Paw Print of Inequality:

Socio Economic Categorization of San Francisco de Quito Neighborhoods

Data Science Capstone Project

Report

Mateo Yerovi

April 2020

Paw Print of Inequality: Socio Economic Categorization of San Francisco de Quito Neighborhoods

Introduction

Business Problem/Research Objective

The actual study wants to pinpoint the living and socio-economic conditions in every neighborhood around the metropolitan district of San Francisco de Quito. With the help of Geolocation tools and some data science techniques like clustering, we can classify Neighborhoods based on socio economic information of each neighborhood. The result of the study could help to answer some important question about the inequality around the city:

- Which neighborhoods have better or worst living conditions (House construction features and basic services like drinking water and home waste management)?
- Difference of Education skills and conditions between neighborhoods
- Teen pregnancy rates
- Branch and job category
- Neighborhoods with similar commercial characteristics
- Ethnic distribution

Target

The audience for this project could be:

- Municipality of the city:
    - The city government could use the results of the study to implement efficient public politics that help to improve the living and socio-economic conditions in those neighborhoods that have deficiencies.
- Business-Stakeholders:
    - Companies could have a better understanding of the target in every neighborhood.
- New business:
    - With this information new entrepreneurs have a better perspective of the characteristics of every single neighborhood around the city and they can match those characteristics and their new business objectives. This make easier the decision of where they can stablish their company.
    - This is more relevant in-service segment (restaurants, coffee shops, bars, nightclubs, hotels, etc.)

Data and Methodology

Data:

The data of this research came from this resource:

- Fourscore API
  - From this resource I am going to take the information of venues of each neighborhood like name and classification (restaurant, coffee shop, gym, etc)
- Geo Data from "http://www.codigopostalecuador.com/quito-1896:
  - This page has information of zip code and location of Quito
  - I will use this web page to make a data frame of latitude and longitude information, with the help of beautifullsoup.
- Socio-Economic Data base (Instituto de la Ciudad, QUITO)
  - In this base we can find some economic demographic and social indicators, like teen pregnancy rate, education attendance, population, density, health coverage, living conditions.

Methodology

1. **Collection of Geo-Data**

Import the Geo data from the web page:

http://www.codigopostalecuador.com/quito-1896

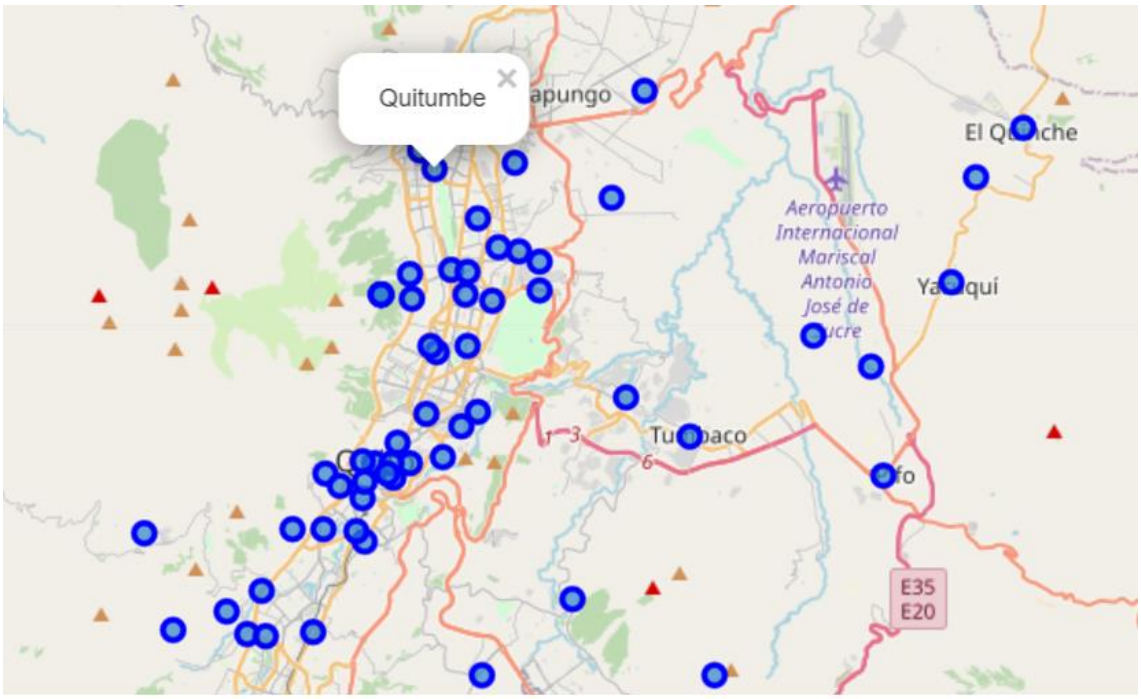
Read and decode the web information with “beutyfullsoup” and transform to “pandas” data frame.

