Análisis Exploratorio de Defunciones en Guatemala 2011-2021

Proyecto 1 - Minería de Datos CC3074

Grupo 7

2026-02-15

Contents

Introducción

Contexto

El presente documento constituye un **análisis exploratorio** de los registros de **defunciones en Guatemala** para el período **2011–2021**.

Situación problemática

Las defunciones son un indicador clave para salud pública y planificación. Sin embargo, su comportamiento **no es uniforme**: puede variar por **departamento**, **área (urbana/rural)**, **sexo**, **acceso a asistencia médica**, y por **patrones temporales** (meses/años). Sin explorar los datos, es difícil priorizar acciones y entender dónde se concentran los mayores riesgos.

Problema científico (enunciado)

¿Qué patrones demográficos, geográficos y temporales se observan en las defunciones registradas en Guatemala (2011–2021) y cómo se relacionan entre sí variables como sexo, edad, área geográfica, asistencia médica y lugar de ocurrencia?

Objetivos preliminares

Objetivo general:
- Describir y analizar patrones y relaciones en las defunciones registradas (2011–2021) mediante técnicas de estadística descriptiva, visualización y agrupamiento.

Objetivos específicos:
- Identificar tendencias temporales (por año y mes) y diferencias por sexo/edad.
- Explorar diferencias geográficas (departamento/área) y su relación con asistencia médica y lugar de ocurrencia.
- Probar supuestos mediante al menos 5 preguntas de investigación apoyadas con tablas, métricas y gráficos.

Carga y Consolidación de Datos

```
archivos <- c(
  "1Sg2Lmx2sWne8RcD9hM6guU73I9eQW8B.csv",
  "QnZLxknSHwbFfpJ8I1frbkoIKfz1BBjd.csv",
  "iY2sN6q3d4ihJpgzr7KgJpQAiEn0bo60.csv",
  "DX2BmYU5m4JfPRhrFHwDRDEs49V7fN5I.csv",
  "bb6vENc1cmPlBToSEEr6HESWdZk6tHFs.csv",
  "20171204152107xRp35JuZin7nN2x88Me8MVcQvyZCnu5K.csv",
  "20181226142907xRp35JuZin7nN2x88Me8MVcQvyZCnu5K.csv",
  "2019112915200690dm3oxU9mTY58hkborwzylm7MJop05q.csv",
  "20201201154851el8puh8r6zutgVKBoRIbazWluzIr25A3.csv",
  "20210930225530FopQpWf6BcBWj8taVS3Q3mRKxgDsvwPe.csv"
)

datos_raw <- archivos %>%
  map_dfr(~ read_csv(.x, col_types = cols(.default = "c"))) %>%
  clean_names()

cat("Datos cargados:", nrow(datos_raw), "observaciones ×", ncol(datos_raw), "variables\n")
```

```
## Datos cargados: 809296 observaciones × 35 variables
```

Descripción General del Dataset

Dimensiones del Dataset

```
data.frame(
  Característica = c("Observaciones", "Variables", "Período"),
  Valor = c(format(nrow(datos_raw), big.mark = ","),
            ncol(datos_raw),
            "2011-2021")
```

```
) %>%
  kable(caption = "Dimensiones del Dataset") %>%
  kable_styling(bootstrap_options = c("striped", "hover"), full_width = FALSE)
```

Table 1: Dimensiones del Dataset

Característica	Valor
Observaciones	809,296
Variables	35
Período	2011-2021

Tipos de Variables

```
vars_numericas_esperadas <- c("anoreg", "mesreg", "diaocu", "edadif")
vars_categoricas <- setdiff(names(datos_raw), vars_numericas_esperadas)

data.frame(
  Tipo = c("Numéricas", "Categóricas", "Total"),
  Cantidad = c(length(vars_numericas_esperadas), length(vars_categoricas), ncol(datos_raw))
) %>%
  kable(caption = "Tipos de Variables") %>%
  kable_styling(bootstrap_options = c("striped", "hover"), full_width = FALSE)
```

Table 2: Tipos de Variables

Tipo	Cantidad
Numéricas	4
Categóricas	31
Total	35

Diccionario de Variables Principales

```
tribble(
  ~Variable, ~Descripción,
  "anoreg", "Año de registro",
  "mesreg", "Mes de registro",
  "depreg", "Departamento de registro",
  "depocu", "Departamento de ocurrencia",
  "sexo", "Sexo del fallecido",
  "edadif", "Edad del fallecido",
  "perdif", "Período de edad",
  "caudef", "Código CIE-10",
  "asist", "Asistencia médica",
  "ocur", "Lugar de ocurrencia"
) %>%
  kable(caption = "Diccionario de Variables Principales") %>%
  kable_styling(bootstrap_options = c("striped", "hover"))
```

Table 3: Diccionario de Variables Principales

Variable	Descripción
anoreg	Año de registro
mesreg	Mes de registro
depreg	Departamento de registro
depocu	Departamento de ocurrencia
sexo	Sexo del fallecido
edadif	Edad del fallecido
perdif	Período de edad
caudef	Código CIE-10
asist	Asistencia médica
ocur	Lugar de ocurrencia

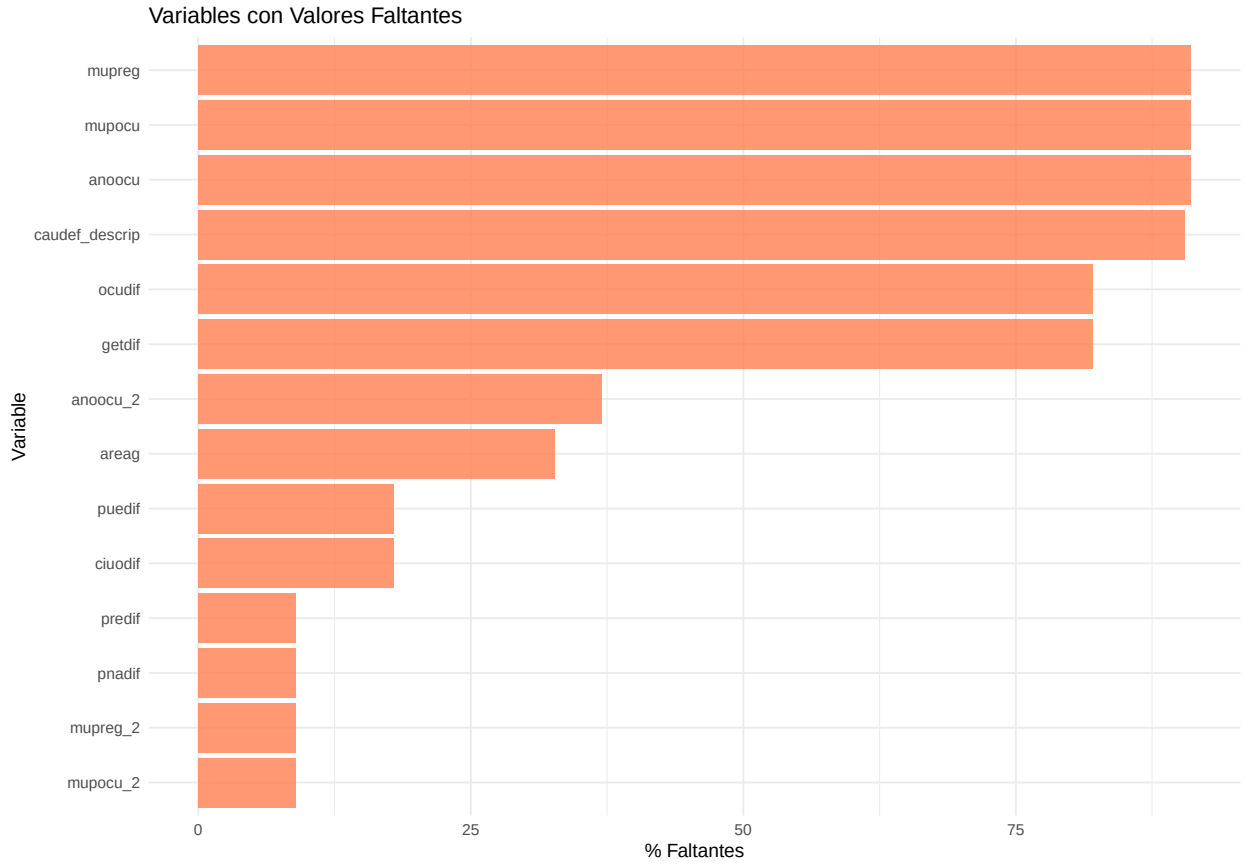
Valores Faltantes

```
na_summary <- datos_raw %>%
  summarise(across(everything(), ~sum(is.na(.) | . == "" | . == " ") / n() * 100)) %>%
  pivot_longer(everything(), names_to = "Variable", values_to = "Porcentaje_NA") %>%
  arrange(desc(Porcentaje_NA)) %>%
  filter(Porcentaje_NA > 0)

na_summary %>%
  head(15) %>%
  mutate(Porcentaje_NA = paste0(round(Porcentaje_NA, 2), "%")) %>%
  kable(caption = "Top 15 Variables con Valores Faltantes",
        col.names = c("Variable", "% Faltantes")) %>%
  kable_styling(bootstrap_options = c("striped", "hover"))
```

Table 4: Top 15 Variables con Valores Faltantes

Variable	% Faltantes
mupreg	91.06%
mupocu	91.06%
anoocu	91.06%
caudef_descrip	90.53%
getdif	82.08%
ocudif	82.08%
anoocu_2	37%
areag	32.7%
puedif	17.92%
ciuodif	17.92%
mupreg_2	8.94%
mupocu_2	8.94%
pnadif	8.94%
predif	8.94%

```r
na_summary %>%
  head(20) %>%
  ggplot(aes(x = reorder(Variable, Porcentaje_NA), y = Porcentaje_NA)) +
  geom_col(fill = "coral", alpha = 0.8) +
  coord_flip() +
  labs(title = "Variables con Valores Faltantes", x = "Variable", y = "% Faltantes") +
  theme_minimal()
```



Limpieza y Transformación

```r
datos <- datos_raw %>%
  mutate(across(where(is.character), ~ str_to_upper(stri_trans_general(.x, "Latin-ASCII")))) %>%
  mutate(across(c(anoreg, mesreg, diaocu, edadif), as.numeric)) %>%
  mutate(edad_anios = case_when(
    str_detect(perdif, "ANO") ~ edadif,
    str_detect(perdif, "MES") ~ edadif / 12,
    str_detect(perdif, "DIA") ~ edadif / 365,
    str_detect(perdif, "HORA") ~ edadif / 8760,
    TRUE ~ NA_real_
  )) %>%
```

```r
  mutate(causa_capitulo = str_sub(caudef, 1, 1)) %>%
  filter(edad_anios <= 115 | is.na(edad_anios))

cat("Registros después de limpieza:", format(nrow(datos), big.mark = ","), "\n")
```

```
## Registros después de limpieza: 809,292
```

Exploración de Variables Numéricas

Resumen Estadístico

```r
datos %>%
  select(anoreg, mesreg, diaocu, edadif, edad_anios) %>%
  summarise(across(where(is.numeric),
                   list(
                     N = ~sum(!is.na(.)),
                     Media = ~mean(., na.rm = TRUE),
                     Mediana = ~median(., na.rm = TRUE),
                     Desv_Std = ~sd(., na.rm = TRUE),
                     Min = ~min(., na.rm = TRUE),
                     Q1 = ~quantile(., 0.25, na.rm = TRUE),
                     Q3 = ~quantile(., 0.75, na.rm = TRUE),
                     Max = ~max(., na.rm = TRUE)
                   ))) %>%
  pivot_longer(everything(),
               names_to = c("Variable", "Estadistica"),
               names_sep = "_(?=[^_]+$)",
               values_to = "Valor") %>%
  mutate(Valor = round(Valor, 2)) %>%
  pivot_wider(names_from = Estadistica, values_from = Valor) %>%
  kable(caption = "Resumen estadístico de variables numéricas") %>%
  kable_styling(bootstrap_options = c("striped", "hover"), full_width = FALSE) %>%
  scroll_box(width = "100%")
```

