Coursera:
Applied data science capstone
Week 5: Data exploitation

Date 20/12/2020

INTRODUCTION

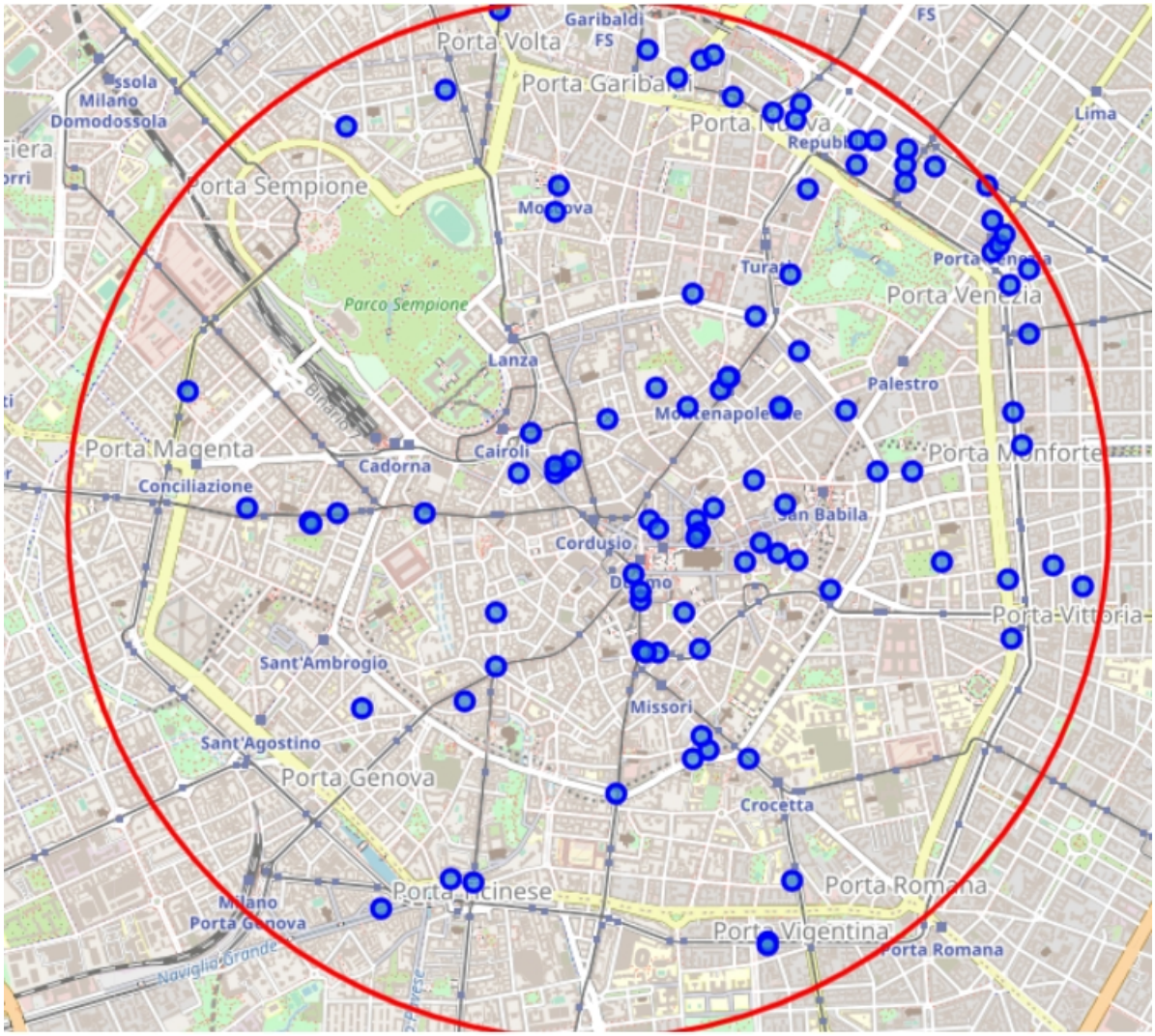
A Taxy company wants to set up its business in Milano city, Italy. The business idea is to be affiliated with all the Hotels in the radius of 2 km from the city center so as to exploit their smart mobility transport service solutions. They want to offer the fastest pick up service to be as effective and fast as possible. For this reason they are looking for the perfect positioning for their headquartier. The goal is to find the place that minimize the average distance from all the Hotels in the very citycenter. Assuming an average speed of 10 km per hour to move in the city (indeed a conservative value (https://www.statista.com/statistics/264703/average-speed-in-europes-15-most-congested-cities/) and neglecting all mobility restrictions and all traffic congestion issues we want to define an estimation time for a taxi service to reach each Hotel also.

