Capstone Project - The Battle of Neighborhoods

Problem & Background

This final project explores the similarity or dissimilarity in aspects from a tourist point of view regarding food, accommodation, beautiful places, and many more. We will explore, segment, and cluster the neighborhoods in New York City and Toronto, as they are famous places in the world. They are diverse in many ways. Both are multicultural as well as the financial hubs of their respective countries.

Tourism is one of the world's major economic sectors and the people most often visits those countries who are rich in heritage and developed enough from a foreign prospective, like friendly environment. Every city is unique in their own way and give something new. And now the information is so common regarding location of every place around the world on your fingertips which make it easier to explore. Therefore, tourists always eager to travel to different places on the basis of available information, and the comparison between the two cities always assist to choose the specific places or according to their choice.

Data Description

For this problem, we'll clean the data, and then read it into a pandas dataframe so that it is in a structured format. Once the data is in a structured format, 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, theaters, art galleries, museums and many more. We'll select one Borough from each city to analyze their neighborhoods. Manhattan from New York and Downtown Toronto from Toronto. We will use k-means clustering algorithm to segment the neighborhoods with similar objects on the basis of each neighborhood data. These objects will be given priority on the basis of foot traffic (activity) in their respective neighborhoods. Finally, we'll use the Folium library to visualize the neighborhoods in cities and their emerging clusters. This will help to locate the tourist's areas and hubs, and then we can judge the similarity or dissimilarity between two cities on that basis.

Methodology

We have selected two cities to explore their neighborhoods. For Downtown Toronto we extracted table of Toronto's Borough from Wikipedia page and cleaned it according to our requirements. Which include eliminating, combining neighborhoods that have some 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 Downtown Toronto and explore its neighborhoods. 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 both neighborhoods as venues and venue categories

For Toronto we select Downtown Toronto, will use following Wikipedia page, https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M, in order to obtain the data that is in the table of postal codes of Canada will transform the data into pandas dataframe. The dataframe will consist of three columns: PostalCode, Borough, and Neighborhood. Only process the cells that have an assigned borough. Eliminate cells with a borough that is 'Not assigned'. Cell has a borough but a 'Not assigned' neighborhood, then the neighborhood will be the same as the borough. Discard other borough that are not Downtown Toronto and combine neighborhoods having geographical coordinates. That's how the data frame will look like after these all changes.

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

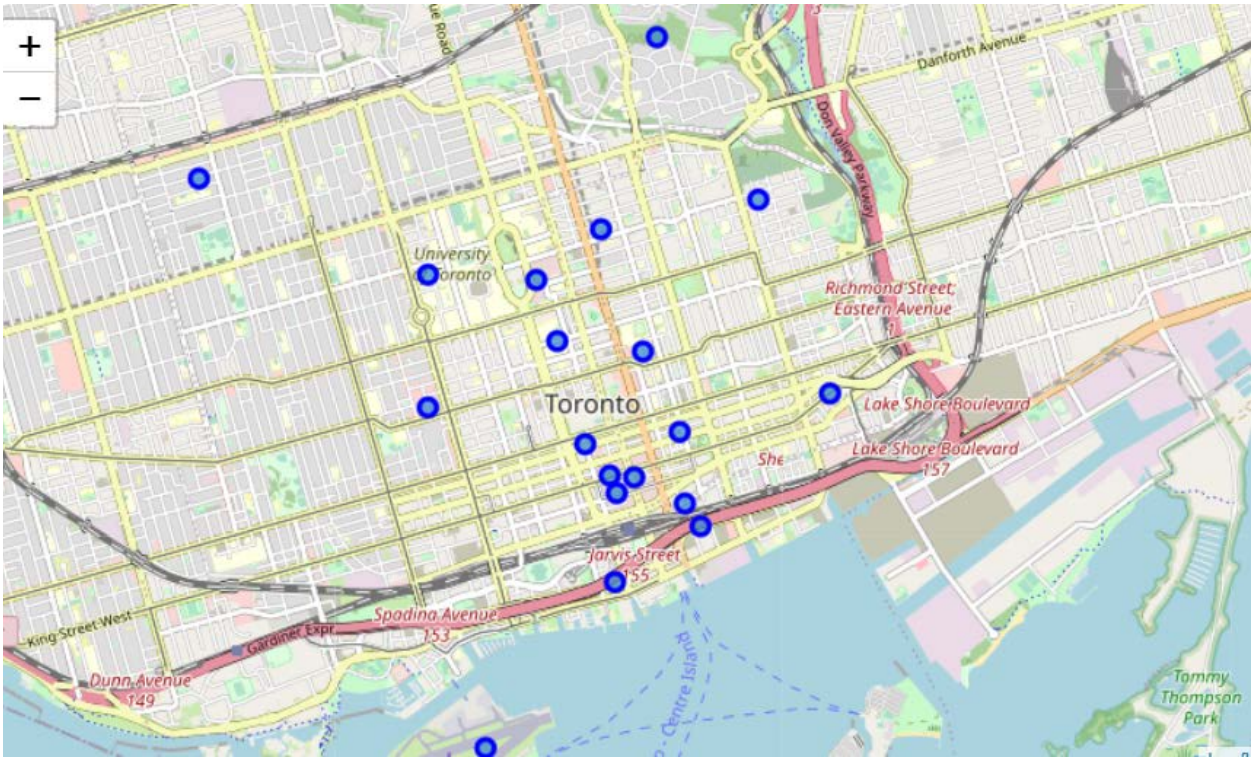
For New York we select Manhattan, will use given data load and transfer it into pandas dataframe and combine neighborhoods having geographical coordinates. Resulting data frame show below.