The output includes some information like: “Postal Code”, “Name of the ‘Parroquias’ (Neighborhoods or districts)”, “Latitude” and “Longitude” of each district inside the Quito County.

	Codigo Postal	Lugar	Latitud	Longitud
1	170151	Alangasi	-0.30505	-78.41344
2	170101	Alfaro (Chimbacalle)	-0.23333	-78.51667
3	170152	Amaguaña	-0.38084	-78.51544
4	170153	Atahualpa (Chabaspamba)	-0.18439	-78.49171
5	170129	Belisario Quevedo	-0.16563	-78.51045

2. **Analyze and understand the data**

Develop a geo map with the latitude and longitude information of each neighborhood, to check if every single point is locating correctly.



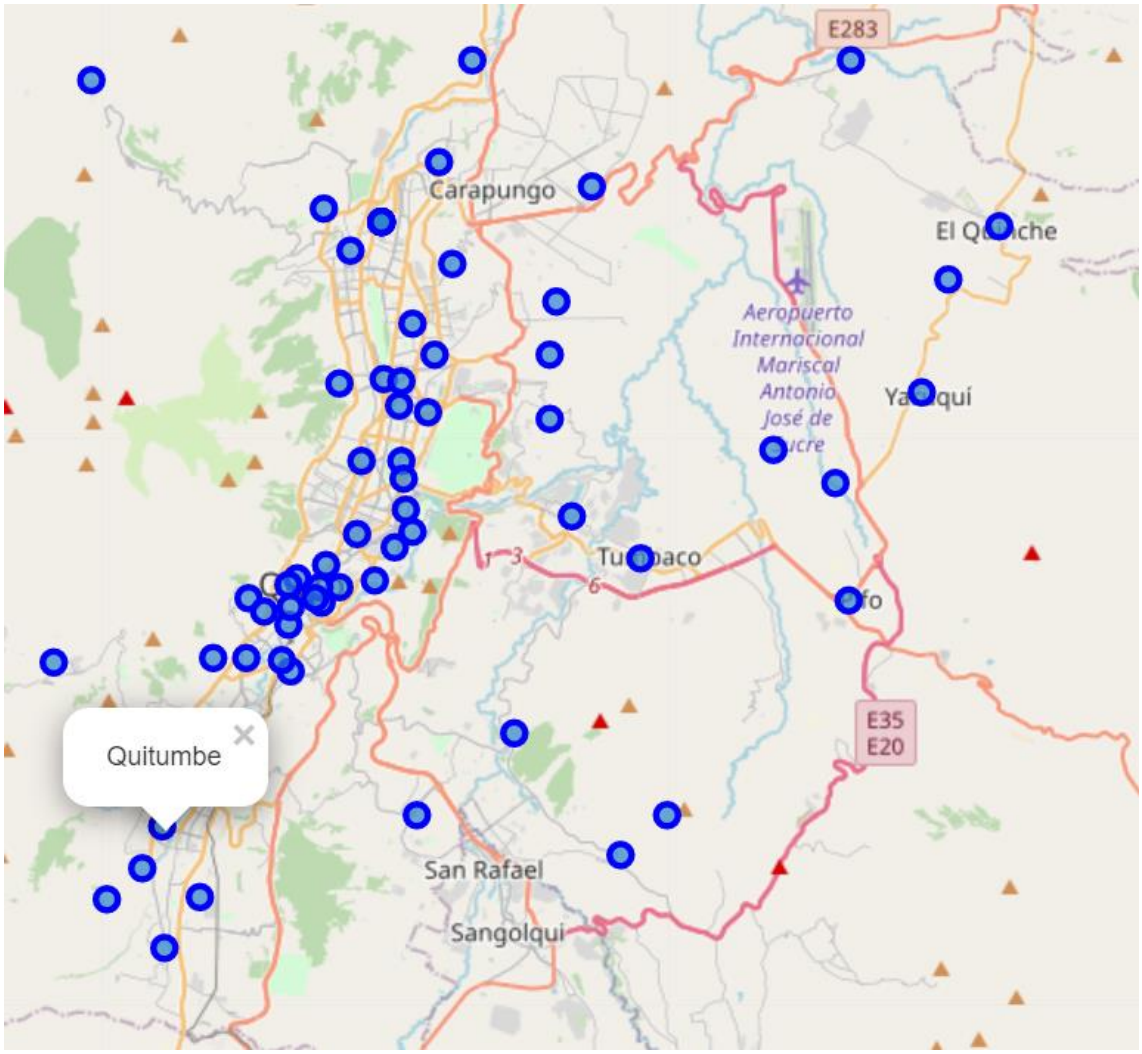
There is a little percent of elements inside de data frame that have wrong geo location information like the point that is highlighted on the map “Quitumbe”.

