Coursera Capstone Project : Applied Data Science

Habib jarboui

Mathematician, France

jarbouihabib@yahoo.fr

Introduction/Business Problem

Data Neighbourhoods

Methodology

Discussion

Conclusion

Introduction

- This project is about finding the best neighbourhoods in the city of Toronto to open a new restaurant of a specific type (for example : chinese, or italian restaurant). This project would interest anyone which wants to open a new restaurant in the city of Toronto, and seeks the best neighbourhoods where the habitants would likely eat in this kind of restaurant, and where the competition is limited (e.g. there is a a reasonable number of existing restaurants of the same type in the neighbourhood).

- When we look for the best place to open a new restaurant in a city like Toronto, we have to gauge people's taste in each neighbourhood of the city. We will a know in what neighbourhood of the city people will likely come and spend money in our new restaurant.
- A good way to gauge people's taste in a specific area is to look into the demographic data of this area. For example, areas with a majority of chinese people would be good for chinese restaurants, and areas with a majority of italian people would be good for opening an italian restaurant, etc.
- With this kind of demographic data associated with different neighbourhoods of Toronto, we can cluster neighbourhoods by demographic data. Thus, we will be able to distinguish the areas where a lot of chinese people live, the areas where a lot of italian people live, and so on, based on the clustering.

Find the best neighbourhoods within a cluster to open the restaurant

- Once the neighbourhoods have been categorised into clusters, and we've got a list of neighbourhoods where people living there would likely want to eat in the restaurant we want to open, we need to find out in which neighbourhoods there is less competition. It means that we have to find out what neighbourhoods contain the lowest number of existing restaurants of the same type as the one we want to open.
- In order to count the number of existing restaurants of the same type in a neighbourhood, we perform a FoursquareAPI explore query. Like that, we obtain the list of venues of each neighbourhood, and we can count the number of restaurants of each type.

Data

- Demographic data from the City of Toronto's open data
- The list of neighbourhoods, and the demographic data associated to each neighbourhood, has been made available by the city of Toronto herehttps://www.toronto.ca/ext/open_data/catalog/data_set_files/2016_neighbourhood_profiles.csv .

- The Toronto demographic dataset contains multiple features such as :
- Citizenship
- Ethnic origin
- Income
- Languages / Mother tongue
- Marital status
- Neighbourhood information
- Work activity
- Etc

For this project, we use the Ethnic origin and the Neighbourhood information for each neighbourhood, in order to cluster the neighbourhoods of Toronto.

- Using the data, we can see :
  - We have the name of each neighbourhood in each column name (starting at position 6)
  - We have the name of each ethnic origin in the Characteristic column
  - The number of people living in each neighbourhood, associated to each ethnic origin name.

