How to open new pizzeria with DBSCAN.

Arif Huseynov.

Turin, 2021.

1. Introduction.

Description & discussion of the background.

Welcome to the opening of our pizzeria! These are the words that the first guests of your restaurant will hear from its smiling owner. Sparkling glasses of Piedmont wine, the landscapes of the endless Alps frozen in the paintings on the wall, and the breathtaking smells of various pizzas everywhere, attracting more and more visitors ...

Indeed, nice dream, but how to achieve it? Well, although the whole answer to this question is a complex strategy, I am going to answer to the part of this question, specifically – how to choose the right place for your new business.

For the current analysis Metropolitan City of Turin is chosen, former capital of Italy and current capital of Piedmont region. All big cities in Italy significantly differ from each other, and Turin is not the exception. This is the bridge between Italian and French cuisine, architecture and culture. Moreover, Turin is so called 'industrial capital' of Italy. Fiat, Lancia, Alfa Romeo, Iveco – did you hear these names? Or Lavazza? Well, these all are originated in Turin.

Of course, we all heard about love of Italians for pizza. If we will search in any city of Italy for the pizzeria, we will find a plenty of restaurants. So, is it worth to open new pizzeria? I guessed, that first input for making the decision can be such part of the city, which is less populated with the current category of restaurants, but highly populated with people. The second input can be the price for rent the premises and the availability of premises in general.

The key word of the problem described above is ***density***.

Data Description.

The main data I need for this project is information about locations of all pizzerias in Turin. Mining process will be divided on two parts:

1. First, I will get information about the administrative division of the city using web scrapping with Beautiful Soup. The source of data is Wikipedia page: https://en.wikipedia.org/wiki/Turin#Administration.
2. Then I need to find geolocation information for every borough (Circoscrizione) with Google Geocoding API.
3. After that for every borough I can explore restaurants under category 'pizza' with Foursquare API. Thus, I will get location information for every interested venue in Turin.
4. The last information I will need after obtaining clusters of pizzerias in Turin is data about premises for rent and their locations. I will use Google search and choose some agency website, then I will try to mine data from it.

2. Methodology.

Basically, we have four general steps to achieve final results:

1. Acquisition and representation of the data about city administrative division.
2. Acquisition and representation of the data about interested venues for each administrative entity.
3. Clustering pizzerias and showing results on the map.
4. Acquisition and representation on the map data about available for rent premises.

Almost in every step we need to acquire some data, make some preparation and analysis if needed, and then represent it on the map. Different libraries will be used for these purposes, in every step I will mention appropriate library name.

As a start we need to represent on the map centroids of every administrative entity of Turin. Turin is divided on the 8 boroughs, called in Italian '*Circoscrizione'*. Every borough also divided on smaller zones called in Italian '*borghi*' or '*quartieri'*. To get the information about names of every zone I used Wikipedia official page about Metropolitan City of Turin - https://en.wikipedia.org/wiki/Turin#Administration. Library used for web-scrapping is '*Beautiful Soup*' (https://pypi.org/project/beautifulsoup4/). The example of results are shown below:

Circoscrizione	Zone
Circoscrizione 1	Centro
Circoscrizione 1	Crocetta
Circoscrizione 2	Santa Rita
Circoscrizione 2	Mirafiori Nord
Circoscrizione 2	Mirafiori Sud
Circoscrizione 3	San Paolo
Circoscrizione 3	Cenisia
Circoscrizione 3	Pozzo Strada
Circoscrizione 3	Cit Turin
Circoscrizione 3	Borgata Lesna

Table 1. Part of dataframe with zonal division of Metropolitan City of Turin.

Overall, we have 8 boroughs and 34 zones. Next step is determination of the zone locations. It is time to use Google Geocoding API, thankfully to Google's gift for the new user subscription I have some free credits, which will be more than enough for this project. Results are shown on the interactive map (), which is available through python Folium library. The points are representing zonal centroids, acquired Geocoding API.

