Clustering Cdmx Neighborhoods

Aldo Cruz Hernandez.
Mexico City. July, 2020

Trying to find the best place to live or to locate a new business is defined by many factors as, the price, the surface size, the services, transportation system, places around it, etc. This time I will focus on discover what kind of places are most common in each Neighborhood of Mexico city. This information is going to help all people who is looking for a new place to rent or buying to decide which neighborhood fits better their needs.

The information also provides the required knowledge to locate a new business according with its category and make it stand out among others.

The goal of this project is to find what makes neighborhood to be similar to another and what makes them to be different. Once we know that, it is possible to group them and look the distribution they have in the city.

Data

For this project I will use data provided by the government site "Datos CDMX" (https://datos.cdmx.gob.mx/pages/home/) to get the neighborhood Names and the location (Latitude and Longitude) of each of them. Once we download the data, the next step is to deal with the missing values, which first, I will try to find the information using Nominatim from geopy. The remain missing data will be drop.

Once we get the Latitude and Longitude for the Neighborhoods, the foursquare request is going to be made and we will get information in JSON format (figure 1).

```
"venue": {
  "id": "49b6e8d2f964a52016531fe3",
  "name": "Russ & Daughters",
  "location": {
    "address": "179 E Houston St",
    "crossStreet": "btwn Allen & Orchard St",
    "lat": 40.72286707707289,
    "lng": -73.98829148466851,
    "labeledLatLngs": [
      {
        "label": "display",
        "lat": 40.72286707707289,
        "lng": -73.98829148466851
      }
    ],
    "distance": 130,
    "postalCode": "10002",
    "cc": "US",
    "city": "New York",
    "state": "NY",
    "country": "United States",
    "formattedAddress": [
      "179 E Houston St (btwn Allen & Orchard St)",
      "New York, NY 10002",
      "United States"
    ]
  },
  "categories": [
    {
      "id": "4bf58dd8d48988d1f5941735",
      "name": "Gourmet Shop",
      "pluralName": "Gourmet Shops",
      "shortName": "Gourmet",
      "icon": {
        "prefix": "https://ss3.4sqi.net/img/categories_v2/shops/food_gourmet_",
        "suffix": ".png"
      },
      "primary": true
```


Figure 1. Foursquare result in JSON format

The process of making requests will be perform for each Neighborhood. In order to work with the information it will be save in a data frame.

We do not need all of this information, the relevant information will be: the venue name, category venue and Neighborhood the venue is located. In order to work with the information, the categories will be transform into dummy values. Then we will group the mean values by neighborhood an sort them in a descending way, that will allows us to know the most frequent venues in each neighborhood.

Methodology

EXPLORATORY DATA

There are 463 unique venues distributed in 1759 neighborhoods which are distributed in 16 Cities (Alcaldias). Due the number of neighborhoods is too large to be seen in a chart, it is easier to see the venues distribution in each City (figure 2).

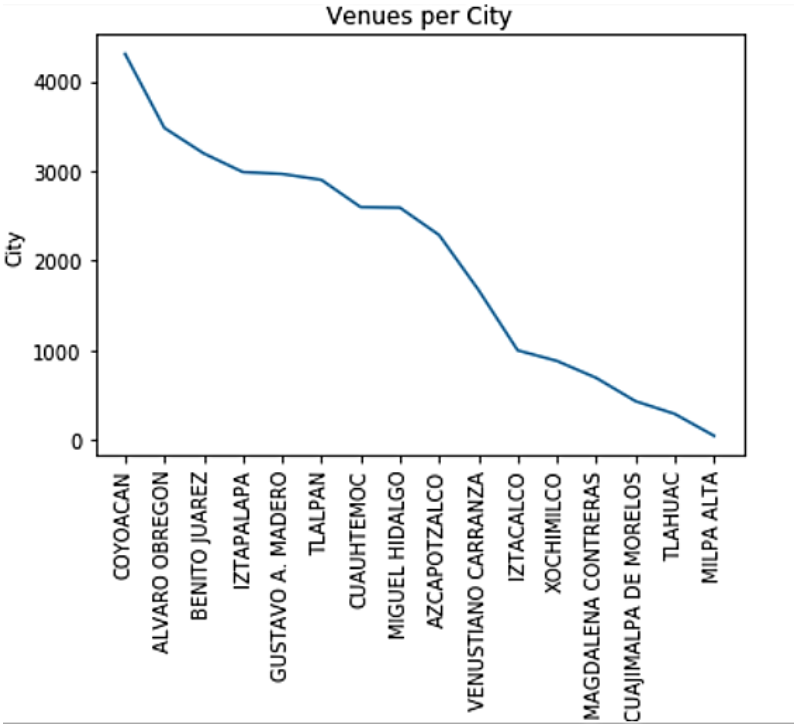

Figure 2. Venues Distribution.