Table 5: Resumen estadístico de variables numéricas

Variable	N	Media	Mediana	Std	Min	Q1	Q3	Max
anoreg	809292	2015.74	2016	NA	2011	2013	2018	2021
anoreg_Desv	NA	NA	NA	2.89	NA	NA	NA	NA
mesreg	0	NaN	NA	NA	Inf	NA	NA	-Inf
mesreg_Desv	NA	NA	NA	NA	NA	NA	NA	NA
diaocu	809292	15.67	16	NA	1	8	23	31
diaocu_Desv	NA	NA	NA	8.82	NA	NA	NA	NA
edadif	802793	54.14	61	NA	0	32	78	115
edadif_Desv	NA	NA	NA	28.11	NA	NA	NA	NA
edad_anios	802793	53.63	61	NA	0	32	78	115

edad_anios_Desv	NA	NA	NA	28.96	NA	NA	NA	NA

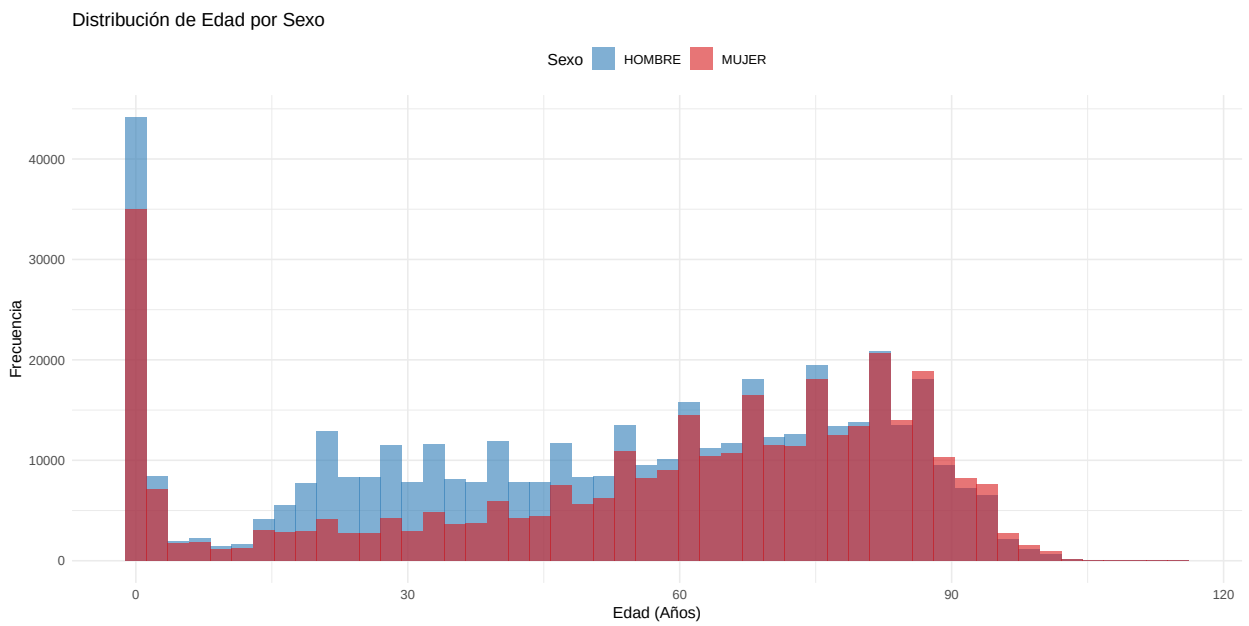
Análisis de Edad

```
datos %>%
  filter(!is.na(edad_anios)) %>%
  summarise(
    N = n(),
    Media = mean(edad_anios),
    Mediana = median(edad_anios),
    Desv_Std = sd(edad_anios),
    Q1 = quantile(edad_anios, 0.25),
    Q3 = quantile(edad_anios, 0.75),
    Mínimo = min(edad_anios),
    Máximo = max(edad_anios)
  ) %>%
  pivot_longer(everything(), names_to = "Estadística", values_to = "Valor") %>%
  mutate(Valor = round(Valor, 2)) %>%
  kable(caption = "Estadísticas de Edad") %>%
  kable_styling(full_width = FALSE)
```

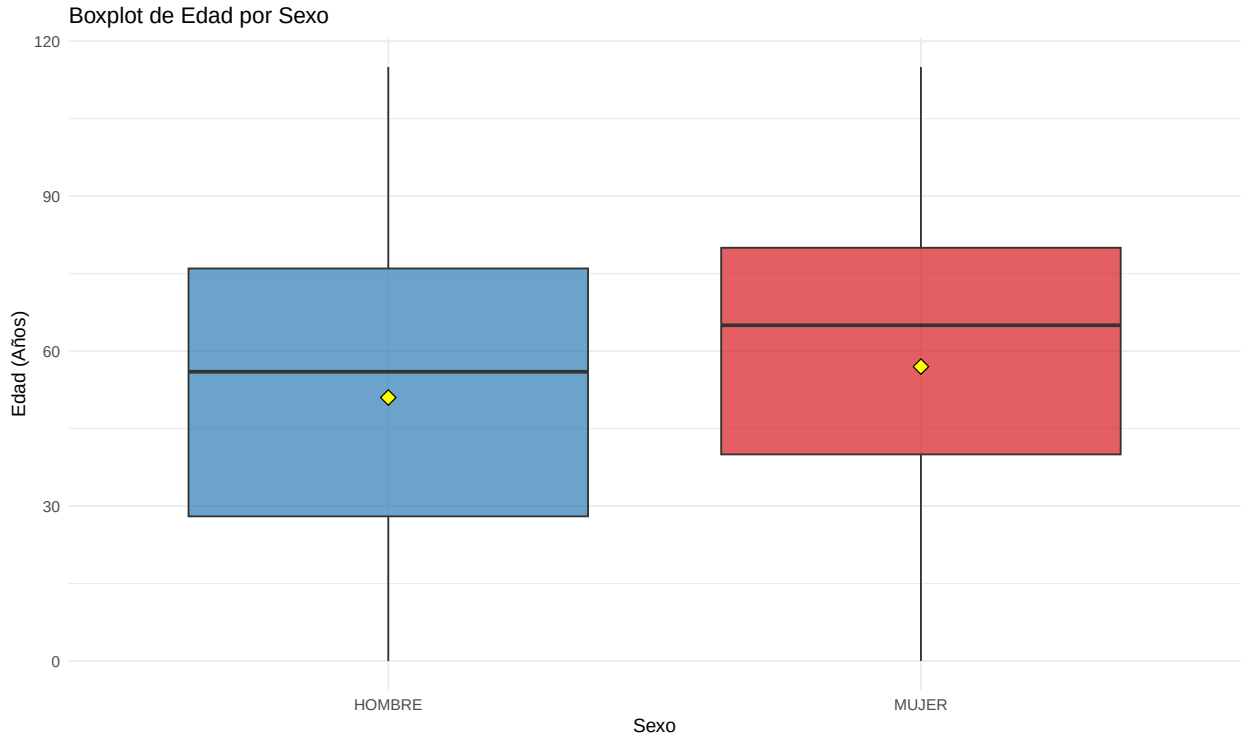
Table 6: Estadísticas de Edad

Estadística	Valor
N	802793.00
Media	53.63
Mediana	61.00
Desv_Std	28.96
Q1	32.00
Q3	78.00
Mínimo	0.00
Máximo	115.00

```
datos %>%
  filter(!is.na(edad_anios), !is.na(sexo)) %>%
  ggplot(aes(x = edad_anios, fill = sexo)) +
  geom_histogram(bins = 50, alpha = 0.6, position = "identity") +
  scale_fill_manual(values = c("HOMBRE" = "#2c7bb6", "MUJER" = "#d7191c"), name = "Sexo") +
  labs(title = "Distribución de Edad por Sexo", x = "Edad (Años)", y = "Frecuencia") +
  theme_minimal() +
  theme(legend.position = "top")
```

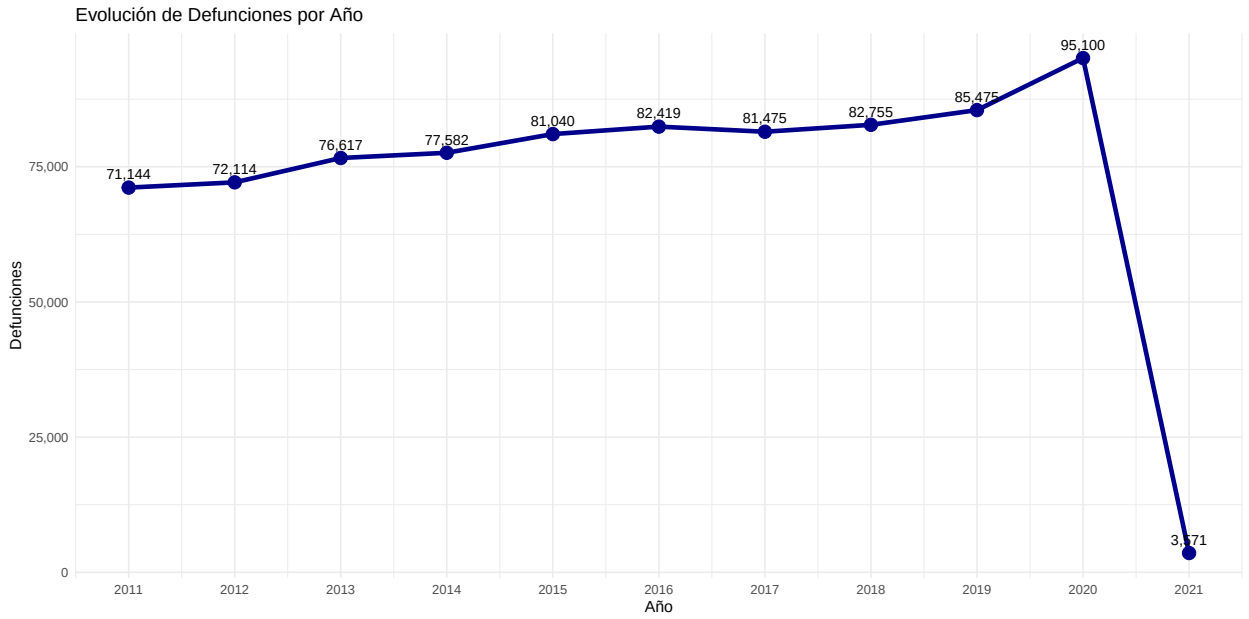



```
datos %>%
  filter(!is.na(edad_anios), !is.na(sexo)) %>%
  ggplot(aes(x = sexo, y = edad_anios, fill = sexo)) +
  geom_boxplot(alpha = 0.7, outlier.alpha = 0.3) +
  scale_fill_manual(values = c("HOMBRE" = "#2c7bb6", "MUJER" = "#d7191c")) +
  stat_summary(fun = mean, geom = "point", shape = 23, size = 3, fill = "yellow") +
  labs(title = "Boxplot de Edad por Sexo", x = "Sexo", y = "Edad (Años)") +
  theme_minimal() +
  theme(legend.position = "none")
```

