Proyecto 1 - Análisis exploratorio

Pablo José Méndez Alvarado

Ciudad de Guatemala, Guatemala

14 de febrero de 2026

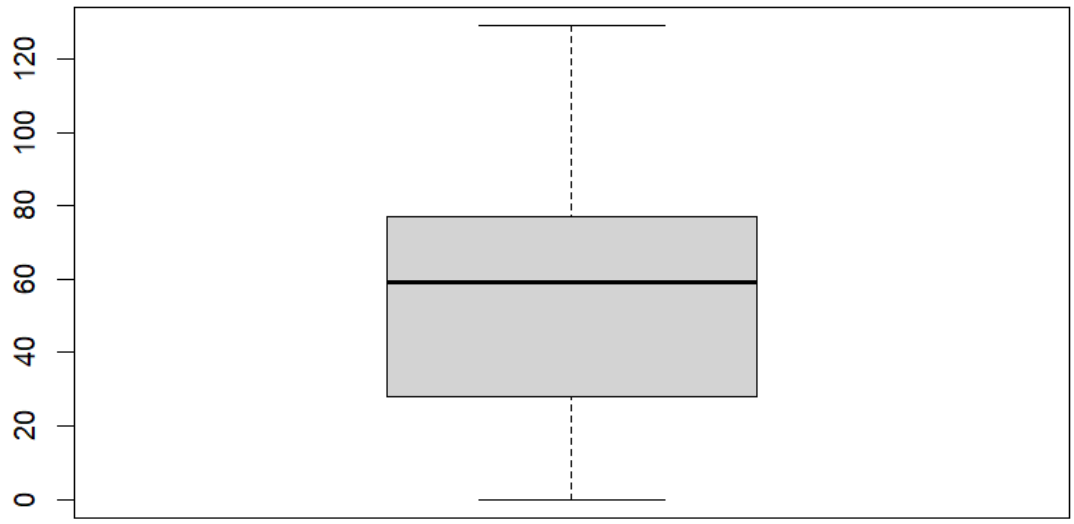
Tabla de contenido

1.

a.

La base de datos analizada contiene 772,150 registros y 36 variables, correspondientes a registros de defunciones. Las variables incluyen información demográfica, geográfica, temporal y características del evento. Estas tienen variables cuantitativas: Edadif, Diaocu, anio; y categóricas: Sexo, Areag, Getdif, Ecidif, Depocu, Mesocu, etc.

b.



Entre las variables cuantitativas más relevantes se encuentra la edad del difunto (Edadif), la cual presenta:

Edad mínima: 0 años

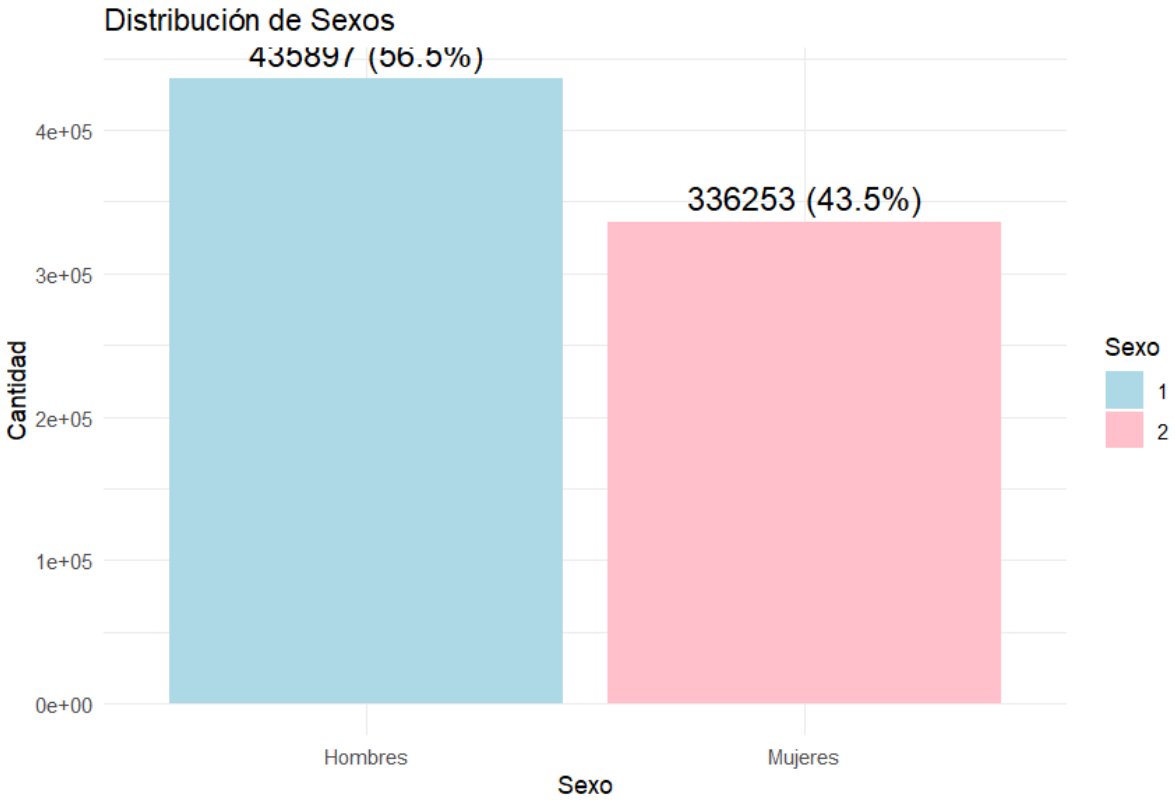
Edad máxima: 129 años

Media: 52.58 años

Mediana: 59 años

Se realizó limpieza de datos eliminando valores codificados como 999, los cuales representaban datos ignorados. El análisis de normalidad mediante la prueba de Shapiro-Wilk (muestra de 5000 datos) arrojó un p-valor < 2.2e-16, lo que indica que la distribución de

edades no sigue una distribución normal. Esto también se observa visualmente en el histograma y el diagrama de caja, donde se aprecia asimetría en la distribución.



La variable sexo fue analizada mediante tablas de frecuencia y gráficos de barras. Los resultados muestran:

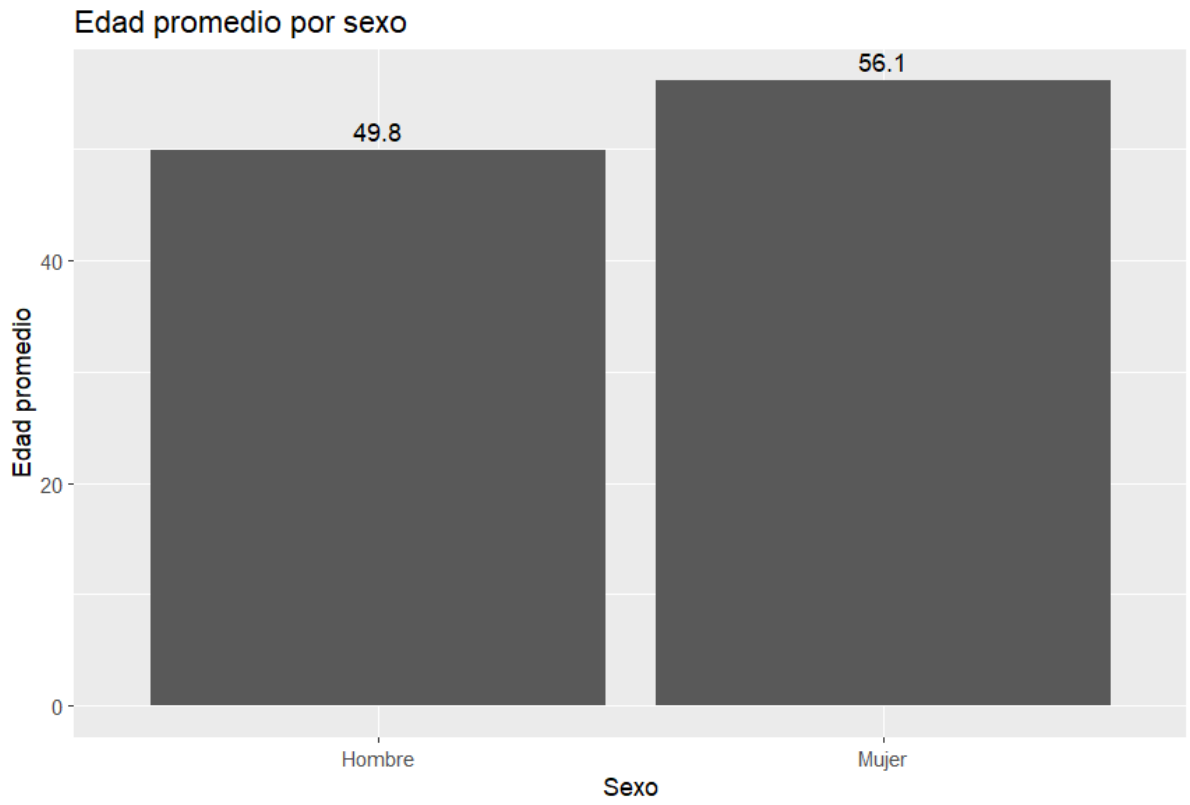
Sexo	Cantidad	Porcentaje
Hombres	435,897	56.5%
Mujeres	336,253	43.5%

