IBM Data Science Course Capstone

A tourist guide based on unsupervised learning clustering techniques

Carlo D'Aloia

1. Business Problem

In this section I am going to discuss my idea for the Capstone project

1.1 Background

First of all I would like to introduce the location of Interest I chose for my project. The business idea will be developed in the city of Salerno (Italy) and its neighborhoods. Salerno is my hometown, I chose it because I know it well and I can help with results interpretation. The city is quite small so I am going to take into account some nearby places. This area is suitable for my goal because tourism is a great source of income and the amount of places will provide an interesting scenario to study.

1.2 Mission

The idea is to create a tourist guide using geodata and machine learning. I am going to cluster the area in order to find the places more suitable for every need, cluster them and then make a comparison between the logical clusters and the geographical area.

The outcome will be a list of clusters grouped by most common venue types and where they are located.

Example:

Suppose you are visiting the city and you want to eat a pizza you can check the clusters in order to find how the Pizza Places are distributed

2. Data

In this section I am going to discuss data that will be used.

2.1 Source

My source of data will be Foursquare API. It is very difficult to find free data for this specific task, so I decided to rely on Foursquare because I know how to use the API and because I was able to find a quite satisfactory amount of data to use

2.2 Structure

It is very important to choose the right features to use in order to find what we need to solve our problem.

The most important features are all the ones related to venues; for example name, category, and address. Latitude and longitude are also very important in order to represent the clusters on the map

So a quick recap:

1. Latitude, Longitude - features related to position
2. Venue name, address - descriptive features
3. Venue type - feature that will be the pivot of clustering

2.3 Feature Selection

Foursquare data contains a lot of useless information for my purpose so I had to select only a few features according to the previous paragraph. In the end I kept only:

1. *venue__name:* name of the venue
2. *venue__location__lat:* latitude of the venue, I need it to represent data
3. *venue__location__lng:* longitude of the venue, I need it to represent data
4. *venue__categories__name:* venue category, it is the feature on which rely clustering process
5. *venue__location__address:* venue address, another feature useful for representation

It is interesting to note that only one feature will be used for classification, but I am going to clarify it later.

2.4 Data Cleaning

Despite data being downloaded from one source, they had a bad structure and some missing values. I chose in the Address column to keep only street names without house numbers. So I had to remove house numbers, fix some case sensitivity issues and modify inaccurate addresses. Following some example:

- via roma, 16 → Via Roma
    - “Via” in italian language means street.
    - This is both an example of house number cleaning and case sensitivity problem.
- Salerno → Via Torrione
    - In this case instead I had to change Salerno (which is the name of the city) with the name of the street according to latitude and longitude.

The dataset was quite small, so I easily managed to clean data.
Finally I drop NaN values.

3. Exploratory Analysis

3.1 Understanding the location

I plotted a map which represents all the dataset. As you can see I included also places nearby the city of Salerno (red circles in the figure). More places means more variety and more interesting outcome, they also enrich the small dataset.

3.2 Investigation on venue categories

In 2.3 I decided to use only one feature, Venue Category to create the clusters.
As shown in the table I calculated the occurrences of each feature for each address.

	Address	Bakery	Bar	Beach	Bowling Alley	Brazilian Restaurant	Burger Joint	Café	Cocktail Bar	Cupcake Shop	Diner	Electronics Store	Event Space	Furniture / Home Store	Garden	Gastropub	Historic Site
0	Corso Garibaldi	0.25	0.25	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0
1	Corso Mazzini	0.00	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0
2	Corso Principe Amedeo	0.00	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0
3	Corso Umberto I	0.00	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.25	0.0	0.0	0.0	0.0
4	Corso Vittorio Emanuele	0.00	0.00	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0

Next step is to identify the most common Venue Categories for each Address.

