Where to Locate a new Italian restaurant in Antwerpen, Belgium

Coursera Capstone Project

Created by: Anna Sukhareva, data scientist
Antwerpen, Belgium
Date: November 04, 2019
Contacts: anna@linefeed.be

Table of Contents

Part 1 : Business Problem & Interests

Problem:

Location influences the success or failure of a restaurant in a host of ways, from attracting enough initial customer interest to being convenient to visit. The restaurant's location is also interrelated to other factors, for instance, the immediate surroundings of the restaurant site, accessibility and right neighborhoods (as competitive). To determine the location for a new restaurant, making the research surrounding businesses is a must, to answer the following questions:

Is there enough room for a new restaurant?

- What are the local trends in that area?
- How location works for surrounding businesses, and what impact will it make on a new business performance?
- Question:
- Can we cluster similar areas of Antwerpen and make a profile of each area?

Who would be interested in this project:

Antwerp is one of the most popular city in West Europe and the second largest metropolitan region in Belgium. The Port of Antwerp is one of the biggest in the world, ranking second in Europe, and within the top 20 globally; the city is also known for its diamond industry and trading.

The potential investors of horeca and food chain businesses would be interested in this project.

Part 2 : Data description & sources

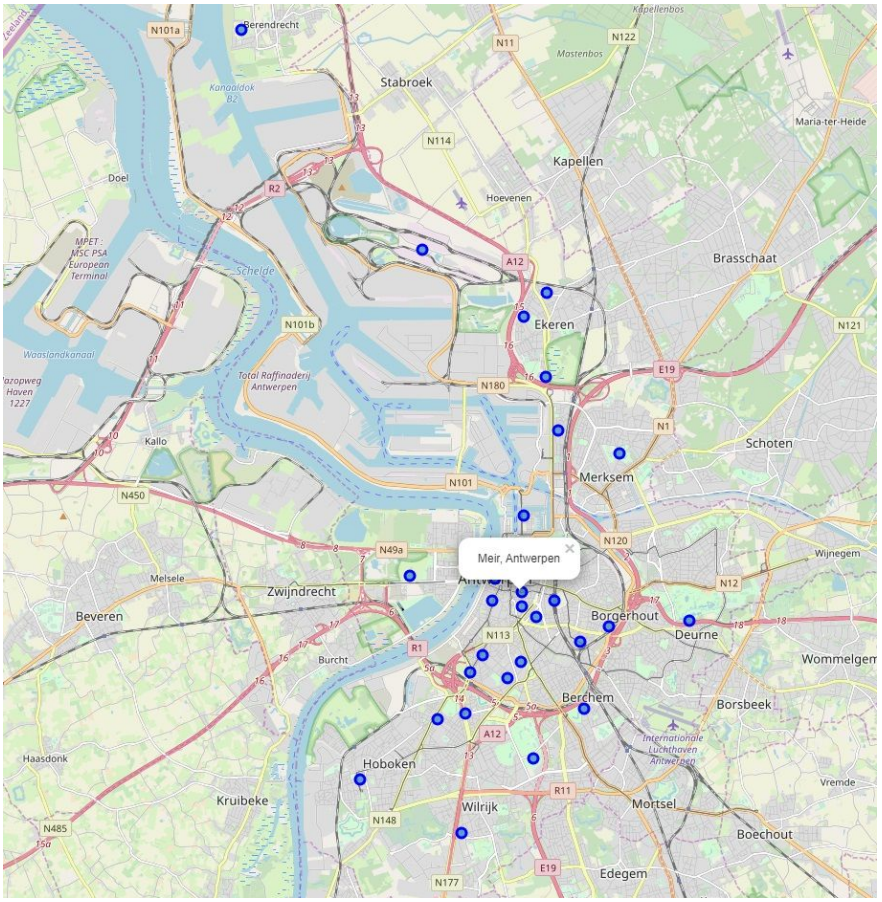
Data that was used to solve the problem and the source of the data:

1. Data about the borough (districts) of the city of Antwerp, provided in CSV format. Source: from Open Data Antwerpen Portal
2. Data about neighborhoods of the biggest borough of Antwerp, called "Antwerpen", Source: Wikipedia
3. Location data of the borough and neighborhoods (latitude, longitude). Source: gathered using Geopy library.
4. Data about all different venues in Antwerpen neighborhoods. Source: gathered using the Foursquare API

Part 3 : Methodology

I collect data about the borough (districts) of the city of Antwerp, from Open Data Antwerpen Portal; then I collect unstructured data about neighborhoods of the biggest borough of Antwerp, called "Antwerpen", using the Wikipedia. Using the Geopy library, I gather location data (latitude, longitude) for each neighborhood. I used the locations data to visualize data, using the Folium library.

The result: 9 boroughs and 31 neighborhoods:



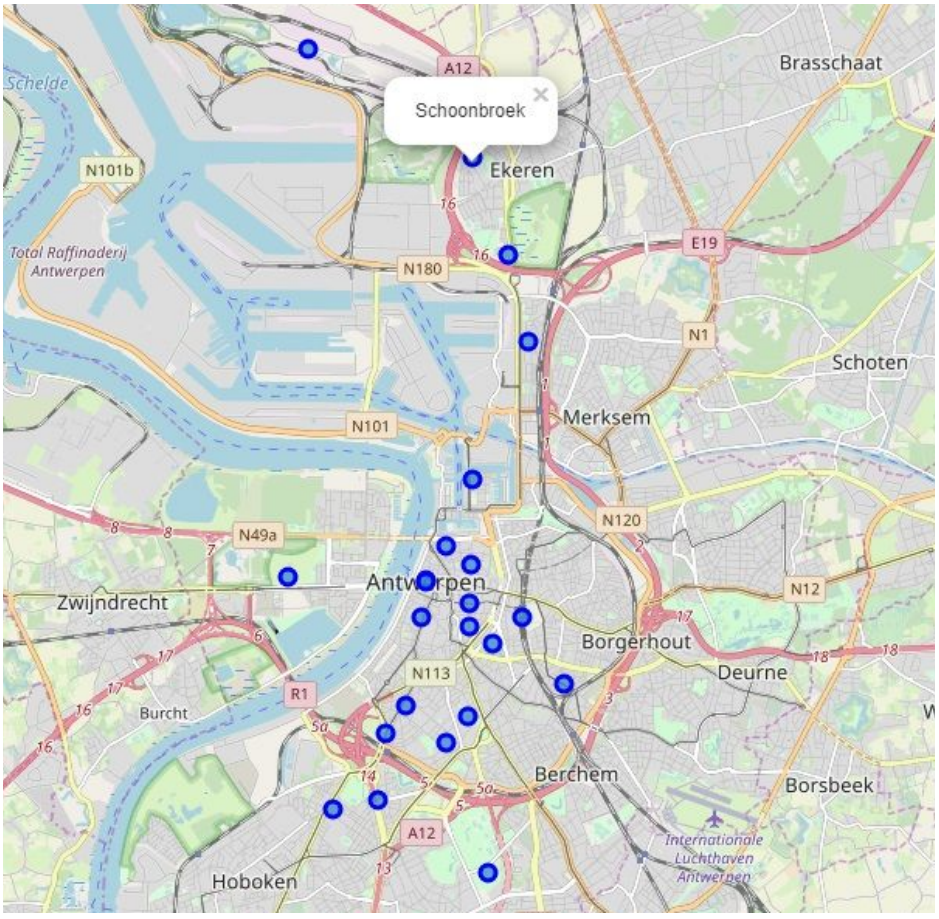
Pic: Borough and neighborhoods of Antwerp, Belgium.

As we're interested in the central district of the city, I narrow the research to the largest borough of Antwer, called 'Antwerpen'.