	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
5	Manhattan	Manhattanville	40.816934	-73.957385

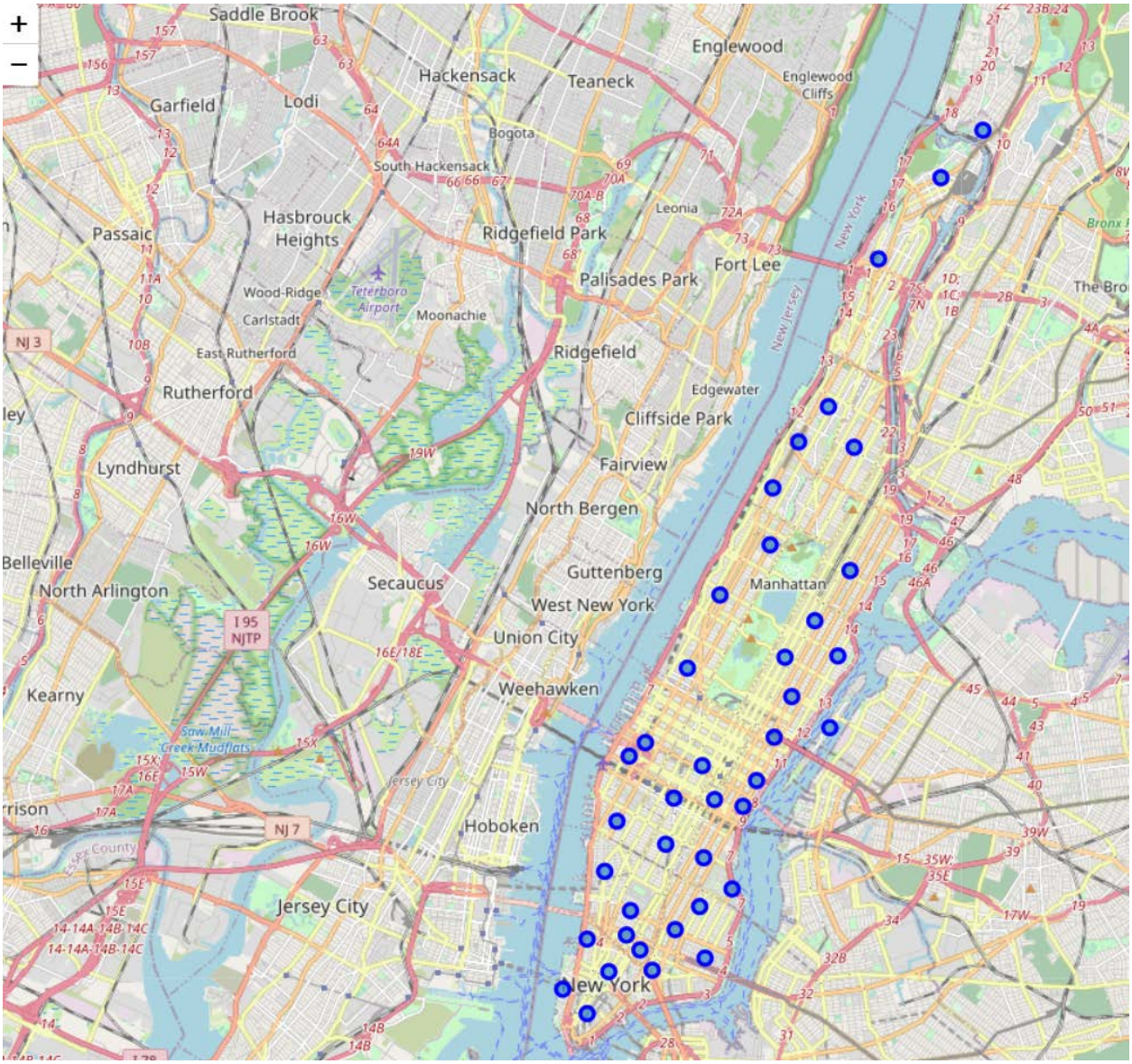
Visualization

Folium is a great visualization library, each blue circle mark reveal the name of the neighborhood and its respective borough.
We visualize the data many times at different stages. In the beginning, we visualize the selected borough neighborhoods so that we can get an idea or confirmation regarding the coordinates of that Borough. The second time after clustered the neighborhoods, we visualize the clusters.

