Comparing Neighborhood similarities between Toronto, ON, and Boston, MA.

Anthony Oppong-Gyebi

Introduction

Background

The US, a cosmopolitan country sees many people around the globe travel in and out annually. Many of those immigrants settle in at varying times to start families, businesses as well as on long vacations. Such immigrants preferably would want to situate in states which presents with similar amenities as their original homes, mainly for ease of adjustment to the new environment. For instance, persons traveling from coastal areas will wish to find homes and jobs in places where they can have access to the beaches as well as enjoy seafood, all year round.

Problem

One major challenge immigrants face choosing neighborhoods which suitably fit their interests and like places where they had migrated from, considering unavailability of comprehensive comparative information for neighborhoods. The aim of this project, therefore, is to highlight the similarities in Toronto and Boston neighborhoods for Toronto, ON residents who intend to travel to Boston, MA and vice versa. This will be addressed segmenting and clustering the neighborhoods using machine learning techniques to compare the two cities.

Data Acquisition and Cleaning

Source

The JSON file for Boston neighborhood will be obtained from the link https://github.com/dj/boston/blob/master/data/boston-neighborhoods.json and web scraping will be used for the link https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M to glean the neighborhoods for Toronto.

Data Cleaning

The datasets contain extra information that must be gotten rid of to make the needed information clean and easy to work with. The Toronto dataset will be extracted from the borough and neighborhoods in the Canada dataset link above into a new dataframe while the Boston

data geojson file will be used to obtain the boroughs and neighborhoods for Boston into another dataframe.

Feature selection

The distinct categories of venue locations in the neighborhoods of the two cities will be used to construct the features of the product. The coordinates of each neighborhood, that's the latitude and longitudes shall be extracted from the datasets and used to obtain the diverse venues in the neighborhoods to bring to bear the similarities. Foursquare API will be used to generate the locations/ amenities within the specified vicinities followed by one-hot encoding to obtain the required features from the distinct categories. A mean of the total number of venues per neighborhood will be determined as representative of the fraction of distinct categories for the neighborhood. K Means clustering will be used to segment the neighborhoods to determine the key similarities between the two cities, Toronto and Boston.

Methodology

The neighborhood coordinates were used in Folium maps to determine the locations and distribution of neighborhoods in the two cities, Toronto and Boston (Fig. 1 and 2).

