Selection of the best place to open a lab

Project for IBM applied Data Science capstone
Giovanni Manfredi 2020/10/19

Introduction

For the final assignment of the IBM applied Data Science course on Coursera, I decided to explore a bit the area of Genova, the city where I live. The problem I invented is the following:

A health company is trying to open a new laboratory that will work on the development of new pharmaceuticals in Genova. In order to do so, they have to find the place that is best suited for their lab. The place they will choose will require the following:

- It must be near public transportation, as Genova is notoriously a very chaotic city and moving by car is a great problem for employees.
- It must be near universities as the lab will exploit synergies with academic institutes and will offer opportunities for master students that are looking for thesis.
- It must be near health institutes to exploit synergy with them.
- It must not cost too much.

The project will develop the request of this hypothetical health company and answer the question giving the best places where the company can open the lab.

I will do this by first exploring all the public transportation stops in genova acquired by open databases. These results will be explored using 4squares to find all the venues near the area and then completed using Italian real estate observatory data.

Data

The data which will be used is supplied by different sources. Venues data is of course offered by 4Squares as also stated by the course requirements. The data contained all the venues near a specific area comprising precise location, venue category, and venue principal information.

Public transport stops are supplied by Genova municipality at http://dati.comune.genova.it/dataset/amt-trasporto-pubblico-nel-comune-di-genova/resource/8dab9db4-c0b1-4146-8e3a-29d9c91d08ef. Each point of the dataset comprised a name of the stop, its coordinates, the municipality, and other information.

Real estate data (updated to 2019) are supplied by Italian Observatory of real estate market (Osservatorio del Mercato Immobiliare – OMI, https://www.agenziaentrate.gov.it/portale/web/guest/aree-tematiche/osservatorio-del-mercato-immobiliare-omi) and were first acquired and cleaned by the OnData civic hacking group (http://blog.ondata.it/i-dati-sulle-quotazioni-immobiliari-dellagenzia-entrate/). Each point of the dataset is related to a specific geographical zone and included a minimum purchase and rent price, a maximum purchase and rent price, plus other information.

The geographical area for the OMI data were also supplied by OMI itself and were first cleaned up by OnData civic hacking group (http://blog.ondata.it/i-dati-sulle-quotazioni-immobiliari-dellagenzia-entrate-i-poligoni-delle-zone-omi/).

Methods

To create a useful dataset, I first used the data supplied by Genova Municipality about public transport stops. Indeed, the proximity to transports requirement is automatically fulfilled checking the venues around stops. The data are quite clear with full coordinates and stop names. Figure 1 reports a picture of the map showing all the stops. They are so dense I will use them as territory mesh points.

