The character of Italian cities

Introduction / Business Problem

Italian cities are small (Wikipedia list): the most populated city is Rome, the capital city, with population close to 3 millions. Then comes Milan, the economic capital, with a population of less than 1.5 millions. The population of all the other Italian cities is below 1 million, with only 15 cities above 200 thousands and more or less 130 other cities (more than 50 thousands inhabitants).

We can think that small cities have a prevailing character or vocation. So, our business problem is:

Business problem: can we use data to grasp the character of Italian cities?

This is a very general study, so it may be of interest for several categories of stakeholders and at the same time it is not specifically aimed at any of these. We identify two groups of stakeholders:

Stakeholders:

- Tourists: suppose you have a list of your preferred Italian cities and you want to visit another Italian city. Then you can choose among the cites like your preferred ones.
- Investors: suppose you did investments in Italian cities. Some of them went very well, some others went less well. If you want to want to invest other money you may chose the cities having a similar character as those were you did the best money.

Data

We will use two main sources of data:

The list of Italian cities

We will take the list from the Italian Wikipedia page. We don't use the English version (linked above) because it doesn't provide the *province* column that we found to be necessary to get the latitude and longitude with *geocoder*. The table in the Wikipedia page lists the 146 Italian cities (data from December 31, 2018) starting from the most populated. For each city the table reports:

- Rank. The rank of the city according to its population.

- City. The (Italian) name of the city.

- Region. Italy is subdivided in 21 regions. The column reports which region the city belongs to.

- Province. Italian regions are further subdivided into provinces. The column reports the province of the city.

- Population. The population of the city.

The Foursquare database of world venues

More specifically, we use the Foursquare API *explore* request that returns ‘a list of recommended venues near the current location’, i.e. near each city center. The foursquare response provides several information about each venue. We will leverage the venue *category*. Foursquare defines more than 900 venue categories, but each category is a subcategory of one of 10 main (or root) categories:

1. Arts & Entertainment.
2. College & University.
3. Event.

4. Food.
5. Nightlife Spot.
6. Outdoors & Recreation.
7. Professional & Other Places.
8. Residence.
9. Shop & Service.
10. Travel & Transport.

For each Italian city, we will determine its distribution of root categories and we will use this datum to cluster the Italian cities.

Methodology

The list of Italian cities was obtained from the table in the Wikipedia page, using the *read_html* function of the *pandas* package. The columns of the obtained *dataframe* were renamed using English names. The spaces in *Population* column were removed and the column was converted to numeric.

The coordinates of the cities were obtained using the *Nominatim* geocoder for OpenStreetMap data as provided by the *geopy* package. The query was composed as <city_name>, <province>, IT. After inspecting the map, a few coordinates values were correct manually.

The Foursquare API request *https://api.foursquare.com/v2/venues/categories* was used to retrieve the categories tree. The resulting JSON response was parsed using a custom function that allowed to obtain a dictionary whose keys were categories IDs and whose values were the root categories' root categories.

The recommended venues nearby the city center were obtained with the Foursquare API request https://api.foursquare.com/v2/venues/explore with parameters

- v=20191201
- ll=[the latitude and longitude of the city]
- radius=3000
- limit=100

The cities with less than 15 recommended venues were removed from the analysis due to lack of statistics.

The resulting venues dataframe was used to compute, for each city, the distribution of the venues among the ten root categories.

Summary statistics and boxplots were used to have an idea on the data on the root venues.

k-means clustering (KMeans algorithm from the scikit-learn sklearn.cluster module) was applied to the root categories distributions, setting the parameters n_clusters=5 and random_state=0.

The *folium* package was used for the visualization of the cities and city clusters on the map.

A plot of the average value of each category in the different clusters allowed to visualize the main differences among the clusters and do interpret them.

Results

Location of the Italian cities

The obtained coordinates of the table of the 146 Italian cities listed in the Wikipedia are shown on the map of Figure 1. We could observe some cluster of cities around the big cities of Milan, Rome, Naples and Bari, but we were interest on clustering by characteristics instead of by position.