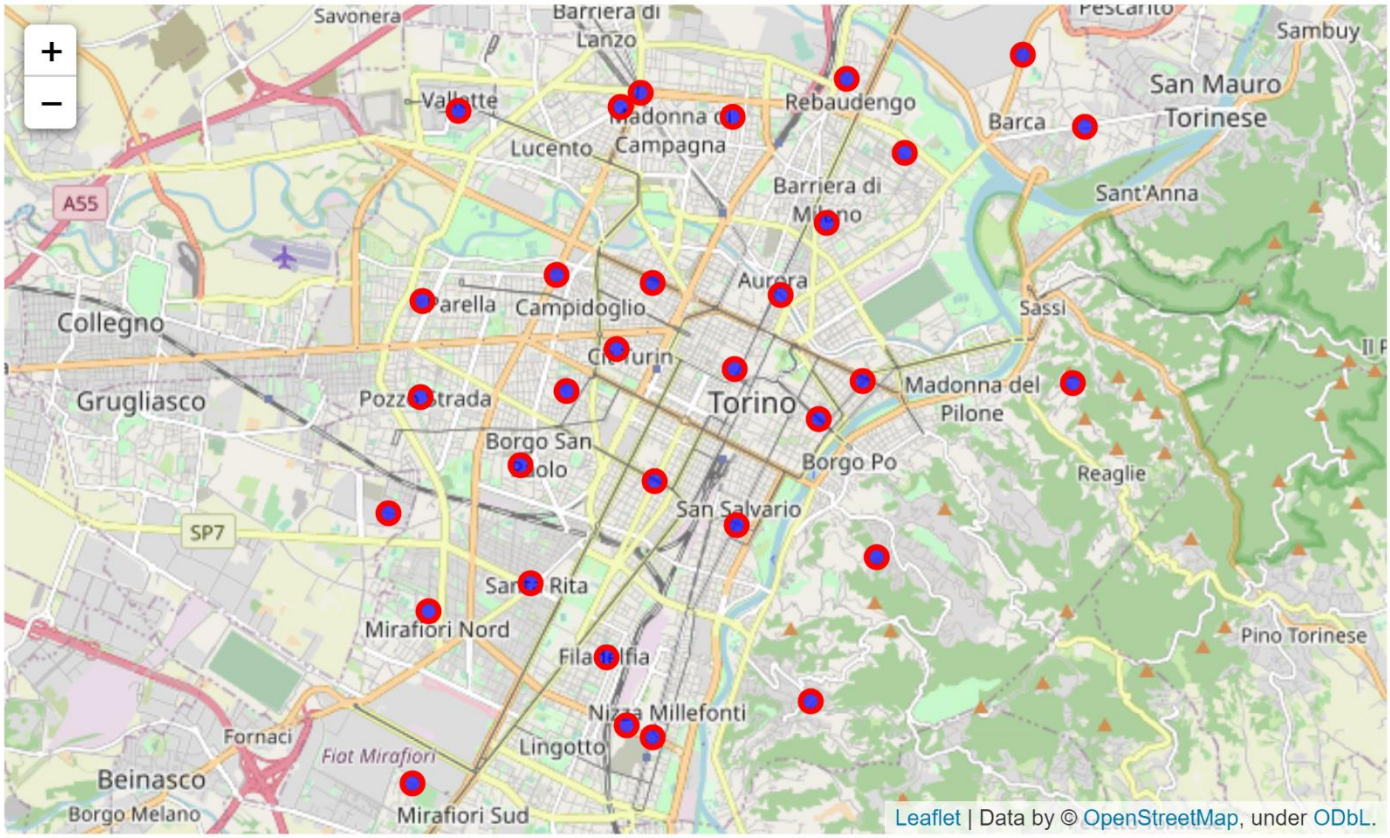

Figure 1. Map of Turin with zonal centroids.