Downtown Toronto



Manhattan



Analysis

We analyze both boroughs neighborhoods through one hot encoding (giving ‘1’ if a venue category is there, and ‘0’ in case of venue category is not there). On the basis of one hot encoding, we calculate mean of the frequency of occurrence of each category and picked top ten venues on that basis for each neighborhood. It means the top venues are showing the foot traffic or the more visited places.

Segmentation

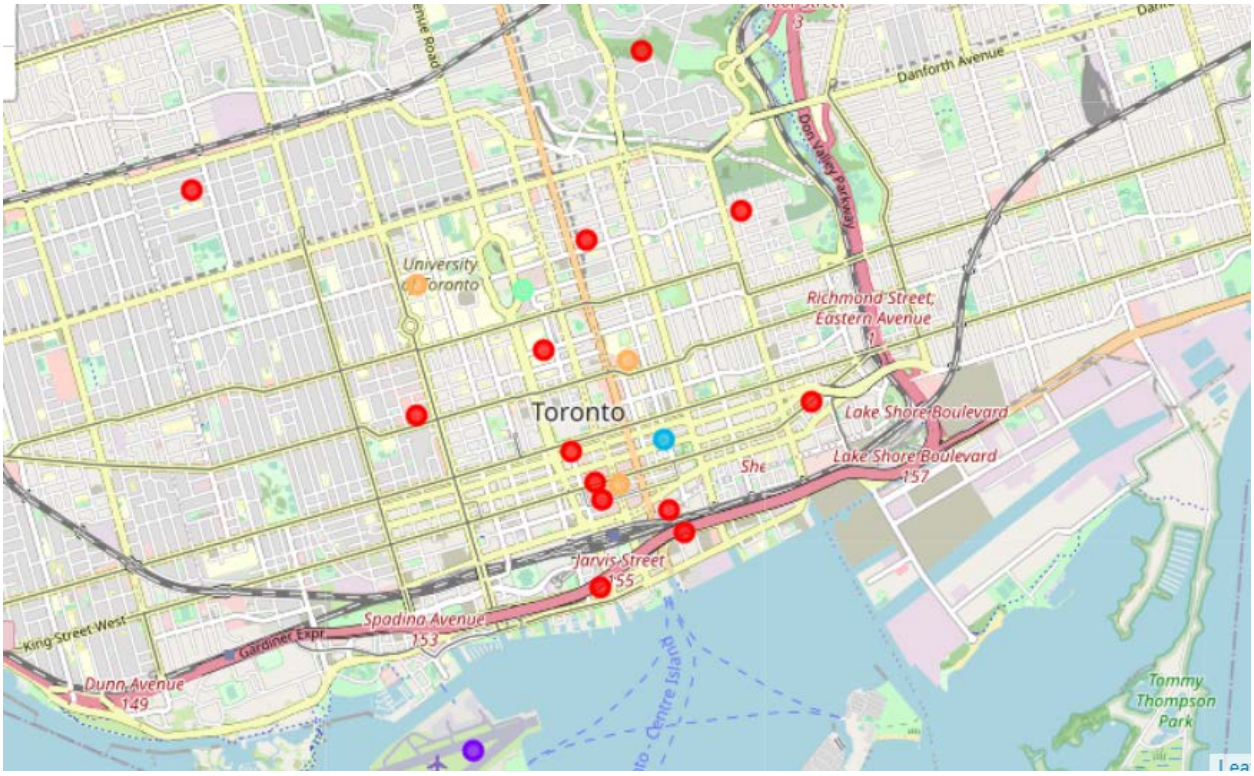
However, for illustration purposes, let's simplify the above map and segment and cluster only the neighborhoods in Downtown Toronto and Manhattan. Will be utilizing the Foursquare API to explore the neighborhoods and segment them.

