Comparing the tourism prospects of two cities: New York and Toronto

Abhishek Rao

May -2020

1. Introduction

1.1 Background

The tourism industry, also known as the travel industry, is linked to the idea of people travelling to other locations, either domestically or internationally, for leisure, social or business purposes. It is closely connected to the hotel industry, the hospitality industry and the transport industry, and much of it is based around keeping tourists happy, occupied and equipped with the things they need during their time away from home.Tourism industry value chain is very large. This value chain of tourism industry is or can be elastic and flexible and can also be much larger and widespread. The broad value chain of tourism comprises of travel and tour services like booking and reservation; transportation industry (international and national travel and transportation); accommodation; hospitality industry; food & beverages; tourism products and destinations and related products and services; local travel and transportation.

Therefore it is advantageous for a tourist to have a detailed comparison of various aspects of travelling between two cities in consideration. It empowers him to optimize his time and money and select the destination that will give him the most diverse experience.

1.2 Problem

This battle of neighbourhoods will be between the neighbourhoods of two cities New York and Toronto. The primary factor that makes this competition interesting is that both the cities are geographically very similar. Both the cities have rivers flowing through them and beautiful flora and fauna which attract millions of visitors each year. Hence we will explore how similar are these cities in terms of food, accommodation, water activates.Moreover, for more specific comparison two boroughs from each country will be selected. Scarborough from Toronto and Manhattan from New York will be the contenders.¶

2. Methodology

For this problem, we will get the services of Foursquare API to explore the data of two cities, in terms of their neighborhoods. The data also include the information about the places around each neighborhood like restaurants, hotels, coffee shops, parks, theatres, art galleries, museums and many more. We selected one Borough from each city to analyse their

neighborhoods. Manhattan from New York and Scarborough from Toronto. We will use machine learning technique; “Clustering” to segment the neighborhoods with similar objects on the basis of each neighborhood data. We will use 5 centres for clustering. These objects will be given priority on the basis of foot traffic (activity) in their respective neighborhoods. The clusters will thus be analysed so that a tourist can easily select between two cities. For example Foursquare API can be used to get information about all the venues located within some radius of a neighbourhood and this method is applied for all the neighborhoods in a bourough.¶

2.1 Data Acquisition

For Scarborough case, we have extracted table of Toronto’s Borough from Wikipedia page. Then we arrange the data according to our requirements. In the arrangement phase, this applied multiple steps including but not limited to, eliminating “Not assigned” values, combine neighborhoods which have same geographical coordinates at each borough and sorted against the concerned borough. For data verification and further exploration, we use Foursquare API to get the coordinates of Scarborough and explore its neighborhoods. The neighborhoods are further characterized as venues and venue categories.¶

For Manhattan, we used a saved data file which is already explored through foursquare API in which we have extracted all the boroughs of New York and then sorted against the concerned borough. Then we explored the Manhattan neighborhoods as venues and venue categories

2.2 Data Cleaning

In case of New York, the data pertaining to the neighbourhoods was downloaded from nyu.edu. All the relevant data was included in “Features” key. This data was transformed into a Pandas dataframe. The relevant information of Manhattan dataframe was extracted as a separate entity. The latitude and longitudinal values of respective neighbourhoodwas obtained from Geo-locator and the tables were combined. Folium Library was used to visualise the neighbourhoods on the Manhattan map. This was the final dataframe that was obtained

	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 Dataframe

The data of Toronto was not available very easily. So it was obtained through Wikipedia, the table was scraped using Beautiful Soup function. The API calls were similar to that of Toronto case.

