Are we what we eat ?

Lucas Délez
lucas.delez@epfl.ch

Stanislas Furrer
stanislas.furrer@epfl.ch

Luis De Lima
luis.carvalho@epfl.ch

Yura Tak
yura.tak@epfl.ch

Abstract

This project aims to study the food inspections in Chicago. Our objective was to try to determine in which areas of Chicago, it is safer to eat. Or at least in which areas, we have less chance to encounter sanitary issues. We did correlate the restaurants present in the dataframe with their geolocalisation in order to give each zone a score.

1 Introduction

We chose to work on the Chicago food inspections dataset for this project provided by the *Chicago Department of Public Health*. This dataframe consists of a list of inspections from 2010 to present with entries such as the outcome of the inspections, the violations committed by the facility, the date of the inspection, the geolocalisation of the place and so on... The inspections concerns different types of facilities like restaurants, schools and even hospitals. Each inspection results in a grade associated with a risk such as 'low risk', 'medium risk' and 'high risk' which are determined by the kind of violations the facility committed. Our original working hypothesis was to try to assess which locations were associated with which type of food and if we could find correlations between the locations and the quality of life in the area. Finally we decided to look at which zone was associated with a higher risk and how it changed overtime.

2 Data Collection

The dataframe is available at kaggle. We used an additional dataframe to get *geojson* data about the areas of Chicago (it can be found here).

3 Dataset Description

Before analysing the data, we played with the dataframe trying to find what we should clean. As it was stipulated on the site providing the dataframe, most of the data was already cleaned. We found mostly names that needed to be standardized (e.g. McDonalds/Mc Donald's).

3.1 Date cleaning

As we could not use the original format directly, we had to choose between converting the dates to *datetime* or converting them to a chosen string format. We decided to keep only the years of the inspections because we considered that it was enough for our future analysis.

3.2 Facility cleaning

We then looked at the facility types provided by the dataframe. We had as an idea to focus our analysis on the facility types and keep only the restaurants, schools and hospitals because we thought that the hygiene in those facilities could reflect on the life quality in the area. The first constatation that told us that it was not the best idea is that the schools were not only public elementary schools but could be any type of learning place. We also realised that the sanitary inspections did not concern the general state of the hospital but were only based on food quality. For those reasons, we chose to focus only our analysis on the restaurants. Even though this decision reduced the size of the dataframe quite dramatically, it was still quite large. In further analysis, we also did try to constitute a dataframe regrouping only the biggest chains of restaurants present in the dataframe such as McDonald's or Subway.

3.3 Violations

From the resumé we could find on the site, we understood that the violations are associated with specific risk factors and that there should be 45 types of violations. The violations labeled 1 to 29 were associated with low risks and resulted in a 'pass' from the inspections. The result 'pass-with-condition' was attributed to facilities having committed a serious violation but that was corrected during the inspection. The violations resulting in a 'fail' are the ones that could not be corrected during the inspection but this does not necessarily mean that the establishment is closed. Of course, an inspection could result in the recognition of more than one violation in the same facility and on the other hand of none. Interestingly, the labeled of the violations we found on the site was wrong and gave no idea of the importance of the violations. We decided to sort the violations and see which ones were truly responsible of the different outcomes. Another difficulty was that the violation system changed in 2018.

3.4 Inspection Analysis

We analysed the inspection count per facility type and per chain of restaurant. We observed a diminution of the inspections for all facilities types and for the big chains of restaurants. The most frequent inspection type observed is the canvass. For restaurant facilities, we observe that most of inspections are related to complaints especially in McDonald's.

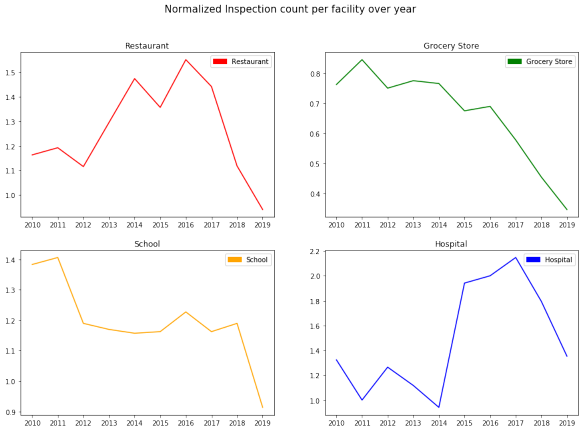

Figure 1: Normalised Inspection count per facility over the years. Red for restaurants, yellow for schools, blue for hospitals and green for grocery stores

3.5 Risk Analysis

We looked at the risk distribution over the years. As you can for example in figure 3.

