IBM Data Science Capstone Project

Analyzing and Clustering Neighborhoods of Milano, Italy

Author: Davide Miglietta

Introduction

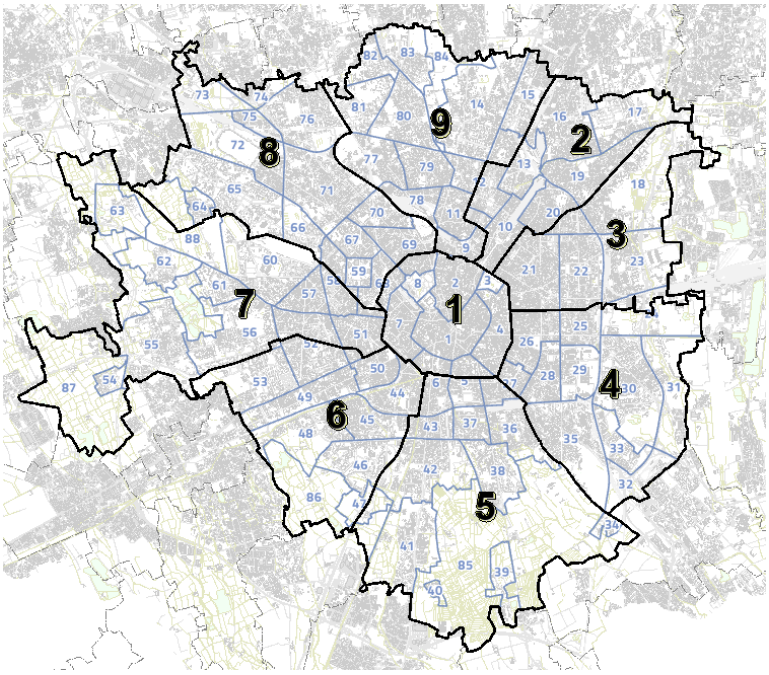
Scope of the project

The scope of this project is to explore the city of Milano and its neighborhoods, by crossing data and information about parks, green area, and venues. The goal is to investigate and find the "most livable" parts of the city. The basic idea is that a neighborhood is more livable and enjoyable by its inhabitants if has a wide variety of services and venues (such as restaurants, bars, theatres) and a big extension of parks and green areas.

Background

Milano is one of the most important cities in Italy, one of the richest and most populated in the country. The city itself counts a population over than 1.3M, while the metropolitan area has more than 3M inhabitants. It is considered a global city with strengths in many fields such as commerce, finance, fashion, education, and many others. In this context this project aims to be a first exploratory analysis to indicate the best areas where to live in the city and also find insights that may lead to actions for improving the others.

Milano is divided in 9 "Municipi" (districts) and 88 "Nuclei di identità locale (NIL)" (neighborhoods), such as represented in the picture.



Data and Data Sources

Data for the analysis have been found on:

- Foursquare, where data containing information about venues has been scraped. The information is about the name, geographical location and category of the venue.
- https://dati.comune.milano.it (official website of the "Comune di Milano"), where it was found data about geographical location and dimension of Municipi, NIL, dog areas and parks.

As first step of the work it has been necessary manipulate the data found at https://dati.comune.milano.it to organize it in dataset with aggregate information useful to the scope. In particular:

-
- For NIL and Municipi, 3 different dataset have been merged, aggregated and cleaned to obtain a more useful one, shown.

Out[60]:

ID_NIL	NIL	NIL_Long	NIL_Lat	NIL_Area_mq	MUN
1	DUOMO	9.186948	45.463707	2.341704e+06	1
2	BRERA	9.188157	45.474252	1.637395e+06	1
3	GIARDINI P.TA VENEZIA	9.200231	45.474564	2.496468e+05	1
4	GUASTALLA	9.201891	45.463219	1.548021e+06	1
5	PORTA VIGENTINA - PORTA LODOVICA	9.192446	45.450950	1.135239e+06	1

- Similar manipulation has been performed on parks and green areas data, to obtain:

Out[18]:

	MUN	park_area_mq	park_name	long_parks	lat_parks
0	6	49230.077148	PARCO DELLE CROCEROSSINE	9.123539	45.450540
1	9	1451.261719	GIARDINO VIA PORRO JENNER	9.179612	45.496733
2	1	351.915039	GIARDINO ROBERTO BAZLEN	9.197675	45.453966
3	2	973.018555	GIARDINO ALDO PROTTI	9.200186	45.493943
4	7	1640.686523	PARCO ANNARUMMA	9.118195	45.460160

