Using foursquare API and Clustering to Identify a suitable location to open a new Bar in Toronto, Canada

Aishik Mukherjee

April 24th, 2020

1. INTRODUCTION:

For this Capstone project, I am creating a hypothetical scenario for a concept manager/owner who wants to open a new Bar in Toronto, Canada. With the purpose in mind, finding the location to open a Bar is one of the most important decisions for this entrepreneur and I am designing this project to help him find the most suitable location. In order to do so, I shall do some analysis on the location data using foursquare API and clustering algorithm.

1.1 Business Problem:

The idea behind this project is that there may be enough number of Bars in Toronto and it might present a great challenge for this entrepreneur who is based in Canada, to choose a suitable location in Toronto to sustain his business where the competition might be low. The objective of this capstone project is to find the most suitable location for the entrepreneur to open a new Bar in Toronto, Canada. By using data science methods and machine learning methods such as clustering, this project aims to provide solutions to answer the business question: In Toronto, if an entrepreneur wants to open a new Bar, where should they consider opening it?

1.2 Target Audience :

The owner/manager is the sole target client who seeks a good location to open a new Bar

2. DATA:

In order to solve this problem, I will be needing some relevant data on which we can work on. Specifically, the list of neighborhoods in Toronto, Canada. Then we would need to analyze and plot the neighborhoods in a map, we will be needing the location and longitude values of the neighborhoods. Then ultimately, we would require venue data related to Bars. This will help us find the neighborhoods that are most suitable to open a new Bar.

3. EXTRACTING DATA:

We will use the Data of neighborhoods in Toronto from wikipedia, which is available here at -
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
The table in the given site provides us with the Postal Codes, Boroughs and Neighborhoods in Toronto, Canada.

Next, we will be needing location data of the neighborhoods we got from wikipedia, specifically the longitude and latitude values. For this, we can use Geocoder package or a file which contains the latitude and longitude values of these neighborhoods(http://cocl.us/Geospatial_data)

Next, we will be needing the venue data related to these neighborhoods. For this we shall be using the foursquare API.

4. METHODOLOGY:

First, I got the list of neighborhoods in Toronto, Canada. This was possible by extracting the list of neighborhoods from wikipedia page (“ https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M ”).
I did the web scraping by utilizing pandas html table scraping method as it was just one table and using this method was easier and more convenient to pull tabular data directly from a web page into dataframe. Next, I cleaned the dataframe and did some operations on them like, getting rid of blank rows with no borough name, formatting the neighborhood column and merging the neighborhoods based on postal code and boroughs.

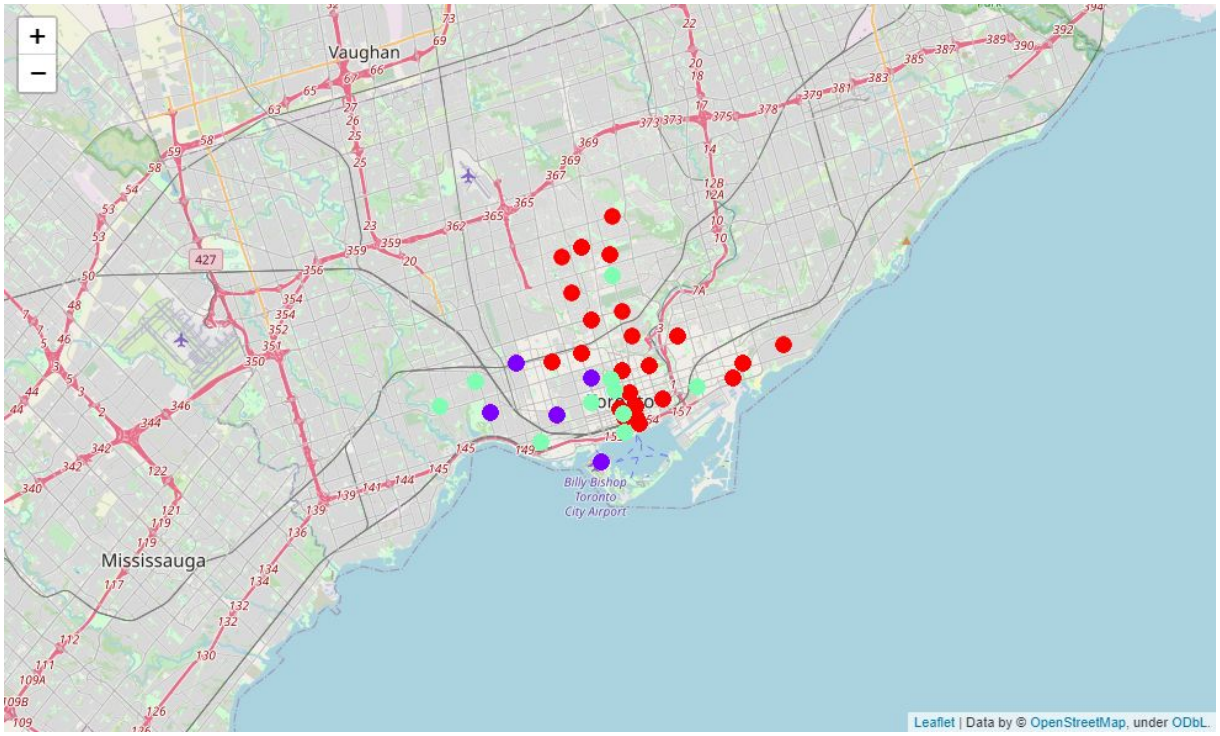
After preprocessing the neighborhood data, I needed to get the coordinates of the boroughs to utilize Foursquare to pull the list of venues near these neighborhoods. To get the coordinates, I tried using Geocoder package but it was not working so I used the csv file provided by IBM team to match the coordinates of Toronto neighborhoods. After successfully gathering all these coordinates, I merged it with the initial dataset and then I visualized the map of Toronto using Folium package to verify whether these were correct coordinates.

