

IBM DATA SCIENCE PROFESSIONAL CERTIFICATE: CAPSTONE PROJECT

THE BATTLE OF NEIGHBORHOODS

Omar Torres Canales

Octubre 4, 2019

INTRODUCTION: BUSINESS PROBLEM

According to the World Health Organization (WHO), between 1975 and 2016 obesity has tripled worldwide. In 2016, more than 1900 million adults 18 and over were overweight, of which more than 650 million were obese. [1]

In Mexico, overweight and obesity affect 70% of Mexicans, which translates into a serious public health problem. WHO recommends that adults do at least 150 minutes of moderate-vigorous physical activity per week. From 2012 to 2016, the proportion of adults who did not comply with this recommendation in Mexico decreased slightly from 16% to 14.4%, being higher in men (15.2% vs. 13.7%) than in women (16.7% vs. 15.0%). [2]

This problem is one of the main aspects that motivate the analysis of this market and this translates into an extraordinary business opportunity. Accordingly, we will develop a project to find prosperous and optimal boroughs to establish gyms in Mexico City. This analysis will deepen into one of them, chosen for its optimal characteristics, we will explore its neighborhoods and find the most common places to later group them according to their behavior patterns.

The study is targeted at gym entrepreneurs and future entrepreneurs who are considering venturing into this sector. The findings will allow to stakeholders to have more information that results in better business decisions.

DATA

To achieve the objective of the project, the following information is required:

- Indicator that measures the prosperity of places [3]
- Boroughs data [4]
- Neighborhoods data [5]
- Population and land area data [6]
- Projections of population growth [7]
- Number and location of existing gyms in Mexico City [8]
- Information of the most common places [9]

NOTE: Consult references at the end of the document.

The population and gym data available in Mexico City will allow us to have a clear picture of current state of this market. The data will be divided by borough and we will get the information of the neighborhoods that integrate each of them. We will choose a borough that we can consider optimal to develop the project. To choose the borough where we will deepen the analysis, we will also consider the most appropriate information we find about the prosperity of the boroughs and we will use k-means clustering algorithm to discover insights and try to choose the best choice.

Finally, we will deepen the analysis of the chosen borough and explore the most common venues in each of the neighborhoods through the Foursquare API and the information obtained will be treated by k-means clustering algorithm to group the neighborhoods into clusters and be able to identify patterns of each cluster.

METHODOLOGY

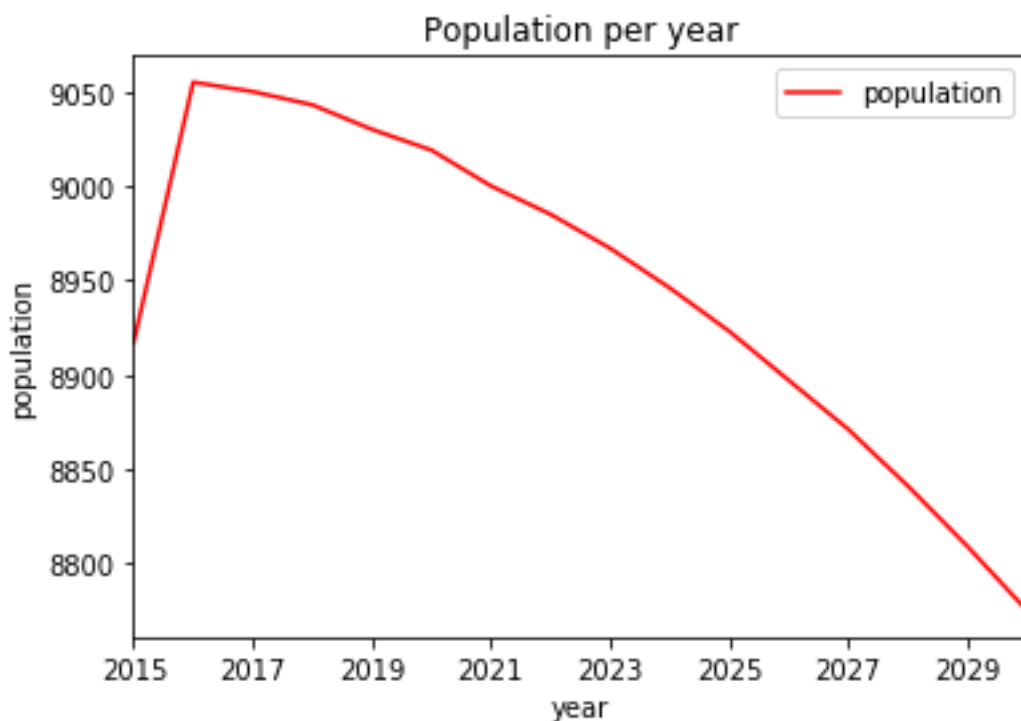
This project will determine which are the most suitable boroughs to establish gyms in Mexico. For the purposes of this analysis, it will consider the population and territorial characteristics, the prosperity indices and the amount of existing gyms.

