UNIVERSIDAD PANAMERICANA
Campus México

UNIVERSIDADPANAMERICANA

Subject: Gestión de Análisis y Diseño de Comercialización(COM145)

Professor: Sarahí Aguilar González

Date of Delivery: 23/05/2022

Season: 1222

Project Name: Patrones en las defunciones previo al inicio del periodo de vacunación de COVID-19 en la CDMX

Miembros del Equipo		
ID	**Nombre**	**Carrera**
0209486	Brian Antonio Aranda Mejía	LITSC
0207950	Gonzalo Ronzón Carniado	LITSC
0203280	Pablo Mieres Noriega	LITSC

Rúbricas				
ID	**2-social**		**7-knowledge**	
	D	**C**	**A**	**JI**

Abstract

An analysis of the data set of death certificates from Mexico was carried out, specifically from Mexico City between the start of the COVID-19 pandemic in national territory and the start of the vaccination scheme in the city. Python data science tools such as pandas and scikit learn were used for this. This analysis entailed data manipulation and visualization as well as the design of an unsupervised Machine Learning model, with which we were unable to reach a precise conclusion about our research question. Thanks to the visualization of data, we were able to realize that the most vulnerable groups are men of advanced age, more precisely those who are between 60 and 80 years of age in the municipalities of Venustiano Carranza, Iztacalco and Azcapotzalco.

Introducción

On February 27, 2020, the first case of COVID-19 was registered in Mexico, starting the pandemic in the national territory. Since then, the country has suffered a great deal economically and socially. [1]

In total, there have been nearly 6 million infections and more than 300,000 deaths. Therefore, it is important to know which groups have been most vulnerable to this virus. [2]

On December 24, 2020, the first dose of the vaccine against SARS-COV-2 was applied, which marked the beginning of the vaccination scheme. [3]

Vaccines have helped greatly reduce the mortality of the virus. For this reason, our research will focus on the period before the start of the vaccination scheme.

The great idea

Among those killed by COVID-19 there are a variety of factors that influence a person to die, so we seek to know the most common patterns. The analysis period will be between the start of the pandemic in Mexico and the start of the vaccination scheme in Mexico City.

Research question

Which groups were the most vulnerable against COVID-19 in the CDMX before the start of the vaccination scheme, taking into account factors such as their sex, age, state and comorbidities?

Relevant data sources:

- Actas de defunción CDMX (Deaths since 2017)

Inconvenience:

- Actas de defunción CDMX
  - Has no comorbidities
  - They are deaths since before the COVID
  - Several states of the republic come

Dependent and independent variables

- **Dependent**
  - N/A
- **Independent**
  - City hall
  - Age
  - Comorbidities
  - Sex
  - Place of death
  - Date of death

Model to implement

In conducting this research we will not predict a variable, but we will find patterns that can define different vulnerable groups in the CDMX. In other words, our model will classify these vulnerable groups, so it is an unsupervised model.

In this sense, the priority will be the inference, since we want to deduce these characteristics to classify the groups and to be able to infer the vulnerable groups.

Development

We found a couple similar projects to ours:

- https://ellis.eu/covid-19/projects#ai-against-covid-19-mila
- https://github.com/sarahiaguilar/fundamentos-cdd/blob/main/notebooks/extra/Workshop_DSci_and_INEGI_RIIA_2021.ipynb
- https://www.frontiersin.org/articles/10.3389/fpubh.2021.602353/full

To answer the research question, we will rely on the following tools:

- Python 3.7.13
- Pandas 1.3.5
- Numpy 1.21.6
- Plotnine 0.6.0
- Scikit learn 1.0.2
- Matplotlib 3.2.2

The first step is to get the data. To do that, we're going to use the death certificates from the Mexican government. The URL of the CSV file is as follows:

https://datos.cdmx.gob.mx/dataset/19e094a0-f1c0-4544-bac6-dd1d5cb8a4de/resource/d683ec6e-171a-4825-a523-2cdbf30f9894/download/defunciones_corte_110322.csv

The way we read and store the data is through the Pandas library. The dataset data comes with the following structure:

1. **Age - float64**

   Age of the deceased.
2. **Sex - String**

Sex of the deceased.

3. **Date of death - String**

   Date of death in format YYYY-MM-DD
4. **State - String**

   The state in which the person died.
5. **Cause - String**

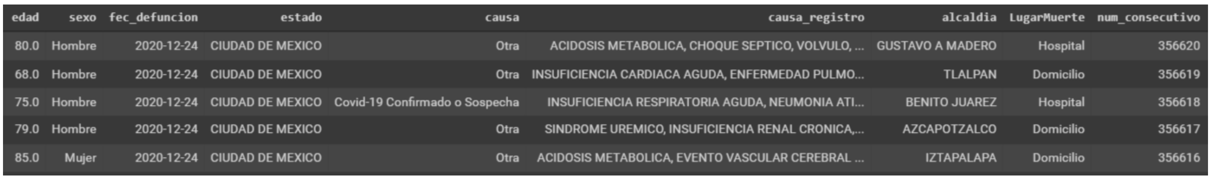
   Determination of whether he died from COVID-19 or other causes.
6. **Cause Record - String**

   Causes of death in comas separate format.
7. **City Hall - String**

   City hall in which he died. In case of being outside the CDMX is in NaN.
8. **Place of Death - String**

   Whether he died at home or in the hospital.
9. **Consecutive number - int64**

   Consecutive number of deaths.



edad	sexo	fec_defuncion	estado	causa	causa_registro	alcaldia	LugarMuerte	num_consecutivo
80.0	Hombre	2020-12-24	CIUDAD DE MEXICO	Otra	ACIDOSIS METABOLICA, CHOQUE SEPTICO, VOLVULO, ...	GUSTAVO A MADERO	Hospital	356620
68.0	Hombre	2020-12-24	CIUDAD DE MEXICO	Otra	INSUFICIENCIA CARDIACA AGUDA, ENFERMEDAD PULMO...	TLALPAN	Domicilio	356619
75.0	Hombre	2020-12-24	CIUDAD DE MEXICO	Covid-19 Confirmado o Sospecha	INSUFICIENCIA RESPIRATORIA AGUDA, NEUMONIA ATI...	BENITO JUAREZ	Hospital	356618
79.0	Hombre	2020-12-24	CIUDAD DE MEXICO	Otra	SINDROME UREMICO, INSUFICIENCIA RENAL CRONICA,...	AZCAPOTZALCO	Domicilio	356617
85.0	Mujer	2020-12-24	CIUDAD DE MEXICO	Otra	ACIDOSIS METABOLICA, EVENTO VASCULAR CEREBRAL ...	IZTAPALAPA	Domicilio	356616

In total we have 85,904 entries in our dataset. Among the columns, we find several that have null values: Age, Cause of Registration and State. In those cases, we fill those null spaces with the string “No data” since we consider all records important.

Once the data is obtained, we can begin to manipulate it to clean it or extract new DataFrames that facilitate the visualization of data and their relationship. In this step we clean and filter the data to be able to analyze it later.

