Clustering the most prominent Italian cities by cultural heritage sites.

Lorenzo Pedrazzetti

August 10, 2020

1 Introduction

1.1 Background

Italy is first country for cultural heritage sites count in the world[1]. This feature enable a very dynamic tourism market[2], that is not limited to natural parks or leisure centres. It is common for many tourists to visit those Italian cities where the density of museums and antique buildings or churches are the most important feature. Given the very high number of attractions of such kind, it is not always trivial for travel companies to guide the customers in deciding how to spend their limited vacation time in Italy and which cities they'd prefer to visit, based on their interest. Furthermore, having a well-structured classification of what the most prominent Italian cities has to offer could also prove beneficial in designing vacation products that genuinely suit the tastes of the customer base.

1.2 Problem

It should be possible to use geospatial data to describe the vicinity of the Italian cities down towns and to classify which venues account for an intrinsic cultural value. Moreover, it would be useful to provide details on which cities share similar "cultural morphology".

1.3 Interests

Travel companies should be the clear target for this extensive study, as they have the means to extract value from the data by designing very specific products: they could propose to the customer a detailed tour of only those Italian cities whose main attractions strike their fancy. Moreover, this dataset could also be used to create customized marketing advertisement from the Italian cities themselves, who would be able to leverage their unique heritage with respect to other cities.

1 https://en.wikipedia.org/wiki/World_Heritage_Sites_by_country
2 https://www.ceicdata.com/en/indicator/italy/visitor-arrivals

2 Data

2.1 Data Sources

A repository of the geospatial coordinates of the most prominent Italian cities can be found at the cited source[3]. This table provides the information needed to leverage the Foursquare API and fetch a list of all the venues in the proximity of each city geographical centre.

2.2 Data Wrangling

The data downloaded from the aforementioned table were available as a Comma Separated Values file, that was easily imported in a Jupyter Notebook. Information of lower interest, such as the population count, were trimmed from the table, as well as any NaN value found in the table and not needed. In the interest of structuring even more the dataset, so that more information could be extracted, the cities listed in the table were split into 3 macro-areas, depending on their latitude: the geographical north, middle and south of Italy. From now on, each step described is assumed to be applied to each one of the 3 datasets.

2.2.1 Scraping the Venues Category List

Obtaining a list of what is considered a cultural attraction was another task to undertake. To do so, the Foursquare webpage listing all the categories was scraped[4]. The data were organized in a typical data structure (Pandas data frame), cleaned of the html refuses, and analysed to extrapolate information about which venues fit the description. To do so, the headers of each category section were recognized and used to find the categories subset needed.

3 https://simplemaps.com/data/it-cities
4 https://developer.foursquare.com/docs/build-with-foursquare/categories/

3 Methodology

First, a Folium map showing the most prominent Italian cities was primed, to reference the three geographical macro-areas.

Figure 1: Most prominent Italian cities by geographical area: north (green), middle (blue), and south (red).

3.1 The Foursquare API

The Foursquare API was called for each city of the tables. The aim of the project was to develop a description of the inner down towns of each city, hence the following parameters were passed to Foursquare:

- venues had to be found only within 5 km of the city's geographical centre
- the returned dataset was limited to the first 200 entries for each city centre

For each call, the API returned a 'JSON' file containing a list of venues that comprehended both cultural and artistic attractions, as well as more frivolous ones. The JSON was then disassembled and the data organized again in a more handy structure, such as a Pandas data frame.

The tables, at this point, featured a list of ~200 venues, with their respective geospatial data and categorization, for each city included in the original list.

In the interest of this project, the resulting datasets was filtered to set apart the cultural attractions and the more frivolous ones. To do so, the trimmed venues categories table described in section 2.2.1 was employed.

3.2 Feature selection

The key feature around which the project developed was the frequency of occurrence of the first 10 most popular cultural venues for each city.

This classification was easily done within the Python script and a sample table is here reported for reference (featuring the first 3 most common venues).

To do this, the one-hot encoding technique was employed to transform categorical values like the venues category into numerical ones, on which inferential statistic is easier. This approach made also possible to apply a machine learning algorithm (as discussed in section 3.3) to the dataset.

A sample of the inferential statistic performed on one of the 3 datasets is here reported.

Table 1: Table featuring the first 3 most common venues for the cities of the northern Italian macro-area.