Next, I used Foursquare API to pull the list of top 100 venues within 500 meters radius. I had created a Foursquare developer account in order to obtain the account ID and API key to pull the required venue data. By using the foursquare API, I pulled the venue data namely- names, categories, latitude and longitude of the venues. Then I ran some analysis on this data, checked how many unique categories I could get. Then, I analyzed each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. This was to prepare the dataset for clustering to be done later. I had looked specifically for bars.

Lastly, I performed the clustering method by using k-means clustering. K-means clustering algorithm identifies k number of centeriods, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and it was highly suited for this project as well. I had clustered the neighborhoods in Toronto into 3 clusters based on their frequency of occurrence for “Bars”. Based on the results (the concentration of clusters), I was able to recommend the ideal location to open the Bar.

5. RESULTS:

We categorized the Toronto neighborhoods into 3 clusters with the help of K-means based on how many Bars were in the neighborhood. The following are the resultant clusters that was formed:



- Cluster 0(Red): Neighborhoods with little or no Bars
- Cluster 1(Blue): Neighborhoods with moderate number of Bars
- Cluster 2(Green): Neighborhoods with significant number of Bars

6. DISCUSSION/RECOMMENDATION:

By the analysis of the above 3 clusters, it is noted that most of the Bars are in Cluster 2 which are around First Canadian Place , Underground city, Harbourfront East , Union Station , and Toronto Islands. The lowest number of bars are in Cluster 0 areas which are Toronto Dominion Centre , Design Exchange, Commerce Court , Victoria Hotel, Garden District, Ryerson, etc. Hence, there are good opportunities to open the bar in these areas of Cluster 0 as competition seems to be low and it could attract the nearby crowd more often than compared to other clusters. Based on the results obtained by the analysis, it can be concluded that the neighborhoods of Cluster 0 might be a good location as there are not a lot of Bars in these areas for people to choose from. Therefore, this project recommends the entrepreneur to open a Bar in a suitable neighborhood of Cluster 0.

7. LIMITATIONS AND SCOPE FOR FUTURE IMPROVEMENT :

In this project, I only take one factor into consideration: the occurrence / existence of Bars in each neighborhood. There are many factors that can be taken into consideration such as population density, income of residents, rent that could influence the decision to open a Bar, Age group of people, etc. However, to put all these data into this project was not possible to do within a short time frame for this capstone project. Future improvements can certainly be taken into consideration of these factors. In addition, I am relying on the existence of Bars only for this project but future research can also be taken into consideration of other variables such as existence of Pubs, Nightclubs, etc in each neighborhood.

8. CONCLUSION:

In this project, we went through the process of identifying the business problem, specifying the data required, extracting and preparing the data, explaining the methodology of performing the machine learning by utilizing k-means clustering and providing best recommendation to the stakeholder.