Battle of restaurants between Downtown Toronto and Manhattan

T. Rashid
Capstone Project

IBM Data Science Professional Certificate

1. Introduction

A great restaurant area must have diverse selection of food. Both Manhattan and Toronto have always been food lover's cities. Both cities are known for their diversity in food and culture. This project is to analyze the restaurants of the Downtown Toronto and Manhattan by classifying them into different associated with the restaurants. The main idea is to explore the restaurants and cluster them together on the basis of top 5 most common cuisine in each neighborhood of Toronto and Manhattan using K-means Clustering and display them on the geographical map.

This analysis can be helpful for somebody who wants to open a restaurant in either of the cities. It can help the person to decide what kind of restaurants are more common in big cities. This research cannot be solely used to decide the opening of the restaurant but it can give the future owner of the restaurant an idea of how the restaurants related to different cuisines are clustered together in the neighborhoods of the two food capital.

2. Data Preparation

We are going to leverage the data from Foursquare's API, of different restaurants that includes the location of the restaurant and category of the cuisine. We will also use some of the data from the CSV files that helps us with the name of the neighborhoods along with their longitudes and latitudes.

We are going to leverage the 'Venue Category' available in the Foursquare's database. Since, there are so many cuisines, we are going to narrow them down to seven main categories i.e. American, Latin, European, Asian, Casual, Middle Eastern and Other. In order to do that we pulled the data that contains keywords like Taco, Pizza, Restaurant, and Sandwich etc. under the 'Venue Category'.

There has been some assumptions made in terms of the data related to restaurants. For example, Indian, Afghani, Japanese and Chinese cuisines have been assigned to 'Asian' category as the countries associated with the cuisines do fall under 'Asian' continent, however, the cuisines are totally different from each other. Also, pizza is considered 'European' and tacos would have fallen under 'Latin' food category but they are put under 'Casual' category.

The analysis will be as good as the data provided. Hence, if some restaurants are not available in the Foursquare API, they won't be included in the analysis.

<u>Information about Toronto data:</u>

1. The Toronto data is available on Wikipedia page that consists of three columns: Postal Code, Community, and Neighborhood.
2. We will convert the data into CSV file and then to dataframe to perform analysis.
3. We will only process the cells that have an assigned borough and remove the cells with a borough that is 'Not assigned'.

4. There exist more than one neighborhood for one postal code area. For example, in the table on the Wikipedia page, you will notice that M5A is listed twice and has two neighborhoods: Harbourfront and Regent Park. These two rows will be combined into one row with the neighborhoods separated with a comma.
5. If a cell has a borough but a 'Not assigned' neighborhood, then we will substitute its name with the name of its neighborhood.

Information about Manhattan data:

1. New York City comprise of 5 boroughs: Bronx, Brooklyn, Manhattan, Queens and Staten Island.
2. For this analysis, we will be using restaurant locations available in Manhattan only.