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

Figure 2: Scarborough Dataframe

2.3 Data Clustering

In both the cases we will use k means clustering with k=4.These objects will be given priority on the basis of foot traffic (activity) in their respective neighborhoods. The clusters will thus be analysed so that a tourist can easily select between two cities. For example Foursquare API can be used to get information about all the venues located within some radius of a neighbourhood and this method is applied for all the neighborhoods in the borough

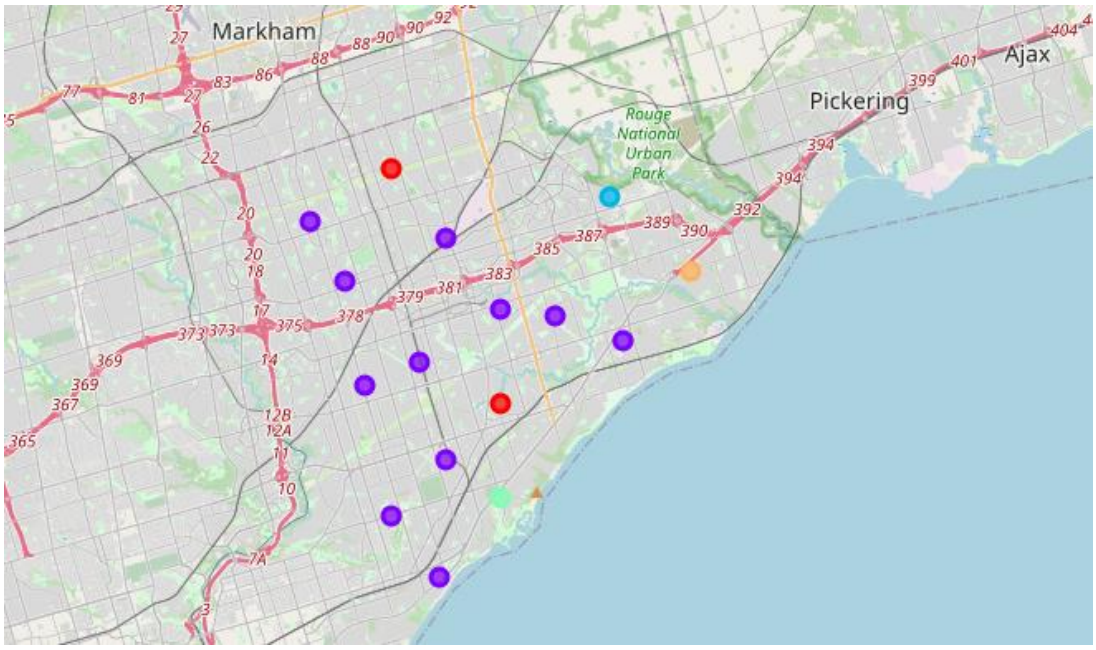

Figure 3: Scarborough Clustering

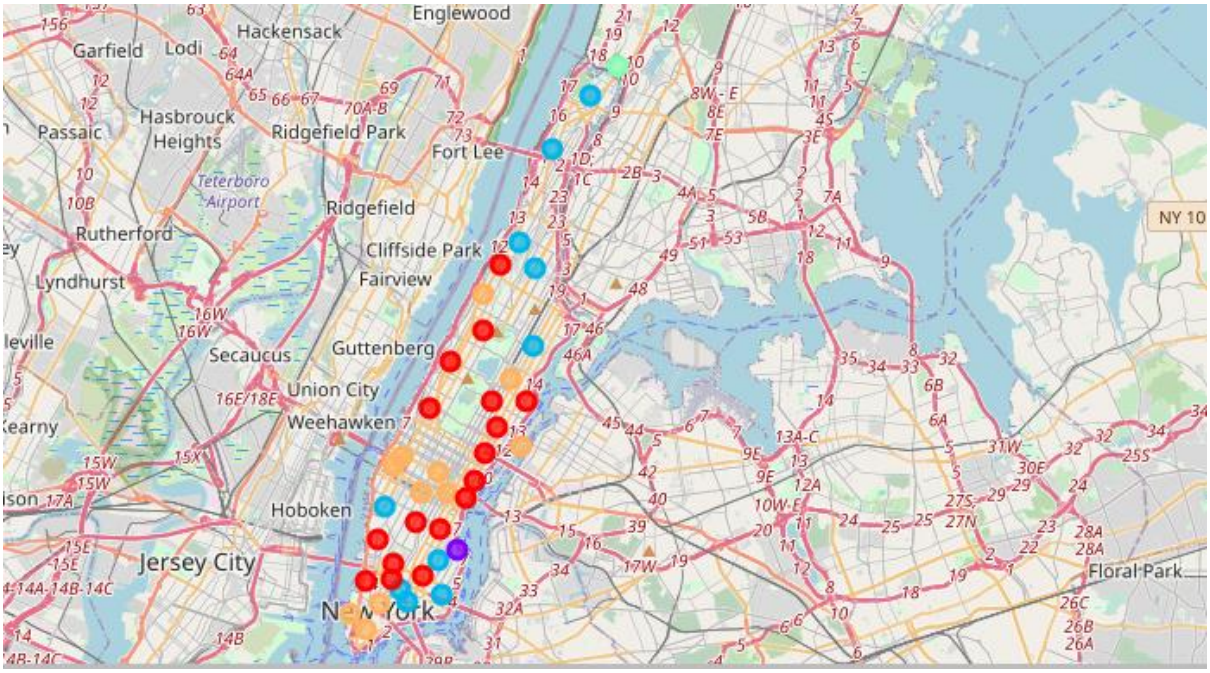

Figure 4: Manhattan Clustering

3. Results

K-means returned 5 clusters and each cluster has its own characteristics: • Cluster 1 in red: groups together neighborhoods with a relatively low score and with a relatively low restaurant count in its perimeter average.Number_of_Restaurant cluster 1 : 12.0 average score cluster 1 : 2.27 • Cluster 2 in blue: groups neighborhoods with a relatively high score and a relatively high number of restaurants also in its perimeter average Number_of_Restaurant cluster 2 : 18.8 average score cluster 2 : 2.39

• Cluster 3 in green: groups neighborhoods with a relatively low score and a relatively high number of restaurants also in its perimeter average Number_of_Restaurant cluster 3: 19.64 average score cluster 3: 2.30 • Cluster 4 in yellow: groups neighborhoods with a relatively high score and a relatively low number of restaurants also in its perimeter average Number_of_Restaurant cluster 4: 13.692307692307692, average score cluster 4: 2.4528401540266525

In case of Scarborough maximum neighborhoods are included in red cluster whereas in Manhattan, maximum neighborhoods are included in orange cluster.

