Finding Similar Neighborhoods between Toronto, ON and Queens, New York

Collins Opoku-Baah

Background

- The world has become one giant global village and thus, each and every minute, people travel from one place to another. The purpose for traveling could be temporary that is for vacation, business, visits etc. or could be permanent e.g. school, work etc.
- When people live in a particular region for a long time, they tend to embrace the cultures (e.g. food, clothing, language etc.) of that region, making it very difficult to transition into other neighborhoods. For example, a person who loves seafood and attend yoga classes will want to move to a new place with such venues in order to continue having their pleasant life experience

Problem

- Finding a place that share similar venue as your current neighborhood can be burdensome considering how developed most cities in the world are currently.

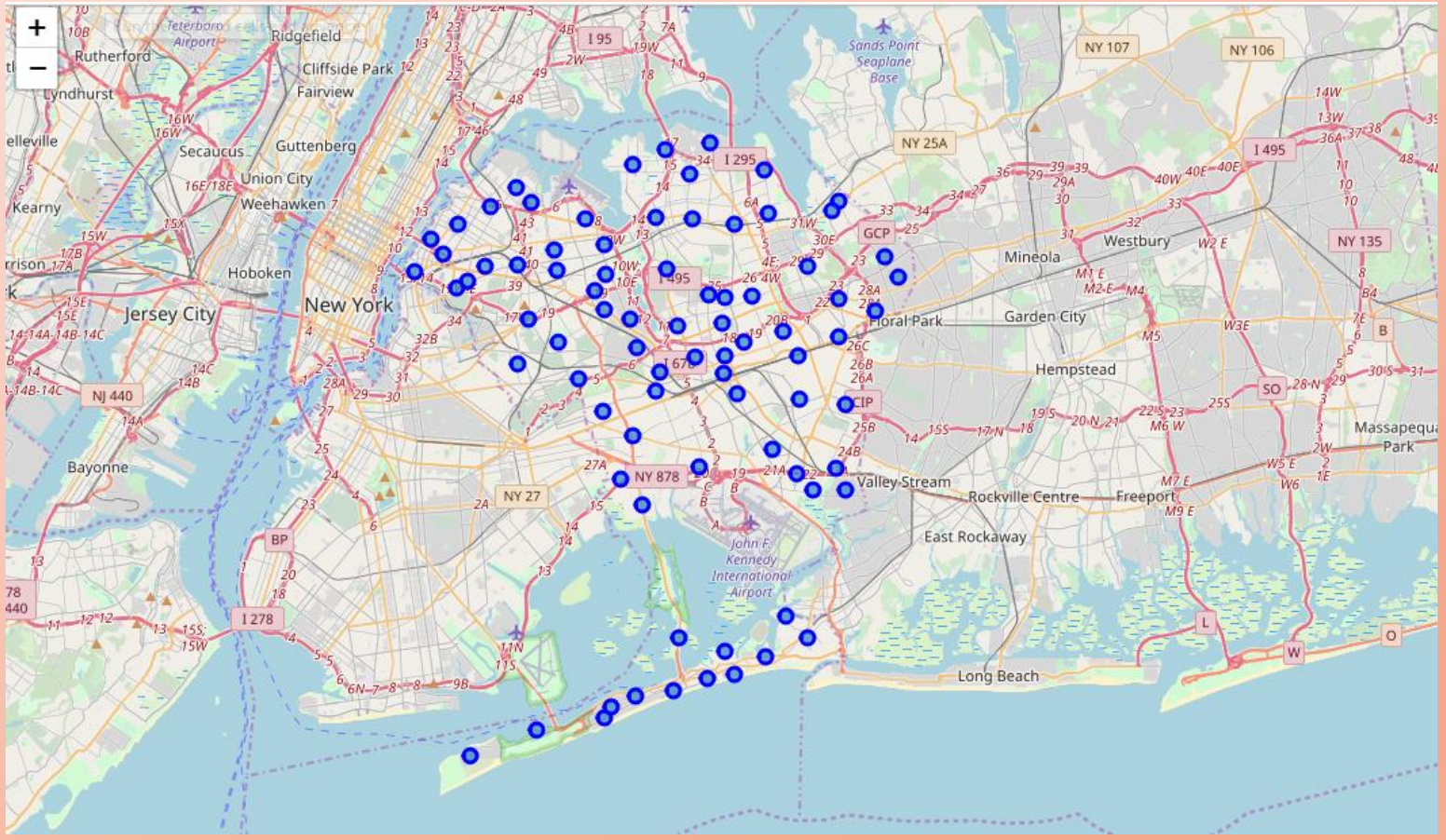
- Hence, this project aims to find neighborhoods between two big cities namely Toronto, ON and Queens, New York that are similar with respect to venues.

- To do this, I will employ machine learning approaches and other techniques to segment and cluster neighborhoods in these two cities

