Potential Retail Line Investment In Toronto and New York

Lim Eng Lee

10 Jun 2020

1. Introduction

New York City and the city of Toronto are among the world busiest financial centres, and the vibrancies in these cities cater for vast amount of opportunities in setting up retail business. However, prior to putting in limited resources and investment, due diligence would be required to ensure that the highest chance of business success.

It would be appropriate to carry out the initial analysis using data as a platform to understand the differences and nature of retail business around the two locations. The primary questions would be as follows:

a. What would be the top 5 retail shops in both Toronto and New York business districts?

b. If someone wants to setup a retail line in Toronto or New York, which retail line would be a safe bet given the current cluster of shops in these cities?

Answering the above questions contribute towards the right direction subsequently during decision making to setup retail business. Here Python would be used as the data analysis platform to sieve out the required answers amongst data gather from various open sources.

2. Data Acquisition and Cleaning

2.1 Data Requirements

In order to identify the popular retail lines and enable higher probability of success when setting up new retail line within these cities, appropriate data would require to be gathered to perform the necessary analysis. The following data would be required to carry out analysis:

a. The list and type of retail shops in Toronto and New York business districts

b. The ratings and reviews on the type of popular retail lines in Toronto and New York

c. If trending data can be obtained, it will be an added advantage in determining the potential retail line that can be invested in.

2.2 Import Relevant Libraries for Python

To carry out analysis of data, the source of data would need to be identified upfront. In this case, data is obtained from existing open sources. Relevant libraries for Python are loaded to enable more efficient data processing.

2.3 Load and Preparing Data

There are two main sets of location data before performing the analysis, namely the data for New York City and City of Toronto. Here we will focus on the main business districts of both the cities, i.e. Manhatten and heart of Toronto.

The dataset for New York city is downloaded from existing available sourc from the internet (URL : https://cocl.us/new_york_dataset). Data is downloaded in JSON format and then converted into python dataframe format. Relevant data on the heart of New York (i.e. Manhattan) is then extracted and process from these data.

The dataset for Toronto is extracted from the existing web page (URL: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M). Data downloaded and processed, followed by getting the coordinates of the relevant location using the geopy functions found in Python library.

3. Exploratory Data Analysis

3.1 Get Data from Foursquare

The Foursquare API is then used to explore neighbourhoods in the heart of New York (i.e. Manhattan) and Toronto. Foursquare API allows the application developers to interact with the Foursquare platform to obtain data related to a location such as retails shops around the neighbourhoods, comments on these shops and the popular venues around the locations. The ***explore*** function under Foursquare API is used to get the most common venue categories in each neighbourhood, and then use this feature to group the neighbourhoods into clusters using the *k*-means clustering algorithm.

With both of the dataset ready, Foursquare functions are then invoked to get information on the top 100 venues within 500m radius from each of the respective neighbourhoods from Manhattan and Toronto obtained and processed in earlier steps.

Finally, the Folium library is used to visualize the neighbourhoods in New York City and their emerging clusters.

4. Predictive Modelling

With the venues' information received from Foursquare, the top 10 venues from the neighbourhoods of each cities is identified as shown in Table 4.1.

4.1 One-Hot Encoding

One-Hot encoding is then used to allow representation of categorical data in numerical forms to allow processing for predications. With the numerical form from One-Hot encoding, the top 5 categories within each neighbourhood of the two cities is identified as in Table 4.2. From the weightage is use to sieve out the tops venues from these locations. The list of top 5 categories for both Manhattan and Toronto are given in below Table.

Table 4.1 – Top 5 Categories for Manhattan and Toronto

S/N	Top Venues	
	Manhattan	**Toronto**
1	Coffee Shop	Coffee Shop
2	Italian Restaurant	Park
3	Park	Café
4	Pizza Place	Restaurant
5	Café	Italian Restaurant

4.2 Clustering

Other than using One-Hot encoding to understand the popularity of venue by categories for each of the neighbourhood, Clustering is used to group the neighbourhoods into five clusters. By clustering, it can be seen that there is a cluster that includes majority of the neighbourhoods for each city. Maps showing respective clustering for Manhattan and Toronto are as given Figure 4.1 and 4.2.