DATA COLLECTION

We make use of data retrived through Fouresquare API to get the cooridnates for every hotel (https://developer.foursquare.com/developer/). We request results having *the tag 'Hotel'* in the Foursquare API and located within a radius of 1500 meters from Cordusio square (https://en.wikipedia.org/wiki/Piazza_Cordusio), this distance is representative of the restricted-traffic zone of the city, the center decimal coordinates are Latitude = 45.465586 and Longitude = 9.185944. Notice that using a free developer account of Foursquare we retrive up to 100 places only so using the 'query' tag is particularly useful in since we have not to filter every type of venues. This provides us a list of 100 Hotels in the area with their coordinates as shown in the next table where the first 10 results are shown.

	name	categories	lat	lng
0	Park Hyatt Milan	Hotel	45.465532	9.188911
1	Room Mate Giulia Hotel	Hotel	45.465250	9.189396
2	BVLGARI Hotel Milano	Hotel	45.470149	9.189318
3	Mandarin Oriental	Hotel	45.469461	9.190876
4	Armani Hotel Milano	Hotel	45.470478	9.192882
5	Starhotels Rosa Grand	Hotel	45.464122	9.193692
6	Four Seasons Hotel Milano	Hotel	45.469372	9.195466
7	The Square	Hotel	45.461003	9.189338
8	HMS Hotel Milano Scala	Hotel	45.469061	9.186865
9	Hotel Spadari al Duomo	Hotel	45.463738	9.187130

Finally we can depict all the Hotels on the Milano city map, we also draw the circle defining the considered urban area.



METHODOLOGY

We want to use the data to define the best location for the headquartier. Lets *x* and *y* be our the Headquartier coordianates, and lets *Hxi* and *Hyi* the coordinates of each Hotel *i*. We can express the average distance d from the headquartier to the hotels using the formula

$$d = \frac{1}{N} \sum_{i}^{N} \sqrt{(x - H_x^i)^2 + (y - H_y^i)^2}$$

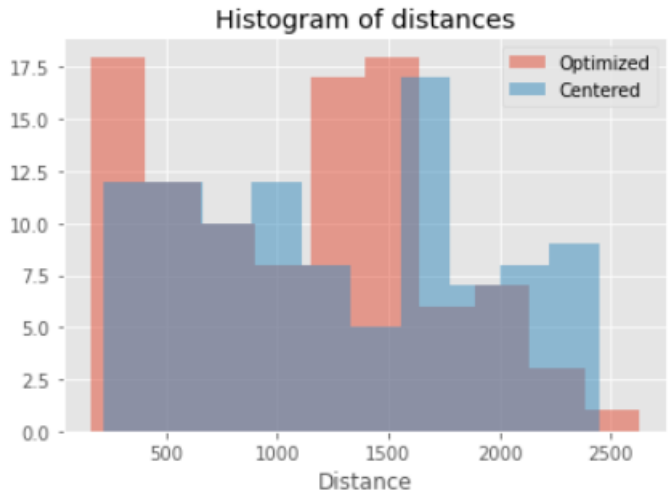
We can use an operational resarch approach to find the optimal solution for the minimization problem of finding the Headquartier coordinates x and y so as that d is minimized. We can do it in Python using it Scipy libraries to numerically solve the problem.

