Housing Sales Prices & Venues Data Analysis of Mexico City

Jesús Alfonso Juárez

February 2020

1.- Introduction

1a.-Description and discussion of the background

Greater Mexico City is the second largest metropolitan area of the western hemisphere and the largest spanish-speaking city in the world with **21.3** million of population. Mexico City has by itself **9 million** people gathered in just **1,485 square kilometers** turning into a high-density zone with **6,000 persons** by square kilometer. Mexico has the history embedded in their walls, originally named **Mexico Tenochitlan** by the aztecs has been witness of many stages from the pre-Hispanic to the modern era. Currently, the city is formed by **16 boroughs.[1]**

Mexico City is considered a megacity which means that is a high population density zone. Thus, there is a restricted supply of commercial and residential real estate. Moreover, the tendency to the vertical urbanization and the new structures of families demand a new approach in the housing sector. The city residents are seeking zones near to their jobs, with the venues that they attend and where the real estate values are lower, and Furthermore, investors are seeking to establish business in the neighborhoods with lower cost and less competition in the district.

Nowadays does not exist a tool that lead investors and city residents to make a data-based decision of the neighborhood to select. Consequently, we can create a map and information chart where the real estate index is placed on Mexico City and each district is clustered according to the venue density.

1b.-Data Description

To solve the problem, we can list the data as below:

- I found the Boroughs Coordinates of Mexico City in the Data Repository of the Mexico City government website [2]. The. geojson has the coordinates of all the districts and boroughs ('Delegaciones') of Mexico City
- Furthermore, I found 2 excel files, that contain the information for the latitude and longitude of Boroughs and Neighborhoods in Mexico City from the Data Repository of the Mexico City government website [2].
- I used Forsquare API to get the most common venues of given Borough ('Delegación') of Mexico City [3].
- The real estate as other markets has a widespread range of prices in similar housing, thus there is a myriad of information regarding the real estate costs. To overcome this issue, we are going to use the latest square meter Housing Sales Price (HSP)

Average for each Borough ('Delegación') of Mexico City retrieve from the real state retail web page [4].

1c.-Data Usage

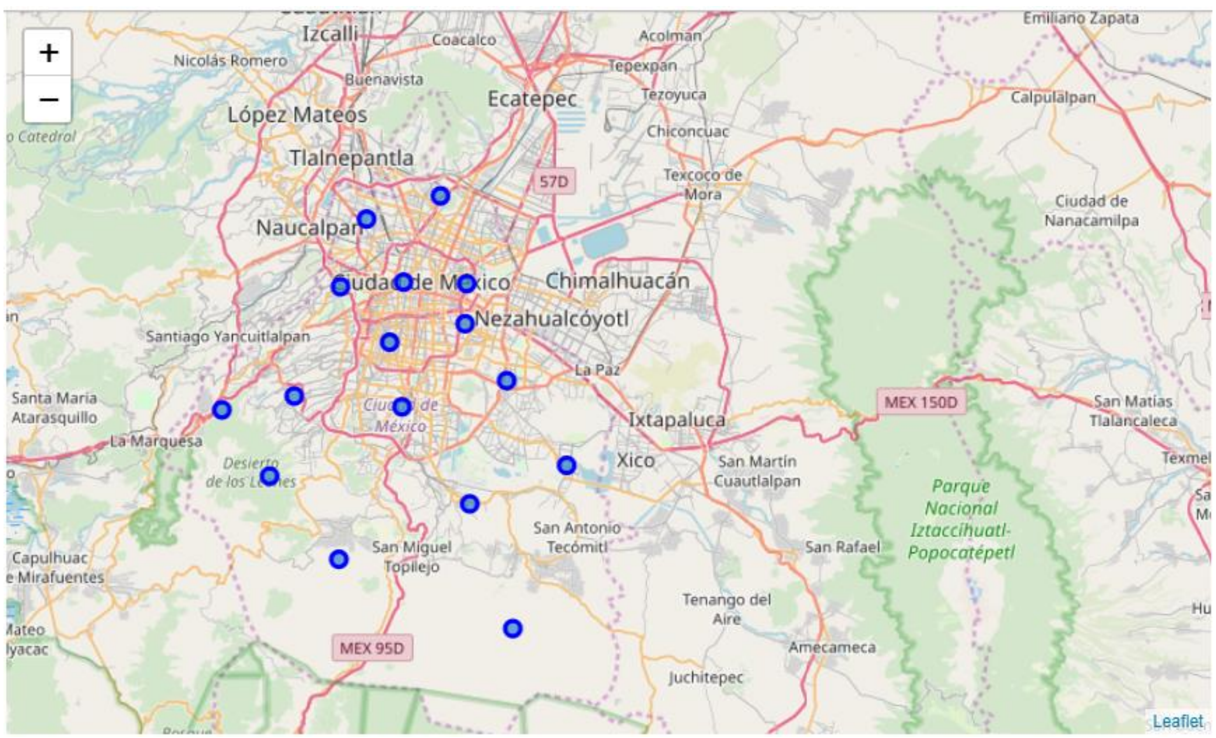
The approach that we are going to take with the different sources of information is:

1.-The geojson file will be uploaded to the Jupyter notebook file.

