Data Science Capstone Project - The Battle of Neighborhoods

Introduction (Business Problem)

In this project I'm going to segment and cluster Brussels-Capital Region[1] in order to help expats choose the best place to settle in. Brussels, officially the Brussels-Capital Region, is a region of Belgium comprising 19 municipalities, including the City of Brussels, which is the capital of Belgium. I will be using the term "Brussels" when I mean Brussels-Capital Region, and the term "City of Brussels", when I mean one of the 19 municipalities being also the capital of Belgium.

Brussels is a very popular destination for expats[2]. More often than not, they move to Brussels looking for better working and living conditions than the ones they can find in their own country. Why is it so? Brussels hosts several international institutions. Think about European Union institutions[3] such as European Commission, Council of the European Union, European Council and European Parliament. The North Atlantic Treaty Organization headquarters[4] are located in Brussels. As well as numerous multinational groups headquarters and subsidiaries.

The number of expats in Brussels is estimated[5] to be 220,000 including 40,000 employees of European Union institutions and about 24,000 journalistes, lobbyists and diplomats gravitating around them. The number of NATO related workers is estimated to be 4,000.

The population of expats is diversified due to numerous factors, such as country of origin, income, age, marital status: are they single, married with children, what is the number and the age of children? In this study, I will segment and cluster the Brussels-Capital Region, to help expats choose the best place for living. It means that I will present different clusters with their majors characteristics and it will be up to expats themselves to perform an informed choice depending on their various expectations.

[1] https://en.wikipedia.org/wiki/Brussels

[2] https://en.wikipedia.org/wiki/Expatriate "An expatriate (often shortened to expat) is a person residing in a country other than their native country. In common usage, the term often refers to professionals, skilled workers, or artists taking positions outside their home country, either independently or sent abroad by their employers, which can be companies, universities, governments, or non-governmental organisations."

[3] https://en.wikipedia.org/wiki/Brussels_and_the_European_Union "Brussels (Belgium) is considered the de facto capital of the European Union, having a long history of hosting a number of principal EU institutions within its European Quarter. The EU has no official capital, and no plans to declare one, but Brussels hosts the official seats of the European Commission, Council of the European Union, and European Council, as well as a seat (officially the second seat but de facto the most important one) of the European Parliament."

[4] https://en.wikipedia.org/wiki/NATO_headquarters "The North Atlantic Treaty Organization is headquartered in a complex in Haren, part of the City of Brussels municipality of Belgium."

[5] https://www.beci.be/les-expats-a-bruxelles-des-attentes-a-satisfaire/ (in French)

Data

As explained in the Introduction, Brussels-Capital Region is made up of 19 municipalities, including the City of Brussels, the capital of Belgium. Remember that I use the term “Brussels” when I mean Brussels-Capital Region. Note: I was hesitating a lot about taking municipality as the basis for the project. Indeed, the first question that comes to mind is if the segmentation by municipality is not a too big subdivision to be suitable to point out main characteristics of it. At the beginning of the project I considered using a smaller subdivision such as a neighborhood - “quartier” in French - for which I found the necessary data on the website: https://monitoringdesquartiers.brussels/.

Finally I went back to the idea of using larger divisions such as municipality. Data over municipalities in Brussels is easily available. In this project I took advantage of BeautifulSoup library to scrap a Wikipedia page containing a table of 19 municipalities in Brussels. Page URL: https://en.wikipedia.org/wiki/List_of_municipalities_of_the_Brussels-Capital_Region.

See below 5 first rows :

	French name	Dutch name	Flag	CoA	post code	Population (1/1/2017)	Area	Population density (km²)	Ref.
1	Anderlecht	Anderlecht			1070	118,241	17.7 km^2 (6.8 sq mi)	6,680	[7]
2	Auderghem	Oudergem			1160	33,313	9.0 km^2 (3.5 sq mi)	3,701	[8]
3	Berchem-Sainte-Agathe	Sint-Agatha-Berchem			1082	24,701	2.9 km^2 (1.1 sq mi)	8,518	[9]
4	Ville de Bruxelles*	Stad Brussel*			1000 1020 1030 1040 1050 1120 1130	176,545	32.6 km^2 (12.6 sq mi)	5,415	[10]
5	Etterbeek	Etterbeek			1040	47,414	3.1 km^2 (1.2 sq mi)	15,295	[11]