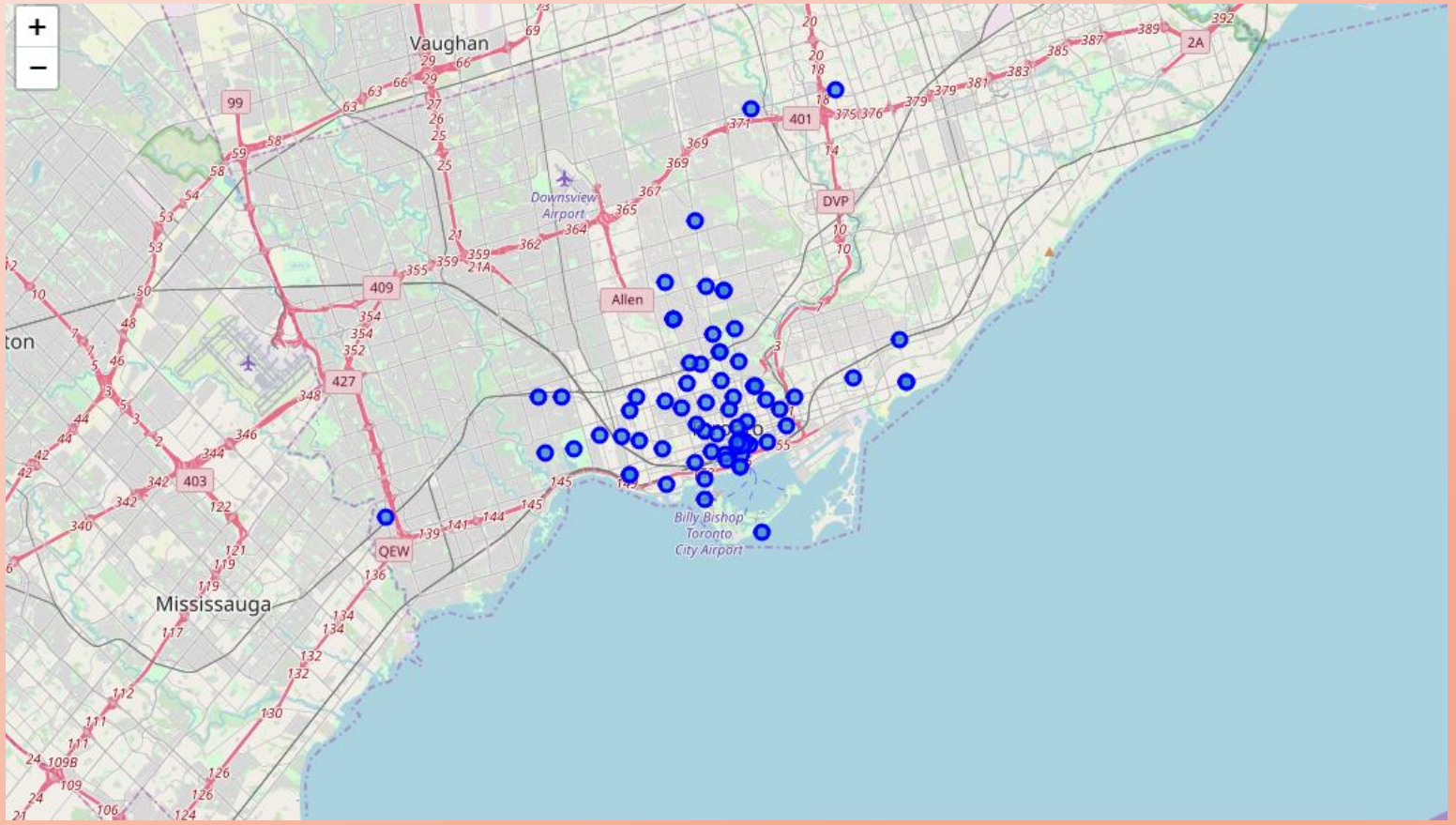
Data Collection and Cleaning

- The data for this project comprised the venue locations that are within a radius of 750 meters around the neighborhoods in the two big cities, which are Toronto, ON, Canada and Queens, New York, USA.
- The neighborhoods for Toronto, ON were obtained by scraping the website, https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- The neighborhoods for Queens, New York were obtained by downloading json file from https://cocl.us/new_york_dataset.
- Data from both cities were cleaned and combined into one dataframe.

Folium map showing the neighborhoods of Queens, New York.



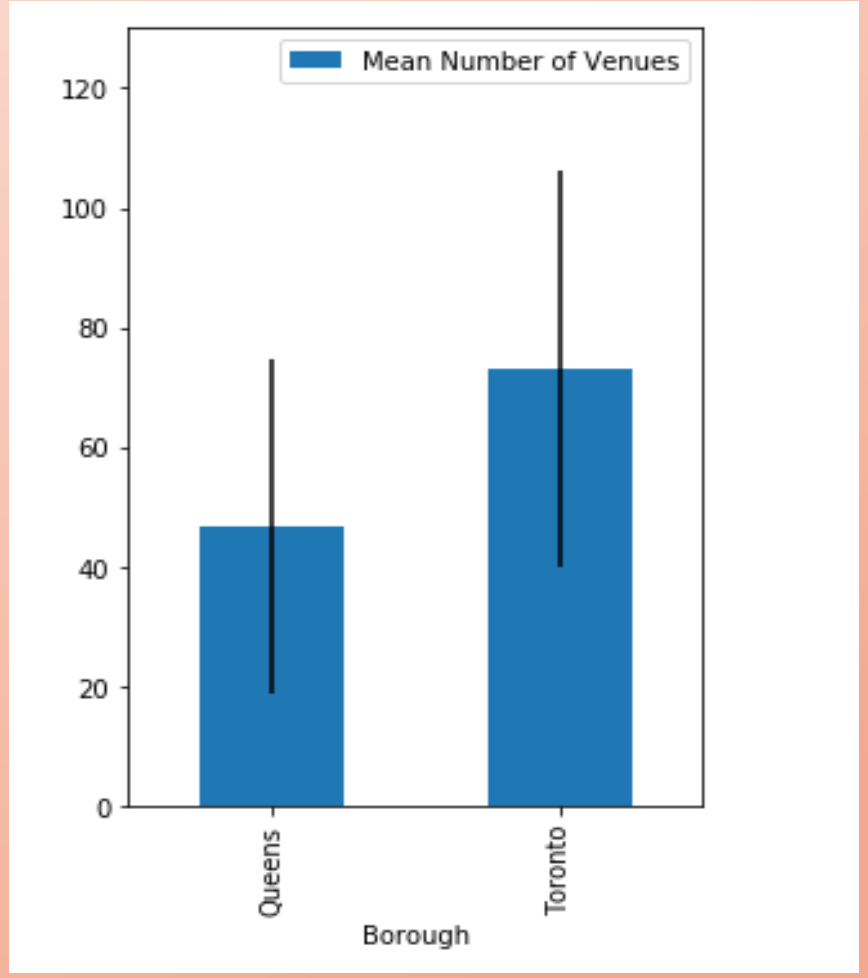
Folium map showing the neighborhoods of Toronto, Ontario.