The most popular venue is Mexican restaurant, which makes sense since it is a Mexican city. The second most popular venue, taco place, is very close to the first one. We can notice an abrupt drop in the other venues (figure 3)

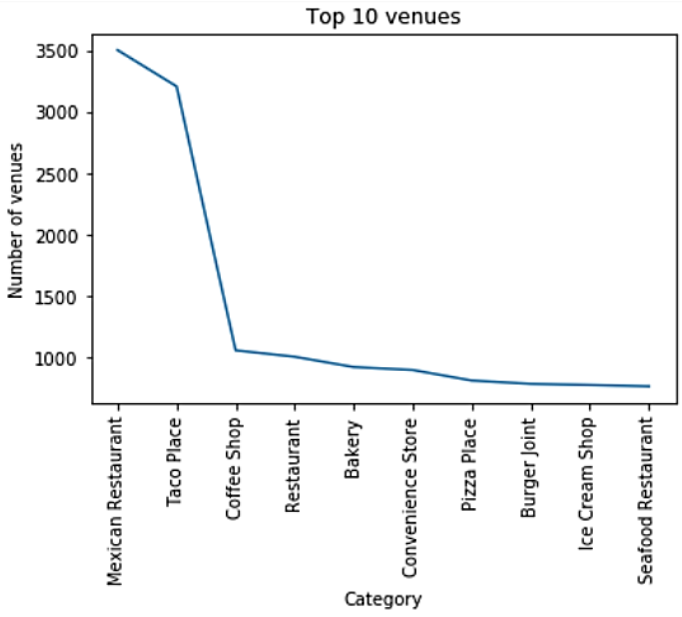

Figure 3. Most common venues.

MODEL SELECTION

The purpose of this project is to help someone or an organization to decide which neighborhood fits better their needs according with the venues in that neighborhood. To accomplish that goal, it is necessary to group the neighborhoods that are similar with each other and then, find underlying patterns.

Since we only have the input vectors without referring to known, or labelled outcomes, unsupervised algorithms are better for the data we are working with. For that reason, the machine learning method that will be used for this project is K-means clustering.

Select he Number of cluster is a subjective decision, for this process is important to consider that selecting a low number of clusters may result in under fitting, in other hand, a higher number of cluster may over fit the model. Another disadvantage of higher number of clusters is that it could be overwhelming for the user.

This time I will chose 7 clusters, it is enough to see differences between clusters and not over fit the model.

Results

This map (see figure 4) represent the 7 clusters.

