Análisis Exploratorio de Datos: Mortalidad en Guatemala (2012-2023)

Curso: Minería de Datos
Universidad del Valle de Guatemala
Fecha: Febrero 2026

---

1. Situación Problemática

Guatemala enfrenta retos estructurales profundos en salud pública y desarrollo social. A pesar de los esfuerzos por modernizar el sistema de estadísticas vitales, los datos del Instituto Nacional de Estadística (INE) revelan disparidades críticas que afectan la esperanza de vida de la población.

Durante el análisis exploratorio de más de 950,000 registros de defunciones entre 2012 y 2023, se han identificado tres ejes de vulnerabilidad:

1. **Acceso Desigual a la Salud:** Un alto porcentaje de defunciones ocurre sin asistencia médica profesional, especialmente en áreas rurales y departamentos con mayor índice de pobreza.
2. **Impacto de la Educación:** Existe una correlación aparente entre la falta de escolaridad y la mortalidad prematura, lo que sugiere que el nivel educativo actúa como un determinante de la longevidad.
3. **Violencia y Accidentalidad:** La alta incidencia de muertes en la vía pública, concentrada en hombres en edad productiva, representa una pérdida social y económica significativa para el país.

La pandemia de COVID-19 exacerbó estas debilidades, poniendo a prueba la capacidad de respuesta del sistema de salud en los departamentos más alejados de la capital.

2. Problema Científico

¿En qué medida el acceso a servicios de salud, el nivel educativo y la ubicación geográfica determinan los patrones de mortalidad y la esperanza de vida en la población guatemalteca durante el periodo 2012-2023?

3. Objetivos

Objetivo General

Analizar los patrones sociodemográficos y geográficos de la mortalidad en Guatemala mediante técnicas de minería de datos y análisis exploratorio, para identificar los factores de riesgo principales que afectan a la población.

Objetivos Específicos

1. **Identificar brechas de asistencia:** Cuantificar la disparidad en el acceso a asistencia médica al momento del fallecimiento entre las áreas urbanas y rurales.
2. **Evaluar determinantes educativos:** Determinar la relación estadística entre el nivel de escolaridad alcanzado y la edad promedio de defunción.
3. **Caracterizar perfiles de riesgo:** Segmentar a la población fallecida mediante algoritmos de clustering para identificar grupos con características comunes de vulnerabilidad.
4. **Validar el impacto de eventos críticos:** Evaluar si el COVID-19 afectó desproporcionadamente a los departamentos con menor infraestructura de salud.

---

4. Descripción de los Datos

4.1 Fuente y Alcance

Los datos provienen del Instituto Nacional de Estadística (INE) de Guatemala, sección de Estadísticas Vitales. Se trabajó con archivos en formato SPSS (.sav) descargados directamente del portal oficial del INE.

Se analizaron 5 conjuntos de datos:

Dataset	Registros	Variables	Período
Defunciones	950,793	28	2012-2023
Nacimientos	4,107,969	44	2012-2023
Matrimonios	842,333	23	2012-2023
Divorcios	71,576	19	2012-2023
Defunciones Fetales	28,626	31	2012-2023

Nota: El año 2016 no se encuentra disponible en la fuente del INE para la mayoría de datasets.

El análisis principal se enfocó en el dataset de **defunciones** (950,793 registros, 28 variables de análisis), que contiene información sobre cada fallecimiento registrado en el

país.

4.2 Variables del Dataset de Defunciones

Las 28 variables se clasifican en:

Variables numéricas (4):

- `anioreg`: Año de registro
- `diaocu`: Día de ocurrencia
- `anioocu`: Año de ocurrencia
- `edadif`: Edad del difunto

Variables categóricas (24):

Variables geográficas (departamento, municipio de registro/ocurrencia/residencia/nacimiento), área geográfica (rural/urbano), sexo, período de edad, pueblo de pertenencia (etnia), estado civil, escolaridad, ocupación, país de nacimiento/residencia, nacionalidad, causa de defunción (CIE-10), asistencia médica, lugar de ocurrencia y certificado de defunción.

4.3 Operaciones de Limpieza y Armonización

Para garantizar la consistencia en el análisis de series temporales (2012-2023), se aplicó un pipeline de limpieza automatizado:

1. **Armonización de Esquemas:** Se mapearon alias de columnas que cambian de nombre entre años (ej: `getdif` -> `puedif`, `ocuhom` -> `ciuohom`). Las columnas inexistentes en años específicos se rellenaron con valores nulos para mantener un esquema canónico uniforme.
2. **Tratamiento Inteligente de Etiquetas:** Se utilizaron metadatos SPSS para identificar labels. Si más del 50% de los valores de una variable tenían etiqueta, se aplicó como variable categórica (reemplazando códigos por descripciones). Si menos del 50% tenían etiqueta, se asumió variable numérica y los códigos etiquetados (valores sentinela como 9, 999) se convirtieron en nulos.
3. **Normalización de Texto:** Estandarización de categorías con variaciones ortográficas (ej: `Garífuna` -> `Garifuna`, `Casado(a)` -> `Casado`, códigos ISO de país como `320.0` -> `Guatemala`).
4. **Formato de Almacenamiento:** Los datos armonizados se almacenaron en formato Parquet con compresión zstd, y se registraron como vistas en DuckDB para consultas SQL eficientes.

4.4 Valores Faltantes

Variable	% Nulos	Observación
`areag` (área geográfica)	50.3%	No disponible en todos los años
`anioocu` (año ocurrencia)	23.9%	Registros sin fecha exacta
`edadif` (edad)	0.6%	Cobertura casi completa
Resto de variables	0.0%	Sin valores faltantes

---

5. Análisis Exploratorio

5.1 Variables Numéricas

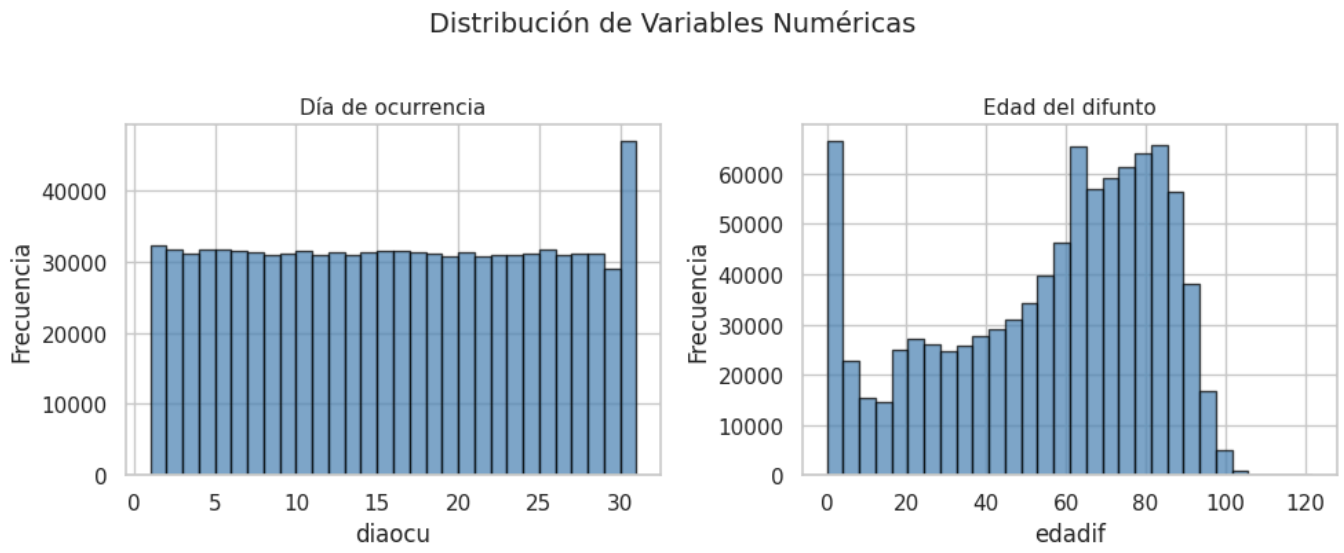
Estadísticas Descriptivas

La variable principal de análisis numérico es `edadif` (edad del difunto):

- **Media:** ~57 años
- **Mediana:** ~62 años
- **Desviación estándar:** ~28 años
- **Asimetría:** -0.536 (sesgada a la izquierda)
- **Rango intercuartílico:** Q1=35, Q3=78 (IQR=43)

