Proyecto 1

Bryan Martínez, Adriana Palacios, Brandon Rivera, Javier Benítez, y Pedro Avila

2026-02-16

Situación Problemática

En los últimos años se ha vuelto común escuchar que el matrimonio y la formación de una familia ya no forman parte de los planes de muchas personas, o que estas decisiones se están postergando a edades más avanzadas. Asimismo, se percibe que las relaciones de pareja tienden a adoptar formas distintas a las tradicionales, ahora vemos a más personas que buscan convivir con una pareja sin matrimonio o bajo compromisos legales.

Sin embargo, estas percepciones sociales no siempre son evidentes o respaldadas con información verídica. En particular, no se conoce con certeza si en Guatemala ha ocurrido una disminución en los matrimonios, un aumento en los divorcios, o cambios significativos en la edad a la que las personas deciden casarse o divorciarse. Tampoco es claro si estos posibles cambios se presentan de manera uniforme en todo el país o si existen departamentos en donde se identifiquen diferencias relevantes.

Al tener acceso a datos oficiales por parte de la INE sobre matrimonios y divorcios en Guatemala, surge la necesidad de analizar estos datos para tratar de responder a estas interrogantes, buscando identificar diferentes patrones a lo largo de la última década y media. Por medio de nuestro análisis seremos capaces de contrastar las percepciones sociales con datos reales, y comprender si realmente ha ocurrido un cambio en la dinámica de los matrimonios y divorcios guatemaltecos.

Problema Científico

No se cuenta con evidencia estadística concluyente sobre si, en Guatemala (2009–2022), han ocurrido cambios significativos en los matrimonios y divorcios, ni si estos cambios se relacionan con la edad de las personas o presentan diferencias relevantes entre departamentos.

Repositorio

https://github.com/Bamo0507/Proyecto1_Mineria

Objetivos

Objetivo General

Analizar los datos de matrimonios y divorcios en Guatemala con el fin de identificar patrones demográficos y diferencias entre departamentos, y evaluar si se han producido cambios en la dinámica de estos eventos a lo largo del período analizado.

Objetivos Específicos

1. Examinar la distribución de los matrimonios y divorcios según los rangos de edad de las personas, con el propósito de identificar posibles variaciones en las edades en las que ocurren estos eventos.
2. Analizar la distribución de matrimonios y divorcios por departamento, con el fin de detectar la existencia de patrones regionales relevantes.

Descripción de los Datos

```
# Algunos departamentos vienen escritos diferente, mayúsculas, minúsculas o con/sin tildes.
norm_depto <- function(x) {
  x %>%
    str_trim() %>%
    str_squish() %>%
    stri_trans_general("Latin-ASCII") %>%  # quita tildes (Petén -> Peten)
    str_to_lower()
}

# Carga de datos
matrimonios_depto <- read_csv("matrimonios_depto_mes.csv", show_col_types = FALSE) %>%
  mutate(departamento = norm_depto(departamento))

matrimonios_edad  <- read_csv("matrimonios_edad.csv", show_col_types = FALSE)

divorcios_depto <- read_csv("divorcios_depto_mes.csv", show_col_types = FALSE) %>%
  mutate(departamento = norm_depto(departamento))

divorcios_edad  <- read_csv("divorcios_edad.csv", show_col_types = FALSE)

# Limpieza de datasets
matrimonios_depto_limpio <- matrimonios_depto %>%
  filter(nivel_geo == "departamento", !is.na(mes))

divorcios_depto_limpio <- divorcios_depto %>%
  filter(nivel_geo == "departamento", !is.na(mes))

matrimonios_edad_limpio <- matrimonios_edad %>%
  filter(
    !str_detect(tolower(edad_hombre_grupo), "ignorado"),
    !str_detect(tolower(edad_mujer_grupo), "ignorado")
  )

divorcios_edad_limpio <- divorcios_edad %>%
  filter(
    !str_detect(tolower(edad_hombre_grupo), "ignorado"),
```

```
    !str_detect(tolower(edad_mujer_grupo), "ignorado")
  )
```

Significado y tipo de cada variable

Para el análisis de los datos, se estará trabajando con cuatro datasets con información desde 2009 hasta 2022:

- matrimonios_depto_mes: matrimonios registrados en cada mes acorde a cada departamento.
- matrimonios_edad: matrimonios que ocurrieron en diferentes rangos de edad.
- divorcios_depto: divorcios almacenados por mes para cada departamento.
- divorcios_edad: divorcios acontecidos en diferentes rangos de edad.

Variables Compartidas en Todos los Datasets anio

- Representa el año en que fue registrada la observación.
- La variable es cuantitativa discreta.

valor

- Número de veces que ocurrió un evento, dependiendo del dataset, este puede representar cantidad de matrimonios o divorcios.
- La variable es cuantitativa discreta.

Variables Compartidas en Datasets con Datos Departamentales nivel_geo

- Se utiliza para indicar si la observación es de un departamento en específico, o si se tiene registrada a nivel nacional.
- La variable es cualitativa nominal.

departamento

- Es el departamento en donde se registró la observación; este valor puede estar vacío cuando el registro corresponde al total nacional.
- La variable es cualitativa nominal.

mes

- Muestra el número de mes al que corresponden los datos de la observación.
- La variable es cualitativa ordinal; aunque se representa como un número en el dataset, se utiliza para indicar la posición del mes dentro del año y no como una magnitud numérica.

Variables Compartidas en Datasets con Datos de Edades edad_mujer_grupo

- Rango de edad al que pertenece la novia/mujer de la observación.
- La variable es cualitativa ordinal.

edad_hombre_grupo

- Rango de edad al que pertence el novio/hombre de la observación.
- La variable es cualitativa ordinal.

Cantidad de Variables y Observaciones

Matrimonios por departamento:

```
dim(matrimonios_depto)
```

```
## [1] 4186    5
```

Se tienen 4186 observaciones y 5 variables.

Matrimonios por rangos de edad:

```
dim(matrimonios_edad)
```

```
## [1] 2191    4
```

Hay 2191 observaciones y 4 variables.

Divorcios por departamento:

```
dim(divorcios_depto)
```

```
## [1] 4186    5
```

Se tienen 4186 observaciones y 5 variables.

Divorcios por rangos de edad:

```
dim(divorcios_edad)
```

```
## [1] 1878    4
```

Hay 1878 observaciones y 4 variables.

Operaciones de Limpieza Realizadas

Al explorar las opciones de descarga de datos disponibles en la plataforma del INE, se observó que gran parte de la información se ofrece en formato .sav, lo cual requiere software de IBM o procesos adicionales para manipularlos. Para facilitar el análisis, se decidió utilizar los libros de Excel disponibles para las estadísticas de matrimonios y divorcios.

Los archivos descargados contenían múltiples hojas y tablas cruzadas que no se encontraban en un formato que se pudiera utilizar directamente para análisis en R. Por esta razón, se realizó un proceso de limpieza. Primero, se seleccionaron únicamente las hojas asociadas a rangos de edad y distribución por departamento en todos los libros de matrimonios y divorcios.

Después, se transformaron las tablas cruzadas a un formato que permitiera el análisis estadístico. La idea era que en cada observación se pudiera identificar sencillamente los datos para un año determinado, logrando reconocer el departamento, rangos de edad de los novios, y la cantidad de un evento (matrimonio o divorcio). Asimismo, se estandarizaron varios datos, pues en varias ocasiones se redactaba de forma distinta, por ejemplo, había matrimonios o divorcios que ocurrían después de los 65 años, y en algunos documentos

aparecía como 65 y mas, y en otros correctamente escrito como 65 y más. También, se normalizaron los nombres de los departamentos para evitar inconsistencias causadas por mayúsculas, espacios y tildes.

Adicionalmente, al analizar datos estadísticos sobre los valores de algunos datasets, se decidió dejar una versión limpia para trabajar únicamente con datos relevantes para los análisis en donde explícitamente se conservaran datos a nivel geográfico de departamento en los datasets de departamento. En los datasets por rangos de edad, se eliminaron los registros clasificados como Ignorado, ya que no aportan información útil para el análisis demográfico.

Finalmente, los datos procesados fueron exportados a archivos CSV para facilitar su manipulación y análisis en R.

Análisis Exploratorio

Exploración de Variables Cuantitativas

```
summary(matrimonios_depto_limpio %>% pull(valor))
```

Estadística Descriptiva

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     8.0   135.0   204.0   285.7   355.0  2619.0
```

```
summary(divorcios_depto_limpio %>% pull(valor))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.00   13.00   22.34   19.00  410.00
```

```
summary(matrimonios_edad_limpio %>% pull(valor))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     5.0    62.0   566.7   282.0 14859.0
```

```
summary(divorcios_edad_limpio %>% pull(valor))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    2.00   24.10   15.25  767.00
```

Para el análisis de variables numéricas se decidió enfocar el estudio únicamente en la variable **valor**, ya que esta representa la cantidad de eventos registrados (matrimonios o divorcios) y es la única variable cuantitativa con significado estadístico. Las demás variables numéricas presentes en los datasets, como el año o el mes, cumplen una función temporal, pero no representan magnitudes susceptibles de análisis de tendencia central o dispersión en este contexto.

Adicionalmente, para evitar distorsiones en las medidas estadísticas, se trabajó únicamente con registros a nivel departamental y se excluyeron los totales nacionales. En el caso de los datasets por rangos de edad, se eliminaron los registros clasificados como *Ignorado*, ya que no aportan información útil para el análisis demográfico.

Bajo estos criterios, se obtuvieron los siguientes resultados descriptivos para la variable **valor**:

Matrimonios por departamento (mensuales)

- Los valores presentan un **mínimo de 8** y un **máximo de 2619 matrimonios**, con una **mediana de 204**. La **media (285.7)** es considerablemente mayor que la mediana, lo que indica una distribución asimétrica hacia la derecha, con la presencia de valores extremos altos. Esto sugiere que existen algunos departamentos con cantidades de matrimonios significativamente mayores al resto, mientras que la mayoría presenta valores más moderados. El máximo de matrimonios mensuales se dio en enero del 2012 en Guatemala.

Divorcios por departamento (mensuales)

- Los valores oscilan entre **0 y 410 divorcios**, con una **mediana de 13** y una **media de 22.34**. Nuevamente, la media supera a la mediana, evidenciando una **asimetría positiva**. Como dato curioso se buscó en el dataset quien era el 410, y fue Guatemala en junio del 2019. En comparación con los matrimonios, los divorcios ocurren en cantidades considerablemente menores, lo cual es coherente con la naturaleza del fenómeno.

Matrimonios por rangos de edad

- Se observa una gran dispersión en los valores, con un **máximo de 14,859 registros**, una **mediana de 62** y una **media de 566.7**. La diferencia marcada entre la mediana y la media confirma la presencia de **grupos de edad con una concentración muy alta de matrimonios**, mientras que otros rangos presentan valores bajos. Esto indica que los matrimonios no se distribuyen uniformemente entre los rangos establecidos por el INE, sino que se concentran en edades específicas. El tope de matrimonios se dio en parejas en donde tanto el novio y la novia estaban entre 20 y 24 años durante el 2021, cuando estuvimos en cuarentena por la mayor parte del tiempo.

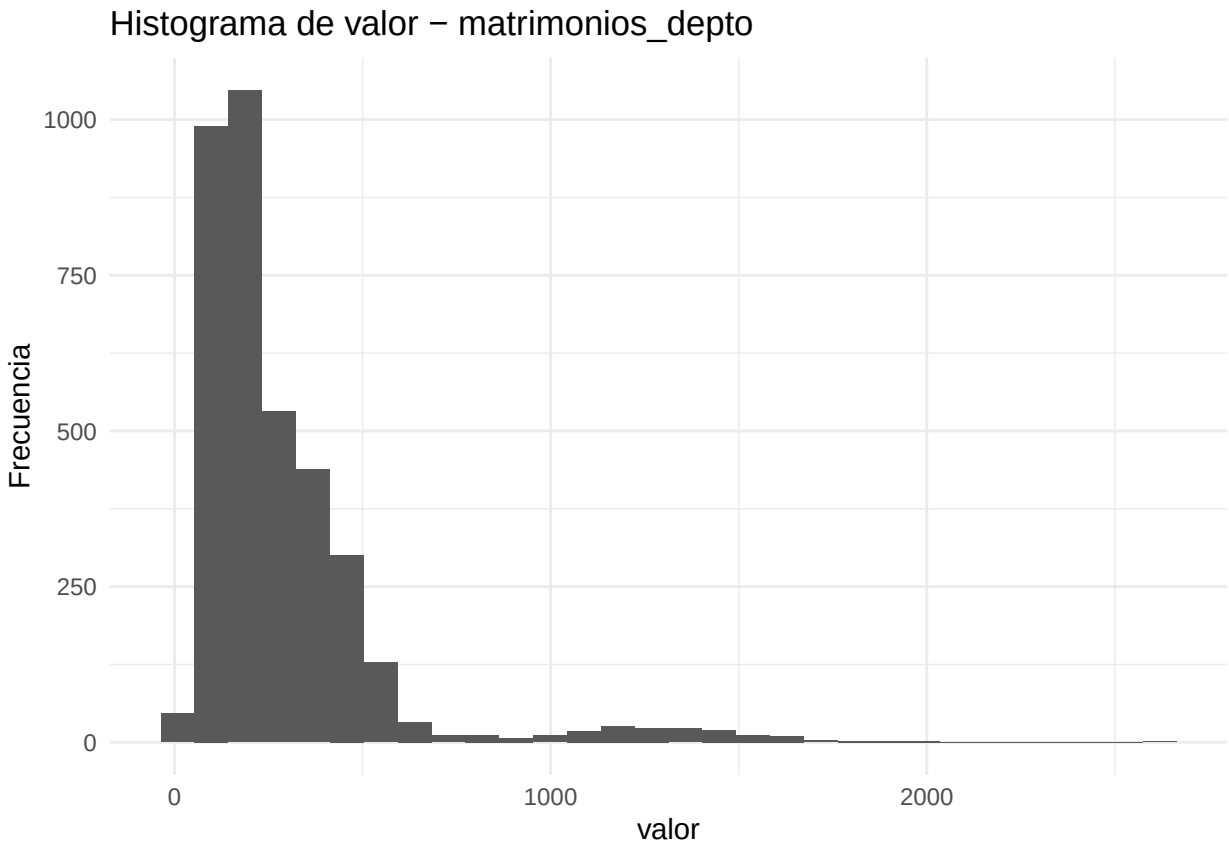
Divorcios por rangos de edad

- Los valores son más bajos en comparación con los matrimonios, con una **mediana de 2** y una **media de 24.1**, y un máximo de **767 divorcios**. La fuerte diferencia entre la media y la mediana refleja una **distribución altamente sesgada**, donde pocos rangos de edad concentran la mayor cantidad de divorcios. Es interesante ver que los 767 divorcios, se dieron con el hombre y mujer estando entre 30 y 34 años, en el año 2022, justo en el año en que estábamos empezando a salir de las restricciones por COVID.

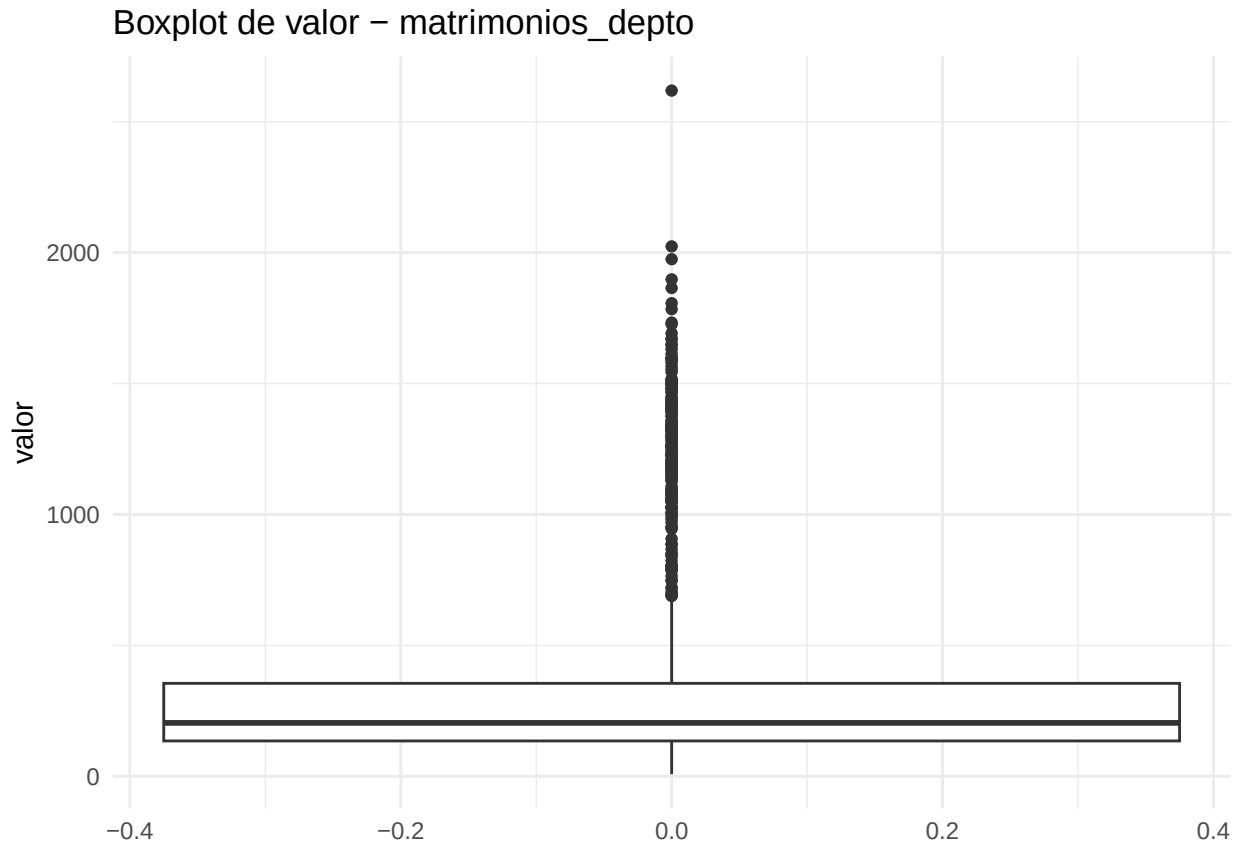
Distribuciones por variables

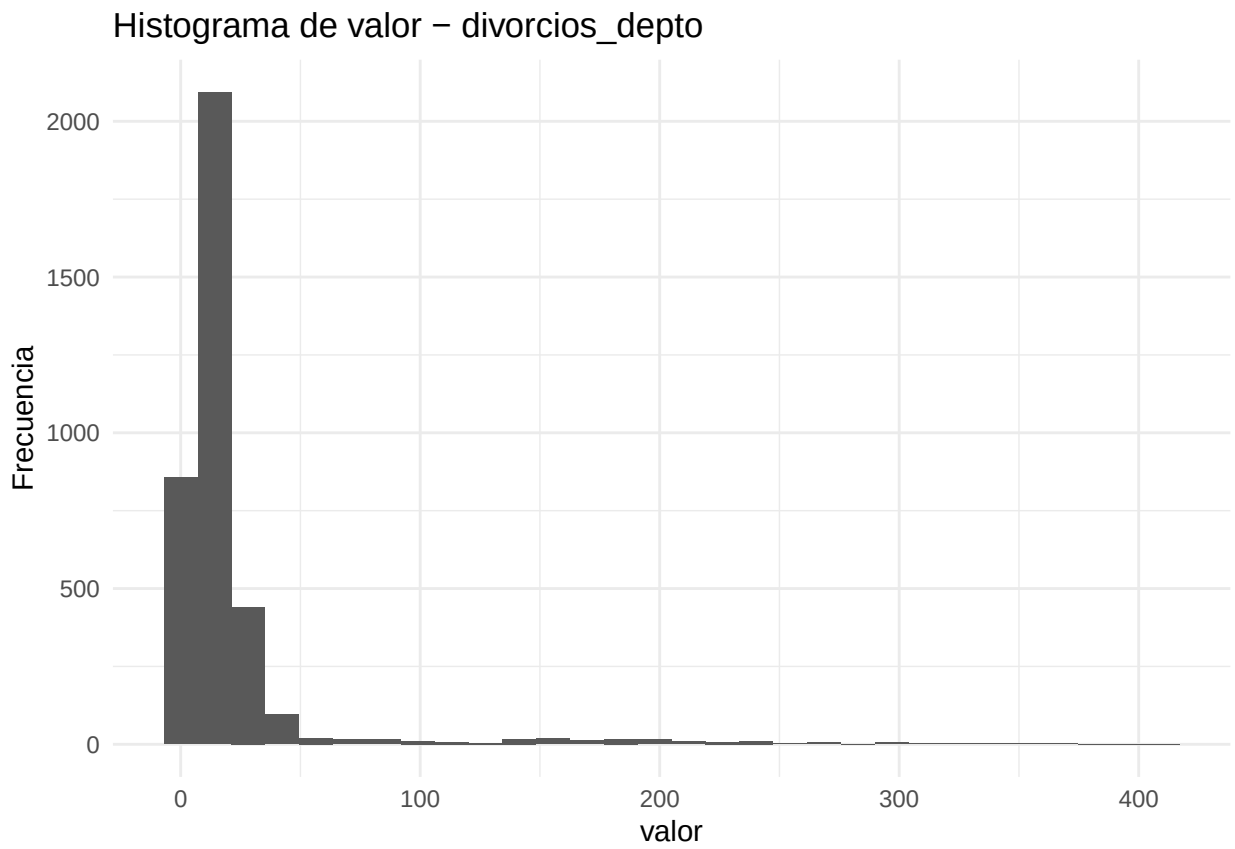
```
each_df <- list(
  matrimonios_depto = matrimonios_depto_limpio,
  divorcios_depto = divorcios_depto_limpio,
  matrimonios_edad = matrimonios_edad_limpio,
  divorcios_edad = divorcios_edad_limpio
```

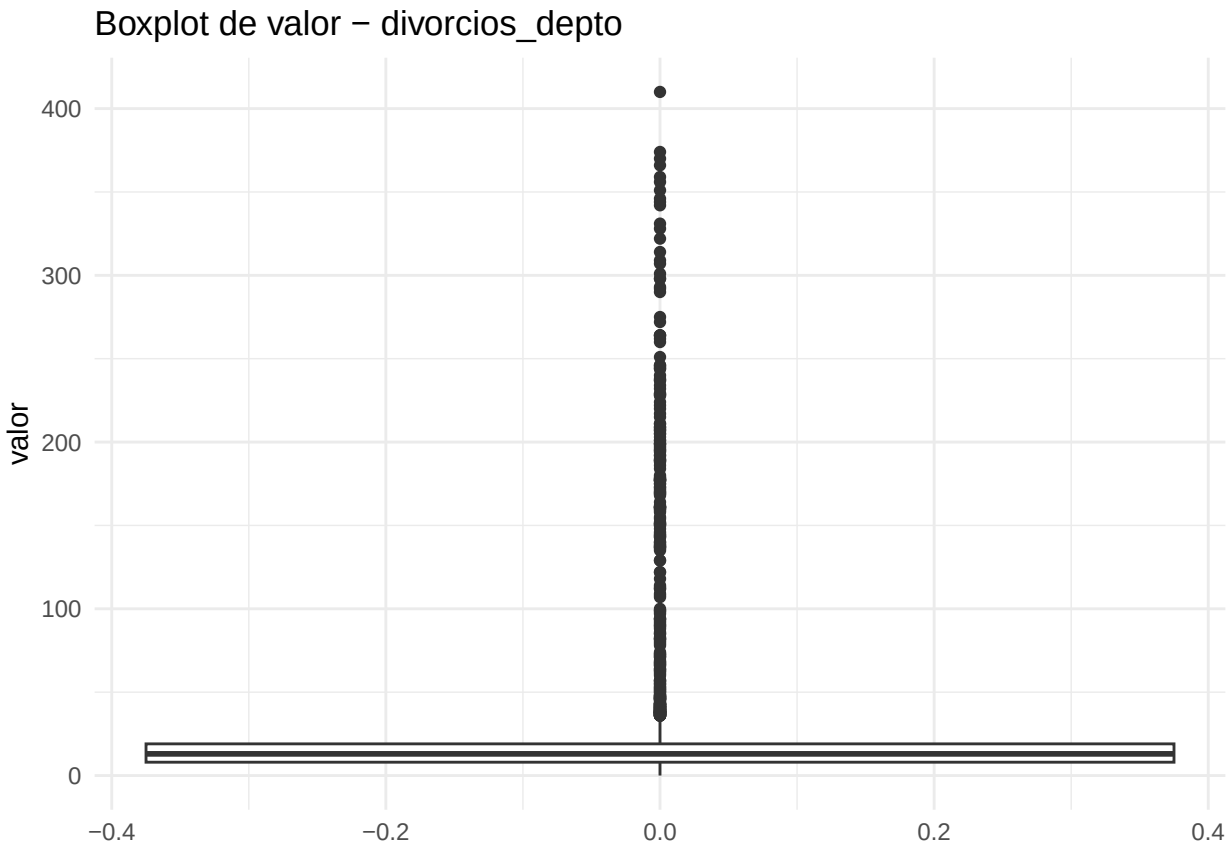
```
)

