IBM Applied Data Science Specialization

Capstone Project

Setting up a Business

Comparing and evaluating Guadalajara, Mexico’s neighborhoods for opening a Gymnastics Business.

Gustavo Santana Velazquez

28 July 2020

Introduction

Business Problem

Location has always been an important factor when starting a business. Despite the rise of technology, virtual communication and targeted marketing; a business's address is still a key factor on the success of the business, especially if it relies on offline sales or provides a service directly to customers.

A business location should consider the following points: if the business relies heavily on customer traffic, if it's convenient for the employees, if the brand visibility is important for growth, competitors and legal implications.

For this project, we'll suppose we have an investor looking for the right location in the Metropolitan Area of Guadalajara, Mexico to open a Gymnastic Academy for children. To do this, we will base on the USA Gymnastics "*How to start a Gymnastics Business*" guide to get the location requisites for this specific kind of business.

We will approach this problem using statistical data from the Metropolitan Area of Guadalajara provided by the government and using data science tools such as geospatial API's and statistical methods to find suitable locations that meet all the characteristics we're looking for.

This project may result interesting to anyone attempting to start a business in Guadalajara, Mexico or to someone just interested on how the different venues are distributed around one of the biggest and most important cities of Latin America.

Description and Background

Guadalajara is a metropolis in Western Mexico and the capital of the state of Jalisco. The metropolitan area of Guadalajara is the second largest of the country and an international center of business, finance, arts and culture; as well as the economic

center of the Bajío region, one of the most productive and developed regions in Latin America[1].

Guadalajara is a global city and one of the country's most important cultural centers. It is home to numerous mainstays of the Mexican culture (tequila, mariachi, birria) and host a variety of international events. It also has some of the most important universities and research institutes of the region.

Guadalajara metropolitan area includes the core municipality of Guadalajara and the surrounding municipalities/boroughs of Zapopan, Tlaquepaque, Tonalá, Tlajomulco de Zúñiga, El Salto, Ixtlahuacán de los Membrillos and Juanacatlán. We will be focusing on Zapopan, Guadalajara and Tlaquepaque for this project, since these are the core of the city and have a higher population density, making it a suitable space for the analysis.

[1] https://en.wikipedia.org/wiki/Guadalajara
Oppong, T. (2018). The Importance of Location in Business. July, 26, 2020. AllTopStartups Website : https://alltopstartups.com/2018/03/15/the-importance-of-location-in-business/

Objectives

To choose the right location we'll base on the USA Gymnastics[2] guide for starting a gymnastics business aimed mostly towards children. We can summarize the recommendations as follows:

1. Choose a location where there are enough children and where the families have enough money to pay for gymnastics. High population density and above average household income levels are key.

   For this point a database provided by the Government of Jalisco will be used to segment and study only the wealthiest neighborhoods in the metropolitan area based on the housing sale prices.

2. Choose a convenient location (very close to residential areas). To extent this point look for an area already supporting other businesses and facilities that cater children and their families (schools, sports facilities, day care centers, music studios). This will give the venue a great exposure and it also provides certainty that these locations can support similar businesses.

   Foursquare API will look for this kind of venues in or nearby the selected neighborhoods.

3. Do not locate the gym close to an already existing gymnastics school unless there's a certainty that the business can do a much better job.

   For this point the user reviews on Foursquare will be the indicator to determine if the existing gyms are a risk.

4. If there's not an area full of child-centered businesses, try to locate an area that draws plenty of families for other reasons (grocery stores, home improvement centers, etc.) that will still give the gym a lot of exposure.

---

[2] Taylor, M. (2015). How to Start a Gymnastics Business. july 26, 2020. USA Gymnastics Website: https://usagym.org/PDFs/Home/Publications/HowToStart/HowToStartGymnasticsBusiness.pdf

5. Beware of man-made barriers that make any location less convenient. Also consider traffic and parking as this can result in loosing potential customers.

This point will be left to anyone interested on future research as it is more related towards choosing the right premises.

