**Analysis of Cereal Nutrition Data**

**Introduction:**

In this report, we conducted a comparative analysis of three cereal companies (Company G, Company K, and Company Q) based on various nutritional variables. This report presents an analysis of nutritional data for 43 types of cereals produced by three companies: General Mills, Kellogg's, and Quaker. The dataset includes eight numerical variables: Calories, Protein, Fat, Sodium, Fiber, Carbohydrates, Sugar, and Potassium. The goal of this analysis is to answer several questions posed by the food-safety and nutrition board, including identifying associations between variables, visualizing the data, determining underlying groups, characterizing dissimilarities, comparing the companies, and exploring other interesting aspects of the dataset.

**Methodology:**

In order to conduct the analysis, we utilized the R programming language with a number of packages, including tidyverse, dplyr, ggplot2, and others. Importing and cleansing the dataset to ensure correct data types and structure. We conducted exploratory data analysis, correlation analysis, data visualization, principal component analysis (PCA), multidimensional scaling (MDS), and a comparison of company means.

**Results and Analysis:**

1. **Associations between Variables:**

**A picture containing screenshot, square, text, pattern

Description automatically generated**

We computed the correlation matrix and identified several notable associations between the variables:

Calories and Fat: There is a positive correlation, suggesting that cereals with higher fat content tend to have higher calorie counts.

Protein and Fiber: A positive correlation indicates that cereals with higher protein content also tend to have higher fiber content.

Carbohydrates and Sodium: A positive correlation suggests that cereals with higher carbohydrate content may also contain higher levels of sodium.

Fiber and Potassium: A strong positive correlation indicates that cereals high in fiber content tend to have higher potassium content.

Calories and Sugar: A positive correlation implies that cereals with higher sugar content tend to have higher calorie counts.

Fiber and Carbohydrates: A strong negative correlation suggests that cereals high in fiber tend to have lower carbohydrate content.

Carbohydrates and Fat: A strong negative correlation indicates that cereals high in carbohydrate content tend to have lower fat content.

Protein and Sugar: A strong negative correlation suggests that cereals high in protein content tend to have lower sugar content.

Carbohydrates are positively correlated to calories and sodium and negatively correlated to every other nutritional characteristic.

Calories are positively correlated to every nutritional characteristic.

Protein, Fiber is negatively correlated to sugar and carbohydrates and positively to every other variable.

Fat and potassium are positively correlated to every nutritional characteristic and negatively with only Carbohydrates.

1. **Graphical Representation of Raw Data:**

**A picture containing text, screenshot, diagram, colorfulness

Description automatically generated** A picture containing text, screenshot, font

Description automatically generated

To graphically represent the raw data for the 43 cereals, we created a scatterplot matrix. Each variable was plotted against each other variable, and brand names were used to color the points.

Based on clusters, it can be seen that all brands seemed to have nearly the same values of calories (except 3 brands), sodium, fiber (except one brand), carbohydrates (except one), sugar, and potassium (except one brand). Potassium and fat seemed to be varied for almost all brands.

Observing clusters of cereals with similar nutritional properties provided insight into the data set.

1. **Basic Underlying Groups:**

Using PCA, we identified the basic underlying cereal groups. The first two principal components (PC1 and PC2) accounted for a substantial amount of the variance. PC1 appeared to represent the overall nutritional value, whereas PC2 indicated the amount of calories, sodium, and sugar. By plotting the cereals in a two-dimensional space, clusters of cereals with similar characteristics were observed.

Almost all come to two clusters, except AllBran, QuakerOatmeal, MueslixCrispylend, Puffed rice, and puffed wheat.

A screenshot of a computer

Description automatically generated with low confidence

1. **Dissimilarity Plot:**

Using dissimilarity measures and MDS, we plotted the cereals on a 2-dimensional map. The two dimensions represented the similarity or dissimilarity between cereals based on their characteristics.

By plotting the cereals on a 2-dimensional map based on their dissimilarity, we can visually identify clusters or groups of cereals that have similar characteristics. Product19, RiceKrispies, CornFlakes, GoldenGrahams, Kix, Crispix, FrostedFlakes, CapNCrunch, SpecialK, HoneyGrahamOhs,TotalCornFlakes, MueslixCrispyBlend, TotalWholeGrain are likely to cluster. Whereas, other brands are not.

A picture containing text, screenshot, diagram

Description automatically generated

1. **Comparison of Companies:**

We compared the average values of the eight variables across the three companies. Significant differences were observed among the companies.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Company** | **Calorie Count** | **Protein (g)** | **Fat (g)** | **Sodium (mg)** | **Fiber (g)** | **Carbohydrates (g)** | **Sugar (g)** | **Potassium (mg)** |
| G | 110.59 | 2.35 | 1.24 | 203.53 | 1.29 | 14.59 | 8.12 | 85.00 |
| K | 111.00 | 2.60 | 0.65 | 185.50 | 2.25 | 15.25 | 7.95 | 91.75 |
| Q | 90.00 | 2.33 | 1.33 | 98.33 | 1.12 | 10.00 | 5.00 | 58.33 |

Based on the mean values for each variable by company, we can observe the following:

From these mean values, we can see that there are differences between the three companies in terms of their average values for the eight variables.

Some notable differences include:

Company K has the highest average calorie count, fiber, Carbohydrates and Potassium content among the three companies.

Company Q has the lowest average calorie count and highest average fat content.

Company G has the highest average sodium and sugar content.

These differences indicate that there are interesting variations in the nutritional composition of cereals produced by the different companies.

1. **Other Interesting Aspects:**

We explored additional aspects of the dataset, including the distribution of calories and protein, as well as scatterplots and correlation matrices. These visualizations provided further insights into the data set.

A picture containing text, screenshot, plot, diagram

Description automatically generated Majority of the items are in 100 to 120 calories only

A picture containing text, diagram, plot, screenshot

Description automatically generated High protein products had higher density

A picture containing screenshot, text, rectangle, line

Description automatically generated

There exists only one with sodium value 0 in the dataset that is FrostedMiniWheats

#Puffed foods have low calories, QuakerOatmeal had lower carbohydrates (1.0), and potassium is so high in AllBran