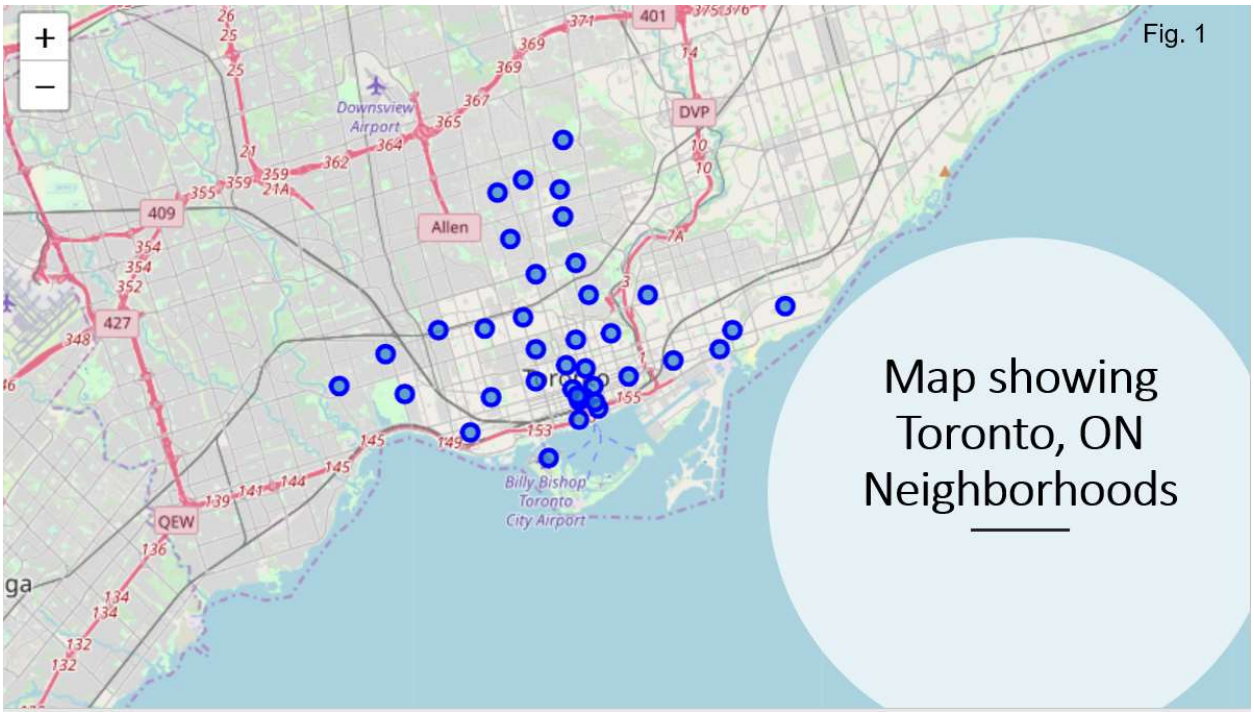
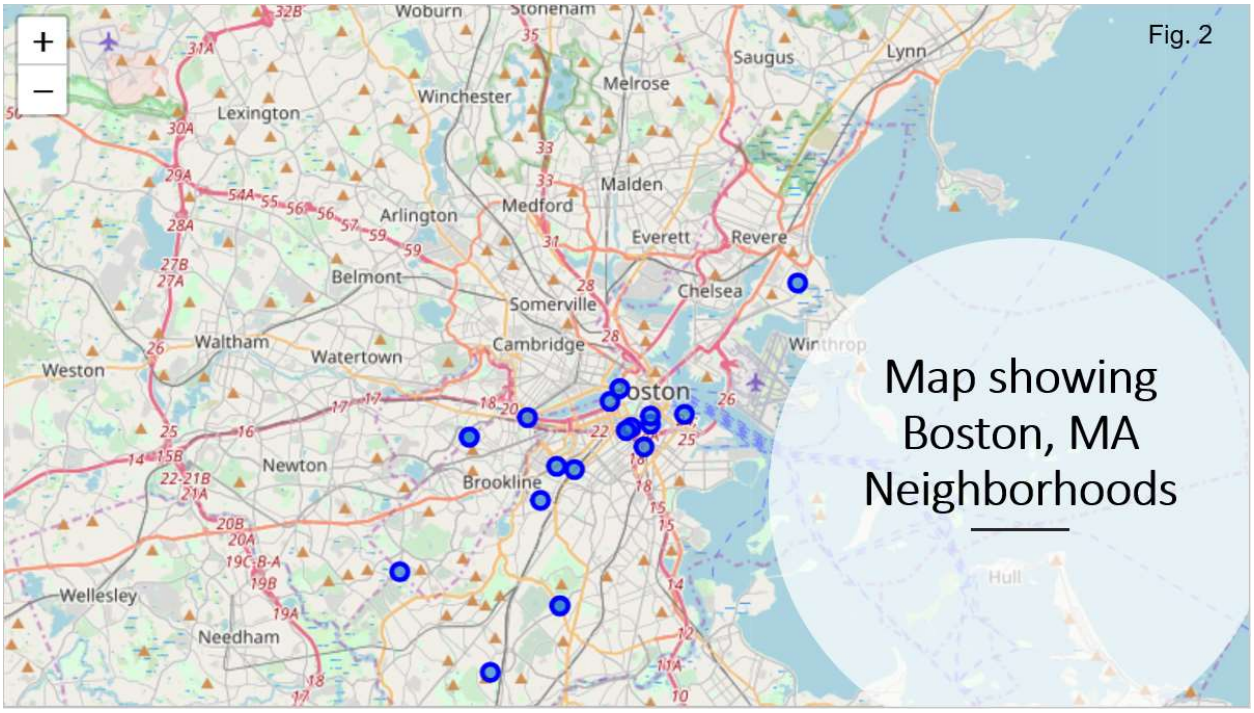


Fig. 1



Foursquare API was further run to explore the various venues present in the neighborhoods of both cities. Averaging the number of venues in the two cities showed a slightly higher mean number of venues in Toronto than Boston. Fig. 3.