plot_valor_hist_box <- function(df, df_name, bins = 30) {
  df <- as.data.frame(df)

  p_hist <- ggplot(df, aes(x = valor)) +
    geom_histogram(bins = bins) +
    labs(
      title = paste("Histograma de valor -", df_name),
      x = "valor",
      y = "Frecuencia"
    ) +
    theme_minimal()

  p_box <- ggplot(df, aes(y = valor)) +
    geom_boxplot() +
    labs(
      title = paste("Boxplot de valor -", df_name),
      y = "valor"
    ) +
    theme_minimal()

  print(p_hist)
  print(p_box)
}

for (df_name in names(each_df)) {
  plot_valor_hist_box(each_df[[df_name]], df_name)
}
```

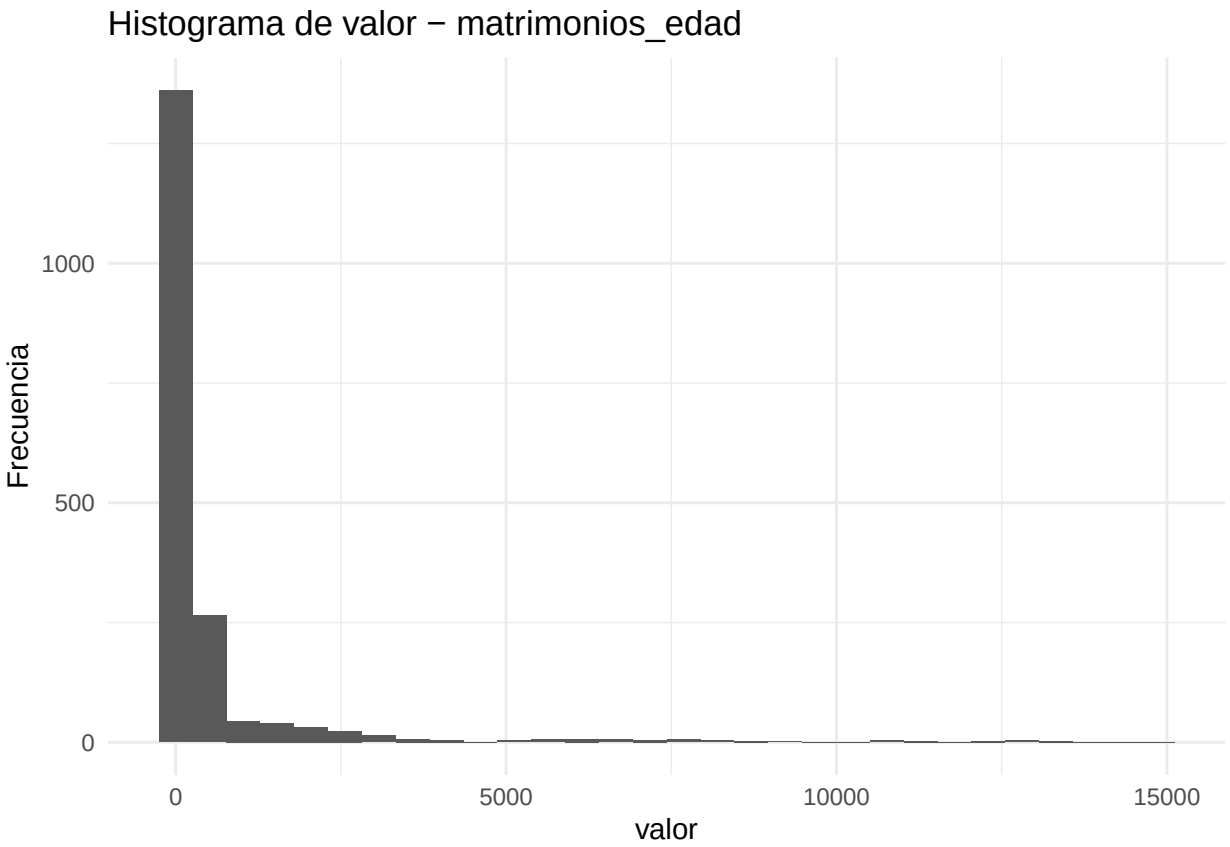


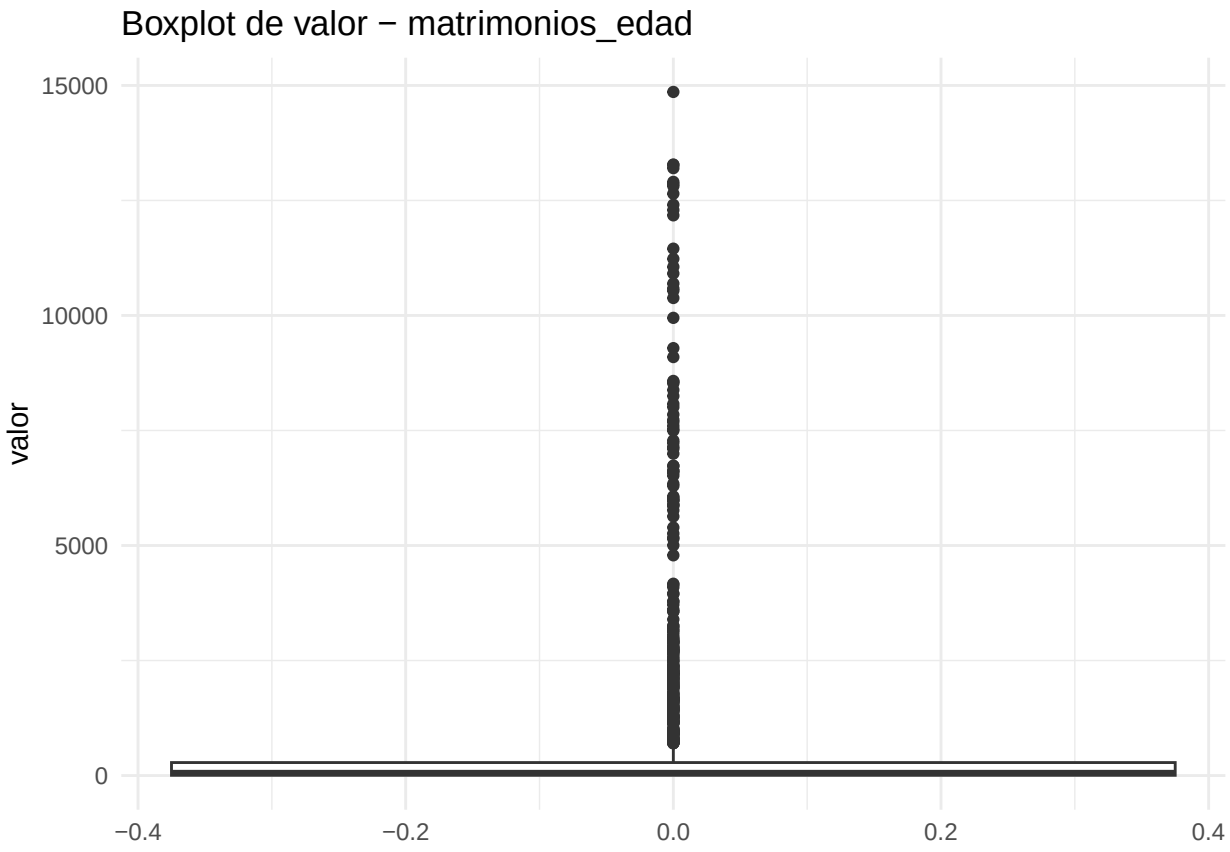
Gráficamente

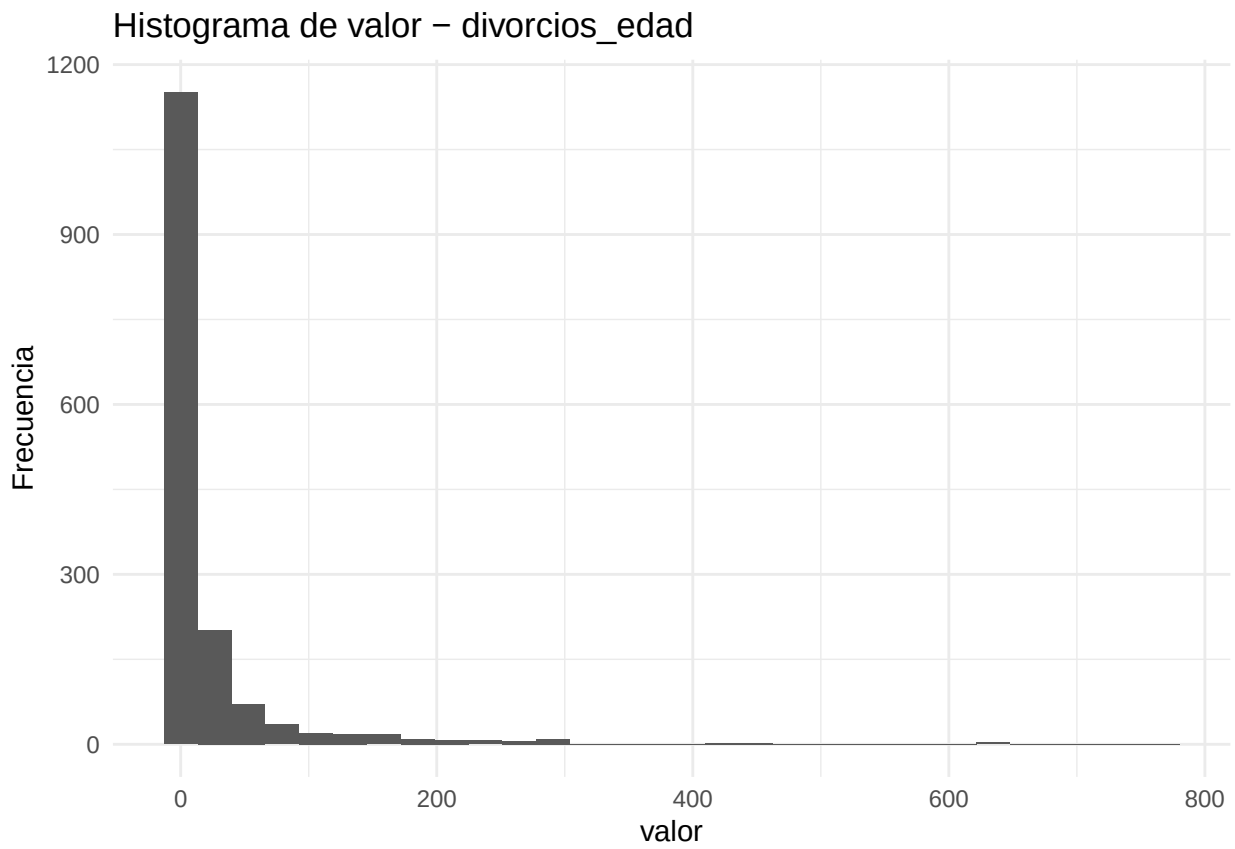


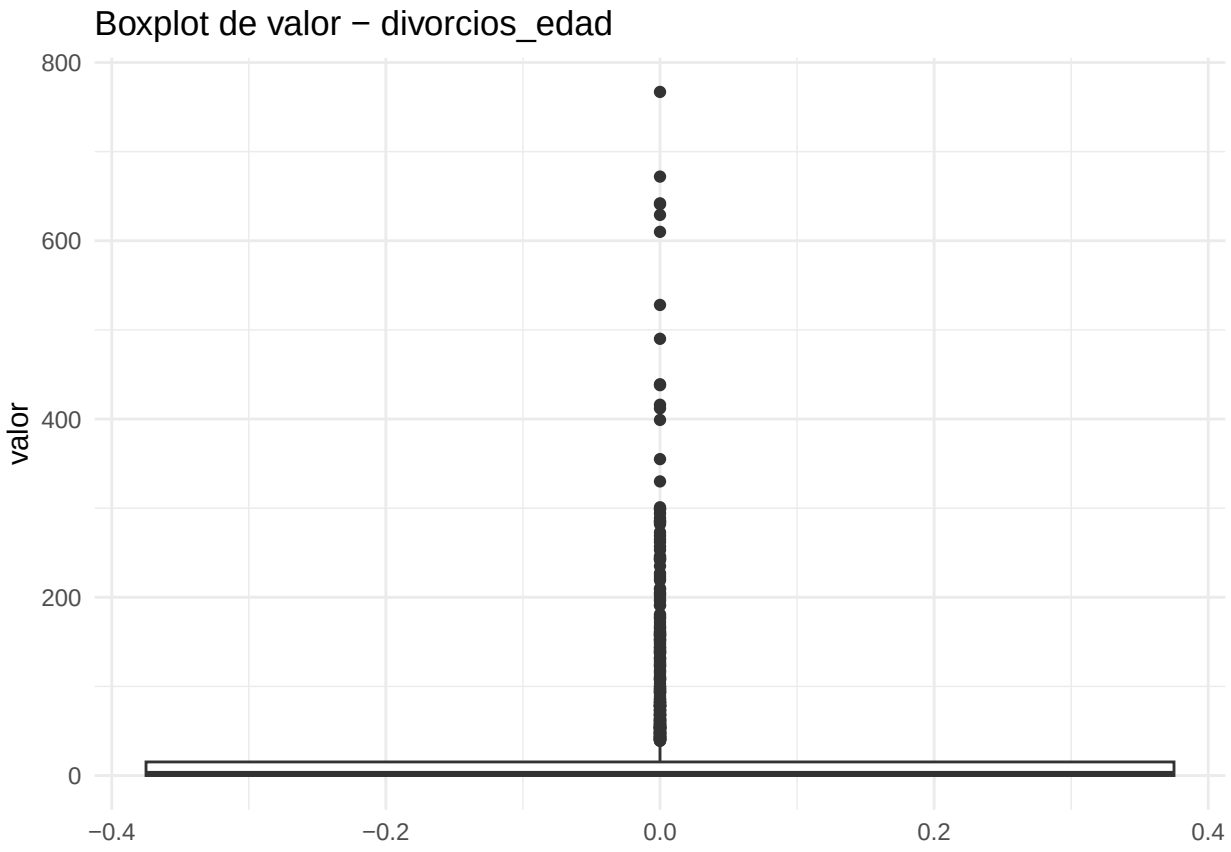

Histograma de valor − divorcios_depto
Frecuencia
2000
1500
1000
500
0
0
100
200
300
400
valor


Boxplot de valor − divorcios_depto
valor
400
300
200
100
0
−0.4
−0.2
0.0
0.2
0.4


Histograma de valor − matrimonios_edad
Frecuencia
1000
500
0
0
5000
10000
15000
valor


Boxplot de valor − matrimonios_edad
valor
15000
10000
5000
0
−0.4
−0.2
0.0
0.2
0.4


Histograma de valor − divorcios_edad
Frecuencia
1200
900
600
300
0
0
200
400
600
800
valor



Distribución de los datos (histogramas)

Luego de obtener medidas de tendencia central y dispersión para la variable valor en cada dataset, se generaron histogramas con el fin de comprender visualmente la forma de distribución de los datos. En los cuatro casos se observa una concentración fuerte hacia valores bajos, con una cola hacia la derecha. La mayoría de observaciones registran cantidades moderadas, mientras que un subconjunto reducido presenta valores mucho más altos. Por ejemplo, en matrimonios por departamento (mensual) una parte considerable de los datos se concentra aproximadamente entre 0 y 800, mientras que en divorcios por departamento (mensual) la mayor densidad se ubica entre 0 y 100. En los datasets por rangos de edad, la concentración hacia valores bajos también es evidente, lo cual sugiere que los conteos altos se concentran en grupos específicos de edad, mientras que muchos rangos (por ejemplo, edades muy bajas o muy altas) presentan valores pequeños o incluso cercanos a cero.

Valores atípicos (boxplots)

Los diagramas de cajas y bigotes refuerzan lo observado en los histogramas: en los cuatro datasets se identifican múltiples valores atípicos, lo cual indica la presencia de meses, departamentos o combinaciones de rangos de edad con conteos significativamente mayores al resto. En matrimonios por departamento (mensual) destacan varios puntos por encima de 800 y algunos cercanos al máximo del dataset; en divorcios por departamento (mensual) también se observan atípicos por encima de 120 y hasta valores cercanos a 400. En los datasets por edad, los outliers son todavía más notorios debido a la concentración en ciertos rangos; por ejemplo, en matrimonios por rangos de edad aparece un valor extremo cercano a 15,000, correspondiente a una combinación de edades altamente frecuente. En conjunto, estos resultados sugieren que los eventos no se distribuyen uniformemente entre departamentos o rangos de edad, sino que existen picos de alta concentración asociados a rangos bastatne comunes.

