2. Data acquisition and cleaning

2. 1. Data acquisition

I will considerer the dataset from the last three years. I get this information from the official data from the government of Buenos Aires: https://data.buenosaires.gob.ar/dataset/mapa-del-delito. It has the date, the neighborhood (comunas), the type of crime, the subtype of crime, the neighborhood, the latitude and the longitude and others variables.

Later I import the CABA_comunas.geojson (https://data.buenosaires.gob.ar/dataset/comunas/archivo/b0b627ac-5b47-4574-89ac-6999b63598ee) to extract the latitude and longitude to create the cloropleth map and later the barrios.geojson (http://cdn.buenosaires.gob.ar/datosabiertos/datasets/barrios/barrios.geojson)

Finally, I use the FOURSQUARE API to get the venues and his category

2.2 Data cleaning

First, I import all the three dataset and takes different actions to have the dataset already for the analyze.

Later I begin to clean the data. This were all steps that I took for each year.

- Drop all the rows that do not have 'Comunas'
- Drop all the unnecessary columns like 'id', 'franja_horario','subtipo_delito'
- Replace all the values that have a parenthesis on the 'Tipo_delitos' Column
- Rename the columns latitude and longitude
- Set the crime date with Argentinian format
- Capitalize the names of the columns
- Convert to the correct type all the columns that it was not when it was imported
- Drop the rows that have more than one registered quantity
- And finally create a new column with the years 'crime

After I concatenate all the three dataset in one to start with the data exploratory. I drop the date because it was not being considered.

I import the CABA_comunes.geojson to have which neighborhood are grouped in 'Comunas' due to I will create a map of the city and show the safest and dangerous neighborhood

And I finally I import barrios.geojson to obtain the average of latitude and longitude for each 'Comuna' and if I see that some point in the map are so far form the center, I manually correct with an arbitrary longitude and latitude that represents the center of it. I have to modify the data for get the most venues in the FOURSQUARE API and obtain the clusters