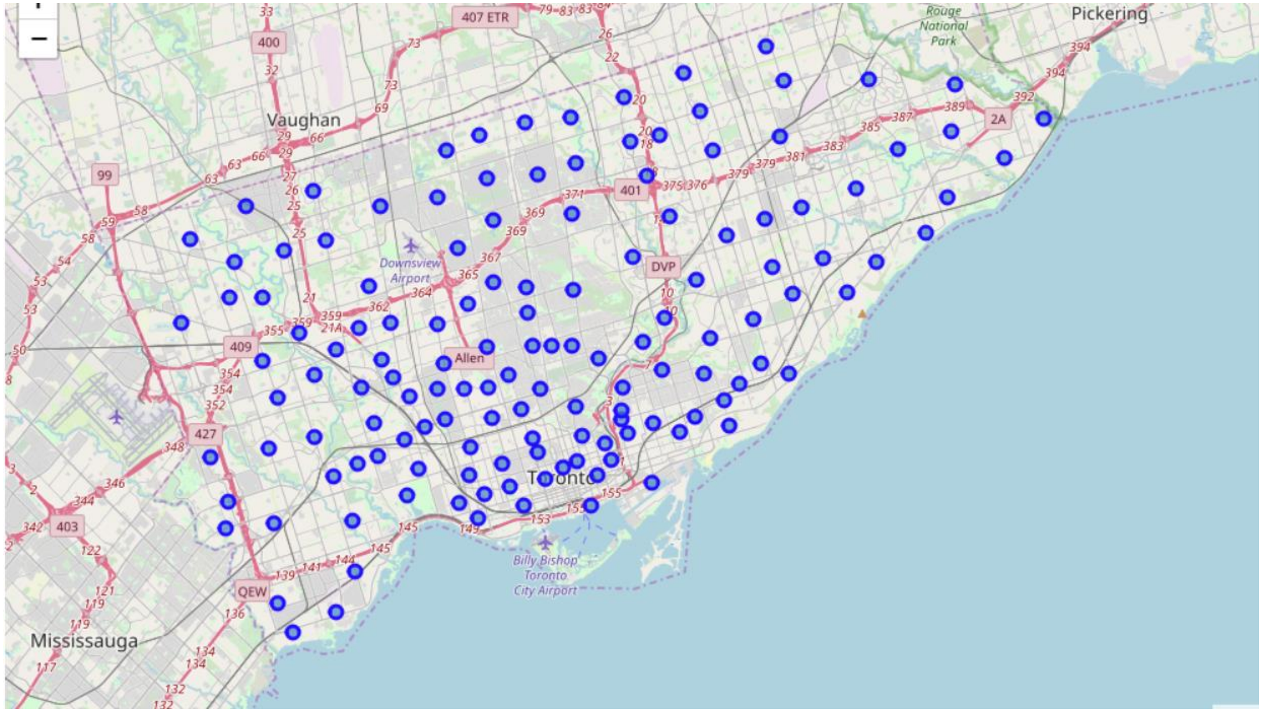
3. Methodology

- As we previously saw, we use the following datasets :
  - A list of general information about the neighbourhoods (Neighbourhood name, Number, and coordinates calculated using the Geocoder package)
  - A list of demographic data about the neighbourhoods, with the number of people of each ethnic origin living in these neighbourhoods

3.1. Data analysis

- A good way to start our analysis is to draw each neighbourhood over the map of the city of Toronto, in order to check if the dataset with the list of neighbourhoods is complete and covers the whole city. For that, we need each neighbourhood' coordinates.
- As we saw, the neighbourhoods' coordinates are not available in the Neighbourhood information data dataset. So we are going to retrieve them using the Geocoder package. We then store each neighbourhood's coordinates into a dataframe, like this :

CDN	City_Area	Latitude	Longitude
129	Agincourt North	43.80930	-79.26707
128	Agincourt South-Malvern West	43.78735	-79.26941
20	Alderwood	43.60496	-79.54116
95	Annex	43.66936	-79.40280
42	Banbury-Don Mills	43.74041	-79.34852