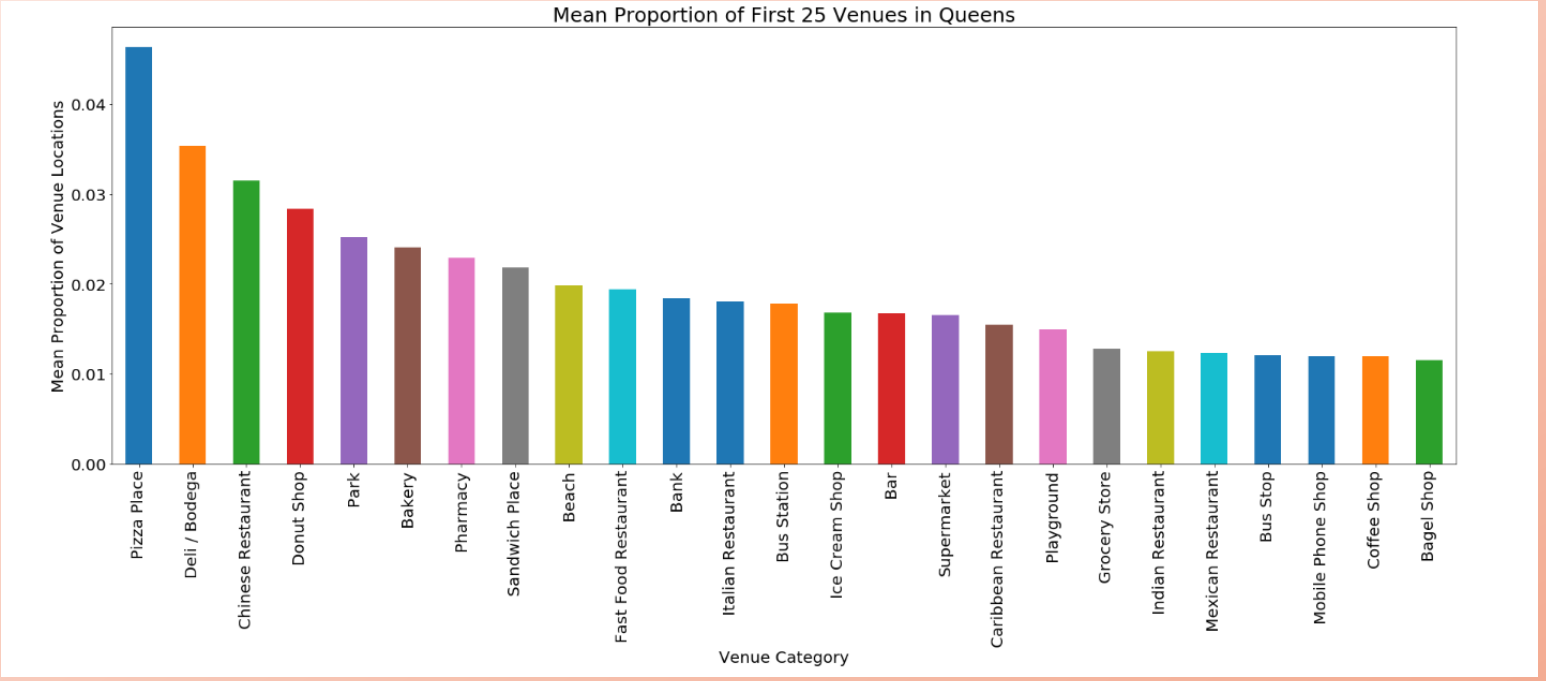
Feature Selection

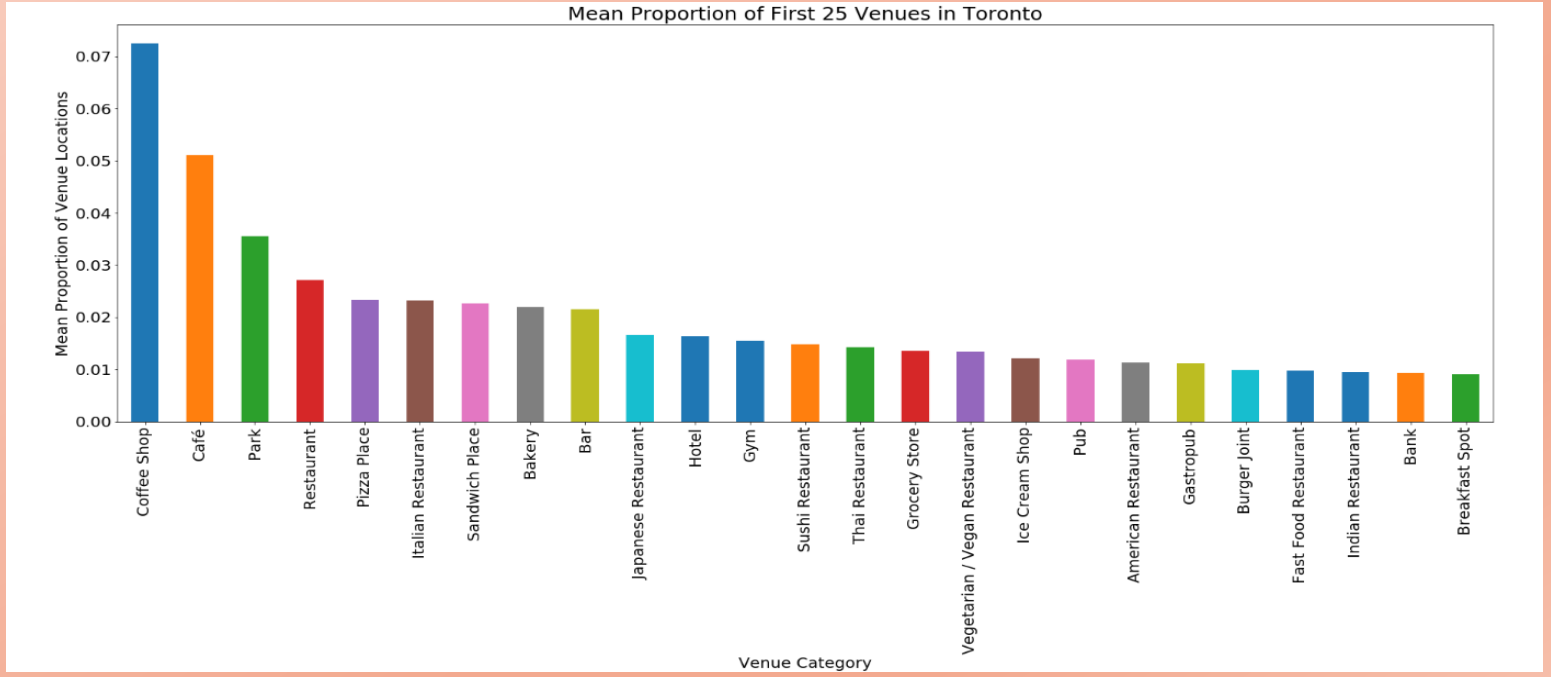
- Latitude and Longitude coordinates for all neighborhoods were obtained using geocoder API.
- A maximum of 100 venues were obtained for within 750 meters of each neighborhood using the foursquare API.
- One-hot encoding is used create features based on the distinct venue categories.
- Data points are the proportion of the venue category in that neighborhood.

Toronto had a higher mean number of venues than that of Queens



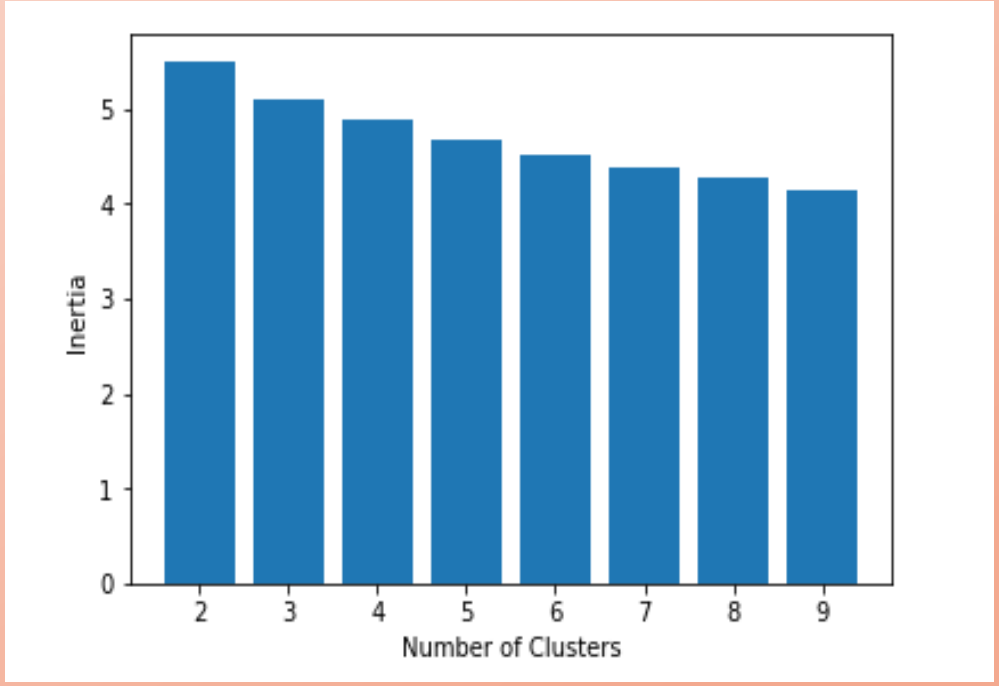
Mean Proportion of First 25 Venues in Queens and Toronto



