CUNDINAMARCA POTENTIAL MARKET 1.0

2021

12 FEBRUARY

Creado por: Gabriel G. Amaya

CONTENT

1. Introduction

1.1. Business Context

A technology company located in Cundinamarca - Colombia has recently developed a new business line know as Analytics - AI, this because the company has seen the change in the trend respect that companies not only need the engineer structure but also the real implementation of AI technologies into the companies.

1.2. Business Problem

The company is new in this line, however based on its own history they know that once they develop a solution it is possible to offer it as a sector solution, in that way they want to know what the most common types of industries and their location in the way is to put their efforts in the most common sector, but also to make a business plan with the second and third common.

2. Data

2.1. Data Sources

In the way to accomplish the goal we need to get the information about all the towns that formed Cundinamarca state, also the location latitude and longitude of each of them, this information is going to use to get the information about venues through foursquare API.

The information about all the towns that formed Cundinamarca state is available by DANE, the type of document is a pdf which has these columns:

a. Codigo_Depto: Code to identify each department.
b. Nombre_Depto: Name of each department.
c. Provincia: Group by each province where town is located inside each department.
d. Codigo_Municipio: Codes of each town.
e. Nombre_Municipio: Name of each town.

Based on it we now know that Cundinamarca is formed by 116 towns and by Bogotá that is a special district, now we need to aggregate latitude and longitude by each town, using geocoder library we get this information.

Using the foursquare API we get the following data set:

a. Codigo_Municipio: Codes of each town.
b. Codigo_Municipio Latitude: Latitude where town is located.
c. Codigo_Municipio Longitude: Longitude where town is located.
d. Venue: Name of the company.
e. Venue Latitude: Latitude where company is located.
f. Venue Longitude: Longitude where company is located.
g. Venue Category: Group or sector where company take part.

2.2. Data Cleaning and Feature selection

Taking to account that information about Cundinamarca is clean we do not need to do more than merge information about each latitude and longitude inside the first data set.

In contrast information about Foursquare requires a little bit cleaning, specially because inside the information we get there are some venues that will create a noise, for that reason we remove the following venues:

a. Park
b. Plaza
c. Farm

The main reason to remove them is that based on the structure of towns in Colombia it is very normal to find in the middle of the town a central park or main plaza, and that farms are very common in these places for that reason we prefer to remove to avoid the possible bias produced by each characteristic. The following table show us the top 12 venues before the cleaning process describe before.

Top 12 venues

Venue Category	Codigo_Municipio
Coffee Shop	11
Fast Food Restaurant	11
Latin American Restaurant	14
Bakery	14
Pizza Place	14
Burger Joint	15
Café	16
Hotel	16
BBQ Joint	25
Plaza	26
Restaurant	45
Park	47

