AI venue type investment model

Carlos Morlan

July 2019

Contenido

Executive Summary

I can tell you by experience that as a full time employee a day will come where you start thinking on becoming your own boss by opening a business. You have enough money, or at least you think that you have it. You are almost an expert on how to manage the business type you have in mind, or at least you believe so. So why not? Well, as soon as you start going into the details of your *original* plan, you realize that you may not be so expert to open the business you choose, but smart enough to learn about it. Or maybe your family and/or friends give you a better option to pick your *future* business category. Now you have more than one option on which business type will help you be the greatest businessman for whom you were born to be.

Now the second question to answer is on which part of my city should I invest my money to open my new business? Sometimes opening a business in a neighborhood where there are no similar

businesses nearby can be a good idea. On the other hand, it is best to open a business near other businesses whose customers can be mine too. For example, opening a pet food store not far away from a veterinary clinic can ensure the success of my business.

As you can see, before going on more details on our business plan, finding a list, book, website or tool where I can identify what types of business exist on every neighborhood in my city will make my life easier to start my new business, right? The intend of this document is to show how Data Science can be implemented to build this tool regardless the city where you are from in the whole world. In particular, I will focus on mine, *Mexico City*.

Introduction

I live in Mexico City, one of the biggest and most populated cities in the world. One of its citizens main concerns is that the country's economy is volatile, you can feel it in the air. A good proxy for the overall stability of a country is the consistency of its economic growth. From my personal point of view, getting more investments is a good way to improve its economic growth. The investments, that can be done by the government or by private companies, should be well planned based on the different communities needs through all the main country's cities.

This capstone project will try to show how Mexico City can attract new invests for Mexico's economy improvement. The Government and new investors should know what are the popular places where the citizens have fun, get dinner or bought supplies. With such information, either of both can make best decisions about the type of business they can open and how well the people will take the new venue. Moreover, if there are popular places with special attributes like a Medical Center, they can start opening required business types near such places like Laboratories, Pharmacies or even a Hotel so the people from outside town can stay there while their patients are receiving treatment. As you can see, small or big investors can use this valuable information to take the path of a successful opportunity and the city communities will also get more and better services: Is a win-win situation.

Data acquisition and cleaning

Data sources

Some ingredients are mandatory before we can make this happen. A couple of data sources have been identified for this project:

- Major city spots identified by zone or neighborhood.
- Popular venues identification based on social networks (www.foursquare.com).

Major City Spots

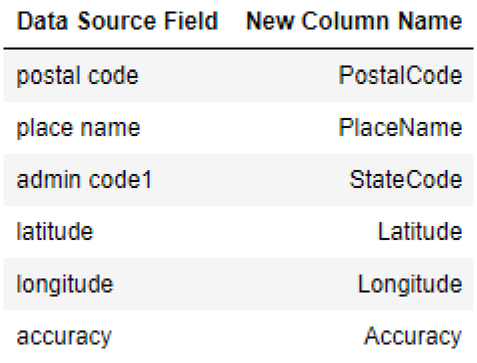
To get this information I will be using a public data set from www.geonames.org site downloaded locally in case the file is removed or changed in the future, check
the References section for more details. The data format is tab-delimited text in utf8 encoding, with the following fields:

- country code : iso country code, 2 characters
- postal code : varchar(20)
- place name : varchar(180)

- admin name1 : 1. order subdivision (state) varchar(100)
- admin code1 : 1. order subdivision (state) varchar(20)
- admin name2 : 2. order subdivision (county/province) varchar(100)
- admin code2 : 2. order subdivision (county/province) varchar(20)
- admin name3 : 3. order subdivision (community) varchar(100)
- admin code3 : 3. order subdivision (community) varchar(20)
- latitude : estimated latitude (wgs84)
- longitude : estimated longitude (wgs84)
- accuracy : accuracy of lat/lng from 1=estimated, 4=geonameid, 6=centroid of addresses or shape

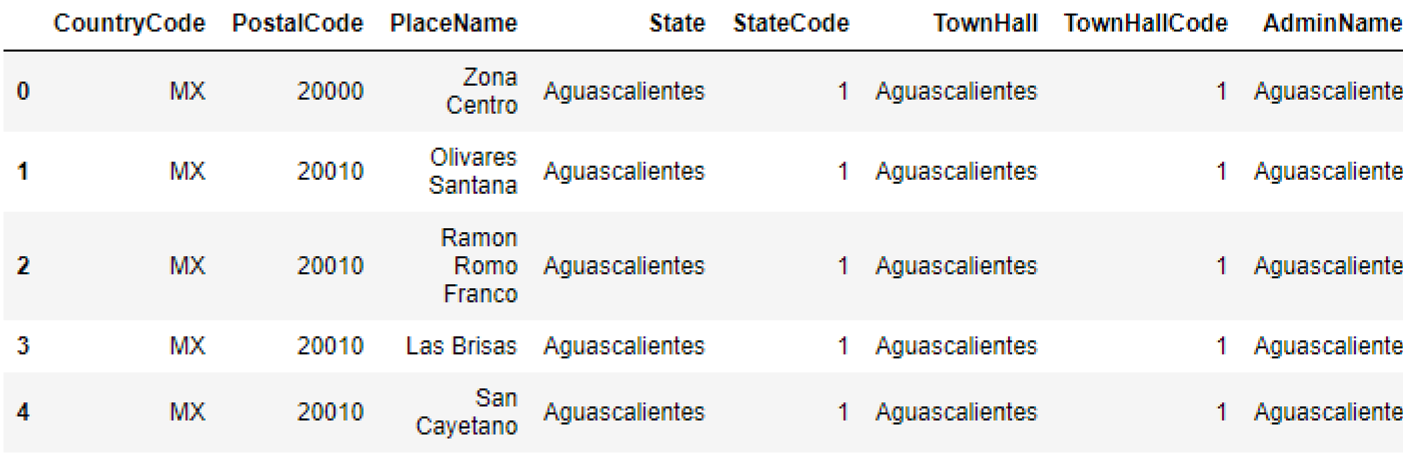
FYI, this file doesn't have column headers.