3. Clean Data
    a. Clean de information

```
Quito_df=Quito.replace([-0.27874,-78.55484],[-0.2188216,-78.5135489])
Quito_df=Quito_df.replace([-0.123360, -78.492230],[-0.2956,-78.556])
Quito_df=Quito_df.replace([-0.221970, -78.512390],[-0.1975,-78.4801])
Quito_df=Quito_df.replace( [-0.164340, -78.457750],[-0.1694004,-78.4354042])
Quito_df=Quito_df.replace( [-0.154240, -78.457750],[ -0.1499,-78.4368])
Quito_df=Quito_df.replace( [-0.151010, -78.464380],[0.1713,-78.4112])
Quito_df=Quito_df.replace( [-0.166670, -78.500000],[ -0.1878,-78.4809])
Quito_df=Quito_df.replace([-0.16563, -78.51045],[-0.1085587,-78.4877381])
Quito_df=Quito_df.replace([-0.184390,-78.491710],[0.167761,-78.360507])
```

	Codigo Postal	Lugar	Latitud	Longitud
1	170151	Alangasi	-0.305050	-78.413440
2	170101	Alfaro (Chimbacalle)	-0.233330	-78.516670
3	170152	Amaguaña	-0.380840	-78.515440
4	170153	Atahualpa (Chabaspamba)	0.167761	-78.360507
5	170129	Belisario Quevedo	-0.108559	-78.487738
6	170102	Benalcazar	-0.182620	-78.481220
7	170154	Calacali	-0.001140	-78.513550
8	170155	Calderon (Carapungo)	-0.097490	-78.422510
9	170120	Carcelen	-0.089710	-78.469920
10	170130	Centro Historico	-0.218822	-78.513549

b. Make a new check after changes with a geo map.



4. Use Foursquare API and elaborate a ranking venue base:
   a. Using clean base of geo location and some specifications like radius 500 and limit of requirements (1000), we can extract the venue information and put it in a data frame.

	Lugar	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Alangasi	-0.305050	-78.413440	Parque De Alangasi	-0.307701	-78.415412	Park
1	Alangasi	-0.305050	-78.413440	Cancha Vieja	-0.306081	-78.415789	Soccer Field
2	Alfaro (Chimbacalle)	-0.233330	-78.516670	Museo Interactivo de Ciencia	-0.236313	-78.516309	Science Museum
3	Alfaro (Chimbacalle)	-0.233330	-78.516670	MIC	-0.231879	-78.515172	Science Museum
4	Alfaro (Chimbacalle)	-0.233330	-78.516670	Mesón de la Recoleta	-0.231333	-78.512786	South American Restaurant
5	Amaguaña	-0.380840	-78.515440	Pista De Patinaje - Castillo De Amaguaña	-0.383567	-78.514819	Ski Trail
6	Belisario Quevedo	-0.108559	-78.487738	Estadio LIGA Deportiva Universitaria - "Casa B...	-0.107717	-78.489075	Soccer Stadium
7	Belisario Quevedo	-0.108559	-78.487738	Ceviche con Faldas	-0.105795	-78.490353	Seafood Restaurant
8	Belisario Quevedo	-0.108559	-78.487738	Metrobus: La Ofelia	-0.109602	-78.488360	Bus Station
9	Belisario Quevedo	-0.108559	-78.487738	Terminal Microrregional La Ofelia	-0.110068	-78.488028	Bus Station

b. Generate dummies of every single venue category for every single venue

	Lugar	Advertising Agency	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant
0	Alangasi	0	0	0	0	0	0	0	0
1	Alangasi	0	0	0	0	0	0	0	0
2	Alfaro (Chimbacalle)	0	0	0	0	0	0	0	0
3	Alfaro (Chimbacalle)	0	0	0	0	0	0	0	0
4	Alfaro (Chimbacalle)	0	0	0	0	0	0	0	0

c. Calculate the percent that the venue category represents in the total number of categories in the neighborhood.