```r
prueba_lilliefors_valor <- function(df, df_name, alpha = 0.05) {
  x <- na.omit(df$valor)
  test <- lillie.test(x)

  decision <- ifelse(
    test$p.value < alpha,
    "Se rechaza normalidad",
    "No se rechaza normalidad"
  )

  tibble(
    dataset = df_name,
    n = length(x),
    p_value = test$p.value,
    conclusion = decision
  )
}

resultados_normalidad <- bind_rows(
  lapply(names(each_df), function(nm) {
    prueba_lilliefors_valor(each_df[[nm]], nm)
  })
)

resultados_normalidad
```

Prueba de Normalidad

```
## # A tibble: 4 x 4
##   dataset              n   p_value conclusion
##   <chr>            <int>     <dbl> <chr>
## 1 matrimonios_depto  3696 7.41e-323 Se rechaza normalidad
## 2 divorcios_depto    3696 0         Se rechaza normalidad
## 3 matrimonios_edad   1855 0         Se rechaza normalidad
## 4 divorcios_edad     1568 0         Se rechaza normalidad
```

Para corroborar lo observado gráficamente (asimetría positiva y presencia de valores atípicos), se aplicó la prueba de normalidad de Lilliefors sobre la variable valor en cada dataset. En todos los casos se obtuvo un p-value < 0.05, por lo que se rechaza la hipótesis de normalidad.

Exploración de Variables Categóricas

En esta sección se analizan las variables categóricas mediante tablas de frecuencia y proporciones. En nuestros datasets, la variable valor representa el conteo de eventos (matrimonios o divorcios) asociado a una categoría específica: ya sea un departamento (en los datasets geográficos) o un rango de edad (en los datasets demográficos). Por lo tanto, tanto los totales como las proporciones describen la concentración relativa de eventos en cada categoría. Es decir, una proporción alta indica que una categoría (por ejemplo, un departamento o un rango de edad) concentra una parte mayor de los registros totales del periodo analizado.

```r
matr_depto_cat <- matrimonios_depto %>%
  filter(nivel_geo == "departamento", is.na(mes))

div_depto_cat <- divorcios_depto %>%
  filter(nivel_geo == "departamento", is.na(mes))

tabla_depto_matr <- matr_depto_cat %>%
  group_by(departamento) %>%
  summarise(total = sum(valor, na.rm = TRUE), .groups = "drop") %>%
  mutate(prop = total / sum(total)) %>%
  arrange(desc(total))

tabla_depto_div <- div_depto_cat %>%
  group_by(departamento) %>%
  summarise(total = sum(valor, na.rm = TRUE), .groups = "drop") %>%
  mutate(prop = total / sum(total)) %>%
  arrange(desc(total))

tabla_depto_matr
```

Análisis de Datasets de Departamentos

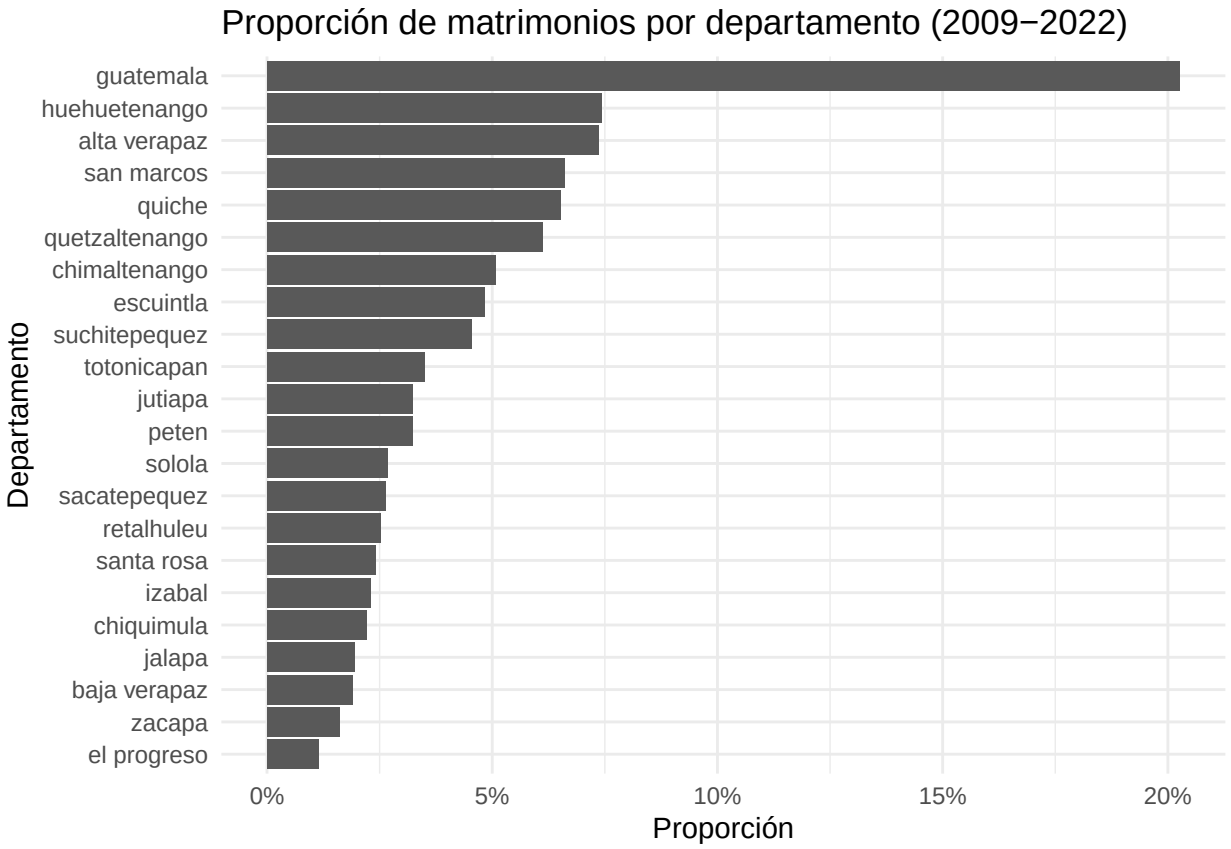
```
## # A tibble: 22 x 3
##    departamento    total   prop
##    <chr>           <dbl>  <dbl>
##  1 guatemala      213948 0.203
##  2 huehuetenango   78339 0.0742
##  3 alta verapaz    77626 0.0735
##  4 san marcos      69837 0.0661
##  5 quiche          68741 0.0651
##  6 quetzaltenango  64494 0.0611
##  7 chimaltenango   53487 0.0507
##  8 escuintla       50966 0.0483
##  9 suchitepequez   47873 0.0453
## 10 totonicapan     37006 0.0350
## # i 12 more rows
```

```r
tabla_depto_div
```

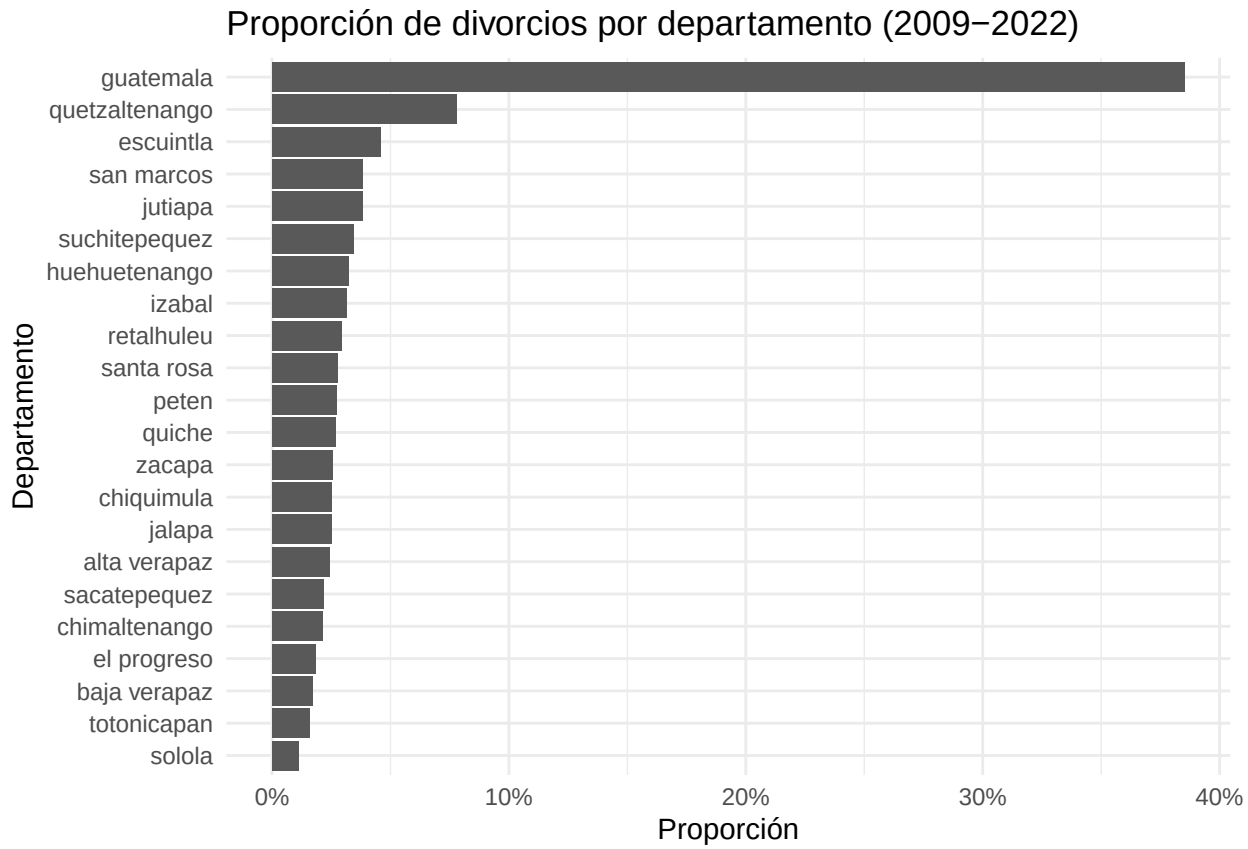
```
## # A tibble: 22 x 3
##    departamento   total   prop
##    <chr>          <dbl>  <dbl>
##  1 guatemala      31813 0.385
##  2 quetzaltenango  6423 0.0778
##  3 escuintla       3768 0.0456
##  4 san marcos      3155 0.0382
##  5 jutiapa         3145 0.0381
##  6 suchitepequez   2826 0.0342
##  7 huehuetenango   2683 0.0325
##  8 izabal          2597 0.0315
##  9 retalhuleu      2436 0.0295
```

```
## 10 santa rosa      2295 0.0278
## # i 12 more rows
```

```r
ggplot(tabla_depto_matr, aes(x = reorder(departamento, prop), y = prop)) +
  geom_col() +
  coord_flip() +
  scale_y_continuous(labels = scales::percent) +
  labs(title = "Proporción de matrimonios por departamento (2009-2022)",
       x = "Departamento", y = "Proporción") +
  theme_minimal()
```




```r
ggplot(tabla_depto_div, aes(x = reorder(departamento, prop), y = prop)) +
  geom_col() +
  coord_flip() +
  scale_y_continuous(labels = scales::percent) +
  labs(title = "Proporción de divorcios por departamento (2009-2022)",
       x = "Departamento", y = "Proporción") +
  theme_minimal()
```



Al analizar la distribución de matrimonios por departamento (2009–2022), se observa una marcada concentración en pocos departamentos. Los tres departamentos con mayor presencia son Guatemala (20.26%), Huehuetenango (7.42%) y Alta Verapaz (7.35%). La diferencia entre Guatemala y el resto es notable: incluso sumando Huehuetenango y Alta Verapaz, no se alcanza la proporción que aporta Guatemala por sí sola. Esto sugiere que Guatemala concentra una fracción muy importante de los matrimonios registrados durante el periodo.

En el caso de divorcios por departamento, la concentración es todavía más marcada en Guatemala. Guatemala registra 31,813 divorcios, lo que corresponde al 38.53% del total, seguido por Quetzaltenango con 6,423 (7.78%) y Escuintla con 3,768 (4.56%). Nuevamente, Guatemala sobresale ampliamente y domina la distribución. Esta diferencia también se aprecia claramente en los gráficos: la barra de Guatemala es muy superior a las demás.

Análisis de Datasets de Rangos de Edad Para el análisis por edades, se construyeron cuatro tablas:

1. matrimonios por rango de edad de mujeres,
2. divorcios por rango de edad de mujeres,
3. matrimonios por rango de edad de hombres,
4. divorcios por rango de edad de hombres.

