UNIVERSIDAD PANAMERICANA
Campus México

UNIVERSIDAD PANAMERICANA

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

Profesor: Sarahí Aguilar González

Fecha de entrega: 23/05/2022

Ciclo: 1222

Nombre del proyecto: 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

Se realizó un análisis del set de datos de actas de defunciones de México, específicamente de la Ciudad de México entre el inicio de la pandemia por COVID-19 en territorio nacional y el inicio del esquema de vacunación en la ciudad. Se usaron herramientas de ciencia de datos en Python, tales como pandas y scikit learn para ello. Este análisis conllevó una manipulación y visualización de datos así como el diseño de un modelo de Machine Learning no supervisado, con el cual no logramos llegar a una conclusión precisa sobre nuestra pregunta de investigación. Gracias a la visualización de datos logramos darnos cuenta que los grupos más vulnerables son los hombres con una edad avanzada, más precisamente aquellos que tienen una edad entre los 60 y 80 años de las alcaldías de Venustiano Carranza, Iztacalco y Azcapotzalco.

Introducción

El 27 de febrero de 2020 se registró el primer caso de COVID-19 en México, dando inicio a la pandemia en territorio nacional. Desde entonces, el país ha sufrido mucho económicamente y socialmente. [1]

En total, se han tenido cerca de 6 millones de contagios y más de 300 mil muertes. Por ello, es importante saber cuáles han sido los grupos más vulnerables a este virus. [2]

El 24 de diciembre de 2020 se aplicó la primera dosis de la vacuna contra el SARS-COV-2, lo cual marcó el inicio del esquema de vacunación en nuestro país. [3]

Las vacunas han ayudado a reducir enormemente la mortalidad del virus. Por esta razón, nuestra investigación se centrará en el periodo antes del inicio del esquema de vacunación.

La gran idea

Entre los fallecidos por COVID-19 existen una gran variedad de factores que influyen a que una persona fallezca, por lo que buscamos conocer los patrones más comunes. El periodo de análisis será entre el inicio de la pandemia en México y el inicio del esquema de vacunación en la Ciudad de México, para evitar sesgar los datos con la protección que proveen las

vacunas.

Pregunta de Investigación

¿Cuáles fueron los grupos más vulnerables frente al COVID-19 en la CDMX antes del inicio del esquema de vacunación, tomando en cuenta factores como su sexo, edad, alcaldía y comorbilidades?

Fuentes de datos relevantes:

- Actas de defunción CDMX (Defunciones desde 2017)

Inconvenientes

- Actas de defunción CDMX
    - No tiene comorbilidades
    - Son muertes desde antes del COVID
    - Vienen varios estados de la república

Variables dependientes e independientes

- **Dependientes**
    - N/A
- **Independientes**
    - Alcaldía
    - Edad
    - Comorbilidades
    - Sexo
    - Lugar de muerte
    - Fecha de defunción

Modelo a implementar

Al realizar esta investigación no vamos a predecir una variable, sino que vamos a encontrar patrones que puedan definir distintos grupos vulnerables en la CDMX. En otras palabras, nuestro modelo va a clasificar dichos grupos vulnerables, por lo que es un modelo no

supervisado.

En este sentido, la prioridad será la inferencia, ya que queremos deducir dichas características para clasificar los grupos y poder inferir los grupos vulnerables.

Desarrollo

Proyectos similares al nuestro:

- 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

Para responder a la pregunta de investigación, nos vamos a apoyar de la siguientes herramientas:

- 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

El primer paso es obtener los datos. Para ello, vamos a utilizar las actas de defunción del gobierno de México. La URL del archivo CSV es el siguiente:

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

La manera en que leemos y almacenamos los datos es mediante la librería de Pandas. Los datos vienen con la siguiente estructura:

1. **Edad - float64**
   Edad de la persona fallecida.
2. **Sexo - String**
   Sexo de la persona fallecida.

3. **Fecha Defunción - String**

   Fecha en la que falleció en formato YYYY-MM-DD
4. **Estado - String**

   Estado en el cual falleció la persona.
5. **Causa - String**

   Determinación de si falleció por COVID-19 o por otras causas.
6. **Causa Registro - String**

   Causas de muerte en formato separado por comas.
7. **Alcaldía - String**

   Alcaldía en la cual falleció. En caso de estar fuera de la CDMX está en NaN.
8. **Lugar Muerte - String**

   Determinación de si murió en casa o en el hospital.
9. **Número Consecutivo - int64**

   Número consecutivo de fallecimiento.

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

En total tenemos 85,904 entradas en nuestro dataset. Entre las columnas, encontramos varias que tienen valores nulos: *Edad*, *Causa Registro* y *Alcaldía*. En esos casos, llenamos esos espacios nulos con el string “Sin dato” ya que consideramos importantes todos los registros.

Una vez obtenidos los datos, podemos comenzar con su manipulación para limpiar los mismos o sacar nuevos DataFrames que faciliten la visualización de datos y la relación de los mismos. En este paso realizamos la limpieza y filtración de datos para poder analizarlos posteriormente.

Análisis exploratorio

Lo primero que queremos hacer es contabilizar las causas de muerte que existen en nuestro dataset. Para ello, iteramos sobre nuestro dataset de defunciones pre vacunación en la Ciudad de México y analizamos las causas de muerte que se asociaron al paciente. Almacenamos todas estas causas de muerte dentro de un diccionario para su posterior análisis.

Si detectamos que alguna de las causas de muerte del paciente tienen alguna palabra clave relacionada con el COVID-19 (COVID, SARS, Neumonía Atípica, Insuficiencia respiratoria, etc.) entonces la agregamos al diccionario bajo una sola llave que denominamos *COVID*, en caso contrario, agregamos la causa textual al diccionario como una llave. Los valores de las llaves son las defunciones asociadas a esa causa.

Un problema con el que nos encontramos al realizar esto, es que los registros varían mucho ya que las mismas causas de muerte se podían escribir de una manera diferente, lo cual dificultó la identificación de defunciones asociadas al COVID-19.