After looking into the data, I chose a better name for each column based on its content and identified which columns can be dropped because are not required for this analysis. The final mapping is shown below:



Data Source Field	New Column Name
postal code	PostalCode
place name	PlaceName
admin code1	StateCode
latitude	Latitude
longitude	Longitude
accuracy	Accuracy

Reading the file content using pandas package looks like this:



	CountryCode	PostalCode	PlaceName	State	StateCode	TownHall	TownHallCode	AdminName
0	MX	20000	Zona Centro	Aguascalientes	1	Aguascalientes	1	Aguascaliente
1	MX	20010	Olivares Santana	Aguascalientes	1	Aguascalientes	1	Aguascaliente
2	MX	20010	Ramon Romo Franco	Aguascalientes	1	Aguascalientes	1	Aguascaliente
3	MX	20010	Las Brisas	Aguascalientes	1	Aguascalientes	1	Aguascaliente
4	MX	20010	San Cayetano	Aguascalientes	1	Aguascalientes	1	Aguascaliente

Data cleaning

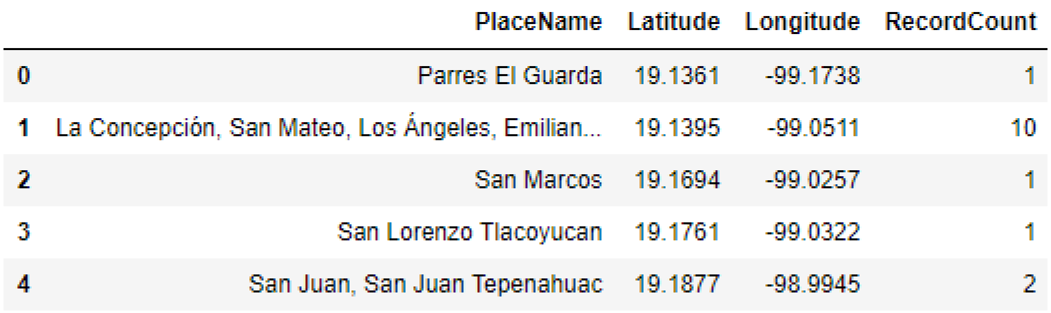
Regardless that the data for every Mexican neighborhood is available, I will apply a couple of filters to the data source because of 2 reasons:

- The analysis that will be done is for Mexico City only (i.e. admin code1 == 9).
- The number of Foursquare API free calls are limited by day and won't suffice to process all the Mexico cities (i.e. 50 unique coordinates will be used, latitude/longitude values).

Whenever this process is ready for production environments the filters can be easily removed to analyze the whole country if needed.

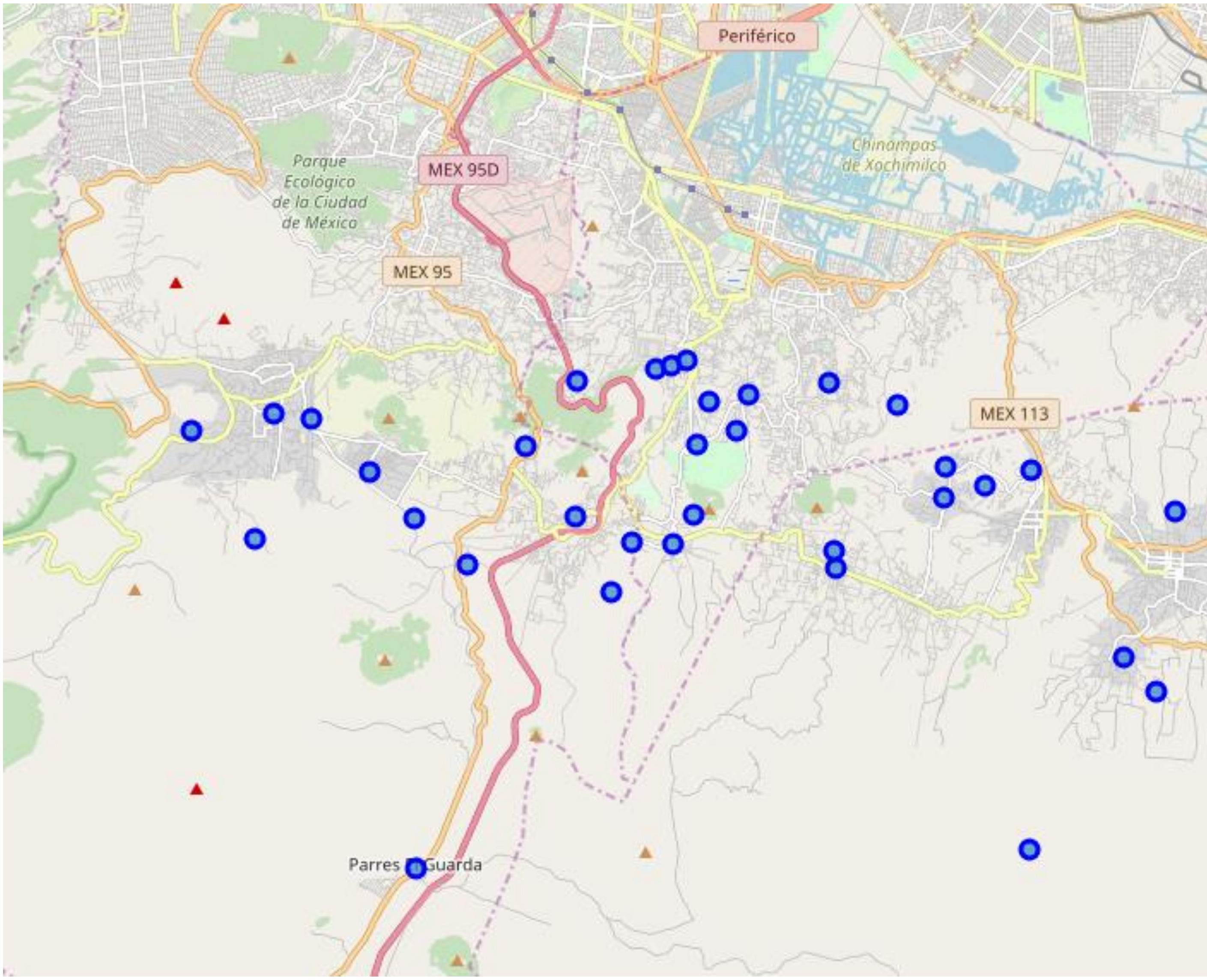
The Major City spots data has several place names for the same latitude and longitude because some of the coordinates are estimated (check Accuracy column definition). Because of this reason, a unique combination for latitude and longitude is identified and a list of place names delimited by comma is created as a new column. I'm leaving a record count column to debug the process, but is not needed for the analysis.

