Introduction/Business Problem

A large restaurant company that owns several brands and operates in several European countries seeks the opportunity in expanding its business in North America and more specifically in Toronto, Canada and New York City, USA.

Both are the largest cities and the financial capitals of their respective countries. They are incredibly famous thus attracting large numbers of tourists and business travellers from all over the world. The natural consequence is that both cities are well developed and multicultural.

The purpose of this project is to analyse and compare the cities of New York and Toronto to identify the most suitable neighbourhoods for our client to expand its business. Furthermore, the purpose is to distinguish and classify the several types of restaurants currently operating in those cities. The results will be used by our client to determine which of its brand(s) will be used for the expansion.

Finally, the results of this project could be used in similar cases where a company requires market analysis and segmentation to make a strategic decision.

Data

For both cities we require the datasets that segment the neighbourhoods providing the necessary coordinates. So, for New York the relevant data will be acquired from link (a) and for Toronto from link (b):

(a) https://geo.nyu.edu/catalog/nyu_2451_34572

(b) https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

For Toronto, a second file was provided containing the geographical coordinates in csv format.

Foursquare's database will be used to acquire all relevant data regarding restaurants based in New York and Toronto including their types and categories. By making the necessary API calls with Python we will collect the data that will be used to cluster and analyse cities' neighbourhoods. We will also define any similarities or dissimilarities between the cities and neighbourhoods.

In addition, from Foursquare the categorization of the venues was acquired in order to analyse the neighbourhoods.

Methodology

Data retrieval

Python code was used to retrieve data from the sources mentioned above. The required data was then stored in json or csv files for later retrieval. This was a necessary step as the large number of calls to Foursquare database was time consuming and,

in some cases, it resulted in the depletion of the limit thus preventing further calls until the time of the reset for the corresponding limits.

The main data structure that was used for data analysis and modelling was the Pandas DataFrame and in some cases plain Python Lists. Python dictionaries were the used to format and prepare the data provided by Foursquare through the API calls.

Data preparation

For New York, the json file was parsed to create a DataFrame with all Boroughs, Neighborhoods and their coordinates (Latitude and Longitude). The below given table shows the first 5 rows of that DataFrame.

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

For Toronto, data regarding Boroughs and Neighborhoods was retrieved from Wikipedia and the merged with the csv file that was provided by the course instructor containing coordinates. So, the resulting DataFrame was almost the same as the one for New York. The difference is that an extra column having the postal codes exists here and Neighborhoods were grouped according to the corresponding postal codes. An example with the first 5 rows of that data follows.

Postcode	Borough	Neighborhood	Latitude	Longitude
M1B	Scarborough	Malvern / Rouge	43.806686	-79.194353
M1C	Scarborough	Rouge Hill / Port Union / Highland Creek	43.784535	-79.160497
M1E	Scarborough	Guildwood / Morningside / West Hill	43.763573	-79.188711
M1G	Scarborough	Woburn	43.770992	-79.216917
M1H	Scarborough	Cedarbrae	43.773136	-79.239476

After that the API calls were made to retrieve the data about the venues. A special function was defined to retrieve the venues and create the corresponding DataFrame. The below given table shows and example for New York. The same was done for Toronto.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

Finally, the data regarding venue categories was retrieved and prepared with the use of API calls and the necessary json manipulation. The following table show and example having the 10 root categories.

id	name	pluralName	shortName	categories	icon.prefix	icon.suffix
4d4b7104d754a06370d81259	Arts & Entertainment	Arts & Entertainment	Arts & Entertainment	[{'id': '56aa371be4b08b9a8d5734db', 'name': 'A...	https://ss3.4sqi.net/img/categories_v2/arts_en...	.png
4d4b7105d754a06372d81259	College & University	Colleges & Universities	College & Education	[{'id': '4bf58dd8d48988d198941735', 'name': 'C...	https://ss3.4sqi.net/img/categories_v2/educati...	.png
4d4b7105d754a06373d81259	Event	Events	Event	[{'id': '52f2ab2ebcbc57f1066b8b3b', 'name': 'C...	https://ss3.4sqi.net/img/categories_v2/event/d...	.png
4d4b7105d754a06374d81259	Food	Food	Food	[{'id': '503288ae91d4c4b30a586d67', 'name': 'A...	https://ss3.4sqi.net/img/categories_v2/food/de...	.png
4d4b7105d754a06376d81259	Nightlife Spot	Nightlife Spots	Nightlife	[{'id': '4bf58dd8d48988d116941735', 'name': 'B...	https://ss3.4sqi.net/img/categories_v2/nightli...	.png
4d4b7105d754a06377d81259	Outdoors & Recreation	Outdoors & Recreation	Outdoors & Recreation	[{'id': '4f4528bc4b90abdf24c9de85', 'name': 'A...	https://ss3.4sqi.net/img/categories_v2/parks_o...	.png
4d4b7105d754a06375d81259	Professional & Other Places	Professional & Other Places	Professional	[{'id': '4e52d2d203646f7c19daa8ae', 'name': 'A...	https://ss3.4sqi.net/img/categories_v2/buildin...	.png
4e67e38e036454776db1fb3a	Residence	Residences	Residence	[{'id': '5032891291d4c4b30a586d68', 'name': 'A...	https://ss3.4sqi.net/img/categories_v2/buildin...	.png
4d4b7105d754a06378d81259	Shop & Service	Shops &	Shops	[{'id': '52f2ab2ebcbc5	https://ss3.4sqi.net/img/categories_v2/shops/d...	.png

