Advanced Data Science Capstone Project

Battle of the Neighborhoods

Introduction

In this project, we analyzed and compared between the neighborhoods of New York City, New York and Toronto, ON, Canada. Our client in North East Asia wants to set up a new North America branch office, and New York City and Toronto were selected as their finalists. The company wants some insight into the neighborhoods and local businesses in these two cities so that its employees can have quality of life while they work or live in the city. We explored the similarities and dissimilarities between certain neighborhoods in these two cities and we will suggest which neighborhoods would fit the culture of the client's employees the best.

Data

We acquired the data from various sources fort this project. The datasets consist of borough names, postal codes, neighborhood names, and their latitude and longitude information.
The following links were used to acquire the data.

1. https://geo.nyu.edu/catalog/nyu_2451_34572

	Borough	Neighborhood	Latitude	Longitude
0	Manhattan	Marble Hill	40.876551	-73.910660
1	Manhattan	Chinatown	40.715618	-73.994279
2	Manhattan	Washington Heights	40.851903	-73.936900
3	Manhattan	Inwood	40.867684	-73.921210
4	Manhattan	Hamilton Heights	40.823604	-73.949688

2. https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
3. http://cocl.us/Geospatial_data

	Postal Code	Borough	Neighborhood	Latitude	Longitude
37	M4E	East Toronto	The Beaches	43.676357	-79.293031
41	M4K	East Toronto	The Danforth West, Riverdale	43.679557	-79.352188
42	M4L	East Toronto	The Beaches West, India Bazaar	43.668999	-79.315572
43	M4M	East Toronto	Studio District	43.659526	-79.340923
44	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790

Foursquare API search feature was used to collect the venue information for each given neighborhood.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
1	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	Coffee Shop
4	Marble Hill	40.876551	-73.91066	Dunkin'	40.877136	-73.906666	Donut Shop

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	The Beaches	43.676357	-79.293031	Glen Manor Ravine	43.676821	-79.293942	Trail
1	The Beaches	43.676357	-79.293031	The Big Carrot Natural Food Market	43.678879	-79.297734	Health Food Store
2	The Beaches	43.676357	-79.293031	Grover Pub and Grub	43.679181	-79.297215	Pub
3	The Beaches	43.676357	-79.293031	Upper Beaches	43.680563	-79.292869	Neighborhood
4	The Danforth West, Riverdale	43.679557	-79.352188	Pantheon	43.677621	-79.351434	Greek Restaurant

We expect that the details about local business venues and their localities would provide insights into how easily the employees of our client access and enjoy the neighborhoods that their new branch office sits in.

Methodology

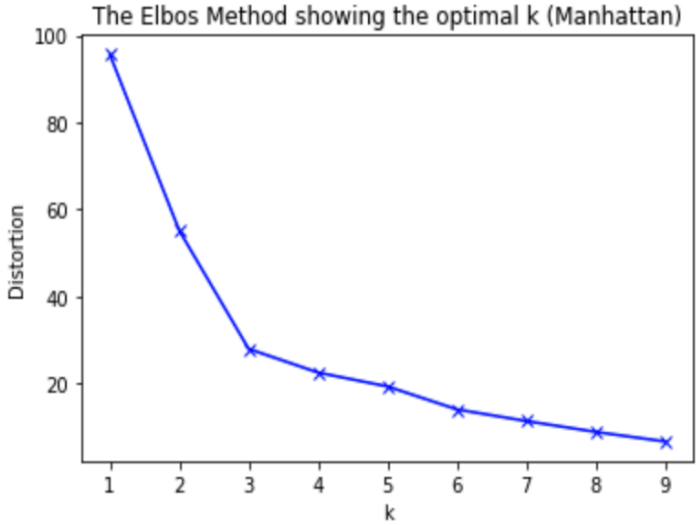
1. Foursquare API search feature was used to enable to collect the nearby venues of the neighborhoods in both New York City and Toronto. Due to the http request limitations, the number of places per neighborhood parameter was set to be 100 and the radius parameter was set to be 500. In order to download and handle New York City's JSON files and Toronto's web data, and retrieve location data based on their latitudes and longitudes, Python's JSON, BeautifulSoup, and Geopy libraries were used.
2. Extensive comparative analysis of two randomly picked neighborhoods, which are Manhattan and Old Toronto, was carried out to derive the desirable insights from the outcomes using Python's scientific libraries Pandas, NumPy, and Scikit-learn.
3. Folium, which is a Python visualization library, was used to visualize the neighborhoods cluster distribution of Manhattan and Old Toronto over the interactive leaflet maps.
4. K-means clustering, an unsupervised machine learning algorithm, was applied to form the clusters of different venue categories residing in and around the neighborhoods in both Manhattan and Old Toronto.

