Project report: Mexico City neighborhood analysis for housing developing.

Submitted by: Adrián Rosales.

Business Case: Find suitable place for construction companies to establish new housing developments.

Summary

Mexico City is an enormous urban agglomeration which has a severe problem of overpopulation concentrated in few areas, construction companies are always looking for new land to develop new housing developments.

Strategy:

On this project we will be using clustering to determine which areas are suitable for new housing developments based on the amount and type of venues that each of them has.

Introduction

As part of the Mexican Valley Metropolitan Zone, Mexico City is the 9th biggest urban agglomeration across the world, and the most populous city in North America. It has almost 8.9 million people living in a land area of 1,485 m^2 with a density of 6,000/km^2, this density reveals the city's biggest problem which is the lack of residence for new inhabitants.

Construction companies have been developing new housing buildings across the city, but the demand is still growing, in order to check for new constructions lands, it is important to know which neighborhoods have a bigger number of preestablished venues and which don't. For those neighborhoods that have bigger number of preestablished venues, companies only need to construct residential building; and for those that don't, companies are looking for constructing multifunctional complexes that includes residential apartments, office section and mall. In order to help companies and government deciding where to establish new buildings it is important to know all neighborhood areas background; one approximation could be using data science to cluster the neighborhoods into 3 different categories.

Data

Requirements

- List of neighborhoods in Mexico City (CDMX) and their coordinates location.
- List of venues for a specific neighborhood area.

Data sources, processing and tools used

- To obtain the list of neighborhoods, we will be using a dataset from the local government which are free and public, and it is available on this link: https://datos.cdmx.gob.mx/explore/dataset/coloniascdmx/table/?dataChart=eyJxdWVyaWVzIjpbeyJjb25maWciOnsiZGF0YXNldCI6ImNvbG9uaWFzY2RteCIsIm9wdGlvbnMiOnt9fSwiY2hhcnRzIjpbeyJhbGlnbk1vbnRoIjp0cnVlLCJ0eXBlIjoiY29sdW1uIiwiZnVuYyI6IkFWRyIsInlBeGlzIjoiZW50aWRhZCIsInNjaWVudGlmaWNEaXNwbGF5Ijp0cnVlLCJjb2xvciI6IiM2NmMyYTUifV0sInhBeGlzIjoibm9tYnJlIiwibWF4cG9pbnRzIjo1MCwic29ydCI6IiJ9XSwidGltZXNjYWxlIjoiIiwiZGlzcGxheUxlZ2VuZCI6dHJ1ZSwiYWxpZ25Nb250aCI6dHJ1ZX0%3D
- To proccess the dataset we will need to use Foursquare API and get venues near 1 km around the neighborhoods, the top ten of venues categories, transform into one hot encoding and merge them with the previous dataset.
- There are 1,812 neighborhoods in CDMX, we will add a column to count the total number of venues for each neighborhood.

Methodology

Every stage to build a data science project is important to get the final goal, but for me the key to achieve good results is the data analysis and processing. On This project the 3 main steps were to acquire, analyze and preprocess the data (these ones are related that is why they count as one) and finally the clustering.

Acquire the Data

Web searching were made before web scrapping, finding a suitable dataset for our purpose which contains geographic data (coordinates) from Mexico City (CDMX) neighborhoods, on a csv file. This made easier this stage of the project.

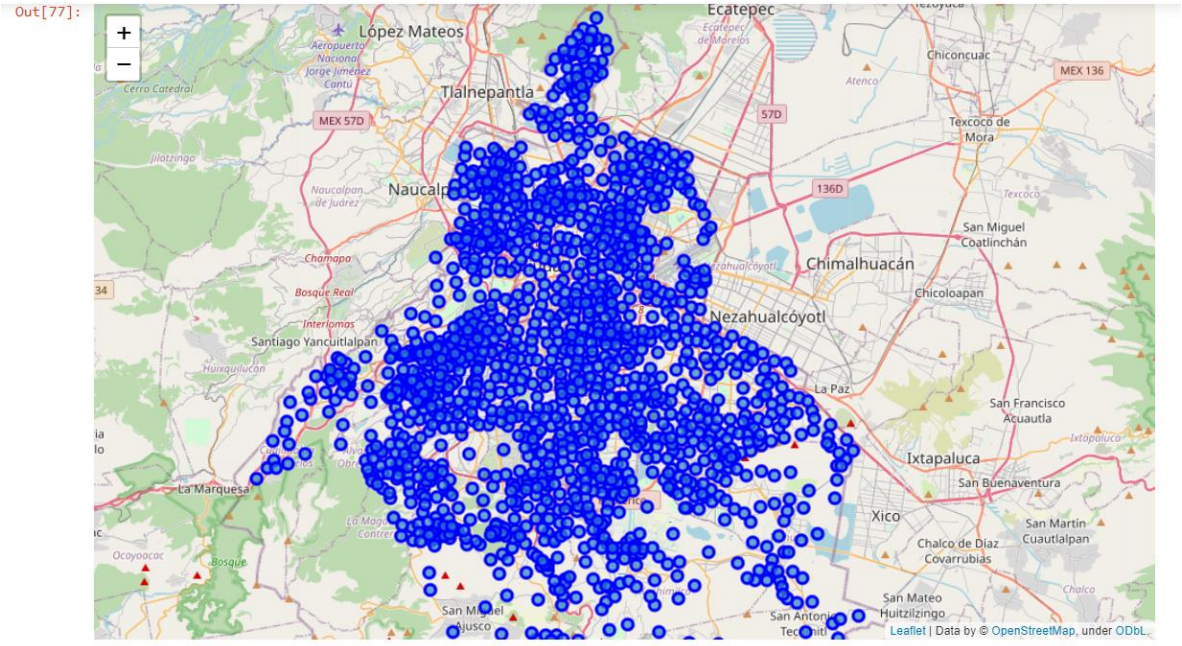
After eliminating unnecessary columns, we import the dataset into a dataframe, looking like this:

Out[15]:

	COLONIA	ALCALDIA	Latitude	Longitude
0	IRRIGACION	MIGUEL HIDALGO	19.4429549298	-99.2099357048
1	MARINA NACIONAL (U HAB)	MIGUEL HIDALGO	19.4466319056	-99.1795110575
2	PEDREGAL DE STO DOMINGO VI	COYOACAN	19.3234027183	-99.1654676133
3	VILLA PANAMERICANA 7MA. SECCIN (U HAB)	COYOACAN	19.304604269	-99.1677617231
4	VILLA PANAMERICANA 6TA. SECCIN (U HAB)	COYOACAN	19.3112238873	-99.1696478642

Using this dataframe, I plot a folium map to get an idea of how neighborhoods are distributed into the city.

Out[77]:



As we may see, this city is really crowded, it has 1812 neighborhoods distributed into 9 boroughs.

Data Analysis and preprocessing

First thing that noticed is that, there already housing developments which are neighborhoods and contains a label "(U HAB)", we eliminated rows containing that string in their neighborhoods' name. Another thing that I noticed was that there were neighborhoods duplicated so, I decided to drop those rows by using pandas function *"drop_duplicates()"*, so after that we use only 1489 rows in our Dataframe.

The main approach of this project to achieve the goal, were by using venues to determine how similar are the neighborhoods and by the number of them. In order to get venues near a specific area, we use Foursquare API and request venues near 1 km around. Getting a new dataframe with 34,388 venues, those venues were grouped with the neighborhoods using its name as a key.

```
In [100]: print(mex_venues.shape)
          mex_venues.head()
```

```
(34388, 7)
```

Out[100]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	IRRIGACION	19.4429549298	-99.2099357048	La Pantera Fresca	19.442104	-99.209314	Ice Cream Shop
1	IRRIGACION	19.4429549298	-99.2099357048	Glorieta de Fátima	19.444397	-99.210792	Park
2	IRRIGACION	19.4429549298	-99.2099357048	Panmex	19.440615	-99.208584	Bakery
3	IRRIGACION	19.4429549298	-99.2099357048	Canchas de Tenis	19.440700	-99.211627	Tennis Court
4	IRRIGACION	19.4429549298	-99.2099357048	Salchichonería San Miguel	19.440234	-99.207833	Deli / Bodega

We use one hot encoding paradigm to stablish the frequency of venues category into a dataframe, and then we add a new column which contains the total number of venues.

```
In [138]: mex_merged.rename(columns={'Venue':'Count'}, inplace=True)
          print(mex_merged.shape)
          mex_merged.head()
```

```
(1457, 466)
```

Out[138]:

	Neighborhood	Count	Zoo Exhibit	ATM	Accessories Store	Adult Boutique	Advertising Agency	African Restaurant	Airport	Airport Service	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentini Restaura
0	10 DE ABRIL	7	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1	10 DE MAYO	20	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
2	12 DE DICIEMBRE	15	0.066667	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
3	15 DE AGOSTO	5	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
4	16 DE SEPTIEMBRE	17	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

If we observe, some rows were missed when we request venues to foursquare, this is because there was not available data, so in order to not affect the final clustering, we add those missing rows by establish the count and the categories frequency as 0.

```
In [140]: print(mex_merged.shape)
          mex_merged.tail()
```

```
(1489, 466)
```

Out[140]:

Neighborhood	Count	Zoo Exhibit	ATM	Accessories Store	Adult Boutique	Advertising Agency	African Restaurant	Airport	Airport Service	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	C
10 DE ABRIL	7	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
10 DE MAYO	20	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
12 DE DICIEMBRE	15	0.066667	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
15 DE AGOSTO	5	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
16 DE SEPTIEMBRE	17	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

Clustering

On this stage we decided to cluster the neighborhoods into 3 categories by using kmeans algorithm, using the scikit-learn library implementation.

