Predicting Investment Options with venue focus in Ecuador

David A. Espinosa G.

Coursera / IBM

Capstone Project “The Battle of the Neighborhoods” (Week 1)

Instructor: Dr. Alex Alkson

November 16, 2020

Predicting Investment Options with Capital Optimization in Quito City

1. Introduction

1.1. Background

Worldwide pandemics caused by SARS-CoV-2, more widely known as Coronavirus or Covid-19, has caused a dramatic impact at all levels, becoming a major threat to people's health, and causing millions to stay quarantined as, *in time*, it was the official recommendation from the World Health Organization (amongst others). Unfortunately, and as a side consequence, global economics was heavily impacted for the sudden stop in the daily commercial *peer to peer* activities, while other businesses (mostly internet based) flourished as a result of the same cause. This hit to economics was even worse in developing countries, which are usually less prone to have contingency plans for these scenarios, having to improvise "on the go" on an already poor economic system. One of these countries is Ecuador, which according to the website of the Human Development Reports by the United Nations, is placed at the 85th rank; additionally and according to the Ecuadorian Ministry of Labor, 67.9% of the work places are considered to be "informal" which groups both freelance or familiar *peer to peer* projects, and street sales of several categories of products (both of which do not have any kind of contract or assured economic income, but rely entirely on their sales). Speaking of the freelance projects, the investors are very meticulous when deciding what to invest in, as the usually end up tied to that investment for long periods of time, and usually rely on *gossip of what's popular at the moment to decide the investment field*, rather than an actual market study. Knowing which are the current popular offers in some town as well as if that town is similar to the one currently living, would increase dramatically the odds of success in the investment process.

1.2. Problem

The data to approach this problem, involves a knowledge of the current venues in Ecuadorian Capital Cities (i.e., Ecuador is divided in Provinces, as Canada is). The principle is simple: investors usually look for towns that have low market competition, high vairety and as similar as their home town. These aspects are to be kept in mind in the research to come.

1.3. Interest

Small investors will be very interested in the outcome of this project, as this could help them to choose a location near to them to start their new businesses, as well as several options for them so they don't "overlap" in their decision.

2. Data Acquisition and Cleaning

2.1. Data sources

For venues search, Foursquare API will be used, which usually returns venues and / or reviewers data based on the geographical coordinates. For the cities, scrapping was done from Wikipedia. And for the geographical coordinates, GeoPy was used.

2.2. Data cleaning

The data downloaded from Wikipedia was not quite difficult to collect, however the size in km^2 of each capital city was not present, and therefore, another scrapping layer was added (see one example here). Even so, there were some cities that had missing information, but since the surface was important, to (later on) define a proportional search radius, the couple of missing samples were manually added. After that, having the city names, the geographical coordinates started. It was soon noticed that, due to the small size of the country, added to the fact that there are (at points) dozens of cities named the exact same way, the convention [Country] / [Province] / [City]

had to be used (and eventually succeeded). Finally, and since capital cities vary widely in size, a proportional measurement per each city was done, basically approximating the surface of each city, to the one of a circumference of the same size; hence, each radius was extracted. After that, the starting dataset was very much clean and ready to continue with the next stages.

2.3. Feature Selection

Much of the original features were discarded, as they were related with each province data, rather than the capital cities; it was soon noticed that the ration between population of the province versus population of the capital city do not bring a linear relation, and thus were discarded. There were other redundancies noted in the previous section, and summarized in Table 1.

Table 1. Simple feature selection during data cleaning.

Kept Features	Dropped Features	Reason for dropping features
Province, City, Population	Población, Área, Densidad, Hab. (2020)[3], Área (km²), Densidad (hab./km²), Cantones, Fundación, Bandera, #	Not directly related with the capital cities themselves
SearchRadius	KM2	Used temporarily to build the approximate size per each city
Latitude, Longitude		

3. Exploratory Data Analysis

3.1. Collection of venue data

Taking each city and their corresponding latitude and longitude, the venue data (Venue, Venue Latitude, Venue Longitude, Venue Category) was obtained, using Foursquare API. To make sure the locations were correct, they were displayed in the map, using GeoPy (Figure 1). Next, having made sure the locations were correct, it was necessary to extract all the venues available in each of those cities. There were fewer venues than expected (643 samples) for a reason explained later in the “Results” section.