Figure 4.1 – Map of Manhattan with clustering of neighbourhoods

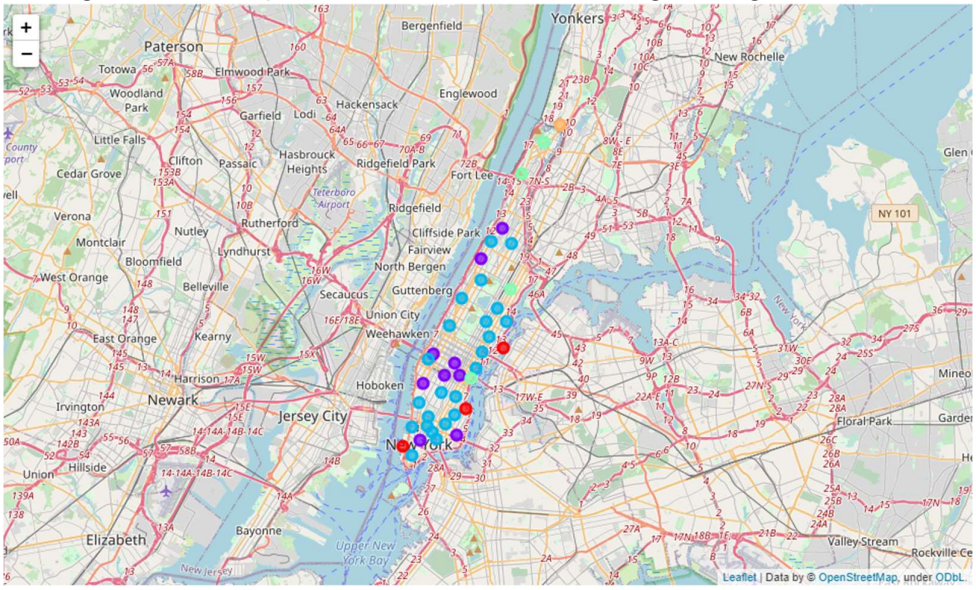
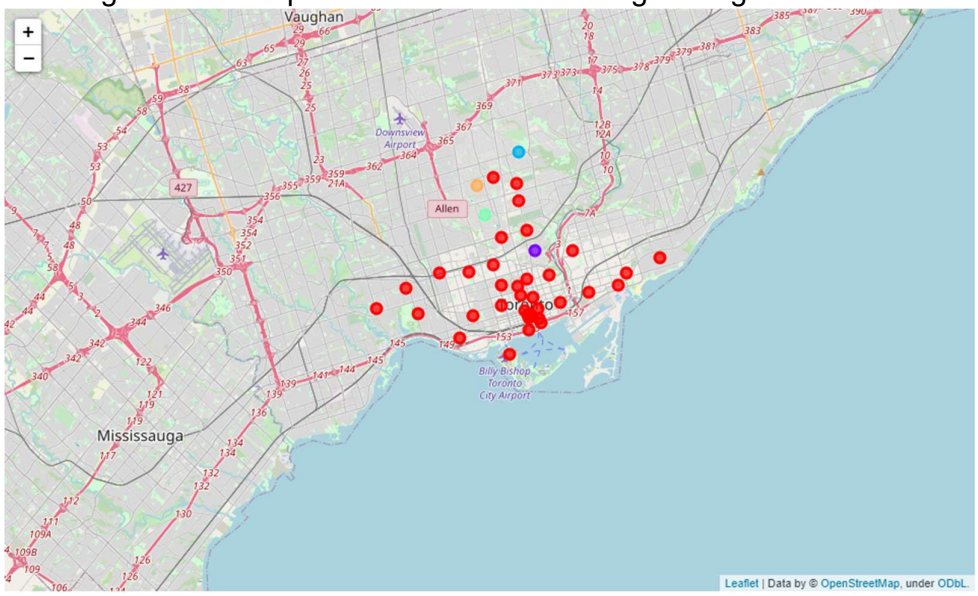


Figure 4.2 – Map of Toronto with clustering of neighbourhoods



Clustering enable focus in the appropriate neighbourhoods in the cities that have better prospects in the top categories that can be considered. The top five categories for Manhattan and Toronto clusters are as shown in Table 4.2 and Table 4.3 respectively.

Table 4.2 – Top 5 Categories for Manhattan Clusters

S/N	Manhattan Cluster				
	0	1	2	3	4
1	Park	Coffee Shop	Italian Restaurant	Mexican Restaurant	Sandwich Place
2	Playground	Hotel	Coffee Shop	Café	Coffee Shop
3	Coffee Shop	Café	Pizza Place	Bakery	Gym
4	Farmers Market	Pizza Place	Café	Deli / Bodega	Yoga Studio
5	Baseball Field	American Restaurant	Bar	Park	Supplement Shop

Table 4.3 – Top 5 Categories for Toronto Clusters

S/N	Toronto Cluster				
	0	1	2	3	4
1	Coffee Shop	Park	Park	Mexican Restaurant	Music Venue
2	Café	Playground	Swim School	Jewelry Store	Ice Cream Shop
3	Park	Trail	Bus Line	Trail	Garden
4	Restaurant	Department Store	Yoga Studio	Sushi Restaurant	Dog Run
5	Italian Restaurant	Event Space	Diner	Yoga Studio	Dessert Shop

From the above tables, the cluster that includes majority of neighbourhoods are Cluster 2 for Manhattan and Cluster 0 for Toronto as highlighted. Both cities have much similarities with eateries in most of the top five categories. For the clusters with majority of neighbourhoods for both cities, there are much similarity with Coffee Shop, Café and Italian Restaurant in the top five categories. Hence, it seems like eateries will be a safe bet in both cities, particularly the popular categories include Coffee Shop, Café and Italian restaurant.

5. Conclusion

From the findings and analysis in paragraph 4, it can be seen that for respective clusters in the cities, different categories of business may have better chances of success if it's aligned to the top categories in that respective cluster. However, it would be easier to start off business in the cluster that includes majority of the neighbourhoods.

It can be seen that although both cities are few hundred miles apart, there are many similarities in the top categories in their neighbourhoods. With these similarities, it there may even be potential growth of chained outlets in both cities. Hence, it gives more options in terms of venue if there's a need to setup any of the top categories in the identified neighbourhoods.