Example in Excel of the Boroughs' coordinates:

	NOMBRE	CLAVE_MUNICIPAL	CVE_ENTIDAD	CVEGEO	Geo Point	Geo Shape
0	Cuauhtémoc	15	9	9015	19.4313734294, -99.1490557562	{"type": "Polygon", "coordinates": [[[-99.1291...
1	Álvaro Obregón	10	9	9010	19.336175562, -99.246819712	{"type": "Polygon", "coordinates": [[[-99.1887...
2	Xochimilco	13	9	9013	19.2451450458, -99.0903636045	{"type": "Polygon", "coordinates": [[[-99.0986...
3	Tláhuac	11	9	9011	19.2769983772, -99.0028216137	{"type": "Polygon", "coordinates": [[[-98.9789...
4	Benito Juárez	14	9	9014	19.3806424162, -99.1611346584	{"type": "Polygon", "coordinates": [[[-99.1367...

Here it is a map with the center of each Borough.



2.-The information retrieve of Metroscubicos website will be compiled in a csv file then uploaded to the Jupyter notebook file.

	Borough	House Alone	Department	House Condominium	Average Square Meter
0	Álvaro Obregón	26003.00	31371.95	25878.55	27751.166667
1	Azcapotzalco	12688.96	17036.16	14084.18	14603.100000
2	Benito Juárez	25097.83	29594.46	25611.02	26767.770000
3	Coyoacán	20755.53	24808.50	23481.23	23015.086667
4	Cuajimalpa de Morelos	26128.21	36919.30	25100.27	29382.593333

3.- The gejson file and the excel files are going to pass through a cleansing stage in order to homologue the Boroughs names, and other fields.

4.- Afterwards, we are going to set a panda's data frame with the neighborhood name, coordinates, Borough, average square meter.

	Neighborhood	CVE_ALC	Borough	Average Square Meter	Lat Center	Lon Center	Latitude	Longitude
0	IRRIGACION	15	MIGUEL HIDALGO	40772.71	19.428062	-99.204567	19.4429549298	-99.2099357048
1	MARINA NACIONAL (U HAB)	15	MIGUEL HIDALGO	40772.71	19.428062	-99.204567	19.4466319056	-99.1795110575
2	MORALES SECCION ALAMEDA (POLANCO)	15	MIGUEL HIDALGO	40772.71	19.428062	-99.204567	19.4337174017	-99.2048231931
3	TORRE BLANCA (AMPL)	15	MIGUEL HIDALGO	40772.71	19.428062	-99.204567	19.454722061	-99.1998072368
4	ARGENTINA ANTIGUA	15	MIGUEL HIDALGO	40772.71	19.428062	-99.204567	19.4555189573	-99.2070212923

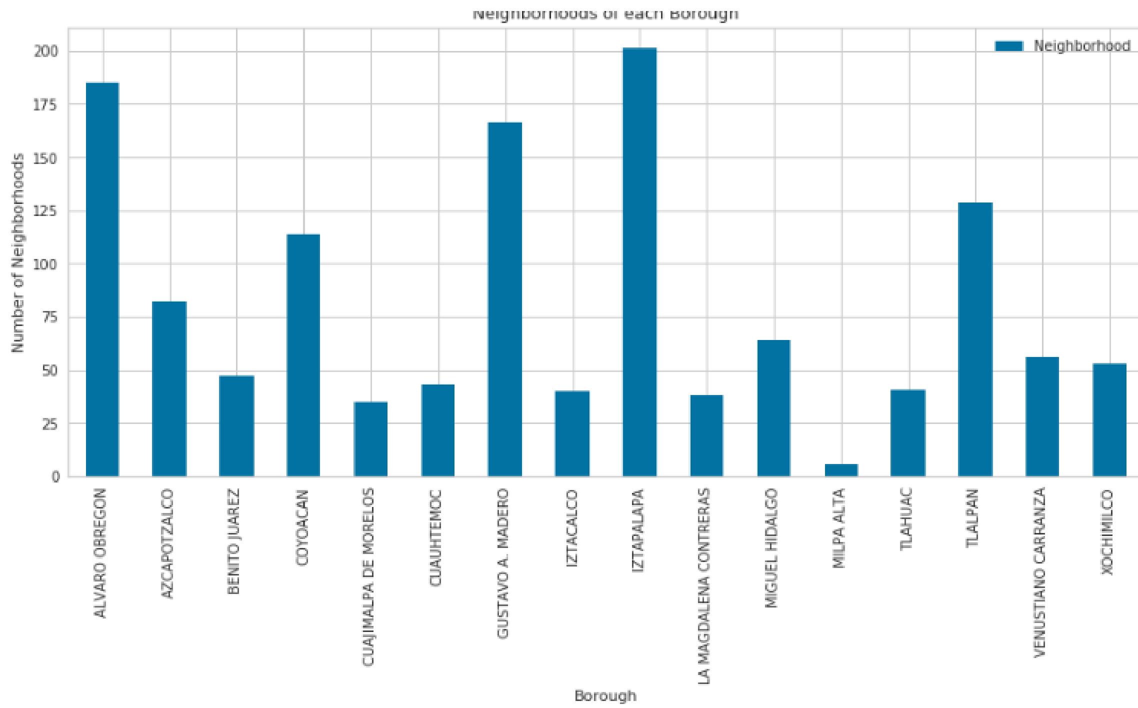
.

5.- Leveraging the foursquare API we are going to retrieve the closest venues to each district in a radio of 500 meters.

