Capstone Project - The Battle of Neighborhoods

Soizic Schneider

1. Introduction
2. description of the problem and discussion of the background

Toronto,is the most populous city of Canada with 2,7 million inhabitants as for 2016. Toronto is concentrating business, financial and cultural activities and is a very multicultural city.

Therefore, this is interesting to see how this multicultural aspect of the city is manifesting itself. The presence of restaurants from different cultures is a good way to find out how multicultural the city really is. How many different countries are represented? And how are these restaurants distributed amongst Toronto different neighborhoods?

2) description of the data and how it will be used to solve the problem

It is possible to find a list of neighborhoods on the Internet. Once this list is scrapped and I have a usable table, I will include geolocalisation aspects. Then, I will generate the list of first venues for restaurants with Foursquare.

The project should include several phases. First, I will analyze the data and find out which country are represented and what are the most common restaurants in Toronto. Then, I will use clustering to find out which neighborhoods have similar profiles and which have not.

3) Interest

This project can be interesting for different kinds of people:

a) business people wanting to open a restaurant need to know more about their competitors before they choose the location for their new restaurants.

b) people who are looking for a restaurant where they can eat want to know which restaurants are present in their area.

II) Data acquisition and cleaning

1. Data sources

I have been using the list of postal codes and neighborhood for Toronto on the following wikipedia page: <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>

I also used the following list of geospatial data for Toronto: <http://cocl.us/Geospatial_data>

Finally, I used Foursquare to generate a list of Toronto’s restaurants.

2) Data cleaning

I combined the data from Wikipedia and the data in the geospatial data list into one table. Then I used Foursquare to obtain the restaurants venues for Toronto and combined them with the already obtained data in one table.

III) Exploratory data analysis

1. Analyzing the different types of restaurants

Toronto is an international city and you can find there restaurants offering dishes from all over the world. Therefore, I grouped together the different kinds of restaurants present in my data set.

I obtained 61 different types of restaurants. However, I noticed that some types of restaurants were very similar. For instance, among the five top types of restaurants, I had both „Japanese restaurants“ and „Sushi Restaurants“. I therefore decided to sort the venues per continents to see better whether there was a continent that was more represented than others.

Here are the categories I defined:

Europe: “Italian Restaurant", "Greek Restaurant", "French Restaurant", "Mediterranean Restaurant", "Portuguese Restaurant", "Eastern European Restaurant", "Tapas Restaurant", "Modern European Restaurant", "German Restaurant", "Belgian Restaurant"

Asia: "Japanese Restaurant", "Sushi Restaurant", "Thai Restaurant", "Chinese Restaurant","Indian Restaurant", "Vietnamese Restaurant", "Korean Restaurant", "Ramen Restaurant", "Asian Restaurant", "Cantonese Restaurant", "Filipino Restaurant", "Tibetan Restaurant", "Afghan Restaurant", "Pakistani Restaurant", "Indonesian Restaurant", "Indian Chinese Restaurant", "Malay Restaurant", "Taiwanese Restaurant", "Sri Lankan Restaurant", "Hong Kong Restaurant", "Shanghai Restaurant", "North Indian Restaurant", "Dim Sum Restaurant", "Hakka Restaurant", "Dumpling Restaurant", "Udon Restaurant", "Hotpot Restaurant"

Africa: “Ethiopian Restaurant", "African Restaurant", "Moroccan Restaurant"

