Analysis of Mexico City's Venues

Marco Antonio Xoca Orozco

14 December, 2020

1 Introduction and business problem

Mexico City is the capital, largest city, and most populous city in North America. This city is also one of the most important and cultural centers in the world.

By 2015 the estimated population was about 9,000,000, also Mexico City, or CDMX has 16 municipalities.

In this project I'll find the information about landmarks, restaurants, shops, and more venues of each municipalities, in order to help people to find the best municipality to visit according at their interest

2 Data

In order to get all the information needed, we can list the data as below:

- List of each municipality in CDMX
- Coordinates (latitude and longitude) of each municipality in CDMX
- List of every venue by municipalities (obtained through Foursquare API)

For the list of each municipality, I scrapped the CDMX's Wikipedia webpage, which is the same table as below:

Demarcaciones territoriales de la Ciudad de México[8]

Alcaldías	Población (2016)	Superficie (km²)	Densidad de habitantes (km²)	Código postal
Álvaro Obregón	727,034	96.17	7.347	01000 - 01999
Azcapotzalco	414,711	33.66	12.635	02000 - 02999
Benito Juárez	385,439	26.63	13.331	03000 - 03999
Coyoacán	620,416	54.4	11.545	04000 - 04999
Cuajimalpa de Morelos	186,391	74.58	2.328	05000 - 05999
Cuauhtémoc	531,831	32.4	16.071	06000 - 06999
Gustavo A. Madero	1,185,772	94.07	12.683	07000 - 07999
Iztacalco	384,326	23.3	16.953	08000 - 08999
Iztapalapa	1,815,786	117	15.563	09000 - 09999
Magdalena Contreras	239,086	74.58	3.069	10000 - 10999
Miguel Hidalgo	372,889	46.99	7.523	11000 - 11999
Milpa Alta	130,582	228.41	507	12000 - 12999
Tláhuac	360,265	85.34[9]	4.032	13000 - 13999
Tlalpan	650,567	312	2.085	14000 - 14999
Venustiano Carranza	430,978	33.4	13.396	15000 - 15999
Xochimilco	415,007	122	3.427	16000 - 16999

Figure 1: Table of CDMX municipalities

For the coordinates of each municipality, I had to use the CDMX's government database and separate the value of latitude and longitude for each municipality, this database is show below:

	NOMBRE	CLAVE_MUNICIPAL	CVE_ENTIDAD	CVEGEO	Geo Point
1	Coyoacán	003	09	09003	19.3266672536, -99.1503763525
2	Miguel Hidalgo	016	09	09016	19.4280623649, -99.2045669144
3	La Magdalena Contreras	008	09	09008	19.2689765031, -99.2684129061
4	Tláhuac	011	09	09011	19.2769983772, -99.0028216137
5	Azcapotzalco	002	09	09002	19.4853286147, -99.1821069423
6	Iztacalco	006	09	09006	19.396911897, -99.094329797
7	Álvaro Obregón	010	09	09010	19.336175562, -99.246819712
8	Xochimilco	013	09	09013	19.2451450458, -99.0903636045
9	Venustiano Carranza	017	09	09017	19.4304954545, -99.0931057959
10	Tlalpan	012	09	09012	19.1983396763, -99.2062207957
11	Cuajimalpa de Morelos	004	09	09004	19.3246343001, -99.3107285253
12	Cuauhtémoc	015	09	09015	19.4313734294, -99.1490557562

Figure 2: Table of CDMX municipality's coordinates

After all the preprocessing, the final data was this:

	borough	population	area	density	postal code	Latitude	Longitude
0	Azcapotzalco	414711	33.66	12.635	02000 - 02999	19.485329	-99.182107
1	Benito Juárez	385439	26.63	13.331	03000 - 03999	19.380642	-99.161135
2	Coyoacán	620416	54.4	11.545	04000 - 04999	19.326667	-99.150376
3	Cuajimalpa de Morelos	186391	74.58	2.328	05000 - 05999	19.324634	-99.310729
4	Cuauhtémoc	531831	32.4	16.071	06000 - 06999	19.431373	-99.149056
5	Gustavo A. Madero	1185772	94.07	12.683	07000 - 07999	19.504065	-99.115864
6	Iztacalco	384326	23.3	16.953	08000 - 08999	19.396912	-99.094330
7	Iztapalapa	1815786	117	15.563	09000 - 09999	19.349166	-99.056799
8	Magdalena Contreras	239086	74.58	3.069	10000 - 10999	19.268977	-99.268413
9	Miguel Hidalgo	372889	46.99	7.523	11000 - 11999	19.428062	-99.204567
10	Milpa Alta	130582	228.41	507	12000 - 12999	19.139457	-99.051095
11	Tlalpan	650567	312	2.085	14000 - 14999	19.198340	-99.206221
12	Tláhuac	360265	85.34[9]	4.032	13000 - 13999	19.276998	-99.002822
13	Venustiano Carranza	430978	33.4	13.396	15000 - 15999	19.430495	-99.093106
14	Xochimilco	415007	122	3.427	16000 - 16999	19.245145	-99.090364
15	Álvaro Obregón	727034	96.17	7.347	01000 - 01999	19.336176	-99.246820