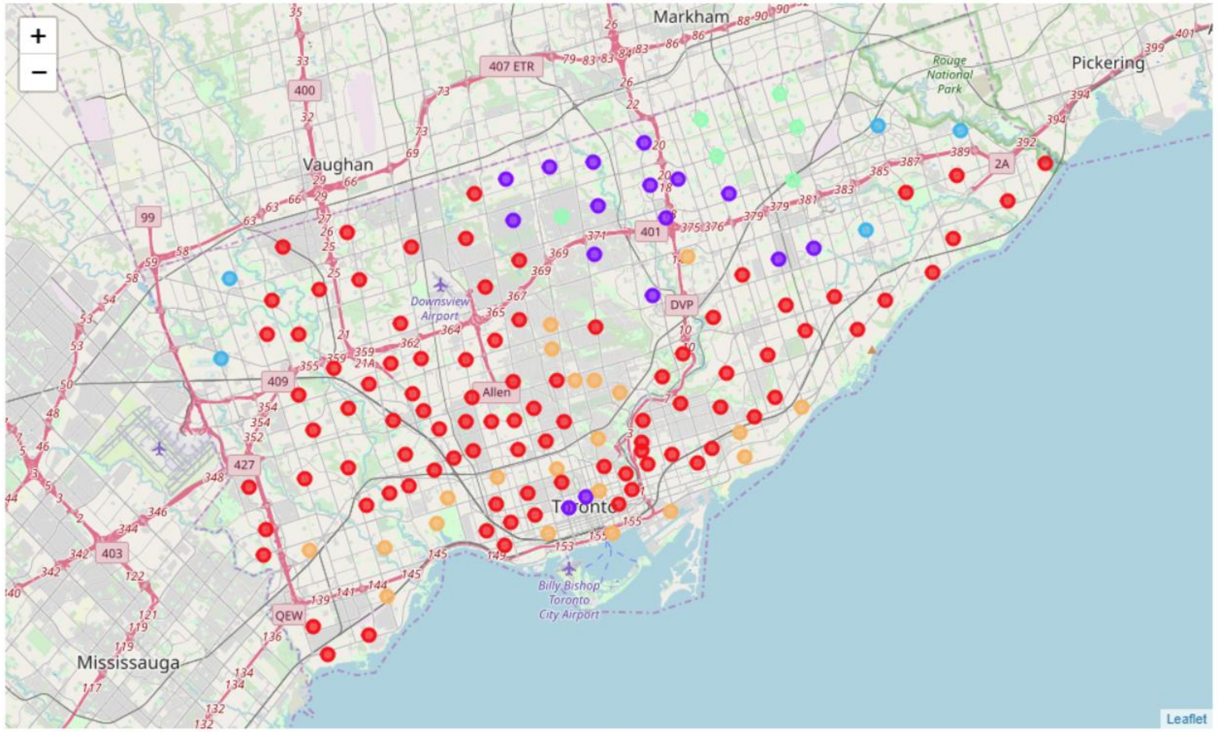


The map of the city is displayed using the Folium package. On this map, we draw a blue circle for each neighbourhood, using the neighbourhoods' coordinates. It is a good way to visualise the position of each neighbourhood in our dataset. It also confirms that the different neighbourhoods are well distributed within the city, and that our dataset covers the whole city (no missing neighbourhood).

3.2. Machine learning algorithm used

- For the clustering, we use a K-Means algorithm. I chose to use a K-Means algorithm, as it is one on the most used algorithm for unsupervised learning and clustering. It is typically used for scenarios like understanding the population demomgraphics, market segmentation, social media trends, anomaly detection, etc... where the clusters are unknown to begin with. It is exactly our scenario, as we want to understand how the neighbourhoods of Toronto are segmented, and the clusters to begin with are unknown in this situation.
  Also, K-Means is one of the simplest clustering algorithm to implement and to run, and is less time consuming than other, more complex algorithms.
- The Elbow method is a method to find the most appropriate number of clusters in a dataset, by running several K-means algorithm and comparing the sum of squared distances of samples to the nearest cluster centre. The more the sum of squared distance is, the further the datapoints are globally from their cluster centre. But we don't have to set K too high, as if K is set to the number of datapoints, then each sample will form its own cluster meaning sum of squared distances equals zero, which is not a good clustering.

3.3. Visualizing the clusters



- We can then visualise the clusters on a Folium map. We display each neighbourhood as a circle on the map, each circle will be coloured according to the cluster they have been categorised into

4. Discussion

We obtain the following results and visualizations

❑ Cluster 0 : European & Canadian (Red colour)

- The Cluster 0 regroups areas higly habited by European and Canadian people.
  We can see English, Italian, Portuguese, French people ...
  Most of them are positioned in almost all the south of Toronto, and in the downtown.

CDN	City_Area	Latitude	Longitude	Cluster Labels	1st Most Common Origin	2nd Most Common Origin	3rd Most Common Origin	4th Most Common Origin	5th Most Common Origin	6th Most Common Origin
20	Alderwood	43.604960	-79.541160	0	English	Canadian	Irish	Scottish	Italian	Polish
57	Broadview North	43.689370	-79.354290	0	English	Irish	Scottish	Greek	Canadian	French
11	Eringate-Centennial-West Deane	43.661910	-79.577380	0	Canadian	English	Italian	Irish	Scottish	Ukrainian
101	Forest Hill South	43.694310	-79.416100	0	Polish	Canadian	Russian	English	Scottish	Irish

Cluster 1 : Asian (Purple colour)

The Cluster 1 regroups areas higly habited Chinese people, and people from others countries in Asia.
We can see that most of them are positioned at the north of Toronto.

CDN	City_Area	Latitude	Longitude	Cluster Labels	1st Most Common Origin	2nd Most Common Origin	3rd Most Common Origin	4th Most Common Origin	5th Most Common Origin	6th Most Common Origin
127	Bendale	43.75963	-79.25739	1	Chinese	East Indian	Filipino	Canadian	English	Scottish
47	Don Valley Village	43.78552	-79.35017	1	Chinese	Filipino	East Indian	Iranian	English	Canadian
126	Dorset Park	43.75533	-79.27746	1	Filipino	East Indian	Chinese	Canadian	Sri Lankan	English
53	Henry Farm	43.77230	-79.34087	1	Chinese	East Indian	Filipino	Iranian	Canadian	English
48	Hillcrest Village	43.80303	-79.35346	1	Chinese	East Indian	Canadian	Iranian	Korean	English

Cluster 2 : Indian (Dark green colour)
The Cluster 2 concentrates areas haghly habited by Indian people.
We can see that these areas are located at the edges of Toronto.

