APPLIED DATA SCIENCE CAPSTONE FROM IBM BY COURSERA

CAPSTONE PROJECT

"Analysis of places to start a pharmacy business"

Author: Carlos A. Evangelista Busso

1. Business Problem:

The present project is born in the middle of a problem caused by the virus called: Coronavirus or Covid-19, which shakes the entire world in 2020, in most countries, quarantine was established by governments to prevent the spread of said virus, keeping people in their homes. Much of the population, and also in my case, was to spend almost the entire period of quarantine inside my home, and even more so due to my work situation in the mining sector that suspended all activities.

The days I stayed at home I had the opportunity to observe with my family the activity carried out by the business located in front of my home, a pharmacy, where merchandise arrived day after day, since their activities had not been suspended because it was the area of health and this was something that we did not notice despite its years of existence. At first glance, it is a very profitable business to stay active even in times of crisis; it was at that moment where together with my family we asked ourselves the following question: What place or areas are profitable to open a pharmacy business?

The city of Metropolitan Lima is the capital of Peru and has about 9.5 million inhabitants [1] and in this there are approximately 198 first-level health establishments [2]. It is in this city that this project is focused and an analysis will be carried out using geolocation to be able to determine the location where a pharmacy business could be started and justify the reasons for it to be profitable.

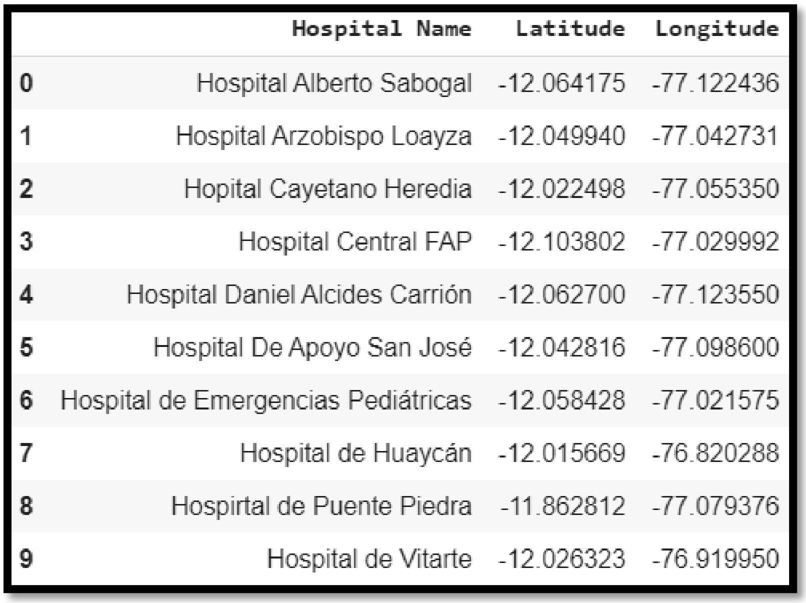
2. Data Description:

To solve the problem, an analysis was made of the areas around the main hospitals in Lima. Logically, it is through these places that people who are treated and subsequently prescribed to acquire medicines by doctors pass through.

- In order to obtain the coordinate information of these places, first we must have a list of the names of the hospitals with which we will work. This data collection was performed through a web search [3] and subsequent use of the method: Web Scraping.
- Once the list of all hospital names has been obtained, information is extracted from the coordinates of each one, this is done using the Google Maps API: Places API [4].
- With the coordinates, the Foursquare API was used to obtain the most common places around each hospital within a radius of 500 meters [5].

