# UNIVERSIDAD DEL VALLE DE GUATEMALA

**Minería de Datos**


# Proyecto 3. Entrega 01 de Avances (EDA)

Karen Pineda

Michelle Mejía

Silvia Illescas

Emilio Reyes

Hansel López

Guatemala, 23 de Marzo de 2025

Titulo

**Situación**

Las defunciones en Guatemala representan un fenómeno complejo que involucra múltiples factores sociales, económicos, geográficos y de salud pública. Comprender las causas más frecuentes de muerte en la población, así como su distribución por edad, sexo, región y condición socioeconómica, es fundamental para guiar decisiones en políticas de salud, prevención de enfermedades y atención médica. Sin embargo, a pesar de contar con registros oficiales, el análisis de estos datos no siempre se realiza con profundidad, lo que limita la capacidad del Estado y de las instituciones para actuar proactivamente ante las principales causas de mortalidad.

**Problema científico:** ¿Qué patrones relevantes pueden identificarse en los registros de defunciones ocurridas en Guatemala entre 2009 y 2021, considerando variables sociodemográficas, geográficas y clínicas, que permitan entender mejor las causas predominantes de muerte en el país?

**Objetivo General**

Analizar los registros de defunciones en Guatemala entre 2009 y 2021 para identificar patrones relevantes asociados a variables sociodemográficas, de acceso a salud y localización geográfica.

**Objetivos específicos**

- Determinar las causas de muerte más frecuentes a nivel nacional y su evolución a lo largo del tiempo.
- Evaluar la distribución de las defunciones por edad, sexo, grupo étnico y nivel educativo.
- Analizar la relación entre el lugar de ocurrencia de la muerte y el acceso a asistencia médica.
- Detectar diferencias regionales y territoriales en la incidencia de las principales causas de defunción.
- Aplicar técnicas de agrupamiento o visualización para identificar perfiles poblacionales o zonas críticas con alta mortalidad.

Repositorio

**Descripción de los datos**

En un primer momento, los datos se encuentran separados en archivos .sav, los cuales contienen los registros originales de defunciones desde el año 2012 hasta el 2021. Se utiliza un script de Python llamado preprocessing.py para recorrer estos archivos, leerlos mediante la librería pyreadstat y combinarlos en un único DataFrame. Este conjunto de datos unificado se guarda en formato CSV (defunciones.csv), creando así una base homogénea para el análisis.

Posteriormente, se procede a la limpieza y transformación de defunciones.csv. En esta etapa, se cargan los datos utilizando un diccionario de tipos, convirtiendo columnas específicas como “Añoreg” y “Edadif” a enteros y el resto a cadenas. Esto permite seleccionar únicamente las columnas relevantes y renombrarlas a nombres descriptivos, como departamento, municipio, sexo, etc. El DataFrame se ordena por año de registro y se guarda en un archivo intermedio llamado defunciones_clean.csv. Además, se convierte la mayoría de las columnas a variables categóricas, lo que optimiza la memoria y facilita el análisis.

A continuación, se realiza una exploración inicial del dataset limpio. Se revisan las dimensiones, se cuenta el número de valores nulos por columna y se obtiene un resumen estadístico de las variables numéricas. Este análisis permite identificar la calidad de los datos y detectar posibles problemas, como valores faltantes o atípicos.

Finalmente, se lleva a cabo la simplificación y el enriquecimiento del dataset. Se utiliza el archivo CIE.csv para construir un diccionario que mapea los códigos ICD-10 a sus descripciones, simplificando la columna “causa” mediante coincidencias exactas o por prefijo y eliminando registros no relevantes. Además, se corrigen inconsistencias en variables como “ocupacion” y se crea la variable “age_group”. El resultado es un archivo final llamado defunciones_simplified.csv, listo para análisis exploratorios y modelado.

Cabe destacar que en la columna de “Age group” se encontraron 3,726 valores nulos, los cuales se trataron asignándoles el valor “Desconocido” para no perder información ni realizar suposiciones. Asimismo, algunas columnas, como departamento, sexo y

etnia, estaban en formato numérico y se convirtieron a variables categóricas para mejorar su manejo y análisis posterior.

