IBM Capstone Project - Where shall I expand my fruit shop to: Toronto or London?

Introduction - The Business Problem

One of our clients has multiple success fruit shops within New York and is seeking to expand their business Internationally. The client would like to either expand their business to Toronto or London, and open multiple shops in the new city.

The client knows that the neighbourhoods their shops are located in in New York are the most likely to return high profits. The client would like to know whether to Toronto or London has more neighbourhoods which are similar to the neighbourhoods their shops are located in in New York. This will inform their decision on where to open the shops.

The aim of this project is therefore to identify whether London or Toronto have more neighbourhoods which are similar to the neighbourhoods that the clients shops are currently located in in New York.

Planned Data Usage

In order to conduct this analysis, I will need access neighbourhood datasets for New York, Toronto and London. I will also need acces to the longitude and latitude for each of these neighbourhoods. This will be obtained using the following websites:

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

I will also need access to data which in some way helps to categorise the types of neighbourhoods. This will be achieved by using https://foursquare.com/ to identify the most common venue types within each neighbourhood.

I would have liked to conduct analysis on the types of people that live in each neighbourhood (e.g. shopping habits - Amazon, viewing habits - Netflix), but this type of data is not freely available.

In reality, there are multiple other factors that would need to be taken into account before making this decision, such as availablity of stock and costs. However, in this report I will only focus on identifying which city would be best based on the number of similar neighbourhoods.

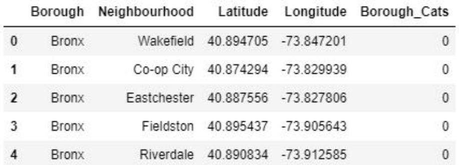
Methodology

The methodology for this notebook has been broken up into 15 steps. For this analysis, I will use K-means clustering to identify similar neighbourhoods in three different cities (New York, Toronto and London).

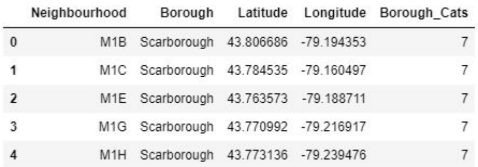
- **Step 1** - Import all of the libraries required for this notebook
- **Step 2** - Create a Dataframe containing the Neighbourhoods and their locations for New York
- **Step 3** - Create a Dataframe containing the Neighbourhoods and their locations for Toronto
- **Step 4** - Create a Dataframe containing the Neighbourhoods and their locations for London
- **Step 5** - Initial Analysis conducted on New York, Toronto and London neighbourhoods
- **Step 6** - Create a map of New York, Toronto and London showing all of their neighbourhoods, and display the boroughs in different colours.
- **Step 7** - Define a function to query Foursquare for each city and set up Foursquare credentials
- **Step 8** - Get Venues close to each neighbourhood in New York and put the results into a dataframe
- **Step 9** - Get Venues close to each neighbourhood in Toronto and put the results into a dataframe
- **Step 10** - Get Venues close to each neighbourhood in London and put the results into a dataframe
- **Step 11** - Conduct initial Analysis on venues returned for New York, Toronto and London
- **Step 12** - Prepare the data for k-means clustering
- **Step 13** - Identify the optimum K value for K means clustering
- **Step 14** - Conduct K means clustering with k = 4
- **Step 15** - Create a map of New York, Toronto and London showing all the cluster associated with each neighbourhood in different colours

Steps 1 - 4: Obtain the data for New York, Toronto and London

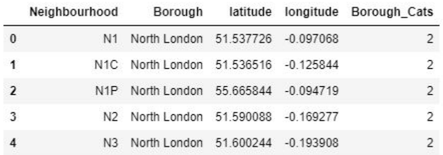
Using a variety of datasources available on the interenet (see Planned Data Usage) data for each of the neighbourhoods in New York, Toronto and London was obtained. The coordinates for the location which was approximately central to each of these neighbourhoods was also obtained. This resulted in three dataframes, one for each of the cities as shown below.



	Borough	Neighbourhood	Latitude	Longitude	Borough_Cats
0	Bronx	Wakefield	40.894705	-73.847201	0
1	Bronx	Co-op City	40.874294	-73.829939	0
2	Bronx	Eastchester	40.887556	-73.827806	0
3	Bronx	Fieldston	40.895437	-73.905643	0
4	Bronx	Riverdale	40.890834	-73.912585	0



	Neighbourhood	Borough	Latitude	Longitude	Borough_Cats
0	M1B	Scarborough	43.806686	-79.194353	7
1	M1C	Scarborough	43.784535	-79.160497	7
2	M1E	Scarborough	43.763573	-79.188711	7
3	M1G	Scarborough	43.770992	-79.216917	7
4	M1H	Scarborough	43.773136	-79.239476	7

