Opening a new subsidiary of company “A” July 28 2019

1) Introduction:

Background

Company "A" is a small-sized company that remodels and rebuilds old houses and then sells them to an exclusive group of customers. The last past years the company made very good profits and now the whole business is growing quickly. As a result, Company “A” is expanding the business in different places. Mr Jones, the owner, wants to open a new branch in The North of Spain, in Navarra province.

The problem

The new subsidiary of Company "A" must be in the best location according to customer preferences and a specific group of facilities near this place. It is well known that large cities would generally supply most customer needs. However, in large cities, the competence of Company "A" is greater and is not the objective of Mr. Jones. Jones believes that a place outside Pamplona (capital of Navarra) could meet all the requirements. Could Data analysis help Company "A" to take better decisions?

2) Data acquisition and cleaning

For this project we need 4 types of information: 1) Demographic Information (data of Places, Names, Population and coordinates), 2) Data frames with customer preferences, 3) Maps of the region in study, 4) Venues near points of Interest.

Data sources:

1) Demographic Information: we scrap or extract data from wikipedia ('https://es.wikipedia.org/wiki/Merindad_de_Pamplona). In this site we obtain the Name of the cities, population, category, and coordinates.
2) Data frame with venues customer preferences. Company “A” did a survey on costumer’s preferences about venues categories. The result is a list of 154 specific categories (category_costumer.csv).
3) Maps of North of Navarra. We use Folium libraries.
4) Venues near points of Interest. To find venues (R=2Km) near these places, we extract data from Foursquare.

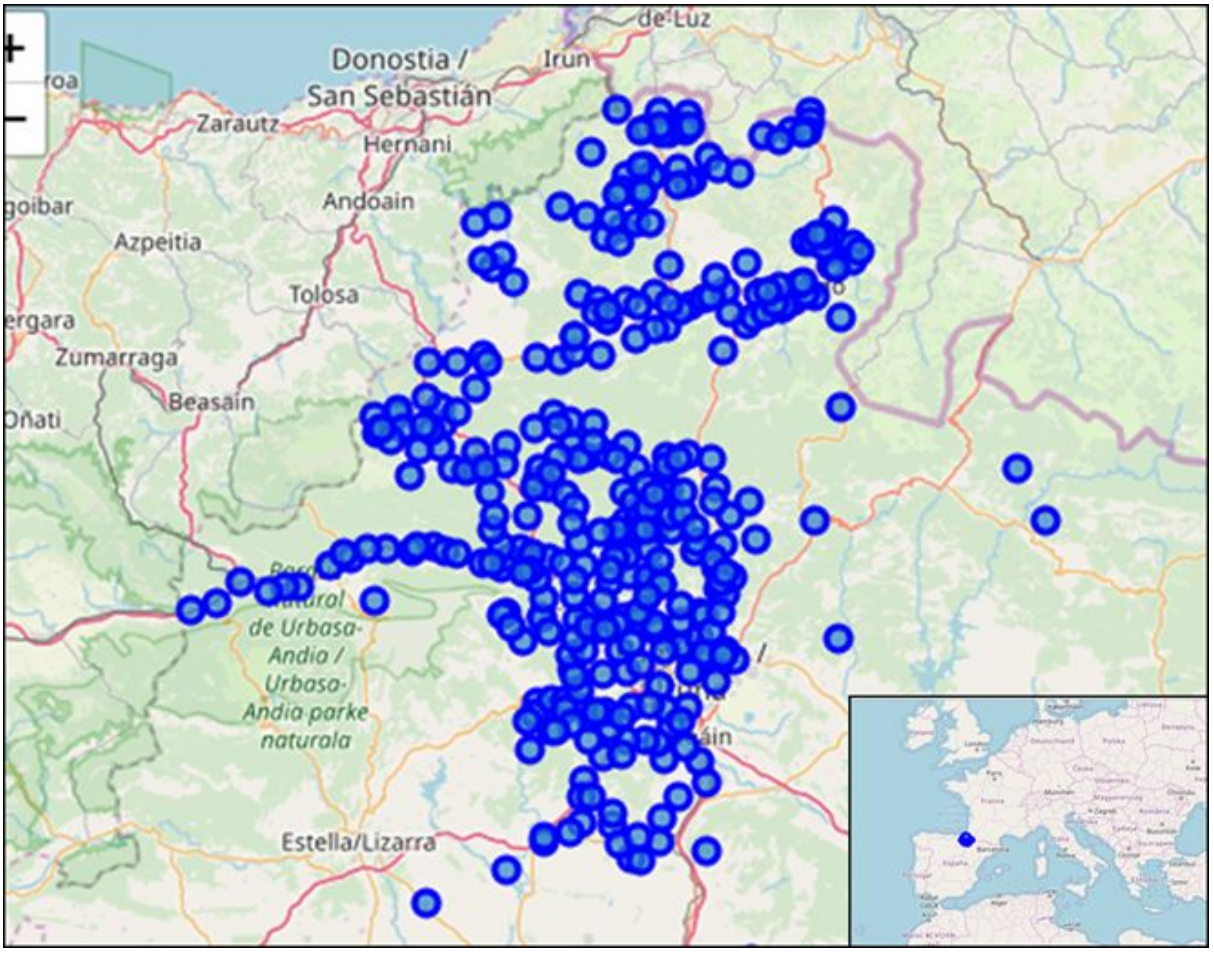
Data cleaning:

First of all we created a Data Frame with relevant information of the places, and delete the missing values (least the 0.5%). The result is a list of 316 unique Village Name, with coordinates and population.

From the Foursquare´s result we got so much irrelevant information for this analysis, around 306 unique categories. So we regrouped and filtered it based on venues categories of the Data Frame of the customer preferences. The result is a Data Frame with relevant venues information near the points of interest into a 2Km radius.

3) Feature selection

The North of Navarra have 392 urban place, hower this includes Villages, Cities and Towns. Based on the population (>30 people) we keep just 316 Places to the analysis. Places with less than 30 are not profitable for the company.



Map of the Point of interest. Places analyzed for company “A” new subsidiary.

From Foursquare we obtain the Venus that match with the customer's preferences in a Radius of 2 Km. The following table is and example of the first 5 Places and the 10 most common category venues for each one.

[13]:

	nombre	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Adiós	Park	Capitol Building	Field	Liquor Store	Yoga Studio	Event Space	Dry Cleaner	EV Charging Station	Electronics Store	Elementary School
1	Adériz	Field	Nightclub	Mountain	Outdoors & Recreation	Campground	Restaurant	Yoga Studio	Event Space	Dry Cleaner	EV Charging Station
2	Aientsa	Hotel	Spanish Restaurant	Yoga Studio	Falafel Restaurant	Driving School	Dry Cleaner	EV Charging Station	Electronics Store	Elementary School	Event Space
3	Aintzialde	Stables	Bar	Falafel Restaurant	Dry Cleaner	EV Charging Station	Electronics Store	Elementary School	Event Space	Factory	Farm
4	Aitzano	Bar	Hotel	Field	Bed & Breakfast	Housing Development	Restaurant	Scenic Lookout	Building	Church	Dessert Shop

3) Data Analysis and clustering

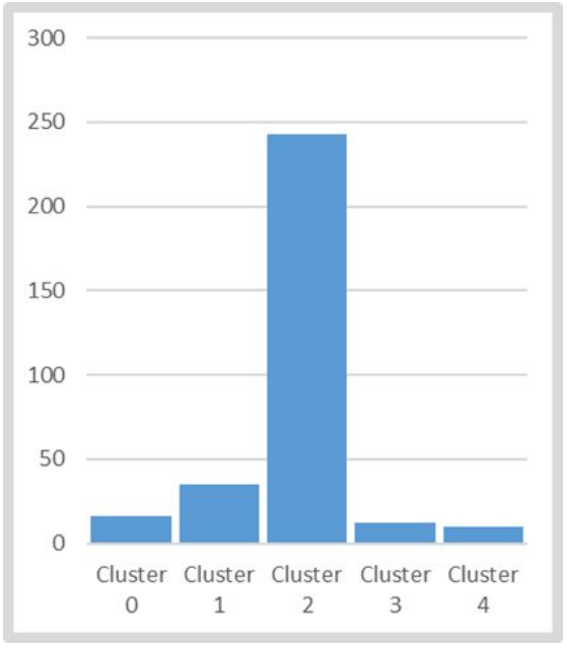
We decided to use a k-mean clustering to model the data and make suggestions based on clusters. We analyzed the best K for the data distribution and the end k=5 makes the best result.