Three groups will be created and each of the 16 boroughs will be part of one of them. We will choose the most appropriate group for our project and then we will explore one of the boroughs of the group chosen to deepen the analysis at neighborhood level. We will choose a borough that fits our needs and promises to be a success for our project.

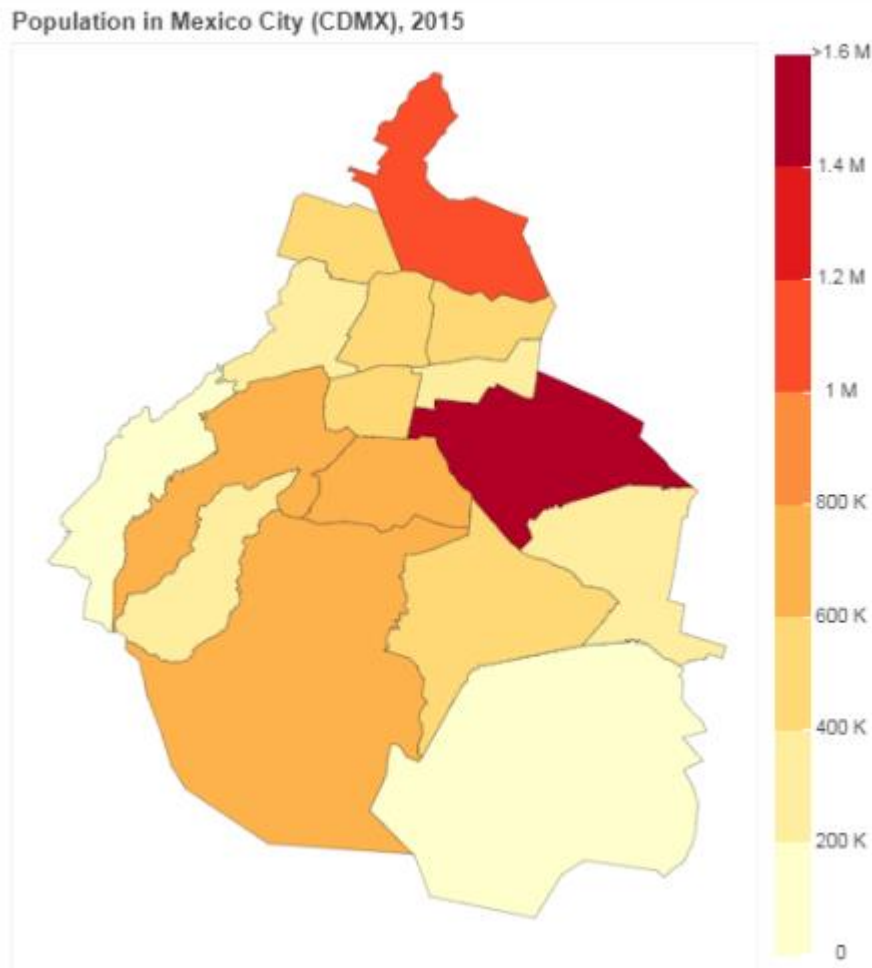
To perform the exploration we will use the location of each of the neighborhoods as a starting point. We will consider as limit the first 100 places provided by Foursquare API in a radius of 500 meters. We will obtain the 10 most common venues per neighborhood and then separate them into 8 groups according to their behavior patterns.