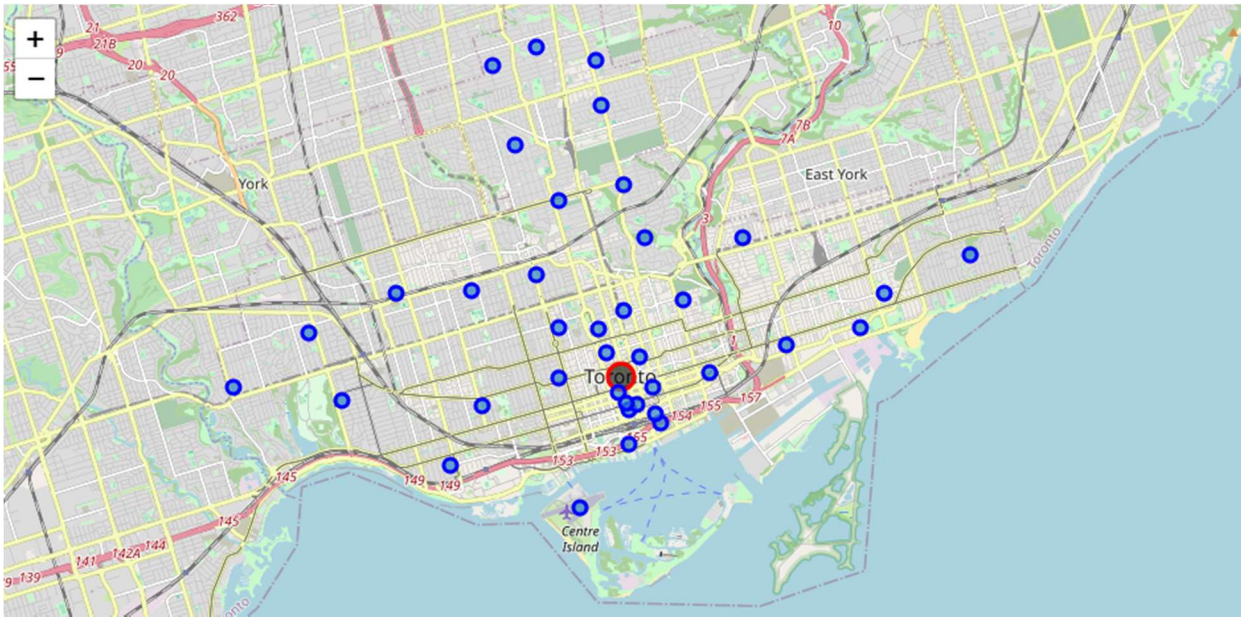
3. Methodology

We will analyze the data using Segmentation and K-means Clustering and visualize it on a geographical map to get better idea how the restaurants are being clustered in the neighborhoods. We will also look into the distribution by categories of restaurants in each of the cities. At the end we will suggest some of the potential ways to find an ideal location for an upcoming restaurant.

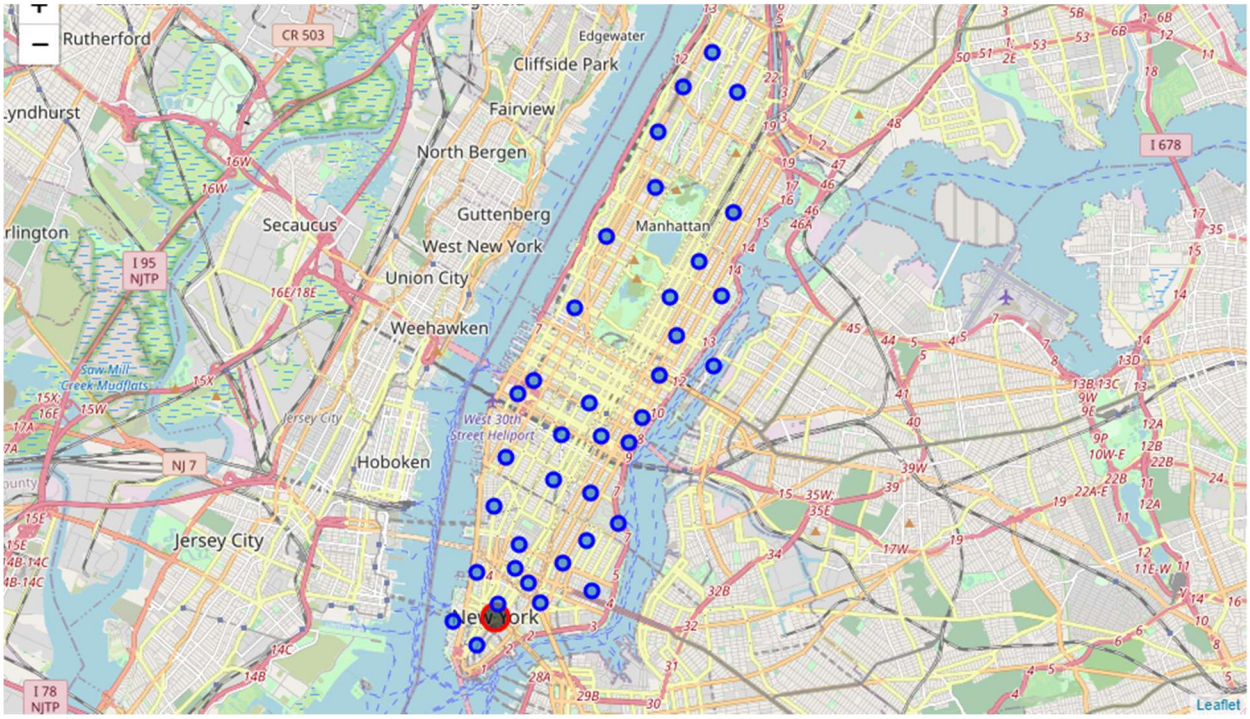
At this point we need to perform some exploratory data analysis to cluster these neighborhoods based on the restaurant categories. We are going to use folium library to visualize the geographical details of Toronto and Manhattan along with their neighborhoods.

