Applied Capstone Foursquare Final Project: Exploring crime data in neighbourhoods in Mexico City

K. Damián E.

City dynamics, a new standpoint that crime prediction could take advantage

- Currently, the crimes are not just solved, also could be predicted. This can be useful for authorities, enthusiastic data scientists, and investors to some business.
- Explication to "...Some neighbourhood is very popular, what are people looking for? What is in there that is not anywhere else? Is there a route with high-density people?"
- How the human mobility across a city affects the crimes? Or vice versa?
- All of above, how is it in Mexico City (aka CDMX)?

Data sources, downloading and cleaning

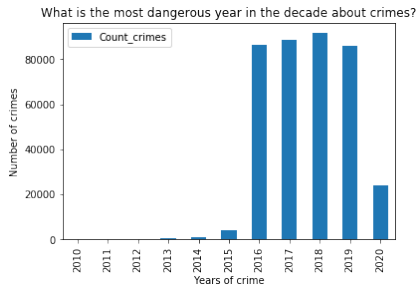
- The dataset was obtained from the **Open Data Mexico City**. In this case it was chosen the *Investigation's folders of CDMX's General of Justice Prosecutor's Office*, then to chose the filters of interest and download the suitable dataset. Crime dataset CDMX.
- The raw dataset has records since beginnigs of XX century. It was preserved the period of 2010-2020.
- The features of interest are: Year, Category of crime, Crime, Weekday, Hour, Borough, Neighbourhood, Latitude and Longitude.

Methodology

- It will be observed the relationship between temporal and geographic features respect to number of crimes.
- Once obtained the most dangerous CDMX's borough, it will use the Foursquare API, in order to study the city dynamics.
- It will arrive K means learning, because it seeks to find a possible structure of data.

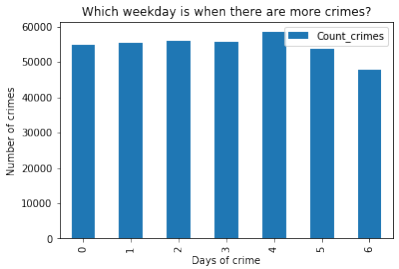
Exploratory analysis

Analyzing the data of all CDMX, it was got that

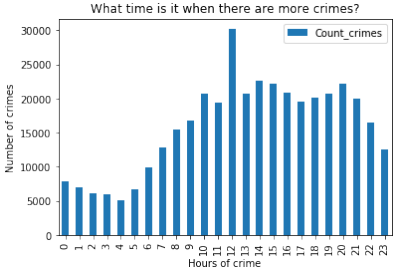


(a) The most dangerous year according to complaints was 2018.

(b) The most dangerous mont is March.



(c) The most dangerous weekday is Friday.



(d) The most dangerous hour is noon.

