Clustering Mexico City locations

Jesus Hernández Serda

March 30, 2020

1 Introduction

Mexico City is one of the most densely populated cities in the world. Its history of rapid growth has molded the geographical distribution of its population. In this brief exercise, we will take a look into this distribution in regards to the type of venues that can be found in some regions. With a classification by venue types, we attempt to give a guide on the different lifestyles of some locations in Mexico City and where to find them. We will restrict our analysis to the central municipalities of Mexico City, *Benito Juárez, Cuauhtémoc* and *Coyoacán*.

2 Data

For this analysis, we gathered geographic data from the Mexican Postal Service in the form of two tables available to the public at `https://datos.gob.mx`. One table contains polygon data for all postal codes in Mexico City with no labels while the other contains type, name, associations and postal office for most of the registered postal codes.

The postal codes table includes the following features.

- **Código Postal**: the postal code,
- **Estado**: the state,
- **Municipio**: the municipality,
- **Ciudad**: the city,
- **Tipo de Asentamiento**: the type of location,
- **Asentamiento**: the name of the location,
- **Clave de Oficina**: an ID for the corresponding postal office,

The city and state features had only one value —*Ciudad de México* —and were redundant and therefore dropped. The postal office ID was also dropped. There are six different types of locations: *suburb, neighborhood, village, equipment, encampment*, and *airport*. Each postal code can contain more than one location. There are 41 postal codes with two different types of locations and

only 3 with three types of location. No postal code has more types of locations than that. After inspecting these locations we decided to drop the Type of location feature because we do not expect the historic denominational quality of the feature to reflect any information that the geolocation of venues would not provide.

After grouping the locations by postal code, we merged these tables in a `geopandas` data frame structure for easy handling of the geographic features, including an easy export to `geojson` files for map renderings. Also, we added a *mean coordinates* feature by computing the euclidean centroid of each polygon.

The new table still had some missing data. There were two postal codes for which we did not get any polygon and those were dropped. There were also 284 polygons for which we only had the postal code but no name or municipality. Since we are analyzing only a few municipalities we decided to infer these missing values. After consulting at the local post office we got the following correspondence of municipalities and the first digits of postal codes.

Municipality	**Postal Code**
Alvaro Obregón	01XXX
Azcapotzalco	02XXX
Benito Juárez	03XXX
Coyoacán	04XXX
Cuajimalpa de Morelos	05XXX
Cuauhtémoc	06XXX
Gustavo A. Madero	07XXX
Iztacalco	08XXX
Iztapalapa	09XXX
La Magdalena Contreras	10XXX
Miguel Hidalgo	11XXX
Milpa Alta	12XXX
Tláhuac	13XXX
Tlalpan	14XXX
Venustiano Carranza	15XXX
Xochimilco	16XXX

With this correspondence we identified municipality for all postal codes, as shown in Figure 1.

Our final table has 1404 entries with the following features:

- **CP**: postal code,
- **Municipio**: municipality of the postal code,
- **Asentamiento**: names of the locations in the postal code separated by commas,
- **geometry**: a MultiPolygon object, a geographic depiction of the postal code,

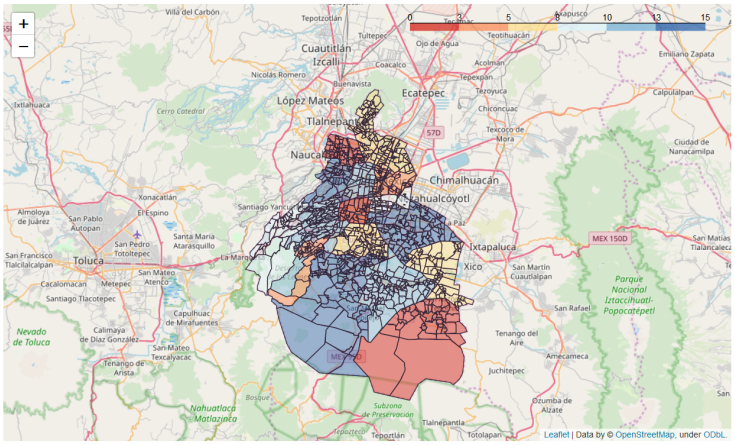


Figure 1: Map of the municipalities of Mexico City.

