Scraping popular online cooking recipes to measure the impact of diet on modern diseases around the world

Romain Caristan (romain.caristan@epfl.ch)
Manuel Faysse (manuel.faysse@epfl.ch)
Rafael Rolli (rafael.rolli@epfl.ch)

Abstract

In our present day society, the amount of variety we are offered when buying food products is unfathomable, and it can become difficult to make conscious and healthy changes to our regime. The first step towards a better lifestyle is acknowledging the impact that nutrition has on our health and our project is centered towards this goal. Using datasets containing cooking recipes from around the world and World Health Organization diseases data, we first associate a nutritious score (NS) indicating the healthiness of each recipe. The NS score is calculated following the equation recommended by the french government during the health law reform of 2016. We then attempt to compare nutritional habits with population health statistics from countries around the world, such as obesity and diabetes to study potential correlations.

1 Introduction

Noncommunicable diseases (NCDs) are the leading cause of death and responsible for an estimated 72 % of the worlds 54.7 million deaths in 2016 (WHO, 2018). Modifiable risk factors such as unhealthy diet, physical inactivity, tobacco use and harmful use of alcohol are major causes of these diseases. Dietary saturated fatty acids and trans-fatty acids are of particular concern as high levels of intake are correlated with increased risk of cardio-vascular conditions. To obtain data provided by popular recipes around the globe, we will start by HTML scraping a cooking recipes website. Subsequently, we will introduce the notion of nutritious score (NS) which allow us to classify any given recipe, given its fat, protein, fiber, sugar and vegetable content, into a simple numerical or letter score in order to determine its healthiness level. We will proceed by using these informations to explore how the different recipes from each country can be clustered by ingredients and cooking style. Finally, we will clean, organize and manipulate the available datasets from the World Health Organization to discuss the prevalence of different diseases in the population and correlate these findings with either a combination of health-defined factors or the factors alone.

2 Data Acquisition

All of the scraping, processing and graphing methods that are briefly described below have code and detailed explanations contained in the project's Jupyter notebook.

2.1 Health statistics dataset

To study the impact of food on health, we chose to focus on overweight and obesity percentages, as well as statistics on diabetes and cardio-respiratory diseases in each region of the world. To that end, we downloaded and treated the public health datasets provided by the World Health Organization.

2.2 Recipe Dataset

The recipe dataset used during this project was generated through a web scraping program we developed. For each international cuisine section of the allrecipes website, it detected the 20 most relevant recipes and scraped the recipe data. The top 20 recipes always seemed like varied and relevant recipes upon inspection, and we felt that though they may not be fully representative of a country's diet, they could at least give relevant indications about some of the eating habits and patterns. This allowed for the dataset to contain the name, the region, the nutritional values, the user rating, as well as the list of all ingredients and quantities needed for each recipe. We then used dictionary matching based approaches to extract the numerical values of all quantities and make sure they were consistent throughout the data. This was simple at times (100mg to grams: 0.1), but more complex in others. For example, to estimate the mass of 1/8 banana, we first detected the fraction and converted it as a floating point number, then we recognized the fact there was no unit associated to this ingredient, so we used a dictionary to match 'banana' with its average weight, before multiplying it with the computed 0.125. It was also useful at times to convert units from imperial to metric units. The final dataset contained a bit under 400 recipes from 18 regions of the world.

2.3 NS calculations

NutriScore as described earlier can be computed given the official french government equation. We compute the negative impact points by penalizing the amount of Calories, sugars, saturated fats and sodium contained in 100 grams of the recipe according to a table and summing them. We then calculate the positive impact points by accounting for proteins and dietary fibers per 100 g, as well as the percentage of weight coming from vegetables. We determine if an ingredient is a vegetable or not by checking if the ingredient contains one element of the vegetable and fruit list we scraped from yet another website. We then substract the positive impact points from the negative ones to get the final numerical nutriscores, which will then be associated to ranges of letter scores (from A most healthy to E least healthy). By analyzing the results,it is possible to verify that the lowest NutriScores in fact correspond to the recipes that seem the healthiest (*Moroccan Lentil Soup, Greek Pasta with Tomatoes and White Beans*, etc) and inversely, the highest Nutriscore do seem to make sense (*Cali's Sinful Creme Brulee , Salt and Pepper Squid, Chef John's Patatas Bravas*, etc)... Nutriscores from recipes in the dataset range from letter grade A to D (-8 to 19), with a mean score of 0.72, and median 1 (B). Histograms of the letter scores for every region are drawn in the notebook, clearly showing that Lebanese and North-African recipes tend to be the healthiest while Canadian and Scandinavian recipes include much more unhealthy alternatives (Figure 1).