```
matr_edad_cat <- matrimonios_edad %>%
  filter(edad_mujer_grupo != "Ignorado", edad_hombre_grupo != "Ignorado")

div_edad_cat <- divorcios_edad %>%
  filter(edad_mujer_grupo != "Ignorado", edad_hombre_grupo != "Ignorado")
```

```r
# Mujer
tabla_matr_mujer <- matr_edad_cat %>%
  group_by(edad_mujer_grupo) %>%
  summarise(total = sum(valor, na.rm = TRUE), .groups = "drop") %>%
  mutate(prop = total / sum(total)) %>%
  arrange(desc(total))

tabla_div_mujer <- div_edad_cat %>%
  group_by(edad_mujer_grupo) %>%
  summarise(total = sum(valor, na.rm = TRUE), .groups = "drop") %>%
  mutate(prop = total / sum(total)) %>%
  arrange(desc(total))

# Hombre
tabla_matr_hombre <- matr_edad_cat %>%
  group_by(edad_hombre_grupo) %>%
  summarise(total = sum(valor, na.rm = TRUE), .groups = "drop") %>%
  mutate(prop = total / sum(total)) %>%
  arrange(desc(total))

tabla_div_hombre <- div_edad_cat %>%
  group_by(edad_hombre_grupo) %>%
  summarise(total = sum(valor, na.rm = TRUE), .groups = "drop") %>%
  mutate(prop = total / sum(total)) %>%
  arrange(desc(total))

tabla_matr_mujer
```

```
## # A tibble: 14 x 3
##    edad_mujer_grupo  total    prop
##    <chr>             <dbl>   <dbl>
##  1 20 - 24          352716 0.336
##  2 25 - 29          207469 0.197
##  3 15 - 19          177653 0.169
##  4 30 - 34          105866 0.101
##  5 35 - 39           55252 0.0526
##  6 40 - 44           32255 0.0307
##  7 Menos de 20       32206 0.0306
##  8 45 - 49           22377 0.0213
##  9 50 - 54           15903 0.0151
## 10 18 - 19           12810 0.0122
## 11 55 - 59           11564 0.0110
## 12 65 y más          10597 0.0101
## 13 60 - 64            8035 0.00764
## 14 Menos de 15        6594 0.00627
```

```r
tabla_div_mujer
```

```
## # A tibble: 13 x 3
##    edad_mujer_grupo total     prop
##    <chr>            <dbl>    <dbl>
##  1 25 - 29           9781 0.259
##  2 30 - 34           8235 0.218
```

```
##  3 20 - 24           6367 0.168
##  4 35 - 39           5203 0.138
##  5 40 - 44           3175 0.0840
##  6 45 - 49           1748 0.0463
##  7 15 - 19           1312 0.0347
##  8 50 - 54            940 0.0249
##  9 55 - 59            526 0.0139
## 10 60 y más           399 0.0106
## 11 Menos de 15         47 0.00124
## 12 18 - 19             39 0.00103
## 13 Menos de 20         16 0.000423
```

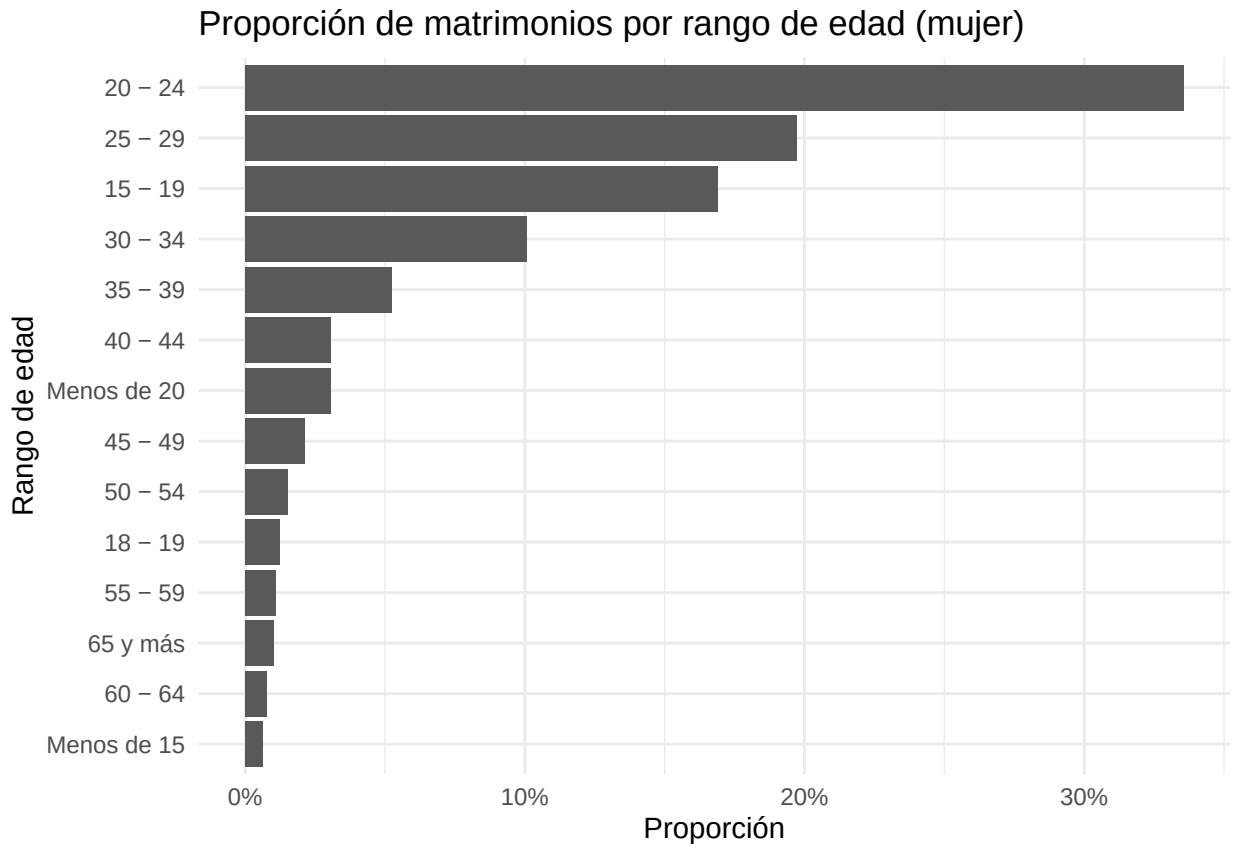
```
tabla_matr_hombre
```

```
## # A tibble: 14 x 3
##    edad_hombre_grupo  total      prop
##    <chr>              <dbl>     <dbl>
##  1 20-24             342479 0.326
##  2 25-29             260128 0.247
##  3 30-34             141959 0.135
##  4 15-19              74117 0.0705
##  5 35-39              73538 0.0699
##  6 40-44              41818 0.0398
##  7 45-49              27244 0.0259
##  8 65 y más           23194 0.0221
##  9 50-54              20770 0.0198
## 10 55-59              16945 0.0161
## 11 60-64              14591 0.0139
## 12 Menos de 20        10177 0.00968
## 13 18-19               4285 0.00408
## 14 Menos de 15           52 0.0000495
```

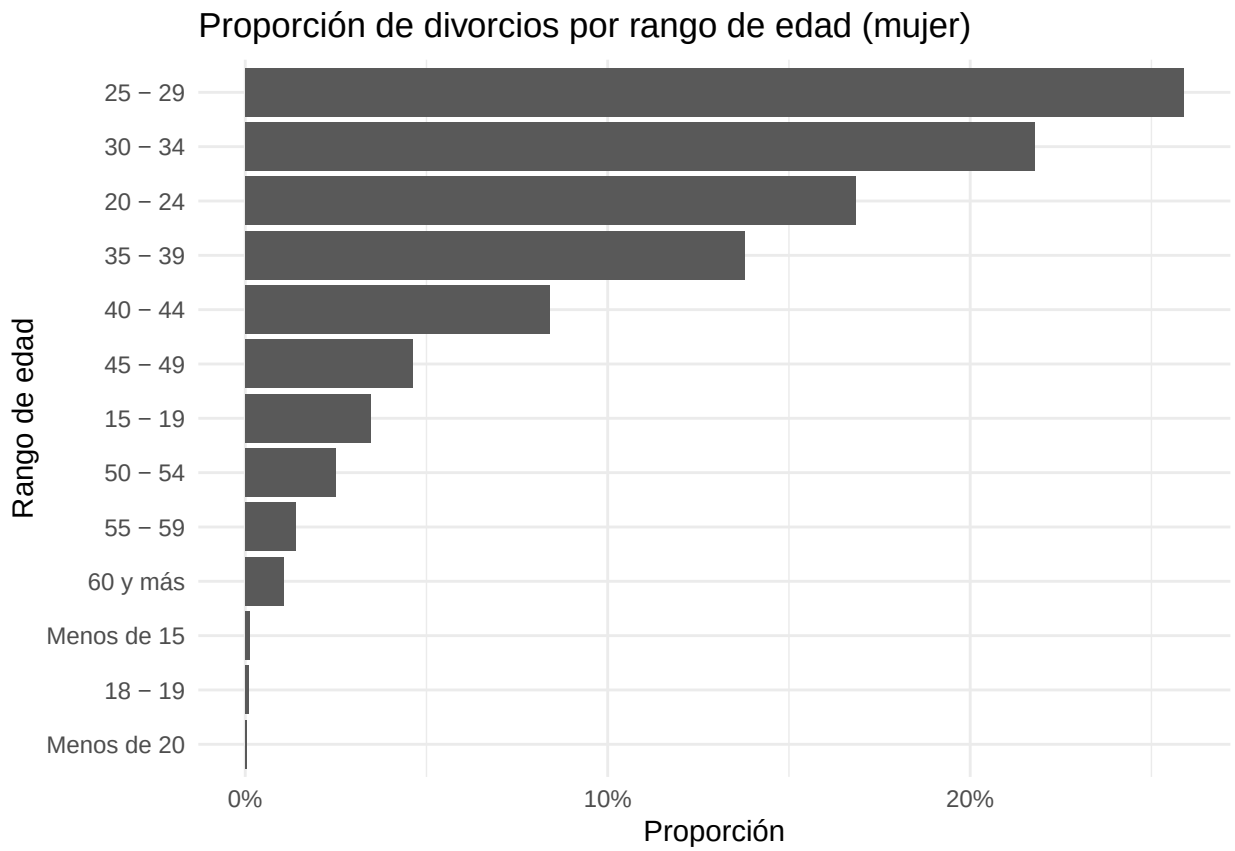
```
tabla_div_hombre
```

```
## # A tibble: 13 x 3
##    edad_hombre_grupo total      prop
##    <chr>             <dbl>     <dbl>
##  1 30-34              9245 0.245
##  2 25-29              8416 0.223
##  3 35-39              6439 0.170
##  4 40-44              4084 0.108
##  5 20-24              3386 0.0896
##  6 45-49              2461 0.0651
##  7 50-54              1439 0.0381
##  8 60 y más           1050 0.0278
##  9 55-59               901 0.0238
## 10 15-19               362 0.00958
## 11 Menos de 20           4 0.000106
## 12 18-19                 1 0.0000265
## 13 Menos de 15           0 0
```

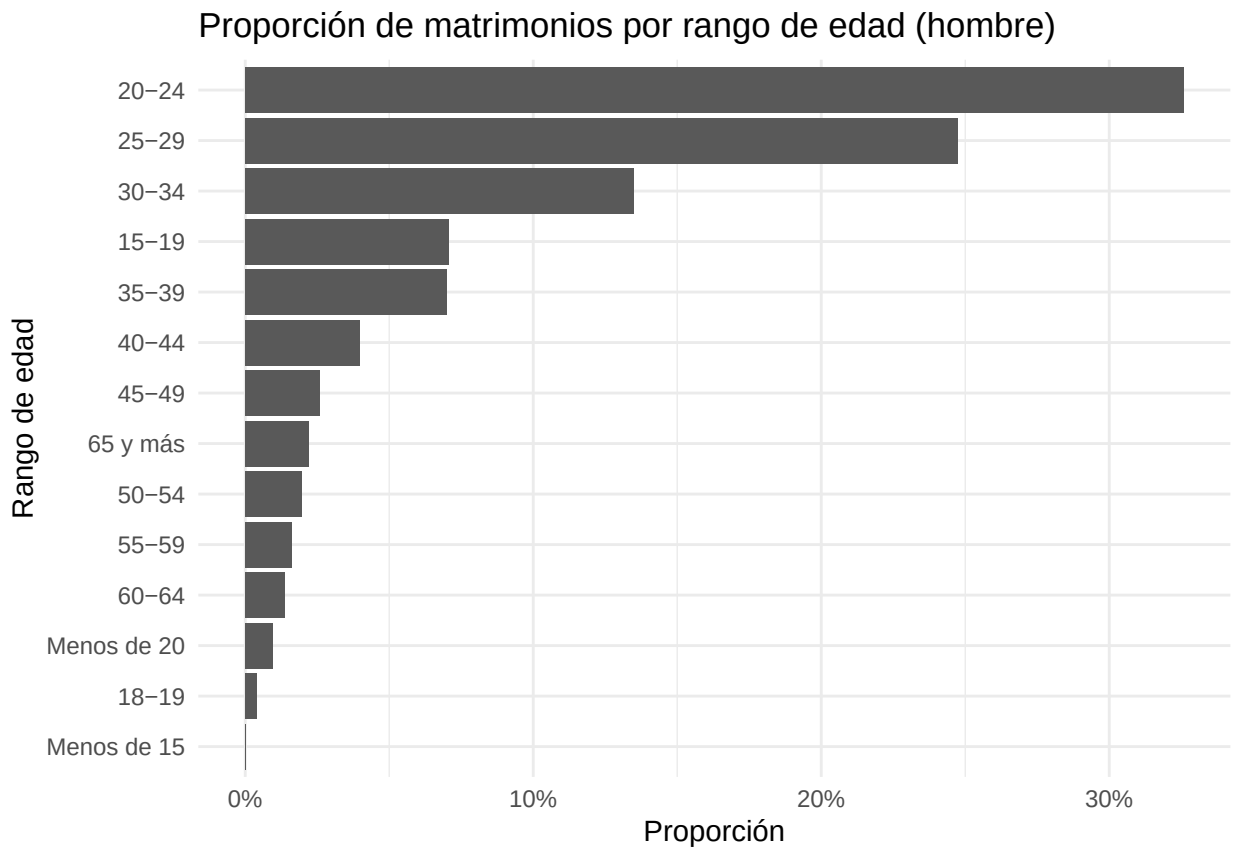
```r
ggplot(tabla_matr_mujer, aes(x = reorder(edad_mujer_grupo, prop), y = prop)) +
  geom_col() + coord_flip() +
  scale_y_continuous(labels = scales::percent) +
  labs(title = "Proporción de matrimonios por rango de edad (mujer)",
       x = "Rango de edad", y = "Proporción") +
  theme_minimal()
```



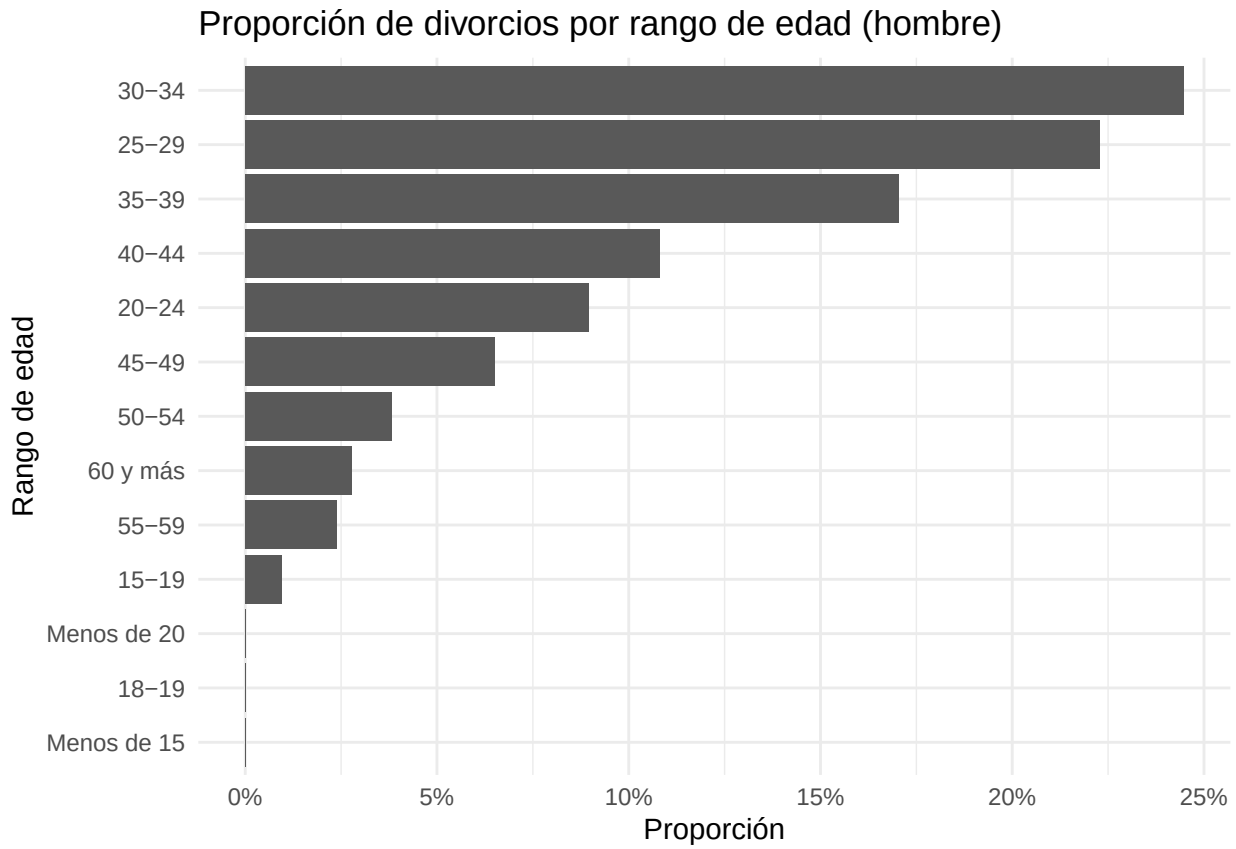