Foursqare API allows me to find some specific venues around some location. Thus, I am able to find venues around every centroid. But since I need to mention radius, I prefer to use higher value in order to avoid some lack of data. However, it is obvious that I will get a lot of duplicated values, since radius of one centroid probably will overlap with other one. After requesting all venues resulting dataframe had the shape [760, 3]. I performed cleaning and got 533 duplicates, my final venues dataframe had the shape [227, 3], which means that we have 227 restaurants with query = ‘pizza’, or probably – pizzerias.

Index	name	latitude	longitude
0	Mister Food Pizza	45.07520871	7.67539729
1	Pizza al taglio	45.07634629	7.670114028
2	Da Aydin Pizza Kebap Bar (Da AydÄ±n)	45.06338399	7.677842081
3	Ciro Pizza & Birra	45.06653633	7.692859613
4	Taglio - La pizza per fetta	45.07341175	7.682489925
5	Pizza menÃ¹	45.07230075	7.682812214
6	Barbaroux Pizza	45.07149774	7.681399326
7	Mido pizza + kebap	45.070401	7.680659
8	Pizza & Cozze	45.069733	7.683348
9	Focacce Pizza Al Taglio	45.07212206	7.682671899
10	Pizza MenÃ¹ Di Licata Domenico	45.07543	7.682868
11	Amata'S Pizza	45.0695405	7.6786769
12	Pizza & Brioches	45.07486981	7.674670023

Table 2. Example of dataframe with venues name and coordinates.

If we plot results on the map:

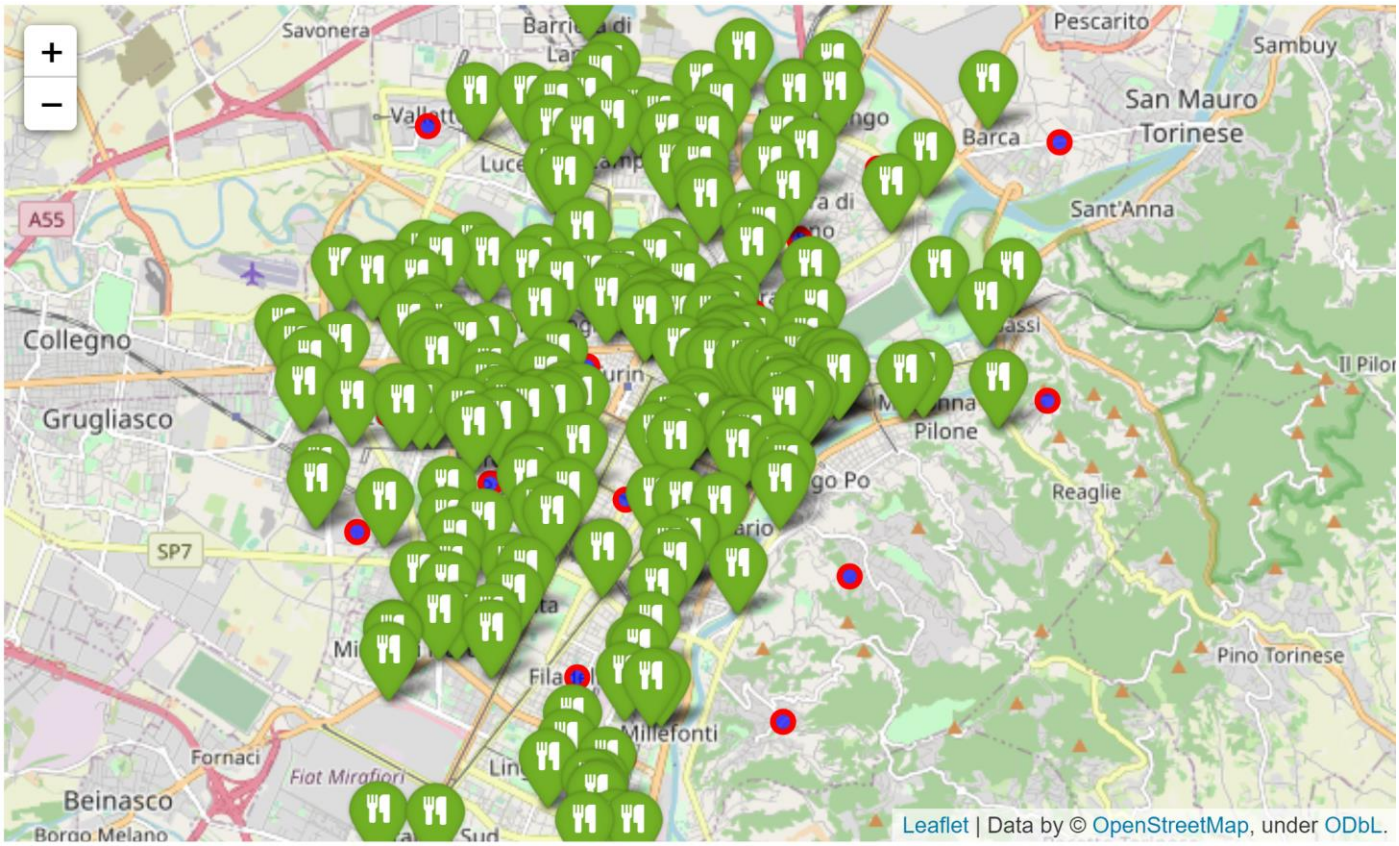

Figure 2. All pizzerias in Turin.

Wow!.. All city is hidden under icons of pizzerias now! Seems to be impossible build some business in this sector, but let's not rush.

My next step was based on the very trivial idea: if I would choose a good place for my restaurant, one of the most important criteria for me will be avoidance of another restaurant nearby in the same category (if it is some Chinese restaurant, then it does not really matter). So I decided to cluster zones according to how their served by nearby restaurants. With other words – instead of K-Means clustering, which solves mostly the problem of centroid allocation, I will use DBSCAN algorithm, which based on density clustering. But this will be only first step, because after I will need to cover resulting clusters into appropriate polygons on the map… And voila! Empty spaces between those clusters will be my first input for making decision. Actually, this is the main objective of this paper.

In order to perform clustering I used Sklearn library, especially its module called ‘cluster’. The main parameters of DBSCAN are 'eps' and 'min_samples'. 'eps' stands for maximum distance between two samples. 'min_samples' - the number of samples in a neighbourhood for a point to be considered as a core point. The core point definition is based on the principle of DBSCAN algorithm. Point is considered as core point if some minimum number of points including itself are in the neighbourhood with this point. Otherwise, if point in neighbourhood with core point, but cannot itself considered as core point, the it will be considered as border point.

I created iterative algorithm to search optimal value of eps. For this I made my first assumption: I thought that even 2 restaurants nearby should be considered as cluster, because they already can be considered as competitive business area for my new venue. So min_samples = 2. Second assumption was for the ‘best’ epsilon search - I could be wrong, but I assumed that the best 'eps' will be such distance, when I will have as less

as possible outliers of my model, and as high as possible clusters. Outliers are mentioned as '-1' in the output of the model. Results of algorithm are shown below:

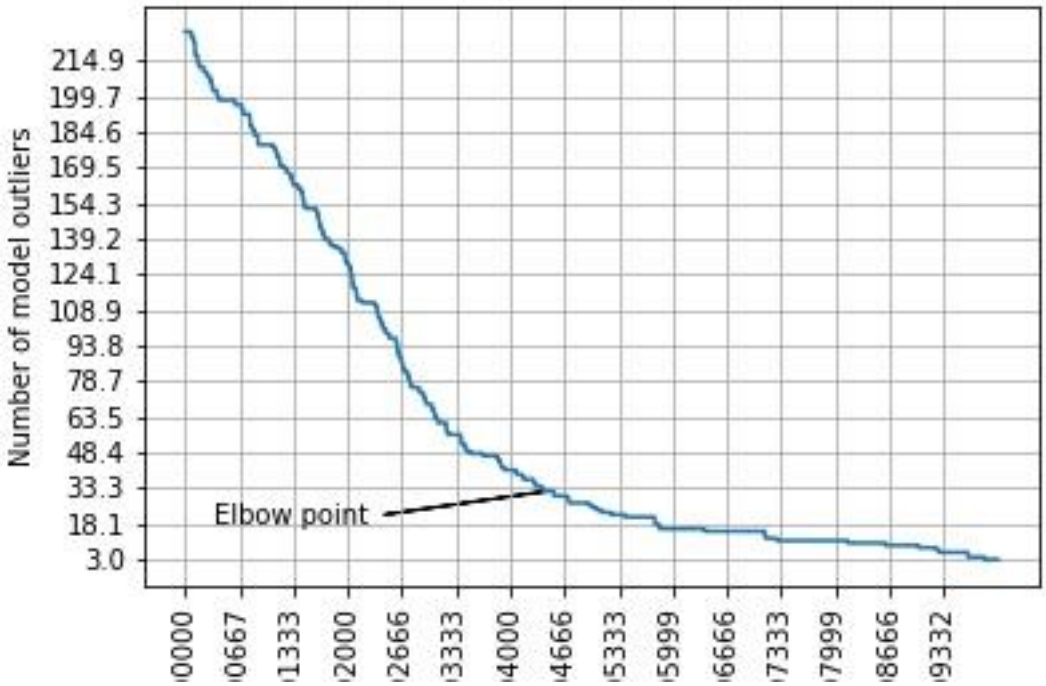

Figure 3. Model errors are decreasing with respect to decrease in the number of clusters.

On the graph we can see, that after some elbow point outliers are decreasing slightly. Therefore we can select this value of epsilon. I rounded it up to 0.005.

I coloured my venues according to their cluster number. Lets see, how it looks on the map (I increased map to show some example of clusters):

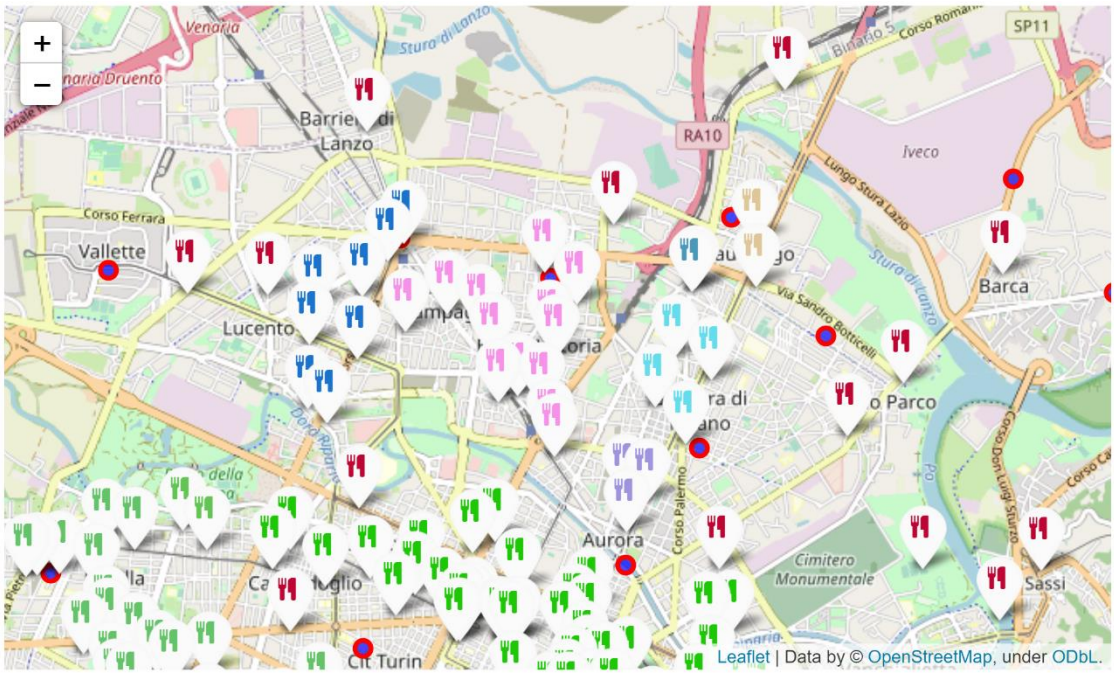

Figure 4. Results of DBSCAN clustering. Outliers are also shown with separate colour.