**Situación problemática a partir de las visualizaciones**

Las visualizaciones generadas evidencian que las defunciones en Guatemala no están distribuidas de forma uniforme ni al azar. Observamos, por ejemplo, una fuerte concentración de muertes en grupos etarios mayores (65+), y una mayor proporción de defunciones en hombres que en mujeres. Además, se identifican desigualdades en variables como escolaridad y etnia, lo cual sugiere una posible relación entre las condiciones socioeconómicas y el riesgo de mortalidad.

Las causas más frecuentes de muerte están dominadas por enfermedades crónicas y respiratorias, así como por condiciones prevenibles con adecuada atención médica. A esto se suma la variabilidad territorial en el registro de defunciones y la asistencia médica recibida, lo que apunta a diferencias regionales en el acceso a servicios de salud.

Estos hallazgos justifican la necesidad de una investigación más profunda para identificar patrones latentes y grupos vulnerables, con el objetivo de generar evidencia que contribuya a políticas públicas más efectivas.

**Preguntas**

¿Cuáles son las principales causas de muerte en diferentes grupos de edad?

¿Hay diferencias en la mortalidad según género, etnia o escolaridad?

¿Existe alguna relación entre la mortalidad y el acceso a servicios de salud?

¿Cómo se compara la mortalidad más actual con la de hace 10 años?, ¿Ha aumentado o disminuido? ¿cuáles podrían ser las posibles razones?

**Conclusiones**

Los datos muestran que la distribución de las defunciones en Guatemala no es homogénea, sino que está influenciada por factores socioeconómicos como la escolaridad y la etnia. La concentración de muertes en personas mayores y en hombres sugiere que ciertos grupos poblacionales enfrentan un mayor riesgo de mortalidad, lo que resalta la importancia de abordar las desigualdades estructurales en salud.

La identificación de patrones de mortalidad relacionados con la edad, el género, la condición socioeconómica y la ubicación geográfica destaca la necesidad de estudios más detallados. Un análisis más profundo permitirá comprender mejor los factores de riesgo y orientar la formulación de políticas públicas que promuevan la equidad en salud y reduzcan las brechas en la atención médica.

**Anexos**

Frecuencia de Sexo:

| | Sexo | Frecuencia |
|---|---|---|
| 0 | Masculino | 480556 |
| 1 | Femenino | 374851 |

Frecuencia de Etnia:

| | Etnia | Frecuencia |
|---|---|---|
| 0 | Ladino o mestizo | 411148 |
| 1 | Maya | 217962 |
| 2 | No especificado | 143746 |
| 3 | Desconocido | 72657 |
| 4 | Extranjero | 9195 |
| 5 | Xinca | 371 |
| 6 | Garífuna | 328 |

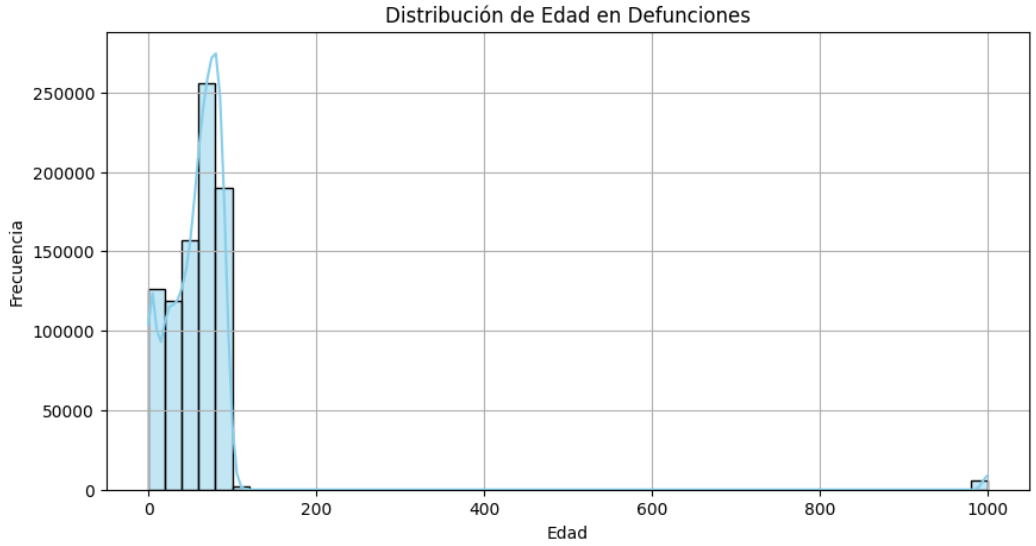
Frecuencia de Departamento:

| | Departamento | Frecuencia |
|---|---|---|
| 0 | Guatemala | 249498 |
| 1 | Quetzaltenango | 54541 |
| 2 | Alta Verapaz | 50846 |
| 3 | Escuintla | 49574 |
| 4 | San Marcos | 49537 |
| 5 | Huehuetenango | 47907 |
| 6 | Quiché | 41807 |
| 7 | Suchitepéquez | 32196 |
| 8 | Chimaltenango | 30097 |