Data Wrangling

The data from the above table was the starting point to prepare a file to be used as input for geopy and Nominatim service in order to get latitude and longitude for each municipality. The “Address” column has been added that was a concatenation of postal code and municipality name so that Nominatim can geolocate every place without ambiguity:

	Municipality	PostalCode	Address
0	Anderlecht	1070	1070, Anderlecht
1	Auderghem	1160	1160, Auderghem

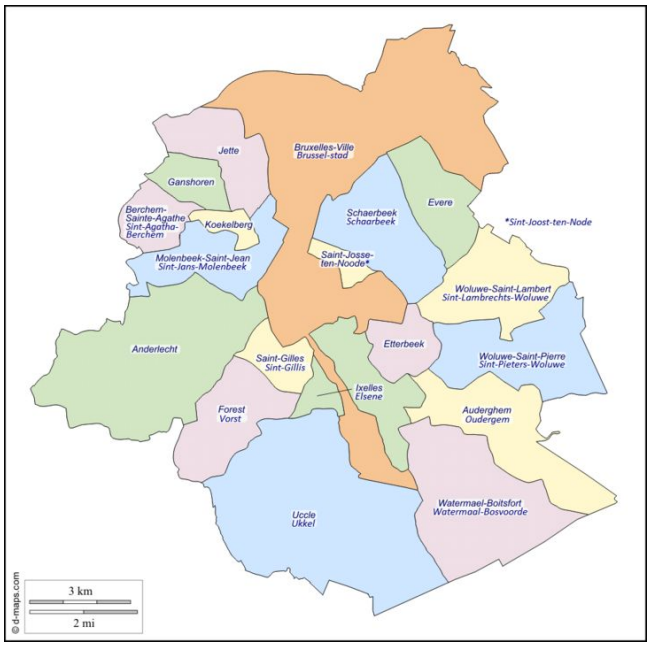
After geocoding with Nominatim and some data cleaning, I ended up with the file to be used in Foursquare API to get venues in each municipality:

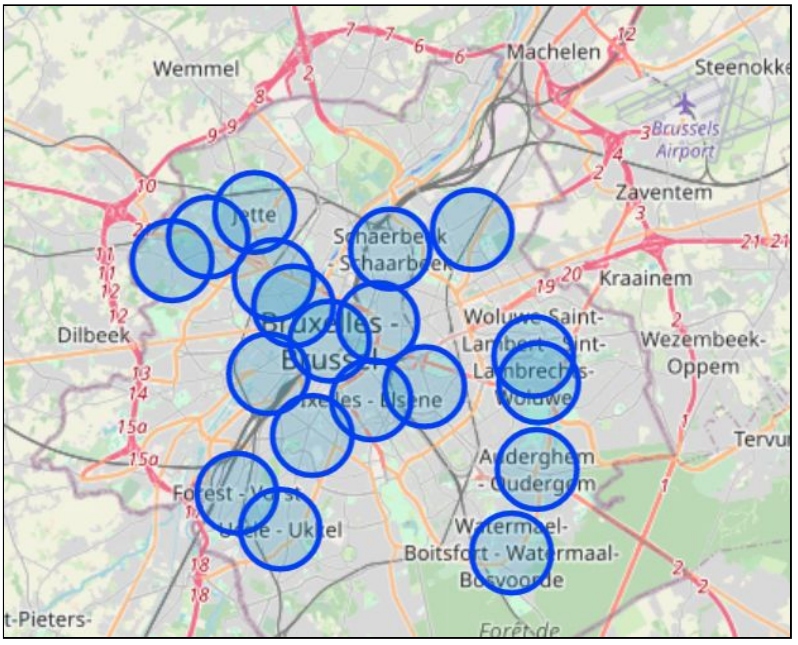
	Municipality	PostalCode	Latitude	Longitude
0	Anderlecht	1070	50.839098	4.329653
1	Auderghem	1160	50.817235	4.426898
2	Berchem-Sainte-Agathe	1082	50.864923	4.294673
3	Bruxelles	1000	50.846557	4.351697
4	Etterbeek	1040	50.836145	4.386174