id	name	pluralName	shortName	categories	icon.prefix	icon.suffix
		Services		7f1066b8b56', 'name': 'A...		
4d4b7105d754a06379d81259	Travel & Transport	Travel & Transport	Travel	[{'id': '4bf58dd8d48988d1ed931735', 'name': 'A...	https://ss3.4sqi.net/img/categories_v2/travel/...	.png

Data understanding

During this stage, the focus was given mainly on the data regarding venues and categories as the geospatial data are straight forward. So, regarding categories we had to understand the depth of the categorization tree and extract the necessary categories. In this case the necessary categories were under the 4th main category which is named **Food**. It has to be noted here that the documentation for the categories was very helpful and the relevant information can be found in this link https://developer.foursquare.com/docs/build-with-foursquare/categories/.

A function was defined to retrieve recursively the categories and create a Pandas DataFrame to store those needed for restaurants' categorization. In the same manner it was used as a test to retrieve the whole categorization tree. The DataFrame was then used to filter the venues separating anything that was under the food category from the others.

For New York, the number of venues collected was 9.751 in total 5.078 of them being under the food category. The following output shows the top 10 categories summing in 2.378 venues or 46,83% of the Total.

TOP 10 restaurant categories in New York

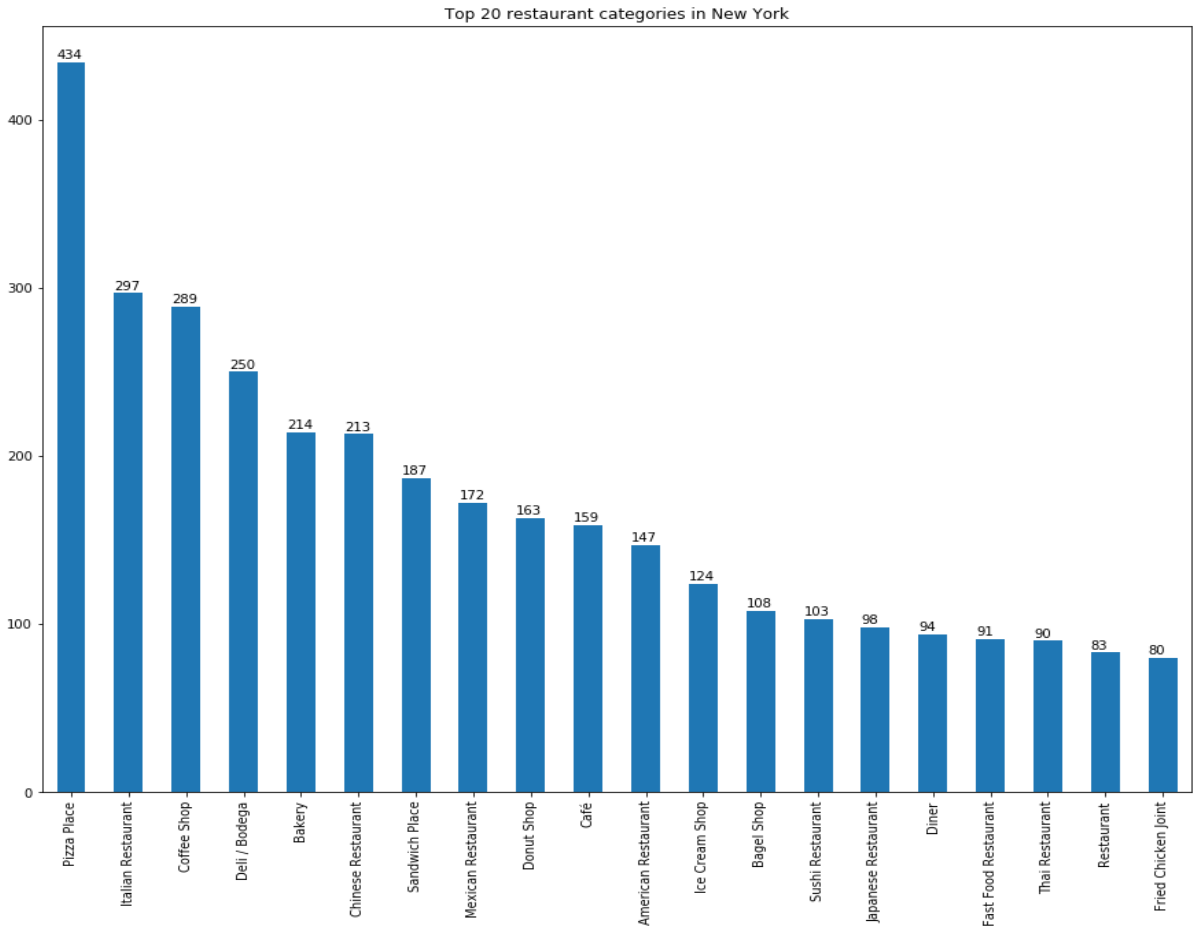
	Venue
Venue Category	
Pizza Place	434
Italian Restaurant	297
Coffee Shop	289
Deli / Bodega	250
Bakery	214
Chinese Restaurant	213
Sandwich Place	187
Mexican Restaurant	172
Donut Shop	163
Café	159
Count of TOP 10	**2378**
46.83% of total	