	Address	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Corso Garibaldi	Italian Restaurant	Pizza Place	Bar
1	Corso Mazzini	Hotel	Waterfront	Ice Cream Shop
2	Corso Principe Amedeo	Italian Restaurant	Pizza Place	Electronics Store
3	Corso Umberto I	Hotel	Plaza	Jazz Club
4	Corso Vittorio Emanuele	Cocktail Bar	Other Great Outdoors	Waterfront

Top 3 is enough for my purpose, I tried also for Top 10 but the last positions were made by many venues with only 1 occurrence so I kept only the 1st, 2nd and 3rd ones.

The last step is to prepare a dataset to be clustered which is the outcome of occurrences without address.

	Bakery	Bar	Beach	Bowling Alley	Brazilian Restaurant	Burger Joint	Café	Cocktail Bar	Cupcake Shop	Diner	Electronics Store	Event Space	Furniture / Home Store	Garden	Gastropub	Historic Site	History Museum
0	0.25	0.25	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0
1	0.00	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0
2	0.00	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0
3	0.00	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.25	0.0	0.0	0.0	0.0	0.0
4	0.00	0.00	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0

4. Clustering

4.1 Algorithm

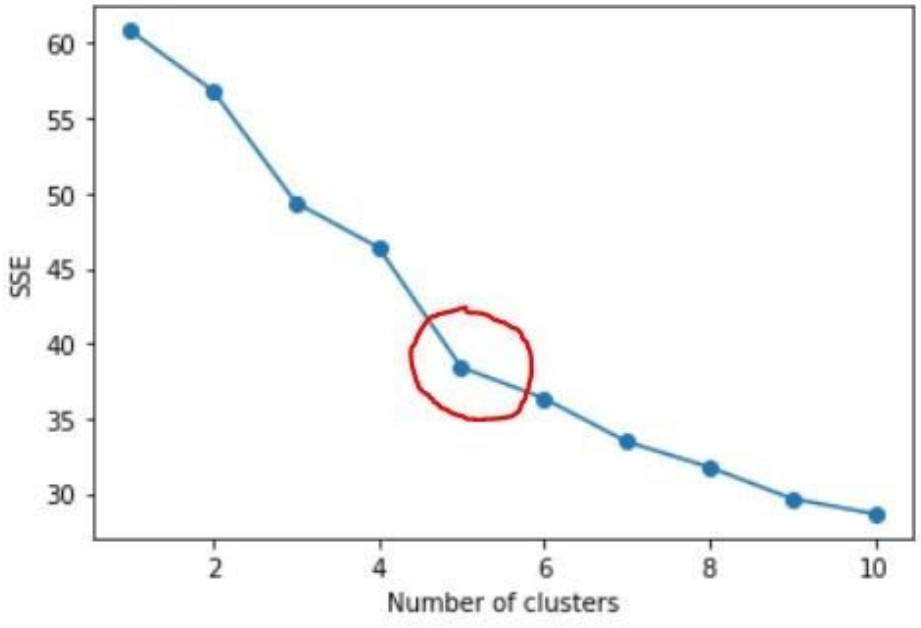
I choose K-Means clustering because it has several advantages:

1. Simple to understand
2. Fast to cluster
3. Easy to implement
4. Always yields a result

4.2 Tuning & Clustering

Despite its neatness K-Means has to be tuned properly. First step is to choose an initialization method between *kmeans++* and *random*, the first one gave the best results so I kept. The second/parallel step is to choose the number of clusters, a good idea is to use the elbow method:

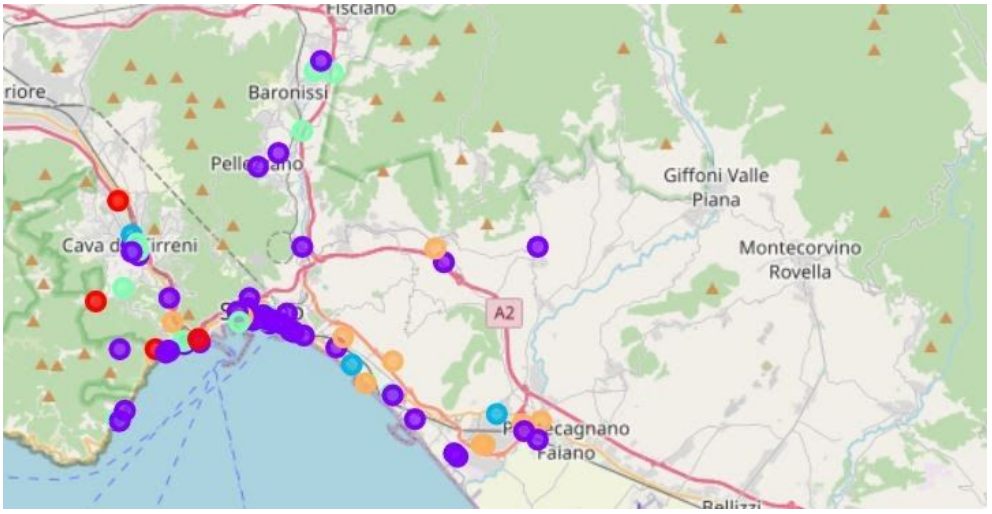
1. Execute the algorithm for a several number of clusters
2. Calculate SSE (Sum of Squared Error) for each execution
3. Plot the result as shown below



According to the plot I estimated 5 as a good number.

Other parameters were not so decisive as the init method and number of clusters.

This is the final outcome:



5. Cluster Analysis

The algorithm produced 5 Clusters:

1. Cluster 0
2. Cluster 1
3. Cluster 2
4. Cluster 3
5. Cluster 4

5.1 Cluster 0

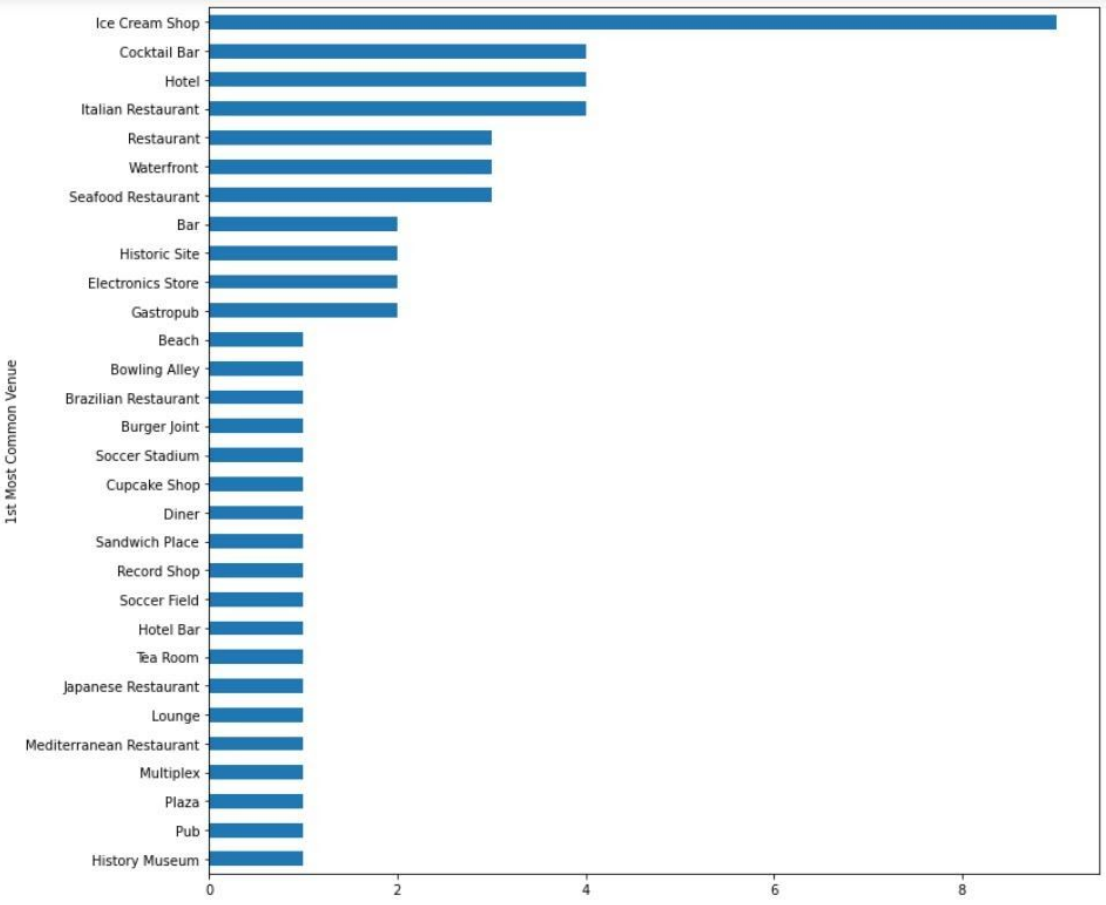
	Address	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
43	Via Nuova Raito	0	Hotel	Waterfront	Ice Cream Shop
68	Corso Mazzini	0	Hotel	Waterfront	Ice Cream Shop
69	Via Enrico De Marinis	0	Hotel	Waterfront	Ice Cream Shop
87	Piazza Risorgimento	0	Hotel	Waterfront	Ice Cream Shop

