Capstone Project - The Battle of Neighbourhoods (Week 1)

By Christopher Luu

Manhattan vs Toronto: Where should I travel to?

The Problem:

If you have ever wondered whether to travel to Manhattan or Toronto (when COVID is over and provided travel is easily accessible as it used to) then this analysis is for you. This project aims to analyse these two cities using FourSquare API and K-means clustering to help identify which of the two locations may be more interesting to travel to, subject to the readers(stakeholders) own preferences.

Background

Given the wide variety of venues available in the vibrant cities, it may be difficult for someone with travel plans to see if they wish to travel to Canada or America. These two cities are listed as a part of the top 25 cities to travel in the world [1], with Toronto being multi-cultural, it presents a diverse range of venues that

travellers could be interested in checking out. Manhattan on the other hand is a borough in New York, a place that is shown commonly in films and TV. Thus, for those wishing to experience Manhattan as they do in movies and films, this could be the place for them to go. As America and Canada are close together, allowing for ease of travel then there is potential to travel to both places as well, rather than being subject to only one location (if you have the money for it).

Data Used

Manhattan data available from: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json

Toronto data available from: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Geospatial data available from: https://cocl.us/Geospatial_data

The above data are used together with Foursquare location data to cluster venues in Manhattan and Toronto, and are presented in a format that allows for ease of access.

These two data sources are also used within the Coursera IBM Data Science Certification capstone projects, and this project aims to extend analysis based on these data. The ease of access comes from the format of the data which is structured, saving much time that is typically used in the data wrangling section of beginning projects.

Image 1 and 2 below show snippets from the Jupyter Notebook of the data used.

	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

Figure 1 Manhattan Data - subset from New York Data

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

Figure 2 Data wrangled Toronto data which has also been cleaned.

As can be seen in figures 1 and 2, both the data wrangled data of Manhattan and Toronto are data frames with common column headers ‘Borough’, ’Neighbourhood’, ’Latitude’ and ‘Longitude’ allowing for ease of analysis with Foursquare API calls.

Methodology

To perform the analysis, K-means clustering was performed as it would be a quick and easy tool to address the problem for the stakeholders. Furthermore, in combination with Foursquare location data on venues, clustering provided the means to give insight on the two cities without delving too deep into the data. The folium package was used to produce visual maps of the data pre- and post-clustering as this package is highly versatile in doing so. Specifically with Foursquare API, only the top 10 venues within a 500meter radius of each latitude and longitude location was used due to the limitations of Foursquare and their daily hard limit on calls. As to not exceed this, a reduced volume of calls would be required, and as such, was performed.

Specific to Toronto, I aimed to produce a dataframe that would contain the same or as close to, shape as the Manhattan data which required little to no data cleaning. From the data wrangling, the final data frames of Manhattan displayed 1 Borough and 40 Neighbourhoods whilst the combined Toronto data contained 1 Borough and 39 Neighbourhoods. The similar data size provided a more standardised approach towards comparing the two clusters at the end. This assessment of the Toronto data can be seen in figure 3 below.

```
In [138]: test = merged_table['Borough'].unique()
          test

Out[138]: array(['North York', 'Downtown Toronto', 'Etobicoke', 'Scarborough',
                 'East York', 'York', 'East Toronto', 'West Toronto',
                 'Central Toronto', 'Mississauga'], dtype=object)

In [139]: #identifying best borough to use
          for i in test:
              a = merged_table[merged_table['Borough']==i]
              print('The dataframe {} has {} boroughs and {} neighborhoods.'.format(i,l

          The dataframe North York has 1 boroughs and 24 neighborhoods.
          The dataframe Downtown Toronto has 1 boroughs and 19 neighborhoods.
          The dataframe Etobicoke has 1 boroughs and 12 neighborhoods.
          The dataframe Scarborough has 1 boroughs and 17 neighborhoods.
          The dataframe East York has 1 boroughs and 5 neighborhoods.
          The dataframe York has 1 boroughs and 5 neighborhoods.
          The dataframe East Toronto has 1 boroughs and 5 neighborhoods.
          The dataframe West Toronto has 1 boroughs and 6 neighborhoods.
          The dataframe Central Toronto has 1 boroughs and 9 neighborhoods.
          The dataframe Mississauga has 1 boroughs and 1 neighborhoods.
```

Figure 3 Assessing the Toronto data to see how to group them.