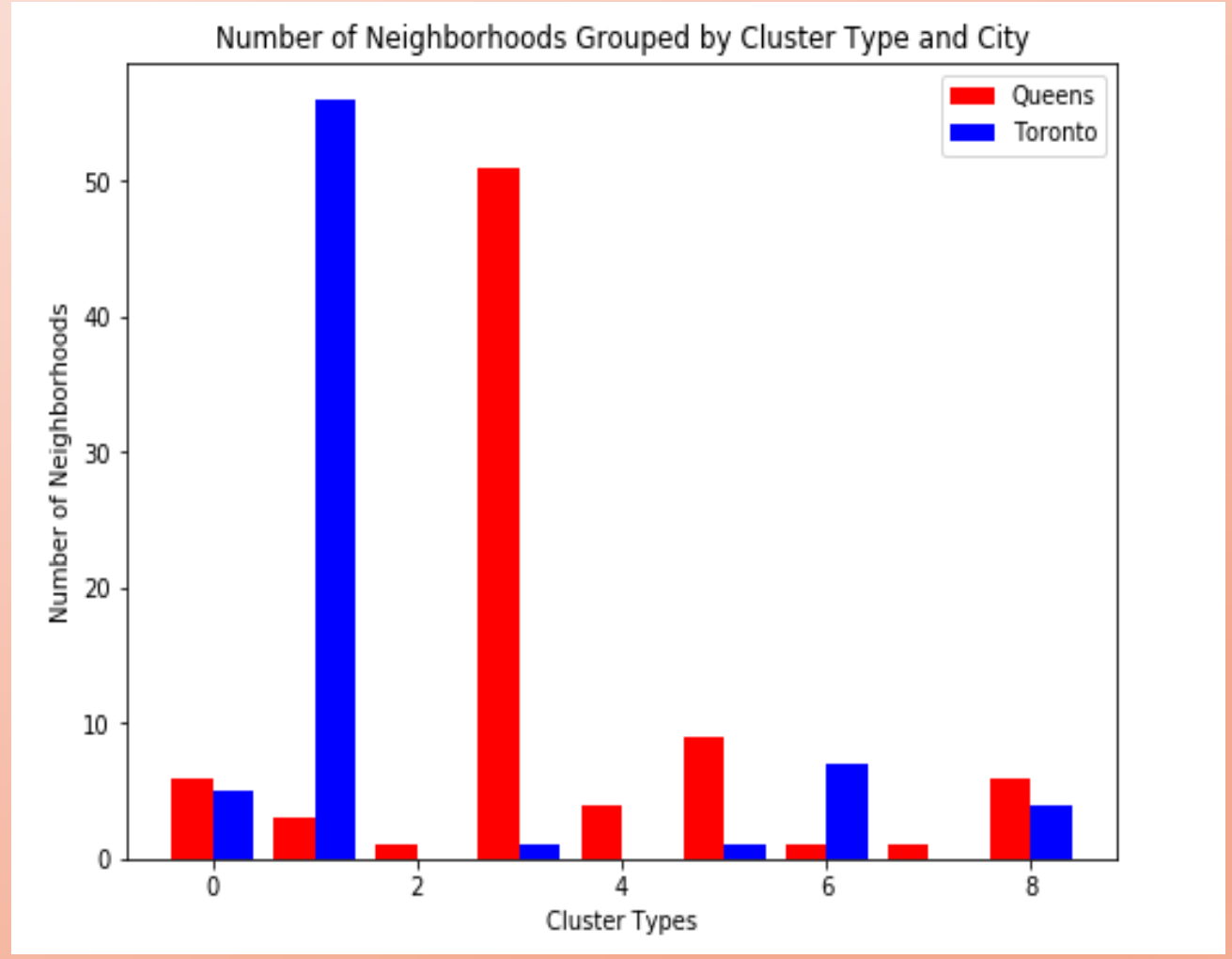


Choosing Best K Means Model

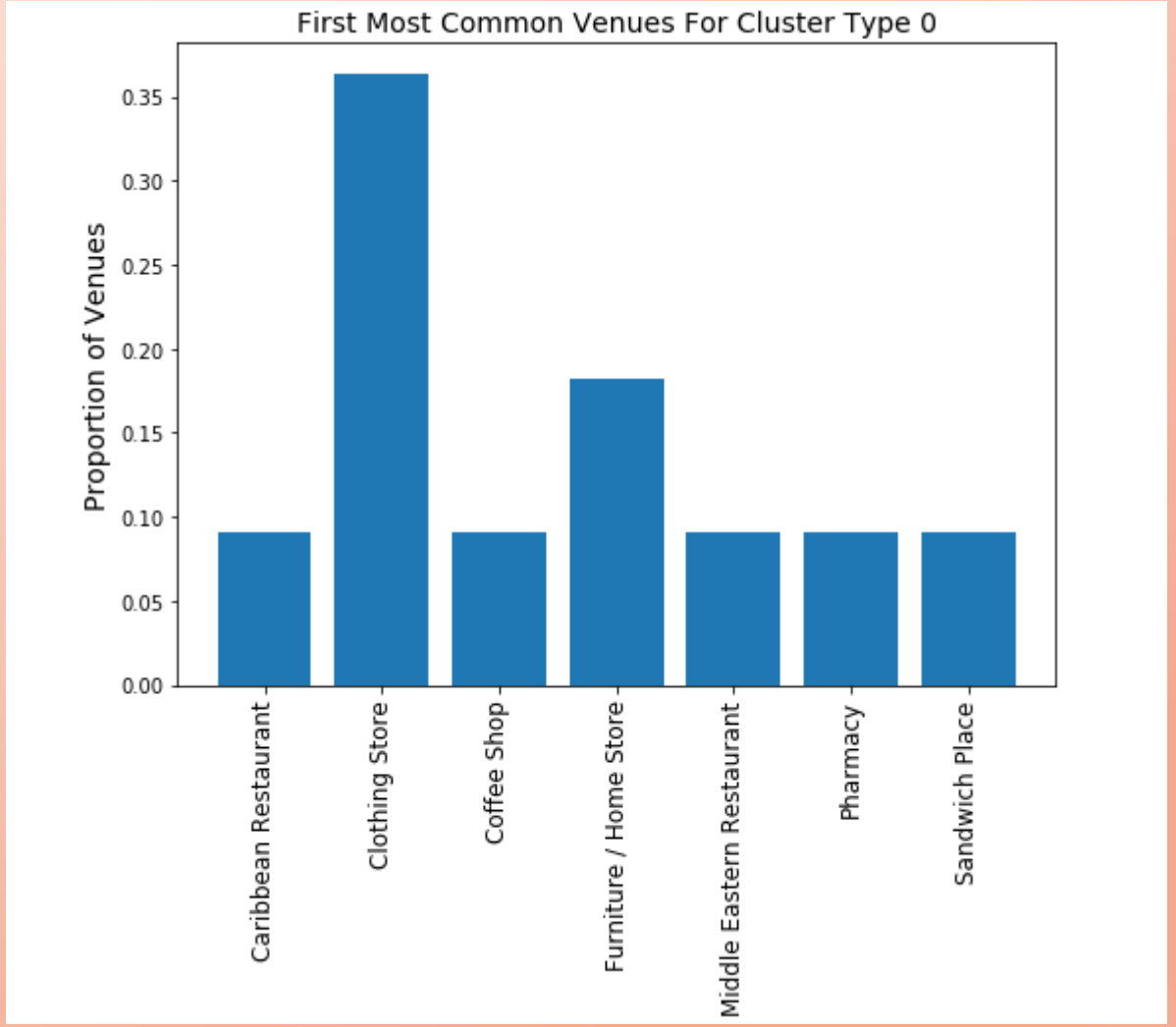
A model with nine clusters seemed to be the best model with the smallest Inertia value amongst the limit set (2-9 clusters)



