Geographical cluster of the possible COVID-19 infection points in México City

Andres J. Ramos

1. Introduction

1.1 Background

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) it's a virus that has the initial outbreak in Wuhan China in December 2019 and has stopped most of the economic and social activities around the world. Seven months after the first outbreak was recorded, some governments have decided to restart global economic activities gradually and to treat the resumption of social activities with great caution. Given that currently social gathering points could be the main sources of contagion to initiate a resurgence, it is important to know which places and establishments will be most frequented based on social behaviors. In Mexico City, for example, there are boroughs that are presenting large outbreaks and want to return to economic and social activities. Therefore, it would be very useful for the health authorities to obtain a classification of the places of social events near the geographical areas that have registered the most infected people, in order to take preventive measures when social coexistence is reactivated.

1.2 The problem

This project aims to classify geographical locations in Mexico City that could be considered a future focus of infection from highest to lowest risk, through the use of the API Foursquare that would provide information on the inflow into establishments around the infected people and the unsupervised K-Means algorithm

1.3 Interest

This information can be useful when health authorities in Mexico City need to establish prevention measures to avoid contagion in public places, as well as to determine which types of meeting places can be opened and which should wait until contagion decreases.

2. Data acquisition and cleaning

2.1 Data source

For this project, the data from Mexico City are obtained from the official URL of the government of México and the author are "Secretaria de Salud de la Ciudad de México".

https://datos.cdmx.gob.mx/explore/dataset/casos-asociados-a-covid-19/table/?disjunctive.resultado.

This dataset provides all cases related to COVID-19 up to 6/2/2020 and is updated regularly every week.

2.2 Important keywords

Due to the increasing rate of infections, health services are insufficient, so the Secretary of Health decided to classify cases as outpatient and inpatient:

- **Outpatient** cases ("Casos Ambulatorios"): refers to non-severe SARs-CoV-2 positive cases where the person is not hospitalized and is asked to undergo quarantine at home.
- **Inpatient** ("Casos Hospitalizados"): Refers to severe SARs-CoV-2 positive cases requiring urgent medical care and hospital isolation.

Also, the Mexican **attitude** front COVID-19 is mostly unbelievable, this represent a serious risk of virus resurgence. Very much outpatients are confusing the symptoms of SARs-CoV-2 with the influenza and they don't take the proper isolation measures.

2.3 Data cleaning

The original data set weighs 1.71 Gb because it has forty columns and sixty-four thousand columns approximately. This includes not only the positive and negative cases, but also the geographical location, clinical conditions of the patient, place of birth, place of hospital where they were treated, registration ID, etc.

The data cleaning aims to create a clean Dataframe that contains:

- Only Positive cases per borough.
- Only outpatient cases
- Categorize the cases with an encoder

First, all the columns that indicate if the patient have a chronic condition has been dropped, this includes: EPOC, diabetes, asthma, hypertension, obesity, etc. Also, the columns that

contain information like birth site, nationality, hospital sector, update date and home state because the data is about Mexico City.

The column "gender" is not necessary because de virus doesn't differentiate the gender.

Second, the rows that contain negative cases, hospitalized cases, and people dead was dropped.

Third, the result dataset only contains the columns "borough", "age" and "result".

	Borough	Age	Result
0	Iztacalco	50	Positivo SARS-CoV-2
1	Gustavo A. Madero	52	Positivo SARS-CoV-2
2	Tlalpan	7	Positivo SARS-CoV-2
3	Iztapalapa	48	Positivo SARS-CoV-2
4	Xochimilco	58	Positivo SARS-CoV-2

Figure 1 df.head of final dataset