Methodology

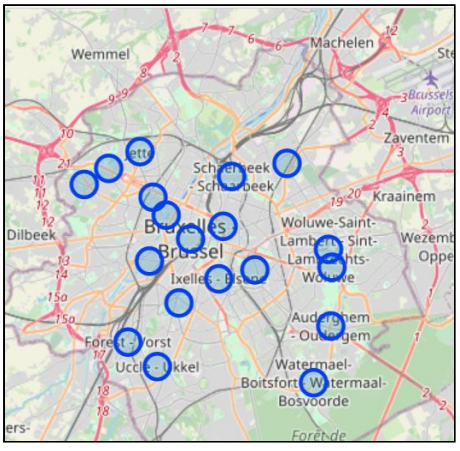
Segmentation and Clustering

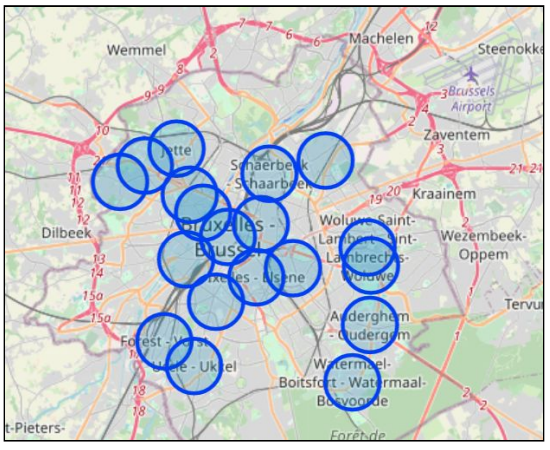
The geopy library and Nominatim service have been used to get coordinates of Brussels-Capital Region and the folium library was used to display a map with municipalities as markers on the map. The ‘folium.Circle’ has been used instead of ‘folium.CircleMarker’ in order to visualize the radius in meters rather than in pixels as it gives us a feeling about overlapping areas.

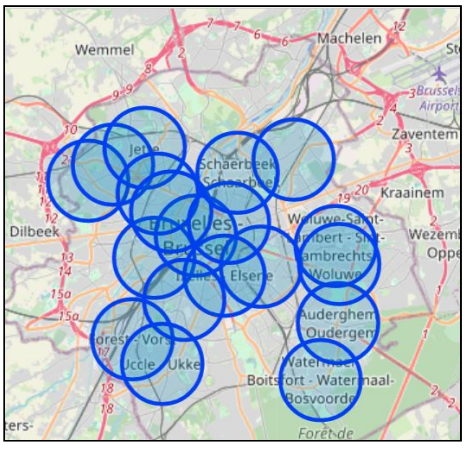




As you can see on the image at the left side, the territorial division of Brussels into 19 municipalities is quite heterogeneous. Areas differ very much in form and size. We are facing here a difficult choice of the circle radius because its value will be used later on to get venues in municipalities from Foursquare API. We understand here, that in fact, we are not going to get venues from municipalities, but from areas defined by the geographical coordinates as a center of a circle and the chosen radius. We can visualize the consequences of this choice when displaying maps using different circle radius: 500 meters, 1000 meters and 1500 meters. If the radius is small, circles are not overlapping a lot but it is a drawback for bigger municipalities as only a reduced area is covert. If the radius is big, a huge overlapping occurs for smaller municipalities which leads to the under-differentiation of clusters.







It appears that the usage of a circle to determine an area is not suitable for most of the municipalities of Brussels. The possible solution here might have been to divide municipalities in smaller groups of similar size and form and apply different circle radius parameters to different groups and merge results together.

The radius finally chosen for this project was 1000 meters - the map from the middle - as a compromise between less areas overlapping and best territorial coverage.