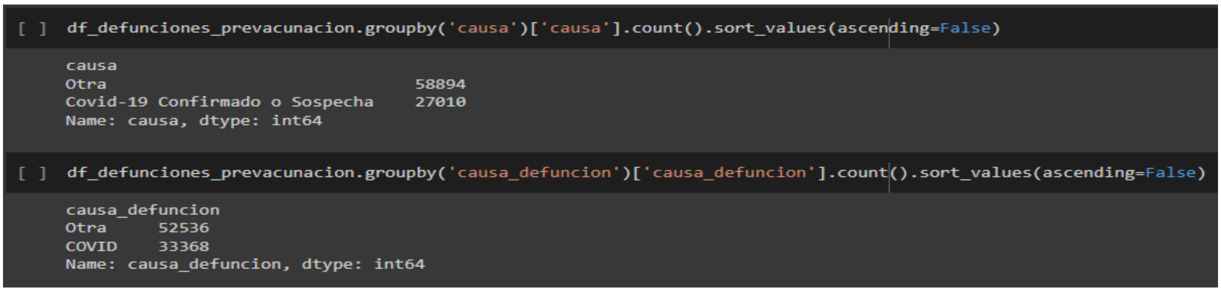
Exploratory analysis

The first thing we want to do is count the causes of death that exist in our dataset. To do this, we iterate on our pre-vaccination deaths dataset in Mexico City and analyze the causes of death associated with the patient. We store all of these causes of death within a dictionary for later analysis.

If we detect that any of the patient's causes of death have a keyword related to COVID-19 (COVID, SARS, Atypical Pneumonia, Respiratory Failure, etc.) then we add it to the dictionary under a single key that we call *COVID*, otherwise, we add the textual cause to the dictionary as a key. The values of the keys are the deaths associated with that cause.

One problem that we encountered when doing this is that the records vary greatly since the same causes of death could be written in a different way, which made it difficult to identify deaths associated with COVID-19.

After this procedure, we continue to perform a feature engineering, where we add a field to the Data Frame that we call *Cause of Death*. This cause is formed taking into account the analysis of keywords that we determined previously, since in several cases with atypical pneumonia it was not counted as a death from COVID. When comparing both numbers, we realize that there are indeed more deaths possibly related to COVID, specifically, we see that there is a difference of 6,358 deaths in our dataset.

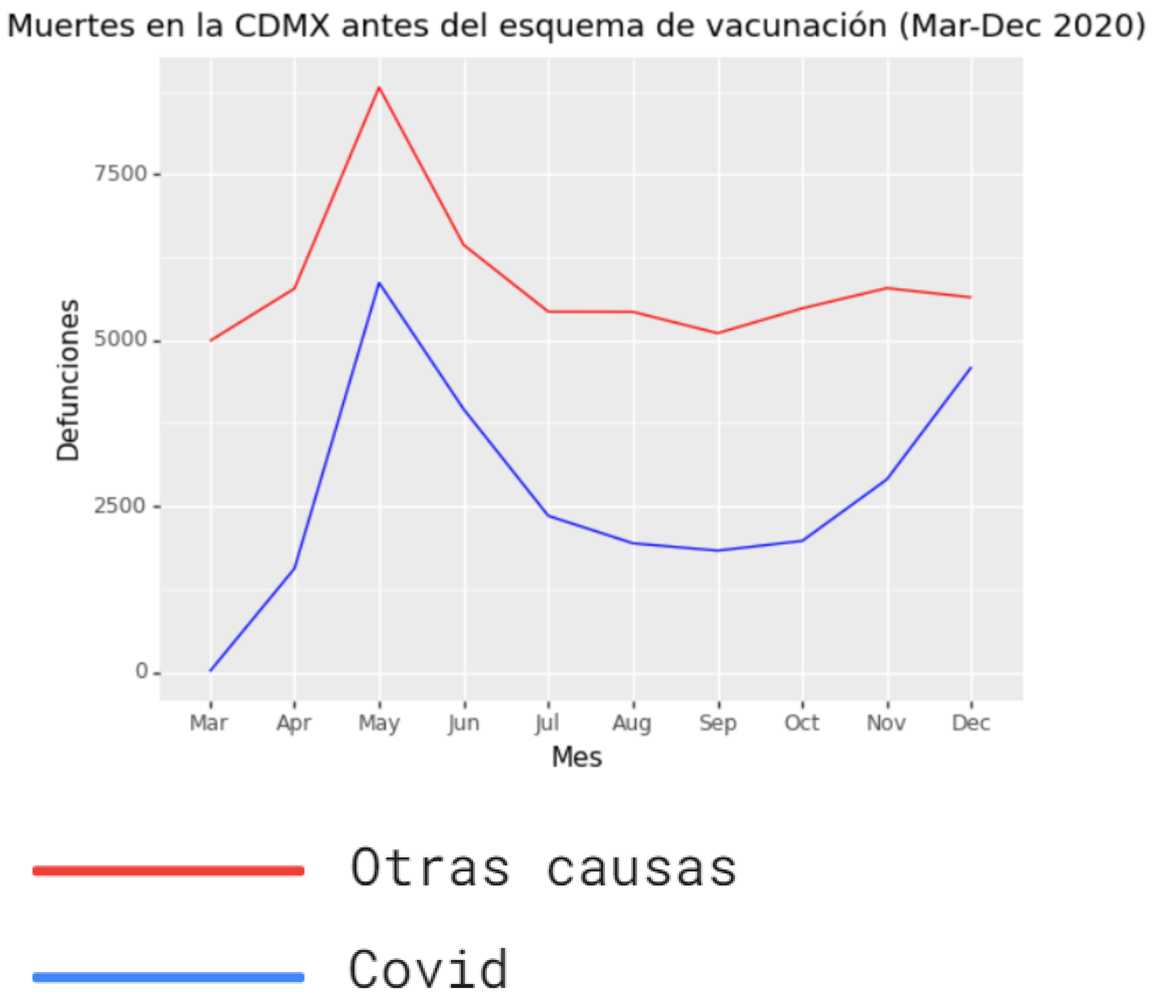


```
[ ]  df_defunciones_prevacunacion.groupby('causa')['causa'].count().sort_values(ascending=False)

     causa
     Otra                             58894
     Covid-19 Confirmado o Sospecha   27010
     Name: causa, dtype: int64

[ ]  df_defunciones_prevacunacion.groupby('causa_defuncion')['causa_defuncion'].count().sort_values(ascending=False)

     causa_defuncion
     Otra     52536
     COVID    33368
     Name: causa_defuncion, dtype: int64
```

The next analysis that we proceed to carry out is that of the dates. We define two dictionaries with all the deaths for each month. In one, they are deaths from COVID, in the other they are from the other causes. With this information, we perform feature engineering to define another column that determines the month of death, in order to view it more easily.