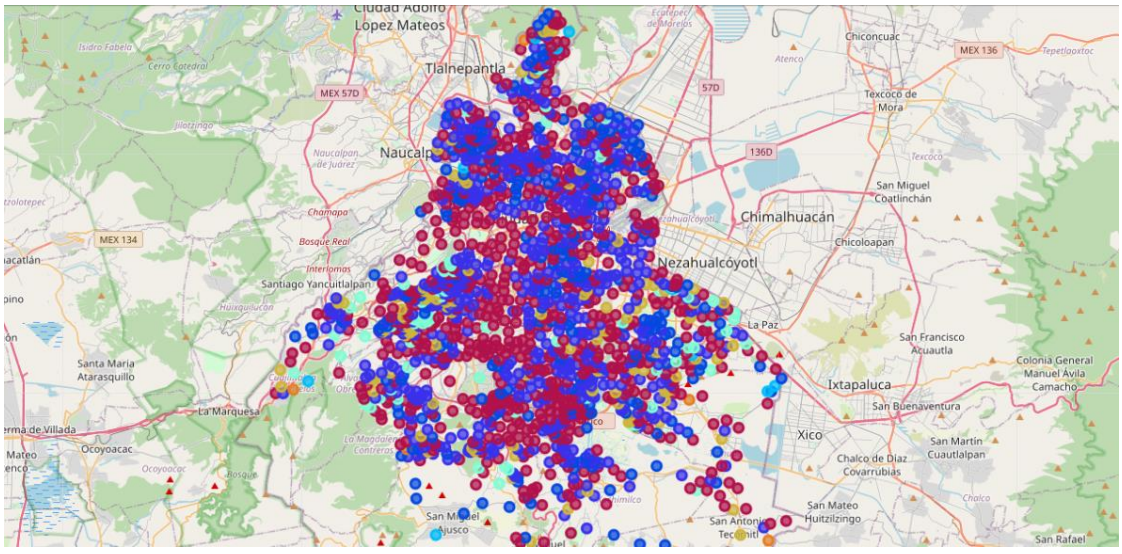

Figure 3. Cluster Map

CLUSTER 0

There are 707 Neighborhoods in this cluster, most of the neighborhoods are located in it (Figure 4), there is a great variety of places, for that reason is hard to say what venue represent this cluster. The neighborhoods are distributed along all the territory.

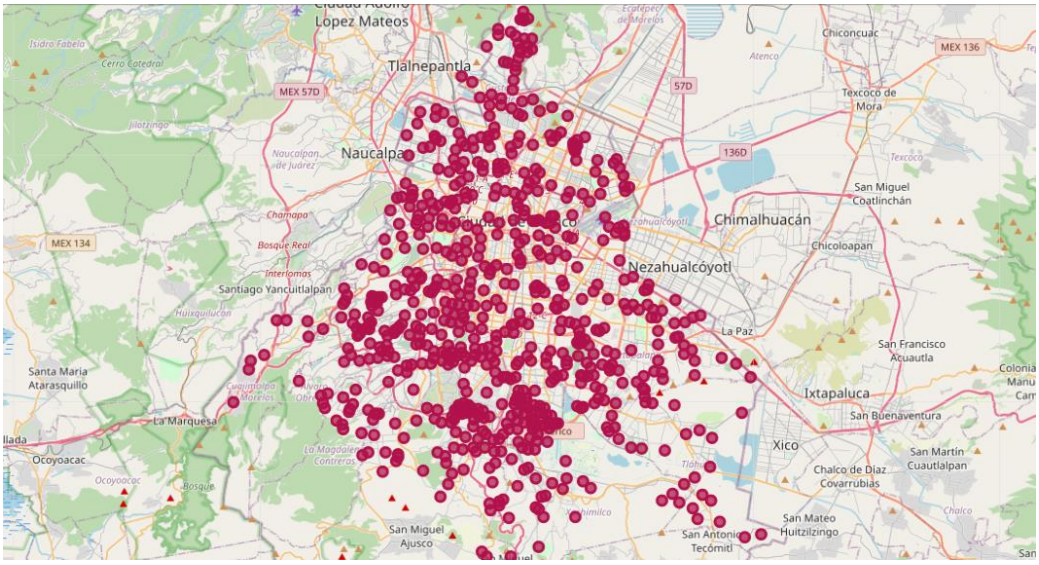

Figure 4. Cluster 0

CLUSTER 1

There are 593 Neighborhoods in this cluster (figure 5). This cluster present a variety of venues, but what stands out are different kinds of restaurants and places to have desserts.

