Analysis of Covid-19 cases in the city of Buenos Aires

IBM Data Science Professional Certificate – Capstone Project

Héctor Moreno García

6/18/20

Table of Contents

1. Introduction

1.1 Background

As it is well known, the COVID-19 pandemic has created an enormous amount of challenges and difficulties worldwide, and Argentina has been no exception. In Argentina, the first case of COVID-19 was reported on March 3, 2020 (1) and two weeks later, a mandatory nation-wide lockdown was announced (2), which remains in effect in its capital, the city of Buenos Aires.

As of June 18, 2020, 17,196 COVID-19 cases have been reported, only in the city of Buenos Aires, which represent 45.9% of the total cases reported in the country. Additionally, 359 people have died in the city due to this pandemic. (3, p. 12)

1.2 Problem

The statistics of infections and deaths that are reported periodically only display the total of cases or, sometimes, the information about gender or age group, but other variables are not reported. The purpose of this project is to collect and analyze different data in order to locate the *comunas* (communes)[1] with the majority of cases and other factors that might have a strong correlation with the reported cases of COVID-19. Then, applying machine learning techniques the *comunas* will be clustered based on the variables shown in the exploratory analysis

1.3 Audience

This analysis may possibly be of use for different social organizations and projects; that said, the audience who might benefit the most from the results are government officials and health experts. These data could provide an additional insight to the plans that are designed to prevent the spread of similar epidemics.

[1] The city of Buenos Aires is administratively divided into fifteen *comunas* (communes).

2. Data

2.1 Data description

The main data collected for this project consisted in the number of COVID-19 cases reported in the city of Buenos Aires, which was scraped from the webpage of the newspaper La Nación (3, 4), and the geographical location of each neighborhood and *comuna* of the city, which was retrieved from Wikipedia (5) and from GeoHack (6), that was fundamental to get the decimal values of the *comuna's* coordinates.

For the exploratory analysis, the following financial and sociodemographic data were obtained from the official webpage of the City of Buenos Aires (7):

- Population density: Total number of people per square kilometer.

- Crime: Number of crimes registered during the year 2019 in the city.

- Overcrowding situation: Percentage of households from the year 2010 to 2018, with respect to their overcrowding situation as follows: No Overcrowding: Less than two people per room; Overcrowding non-critical: two to three people per room; Overcrowding critical: more than three people per room.

- Employment rate: Employment and unemployment rates in 2019 for each *comuna* in the city.

- Subway data: Location of subway stations and passengers per station in 2019.

- Health budget: Financial resources allocated to health topics in 2019 in the city.

Additionally, **the Foursquare API** was used to determine the similarities and most common venues of the clusters obtained from the machine learning algorithms.

2.2 Preprocessing and data cleaning

2.2.1 Main data

As said above, the main data collected consisted in the COVID-19 cases in the city of Buenos Aires and the geographical location of each *barrio* (neighborhood) and *comunas.*

For the geographical data, two data frames were created consisting basically of the geographical coordinates for each *barrio* and *comuna* of the city of Buenos Aires.

	barrio	comuna	latitude	longitude
0	AGRONOMIA	15	-34.595041	-58.494293
1	ALMAGRO	5	-34.600000	-58.416667
2	BALVANERA	3	-34.610500	-58.397600
3	BARRACAS	4	-34.650000	-58.383333
4	BELGRANO	13	-34.562500	-58.458333

	comuna	latitude	longitude
0	1	-34.608196	-58.377771
1	2	-34.590556	-58.390556
2	3	-34.613754	-58.403687
3	4	-34.648503	-58.390695
4	5	-34.618314	-58.420882

Figure 1. Barrios and comuna data frames showing the first five rows

The data related to the COVID-19 cases was scraped from the webpage of La Nación, and after dropping unnecessary rows and matching the names of the barrios, the data was merged with the above data frames.

	comuna	latitude	longitude	Total Cases	Total Population	Cases per 100,000
0	1	-34.608196	-58.377771	3571	206227	1731.59
1	2	-34.590556	-58.390556	432	158648	272.30
2	3	-34.613754	-58.403687	1166	187799	620.88
3	4	-34.648503	-58.390695	2280	217605	1047.77
4	5	-34.618314	-58.420882	658	179366	366.85

Figure 2. Main data frame

2.2.2 Data for exploratory analysis

The population density data frame includes the area in squared kilometers of each *comuna* and the number of people per squared kilometer.

	comuna	Total Cases	Total Population	area km	Cases per 100,000	Nr. people/km2
0	1	3571	206227	17.376	1731.587	11868.497
1	2	432	158648	6.317	272.301	25114.453
2	3	1166	187799	6.386	620.877	29407.924
3	4	2280	217605	21.684	1047.770	10035.279
4	5	658	179366	6.661	366.848	26927.789