Results

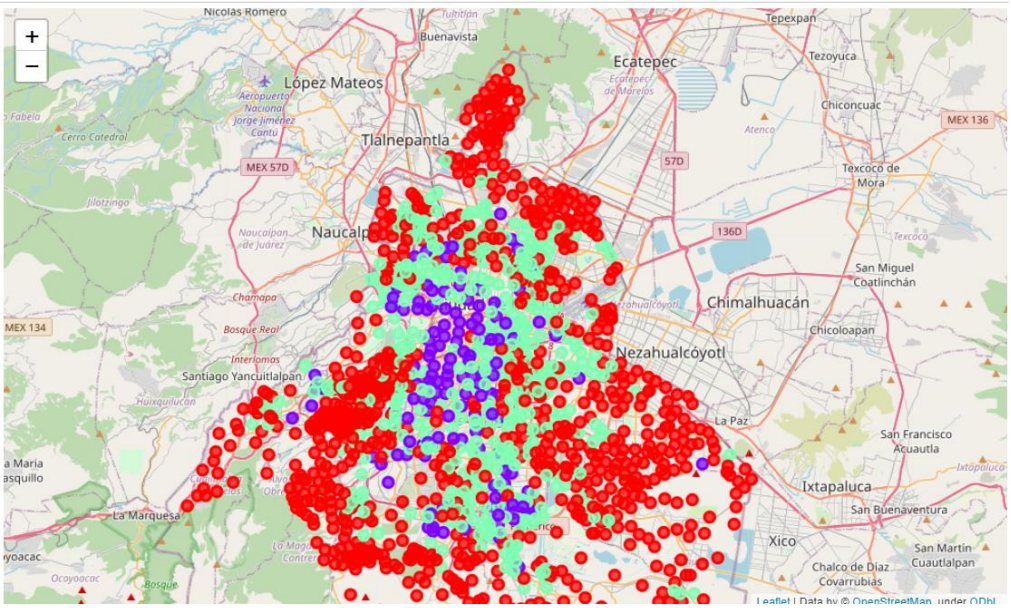
The 3 cluster labels were distributed like this:

```
In [151]: mex_merged2['Cluster Labels'].value_counts()

Out[151]: 0    897
          2    456
          1    136
          Name: Cluster Labels, dtype: int64
```

And by merging the labels with their respected neighborhood and coordinates, we proceed to plot a new folium map.

Out[154]:



Purple = Cluster Label ‘0’

Green = Cluster Label ‘1’

Red = Cluster Label ‘2’

	Neighborhood	ALCALDIA	Latitude	Longitude	Cluster Labels	Count	Zoo Exhibit	ATM	Accessories Store	Adult Boutique	Advertising Agency	African Restaurant	Airport	A Se
0	IRRIGACION	MIGUEL HIDALGO	19.4429549298	-99.2099357048	2	42	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
2	PEDREGAL DE STO DOMINGO VI	COYOACAN	19.3234027183	-99.1654676133	0	19	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
5	SANTA CRUZ AVIACION	VENUSTIANO CARRANZA	19.4223039522	-99.0972147602	2	36	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
6	MAGDALENA MIXHUCA	VENUSTIANO CARRANZA	19.407130147	-99.1181313602	2	28	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
7	COPILCO UNIVERSIDAD	COYOACAN	19.3361172512	-99.1826102146	1	73	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

Discussion

As we may observe there are 3 different types of neighborhoods, the ones that are close to the borders are colored with red, it is obvious that those areas are cheaper lands and have limited number of venues. But the interesting thing occurs analyzing ‘0’ (purple areas) and ‘1’ (green areas) labels, having the background of living in this city and observing some of the pop well know neighborhoods labeled with 0 and 1, the 0’s are for neighborhoods that has certain commodities but are not crowded, this includes exclusive zones for the richest people. Areas labeled with ‘1’ are populated zones which has lot of offices buildings and venues.

Conclusion

Taking into consideration the previous discussion, it is obvious that if a construction company is looking for cheaper lands they need to locate their buildings into a red zone, but if a company wants to make better profits it needs to concentrate its effort in locate areas in green and red zones, the purple zones are suitable for multifunctional complexes because of their lack of venues and are far from crowded zones and the green ones are suitable just for housing building they are already crowd with offices.

Future work

We could analyze the results of working with a classification paradigm looking for some well know neighborhoods, famous by their good quality of life or by its poor quality an using an expert to label these neighborhoods in order to obtain a train set which could lead us to determine every other neighborhood label into the city and the metropolitan area that includes another boroughs that were not listed in this project.