Frecuencia de Escolaridad:

| | Escolaridad | Frecuencia |
|---|---|---|
| 0 | Sin escolaridad | 453598 |
| 1 | Primaria | 233105 |
| 2 | Diversificado | 57398 |
| 3 | No especificado | 56472 |
| 4 | Básico | 40381 |
| 5 | Universitaria | 14157 |
| 6 | Postgrado | 296 |

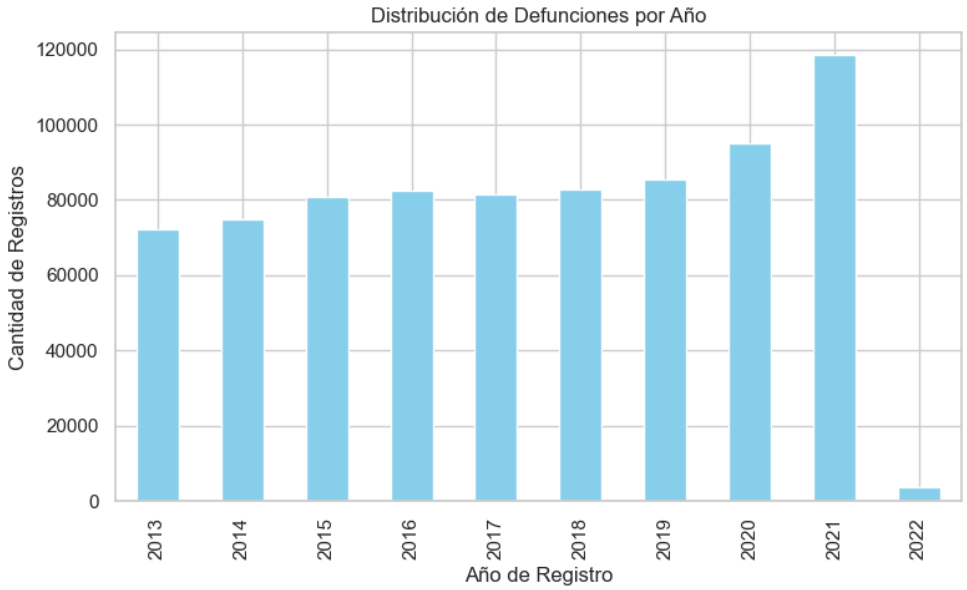
Este análisis muestra la distribución de las defunciones según variables clave como sexo, etnia, departamento y escolaridad. Se observa que la mayoría de las muertes corresponden a personas sin escolaridad y del sexo masculino. En cuanto a etnia, el grupo ladino o mestizo presenta el mayor número de defunciones, seguido por población maya. Departamentos como Guatemala y Alta Verapaz concentran una gran cantidad de casos, lo que podría reflejar tanto una mayor densidad poblacional como desigualdades en el acceso a servicios. Estos patrones ofrecen información valiosa sobre las brechas sociales y territoriales que inciden en la mortalidad en el país.




Este análisis muestra cómo se distribuyen las defunciones según la edad. La media y mediana rondan los 62 años, lo que indica una alta mortalidad en adultos mayores. Sin