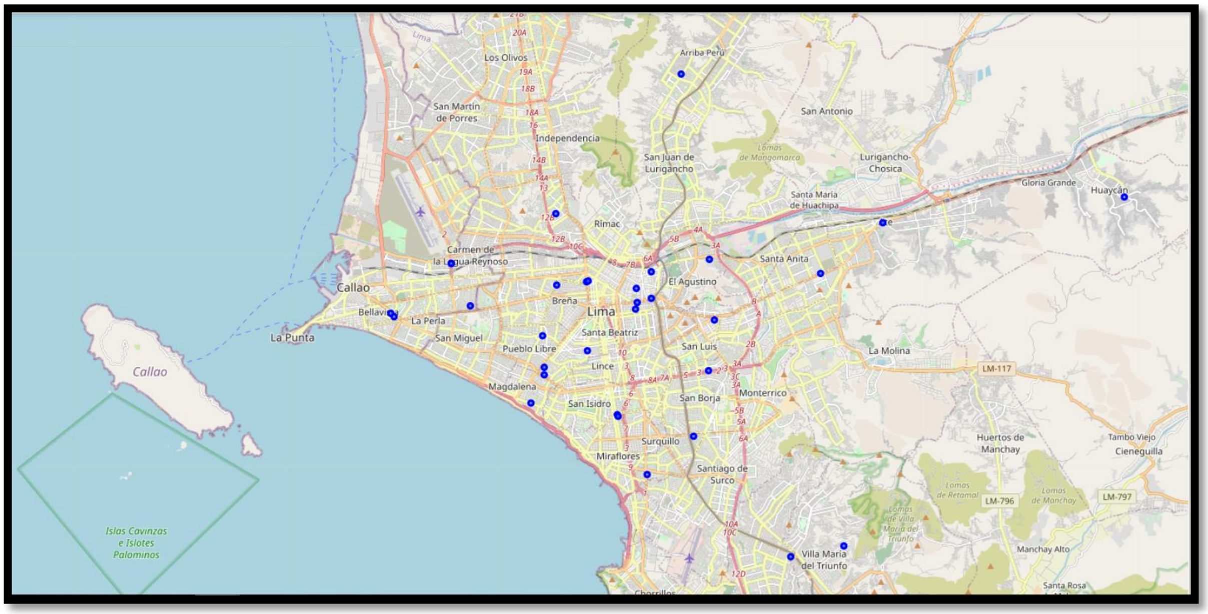
3. Methodology:

Using the data set collected with "Places API Google", we proceed with the preparation of data. A Dataframe is formed with the following columns: "Hospital Name", "Latitude" and "Longitude".

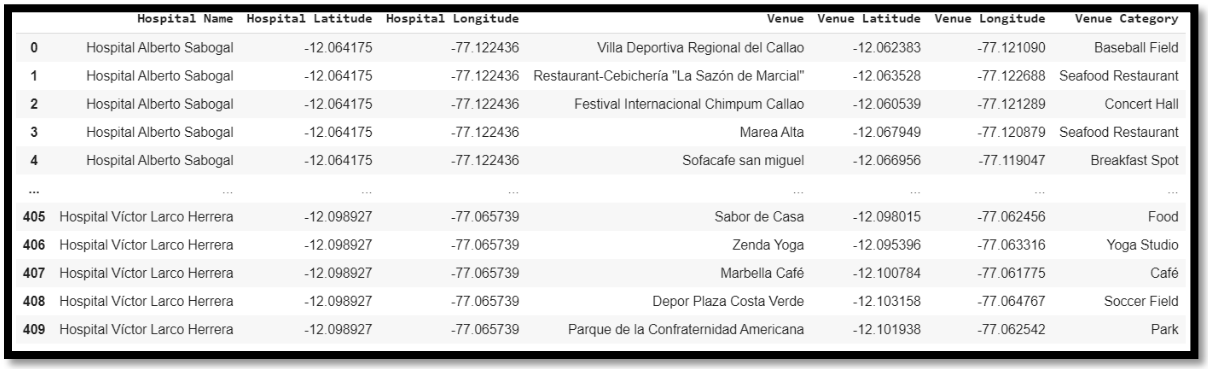


	Hospital Name	Latitude	Longitude
0	Hospital Alberto Sabogal	-12.064175	-77.122436
1	Hospital Arzobispo Loayza	-12.049940	-77.042731
2	Hopital Cayetano Heredia	-12.022498	-77.055350
3	Hospital Central FAP	-12.103802	-77.029992
4	Hospital Daniel Alcides Carrión	-12.062700	-77.123550
5	Hospital De Apoyo San José	-12.042816	-77.098600
6	Hospital de Emergencias Pediátricas	-12.058428	-77.021575
7	Hospital de Huaycán	-12.015669	-76.820288
8	Hospirtal de Puente Piedra	-11.862812	-77.079376
9	Hospital de Vitarte	-12.026323	-76.919950

To visualize the previous information on a map and verify the correct location of the hospitals, the Python **Folium** library was used.