In the following maps, the blue dots represents the Neighborhoods and the black dot with the red circle shows the coordinates of Toronto and Manhattan.

Toronto:



Manhattan:



We will leverage the Foursquare API to explore the restaurants in the neighborhoods, distance from the neighborhood, category of the restaurant etc. For this analysis we restricted the list of restaurants to 100 venues and 500 meter radius for each neighborhood from their given latitude and longitude.

Here is the head of the data after the coordinates were merged together with the list of restaurants obtained using Foursquare API. Please note that the restaurants were divided into seven main categories i.e. American, Latino, Euro, Asian, Casual, Middle Eastern and Other by creating a separate column called as 'categories_class'.

Toronto:

[25]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	distance	Venue Category	categories_class
0	Regent Park, Harbourfront	43.654260	-79.360636	Impact Kitchen	43.656369	-79.356980	376	Restaurant	American
1	Regent Park, Harbourfront	43.654260	-79.360636	El Catrin	43.650601	-79.358920	430	Mexican Restaurant	Latin
2	Regent Park, Harbourfront	43.654260	-79.360636	Cluny Bistro & Boulangerie	43.650565	-79.357843	468	French Restaurant	European
3	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	Nando's	43.661728	-79.386391	257	Portuguese Restaurant	European
4	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	Mercatto	43.660391	-79.387664	258	Italian Restaurant	European
...	...	...	...	...	...	...	...	...	...

Manhattan:

[44]:

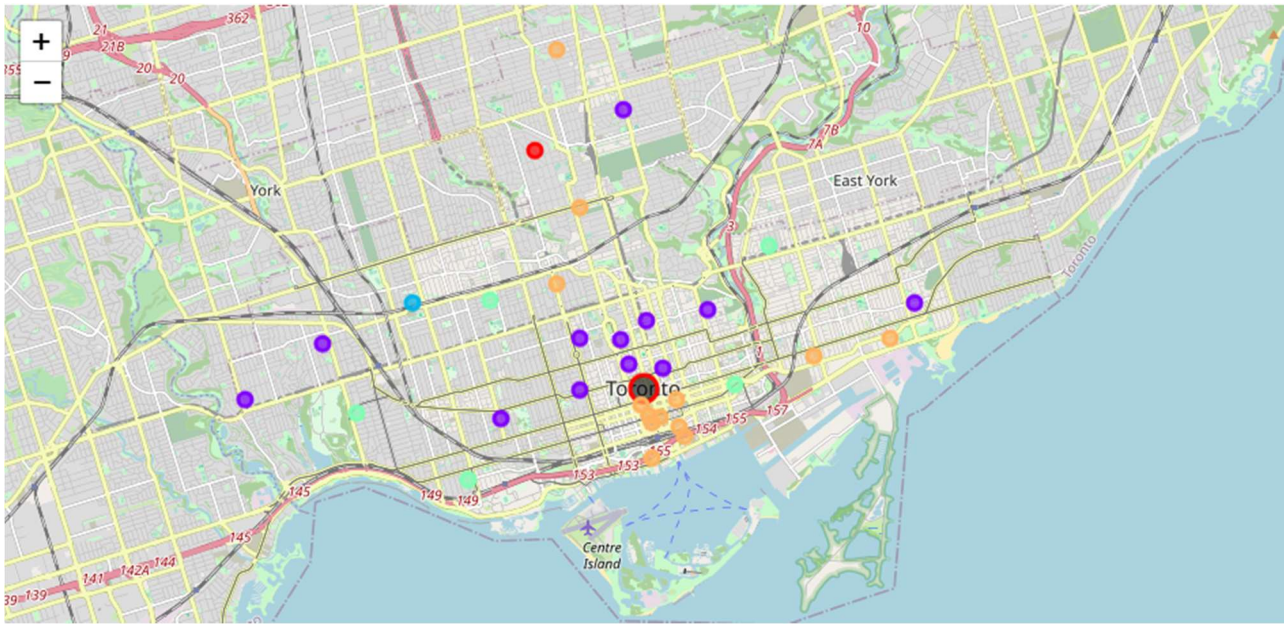
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	distance	Venue Category	categories_class
0	Marble Hill	40.876551	-73.910660	Land & Sea Restaurant	40.877885	-73.905873	429	Seafood Restaurant	American
1	Chinatown	40.715618	-73.994279	Spicy Village	40.717010	-73.993530	167	Chinese Restaurant	Asian
2	Chinatown	40.715618	-73.994279	Kiki's	40.714476	-73.992036	228	Greek Restaurant	European
3	Chinatown	40.715618	-73.994279	Wah Fung Number 1 Fast Food 華豐快餐店	40.717278	-73.994177	184	Chinese Restaurant	Asian
4	Chinatown	40.715618	-73.994279	Xi'an Famous Foods	40.715232	-73.997263	255	Chinese Restaurant	Asian
...	...	...	...	...	...	...	...	...	...

