IBM Applied Data Science Capstone Project

Venues Data Analysis of the cities with the highest population in Ecuador.

Daniela Sánchez

1. Introduction

1.1. Background

Tourism is one of the most important economic and cultural activities for any country. This industry promotes the economic reactivation of the site and generates sources of employment. Tourism has a favorable impact on sectors such as construction, hotels, gastronomy, and transport. Tourism activities creates demand across diverse industries.

Ecuador is a small country with an estimated population of 17 million, but it is also one of seventeen megadiverse countries in the world according to Conservation International. The country is divided into four regions: the Andes, the Amazon, the Pacific Coast, and the Galapagos. The diversity of its four regions has resulted in hundreds of thousands of species of flora and fauna. From beaches to volcanoes, you can expect to find almost any kind of environment here [1]. The three most populated cities (Quito, Guayaquil and Cuenca) are attractive destinations for vacation or retirement.

1.2. Problem

The objective of this project is to analyze the most common venues of the 20 most populous cities in Ecuador. Due to the diversity of places to visit, determine if there is a relation between the cities and the categories of the venues. This analysis could contribute to identify a business opportunity.

1.3. Interest

People who are interested to open a business that has relation with the tourism industry. Others who are interested in this analysis could be travelers or tourists who want to know the most common venues that they can enjoy in each analyzed city.

2. Data

To achieve the mentioned problem the following data was considered:

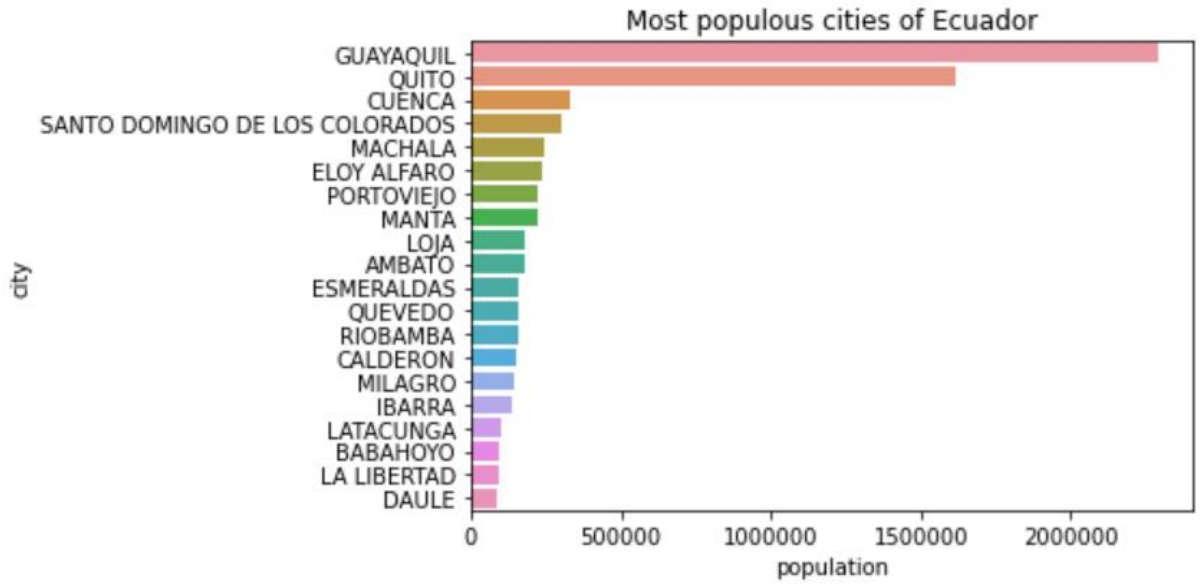
- From the National Institute of Statistics and Censuses of Ecuador (INEC), I obtained the data of the Ecuadorian population distributed by provinces, cantons (local government area) and cities. This dataset was obtained from the last census carried out in Ecuador [2]. I cleaned the data and reduced the data to the most populous cities.
- The latitude and longitude coordinates of the cities was obtained with the python's package geocoder.
- I used **Foursquare API** to get the most common venues of the selected cities.

