COM480
Data Visualization

Milestone 1

By:
Siffert Jean - 366682
Rein Gundersen Bentdal- 366666
Faucher Laurent - 366681
Insight Infusion

April 7th 2023

Dataset

We chose to use the dataset from the Open Food Facts association [1]. It's an open-source and crowdsourced database that contains a lot of information about approximately 2 million food products. It gathers information such as ingredients, origins, additives, nutritional quality, ecoscore, CO2 footprint, packaging, etc.

We retrieved the entire database for a total of 2,868,223 products. By playing with the data, we realized that a lot of products had some missing fields since users have the option to leave some of them blank while filling out the form for a new product. Nevertheless, every product has a completeness field, so we can filter the products to keep the ones with the most information. The following chart shows the number of products by completeness range :

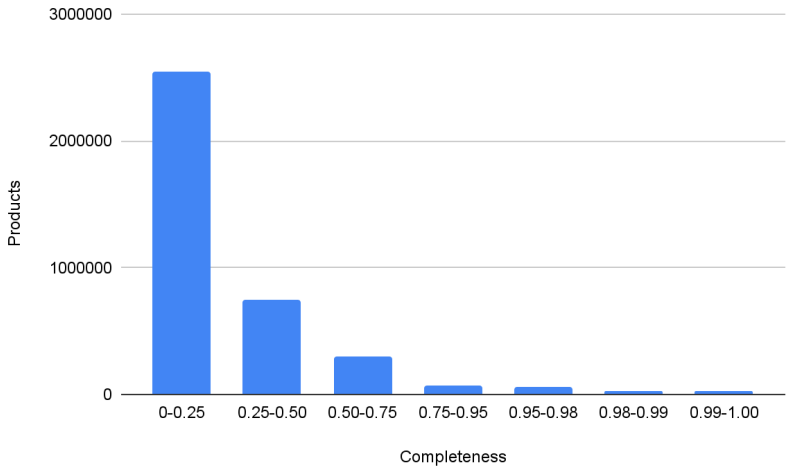

Figure 1 : Number of products by data completeness range

There are 69,622 products that have a completeness of 95% or more. We believe that this represents enough data to perform meaningful analysis. The completeness index is based on all the fields, but many of them are not useful to us. Moreover, some fields have a repetitive meaning that can lower the completeness index (e.g. "*origins*", "*origins_tags*" and "*origins_en*"). We will do a data cleaning to combine similar fields, infer information from them and keep only the relevant ones for our analysis.

In addition, since it is an open source dataset and anyone can add a product, the quality of the data is not optimal. For instance, we have to trust the community and assume that the information on every product is true, even though it doesn't necessarily come from a reliable source. Also, the users who filled the information don't always use the same format. As a result, the data is not captured consistently, and we will need to map the user inputs to their true meaning.

OpenFoodFacts is a platform that is originally from France. Since most of the contributors are French, the data on food products has a French bias. Therefore, the visualization will be more focused on the French food market.

Problematic

We often tend to take the food we eat for granted, but it has a huge impact on our health and that of our planet. Firstly, a lot of the food that we buy at the supermarket has been transformed and contains additives and modified ailments that can have impacts on our health.

Secondly, two of the most influential sectors in global greenhouse gas emissions are agriculture and transportation, which are intrinsically linked to the food industry. For instance, transportation is responsible for 14% of worldwide emissions, while agriculture is responsible for 24% of emissions as well [2].

That being said, we would like to develop visualizations to help supermarket customers shop for products that are good for them and the environment. Our visualization should point out the extremes in ingredients to show which products are the best or the worst. For instance, one of the features we would like to add to our project is a world map that shows the origin countries of the ingredients in a product. This would tell the shoppers if a lot of transportation was involved in the making of a specific product. We would also like to make some visualizations about food additives and their effects, because they are often hidden from consumers.

Exploratory Data analysis

To begin the exploratory data analysis, we chose the dataset with data whose completeness index is above 95%. This dataset contains 69622 products. We developed a Jupyter notebook using Python to play with the data and show insights. Here are some basic statistics and insights we discovered while digging into the data.

