Fast-food presence and Criminality in San-Francisco

1. Introduction: Business Problem

1.1 Background - As a product owner moving to San Francisco on early 2020, I have at heart to find a safe location for myself and my family.

1.2 Business Problem - This led to a sort of tongue-in-cheek idea: how about finding the correlation between criminality and what we could get out of Foursquare, a service absolutely not designed to return geopolitical and societal insights. This would serve multiple purposes:

Primary

- Getting to know the type of neighborhood with a glance at the venues in the streets.

Secondary

- Shouldering the idea that information is not data but what we get out of it.
- The difference between correlation and causation in action.
- Working with nonoptimal/indirectly correlated data for a case, as most real-life projects.
- Showing the importance of ethics and a way fake news could be born.

1.3 Interest - This simple exploratory study could benefit people searching for a safe place to live, commute and work in San Francisco with a quick glance at the neighbourhood. This could also come handy for realtors but the underlying purpose of it is to show how a bad use of data could result in damages for companies in terms of image and the proliferation of fake news.

Baptiste Marien - December 2019

2. Data acquisition and cleaning

2.1 Sources - To complete this task and validate our assumptions, we will use:

- The **venues database in Foursquare** (2019 data) [Source | Documentation]
- The **Census Tract boundaries** (2010 data) as one of our data, Average Income per Household is per census tract [Source | File], this is a cleaned up version of what could be found on the geo.census.gov website
- The **Average Income per Household** (2013-2017) [Source | File]
- The **Crime report CSV file** (2016) provided by IBM [File]
- The **Geolocation tool** provided by the Census Bureau will come handy [Tool | Documentation]

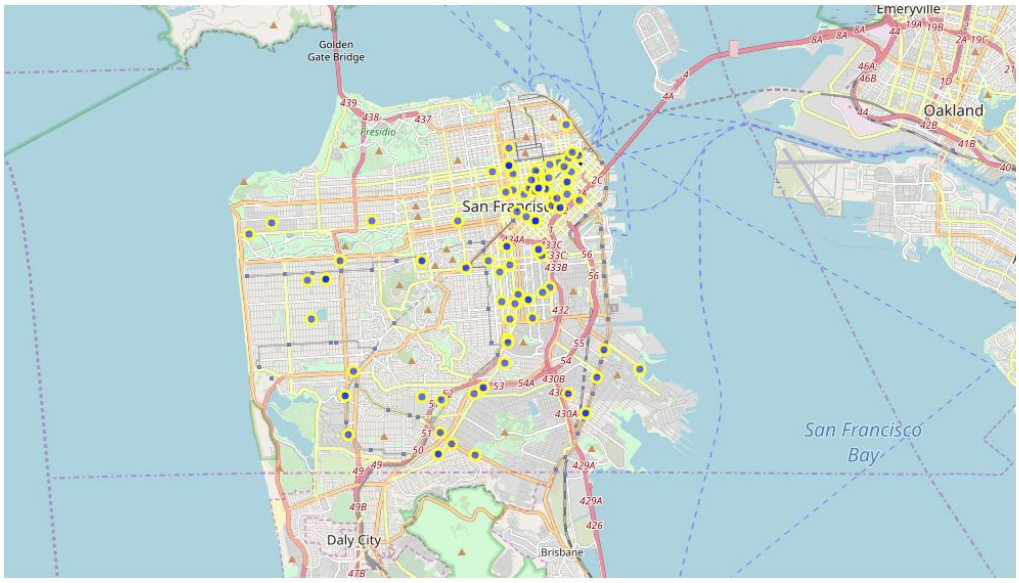
All of this to correlate crimes with a type of venue and plot it all on the map. Then we might want to find an inverse correlation with another type of venue. Finally we will compare it to a first hand correlation with crimes, namely a poverty index (reference) under the form of the Average Income per Household.

3. Methodology

The thick of this study lies here. It is all about aggregating data from different sources with different formats and coverage.