The final data frame, major_city_spots, needed for this model has the following columns:



	PlaceName	Latitude	Longitude	RecordCount
0	Parres El Guarda	19.1361	-99.1738	1
1	La Concepción, San Mateo, Los Ángeles, Emilian...	19.1395	-99.0511	10
2	San Marcos	19.1694	-99.0257	1
3	San Lorenzo Tlacoyucan	19.1761	-99.0322	1
4	San Juan, San Juan Tepenahuac	19.1877	-98.9945	2

Below you can see a Mexico City map where each data frame place has a marker.



Popular Venues Identification

A developer account was created in Foursquare to get access to the available endpoints to get the popular venues from a particular location and radius, check the References section for more details.

FYI For this project a 500 meter radius will be used unless something different is noted.

One single Foursquare endpoint will be used to get the places nearby a specific location defined by its latitude and longitude. The search API call will return a JSON object that should be read to identify the venue's attributes: name, latitude, longitude and category. The category is what will identify the business type of the venue. Take into account that the identified venues are those which the Foursquare users have been checked-in a visit.

A couple of Python functions were created, searchNearbyVenues() to read the JSON object returned by the Foursquare API and return_most_common_venues() to sort the most common venues nearby a specific place order by its frequency (check-in records). In particular, the first function manage the exception whenever for the specific place there are no venues around.

This is the API call output for the 6th record of a sample location, take into account that the client_id, client_secret and API version are mandatory parameters.

```
{'id': '522e010f11d2740eb812b26f',
 'name': 'Laboratorio Biologia A-23A',
 'location': {'lat': 19.353087092842518,
  'lng': -99.15553406100847,
  'labeledLatLngs': [{'label': 'display',
    'lat': 19.353087092842518,
    'lng': -99.15553406100847}],
  'distance': 31,
  'cc': 'MX',
  'country': 'México',
  'formattedAddress': ['México']},
 'categories': [{'id': '4bf58dd8d48988d1a5941735',
   'name': 'College Lab',
   'pluralName': 'College Labs',
   'shortName': 'Lab',
   'icon': {'prefix':
'https://ss3.4sqi.net/img/categories_v2/education/lab_',
    'suffix': '.png'},
   'primary': True}],
 'referralId': 'v-1564269326',
 'hasPerk': False}
```

Methodology

With the acquired Data Science knowledge from all the courses included in the IBM Data Science Professional Coursera track an AI venue type investment model will be created. What information will be shown by this model? First of all, a map that will display 7 different clusters that group neighborhoods with similar venue types. As learned through all the past weeks, data visualization is always more advantageous to communicate ideas. Secondly, for each cluster a list of most common venue types will be shown ordered by it's frequency. With these lists, the Government or private investors can decide where to open a new business (on which neighborhood) and of what type (based on the venue's category from the social network used). They can also notice if there is any type of missing business required on the specific neighborhoods.

Data Science Tools & Algorithms

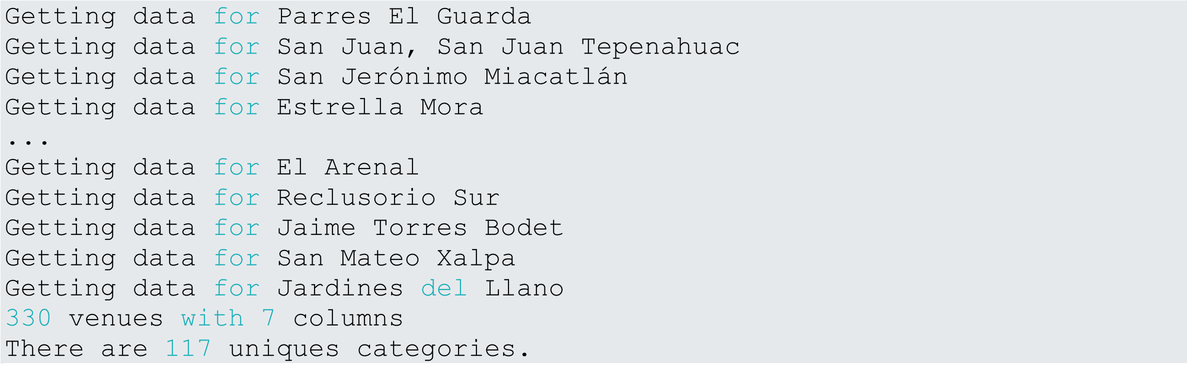
Python Notebook was created in Skills Network Labs framework for this project, check the References section for more details.

k-means clustering method will be used to group the data identifying the 10 more common venues in 500 meters radius for each identified zone/neighborhood. After identifying all the venues, the data for each cluster will be grouped by category's venue to show the most frequent categories.

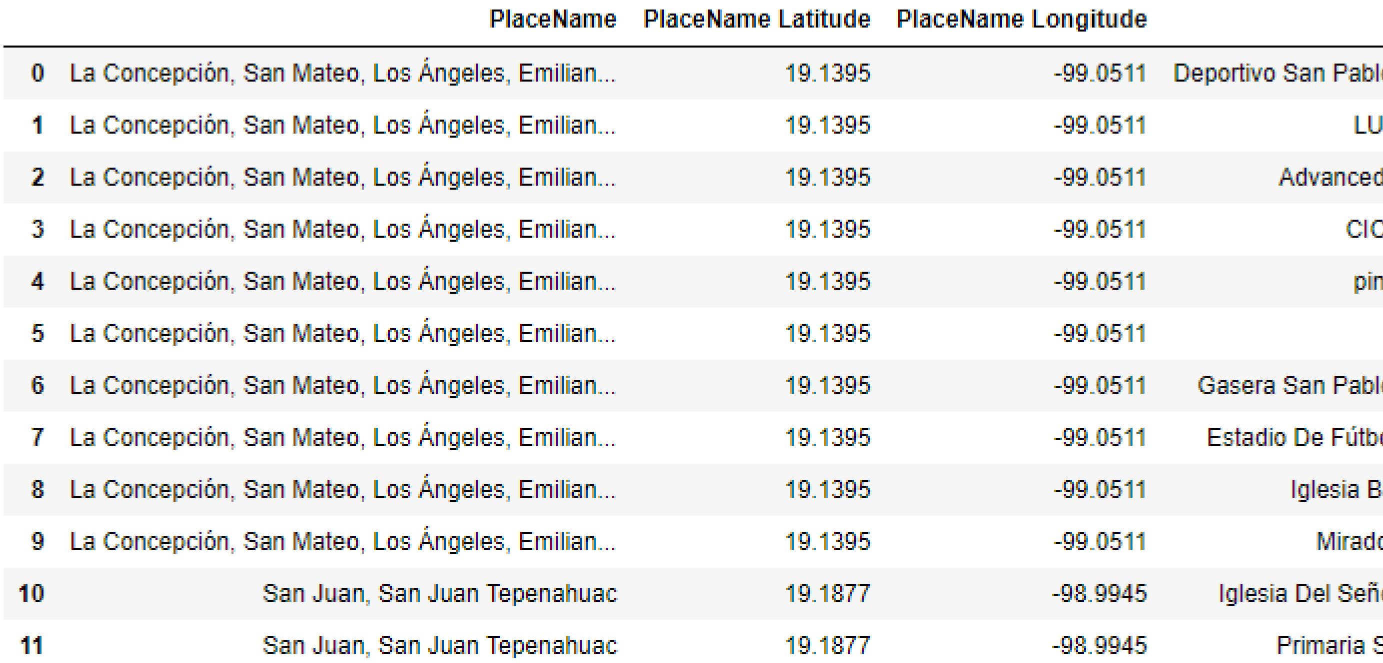
FYI A matrix of 10 x n will be created by the algorithm (number of the most common venues times the number of neighborhoods) for each cluster