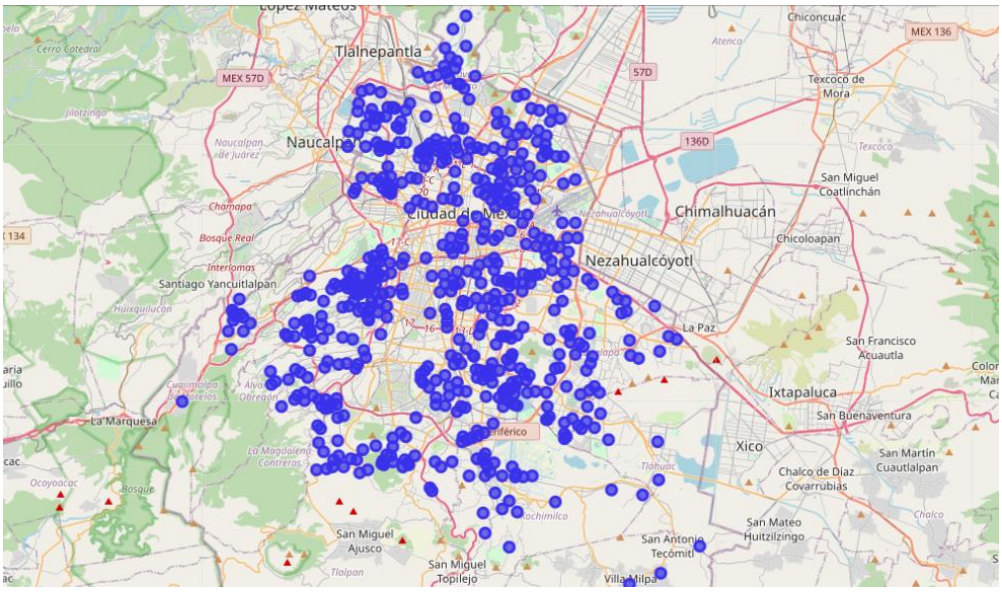

Figure 5. Cluster 1

CLUSTER 2

There are 250 Neighborhoods in this cluster. It is composed by places to eat, especially Mexican restaurants, coffee shops, pizza places, taco places. Cluster 2 locations are indicated blue dots (figure 6)

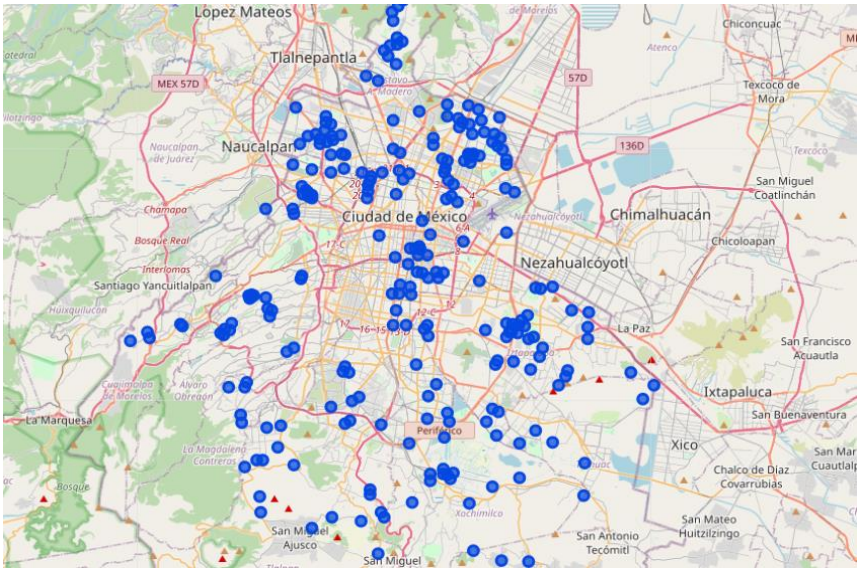

Figure 6. Cluster 2

CLUSTER 3

There are 11 Neighborhoods in this cluster. Cluster 3 is composed mainly by one venue, which is "park", and in some cases for another venue. This Neighborhoods are located in the city outsides. Cluster 3 locations are indicated with light blue dots (figure 7)

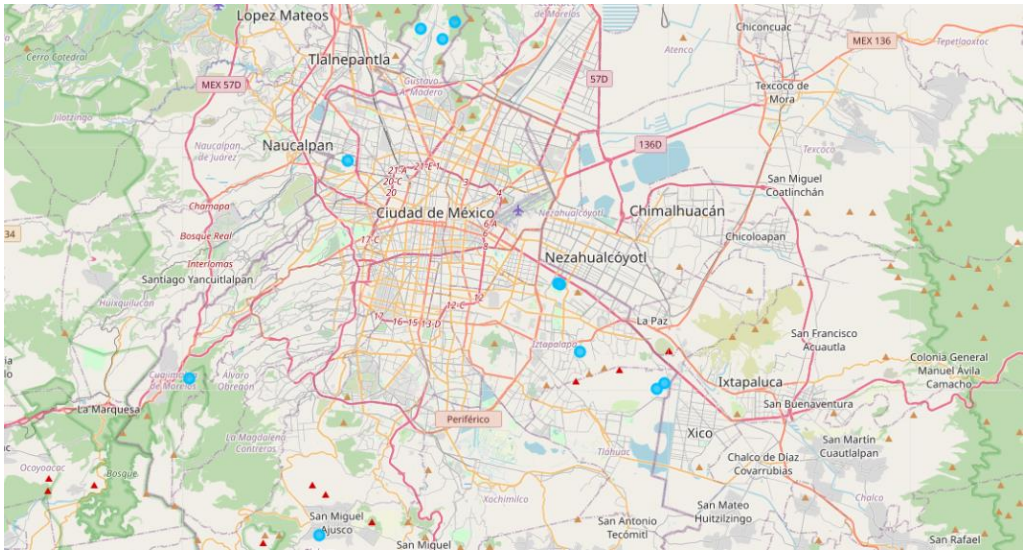

Figure 7 Cluster 3

CLUSTER 4

There are 81 Neighborhoods in this cluster. The main venues in this category are Fast food, coffee shops, and burger joints. It is important to notice that there are few venues per Neighborhood (around 5). Cluster 4 locations are indicated with cyan dots (figure 8)