ANALYSIS

Population behavior of Mexico City over time.



This is the map of Mexico City, the darkest colors represent the boroughs with the most population. The most populated boroughs are Iztapalapa (1.8 million) and Gustavo A. Madero (1.1 million).



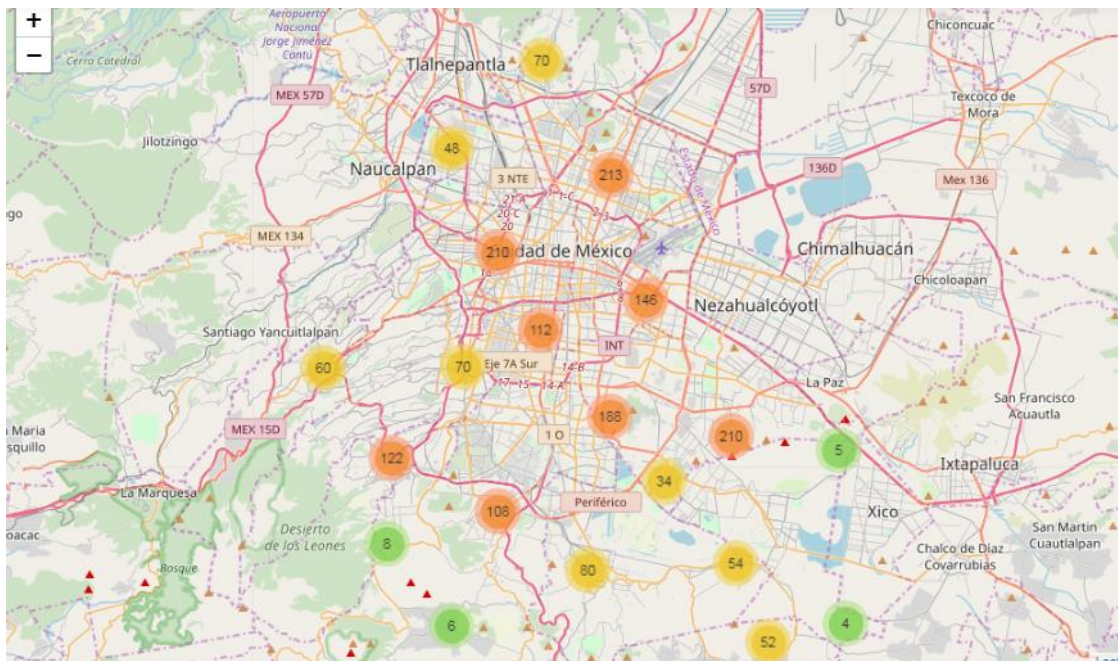
How many gyms are there in Mexico City?

- 1950 private sector gyms
- 175 public sector gyms

Gyms by borough

BOROUGH	GYMS
IZTAPALAPA	380
GUSTAVO A MADERO	283
TLALPAN	164
CUAUHTEMOC	153
COYOACAN	152
TLAHUAC	130
XOCHIMILCO	127
BENITO JUAREZ	116
ALVARO OBREGON	110
MIGUEL HIDALGO	104
VENUSTIANO CARRANZA	101
IZTACALCO	84
AZCAPOTZALCO	71
LA MAGDALENA CONTRERAS	58
CUAJIMALPA DE MORELOS	49
MILPA ALTA	43

This is a map of existing gyms inside of Mexico City (sample of 1,800 gyms). This will be done using clusters for better viewing and each cluster is then represented by the number of gyms in that area.