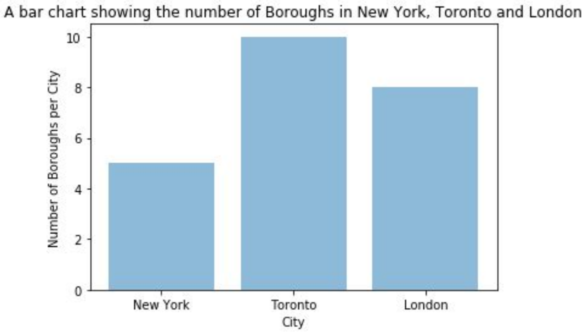


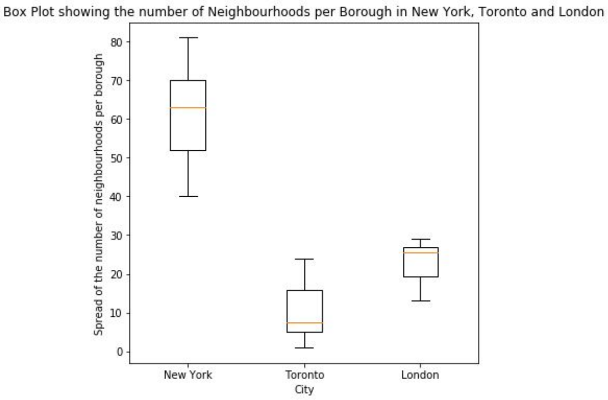
	Neighbourhood	Borough	latitude	longitude	Borough_Cats
0	N1	North London	51.537726	-0.097068	2
1	N1C	North London	51.536516	-0.125844	2
2	N1P	North London	55.665844	-0.094719	2
3	N2	North London	51.590088	-0.169277	2
4	N3	North London	51.600244	-0.193908	2

Step 5: Initial Analysis on the Neighbourhoods in New York, Toronto and London

Initial analysis was conducted to identify how many boroughs and neighbourhoods were located in New York, Toronto and London. THis analysis shows that:

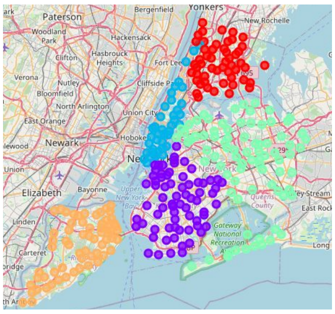
1. New York has much fewer Boroughs than Toronto and London.
2. Each borough in New York has a higher number of Neighbourhoods within them than Toronto and London.

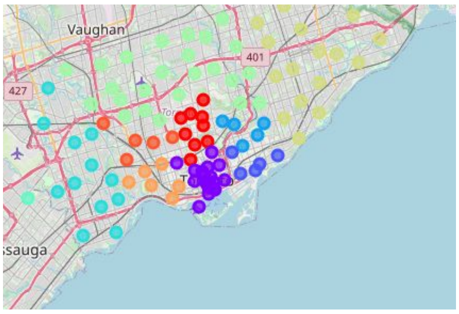


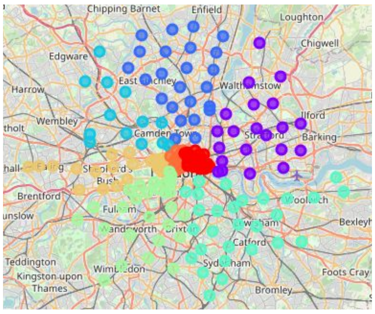


Step 6: Creating maps to show the neighbourhoods in New York, Toronto and London

Maps were also created to display the distribution of neighbourhoods within each of the cities. These maps are coloured to display the different boroughs each of the neighbourhoods belongs to.

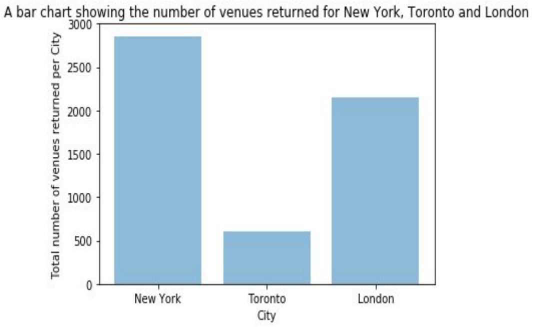


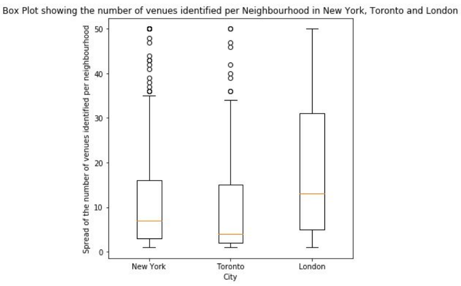


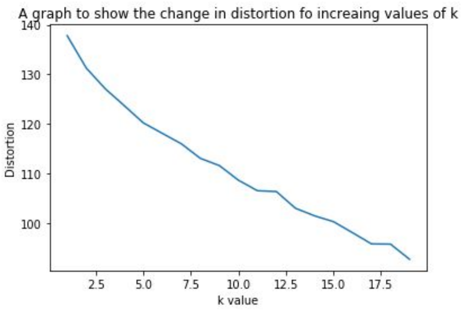


Step 7-11: Obtaining & Initial analysis on venus data for New York, Toronto and London