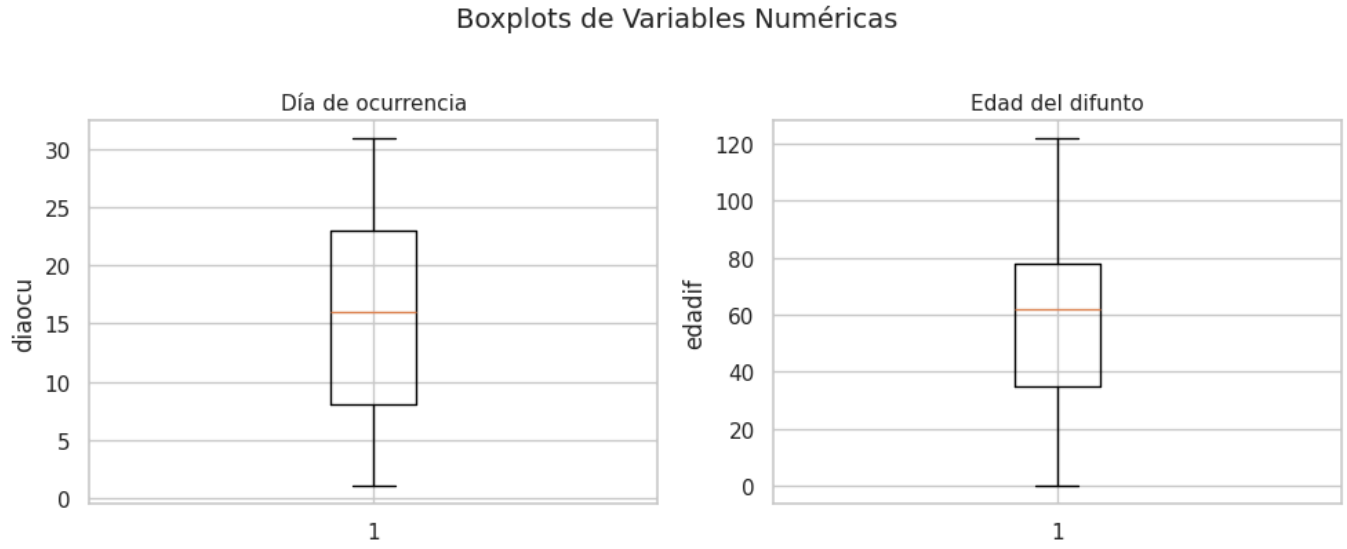
No se detectaron outliers con el método IQR (1.5xIQR), lo cual es coherente con el amplio rango natural de edades de fallecimiento.

Histogramas



La distribución de edad de fallecimiento es bimodal: presenta un pico en edades avanzadas (70-85 años) y un pico secundario en la infancia (mortalidad neonatal), patrón típico de países en desarrollo.

Boxplots



El boxplot de `edadif` confirma la asimetría negativa y el amplio rango intercuartílico. No se identifican valores atípicos fuera de los límites IQR.

Distribución

El test de **Shapiro-Wilk** (muestra de 5,000 observaciones) rechaza la hipótesis de normalidad para ambas variables numéricas:

- `edadif`: W=0.9322, p=6.88e-43 -> **No normal**
- `diaocu`: W=0.9543, p=6.11e-37 -> **No normal**

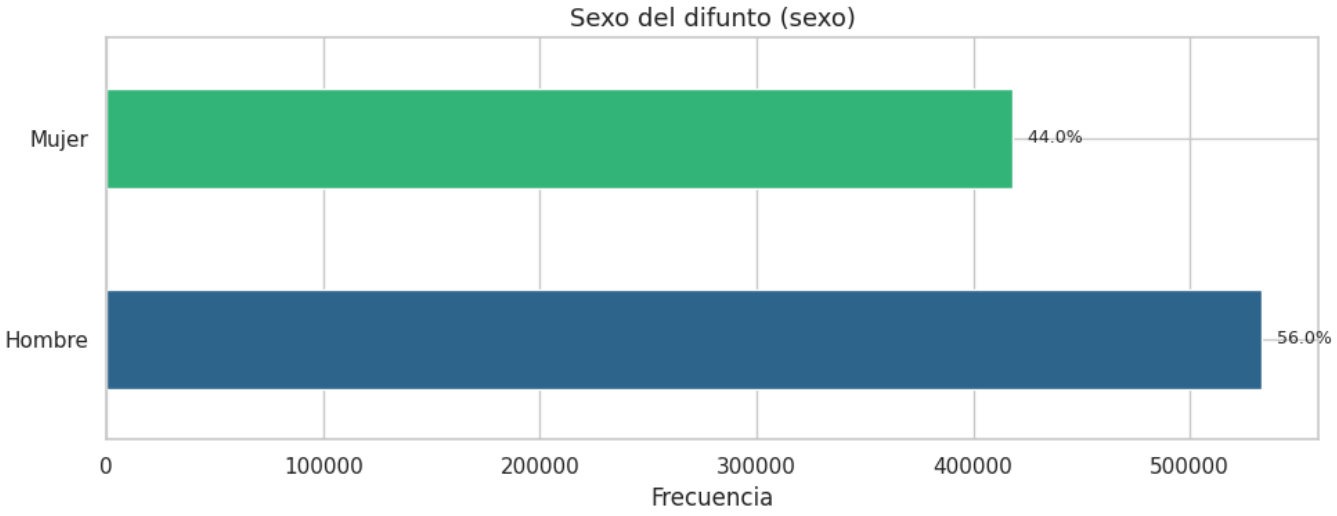
Dado que las variables no siguen una distribución normal, se utilizan pruebas no paramétricas (Kruskal-Wallis) en las hipótesis que involucran comparación de grupos.