Figure 3. Population density data frame

The crime data frame includes the number of crimes that were reported for each *comuna* and the crime rate per 100,000 people.

	comuna	Total Cases	Total Population	Cases per 100,000	number of crimes	Nr. crimes per 100,000
0	1	3571	206227	1731.59	18876	9153.02
1	2	432	158648	272.30	5589	3522.89
2	3	1166	187799	620.88	11135	5929.21
3	4	2280	217605	1047.77	9921	4559.18
4	5	658	179366	366.85	6720	3746.53

Figure 4. Crime reported data frame

The data frame related to the overcrowding situation includes the percentage of households which are either not overcrowded or overcrowded.

	comuna	Total Cases	Total Population	Cases per 100,000	No Overcrowding	Overcrowding
0	1	3571	206227	1731.59	79.69	20.29
1	2	432	158648	272.30	94.91	5.09
2	3	1166	187799	620.88	85.53	14.46
3	4	2280	217605	1047.77	79.72	20.28
4	5	658	179366	366.85	92.61	7.40

Figure 5. Overcrowding situation data frame

The employment situation data frame contains the following employment indicators: Economic activity rate (EAR), Employment rate (ER), Unemployment rate (UR), Time-related underemployment (TRU).

	comuna	Total Cases	Total Population	Cases per 100,000	EAR	ER	UR	TRU
0	1	3571	206227	1731.59	66.8	61.0	8.7	9.4
1	2	432	158648	272.30	65.4	61.7	5.7	10.2
2	3	1166	187799	620.88	63.4	58.9	7.1	11.0
3	4	2280	217605	1047.77	61.5	55.1	10.5	18.0
4	5	658	179366	366.85	65.9	62.2	5.6	10.1

Figure 6. Employment situation data frame

The subway-passengers data frame shows the total number of subway passengers in 2019 for each *comuna*.

	comuna	Total Cases	Total Population	Cases per 100,000	Passengers 2019
0	1	3571	206227	1731.59	119825201
1	2	432	158648	272.30	23368599
2	3	1166	187799	620.88	49166476
3	4	2280	217605	1047.77	11705224
4	5	658	179366	366.85	20077830

Figure 7. Subway-passengers data frame

The health budget data frame includes the financial resources (in million ARS) allocated to health topics.

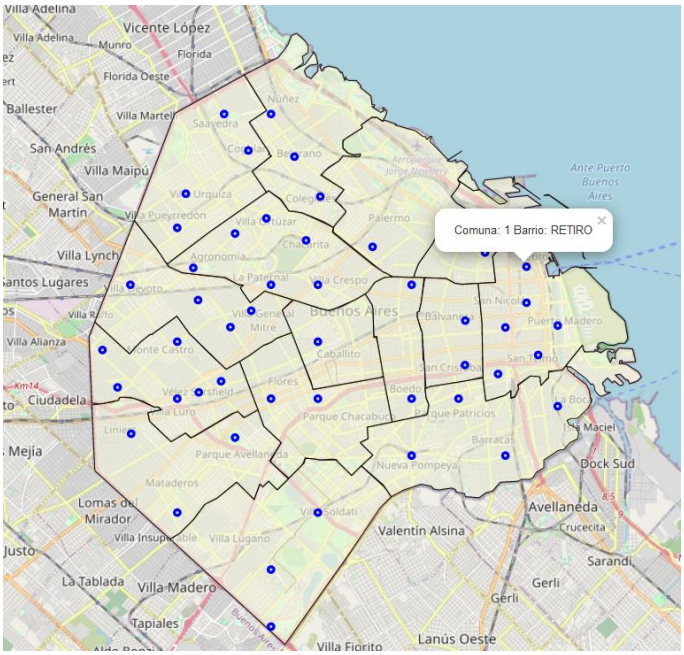
	comuna	Total Cases	Total Population	Cases per 100,000	Given budget (in Mill. ARS)
0	1	3571	206227	1731.59	2711.15
1	2	432	158648	272.30	2607.56
2	3	1166	187799	620.88	3629.11
3	4	2280	217605	1047.77	20292.03
4	5	658	179366	366.85	352.54

