What's in my food?

Zoé Baraschi
zoe.baraschi@epfl.ch

Simon Maulini
simon.maulini@epfl.ch

Abstract

This paper is an analysis of the correlation between food composition - in terms of nutriments - and population health. It is based on the Open Food Facts dataset available online. The paper focuses on the following diseases : diabetes, obesity and heart attacks. The main idea is to highlight links between the diseases and the composition of the food grouped by countries.

1 Introduction

The alimentation question is one of the most important issues of the 21st century. Indeed, the number of health problems linked to the over-consumption of sugar and fat in the rich countries has been increasing in the past years. Simultaneously, the population of many developing countries is hurt by the lack of food. We have to change our food habits and it can begin only if the population acknowledges the urgency of the problem.

We choose to focus our study on the following diseases : diabetes, obesity and heart attacks. The main goal of our work is to try to show where people are more healthy and watch if there is a link with the quality of their food.

2 Data collection & cleaning

We start from the dataset of Open Food Facts (Open Food Fact, 2018). In that file there is a lot of information on many food products like the country where they were sold and their composition. To be able to work properly, we needed some extra information, such as country names (IP2Location Country Multilingual Database, 2016) and their populations. The main dataset contains 685'395 entries but a lot of information is missing for many entries.

That led us to the first tricky part : extract relevant data from the initial data set. Concerning the food composition we chose to focus the majority of our study on the sugar and the fat nutriments because there is a lot of rows containing information about that.

Concerning the location of the food products, we had to apply many operations to be able to get similar format for the entire data set. We chose to use the "ISO 3166-1 alpha-2" code to classify products by countries. If a product is sold in many countries, we duplicate the row for each of the countries.

Information about the different diseases was not present in the main dataset, so we had to search the internet for legitimate sites to find them. For the heart attacks we found our data in a paper available on the National Center for Biotechnology Information website (IHD Information, 2013). Concerning the obesity prevalence, we found data from the website *https://ourworldindata.org* which results' are based on the data of "World Health Organization Global Health Observatory" (World Health Organization Global Health Observatoryn, 2016). Finally, data about diabetes prevalence are from the "International Diabetes Federation" website (Internation Diabetes Federation, 2017).

3 Analysis

First, it is important to notice that our tool for displaying maps contains 3 little strange behaviors. Indeed Kosovo, the north of Somalia and a part of Morrocco is always displayed with the highest value. We did not success to fix them but we can nonetheless use our maps to show some tendencies. We use a JSON file (JSON Countries file, 2018) to be able to plot the maps. It is possible that the polygons contains some little drawing errors and lead to this strange behavior.

3.1 Sugar vs Fat

We decide to start this report with the sugar and fat nutriments. We will see below that a strong links exists between those two factors.

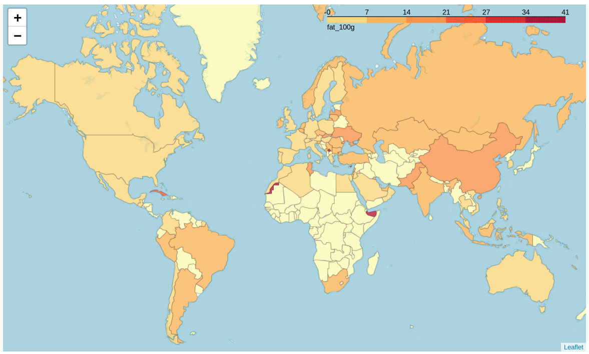


Figure 1: Fat by countries

As we can see on the figure 1 the common idea ”Americans eat much more fat than the rest of the world” is not really true. Indeed, Asian countries eat more fat and the southern part of the American continent shows a bigger consumption of fat than the north.

