Where should I go?

Analysis of the state cities of Colima

Alexis González Zambrano | Coursera Capstone Project | October 01, 2019

Report Content

1. Introduction

1.1 Background

"In the last decade, no country in the hemisphere has experienced an increase as large as Mexico in its homicide rate."

"The increase in violence is related to the so-called "war on drugs" and marks an upward trend."

These are some of the conclusions of the "Organized Crime and Justice in Mexico (2006 – 2010)" report of the Justice in Mexico program, developed by the University of San Diego, United States. For this reason, and many others, many businesses have found themselves in need of closing or moving out of state or city.

Now suppose that a person, who owns a dive center, is in a tourist city that has had a great increase in insecurity. This person has seen in the NISG (National Institute of Statistics and Geography) that one of the States with the lowest rate of insecurity in the country is the State of Colima and has decided to move there. But he doesn't know in which city of the state of Colima open his diving business so that his business can prosper.

1.2 Problem

Data that might contribute to determining in which city of the state of Colima it is convenient to open a business by finding, for example, in which cities in Colima have tourism activity related to the diving business?

For this we will use the information provided by Foursquare, since, from the cities of the State of Colima we can identify which are the most popular places and based on that information create clusters with the Clustering-k-means procedure and see which cities are similar, what type of trade is the one that develops the most, and can infer, in a better way, where it is more advisable to start a Diving Business.

This project seeks to recommend a city to people looking to open a business based on data that provide information related to the behavior of the local economy.

1.3 Interest

This type of problem turns out to be very common out there, since there are many people who seek to move from their respective cities for several reasons and need to find cities where their businesses can thrive.

More specifically, the solution I propose is aimed to those people who are looking to move or grow their business and need to know if the city or state , that they have in mind, is ideal for their business to thrive.

2. Data acquisition and cleaning

2.1 Data sources

In order to perform the analysis with the information provided by the Forsquare platform, we need to find geographic information about the state of Colima.

So essentially we will be working with the following information sources:

1) Database with information on the State of Colima regarding the postal codes of the town. As well as latitude, longitude and cities associated with those postal codes.

2) Data obtained through the Foursquare platform referenced to the longitudes and latitudes of each postal code in the State of Colima.

The first database can be found on the internet, in my country's portal, (https://datos.gob.mx/busca/dataset/bienes-inmuebles-del-patrimonio-del-gobierno-del-estado-de-colima-en- el-ano-2018). It is public and free information.

2.2 Additional considerations

The state of Colima is one of the smallest in Mexico (#27 of 32) and one of the states with the lowest number of habitants (#30 of 32).

Its cities, although rich in culture, are very small. So an analysis of a single city would not make much sense since the information obtained can be very poor.

According to INEGI, 70% of its economic activity belongs to the tertiary sector (services and products). So an analysis that seeks a tourist behavior in the state economy makes sense.

2.3 Data cleaning

Our first database is composed, in Excel format, of 16 columns and 564 rows. But this database shows us the real estate assets of the state of Colima in the year 2018.

The interesting thing about this database is that it relates each postal code to a latitude and longitude and just like in the “Segmenting and Clustering Neighborhoods in Toronto” project we can obtain a similar table where for each city we relate a postal code, latitude and longitude.

But before arriving at that table I had to clean the database: First I deleted all the rows that had the same postal code, same city and same municipality, since given the origin of the information we had repeated values because obviously the state government can have more than one real estate in the same city and with an address associated with the same postal code; Second, I removed all the columns that did not contain relevant information for the analysis and only kept the columns that had the name of the city, municipality, postal code, latitude and longitude; Finally, as in the “Segmenting and Clustering Neighborhoods in Toronto” project, I put together the cities that shared the postal code but were different in the municipalities.

The resulting dataframe, after cleaning, consists of 21 rows and 5 columns.

3. Exploratory Data Analysis

3.1 Foursquare information

With the resulting dataframe I used a procedure similar to the one carried out in the “Segmenting and Clustering Neighborhoods in Toronto” project and obtained the 10 most popular types of places in each city within a 10km radius.