Figure 8. Health budget data frame

3. Methodology

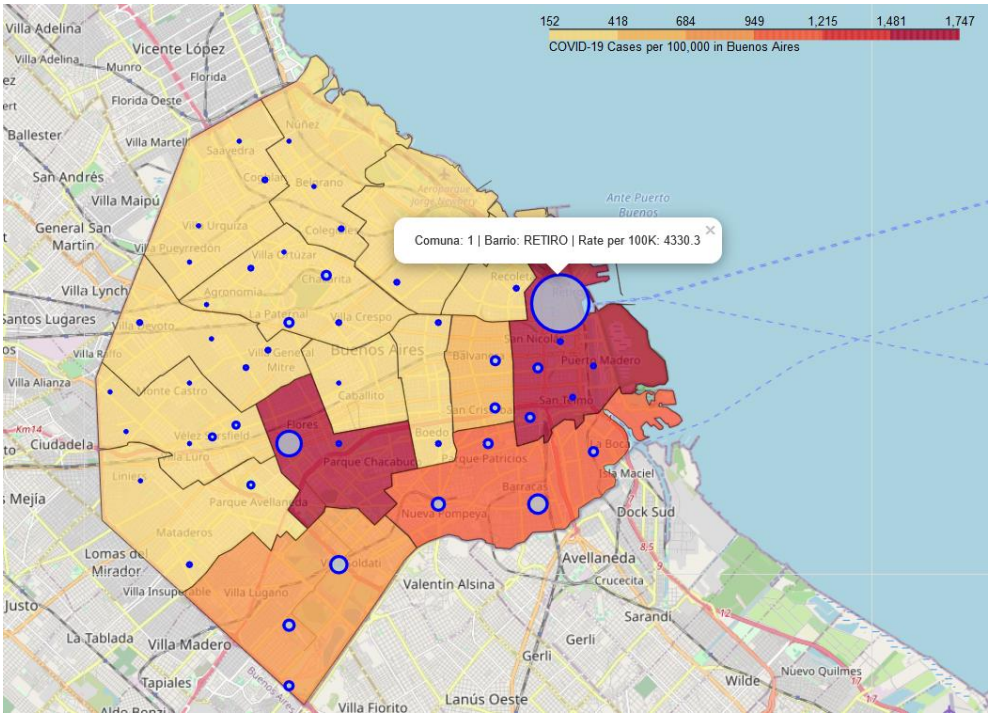
3.1 Exploratory Analysis

As it was mentioned, the city of Buenos Aires comprises 48 *barrios* that are part of one of the 15 *comunas* in which the city is administratively divided.



Map 1. The comunas with their barrios

The below heatmap shows the *comunas* that have been affected the most by the COVID-19 pandemic and which *barrios* have the highest rate of cases per 100,000 people.



Map 2. Rate of COVID-19 cases per 100.000

The data in section 2.2.2 had to be analyzed to get a better understanding and identify whether they have a strong or weak correlation to the number of COVID-19 cases in the *comunas* of the city of Buenos Aires, specifically the rate of COVID-19 cases per 100,000 people.

The Pearson Correlation was used to determine whether a variable would be used for the clustering algorithm. The threshold value was set to 0.60 for positive correlations and -0.60 for negative correlations.

Population density

It can be observed that the variables area km and Nr. people/km2 do not have a strong correlation with the rate of COVID-19 cases per 100,000 people, and they have a value below the defined threshold.

	comuna	Total Cases	Total Population	area km	Cases per 100,000	Nr. people/km2
comuna	1.000000	-0.501828	0.139809	0.281266	-0.532641	-0.461997
Total Cases	-0.501828	1.000000	0.547773	0.352946	0.996651	-0.197147
Total Population	0.139809	0.547773	1.000000	0.448498	0.494380	-0.282402
area km	0.281266	0.352946	0.448498	1.000000	0.347672	-0.932730
Cases per 100,000	-0.532641	0.996651	0.494380	0.347672	1.000000	-0.190023
Nr. people/km2	-0.461997	-0.197147	-0.282402	-0.932730	-0.190023	1.000000

Figure 9. Pearson correlation - Population density

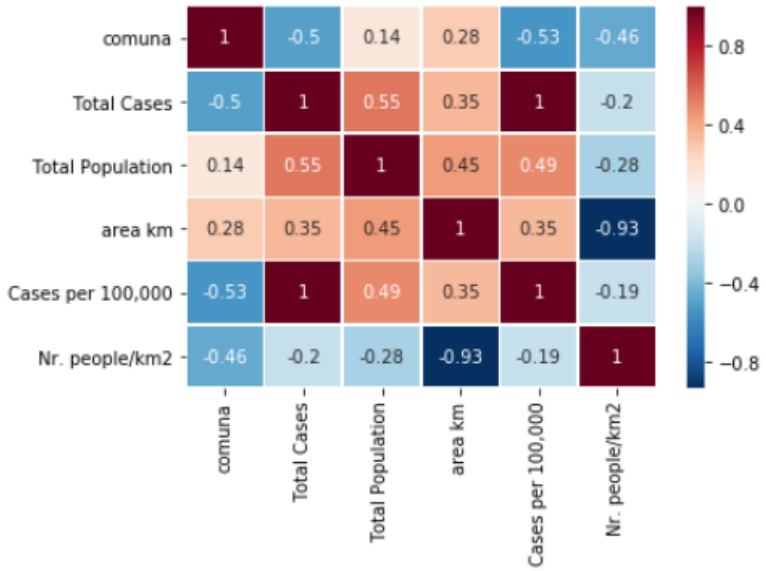


Figure 10. Pearson correlation heatmap - Population density

Crime reported

It can be observed that the variable number of crimes has a strong correlation with the rate of COVID-19 cases per 100,000 people, and it has a value above the defined threshold.

