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

K. Damián E.

In this report will be justified and explicated the task of exploring crime data in Mexico City (aka CDMX) using the knowledge acquired in this course and others.

1. Introduction

1.1. Background

Crime is a phenomenon present since the beginning of the times and everywhere. A few years ago the crimes just could be solved, but currently, crimes also could be predicted. Traditionally the modus operandi of crime prediction is to use historical patterns (empirical knowledge), geographic information systems and demographic variables e.g. sex, income, age, profession, but these variables only change slowly in the time. Nevertheless, this modus operandi ignores other short-term variations of factors which are relevant to the occurrence in crime events.

These factors that provided a new and interesting standpoint to *modus operadi* are inside of the city dynamics i.e. human mobility across a city, e.g. "...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?", currently the repository of data and social media allow us to analyze the mentioned factors.

1.2. Problem

For the purpose to analyze city dynamics and criminal behaviour, to discover all of this new standpoint can offer, it will be described the process to explore crime data in CDMX with libraries of Python to manage data as *pandas*, visualization data as *folium*, the Foursquare API to get data about venues and their categories around the crime locations, and exploring the relationship between the number of crimes, type of crimes and their location, in the ultimate stage the neighbourhoods will be studied with an unsupervised machine learning methods, K means clustering.

1.3. Interested sectors

The crime analysis with a new approach of a city so heterogeneous, it is an excellent opportunity to observe the dynamics of its residents. That would be a business problem for distinct justice organizations in Mexico City, national authorities, security private companies, data science consultancy as a task given by restaurants or hotels that would want to be established in security zones.

2. Data

2.1. Data sources

The dataset was obtained from the «Open Data Mexico City» which has a different type of data of CDMX as mobility, development, environment, health, justice, finances, etc. In this case it was chosen the section **Justice and security**, and it was chosen the section **Investigation's folders of CDMX's General of Justice Prosecutor's Office** (all of that in Spanish), then it can go to the section *exploring data*, to chose the filters of interest and download the suitable dataset.

2.2. Data downloading and cleaning

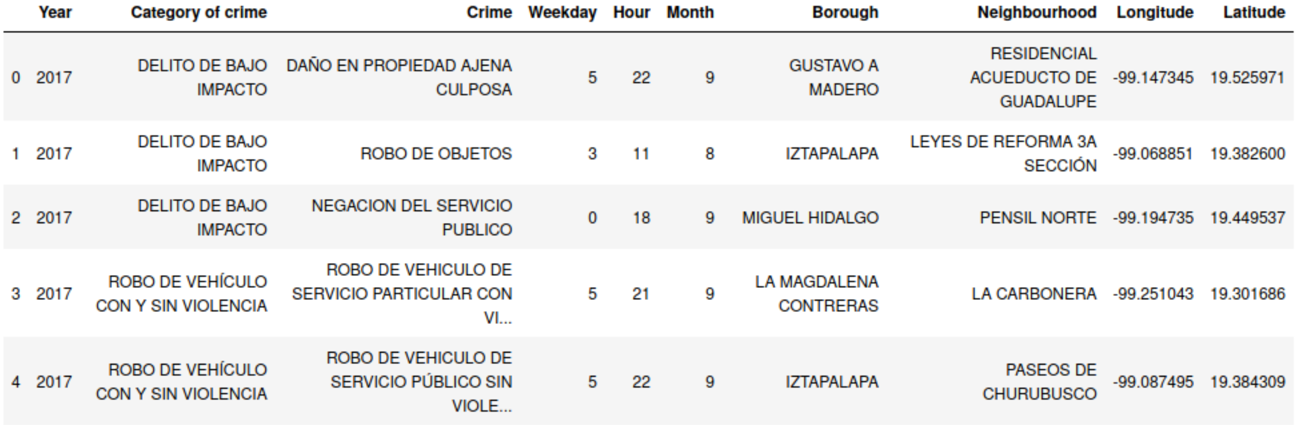
It was downloaded a csv file with all the registered complaints in CDMX, whose features were

- Year
- Date of crime

- Category of crime
- Type of crime
- Date when complaint was done
- Boroughs
- Neighbourhoods
- Public department correspondent
- Geographic coordinates
- ...

the first modification to the original dataset was filter by year, just keeping the period 2010-2020, and preserve the features of interest.

The final data set it is showed in the figure 1.