2.4 Cooking Similarities and Centralness

In order to go further than simple mathematical analysis, we used natural language processing to find similarities between the ingredients used in each country's culinary traditions. Using Google's Universal Sentence Encoder based on a Deep Averaging Network model that we ran with the Keras API for Tensorflow, we encoded the total ingredient list of each country as a 512 feature vector, usable to determine semantic similarities. We then computed a similarity (correlation) matrix (Figure 2) filled with the dot products of each region's vector with each of the other region's vector.

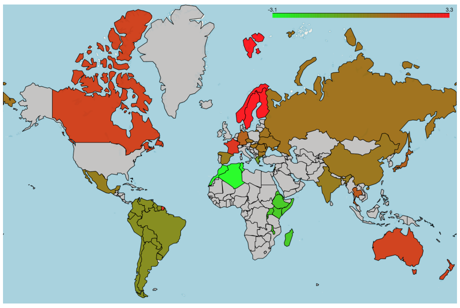


Figure 1: Mean Nutriscore of top 20 recipes per region (red is more unhealthy)

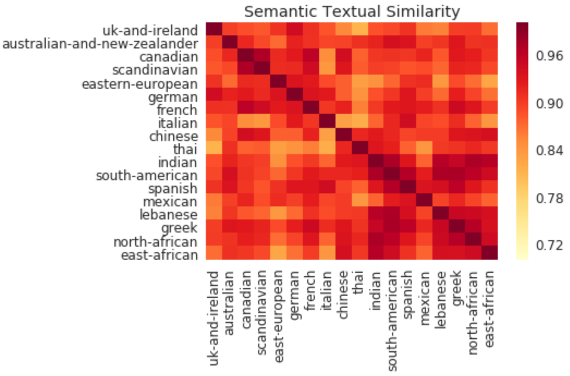


Figure 2: Several trends are observable. Mediterranean cuisine (Greek, North African, Lebanese, Spanish) are highly correlated, and this is also true for Latin cooking (Spanish, South American, Mexican), northern cuisine (Canadian, Scandinavian), as well as the German, eastern European and UK group. Inversely, Thai food is very different from most of the previously cited regions, and eastern European is far from Indian or Mediterranean cuisine.

A measure of centralness was also calculated as the sum of the centered distances and indicates Greek and South-American ingredients to be the most central. Indian ingredients surprisingly tend to be pretty similar to those of Latin and Mediterranean regions.

3 Results

3.1 Country clusters by nutritional values

Several attempts were made to cluster countries based on their nutritional values. Different machine learning algorithms such as PCA (Figure

3), K-Nearest Neighbor, DBScan, decision trees were attempted to try and find means of clustering data-points according to their respective class and thus being able to find similarities between regions. However, no significant result was found... This was to be expected since every country has very varied recipe types and cooking styles, but it is interesting to note that it is not achievable to find links between regions cooking styles in this way.

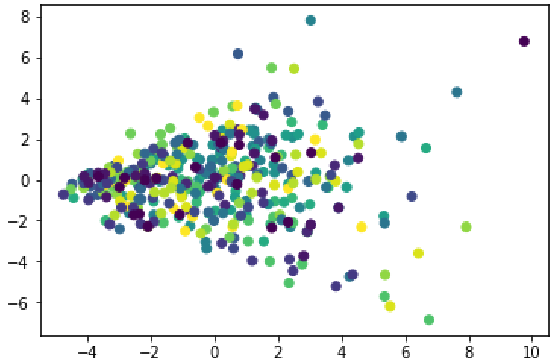


Figure 3: Principal Component Analysis is performed on the normalized numerical values of the nutritional values of every recipe. The projection of the high-dimensional data on the two first eigenvectors show no possible separation of data according to their class.