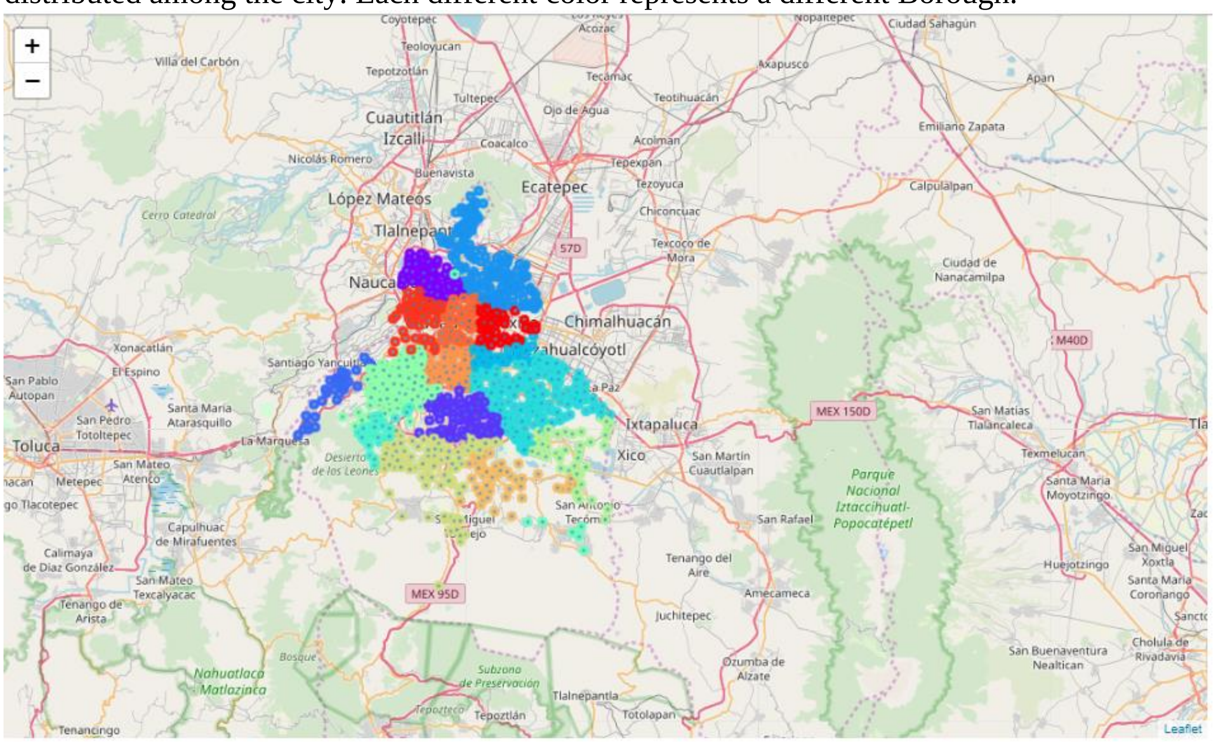
6.- Finally, we are going to transform this last panda's data frame, establishing the venues categories as columns with the get_dummies method and grouping by the neighborhood. The resultant data frame will be our input for the k-mean cluster method.

2.- Methodology

First of all, it was made an exploratory analysis of the distribution of the Neighborhoods (1800) in the Boroughs (16)



Subsequently, a folium map was built to view graphically how the neighborhoods are distributed among the city. Each different color represents a different Borough.



As result of this first approach, the scope of the study was limit to a sample of 12.5% neighborhoods of the Mexico City (225 of 1800). The main reason is the limit of calls that the Foursquare API set us.

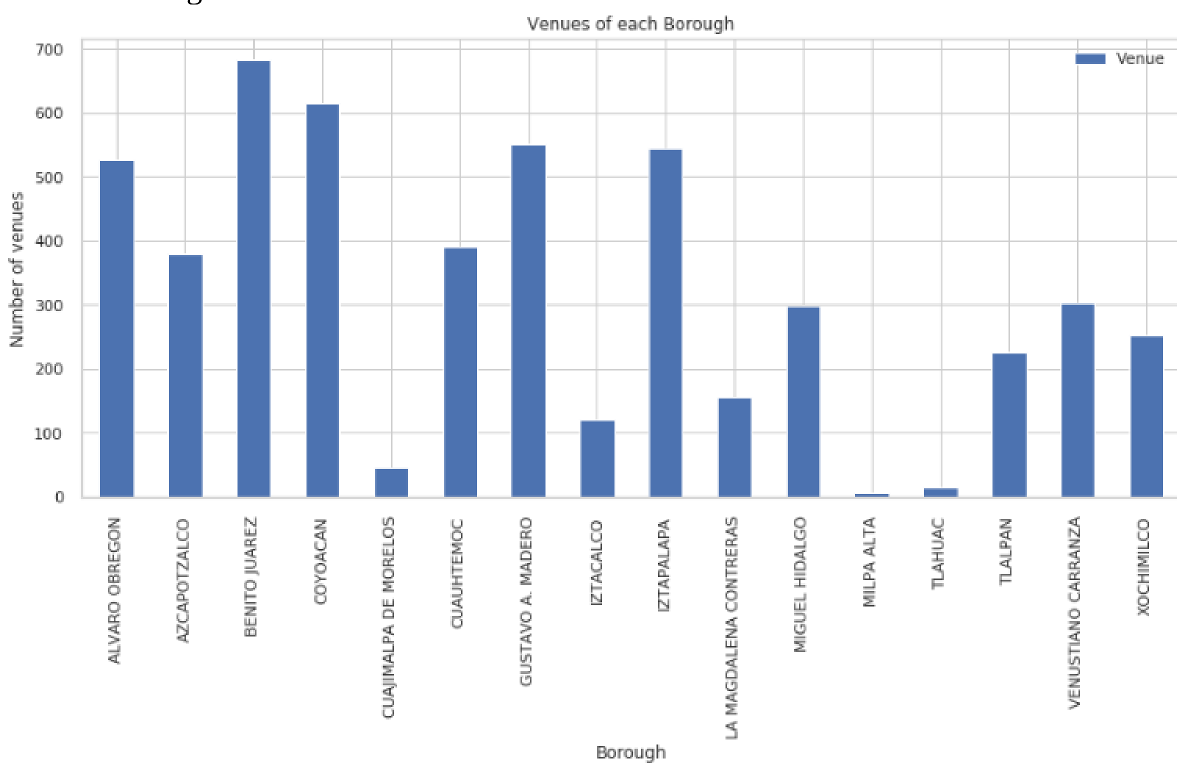
	Neighborhood	CVE_ALC	Borough	Average Square Meter	Lat Center	Lon Center	Latitude	Longitude
779	TECACALANCO	12	XOCHIMILCO	15852.443333	19.245145	-99.090364	19.240526	-99.063652
1171	EL PIRU (FRACC)	9	ALVARO OBREGON	27751.166667	19.336176	-99.246820	19.380379	-99.217964
280	ARTES GRAFICAS	16	VENUSTIANO CARRANZA	11776.340000	19.430495	-99.093106	19.411346	-99.125870
760	SAN LORENZO LA CEBADA II	12	XOCHIMILCO	15852.443333	19.245145	-99.090364	19.279520	-99.120590
1210	LA ANGOSTURA	9	ALVARO OBREGON	27751.166667	19.336176	-99.246820	19.333006	-99.232437