3.1 Acquisition - To complete this task and validate our assumptions. Starting with the crime report, each of the incidents will be qualified with a census block ID using the census tract boundaries and a point to polygon detection library (the geolocation tool could be used but at the risk of being URL banned) then grouped. We are doing so because our test sample is the average income per household that comes per census block. Using the same process, we will list all the venues we are interested in using a spatial coverage strategy to limit the 100 venues per request imposed by the Foursquare API.

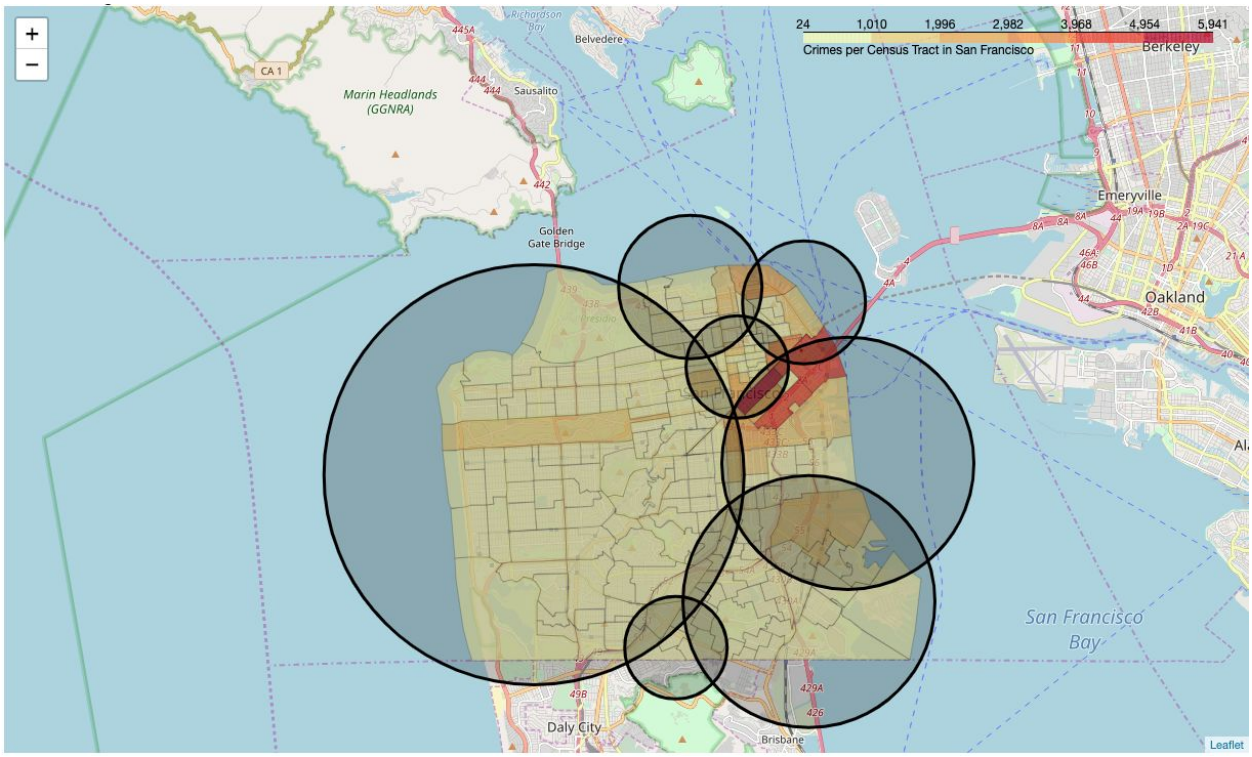
3.2 Crimes - Those came already cleaned, we just made a selection of which crime would be a direct threat or a degradation of the urban space. After dropping useless column, it became really easy to plot.



From there we just have to attribute each point to a census which could be done using the geolocation tool hosted on a .gov or making use of a bounding box detection library called Shapely.
A quick glance at the map of the crimes that came back without a census block attributed tells us that we can easily get rid of them as they are not part of our definition of San Francisco (there is no miss inside our census block map). We then have a number of crimes per census block.