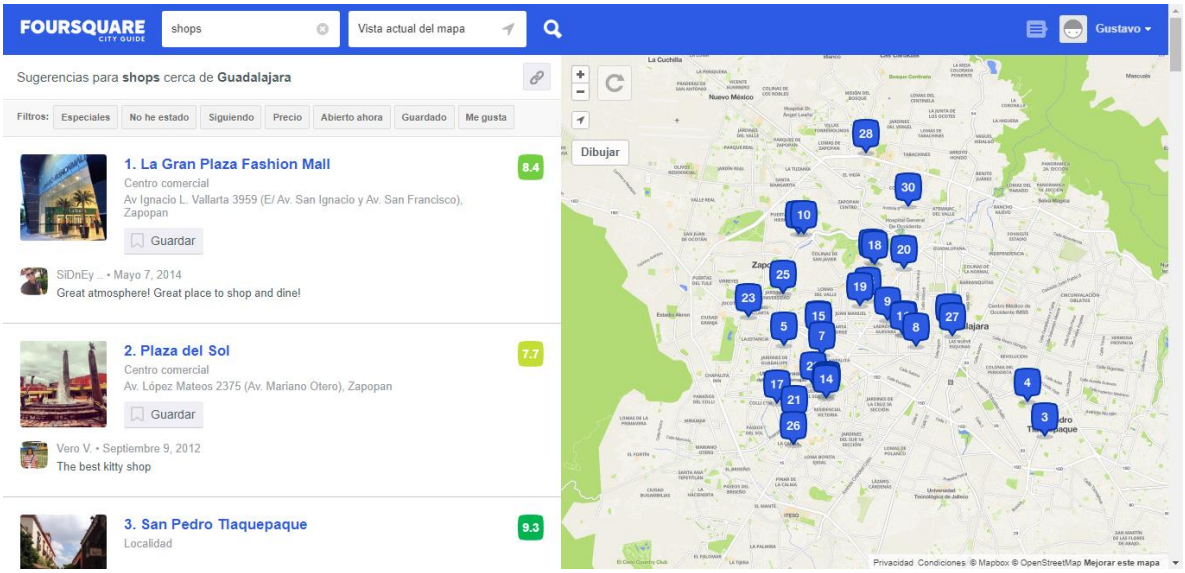
Data Sources

Foursquare API:

As this work is based on the nearby venues for each neighborhood, Foursquare API will allow us to retrieve the information for each of these venues, categorizing it and showing the reviews left by customers.

The key venues we'll be looking for are child-centered businesses and venues aimed to families or parents. It is also important to look for existing competition before to evaluate the risk.

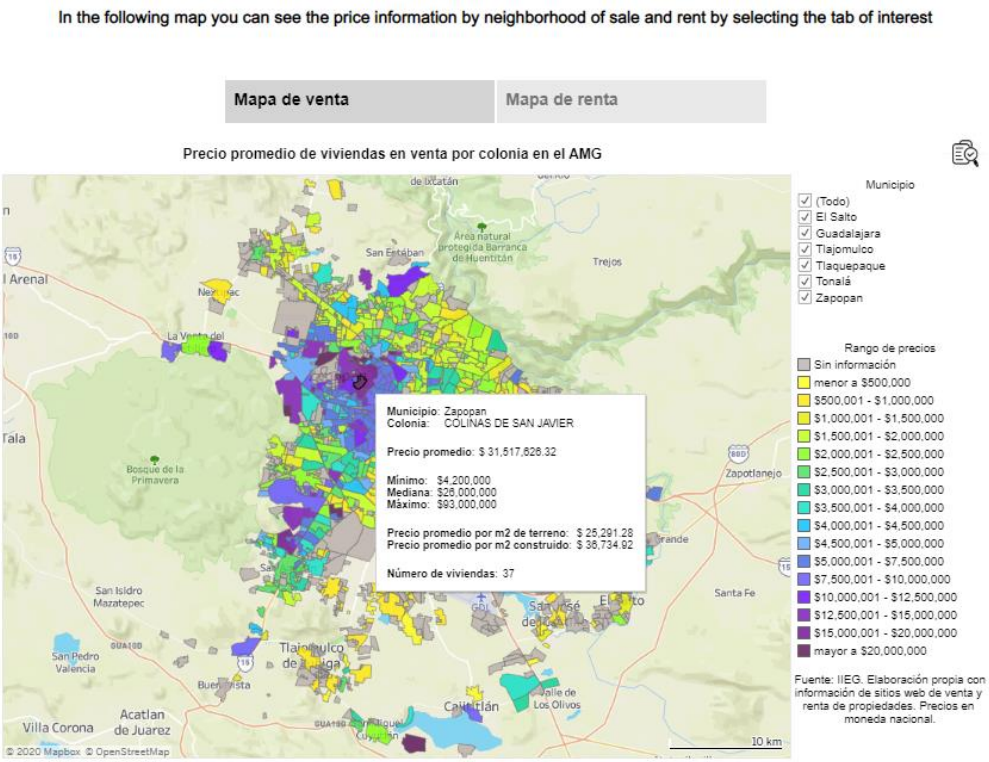
https://es.foursquare.com/explore?mode=url&ne=20.77056%2C-103.271656&q=shops&sw=20.593259%2C-103.500996