City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Alessandria	Plaza	History Museum	Historic Site
Ancona	Monument / Landmark	Plaza	Park
Aosta	Historic Site	Church	Mountain
Arezzo	Stables	Park	Flea Market
Asti	Plaza	Art Museum	History Museum

As it can be seen, the table only list those venues that, in my personal opinion, have a cultural value for a possible customer.

3.3 Machine Learning

A key aspect of the project was to understand how the different cities in the different geographical areas could be associated in terms of their cultural heritage. To do so a machine learning algorithm was implemented, specifically, the *k-means* clustering.

3.3.1 K-means clustering

K-means clustering is a ML (machine learning) method that aims at partitioning n observations into k clusters, in which each cluster has common features[5]. It originated in the field of signal processing but the open Python library scikit-learn[6] features method that can be freely implemented in many scenarios.

In this project, the algorithm was employed to find clusters of Italian cities (belonging to the same geographical macro-area) in terms of frequency of occurrence of the first 10 cultural venues.

As mentioned before, the datasets were first turned into dummies numerical values from categorical ones, to allow for an easier implementation of the method.

Moreover, the parameters passed to the method were the following ones:

- an initial clusters number of 5 labels
- a seed for the pseudo-random distribution to start with equal to 0

In the following section, graphical exemplification of the output will be given and discussed; here a sample table will be reported for reference, showing the allocated cluster labels along with the corresponding top venues and the geospatial data (only top 2 venues showed to fit the page size).

Table 2: Table reporting geospatial data, ML labels and top venues for the northern Italian macro-area.

City	CityLat	CityLng	ClusterLabels	1st Most Common Venue	2nd Most Common Venue
Alessandria	44.9	8.616667	**0**	Plaza	History Museum
Ancona	43.633333	13.5	**2**	Monument / Landmark	Plaza
Aosta	45.733333	7.333333	**0**	Historic Site	Church
Arezzo	43.416667	11.883333	**4**	Stables	Park
Asti	44.9	8.2	**0**	Plaza	Art Museum
Belluno	46.145	12.221389	**2**	Movie Theatre	Plaza
Bergamo	45.683333	9.716667	**2**	Plaza	Park
Biella	45.566667	8.05	**1**	Park	Art Gallery
Bologna	44.483333	11.333333	**2**	Plaza	Church

5 https://en.wikipedia.org/wiki/K-means_clustering
6 https://scikit-learn.org/stable/modules/generated/sklearn.cluster.k_means.html?highlight=k%20means#sklearn.cluster.k_means

3.4 Exploring the clusters

To obtain a clear view on the clustering feat meaning, it was necessary to explore each cluster for each macro-area.
To avoid being pedantic with the data report, only a small number of the results were shown in this section, as a reference; a more complete discussion is reported in 5.

Table 3: Northern macro-area, cluster label 0 description

	CulturalVenue
1st Most Common Venue	Plaza
2nd Most Common Venue	Art Gallery
3rd Most Common Venue	Theater
4th Most Common Venue	Bookstore
5th Most Common Venue	Theater
6th Most Common Venue	Park
7th Most Common Venue	Bookstore
8th Most Common Venue	Convention Center
9th Most Common Venue	Distillery
10th Most Common Venue	Event Space

4 Results and Discussion

In this section, the data produced by the implementation of the methodology above is presented.

Let's see how the different macro areas were clustered by the ML algorithm:

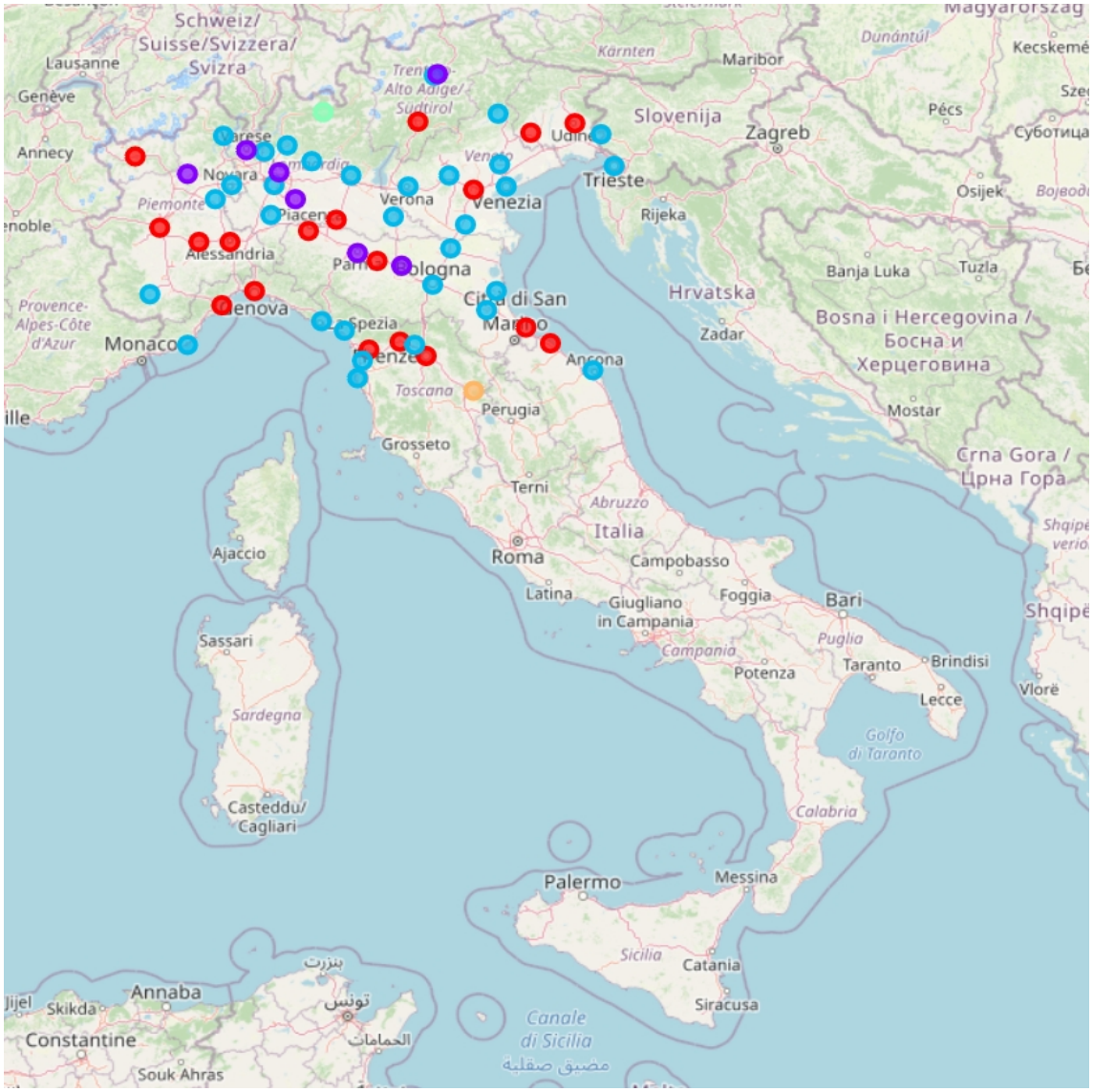


Figure 2: Cluster display of the northern Italian cities: cluster0 (red), cluster1 (purple), cluster2 (blue), cluster3 (green), and cluster4 (yellow).

To better understand the meaning of the clusters let's see which are the 10 most frequent cultural venues for each of them:

Table 4: Summary of the top 10 cultural venues in each northern cluster.

	CulturalVenues, Cluster0	Cluster1	Cluster2	Cluster3	Cluster4
1st Most Common Venue	*Plaza*	*Park*	*Plaza*	*Plaza*	*Stables*
2nd Most Common Venue	*Art Gallery*	*Park*	*Park*	*College Library*	*Park*
3rd Most Common Venue	*Theater*	*Bookstore*	*Historic Site*	*Concert Hall*	*Flea Market*
4th Most Common Venue	*Bookstore*	*Plaza*	*Theater*	*Convention Center*	*Concert Hall*
5th Most Common Venue	*Theater*	*Wedding Hall*	*Bistro*	*Distillery*	*Convention Center*
6th Most Common Venue	*Park*	*Historic Site*	*Bookstore*	*Event Space*	*Distillery*
7th Most Common Venue	*Bookstore*	*Distillery*	*Castle*	*Factory*	*Event Space*
8th Most Common Venue	*Convention Center*	*Wedding Hall*	*Concert Hall*	*Farm*	*Factory*
9th Most Common Venue	*Distillery*	*Factory*	*Convention Center*	*Field*	*Farm*
10th Most Common Venue	*Event Space*	*Farm*	*Distillery*	*Flea Market*	*Field*

Let’s see now how the rest of Italy is clustered.