Detailed results & discussion

Now that the Major Spots are identified and the functions are ready to be used let's get all the city venues based on the Major City Spots data frame.

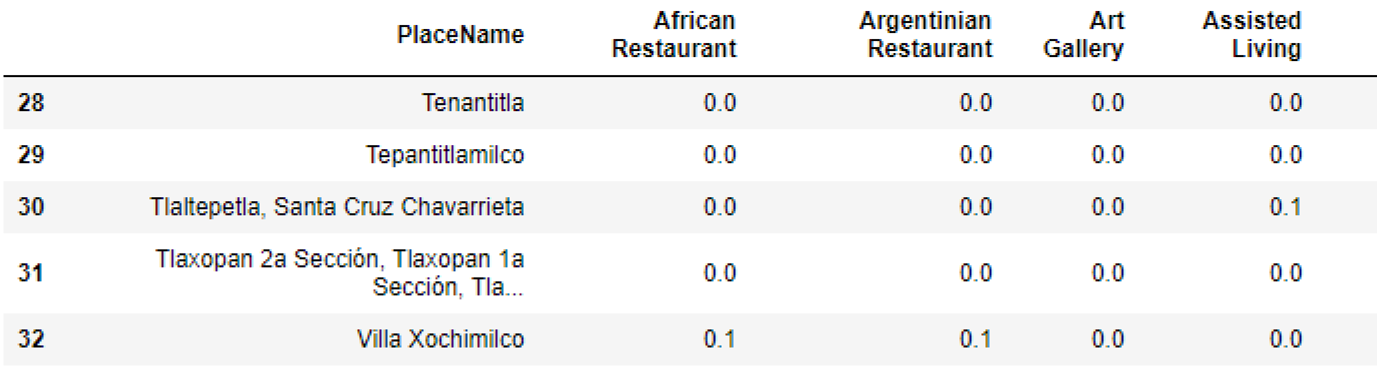


```
Getting data for Parres El Guarda
Getting data for San Juan, San Juan Tepenahuac
Getting data for San Jerónimo Miacatlán
Getting data for Estrella Mora
...
Getting data for El Arenal
Getting data for Reclusorio Sur
Getting data for Jaime Torres Bodet
Getting data for San Mateo Xalpa
Getting data for Jardines del Llano
330 venues with 7 columns
There are 117 uniques categories.
```



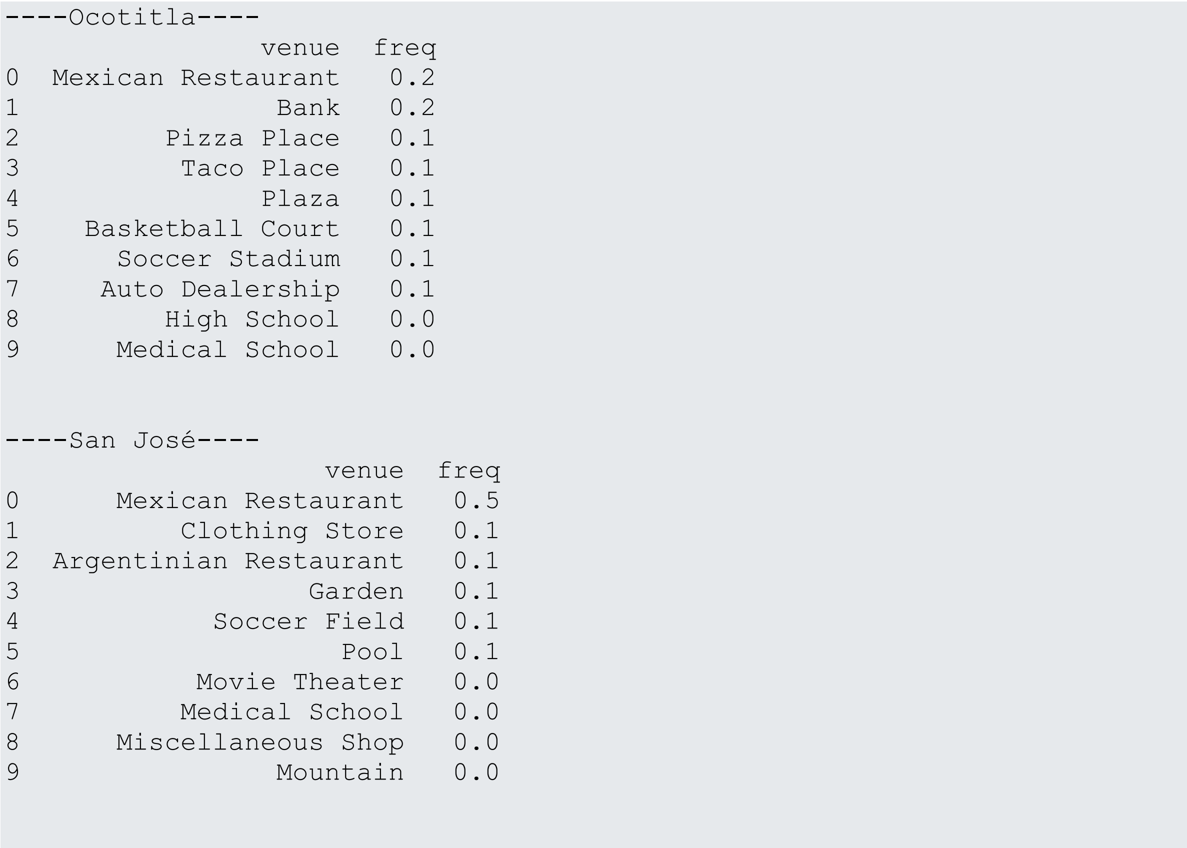
	PlaceName	PlaceName Latitude	PlaceName Longitude	
0	La Concepción, San Mateo, Los Ángeles, Emilian...	19.1395	-99.0511	Deportivo San Pabl
1	La Concepción, San Mateo, Los Ángeles, Emilian...	19.1395	-99.0511	LU
2	La Concepción, San Mateo, Los Ángeles, Emilian...	19.1395	-99.0511	Advanced
3	La Concepción, San Mateo, Los Ángeles, Emilian...	19.1395	-99.0511	CIC
4	La Concepción, San Mateo, Los Ángeles, Emilian...	19.1395	-99.0511	pin
5	La Concepción, San Mateo, Los Ángeles, Emilian...	19.1395	-99.0511	
6	La Concepción, San Mateo, Los Ángeles, Emilian...	19.1395	-99.0511	Gasera San Pabl
7	La Concepción, San Mateo, Los Ángeles, Emilian...	19.1395	-99.0511	Estadio De Fútb
8	La Concepción, San Mateo, Los Ángeles, Emilian...	19.1395	-99.0511	Iglesia B
9	La Concepción, San Mateo, Los Ángeles, Emilian...	19.1395	-99.0511	Mirado
10	San Juan, San Juan Tepenahuac	19.1877	-98.9945	Iglesia Del Señ
11	San Juan, San Juan Tepenahuac	19.1877	-98.9945	Primaria S