	Lugar	Advertising Agency	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	BBQ Joint
0	Alangasi	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.000000
1	Alfaro (Chimbacalle)	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.000000
2	Amaguaña	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.000000
3	Belisario Quevedo	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.000000
4	Benalcazar	0.0	0.000000	0.000000	0.015152	0.000000	0.0	0.000000	0.000000	0.0	0.015152

d. Find the 10 more common venues in each neighborhood.

	Lugar	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
0	Alangasi	Park	Soccer Field	Seafood Restaurant	Food & Drink Shop	Flea Market	Fire Station	Fast Food Restaurant	Farmers Market
1	Alfaro (Chimbacalle)	South American Restaurant	Science Museum	Comedy Club	Food & Drink Shop	Food	Flea Market	Fire Station	Fast Food Restaurant
2	Belisario Quevedo	Bus Station	Soccer Stadium	Seafood Restaurant	Farmers Market	Food & Drink Shop	Food	Flea Market	Fire Station
3	Benalcazar	Italian Restaurant	Hotel	Bakery	Coffee Shop	Ice Cream Shop	French Restaurant	Japanese Restaurant	Fast Food Restaurant
4	Calderon (Carapungo)	Pizza Place	Chinese Restaurant	Park	Wings Joint	Farm	Food	Flea Market	Fire Station

5. Import Socio-economic Data
    a. Import the data from a excel file and put it in a data frame

```
Shape of Socio econmic information data frame: (65, 67)
```

	Parroquias	Poblacion	Densidad	Proporción_Mujer	Promedio_edad	Promedio_per_hogar	Madres_Adolscentes	PEA	
0	Alangasí	0	0.000000	0.000000	0.000000	0.000000	0.000000	0	
1	Amaguaña	31106	44.363608	0.505079	28.307336	3.830000	0.032300	14158	
2	Atahualpa	1901	15.200114	0.501841	34.276696	3.358657	0.044444	840	
3	Belisario Quevedo	45370	8345.680223	0.527551	32.628675	3.100000	0.022811	24008	2

    b. Divide the base in different segments

The Data Base present 65 observations and 67 variables that are indicators of different classifications like:

- Education
- Living Conditions
- Job
- Social problems like:
    - Teen pregnancy

For thar reason the next step is divide the raw data in each category to make a better analysis

Examples:

EDUCATION

	Parroquias	Alfabetismo	Años_Escolaridad	Asistencia_basica	Asistencia_Bachiller	Asistencia_Superior
0	Alangasí	0.000000	0.000000	0.000000	0.000000	0.000000
1	Amaguaña	0.993509	9.110869	0.969914	0.845095	0.301080
2	Atahualpa	0.986971	7.467333	0.953039	0.722222	0.191045
3	Belisario Quevedo	0.997355	12.262085	0.979246	0.892808	0.454538
4	Calacalí	0.981928	8.243497	0.957659	0.792627	0.218310