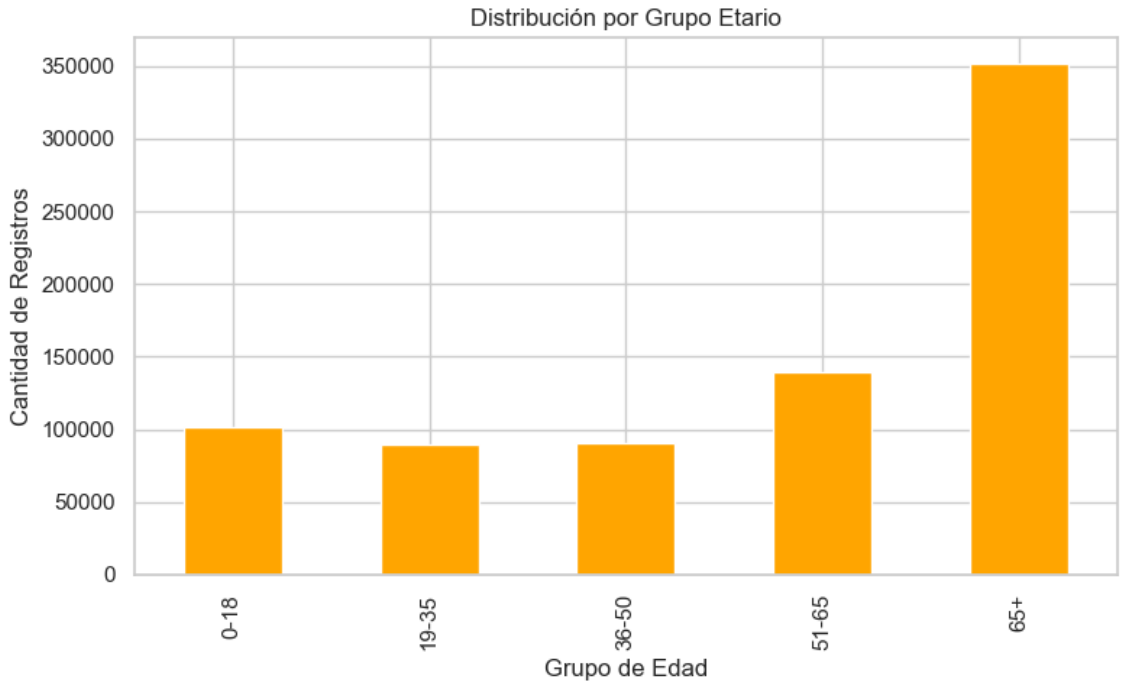
embargo, la moda de 1 año revela un número preocupante de muertes en la infancia. Además, se observan valores atípicos que podrían indicar errores en el registro de datos.


El gráfico muestra las principales causas de muerte en Guatemala entre 2012 y 2021. Las más frecuentes son enfermedades crónicas como infarto al miocardio, diabetes y cirrosis, seguidas por neumonía, COVID-19 y muerte sin asistencia médica. Esto evidencia tanto una alta carga de enfermedades no transmisibles como deficiencias en el acceso a servicios de salud.

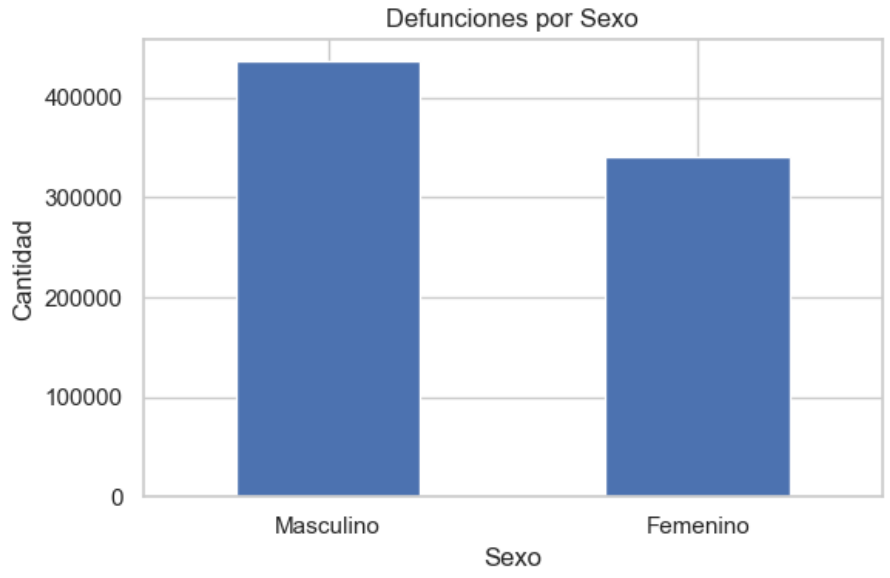



Se muestra una tendencia general creciente en la cantidad de defunciones registradas entre 2013 y 2021. Destacan dos incrementos significativos: uno en 2020 y otro aún más marcado en 2021, lo cual podría estar asociado al impacto de la pandemia por COVID-

