US Cereal
Exploratory Analysis

Yeva Kramarova, ykramarova@bellarmine.edu

I. INTRODUCTION

US Cereal: Nutritional and Marketing Information of US cereal. The data set is build-in into R and it can be accessed through a 'MASS' package in R. I chose it because as being international, it was a cultural shock for me to learn how Americans eat so much cereal.

II. DATA SET DESCRIPTION

The US Cereal data frame has 65 rows and 11 columns. The data come from the 1993 ASA Statistical Graphics Exposition, and are taken from the mandatory F&DA food label. The data have been normalized here to a portion of one American cup.

This data frame contains the following columns:

- mfr: manufacturer, represented by its first initial: G=General Mills, K=Kelloggs, N=Nabisco, P=Post, Q=Quaker Oats, R=Ralston Purina.
- calories: number of calories in one portion.
- protein: grams of protein in one portion.
- fat: grams of fat in one portion.
- sodium: milligrams of sodium in one portion.
- fibre: grams of dietary fibre in one portion.
- carbo: grams of complex carbohydrates in one portion.
- sugars: grams of sugars in one portion.
- shelf: display shelf (1,2,3, counting from the floor).
- potassium: grams of potassium in one portion.
- vitamins: vitamins and minerals (none, enriched, or 100%)

Table 1: Data Types and Missing Data

Variable Name	*Data Type*	*Missing Data (%)*
mfr	factor	0%
calories	num	0%
protein	num	0%
fat	num	0%
sodium	num	0%
fibre	num	0%
carbo	num	0%
sugars	num	0%
shelf	int	0%
potassium	num	0%
vitamins	factor	0%

This data set does not contain any missed values. Nearly every variable is either a number or an integer, with only two categorical (factor) variables.

III. Data Set Summary Statistics

The following section dives deeper into the statistical analysis of the variables of the data set.

Table 2: Summary Statistics for US Cereal

```
> summary(data)
 mfr       calories        protein            fat            sodium
 G:22   Min.   : 50.0   Min.   : 0.7519   Min.   :0.000   Min.   :  0.0
 K:21   1st Qu.:110.0   1st Qu.: 2.0000   1st Qu.:0.000   1st Qu.:180.0
 N: 3   Median :134.3   Median : 3.0000   Median :1.000   Median :232.0
 P: 9   Mean   :149.4   Mean   : 3.6837   Mean   :1.423   Mean   :237.8
 Q: 5   3rd Qu.:179.1   3rd Qu.: 4.4776   3rd Qu.:2.000   3rd Qu.:290.0
 R: 5   Max.   :440.0   Max.   :12.1212   Max.   :9.091   Max.   :787.9
     fibre            carbo           sugars          shelf
 Min.   : 0.000   Min.   :10.53   Min.   : 0.00   Min.   :1.000
 1st Qu.: 0.000   1st Qu.:15.00   1st Qu.: 4.00   1st Qu.:1.000
 Median : 2.000   Median :18.67   Median :12.00   Median :2.000
 Mean   : 3.871   Mean   :19.97   Mean   :10.05   Mean   :2.169
 3rd Qu.: 4.478   3rd Qu.:22.39   3rd Qu.:14.00   3rd Qu.:3.000
 Max.   :30.303   Max.   :68.00   Max.   :20.90   Max.   :3.000
   potassium          vitamins
 Min.   : 15.00   100%    : 5
 1st Qu.: 45.00   enriched:57
 Median : 96.59   none    : 3
 Mean   :159.12
 3rd Qu.:220.00
 Max.   :969.70
```

The dataset provides summary statistics for various nutritional characteristics of cereals from different manufacturers. The average calorie content is about 149.4, with a maximum of 440 calories, indicating a wide range in calorie levels across cereals. Sodium content varies significantly as well, with values ranging from 0 to 787.9 mg, suggesting that some cereals are much higher in sodium than others. Additionally, fiber content varies greatly, with a maximum of 30.3 g and a mean of 3.87 g, highlighting differences in fiber levels among the products. Finally, most cereals are enriched with vitamins, though a few have no added vitamins.

Next, the proportion of each categorical variables. In the US Cereal Dataset there is two categorical (factor) variables: mfr and vitamins.