LIVING CONDITIONS

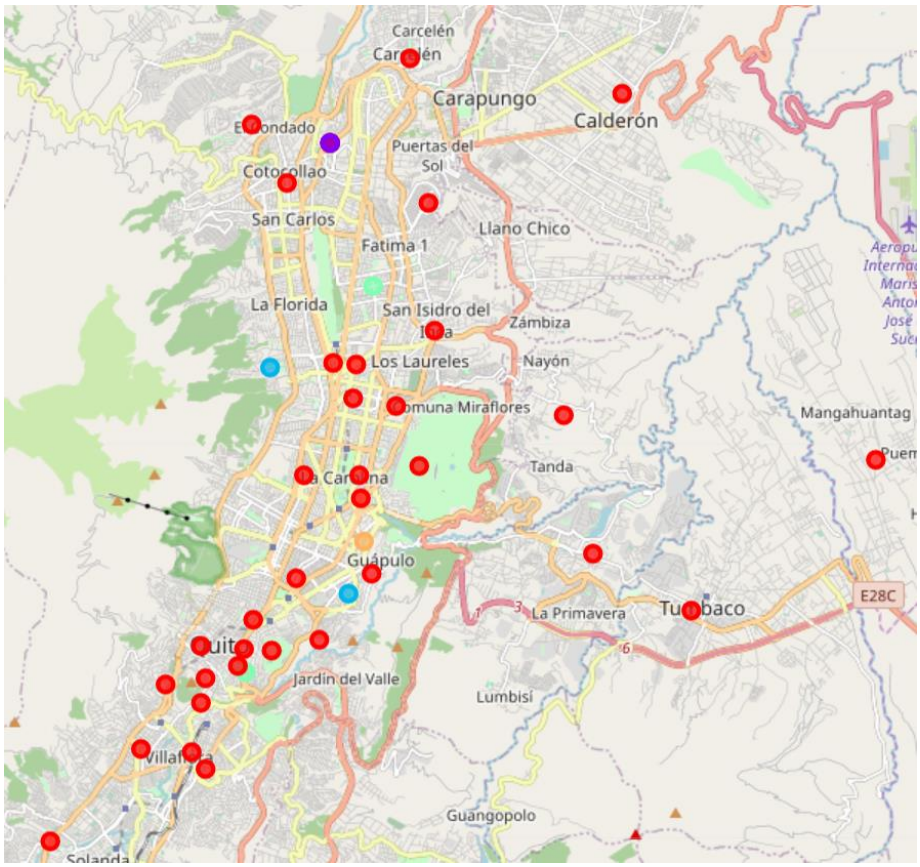
	Parroquias	Vienda_Mal	Vivienda_Buen	Agua_potable	Viviendas_inadecuadas	Salubridad_inadecuadas
0	Alangasí	0.000000	0.000000	0.000000	0.000000	0.000000
1	Amaguaña	0.029371	0.424822	0.815523	0.085461	0.139845
2	Atahualpa	0.066071	0.276786	0.667857	0.132509	0.367491
3	Belisario Quevedo	0.007102	0.582747	0.945456	0.014964	0.033686
4	Calacalí	0.037500	0.322115	0.715385	0.054545	0.216268

6. K means clusters:
    a. Conduct k-means clustering for every segment (e.g venue categories, education indicators, living conditions characteristics, etc.), using the mean frequency of occurrence to create a centroid for each postal code. The k-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as distinct as possible.

Number of clusters: 5

Venues

	Lugar	Latitud	Longitud	Cluster Labels	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
0	Alangasi	-0.305050	-78.413440	0	Park	Soccer Field	Seafood Restaurant	Food & Drink Shop	Flea Market	Fire Station	Fast Food Restaurant	Farmers Market
1	Alfaro (Chimbacalle)	-0.233330	-78.516670	0	South American Restaurant	Science Museum	Comedy Club	Food & Drink Shop	Food	Flea Market	Fire Station	Fast Food Restaurant
2	Amaguaña	-0.380840	-78.515440	3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	Belisario Quevedo	-0.108559	-78.487738	0	Bus Station	Soccer Stadium	Seafood Restaurant	Farmers Market	Food & Drink Shop	Food	Flea Market	Fire Station
4	Benalcazar	-0.182620	-78.481220	0	Italian Restaurant	Hotel	Bakery	Coffee Shop	Ice Cream Shop	French Restaurant	Japanese Restaurant	Fast Food Restaurant



