IBM Capstone Project

Which city should I expand my fruit shop to:
Toronto or London?

Introduction

Client:

Our client is a fruit shop owner who owns multiple shops in Soho, Greenwich Village and Little Italy in New York City

The Problem:

The client would like to expand their business to either Toronto or London.

The client knows that the Neighbourhoods their shops are located in in New York are the most profitable for his business

The client would like to know whether Toronto or London have more neighbourhoods similar to the neighbourhoods their shops are located in in New York.

Data Required

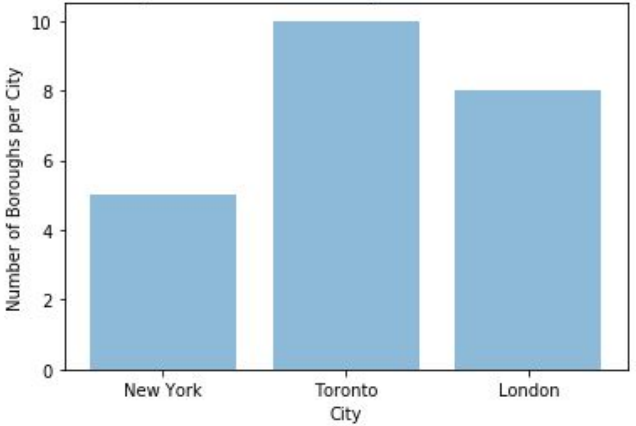
A: Identification of the neighbourhoods and locations of the neighbourhoods in New York, Toronto and London:

1. **New York**
    a. Post Codes: https://cocl.us/new_york_dataset
    b. Locations: https://cocl.us/new_york_dataset
2. **Toronto**
    a. Post Codes: https://en.wikipedia.org/w/index.php?title=List_of_postal_codes_of_Canada:_M&oldid=945633050.
    b. Locations: http://cocl.us/Geospatial_data
3. **London**
    a. Post Codes: https://en.wikipedia.org/wiki/List_of_postcode_areas_in_the_United_Kingdom (multiple links from this)
    b. Locations: https://www.freemaptools.com/download-uk-postcode-lat-lng.htm (will average accross each postcode within a region)

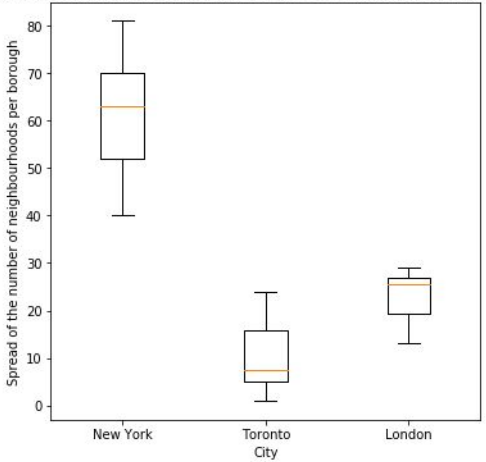
B: Data which enables characteristics of each neighbourhood to be identified. This will be obtained by finding the types of venues within each neighbourhood, from foursquare.com..

Initial analysis on Boroughs & Neighbours

A bar chart showing the number of Boroughs in New York, Toronto and London



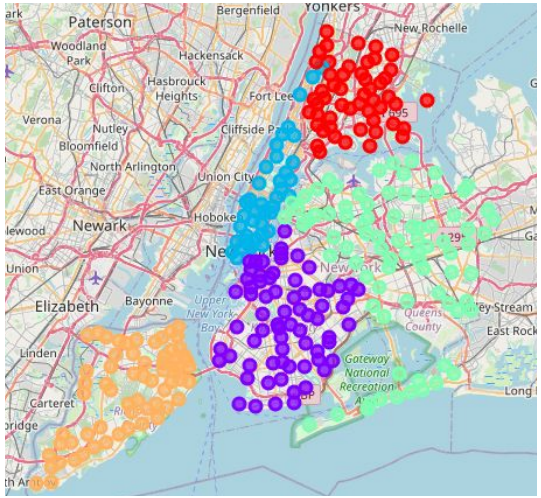
Box Plot showing the number of Neighbourhoods per Borough in New York, Toronto and London