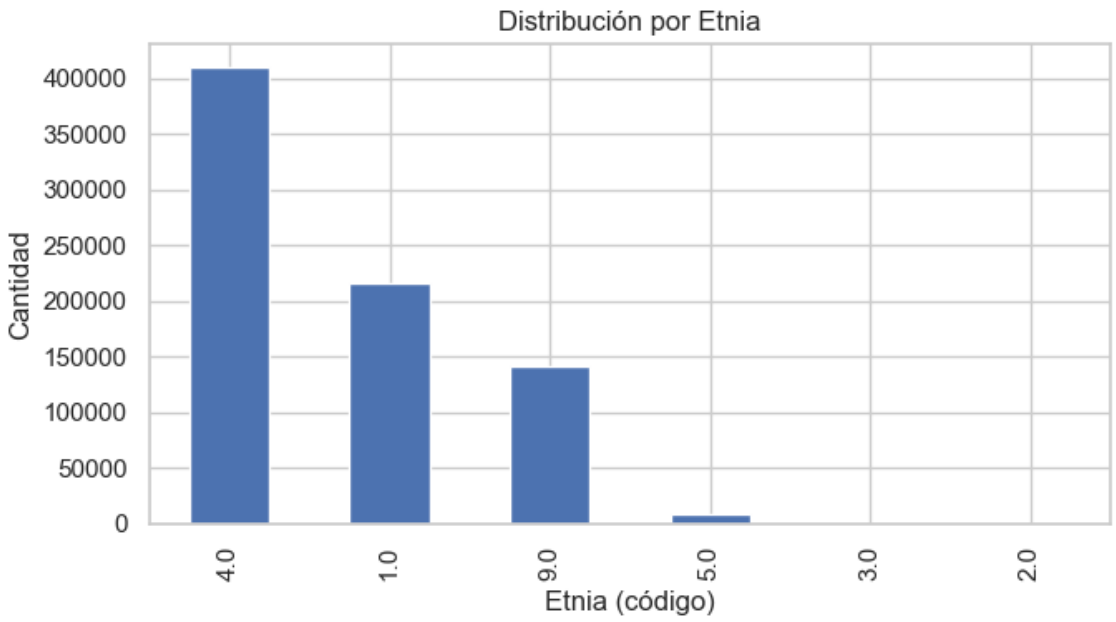
19. En contraste, en 2022 se observa una caída abrupta, probablemente debido a datos aún no consolidados o parciales de ese año.




la mayor concentración de defunciones ocurre en el grupo de personas mayores de 65 años, lo que es coherente con la mayor vulnerabilidad de este segmento poblacional. Los demás grupos etarios presentan cantidades similares, excepto el grupo de 51 a 65 años, que muestra un leve incremento. Esto resalta la necesidad de enfocar políticas de salud pública principalmente en adultos mayores.