Based on this data frame, city_venues_grouped data frame was generated to identify the venues categories for each major spot. You can think about it as a matrix of n rows and m columns where n is the number of distinct major spots and m is the number of categories (in this case a 33 x 118 matrix).



	PlaceName	African Restaurant	Argentinian Restaurant	Art Gallery	Assisted Living
28	Tenantitla	0.0	0.0	0.0	0.0
29	Tepantitlamilco	0.0	0.0	0.0	0.0
30	Tlaltepetla, Santa Cruz Chavarrieta	0.0	0.0	0.0	0.1
31	Tlaxopan 2a Sección, Tlaxopan 1a Sección, Tla...	0.0	0.0	0.0	0.0
32	Villa Xochimilco	0.1	0.1	0.0	0.0

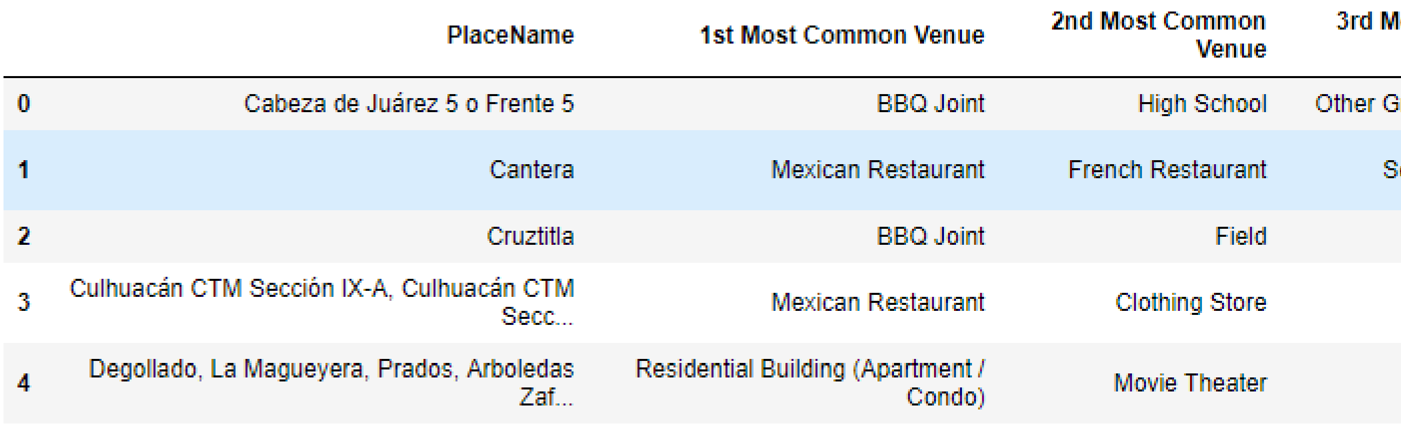
Let's display the data in a different way. For a couple of Major City Spots the venue category frequency will be shown.



```
----Ocotitla----
                venue  freq
0  Mexican Restaurant   0.2
1                Bank   0.2
2         Pizza Place   0.1
3          Taco Place   0.1
4               Plaza   0.1
5    Basketball Court   0.1
6      Soccer Stadium   0.1
7     Auto Dealership   0.1
8         High School   0.0
9      Medical School   0.0


----San José----
                    venue  freq
0      Mexican Restaurant   0.5
1          Clothing Store   0.1
2  Argentinian Restaurant   0.1
3                  Garden   0.1
4            Soccer Field   0.1
5                    Pool   0.1
6           Movie Theater   0.0
7          Medical School   0.0
8      Miscellaneous Shop   0.0
9                Mountain   0.0
```

