Healthy food: from packaging to consumption

Matthieu Buot de l'Épine, Valérian Rey and Adrien Vandenbroucque
Department of Computer Science, EPFL, Switzerland

***Abstract*—This document contains a summary of the different steps we took when analyzing the Open Food Facts dataset. We will explain precisely how we first cleaned the dataset and performed many conversions. In a second part, we explain the analyses that we made together with the meaningful insights we got.**
Finally, we will show the results, and see the conclusions we can draw from them with respect to our problematic.

I. Introduction

One of the highest stakes of the century is to limit the damages done to natural resources. Understanding where the food is produced, packaged and consumed can show insights about the current issues in the food industry.

For example in some countries, there is a tendency to favor local products. We want to see if this is really the case and see more generally which countries import-export where in order to better understand their food consumption.

Also, today, people start to buy more and more organic products, because they want to eat healthy and make a positive impact on the environment. At the same time, packaging has become an important matter and there is a need to reduce our plastic waste. Nutrition which is directly related to health, is also a big concern for people today. Now can we see these new trends in the Open Food Facts dataset? And also, are there any specific characteristics in the packaging of organic products?

Our project is motivated by the idea of understanding these tendencies in the marketing behind the food industry, and get meaningful visualizations of these trends.

II. Dataset

A. Open Food Facts

1) Brief description: "Open Food Facts is a food products database made by everyone, for everyone."

2) What the dataset contains: The dataset we used is in CSV format, and lists hundreds thousands of products. The information that were useful for our project are the following:

- Product names
- Origin countries
- Destination countries
- Packaging
- Labels
- Image URL
- Nutrition values

3) Statistics about the dataset: The thing about the Open Food Facts dataset is that it is continuously updated, so we cannot give the exact number of products in the database. At the time of the writing of this report, there are around 700'000 products.

4) A note about the dataset: An important point that needs to be made right at the start is that this dataset was created in France, and so many of the products added come from France and Europe. Thus we note here that this may introduce significant biases in the analyses, and we therefore ask the reader to be careful not to jump to conclusions. Also, the products are not independent entries. If a country has 100 products, and all of these were given by the same person, it induces a huge bias (for example if this person is vegan, most of the products he will include will have the vegan label, which will give rise to an outlier when analyzing the proportion of vegan products per country).

III. Data Cleaning

The biggest part of our work surely went for the cleaning of the dataset. Indeed, since the dataset is "made by everyone, for everyone", the format of the different columns of the CSV happens to be quite chaotic. The contributors input data that often isn't in the format that was expected. For example, many languages are used in the dataset. We explain here the main things we changed in order to make use of most of the data at our disposal.

A. Proportion of missing values

We were quite surprised to notice that for most of the columns, the proportion of missing values exceeded 50%. Actually, only 30 over the 173 features available had less than 50% of missing values. So in a first step, we had to carefully choose the columns we wanted to keep in order to perform our analysis. This way, we can use as much data as possible in order to get better and less biased results.

B. Countries

After choosing the columns to keep, the next step was to clean the features themselves. For those containing country names, a lot of work had to be done. Not only contributors wrote in many languages, so we end up with values like "*Royaume-Uni*", "*United Kingdom*" or "*UK*" which all represent one single country. But sometimes, they also indicated the exact region, like "*California*", instead of a country name.

We designed a way to efficiently translate most of them to English and to the corresponding alpha-3 country code with the help of the library `PyCountry`.

C. Packaging and labels

For the analysis of the packaging, we also needed to translate the terms to a single language, otherwise when grouping by packaging, we ended up with different groups (one in each language) for one type of packaging. For example, for the plastic type of packaging, we could have "*plastico*" (Spanish), "*plastique*" (French), or "*Kunststoff*" (German).
In order to clean this, we looked at the 12 most used languages in the dataset, and created functions to select products corresponding to each of the main packaging materials (plastic, cardboard, glass and metal) for those languages. Likewise, we also defined functions to select all the products with interesting labels (among organic, vegan, halal, gluten-free and fair-trade).

D. Columns containing lists

For columns containing information about countries (origin and sell place), it was often the case that there are multiple values in the form of a list, like "*Germany, France*" Since we needed all countries in those lists for our analysis, we created a function that duplicate lines when there are multiple countries, one for each. Indeed, when grouping by countries, we want that if a product is sold in both Germany and France, then it appears in both groups. Note that we have thought of using the same technique for labels and packaging, which are often lists as well. But this would be wrong when counting the proportion of products with each packaging / label. For example with a single product with packaging "plastic, cardboard", if we duplicate the line into a product with packaging "plastic" and a product with packaging "cardboard", we would mistakenly say that half of the products contain plastic and half of the products contain cardboard, while actually the only product contains both.

IV. Analysis

We now turn to the actual analyses we performed on the cleaned dataset. Note that, since it contains many missing values in different columns, we split it into multiple smaller datasets (one without any spatial information, one with sell place only, and one with origin and sell place) so that for each of our analyses, we can use as much data as possible and don't remove any useful information.

