Finding alternative cities to emigrate from Venezuela

IBM Applied Data Science Capstone Project

Heberto Alarcon

June 2020

Introduction

Background

A socioeconomic and political crisis that began in Venezuela during the presidency of Hugo Chávez has continued into the presidency of Nicolás Maduro (still 2020). It is marked by hyperinflation, escalating starvation,[1] disease, crime and mortality rates, resulting in massive emigration from the country[2]. Over the past 2 decades, many Venezuelans (Around 4-6 millions[3]) have flee the country searching a better lifestyle than the precarious one Venezuelan regime offers. Such rapid growth of emigrants, however, has caused a general sense of crowding in certain countries (like Colombia, Peru, etc.[4]).

Problem

The steep rise in the cost of living, insecurity and low wages are pushing Venezuelans to seek alternative places to live in. The question for this subset of people that leave the country is ***how to even get started browsing prospective places to move***, as Spanish speaking countries alone are 22 cities (Mexico, Colombia, Spain, Argentina, Peru, **Venezuela**, Chile, Guatemala, Ecuador, Cuba, Bolivia, Replica Dominicana, Honduras, El Salvador, Paraguay, Nicaragua, Costa Rica, Puerto Rico, Panama, Uruguay, Guinea Equatorial and Belize).

To answer this question, we'll start with the assumption that potential Venezuelan emigrants looking to move are still interested in in living in Spanish speaking country and seek to find alternative cities with similar amenities as their current one (In this case, Maracaibo city). Given this scope, we can sample the superset of biggest cities within the Spanish speaking countries to create a kind of "fingerprint" of popular venues (such as certain types of restaurants, stores and natural areas) for each city, and then use this to identify potential similarities with other cities. The findings of this exercise could then be used as a recommendation guide for further, in-person demographics research.

Audience

The primary audience of this study might include Venezuelan immigrants as well as other Latin American emigrants planning on leaving their country. The findings could also be used by Latin American entrepreneurs looking to open new businesses or even a way of fostering outreach and partnerships among Spanish speaking municipal chambers of commerce.

Data

Sources

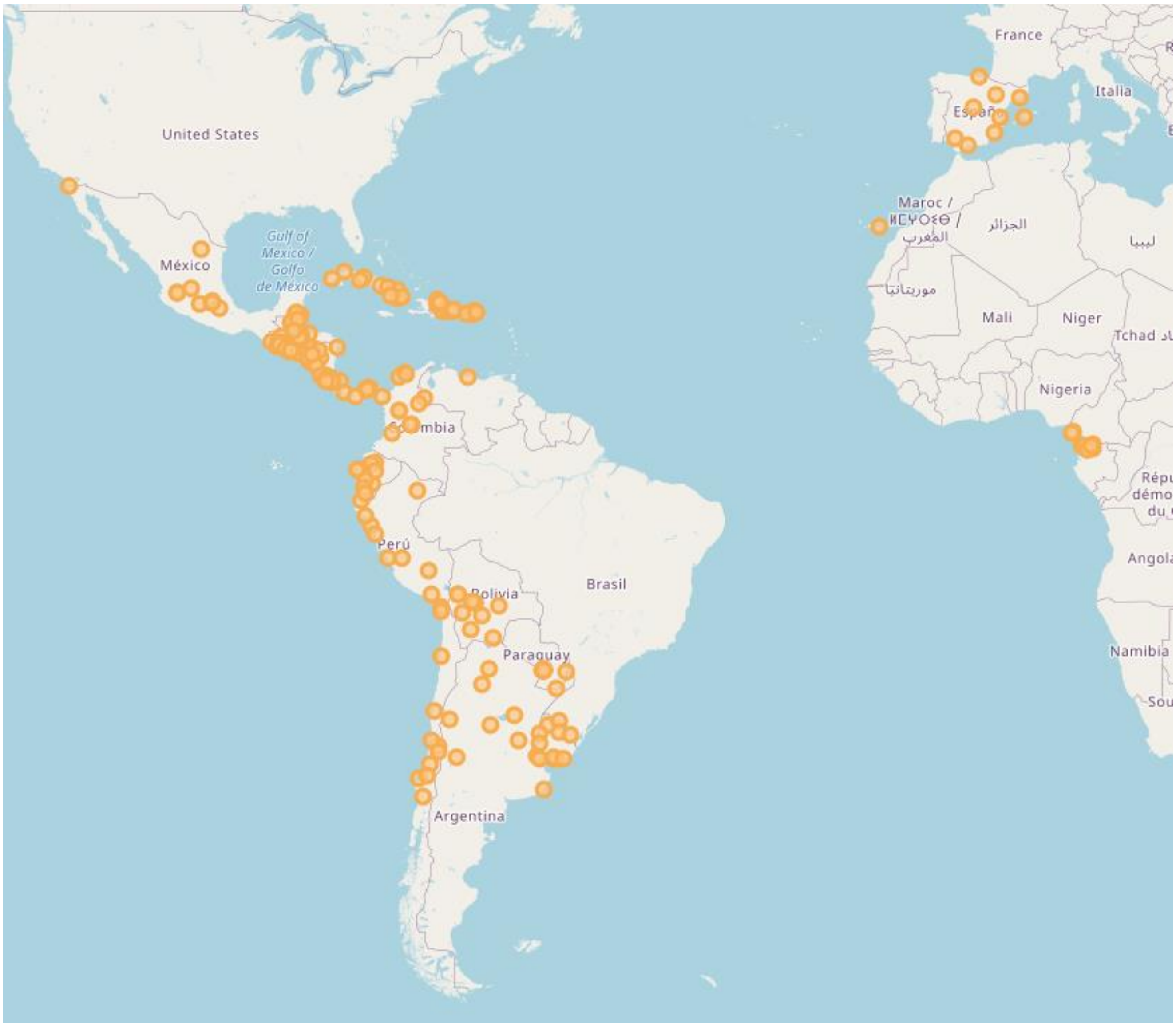
To obtain a list of the biggest cities with all Spanish speaking countries, we'll scrape Wikipedia for the list of Spanish speaking countries.[5]. We'll use the Foursquare venue recommendation API [5] to obtain a list of the most popular venues for each cities and query location data (latitude/ longitude) using the Nominatim in order to map all the cities and visualize the clusters.