Geographical and Statistical Information Institute of Jalisco (IIEG)

IIEG[3] offers statistical data from Jalisco. For this project an Excel database containing a price list of 17,700 houses on the metropolitan area will be used to retrieve and segment the neighborhoods.

NOTE: This database was built on April 2020, so the exchange rate from the end of that month (24.207 MXN/USD) was considered.



Geopy Library

The search API (https://nominatim.openstreetmap.org/) returns the coordinates of a given address. It will be used to get the longitude and latitude for each neighborhood after cleaning the database.

[3] N.A.. (2020). Housing supply in the main municipalities of the Guadalajara Metropolitan Area. July 25, 2020, de Instituto de Información Estadística y Geográfica de Jalisco Sitio web: https://iieg.gob.mx/ns/?page_id=11967

Methodology

Data Acquisition and Cleaning

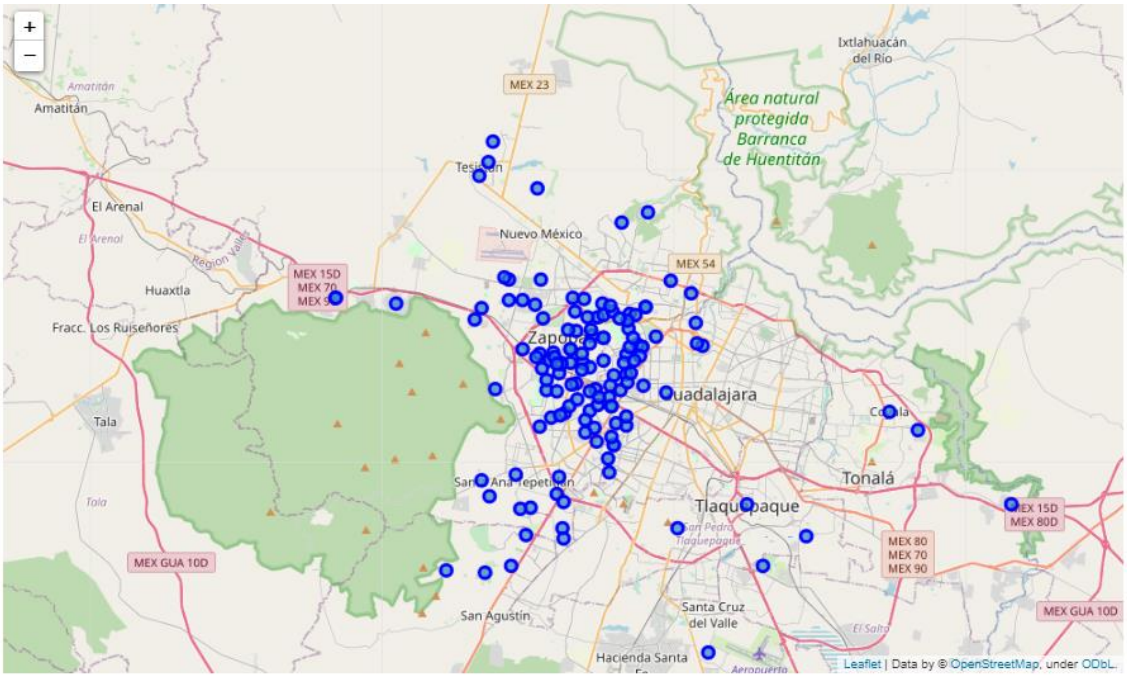
As mentioned, the first step is to download and clean the Excel the target neighborhoods. The segmentation was made based on their average selling price and only the 150 wealthiest neighborhoods were selected.