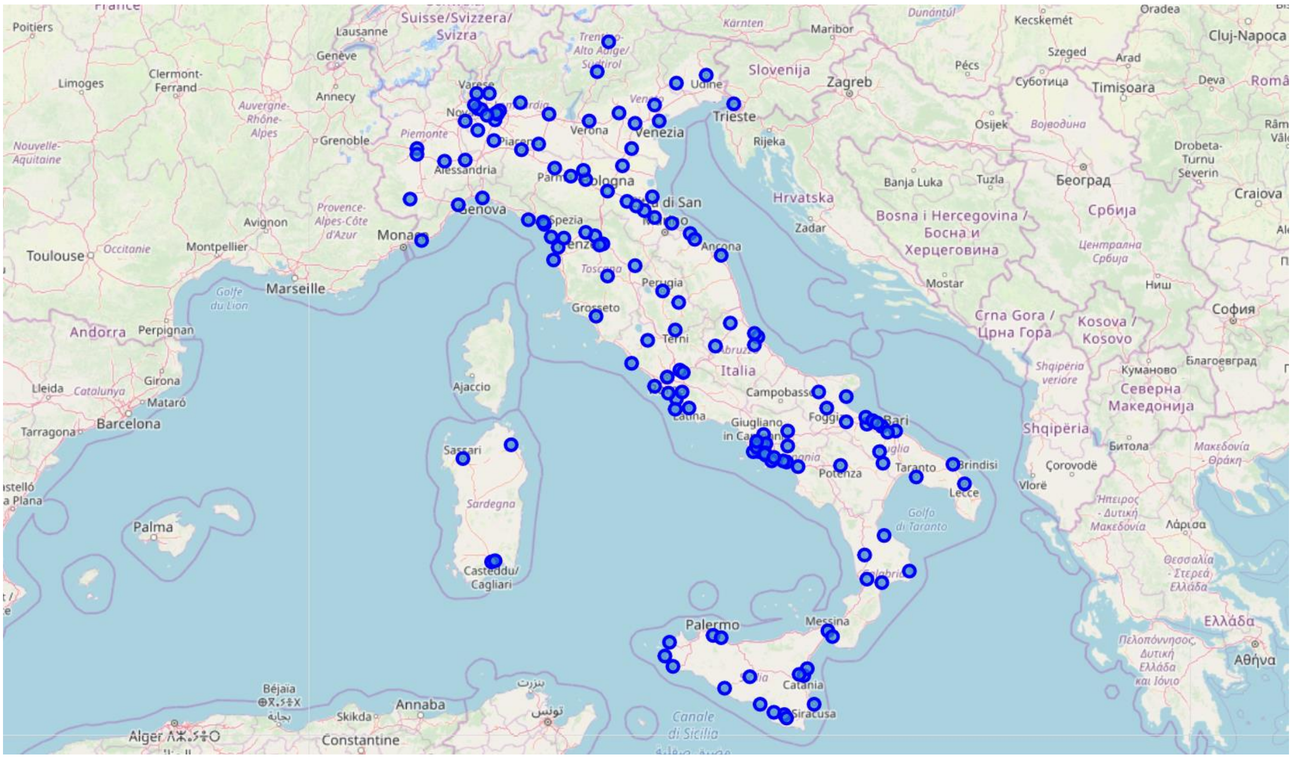

Figure 1. The map Italian cities.

Venues

Table 1 and Figure 2 represent, for a summary statistic for each of the root categories. The *Event* category did not appear in the results, as it is reserved to a specific query. Also no venue of *Residence* category was returned. The most common venue category was almost always the *Food* category, with an usual frequency of 50 % to 60 %. Then there were three categories with median frequency around 10 %: *Nightlife Spot*, *Outdoors & Recreation* and *Shop & Service*. Then came *Arts & Entertainment* and *Travel & Transport* with a median frequency of 6 %. Finally, less represented among the recommended places, came the *College & University* and *Professional & Other Places* categories.

	mean	std	min	**25%**	**50%**	**75%**	max
Arts & Entertainment	7%	4%	0%	**4%**	**6%**	**9%**	21%
College & University	0.0%	0.2%	0.0%	**0.0%**	**0.0%**	**0.0%**	2.2%
Food	53%	9%	22%	**49%**	**54%**	**58%**	80%
Nightlife Spot	10%	6%	0%	**6%**	**9%**	**13%**	36%
Outdoors & Recreation	11%	5%	0%	**8%**	**11%**	**14%**	37%
Professional & Other Places	1%	2%	0%	**0%**	**0%**	**2%**	12%
Shop & Service	11%	9%	0%	**6%**	**10%**	**15%**	56%
Travel & Transport	7%	5%	0%	**3%**	**6%**	**8%**	29%