Se observa una mayor proporción de defunciones en hombres en comparación con mujeres.

c.

Se realizó un análisis comparativo de la edad promedio de defunción según sexo y de la evolución del número de defunciones a lo largo del tiempo.

Edad promedio por sexo



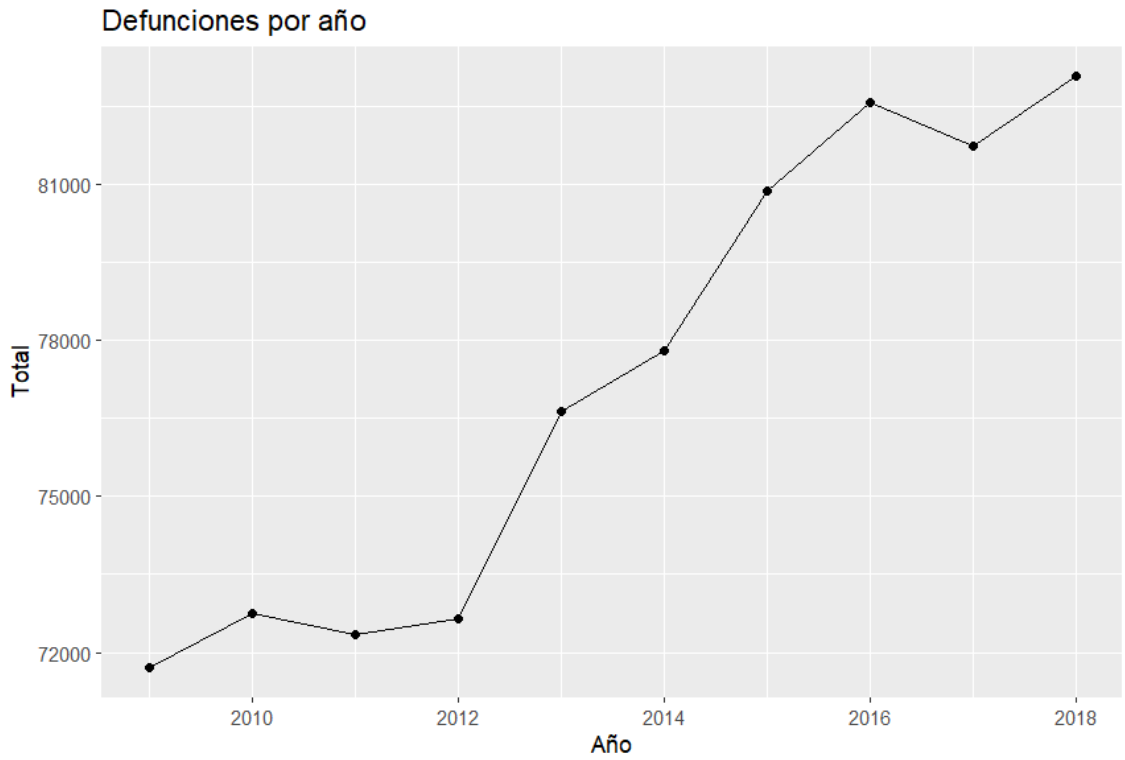
Los resultados muestran que:

Edad promedio hombres: 49.84 años

Edad promedio mujeres: 56.13 años

Esto indica que, en promedio, las mujeres fallecen a edades mayores que los hombres. La diferencia observada sugiere posibles diferencias en condiciones de salud, exposición a riesgos o factores biológicos y sociales entre ambos sexos.

Defunciones por año

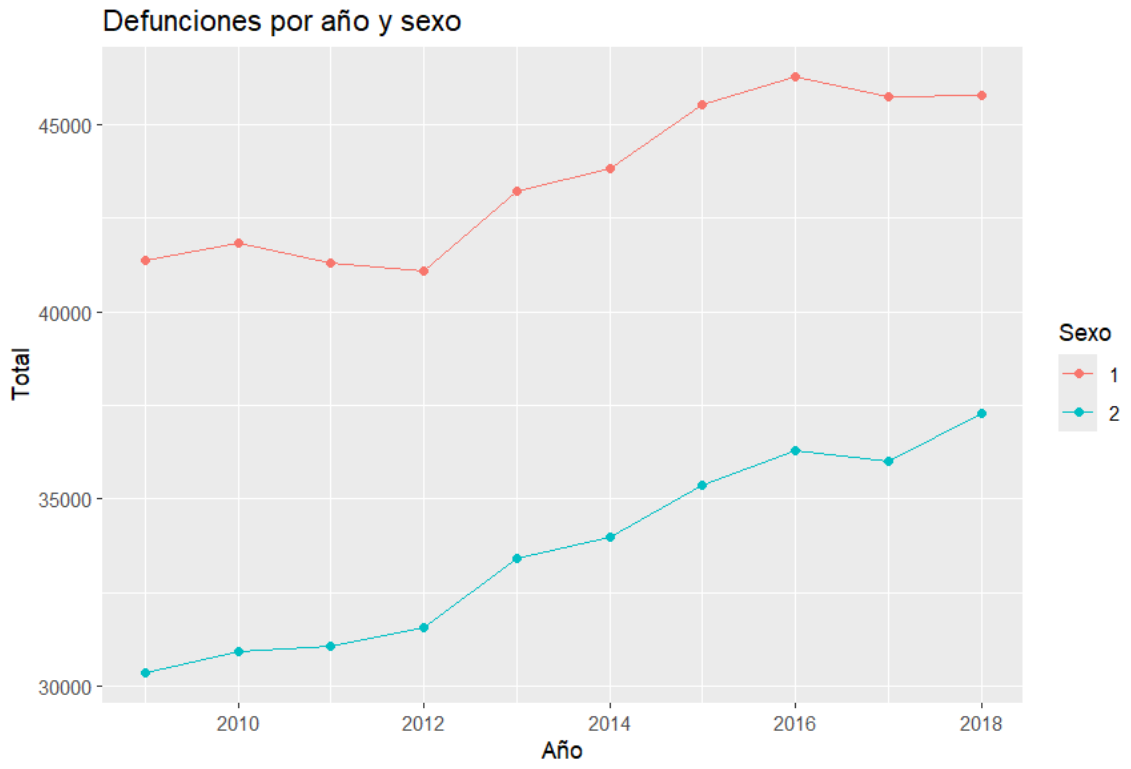


Año	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Total	71707	72748	72354	72657	76639	77807	80876	82565	81726	83071

El análisis temporal muestra la siguiente evolución del total de defunciones:

Se observa una tendencia general al aumento de defunciones a lo largo del tiempo, con pequeñas variaciones entre años. El valor más alto se presenta en 2018.