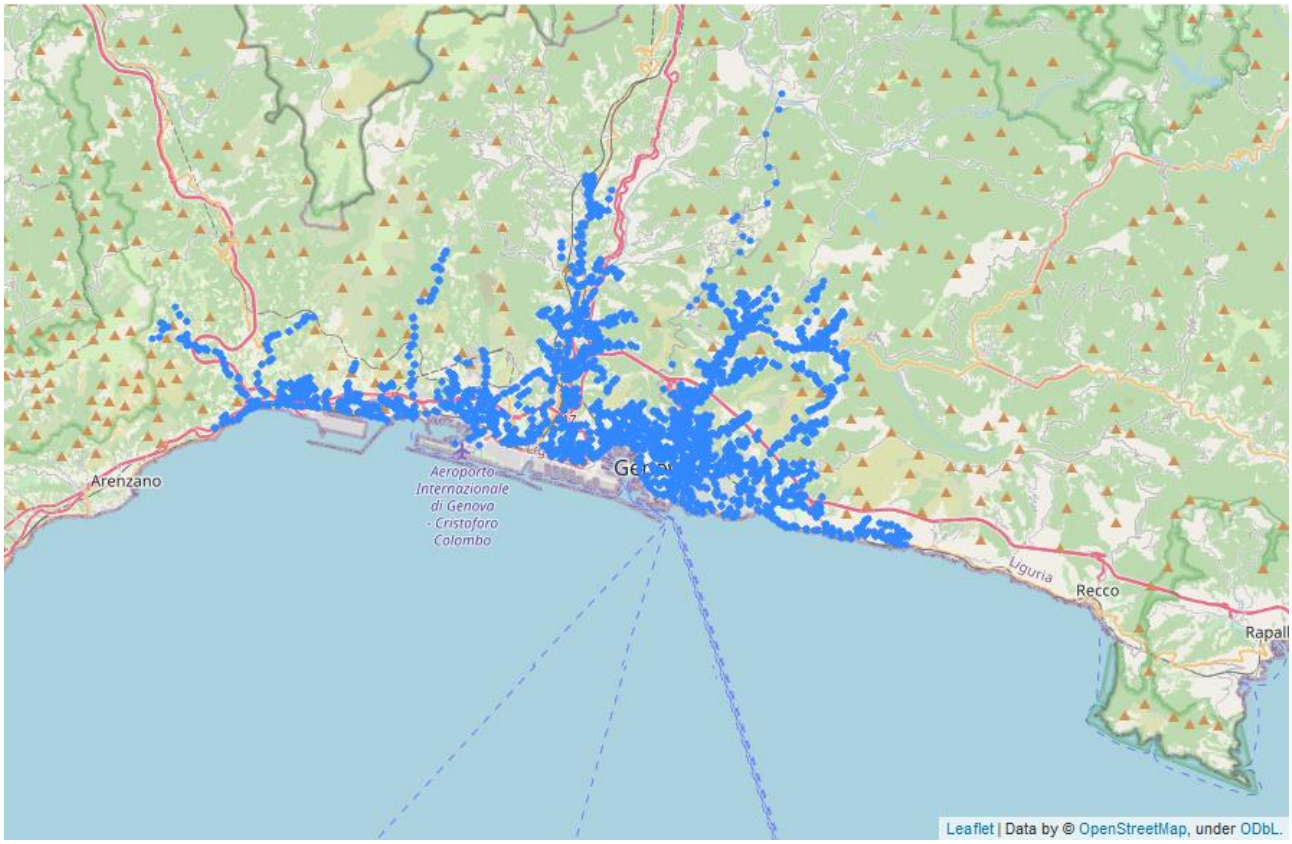

Figure 1 Public transport stops map

An estimated price of purchasing a laboratory building is obtained using OMI data as told above. The data supplied by Italian agency report a minimum and maximum price. I supposed the median price is the mean of the two (probably not true but not so distant I guess).

For each transport stop I extracted the number of venues of specific categories using Foursquare. This task was quite delicate. Foursquare gives the possibility to use explore or search. If I use explore as suggested in all exercises in the course, the result is capped to 50 and Foursquare decide by itself the type of the venues reported. Since I have no control over the selection algorithm, I have no idea if the data will be biased or not. To avoid this, I used a call to Foursquare API for each category. The idea is to never have a result over 50 venues per call. This grants the absence of biases. I call the API for each one of the following categories:

- Entertainment
- University
- School
- Food
- Nightlife
- Work
- Government
- Conventions
- Fitness
- Outdoor
- Health
- Residence
- Shops
- Hotels
- Airport

- Transports

Much of these categories will not be used for the main goal of the problem but they might be interesting for future evolutions or for you reader to play if you want.

Once the venues have been extracted, they have been used to count the number of venues of each category that are no more distant then 500 m from a transportation stop.

Using the coordinates of each stop, the mid price of laboratory buildings in the area has been found and added to the dataset.

The total number of venues for each stop has been calculated.

In the end, the final dataset is composed by 3053 rows and 22 columns.

An exploratory analysis of stops clustering has been conducted using a KNN algorithm to have an idea about the goodness of the data. The clustering have been conducted not on the numbers of venues but on the fraction of venues respect to their total number. For example, if a stop has 2 Food venues and 2 Entertainment venues, it has then 50% Food, 50% Entertainment and 4 total. The cost of buildings has been normalized 0 to 1 before clustering and the total number of venues has been ignored (it was included in a first version but excluded as it basically distinguish by itself central and peripheral areas). All samples having less than 50 total venues have been excluded by clustering and considered as a special cluster of points having not enough data.

The finding of the best place to build the lab is conducted giving a score to the number of venues present in the area and normalizing by a function of the price.