We can see that we only have a single entry for low risk in the map above which means these chains are strictly rated. There is some kind of pattern that seems to appear in the distribution of medium risk and high risk restaurants. For example, more medium risks by the lake and more high risks by the suburbs. Some of the restaurants are simply fast foods with bad rating independent of location, (i.e. Subway). All restaurants are not present in every community which means that we are not taking into account high quality restaurants that will be only present in certain locations. For a better study, we should look at a more complete dataset where we can find more types: restaurants (but all of them), grocery store, schools and hospitals.

4 Methods

Taking all this information into account, we decided to focus on the attributions of risk scores on different Chicago areas overtime. We used an additional dataframe (see 2) to get the *geojson* shapes of the zones. This allowed us, using the longitude and latitude coordinates present in the food inspection dataframe, to situate in which zone of the second dataframe every facility is present. To calculate the risk score for each community or area we used the following formula:

$$\frac{High + Medium * 0.5}{High + Medium + Low} \tag{1}$$

We considered this formula to be interesting because it should give us the risk associated with sanitary issues in each area. In theory, a community that only has low risk restaurants will have a score of 0 (the lowest possible risk) but a community that only high risk food facilities will get the maximum score of 1.

You can find the resulting heatmaps in the figure 4 below.

5 Discussion

We noticed that using only the top five restaurant chains wouldn't give us a real score representing each area because most of their facilities share the same status and quite incredibly, those important chains are quite comparable in their scores. We decided to filter our data by picking only once every restaurant in an area. This gives small restaurants, those that are not part of a big chain, a heavier weight in our score calculation. To validate this approach we will also plot the variance and mean of the risks of these restaurant chains (figure 5).

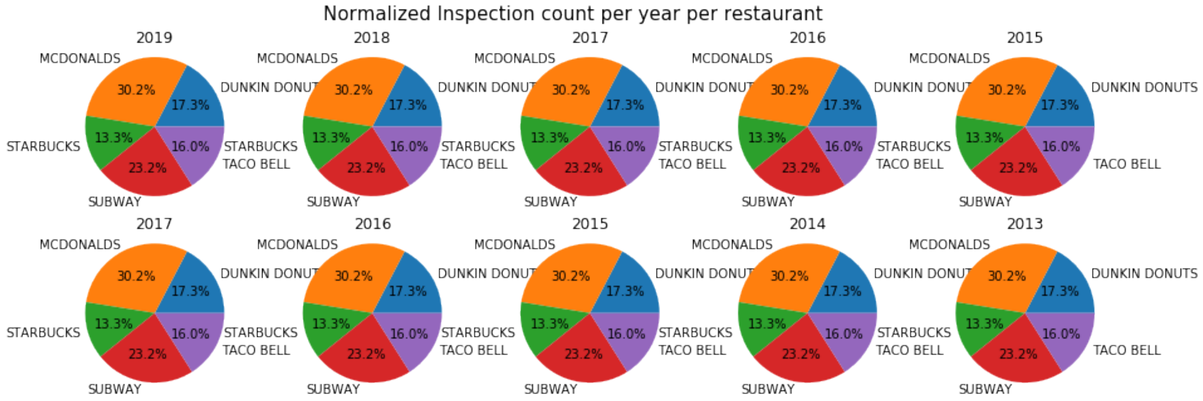

Figure 2: Normalised Inspection Count per year per restaurant. We see the relative evolution of inspection over the years.