Education

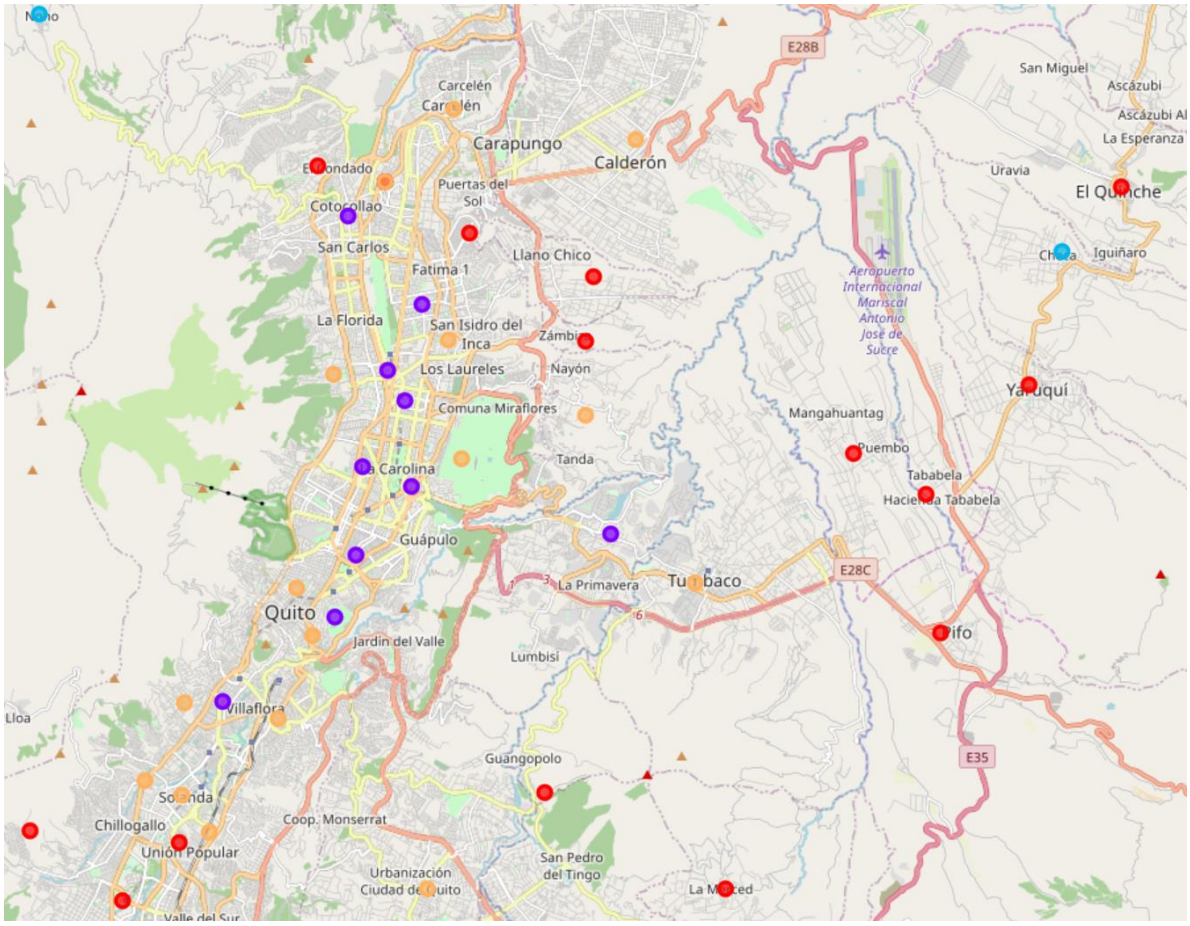
Data

	Parroquias	Latitud	Longitud	Cluster Labels	Alfabetismo	Años_Escolaridad	Asistencia_basica	Asistencia_Bachiller	Asistencia_Superior
0	Alangasí	-0.305050	-78.413440	3	0.000000	0.000000	0.000000	0.000000	0.000000
1	Amaguaña	-0.380840	-78.515440	0	0.993509	9.110869	0.969914	0.845095	0.301080
2	Atahualpa	0.167761	-78.360507	2	0.986971	7.467333	0.953039	0.722222	0.191045
3	Belisario Quevedo	-0.108559	-78.487738	1	0.997355	12.262085	0.979246	0.892808	0.454538
4	Calacalí	-0.001140	-78.513550	0	0.981928	8.243497	0.957659	0.792627	0.218310

Cluster

Number of Cluster	Education Level	Color in the map
1	High Level	Purple
4	Medium-High Level	Orange
0	Medium Level	Red
2	Low Level	Blue
3	Very Low Level	Green

