Capstone Project-The battle of neighborhood

1. Introduction/Business Problem

Toronto is the provincial capital of Ontario and the most populous city in Canada, with a population of 2,731,571 in 2016. Current to 2016, the Toronto census metropolitan area (CMA), of which the majority is within the Greater Toronto Area (GTA), held a population of 5,928,040, making it Canada's most populous CMA. The city is the anchor of the Golden Horseshoe, an urban agglomeration of 9,245,438 people (as of 2016) surrounding the western end of Lake Ontario. Toronto is an international centre of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world.

The idea of this project is about open a new Chinese restaurant in Toronto. By using data science methods and machine learning methods such as clustering, I will help people to choose the right location by providing a cluster map of Toronto and the data about income and population of each neighborhood.

2. Data

There are two kinds of Data required:

(1) List of Toronto's neighborhoods with their Latitude and Longitude.

Description: Scrapping of Toronto neighborhoods via Wikipedia, getting Latitude and Longitude data of these neighborhoods via Geocoder package, using Foursquare API to get venue data related to these neighborhoods.

(2) Toronto's 2016 Census.

Description: Comparing the population and income of each cluster and find the best location for restaurant. The website: https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/#8c732154-5012-9afe-d0cd-ba3ffc813d5a contains Population, Average income for each of the Neighborhood.

3. Methodology

First, I find the list of neighborhoods in Toronto from our lab resource ("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 is easier and more convenient to pull tabular data directly from a web page into dataframe. However, it is not enough because there is only a list of neighborhood names and postal codes. So I have to find their coordinates from Foursquare to pull the list of venues near these neighborhoods. After gathering all these coordinates, I visualized the map of Toronto using Folium package to verify whether these are correct coordinates.

After that, I use Foursquare API to pull the list of top 100 venues within 500 meters radius. From Foursquare, I am able to pull the names, categories, latitude and longitude of the venues. With this data, I checked how many unique categories that I have from these venues.

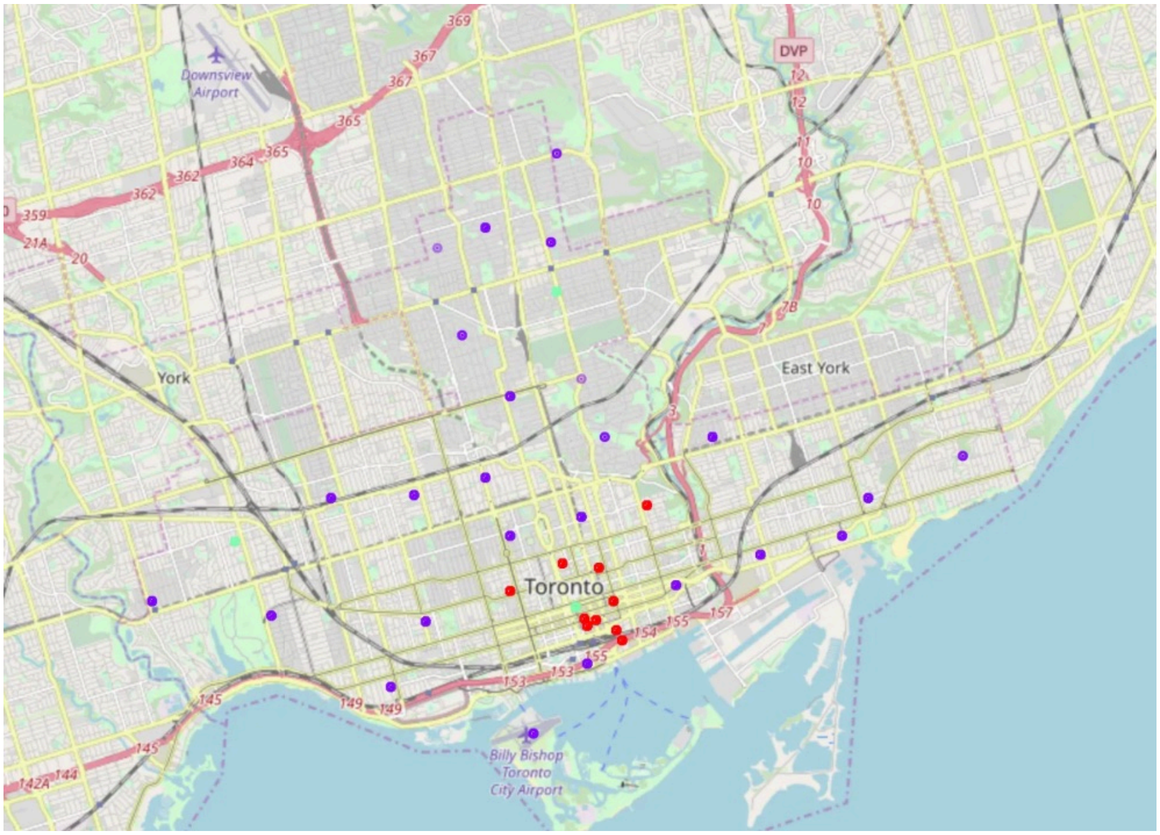
Then, I analyze each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category, which is prepare for the clustering.

After that, 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 is highly suited for this project as well. I have clustered the neighborhoods in Toronto into 3 clusters based on their frequency of occurrence for "Asian".

Finally, I also take local income and population into account. In order to do that I've used the 2016 Census information to display the wealthier and more populational neighborhoods and Foursquare data to display the current restaurants in each region.