	comuna	Total Cases	Total Population	Cases per 100,000	number of crimes	Nr. crimes per 100,000
comuna	1.000000	-0.501828	0.139809	-0.532640	-0.452757	-0.528569
Total Cases	-0.501828	1.000000	0.547773	0.996651	0.699621	0.631615
Total Population	0.139809	0.547773	1.000000	0.494380	0.484019	0.290636
Cases per 100,000	-0.532640	0.996651	0.494380	1.000000	0.708149	0.653862
number of crimes	-0.452757	0.699621	0.484019	0.708149	1.000000	0.977146
Nr. crimes per 100,000	-0.528569	0.631615	0.290636	0.653862	0.977146	1.000000

Figure 11. Pearson correlation - Crime reported

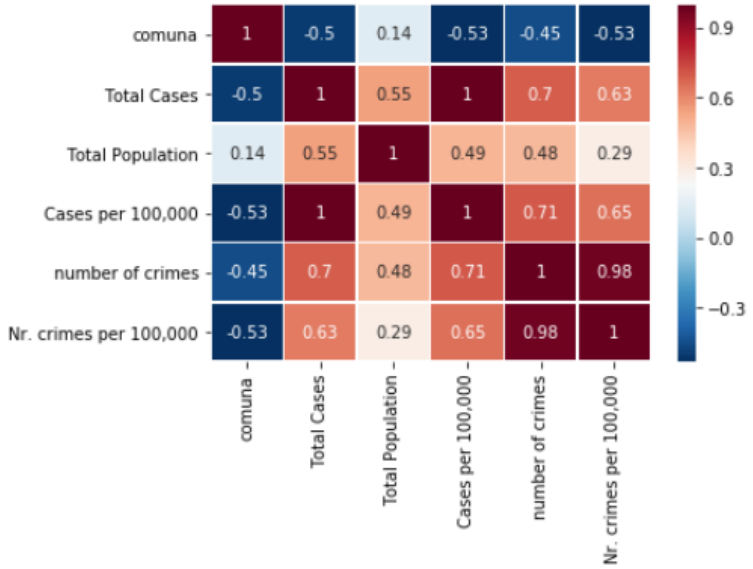


Figure 12. Pearson correlation heatmap - Crime reported

Overcrowding situation

It can be observed that the variables No Overcrowding and Overcrowding have a strong correlation with the rate of COVID-19 cases per 100,000 people, and both have a value above the defined threshold. However, these variables are inversely proportional to each other and only one will be selected for the clustering algorithm.

	comuna	Total Cases	Total Population	Cases per 100,000	No Overcrowding	Overcrowding
comuna	1.000000	-0.501828	0.139809	-0.532640	0.480693	-0.480733
Total Cases	-0.501828	1.000000	0.547773	0.996651	-0.760066	0.759818
Total Population	0.139809	0.547773	1.000000	0.494380	-0.273772	0.273984
Cases per 100,000	-0.532640	0.996651	0.494380	1.000000	-0.790428	0.790152
No Overcrowding	0.480693	-0.760066	-0.273772	-0.790428	1.000000	-0.999998
Overcrowding	-0.480733	0.759818	0.273984	0.790152	-0.999998	1.000000