Running K-Means algorithm and labeling the boroughs.

	population2015_th	population_per1	population_per_square_kilometer1	land_area1	prosperity_index	Labels
0	677.0	7.6	2153.0	21.0	54.3	0
1	427.0	4.8	12617.0	2.3	53.7	1
2	400.0	4.5	11937.0	2.2	56.5	1
3	1828.0	20.5	16152.0	7.6	53.2	2
4	390.0	4.4	16902.0	1.5	54.4	1
5	364.0	4.1	7856.0	3.1	61.5	1
6	244.0	2.7	3846.0	4.2	53.4	0
7	608.0	6.8	11284.0	3.6	57.1	1
8	138.0	1.5	463.0	20.0	49.7	0
9	533.0	6.0	16374.0	2.2	55.7	1
10	750.0	8.4	7821.0	6.4	56.6	1
11	416.0	4.7	3645.0	7.6	52.2	0
12	362.0	4.1	4212.0	5.7	52.1	0
13	417.0	4.7	15633.0	1.8	57.4	1
14	199.0	2.2	2788.0	4.8	55.2	0
15	1164.0	13.1	13247.0	5.9	54.1	2

ID	BOROUGH
0	Tlalpan
1	Venustiano Carranza
2	Azcapotzalco
3	Iztapalapa
4	Iztacalco
5	Miguel Hidalgo
6	La Magdalena Contreras
7	Coyoacán
8	Milpa Alta
9	Cuauhtémoc
10	Álvaro Obregón
11	Xochimilco
12	Tláhuac
13	Benito Juárez
14	Cuajimalpa de Morelos
15	Gustavo A. Madero

Centroid Value

	population2015_th	population_per1	population_per_square_kilometer1	land_area1	prosperity_index
Labels					
0	339.333333	3.8000	2851.166667	10.5500	52.816667
1	486.125000	5.4625	12553.000000	2.8875	56.612500
2	1496.000000	16.8000	14699.500000	6.7500	53.650000