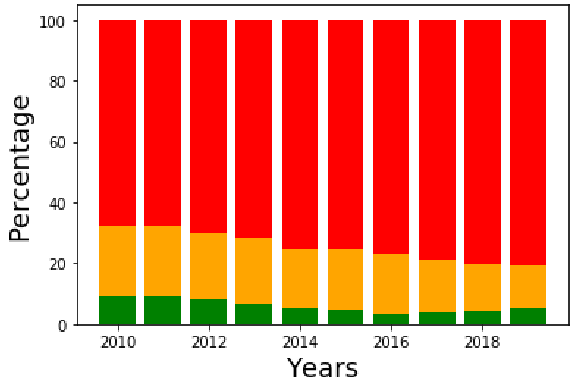

Figure 3: The evolution of the three levels of Risks over the years. As the choice of the colors suggest, red is associated with a 'high-risk', yellow with 'medium-risk' and finally green with 'low-risk'.

We see in figure 5 that big restaurant chains, such as *Subway* or *Dunkin Donuts*, can have a lot of restaurants in the same community and that the variance of risk between those restaurants are very low. To study the risk of each area, we should consider each restaurant present in the community and not be biased by the risk associated with those big restaurants chains. So we decided to consider those big restaurant chains, which are largely present more than once in each community, as one single abstract facility with a rounded normalized risk.

6 Conclusions

You can find the final map at figure 6. We observe that the zones with higher risk scores tend to be in the best places for fast foods (downtown, business areas, highest frequentation areas, tourism oriented places) and that the lower scores are concentrated further from the heart of the city in more residential areas. We see also a drastic color change for a very small area in the middle of zones with lower risk scores, this kind of result may be ignored due to the fact that it is too small so it only needs to have for example a *McDonald's* or a *Burger King* to have its risk score heavily raised. The score system is not perfect but can definitely discriminate different areas and give an idea of where you can find the main areas of the city like the most visited ones, the residential zones, touristic parks, suburbs.