Venue number Is summed using the following weight vector:

Airports	0.5
Conventions	0
Entertainment	0
Fitness	0
Food	1
Government	0
Health	10
Hotels	0.5
Nightlife	0
Outdoor activities	0
Residence	1
School	1
Shops	0.5
Transports	2
University	7
Work activities	1
Entertainment	0

The sum is then divided by the logarithm of price (places with price = 0 are considered having a score of 0 as basically they are not reliiable).

The areas with the best score are then calculated.

Results and discussion

The bus stops on the Genova territory represent a good mesh of the same. I used them to map the territory. I was not able to retrieve a highly detailed map of real estate prices but using OMI zones is a good approximation (Figure 2).

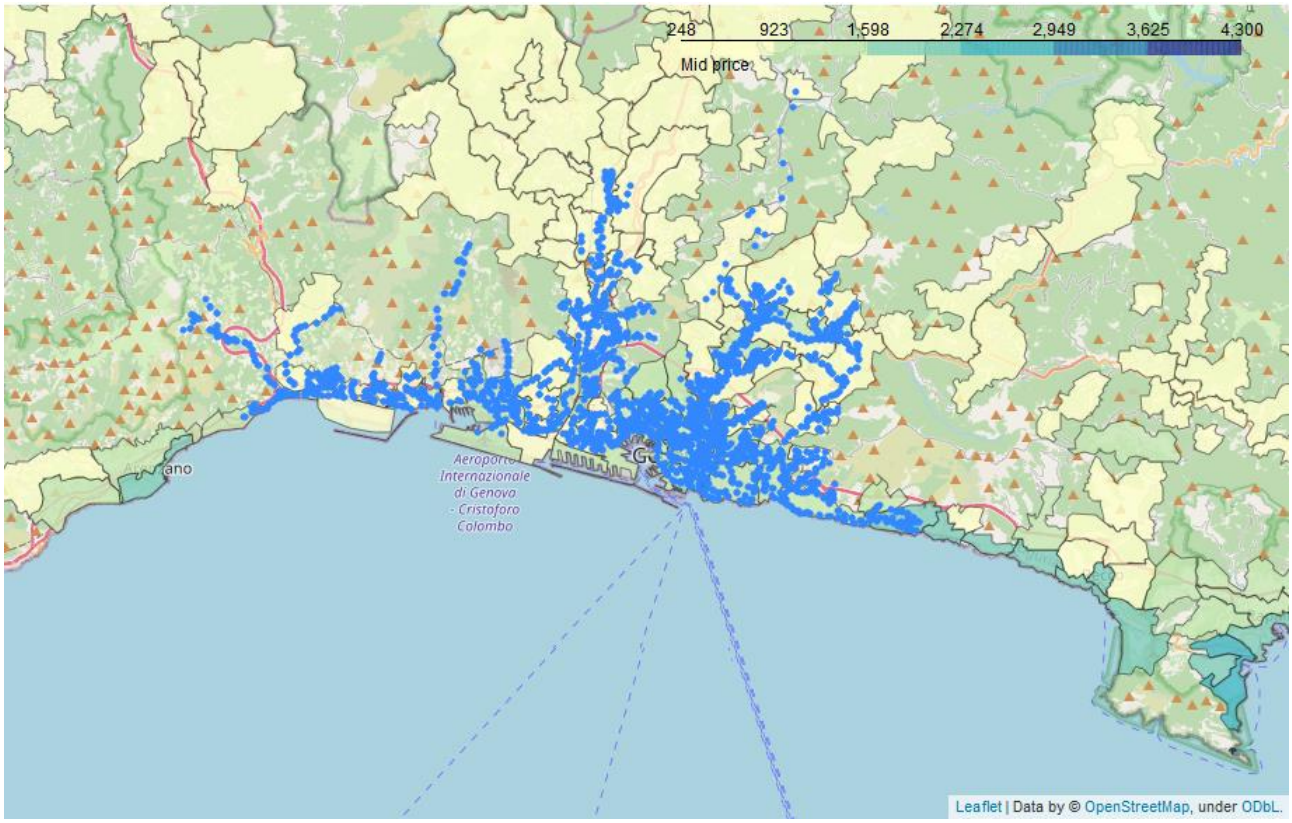

Figure 2 Map of Genova transportation stops and price range areas shown as choropleth map.