3.2 Health correlations

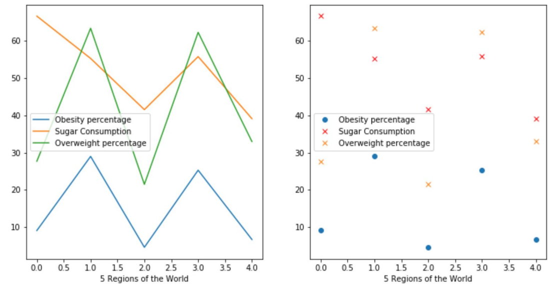
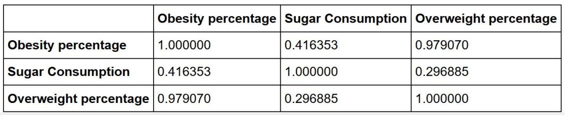


Figure 4: Linear (left) and scatter (right) plots representing the consumption of sugar by the world's continents. In addition, overweight and obesity prevalence in percentage are also shown.

We begin to study the influence of sugar consumption on health-related markers as shown on figure 4 where we observe a linear and scatter plot of the consumption of sugar by world continent and it's punctual representation by a scatter plot. There is no real correlation between these variables as seen on figure 4 where the Pearson correlation is shown. A value of p = 0.42 between obesity and sugar consumption is a good indicator however we do not possess enough data points to draw any conclusions. As a blank measurement for the validity of our results, we notice the almost perfect correlation (p = 0.98) between overweight and obesity percentages.(Figure 5)



	Obesity percentage	Sugar Consumption	Overweight percentage
Obesity percentage	1.000000	0.416353	0.979070
Sugar Consumption	0.416353	1.000000	0.296885
Overweight percentage	0.979070	0.296885	1.000000

Figure 5: Correlation table between the different variables.

We proceed to study how the protein content and ratio of saturated fats to total calories affect the same health markers. (Figure 6)

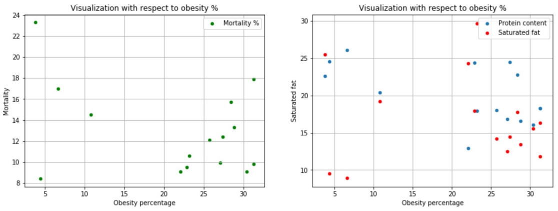


Figure 6: Left: Mortality vs Obesity plot, serves as validation measurement. Right: Saturated fat and protein content vs Obesity percentage per region of the world.

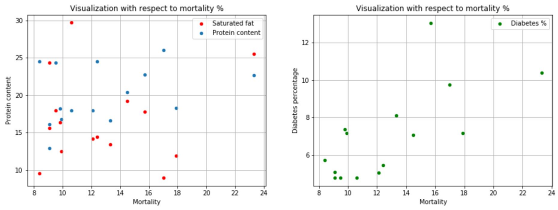
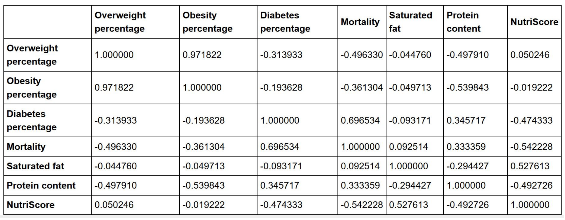


Figure 7: Left: Saturated fat and protein content vs Mortality percentage per region of the world. Right: Mortality vs Diabetes prevalence plot, serves again as validation measurement.

In both figure 6 and 7 we plot two of the available health-markers against each other that serve as validation measurement. On the other plot we see that no clear clustering or possible regression line fit on either of the figures. However, we would be able to fit a regression line on figure 6 (left). It is important to notice that the diabetes and mortality would be expected to have a high positive correlation however the Pearson coefficient value is relatively low (Figure 11). We will further discuss this phenomenon under the Discussion section. In general, we cannot draw any strong conclusions from these measurements due to a lack of data and generalization of the health-related markers by country that do not take into account several other factors that affect the prevalence of these diseases on every region of the world.

In figure 9 we visualize the mean value of the nutriscore with respect to obesity, diabetes and mortality for the different regions of the world.



	Overweight percentage	Obesity percentage	Diabetes percentage	Mortality	Saturated fat	Protein content	NutriScore
Overweight percentage	1.000000	0.971822	-0.313933	-0.496330	-0.044760	-0.497910	0.050246
Obesity percentage	0.971822	1.000000	-0.193628	-0.361304	-0.049713	-0.539843	-0.019222
Diabetes percentage	-0.313933	-0.193628	1.000000	0.696534	-0.093171	0.345717	-0.474333
Mortality	-0.496330	-0.361304	0.696534	1.000000	0.092514	0.333359	-0.542228
Saturated fat	-0.044760	-0.049713	-0.093171	0.092514	1.000000	-0.294427	0.527613
Protein content	-0.497910	-0.539843	0.345717	0.333359	-0.294427	1.000000	-0.492726
NutriScore	0.050246	-0.019222	-0.474333	-0.542228	0.527613	-0.492726	1.000000

Figure 8: Correlation table between the different variables.