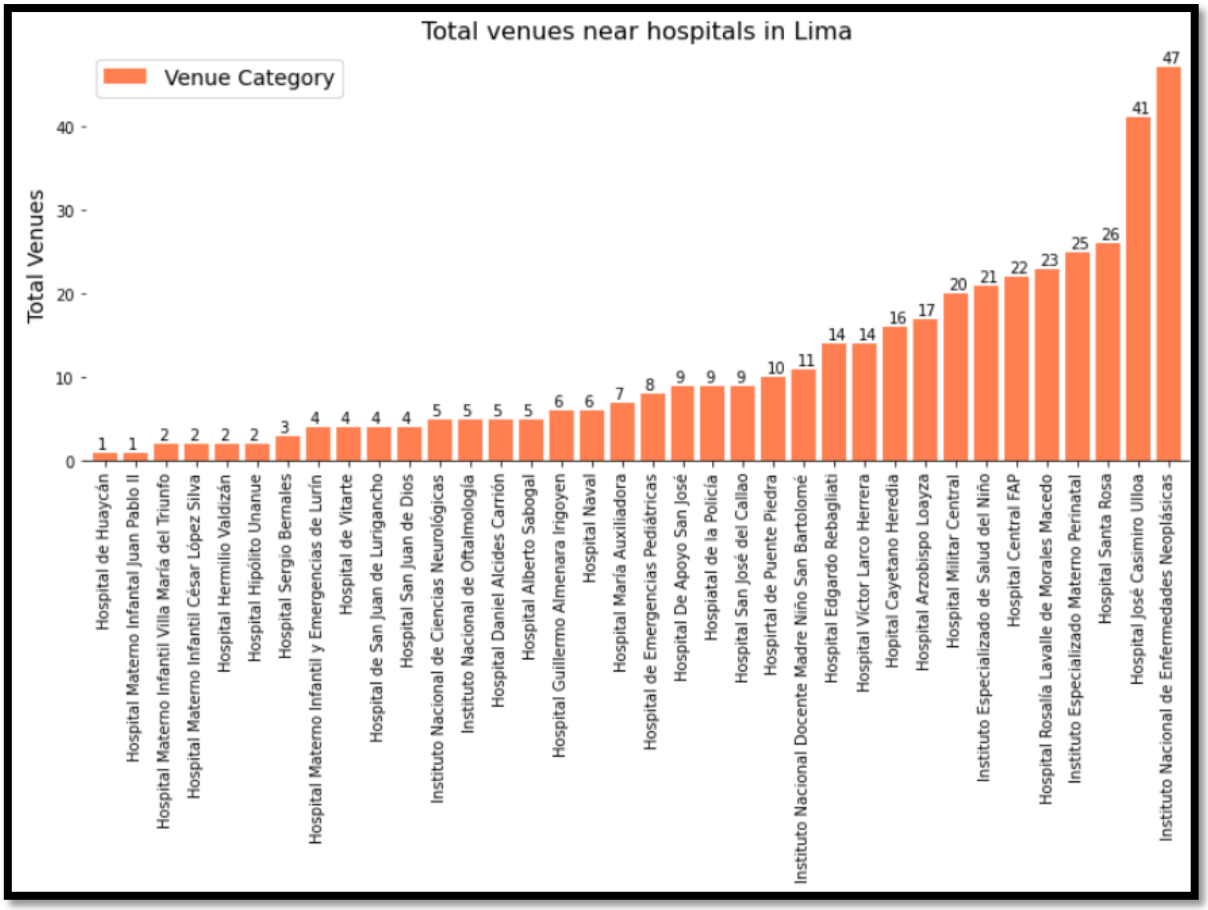


We obtain the places around each hospital such as: parks, cafes, restaurants, etc. For this, Foursquare API was used with the parameters of 100 venues as a limit and a radius of 500 meters. A Dataframe was formed with the most important data from the information obtained.



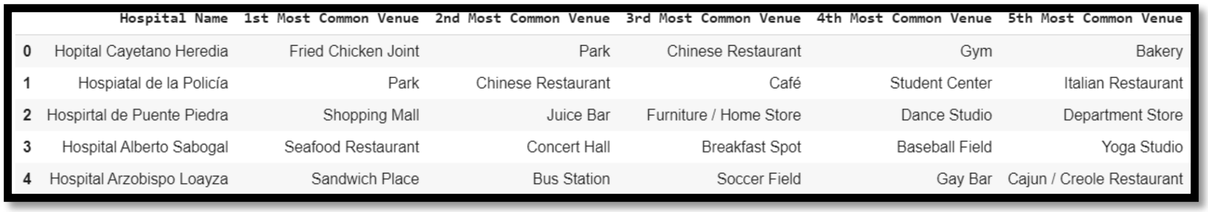
	Hospital Name	Hospital Latitude	Hospital Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Hospital Alberto Sabogal	-12.064175	-77.122436	Villa Deportiva Regional del Callao	-12.062383	-77.121090	Baseball Field
1	Hospital Alberto Sabogal	-12.064175	-77.122436	Restaurant-Cebichería "La Sazón de Marcial"	-12.063528	-77.122688	Seafood Restaurant
2	Hospital Alberto Sabogal	-12.064175	-77.122436	Festival Internacional Chimpum Callao	-12.060539	-77.121289	Concert Hall
3	Hospital Alberto Sabogal	-12.064175	-77.122436	Marea Alta	-12.067949	-77.120879	Seafood Restaurant
4	Hospital Alberto Sabogal	-12.064175	-77.122436	Sofacafe san miguel	-12.066956	-77.119047	Breakfast Spot
...	...	...	...	...	...	...	...
405	Hospital Víctor Larco Herrera	-12.098927	-77.065739	Sabor de Casa	-12.098015	-77.062456	Food
406	Hospital Víctor Larco Herrera	-12.098927	-77.065739	Zenda Yoga	-12.095396	-77.063316	Yoga Studio
407	Hospital Víctor Larco Herrera	-12.098927	-77.065739	Marbella Café	-12.100784	-77.061775	Café
408	Hospital Víctor Larco Herrera	-12.098927	-77.065739	Depor Plaza Costa Verde	-12.103158	-77.064767	Soccer Field
409	Hospital Víctor Larco Herrera	-12.098927	-77.065739	Parque de la Confraternidad Americana	-12.101938	-77.062542	Park