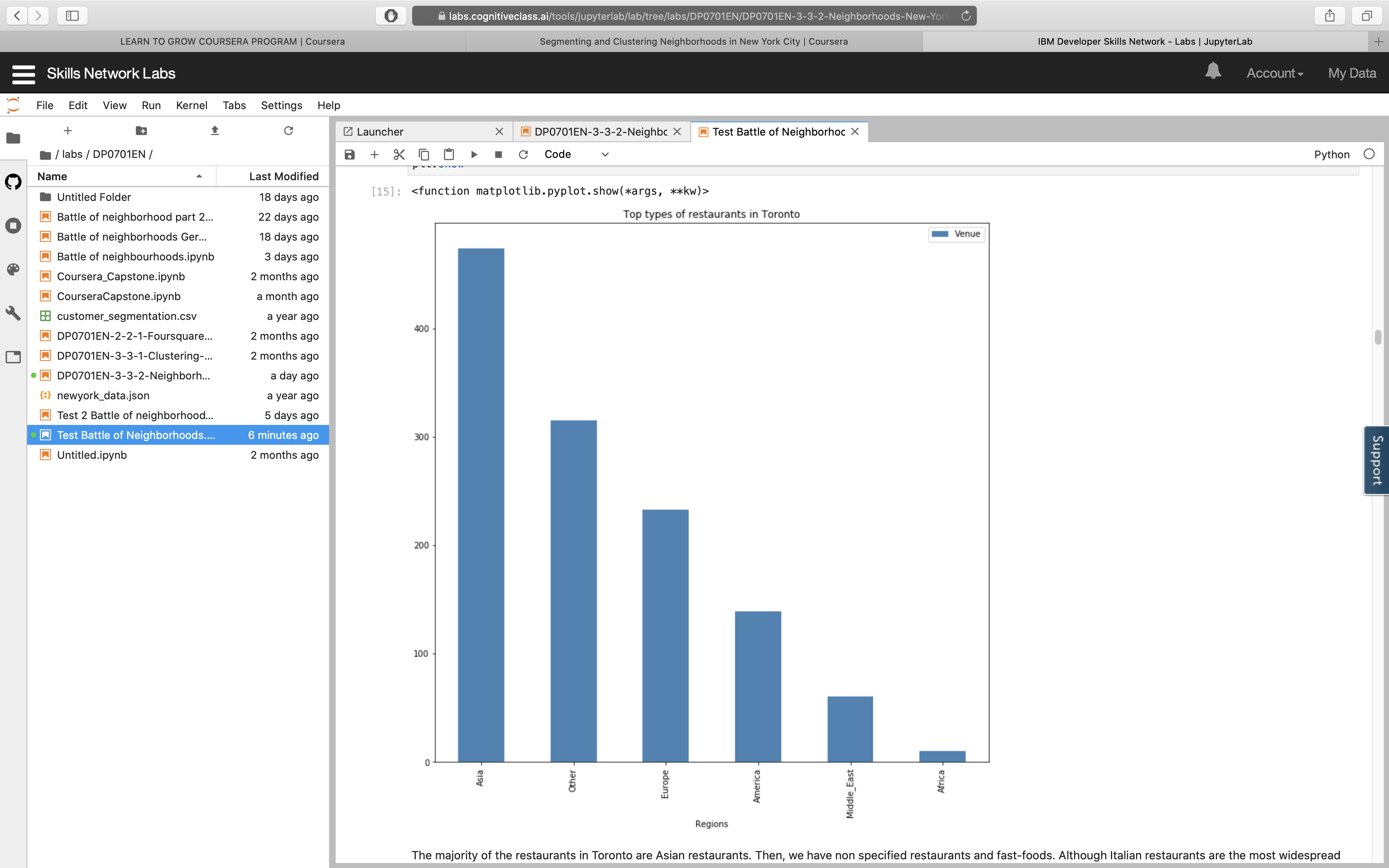
America: “American Restaurant", "Mexican Restaurant", "Caribbean Restaurant", "Latin American Restaurant", "New American Restaurant", "Brazilian Restaurant", "South American Restaurant", "Empanada Restaurant", "Hawaiian Restaurant", "Southern / Soul Food Restaurant"

Middle\_East: “Middle Eastern Restaurant", "Falafel Restaurant", "Turkish Restaurant", "Jewish Restaurant", "Persian Restaurant", "Syrian Restaurant"

Other: “Restaurant", "Fast Food Restaurant", "Vegetarian / Vegan Restaurant", "Seafood Restaurant", "Comfort Food Restaurant", "Gluten-free Restaurant","Theme Restaurant"

I am aware that the categories I chose are broad. I first wanted to define smaller geographical areas such as South-East Asia or Mediterranean area but it was first not always easy to sort venues into these categories because of venues such as „Indian Chinese Restaurant“ or „Dumpling Restaurants“. Moreover, for continents with less venues like Africa, it made little sense to sort venues between, for example Maghreb and Central Africa because there were not enough venues. Venues that could not be affected clearly to a continent are in the category „Other“. I wondered whether I should sort fast food restaurants in „America“ or in „Other“ but finally chose „Other“ because I considered that not only America has fast food and I could not know for sure what was the type of cooking behind each fast food restaurant in the data.