Resource: Own develop

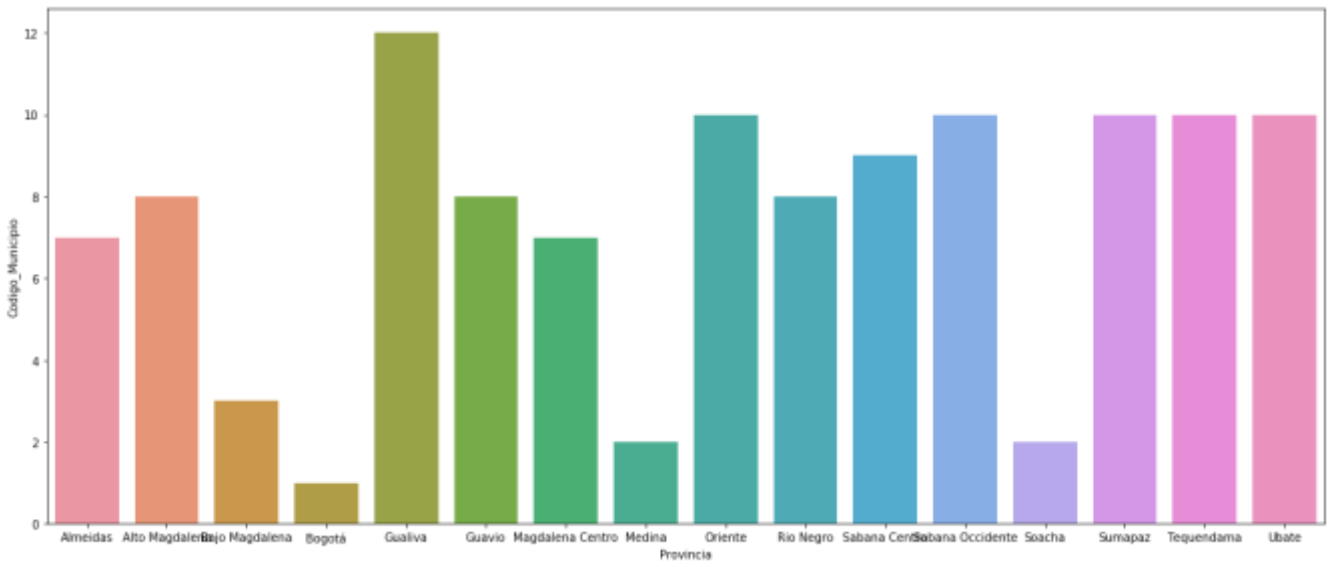
Once both data set were cleaning and merge, we have all the resources to analyze in deep the information and create our cluster.

3. EDA

First, we try to understand how the 116 towns are divided by provinces, in this line we create a bar plot when we get the following elements:

a. Gualiva is the province that has the greatest number of towns following by Oriente, Sabana Occidente, Sumapaz, Tequendama and Ubate.
b. Bajo Magdalena, Medina and Soacha are the towns with lesser number of towns.
c. Bogotá is a special district in that way there do not have more towns or cities inside it.

Distribution of towns by each Province



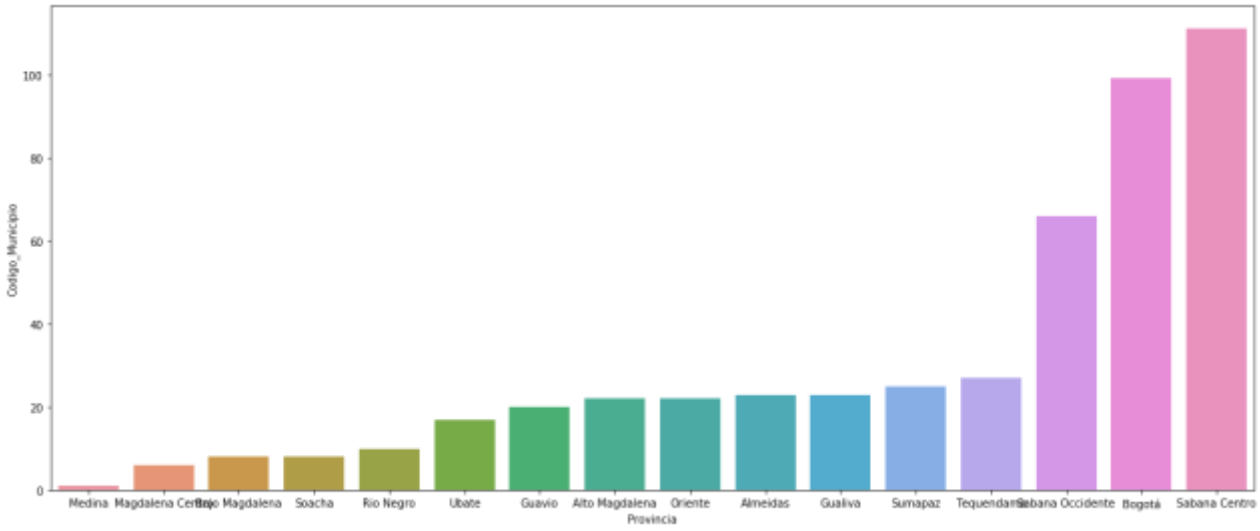
Resource: Own develop

Second, we get the information about venues inside each province, it is important to highlight that based on latitude and longitude that normally is in the middle of the town we get the venues as a 1 km around, for that reason Bogotá has not all the places, because the latitude and longitude is around 26 avenue, similar things will happen with other large towns.