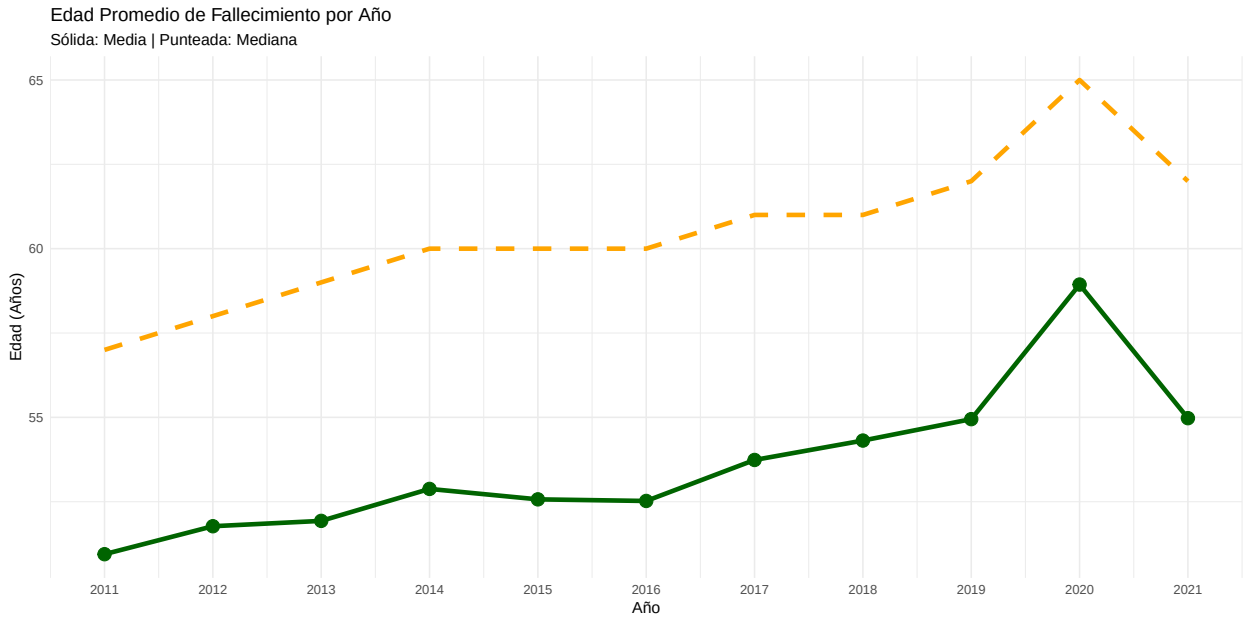


Evolución Temporal

```
datos %>%
  count(anoreg) %>%
  ggplot(aes(x = anoreg, y = n)) +
  geom_line(color = "darkblue", size = 1.5) +
  geom_point(color = "darkblue", size = 4) +
  geom_text(aes(label = comma(n)), vjust = -0.8, size = 3.5) +
  scale_x_continuous(breaks = seq(min(datos$anoreg, na.rm = TRUE),
                                  max(datos$anoreg, na.rm = TRUE), 1)) +
  scale_y_continuous(labels = comma) +
  labs(title = "Evolución de Defunciones por Año", x = "Año", y = "Defunciones") +
  theme_minimal()
```




```
datos %>%
  filter(!is.na(edad_anios), !is.na(anoreg)) %>%
  group_by(anoreg) %>%
  summarise(edad_promedio = mean(edad_anios), edad_mediana = median(edad_anios)) %>%
  ggplot(aes(x = anoreg, y = edad_promedio)) +
  geom_line(color = "darkgreen", size = 1.5) +
  geom_point(color = "darkgreen", size = 4) +
  geom_line(aes(y = edad_mediana), color = "orange", size = 1.5, linetype = "dashed") +
  scale_x_continuous(breaks = seq(min(datos$anoreg, na.rm = TRUE),
                                  max(datos$anoreg, na.rm = TRUE), 1)) +
  labs(title = "Edad Promedio de Fallecimiento por Año",
       subtitle = "Sólida: Media | Punteada: Mediana",
       x = "Año", y = "Edad (Años)") +
  theme_minimal()
```




```
datos %>%
  filter(!is.na(mesreg)) %>%
  count(mesreg) %>%
  mutate(mes_nombre = month.name[mesreg]) %>%
  ggplot(aes(x = reorder(mes_nombre, mesreg), y = n)) +
  geom_col(fill = "steelblue", alpha = 0.8) +
  scale_y_continuous(labels = comma) +
  labs(title = "Distribución por Mes de Registro", x = "Mes", y = "Frecuencia") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Exploración de Variables Categóricas

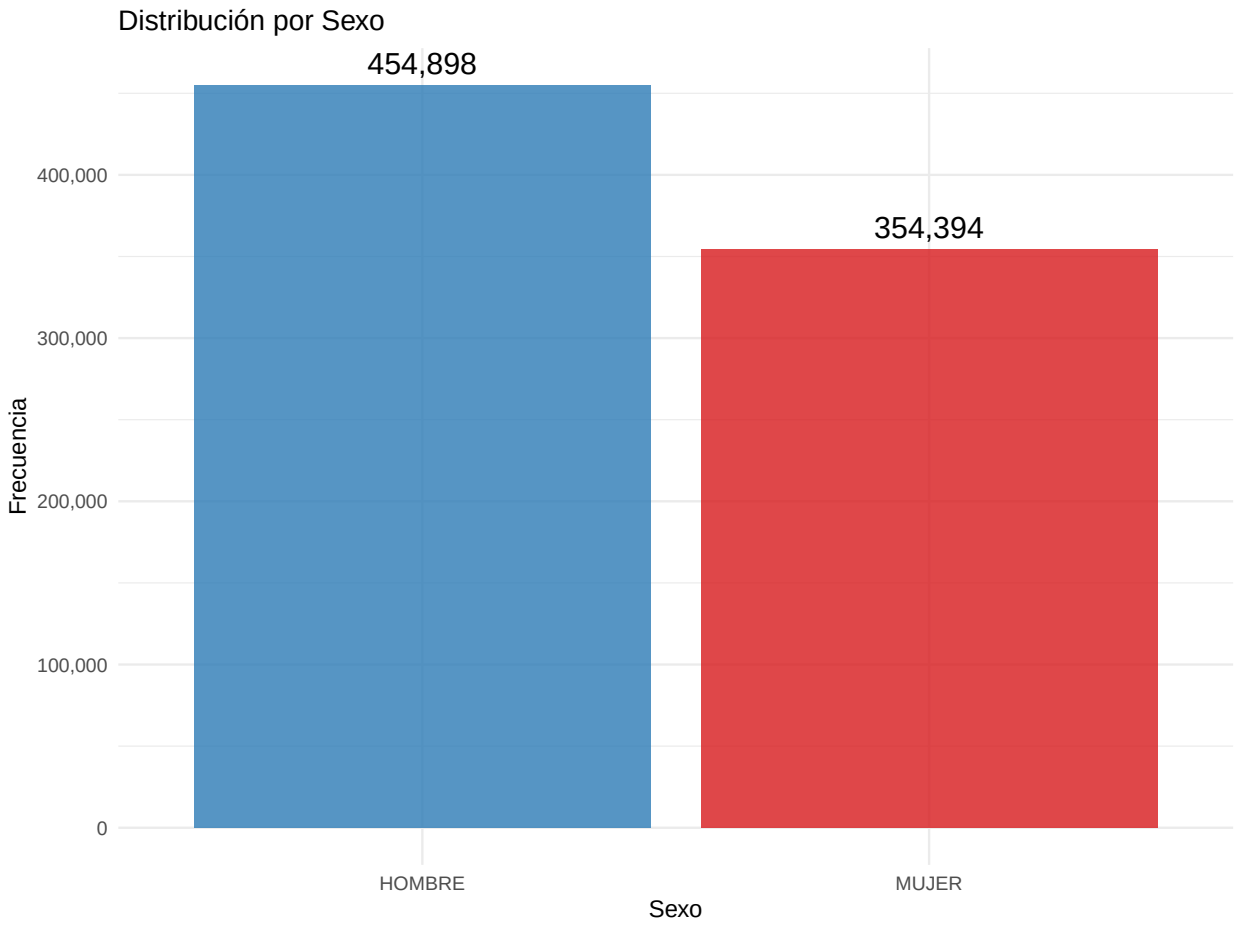
Sexo

```
tabla_sexo <- datos %>%
  count(sexo) %>%
  drop_na() %>%
  mutate(porcentaje = paste0(round(n / sum(n) * 100, 2), "%"))

tabla_sexo %>%
  kable(caption = "Distribución por Sexo",
        col.names = c("Sexo", "Frecuencia", "Porcentaje"),
        format.args = list(big.mark = ",")) %>%
  kable_styling(full_width = FALSE)
```

Table 7: Distribución por Sexo

Sexo	Frecuencia	Porcentaje
HOMBRE	454,898	56.21%
MUJER	354,394	43.79%

```
tabla_sexo %>%
  ggplot(aes(x = sexo, y = n, fill = sexo)) +
  geom_col(alpha = 0.8) +
  geom_text(aes(label = comma(n)), vjust = -0.5, size = 5) +
  scale_fill_manual(values = c("HOMBRE" = "#2c7bb6", "MUJER" = "#d7191c")) +
  scale_y_continuous(labels = comma) +
  labs(title = "Distribución por Sexo", x = "Sexo", y = "Frecuencia") +
  theme_minimal() +
  theme(legend.position = "none")
```



Área Geográfica

```
datos %>%
  count(areag) %>%
  drop_na() %>%
  mutate(porcentaje = paste0(round(n / sum(n) * 100, 2), "%")) %>%
  kable(caption = "Distribución por Área",
        col.names = c("Área", "Frecuencia", "Porcentaje"),
        format.args = list(big.mark = ",")) %>%
  kable_styling(full_width = FALSE)
```

Table 8: Distribución por Área

Área	Frecuencia	Porcentaje
IGNORADO	10,171	1.87%
RURAL	239,386	43.95%
URBANO	295,064	54.18%

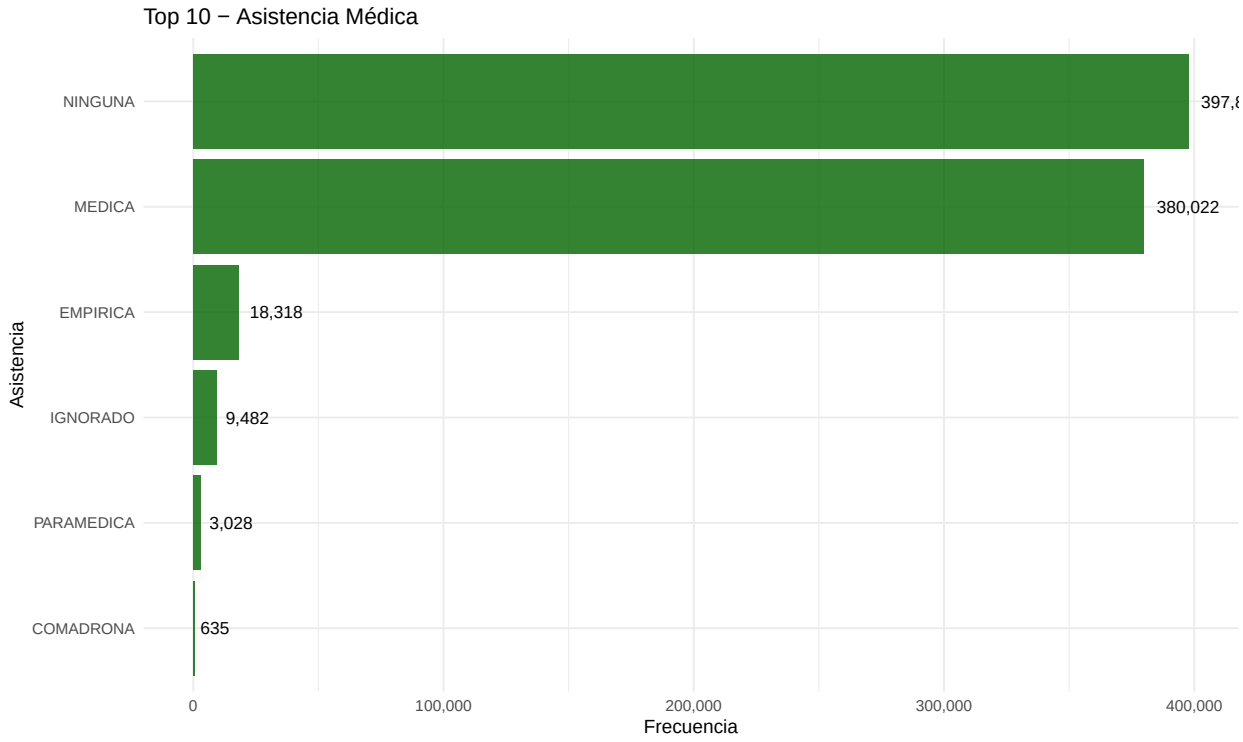
Asistencia Médica

```
datos %>%
  count(asist) %>%
  drop_na() %>%
  arrange(desc(n)) %>%
  head(10) %>%
  mutate(porcentaje = paste0(round(n / sum(n) * 100, 2), "%")) %>%
  kable(caption = "Top 10 - Asistencia Médica",
        col.names = c("Asistencia", "Frecuencia", "Porcentaje"),
        format.args = list(big.mark = ",")) %>%
  kable_styling()
```

Table 9: Top 10 - Asistencia Médica

Asistencia	Frecuencia	Porcentaje
NINGUNA	397,807	49.15%
MEDICA	380,022	46.96%
EMPIRICA	18,318	2.26%
IGNORADO	9,482	1.17%
PARAMEDICA	3,028	0.37%
COMADRONA	635	0.08%