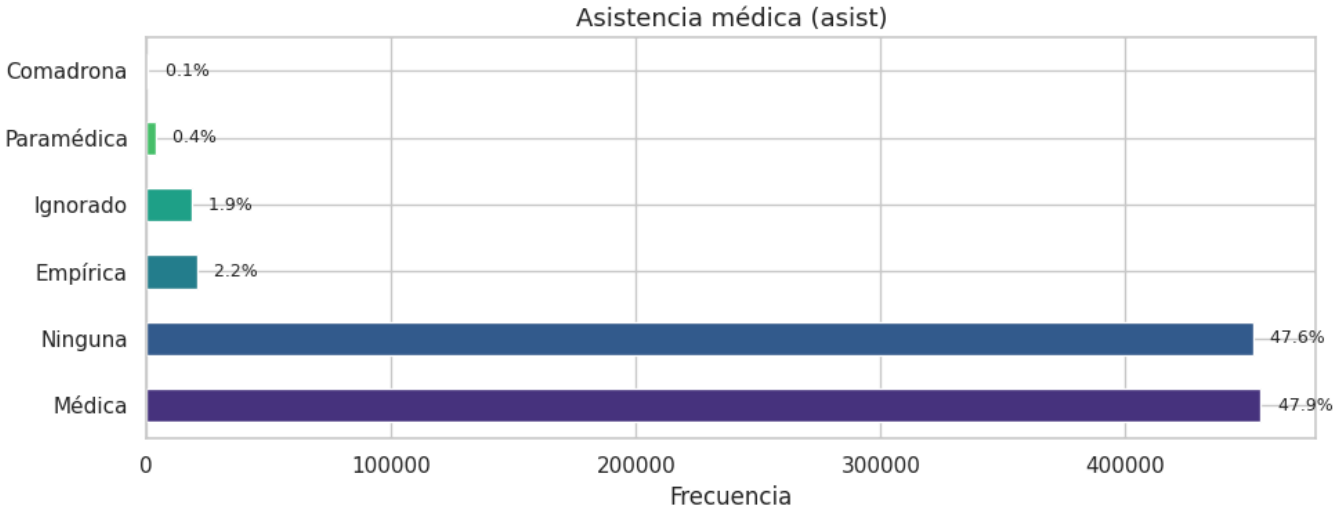
5.2 Variables Categóricas

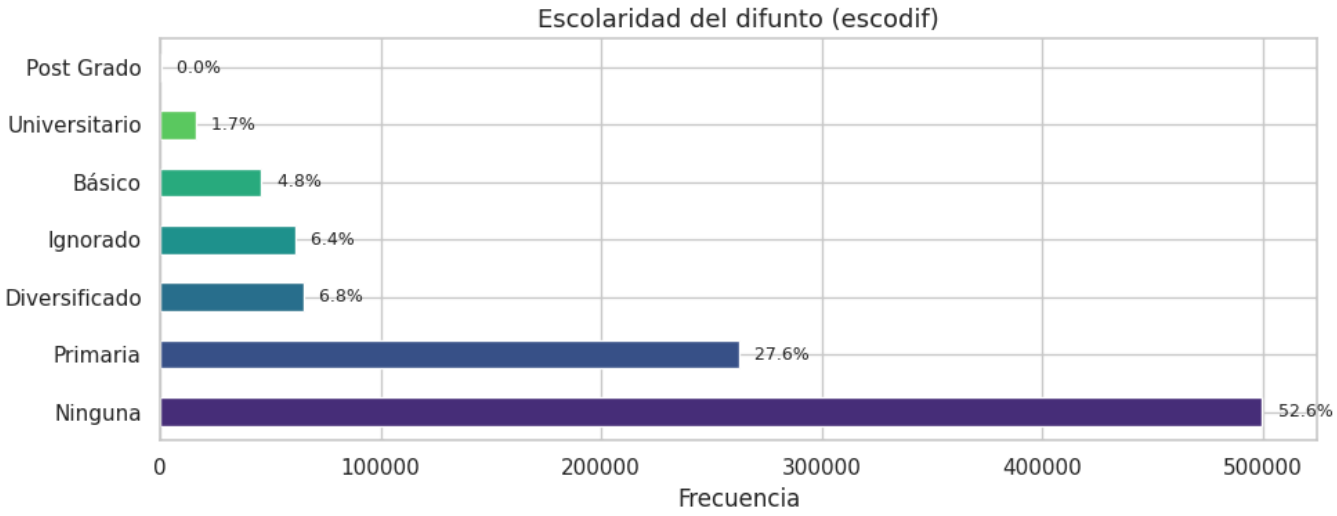
Las tablas de frecuencia revelan los patrones más relevantes:

- **Sexo:** Hombre (54.4%) vs Mujer (45.6%)
- **Área geográfica:** Urbano (54.5%) vs Rural (43.7%) vs Ignorado (1.9%)
- **Asistencia médica:** Ninguna (50.7%) es la categoría más frecuente, seguida de Médica (42.3%)
- **Escolaridad:** Ninguna (52.2%) y Primaria (30.3%) dominan

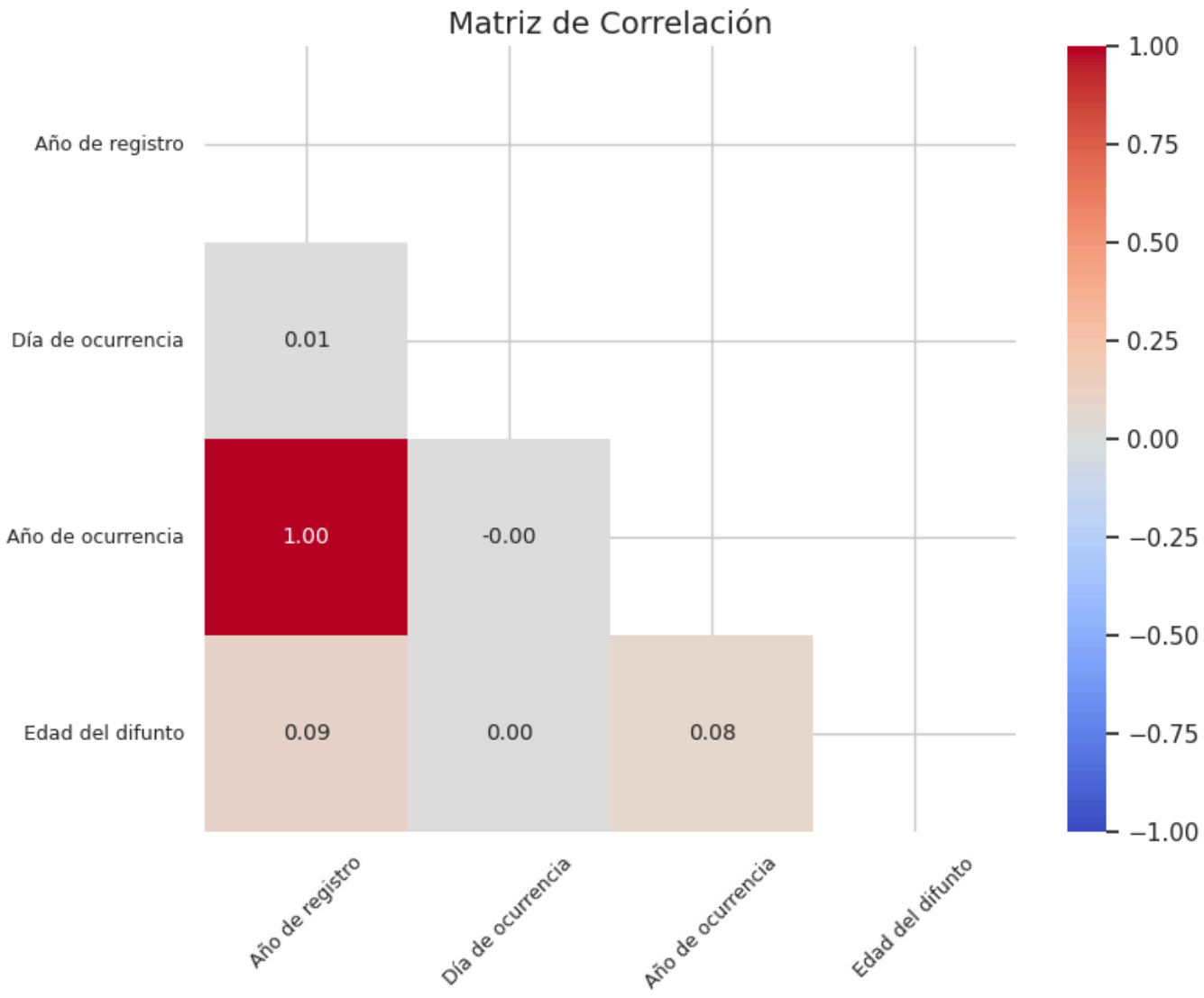
- **Pueblo de pertenencia:** Ladino/Mestizo (47.2%) y Maya (27.9%) son los principales grupos



