Do you buy healthy ?

Applied Data Analysis project

EPFL, January 2020

Marc Bickel, Lorenzo Donadio, Samuel Dubuis and Cécile Trottet

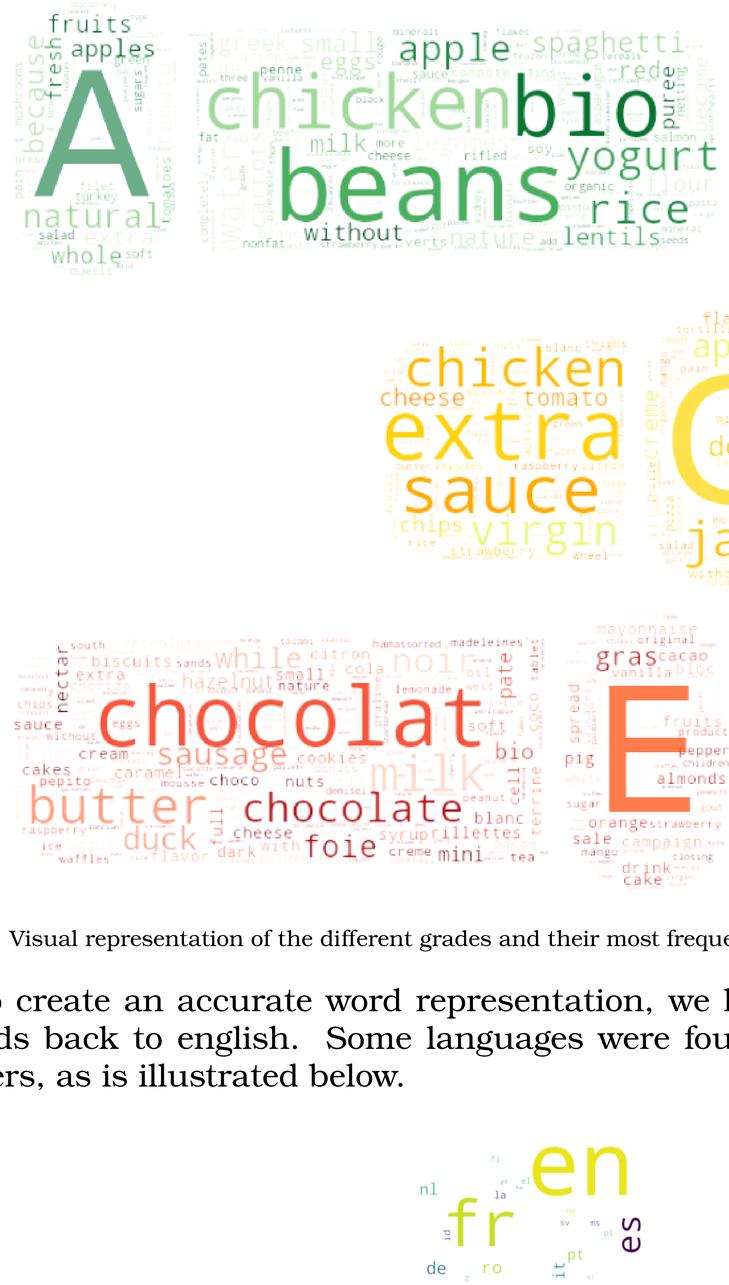
Abstract

How do we assess the quality of the most popular groceries bought online ? The Open Food Facts dataset provided us with the nutritional values of a myriad of products. After extracting the attributes most relevant to our analysis, we matched our findings with the Instacart dataset, containing real world orders.

Step 1 : Word Representation

The nutri-score attribute in Open Food Facts ranges from A to E, A corresponding to the healthiest products and E the worst.

The very first step was to find for each nutritional grade, the most present words in the products to create a word representation of each grade. We created dictionaries of the most important words with their number of occurrences for each grade.



Visual representation of the different grades and their most frequent words in Open Food Facts.

To create an accurate word representation, we had to translate all these words back to english. Some languages were found more frequently than others, as is illustrated below.