CDN	City_Area	Latitude	Longitude	Cluster Labels	1st Most Common Origin	2nd Most Common Origin	3rd Most Common Origin	4th Most Common Origin	5th Most Common Origin	6th Most Common Origin
132	Malvern	43.80977	-79.22084	2	East Indian	Sri Lankan	Filipino	Chinese	Jamaican	Canadian
2	Mount Olive-Silverstone-Jamestown	43.74721	-79.58826	2	East Indian	Iraqi	Jamaican	Canadian	Somali	Italian
131	Rouge	43.80766	-79.17405	2	East Indian	Sri Lankan	Canadian	Filipino	Jamaican	English
1	West Humber-Clairville	43.71451	-79.59292	2	East Indian	Jamaican	Canadian	Filipino	Italian	Punjabi
137	Woburn	43.76730	-79.22823	2	East Indian	Canadian	Sri Lankan	Chinese	Filipino	English

Cluster 3 : Chinese (Light green colour)
The Cluster 3 also regroups areas higly habited by asian people, the most common ethnic origin is Chinese.
We can see that most of them are positioned at the north east of Toronto, next to the cluster 1. This cluster is highly similar to the cluster 1.

CDN	City_Area	Latitude	Longitude	Cluster Labels	1st Most Common Origin	2nd Most Common Origin	3rd Most Common Origin	4th Most Common Origin	5th Most Common Origin	6th Most Common Origin
129	Agincourt North	43.80930	-79.26707	3	Chinese	Sri Lankan	East Indian	Filipino	Canadian	English
128	Agincourt South-Malvern West	43.78735	-79.26941	3	Chinese	East Indian	Filipino	Sri Lankan	Canadian	English
117	L'Amoreaux	43.79726	-79.31220	3	Chinese	East Indian	Canadian	Sri Lankan	Filipino	English
130	Milliken	43.82280	-79.27694	3	Chinese	Sri Lankan	East Indian	Filipino	Canadian	Tamil
51	Willowdale East	43.77248	-79.40039	3	Chinese	Iranian	Korean	East Indian	English	Canadian

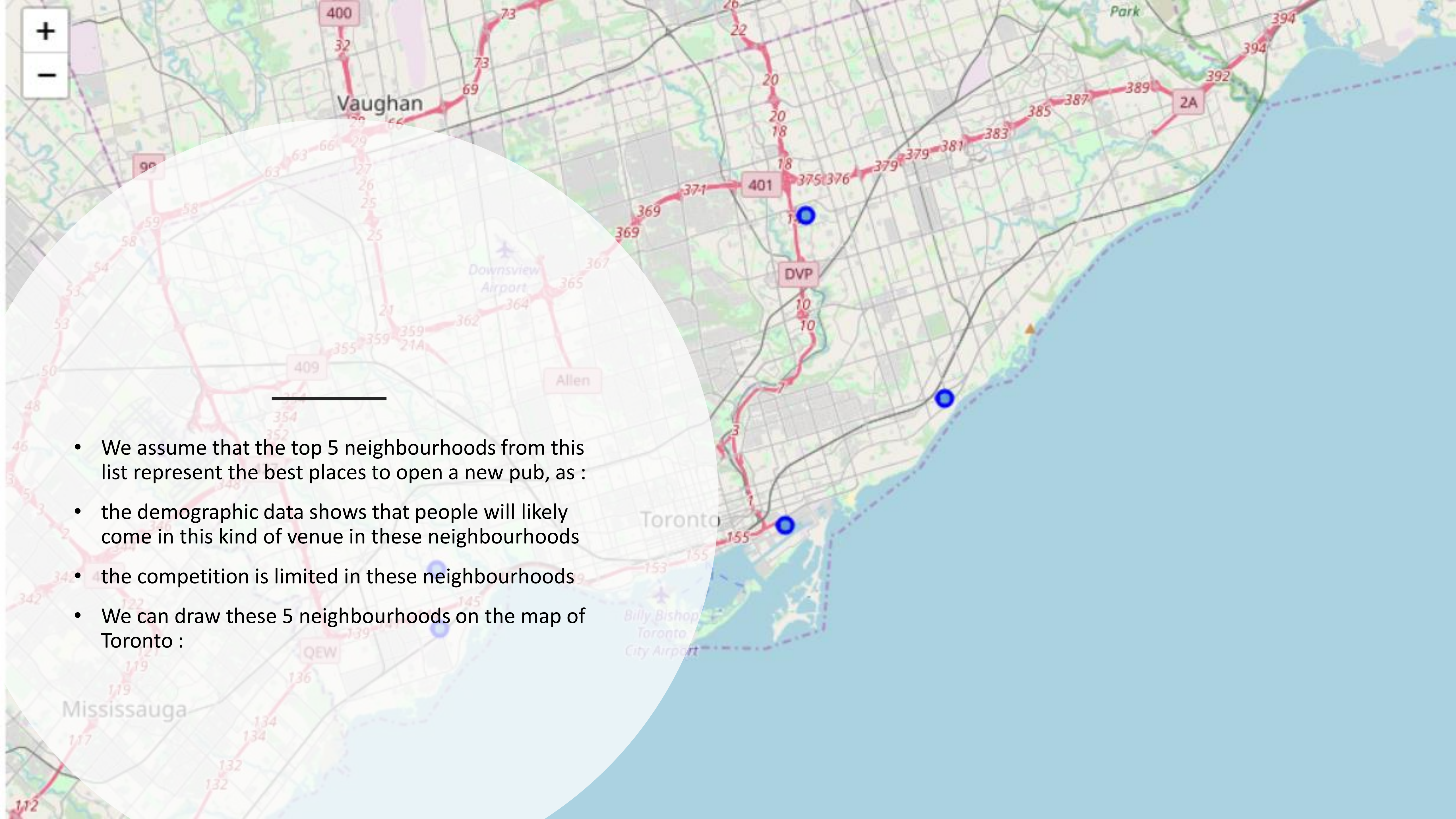
Cluster 4 : Irish, Scottish & English (Yellow colour)
The Cluster 4 regroups areas higly habited by English, Irish, Scottish and Canadian people.
We can also see that there are a lot of people from other european countries as well, such as French, German, Polish, ...
Most of these neighbourhoods are positioned at the south and in the downtown of Toronto.