Table 3.a: Frequency and Proportion within Manufacturer Variable

	Frequency.Var1	Frequency.Freq	Proportion.Var1	Proportion.Freq
1	G	22	G	0.33846154
2	K	21	K	0.32307692
3	N	3	N	0.04615385
4	P	9	P	0.13846154
5	Q	5	Q	0.07692308
6	R	5	R	0.07692308

Table 3.b: Frequency and Proportion within Vitamins Variable

	Frequency.Var1	Frequency.Freq	Proportion.Var1	Proportion.Freq
1	100%	5	100%	0.07692308
2	enriched	57	enriched	0.87692308
3	none	3	none	0.04615385

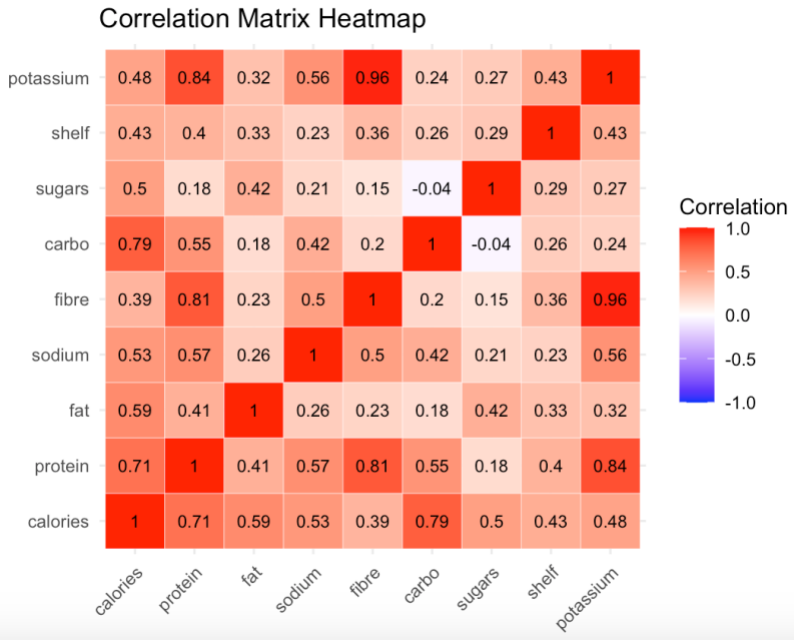
Table 4 shows the correlation between all the numerical variables.

Table 4: Correlation Table

	calories	protein	fat	sodium	fibre	carbo
calories	1.0000000	0.7060105	0.5901757	0.5286552	0.3882179	0.78872268
protein	0.7060105	1.0000000	0.4112661	0.5727222	0.8096397	0.54709029
fat	0.5901757	0.4112661	1.0000000	0.2595606	0.2260715	0.18285220
sodium	0.5286552	0.5727222	0.2595606	1.0000000	0.4954831	0.42356172
fibre	0.3882179	0.8096397	0.2260715	0.4954831	1.0000000	0.20307489
carbo	0.7887227	0.5470903	0.1828522	0.4235617	0.2030749	1.00000000
sugars	0.4952942	0.1848484	0.4156740	0.2112437	0.1489158	-0.04082599
shelf	0.4263400	0.3963311	0.3256975	0.2341275	0.3578429	0.26045989
potassium	0.4765955	0.8417540	0.3232754	0.5566426	0.9638662	0.24204848

	sugars	shelf	potassium
calories	0.49529421	0.4263400	0.4765955
protein	0.18484845	0.3963311	0.8417540
fat	0.41567397	0.3256975	0.3232754
sodium	0.21124365	0.2341275	0.5566426
fibre	0.14891577	0.3578429	0.9638662
carbo	-0.04082599	0.2604599	0.2420485
sugars	1.00000000	0.2900511	0.2718335
shelf	0.29005112	1.0000000	0.4262529
potassium	0.27183347	0.4262529	1.0000000

The correlation matrix below will help us to see whether there are any strong or weak, positive or negative correlations between the numerical variables. A correlation heatmap visually displays the strength and direction of relationships between multiple variables in a dataset. Each cell in the heatmap represents the correlation coefficient between two variables, which ranges from -1 to +1. The closer the number to +1, the stronger the relationship between the variables. The closer the number to -1, the stronger the negative relationship between the variables.