```
datos %>%
  count(asist) %>%
  drop_na() %>%
  arrange(desc(n)) %>%
  head(10) %>%
  ggplot(aes(x = reorder(asist, n), y = n)) +
  geom_col(fill = "darkgreen", alpha = 0.8) +
  geom_text(aes(label = comma(n)), hjust = -0.2, size = 3.5) +
  coord_flip() +
  scale_y_continuous(labels = comma) +
  labs(title = "Top 10 - Asistencia Médica", x = "Asistencia", y = "Frecuencia") +
  theme_minimal()
```



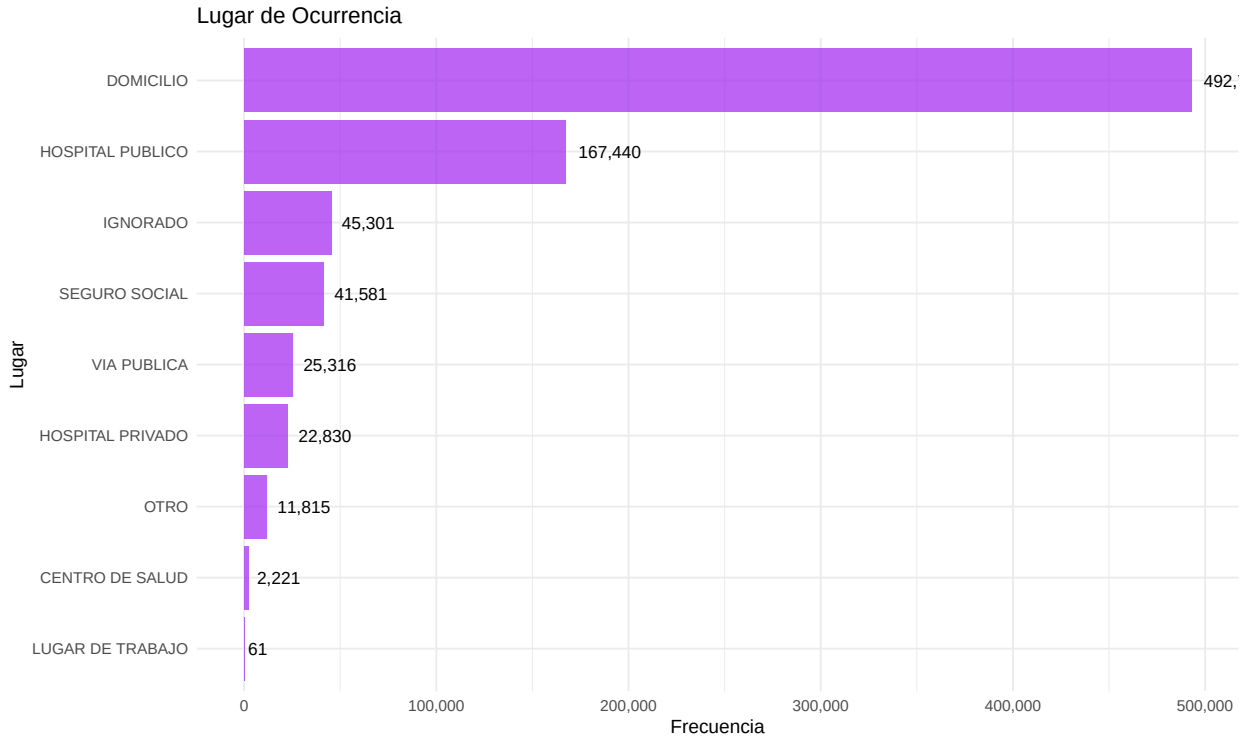
Lugar de Ocurrencia

```
datos %>%
  count(ocur) %>%
  drop_na() %>%
  mutate(porcentaje = paste0(round(n / sum(n) * 100, 2), "%")) %>%
  kable(caption = "Lugar de Ocurrencia",
        col.names = c("Lugar", "Frecuencia", "Porcentaje"),
        format.args = list(big.mark = ",")) %>%
  kable_styling()
```

Table 10: Lugar de Ocurrencia

Lugar	Frecuencia	Porcentaje
CENTRO DE SALUD	2,221	0.27%
DOMICILIO	492,727	60.88%
HOSPITAL PRIVADO	22,830	2.82%
HOSPITAL PUBLICO	167,440	20.69%
IGNORADO	45,301	5.6%
LUGAR DE TRABAJO	61	0.01%
OTRO	11,815	1.46%
SEGURO SOCIAL	41,581	5.14%
VIA PUBLICA	25,316	3.13%

```r
datos %>%
  count(ocur) %>%
  drop_na() %>%
  ggplot(aes(x = reorder(ocur, n), y = n)) +
  geom_col(fill = "purple", alpha = 0.7) +
  geom_text(aes(label = comma(n)), hjust = -0.2, size = 3.5) +
  coord_flip() +
  scale_y_continuous(labels = comma) +
  labs(title = "Lugar de Ocurrencia", x = "Lugar", y = "Frecuencia") +
  theme_minimal()
```



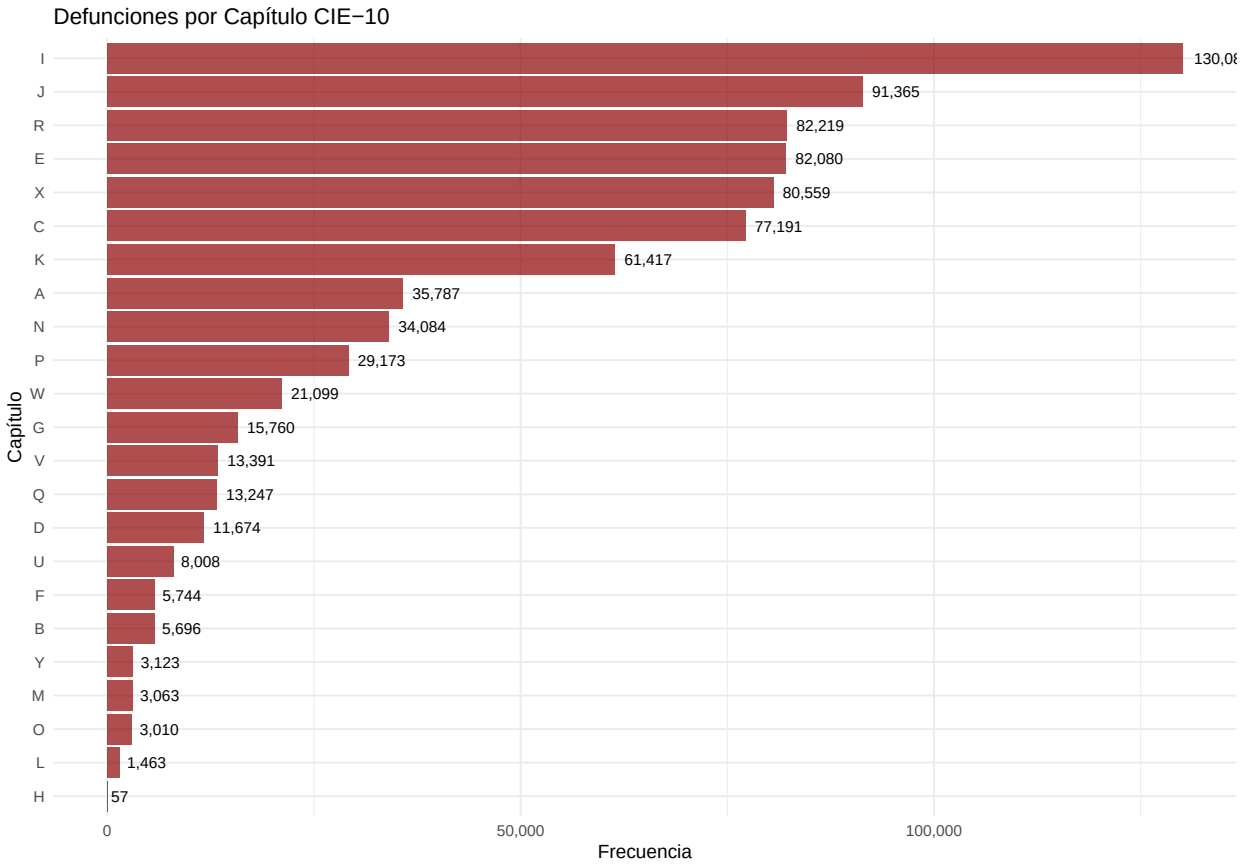
Causas de Muerte (CIE-10)

```r
datos %>%
  count(causa_capitulo) %>%
  drop_na() %>%
  arrange(desc(n)) %>%
  mutate(porcentaje = paste0(round(n / sum(n) * 100, 2), "%")) %>%
  kable(caption = "Capítulos CIE-10",
        col.names = c("Capítulo", "Frecuencia", "Porcentaje"),
        format.args = list(big.mark = ",")) %>%
  kable_styling()
```

Table 11: Capítulos CIE-10

Capítulo	Frecuencia	Porcentaje
I	130,082	16.07%
J	91,365	11.29%
R	82,219	10.16%
E	82,080	10.14%
X	80,559	9.95%
C	77,191	9.54%
K	61,417	7.59%
A	35,787	4.42%
N	34,084	4.21%
P	29,173	3.6%
W	21,099	2.61%
G	15,760	1.95%
V	13,391	1.65%
Q	13,247	1.64%
D	11,674	1.44%
U	8,008	0.99%
F	5,744	0.71%
B	5,696	0.7%
Y	3,123	0.39%
M	3,063	0.38%
O	3,010	0.37%
L	1,463	0.18%
H	57	0.01%

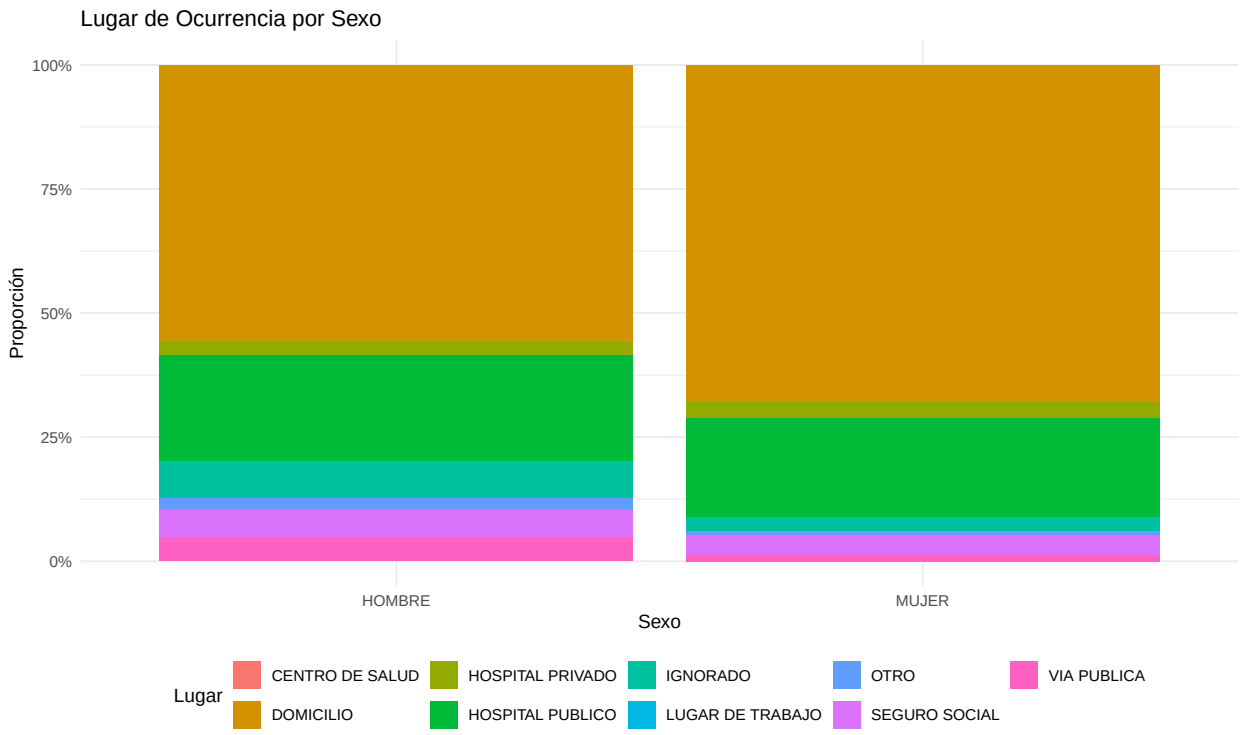
```
datos %>%
  count(causa_capitulo) %>%
  drop_na() %>%
  ggplot(aes(x = reorder(causa_capitulo, n), y = n)) +
  geom_col(fill = "darkred", alpha = 0.7) +
  geom_text(aes(label = comma(n)), hjust = -0.2, size = 3) +
  coord_flip() +
  scale_y_continuous(labels = comma) +
  labs(title = "Defunciones por Capítulo CIE-10", x = "Capítulo", y = "Frecuencia") +
  theme_minimal()
```



Relaciones entre Variables

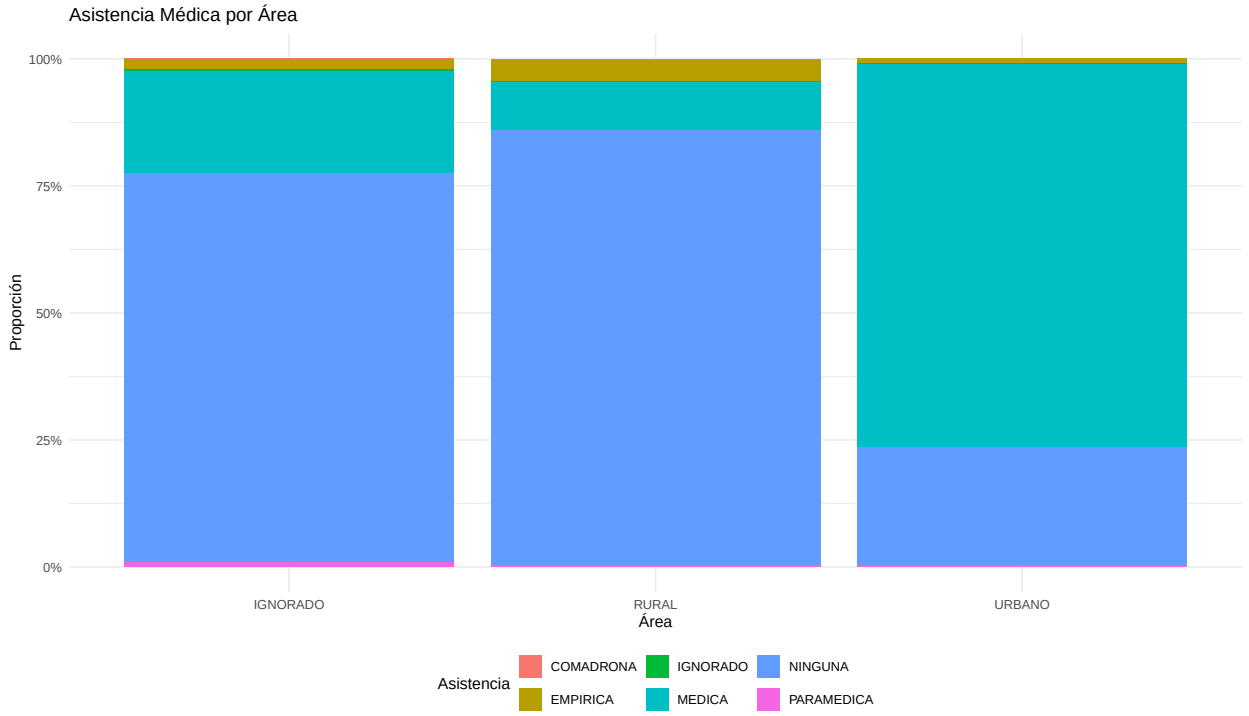
Sexo vs Lugar de Ocurrencia