Figura: Bar charts of crime's numbers vs temporal features

Geographical features and number of crimes

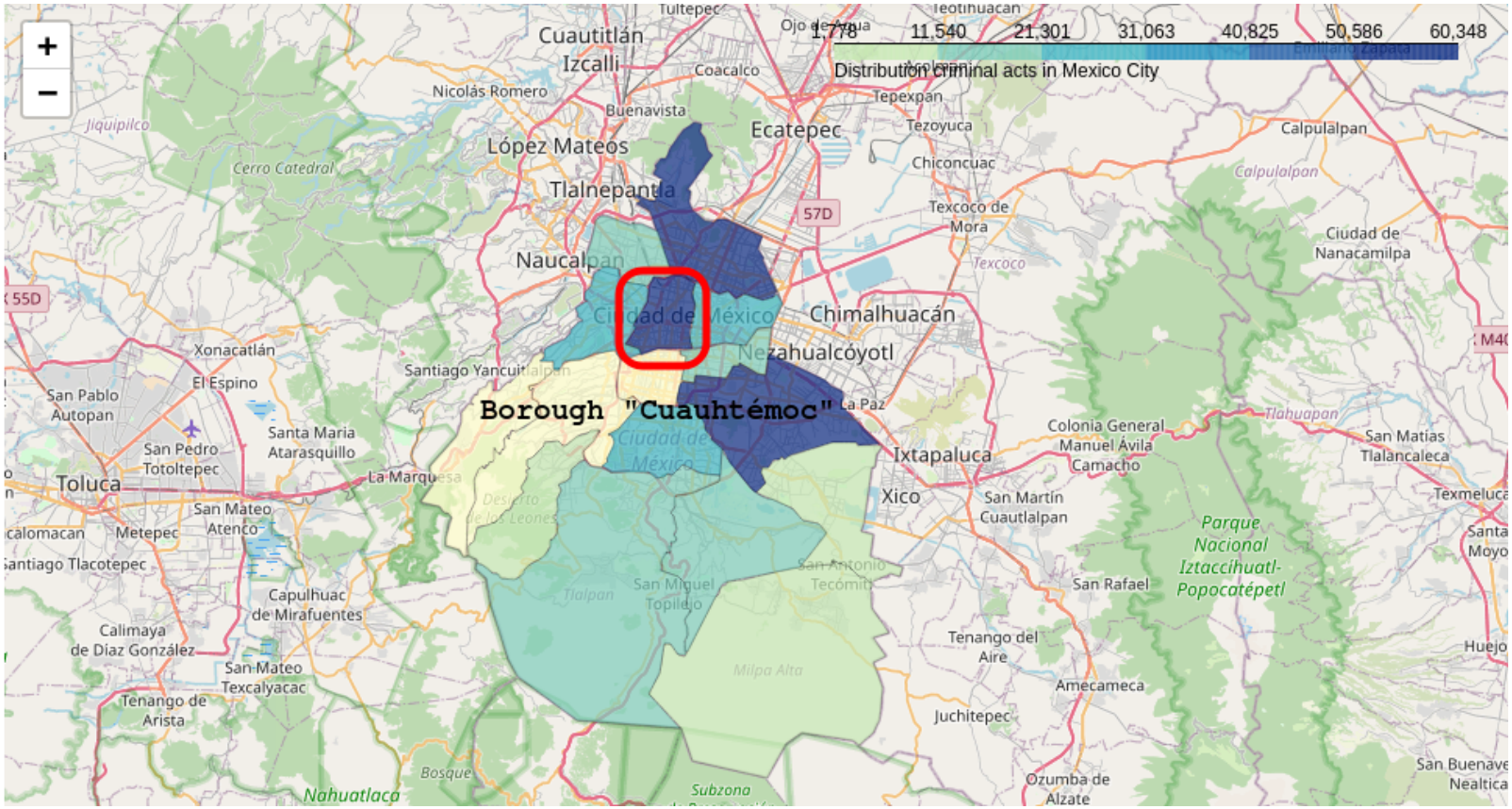


Figura: This is the map of CDMX, it notes that according to the crime's numbers, the north-west region is the most dangerous. The remarked borough has the biggest number of crimes.

In object to keep computational performance, henceforth only it was used the data of borough Cuauhtémoc and year 2020.