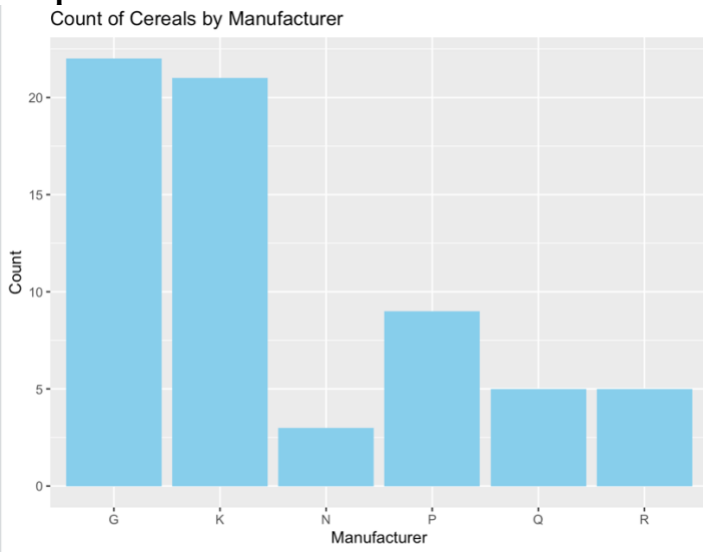


By looking at the correlation matrix, we can see that there are some strong correlated variables. Potassium and fibre are strong positively correlated. Protein and potassium also have a high correlation. Not surprisingly, carbohydrates and calories have high correlation too. Surprisingly, there are just few negative correlations between any of the variables.

IV. DATA SET GRAPHICAL EXPLORATION

In this section, different kinds of graphs will help us to become more familiar with the dataset and to discuss interesting distributions, anomalies, or imbalances of the dataset.

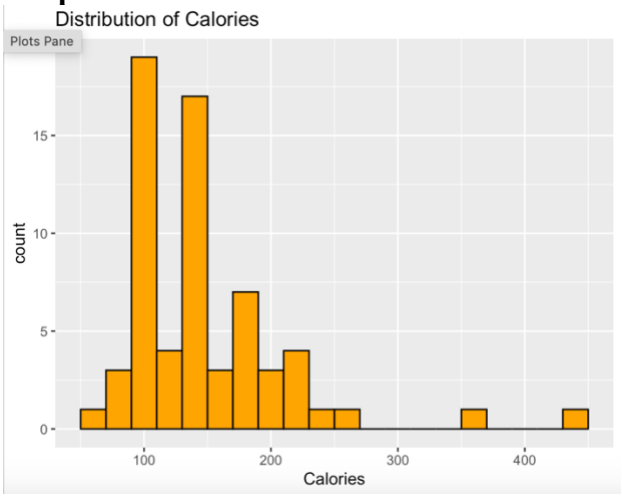
Graph 1: Bar Plot for Manufacturer Count



Based on this bar plot, it is easy to say that General Mills and Kelloggs produces the most variety of the cereal. It also appears that Nabisco produces the least variety of cereal.

The graph below shows the distribution of calories. the histogram groups the calorie values into bins of 20-calorie increments (e.g., 0–20, 20–40, etc.). The **height of each bar** represents the number of cereals that fall within that specific calorie range. For instance, if a bar over the range 100–120 has a count of 5, it means there are 5 cereals with calorie values between 100 and 120.