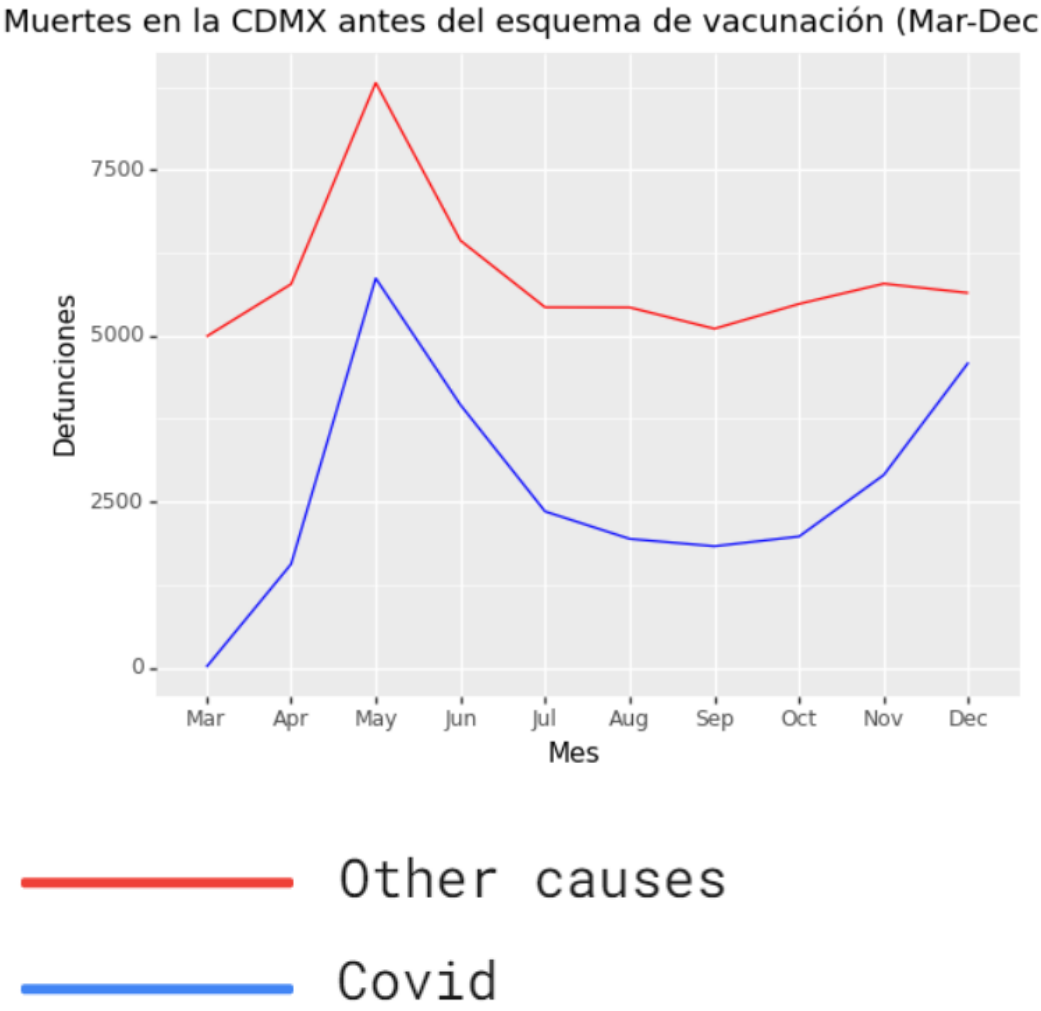
Posterior a este procedimiento, proseguimos a realizar un feature engineering, en donde agregamos un campo al Data Frame que denominamos *Causa Defunción*. Esta causa se determina tomando en cuenta el análisis de palabras clave que determinamos anteriormente, ya que en varios casos con neumonías atípicas no se contabilizaba como defunción por COVID. Al comparar ambos números nos damos cuenta que en efecto hay más muertes posiblemente relacionadas al COVID, en específico, vemos que hay una diferencia de 6,358 defunciones en nuestro 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
```

El siguiente análisis que pasamos a realizar es el de las fechas. Definimos dos diccionarios con todas las defunciones por cada mes. En uno, son defunciones por COVID, en el otro son por las otras causas. Con esta información, realizamos feature engineering para definir otra columna que determina el mes de defunción, para poder visualizarlo con mayor facilidad.

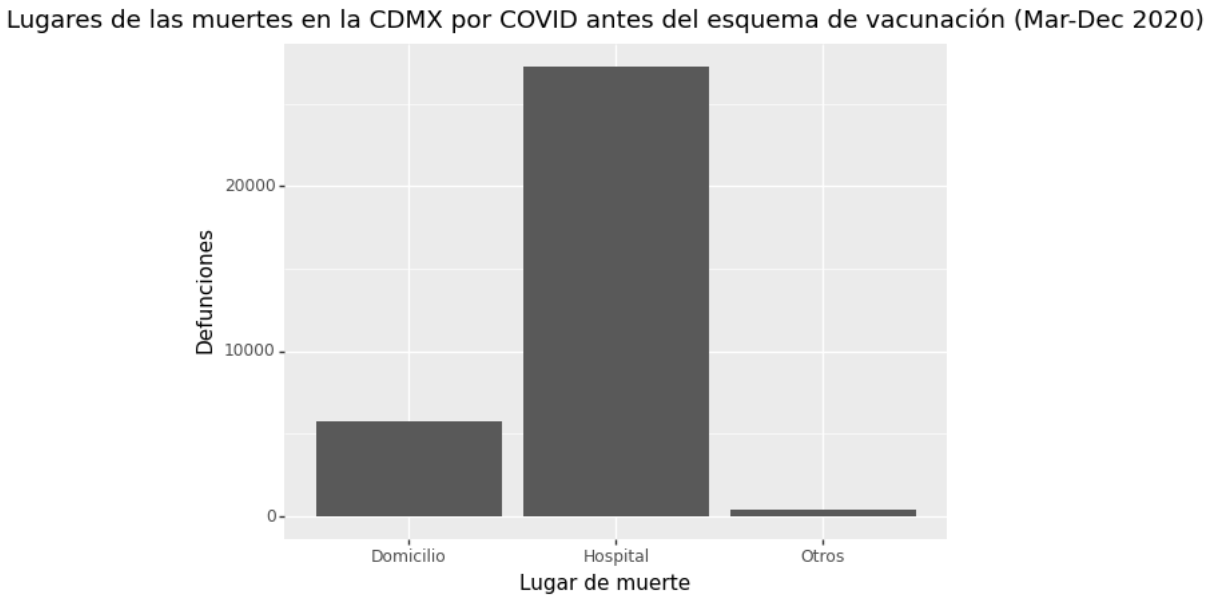
Visualización de datos