The approach that we have establish to solve the problem is to retrieve the nearby venues in a range of 500 m from center of each Neighborhood and limit to 100 calls per Neighborhood.

[116]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	EL PIRU (FRACC)	19.380379	-99.217964	Feli Pizzas	19.380835	-99.214226	Pizza Place
1	EL PIRU (FRACC)	19.380379	-99.217964	Taqueria El Guero	19.377413	-99.219273	Taco Place
2	EL PIRU (FRACC)	19.380379	-99.217964	Alitas Bbq	19.378107	-99.217899	Wings Joint
3	EL PIRU (FRACC)	19.380379	-99.217964	Mercado De Los Domingos (Capula)	19.379685	-99.215438	Market
4	ARTES GRAFICAS	19.411346	-99.125870	El Huarache De Jamaica	19.409581	-99.124144	Mexican Restaurant

Subsequently, it was made an analysis of the distribution of the retrieved venues in the different Boroughs.



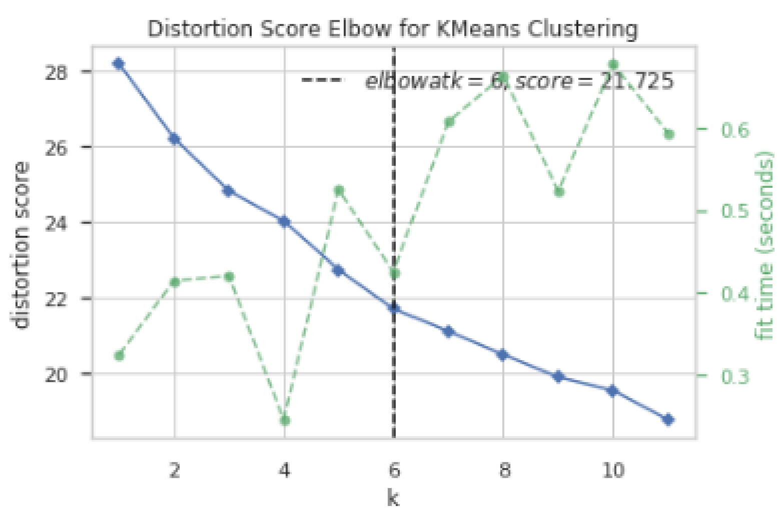
Third step in our analysis will be calculation of the top 10 most repeat it venues in each neighborhood, after it will transform with dummies values in order to train the Machine Learning model.

	Neighborhood	Zoo	ATM	Accessories Store	African Restaurant	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	...	Warehouse Store	Water Park	Waterfall	Whisky Bar	Wine Bar	Winery	Wings Joint
0	2A AMPLIACION SANTIAGO ACAHUALTEPEC I	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	2A AMPLIACION SANTIAGO ACAHUALTEPEC II	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	6 DE JUNIO	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	7 DE JULIO	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	ABRAHAM GONZALEZ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0