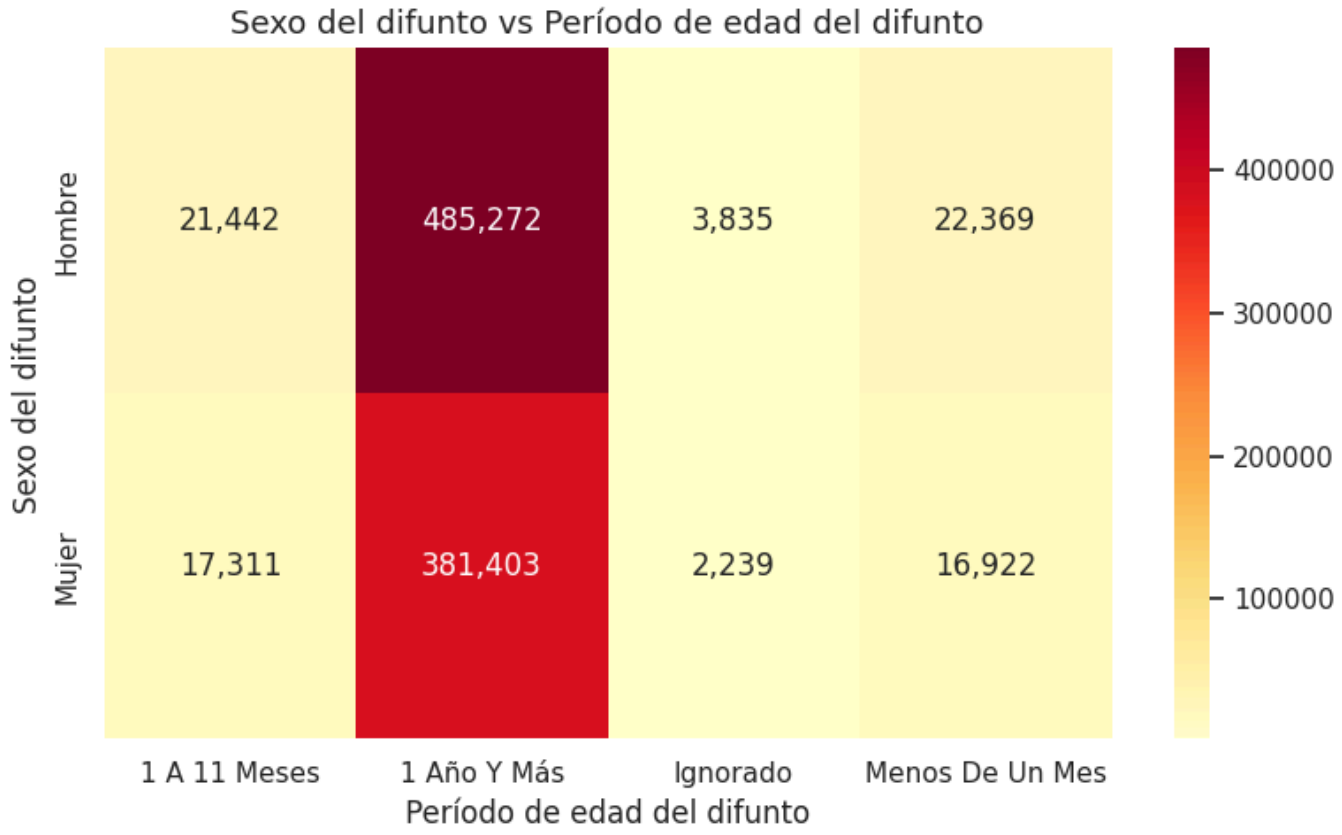




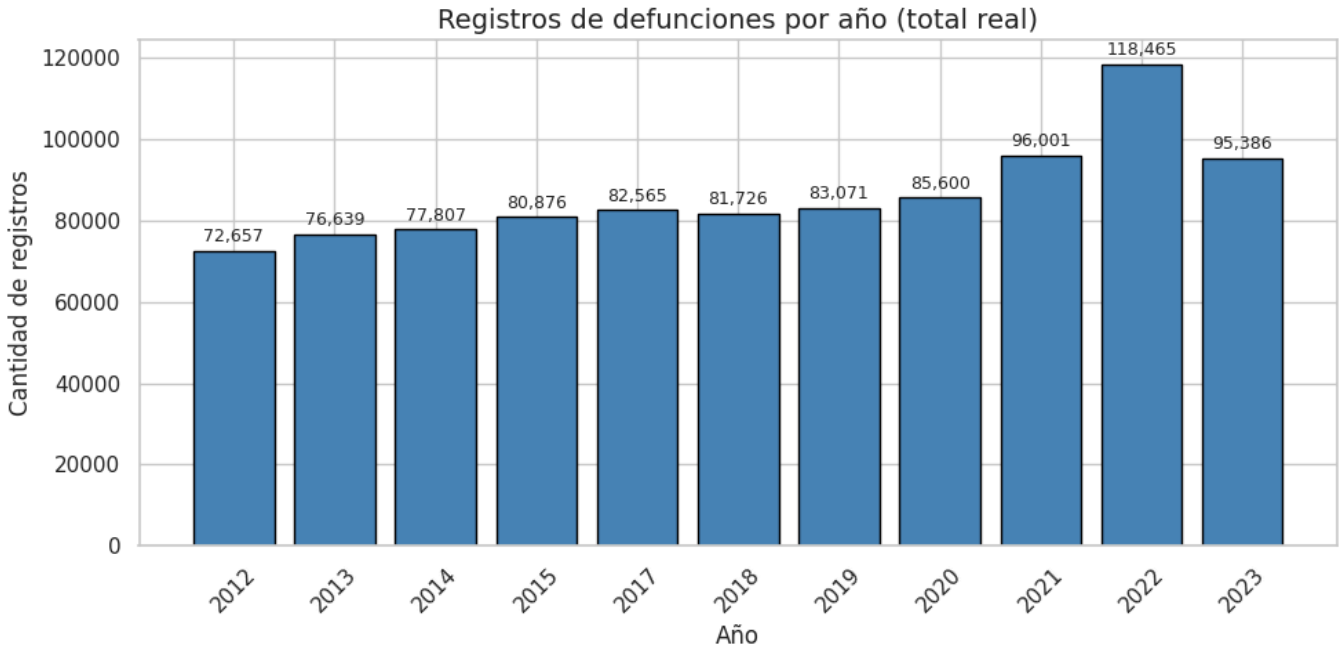
5.3 Correlaciones



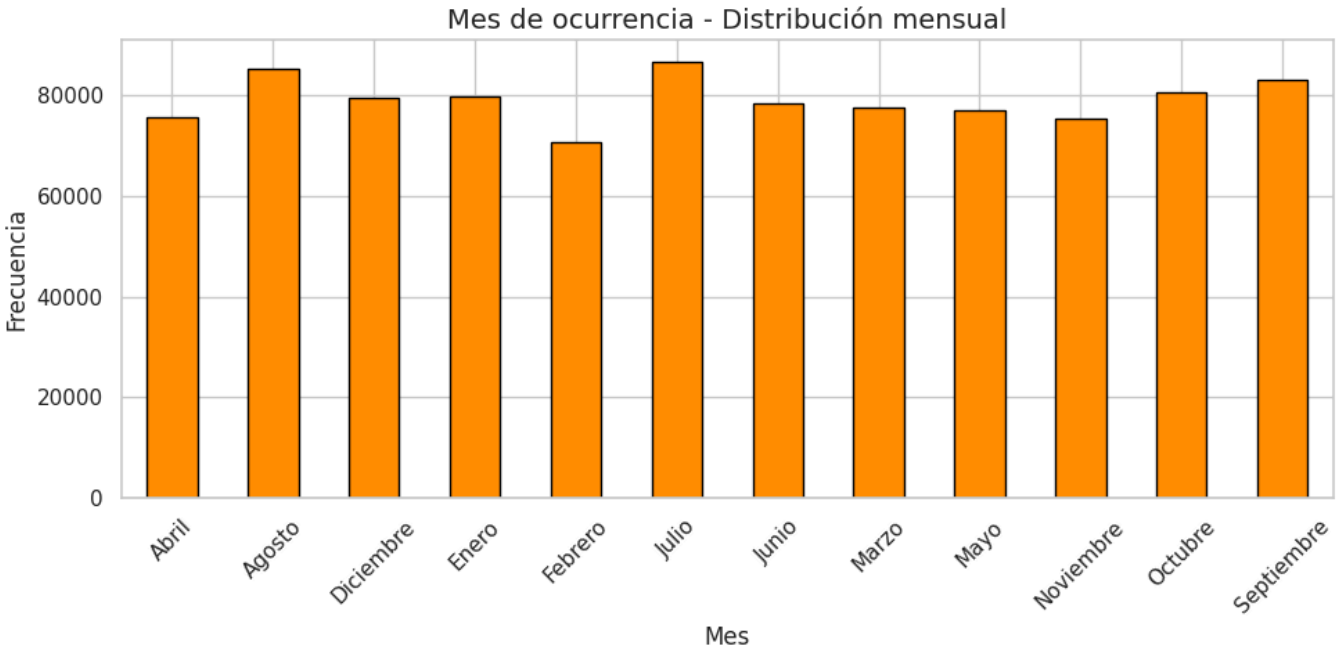
Con solo 2 variables numéricas significativas (edad y día de ocurrencia), la matriz de correlación no revela relaciones lineales fuertes. El análisis más rico se obtiene mediante tablas cruzadas de variables categóricas.



5.4 Análisis Temporal



Los registros por año muestran una tendencia creciente en el número de defunciones registradas, con un pico notable en 2020-2021 coincidente con la pandemia de COVID-19.



La distribución mensual es relativamente uniforme, sin estacionalidad marcada.