3.3 Average Income per Household - Comes as is there was no special transformation except getting the freshest value per census block among the four years available in the report. This set of data is then merged to the crime dataset.

3.3 Fast-food and Yoga venues listing - The Foursquare API comes with two limitations: a usage limitation and a per request limitation. A good balance between the amount of queries and the radius of them was necessary. We could have packed the map with overlapping uniform circles but this would have been suboptimal for API usage. Instead we used this tool in conjunction with a few iterations to make sure that each circle returns less than 100 venues, but enough to be the largest possible and overlapping its neighbours. Here is a visualisation.



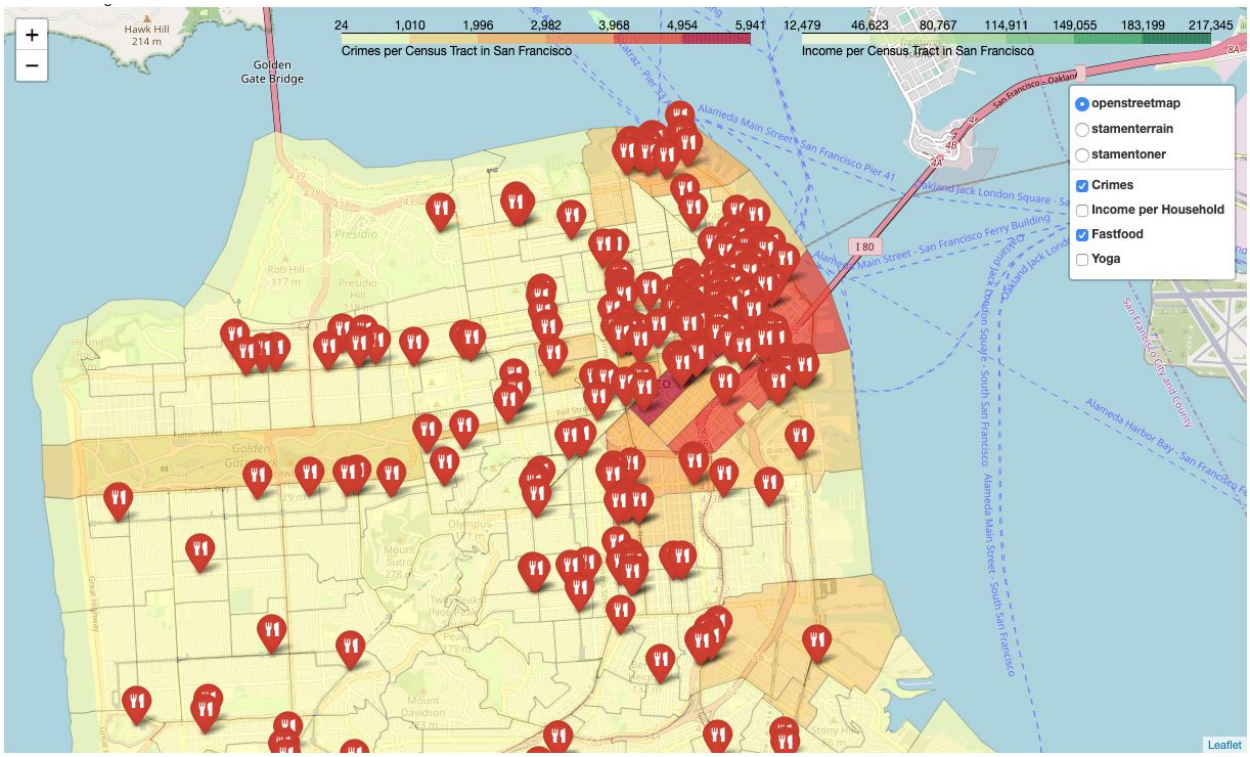
Cleaning came in two forms: removing the venues returned multiple times due to the overlap. This was solved by removing duplicates using the unique ID as index. The second step was to remove the venues out of our census block coverage. Attributing each venue a census block first then ditching the venues with no value did this exact job.

We then grouped each venue per type and per block as merging them to the crimes and incomes dataset. The exploratory analysis was then mostly done. Returning this interactive map.