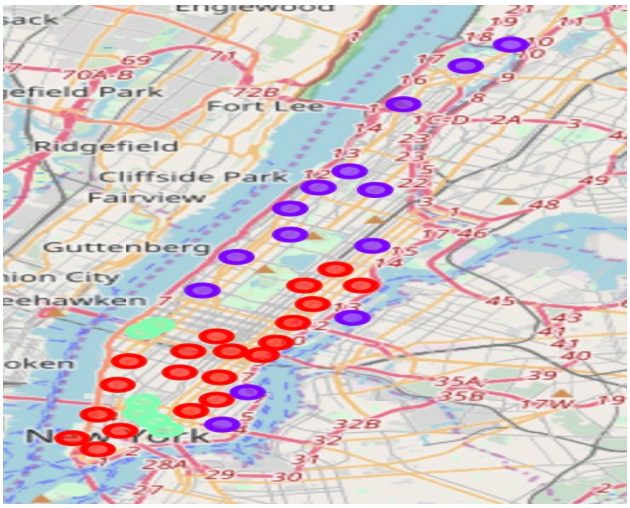
Results

Manhattan in New York City

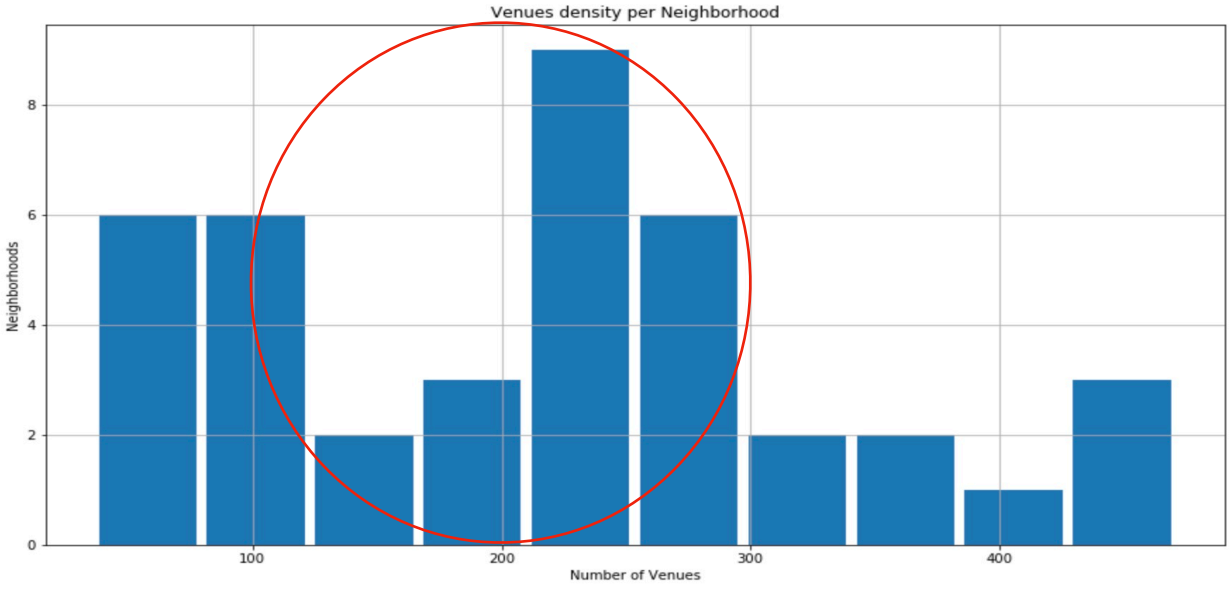
K-means clustering was used to group the neighborhoods in Manhattan into 3 clusters since the elbow method gave us 3 as the optimal number of clusters. The first cluster has 20 neighborhoods and the most common venues are coffee shops, restaurants, gym/fitness centers, and parks. The second cluster has 14 neighborhoods and its venues are mostly are

restaurants, coffee shops, bar/lounge and parks. And the third cluster includes 6 neighborhoods and its venues are restaurants, bakery, theater and clothing store.