Table 1. Statistics of the distribution of the venues among the

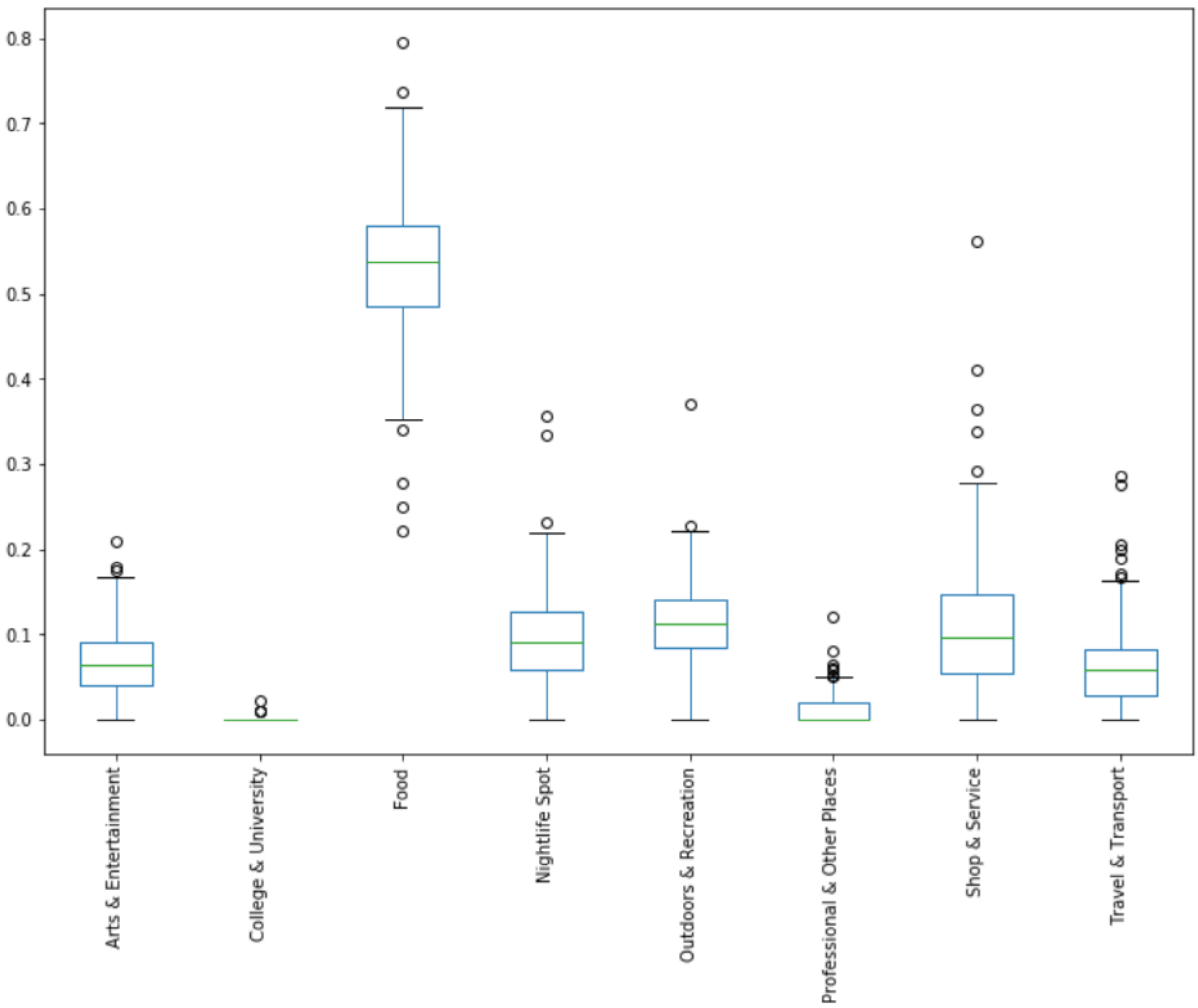

Figure 2. For each root category, the boxplot of the fraction of venues of that category in the Italian cities.