Preparation

First, obtain a list of cities by scraping the Wikipedia pages on those topics. The lists of cities on those pages are structured in tables, so we can easily use Pandas to read in the HTML table and convert it to a dataframe. We'll set up a new dataframe to store the location data for each Latin American city.

- Mexico
- Colombia
- Spain
- Argentina
- Peru
- **Venezuela**
- Chile
- Guatemala
- Ecuador
- Cuba
- Bolivia
- Replica Dominicana
- Honduras
- El Salvador
- Paraguay
- Nicaragua
- Costa Rica
- Puerto Rico
- Panama
- Uruguay
- Guinea Equatorial
- Belize

I then ran the master city dataframe through the Geocoding to look up the location (in terms of longitude and latitude) of each city. This is called geocoding. Using the Folium mapping library, I was then able to render the full set of cities.



Next, I counted the number of venues for each city. Some cities (for example, very poor cities in Africa) have very few venue entries on Foursquare. After testing different limits, I found that a city requires at least about 10 venue entries in order to have an adequate venue "profile" for meaningful clustering results with other cities. Given that, I dropped cities with fewer than 50 venues for the remainder of the study. This was a necessary step for further analysis; however, it drastically truncated the list of 158 Spanish speaking cities down to only 62 cities. Even so, the 62 remaining cities represented 326 unique venue categories, which seemed suitable for expressing interesting clustering patterns, as I later discovered.

Methodology

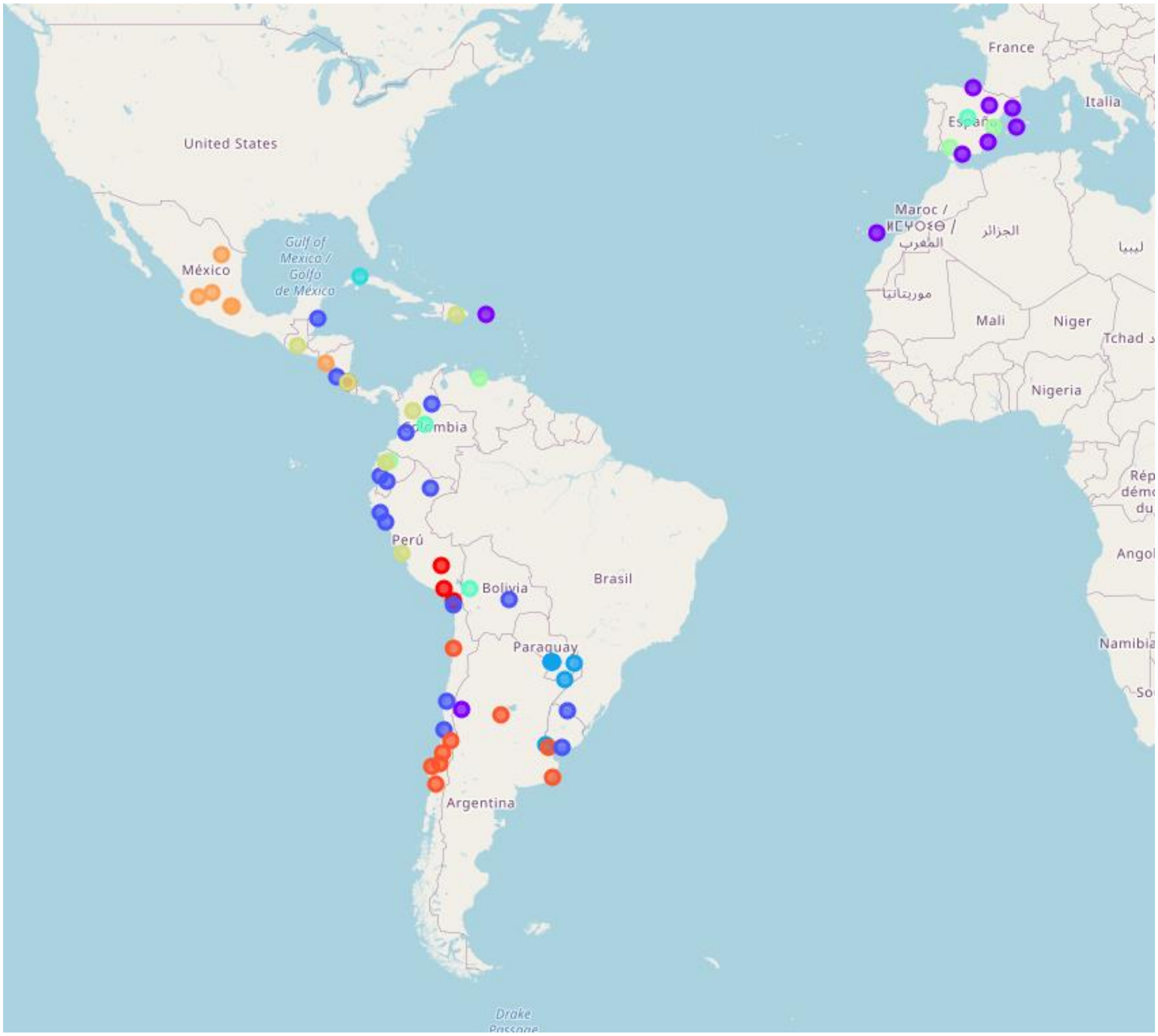
To analyze the data, I used a popular unsupervised machine learning algorithm called k-means clustering to partition observations into a specified number of clusters in order to discover underlying patterns. Specifically, I used the top 5 venue categories for each city (based on occurrences in the dataset) as each city's vector profile for finding similarities with other cities.