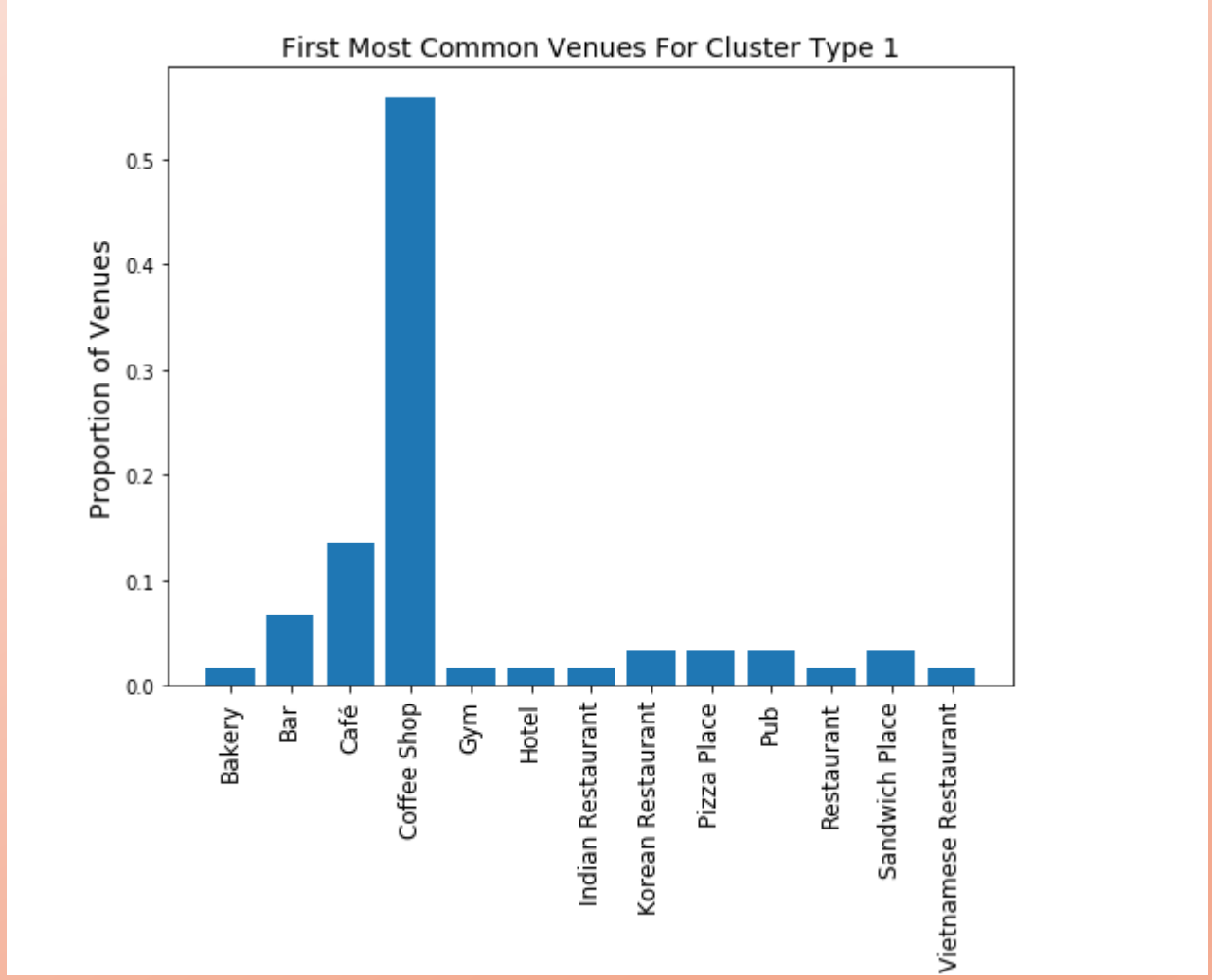
Toronto and Queens both had neighborhoods that shared similar neighborhood structures in 6 out of nine clusters