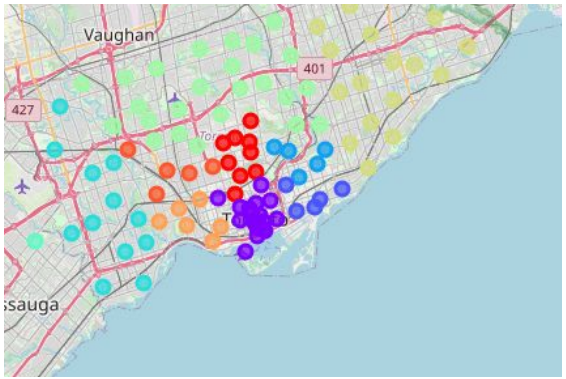


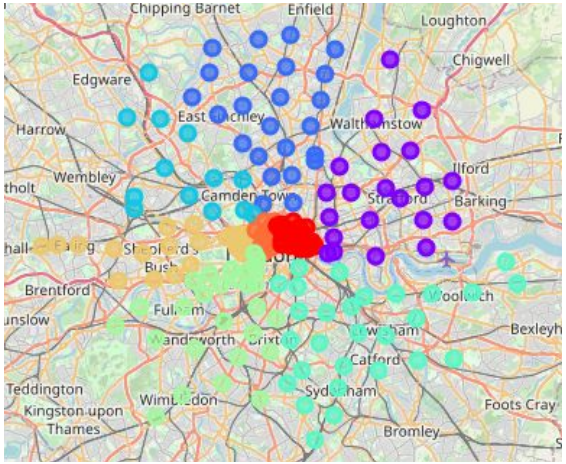
Initial analysis shows that:

1. New York has much fewer Boroughs than Toronto and London.
2. Each borough in New York has a higher number of Neighbourhoods within them than Toronto and London.

Mapping the Neighbourhoods in New York, Toronto and London





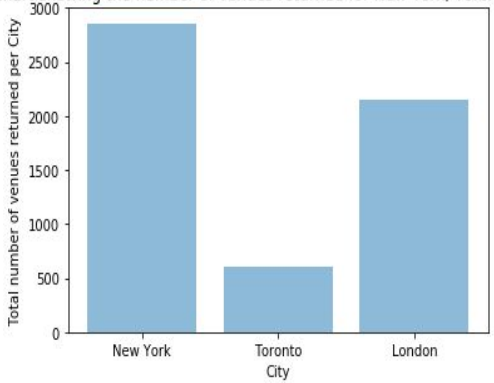


Maps of New York, Toronto and London respectively. The colours distinguish different boroughs within each of the cities.

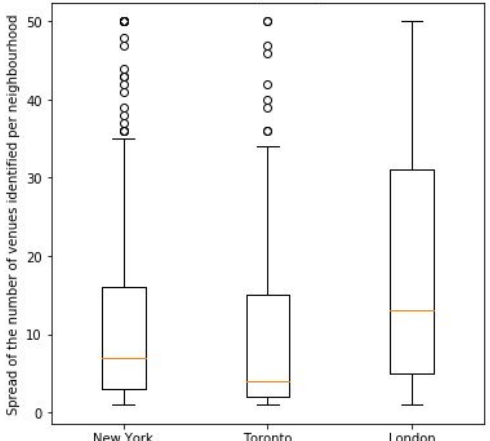
Venues data obtained from FourSquare Radius = 250, Limit = 50

Foursquare was queried to identify venues within a 250 meter radius of each neighbourhood, with the total limit set to 50. These plots show that the most venues were obtained for New York City and the fewest for Toronto. The plots also show that the median number of venues returned per neighbourhood for New York and Toronto was less than 10. This resulted in high distortion values when clustering, and I concluded additional data would be required.