Step 2 : Open Food Facts to Instacart

For a given product j in Instacart, we computed its affiliation to the different nutrition grades in the following way :

$$Grade_{i,j} = \frac{N_{i,j}}{Length\left(dict_i\right) \times Length\left(name_j\right)},$$

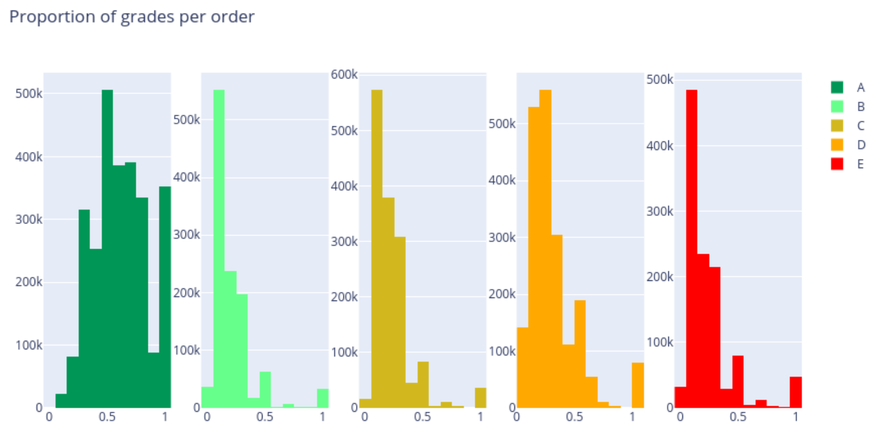
with $i \in [A, B, C, D, E]$ and $N_{i,j}$ is the number of words in the name of product j belonging to dictionary i. The grades were then normalized, and the highest computed score shows to which grade a product belongs the most. We also computed a weighted version of the nutrition score.

Step 3 : How do people buy ?

With the Instacart nutrition grades in hand, we focused on two distinct aspects of the purchases. We analyzed the overall content of the orders and the grades of the most sought after products.

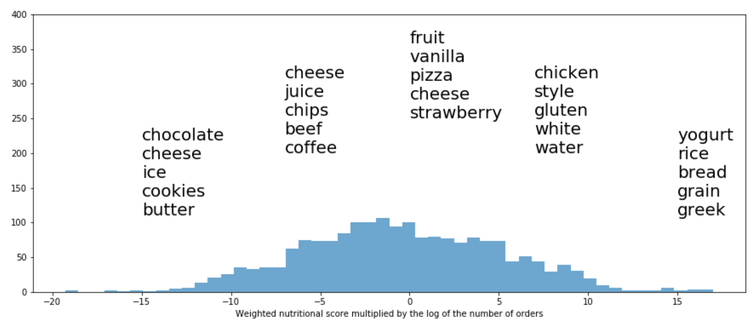
Distribution of the grades in the orders

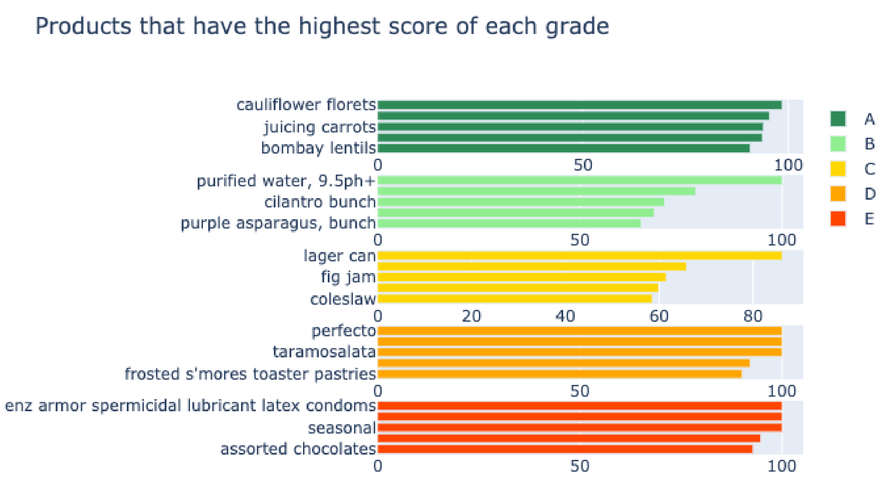
The following plots provide a closer insight into what people put into their baskets. They show the proportion of a given grade inside the orders.



Food rankings

These rankings show which products have the most impact on the consumers health in terms of nutrition score and number of orders.





Fun Facts

Some words that appear the most often are not always food related. However, one can easily imagine why these words come up as related to a specific grade. "**without**" is quite big in grade A, while "**extra**" appears as an important word in grade D. In this same representation of the instacart products "**large lemon**" is quite surprisingly one of the largest words. This misclassification of a healthy product, is probably due to the fact that the adjective "large" is often associated with unhealthy products.