The Battle of Neighborhoods

Coursera Applied Data Science Capstone - Final Project

Milan, Piazza del Duomo (Cathedral Square)

1. Introduction and Business Problem

Milan is the second biggest city in Italy, and it is considered a leading alpha global city in terms of art, commerce, design, education, entertainment, fashion, finance, healthcare, media, services, research and tourism.

In this context, the Business Problem that we want to address is opening a new wine bar, meant as a café selling a wide selection of wines as well as liquors and beers, paired with gourmet food.

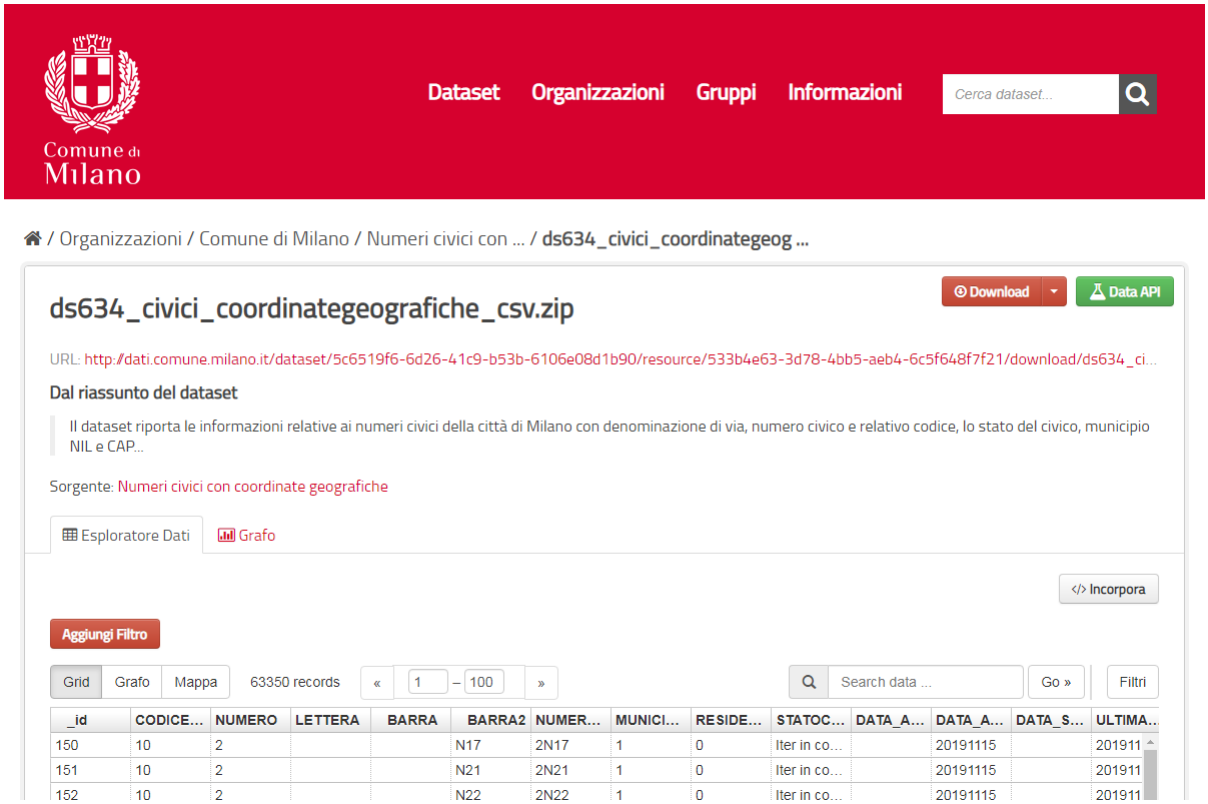
First step of analysis that we want to perform is focused on studying neighborhoods of Milan by means of an unsupervised machine learning (ML) algorithm called K-Means clustering. Main goal of this analysis is outlining an idea about which location might be most suitable for a new wine bar, taking into account average level of market competitors (we expect an area with high to top level restaurants) and presence of specific competitors in terms of wine bars.

Selecting a location where we can expect a lifestyle of wine and food appreciation we will be able to quickly reach a breakeven point of the investment. In addition to that, we expect starting a wine bar less expensive than starting a full-service restaurant and, therefore, profits could be higher.

2. Data

Data about Milan neighborhood is scraped out of a quite large database of Milan addresses, available from Municipality website:

http://dati.comune.milano.it/it/dataset/ds634-numeri-civici-coordinate/resource/533b4e63-3d78-4bb5-aeb4-6c5f648f7f21



Screenshot of Milan neighborhoods data source

An API to query this database is available, hence it is used to import data into a Jupyter Notebook.