---

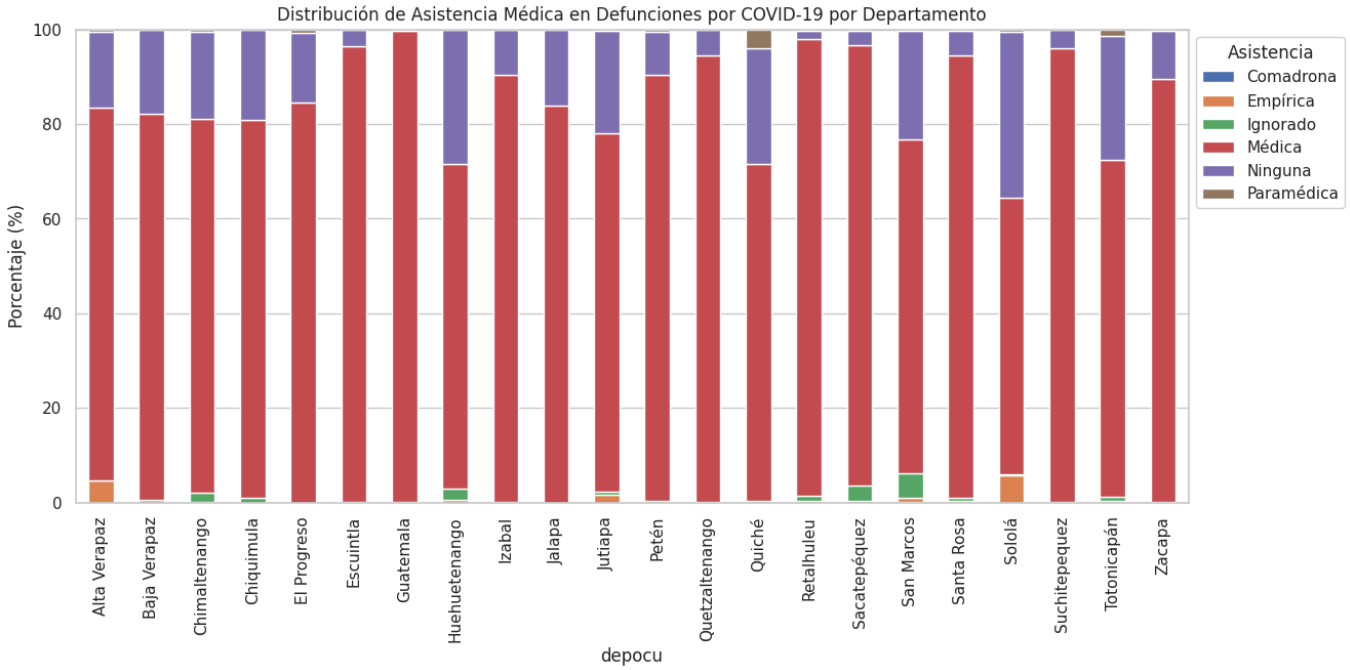
6. Hipótesis de Investigación

Para cada hipótesis se definió una hipótesis nula (H0) y una alternativa (H1), evaluadas con nivel de significancia alpha = 0.05.

H1: Impacto del COVID-19 en departamentos con menor acceso a salud

Supuesto: Los departamentos con mayor índice de "Sin Asistencia Médica" mostraron una tasa de mortalidad por COVID-19 más alta.

- **H0:** La proporción de defunciones COVID-19 sin asistencia médica es independiente del departamento.
- **H1:** La proporción difiere significativamente entre departamentos.
- **Prueba:** Chi-cuadrado de independencia



Resultados: De las 24,579 defunciones por COVID-19 (códigos CIE-10 U07):

- Sololá: 35.0% sin asistencia (el más alto)
- Huehuetenango: 28.3% sin asistencia
- Totonicapán: 26.1% sin asistencia
- Guatemala capital: 0.3% sin asistencia (el más bajo)

Prueba estadística: X2 = 3,322.04, gl = 21, p < 0.001

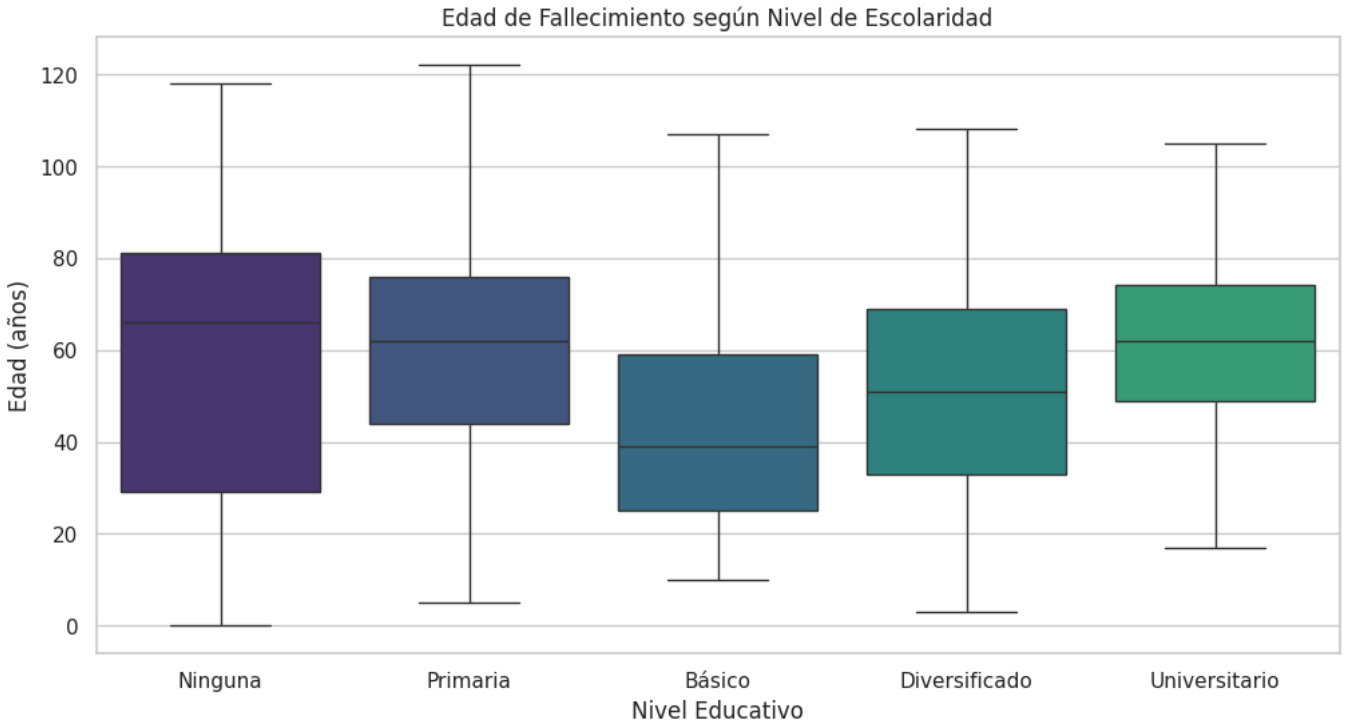
Decisión: Se rechaza H0. La desasistencia médica en muertes COVID depende significativamente del departamento. Los departamentos del altiplano occidental absorbieron la pandemia con recursos mínimos.

Conclusión: CONFIRMADA.

H2: Escolaridad y edad de fallecimiento

Supuesto: Existe una brecha de al menos 10 años en la mediana de edad de fallecimiento entre personas sin escolaridad y aquellas con nivel universitario.

- **H0:** La distribución de edad de defunción es igual para todos los niveles educativos.
- **H1:** Al menos un nivel educativo tiene una distribución significativamente diferente.
- **Prueba:** Kruskal-Wallis H (no paramétrica, dado que `edadif` no sigue distribución normal)



Resultados -- Medianas por nivel educativo:

Nivel	Mediana (años)
Post Grado	70.0
Ninguna	66.0
Primaria	62.0
Universitario	62.0
Diversificado	51.0
Básico	39.0

La brecha Ninguna vs Universitario es de solo 4 años (no 10). Se descubre una **paradoja**: las personas sin escolaridad tienen mediana *mayor* que quienes alcanzaron nivel básico (66 vs 39 años). Esto se explica por un efecto de composición: "Básico" incluye hombres jóvenes fallecidos por causas externas (violencia), mientras que "Ninguna" incluye adultos mayores rurales.

