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

The crime is a phenomenon present since 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 it is 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 is 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 to analyse the mentioned factors.

1.2. Problem

For the purpose of make an analysis between city dynamics and crime behaviour and 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 ultimate stage the neighbourhoods will be studied with a unsupervised machine learning methods, K means clustering.

1.3. Interested sectors

The crime analysis 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.