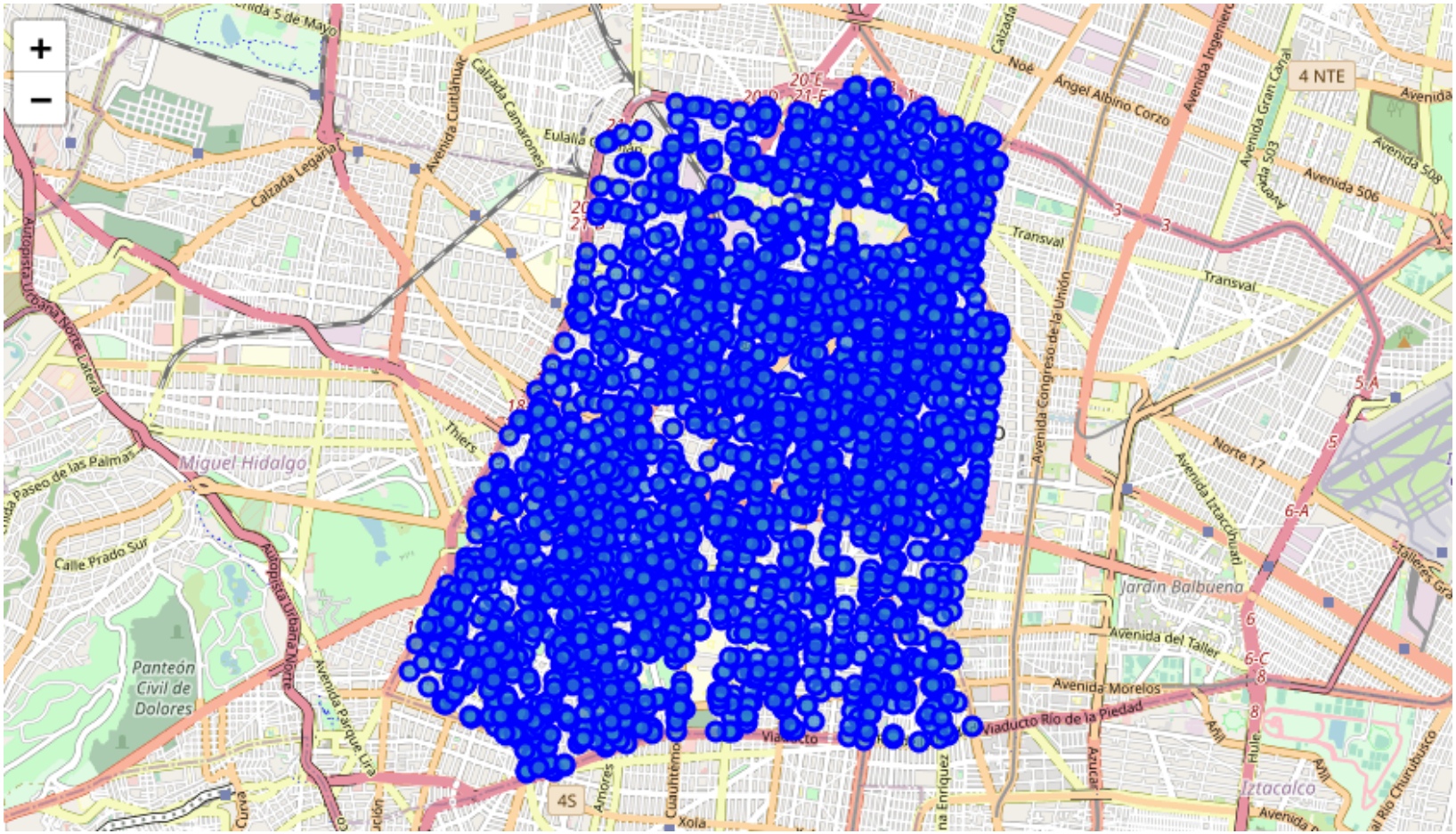


Figura: The crimes of the borough Cuauhtemoc.

Foursquare big entrance

The Foursquare API will give us the information of category of venues surrounding crime locations. A brief explication of the process, next.

- Getting credentials, and define an exploring url with a radius of 500 meters of all neighbourhoods belongings to Cuauhtémoc borough, and a limit of 30 venues.
- Obtaining a JSON file and convert in *pandas* dataframe.
- To know the different categories of the venues obtained, and calculate their frequency.
- Finally create a new dataframe that displays the top ten of the venues of each neighbourhood.

The final dataframe is

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	ALGARIN	Mexican Restaurant	Taco Place	Bakery	Advertising Agency	Coffee Shop	Paper / Office Supplies Store	Seafood Restaurant	Brewery	Food Truck	Steakhouse
1	AMPLIACIÓN ASTURIAS	Mexican Restaurant	Bakery	Argentinian Restaurant	Print Shop	Dessert Shop	Café	Ice Cream Shop	Video Game Store	Coffee Shop	Bed & Breakfast
2	ASTURIAS	Mexican Restaurant	Bar	Bakery	Liquor Store	Latin American Restaurant	Music Venue	Market	Martial Arts Dojo	Flower Shop	Grocery Store
3	ATLAMPA	Restaurant	Taco Place	Park	Mexican Restaurant	Food Truck	Coffee Shop	Burger Joint	Candy Store	Bakery	Bridge
4	BUENAVISTA	Mexican Restaurant	Dessert Shop	Ice Cream Shop	Coffee Shop	Japanese Restaurant	Garden	Burger Joint	Museum	Brewery	Mediterranean Restaurant

Figura: The above dataframe is the obtained by Foursquare API.