Starting from a 633550 rows database, data is filtered and cleaned, in order to obtain a new database with neighborhoods names and average coordinates of their addresses' ones.

This new databased serves as a starting point to get data about all available venues from Foursquare API, by means of an API get request, and within a radius of 500 m.

Data is then formatted by means of one-hot encoding with the categories of each venue. Then, venues are grouped by neighborhoods, evaluating the mean of each venue occurrence.

Similarities among neighborhoods are then evaluated by means of K-Means clustering technique, selecting a number of six clusters.

3. Methodology

In order to get data from Milan Municipality website, a request command is used. Data is then converted into json strings and normalized pandas dataframe.

```
In [2]: response=request(url='http://dati.comune.milano.it/it/api/3/action/datastore_search?resource_id=533b4e63-3d78-4bb5-aeb4-6c5f648f7f21
addresses=response.json()
jtopy=json.dumps(addresses) #json.dumps take a dictionary as input and returns a string as output
data=json.loads(jtopy)
data=json_normalize(data['result']['records'])
data.head()
```

Out[2]:

	ANNCSU	ANNO_SOPPRESSIONE	BARRA	BARRA2	CAP	CODICE_VIA	DATA_APPLICAZIONE	DATA_ATTIVAZIONE	DATA_INTITOLAZIONE	DATA_
0	VITTORIO EMANUELE SECONDO	0.0		N17	20121.0	10	NaN	20191115.0	0.0	
1	VITTORIO EMANUELE SECONDO	0.0		N21	20121.0	10	NaN	20191115.0	0.0	
2	VITTORIO EMANUELE SECONDO	0.0		N22	20121.0	10	NaN	20191115.0	0.0	
3	VITTORIO EMANUELE SECONDO	0.0		N04	20121.0	10	NaN	20191114.0	0.0	
4	VITTORIO EMANUELE SECONDO	0.0		N01	20121.0	10	NaN	20200123.0	0.0	

5 rows × 39 columns

```
In [3]: data.shape
Out[3]: (63350, 39)
```

Screenshot of data scraping from Milan Municipality website on Jupiter Notebook

In order to get the correct level of granularity of data we need to get rid of addresses, and to extract from this database the names of distinct neighborhoods and their geographic coordinates.

In order to do that, data is grouped by "NIL" (Nuclei di Identità Locale), literally "Local identity units, that is neighborhoods, and the average of latitude and longitude values relevant to involved addresses is calculated for each of them.

```
In [7]: neighborhood=data_filtered.rename(columns = {'NIL':'Neighborhood','LAT_WGS84':'Latitude','LONG_WGS84':'Longitude'})
neighborhood
```

Out[7]:

	Neighborhood	Latitude	Longitude
1	ADRIANO	45.511738	9.245336
2	AFFORI	45.514814	9.172760
3	ASSIANO	45.455063	9.062976
4	BAGGIO - Q.RE DEGLI OLMI - Q.RE VALSESIA	45.461175	9.088395
5	BANDE NERE	45.460703	9.138751
6	BARONA	45.432688	9.154456
7	BICOCCA	45.517449	9.210782
8	BOVISA	45.501859	9.165413
9	BOVISASCA	45.516964	9.154238
10	BRERA	45.473413	9.187425

Screenshot of neighborhoods dataframe ready to be used

An API get request is then used to get data relevant to venues available from Foursquare database.

```
In [11]: print(milan_venues.shape)
         milan_venues.head()
```

```
(2406, 7)
```

Out[11]:

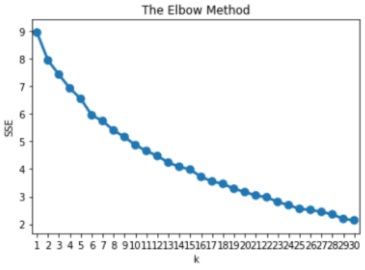
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	ADRIANO	45.511738	9.245336	Cargo	45.508406	9.243159	Furniture / Home Store
1	ADRIANO	45.511738	9.245336	Osteria Ai 3/4	45.508452	9.243858	Italian Restaurant
2	ADRIANO	45.511738	9.245336	Zan Zara Zan	45.508652	9.242995	Café
3	ADRIANO	45.511738	9.245336	Unieuro	45.514146	9.244038	Electronics Store
4	ADRIANO	45.511738	9.245336	Esselunga	45.513782	9.244425	Supermarket