Information about venues scraped from Foursquare:

In [176]:
```
1 print(milano_venues.shape)
2 milano_venues.head()
```

```
(2212, 5)
```

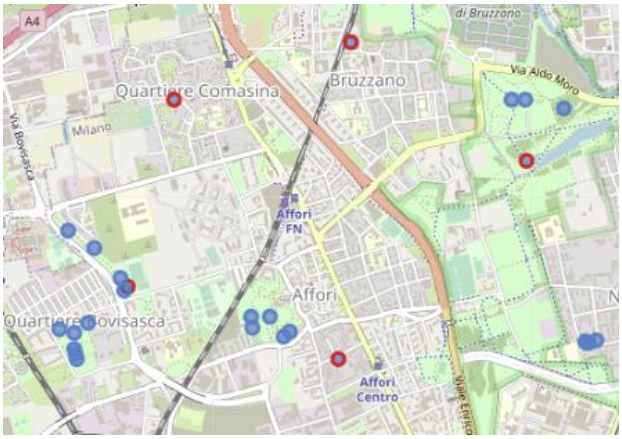
Out[176]:

	NIL	Venue	Venue Latitude	Venue Longitude	Venue Category
0	DUOMO	Piz	45.462163	9.185767	Pizza Place
1	DUOMO	Starbucks Reserve Roastery	45.464920	9.186153	Coffee Shop
2	DUOMO	Piazza del Duomo	45.464190	9.189527	Plaza
3	DUOMO	Ciacco. Gelato senz'altro	45.463704	9.186796	Ice Cream Shop
4	DUOMO	Venchi	45.465214	9.187340	Ice Cream Shop

Methodology and data manipulation

The analysis was performed in Jupyter Notebook, using Python as programming language.

As mentioned, the very first step was to find data and manipulate it to have more useful and manageable ones. For instance, information about NIL coordinates and area were on different datasets; also, the "match" between NIL and Municipi was in a third dataset. From the three datasets were extract the needed information and aggregate in the previously shown dataframe. To check this phase success, a plot of the city using Folium has been performed. The goal was to check the correct position of the NIL using the given coordinates and correct assignation of the NIL to the Municipio.



After that it has been processed data about parks and green areas. The data found had aggregate information about parks, their location and to which Municipio they belong. However, the goal of the project was to analyze more in detail the neighborhoods instead of the districts. For this reason, using geographical location of the parks the have been assigned to a NIL. To do this, a KNN has been used, with K=1. The idea is that a park belongs to the closest NIL (of which are available coordinates of the centroid). The park themselves were already divided in different parts (each part with its area and coordinates). That resulted to be particularly useful since, in this, way bigger parks "shared" between neighborhoods have been assigned automatically to different NIL. The result was quite accurate and only very few adjustments have been performed. A map of the city, using Folium, have been produced to compare park's location and assignation to the NIL.

After that, with the data manipulated it has been possible make a first classification and clustering of the neighborhoods. It has been calculated the number of parks, the extension of all the green areas compared to the total area of the NIL and the extension of the bigger park in the neighborhood. In the dataframe shown below it is possible to see also information about dog areas. However, for this specific classification, it has not been used.

Out[1949]:

ID_NIL	NIL_Area_mq	num_dog_area	bigger_dog_area_mq	%_dog_area	num_of_parks	bigger_park_mq	%_parks_area
1	2.341704e+06	1.0	1567.106991	0.066922	1.0	106.402832	0.004544
2	1.637395e+06	2.0	1256.636866	0.083270	1.0	4665.222656	0.284917
3	2.496468e+05	2.0	7942.425550	5.046969	2.0	192970.566406	85.020895
4	1.548021e+06	3.0	549.244339	0.064384	2.0	33585.192383	2.634846
5	1.135239e+06	6.0	4703.437904	0.927938	4.0	73821.117188	8.532211

The necessary data (left) has been scaled using StandardScaler (right).