RESULTS

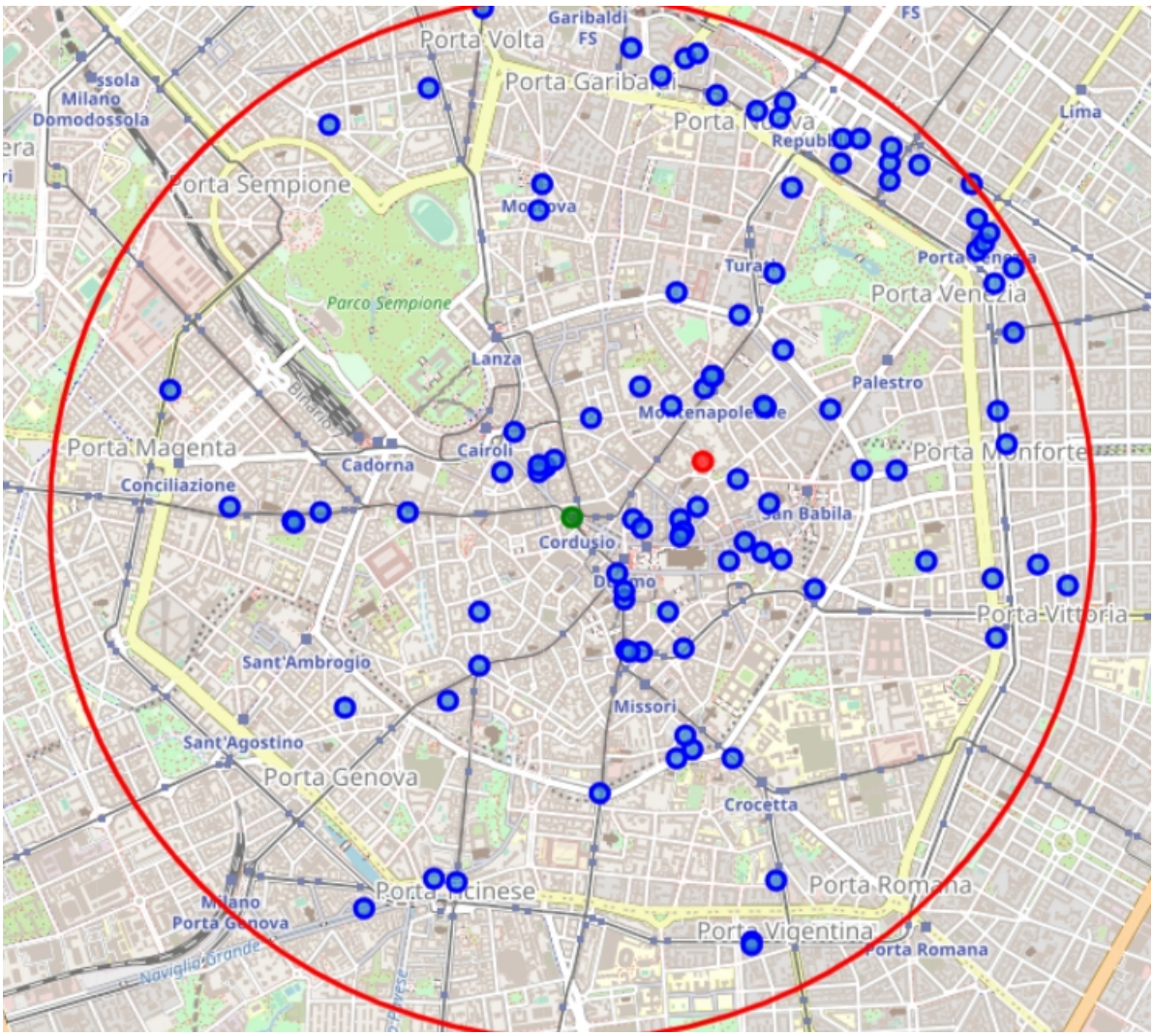
The coordinates of the optimal position are Latitude 45.46756174 and Longitude 9.19235717 that allows a minimun average distance of 1102 meters. The next table present an extract of the calculated distances

	name	distance
0	Park Hyatt Milan	399.93
1	Room Mate Giulia Hotel	375.67
2	BVLGARI Hotel Milano	399.13
3	Mandarin Oriental	240.82
4	Armani Hotel Milano	296.29
5	Four Seasons Hotel Milano	359.68
6	Starhotels Rosa Grand	368.93
7	The Square	721.98
8	HMS Hotel Milano Scala	569.28
9	Palazzo Parigi	594.91

Notice that, considering an average speed of 10 km per hour the average travelling time is about 7 minutes while we can calculate de estimated travelling time for every hotel. Though the result may seems naive we can confrontate the fact that placing the headquartier in the geographical city center may not be the most conveniet choice. The histogram below shows the Hotels distances in respect to both the minimum average location and the geographical centered location.



We notice that for the optimized case the mass of the histogram is more concentrated around the average value, with a spike on the left (i.e. closer positions) and that it presents a lighter right tail. Numerically, the optimized placement allows an average distance of about 200 meter less in respect to the geographical center. Though this may seem negligible it may be not while considering a consistent amount of calls.



In the picture the Red spot is the optimized placement while the red one is the geographical center.

DISCUSSION

The best place to set a fast taxi service for hotels in Milano city center is not in the real center of the city. The result is not surprising, we can see from the map that the hotels are located in the north-west and the center area, hovewer this conclusion is more solid when supported by data. In a more precise study one should take into account road distance instead of geographical distances, traffic limitations and congestion should be considered too. Another important thing one could consider is the number of calls each hotel makes in order to weight their importance: the latter implementation is straightforward by adding a weighting coefficient Wi to the formula to represent the number of calls associated to each hotel.

$$d = \frac{1}{N} \sum_{i}^{N} W_i \sqrt{(x - H_x^i)^2 + (y - H_y^i)^2}$$

CONCLUSIONS

It has been shown that is possible to use data and python to solve a particular problem that won't be easly solved differently. In this case we used hotel location coordinates to get conclusions about the optimal positioning of a shared service. The same approach can be used for more particular use cases and can be more and more effective the more rich data one can use. Machine learning techniques can be used to derive conclusions and optimization can be used to get numerical results out of cumbersome problems.