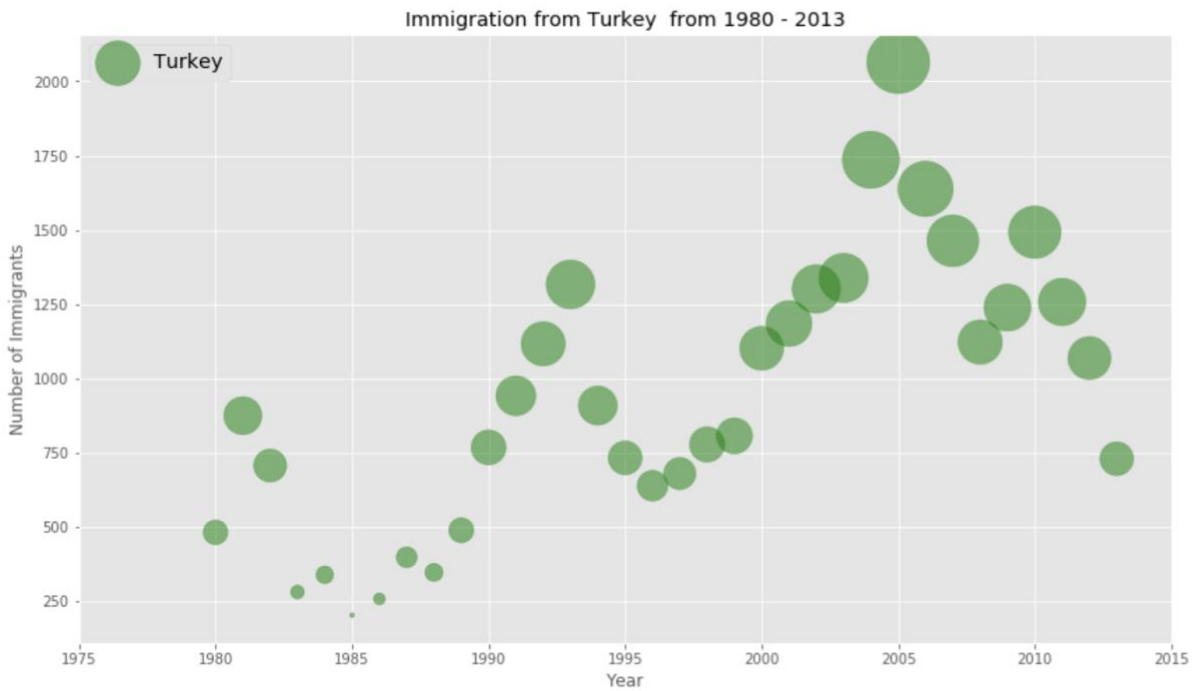
Introduction

Background and problem

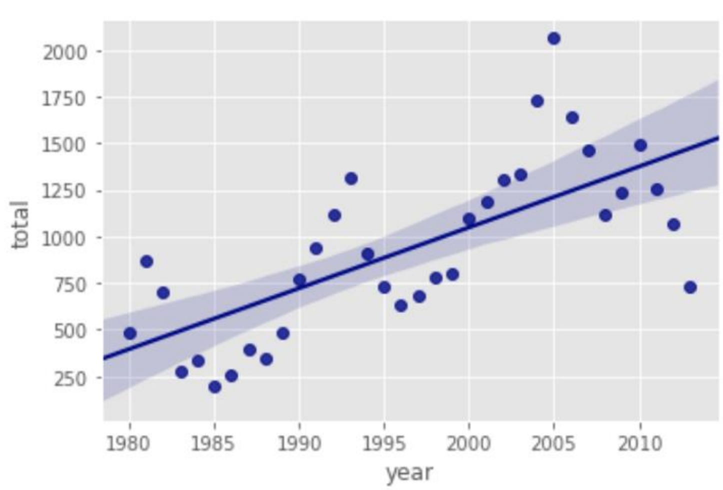
People have moved from their countries for centuries, for all sorts of reasons such as escaping conflict, violence, poverty, or past or future persecution based on race, religion, nationality, and/or membership in a particular social group or political opinion. Nowadays, people also migrate to developed countries to seek superior healthcare, education, jobs and business opportunities. In 2013, the percentage of international migrants worldwide increased by 33% with 59% of migrants targeting developed regions (https://en.wikipedia.org/wiki/Human_migra
-tion). Many people go through difficult period in their adaptation which may be related to having lived under very different conditions and circumstances in their countries of origin. I firmly believe that similarities between the neighborhoods of people's home and host countries might help them better adapt to the host country. Therefore, for this project I have picked one of the top migration corridors: Turkey – Canada and clustered neighborhoods by venues in their economic centers: Istanbul and Toronto. Clustering of neighborhoods might help migrants to
find similar neighborhoods in Toronto and accelerate the adaptation process in their host country.

