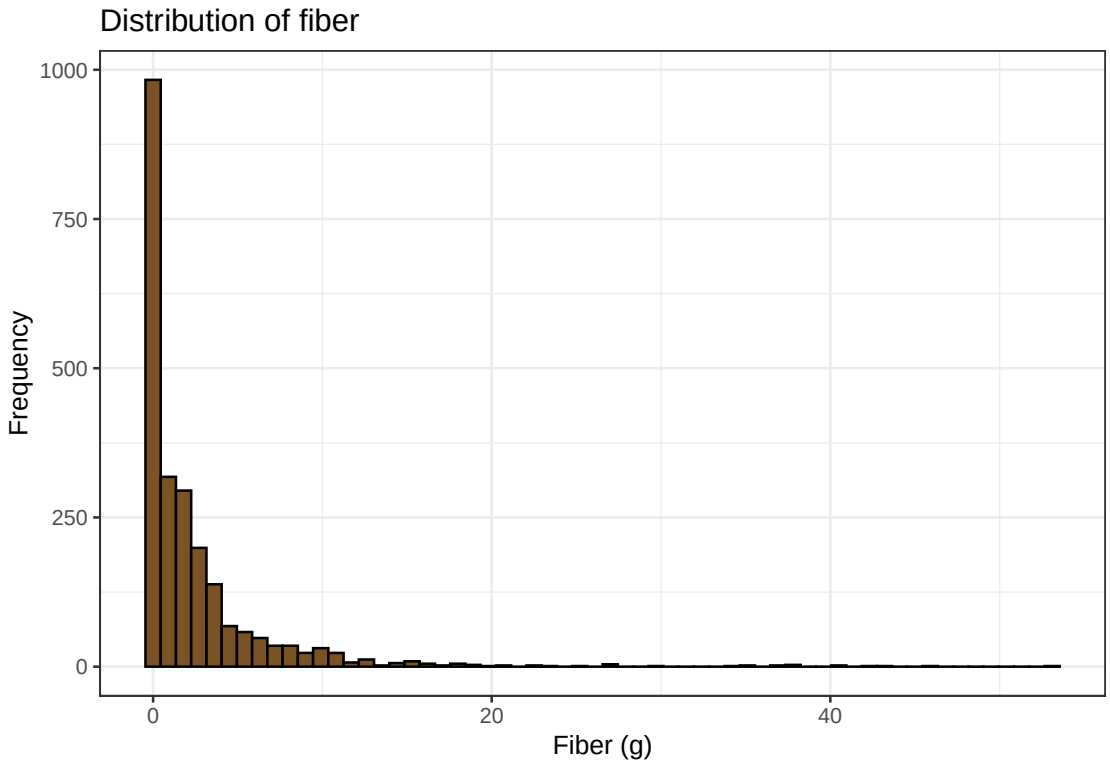


A moderate, negative correlation between total lipid content and water was found (-0.45999). Foods with high quantities of total lipids generally have lower water content than those with low quantities of total lipids.