Allergens

48% of products (33597) contain allergens and these products contain an average of 1.79 allergenic substances per product. The following graph shows the 10 most present allergens :

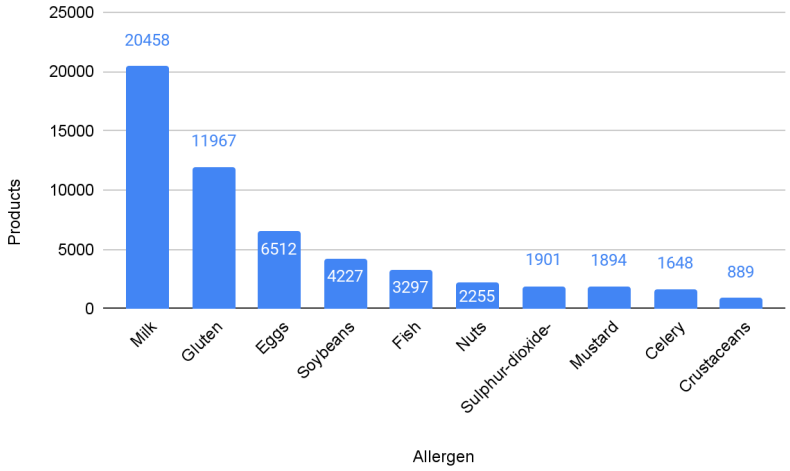

Figure 2 : 10 most common allergens in food products

Additives

32882 food products, i.e. 47% of the sample contain additives. Furthermore, a product that contains additives is composed of an average of 3.12 additives. The following graph shows the 10 most present additives :

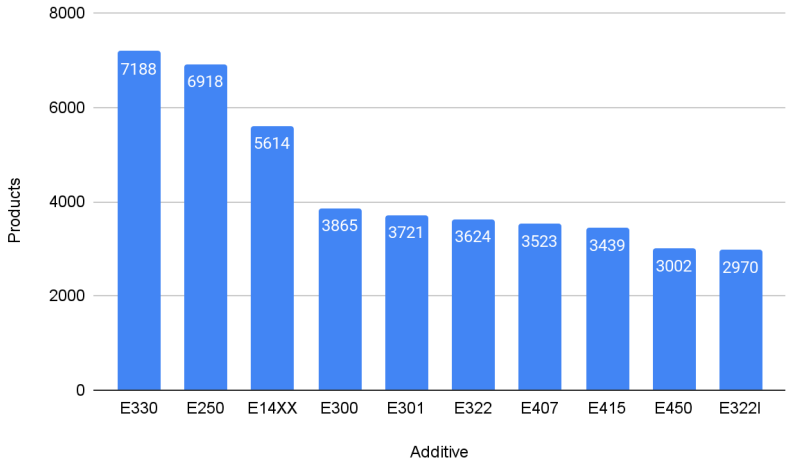

Figure 3 : 10 most common additives in food products

Nutriscore (nutritive score)

We know the nutriscore (score noting the quality of the product for a healthy and balanced diet in France) of 84% of products (58537). The following graph shows the distribution of nutriscores (A being the best) :

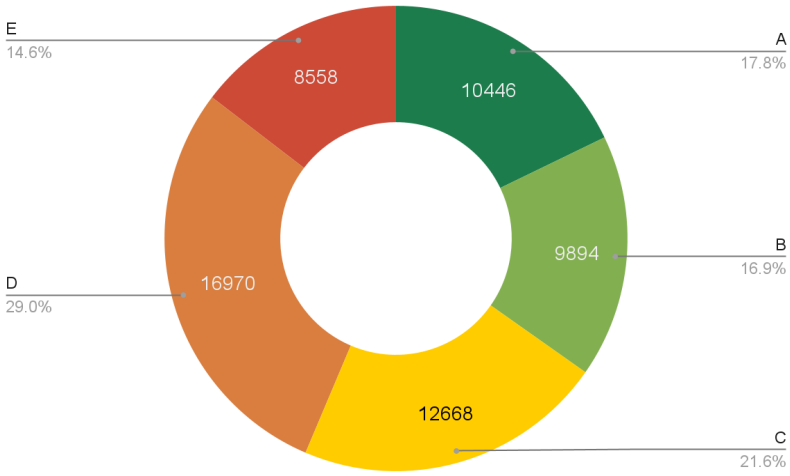

Figure 4 : Distribution of nutriscores in products

Brands

The products in the database are sold by over 21k different brands. The following graph shows the distribution of the 10 most present brands :