In that way we collect the information and at difference with the previous chart here the provinces with more information are:

a. Sabana Centro: Composed by 9 towns
b. Bogotá that is a special district
c. Sabana Occidente: Composed by 10 towns

Quantities of venues collected by each Province.



Resource: Own develop

4. Predictive Modeling

4.1. Assumptions and model

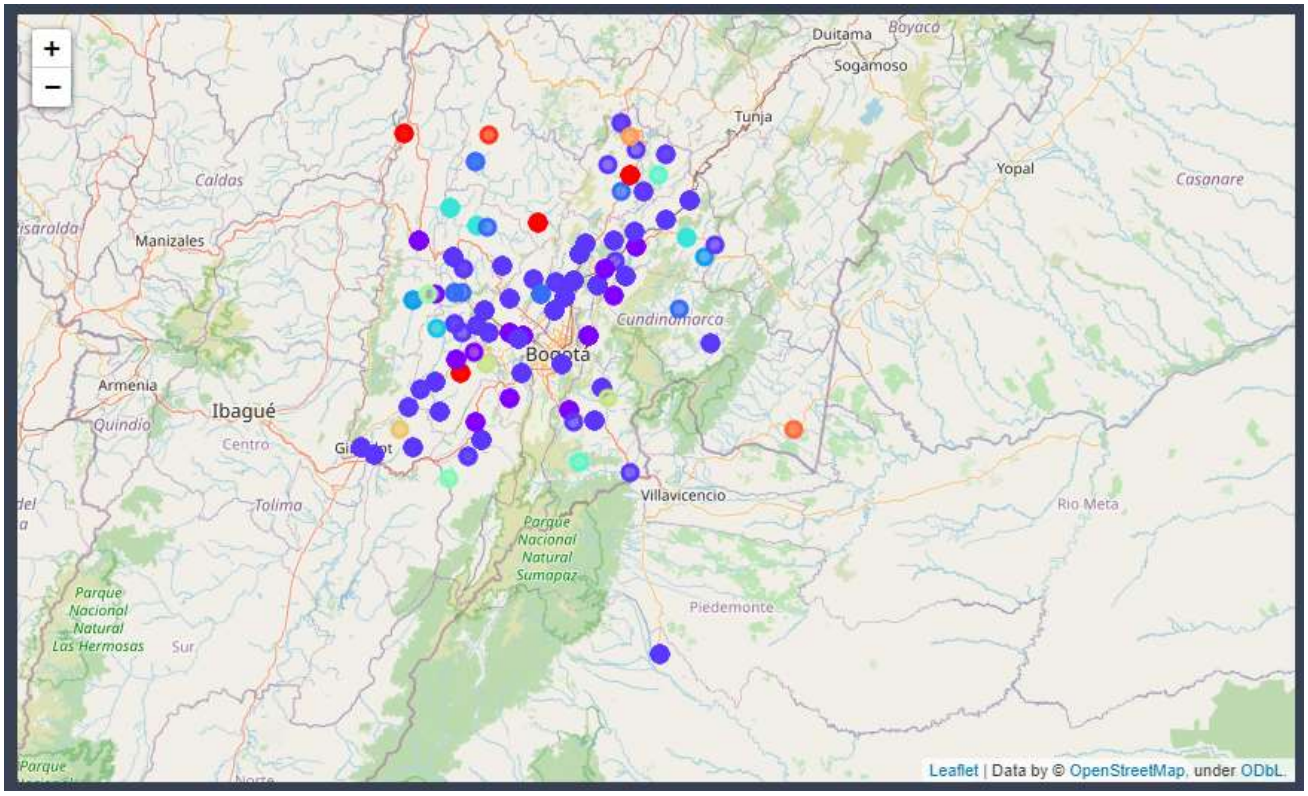
Based on information, we decided to create a K-means cluster to know which towns are like others, this would help the company to develop a product that can be offered in more than one town, for that reason an trying to avoid a great generalization we decided to create a 15 cluster with the expectative that almost each cluster has around 7 towns inside.