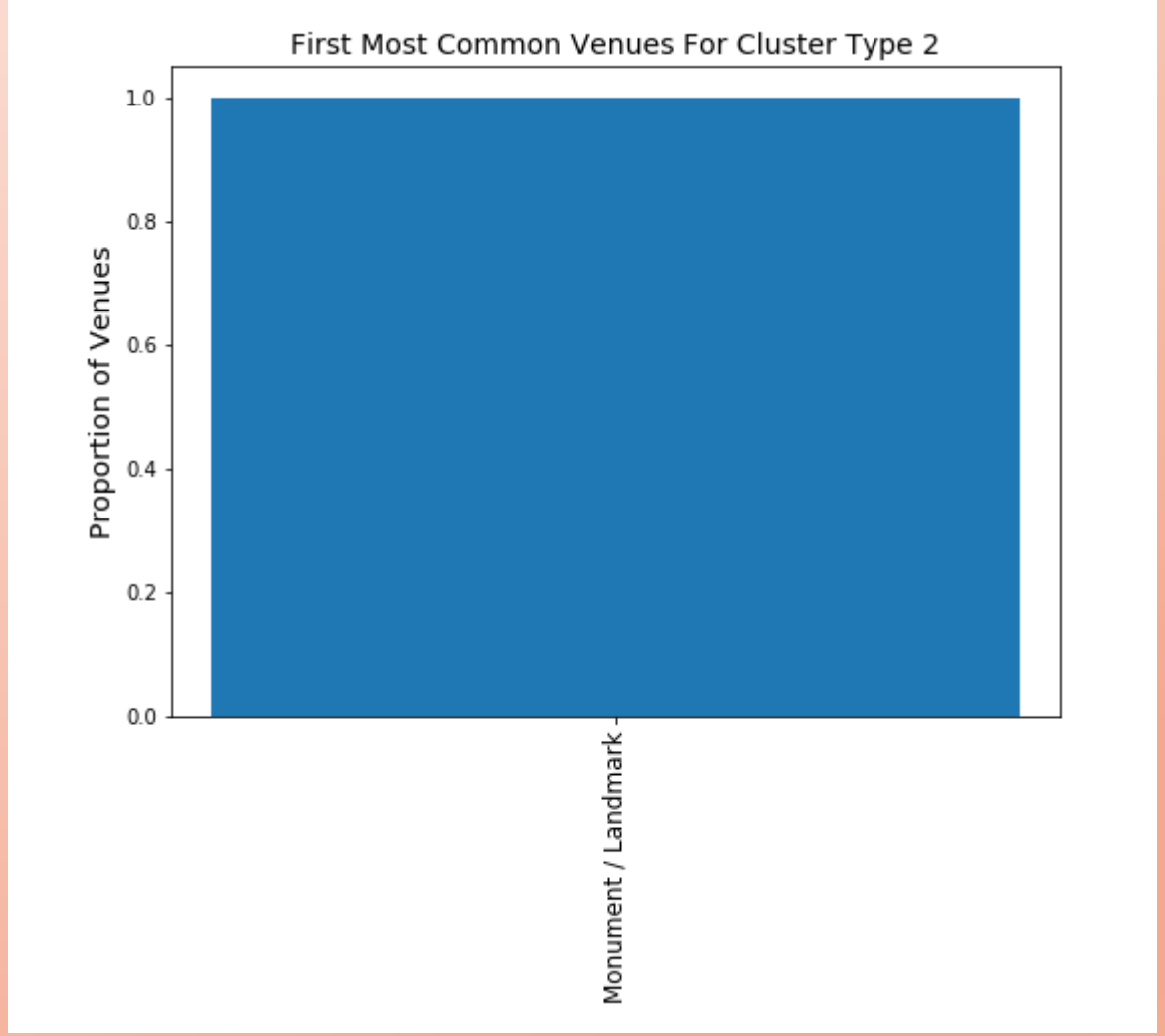
Cluster Type 0



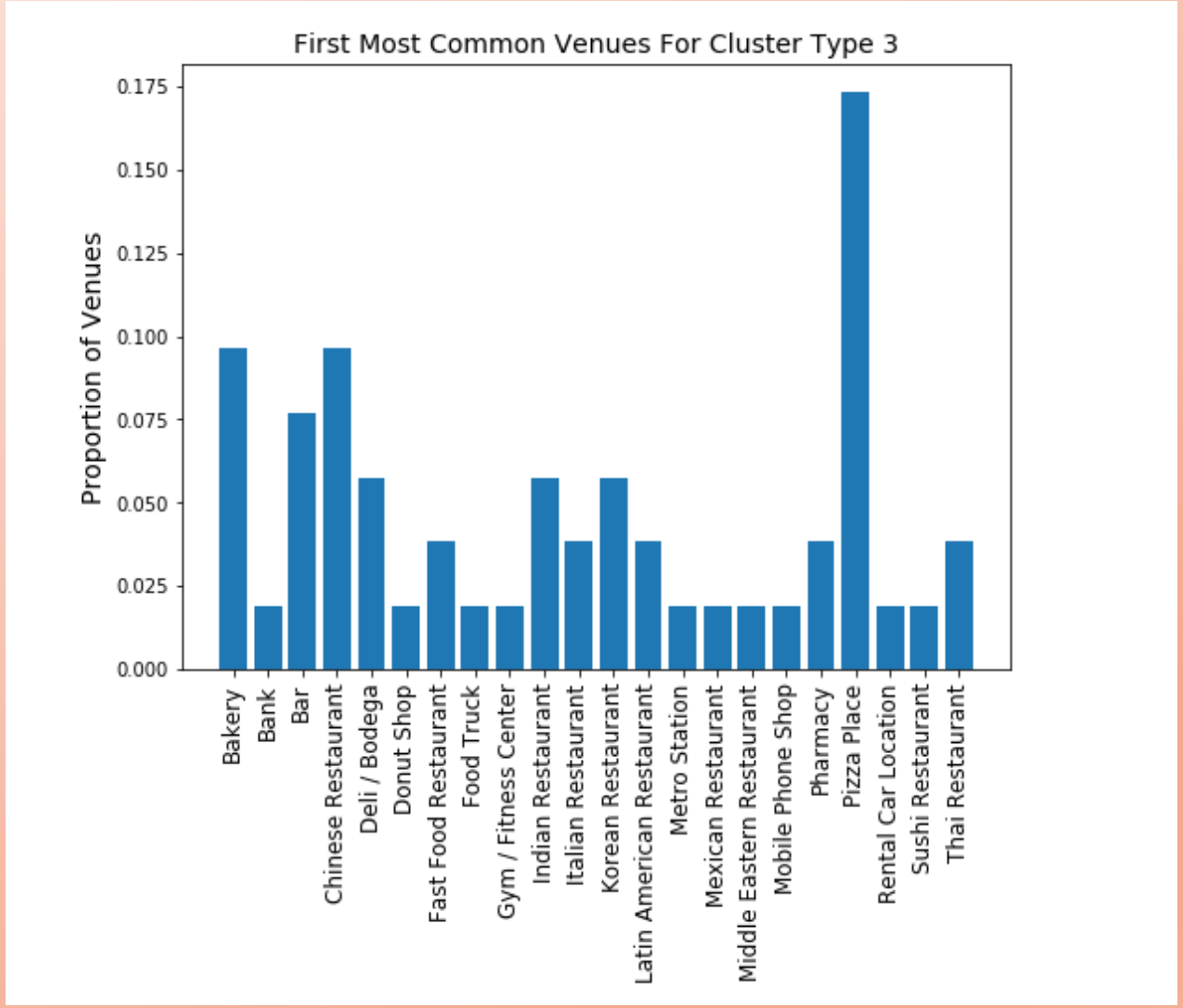
Cluster Type 1



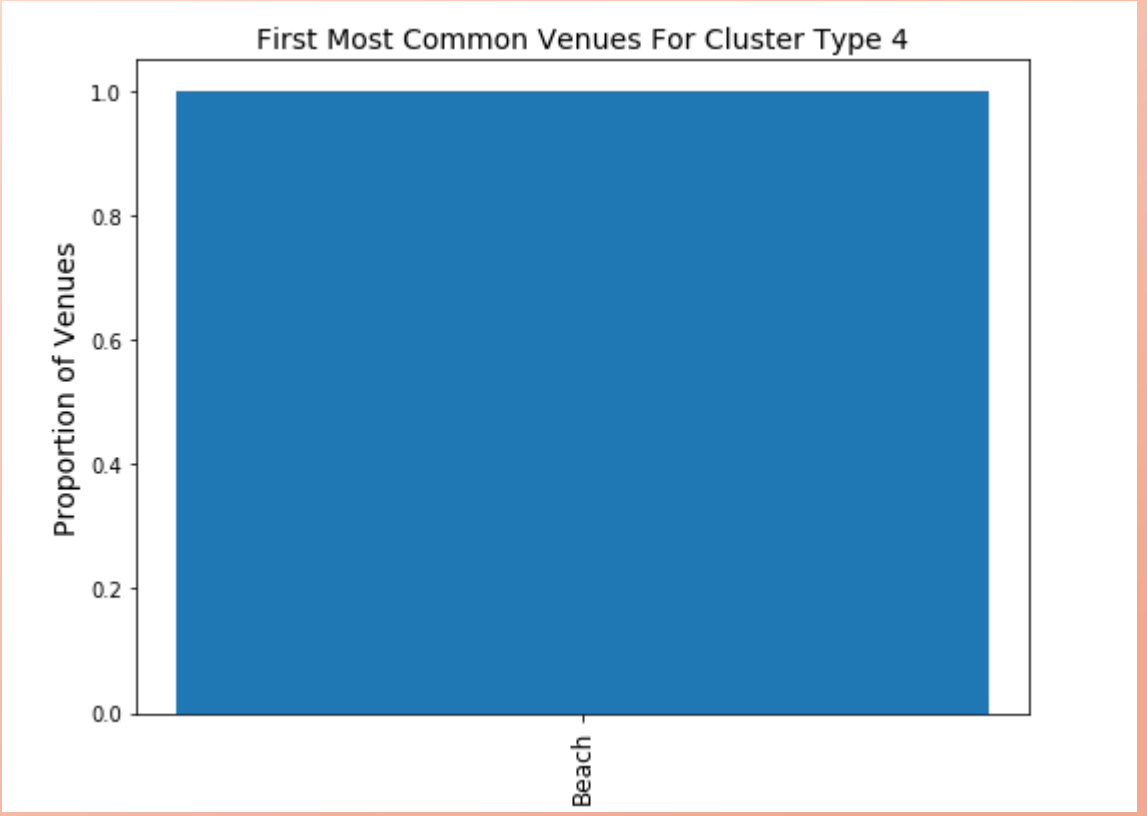
Cluster Type 2



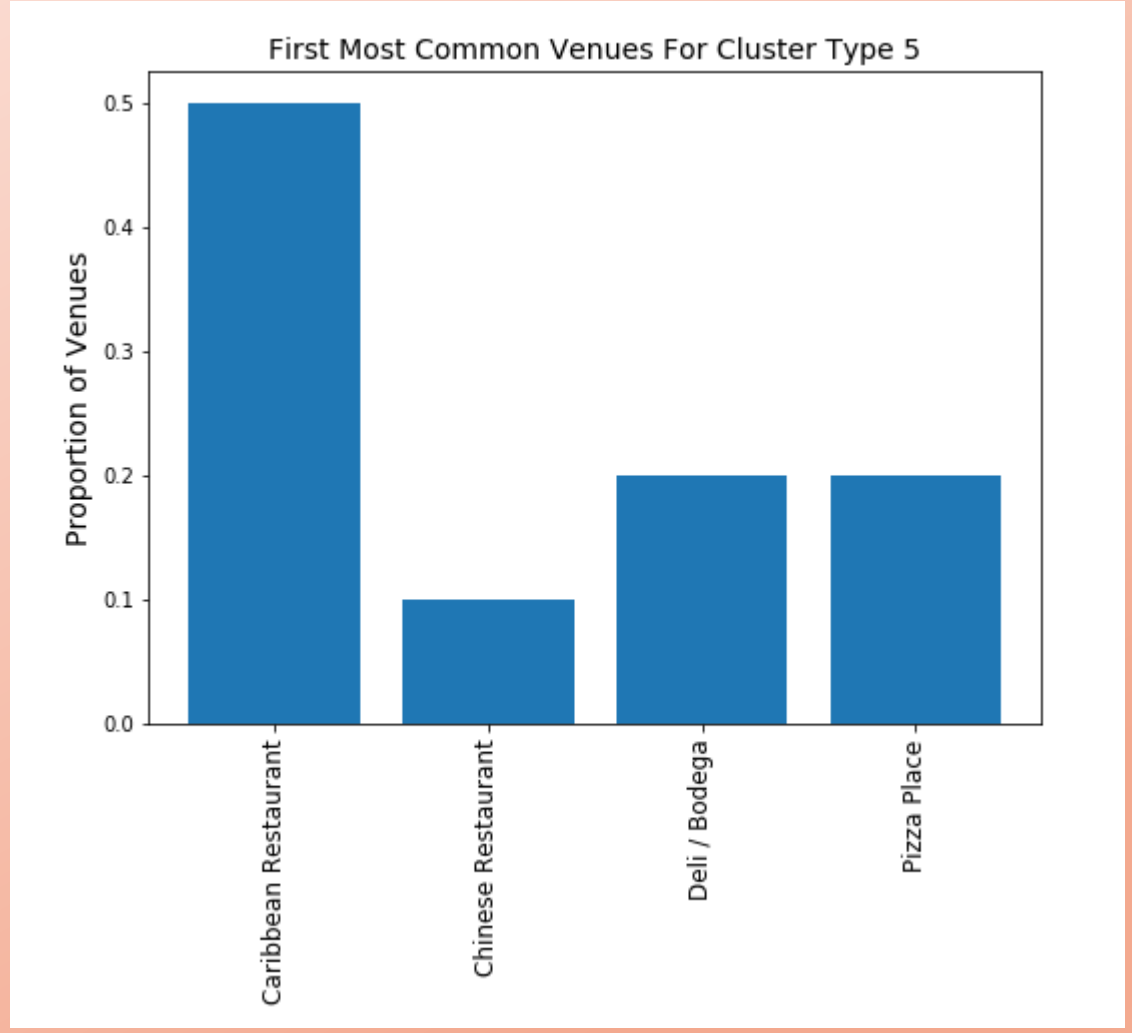
Cluster Type 3



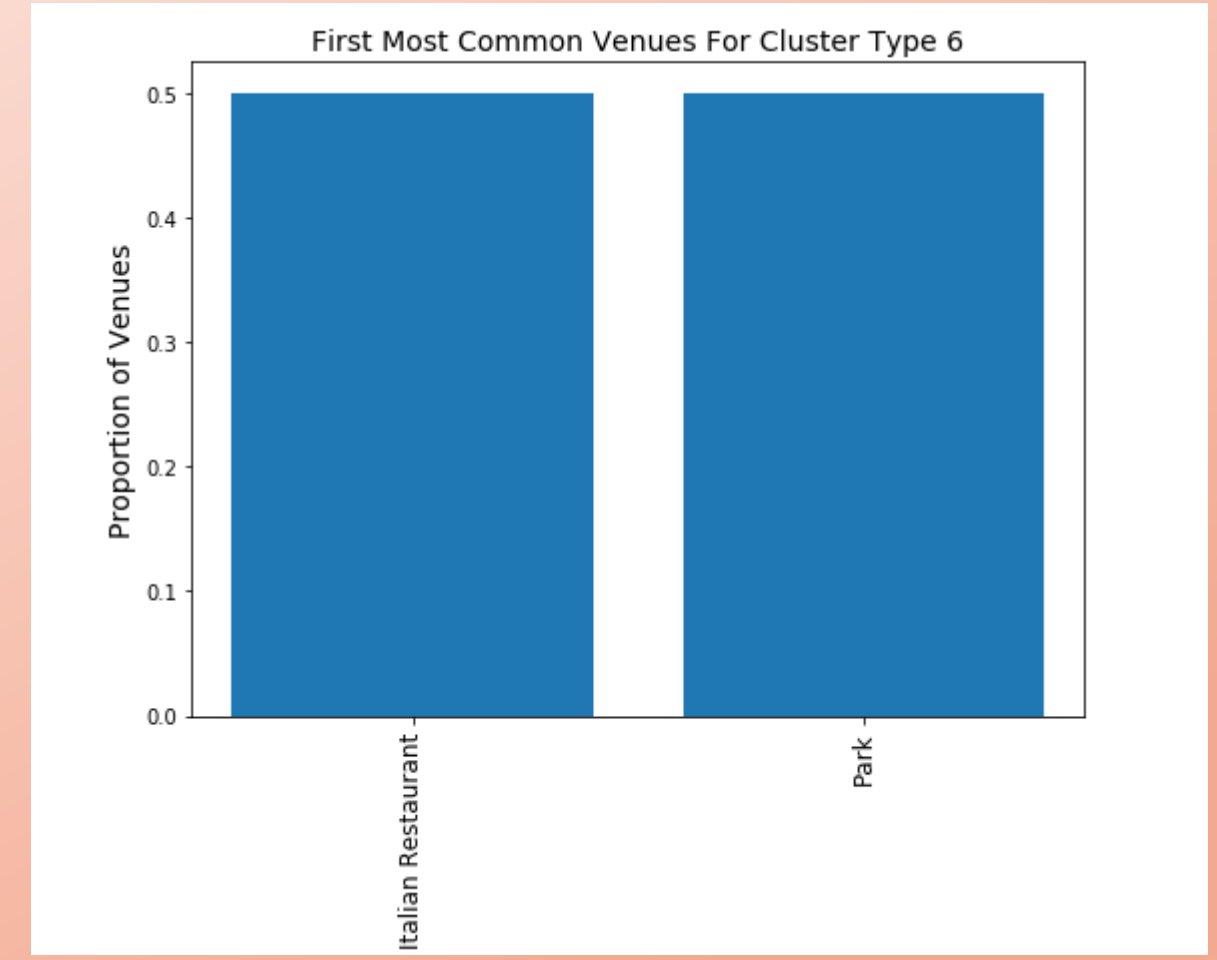
Cluster Type 4



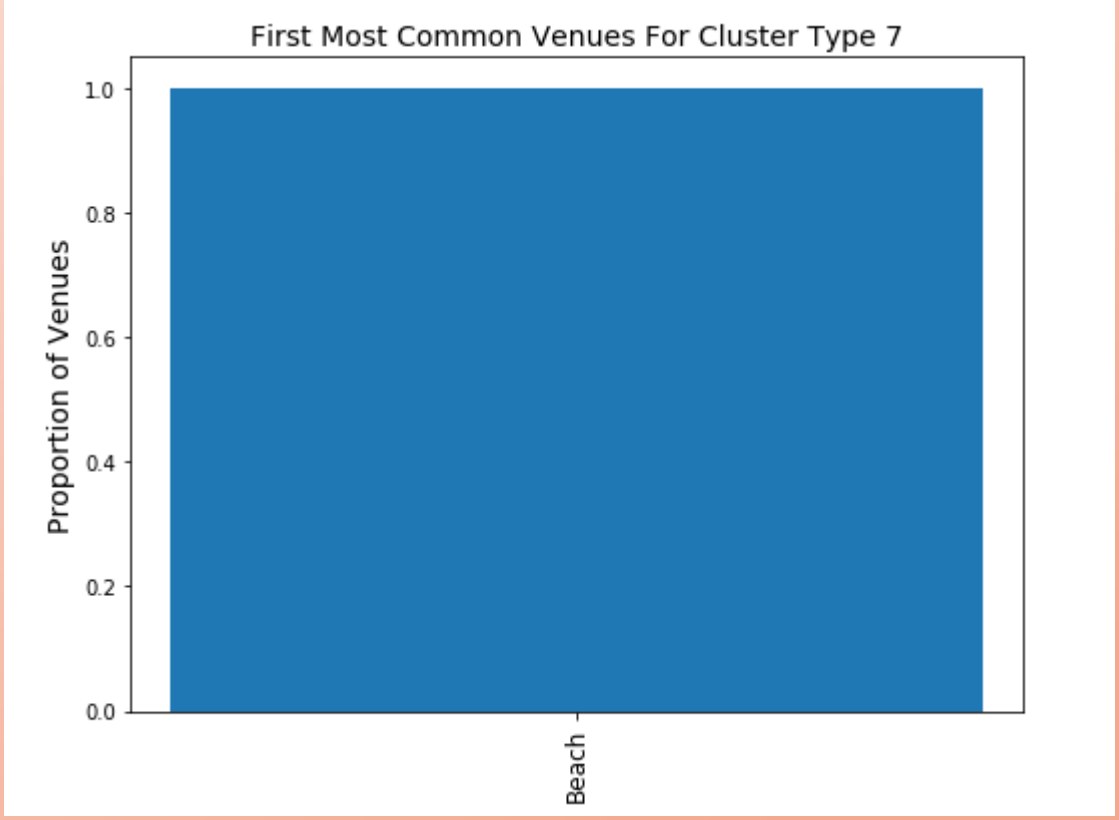