City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
BUENAVISTA	Mexican Restaurant	Convenienc e Store	Brewery	Airport	Seafood Restaurant	Garden	Hotel	Food & Drink Shop	Lounge	Park
CHANDIABLO	Beach	Resort	Mexican Restaurant	Breakfast Spot	Seafood Restaurant	Golf Course	Café	Burger Joint	Plaza	Convenienc e Store
CIUDAD DE ARMERÍA	Convenienc e Store	Beach	Taco Place	Seafood Restaurant	Hotel	Steakhouse	Pizza Place	Sandwich Place	Sculpture Garden	Mexican Restaurant
CIUDAD DE VILLA DE ÁLVAREZ	Taco Place	Mexican Restaurant	Seafood Restaurant	Park	Pizza Place	Ice Cream Shop	Argentinian Restaurant	Bar	Hotel	Restaurant
COFRADÍA DE SUCHITLÁN	Mexican Restaurant	Pizza Place	Taco Place	Coffee Shop	Restaurant	Plaza	Convenience Store	Bar	Bakery	Café

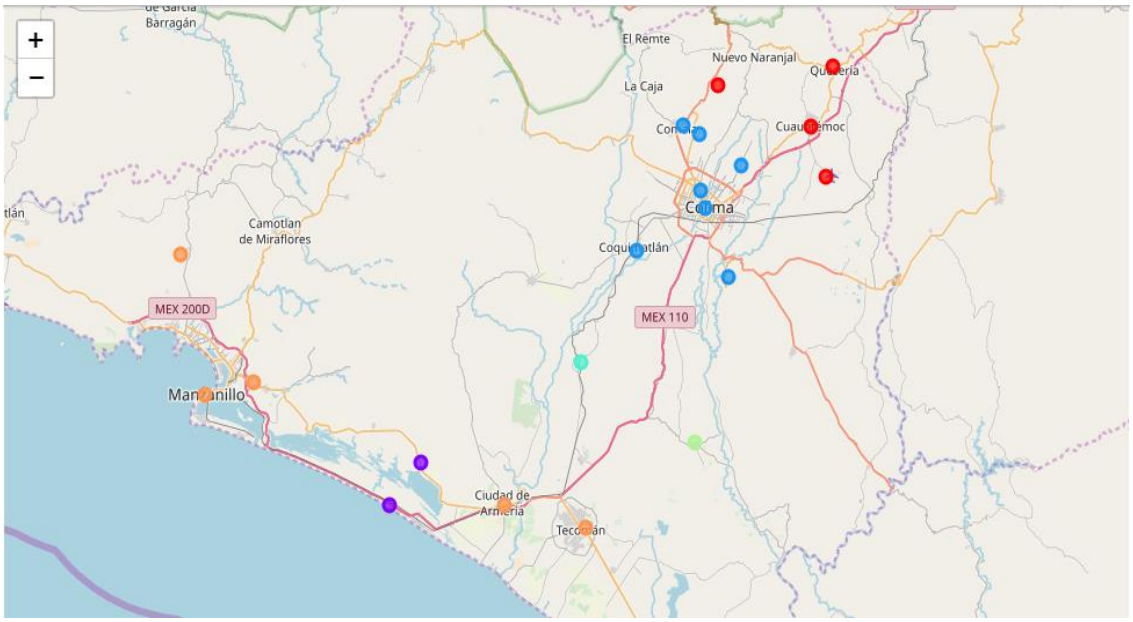
Using the machine learning method K-Cluster-Means, six Clusters were created for all the cities in the State of Colima.

It is important to mention that in order to choose the cluster number needed for a more precise analysis I had to test different sizes and decided that six cluster was ideal to represent the economy of each city in each cluster group.

3.2 Colima Map Analysis

Since each city has a particular cluster assigned, it is time to visualize what our methodology is really doing and why it does what it does.

To do that we create a map with different color markers for each cluster shown below:



With the help of the map it is obvious that of the six clusters, that we create, the most numerous is the cluster represented by the color blue, then the cluster represented by the color red and color orange, followed by purple and finally we have two individual clusters

An early analysis would lead us to think that the cities near the coast (beach) would have more things in common and we may come to think that they should belong to the same cluster.It is clear with the map that this does not happen and in fact it has a very reasonable justification.