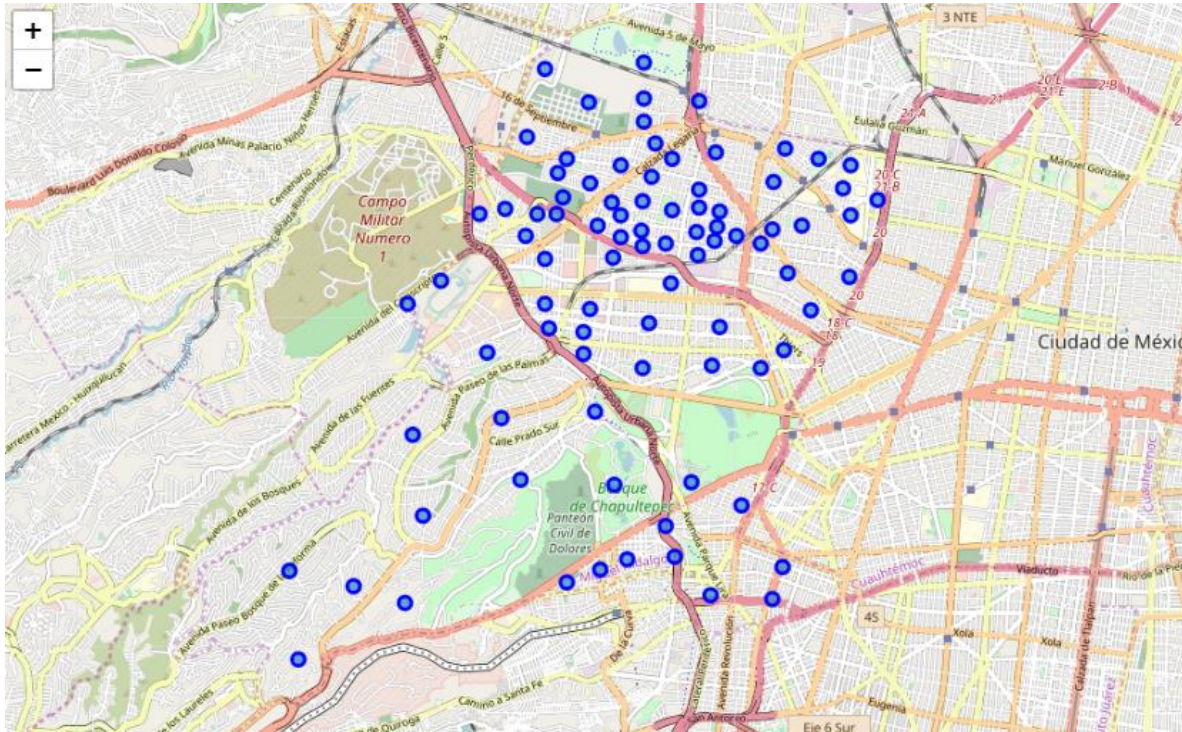
Analizing clusters

LABEL 0	LABEL 1	LABEL 2
Population: Low	Population: Middle	Population: High
Percentage of population: Low	Percentage of population: Middle	Percentage of population: High
Population per square kilometer: Low	Population per square kilometer: Middle	Population per square kilometer: High
Large area of land: High	Large area of land: Low	Large area of land: Middle
Prosperity Index: Low	Prosperity Index: High	Prosperity Index: Middle

WE CHOOSE THE "LABEL 1" GROUP MAINLY BECAUSE IT HAS THE HIGHEST PROSPERITY INDEX AND ITS POPULATION FEATURES ARE ACCEPTABLE TO DEVELOP THE PROJECT.

WITHIN "LABEL 1", WE IDENTIFY MIGUEL HIDALGO AS THE BOROUGH WITH THE HIGHEST PROSPERITY INDEX AND IT IS THE SEVENTH BOROUGH WITH THE LEAST AMOUNT OF EXISTING GYMS.

Mapping neighborhoods in Miguel Hidalgo



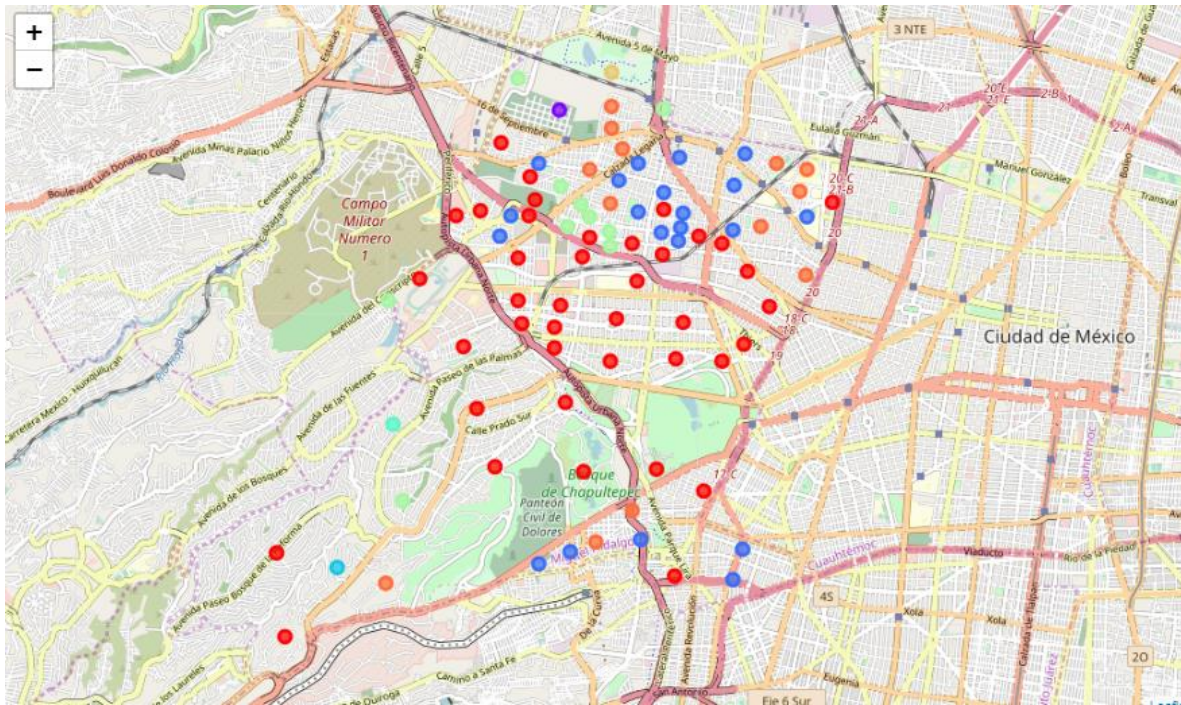
There are 88 neighborhoods in Miguel Hidalgo:

- LOMAS DE REFORMA
- DANIEL GARZA (AMPL)
- IGNACIO MANUEL ALTAMIRANO
- LEGARIA
- LEGARIA (U HAB)
- TORRE BLANCA
- UN HOGAR PARA NOSOTROS
- PENSIL SUR
- ANAHUAC PERALITOS
- AGRICULTURA
- ANAHUAC DOS LAGOS
- MODELO PENSIL
- POPO
- REFORMA PENSIL
- VENTURA PEREZ DE ALBA
- 16 DE SEPTIEMBRE

- POLANCO REFORMA (POLANCO)
- MOLINO DEL REY
- BOSQUES DE LAS LOMAS
- ESCANDON I
- ESCANDON II
- POPOTLA I
- SAN MIGUEL CHAPULTEPEC I
- BOSQUES DE CHAPULTEPEC (POLANCO)
- LOMAS DE BEZARES
- REFORMA SOCIAL
- LOMAS DE SOTELO
- SAN LORENZO TLALTENANGO
- PALMITAS (POLANCO)
- 10 DE ABRIL
- SAN MIGUEL CHAPULTEPEC II
- ANAHUAC II
- LOMAS DE CHAPULTEPEC
- LOMAS DE REFORMA (LOMAS DE CHAPULTEPEC)
- DEL BOSQUE (POLANCO)
- 5 DE MAYO
- PLUTARCO ELIAS CALLES
- CHAPULTEPEC MORALES (POLANCO)
- ANZURES
- CUAUHEMOC PENSIL
- GRANADA (AMPL)
- LOMAS VIRREYES (LOMAS DE CHAPULTEPEC)
- PERIODISTA
- SAN JOAQUIN
- VERONICA ANZURES
- SANTO TOMAS
- TATA LZARO
- ANAHUAC MARIANO ESCOBEDO
- ANAHUAC LAGO NORTE
- ANAHUAC LAGO SUR
- LOMAS ALTAS
- DEPORTIVA PENSIL
- TORRES TOREO
- AMERICA

- FRANCISCO I MADERO
- LOMAS DE BARRILACO (LOMAS DE CHAPULTEPEC)
- MORALES SECCION PALMAS (POLANCO)
- PENSIL NORTE
- PENSIL SAN JUANICO
- RINCON DEL BOSQUE
- TLAXPANA
- AHUEHUETES ANAHUAC
- LOMA HERMOSA (CONJ HAB)
- HUICHAPAN
- MILITAR 1 K LOMAS DE SOTELO (U HAB)
- OBSERVATORIO
- MANUEL AVILA CAMACHO
- POPO (AMPL)
- IRRIGACION
- MARINA NACIONAL (U HAB)
- MORALES SECCION ALAMEDA (POLANCO)
- TORRE BLANCA (AMPL)
- ARGENTINA ANTIGUA
- ANAHUAC LOS MANZANOS
- ANGEL ZIMBRON
- CHAPULTEPEC POLANCO (POLANCO)
- DANIEL GARZA
- LOS MORALES (POLANCO)
- NEXTITLA
- SAN DIEGO OCOYOACAC
- TACUBA
- TACUBAYA
- NUEVA ARGENTINA (ARGENTINA PONIENTE)
- BOSQUE DE CHAPULTEPEC I II Y III SECCIONES
- MEXICO NUEVO
- ANAHUAC I
- GRANADA
- POPOTLA II

Clustering neighborhoods considering its 10 most common venues



RESULTS AND DISCUSSION

There are 16 boroughs in Mexico City and it represents 7.5% of the national population. The population projections of the boroughs towards the year 2030 show that the trend is down, until 2015 there were almost 9 million people.