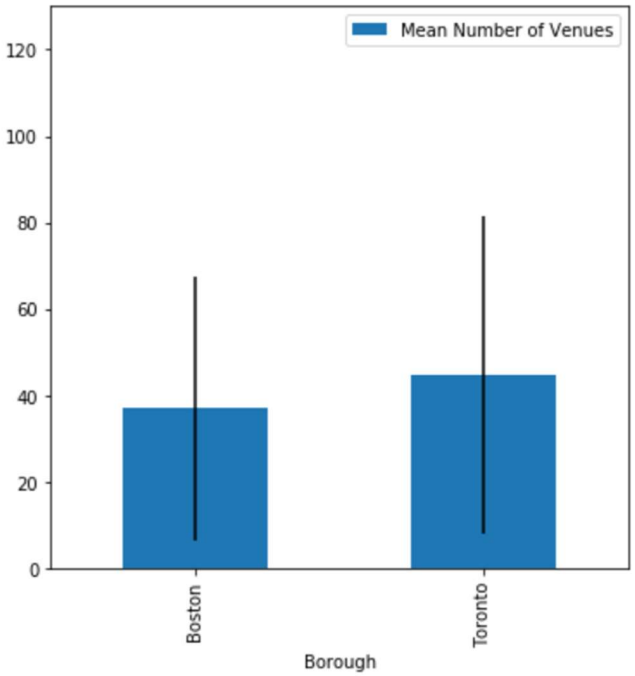


Fig. 3 Mean number of venues between Boston and Toronto

One-hot encoding was employed to generated charts which showed the first 25 most common venues per city making use of the data from each of the neighborhoods. Toronto has predominantly Coffee shops, Parks and Cafes as the three most common venues while Boston presents with Bars, Middle Eastern Restaurants and Dance Studios. The divergence in the distribution is possibly due in part to the climate, lifestyle and population diversity differences in the two cities.