Para la visualización de datos utilizamos exclusivamente ggplot. La primera gráfica que desplegamos es la de muertes por mes en la CDMX antes del esquema de vacunación.

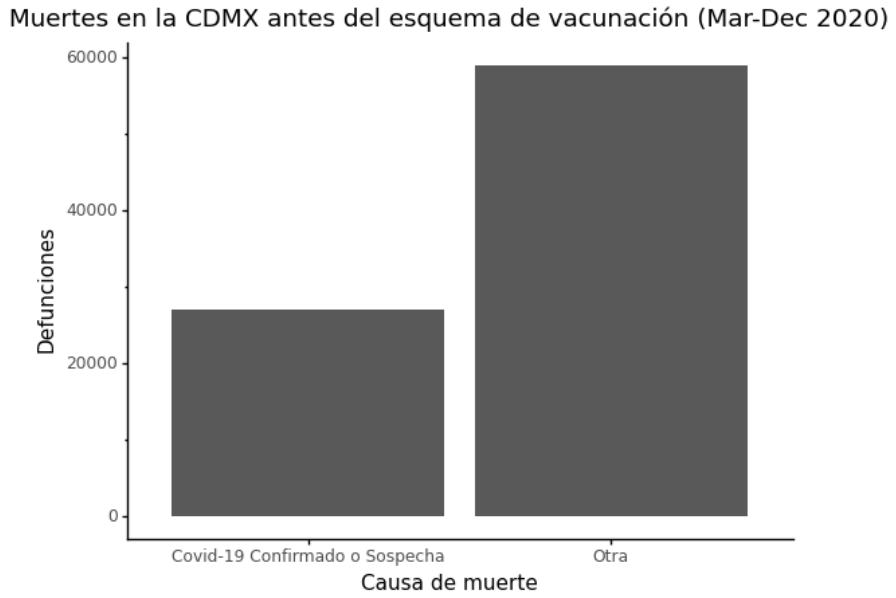


Con esta visualización nos damos cuenta de algo curioso, que es que las muertes por COVID parecen estar relacionadas con las muertes por otras causas, ya que ambas tienen la misma forma de curva aproximadamente, pero con diferentes cantidades.

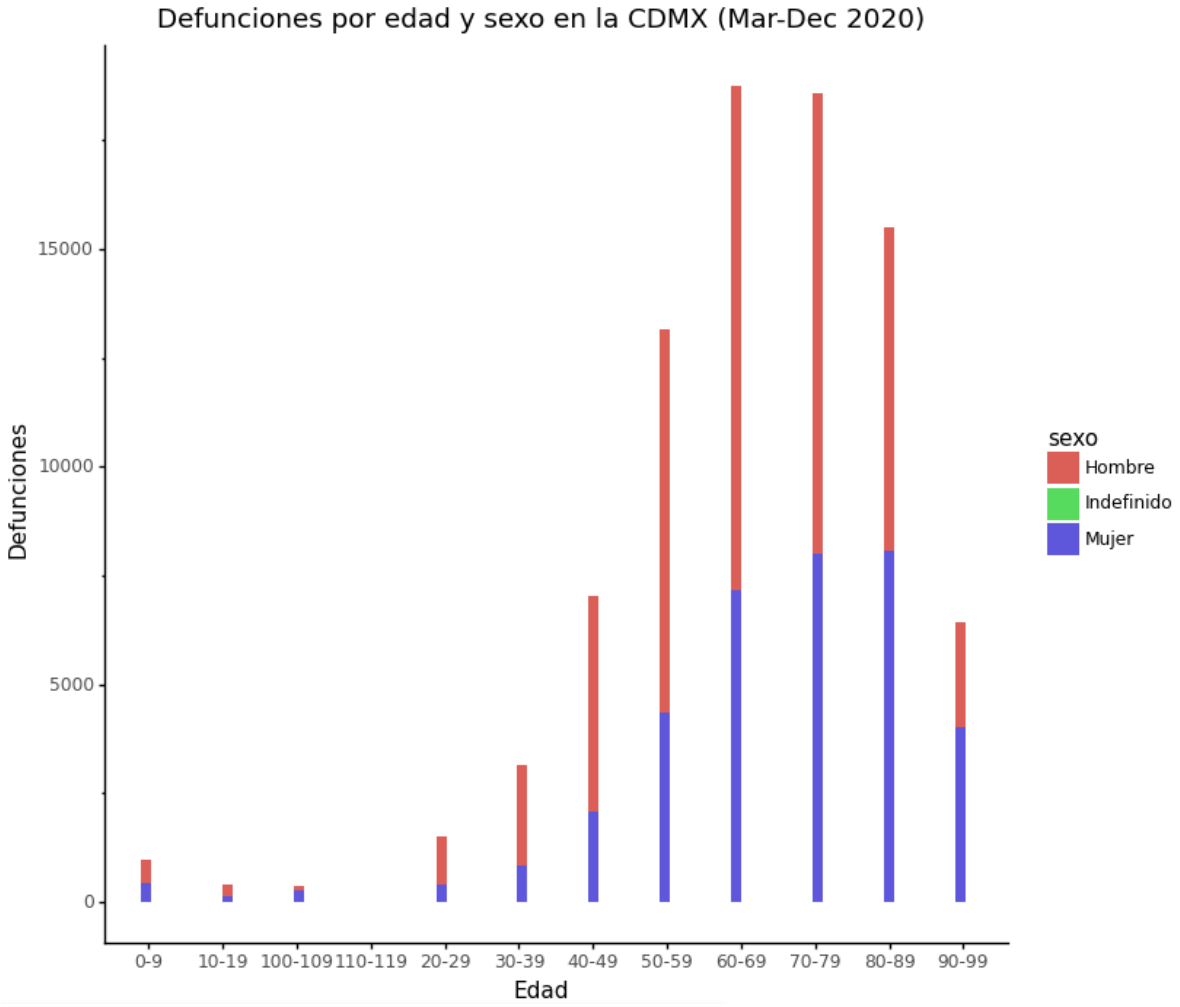
Posteriormente, graficamos las defunciones que ocurrieron en las tres opciones de lugares: Domicilio, Hospital u Otras.



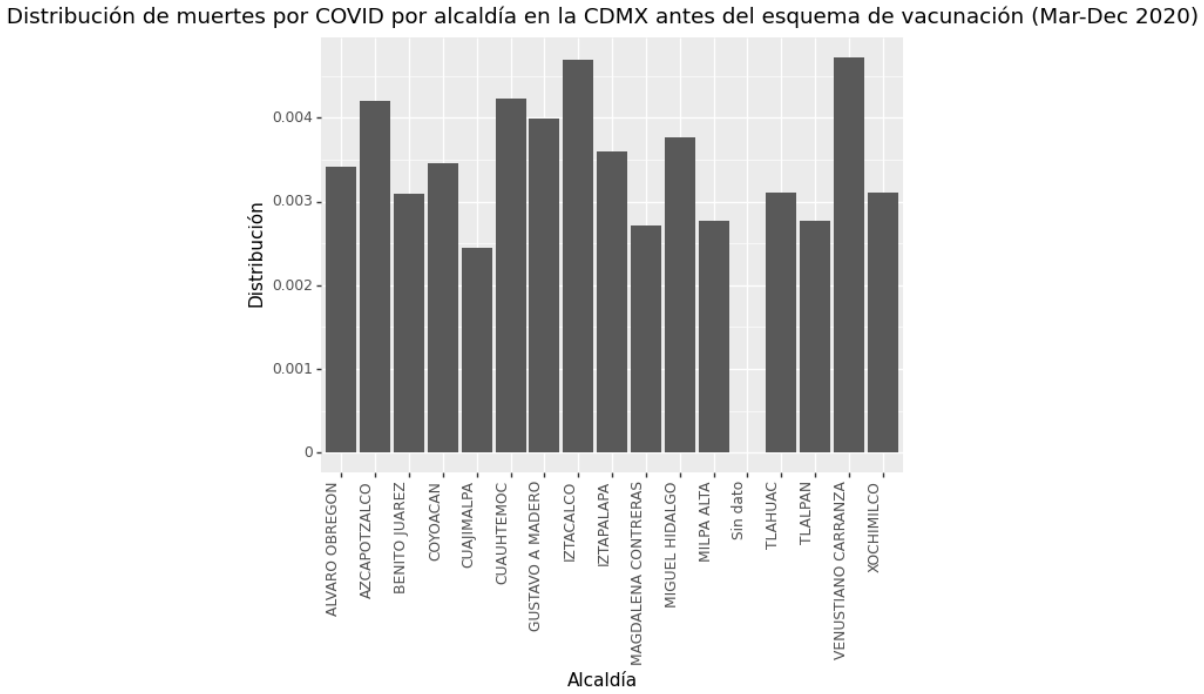
Otra visualización que consideramos relevante es la comparación entre la cantidad de muertes por COVID y por otras causas. Con ello, nos dimos cuenta que la cantidad de muertes por COVID llega casi a la mitad de las muertes por todas las demás causas.



La siguiente visualización fue la relación entre la edad y el género en cuanto al número de defunciones. Evidentemente, encontramos que la mayor concentración de defunciones está entre los 60 y 80 años. En cuanto al género, parece ser que los hombres fueron más vulnerables.