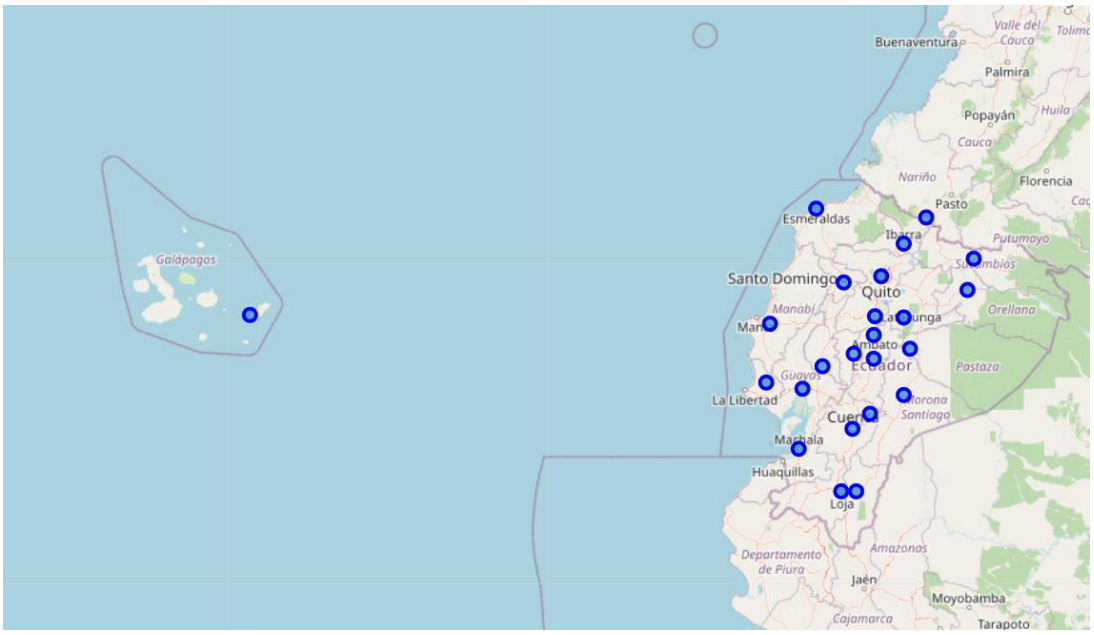

Figure 1. Preliminary location of Ecuadorian Capital Cities.

After some quick inspection, it was seen that there were some non-profitable venues, that were filtered also, leaving a final result of 465 samples.

465 venues found

	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Cuenca	-2.89704	-79.003948	Tiestos	-2.900844	-79.001314	South American Restaurant
1	Cuenca	-2.89704	-79.003948	Mediterraneo	-2.901106	-79.004264	Italian Restaurant
2	Cuenca	-2.89704	-79.003948	El Mercado Cuenca	-2.901356	-79.005226	Comfort Food Restaurant
3	Cuenca	-2.89704	-79.003948	Hotel Santa Lucia	-2.897900	-79.002874	Hotel
4	Cuenca	-2.89704	-79.003948	Mangiare Bene	-2.895692	-79.009809	Italian Restaurant

Figure 2. Dataset after venue collection stage

3.2. One hot encoding and aggregating function

The target of this study is to set groups of cities, based on the venue offer; since there is no preset label, the most proper approach would be a clustering algorithm. To set sail into this task, and based on the information obtained in the previous stage (Figure 2), as well as keeping in mind that we want to obtain the type of city based on the venue category, then for this stage we will start by encoding such venue category with one-hot encoding.

	City	American Restaurant	Argentinian Restaurant	Arts & Crafts Store	Arts & Entertainment	Asian Restaurant	BBQ Joint	Bagel Shop	Bakery	Bar	...	Sporting Goods Shop	Sports Bar	Steakhouse	Supermarket	Sushi Restaurant
0	Cuenca	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
1	Cuenca	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
2	Cuenca	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
3	Cuenca	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
4	Cuenca	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0

Figure 3. One-Hot Encoding detail result

For each city, we will group the by city, to the proceed to extract the mean per each venue category; the result would be obtaining a proportional weight (percentage actually) per each category, to easily visualize later on.