The first step was to calculate the average frequency for each venue category across each city. Using a Pandas dataframe I converted each venue category into a boolean (yes/no) column using the Onehot encoding method, verifying that new dataframe's column count equaled the number of unique

venue categories (326) I identified during data preparation. Next, I grouped rows by city mean of frequency for each category, and used that to find the five most common venues for each city. With that I were ready to apply the K-means clustering algorithm. After trying out different k values (where k= number of clusters), I found the clusters to be most meaningful and interesting with around k=10. The output of the K-means algorithm is an array of cluster assignments for each row in the dataframe. With that I then stitched the cluster labels back into the dataframe and combined city location data in order to print out and visualize the results.

Results and discussions

I used the Python Folium library to render the clusters, using a distinct color for each. At first glance, the results look promising in terms of holding some patterns about the dataset. The clusters seem generally dispersed geographically and balanced in terms of member count.



Conclusion

Starting from a list of 210 total cities across all Spanish speaking countries, I found 158 cities with Foursquare venue data. A Foursquare query of venues in those cities yielded 5935 venues, however, it was necessary to filter out cities with fewer than 50 venues, as their data profile later proved insufficient for meaningful clustering. After filtering out those cities, only 62 cities remained—less than 30% of the original group of cities.

The 62 cities used in the final analysis represented 4272 venues and 326 unique venue types. I used the k-means clustering algorithm to group them into ten distinct clusters, however only four of those clusters were truly meaningful in terms of revealing insights among the dataset that I could use to answer the original question of the immigrant's problem: how can Venezuelans identify similar cities as prospective places to move? The results of the analysis certainly provide one answer to the question:

City	Country	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	
32	Valencia	Spain	Plaza	Spanish Restaurant	Italian Restaurant	Restaurant	Tapas Restaurant
33	Seville	Spain	Tapas Restaurant	Plaza	Hotel	Ice Cream Shop	Spanish Restaurant
70	Quito	Ecuador	Restaurant	Hotel	Plaza	History Museum	Church
210	Caracas	Venezuela	Plaza	Bakery	Pharmacy	Theater	Historic Site

- Seville
- Quito
- Valencia

The next step was to run the list of English-speaking countries' cities through the Foursquare API to query the top venues in each city (according to the ratings of Foursquare users). I abstracted the individual venues by filtering the returned data based on the general category of each venue.

Footnotes

[1] https://www.theguardian.com/world/2020/feb/24/venezuela-hungry-food-insecure-un-world-food- program

[2] https://www.nytimes.com/2018/11/01/magazine/venezuela-inflation-economics.html?rref=collection

[3] %2Fsectioncollection% 2Fmagazine&action=click&contentCollection=magazine&region=rank&module=package&version=highligh t s&contentPlacement=3&pgtype=sectionfront

[4] https://en.wikipedia.org/wiki/Venezuelan_refugee_crisis#:~:text=The%20Venezuelan%20migration % 20and%20refugee,because%20of%20the%20Bolivarian%20Revolution.

[5] https://r4v.info/en/situations/platform

[6] https://en.wikipedia.org/wiki/List_of_cities_by_GDP

[7] https://developer.foursquare.com/docs/api/venues/explore

[8] https://developer.mapquest.com/documentation/geocoding-api/