Applied Data Science Capstone Project

Introduction

This project is the final assignment for Coursera's Applied Data Science Capstone course. The course is part of a series for the IBM Data Science Professional Certificate. The concept of the project is a hypothetical scenario where an entrepreneur or restauranteur wants to open a Cuban restaurant in Toronto, Canada. In the culinary industry, location is pivotal when opening a new restaurant. With this idea in mind, finding the location to open such a Cuban restaurant is one of the chief decisions for this restauranteur and I am designing this project to find the optimal locale.

Business Problem

The objective of this capstone project is to find the most suitable location for the restauranteur to open a new Cuban restaurant in Toronto, Canada. By using data science methods and tools along with machine learning algorithms such as clustering, this project aims to provide solutions to answer the business question: In Toronto, if a restauranteur wants to open a Cuban restaurant, where should they consider opening it?

Target Audience

A restauranteur who wants to find the location to open an authentic Cuban restaurant.

Data

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

- List of neighborhoods in Toronto, Canada
- Latitude and Longitude of these neighborhoods
- Venue data related to Cuban restaurants

Data Extraction

- Scrape list of Toronto neighborhoods from Wikipedia
- Install Geocoder package to obtain geospatial data (longitudes and latitudes) of these neighborhoods
- Use Foursquare API to get venue data related to these neighborhoods

Methodology

First, I needed to get the list of neighborhoods in Toronto, Canada. This is possible by extracting the list of neighborhoods from Wikipedia:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

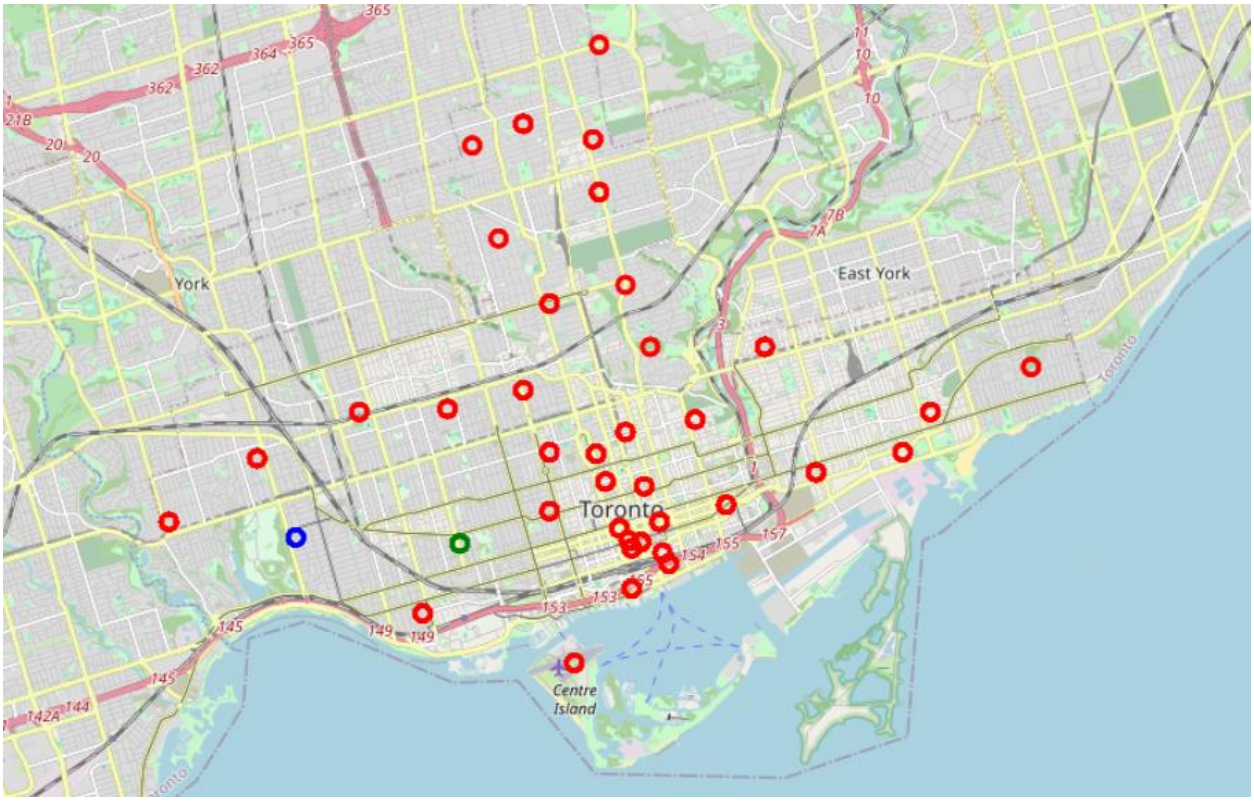
I scraped the webpage by utilizing pandas HTML table scraping method. However, it is only a list of neighborhood names and postal codes. I needed to get their coordinates to utilize Foursquare to pull the list of venues near these neighborhoods. To get the coordinates, I tried using Geocoder Package but encountered some problems so I used the CSV file provided by

IBM team to match the coordinates of Toronto neighborhoods. After gathering these coordinates, I visualized the map of Toronto using Folium package to verify whether these are correct coordinates. Next, I used the Foursquare API to pull the list of top 100 venues within 500 meters radius. I created a Foursquare developer account in order to obtain an account ID and API key to pull the data. From Foursquare, I was able to pull the names, categories, latitude, and longitude of the venues. With this data, I could also check how many unique categories I can get from these venues. Then, I analyzed each neighborhood by grouping the rows of neighborhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later.

Here, I made a justification to specifically look for "Cuban restaurants". Lastly, I performed the clustering method by using k-means clustering. K-means 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. I clustered the neighborhoods in Toronto into 3 clusters based on their frequency of occurrence for "Cuban food." Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the restaurant.

Result

Clusters



The results from k-means clustering show that we can categorize Toronto neighborhoods into 3 clusters based on how many Indian restaurants are in each neighborhood:

- Cluster 0: Neighborhoods with no Cuban restaurants

- Cluster 1 & 2: Neighborhoods with more number of Cuban restaurants

The results are visualized in the above map with Cluster 0 in red, Cluster 1 in blue, and Cluster 2 in green.

Recommendations

There are only two Cuban restaurants in the Toronto area in Cluster 1 and Cluster 2. There is both good news and bad news. There is very little competition but there might be little interest in a new Cuban restaurant. It seems Cluster 0 might be the optimal location as there are no Cuban restaurants in these areas. Therefore, this project recommends the entrepreneur to open an authentic Cuban restaurant in these locations.