Otra gráfica importante fue la de la relación de muertes por COVID por alcaldía. Aquí dividimos las defunciones por alcaldía entre la población de esa alcaldía para sacar las alcaldías con más porcentaje de muertes en relación a su población.



Implementación de modelo de Machine Learning

Como mencionamos al principio de este documento, nuestro modelo de Machine Learning va a ser uno no supervisado, ya que no buscamos predecir una variable, sino encontrar características comunes entre grupos. Es por ello que vamos a realizar dos procedimientos: Hacer un Análisis de los Componentes Principales (PCA) y hacer Clustering con los k means.

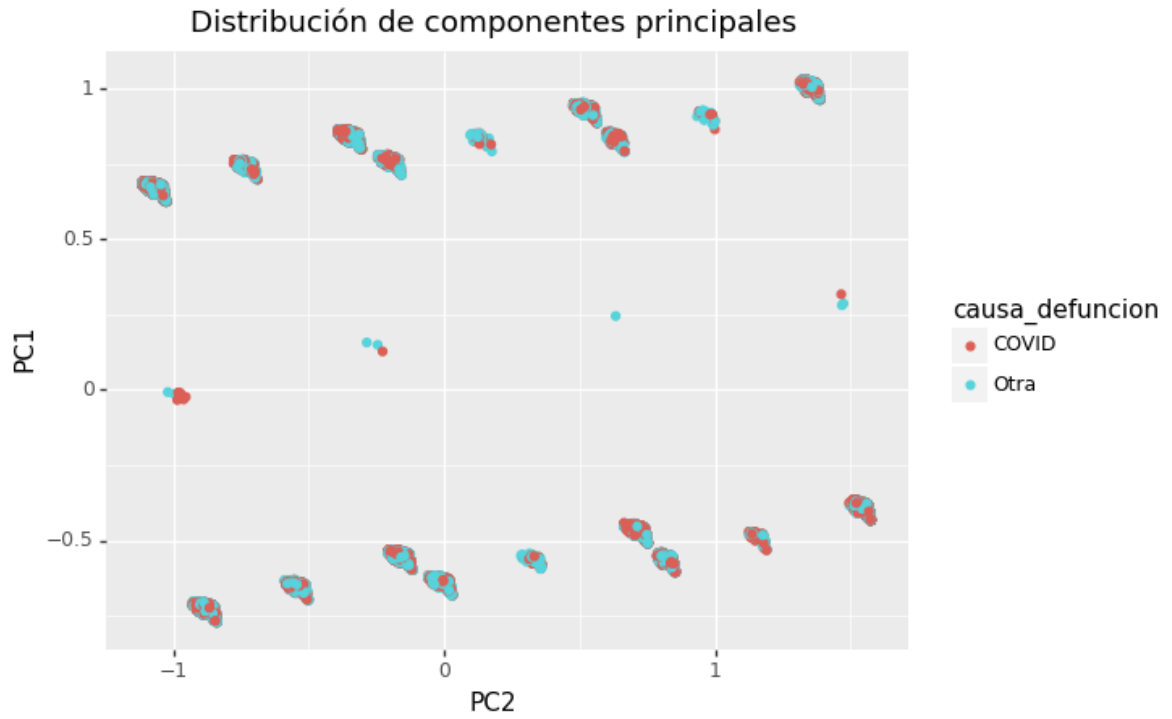
El primer paso para ello es separar nuestros datos en un conjunto de entrenamiento y otro de pruebas. La relación que usaremos será 80% y 20% respectivamente, con un random state de 1 para poder replicar los resultados. Para realizar estas operaciones vamos a necesitar la librería de Scikit learn.