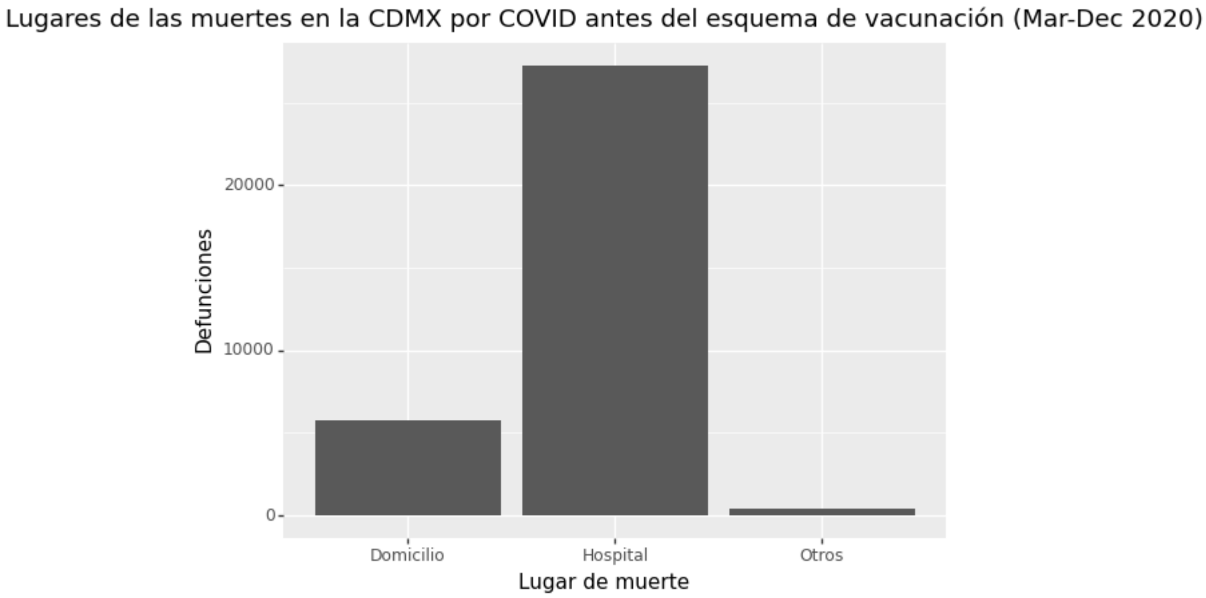
Data visualization

We use ggplot exclusively for data visualization. The first graph that we display is that of deaths per month in the CDMX before the vaccination scheme.

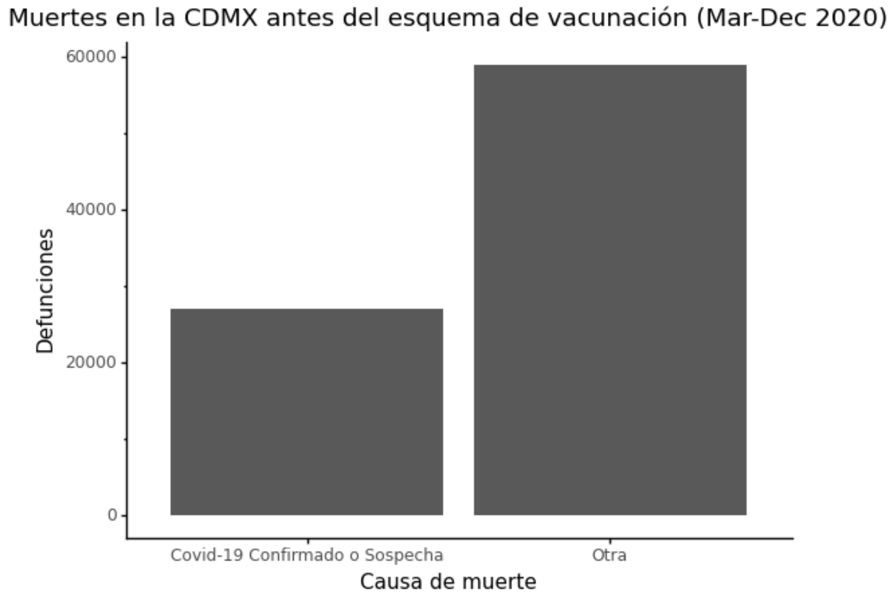


With this visualization we realize something curious, which is that deaths from COVID seem to be related to deaths from other causes, since both have approximately the same curve shape, but with different amounts.

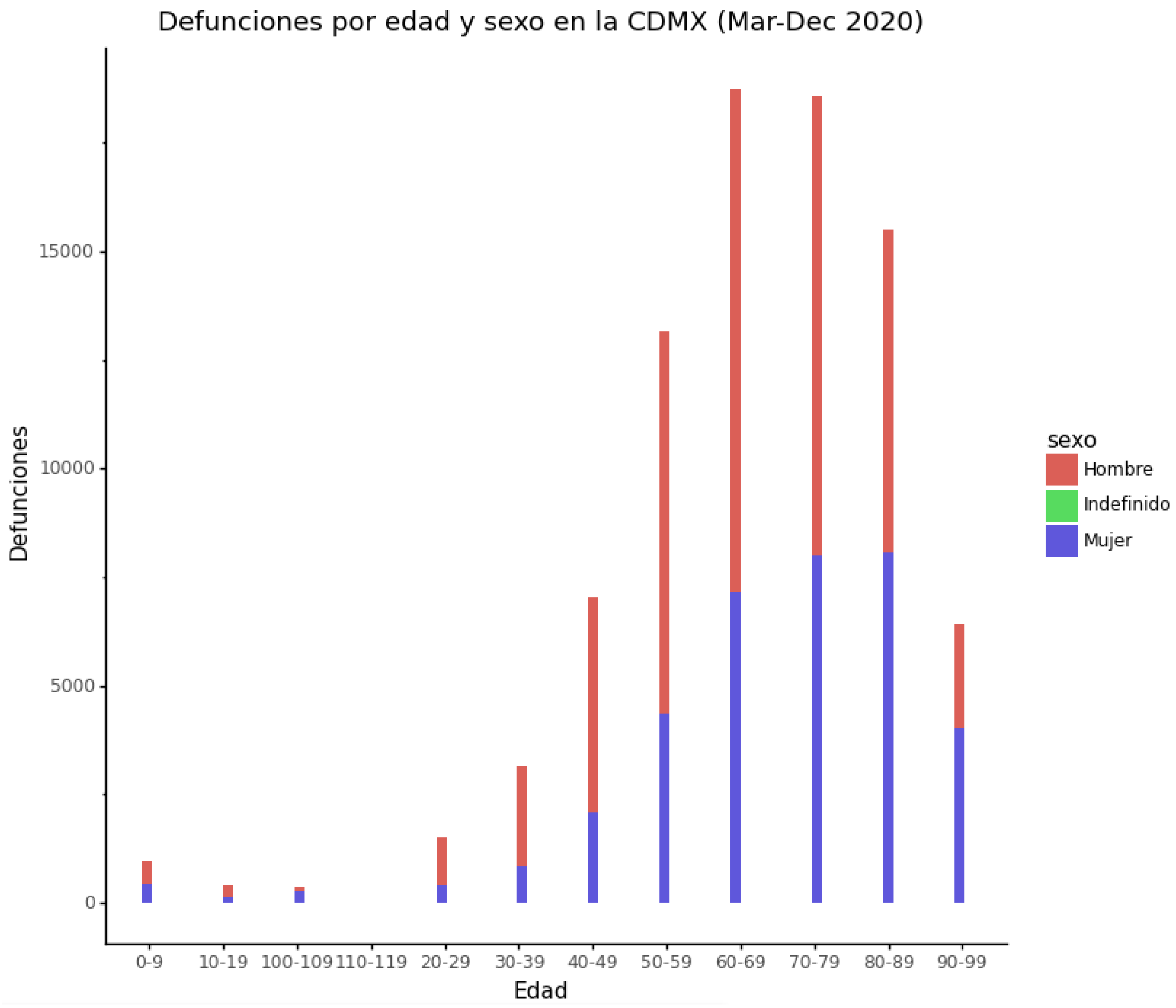
Subsequently, we graph the deaths that occurred in the three options of places: Home, Hospital or Other.