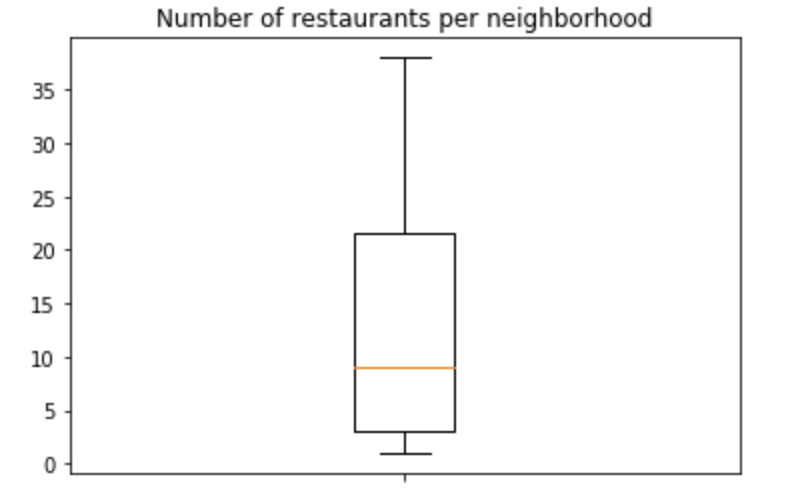
I finally made a histogram representing the number of venues per category:



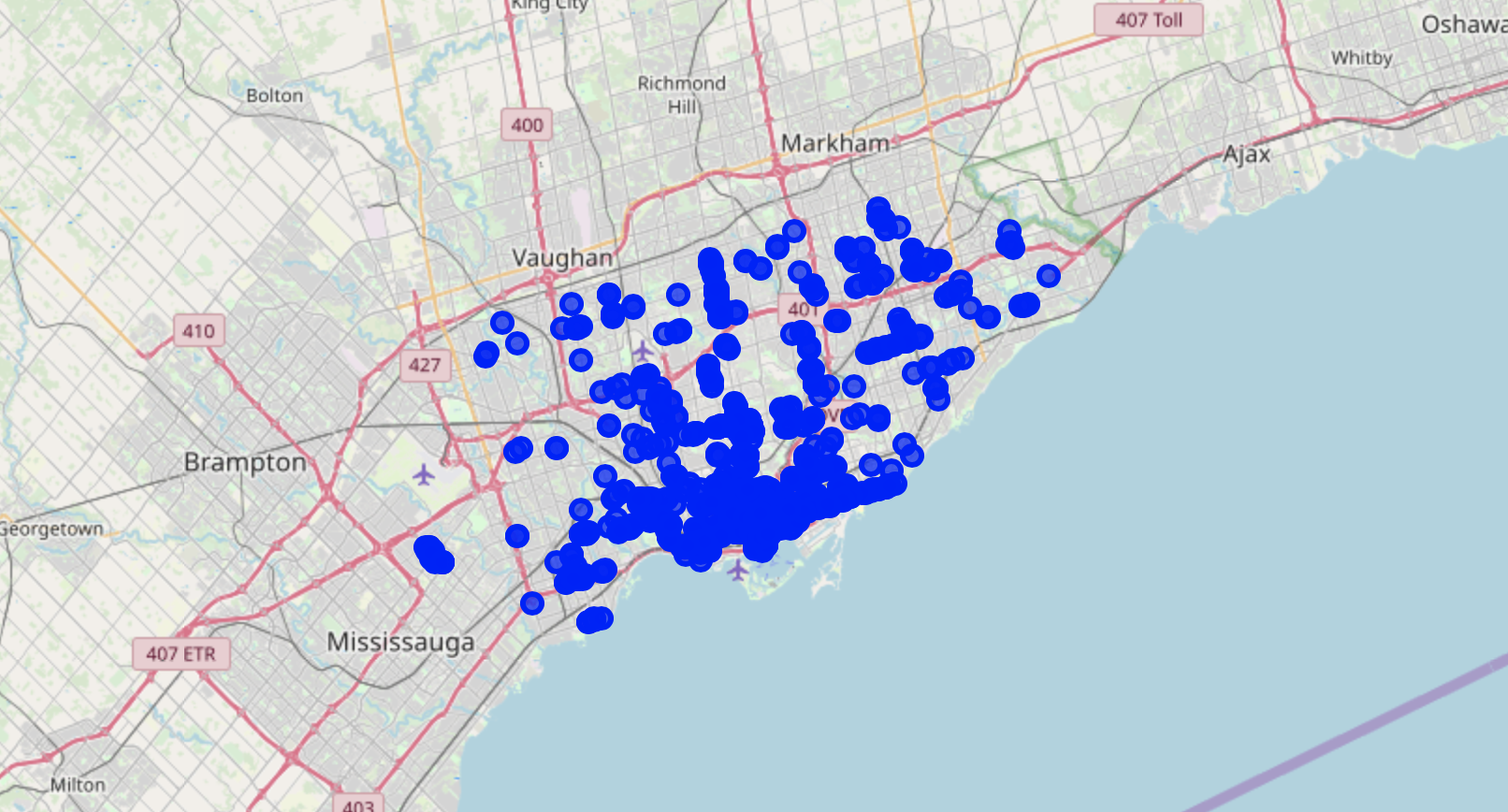
The majority of the restaurants in Toronto are Asian restaurants. Then, we have non specified restaurants and fast-foods. Although Italian restaurants are the most widespread type of restaurants in Toronto, European restaurants are only the third group of restaurants present in Toronto.