3. Methodology

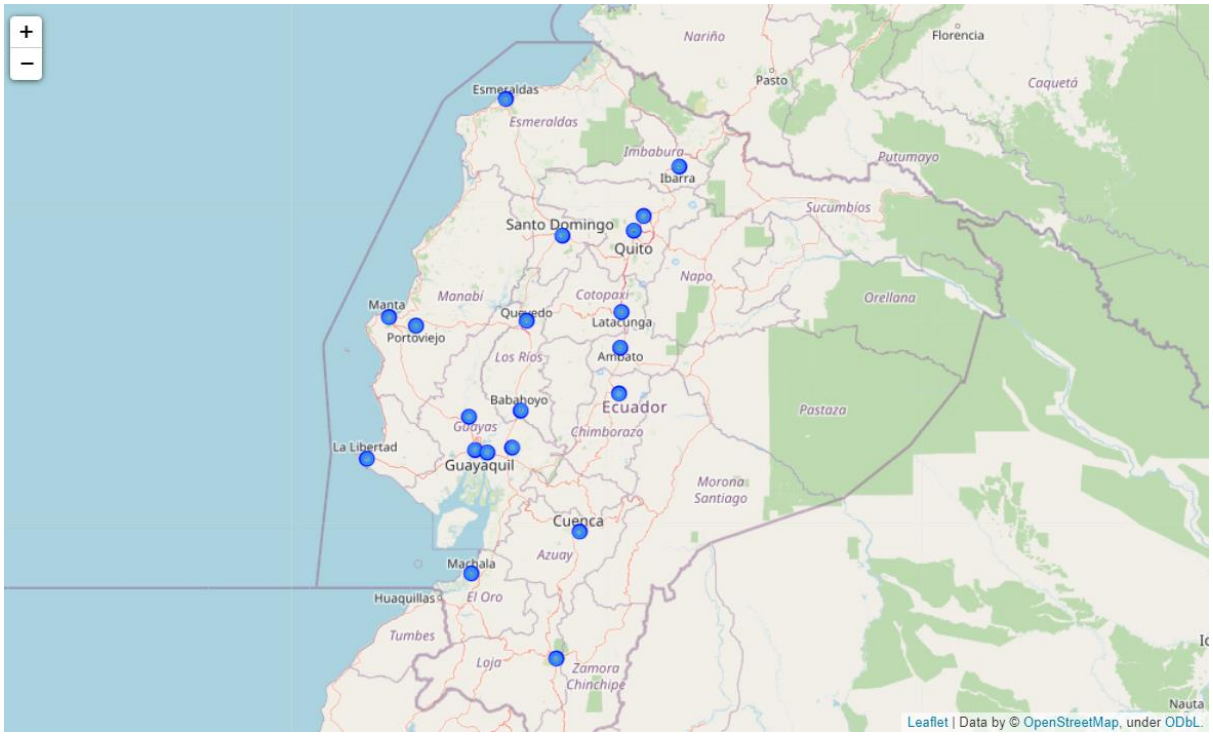
First, I loaded the collected data from the National Institute of Statistics and Censuses of Ecuador. After removing the missing values and sort the data by population, the twenty cities with the highest population were selected for the analysis. To get the latitude and longitude of the selected cities, the python's package geocoder was used. The following dataset was obtained:

	Latitude	Longitude	Cities	Population
0	-2.15960	-79.92830	GUAYAQUIL	2291158.0
1	-0.20565	-78.50888	QUITO	1619146.0
2	-2.89310	-78.99410	CUENCA	331888.0
3	-0.25000	-79.15000	SANTO DOMINGO DE LOS COLORADOS	305632.0
4	-3.26267	-79.96053	MACHALA	241606.0

There is a highly correlation or impact between the population of a city and the tourism. The population distribution of the selected cities is the following:



As we can evidence in the plot, the most populous cities of Ecuador are Guayaquil, Quito, and Cuenca. And the twenty selected cities are in the map as shown below.



For each city, the venues were retrieved using the Foursquare API. The search was performed for venues within a radius of 5000 meters. Due to the analysis was related with tourism, the categories for venues chosen were Arts & Entertainment, Nightlife Spot, Food, Travel & Transport. These categories are available in the Foursquare website [3]. There was a limitation with the venues

retrieved by Foursquare API, the cities chosen for further analysis were the ones that at least ten venues were obtained. Next, the results were grouped by city and calculated the mean of the frequency of occurrence of each venue category. This frequencies dataset was the input for the clustering section.