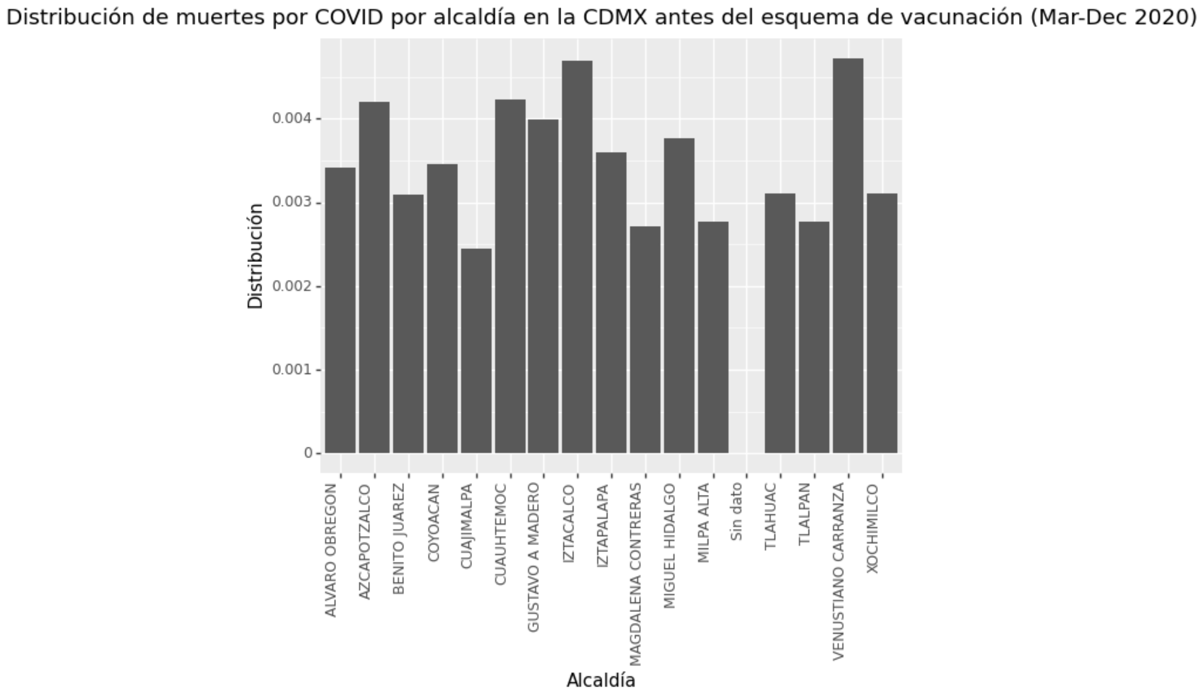
Another visualization that we consider relevant is the comparison between the number of deaths from COVID and from other causes. With this, we realized that the number of deaths from COVID reaches almost half of the deaths from all other causes.



The next visualization was the relationship between age and gender in terms of the number of deaths. Obviously, we find that the highest concentration of deaths is between 60 and 80 years. Regarding gender, it seems that men were more vulnerable.



Another important graph was that of the relationship of deaths from COVID by state. Here we divide the deaths by state by the population of that state to obtain the state with the highest percentage of deaths in relation to their population.



Implementation of Machine Learning model

As we mentioned at the beginning of this document, our Machine Learning model is going to be an unsupervised one, since we are not looking to predict a variable, but to find common characteristics between groups. That is why we are going to carry out two procedures: A Principal Component Analysis (PCA) and Clustering with k means.

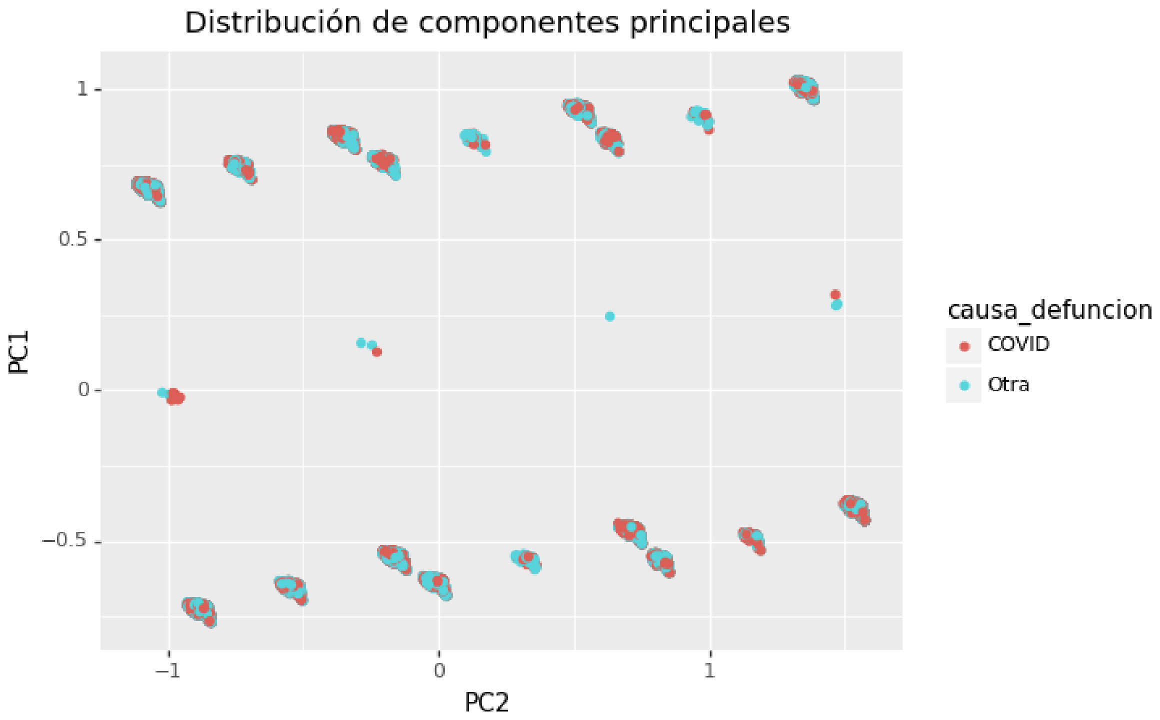
The first step in doing this is to split our data into a training set and a testing set. The ratio that we will use will be 80% and 20% respectively, with a random state of 1 to be able to replicate the results. To perform these operations we will need the Scikit learn library.

Once the data has been partitioned, we have to use the One Hot Encoder to encode our categorical variables (which is basically all of our dataset) into an array of pure 0's and 1's so that the model can properly identify the variables and their values in each row.

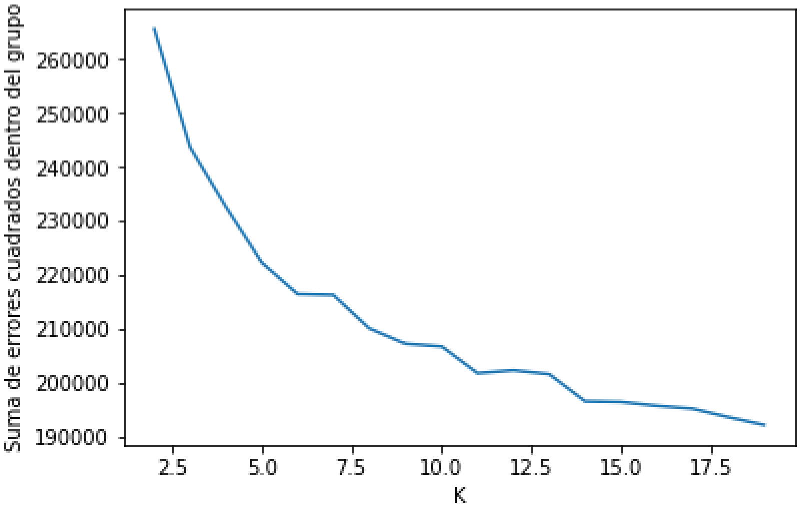
The result of this encoding operation gives us a *Sparse Matrix*. However, this type of array cannot be easily graphed or used for PCA, so we convert it to a numpy array first.