4. Discussions

4.1 Scarborough

1. Cluster 1 which primarily would be an ideal choice for families to stay as it has all the essentials right from playground to business service.
2. Cluster 2 consists of a lot of Asian food restaurants hence it would create conflict amongst each other.
3. Cluster 4 and 5 would definitely be suitable to elite visitors which consists of golf course and clothing stores

4.2 Manhattan

1. Cluster 1 consists of a lot of coffee shops and cafe and hence it would be more useful for the neighborhoods to try to open something different like a bar or bookstores.
2. Cluster 2 is suitable for visitors interested in water sports.
3. Cluster 3 consists of a lot of bars; hence opening some coffee shops would be more suitable.
4. Cluster 5 would be an ideal choice for families to stay as it has all the essentials right from playground to business service

5.Conclusion

Scarborough in spite of having rivers doesn't use it efficiently for the water activities. Adding these features would definitely help neighborhoods generate more revenue

Manhattan in general consists of a lot of bars, hence it would be better if the try to explore something recreational to provide diversity.

We can rely on the results quoted before even if it remains imprecise and that because of the lack of data provided by the foursquare API, a premium account will give us the possibility of seeing the note of the people and it will facilitate us the work better, by using a far-off regression we will get the right selection of a neighborhood and the result will be more accurate. Also later we will consider more data to reinforce our choice like, the trade surrounding the restaurants, the crime scene of neighborhoods and others.