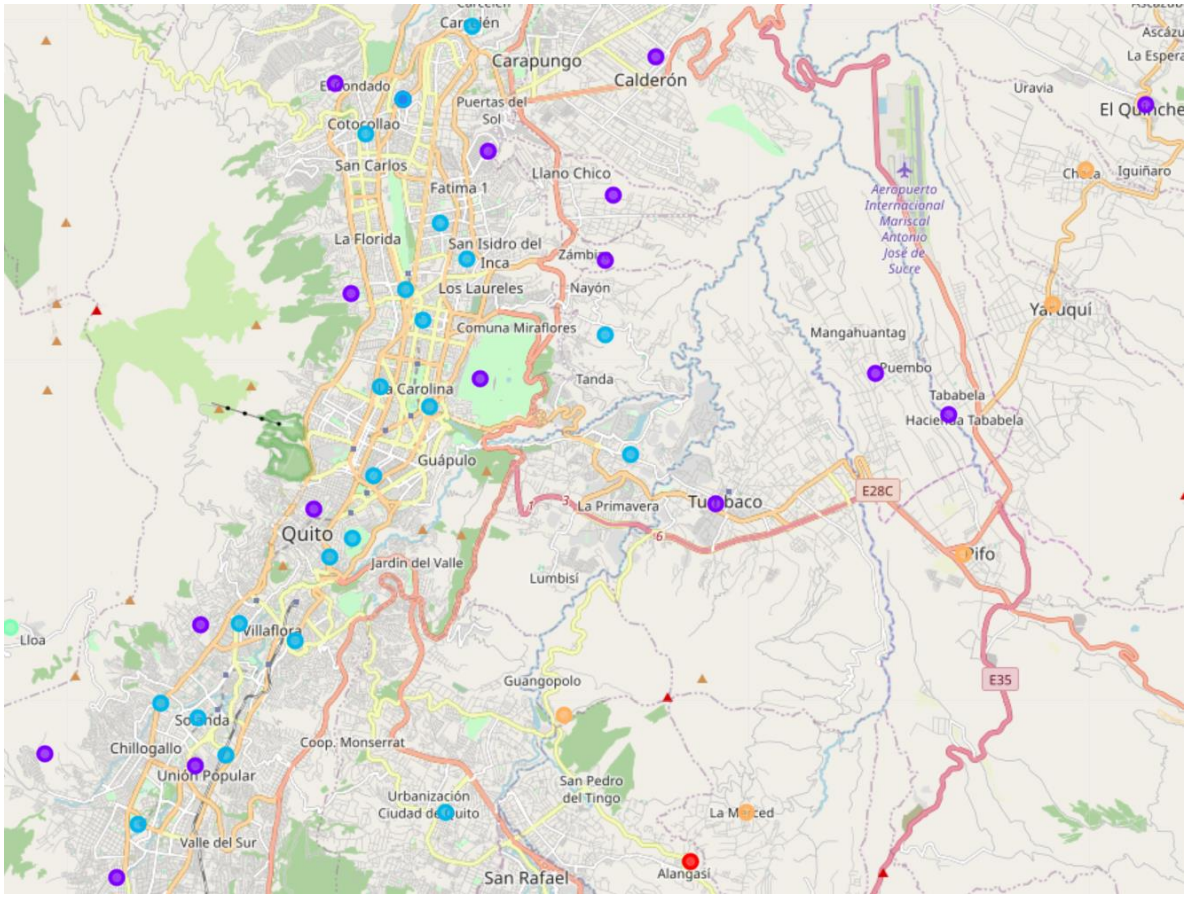
Map



Living Conditions

	Parroquias	Latitud	Longitud	Cluster Labels	Vienda_Mal	Vivienda_Buen	Agua_potable	Viviendas_inadecuadas	Salubridad_inadecuadas
0	Alangasí	-0.305050	-78.413440	0	0.000000	0.000000	0.000000	0.000000	0.000000
1	Amaguaña	-0.380840	-78.515440	4	0.029371	0.424822	0.815523	0.085461	0.139845
2	Atahualpa	0.167761	-78.360507	3	0.066071	0.276786	0.667857	0.132509	0.367491
3	Belisario Quevedo	-0.108559	-78.487738	2	0.007102	0.582747	0.945456	0.014964	0.033686
4	Calacalí	-0.001140	-78.513550	4	0.037500	0.322115	0.715385	0.054545	0.216268