	Year	Category of crime	Crime	Weekday	Hour	Month	Borough	Neighbourhood	Longitude	Latitude
0	2017	DELITO DE BAJO IMPACTO	DAÑO EN PROPIEDAD AJENA CULPOSA	5	22	9	GUSTAVO A MADERO	RESIDENCIAL ACUEDUCTO DE GUADALUPE	-99.147345	19.525971
1	2017	DELITO DE BAJO IMPACTO	ROBO DE OBJETOS	3	11	8	IZTAPALAPA	LEYES DE REFORMA 3A SECCIÓN	-99.068851	19.382600
2	2017	DELITO DE BAJO IMPACTO	NEGACION DEL SERVICIO PUBLICO	0	18	9	MIGUEL HIDALGO	PENSIL NORTE	-99.194735	19.449537
3	2017	ROBO DE VEHÍCULO CON Y SIN VIOLENCIA	ROBO DE VEHICULO DE SERVICIO PARTICULAR CON VI...	5	21	9	LA MAGDALENA CONTRERAS	LA CARBONERA	-99.251043	19.301686
4	2017	ROBO DE VEHÍCULO CON Y SIN VIOLENCIA	ROBO DE VEHICULO DE SERVICIO PÚBLICO SIN VIOLE...	5	22	9	IZTAPALAPA	PASEOS DE CHURUBUSCO	-99.087495	19.384309

Figura 1: Clean dataset to work, the features of interest are: Year, Category of crime, Crime, Weekday, Hour, Borough, Neighbourhood, Latitude and Longitude.

Also the missing values were erased. The geographic coordinates will be used to obtain data of Foursquare, specifically venues and their category.

3. Methodology

In this part begins the exploration, so it is groping because it does not have a clue.

The first stage it will be observed the relationship between temporal features and number of crimes, after, it will be observed the relationship between geographic features and number of crimes. Also, the popular wisdom will be tested, e.g. how true is there are more crimes at night than at day?. In this part of the process the type of crimes, it is not taken into account.

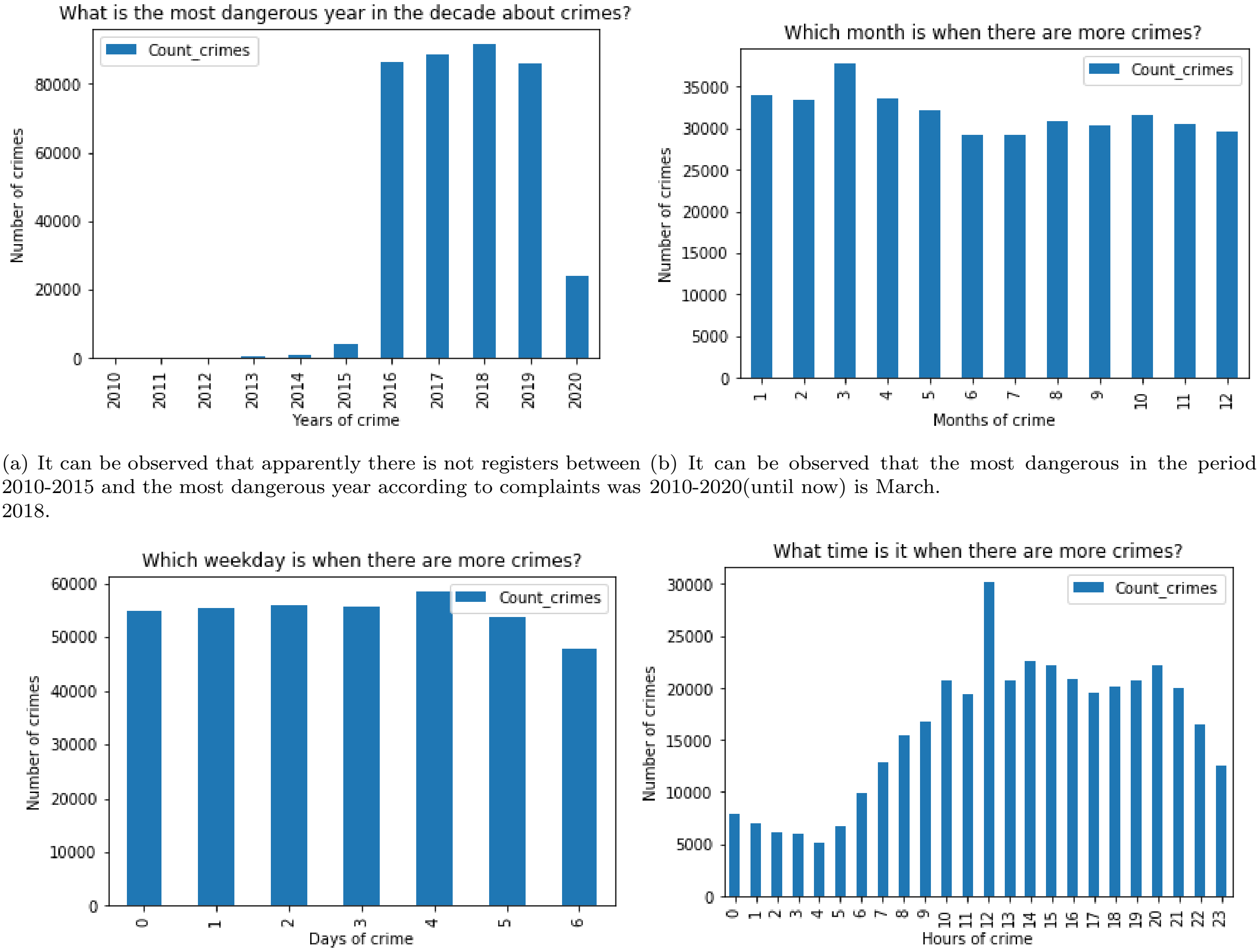
Once chosen the most dangerous neighbourhood, it will use the Foursquare API, in order to study the city dynamics, knowing the category of venues of the locations of the crimes.