Migration facts

According to Canadian Census data, in the period of 1980 – 2013 more than 30.000 Turkish nationals migrated to Canada for different reasons. The majority of Turkish Canadians live in Ontario, mostly Toronto. The graph below shows the trend of immigration from Turkey to Canada in the period between 1980 and 2013:



It is seen from the graph that immigration from Turkey to Canada peaked in 2005 with more that 2.000 migrants. The potential reason of decrease in immigration to Canada after 2005 might be that Turkey's EU membership negotiations officially launched after intense bargaining. The process of softening EU's immigration policies for Turkish citizens might cause them to prefer EU countries over Canada for migration. Nevertheless, the overall trend of migration from Turkey to Canada is upward:



Data

Data of this project come from different sources. Migration data was retrieved from Statistics Canada (https://www12.statcan.gc.ca/census-recensement/index-eng.cfm?MM=1). After cleaning data and extracting yearly data for Turkey, I used Python "matplotlib" and "seaborn" libraries to obtain visualizations above.

The next task was to obtain data of Toronto's neighborhoods. To do so, I used several libraries to scrape Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) which contained the list of postcodes in Toronto (*i.e. BeautifulSoup4, lxml, html5lib*). Then I downloaded coordinates of each postcodes from *colc.us* and constructed a final data set which looked like this:

	Postcode	Borough	Neighbourhood	Postal Code	Latitude	Longitude
0	M3A	North York	Parkwoods	M3A	43.753259	-79.329656
1	M4A	North York	Victoria Village	M4A	43.725882	-79.315572
2	M5A	Downtown Toronto	Harbourfront	M5A	43.654260	-79.360636
3	M5A	Downtown Toronto	Regent Park	M5A	43.654260	-79.360636
4	M6A	North York	Lawrence Heights	M6A	43.718518	-79.464763
5	M6A	North York	Lawrence Manor	M6A	43.718518	-79.464763
6	M7A	Queen's Park	Queen's Park	M7A	43.662301	-79.389494
7	M9A	Etobicoke	Islington Avenue	M9A	43.667856	-79.532242
8	M1B	Scarborough	Rouge	M1B	43.806686	-79.194353
9	M1B	Scarborough	Malvern	M1B	43.806686	-79.194353
10	M3B	North York	Don Mills North	M3B	43.745906	-79.352188
11	M4B	East York	Woodbine Gardens	M4B	43.706397	-79.309937
12	M4B	East York	Parkview Hill	M4B	43.706397	-79.309937

Next, I needed to obtain the list of Istanbul's neighborhoods with latitudes and longitudes. I extracted Istanbul's 39 neighborhoods with coordinates from a web page (https://worldpostalcode.com/turkey/istanbul/). I couldn't automate this process using Python's scrapping libraries, as every neighborhood was a hyperlink and it was needed to click on it to see its coordinates. Therefore, I manually collected coordinates of Istanbul's 39 neighborhoods:

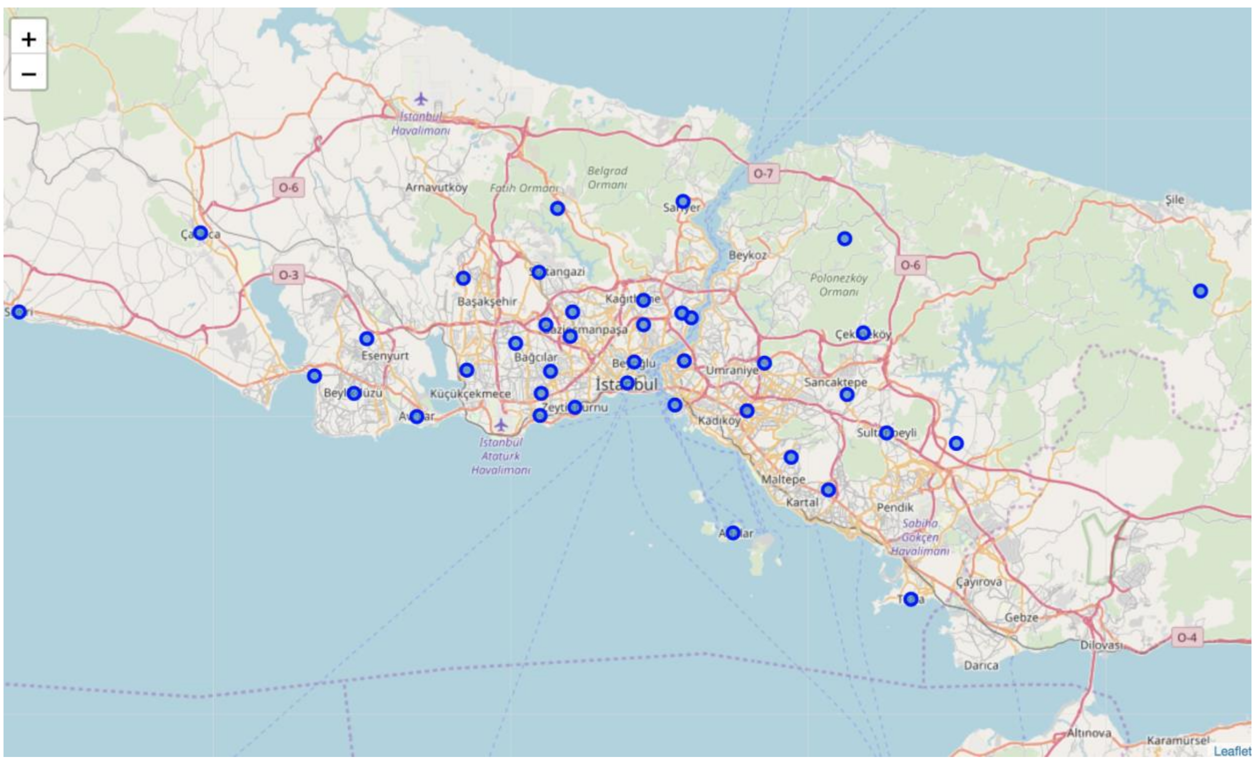
	Neighbourhood	Latitude	Longitude
0	Adalar	40.876259	29.091027
1	Bagcilar	41.045061	28.833649
2	Bayrampasa	41.051248	28.898469
3	Beyoglu	41.028387	28.974045
4	Esenler	41.062002	28.869538
5	Gaziosmanpasa	41.073421	28.901556
6	Kartal	40.915036	29.203717
7	Sancaktepe	40.999438	29.225250
8	Sisli	41.061672	28.984261
9	Umraniye	41.027219	29.127459

Lastly, I extracted data on venues in both cities from Foursquare. I applied two criteria for venue data: venues had to be within 500m radius of the neighborhood and the maximum number of venues around the neighborhood was limited to 100. The final data contained names and coordinates of neighborhoods, and names, coordinates and categories of venues in both cities.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Adalar	40.876259	29.091027	Merit Halki Palace Hotel	40.878802	29.090974	Hotel
1	Adalar	40.876259	29.091027	İnönü Evi Müzesi	40.878251	29.093647	History Museum
2	Adalar	40.876259	29.091027	L'isola Guesthouse	40.877038	29.096136	Bed & Breakfast
3	Adalar	40.876259	29.091027	Aqua Green Beach	40.880498	29.090354	Beach
4	Adalar	40.876259	29.091027	Huseyin Rahmi Gurpinar Muzesi	40.877224	29.092228	Museum
5	Adalar	40.876259	29.091027	Asaf Beach Club	40.879211	29.088317	Surf Spot
6	Adalar	40.876259	29.091027	heybeliada tenis kortu	40.878337	29.091221	Tennis Court
7	Adalar	40.876259	29.091027	heybeliada cicekli orman	40.878291	29.090232	Mountain
8	Adalar	40.876259	29.091027	Heybeliada Bisiklet Turu	40.878426	29.091641	Bike Rental / Bike Share
9	Adalar	40.876259	29.091027	Heybeliada Bayraktepe	40.874992	29.094010	Scenic Lookout