The following table shows the merged Data frame with whole relevant information, cluster categories and how many features are within each one.

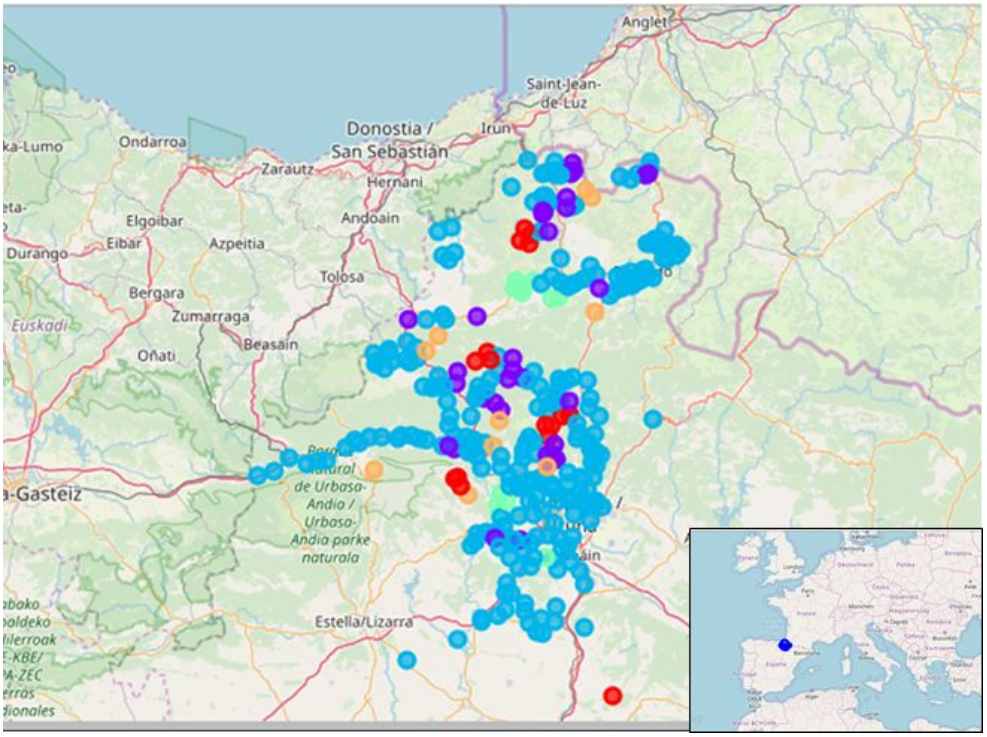
[23]:

	Village_Name	Municip	Category	Population	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
12	Aldatz	Larráun	Concejo	130	43.009755	-1.859486	1	Restaurant	Field	Beach	Yoga Studio	Driving School
13	Alkaiaga	Lesaca	Barrio	201	43.279708	-1.694239	2	Bar	Diner	Spanish Restaurant	Hotel Bar	Building
14	Alkainzuriain	Goizueta	Barrio	24	43.160850	-1.874923	2	Cemetery	Gas Station	Yoga Studio	Dog Run	Driving School
15	Almandoz	Baztán	Lugar	198	43.090510	-1.605399	4	Mountain	Food	Farm	Dry Cleaner	EV Charging Station
16	Alkerdi	Urdax	Barrio	85	43.276305	-1.528310	2	Restaurant	Spanish Restaurant	Bar	History Museum	Food
17	Altsasu/Alsasua	Alsasua	Villa	7612	42.894966	-2.168510	2	Bar	Building	Office	Restaurant	Electronics Store

We can see the data distribution in clusters:

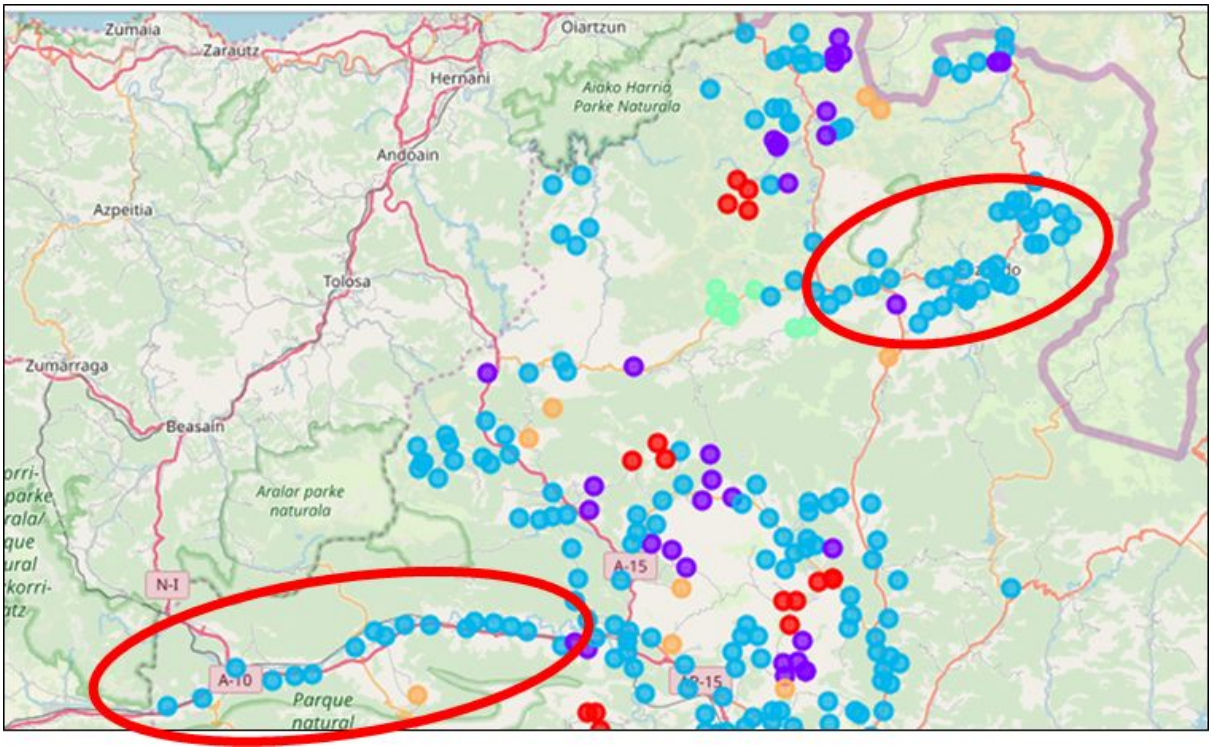


Finally we plot the cluster result, in different colors for each place of interest.



In the introduction of the problem we explain that Pamplona (Capital City of Navarra) is not a strategic place for company “A” because of too much competition. In this stage of the

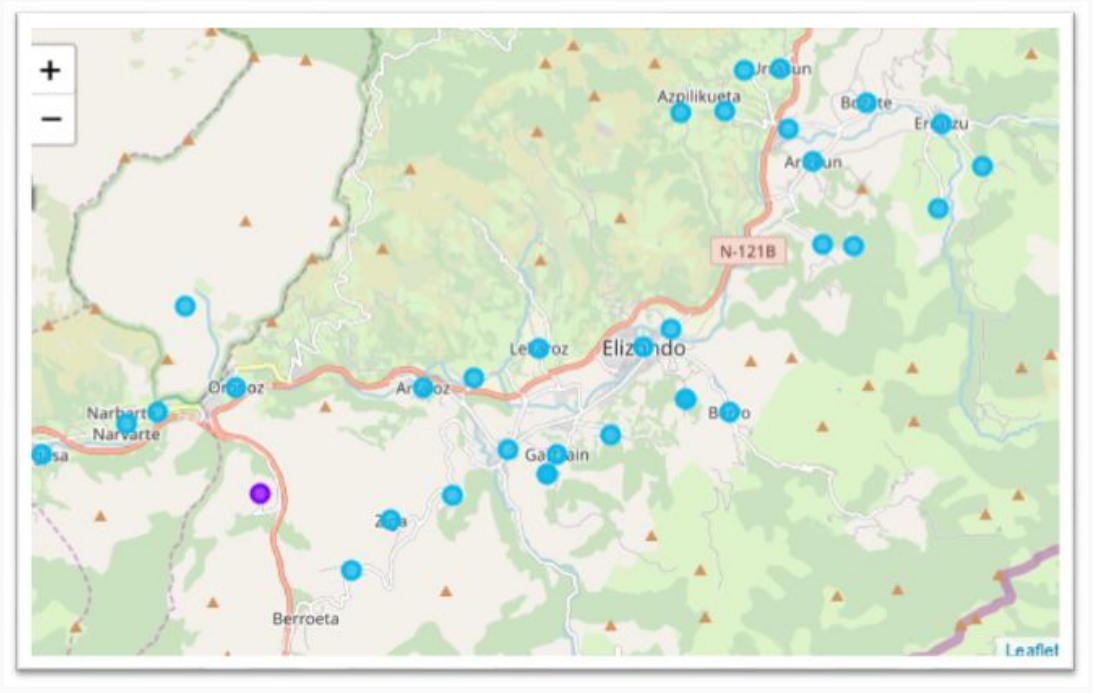
analysis and according to the clustering result, we decide to focus just in places in the North of Pamplona.