Figure 3: Table of CDMX municipality's coordinates after the preprocesing

3 Methodology

After all the preprocessing of the data and the cleaning of the dataframe, using geolocator's library, I got the geographical coordinates of Mexico City, which were: longitude = 19.4326296 and latitude = -99.1331785

With these coordinates, and using Folium's package, I made a map of the CDMX, with labels at each municipality. The map result is show belong:

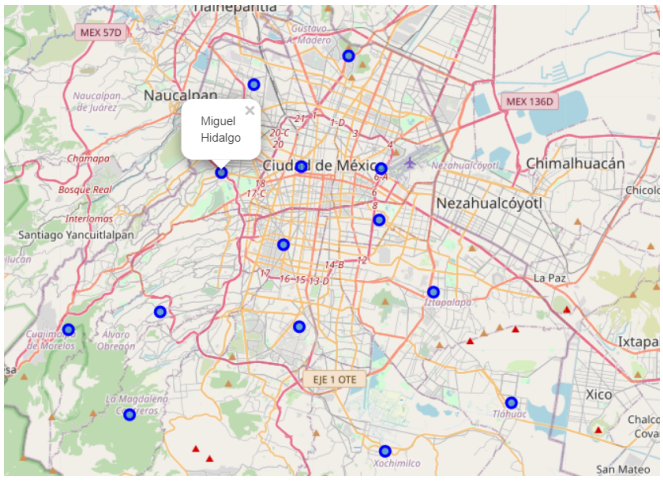

Figure 4: Map of CDMX with labeled municipalities

Once having the CDMX's map, with the help of Foursquare API I got all the venues available by municipality (neighborhood), and venue category, and put them on this dataframe:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Azcapotzalco	19.485329	-99.182107	Neko Café	19.484152	-99.183326	Japanese Restaurant
1	Azcapotzalco	19.485329	-99.182107	Café Revolución	19.484433	-99.183248	Breakfast Spot
2	Azcapotzalco	19.485329	-99.182107	Centro Verde Azcapotzalco	19.487757	-99.182125	Garden
3	Azcapotzalco	19.485329	-99.182107	Frody	19.482723	-99.182390	Ice Cream Shop
4	Azcapotzalco	19.485329	-99.182107	Café ONCE28	19.484427	-99.185720	Breakfast Spot

Figure 5: Dataframe header of every venue in CDMX's municipalities

With that dataframe, I counted how many venues were returned for each municipality, and the unique categories, which were 122 unique venue's categories

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Azcapotzalco	20	20	20	20	20	20
Benito Juárez	27	27	27	27	27	27
Coyoacán	28	28	28	28	28	28
Cuajimalpa de Morelos	1	1	1	1	1	1
Cuauhtémoc	98	98	98	98	98	98

Figure 6: Dataframe header of categories returned by municipality

3.1 Clustering

Before start to clustering all the venues data, I had to apply the one-hot encoding, to determinate the frequency of each venue's category by municipality. The frequency dataframe is show below:

	Neighborhood	Airport	Airport Service	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	BBQ Joint	Bakery	Bar
0	Azcapotzalco	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.050000	0.000000
1	Benito Juárez	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.037037	0.000000	0.037037
2	Coyoacán	0.0	0.0	0.000000	0.035714	0.000000	0.000000	0.035714	0.0	0.000000	0.000000	0.000000
3	Cuajimalpa de Morelos	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000
4	Cuauhtémoc	0.0	0.0	0.020408	0.030612	0.010204	0.020408	0.010204	0.0	0.000000	0.030612	0.030612

Figure 7: Header's frequency of category's venue after the one-hot encoding

Using the one-hot encoding, and with the frequency dataframe, I made the follow dataframe with the top 10 most common venue by municipality. This would help to a visitor to determinate which one municipality it is better to visit