- **mean coords**: a tuple with the coordinates of the polygon's centroid,
- **m longitude**: mean latitude of the polygon coordinates,
- **m latitude**: mean longitude of the polygon coordinates.

Using the Foursquare API we gathered information about the venues in each postal code of the three central municipalities of Mexico City: *Benito Juárez, Coyoacán* and *Cuauhtémoc*.

We requested up to 150 venues within a radius of 650 meters around the centroid of each postal code. The results were stored in a dataframe for each municipality with the following features:

- **CP**: postal code,
- **Venue**: name of the venue,
- **Venue Latitude**: latitude coordinate of the venue,
- **Venue Longitude**: longitude coordinate of the venue,
- **Venue Category**: type of venue.

3 Methodology

Postal codes vary in size throughout the city. We chose a radius of 650 meters for our queries to include a representative part for most of the locations, but there is definitely an overestimate.

To eliminate redundancies, we matched each venue's coordinates against the corresponding polygon to certify whether or not they belong to the polygon. The results of the Foursquare API query were filtered by testing if each venue is actually inside the polygon in turn. In Figure 2 we can see an example of

Figure 2: Example of a polygon and the queried venue data.

a polygon and also a map with the corresponding venues marked in red when they're not inside the polygon and green when they do.

For the classification, we apply a one-hot transformation to the Category feature of the venues and build a *probability vector* for the venue types for each postal code. For the actual classification, we used two unsupervised clustering algorithms: K-means and agglomerative hierarchical clustering.

We ran some testing using inertia and silhouette scores respectively to choose the best parameters. The inertia test for the K-means algorithm were inconclusive for the three municipalities, see Figure 3. The silhouette test for hierarchical clustering suggested some optimal values for each municipality. We ran the hierarchical clustering algorithm with the parameter $K = 4, 3, 2$ for *Benito Juárez, Cuauhtémoc* and *Coyoacán* respectively. See Figure 4.

4 Results

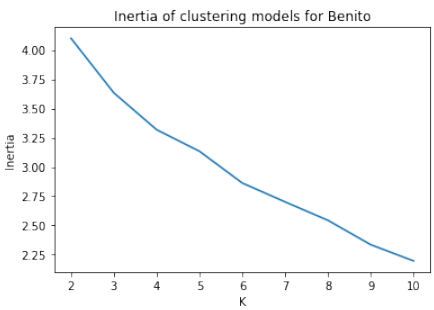
As the inertia test may suggest, for all three municipalities we got one big cluster with most of the postal codes and the rest are some kind of outliers.

We displayed the top ten venue types for each postal code. Nevertheless, there are some postal codes that have less than ten venues available in the queried data. The tables with the top venue types can be seen in the appended notebook.

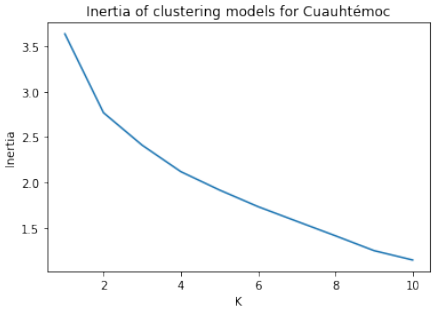
4.1 Clustering Benito Juárez

For Benito Juárez we have 4 clusters, see Figure 5. The first cluster contains 11 postal codes. The most common type of venue is Mexican Restaurant and the second is Taco Place.