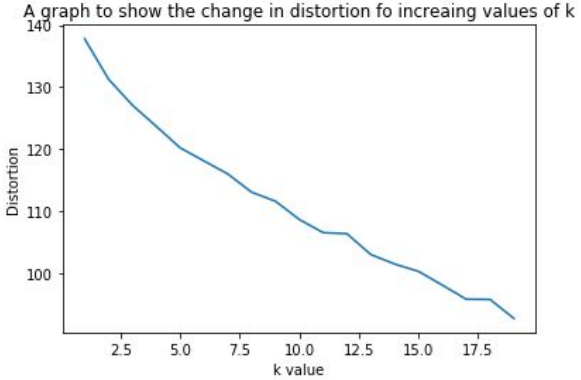
A bar chart showing the number of venues returned for New York, Toronto and London



Box Plot showing the number of venues identified per Neighbourhood in New York, Toronto and London



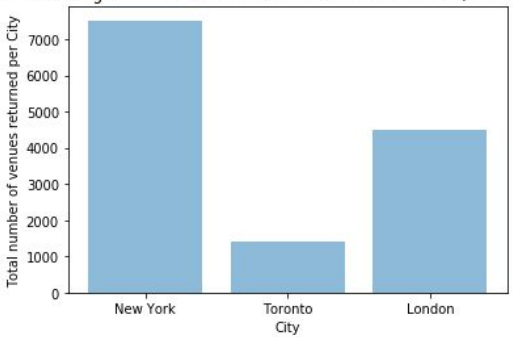
A graph to show the change in distortion fo increaing values of k



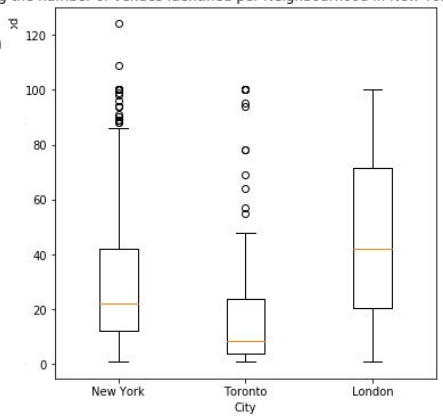
Venues data obtained from FourSquare Radius = 500, Limit = 100

Foursquare was queried again to identify venues within a 500 meter radius of each neighbourhood with the total limit set to 100. These plots show that again alot more venues were identified in New York than in Toronto. The plots also show that the median number of venues did not massively increase, do to multiple venues having no venues identified. This again resulted in high distortion when clustering, and I decided to only include neighbourhoods where more that 15 venues had been identified.

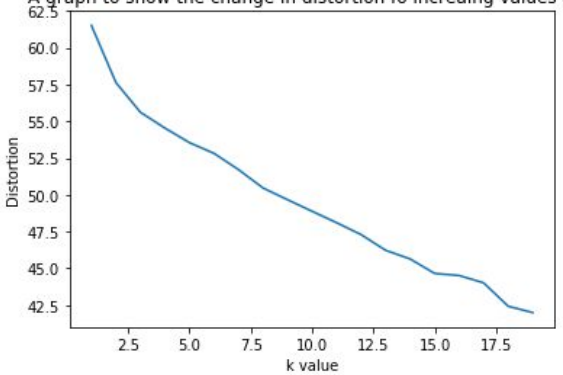
A bar chart showing the number of venues returned for New York, Toronto and London



Box Plot showing the number of venues identified per Neighbourhood in New York, Toronto and London



