THE BATTLE OF NEIGHBOURHOODS (WEEK 2)

Russian Language in Toronto

Language distribution in Toronto: Russian case

- Canada is a great place for immigrants and it has the highest immigration rate among developed countries.
- Toronto's linguistic diversity is well-established and Russian language is popular in many neighbourhoods.
- People who don’t know English at all, at first time of adaption in a foreign country tend to choose a neighbourhood where its homeland language strongly presented.

Data Sources

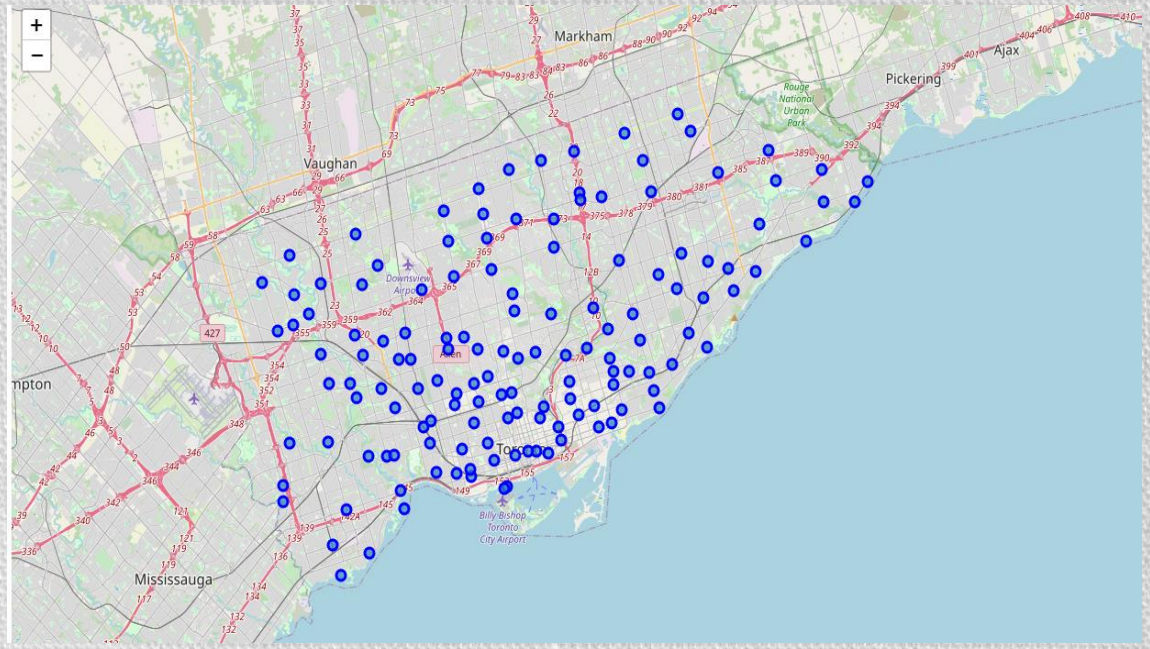
To solve this problem, the following sources of data will be used for analysis:

- Geojson file of Toronto neighborhoods from Github.
- Information about languages in Toronto's neighborhoods at https://www.toronto.com/ and https://open.toronto.ca/. I found there a csv file with each neighbourhoods' profile and filtered out dataset to the numbers of Russian language speakers.
- And, finally Foursquare API provided me with venues list to get the most common ones, including Russian cafes too.

Methodology

- My main data consist of the following information: Neighbourhood, Number of Russian speaking people, Latitude and Longitude.