```
datos %>%
  filter(!is.na(sexo), !is.na(ocur)) %>%
  ggplot(aes(x = sexo, fill = ocur)) +
  geom_bar(position = "fill") +
  labs(title = "Lugar de Ocurrencia por Sexo",
       y = "Proporción", x = "Sexo", fill = "Lugar") +
  scale_y_continuous(labels = percent) +
  theme_minimal() +
  theme(legend.position = "bottom")
```



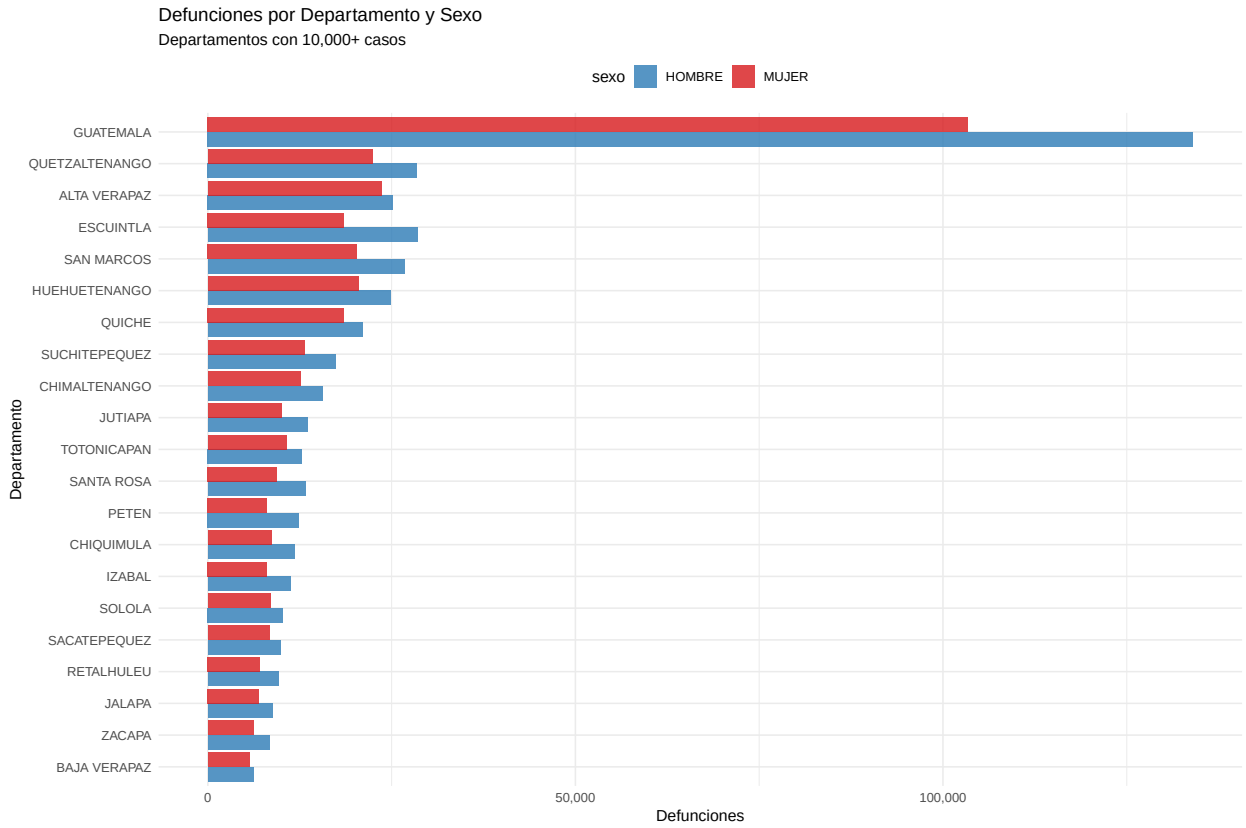
Área vs Asistencia

```
datos %>%
  filter(!is.na(areag), !is.na(asist)) %>%
  count(areag, asist) %>%
  group_by(areag) %>%
  mutate(prop = n / sum(n)) %>%
  ggplot(aes(x = areag, y = prop, fill = asist)) +
  geom_col(position = "fill") +
  labs(title = "Asistencia Médica por Área", x = "Área", y = "Proporción", fill = "Asistencia") +
  scale_y_continuous(labels = percent) +
  theme_minimal() +
  theme(legend.position = "bottom")
```



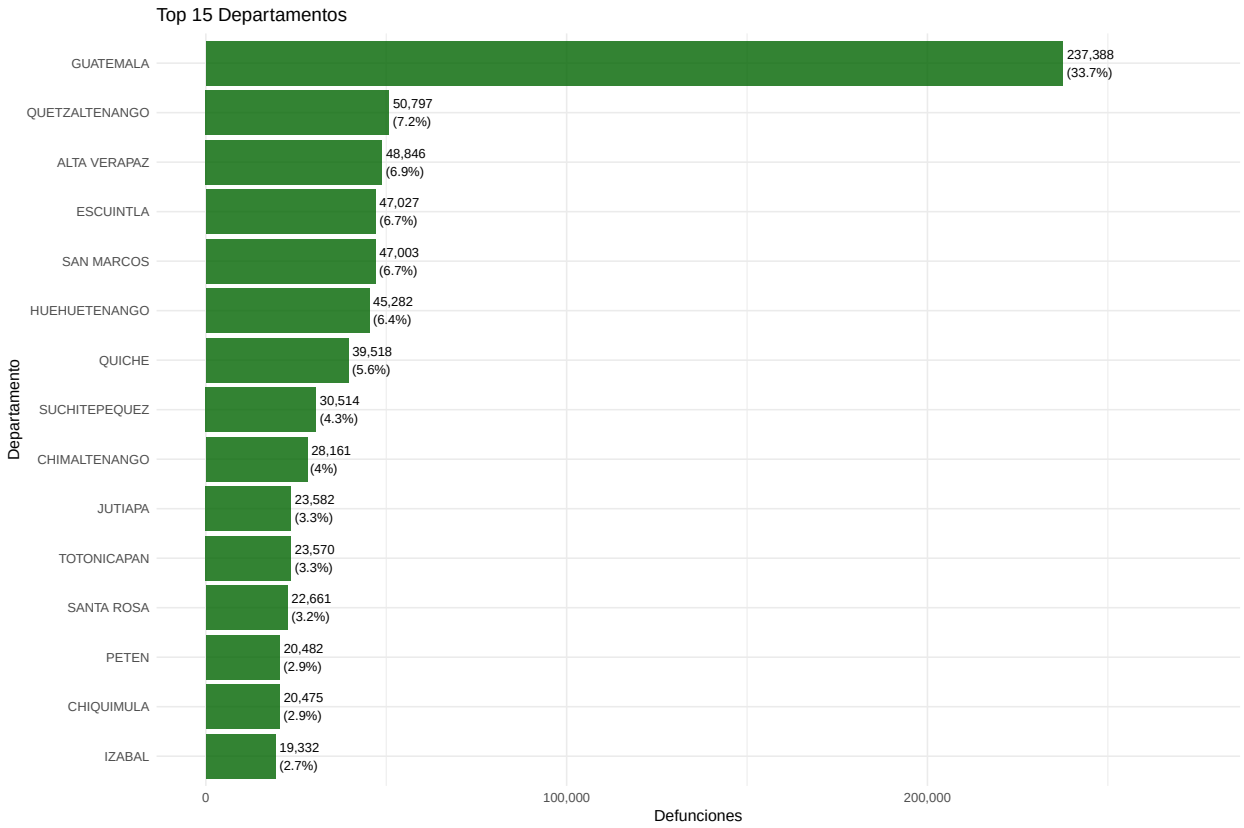
Departamento vs Sexo

```
datos %>%
  filter(!is.na(depocu), !is.na(sexo)) %>%
  count(depocu, sexo) %>%
  group_by(depocu) %>%
  filter(sum(n) >= 10000) %>%
  ungroup() %>%
  ggplot(aes(x = reorder(depocu, n), y = n, fill = sexo)) +
  geom_col(position = "dodge", alpha = 0.8) +
  coord_flip() +
  scale_fill_manual(values = c("HOMBRE" = "#2c7bb6", "MUJER" = "#d7191c")) +
  scale_y_continuous(labels = comma) +
  labs(title = "Defunciones por Departamento y Sexo",
       subtitle = "Departamentos con 10,000+ casos",
       x = "Departamento", y = "Defunciones") +
  theme_minimal() +
  theme(legend.position = "top")
```



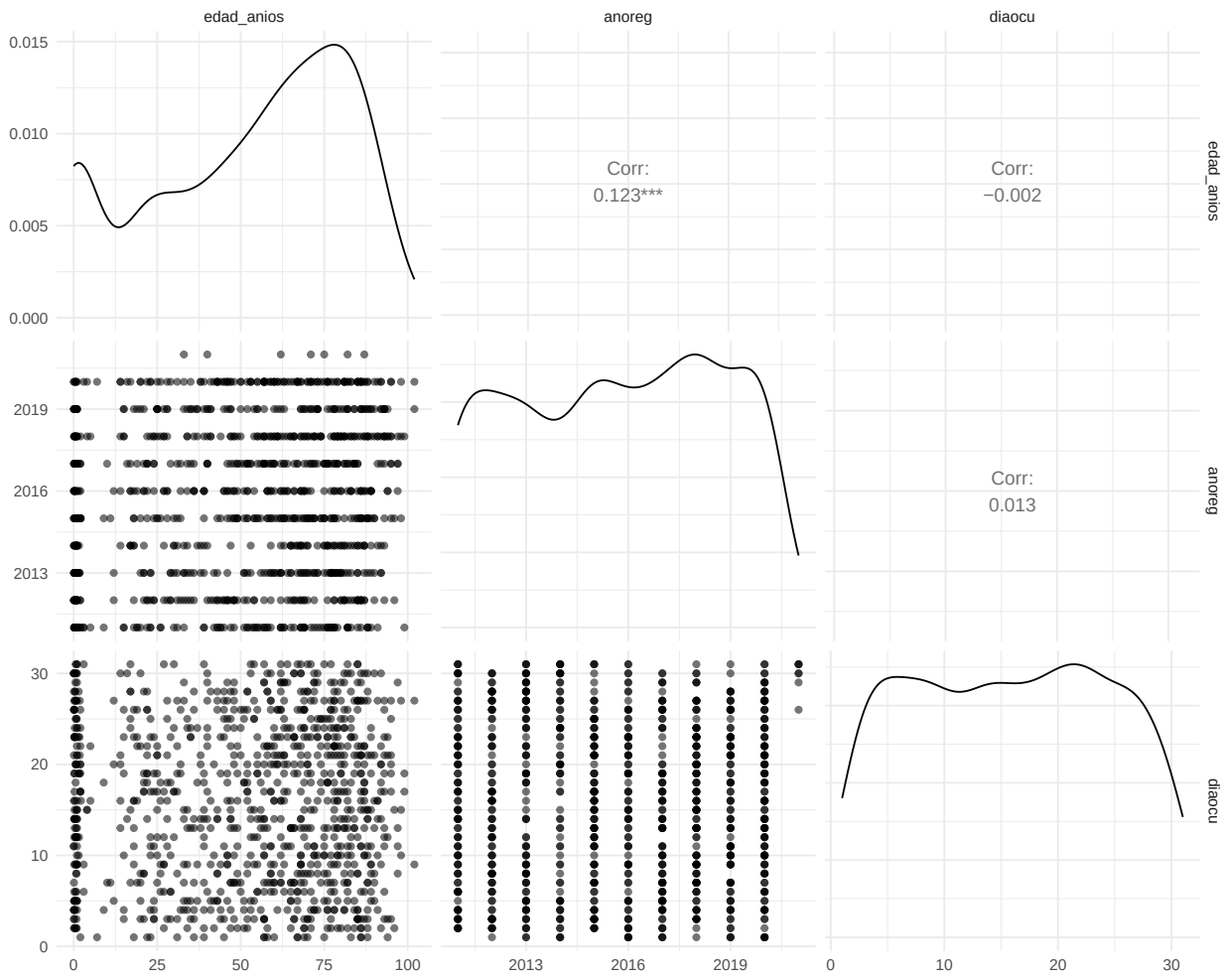
Top Departamentos