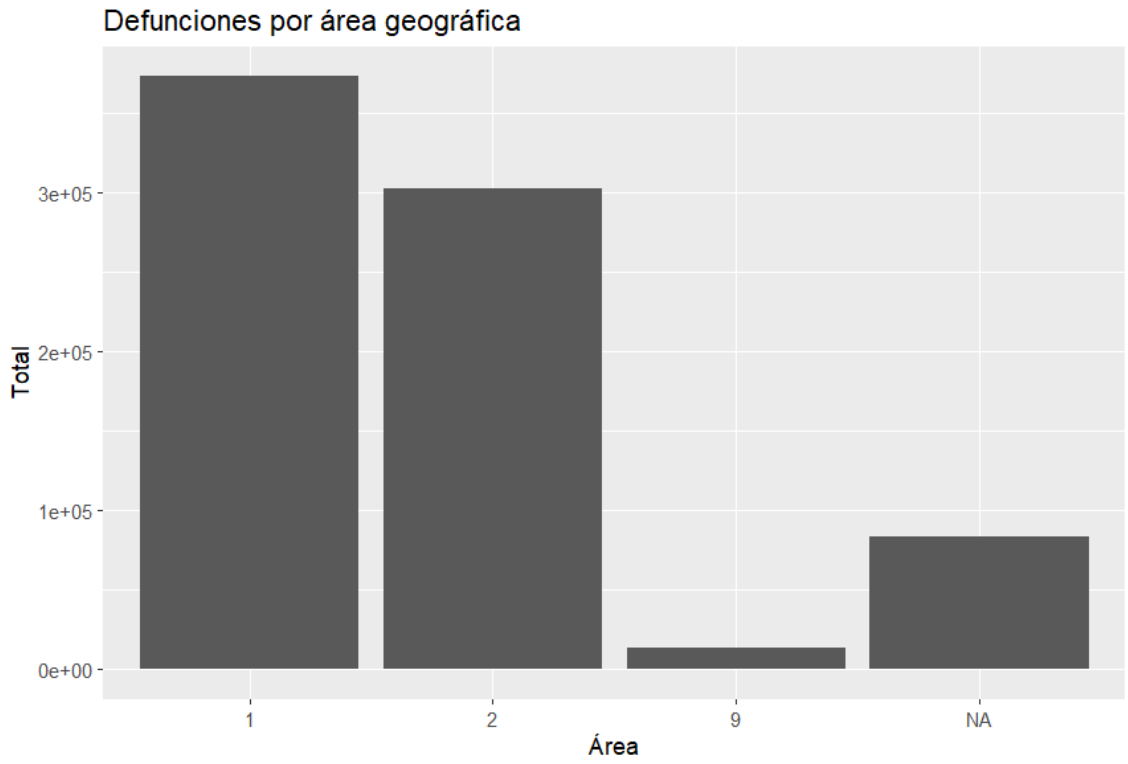
Defunciones por año y sexo



Al analizar conjuntamente año y sexo, se observa que:

En todos los años, el número de defunciones en hombres es mayor que en mujeres. Ambos sexos muestran una tendencia creciente a lo largo del tiempo. El incremento es relativamente paralelo entre hombres y mujeres. Esto sugiere que los factores que explican el aumento general de defunciones afectan de forma similar a ambos sexos, aunque manteniendo la diferencia en magnitud.

Defunciones por área geográfica



La mayor proporción de defunciones ocurre en el área geográfica 1 Urbano (54.13%), seguida del área 2 Rural (43.89%), el resto fueron ignorados o no están especificados.

Esto podría explicarse por:

Mayor concentración poblacional en esa área

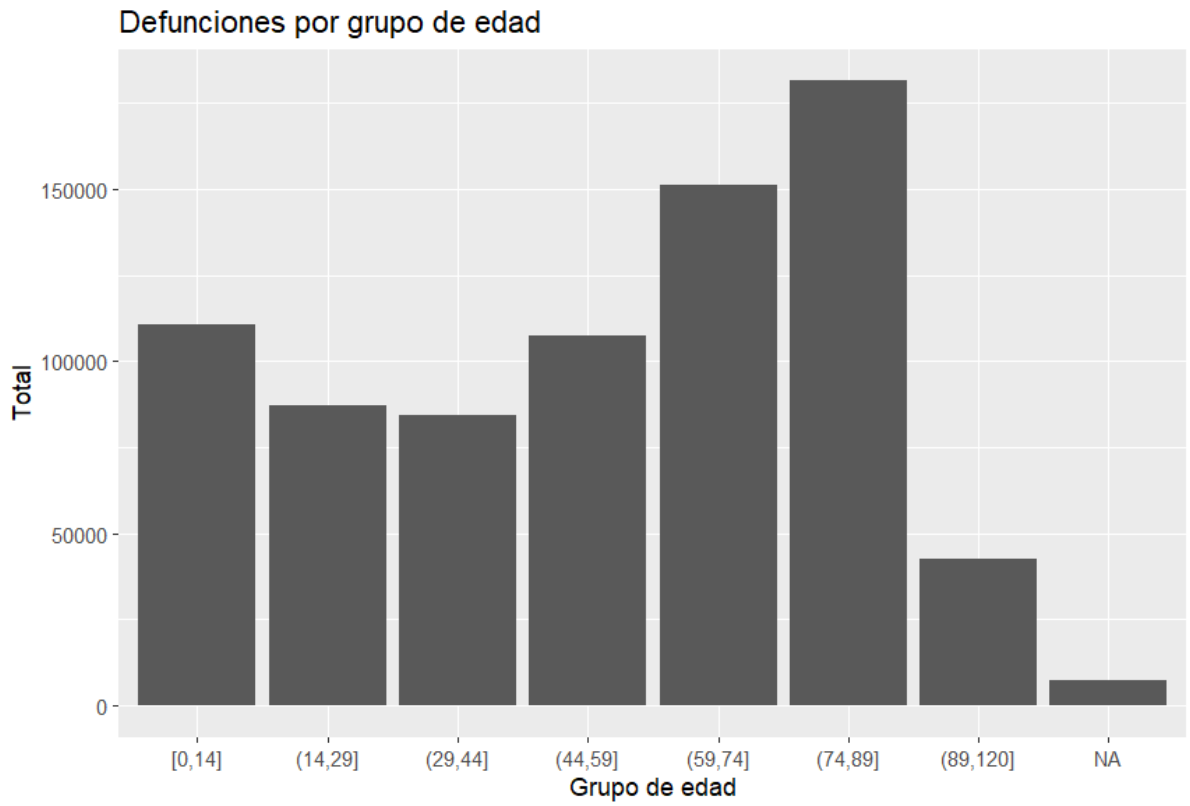
Diferencias en acceso a servicios de salud

Factores socioeconómicos

Diferencias en estilos de vida

La categoría ignorada es relativamente baja (≈2%), lo que sugiere buena calidad en el registro de datos.