Clustering Neighborhoods

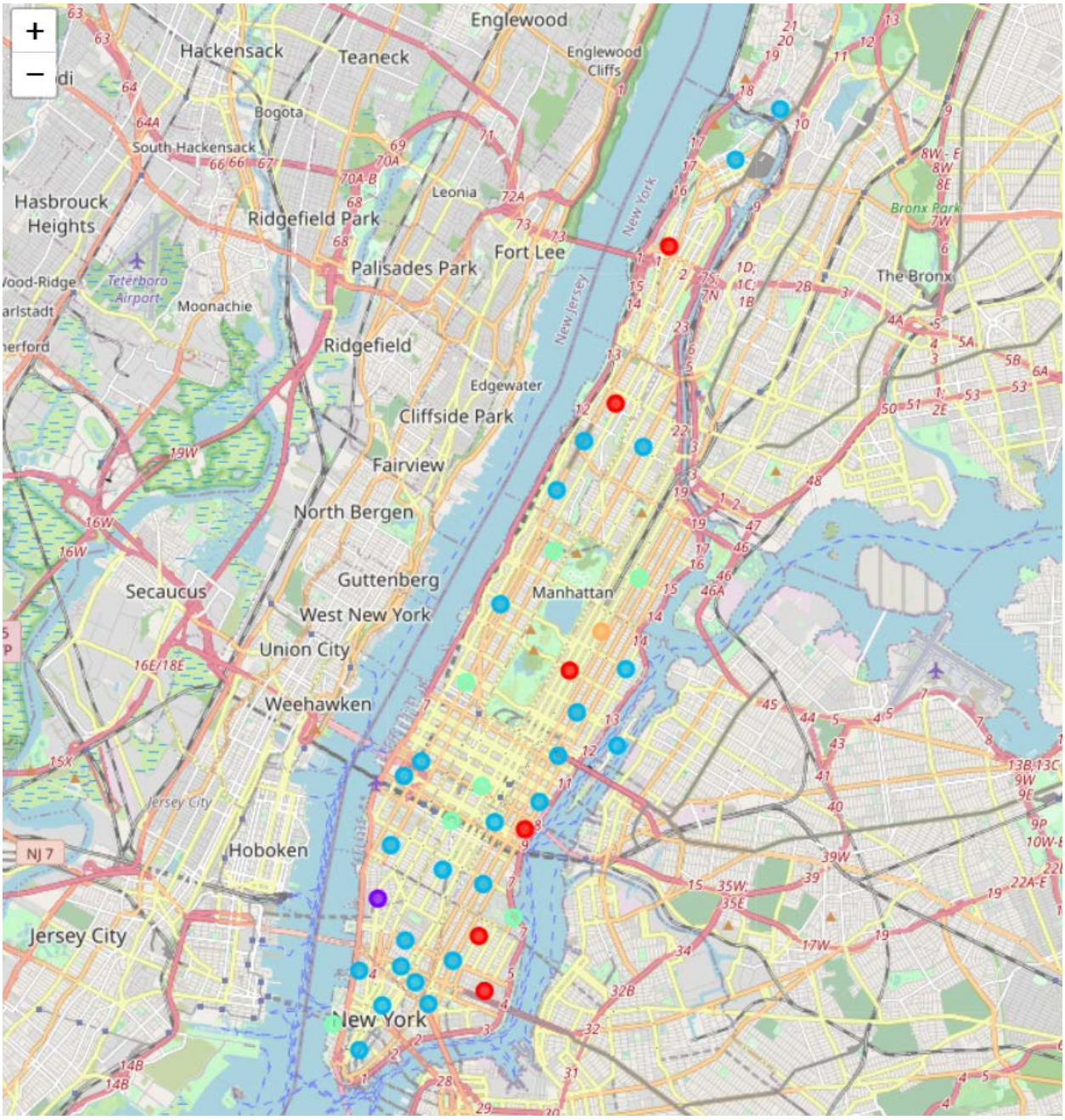
Now we applied Machine Learning Technique “Clustering” to segment the neighborhoods in similar objects cluster. This will help to analyze from Tourist perspective and we can easily extract the Tourist places which are present on one of the clusters.

Then, we can examine each cluster and determine the discriminating venue categories that distinguish each cluster.

Downtown Toronto



Manhattan



Result

After clustering the data of the respective neighborhoods, both cities (Boroughs) have venues which can be explored and attract the Tourists. The neighborhoods are much similar in features like Theaters, opera houses, food places, clubs, museums, parks etc. As far as concern to dissimilarity, it differs in terms of some unique places like historical places and monuments.

Observations & Recommendations

When we compare the tourist places, we observe that the historical place is only situated in Downtown Toronto and the Monument or landmark venue is in Manhattan neighborhoods. Similarly, Airport facility, Harbor, Sculpture garden and Boat or ferry services are also available in Downtown Toronto while venues like Nightlife, Climbing gym and Museums are present in Manhattan. As far as concern to recommendations, we recommend Downtown Toronto Neighborhoods will be considered first to visit. The tourists have an easily travelling access due to Airport facility, which not only saves time but also helps to save money. This saved money can be utilized to explore more, the attracting venues.

Conclusion

The downtown Toronto and Manhattan neighborhoods have more like similar venues. As we know that every place is unique in its own way, so that’s argument is present in both neighborhoods. The dissimilarity exists in terms of some different venues and facilities but not on a larger extent.