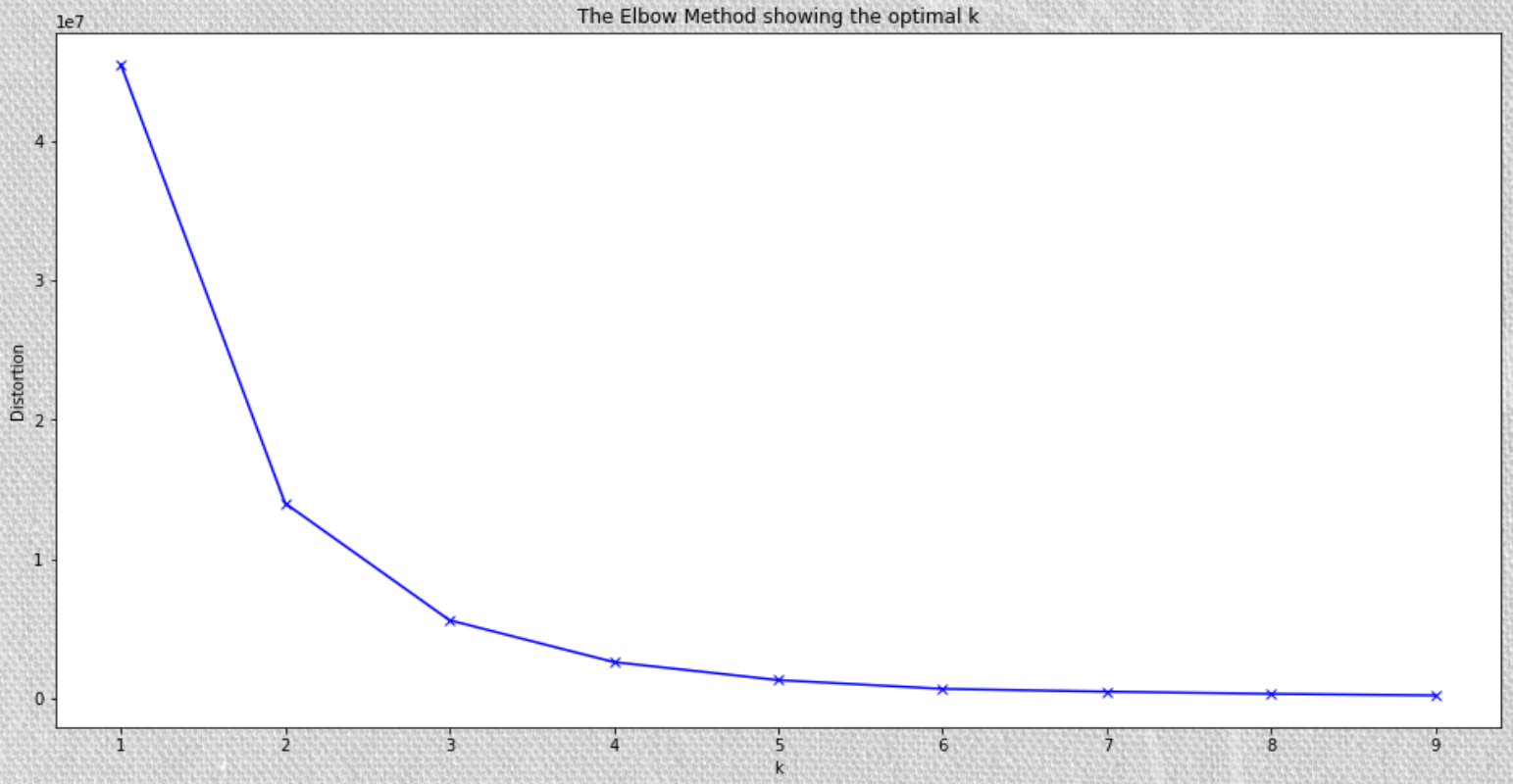
	Neighborhood	Number	Latitude	Longitude
0	Westminster-Branson	5755.0	43.770392	-79.442111
1	Newtonbrook West	2645.0	43.781663	-79.415981
2	Bathurst Manor	1545.0	43.755480	-79.438390
3	Willowdale East	1220.0	43.766694	-79.388044
4	Waterfront Communities-The Island	1150.0	43.635298	-79.394945



- On the left you can see a map of Toronto city with its neighbourhoods (overall there are 140 of them) superimposed on it.

Clustering – finding optimal K

- **K-means clustering** help us to group data into similar ones (clusters) and dissimilar them from the other ones.
- To define optimal K, popular technique called **the Elbow method** was used, which runs model to find the optimal k (the point of inflection on the curve) from given range. In our case it is 3.