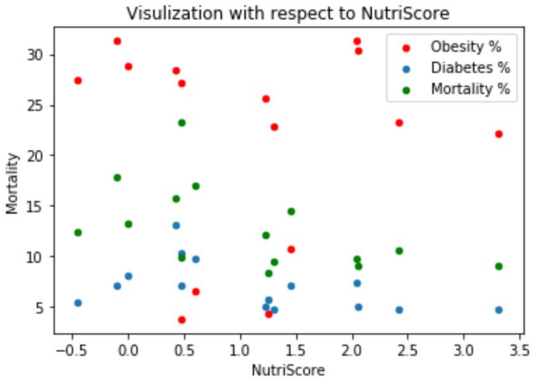


Figure 9: Obesity, Diabetes and Mortality rates plotted against the mean value of NS per region

Again the values appear to be distributed over the whole range.

We decided to quantify the influence of two variables at the same time. In figure 10, we visualize the influence of a high consumption of saturated fats and sugar on obesity, diabetes and mortality. There are a couple of extreme values, specially on the lower spectrum of obesity prevalence that are due to the fact that some countries (African and Asian) possess a lower percentage of obese people in their population not because of a healthier diet but because of poverty issues, this will be discussed further.

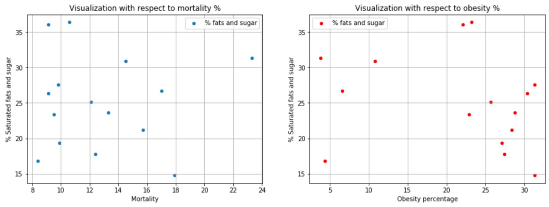
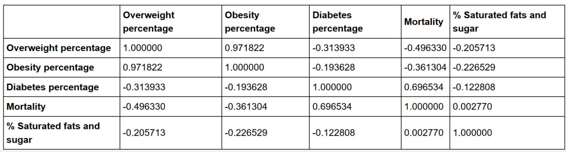


Figure 10: Plot of the compound effect of a diet higher in saturated fats and sugar with respect to mortality and obesity prevalence on the different regions of the world.

4 Discussion

The different analysis we conducted lead to show it is difficult to truly find correlations between traditional recipes found online and health markers in different regions of the world. This may be true because we do not have enough information about the true diet of the population, and we merely tried to estimate the eating habits in function of the region's most popular recipes. Another important factor to mention is the fact that the most *relevant* recipes for each region that the website provides target mainly US residents, whose diet may differ from each region's natives. For example, there is no clear evidence that higher consumption of sugar found in the most popular dishes of different regions play a role on obesity or the overweight percentage in the same region.



	Overweight percentage	Obesity percentage	Diabetes percentage	Mortality	% Saturated fats and sugar
Overweight percentage	1.000000	0.971822	-0.313933	-0.496330	-0.205713
Obesity percentage	0.971822	1.000000	-0.193628	-0.361304	-0.226529
Diabetes percentage	-0.313933	-0.193628	1.000000	0.696534	-0.122808
Mortality	-0.496330	-0.361304	0.696534	1.000000	0.002770
% Saturated fats and sugar	-0.205713	-0.226529	-0.122808	0.002770	1.000000

Figure 11: Correlation table between health markers and compound effect of saturated fats and sugar.

Is the nutriscore a relevant indicator of health in countries around the world? In a nutshell, no, or at least, not the nutriscore of the most popular recipes given by an online cooking website. It is clear that a healthy diet and regular exercise are essential for a long, fulfilling life. It is however not always the quality of the food and its macro-nutrients that determine health-related conditions all over the world, socio-economic impact on health is extremely significant. Several regions of the world have a low obesity and low consumption of sugars and saturated fats however their mortality is on the other side of the spectrum. Regions of the world, namely Africa, Asia and South America suffer from extreme poverty and food is not available to them. It is thus, very tricky to get any interesting correlations from these datasets, socio-economic gap and the unequal access to great health infrastructures are too important to be ignored. A further study should probably consider to add this data, or to focus on regions with similar industrialization levels. However, without clear correlations between recipe data and health markers, some interesting conclusions can still be obtained. The case study of Canada (Appendix A) shows interesting information of how high consumption of fats and sugars, and low quality of food (measured by the nutriscore) only has a limited impact on resident's health, when comparing the Canadian health marker data to the rest of the world's.

