Capstone Project

The Battle of the Neighborhoods

Applied Data Science Capstone by IBM/Coursera

Introduction: Business Problem

In this project we will try to find an optimal location for a zumba club. Specifically, this report will be targeted to stakeholders interested in opening an **Zumba Fitness Club** in **New York**, The United-States.

Since there are lots of Fitness Club in New York we will try to detect **locations that are not already crowded with gymnasium**. We are also particularly interested in **areas with no Zumba Fitness Club in vicinity**. We would also prefer neighborhoods **with high percentage of Hispanic population**, assuming that first two conditions are met.

We want to consider also other factors in our decision making, particularly socio-economic factors. For that we would opt for neighborhoods where **unemployment rate is acceptable**, **average income is medium to high** and where **average rent price is acceptable**.

We will use our data science powers to generate a few most promising neighborhoods based on this criteria. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by stakeholders.

Data

Based on definition of our problem, factors that will influence our decision are:

- Number of existing Fitness Club in the neighborhood (any type of gym)
- Number of and distance to Zumba Club in the neighborhood, if any
- Percentage of Hispanic population
- Unemployment rate, average income, average rent price

Following data sources will be needed to extract/generate the required information:

- The number of gym or fitness club and their type and location in every neighborhood will be obtained using **Foursquare API**
- The list of the neighborhoods will be extracted from **here** but I already convert it into an excel file.
- Then we will use the **Shapely** library of python to compute the Coordinates of the centroid of the multipolygon of each neighborhood. The geojson file for that was obtained from **here**
- The number of Hispanic population will be obtained from csv files downloaded from **here** and the The percentage will be obtained by directly calculating it.
- The Socio-economic data will be obtained from csv file downloaded from **here**. It contains the Labor force participation rate, the Unemployment rate, the Median household income, and the Median rent for each neighborhood.