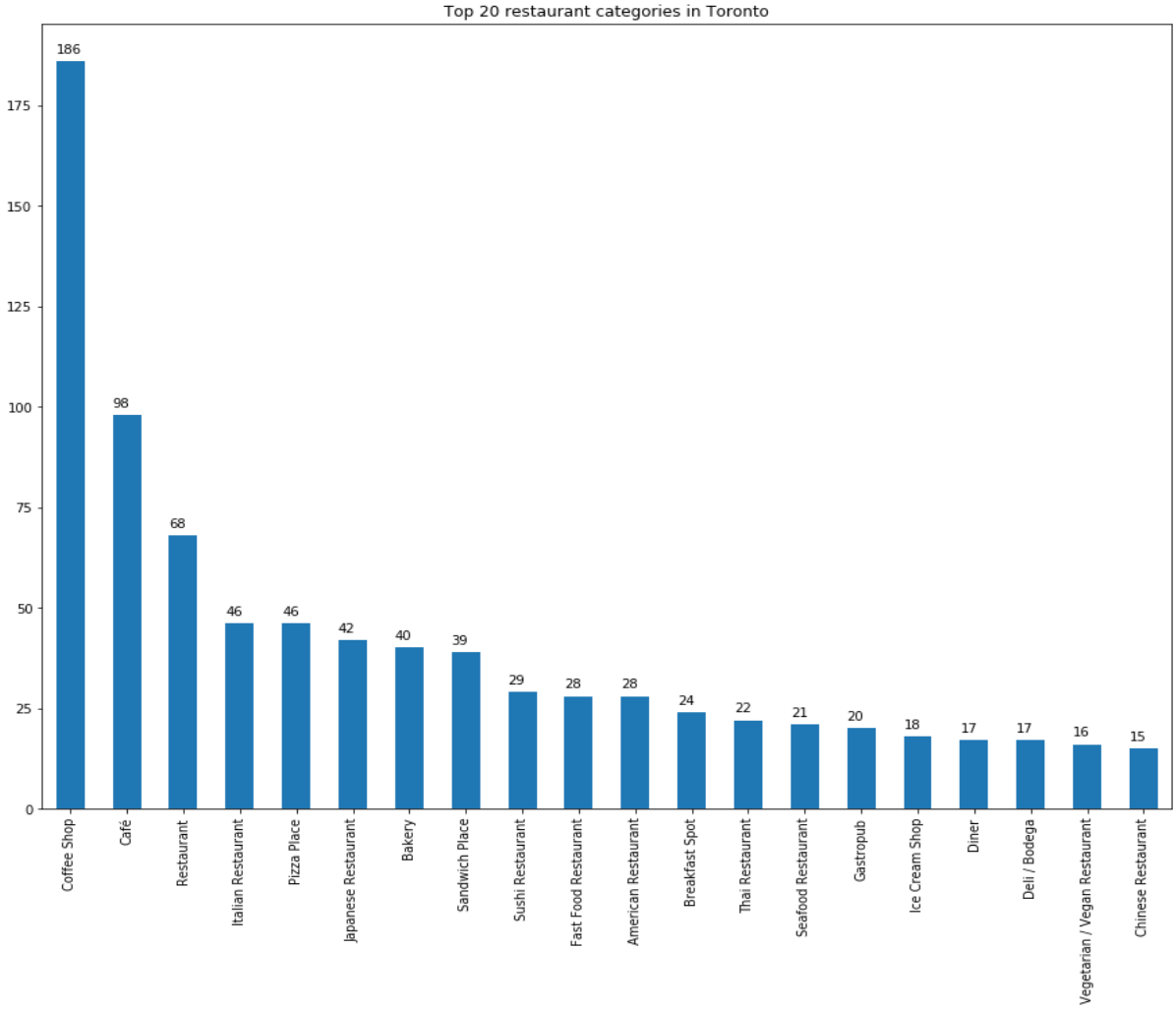
For Toronto, the number of venues collected was 2.110 in total 1.161 of them being under the food category. The following output shows the top 10 categories summing in 622 venues or 53,57% of the Total.