	Neighborhood	Borough	Avg Price
0	PALMIRA	ZAPOPAN	2272070.06
1	LOS FRAILES	ZAPOPAN	1982897.51
2	LOMAS DEL BOSQUE	ZAPOPAN	1397685.70
3	COLINAS DE SAN JAVIER	ZAPOPAN	1302004.64
4	ZOTOGRANDE	ZAPOPAN	1241280.10
5	SAN LUCAS EVANGELISTA	TLAJOMULCO	1198000.58
6	PONTEVEDRA	ZAPOPAN	1162198.26
7	SAN MIGUEL DE LA COLINA	ZAPOPAN	1106429.27
8	LAS LOMAS GOLF HABITAT	ZAPOPAN	1015533.47
9	ATLAS COLOMOS	ZAPOPAN	1011543.10

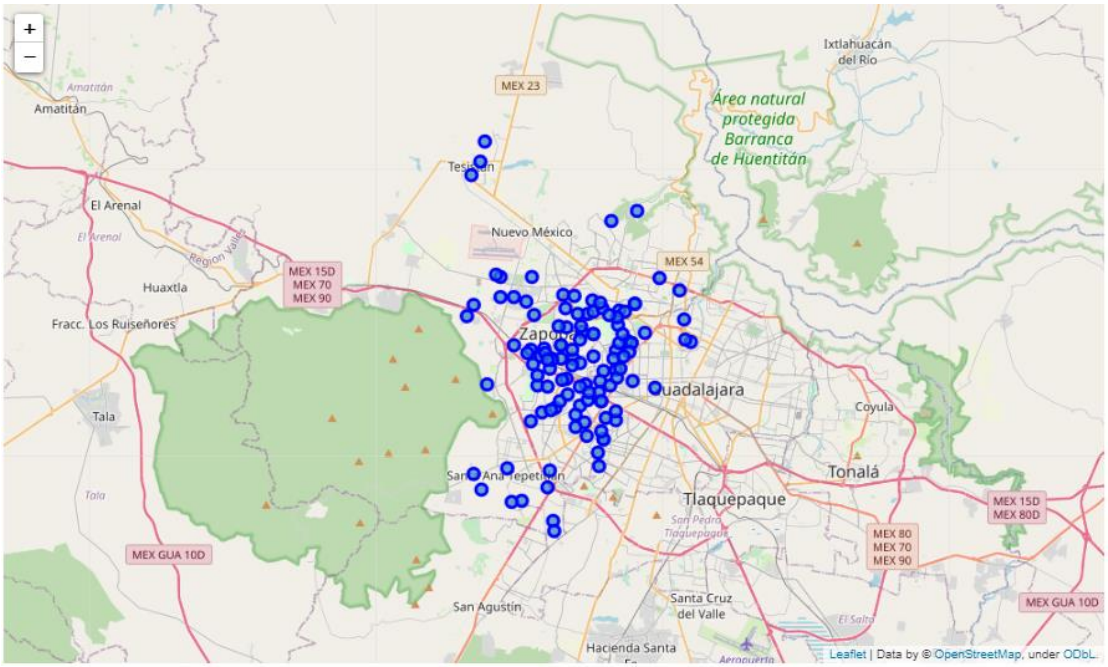
The latitude and longitude of each neighborhood is needed to use Foursquare and folium. Geopy was used in this case to retrieve the coordinates, after making some modifications to some of the rows to be able to find all of them. We get the following data frame after adding this information:

	Neighborhood	Borough	Avg Price	Latitude	Longitude
0	PALMIRA	ZAPOPAN	2272070.06	20.60	-103.43
1	LOS FRAILES	ZAPOPAN	1982897.51	20.72	-103.41
2	LOMAS DEL BOSQUE	ZAPOPAN	1397685.70	20.71	-103.41
3	COLINAS DE SAN JAVIER	ZAPOPAN	1302004.64	20.70	-103.40
4	ZOTOGRANDE	ZAPOPAN	1241280.10	20.72	-103.40
5	SAN LUCAS EVANGELISTA	TLAJOMULCO	1198000.58	20.41	-103.36
6	PONTEVEDRA	ZAPOPAN	1162198.26	20.71	-103.41
7	SAN MIGUEL DE LA COLINA	ZAPOPAN	1106429.27	20.72	-103.39
8	LAS LOMAS GOLF HABITAT	ZAPOPAN	1015533.47	20.71	-103.44
9	ATLAS COLOMOS	ZAPOPAN	1011543.10	20.72	-103.40
10	AYAMONTE	ZAPOPAN	981381.42	20.67	-103.47

Now the map of the metropolitan area of Guadalajara and its wealthiest neighborhoods can be displayed using folium.



