Where to Expand?

Ralph Bou Ghanem

June 07, 2020

1. Introduction

1.1 Background

Having to decide where to expand your business is one of the most risky decisions. Are you a coffee shop owner? do you want to expand your business and maybe start a franchise? This report will compare the most common venues between Manhattan and Toronto. and will let you decide based on a thorough data studies.

1.2 Problem

To expand a business or a franchise is a big investment. With big investments comes big risks. If a decision was taken wrong, all your investment, your money, and most importantly your time will be wasted.

1.3 Interest

Obviously, coffee shop franchise owners who are considering to open one or more branches in Manhattan or Toronto, will be interested to read this report that will help them take the correspondent decision.

2. Data acquisition and cleaning

2.1 Data sources

The data of Manhattan New york was taken from the below mentioned link:

https://cocl.us/new_york_dataset

The data of Toronto was taken from Wikipedia page from. Link mentioned below:

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Data about the venues were imported from foursquare

2.2 Data cleaning

Data downloaded from multiple sources, were cleaned firstly by removing the non assigned values in each column after transforming the data to a dataframe using pandas library. After that, non essential columns were removed and coordinates columns were added based on the postal code of each borough. After cleaning the data of NY we got the below table:

Out[146]:

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

2.3 Feature selection

After data cleaning, I chose from NY Manhattan and from Canada Toronto. For Toronto I get the below Table:

	Postal Code	Borough	Neighborhood	Latitude	Longitude
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
9	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937
15	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418
19	M4E	East Toronto	The Beaches	43.676357	-79.293031

3. Exploratory Data Analysis

My target is to explore the top 5 venues in each neighborhood/city. That is why, and after getting the needed data from foursquare, we were able to get the below table for the 5 venues in Toronto:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Berczy Park	Coffee Shop	Farmers Market	Concert Hall	Cocktail Bar	Museum
1	Brockton, Parkdale Village, Exhibition Place	Coffee Shop	Pet Store	Furniture / Home Store	Italian Restaurant	Gym
2	Business reply mail Processing Centre, South C...	Garden Center	Auto Workshop	Pizza Place	Burrito Place	Restaurant
3	CN Tower, King and Spadina, Railway Lands, Har...	Airport Lounge	Coffee Shop	Bar	Airport	Airport Food Court
4	Central Bay Street	Coffee Shop	Gastropub	Italian Restaurant	Modern European Restaurant	Middle Eastern Restaurant

And the below table for the 5 venues in Manhattan:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Battery Park City	Park	Hotel	Gym	Coffee Shop	Memorial Site
1	Carnegie Hill	Coffee Shop	Pizza Place	Café	Bookstore	Gym / Fitness Center
2	Central Harlem	African Restaurant	Chinese Restaurant	Gym / Fitness Center	American Restaurant	Cosmetics Shop
3	Chelsea	Art Gallery	Coffee Shop	Ice Cream Shop	Café	American Restaurant
4	Chinatown	Chinese Restaurant	Bakery	Cocktail Bar	Bubble Tea Shop	Ice Cream Shop

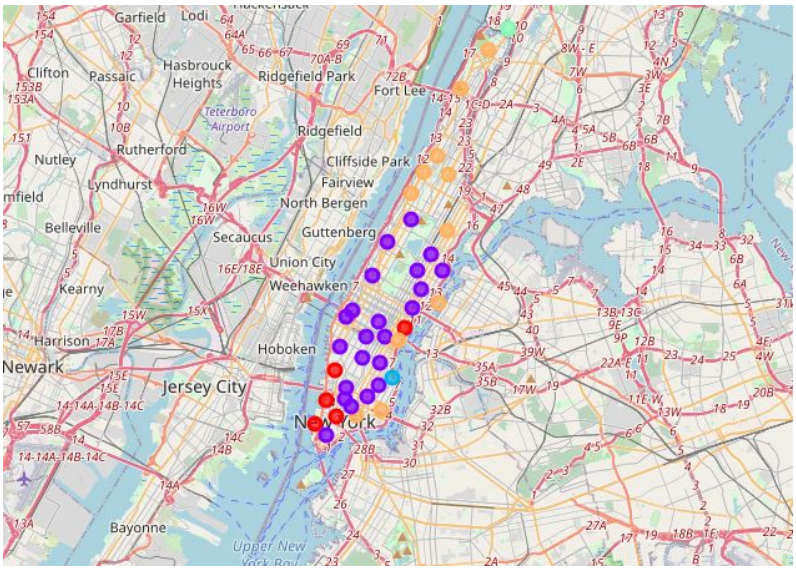
4. Predictive Modeling

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. It is popular for cluster analysis in data mining. k-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. For instance, better Euclidean solutions can be found using k-medians and k-medoids.