The result: 23 neighborhoods:

	Borough	Neighborhood	latitude	longitude
0	Antwerpen	Schoonbroek	51.279933	4.409993
1	Antwerpen	Eilandje	51.235300	4.409900
2	Antwerpen	Zuid	51.199941	4.390926
3	Antwerpen	Zurenborg	51.206853	4.430287
4	Antwerpen	Theaterbuurt	51.214919	4.409325
5	Antwerpen	Tentoonstellingswijk	51.190651	4.388902
6	Antwerpen	Stadspark	51.212480	4.414373
7	Antwerpen	Sint-Andries	51.216174	4.398647
8	Antwerpen	Schipperskwartier	51.225922	4.404064
9	Antwerpen	Middelheim	51.180541	4.413492
10	Antwerpen	Meir	51.218062	4.409403
11	Antwerpen	Markgrave	51.198639	4.404020
12	Antwerpen	Historisch Centrum	51.221200	4.399800
13	Antwerpen	Rozemaai	51.266475	4.417845
14	Antwerpen	Universiteitswijk	51.223500	4.409600
15	Antwerpen	Linkeroever	51.221673	4.368960
16	Antwerpen	Kiel	51.189501	4.379055
17	Antwerpen	Harmonie	51.202394	4.408989
18	Antwerpen	Den Dam	51.221110	4.399708
19	Antwerpen	Centraal Station	51.216083	4.421038
20	Antwerpen	Brederode	51.203865	4.395338
21	Antwerpen	Antwerpen Noord	51.294924	4.373469
22	Antwerpen	Luchtbal	51.254337	4.422190

Pic: Borough and neighborhoods of Antwerp, Belgium



Pic: Neighborhoods of the borough Antwerp

Then I request data about all different venues in Antwerpen neighborhoods, using the Foursquare API; data is provided as JSON file.
I structured data to dataframe.

The result: dataframe with 1139 venues, with name, category and coordinates for each of them:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Schoonbroek	51.279933	4.409993	't Aperoke	51.278670	4.410173	Bar
1	Schoonbroek	51.279933	4.409993	Zwembad De Schinde	51.279817	4.413643	Pool
2	Schoonbroek	51.279933	4.409993	Putten van Ekeren	51.282588	4.404558	Park
3	Schoonbroek	51.279933	4.409993	Halte Ekeren Akkerstraat	51.279010	4.410192	Bus Stop
4	Schoonbroek	51.279933	4.409993	Halte Antwerpen Oorderseweg	51.281993	4.411528	Bus Stop

Then I grouped them by category; there was 196 uniques categories.

Neighborhood	
Antwerpen Noord	1
Brederode	39
Centraal Station	64
Den Dam	100
Eilandje	43
Harmonie	23
Historisch Centrum	100
Kiel	26
Linkeroever	16
Luchtbal	23
Markgrave	29
Meir	100
Middelheim	8
Rozemaai	5
Schipperskwartier	95
Schoonbroek	11
Sint-Andries	100
Stadspark	69
Tentoonstellingswijk	25
Theaterbuurt	100
Universiteitswijk	87
Zuid	25
Zurenborg	50

Pic: How many venues for each neighborhood?

I used “hot coding” technique to get an average frequency of occasionality of different categories of venues in each neighborhood, and grouped them by neighborhood.

Then I got 5 top venues for each neighborhood, f.e.:

```
----Meir----
            venue  freq
0  Clothing Store  0.15
1        Boutique  0.07
2     Coffee Shop  0.06
3         Theater  0.04
4  Cosmetics Shop  0.04


----Historisch Centrum----
               venue  freq
0                Bar  0.10
1        Coffee Shop  0.08
2       Cocktail Bar  0.07
3         Restaurant  0.06
4  French Restaurant  0.05


----Rozemaai----
                venue  freq
0           Bus Stop   0.2
1        Auto Garage   0.2
2           Pharmacy   0.2
3  Light Rail Station   0.2
4        Sports Club   0.2
```