Figure 13. Pearson correlation - Overcrowding situation

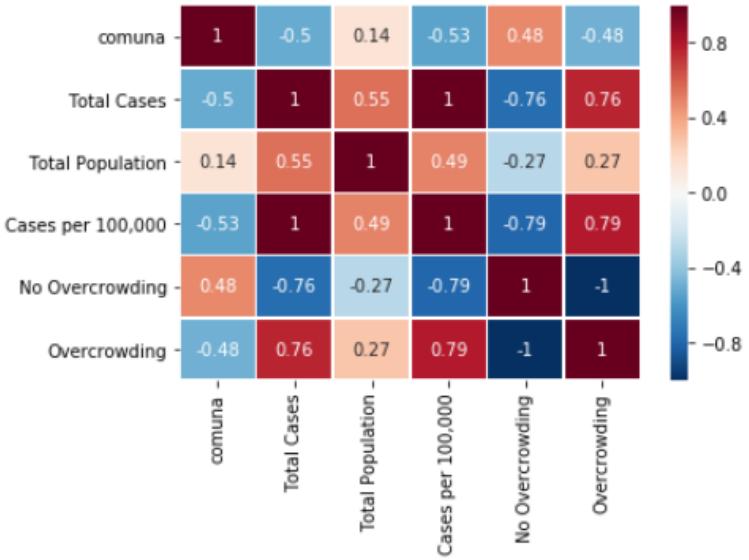


Figure 14. Pearson correlation heatmap - Overcrowding situation

Employment situation

It can be observed that the Employment variables (EAR, ER, UR, TRU) do not have a strong correlation with the rate of COVID-19 cases per 100,000 people, and they have a value below the defined threshold.

	comuna	Total Cases	Total Population	Cases per 100,000	EAR	ER	UR	TRU
comuna	1.000000	-0.501828	0.139809	-0.532640	0.201410	0.137483	-0.004705	-0.132844
Total Cases	-0.501828	1.000000	0.547773	0.996651	-0.262872	-0.337758	0.379761	0.358490
Total Population	0.139809	0.547773	1.000000	0.494380	0.094443	0.039933	0.063705	0.040490
Cases per 100,000	-0.532640	0.996651	0.494380	1.000000	-0.291778	-0.369389	0.410369	0.374098
EAR	0.201410	-0.262872	0.094443	-0.291778	1.000000	0.955940	-0.684111	-0.668330
ER	0.137483	-0.337758	0.039933	-0.369389	0.955940	1.000000	-0.867013	-0.767138
UR	-0.004705	0.379761	0.063705	0.410369	-0.684111	-0.867013	1.000000	0.767704
TRU	-0.132844	0.358490	0.040490	0.374098	-0.668330	-0.767138	0.767704	1.000000

Figure 15. Pearson correlation - Employment situation

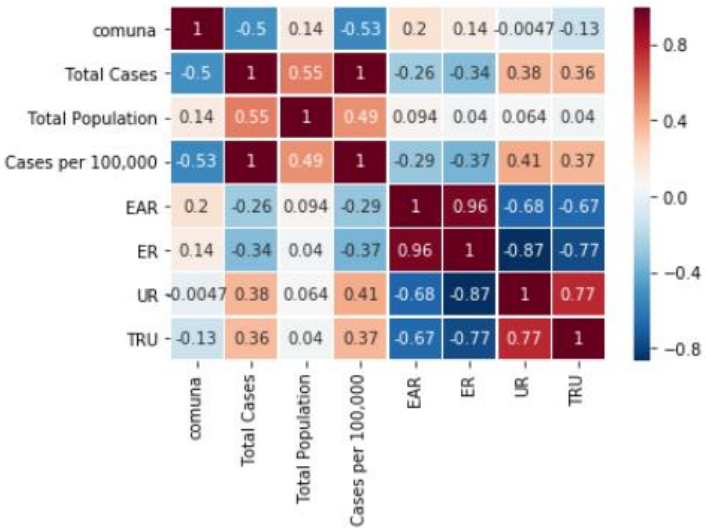


Figure 16. Pearson correlation heatmap - Employment situation

Subway passengers

It can be observed that the variable Passengers 2019 has a moderate correlation with the rate of COVID-19 cases per 100,000 people, and it has a value above the defined threshold.

	comuna	Total Cases	Total Population	Cases per 100,000	Passengers 2019
comuna	1.000000	-0.501828	0.139809	-0.532640	-0.482099
Total Cases	-0.501828	1.000000	0.547773	0.996651	0.596052
Total Population	0.139809	0.547773	1.000000	0.494380	0.297906
Cases per 100,000	-0.532640	0.996651	0.494380	1.000000	0.612131
Passengers 2019	-0.482099	0.596052	0.297906	0.612131	1.000000

Figure 17. Pearson correlation - Subway passengers

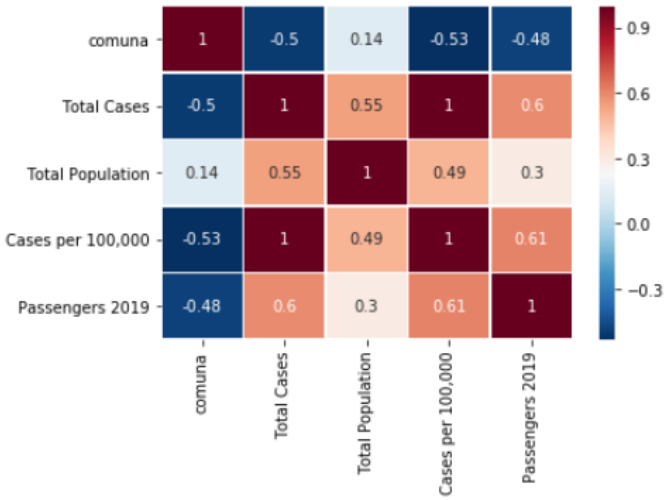


Figure 18. Pearson correlation heatmap - Subway passengers

Health budget

It can be observed that the variables Given budget does not have a strong correlation with the rate of COVID-19 cases per 100,000 people, and it has a value below the defined threshold.