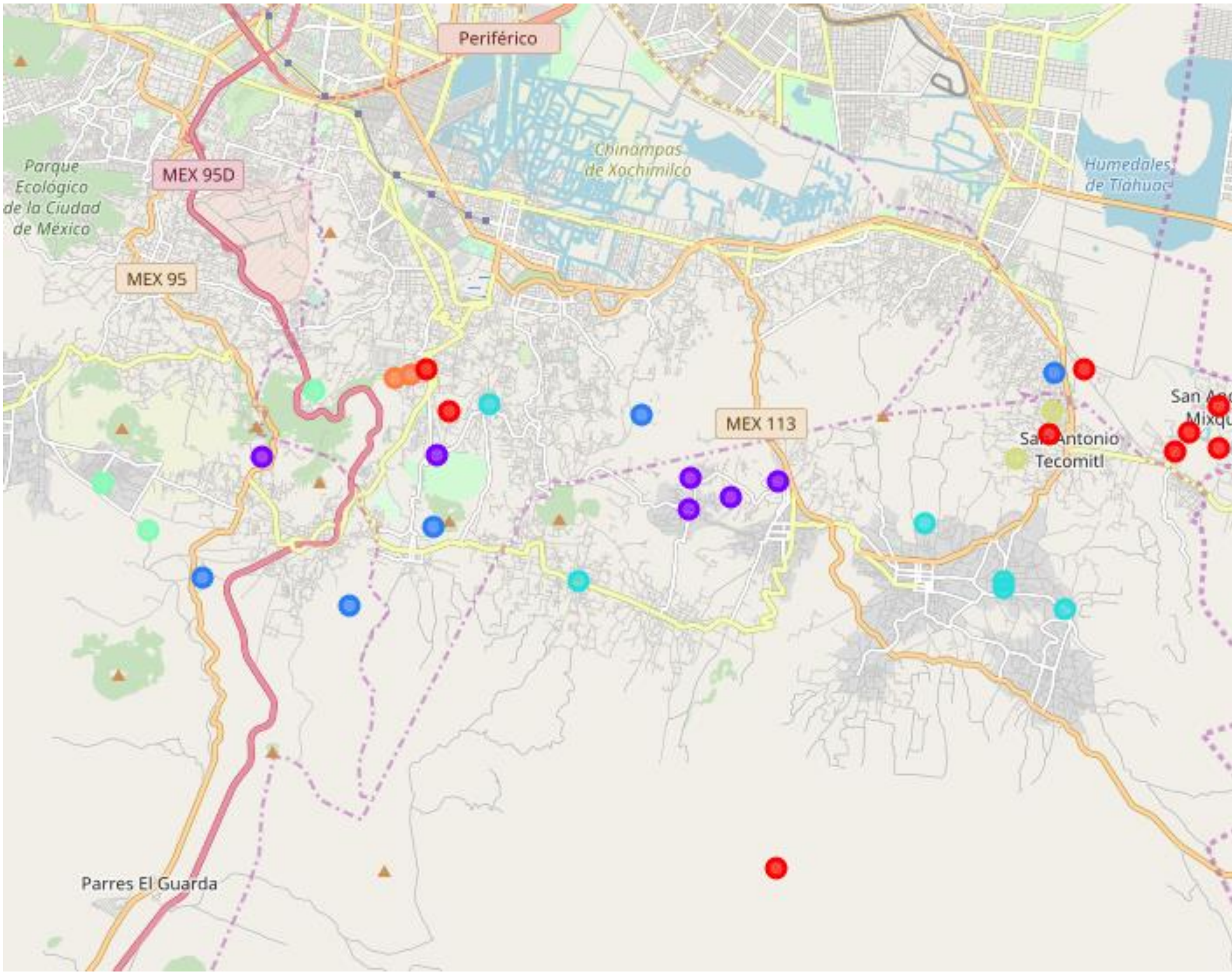
The next step is to create the final data frame city_venues_sorted to visualize for each Major Spot the 10th most common venues. Later I will add the cluster associated to each row.



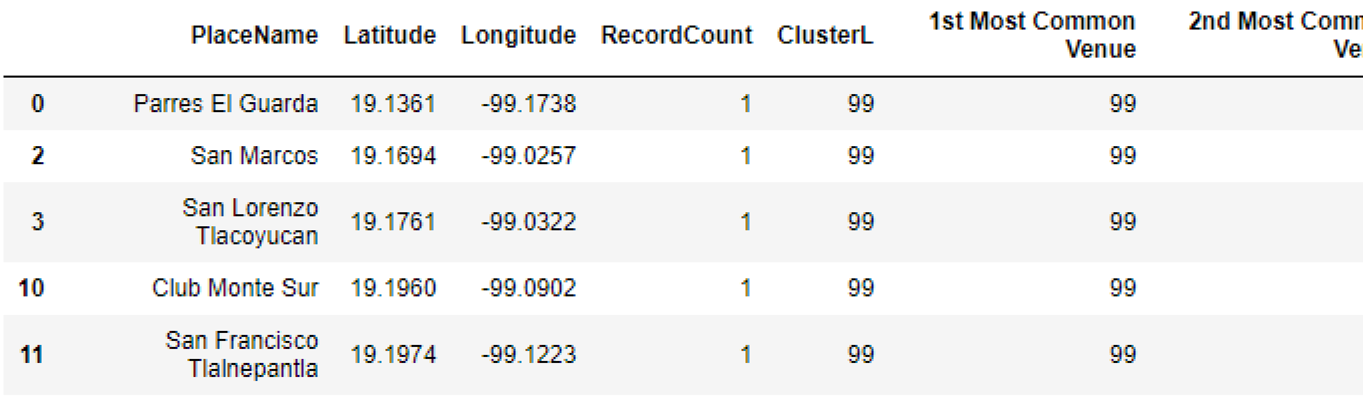
	PlaceName	1st Most Common Venue	2nd Most Common Venue	3rd M
0	Cabeza de Juárez 5 o Frente 5	BBQ Joint	High School	Other G
1	Cantera	Mexican Restaurant	French Restaurant	S
2	Cruztitla	BBQ Joint	Field	
3	Culhuacán CTM Sección IX-A, Culhuacán CTM Secc...	Mexican Restaurant	Clothing Store	
4	Degollado, La Magueyera, Prados, Arboledas Zaf...	Residential Building (Apartment / Condo)	Movie Theater	

k-means Clustering

Everything is now set to identify 7 clusters to group the venues and display them in a map. The first step is to drop the place name from the data because otherwise it will only cause noise. Now let's use the existing city_venues_sorted data frame and add a new column to identify on which cluster every major spot will be grouped. Finally, to make the data frame more readable a join with the city_major_spots was done to show the latitude and longitude as well.