A. Analyzing import-export

Since we wanted to better understand how the food travels, we decided to develop an easy way to visualize, for any country, the places where it exports and from where it imports.
In order to achieve this, we made good use of the cleaned dataset and used it to group the products by origin and destination countries. This way, we were able to retrieve the count of products for each country where a specified country imports or exports.
This lead us to the kind of visualization shown in Figure 1.

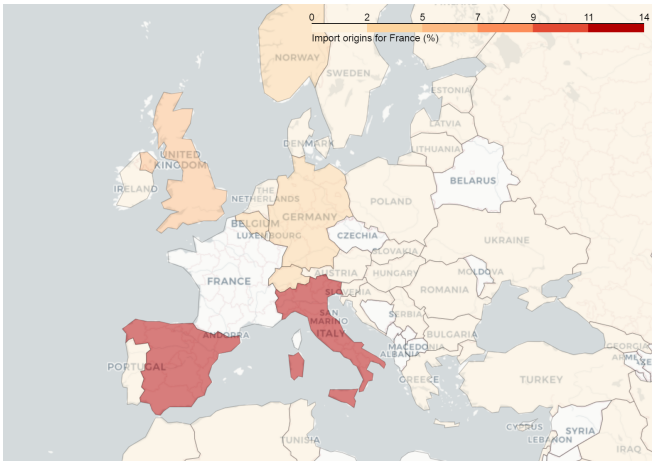

Figure 1. Countries in Europe from where France imports (France excluded)

With the import-export analysis, getting information about the proportion of products that are local in a country is easily done. We provide the results we got for Switzerland in Figure 2.

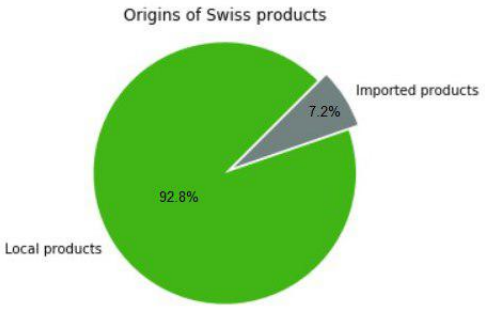

Figure 2. Proportion of products that are local vs imported in Switzerland

B. Analyzing the packaging

Another part of the project was focused on the packaging. Some questions were: What is the most used type of packaging around the world, and in what percentage? We obviously expected the plastic to be number one, and the results are shown in the next section.
To perform this analysis, we used this time the cleaned dataset containing the products' sell place. In the same spirit as for the countries, we grouped by countries to see which

one uses what type of packaging. We also computed the proportion of the different types used around the world, and we included in this report results for the plastic type. We only displayed the proportion for countries that have at least 100 products in the dataset. The others are colored in grey (Figure 4).

C. The link with organic products

A big part of our project was to perform the analyses we made just before about import-export and packaging, but specifically on organic products. The methods we used are the same but this time we included a confidence interval (99.99%). This would be true only if the product entries were independent, which is (like explained above) unfortunately not the case. Thus we do not really reach that level of confidence, but we come close to it.

We now want to highlight another big analysis of our project: the comparison of dominant colors in the images of organic products versus non organic products. In order to perform such a work, we first needed to retrieve all the images of products (with the right labels) using their URLs, which was available through a JSON API on the Open Food Facts website. While exploring, we saw that the pictures sometimes were taken by a smart phone, and so the background colors would be the main ones. So we decided to first detect the text on the images using the `OpenCV` library [1], and then perform the color detection around this area, since it will indeed contain the color of the package and not noise. Another advantage of this technique is that pictures which actually don't show a package are ignored. We provide an example of the results we get on one image in Figure 3.
From there, we used the *KMeans* algorithm to compute the two main colors in each image (text and background) and combined all the results. We then decided to filter these colors and throw away those which had a low Saturation or Value in their *HSV* representation because we had too many whites and blacks, which we considered as outliers as they are not usually a design choice but more of a default option. We applied *KMeans* again on the remaining colors, this time with 100 clusters to create a spectrum of the dominant colors for the organic and non-organic products packages, which are presented in Figure 6.

V. Results

A. Using Folium to display the results

For a cleaner and easier visualization of our results, we decided to use the `Folium` library in order to display world maps and color intensities to show information for each country. For example, in this way we could display per country the proportion of products that use plastic packaging, as we can see in Figure 4.

Figure 3. Result of the text detection, and the two dominant colors that were extracted

B. Results for import-export

Our functions let us easily display world maps showing, for a given country, the places where they import and export (the country itself is excluded). An example of the result we got for the countries from which France import is displayed in Figure 1. Note that we only included the Europe in order for the map to be more readable. One can see that for France, imports mainly come from Spain and Italy, followed by UK and Germany.
We could also observe that for Switzerland, more than 92% of the products are actually manufactured and sold in Switzerland, which shows that the country tries to eat as local ass possible. However, note that it doesn't take into account the volume of sales for each product.