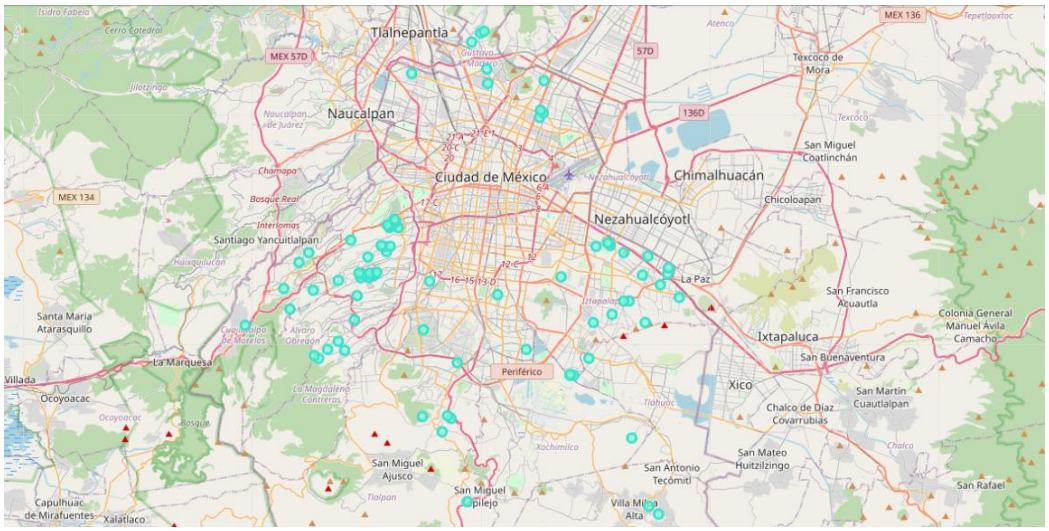

Figure 8 Cluster 4

CLUSTER 5

There are 112 neighborhoods in cluster 5. It is mainly composed by taco places. Cluster 5 locations are indicated with yellow dots (figure 9)

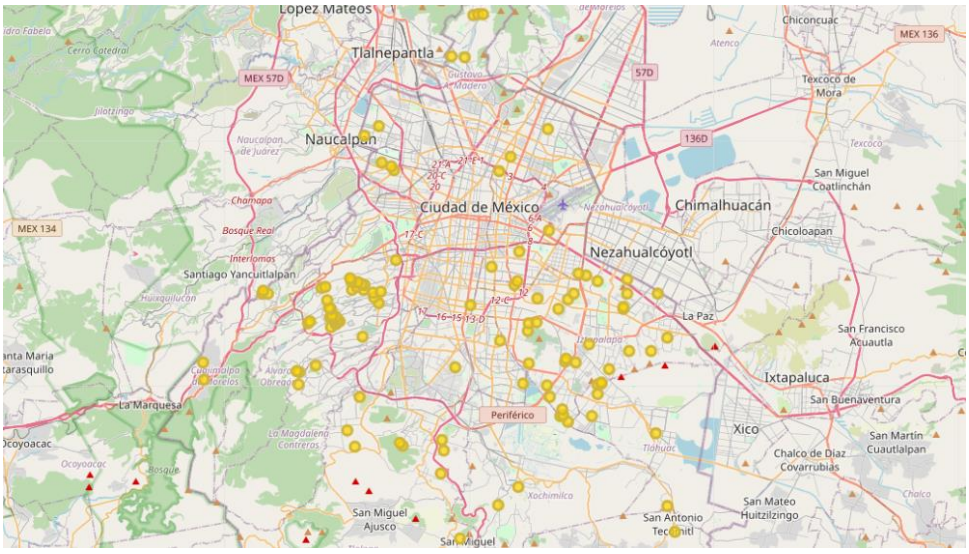

Figure 9 Cluster 5

CLUSTER 6

Cluster 6 is composed by mountains, there are just 5 Neighborhoods classified in this cluster. *(Figure 10).* This neighborhoods are located in the outsides Neighborhoods