INEGI has a record of 2125 gyms throughout the city, Iztapalapa being the borough with more existing gyms, 1950 from the private sector and 175 from the public sector. Iztapalapa is the borough with the largest number of gyms (380) and Milpa Alta has the least (43), despite being one of the largest boroughs.

After running the K-means algorithm, the boroughs that belong to each of three groups are identified according to their characteristics they have in common. "Label 0" is identified as low, "Label 1" as medium and "Label 2" as high. But it is important to mention that the label marked 1 has the highest prosperity index.

To deepen the analysis we chose Miguel Hidalgo because it is the borough with the highest prosperity index, acceptable population characteristics according to its size and is the seventh borough with the least amount of existing gyms.

Later we found that Miguel Hidalgo has 88 neighborhoods. We use the Foursquare API to explore each of the neighborhoods and thus find its 10 most common venues within a radius of 500 meters. With the information, we proceeded to run the K-means algorithm again to identify them in 8 different groups according to their behavior patterns.

Analyzing the groups, we identify clusters 0, 2, 5 and 7 as the most commercial areas and specifically cluster 0 is the one with the most gyms working (15), followed by cluster 2 (10).

CONCLUSION

For the purposes of having an even more enriched analysis, we could include, for example, data on public safety or the cost of renting business premises. However, with the data we obtained we can have an estimate of which are the most appropriate boroughs to establish gyms in Mexico. The group identified as "Label 1" has very interesting characteristics because although the boroughs that compose it do not have as much population as the other groups, they are also not the largest in the territory and have a high prosperity rate compared to the other two groups.

Talking about Miguel Hidalgo, the borough explored in the development of this project, we discovered that there are highly commercial areas, mainly in Polanco or in the surrounding neighborhoods identified by that name. Cluster 0 is the group with the highest commercial offer and 15 of its gyms are among the most common venues. Likewise, we consider cluster 2 as very promising to implement our project, it has 10 gyms within the most common venues. I consider it necessary to evaluate the costs of renting commercial premises and also assess crime rates.

REFERENCES

1. **Obesity and overweight. World Health Organization**
<https://www.who.int/es/news-room/fact-sheets/detail/obesity-and-overweight>
2. **Encuesta Nacional de Salud y Nutrición de Medio Camino 2016 (ENSANUT 2016).** http://transparencia.insp.mx/2017/auditorias-insp/12701_Resultados_Encuesta_ENSANUT_MC2016.pdf
3. **Índice de las Ciudades Prósperas, CPI, México 2018—ONU**
<https://onuhabitat.org.mx/index.php/indice-de-las-ciudades-prosperas-cpi-mexico-2018>
4. **Alcaldías—Gobierno de la Ciudad de México**
<https://datos.cdmx.gob.mx/explore/dataset/alcaldias/table/>
5. **Colonias—Gobierno de la Ciudad de México**
<https://datos.cdmx.gob.mx/explore/dataset/coloniascdmx/table/>
6. **Panorama sociodemográfico de Ciudad de México 2015—INEGI**
<https://www.inegi.org.mx/app/biblioteca/ficha.html?upc=702825082178>
7. **Proyecciones de la Población de los Municipios de México, 2015–2030 (base 1)—CONAPO**
<https://datos.gob.mx/busca/dataset/proyecciones-de-la-poblacion-de-mexico-y-de-las-entidades-federativas-2016-2050/resource/efd9c27a-c8f5-4980-ab4d-e29c6e7638c8>
8. **Directorio Estadístico Nacional de Unidades Económicas CDMX 2019/04—INEGI**
<https://www.inegi.org.mx/app/descarga/?ti=6#saladeprensa>
9. **Foursquare API**
<https://developer.foursquare.com/>

Here is a link to my code in Github:

https://github.com/otorresca23/capstone-project/blob/master/capstone_project_week2.ipynb