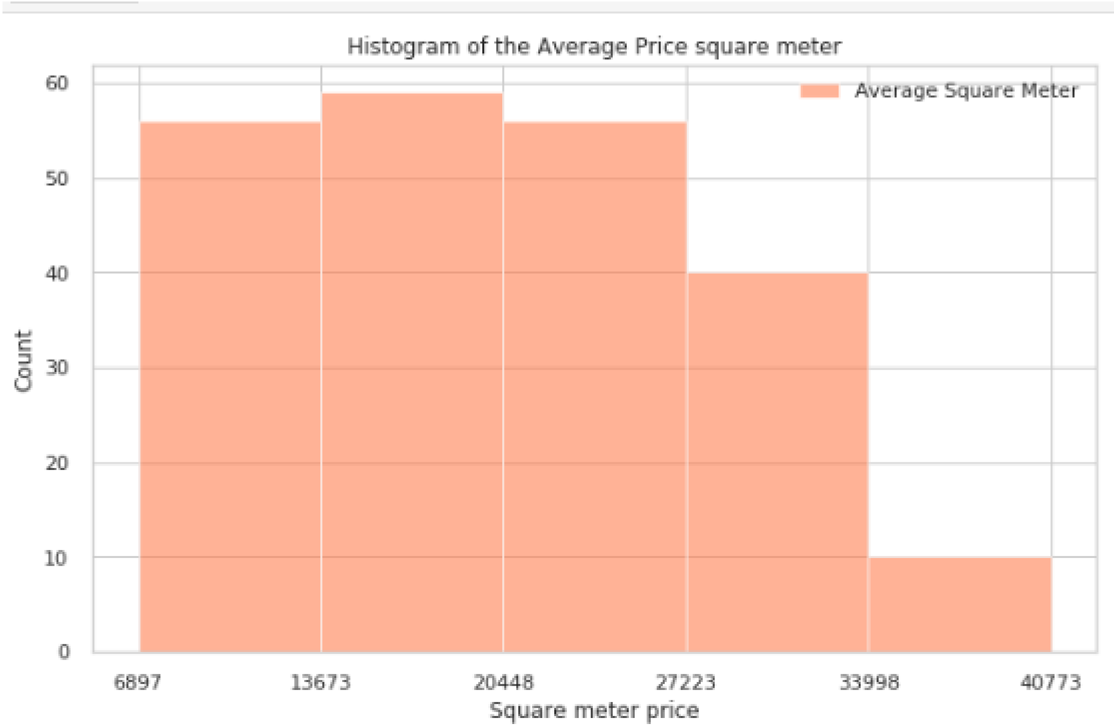
Subsequently, the data frame of the Dummies of the top 10 venues per neighborhood will be our raw material to train our model.

	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	2A AMPLIACION SANTIAGO ACAHUALTEPEC I	Bar	Skate Park	Seafood Restaurant	Gym	BBQ Joint	Pharmacy	Fast Food Restaurant	Sandwich Place	Taco Place	Chinese Restaurant
1	2A AMPLIACION SANTIAGO ACAHUALTEPEC II	Convenience Store	Soccer Stadium	Shopping Mall	Gym	Coffee Shop	Taco Place	Health & Beauty Service	Farmers Market	Burger Joint	Fast Food Restaurant
2	6 DE JUNIO	Moving Target	Shopping Mall	Food Truck	Park	Farm	Event Service	Event Space	Exhibit	Fabric Shop	Factory
3	7 DE JULIO	Pizza Place	Restaurant	Café	BBQ Joint	Diner	Rental Car Location	Coffee Shop	Donut Shop	Salad Place	Empanada Restaurant
4	ABRAHAM GONZALEZ	Mexican Restaurant	Taco Place	Burger Joint	Farmers Market	Bar	Restaurant	General Entertainment	Flea Market	Food Truck	Housing Development

The chosen method to work is the K-means algorithm, the reason is that we are working with unlabeled categories, and since we are trying to classify our Neighborhoods with similar venues Cluster Means is the best approach.
To train the model we need to optimize the cluster number, this is the reason we use the Elbow Distortion Score to identify the best number of clusters.

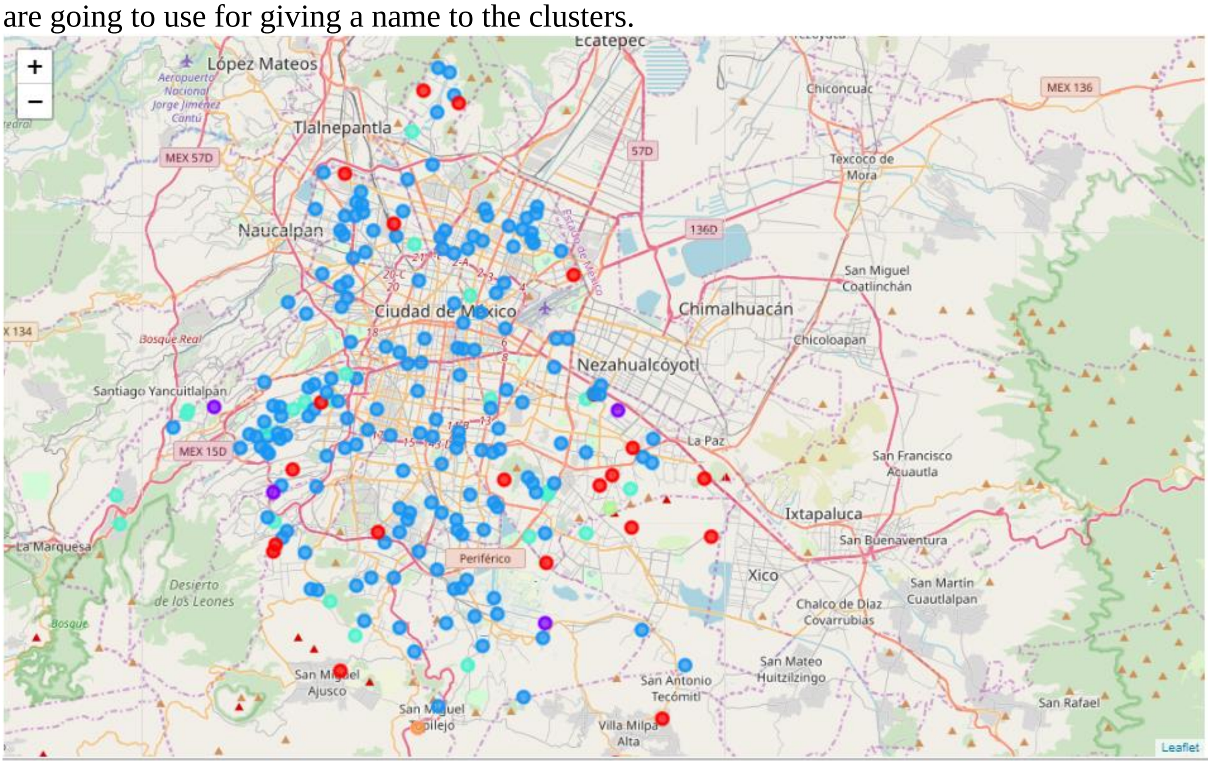


In the fourth step we are going to establish a Type Zone base on the distribution of the average price of square meter for each Borough