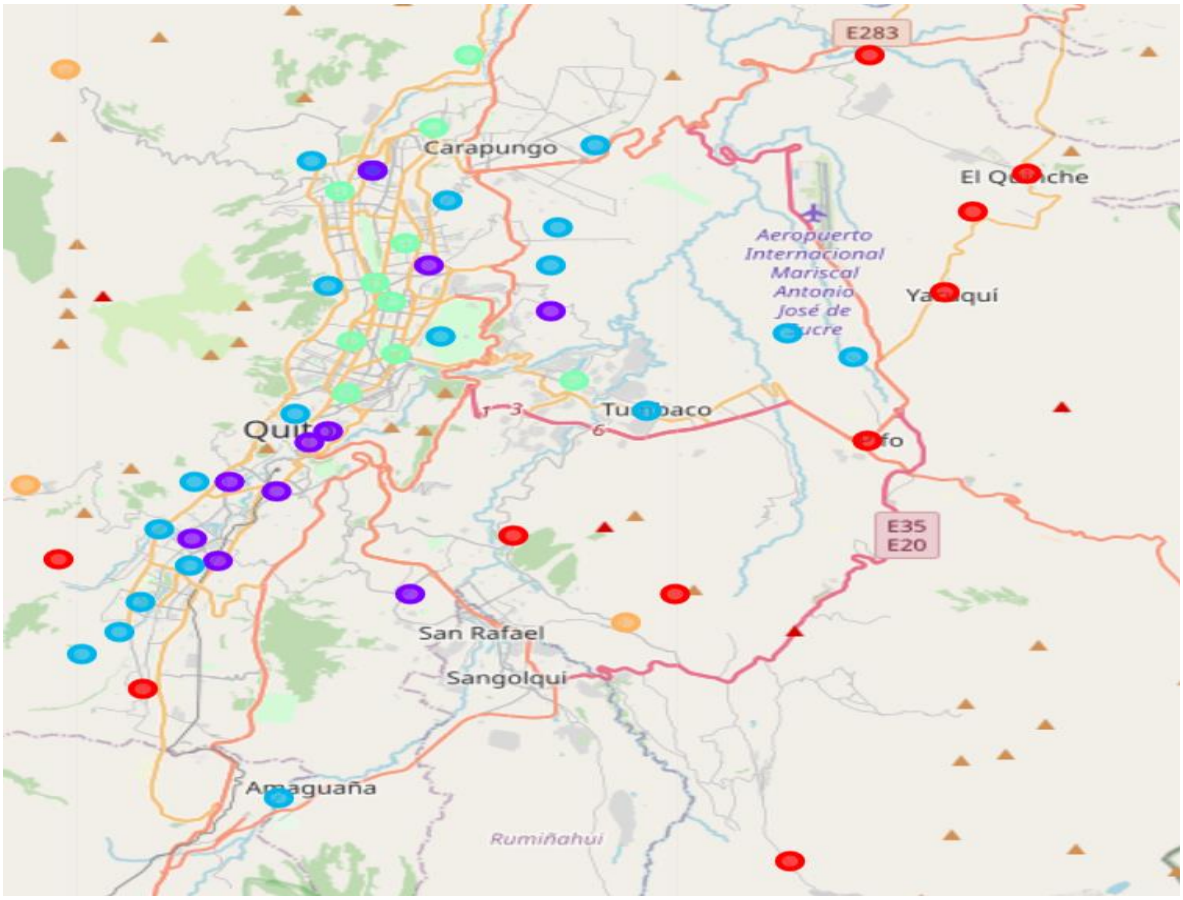
Number of Cluster	Living Conditions	Color in the map
2	High Conditions	Blue
1	Medium-High Conditions	Purple
4	Medium Conditions	Orange
3	Low Conditions	Green
0	Very Low Conditions	Red



Health Coverage

	Parroquias	Latitud	Longitud	Cluster Labels	Seguro_Privado	IESS	IESS_campesino	Jubilados
0	Alangasí	-0.305050	-78.413440	4	0.000000	0.000000	0.000000	0.000000
1	Amaguaña	-0.380840	-78.515440	2	0.047515	0.436346	0.001245	0.302326
2	Atahualpa	0.167761	-78.360507	4	0.026302	0.172121	0.344242	0.147368
3	Belisario Quevedo	-0.108559	-78.487738	1	0.128477	0.530531	0.001318	0.390399
4	Calacalí	-0.001140	-78.513550	0	0.046213	0.304348	0.006293	0.113772

Number of Cluster	Health Coverage	Color in the map
2	High Coverage	Green
1	Medium-High Coverage	Purple
4	Medium Coverage	Green
3	Low Coverage	Orange
4	Very Low Coverage	Red



Conclusion and discussion

After the entire analysis and the socio-economic categorization of different districts around Quito city. We can identify each one that present lack in one or more socio-economic conditions. The study found that the districts that are more isolated from the center of the metropolitan city have less educations skills, living conditions and health coverage. Most of this neighborhood are in the rural or urban-rural area.

It is very alarming the fact that in most of this remote areas people are living without appropriate conditions. They do not have drinking and clean water; good waste manages or structurally safe houses. But also have lack in education attendance and alphabetize index.

Despite this area have more risk to have accidents or health problem, following the logic of their quality of life, they have the very low health coverage.

All this data highlights the imperative necessity of promote more public politics that have the objective of improve their living conditions. The municipality could focus and redirect the budget of every year to develop more efficient projects: like construction materials subsidies, more public service coverage with the help of the private and the community sectors and create commissions of specific clusters of districts that have the task of make more search and receive information from the people of that places with the objective of have a better understanding of the necessity of the people and take better decisions with less cost.