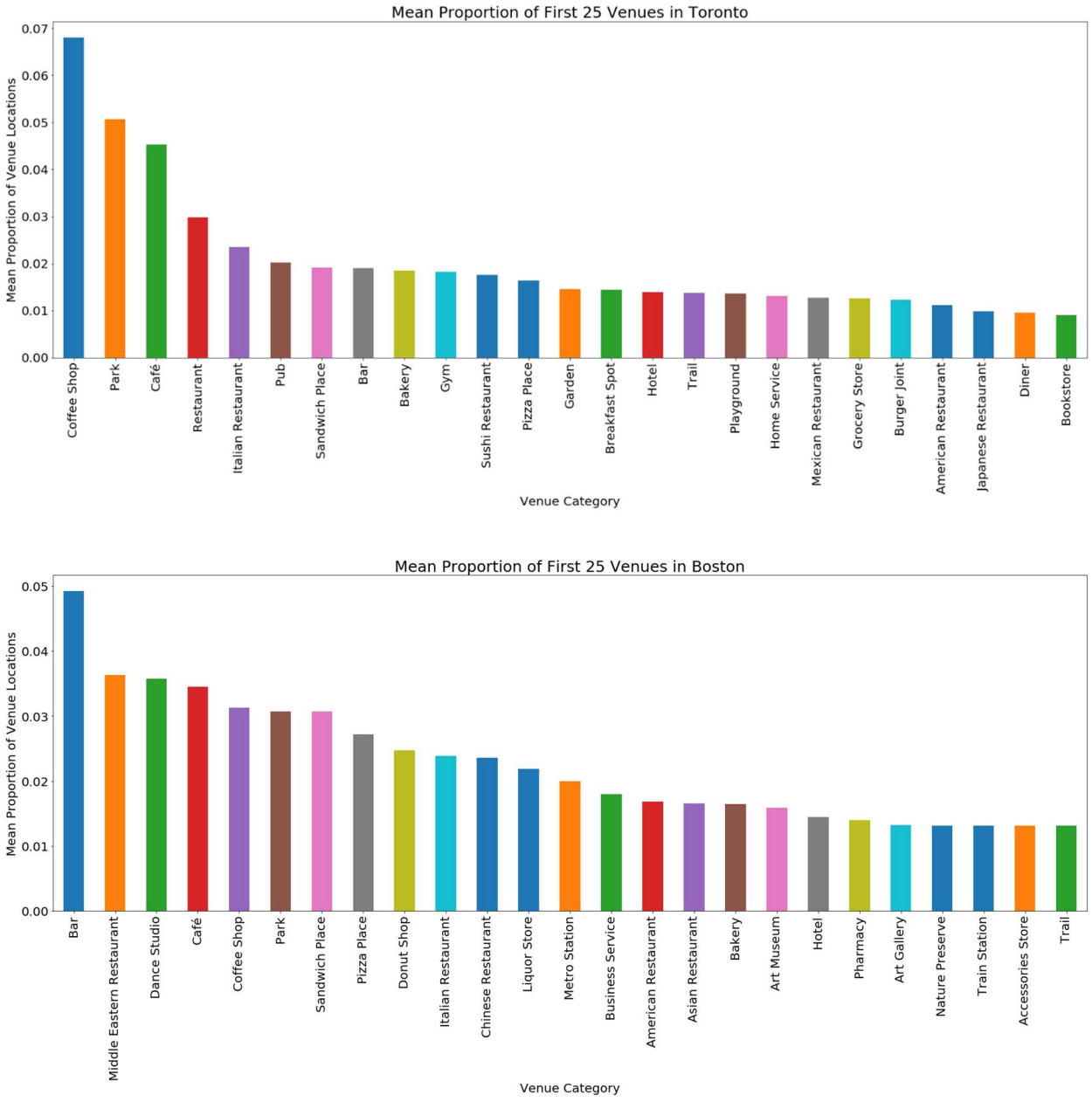


Fig. 4 Mean Proportion of First 25 most common Venues in Boston and Toronto

Results/ Discussion

Clustering and Segmentation

K Means clustering approach was used to segment and cluster the neighborhoods by way of determining similar neighborhoods between the two cities. K Means model was run with cluster numbers from 2 to 9 with inertia value of each cluster measured to show how close the individual points are within the clusters were compared. The outcome of the model followed a stepwise downward trend (decreasing inertia) with nine clusters having the least inertia. This observation suggests that increasing the clusters might further decrease the inertia (Fig. 5).