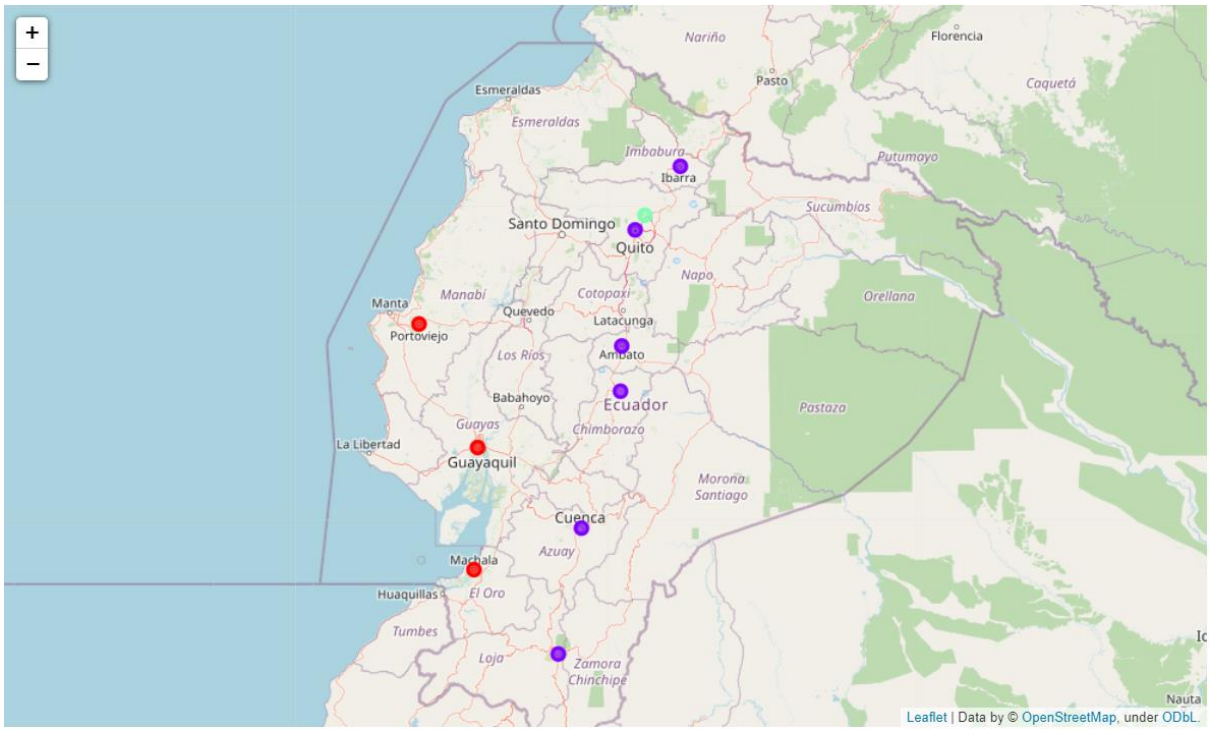
The five most common venues for the selected cities are:

	Cities	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	AMBATO	Hotel	Mexican Restaurant	Pizza Place	Restaurant	Seafood Restaurant
1	CALDERON	Seafood Restaurant	Pizza Place	Zoo	Italian Restaurant	Coffee Shop
2	CUENCA	Restaurant	Italian Restaurant	BBQ Joint	Hotel	Latin American Restaurant
3	GUAYAQUIL	Seafood Restaurant	Coffee Shop	Latin American Restaurant	Café	Pizza Place
4	IBARRA	Hotel	Pizza Place	Bar	Seafood Restaurant	Coffee Shop
5	LOJA	Seafood Restaurant	Fast Food Restaurant	Pizza Place	Café	Restaurant
6	MACHALA	Seafood Restaurant	Fast Food Restaurant	Hotel	American Restaurant	Restaurant
7	PORTOVIEJO	Seafood Restaurant	Mexican Restaurant	Latin American Restaurant	Breakfast Spot	Hotel
8	QUITO	Coffee Shop	Restaurant	Hotel	Café	Italian Restaurant
9	RIOBAMBA	Hotel	Fast Food Restaurant	American Restaurant	Restaurant	Mexican Restaurant

Finally, the frequencies dataset calculated was cluster to determine if there is a relation or pattern between the most common venues and the cities of the two Ecuadorian regions. K-Means algorithm was used because is one of the most common cluster method of unsupervised learning. The silhouette coefficient method was used to determine the optimal number of clusters. The result was 3 clusters.