After taking a look at the map, we can see that some of the locations are very far from the city, this means that they are either ranches or luxurious mansions on the outsides. These locations were dropped as they aren't adequate for a Gymnastics Business. As a result, we now have 120 neighborhoods to work with.



Evaluating Competition

One of the objectives is to determine if the existing Gymnastic Businesses represent a risk. To do this, the Foursquare API was used to retrieve all the venues related to 'Gymnastics' on every neighborhood. The following data frame was generated:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue ID	Venue	Venue Latitude	Venue Longitude	Venue Category
0	JARDINES DE LOS ARCOS	20.67	-103.40	4fb1b3aee4b0b5bcd2c46d3e	Gimnasia Artistica Ling	20.67	-103.40	Athletics & Sports
1	LOMAS PROVIDENCIA	20.69	-103.39	55db9d82498e77847671f6a8	Volaré Academia de Gimnasia	20.70	-103.39	Sports Club
2	PROVIDENCIA 4a SECCION	20.70	-103.39	55db9d82498e77847671f6a8	Volaré Academia de Gimnasia	20.70	-103.39	Sports Club
3	JARDINES DE SAN IGNACIO	20.67	-103.40	4fb1b3aee4b0b5bcd2c46d3e	Gimnasia Artistica Ling	20.67	-103.40	Athletics & Sports
4	RESIDENCIAL CHAPALITA	20.66	-103.43	4c05034473a8c9b6f9ca96e0	Club Atlas Chapalita	20.66	-103.43	Athletics & Sports
5	GUADALUPE JARDIN	20.66	-103.43	4c05034473a8c9b6f9ca96e0	Club Atlas Chapalita	20.66	-103.43	Athletics & Sports

As shown, there are three Gymnastics Schools close to six different neighborhoods. Nevertheless, only one of these (according to the available data on Foursquare) represents a risk:

```
Gimnasia Artistica Ling
        Number of Tips: 0
        Likes: 6
        Dislikes: False
        This venue has not been rated yet.

Volaré Academia de Gimnasia
        Number of Tips: 0
        Likes: 2
        Dislikes: False
        This venue has not been rated yet.

Club Atlas Chapalita
        Number of Tips: 23
        Likes: 288
        Dislikes: False
        Rating: 8.3
```

Therefore, the neighborhoods close to "Club Atlas Chapalita" were removed from the main data frame, leaving 118 neighborhoods.

Getting the nearby venues for each neighborhood

A function to get all the venues with their names, locations and category using Foursquare was used on the 118 neighborhoods. 3,392 venues in 268 unique categories were returned.

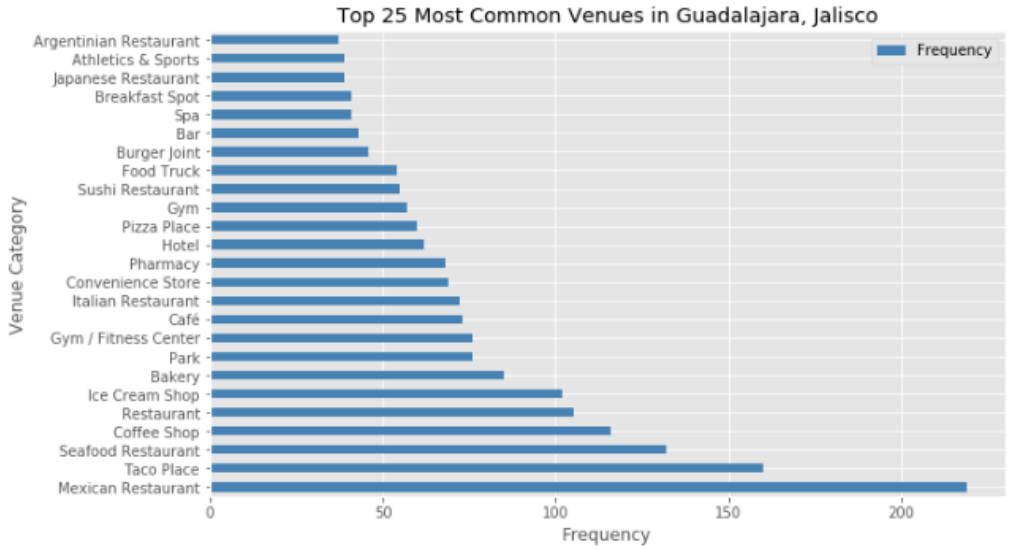
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	PALMIRA	20.60	-103.43	Sabra Dios Donde	20.59	-103.43	Farm
1	PALMIRA	20.60	-103.43	Agua Blanca Country Club Centro De Espectaculo...	20.60	-103.43	Speakeasy
2	PALMIRA	20.60	-103.43	Spa Can Camp	20.60	-103.43	Spa
3	LOS FRAILES	20.72	-103.41	Decathlon	20.72	-103.42	Sporting Goods Shop
4	LOS FRAILES	20.72	-103.41	Hiperlumen	20.72	-103.41	Paper / Office Supplies Store
5	LOS FRAILES	20.72	-103.41	The Home Depot	20.72	-103.41	Hardware Store
6	LOS FRAILES	20.72	-103.41	Los Arcos	20.72	-103.41	Seafood Restaurant