TOP 10 restaurant categories in Toronto

	Venue
Venue Category	
Coffee Shop	186
Café	98
Restaurant	68
Pizza Place	46
Italian Restaurant	46
Japanese Restaurant	42
Bakery	40
Sandwich Place	39
Sushi Restaurant	29
American Restaurant	28
Count of TOP 10	**622**
53.57% of total	

The following two bar charts show graphically the top 20 restaurant categories for New York and Toronto with the respective counters. The totals are for New York 3.396 or 66,88% or the total number of restaurants and for Toronto 820 or 70,63 of the total number of restaurants



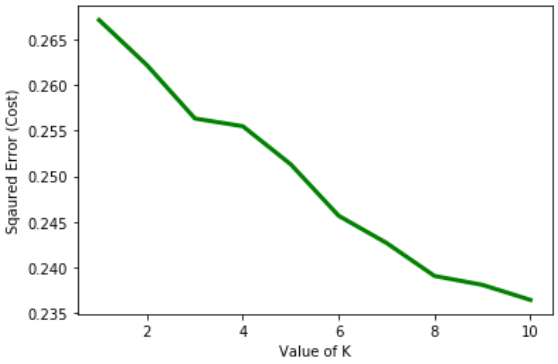


Modelling

In this case the most suitable data modelling was Clustering which is one of the most well-known unsupervised machine learning algorithms. It was used as it was of vital importance to group neighborhoods according to their similarities.

The preparation for clustering included the procedure to encode the data frames and then group the encoded data frames by Neighborhood. The final data structures (one for each city) were used to define the k for clustering and fit the k-means clustering algorithm. The resulting data frames can be seen in the Jupyter notebook that is included as the project's deliverable.

The next step was to define the number of clusters for each city by using the Elbow method and silhouette analysis. As it appears in the next graph for New York the elbow of the curve appears for k=3 and k=4.

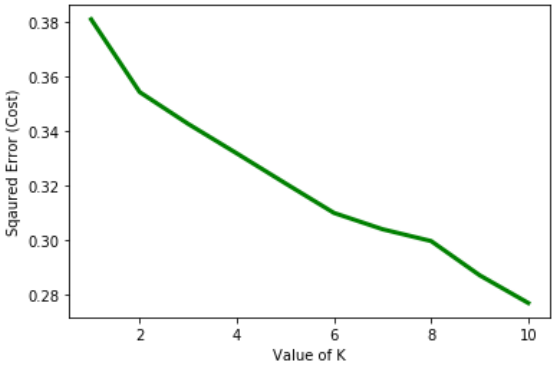


The silhouette analysis for New York is shown below:

For n_clusters = 2, silhouette score is 0.5236801422751508)
For n_clusters = 3, silhouette score is 0.09529805887971099)
For n_clusters = 4, silhouette score is 0.5265485356597408)
For n_clusters = 5, silhouette score is **0.030516352395696653**)
For n_clusters = 6, silhouette score is 0.033134899470598195)
For n_clusters = 7, silhouette score is 0.033259823939917985)
For n_clusters = 8, silhouette score is 0.03500437773899958)
For n_clusters = 9, silhouette score is 0.03699078931763086)
For n_clusters = 10, silhouette score is 0.03175896567437805

From the above given analysis, it seems that the value of 5 has the lowest score which means that we should group our venues in 5 clusters.

For Toronto there is not any apparent elbow with a rather small exception for k=2.



