DATA SCIENCE FOR RETAIL LOCATION SELECTION IN MILAN

Applied Data Science Capstone

IBM Data Science Professional Certificate

Author: Manuel Cirulli

Date: 31/01/2021

Project Definition

- Problem: In the retail industry, the selection of a new location is an important strategic decision that can largely determine the success of the new store and affect the company's bottom line for years to come.

- Context: With a population of around 1.4 million people, over 3.26 million in the metropolitan area, and over 5.2 million people in its urban area, Milan is the second-most populous city in Italy after Rome and one of the largest urban areas in the EU.

- Project goal: to identify the best areas to open a new supermarket in the city of Milan using open source data science tools and leveraging the Foursquare API.

Data

- The data I used for this project is all available online.
- General data about Milan city population, boroughs, and districts:
- https://en.wikipedia.org/wiki/Municipalities_of_Milan
- Milan Open Data website:
- https://dati.comune.milano.it/
- Link to the JSON file with coordinates for each borough on Milan Open data portal:
- https://dati.comune.milano.it/dataset/a7f54a9a-8331-4825-bc52-6a69a10b0bd3/resource/162ffc23-419f-420d-b14f-bccfe28d920a/download/municipi_sedi__final.json
- **Foursquare API:**
- https://developer.foursquare.com/

Milan City Boroughs: name, population, area, districts

```
[9]: #show dataframe df
     df
```

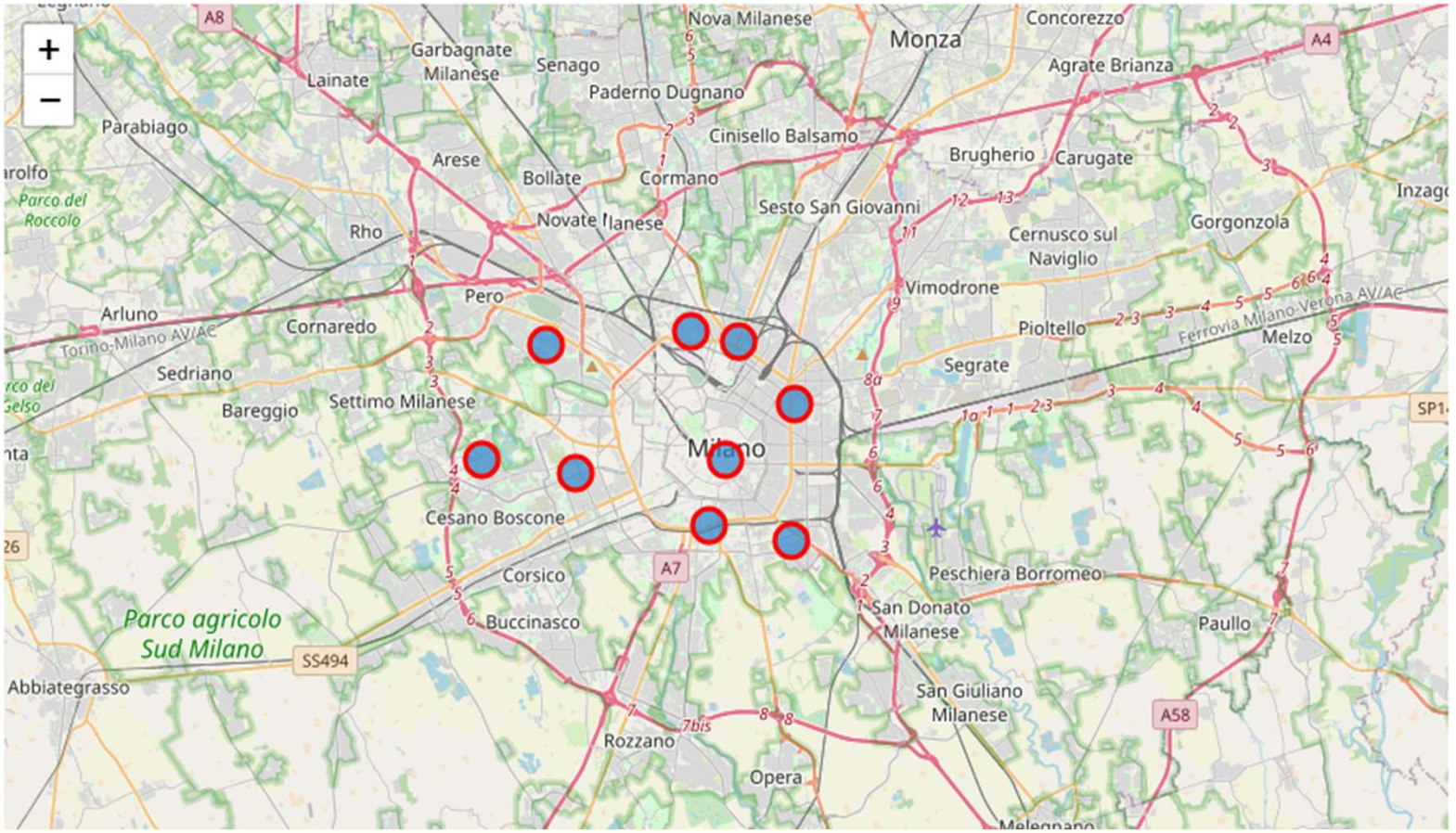
[9]:

	Borough	Name	Area(km2)	Population(2014)	Population density(inhabitants/km2)	Quartieri (districts)
0	1.0	Centro storico	9.67	96315.0	11074	Brera, Centro Storico, Conca del Naviglio, Gua...
1	2.0	Stazione Centrale, Gorla, Turro, Greco, Cresce...	12.58	153109.0	13031	Adriano, Crescenzago, Gorla, Greco, Loreto, Ma...
2	3.0	Città Studi, Lambrate, Porta Venezia	14.23	141229.0	10785	Casoretto, Cimiano, Città Studi, Dosso, Lambra...
3	4.0	Porta Vittoria, Forlanini	20.95	156369.0	8069	Acquabella, Calvairate, Castagnedo, Cavriano, ...
4	5.0	Vigentino, Chiaravalle, Gratosoglio	29.87	123779.0	4487	Basmetto, Cantalupa, Case Nuove, Chiaravalle, ...
5	6.0	Barona, Lorenteggio	18.28	149000.0	8998	Arzaga, Barona, Boffalora, Cascina Bianca, Con...
6	7.0	Baggio, De Angeli, San Siro	31.34	170814.0	6093	Assiano, Baggio, Figino, Fopponino, Forze Arma...
7	8.0	Fiera, Gallaratese, Quarto Oggiaro	23.72	181669.0	8326	Boldinasco, Bullona, Cagnola, Campo dei Fiori,...
8	9.0	Porta Garibaldi, Niguarda	21.12	181598.0	9204	Affori, Bicocca, Bovisa, Bovisasca, Bruzzano, ...

Milan City Boroughs map using *folium*

```
map_milan
```

[26]:



Measuring the attractiveness of each borough to retailers

- ✓ *Main Assumption:* given two areas with the same population, the area with fewer options to buy would be more attractive for a potential food retailer.