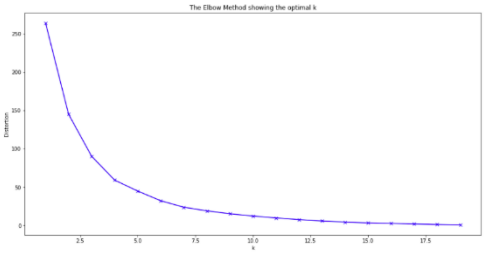
Out[1951]:

ID_NIL	num_of_parks	bigger_park_mq	%_parks_area
1	1.0	106.402832	0.004544
2	1.0	4665.222656	0.284917
3	2.0	192970.566406	85.020895
4	2.0	33585.192383	2.634846
5	4.0	73821.117188	8.532211

Out[1956]:

ID_NIL	num_of_parks	bigger_park_mq	%_parks_area
1	-0.021115	-0.498672	-0.425301
2	-0.021115	-0.468588	-0.407084
3	0.907933	0.774061	5.098584
4	0.907933	-0.277741	-0.254399
5	2.766029	-0.012220	0.128779

K-mean classification has been performed and the correct K has been chosen using the elbow method. The K picked was 7. So, the results is to divide the city in 7 clusters by the previously shown “green variables”. For example, a cluster obtained was the following:

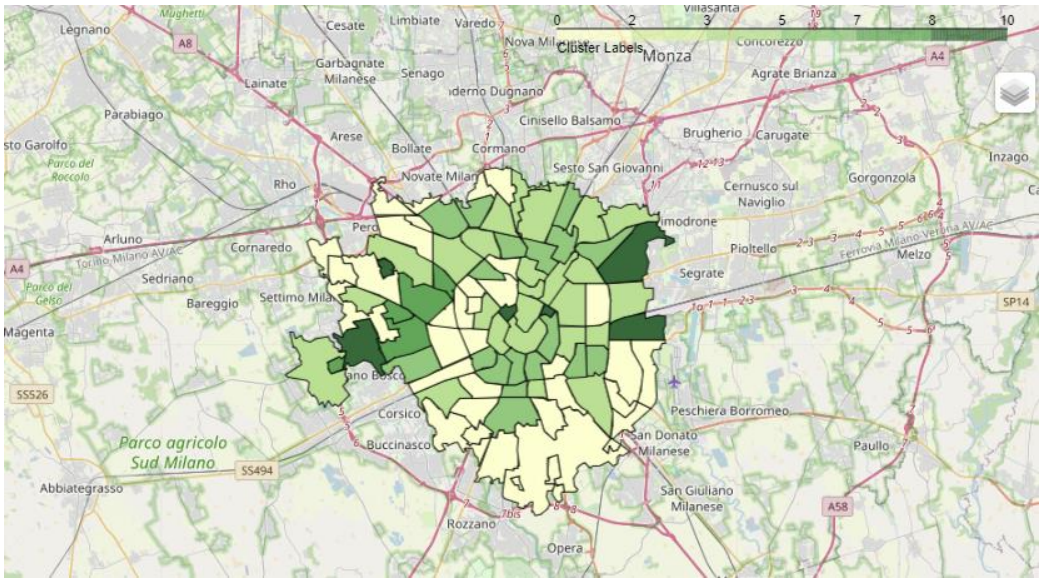


In [1966]:
```
1 cluster0 #bigger parks in the city, big % of the NIL is green - choropleth label=8
```

Out[1966]:

ID_NIL	Cluster Labels	num_of_parks	bigger_park_mq	%_parks_area
18	0	1.0	773404.921875	15.555303
24	0	2.0	545306.854492	22.274588
55	0	4.0	852748.097656	28.018331

Each cluster is representing a group of neighborhoods with less or more “green characteristics”. For this reason, a “mark” (choropleth label) has been assigned to each one. The higher the label, the greener the cluster. That was necessary for representing the insights obtained with a choropleth map, shown below. The darker the green, the greener the NIL. To make this map, a .geojson file with geographical coordinates has been used.