Defunciones por grupo de edad



Se observa una clara concentración de defunciones en edades avanzadas, especialmente en los grupos:

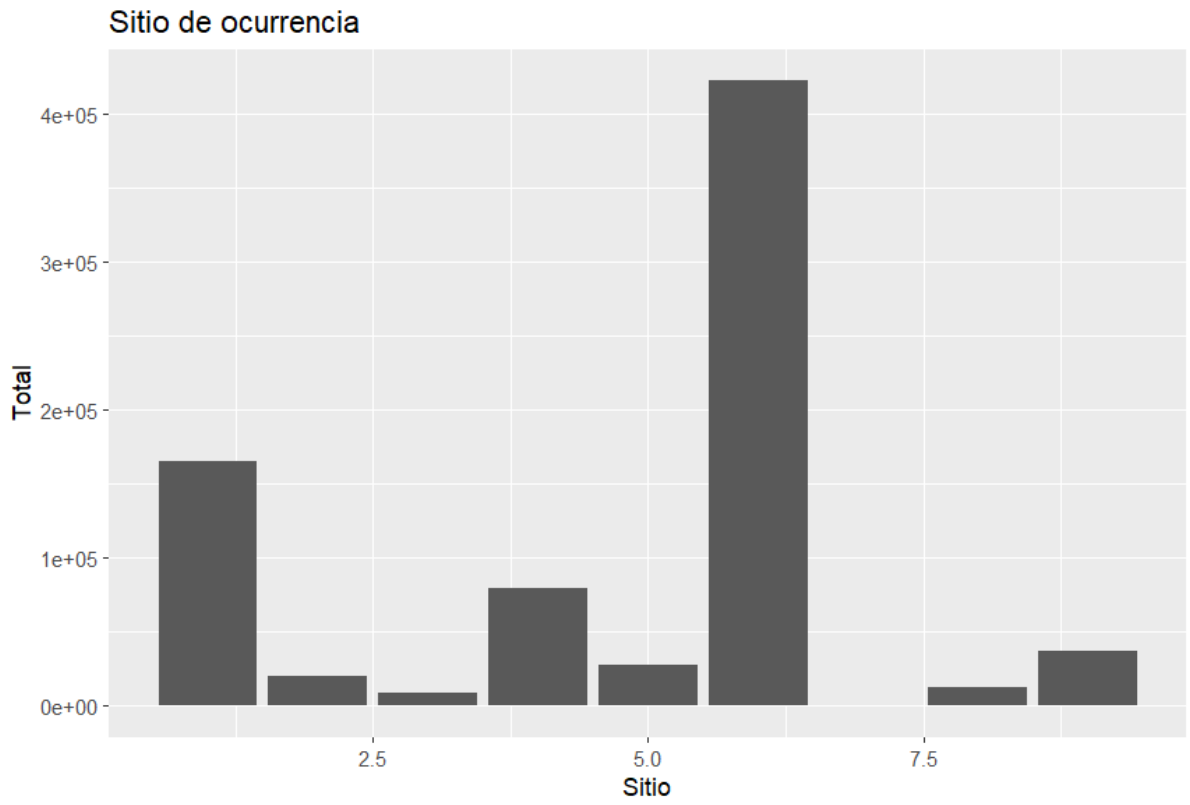
60–74 años

75–89 años

Esto confirma el patrón esperado en poblaciones donde la mortalidad se concentra en edades adultas mayores.

El grupo de 0–14 años también presenta un número considerable de defunciones, lo cual podría requerir análisis específico para identificar causas asociadas.

Sitio de orurrencia



Código	1	2	3	4	5	6	7	8	9
Significado	Hospital público	Hospital privado	Otros serivicios de salud pública	IGSS	Vía Pública	Domicilio	Lugar de trabajo	Otro	Ignorado

Más del 50% de las defunciones fueron en el sitio clasificado como categoría 6, seguido por las categorías 1 y 4. Esto puede reflejar:

- Concentración de defunciones en centros de atención médica
- Diferencias en acceso a servicios de emergencia
- Variaciones en atención hospitalaria vs domiciliaria

El análisis del sitio de ocurrencia es importante para la planificación de servicios de salud.

d.

Para responder las preguntas se hizo un análisis de los datos en los años siguientes (2019-2022)

Pregunta 1

¿Las defunciones han aumentado con el tiempo (2009–2018)?

Supuesto inicial

Se espera que las defunciones aumenten debido al crecimiento poblacional y envejecimiento.

Respuesta:

Año	2019	2020	2021	2022
Muertes	85600	96001	118465	95386

El análisis de las defunciones entre 2019 y 2022 muestra un incremento importante entre 2019 y 2021, pasando de aproximadamente 85,600 defunciones en 2019 a 118,465 en 2021. Posteriormente, en 2022 se observa una disminución hasta 95,386 defunciones. Este comportamiento indica que las defunciones no aumentan de manera constante en el tiempo, sino que presentan variaciones importantes. El incremento observado entre 2020 y 2021 podría estar asociado a factores epidemiológicos extraordinarios, como pudo ser la pandemia de covid19, mientras que la disminución posterior sugiere una normalización parcial de la mortalidad. El supuesto no se confirma, pues el número de muertes no es lineal, al menos hasta cierto punto.

Pregunta 2

¿Existe mayor mortalidad en hombres que en mujeres?

Supuesto inicial

Se cree que los hombres presentan mayor mortalidad por factores conductuales y ocupacionales.

Respuesta:

Hay más defunciones masculinas, hay una diferencia de aproximadamente 12 puntos porcentuales, esto es bastante fuerte estadísticamente. Se observa una mayor proporción de defunciones en hombres (56.0%) en comparación con mujeres (44.0%). Este resultado coincide con patrones epidemiológicos observados internacionalmente, donde los hombres presentan mayor mortalidad asociada a factores ocupacionales, conductuales y de exposición a riesgos. Por lo que el supuesto es confirmado.

Pregunta 3

¿La edad promedio de defunción aumenta con el tiempo?

Supuesto inicial

Se espera aumento por mejoras en salud pública y esperanza de vida.

Respuesta:

Si la línea sube, la esperanza de vida mejora; si baja, puede haber eventos externos; si fluctua, hay estabilidad. La edad promedio de defunción muestra una variación temporal entre 2019 y 2021, bajando en 2022. La tendencia observada sugiere cambios en la estructura de mortalidad que podrían estar asociados a factores demográficos, sanitarios o epidemiológicos, como la pandemia Covid19. El supuesto se confirma parcialmente.

Pregunta 4

¿La mayoría de las defunciones ocurre en edades avanzadas?

Supuesto inicial

Se espera concentración en adultos mayores.

Respuesta:

Grupo	0-14	15-29	30-44	45-59	60-74	75-89	90+
%	8.7	8.4	10.4	15.6	24.5	25.3	7.1

Adultos mayores (60+) = 24.5 + 25.3 + 7.1 ≈ 56.9%, que es más de la mitad. La mayor concentración de defunciones se observa en edades mayores de 60 años, representando aproximadamente el 57% del total de fallecimientos. Esto confirma el patrón esperado de mortalidad asociado al envejecimiento poblacional. El supuesto se confirma totalmente.