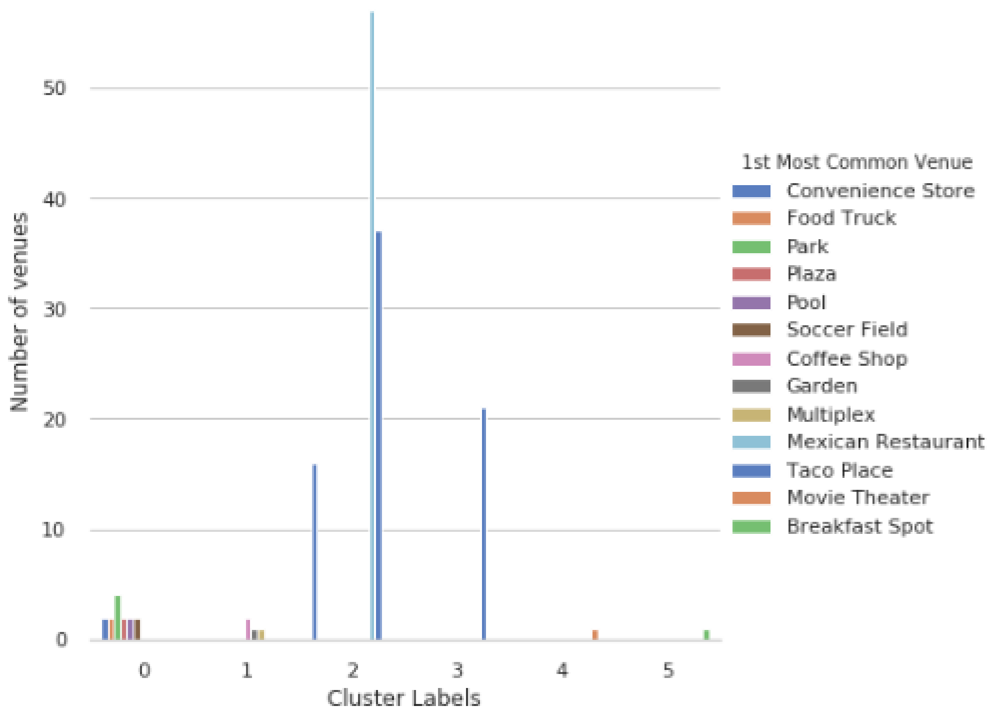
	Min_Range	Max_Range	Type of Zone
0	6897	13673	Low Level HSP
1	13673	20448	Bot Mid Level HSP
2	20448	27223	Top Mid Level HSP
3	27223	33998	High Level HSP
4	33998	40773	Top Level HSP

Moreover, a deep study of the venues that form each cluster and a map view are tools that we are going to use for giving a name to the clusters.