The silhouette analysis for New York is shown below:

For n_clusters = 2, silhouette score is 0.3018523882819826)
For n_clusters = 3, silhouette score is 0.2916324955210721)
For n_clusters = 4, silhouette score is 0.3003125718298839)
For n_clusters = 5, silhouette score is 0.30596654368247284)
For n_clusters = 6, silhouette score is 0.315103384915657)
For n_clusters = 7, silhouette score is 0.15138692180971608)
For n_clusters = 8, silhouette score is **0.10512880922234978**)
For n_clusters = 9, silhouette score is 0.14205795339101646)
For n_clusters = 10, silhouette score is 0.11994526058467776)

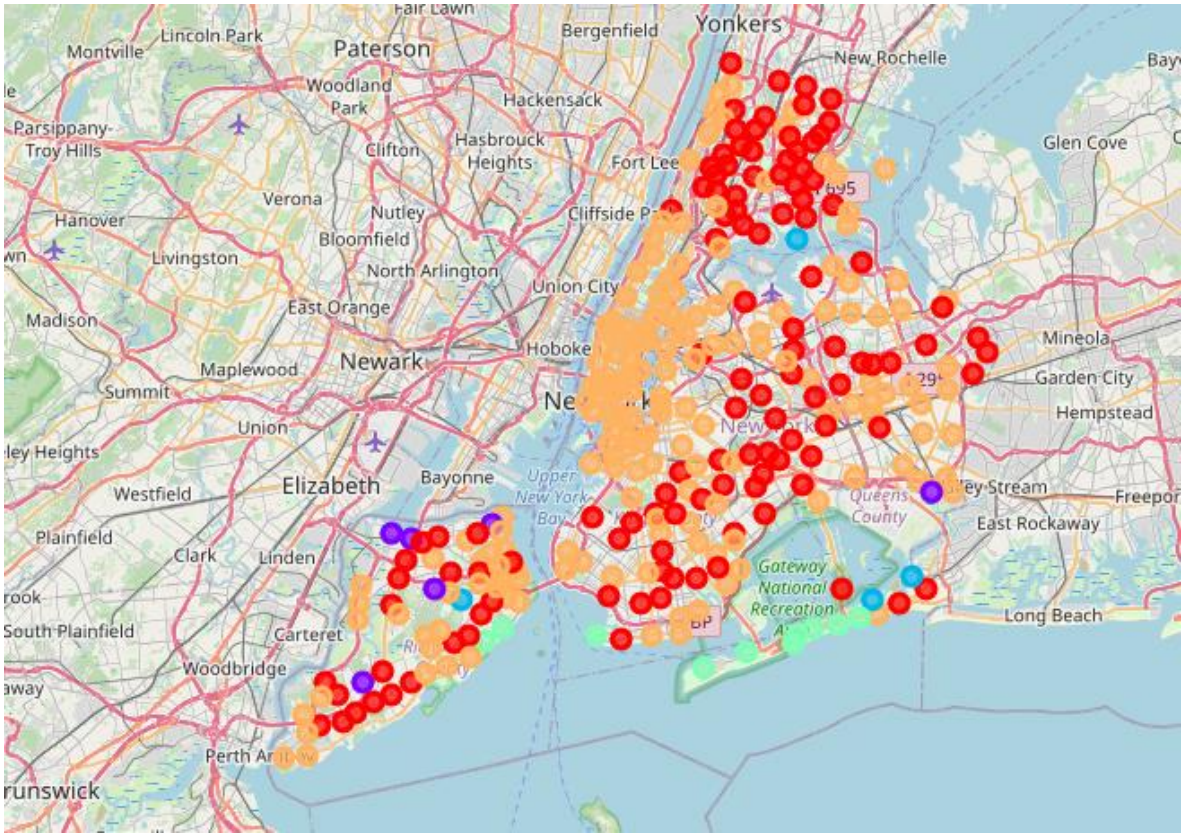
From the above given analysis, it seems that the value of 8 has the lowest score which means that we should group our venues in 8 clusters.

After further analysis which was based on trying several values for k (3, 5 and 8 for New York and 2, 5 and 8 for Toronto) it was found that 8 clusters were too many for both cities. Thus, the final number of clusters was decided to be 5 for both New York and Toronto. The number of neighborhoods per cluster are shown below:

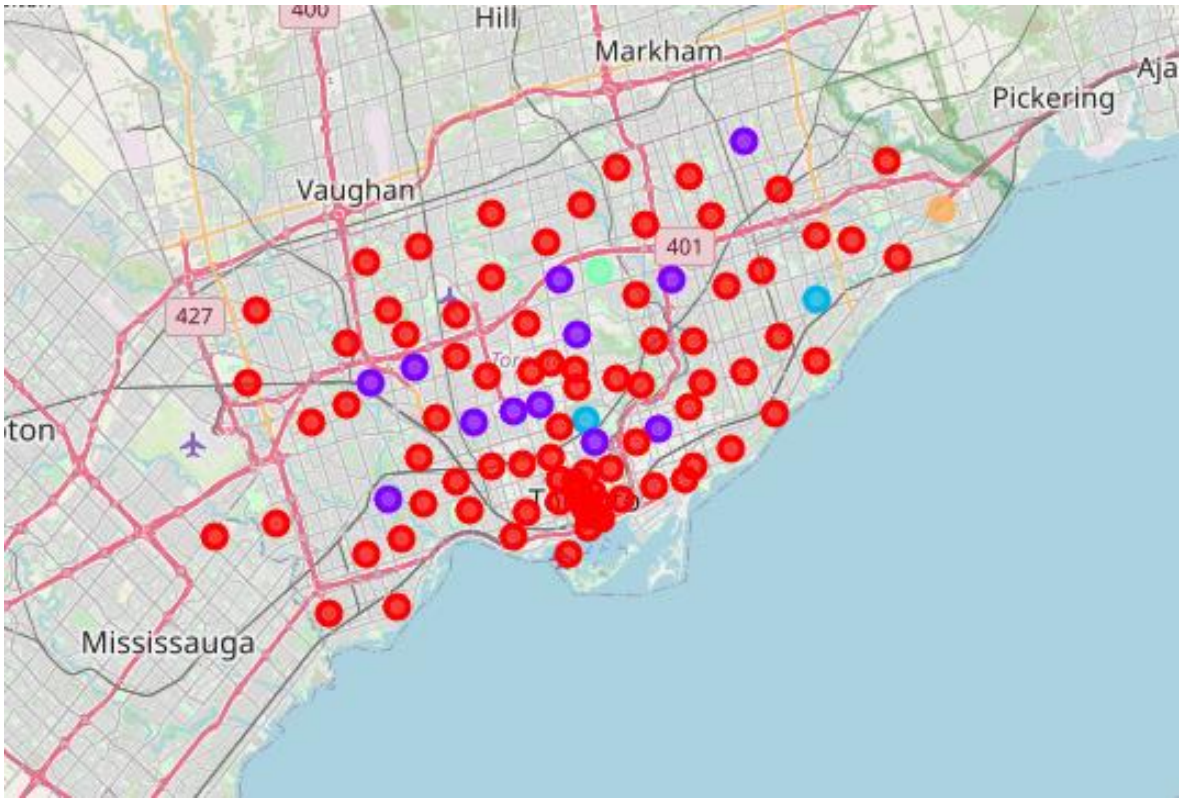
NY Cluster 1 117
NY Cluster 2 6
NY Cluster 3 4
NY Cluster 4 10
NY Cluster 5 167

TO cluster 1 82
TO cluster 2 12
TO cluster 3 2
TO cluster 4 1
TO cluster 5 1

New York's dominant clusters 1 and 5 appear in red and orange below.



Toronto's dominant clusters 1 and 2 appear in red and purple below



Clustering Analysis

It can be clearly seen that in both cities neighborhoods present wide similarities as both have two dominant clusters which are 1 and 5 for New York and 1 and 2 for Toronto so let us examine those clusters in a little bit more depth.

New York's clusters 1 and 5 present a wide variety in venues both for restaurants and other categories. We have included for those clusters the first 10 rows each row having the 10 most common categories on that cluster.

Cluster 1

Neighborhood	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
Wakefield	Pharmacy	Pizza Place	Gas Station	Sandwich Place	Dessert Shop	Donut Shop	Laundromat	Ice Cream Shop	Falafel Restaurant	Exhibit
Co-op City	Bus Station	Park	Grocery Store	Bagel Shop	Pizza Place	Discount Store	Restaurant	Pharmacy	Fried Chicken Joint	Ice Cream Shop
Eastchester	Caribbean Restaurant	Deli / Bodega	Diner	Bus Station	Donut Shop	Fast Food Restaurant	Chinese Restaurant	Metro Station	Bowling Alley	Seafood Restaurant

Neighborhood	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
Kingsbridge	Pizza Place	Bar	Bakery	Sandwich Place	Mexican Restaurant	Supermarket	Latin American Restaurant	Pub	Pharmacy	Donut Shop
Woodlawn	Pub	Deli / Bodega	Pizza Place	Playground	Grocery Store	Bar	Donut Shop	Trail	Train Station	Park
Norwood	Pizza Place	Pharmacy	Park	Chinese Restaurant	Deli / Bodega	Bank	Supermarket	Mexican Restaurant	Sandwich Place	Restaurant
Baychester	Donut Shop	Electronics Store	Supermarket	Mattress Store	Shopping Mall	Bank	Men's Store	Mexican Restaurant	Sandwich Place	Pet Store
Pelham Parkway	Italian Restaurant	Bus Station	Chinese Restaurant	Pizza Place	Eye Doctor	Frozen Yogurt Shop	Sandwich Place	Performing Arts Venue	Track	Donut Shop
Bedford Park	Diner	Mexican Restaurant	Chinese Restaurant	Spanish Restaurant	Sandwich Place	Deli / Bodega	Pizza Place	Bus Station	Supermarket	Grocery Store
University Heights	Pizza Place	Burger Joint	Pharmacy	Donut Shop	Sandwich Place	Latin American Restaurant	Bakery	History Museum	Fast Food Restaurant	Bank