The number of venues found for each neighborhood ranged from 100 (which was the imposed limit) to 1.

Neighborhood						
DON BOSCO VALLARTA	100	100	100	100	100	100
PROVIDENCIA 4a SECCION	100	100	100	100	100	100
CHAPALITA ORIENTE	100	100	100	100	100	100
JARDINES DE SAN IGNACIO	100	100	100	100	100	100
LADRON DE GUEVARA	93	93	93	93	93	93
...	...	...	...	...	...	...
AYAMONTE	3	3	3	3	3	3
ARAUCA I	3	3	3	3	3	3
VALLE ESMERALDA	3	3	3	3	3	3
LOS PINOS	1	1	1	1	1	1
SAN WENCESLAO	1	1	1	1	1	1

115 rows × 6 columns

The most common venues are food related, nevertheless gyms and parks are also on the top 10. As mentioned before, the Gymnastic Business location depends on exposure and parent-oriented or family-oriented venues that offer the parents something to do while their son is taking the class, therefore the collected venues are adequate for this study.



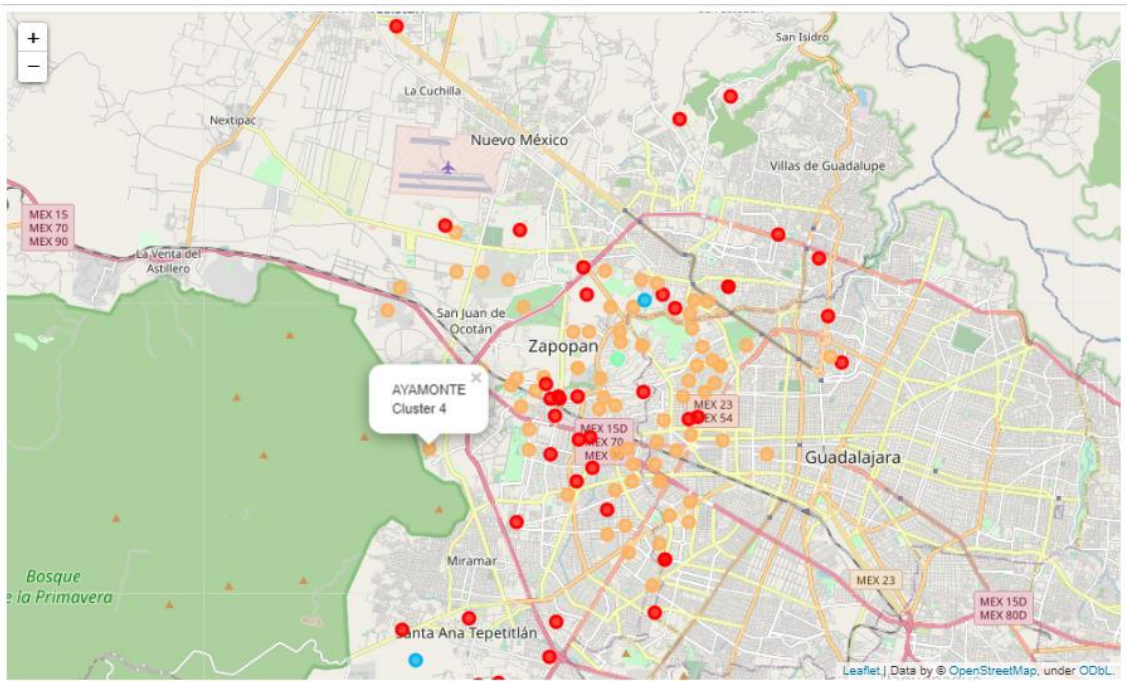
Finally, a data frame was created with the top 10 most common venues for each neighborhood:

	Neighborhood	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
0	AGRICOLA	Convenience Store	Restaurant	Lounge	Liquor Store	Construction & Landscaping	Salad Place	Taco Place	Butcher	Food Court	Mexican Restaurant
1	ALTAMIRA	Dog Run	Coffee Shop	Lounge	Burrito Place	Boutique	Food	Pizza Place	Gym / Fitness Center	Medical Center	Vegetarian / Vegan Restaurant
2	AMERICANA	Café	Coffee Shop	Mexican Restaurant	Tea Room	Breakfast Spot	Gastropub	Beer Bar	Bakery	Japanese Restaurant	Pizza Place
3	ARAUCA I	Sports Club	Food Truck	Pool	Yucatecan Restaurant	Flea Market	Farmers Market	Fast Food Restaurant	Film Studio	Fish & Chips Shop	Food
4	ATLAS COLOMOS	Bar	Athletics & Sports	Gym / Fitness Center	Shoe Store	Farm	Farmers Market	Fast Food Restaurant	Film Studio	Fish & Chips Shop	Yucatecan Restaurant

Out of the 118 neighborhoods, 3 had no venues, so they were dropped from the dataset, leaving 115 neighborhoods for clustering.

Analysis: K-means Clustering

The data is now cleaned up and ready for clustering. For the purpose of this worked 5 clusters will be set. The resulting clusters can be visualized in the following map:



Analyzing each cluster

- Cluster 0 (Red)

  40 neighborhoods were clustered in this group. Even though there are a lot of restaurants as the most common venue for these neighborhoods, there are also a great number of nearby venues that offer the families something to do while their child is taking the gymnastics classes (which is what we're looking for). Examples of these venues are: Parks, Athletics & Sports, museums, music venues, shopping malls, spas, café, etc.

- Cluster 1 (Purple)

Only one neighborhood is contained here (Los Pinos) and the most common venues are: Farm, Yucatecan Restaurant and Fabric Shop. Considering what we're looking for we can discard this cluster.

- Cluster 2 (Blue)

Two neighborhoods are included on this cluster (Atlas Colomos and Ciudad Bugambilias) with the main categories being Athletics & Sports and Restaurants. Although it holds some similarity with the categories from Cluster 0, it offers less activities.

- Cluster 3 (Green)

Holds one neighborhood (San Wenceslao) the most common venues are Pool, Yucatecan Restaurant and Flea Market.

- Cluster 4 (Orange)

Contains 74 neighborhoods, even though the most common venues for some of the neighborhoods offer activities similar to cluster 0, this cluster is mostly composed of food venues with some exceptions such as Puerta Aqua that has Athletic & Sports, Park, Gym Pool and Dance Studio as its most common venues. Without a doubt this cluster holds suitable neighborhoods for the purpose of this project.

Results

The above methodology allowed us to evaluate some of the objectives proposed by the USA Gymnastics Guide. First, we separated the 150 wealthiest neighborhoods in Guadalajara, which presumably have an above average income that would allow their children to attend out Gymnastics Gym.

Afterwards, we used the Foursquare API to research the existing business competition. We used the rating, tips and number of likes to determine if these gyms were a risk. Foursquare currently holds very little information for some of these venues, so further evaluation is recommended.

We then got the nearby venues for each neighborhood. Most of the results were food related venues. Then we proceeded to prepare the data for clustering and ran the K-means Clustering Algorithm with the objective to find a cluster that held the neighborhoods with the right qualities for opening the Gymnastics Business. The algorithm was run several times and the results were the same. Three out of the five clusters returned two or one neighborhood. The other two contained restaurants for the most part but also venues that offered different kind of activities, which is what we were looking for.