Prueba estadística: H = 18,992.25, p < 0.001

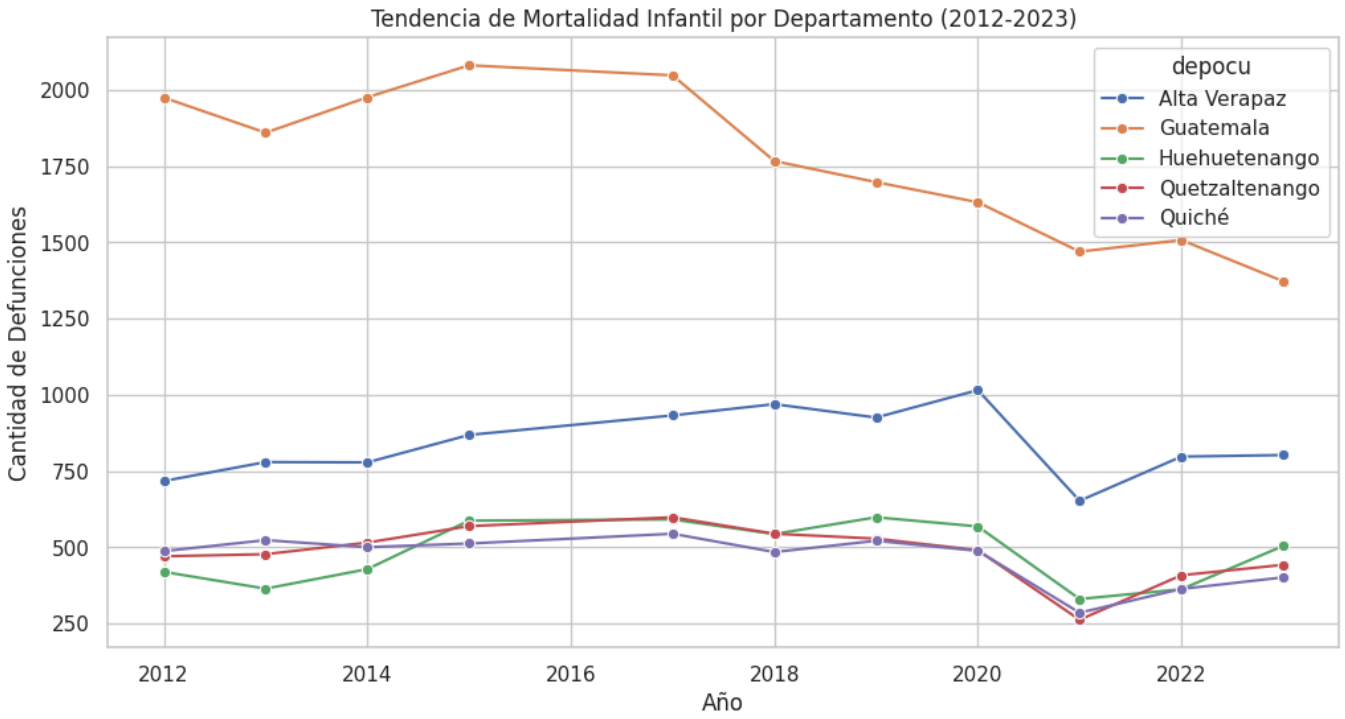
Decisión: Se rechaza H0. Las diferencias entre niveles son significativas, aunque la relación no es lineal como se esperaba.

Conclusión: PARCIALMENTE CONFIRMADA.

H3: Mortalidad infantil en descenso

Supuesto: La mortalidad infantil ha disminuido a nivel nacional, pero con disparidades departamentales persistentes.

- **H0:** La proporción de mortalidad infantil no cambió entre 2012 y 2021.
- **H1:** La proporción disminuyó significativamente.
- **Prueba:** Chi-cuadrado de proporciones



Resultados:

- 2012: 9.80% del total de defunciones (7,121 de 72,657)
- 2021: 5.29% del total de defunciones (5,075 de 96,001)
- Reducción: 28.8% en muertes absolutas

Los departamentos con mayor volumen (Guatemala, Huehuetenango, Alta Verapaz) concentran la mayor parte de las muertes infantiles, con disparidades persistentes.

Prueba estadística: $X2 = 1{,}255.77$, $gl = 1$, $p < 0.001$

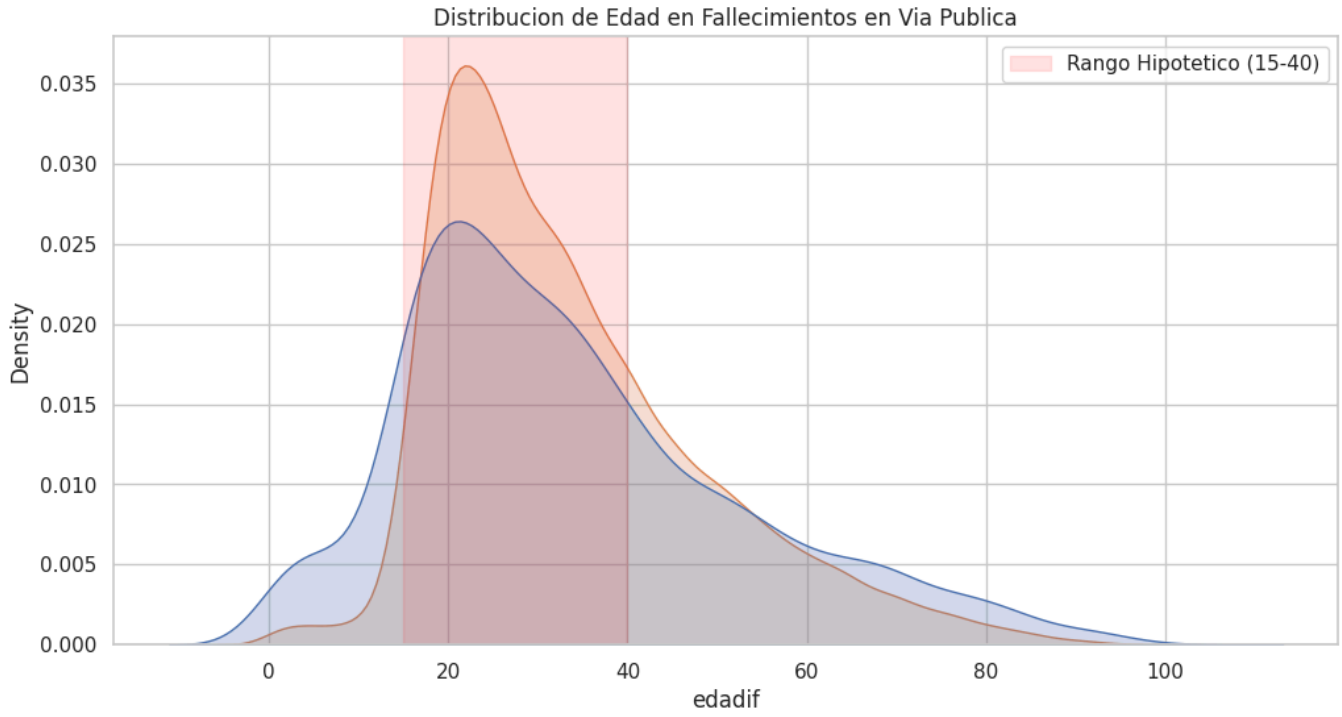
Decisión: Se rechaza H0. La reducción es estadísticamente significativa.

Conclusión: CONFIRMADA.

H4: Muertes en vía pública y hombres jóvenes

Supuesto: Más del 70% de las muertes en vía pública corresponden a hombres entre 15 y 40 años.

- **H0:** La proporción de hombres jóvenes (15-40) en vía pública es menor o igual al 50%.
- **H1:** La proporción es > 50%.
- **Prueba:** Test z de proporción (una cola)



Resultados:

- Total defunciones en vía pública: 25,064
- Hombres: 21,411 (85.4%)
- Hombres jóvenes (15-40 años): 14,911 (59.5% del total)
- Si solo consideramos hombres: 69.6% son jóvenes (15-40)

El porcentaje (59.5%) no alcanza el 70% supuesto, pero la concentración en este grupo demográfico es clara e inequívoca.