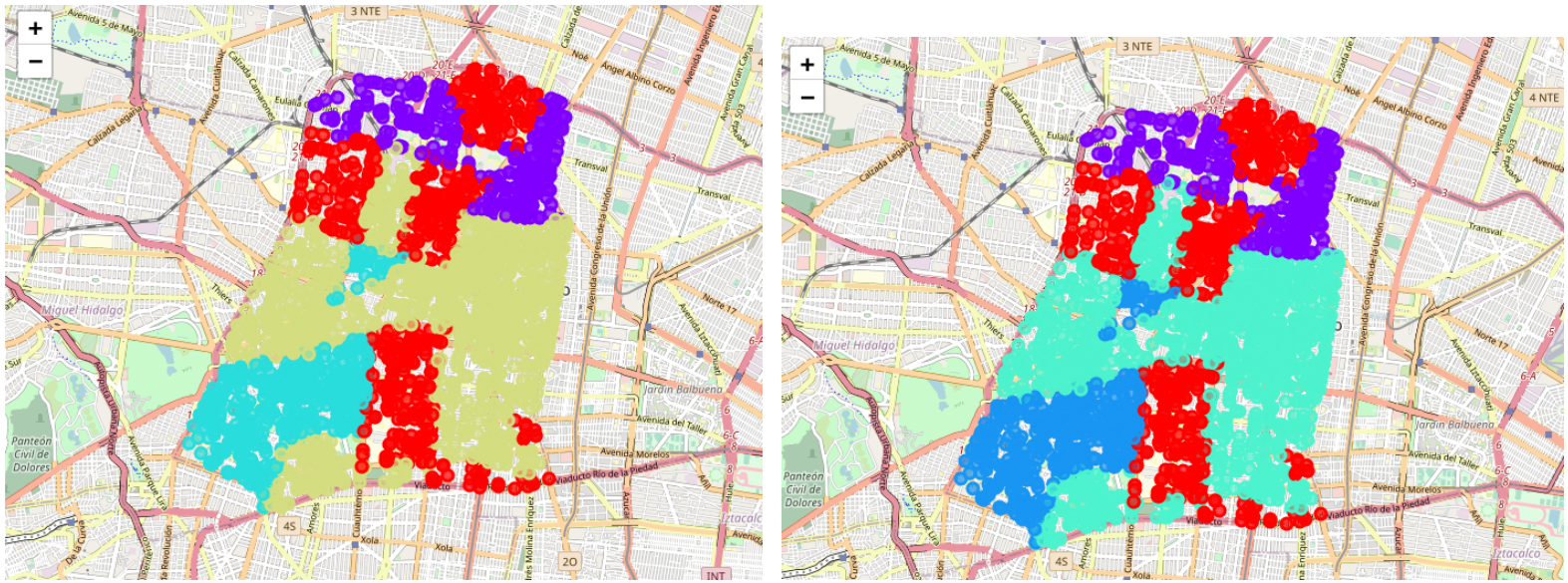
Machine learning to the rescue

- The use of machine learning in this work is to give an answer a certaing questions than descriptive and inferential statistics can not. It means, for example to ask, "Where is easier been a victim of a robbery, in a street market, in a food stand, or a formal restaurant?
- Why K means clustering? Because of is an unsupervised method and this work is about exploring data, not making predictions. Also there is not a "true dataset" to try train phase a methods as supervised learning.

K means clustering

For starting, 4 clusters were configurated. *Note: The cluster labels were applied only to foursquare dataframe showed, so the data that has the bigger importance in this particular analysis is the category of venues, i.e. the motive of certain journeys than shape city dynamics.* Following the above process, the dataframe with clusters was merged with the dataframe of crime in Cuauhtémoc. Finally these clusters were projected on a map.

Even though the K means clusters could have a non perfect forms, how could do we know that the number of clusters is the best? Play with numbers an to observe what happens.



(a) Crime distribution in Cuauhtémoc organized by 4 clusters.

(b) Crime distribution in Cuuhtémoc organized by 6 clusters. Observe the change in the colors.

Figura: Two different executions of clustering. Nevertheless apparently only 4 clusters are effective.

Elbow method

What can be done to get the best clustering? Comparing k vs *inertia*, the inertia is measure of how internally coherent clusters are. This is showed below.