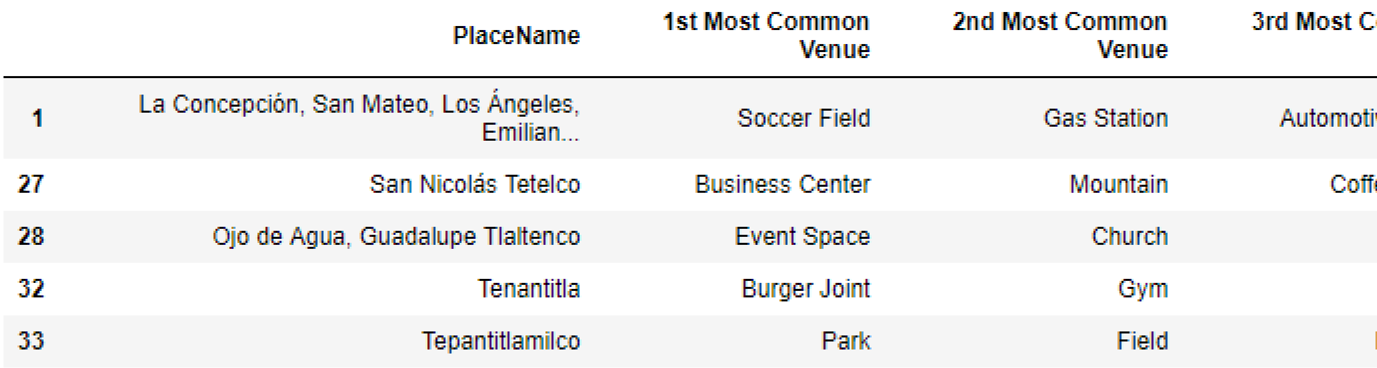
An important explanation must be added to this report. In the k-means clustering iteration process some of the centroids can *die*, meaning that a particular Major Spot can't be grouped to any other because there are not enough information to do so. This scenario happened in this project and those places were removed without impacting the model, check the References section for more details. A data frame city_venues_wo_cluster was created to make a separate analysis about those places in a different project.



	PlaceName	Latitude	Longitude	RecordCount	ClusterL	1st Most Common Venue	2nd Most Comm Ve
0	Parres El Guarda	19.1361	-99.1738	1	99	99	
2	San Marcos	19.1694	-99.0257	1	99	99	
3	San Lorenzo Tlacoyucan	19.1761	-99.0322	1	99	99	
10	Club Monte Sur	19.1960	-99.0902	1	99	99	
11	San Francisco Tlalnepantla	19.1974	-99.1223	1	99	99	

The final output

Now for each cluster the most common venues nearby a particular major spot are identified. The investors actually can use this information to make their decisions, they have 7 different groups of venues and at least 10 most common venues for each one, this give them more than 70 different options to chose from. In particular, if you remember the beginning of the Detailed results & discussion section, they will have 330 different venues. Let's see how the first cluster looks like:



	PlaceName	1st Most Common Venue	2nd Most Common Venue	3rd Most C
1	La Concepción, San Mateo, Los Ángeles, Emilian...	Soccer Field	Gas Station	Automoti
27	San Nicolás Tetelco	Business Center	Mountain	Coffe
28	Ojo de Agua, Guadalupe Tlaltenco	Event Space	Church	
32	Tenantitla	Burger Joint	Gym	
33	Tepantitlamilco	Park	Field	

Finally, let's check what is the venue distribution by cluster.

```
             Cluster 0: 9 rows --> 18.0% places included
             Cluster 1: 6 rows --> 12.0% places included
             Cluster 2: 5 rows --> 10.0% places included
             Cluster 3: 6 rows --> 12.0% places included
             Cluster 4: 3 rows --> 6.0% places included
             Cluster 5: 2 rows --> 4.0% places included
             Cluster 6: 2 rows --> 4.0% places included
Venues without cluster: 17 rows --> 34.0% places included
```

An additional way of grouping the data will be done in case the investors want to know the frequency for each venue category regardless in which position of the most common venues appear. E.g. if you check cluster0 data shown above, you will notice that Beer Garden venue category appears as the fourth most common venue a couple of times and one more time as the eighth most common venue. On the other hand, Burger Joint venue category appears only one time in the whole data set as the most common venue for Tenantitla Major Spot. Some investors may give more weight to their decision to a venue category that seems more successful because appears more than once in the cluster than another that may be one of a kind.

Let's group the results in cluster0, count how many times every venue category is present in the most common venues for all the Major Spots and show the new category list order by the sum of the record count. This is the result of the venue categories in cluster0:

```
9 records in cluster 0, 5 top venues = 90 entries
Plaza                           5.0
Church                          4.0
Park                            4.0
Beer Garden                     3.0
Cemetery                        3.0
Coffee Shop                     3.0
Field                           3.0
Elementary School               2.0
Soccer Field                    2.0
Public Art                      2.0
Cultural Center                 2.0
Stables                         2.0
Gym                             2.0
Mexican Restaurant              2.0
Athletics & Sports              2.0
Taco Place                      2.0
Cupcake Shop                    1.0
Dessert Shop                    1.0
Dance Studio                    1.0
Dentist's Office                1.0
College Library                 1.0
Event Space                     1.0
College Rec Center              1.0
University                      1.0
College Classroom               1.0
College Auditorium              1.0
College Academic Building       1.0
Food & Drink Shop               1.0
Business Center                 1.0
Burger Joint                    1.0
Brewery                         1.0
Breakfast Spot                  1.0
BBQ Joint                       1.0
Automotive Shop                 1.0
City Hall                       1.0
General Entertainment           1.0
Food Court                      1.0
Miscellaneous Shop              1.0
Student Center                  1.0
Soccer Stadium                  1.0
Shipping Store                  1.0
Sandwich Place                  1.0
Salon / Barbershop              1.0
Pizza Place                     1.0
Outdoors & Recreation           1.0
Outdoor Supply Store            1.0
Museum                          1.0
Movie Theater                   1.0
```

```
Mountain                        1.0
Medical School                  1.0
Frozen Yogurt Shop              1.0
Medical Center                  1.0
Market                          1.0
Library                         1.0
Lake                            1.0
Juice Bar                       1.0
Ice Cream Shop                  1.0
History Museum                  1.0
High School                     1.0
Government Building             1.0
General College & University    1.0
Gas Station                     1.0
Assisted Living                 1.0
Name: PlaceName, dtype: float64
```