Pic: 5 top venues for the neighborhood

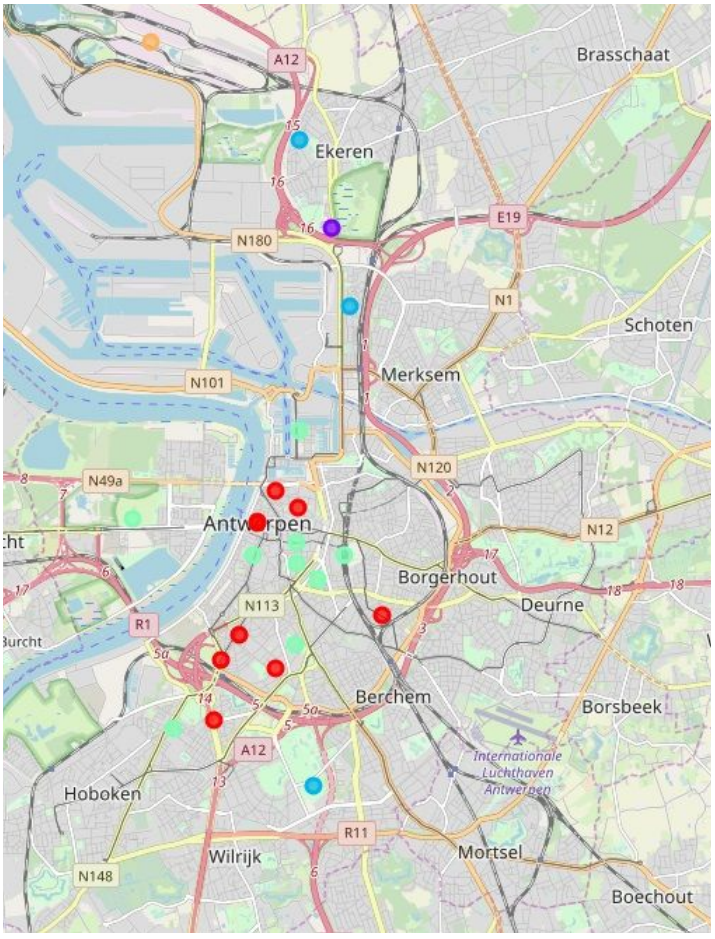
Then I structured a new dataframe with top 10 most common venues for each neighborhood:

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Antwerpen Noord	Train Station	Zoo Exhibit	Donut Shop	Flower Shop	Flea Market
Brederode	Coffee Shop	Bar	Pub	Friterie	Middle Eastern Restaurant
Centraal Station	Zoo Exhibit	Asian Restaurant	Italian Restaurant	Thai Restaurant	Coffee Shop
Den Dam	Bar	Coffee Shop	Cocktail Bar	French Restaurant	Restaurant
Eilandje	Nightclub	Sandwich Place	Coffee Shop	Restaurant	Beach Bar

Pic: Top 10 most common venues for each neighborhood, head

I used unsupervised machine learning technique - clustering, with KMeans method to cluster all neighborhood, using the venues, to make a business profile of each cluster, that determines its business environment, competition level, accessibility and other features.

The result: 5 clusters:



Pic: clustered neighborhoods

where:
cluster 0 - red
cluster 1 - purple
cluster 2 - blue
cluster 3 - green
cluster 4 - orange

Finally,
I analyze each cluster to make a conclusion if it is possible an appropriate location to open a new Italian restaurant, for example, that's how cluster 0 looks like:

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Zuid	Bar	Pizza Place	Restaurant	Coffee Shop	Moving Target
Zurenborg	Bar	Restaurant	Bistro	Pizza Place	Bakery
Tentoonstellingswijk	Park	Bakery	Grocery Store	Pharmacy	Supermarket
Schipperskwartier	Bar	Restaurant	Coffee Shop	Italian Restaurant	Gay Bar
Markgrave	Sandwich Place	Italian Restaurant	Supermarket	Belgian Restaurant	Restaurant
Historisch Centrum	Bar	Coffee Shop	Cocktail Bar	Restaurant	French Restaurant
Universiteitswijk	Bar	Coffee Shop	Sandwich Place	Asian Restaurant	Soup Place
Den Dam	Bar	Coffee Shop	Cocktail Bar	French Restaurant	Restaurant
Brederode	Coffee Shop	Bar	Pub	Friterie	Middle Eastern Restaurant