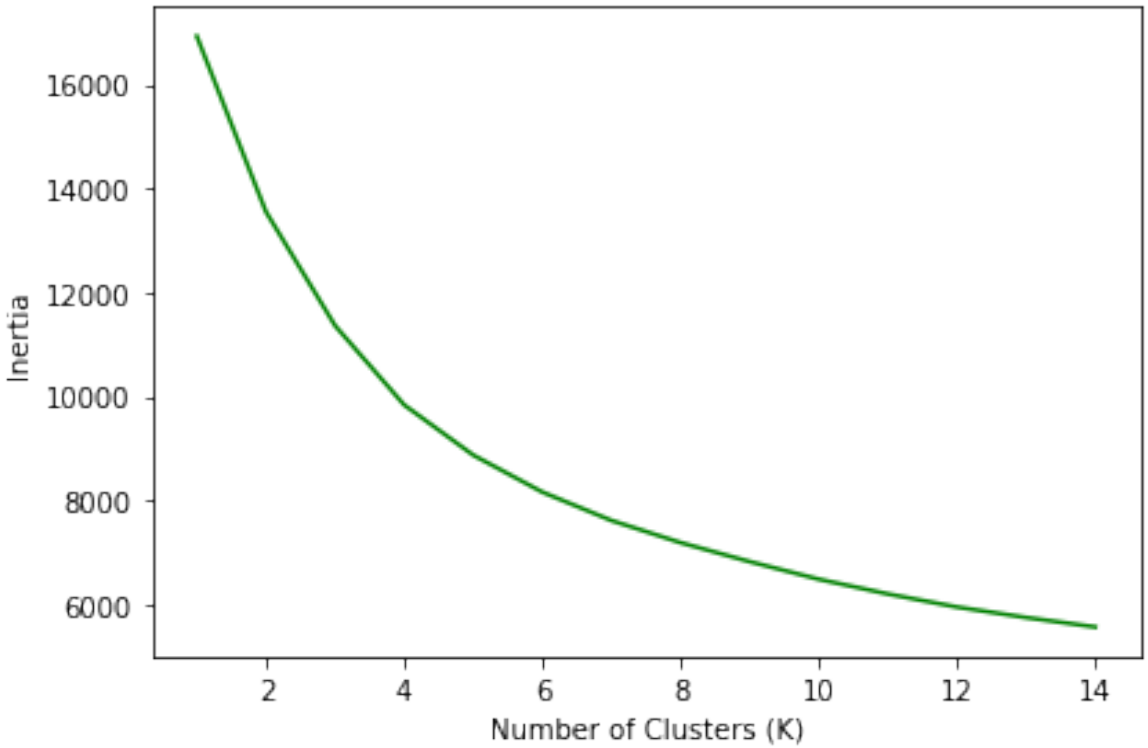


Figura: Curve of elbow method, observe that 'knee of the curve' is when K = 4, therefore this configuration gets the best clustering.

Results

Cluster 1

Figura: In the cluster one, there are seven neighbourhoods, of which the most dangerous are Buenavista and Santa Maria La Ribera. This cluster has zones mostly about food places and entertainment, also in above mentioned neighbourhoods there is a considerable geographic mobility and there a lot of means of transport as underground, trains, buses for rapid transit, etc.
The two most recurrent crimes in this cluster are "Low-impact felony"(this according to CDMX's regulation, are crimes like threats, fights, vandalism), and "Stealing and robbery to passer-by"

Cluster 2

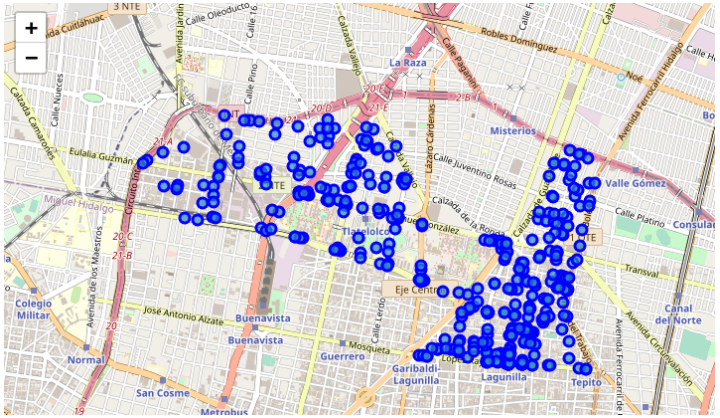


Figura: In the cluster two, there are eight neighbourhoods, of which the most dangerous are Morelos and Obrera.This cluster has zones more assorted besides food and entertainment has a lot of store, also in the first above mentioned neighbourhood is worth to highlight that it has one of the principal street market of CDMX.
The two most recurrent crimes in this cluster are "Low-impact felony"(this according to CDMX's regulation, are crimes like threats, fights, vandalism), and "Stealing and robbery to passer-by"

