Capstone Project - The Battle of Neighborhoods

Business Problem

Introduction:

In this is a capstone project for Applied Data Science Capstone for IBM DAta Science Professional Certificate. I will try to replicate a hypothetical situation where an entrepreneur is interested in opening a refined Spanish restaurant in the city of Toronto. Spanish cuisine, more specifically, mediterranean gastronomy is a very popular and world wide famous for its balanced diet and excellent tastes. Entrepreneur are constantly trying to find new markets to establish and thanks to data analytics we can give insights of possible opportunities.

Business goal:

Find the best possible location for a refined Spanish restaurant in the city of Toronto. Thanks to data science methodologies and tools such as machine learning algorithms (clustering) I hope to give answer to the question: “Where do I have to locate my refined Spanish restaurant?”.

Target audience:

Any entrepreneur/invester who is looking foward to opening a new business in the city of Toronto.

Data:

Selection criteria

For the purposes of this project, the definition of a good location is one that has the absence or fewest identical or similar commercial competitor presence and potentially high income (is normally associated to high housing prices, where upper class people are stablished):

1. Find location with similar restaurant type competitors (italian, greek)
2. Compare average housing prices

Work flow

To solve this problem we will need to:

1. List neighborhoods in Toronto city.
2. Location of the different neighborhoods (latitude, longitude).
3. Explore venue data about spanish restaurants, helping us locate a better location.
4. Explore average house pricing of the different neighborhoods.
5. Locate more optimum location for restaurant opening.

Libraries that we will be using:

Pandas: For creating and manipulating dataframes

Folium: Python visualization library would be used to visualize the neighborhoods cluster distribution of using interactive leaflet map.

Scikit Learn: For importing k-means clustering

JSON: Library to handle JSON files

Geopy: To retrieve Location Data

Requests: Library to handle http requests

Matplotlib: Python Plotting Module

For the classification algorithm that we will use will be k-means clustering algorithm. This will help us segment and group similar locations and neighborhoods.

Sources of data:

The data will be obtained using wikipedia by data scraping techniques thanks to the help of Beautiful Soup packages. Longitude and latitude for the different neighborhoods will be given by the use of Geocoder package. After we have the coordinates we will be able to search for the different venue data thanks to the use of the Foursquare API. K-means algorithm will helps to cluster the different neighborhoods.