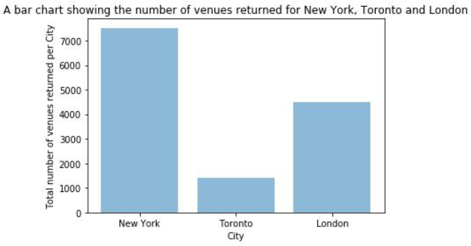
Foursquare was queried to identify venues within a 250 meter radius of each neighbourhood, with the total limit set to 50. These plots show that the most venues were obtained for New York City and the fewest for Toronto. The plots also show that the median number of venues returned per neighbourhood 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.

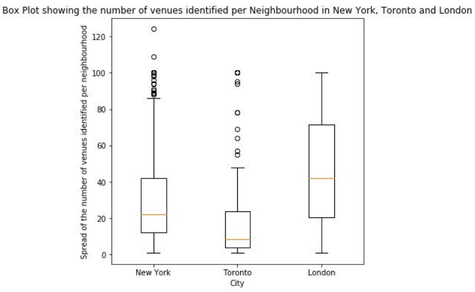


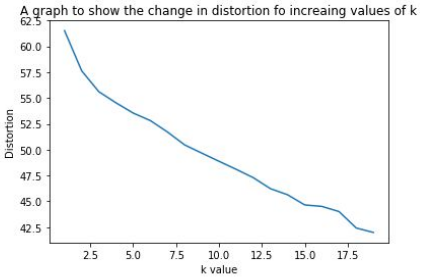




Foursquare was queried again to identify venues within a 500 meter radius of each neighbourhood with the total limit set to 100. These plots show that again alot more venues were identified in New York than in Toronto. The plots also show that the median number of venues did not massively increase, do 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.



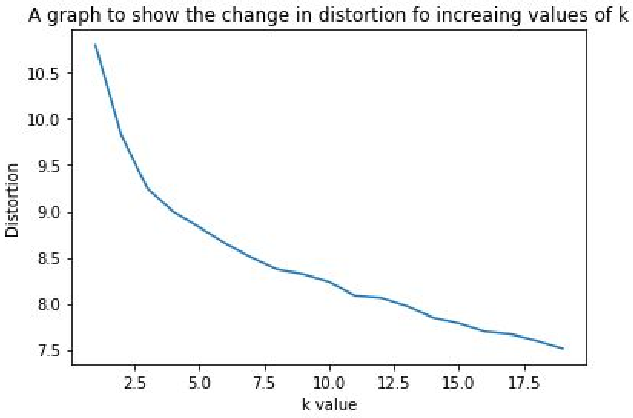




Step 12-13: Identifying the optimum value for k for k means clustering.

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

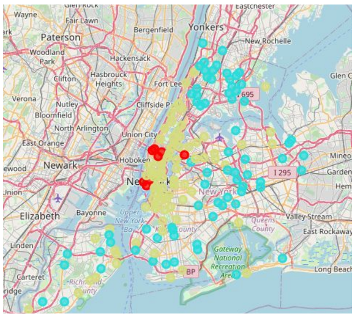
I conducted k-means clustering through a variety of k values, to find the optimum k value, and found the “elbow point” in the distortion was when k = 4. Therefore 4 values were used to conduct this analysis.

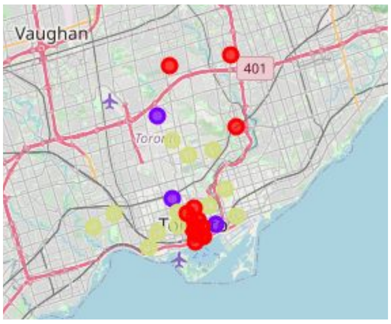


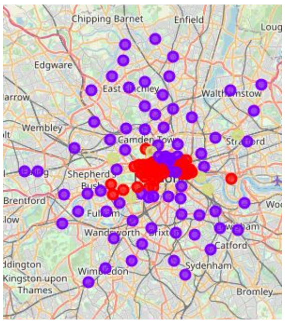
Results and Discussion

Four distinct neighbourhood types were found in the cities New York, Toronto and London. These were plotted onto maps to show the distribution of these neighbourhood types geographically. The main 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.