Clustering

The k-means clustering is visualized on the map in Figure 3. This map shows that the clusters aren't related to the geographic position of cities. On the other hand the nineteen removed cities (cities with less than 15 recommended venues) are all in the south of Italy, which is lagging behind with respect to the richer north and has seen a lot of emigration of the youngest, also in recent years.

In order to interpret the clustering, we plotted the mean distribution of the venues root categories in each cluster Figure 4. As expected, the principal venue category for all the clusters is *Food*. It is instead interceding to look at who comes second:

- Cluster 0 (36 cities) who comes second is Outdoors & Recreation, meaning that there are several opportunities for exploring natural beauties around these cities.
- Cluster 1 (32 cities) Shops & Services is the second most important category here. Should be a good city to eat and buy typical Italian stuff.
- Cluster 2 (26 cities) There are many Nightlife spots: are perfect for having fun in the night!
- Cluster 3 (25 cities). I would say food and food again. You can eat quite a lot there.
- Cluster 4 (8 cities). These cities are filled with shops, so much that the number of Food places is below the average.

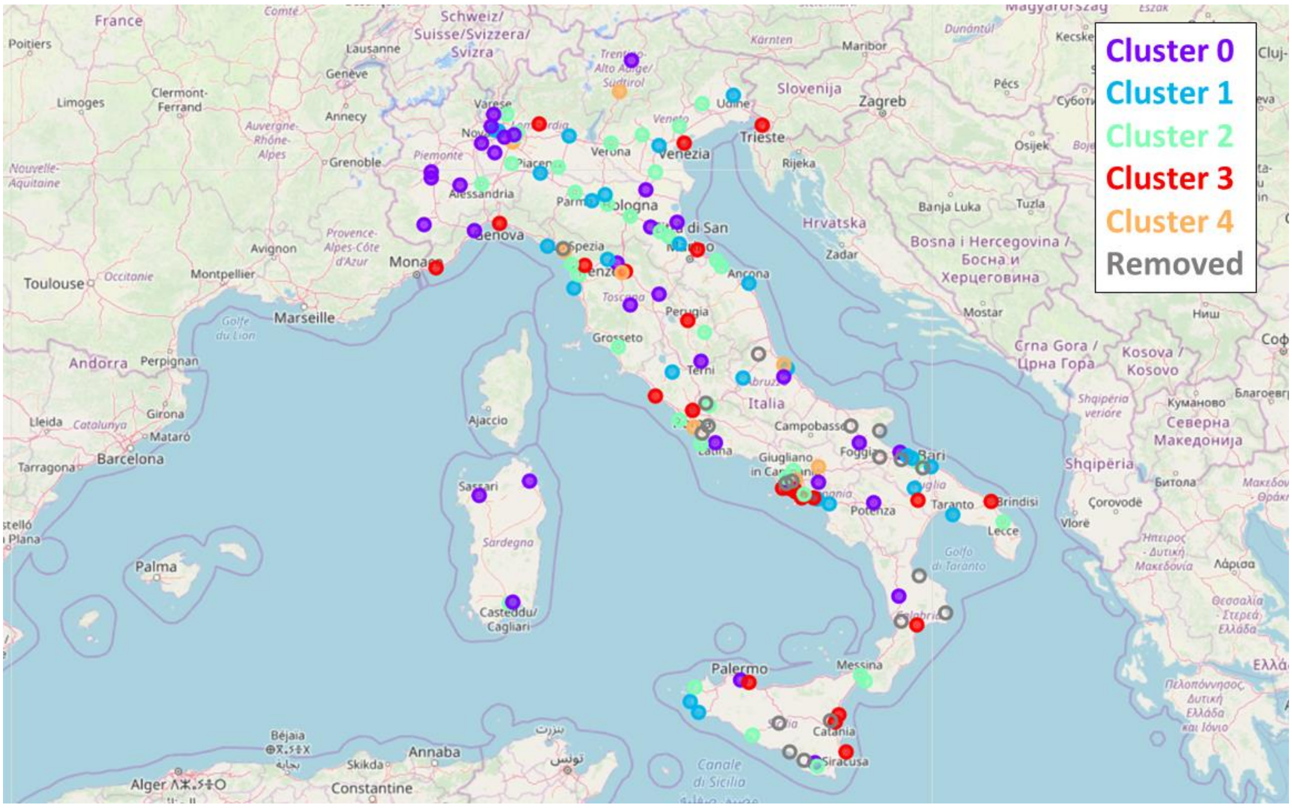

Figure 3. k-means clustering of Italian cities (5 cluster) and cities removed from clustering analysis due to the lack of data (grey circles).