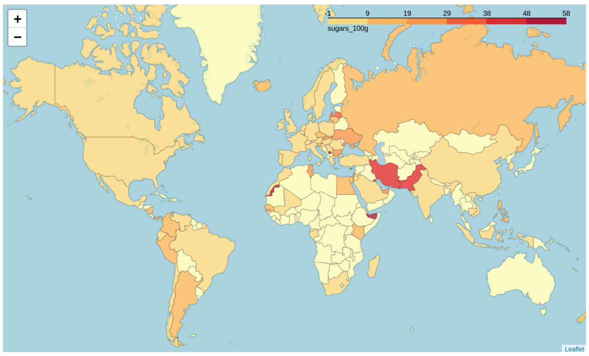


Figure 2: Sugar by countries

The distribution of the sugar consumption looks pretty similar as the one for the fat, except for two outliers (Iran and Pakisan). This visual result lead us to investigate further.

We decided to try to find whether there is a correlation between those two metrics. We started by making a scatter plot (figure 3). It is hard to see a real correlation on the plot but we can suppose that the result is biased due to the big differences between the lifestyle in the different continents.

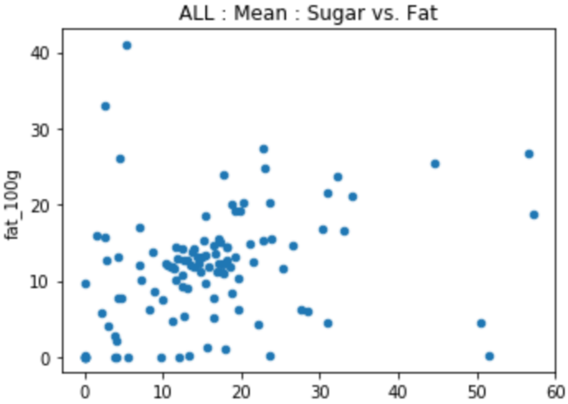


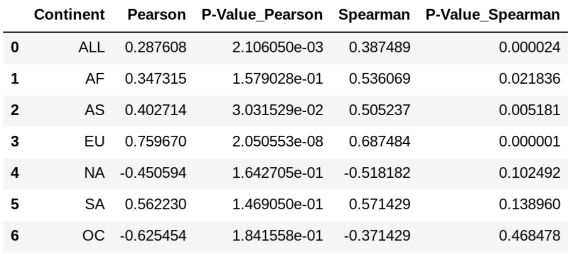
Figure 3: Correlation between Sugar and Fat

We thus decided to group the countries by continents to check if we could find more correlations.

When we look the results (figure 4) we have a case of the Simpson’s paradox: we can see that the correlation for the mean/median for all continents is around 0.3-0.4, which is not very significant.

When we look at continents on their own, we can see that Europe has a high Pearson correlation (0.759 for the mean and 0.727 for the median. The p-value is much smaller than 0.05, so the result is significant. The Spearman correlation is also quite high for the mean (0.688) but not so much for the median (0.54). It makes sense when we look at the correlation plot. For the mean, there is a clear monotonic (always increasing) and a linear relationship. The Spearman correlation is also quite strong in this case, thanks to the monotonic relationship. We can see a strong linear relationship for the two groups in the median, but a weaker monotonic relationship, which explains the weaker Spearman correlation.

Asia comes next in line with respect to correlation coefficients, having a moderate positive correlation. We can spot some linear relationships in both plots, but the outliers influence the monotonicity of the relationship.



	Continent	Pearson	P-Value_Pearson	Spearman	P-Value_Spearman
0	ALL	0.287608	2.106050e-03	0.387489	0.000024
1	AF	0.347315	1.579028e-01	0.536069	0.021836
2	AS	0.402714	3.031529e-02	0.505237	0.005181
3	EU	0.759670	2.050553e-08	0.687484	0.000001
4	NA	-0.450594	1.642705e-01	-0.518182	0.102492
5	SA	0.562230	1.469050e-01	0.571429	0.138960
6	OC	-0.625454	1.841558e-01	-0.371429	0.468478

Figure 4: Correlation between Sugar and Fat

If we render a scatter plot only for Europe, we can clearly see the positive linear correlation.