Prueba estadística: $z = 30.05$, $p < 0.001$

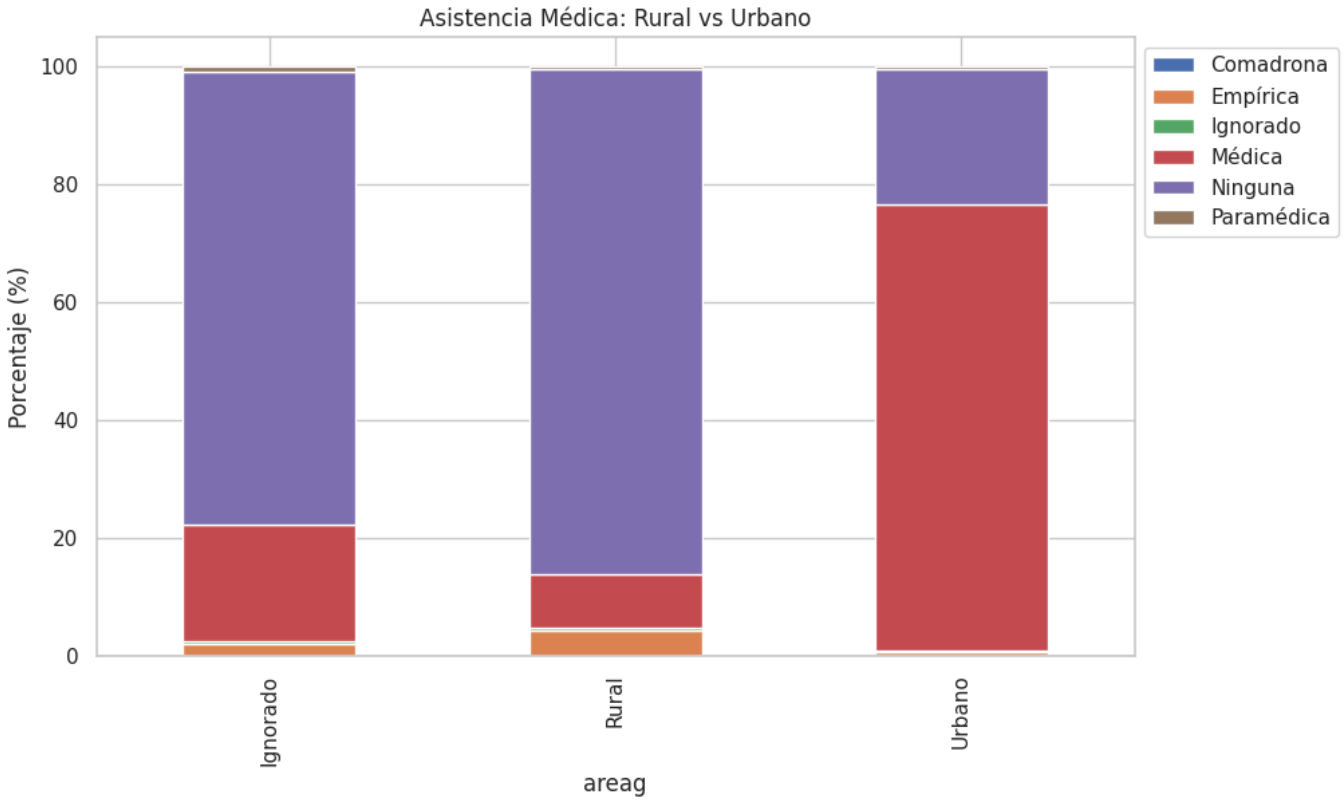
Decisión: Se rechaza H0. La proporción es significativamente mayor al 50%.

Conclusión: CONFIRMADA.

H5: Brecha rural-urbana en asistencia médica

Supuesto: La proporción de personas que fallecen sin asistencia médica en el área rural es al menos el doble que en el área urbana.

- **H0:** La proporción de defunciones sin asistencia es independiente del área geográfica.
- **H1:** La proporción es significativamente mayor en el área rural.
- **Prueba:** Chi-cuadrado de independencia



Resultados:

Área	Sin asistencia	Con asistencia médica
Rural	85.6%	9.2%
Urbano	23.1%	75.6%

La brecha es de **3.7 veces** (85.6% vs 23.1%), superando ampliamente el factor de 2x supuesto.

Prueba estadística: X2 = 178,929.08, gl = 1, p < 0.001

Decisión: Se rechaza H0. La asociación es estadísticamente significativa.

Conclusión: CONFIRMADA CON CRECES.

---

7. Clustering

7.1 Preparación

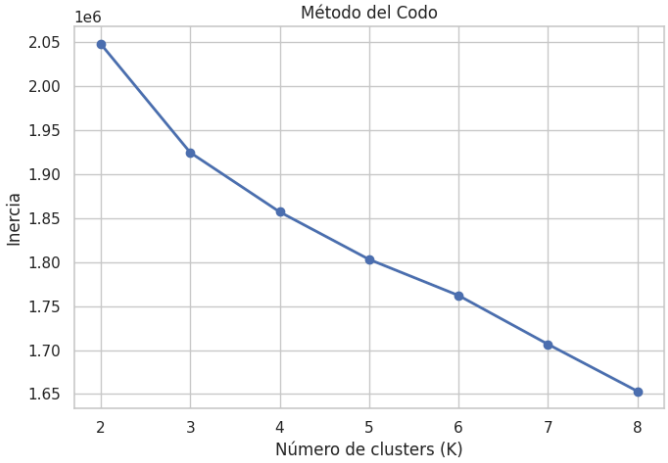
Se seleccionaron 10 variables para el agrupamiento:

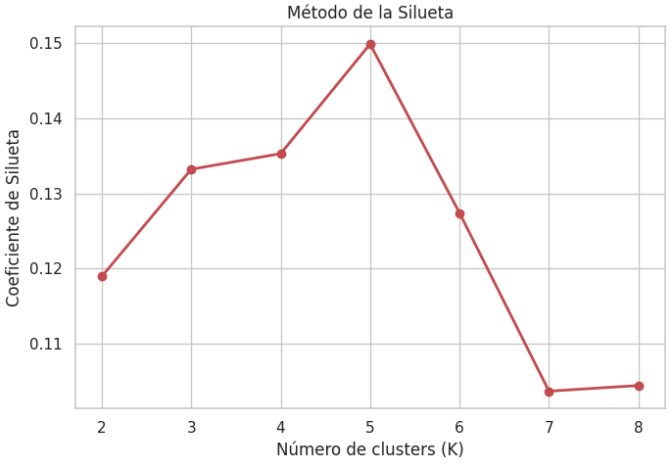
- **Numéricas (2):** día de ocurrencia, edad del difunto
- **Categóricas (8):** sexo, período de edad, pueblo de pertenencia, estado civil, escolaridad, asistencia médica, lugar de ocurrencia, certificado de defunción

Las variables categóricas se codificaron con one-hot encoding y todas se estandarizaron con StandardScaler. Se utilizó una muestra de 50,000 registros para el cálculo de K óptimo, y el clustering final se aplicó al dataset completo.

7.2 Selección de K

Se evaluaron K=2 a K=8 con los métodos del codo e índice de silueta:



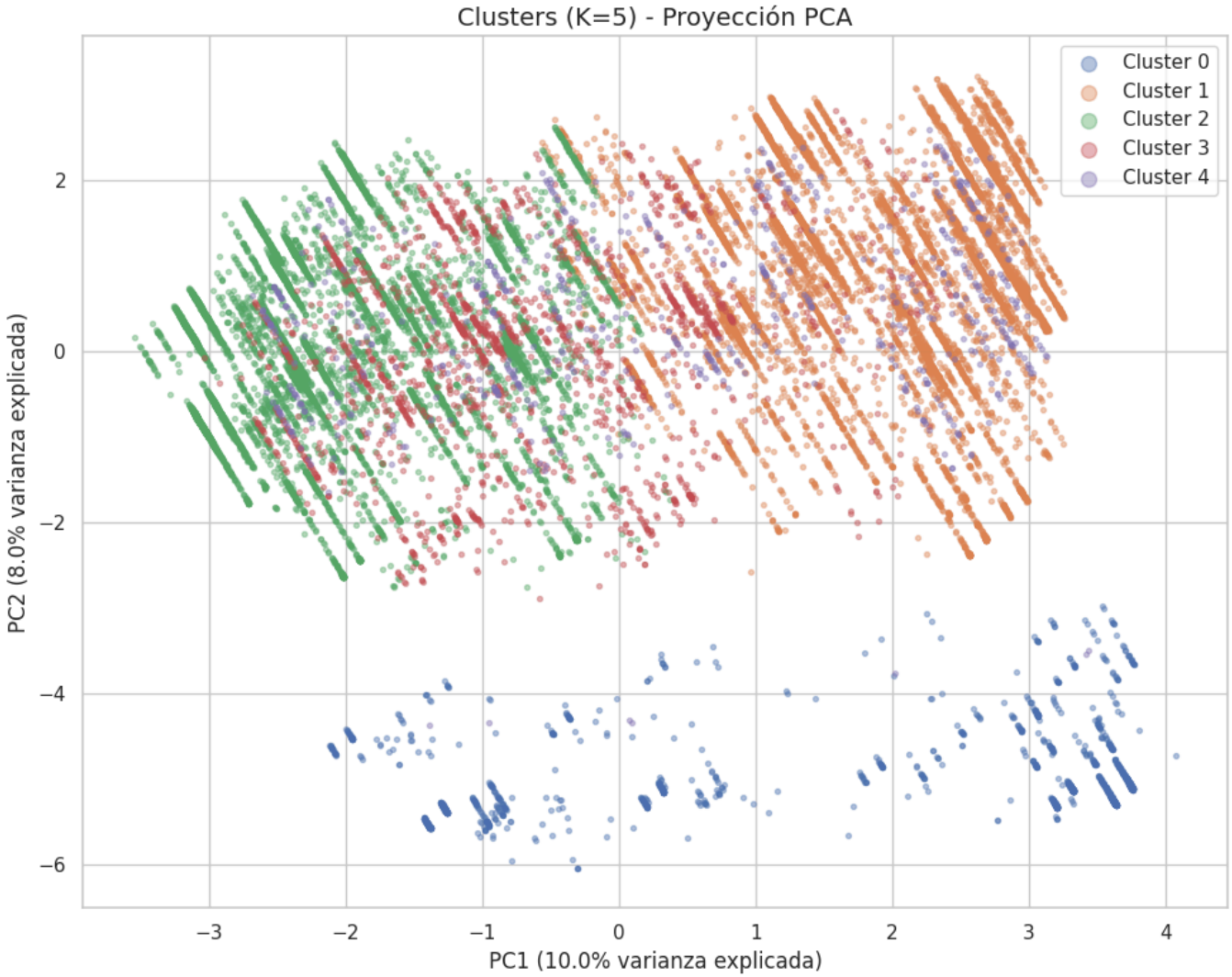


K	Silueta
2	0.1190
3	0.1332
4	0.1353
5	**0.1499**
6	0.1274
7	0.1037
8	0.1045

Se seleccionó **K=5** por maximizar el coeficiente de silueta (0.1499). El valor relativamente bajo es esperado en datos sociodemográficos donde las fronteras entre perfiles no son

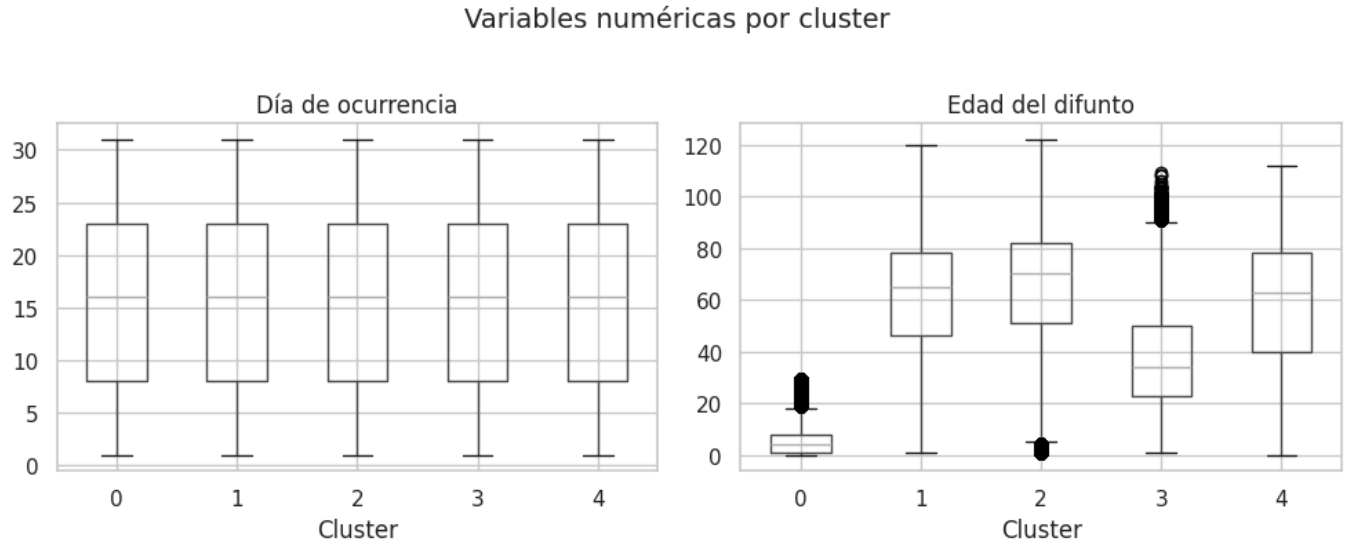
discretas.

7.3 Visualización con PCA



La reducción a 2 componentes principales permite observar la separación entre clusters, especialmente el cluster de mortalidad infantil (cluster 0) y el de jóvenes en vía pública (cluster 3).

7.4 Interpretación de Clusters



Cluster	Nombre	Edad Media	Sexo Dom.	Asistencia	Lugar	% Total
0	Mortalidad Infantil/Neonatal	6.1	Hombre (56.1%)	Médica (64.0%)	Hospital Público (49.7%)	8.2%
1	Adultos Con Asistencia Médica	60.6	Hombre (54.1%)	Médica (99.0%)	Domicilio/Hospital	39.7%
2	Adultos Mayores Rurales (Sin Asistencia)	63.7	Hombre (52.0%)	Ninguna (90.3%)	Domicilio (99.5%)	40.3%
3	Jóvenes -- Muerte Violenta	38.5	Hombre (82.2%)	Ninguna (78.2%)	Vía Pública (27.7%)	8.6%
4	Adultos No Indígenas (Mixto)	58.5	Hombre (59.0%)	Médica (50.8%)	Domicilio (55.3%)	3.3%

Observaciones:

- El cluster 2 es el más grande (40.3%): adultos mayores que fallecen en domicilio sin asistencia, perfil típico del área rural. Junto con el cluster 3, representan el 48.9% de las defunciones sin acceso a salud.
- El cluster 3 (82.2% hombres, edad media 38.5) confirma el hallazgo de la H4 sobre mortalidad violenta en hombres jóvenes.
- El cluster 0 agrupa correctamente las defunciones de menores de 1 año (100% solteros, 99.8% sin escolaridad).

- El cluster 1 (39.7%) es el grupo más grande con acceso a salud: adultos que fallecieron en hospitales públicos, privados o seguro social.

---

8. Hallazgos y Conclusiones

8.1 Resumen de Hallazgos

1. **Brecha de asistencia médica rural-urbana:** El hallazgo más contundente. El 85.6% de las defunciones rurales ocurre sin asistencia médica (3.7x la tasa urbana de 23.1%). Confirmado estadísticamente ($X2 = 178,929$, $p < 0.001$).
2. **COVID-19 exacerbó desigualdades existentes:** Los departamentos del altiplano occidental (Sololá, Huehuetenango, Totonicapán) absorbieron la pandemia con menos del 35% de asistencia médica en muertes COVID.
3. **Paradoja educativa:** La relación escolaridad-longevidad no es lineal. Personas sin escolaridad (mediana 66 años) viven más que personas con educación básica (mediana 39 años), debido al sesgo de mortalidad violenta en jóvenes con educación básica.
4. **Mortalidad violenta masculina:** 85.4% de las muertes en vía pública son hombres, con concentración en el rango 15-40 años (59.5%).
5. **Mortalidad infantil en descenso:** Reducción de 9.80% a 5.29% del total de defunciones (2012-2021), aunque con disparidades departamentales persistentes.

8.2 Nombres de los Clusters

ID	Nombre del Grupo	Característica Principal
0	Mortalidad Infantil/Neonatal	Menores de 1 año en hospitales
1	Adultos Con Asistencia Médica	Fallecimiento institucionalizado
2	Adultos Mayores Rurales	Sin asistencia, en domicilio
3	Jóvenes -- Muerte Violenta	Hombres en vía pública
4	Adultos No Indígenas	Asistencia mixta

8.3 Siguientes Pasos

1. **Análisis de causas específicas:** Aplicar técnicas de NLP sobre los códigos CIE-10 para agrupar causas de muerte en categorías epidemiológicas.
2. **Normalización poblacional:** Integrar datos del censo INE para calcular tasas de mortalidad por cada 100,000 habitantes, eliminando el sesgo por tamaño de

departamento.
3. **Modelado predictivo:** Desarrollar modelos de clasificación para predecir el riesgo de fallecer sin asistencia médica, usando variables sociodemográficas como predictores.

Nota: Todos los análisis fueron realizados con Python (pandas, scipy, scikit-learn) sobre datos crudos del INE. No se utilizaron paquetes de análisis exploratorio automático. El código fuente está disponible en el repositorio de GitHub del proyecto.