Methodology

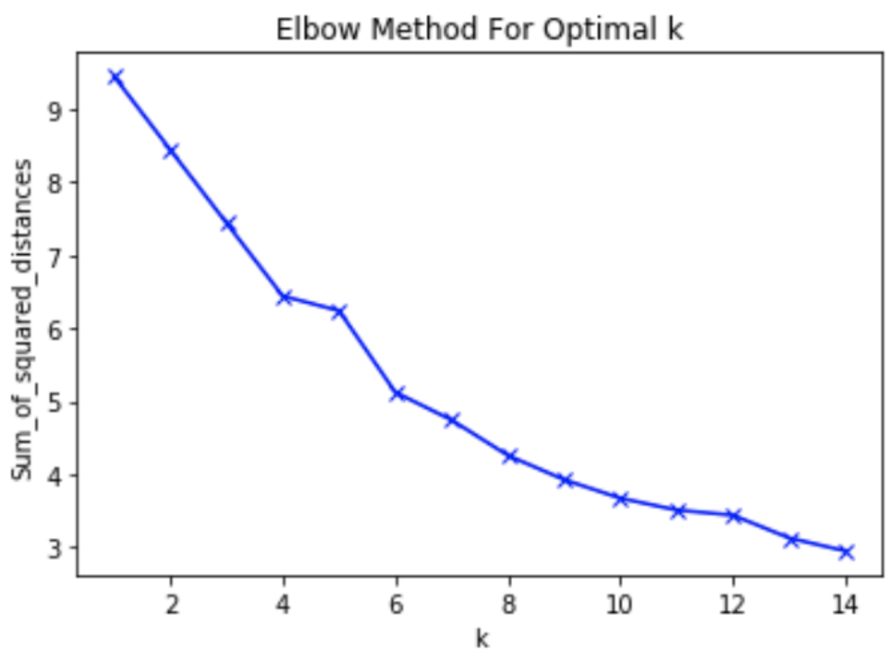
I used folium - map rendering library to plot Istanbul's has 39 neighborhoods. The step was to obtain top venues and sort them by frequency in Istanbul and Toronto. To do so, I first grouped



venues in both cities separately and grouped them by frequency. Then I placed the results into new data frame. Here the problem was that some common categories in one city did not exist in another. For example, Ethiopian Restaurant and Gay Bar categories did not exist in Istanbul. Therefore, I replaced nan values with zero in the final data frame. The last step before starting clustering was to group neighborhoods by frequency of venue categories and obtain top 10 venue categories in each neighborhood.

	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
0	Adalar	Café	Scenic Lookout	Surf Spot	Beach	Bed & Breakfast	Museum	Tennis Court	Harbor / Marina	University	Bike Rental / Bike Share
1	Arnavutkoy	Seafood Restaurant	Café	Restaurant	Lounge	Cocktail Bar	Boat or Ferry	Lighthouse	Music Venue	Coffee Shop	Pastry Shop
2	Atasehir	Café	Bistro	Restaurant	Gym Pool	Doner Restaurant	Bakery	Coffee Shop	Steakhouse	Kebab Restaurant	Basketball Stadium
3	Avcilar	Café	Gym / Fitness Center	Dessert Shop	Breakfast Spot	Bar	Restaurant	Coffee Shop	Turkish Restaurant	Gym	Art Gallery
4	Bagcilar	Café	Department Store	Gym	Wings Joint	Turkish Restaurant	Steakhouse	BBQ Joint	Dessert Shop	Auto Garage	Stationery Store

I used two methodologies to cluster neighborhoods: K – means and hierarchical. For k -means I visualized accuracy of the model using different K's and decided 6 gave the best accuracy to the model. As it is seen from the graph below, an elbow appears when k equals 6.