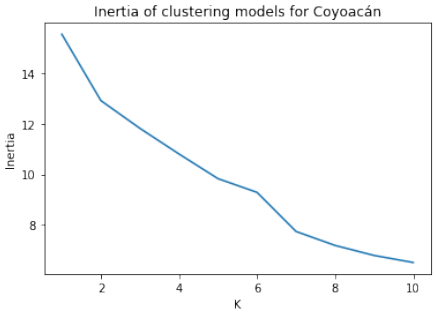
Locations in the second cluster have many restaurants among their top popular venue types, but there is a little more variety in the cuisine styles, including Italian, Japanese and vegetarian. Mexican Restaurant is the top venue type in this cluster. Taco Place and Coffee Shop are the most popular venue types after Mexican restaurants. There is also a significant amount of bakeries, ice cream



(a) Benito Juárez

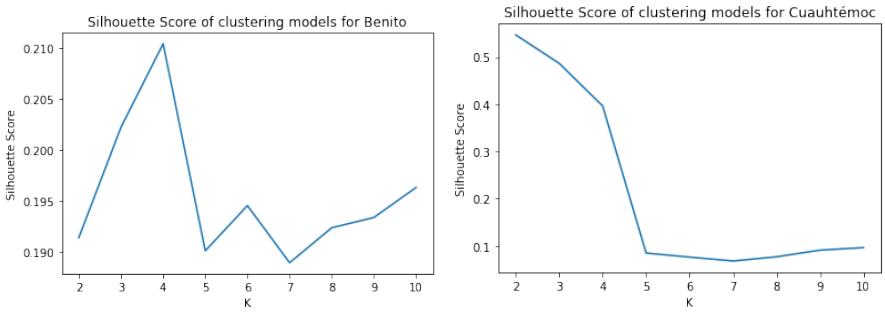


(b) Cuauhtémoc

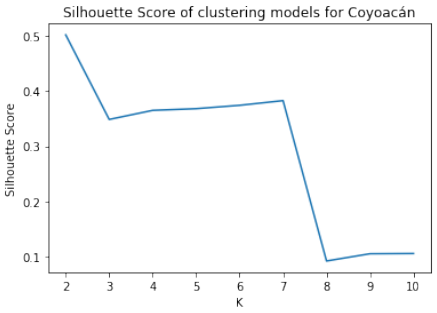


(c) Coyoacán

Figure 3: Inertia tests for K-means.



(a) Benito Juárez (b) Cuauhtémoc



(c) Coyoacán

Figure 4: Silhouette tests for hierarchical clustering.

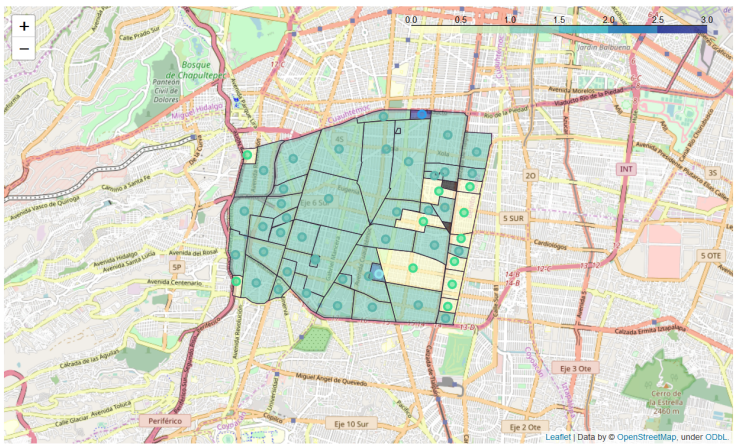


Figure 5: Clustered Benito Juárez.

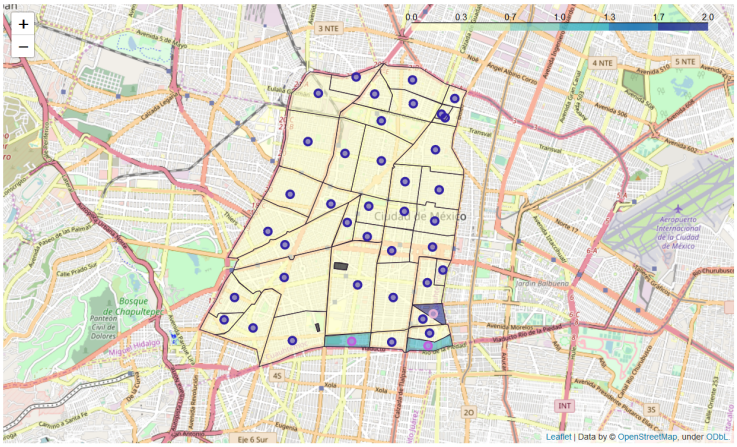


Figure 6: Clustered Cuauhtémoc.

shops, seafood restaurants and gyms in some of the top places for postal codes in this cluster. This cluster contains 39 out of the 54 postal codes in Benito Juárez.