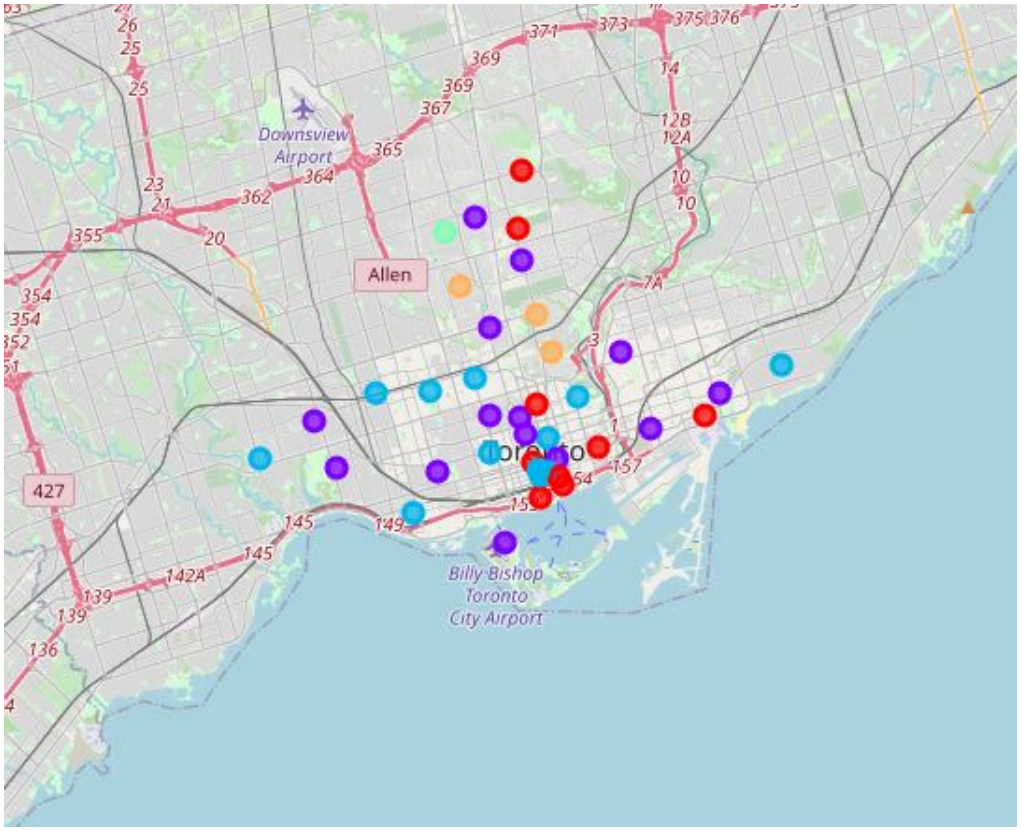
The problem is computationally difficult (NP-hard); however, efficient heuristic algorithms converge quickly to a local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both k-means and Gaussian mixture modeling. They both use cluster centers to model the data; however, k-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

The algorithm has a loose relationship to the k-nearest neighbor classifier, a popular machine learning technique for classification that is often confused with k-means due to the name. Applying the 1-nearest neighbor classifier to the cluster centers obtained by k-means classifies new data into the existing clusters. This is known as nearest centroid classifier or Rocchio algorithm.

The clustering for this project was visualized on two maps, one for Toronto and Other for Manhattan:



Toronto:


Downsview Airport
Allen
427
Billy Bishop Toronto City Airport

6. Conclusion

What we can see from the above study, is that in both cities, we can see an interest in coffee shops since it is between the top 5 venues. However, starting the business in Toronto has a higher possibility of success taking into considering that your business presents high end coffee quality and can make its way to the top 5 venues. In Manhattan, a large quantity of coffee shops is already on the top 5 which will make starting the business in Manhattan more competitive.