```
datos %>%
  count(depocu) %>%
  drop_na() %>%
  arrange(desc(n)) %>%
  head(15) %>%
  mutate(porcentaje = n / sum(n) * 100) %>%
  ggplot(aes(x = reorder(depocu, n), y = n)) +
  geom_col(fill = "darkgreen", alpha = 0.8) +
  geom_text(aes(label = paste0(comma(n), "\n(", round(porcentaje, 1), "%)")),
            hjust = -0.1, size = 3) +
  coord_flip() +
  scale_y_continuous(labels = comma, limits = c(0, max(datos %>% count(depocu) %>% pull(n), na.rm = TRU
  labs(title = "Top 15 Departamentos", x = "Departamento", y = "Defunciones") +
  theme_minimal()
```



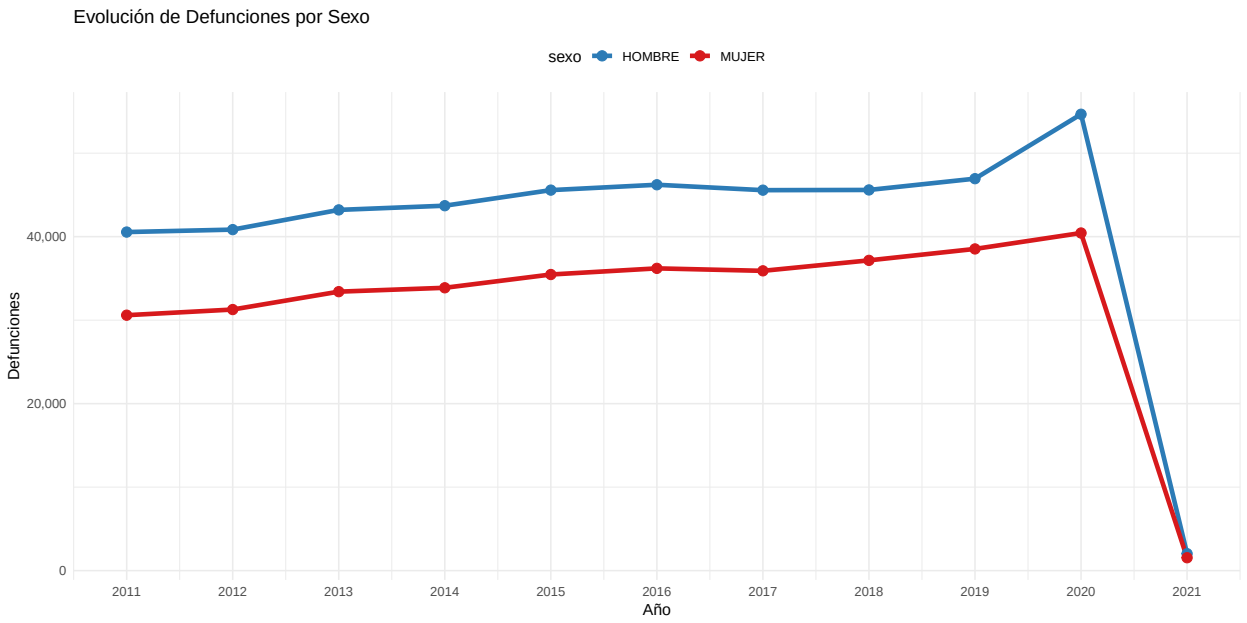
Correlaciones

```
set.seed(42)
datos %>%
  select(edad_anios, anoreg, diaocu) %>%
  drop_na() %>%
  slice_sample(n = min(1000, nrow(.))) %>%
  ggpairs(aes(alpha = 0.4)) +
  theme_minimal()
```



Evolución por Sexo

```
datos %>%
  filter(!is.na(anoreg), !is.na(sexo)) %>%
  count(anoreg, sexo) %>%
  ggplot(aes(x = anoreg, y = n, color = sexo)) +
  geom_line(size = 1.5) +
  geom_point(size = 3) +
  scale_color_manual(values = c("HOMBRE" = "#2c7bb6", "MUJER" = "#d7191c")) +
  scale_x_continuous(breaks = seq(min(datos$anoreg, na.rm = TRUE),
                                  max(datos$anoreg, na.rm = TRUE), 1)) +
  scale_y_continuous(labels = comma) +
  labs(title = "Evolución de Defunciones por Sexo", x = "Año", y = "Defunciones") +
  theme_minimal() +
  theme(legend.position = "top")
```



Preguntas de investigación (supuestos a validar)

A continuación se plantean **5 preguntas** basadas en supuestos comunes sobre el fenómeno. En cada una se indica el supuesto y se valida/refuta con análisis de datos.

> Nota: Las conclusiones dependen de los resultados que generen las tablas y gráficos al ejecutar el documento.

P1. ¿Existen diferencias en la edad al fallecer entre hombres y mujeres?

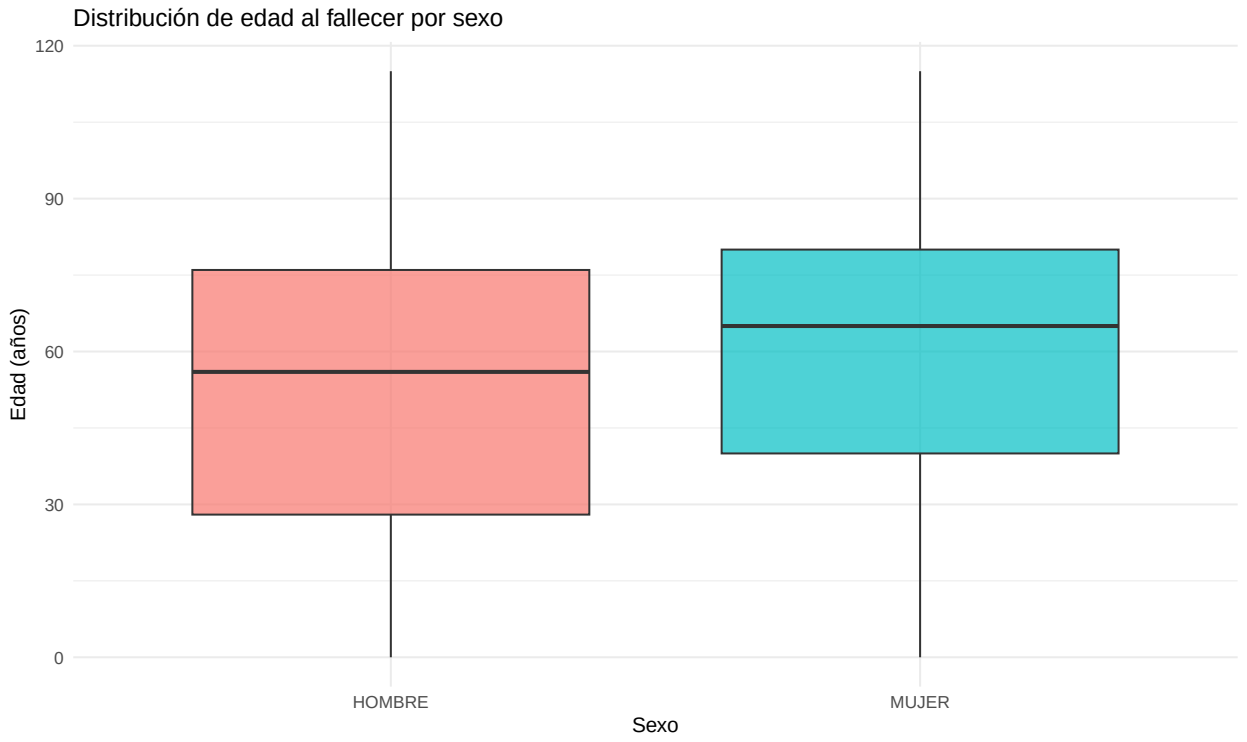
Supuesto: La distribución de edad difiere por sexo (por ejemplo, mayor edad promedio en mujeres).

```
datos %>%
  filter(!is.na(edad_anios), !is.na(sexo)) %>%
  group_by(sexo) %>%
  summarise(
    N = n(),
    media = mean(edad_anios, na.rm = TRUE),
    mediana = median(edad_anios, na.rm = TRUE),
    q1 = quantile(edad_anios, 0.25, na.rm = TRUE),
    q3 = quantile(edad_anios, 0.75, na.rm = TRUE)
  ) %>%
  arrange(desc(N)) %>%
  kable(caption = "Edad (años) por sexo") %>%
  kable_styling(bootstrap_options = c("striped", "hover"), full_width = FALSE)
```

Table 12: Edad (años) por sexo

sexo	N	media	mediana	q1	q3
HOMBRE	450799	51.00790	56	28	76
MUJER	351994	56.99476	65	40	80

```
datos %>%
  filter(!is.na(edad_anios), !is.na(sexo)) %>%
  ggplot(aes(x = sexo, y = edad_anios, fill = sexo)) +
  geom_boxplot(alpha = 0.7, outlier.alpha = 0.2) +
  labs(title = "Distribución de edad al fallecer por sexo", x = "Sexo", y = "Edad (años)") +
  theme(legend.position = "none")
```



P2. ¿La asistencia médica se asocia con el área (urbana/rural)?

Supuesto: En áreas rurales hay menor proporción de asistencia médica.

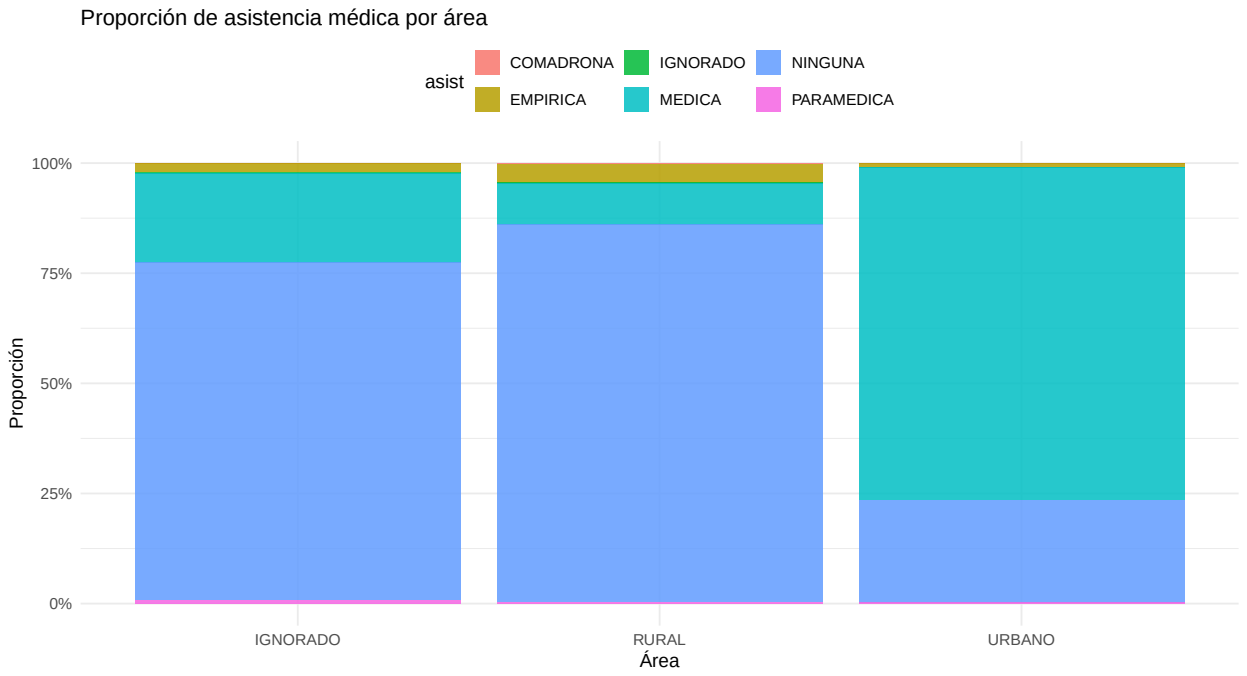
```
tabla <- datos %>%
  filter(!is.na(areag), !is.na(asist)) %>%
  count(areag, asist) %>%
  group_by(areag) %>%
  mutate(prop = n / sum(n)) %>%
  ungroup()