3.3 Deep Cluster Analysis

It is time to analyze each cluster one by one.

Cluster 1: The first cluster represented by the color red has four cities, where all located in the upper part of the state. Among the most famous places stand out the Mexican restaurants, Convenience Stores,Plazas, Breweries and Parks.

Cluster 2: The second cluster represented by the color purple has two cities, both located at the bottom of the state. Among the most famous places stand out Surf Spots, Hotels, Beaches and Seafood Restaurants.

Cluster 3: The third cluster represented by the color blue has 7 cities, where all are located in the center and upper part of the state. Among the most famous places stand out Mexican food restaurant, Taco Places, Parks and Seafood Restaurants.

Cluster 4: The fourth cluster represented by the Light Blue color has only one city. Among the most famous places in the city stand out rivers, Mexican food restaurants, IT Services and Casinos.

Cluster 5: The fifth cluster represented by the color green has only one city. Among the most famous places in the city stand out caves, parks, Mexican food restaurants and Wings Joints.

Cluster 6 : The sixth and last cluster represented by the color orange has five cities. Among the most famous places stand out Beaches, Seafood Restaurants, Hotels and Mexican food restaurants.

Before continuing I would like to make a little clarification that can be obvious, the reason why Mexican food restaurants are found in literally all our clusters is because Colima is a Mexican state and obviously in Mexico there are Mexican food restaurants.

After this small individual analysis, we can separate the cities of Colima into two large groups. Those that are tourist cities and those that are not so much, since, if we analyze the most popular places we will realize that in clusters 1,3, 4 and 5 the most popular places are in general parks, convenience stores, caves, rivers and restaurants of Mexican food.

On the other hand in clusters two and six the most popular places are Hotels, Beaches, Seafood Restaurants and Surf Spots.

Trying to answer the initial question, we can now infer that the cities that we are going to recommend to open a diving business must belong to the second or sixth cluster; since, they are those that present a greater tourist activity in comparison with the other clusters. But to answer the question more accurately we must be able to choose in which cluster (2 or 6) it is more convenient to open a diving business and why?

In order to answer this question we must analyze both clusters a little more thoroughly.

From the position in which most of the cities are in the sixth cluster, it can be said that these cities are tourist but do not have tourism related to aquatic activities and the reason is essentially the type of cities they are. they are cities like manzanillo, with a lot of history and related to trade and not necessarily to tourism for pleasure, but rather business tourism.

Therefore, it is not uncommon that although they have coasts ,they are tourist centers oriented equally to cultural or business tourism and not only to “aquatic” tourism.

Finally, if we analyze in depth the cluster two, we can see that both cities are very close to the coast and for both cities their most popular places are all related to aquatic and beach tourism.

4. Results section

The results obtained from the analysis performed were the following:

The machine learning clustering-k-meaning method is a great tool for identifying groups that look alike. But it is only that, a tool, it does not work if the results obtained after the application of the method are not analyzed.

With the analysis of each cluster we realized that the location, the most visited places and the proximity of the cities does not imply that they belong to a particular group. Influence many other factors such as the history of the city and the development of the local economy to give some examples.

5. Discussion section

Returning to the objective of the report, we can now recommend with certainty, to our client, that the group of cities that he must have in mind, to open his dive business, must belong to cluster number two, since the development of the local economy matches the type of deal.

This recommendation has very big implications, since we must remember that, even if this person leaves his city, he will really be able to continue with his type of business, in which he has experience. So the change of city does not imply a decrease in the success in the business.

The analysis of data in this way is an excellent complement for business decisions; since assumptions are removed and just data is analyzed.

6. Conclusion

We can now understand how with information as “simple” as the most visited places in an area, we come to a recommendation with great implications and that answers the million dollar question. Where should I open my business and why?

We can conclude then that with creativity and data we can answer complicated questions with enormous implications. That is what a data scientist does.