2) Repartition of the restaurant’s through the city

After analyzing the different kinds of restaurants, I wanted to check how they are located in the city. To do that, I first grouped venues per neighborhood. I sorted the values form the highest number of venues to the lowest number of venues and selected the first ten venues. I could see that the neighborhood with the most restaurants was Willowdale South with 40 restaurants.

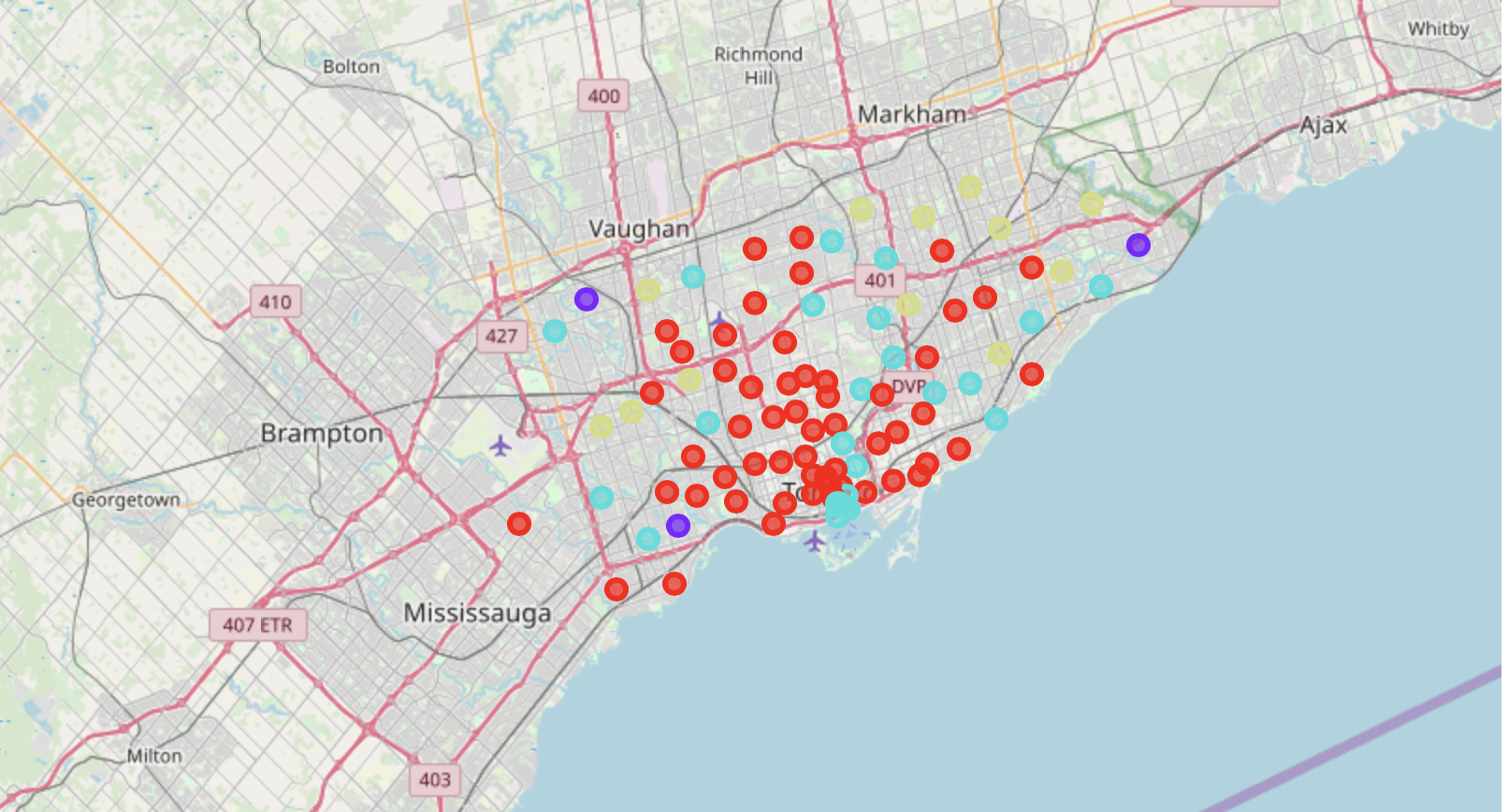
However, this list gave only limited information on the repartition of restaurants in the city. I chose to use a box plot to see better how the restaurants are distributed in different neighborhoods.