References

Allrecipes *World Cuisine* https://www.allrecipes.com/recipes/86/world-cuisine/

Google *Universal Sentence Encoder* https://tfhub.dev/google/universal-sentence-encoder/2

Global Health Observatory. *Prevalence of obesity among adults, BMI 30, crude Estimates by country.* http://apps.who.int/gho/data/node.main.BMI30C?

Global Health Observatory. *Prevalence of overweight among adults, BMI 25, crude Estimates by country.* http://apps.who.int/gho/data/node.main.BMI30C?

International Diabetes Federation, Diabetes Atlas. *Diabetes prevalence (% of population ages 20 to 79).* https://data.worldbank.org/indicator/SH.STA.DIAB.ZS

Project File *Jupyter Notebook with data scraping and treatment* https://github.com/ada-food-recipes/ada-food-recipes.github.io/

SANTE PUBLIQUE FRANCE *Nutri-Score* https://www.santepubliquefrance.fr/Sante-publique-France/Nutri-Score

World Health Organization *Saturated fatty acid and trans-fatty intake for adults and children* https://www.who.int/nutrition/topics/sfa-tfa-public-consultation-4may2018

A Case Study: Canada

To illustrate how the food recipes from a well-established and developed country affect health markers, we decided to prepare a case study about Canada. The nutriscore, fat and sugar consumption will be taken into account to find potential correlation for our study in a country where the socio-economic impact on health is not affected by a high number of poor people.

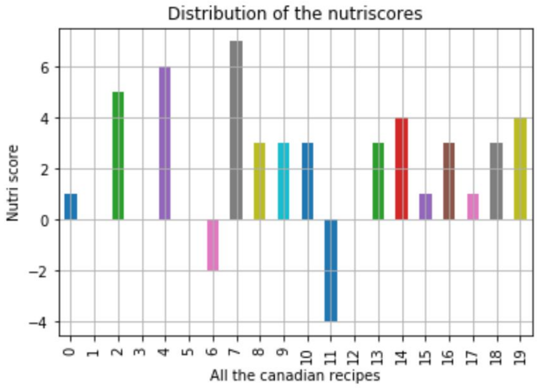


Figure 12: Distribution of the nutriscore for all Canadian recipes available in our datasets.

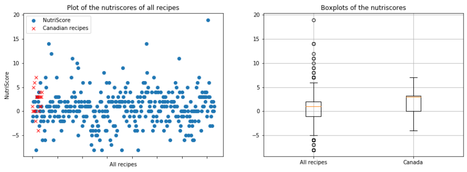
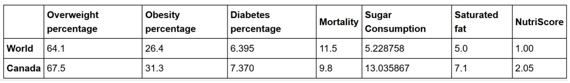


Figure 13: Visualization of all the nutriscores of all countries available in the dataset against the Canadian (in red).

It is not possible to predict health-related markers with high level of accuracy with our data. However, we observe that Canada has a high consumption of saturated fat and sugar (well above the median), its nutriscore (remember: the lowest it is, the better and healthier the food is) shows an unhealthy trend and above all, the health-related overweightness, obesity and diabetes markers are also above the median as presented on figure 14. This simple but insigthful case study shows that if we wanted to predict at a high level of accuracy the health conditions of a country from its traditional food/recipes we would need also a socio-economic study to eliminate such factors and focus exclusively on the correlation between the quality of available food and different health conditions. As an example, China has a small consumption of sugar and saturated fats but the fact that the socio-economic atmosphere is not the best, affected our studies. Some of the others African and Asian countries presented high levels of sugar consumption but low obesity due to the fact that poverty is very common in these regions.



	Overweight percentage	Obesity percentage	Diabetes percentage	Mortality	Sugar Consumption	Saturated fat	NutriScore
World	64.1	26.4	6.395	11.5	5.228758	5.0	1.00
Canada	67.5	31.3	7.370	9.8	13.035867	7.1	2.05

Figure 14: Correlation table and comparison studies between world and Canada.