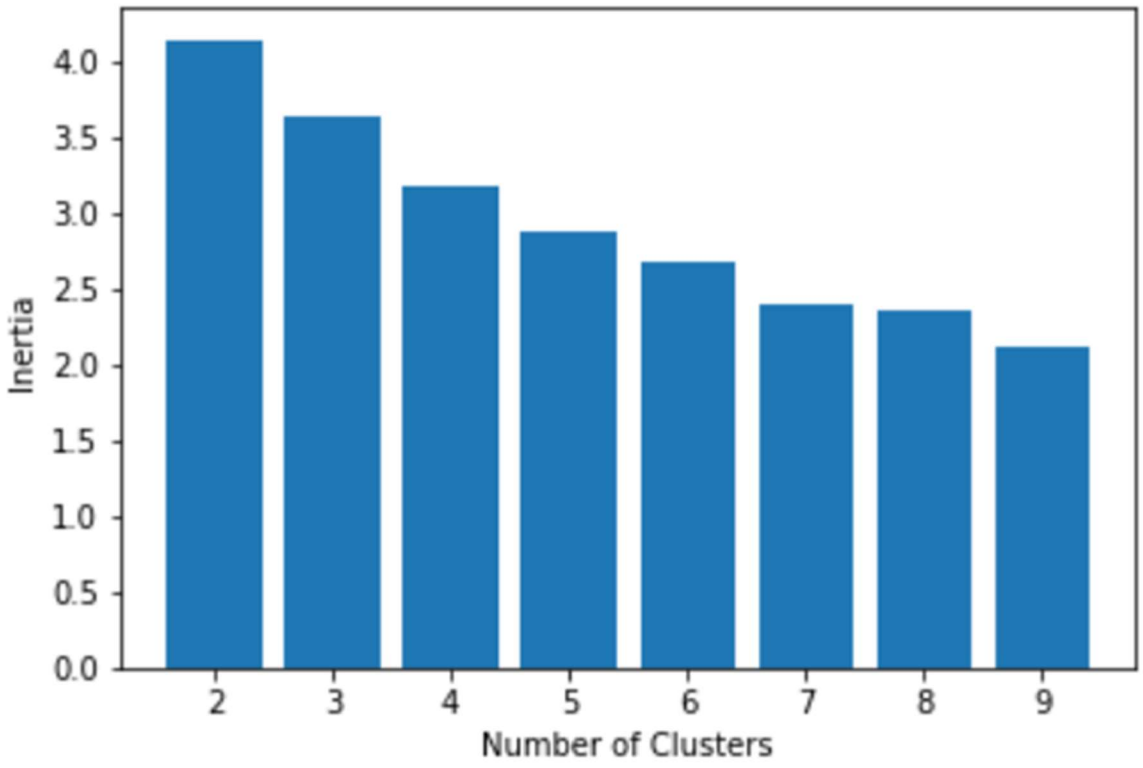


Fig. 5 Choosing Best K Means Model

Further analysis of the model with nine clusters revealed that only two out of the nine clusters had neighborhoods in both Toronto and Boston while the remaining seven had only neighborhoods in either Boston or Toronto. An observation of similar neighborhoods in each city was made even though the neighborhoods in Boson formed different clusters from Toronto's.