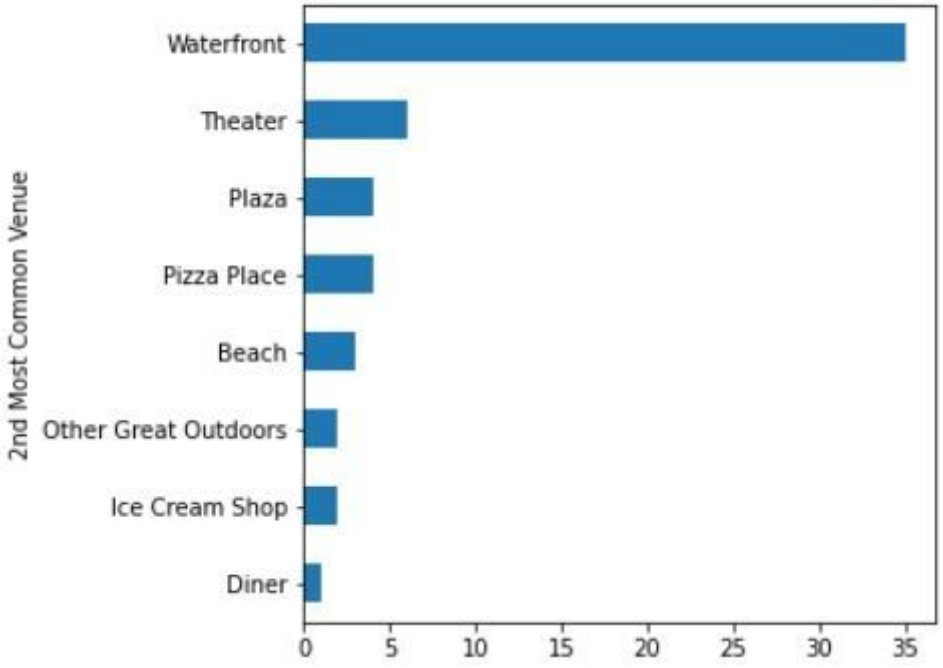
Cluster 0 due to its size is straightforward to analyze, if a tourist is looking for an hotel He can check Cluster 0. These places are outside the center of the city so It is plausible to find hotels, while due to the presence of the coast and summer tourism Waterfront and Ice Cream Shop are predictable places to find here.