The middle:

Figure 3: Cluster display of the middle Italian cities: cluster0 (red), cluster1 (purple), cluster2 (blue), cluster3 (green), and cluster4 (yellow).

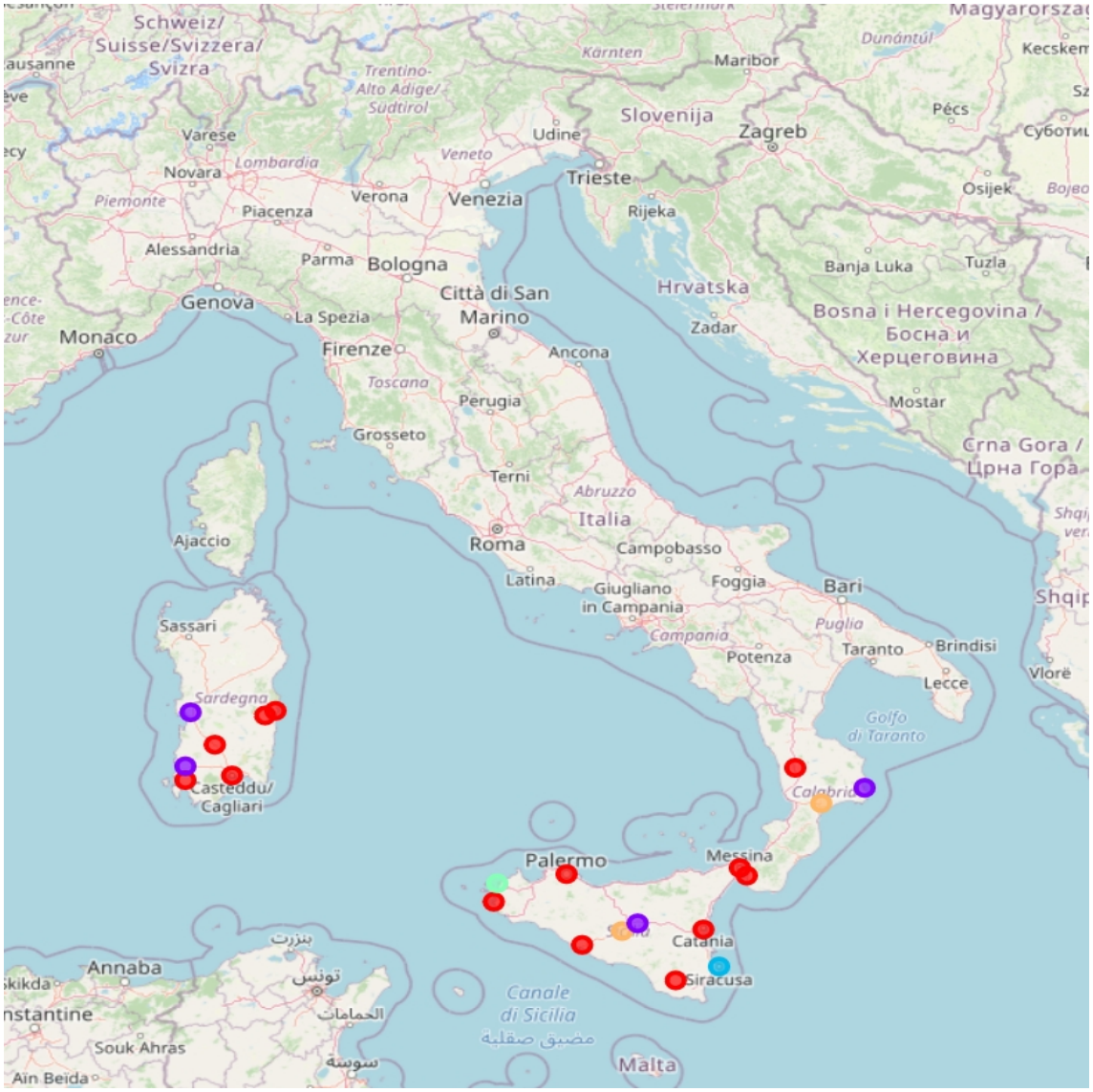
Let's now see the clusters structure:

Table 5: Summary of the top 10 cultural venues in each middle cluster.