	comuna	Total Cases	Total Population	Cases per 100,000	Given budget (in Mill. ARS)
comuna	1.000000	-0.501828	0.139809	-0.532640	-0.257008
Total Cases	-0.501828	1.000000	0.547773	0.996651	0.342426
Total Population	0.139809	0.547773	1.000000	0.494380	0.463182
Cases per 100,000	-0.532640	0.996651	0.494380	1.000000	0.305201
Given budget (in Mill. ARS)	-0.257008	0.342426	0.463182	0.305201	1.000000

Figure 19. Pearson correlation - Health budget

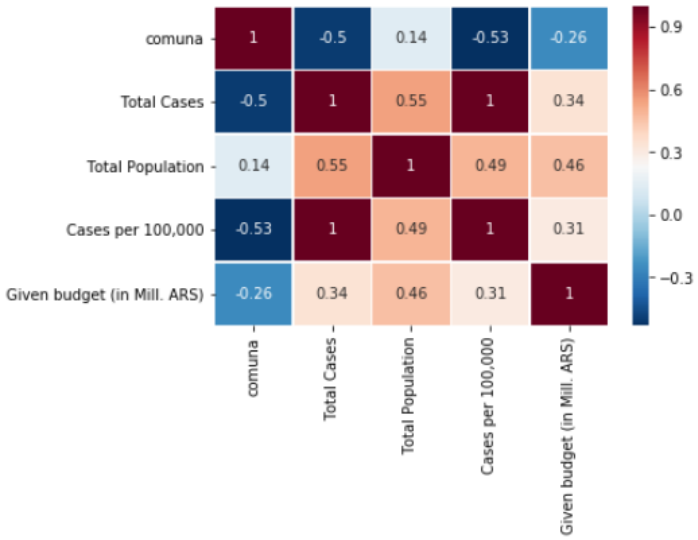


Figure 20. Pearson correlation heatmap - Health budget

3.2 Data selection

After analyzing the above-mentioned variables regarding their respective correlation with the rate of COVID-19 cases per 100,000 people and the defined threshold, the following variables were selected as input to the clustering algorithm:

Variable	Correlation value	Threshold
Not in overcrowding situation (avg. 2010-2019)	-0.79	-0.60
Number of reported crimes in 2019	0.71	0.60
Subway passengers in 2019	0.61	0.60

Table 1. Variables selected

3.3 Modeling

The k-means algorithm was used to cluster the *comunas* of the city of Buenos Aires in order to identify how similar these are with respect to the other ones inside the same cluster and how different they are with respect to the *comunas* in other clusters.

The k-means algorithm has the particularity that the number of clusters to be used needs to be given so it can run the code. Therefore, before training the model the best K (optimal number of clusters) had to be identified using the elbow method and the normalized data of the data frame containing the variables: Cases per 100,000, No overcrowding, Number of crimes, and Passengers 2019. As a result, the best K was 3.

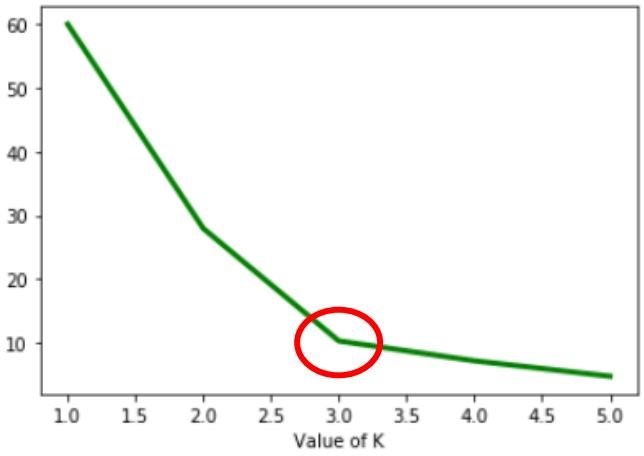


Figure 21. Elbow method - Identifying the best K

Once the best K was identified, the clustering algorithm was executed and the 15 *comunas* of the city of Buenos Aires were grouped in three different clusters.