The combination of Downtown Toronto, East Toronto, West Toronto and Central Toronto formed the new dataframe that would be assessed, and was given a new borough name of "Toronto Main" which led to the 1 Borough, 39 Neighbourhood data set to be processed and clustered.

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M5A	Toronto Main	Regent Park, Harbourfront	43.654260	-79.360636
1	M7A	Toronto Main	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
2	M5B	Toronto Main	Garden District, Ryerson	43.657162	-79.378937
3	M5C	Toronto Main	St. James Town	43.651494	-79.375418
4	M5E	Toronto Main	Berczy Park	43.644771	-79.373306
5	M5G	Toronto Main	Central Bay Street	43.657952	-79.387383

Figure 4 Borough set to Toronto Main

Results

Manhattan Clustering

Figure 5 Pre-clustered Manhattan

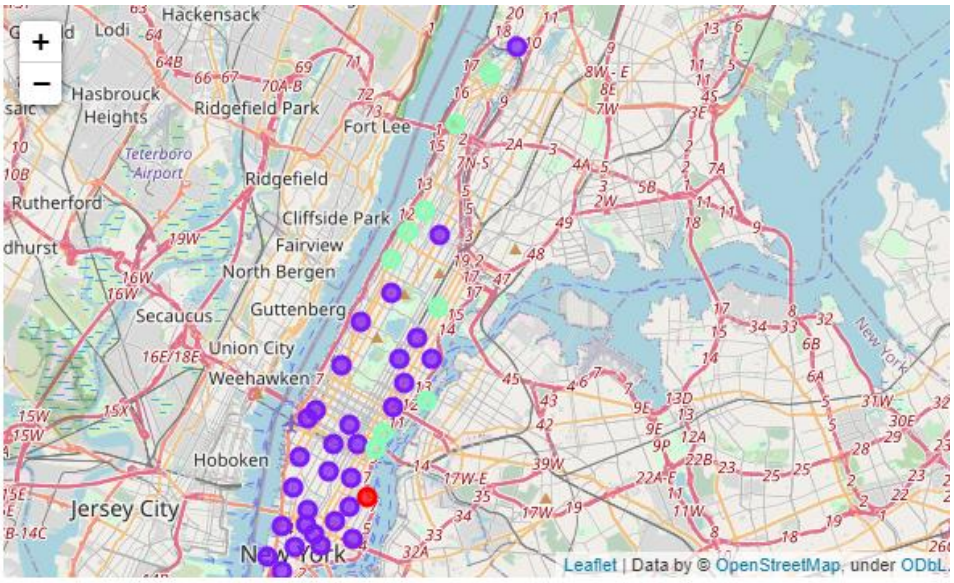

Figure 6 Post-Clustered Manhattan, with 3 clusters.

Manhattan venues by cluster

```
Park     1
Name: 1th Most Common Venue,
```

Figure 7 Cluster 0. Most common venue, only 1 data point.

```
Italian Restaurant      8
Coffee Shop             6
Bar                     3
Café                    2
Gym / Fitness Center    1
Hotel                   1
Theater                 1
Gym                     1
American Restaurant     1
Park                    1
Clothing Store          1
Art Gallery             1
Korean Restaurant       1
Chinese Restaurant      1
Plaza                   1
Name: 1th Most Common Venue,
```

Figure 8 Cluster 1. Most clusters are of this type for Manhattan.

```
Park                 3
Café                 2
Coffee Shop          2
Pizza Place          1
Mexican Restaurant   1
Name: 1th Most Common Venue,
```

Figure 9 Third and final cluster for Manhattan.