Following this, we proceed to perform the PCA with two components. This PCA is the one we are going to graph.

The result of this visualization is the following:

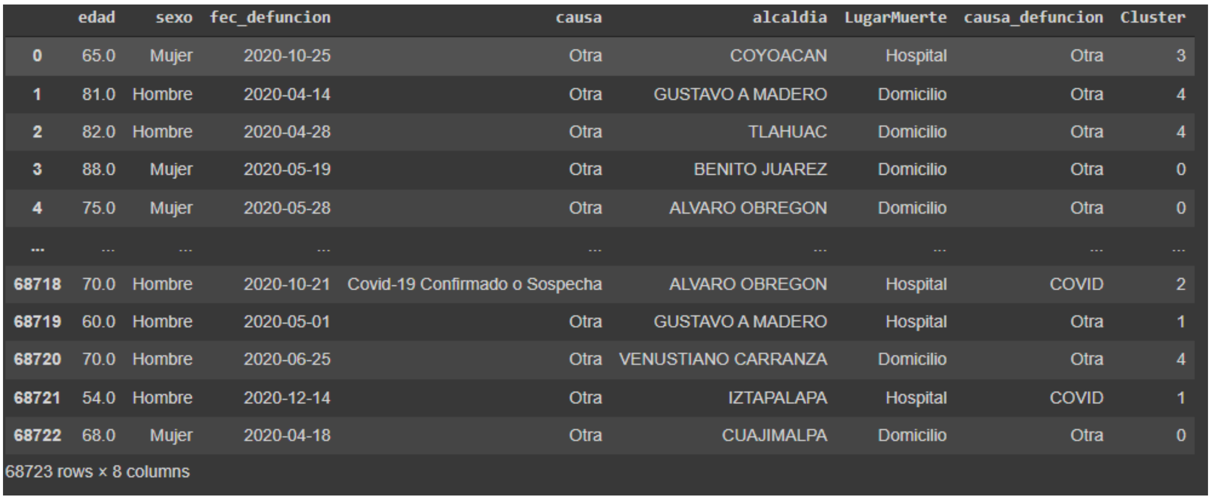


Once we have the PCA, we can move on to clustering. To do this, we are going to start with a method called the *elbow method*, where we have to plot the sum of squared errors within the group. At the point of the graph in which an "elbow" is formed we can see the supposed ideal number of clusters. For this simple plot we use MatPlotLib.



There you can see that the "elbow" is formed in the K of value 6, so we will use 6 clusters.

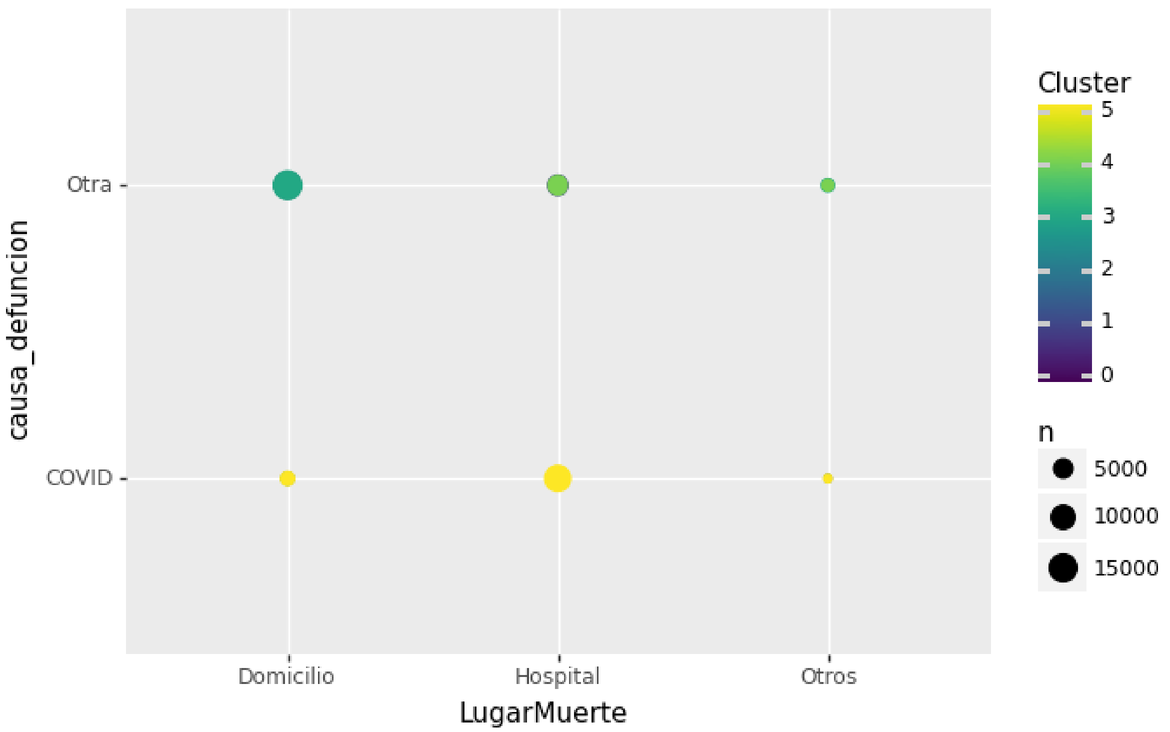
Once we have the clusters, we are going to merge the result of our One Hot Encoder with the clusters, associating the ID to the cluster.