	comuna	Cases per 100,000	No Overcrowding	number of crimes	Passengers 2019	Cluster
0	1	1731.59	79.69	18876	119825201	0
1	2	272.30	94.91	5589	23368599	1
2	3	620.88	85.53	11135	49166476	2
3	4	1047.77	79.72	9921	11705224	2
4	5	366.85	92.61	6720	20077830	1

Figure 22. Cluster labels assigned to the comunas

Foursquare API

The Foursquare API was used to determine the similarities and most common venues of the clusters obtained from the machine learning algorithms. Using a radius value of 1000, the venues for each *barrio* were first extracted, then the results were grouped by *comuna*, and the mean of the frequency of occurrence of each venue category was calculated. Finally, the cluster labels from the k-means algorithm were included and used to group the results of the data frame.

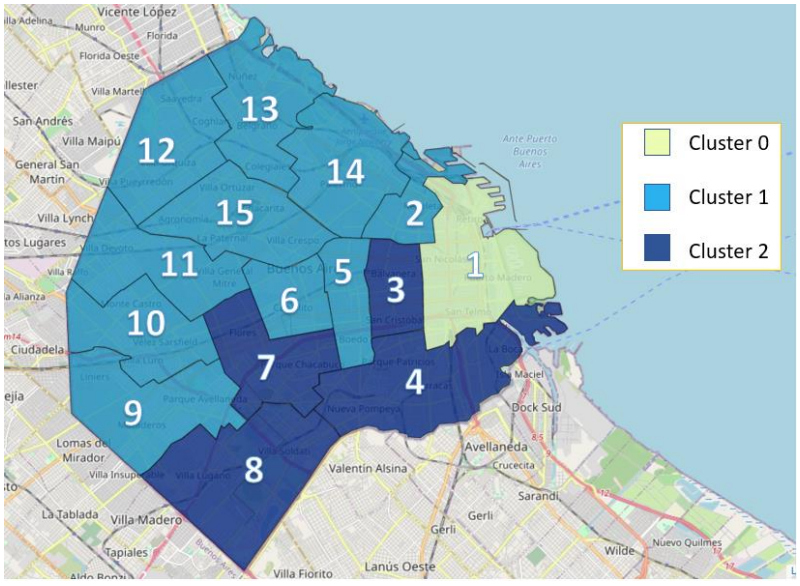
	Cluster	Accessories Store	Adult Boutique	American Restaurant	Amphitheater	Arcade	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	...
0	0	0.000	0.000000	0.001953	0.000	0.000000	0.078125	0.007812	0.003906	0.000000	...
1	1	0.002	0.000000	0.001373	0.001	0.000931	0.064486	0.003491	0.001333	0.000246	...
2	2	0.000	0.001582	0.000000	0.000	0.000000	0.054780	0.004934	0.001645	0.003227	...

3 rows × 242 columns

Figure 23. Mean of occurrence of each type of venue grouped by cluster

4. Results

As it was previously shown in Figure 22, the 15 *comunas* of the city were segmented into three different clusters, based on the following criteria: COVID-19 cases per 100.000 people, percentage of households in non-overcrowded situation, number of crimes registered in 2019, and number of subway passengers in 2019; as shown in the map below.



Map 3. Comunas grouped by cluster

Grouping by cluster labels the final data frame (Figure 22) and obtaining the mean for each column, the following observations can be inferred:

- Cluster 0 is composed solely of *comuna* 1
    - Rate of COVID-19 cases per 100,000 people: 1,731.6
    - Percentage of No Overcrowding: 79.7%
    - Number of crimes registered in 2019: 18,876
    - Number of subway passengers in 2019: 119,825,201

- Cluster 1 is composed of the following *comunas*: 2, 5, 6, 9, 10, 11, 12, 13, 14, and 15
    - Rate of COVID-19 cases per 100,000 people: 262.8
    - Percentage of No Overcrowding: 93.5%
    - Number of crimes registered in 2019: 6,329.6
    - Number of subway passengers in 2019: 13,752,521

- Cluster 2 is composed of the following *comunas*: 3, 4, 7, and 8
    - Rate of COVID-19 cases per 100,000 people: 1,028.2
    - Percentage of No Overcrowding: 82.7%
    - Number of crimes registered in 2019: 8,724.8
    - Number of subway passengers in 2019: 20,997,522.5

As presented in figure 23, the **foursquare** results show the frequency of occurrence of each type of venue in every cluster. In the following table, the most recurrent venues are displayed.