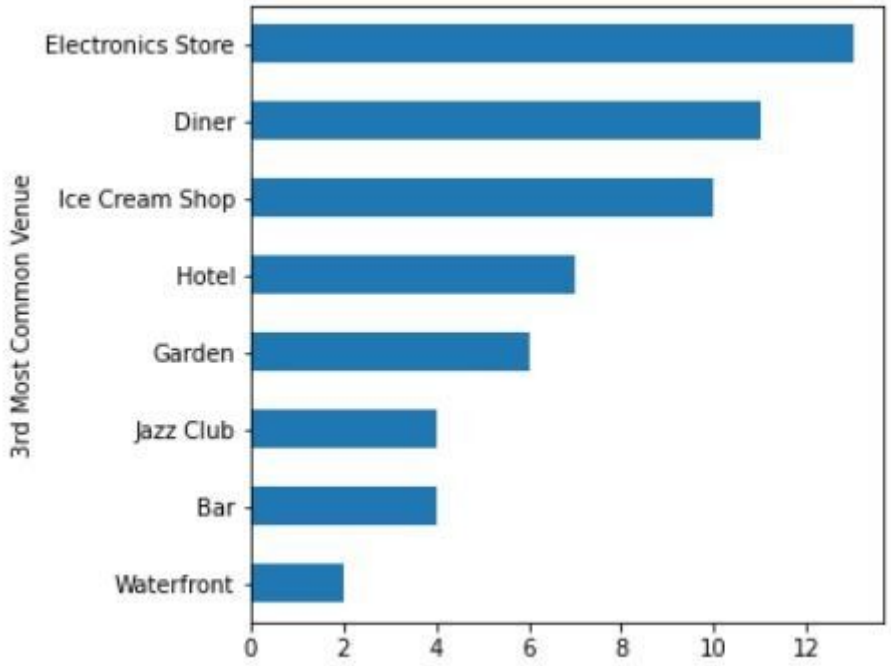
5.2 Cluster 1

Cluster 1 biggest one, so I plot the values. According to the charts readable at the end of the paragraph here there is a more diversified distribution of places. This cluster corresponds more or less to the city center and it includes not only tourism places.

In the first plot there is a huge amount of places to eat and drink (Ice Cream Shops, Cocktail Bars, Italian Restaurant, Restaurant and so on), Hotels and Waterfront. In the second plot Waterfront is the most common place, which means that Waterfront is the 2nd most common place to find in the cluster. In the third plot there is Electronics Store, so in this cluster appears a venue type not strictly related to tourism.


1st Most Common Venue
Ice Cream Shop
Cocktail Bar
Hotel
Italian Restaurant
Restaurant
Waterfront
Seafood Restaurant
Bar
Historic Site
Electronics Store
Gastropub
Beach
Bowling Alley
Brazilian Restaurant
Burger Joint
Soccer Stadium
Cupcake Shop
Diner
Sandwich Place
Record Shop
Soccer Field
Hotel Bar
Tea Room
Japanese Restaurant
Lounge
Mediterranean Restaurant
Multiplex
Plaza
Pub
History Museum
0
2
4
6
8


2nd Most Common Venue
Waterfront
Theater
Plaza
Pizza Place
Beach
Other Great Outdoors
Ice Cream Shop
Diner
0
5
10
15
20
25
30
35



5.3 Cluster 2

	Address	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
31	Piazza Caduti di Brescia	2	Café	Waterfront	Electronics Store
59	Via Vittorio Veneto	2	Café	Waterfront	Electronics Store
71	Via Delle Calabrie	2	Café	Waterfront	Electronics Store

Like Cluster 0 this cluster is also straightforward, I can infer by checking the map and looking at the venues that places in this cluster are outside of tourist scope. In the map they are located in residential neighborhoods and in fact they include places you can find in a residential neighborhood of a city on the sea.

5.4 Cluster 3

Looking at map these places are still in the city center (except for a few outliers) but they are not directly near to the sea. They share in common the presence of an Italian Restaurant which is plausible for a city based on tourism, the same assumption can be made for Hotels. It is also interesting to find a high amount of Ice Cream Shops and few places not related to tourism like Electronics Stores and Home Stores.