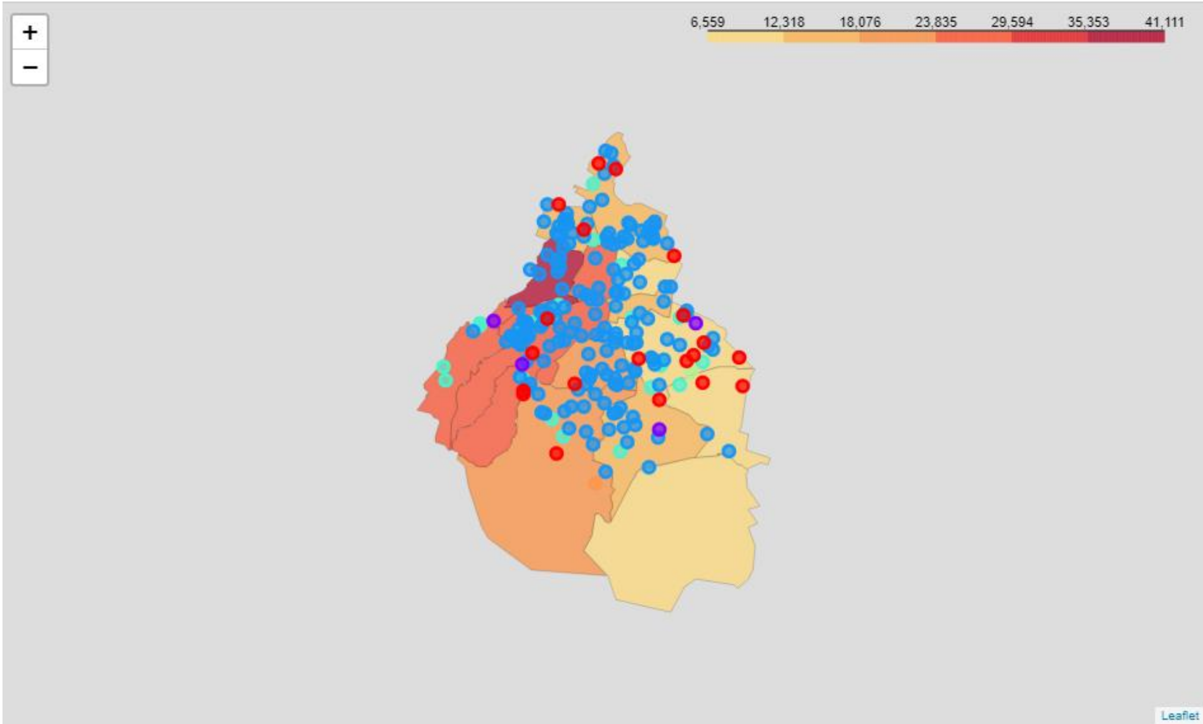


	Cluster Labels	1st Most Common Venue	Neighborhood
0	0	Convenience Store	2
1	0	Food Truck	2
2	0	Park	4
3	0	Plaza	2
4	0	Pool	2
5	0	Soccer Field	2
6	1	Coffee Shop	2
7	1	Garden	1
8	1	Multiplex	1
9	2	Convenience Store	16
10	2	Mexican Restaurant	57
11	2	Taco Place	37
12	3	Taco Place	21
13	4	Movie Theater	1
14	5	Breakfast Spot	1

	Cluster	Zone_Name
0	0	Park & Convience Store
1	1	Cofee Shop
2	2	Mexican Restaurant and Taco Place
3	3	Pure Taco Place
4	4	Movie Theather
5	5	Breakfast Spot

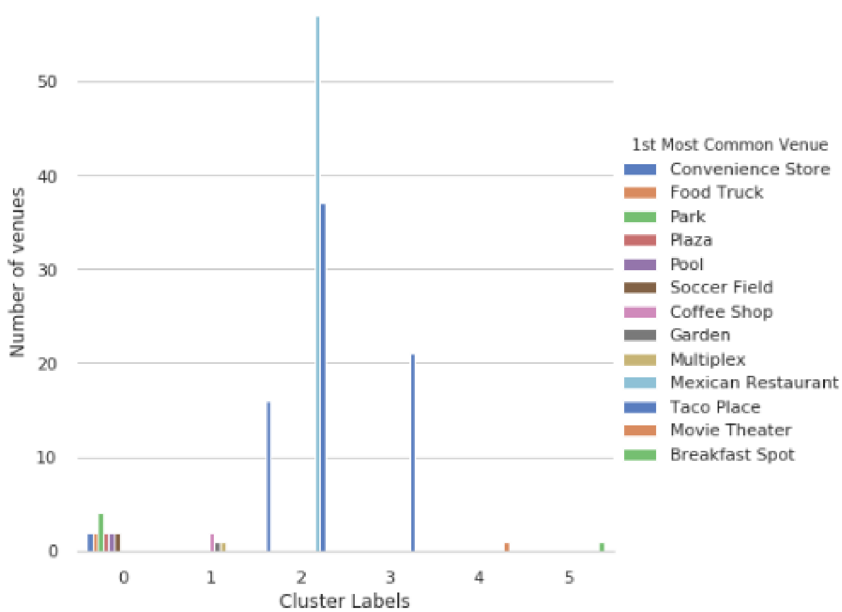


Finally, it will be build graphs of distribution of the average house price to establish the limits and two maps: one with the cluster information only, and other a choropleth map of Mexico City, that color the city base on the average price of the square meter in each Neighborhood and it is label with the information of the Price zone type, the Cluster name, the top 3 venues of the neighborhood and the name of the Neighborhood.



3.- Results and Discussion

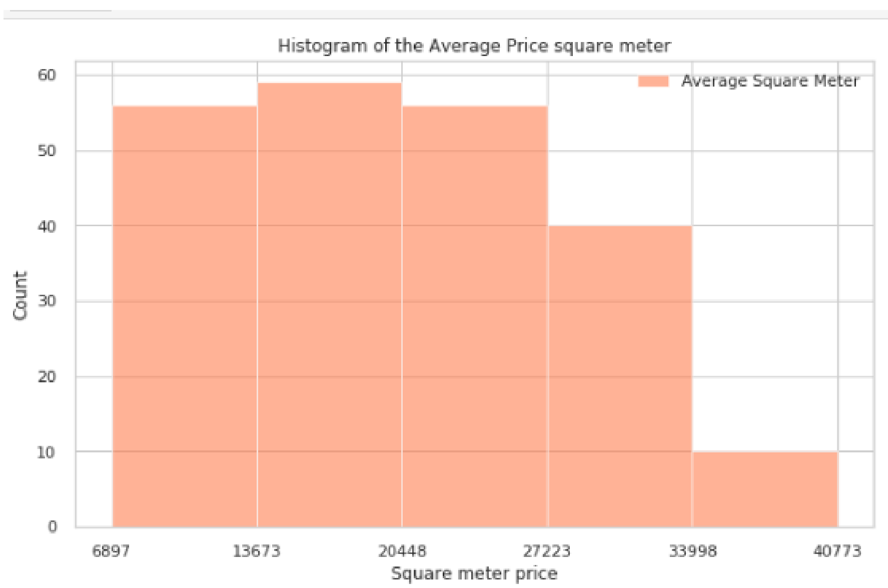
- The best way to analyze our results would be with the map, and the graphs

- 

The Cluster bar graph give us two important information about our sample. First how our clusters are formed, thus we can assign a specific name to each cluster:

	Cluster	Zone_Name
0	0	Park & Convience Store
1	1	Cofee Shop
2	2	Mexican Restaurant and Taco Place
3	3	Pure Taco Place
4	4	Movie Theather
5	5	Breakfast Spot

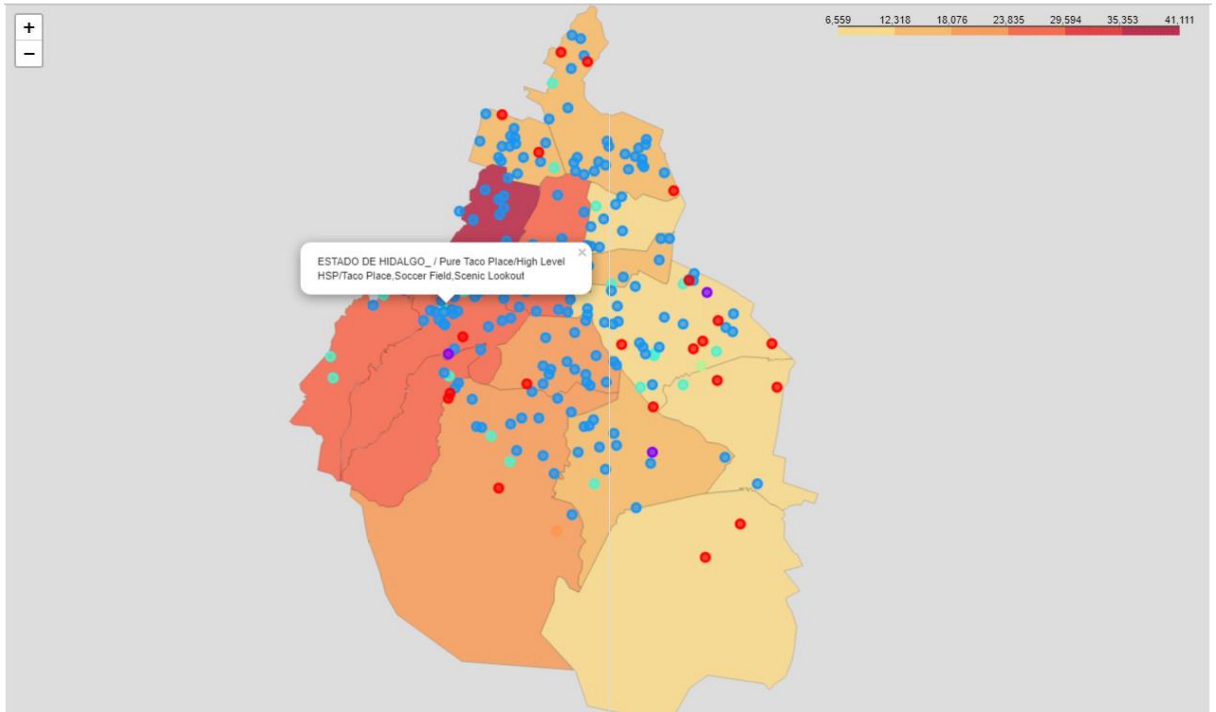
Second it shows that the cluster number two is the most competed, it has two of the most common places in Mexico City Taco Places and Mexican Restaurant, therefore for an investor these cluster is a competed zone for these venues. Furthermore, clusters as the 0, 1st, 4th and 5th are a complete different offer maybe people interested in art would like the 4th cluster or people who like to have close a convenience store would prefer the 0 cluster or people who likes to take a coffee cup will choose the 1st cluster.

- 

The histogram chart gave us valuable information regarding the distribution of housing in Mexico City it seems uniform in the first 3 segments, but as expected the luxury segments are less frequent. Consequently, we can classify each zone price.

	Min_Range	Max_Range	Type of Zone
0	6897	13673	Low Level HSP
1	13673	20448	Bot Mid Level HSP
2	20448	27223	Top Mid Level HSP
3	27223	33998	High Level HSP
4	33998	40773	Top Level HSP

It is very likely that the less payed employees would like the Low Level Housing, while mid salary employees would prefer the Bot Mid-Range or the Top Mis Range Housing and the well payed employees or management position look for High Level or Top Level Housing.



The Choropleth map gave meaningful information regarding the places, first each color point represent a different cluster, the color of the surface tell us which are the Housing Zone and finally the labels in each point tell us the Neighborhood, the Cluster Name, the Housing Price Zone and the top 3 venues in the zone.

The project was bounded to a sample space of 225 Neighborhoods, nevertheless working with the whole space the 1800 Neighborhoods will give a deeper and more complete study, certainly we would fine different Clusters but globally this was a good first approach for citizens and business man to make data-based decisions in Real Estate subjects.

4.- Conclusion

The purpose of this project was to present a tool for investors and citizens in order to take real estate business decisions. The aim of this study was accomplished with the delivery of the maps. Mainly, the last map that shows in color the expensive zones, the low-cost zones. Furthermore, the labels of the map give important information such as the Neighborhood name, the cluster name, the Housing type Zone and the top 3 venues.

An investor could use the map to look for the zones where restaurants, coffee shops or any business are not saturated such a look for the right zone to establish a business. Whereas, the common citizen will use the map to look for the best places to live base on the venues in the zone, the price of the zone and the vicinity to their work.

5.- References

[1] Mexico City recover from Wikipedia February 2020

[2] Coordinates of neighborhoods in Mexico City recover from CDMX government website February 2020

[3] Foursquare API

[4] Housing square meter average sales prices of each Borough recover from Metroscubicos February 2020