In the final stage, it will be used an unsupervised machine learning method, because it seeks to find a possible structure of data. This is the step that mixes all datasets, the crime data, the Foursquare data and it gives us an order and it tells us the reason by which certain neighbourhood are attractive to people and how is the crime in these.

3.1. Exploratory analysis

3.1.1. Temporal Features

First of all, a few bar charts were done, in the figure 2 it can be seen the relationship bewtween number of crimes and temporal features.



(a) It can be observed that apparently there is not registers between 2010-2015 and the most dangerous year according to complaints was 2018.

(b) It can be observed that the most dangerous in the period 2010-2020(until now) is March.

(c) It can be observed that popular wisdom is right here, the most dangerous weekday is Friday.

(d) It can be observed, that popular wisdom is not right here, the most dangerous hour is noon.

Figura 2: Bar charts of crime's numbers vs temporal features

3.1.2. Descriptive and Inferential Statistics

Later the bar charts, a descriptive study was carried out with the command *df.describe()*, is was obtained the means, counts, percentils, but them can not give a useful information. So, it was used the Pearson's correlation, this it shows in figure 3.

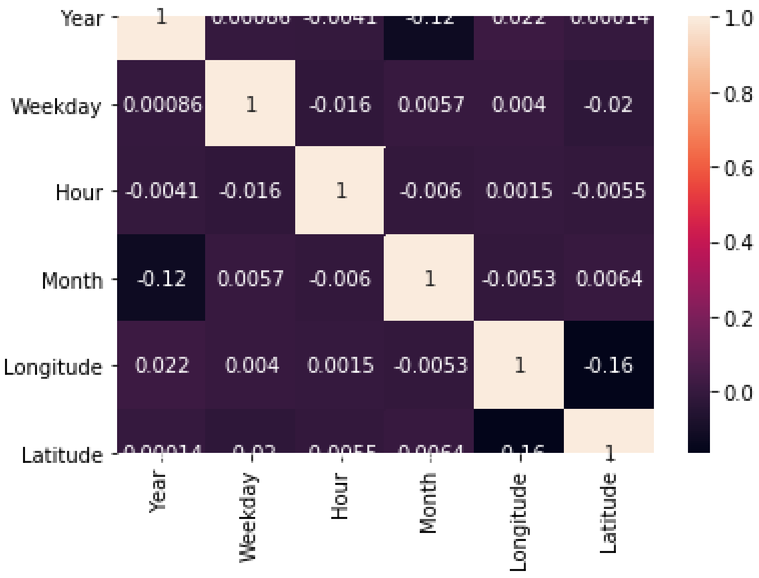


Figura 3: The Pearson's correlation was done only with raw numerical features. As it can be observed, this correlation is not conclusive.