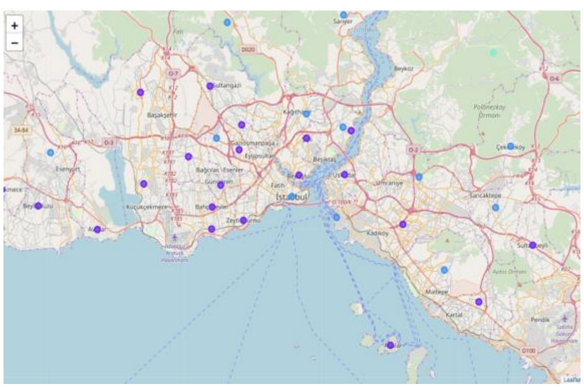


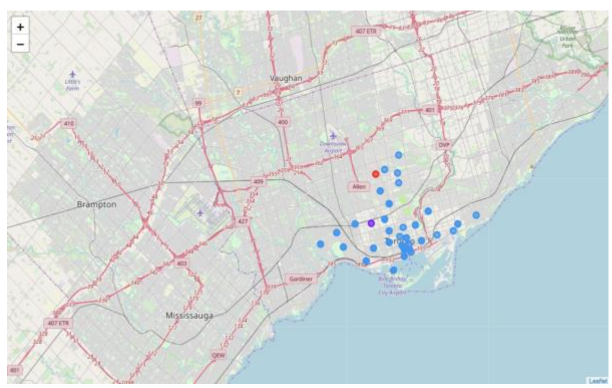
Hierarchical clustering also split neighborhoods into 6 clusters. The advantage of hierarchical clustering is that it does not require any input for the number of clusters, instead algorithm choose the number of clusters based data. However, I needed to limit the maximum number of clusters to avoid much detailed clustering.

After clustering all neighborhoods in Istanbul and Toronto, I assigned all cluster labels to the original data frame and ended up a final data frame which contained names, coordinates, cluster label and top 10 common venue categories of neighborhoods.

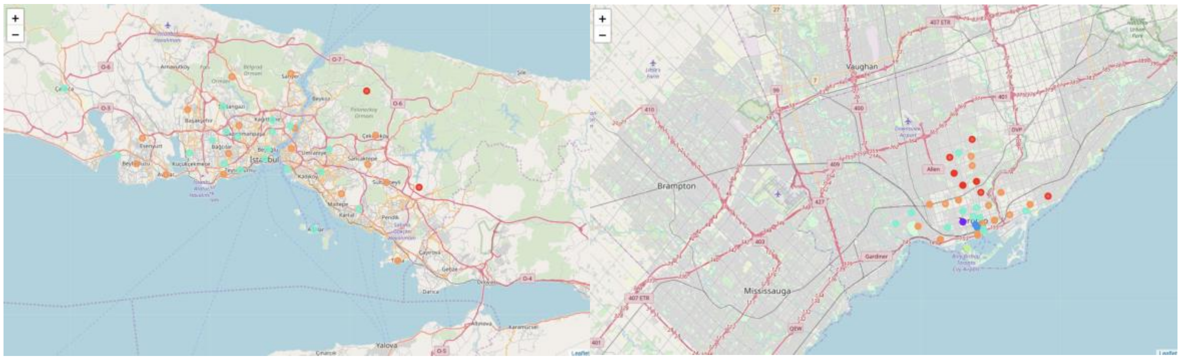
Results

The maps below which were visualized using k – means clustering methodology show similar neighborhoods in the same color. Some colors are not present in both maps: red, light green and orange. For example, red color does not exist in Istanbul's map which means that the red neighborhood in Toronto is not comparable to any neighborhood in Istanbul. The red neighborhood is so close to the airport, therefore there might be some airport related services that are not present in Istanbul's venue categories. Moreover, two light green colored neighborhoods are only in Toronto, one orange colored neighborhood is only in Istanbul.





Hierarchical clustering methodology provides slightly different results, especially for Toronto's neighborhoods.



As both k – means and hierarchical methodologies had pros and cons, I could not make a choice between two, therefore I reported both results.

Discussion

The results imply that there are some similarities between Istanbul and Toronto based on venue categories in the vicinity of neighborhoods. The purpose of this project was to give some insights to migrants from Istanbul which might help them to find a similar neighborhood in Toronto. In this project, the term of "similarity" was limited, because we had to use Foursquare data. Future research can extend this concept by including some other elements of daily life to cluster neighborhoods. For example, house prices, schools, density, noise information of neighborhoods might produce more accurate data about similarities between neighborhoods of two cities.

Conclusion

Every year thousands of Turkish nationals migrate to Canada, mostly Toronto. Most of them face difficulties to assimilate in terms of socioeconomic attainment, social relations, and cultural beliefs. I argue that if the neighborhood in Toronto is similar to one in Istanbul, these people will be better able to find an overall pattern of intergenerational assimilation.