- ✓ To assess the attractiveness of each borough for a food retailer searching for a location to open a supermarket, the following data was needed:

- ➢ Potential customers: population of each borough;

- ➢ Competition: number of supermarkets and other food stores in each borough;

Getting a list of Milan venues using the Foursquare API

- Using the *getNearbyVenues* function, The Foursquare API call returned a total of 881 venues in Milan

[30]:
```
#get the venues in Milan
Milan_venues = getNearbyVenues(names=df3['Municipio'],
                               latitudes=df3['LAT_Y_4326'],
                               longitudes=df3['LONG_X_4326']
                              )
```

```
Municipio 1 - Centro storico
Municipio 2
Municipio 3
Municipio 4
Municipio 5
Municipio 6
Municipio 7
Municipio 8
Municipio 9
```

[31]:
```
#visualise the shape of the Milan_venues dataframe

Milan_venues.shape
```

[31]:
```
(881, 7)
```

Getting a list of Milan venues using the Foursquare API

- Irrelevant venue categories were eliminated from the dataframe:

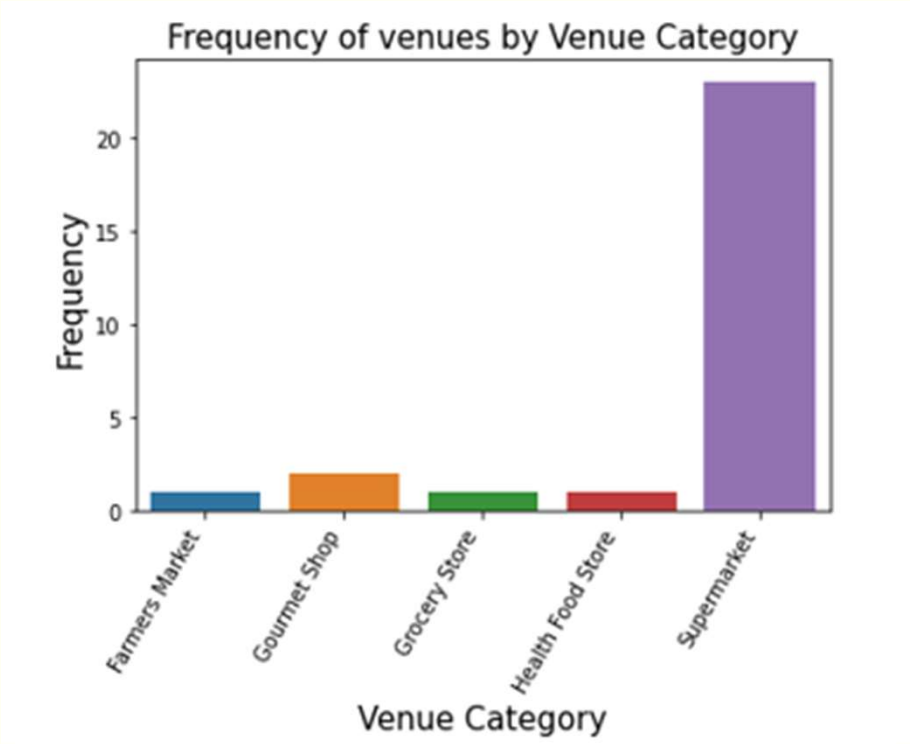
```
[39]: #create a dataframe for each one of these venues' categories
Milan_venues_grocerystores =Milan_venues[Milan_venues['Venue Category'].str.contains('Grocery Store')].reset_index(drop=True)
Milan_venues_healthfood= Milan_venues[Milan_venues['Venue Category'].str.contains('Health Food Store')].reset_index(drop=True
Milan_venues_farmersmarket= Milan_venues[Milan_venues['Venue Category'].str.contains('Farmers Market')].reset_index(drop=True
Milan_venues_markets= Milan_venues[Milan_venues['Venue Category'].str.contains('Farmers Market')].reset_index(drop=True)
Milan_venues_gourmet = Milan_venues[Milan_venues['Venue Category'].str.contains('Gourmet Shop')].reset_index(drop=True)

[40]: #merge to create a dataframe that includes all venues
Milan_venues_all = pd.concat([Milan_venues_Supermarkets,Milan_venues_grocerystores,Milan_venues_healthfood,Milan_venues_farme

[41]: Milan_venues_all.shape

[41]: (37, 8)
```

- After removing duplicates, the list of relevant venues was reduced to 28:

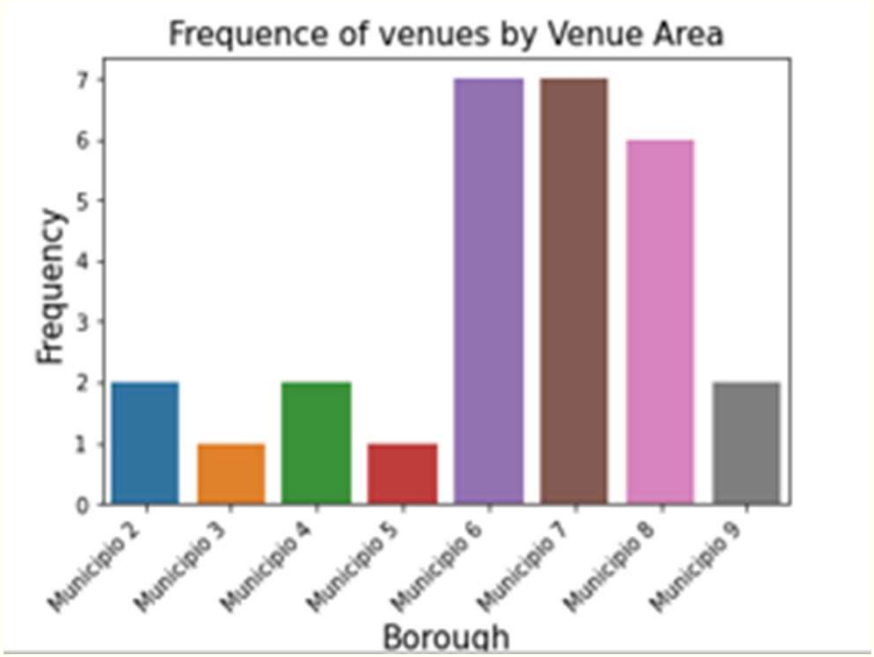
```
[43]: #remove duplicated venues by removing all venues with the same values of Latitude + Longitude

Milan_venues_all = Milan_venues_all.drop_duplicates(subset=['Latitude + Longitude'])

Milan_venues_all.shape

[43]: (28, 9)
```