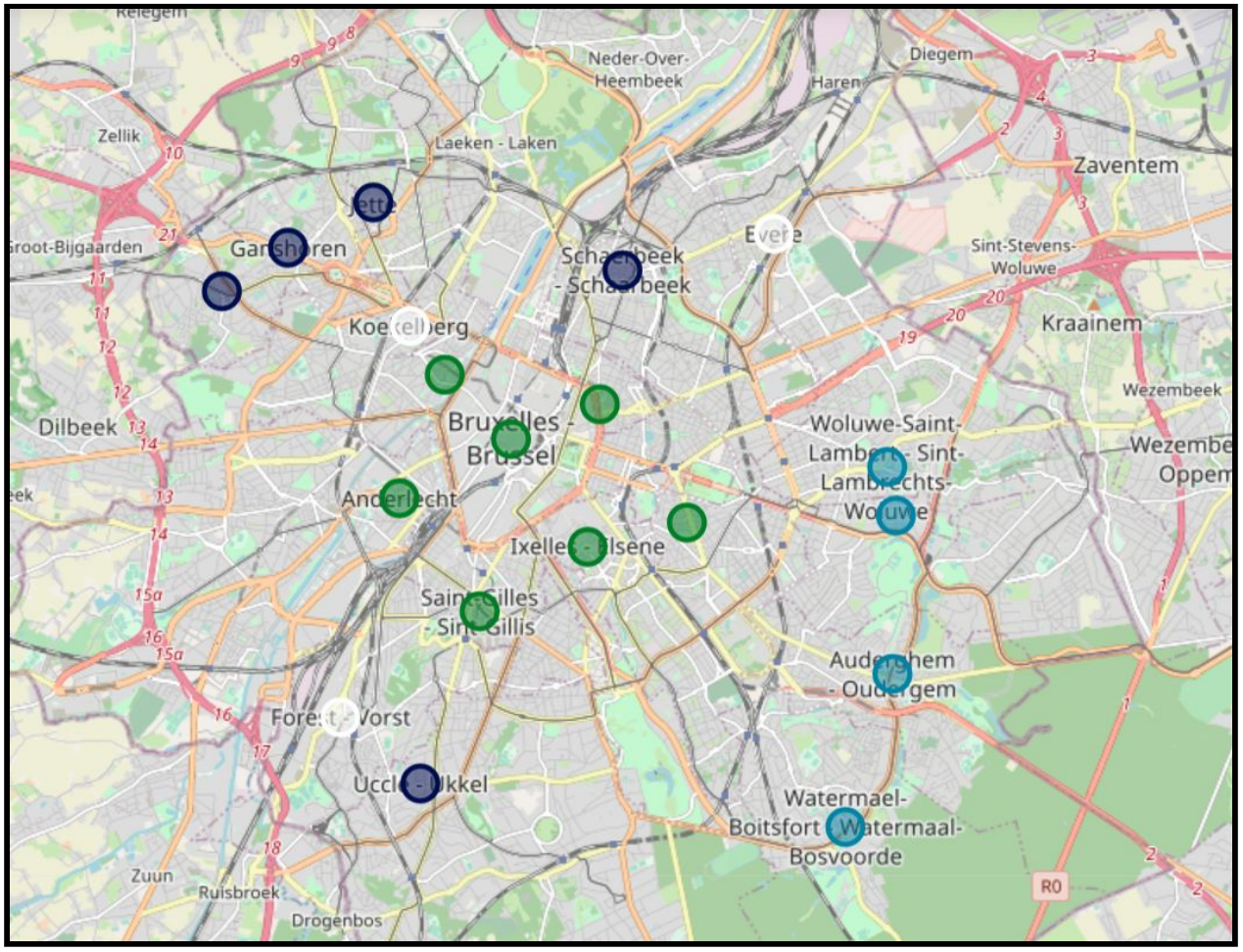
The Foursquare API has been used to explore the municipalities and segment them. The term neighborhood will be used instead of municipality from this point on. I started by exploring the first neighborhood in neighborhoods dataframe. It returned ‘Anderlecht’ neighborhood. Then I got the top 100 venues that are in ‘Anderlecht’ within a radius of 1000 meters. Afterwards the same method has been applied to get all the neighborhoods (municipalities) in Brussels-Capital and put them in a new dataframe called ‘brussels_venues’. There were 243 unique categories found. See below examples of categories:

```
array(['Plaza', 'Concert Hall', 'Toy / Game Store', 'Shopping Mall',
       'Chocolate Shop', 'Bar', 'Cheese Shop', 'Dessert Shop', 'Beer Bar',
       'Café', 'City Hall', 'Hotel', 'Clothing Store', 'Bookstore',
       'Historic Site', 'Italian Restaurant', 'Middle Eastern Restaurant',
       'Sandwich Place', 'Herbs & Spices Store', 'Fish & Chips Shop',
       'Gastropub', 'Belgian Restaurant', 'Wine Bar', 'Beer Store',
       'Coffee Shop', 'French Restaurant', 'Record Shop', 'Escape Room',
       'Bakery', 'Bubble Tea Shop', 'Theater', 'Comic Shop',
       'Indie Movie Theater', 'Portuguese Restaurant', 'Burger Joint',
       'Moroccan Restaurant', 'Supermarket', 'Hotel Bar',
       'Kitchen Supply Store', 'Cocktail Bar', 'Tea Room', 'Opera House',
       'Rooftop Bar', 'Thai Restaurant', 'Brazilian Restaurant',
       'Scenic Lookout', 'Eastern European Restaurant',
       'Department Store', 'Pub', 'Cosmetics Shop'], dtype=object)
```