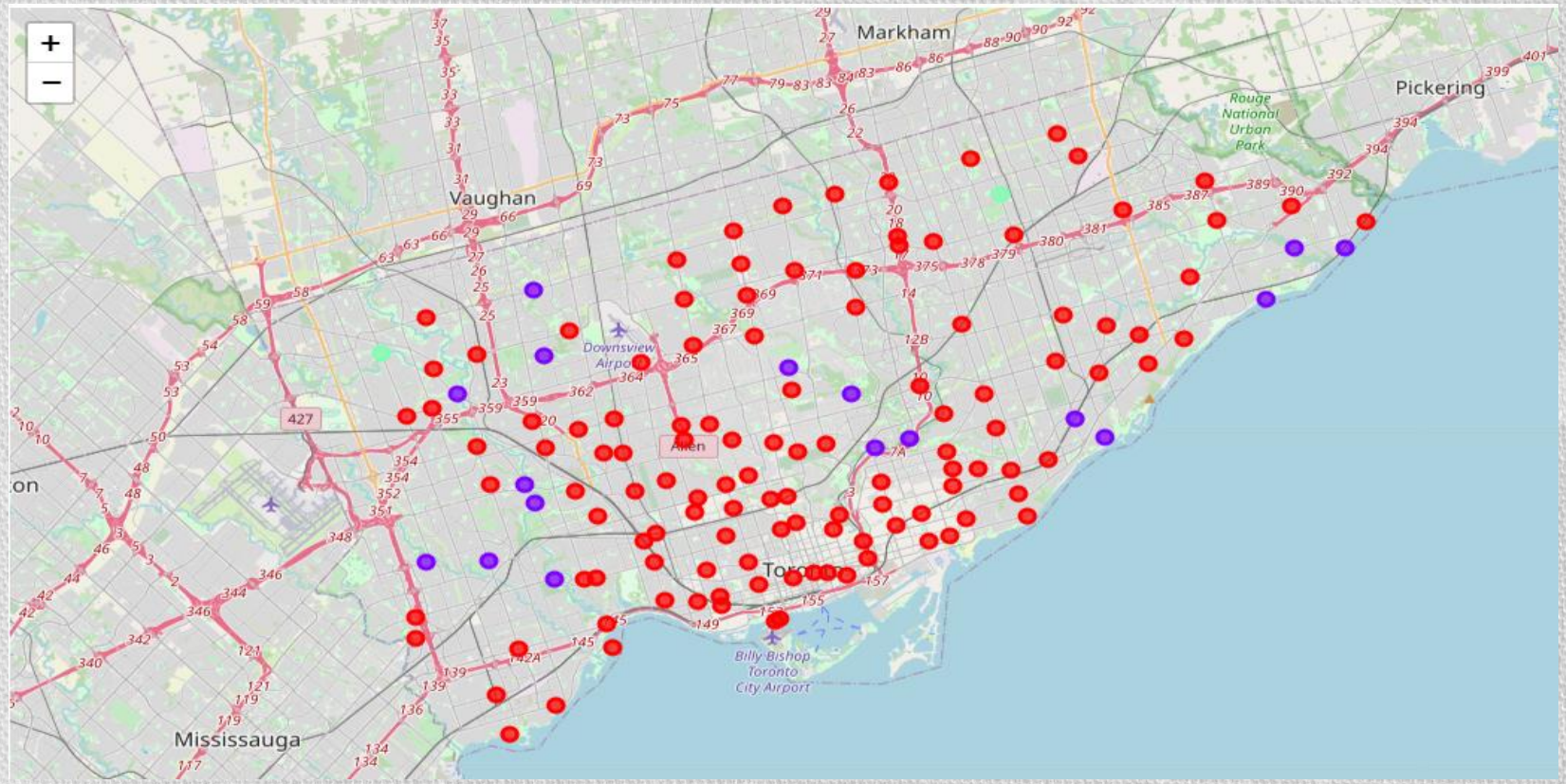


Clustering - Results

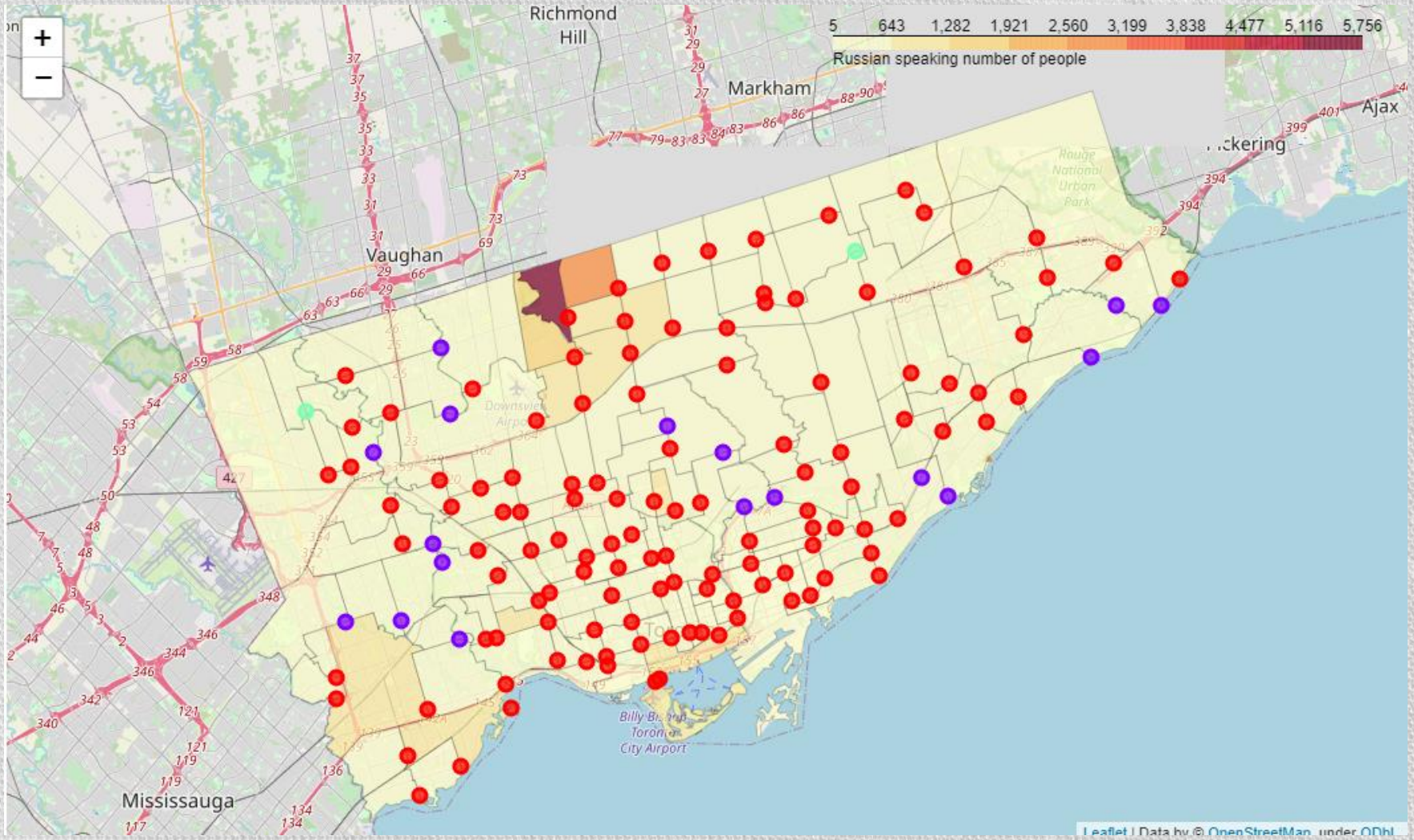
Examining each cluster shows the following results:

- **Cluster 0** – popular places: coffee shops, café and Italian cuisine.
- **Cluster 1** – popular places: primarily park zones.
- **Cluster 2** – popular places: small retail shops, women's store and farm.

Applying K-means algorithm provided with such clustered map of Toronto below.



Russian speaking people mostly live in the northern part of the city (North York borough)



Conclusion

- **Cluster 0** (especially North York neighbourhoods) will be optimal one for immigrants for whom Russian is mother tongue or who speak only Russian. At least for the adaptation period.
- Ideally, there is a room to improve on this problem considering other factors. This research takes those factors all else equal concentrating only on language preferences.
- Finally, the language distribution map would be useful for those who target certain groups as their customer audience. For example, I didn’t manage to find Russian restaurant in North York (the top Russian speaking area) and this will be a good food for thought to open it there.