In order to visualize the total number of venues found with the Foursquare API, the following bar chart is used, in which it can be seen that the hospitals with the fewest places nearby (1) are: "Hospital de Huaycán" and "Hospital Materno Infantil Juan Pablo II". On the other hand, the Hospital with the most nearby places (47) is "National Institute of Neoplastic Diseases".



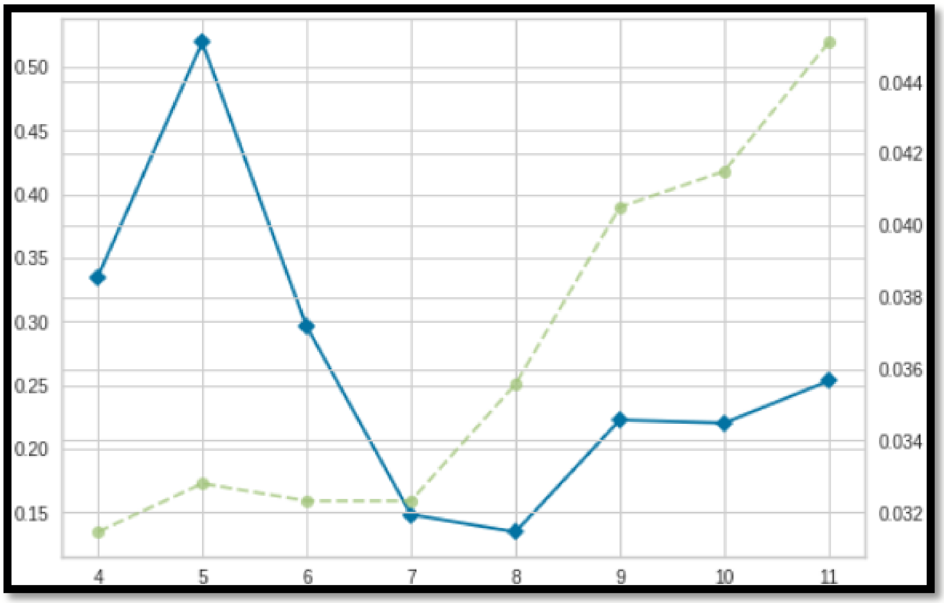
The amount of nearby venues that was obtained does not necessarily coincide with reality, this will depend on the quality of the information that is available, as well as the various sources that are consulted. In this work, Foursqueare API was used for this and more detailed information could be collected from other sources and thereby increase the precision of the analysis.

A Dataframe was performed showing the 5 most common places located near each hospital.