To complete the analysis, the obtained data from Foursquare has been explored and analyzed.

```
1 milano_venues['NIL'].value_counts()
```

GIARDINI P.TA VENEZIA	100
PORTA TICINESE - CONCA DEL NAVIGLIO	100
BRERA	95
PORTA GARIBALDI - PORTA NUOVA	94
STAZIONE CENTRALE - PONTE SEVESO	87
PTA ROMANA	79
DUOMO	75

```
1 milano_venues['Venue Category'].value_counts()
```

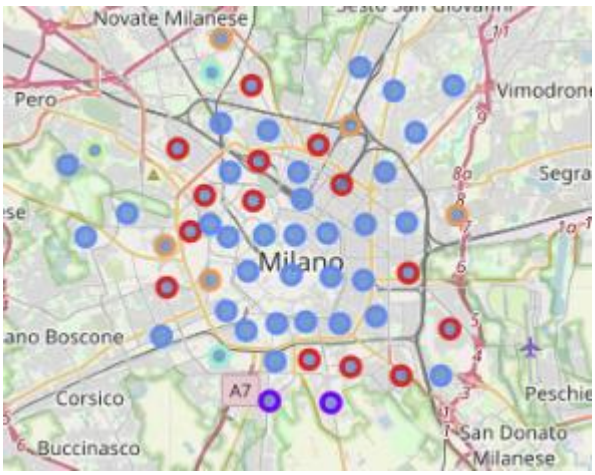
Italian Restaurant	216
Café	144
Pizza Place	124
Hotel	85
Ice Cream Shop	69
Cocktail Bar	62
Plaza	59
Supermarket	45

Out[188]:

	NIL	Art Gallery	Art Museum	Asian Restaurant	Athletics & Sports	Bakery	Bar	Bed & Breakfast	Beer Bar	Bistro	Bookstore	Boutique	Breakfast Spot	Brewery	Burger Joint	Bus Stop
0	ADRIANO	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0
1	AFFORI	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0
2	BANDE NERE	0.0	0.0	0.0	0.076923	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0
3	BICOCCA	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.041667	0.0	0.041667	0.0
4	BOVISA	0.0	0.0	0.0	0.000000	0.1	0.1	0.0	0.0	0.0	0.1	0.0	0.000000	0.1	0.000000	0.0

It has been performed a K-mean classification of the neighborhoods, but not all the data has been used. For example, the three most common venues were not included since spread all over the city and with a quite higher number compared to the others. That was not adding much information to the classification and it has been decided to look for insights “behind the surface”. The frequency of each category in each NIL has been determined. Again, the K for the classification has been determined by the elbow method and the chosen one was 6.

Six clusters have been obtained, showed in the map.



For example, Cluster number 3 (blue dots) was the one with a wider variety of venues indicating more choices for the population.

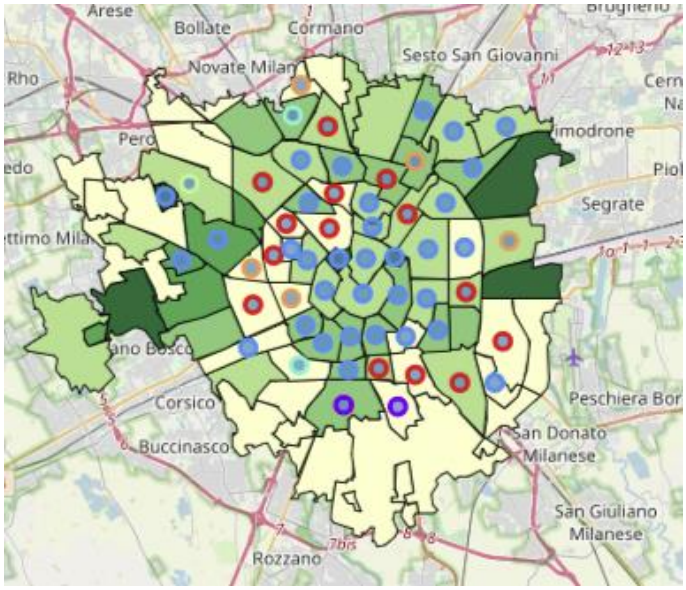
```
cluster3 #various venues
```

Cluster Labels	NIL	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
3	ADRIANO	Plaza	Supermarket	Clothing Store
3	BICOCCA	Steakhouse	Plaza	Sandwich Place
3	BOVISA	Piadineria	Vegetarian / Vegan Restaurant	Gym
3	BRERA	Ice Cream Shop	Japanese Restaurant	Wine Bar

Results

The results of the project were obtained by unifying the previously shown analysis.

A map of Milano grouping the two classification is shown here. The “best match” between the two classification are the blue dots and the darker green NIL, indicating many venues and variety and greener neighborhoods.



Conclusions

As visible in the map, this report concludes that the better part where to live in Milano are the center of the city, the west suburbs, and the north-east suburbs