```r
ggplot(tabla_div_mujer, aes(x = reorder(edad_mujer_grupo, prop), y = prop)) +
  geom_col() + coord_flip() +
  scale_y_continuous(labels = scales::percent) +
  labs(title = "Proporción de divorcios por rango de edad (mujer)",
       x = "Rango de edad", y = "Proporción") +
  theme_minimal()
```




```
ggplot(tabla_matr_hombre, aes(x = reorder(edad_hombre_grupo, prop), y = prop)) +
  geom_col() + coord_flip() +
  scale_y_continuous(labels = scales::percent) +
  labs(title = "Proporción de matrimonios por rango de edad (hombre)",
       x = "Rango de edad", y = "Proporción") +
  theme_minimal()
```




```
ggplot(tabla_div_hombre, aes(x = reorder(edad_hombre_grupo, prop), y = prop)) +
  geom_col() + coord_flip() +
  scale_y_continuous(labels = scales::percent) +
  labs(title = "Proporción de divorcios por rango de edad (hombre)",
       x = "Rango de edad", y = "Proporción") +
  theme_minimal()
```



Rangos de edad en mujeres

En matrimonios, el rango más frecuente para mujeres es 20–24 años, con 352,716 registros (33.55%). Le siguen 25–29 años con 207,469 (19.73%), y 15–19 años con 177,653 (16.90%). Esto muestra una concentración fuerte en edades jóvenes-adultas, particularmente en el rango de 20–24 años.

En divorcios, la distribución se desplaza hacia edades mayores: el rango más frecuente es 25–29 años con 9,781 (25.88%), seguido de 30–34 años con 8,235 (21.79%), y luego 20–24 años con 6,367 (16.85%). En comparación con matrimonios, los divorcios aparecen con mayor frecuencia en rangos que corresponden a edades ligeramente más avanzadas.

Rangos de edad en hombres

En matrimonios, el rango más frecuente para hombres también es 20–24 años, con 342,479 (32.58%), seguido por 25–29 años con 260,128 (24.74%), y en tercer lugar 30–34 años con 141,959 (13.50%). La concentración se mantiene en el periodo de adultez temprana, aunque se nota un mayor peso relativo en rangos ligeramente superiores en comparación con mujeres (por ejemplo, 30–34 aparece en el top 3).

En divorcios, el pico principal está en edades más altas: 30–34 años con 9,245 (24.47%), seguido por 25–29 años con 8,416 (22.27%), y luego 35–39 años con 6,439 (17.04%). Esto refuerza la idea de que los divorcios tienden a concentrarse en rangos posteriores a los que concentran matrimonios.

Finalmente, los gráficos permiten detectar un aspecto relevante: en la distribución de matrimonios por edad de mujeres aparecen registros en el grupo “Menos de 15”, mientras que en los hombres no aparecen registros menores de 15. Esto es alarmante, porque se evidencia la posible presencia de matrimonios con mujeres menores de edad con hombres mayores de edad.

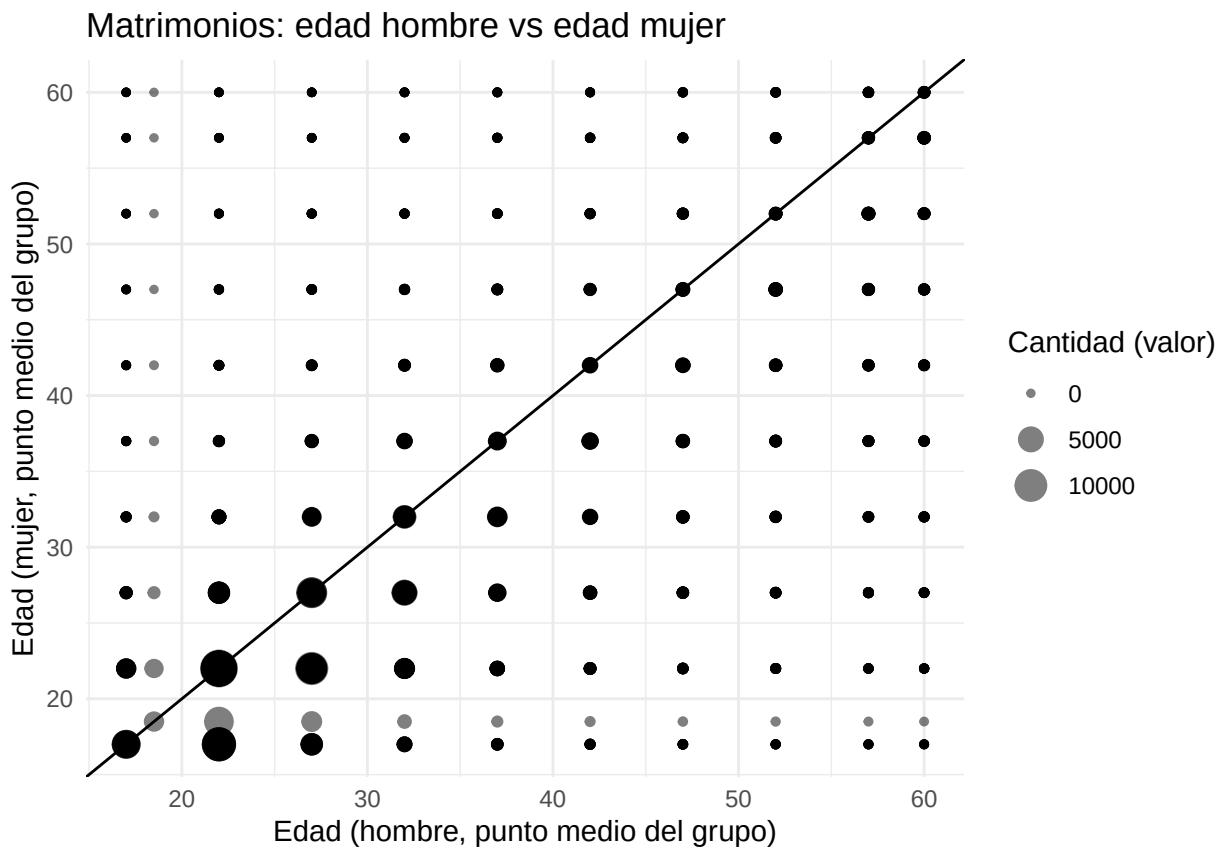
Relaciones entre Variables

```
matrimonios_depto_clean <- filter(matrimonios_depto, nivel_geo == "departamento" & !is.na(mes))
divorcios_depto_clean <- filter(divorcios_depto, nivel_geo == "departamento" & !is.na(mes))

to_mid_age <- function(x){
  x <- str_trim(x)
  if (x %in% c("Ignorado", "Menos de 15")) return(NA_real_)
  if (str_detect(x, "60")) return(60) # "60 y más"
  nums <- str_extract_all(x, "\\d+")[[1]]
  if (length(nums) == 2) return(mean(as.numeric(nums)))
  NA_real_
}
```

1. Relacion entre edades de mujeres y hombres en la decision de casarse.

- Al tratarse de datos con valores comunes, no se puede utilizar un gráfico de dispersión "común y corriente", así que se decidió usar uno de burbujas (tamaño proporcional al conteo). En este gráfico podemos observar que las burbujas más grandes no se encuentran sobre la diagonal de relación, sino más bien en la parte inferior del diagrama, lo cual sugiere que no hay una relación fuerte entre la edad del hombre y la de la mujer al momento de casarse.

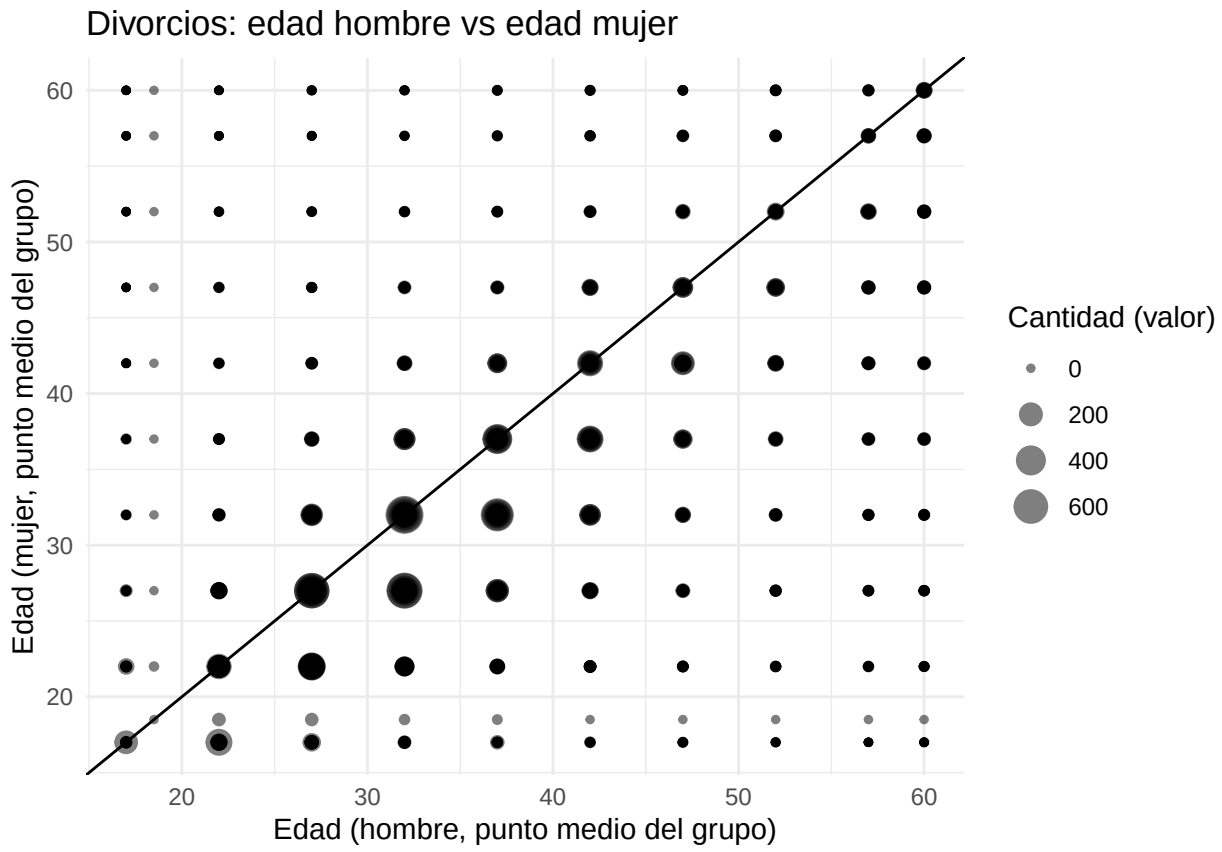
```
matrimonios_edad_plot <- matrimonios_edad %>%
  mutate(
    edad_h = sapply(edad_hombre_grupo, to_mid_age),
    edad_m = sapply(edad_mujer_grupo, to_mid_age)
  ) %>%
  filter(!is.na(edad_h), !is.na(edad_m))

ggplot(matrimonios_edad_plot, aes(x = edad_h, y = edad_m)) +
  geom_abline(slope = 1, intercept = 0) +
  geom_point(aes(size = valor), alpha = 0.5) +
  labs(
    title = "Matrimonios: edad hombre vs edad mujer",
    x = "Edad (hombre, punto medio del grupo)",
    y = "Edad (mujer, punto medio del grupo)",
    size = "Cantidad (valor)"
  ) +
  theme_minimal()
```



2. Relacion entre edades de mujeres y hombres en la decision de divorciarse

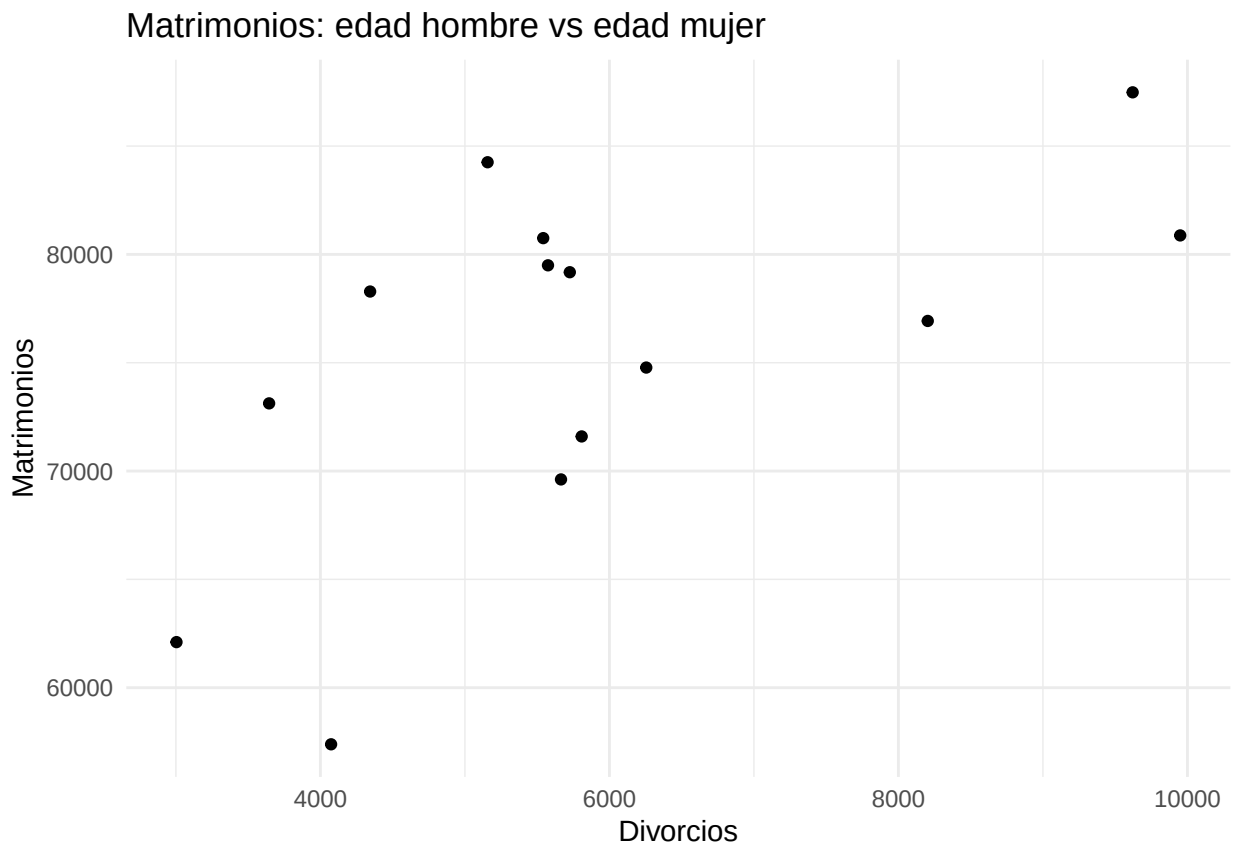
- Al tratarse de datos con valores comunes, no se puede utilizar un gráfico de dispersión "común y corriente", así que se decidió usar uno de burbujas (tamaño proporcional al conteo). En este gráfico podemos observar que las burbujas más grandes sí se encuentran sobre la diagonal de relación, aunque con cierta dispersión, lo cual sugiere que existe una relación moderada entre la edad del hombre y la de la mujer al momento de divorciarse.

```
divorcios_edad_plot <- divorcios_edad %>%
  mutate(
    edad_h = sapply(edad_hombre_grupo, to_mid_age),
    edad_m = sapply(edad_mujer_grupo, to_mid_age)
  ) %>%
  filter(!is.na(edad_h), !is.na(edad_m))

ggplot(divorcios_edad_plot, aes(x = edad_h, y = edad_m)) +
  geom_abline(slope = 1, intercept = 0) +
  geom_point(aes(size = valor), alpha = 0.5) +
  labs(
    title = "Divorcios: edad hombre vs edad mujer",
    x = "Edad (hombre, punto medio del grupo)",
    y = "Edad (mujer, punto medio del grupo)",
    size = "Cantidad (valor)"
  ) +
  theme_minimal()
```



3. La relación entre la cantidad de matrimonios y la cantidad de divorcios es de 0.603 (coeficiente de correlación de Pearson), lo cual indica una relación moderada. En general, a mayor tasa de matrimonios, mayor tasa de divorcios, aunque con variabilidad considerable.