An idea to improve the resolution of real estate prices is to scrape data from real estate trading sites like immobiliare.it, casa.it, etc. In this work I reduced the resolution of this data but it could be a future augmentation. To explore the data I looked at clustering in the composition of the venues at the various stops. I excluded stops having a too small total number of venues (at least a total number of 50 venues) as they are not reliable enough to be considered.

The clustering was quite astonishing, it individuated the old town centre from Brignole to San Teodoro, it clustered Foce residential area by itself, it clustered Marassi and Righi as the same kind of place (they indeed are mid class residential areas). Sampierdarena and Cornigliano are low-class residential areas and are clustered by themselves. San Martino hospital distinguish itself as well as the area of Borgoratti/Sturla which are peripheral residential areas. Astonishing result knowing living in Genova and knowing its quarters.

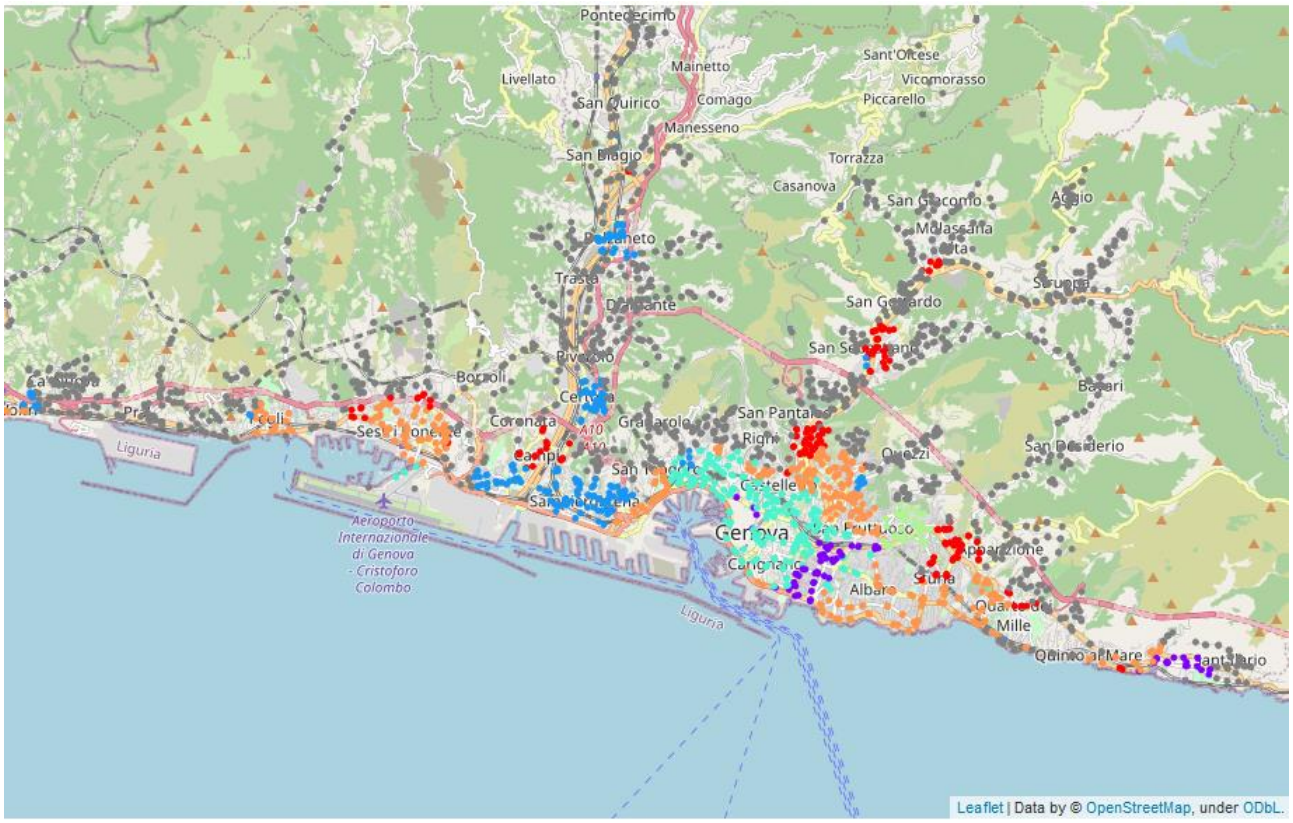

Figure 3 Clustering of Genova public transportation stops. The different colors represent different clusters.