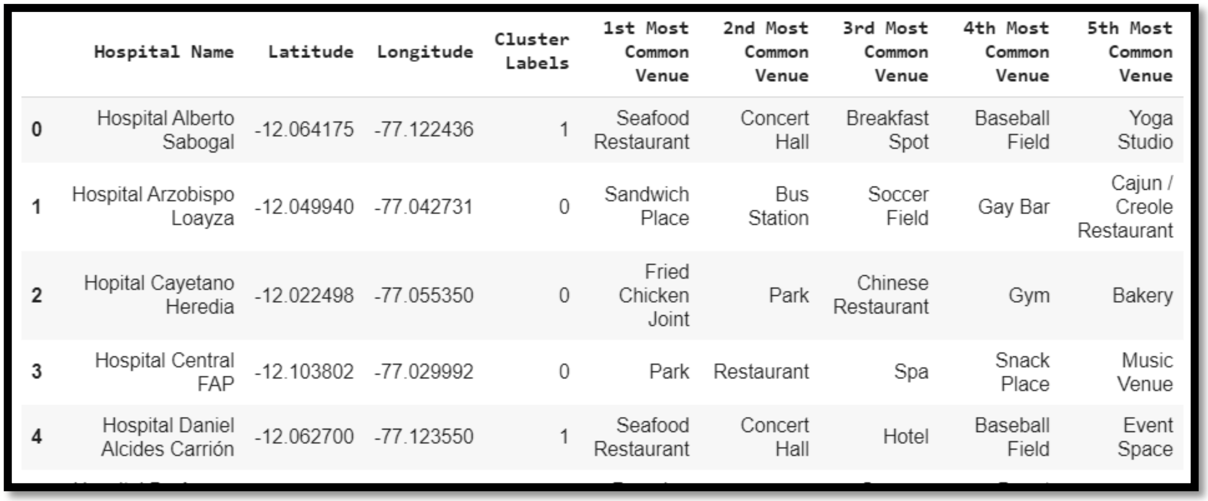


	Hospital Name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Hopital Cayetano Heredia	Fried Chicken Joint	Park	Chinese Restaurant	Gym	Bakery
1	Hospiatal de la Policía	Park	Chinese Restaurant	Café	Student Center	Italian Restaurant
2	Hospirtal de Puente Piedra	Shopping Mall	Juice Bar	Furniture / Home Store	Dance Studio	Department Store
3	Hospital Alberto Sabogal	Seafood Restaurant	Concert Hall	Breakfast Spot	Baseball Field	Yoga Studio
4	Hospital Arzobispo Loayza	Sandwich Place	Bus Station	Soccer Field	Gay Bar	Cajun / Creole Restaurant

Having this information, a Dataframe is formed with the Pandas (Python) get_dummies method and applied to “Venue Category” column. The unsupervised K-means algorithm is then used to group the hospitals. For this, an analysis is carried out with the “Elbow” method and the number of groupings for the algorithm is determined, this results: 7, as shown in the following graph:

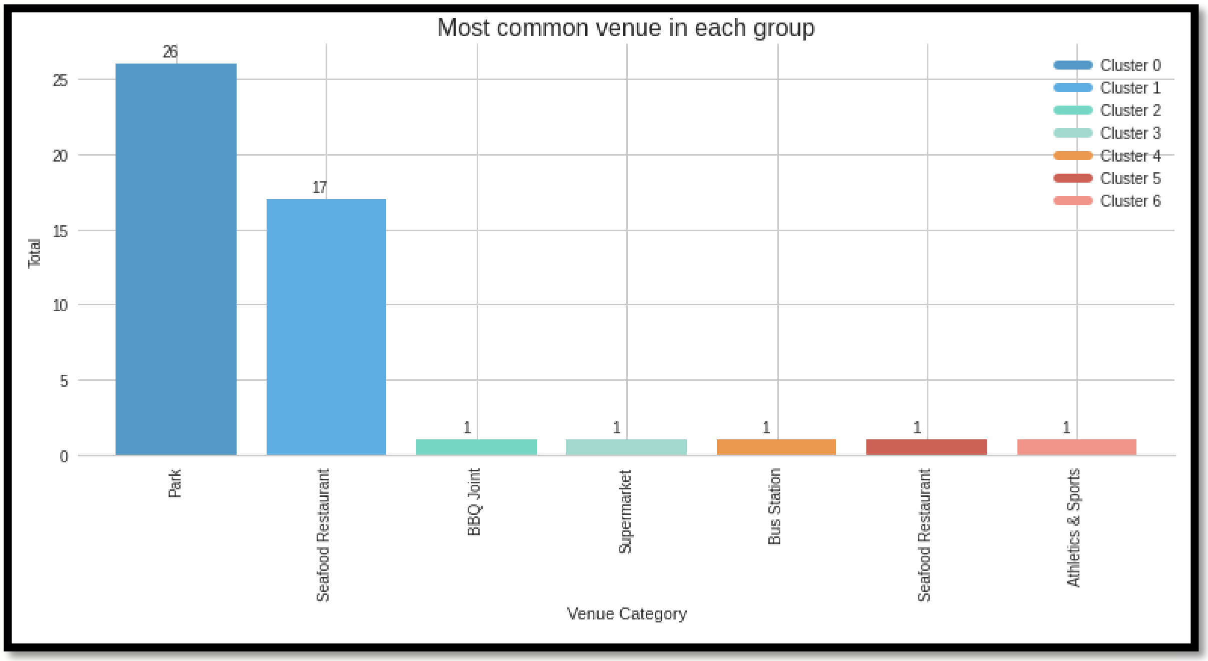


After grouping by K-means, the table is generated with the column "Cluster Labels", shown in the following image:



	Hospital Name	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Hospital Alberto Sabogal	-12.064175	-77.122436	1	Seafood Restaurant	Concert Hall	Breakfast Spot	Baseball Field	Yoga Studio
1	Hospital Arzobispo Loayza	-12.049940	-77.042731	0	Sandwich Place	Bus Station	Soccer Field	Gay Bar	Cajun / Creole Restaurant
2	Hopital Cayetano Heredia	-12.022498	-77.055350	0	Fried Chicken Joint	Park	Chinese Restaurant	Gym	Bakery
3	Hospital Central FAP	-12.103802	-77.029992	0	Park	Restaurant	Spa	Snack Place	Music Venue
4	Hospital Daniel Alcides Carrión	-12.062700	-77.123550	1	Seafood Restaurant	Concert Hall	Hotel	Baseball Field	Event Space