Frequency of venues by category vs borough

Frequency of Venues by Category

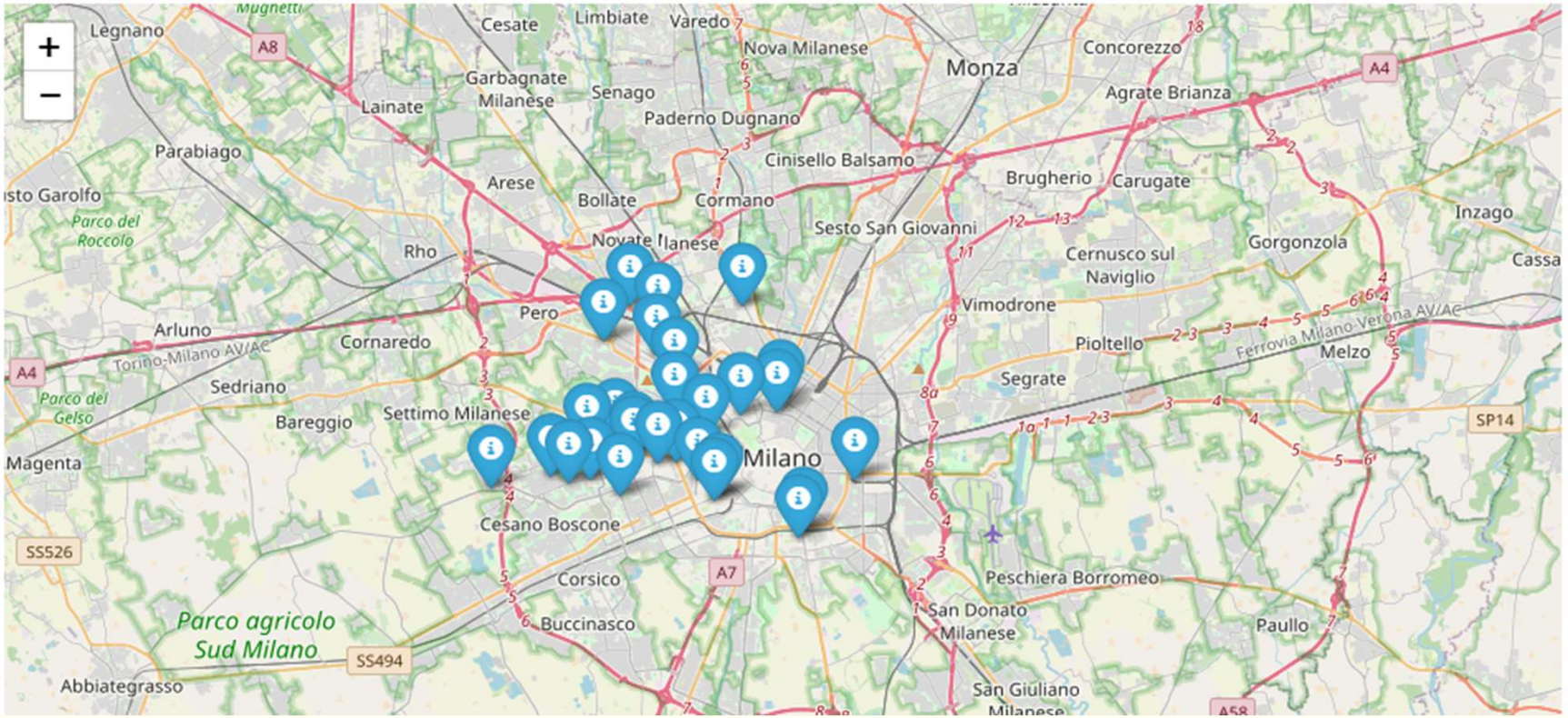


Frequency of Venues by Borough



Folium map of Milan showing relevant venues

```
[56]: map_venues_all
```

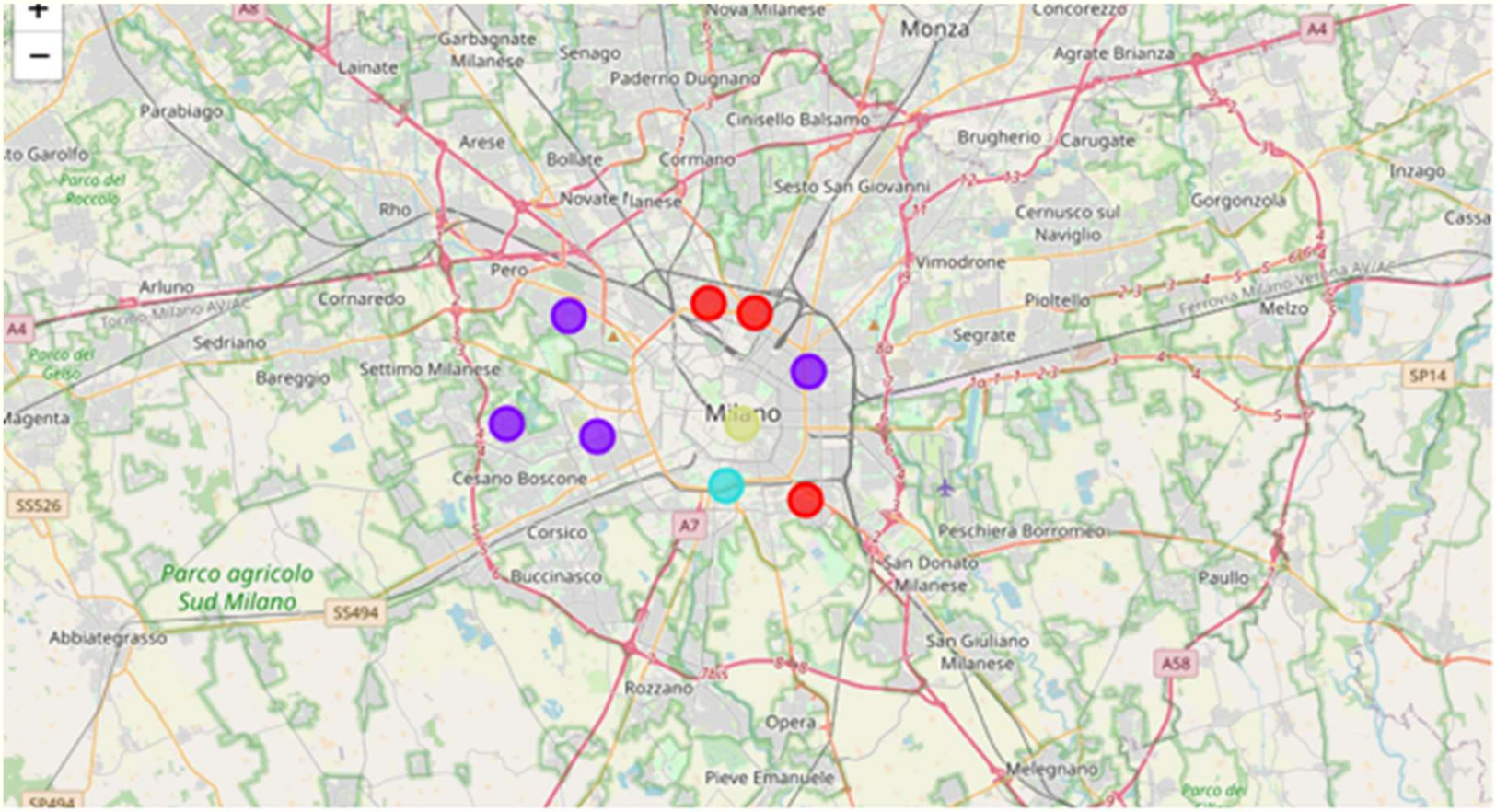


Clustering boroughs using K-means algorithm (k = 3)

Clustering allowed to differentiate boroughs based on the percentace of supermarkets in each borough. Boroughs were divided into 3 clusters. *Municipio 1 was excluded by the clustering algorithm because it contains no venues, so it was added afterwards to form a cluster of its own.*

- CLUSTER 0: 50% of venues in this cluster are supermarkets. The cluster includes 3 boroughs: *Municipio 2, Municipio 4, Municipio 9*.
- CLUSTER 1: In this borough, more than 50% of venues are supermarkets. The cluster includes four boroughs: *Municipio 3, Municipio 6, Municipio 7, Municipio 8*.
- CLUSTER 2: there are no supermarkets among the venues of this cluster. The cluster only includes one borough, *Municipio 5*.
- CLUSTER 3: This cluster only includes one borough (*Municipio 1*). There are no food stores among the venues of this cluster.

Folium map with the clusters



Folium map with the clusters

- CLUSTER 0 (RED): 50% of venues in this cluster are supermarkets. The cluster includes 3 boroughs: *Municipio 2, Municipio 4, Municipio 9*.
- CLUSTER 1 (VIOLET): In this borough, more than 50% of venues are supermarkets. The cluster includes four boroughs: *Municipio 3, Municipio 6, Municipio 7, Municipio 8*.
- CLUSTER 2 (BLUE): there are no supermarkets among the venues of this cluster. The cluster only includes one borough, *Municipio 5*.
- CLUSTER 3 (YELLOW): This cluster only includes one borough (*Municipio 1*). There are no food stores among the Foursquare venues of this cluster.