tabla %>%
  arrange(areag, desc(n)) %>%
```

```
  kable(caption = "Frecuencias y proporciones: Área vs Asistencia médica") %>%
  kable_styling(bootstrap_options = c("striped", "hover"), full_width = FALSE)
```

Table 13: Frecuencias y proporciones: Área vs Asistencia médica

areag	asist	n	prop
IGNORADO	NINGUNA	7788	0.7657064
IGNORADO	MEDICA	2055	0.2020450
IGNORADO	EMPIRICA	206	0.0202537
IGNORADO	PARAMEDICA	83	0.0081605
IGNORADO	IGNORADO	33	0.0032445
IGNORADO	COMADRONA	6	0.0005899
RURAL	NINGUNA	204915	0.8560024
RURAL	MEDICA	22367	0.0934349
RURAL	EMPIRICA	10230	0.0427343
RURAL	PARAMEDICA	896	0.0037429
RURAL	IGNORADO	599	0.0025022
RURAL	COMADRONA	379	0.0015832
URBANO	MEDICA	222758	0.7549481
URBANO	NINGUNA	68453	0.2319937
URBANO	EMPIRICA	2351	0.0079678
URBANO	PARAMEDICA	1005	0.0034060
URBANO	IGNORADO	440	0.0014912
URBANO	COMADRONA	57	0.0001932

```
datos %>%
  filter(!is.na(areag), !is.na(asist)) %>%
  count(areag, asist) %>%
  group_by(areag) %>%
  mutate(prop = n / sum(n)) %>%
  ggplot(aes(x = areag, y = prop, fill = asist)) +
  geom_col(position = "fill", alpha = 0.85) +
  scale_y_continuous(labels = percent_format()) +
  labs(title = "Proporción de asistencia médica por área", x = "Área", y = "Proporción") +
  theme(legend.position = "top")
```



P3. ¿El lugar de ocurrencia (hogar/hospital/otro) cambia según sexo?

Supuesto: Hay diferencias por sexo en el lugar donde ocurre la defunción.

```
datos %>%
  filter(!is.na(sexo), !is.na(ocur)) %>%
  count(sexo, ocur) %>%
  group_by(sexo) %>%
  mutate(prop = n / sum(n)) %>%
  ungroup() %>%
  arrange(sexo, desc(prop)) %>%
  kable(caption = "Lugar de ocurrencia por sexo (proporciones)") %>%
  kable_styling(bootstrap_options = c("striped", "hover"), full_width = FALSE)
```

Table 14: Lugar de ocurrencia por sexo (proporciones)

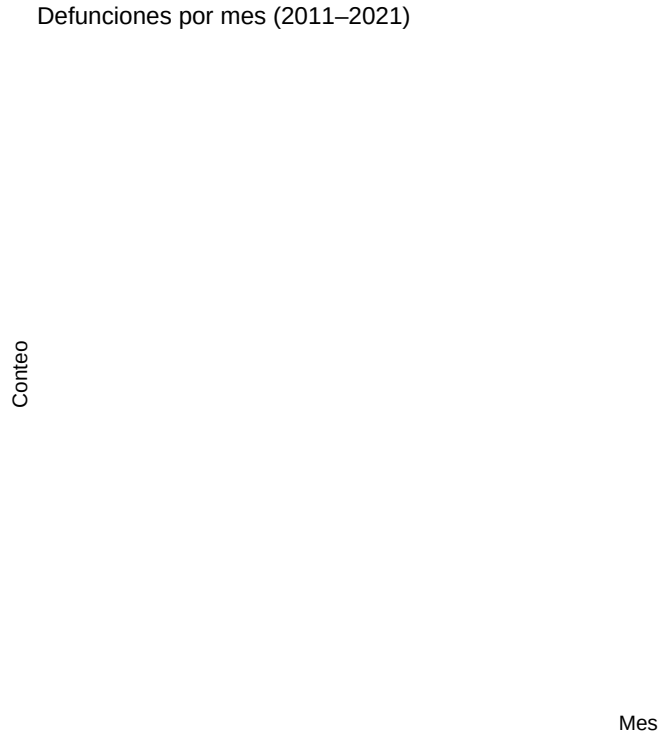
sexo	ocur	n	prop
HOMBRE	DOMICILIO	252690	0.5554872
HOMBRE	HOSPITAL PUBLICO	96777	0.2127444
HOMBRE	IGNORADO	34780	0.0764567
HOMBRE	SEGURO SOCIAL	26503	0.0582614
HOMBRE	VIA PUBLICA	21621	0.0475293
HOMBRE	HOSPITAL PRIVADO	11721	0.0257662
HOMBRE	OTRO	9352	0.0205585
HOMBRE	CENTRO DE SALUD	1401	0.0030798
HOMBRE	LUGAR DE TRABAJO	53	0.0001165
MUJER	DOMICILIO	240037	0.6773168
MUJER	HOSPITAL PUBLICO	70663	0.1993911

MUJER	SEGURO SOCIAL	15078	0.0425459
MUJER	HOSPITAL PRIVADO	11109	0.0313465
MUJER	IGNORADO	10521	0.0296873
MUJER	VIA PUBLICA	3695	0.0104262
MUJER	OTRO	2463	0.0069499
MUJER	CENTRO DE SALUD	820	0.0023138
MUJER	LUGAR DE TRABAJO	8	0.0000226

P4. ¿Hay estacionalidad en las defunciones (meses con más registros)?

Supuesto: Existen meses del año con mayor cantidad de registros.

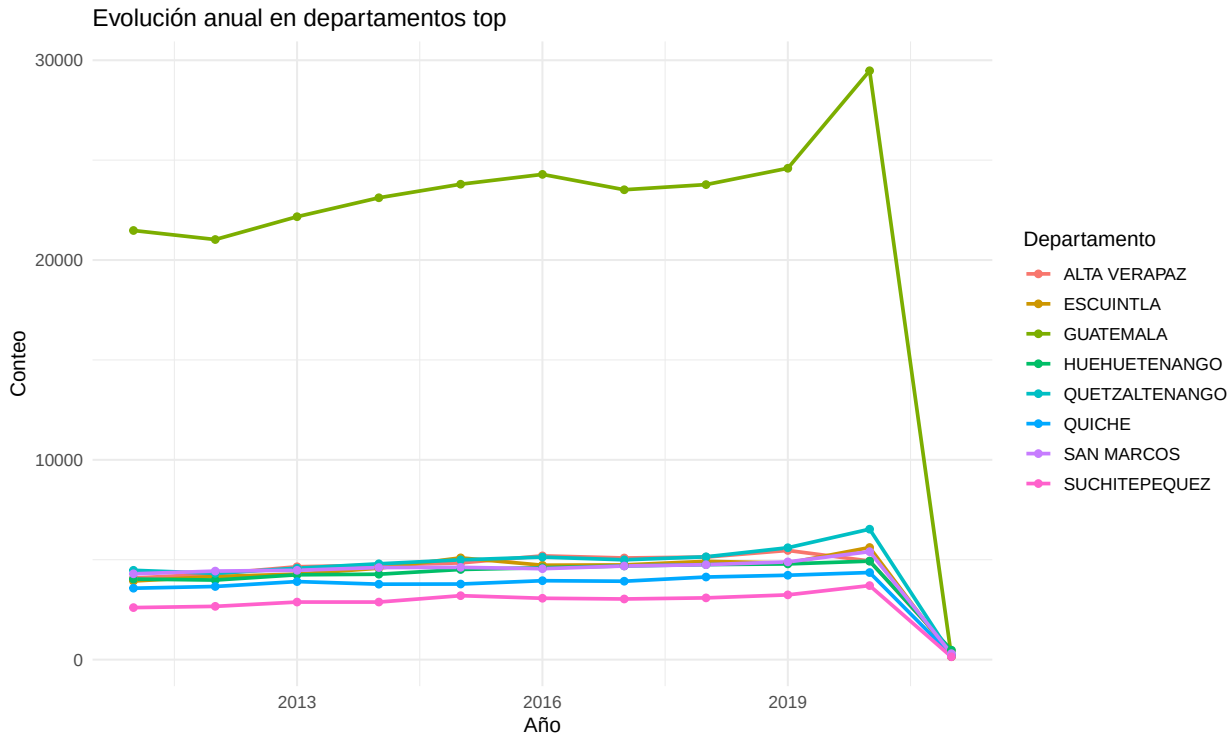
```
datos %>%
  filter(!is.na(mesreg)) %>%
  count(mesreg) %>%
  mutate(mesreg = as.integer(mesreg)) %>%
  arrange(mesreg) %>%
  ggplot(aes(x = mesreg, y = n)) +
  geom_col(alpha = 0.85) +
  scale_x_continuous(breaks = 1:12) +
  labs(title = "Defunciones por mes (2011-2021)", x = "Mes", y = "Conteo")
```



P5. ¿Los departamentos con más registros se mantienen en el tiempo?

Supuesto: Los “top” departamentos se repiten año con año.

```r
top_deps <- datos %>%
  filter(!is.na(depocu), !is.na(anoreg)) %>%
  count(depocu, sort = TRUE) %>%
  slice_head(n = 8) %>%
  pull(depocu)

datos %>%
  filter(depocu %in% top_deps, !is.na(anoreg)) %>%
  count(anoreg, depocu) %>%
  ggplot(aes(x = as.integer(anoreg), y = n, color = depocu)) +
  geom_line(linewidth = 1) +
  geom_point() +
  labs(title = "Evolución anual en departamentos top", x = "Año", y = "Conteo", color = "Departamento") +
  theme(legend.position = "right")
```



Clustering (agrupamiento) e interpretación

En esta sección se realiza un agrupamiento para identificar **perfiles** de defunciones.
Como el dataset incluye variables categóricas y numéricas, se usa **distancia de Gower** (apta para datos mixtos) y **PAM** (k-medoids).

Selección de variables para clustering

```r
library(dplyr)
```

```r
df_clust <- datos %>%
  transmute(
    edad_anios = edad_anios,
    sexo  = factor(sexo),
    areag = factor(areag),
    asist = factor(asist),
    ocur  = factor(ocur)
  ) %>%
  drop_na()

# One-hot encoding + escalado
X <- model.matrix(~ . - 1, data = df_clust)
X_scaled <- scale(X)

cat("Filas clustering:", nrow(X_scaled), " | Features:", ncol(X_scaled), "\n")
```

```
## Filas clustering: 539160  | Features: 18
```

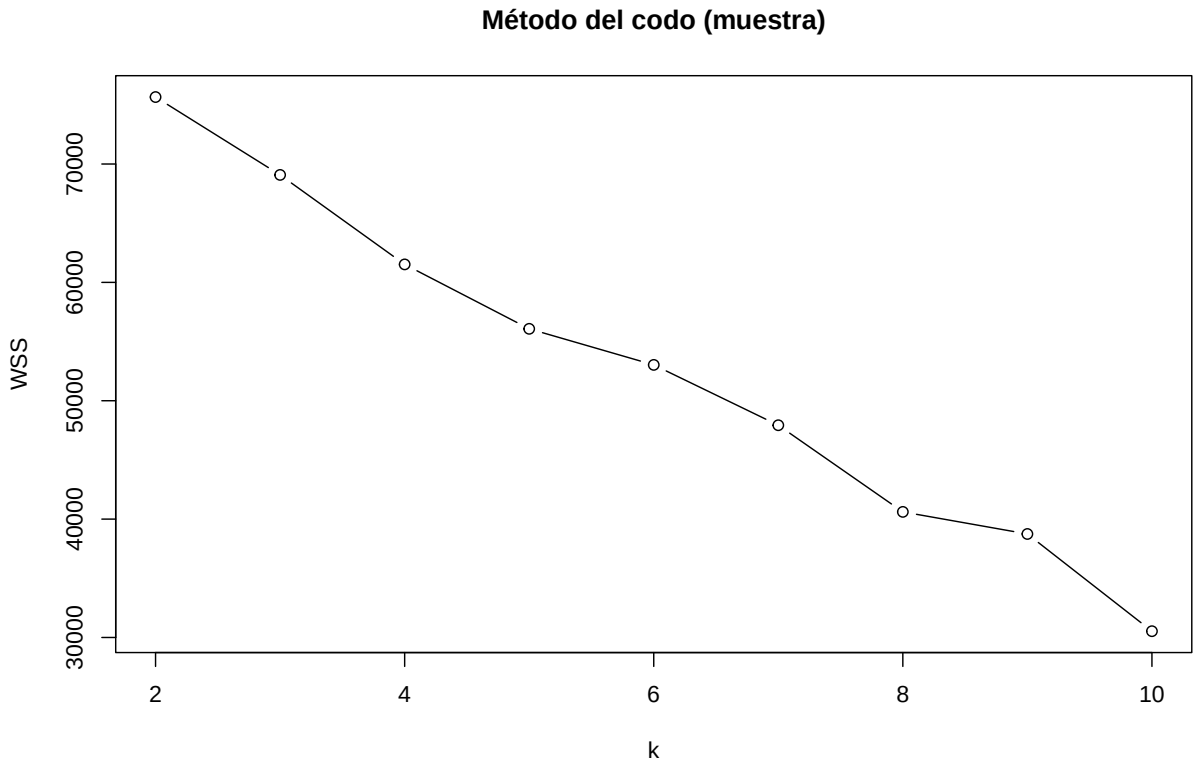
Elegir número de clusters (k) con silueta

