Cereals – short report about findings

On the 25 of April 2021, I have performed analysis on dataset called Cereals. You can find all license data in the README file on GitHub.

In the dataset itself, I had 77 observations/rows and 16 variables. The goal was to try to predict the rating of the cereals in the dataset, with the help of other variables. So, let us start with variable statistics in the dataset/sample.

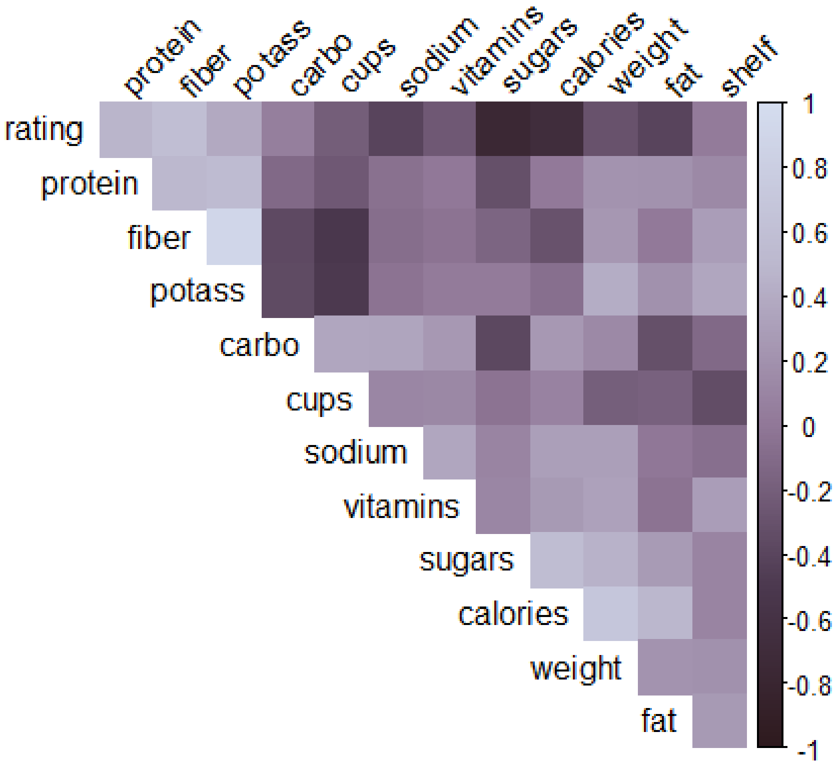
We all know that not all cereals we eat for breakfast (or lunch, I don't judge) aren't the healthiest thing to eat. Let's see the average stats of cereals in this sample.

Variable	Average value
Calories	106,9 ±19,48
Protein	2,545 grams ± 1,09 grams
Fat	1,013 grams ± 1 gram
Sodium	159,7 milligrams ± 83,83 mg
Fiber	2,152 grams ± 2,38 grams
Carbo	14,62 grams ± 4,18 grams
Sugars	6,948 grams ± 4,40 grams
Potass(ium)	96,13 milligrams ± 71,21 mg
Vitamins	28,25 ± 22,34 (FDA recommended, unit unknown)
Weight (in ounces of one serving) – 1 ounce is 28,34 grams	1,03 ± 0,15 ounces [29,19 ± 4,25 g]

The rating itself goes from 0-100, but here it goes from 18,04 to 93,70, with average value of 42,67 ± 14,04. That means that the majority of cereals in the sample were underrated (lower than 50). Let's see why is that or what influences the rating so much, from the POV of those variables in the table above.

I'm going to try to predict the rating, with the help of other significant variables in the dataset. For that, I'm using regression analysis and correlation analysis. The details of both analysis, you can read in the cor+reg – new PDF file on the GitHub.

Below you can see the correlation matrix – the darker the color, the stronger the correlation it is.



Essentially, correlation is a measure which shows how are two variables connected, or what is their relationship. It can be positive (when i.e. fiber level grows, the rating of cereals also grows) or negative (i.e. when sugar level grown, the rating of cereals drops). It can be from -1 to +1, but the strongest correlations start from 0.2. I've checked all correlations, and the picked only 3-4 variables and will continue with regression model.

Regression model result is a linear equation with which you can predict a value. Here, I'm trying to predict the rating of the cereal, with the help of other variables in the dataset. Regression-wise, you can find more details in the cor+reg – new PDF file. Only two regression models were valid. Those are:

Y value	X value	Equation	Example
Rating	sugars	Y= 59.584 – 2,435x	Rating = 59,584 – 2,435 * 6,948 = 42,66
Rating	Sodium	Y= 53,4025 – 0,06724x	Rating = 53,4025 – 0,06724 * 159,7 = 42,66