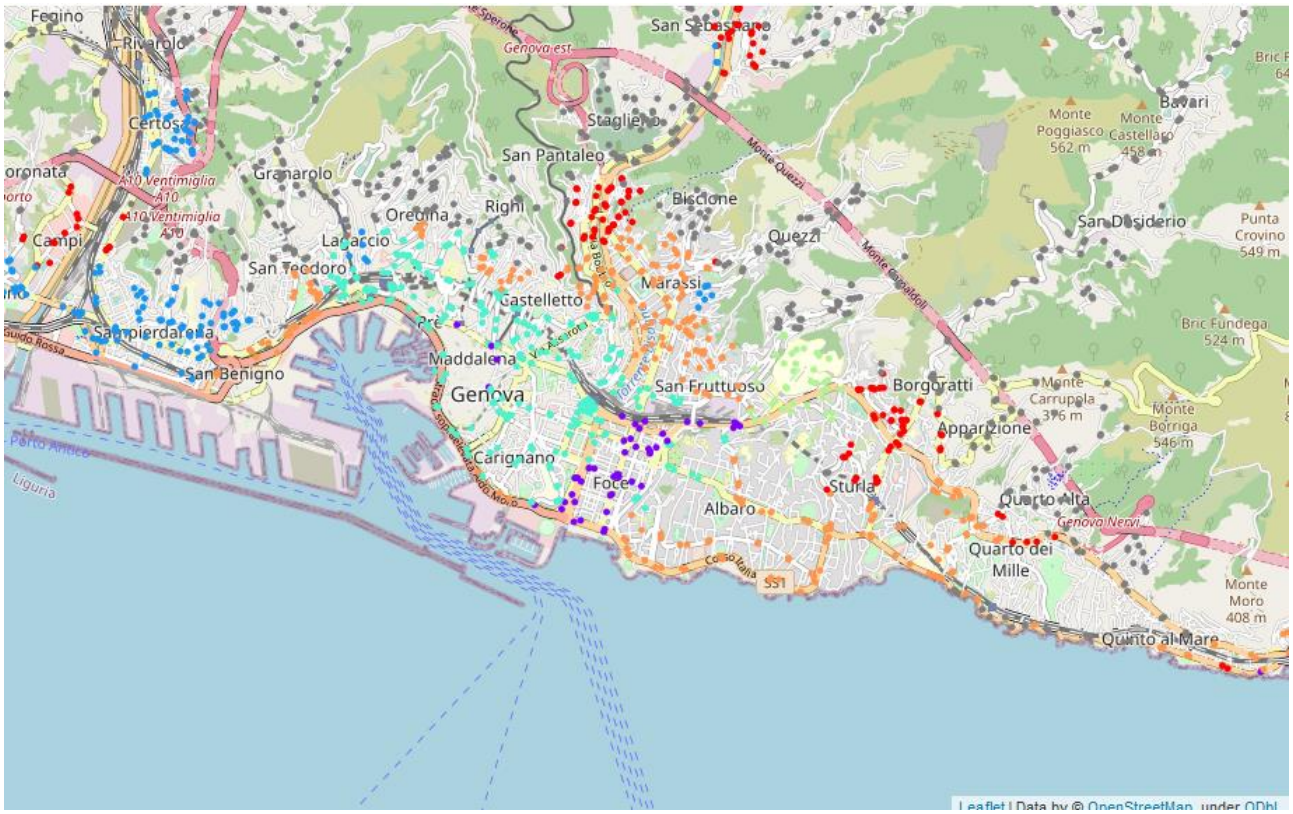

Figure 4 Map of Figure 3 at different zoom level.

The best place to construct a lab has been found giving a score to each transportation stop. The score is given according to the procedure explained in the method section and takes in account all the venues around the transport stop and the mid price of estates in the correspondent OMI area.