Pregunta 5

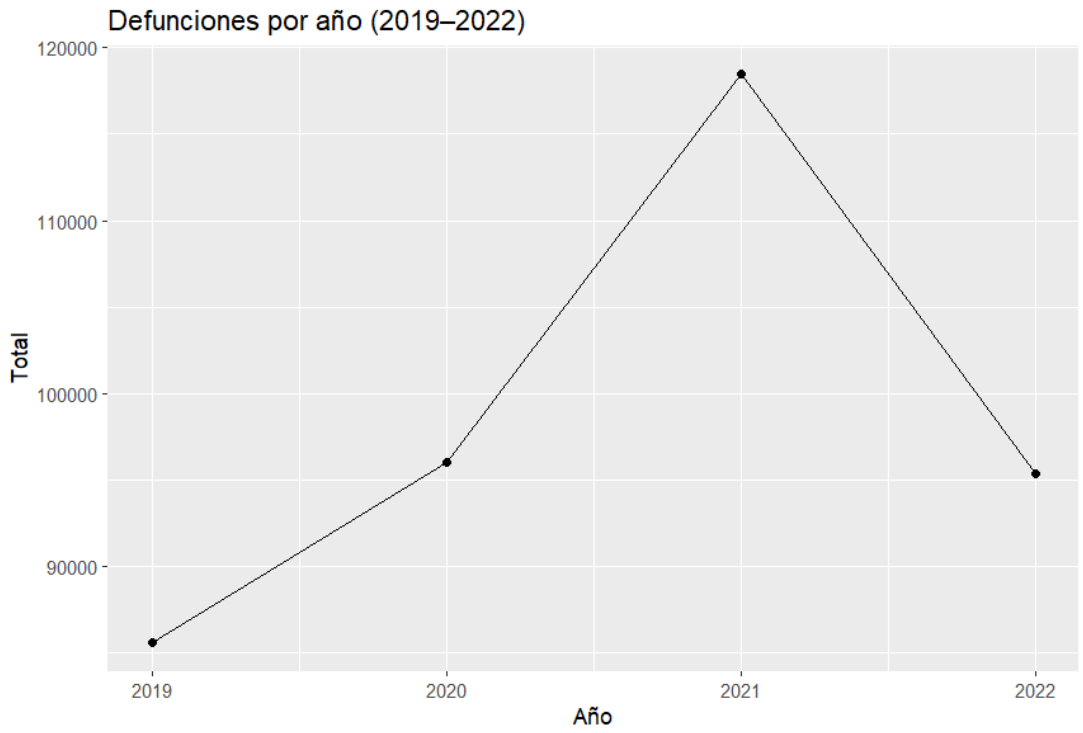
¿Existen diferencias entre área urbana y rural?

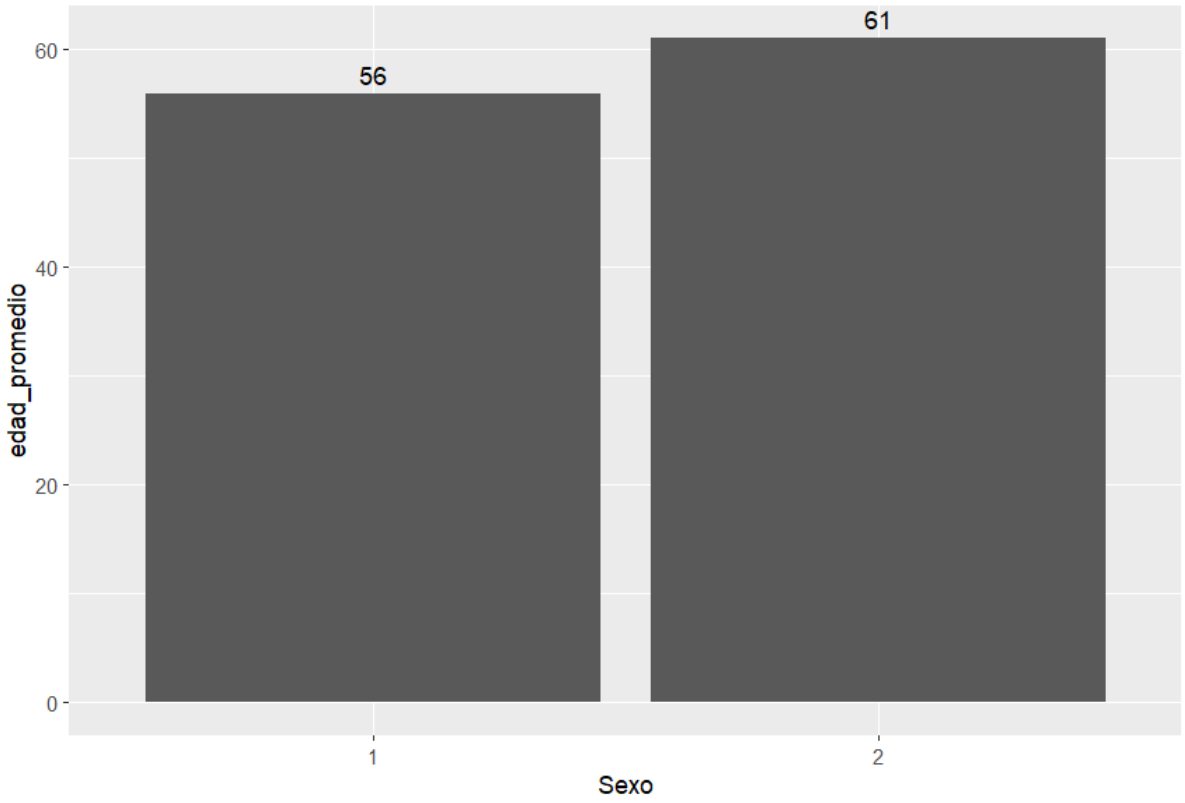
Supuesto inicial

Se cree que las defunciones se concentran en zonas urbanas.

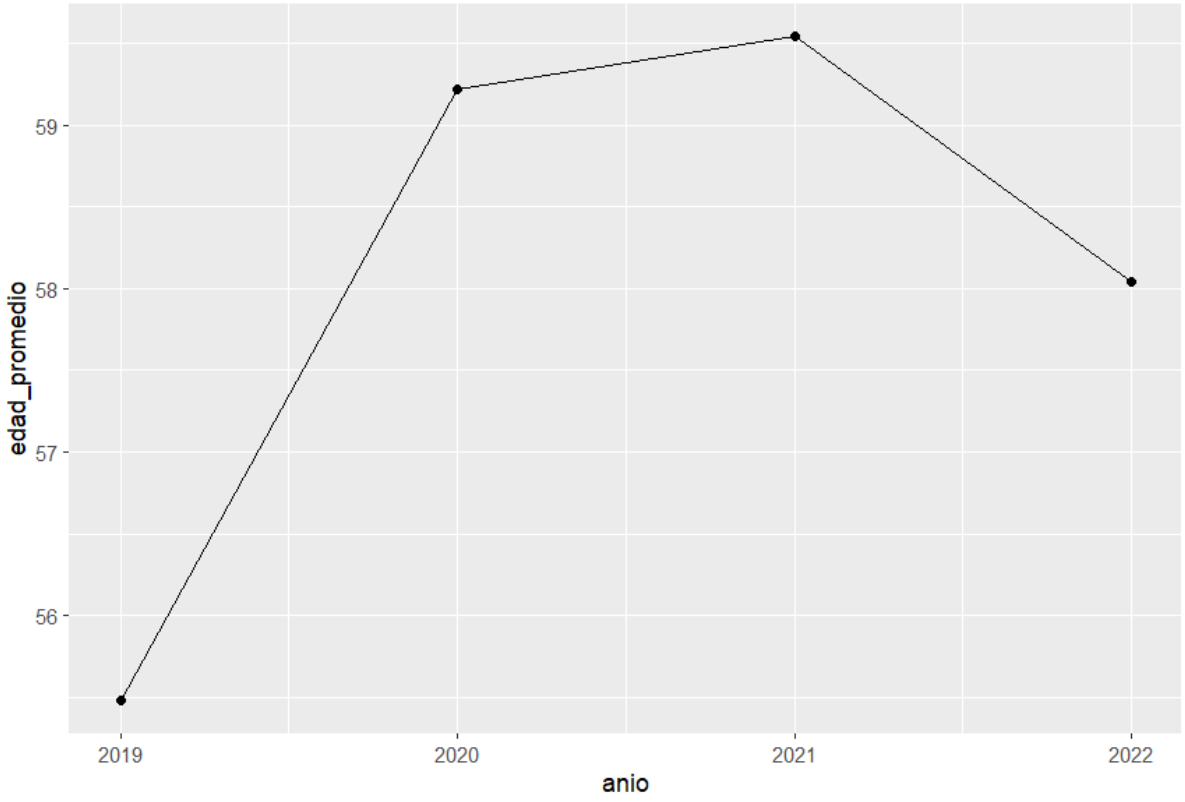
Área	%
Altamente urbano	28.99
Urbano mixto	14.76
Mayormente rural	56.25