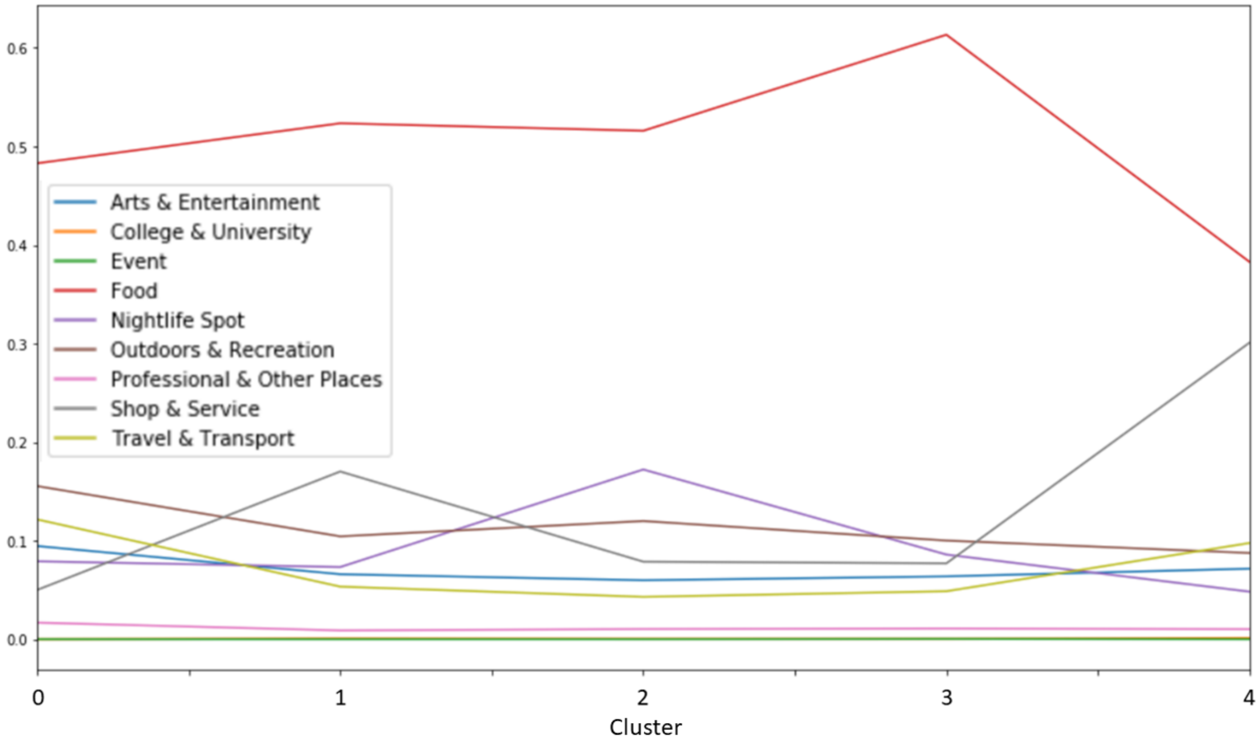

Figure 4. Mean frequency of the root categories in each cluster.

Discussion

The first observation that we can make is the choice to use the root categories, made the result interpretable, while retaining enough discrimination power. As for the choice of the clustering algorithm, the k-means gave good results, but another possible choice could have been the DBSCAN algorithm which could be run on all the cities as it is capable to recognize outliers.

Conclusions

The clustering of Italian cities was done with two groups of stakeholders in mind: tourists and investors. The information obtained during the analysis gave some good indications. For example if you are a tourist that likes night life, or an investor that wants to open a new pub, you should look at the cities of cluster 2. It is also clear that these are very generic indication and a more specific business problem and data analysis should be developed in order to get more insight. For example, in the case of investors, economic data could be included in the analysis.