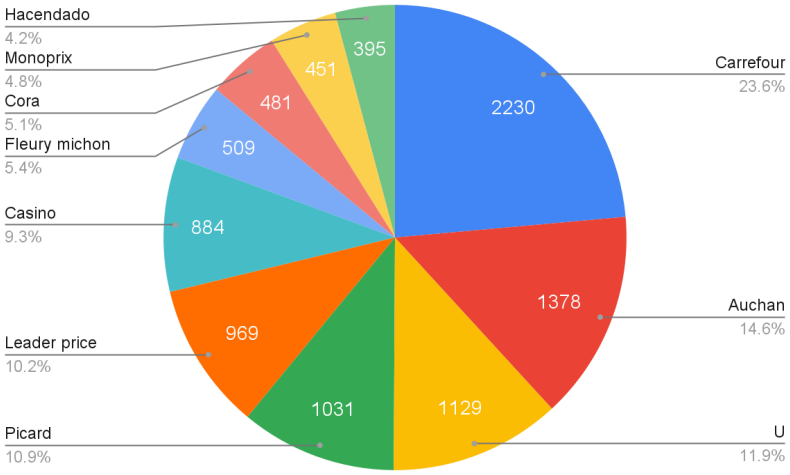

Figure 5 : 10 most common brands in the dataset

Origins of ingredients

We have set up a data processing pipeline to translate the sometimes vague location information (different languages, regions or very local areas) into country names that we can easily exploit. The following map, made with D3.js shows a first draft of a visualization of the countries of origin of the ingredients :

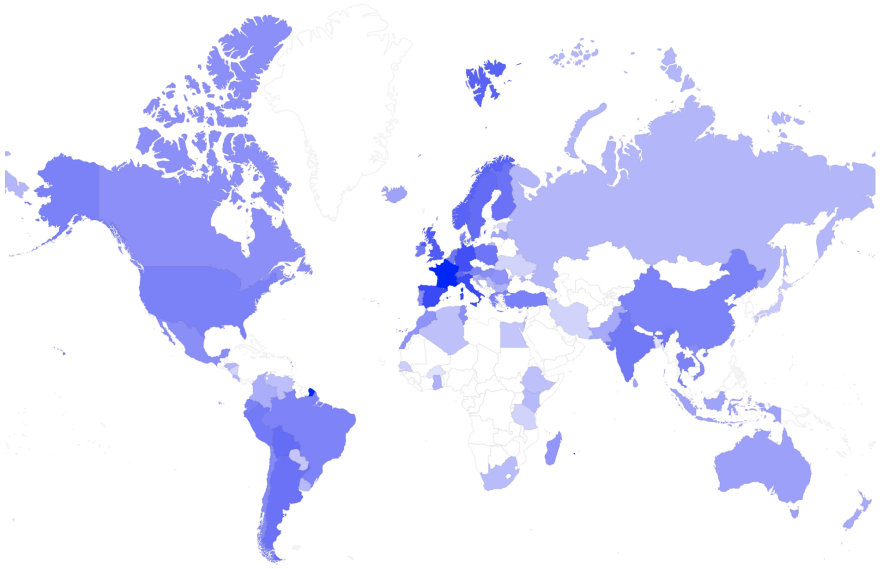

Figure 6 : Map showing the distribution of the countries of origin of the ingredients

Related work

The database from the Open Food Facts association is at the heart of a fast-growing application named Yuka. Yuka is a mobile application developed in France in 2017 that allows supermarket customers to scan and obtain information on the healthiness of a product [3]. The app can also help you choose a healthier product by proposing better alternatives to less healthier products. To let the customer know if the product is good for them or not the application provides a score between 0 and 100 based on the nutritional information.

Our project would be different because, first of all, we would like to tackle more the environmental impact of products. Also because Yuka only gives information about products, while we will create more global visualizations with intuitive graphs using the D3 library.

The website from the Open Food Facts association itself allows users to search for products and see the information, but the visualization is again limited and does not allow a good comparison of the products between them to get an overview.

We want to create visualizations that provide a more holistic view of the data to identify trends and key points that influence the products that make up our daily diet. To do this, we are going to use the D3.js library as well as the techniques seen during the COM-480 courses.

For our approach we are mainly inspired by some examples that are displayed on the https://observablehq.com/@d3/gallery website.

Here are a few graphs that caught our attention and that inspire us in our project:

- https://observablehq.com/@d3/force-directed-graph
- https://observablehq.com/@d3/world-choropleth
- https://observablehq.com/@d3/treemap
- https://observablehq.com/@d3/hierarchical-edge-bundling/2

References

[1] Open Food Facts. (2023). [Online]. Available : https://fr.openfoodfacts.org/
[2] United States Environmental Protection Agency. (2014) Global Emissions by Economic Sector. [Online]. Available : https://www.epa.gov/ghgemissions/global-greenhouse-gas-emissions-data
[3] Yuka. (2023). [Online]. Available : https://yuka.io/en/