We could say that a neighborhood belonging to cluster 0 would make an adequate candidate for the purpose of this project, nevertheless, some of the neighborhoods of cluster 4 may also meet the requirements. As mentioned, when evaluating each cluster, these neighborhoods offer family-activities as well as places to shop, exercise or relax. This makes sense, since most of the neighborhoods included in Cluster 0 are located around shopping malls and commercial areas, thus the wide range of offered activities.

It is important to say that this project has some limitations relating Foursquare API that will be mentioned below.

Discussion

This project was made using various data science skills: working with API (Foursquare), map visualization (Folium), data manipulation (Pandas & Numpy) and statistical model (K-means clustering).

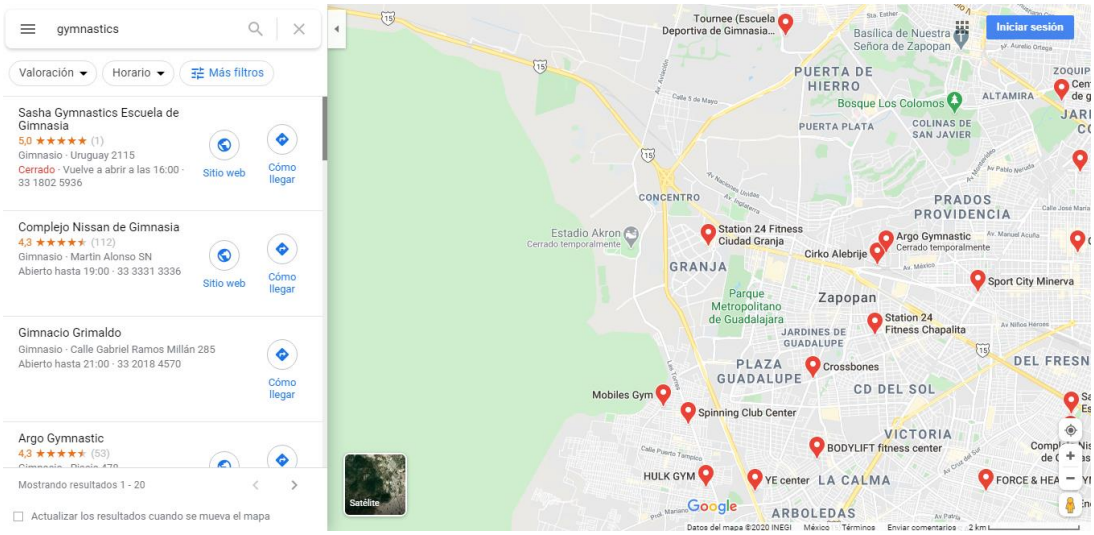
Limitations and Suggestions for Future Research

1. Foursquare's API insufficient information

   When using the K-means clustering, it was evident that most of the neighborhoods fell into two different clusters and the rest of them held only one or two neighborhoods. I consider this problem is because Foursquare is missing an important number of venues. Almost all the returned locations were restaurants, we could see that when reviewing the clusters.

   Child-oriented businesses, which is what we were looking for in this project, were scarce and lacked information. This represented a big problem when evaluating the competition as well, since only three venues that fell into the "Gymnastics" category were returned and only one of them had enough information. To this we can add that when looking for Gymnastics in Google Maps, the range of venues will be three times wider, with much more user reviews, making this a better option for this kind of queries.

   Another issue with this API is the search keywords when looking for a specific venue category. For projects with a similar objective as this one multiple searches with different synonyms will be necessary in order to get all the scouted venues.



2. A larger comparative study is recommended using different tools

   As Mexico is a developing country, tools such as Foursquare may not be that reliable when conducting studies like this one. A more complete search engine would provide more complete and updated results. When opening a business this is crucial since every little factor may have a huge impact on its success.

Conclusions

For this study, I analyzed the neighborhoods in the metropolitan area of Guadalajara with the objective of finding an adequate location to place a Gymnastics Business, based on the spending capacity and nearby venues. The Foursquare query and K-Means Clustering approach used in this project gave an idea of which neighborhoods meet these requirements.

As mentioned on the Discussion section, there are some limitations for this project that demand forward investigation and the use of other tools that include a more complete database of the venues in Mexico’s cities.