Cluster 5

Neighborhood	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
Fieldston	Music Venue	River	Bus Station	Plaza	Farmers Market	English Restaurant	Entertainment Service	Ethiopian Restaurant	Event Service	Event Space
Riverdale	Park	Bus Station	Bank	Food Truck	Gym	Locksmith	Plaza	Home Service	Falafel Restaurant	Entertainment Service
Marble Hill	Sandwich Place	Coffee Shop	Gym	Yoga Studio	Video Game Store	Donut Shop	Supplement Shop	Bank	Pharmacy	Kids Store
Williamsbridge	Soup Place	Bar	Nightclub	Caribbean	Fast Food	Ethiopian	Event Service	Event Space	Exhibit	Eye Doctor

Neighborhood	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
				Restaurant	Restaurant	Restaurant				
City Island	Harbor / Marina	Seafood Restaurant	Park	Thrift / Vintage Store	Music Venue	Diner	Grocery Store	Liquor Store	Baseball Field	Bar
West Farms	Bus Station	Park	Outdoors & Recreation	Sandwich Place	Donut Shop	Liquor Store	Lounge	Diner	Playground	Bank
Port Morris	Furniture / Home Store	Brewery	Spanish Restaurant	Music Venue	Peruvian Restaurant	Donut Shop	Latin American Restaurant	Restaurant	Distillery	Storage Facility
Throgs Neck	Coffee Shop	Sports Bar	Asian Restaurant	Italian Restaurant	Bar	Pizza Place	American Restaurant	Deli / Bodega	Juice Bar	Factory
Country Club	Sandwich Place	Vegetarian / Vegan Restaurant	Playground	Farm	Empanada Restaurant	English Restaurant	Entertainment Service	Ethiopian Restaurant	Event Service	Event Space
Spuyten Duyvil	Park	Bank	Intersection	Thai Restaurant	Tennis Stadium	Pharmacy	Falafel Restaurant	English Restaurant	Entertainment Service	Ethiopian Restaurant

Same for Toronto clusters 1 and 2 present a wide variety in venues both for restaurants and other categories. We have included for those clusters the first 10 rows each row having the 10 most common categories on that cluster.

Cluster 1

Neighborhood	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
Malvern / Rouge	Fast Food Restaurant	Women's Store	Curling Ice	Eastern European Restaurant	Drugstore	Donut Shop	Doner Restaurant	Dog Run	Distribution Center	Discount Store
Guildwood / Morningside / West Hill	Electronics Store	Mexican Restaurant	Medical Center	Breakfast Spot	Intersection	Bank	Rental Car Location	Dim Sum Restaurant	Department Store	Dessert Shop

Neighborhood	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
Woburn	Coffee Shop	Korean Restaurant	Insurance Office	Deli / Bodega	Department Store	Dessert Shop	Dim Sum Restaurant	Diner	Discount Store	Distribution Center
Cedarbrae	Bakery	Bank	Athletics & Sports	Fried Chicken Joint	Hakka Restaurant	Caribbean Restaurant	Gas Station	Thai Restaurant	Dessert Shop	Deli / Bodega
Kennedy Park / Ionview / East Birchmount Park	Discount Store	Coffee Shop	Department Store	Hobby Shop	Train Station	Women's Store	Deli / Bodega	Dessert Shop	Dim Sum Restaurant	Diner
Golden Mile / Clairlea / Oakridge	Bus Line	Bakery	Bus Station	Metro Station	Soccer Field	Park	Intersection	Ice Cream Shop	Dim Sum Restaurant	Department Store
Cliffside / Cliffcrest / Scarborough Village West	American Restaurant	Motel	Women's Store	Deli / Bodega	Department Store	Dessert Shop	Dim Sum Restaurant	Diner	Discount Store	Curling Ice
Birch Cliff / Cliffside West	Skating Rink	College Stadium	General Entertainment	Café	Discount Store	Deli / Bodega	Department Store	Dessert Shop	Dim Sum Restaurant	Diner
Dorset Park / Wexford Heights / Scarborough To...	Indian Restaurant	Chinese Restaurant	Vietnamese Restaurant	Pet Store	Women's Store	Deli / Bodega	Department Store	Dessert Shop	Dim Sum Restaurant	Diner
Wexford / Maryvale	Smoke Shop	Middle Eastern Restaurant	Sandwich Place	Breakfast Spot	Bakery	Auto Garage	Shopping Mall	Department Store	Drugstore	Donut Shop