3.1.3. Geographical features

Afterwards, it will be observed the relationship between geographical features and number of crimes. Since CDMX is a huge city, just the most dangerous borough was studied. This projection can be seen in the figure 4.

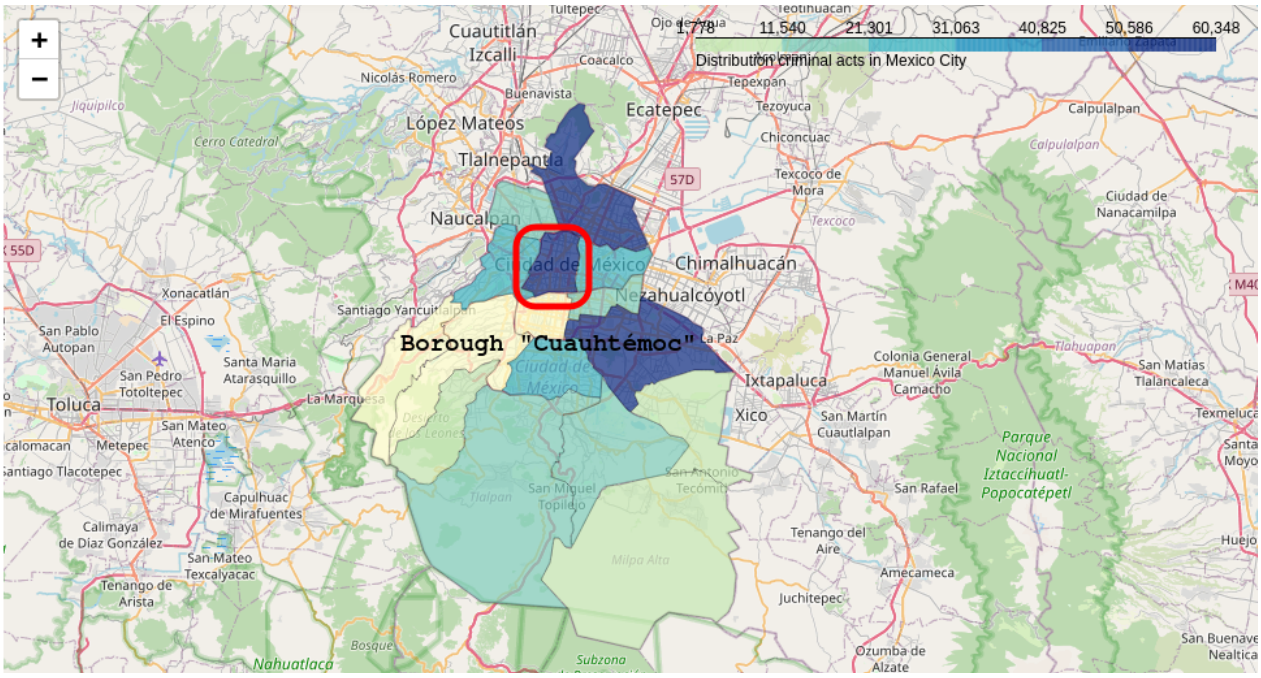


Figura 4: This is the map of CDMX, it notes that according to the crime's numbers, the north-west region is the most dangerous. The remarked borough has the biggest number of crimes.

It would be great to get the information of venues of all crimes but is a huge work. In object to keep computational performance, henceforth only it was used the data of borough Cuauhtémoc and year 2020. A projection in a map it is showed in figure 5.