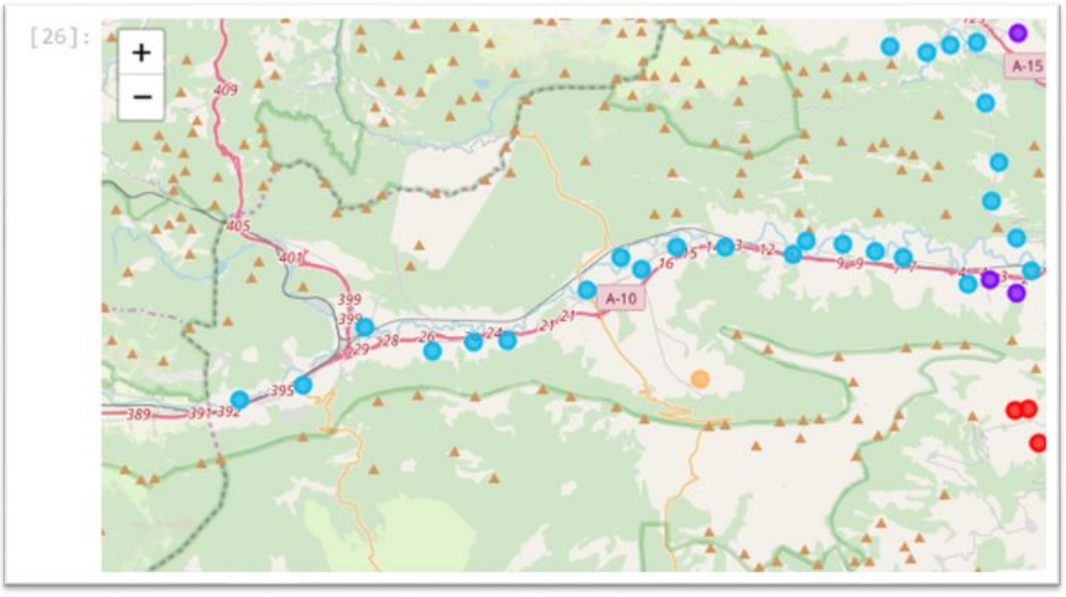
4) Conclusion

- The cluster distribution shows a homogeneous distribution of venues categories in all the region. Places in Cluster 2 (Blue points) is the most common for North of Navarra and supply the customer and Company “A” requirement.
- Because of the homogenous of the data is not easy to find a specific place to open the new subsidiary. So many places have more or less the same venues categories and facilities. However we can identify two areas (red marks) where lots of blue points are concentrated.
- We suggest expanding the study with economic profitability data, focusing just on these two regions. Region (1) = Close to “Elizondo” town. Region (2)= And some point in the A10 High way.

Area 01 “ Close Elizondo Town”



Area 02 “ Middle of the way “A10” between Vitoria City and Pamplona City.



5) Recommendation

- In order to select the best place to Company “A” is necessary to complement this analysis with economic features.