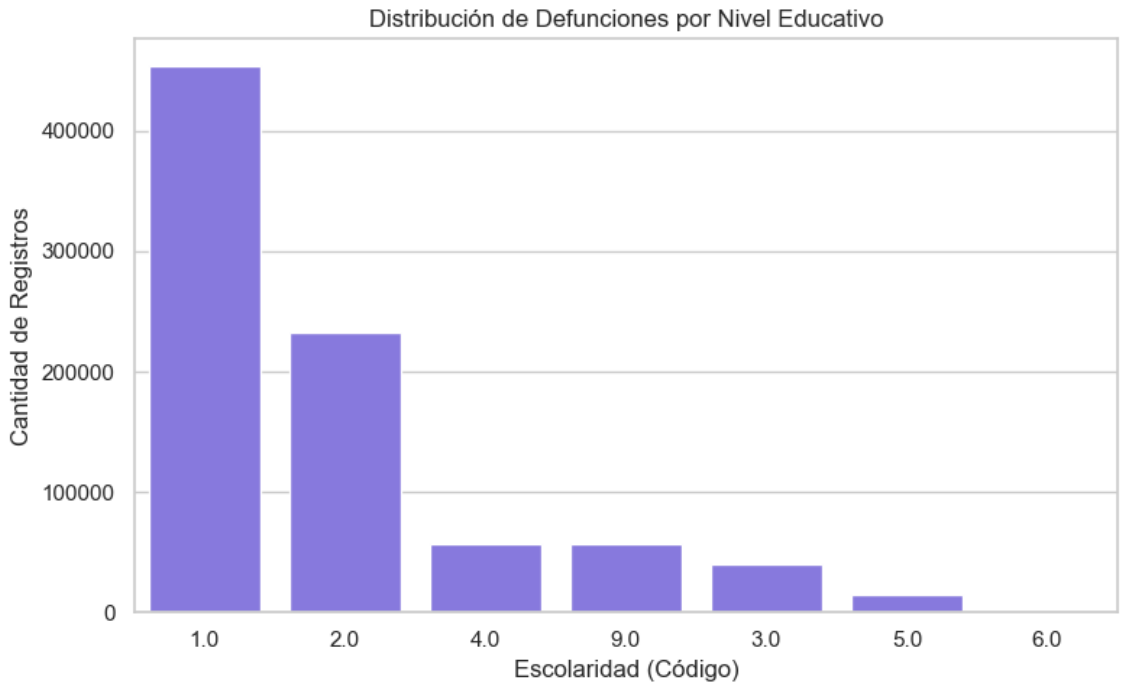


Se muestra una mayor cantidad de defunciones en hombres en comparación con mujeres. Esta diferencia podría explicarse por factores como mayor exposición a riesgos laborales, menor asistencia médica o condiciones de salud menos atendidas en la población masculina.




| | | |
|---|---|---|
| Pueblo de pertenencia del difunto(a) | 1 | Maya |
| | 2 | Garífuna |
| | 3 | Xinka |
| | 4 | Mestizo / Ladino |
| | 5 | Otro |
| | 9 | Ignorado |

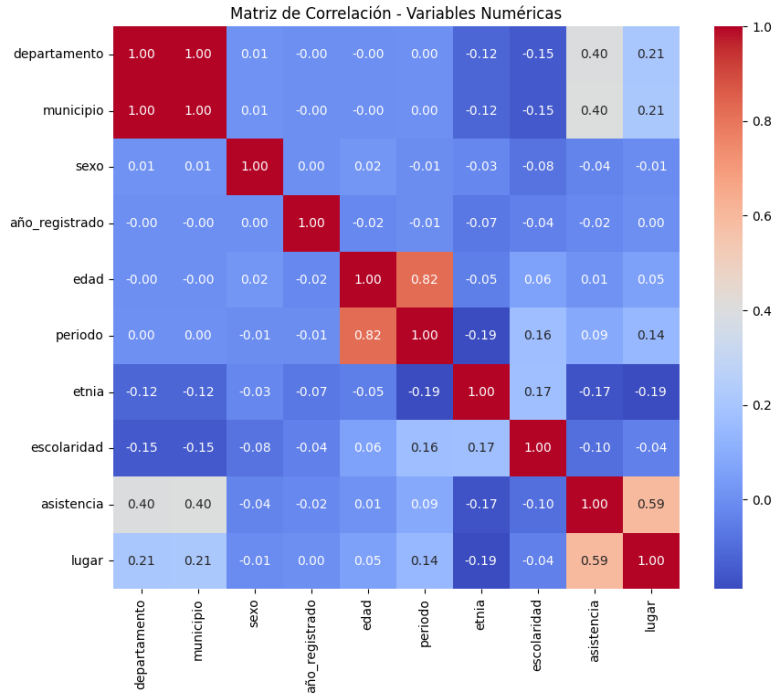
Se observa una concentración alta de defunciones en el grupo con código 4.0, que corresponde a la población ladina o mestiza. Le siguen los grupos étnicos indígenas (1.0, 9.0). Esta distribución refleja tanto la composición demográfica del país como posibles desigualdades en acceso a salud o condiciones socioeconómicas.




| | | |
|---|---|---|
| Escolaridad del difunto(a) | 1 | Ninguno |
| | 2 | Primaria |
| | 3 | Básica |
| | 4 | Diversificado |
| | 5 | Universitario |
| | 6 | Post grado |
| | 9 | Ignorado |