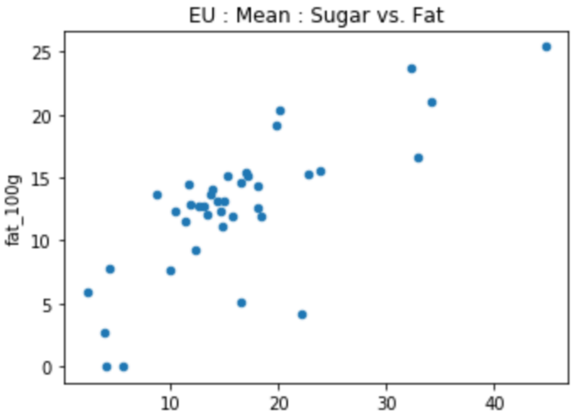
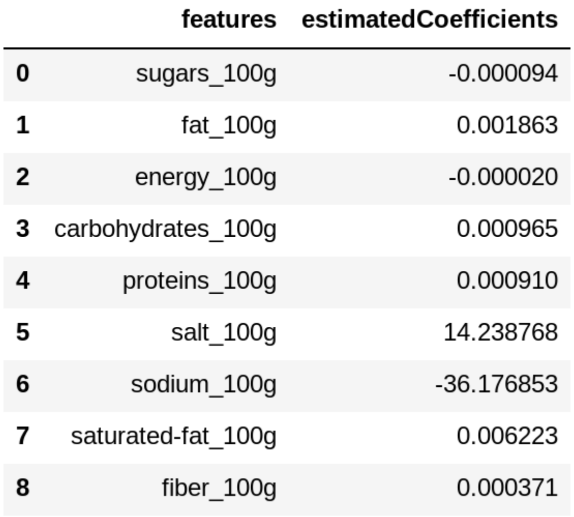


Figure 5: Correlation between Sugar and Fat for EU

3.2 What about heart attacks ?

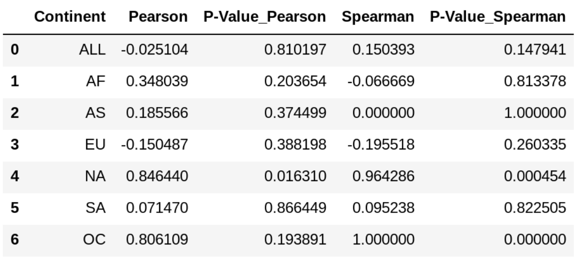
To identify the most preponderant features concerning the coronary artery disease (IHD), we decided to run a linear regression on the dataset, concatenated with the average heart attack prevalence by countries. The results are displayed in figure 6.



	features	estimatedCoefficients
0	sugars_100g	-0.000094
1	fat_100g	0.001863
2	energy_100g	-0.000020
3	carbohydrates_100g	0.000965
4	proteins_100g	0.000910
5	salt_100g	14.238768
6	sodium_100g	-36.176853
7	saturated-fat_100g	0.006223
8	fiber_100g	0.000371

Figure 6: Results of the linear regression for IHD

We see a strange behavior. Indeed the salt and the sodium seem to have a real impact on the IHD prevalence but one with a positive coefficient and the other with a negative one. To further investigate, we applied the same method as in the previous sections to the sodium metric. We grouped the countries by continents to see if the correlation is bigger in some parts of the world. As we can see on figure 7 North America is more affected by the presence of sodium in food.



	Continent	Pearson	P-Value_Pearson	Spearman	P-Value_Spearman
0	ALL	-0.025104	0.810197	0.150393	0.147941
1	AF	0.348039	0.203654	-0.066669	0.813378
2	AS	0.185566	0.374499	0.000000	1.000000
3	EU	-0.150487	0.388198	-0.195518	0.260335
4	NA	0.846440	0.016310	0.964286	0.000454
5	SA	0.071470	0.866449	0.095238	0.822505
6	OC	0.806109	0.193891	1.000000	0.000000

Figure 7: Correlations between Salt and IHD

The strange result is that the correlation in North America seems positive despite the fact that the sodium coefficient for the linear regression is negative (see on figure 8). We must also note that there too few countries to make a generalization.

