Manhattan & Toronto Community Differences

Applied Data Science Capstone

Ximing Wang
30/09/2018

Introduction

The behaviour within different neighbourhood normally varies. This is largely due to many aspects: geographical density, living standard of residents, age of residents, local weather etc. The aim of this report is to compare the community within Manhattan and Toronto, seeking to find the standardised living patterns within the two districts and then trying to rationalize the results based on empirical and historical evidences.

The report will be split into six sections. Next section details the data to be used for the analysis. Including the format of the data, the origin of data, and the standardization of data. In methodology, any exploratory data analysis and machine learning used will be discussed. All result outputs will be stated and analysed in result section and then discussed based on historical and empirical evidence in discussion section. The last section is conclusion, which will draw a short conclusion of the whole experiment and findings.

Data Source

In order to find out the community composition of the neighbourhoods within the two districts. The following data will be needed:

1. The community information of the two districts
2. Community neighbourhood information
3. Categorised venues around neighbourhood

Once the data are gathered, they will be pre-processed into standardised format in order for later use. The raw data from this report mainly come from four resources. The Manhattan data from IBM online cloud database, Toronto data extracted from Wikipedia page and geographical data from folium library and venue data from Foursquare. The first two data are for mapping with the latter two data. The first two data contain information for the two districts of their borough, neighbourhood and postal code, with which could be mapped with geographical attribute from folium library. The folium library will also be used later for neighbourhood clustering visualization purpose.

The data will be stored as Python dataframe format. Once the neighbourhood and geographical data has been stored as dataframe. They will be mapped to venue data gathered from the online source of Foursquare. These data contain the information about the venues within each specific geographical location within a pre-determined radius. After merging the borough data with venue data, this will generate the main data to be worked on.

Methodology

In this section, the process of the whole experiment and method of analysis will be detailed. The first step would be to gather data from the resource mentioned. The data gathered from IBM is already formatted in clean dataframe format, but Toronto data extracted from html was not. Multiple operations within Python Pandas library could be applied to the dataset to transform all raw data into the format as shown in Figure 1 and this will be the standardised format for the analysis purpose.

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M4E	East Toronto	The Beaches	43.676357	-79.293031
1	M4K	East Toronto	The Danforth West, Riverdale	43.679557	-79.352188
2	M4L	East Toronto	The Beaches West, India Bazaar	43.668999	-79.315572
3	M4M	East Toronto	Studio District	43.659526	-79.340923
4	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790

Figure 1: Data format example for Toronto

Once the data is structured, geographical information about the neighbourhood will be mapped to the cloud database of Foursquare. The mapped information will be a dataframe containing relationship between each neighbourhood and the venues around the specific neighbourhood. Here, the radius was set to 500 to include venues within this distance.

The next step would be to count the occurrence of different venues within the whole area and rank them in descending order. There are around 300 different venue categories, each occurrence will be counted as 1 is the venue is included in the pre-defined radius of the neighbourhood. The count is added as the total frequency of the venue within the district.

The final step would be using machine learning to find the relationship between clusters and venues, in order to determine whether some venues are specifically popular within clusters of certain property. The clustering method used here was K-mean, with K initialized to 5. The clusters will be generated and visualized on the district map as shown in Figure 2

Figure 2: Cluster Visualization

Results

The following are results for the top 10 most frequent venues within Manhattan (Left) and Toronto (Right). The frequency column counts the occurrence of each venue within the district, as a sum of all neighbourhood, as shown in Figure 3.

	Frequency
Italian Restaurant	141
Coffee Shop	127
American Restaurant	80
Bakery	77
Café	74
Pizza Place	67
Mexican Restaurant	63
Park	61
Hotel	59
Bar	57

	Frequency
Coffee Shop	156
Café	99
Restaurant	50
Hotel	43
Italian Restaurant	43
Bar	40
Bakery	36
Park	33
American Restaurant	30
Japanese Restaurant	30

Figure 3: Frequency of occurrence of venues, Manhattan (L) & Toronto (R)

The clustering results were generated by finding similar venues within the pre-defined radius of 500. Each dot indicates a neighbourhood and dots of same colour indicate neighbourhoods within the same cluster. Results shown in Figure 4 & 5.

