COURSERA CAPSTONE

IBM Applied Data Science Capstone

Opening Restaurant in Toronto, Canada

By : Bernardo Andrey Panggabean

July 2020

Introduction

Restaurant is the a place where cooked food is sold to the public. Many people visit to Restaurant because hang out with friends, meet family or many more. Many businessman visit to Restaurant because want to client and establish cooperation. So, Location is an important aspects.We have to know if we have any competitor in the area, the crowd of the area, rent cost of the area etc. Of course, as with many business decision, opening a new restaurants requires serious consideration and a lot more complicated than it seems. Particularly, the location of the restaurants is one of the most important decisions that will determine whether the restaurants will be a success or a failure. If, restaurants is built in quiet place, so restaurants will be a failure.

Business Problem

The objective of this capstone project is to analyze and select the best location in Toronto, Canada to open a Restaurants. Using data scince methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question

Target Audience of This Project

This project is particularly useful to property developers and investors looking to open or invest in new restaurants in Toronto, Canada. This project is timely as the city is currently suffering from oversupply of Restaurants. Recently, there are a lot of restaurants are opened in Toronto. Chairman Specialty Restaurants Association of Canada predict the number of Restaurants in Toronto will increase to 20% by the end of the year. So this project will be useful for anyone who want to open Restaurants in Toronto.

Data

To solve the problem, we will need the following data:

- List of districts in Toronto. This defines the scope of this project which is to confined to the city of Toronto, the capital city of Canada.
- Latitude and longitude coordinates of those districts. This is required in order to plot the map and get the venue data from foursquare.

- Venue data, particularly data related to restaurants. We will use this data to perform clustering on the districts.

Sources of data and methods to extract them
This Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) contains a list of districts in Canada. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordiinates of the districts using Python. After that, we will use Foursquare API to get the venue data for those districts. Foursquare has one of the largest database of 105+ million places and is used by over 125000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Restaurants category in order to help us solve the business problem described above. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), machine learning (K-means clustering) and map visualization (Folium).

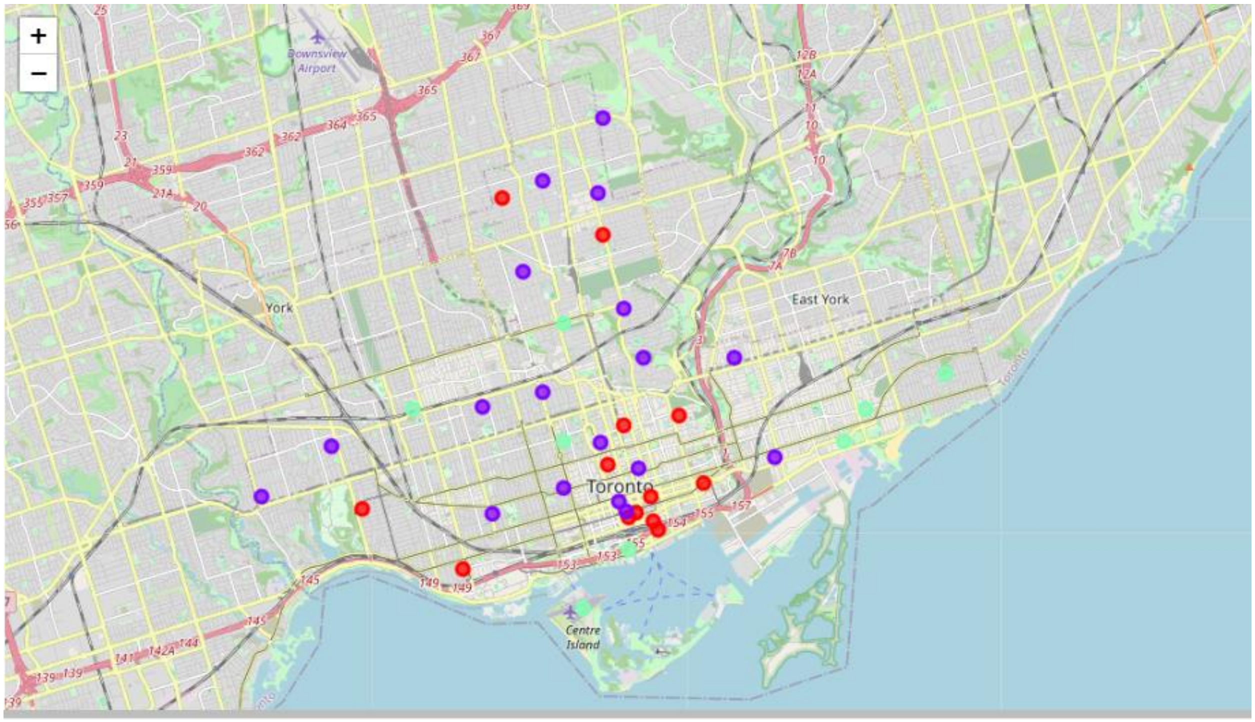
Methodology

First, We must to have the list of district in the city Toronto, Canada. And I get list of district Canada from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. I will do web scrapping using Python requests and beautifulsoup package to extract the list of districts data. I need coordinates in the form of latitude and longitude in order to be able to plot it and get venue data from Foursquare API. Then, I use Geocoder package that will to convert address to geograpichal coordinates in the form of latitude and longitude. After that I will connect to Foursquare with input CLIENT_ID and CLIENT_SECRET to get nearby venues with radius 2000 meters and Limit 100.After I get nearby venues from Toronto. And Venue category from Toronto Area, I get unique venue category and I choose Restaurants. And I create new dataframe that contain Neighborhood and Restaurants. And I set number of clusters is 3 with using KMeans Clustering. KMeans Clustering algorithm identifies k number of centroids and then allocates every data point to the nearest cluster while keeping the centroids as small as possible. After that I visualization using folium package. After Visualization, I get area with high and low concentration of Restaurants.

Results

The results from the k-means clustering show that we can categorize the districts into 3 clusters based on the frequency of occurrence for restaurants:

- Cluster 0: Districts with high number to no existence of restaurants
- Cluster 1: Districts with moderate number of restaurants
- Cluster 2: Districts with low concentration of restaurants



The results of the clustering are visualized in the map below with cluster 0 in red, cluster 1 in purple, and cluster 2 in mint green.

Conclusion

As observations noted from the map in the results section, most of the restaurants are concentrated in the South area of Toronto, with the highest number in cluster 0 and moderate number in cluster 1. On the other hand, cluster 2 has very low number to no restaurant in the districts. This represents a great opportunity and high potential areas to open new restaurant in cluster as there is very little to no competition from existing restaurant.

Discussion

In this project, we only consider the number of restaurants while there are many other factors such as population and income of residents that could influence the location decision of a new restaurants, the density of the districts, the environment around the districts, etc. Can try Hierarchical Clustering to perform this project and that 2 method that have high accurate.In addition, this project made by the free sandbox tier account of foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more optimal results.