Also, as we explain before we clean the data set removing the most common venues which would be make difficult to the algorithm to classify, finally we assume that the point we select (latitude and longitude) summarize the characteristics of each town.

4.2. Results

The follow image is the result of cluster model, as we see the towns has been classified by each characteristic into 15 clusters.

Cluster results inside the Cundinamarca´s map



Resource: Own develop

About the cluster result we get:

- ✓ We see that distribution is not equal through different clusters, this will be explained because most towns in Cundinamarca have similar structures, like a church, a central plaza or central park and around it the people develop each company.
- ✓ The clusters that get the most quantities of information are cluster 0, 1, 2, 3 and 6, we are going to analyze in deep each cluster, in the way to help the company to know what are the characteristics that they have, and the company will focus on develop one solution for them.
- ✓ There are 6 cluster than only have 1 town, this is not a good result because that means that cluster can not summarize the data in a good way.

Table of cluster distribution towns

Cluster Labels	Codigo_Municipio
0	14
1	51
2	381
3	12
4	3
5	1
6	13
7	2
8	1
9	3
10	3
11	1
12	1
13	1
14	1

Resource: Own develop

Now we are going to analyze the main clusters, the results are:

- ✓ Cluster 0 is based on food venues, the most common are Latin American Restaurant, Soup place, Food court and Diner places.
- ✓ Cluster 1 has multiple kinds of places like multiplex, shops, foods, mountain, and history, in that way it will be a good idea to filter these places and run again a cluster to get better ideas about kinds of venues and segmentation for each group of towns.
- ✓ Cluster 2 is more related with restaurants and coffee places, also another important thing to highlight is related with Bodega which will give us to think that here we will find some medium and big companies.
- ✓ Cluster 3 is more related with home shops and wings joint.
- ✓ Cluster 6 is more related with outdoor activities like rafting, mountain, lake, and pool.

5. Conclusion

5.1. General

Cluster towns is a good idea to learn more about common places when the strategy is to develop a possible common solution with easy implementation, it would be seen like a potential market analysis.

Based on cluster results we would say:

- ✓ The company will focus on analyze in deep the idea to make a solution for food companies, specially for restaurants where they can create a technological solution for getting orders but also to learn more about their clients.
- ✓ The company will evaluate in deep the cluster 2 specially venues bodegas, because in that places they will find medium and big companies that would be clustering on similar purposes.
- ✓ The company would see what kind of venues are more interesting, that would help to create better cluster and summarize, because as we see before the cluster 1 has a lot of different places were the company will create a solution like multiplex and history, but this cluster do not give the possibility to know the real potential market that company will have.
- ✓ Finally, the company will think how since technology they can help outdoor activities, however it is important to say that normally when people want to do these activities they prefer to be disconnected (not use technology) in that way this would be a risk idea.

5.2. Recommendations

The main recommendations we make for the company are:

- ✓ Based on this first approach to potential market the company will help to create filters based on their interest, these filters will help to run again the model and would create better clusters and potential market.
- ✓ Independent the sector that they select to create a solution, it is very important to make a evaluation about needs, movements and the client disposition to acquire and implement a new TI solution, because as we discuss before venues related with outdoor activities not always are related with technology connection.