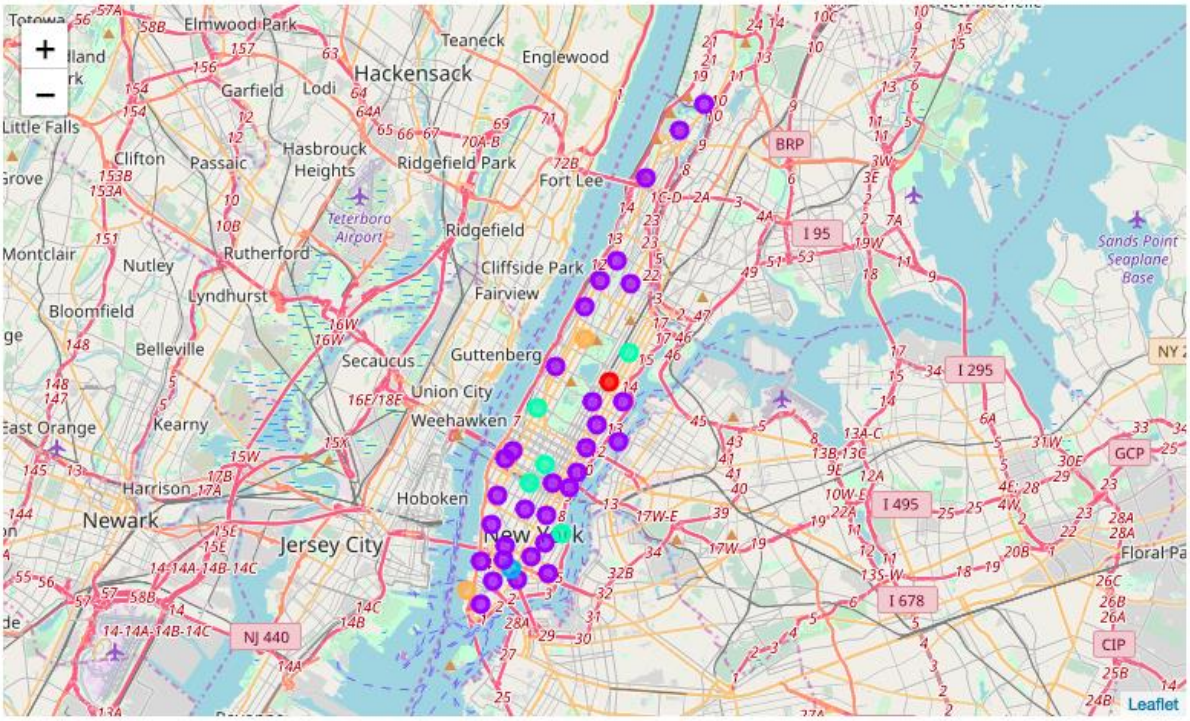

Figure 4: Clustering results for Manhattan

Figure 5: Clustering results for Toronto

Discussion

From Figure 3 in the previous section. It is clear to see that coffee shops ranks high in both districts. By summing the top ten counts and divided by the occurrence of coffee shops, it could be seen that 15.8% (127/806 in Manhattan) and 27.9% (156/560 in Toronto) of top ten lists are coffee shops. This should not be surprised because coffee is normally consumed by working people who wants to get energized, especially those working in financial or service industries. Manhattan and Toronto are both areas crowded with those careers, therefore the consumption of coffee is huge, leading to many coffee businesses within the area. It is quite surprise to see Italian restaurant ranks top within Manhattan. Through research, it has shown that United State cuisine actually was largely influenced by Italian cuisine, and multiple recipes within Italian cuisine such as pizza and Macaroni.

Further analysis could be done by displaying the detailed ranking within each cluster. One cluster example as shown in Figure 6.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
7	East Harlem	Mexican Restaurant	Bakery	Latin American Restaurant	Thai Restaurant	Deli / Bodega	Fast Food Restaurant	Beer Bar	Taco Place	Street Art	Pizza Place
13	Lincoln Square	Gym / Fitness Center	Theater	Plaza	Concert Hall	Italian Restaurant	French Restaurant	Indie Movie Theater	Bakery	Café	Opera House
15	Midtown	Hotel	Coffee Shop	Clothing Store	Steakhouse	Theater	Spa	Sporting Goods Shop	Cocktail Bar	Food Truck	Park
33	Midtown South	Korean Restaurant	Coffee Shop	Hotel Bar	Cosmetics Shop	Japanese Restaurant	Boutique	Italian Restaurant	Cocktail Bar	Hotel	Gym / Fitness Center
37	Stuyvesant Town	Bar	Park	Basketball Court	Baseball Field	Gas Station	Playground	Harbor / Marina	Tennis Court	Cocktail Bar	Coffee Shop

Figure 6: Clustering example for Manhattan with 4 members

From the analysis of the detailed clusters, it could be seen that although both districts are divided into five clusters, around 90% of all neighbourhoods are within one specific cluster. This may indicate the spread and popular venues are quite similar within the district. Among the largest clusters within the two districts, it could be easily seen that coffee shops are around 1st or 2nd in nearly all neighbourhoods. As Figure 6 is an example of a small clusters, it could be seen that within this small cluster, the popular venues such as coffee shop and Italian restaurants are not within high ranks. It could be deduced that this might be a cluster containing neighbourhoods of residential area.

Conclusion

The report started by gathering neighbourhood and geographical data for Manhattan and Toronto, mapped to venue data sourced from Foursquare by grouping neighbourhood attribute. The first analysis found the top ten most frequent venues within the districts. The results were analysed and empirical reasons of the occurrence of some high-ranking venues were given in the discussion section. Then the data were used to form clusters by K-mean method and the clusters were visualized on the districts' maps. It was deduced that there are extremely large clusters in both districts, surrounded by coffee shops and restaurant which may indicate Manhattan and Toronto are districts with large portion of financial districts. Other small areas without those venues may indicate a spread of residential areas.