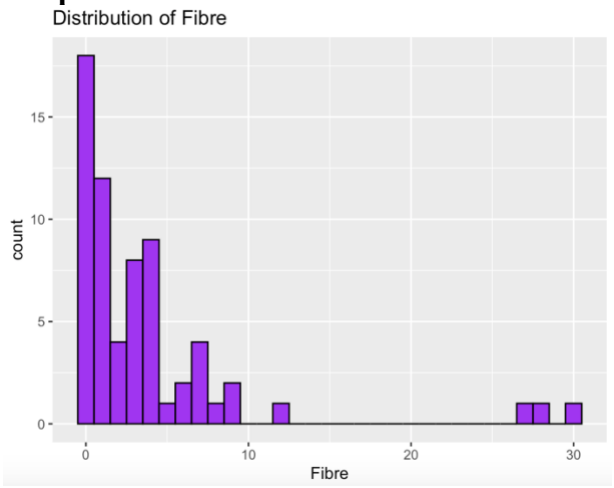
Graph 2: Distribution of Calories



Based on the graph above, we could conclude that the amount of calories per a cup is usually in the range of 100-150 calories. There are few outliers that have 300+ calories per cup of cereal.

The next graph shows the distribution of the fiber in cereal. We can see that fiber, unfortunately, is not added much into the production of cereal.

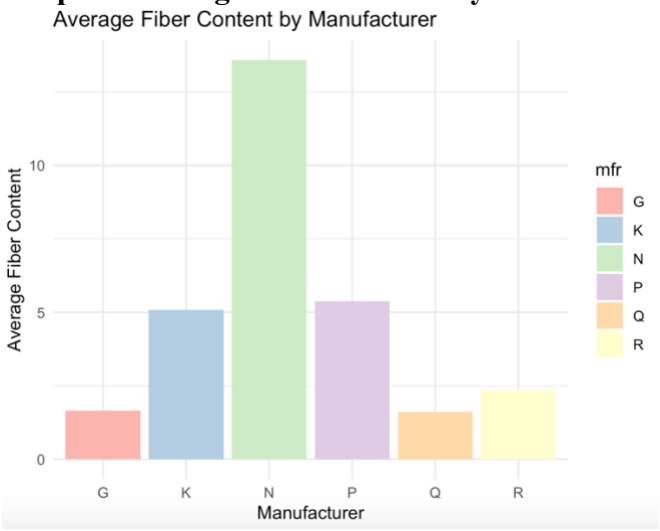
Graph 3: Distribution of Fibre



Based on the graph above, it could be concluded that fiber is not widely used in cereal production. However, fiber helps regulate the body's use pf sugars, helping to keep hunger and

blood sugar in check. Therefore, fiber is an important nutrition in our daily food consumption. So, I decided to create a graph that would show which manufacturers use the most fiber in their cereal production.

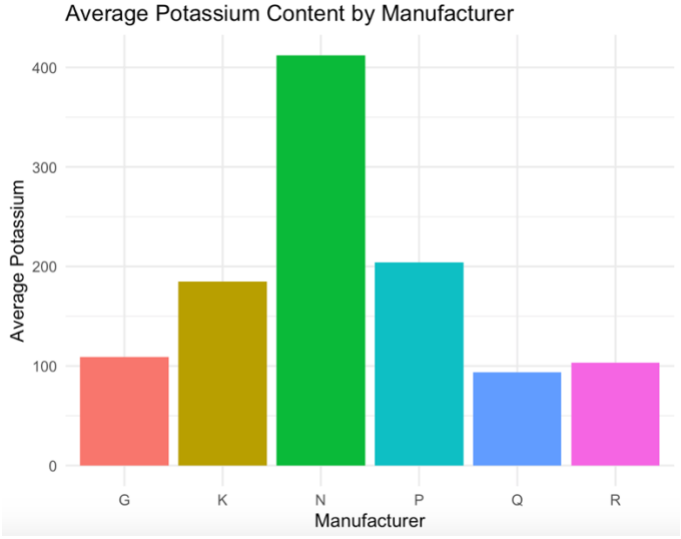
Graph 4: Average Fiber Content by Manufacturer



Here, it is easy to see that Nabisco (N) manufacturer has a much higher average fiber content than any other manufacturer.

Also, recall the correlation matrix that showed us that potassium has several strong correlations with other variables such as fiber and protein. As was mentioned earlier, fibre is an extremely important nutrition, and protein is too. Potassium is also one of the highly important nutrition in our lives since it decreases the risk of the cardiovascular diseases. So, the graph below also shows what manufacturer uses potassium in their cereal production the most.