	City	American Restaurant	Argentinian Restaurant	Arts & Crafts Store	Arts & Entertainment	Asian Restaurant	BBQ Joint	Bagel Shop	Bakery	Bar	...	Sporting Goods Shop	Sports Bar	Steakhouse	Supermarket
0	Ambato	0.000000	0.000000	0.000000	0.000000	0.0	0.052632	0.0	0.000000	0.026316	...	0.0	0.0	0.026316	0.000000
1	Babahoyo	0.000000	0.000000	0.000000	0.000000	0.0	0.250000	0.0	0.000000	0.000000	...	0.0	0.0	0.000000	0.000000
2	Cuenca	0.016129	0.016129	0.032258	0.016129	0.0	0.064516	0.0	0.016129	0.000000	...	0.0	0.0	0.016129	0.016129
3	Esmeraldas	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.0	0.000000	0.000000	...	0.0	0.0	0.000000	0.000000
4	Guaranda	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.0	0.000000	0.285714	...	0.0	0.0	0.000000	0.000000

Figure 4. Dataset result after applying aggregating function (mean value per city)

3.3. Preliminary Visualization of Top Venues and Reports

Based from he obtained dataset, it was possible to obtain some preliminary results, for example, we obtained the top (most popular) venues present in Quito – City (capital city of Ecuador); as it is a heavily touristic city, hotels and several categories of restaurants were present, however noticing the presence of supermarkets also (Figure 5).

With the intention of possibly automating the report generation in the future, functions for getting top venues and reports, were also implemented. We obtained some samples to test these functions (Figure 6).

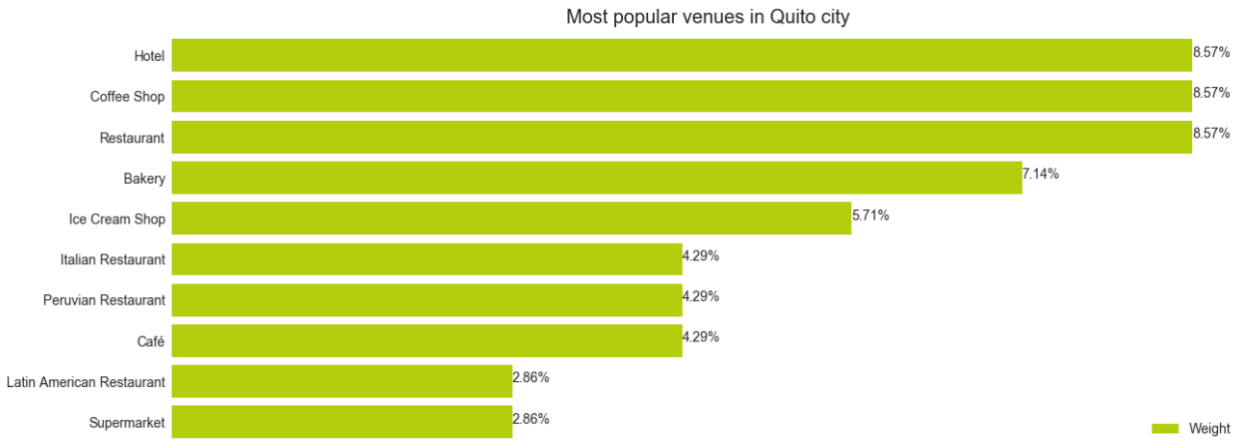

Figure 5. Top (Most Popular) venues in Quito City

	City	Top Venue # 1	Top Venue # 2	Top Venue # 3	Top Venue # 4
0	Ambato	Hotel	Mexican Restaurant	Latin American Restaurant	Pizza Place
1	Babahoyo	Pharmacy	BBQ Joint	Burger Joint	Fast Food Restaurant
2	Cuenca	Italian Restaurant	Restaurant	BBQ Joint	Latin American Restaurant
3	Esmeraldas	Hotel	Sandwich Place	Seafood Restaurant	Women's Store
4	Guaranda	Bar	Mountain	Coffee Shop	Cocktail Bar

Figure 6. Output of function to get the Top (Most Popular) venues in Ecuador

These functions were not only capable of achieving the Top Venues, but also the Interesting Venues, which are the ones less popular, for investment purposes, very interesting, as they would virtually have way less competition within the mentioned city.