```
number_of_divorcios_por_anio <- divorcios_depto_clean |>
  group_by(anio) |>
  summarise(total_divorcios = sum(valor))

number_of_matrimonios_por_anio <- matrimonios_depto_clean |>
  group_by(anio) |>
  summarise(total_divorcios = sum(valor))

matrimonios_vs_divorcios <- number_of_matrimonios_por_anio |>
  inner_join(number_of_divorcios_por_anio, by = c("anio"))
colnames(matrimonios_vs_divorcios) <- c("anio", "total_matrimonios", "total_divorcios")

ggplot(matrimonios_vs_divorcios, aes(x = total_divorcios, y = total_matrimonios)) +
  geom_point(na.rm = TRUE) +
  labs(
    title = "Matrimonios: edad hombre vs edad mujer",
    x = "Divorcios",
    y = "Matrimonios"
  ) +
  theme_minimal()
```




```
correlation_coefficient <- cor(matrimonios_vs_divorcios$total_matrimonios, matrimonios_vs_divorcios$tota
print(paste("Coefficiente de Correlacion entre Matrimonios and Divorcios:", correlation_coefficient))
```

```
## [1] "Coefficiente de Correlacion entre Matrimonios and Divorcios: 0.603599487196953"
```

```
print(paste("Porcentaje de :", correlation_coefficient))
```

```
## [1] "Porcentaje de : 0.603599487196953"
```

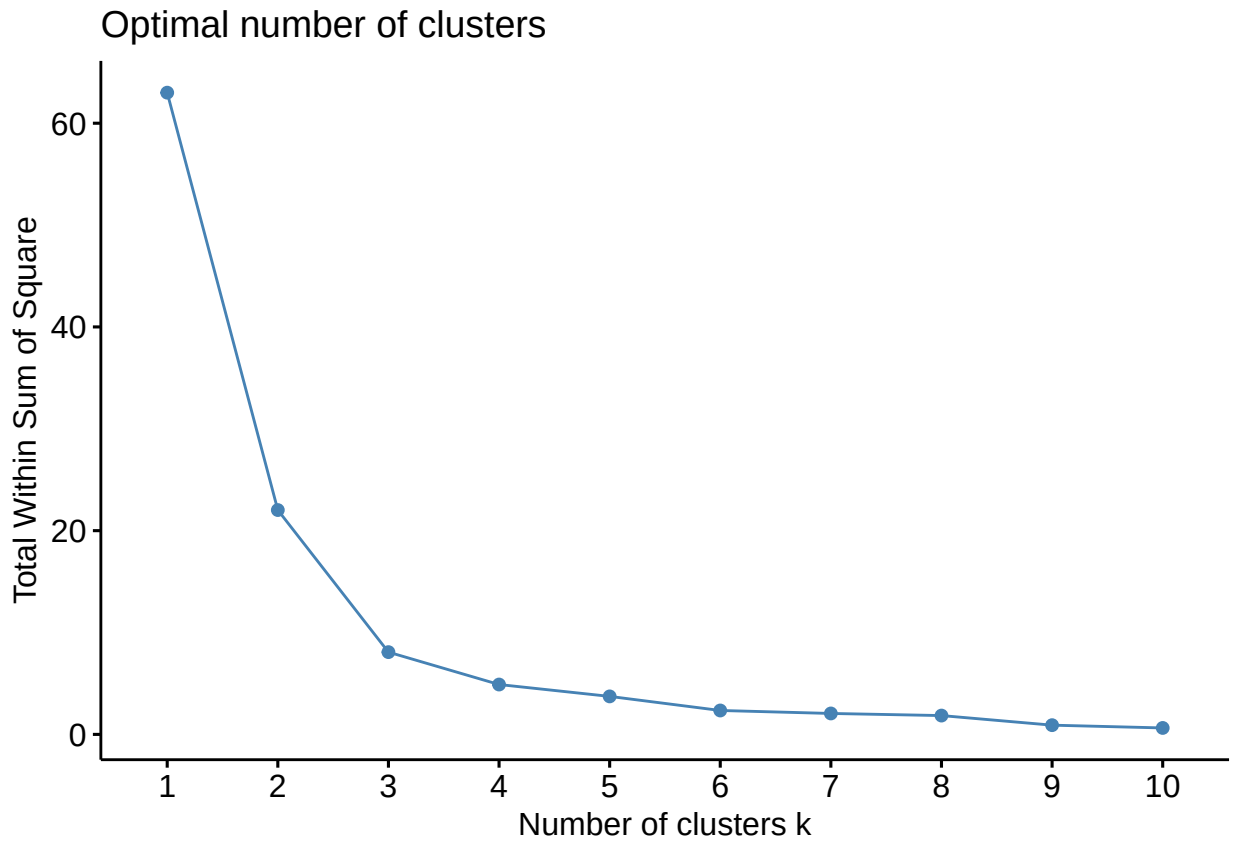
Clustering

Número de Clusters

```
matrimonios_anual <- matrimonios_depto %>%
  filter(
    nivel_geo == "departamento",
    is.na(mes),
    !is.na(departamento)
  ) %>%
  mutate(departamento = norm_depto(departamento)) %>%
  group_by(anio, departamento) %>%
  summarise(matrimonios = sum(valor, na.rm = TRUE), .groups = "drop")
```

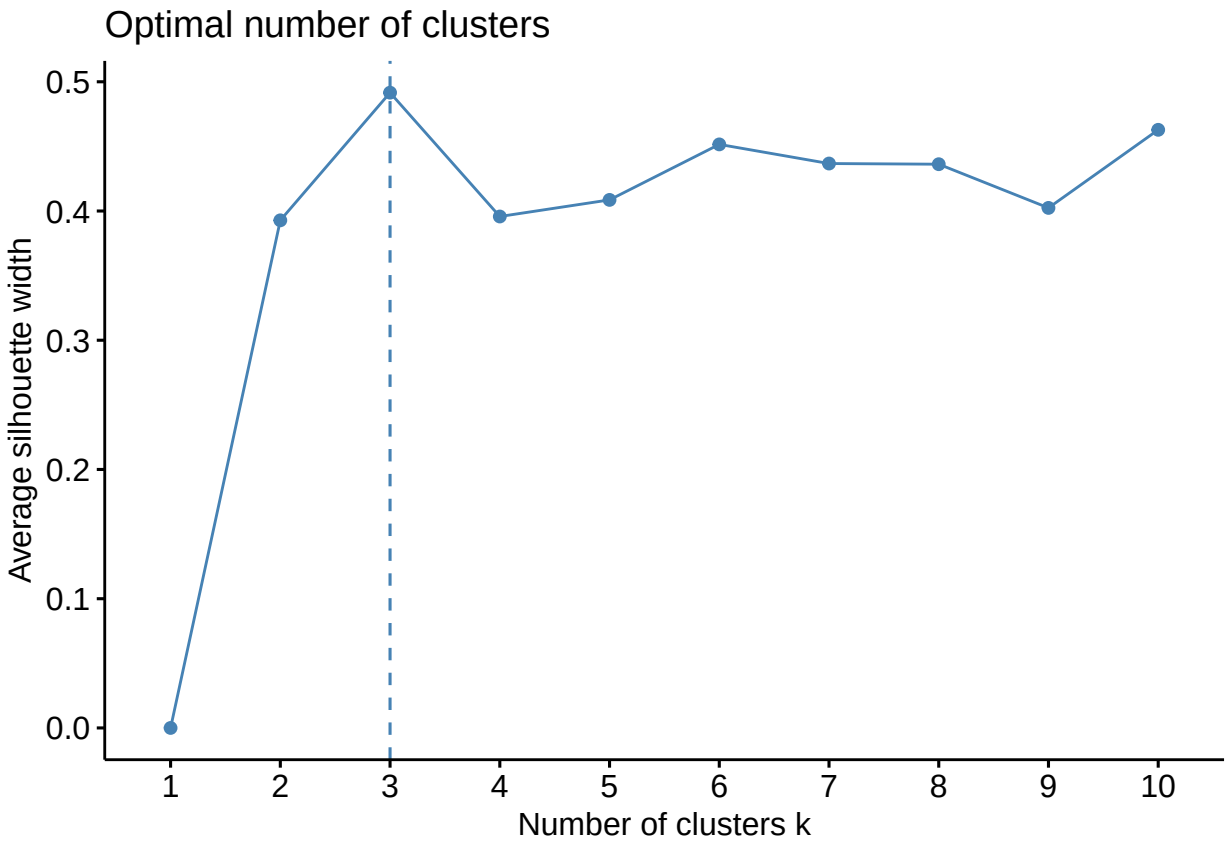
```r
divorcios_anual <- divorcios_depto %>%
  filter(
    nivel_geo == "departamento",
    is.na(mes),
    !is.na(departamento)
  ) %>%
  mutate(departamento = norm_depto(departamento)) %>%
  group_by(anio, departamento) %>%
  summarise(divorcios = sum(valor, na.rm = TRUE), .groups = "drop")

matrimonios_divorcios_depto_anual <- full_join(
  matrimonios_anual,
  divorcios_anual,
  by = c("anio", "departamento")
) %>%
  mutate(
    ratio_div_matr = divorcios / matrimonios
  )

depto_stats <- matrimonios_divorcios_depto_anual %>%
  group_by(departamento) %>%
  summarise(
    avg_matrimonios = mean(matrimonios, na.rm = TRUE),
    avg_divorcios   = mean(divorcios, na.rm = TRUE),
    ratio_div_matr  = avg_divorcios / avg_matrimonios,
    .groups = "drop"
  )

dept_labels <- depto_stats$departamento
depto_stats_num <- depto_stats %>% select(-departamento)
depto_stats_num_scaled <- scale(depto_stats_num)
```

```r
fviz_nbclust(depto_stats_num_scaled, kmeans, method = "wss")
```



Utilizando el método del codo, se logra identificar que el punto en donde se identifica que el gráfico deja de ser pronunciado y empieza a estabilizarse es en 3.

```
fviz_nbclust(depto_stats_num_scaled, kmeans, method = "silhouette")
```



Para garantizar una selección adecuada de clusters, se decidió hacer la prueba por método de la silueta. Al realizarlo, vemos que el punto óptimo señalado es 3. Tomando en consideración ambos métodos, para nuestro clustering utilizaremos k=3.

Agrupamiento y Tendencia

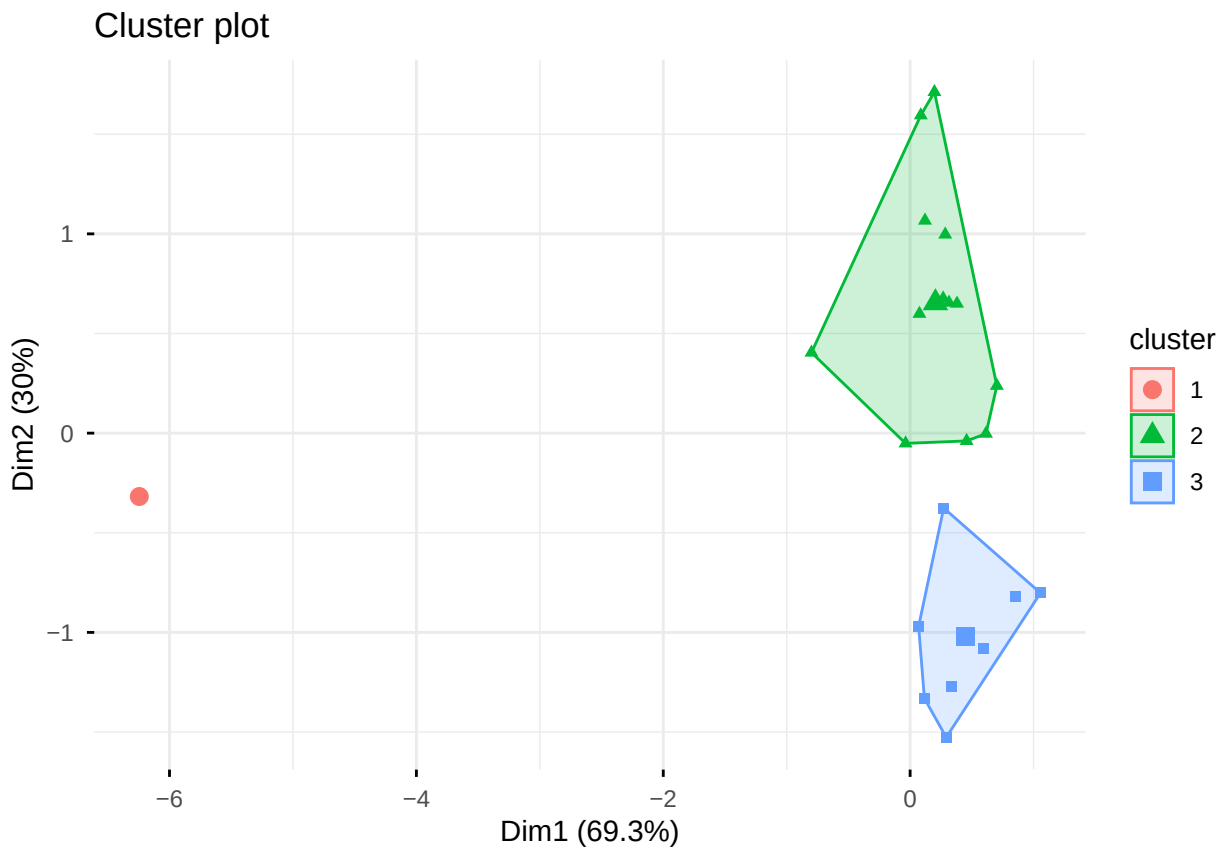
Para este caso se estará trabajando con k-means para agrupar los elementos.

```
set.seed(56)
kmeans_model <- kmeans(depto_stats_num_scaled, centers = 3, nstart = 25)
sil <- silhouette(kmeans_model$cluster, dist(depto_stats_num_scaled))
mean(sil[, 3])
```

```
## [1] 0.4915747
```

Para evaluar la tendencia al agrupamiento se encontró el coeficiente promedio de silueta para k=3 y este fue de 0.4915, lo cual indica una separación fuerte. Esto indica que nuestros datos pueden ser separados en clusters.

```
fviz_cluster(kmeans_model, data = depto_stats_num_scaled,
             geom = "point",
             ellipse.type = "convex",
             ggtheme = theme_minimal())
```




```
depto_clusters <- depto_stats %>%
  mutate(cluster = kmeans_model$cluster) %>%
  arrange(cluster, desc(ratio_div_matr))

depto_clusters
```

```
## # A tibble: 22 x 5
##    departamento   avg_matrimonios avg_divorcios ratio_div_matr cluster
##    <chr>                    <dbl>         <dbl>          <dbl>   <int>
##  1 guatemala                15282         2272.          0.149       1
##  2 el progreso                860          109.          0.127       2
##  3 zacapa                    1213.         151.          0.124       2
##  4 izabal                    1741          186.          0.107       2
##  5 jalapa                    1463.         149.          0.102       2
##  6 quetzaltenango            4607.         459.          0.0996      2
##  7 jutiapa                   2434.         225.          0.0923      2
##  8 retalhuleu                1898.         174           0.0917      2
##  9 santa rosa                1816.         164.          0.0902      2
## 10 chiquimula                1672.         149.          0.0891      2
## # i 12 more rows
```

En este caso se obtuvieron tres grupos. El primer grupo contiene únicamente al departamento de Guatemala, el cual presenta valores promedio de matrimonios y divorcios considerablemente más altos que el resto. En particular, registra aproximadamente 15,282 matrimonios promedio y 2,272 divorcios promedio. En contraste, los demás departamentos presentan promedios notablemente menores (aprox. 800 a 6,000 matrimonios y 65 a 458 divorcios). Esto sugiere que Guatemala se comporta de manera diferenciada.

Calidad de Agrupamiento

```
summary(sil)
```

```
## Silhouette of 22 units in 3 clusters from silhouette.default(x = kmeans_model$cluster, dist = dist(d
##  Cluster sizes and average silhouette widths:
##         1        13         8
## 0.0000000 0.4761679 0.5780577
## Individual silhouette widths:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.3491  0.5986  0.4916  0.6574  0.7148
```

El coeficiente promedio de silueta obtenido fue 0.4916 lo que indica un agrupamiento aceptable. Al analizar por cluster, se observa que un grupo presenta un valor promedio de 0.4762 (cluster bueno), otro grupo de 0.5781 (cl.uster muy bueno), y otro que está en cero. Este tercer cluster presenta un valor promedio de 0.00, pero esto se debe a que es un cluster unitario, y es que este elemnto pertenece a un caso atípico que es el departamento de Guatemala, el cual tiene magnitudes bastante superiores a las presentadas a comparación de todos los demás departamentos, por lo que fue separado a un grupo propio.

Interpretación de grupos

Cluster 1

Aquí solo tenemos a Guatemala es el caso atípico, tenemos promedios muchos más altos en matrimonios y divorcios, así como un ratio mayor al de los demás con 0.149.

Cluster 2

Tiene una proporción media alta de divorcios con respecto a matrimonios (entre 0.065 y 0.127).

Departamentos pertenecientes:

- El Progreso
- Zacapa
- Izabal
- Jalapa
- Quetzaltenango
- Jutiapa
- Retalhuleu
- Santa Rosa
- Chiquimula
- Escuintla
- Baja Verapaz
- Petén
- Sacatepéquez

Cluster 3

Departamentos con una baja proporción de divorcios respecto a matrimonios (valores menores a 0.059). Cabe destacar que en este grupo aparecen varios departamentos con algunos de los promedios más altos de matrimonios como Huehuetenango y Alta Verapaz.

- Suchitepéquez
- San Marcos
- Totonicapán
- Huehuetenango
- Chimaltenango
- Sololá
- Quiché
- Alta Verapaz

Preguntas de Investigación

1. ¿Se ha observado un cambio en la edad promedio en la que las personas contraen matrimonio en Guatemala, indicando que se casan a mayor edad?

Se puede observar que se compararon las combinaciones de edad más frecuentes en matrimonios entre los periodos 2009-2012 y 2019-2022. Los resultados muestran que en ambos periodos predominan las parejas en donde el hombre y la mujer tienen entre 20 y 24 años, seguidas por los rangos entre 25 y 29 años. Aunque en el periodo de 2019-2022 se pueden observar algunos rangos ligeramente mayores estos no superan a los grupos principales. Por lo tanto no se identifica un cambio significativo hacia edades más altas en los matrimonios. Por consiguiente no se observa un cambio en la edad promedio en que las personas contraen matrimonio en Guatemala.