Graph 5: Average Potassium Content by Manufacturer

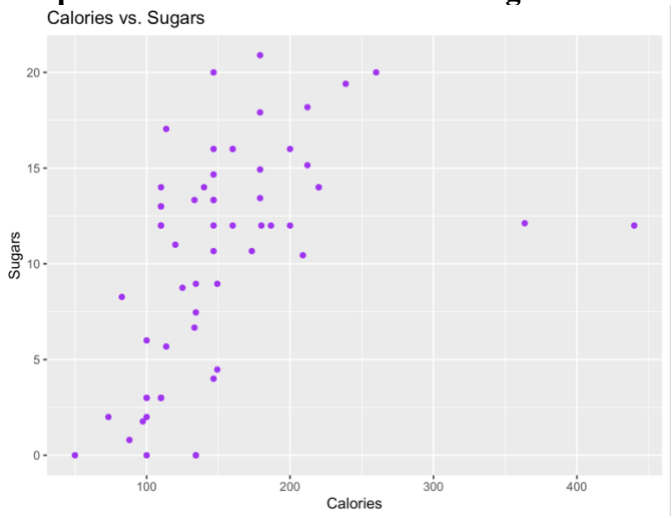


Here, it is easy to see that Nabisco (N) manufacturer has a much higher average potassium content than any other manufacturer.

Nabisco may produce fewer types of cereals, but they lead in average amounts of healthy nutrients, demonstrating that quality outweighs quantity.

Next, we are going to take a look into some scatter plots with trend lines.

Graph 6: Scatter Plot Calories vs. Sugars

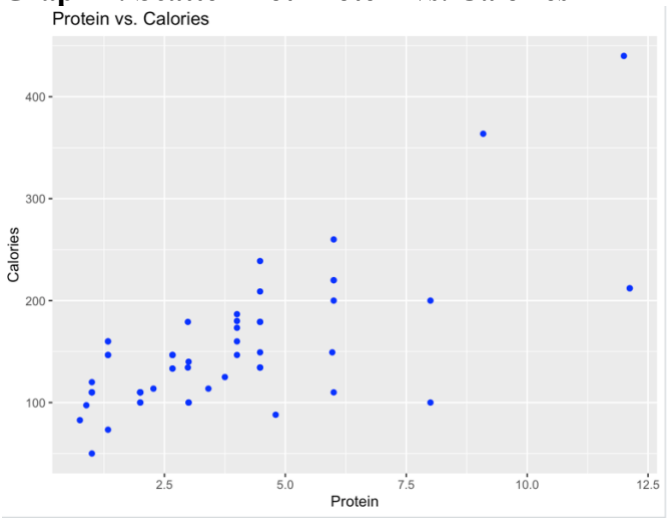


The graph below shows relationship between calories and sugars in a dataset.

It is easy to see that there is a positive relationship between the calories and sugars.

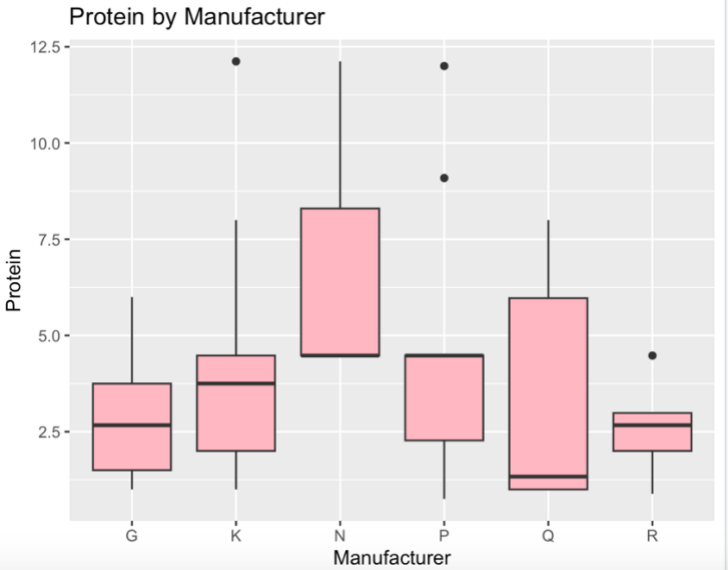
Another scatter plot showing a positive relationship between protein and calories.