Cluster 2

Neighborhood	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
Milliken / Agincourt North / Steeles East / L'...	Park	Playground	Women's Store	Diner	Dance Studio	Deli / Bodega	Department Store	Dessert Shop	Dim Sum Restaurant	Distribution Center
York Mills West	Park	Bank	Convenience Store	Women's Store	Discount Store	Department Store	Dessert Shop	Dim Sum Restaurant	Diner	Distribution Center
Parkwoods	Park	Food & Drink Shop	Women's Store	Diner	Dance Studio	Deli / Bodega	Department Store	Dessert Shop	Dim Sum Restaurant	Discount Store
East Toronto	Park	Convenience Store	Women's Store	Discount Store	Deli / Bodega	Department Store	Dessert Shop	Dim Sum Restaurant	Diner	Distribution Center
Lawrence Park	Bus Line	Park	Swim School	Women's Store	Dim Sum Restaurant	Dance Studio	Deli / Bodega	Department Store	Dessert Shop	Discount Store
Rosedale	Park	Trail	Playground	Discount Store	Dance Studio	Deli / Bodega	Department Store	Dessert Shop	Dim Sum Restaurant	Diner
Forest Hill North & West	Jewelry Store	Park	Sushi Restaurant	Trail	Drugstore	Donut Shop	Doner Restaurant	Dog Run	Distribution Center	Eastern European Restaurant
Humewood-Cedarvale	Park	Trail	Hockey Arena	Field	Diner	Dance Studio	Deli / Bodega	Department Store	Dessert Shop	Dim Sum Restaurant
Caledonia-Fairbanks	Park	Women's Store	Pool	Diner	Dance Studio	Deli / Bodega	Department Store	Dessert Shop	Dim Sum Restaurant	Discount Store
North Park / Maple Leaf Park / Upwood Park	Construction & Landscaping	Park	Bakery	Women's Store	Department Store	Dessert Shop	Dim Sum Restaurant	Diner	Discount Store	Distribution Center

In the notebook that is included with the deliverables we have done some more analysis on the neighborhoods. These sections were not mentioned here intentionally as it would assume the presentations of more tables that would make this report unnecessarily big.

Results

After the analysis and clustering of the given data the desired questions about which restaurants should the client open and in which neighborhood are now ready to be answered.

The client would like to open restaurants that present high levels of popularity in both cities as its brands are strong and they would like to target the largest proportion of population possible. Thus, the restaurant types that match the requirements are Pizza Places, Italian restaurants and Coffee Shops. Pizza Place and Italian Restaurant are the two categories that were chosen.

Venue Category	NY	TO	TOTAL
Pizza Place	**434**	**46**	**480**
Coffee Shop	289	186	475
Italian Restaurant	**297**	**46**	**343**
Deli / Bodega	250	17	267
Café	159	98	257
Bakery	214	40	254
Chinese Restaurant	213	15	228
Sandwich Place	187	39	226
Mexican Restaurant	172	12	184
American Restaurant	147	28	175

As far as the second question is concerned, we found that the client has I wide variety of neighborhoods to choose from in order to pick the place for the new restaurants. As a matter of fact, they can choose many different neighborhoods to open the required number of places. For New York Bronx and Manhattan should be included and for Toronto Downtown is the strongest candidate.

Discussion

Now that we have reached the end of the analysis and presented some results, we should discuss about some of the research and analysis limitations and further steps.

The greatest limitation was that further data for venues required premium calls to Foursquare database that could not be performed without paying having in mind the multitube of venues. If someone has the resources to collect this data, the results of the analysis can be much more targeted.

Another issue that was found is that even though clustering algorithm is great in grouping similar objects (neighborhoods in our case) it cannot clearly distinguish the characteristics (venues in our case) that result in those similarities.

Finally, an extra set of data regarding the criminal rates and their occurrences would clearly help in selecting the best neighborhoods.

Conclusion

Overall, it as a great experience. We had the opportunity to work as data scientists on a project that gave us the chance to experiment on a large dataset and use a wide variety of tools for data gathering, formatting and analysis. Furthermore, we used mapping libraries and bar charts for visualisation. To sum up the findings of that project can be used as a basis for further analysis and / or in similar cases. Both New York and Toronto are great cities to expand business in a wide range of activities. Their multiculturality is apparent that is combined with the large numbers of population and visitors. I hope that you enjoyed reading this and that it will help someone in taking that some steps forward.