	Address	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
3	Via Masuccio Salernitano	3	Italian Restaurant	Ice Cream Shop	Hotel
10	Via Antonio Genovesi	3	Italian Restaurant	Ice Cream Shop	Hotel
40	Via Picenza	3	Italian Restaurant	Ice Cream Shop	Hotel
44	Via Taiani Diego	3	Italian Restaurant	Restaurant	Diner
47	Piazza S.Francesco	3	Italian Restaurant	Ice Cream Shop	Hotel
54	Via Gioacchino Trezza	3	Italian Restaurant	Ice Cream Shop	Hotel
60	Via Sant Andrea	3	Italian Restaurant	Ice Cream Shop	Hotel
72	Corso Principe Amedeo	3	Italian Restaurant	Pizza Place	Electronics Store
75	Via Salvatore Allende	3	Italian Restaurant	Furniture / Home Store	Ice Cream Shop
78	Via Porto	3	Italian Restaurant	Ice Cream Shop	Hotel

5.5 Cluster 4

Cluster 4 is very similar to Cluster 3 but it is distributed on a wider range of the map. The most common venue is Pizza Place, pizza is a typical meal so I expected to find it while the presence of the waterfront means that this place are still near to the sea but like Cluster 2 the presence of Electronics store suggests us that this cluster is distributed to residential areas.

	Address	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
2	Via Romualdo II	4	Pizza Place	Waterfront	Electronics Store
5	Piazza Alfano I	4	Pizza Place	Waterfront	Electronics Store
37	Via Pietro del Pezzo	4	Pizza Place	Waterfront	Electronics Store
38	Via Santa Teresa	4	Pizza Place	Waterfront	Electronics Store
39	Lungomare Colombo	4	Pizza Place	Waterfront	Electronics Store
46	Via Matteo Lecce	4	Pizza Place	Waterfront	Electronics Store
50	Via Sant'Eustachio	4	Pizza Place	Waterfront	Electronics Store
52	Via Molina di Vietri	4	Pizza Place	Waterfront	Electronics Store
61	Uscita Autostradale SA-NA	4	Pizza Place	Waterfront	Electronics Store
65	Via Scavata Case Rosse, 88	4	Pizza Place	Waterfront	Electronics Store
77	Via Tiberio Claudio Felice	4	Warehouse Store	Pizza Place	Electronics Store
79	Via Tevere	4	Pizza Place	Waterfront	Electronics Store

5.6 Quick Recap

M.C.V. : Most Common Venue

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1st M.C.V.	Hotel	Ice Cream Shop	Cafè	Italian Restaurant	Pizza Place
2nd M.C.V.	Waterfront	Waterfront	Waterfront	Ice Cream Shop	Waterfront
3rd M.C.V.	Ice Cream Shop	Electronics Store	Electronics Store	Hotel	Electronics Store

Table contains elements which occur as the majority as 1st, 2nd and 3rd values, check previous paragraphs for more detailed information.

In green I outlined the clusters which could be more interesting for a tourist while in white there are the clusters which contain places located in residential neighborhoods. Cluster 1 looks very similar to Cluster 2 and Cluster 4 but I put in the green group due to its high variety of places which are suitable for tourists.

6. Conclusions

In this study I analyzed venues in the city of Salerno in order to find which kind of attractiveness can find a tourist who visits the city, so this study could be considered as a guide for those who want to have a holiday in the city. In this paragraph I would like to outline the profile of a tourist who could be interested in spending her vacation in Salerno. The city is located near to Mediterranean sea in the south of Italy, this means that the city is characterized by a long hot summer. After this premises I can recommend this city to tourists who:

- are interested in food and wine tourism, because the city offers a wide range of restaurants and pizza places.
- like seasides and sunny places, the presence of waterfront and ice cream shops helps to enjoy the stay despite the very high temperatures.

Furthermore the tourist has a good variety of choice in terms of accomodation, in fact there hotel both in the center of the city and outside of it.

7. Further Developments

I was able to achieve a quite coherent result, but I was limited by the small dataset. To keep this study valid also for the future it is important to find a bigger and up-to-date dataset which maps better not only to restaurants and similar but also other places of interest like historical sites and so on.

Carlo D'Aloia