CDN	City_Area	Latitude	Longitude	Cluster Labels	1st Most Common Origin	2nd Most Common Origin	3rd Most Common Origin	4th Most Common Origin	5th Most Common Origin	6th Most Common Origin
95	Annex	43.66936	-79.40280	4	English	Irish	Scottish	Canadian	German	French
122	Birchcliffe-Cliffside	43.69472	-79.26460	4	English	Irish	Canadian	Scottish	French	German
75	Church-Yonge Corridor	43.66024	-79.37868	4	English	Irish	Scottish	Chinese	Canadian	French
93	Dovercourt-Wallace Emerson-Junction	43.66604	-79.43687	4	Portuguese	English	Canadian	Irish	Scottish	Chinese
62	East End-Danforth	43.68415	-79.29911	4	English	Irish	Scottish	Canadian	French	German

Analyse each neighbourhood's competition

- Let's say we want to open an irish pub. We are going to use the cluster 4 in order to find the best neighbourhood for this will.
  In order to analyse the competition for each neighbourhood, we are going to retrieve the list of existing venues of the type pub, in the neighbourhoods categorised as cluster 4. For this task, we use FoursquareAPI.
- We build a dataframe as such (top 5 rows which represent 5 venues with the Pub category) :

CDN	Area Latitude	Area Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
95	43.66936	-79.40280	The Madison Avenue Pub	43.667947	-79.403486	Pub
95	43.66936	-79.40280	Duke of York	43.669186	-79.397527	Pub
75	43.66024	-79.37868	Churchmouse & Firkin	43.664632	-79.380406	Pub
62	43.68415	-79.29911	Grover Pub and Grub	43.679181	-79.297215	Pub
62	43.68415	-79.29911	Mullins Irish Pub	43.680348	-79.289370	Pub



- We assume that the top 5 neighbourhoods from this list represent the best places to open a new pub, as :
- the demographic data shows that people will likely come in this kind of venue in these neighbourhoods
- the competition is limited in these neighbourhoods
- We can draw these 5 neighbourhoods on the map of Toronto :

5.Conclusion

- In this project, we managed to cluster the city of Toronto using demographic data by neighbourhoods. This helps us identify which neighbourhoods are the most adequate for opening a new restaurant of a specific type.
- Then, we managed to identify the neighbourhoods with the less competition within these adequate neighbourhoods, in order to optimise the performance of this new business.
- Food service contractors can use similar data analysis in order to find the best spots to open a new restaurant.