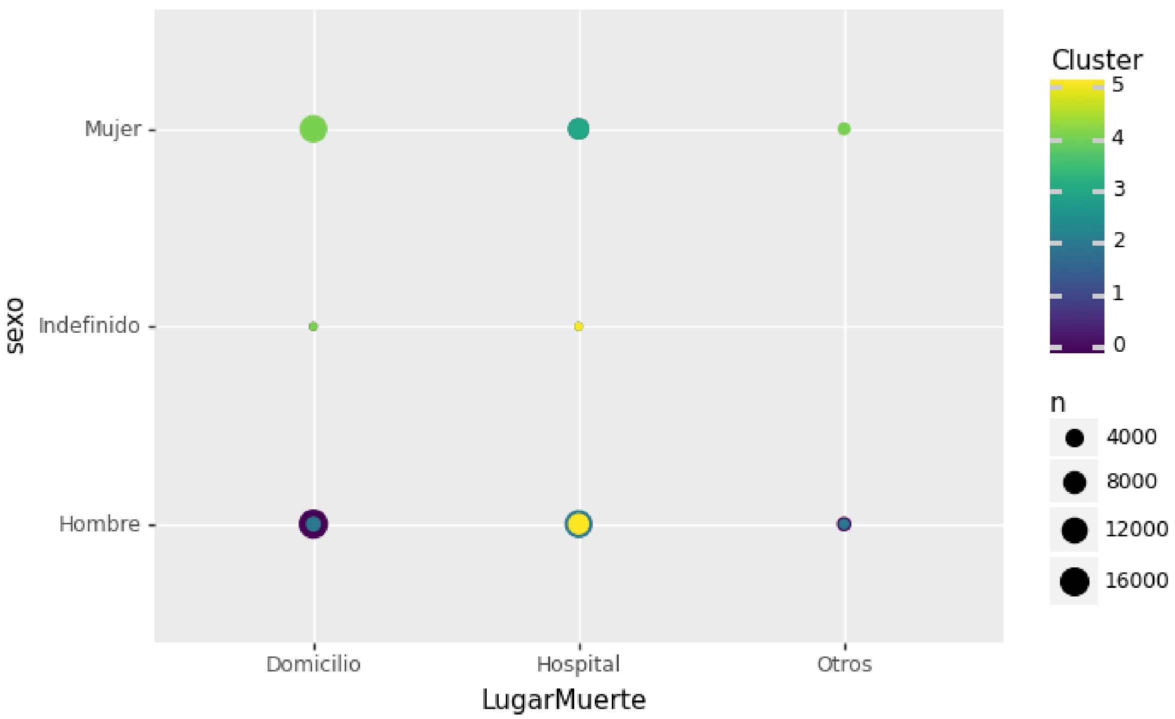


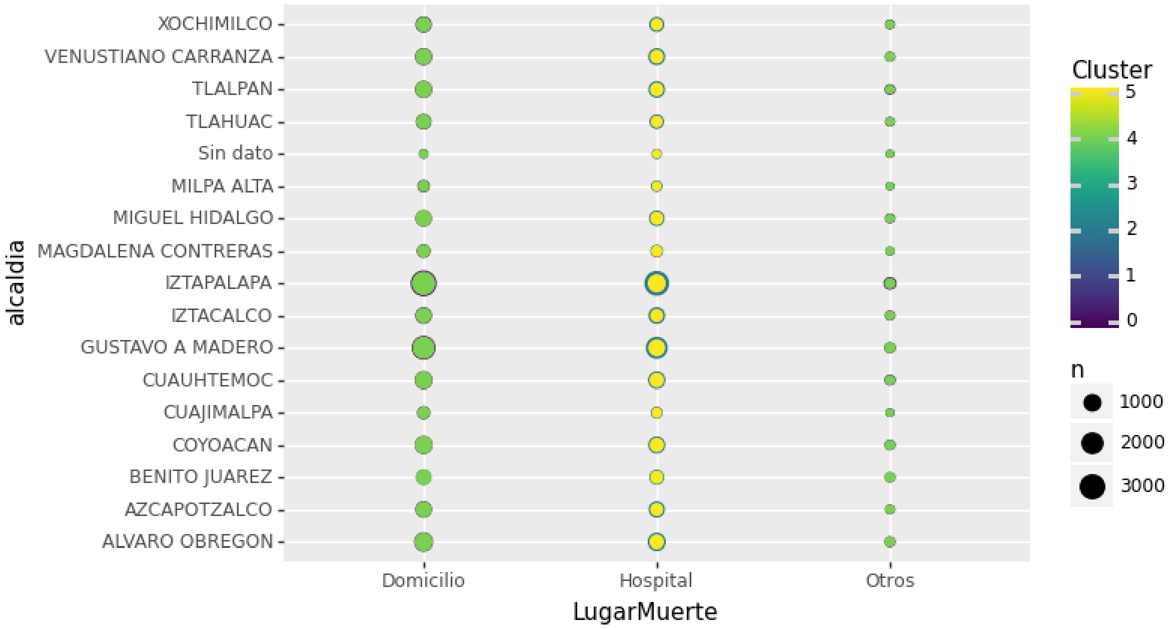
	edad	sexo	fec_defuncion	causa	alcaldia	LugarMuerte	causa_defuncion	Cluster
0	65.0	Mujer	2020-10-25	Otra	COYOACAN	Hospital	Otra	3
1	81.0	Hombre	2020-04-14	Otra	GUSTAVO A MADERO	Domicilio	Otra	4
2	82.0	Hombre	2020-04-28	Otra	TLAHUAC	Domicilio	Otra	4
3	88.0	Mujer	2020-05-19	Otra	BENITO JUAREZ	Domicilio	Otra	0
4	75.0	Mujer	2020-05-28	Otra	ALVARO OBREGON	Domicilio	Otra	0
...	...	...	...	...	...	...	...	...
68718	70.0	Hombre	2020-10-21	Covid-19 Confirmado o Sospecha	ALVARO OBREGON	Hospital	COVID	2
68719	60.0	Hombre	2020-05-01	Otra	GUSTAVO A MADERO	Hospital	Otra	1
68720	70.0	Hombre	2020-06-25	Otra	VENUSTIANO CARRANZA	Domicilio	Otra	4
68721	54.0	Hombre	2020-12-14	Otra	IZTAPALAPA	Hospital	COVID	1
68722	68.0	Mujer	2020-04-18	Otra	CUAJIMALPA	Domicilio	Otra	0