```r
library(cluster)

set.seed(123)
n_samp <- min(5000, nrow(X_scaled))
idx <- sample(seq_len(nrow(X_scaled)), n_samp)
X_s <- X_scaled[idx, ]

ks <- 2:10
wss <- numeric(length(ks))
sil_avg <- numeric(length(ks))

for(i in seq_along(ks)){
  k <- ks[i]
  km <- kmeans(X_s, centers = k, nstart = 20)
  wss[i] <- km$tot.withinss

  sil <- silhouette(km$cluster, dist(X_s))
  sil_avg[i] <- mean(sil[, 3])
}

k_eval <- data.frame(k = ks, WSS = wss, silueta_promedio = sil_avg)
k_eval
```

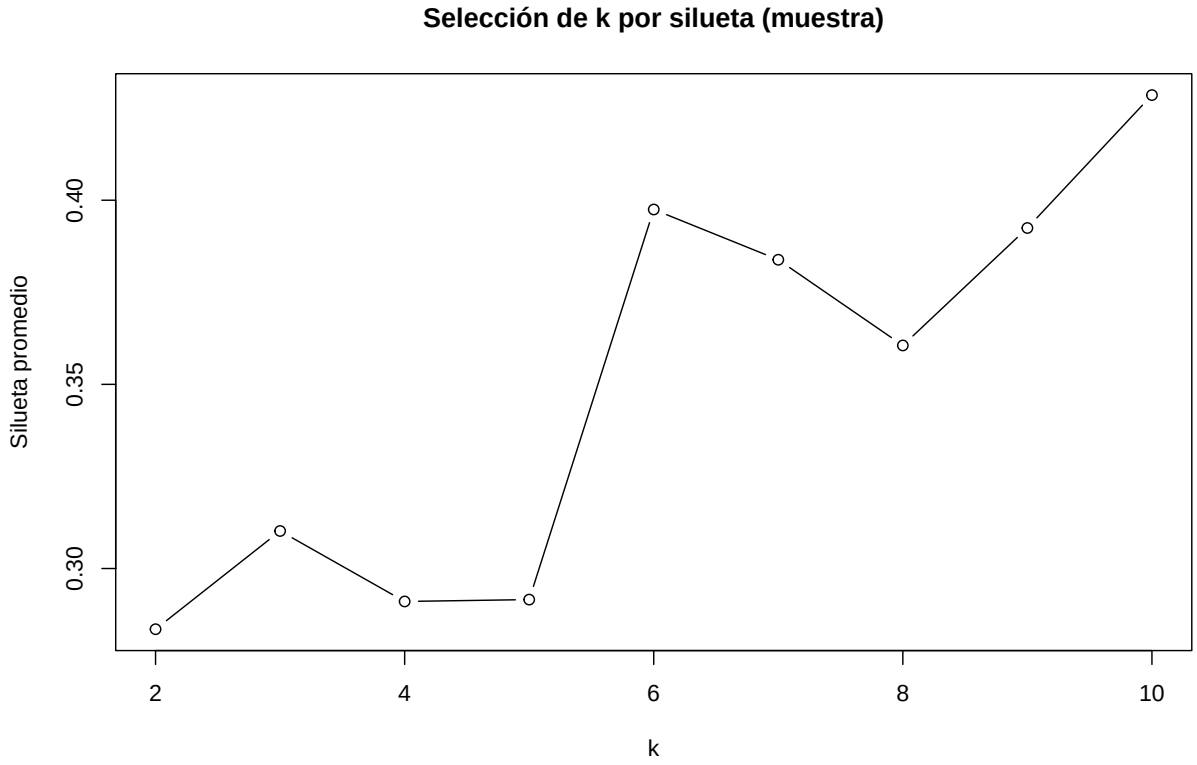
```
##    k      WSS silueta_promedio
## 1  2 75657.01        0.2835185
## 2  3 69078.45        0.3101856
## 3  4 61519.76        0.2910505
## 4  5 56075.78        0.2915601
## 5  6 53028.86        0.3974797
## 6  7 47931.78        0.3838330
## 7  8 40610.06        0.3605648
```

```
## 8  9 38743.85        0.3924519
## 9 10 30530.82        0.4285451
```

```r
plot(k_eval$k, k_eval$WSS, type="b",
     xlab="k", ylab="WSS", main="Método del codo (muestra)")
```




```r
plot(k_eval$k, k_eval$silueta_promedio, type="b",
     xlab="k", ylab="Silueta promedio", main="Selección de k por silueta (muestra)")
```



Ajuste final y perfil de clusters

```
set.seed(123)
k_opt <- k_eval$k[which.max(k_eval$silueta_promedio)]
cat("k óptimo:", k_opt, "\n")
```

```
## k óptimo: 10
```

```
km_final <- kmeans(X_scaled, centers = k_opt, nstart = 30)
df_clust$cluster <- factor(km_final$cluster)

# Resumen numérico (edad) por cluster
res_num <- df_clust %>%
  group_by(cluster) %>%
  summarise(
    N = n(),
    edad_media = mean(edad_anios, na.rm = TRUE),
    edad_mediana = median(edad_anios, na.rm = TRUE),
    .groups = "drop"
  ) %>% arrange(desc(N))

res_num
```

```
## # A tibble: 10 x 4
```

```
##    cluster      N edad_media edad_mediana
##    <fct>    <int>      <dbl>        <dbl>
##  1 3       123123       58.6           67
##  2 5       110778       37.3           38
##  3 4       108249       55.2           65
##  4 7        58395       62.9           69
##  5 10       50924       68.0           73
##  6 6        29688       42.9           38
##  7 1        28116       35.4           31
##  8 9        26887       51.1           60
##  9 8         1964       45.6           47
## 10 2         1036       43.5           39
```

```
# Top categorías por cluster
top_cat <- function(var){
  df_clust %>%
    count(cluster, .data[[var]], sort = TRUE) %>%
    group_by(cluster) %>%
    slice_head(n = 3) %>%
    ungroup()
}

top_cat("sexo")
```

```
## # A tibble: 17 x 3
##    cluster sexo        n
##    <fct>   <fct>   <int>
##  1 1       HOMBRE  23576
##  2 1       MUJER    4540
##  3 2       HOMBRE    789
##  4 2       MUJER     247
##  5 3       MUJER  123123
##  6 4       HOMBRE 108249
##  7 5       HOMBRE  64104
##  8 5       MUJER   46674
##  9 6       HOMBRE  22361
## 10 6       MUJER    7327
## 11 7       HOMBRE  58395
## 12 8       HOMBRE   1229
## 13 8       MUJER     735
## 14 9       HOMBRE  17024
## 15 9       MUJER    9863
## 16 10      MUJER   43134
## 17 10      HOMBRE   7790
```

```
top_cat("areag")
```

```
## # A tibble: 28 x 3
##    cluster areag         n
##    <fct>   <fct>     <int>
##  1 1       RURAL     13845
##  2 1       URBANO    13302
##  3 1       IGNORADO    969
```

```
##  4 2       RURAL       581
##  5 2       URBANO      426
##  6 2       IGNORADO     29
##  7 3       RURAL     97570
##  8 3       URBANO    23935
##  9 3       IGNORADO   1618
## 10 4       RURAL    106282
## # i 18 more rows
```

```
top_cat("asist")
```

```
## # A tibble: 26 x 3
##    cluster asist         n
##    <fct>   <fct>     <int>
##  1 1       NINGUNA   25121
##  2 1       MEDICA     2833
##  3 1       EMPIRICA    156
##  4 2       IGNORADO   1036
##  5 3       NINGUNA  107419
##  6 3       MEDICA     9588
##  7 3       EMPIRICA   5861
##  8 4       NINGUNA   93075
##  9 4       MEDICA    10063
## 10 4       EMPIRICA   4996
## # i 16 more rows
```

```
top_cat("ocur")
```

```
## # A tibble: 24 x 3
##    cluster ocur                n
##    <fct>   <fct>           <int>
##  1 1       VIA PUBLICA     19179
##  2 1       OTRO             8878
##  3 1       LUGAR DE TRABAJO   59
##  4 2       DOMICILIO         409
##  5 2       IGNORADO          273
##  6 2       VIA PUBLICA       144
##  7 3       DOMICILIO      123032
##  8 3       CENTRO DE SALUD    85
##  9 3       HOSPITAL PUBLICO    6
## 10 4       DOMICILIO      108176
## # i 14 more rows
```

Conclusiones

Hallazgos Principales

Variables Numéricas:

- Edad promedio: ~53.6 años

- Distribución bimodal (infantil y adulto mayor)
- Incremento notable 2020-2021

Variables Categóricas:

- Mayor mortalidad masculina (~59%)
- Concentración urbana
- Causas mal definidas predominantes

Relaciones:

- Área geográfica asociada con asistencia médica
- Departamentos poblados concentran más casos absolutos
- Patrón consistente de mayor mortalidad masculina

```
sessionInfo()
```

```
## R version 4.3.1 (2023-06-16 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19045)
##
## Matrix products: default
##
##
## locale:
## [1] LC_COLLATE=Spanish_Guatemala.utf8  LC_CTYPE=Spanish_Guatemala.utf8
## [3] LC_MONETARY=Spanish_Guatemala.utf8 LC_NUMERIC=C
## [5] LC_TIME=Spanish_Guatemala.utf8
##
## time zone: America/Guatemala
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] cluster_2.1.4    scales_1.4.0     kableExtra_1.4.0 stringi_1.8.7
##  [5] GGally_2.4.0     janitor_2.2.1    lubridate_1.9.4  forcats_1.0.1
##  [9] stringr_1.6.0    dplyr_1.1.4      purrr_1.0.4      readr_2.1.5
## [13] tidyr_1.3.1      tibble_3.2.1     ggplot2_4.0.1    tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] utf8_1.2.3         generics_0.1.4     xml2_1.3.8         hms_1.1.3
##  [5] digest_0.6.37      magrittr_2.0.3     evaluate_1.0.5     grid_4.3.1
##  [9] timechange_0.3.0   RColorBrewer_1.1-3 fastmap_1.2.0      fansi_1.0.4
## [13] viridisLite_0.4.2  cli_3.6.1          crayon_1.5.2       rlang_1.1.5
## [17] bit64_4.6.0-1      withr_3.0.2        yaml_2.3.10        parallel_4.3.1
## [21] tools_4.3.1        tzdb_0.5.0         ggstats_0.12.0     vctrs_0.6.5
## [25] R6_2.5.1           lifecycle_1.0.3    snakecase_0.11.1   bit_4.6.0
## [29] vroom_1.6.5        pkgconfig_2.0.3    pillar_1.9.0       gtable_0.3.6
```

```
## [33] glue_1.6.2         systemfonts_1.2.2  xfun_0.52          tidyselect_1.2.1
## [37] rstudioapi_0.18.0  knitr_1.51         farver_2.1.2       htmltools_0.5.8.1
## [41] labeling_0.4.3     rmarkdown_2.30     svglite_2.1.3      compiler_4.3.1
## [45] S7_0.2.0
```

Siguientes pasos

- Refinar limpieza (valores faltantes y categorías poco informativas).
- Validar preguntas adicionales (por ejemplo, causas CIE-10 por grupo etario/departamento).
- Profundizar en los clusters: nombrar grupos, comparar distribuciones y generar hipótesis nuevas.
- Preparar el informe final en PDF **sin código**, resumiendo hallazgos con tablas y gráficos clave.