The last two clusters contain only one postal code each. There were two postal codes missing because they had no venues left after filtering.

The third cluster contains a residential zone and the most popular venues are Italian Restaurant and Restaurant. The last cluster contains another residential zone and their most popular venues are Argentinian Restaurant, Food Truck, and Bar. There are no more venue types in any of these locations.

4.2 Clustering Cuauhtémoc

For Cuauhtémoc, we have 3 clusters, see Figure 6. The first cluster is the biggest one with 38 out of the 43 locations in Cuauhtémoc. The most popular type of venue is Mexican Restaurant with Taco Place in second. The top venue types are in Figure 8.

The second cluster contains two postal codes with only two venue types each. The most popular type of venue is Convenience Store and the other venues are Argentinian Restaurant and Hotel Bar, one in each location.

The last cluster contains only one postal code with two venue types, Mexican Restaurant and Burrito Place.

4.3 Clustering Coyoacán

For Coyoacán we have 2 clusters, see Figure 7. The first cluster contains 78 out of the 97 postal codes in Coyoacán. The top venue types of this cluster are plotted in Figure 10. The most popular is Mexican Restaurant with Taco Place as a close second. Other popular venues are Coffee Shop, Ice Cream Shop, Restaurant.

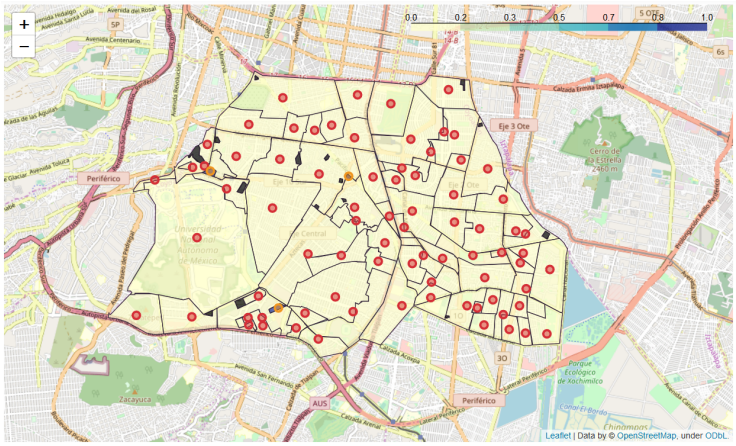


Figure 7: Clustered Coyoacán.

The second cluster contains only three locations. The rest of the locations have no venues left after filtering. On this second cluster, there is only one venue type, Taco Place. The locations in this cluster are all housing complexes.

5 Observations

After clustering each municipality we also tested them together for clustering but both tests suggested there were no clusters. That is also the broad result of this exercise, all three cases showed one big cluster and some smaller outlier clusters. But is in these outliers that we can find key points in the geographic distribution of Mexico City's *lifestyles*.

Mexico City's history of rapid growth is reflected in the layered distribution of these lifestyles. Cuauhtémoc contains Mexico City's main square spanning nine postal codes, which are all full of shops that sell a plethora of products. Looking at the top venue types we can see mostly food related venues, while all this variety of shops were left as not significant to the clustering algorithms implemented here. Unless the number of clusters was picked to be as the number of postal codes. All the shops in the first cluster of Cuauhtémoc are in the tables on the notebook.