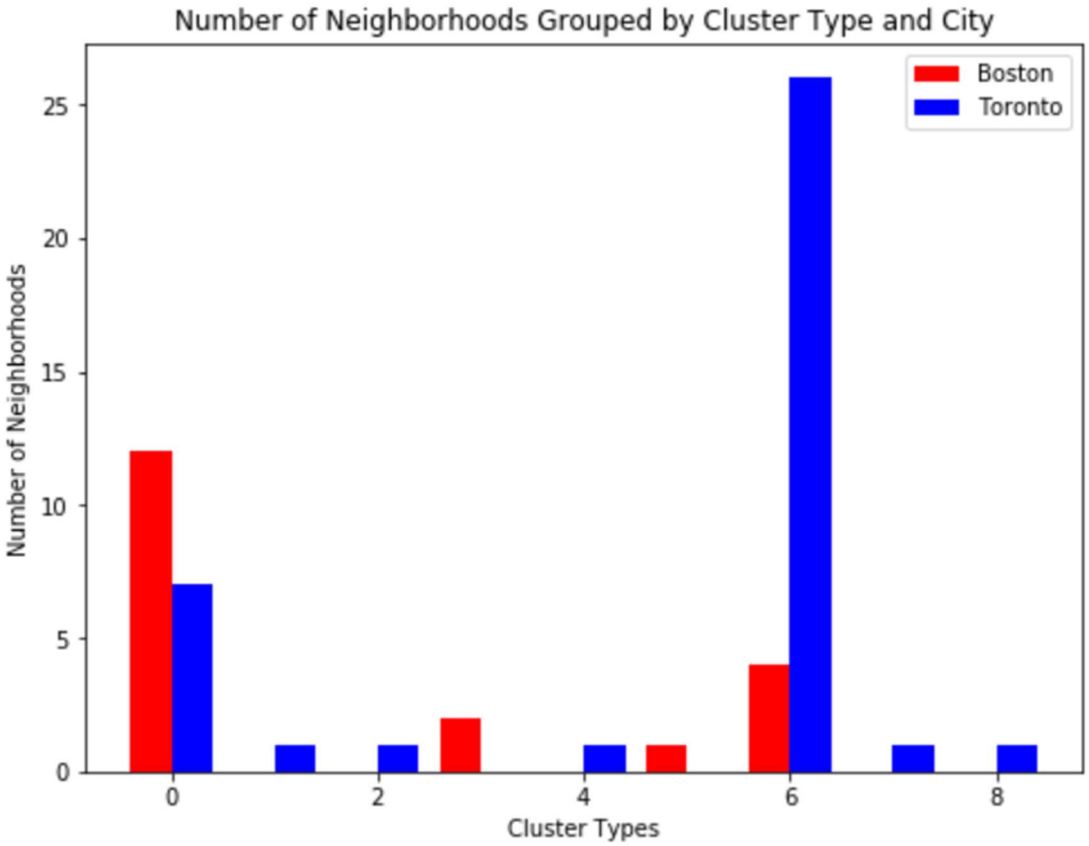
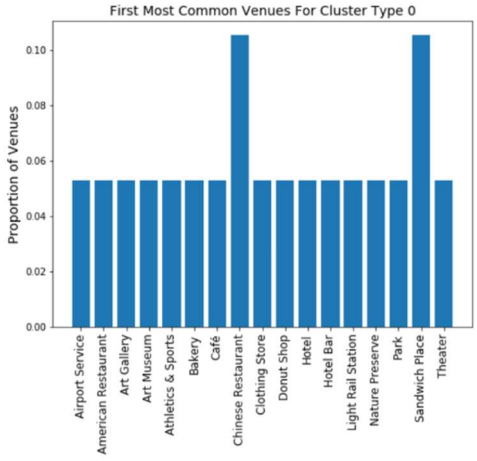
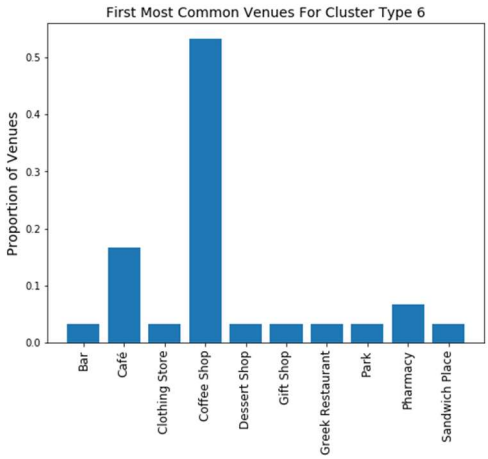


Fig. 6 Number of Neighborhoods by Cluster and City

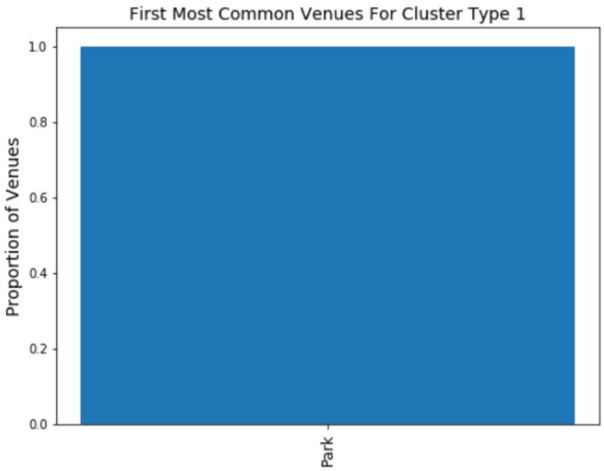
A comparison between the cluster types showed significant differences in the distribution of venues. The most common venue in Cluster Type 0 was Chinese Restaurants and Sandwich