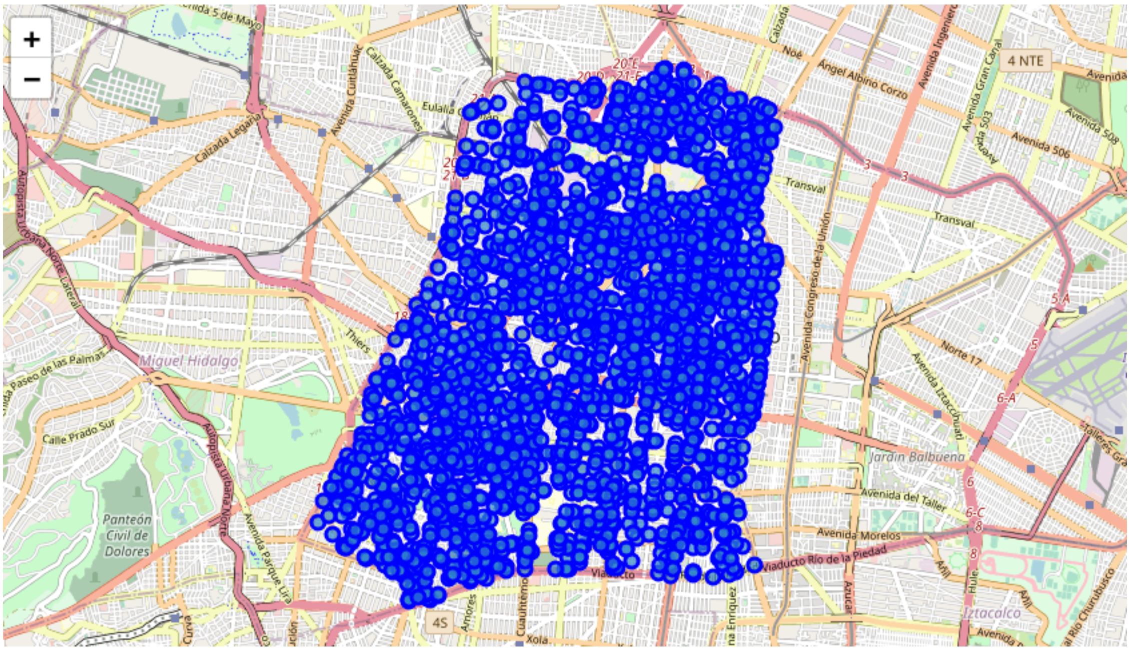


Figura 5: The crimes of the borough Cuauhtemoc.

3.2. Foursquare big entrance

3.2.1. Getting data

It is time to get the information that Foursquare offers, firstly it was defined the credentials, and after a few attempts, I came to the conclusion that is better to work with the API when the dataframe does not have repeated values.

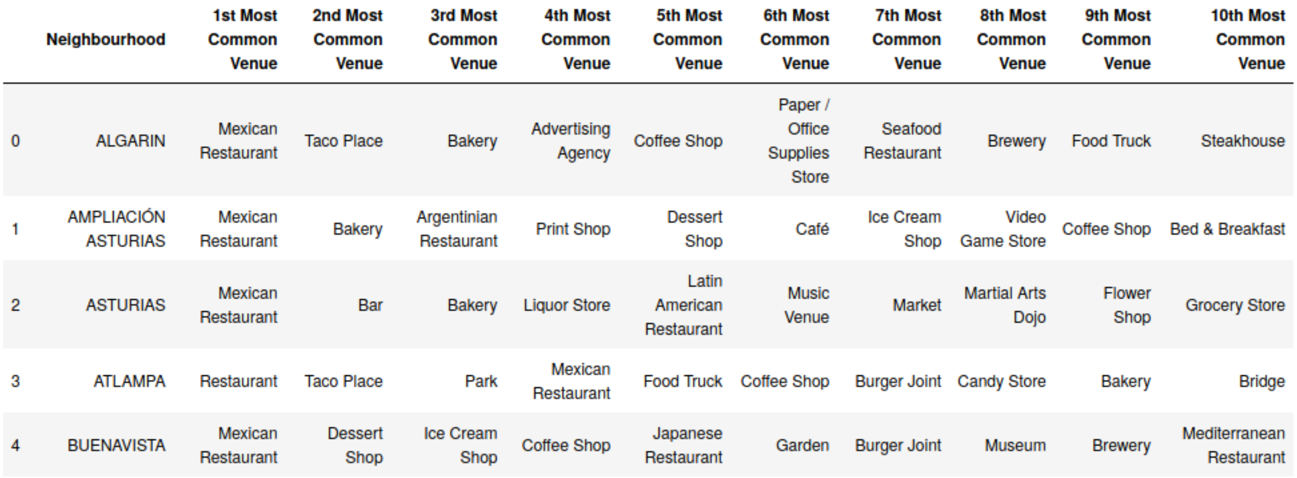
The url was defined with a radius of 500 meters of all neighbourhoods belongings to Cuauhtémoc borough, and a limit of 30 venues.

