This is a report for the Final Capstone of IBM Data Science Professional Certificate. It covers all the process from the problem upto the conclusion. Notebook and presentation are also available.

Where to put a chocolate shop in Argentina

IBM Data Science Professional Certificate Final Capstone

Agustín Ianchina

Content

Introduction

The business problem to face is the following:

"An investment group, experienced on chocolatery industry, wants to invest in Argentina. Since they are not familiar with this industry in the country, they are trying to figure out where is it a more convenient location to invest on and of course, opening a new chocolate shop."

The work will be done under the paradigm that clusters of some specific business attract customers looking for that specific product or service. Thus, though financial or economic analysis will not be performed, it will be assumed that clusters are preferable locations vs outliers.

In order to find out where should be more rentable to open a new chocolate shop, current and actual clusters need to be differentiated from outliers. The current clusters will show where are the current productive areas. The outliers will show isolated chocolate shops, which of course means that these are not productive areas for chocolate shops.

The clustering algorithm used is DBSCAN. This algorithm allows to differentiate outliers from clusters.

Data

The data used is the following:

1) Wikipedia for all cities from Argentina.
2) ArcGIS library to geocode Argentinian cities.
3) Foursquare to find all chocolate shops.

Methodology

In this section it will be described what has been the methodology used for the analysis.

Cities from Argentina are got from Wikipedia. A dataframe will be created with all cities from Argentina, and thereafter, they are geolocated using ArcGIS. This results in a DataFrame with all cities, and their latitude and longitude. A map is added to the notebook to see them.

Dataframe with Argentinian cities extracted from Wikipedia.

	City	Province
0	Buenos Aires	Autonomous city
1	Córdoba	Córdoba
2	Rosario	Santa Fe
3	La Plata	Buenos Aires
4	Mar del Plata	Buenos Aires

Dataframe with Argentinian cities Geocoded.

	City	Province	Latitude	Longitude
0	Buenos Aires	Autonomous city	-34.60849	-58.37344
1	Córdoba	Córdoba	-31.40718	-64.18571
2	Rosario	Santa Fe	-32.96780	-60.65924
3	La Plata	Buenos Aires	-34.91393	-57.94636
4	Mar del Plata	Buenos Aires	-37.99741	-57.54846

Map with Argentinian cities

This dataframe, is used in the following manner: each city with its latitude and longitude, is used to perform a search request to Foursquare API. The search request consists in looking for chocolate shops, within a radius of each city. Chocolate Shop is an actual Foursquare category id. Since the location is the relevant information we need from Foursquare, latitude and longitude from each chocolate shop are extracted.

A function is used to create a dataframe which has for each city, all the chocolate shops with their latitude and longitude. A map is added to the notebook to see them.

The radius used is 50km.

Dataframe with all chocolate shops geocoded.

	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude
0	Buenos Aires	-34.60849	-58.37344	Havanna	-34.606191	-58.375046
1	Buenos Aires	-34.60849	-58.37344	Rapa Nui	-34.599218	-58.513947
2	Buenos Aires	-34.60849	-58.37344	Havanna	-34.615760	-58.456093
3	Buenos Aires	-34.60849	-58.37344	Rapa Nui	-34.764566	-58.401615
4	Buenos Aires	-34.60849	-58.37344	Havanna	-34.611148	-58.363783

Map with All chocolate shops.

This last dataframe is used to run the DBSCAN cluster. A map is added to see the clusters distributed all along the country.

The parameters for the DBSCAN are eps = 0.5, and minimum samples are 3.

Dataframe including Clus_Db. This column has the cluster value for each chocolate shop.

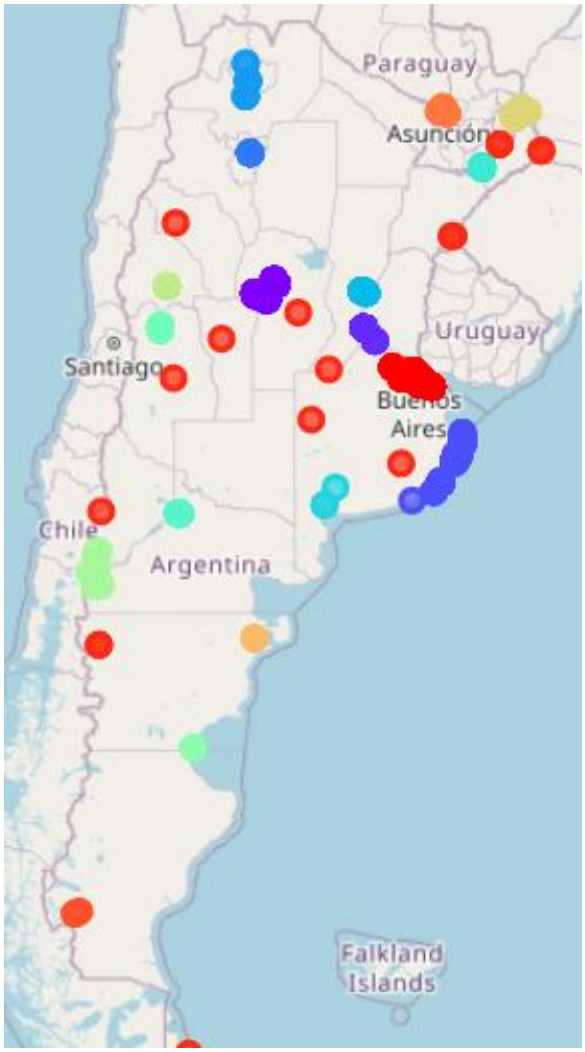
	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Clus_Db
0	Buenos Aires	-34.60849	-58.37344	Havanna	-34.606191	-58.375046	0
1	Buenos Aires	-34.60849	-58.37344	Rapa Nui	-34.599218	-58.513947	0
2	Buenos Aires	-34.60849	-58.37344	Havanna	-34.615760	-58.456093	0
3	Buenos Aires	-34.60849	-58.37344	Rapa Nui	-34.764566	-58.401615	0
4	Buenos Aires	-34.60849	-58.37344	Havanna	-34.611148	-58.363783	0

All cluster values. '-1' corresponds to the outliers.

```
clusters
```

```
array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, -1, 12, 13, 14, 15,
       16, 17, 18], dtype=int64)
```

Map showing all the clusters. Since it is a folium map, it can be zoomed in and out to better understand the distribution.



Conclusion and further analysis

A map is the outcome of the analysis. The investor group now has a clear picture of where the chocolate shop clusters in Argentina.

The next step is to perform economic and financial analysis and try to figure out where is more likely to have a successful chocolate shop.