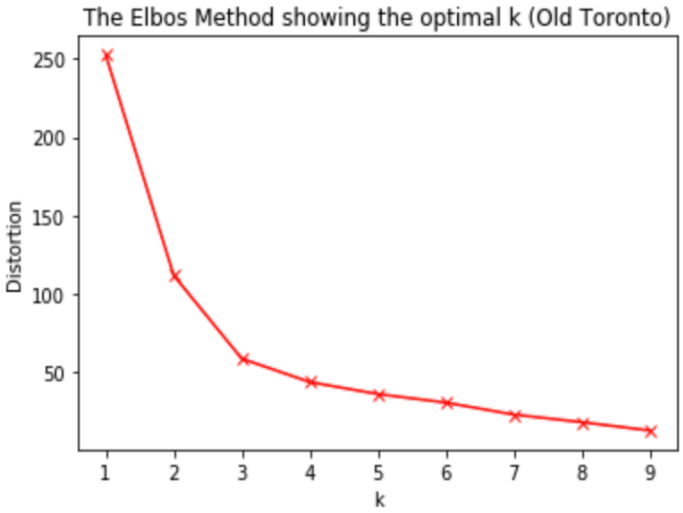


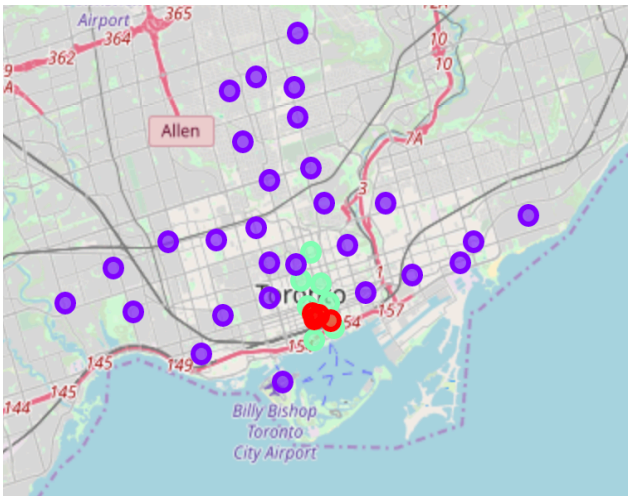
The density distribution shows that the over 50% of the neighborhoods presents a density between 100 and 270 venues per neighborhoods.



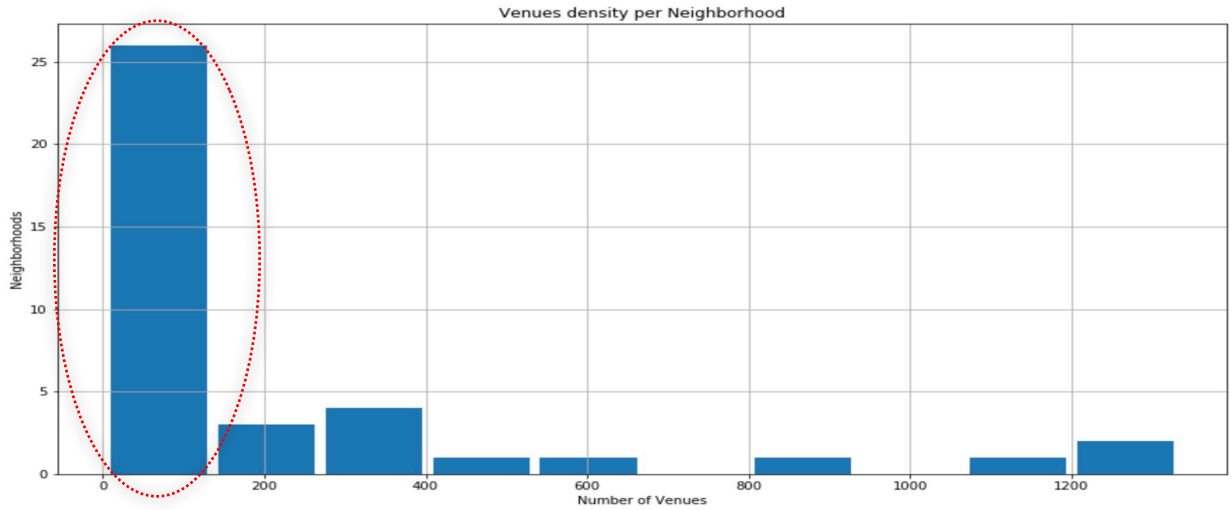
Old Toronto in Toronto, ON, Canada

K-means clustering was used to group the neighborhoods in Old Toronto into 3 clusters since the elbow method also gave us 3 as the optimal number of clusters. The first cluster has 28 postal codes and the most common venues are coffee shops/cafes, restaurants, grocery stores, bars/pub and parks. The second cluster has 4 postal codes and its venues are mostly are coffee shops/cafes. And the third cluster includes 7 postal codes and its venues are restaurants and coffee shops.





The density distribution shows that over 50% of the postal codes presents a density between 0 and 50 venues per postal codes.



Discussion

New York City has 5 boroughs and 306 neighborhoods and its geographical coordinate are 40.7127281, -74.0060152. In Manhattan, we found 3304 venues in 335 venue categories in 40 neighborhoods.

Toronto has 11 boroughs and 103 postal codes and its geographical coordinate are 43.653963, -79.387207. In Old Toronto, we found 1718 venues in 234 venue categories in 39 listed postal codes.

Given that each neighborhood has different radius passed to the venues request, we considered to present the venues per neighborhood in terms of density by distance. In Manhattan, over 50% of the neighborhoods have venues density between 100 and 270 venues per neighborhood, and in Old Toronto, over 50% of the postal codes have venues density between 0 and 50 venues per postal code.

Based on the elbow method, the optimal number of clusters are 3 for both Manhattan and Old Toronto. And the most common venues in both cities that we got are coffee shops or cafes, restaurants, parks, bars or pubs, and gym or fitness center.
Based on our analysis, Manhattan has more distinct venues in more venue categories overall and it also has more venues per neighborhood than Old Toronto as well.

Conclusion

Based on the quantity of venues, variety of venue categories, and venue density, we recommend Manhattan over Old Toronto as a location for the future branch office. Manhattan offers a lot more choices for coffee shops, restaurants, bars, parks, gym, and many other venues for office workers to use and enjoy conveniently while they work for this company.