Graph 7: Scatter Plot Protein vs. Calories



A box plot below shows the shows the distribution of protein content across cereals for each manufacturer. The protein content is in grams for each cereal. This boxplot helps compare protein content across manufacturers, showing differences in protein ranges, medians, and outliers for each brand, which can reveal if some manufacturers tend to produce higher-protein cereals than others.

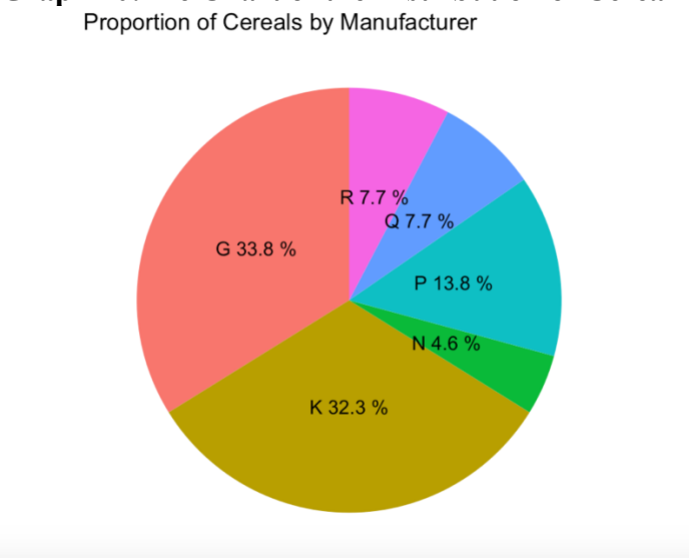
Graph 9: Boxplot Showing Protein by Manufacturer



Based on the boxplot, we again, can see that Nabisco stands out as a manufacturer that tends to produce higher-protein cereals.

Below is the pie chart that visualizes the distribution of categorical variables in the data set. This pie chart visualizes the proportion of cereals produced by each manufacturer.

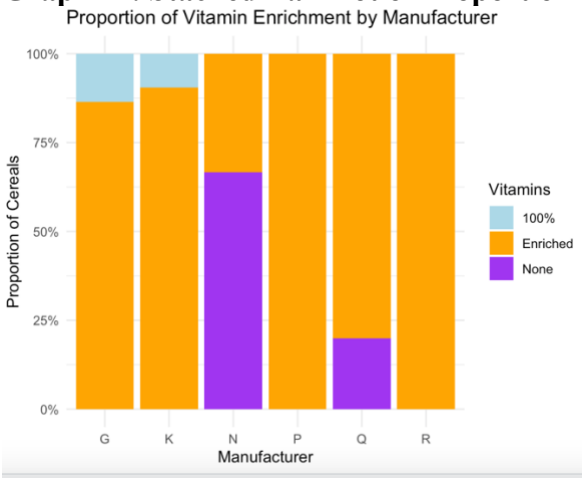
Graph 10: Pie Chart of the Distribution of Cereal Production by Manufacturer



General Mills produce the most cereal.

Lastly, a stacked bar plot that shows the proportion of cereals with each vitamin enrichment level by manufacturer. This way, we can compare how manufacturers differ in their use of vitamin enrichment.

Graph 11: Stacked Bar Plot of Proportion of Vitamin Enrichment by Manufacturer



It appears that General Mills, Kelloggs, and Ralston Purina show the best results as for the vitamin enrichment in their cereal production.

V. SUMMARY OF FINDINGS

The US cereal data set showed us the information about the US cereal manufacturers, together with the nutritional contents of them such as carbohydrates, sugars, calories, protein, fiber, potassium, vitamins, fat and sodium. The data set is from the 1993 data. The graphs did a great job indicating what manufacturers are doing the best job with enriching their cereal with the highest amount of the most vital nutrition. Nabisco showed the best results when the protein, fiber and potassium was being compared among the manufacturers, whereas being the manufacturer with the lowest count of the cereal as shown in the Graph 1. General Mills is the producer of the highest count of cereal, however, contains some of lowest amounts of potassium, protein and fiber. But, General Mills together with Kelloggs and Ralston Purina have high amounts of vitamins in their cereal.