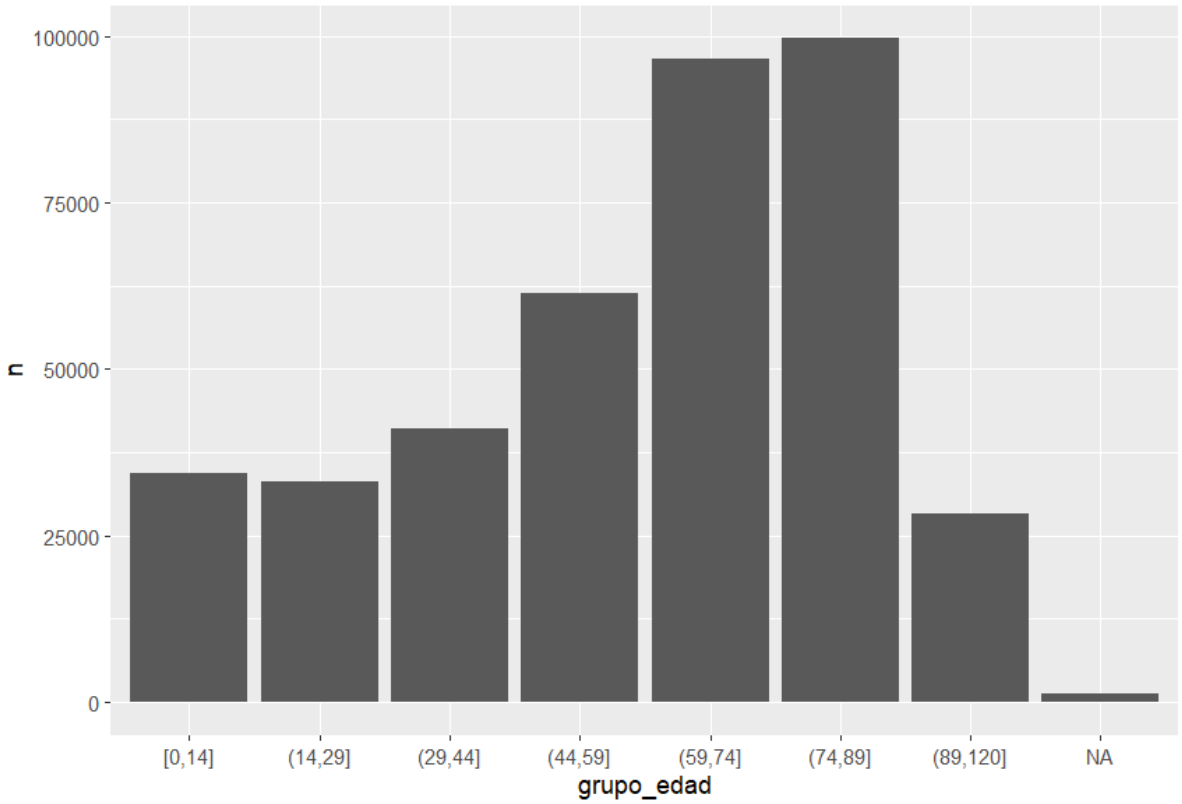
e.

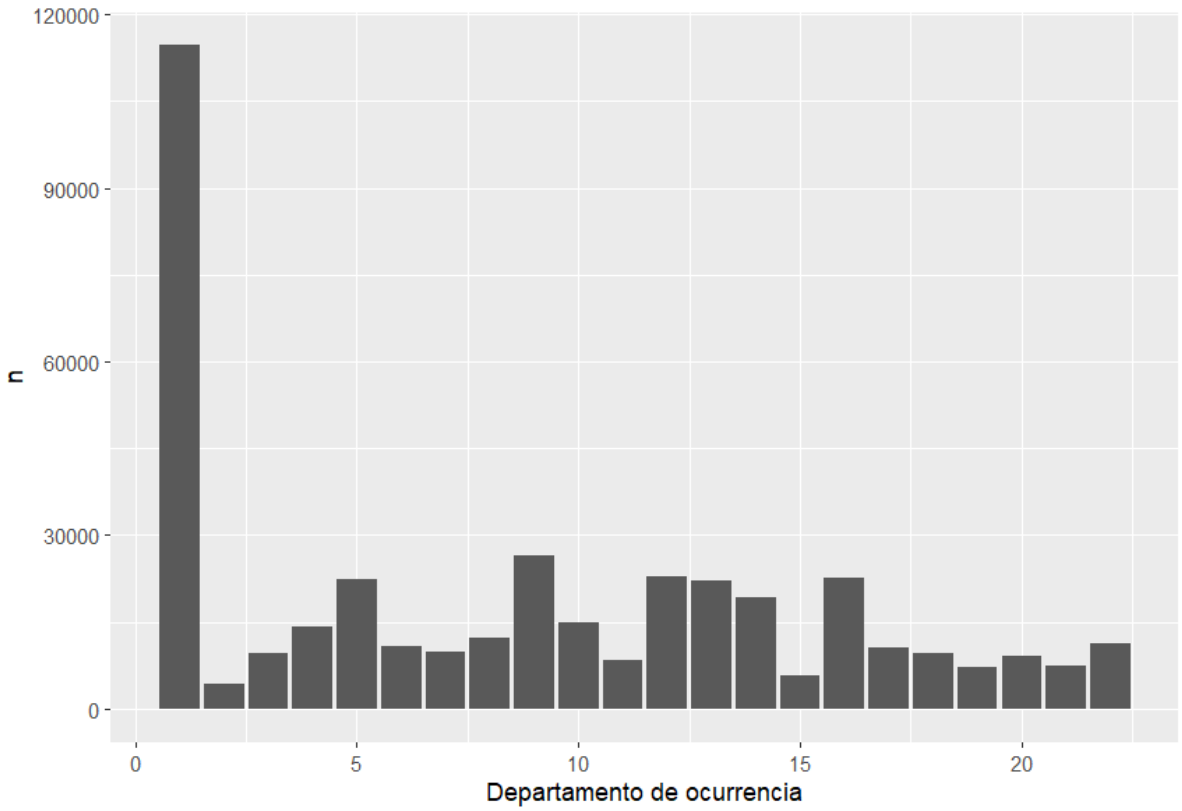

Defunciones por año (2019–2022)
120000
110000
100000
90000
Total
2019
2020
2021
2022
Año