4. Results

The Clusters



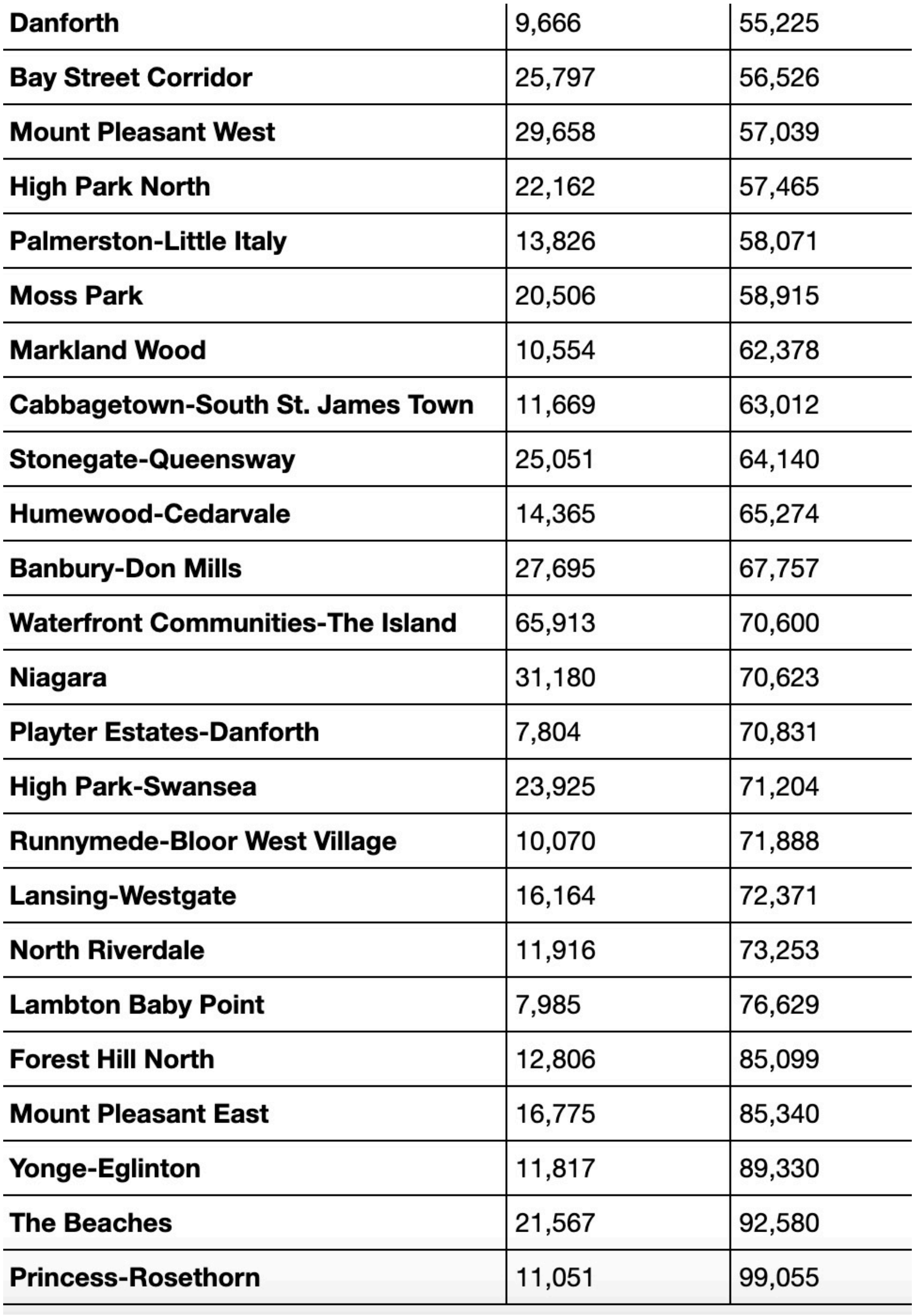
K-means clustering show that we can categorize Toronto neighborhoods into 3 clusters based on how many Asian restaurants are in each neighborhood:

Cluster 0(Red Dots): Neighborhoods with little or no Asian restaurants

Cluster 1(Purple Dots): Neighborhoods with no Asian restaurants

Cluster 2(Light Green): Neighborhoods with high number of Asian restaurants

The Population and Income



Danforth	9,666	55,225
Bay Street Corridor	25,797	56,526
Mount Pleasant West	29,658	57,039
High Park North	22,162	57,465
Palmerston-Little Italy	13,826	58,071
Moss Park	20,506	58,915
Markland Wood	10,554	62,378
Cabbagetown-South St. James Town	11,669	63,012
Stonegate-Queensway	25,051	64,140
Humewood-Cedarvale	14,365	65,274
Banbury-Don Mills	27,695	67,757
Waterfront Communities-The Island	65,913	70,600
Niagara	31,180	70,623
Playter Estates-Danforth	7,804	70,831
High Park-Swansea	23,925	71,204
Runnymede-Bloor West Village	10,070	71,888
Lansing-Westgate	16,164	72,371
North Riverdale	11,916	73,253
Lambton Baby Point	7,985	76,629
Forest Hill North	12,806	85,099
Mount Pleasant East	16,775	85,340
Yonge-Eglinton	11,817	89,330
The Beaches	21,567	92,580
Princess-Rosethorn	11,051	99,055

5. Discussion

Most of Asian restaurants are in Cluster 2 which is around Adelaide, King, Richmond areas and lowest (close to zero) in Cluster 1 areas which are North Toronto West and Parkdale areas. Also, there are good opportunities to open near Chinatown, St James town as the competition seems to be low. Looking at nearby venues, it seems Cluster 1 might be a good location because there are not a lot of Asian restaurants in these areas.

Also, the majority of the restaurants grouped on main streets and on the south of the city, although some of the richest neighborhoods are up to the north.

6. Conclusion

This report may be helpful for someone planning on opening a restaurant in Toronto, by comparing the current offers and neighborhoods profiles, local income and populations. However, it may not cover all variables such as access to public transportation.