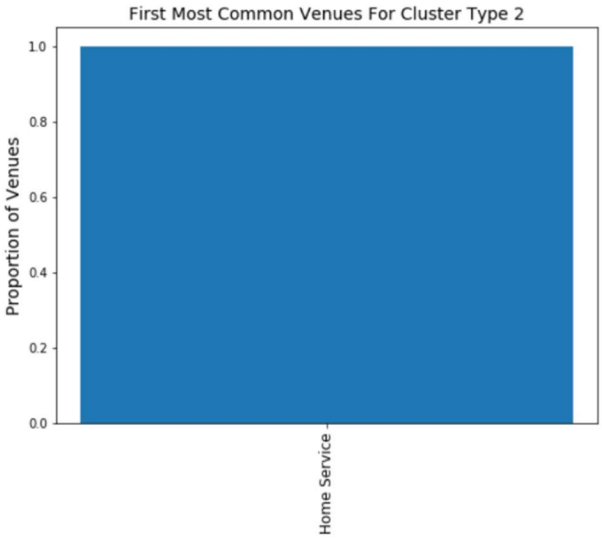


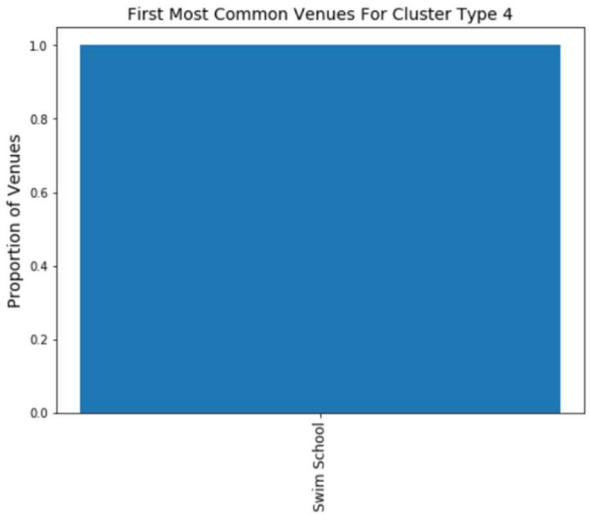


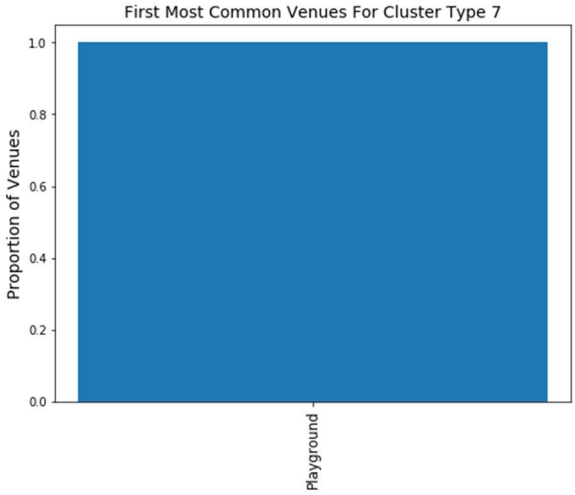
Place suggesting an area with a high density of food joints and will be good places for travelers and settler who would want to try some Asian foods. The other Cluster Type where similar venues were drawn from both cities was Cluster 6 with the most common venue being Coffee Shop and it is predominantly found in Toronto neighborhoods rather than Boston.

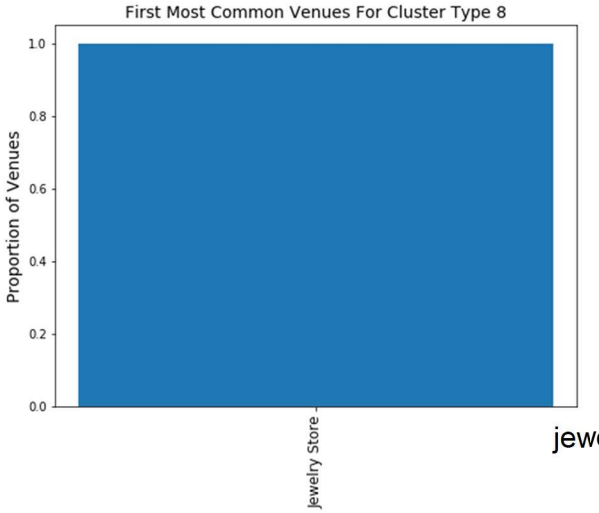
Cluster Types 1, 2, 3, 4, 5, 7 and 8 were the cluster types with only neighborhoods in either Boston or Toronto. Cluster Types 1, 2, 4, 7 and 8 are found within only Toronto neighborhoods with the most occurring being Park, Home Service, Swim School, Playground and Jewelry Store respectively.