El numero 1 representando a los hombres y el 2 a las mujeres

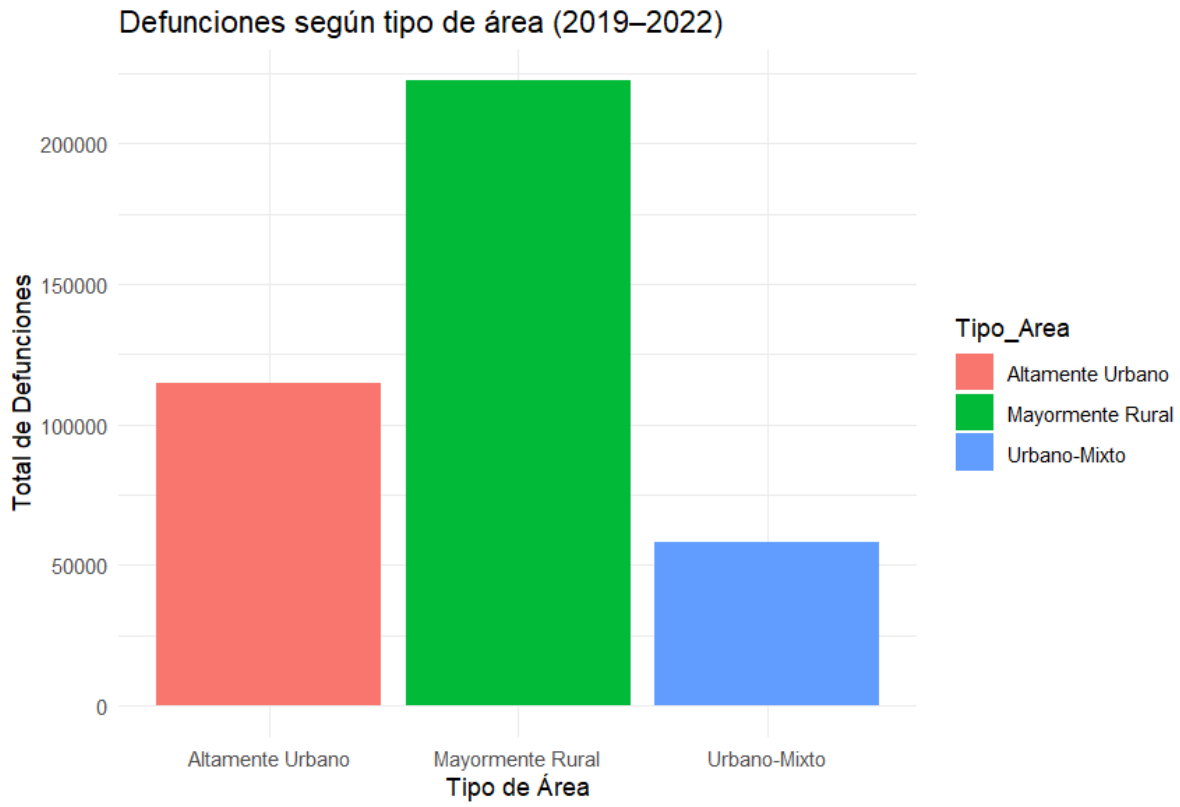




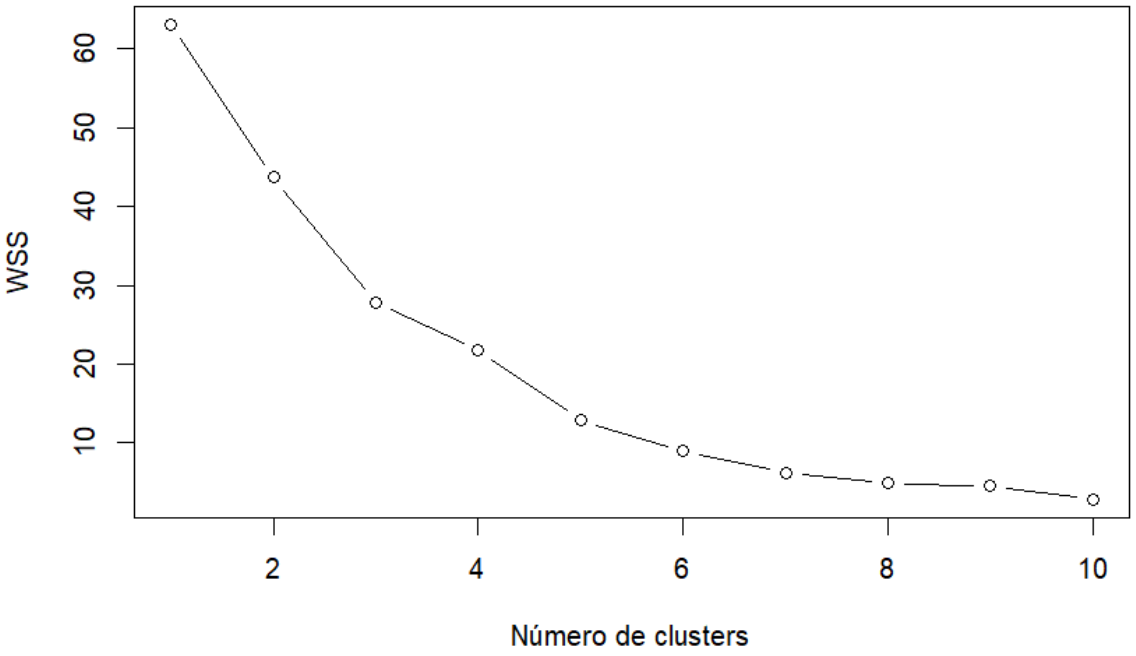


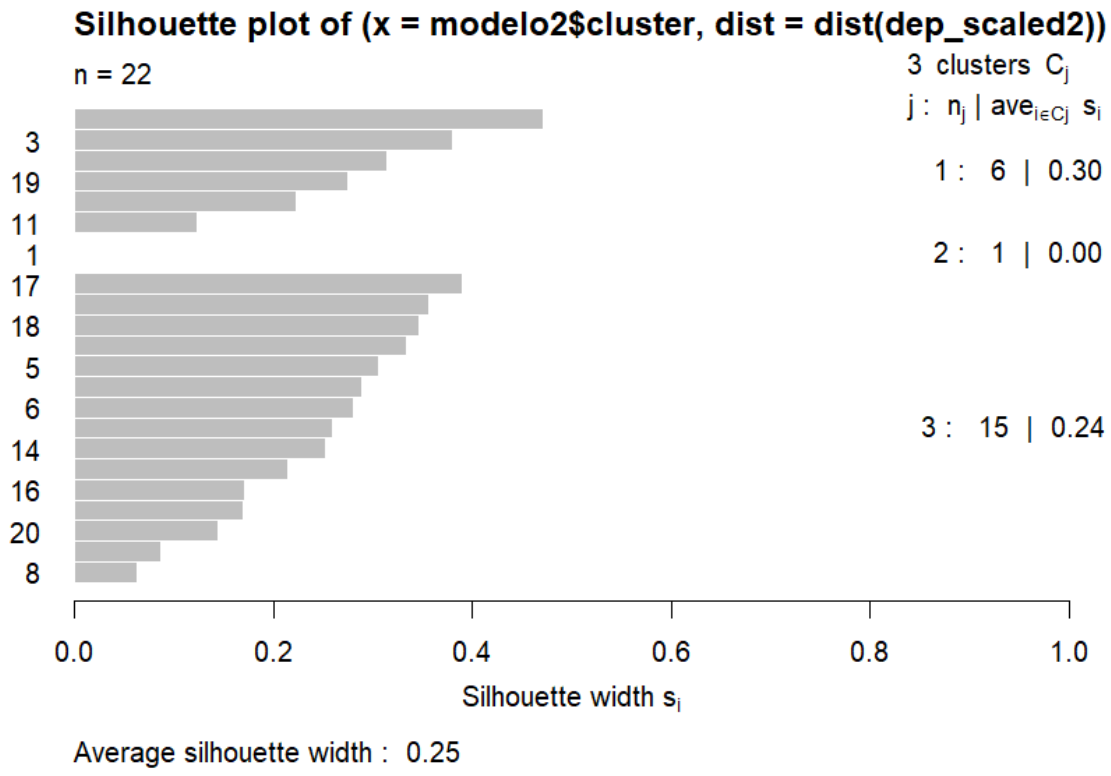
Los datos a partir del 2019 utilizan el número del departamento, dejan de usar el área geográfica como urbano y rural. Utilizando ahora el departamento de ocurrencia, por lo que

los datos tomados a partir de ahora serán modificados según donde predomina más el área urbana de la rural, urbano-mixto y donde predomina más el área rural del urbano,



f.





Método del codo (WSS vs número de clusters)

La primera gráfica muestra la suma de cuadrados dentro de los clusters (WSS) para distintos números de clusters. Observaciones clave:

- La WSS disminuye al aumentar los clusters, como es esperado.
- El "codo" se nota alrededor de 3 clusters, porque después de ese punto la reducción de WSS es mucho menor.
- Esto indica que 3 clusters es un buen compromiso entre complejidad y ajuste.

El análisis del codo sugiere que los datos se pueden agrupar en 3 clusters principales, ya que agregar más clusters no reduce significativamente la variación dentro de los grupos.