	CulturalVenues, Cluster0	**Cluster1**	**Cluster2**	**Cluster3**	**Cluster4**
1st Most Common Venue	*Plaza*	*Plaza*	*Park*	*Movie Theater*	*Castle*
2nd Most Common Venue	*Theater*	*Plaza*	*History Museum*	*Plaza*	*Theater*
3rd Most	*Castle*	*Plaza*	*Harbor /*	*Waterfront*	*Waterfront*

Common Venue			*Marina*		
4th Most Common Venue	*Harbor / Marina*	*Bookstore*	*Garden*	*Harbor / Marina*	*Harbor / Marina*
5th Most Common Venue	*Garden*	*Waterfront*	*Garden*	*Garden*	*Garden*
6th Most Common Venue	*Fountain*	*Castle*	*Cultural Center*	*Fountain*	*Fountain*
7th Most Common Venue	*Cultural Center*	*Garden*	*Cultural Center*	*Cultural Center*	*Cultural Center*
8th Most Common Venue	*Convention Center*	*Fountain*	*Convention Center*	*Convention Center*	*Convention Center*
9th Most Common Venue	*City Hall*	*Cultural Center*	*Church*	*City Hall*	*City Hall*
10th Most Common Venue	*Church*	*Convention Center*	*Cemetery*	*Church*	*Church*

Finally, the South:



And the cluster structure:

Table 6: Summary of the top 10 cultural venues in each southern cluster.

	CulturalVenue, Cluster0	**Cluster1**	**Cluster2**	**Cluster3**	**Cluster4**
1st Most Common Venue	*Plaza*	*Plaza*	*Scenic Lookout*	*River*	*Park*
2nd Most Common Venue	*Plaza*	*Theme Park*	*Theme Park*	*Theme Park*	*Park*
3rd Most Common Venue	*Theme Park*	*Garden*	*Garden*	*Garden*	*Museum*
4th Most Common Venue	*Theater*	*Amphitheater*	*Amphitheater*	*Amphitheater*	*Amphitheater*

5th Most Common Venue	*Farm*	*Art Gallery*	*Art Gallery*	*Art Gallery*	*Art Gallery*
6th Most Common Venue	*Garden*	*Art Museum*	*Art Museum*	*Art Museum*	*Event Service*
7th Most Common Venue	*Bistro*	*Bistro*	*Bistro*	*Bistro*	*Church*
8th Most Common Venue	*Art Gallery*	*Bookstore*	*Bookstore*	*Bookstore*	*Castle*
9th Most Common Venue	*Fountain*	*Boutique*	*Boutique*	*Boutique*	*Boutique*
10th Most Common Venue	*Cafeteria*	*Cafeteria*	*Cafeteria*	*Cafeteria*	*Cafeteria*

5 Discussion

So, it looks like plazas are them most interesting attraction in the north. If a customer is more interested in parks and books, a city like Lodi (cluster 1) seems more appropriate. Finally, if the customer is interested in horses and open markets, he should definitely check out Arezzo, the only city in the fourth cluster.

Again, by checking the table it's easy to see the plazas that take a dominant role also in the middle of Italy. But here, castle lovers are that visit Barletta are well poised for a dreamy vacation.

By reviewing the clusters in the south of Italy it is clear how theatres and amphitheatres are among the most frequent venues you could see by travelling around Palermo (cluster 1), Tortolì (cluster 2), Villacidro (cluster 3), and Catanzaro (cluster 4).

It appears clear how segmenting the cities into 3 different macro-areas helped the ML algorithm in terms of clusters diversity: each section of the Country has cluster that better suit the cultural heritage found there. In this way, any recommendation based on these datasets should be accurate.

6 Conclusion

In this project, the most prominent Italian cities were segmented and clustered to identify those that share common ground in terms of cultural heritage. To start, I divided the data set by different values of latitude, to better understand the cities outline. Then, the geospatial data obtained were used to iteratively call the Foursquare API to locate the most interesting venues in each city. With the aim of filtering out the most frivolous attractions, a list of venues categories was scraped by the Foursquare documentation page and used to structure the dataset obtained via the API.

Once the entries were sufficiently organized, a k-means ML algorithm was implemented to cluster the cities' venue and create an in depth map of which cities shared some common venues structure. This was done by analysing the frequency of occurrence of the venues categories and by sorting them out in tables that clearly explain what to expect from each city belonging to each cluster for each macro-area.

These data can be used by travel companies to design products tailored on customers tastes for culture and make more efficient selection of which of the many interesting Italian cities are worth visiting.