```
#Periodos matrimonios
matr_periodos <- matrimonios_edad_limpio %>%
  mutate(
    periodo = case_when(
      anio >= 2009 & anio <= 2012 ~ "2009-2012",
      anio >= 2019 & anio <= 2022 ~ "2019-2022",
      TRUE ~ NA_character_
    )
  ) %>%
  filter(!is.na(periodo))

# tabla de matrimonios
tabla_top10 <- matr_periodos %>%
  group_by(periodo, edad_mujer_grupo, edad_hombre_grupo) %>%
  summarise(total_matrimonios = sum(valor, na.rm = TRUE), .groups = "drop") %>%
  arrange(periodo, desc(total_matrimonios)) %>%
  group_by(periodo) %>%
  slice_head(n = 10)

tabla_top10
```

```
## # A tibble: 20 x 4
## # Groups:   periodo [2]
##    periodo   edad_mujer_grupo edad_hombre_grupo total_matrimonios
##    <chr>     <chr>            <chr>                         <dbl>
##  1 2009-2012 20 - 24          20-24                         45282
##  2 2009-2012 15 - 19          20-24                         41155
##  3 2009-2012 20 - 24          25-29                         28886
##  4 2009-2012 15 - 19          15-19                         24695
##  5 2009-2012 25 - 29          25-29                         20847
##  6 2009-2012 25 - 29          30-34                         13096
##  7 2009-2012 15 - 19          25-29                         11826
##  8 2009-2012 25 - 29          20-24                          9809
##  9 2009-2012 30 - 34          30-34                          8947
## 10 2009-2012 20 - 24          30-34                          8003
## 11 2019-2022 20 - 24          20-24                         51572
## 12 2019-2022 20 - 24          25-29                         34679
## 13 2019-2022 25 - 29          25-29                         29346
## 14 2019-2022 Menos de 20      20-24                         18887
## 15 2019-2022 25 - 29          30-34                         17253
## 16 2019-2022 30 - 34          30-34                         12785
## 17 2019-2022 25 - 29          20-24                         11231
## 18 2019-2022 20 - 24          30-34                          8967
## 19 2019-2022 30 - 34          35-39                          7950
## 20 2019-2022 18 - 19          20-24                          7497
```

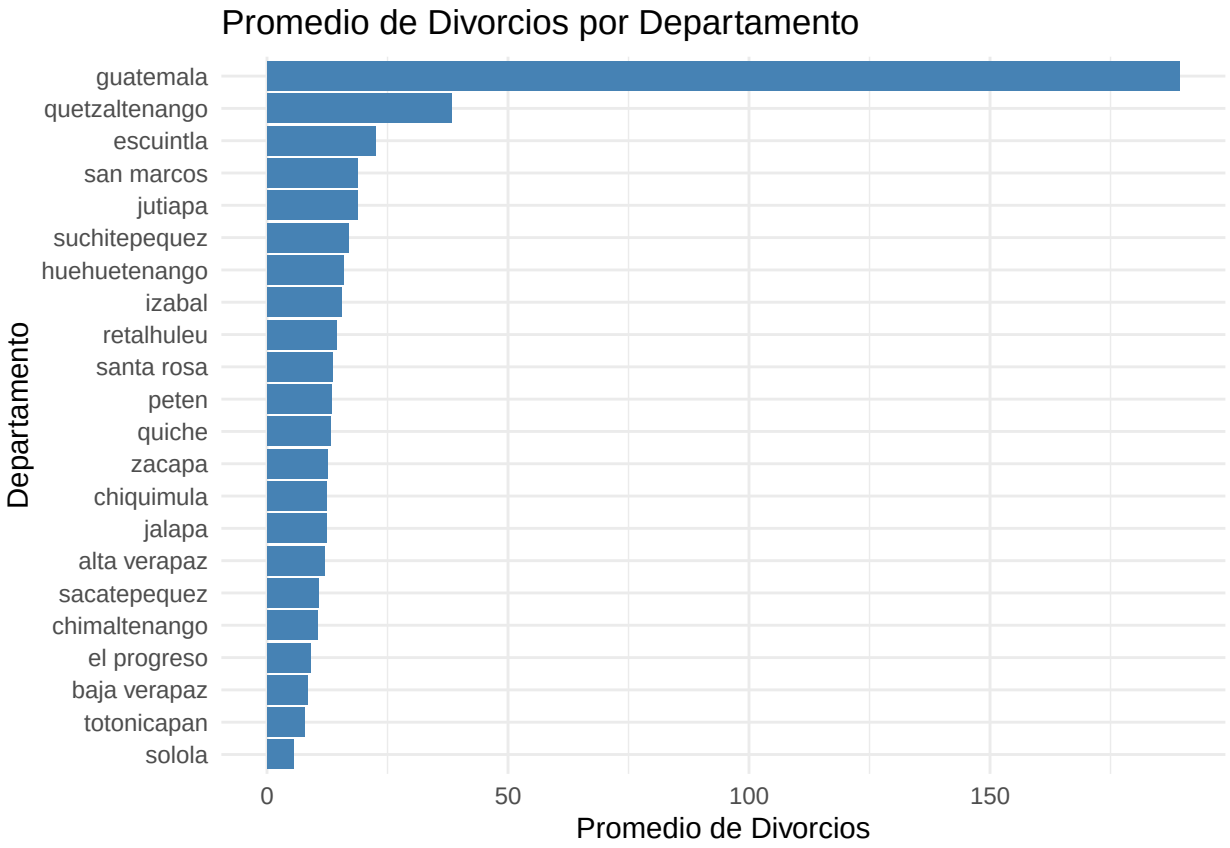
---

2. ¿La tasa promedio de divorcios por cada 100 matrimonios es mayor en el departamento de Guatemala que en el interior del país?

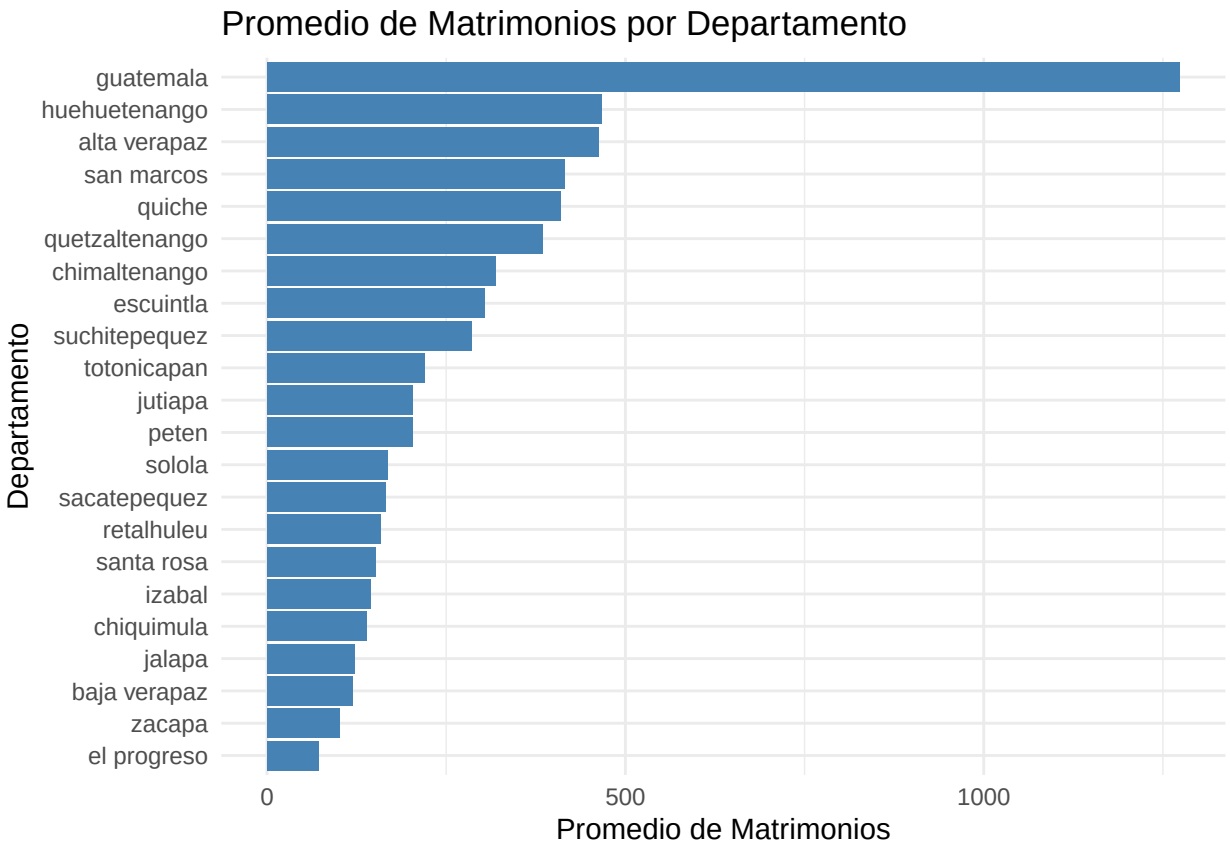
- La taza de divorcios por cada 100 matrimonios es mucho mayor en la Ciudad de Guatemala a comparacion del interior del país, esto siendo una diferencia casi del doble con 14.87 en guatemala en promedio y 6.03 en el interior de promedio, esto posiblemente nos puede indicar que en la capital el divorcio no es tan mal visto a comparacion del interior.
- Adicionalmente, para descartar posibles desequilibrios por los promedios, si observamos los graficos de barra podemos confirmar estos numeros viendo que incluso los divorcios de Quetzaltenango no llegan ni siquiera cerca de los que ocurren el la Capital.

```
divorcios_por_depto <- divorcios_depto_clean |>
  mutate(departamento = stri_trans_general(departamento, "Latin-ASCII")) |>  # quita tildes
  group_by(departamento) |>
  summarise(promedio_divorcios = mean(valor, na.rm = TRUE), .groups = "drop") |>
  arrange(desc(promedio_divorcios))

matrimonios_por_depto <- matrimonios_depto_clean |>
  mutate(departamento = stri_trans_general(departamento, "Latin-ASCII")) |>  # quita tildes
  group_by(departamento) |>
  summarise(promedio_matrimonios = mean(valor, na.rm = TRUE), .groups = "drop") |>
  arrange(desc(promedio_matrimonios))

ggplot(divorcios_por_depto, aes(x = reorder(departamento, promedio_divorcios), y = promedio_divorcios)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +
```

```
  labs(title = "Promedio de Divorcios por Departamento", x = "Departamento", y = "Promedio de Divorcios
  theme_minimal()
```




```
ggplot(matrimonios_por_depto, aes(x = reorder(departamento, promedio_matrimonios), y = promedio_matrimo
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +
  labs(title = "Promedio de Matrimonios por Departamento", x = "Departamento", y = "Promedio de Matrimo
  theme_minimal()
```




```
# promedio de guatemala vs el resto del pais
divorcios_guatemala <- divorcios_depto_clean |>
  filter(departamento == "guatemala") |>
  summarise(promedio_divorcios_guatemala = mean(valor, na.rm = TRUE))

divorcios_resto <- divorcios_depto_clean |>
  filter(departamento != "guatemala") |>
  summarise(promedio_divorcios_resto = mean(valor, na.rm = TRUE))

matrimonios_guatemala <- matrimonios_depto_clean |>
  filter(departamento == "guatemala") |>
  summarise(promedio_matrimonios_guatemala = mean(valor, na.rm = TRUE))

matrimonios_resto <- matrimonios_depto_clean |>
  filter(departamento != "guatemala") |>
  summarise(promedio_matrimonios_resto = mean(valor, na.rm = TRUE))

# tasa de divorcios por cada 100 matrimonios
tasa_divorcios_guatemala <- (divorcios_guatemala$promedio_divorcios_guatemala / matrimonios_guatemala$pr
tasa_divorcios_resto <- (divorcios_resto$promedio_divorcios_resto / matrimonios_resto$promedio_matrimoni

print(paste("Tasa de Divorcios por cada 100 Matrimonios en Guatemala:", round(tasa_divorcios_guatemala,
```

```
## [1] "Tasa de Divorcios por cada 100 Matrimonios en Guatemala: 14.87"
```

```
print(paste("Tasa de Divorcios por cada 100 Matrimonios en el resto del país:", round(tasa_divorcios_res
```

```
## [1] "Tasa de Divorcios por cada 100 Matrimonios en el resto del país: 6.03"
```

---

3. ¿Se observa un incremento en el número promedio anual de divorcios en años recientes (2019–2022) frente a hace una década (2009–2012), lo cual podría sugerir cambios sociales?

- Si observamos las tablas y el gráfico de barras podemos observar que los divorcios de 2019-2022 han incrementado el doble que hace 10 años (2009-2012) lo cual posiblemente nos podría indicar que sí ha habido un cambio social sobre la aceptación del divorcio.

```
# divorcios de 2019 a 2022
divorcios_2019_2022 <- divorcios_depto_clean |>
  filter(anio >= 2019 & anio <= 2022) |>
  group_by(anio) |>
  summarise(total_divorcios = sum(valor, na.rm = TRUE), .groups = "drop")

# divorcios de 2009 a 2012
divorcios_2009_2012 <- divorcios_depto_clean |>
  filter(anio >= 2009 & anio <= 2012) |>
  group_by(anio) |>
  summarise(total_divorcios = sum(valor, na.rm = TRUE), .groups = "drop")

divorcios_2019_2022
```

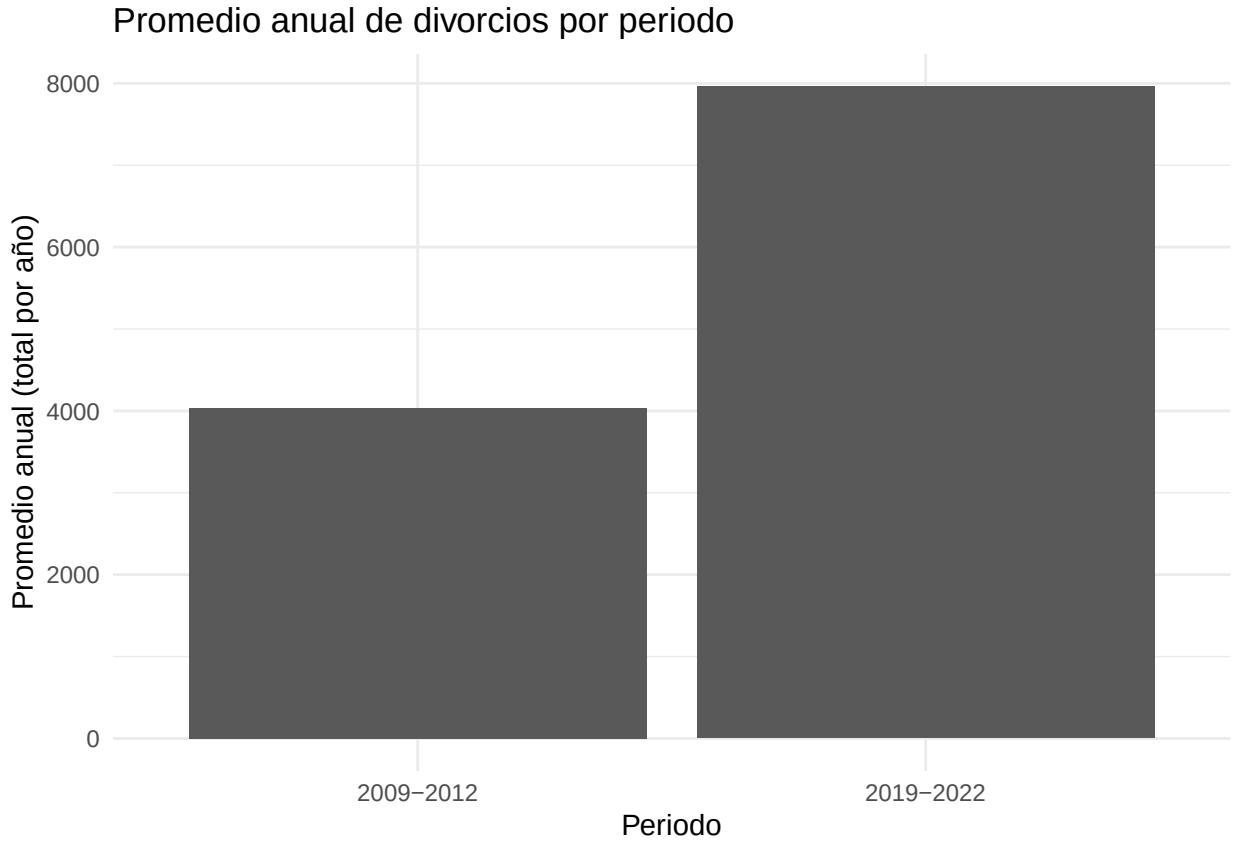
```
## # A tibble: 4 x 2
##    anio total_divorcios
##   <dbl>           <dbl>
## 1  2019            8203
## 2  2020            4074
## 3  2021            9621
## 4  2022            9950
```

```
divorcios_2009_2012
```

```
## # A tibble: 4 x 2
##    anio total_divorcios
##   <dbl>           <dbl>
## 1  2009            3004
## 2  2010            3645
## 3  2011            4344
## 4  2012            5157
```

```
# promedios por anios
div_periodo <- divorcios_depto_clean |>
  filter((anio >= 2009 & anio <= 2012) | (anio >= 2019 & anio <= 2022)) |>
  group_by(anio) |>
  summarise(total_anual = sum(valor, na.rm = TRUE), .groups = "drop") |>
```