ALCALDIA	Neighborhood	0	1	2	3	4	5	6	7	8	9
TLAHUAC	EMILIANO ZAPATA 1A	Mountain	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
TLALPAN	BOSQUES DEL PEDREGAL	Mountain	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
TLAHUAC	SANTIAGO ZAPOTITLAN (PBLO)	Mountain	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
CUAJIMALPA DE MORELOS	CRUZ BLANCA	Mountain	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
GUSTAVO A. MADERO	CHALMA DE GUADALUPE I	BBQ Joint	Mountain	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure 10 Cluster 6 neighborhoods

The cluster 6 locations are indicated with orange dots (figure 11)

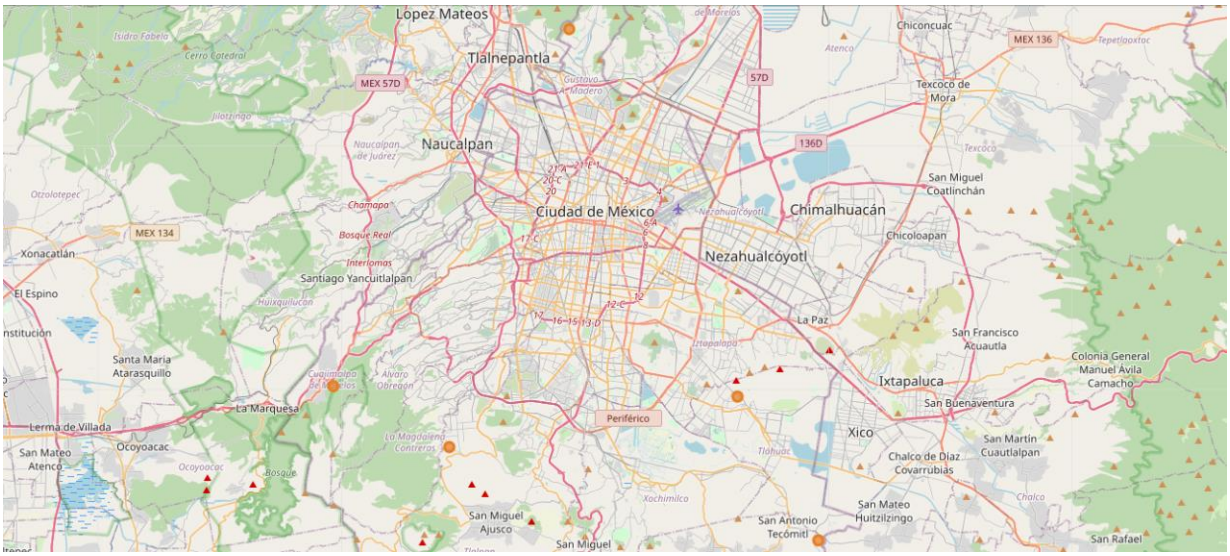

Figure 11 Cluster 6

Discussion

The first two clusters (0 and 1) seem to be under fitted due they have a great variety of venues, the only venues that stand out are places to eat, nevertheless this result is expected since the most common venue in the whole city is "restaurant". This outcome could be fixed by adding more clusters.

The other clusters present greater differences between each other, which is what we was looking for

Conclusion.

Cluster 0 is the most congested, it is also the cluster with more neighborhoods. In those places you can find almost everything you are looking for. It could be improve the segmentation adding more clusters to the model, but right now is very ambiguous.

Cluster 1 is defined by places you can eat everything you want, from the typical Mexican food to foreign restaurants. It is located in crowded places, which make a good idea if you are looking for clients, the disadvantage is that is harder to stand out among the competency.

Cluster 2 is represented by Mexican restaurants and other places to eat. These Neighborhoods are located near each other.

Cluster 5 is mainly composed by taco places, if you are planning to sell this kind of food, you better avoid those neighborhoods. Otherwise, if what you want is to find different varieties of this typical food, I highly recommend you to go and visit them.

The clusters whit less venues are cluster 6 and 3, it is easy to notice what kind of neighborhood they are.

For the number 3; it is composed by parks. If you are planning to do and activity that fist better with nature and big spaces, you should be interested in this cluster. The main disadvantage in this neighborhood are that is harder to access to them due they are located in the city outsides, another important disadvantage the lack of other venues, be prepared to travel longer distances if you want a different activity to do.

Cluster 6 is represented by mountains, there are located in the city outsides and there is nothing more but nature.