4. Clustering

4.1. Clustering process and report

For clustering, to determine a rather good value for k, we have used the K-Elbow method, facilitated by Yellowbrick Library, to speed up analysis process (Figure 7). The optimal value obtained was k = 4. With this parameter, we decided to use K-Means algorithm to have the results needed, visualized on the map (Figure 8).

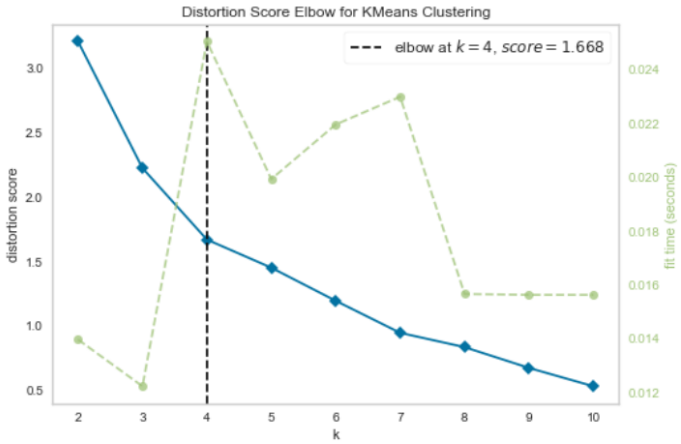

Figure 7. Calculation of "K" in K-Means algorithm, using Yellowbrick Library

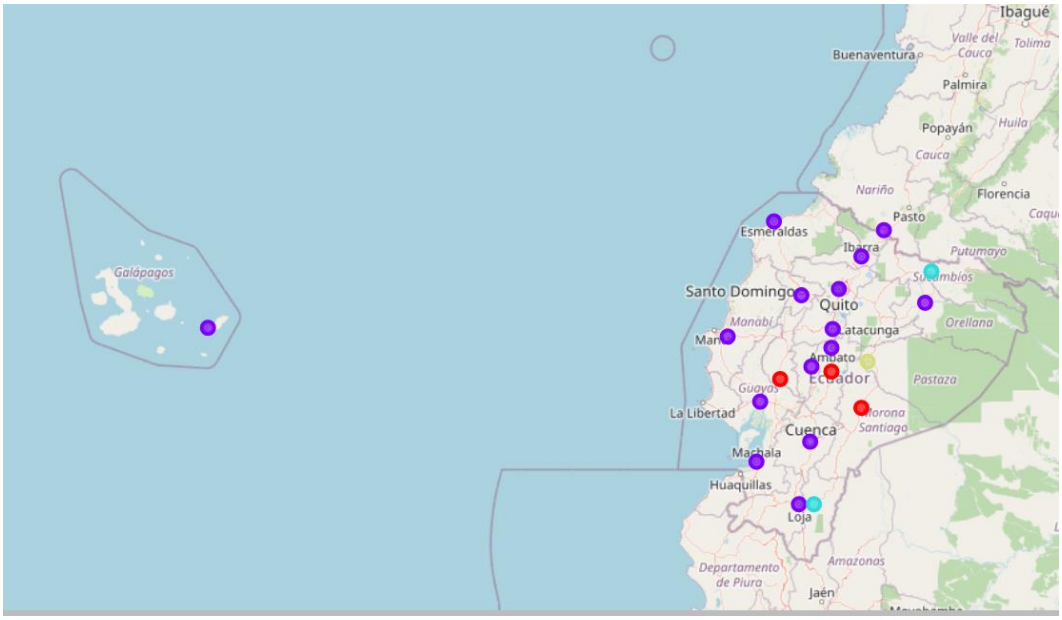

Figure 8. Visualization of Clusters obtained, via GeoPy

For the last step, a summary report was obtained, merging several preliminary dataframes.

	Province	City	TypeOfCity	Population	Latitude	Longitude	Top Venue # 1	Top Venue # 2	Top Venue # 3	Top Venue # 4
0	Azuay	Cuenca	1	329 928	-2.897040	-79.003948	Italian Restaurant	Restaurant	BBQ Joint	Latin American Restaurant
1	Bolívar	Guaranda	1	23 874	-1.592290	-79.001561	Bar	Mountain	Coffee Shop	Cocktail Bar
2	Carchi	Tulcán	1	53 558	0.811929	-77.717108	Chinese Restaurant	Food & Drink Shop	Latin American Restaurant	Hotel
3	Chimborazo	Riobamba	0	146 324	-1.673148	-78.648646	Hotel	Fast Food Restaurant	Burger Joint	Bar
4	Cotopaxi	Latacunga	1	63 842	-0.933621	-78.615049	Restaurant	Pizza Place	Hotel	Fried Chicken Joint

Figure 9. Summary report after clusterization

4.2. Cluster analysis

After the clusterization process, it is also important to analyze about the differences between these clusters; some graphs were made, and the results will be reported in the next section.