Cluster Type 5



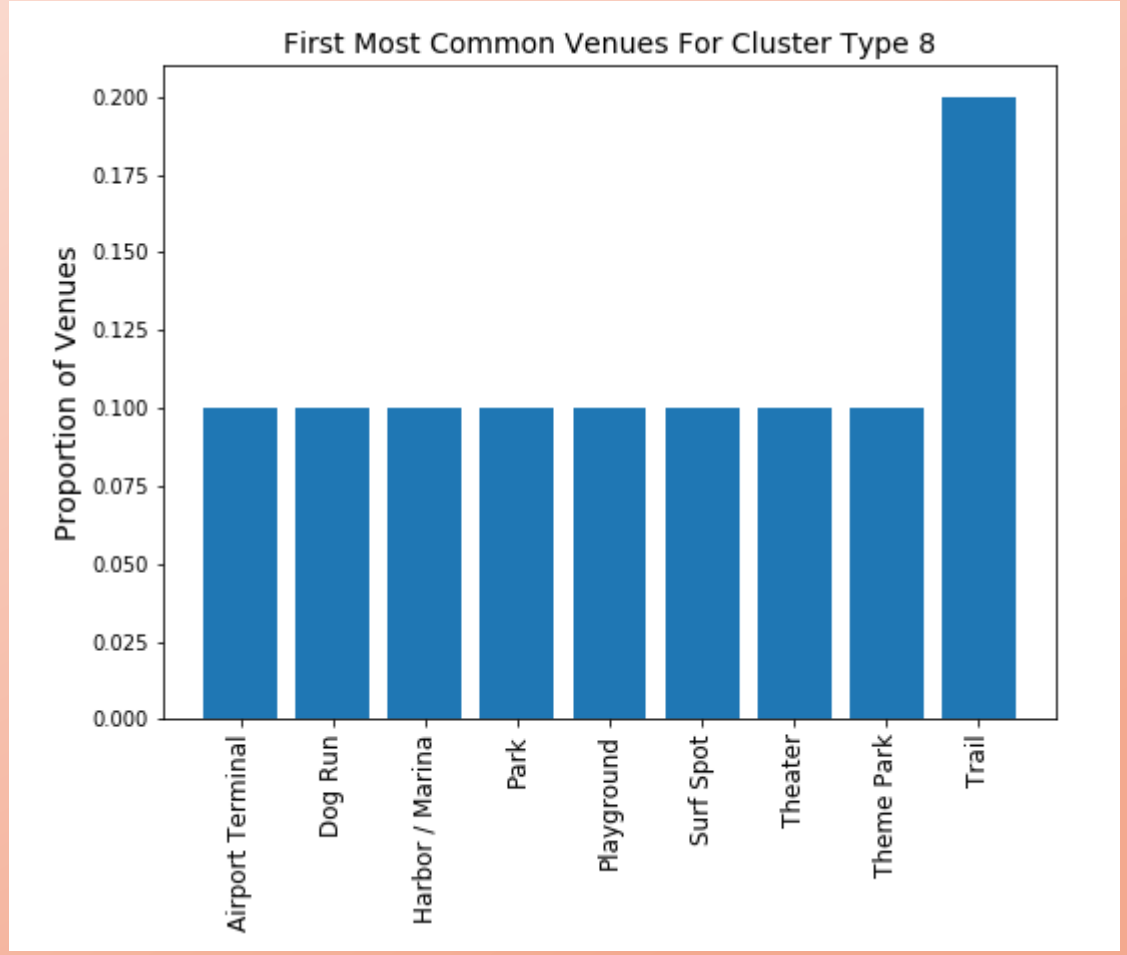
Cluster Type 6



Cluster Type 7



Cluster Type 8



Conclusion

- This project showed that clustering models and location data can be used to determine places that share similar venue categories. This will be useful for people who are moving from one place to another and wanting either something similar to or different from their previous locations.

Future Directions

- In the future, it will be interesting to include other information on these neighborhoods such as climate, socioeconomic status, crime rates, prices of housing etc. This will enrich the model and discover interesting similarities between the neighborhoods.