Cluster 3

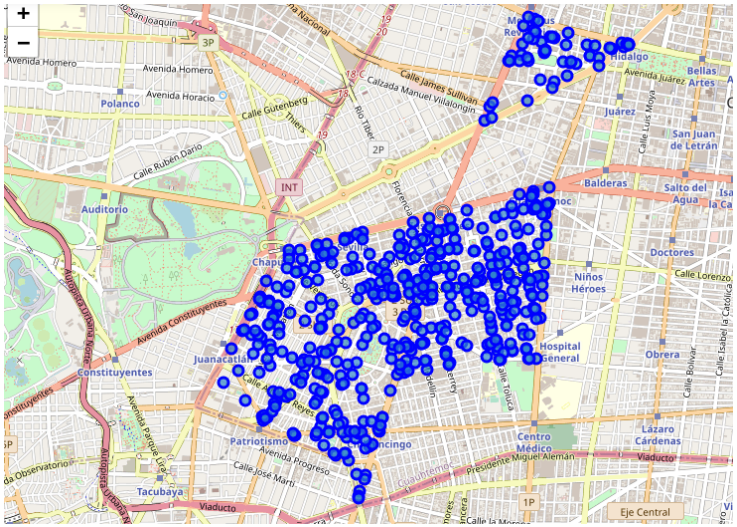


Figura: In the cluster three, there are six neighbourhoods, of which the most dangerous are Guerrero and Doctores.This cluster has no mainly sector in particular.
The two most recurrent crimes in this cluster are "Low-impact felony"(this according to CDMX's regulation, are crimes like threats, fights, vandalism), and "Stealing and robbery to passer-by"

Cluster 4

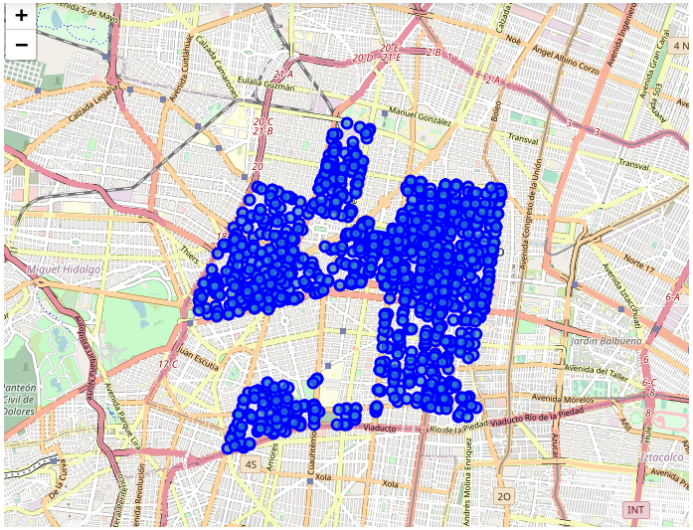


Figura: In the cluster four, there are fourteen neighbourhoods, of which the most dangerous are Centro and Juarez. The first neighbourhood is the most famous in CDMX, and it is an incredibly assorted place, it can be found food places, day/night entertainment, the sightseeing places, malls, all sort of shop, this is worth to highlight because of the number of crimes is the biggest of the analyzed neighbourhoods.
Also this cluster contains a few of the cosmopolitan neighbourhoods of CDMX.
The two most recurrent crimes in this cluster are "Low-impact felony"(this according to CDMX's regulation, are crimes like threats, fights, vandalism), and "Stealing and robbery to passer-by"

Discussion

Possibles improvements:

- Implementing POI (*Points Of Interest*) or mesh blocks.
- To use check-ins proportionated by Foursquare to consider tendencies and newer standpoint.
- New metrics can be used in the exploratory analysis. Also, investigate the relationship with types of crimes.

Another directions.

- In the technique aspect, dataframe in different category venues with one hot encoding can be very useful in training supervised methods, and the types of crimes (many) can be worked with decision trees or neural networks.
- Another interesting analysis would be interesting include as features the income and the social stratum of the plaintiffs and how is the relationship with the surrounding and type of crimes committed.

Conclusions

- The crime in CDMX is not uniform in the matter of distribution in their boroughs. In its most dangerous borough (Cuauhtemoc) it was identified the principal type of crimes committed and the behaviour in their neighbourhoods, resulting in the most dangerous the neighbourhood Centro.
- What about evaluation metrics? The standard evaluation metrics, like Jaccard similarity score, F1-score, even LogLoss, are used when we have predicted data. In this case, prediction data was not done, only exploring the neighbourhoods so evaluation metrics in this work, are not necessary. In this case, just optimizing was possible to be done, as elbow method.

THE END!

Finally, this work gives a view about the utility of Foursquare and can be improved, all suggestions are welcome. Thanks for your attention!!