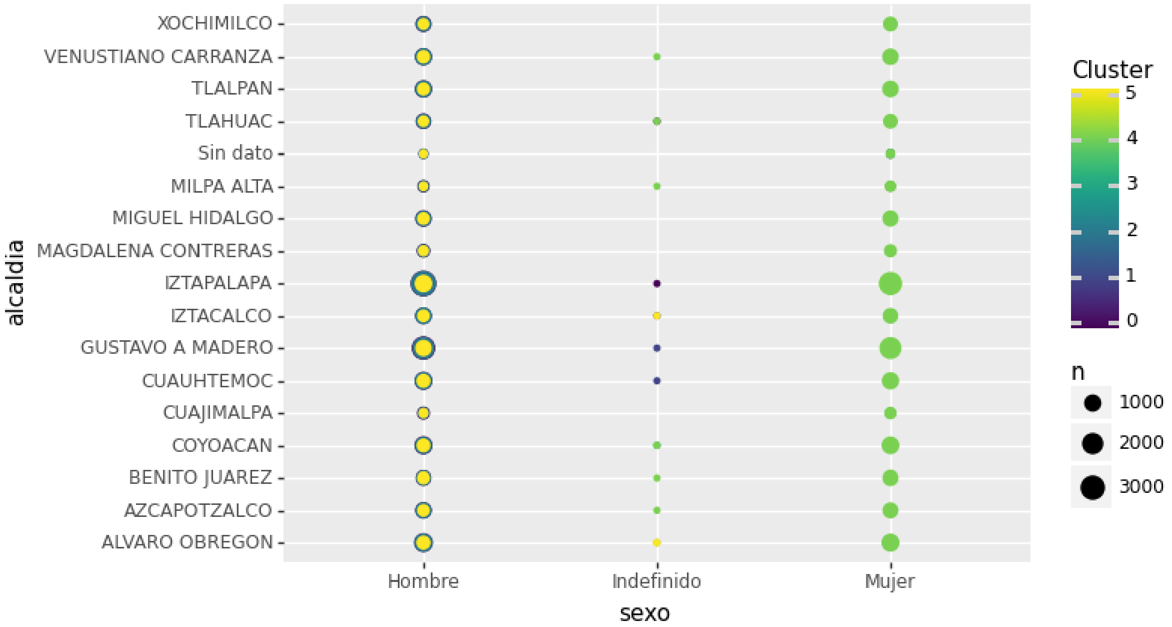
68723 rows × 8 columns

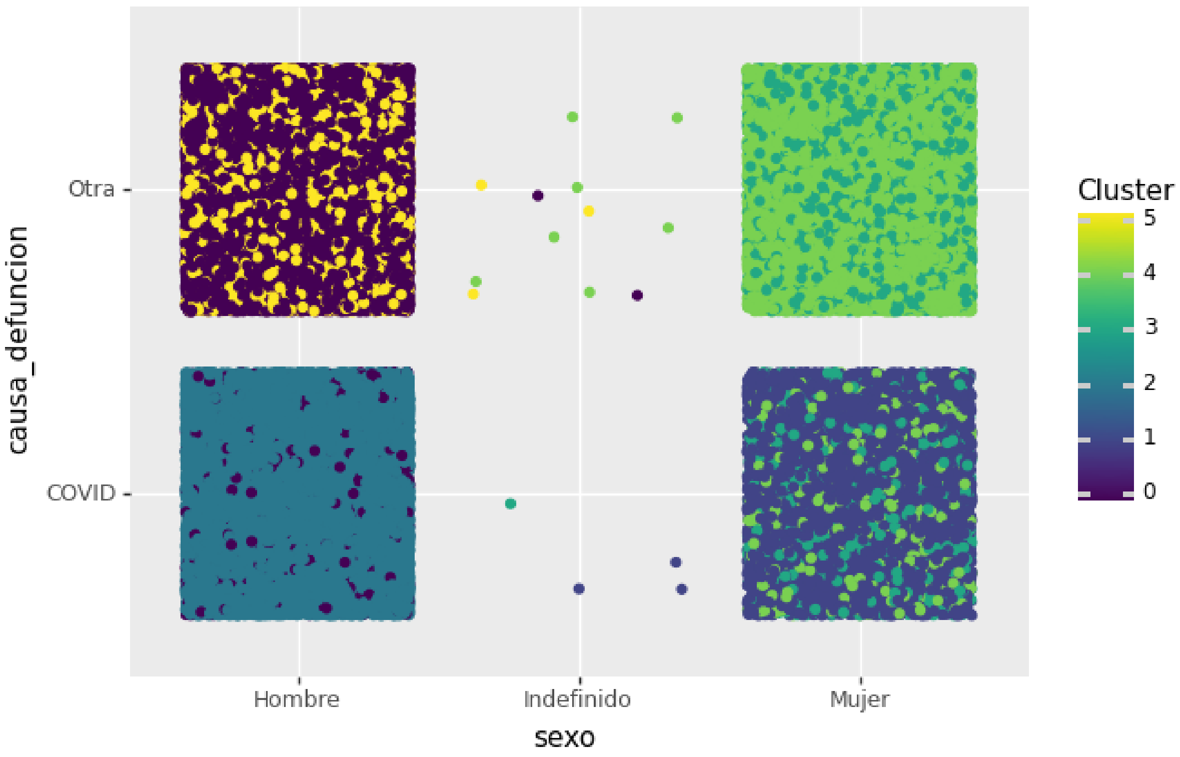
With this result, we can start plotting based on the clusters.


causa_defuncion
Otra
COVID
Domicilio
Hospital
Otros
LugarMuerte
Cluster
5
4
3
2
1
0
n
5000
10000
15000

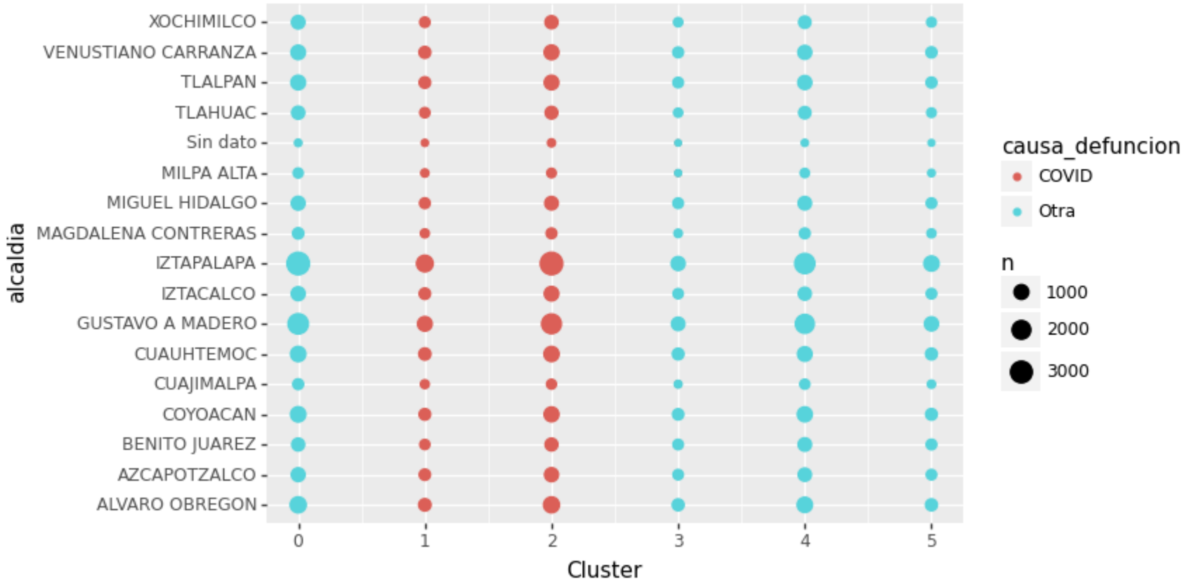

sexo
Mujer
Indefinido
Hombre
Domicilio
Hospital
Otros
LugarMuerte
Cluster
5
4
3
2
1
0
n
4000
8000
12000
16000


XOCHIMILCO
VENUSTIANO CARRANZA
TLALPAN
TLAHUAC
Sin dato
MILPA ALTA
MIGUEL HIDALGO
MAGDALENA CONTRERAS
IZTAPALAPA
IZTACALCO
GUSTAVO A MADERO
CUAUHTEMOC
CUAJIMALPA
COYOACAN
BENITO JUAREZ
AZCAPOTZALCO
ALVARO OBREGON
alcaldia
Domicilio
Hospital
Otros
LugarMuerte
Cluster
5
4
3
2
1
0
n
1000
2000
3000

