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
    a. Import the Geo data from the web page and transform to a data frame that includes “Parroquias”(Neighborhoods),’Latitude’ and ‘Longitude’ columns.
2. Analyze and understand the data
    a. Develop a geo map with the latitude and longitude information of each neighborhood, to check if every single point is locating correctly.
3. Clean Data
    a. If one or more neighborhoods have wrong geo information, correct them.
    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 (500), we can extract the venue information and put it in a data frame.
    b. Group the information by the name of the neighborhoods
    c. Generate dummies of every single venue category for every single venue
    d. Calculate the percent that the venue category represents in the total number of categories in the neighborhood.
    e. Find the 10 more common venues in each neighborhood.
5. Import Socio-economic Data
    a. Import the data and put it in a data frame

    b. Divide the base in different segments and data frames like education, living conditions, economic activity, ethnic distribution, teen pregnancy, etc.
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.