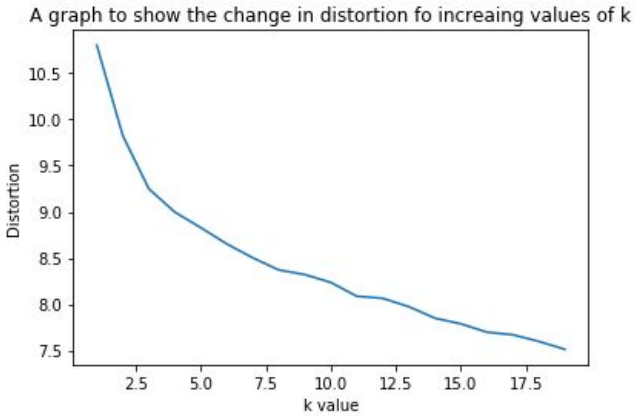
A graph to show the change in distortion fo increaing values of k



Finding the Optimum K Value and conducting K-means clustering

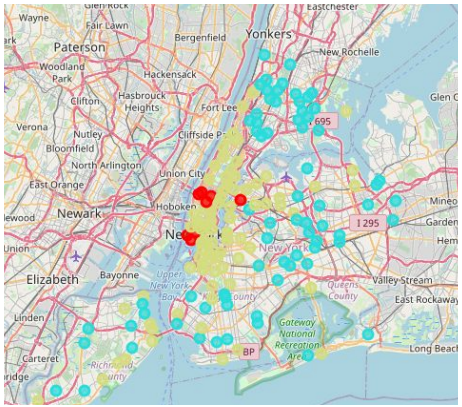
The data therefore used was obtained from Foursqure with a radius of 500 meters and a limit set to 100 venues for each radius. I only included neighbourhoods which had more than 15 venues identified in them, to prevent bias within the dataset.

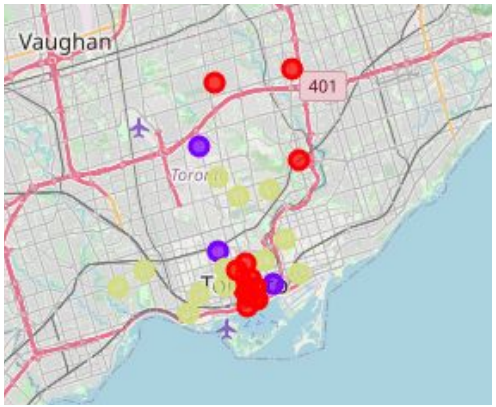
I conducted k-means clustering to find the optimum k value, and found the “elbow point” was k = 4 in order to find the predicted cluster for each neighbourhood.

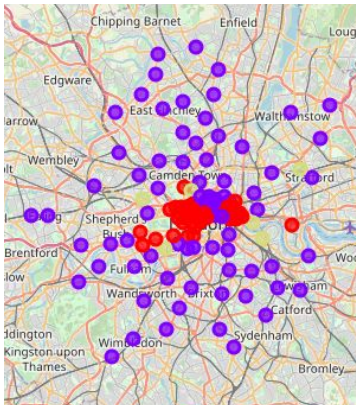


Cluster Labels	Neighbourhood
2	Allerton
3	Arrochar
3	Astoria

Mapping the Neighbourhood clusters in New York, Toronto and London







Maps of New York, Toronto and London respectively. The colours distinguish different clusters each neighbourhood is assigned to within each of the cities.

Discussion

The results show:

1. Each of the three cities have a similar central neighbourhoods (colour red in the maps) this is the location of each cities financial district.
2. As you move away from the centre of the cities New York and London are very differnt. New York has more restaurant's/bars (coloured green) while London has more recreational neighbourhoods containing gyms/parks (coloured purple). Toronto is more of a hybrid of New York and London.

Conclusion

The customer has fruit shops within Greenwich Village, Little Italy and Soho in New York city. These neighbourhoods all fall within Cluster number 3.

London only has 6 neighbourhoods which fall within cluster 3, while Toronto has 12 neighbourhoods which fall within cluster 3.

We therefore recommend the customer expands their business to Toronto as there are more neighbours similar to those that have been successful in New York.

City	Cluster Labels	
London	0	43
	1	85
	3	6
New York	0	8
	2	71
	3	81
Toronto	0	15
	1	3
	3	12