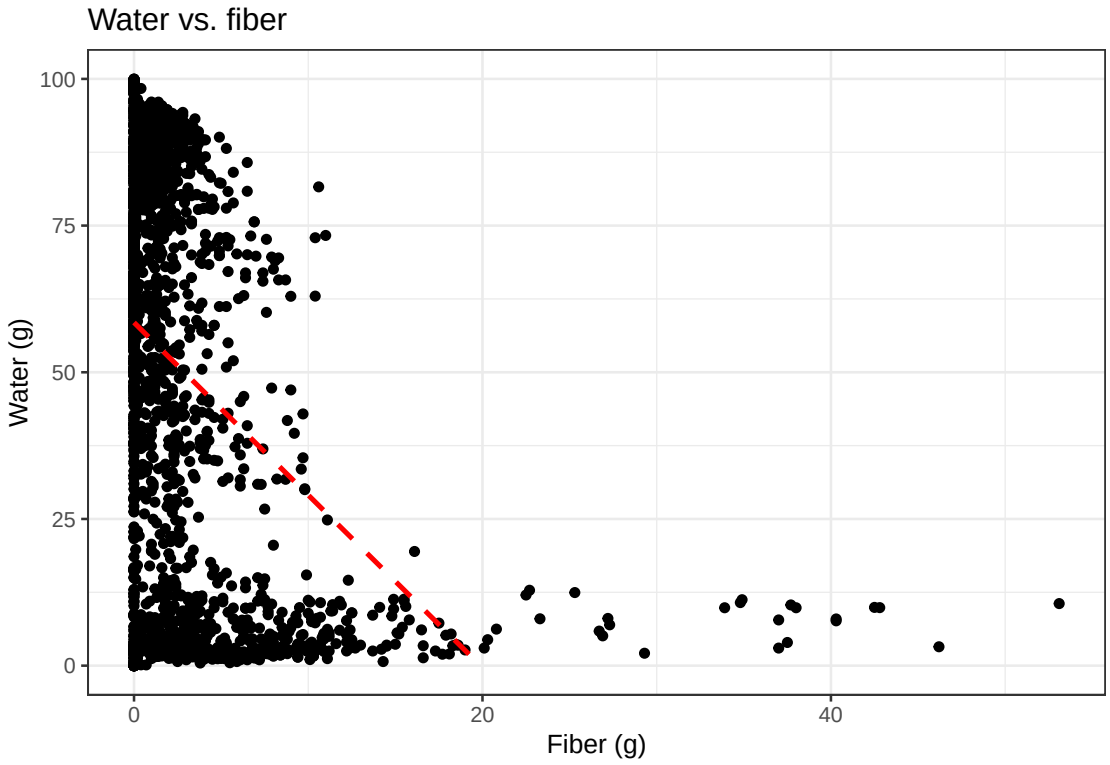


Water vs. Fiber

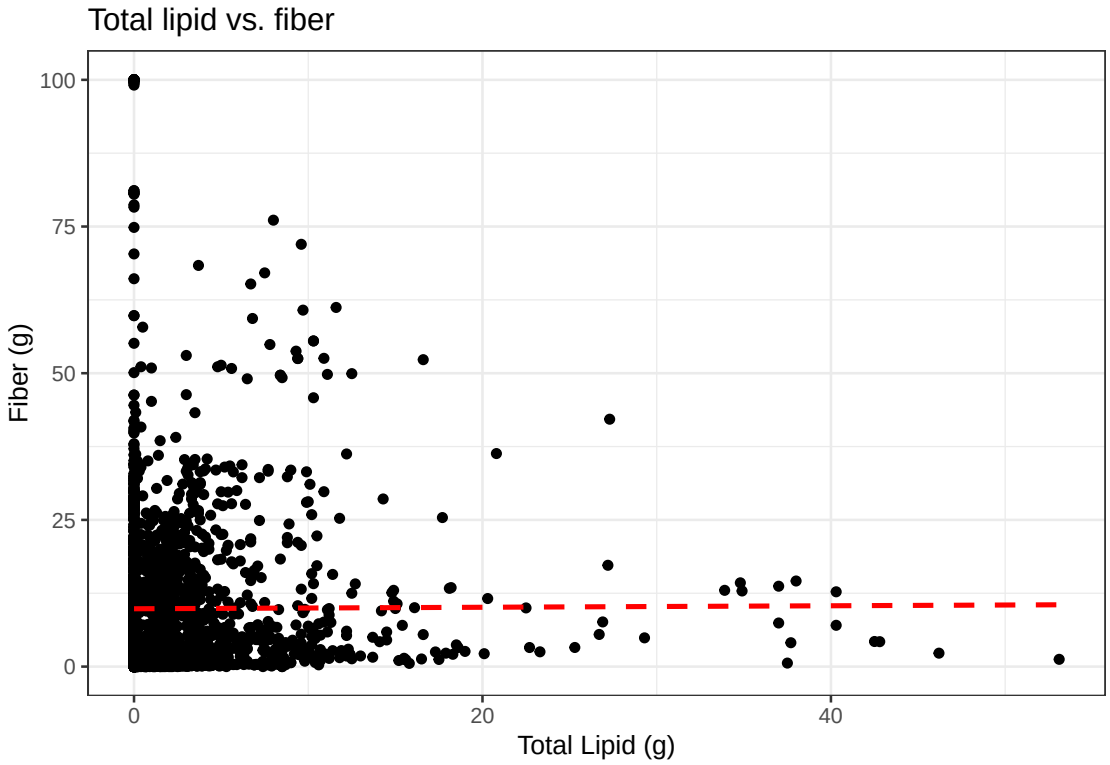
Fiber is the other predictor that was investigated in relation to the water content. Its distribution is similar to that of total lipid content, but it ranges from 0 g to 53.1 g and has a median of 1 g. The data is concentrated on the lower side of the distribution, so there are more ingredients (in the data) with low fiber than with high fiber.



Fiber was found to have a moderate, negative correlation with water content (-0.39566). This reveals that ingredients abundant with fiber tend to have low quantities of water whilst those that don't have much fiber usually contain high water amounts.



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). Hence, the total lipid content does not influence the fiber in a food item, and vice versa.



Debugging

The debugging for this project was nothing complex. When an error or incorrect output was encountered, the code was reviewed. If the issue still persisted, it would be searched on Google or fixed using print statements. Processing the data created many more issues than visualizing it did, especially when creating the new columns. Specifically, when the code-writing process first started, the working directory was not set to the correct location, which needed to be resolved in order to read the data into R. Additionally, some labels were missed when manually categorizing `Category` into `Category.Broad`, and looking through the 2000+ ingredients would not be efficient. Thus, an if statement was placed within the loop that labeled the data to collect any missing values. Then, these missing labels were added to the CSV file storing the others. Issues with graphs were solved by using Google, which led to Stack Overflow.

Conclusion

Much of the effort was concentrated in data processing, which is one of the most important steps in the analytic process. However, the visualizations give an intuitive interpretation of the numbers, so this is also crucial. There were many outliers in this ingredients data set, but relationships between variables were still able to be found. Firstly, fruits and vegetables tend to have higher total vitamin content than meat and dairy products. Secondly, high quantities of copper in ingredients are associated with high quantities of Vitamin B12, and high levels of zinc are related to high levels of both Vitamin B12 and Vitamin B6. Lastly, high total lipid and fiber content are both correlated with low water content (independently from each other). This information can be used to efficiently construct diets for various purposes because knowing which types of ingredients one's targeted nutrients are prevalent in will allow accurate predictions of results. Future research could hopefully prove causations for the correlations discovered in this analysis or find more complex, multivariate relationships between nutrients.

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] R. Obeid, S. G. Heil, M. M. A. Verhoeven, E. G. H. M. van den Heuvel, L. C. P. G. M. de Groot, and S. J. P. M. Eussen, "Vitamin B12 Intake From Animal Foods, Biomarkers, and Health Aspects," *Frontiers in Nutrition*, vol. 6, 2019, doi: 10.3389/fnut.2019.00093. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnut.2019.00093

[5] "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