```r
  mutate(periodo = ifelse(anio <= 2012, "2009-2012", "2019-2022")) |>
  group_by(periodo) |>
  summarise(promedio_anual = mean(total_anual), .groups = "drop")

ggplot(div_periodo, aes(x = periodo, y = promedio_anual)) +
  geom_col() +
  labs(title = "Promedio anual de divorcios por periodo",
       x = "Periodo", y = "Promedio anual (total por año)") +
  theme_minimal()
```



4.¿Los departamentos con menor cantidad de matrimonios presentan una proporción considerablemente menor de divorcios?

Contrario a la hipótesis inicial, al agrupar los departamentos según su volumen total de matrimonios (2009–2022), se observa que el grupo de menor volumen presenta la tasa promedio de divorcios más alta (12.63 divorcios por cada 100 matrimonios), mientras que el grupo de mayor volumen presenta la más baja (6.12).

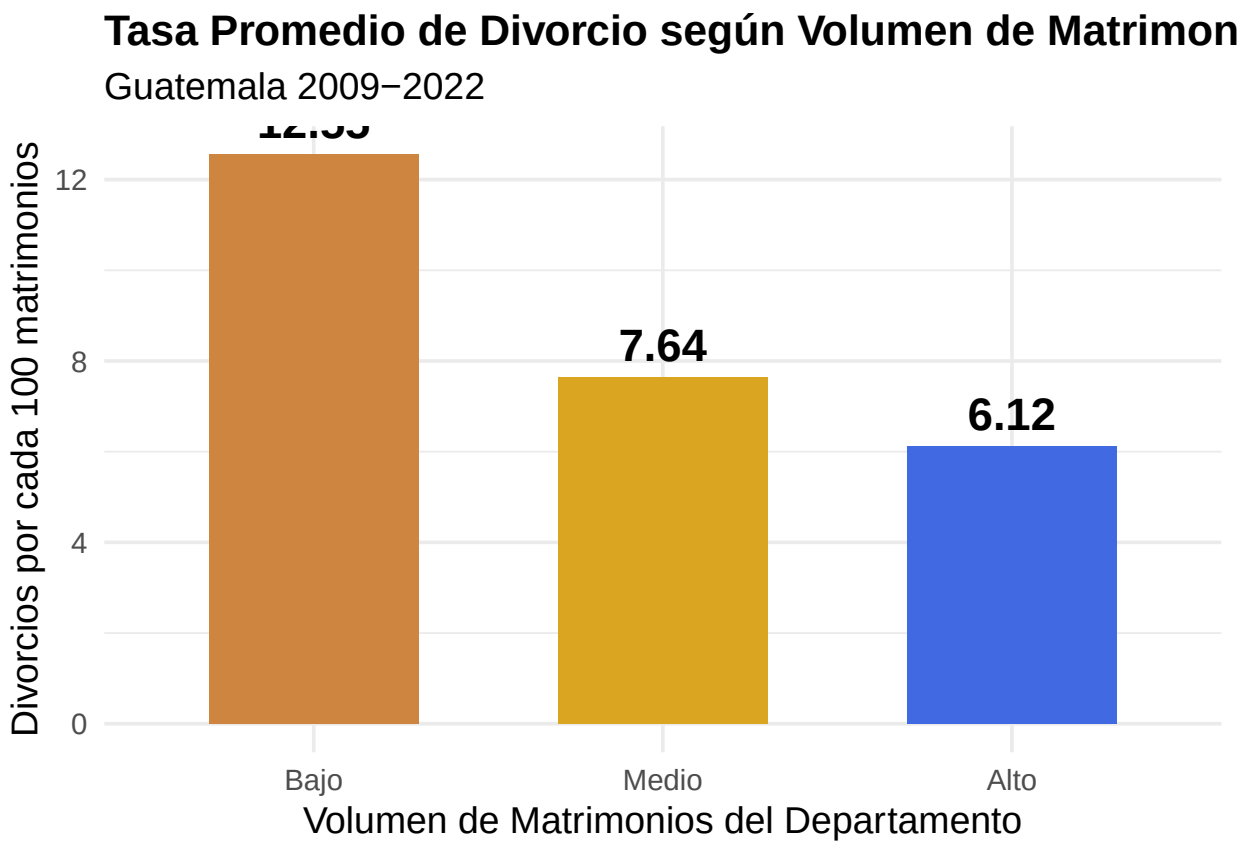
Esto contradice la creencia de que donde hay menos matrimonios también hay proporcionalmente menos divorcios. Sin embargo, este resultado debe interpretarse con cautela, ya que el grupo “Bajo” contiene únicamente 2 departamentos y la categorización se basa en volumen de matrimonios, no en un indicador directo de urbanización.

```r
# Calcular totales y tasas por departamento
analisis_rural_urbano <- matrimonios_depto %>%
  filter(nivel_geo == "departamento", is.na(mes)) %>%
  group_by(departamento) %>%
  summarise(total_matrimonios = sum(valor, na.rm = TRUE), .groups = "drop") %>%
  left_join(
    divorcios_depto %>%
      filter(nivel_geo == "departamento", is.na(mes)) %>%
      group_by(departamento) %>%
      summarise(total_divorcios = sum(valor, na.rm = TRUE), .groups = "drop"),
    by = "departamento"
  ) %>%
  mutate(
    total_divorcios = tidyr::replace_na(total_divorcios, 0),
    tasa_divorcio_por_100 = (total_divorcios / total_matrimonios) * 100,
    categoria = case_when(
      total_matrimonios > 40000 ~ "Alto",
      total_matrimonios > 20000 ~ "Medio",
      TRUE ~ "Bajo"
    ),
    categoria = factor(categoria, levels = c("Bajo", "Medio", "Alto"))
  )

# Tabla resumen
tabla_resumen <- analisis_rural_urbano %>%
  group_by(categoria) %>%
  summarise(
    n_departamentos = n(),
    promedio_tasa_divorcio = round(mean(tasa_divorcio_por_100), 2),
    .groups = "drop"
  )

print(tabla_resumen)
```

```
## # A tibble: 3 x 3
##   categoria n_departamentos promedio_tasa_divorcio
##   <fct>               <int>                  <dbl>
## 1 Bajo                    2                   12.6
## 2 Medio                  11                    7.64
## 3 Alto                    9                    6.12
```

```r
# Gráfico de barras
ggplot(tabla_resumen, aes(x = categoria, y = promedio_tasa_divorcio, fill = categoria)) +
  geom_bar(stat = "identity", width = 0.6) +
  geom_text(aes(label = promedio_tasa_divorcio),
            vjust = -0.5, size = 6, fontface = "bold") +
  labs(
    title = "Tasa Promedio de Divorcio según Volumen de Matrimonios",
    subtitle = "Guatemala 2009-2022",
    x = "Volumen de Matrimonios del Departamento",
    y = "Divorcios por cada 100 matrimonios"
  ) +
  theme_minimal(base_size = 14) +
```

```
  scale_fill_manual(values = c("Bajo" = "#CD853F", "Medio" = "#DAA520", "Alto" = "#4169E1")) +
  theme(legend.position = "none",
        plot.title = element_text(face = "bold", size = 16))
```



5. ¿Sera que los departamentos más frios tienden a tener una proporción de divorcios más bajos respecto al matrimonio?

Al momento que llegamos a ver los clusters podemos ver una tendencia curiosa, aquellos que están en el cluster tres son departamentos frios respecto a todos los demas.

Para eso podriamos juntar solo aquellos valores que son parte de estos dos clusters, y viendo cuales son los valores mas bajos entre ellos y si son considerados de lugares los cuales normalmente se consideran mas frios en el pais de guatemala.

```
cluster_data_set = depto_clusters

cluster_data_set = cluster_data_set[order(cluster_data_set$ratio_div_matr),]

ratio_df = cluster_data_set[c("ratio_div_matr", "departamento")]

ratio_df
```

```
## # A tibble: 22 x 2
##    ratio_div_matr departamento
##             <dbl> <chr>
```

```
##  1         0.0258 alta verapaz
##  2         0.0320 quiche
##  3         0.0321 solola
##  4         0.0327 chimaltenango
##  5         0.0342 huehuetenango
##  6         0.0354 totonicapan
##  7         0.0452 san marcos
##  8         0.0590 suchitepequez
##  9         0.0645 sacatepequez
## 10         0.0661 peten
## # i 12 more rows
```

Aquí ya empezamos a mirar una tendencia, por ejemplo los 5 departamentos con la proporcion de divorcios matrimonio son de lugares frios como Alta Verapaz, Solola, Chimaltenango, Huehuetenango, Totonicapan y San Marcos.

Ahora lo que podemos hacerlo es separarlos en departamentos frios y calientes solo entre 2 y el 3.

```
  # Vector de departamentos fríos
departamentos_frios <- c("quiche",
                         "solola",
                         "chimaltenango",
                         "huehuetenango",
                         "totonicapan",
                         "quetzaltenango",
                         "san marcos",
                         "alta verapaz",
                         "baja verapaz",
                         "sacatepequez")

dept_frios = filter(ratio_df,departamento %in% departamentos_frios)

resto_depts = filter(ratio_df,!departamento %in% departamentos_frios)

dept_frios
```

```
## # A tibble: 10 x 2
##    ratio_div_matr departamento
##             <dbl> <chr>
##  1         0.0258 alta verapaz
##  2         0.0320 quiche
##  3         0.0321 solola
##  4         0.0327 chimaltenango
##  5         0.0342 huehuetenango
##  6         0.0354 totonicapan
##  7         0.0452 san marcos
##  8         0.0645 sacatepequez
##  9         0.0706 baja verapaz
## 10         0.0996 quetzaltenango
```

```
resto_depts
```

```
## # A tibble: 12 x 2
```

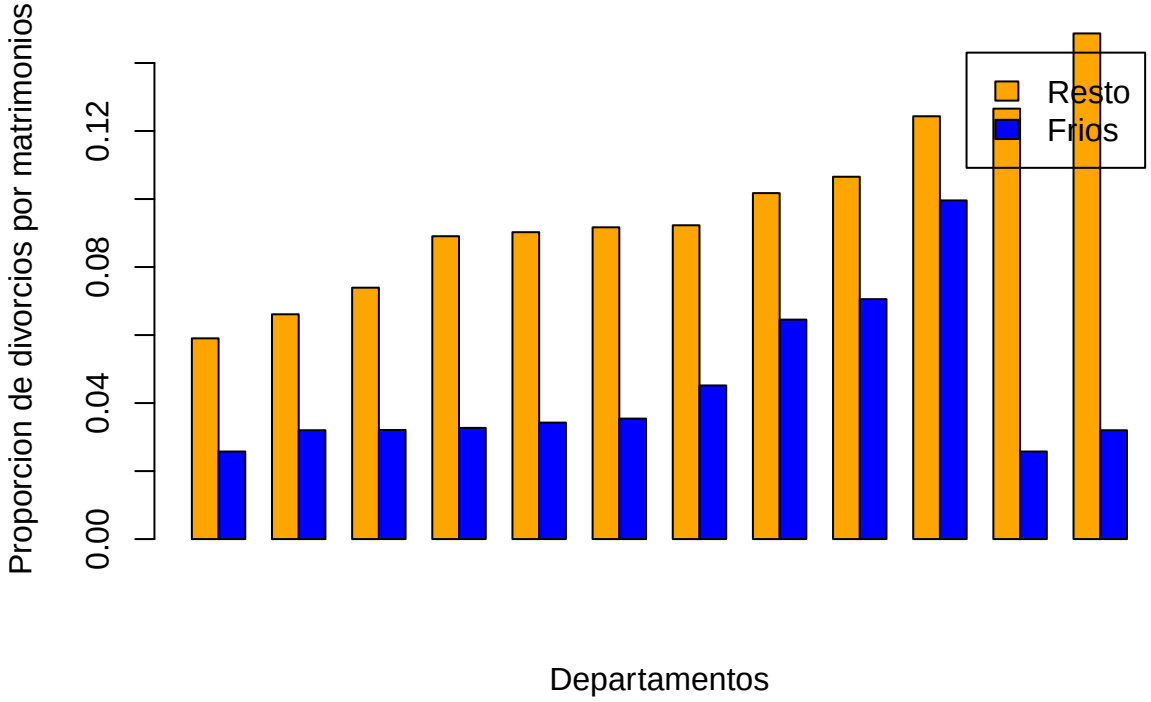
```
##    ratio_div_matr departamento
##             <dbl> <chr>
##  1         0.0590 suchitepequez
##  2         0.0661 peten
##  3         0.0739 escuintla
##  4         0.0891 chiquimula
##  5         0.0902 santa rosa
##  6         0.0917 retalhuleu
##  7         0.0923 jutiapa
##  8         0.102  jalapa
##  9         0.107  izabal
## 10         0.124  zacapa
## 11         0.127  el progreso
## 12         0.149  guatemala
```

Ahora que ya tenemos los dos grupos separados, ahora 'plotemos' estos valores uno encima de otro, para ver cua de los dos grupos tienden a tener una proporcion de divorcios respecto a matrimonios mejor (mas bajo).osea

```r
# Combine ratios into a matrix
datos <- rbind(
  Resto = resto_depts$ratio_div_matr,
  Frios = dept_frios$ratio_div_matr
)
```

```
## Warning in rbind(Resto = resto_depts$ratio_div_matr, Frios =
## dept_frios$ratio_div_matr): number of columns of result is not a multiple of
## vector length (arg 2)
```

```r
# Create grouped barplot
barplot(
  datos,
  beside = TRUE,
  col = c("orange", "blue"),
  xlab = "Departamentos",
  ylab = "Proporcion de divorcios por matrimonios",
  legend.text = TRUE
)
```



Como podemos llegar a ver si llega a ver esa tendencia, quitando la capital los departamentos mas frios tienden a tener una proporcion de divorcios por matrimonios bastante menor.

Hallazgos y Conclusiones

Resumen de Hallazgos en Análisis Exploratorio

Estructura del Conjunto de Datos

- Se trabajó con 4 datasets (2009–2022): matrimonios por departamento/mes, matrimonios por rangos de edad, divorcios por departamento/mes y divorcios por rangos de edad.
- Cada dataset contiene una variable clave de conteo (valor) que representa la cantidad de eventos (matrimonios o divorcios) para una combinación de variables categóricas (año/mes/departamento/edad).

Variables Cuantitativas

- Se analizó principalmente valor, ya que las demás variables numéricas (año/mes) funcionan como identificadores temporales y no como magnitudes interpretables para tendencia central/dispersión.
- En general, valor muestra asimetría positiva (media > mediana) y presencia de valores atípicos, especialmente en departamentos con mayor concentración poblacional.

Distribución y Normalidad

- Los histogramas y boxplots muestran concentración de valores bajos con colas largas (outliers) en los cuatro datasets.
- La prueba de Lilliefors rechaza normalidad para valor en los datasets analizados (p-value < 0), confirmando que no siguen distribución normal.

Variables Cualitativas

- Por departamento, Guatemala domina tanto en matrimonios como divorcios, mostrando una diferencia marcada frente al resto.
- Los matrimonios se concentran principalmente en edades jóvenes adultas (por ejemplo, 20–24 y 25–29).
- Los divorcios tienden a concentrarse en rangos más altos que los matrimonios (por ejemplo, 25–29, 30–34 y 35–39, dependiendo del sexo).

Relaciones entre variables

- En el análisis por edades (gráfico de burbujas), se observan concentraciones fuertes en combinaciones específicas de edades
- La relación entre total anual de matrimonios y divorcios muestra una correlación positiva moderada (cercano a 0.60): cuando hay más matrimonios en un año, también suelen registrarse más divorcios, aunque con variabilidad.

Resumen de Clustering

Cluster 1 — "Outlier (Guatemala)"

- Característica principal: promedios de matrimonios y divorcios muy superiores al resto, y ratio relativamente alto.

Cluster 2 — "Alta proporción de divorcios"

- Característica principal: ratio divorcios/matrimonios medio–alto (aprox. 0.065 a 0.127).

Cluster 3 — "Baja proporción de divorcios" - Característica principal: ratio divorcios/matrimonios bajo (menor a ~0.059), aunque algunos tienen altos promedios de matrimonios.

Conclusiones con Base a Objetivos

Analizar los datos de matrimonios y divorcios en Guatemala para identificar patrones demográficos, diferencias entre departamentos y cambios en la dinámica a lo largo del período.

- El análisis exploratorio muestra patrones claros de concentración tanto por edad como por departamento, por lo que sí existen diferencias relevantes entre regiones y grupos etarios.
- A nivel temporal, los datos sugieren variaciones en la dinámica de divorcios al comparar periodos (por ejemplo, 2019–2022 vs 2009–2012), lo cual es consistente con la hipótesis de cambios sociales en años recientes.

- En la relación global, se observó una asociación moderada entre matrimonios y divorcios, en años con más matrimonios, tienden a registrarse más divorcios, aunque con variabilidad.

Examinar la distribución de matrimonios y divorcios según rangos de edad, identificando variaciones en las edades donde ocurren estos eventos.

- Los matrimonios no se distribuyen uniformemente entre rangos de edad: se concentran en rangos específicos (por ejemplo, en el análisis de frecuencias, los mayores registros se ubican en edades jóvenes-adultas como 20–24 y 25–29).
- Los divorcios tienden a concentrarse en rangos mayores respecto a los matrimonios (por ejemplo, rangos como 25–29, 30–34 y 35–39 aparecen con alta presencia), lo que sugiere que el divorcio ocurre con mayor frecuencia después de algunos años de vida en pareja.

Analizar la distribución de matrimonios y divorcios por departamento para detectar patrones regionales relevantes.

- Existe una concentración fuerte de eventos en el departamento de Guatemala, que destaca ampliamente tanto en matrimonios como en divorcios (y por proporciones, especialmente en divorcios).
- Al comparar "capital vs interior", la tasa promedio de divorcios por cada 100 matrimonios fue mayor en Guatemala (14.87) que en el resto del país (6.03), confirmando diferencias regionales relevantes.
- El clustering reforzó la existencia de patrones departamentales:
  - Guatemala se comporta como caso atípico.

Potencial Plan de Pasos a Seguir

Tras analizar los patrones de matrimonios y divorcios en Guatemala (2009–2022) a nivel geográfico y demográfico, se identificó que el departamento de Guatemala presenta un comportamiento marcadamente distinto al resto. Por ello, un siguiente paso sería profundizar su análisis y formular hipótesis explicativas (por ejemplo, diferencias en urbanización, acceso a servicios legales, concentración económica, etc), para luego contrastarlas con evidencia adicional.

En cuanto a los matrimonios, los registros se concentran principalmente en el rango de 20 a 30 años, por lo que conviene explorar con más detalle si esta concentración tiene factores económicos detrás, como que la mayoría de personas empiezan a laburar en esta rango, o si existen más factores.

Respecto a los divorcios, sería relevante analizar su evolución en años recientes más a fondo, con el fin de comprender qué cambios culturales han ocurrido que hagan que se presente este fenómeno.