Pic: cluster 0

The results: By examining the clusters, we identified 5 clusters with following profiles:

Part 4 : Results section

Cluster 1, red

- Includes: 8 neighborhoods (Zuid, Zurenborg, Tentoonstellingswijk, Schipperskwartier, Markgrave, Historisch Centrum, Universiteitswijk, Den Dam, Brederode).
- Business environment: The most popular categories of venues are bar, coffee shops and restaurants.
- Competitors: Hight competitors area; competitors are with different horeca type (coffee bars, bar, restaurant, fast food restaurants). Italian restaurants meets 5 times in top of 'venues categories '.
- Accessibility: Foot traffic.

Cluster 2, purple

- Includes: 1 neighborhood (Rozemaai).
- Business environment: The most popular categories are sport clubs, pharmacy, bus stops; we can consider the cluster as 'uptown'.
- Competitors: Low competitors area.
- Accessibility: Very good; train station, wide net of bus stations.

Cluster 3, blue

- Includes: 3 neighboorhoods (Schoonbroek, Middelheim, Luchtbal)
- Business environment: The most popular categories are sport clubs, parks, horeca, health & beauty service.
- Competitors: Middle competitors area.
- Accessibility: Very good; wide net of bus stations, foot traffic.

Cluster 4, green

- Includes: 9 neighborhoods (Eilandje, Theaterbuurt, Stadspark, Sint-Andries, Meir, Linkeroever, Kiel, Harmonie, Centraal Station)
- Business environment: The most popular categories are horeca, clothing stores & boutique.
- Competitors: Highly competitors area.
- Accessibility: Foot traffic.

Cluster 5, orange

- Includes:1 neighborhood (Antwerpen Noord).
- Business environment: The most popular categories are sport clubs, pharmacy, auto garage; we can consider the cluster as 'uptown'.
- Competitors: Low competitors area.
- Accessibility: Good, a wide net of bus stops.

Part 5 : Discussion section

As features, to cluster the neighborhoods we used:

- Categories of venues (as horeca, parks, sport clubs),
- Types of horeca (restaurants, coffee shops, bars),
- Restaurants cousines (Italian, Azian, Marrocance...etc),
- Accessibility (bus and train stations).

Although the K-Means method works properly, we can also notice that the clusters is not precise enough; we would recommend adding more features like population and average income.

As well, as Foursquare returns only 1139 venues, we can conclude that this platform is less popular in Belgium. We would recommend along Foursquare using alternative geolocation platforms, as Google Maps API.

Part 6 : Conclusion section

By examining the clusters, we identified 5 clusters.

Two of clusters them have high competitive level, having horeca at the top of all venues, with presents all types of horeca, including range of Italian restaurants. We consider the areas are not having a place for a new Italian restaurant.

Also, two clusters we determine as residential area category uptown, with calm environment; those two does not provide high interest for our purposes. We consider the areas are not appropriate for opening a new Italian restaurant.

As an appropriate location for a new Italian restaurant, we would suggest the cluster having following neighborhoods: **Schoonbroek, Middelheim, Luchtbal**.

- The most popular categories of venues in those areas are sport clubs, parks, horeca, health & beauty service; it may be considered as areas with high families traffic, especially at the weekend.
- It has a good accessibility: wide net of bus stations, and foot traffic.
- Although it still have range of horeca venues, restaurants don't taking a leader positions in comparing with other categories of venues; as well there are a wide range of presented horeca categories (like cafe, bakkery, fritery, bar). There were no Italian restaurant spotted in those areas.

Based on the above, we consider Schoonbroek, Middelheim, Luchtbal as the most appropriate areas in Antwerpen borough of the city of Antwerp, Belgium, to open an Italian restaurant.

End of the report.