Here are the results for the remaining 6 clusters:

```
6 records in cluster 1, 10 top venues = 60 entries
Mexican Restaurant         6.0
Clothing Store             4.0
Country Dance Club         4.0
Argentinian Restaurant     3.0
Cupcake Shop               3.0
Convenience Store          3.0
Soccer Field               3.0
Dentists Office           3.0
Comfort Food Restaurant    2.0
Dessert Shop               2.0
Market                     2.0
Garden                     2.0
Funeral Home               2.0
Dance Studio               2.0
College Library            1.0
College Rec Center         1.0
Business Service           1.0
Cultural Center            1.0
Brewery                    1.0
Strip Club                 1.0
Distribution Center        1.0
Farm                       1.0
Street Fair                1.0
Frozen Yogurt Shop         1.0
Housing Development        1.0
Mountain                   1.0
Pizza Place                1.0
Pool                       1.0
Rest Area                  1.0
Scenic Lookout             1.0
Shipping Store             1.0
Stables                    1.0
French Restaurant          1.0
Name: PlaceName, dtype: float64

5 records in cluster 2, 10 top venues = 50 entries
Housing Development                         4.0
Other Great Outdoors                        4.0
```

```
Movie Theater                               3.0
Rest Area                                   2.0
Field                                       2.0
High School                                 2.0
BBQ Joint                                   2.0
Bank                                        1.0
Breakfast Spot                              1.0
Church                                      1.0
Hardware Store                              1.0
Convenience Store                           1.0
Cultural Center                             1.0
Cupcake Shop                                1.0
Assisted Living                             1.0
Art Gallery                                 1.0
Dog Run                                     1.0
Festival                                    1.0
Argentinian Restaurant                      1.0
General College & University                1.0
College Rec Center                          1.0
Trail                                       1.0
Temple                                      1.0
Italian Restaurant                          1.0
Medical Center                              1.0
Mexican Restaurant                          1.0
Outdoors & Recreation                       1.0
Park                                        1.0
Pharmacy                                    1.0
Public Art                                  1.0
Residential Building (Apartment / Condo)    1.0
Salon / Barbershop                          1.0
Soccer Field                                1.0
Spa                                         1.0
Speakeasy                                   1.0
Student Center                              1.0
Taco Place                                  1.0
African Restaurant                          1.0
Name: PlaceName, dtype: float64

6 records in cluster 3, 10 top venues = 60 entries
Mexican Restaurant    6.0
Auto Dealership       4.0
Pizza Place           4.0
Taco Place            4.0
Plaza                 4.0
Bank                  4.0
Café                  2.0
Soccer Stadium        2.0
Scenic Lookout        2.0
Church                2.0
Office                2.0
Internet Cafe         2.0
Dry Cleaner           2.0
Dentist's Office      2.0
Dance Studio          2.0
Campaign Office       1.0
Country Dance Club    1.0
Basketball Court      1.0
```

```
Bar                   1.0
Coffee Shop           1.0
Convenience Store     1.0
University            1.0
Cultural Center       1.0
Food & Drink Shop     1.0
Recreation Center     1.0
Restaurant            1.0
Speakeasy             1.0
Stables               1.0
Street Fair           1.0
Student Center        1.0
Ice Cream Shop        1.0
Name: PlaceName, dtype: float64

3 records in cluster 4, 10 top venues = 30 entries
Field                      3.0
Scenic Lookout             3.0
Convenience Store          3.0
School                     2.0
Mexican Restaurant         2.0
Italian Restaurant         2.0
Taco Place                 2.0
Country Dance Club         1.0
Campground                 1.0
Clothing Store             1.0
College Rec Center         1.0
Comfort Food Restaurant    1.0
Toll Booth                 1.0
Distribution Center        1.0
Electronics Store          1.0
French Restaurant          1.0
Hardware Store             1.0
Mountain                   1.0
Salon / Barbershop         1.0
Assisted Living            1.0
Name: PlaceName, dtype: float64
2 records in cluster 5, 10 top venues = 20 entries
BBQ Joint                  2.0
General Entertainment      2.0
Diner                      2.0
Taco Place                 1.0
Distribution Center        1.0
Bar                        1.0
Cafeteria                  1.0
College Classroom          1.0
Comfort Food Restaurant    1.0
Event Space                1.0
Soccer Field               1.0
Field                      1.0
Food Truck                 1.0
Garden                     1.0
Nail Salon                 1.0
Park                       1.0
Auto Garage                1.0
Name: PlaceName, dtype: float64
```

```
2 records in cluster 5, 10 top venues = 20 entries
Stationery Store           2.0
Outdoors & Recreation      2.0
Food Stand                 2.0
Festival                   2.0
University                 1.0
Taco Place                 1.0
Stables                    1.0
Pool                       1.0
Pharmacy                   1.0
Outdoor Sculpture          1.0
Miscellaneous Shop         1.0
Health & Beauty Service    1.0
Food Truck                 1.0
Diner                      1.0
Dessert Shop               1.0
Cupcake Shop               1.0
Name: PlaceName, dtype: float64
```

The results of this project have shown how Data Science algorithms and tools can help to find hidden information in large data sets. Is really important to notice that with the generated information a lot of different analysis can be done, defining clear objectives a big bunch of questions will be answered with certainty and confidence. Maybe some of the results were expected like the one that shows that a *Mexican Restaurant or a Taco Place* are part of the top venue categories because they appear in at least 5 of the 7 clusters, but we are Mexicans, right? There are some other results that were difficult to anticipate, e.g. cluster3 is a good place to start an *Auto Dealership* business. Besides, opening a *Sporting Goods Shop* nearby a *Soccer Field* venue would be a good idea to start a new business. A *Soccer Field* venues appears in 4 of the 7 clusters and there is no trending *Sporting Goods Shop* close those. Long story short, the available information can be used to make better decisions.

Conclusions

I hope you agree with me that the analysis done in this project can be applied broadly if and only if the essential data is present for the city or group of cities where the model wants to be applied. Adding more available data to the model in the future, like real estate costs or criminal rate, will make it superior. On the other hand, the developed code is easy to follow to make a similar analysis for any part of the world.

> *In the business people with expertise, experience and evidence will make more profitable decisions than people with instinct, intuition and imagination.*
>
> *Amit Kalantri* https://www.goodreads.com/quotes/tag/data-science

References

- Volatile economies article
- Geolocation Mexico Postal Codes
- Foursquare endpoints
- Data Science framework
- k-means clustering
- Why K-means centroids contain NAN?
- AI Venue Type Investment Model GitHub Code