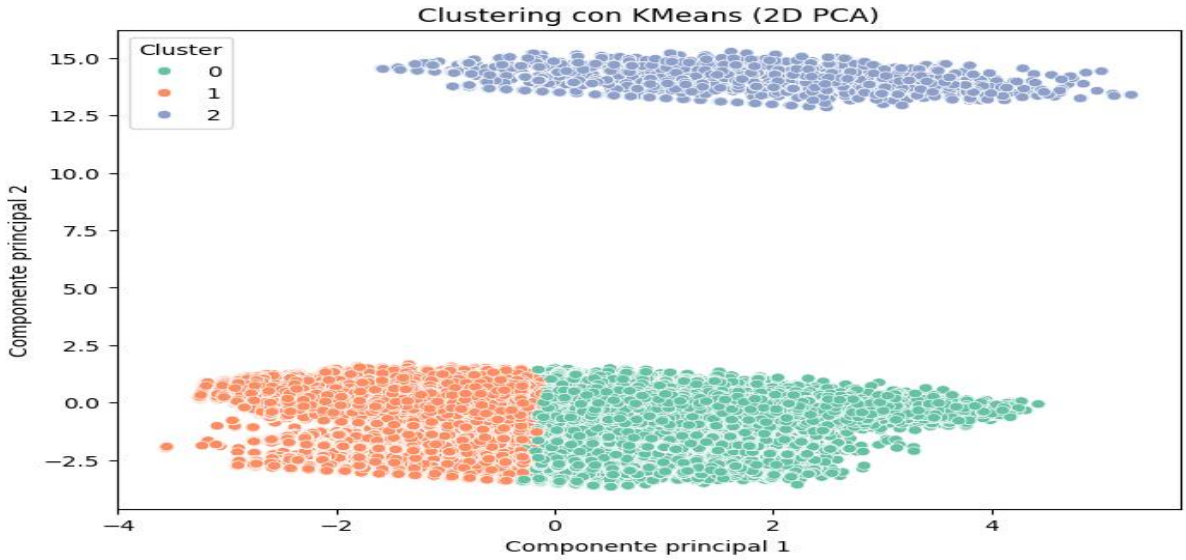
La gráfica muestra que la mayoría de las personas fallecidas registradas tenían un nivel educativo bajo. El grupo más grande corresponde a quienes no tenían ningún nivel de escolaridad formal (código 1), seguido por quienes solo alcanzaron educación primaria (código 2).

Los niveles de educación media y superior (básica, diversificado, universitario y postgrado) aparecen con mucha menor frecuencia, lo cual podría reflejar desigualdades estructurales: personas con menor educación pueden estar expuestas a condiciones de vida más precarias, menor acceso a servicios de salud y, por tanto, mayor riesgo de mortalidad.




Se realizó una matriz de correlación entre las variables numéricas del conjunto de datos con el fin de identificar relaciones significativas que pudieran sugerir asociaciones o posibles predictores. Entre los hallazgos más relevantes se observó una fuerte correlación positiva entre las variables edad y periodo (r = 0.82), lo cual es coherente, ya que la variable periodo agrupa los datos por rangos de edad. Asimismo, se encontró una correlación moderada entre asistencia médica y el lugar donde ocurrió la defunción (r = 0.59), lo cual indica que las muertes en establecimientos de salud están estrechamente asociadas con la presencia de atención profesional.

Por otro lado, las variables departamento y municipio presentaron una correlación perfecta, lo cual era esperable debido a su relación jerárquica territorial. Variables como etnia, sexo y escolaridad mostraron correlaciones más bajas con el resto de las variables, pero su inclusión sigue siendo relevante para análisis multivariados más complejos.




El resultado del clustering reveló tres grupos bien diferenciados de observaciones. La separación clara entre los clusters sugiere la existencia de perfiles diferenciados dentro de las defunciones registradas. Por ejemplo, un grupo puede estar caracterizado por menor acceso a asistencia médica y bajo nivel educativo, mientras que otro podría representar defunciones en contextos institucionales con mayor asistencia profesional. Estos grupos identificados permiten dirigir futuros análisis hacia la comprensión de factores estructurales asociados a la mortalidad, y pueden ser clave para el diseño de intervenciones más focalizadas en salud pública.

Enlace a la presentación:
https://www.canva.com/design/DAGiqPcPpe4/q_OszetLSXkPkV7dEwrw_g/view?utm_content=DAGiqPcPpe4&utm_campaign=designshare&utm_medium=link2&utm_source=uniquelinks&utlId=h0346d7dc7f

Enlace al repositorio: https://github.com/michellemej22596/Proyecto3-MineriadeDatos.git