For this analysis we will be using unsupervised k-means algorithm to cluster the neighborhoods. First, we will group rows by neighborhood and then take the mean of the frequency of occurrence of each of the restaurant category. This will help us cluster the neighborhoods together. We are going to cluster the neighborhoods into 5 clusters and then we will define each cluster.

4. Results

Here are the results based on our analysis. The observation is based on the top 5 most common restaurant categories in a particular neighborhood.

Toronto:



The neighborhoods of Toronto clustered together using k-means algorithm are based on the categories and frequency of occurrence of the restaurants defined as follows:

Cluster #0: Asian restaurants

Cluster #1: Middle Eastern restaurants (The neighborhood seems unique/outlier)

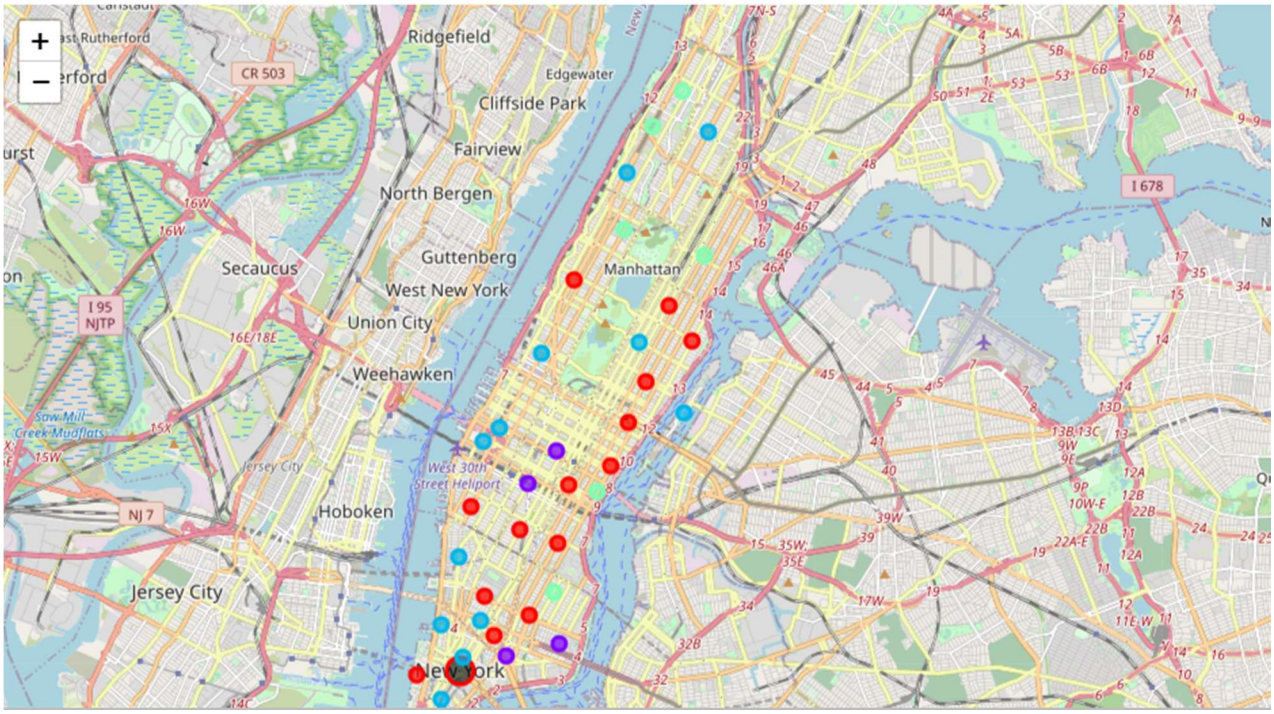
Cluster #2: European restaurants

Cluster #3: American restaurants

Cluster #4: Asian restaurants (The neighborhood seems unique/outlier)

Based on the clusters #0 and #3, Asian and American restaurants are mostly concentrated together in the neighborhoods close to each other.

Manhattan:



The neighborhoods of Manhattan clustered together using k-means algorithm are based on the categories and frequency of occurrence of the restaurants defined as follows:

Cluster #0: Asian restaurants (American and Latin being 2nd most common)

Cluster #1: American and European restaurants

Cluster #2: Asian and Latin restaurants

Cluster #3: American restaurants (The neighborhood seems unique/outlier)

Cluster #4: Asian restaurants (European being 2nd most common)

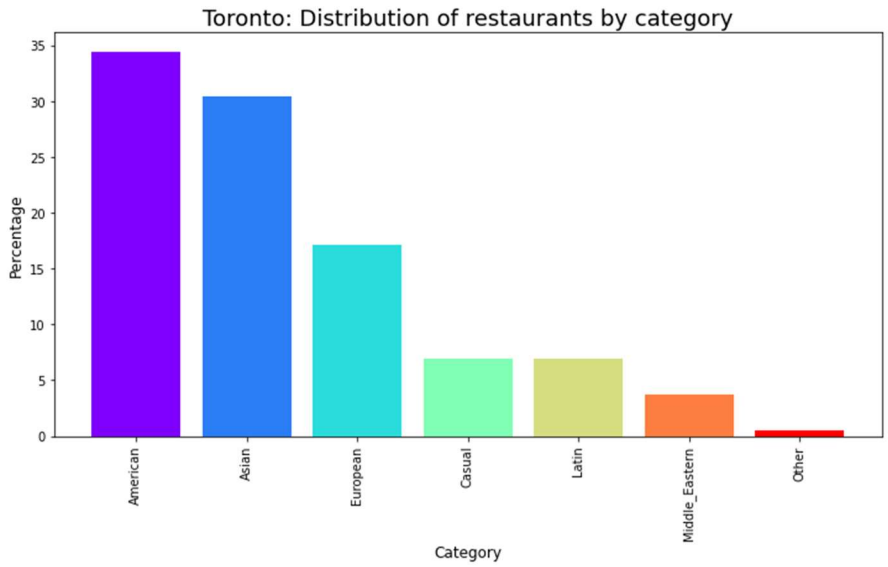
Based on the cluster #4, Asian restaurants are spread out evenly in most of the Manhattan's neighborhoods along with European restaurants.

5. Discussion

Here are some charts that shows the distribution of the restaurants by category in the city of Toronto and Manhattan.