Benito Juárez is right at the south of Cuauhtémoc and it shows a slight change in its top venue types. While the top venue types are quite similar, the relevance of each of the venues changes. In this municipality, we can find zones with many buildings with office spaces, expensive apartments, and shopping centers. These locations are usually surrounded by food places and that is reflected in the results. Note that there are almost as many Coffee Shops as Taco Places, this can lead us to think about the people that gather in these places being either residents or workers that spend some of their leisure time in the vicinity of their workplace.

Coyoacán is the southernmost of the three municipalities. The main clus-

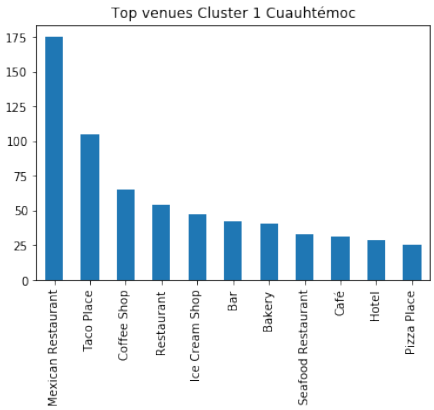


Figure 8: Top popular venue types for cluster 1 in Cuauhtémoc.

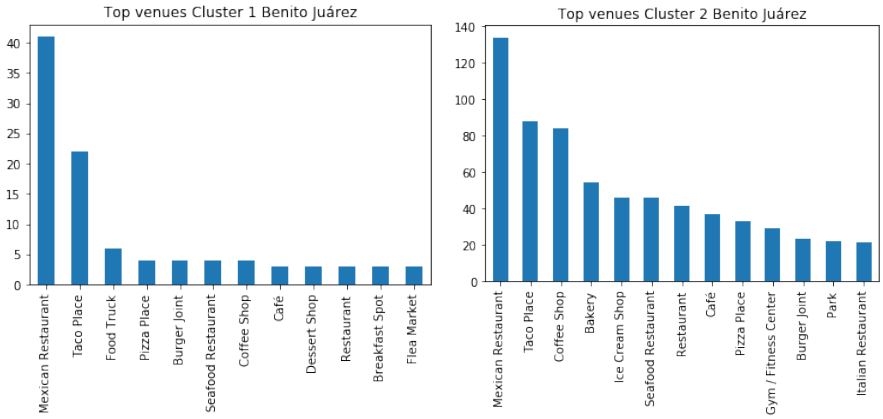


Figure 9: Top popular venue types for clusters 1 and 2 in Benito Juárez.

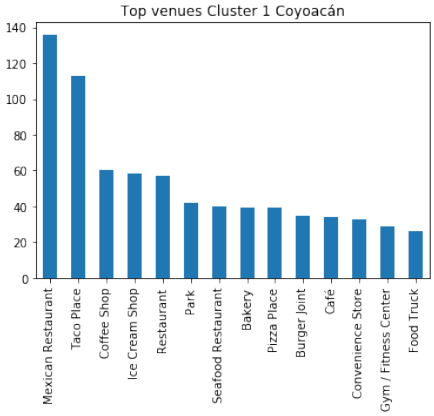


Figure 10: Top popular venue types for cluster 1 in Coyoacán.

ter is similar to the previous big clusters, but it shows venue types like Park, Convenience Store and Food Truck in the top popular. The second cluster of Coyoacán, the outlier, consists only of housing complexes, usually planned to be a solution for a housing problem for city workers during one of the many growth stages of Mexico City.

6 Conclusion

Mexico City's lifestyle as measured in this exercise appears to be homogeneous in regards to the venue types we can find in these three central municipalities. The results of this clustering exercise showed a glimpse of the multilayered history of growth of the city, having a central location full of diversity, a somewhat *business oriented* outer layer and an even further outer layer with hints of being once deemed as marginal. These three municipalities are just a small part of the city which nowadays extends beyond its boundaries into the neighboring State of Mexico. These motifs of centralization of the activities, followed by expansion and development to later lead to housing problems repeat all throughout the history of Mexico City. Any further insight would require an analysis beyond this exercise's scope.

The code for this exercise can be accessed at `https://github.com/jjhsaq/Clustering-Mexico-City/blob/master/Clustering%20CDMX.ipynb`.