	Cluster	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	0	Hotel	Argentinian Restaurant	Coffee Shop	Café	Theater	Italian Restaurant	Bar	Restaurant	BBQ Joint	Hostel
1	1	Argentinian Restaurant	Pizza Place	Café	Ice Cream Shop	Bakery	Coffee Shop	Plaza	Restaurant	BBQ Joint	Hotel
2	2	Pizza Place	Café	Argentinian Restaurant	Ice Cream Shop	Racetrack	Supermarket	Recreation Center	Gym	Plaza	Restaurant

Figure 24. Most common venues in each cluster

5. Discussion

Based on the results, it can be determined that Cluster 0 has the highest rate of COVID-19 cases, the highest number of crimes registered and subway passengers during the last year; the data of the same cluster also show the highest percentage of households facing an overcrowding situation. On the other hand, the lowest rate of COVID-19 cases corresponds to Cluster 1, which comprises the majority of the *comunas*, and displays the lowest number of crimes and subway passengers and the highest percentage of households not overcrowded. Finally, the data obtained for Cluster 2, positions its values in between the previous clusters.

Regarding the foursquare results, hotels and hostels, restaurants, coffee places, theaters, and bars are displayed as the most common venues in Cluster 0. It should be noted that these types of venues normally gather a significant amount of people per day, leading to potentially large crowds in confined spaces, thus facilitating the spread of a disease, in this case the COVID-19.

6. Conclusion

By using a machine learning algorithm, the 15 *comunas* of the city of Buenos Aires could be grouped into three different clusters, based on the rate of COVID-19 cases per 100, 000 people and other social variables, that have not been covered in the health reports presented by the government offices and traditional media.

The data were analyzed, and important correlations were revealed. From the three clusters identified, Cluster 0 displays the highest rate of COVID-19 cases, so it might be considered as a potential infection epicenter in the city; this cluster also encompass the highest number of reported crimes and the highest proportion of overcrowded households, which may indicate a direct relationship between the spread of the COVID-19 infection and socioeconomic variables that affect the quality of living. Additionally, Cluster 0 has by far the largest number of subway passengers which then transit to other areas of the city, making it a potential risk for the transmission of the virus; it is also noted that its most common venues are places such as hotels, restaurants, and theaters; that is, indoor locations where large groups of people come together, aggravating the risk of infection.

The results shown in this project may be beneficial for social and health organizations, as well as government offices. Further collection of data related to the number of passengers on different types of public transportation, as well as data associated to the number of people in geriatric centers, hospitals and schools, would enrich the analysis and provide more variables for the clustering algorithm, facilitating further research.

This tool has the potential to uncover valuable insights regarding the relationship between variables that might not seem relevant at the beginning, but that with further scrutiny might reveal correlations that may aid future explorations and social interventions.

7. References

1. Confirmaron el primer caso de coronavirus en la Argentina [Internet]: Infobae [cited 2020 June 17]. Available from: https://www.infobae.com/coronavirus/2020/03/03/confirmaron-el-primer-caso-de-coronavirus-en-la-argentina/

2. Garrison, C. Argentina announces mandatory quarantine to curb coronavirus [Internet]. Buenos Aires (AR): Reuters; 2020 [cited 2020 June 17]. Available from: https://www.reuters.com/article/us-health-coronavirus-argentina/argentina-announces-mandatory-quarantine-to-curb-coronavirus-idUSKBN216446

3. Equipo de Epidemiología del Nivel Central de Abordaje de COVID19. Boletín epidemiológico semanal - Ciudad Autónoma de Buenos Aires [Internet]. Buenos Aires (AR): Ministerio de Salud de la Ciudad Autónoma de Buenos Aires; 2020 June 19 [cited 2020 June 25]. (52 p.) Report No. 199 Año V. Available from: https://www.buenosaires.gob.ar/sites/gcaba/files/bes_200_se_23_vf.pdf

4. La distribución de la pandemia en Capital Federal [Internet]. Buenos Aires (AR): La Nación; 2020 [cited 2020 June 17]. Available from: https://www.lanacion.com.ar/sociedad/coronavirus-caba-mapa-pandemia-comuna-comuna-nid2364565

5. Comunas de la ciudad de Buenos Aires [Internet]: Wikipedia; [updated 2020 April 09; cited 2020 June 17]. Available from: https://es.wikipedia.org/wiki/Comunas_de_la_ciudad_de_Buenos_Aires

6. Comuna 1 (Ciudad de Buenos Aires) [Internet]: GeoHack; [cited 2020 June 17]. Available from: https://tools.wmflabs.org/geohack/geohack.php?language=es&pagename=Comuna_1_(Ciudad_de_Buenos_Aires)&params=-34.600022222222_N_-58.386916666667_E_type:city

7. Datasets [Internet]: Buenos Aires (AR): Buenos Aires Ciudad; [cited 2020 June 17]. Available from: https://data.buenosaires.gob.ar/dataset