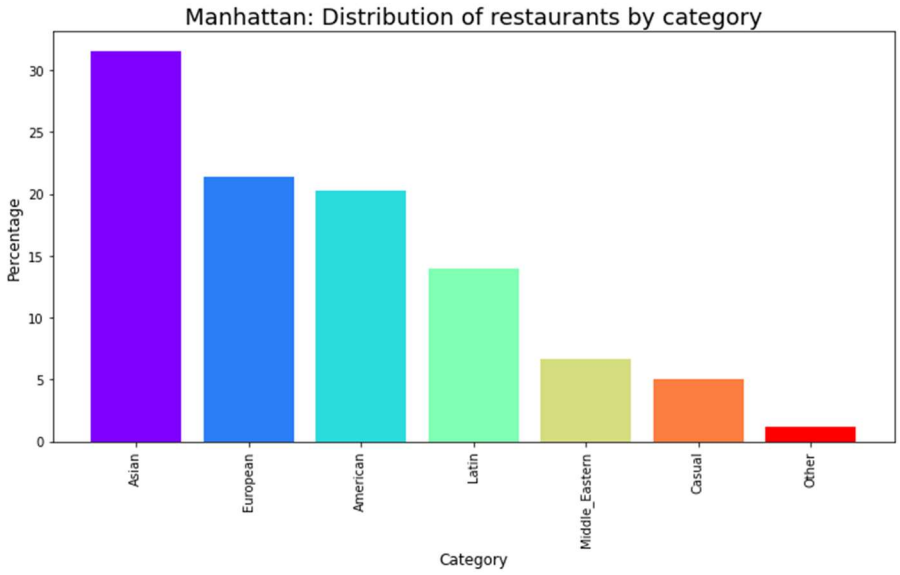
Toronto:

Based on the chart below we can conclude that Toronto has majority of American (~34%) restaurants followed by Asian (~30%) and European (~17%) restaurants.