Last bit consists of a correlation matrix, plotting for the most relevant index and correlation evaluation using the Pearson coefficient and the P Value.

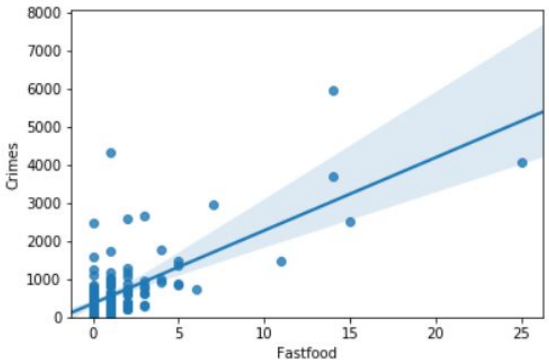
4. Results

4.1 Final Map - We could observe that the crime and income maps are not fully overlapping, the yoga ventures are actually scattered more or less in a homogeneous fashion but that the fast food ventures are packed in the zones of high crime frequency.



4.2 Correlations - Here is our findings:

	Crimes	Income	Fastfood	Yoga
Crimes	1.000000	-0.158367	0.708502	0.486440
Income	-0.158367	1.000000	-0.069147	0.133501
Fastfood	0.708502	-0.069147	1.000000	0.473737
Yoga	0.486440	0.133501	0.473737	1.000000



- Pearson for Fast-food venues presence and Crimes is 0.7085020108512083 with a P-value of P = 6.970930586185497e-31
- Pearson for Yoga venues presence and Crimes is 0.4864399942517627 with a P-value of P = 6.436520775576165e-13
- Pearson for Average Income per Household and Crimes is -0.15836724300092672 with a P-value of P = 0.027420251750343953

Since the p-value is < 0.001, the correlation between the presence of fast-food ventures and criminality within the same neighborhood is statistically significant, and the linear relationship is quite strong (~0.708, close to 1)

5. Conclusion

Our analysis shows surprising results for those data under those vintages for the city of San Francisco. Our tongue-in-cheek assumption happened to be verified, **there is a solid correlation between the density of fast-food ventures in an area and criminality level in San Francisco**. This is a correlation, not a causation.

More interesting is the fact that for San Francisco, applied to the census blocks the **Average Income per Household shows no obvious inverse correlation**. Criminality looks completely unrelated. This is very surprising as this was our gauging index.

Also the presence of **yoga studios shows a very mild correlation with criminality**. This is a plain proof for a **wrong assumption** showing that certainly population density is a stronger factor than population type in the business plan to establish those shops.

6. Discussion

While we found something odd about the city of San Francisco, we could only conclude that this city is not typical. Which is certainly why it is the subject of many studies.

Back to our goals:

Primary

- **For a short span of time, if both the economic and societal profile of San Francisco does not change as well as most Fast Food chain business model remains consistent, a high concentration of those could translate in a potentially more dangerous area**. But this is a second hand indicator and could be only valid at the time of the study.

Secondary

- Shouldering the idea that information is not data but what we get out of it. This is the way we decided to work our data. Our data sources are not absolutely leading in any case to those solutions. A different type of shop used, a different approach or territorial partition would provide **different information with the same data sets**.
- The difference between correlation and causation in action. Are happy meals coming with concealed firearms? No.
- Working with nonoptimal/indirectly correlated data for a case, as most real-life projects. Our results could be absolutely biased because we chose to work on non optimal data. Far from the origin of our problem. **Not crossing information in that case can't lead to an actual firm conclusion but at most to data oddities and peculiar data or mathematical facts**.
- Showing the importance of ethics and a way fake news could be born. The most important conclusion in the end is that ethics are the key factor in data analysis. People believe in the power of data, reports and images based on official reports and datasets labelled ".gov" but it is surprisingly easy to cherry pick what conveniently concurs to what we want to demonstrate, true or not. **From data selection, cleaning, parsing, methods to interpretation, there are so many ways of stamping a fake/inaccurate information as legit**.