The median is 10, which means that 50% of the neighborhoods have less than 10 restaurants. 25% of the neighborhoods have more than twenty restaurants, the maximum being 40 for Willowdale South. On the other hand, 25% of the neighborhoods have less than 3 restaurants.

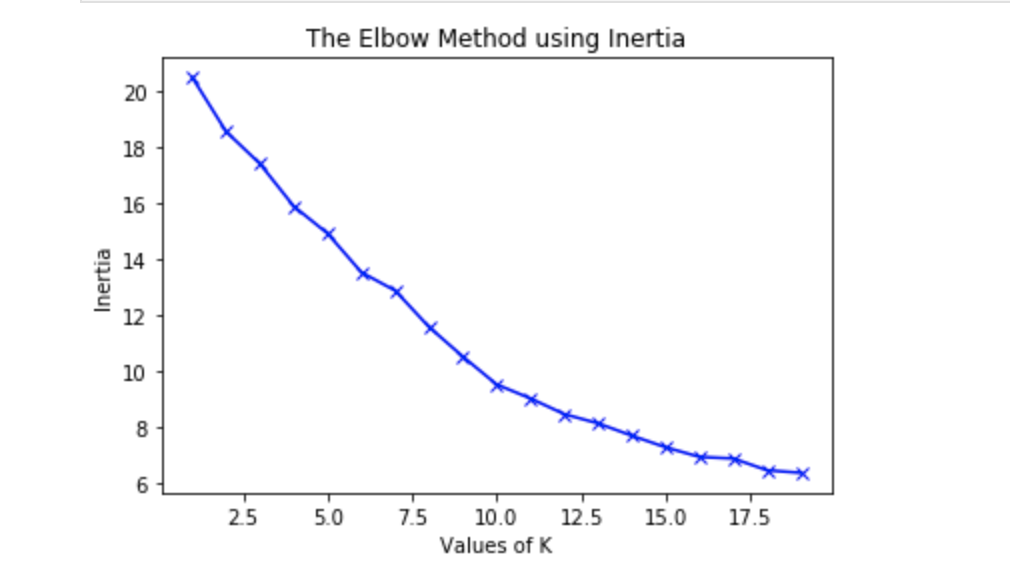
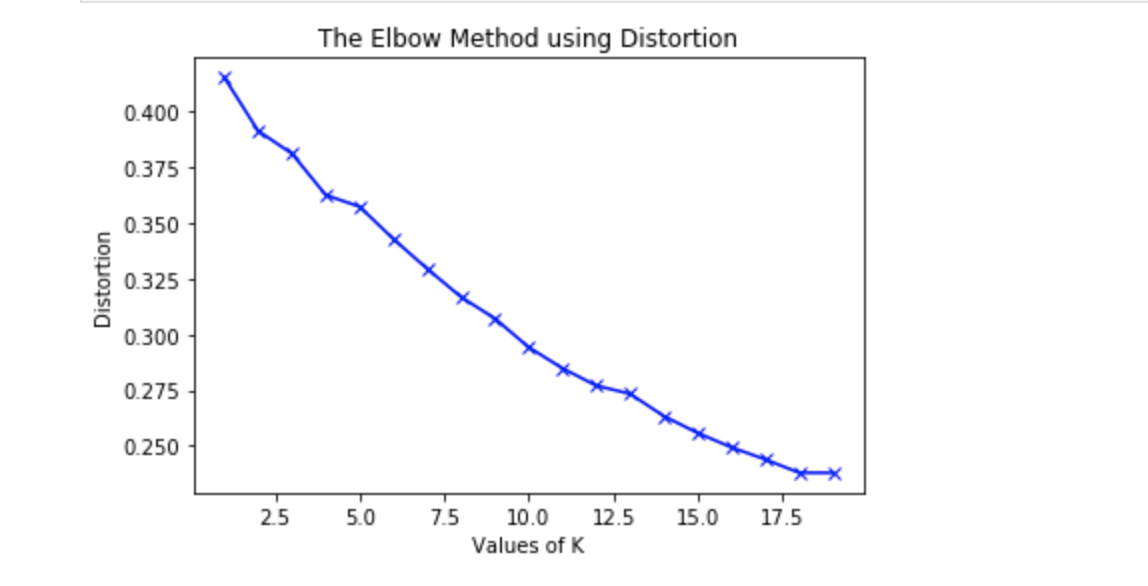
2) Localization of the restaurants

A visualization of the different restaurants of Toronto on a map shows that indeed the neighborhoods with most restaurants are localized in the center of the city, whereas the neighborhoods with less restaurants are rather in the outskirts, just as we could expect.

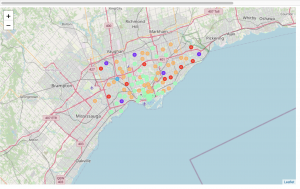
III) Clustering

I want to compare the city’s neighborhoods. For each neighborhood, I have the list of the most prevalent types of restaurants. Clustering will help me to see which neighborhoods have the same types of restaurants.

For k-means clustering, I first have to find the right k. To find out which is the best k, I will use the Elbow method as well as the Silhouette method.

For the Elbow method, I am using both distortion and inertia.

I am also using the Silhouette method to confirm the results already found with the Elbow method. Based on the different results obtained, I chose 4 as a value for k.



The neighborhoods are represented on the map with dots of different colors according to the clusters:

* Cluster 0 : This cluster only contains Etobicoke as this is the only neighborhood for which the first most common venue is the African restaurant scategory
* Cluster 1: This is the second biggest cluster. All the neighborhoods have as first most common venue the Asian restaurants category
* Cluster 2 : This is the largest cluster, with a high representation of Asian, other and America categories.
* Cluster 3 : This is a small cluster, with a high representation of restaurants of the Europe category
* Cluster 4: This is a small cluster, with a high representation of restaurants of the Other category

IV) Conclusion

This project enabled us to have a better idea of the repartition of restaurants in Toronto. As expected, the central neighborhoods had both a higher number of restaurants and most diversity. In the outskirts of the city, we had less restaurants and we could find more fast-food restaurants. Interestingly, although the most widespread type of restaurants was Italian restaurants, the continent that was most represented among Toronto restaurants was not Europe but Asia.