Thereafter, it was replicated the function learned in another notebook, this function to use the url, then the GET request was sent and the data was examined by groups and captured by the item keys. Following that, it gets the category of venues using names, latitude, and longitude, at last the obtained JSON file was converted into a dataframe.

3.2.2. Exploring neighbourhoods

The purpose of this part is to know the different categories of the venues obtained. The approach was to apply one hot encoding technique, then to get the mean of each category and to get the frequency of each category of venue in each neighbourhood of Cuauhtemoc.

Later than, a function that ordered those frequencies in ascending order was created, then at the close of a new dataframe that displays the top ten of the venues of each neighbourhood. The resulting dataframe it is showed in figure 6.



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

3.3. K means clustering

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?

3.3.1. Justification

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.

3.3.2. Performance and optimization

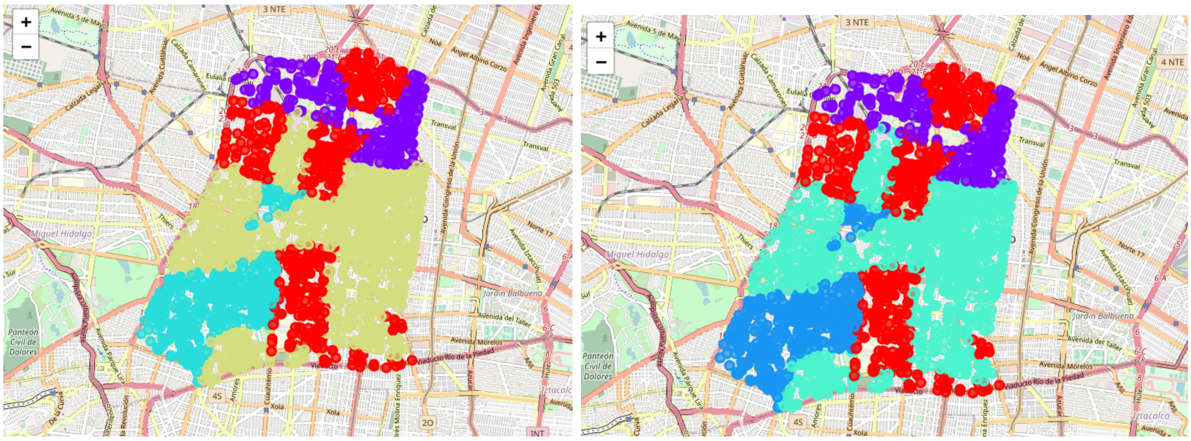
For starting, I chose to implement 4 clusters. This part is tricky. The cluster labels were applied only to dataframe showed in the figure 6, so the data that has the bigger importance in this particular analysis is the category of venues, i.e. the motive of certain journeys than shape city dynamics.

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? A first way is just playing with numbers an to observe what happens. It can be showed in figure 7.

What can be done to get the best clustering? The Elbow method is the answer. The curva was obtained with help of a loop comparing k vs *inertia*, the inertia is measure of how internally coherent clusters are. This is showed in figure 8.

4. Results

In this section the cluster will be examined. Due to the types of crimes were a big number (200), I considered more practical just work with the categories of crimes.



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

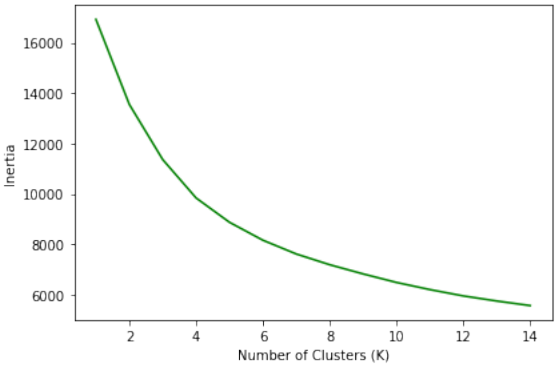


Figura 8: Curve of elbow method, observe that 'knee of the curve' is when K = 4, therefore this configuration gets the best clustering.

In each cluster it was obtained the number of crimes per neighbourhood and with the knowledge of the categories of venues correspondent to the clusters, it was determined the city dynamics.