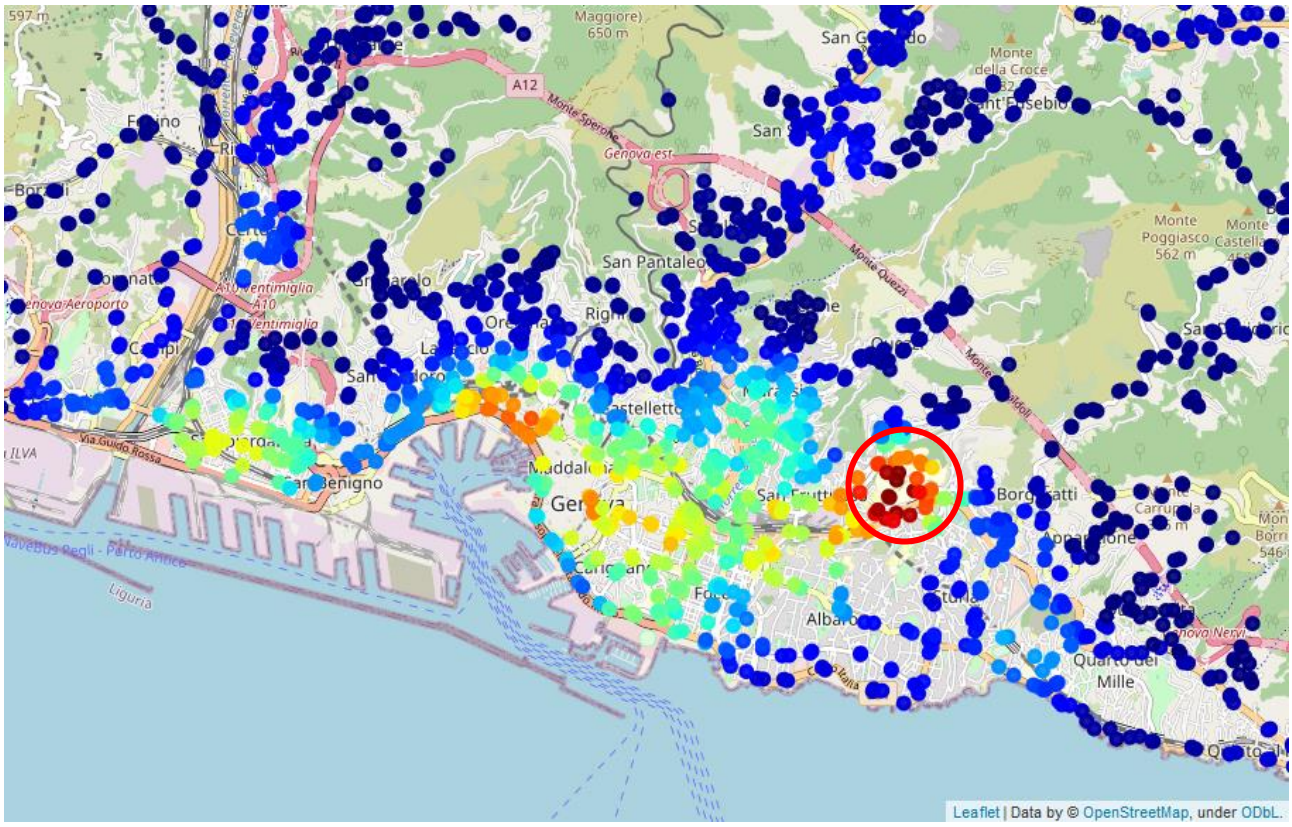

Figure 5 Map showing the suitability score of each point in the map. The colors represent the score. Blue means low score, red means high score. The red circle indicates the best area to build.

Visual inspection of the data on the map clearly suggest that, according to the score metric, the best place to construct the lab is in the San Martino area.

Conclusion

In conclusion in this assignment I collected the data about Genova metropolitan area public transport stops. I used them as a mesh to find the distribution of venues of different categories on the territory and finally evaluated the place which is best suited to build the lab according to the score reported above. Possible evolutions of the work consist in:

- Enhance real estate prices resolution scraping data from specialized sites.
- Enhance the quality of collected data separating venue acquisition phase from assignment to transport stop.
- Improve the definition of the score to find the best place to build a laboratory.
- In the optic of using the method in a real environment, there is the possibility to study every place where the pharma company already has a lab, train the score parameters on every city and then apply it again on Genova.