4. Results

The following map shows the result of the clustering. Cluster 0 is represented with red, Cluster 1 with purple color and Cluster 2 with mint green color.



The main venue category found across clusters is related with gastronomy. The cluster 0 contained only cities from the Coastal region and due to the geography of the zone the most common venues are seafood restaurants.

	Cities	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	ClusterLabels	Latitude	Longitude	Population
3	GUAYAQUIL	Seafood Restaurant	Coffee Shop	Latin American Restaurant	Café	Pizza Place	0	-2.15960	-79.92830	2291158.0
6	MACHALA	Seafood Restaurant	Fast Food Restaurant	Hotel	American Restaurant	Restaurant	0	-3.26267	-79.96053	241606.0
7	PORTOVIEJO	Seafood Restaurant	Mexican Restaurant	Latin American Restaurant	Breakfast Spot	Hotel	0	-1.05563	-80.45318	223086.0

Whereas in cluster 1, we can find three main categories in the most common venues like Hotel, Seafood Restaurants, and Coffee Shop. Also, all the cities presented in this cluster belong to the Andes region.

	Cities	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	ClusterLabels	Latitude	Longitude	Population
0	AMBATO	Hotel	Mexican Restaurant	Pizza Place	Restaurant	Seafood Restaurant	1	-1.24917	-78.62997	178538.0
2	CUENCA	Restaurant	Italian Restaurant	BBQ Joint	Hotel	Latin American Restaurant	1	-2.89310	-78.99410	331888.0
4	IBARRA	Hotel	Pizza Place	Bar	Seafood Restaurant	Coffee Shop	1	0.36509	-78.10613	139721.0
5	LOJA	Seafood Restaurant	Fast Food Restaurant	Pizza Place	Café	Restaurant	1	-4.02156	-79.20295	180617.0
8	QUITO	Coffee Shop	Restaurant	Hotel	Café	Italian Restaurant	1	-0.20565	-78.50888	1619146.0
9	RIOBAMBA	Hotel	Fast Food Restaurant	American Restaurant	Restaurant	Mexican Restaurant	1	-1.66378	-78.64094	156723.0

Finally, the Cluster 2 is formed only by one observation. Calderon is a rural zone belong to the Andes region, located near the city of Quito. In the last years, the population of this zone has increased, and we can observe a slightly different pattern for venue categories compared with other clusters.

	Cities	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	ClusterLabels	Latitude	Longitude	Population
1	CALDERON	Seafood Restaurant	Pizza Place	Zoo	Italian Restaurant	Coffee Shop	2	-0.08083	-78.42219	152242.0

5. Discussion

Each region of Ecuador has their own attractive to offer to the tourists. Its "four worlds", the Galapagos, the coast, the Andes and Amazonia, are incredibly diverse and possess a unique multiculturalism [4]. Gastronomy is an important aspect of the tourism industry and it was evident as the top business category in Ecuador. In this analysis, we evidence a lack of venues related to transportation services, this could be a potential business opportunity to connect hotels, the popular restaurants with the amazing landscapes that Ecuador offers. The develop of a personalized and affordable transportation service could generate a positive impact in tourism.

In this project, we only consider ten cities with the highest population for the analysis and only certain venues categories. It is important to remark that there was a limitation in the information retrieved by Fourquare API. This limitation could generate some inconsistencies in the analysis. Also, there are other important factors to consider that can impact the tourism industry and the selection of a business opportunity.

6. Conclusion

Purpose of this project was to identify the most common venues of the cities with the highest population in Ecuador. This analysis is oriented to the tourism industry, so we only review the following main categories of venues: Arts & Entertainment, Nightlife Spot, Food, Travel & Transport.

Clustering of venues' frequencies was then performed in order to identify if there is a relation between the common venues and the regions of Ecuador (the cities analyzed correspond to two regions: the Andes and the Pacific Coast). Three clusters were obtained, and we could confirm the relation across the most common venues and the cities of each region. The cluster 2 only contains one rural zone called Calderon near the city of Quito. In the last four years, Calderon went from a rural zone to a strong urban development pole.

Across all the clusters the gastronomy or food is the most common category venue and we could evidence a lack of venues related with transportation services like bike or car rental venues.

References

1. Ecuador – Wikipedia
2. INEC
3. Foursquare categories
4. Tourism Ecuador