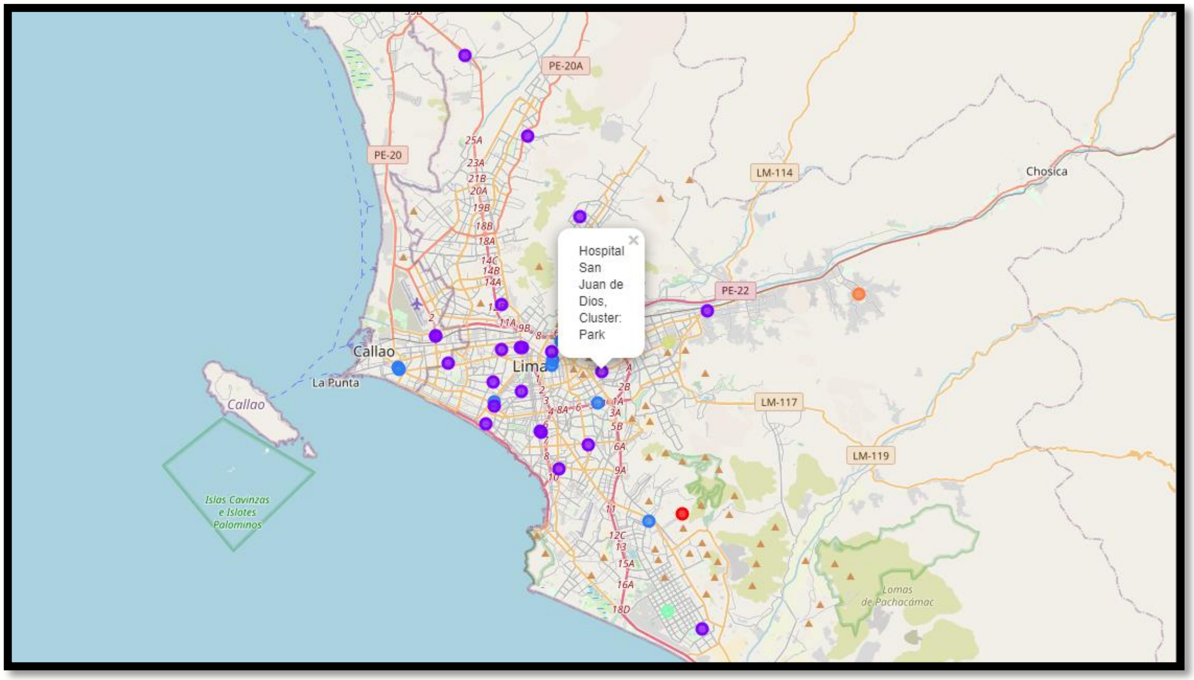
In addition to the previous table, the most common venue in each of the groups is determined; we can take this data as representative of each group. The result is visualized in the following graph:



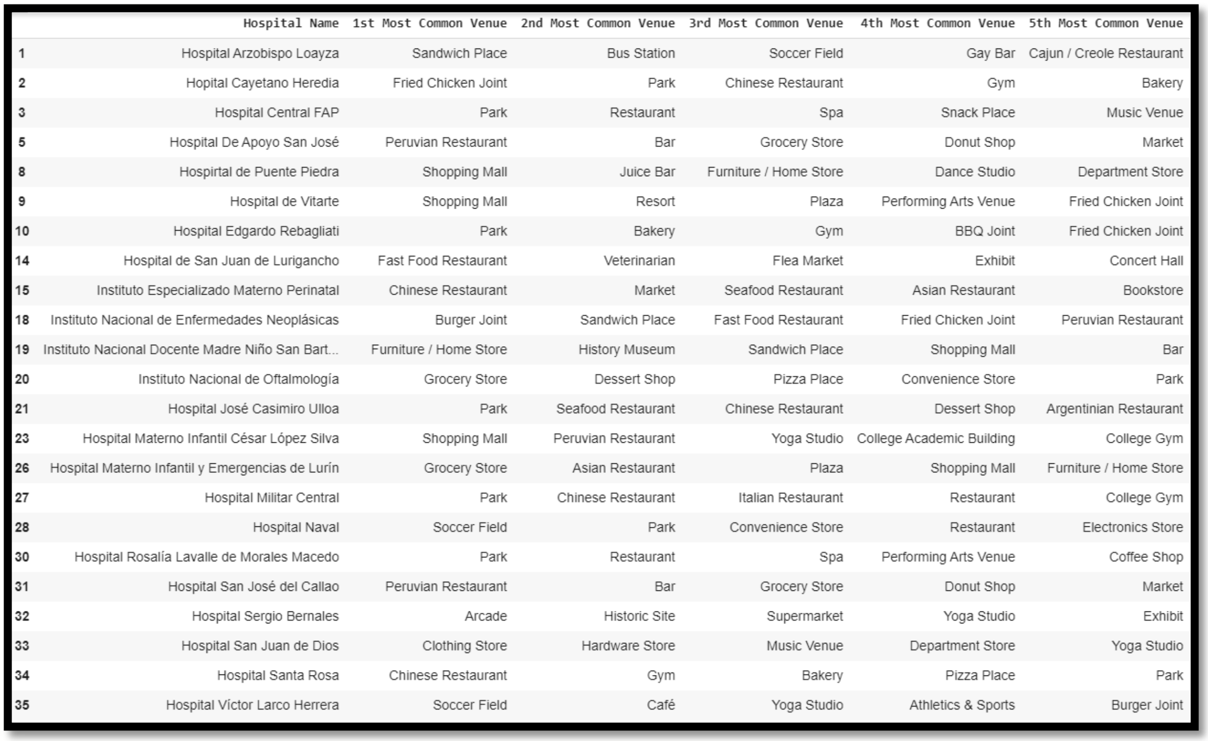
It can be seen from the previous graph that clusters 0 and 1 have the most common places well defined, the representative place of cluster 0 is "Park" and for cluster 1 it is "Seafood Restaurant".

4. Results:

To observe the results of the clustering by K-means, we generated a map with colored marks that represent each of the clusters. The label of each point shows the name of the hospital and the representative venue of the cluster.

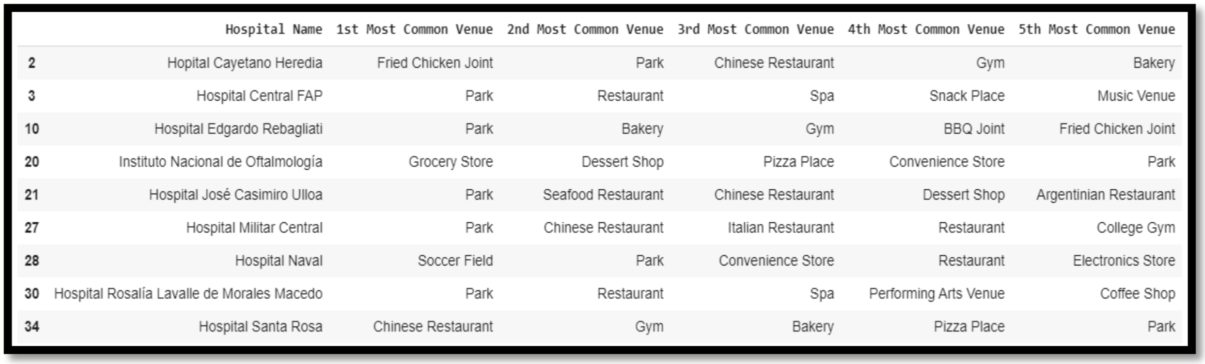


In the previous section, it was observed that "cluster 0" is where there are areas with the highest concentration of people, such as parks, which is why this group is taken for the analysis. The information corresponding to this cluster is separated and shown in the following table:



	Hospital Name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Hospital Arzobispo Loayza	Sandwich Place	Bus Station	Soccer Field	Gay Bar	Cajun / Creole Restaurant
2	Hopital Cayetano Heredia	Fried Chicken Joint	Park	Chinese Restaurant	Gym	Bakery
3	Hospital Central FAP	Park	Restaurant	Spa	Snack Place	Music Venue
5	Hospital De Apoyo San José	Peruvian Restaurant	Bar	Grocery Store	Donut Shop	Market
8	Hospirtal de Puente Piedra	Shopping Mall	Juice Bar	Furniture / Home Store	Dance Studio	Department Store
9	Hospital de Vitarte	Shopping Mall	Resort	Plaza	Performing Arts Venue	Fried Chicken Joint
10	Hospital Edgardo Rebagliati	Park	Bakery	Gym	BBQ Joint	Fried Chicken Joint
14	Hospital de San Juan de Lurigancho	Fast Food Restaurant	Veterinarian	Flea Market	Exhibit	Concert Hall
15	Instituto Especializado Materno Perinatal	Chinese Restaurant	Market	Seafood Restaurant	Asian Restaurant	Bookstore
18	Instituto Nacional de Enfermedades Neoplásicas	Burger Joint	Sandwich Place	Fast Food Restaurant	Fried Chicken Joint	Peruvian Restaurant
19	Instituto Nacional Docente Madre Niño San Bart...	Furniture / Home Store	History Museum	Sandwich Place	Shopping Mall	Bar
20	Instituto Nacional de Oftalmología	Grocery Store	Dessert Shop	Pizza Place	Convenience Store	Park
21	Hospital José Casimiro Ulloa	Park	Seafood Restaurant	Chinese Restaurant	Dessert Shop	Argentinian Restaurant
23	Hospital Materno Infantil César López Silva	Shopping Mall	Peruvian Restaurant	Yoga Studio	College Academic Building	College Gym
26	Hospital Materno Infantil y Emergencias de Lurín	Grocery Store	Asian Restaurant	Plaza	Shopping Mall	Furniture / Home Store
27	Hospital Militar Central	Park	Chinese Restaurant	Italian Restaurant	Restaurant	College Gym
28	Hospital Naval	Soccer Field	Park	Convenience Store	Restaurant	Electronics Store
30	Hospital Rosalía Lavalle de Morales Macedo	Park	Restaurant	Spa	Performing Arts Venue	Coffee Shop
31	Hospital San José del Callao	Peruvian Restaurant	Bar	Grocery Store	Donut Shop	Market
32	Hospital Sergio Bernales	Arcade	Historic Site	Supermarket	Yoga Studio	Exhibit
33	Hospital San Juan de Dios	Clothing Store	Hardware Store	Music Venue	Department Store	Yoga Studio
34	Hospital Santa Rosa	Chinese Restaurant	Gym	Bakery	Pizza Place	Park
35	Hospital Víctor Larco Herrera	Soccer Field	Café	Yoga Studio	Athletics & Sports	Burger Joint

From the table above, extract the hospitals that have parks among their most common places. We show the results:



	Hospital Name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
2	Hopital Cayetano Heredia	Fried Chicken Joint	Park	Chinese Restaurant	Gym	Bakery
3	Hospital Central FAP	Park	Restaurant	Spa	Snack Place	Music Venue
10	Hospital Edgardo Rebagliati	Park	Bakery	Gym	BBQ Joint	Fried Chicken Joint
20	Instituto Nacional de Oftalmología	Grocery Store	Dessert Shop	Pizza Place	Convenience Store	Park
21	Hospital José Casimiro Ulloa	Park	Seafood Restaurant	Chinese Restaurant	Dessert Shop	Argentinian Restaurant
27	Hospital Militar Central	Park	Chinese Restaurant	Italian Restaurant	Restaurant	College Gym
28	Hospital Naval	Soccer Field	Park	Convenience Store	Restaurant	Electronics Store
30	Hospital Rosalía Lavalle de Morales Macedo	Park	Restaurant	Spa	Performing Arts Venue	Coffee Shop
34	Hospital Santa Rosa	Chinese Restaurant	Gym	Bakery	Pizza Place	Park

This list represents the set of options of places proposed in this work for the opening of a pharmacy business.

5. Discussion:

This work analyzed the places that are in the area around the main hospitals. Therefore, this work represents one of the different methods that exist to respond to the problem, the results always depend on the type of information collected and the quality of it.

The clustering was done with the data extracted from Foursqueare API and possibly this data does not actually represent the nearby places that exist in each hospital, this can vary for example using other types of APIs such as Google Places API, which is not completely free and it has a cost for use.

In the final part of the work, a list of the proposed places for the solution of the problem was presented. In order to choose a place, you can take into account the district where each hospital is located and select one of them, for example: the most central location in the city of Lima, given that it is the area where most people pass.

6. Conclusion:

The areas around hospitals represent places where there is a large number of businesses and the one that stands out the most is that of pharmacies. For this reason, it is the starting point to carry out an analysis on the opening of a business of this type.

The updated information is very useful and for this analysis by geolocation much more, it provides a better panorama of reality and with it a better result is obtained.

Carlos Evangelista.

Lima-Peru

04 May 2020

7. References:

- [1] "LIMA METROPOLITANA ESTADISTICAS", Mimp.gob.pe, 2020. [Online]. Available: https://www.mimp.gob.pe/adultomayor/regiones/Lima_Metro2.html.
- [2] M. Lopez, "Aniversario de Lima: ¿cuántos establecimientos de salud funcionan en la capital y cuáles son sus carencias?". [Online]. Available: https://rpp.pe.
- [3] "Hospitales en Lima | Inicia", Inicia. [Online]. Available: https://inicia.pe/hospitales-lima.
- [4] "Overview | Places API | Google Developers", Google Developers, 2020. [Online]. Available: https://developers.google.com/places/web-service/intro.
- [5] [Online]. Available: https://foursquare.com/.