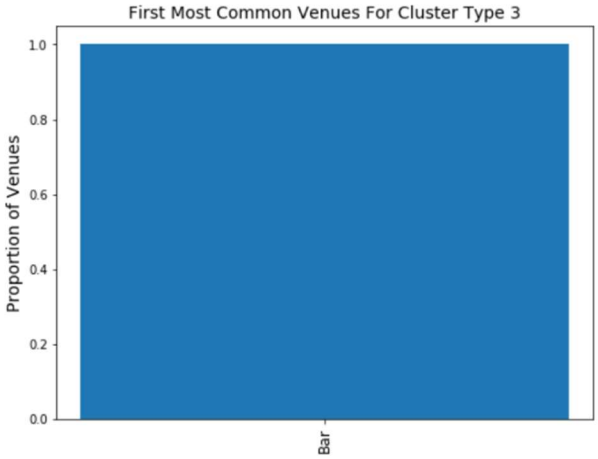


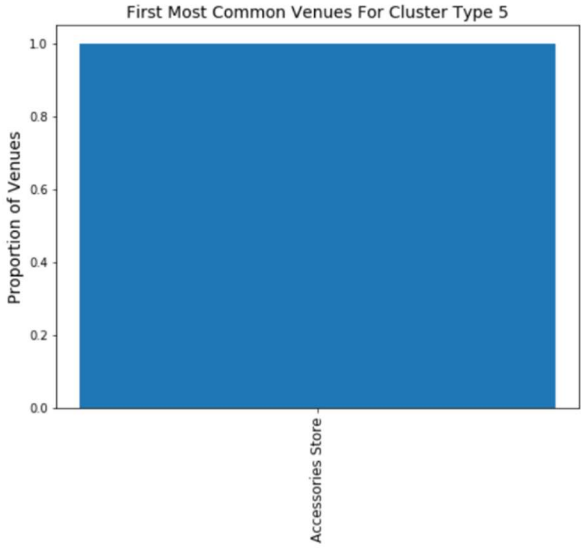




The presence of numerous parks and playgrounds in Toronto Neighborhoods is an indication for the availability of recreational stuff to take advantage while in Toronto compared with Boston. There seem to be many jewelry stores in Toronto Neighborhoods as well compared with Boston, which suggests traveler from Toronto to Boston are not likely to come by jewels as they would in their neighborhoods.

Cluster Types 3 and 5 are found within only Boston neighborhoods with the most occurring being Bar and Accessories Store respectively. Boston has more bars as common venues across its neighborhoods suggesting more of a lifestyle of the inhabitants of Boston compared with Toronto.





Conclusion

We have seen from this project that two cities can be compared using Foursquare API and machine learning strategies to reveal any similarities in the amenities available in the

neighborhoods. Comparing Toronto and Boston neighborhoods revealed that the most common venues in Toronto neighborhoods are Coffee shops, Parks and Cafes as opposed to Bars, Middle Eastern Restaurants and Dance Studio in Boston neighborhoods. Boston and Toronto shared similarities in the amenities and venues after clustering and segmenting. These included Airport Services, American Restaurants, Art Gallery and Sandwich Places. The relevance of this strategy is to inform immigrants about the convergence or divergence of their new/prospective environments compared with their original neighborhoods.

Future Direction

Moving forward, there will be a need to further explore the subtle similarities and differences like mortgage rates, seasonal weather conditions and rate of crime. Upgrading the model to capture these parameters will underscore a broad-based analysis of the details of any two cities to provide adequate information to immigrants.