Good, but not excellent – what is the area of influence, or with other words – where I must not start my business? Here comes ‘alphashape’ library, which allow to create special shape of polygon of the map – alpha shape. According to Wikipedia, alpha shape, or α-shape, is a family of piecewise linear simple curves in the Euclidean plane associated with the shape of a finite set of points. The algorithm of alpha-shape tries to create borders of the set of points in such manner, that they will perfectly fit the shape of this set.

I used this algorithm to plot with folium perfect polygons of clusters without venues, except outlier pizzerias. Thus, I achieved my final goal – I am able to see on the map areas of city, that are less occupied by pizzerias:

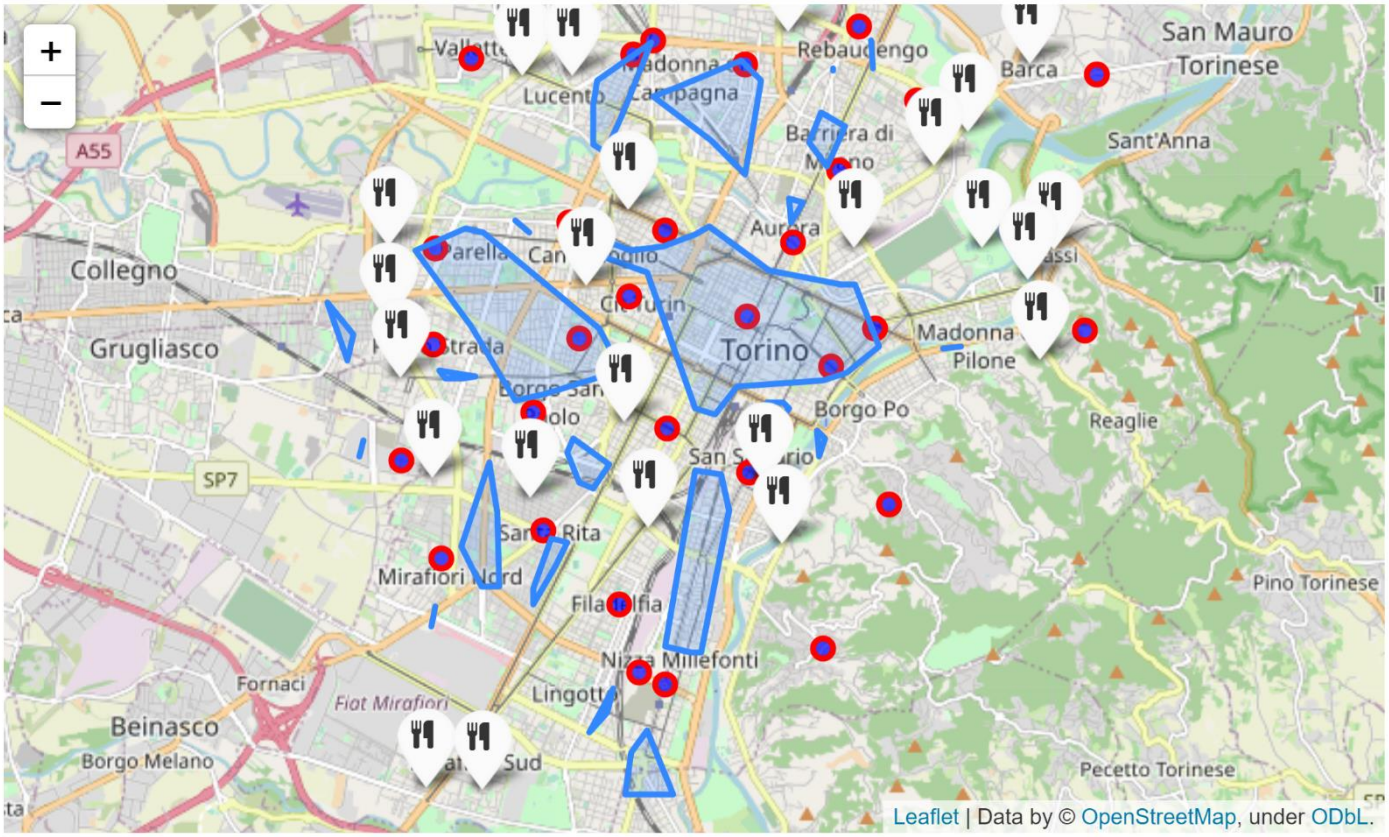

Figure 5. Blue polygons represent areas, occupied by pizzerias. Icons are outliers of our model, however we also need to see them for decision making process.

3. Result.

There are few areas in Turin, less occupied by pizzerias. But one among them is especially catches my attention: area between Lingotto and Filadelfia. Last years, Turin was dynamically growing towards south, new metro stations were opened and household prices started to increase in this zone:

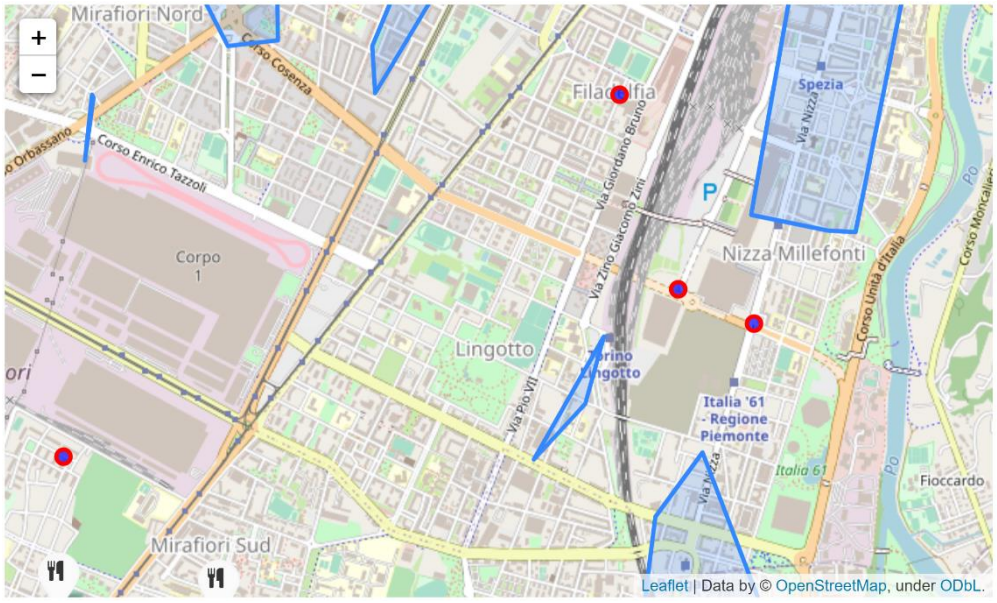

Figure 6. The area I suggest for further analysis.

4. Short discussion

As I told before, I considered my analysis as a first step for choosing the best location of the business, though in real life this case will need additional complex plan, such as search of premises available for rent, rental prices, number of households, and maybe some additional surveys.

I could add some locations of available places for rent, but unfortunately there is not enough data about that for Turin. In such case I can built more powerful and complex search tool using Google Search, but it requires some time.

5. Conclusion.

This project was aimed to show simple and real-life usage of DBSCAN algorithm and some useful libraries for geospatial analysis. There are still a lot of ways to improve and achieve more precise results for the final goal of new business allocation.

Thank you for your attention and patience, definitely invite me to your restaurant, if you will open one in Turin!

Sincerely yours,

Arif Huseynov.