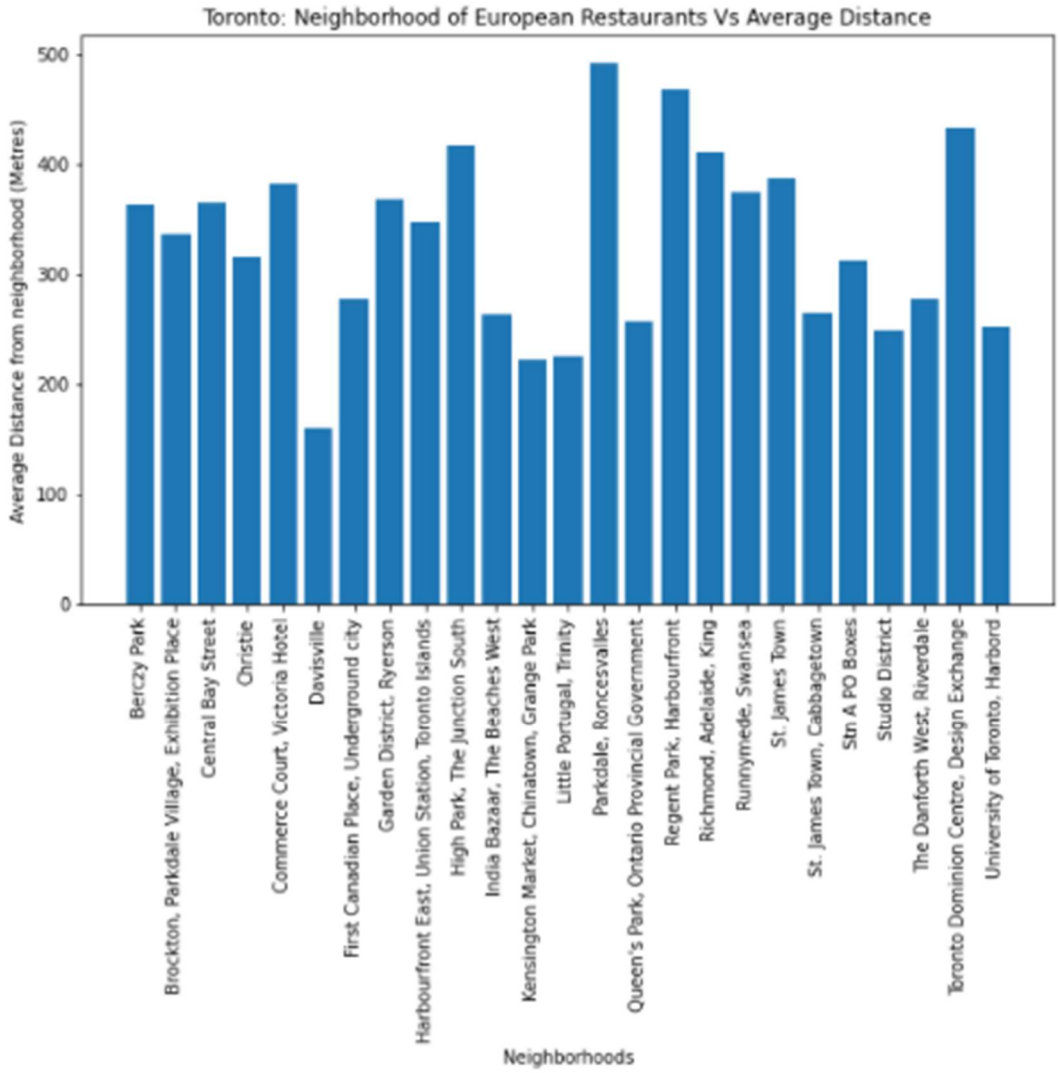


Manhattan:

Based on the above chart we can conclude that Manhattan has majority of Asian (~32%) restaurants followed by European (~21%) and American (~20%) restaurants.



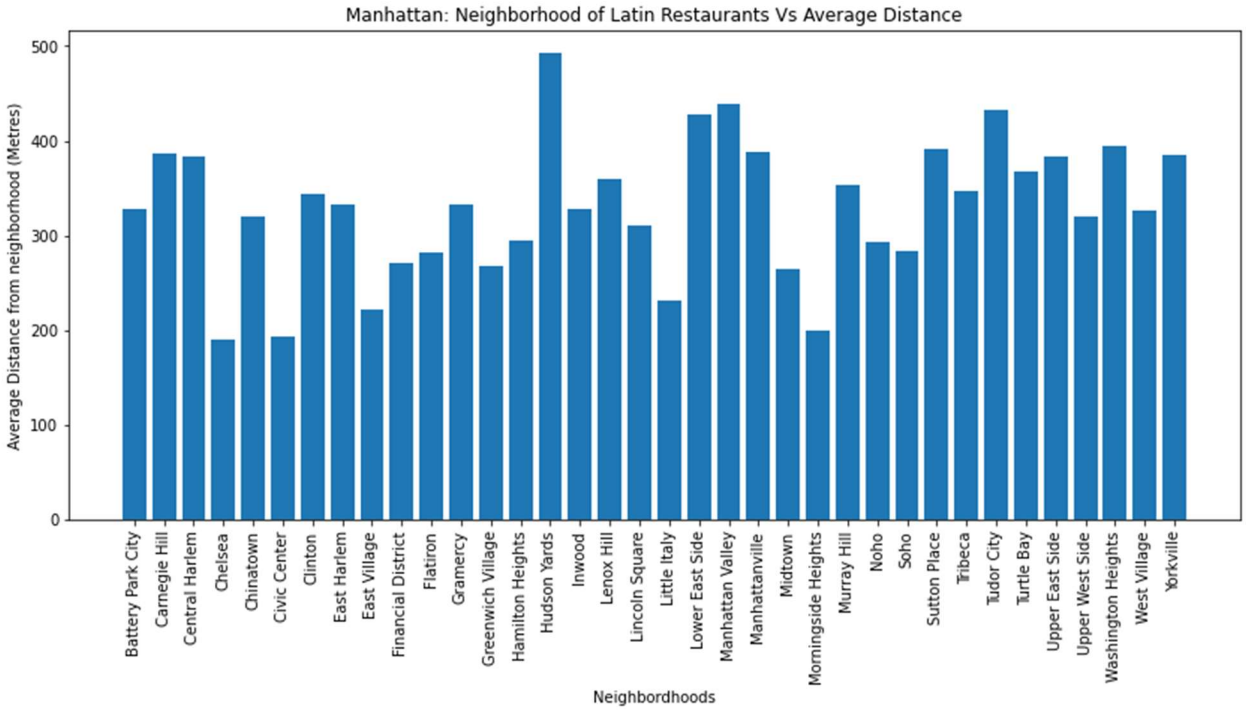
The following chart shows the average distance between the neighborhood and European restaurants. Let's say if there is a future restaurant owner and wants to open a European restaurant in Toronto then he/she would need to look at the following neighborhoods as shown. Then check the cluster that is assigned to that neighborhood and see if the European restaurants are common within that neighborhood. For example, Cluster #2 under Toronto has a hub of European restaurants. There is a great potential to open a European restaurant as it is the most common restaurant category. Another way of looking at is that Cluster #2 has enough European restaurants so maybe it is not wise to open a European restaurant there. There is some potential in Cluster #0 to open European restaurant as it is the second most common type of restaurant.



	Neighbourhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Queen's Park, Ontario Provincial Government	0.0	European	Asian	Latin	Casual	American
2	Garden District, Ryerson	0.0	Asian	European	Casual	American	Middle_Eastern
5	Central Bay Street	0.0	Asian	European	American	Middle_Eastern	Casual
10	Little Portugal, Trinity	0.0	Asian	American	European	Latin	Other
14	India Bazaar, The Beaches West	0.0	European	Casual	Asian	American	Other
18	High Park, The Junction South	0.0	Latin	Casual	Asian	European	American
22	Davisville	0.0	Asian	European	American	Other	Middle_Eastern
23	University of Toronto, Harbord	0.0	Asian	European	American	Other	Middle_Eastern
24	Runnymede, Swansea	0.0	European	Asian	American	Middle_Eastern	Latin
25	Kensington Market, Chinatown, Grange Park	0.0	Asian	Latin	American	Casual	European
28	St. James Town, Cabbagetown	0.0	Asian	American	European	Latin	Other
30	Church and Wellesley	0.0	Asian	American	Middle_Eastern	Latin	Casual

The following chart shows the average distance between the neighborhood and Latin restaurants. Let's say if there is a future restaurant owner and wants to open a Latin restaurant in Manhattan then he/she would

need to look at the following neighborhoods as shown. Then check the cluster that is assigned to that neighborhood and see if the Latin restaurants are common within that neighborhood. For example, Cluster #2 under Manhattan is packed with Latin restaurants. So, may be it is not a good idea to open another Latin restaurant within the same cluster of neighborhoods. But seems like there is a great potential in Cluster #0 with Latin restaurants being second common in that neighborhood.



	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Chinatown	0.0	Asian	American	European	Latin	Other
15	Midtown	0.0	Asian	Latin	Middle_Eastern	European	Casual
20	Lower East Side	0.0	Asian	Latin	Middle_Eastern	European	American
33	Midtown South	0.0	Asian	American	Casual	Middle_Eastern	European

6. Conclusion

Overall, American and Asian are the most common type of category of restaurants in Toronto and Manhattan, respectively. Based on the analysis we can conclude that Toronto has Asian and American restaurants that are concentrated together in the neighborhoods. However, Manhattan has Asian restaurants that are spread out evenly in most of the neighborhoods along with European restaurants. The analysis performed is really a high level analysis using k-means clustering. This can be used as a starting point if someone is trying to open a restaurant in either of the cities.

The analysis is as good as the data. We noticed that there are some restaurants that cannot be found in Foursquare API but we see it on Google. Basically, the analysis can generate different results if all the restaurants are to be included. Also, the analysis is highly dependent on the accuracy of the latitudes and longitudes of the neighborhoods as well as the neighborhoods associated with the restaurants. Therefore, the project can be leveraged for the initial thoughts of opening a restaurant as we require more data to get comfortable with the ideal location of the restaurant.

In terms of comparing the neighborhoods of both the cities it would be a great idea to include some other metrics like foot traffic, ‘likes’ associated with the restaurants, ratings, tips, etc. Then we can cluster the restaurants together and analyze using Logistic Regression or Multiple Linear Regression that if the metrics chosen are better predictor to analyze the category of restaurants of Toronto and Manhattan.

For further analysis, we can analyze population density, crime rate, and average income of the population of the neighborhoods of both the cities. These metrics play a vital role in starting up new business in a particular neighborhood.