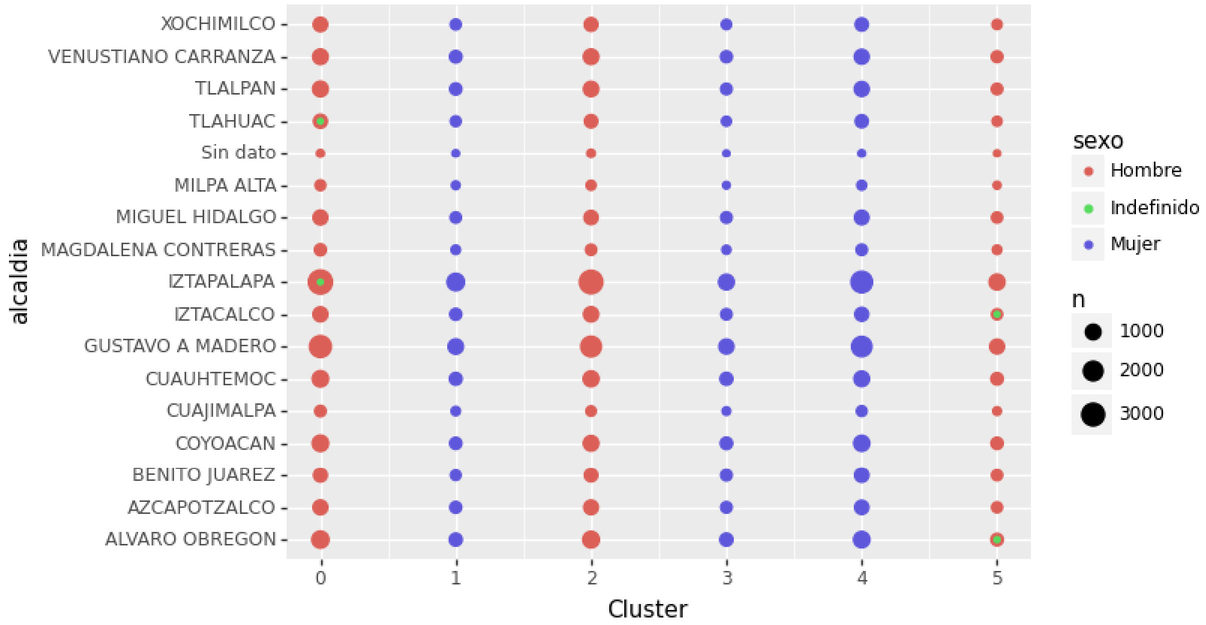

XOCHIMILCO
VENUSTIANO CARRANZA
TLALPAN
TLAHUAC
Sin dato
MILPA ALTA
MIGUEL HIDALGO
MAGDALENA CONTRERAS
IZTAPALAPA
IZTACALCO
GUSTAVO A MADERO
CUAUHTEMOC
CUAJIMALPA
COYOACAN
BENITO JUAREZ
AZCAPOTZALCO
ALVARO OBREGON
alcaldia
Hombre
Indefinido
Mujer
sexo
Cluster
5
4
3
2
1
0
n
1000
2000
3000

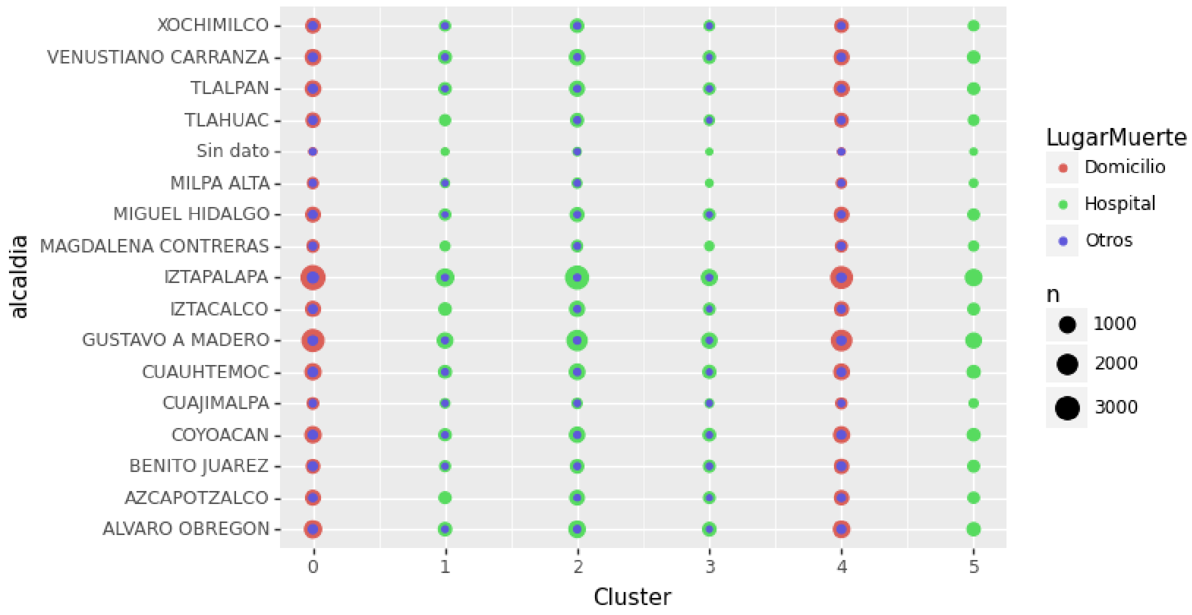


causa_defuncion
Otra
COVID
Hombre
Indefinido
Mujer
sexo
Cluster
5
4
3
2
1
0



alcaldia
XOCHIMILCO
VENUSTIANO CARRANZA
TLALPAN
TLAHUAC
Sin dato
MILPA ALTA
MIGUEL HIDALGO
MAGDALENA CONTRERAS
IZTAPALAPA
IZTACALCO
GUSTAVO A MADERO
CUAUHTEMOC
CUAJIMALPA
COYOACAN
BENITO JUAREZ
AZCAPOTZALCO
ALVARO OBREGON
0
1
2
3
4
5
Cluster
causa_defuncion
COVID
Otra
n
1000
2000
3000





Social implications and consequences

Today, around the world, we continue to face a crisis as a result of COVID-19. During the confinement stage, there was no vaccine to deal with this pandemic, which has been characterized by the WHO as being highly contagious and leading millions of people to death, so the only ways in which it has been able to contain the spread of the virus have been with sanitization, hygiene and social distancing measures.

This health crisis, caused by COVID-19, has revealed the inequality and fragility of economic, health and educational systems throughout the world, even in those nations with high levels of development.

A serious consequence was that several students had to drop out of school due to lack of financial resources to obtain a computer and pay for the internet on a monthly basis in order to be able to take their classes online.

In turn, the cost to treat this disease is extremely expensive, so not many had the financial capacity to go to the hospital and treat it. Also, there was a lot of hospital saturation since there were no beds for the sick, so many could not be treated in time.

With this project, it was possible to somewhat determine groups of people vulnerable to this disease, which could help us mitigate the impact on them through stricter precautionary measures near them.

Conclusions

With this model, we were unable to reach a precise conclusion about the vulnerable groups affected by COVID. The analysis by PCA failed to give significant insights, while the clustering gave us a small idea of the groups within our dataset, however, it was not definitive by any means.

The exploratory analysis was the one that gave us the most insights regarding vulnerable groups, since it indicated that older people are the most susceptible to this virus. Another couple of insights we obtained is that men were more affected by this virus, and that the colonies that suffered more deaths in relation to their population were Venustiano Carranza, Iztacalco and Azcapotzalco.

Bibliografía.

[1] Saénz, C. (2021). *'Línea de tiempo COVID-19'; a un año del primer caso en México*. Capital 21. consultado de: https://www.capital21.cdmx.gob.mx/noticias/?p=12574#:~:text=El%2027%20de%20febrero%202020,suma%2058%20d%C3%ADas%20de%20pandemia.

[2] JHU CSSE. (2022). *COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University*. GitHub. consultado de: https://github.com/CSSEGISandData/COVID-19

[3] Gob. MX (2022). *Calendario de Vacunación*. GobMX. consultado de: http://vacunacovid.gob.mx/wordpress/calendario-vacunacion/#:~:text=El%2024%20de%20diciembre%20inici%C3%B3,de%20la%20Ciudad%20de%20M%C3%A9xico.