C. Results for packaging

1) What type of packaging is used around the world?: From our analysis, we got that the most used types of packaging are plastic, cardboard and glass. We then spent more time on analyzing results about plastic since this was one of our main focus.
As we can see in Figure 4, the proportion of packaging using plastic is very high in every country. This is not a surprising fact as we can see that in our every day's life. People are today trying to change this tendency and promote material that we can recycle. This tendency is strong for organic products, as we can see on Figure 5. The proportion of plastic drops, while the use of other materials like cardboard or glass goes up, suggesting that change is taking place. Note that the total can exceed 100%, because many packages actually contain multiple materials (plastic and cardboard for example).

2) Dominant colors of the packaging: The results for this part were pretty interesting. Indeed, since the images of products were often of bad quality and very noisy, we expected bad results. However, using text detection on the

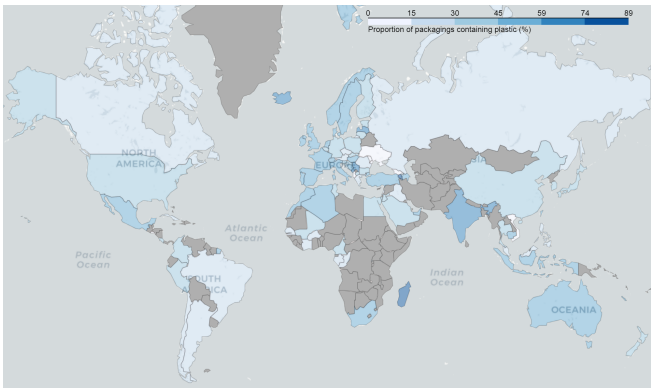

Figure 4. Proportion of packaging containing plastic in the World

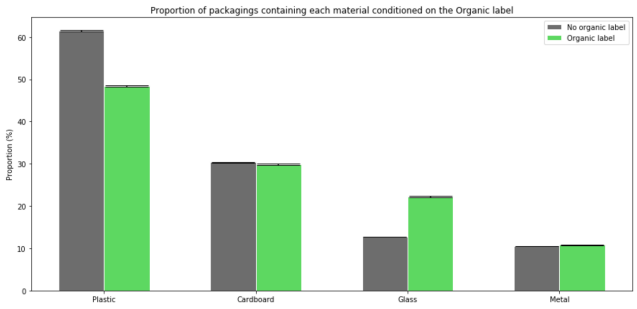

Figure 5. Comparison of the main types of packaging for organic and non-organic products

images and performing the color detection around them let us get the colors spectrum presented in Figure 6.
The first one represent the main colors used among all products, and the second one the main colors but for organic products. As one can imagine, we can see that for organic products, the green color is more present, and we can actually observe it on the spectrum. In general, there is a tendency to use less "flashy" colors for organic products, so we end up with darker tones for the colors, which seem to indicate for the consumer that the product is healthier.
As pointed out in this article [2], color really is a key factor when it comes to buying food, and the spectra we present provide a way to easy visualize it.

VI. Conclusion

The import-export maps let us easily visualize how food travels from one country to another. With the example of France, we were able to see that many products actually come from Spain or Italy. This also let us see what proportion of products are actually manufactured and consumed directly in France. When displaying for many countries, one can easily understand how the import/export takes place in the world.

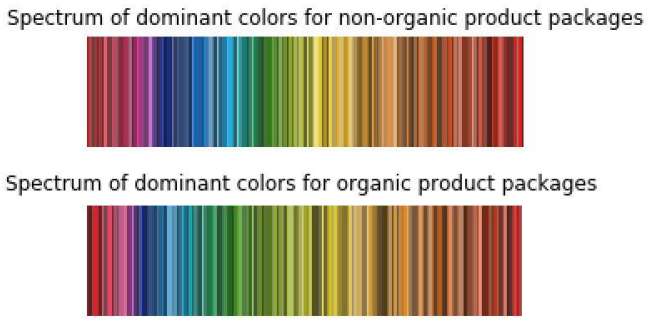

Figure 6. Comparison of the main types of packaging for organic and non-organic products

Plastic is the most widely used type of packaging, with a proportion higher than 45% in most countries. For organic products, plastic is also very often used, but the proportion is smaller, which indicates how organic products are also concerned about environment when it comes to packaging.

The importance of the colors of the packaging is higher than we think. For organic products, the green color is used very often, and dark tones are also widely used. This suggest that the consumer has linked the concept of healthy and organic food to these tones, and those are actually used for the marketing of many products.

References

[1] "Opencv text detection (east text detector)," 2018. [Online]. Available: https://www.pyimagesearch.com/2018/08/20/opencv-text-detection-east-text-detector/

[2] "How food packaging color influences consumer behavior," 2016. [Online]. Available: https://hartdesign.com/industry-news/food-packaging-color-influences-consumer-behavior/