Una vez que se dividieron los datos, tenemos que utilizar el One Hot Encoder para codificar nuestras variables categóricas (Que básicamente son todas las de nuestro set de datos) en una matriz de puros 0 y un 1 para que el modelo pueda identificar bien las variables y sus valores en cada fila.

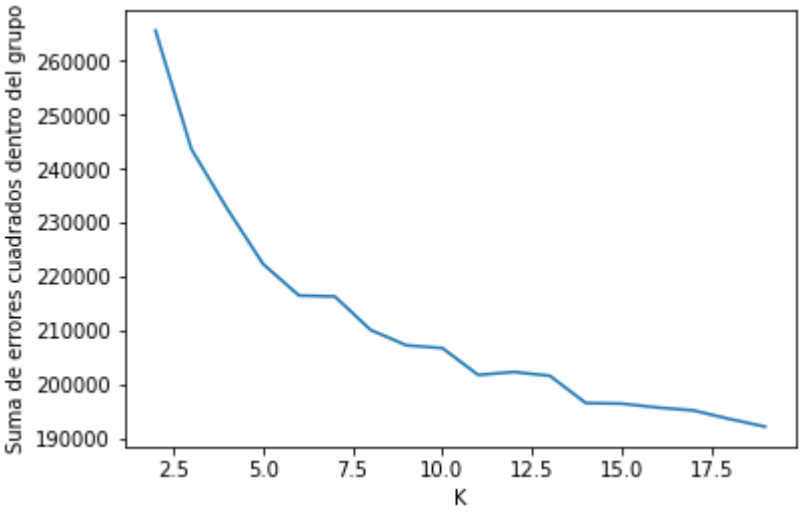
El resultado de esta operación de codificación nos da como resultado un *Sparse Matrix*. Sin embargo, este tipo de matriz no se puede graficar o usar fácilmente para el PCA, por lo que la pasamos a un arreglo de numpy antes que nada.

Seguido de esto, pasamos a realizar el PCA con dos componentes. Este PCA es el que vamos a graficar.

El resultado de dicha visualización es la siguiente:



Una vez que tenemos el PCA, podemos pasar a el clustering. Para ello, vamos a comenzar con un método llamado *método del codo*, en donde tenemos que graficar la suma de los errores cuadrados dentro del grupo. En el punto de la gráfica que se forme un "codo" podemos ver el supuesto número ideal de clusters. Para esta simple gráfica usamos MatPlotLib.



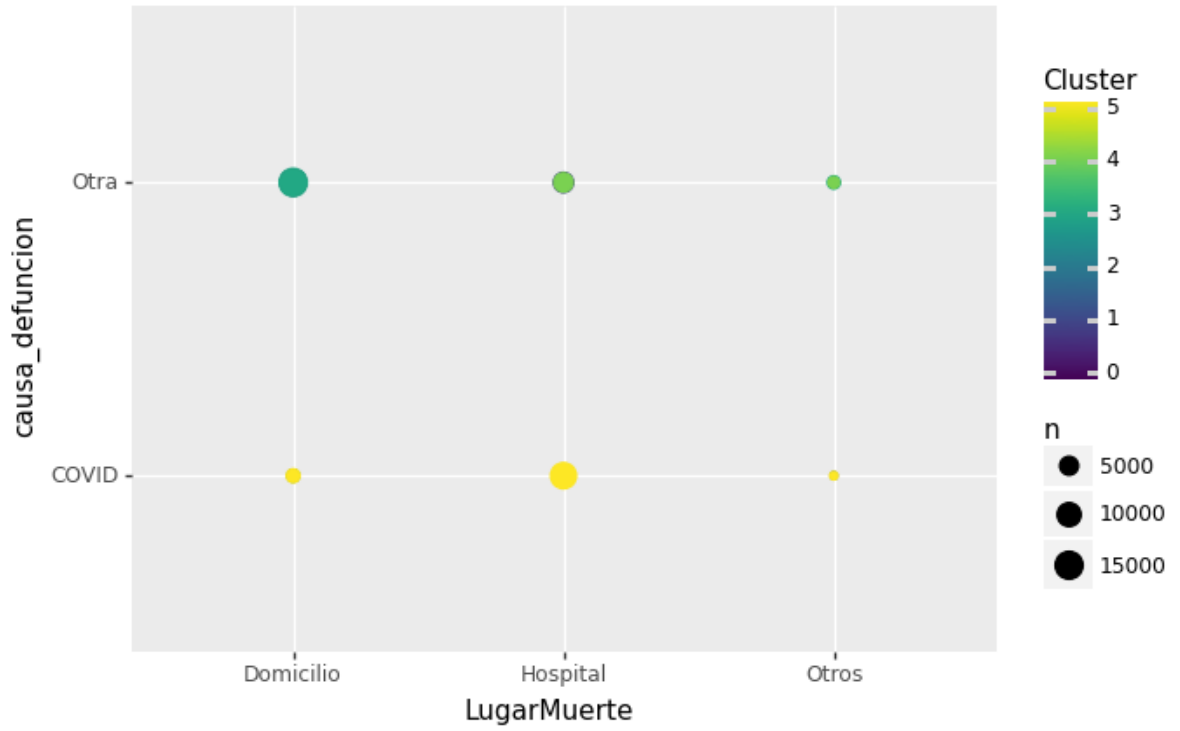
Ahí se puede apreciar que el "codo" se forma en la *K* de valor 6, por lo que vamos a usar 6 clusters.