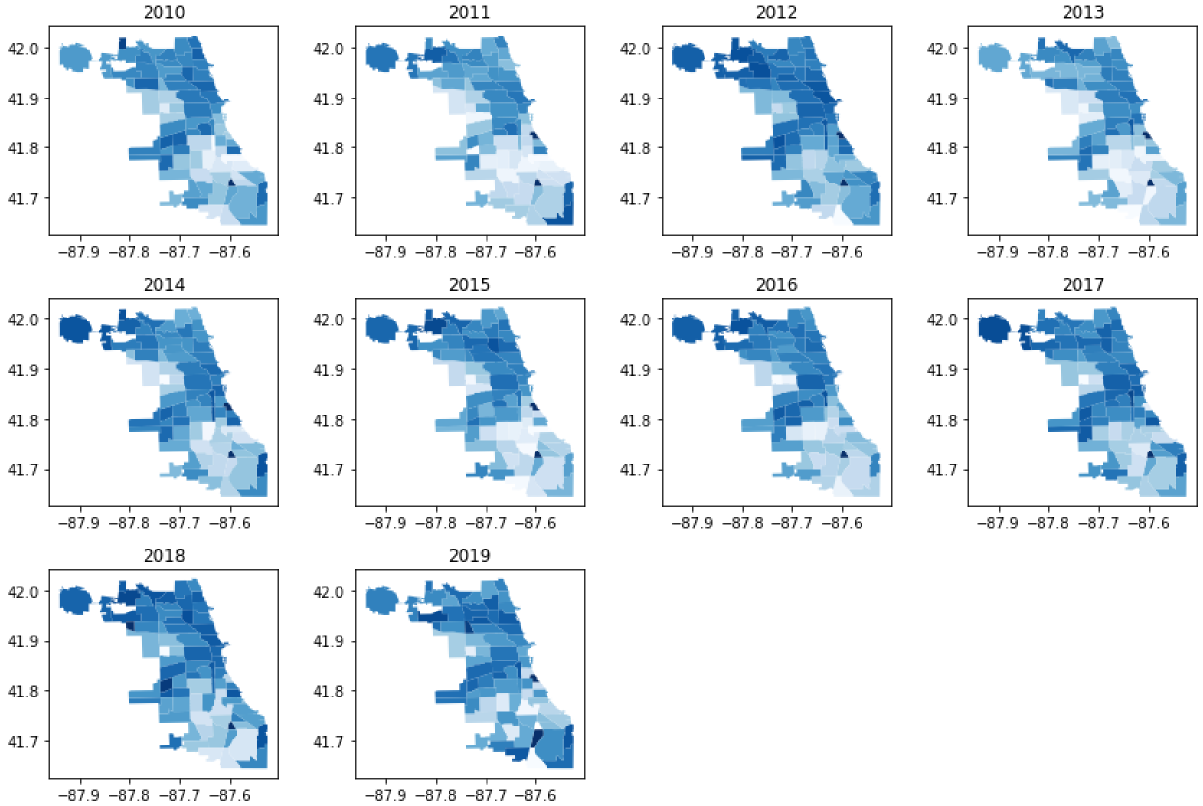

Figure 4: Risk score associated with each year. You can observe the evolution of the score over the years. Deep blue corresponds to the maximum score while white is associated with a low score.