	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	Azcapotzalco	Mexican Restaurant	Ice Cream Shop	Taco Place	Seafood Restaurant	Breakfast Spot	Burrito Place	Bakery	Food Court	Japanese Restaurant	Garden
1	Benito Juárez	Pizza Place	Mexican Restaurant	Coffee Shop	Pet Store	Southern / Soul Food Restaurant	Greek Restaurant	IT Services	Ice Cream Shop	Dog Run	Deli / Bodega
2	Coyoacán	Taco Place	Seafood Restaurant	Ice Cream Shop	Gym / Fitness Center	Fast Food Restaurant	Park	Coffee Shop	Mexican Restaurant	Pizza Place	Pool
3	Cuajimalpa de Morelos	Park	Women's Store	Comfort Food Restaurant	Convenience Store	Cupcake Shop	Cycle Studio	Deli / Bodega	Design Studio	Dessert Shop	Diner
4	Cuauhtémoc	Mexican Restaurant	Taco Place	Deli / Bodega	Hotel	Coffee Shop	Art Gallery	Bar	Restaurant	Café	Bakery

Figure 8: Header of top 10 most common venues by municipality

Also, in order to visualize better this information, a bar plot was made:

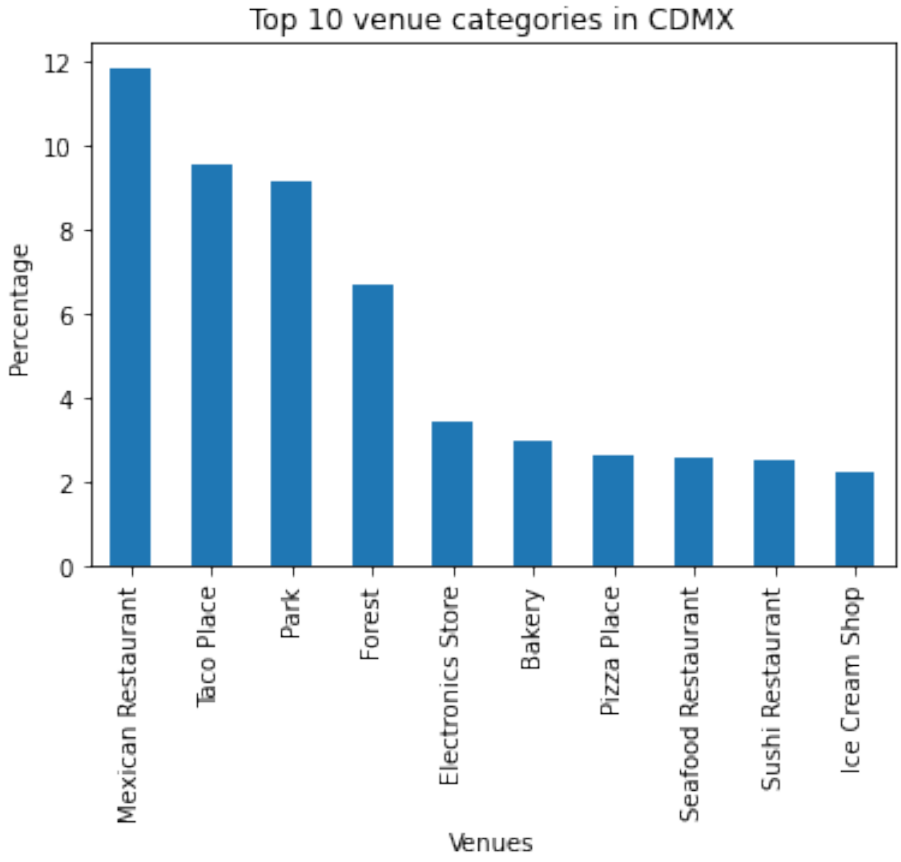

Figure 9: Bar plot of most common venues in all CDMX

3.1.1 Elbow's method

One step just before to start clustering, is to know how many clusters (or the k value) are the optimal for this problem. So, I made a plot of the sum of squared distances with different k-values:

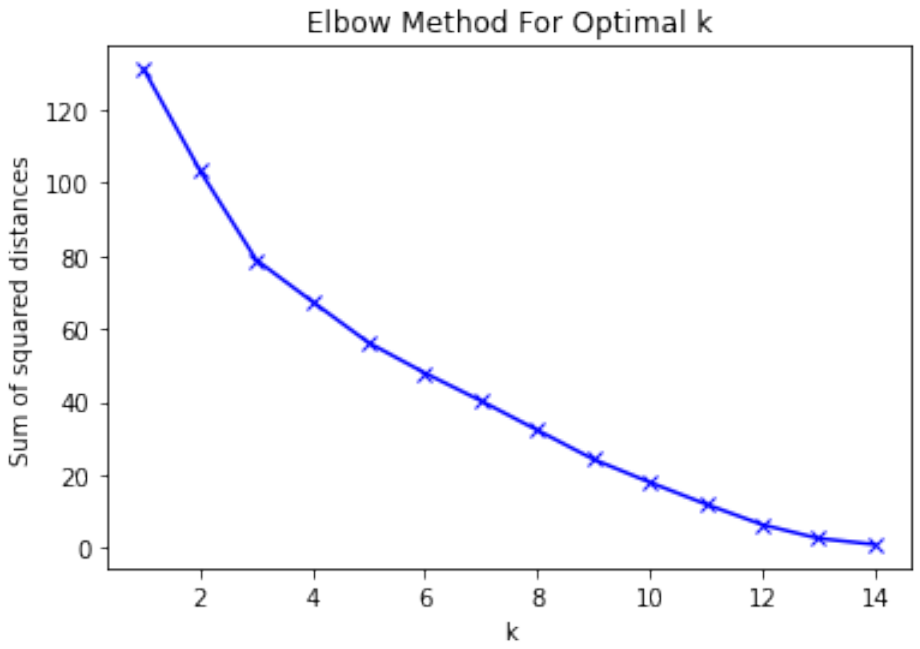

Figure 10: Elbow's method to find optimal k-value

with Fig 10 one can notice that the optimal k-value is 3.

3.2 Clustering with k=3

Using the KMeans library, I divided all the venues data in 3 clusters. The dataframe with each labeling is below:

	borough	population	area	density	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
0	Azcapotzalco	414711	33.66	12.635	19.485329	-99.182107	0	Mexican Restaurant	Ice Cream Shop	Taco Place	Seafood Restaurant
1	Benito Juárez	385439	26.63	13.331	19.380642	-99.161135	0	Pizza Place	Mexican Restaurant	Coffee Shop	Pet Store
2	Coyoacán	620416	54.4	11.545	19.326667	-99.150376	0	Taco Place	Seafood Restaurant	Ice Cream Shop	Gym / Fitness Center
3	Cuajimalpa de Morelos	186391	74.58	2.328	19.324634	-99.310729	1	Park	Women's Store	Comfort Food Restaurant	Convenience Store
4	Cuauhtémoc	531831	32.4	16.071	19.431373	-99.149056	0	Mexican Restaurant	Taco Place	Deli / Bodega	Hotel

Figure 11: Dataframe with clusters labels

Finally, using Folium library I made a cluster map of CDMX with 3 clusters and their respective labels:

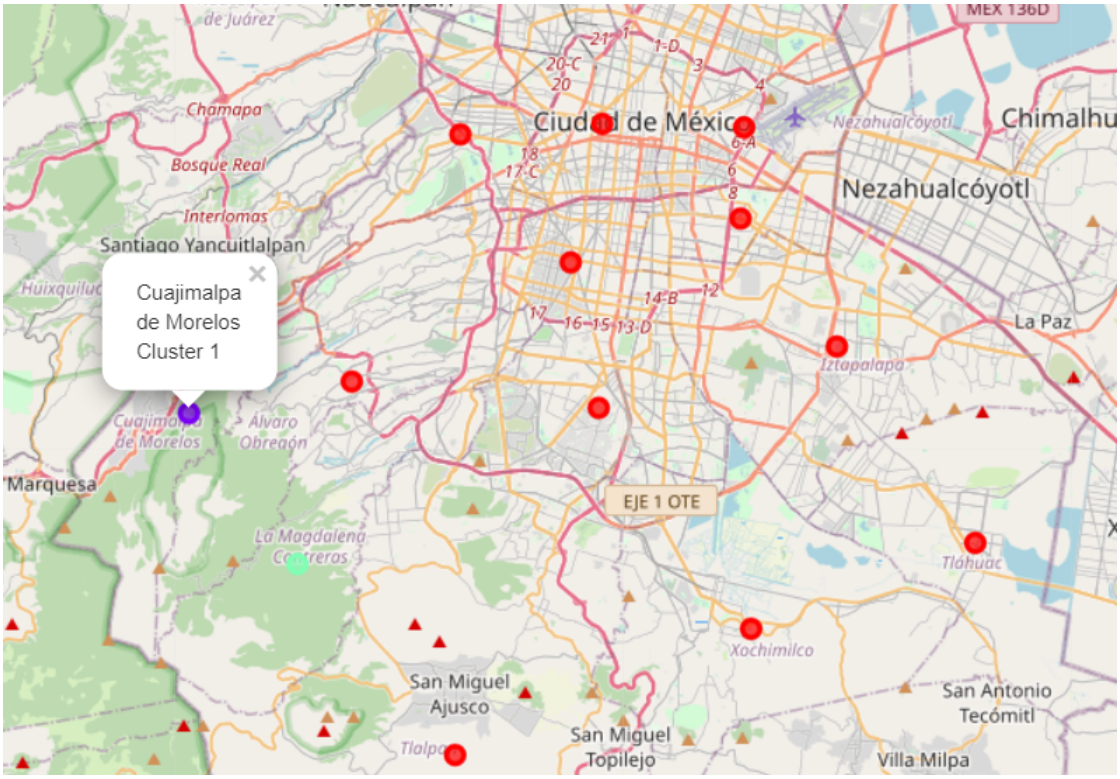

Figure 12: Map of CDMX's venues in 3 clusters

4 Results

Seeing all the tables and maps, we can notice that almost all the venues are just in one cluster, which has a big relation in the urbanized area.

Also we can explore venues by cluster, like the venues in cluster 1 order by common:

	population	Longitude	Cluster Labels	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
0	414711	-99.182107	0	Mexican Restaurant	Ice Cream Shop	Taco Place	Seafood Restaurant	Breakfast Spot	Burrito Place	Bakery	Food Court
1	385439	-99.161135	0	Pizza Place	Mexican Restaurant	Coffee Shop	Pet Store	Southern / Soul Food Restaurant	Greek Restaurant	IT Services	Ice Cream Shop
2	620416	-99.150376	0	Taco Place	Seafood Restaurant	Ice Cream Shop	Gym / Fitness Center	Fast Food Restaurant	Park	Coffee Shop	Mexican Restaurant
4	531831	-99.149056	0	Mexican Restaurant	Taco Place	Deli / Bodega	Hotel	Coffee Shop	Art Gallery	Bar	Restaurant
5	1185772	-99.115864	0	Performing Arts Venue	Hot Dog Joint	Fast Food Restaurant	Scenic Lookout	Athletics & Sports	Movie Theater	Park	Shopping Mall

Figure 13: Dataframe of venues in cluster 1 by top 10 of common.

In cluster 2, we have this top common venues:

	population	Longitude	Cluster Labels	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
3	186391	-99.310729	1	Park	Women's Store	Comfort Food Restaurant	Convenience Store	Cupcake Shop	Cycle Studio	Deli / Bodega	Design Studio	Dessert Shop	Diner

Figure 14: Dataframe of venues in cluster 2 by top 10 of common.

And finally, in cluster 3 we have:

	population	Longitude	Cluster Labels	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
8	239086	-99.268413	2	Forest	Women's Store	Fast Food Restaurant	Convenience Store	Cupcake Shop	Cycle Studio	Deli / Bodega	Design Studio	Dessert Shop	Diner

Figure 15: Dataframe of venues in cluster 3 by top 10 of common.

As we can see in the last three figures, if we like to eat Mexican food, pizza or go to arts venue, we definitely should in the cluster 1 areas. Otherwise, if we like to visit parks we should go to cluster 2 areas. Finally if we like to go to the forest, we should go to cluster's 3 areas.

Also, with Fig 9 we can notice that in all CDMX, the most common venues are, obviously, Mexican restaurants, followed by Taco places, parks and forest, being such a good place to free activities and good food places.

5 Discussions

Being Mexico City one of the bigger cities in the world, probably it is better to analyze venues by municipality. Also, foursquare has not a lot of data of CDMX's venues, so it would be a great idea using another venues API, like the Google's one, or a local government API to get all the necessary data to a better analysis

6 Conclusion

With all this analysis we can conclude that, according to our likes, we can visit different municipalities of CDMX. If we like all kind of restaurants, we should go to any area in cluster 1 (labeled by red in Fig 13), if we like to go to a park or shopping, we should go to any area in cluster 2, showed by the violet label, also in Fig 13. Finally, if we like to go to a forest, shopping, or food restaurant, we should go to any area in cluster 3.

This kind of analysis is very useful if we like to go and visit a new city, but we have no idea what to expect, and with this we can cluster our favorite places and planning in a better way our travel.

Also, this analysis is very straightforward, and you don't need to be the grater programmer of all time in order to do a very well-made analysis and clustering of venues