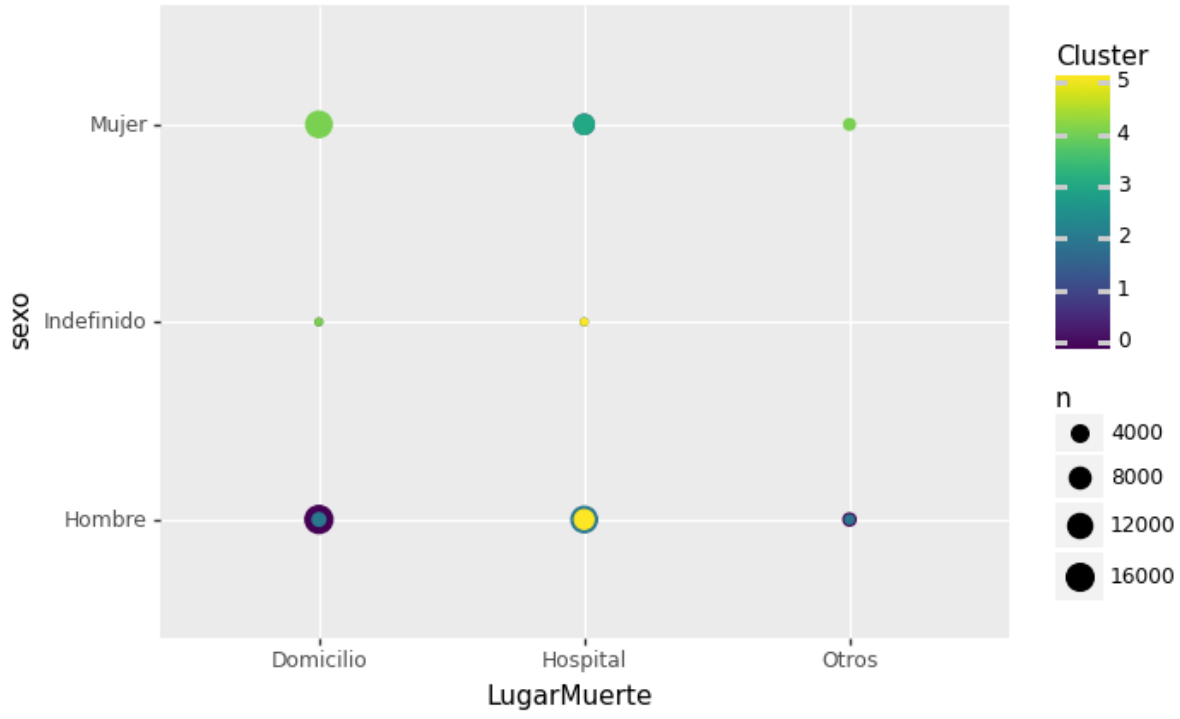
Al ya tener los clusters, vamos a relacionar el resultado de nuestro One Hot Encoder con los clusters, asociando el ID al cluster.

	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

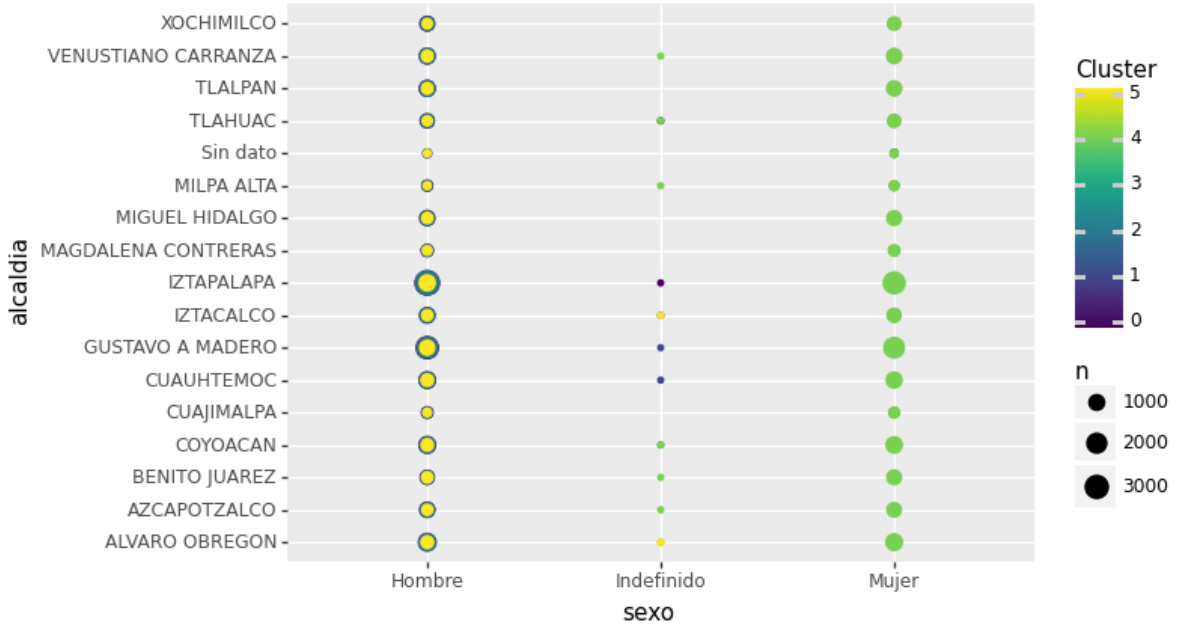
68723 rows × 8 columns

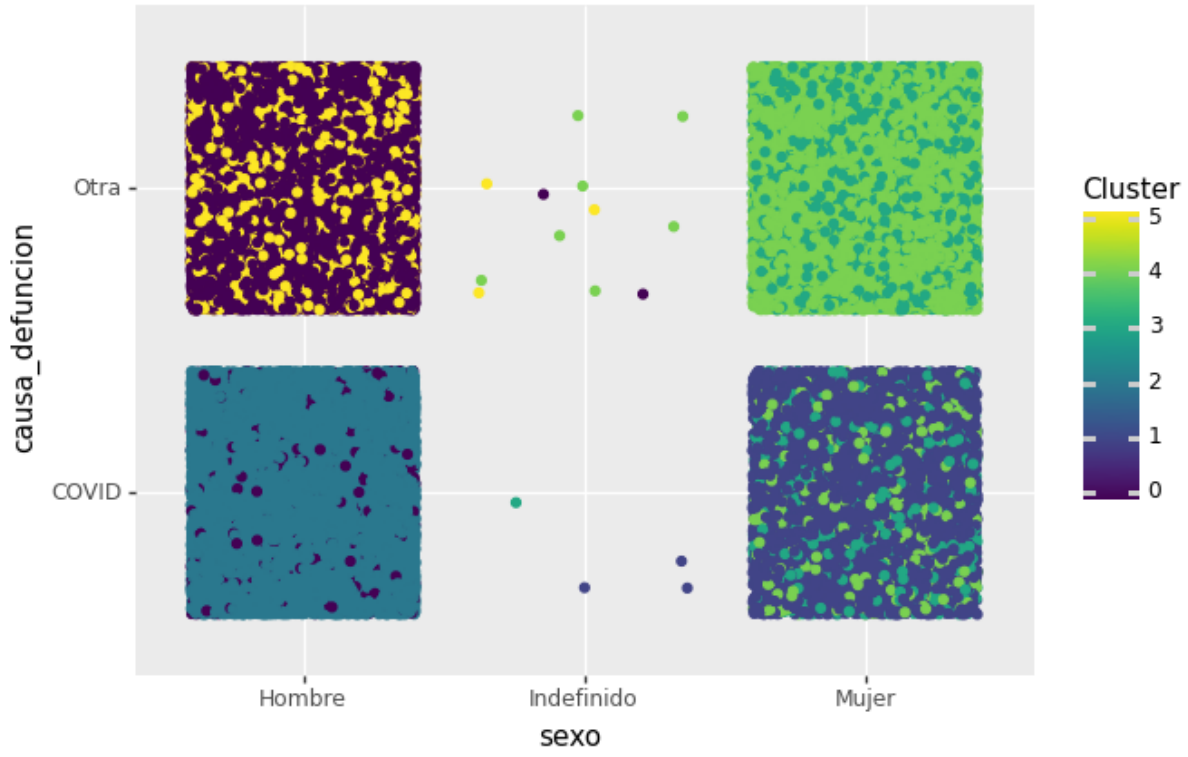
Con este resultado, podemos empezar a graficar con base en los clusters.

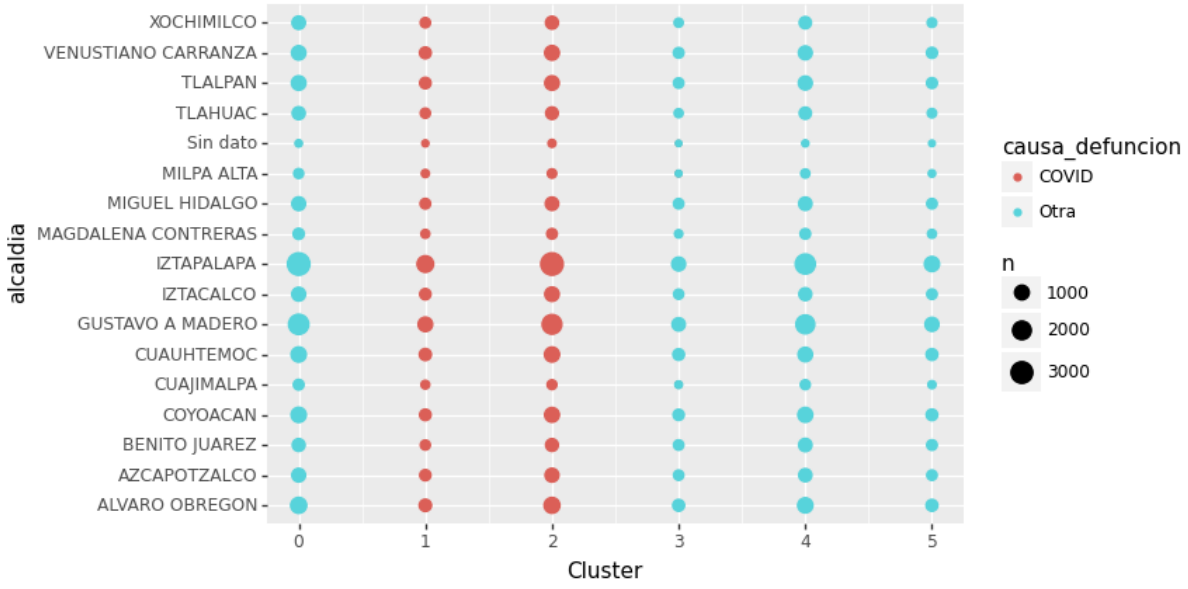

causa_defuncion
Otra
COVID
Domicilio
Hospital
Otros
LugarMuerte
Cluster
5
4
3
2
1
0
n
5000
10000
15000

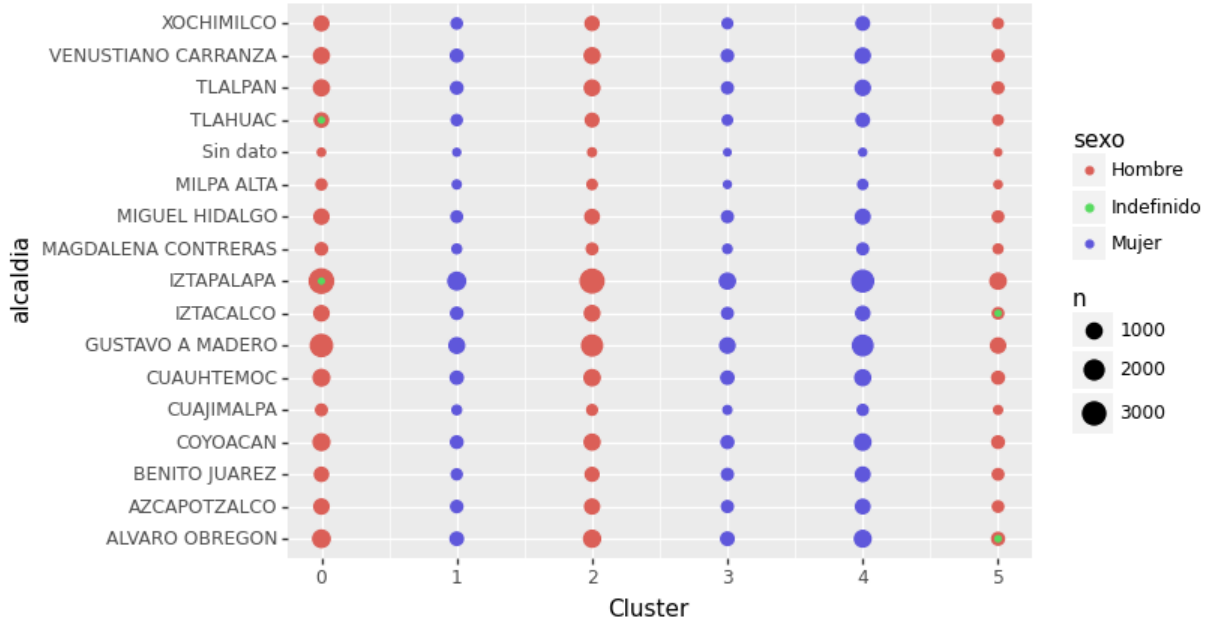

sexo
Mujer
Indefinido
Hombre
Domicilio
Hospital
Otros
LugarMuerte
Cluster
5
4
3
2
1
0
n
4000
8000
12000
16000

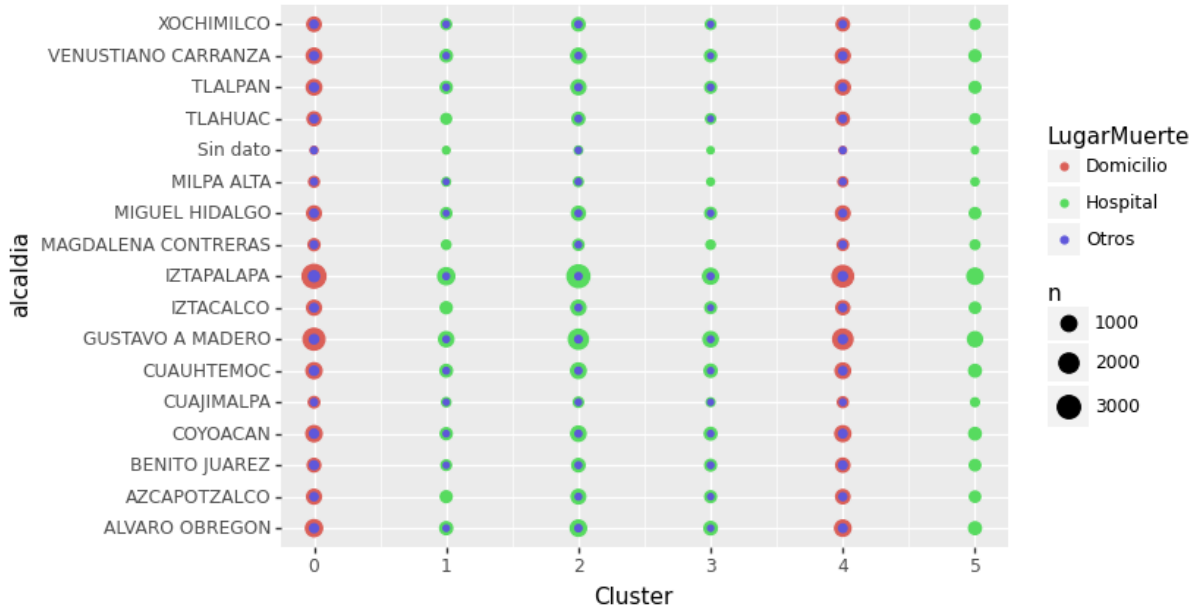

XOCHIMILCO
VENUSTIANO CARRANZA
TLALPAN
TLAHUAC
Sin dato
MILPA ALTA
MIGUEL HIDALGO
MAGDALENA CONTRERAS
IZTAPALAPA
IZTACALCO
GUSTAVO A MADERO
CUAUHTEMOC
CUAJIMALPA
COYOACAN
BENITO JUAREZ
AZCAPOTZALCO
ALVARO OBREGON
alcaldia
Domicilio
Hospital
Otros
LugarMuerte
Cluster
5
4
3
2
1
0
n
1000
2000
3000


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