Toronto

Figure 10 Pre-clustered Toronto

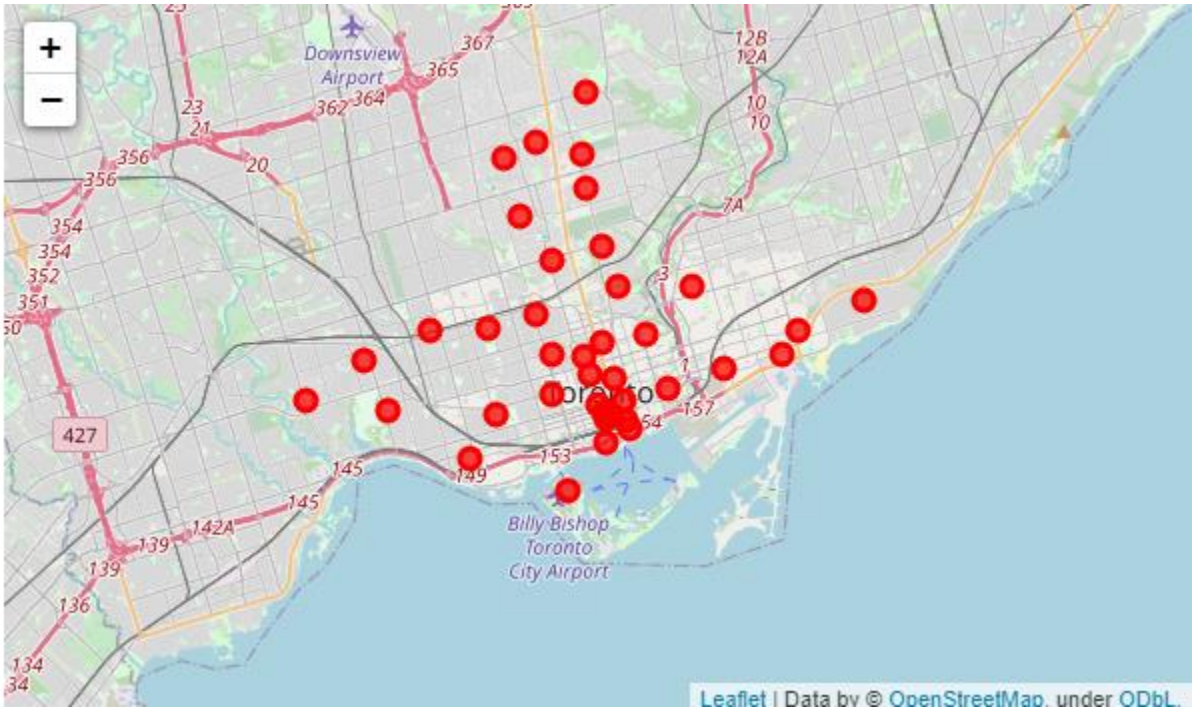

Figure 11 Post-Clustered Toronto, with 1 clusters.

```
Venue CategorySpa    19
Name: 1st Most Common Venue,
```

Figure 12 Most common venues for the only cluster in Toronto

Discussion

From the 2 cluster maps generated, it can be seen that only 1 cluster is similar between Toronto and Manhattan, which is cluster 0 (coloured red on the map). However, it should be noted that Toronto only displays 1 cluster which indicates that its venue choice is not as diversified as well as Manhattan which produced 3 clusters. As both cities were analysed using the same methods and parameters (i.e. k = 3 for k means clustering and almost same number of neighbourhoods for both cities) then what is expected would have been a fair comparison. The results suggest that Manhattan presents higher variety in terms of venues when compared to Toronto which is mostly the same venues that are rated highly consistently.

Analysing the results of Manhattan in cluster 0, the single data point shows the result of park which does not provide a lot of information. In cluster 1 we can see that the most common venue are Italian restaurants and coffee shops coming in second. Similarly we see that in cluster 2, the most common venue is parks which aligns with cluster 0, possibly indicating that the clusters could have been increased to classify parks into their own clusters.

For Toronto, the venue of Spa is ranked number 1 across all neighbourhoods with the rest which can be seen in the table above.

Comparing the cluster tables between Manhattan and Toronto it can clearly be seen that Manhattan is more food centric whilst Toronto is more exploration and tourism based as their History museum is ranked 4th most common across all neighbourhoods. This suggests that travellers more focused on food would be more inclined to choose Manhattan as their destination of choice, whilst those aiming for exploration and learning about the culture, tradition and history of the city they are in, would choose Toronto.

Conclusion

Using KMeans clustering on a subset of data representing Manhattan and Toronto, this project provides a finding that suggests that travellers with an aim to eat food would be inclined to travel to Manhattan as presented by the cluster tables for Manhattan presenting high re-occurence of food locations. Alternatively, travellers interested in exploration, relaxation and learning about the history of a city would be more inclined to choosing Toronto as presented by the cluster it presented with Spa ranking 1st and History Mueseum ranking fourth.

[1] https://www.afar.com/magazine/best-cities-in-the-world