The next step was to use the one-hot encoding technique that converts categorical values into dummies necessary for machine learning.

Then I clustered neighborhoods using Machine Learning technique, namely k-means. I chose to work with 4 clusters. Ideally, the number of clusters should have been determined by the appropriate method. The most known methods being “elbow”, “silhouette” and “gap statistics”.

Finally I visualized the resulting clusters on the map using Folium and proceeded with clusters exploration.



I examined each cluster and determined the discriminating venue categories that distinguish each cluster. Based on the defining categories, I made some general recommendations for expats so they can choose which neighborhoods best meet their expectations for settling themselves in Brussels for living and working.

Results

The clustering process returned 4 clusters of different sizes. For each cluster we display 10 most common venues per neighborhood so we can explore them. We select the cluster using ‘Cluster Labels’ that range from 0 to 3.

The first cluster (cluster label 0) (green marker on the map) has 7 municipalities:

	Municipality	Longitude	Cluster Labels	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	Anderlecht	4.329653	0	Sandwich Place	Hotel	Coffee Shop	French Restaurant	Supermarket	Greek Restaurant	Bar	Seafood Restaurant	Food Court	Metro Station
3	Bruxelles	4.351697	0	Chocolate Shop	Coffee Shop	Bar	Plaza	Bookstore	Beer Bar	Hotel	Italian Restaurant	Seafood Restaurant	Sandwich Place
4	Etterbeek	4.386174	0	Italian Restaurant	Bakery	Plaza	History Museum	Greek Restaurant	Coffee Shop	Sandwich Place	Hotel	Restaurant	Bar
8	Ixelles	4.366828	0	Boutique	Hotel	Italian Restaurant	Vegetarian / Vegan Restaurant	Bar	Tea Room	Bakery	Coffee Shop	Sandwich Place	Wine Bar
11	Molenbeek-Saint-Jean	4.338636	0	Bar	Seafood Restaurant	Hotel	Belgian Restaurant	French Restaurant	Theater	Restaurant	Plaza	Bakery	Burger Joint
12	Saint-Gilles	4.345484	0	Bar	Italian Restaurant	Brasserie	French Restaurant	Bakery	Restaurant	Plaza	Pizza Place	Hotel	Coffee Shop
13	Saint-Josse-ten-Noode	4.369163	0	Italian Restaurant	Hotel	Concert Hall	Sandwich Place	Bookstore	Pizza Place	Park	Plaza	Turkish Restaurant	Japanese Restaurant

The second cluster (cluster label 1) (indigo marker on the map) has 5 municipalities:

	Municipality	Longitude	Cluster Labels	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
2	Berchem-Sainte-Agathe	4.294673	1	Electronics Store	Supermarket	Restaurant	Bar	Greek Restaurant	Gym	Cosmetics Shop	Furniture / Home Store	Brasserie	Plaza
7	Ganshoren	4.307798	1	Bar	Restaurant	Italian Restaurant	Plaza	Pizza Place	Park	Supermarket	Chinese Restaurant	Hockey Field	Electronics Store
9	Jette	4.324570	1	Bar	Plaza	Bakery	Pizza Place	Sandwich Place	Italian Restaurant	Park	Snack Place	Convenience Store	Platform
14	Schaerbeek	4.373712	1	Snack Place	Supermarket	Bar	Turkish Restaurant	Gym / Fitness Center	Tram Station	Italian Restaurant	Coffee Shop	Diner	Music Venue
15	Uccle	4.333844	1	Supermarket	French Restaurant	Sandwich Place	Plaza	Bar	Bakery	Park	Pizza Place	Chinese Restaurant	Italian Restaurant