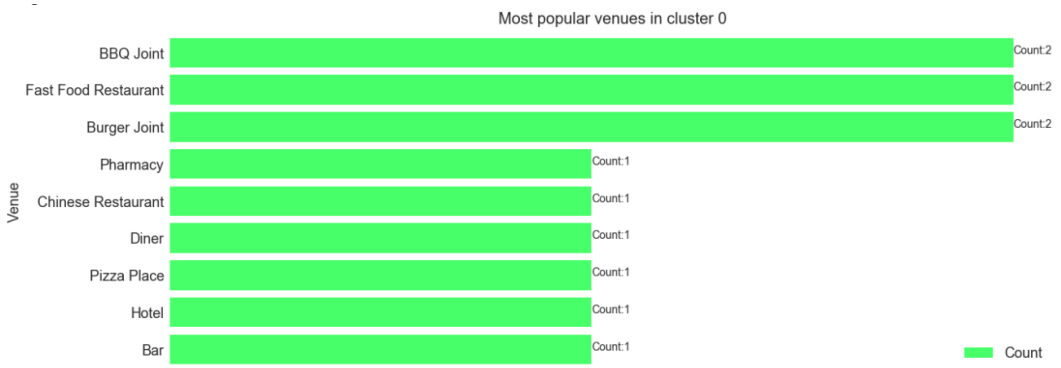

Figure 10. Graphical report: Cluster Type 0

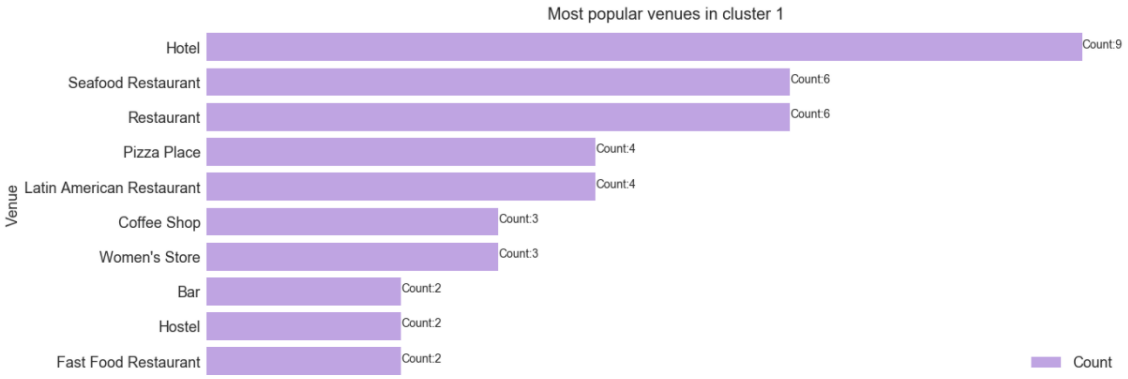

Figure 11. Graphical report: Cluster Type 1

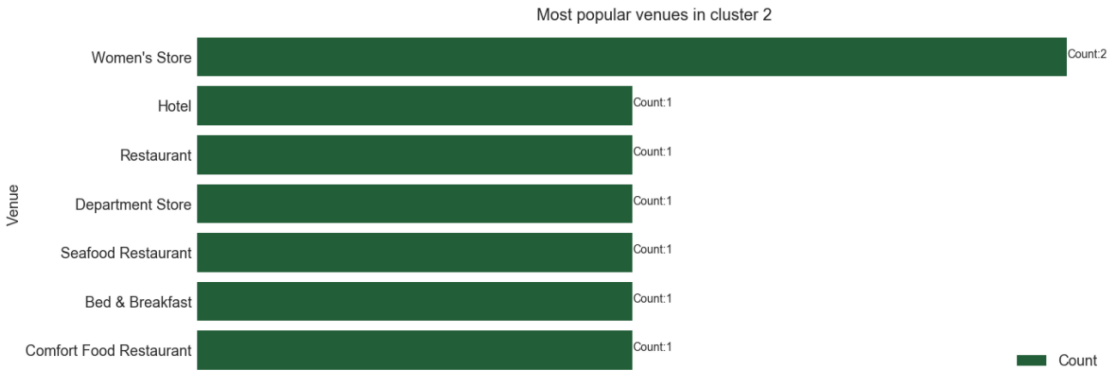

Figure 12. Graphical report: Cluster Type 2

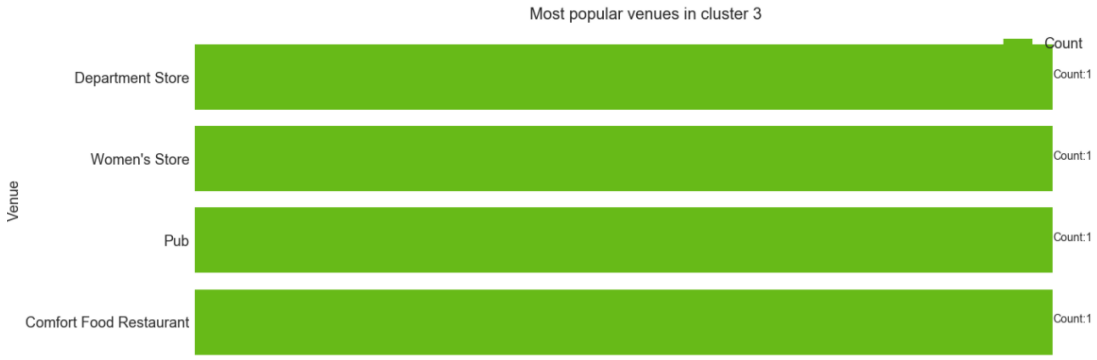

Figure 13. Graphical report: Cluster Type 3

5. Conclusion

This market analysis shows that restaurants are quite popular in all of the cities, however not showing any trend, in any cluster. After the clustering process, we got 4 "City Types", where "Type 0" belonged to big cities which offer lots of different venue categories but without one of relevance; "Type 1" belongs to smaller cities basically focused on tourism (heavy presence of hotels & restaurants alone); "Type 2" belongs to small cities which have a quite decent offer of several venue categories, and "Type 3" that are the remaining group, which is small cities with a reduced venue offer. We also saw that cities belonging to "Type 0" are the ones more popular in Ecuador, probably linked to the high population present in those cities.

If an investor plans to start a new business and wants to obtain revenue as quick as possible, the most logic option would be moving to those cities different than "Type 0", because competition will be smaller than the ones in "Type 0". Plus, to provide hints as to what to choose, we could tweak the created function "get_ecu_top_venues_report", setting its argument "popular_venues" to "False".

This market analysis wanted to provide general hints about the most popular venues already present in the Ecuadorian Capital cities, and therefore obtain _interesting_ venue categories what to invest on; it also provided some ideas about the "Market Types" each of those cities has, so the

new investor is generally aware about what kind of market his home town is, what would he / she need pick as venue category, to start a new business, and even enables to grasp knowledge about other cities's market, if he feel daring and wants to start over in a different side.

6. Future directions

Some facts that impacted however to this study, and are prone of improvement in future studies. are:

1. Ecuadorian People is not used to take time to provide deep feedback about the places they visit, and so the database of existing venues was relatively small, even when this might not be the actual reality in those markets. However, the same principles applied on this study, could be extrapolated to bigger cities (or ones with more data).

2. Specifically speaking about the general categories "Restauran" & "Hotel", we saw that their presence is larger than other kinds of venues; if something more "specific" would need to be grasped (i.e., what type of restauran is more popular in some city), then a further study should be done, focusing the efforts in that specific city, and that specific venue category.

3. The previous suggestion is limited by the first one, so this study would limited to be done in those cities that have a large population, and a high venue offer (i.e., according to this study, only those cities of "Type 0" and "Type 2").

4. There are several clustering algorithms, that work well however with large numbers of features. An study embracing these algorithms, combined with cities that generate a larger number of venues, could obtain more interesting insights.