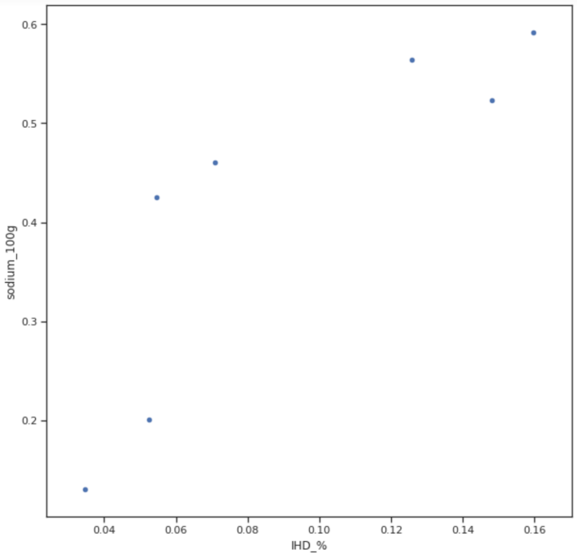


Figure 8: Correlations between Salt and IHD for North America

3.3 Diabetes and obesity : Hard to find a causality

We decided to look at the obesity and the diabetes prevalence on the world and linked them with available features of our dataset. We could expect a high correlation between these two diseases and the sugar metric and maybe the fat metric for the obesity prevalence. Unfortunately as we can see on figure 9 & 10 there is no real correlation between the chosen features and diseases.

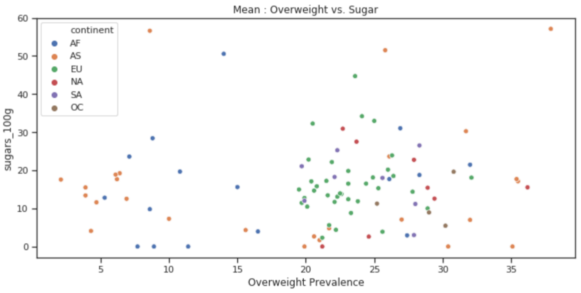


Figure 9: Correlations between Obesity and Sugar

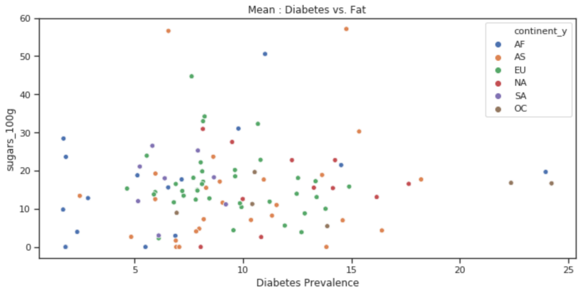


Figure 10: Correlations between Diabetes and Sugar

We do not really know why this is the case, but we can make a strong assumption for diabetes. The data we downloaded from the International Diabetes Federation (Internation Diabetes Federation, 2017) unfortunately does not differentiate between diabetes type 1 and type 2. Type 1 diabetes is not linked to a person's alimentation, since it is usually a genetic disease. It would be interesting to separate the two types to get more relevant results. Unfortunately, we came upon this realization too late to try to find more legitimate data on these diseases.

4 Conclusion

Our main dataset contained many entries but as we said previously, there were a lot of missing features. A significant portion of our working time was spent cleaning the data from various data sets and merging the sets together in order to put them to good use. We extracted a lot of metrics and computed many procedures to try to expose characteristics, using basic statistical methods such as Pearson and Spearman correlations, p-values, scatter plots and linear regression. It was challenging to get relevant results from our data, though the absence of results is a result in itself. Nevertheless, it was difficult to summarize all of the results in a single report. We encourage you to take a look at our notebook (Project repository, 2018). You will find all of our results for a few more metrics and better understand the context of this work.

References

Open Food Fact : L'information alimentaire ouverte https://world.openfoodfacts.org/data

Mortality from ischaemic heart disease by country, region, and age: Statistics from World Health Organisation and United Nations https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3819990/

Share of adults who are overweight https://ourworldindata.org/obesity

IDF Diabetes Atlas - 8th Edition http://diabetesatlas.org/resources/2017-atlas.html

IP2Location Country Multilingual Database https://www.ip2location.com/

JSON File for folium https://github.com/parulnith/Visualising-Geospatial-data-with-Python/blob/master/world-countries.json

Project repository : What is in my food? https://github.com/baraschi/ada_project_bamash