The third cluster (cluster label 2) (blue marker on the map) has 4 municipalities:

	Municipality	Longitude	Cluster Labels	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
1	Auderghem	4.426898	2	Italian Restaurant	Bakery	Pharmacy	Pizza Place	Plaza	Sandwich Place	Park	Fast Food Restaurant	Restaurant	French Restaurant
16	Watermael-Boitsfort	4.417644	2	Restaurant	Bus Stop	Italian Restaurant	French Restaurant	Park	Ice Cream Shop	Gastropub	Chinese Restaurant	Farmers Market	Event Service
17	Woluwe-Saint-Lambert	4.425673	2	Italian Restaurant	French Restaurant	Supermarket	Sushi Restaurant	Fast Food Restaurant	Restaurant	Park	Gourmet Shop	Asian Restaurant	Bakery
18	Woluwe-Saint-Pierre	4.427464	2	Italian Restaurant	Park	Restaurant	Pharmacy	Supermarket	French Restaurant	Sandwich Place	Sushi Restaurant	Belgian Restaurant	Bistro

The fourth cluster (cluster label 3) (white marker on the map) has 3 municipalities:

	Municipality	Longitude	Cluster Labels	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
5	Evere	4.403418	3	Park	Bar	Snack Place	Supermarket	Sandwich Place	Tram Station	Paper / Office Supplies Store	Hotel	Brasserie	Train Station
6	Forest	4.318119	3	Park	Supermarket	Sandwich Place	Furniture / Home Store	Brasserie	Bus Stop	Pet Store	Bookstore	Snack Place	Bus Station
10	Koekelberg	4.331550	3	Supermarket	Snack Place	Sandwich Place	French Restaurant	Convenience Store	Hostel	Gym / Fitness Center	Italian Restaurant	Park	History Museum

Discussion

The first cluster (cluster label 0, green marker on the map, 7 municipalities)

All municipalities from this cluster are regrouped in the center or near the center of Brussels. They are adjacent one to another. With its 7 municipalities it is the largest cluster. We notice that the most common venues are restaurants serving dishes from all over the world, bars, fast-foods and hotels. It suggests lively urban areas and the presence of many touristes. These kinds of places might be attractive to younger expats, singles or couples without children enjoying to go out and have fun with friends after their work, without having to do long distances to get there.

The second cluster (cluster label 1, indigo marker on the map, 5 municipalities)

The majority of municipalities from this cluster - 4 of them - are located outside the center, in the northern part of Brussels. They are adjacent one to another. One municipality being located outside as well but at the extreme southern part and not adjacent to others. We count less restaurants and bars than in the first cluster and we start to have among the most common venues parks, gyms and even a hockey field. Supermarkets are also present. The municipalities from this cluster being located outside the center and well equipped for day to day life suggest that they are better suited for expats families with children.

The third cluster (cluster label 2, blue marker on the map, 4 municipalities)

All municipalities from this cluster are located outside the center and close to The Sonian Forest at the southeast edge of Brussels. They are adjacent one to another. The most common venues are parks and all kinds of restaurants. There are also bakeries, pharmacies, supermarkets and even an ice cream shop. These municipalities are very well suited for expats who enjoy green areas and value the near presence of the large forest where they can go walking or biking after work.

The fourth cluster (cluster label 3, white marker on the map, 3 municipalities)

All 3 municipalities from this cluster are spread from nord to south and are not adjacent one to each other.
The most common venues are supermarkets and fast-foods. It is difficult to distinguish dominating characteristics here and furthermore give any general recommendations.

Conclusion

After exploring all clusters, there is a question that comes to mind: to what extent the Foursquare application is suitable to solve the problem we are trying to address in this project, namely, give recommendations about the best place for living for the expats arriving for work in Brussels based on their expectations?

Foursquare contains a lot of valuable information gathered from application’s users about places they have visited. It concerns mainly the best places to go out, to do shopping, to do sports. In the project we take into account 10 most common venues for each of 19 municipalities and we notice that restaurants of all kinds as well as bars, coffee shops and fast-foods are among the most visited venues. The predominance of food and drink related places limits the interest of clustering based on 10 most common venues to solve the problem stated in this project.

Nevertheless, we ended up with 4 clusters, from which 3 were well differentiated and allowed us to give some recommendations for expats coming to Brussels for work.