4.1. Cluster 1

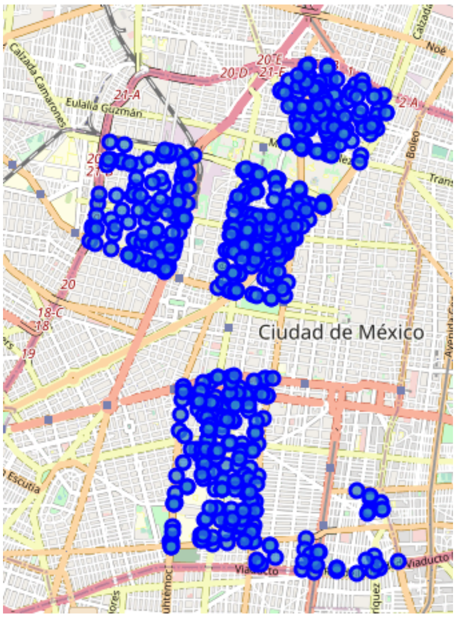


Figura 9: Cluster one

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".

4.2. Cluster 2

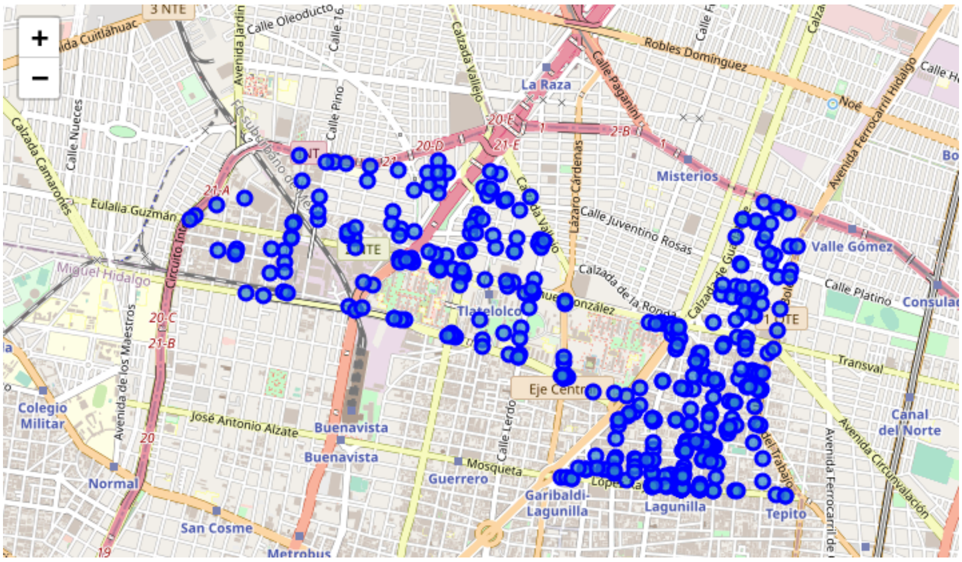


Figura 10: Cluster one

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".