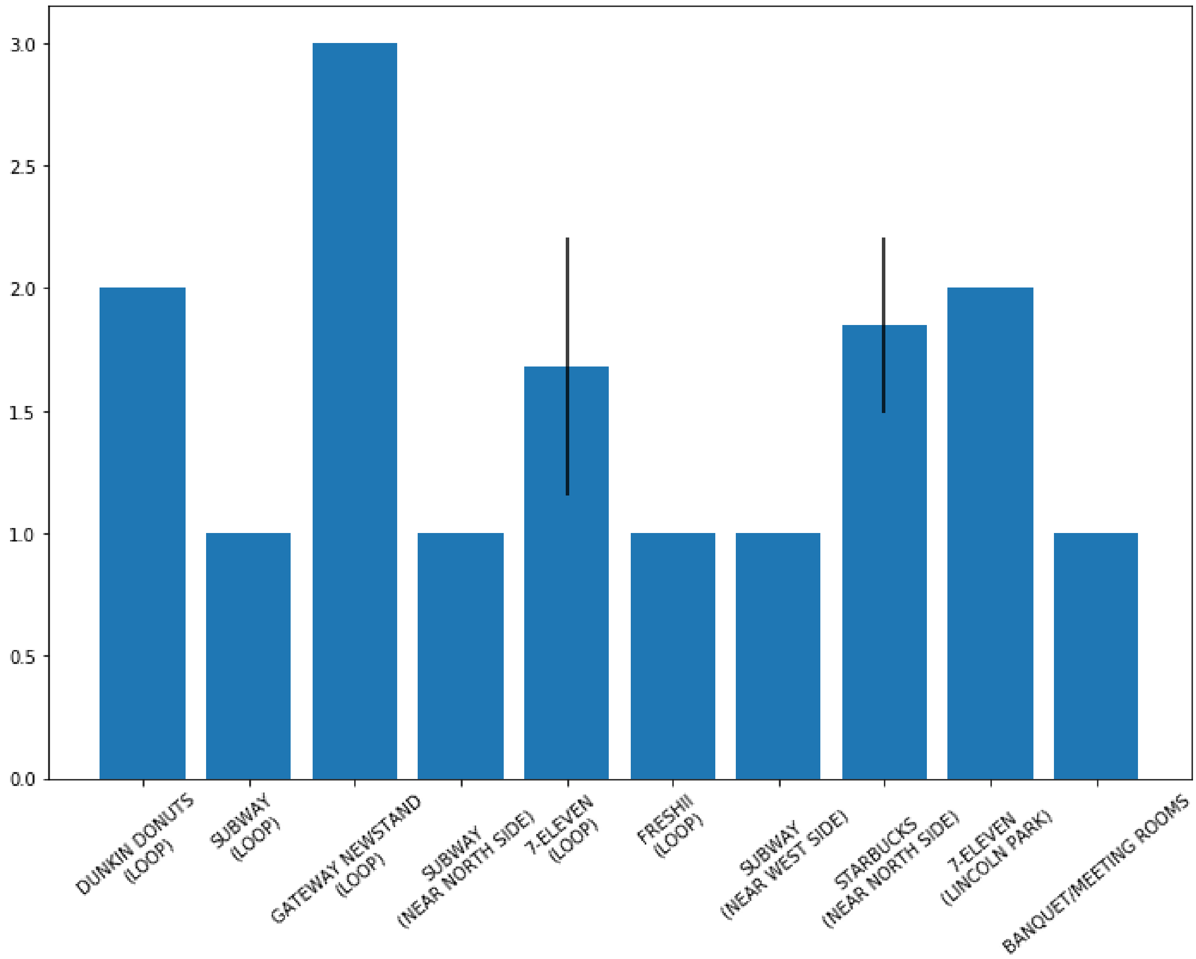

Figure 5: Risk associated with the restaurant chains that are the most present in a single area. You can find the name of the restaurant with the name of its community below.

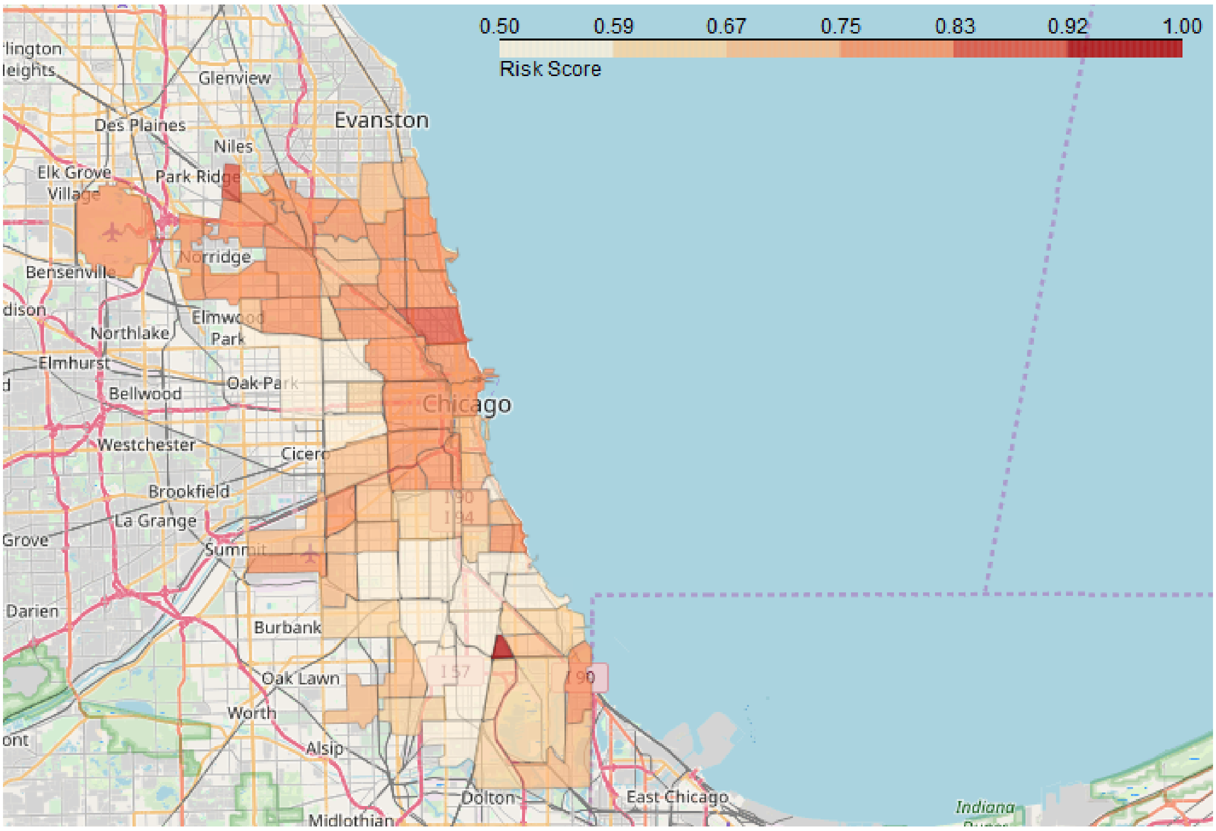

Figure 6: Final Heatmap showcasing the new ranking system. Screen capture of the follium heatmap you can find in our jupyter.