Screenshot of venues dataframe

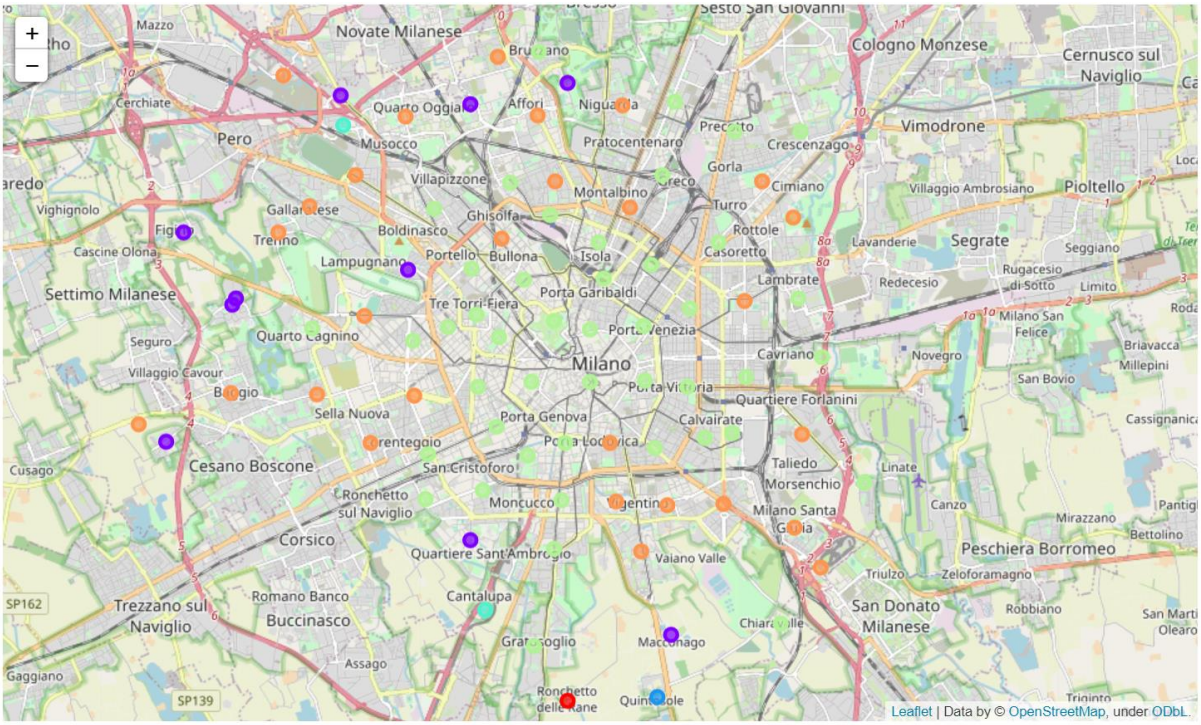
After that, venues are grouped by neighborhood, one-hot encoded and normalized by means of an average calculation of one-hot encoded matrix.

A dataframe with top ten venues for each neighborhood is then created.

In order to perform K-Means clustering analysis, we evaluate optimum k using the elbow method: as per the picture on the right it appears that there are no evident elbows. Anyway, there is a certain change in the curve slope at k=6, hence this value is selected to perform the analysis.



K-Means clustering is then performed, and resulting data are visualized on a map, by means of Folio library.



Map of identified neighborhood clusters

Finally, clusters are examined in order to understand possible patterns.

4. Results

From clusters examination following descriptions arise:

Cluster #	Description
Cluster 1	Pizza and Café
Cluster 2	B&Bs
Cluster 3	Playgrounds
Cluster 4	Shops
Cluster 5	Hotels
Cluster 6	Restaurants

It appears that cluster 6 is the most suitable one for addressed business problem. Indeed, a restaurant is the most frequent venue in 67 % of the neighborhoods and one of first two most frequent venues in 87 % of the neighborhoods.

Anyway, wine bars are one of the five most frequent venues in only less than 5 % of neighborhoods, and this makes us confident that a new wine bar should not be experience a fierce competition among existing venues.

5. Discussion

It is to be noted that this is only an initial analysis to determine in which area of Milan opening a new wine bar could lead to better results.

Further market analysis should be performed in order to take into account demographic parameters and financial statements of existing wine bars similar to the one that we are planning to open.

In addition to that, neighborhoods relevant to Cluster 6 are mainly located in the city center, where the cost of rents are, on average, quite high. Hence, this point should also be taken into proper account.

6. Conclusion

Using python language and K-Means machine learning algorithm allowed to perform an interesting analysis about Milan neighborhood, highlighting that those located in the city center (inside beltway) are almost all referring to the same cluster, hence very similar each other.

Hence locating new wine bar in this area is to be suggested, even if further analysis highlighted in above paragraph are still to be carried out.