4.3. Cluster 3

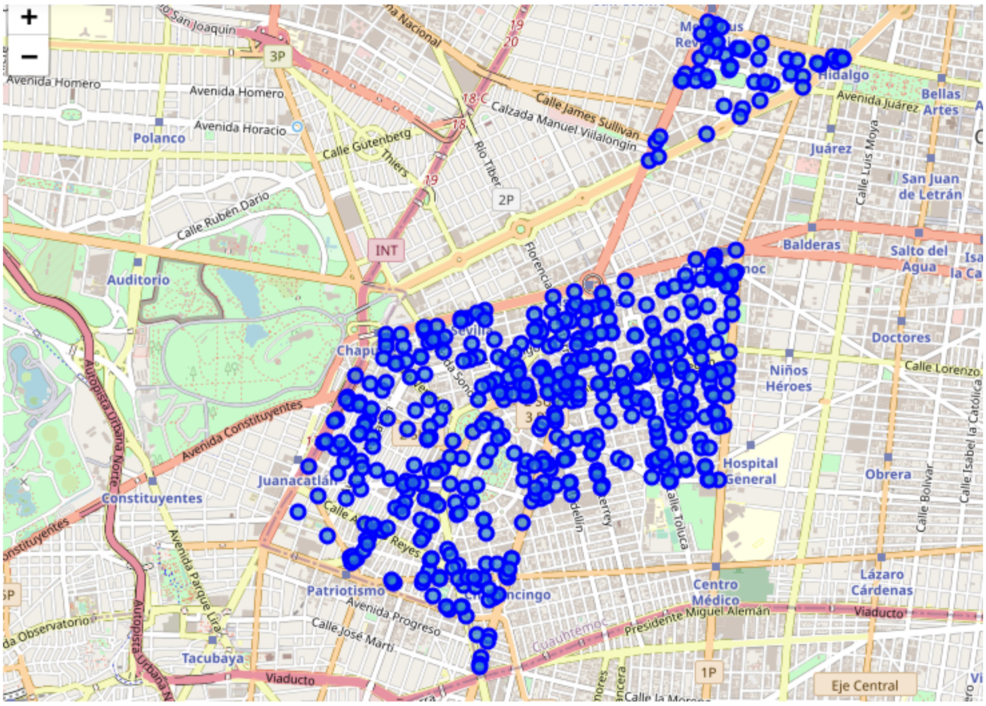


Figura 11: Cluster three

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”.

4.4. Cluster 4

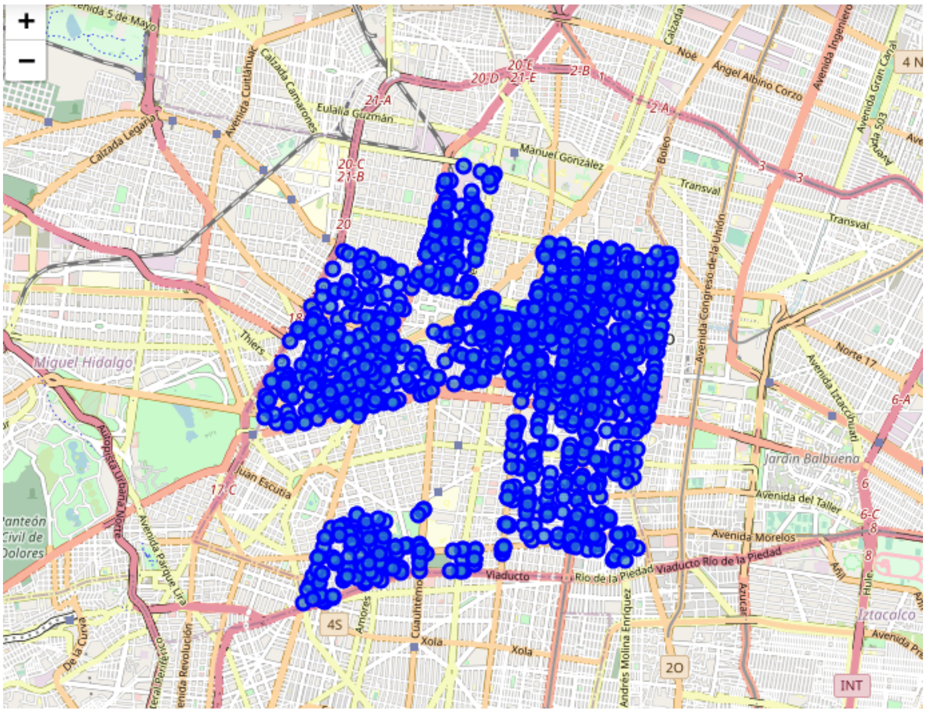


Figura 12: Cluster four

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”

5. Discussion

Suppose that the target public of this work is divided into two parts, the authorities, security, and enthusiastic data scientists (team A) in a part and the investors for business in the other (team B).

The cluster four is the most interesting to the team A because the studies of crime exploring and prediction can go deeper, e.g. implement POI (*Points Of Interest*) or mesh blocks, to use check-ins proportionated by Foursquare to consider tendencies and newer standpoint, also new metrics can be used in the exploratory analysis as venues equitability indexes or density of crimes by day. Also, investigate the relationship with types of crimes.

On the other side, cluster one could be of interest to team B, because it is a zone with considerable mobility and is attractive to the business sector.

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.

6. 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. Also was identified the category of venues that there are surrounding. It would be interesting to repeat this analysis in other boroughs and neighbourhoods of CDMX.

In this work, the heavy part in machine learning was the short-dynamics of the city (type venues), because the clusters were applied to the cuauhtemoc merged dataframe (Foursquare data) because the principal purpose is exploring the neighbourhoods.

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.

To jump and make predictions will be necessary supervised methods of machine learning and give more priority to traditional features (as in the first analysis), for example with the knowledge of crime in the period 2010-2020, can we predict the crime in 2021-2024?

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.

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