Análisis de silueta

La segunda gráfica es un diagrama de silueta:

- Cada barra representa un departamento y qué tan bien se ajusta a su cluster (ancho de la barra = silueta).
- Los valores varían de -1 a 1: valores cercanos a 1 indican que el departamento está bien asignado, valores cercanos a 0 indican que está en el límite, y valores negativos indican posible asignación incorrecta.
- Tu promedio de silueta es 0.25, lo que es moderado: los clusters existen, pero no están muy bien separados.
- Observamos que algunos departamentos (ej. cluster 2 con un solo departamento) tienen silueta de 0, lo que indica que este cluster está aislado pero pequeño.

La estructura de clusters es débilmente definida, con un cluster pequeño muy específico y otros dos más grandes con solapamiento moderado. Esto sugiere que los clusters capturan tendencias generales, pero no diferencias muy marcadas entre todos los departamentos.

Resumen de los clusters

La tabla aggregate muestra las características promedio de cada cluster:

Cluster	#Dep	total_def promedio	edad_prom	prop_hombres
1	12	7757	62.07	0.55
2	1	114646	68.53	0.56
3	12	15617	57.57	0.56

Intepretación:

Cluster 2: Solo un departamento (Guatemala). Muy grande en población de defunciones, edad media menor que cluster 1, proporción de hombres similar. Es claramente un cluster outlier, que refleja el tamaño especial de la capital.

Cluster 1: Departamentos con menor cantidad de defunciones, pero edad promedio más alta. Podrían ser departamentos más pequeños o con población más envejecida.

Cluster 3: Departamentos de tamaño intermedio, con edad promedio ligeramente menor que cluster 1, proporción de hombres similar.

Los clusters capturan diferencias en tamaño del departamento y edad promedio, mientras que la proporción de hombres es bastante homogénea entre clusters. El cluster 2 destaca por su magnitud extrema y representa un caso especial.

Conclusión general:

1. El método del codo sugiere usar 3 clusters.
2. La silueta promedio baja indica que la separación no es muy fuerte, pero hay tendencias claras.
3. Los clusters se interpretan como:

- Cluster 1: Departamentos pequeños con población más envejecida.
- Cluster 2: Departamento único, muy grande (Guatemala).
- Cluster 3: Departamentos medianos, población algo más joven que cluster 1.

4. La proporción de hombres es similar en todos los clusters, por lo que no aporta mucho a la segmentación.

2.

a.

La mortalidad en Guatemala representa un desafío significativo para el sistema de salud pública y el desarrollo socioeconómico del país. Según los datos del Instituto Nacional de Estadística (INE), entre 2009 y 2022 se registraron más de 1 millón de defunciones, con una tendencia general al aumento, especialmente marcada entre 2019 y 2021, posiblemente influenciada por eventos epidemiológicos como la pandemia de COVID-19. Esta situación se agrava por disparidades demográficas: una mayor proporción de muertes en hombres (56.5%), concentración en edades adultas mayores (más del 50% en personas de 60 años o más), y diferencias geográficas, con mayor incidencia en áreas mayormente rurales (56.25%) comparado con urbanas. Estos patrones sugieren problemas subyacentes como acceso desigual a servicios de salud, exposición a riesgos ocupacionales y conductuales en hombres, envejecimiento poblacional, y vulnerabilidades en zonas rurales. Sin un análisis profundo, estas tendencias podrían perpetuar desigualdades y sobrecargar los recursos sanitarios, lo que justifica una investigación para identificar causas y proponer intervenciones.

b.

Problema científico: ¿Cuáles son los patrones demográficos, temporales y geográficos de la mortalidad en Guatemala entre 2009 y 2022, y cómo influyen factores como el sexo, la edad y el área geográfica en estas tendencias?

Objetivos:

- Objetivo general: Analizar los patrones de mortalidad en Guatemala a partir de datos vitales del INE para identificar tendencias y disparidades que permitan proponer recomendaciones para políticas de salud pública.
- Objetivos específicos:
    - Identificar y cuantificar las diferencias en tasas de mortalidad por sexo, edad y área geográfica, midiendo proporciones y promedios a lo largo del período estudiado.
    - Evaluar el impacto temporal de eventos externos (como pandemias) en el número y características de defunciones, mediante análisis de series temporales y clustering por departamentos.

Estos objetivos son medibles mediante técnicas estadísticas descriptivas y exploratorias aplicadas a los datos disponibles, y alcanzables dentro del alcance de esta investigación inicial.

c.

Los datos provienen de las bases de estadísticas vitales del Instituto Nacional de Estadística (INE) de Guatemala, específicamente los conjuntos de defunciones anuales desde 2009 hasta 2022, disponibles en formato .sav (SPSS). Se utilizaron 14 archivos individuales (uno por año), que en total contienen aproximadamente 1,100,000 observaciones y 36 variables por archivo. Las variables clave incluyen demográficas (Edadif: edad del difunto, Sexo: 1=hombre, 2=mujer), geográficas (Areag: área urbana/rural hasta 2018, Depocu: departamento de ocurrencia desde 2019), temporales (anio: año de defunción, agregado manualmente), y del evento (Ocur: sitio de ocurrencia, Getdif: grupo étnico, etc.). Las variables son mayoritariamente categóricas (factores), con algunas cuantitativas (Edadif, Diaocu).

El estado inicial de los datos era crudo, con valores missing codificados como 999 (en Edadif) e inconsistencias en codificación geográfica entre períodos (cambio de Areag a Depocu). Se realizaron las siguientes operaciones de limpieza y transformación:

- Lectura y unión de archivos usando librerías haven y dplyr en R: se crearon datasets combinados (datos para 2009-2018, datos2 para 2019-2022).
- Conversión de tipos: Edadif a numérico, Sexo/Areag/Getdif/Ecidif a factores.
- Manejo de missing: Reemplazo de 999 en Edadif por NA, para evitar sesgos en resúmenes.

- Creación de variables derivadas: anio (agregado por año), grupo_edad (cortes por rangos de edad), Tipo_Area (clasificación urbana/rural/mixta basada en Depocu para 2019-2022).
- Escalado de datos para clustering (usando scale() en variables numéricas como total_def, edad_prom, prop_hombres).

No se detectaron duplicados significativos ni outliers extremos que requirieran eliminación, pero se documentaron en gráficos (ej. boxplot de Edadif).

Estos datos permiten responder el problema científico al proporcionar una cobertura temporal amplia y variables para analizar patrones demográficos y geográficos.

d.

El análisis exploratorio revela patrones claros en la mortalidad guatemalteca: un aumento general de defunciones (de ~71,000 en 2009 a ~118,000 en 2021, con pico posiblemente por COVID-19), mayor incidencia en hombres (56.5%) y edades avanzadas (57% en >60 años), y concentración en áreas rurales (56.25%). Las edades no siguen distribución normal (asimetría hacia mayores), y el clustering por departamentos identifica tres grupos:

- Cluster 1: "Departamentos pequeños envejecidos" (12 departamentos, bajo total_def, alta edad_prom ~62 años).
- Cluster 2: "Departamento metropolitano outlier" (1 departamento: Guatemala, alto total_def ~114,000, edad_prom ~68 años).
- Cluster 3: "Departamentos medianos jóvenes" (12 departamentos, total_def intermedio, edad_prom ~57 años).

Estos hallazgos confirman parcialmente supuestos iniciales (ej. mayor mortalidad en hombres y adultos mayores), pero refutan otros (ej. aumento lineal de defunciones, concentración exclusiva en urbanas). Las disparidades sugieren necesidades en políticas de salud focalizadas en hombres, rurales y envejecimiento.

Enlace del informe del proyecto

Project1.docx

Enlace al repositorio de github

https://github.com/Paul-1511/project1/tree/main