Implicaciones sociales y sus consecuencias

Hoy en día, alrededor del mundo, nos seguimos enfrentando a una crisis a consecuencia del COVID-19. Durante la etapa del confinamiento no se contó con alguna vacuna para poder hacerle frente a esta pandemia el cual ha sido caracterizada por la OMS por ser altamente contagiosa y llevar a millones de personas a la muerte por lo que, las únicas formas en las que

ha podido contener la propagación del virus han sido con medidas de sanitización, higiene y el distanciamiento social.

Esta crisis sanitaria, causada por el Covid-19, ha develado la desigualdad y fragilidad de los sistemas económicos, sanitarios y educativos en todo el mundo, incluso en aquellas naciones con altos niveles de desarrollo.

Una grave consecuencia fue que varios estudiantes tuvieron que dejar los estudios por falta de recursos económicos para poder obtener una computadora y pagar internet de manera mensual con el fin de poder tomar sus clases en línea.

A su vez el costo para tratar dicha enfermedad es extremadamente caro por lo que no muchos tenían la capacidad económica para ir al hospital y tratarlo. También, había mucha saturación hospitalaria ya que no había camas para los enfermos por lo que muchos no pudieron ser atendidos a tiempo.

Con este proyecto, se pudieron determinar grupos de personas vulnerables ante esta enfermedad, lo cual podría ayudarnos a mitigar el impacto en ellos mediante medidas de precaución más estrictas cerca de los mismos.

Conclusiones

Con este modelo, no pudimos llegar a una conclusión precisa sobre los grupos vulnerables afectados por el COVID. El análisis por PCA no logró dar insights significativos, mientras que el clustering nos dió una pequeña idea de los grupos dentro de nuestro dataset, sin embargo, no fue definitivo de ninguna manera.

El análisis exploratorio fue el que más insights nos dió en cuanto a los grupos vulnerables, ya que nos indicó que las personas de mayor edad son las más susceptibles ante este virus. Otro par de insights que obtuvimos es que los hombres fueron más perjudicados por este virus, y que las colonias que sufrieron más muertes en relación a su población fueron Venustiano Carranza, Iztacalco y 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.